# ANÁLISE DA ESTABILIDADE TERMODINÂMICA E DE PARÂMETROS ESTRUTURAIS DE DNA E RNA POR MODELOS MESOSCÓPICOS

TAUANNE DIAS AMARANTE

# ANÁLISE DA ESTABILIDADE TERMODINÂMICA E DE PARÂMETROS ESTRUTURAIS DE DNA E RNA POR MODELOS MESOSCÓPICOS

Tese apresentada ao Programa de Pós-Graduação em Física do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Física.

Orientador: Gerald Weber

Belo Horizonte

Março de 2015

Amarante, Tauanne Dias

D1234p Análise da estabilidade termodinâmica e de parâmetros estruturais de DNA e RNA por modelos mesoscópicos / Tauanne Dias Amarante. — Belo Horizonte, 2015
xvi, 83 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais
Orientador: Gerald Weber

1. Física — Teses. 2. — Teses. I. Orientador. II. Título.

CDU 519.6*82.10

# [Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

Se a imagem da folha de aprovação precisar ser ajustada, use:
`approval=[`*ajuste*`][`*escala*`]{`*nome do arquivo*`}`
onde *ajuste* é uma distância para deslocar a imagem para baixo
e *escala* é um fator de escala para a imagem. Por exemplo:
`approval=[-2cm][0.9]{`*nome do arquivo*`}`
desloca a imagem 2cm para cima e a escala em 90%.

*Aos meus pais, por proporcionarem as condições ideais de temperatura, pressão e raciocínio para o nascimento de uma cientista.*

# Agradecimentos

Ao Gerald, pelo conhecimento, pela amizade, pela paciência e pelo incentivo constante. Pelos ensinamentos, cafés e conversas e pela confiança compartilhada através do olhar cada vez que a minha ousava titubear. O mundo acadêmico seria mais macio e feliz se todos pudessem ter o privilégio de ter um orientador como ele. Com as piadas de bônus.

Aos amigos do grupo e do laboratório, pelo sorriso diário.

Ao Jon Essex, Eugen Stulz e Jeremy Frey pela recepção em Southampton.

Ao Pedro, Frank, Nawel e amigos por terem transformado Southampton em lar.

Aos meus pais que me ensinaram o amor ao saber e à leitura. Ao meu pai por me instigar a argumentar, desde os 2 anos de idade. À minha mãe por mostrar que a única ambição que se deve ter é a de conhecimento.

Ao Juliano por sempre ter respondido as minhas indagações com todo conhecimento que detinha, apesar da nossa diferença de 6 anos de idade.

Ao Claret por me ensinar a ler pessoas e rir das minúcias.

Ao Lorenzo pelo amor e cuidado e por, mesmo fazendo miçanga, ser um discípulo de Popper e prezar pelo conhecimento científico.

À Mabel pela animação contagiante, pela independência modelo e por cada LEGO.

À Kiss e à Bella pelo carinho e pelas pequenas. O existir delas faz o meu ser mais belo.

Ao Saramago, Fernando Pessoa e Jostein Gaarder por terem preservado minha sanidade e minha poesia.

Ao CNPq e à Fapemig pelo financiamento.

*"Ninguém deve contentar-se com o que lhe dizem. Deve averiguar se é verdade.*
*Saber se é a única verdade e compará-la com a verdade dos demais.*
*Há que procurar sempre o outro lado de tudo."*
(José Saramago)

# Resumo

**Palavras-chave:** Modelo Peyrard-Bishop, DNA, RNA.

Modelos mesoscópicos como o proposto por Peyrard e Bishop apresentam uma relevância significativa no estudo da estabilidade térmica de DNA e RNA. A sua simplicidade computacional permite a análise desses oligonucleotideos por períodos mais longos e de importância fisiológica e experimental inacessíveis por técnicas mais sofisticadas. Recentemente houve bastante progresso na parametrização dos pares de bases canônicos CG e AT em DNA e CG e AU em RNA para este modelo. No entanto, por serem mais estáveis do que pares não-canônicos, essas bases são relativamente simples de modelar. Nosso estudo viabilizou a aplicação do modelo PB além das bases canônicas, bem como a investigação de parâmetros estruturais. O primeiro sistema que foi estudado é o de guanina-uracila (GU) em RNA. GU desempenha um papel biológico importante, atuando como um local de reconhecimento para biomoléculas, além de ser o par não complementar mais comum em RNA. Devido à não isostericidade do par GU, sua estabilidade é influenciada pelo contexto da sequência. GU apresenta diversas possibilidades conformacionais, inclusive assumindo um número diferente de ligações hidrogênio dependendo das bases vizinhas. Para considerar todas as possibilidades de contexto de GU em RNA é necessário lidar com uma quantidade muito grande de parâmetros. Esse problema é contornado ao realizarmos a otimização de parâmetros correlacionando com dados de temperatura de desnaturação obtidos da literatura. Desenvolvemos uma estratégia de redução do espaço de busca de parâmetros que possibilitou a determinação de grupos de configuração de GU ordenados por intensidade de ligação de hidrogênio. Em comparação com dados experimentais da literatura obtivemos uma grande concordância nas ligações de hidrogênio, em especial com as técnicas de NMR. Assim, pudemos não apenas concluir a parametrização de GU em RNA mas também demonstrar que a técnica permite prever ligações de hidrogênio de pares de base não-canônicos. No segundo tema nós abordamos a limitação da Hamiltoniana 2D Peyrard-Bishop que não inclui qualquer parâmetro estrutural.

Mostramos que é possível partir de uma formulação Hamiltoniana helicoidal e obter uma Hamiltoniana modificada em 2D que inclui a informação sobre o passo da hélice de DNA, tecnicamente chamado de *rise*. Nesse estudo também usamos o método de otimização correlacionando com dados de desnaturação disponíveis na literatura para obter valores de *rise*. Para realizar a comparação dos nossos resultados com os obtidos por medidas experimentais, desenvolvemos uma ferramenta de pesquisa que acessa a base de dados no Nucleic Acids Database (NDB) e seleciona as sequências de interesse para obter valores de *rise* oriundos de raios-x e NMR. A concordância dos nossos resultados de *rise* foi em geral semelhante com os do NDB com exceção de AT seguido de TA em DNA. Além disto, pudemos estudar a variação do *rise* em função de concentração salina. Os nossos resultados evidenciam a possibilidade de realizar estudos estruturais em oligonucleotideos usando temperaturas de desnaturação.

# Abstract

**Keywords:** Peyrard-Bishop model, DNA, RNA.

Mesoscopic models, like the one proposed by Peyrard and Bishop, are important for the study of the thermal stability of DNA and RNA. Its computational simplicity allows to analyse these oligonucleotides over longer time scales of physiological and experimental importance which are not accessible by more sophisticated techniques. Recently, there has been progress in the parametrization of canonical base pairs for this model. However, as they are more stable than non-canonical pairs, these bases are relatively easy to model. Our study aimed at applying the PB model beyond the canonical bases, as well as to investigate structural parameters. The first system we analysed was guanine-uracil (GU) in RNA. GU has an important biological role, acting as a recognition site for biomolecules and is also the most common non-complementary base pair in RNA. Due to the non-isostericity of GU, its stability is influenced by the sequence context. GU has several conformational possibilities and may even have different hydrogen bonds depending on neighbouring bases. To consider all context possibilities for GU in RNA it becomes necessary to deal with a very large number of parameters. This problem is dealt with by optimizing the parameters correlating with melting temperature data from the literature. We developed a strategy to reduce the parameter search space which allowed us to determine GU configuration groups sorted by hydrogen bond intensity. When comparing with experimental results from the literature we obtain a good agreement for the hydrogen bonds, especially from NMR. Therefore, we not only were able to conclude the parametrization of GU in RNA but also to demonstrate that the technique allows to predict hydrogen bonds for non-canonical base pairs. In the second subject we approached a limitation of the 2D Peyrard-Bishop Hamiltonian which is the lack of structural parameters. We showed that it is possible to start from a helicoidal Hamiltonian and obtain a modified 2D Hamiltonian which includes information about the helix step, technically known as rise. In this study we also use the optimization method where we correlate melting

temperature data from the literature to obtain the values for rise in DNA. To perform the comparison of our results with those obtained from experimental measurements we developed a query tool which accesses the Nucleic Acids Database (NDB) and selects sequences of interest to obtain values of rise coming from X-ray and NMR. In general our results agreed well with those from the NDB except for AT followed by TA in DNA. Furthermore, we were able to study the variation of rise with salt concentration. Our results show the possibility of performing structural studies in oligonucleotides using melting temperatures.

# Lista de Figuras

# Lista de Tabelas

# Sumário

# Capítulo 1

# Introdução

## 1.1 Apresentação

O objetivo deste trabalho é investigar a possibilidade de obter informações de interações intramoleculares de RNA e DNA modelando medidas experimentais de temperaturas de desnaturação por modelos mesoscópicos do tipo Peyrard-Bishop (PB). Em particular, procuramos responder se é possível obter informações, tais como ligações de hidrogênio, sobre guanina-uracila que é um par de base comum em RNA e sujeito a uma diversidade de conformações. O segundo objetivo é determinar se é possível alterar o Hamiltoniano PB e assim obter parâmetros estruturais em DNA, também por medidas de temperatura.

Esta tese está organizada da seguinte forma: no primeiro capítulo introduzimos as principais características da estrutura do DNA e discutimos os experimentos de espectroscopia ultravioleta que permitem acessar o comportamento termodinâmico dessas moléculas. Em seguida, discutimos os diferentes tipos de modelos empregados no estudo de ácidos nucleicos e introduzimos o modelo Peyrard-Bishop original e as várias modificações desse modelo. Por fim, estabelecemos a metodologia adotada que permite obter a parametrização dessa classe de modelos a partir de dados de temperatura de desnaturação.

No capítulo 2, exploramos o estudo do par guanina-uracila em RNA. Primeiramente apresentamos a importância biológica desse par não-complementar em RNA. Posteriormente, discutimos a estrutura do RNA e as geometrias que a conformação dos pares de base podem assumir. Introduzimos a questão da dependência de contexto de sequência da estabilidade do par GU e a divergência das medidas experimentais quanto ao número de ligações de hidrogênio para uma determinada configuração. Finalmente apresentamos como aplicamos o modelo Peyrard-Bishop para abordar esse par não-

canônico e a influência dos pares vizinhos. Por último estabelecemos a comparação entre os parâmetros obtidos e os dados disponíveis na literatura, tanto experimentais quanto teóricos.

No capítulo 3, comentamos a respeito da existência de modelos PB tridimensionais. Subsequentemente, apresentamos o desenvolvimento analítico de um novo modelo 2D do tipo PB incluindo informações estruturais a partir de um modelo helicoidal existente. Usando o mesmo método de equivalência termodinâmica para otimizar os parâmetros a partir de temperaturas de desnaturação, obtivemos os parâmetros associados a esse modelo bidimensional estrutural incluindo o passo da hélice (*rise*). Em seguida, discutimos as técnicas experimentais que fornecem informações estruturais de DNA e os programas computacionais que interpretam esses dados permitindo inferir o valor do *rise*. Por fim realizamos a comparação dos dados de *rise* de medidas de raio-X e NMR extraídos da base de dados NDB e calculados através do software 3DNA com os nossos parâmetros.

Apresentamos as nossas conclusões no capítulo 4 e comentamos brevemente sobre os outros trabalhos realizados, em particular o estudo da estabilidade das terminações de sequências canônicas de RNA e DNA.

## 1.2 Motivação

A compreensão das propriedades termodinâmicas do DNA é de grande importância em aplicações biológicas como, por exemplo, PCR (reação de polimerase em cadeia, técnica utilizada no sequenciamento de DNA) [1], produção de sondas [2], *gene arrays* [3] e polimorfismos de nucleotídeo único (*single nucleotide polymorphisms* ou SNP) [4]. Todas estas técnicas requerem um conhecimento preciso de temperaturas de desnaturação de DNA ou RNA. Os modelos que são objeto desta tese podem prover estas temperaturas e também informações complementares, como localização de abertura de pares de bases. Contudo, as parametrizações ainda são insuficientes para cobrir todos os espectros de aplicações. Por exemplo o par guanina-uracila, que será visto no capítulo 2, ainda não havia sido parametrizado para o modelo Peyrard-Bishop, o que impedia usar esse modelo em aplicações envolvendo RNA natural, como o RNA mensageiro (mRNA).

O modelo em estudo nesta tese também permite a obtenção de informações sobre a interação intramolecular como ligações de hidrogênio. Atualmente as principais técnicas experimentais que medem estas interações diretamente são NMR e raios-X. O desenvolvimento do modelo Peyrard-Bishop para extrair estas informações de medidas de desnaturação oferece a oportunidade de complementar estas técnicas ou até mesmo

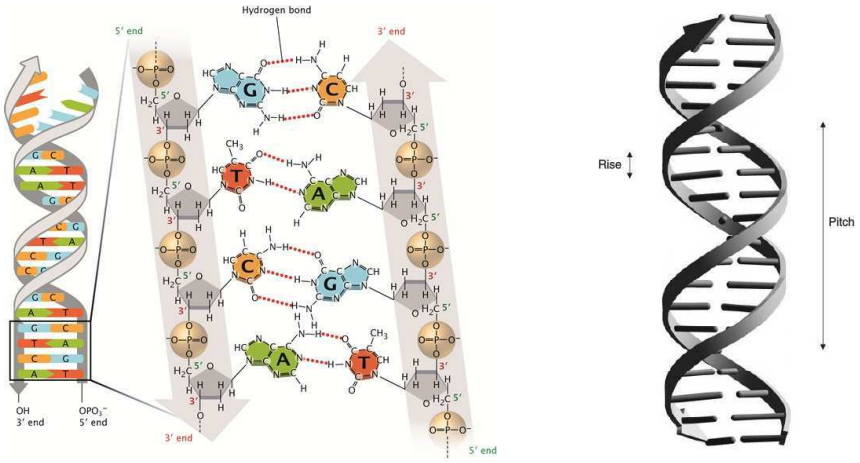



**Figura 1.1**
Esquemas de estrutura do DNA. (a) Exibe a composição do DNA e a direcionalidade das fitas do DNA: as diferentes extremidades de uma fita são chamadas *3'-end* e *5'-end* e se referem ao átomo de carbono da desoxirribose ao qual o fosfato está ligado. (b) Mostra a geometria helicoidal e destaca o passo de hélice (*rise*) que no B-DNA possui valor médio de 3,4 Å. Figuras retiradas respectivamente de Pray [5] e de Neidle [6].

produzir informações inéditas. Esta parte é explorada nos capítulos 2 e 3.

## 1.3 A estrutura do DNA

O ácido desoxirribonucleico (DNA) é um polímero cujos monômeros são os nucleotídeos. Cada nucleotídeo que o compõe contém um grupo fosfato, um açúcar do tipo desoxirribose e uma das quatro bases nitrogenadas: adenina (A), timina (T), citosina (C) e guanina (G) (figura 1.1). Essas bases são classificadas em dois grupos, T e C são chamadas pirimidinas enquanto A e G são purinas. A estrutura de DNA mais estável consiste na hélice dupla dextrógira denominada B-DNA. As fitas que formam a dupla hélice são unidas através de ligações de hidrogênio entre as bases que as compõem, a saber, o par de base AT é formado por duas ligações e o par CG é formado por três ligações de hidrogênio.

Existe uma variedade de interações possíveis entre os pares de bases dos ácidos nucleicos. Há uma classificação que considera apenas aquelas que formam ao menos duas ligações de hidrogênio entre os pares de base que os organiza em doze famílias. Os pares complementares AT e CG apresentados estão classificados na família Watson-Crick e são isostéricos entre si. Dizer que pares de base são isostéricos entre si significa que as posições e distâncias entre os carbonos C1′ são muito similares [7]. Na estru-

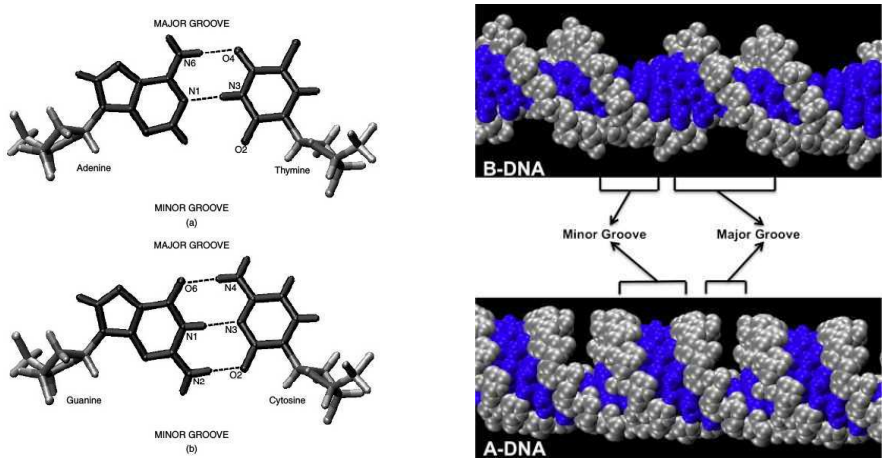



**Figura 1.2**
a) *Major* e *minor groove* identificados nos pares de base AT e CG. Figura retirada Philip E. Bourne [9]. b) *Major* e *minor groove* na dupla hélice B-DNA (superior) e A-DNA (inferior). Os átomos das bases estão exibidos em azul e a estrutura de açúcar-fosfato em cinza. No caso do A-DNA o *major groove* é mais estreito e profundo e, assim, inacessível. Figura retirada de Halder and Bhattacharyya [8]

tura do B-DNA, os pares de base se posicionam quase perpendicularmente ao eixo da hélice. Consequentemente, a separação entre os pares de base coincide com o passo da hélice 3,4 Å, conhecido como *rise*.

No arranjo Watson-Crick, os grupos de açúcar se ligam às bases no mesmo lado do par de base e, assim, definem a posição das duas estruturas de açúcar fosfato no DNA. Quando os pares se empilham sucessivamente na hélice, a separação entre os açúcares devido a essa assimetria dá origem a duas reentrâncias de dimensões diferentes chamadas de sulco menor e sulco maior (*minor groove* e *major groove*), como pode ser visto na figura 1.2. Na estrutura B-DNA há um número maior de grupos funcionais expostos em direção ao major *groove*. Como esses grupos funcionais dependem da sequência, tal característica permite que as proteínas localizem e identifiquem as bases da sequência [8].

Em uma fita, os nucleosídeos individuais se unem linearmente através dos grupos fosfatos ligados às posições $3'$ e $5'$ dos açúcares. Essa constituição assimétrica dos nucleotídeos confere direcionalidade a fita, como pode ser visualizado na figura 1.1, a denominação terminação $5'$ de uma fita faz referência a existência da hidroxila livre na posição $5'$ e, de maneira análoga, chamamos terminal $3'$ a extremidade da fita que apresenta uma hidroxila livre na posição $3'$. Quando os pares são do tipo cis-Watson-Crick, as fitas se posicionam de forma anti-paralela, ou seja uma fita $5' \rightarrow 3'$

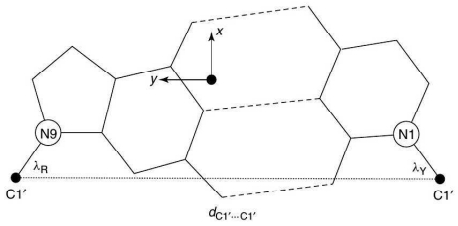



**Figura 1.3**
Parâmetros dos pares de bases idealizados, $d_{C1'...C1'}$ e $\lambda$, usados para deslocar e girar as bases complementares na otimização do sistema de referência padrão para A e B-DNA. A origem é indicada pelo ponto preto e os eixos x e y apontam nas direções exibidas. Sistema de referência proposto por Olson et al. [10], figura retirada do artigo em questão.

se combina com a outra fita $3' \rightarrow 5'$. Ao escrever uma sequência os grupos fosfatos podem ser representados por "p" ou omitidos.

A ligação entre o açúcar e a base é chamada de ligação glicosídica. Em relação ao ângulo glicosídico, observações experimentais indicam que há duas configurações de energia mais baixas. Na conformação *anti* as ligações de hidrogênio dos pares Watson-Crick apontam para fora do anel do açúcar, enquanto na conformação *syn* as ligações de hidrogênio apontam na direção do açúcar [9].

Como as ligações de hidrogênio e o esqueleto de açúcar fosfato não conferem uma estrutura rígida ao DNA, o DNA exibe flexibilidade tanto entre as bases de um par quanto na transversal, entre os pares de bases. Tal flexibilidade depende da constituição das bases que estão interagindo, da natureza da sequência. Para descrever essa conformação estrutural, uma série de parâmetros rotacionais e translacionais intra-par de base e inter-pares de base foram definidos pelos pesquisadores no acordo de Cambridge, em 1989 [11]. Várias abordagens matemáticas e, portanto, algoritmos diferentes foram desenvolvidos para a determinação desses parâmetros [12–17], contudo uma série de divergências foram observada nos valores obtidos [18]. Posteriormente essa discordância foi atribuída principalmente a diferença nas várias referências adotadas [18] e um sistema de referência de coordenadas padrão foi proposto para resolver essas ambiguidades [10]. A figura 1.3 exibe o sistema de referência: em um par Watson-Crick ideal o eixo horizontal é posicionado em direção ao *major groove*, o eixo vertical é paralelo ao vetor C1′...C1′.

Os principais métodos experimentais de investigação da estrutura de oligonucleotídeos são a cristalografia de raio-X e a ressonância magnética nuclear (NMR). Desde o desvendamento da dupla hélice por Watson e Crick [19] a partir das medidas de raio-X realizadas por Franklin e Wilkins [20], uma série de outras conformações e características estruturais foram elucidadas, como a existência de multihélice [21].

Além da forma B, existe uma variedade de conformações que o DNA pode assu-

mir, sendo as outras formas canônicas mais comuns A-DNA e Z-DNA. A conformação A-DNA também consiste em uma hélice dextrógira, nessa forma a presença de açúcar na forma C3'-endo provoca a aproximação dos grupos fosfatos das fitas (5,9 Å comparado com 7,0 Å da forma B). Como consequência, A-DNA apresenta pares de bases girados e inclinados em relação ao eixo da hélice. O passo de hélice, *rise* também é afetado e mede 2,56 Å em contraste com o B-DNA canônico 3,4 Å e uma volta completa da hélice contém 11 pares de base, enquanto na forma B são 10 pares. Os sulcos também são distintos, conforme mostrado na figura 1.2. A diferença de largura entre o *major* e *minor groove* não é tão pronunciada no A-DNA quanto no B-DNA. Por outro lado, enquanto no B-DNA eles apresentam praticamente a mesma profundidade no A-DNA o *major groove* é estreito e profundo e o *minor* é largo e raso. A forma Z-DNA, descoberta em 1979 [22], possui orientação espiralada à esquerda e a cadeia de fosfato apresenta o formato peculiar de zigue-zague.

## 1.4 A desnaturação de DNA

Quando a molécula de DNA é submetida a um aumento gradual de temperatura, as ligações de hidrogênio que mantém as bases ligadas são progressivamente rompidas. O processo de separação total das duas fitas da hélice é chamado de desnaturação. A temperatura de desnaturação é definida como aquela em que metade dos duplexos estão dissociados e é obtida experimentalmente através de espectrofotometria ultravioleta (UV), pois a absorção de UV aumenta com o rompimento das bases [23]. O aumento observado na absorção UV na faixa de 260 nm se deve à transição eletrônica $\pi - \pi*$ nas bases purinas e pirimidinas, representando a mudança na configuração eletrônica dessas bases decorrente da redução do empilhamento e do pareamento [24]. É possível predizer a fração de sequências dissociadas através da medida de absorção UV pois o percentual de hipercromicidade, o aumento da absorção com a desnaturação da hélice, varia aproximadamente linearmente com o número de bases desemparelhadas [25].

Experimentos de desnaturação de DNA são feitos em solução salina variando progressivamente a temperatura e observando a absorção UV. A figura 1.4 (painel esquerdo) mostra um exemplo deste tipo de curva de absorção. A temperatura de desnaturação $T_m$ é determinada pelo ponto de inflexão da curva de absorção, para identificá-lo, basta fazer a primeira derivada desta curva e reconhecer o máximo [26]. O valor exato de $T_m$ pode variar um pouco dependendo do tipo de normalização adotada e outros ajustes, condições que variam dependendo do grupo experimental [27].

A dissociação da hélice dupla, que vamos chamar de $X \cdot Y$, em duas fitas simples

$X$ e $Y$ pode ser descrita pela reação reversível $X \cdot Y \rightleftharpoons X + Y$ [23, 28]. Ou seja, é um processo que envolve apenas dois estados (hélice $X \cdot Y$ e fitas simples desnaturadas $X$ e $Y$) e a temperatura de desnaturação é definida como sendo aquela em que metade das fitas está formando hélices duplas e a outra metade está desnaturada. Assim, no equilíbrio entre estes dois estados é possível usar a a equação de Van't Hoff que estabelece uma relação entre a energia livre de Gibbs, $\Delta G$, a temperatura e a constante de equilíbrio $K_{eq}$ dessa dissociação

$$\Delta G = -RT \ln(K_{eq}) \tag{1.1}$$

onde $K_{eq} = [X][Y]/[X \cdot Y]$, os colchetes presentes nessa equação indicam concentração. Seja $f$ a fração de duplexos dissociados e $C_t = [X] + [Y] + 2[X \cdot Y]$ a concentração total de fitas simples. Para sequências não-auto-complementares, $f = [X]/[X \cdot Y]$ e

$$K_{eq} = \frac{[X]^2}{[X \cdot Y]} = \frac{f^2 C_t}{2(1-f)}, \tag{1.2}$$

para $T = T_m$, a fração de sequências dissociadas $f = 1/2$, ou seja $K_{eq} = C_t/4$. Para sequências auto-complementares ($X = Y$), $f = [X]/2[X \cdot X]$ e

$$K_{eq} = \frac{[X][Y]}{[X \cdot Y]} = \frac{2f^2 C_t}{(1-f)}, \tag{1.3}$$

assim, para $T = T_m$, $f = 1/2$, temos $K_{eq} = C_t$. Podemos então resumir a dependência de $T_m$ com as grandezas termodinâmicas por

$$\frac{1}{T_m} = \begin{cases} -\frac{R}{\Delta H} \ln C_t + \frac{\Delta S - R \ln 4}{\Delta H} & \text{não-auto-complementar} \\ -\frac{R}{\Delta H} \ln C_t + \frac{\Delta S}{\Delta H} & \text{auto-complementar} \end{cases} \tag{1.4}$$

Portanto, analisando a dependência da temperatura de desnaturação com a concentração da fitas de DNA $C_t$, é possível extrair informações de entalpia, $\Delta H$, e entropia, $\Delta S$, a partir da equação 1.1 e da relação $\Delta G = \Delta H - T\Delta S$. Isto é feito a partir do gráfico $1/T_m$ por $\ln C_t$ (gráfico de Van't Hoff), a regressão linear possibilita a obtenção das grandezas termodinâmicas $\Delta H$ e $\Delta S$. Um exemplo deste ajuste de Van't Hoff pode ser visualizado na figura 1.4. Quando uma sequência de DNA se ajusta linearmente desta maneira se diz que o processo de desnaturação é caracterizado por dois estados. Há casos, no entanto, em que o gráfico tipo Van't Hoff desvia significativamente da relação linear, o que ocorre com maior frequência para sequências longas ou ricas em AT [29]. Neste caso a sequência é chamada de *non-two-state* e usualmente excluída de

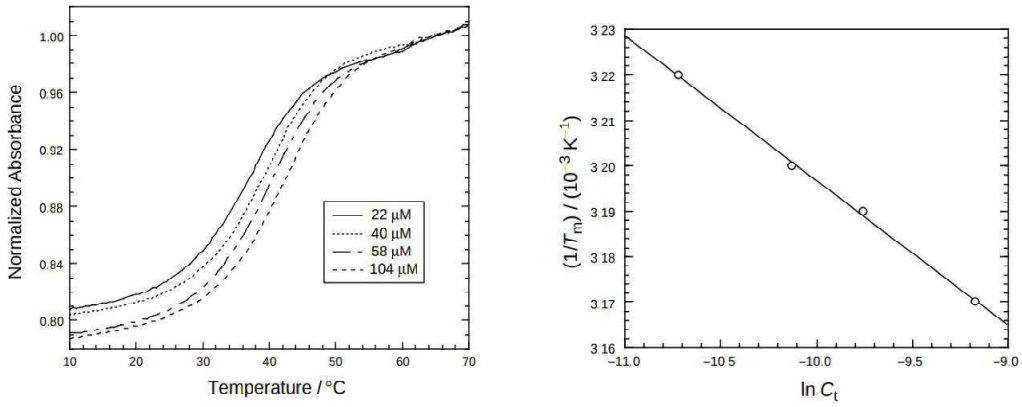



**Figura 1.4**
Curvas experimentais de absorção UV em DNA. Esquerda: curvas experimentais de absorção UV versus temperatura para quatro concentrações das fitas complementares $d(CAAAAAAAG)$ e $d(GTTTTTTTC)$. As curvas estão normalizadas para indicar 1 a 65 °C. Direita: ajuste tipo van't Hoff $1/T_m \times \ln C_t$ onde $T_m$ é a temperatura de desnaturação e $C_t$ a concentração total de fitas simples geralmente dada em $\mu$M. Figure retirada de Howard [23].

análises termodinâmicas [29].

Uma questão interessante foi levantada em relação às sequências longas, já que essas sequências podem existir em estados intermediários, no qual apresentam abertura parcial interna (bolhas) ou nas terminações [30]. A intensidade de absorbância UV indica a quantidade de pares abertos, contudo não é capaz de discernir em quais sequências o par aberto ocorre, afinal a medida representa a propriedade do ensemble. Ou seja, uma intensidade de 50% pode corresponder a situação em que metade das sequências da solução se encontra no estado simples e metade como fitas duplas, bem como pode estar associada à situação em todas as moléculas estão parcialmente (50%) abertas.

Buscando diferenciar essas situações, foi desenvolvido um processo experimental [31, 32], utilizando sequências auto-complementares nas terminações. Devido à complementaridade, a fita única é capaz de se dobrar nela mesma formando um *hairpin* (formato de grampo com um loop no centro, na região não complementar). Partindo de um estado em que todas as moléculas são duplexos e aumentando gradualmente a temperatura, uma amostra é retirada em diferentes temperaturas e resfriada. Ao resfriar a amostra retirada da solução as moléculas parcialmente abertas refazem as ligações de hidrogênio restantes, enquanto as fitas simples se dobram nelas mesmas, e assim é possível identificar a fração de fitas simples.

## 1.5 Modelos para determinar temperaturas de desnaturação

O procedimento experimental da seção anterior evidencia que a obtenção de temperaturas de desnaturação é um procedimento bastante trabalhoso. Assim, considerando o grande número de combinações de pares de bases que DNA pode ter, medir todas as combinações possíveis é impraticável. Dessa forma, se fazem necessários modelos que possam ajudar a prever a temperatura de desnaturação de sequências desconhecidas partindo de um conjunto limitado de temperaturas medidas.

O modelo mais simples que realiza esta tarefa, e que continua sendo extensivamente usado por bioquímicos, é o de próximos vizinhos (*nearest-neighbour* ou NN) que associa incrementos de energia livre a cada par de base de uma sequência. O valor desse incremento depende do primeiro par adjacente e é obtido a partir da regressão de dados de concentração versus temperatura de desnaturação para várias sequências [33–35]. A nomenclatura introduzida neste modelo é recorrente na discussão do nosso trabalho e por isto vamos rever a sua ideia básica aqui.

A conceito do modelo NN é subdividir a sequência em partes de próximos vizinhos como ilustrado neste exemplo

$$\begin{matrix}5'\text{-CGATGA-}3'\\3'\text{-GCTACT-}5'\end{matrix} \rightarrow \begin{matrix}5'\text{-CG-}3'\\3'\text{-GC-}5'\end{matrix} + \begin{matrix}5'\text{-GA-}3'\\3'\text{-CT-}5'\end{matrix} + \begin{matrix}5'\text{-AT-}3'\\3'\text{-TA-}5'\end{matrix} + \begin{matrix}5'\text{-TG-}3'\\3'\text{-AC-}5'\end{matrix} + \begin{matrix}5'\text{-GA-}3'\\3'\text{-CT-}5'\end{matrix}$$

onde os seis pares de base formam cinco próximos vizinhos. Note que há a possibilidade de ter próximos vizinhos equivalentes por razões de simetria. Por exemplo

$$\begin{matrix}5'\text{-TG-}3'\\3'\text{-AC-}5'\end{matrix}$$

equivale a

$$\begin{matrix}5'\text{-CA-}3'\\3'\text{-GT-}5'\end{matrix}$$

Assim, para DNA contendo somente pares AT e CG, essa divisão resulta em 10 possibilidades de próximos vizinhos.

A notação desta decomposição é frequentemente simplificada para

$$\begin{matrix}5'\text{-CGATGA-}3'\\3'\text{-GCTACT-}5'\end{matrix} \rightarrow \text{CG/GC+GA/CT+AT/TA+TG/AC+GA/CT}$$

ou na forma base-fosfato-base

$$\begin{matrix}5'\text{-CGATGA-}3'\\3'\text{-GCTACT-}5'\end{matrix} \rightarrow \text{CpG+GpA+ApT+TpG+GpA}$$

onde se omite a fita complementar.

Enfim, as variações de entalpia e entropia, obtidas experimentalmente pela análise do gráfico de $T_m^{-1} \times \ln C_t$ da equação (1.4), podem ser escritas em contribuições individualizadas de próximos vizinhos de pares de base. Continuando no exemplo anterior, a variação de entropia experimental $\Delta S$ seria

$$\Delta S = \Delta S_{\mathrm{CpG}} + 2\Delta S_{\mathrm{GpA}} + \Delta S_{\mathrm{ApT}} + \Delta S_{\mathrm{TpG}}$$

onde o termo relativo a GpA aparece duas vezes. Para a variação de entalpia pode-se proceder da mesma maneira. De posse de valores de próximos vizinhos agora basta realizar a soma e substituir na equação (1.4) para obter uma predição de $T_m$.

Contudo, embora faça previsões boas de temperatura de desnaturação para sequências desconhecidas, esse tipo de modelo não revela nada sobre as interações intramoleculares [36]. O modelo NN também não é útil quando a estrutura varia para situações não cobertas por experimentos. Isto motivou o desenvolvimento de outros modelos que vamos discutir a seguir.

O modelo mecânico-estatístico proposto por Poland and Scheraga, conhecido como modelo Poland-Scheraga, usando a notação da revisão [38], assume uma fita de $L$ bits para representar o DNA. Ou seja, de forma análoga ao modelo de Ising, o DNA é tratado como uma sequência de dois estados, a cada "1" está associado um estado ligado, ao qual é atribuído um peso estatístico q=$e^{-\epsilon/k_B T}$, e "0" ao estado aberto, ou seja em que as bases não formam ligações de hidrogênio, ao qual se associa um peso entrópico. Esse modelo é apropriado para sequências longas [39] e é bastante usado na predição de temperaturas de desnaturação [40, 41]. Mas de maneira semelhante ao modelo NN, não permite inferir muito sobre os processos intramoleculares.

Já que estes modelos não caraterizam as interações intramoleculares, por que então insistir com estes modelos simplificados e não calcular as temperaturas de desnaturação por modelos mais completos como dinâmica molecular atomística? A razão aqui é o alto custo computacional que restringe esta técnica em simular a interação entre poucos pares de base e por período curto de tempo [42]. Isto é, o tempo coberto pela simulação não é longo o suficiente para permitir uma variação contínua de temperatura e assim observar a desnaturação.

Um modelo que permite descrever as interações intramoleculares, como ligação de hidrogênio, e é simples o suficiente para permitir o cálculo de temperaturas de desnaturação é o modelo Peyrard-Bishop [43]. Diferente do modelo Poland-Scheraga, este modelo apresenta a vantagem de compreender os estados intermediários individuais através da variável de abertura $y_n$ em vez de considerar apenas estados abertos e

fechados. Além disto as interações intramoleculares podem ser descritas por potenciais que tem por parâmetros grandezas físicas que podem ser diretamente associadas a ligações de hidrogênio e interação de empilhamento. A possibilidade de calcular temperaturas de desnaturação foi realizadas através da introdução de um conceito de equivalência [44]. Além disto foi demonstrada a possibilidade de obter os parâmetros de ligações de hidrogênio e interação de empilhamento em DNA [45] e RNA [46]. Nas seções seguintes descrevemos o modelo em detalhe bem como o método de equivalência que é a base do nosso trabalho.

## 1.6 Modelos Peyrard-Bishop

### 1.6.1 Modelo PB original para sequências homogêneas

O modelo proposto por Peyrard and Bishop [43], ao qual vamos nos referir como modelo PB original, simplifica o DNA ao representá-lo por uma estrutura bidimensional plana, na qual cada nucleotídeo é interpretado como uma partícula. As interações de empilhamento entre as bases dos pares vizinhos são modeladas por um potencial harmônico e as ligações de hidrogênio formadas entre as bases de um par são descritas através de um potencial de Morse. Dessa forma o hamiltoniano associado ao modelo é descrito como

$$H = \sum_n \frac{1}{2}m(\dot{u_n}^2 + \dot{v_n}^2) + \frac{1}{2}k(u_n - u_{n-1})^2 + \frac{1}{2}k(v_n - v_{n-1})^2 + D(e^{-\frac{a}{\sqrt{2}}(u_n - v_n)} - 1)^2 \quad (1.5)$$

onde $u_n$ e $v_n$ são as variáveis responsáveis por descrever as posições das bases do par $n$ (exibidos no esquema à esquerda da figura 1.5), $m$ representa a massa das bases, $k$ é a constante elástica associada ao empilhamento e $D$ e $a$ parâmetros associados ao potencial de Morse.

O hamiltoniano da equação 1.5 pode ser reescrito em termos das novas variáveis $x_n = (u_n + v_n)/\sqrt{2}$ e $y_n = (u_n - v_n)/\sqrt{2}$, como mostrado na figura 1.5

$$H = \sum_n \frac{{p_n}^2}{2m} + \frac{1}{2}k(x_n - x_{n-1})^2 + \sum_n \frac{{q_n}^2}{2m} + \frac{1}{2}k(y_n - y_{n-1})^2 + D\left(e^{-y_n/\lambda} - 1\right)^2 \quad (1.6)$$

onde $p_n = m\dot{x}_n$ e $q_n = m\dot{y}_n$ e $\lambda = 1/a$. Esta Hamiltoniana permite o cálculo da função de partição

$$Z = \int \prod_n dp_n dq_n dx_n dy_n e^{-\beta H} \quad (1.7)$$

Esta função de partição é separável em termos das variáveis que a compõem, podendo

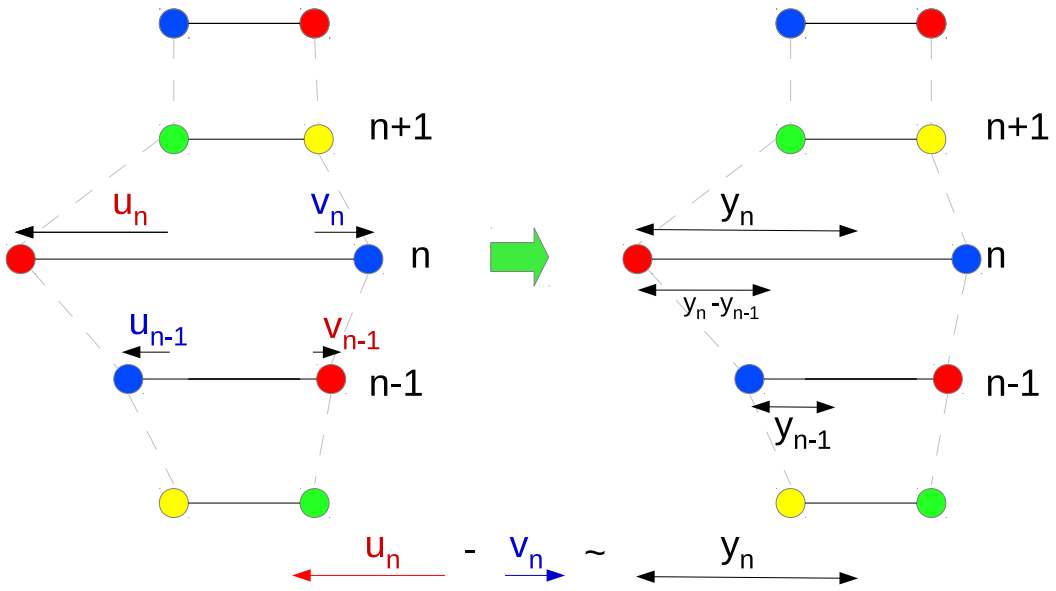



**Figura 1.5**
Identificação das variáveis iniciais $u_n$ e $v_n$ do modelo Peyrard-Bishop que medem a posição de cada base do par $n$ (para facilitar a visualização optamos por adotar como referência a posição de equilíbrio de cada base). Na figura da direita é exibida a variável final $y_n = \frac{(u_n - v_n)}{\sqrt{2}}$ responsável por mensurar o deslocamento do par de base em relação à posição de equilíbrio, ou seja a abertura.

ser fatorada em $Z = Z_p Z_q Z_x Z_y$. O termo função da variável $x_n$, $Z_x$, assim como os termos dependente dos momentos, $Z_p$ e $Z_q$, apresenta forma gaussiana e pode ser integrado diretamente, de acordo com

$$\int e^{-az^2} dz = \sqrt{\pi/a}.$$

Portanto a função partição simplifica para

$$Z_p = Z_q = (2\pi m k_B T)^{\frac{N}{2}}$$

e

$$Z_x = \left(\frac{2\pi k_B T}{k}\right)^{\frac{N}{2}}.$$

Desta maneira resta apenas resolver a função de partição $Z_y$ onde podemos usar a técnica de integral de transferência [47].

$$Z_y = \int \exp\left[-\beta\left(\sum \frac{1}{2}k(y_n - y_{n-1})^2 + D\left(e^{-y_n/\lambda} - 1\right)^2\right)\right] \prod_n dy_n \qquad (1.8)$$

Considerando condições periódicas de contorno (sequência em círculo de forma que

$n = N + 1 = 1$)

$$Z_y = \int \exp\Bigg[-\beta\Bigg(\frac{1}{2}k(y_1 - y_N)^2 + D\left(e^{-y_1/\lambda} - 1\right)^2 + \frac{1}{2}k(y_2 - y_1)^2 + D\left(e^{-y_2/\lambda} - 1\right)^2 ... + \frac{1}{2}k(y_N - y_{N-1})^2 + D\left(e^{-y_N/\lambda} - 1\right)^2\Bigg)\Bigg] dy_1 dy_2 ... dy_N \tag{1.9}$$

Seja uma função

$$K(y_a, y_b) = \exp\left[-\beta\left(\frac{1}{2}k(y_b - y_a)^2 + \frac{1}{2}D\left(e^{-y_a/\lambda} - 1\right)^2 + \frac{1}{2}D\left(e^{-y_b/\lambda} - 1\right)^2\right)\right] \tag{1.10}$$

é fácil perceber que ela possui a seguinte propriedade $K(y_a, y_b) = K(y_b, y_a)$. Reescrevendo a função de partição em termos dessa função temos

$$Z_y = \int dy_1 dy_2 ... dy_N K(y_1, y_2) K(y_2, y_3) ... K(y_N, y_1) \tag{1.11}$$

Dado que $K(y_a, y_b) > 0$ é simétrico, se assumirmos que

$$\|K(y_a, y_b)\| \equiv \left(\int\int [K(y_a, y_b)]^2 dy_a dy_b\right)^{1/2} < \infty \tag{1.12}$$

a integral da equação

$$\int K(y_a, y_b)\varphi(y_a)\ dy_a = \lambda\varphi(y_b) \tag{1.13}$$

possui um *kernel* positivo do tipo Hilbert Schmidt. E, assim, possui um conjunto completo associado de autovalores positivos e autovetores ortonormais. Ordenando os autovalores em ordem decrescente e denominando $\lambda_1$, $\lambda_2, \ldots$ etc os autovalores associados aos autovetores $\varphi_1(y)$, $\varphi_2(y)$, o *kernel* pode ser expandido nesta base como

$$K(y_a, y_b) = \sum_n \lambda_n \varphi_n(y_a)\varphi_n(y_b) \tag{1.14}$$

Usando a relação de completeza e a ortonormalidade

$$Z = \sum_{n=1}^{\infty} \lambda_n^N. \tag{1.15}$$

Note que este resultado fechado é limitado a uma sequência homogênea de DNA.

Com o resultado da equação (1.15) podemos agora calcular uma grandeza de interesse que é a abertura média das ligações de hidrogênio $\langle y \rangle$. O valor desta abertura

determina se a hélice dupla está desnaturada ou não. A abertura média pode ser obtida por

$$\langle y \rangle = \frac{1}{Z} \sum_n \langle n|y|n \rangle \lambda_n^N \tag{1.16}$$

sendo que

$$\langle n|y|m \rangle = \int \varphi_n(y) y \varphi_m(y)\ dy \tag{1.17}$$

onde $\varphi_n$ são os autovetores e $\lambda_n$ os autovalores associados

No limite termodinâmico, em que o tamanho $N$ da cadeia tende ao infinito, ficamos apenas com o menor termo $n = 0$, o que permite uma simplificação maior ainda

$$\langle y \rangle = \int \varphi_1(y) y \varphi_1(y)\ dy \tag{1.18}$$

A versão original do modelo PB, descrita nesta seção, considera o DNA homogêneo, ou seja, a constante de empilhamento $k$ e parâmetros de Morse $D$ idênticos para todas as bases. Isto equivale, por exemplo, a ter uma sequência de DNA onde um lado da hélice é composto somente por C e o outro lado somente por G. Zhang et al. [47] desenvolveram um método que permite aplicar o modelo PB a sequências heterogêneas, ou seja, com uma sequência arbitrária de pares de bases. Isso pôde ser alcançado expandindo a função de partição na base de autovetores da função de partição homogênea. Essa abordagem atribuiu aos modelos mesoscópicos a capacidade de diferenciar os pares de bases e, assim, capturar de forma mais realista as propriedades das sequências. Este método será apresentado na seção 1.7.

### 1.6.2 Modificação da Hamiltoniana PB

Uma das grandes vantagens do modelo PB é permitir alterar a Hamiltoniana (1.5) modificando os potenciais para incluir interações de interesse. Esta ideia será usada no capítulo 3 e nesta seção vamos rever algumas modificações relatadas na literatura. Vários estudos e versões modificadas do modelo PB foram sugeridas, Goldman and Olson [48] por exemplo introduziram uma versão quântica estatística do modelo PB para estudar o papel da massa no processo de desnaturação e usaram teoria perturbativa para tratar acréscimos aleatórios na temperatura.

Conforme visto, ao integrar a função de partição, a parte dependendo de $x_n$ é diretamente "eliminada" e a dependência do modelo fica restrita à variável $y_n$. Por esse motivo, é comum se referirem ao modelo PB como um modelo unidimensional. Devido a essa redução de variáveis, usualmente os modelos mesoscópicos subsequentes ao original modificam apenas os potenciais associados a variável $y_n$ e são apresentados

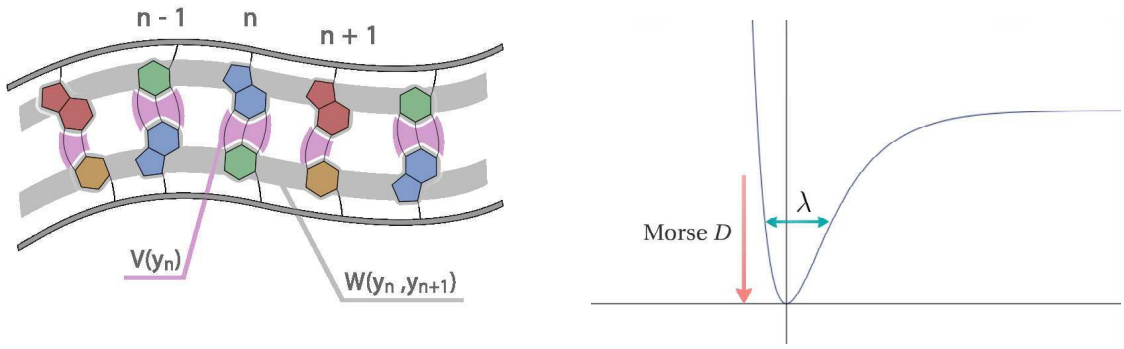



**Figura 1.6**
Figura esquemática de DNA definindo os potencias do modelo PB. a) No modelo mesoscópico as interações químicas são traduzidas em potenciais: $W(y_n, y_{n+1})$ se refere ao potencial de empilhamento e $V(y_n)$ é o potencial associado às ligações de hidrogênio. b) Exibe o potencial de Morse responsável por descrever o comportamento das pontes de hidrogênio no modelo PB.

exibindo apenas a parte do hamiltoniano dependente dessa variável

$$H = \sum_n \left( \frac{1}{2} m \dot{y_n}^2 + V(y_n) + W(y_n, y_{n+1}) \right). \tag{1.19}$$

De acordo com o que está mostrado na figura esquemática 1.6, $y_n$ mede o deslocamento em relação a posição de equilíbrio dos nucleotídeos de um par de base $n$, $W(y_n, y_{n+1})$ se refere ao potencial de empilhamento que caracteriza a interação entre pares vizinhos — engloba sobreposição dos orbitais elétron $\pi$ dos anéis orgânicos que formam as bases e o acoplamento das estruturas açúcar-fosfato — e $V(y_n)$ é o potencial associado as ligações de hidrogênio que unem os nucleotídeos de um par. Os modelos mesoscópicos do tipo Peyrard-Bishop são uma classe de modelos que se destacam na descrição do mecanismo de desnaturação por possuir parâmetros com interpretação física direta.

Vamos identificar os termos da equação 1.19 com os potenciais adotados pelo modelo PB original, para promover uma explicação mais detalhada dessa escolha, assim como incluir a abordagem de sequências heterogêneas, isto é, os parâmetros mencionados anteriormente vão receber subíndices para denotar a dependência da sequência. O potencial de Morse

$$V(y_n) = D_n \left( e^{-y_n/\lambda_n} - 1 \right)^2, \tag{1.20}$$

escolhido para descrever as ligações de hidrogênio, veja a figura 1.6b, possui o comportamento qualitativo desejado para representar a situação física. Para $y < 0$ o potencial é repulsivo, possui um mínimo estável em $y = 0$ correspondente à posição de equilíbrio

das bases. Para $y$ suficientemente grande o potencial é constante e está associado à ausência de atração quando as bases estão distantes. Contudo, infelizmente a dinâmica não é descrita satisfatoriamente por esse potencial. Ao comparar com os valores inferidos por experimentos de troca próton-deutério, tanto o tempo médio de abertura após uma flutuação quanto o tempo que o par de base permanece fechado entre duas flutuações de abertura são ordens de magnitude inferiores ao experimental [49] .

A aproximação harmônica para a interação de empilhamento

$$W(y_n, y_{n+1}) = \frac{k_{n,n+1}}{2} \left(y_n - y_{n+1}\right)^2 \tag{1.21}$$

introduz uma interação tipo próximos vizinhos semelhante ao que discutimos no início da seção 1.5. Como veremos em maior detalhe no capítulo 3 trata-se de uma simplificação que elimina um parâmetro estrutural. Porém, mais importante foi a observação por Dauxois et al. [50] que este potencial não causa uma transição abrupta nas aberturas médias $\langle y \rangle$. Isto motivou a proposta de uma variação do potencial de próximos vizinhos que reproduz este tipo de transição.

$$W(y_n, y_{n+1}) = \frac{k}{2} \left(1 + \rho e^{-\alpha(y_n + y_{n+1})}\right) \left(y_n - y_{n+1}\right)^2. \tag{1.22}$$

A escolha da forma do potencial foi motivada pela ideia de que a energia de empilhamento está associada ao par de base e não às bases individuais e, sendo assim, a separação das bases e consequente rompimento da ligação de hidrogênio, deve ser seguida de uma reorganização eletrônica e redução na interação de empilhamento [50, 51]. Esse modelo ficou conhecido como PBD.

Uma transição abrupta também pode ser obtida alterando o potencial de Morse ao invés do potencial de empilhamento [44]. A motivação desse potencial proposto foi a tentativa de simular a interação com o solvente,

$$V(y_n) = V_{\text{Morse}}(y_n) - f_s D \left[\tanh(y_n/\lambda_s) + 1\right]. \tag{1.23}$$

Essa modificação insere uma barreira no potencial de Morse sugerindo que quando um par de base é rompido ele passa a interagir com as moléculas de água da solução a partir de ligações de hidrogênio. Dessa forma, para refazer a formação é necessário que as ligações com as moléculas de água sejam desfeitas, essa "dificuldade" a ser transposta é representada pela barreira no potencial. De maneira análoga, com o mesmo propósito de capturar essa interação com o solvente outro grupo propôs adicionar uma barreira

gaussiana no potencial de Morse[52, 53]

$$V(y_n) = D(e^{-ay_n} - 1)^2 + Ge^{-(y_n - y_0)^2/b} \tag{1.24}$$

Nesses estudos, em conjunto com o potencial acima, o potencial de empilhamento proposto por PBD foi adotado.

Também é possível modelar um termo adicional de força sobre o DNA como é realizado experimentalmente por pinças óticas. Esta técnica experimental tem se mostrado muito útil para investigar as propriedades mecânicas do DNA, uma das aplicações consiste no estudo do efeito de fármacos no comprimento de persistência do DNA, por exemplo psoralen [54] e ciclo-dextrina [55]. Para realizar a abertura mecânica aplica-se uma força perpendicular à direção da hélice dupla. A abertura mecânica das hélices é considerada introduzindo um termo de força externa no hamiltoniano [53, 56, 57].

$$H = \sum_n \left( \frac{1}{2} m \dot{y_n}^2 + V(y_n) + W(y_n, y_{n+1}) \right) - F y_e. \tag{1.25}$$

onde $y_e$ é a extensão decorrente da aplicação da força. Singh and Singh [57] por exemplo usaram esta Hamiltoniana para estudar o efeito do comprimento da sequência na desnaturação térmica e na abertura mecânica das dupla hélice do DNA usando o modelo PBD equação 1.22.

Como mostramos, existe uma flexibilidade bastante grande em modificar a Hamiltoniana dependendo do tipo de interação que se pretende estudar. Outros exemplos incluem trocar a interação de empilhamento por entalpias como feito por Joyeux and Buyukdagli [58] ou ir além da interação de próximos vizinhos Rapti [59]. Esta flexibilidade será explorada no capítulo 3, no qual estudaremos um potencial modificado que contém um parâmetro estrutural do DNA.

## 1.7 Método de otimização por equivalência termodinâmica

Apresentamos aqui o método principal usado nesta tese que permite calcular temperaturas de desnaturação em DNA e RNA e também obter parâmetros associados a ligação de hidrogênio e interações de empilhamento.

O método de equivalência termodinâmica permitiu o cálculo de temperaturas de desnaturação em DNA sem a necessidade de calcular a abertura média $\langle y \rangle$ em toda a faixa de temperatura [44]. Com a simplificação do cálculo e a consequente eficiência

computacional, foi possível desenvolver um novo método que extrai parâmetros associados a ligação de hidrogênio e interações de empilhamento diretamente de temperaturas de desnaturação [45]. A técnica de equivalência termodinâmica dispensa a necessidade de calcular a temperatura de desnaturação para comparar as propriedades térmicas de duas sequências. Mostrou-se que tais propriedades podem ser acessadas através do "índice de desnaturação" promovendo uma eficiência computacional que viabiliza a otimização dos parâmetros.

Conforme visto na seção 1.6.1, cada modelo PB tem um Hamiltoniano associado descrito por

$$H = \sum_n \left( \frac{1}{2} m \dot{y_n}^2 + V(y_n) + W(y_n, y_{n+1}) \right) \tag{1.26}$$

Cuja função de partição para uma sequência de tamanho $N$ é dada por

$$Z = \int e^{-\left[\beta \sum_{n=1}^{N} \left( \frac{1}{2} m \dot{y_n}^2 + V(y_n) + W(y_n, y_{n+1}) \right)\right]} dy_1 ... dy_N \tag{1.27}$$

Para possibilitar a integração no caso de sequências não-homogêneas Zhang et al. [47] propuseram que essa função de partição fosse expandida em funções ortonormais, o que resulta em

$$Z = \mathrm{Tr}\,(C^{(1,2)} C^{(2,3)} \dots C^{(N,1)}), \tag{1.28}$$

onde cada $C^{(n,n+1)}$ identifica uma matriz que representa a interação entre os pares de base vizinhos $n$ e $n+1$. A matriz $C^{(N,1)}$ está associada a condição de fronteira que relaciona o primeiro e o último par. Ao considerar apenas os dois pares complementares CG ($s$) e AT ($w$), há quatro tipos de vizinhos e, assim, quatro matrizes correspondentes $C^{(w,w)}$, $C^{(w,s)}$, $C^{(s,w)}$ e $C^{(s,s)}$. Decidiu-se adotar como conjunto de bases ortonormais as bases associadas a sequência homogênea de CG, obtidas através da técnica de integral de transferência [44]. Nessa base $C^{(s,s)}$ se torna uma matriz diagonal $\Lambda$ na qual os elementos $\lambda_i$ são os autovalores das autofunções associadas a sequência homogênea de CG. Definindo uma matriz $\Delta^{(a,b)}$ que mensura a diferença entre uma matriz associada a interação dos vizinhos do tipo$(a, b)$ com a matriz de interação entre a os vizinhos $(s, s)$

$$\Delta^{(a,b)} = C^{(a,b)} - \Lambda, \tag{1.29}$$

A função pode ser reescrita como

$$Z = \mathrm{Tr}\,[(\Lambda + \Delta^{(1,2)})(\Lambda + \Delta^{(2,3)}) \dots (\Lambda + \Delta^{(N,1)})]. \tag{1.30}$$

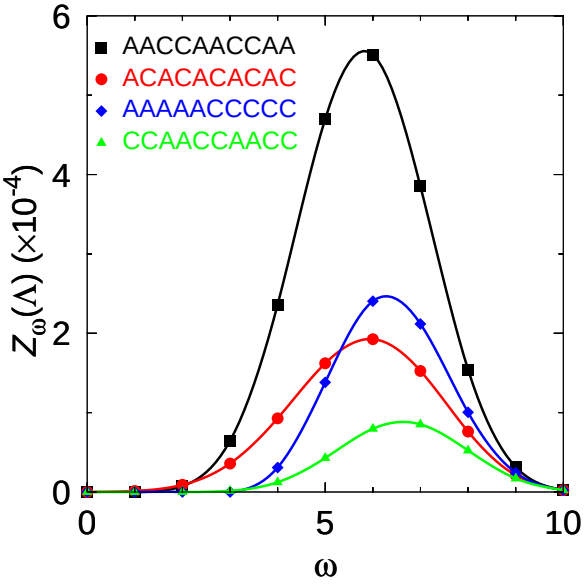



**Figura 1.7**
Função partição $Z_\omega(\Lambda)$ em função de $\omega$. Sequências de tamanho $N = 10$ com conteúdo de pares de base do tipo CG entre 40 e 60% ($d(AACCAACCAA)$ (quadrados pretos), $d(AC)_5$ (círculos vermelhos), $d(AAAAACCCCC)$ (diamantes azuis) e $d(CCAACCAACC)$ (triângulos verdes). Valores foram calculados para temperatura de 370 K. As curvas são as regressões gaussianas $\mathcal{F}(\omega)$. Figura retirada da referência 44.

Ao desenvolver as multiplicações e empregar propriedades de traço, obtém-se

$$Z = \sum_{\omega=0}^{N} Z_\omega(\Lambda) = \sum_{\omega=0}^{N} \mathrm{Tr}[M(\Lambda^\omega)], \tag{1.31}$$

onde $M(\Lambda^\omega)$ são todos os termos contendo $\omega$ multiplicações da matriz $\Lambda$. É fácil perceber que, para uma sequência homogênea de CG, apenas haveria termos em $\Lambda^N$, já que todas matrizes $\Delta$ seriam nulas.

A função de partição $Z_\omega(\Lambda)$ (usando o modelo PB modificado por Dauxois et al. [50]) foi obtida numericamente para várias sequências curtas com conteúdo similar de pares CG (40–60%). A figura 1.7 exibe $Z_\omega(\Lambda)$como função do parâmetro $\omega$. Observou-se que essa dependência apresenta comportamento gaussiano, as curvas mostradas $\mathcal{F}(\omega)$ são regressões gaussianas dos valores obtidos para $Z_\omega(\Lambda)$. De fato, conforme o teorema de limite central, $\lim_{N\to\infty} Z_\omega(\Lambda) = \mathcal{F}(\omega)$. Embora a função de partição tenha forte dependência com a temperatura, o valor $\omega_{\max}$ (obtido através da função gaussiana $\mathcal{F}(\omega)$ não apresenta esse comportamento.

Na figura 1.8 pode se observar que além de não depender fortemente da temperatura, a ordem das sequências de diferentes composições não é afetada. O parâmetro de ordem $\omega_{\max}$ indica a quantidade de matrizes $\Delta$ não nulas. Para facilitar o entendimento, uma sequência qualquer de tamanho N, se for homogênea de CG apresenta o valor mais alto de $\omega_{\max}$, se for homogênea de AT o valor mais baixo e, portanto, qualquer composição heterogênea vai possuir um valor intermediário.

Os dados exibidos na figura 1.9 demonstram que $\omega_{\max}$ captura corretamente a ordem das temperaturas de desnaturação experimentais [60] das sequências. Muito

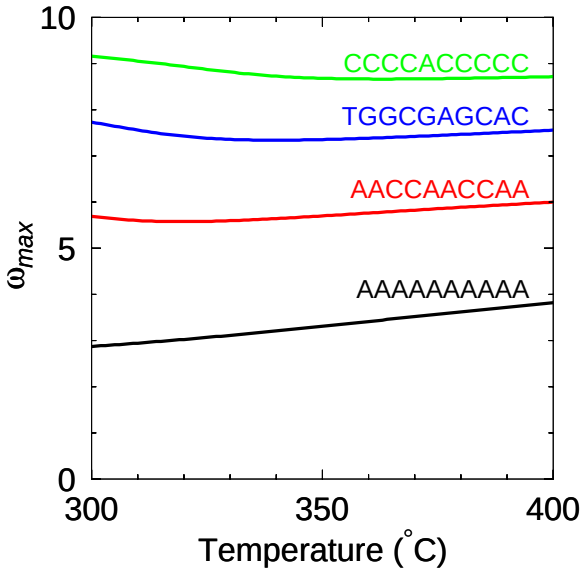



**Figura 1.8**
Parâmetro $\omega_{\max}$ versus temperatura. Vários exemplos de sequências com 10 pares de base de comprimento foram consideradas. Figura retirada da referência 44.

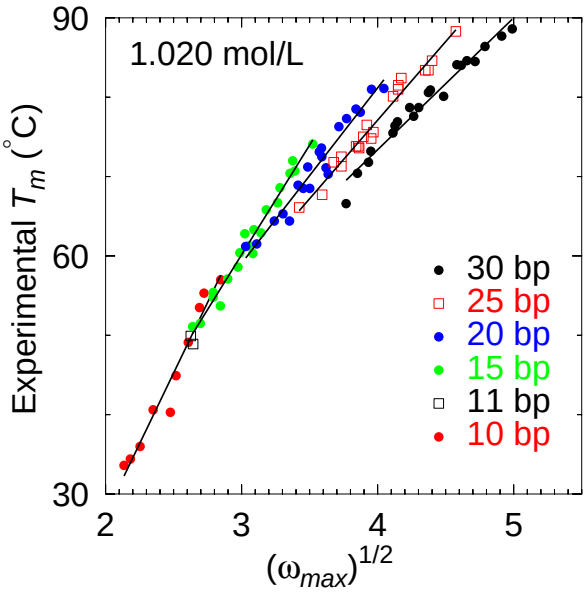



**Figura 1.9**
Temperaturas de desnaturação experimentais $T_m$ em função do índice de desnaturação $\tau = \omega_{\max}^{1/2}$. O tamanho das sequências varia de $N = 10$ a $N = 30$ e os dados experimentais foram realizados a concentração salina $[Na^+]$ de 1020mM [60]. A regressão linear é realizada para grupos de mesmo tamanho $N$ e possui desvios padrões de 1.7, 1.1, 1.3, 0.9 e 1.0 °C em ordem crescente de $N$. Figura retirada da referência 44.

mais do que isso, como as regressões lineares mostraram, há um comportamento linear da temperatura com $\omega_{\max}^{1/2}$. Adicionalmente, a inclinação das regressão linear depende de $N^{1/2}$. Devido a essa relação linear entre a temperatura de desnaturação e $\omega_{\max}^{1/2}$, esse parâmetro pode ser usado como um medidor de "equivalência termodinâmica", portanto foi denominado índice de desnaturação, $\tau$.

Portanto, esse método de otimização consiste em calcular, a partir da função de partição associada ao Hamiltoniano do modelo PB, um índice adimensional de desnaturação $\tau_i$ para cada sequência $i$. Esse processo é usado para cada conjunto de parâmetros $P = \{p_1, p_2 \dots p_L\}$ do hamiltoniano com o objetivo de prever as temperaturas de desnaturação $T_i'$. Para clarificar essa notação no caso do modelo PB original esse conjunto de parâmetros $P$ se refere aos parâmetros $D$s $k$s e $\lambda$s.

$$T_i'(P) = a_0(N) + a_1(N)\tau_i(P), \tag{1.32}$$

na qual os coeficientes são dependentes do tamanho $N$ da sequência. Isto é, para cada grupo de sequências de tamanho $N$, são obtidos dois coeficientes de regressão $a_0(N)$ e $a_1(N)$. Como foi observado que esses coeficientes apresentam dependência essencialmente linear com $N^{1/2}$ [44], uma nova regressão linear é realizada

$$a_k(N) = b_{0,k} + b_{1,k}N^{1/2}, \quad k = 0, 1. \tag{1.33}$$

As temperaturas de desnaturação previstas $T_i'$, por sua vez, são comparadas com as temperaturas experimentais $T_i$ extraídas da literatura.

$$\chi^2 = \sum_{i=1}^{N} \left(T_i' - T_i\right)^2. \tag{1.34}$$

A minimização do $\chi^2$ é realizada numericamente variando os parâmetros através do método Nelder-Mead, também conhecida como método *downhill simplex* [61]. Obtidos os parâmetros, é possível estimar a temperatura de desnaturação assim como estudar a abertura parcial da hélice de uma sequência arbitrária. Esses perfis de abertura permitem acessar a dinâmica da desnaturação.

## 1.8 Conclusão

Introduzimos os principais conceitos usados neste trabalho que são a estrutura de DNA, a desnaturação térmica, o modelo Peyrard-Bishop e o método da equivalência termodinâmica. Nos próximos dois capítulos vamos mostrar como aplicamos estes conceitos a pares guanina-uracila em RNA e na obtenção de informações estruturais em DNA.

# Capítulo 2

# Estudo da estabilidade do par Guanina-Uracila em RNA

## 2.1 Introdução

O *mismatch* Guanina-Uracila (GU) foi descrito inicialmente por Crick na "hipótese de *wobble*" [62], na qual ele sugeriu que esse par se formava a partir de duas ligações de hidrogênio. Guanina-Uracila é o par não-complementar mais frequente no RNA e sua ocorrência no RNA não é acidental, apresentando importante papel biológico. O par GU está presente em quase todas classes de RNA de organismos abrangendo os três domínios filogenéticos. Como o RNA em dupla hélice adota a conformação A, não há a variação necessária nos sulcos (major e minor *groove*) para que os ligantes reconheçam, de forma que mecanismo utilizado para a identificação são os pares não-canônicos [8]. A existência do *mismatch* GU é biologicamente funcional, por possuir propriedades químicas e estruturais únicas, o par GU se torna alvo para o reconhecimento de biomoléculas [63]. Uma evidência da importância desse *mismatch* é a conservação da posição ao longo da evolução. O processo de reconhecimento da alanyl-tRNA sintetase (AlaRS), enzima que associa o aminoácido alanina ao seu tRNA, ocorre através de um único *mismatch* GU que está presente na mesma posição na maioria dos organismos. A troca do *mismatch* GU no tRNA AlaRS por um par Watson-Crick reduz a taxa de catálise de Ala [64].

Recentemente, o nosso grupo aplicou o modelo Peyrard-Bishop, juntamente com o método de equivalência termodinâmica para estudar a estabilidade térmica de RNA [46]. Usando um conjunto de dados de temperatura de desnaturação [65] foi possível calcular o valor das ligações de hidrogênio para os pares citosina-guanina (CG) e

adenina-uracila (AU). No caso de CG o valor das ligações de hidrogênio coincidiu perfeitamente com cálculos independentes de CG para DNA [45]. Já para AU este valor foi maior do que AT em DNA, confirmando medidas de NMR [66]. Os bons resultados para pares canônicos em RNA nos motivou a estender este tipo de tratamento para GU em RNA.

Neste capítulo apresentamos o estudo da estabilidade térmica do par Guanina-Uracila em RNA usando o modelo Peyrard-Bishop. Primeiro vamos revisar aspectos gerais da estrutura do RNA, em seguida discutimos as peculiaridades de GU em RNA. Posteriormente discutimos as medidas experimentais existentes de NMR e raios-X. Finalizamos descrevendo o nosso procedimento teórico e os resultados, em que mostramos que a variabilidade do número de ligações de hidrogênio pode ser bem descrita pelo modelo mesoscópico. Estes resultados foram publicados em Amarante and Weber [67].

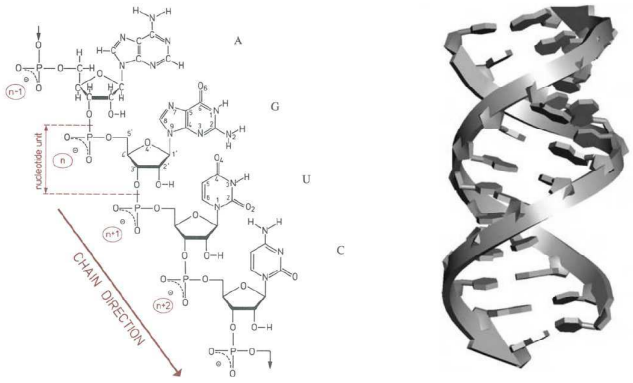



**Figura 2.1**
(a) Nucleotídeos que compõem o RNA e a direcionalidade da fita $5' \rightarrow 3'$, retirada de Saenger [25]. (b) Figura esquemática mostrando a estrutura de dupla hélice de r(UUAUAUAUAUAUAA) onde é possível notar a inclinação dos pares de base em relação ao eixo vertical da hélice característica do A-RNA. Extraída de Neidle [6].

### 2.1.1 Estrutura do RNA

Assim como o DNA, o ácido ribonucleico, RNA, é um polímero composto por nucleotídeos. O que diferencia os nucleotídeos que o compõem é o tipo do açúcar, no RNA é ribose, e uma das quatro bases nitrogenadas. Além da adenina (A), citosina (C) e guanina (G), o RNA contém uracila (U).

Essas bases se combinam formando pares, sendo os de maior afinidade e, portanto, observados com maior frequência, citosina associada à guanina, CG, através de três ligações de hidrogênio e a base adenina associada à uracila, AU, formando duas ligações de hidrogênio. Quando essas bases complementares se pareiam na configuração de maior estabilidade, os chamados pares Watson-Crick canônicos, exibem uma isostericidade que no RNA origina a dupla hélice do tipo A-forma, como mostrado na figura 2.1. A partir de imagens de cristalografia, foi observado que essas estruturas podem exibir uma variedade de combinações e geometrias na formação dos pares de base, resultando assim em diferenças de estabilidade. Embora a maioria dos pares sejam do tipo Watson-Crick canônicos, a existência dos pares não-Watson-Crick são determinantes na estrutura terciária do RNA.

A princípio não é possível prever a orientação relativa das bases apenas observando a estrutura química das bases. A nomenclatura proposta por Leontis and Westhof [68] se baseia nas fronteiras de interação das bases para descrever sem ambiguidade as formas possíveis de configuração dos pares de bases. Existem três regiões que possibilitam o estabelecimento de ligação de hidrogênio: fronteira Watson-Crick,

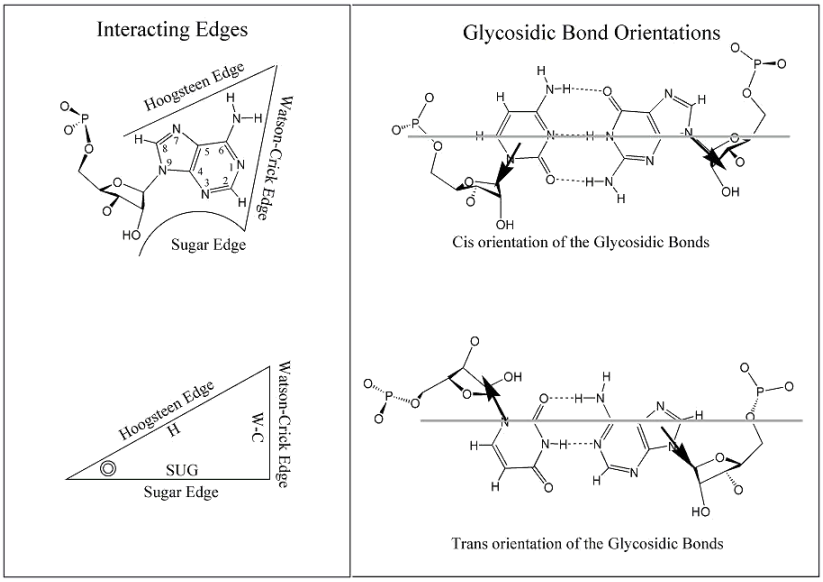



**Figura 2.2**
Geometria dos pares de base e as fronteiras de interação. Canto superior esquerdo: as três fronteiras de interação disponíveis para ligações de hidrogênio (Watson-Crick, Hoogsteen e *sugar-edge*) identificadas em uma adenosina. Canto inferior esquerdo: representação de uma base de RNA como um triângulo onde o círculo no canto, definido pelas fronteiras de açúcar e Hoogsteen, indica a posição da ribose. Direita: geometrias Cis e Trans de pareamento das bases. Figura adaptada de Leontis and Westhof [68].

| $N$ | Orientação da ligação glicosídica | Fronteiras interagentes | Orientação da fita |
|---|---|---|---|
| 1 | Cis | Watson-Crick/Watson-Crick | Antiparalela |
| 2 | Trans | Watson-Crick/Watson-Crick | Paralela |
| 3 | Cis | Watson-Crick/Hoogsteen | Paralela |
| 4 | Trans | Watson-Crick/Hoogsteen | Antiparalela |
| 5 | Cis | Watson-Crick/Sugar Edge | Antiparalela |
| 6 | Trans | Watson-Crick/Sugar Edge | Paralela |
| 7 | Cis | Hoogsteen/Hoogsteen | Antiparalela |
| 8 | Trans | Hoogsteen/Hoogsteen | Paralela |
| 9 | Cis | Hoogsteen/Sugar Edge | Paralela |
| 10 | Trans | Hoogsteen/Sugar Edge | Antiparalela |
| 11 | Cis | Sugar Edge/Sugar Edge | Antiparalela |
| 12 | Trans | Sugar Edge/Sugar Edge | Paralela |

**Tabela 2.1**
12 geometrias possíveis com duas ligações de hidrogênio para RNA.

fronteira Hoogsteen, e a fronteira do açúcar (*sugar-edge*). Além disso, referente à ligação glicosídica, as bases podem se orientar de maneira *cis* ou *trans* em relação à ligação de hidrogênio como mostrado na figura 2.2. Há, portanto, 12 geometrias possíveis com duas ligações de hidrogênio e cada geometria de par de base é nomeada a partir das fronteiras que interagem, conforme apresentado na tabela 2.1.

Para caracterizar a orientação relativa das bases e das fitas, Leontis and Westhof propuseram um esquema bidimensional em que a base é descrita por um triângulo retângulo no qual cada face representa uma fronteira de interação. A orientação do esqueleto de açúcar fostato em relação ao plano da página é indicada pela presença de uma cruz (5′ para 3′) ou de um círculo (3′ para 5′) mostrado na figura 2.2.

Além dessas conformações da tabela 2.1, nas quais o maioria das estruturas observadas se enquadram, há pares que apresentam pontes de hidrogênio bifurcadas e, dessa forma, se encontram em um estado intermediário das geometrias envolvendo duas fronteiras descritas. Nos pares bifurcados, dois átomos de hidrogênio se associam a um único átomo aceitador. Outra estrutura anômala observada ocorre quando há inserção de uma molécula de água no interior do par de bases, devida geralmente a rotação de uma base em relação a outra.

### 2.1.2 A estrutura de Guanina-Uracila

Em contraste com os pares complementares, os pares GU são levemente não auto-isostéricos, ou seja, um par GU não é isostérico em relação a UG [68, 69]. O par GU também apresenta uma leve não-isostericidade em relação ao pares cis-Watson-Crick/Watson-Crick complementares. Isostericidade entre pares de base quer dizer que as posições e as distâncias entre os átomos de carbono C1′ (figura 1.3) desses pares são muito semelhantes [7]. A distância C1′...C1′ no par de base GC (e CG) é de 10.77 Å e no par AU (e UA) essa distância é de 10.69 Å, enquanto no par GU essa distância mede 10.30 Å [8].

Algumas medidas de ressonância indicaram que a maioria dos *mismatches* GU apresenta duas ligações de hidrogênio [70], estando de acordo com a hipótese de Crick [62]. Contudo, medidas posteriores de ressonância [71] mostraram que o número de ligações de hidrogênio pode variar dependendo do contexto e é possível que seja o fator responsável pela variação de estabilidade observada. O par GU é capaz de formar uma ou duas ligações de hidrogênio de maneira estável em diferentes direções, algumas possibilidades de conformação do par podem ser visualizadas na figura 2.3.

Assim como os pares canônicos AU e GC, a configuração mais estável do *wobble* GU ocorre com a geometria cis Watson-Crick/Watson-Crick. Adicionalmente, observou-se que no caso do GU a estabilidade depende dos pares vizinhos, ou seja, do contexto das bases na sequência em que está inserido [70, 72, 73]. Em particular, quando dois *mismatches* GU estão adjacentes em *tandem* simétrico a estabilidade depende tanto da direção, 5′−UG−3′ é mais estável que 5′−GU−3′, quanto dos pares vizinhos, isto é, 5′G> 5′C> 5′U $\geq$ 5′A [73]. O contexto 5′GGUC3′ é uma exceção,

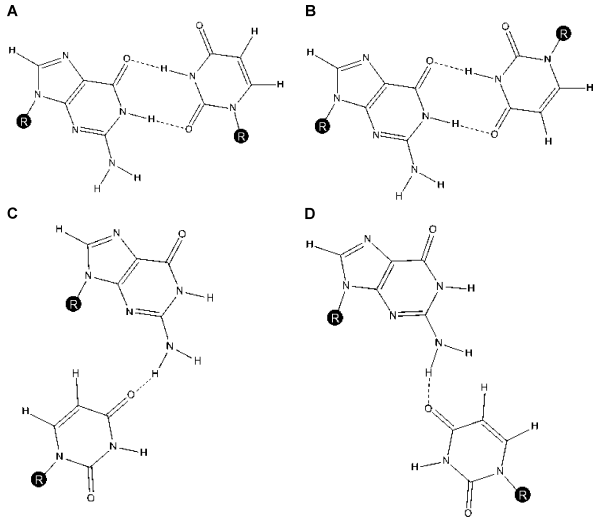



**Figura 2.3**
Exemplos de conformações do par GU com uma ou duas ligações de hidrogênio. (A) cis Watson-Crick/Watson-Crick, (B) trans Watson-Crick/Watson-Crick, (C) cis Hoogsteen/Sugar edge, and (D) trans Hoogsteen/Sugar edge. Figura retirada de Nguyen and Schroeder [75].

exibindo estabilidade semelhante ao 5′GUGC3′ [74]. Quando o par GU está presente no final da hélice, o par contendo G na terminação 5′ exibe um empilhamento mais estável com o par de base que o antecede em comparação com a situação em que G se situa no lado 3′. Essa estabilidade também se evidencia pela frequência superior com que essa configuração é observada em rRNAS [69].

No trabalho realizado por Sugimoto et al. [72], a estabilidade termodinâmica desse *mismatch* foi explorada a partir de medidas de temperatura de desnaturação realizadas em um conjunto de sequências, cada sequência contendo o par GU foi comparada com uma sequência igual que diferia apenas por conter AU na posição do GU. As temperaturas das sequências contendo o par GU foram inferiores às das sequências equivalentes com o par AU, apesar de, *a priori*, possuírem conteúdo similar de interações de hidrogênio.

### 2.1.3 Definição de contexto e notação

A princípio, como o modelo PB não diferencia a estrutura química das bases, não é possível determinar a partir de análises com esse modelo a orientação de uma base em relação a outra do mesmo par. No entanto, esse modelo mesoscópico é capaz de captar a diferença entre essas interações, através dos parâmetros associados aos potenciais de Morse. Isso é possível porque cada tipo de interação possui uma determinada estabilidade que será traduzida na intensidade dos parâmetros do modelo. Assim, para poder estudar GU em RNA com o modelo PB será necessário introduzir uma nova notação de pares de base para poder captar a dependência com os vizinhos do par GU.

Em alguns experimentos, houve evidência de que, em determinados casos, a dependência da estabilidade do *mismatch* GU extrapola os primeiros vizinhos. Em par-

ticular, foram observados [72] incrementos de energia distintos ao substituir GU por AU nos casos r(AUGCGU) → r(AUGCAU) e r(AUGCGCGU) → r(AUGCGCAU).

No entanto, a maioria dos dados da literatura indicam que os primeiros pares de bases que ladeam GU e a orientação da fita são os fatores mais determinantes na estabilidade desse *mismatch*.

Portanto, baseando a definição de contexto nesses fatores, nesse trabalho os pares GU são unicamente descritos a partir de trímeros de contexto NGN/NUN ou NUN/NGN, sendo N qualquer outra base. Como exemplo, o trímero de contexto GGC/CUG, representa um *mismatch* GU ladeado por GC na lateral 5′ e por CG no lado 3′. O par GU está sublinhado para facilitar a identificação. No caso da presença de GU no término da sequência, a notação adotada especifica um dímero terminal identificado como NG*/NU*, em que ** caracteriza o término da hélice e atua como um pseudo-par de um trímero.

Vamos ilustrar o desmembramento de uma sequência nos trímeros correspondentes em uma sequência contendo GU na terminação 5′ e um *tandem* interno, isto é dois pares GU lado a lado.

5′-UCAUGUGG-3′
3′-GGUGUACC-5′

a decomposição inicial em trímeros é dada por,

*UC/*GG, UCA/GGU CAU/GUG, AUG/UGU, UGU/GUA, GUG/UAC
UGG/ACC GG*/CC*

onde, conforme descrito, cada * representa as terminações das fitas. Há em seguida uma identificação de trímeros equivalentes devido à simetria, e uma consequente redução de tipos de trímeros. Por exemplo, como os trímeros

5′-AUG-3′ 5′-UGU-3′
3′-UGU-5′ 3′-GUA-5′

são equivalentes, apenas mantemos o trímero

5′-AUG-3′
3′-UGU-5′

A opção de manter AUG/UGU, em vez de UGU/GUA, se baseia unicamente na ordem alfabética e essa regra foi adotada sempre na ocorrência de redução por simetria. Para dímeros terminais, essa redução é feita de maneira análoga

5′-*UC-3′ ⇒ 5′-GG*-3′
3′-*GG-5′ 3′-CU*-5′

optamos por manter os dímeros com */* no lado $3'$, já que caracteres como asterisco * antecedem as letras na ordem alfabética. Portanto, após todas as reduções, os trímeros que compõe nosso exemplo são

G$\underline{\text{G}}$*/C$\underline{\text{U}}$*, UCA/GGU, CAU/GUG, A$\underline{\text{U}}$G/U$\underline{\text{G}}$U, A$\underline{\text{U}}$G/U$\underline{\text{G}}$U, CAU/GUG, GG*/CC*

Para considerar a diferença de estabilidade existente para pares GU situados em diferentes contextos, vamos atribuir uma constante de Morse única para cada *mismatch* GU em um trímero de contexto ou grupo de trímeros. A notação escolhida associa um parâmetro de Morse $D^{\alpha}$ independente para cada par GU em um trímero específico, onde $\alpha$ denota cada trímero de contexto. Para referir ao *mismatch* GU nesse contexto usamos $\text{GU}^{\alpha}$. Em relação aos parâmetros associados ao potencial de Morse $\text{GU}^{\alpha}$ é equivalente $\text{UG}^{\alpha}$, já que esse potencial modela a ligação de hidrogênio. É conveniente introduzir também a notação adotada para os grupos de trímeros de contexto que serão definidos posteriormente nesse trabalho, a cada grupo de trímeros de contexto $\beta$ será associado um parâmetro de Morse $D^{\beta}$.

Adotamos uma notação simplificada para identificar os pares vizinhos considerando o contexto do GU. Por exemplo, para descrever o par $\text{GU}^{A}$ seguido de $\text{UG}^{B}$, em vez de usar a notação usual, $5'$-GU-$3'$/$3'$-UG-$5'$ que resulta em $5'$-$\text{G}^{A}\text{U}^{B}$-$3'$/$3'$-$\text{U}^{A}\text{G}^{B}$-$5'$, esses pares serão identificados $\text{GU}^{A}\text{pUG}^{B}$,

## 2.2 Métodos

### 2.2.1 Modelo

O modelo mesoscópico escolhido para investigar a estabilidade do *mismatch* GU foi o modelo PB original[43] com empilhamento harmônico. Além de possuir menos parâmetros em relação às versões modificadas, para o RNA canônico os parâmetros foram obtidos para esse mesmo modelo [46]. Os parâmetros foram otimizados através do método descrito na seção 1.7. Para garantir a convergência da integral da função de partição [47] um pequeno ângulo (0.01 rad) $\theta$ foi introduzido no potencial de empilhamento

$$W(y_i, y_{i+1}) = \frac{k_{i,i+1}}{2}\left(y_i^2 - 2y_i y_{i+1}\cos\theta + y_{i+1}^2\right), \tag{2.1}$$

As ligações de hidrogênio, conforme dito anteriormente, são descritas através do potencial de Morse, Eq. (1.20)

$$V(y_n) = D_n\left(e^{-y_n/\lambda_n} - 1\right)^2$$

Onde os parâmetros $D$ e $\lambda$ dependem do par de base $i$, e a constante elástica está associada à interação entre os pares consecutivos $i$ e $i+1$.

### 2.2.2 Conjunto de dados experimentais de temperatura

Os dados de temperatura de desnaturação foram retirados de Chen et al. [76]. Esse artigo reúne uma base de dados com 80 sequências contendo todas as combinações de contexto possíveis de trímeros que incluem o *mismatch* GU ladeado por pares canônicos, AU e CG, considerando todas possibilidades de orientação. Essa base é o resultado da expansão de uma base de dados anterior Mathews et al. [77] construída a partir da coleção de dados de temperatura de desnaturação publicados. Chen et al. expandiram tal base através da realização de medidas de oligonucleotídeos elaborados para que o conjunto de sequências final contivesse todas as possibilidades de trímeros com GU. Cabe destacar que, como as temperaturas de desnaturação foram obtidas por grupos diferentes, é difícil estabelecer uma estimativa consistente da incerteza experimental associada a esse conjunto.

As sequências extraídas de Chen et al. [76], usadas para otimizar os parâmetros associados com o *mismatch* GU estão mostradas na tabela B.1, página 81 do apêndice B. A notação dos *mismatches* GU corresponde ao adotado após a otimização dos parâmetros, conforme descrito na seção 2.2.3. Esta base de dados inclui tanto sequências auto-complementares como não auto-complementares. Intuitivamente, em uma solução com a mesma concentração de fitas de RNA, se a sequência for auto-complementar, terá o dobro da probabilidade de encontrar a sequência complementar correspondente em comparação com o caso não auto-complementar. Para compensar isso a disponibilidade de fitas no caso não auto-complementar deve ser duas vezes a quantidade da situação auto-complementar.

Antes de podermos usar os dados de Chen et al. [76] precisamos realizar um ajuste das temperaturas em função da concentração $C_t$. As temperaturas de desnaturação para sequências auto-complementares (*self-complementary*), formada por duas fitas idênticas anti-paralelas, é dada por

$$T^{\mathrm{sc}} = \frac{\Delta H}{\Delta S + R \ln C_t} \tag{2.2}$$

e de não-auto-complementares (*non-self-complementary*), formadas por duas fitas diferentes, é dada por

$$T^{\mathrm{nsc}} = \frac{\Delta H}{\Delta S + R \ln C_t/4} \tag{2.3}$$

observe a diferença por um fator 4 no termo das concentrações entre as duas equações. Por esta razão, as tabelas de temperaturas de desnaturação publicadas em geral fornecem as temperaturas de sequências auto-complementares em concentração de 100 $\mu$M ($\mu$mol/L) e as não-auto-complementares em 400 $\mu$M, de tal maneira que os termos $R \ln C_t$ e $R \ln C_t/4$ resultem em valores idênticos e as entalpias e entropias possam ser comparadas entre sí.

No entanto, Xia et al. [65] observaram que isto ignora uma redução de entropia em $-R \ln(2)$ de sequências auto-complementares, já que estas são simétricas ao longo do eixo da hélice. Assim, introduzindo $-R \ln(2)$ na equação (2.2) temos

$$T^{sc} = \frac{\Delta H}{\Delta S - R \ln(2) + R \ln C_t} = \frac{\Delta H}{\Delta S + R \ln C_t/2} \tag{2.4}$$

Desta maneira, devemos considerar um fator de 2 e não de 4 para o ajuste das concentrações $C_t$. Em função desta observação, as temperaturas foram recalculadas para as sequências não-auto-complementares para o dobro da concentração das sequências auto-complementares, isto é, 200 $\mu$M, como realizado para RNA canônico por Xia et al. [65]. Isto também fica consistente com as parametrizações de pares CG e AU que serão usadas aqui [46].

### 2.2.3 Procedimento de minimização

Ao considerar a variedade de contextos de trímeros que contém GU, o número total de parâmetros associados a este *mismatch* é igual a 114 (40 parâmetros de Morse e 74 constantes de empilhamento). No entanto, temos apenas 80 sequências disponíveis para realizar a optimização dos parâmetros, ou seja o número de parâmetros excede o número de sequências. Por outro lado, mesmo em contextos diferentes muitos trímeros são equivalentes entre si, portanto o número de parâmetros pode ser reduzido consideravelmente se considerarmos algumas equivalências entre trímeros. A questão que fica é como estabelecer estas equivalências. Há duas estratégias possíveis, uma delas é levantar da literatura quais destes trímeros podem ser considerados equivalentes e a outra é usar o próprio procedimento de otimização para estabelecer estas equivalências. O conhecimento sobre os trímeros relacionados a GU na literatura é incompleto e os experimentos usados para seu estudo são variados, desta maneira estabelecer a equivalência pelo levantamento da literatura poderia introduzir um viés artificial nestas equivalências. Sendo assim, optamos por estabelecer a equivalência pelo próprio método de minimização mantendo constantes os 74 parâmetros de empilhamento ao que chamamos de rodada de minimização 1 (*minimization round 1* ou MR1). Em se-

guida, estudamos os resultados de MR1 e estabelecemos um conjunto de equivalências para reduzir o número de parâmetros de Morse que partimos para a minimização MR2. Como vários conjuntos de equivalência são possíveis, estabelecemos como critério que será aceito aquele que perturba menos os resultados de MR1, isto será discutido em detalhe na seção 2.3, página 38. As demais minimizações passam a incluir agora um conjunto reduzido de parâmetros de empilhamento e outros refinamentos para estabelecer a incerteza associada ao erro experimental dos dados de desnaturação.

Durante todas as rodadas de minimização os parâmetros $\lambda$ do potencial de Morse da equação (1.20) foram mantidos constantes em 0.03 nm. Isto vem da nossa experiência de que este parâmetro tem muito pouca influência sobre os resultados de minimização [45]. Para os parâmetros associados aos pares de base complementares AU e CG foram utilizados os parâmetros obtidos recentemente para o modelo PB [46].

Um problema importante do método de minimização é a ocorrência de mínimos locais, isto é, nenhum algoritmo de minimização é capaz de garantir que o valor mínimo encontrado seja de fato um mínimo global [61]. Para contornar este problema nós repetimos as minimizações do $\chi^2 = \sum_{i=1}^{N} (T_i' - T_i)^2$ (equação 1.34) sempre partindo de parâmetros iniciais diferentes. Nas rodadas iniciais (MR1–MR4) os parâmetros de entrada foram randomizados ao redor de um dado valor $p$ no intervalo $[0.5p, 1.5p]$. Dessa forma, para cada passo da minimização, o mínimo global é alcançado independentemente a partir de caminhos diferentes.

#### 2.2.3.1 Primeira rodada de minimização (MR1)

Decidimos considerar inicialmente que cada *mismatch* GU em trímeros de contexto diferentes possui um padrão único de ligação de hidrogênio. Isto é, para evitar que os resultados fossem enviesados por uma suposição de que determinados *mismatches* GU em contextos diferentes compartilham o mesmo número de ligações de hidrogênio, nessa etapa cada GU contido em trímeros distintos foi associado a um potencial de Morse diferente. Portanto, para cada trímero de contexto $\alpha$ presente nas sequências, um parâmetro Morse $D^{\alpha}$ foi atribuído, resultando em 40 parâmetros.

Como o conjunto de sequências da base de dados contém apenas 80 sequências, para que existam dados suficientes para realizar a otimização, nessa etapa todas as 74 constantes elásticas associadas ao GU foram fixadas em 2.5 eV/nm$^2$.

Foram realizadas 300 minimizações para que mínimos locais fossem evitados, cada uma dessas a partir de parâmetros iniciais de Morse $D$ diferentes, com um valor entre 15 meV e 45 meV. O valor $p = 30$ meV, ao redor do qual os parâmetros de Morse foram variados em $\pm 50\%$, corresponde ao potencial de Morse calculado para o par AU [46],

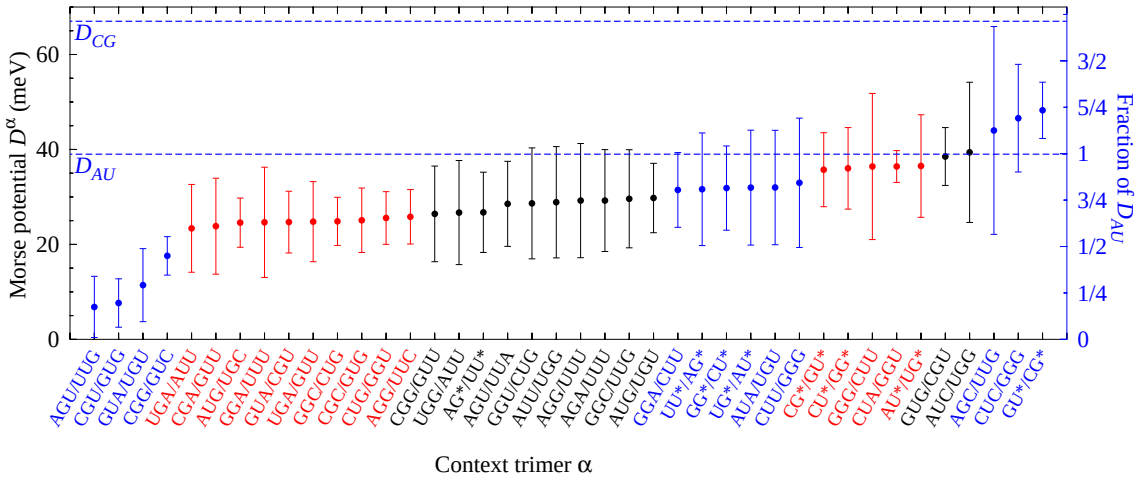



**Figura 2.4**
Parâmetros de Morse $D^{\alpha}$ médios obtidos para cada trímero de contexto $\alpha$ a partir da etapa de minimização MR1. Os trímeros são exibidos em ordem crescente de $D^{\alpha}$. Parâmetros de Morse associados aos pares AU e CG são mostrados como referência e representados por linhas azuis tracejadas foram obtidos da Ref. 46. Os valores exibidos em azul no eixo à direita são frações do parâmetro de Morse do par canônico AU $D_{AU}$. As cores dos trímeros no eixo horizontal são referentes ao agrupamento I de contextos usado em MR2.

já que se espera que GU apresente estabilidade comparável a esse par na maioria dos contextos.

Para MR1, o valor final obtido para o quadrado da diferença total foi $\chi^2 = 1453\ ^{\circ}C^2$, correspondendo a aproximadamente 6000 h em processadores de 2 GHz.

Os valores obtidos para os parâmetros de Morse associado a cada trímero estão exibidos na Fig. 2.4. A barra de erro mostrada na Figura 2.4 está associada à dificuldade numérica em realizar um número finito de rodadas de minimizações para percorrer um espaço de parâmetros de 40 dimensões para um modelo não linear.

#### 2.2.3.2 Segunda rodada de minimização (MR2) — Agrupamento I

Observando os resultados obtidos na etapa MR1, figura 2.4, verificamos que determinados trímeros de contexto mostram valores similares de parâmetros de Morse, indicando que possuem o mesmo número de ligações de hidrogênio. A partir dessa análise, os trímeros foram reunidos em 7 grupos de contexto, W, M1, M2, S1, S2, S3, S4, denominado agrupamento I.

Ao realizar esse agrupamento, obtemos um aumento considerável do número de ocorrência de GU por trímero de contexto nas sequências das base de dados como mostrado na tabela 2.2. Adicionalmente, o agrupamento reduz o número de parâmetros de Morse para 7. Nessa etapa de minimização MR2, atribuímos como valor inicial

| Trímero de contexto | $n$ | Agrupamento I | Agrupamento II |
|---|---|---|---|
| AGU/UUG | 6 | 23 (W) | 23 (W) |
| CGU/GUG | 8 | | |
| GUA/UGU | 5 | | |
| CGG/GUC | 4 | | |
| UGA/AUU | 4 | 39 (M1) | 39 (M1) |
| CGA/GUU | 2 | | |
| AUG/UGC | 5 | | |
| GGA/UUU | 5 | | |
| GUA/CGU | 4 | | |
| UGA/GUU | 4 | | |
| GGC/CUG | 2 | | |
| CGC/GUG | 4 | | |
| CUG/GGU | 5 | | |
| AGG/UUC | 4 | | |
| CGG/GUU | 5 | 35 (M2) | 35 (M2) |
| UGG/AUU | 2 | | |
| AG*/UU* | 4 | | |
| AGU/UUA | 4 | | |
| GGU/CUG | 1 | | |
| AUU/UGG | 1 | | |
| AGG/UUU | 2 | | |
| AGA/UUU | 2 | | |
| GGC/UUG | 6 | | |
| AUG/UGU | 8 | | |
| GGA/CUU | 8 | 23 (S1) | 45 (SA) |
| UU*/AG* | 4 | | |
| GG*/CU* | 4 | | |
| UG*/AU* | 4 | | |
| AUA/UGU | 2 | | |
| CUU/GGG | 1 | | |
| CG*/GU* | 4 | 22 (S2) | |
| CU*/GG* | 4 | | |
| GGG/CUU | 4 | | |
| CUA/GGU | 4 | | |
| AU*/UG* | 6 | | |
| GUG/CGU | 5 | 8 (S3) | 18 (SB) |
| AUC/UGG | 3 | | |
| AGC/UUG | 3 | 10 (S4) | |
| CUC/GGG | 3 | | |
| GU*/CG* | 4 | | |

**Tabela 2.2**
Número de ocorrências $n$ de *mismatches* GU em cada grupo de contexto. Os trímeros são exibidos em ordem ascendente de potencial de Morse de MR1 que também é a ordem exibida da figura 2.4. Também estão mostrados os agrupamentos tentativa e número de trímeros de contexto contidos em cada grupo.

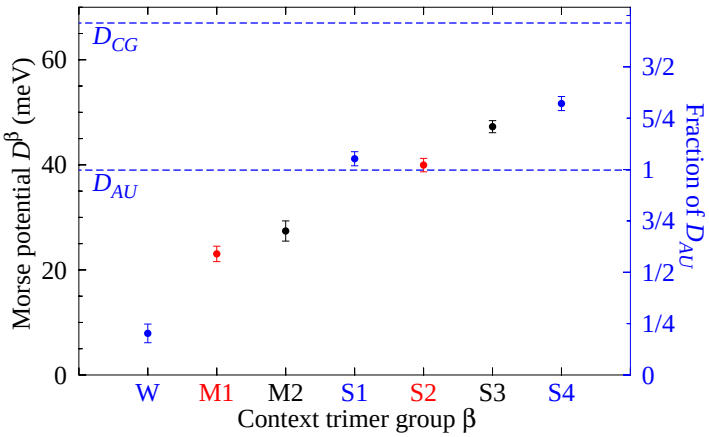



**Figura 2.5**
Potenciais de Morse $D^\beta$ médios obtidos para cada grupo trímero de contexto $\beta$ a partir de MR2. Os grupos de contextos são do agrupamento I exibidos na Tabela 2.2.

para cada um desses parâmetros a média dos valores obtidos para os contextos que o compõe obtidos na minimização anterior. Assim como realizado na MR1, esses valores de entrada foram randomizados em ±50%. No entanto, como o objetivo dessa etapa era de testar o agrupamento proposto, o processo de minimização foi realizado apenas 40 vezes O tempo computacional utilizado foi de 280 h e o $\chi^2 = 1426\ °C^2$. Novamente mantivemos as constantes de empilhamento fixas em 2.5 eV/nm$^2$. Os parâmetros otimizados obtidos em MR2 estão mostrados na figura 2.5. Ressaltamos que a barra de erro exibida está associada à dificuldade intrínseca à convergência numérica e, dessa forma, são inferiores às mostradas na figura 2.4 pois os parâmetros de Morse dessa etapa são variados ao redor dos valores que foram obtidos na otimização anterior.

#### 2.2.3.3 Terceira rodada de minimização (MR3) — Agrupamento II

Ao analisar a Figura 2.5, notamos que alguns grupos de contexto ainda apresentavam potenciais de Morse similares sugerindo que poderíamos reduzir ainda mais o número de parâmetros de Morse. Esse reagrupamento, denominado agrupamento II, foi definido reunindo os trímeros de contexto dos grupos S1 e S2 no novo grupo SA e os contextos dos grupos S3 e S4 no novo grupo SB. Em consequência, o agrupamento II reduz o número de parâmetros de Morse $D^\beta$ para 5, conforme tabela 2.2. Usando como valores de entrada os valores médios dos parâmetros de Morse de MR2, realizamos novamente o procedimento de minimização randomizando os valores iniciais. As 300 rodadas dessa etapa de minimização confirmaram a estabilidade do reagrupamento proposto, como pode-se observar na figura 2.6. Foram necessárias 1500 h de tempo computacional que resultaram em $\chi^2 = 1431\ °C^2$.

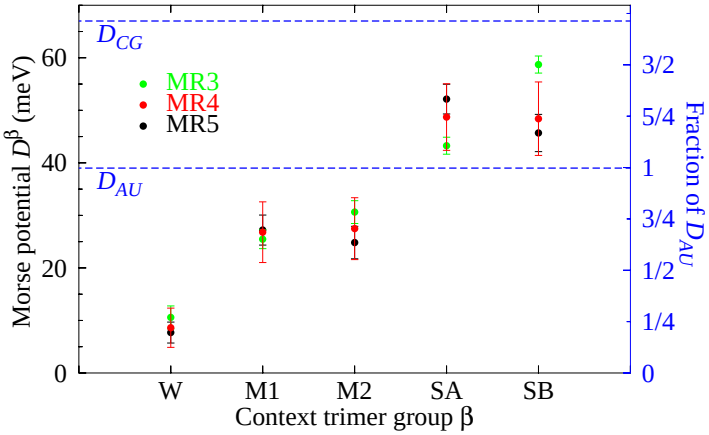



**Figura 2.6**
Potenciais de Morse $D^\beta$ médios para as rodadas de minimização MR3, MR4 e MR5. Agrupamento II conforme tabela 2.2.

#### 2.2.3.4 Quarta rodada de minimização (MR4) — inclusão das constantes de empilhamento

Após a otimização sistemática dos parâmetros de Morse que permitiu a identificação e agrupamento dos trímeros de contexto por estabilidade, foi possível incluir nessa etapa de minimização as constantes de empilhamento associadas. Portanto, utilizamos os valores obtidos em MR3 como entrada para os parâmetros $D^\beta$ e 2.5 eV/nm$^2$ para todas as 40 constantes de empilhamento. Em seguida, o mesmo procedimento anterior de randomizar esses parâmetros iniciais em ±50%, no entanto, dessa vez variando tanto $D$ quanto $k$, em um total de 45 parâmetros. Para MR4, o valor final do $\chi^2 = 1023$ °C$^2$ demandando 3600 h de tempo computacional. A variação dos parâmetros de Morse são mostradas na figura 2.6.

#### 2.2.3.5 Quinta rodada de minimização (MR5) — influência do erro experimental

Como toda medida experimental, os dados de temperatura de desnaturação utilizados para realizar as otimizações possuem uma incerteza experimental associada. Buscando obter uma estimativa do erro associados aos nossos parâmetros em consequência dessa incerteza, uma nova minimização foi realizada conforme o procedimento feito por Weber et al. [45]. Em vez de variar os parâmetros iniciais, os dados de temperatura de desnaturação são randomizados levando em consideração o erro experimental. Agora, cada rodada envolve a adição de uma quantia aleatória $\pm\delta T_i$ aos dados de temperatura de desnaturação extraídos da literatura $T_i$. Cada $\delta T_i$ é determinado de tal maneira que siga uma distribuição gaussiana de forma que o desvio padrão resultante da base de dados coincida com a incerteza experimental de 1.3 °C [65, 76]. Esse procedimento foi realizado 300 vezes, correspondendo a 3900 h de tempo computacional e $\chi^2 = 920$ °C$^2$. A figura 2.6 demonstra que os agrupamentos foram consistentes, já que os resultados

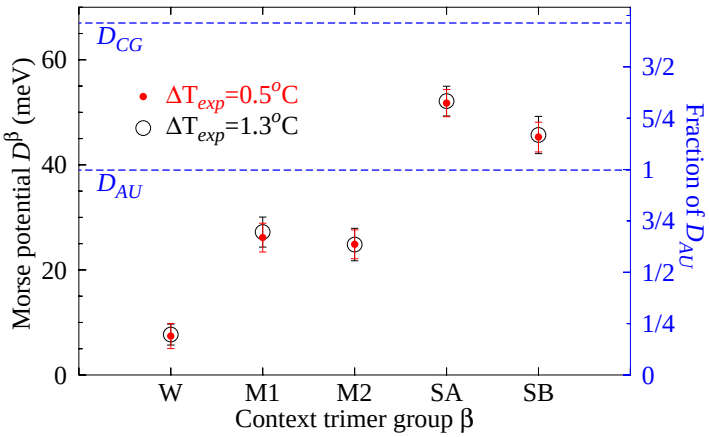



**Figura 2.7**
Comparação entre os potenciais de Morse $D^\beta$ médios obtidos considerando diferentes incertezas experimentais para as etapas de minimização MR5 (círculos pretos) e MR5′ (pontos vermelhos)

das constantes de Morse não exibiram uma variação significativa de uma etapa para a outra.

Idealmente o erro experimental usado para estimar o erro dos parâmetros deveria ser o erro relacionado com as medidas de desnaturação das sequências utilizadas para otimizar esses parâmetros. No entanto Chen et al. [76] não relata essa incerteza experimental, possivelmente por reunir uma série de medidas experimentais de temperatura de desnaturação realizada por outros grupos. Buscando contornar essa situação e sermos consistentes, o valor 1.3 °C escolhido é o mesmo que utilizado no trabalho de obtenção de parâmetros para os pares de base canônicos de RNA [46]. Para investigar o efeito de incerteza experimental menor repetimos todo procedimento realizado em MR5 para um valor diferente de erro, 0.5 °C, a essa etapa de teste nos referimos como MR5′.

Os parâmetros finais de MR5 e MR5′, com as respectivas estimativas de erro, estão exibidos nas figuras 2.7 e 2.8 respectivamente. Essencialmente o único efeito constatado foi foi uma pequena redução na barra de erro para os parâmetros de MR5′ em relação ao MR5, mas não houve alteração nos valores médios dos parâmetros calculados.

### 2.2.4 Teste de convergência

Devido à limitação de dados de temperatura de desnaturação, como descrito, o procedimento adotado para realizar a minimização e definir os agrupamentos se baseou em fixar as constantes elásticas na primeira rodada MR1. Com o objetivo de explorar se essa definição provocou um viés nos resultados, realizamos um teste que consistiu em uma minimização alternativa. Nessa rodada de minimização, os valores dos potenciais de Morse $D$ obtidos na última minimização MR5 foram atribuídos aos parâmetros $D$ correspondentes sem agrupamento, isto é, distinguindo todos os 40 contextos separadamente. Embora a situação de distinção dos contextos seja análoga à MR1, em vez

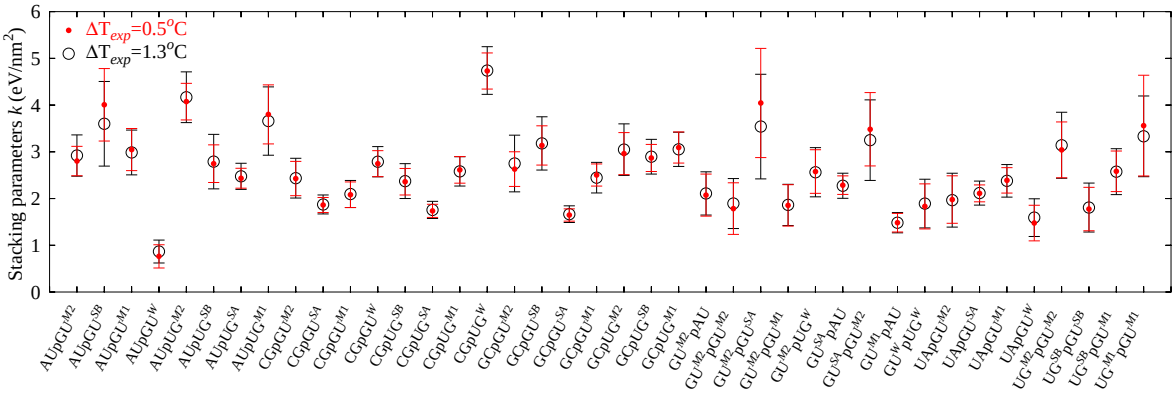



**Figura 2.8**
Parâmetros de empilhamento obtidos ao considerar diferentes erros experimentais nas etapas de minimização MR5 (círculos pretos) e MR5′ (pontos vermelhos)

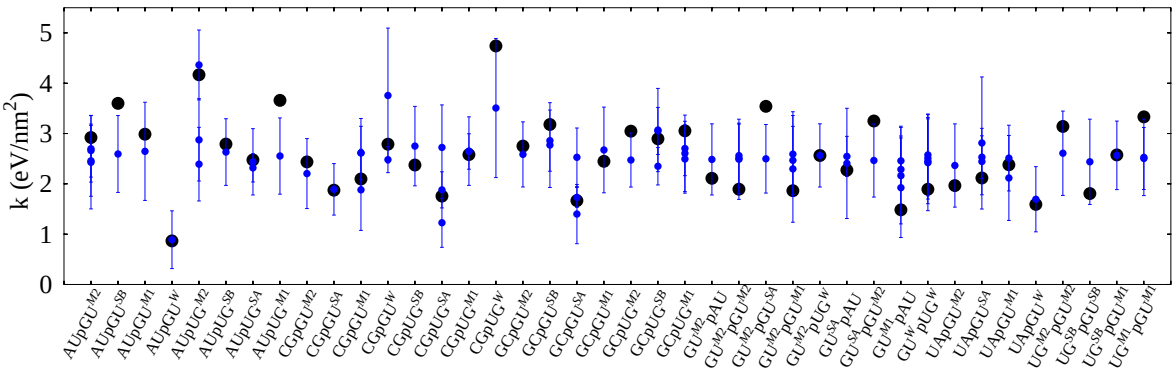



**Figura 2.9**
Comparação das constantes elásticas $k$ obtidas no MR5 (preto) e no teste de convergência (azul).

de variar os parâmetros $D$, optamos por mantê-los fixos e randomizar as 74 constantes elásticas associadas. Conforme mostrado na figura 2.9, a maioria dos valores de $k$ considerando todos os contextos (em azul) exibe um valor consistente com os $k$ em que ficou "alocado", associado aos grupo de contexto determinado no agrupamento II (exibidos no eixo horizontal e com os valores mostrados em preto).

## 2.3 Discussão dos resultados

Os resultados obtidos para os parâmetros associados ao GU mostraram que, apesar da simplicidade, o modelo PB é suficientemente sensível para determinar a dependência da estabilidade desse par *wobble* em relação ao contexto de sequência em que se insere. Notadamente, o potencial de Morse foi capaz de assimilar as diferenças das ligações

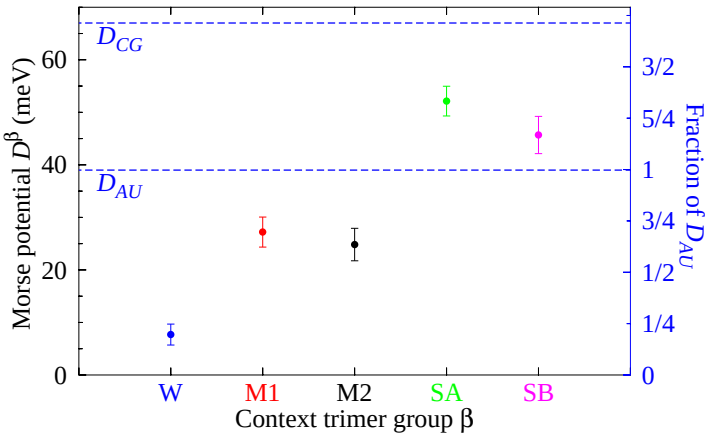



**Figura 2.10**
Potenciais de Morse $D^\beta$ médios obtidos para cada grupo trímero de contexto $\beta$ de MR5. Grupos de contexto são do agrupamento II mostrado na tabela 2.2. Os demais elementos da figura são como os exibidos na figura 2.4.

de hidrogênio. Conforme apresentado na introdução deste capítulo, essas diferenças se devem às diferentes conformações que as bases podem assumir uma em relação a outra. Infelizmente não é possível inferir a partir do valor do parâmetro de Morse $D$ qual a configuração que o par teria, conforme a tabela 2.1, naquele trímero de contexto. Contudo, a comparação dos valores obtidos entre eles e com os valores calculados para os pares de base canônicos de RNA [46] permite uma estimativa do número de ligações de hidrogênio. A fim de mensurar a qualidade da otimização dos parâmetros, calculamos o desvio médio de previsão de temperatura usando os parâmetros finais de MR5 $\Delta T = 2.7$ °C. Esse valor é levemente inferior ao desvio de previsão do modelo de próximos vizinhos (NN), $\Delta T = 3.0$ °C, calculado a partir da tabela 2 da referência 76.

Os valores finais para os parâmetros estão associados a cinco grupos de contextos definidos no agrupamento II, tabela 2.2, e obtidos na etapa MR5. São os cinco parâmetros de Morse mostrados na figura 2.10 e 40 constantes de empilhamento exibidas nas tabelas 2.3 e 2.4.

## 2.3.1 Comparação com dados experimentais

### 2.3.1.1 GU em tandem

Embora a configuração mais estável de GU seja a cis Watson-Crick/Watson-Crick apresentando duas ligações de hidrogênio, a não isostericidade desse par faz com que a direção da fita e os pares vizinhos afetem sua estabilidade. Esse efeito na estabilidade é acentuado em contextos contendo dois pares GU vizinhos, isto é um GU seguido de outro GU, configuração usualmente referida como GU em *tandem*. Este é uma das situações mais estudadas tanto experimentalmente como teoricamente [69–71, 73–76, 78, 79, 81, 83–86]

Especificamente para o *tandem* GUpUG, há uma divergência em relação ao pa-

| NN | $k$ | NN | $k$ | NN | $k$ |
|---|---|---|---|---|---|
| AUpGU$^{M2}$ | 2.9(4) | AUpGU$^{SB}$ | 3.6(9) | AUpGU$^{M1}$ | 3.0(5) |
| AUpGU$^{W}$ | 0.9(2) | AUpUG$^{M2}$ | 4.2(5) | AUpUG$^{SB}$ | 2.8(6) |
| AUpUG$^{SA}$ | 2.5(3) | AUpUG$^{M1}$ | 3.7(7) | CGpGU$^{M2}$ | 2.4(4) |
| CGpGU$^{SA}$ | 1.9(2) | CGpGU$^{M1}$ | 2.1(3) | CGpGU$^{W}$ | 2.8(3) |
| CGpUG$^{SB}$ | 2.4(4) | CGpUG$^{SA}$ | 1.8(2) | CGpUG$^{M1}$ | 2.6(3) |
| CGpUG$^{W}$ | 4.7(5) | GCpGU$^{M2}$ | 2.7(6) | GCpGU$^{SB}$ | 3.2(6) |
| GCpGU$^{SA}$ | 1.7(2) | GCpGU$^{M1}$ | 2.4(3) | GCpUG$^{M2}$ | 3.0(5) |
| GCpUG$^{SB}$ | 2.9(4) | GCpUG$^{M1}$ | 3.1(4) | GU$^{M2}$pAU | 2.1(5) |
| GU$^{SA}$pAU | 2.2(3) | GU$^{M1}$pAU | 1.5(2) | UApGU$^{M2}$ | 2.0(6) |
| UApGU$^{SA}$ | 2.1(3) | UApGU$^{M1}$ | 2.4(3) | UApGU$^{W}$ | 1.6(4) |

**Tabela 2.3**
Parâmetros de empilhamento finais $k$ em eV/nm$^2$ associados a *mismatches* GU únicos, obtidos em MR5.

| motivo | NN | $k$ | Medidas independentes das ligações de hidrogênio |
|---|---|---|---|
| UGpGU | UG$^{SB}$pGU$^{SB}$ | 1.8(5) | 2 ligações de hidrogênio NMR [71, 74], MD [78] |
| | UG$^{SB}$pGU$^{M1}$ | 2.6(5) | |
| | UG$^{M2}$pGU$^{M2}$ | 3.1(7) | 2 ligações de hidrogênio NMR [70, 79], X-ray [80], MD [78] |
| | UG$^{M1}$pGU$^{M1}$ | 3.3(9) | 2 ligações de hidrogênio NMR [74], MD [78] |
| GUpUG | GU$^{W}$pUG$^{W}$ | 1.9(5) | 1 ligação de hidrogênio NMR [70, 71], MD [78] 2 ligações de hidrogênio NMR [79], raio-X [81] |
| | GU$^{M2}$pUG$^{W}$ | 2.6(5) | |
| GUpGU | GU$^{M2}$pGU$^{M2}$ | 1.9(5) | 2 ligações de hidrogênio X-ray [82, 83] |
| | GU$^{M2}$pGU$^{M1}$ | 1.9(4) | |
| | GU$^{SA}$pGU$^{M2}$ | 3.2(8) | |
| | GU$^{M2}$pGU$^{SA}$ | 3.5(1) | |

**Tabela 2.4**
Constantes de empilhamentos $k$ em eV/nm$^2$ para GU em *tandem*, obtidas em MR5. As incertezas estão descritas em notação compacta. Também estão exibidas as referências que determinam independentemente o número de ligações de hidrogênio para cada configuração.

drão de ligações de hidrogênio que esta formação apresenta. Medições de NMR realizadas por He et al. [70] indicaram a possibilidade de ligações de hidrogênio mais fracas para alguns *mismatches* em *tandem* GUpUG, sugerindo talvez apenas uma ligação. Outras medidas, no entanto McDowell and Turner [74], McDowell et al. [79] apontaram para a hipótese de que os pares *wobble* GU, mesmo na configuração tandem, formam duas pontes de hidrogênio. Dados posteriores de NMR realizados pelo mesmo grupo [71] sugeriram apenas uma ligação de hidrogênio para a sequência específica r(GGCGUGCC)$_2$ contendo o padrão de *tandem* GUpUG. Contrastando com este resultado, ao analisar essa mesma sequência através de cristalografia de raio-X Jang et al. [81] inferiram duas ligações de hidrogênio e atribuíram a divergência com o padrão único obtido por Chen et al. [71] às condições experimentais distintas e em limitações da técnica.

Para realizar a análise da intensidade das ligações de hidrogênio com nossos resultados, é necessário identificar os trímeros de contexto associados a cada motivo e, assim, reconhecer o grupo de contexto e os parâmetros respectivos para os pares GU em questão. Para os *mismatches* em tandem essa correspondência está exibida na tabela 2.5. Apenas duas configurações do tipo *tandem* GUpUG aparecem no conjunto de dados para este motivo, um é GU$^W$pUG$^W$ no qual ambos os pares de bases estão no grupo mais fraco W de 8 meV, em contraste com o valor de 39 meV para duas ligações de hidrogênio AU [46]. A sequência r(GGCGUGCC)$_2$ estudada por Chen et al. [71] é exatamente deste tipo, como são também as três sequências de He et al. [70]. A outra configuração, GU $^{M2}$p$UG^W$, apresenta um par de base do tipo W e um com intensidade do potencial de Morse média M2 de 25 meV, para a qual não temos conhecimento de quaisquer medidas experimentais independentes de raio-X ou NMR. Todas as medidas que encontramos na literatura para esta formação apresentam informações para GUpUG flanqueado por pares idênticos em ambos os lados, o que implica os dois pares GU pertencerem ao mesmo grupo de contexto.

Com os novos parâmetros obtidos para GU, somos capazes de estudar os perfis de abertura de sequências, analisando o $\langle y \rangle$ para uma temperatura de interesse usando o software TfReg desenvolvido pelo nosso grupo [87] (veja apêndice A). Portanto, decidimos explorar o comportamento das sequências estudadas em Chen et al. [71] e as sequências canônicas análogas, devido aos resultados distintos para o padrão de ligação de hidrogênio expostos anteriormente. É possível notar na figura 2.11 que o potencial de Morse reduzido tem uma influência dramática nos perfis de abertura. Comparando o grupo forte SB, figura 2.11a, com o grupo mais fraco W dos potenciais Morse na figura 2.11b notamos uma diferença de 6 vezes no perfil médio de abertura $\langle y \rangle$. Observe que os parâmetros de empilhamento para GU em *tandem* quase não apresentam

| | par de base vizinho/trímero de contexto (grupo) | | | |
|---|---|---|---|---|
| NN | 5′G | 5′C | 5′U | 5′A |
| UGpGU | 5′-**G**UG / CGU (SB) | 5′-**C**UG / GGU (M1) | UGA / GU**U**-5′ (M1) | 5′-**A**UG / UGU (M2) |
| GUpUG | ∗5′-**G**GU / CUG (M2) | 5′-**C**GU / GUG (W) | GUA / UG**U**-5′ (W) | 5′-**A**GU / UUG (W) |

**Tabela 2.5**
Identificação dos trímeros associados aos *mismatches* GU em tandem simétrico de acordo com a direção e os pares de base vizinhos. Trímeros marcados com * nenhuma sequência da base de dados, referência 76, contém esse trímero de contexto apresentando GU *tandem* simétrico.

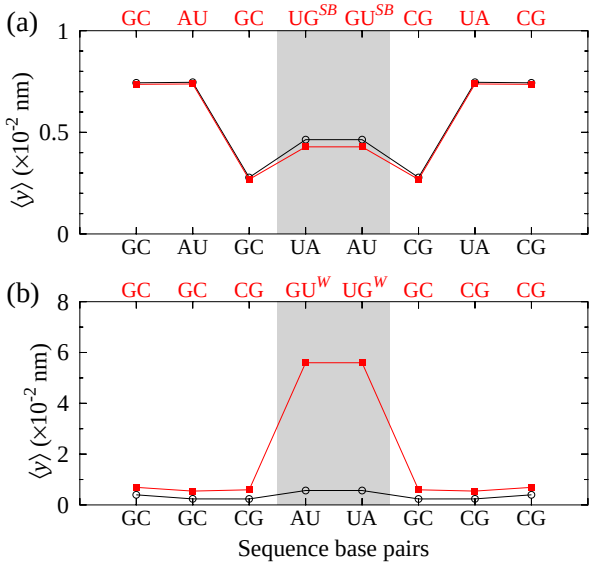



**Figura 2.11**
Perfil de abertura médio calculado para sequências contendo o tandem simétrico de mismatches GUpUG. São mostrados os perfis de abertura (em vermelho, eixos horizontais superiores) para as sequências (a) r(GGCGUGCC)$_2$ e (b) r(GAGUGCUC)$_2$ analisadas por Chen et al. [71]. As sequências canônicas análogas (círculos pretos, eixos horizontais interiores) foram obtidas a partir da substituição de G por A. A área sombreada destaca as posições onde as sequências contendo *mismatch* GU e as sequências canônicas diferem. Os perfis de abertura foram calculados a T= 150 K.

variação, consulte a tabela 2.4. A comparação com as sequências canônicas destaca a diferença de estabilidade em relação ao par AU no caso da sequência apresentando um *mismatch* do grupo fraco W.

As outras duas formações duplas, o *tandem* simétrico UGpGU e o assimétrico GUpGU, dependendo dos pares vizinhos, apresentam potenciais de Morse com intensidade mediana (M2, M1) ou fortes (SB) como mostrados na tabela 2.4 da página 40. Isto é consistente com os dados de NMR experimentais para UGpGU [71, 73, 79] e com os dados de difração de raios-X disponíveis para GUpGU [82, 83], que atribuem unanimemente duas ligações de hidrogênio para esta formação em tandem. Em particular, a partir da análise da estrutura cristalina de uma sequência, Kondo et al. [88] relatam o papel desempenhado por moléculas de água na estabilização dos pares *wobble* UGpGU que poderiam explicar porque observamos valores maiores para $D$ do que aqueles para AU. Além disso, a estabilidade do tandem UGpGU pode ser investigada mais detalhadamente em termos dos pares de bases que o flanqueiam [73, 80]. Por exemplo, a

ocorrência mais estável desse motivo se dá quando o UGpGU é ladeado simetricamente por pares de bases GC

5′-UCAUGUGG-3′
3′-GGUGUACC-5′

Na tabela 2.5 estão exibidos todos os trímeros associados aos GU tandem identificados pela base que o flanqueia na direção 5′. Dessa forma podemos observar que a estabilidade dos parâmetros de Morse relacionados à formação UGpGU obedece a tendência prevista $5'G > 5'C > 5'U \geq 5'A$ [73, 80], lembrando que o potencial de Morse para M2 é levemente menor que para M1. Uma tendência semelhante é observada para o tandem GUpUG para o qual 5′G apresenta potencial de Morse maior que para os outros contextos. No entanto, é necessário salientar que a situação desse motivo flanqueado por 5′G é inferida apenas de potenciais Morse já que não existe uma sequência real com essa configuração no conjunto de dados.

#### 2.3.1.2 GU terminal

Há um consenso geral de que pares terminais GU estabilizam a hélice [69, 75, 86, 89]. Contudo, não é claro se essa estabilização da fita dupla se deve à ligação de hidrogênio ou devido a interações de empilhamento. Os nossos resultados concordam em grande parte com a estabilidade aumentada que se observa experimentalmente, 8 pares GU terminais se encontram nos grupos de contexto com parâmetros de Morse maiores, SA (52 meV) e SB (46 meV), exceto por AG*/UU* situado no grupo intermediário M2 (25 MeV). Dessa forma nossos dados sugerem que, para os contextos contendo GU terminal identificados nos grupos SA e SB, esta estabilização pode ser atribuída a um aumento da ligação de hidrogênio.

Foi sugerido que a interação de empilhamento desempenha um papel na estabilidade dos *mismatches* GU terminais [69, 89]. Por apresentar uma sobreposição de empilhamento mais elevada, aqueles que possuem G na terminação 5′ são considerados mais estáveis [69, 89]. Na tabela 2.6 mostramos os pares terminais, selecionados da tabela 2.3, agrupados de acordo com a posição da base G (5′ ou 3′) e além disso separados por tipo de base que os flanqueiam. Nos nossos resultados, a constante de empilhamento para 5′G supera o da posição oposta 3′G para os casos em que os pares de base flanqueadores são GC, CG and UA, concordando em sua maioria com os resultados experimentais [69, 89]. No entanto, para o caso em que o par de base vizinho é o AU, se observa um empilhamento menor para 5′G. A estabilidade nessa situação parece ser compensada por um potencial muito maior, situado no grupo de Morse SA.

| BP | 3′-end | NN | $k$ | 5′-end | NN | $k$ |
|---|---|---|---|---|---|---|
| AU | AG*/UU* | AUpGU$^{M2}$ | 2.9 | UU*/AG* | GU$^{SA}$pAU | 2.2 |
| GC | GG*/CU* | GCpGU$^{SA}$ | 1.7 | CU*/GG* | CGpUG$^{SA}$ | 1.8 |
| CG | CG*/GU* | CGpGU$^{SA}$ | 1.9 | GU*/CG* | GCpUG$^{SB}$ | 2.9 |
| UA | UG*/AU* | UApGU$^{SA}$ | 2.1 | AU*/UG* | AUpUG$^{SA}$ | 2.5 |

**Tabela 2.6**
Identificação das constantes de empilhamento $k$ (eV/nm$^2$) associadas aos trímeros de contexto contendo GU terminal. Cada linha exibe o par de base (BP) que ladeia o GU terminal. Estes parâmetros constam também na tabela 2.3.

#### 2.3.1.3 GU interno

Os *mismatches* GU únicos internos aparentemente apresentam menor dependência do contexto na estabilidade. Provavelmente isso se deve a não isostericidade do par GU ser leve, ao ser flanqueado por pares complementares em ambos os lados deve normalizar sua estabilidade. As medições iniciais [70] sugeriram o padrão de duas ligações de hidrogênio para GU único em todos os contextos, o que foi confirmado por meio de medições de raios-X para alguns contextos específicos [90]. Nossos resultados endossam essa observação para a maioria dos casos como mostrado na figura 2.10 da página 39. Contudo obtivemos alguns valores anômalos para a configuração de GU individual interno. O trímero C<u>G</u>G/G<u>U</u>C, como pode ser verificado na tabela 2.2, da página 34, se situa no grupo W que apresenta um parâmetro de Morse muito baixo, de 8 meV. Isto sugere uma única ligação de hidrogênio para este contexto GU particular. Ao buscar dados experimentais com sequências apresentando essa configuração, encontramos uma estrutura cristalina analisada por Kondo et al. [88] contendo um GU neste contexto particular, que em contraste com nossos resultados prevê duas ligações de hidrogênio. Essa disparidade pode estar associada à cristalização da amostra de RNA para realização da medida de raios-X, enquanto os dados experimentais utilizados aqui, dos quais nossos parâmetros são derivados, são realizados com RNA em solução.

Já os resultados associados à interação de empilhamento do par GU único, a constante elástica do empilhamento para AUpGU$^W$ se destaca por apresentar um valor muito menor do que a média como mostrado na tabela 2.3. Não obstante, uma configuração semelhante, obtida apenas revertendo a direção do par GU, AUpUG$^W$, exibe um parâmetro de empilhamento mais forte que a média. Infelizmente, não temos conhecimento de quaisquer medições independentes que poderiam ser usadas para comparação nesses casos, já que a maioria das medições estruturais não fornecem uma estimativa das forças de interação de empilhamento.

### 2.3.2 Comparação com outros resultados computacionais

As simulações computacionais baseadas em dinâmica molecular e cálculos de mecânica quântica feitas pelo grupo de Pan et al. [78] também analisaram o caso r(GGCGUGCC)$_2$ e estão de acordo com a hipótese de uma única ligação de hidrogênio. No entanto, adicionalmente sugerem que a estabilidade menor do motivo GUpUG poderia também estar associada às interações de empilhamento. Quanto a essa possibilidade levantada por Pan et al. [78] de que as interações de empilhamento poderiam ser responsáveis por uma estabilidade reduzida de GUpUG, ao confrontar com nossos resultados de constantes de empilhamento, não encontramos nenhuma diferença em particular em relação a outras formações tipo *tandem* como pode ser visto na tabela 2.4. Na verdade, os parâmetros de empilhamento não mostram diferença sequer em relação aos pares de bases canônicas CG ou AU [46]. Portanto, as interações de empilhamento não parecem ser a causa primária para a instabilidade da formação GUpUG restando apenas a hipótese de uma única ligação de hidrogênio como explicação plausível.

Como os cálculos de DFT existentes modelam apenas o par GU isolado [91–96], não podemos fazer uma análise de dependência de contexto em relação a esses resultados. Resta tentar comparar com nossos dados para o *mismatch* GU único. Embora eles sejam capazes de considerar diversos tipos de geometrias de pares de bases, expostas na introdução, tais como cis Watson-Crick/*sugar edge* [93] ou *sugar edge*/*sugar edge* [92], eles não conseguem especificar qual dessas conformações será assumida em uma determinada sequência. Por outro lado, embora o nosso modelo seja capaz de predizer a estabilidade do GU dependendo do contexto, como dito anteriormente nossos resultados não possuem detalhamento suficiente para inferir a geometria de pares de bases tornando impraticável a comparação direta com os resultados de DFT.

## 2.4 Conclusão

Conforme exposto, NMR e cristalografia algumas vezes chegam a resultados conflitantes. Portanto, entendemos que ter um terceiro método mais simples e completamente independente é muito oportuno para resolver esses conflitos. Nosso método proporciona uma avaliação independente das ligações de hidrogênio e interações de empilhamento. Além disso nossos resultados são derivados de temperaturas de desnaturação que são, em princípio, muito mais fáceis de realizar do que experiências de difração de raio-X ou NMR. Dessa forma, há geralmente mais dados de temperatura de desnaturação, abrangendo mais configurações de sequência. Isto nos permitiu, por exemplo, estabele-

cer previsões para configurações GU que atualmente não foram investigados por outras técnicas.

# Capítulo 3

# Modelo Estrutural Bidimensional

## 3.1 Introdução

Conforme visto no capítulo 1, o modelo Peyrard-Bishop (PB) original assume dois graus de liberdade para cada par de base, confinando o DNA a uma estrutura bidimensional. Para o estudo da estabilidade térmica do DNA este modelo 2D é bastante razoável, já que próximo da temperatura de desnaturação é necessário que a hélice tenha sido completamente desfeita (*unwinding*). Porém, na formulação original da equação (1.5), essa simplificação também eliminou todas as informações estruturais associadas à hélice.

Neste capítulo mostraremos que tal simplificação é, na verdade, desnecessária, mesmo a Hamiltoniana 2D pode conter parâmetros relacionados à estrutura do DNA. Para isso vamos partir de formulações 3D da Hamiltoniana e chegar a uma forma modificada do modelo Peyrard-Bishop que contém um dos parâmetros estruturais mais importantes que é o passo da hélice. O interesse biológico no parâmetro passo de hélice extrapola a simples caracterização da estrutura do DNA. Essa distância é o passo relacionado ao movimento de enzimas (motores moleculares) ao longo do DNA, como a RNA polimerase, responsável pela transcrição do DNA em RNA, e a DNA polimerase, pela replicação de DNA [97].

A formulação 2D da Hamiltoniana com informações estruturais permite aplicar todo ferramental de cálculo de parâmetros já desenvolvido que usa o conceito de equivalência termodinâmica descrito na seção 1.7 da página 17. Isto por sua vez leva à pergunta se é possível obter parâmetros estruturais a partir de dados de temperatura de desnaturação.

Existem várias maneiras de modelar a influência da hélice em modelos mesoscópicos [100–103], mas no contexto do modelo PB os mais importantes são de Barbi et al. [104] e Cocco and Monasson [99]. Eles introduzem vários elementos para descre-

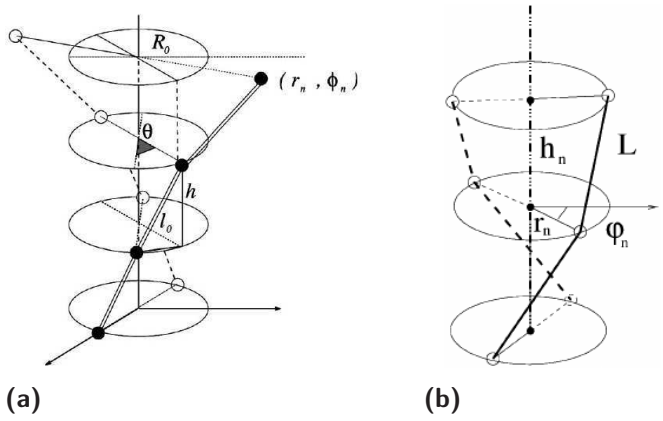



**Figura 3.1**
Representação esquemática do modelo helicoidal (a) de planos fixos ($h$ constante) de Barbi et al. [98] e (b) com ligação de fosfato fixa ($L$ constante) de Cocco and Monasson [99]. Figura da referência 98

ver a dupla hélice dentro do contexto do modelo PB e ao mesmo tempo procurando manter o número de parâmetros ao mínimo. No modelo proposto por Barbi et al. [98] a distância $h$ entre os planos de pares de base, conhecida como *rise*, é mantida constante como mostrado na figura 3.1a. Já modelo de Cocco and Monasson [99] essa distância é variável, mas o comprimento da ligação de fosfato $L$ é fixo como mostrado na figura 3.1b.

Aqui vamos usar o modelo de Barbi et al. [98] que mantém a distância $h$ constante para realizar a aproximação para a Hamiltoniana 2D. Este procedimento está descrito na seção 3.2. Uma vez obtida a nova Hamiltoniana 2D, nosso objetivo passa a ser a aplicação do método de otimização descrito na seçãos 1.7 para obter o passo da hélice a partir de dados de desnaturação de DNA. A motivação disso é que o passo da hélice geralmente é considerado como sendo 3,4 Å, mas na realidade esse parâmetro depende do contexto da sequência [10]. O processo de otimização do novo parâmetro estrutural está descrito na seção 3.3. Finalmente, realizamos um estudo completo para comparar os parâmetros calculados com dados experimentais que detalhamos na seção 3.4.

## 3.2 Modificação do Hamiltoniano

Seria interessante aplicar a parametrização para esses Hamiltonianos helicoidais tridimensionais para obter mais informações estruturais sobre o DNA. Contudo, nosso método de obtenção dos parâmetros baseado na técnica de equivalência termodinâmica atualmente está apto a lidar apenas com hamiltonianas bidimensionais. Vislumbramos

na adaptação de modelos 3D torsionais para formatos 2D simplificados, uma opção para obter os parâmetros de DNA com os métodos existentes. Nesse trabalho, consideramos o modelo proposto por Barbi et al. [98]. A Lagrangiana do modelo helicoidal proposto por Barbi et al. [98] é descrita por

$$L = m\sum_n(\dot{r_n^2} + r_n^2\dot{\phi_n^2}) - D\sum_n\left(\exp[-a(r_n - R_0)] - 1\right)^2$$
$$-k\sum_n(l_{n,n-1} - l_0)^2 - S\sum_n(r_n - r_{n-1})^2\exp[-b(r_n + r_{n-1} - 2R_0)]$$

onde

$$l_0 = \sqrt{h^2 + 4R_0^2 sen^2(\theta/2)} \tag{3.1}$$

e

$$l_{n,n-1} = \sqrt{h^2 + r_{n-1}^2 + r_n^2 - 2r_{n-1}r_n\cos(\phi_n - \phi_{n-1})}. \tag{3.2}$$

são respectivamente a distância de equilíbrio e a distância real entre as bases vizinhas $n$ e $n-1$ de uma fita, onde $R_0$ é a distância de equilíbrio entre as bases que compõem um par e $r_n$ é a distância de fato entre as bases do par $n$. O ângulo de rotação de um par de base $n$ em relação ao par anterior $n-1$, chamado de ângulo de torção, é descrito por $\phi_n - \phi_{n-1}$. O ângulo de torção de equilíbrio da hélice é descrito por $\theta$. Na conformação B-DNA, o valor de $\theta$ à temperatura ambiente é dado por $2\pi/10.4$.

Obtivemos o Hamiltoniano a partir dessa Lagrangiana, da seguinte forma

$$H = p_\alpha\dot{q}^\alpha - L \tag{3.3}$$

$$p_{r_n} = \frac{\partial L}{\partial\dot{r_i}} = 2m\dot{r_i} \qquad \dot{r_i} = \frac{p_{r_i}}{2m} \tag{3.4}$$

$$p_{\phi_i} = \frac{\partial L}{\partial\dot{\phi_i}} = 2m\dot{\phi_i}r_i^2 \qquad \dot{\phi_i} = \frac{p_{\phi_i}}{2mr_i^2} \tag{3.5}$$

Portanto o Hamiltoniano é dado por

$$H = \sum_n\frac{p_{r_n}^2}{4m} + \frac{p_{\phi_n}^2}{4m{r_n}^2} + D\sum_n(\exp[-a(r_n - R_0)] - 1)^2$$
$$+k\sum_n(l_{n,n-1} - l_0)^2 + S\sum_n(r_n - r_{n-1})^2\exp[-b(r_n + r_{n-1} - 2R_0)]$$

Se fazemos $\phi_i = 0$ para todo $i \leq N$ e $\theta = 0$, e ignoramos o último termo que modifica o potencial de empilhamento, o modelo helicoidal se torna plano e é possível compará-lo com o modelo PB. Usando as variáveis de PB no modelo de Barbi et al.

[98] (com $\phi_i = 0$ e $\theta$=0) (veja figura 3.2):

$$l_{n,n-1} = \sqrt{h^2 + ((r_n - R_0) - (r_{n-1} - R_0))^2} \tag{3.6}$$

$$= \sqrt{h^2 + \left(\frac{y_n - y_{n-1}}{\sqrt{2}}\right)^2} \tag{3.7}$$

$$= \sqrt{h^2 + \frac{1}{2}(y_n - y_{n-1})^2} \tag{3.8}$$

$$2(r_n - R_0) = \sqrt{2}y_n \tag{3.9}$$

$$\dot{r_n} = = \frac{\sqrt{2}}{2}\dot{y_n} \tag{3.10}$$

portanto

$$p_{r_n} = 2m\dot{r_n} = \sqrt{2}m\dot{y_n} = \sqrt{2}q_n \tag{3.11}$$

Reescrevendo o modelo proposto por Barbi et al. [98] e "anulando os ângulos", obtivemos um novo Hamiltoniano 2D

$$H = \sum_n \frac{q_n^2}{2m} + D\sum_n (\exp[-a\sqrt{2}y_n] - 1)^2 + k\sum_n \left(\sqrt{h^2 + \frac{1}{2}(y_n - y_{n-1})^2} - h\right)^2 \tag{3.12}$$

Como o Hamiltoniano agora é dependente de uma única variável, $y$, somos capazes de aplicar o método de otimização dos parâmetros descrito na seção 1.7 da página 17. O novo Hamiltoniano, que vamos chamar de Modelo Estrutural Bidimensional (MEB), inclui informação estrutural: a distância entre pares de base, o "passo da hélice".

Escrevendo os termos de potencial de ligação de hidrogênio e de empilhamento separadamente e introduzindo a possibilidade de termos parâmetros diferentes para cada tipo de próximos vizinhos, temos

$$V(y_n) = D_n(e^{-a'y_n} - 1)^2 \tag{3.13}$$

onde $a' = a\sqrt{2}$, e

$$W(y_n, y_{n-1}) = k_{\alpha\beta}\left(\sqrt{h_{\alpha\beta}^2 + \frac{1}{2}(y_n - y_{n-1})^2} - h_{\alpha\beta}\right)^2 \tag{3.14}$$

onde $\alpha\beta$ representa a configuração de próximos vizinhos entre o sítio $n-1$ e $n$. No caso de pares canônicos, agora além das 10 constantes de empilhamento $k_{\alpha\beta}$ existem mais 10 parâmetros de *rise* $h_{\alpha\beta}$.

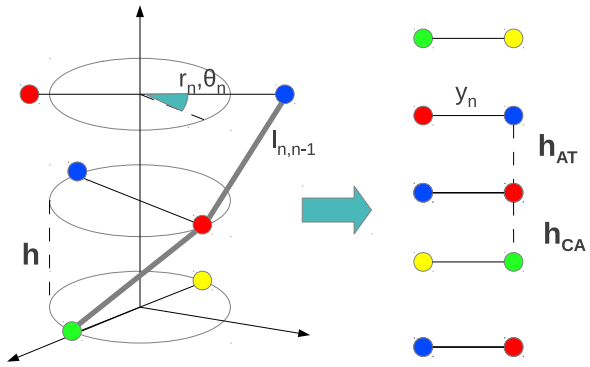



**Figura 3.2**
Esquema mostrando as variáveis e graus de liberdade da geometria helicoidal adotada por Barbi et al. [104]. A seta representa a ação de zerar os ângulos para obter a geometria adotada pelo Modelo Estrutural Bidimensional proposto, envolvendo o parâmetro passo de hélice $h$.

### 3.2.1 Comparação do MEB com o PB original

Comparando os elementos do Hamiltoniano do Modelo estrutural bidimensional, equações (3.13) e (3.14), com o Hamiltoniano original PB:

| Modelo Estrutural Bidimensional (MEB) | Modelo Peyrard-Bishop (PB) |
|---|---|
| $V(y_n) = D_n(e^{-\frac{1}{\lambda'}y_n} - 1)^2$ | $V(y_n) = D_n\left(e^{-\frac{1}{\lambda}y_n} - 1\right)^2$ |
| $W(y_n, y_{n-1}) = k_{\alpha\beta}\left(\sqrt{h_{\alpha\beta}^2 + \frac{1}{2}(y_n - y_{n-1})^2} - h_{\alpha\beta}\right)^2$ | $W(y_n, y_{n-1}) = \frac{k_{\alpha\beta}}{2}(y_n - y_{n-1})^2$ |

observamos que o potencial relativo às pontes de hidrogênio permaneceu sendo caracterizado pelo potencial de Morse. O potencial que descreve o empilhamento, por sua vez, inclui um novo parâmetro, $h_{\alpha\beta}$, que descreve o passo da hélice. Note que no caso particular em que se faz $h_{\alpha\beta} = 0$ recupera-se o potencial de empilhamento original do modelo PB. Nos potenciais acima substituímos $a = 1/\lambda$ e $a' = 1/\lambda'$ para ficar consistente com a notação adotada anteriormente.

É interessante ressaltar que no artigo original do modelo Peyrard-Bishop [43] já foi assumido uma aproximação importante ao simplificar o termo de empilhamento. A interação de empilhamento entre a bases vizinhas $u_n$ e $u_{n-1}$ é escrita simplesmente como um potencial harmônico $\frac{1}{2}k(u_n - u_{n-1})^2$, veja também a equação (1.5). No entanto, quando observamos que a interação ocorre ao longo da fita mas o deslocamento é transversal, veja a figura 3.3, o termo de empilhamento deveria ser diferente. Chamando de $h$ a distância de equilíbrio entre dois pares de bases vizinhos, o deslocamento dessa distância decorrente da movimentação de $u_n$ e $u_{n-1}$ na direção transversal das bases $n$ e $n-1$ respectivamente, é dado por

$$\Delta h = \sqrt{h^2 + (u_n - u_{n-1})^2} - h \tag{3.15}$$

Ou seja o potencial harmônico associado a esse deslocamento deveria ser descrito por $\frac{1}{2}k\left(\sqrt{h^2 + (u_n - u_{n-1})^2} - h)\right)$. É interessante notar a semelhança deste termo com o

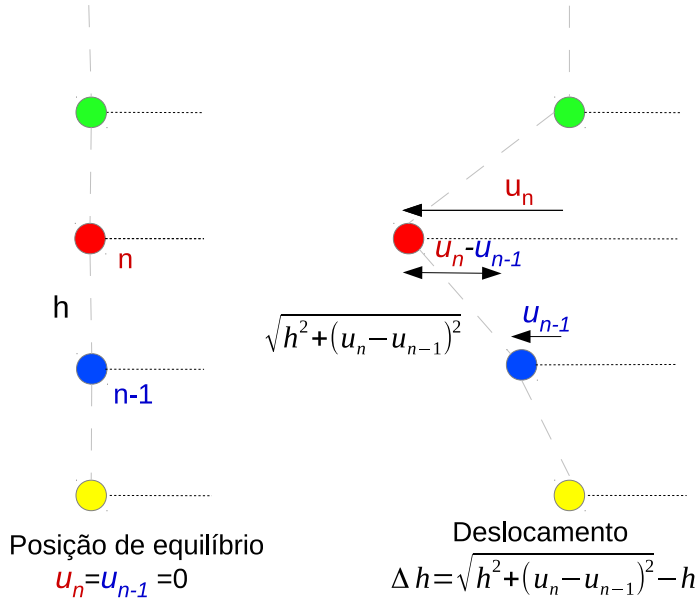



**Figura 3.3**
Esquema do deslocamento do empilhamento ao longo de uma das fitas do DNA. À direita a fita com todas as bases em posição de equilíbrio. À esquerda a fita com o deslocamento transversal.

potencial de empilhamento que obtivemos na equação (3.12).

## 3.3 Otimização dos parâmetros estruturais

O conjunto de dados experimentais escolhidos para realizar a otimização dos parâmetros do modelo foi retirados de Owczarzy et al. [60]. Esse conjunto consiste em 93 sequências cujas temperaturas de desnaturação foram medidas para cinco concentrações salinas diferentes, a saber 69 mM, 119 mM, 220 mM, 621 mM e 1020 mM de sódio $Na^+$. O erro experimental associado às medidas de temperatura por absorção ultravioleta é de 0.3 °C. Esse mesmo conjunto foi utilizado por Weber et al. [45] para obter os parâmetros do modelo PB não-modificado.

### 3.3.1 Procedimento de minimização

A otimização dos parâmetros foi realizada separadamente para cada concentração salina. Como as sequências utilizadas apenas envolvem pares canônicos, é necessário otimizar dez constantes elásticas $k$, dez passos de hélice $h$ e quatro parâmetros associados ao potencial de Morse (dois $D$ e dois $\lambda$), totalizando vinte e quatro parâmetros. Para obter os valores do passo da hélice $h_{\alpha\beta}$, a minimização da equação 1.34 na seção 1.7, foi realizada em três etapas que estão descritas à seguir. É necessário destacar que vários procedimentos de minimização foram testados para estudar a convergência antes que o processo aqui exposto fosse adotado. Inicialmente variamos todos os parâmetros simultaneamente, no entanto, como as variáveis $k$ e $h$ não são independentes, a

minimização não apresentava resultados consistentes. No procedimento final tomamos o cuidado de variar esses parâmetros em minimizações distintas.

**Primeira etapa de minimização (MR1)** Para a primeira rodada, os valores iniciais dos parâmetros Morse $D$, $\lambda$ e constantes elásticas $k$, foram randomizados em até 20% em relação aos valores obtidos anteriormente para o modelo PB original [45]. O passo de hélice foi mantido fixo em $h = 3.4$ Å. Para cada uma das concentrações salinas esse procedimento foi realizado 150 vezes. A tabela 3.1 mostra os valores de $\Delta T$ e $\chi^2$ após esta etapa. Os resultados dessa etapa essencialmente não divergiram dos valores do parâmetros do modelo PB [45]. Isso indica que o valor de $h = 3.4$ Å no modelo MEB em conjunto com os valores do modelo PB representa um resultado estável, um mínimo local.

**Segunda etapa de minimização (MR2)** Na segunda etapa, os parâmetros Morse $D$, $\lambda$ e constantes elásticas $k$ foram mantidos constantes, atribuiu-se a esses o valor médio das rodadas da etapa anterior. O passo de hélice, por sua vez, foi variado em até 20% ao redor do valor $h = 3.4$ Å anterior. Foram realizadas 300 rodadas nessa etapa. A tabela 3.1 mostra os valores de $\Delta T$ e $\chi^2$ após esta etapa.

**Terceira etapa de minimização (MR3)** Para estimar o erro associado a cada parâmetro, nessa etapa em vez de variar os parâmetros iniciais, as temperaturas de desnaturação foram variadas através da adição de uma quantia aleatória $\pm\delta T_i$ de tal maneira que as temperaturas modificadas estejam na faixa do erro experimental. A incerteza experimental relatada por Owczarzy et al. [60] para as temperaturas de desnaturação é de 0.3 °C. Como buscamos investigar o erro associado a cada parâmetro, nessa etapa todos os parâmetros foram variados: Morse $D$, $\lambda$, constantes elásticas $k$ e passo de hélice $h$. Um total de 300 rodadas foram realizadas nessa etapa. É interessante destacar que realizamos uma etapa de minimização de teste variando apenas os parâmetros $h$ para estudar se a relação com o parâmetro de empilhamento estava afetando o valor final desses parâmetros nessa otimização variando os dados experimentais. Observamos que praticamente não houve alteração nos valores médios dos parâmetros, validando essa etapa MR3.

### 3.3.2 Resultados e discussão

Analisando o potencial de empilhamento da equação (3.14), observamos que há uma dependência entre as duas constantes $k$ e $h$, o que impede que sejam minimizadas

| | $\Delta T$ (°C) | | | $\chi^2$ (°C$^2$) | | |
|---|---|---|---|---|---|---|
| [$Na^+$] (mM) | MR1 | MR2 | MR3 | MR1 | MR2 | MR3 |
| 69 | 0.841603 | 0.836828 | 0.816806 | 123.208 | 121.959 | 117.309 |
| 119 | 0.848872 | 0.845658 | 0.834405 | 119.297 | 118.687 | 115.939 |
| 220 | 0.792025 | 0.795403 | 0.795574 | 108.633 | 106.981 | 111.275 |
| 621 | 0.889651 | 0.890475 | 0.833005 | 121.591 | 122.835 | 109.627 |
| 1020 | 0.883649 | 0.867236 | 0.851968 | 127.057 | 123.776 | 121.451 |

**Tabela 3.1**
Evolução de $\Delta T$ e $\chi^2$ ao longo das três etapas de minimização.

simultaneamente. Por isto a otimização foi realizada em três etapas. Na primeira, MR1, mantivemos o *rise* fixo em 3.4 Å [105] e minimizamos os demais parâmetros $D$ $\lambda$ e $k$. Em seguida, na MR2, mantivemos os parâmetro de Morse ($D$ e $\lambda$) e de empilhamento ($k$) constantes nos valores resultantes de MR1, e variamos apenas os parâmetros $h$. Na última otimização, MR3, variando todos os parâmetros fazemos a estimativa da influência do erro experimental sobre os parâmetros otimizados. No geral há uma discreta redução do valor de $\chi^2$ durante as três minimizações como mostrado na tabela 3.1.

Na figura 3.4 mostramos os valores resultantes de $D$ e $\lambda$ que caraterizam o potencial de ligação de hidrogênio. Também mostramos nesta figura os parâmetros tais como obtidos pelo modelo original PB como curvas cinzas. Praticamente não se observa diferenças entre os dois modelos para estes parâmetros. Já para as constantes de empilhamento $k$, observa-se algumas variações importantes como mostrado na figura 3.5. Em especial para a situação em que ambos os próximos vizinhos tem forte ligação de hidrogênio, figura 3.5a, observamos um aumento para vizinhos tipo GpC (GC seguido de GC, curva preta). Este aumento é acompanhado de uma importante diminuição no *rise* de GpC como mostrado na figura 3.6a. Para vizinhos tipo CpC há uma diminuição de $k$ (figura 3.5a, curva vermelha), no entanto o *rise* correspondente mostra pouca variação, figura 3.6a. Para vizinhos fracamente ligados, mostrados na figura 3.5c, observamos pouca alteração em relação ao modelo não-modificado PB, no entanto para ApT (AT seguido e TA) há um forte aumento de *rise* como mostrado na figura 3.6c (curva preta) que chega a um valor médio próximo de 6 Å.

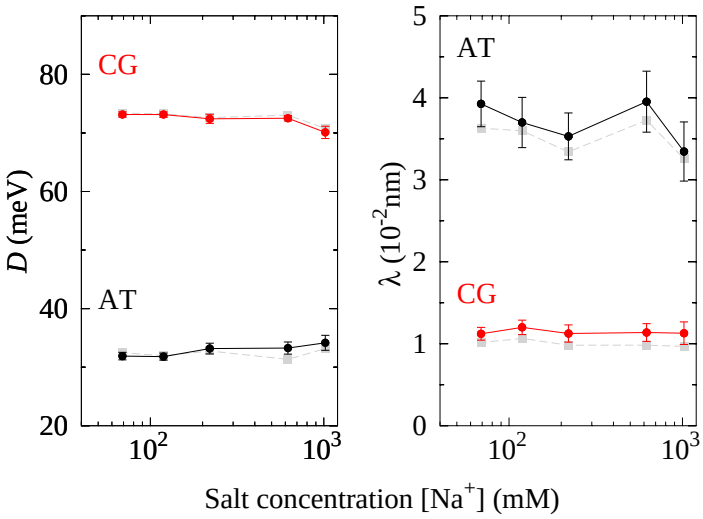



**Figura 3.4**
Dependência das constantes associadas ao potencial de Morse $D$ e $\lambda$ com a concentração salina. Linhas tracejadas em cinza são os parâmetros correspondentes para o modelo não-modificado da referência 45.

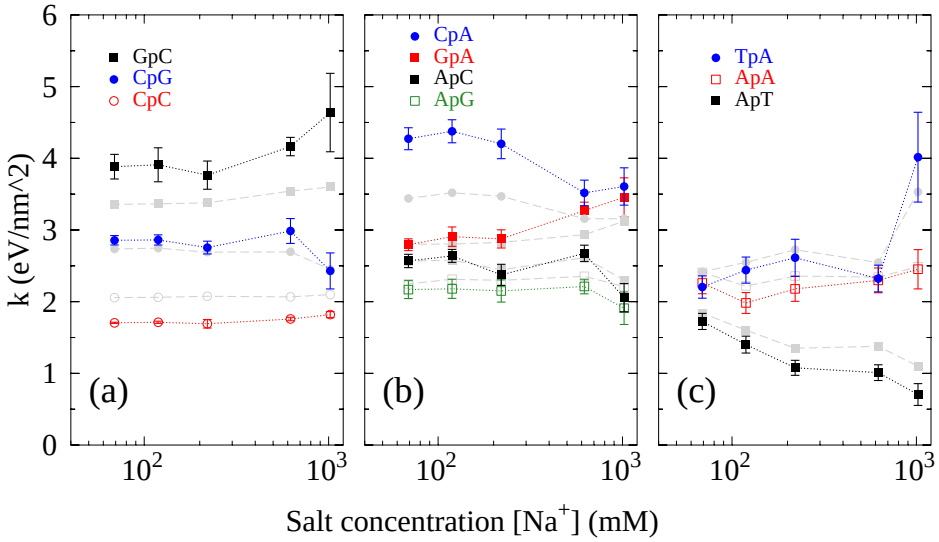



**Figura 3.5**
Dependência da constante de empilhamento do MEB com a concentração salina. Os painéis são ordenados por número de ligações de hidrogênio por par de próximos vizinhos, (a) quatro, (b) cinco e (c) seis ligações. Linhas tracejadas em cinza são os parâmetros correspondentes para o modelo não-modificado da referência 45.

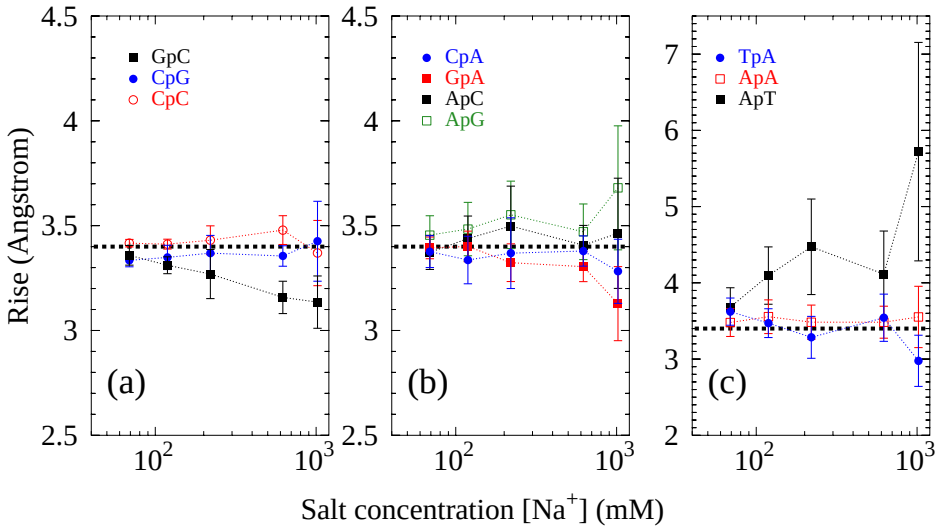



**Figura 3.6**
Dependência do passo de hélice (*rise*) com a concentração salina. Os símbolos e cores são iguais ao da figura 3.5. A linha pontilhada indica a posição de 3.4 Å comumente adotada como valor médio de *rise* [105]. A escala do painel (c) está ampliada em relação aos painéis (a) e (b).

## 3.4 Comparação dos parâmetros estruturais otimizados com medidas experimentais de raio-X e NMR

Os resultados da seção anterior mostram importantes variações do parâmetro de *rise* com o tipo de próximo vizinhos e também com a concentração salina, como visto na figura 3.6. De fato, apenas em poucas situação este parâmetro ficou em 3.4 Å que é o valor usualmente adotado a partir do trabalho pioneiro de Yanagi et al. [105]. Mas se levarmos em consideração medidas mais recentes e detalhadas, de quanto realmente varia este parâmetro? Existe alguma dependência conhecida com o tipo de próximo vizinho?

Grande parte do avanço na investigação da estrutura dos ácidos nucleicos se deve ao refinamento da técnica de cristalografia de raio-X. Uma simples consulta a base de dados NDB (*Nucleic Acids Database*) mostra que a cristalografia ainda é o método dominante, das 5435 estruturas disponíveis para DNA 4639 foram obtidas através de raio-X e as 796 restantes através de NMR. Como com o nosso modelo MEB incluiu a variável estrutural do passo da hélice e foi possível fazer uma estimativa independente para esse parâmetro, antes de comparar os dados obtidos com os valore experimentais é interessante apresentar de maneira breve esses métodos estruturais.

Uma das grandes vantagens do NMR é a determinação das estruturas em solução, estado em que as moléculas de DNA se encontram no meio fisiológico. Por não existir necessidade de cristalizar a molécula, um dos fatores limitantes da cristalografia de raio-X, os resultados de NMR permitem investigar a flexibilidade de DNA e RNA. Em vez de contrastantes esses métodos devem ser complementares, ambos possibilitam, por exemplo, o estudo da dependência da conformação estrutural em relação à sequência. No entanto, algumas vezes há divergência nos resultados, conforme discutido no caso do par GU.

O primeiro desafio para realizar a comparação dos valores de passo de hélice obtidos com os dados disponíveis consiste na inviabilidade de retirar essa informação diretamente de experimentos de raio-X e NMR. É necessária uma interpretação computacional desses dados experimentais para que informações estruturais como o passo de hélice sejam extraídas.

Há uma série de abordagens construídas para analisar e comparar estruturas tridimensionais de ácidos nucleicos (CEHS [12],CompDNA [13] ,Curves [14], FREEHELIX [15], NGEOM [16], NUPARM [17] e RNA [106]. No entanto, os programas de computador desenvolvidos para reconstruir a conformação da cadeia geram resultados inconsistentes [18]. A discrepância se revela na descrição dos parâmetros da estrutura, que podem ser classificados em dois grupos: parâmetros relacionados ao par de base (*shear*, *stretch*, *stagger*, *buckle*, *propeller twist* e *opening*) e parâmetros associados ao dímero formado por dois pares vizinhos, chamados parâmetros de *step*. (*shift*, *slide*, *rise*, *tilt*, *roll*, *twist*) como mostrado na figura 3.7.

Lu and Olson, ao comparar a implementação de sete algoritmos usados para determinação desses valores, identificaram que tal divergência se devia principalmente à adoção de sistemas de referência diferentes. Ao escolher o mesmo sistema de referência para todos as sete abordagens computacionais foi observada uma descrição da estrutura similar entre elas. Segundo Lu and Olson o posicionamento dos sistemas de referência nas fronteiras interiores e exteriores das bases complementares exagera o passo de hélice para os dímeros distorcidos. As divergências foram ilustradas a partir da comparação desses parâmetros usando dados de estrutura cristalinas para algumas sequências e a diferença observada foi maior para *rise* associado ao *step* ApT. Existe também um conflito de definição de parâmetros, o passo de hélice (*rise*) em particular é definido como o como o componente $z$ do vetor translacional que conecta as origens de pares de base sequenciais para todos os programas com exceção do *Curves*. Todos programas definem *shift* e *slide* consistentemente como as componentes $x$ e $y$ desse vetor. Portanto, após adotar o mesmo sistema de referência apenas os resultados para *rise* associados ao *Curves* devido a definição não usual adotada diverge dos demais.

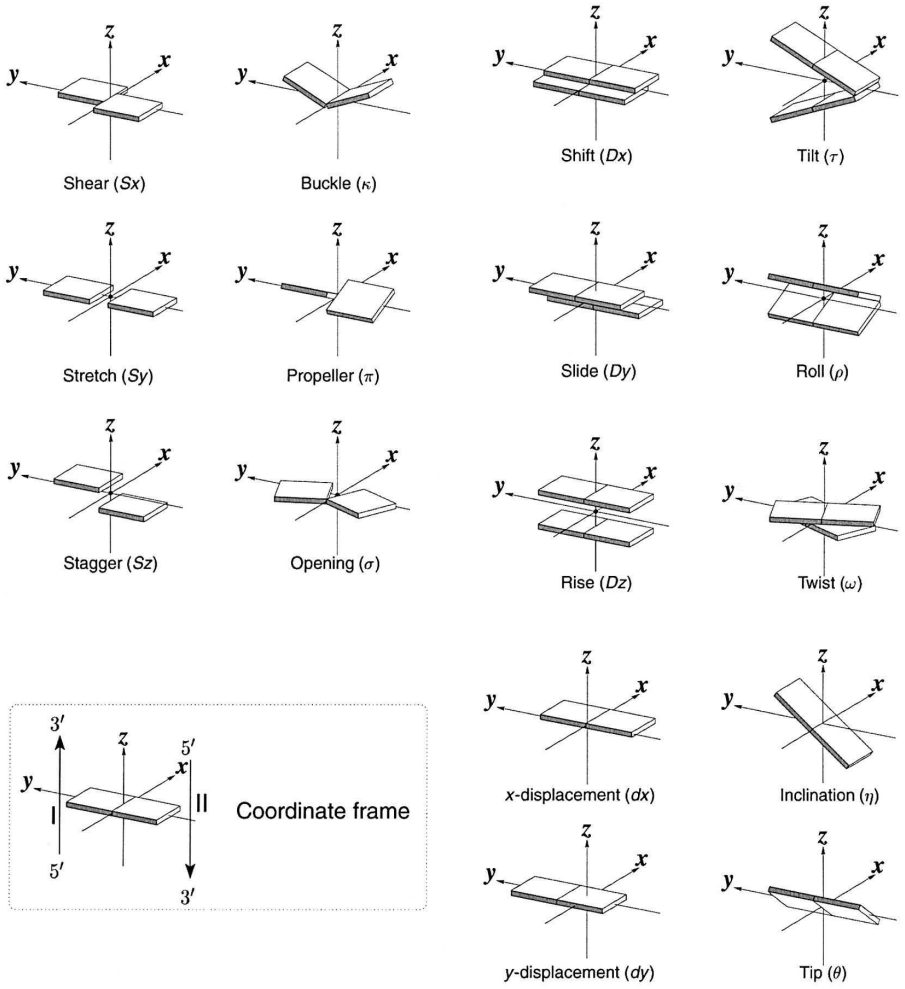



**Figura 3.7**
Sistema de referência e parâmetros estruturais de par de base e entre pares vizinhos (*step*). Figura retirada de Lu and Olson [107].

Diante dessa análise mostrou-se necessário um sistema de referência padrão para descrever o arranjo tridimensional das bases e dos pares de base das estruturas de ácidos nucleicos. Do contrário, tendo em vista as discrepâncias relatadas, a comparação dessas estruturas se torna impraticável. A fim de solucionar essa questão, um grupo de pesquisadores envolvidos com o desenvolvimento e utilização desses programas propôs um sistema de referência padrão [10] cujas coordenadas das bases foram baseadas em estruturas cristalinas de alta resolução. A escolha do sistema de coordenadas estabelece que bases complementares formem um par de base Watson-Crick planar no estado não distorcido com distâncias das ligações de hidrogênio e dos ângulos virtuais obedecem os valores medidos nas estruturas cristalinas de pequenas moléculas. O sistema de referência proposto teve êxito em proporcionar análises conformacionais da estrutura

helicoidal independentes do programa computacional utilizado [10].

Em resumo, portanto, observou-se que o valor do passo da hélice depende do pacote de software particular utilizado para interpretar os dados de difração de raios-x e NMR e seus resultados diferem significativamente uns dos outros. Conforme visto, tais diferenças se devem principalmente às diferentes formas de posicionar a estrutura de referência na estrutura helicoidal do DNA. Essa discrepância pode ser observada para a sequência $d(CGATCGATCG)$ no trabalho inicial realizado com o modelo estrutural bidimensional [108], que exibe a comparação do resultados de passo de hélice obtidos a partir da nossa primeira tentativa de minimização com os valores calculados por diferentes softwares CEHS, NEWHELIX e 3DNA. Tal programa 3DNA [107, 109] foi desenvolvido pelo grupo de pesquisadores responsável pela comparação dos diversos programas e elaboração do sistema de referência padrão [10]. O 3DNA faz uso desse sistema de referência recomendado para a descrição da geometria do par de base e utiliza um esquema rigoroso para calcular os parâmetros estruturais e reconstruir essa estrutura. Em uma das rotinas desse programa os ácidos nucleicos são representados por blocos retangulares conforme exibido na figura 3.7. Para realizar a comparação com nossos resultados optamos por utilizar os dados de parâmetros estruturais das sequências disponíveis na base de dados Nucleic Acids Database (NDB) [110, 111]. Esses parâmetros estruturais contidos na base NDB foram calculados através do pacote 3DNA [107, 109].

### 3.4.1 Extração dos dados experimentais

A fim de comparar os resultados previstos pelo nosso modelo com medidas experimentais do passo da hélice, extraímos dados obtidos através de medidas de raio-X e e de espectroscopia NMR da base de dados *Nucleic Acid Database* (NDB) [110, 111]. Como essa base reúne dados que envolvem diferentes conformações de DNA e pareamentos diversos, é necessário filtrar os dados adequados para possibilitar a comparação. Embora exista uma ferramenta de busca de estruturas no sistema via interface *web*, a seleção é deficiente e não identifica corretamente as estruturas de interesse. Desenvolvemos um programa em *Perl* para selecionar apenas as sequências B-DNA e pares Watson-Crick. A partir dos códigos de identificação (ID NDB) das estruturas fornecidos após utilizar o filtro do sistema, esse programa conseguiu baixar os parâmetros de par de base e de *step*. Em seguida, através de uma análise da sequência, selecionamos apenas as estruturas B-DNA canônicas. Foram selecionadas medidas de raio-X para 290 sequências e medidas de NMR associadas a 163 sequências sequências de B-DNA canônicos. Separadamente para cada tipo de método experimental, calculamos desse conjunto de

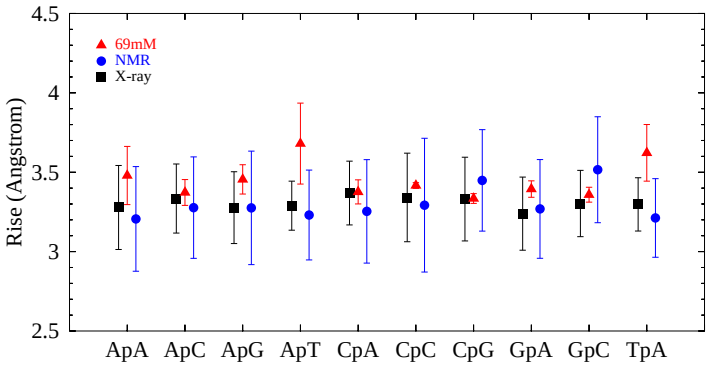



**Figura 3.8**
Comparação entre os valores do parâmetro passo de hélice (*rise*) calculados na concentração salina de 69 mM (vermelho) com a média dos valores do passo de hélice medidos em diversas sequências através de raios-X e analisados com 3DNA (preto) e NMR (azul). Os resultados foram levemente deslocados na horizontal para facilitar a visualização.

sequências a média e o desvio-padrão para cada tipo de passo (apenas relembrando a notação de step: ApC se refere ao passo entre o par AT e o par CG).

### 3.4.2 Discussão

Comparamos os valores obtidos para cada passo de hélice através do parâmetro $h$ do nosso modelo com as médias calculadas das medidas de raio-x e NMR das sequências selecionadas pelo método descrito na seção anterior. Nossos resultados concordam em grande parte com os resultados do 3DNA-NDB dentro da incerteza dada pelos desvios-padrão, figuras 3.8 e 3.9. O valor previsto com MEB que mais difere da média dos dados de raio-x está associado ao passo ApT, mas esse também é o que apresenta a maior incerteza do modelo. É importante enfatizar que a boa concordância não era óbvia e que o modelo foi capaz de estimar parâmetros estruturais do DNA partindo apenas de dados experimentais de temperatura de desnaturação.

Ao contrário das medidas de difração de raios-x, temperaturas de desnaturação de DNA podem ser medidas ao longo de uma gama de concentrações salinas diferentes. Dessa forma, o nosso método abre a possibilidade de analisar a mudança de parâmetros estruturais como uma função da concentração salina. É necessário ressaltar que cada minimização foi feita separadamente para cada concentração salina, ou seja cada otimização é baseada em um conjunto de dados independente de temperatura de desnaturação. Portanto, o comportamento consistente de $h$ observado para cada grupo de próximos vizinhos, em diferentes concentrações, mostra que o modelo produz resultados robustos. Em particular, verificou-se que o passo de hélice varia mais fortemente para pares de bases fracamente ligados AT consecutivos e com a concentração salina mais elevada figura 3.6. Essa observação nos levou a questionar se o valor de *rise* resultante do modelo corresponde à estrutura de dupla hélice estável ou se de certa forma captura o processo de desnaturação. Isto é, dado que AT é o par de base menos estável, a região

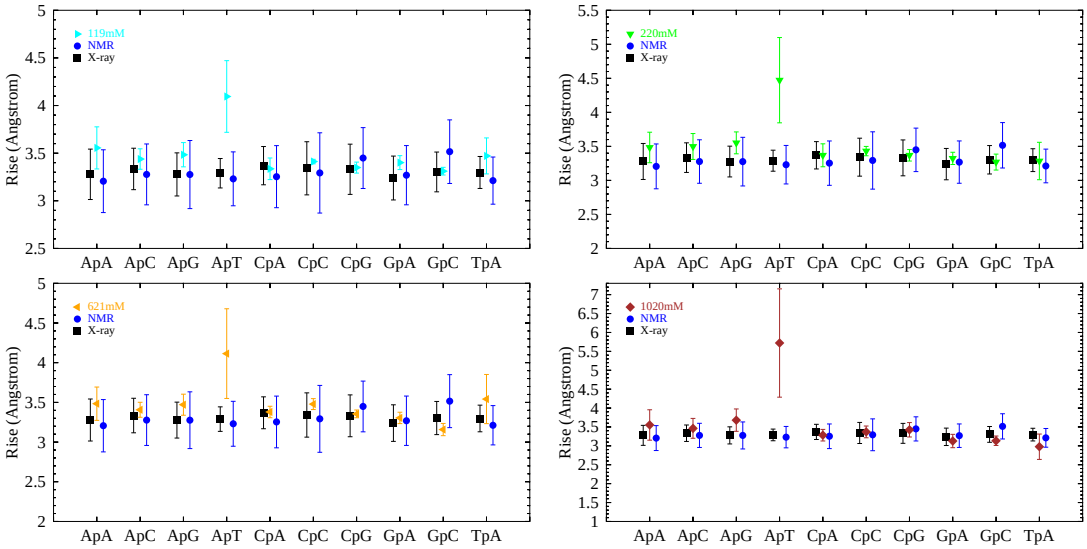



**Figura 3.9**
Comparação entre os valores do parâmetro passo de hélice (*rise*) calculados na concentração salina de 119 mM (ciano), 220 mM (verde), 621 mM (laranja) e 1020 mM (marrom) com a média dos valores do passo de hélice medidos em diversas sequências através de raios-X e analisados com 3DNA (preto) e NMR (azul). Os resultados foram levemente deslocados na horizontal para facilitar a visualização.

que possui dois pares AT consecutivos é candidata a se romper primeiro no processo de desnaturação. Uma vez que esse par se abre, a distância efetiva em relação ao par de base vizinho é acentuada. Em Perkins [97], o valor relatado para o *rise* da dupla hélice é de 3,4 Å enquanto para a fita simples é de 5,4 Å. Ao comparar com os valores obtidos para $h$ na figura 3.6, vemos que ApT, assume valores intermediários.

É interessante fazer um paralelo com os resultados obtidos para os parâmetros associados aos pares de base terminal [112]. O parâmetro de Morse obtido para AT foi o que mais variou com concentração salina. A hipótese que levantamos foi a de ligações de hidrogênio de menor intensidade nas concentrações salinas mais baixas. No entanto, também seria razoável assumir que essas extremidades estariam iniciando um processo de desnaturação. É curioso, todavia, que no caso do *rise* os valores mais altos obtidos estejam associados às concentrações salinas maiores já que a presença do sal deveria contribuir para a estabilidade da hélice.

## 3.5 Conclusão

A partir da adaptação de um modelo mesoscópico tridimensional, foi possível incluir o parâmetro estrutural do passo de hélice anteriormente omitido na aproximação do

modelo PB original. Ao combinar o modelo com a técnica de equivalência termodinâmica, otimizamos os valores do passo de hélice usando temperaturas de desnaturação. Como havia disponibilidade de conjuntos de dados de desnaturação para sequências em soluções de diferentes concentrações salinas, pudemos obter independentemente os parâmetros de *rise* para cada uma dessas concentrações. O comportamento semelhante do passo de hélice observado comparando essas diferentes otimizações confere confiabilidade aos parâmetros. Para realizar a comparação com dados experimentais, com um programa que desenvolvemos, selecionamos da base NDB dados *rise* de B-DNA canônicos interpretados de medidas de raio-X e NMR. O conjunto de próximos-vizinhos ApT exibiu a maior dependência com a concentração salina, apresentando valores mais elevados de *rise* para concentrações maiores. ApT também foi o que mais divergiu do valor médio dos dados de NMR e raio-X e do valor de 3.4 Å.

# Capítulo 4

# Conclusão

Os trabalhos realizados evidenciam a versatilidade de modelos do tipo Peyrard-Bishop. O modelo PB se destaca em relação aos outros modelos simplificados, em particular o modelo de próximos vizinhos, pela capacidade de distinguir o tipo de interação intramolecular: entre as bases de um par (ligação de hidrogênio) ou entre os pares de base (empilhamento). Embora seu nível de detalhamento não proporcione a descrição da estrutura atômica dos ácidos nucleicos, exatamente por isso o tempo computacional requerido permite estudar moléculas com comprimento de interesse por períodos longos, assim como uma variedade de sequências de diferentes composições. O uso dessa classe de modelos mesoscópicos associada à técnica de equivalência termodinâmica possibilitou a abordagem de propriedades físicas de DNA e RNA a partir de dados experimentais de temperatura de desnaturação. O foco da pesquisa foi o estudo da estabilidade do par guanina-uracila em RNA e a proposta de um modelo modificado que inclui o parâmetro estrutural *rise* em DNA. Adaptamos o modelo PB original em uma tentativa de explorar a influência do contexto no *mismatch* guanina-uracila sugerida experimentalmente por medidas de desnaturação, de raio-X e NMR. Para absorver as diferentes conformações que os pares GU podem assumir, introduzimos a dependência de contexto da sequência nos parâmetros desse modelo. Nosso objetivo era esclarecer a divergência dos dados experimentais em relação ao número de ligações de hidrogênio desse par em certas configurações, bem como fornecer os parâmetros associado a esse par. Seria um modelo mesoscópico capaz de distinguir as nuances da estabilidade do par GU? Nossos resultados indicam que sim, além de fornecerem previsões para contextos diversos. Fomos capazes de prever a diferença do padrão da ligação de hidrogênio do GU de maneira consistente com os dados experimentais disponíveis para algumas determinadas configurações. Em particular, a previsão de uma única ligação de hidrogênio para configurações tandem GUpUG concorda com as medições de NMR.

Além disso, os nossos resultados sugerem que a estabilidade dos pares terminais GU relatada experimentalmente pode ser atribuída às interações entre os pares de ligações de hidrogênio. Ao parametrizar o *mismatch* GU, presente em todos os tipos de RNA, disponibilizamos uma ferramenta independente para a análise da estabilidade de sequências quaisquer contendo esse par não canônico. Essa funcionalidade tem um potencial interesse para o desenvolvimento de sondas. Concomitantemente, ao adaptar o software para lidar com GU, expandimos a possibilidade de investigar outros pares cujas ligações de hidrogênio são dependentes de contexto.

Para reduzir ao mínimo o número de variáveis, o modelo PB original assumiu uma aproximação que elimina a existência do parâmetro associado ao passo da hélice no hamiltoniano. Ao planificar o modelo de Barbi tridimensional analiticamente, recuperamos a dependência com esse parâmetro estrutural. Essa distância corresponde ao passo de motores moleculares ao se moverem pelo DNA, por exemplo as enzimas RNA polimerase e DNA polimerase. De posse desse novo hamiltoniano que inclui o passo da hélice, decidimos usar a metodologia de parametrização a partir de temperaturas de desnaturação para obter esse parâmetro. Como o conjunto de dados extraídos da literatura contém temperaturas para diferentes concentrações salinas, foi possível analisar a dependência do passo de hélice com a concentração salina. Adicionalmente comparamos os dados obtidos com os dados experimentais (raio-X e NMR) da média do passo de hélice para cada configuração de próximos vizinhos. Os resultados são consistentes com os valores experimentais com exceção do passo ApT. Além do valor do passo de hélice para ApT divergir do valor médio dos dados de NMR e raio-X e do valor de 3.4 Å, ele exibiu a maior dependência com a concentração salina, apresentando valores mais elevados de *rise* para concentrações maiores.

## 4.1 Outros trabalhos realizados no âmbito deste projeto

**Estabilidade das terminações das hélices em RNA e DNA** A investigação do *mismatch* GU com o modelo Peyrard-Bishop, ao considerar a dependência de contexto, permitiu pela primeira vez diferenciar a estabilidade de um par interno com os pares situados nas terminações com modelos mesoscópicos. Conforme discutido na seção 2.3.1.2, página 43, os valores obtidos para os parâmetros de Morse associados aos pares terminais foram superiores até ao par canônico AU para a maioria dos contextos. Essa característica indica que a existência do par GU terminal confere uma estabilização maior à hélice e sugere que essa estabilidade deve ser atribuída às ligações de

hidrogênio. Todavia, como não havia um estudo da estabilidade dos pares terminais com modelos mesoscópicos surgiu o questionamento se os valores observados realmente traduziam o comportamento do GU ou se isso era um efeito associado ao próprio modelo adotado. Com o intuito de estudar se a estabilidade das ligações de hidrogênio nas extremidades é um efeito real ou está associado a um artefato do modelo, propusemos e realizamos o estudo dos pares terminais em DNA e RNA canônicos.

Empregamos uma adaptação semelhante à utilizada no estudo do GU, do capítulo 2 para estudar as terminações dos oligonucleotídeos. Apenas tratamos as bases nas pontas como se fossem bases diferentes para possibilitar a distinção dos pares internos em relação aos terminais. Contudo, como por enquanto não há interesse em estudar o contexto, não foi necessária a separação em trímeros. Para identificar se o aumento de estabilidade era um artefato do modelo, investigamos separadamente DNA e RNA usando, respectivamente, as bases de temperatura de desnaturação retiradas de [60, 65]. Os resultados referentes ao estudo dos terminais em DNA foram publicados em Ferreira et al. [112] e os resultados para RNA estão sendo preparados para a publicação.

**Projeto em colaboração com grupo da Medicina/Endocrinologia da UFRGS**
Nesta colaboração com o grupo de endocrinologia do Hospital de Clínicas da UFRGS, nós auxiliamos no estudo de estabilidade de RNA mensageiro (mRNA) relativo ao de polimorfismos de nucleotídeo único (SNPs) no gene RET e a susceptibilidade ou progressão do carcinoma medular de tireóide (CMT) hereditária ou esporádica. Nossa primeira tentativa para estudar a estabilidade foi analisar os perfis de abertura das variantes e comparar com *wild type*. A ideia inicial era comparar a abertura média de uma sequência com as demais. Nas sequências de RNA de interesse, o par GU ocorre com frequência, no entanto, só tínhamos os parâmetros necessários — Morse e constantes de empilhamento — para analisar a abertura com o software do nosso grupo TfReg para os pares canônicos (obtidos em [46]). A necessidade de parametrização do *mismatch* GU motivou o estudo de estabilidade desse par descrito no capítulo 2. Contudo, a presença de estruturas diferentes de hélice dupla como *loops* e *hairpins* inviabilizou a análise da forma pretendida.

A abordagem que permitiu que analisássemos a estabilidade das variantes consistiu em considerar as estruturas subótimas geradas pelo software RNAfold. Estruturas sub-ótimas são definidas como aquelas que tem uma energia que difere dentro de uma faixa determinada da energia da estrutura ótima. Utilizamos a sequência da variante *wild type* e de cada uma das variantes polimórficas para gerar um conjunto de estruturas sub-ótimas. A sequência do mRNA associado ao RET possui aproximadamente

5000 pares de bases de comprimento. Dessa forma, cada um dos conjuntos tem muitos resultados possíveis para a estrutura mais estável, com valores muito próximos de energia livre de Gibbs. Seria impraticável, em termos computacionais, gerar todas as estruturas possíveis de baixa energia em uma dada faixa de energia. Em vez disso, 380 estruturas foram geradas aleatoriamente para cada tipo de variante polimórfica, com probabilidades iguais ao peso Boltzman de sua energia. Foram calculadas as energias livres de cada estrutura e o número de pares em dupla-hélice com o pacote Vienna RNA [113]. As estruturas sub-ótimas associadas à variante mais agressiva na média apresentaram maior número de regiões de fita dupla e menor energia livre, indicando maior estabilidade. Este trabalho foi publicado em Ceolin et al. [114].

# Referências


[1] Tobias Mann, Richard Humbert, Michael Dorschner, John Stamatoyannopoulos, and William Stafford Noble. A thermodynamic approach to PCR primer design. *Nucleic Acids Research*, 37(13):e95–e95, 2009.

[2] N. Haslam, N. Whiteford, Gerald Weber, Adam Prügel-Bennett, Jonathan W. Essex, and C. Neylon. Optimal probe length varies for targets with high sequence variation: Implications for probe library design for resequencing highly variable genes. *PLoS ONE*, 3:e2500, 2008. doi: http://dx.doi.org/10.1371/journal.pone.0002500.

[3] Jef Hooyberghs, Marco Baiesi, Alessandro Ferrantini, and Enrico Carlon. Breakdown of thermodynamic equilibrium for DNA hybridization in microarrays. *Physical Review E*, 81(1):012901, 2010.

[4] Xiaoting Qian, Dan Pu, Bicheng Liu, and Pengfeng Xiao. Effect of oligonucleotide probes substituted by deoxyinosines on the specificity of SNP detection on the DNA microarray. *Electrophoresis*, 36(2):263–270, 2015.

[5] L Pray. Discovery of DNA structure and function: Watson and Crick. *Nature Education*, 1(1), 2008.

[6] Stephen Neidle. *Principles of Nucleic Acid Structure.* . Elsevier; Academic Press, 1st ed edition, 2008. ISBN 0123695074,1865843830,9780123695079. URL `http://gen.lib.rus.ec/book/index.php?md5=03D3BD1EA9D141AF1DB919A33FFBF2BB`.

[7] Eric Westhof. Isostericity and tautomerism of base pairs in nucleic acids. *FEBS letters*, 588(15):2464–2469, 2014.

[8] Sukanya Halder and Dhananjay Bhattacharyya. RNA structure and dynamics: a base pairing perspective. *Progress in biophysics and molecular biology*, 113(2): 264–283, 2013.

[9] Helge Weissig Philip E. Bourne. *Structural bioinformatics*. Methods of Biochemical Analysis, V. 44. Wiley-Liss, 1 edition, 2003. ISBN 9780471201991,0471201995. URL `http://gen.lib.rus.ec/book/index.php?md5=33624638679B7FFAE2C4D9B2443B93B0`.

[10] Wilma K. Olson, Manju Bansal, Stephen K. Burley, Richard E. Dickerson, Mark Gerstein, Stephen C. Harvey Udo Heinemann, Xiang-Jun Lu, Stephen Neidle, Zippora Shakked Heinz Sklenar, Masashi Suzuki, Chang-Shung Tung, Eric Westhof Cynthia Wolberger, and Helen M. Berman. A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, 313: 229–237, 2001.

[11] Richard E Dickerson. Definitions and nomenclature of nucleic acid structure parameters. *Journal of Biomolecular Structure and Dynamics*, 6(4):627–634, 1989.

[12] MA El Hassan and CR Calladine. The assessment of the geometry of dinucleotide steps in double-helical DNA; a new local calculation scheme. *Journal of Molecular Biology*, 251(5):648–664, 1995.

[13] Andrey A Gorin, Victor B Zhurkin, and K Wilma. B-DNA twisting correlates with base-pair morphology. *Journal of molecular biology*, 247(1):34–48, 1995.

[14] Richard Lavery and Heinz Sklenar. Defining the structure of irregular nucleic acids: conventions and principles. *Journal of Biomolecular Structure and Dynamics*, 6(4):655–667, 1989.

[15] Richard E Dickerson. DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic acids research*, 26(8):1906–1926, 1998.

[16] Chang-Shung Tung, Dikeos Mario Soumpasis, and Gerhard Hummer. An extension of the rigorous base-unit oriented description of nucleic acid structures. *Journal of Biomolecular Structure and Dynamics*, 11(6):1327–1344, 1994.

[17] Manju Bansal, Dhananjay Bhattacharyya, and B Ravi. NUPARM and NUCGEN: software for analysis and generation of sequence dependent nucleic acid structures. *Computer applications in the biosciences: CABIOS*, 11(3):281–287, 1995.

[18] Xiang-Jun Lu and Wilma K Olson. Resolving the discrepancies among nucleic acid conformational analyses. *Journal of Molecular Biology*, 285(4):1563–1575, 1999.

[19] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[20] Rosalind E Franklin and Raymond G Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171:740–741, 1953.

[21] Alexander Rich. DNA comes in many forms. *Gene*, 135(1):99–109, 1993.

[22] AH Wang, Gary J Quigley, Francis J Kolpak, James L Crawford, Jacques H Van Boom, Gijs van der Marel, and Alexander Rich. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*, 282(5740): 680–686, 1979.

[23] Kathleen P Howard. Thermodynamics of DNA Duplex Formation: A Biophysical Chemistry Laboratory Experiment. *Journal of Chemical Education*, 77(11):1469, 2000.

[24] VA Bloomfield, DM Crothers, and I Tinoco Jr. Electronic and vibrational spectroscopy. *Nucleic Acids, Structures, Properties, and Functions*, pages 165–221, 1999.

[25] Wolfram Saenger. *Principles of Nucleic Acid Structure*. Springer Advanced Texts in Chemistry. Springer New York, 1984. ISBN 978-0-387-90761-1,978-1-4612-5190-3. URL `http://gen.lib.rus.ec/book/index.php?md5=bd7cbb2d2f9c13de92880f686accfff6`.

[26] Lauren A Levine, Matthew Junker, Myranda Stark, and Dustin Greenleaf. A DNA melting exercise for a large laboratory class. *Journal of Chemical Education*, 92:1928–1931, 2015. doi: 10.1021/acs.jchemed.5b00049.

[27] Richard Owczarzy. Melting temperatures of nucleic acids: discrepancies in analysis. *Biophysical Chemistry*, 117(3):207–215, 2005.

[28] Sherrie Schreiber-Gosche and Robert A Edwards. Thermodynamics of oligonucleotide duplex melting. *Journal of Chemical Education*, 86(5):644, 2009.

[29] John SantaLucia, Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA*, 95(4):1460–1465, 1998. URL `http://www.pnas.org/cgi/content/abstract/95/4/1460`.

[30] Grégoire Altan-Bonnet, Albert Libchaber, and Oleg Krichevsky. Bubble dynamics in double-stranded DNA. *Physical Review Letters*, 90(13):138101, 2003.

[31] A Montrichok, G Gruner, and G Zocchi. Trapping intermediates in the melting transition of DNA oligomers. *EPL (Europhysics Letters)*, 62(3):452, 2003.

[32] Yan Zeng, Awrasa Montrichok, and Giovanni Zocchi. Bubble nucleation and cooperativity in DNA melting. *Journal of molecular biology*, 339(1):67–75, 2004.

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

[34] John SantaLucia, Jr., H T Allawi, and P A Seneviratne. Improved nearest-neighbour parameters for predicting DNA duplex stability. *Biochem.*, 35:3555–3562, 1996.

[35] Jr. SantaLucia, John and Donald Hicks. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, 33:415–440, 2004.

[36] Ilyas Yildirim and Douglas H Turner. RNA challenges for computational chemists. *Biochem.*, 44(40):13225–13234, 2005.

[37] Douglas Poland and Harold A Scheraga. Kinetics of the helix—coil transition in polyamino acids. *The Journal of Chemical Physics*, 45(6):2071–2090, 1966.

[38] Marco Baiesi and Enrico Carlon. Models of DNA denaturation dynamics: universal properties. *arXiv preprint arXiv:1402.6492*, 2014.

[39] M. Peyrard, S. Cuesta-López, and D. Angelov. Experimental and theoretical studies of sequence effects on the fluctuation and melting of short DNA molecules. *J. Phys.: Condens. Matter*, 21:034103, 2009.

[40] Roumen A. Dimitrov and Michael Zuker. Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys. J.*, 87:215–226, 2004.

[41] Daniel Jost and Ralf Everaers. Genome wide application of DNA melting analysis. *Journal of Physics: Condensed Matter*, 21(3):034108, 2009.

[42] Rodrigo Galindo-Murillo, Daniel R Roe, and Thomas E Cheatham III. On the absence of intrahelical DNA dynamics on the $\mu$s to ms timescale. *Nature Communications*, 5, 2014.

[43] M. Peyrard and A. R. Bishop. Statistical mechanics of a nonlinear model for DNA denaturation. *Phys. Rev. Lett.*, 62(23):2755–2757, 1989.

[44] Gerald Weber, Niall Haslam, Nava Whiteford, Adam Prügel-Bennett, Jonathan W. Essex, and Cameron Neylon. Thermal equivalence of DNA duplexes without melting temperature calculation. *Nat. Phys.*, 2:55–59, 2006. doi: 10.1038/nphys189.

[45] Gerald Weber, Jonathan W. Essex, and Cameron Neylon. Probing the microscopic flexibility of DNA from melting temperatures. *Nat. Phys.*, 5:769–773, 2009. doi: 10.1038/nphys1371.

[46] Gerald Weber. Mesoscopic model parametrization of hydrogen bonds and stacking interactions of RNA from melting temperatures. *Nucleic Acids Res.*, 41: e30, 2013. doi: 10.1093/nar/gks964. URL `http://nar.oxfordjournals.org/content/41/1/e30`.

[47] Yong-Li Zhang, Wei-Mou Zheng, Ji-Xing Liu, and Y. Z. Chen. Theory of DNA melting based on the Peyrard-Bishop model. *Phys. Rev. E*, 56(6):7100–7115, 1997.

[48] Carla Goldman and Wilma K Olson. DNA denaturation as a problem of translational-symmetry restoration. *Physical Review E*, 48(2):1461, 1993.

[49] Michel Peyrard, Santiago Cuesta-López, and Guillaume James. Modelling DNA at the mesoscale: a challenge for nonlinear science? *Nonlinearity*, 21(6):T91, 2008.

[50] T. Dauxois, M. Peyrard, and A. R. Bishop. Entropy-driven DNA denaturation. *Phys. Rev. E*, 47(1):R44–R47, 1993.

[51] T. Dauxois and M. Peyrard. Entropy-driven transition in a one-dimensional system. *Phys. Rev. E*, 51(5):4027–4040, 1995.

[52] AE Bergues-Pupo, JM Bergues, and F Falo. Unzipping of DNA under the influence of external fields. *Physica A: Statistical Mechanics and its Applications*, 396:99–107, 2014.

[53] Ana Elisa Bergues-Pupo, Fernando Falo, and Alessandro Fiasconaro. Resonant optimization in the mechanical unzipping of DNA. *EPL*, 105(6):68005, 2014.

[54] MS Rocha, AD Lúcio, SS Alexandre, RW Nunes, and ON Mesquita. DNA-psoralen: single-molecule experiments and first principles calculations. *Applied Physics Letters*, 95(25):253703, 2009.

[55] PS Alves, ON Mesquita, and Marcio S Rocha. Controlling Cooperativity in $\beta$-Cyclodextrin–DNA Binding Reactions. *The journal of physical chemistry letters*, 6(18):3549–3554, 2015.

[56] Amar Singh and Navin Singh. Effect of salt concentration on the stability of heterogeneous DNA. *Phys. A (Amsterdam, Neth.)*, 419:328–334, 2015.

[57] Amar Singh and Navin Singh. Pulling DNA: The Effect of Chain Length on the Mechanical Stability of DNA Chain. In *Macromolecular Symposia*, volume 357, pages 64–69. Wiley Online Library, 2015.

[58] Marc Joyeux and Sahin Buyukdagli. Dynamical model based on finite stacking enthalpies for homogeneous and inhomogeneous DNA thermal denaturation. *Phys. Rev. E*, 72:051902, 2005.

[59] Z Rapti. Stationary solutions for a modified peyrard-bishop dna model with up to third-neighbor interactions. *The European Physical Journal E*, 32(2):209–216, 2010.

[60] Richard Owczarzy, Yong You, Bernardo G. Moreira, Jeffrey A. Manthey, Lingyan Huang, Mark A. Behlke, and Joseph A. Walder. Effects of sodium ions on DNA duplex oligomers: Improved predictions of melting temperatures. *Biochem.*, 43: 3537–3554, 2004.

[61] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, 1988.

[62] F. H. C. Crick. Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, 19(2):548–555, 1966.

[63] Gabriele Varani and William H McClain. The G·U wobble base pair. *EMBO Rep.*, 1(1):18–23, 2000.

[64] Masahiro Naganuma, Shun-ichi Sekine, Yeeting Esther Chong, Min Guo, Xiang-Lei Yang, Howard Gamper, Ya-Ming Hou, Paul Schimmel, and Shigeyuki Yokoyama. The selective tRNA aminoacylation mechanism based on a single G·U pair. *Nature*, 510(7506):507–511, 2014.

[65] Tianbing Xia, John SantaLucia, Jr., Mark E. Burkard, Ryszard Kierzek, Susan J. Schroeder, Xiaoqi Jiao, Christopher Cox, and Douglas H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochem.*, 37:14719–14735, 1998.

[66] I. Vakonakis and A. C. LiWang. N1· · ·N3 hydrogen bonds of A:U base pairs of RNA are stronger than those of A:T base pairs of DNA. *J. Am. Chem. Soc.*, 126 (18):5688–5689, 2004.

[67] Tauanne D. Amarante and Gerald Weber. Evaluating hydrogen bonds and base stackings of single, tandem and terminal GU in RNA mismatches with a mesoscopic model. *Journal of Chemical Information and Modeling*, 56(1):101–109, 2016. doi: 10.1021/acs.jcim.5b00571. URL `http://dx.doi.org/10.1021/acs.jcim.5b00571`.

[68] Neocles B Leontis and Eric Westhof. Geometric nomenclature and classification of RNA base pairs. *Rna*, 7(4):499–512, 2001.

[69] Benoît Masquida and Eric Westhof. On the wobble GoU and related pairs. *RNA*, 6(01):9–15, 2000.

[70] Liyan He, Ryszard Kierzek, John SantaLucia, Jr, Amy E Walter, and Douglas H Turner. Nearest-neighbor parameters for G·U mismatches: 5′GU3′/3′UG5′is destabilizing in the contexts CGUG/GUGC, UGUA/AUGU, and AGUU/UUGA but stabilizing in GGUC/CUGG. *Biochem.*, 30(46):11124–11132, 1991.

[71] Xiaoying Chen, Jeffrey A McDowell, Ryszard Kierzek, Thomas R Krugh, and Douglas H Turner. Nuclear magnetic resonance spectroscopy and molecular modeling reveal that different hydrogen bonding patterns are possible for G·U pairs: One hydrogen bond for each G·U pair in r(GGC<u>GU</u>GCC)$_2$ and two for each G·U pair in r(GAG<u>UG</u>CUC)$_2$. *Biochem.*, 39(30):8970–8982, 2000.

[72] Naoki Sugimoto, Ryszard Kierzek, Susan M Freier, and Douglas H Turner. Energetics of internal GU mismatches in ribooligonucleotide helixes. *Biochem.*, 25(19): 5755–5759, 1986.

[73] Ming Wu, Jeffrey A McDowell, and Douglas H Turner. A periodic table of tandem mismatches in RNA. *Biochem.*, 34(10):3204–3211, 1995.

[74] Jeffrey A McDowell and Douglas H Turner. Investigation of the structural basis for thermodynamic stabilities of tandem GU mismatches: Solution structure of (rGAG<u>GU</u>CUC)$_2$ by two-dimensional NMR and simulated annealing. *Biochem.*, 35(45):14077–14089, 1996.

[75] Mai-Thao Nguyen and Susan J. Schroeder. Consecutive terminal GU pairs stabilize RNA helices. *Biochem.*, 49(49):10574–10581, 2010.

[76] Jonathan L Chen, Abigael L Dishler, Scott D Kennedy, Ilyas Yildirim, Biao Liu, Douglas H Turner, and Martin J Serra. Testing the nearest neighbor model for canonical RNA base pairs: Revision of GU parameters. *Biochem.*, 51(16): 3508–3522, 2012.

[77] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288(5):911–940, 1999.

[78] Yongping Pan, U Deva Priyakumar, and Alexander D MacKerell. Conformational determinants of tandem GU mismatches in RNA: Insights from molecular dynamics simulations and quantum mechanical calculations. *Biochem.*, 44(5): 1433–1443, 2005.

[79] Jeffrey A McDowell, Liyan He, Xiaoying Chen, and Douglas H Turner. Investigation of the structural basis for thermodynamic stabilities of tandem GU wobble pairs: NMR structures of (rGGAGUUCC)$_2$ and (rGGAUGUCC)$_2$. *Biochem.*, 36 (26):8030–8038, 1997.

[80] Roopa Biswas, Markus C Wahl, Changill Ban, and Muttaiya Sundaralingam. Crystal structure of an alternating octamer r(GUAUGUA)dC with adjacent G·U wobble pairs. *J. Mol. Biol.*, 267(5):1149–1156, 1997.

[81] Se Bok Jang, Li-Wei Hung, Mi Suk Jeong, Elizabeth L Holbrook, Xiaoying Chen, Douglas H Turner, and Stephen R Holbrook. The crystal structure at 1.5 Å resolution of an RNA octamer duplex containing tandem G·U basepairs. *Biophys. J.*, 90(12):4530–4537, 2006.

[82] Jaishree Trikha, David J Filman, and James M Hogle. Crystal structure of a 14 bp RNA duplex with non-symmetrical tandem G·U wobble base pairs. *Nucleic Acids Res.*, 27(7):1728–1739, 1999.

[83] Junpeng Deng and Muttaiya Sundaralingam. Synthesis and crystal structure of an octamer RNA r(guguuuac)/r(guaggcac) with G·G/U·U tandem wobble base pairs: Comparison with other tandem G·U pairs. *Nucleic Acids Res.*, 28(21): 4376–4381, 2000.

[84] Prakash Ananth, Gunaseelan Goldsmith, and Narayanarao Yathindra. An innate twist between Crick's wobble and Watson-Crick base pairs. *RNA*, 19(8):1038–1053, 2013.

[85] Darui Xu, Theresa Landon, Nancy L Greenbaum, and Marcia O Fenley. The electrostatic characteristics of G·U wobble base pairs. *Nucleic Acids Res.*, 35 (11):3836–3847, 2007.

[86] Ryuji Utsunomiya, Kyoko Suto, Dhakshnamoorthy Balasundaresan, Akiyoshi Fukamizu, Penmetcha KR Kumar, and Hiroshi Mizuno. Structure of an RNA duplex r(GGCG$_{Br}$UGCGCU)$_2$ with terminal and internal tandem G·U base pairs. *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 62(3):331–338, 2006.

[87] Gerald Weber. TfReg: Calculating DNA and RNA melting temperatures and opening profiles with mesoscopic models. *Bioinformatics*, 29:1345–1347, 2013. doi: 10.1093/bioinformatics/btt133. URL `http://bioinformatics.oxfordjournals.org/content/29/10/1345`.

[88] Jiro Kondo, A-C Dock-Bregeon, Dagmar K Willkomm, Roland K Hartmann, and Eric Westhof. Structure of an A-form RNA duplex obtained by degradation of 6S RNA in a crystallization droplet. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.*, 69(6):634–639, 2013.

[89] H Mizuno and M Sundaralingam. Stacking of crick wobble pair and Watson-Crick pair: Stability rules of GU pairs at ends of helical stems in tRNAs and the relation to codon-anticodon wobble interaction. *Nucleic Acids Res.*, 5(11): 4451–4462, 1978.

[90] Uwe Mueller, Harald Schuebel, Mathias Sprinzl, and Udo Heinemann. Crystal structure of acceptor stem of tRNA$^{Ala}$ from *Escherichia coli* shows unique G·U wobble base pair at 1.16 å resolution. *RNA*, 5(05):670–677, 1999.

[91] M Brandl, M Meyer, and J Sühnel. Water-mediated base pairs in RNA: A quantum-chemical study. *J. Phys. Chem. A*, 104(47):11177–11187, 2000.

[92] Judit E Šponer, Jerzy Leszczynski, Vladimír Sychrovskỳ, and Jirí Šponer. Sugar edge/sugar edge base pairs in RNA: Stabilities and structures from quantum chemical calculations. *J. Phys. Chem. B*, 109(39):18680–18689, 2005.

[93] Judit E Šponer, Nad'a Špačková, Petr Kulhánek, Jerzy Leszczynski, and Jirí Šponer. Non-Watson-Crick base pairing in RNA. quantum chemical analysis of the cis Watson-Crick/sugar-edge base pair family. *J. Phys. Chem. A*, 109(10): 2292–2301, 2005.

[94] Judit E Šponer, Nad'a Špačková, Jerzy Leszczynski, and Jirí Šponer. Principles of RNA base pairing: Structures and energies of the trans Watson-Crick/sugar-edge base pairs. *J. Phys. Chem. B*, 109(22):11399–11410, 2005.

[95] Ross EA Kelly and Lev N Kantorovich. Planar heteropairing possibilities of the DNA and RNA bases: An ab initio density functional theory study. *J. Phys. Chem. C*, 111(10):3883–3892, 2007.

[96] Dhananjay Bhattacharyya, Siv Chand Koripella, Abhijit Mitra, Vijay Babu Rajendran, and Bhabdyuti Sinha. Theoretical analysis of noncanonical base pairing interactions in RNA molecules. *J. Biosci.*, 32(1):809–825, 2007.

[97] Thomas T Perkins. Ångström-Precision Optical Traps and Applications*. *Biophysics*, 43, 2014.

[98] M. Barbi, S. Lepri, Michel Peyrard, and Nikos Theodorakopoulos. Thermal denaturation of a helicoidal DNA model. *Phys. Rev. E*, 68:061909, 2003.

[99] Simona Cocco and Remi Monasson. Statistical mechanics of torque induced denaturation of DNA. *Phys. Rev. Lett.*, 83:5178–81, 1999.

[100] Abhijit Sarkar and John F Marko. Removal of DNA-bound proteins by DNA twisting. *Physical Review E*, 64(6):061909, 2001.

[101] Mariano Cadoni, Roberto De Leo, and Sergio Demelio. Soliton propagation in homogeneous and inhomogeneous models for DNA torsion dynamics. *Journal of Nonlinear Mathematical Physics*, 18(supp02):287–319, 2011.

[102] Jae-Hyung Jeon and Wokyung Sung. An effective mesoscopic model of double-stranded DNA. *Journal of Biological Physics*, 40(1):1–14, 2014.

[103] D Lacitignola, G Saccomandi, and I Sgura. Parametric resonance in a mesoscopic discrete DNA model. *Acta Applicandae Mathematicae*, 132(1):391–404, 2014.

[104] Maria Barbi, Simona Cocco, and Michel Peyrard. Helicoidal model for DNA opening. *Phys. Lett. A*, 253:358–369, 1999.

[105] Kazunori Yanagi, Gilbert G Privé, and Richard E Dickerson. Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. *Journal of Molecular Biology*, 217(1):201–214, 1991.

[106] Marla S Babcock, Edwin PD Pednault, and Wilma K Olson. Nucleic acid structure analysis: mathematics for local Cartesian and helical structure parameters that are truly comparable between structures. *Journal of molecular biology*, 237 (1):125–156, 1994.

[107] Xiang-Jun Lu and Wilma K Olson. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protocols*, 3(7):1213–1227, 2008.

[108] Tauanne D Amarante and Gerald Weber. Analysing DNA structural parameters using a mesoscopic model. *J. Phys.: Conf. Ser.*, 490(1):012203, 2014. doi: 10.1088/1742-6596/490/1/012203. URL `http://iopscience.iop.org/1742-6596/490/1/012203`.

[109] Xiang-Jun Lu and Wilma K Olson. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31(17):5108–5121, 2003.

[110] Helen M Berman, Wilma K Olson, David L Beveridge, John Westbrook, Anke Gelbin, Tamas Demeny, Shu-Hsin Hsieh, AR Srinivasan, and Bohdan Schneider. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysical journal*, 63(3):751, 1992.

[111] Buvaneswari Coimbatore Narayanan, John Westbrook, Saheli Ghosh, Anton I Petrov, Blake Sweeney, Craig L Zirbel, Neocles B Leontis, and Helen M Berman. The Nucleic Acid Database: new features and capabilities. *Nucleic acids research*, 42(D1):D114–D122, 2014.

[112] Izabela Ferreira, Tauanne D. Amarante, and Gerald Weber. DNA terminal base pairs have weaker hydrogen bonds especially for AT under low salt concentration. *The Journal of Chemical Physics*, 143:175101, 2015. doi: 10.1063/1.4934783.

[113] Ivo L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res.*, 31:3429–3431, 2003.

[114] Lucieli Ceolin, Mirian Romitti, Débora Rodrigues Siqueira, Carla Vaz Ferreira, Jessica Oliboni Scapineli, Beatriz Assis-Brazil, Rodolfo Vieira Maximiano, Tauanne Dias Amarante, Miriam Celi de Souza Nunes, Gerald Weber, and Ana Luiza Maia. Effect of 3′ UTR RET variants on RET mRNA secondary structure and disease presentation in medullary thyroid carcinoma. *PloS One*, 11(2):e0147840, 2016.

[115] Gerald Weber. Optimization method for obtaining nearest-neighbour DNA entropies and enthalpies directly from melting temperatures. *Bioinformatics*, 2014. doi: 10.1093/bioinformatics/btu751. URL `http://bioinformatics.oxfordjournals.org/content/early/2014/12/06/bioinformatics.btu751.abstract`.

# Apêndice A

# Softwares TfReg e Varpar

Dois softwares de implementação de diversos modelos do tipo PB são utilizados pelo nosso grupo de pesquisa para obter propriedades físicas de DNA e RNA. Ambos empregam a técnica de matriz de transferência e o método de equivalência termodinâmica para o cálculo da função de partição clássica. O software TfReg [115] está disponível online, no momento na versão 3.0, em `https://sites.google.com/site/geraldweberufmg/tfreg` As principais funções do TfReg são a previsão de temperatura de sequências e o perfil de abertura.

O software Varpar ainda não foi disponibilizado publicamente. Com o Varpar é possível obter os parâmetros dos modelos a partir de um conjunto de dados de temperatura de desnaturação experimental. A otimização dos parâmetros é realizada usando o método de Nelder-Meder

O Modelo Estrural Bidimensional foi implementado em ambos os softwares. Esses softwares foram adaptados para possibilitar a dependência de contexto do estudo de estabilidade do par GU [67] e tal adaptação foi útil para estudar as terminações do DNA [112].

# Apêndice B

# Dados de desnaturação com par guanina-uracila

**Tabela B.1**
Sequências usadas para otimizar os parâmetros associados a GU, adaptadas de Chen et al. [76]. Os mismatches GU foram substituídos pela letra correspondente ao agrupamento-II, colchetes correspondem a G e chaves correspondem a U. A temperatura experimental $T$, a temperatura prevista com nosso modelo $T'$ e o iníce de desnaturação $\tau$ são exibidos. Temperaturas estão em °C, $\tau$ é adimensional e a concentração está em $\mu$M. As sequências autocomplementares foram recalculadas para 200 $\mu$M, veja seção 2.2.2.

| Sequência (5'→ 3') | Sequência complementar (3'→5') | $C_t$ | $T$ | $T'$ | $\tau$ |
|---|---|---|---|---|---|
| A{M1}GC[M1]U | U[M1]CG{M1}A | 100 | 19.5 | 24.5763 | 1.68642 |
| A[M1]GC{M1}U | U{M1}CG[M1]A | 100 | 24.2 | 25.2337 | 1.69529 |
| GUC[W]{W}AC | CAG{W}[W]UG | 200 | 24.8427 | 27.1549 | 1.74299 |
| {M2}UGCA[M2] | [M2]ACGU{M2} | 100 | 25.3 | 27.5081 | 1.72598 |
| C{W}GC[W]G | G[W]CG{W}C | 100 | 26.8 | 32.2711 | 1.79025 |
| G[M1]CG{M1}C | C{M1}GC[M1]G | 100 | 29 | 30.4371 | 1.7655 |
| GU[M1]AAUU{M1}AC | CA{M1}UUAA[M1]UG | 100 | 32.8 | 32.4453 | 1.92328 |
| [SA]UGCA{SA} | {SA}ACGU[SA] | 100 | 33.1 | 35.8652 | 1.83875 |
| GC[W]{W}GC | CG{W}[W]CG | 100 | 33.2 | 31.2948 | 1.77708 |
| AU[M1]CG{M1}AU | UA{M1}GC[M1]UA | 100 | 34.4 | 29.3574 | 1.80194 |
| U[SA]GC{SA}A | A{SA}CG[SA]U | 100 | 35 | 33.6712 | 1.80914 |
| C[W]GC{W}G | G{W}CG[W]C | 100 | 35.7 | 32.2711 | 1.79025 |
| CCA[W]{W}UGG | GGU{W}[W]ACC | 100 | 37.1 | 38.2 | 1.94268 |
| GCA[W]{W}UGC | CGU{W}[W]ACG | 100 | 38.1 | 39.6948 | 1.96647 |
| GGU[W]{W}ACC | CCA{W}[W]UGG | 100 | 38.3 | 42.9898 | 2.01891 |
| GUC[W]{W}GAC | CAG{W}[W]CUG | 100 | 38.7 | 41.3452 | 1.99274 |
| A[M2]UCGA{M2}U | U{M2}AGCU[M2]A | 100 | 38.9 | 36.9974 | 1.92354 |
| AUG{SB}[SB]CAU | UAC[SB]{SB}GUA | 100 | 39.5 | 40.2845 | 1.97585 |
| GGA[W]{W}UCC | CCU{W}[W]AGG | 100 | 40.2 | 41.9742 | 2.00275 |
| UCC[M1]CC | AGG{M1}GG | 200 | 40.3318 | 46.4937 | 1.98217 |
| {M2}UACGUA[M2] | [M2]AUGCAU{M2} | 100 | 40.5 | 35.8652 | 1.90552 |
| CG[SA]AU{SA}CG | GC{SA}UA[SA]GC | 100 | 40.8 | 43.223 | 2.02262 |
| G{M1}AGCU[M1]C | C[M1]UCGA{M1}G | 100 | 40.9 | 46.5802 | 2.07605 |
| {SA}AUGCAU[SA] | [SA]UACGUA{SA} | 100 | 41 | 41.6759 | 1.998 |
| GC{M1}[M1]GC | CG[M1]{M1}CG | 100 | 41.5 | 42.418 | 1.92717 |
| CCU[W]{W}AGG | GGA{W}[W]UCC | 100 | 42 | 40.6179 | 1.98116 |
| CGU{M1}[M1]ACG | GCA[M1]{M1}UGC | 100 | 42.4 | 38.0382 | 1.9401 |
| CC[M1]AAUU{M1}GG | GG{M1}UUAA[M1]CC | 100 | 42.6 | 42.2563 | 2.00724 |
| [SA]AUGCAU{SA} | {SA}UACGUA[SA] | 100 | 42.6 | 44.3707 | 2.14872 |
| GAG[M2]{W}GAG | CUC{M2}[W]CUC | 200 | 43.1975 | 42.1623 | 2.00574 |
| GCU[M2][M2]UGC | CGA{M2}{M2}ACG | 200 | 44.138 | 48.9347 | 2.11353 |
| CU[SA]GC{SA}AG | GA{SA}CG[SA]UC | 100 | 44.4 | 51.8895 | 2.16056 |
| CG[SA]AAUU{SA}CG | GC{SA}UUAA[SA]GC | 100 | 44.7 | 45.7897 | 2.17555 |
| GA[M2][M2]CGC[M2][M1]AG | CU{M2}{M2}GCG{M2}{M1}UC | 200 | 45.0211 | 48.8709 | 2.30673 |
| CUC[M1]CUC | GAG{M1}GAG | 200 | 45.5732 | 44.3074 | 1.99424 |
| CUC[M2][M2]CUC | GAG{M2}{M2}GAG | 200 | 46.1708 | 44.3364 | 2.04034 |
| [SA]UCUAGA{SA} | {SA}AGAUCU[SA] | 100 | 46.3 | 44.7054 | 2.04622 |
| CCA{M2}[M2]UGG | GGU[M2]{M2}ACC | 100 | 46.5 | 50.6185 | 2.14033 |
| [SA]UAGCUA{SA} | {SA}AUCGAU[SA] | 100 | 47.4 | 45.5013 | 2.05888 |
| GUG[SB]UCG | CAC{SB}AGC | 200 | 47.4478 | 48.4049 | 2.05426 |

Table B.1 (continuação)

| | | | | | |
|---|---|---|---|---|---|
| CA{M2}[M2]UGC | GU[M2]{M2}ACG | 200 | 47.4953 | 38.5189 | 1.90945 |
| CA[M2]UCGA{M2}UG | GU{M2}AGCU[M2]AC | 100 | 47.5 | 49.7575 | 2.25056 |
| GGU{M1}[M1]ACC | CCA[M1]{M1}UGG | 100 | 47.6 | 47.8451 | 2.09619 |
| CAGA[M2][M1]AGAC | GUCU{M2}{M1}UCUG | 200 | 47.8778 | 48.6239 | 2.22913 |
| CG[SA]AUAU{SA}CG | GC{SA}UAUA[SA]GC | 100 | 48.3 | 47.3834 | 2.20567 |
| {SA}CCGG[SA] | [SA]GGCC{SA} | 100 | 48.5 | 49.3631 | 2.02089 |
| GA[M2]AGCU{M2}UC | CU{M2}UCGA[M2]AG | 100 | 48.9 | 51.1247 | 2.2764 |
| GGA{M2}[M2]UCC | CCU[M2]{M2}AGG | 100 | 49 | 53.8958 | 2.19249 |
| GCA{M2}[M2]UGC | CGU[M2]{M2}ACG | 100 | 49.2 | 52.267 | 2.16656 |
| GAGU[M2][M1]AGAG | CUCA{M2}{M1}UCUC | 200 | 49.5964 | 47.8979 | 2.2154 |
| GU{M1}AGCU[M1]AC | CA[M1]UCGA{M1}UG | 100 | 50.4 | 48.7543 | 2.23159 |
| GAG{SB}[M1]GAG | CUC[SB]{M1}CUC | 200 | 51.2122 | 50.2511 | 2.13448 |
| GA[M1]GAUC{M1}UC | CU{M1}CUAG[M1]AG | 100 | 51.3 | 51.3452 | 2.28057 |
| GAG{SB}[SB]CUC | CUC[SB]{SB}GAG | 100 | 51.6 | 54.082 | 2.19545 |
| CUG[SA]AU{SA}CAG | GAC{SA}UA[SA]GUC | 100 | 51.8 | 50.7989 | 2.27024 |
| [SA]GCGC{SA} | {SA}CGCG[SA] | 100 | 52.9 | 52.0886 | 2.05767 |
| GCU{M1}{M2}GC[M2][M1]AGC | CGA[M1][M2]CG{M2}{M1}UCG | 100 | 54.4 | 55.4931 | 2.54146 |
| {SA}GGCC[SA] | [SA]CCGG{SA} | 100 | 54.7 | 57.468 | 2.13026 |
| GGC[W]{W}GCC | CCG{W}[W]CGG | 100 | 55 | 58.5459 | 2.2665 |
| GU[SA]UGCA{SA}AC | CA{SA}ACGU[SA]UG | 100 | 55 | 57.0441 | 2.3883 |
| CG[SA][M2]CG{M2}{SA}CG | GC{SA}{M2}GC[M2][SA]GC | 100 | 56 | 51.4672 | 2.15383 |
| {SA}CACGUG[SA] | [SA]GUGCAC{SA} | 100 | 56 | 54.6629 | 2.34329 |
| [SB]CCGG{SB} | {SB}GGCC[SB] | 100 | 57 | 54.3054 | 2.08758 |
| [SA]ACGCGU{SA} | {SA}UGCGCA[SA] | 100 | 57.4 | 54.7529 | 2.20613 |
| CCA[SB]C[M1]UCCU | GGU{SB}G{M1}AGGA | 200 | 57.763 | 62.6242 | 2.49379 |
| GCG[SB]GAC | CGC{SB}CUG | 200 | 58.1586 | 59.2819 | 2.21359 |
| [SB]CAGCUG{SB} | {SB}GUCGAC[SB] | 100 | 58.3 | 58.3378 | 2.26318 |
| A{M1}GCGC[M1]U | U[M1]CGCG{M1}A | 100 | 58.6 | 50.9264 | 2.14523 |
| GG{M2}{SA}CG[SA][M2]CC | CC[M2][SA]GC{SA}{M2}GG | 100 | 59.8 | 61.7409 | 2.4771 |
| GAC[M1]CCAG | CUG{M1}GGUC | 200 | 60.1333 | 55.633 | 2.22014 |
| [SA]GAGCUC{SA} | {SA}CUCGAG[SA] | 100 | 61.1 | 58.5181 | 2.26605 |
| {SA}GACGUC[SA] | [SA]CUGCAG{SA} | 100 | 61.6 | 55.8733 | 2.22396 |
| {SA}ACCGGU[SA] | [SA]UGGCCA{SA} | 100 | 63.1 | 57.879 | 2.25588 |
| CAG[SB]GCUC | GUC{SB}CGAG | 200 | 64.0698 | 61.3595 | 2.31128 |
| CG[SB]UGCA{SB}CG | GC{SB}ACGU[SB]GC | 100 | 64.1 | 64.2161 | 2.52389 |
| GGC{M1}[M1]GCC | CCG[M1]{M1}CGG | 100 | 65.9 | 66.0809 | 2.38642 |
| CA[SB]CGCG{SB}UG | GU{SB}GCGC[SB]AC | 100 | 66.2 | 64.958 | 2.53791 |
| GUCG[SB]GCC | CAGC{SB}CGG | 200 | 68.5737 | 68.6522 | 2.42734 |
| UC[M1]CCAGAGG | AG{M1}GGUCUCC | 200 | 70.9226 | 65.0518 | 2.53969 |
| GGC[M2][SA]GGC | CCG{M2}{SA}CCG | 200 | 71.7107 | 69.3753 | 2.43885 |

# Apêndice C

# Artigos publicados

Íntegra dos artigos publicados neste projeto.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

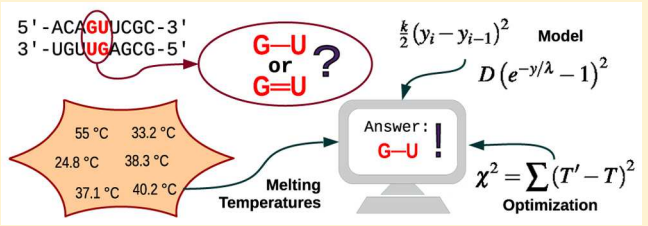
# Evaluating Hydrogen Bonds and Base Stacking of Single, Tandem and Terminal GU Mismatches in RNA with a Mesoscopic Model

Tauanne D. Amarante* and Gerald Weber*

Departamento de Física, Universidade Federal de Minas Gerais, 31270-901 Belo Horizonte-MG, Brazil

Supporting Information

**ABSTRACT:** Guanine−Uracil (GU) mismatches are crucial to the stability of the RNA double helix and need to be considered in RNA folding algorithms for numerous biotechnological applications. Yet despite its importance, many aspects of GU base pairs are still poorly understood. There is also a lack of parametrization which prevents it to be considered in mesoscopic models. Here, we adapted the mesoscopic Peyrard−Bishop model to deal with context-dependent hydrogen bonds of GU mismatches and calculated the model parameters related to hydrogen bonding and base stacking from available experimental melting temperatures. The context-dependence causes a proliferation of parameters which made the problem computationally very demanding. We were able to overcome this problem by systematically regrouping the parameters during the minimization procedure. Our results not only provide the much needed parametrization but also answer several questions about the general properties of GU base pairs, as they can be associated straightforwardly to hydrogen bonding and base stacking. In particular, we found a very small Morse potential for tandem 5′-GU-3′, which confirms a single hydrogen bond for this configuration, answering a long-standing question over conflicting experimental findings. Terminal GU base pairs are known to increase the duplex stability, but it is not clear why. Our results suggest that the increased terminal stability is mostly due to stronger hydrogen bonding.




## INTRODUCTION

The mismatch GU is the most commonly found base pair besides AU and CG in RNA. The unique chemical and structural properties of GU wobble pairs make them special sites for recognition of some biomolecules.[1] The conservation of this motif in specific sites along evolution is more evidence of its functional importance. For example, most living organisms present one GU mismatch at the third position of $tRNA^{Ala}$ that allows the recognition by the enzyme that attaches the amino acid alanine to its tRNA.[2] The GU mismatch is also linked to the RNA catalysis function; for example, in nearly all organisms, in group I self-splicing introns, there is a GU pair at the site of cleavage.[1]

In the wobble hypothesis, Crick proposed that GU could form two hydrogen bonds similar to the canonical AU pair.[3] This was confirmed by early spectroscopic measurements which obtained GU free energies similar to AU.[4,5] However, later experimental studies showed that the stability of internal GU mismatches depends very much on the neighbor context.[6] In particular, thermal stability of GU symmetric tandem base pairs depends on the direction in which they are arranged, with 5′-UG-3′ being generally more stable than 5′-GU-3′. The stability is also influenced by flanking Watson Crick base pairs.[7] Early NMR studies[6] did not attribute this change in stability to a difference in hydrogen-bonding pattern. However, later NMR experiments concluded that the symmetric tandem GU base pair may have either one or two hydrogen bonds depending on mismatch sequence and flanking pairs.[8]

In terms of theoretical studies, a comprehensive analysis of nearest-neighbor (NN) parameters for GU was carried out recently by Chen et al.[9] However, the NN model reveals very little about the intramolecular interactions due to its fundamental limitation of not being able to separate the hydrogen bonds from the stacking interactions.[10] At the other extreme of the theoretical complexity are molecular dynamics simulations which, due to finite computational resources, are typically limited to the analysis of just a few sequences.[11,12] Even more restricted are density functional theroy (DFT) calculations which study the interactions of isolated GU wobble pairs but do not include the RNA backbone and no stacking interaction.[13−18] Mesoscopic approaches such as the Peyrard−Bishop model[19] overcome the limitations of nearest-neighbor models by treating separately the hydrogen bonds and stacking parameters. The combination of this fundamental property of mesoscopic models with experimental melting temperatures provides a way to calculate the strength of hydrogen bonds,[20] which is a difficult property to measure[21] or calculate.[22,23] This has enabled us to obtain fundamental insights into DNA,[20] RNA,[24] and more recently deoxyinosine.[25] These recent advances are a major motivation to extend this type of analysis to GU mismatches.

The simplified model proposed by Peyrard and Bishop[19] consists of a 2D Hamiltonian that takes into account the

Received: June 19, 2015
Published: December 1, 2015





DOI: 10.1021/acs.jcim.5b00571
J. Chem. Inf. Model. 2016, 56, 101−109

stacking interaction and hydrogen bond as separate potentials. The model Hamiltonian is easily modified to include different aspects of nucleotide interactions and can be used either in the framework of equilibrium physical statistics or dynamics. Some examples of recent applications are the analysis of cyanobacterial promoters,[26] protein induced DNA bubbles,[27] and fast prediction of bubble openings.[28] Modified Hamiltonians were proposed to add additional barriers for base pair to model A-DNA[29] or the unzipping induced by force.[30] Further modifications of the model Hamiltonian include the description of structural parameters such as the rise of the helical steps[31] and salt-dependent Morse potentials.[32]

Here, we use the Peyrard–Bishop model to obtain the hydrogen bond and stacking parameters for GU base pairs in RNA from published melting temperatures.[9] In comparison to our previous work on RNA CG and AU base pairs,[24] this represents a considerable challenge since we cannot assume *a priori* a uniform hydrogen bond strength for GU mismatches. In other words, we are not dealing with a single value for hydrogen bonds as previously for CG or AU.[24] Instead, we need to consider the possibility of multiple values for hydrogen bond strengths depending on context, increasing the parameter searching space dramatically to 114 parameters (40 Morse potentials and 74 stacking parameters). Not only does this represent a huge computational challenge for a nonlinear minimization, it also exceeds the number of sequences in the data set.

To overcome the challenges represented by the large number of parameters, we approached the problem in several steps. First we optimized only the 40 Morse potentials independently and kept all stacking parameters constant. Then we gradually reduced the number of Morse potentials into groups of similar strengths until we ended with just five different potentials. By reducing the number of Morse potentials we were able to reduce also the number of stacking parameters from 74 to 40, that is, reducing the problem to optimizing just 45 parameters in total.

This process was successful in optimizing the GU Morse potentials into the known groups of one or two hydrogen bonds. We also obtained larger hydrogen bonds for terminal GUs which are known for their increased stability. In contrast to the variable hydrogen bonds, the GU stacking parameters are very similar in order of magnitude to RNA Watson–Crick base pairs. We found overall agreement with independent experimental measurements such as NMR, and we believe we were able to settle a specific question about the possibility of a single hydrogen bond for tandem GU.

## MATERIALS AND METHODS

**The Mesoscopic Model.** We used the model proposed by Peyrard and Bishop[19] with harmonic stacking interaction, which is the model with the smallest number of parameters. This simple model has provided good results in a variety of situations.[20,24,25]

The main components of this model are the hydrogen bond represented by a Morse potential

$$V(y_i) = D_i(e^{-y_i/\lambda_i} - 1)^2 \quad (1)$$

and the stacking interaction

$$w(y_i, y_{i+1}) = \frac{k_{i,i+1}}{2}(y_i^2 - 2y_iy_{i+1}\cos\theta + y_{i+1}^2) \quad (2)$$

where the parameters $D$ and $\lambda$ depend on the base pair $i$ and the elastic constant $k$ is related to the interaction between subsequent pairs $i$ and $i + 1$. The small angle $\theta$ (0.01 rad) was introduced to avoid numerical problems in the partition function integral.[33]

Equations 1 and 2 form the model Hamiltonian which is used in the classical partition function.[19] The partition function is calculated numerically using the transfer integral technique.[33] For the integration of eq 14 of ref 34, we used 400 points over the interval $y_{min} = -0.1$ nm to $y_{max} = 20.0$ nm, and a cutoff of $P = 10$ of eq 22 of ref 34. The calculation of the thermal index $\tau$ is carried out at 370 K. Please note that this temperature is unrelated to the temperatures obtained from the regression method. For further details on the model implementation, please see refs 20, 34, and 35.

**Experimental Data Used.** Experimental melting temperature data were taken from the comprehensive review by Chen et al.[9] They reported an expanded database of 80 sequences that provides all possible combinations of base triplets containing GU pair flanked by canonical pairs, that is AU and CG, in different orientations. This broad set of melting measurements was achieved by adding some oligonucleotides designed to extend a previous database reported by Mathews et al.[36] These melting temperatures were retrieved from different groups, which makes it difficult to estimate a consistent experimental uncertainty for this set, especially as this was not explicitly provided. We recalculated the melting temperatures of self-complementary sequences to 200 $\mu$M, following the same approach as used by Xia et al.[37] and to be consistent with our previous calculations for RNA,[24] see Supporting Table S1.

**Sequence Decomposition and Notation.** In this work, the GU base pairs are uniquely described in terms of context trimers NGN/NUN or NUN/NGN, where N stands for any base. For instance, a GU flanked by UA on the 5′ side and a AU on the 3′ side is identified by the trimer context UGA/AUU. For clarity, the central GU base pairs are underlined. If the GU mismatch is located at one of the terminals of the helix, this will be indicated as a terminal-dimer NG*/NU*, which we will treat simply as yet another trimer with a pseudobase pair **. Please note that the GU base pair may be flanked by another GU for the case of tandem mismatches. The following example illustrates the procedure adopted in this work. Consider a sequence containing a GU at the 5′ terminal and an internal tandem GU

5′-UCAUGUGG-3′

3′-GGUGUACC-5′

its initial decomposition into trimers would be

*UC/*GG, UCA/GGU, CAU/GUG, AUG/UGU, UGU/GUA, GUG/UAC, UGG/ACC GG*/CC*

where the asterisk represents the terminal of the helix. All trimers are subsequently symmetry-reduced. For instance, out of the two symmetry equivalent trimers,

5′-AUG-3′ 5′-UGU-3′

3′-UGU-5′ 3′-GUA-5′

we retain only the one with lowest lexical order

5′-AUG-3′

3′-UGU-5′



DOI: 10.1021/acs.jcim.5b00571
J. Chem. Inf. Model. 2016, 56, 101–109

that is, we keep AUG/UGU because it alphabetically precedes UGU/GUA. Therefore, the UGU/GUA trimer of our example sequence will be replaced by its equivalent AUG/UGU.

Similarly, for terminal dimers the symmetry lexical reduction results in

$$\begin{matrix} 5'\text{-}*UC\text{-}3' \\ 3'\text{-}*GG\text{-}5' \end{matrix} \Rightarrow \begin{matrix} 5'\text{-}GG*\text{-}3' \\ 3'\text{-}CU*\text{-}5' \end{matrix}$$

therefore, all terminal-dimers considered here will be shown with */* at the 3′ side, since * has a higher lexical order than A, C, G, or U.

After applying all symmetry reductions, the final trimer sequence of our example would be

GG*/CU*, UCA/GGU, CAU/GUG, AUG/UGU, AUG/UGU, CAU/GUG, GG*/CC*

Independent Morse potentials will be attributed for each context trimer or groups of context trimers. For example, we may consider an independent Morse parameter $D^{\mathrm{UGA/AUU}}$ uniquely for UGA/AUU contexts. Alternatively, we may use a parameter $D^{X}$ for a group X containing a collection of context trimers which will be independent of all other context groups. On occasion, we will refer to a generic Morse potential $D^{\alpha}$ where $\alpha$ stands for specific context trimers or to $D^{\beta}$ where $\beta$ stands for context groups.

*Base-Pair Notation.* Context specific base pairs will be shown with an added superscript $\alpha$, that is, $GU^{\alpha}$. From the point of view of base-pair parameters such as the Morse potential $D^{\alpha}$, $GU^{\alpha}$ is equivalent to $UG^{\alpha}$.

*Nearest-Neighbor Notation.* Adapting the typical intrastrand notation 5′-GU-3′/3′-UG-5′ to the notation above would lead to something as unwieldy as 5′-$G^{A}U^{B}$-3′/3′-$U^{A}G^{B}$-5′. Therefore, we preferred to keep a base-pair oriented notation $GU^{A}pUG^{B}$, that is, a base pair $GU^{A}$ followed by another $UG^{B}$, and drop the redundant 5′ and 3′ notation altogether.

**Melting Temperature Prediction.** Given a set of tentative model parameters $P = \{p_1, p_2,...p_L\}$ consisting of Morse potentials and stacking parameters, we calculate an adimensional melting index $\tau_i(P)$ for each sequence $i$ from the partition function of the Peyrard–Bishop Hamiltonian.[38] The melting temperature $T_i'(P)$ resulting from the tentative set of parameters $P$ is then obtained from the following linear equation,

$$T_i'(P) = a_0(N) + a_1(N)\tau_i(P) \tag{3}$$

where the regression coefficients are dependent on the sequence length $N$

$$a_k(N) = b_{0,k} + b_{1,k}N^{1/2},\ k = 0,\ 1 \tag{4}$$

since we have found that the coefficients $a_{0,1}$ are essentially linear with $N^{1/2}$.[38]

**Minimization Procedure.** *Optimization Method.* Here, we briefly outline the optimization method used to obtain the model parameters, which is described in detail in refs 20 and 38. A similar method was also used successfully to calculate the parameters for the Gibbs free energy nearest-neighbor model.[39]

For each tentative set of model parameters $P_j$, we calculate the predicted melting temperatures $T_i'(P_j)$, eq 3, and compare them to the experimental temperatures $T_i$. The model parameters $(P_j)$ are then varied until we minimize the squared differences

$$\chi_j^2 = \sum_{i=1}^{N} [T_i'(P_j) - T_i]^2 \tag{5}$$

The minimization is implemented numerically by the Nelder–Mead or downhill simplex method.[20] To refine the optimized parameters, the minimization is repeated two more times using as new starting points the parameters from the previous round.

Occasionally, we also refer in this work to an average melting temperature deviation

$$\langle \Delta T \rangle = \frac{1}{N}\sum_{i=1}^{N} |T_i' - T_i| \tag{6}$$

Due to the large number of possible GU mismatch contexts, the minimization procedure of eq 5 was carried out in five separate minimization rounds.

*Initial Parameters.* For the first four rounds (MR1–MR4), we vary the initial parameters randomly around a given value $p$ in the interval of $[0.5p,1.5p]$ such that for every minimization step we try to approach the global minima from a different direction.

*Minimization Round 1 (MR1).* In this step, we considered that the hydrogen bonding pattern for a GU mismatch is unique for each trimer context. In other words, an independent Morse parameter $D^{\alpha}$ was associated with each of the 40 different trimer contexts $\alpha$ present in the data set. Stacking parameters associated with GU were fixed at 2.5 eV/nm$^2$. The $\lambda$ parameters were kept constant at 0.03 nm for all minimization rounds. For the remaining AU and CG base pairs, we used the RNA parameters recently calculated for the PB model in ref 24. In order to avoid local minima during the minimization of eq 5, the procedure was carried out 300 times, each time with different set of initial parameters,[20] as described in the previous section. These initial parameters $D$ were randomly chosen between 15 and 45 meV, that is, ±50% of Morse potential calculated for AU.[24] The final total squared difference for MR1 was $\chi^2$ = 1453 °C$^2$ and required on the order of 6000 h on 2 GHz processors.

*Minimization Round 2 (MR2).* From the results of MR1, the trimer contexts were grouped together into seven context groups, W, M1, M2, S1, S2, S3, and S4, chosen by the similarity of their calculated Morse potentials. This time we used as initial values for the Morse potentials the averaged values for each context group. The minimization was repeated again as for MR1, but only 40 times as this step was only to test the initial arrangement. The final total squared difference for MR2 was $\chi^2$ = 1426 °C$^2$ and required 280 h of computation time.

*Minimization Round 3 (MR3).* We evaluated again the resulting Morse potentials from MR2 and identified the possibility of reducing further the number of context groups by joining groups S1 and S2 into SA and groups S3 and S4 into SB. We used the averaged Morse potentials as initial values. This minimization was carried out 300 times. This round resulted in $\chi^2$ = 1431 °C$^2$ and took 1500 h of computation time.

*Minimization Round 4 (MR4).* After verifying the results from MR3, we allowed the stacking constants to vary as well, adding further 40 parameters to the minimization. The final value for the total squared difference was $\chi^2$ = 1023 °C$^2$ and required 3600 h of computation time.

*Minimization Round 5 (MR5).* To obtain an error estimate of the influence of the experimental uncertainty,[20] we carried out one final minimization. This uses the averaged results of MR4, but instead of varying the initial parameters we now


DOI: 10.1021/acs.jcim.5b00571
J. Chem. Inf. Model. 2016, 56, 101–109

varied the experimental melting temperatures. For each round, a small random amount $\delta T_i$ (positive or negative) was added to the reported melting temperature $T_i$. The random $\delta T_i$ follows a Gaussian distribution such that the resulting standard deviation of the modified set matches the experimental uncertainty of 1.3 °C. This procedure was repeated 300 times and provides an estimate of the error for each parameter involved in the minimization. The final parameters, obtained by averaging these runs, correspond to $\chi^2$ = 920 °C$^2$ with 3900 h of calculation time. For further assessment of the influence of the experimental uncertainty, the complete calculation of MR5 was carried out again with 0.5 °C deviation, which we refer to as MR5′.

## RESULTS

The experimental data set used here[9] contains 80 sequences with 32 unique context trimers and eight terminal dimers containing a GU mismatch as shown in Table 1. In the Materials and Methods, we describe the details on how the sequences are divided into context trimers and terminal dimers.

**Table 1. Number of Occurrences *n* of GU Mismatches Per Context Trimers or Context Group[a]**

| Trimer context | *n* | Arrangement I | Arrangement II |
|---|---|---|---|
| AGU/UUG | 6 | 23 (W) | 23 (W) |
| CGU/GUG | 8 | | |
| GUA/UGU | 5 | | |
| CGG/GUC | 4 | | |
| UGA/AUU | 4 | 39 (M1) | 39 (M1) |
| CGA/GUU | 2 | | |
| AUG/UGC | 5 | | |
| GGA/UUU | 5 | | |
| GUA/CGU | 4 | | |
| UGA/GUU | 4 | | |
| GGC/CUG | 2 | | |
| CGC/GUG | 4 | | |
| CUG/GGU | 5 | | |
| AGG/UUC | 4 | | |
| CGG/GUU | 5 | 35 (M2) | 35 (M2) |
| UGG/AUU | 2 | | |
| AG*/UU* | 4 | | |
| AGU/UUA | 4 | | |
| GGU/CUG | 1 | | |
| AUU/UGG | 1 | | |
| AGG/UUU | 2 | | |
| AGA/UUU | 2 | | |
| GGC/UUG | 6 | | |
| AUG/UGU | 8 | | |
| GGA/CUU | 8 | 23 (S1) | 45 (SA) |
| UU*/AG* | 4 | | |
| GG*/CU* | 4 | | |
| UG*/AU* | 4 | | |
| AUA/UGU | 2 | | |
| CUU/GGG | 1 | | |
| CG*/GU* | 4 | 22 (S2) | |
| CU*/GG* | 4 | | |
| GGG/CUU | 4 | | |
| CUA/GGU | 4 | | |
| AU*/UG* | 6 | | |
| GUG/CGU | 5 | 8 (S3) | 18 (SB) |
| AUC/UGG | 3 | | |
| AGC/UUG | 3 | 10 (S4) | |
| CUC/GGG | 3 | | |
| GU*/CG* | 4 | | |

[a]The trimers contexts listed in ascending Morse potential order from MR1, which is also the same order shown in Figure 1. Also shown are the tentative context arrangements and respective number of context trimers contained within each group.

For the remaining discussion, we will consider the terminal dimers as yet another trimer with a */* representing a helix terminal as described in the Materials and Methods.

Unlike our previous parametrization for RNA,[24] we cannot attribute a uniform Morse potential $D$ for GU mismatches. One possibility would be to collect information on GU stability trends from the literature to form groups of contexts which would reduce the number of parameters to minimize. Unfortunately, information about GU stability trends, hydrogen bonding, and stacking interaction is available only for few sequence contexts, and the experimental techniques and conditions differ significantly. Furthermore, we would be introducing an undesirable bias into our calculations. Therefore, we decided to consider an independent Morse potential for each of the 40 possible trimer contexts. In other words, each GU Morse potential is considered independent from all others. While this is interesting, as it prevents grouping biases for the initial minimization, it is challenging to minimize over such a large number of parameters. This is the reason why it was necessary to keep all stacking parameters at a single constant value for the initial minimization (MR1), as otherwise we would be adding a further 74 parameters to the searching space. For this same reason, it would not be feasible to use more complex potentials which require more parameters, such as including the rise distance proposed in ref 31. Also, the Morse potential width $\lambda$ was kept constant for all GU contexts and for all minimization rounds. This was based on our observation[20,24] that $\lambda$ has no significant influence on the final $\chi^2$ value and consequently does not influence the values of $D$ and $k$.

We performed the first round of minimizations (MR1) by letting all 40 Morse potentials vary freely. In Figure 1, we show the averaged Morse potentials resulting from the first minimization round MR1. Most Morse potentials resulted in a range of 25 to 50 meV, which supports the early notions of two hydrogen bonds for most GU mismatches.[3] For comparison, the Morse potential for AU was estimated as 38 meV for the same type of mesoscopic model.[24] However, there is a group of GU contexts with considerably smaller Morse potentials, in the range of 8−20 meV, which suggest much weaker hydrogen bonds.

The error bars shown in Figure 1 are not indicative of the statistical uncertainty of the minimization but represent the numerical difficulty in performing a finite number of minimization rounds over a 40-dimensional parameter space of a nonlinear model. In other words, if we were given an unlimited amount of time, computer resources, or perhaps a more efficient minimization algorithm, these error bars should tend to zero; that is, they should all tend to the same global minimum. Nevertheless, they are helpful in providing guidance for our first attempt in grouping together trimer contexts with similar Morse potentials. Another source of numerical difficulty is that by considering 40 different Morse parameters there is only a reduced number of occurrences of GU mismatches for each trimer context as shown in Table 1.

From the analysis of Figure 1, we selected a tentative arrangement of seven trimer context groups (see color coding), called arrangement I. This increases considerably the number of GU occurrences per trimer group as shown in Table 1 and reduces the number of Morse potentials to 7. The minimization MR2 is the first test of this arrangement where the stacking parameters are still kept constant. The optimized Morse parameters for MR2 are shown in Figure 2. The context groups W (W = weak), M1, and M2 (M = medium) are close to the


DOI: 10.1021/acs.jcim.5b00571
J. Chem. Inf. Model. 2016, 56, 101−109

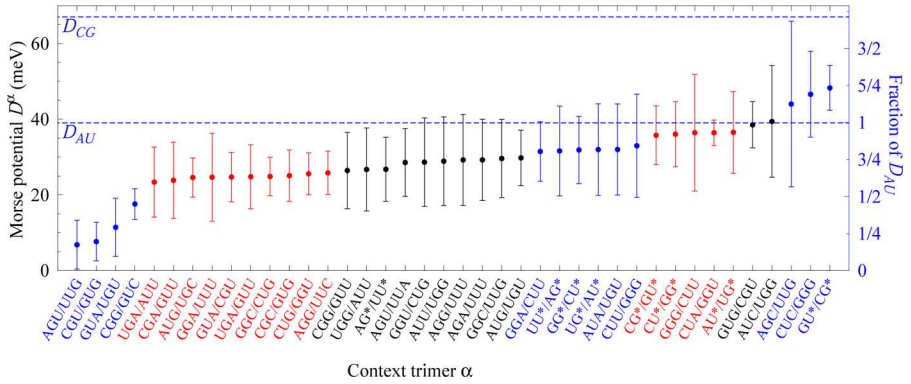



**Figure 1.** Average Morse potentials $D^{\alpha}$ obtained for each trimer context $\alpha$ from MR1. Trimers are shown in order of increasing $D^{\alpha}$. AU and CG Morse parameters are from ref 24 and are shown as blue dashed lines for reference. The left blue scale shows the Morse potentials as fractions of the AU Morse parameter $D_{AU}$. The colors of the trimers on the horizontal axis refer to context arrangement I used for MR2.

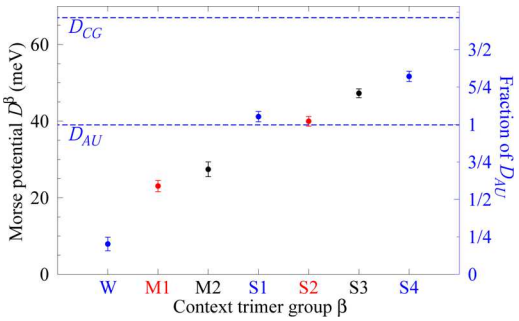



**Figure 2.** Average Morse potentials $D^{\beta}$ obtained for each trimer context group $\beta$ from MR2. Context groups are from arrangement I shown in Table 1. Remaining figure elements are as in Figure 1.

average values of the individual contexts of Figure 1. However, for context groups S1−S4 (S = strong), we observe a somewhat larger Morse potential than for the individual contexts. The error bars of Figure 2, as for Figure 1, are representative of the numerical nonconvergence of the multidimensional minimization. However, they are now considerably smaller since we started out the minimization with a much better initial knowledge of the Morse potentials.

The analysis of Figure 2 suggests that further grouping should be possible. Therefore, we decided to join groups S1 and S2 into group SA and groups S2 and S4 into group SB for the similarity of their Morse potentials. This is now arrangement II, see Table 1, which reduces the Morse potentials to just 5. A more extensive minimization MR3 was carried out which confirmed the stability of this new arrangement. After this round, we now included the stacking parameters into the minimization which increases the searching space to 45 parameters and carried out round MR4. The resulting Morse potentials of MR3 and MR4 are shown in Supporting Figure S1. We also note that the merit function $\chi^2$ gradually reduces from 1453 °C$^2$ for MR1 to 1431 °C$^2$ for MR3. The stability of $\chi^2$ indicates that at no point was the system under-determined, that is, that the number of data points was sufficient such that no parameter has become a function of any other parameter.

Recalling that the stacking interaction parameter $k$ was kept fixed during MR1−MR4, there is now the question of whether the arrangements could have had a different outcome if a different fixed value for $k$ had been selected. Fortunately, the form of the Hamiltonian, composed of the sums of eqs 1 and 2, assures that as long as $k$ is fixed for all Morse potentials, the arrangement will not change. A different value of $k$ would simply shift uniformly all the Morse potentials resulting in the same arrangement.

With the results of MR4, we performed the final minimization MR5 which differs from the previous ones by varying the experimental data set, see the Materials and Methods for details. Figure 3 shows the final Morse potential

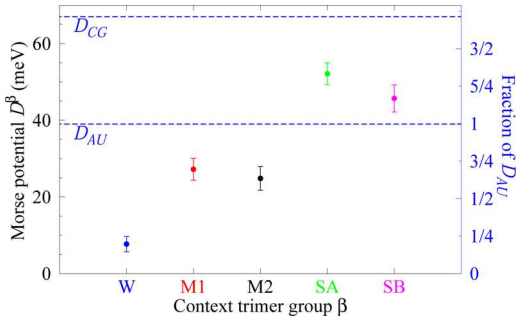



**Figure 3.** Average Morse potentials $D^{\beta}$ obtained for each trimer context group $\beta$ from MR5. Context groups are from arrangement II shown in Table 1. Remaining figure elements are as in Figure 1.

for the remaining five context groups, and Tables 2 and 3 show the 40 stacking parameters for arrangement II and MR5. MR5 differs from MR4 in that we now randomly change the melting temperatures by small amounts, which allows us to evaluate the effect of the estimated 1.3 °C experimental uncertainty on the

**Table 2. Final Stacking Parameters $k$ in eV/nm$^2$, for Nontandem GU Mismatches, from MR5**[a]

| NN | $k$ | NN | $k$ | NN | $k$ |
|---|---|---|---|---|---|
| AUpGU$^{M2}$ | 2.9(4) | AUpGU$^{SB}$ | 3.6(9) | AUpGU$^{M1}$ | 3.0(5) |
| AUpGU$^{W}$ | 0.9(2) | AUpUG$^{M2}$ | 4.2(5) | AUpUG$^{SB}$ | 2.8(6) |
| AUpUG$^{SA}$ | 2.5(3) | AUpUG$^{M1}$ | 3.7(7) | CGpGU$^{M2}$ | 2.4(4) |
| CGpGU$^{SA}$ | 1.9(2) | CGpGU$^{M1}$ | 2.1(3) | CGpGU$^{W}$ | 2.8(3) |
| CGpUG$^{SB}$ | 2.4(4) | CGpUG$^{SA}$ | 1.8(2) | CGpUG$^{M1}$ | 2.6(3) |
| CGpUG$^{W}$ | 4.7(5) | GCpGU$^{M2}$ | 2.7(6) | GCpGU$^{SB}$ | 3.2(6) |
| GCpGU$^{SA}$ | 1.7(2) | GCpGU$^{M1}$ | 2.4(3) | GCpUG$^{M2}$ | 3.0(5) |
| GCpUG$^{SB}$ | 2.9(4) | GCpUG$^{M1}$ | 3.1(4) | GU$^{M2}$pAU | 2.1(5) |
| GU$^{SA}$pAU | 2.2(3) | GU$^{M1}$pAU | 1.5(2) | UApGU$^{M2}$ | 2.0(6) |
| UApGU$^{SA}$ | 2.1(3) | UApGU$^{M1}$ | 2.4(3) | UApGU$^{W}$ | 1.6(4) |

[a]Calculated uncertainties are shown in compact notation.



DOI: 10.1021/acs.jcim.5b00571
J. Chem. Inf. Model. 2016, 56, 101−109

Table 3. Final Tandem Stacking Parameters $k$ in eV/nm$^2$, from MR5[a]

| motif | NN | $k$ | Independent measurements of hydrogen bonds |
|---|---|---|---|
| UGpGU | $UG^{SB}pGU^{SB}$ | 1.8(5) | 2 hydrogen bonds NMR,[8,40] MD[11] |
| | $UG^{SB}pGU^{M1}$ | 2.6(5) | |
| | $UG^{M2}pGU^{M2}$ | 3.1(7) | 2 hydrogen bonds NMR,[6,41] X-ray,[42] MD[11] |
| | $UG^{M1}pGU^{M1}$ | 3.3(9) | 2 hydrogen bonds NMR,[40] MD[11] |
| GUpUG | $GU^{W}pUG^{W}$ | 1.9(5) | 1 hydrogen bond NMR,[6,8] MD[11] 2 hydrogen bonds NMR,[41] Xray[43] |
| | $GU^{M2}pUG^{W}$ | 2.6(5) | |
| GUpGU | $GU^{M2}pGU^{M2}$ | 1.9(5) | 2 hydrogen bonds X-ray[44,45] |
| | $GU^{M2}pGU^{M1}$ | 1.9(4) | |
| | $GU^{SA}pGU^{M2}$ | 3.2(8) | |
| | $GU^{M2}pGU^{SA}$ | 3.5(1) | |

[a]Calculated uncertainties are shown in compact notation. Also shown are references which independently determined the number of hydrogen bonds for each stacking NN configuration.

Table 4. Correspondence between the Flat Nearest-Neighbor (NN) Notation Used in This Work and Elsewhere in the Literature

| NN | structure | groups | equivalent notation and reference |
|---|---|---|---|
| UGpGU | 5′-UG-3′<br>3′-GU-5′ | M1, M2, SB | motif I[45−47] |
| | | | 5′-UG-3′[11] 5′UG3′[8] |
| | | | 5′UG/3′GU[9,48] |
| | | | 5′UG/GU3′[49] |
| | | | 5′U·G/G·U3′[12] |
| | | | U·G/G·U[45] |
| | | | 5′-UG-3′/3′-GU-5′[45,46] |
| GUpUG | 5′-GU-3′<br>3′-UG-5′ | W, M2 | motif II[45−47] |
| | | | 5′-GU-3′[11] 5′GU3′[8] |
| | | | 5′GU/3′UG[9,43,48] |
| | | | 5′G·U/U·G3′[12] |
| | | | G·U/U·G[45] |
| | | | 5′-GU-3′/3′-UG-5′[45,46] |
| GUpGU | 5′-GG-3′<br>3′-UU-5′ | M1, M2, SA | motif III[45−47] |
| | | | 5′GG/3′UU[9,48] |
| | | | G·G/U·U[45] |
| | 5′-UU-3′<br>3′-GG-5′ | | 5′-UU-3′/3′-GG-5′[45,46] |
| | | | 5′UU/3′GG[9,48] |

resulting parameters. Therefore, the error bars shown in Figure 3, as well as the displayed uncertainties of Tables 2 and 3, are now of statistical significance. One should note that since the data set comes from various sources, no explicit experimental uncertainty was given.[9] Therefore, we kept the value of 1.3 °C from our previous work on canonical RNA[24] for consistency. However, a smaller uncertainly value basically just reduces the size of the error bars as shown in Supporting Figures S2 and S3, comparing minimization MR5 and MR5′. The average prediction deviation $\Delta T$ for the final parameters of MR5 is 2.7 °C, which is moderately smaller than the prediction deviation of 3.0 °C evaluated for the NN model (calculated from Table 2 of ref 9).

## DISCUSSION

**Tandem GU.** Tandem mismatches are possibly the most extensively studied GU group,[6−9,11,12,40,41,43,45−49] which gave rise to a confusing variety of notations. To aid the following discussion, we compiled the correspondence of our flat notation and some common forms found in the literature in Table 4.

Early NMR measurements by He et al.[6] hinted at the possibility of weaker hydrogen bonds for some GUpUG tandem mismatches, perhaps even with just a single bond. Other measurements, however,[40,41] essentially confirmed the long held view that all GU mismatches, even in tandem configuration, form two hydrogen bonds. A few years later, new NMR data by the same group[8] revived the idea of a single hydrogen bond for a specific sequence $r(GGCGUGCC)_2$ with a GUpUG tandem mismatch. X-ray crystal structure analysis by Jang et al.,[43] however, was unable to confirm this and attributed their difficulties on different experimental conditions as well as on limitations of NMR technique. On the other hand, molecular dynamics and quantum mechanical calculations by Pan et al.[11] appeared supportive of the one hydrogen bond hypothesis but also pointed out that the lower stability of GUpUG could be also due to stacking interactions.

Our results largely support the single hydrogen bond hypothesis for this tandem GU motif. Only two tandem GUpUG configurations appear in the data set for this motif; one is $GU^{W}pUG^{W}$ where both base pairs are in the weakest W group with a mere 8 meV, in stark contrast to the 39 meV for two-hydrogen bonded AU.[24] The $r(GGCGUGCC)_2$ sequence studied by Chen et al.[8] is exactly of this type, as are three of the sequences by He et al.[6] The other is $GU^{M2}pUG^{W}$ with one W-type base pair and a medium-strength M2 Morse potential of 25 meV for which we are not aware of any independent experimental NMR or X-ray measurements. Concerning the possibility raised by Pan et al.[11] that stacking interactions could be responsible for a reduced stability of GUpUG, we found no particular difference in regard to other tandem motifs as shown in Table 3. In fact, the stacking parameters do not show any particular difference to canonical CG or AU base pairs either.[24] Therefore, stacking interactions do not appear to be the primary cause for GUpUG instability, which leaves only a reduced hydrogen bond as a plausible explanation.

The reduced Morse potential alone has a dramatic influence on opening profiles as exemplified in Figure 4, calculated with the new GU parameters using our free-software TfReg.[35] Comparing the strong SB group, Figure 4a, to the weakest W Morse potentials in Figure 4b, we notice a 6-fold difference in the average opening profile $\langle y \rangle$. Note that the stacking parameters for both tandem mismatches are virtually the same, see Table 3.

The other two tandem motifs, the symmetric UGpGU and the nonsymmetric GUpGU, have Morse potentials in the medium (M2, M1) to strong (SB) groups as shown in Table 3. This is consistent with the experimental NMR data for UGpGU[7,8,41] and X-ray diffraction for GUpGU,[44,45] which unanimously attribute two hydrogen bonds for these tandem motifs. In particular, Kondo et al.[50] reported the role played by water molecules in the stabilization of the wobble pairs UGpGU in tandem that could explain why we observed $D$ values larger than those for AU. In addition, the stability of UGpGU tandems can be further refined in terms of flanking base pairs.[7,42] For instance, the stablest UGpGU is the one flanked symmetrically by GC base pairs


DOI: 10.1021/acs.jcim.5b00571
J. Chem. Inf. Model. 2016, 56, 101−109

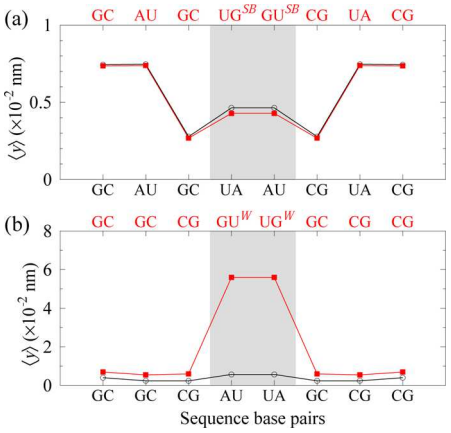



**Figure 4.** Average opening profile calculated for sequences containing symmetric tandem GUpUG mismatches. Shown are the opening profile (red boxes, upper horizontal axes) for sequences (a) r(GGCGUGCC)$_2$ and (b) r(GAGUGCUC)$_2$ analyzed by Chen et al.[8] The canonical sequence analogues (black circles, lower horizontal axes) were obtained by replacing G with A. The shaded area highlights the sequence positions where mismatch and canonical sequences differ. Opening profiles were calculated at $T$ = 150 K; note that this temperature is unrelated to the melting temperatures.

5′-**G**UGC-3′

3′-CGU**G**-5′

In Table 5, we show all trimers associated with a 5′ flanking base which follows closely the predicted trend 5′ G > 5′ C > 5′ U ≥ 5′ A,[7,42] remembering that the M2 Morse potential is somewhat smaller than for M1. A similar trend is observed for GUpUG tandems where 5′ G has higher Morse potentials than the other contexts. However, we should point out that the 5′-G flanking base is inferred only from Morse potentials, as there is no actual sequence with such configuration in the data set.

**Terminal GU.** For terminal GU pairs, there is a general consensus that they stabilize the helix.[47−49,51] What is less clear is where this stabilization comes from, if due to hydrogen bonding or due to stacking interactions. Our results place the eight terminal GU pairs into the context groups with highest Morse potential, SA (52 meV) and SB (46 meV), with the only exception being AG*/UU*, which lies in the intermediate group M2 (25 meV). This suggests that, for the SA and SB groups, this stabilization is due to an increased hydrogen bonding.

It was suggested that the stacking interaction plays a role in the stability of GU terminal mismatches as well, in particular, those with G at the 5′ position are reportedly more stable due to larger stacking overlap.[47,51] In Table 6, we listed the terminal GU mismatches according to whether the position of the G base is at either 5′ or 3′. The table is further organized such that each row shows the same type of flanking base pair (BP), respecting the strand direction. For our results, this is the case for GC, CG, and UA flanking pairs as shown in Table 6, largely reflecting the current consensus.[47,51] For the AU flanking pair, the stacking parameter is smaller for G at 5′, but this appears to be compensated by a much larger Morse potential in the SA group.

**Table 6. Association of Terminal GU Trimers and Stacking Groups**[a]

| BP | 3′-end | NN | $k$ | 5′-end | NN | $k$ |
|---|---|---|---|---|---|---|
| AU | AG*/UU* | AUpGU$^{M2}$ | 2.9 | UU*/AG* | GU$^{SA}$pAU | 2.2 |
| GC | GG*/CU* | GCpGU$^{SA}$ | 1.7 | CU*/GG* | CGpUG$^{SA}$ | 1.8 |
| CG | CG*/GU* | CGpGU$^{SA}$ | 1.9 | GU*/CG* | GCpUG$^{SB}$ | 2.9 |
| UA | UG*/AU* | UApGU$^{SA}$ | 2.1 | AU*/UG* | AUpUG$^{SA}$ | 2.5 |

[a]Each row is for a flanking base pair (BP). Stacking parameters $k$ (eV/nm$^2$) are repeated from Table 2.

**Single Mismatches.** In contrast to tandem and terminal GU mismatches, single GU has seemingly not attracted that much attention, possibly, because early measurements[6] gave no indication of anything but two hydrogen bonds for single GU mismatches, which was confirmed by X-ray measurements for some specific contexts.[52] Indeed, this would appear to be mostly the case for our results as well, as shown in Figure 3. However, a closer inspection of group W, Table 1, shows that a nontandem trimer CGG/GUC also appears in this group with a very low average Morse potential of 8 meV. This would suggest a single hydrogen bond for this particular GU context. In contrast, the crystal structure analyzed by Kondo et al.[50] contains a single mismatch in this particular context and predicts two hydrogen bonds. One possible reason for this disagreement could be the crystallization of the RNA sample for performing X-ray diffraction experiments, while the experimental data used here are for RNA in solution.

For stacking parameters, AUpGU$^W$ also stands out with a much smaller than usual value as shown in Table 2. However, reversing the GU pair as in AUpUG$^W$ shows a stronger than average stacking parameter. While these stacking parameters are nowhere as extreme as recently calculated for deoxyinosine mismatches,[25] they still could influence the melting cooperativity in some important ways. Unfortunately, we are not aware of any independent measurements that could be used for comparison in this case as most structural measurements do not provide an estimate of stacking interaction strengths.

Single mismatches would be the only situation where we could possibly try a comparison to DFT calculations,[13−18] however, they do not consider the RNA backbone and do not include the sequence context. In other words, they are single mismatches without the flanking base pairs. While they are able to consider several type of base pair geometries such as cis Watson−Crick/sugar edge[15] or sugar edge/sugar edge,[14] they

**Table 5. Identification of the Trimers Associated to GU Mismatches in Symmetric Tandem According to the Direction and the Flanking Base Pairs**

| | flanking base-pairs/trimer context (group) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| NN | 5′G | | 5′C | | 5′U | | 5′A | |
| UGpGU | 5′-**G**UG<br>CGU | (SB) | 5′-**C**UG<br>GGU | (M1) | UGA<br>GU**U**-5′ | (M1) | 5′-**A**UG<br>UGU | (M2) |
| GUpUG | [a]5′-**G**GU<br>CUG | (M2) | 5′-**C**GU<br>GUG | (W) | GUA<br>UG**U**-5′ | (W) | 5′-**A**GU<br>UUG | (W) |

[a]No sequence contains this trimer associated with a symmetric tandem GU in the data set of ref 9.



DOI: 10.1021/acs.jcim.5b00571
J. Chem. Inf. Model. 2016, 56, 101−109

cannot specify which type of conformation will be assumed for a given sequence. On the other hand, even though our model can predict the stability of the GU mismatch depending on context, our results for stacking interactions are not detailed enough to infer this base-pair geometry, which means that a direct comparison of our results to the DFT calculations is actually not possible. On a broader basis, some DFT calculations suggest a possibility of up to three hydrogen bonds;[17] however, our SA/SB groups do not confirm this.

## CONCLUSION

Here, we applied successfully a mesoscopic model to GU mismatches. The method involved considering multiple values for Morse potentials depending on the flanking base pair of the GU pair. This provides a way to obtain estimates of hydrogen bond and stacking interaction strengths which are independent from the traditional NMR and crystallographic measurements. This is of importance since there are cases where these measurements provide conflicting results, and a third experimentally derived method could be helpful to resolve those questions. For instance, we confirmed a single hydrogen bond for GUpUG tandem configurations in agreement with NMR measurements[8] while X-ray measurements were suggesting two hydrogen bonds.[43] In some cases, we were able to provide predictions for which, to our knowledge, there are currently no NMR or X-ray measurements available. These encouraging results pave the way to apply the method to other GU configurations such as GU flanked by mismatches[53] or multiple terminal GU.[48] However, at present, our analysis does not cover sequence position dependence of the GU parameters, which would require a considerable number of additional experimental data. Combining the new RNA parameters for GU with the previously calculated AU and CG[24] allows a more comprehensive application of the Peyrard−Bishop model to this important molecule.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00571.

Supporting Table S1 and Figures S2, S2, and S3 (PDF)

## AUTHOR INFORMATION


### Corresponding Authors

*Phone: +55 31 3409 5633. Fax: +55 31 3409 5600. E-mail: tauamarante@gmail.com.

*Phone: +55 31 3409 5633. Fax: +55 31 3409 5600. E-mail: gweberbh@gmail.com


### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This work was supported by Fundação de Amparo a Pesquisa do Estado de Minas Gerais (Fapemig); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Coordenação de Aperfeiçoamento de Nível Superior (Capes).

## ABBREVIATIONS

PB, Peyrard−Bishop; NMR, nuclear magnetic resonance; MR, minimization round; NN, nearest-neighbor; BP, base pair; MD, molecular dynamics; W, weak; M, medium; S, strong

## REFERENCES


(1) Varani, G.; McClain, W. H. The G·U Wobble Base Pair. *EMBO Rep.* **2000**, *1*, 18−23.

(2) Naganuma, M.; Sekine, S.-i.; Chong, Y. E.; Guo, M.; Yang, X.-L.; Gamper, H.; Hou, Y.-M.; Schimmel, P.; Yokoyama, S. The Selective tRNA Aminoacylation Mechanism Based on a Single G·U Pair. *Nature* **2014**, *510*, 507−511.

(3) Crick, F. H. C. Codon-Anticodon Pairing: the Wobble Hypothesis. *J. Mol. Biol.* **1966**, *19*, 548−555.

(4) Freier, S. M.; Kierzek, R.; Caruthers, M. H.; Neilson, T.; Turner, D. H. Free Energy Contributions of G·U and other Terminal Mismatches to Helix Stability. *Biochemistry* **1986**, *25*, 3209−3213.

(5) Sugimoto, N.; Kierzek, R.; Freier, S. M.; Turner, D. H. Energetics of Internal GU Mismatches in Ribooligonucleotide Helixes. *Biochemistry* **1986**, *25*, 5755−5759.

(6) He, L.; Kierzek, R.; SantaLucia, J., Jr; Walter, A. E.; Turner, D. H. Nearest-Neighbor Parameters for G·U Mismatches: 5′GU3′/3′UG5′ is Destabilizing in the Contexts CGUG/GUGC, UGUA/AUGU, and AGUU/UUGA but Stabilizing in GGUC/CUGG. *Biochemistry* **1991**, *30*, 11124−11132.

(7) Wu, M.; McDowell, J. A.; Turner, D. H. A Periodic Table of Tandem Mismatches in RNA. *Biochemistry* **1995**, *34*, 3204−3211.

(8) Chen, X.; McDowell, J. A.; Kierzek, R.; Krugh, T. R.; Turner, D. H. Nuclear Magnetic Resonance Spectroscopy and Molecular Modeling Reveal that Different Hydrogen Bonding Patterns are Possible for G·U Pairs: One hydrogen Bond for Each G·U Pair in r(GGCGUGCC)$_2$ and Two for Each G·U Pair in r(GAGUGCUC)$_2$. *Biochemistry* **2000**, *39*, 8970−8982.

(9) Chen, J. L.; Dishler, A. L.; Kennedy, S. D.; Yildirim, I.; Liu, B.; Turner, D. H.; Serra, M. J. Testing the Nearest Neighbor Model for Canonical RNA Base Pairs: Revision of GU Parameters. *Biochemistry* **2012**, *51*, 3508−3522.

(10) Yildirim, I.; Turner, D. H. RNA Challenges for Computational Chemists. *Biochemistry* **2005**, *44*, 13225−13234.

(11) Pan, Y.; Priyakumar, U. D.; MacKerell, A. D. Conformational Determinants of Tandem GU Mismatches in RNA: Insights from Molecular Dynamics Simulations and Quantum Mechanical Calculations. *Biochemistry* **2005**, *44*, 1433−1443.

(12) Ananth, P.; Goldsmith, G.; Yathindra, N. An Innate Twist Between Crick's Wobble and Watson-Crick Base Pairs. *RNA* **2013**, *19*, 1038−1053.

(13) Brandl, M.; Meyer, M.; Sühnel, J. Water-mediated Base Pairs in RNA: A Quantum-Chemical Study. *J. Phys. Chem. A* **2000**, *104*, 11177−11187.

(14) Šponer, J. E.; Leszczynski, J.; Sychrovskỳ, V.; Šponer, J. Sugar Edge/Sugar Edge Base Pairs in RNA: Stabilities and Structures from Quantum Chemical Calculations. *J. Phys. Chem. B* **2005**, *109*, 18680−18689.

(15) Šponer, J. E.; Špačková, N.; Kulhánek, P.; Leszczynski, J.; Šponer, J. Non-Watson-Crick Base Pairing in RNA. Quantum chemical Analysis of the Cis Watson-Crick/Sugar-Edge Base Pair Family. *J. Phys. Chem. A* **2005**, *109*, 2292−2301.

(16) Šponer, J. E.; Špačková, N.; Leszczynski, J.; Šponer, J. Principles of RNA Base Pairing: Structures and Energies of the Trans Watson-Crick/Sugar-Edge Base Pairs. *J. Phys. Chem. B* **2005**, *109*, 11399−11410.

(17) Kelly, R. E.; Kantorovich, L. N. Planar Heteropairing Possibilities of the DNA and RNA bases: An ab Initio Density Functional Theory study. *J. Phys. Chem. C* **2007**, *111*, 3883−3892.

(18) Bhattacharyya, D.; Koripella, S. C.; Mitra, A.; Rajendran, V. B.; Sinha, B. Theoretical Analysis of Noncanonical Base Pairing Interactions in RNA Molecules. *J. Biosci.* **2007**, *32*, 809−825.

(19) Peyrard, M.; Bishop, A. R. Statistical Mechanics of a Nonlinear Model for DNA denaturation. *Phys. Rev. Lett.* **1989**, *62*, 2755−2757.

(20) Weber, G.; Essex, J. W.; Neylon, C. Probing the Microscopic Flexibility of DNA from Melting Temperatures. *Nat. Phys.* **2009**, *5*, 769−773.

(21) Pervushin, K.; Ono, A.; Fernández, C.; Szyperski, T.; Kainosho, M.; Wüthrich, K. NMR Scalar Couplings Across Watson-Crick Base




DOI: 10.1021/acs.jcim.5b00571
*J. Chem. Inf. Model.* 2016, 56, 101−109

Pair Hydrogen Bonds in DNA Observed by Transverse Relaxation-Optimized Spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 14147−14151.
(22) Kenny, P. W. Hydrogen Bonding, Electrostatic potential, and Molecular Design. *J. Chem. Inf. Model.* **2009**, *49*, 1234−1244.
(23) Szatyłowicz, H.; Sadlej-Sosnowska, N. Characterizing the Strength of Individual Hydrogen Bonds in DNA Base Pairs. *J. Chem. Inf. Model.* **2010**, *50*, 2151−2161.
(24) Weber, G. Mesoscopic Model Parametrization of Hydrogen Bonds and Stacking Interactions of RNA from Melting Temperatures. *Nucleic Acids Res.* **2013**, *41*, e30.
(25) Maximiano, R. V.; Weber, G. Deoxyinosine Mismatch Parameters Calculated with a Mesoscopic Model Result in Uniform Hydrogen Bonding and Strongly Variable Stacking Interactions. *Chem. Phys. Lett.* **2015**, *631*−*632*, 87−91.
(26) Tapia-Rojo, R.; Mazo, J. J.; Hernández, J. Á.; Peleato, M. L.; Fillat, M. F.; Falo, F. Mesoscopic Model and Free Energy Landscape for Protein-DNA Binding Sites: Analysis of Cyanobacterial Promoters. *PLoS Comput. Biol.* **2014**, *10*, e1003835.
(27) Traverso, J. J.; Manoranjan, V. S.; Bishop, A.; Rasmussen, K. Ø.; Voulgarakis, N. K. Allostery through Protein-Induced DNA Bubbles. *Sci. Rep.* **2015**, *5*, 9037.
(28) Zrimec, J.; Lapanje, A. Fast Prediction of DNA Melting Bubbles using DNA Thermodynamic Stability. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2015**, *12*, 1137.
(29) Valle-Orero, J.; Wildes, A.; Theodorakopoulos, N.; Cuesta-López, S.; Garden, J.; Danilkin, S.; Peyrard, M. Thermal Denaturation of A-DNA. *New J. Phys.* **2014**, *16*, 113017.
(30) Bergues-Pupo, A. E.; Falo, F.; Fiasconaro, A. Resonant Optimization in the Mechanical Unzipping of DNA. *EPL* **2014**, *105*, 68005.
(31) Amarante, T. D.; Weber, G. Analysing DNA Structural Parameters using a Mesoscopic Model. *J. Phys.: Conf. Ser.* **2014**, *490*, 012203.
(32) Singh, A.; Singh, N. Effect of Salt Concentration on the Stability of Heterogeneous DNA. *Phys. A (Amsterdam, Neth.)* **2015**, *419*, 328−334.
(33) Zhang, Y.-L.; Zheng, W.-M.; Liu, J.-X.; Chen, Y. Z. Theory of DNA melting based on the Peyrard-Bishop model. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1997**, *56*, 7100−7115.
(34) Weber, G.; Haslam, N.; Essex, J. W.; Neylon, C. Thermal Equivalence of DNA Duplexes for Probe Design. *J. Phys.: Condens. Matter* **2009**, *21*, 034106.
(35) Weber, G. TfReg: Calculating DNA and RNA Melting Temperatures and Opening Profiles with Mesoscopic Models. *Bioinformatics* **2013**, *29*, 1345−1347.
(36) Mathews, D. H.; Sabina, J.; Zuker, M.; Turner, D. H. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. *J. Mol. Biol.* **1999**, *288*, 911−940.
(37) Xia, T.; SantaLucia, J., Jr.; Burkard, M. E.; Kierzek, R.; Schroeder, S. J.; Jiao, X.; Cox, C.; Turner, D. H. Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs. *Biochemistry* **1998**, *37*, 14719−14735.
(38) Weber, G.; Haslam, N.; Whiteford, N.; Prügel-Bennett, A.; Essex, J. W.; Neylon, C. Thermal Equivalence of DNA Duplexes Without Melting Temperature Calculation. *Nat. Phys.* **2006**, *2*, 55−59.
(39) Weber, G. Optimization Method for Obtaining Nearest-Neighbour DNA Entropies and Enthalpies Directly from Melting Temperatures. *Bioinformatics* **2015**, *31*, 871−877.
(40) McDowell, J. A.; Turner, D. H. Investigation of the Structural Basis for Thermodynamic Stabilities of Tandem GU Mismatches: Solution Structure of $(rGGAGUUCC)_2$ by Two-Dimensional NMR and Simulated Annealing. *Biochemistry* **1996**, *35*, 14077−14089.
(41) McDowell, J. A.; He, L.; Chen, X.; Turner, D. H. Investigation of the Structural Basis for Thermodynamic Stabilities of Tandem GU Wobble Pairs: NMR Structures of $(rGGAGUUCC)_2$ and $(rGGAUGUCC)_2$. *Biochemistry* **1997**, *36*, 8030−8038.
(42) Biswas, R.; Wahl, M. C.; Ban, C.; Sundaralingam, M. Crystal Structure of an Alternating Octamer r(GUAUGUA)dC with Adjacent G·U Wobble Pairs. *J. Mol. Biol.* **1997**, *267*, 1149−1156.
(43) Jang, S. B.; Hung, L.-W.; Jeong, M. S.; Holbrook, E. L.; Chen, X.; Turner, D. H.; Holbrook, S. R. The Crystal Structure at 1.5 Å Resolution of an RNA Octamer Duplex Containing Tandem G·U Basepairs. *Biophys. J.* **2006**, *90*, 4530−4537.
(44) Trikha, J.; Filman, D. J.; Hogle, J. M. Crystal Structure of a 14 bp RNA Duplex with Non-Symmetrical Tandem G·U Wobble Base Pairs. *Nucleic Acids Res.* **1999**, *27*, 1728−1739.
(45) Deng, J.; Sundaralingam, M. Synthesis and Crystal Structure of an Octamer RNA r(guguuuac)/r(guaggcac) with G·G/U·U Tandem Wobble Base Pairs: Comparison with other Tandem G·U Pairs. *Nucleic Acids Res.* **2000**, *28*, 4376−4381.
(46) Xu, D.; Landon, T.; Greenbaum, N. L.; Fenley, M. O. The Electrostatic Characteristics of G·U Wobble Base Pairs. *Nucleic Acids Res.* **2007**, *35*, 3836−3847.
(47) Masquida, B.; Westhof, E. On the Wobble GoU and Related Pairs. *RNA* **2000**, *6*, 9−15.
(48) Nguyen, M.-T.; Schroeder, S. J. Consecutive Terminal GU Pairs Stabilize RNA Helices. *Biochemistry* **2010**, *49*, 10574−10581.
(49) Utsunomiya, R.; Suto, K.; Balasundaresan, D.; Fukamizu, A.; Kumar, P. K.; Mizuno, H. Structure of an RNA duplex $r(GGCG_{Br}UGCGCU)_2$ with Terminal and Internal Tandem G·U Base Pairs. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, *62*, 331−338.
(50) Kondo, J.; Dock-Bregeon, A.-C.; Willkomm, D. K.; Hartmann, R. K.; Westhof, E. Structure of an A-form RNA Duplex Obtained by Degradation of 6S RNA in a Crystallization Droplet. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.* **2013**, *69*, 634−639.
(51) Mizuno, H.; Sundaralingam, M. Stacking of Crick Wobble Pair and Watson-Crick Pair: Stability Rules of GU Pairs at Ends of Helical Stems in tRNAs and the Relation to Codon-Anticodon Wobble Interaction. *Nucleic Acids Res.* **1978**, *5*, 4451−4462.
(52) Mueller, U.; Schuebel, H.; Sprinzl, M.; Heinemann, U. Crystal Structure of Acceptor Stem of $tRNA^{Ala}$ From *Escherichia coli* Shows Unique G·U Wobble Base Pair at 1.16 Å Resolution. *RNA* **1999**, *5*, 670−677.
(53) Davis, A. R.; Znosko, B. M. Thermodynamic Characterization of Naturally Occurring RNA Single Mismatches with GU Nearest Neighbors. *Biochemistry* **2008**, *47*, 10178−10187.


DOI: 10.1021/acs.jcim.5b00571
*J. Chem. Inf. Model.* 2016, 56, 101−109

# Analysing DNA structural parameters using a mesoscopic model

Journal of Physics: Conference Series **490** (2014) 012203 doi:10.1088/1742-6596/490/1/012203

# Analysing DNA structural parameters using a mesoscopic model

**Tauanne D Amarante and Gerald Weber**

Universidade Federal de Minas Gerais - Instituto de Ciências Exatas - Departamento de Física, Av. Antônio Carlos, 6627, Belo Horizonte, Minas Gerais CEP 31270-901, Brazil, Tel 55-31-3409-5633 Fax 55-31-3409-5600

E-mail: tauamarante@gmail.com

**Abstract.** The Peyrard-Bishop model is a mesoscopic approximation to model DNA and RNA molecules. Several variants of this model exists, from 3D Hamiltonians, including torsional angles, to simpler 2D versions. Currently, we are able to parametrize the 2D variants of the model which allows us to extract important information about the molecule. For example, with this technique we were able recently to obtain the hydrogen bonds of RNA from melting temperatures, which previously were obtainable only from NMR measurements. Here, we take the 3D torsional Hamiltonian and set the angles to zero. Curiously, in doing this we do not recover the traditional 2D Hamiltonians. Instead, we obtain a different 2D Hamiltonian which now includes a base pair step distance, commonly known as rise. A detailed knowledge of the rise distance is important as it determines the overall length of the DNA molecule. This 2D Hamiltonian provides us with the exciting prospect of obtaining DNA structural parameters from melting temperatures. Our results of the rise distance at low salt concentration are in good qualitative agreement with those from several published x-ray measurements. We also found an important dependence of the rise distance with salt concentration. In contrast to our previous calculations, the elastic constants now show little dependence with salt concentrations which appears to be closer to what is seen experimentally in DNA flexibility experiments.

## 1. Introduction

Many simplified models were proposed to describe the denaturation of DNA. A specific class of such models, called mesoscopic models have the advantage that their parameters have straightforward physical interpretation and can be compared directly with experimental data. One of these models, the Peyrard-Bishop (PB) model [1], is a simple mechanical model for DNA that associates two degrees of freedom for each base pair confining the molecule to a plane. These two degrees of freedom are further reduced to a single variable by linking the strand separation to the stacking distance [1]. This results into a powerful and computationally efficient method which allows us to extract relevant DNA parameters, such as hydrogen bonds, from experimental melting temperatures [2, 3]. Therefore, it is tempting to apply such methods to 3D helicoidal forms of the Peyrard-Bishop Hamiltonians [4, 5] to obtain further structural information about DNA. However, the added degrees of freedom represented by 3D Hamiltonians are not yet tractable by current melting temperature fitting methods. Nevertheless, one possibility is to adapt the 3D torsional models to a simpler 2D format and use them with the existing methods to retrieve DNA parameters.

Here we use the angular forms of the Peyrard-Bishop Hamiltonians proposed by Barbi *et al.* [5] and reduce them to their flattened 2D version by setting all torsional angles to zero. The resulting zero-angle (ZA) 2D Hamiltonian differs from the original Peyrard-Model Hamiltonian mainly by the inclusion of a helicoidal step distance $h$. The received wisdom is that this distance stays fixed at 3.4 Å for B-DNA. However, the base step, better known as *rise* distance, is far from constant and shows important dependencies with nearest neighbour context [6, 7, 8]. While for calculation purposes adding yet another fixed parameter to the Hamiltonian would appear to be of little advantage, here we have the opportunity to advance our understanding of DNA structural parameters by taking this additional parameter into account.

## 2. Model

Our interest is the study of the structural rise distance $h$. However this distance is not present in the original formulation of the Peyrard-Bishop model [1]. The 3D helicoidal variants of this Hamiltonian do incorporate the rise distance $h$, however melting temperature fitting methods [3] can presently deal only with 2D Hamiltonians. Therefore, our approach is to start from a 3D helicoidal Hamiltonian and then simplify this to a 2D Hamiltonian which can be used with melting temperature fitting methods.

The 3D Hamiltonian, derived from the Lagrangean proposed by Barbi *et al.* [5] without the last potential that models the stacking interaction, is

$$\begin{aligned} H_{n,n-1} &= \frac{p_{r_n}^2}{4m} + \frac{p_{\phi_n}^2}{4m{r_n}^2} + D(\exp[-a(r_n - R_0)] - 1)^2 \\ &\quad + k(l_{n,n-1} - l_0)^2. \end{aligned} \tag{1}$$

Rewriting this torsional Hamiltonian [5] in the notation of the original PB model [1] and setting the angles to zero ($\phi_i = 0$ and $\theta_0 = 0$), we obtain a 2D Hamiltonian

$$H_{n,n-1} = \frac{q_n^2}{2m} + D(\exp[-a\sqrt{2}y_n] - 1)^2 + k_{\alpha\beta}\left(\sqrt{h_{\alpha\beta}^2 + \frac{1}{2}(y_n - y_{n-1})^2} - h_{\alpha\beta}\right)^2 \tag{2}$$

where we used the distance between the bases of a pair $2(r_n - R_0) = \sqrt{2}y_n$, and the distance between neighbouring base pair

$$l_{n,n-1} = \sqrt{h_{\alpha\beta}^2 + (r_n - r_{n-1})^2}. \tag{3}$$

The distance between the planes of subsequent base pairs $h_{\alpha\beta}$ is now no longer a fixed value for the whole molecule but has become context dependent. For the remainder of this work we will call this new Hamiltonian the zero-angle (ZA) model.

With the Hamiltonian being now dependent on a single variable $y$, we are able to apply the framework of parameter minimization developed previously [3]. To obtain the values for rise distance $h_{\alpha\beta}$ we optimize these parameters by comparing the predicted melting temperature with the experimental data of DNA melting [9]. Even considering canonical base pairs only, it was necessary to vary ten elastic force constants $k$, ten step rise $h$ and two pairs of Morse potential parameters $D$ in a total of twenty-four parameters. Due to the large number of parameters the minimization was carried out in two steps. First we used a fixed rise distance $h = 3.4$ Å, and minimized all remaining parameters. In the second round of minimization we relaxed the constraint on the rise distance $h$, obtaining thus the final nearest-neighbour dependent values of rise distances.

## 3. Results

Figure 1 shows how the parameter rise depends on salt concentration for the DNA sequence CGATCGATCG according the ZA model. The influence of salt concentration is more pronounced for ApT steps, which are less hydrogen bonded that the other types of nearest- neighours.

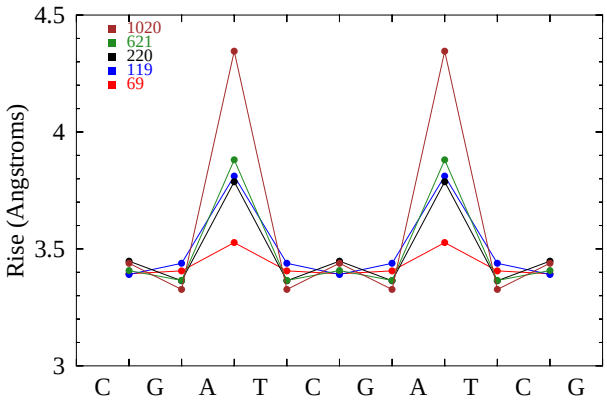



**Figure 1.** Rise distance $h$ profile for the DNA sequence CGATCGATCG as function for all salt concentrations reported in [9].

The comparison of our rise distances $h$, obtained from the modified Peyrard-Bishop ZA Hamiltonian, with experimentally determined values is not straightforward. The DNA structural parameters obtained from experimental techniques, such as X-ray diffraction and NMR, require extensive additional theoretical and algorithmic modelling. Figure 2 shows the comparison of rise distances obtained from our ZA model and various models for x-ray diffraction: CEHS [6], NEWHELIX, 3DNA [7, 8]. These models used experimental data of X-ray images [10] to estimate the step parameters and arrive at seemingly conflicting results. According to Lu and Olson [11], this inconsistence is mainly due to different ways of describing the reference frame of the DNA helical structure. We noticed that although the qualitative behaviour of our ZA model is similar to CEHS and NH, quantitatively it is closer to the results of the newer 3DNA model [8].

We also compared calculated rise distances to experimental data of NMR measurements analysed by 3DNA for the sequence GCGCATGCTACGCG [12], shown in Fig. 3. In contrast to the X-ray data of Fig. 2, the NMR show a marked tendency towards smaller rise steps which are not followed by the results of the ZA model.

## 4. Conclusion

We presented a 2D mesoscopic model as new way to calculate the structural step parameters of a DNA sequence. The main difference here is the origin of the experimental data which are melting temperatures instead of X-ray diffraction or nuclear magnetic resonance. Comparison of our results with existing experimental measurements (X-ray, NMR) point toward the need to a consistent approach in the interpretation of the reference frame of the DNA helical structure. Further work is in progress for a complete comparison of our results and existing results in the Protein Database (PDB) as well as for the calculation of rise distances for RNA.

## Acknowledgments

This work was supported by Fundação de Amparo a Pesquisa do Estado de Minas Gerais (Fapemig); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); National Institute of Science and Technology for Complex Systems.

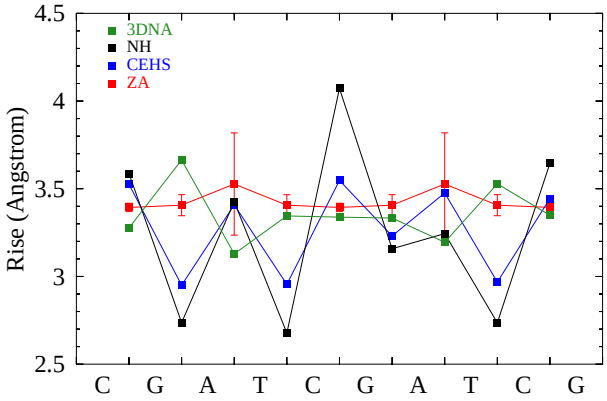



**Figure 2.** Rise distance $h$ profile comparing the ZA model with various models for x-ray diffraction (CEHS, NEWHELIX and 3DNA).

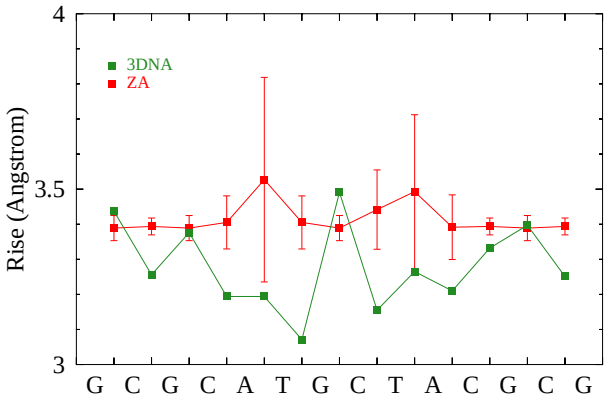



**Figure 3.** Comparison of the modified Peyrard-Bishop ZA model and 3DNA models for the DNA sequence GCGCATGCTACGCG. These 3DNA results are based on NMR experimental data.

## References

[1] Peyrard M and Bishop A R 1989 *Phys. Rev. Lett.* **62** 2755–2757
[2] Weber G, Haslam N, Whiteford N, Prügel-Bennett A, Essex J W and Neylon C 2006 *Nature Physics* **2** 55–59
[3] Weber G, Essex J W and Neylon C 2009 *Nature Physics* **5** 769–773
[4] Cocco S and Monasson R 1999 *Phys. Rev. Lett.* **83** 5178–81
[5] Barbi M, Cocco S and Peyrard M 1999 *Phys. Lett. A* **253** 358–369
[6] El Hassan M and Calladine C 1995 *Journal of Molecular Biology* **251** 648–664
[7] Lu X J and Olson W K 2003 *Nucleic Acids Research* **31** 5108–5121
[8] Lu X J and Olson W K 2008 *Nature Protocols* **3** 1213–1227
[9] Owczarzy R, You Y, Moreira B G, Manthey J A, Huang L, Behlke M A and Walder J A 2004 *Biochem.* **43** 3537–3554
[10] Grzeskowiak K, Yanagi K, Prive G and Dickerson R 1991 *Journal of Biological Chemistry* **266** 8861–8883
[11] Lu X J and Olson W K 1999 *Journal of Molecular Biology* **285** 1563–1575
[12] Ghosh A, Kar R K, Jana J, Biswas A, Ghosh S, Kumar D, Chatterjee S and Bhunia A 2013 Structural insights of DNA duplex stabilization by potent antimicrobial peptide indolicidin rSCB Protein Databank NDB ID:2M2C URL `http://ndbserver.rutgers.edu/service/ndb/atlas/summary?searchTarget=2M2C`