Unioeste - Universidade Estadual do Oeste do Paraná
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
Colegiado de Informática
Curso de Bacharelado em Informática

Programação dinâmica aplicada a problemas de *RNA - folding*

Renata Rockenbach

CASCAVEL
2009

RENATA ROCKENBACH

PROGRAMAÇÃO DINÂMICA APLICADA A PROBLEMAS DE *RNA - FOLDING*

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Informática, do Centro de Ciências Exatas e Tecnológicas da Universidade Estadual do Oeste do Paraná - Campus de Cascavel

Orientador: Prof. Reginaldo Aparecido Zara

CASCAVEL
2009

RENATA ROCKENBACH

PROGRAMAÇÃO DINÂMICA APLICADA A PROBLEMAS DE *RNA - FOLDING*

Monografia apresentada como requisito parcial para obtenção do Título de Bacharel em Informática, pela Universidade Estadual do Oeste do Paraná, Campus de Cascavel, aprovada pela Comissão formada pelos professores:

_____

Prof. Reginaldo Aparecido Zara (Orientador)
Colegiado de Informática, UNIOESTE

_____

Prof. Adriana Postal
Colegiado de Informática, UNIOESTE

_____

Prof. André Luiz Brun
Colegiado de Informática, UNIOESTE

Cascavel, 10 de dezembro de 2009

DEDICATÓRIA

Dedico este trabalho de conclusão de curso a meus pais, João e Edi, que sempre me apoiaram e incentivaram nos estudos.

“Os nossos pais amam-nos porque somos seus filhos, é um fato inalterável. Nos momentos de sucesso, isso pode parecer irrelevante, mas nas ocasiões de fracasso, oferecem um consolo e uma segurança que não se encontram em qualquer outro lugar.”

(Bertrand Russell)

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, por ter me concedido a oportunidade de estudar e concluir minha graduação. Agradeço a meus pais, João e Edi, e a minhas irmãs, Rejane e Ana Paula, pelo amor que sempre me dedicaram. Agradeço ao meu namorado, Yuri, pelo carinho e ajuda durante os anos de faculdade. Agradeço ao meu orientador, professor Reginaldo, pelas lições que me ensinou, sempre com paciência e dedicação. Agradeço também aos demais professores do Colegiado de Informática, pelos conhecimentos repassados a mim durante os anos de graduação. Agradeço aos meus colegas da faculdade: Osmar, Fabio, Gustavo, Danielly, Eliana e Pétterson, pela amizade e ajuda nos trabalhos acadêmicos. Por fim, agradeço à Fundação Araucária e ao PIBIC/Unioeste/PRPPG, pelas bolsas de estudo que recebi desenvolvendo projetos de pesquisa.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

DNA	*DeoxyriboNucleic Acid* ou Ácido Desoxi-ribonucléico
RNA	*RiboNucleic Acid* ou Ácido Ribonucléico
tRNA	RNA Transportador
mRNA	RNA Mensageiro
tmRNA	RNA Transportador e Mensageiro
rRNA	RNA Ribossômico
LMA	*Loop Matching Algorithm*

Lista de Símbolos

A Adenina
C Citosina
G Guanina
T Timina
U Uracila

Sumário

Resumo

Neste trabalho são apresentados conceitos sobre biologia molecular a fim de explicar o mecanismo de dobradura de biopolímeros do tipo RNA. Sabe-se que a estrutura dos biopolímeros pode ser determinada com base em cálculos teóricos, por meio de uma função de partição construída de maneira iterativa. Sendo assim, a partir da definição das funções e estruturas de uma molécula de RNA, são descritos métodos físicos e computacionais para a predição de sua estrutura, com enfoque no algoritmo de programação dinâmica de Nussinov-Jacobson, o qual investiga a estrutura ótima da molécula fazendo uso de recursão. Após a análise e implementação desse algoritmo, adiciona-se a ele um modelo termodinâmico, a fim de analisar o comportamento da molécula de RNA como função da temperatura à qual ela está submetida. Os resultados obtidos mostram que é possível monitorar a formação de pares de bases variando a temperatura. Com a introdução do modelo termodinâmico, observa-se que o polímero pode assumir duas fases distintas e a temperatura de transição entre essas fases pode ser calculada com o algoritmo modificado.

Palavras-chave: RNA, função de partição, programação dinâmica.

Capítulo 1

Introdução

Uma molécula de RNA é um polímero linear e heterogêneo, cuja estrutura primária consiste de sequências de quatro bases nitrogenadas chamadas Adenina (A), Guanina (G), Uracila (U) e Citosina (C) [20][34]. Devido à variação de parâmetros físicos em seu ambiente, a molécula dobra-se, formando estruturas de ordem superior (secundárias e terciárias), com regiões de dupla hélice intercaladas por porções lineares. As regiões de dupla hélice, chamadas de estruturas secundárias, formam-se pelo emparelhamento de bases energicamente favorecidas (ditas complementares), que são C–G e A–U.

Modelos físicos e computacionais têm sido utilizados para investigar a formação da estrutura secundária das moléculas e favorecer o número de bases emparelhadas [40]. Neste trabalho, é analisado e implementado o algoritmo de dobradura de RNA (*RNA-folding*) de Nussinov-Jacobson, o qual fornece a estrutura ótima da sequência utilizando técnicas de programação dinâmica [1][13][14][22]. Nesse algoritmo, dada uma sequência de RNA, são testadas as diversas formas de dobradura a fim de obter a melhor conformação molecular, determinada por meio da maximização do número de bases emparelhadas.

Também é possível realizar uma análise termodinâmica das estruturas de RNA, considerando que a molécula possui uma energia livre que determina a sua capacidade de formar pares a uma dada temperatura. Sendo assim, um tratamento termodinâmico é introduzido ao algoritmo de Nussinov-Jacobson, a fim de analisar a conformação da molécula como função da temperatura. São realizados testes com o algoritmo modificado, avaliando o comportamento da energia e as transições de fase dos biopolímeros mediante a variação de temperatura.

1.1 Organização do trabalho

O Capítulo 2 deste trabalho apresenta conceitos sobre biologia molecular, destacando as características, funções e estrutura dos biopolímeros tipo RNA e justificando o estudo de tais moléculas.

O Capítulo 3 descreve o cálculo teórico para a determinação da estrutura secundária da molécula de RNA. Ele destaca as abordagens adotadas para a escolha da melhor estrutura e as restrições relevantes para a formação dos pares de bases. É mostrado como são construídas todas as possíveis configurações da molécula e como é definida a função de partição, de maneira iterativa.

O Capítulo 4 apresenta os métodos físicos e computacionais utilizados na predição da estrutura secundária do RNA, enfatizando a utilização das técnicas de programação dinâmica. É descrito todo o funcionamento do algoritmo de Nussinov-Jacobson, que fornece a estrutura ótima da cadeia.

O Capítulo 5 explica como foi introduzido o tratamento termodinâmico no algoritmo de Nussinov-Jacobson, quais foram os testes realizados com o algoritmo modificado e quais as grandezas avaliadas.

Por fim, o capítulo 6 apresenta as considerações finais deste trabalho.

Capítulo 2

Biologia molecular

Este capítulo apresenta uma introdução sobre Biologia Molecular. A seção 2.1 traz conceitos básicos sobre os biopolímeros DNA, RNA e proteínas. A seção 2.2 apresenta o Dogma Central da Biologia Molecular, descrito por Francis Crick [8][9], que explica a transferência de informações entre as diferentes classes de biopolímeros.

Uma vez que este trabalho investiga moléculas de RNA, a seção 2.3 aborda justificativas para o estudo desse tipo de polímero, enquanto a seção 2.4 explica sua estrutura. Por fim, a seção 2.5 ilustra as diferentes formas de representação das moléculas de RNA.

2.1 Conceitos básicos

Um polímero linear é uma macromolécula constituída por uma repetição sequencial de unidades estruturais menores, denominadas monômeros. Os monômeros podem ser quimicamente iguais ou diferentes. Quando os monômeros são diferentes, a molécula é chamada de heteropolímero e, quando são iguais, de homopolímero. Biopolímeros como as proteínas, o DNA e o RNA são heteropolímeros. Nas cadeias de proteína os monômeros são chamados de aminoácidos, e podem ser de 20 tipos diferentes. As cadeias de DNA e RNA também são formadas por monômeros de tipos diferentes, chamados nucleotídeos [32].

O RNA (*RiboNucleic Acid* ou Ácido Ribonucléico) é um polímero linear e heterogêneo. Cada monômero, denominado nucleotídeo, é composto por um grupo fosfato (molécula com um átomo de fósforo cercado por 4 oxigênios), uma molécula de açúcar ribose e uma base nitrogenada (anel heterocíclico de átomos de carbono e nitrogênio). As bases nitrogenadas são quimicamente classificadas como purinas (Adenina (A) e Guanina (G)) ou pirimidinas (Uracila

(U) e Citosina (C)) [20][34]. As cadeias de DNA (*DeoxyriboNucleic Acid* ou Ácido Desoxiribonucléico) possuem composição química similar, porém com um tipo diferente de açúcar, a desoxirribose, e a base Timina (T) em substituição à Uracila [20][32].

Apesar da estrutura química e espacial dos monômeros, as moléculas de biopolímeros podem ser representadas por polímeros lineares. Esta representação é útil quando investiga-se as propriedades da cadeia em uma escala de comprimento maior do que o comprimento dos seus monômeros constituintes.

Os polímeros lineares podem ser descritos como uma rede unidimensional flexível imersa em um espaço tridimensional. As ligações covalentes entre átomos de carbono consecutivos determinam o arranjo espacial do segmento limitado da cadeia, que possui caráter unidimensional. Entretanto, em virtude da variação de parâmetros físicos no ambiente, tais como a temperatura, a concentração de sais, entre outros, o polímero pode dobrar-se e provocar o emparelhamento de bases nitrogenadas energicamente favorecidas [32].

As bases que podem ser emparelhadas são chamadas de complementares. Elas são ligadas por meio de pontes de hidrogênio e encaixam-se perfeitamente, contribuindo para a estabilidade da estrutura. O pareamento de bases se dá de maneira padronizada, sendo que as pontes de hidrogênio são estabelecidas entre uma base purina e uma pirimidina, desde que sejam complementares [11].

Na molécula de RNA, os pares possíveis são C–G e A–U. O par G–U, apesar de menos estável, também é admitido. Na molécula de DNA, as bases complementares são C–G e A–T [20][34]. A citosina forma três pontes de hidrogênio com a guanina enquanto a adenina forma duas pontes com a timina [11]. Devido ao arranjo espacial das bases complementares emparelhadas, a região emparelhada assume a forma de dupla hélice, como ilustrado na Figura 2.1a.

Embora apresentem composição química similar, o RNA e o DNA possuem estruturas muito diferentes. Na estrutura de dupla hélice natural do DNA, existem duas cadeias longas perfeitamente complementares na sequência, com uma estrutura secundária regular e simples. Já o RNA geralmente ocorre como uma única cadeia, geralmente menor do que o DNA e com uma enorme variedade de estruturas secundárias, onde os pares de bases formados são intramoleculares [11][20].

A estrutura do DNA foi descoberta em 1953 por Francis Crick e James Watson na Universidade de Cambridge, na Inglaterra, por isso refere-se às bases emparelhadas como *hélices do tipo Watson-Crick* [32][34]. A Figura 2.1 apresenta porções de uma molécula de RNA e de DNA, a fim de ilustrar a estrutura e o emparelhamento de bases complementares dos dois tipos de moléculas.

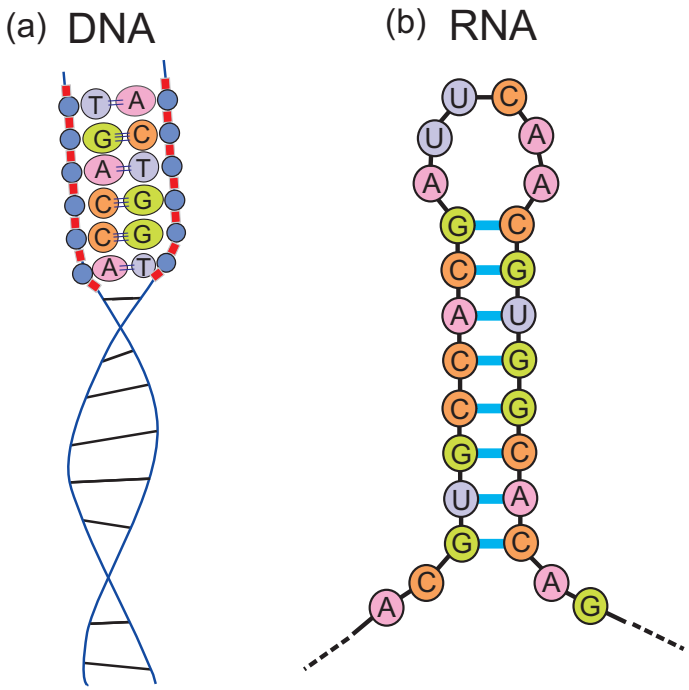

Figura 2.1: (a) Emparelhamento de bases em uma molécula de DNA (b) Emparelhamento de bases em uma molécula de RNA

2.2 O Dogma Central da Biologia Molecular

As células presentes nos organismos constituem a unidade fundamental da vida. As proteínas e os ácidos nucléicos DNA e RNA desempenham um papel importante dentro das células. O DNA e o RNA armazenam, em forma de código, toda informação de que uma célula necessita durante a sua vida e a de seus descendentes. As proteínas utilizam essa informação traduzida para permitir que a célula execute todo o trabalho necessário à sobrevivência do organismo [11].

A Figura 2.2 ilustra o Dogma Central da Biologia Molecular, enunciado por Francis Crick, em 1958 [8] e ampliado por ele em 1970 [9]. Esse dogma é um conceito que descreve a transferência de informação entre as três principais classes de biopolímeros: DNA, RNA e proteínas

[30]. O principal problema é formular regras gerais para a transferência de informação de um polímero com um alfabeto definido para outro [8].

As transferências diretas de informação estão agrupadas em três categorias, descritas a seguir [9][30].

- **Transferências gerais:** ocorrem, normalmente, nas células. Na Figura 2.2, estão representadas por setas contínuas. São as transferências de informação *DNA→DNA*, *DNA→RNA*, *RNA→Proteína*.

- **Transferências especiais:** ocorrem apenas sob condições anormais. Na Figura 2.2, estão representadas por setas pontilhadas. São as transferências de informação *RNA→RNA*, *RNA→DNA*, *DNA→Proteína*.

- **Transferências desconhecidas:** acredita-se que nunca ocorrem e, portanto, são omitidas da Figura 2.2. Referem-se às transferências de informação *Proteína→Proteína*, *Proteína→DNA* e *Proteína→RNA*.

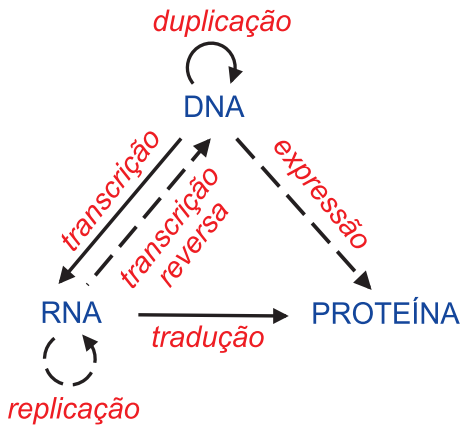

Figura 2.2: Representação geral da transferência de informação dentro do Dogma Central da Biologia Molecular [8].

O DNA contém o código essencial necessário para a síntese de proteínas e aminoácidos. A síntese de proteínas é feita por meio da transferência geral de informação do DNA para o RNA, conhecida como *transcrição*, e da transferência geral de informação do RNA para a proteína, conhecida como *tradução* [30].

A fase de transcrição resulta na síntese do RNA. Nessa fase, a molécula de DNA abre-se e a enzima RNA polimerase se liga a uma extremidade de uma das cadeias do DNA. Uma porção dessa cadeia é transcrita em uma cadeia de RNA, que é a cópia complementar do DNA. Essa cadeia simples de RNA é comumente chamada de RNA Mensageiro (mRNA). Ao final do processo, o RNA se desprende do DNA, que volta a se fechar [11][30].

A fase de tradução resulta na síntese de proteína. Nessa fase, o mRNA encontra-se com uma enzima ribozima, no citoplasma da célula, e procede lendo cada conjunto de três nucleotídeos (um códon) para convertê-los em aminoácidos. Estes aminoácidos são montados juntos utilizando as informações do mRNA como modelo, para formar uma proteína específica usada no organismo celular [30].

O resultado final da síntese de proteína não é possível sem o auxílio de formas variadas de RNA, que realizam os processos de transcrição e tradução. Além do mRNA, existem os RNAs não codificados, que são estruturas de RNA que não são traduzidas em proteínas específicas. Entre eles, estão o RNA Transportador (tRNA) e o RNA Ribossômico (rRNA), ambos envolvidos no processo de tradução [30].

As informações sobre as características genéticas de um ser vivo estão codificadas na sequência de nucleotídeos do DNA. Sendo assim, quando um progenitor transmite suas características genéticas aos seus descendentes, ele precisa fornecer a eles uma cópia do seu DNA. A formação de cópias do DNA é feita por meio do processo de *duplicação* do DNA.

A duplicação consiste na formação de duas cadeias-filhas de DNA a partir da cadeia progenitora, permitindo a transmissão das características hereditárias, bem como a sua conservação. A molécula de DNA abre sob a ação da enzima DNA polimerase. Cada uma dessas cadeias serve de molde para a formação de uma cadeia complementar [31].

Deste modo, a duplicação é um processo semiconservativo pois, de uma molécula de DNA, formam-se duas outras iguais a ela, sendo que cada DNA recém-formado possui uma das cadeias da molécula mãe. O resultado do processo de duplicação é a formação de duas moléculas de DNA idênticas entre si e idênticas à molécula original [31].

Além da tradução, transcrição e duplicação, existem os processos de transcrição reversa, replicação e expressão. Na fase de *transcrição reversa*, o DNA pode ser sintetizado utilizando-se o RNA como molde, com o auxílio da enzima transcriptase reversa. Na fase de *replicação*, o

RNA é “copiado” em RNA, sob a ação da enzima replicase. Na fase de *expressão*, a informação contida no DNA é transformada diretamente em proteína [18].

2.3 Por que estudar o RNA?

As proteínas atuam como catalisadores bioquímicos. O DNA atua no armazenamento de informação genética. Como descrito anteriormente, o RNA também possui funções específicas, associadas à transcrição do DNA e à síntese de proteínas. De acordo com Higgs [20], existem basicamente três tipos de RNA (transportador, mensageiro e ribossômico), que diferem uns dos outros em termos de tamanho, função e estrutura.

- **RNA Mensageiro (mRNA):** uma molécula de mRNA é uma cópia de uma das cadeias do DNA, geralmente formada por milhares de monômeros. O RNA mensageiro (mRNA) é responsável pela transcrição da informação genética contida no DNA. Há, ainda, o tmRNA, um tipo de RNA que apresenta características tanto do tRNA quanto o mRNA.

- **RNA Transportador (tRNA):** é uma molécula pequena, que possui aproximadamente o mesmo tamanho de uma proteína (cerca de 350 aminoácidos). Ela assume uma forma de trevo muito bem definida. Em uma das extremidades, liga-se um aminoácido específico. As três bases do meio da molécula constituem o anticódon, que emparelha-se com as bases complementares do RNA mensageiro (mRNA), o códon. Existe pelo menos um tipo específico de molécula de tRNA para cada um dos 20 aminoácidos comumente encontrados nas proteínas. A principal função do tRNA é conduzir os aminoácidos até os ribossomos, onde ocorrerá a síntese de proteínas.

- **RNA Ribossômico (rRNA):** os ribossomos são partículas responsáveis pela síntese de proteínas. As moléculas de RNA ribossômico (rRNA) são responsáveis por parte da atividade catalítica do ribossomo, além de auxiliar no acoplamento do mRNA e do tRNA e nos modelos estruturais das proteínas sintetizadas.

O estudo do RNA traz muitos benefícios importantes [30]. Além das funções essenciais desempenhadas pelo mRNA, tRNA e rRNA, existem os microRNAs, que desempenham funções

reguladoras. A análise da dobradura de uma molécula de RNA pode estar envolvida na regulagem da expressão do gene, principalmente porque os genomas para muitos vírus são codificados em RNA.

A predição da estrutura do RNA influencia e auxilia na definição da função exercida pela molécula. A partir da conformação do polímero de RNA, pode-se entender os mecanismos das reações biológicas e, assim, desenvolver novas drogas que bloqueiem processos biológicos indesejáveis [7].

Além disso, a predição correta da estrutura do RNA pode fornecer pistas para a cura e classificação de doenças, incluindo muitas que são baseadas em vírus RNA. Essa é uma importante área de estudo, uma vez que os genomas do vírus RNA codificam não só as proteínas necessárias para o seu material genético, mas também as proteínas necessárias para a reprodução do vírus [30].

2.4 Estrutura do RNA

Os biopolímeros são caracterizados por uma hierarquia de estruturas: primária, secundária, terciária e, algumas vezes, quaternária (para as proteínas). No caso do RNA, a estrutura *primária* é dada pela sequência de nucleotídeos que contém todas as informações para a reconstrução da fórmula química da molécula [32].

A alteração das condições do ambiente onde está a molécula (concentração de sais, temperatura, entre outras) faz com que o polímero dobre-se, formando uma estrutura chamada *secundária*. Esta estrutura é constituída por porções lineares intercaladas com regiões de dupla hélice, resultantes do emparelhamento das bases complementares. Por fim, a conformação molecular evolui para uma estrutura *terciária*, biologicamente ativa (ou nativa) [20][32]. Uma estrutura é biologicamente ativa quando exerce uma função específica sobre um determinado ser vivo, como é o caso do RNA em sua forma terciária.

As moléculas de RNA têm potencial para formar estruturas helicoidais em qualquer lugar onde haja duas partes de sequência que sejam complementares. Na estrutura secundária, entre as regiões de dupla hélice aparecem porções de cadeias simples, com bases desemparelhadas. Pares isolados de bases geralmente são instáveis, portanto as hélices normalmente consistem de pelo menos dois pares consecutivos [20].

A estrutura secundária da molécula pode ser dividida em subestruturas, apresentadas na Figura 2.3. Dependendo de sua forma, elas são chamadas de região de dupla hélice (a), laço *hairpin* (b), laço saliente (c), laços interiores (d) e laço de multi-ramificações (e).

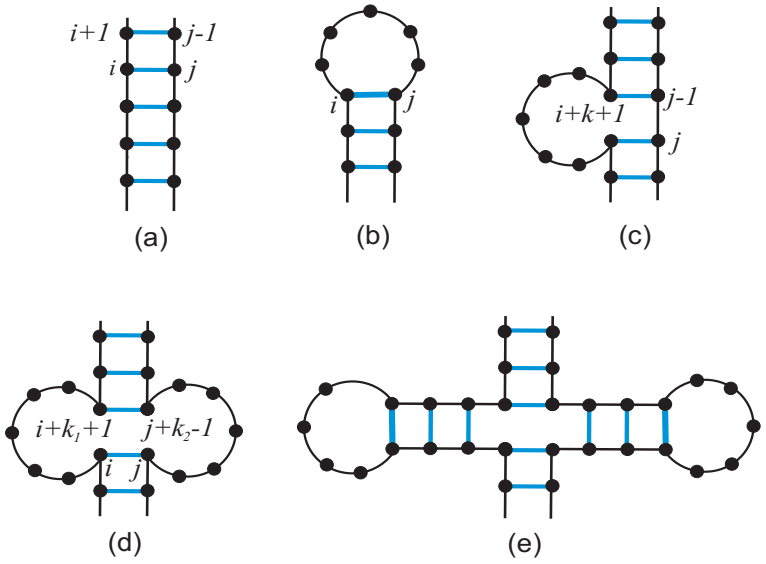

Figura 2.3: Subestruturas presentes na estrutura secundária do RNA

O chamado *hairpin* é um laço que conecta as duas extremidades de uma região de dupla hélice. O laço saliente aparece no meio de uma região de dupla hélice, em uma das cadeias, enquanto o laço interior aparece nas duas cadeias. O laço de multi-ramificações conecta três ou mais regiões de dupla hélice [20].

2.5 Representação do RNA

Observações e análises de estruturas secundárias do RNA mostram que é possível predizer a conformação terciária da molécula a partir da sua estrutura secundária completa [4][23]. Além disso, as estruturas secundárias do RNA podem ser abordadas com análise matemática, desde que a sua formação siga regras combinatórias simples [34]. Por isso, são realizadas diversas pesquisas, por meio de modelos físicos e computacionais, a fim de determinar a melhor estrutura secundária a partir da primária e, então, investigar a formação terciária. No Capítulo 4 serão descritos os diversos métodos utilizados para a predição da estrutura.

A representação convencional da estrutura secundária da molécula de RNA é um grafo pla-

nar onde os nós representam os nucleotídeos individuais e as arestas as conexões entre vizinhos da estrutura principal e das bases emparelhadas [34]. No entanto, a molécula de RNA pode ser representada graficamente de diversas maneiras, conforme descrito a seguir.

2.5.1 Representação utilizando siglas e identificando as extremidades da cadeia

Para representar uma molécula de RNA, pode-se utilizar as siglas correspondentes às bases nitrogenadas que formam a sequência e adotar uma convenção para identificar as extremidades da cadeia, isto é, a primeira e a última base (ou monômero) da cadeia. A indicação 5' refere-se ao primeiro monômero da sequência e 3' refere-se ao último monômero. A Figura 2.4 mostra uma cadeia de RNA que começa com o monômero mais à esquerda sendo precedido por 5' e termina com o monômero mais à direita, sucedido por 3' [34].

5' UUCGGCCGAUGGGCUGCCUAGCCGAGAUCCGGU 3'

Figura 2.4: Representação de uma cadeia de RNA utilizando siglas e identificando as extremidades

2.5.2 Representação gráfica bidimensional destacando as subestruturas da estrutura secundária

A Figura 2.5 reproduz a molécula de RNA apresentada na Figura 2.4, porém a mesma está dobrada e seus monômeros são representados por círculos que estão ligados, assumindo a estrutura secundária. Seguindo a convenção adotada na Figura 2.4, as extremidades da cadeia estão identificadas por 5' e 3'. No desenho, observa-se a presença de porções lineares intercaladas com bases emparelhadas e procura-se destacar essas subestruturas presentes na conformação, bem como laços *hairpin* e finais livres (bases desemparelhadas) [34].

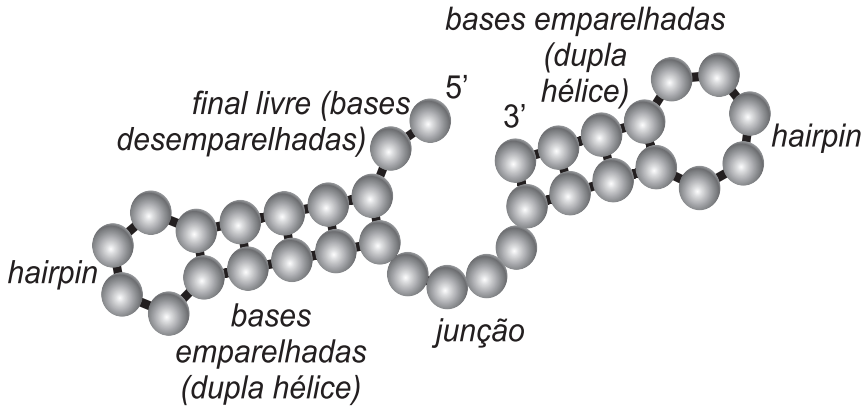

Figura 2.5: Representação gráfica bidimensional de uma cadeia de RNA destacando as subestruturas da estrutura secundária

2.5.3 Representação gráfica bidimensional utilizando siglas e identificando as extremidades da cadeia

A Figura 2.6 apresenta a mesma representação e sequência utilizada na Figura 2.5, porém, em vez de destacar as subestruturas da conformação, ela destaca as bases nitrogenadas que compõem a sequência, identificando as extremidades da cadeia [34].

Fazendo uma análise mais detalhada da Figura 2.5, observa-se a conformação assumida pela molécula, de acordo com sua dobradura. O círculo logo abaixo do 5' representa o primeiro monômero da sequência, que é a base U, a qual não está emparelhada com nenhuma outra base. A seguir, outra base U, que também está desemparelhada. As 5 bases seguintes formam uma porção da cadeia com bases emparelhadas: a primeira base da sequência *ab* (CGGCC) emparelha-se com a última base da sequência *cd* (GGCUG). Conectando as duas extremidades dessa região de dupla hélice estão as bases GAUG, na sequência *bc*. A sequência *ad* forma um laço *hairpin*.

As quatro bases (CCUA) do trecho *de* constituem uma junção entre dois laços *hairpin* (*ad* e *eh*). O segundo laço *hairpin* tem início com a região de dupla hélice, onde a primeira base da sequência *ef* (GCCG) emparelha-se com a última base da sequência *gh* (CGGU). As bases AGAUC, no trecho *fg*, conectam as duas extremidades desta região de dupla hélice.

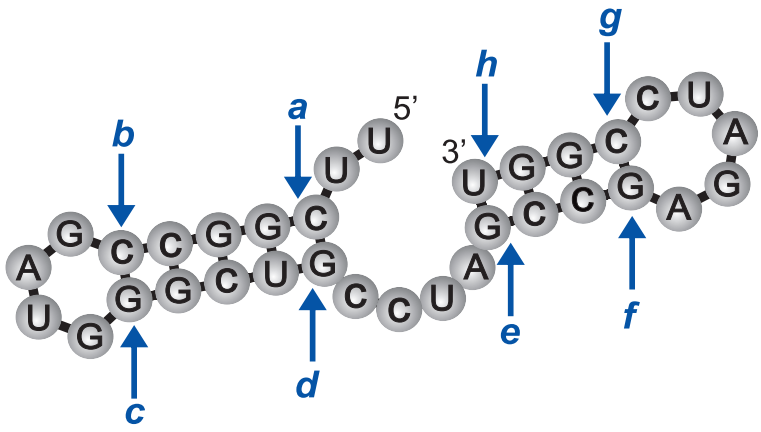

Figura 2.6: Representação gráfica bidimensional de uma cadeia de RNA utilizando siglas e identificando as extremidades da cadeia

2.5.4 Representação simbólica

A Figura 2.7 mostra uma representação simbólica da cadeia de RNA utilizada nas figuras anteriores. O ponto final . representa uma base desemparelhada. O parênteses aberto (representa uma base emparelhada com outra que está em alguma posição depois dela na cadeia, na direção 5'→3'. O parênteseses fechado) representa uma base emparelhada com outra que está em alguma posição antes dela na cadeia, na direção oposta, 3'→5' [34].

Transformando a representação apresentada na Figura 2.4 em uma representação simbólica, a Figura 2.7 apresenta o início da cadeia com dois pontos consecutivos, correspondentes às duas primeiras bases desemparelhadas (UU). Em seguida, a sequência *ab* com as 5 bases (CGGCC) que formam par com a sequência *dc*, representadas por (. Depois, as bases soltas (GAUG) representadas por pontos, seguidas das bases (GGCUG) que complementam a região de dupla hélice (trecho *cd*), representadas por).

Os quatro pontos consecutivos seguintes representam as bases (CCUA), presentes na sequência *de*, que fazem a junção entre os dois laços *hairpin* (*ad* e *eh*). Por fim, no trecho *ef*, as bases que iniciam a região de dupla hélice (GCCG), representadas por (; as bases soltas (AGAUC) (trecho *fg*), representadas por pontos consecutivos; e as bases (CGGU) complementares da região de dupla hélice, na sequência *hg*.

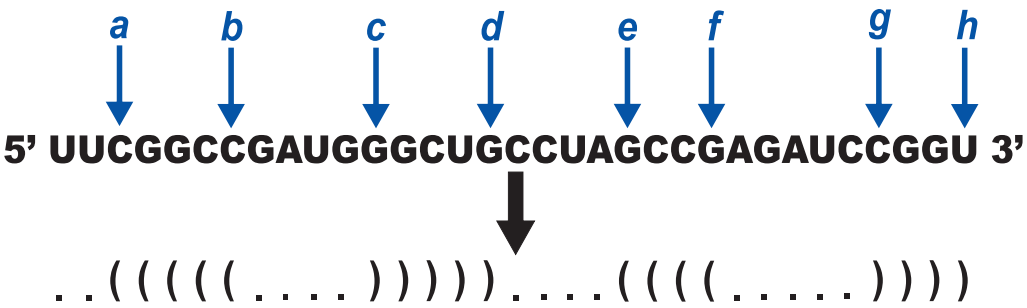

Figura 2.7: Representação simbólica de uma cadeia de RNA

2.5.5 Representação utilizando arcos e siglas

É possível a representação de uma cadeia de RNA com o uso de arcos, a qual será muito usada neste trabalho, pois permite identificar claramente as bases emparelhadas. Utiliza-se as siglas correspondentes à sequência de bases e une-se os pares emparelhados por meio de arcos, como mostra a Figura 2.8.

Figura 2.8: Representação de uma cadeia de RNA utilizando arcos e siglas

2.5.6 Representação utilizando arcos e símbolos

A Figura 2.9 apresenta a representação com arcos, tal como a Figura 2.8, porém, em vez das siglas, utiliza-se os símbolos introduzidos na representação simbólica (Figura 2.7).

Figura 2.9: Representação de uma cadeia de RNA utilizando arcos e símbolos

Capítulo 3

Determinação da Estrutura Secundária de Biopolímeros

Este capítulo descreve o cálculo teórico para a determinação da estrutura secundária de uma molécula de RNA com n monômeros. A seção 3.1 apresenta as abordagens adotadas para a determinação da melhor estrutura. A seção 3.2 considera algumas restrições relevantes na formação dos pares de bases. Na seção 3.3, é mostrado como, sob certas condições, pode-se construir todas as possíveis configurações de uma molécula de RNA. Isso é feito por meio de uma função de partição construída de maneira iterativa, partindo de um monômero, acrescentando um novo monômero à cadeia a cada passo da iteração e avaliando as novas configurações possíveis resultantes da inserção desse novo monômero.

3.1 Abordagens utilizadas

Um dos objetivos da análise da estrutura secundária do RNA é descrever as propriedades biológicas de uma população de moléculas em uma solução. Para isso, é preciso definir as propriedades estruturais predominantes nesta população, as quais são mais suscetíveis a influenciar as reações químicas em que entram as moléculas [6].

A determinação da estrutura secundária do RNA tem sido realizada considerando diversas abordagens, entre as quais destacam-se a maximização do número de pares de bases e a minimização de energia.

- **Maximizar o número de pares de bases:** uma vez que o emparelhamento de bases contribui para a estabilidade da estrutura, é conveniente a existência de muitos pares,

para que a conformação torne-se mais estável. Desta forma, esta abordagem busca uma estrutura ótima que maximize o número de pares de bases complementares na sequência de RNA.

- **Minimizar a energia livre da estrutura:** a configuração assumida pela molécula depende da temperatura a qual ela está submetida, bem como da sua energia interna. Em geral, a energia interna é obtida a partir da energia proveniente da formação de bases emparelhadas e da energia de formação de subestruturas, como laços e *hairpins*.

  Um sistema físico pode ser descrito por uma função termodinâmica conhecida como energia livre de Gibbs [21]. O estado estável do sistema é aquele para o qual a energia livre de Gibbs é mínima. Desta forma a configuração estável de uma dada molécula de RNA a uma temperatura T é aquela para a qual a energia livre de Gibbs é mínima.

  Um dos métodos de predição da estrutura secundária de uma sequência de RNA é construir, a partir de subestruturas, uma configuração para a qual a energia livre de Gibbs seja a menor possível.

Essas duas são as abordagens mais utilizadas, porém elas apresentam dois problemas básicos. Em primeiro lugar, o espaço de configuração do RNA sobre o qual a busca é executada é extremamente grande e, até pouco tempo, nenhum método sistemático de busca por todo o espaço tinha sido proposto. O segundo problema é a atribuição de energias livres adequadas aos vários componentes subestruturais [43].

Na solução, uma molécula de RNA pode dobrar-se de diversas maneiras diferentes, resultando em um número enorme de estruturas secundárias possíveis. Esse número aumenta conforme o comprimento do polímero, alcançando proporções astronômicas para moléculas de tamanho moderado [47].

Embora o RNA seja um heteropolímero, pode-se, sem perda de generalidade, descrever a construção de um homopolímero e, depois, introduzir a heterogeneidade das bases. Desta forma, a princípio, considera-se que todos os emparelhamentos possíveis (GC/CG, AU/UA e GU/UG) fornecem a mesma energia à cadeia. Sendo assim, a energia interna da molécula é igual ao somatório da energia proveniente de cada emparelhamento (chamada energia de contato), independente de quais bases estejam emparelhadas.

3.2 Restrições para a formação de emparelhamentos

Nesta seção são descritos alguns critérios para a construção de estruturas secundárias do RNA. Para isso, busca-se construir estruturas de forma iterativa, partindo de um monômero e acrescentando um novo monômero à cadeia a cada passo da iteração. Essa construção inicia-se estabelecendo critérios e restrições para a formação dos pares de bases.

1. **Contatos são possíveis apenas entre monômeros que não são primeiros vizinhos:** a Figura 3.1 mostra que um monômero não pode emparelhar-se com seu vizinho mais próximo na sequência. O arco com linha pontilhada representa um emparelhamento permitido e o arco com linha contínua representa um contato proibido.

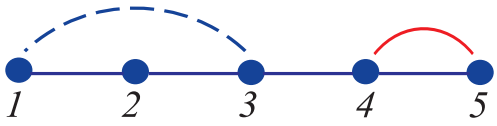

Figura 3.1: Emparelhamento permitido (arco pontilhado) e proibido (arco contínuo) na configuração da molécula de RNA

2. **Não são permitidos pseudo-nós na configuração:** um pseudo-nó do RNA consiste em um par de bases não aninhadas entre um laço de uma haste e resíduos fora desta haste [14], conforme mostra a Figura 3.2.

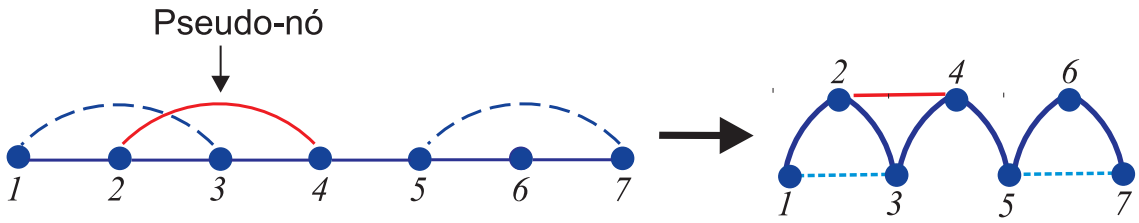

Figura 3.2: Pseudo-nó (arco com linha contínua)

3.3 Construção das possíveis configurações de uma molécula de RNA

Uma vez que não é permitido um monômero emparelhar-se com o vizinho mais próximo, a molécula precisa de pelo menos 3 monômeros para formar um emparelhamento. A construção

das possíveis configurações de uma sequência de bases é realizada por meio de um processo iterativo.

O procedimento iterativo consiste em construir todas as configurações possíveis (obedecidas as restrições já descritas) para uma sequência de n monômeros; adicionar um novo monômero para obter uma sequência de comprimento $(n+1)$ e computar o número de configurações desta nova sequência.

A Figura 3.3 ilustra o processo de construção das possíveis configurações de homopolímeros, onde n é o número de monômeros e N o número de configurações diferentes que o homopolímero pode assumir, apresentando todos os $N(n)$ de forma gráfica, para melhor compreensão. Para ilustrar que o procedimento é iterativo, as configurações da sequência com n monômeros são apresentadas na cor azul, enquanto as configurações da sequência com $(n+1)$ monômeros são apresentadas na cor vermelha.

n	*Configurações possíveis*	*N(n)*
0		0
1		1
2		1
3		2
4		4
5		8
6		17

Figura 3.3: Possíveis estruturas que um homopolímero pode assumir, conforme o número n de monômeros

- Com 0 monômeros a molécula apresenta 0 configurações possíveis (não existe molécula).
- Com 1 monômero a molécula apresenta 1 configuração possível, ou seja, somente uma base.
- Com 2 monômeros a molécula apresenta 1 configuração possível, visto que não pode ocorrer emparelhamento.
- Com 3 monômeros a molécula apresenta 2 configurações possíveis: em uma delas a molécula está estendida e existe a formação de um contato entre as bases das extremidades. Para que esse contato ocorra, a molécula deve dobrar-se.

- Com 4 monômeros a molécula apresenta 4 configurações possíveis. Uma análise da figura revela que duas destas configurações são exatamente as configurações possíveis para $n = 3$. Além disso, existem 2 configurações adicionais.

- Com 5 monômeros a molécula apresenta 8 configurações possíveis: as 4 configurações anteriores e mais 4 configurações adicionais.

- Com 6 monômeros a molécula apresenta 17 configurações possíveis: as 8 configurações anteriores e mais 9 novas configurações.

Desta forma, pode-se calcular o número de configurações possíveis para uma molécula com n monômeros, baseando-se no número de configurações para uma molécula com $(n-1)$ monômeros. Dada uma sequência com n bases, com um número de configurações possíveis $N(n)$, a adição de um novo monômero produz uma sequência cujo número de configurações é dado pela seguinte equação iterativa:

$$N(n+1) = N(n) + Na \tag{3.1}$$

onde Na é o número de configurações adicionais que surgem quando um monômero é adicionado à sequência.

Pode-se perceber que a adição de um monômero faz aumentar rapidamente o número de possíveis configurações.

Para avaliar detalhadamente a quantidade Na considere a Figura 3.4. Se o monômero 1 e $(n+1)$ forem emparelhados, podem ocorrer $N(n-1)$ configurações internas ao arco. Por exemplo, $Na = 0$ para $n = 2$ e $Na = 2$ para $n = 4$. Se o monômero k emparelhar-se com o monômero $(n+1)$, o número de configurações internas possíveis será $N(n-k)$.

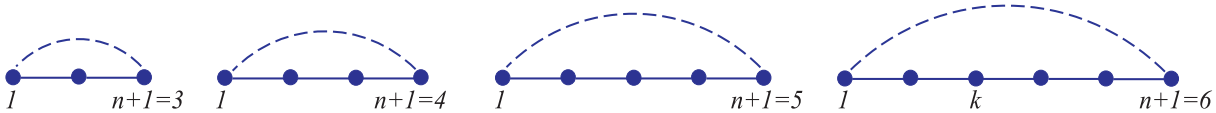

Figura 3.4: Conjunto de configurações com diferentes números de monômeros

Se $3 < k < n$ serão possíveis $N(k-1)$ configurações externas, como ilustra a Figura 3.5. Neste caso, podem existir o par $(1,3)$ e o par $(4,6)$. Se o par $(1,3)$ existe e o par $(4,6)$ não,

obtém-se uma configuração diferente daquela obtida se o par $(4, 6)$ existisse. Isso significa que é necessário testar todas as possíveis configurações de pares.

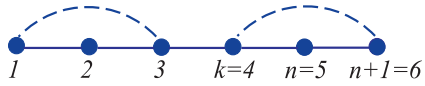

Figura 3.5: Uma das possíveis configurações externas ao par $(4, 6)$ para uma molécula com 6 monômeros

Sendo assim, o número de configurações adicionais Na é dado pelo produto:

$$Na = N(k-1)N(n-k) \tag{3.2}$$

A equação 3.2 pode ser generalizada, obtendo-se uma equação geral (3.3) para o número de configurações que a molécula pode assumir, conforme o número de monômeros n que ela possui:

$$N(n+1) = N(n) + \sum_{k=1}^{n-1} N(k-1)N(n-k) \tag{3.3}$$

A equação 3.3 fornece o número de configurações que podem ser geradas para um homopolímero de $(n+1)$ monômeros, calculadas a partir de um homopolímero de n monômeros. No entanto, ela nada diz sobre quais monômeros estão emparelhados. Ainda assim, é possível estendê-la para este caso.

Considere uma sequência $S = (S_1, S_2, S_3, ..., S_n)$ de tamanho n, apresentada na Figura 3.6. Considere uma subsequência entre as posições i e j, havendo um emparelhamento entre k e j.

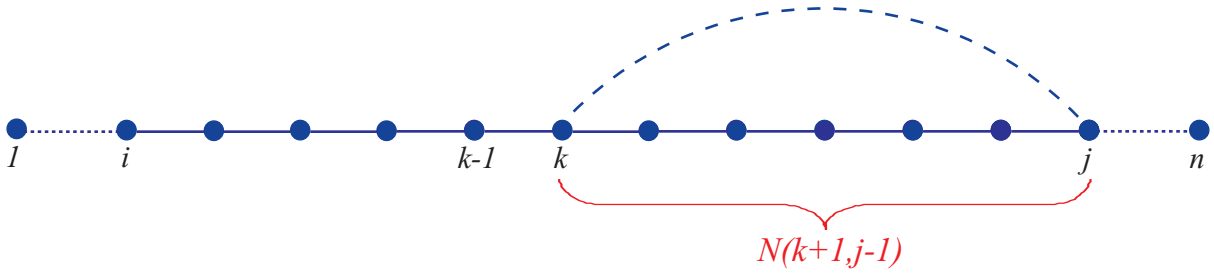

Figura 3.6: Sequência S de tamanho n

O número de estruturas secundárias possíveis para a subsequência entre i e j pode ser escrito como:

$$N(i,j) = N(i,j-1) + \sum_{k=1}^{j-1} \overbrace{N(i,k-1)}^{\text{Externas}} \overbrace{N(k+1,j-1)}^{\text{Internas}} \tag{3.4}$$

onde:

- $N(k+1, j-1)$ = Número de configurações internas entre os monômeros k e j
- $N(i, k-1)$ = Número de configurações externas entre os monômeros i e $(k-1)$

A formação de um par de bases emparelhadas se dá pela formação de ligações químicas do tipo pontes de hidrogênio. As pontes de hidrogênio são formadas por afinidade energética e, em comparação com outras ligações químicas, são extremamente fracas, podendo ser quebradas pela própria agitação térmica do sistema. A manutenção da ligação é uma competição entre a energia química e a energia térmica: se a energia térmica for muito maior que a energia das pontes de hidrogênio, a ligação pode ser quebrada e o par é desfeito.

Na Mecânica Estatística, a formação de ligações é tratada de forma probabilística. A probabilidade de formação de ligação é dada pelo chamado peso de Boltzmann, que é obtido por:

$$p = e^{\frac{-\varepsilon}{\varepsilon_T}} \tag{3.5}$$

onde ε é a energia de ligação e $\varepsilon_T = K_B T$ é a energia térmica (K_B é uma constante chamada constante de Boltzmann e T é a temperatura do sistema) [21].

Sendo $\beta = \dfrac{1}{K_B T}$, a probabilidade de formação de ligação pode ser calculada como:

$$p = e^{-\beta\varepsilon(i,j)} \tag{3.6}$$

Adotando uma energia ε para cada contato (emparelhamento) na sequência, pode-se reescrever $N(i,j)$ como a função $Z(i,j)$ da subsequência $(S_i ... S_j)$, apresentada na equação 3.7.

$$Z(i,j) = Z(i,j-1) \sum_{k=1}^{j-1} Z(i,k-1) e^{-\beta\varepsilon(i,j)} Z(k+1,j-1) \tag{3.7}$$

Em Mecânica Estatística, $Z(i,j)$ é conhecida como Função de Partição [21].

Por fim, define-se uma função de partição Z global para a molécula, apresentada na equação 3.8. Dada uma sequência com n bases, ela considera o conjunto de todas as estruturas possíveis de tamanho n, denotado por $\Omega(n)$, e a energia de cada estrutura S, denotada por $E(S)$.

$$Z(n) = \sum_{S \in \Omega(n)} e^{-\beta\varepsilon(S)} \tag{3.8}$$

Para melhor compreensão das equações definidas anteriormente, considere um exemplo prático. Seja uma molécula de tamanho $n = 15$, apresentada na Figura 3.7.

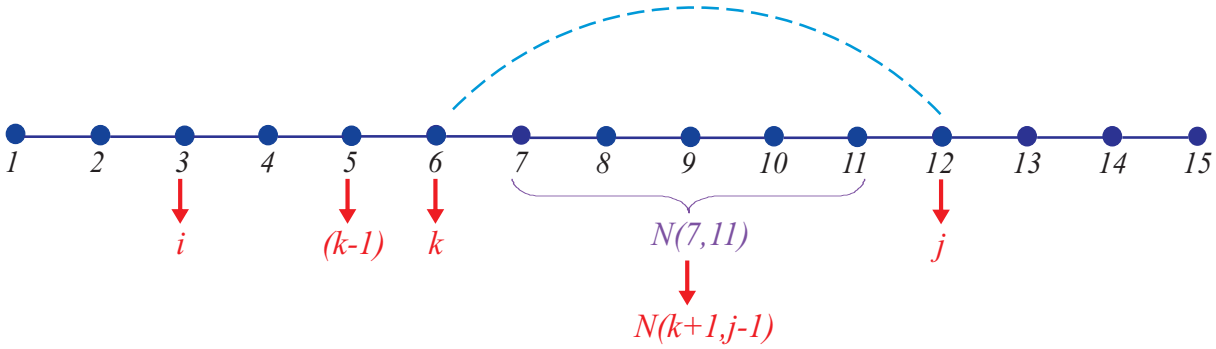

Figura 3.7: Exemplo prático com uma sequência de tamanho 15

Desta forma, sabe-se que:

- $n = 15$
- $i = 3$
- $j = 12$
- $k = 6$

Para obter o número de possíveis configurações quando um novo monômero é adicionado à cadeia, aplica-se a equação 3.3:

$$\begin{aligned} N(16) \;=\; & N(15) + [N(0)N(14)] + [N(1)N(13)] + [N(2)N(12)] + \\ & [N(3)N(11)] + [N(4)N(10)] + [N(5)N(9)] + [N(6)N(8)] + \\ & [N(7)N(7)] + [N(8)N(6)] + [N(9)N(5)] + [N(10)N(4)] + \\ & [N(11)N(3)] + [N(12)N(2)] + [N(13)N(1)] \end{aligned} \tag{3.9}$$

A equação 3.9 mostra que, para calcular $N(16)$, todas as configurações $N(i \leqslant 15)$ já devem ter sido contadas, por isso o cálculo é recursivo.

Da mesma forma, pode-se calcular todas as configurações possíveis especificamente entre os monômeros 3 e 12 (sequência de tamanho 9) aplicando-se a equação 3.4:

$$\begin{aligned} N(3,12) &= N(3,11) + [N(3,0)N(2,11)] + [N(3,1)N(3,11)] + \\ &\quad [N(3,2)N(4,11)] + [N(3,3)N(5,11)] + [N(3,4)N(6,11)] + \\ &\quad [N(3,5)N(7,11)] + [N(3,6)N(8,11)] + [N(3,7)N(9,11)] + \\ &\quad [N(3,8)N(10,11)] + [N(3,9)N(11,11)] + [N(3,10)N(12,11)] \end{aligned} \quad (3.10)$$

Neste caso, todas as configurações internas à sequência $(3,12)$ já devem ter sido computadas.

No próximo capítulo, será apresentado o algoritmo de Nussinov-Jacobson, para a implementação computacional desses cálculos.

Capítulo 4

Algoritmo para a Predição da Estrutura Secundária do RNA

Este capítulo apresenta os métodos físicos e computacionais aplicados na predição da estrutura secundária do RNA. A seção 4.1 descreve os diversos métodos utilizados para este fim. A seção 4.2 enfatiza e justifica a utilização de algoritmos de programação dinâmica na predição das estruturas, com base no cálculo teórico descrito no capítulo anterior. Por fim, a seção 4.3 explica o funcionamento do algoritmo de Nussinov-Jacobson, que será utilizado neste trabalho.

4.1 Pesquisas e experimentos

O progresso da determinação da estrutura secundária do RNA tem sido maior do que o da terciária, sobre a qual ainda há poucas informações experimentais [20]. Sabe-se que as estruturas possuem uma certa hierarquia: a primária serve como base para a construção da secundária que, por sua vez, define a terciária. Assim, a determinação da estrutura secundária é um passo necessário para a reconstrução da molécula de forma iterativa.

Observações e análises de estruturas secundárias do RNA mostram que elas servem como base para a predição da conformação terciária da molécula [4][23]. Por isso, são realizadas diversas pesquisas a fim de determinar a melhor estrutura secundária a partir da primária e, então, investigar possíveis elementos da estrutura terciária que sejam compatíveis com ela [40].

A formação da estrutura secundária do RNA pode ser investigada por meio de métodos físicos ou computacionais. Os métodos físicos são eficientes apenas para determinar um pequeno número de moléculas, uma vez que eles estão propensos a erros e são indicados somente para sequências curtas, com algumas centenas de nucleotídeos ou menos [30]. A seguir, são apresen-

tados alguns dos métodos físicos de determinação da estrutura do RNA, com suas respectivas restrições.

- **Cristalografia de raios-X:** determina a estrutura do RNA por meio da cristalização da molécula. Entretanto, alguns pesquisadores consideram o meio cristalino muito diferente das células (meio em que a molécula exerce sua função), embora haja evidências de que o meio cristalino tem muita água e algumas enzimas são ativas nele [7].

- **Ressonância Magnética Nuclear:** utiliza um campo eletromagnético para investigar os estados dos *spins* (orientação) dos vários núcleos atômicos. Essa técnica permite determinar a estrutura da molécula em uma solução, que é um meio similar à celula. A imagem da ressonância magnética, resultante de sinais de frequência de rádio, reflete o ambiente químico em que se encontra o núcleo da molécula. No entanto, existe um limite físico no tamanho das moléculas analisadas, pois o tempo de relaxação dos *spins* nucleares (taxa com a qual o núcleo volta ao seu estado de menor energia [17]) vai diminuindo com o aumento do tamanho, o que dificulta a observação e a resolução espectral [7].

Diante das limitações apresentadas pelos métodos físicos, conclui-se que os métodos computacionais são mais apropriados para a predição da estrutura do RNA. A maioria deles são baseados em algoritmos estocásticos (como a técnica da dobra cinética), métodos comparativos, recozimento simulado, algoritmos de programação dinâmica e, mais recentemente, algoritmos evolucionários [30]. Os principais métodos computacionais são:

- **Algoritmos estocásticos:** um algoritmo estocástico é um método que combina mudanças aleatórias com mudanças probabilísticas e, portanto, dificilmente repete um resultado de um experimento para outro. Para se ter uma média do seu desempenho, é necessário avaliar o resultado médio de vários experimentos [29]. O objetivo da programação estocástica é auxiliar na tomada de deciões, encontrando soluções ótimas em problemas que envolvam dados com incerteza.

  A técnica da dobra cinética é um algoritmo estocástico que simula a cinética (movimentação) de moléculas de RNA por meio de trajetórias estocásticas modeladas pela formação, dissociação e deslocamento de pares de bases individuais [30]. Essse método tem tido sucesso na predição da estrutura secundária do RNA e é abordado em [15].

- **Métodos Comparativos:** a análise comparativa de sequências busca obter uma solução ótima a partir de informações sobre sequências de RNA relacionadas filogeneticamente. Existem diferentes formas de análise comparativa de sequências para a predição da estrutura secundária do RNA. Há métodos baseados no alinhamento das sequências, na combinação das sequências e no alinhamento das estruturas.

  A análise comparativa de sequências pode ser combinada com a minimização da energia da estrutura, como é feito no software *Dynalign*, que usa análise comparativa combinada com o algoritmo de dobradura Nussinov-Jacobson [26]. Uma limitação dos métodos comparativos é que eles requerem um conjunto de entrada de duas ou mais sequências de RNA filogeneticamente relacionadas ou muito similares para chegar a uma solução.

- **Recozimento Simulado (*Simullated Annealing - SA*):** o recozimento simulado é uma técnica que simula o processo de recozimento de materiais e é utilizada na resolução de problemas de otimização de grande porte. Essa técnica evita os ótimos locais, pois admite soluções de piora, aceitando movimentos para soluções que degradam o valor da função objetivo. As soluções de piora são aceitas com uma certa probabilidade, que depende de um parâmetro chamado temperatura [38]. A probabilidade de aceitar um determinado movimento decresce com a temperatura, portanto, quanto menor a temperatura, menor a chance dos movimentos degradantes serem aceitos [12].

  São feitas buscas heurísticas SAs com a dobradura da molécula de RNA, procurando encontrar a estrutura de menor energia livre. Estudos detalhados do comportamento desse algoritmo são apresentados em [41][42]. Experimentos na predição da estrutura secundária do RNA têm mostrado que os resultados obtidos com algoritmos SAs são superiores aos obtidos com algoritmos evolucionários, utilizando o mesmo modelo termodinâmico [30].

- **Algoritmos Evolucionários (*Evolutionary Algorithms - EAs*):** a computação evolucionária consiste em uma máquina otimizada capaz de aprender, baseada nos moldes dos mecanismos de evolução biológica e seleção natural. Um algoritmo evolucionário apresenta uma estratégia de otimização estocástica, porém enfatiza o relacionamento comportamental entre progenitores e sua descendência em vez de tentar emular operadores genéticos específicos observados na natureza. Desta forma, um algoritmo evolucionário é

similar a um algoritmo estocástico, porém, em vez de implementar cruzamentos entre os indivíduos de uma população, ele confia na aptidão para sobrevivência e mutação [37]. Esses algoritmos estão entre as abordagens mais recentes para a predição da estrutura secundária do RNA. Eles buscam simular um percurso de dobra natural utilizando uma abordagem baseada na população. Destaca-se, neste contexto, o software *RnaPredict*, um algoritmo genético baseado em permutação, e o *P-RnaPredict*, ambos desenvolvidos no laboratório do Dr. Wiese [10][19]. A principal desvantagem dos algoritmos evolucionários é o tempo de execução, que é demorado em relação aos algoritmos de programação dinâmica [30].

- **Algoritmos de Programação Dinâmica (*Dynamic Programming Algorithms - DPAs*):** esses algoritmos têm sido utilizados para otimizar diversas características da estrutura secundária do RNA. O algoritmo de programação dinâmica Nussinov-Jacobson é o primeiro exemplo de otimização por meio da maximização da formação de pares de bases [28]. Outro *software* que tem se destacado é o DPA de Zuker, chamado *Mfold* [46][48][49][50], que incorpora vários parâmetros termodinâmicos a fim de refinar a precisão da predição. Nesta monografia, são discutidos aspectos relacionados ao algoritmo de Nussinov-Jacobson.

4.2 Algoritmos de programação dinâmica

De maneira geral, os algoritmos que investigam a formação da estrutura secundária do RNA simulam a dobra do polímero, procurando obter a melhor conformação. Sendo assim, com base em análises e experimentos, foram criados diferentes algoritmos de dobradura de biopolímeros, os quais buscam obter uma estrutura ótima, que maximiza o número de bases emparelhadas e minimiza a energia.

O estudo de tais algoritmos é de importância considerável, justamente para compreender como essa estrutura ótima é encontrada, ou seja, quais são os critérios para a sua escolha, e como o algoritmo trabalha para esse fim.

Existem muitas maneiras de dobrar a molécula para formar a estrutura secundária. O número de estruturas secundárias diferentes cresce exponencialmente com o tamanho da sequência, o que limita o algoritmo a processar apenas cadeias curtas. Os problemas de dobradura do

RNA são conhecidos como problemas de *RNA-folding*. Os cálculos de dobradura de RNA muitas vezes requerem um grande poder computacional, por esse motivo é necessário recorrer às técnicas de programação dinâmica, que permitem processar cadeias longas recursivamente, a partir da soma de partes menores.

As técnicas de programação dinâmica, desenvolvidas inicialmente por Richard Bellman, em 1957 [2], são usadas para obter a otimização na solução de problemas por meio da minimização ou maximização de algum parâmetro. Elas envolvem a decomposição de um problema particular em diversas partes, tratando cada parte como um problema menor a ser resolvido em um determinado estágio. A solução ótima final é obtida a partir da solução desses subproblemas. Dividir um problema grande em subproblemas faz com que a complexidade do problema inicial seja reduzida [30].

Na prática, valores de predição em sequências de RNA mostram que os programas de dobradura de RNA atuais conseguem aproximadamente entre 50 e 70% de pares de bases corretos, em média. Isso é útil para muitos propósitos, mas não engloba uma série de aplicações. Por outro lado, os algoritmos de programação dinâmica para dobradura de RNA são garantidos para fornecer a estrutura ótima matematicamente [14]. Este é mais um motivo para a utilização de programação dinâmica nos algoritmos de predição da estrutura secundária do RNA.

No processo de predição da estrutura secundária do RNA, é necessário atribuir uma energia livre para cada possível conformação da sequência proposta, e comparar as estabilidades termodinâmicas. A energia livre de uma determinada estrutura pode ser calculada utilizando parâmetros considerados em alguns modelos de predição [24][25][45].

Os modelos assumem que a energia de um RNA depende das bases emparelhadas próximas [16]. Assim, a energia livre total da molécula é estimada pela combinação de termos de energia provenientes de diversas partes da estrutura secundária, tanto das porções lineares quanto das regiões de dupla hélice [16][20].

Este trabalho apresenta a análise e implementação de um algoritmo de dobradura de RNA (*RNA-folding*) desenvolvido por Nussinov e Jacobson [1][13][22], o qual fornece a estrutura secundária ótima da cadeia. Esse algoritmo, também conhecido como *Loop Matching Algorithm* (LMA) [16], utiliza técnicas de programação dinâmica que fornecem o valor da estrutura ótima recursivamente, como uma função dos valores ótimos de subsequências menores [14]. Dada

uma sequência de RNA, são testadas as diversas formas de dobradura a fim de obter a melhor conformação molecular. A estrutura ótima é determinada por meio da maximização do número de pares de bases.

4.3 Algoritmo de Nussinov-Jacobson

De maneira simplificada, pode-se dizer que o algoritmo de Nussinov-Jacobson é a "receita" computacional para a obtenção das configurações indicadas na Figura 3.3 e representadas matematicamente pela equação 3.4. Tanto o número total quanto as configurações para uma sequência de comprimento n dependem do conhecimento das configurações da sequência $(n-1)$, uma vez que a inserção de um monômero aumenta o número total de configurações por uma quantidade Na, conforme descrito anteriormente.

O algoritmo de programação dinâmica Nussinov-Jacobson é um dos primeiros e mais simples métodos para a otimização da estrutura secundária do RNA por meio da maximização do número de pares de bases. Em geral, encontrar a estrutura com o número máximo de bases emparelhadas é a forma mais básica de predizer a estrutura ótima, uma vez que essa abordagem não leva em consideração alguns parâmetros termodinâmicos associados à estrutura secundária do RNA. A aplicação do algoritmo requer a especificação de um conjunto de regras que levam em conta parâmetros termodinâmicos e mecânicos da molécula.

O algoritmo de Nussinov-Jacobson provê uma forma eficiente para a execução de todo o procedimento, construindo as estruturas a partir de subsequências menores, fazendo uso de recursão. Em alguns casos, o algoritmo pode encontrar mais de uma estrutura para uma determinada molécula [14] e critérios físicos e biológicos devem ser utilizados na interpretação dos resultados.

Para executar a recursão de forma eficiente, é preciso ter certeza de que, ao tentar computar uma determinada sequência, todas as pontuações das subsequências menores já estão calculadas, conforme a equação 3.4. Por isso, as técnicas da programação dinâmica são úteis.

No algoritmo de Nussinov-Jacobson, a sequência é armazenada em uma matriz. Se n é o número total de bases da sequência, armazenar a matriz requer memória proporcional a n^2 [14]. Isso não é exagerado nos dias atuais; dobrar uma sequência com 1000 bases, por exemplo, requer alguns megabytes. No entanto, para encontrar possíveis bifurcações na sequência, o

algoritmo requer tempo proporcional a n^3 [14]. Desta forma, em relação à sua complexidade computacional, o algoritmo de Nussinov-Jacobson executa em um tempo $O(n^3)$ e requer espaço na memória $O(n^2)$ [16].

4.3.1 Funcionamento do algoritmo

O pseudo-código do algoritmo de Nussinov-Jacobson é apresentado no trabalho de Durbin [13]. A implementação do algoritmo já foi desenvolvida e pode ser encontrada em [1][22][35]. Com base na implementação de [35], foi possível analisar o funcionamento do algoritmo.

O algoritmo de Nussinov-Jacobson é dividido em dois estágios. O primeiro consiste no preenchimento da matriz, que equivale à construção de todas as configurações possíveis para a sequência. O segundo estágio é um procedimento de *tracebacking*, o qual busca, entre todas as configurações geradas, aquela com o maior número de bases emparelhadas, chamada de configuração ótima. Com base no trabalho de Poonian [30], esses procedimentos são descritos a seguir.

Estágio 1 - Preenchimento da matriz

O algoritmo de Nussinov-Jacobson utiliza uma matriz $\gamma(n, n)$ armazenada na memória principal para computar as estruturas geradas para uma sequência de comprimento n. A matriz é dividida em duas partes, de forma triangular, sendo as partes superior e inferior separadas por $\gamma(i, i)$. Cada elemento da matriz na parte superior representa uma única subsequência que começa com o nucleotídeo i e termina com o nucleotídeo j. Cada célula armazena um valor que indica o número máximo de pares de bases possíveis para essa subsequência. A parte inferior (onde $j < i$) é anulada e não é utilizada no procedimento de preenchimento da matriz. Esse espaço pode ser útil para posteriores modificações no algoritmo.

Conforme será visto a seguir, a matriz triangular superior será preenchida obedecendo a um conjunto de regras de formação de pares e restrições geométricas. Diferentes conjuntos de regras de emparelhamento e de restrições podem ser usados, refletindo diferentes propriedades da molécula. Neste trabalho, será utilizado o seguinte conjunto de regras e restrições:

$$\gamma(i, i) = 0 \tag{4.1}$$

$$\gamma(i, i \pm 1) = 0 \tag{4.2}$$

Segundo a equação 4.1, uma base não pode formar par consigo mesma. Segundo a equação 4.2, uma base não pode formar par com seus vizinhos imediatamente subsequentes.

A seguir, o algoritmo prevê a fase de preenchimento a matriz, segundo um conjunto de regras de emparelhamento. Essas regras baseiam-se na equação 3.4 e na Figura 3.3, e incorporam a heterogeneidade da sequência.

$$\gamma(i,j) = max \begin{cases} \gamma(i+1, j) & \text{se } i \text{ desemparelhada} & \text{(a)} \\ \gamma(i, j-1) & \text{se } j \text{ desemparelhada} & \text{(b)} \\ max_{i<k<j}\left[\gamma(i,k) + \gamma(k+1,j)\right] & \text{se bifurcação} & \text{(c)} \\ \gamma(i+1, j-1) + \delta(i,j) = \left\{ \begin{array}{ll} 1 & \text{se } (i,j) \text{ emparelhadas} \\ 0 & \text{se } (i,j) \text{ desemparelhadas} \end{array} \right\} & \text{(d)} \end{cases} \tag{4.3}$$

O algoritmo atribui um valor para cada par de bases formado e armazena a soma total dos valores como sendo a "pontuação" (ou peso) dessa estrutura. Ao final, a estrutura com maior pontuação, isto é, a que possuir o maior número de pares de bases, é a ótima.

Na formação dos emparelhamentos, a probabilidade de formar os pares GC/CG e AU/UA é a mesma, e recebe pontuação de valor 1 [30]. Nesse algoritmo, o emparelhamento GU/UG não é considerado válido, uma vez que é muito instável, podendo ser desfeito facilmente.

As pontuações das sequências possíveis são armazenadas na matriz triangular. Inicializa-se com subsequências de tamanho 1 ou 2, as quais não possuem pares de bases e, portanto recebem pontuação zero. São acrescentadas bases à subsequência inicial, obtendo subsequências cada vez maiores, até alcançar o tamanho desejado. Em seguida, recupera-se a estrutura ótima por meio de um procedimento de *tracebacking* [14][16].

Primeiramente, a matriz $\gamma(n, n)$ é inicializada com as equações 4.1 e 4.2, ou seja, três de suas diagonais recebem valor zero. As equações 4.1 e 4.2 são utilizadas para garantir que, mais tarde, durante o preenchimento da matriz, não ocorra uma referência a um elemento não preenchido.

Nos elementos da matriz são armazenadas as pontuações para cada estrutura obtida segundo a equação 4.3. A primeira diagonal preenchida contém subsequências de tamanho 3, a segunda contém subsequências de tamanho 4, e assim por diante. Conforme a progressão de cada diago-

nal, o último elemento preenchido da matriz é o do canto superior direito, $\gamma(0, n-1)$. O valor desse elemento fornece o número máximo de bases emparelhadas.

A ideia principal para o preenchimento da matriz e a definição das subsequências é que existem quatro maneiras possíveis de obter a melhor estrutura das subsequências entre as bases (i, j) [5]. Os quatro casos, ilustrados na Figura 4.1, são:

1. Adicionar uma base i desemparelhada à melhor estrutura da subsequência $[i+1, j]$.
2. Adicionar uma base j desemparelhada à melhor estrutura da subsequência $[i, j-1]$.
3. Combinar duas subestruturas ótimas $[i, k]$ e $[k+1, j]$ (bifurcação).
4. Adicionar as bases emparelhadas i e j à melhor estrutura da subsequência $[i+1, j-1]$.

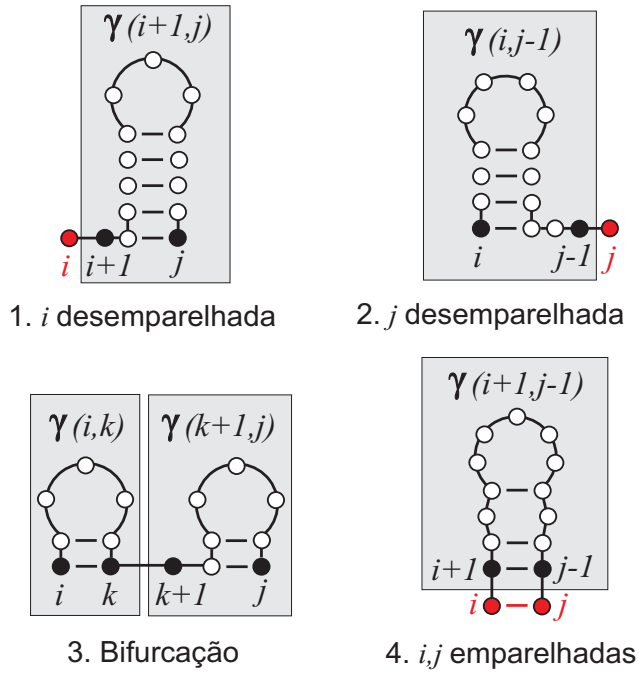

Figura 4.1: Quatro possíveis maneiras de se obter a melhor estrutura de uma subsequência.

Durante o estágio de preenchimento da matriz, é utilizada a parte superior da matriz $\gamma(n, n)$ (onde $j > i$), seguindo o conjunto de regras e restrições descrito na equação 4.3. Utilizando essa abordagem, um elemento da matriz é preenchido por vez, diagonalmente, começando do elemento do canto superior esquerdo da matriz até o elemento do canto inferior direito.

Estágio 2 - Procedimento de *Tracebacking*

O estágio de preenchimento é utilizado para construir a parte triangular superior da matriz $\gamma(n, n)$. O elemento do canto superior direito da matriz, $\gamma(0, n-1)$, contém o valor que indica o número máximo de pares de bases para a sequência. É necessário um procedimento para recuperar e reconstruir a estrutura que corresponde à configuração ótima entre todas as configurações geradas. Para esse propósito, utiliza-se um procedimento de *tracebacking*, que pode ser feito com uma implementação recursiva ou com uma implementação baseada em pilha.

Neste trabalho, utiliza-se um algoritmo de *tracebacking* baseado em pilha. Ele é inicializado inserindo na pilha as posições $i = 0$ e $j = (n-1)$ da matriz, e é finalizado quando a pilha estiver vazia ou quando $i \geqslant j$, correspondente ao *tracebacking* que atingiu a última subsequência da matriz.

Todos os valores preenchidos da matriz são analisados, começando pelo elemento do canto superior direito. A cada elemento $\gamma(i, j)$, é feita uma avaliação com base nos valores já calculados para o preenchimento, por isso o procedimento é chamado de *tracebacking*, já que consiste em um retrocesso nas posições preenchidas da matriz. A partir dessa avaliação, é inserida na pilha uma base desemparelhada i ou j, uma bifurcação ou um par de bases (i, j). Se, em algum ponto, não existir mais do que uma condição máxima plausível para algum $\gamma(i, j)$, então há a possibilidade de alternar estruturas com o número máximo de pares de bases.

Inicialmente, a pilha está vazia. Empilha-se $i = 0$ e $j = (n-1)$. A partir daí, tem início do procedimento de *tracebacking*. Enquanto a pilha não estiver vazia, o elemento do topo é desempilhado e é avaliado por meio de um conjunto de testes para definir o próximo passo. Sendo (i, j) a posição do elemento desempilhado, o procedimento continua enquanto $i < j$.

Os testes feitos com o elemento $\gamma(i, j)$ se baseiam no conjunto de regras utilizadas para o preenchimento da matriz, apresentado anteriormente, na equação 4.3. Cada elemento da matriz analisado irá se enquadrar em um dos quatro casos mostrados na Figura 4.1. Se for o quarto caso, em que as bases i e j estão emparelhadas, então o par (i, j) será armazenado em uma outra pilha, reservada apenas para os pares de bases. A seguir, são apresentados os testes feitos com cada elemento da matriz e a decisão a ser tomada em cada caso.

- **Teste 1 –** Se $\gamma(i, j) = \gamma(i+1, j)$: empilha-se $(i+1, j)$;

- **Teste 2 –** Se $\gamma(i,j) = \gamma(i,j-1)$: empilha-se $(i,j-1)$;
- **Teste 3 –** Para k de $i+1$ até $j-1$, se $\gamma(i,k)+\gamma(k+1,j) = \gamma(i,j)$: empilha-se $(k+1,j)$ e (i,k);
- **Teste 4 –** Se $\gamma(i,j) = \gamma(i+1,j-1)+\delta(i,j)$: empilha-se $(i+1,j-1)$ e coloca-se (i,j) na pilha de pares de bases.

Desta forma, a cada passo do procedimento, o elemento do topo da pilha é desempilhado e avaliado, enquanto a pilha não estiver vazia e $i < j$. Quando uma dessas duas condições não for obedecida, o processo acabará e a pilha de pares de bases irá conter os pares formados na estrutura ótima.

4.3.2 Restrições para as possíveis estruturas

A construção de estruturas secundárias de RNA obedecem a alguns critérios e restrições na formação dos pares de bases. Uma restrição é que não são consideradas interações ou modificações de bases com outras moléculas. Outra restrição estabelece que bases imediatamente vizinhas não podem ser emparelhadas. Além disso, os algoritmos de programação dinâmica discutidos não tratam os chamados pseudo-nós, os quais não são considerados um emparelhamento aceitável [14][16], pois aumentam a complexidade do algoritmo e violam a definição recursiva da pontuação ótima.

Existem algoritmos de dobradura de RNA que lidam com pseudo-nós, mas cada um deles possui pelo menos uma séria limitação. Alguns garantem soluções ótimas de acordo com o modelo termodinâmico, mas são ineficientes para a maioria das aplicações práticas. Outros são eficientes, porém não são comprovadamente ótimos [14].

4.3.3 Exemplo detalhado de aplicação do algoritmo de Nussinov-Jacobson

Para melhor compreensão do algoritmo de Nussinov-Jacobson, será apresentado um exemplo detalhado de sua aplicação. Considere uma cadeia de RNA com a sequência 5' ACAGU 3', de tamanho $n = 5$. Primeiramente, procede-se o estágio 1, correspondente ao preenchimento da matriz $\gamma(n,n)$. O primeiro passo é criar a matriz $\gamma(5,5)$, como ilustra a Figura 4.2.

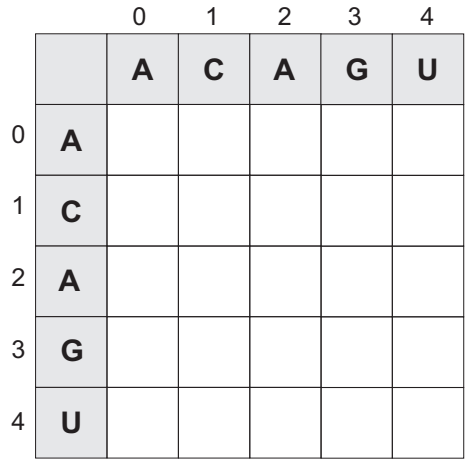

Figura 4.2: Passo 1 - Estágio de preenchimento: representação matricial $\gamma(5,5)$ da sequência de RNA

Depois disso, a matriz é inicializada, conforme as equações 4.1 e 4.2. A Figura 4.3 ilustra a matriz após esse procedimento, com três de suas diagonais zeradas. Como a parte inferior da matriz é anulada e inutilizada, suas células são pintadas de cor escura. A partir daí, os elementos da parte superior serão preenchidos, diagonalmente, iniciando pelo elemento superior mais à esquerda da matriz até o elemento inferior mais à direita, na direção indicada pelas setas.

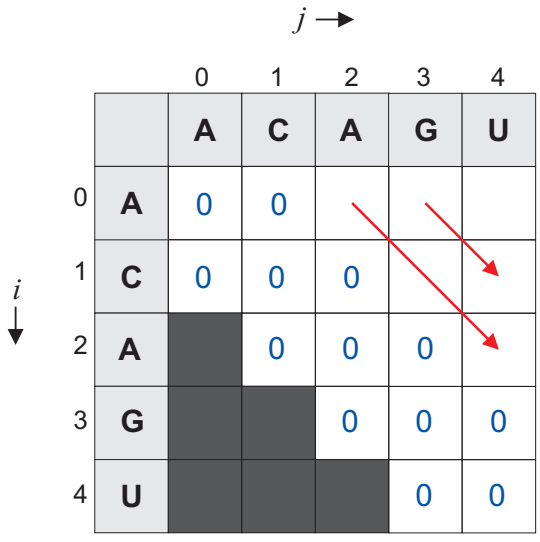

Figura 4.3: Passo 2 - Zerar três diagonais da matriz

A primeira diagonal a ser preenchida, $\gamma(0,2)...\gamma(2,4)$, corresponde a todas as subsequências de tamanho 3. A próxima diagonal $\gamma(0,3)...\gamma(1,4)$ corresponde a todas as subsequências de

tamanho 4, e assim por diante. Utilizando a direção indicada na Figura 4.3, a operação de preenchimento é aplicada para todas as subsequências de tamanho 3 a $n = 5$, aplicando as regras definidas na equação 4.3.

A seguir, é apresentada a sequência de passos realizada para o preenchimento, iniciando pela posição $\gamma(0, 2)$ da matriz e completando a diagonal até a posição $\gamma(2, 4)$. Para cada posição, considera-se os quatro possíveis casos apresentados na equação 4.3 e escolhe-se o que possuir o valor máximo de pares de bases.

- **Posição** $\gamma(0, 2)$: considera-se os quatro possíveis casos da equação 4.3 para $i = 0$, correspondente à base A, e $j = 2$, correspondente à base A.

  1. *Caso (a):* com base na equação, a posição $\gamma(0, 2)$ irá receber o valor da posição $\gamma(1, 2)$, que é 0. Assim, para o caso (a), $\gamma(0, 2) = 0$.
  2. *Caso (b):* com base na equação, a posição $\gamma(0, 2)$ irá receber o valor da posição $\gamma(0, 1)$, que é 0. Assim, para o caso (b), $\gamma(0, 2) = 0$.
  3. *Caso (c):* considerando $k = 1$ (correspondente à base U), aplicando a equação, a posição $\gamma(0, 2)$ irá receber o valor da posição $\gamma(0, 1) = 0$ somado ao valor da posição $\gamma(2, 2) = 0$, que é 0. Assim, para o caso (c), $\gamma(0, 2) = 0$.
  4. *Caso (d):* com base na equação, a posição $\gamma(0, 2)$ irá receber o valor da posição $\gamma(1, 1) = 0$, mais o valor 0, pois A e G não formam um par aceitável. Assim, para o caso (d), $\gamma(0, 2) = 0$.

  Por fim, calculando o valor máximo entre os quatro casos, a posição $\gamma(0, 2)$ recebe o valor 0.

- **Posição** $\gamma(1, 3)$: considera-se os quatro possíveis casos da equação 4.3 para $i = 1$, correspondente à base C, e $j = 3$, correspondente à base G.

  1. *Caso (a):* com base na equação, a posição $\gamma(1, 3)$ irá receber o valor da posição $\gamma(2, 3)$, que é 0. Assim, para o caso (a), $\gamma(1, 3) = 0$.
  2. *Caso (b):* com base na equação, a posição $\gamma(1, 3)$ irá receber o valor da posição $\gamma(1, 2)$, que é 0. Assim, para o caso (b), $\gamma(1, 3) = 0$.

3. *Caso (c):* considerando $k = 2$ (correspondente à base G), aplicando a equação, a posição $\gamma(1,3)$ irá receber o valor máximo da posição $\gamma(1,2) = 0$ somado ao valor da posição $\gamma(3,3) = 0$, que é 0. Assim, para o caso (c), $\gamma(1,3) = 0$.

4. *Caso (d):* com base na equação, a posição $\gamma(1,3)$ irá receber o valor da posição $\gamma(2,2) = 0$, mais o valor 1, pois U e A formam um par aceitável. Assim, para o caso (d), $\gamma(1,3) = 1$.

Por fim, calculando o valor máximo entre os quatro casos, a posição $\gamma(1,3)$ recebe o valor 1.

- **Posição** $\gamma(2,4)$: considera-se os quatro possíveis casos da equação 4.3 para $i = 2$, correspondente à base A, e $j = 4$, correspondente à base U.

1. *Caso (a):* com base na equação, a posição $\gamma(2,4)$ irá receber o valor da posição $\gamma(3,4)$, que é 0. Assim, para o caso (a), $\gamma(2,4) = 0$.

2. *Caso (b):* com base na equação, a posição $\gamma(2,4)$ irá receber o valor da posição $\gamma(2,3)$, que é 0. Assim, para o caso (b), $\gamma(2,4) = 0$.

3. *Caso (c):* considerando $k = 3$ (correspondente à base A), aplicando a equação, a posição $\gamma(2,4)$ irá receber o valor da posição $\gamma(2,3) = 0$ somado ao valor da posição $\gamma(4,4) = 0$, que é 0. Assim, para o caso (c), $\gamma(2,4) = 0$.

4. *Caso (d):* com base na equação, a posição $\gamma(2,4)$ irá receber o valor da posição $\gamma(3,3) = 0$, mais o valor 0, pois G e U não formam um par aceitável. Assim, para o caso (d), $\gamma(2,4) = 0$.

Por fim, calculando o valor máximo entre os quatro casos, a posição $\gamma(2,4)$ recebe o valor 0.

Agora, procede-se preenchendo a próxima diagonal, $\gamma(0,3)...\gamma(1,4)$.

- **Posição** $\gamma(0,3)$: considera-se os quatro possíveis casos da equação 4.3 para $i = 0$, correspondente à base A, e $j = 3$, correspondente à base G.

1. *Caso (a):* com base na equação, a posição $\gamma(0,3)$ irá receber o valor da posição $\gamma(1,3)$, que é 1. Assim, para o caso (a), $\gamma(0,3) = 1$.

2. *Caso (b):* com base na equação, a posição $\gamma(0,3)$ irá receber o valor da posição $\gamma(0,2)$, que é 0. Assim, para o caso (b), $\gamma(0,3) = 0$.

3. *Caso (c):* considerando $k = 1$ (correspondente à base U), aplicando a equação, a posição $\gamma(0,3)$ irá receber o valor entre as posições $\gamma(0,1) = 0$ somado ao valor da posição $\gamma(2,3) = 0$, que é 0. Considerando $k = 2$ (correspondente à base G), a posição $\gamma(0,3)$ irá receber o valor da posição $\gamma(0,2) = 0$ somado ao valor da posição $\gamma(3,3) = 0$, que é 0. Assim, calculando o valor máximo entre $k = 1$ e $k = 2$, para o caso (c), $\gamma(0,3) = 0$.

4. *Caso (d):* com base na equação, a posição $\gamma(0,3)$ irá receber o valor da posição $\gamma(1,2) = 0$, mais o valor 0, pois A e A não formam um par aceitável. Assim, para o caso (d), $\gamma(0,3) = 0$.

Por fim, calculando o valor máximo entre os quatro casos, a posição $\gamma(0,3)$ recebe o valor 1.

- **Posição** $\gamma(1,4)$: considera-se os quatro possíveis casos da equação 4.3 para $i = 1$, correspondente à base C, e $j = 4$, correspondente à base U.

1. *Caso (a):* com base na equação, a posição $\gamma(1,4)$ irá receber o valor da posição $\gamma(2,4)$, que é 0. Assim, para o caso (a), $\gamma(1,4) = 0$.

2. *Caso (b):* com base na equação, a posição $\gamma(1,4)$ irá receber o valor da posição $\gamma(1,3)$, que é 1. Assim, para o caso (b), $\gamma(1,4) = 1$.

3. *Caso (c):* considerando $k = 2$ (correspondente à base G), aplicando a equação, a posição $\gamma(1,4)$ irá receber o valor entre as posições $\gamma(1,2) = 0$ somado ao valor da posição $\gamma(3,4) = 0$, que é 0. Considerando $k = 3$ (correspondente à base A), a posição $\gamma(1,4)$ irá receber o valor da posição $\gamma(1,3) = 1$ somado ao valor da posição $\gamma(4,4) = 0$, que é 1. Assim, calculando o valor máximo entre $k = 2$ e $k = 3$, para o caso (c), $\gamma(1,4) = 1$.

4. *Caso (d):* com base na equação, a posição $\gamma(1,4)$ irá receber o valor da posição $\gamma(2,3) = 0$, mais o valor 0, pois U e U não formam um par aceitável. Assim, para o caso (d), $\gamma(1,4) = 0$.

Por fim, calculando o valor máximo entre os quatro casos, a posição $\gamma(1,4)$ recebe o valor 1.

Agora, procede-se preenchendo a última diagonal, correspondente ao elemento $\gamma(0,4)$ da matriz.

- **Posição** $\gamma(0,4)$: considera-se os quatro possíveis casos da equação 4.3 para a posição $i = 0$, correspondente à base A, e $j = 4$, correspondente à base U.

  1. *Caso (a):* com base na equação, a posição $\gamma(0,4)$ irá receber o valor da posição $\gamma(1,4)$, que é 1. Assim, para o caso (a), $\gamma(0,4) = 1$.
  2. *Caso (b):* com base na equação, a posição $\gamma(0,4)$ irá receber o valor da posição $\gamma(0,3)$, que é 1. Assim, para o caso (b), $\gamma(0,4) = 1$.
  3. *Caso (c):* considerando $k = 1$ (correspondente à base U), aplicando a equação, a posição $\gamma(0,4)$ irá receber o valor entre as posições $\gamma(0,1) = 0$ somado ao valor da posição $\gamma(2,4) = 0$, que é 0. Considerando $k = 2$ (correspondente á base G), a posição $\gamma(0,4)$ irá receber o valor da posição $\gamma(0,2) = 0$ somado ao valor da posição $\gamma(3,4) = 0$, que é 0. Considerando $k = 3$ (correspondente à base A), a posição $\gamma(0,4)$ irá receber o valor da posição $\gamma(0,3) = 1$ somado ao valor da posição $\gamma(4,4) = 0$, que é 1. Assim, calculando o valor máximo entre $k = 1$, $k = 2$ e $k = 3$, para o caso (c), $\gamma(0,4) = 1$.
  4. *Caso (d):* com base na equação, a posição $\gamma(0,4)$ irá receber o valor da posição $\gamma(1,3) = 1$, mais o valor 1, pois A e U formam um par aceitável. Assim, para o caso (d), $\gamma(0,4) = 2$.

  Por fim, calculando o valor máximo entre os quatro casos, a posição $\gamma(0,4)$ recebe o valor 2.

A matriz resultante após o estágio de preenchimento é ilustrada na Figura 4.4.

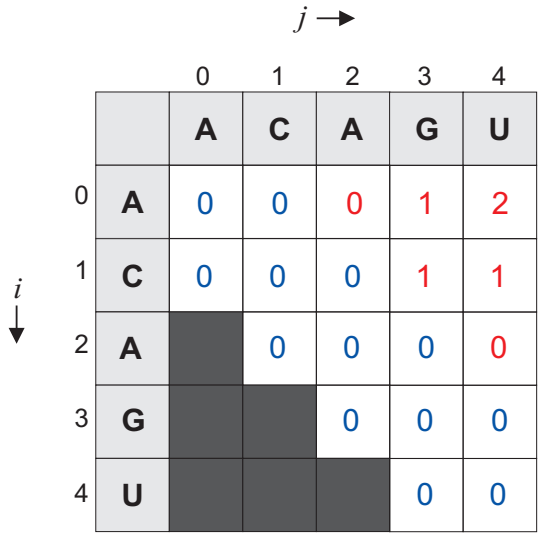

Figura 4.4: Passo 3 - Matriz resultante após o estágio de preenchimento

O valor do elemento do canto superior direito da matriz, $\gamma(0, 4)$, indica o número máximo de pares de bases para a sequência, neste caso, 2 pares.

Nesta etapa, inicia-se o estágio 2, correspondente ao procedimento de *tracebacking*, que escolhe a estrutura ótima entre todas as geradas. Conforme a Figura 4.5, a pilha, inicialmente vazia, recebe as posições $i = 0$ e $j = 4$ da matriz.

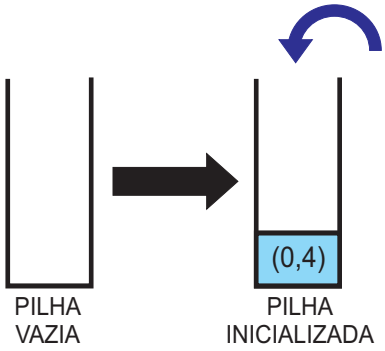

Figura 4.5: Passo 4 - Estágio de *tracebacking*: inicialização da pilha

Inicia-se o procedimento de *tracebacking*. Enquanto a pilha não estiver vazia e $i < j$, aplica-se os testes apresentados na seção 4.3.1 ao elemento do topo da pilha.

O elemento do topo da pilha é $e = (0, 4)$. Como a pilha não está vazia e $0 < 4$, $e = (0, 4)$ é desempilhado e avaliado.

- **Teste 1:** $\{\gamma(0,4) = 2\} \neq \{\gamma(1,4) = 1\}$, então $e = (0,4)$ não encaixa-se no caso 1.
- **Teste 2:** $\{\gamma(0,4) = 2\} \neq \{\gamma(0,3) = 1\}$, então $e = (0,4)$ não encaixa-se no caso 2.
- **Teste 3:** considerando $k = 1$, $\{[\gamma(0,1) = 0] + [\gamma(2,4) = 0] = 0\} \neq \{\gamma(0,4) = 2\}$. Considerando $k = 2$, $\{[\gamma(0,2) = 0] + [\gamma(3,4) = 0] = 0\} \neq \{\gamma(0,4) = 2\}$. Considerando $k = 3$, $\{[\gamma(0,3) = 1] + [\gamma(4,4) = 0] = 1\} \neq \{\gamma(0,4) = 2\}$. Então $e = (0,4)$ não encaixa-se no caso 3.
- **Teste 4:** $\{\gamma(0,4) = 2\} = \{[\gamma(1,3) = 1] + [\delta(\gamma(0,4) = 1) = 2]\}$, então $e = (0,4)$ encaixa-se no caso 4, portanto empilha-se $(1,3)$ e coloca-se $(0,4)$ na pilha de pares de bases, como ilustra a Figura 4.6.

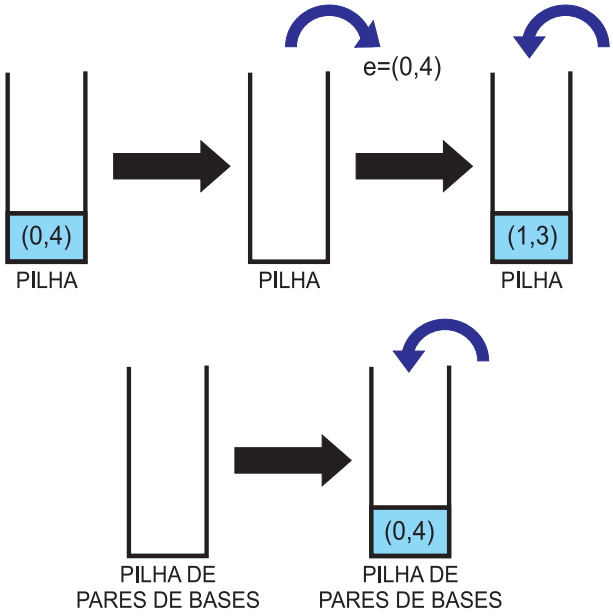

Figura 4.6: Passo 5(a) - Operações na pilha após a avaliação do elemento $(0,4)$

O elemento do topo da pilha é $e = (1,3)$. Como a pilha não está vazia e $1 < 3$, $e = (1,3)$, é desempilhado e avaliado.

- **Teste 1:** $\{\gamma(1,3) = 1\} \neq \{\gamma(2,3) = 0\}$, então $e = (1,3)$ não encaixa-se no caso 1.
- **Teste 2:** $\{\gamma(1,3) = 1\} \neq \{\gamma(1,2) = 0\}$, então $e = (1,3)$ não encaixa-se no caso 2.

- **Teste 3:** considerando $k = 2$, $\{[\gamma(1,2) = 0] + [\gamma(3,3) = 0] = 0\} \neq \{\gamma(1,3) = 1\}$, então $e = (0,4)$ não encaixa-se no caso 3.
- **Teste 4:** $\{\gamma(1,3) = 1\} = \{[\gamma(2,2) = 0] + [\delta(\gamma(1,3) = 1) = 2]\}$, então $e = (1,3)$ encaixa-se no caso 4, portanto empilha-se $(2,2)$ e coloca-se $(1,3)$ na pilha de pares de bases, como ilustra a Figura 4.7.

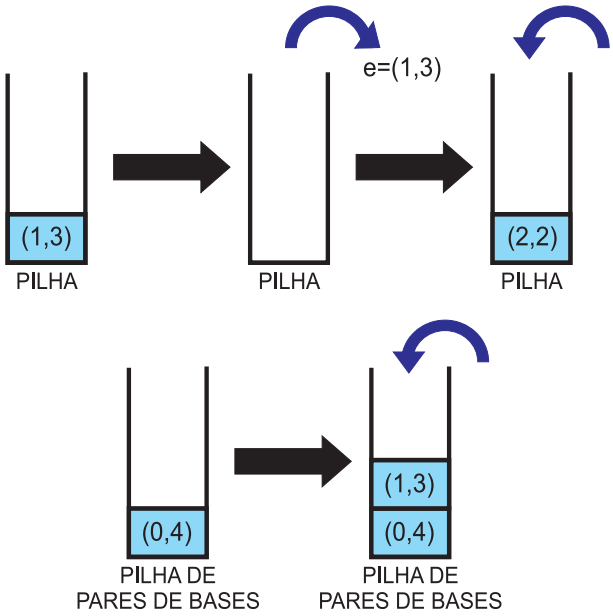

Figura 4.7: Passo 5(b) - Operações na pilha após a avaliação do elemento $(1,3)$

O elemento do topo da pilha é $e = (2,2)$. Como $i = j$, o procedimento acaba. A pilha de pares de bases contém os pares para a sequência, basta desempilhar. Para o exemplo em questão, a estrutura ótima final será: A(0)–U(4), C(1)–G(3) e A(2) desemparelhada, conforme ilustra a Figura 4.8. Um outro emparelhamento possível (mas não ótimo) teria pontuação 1 e seria: A(2)–U(4) com as bases A(0), C(1) e G(3) desemparelhadas.

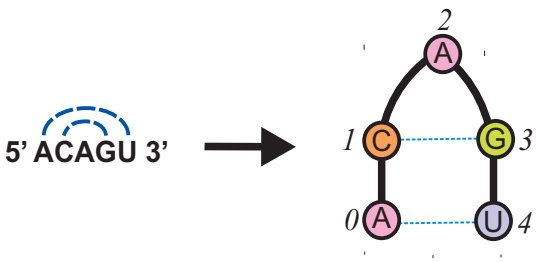

Figura 4.8: Configuração ótima para a sequência 5' ACAGU 3'

Capítulo 5

Modelo Termodinâmico para o Algoritmo de Nussinov-Jacobson

Este capítulo descreve os procedimentos de introdução de efeitos de temperatura no algoritmo de Nussinov-Jacobson. A seção 5.1 apresenta conceitos sobre energia livre como função da temperatura, dentro do modelo termodinâmico. A seção 5.2 descreve o modelo do vizinho mais próximo, que é utilizado para o cálculo das pontuações de energia. A seção 5.3 explica como o tratamento termodinâmico foi introduzido no algoritmo de Nussinov-Jacobson. A seção 5.4 apresenta os cálculos da energia e do calor específico, os quais foram as grandezas avaliadas no modelo implementado.

5.1 Energia livre

Além da maximização do número de pares de bases, a predição da estrutura secundária ótima da molécula de RNA pode ser realizada por meio da minimização da energia livre de Gibbs [19][36]. Esta é definida como a quantidade de energia disponível em um sistema para trabalhar sob condições ideais de temperatura e pressão [19]. Para a predição da estrutura ótima do RNA, assume-se que a molécula possui uma energia livre, que determina o seu potencial para formar pares de bases e, desta forma, liberar energia. Uma vez que as moléculas são estáveis se sua energia livre for pequena, a estrutura secundária ótima pode ser encontrada quando a molécula possuir energia global livre mínima.

No entanto, essa abordagem prevê a escolha de uma única estrutura ótima, enquanto a realidade física é que as moléculas em solução irão alternar entre diferentes conformações estruturais. Para amenizar esse problema e obter uma boa predição da melhor estrutura, é necessário

descrever as propriedades estatísticas de toda a população de moléculas, fazendo uso de ferramentas teóricas e computacionais no domínio da termodinâmica estatística [6].

Adotando um modelo termodinâmico, pode-se representar a energia livre de Gibbs ΔG como uma função da entalpia ΔH (quantidade de energia possuída por um sistema termodinâmico para ser transferida entre ele e o ambiente), da temperatura T (quantidade de energia cinética média de um sistema) e da entropia ΔS (quantidade da desordem relativa da energia de um sistema) [10][19], como mostra a equação 5.1.

$$\Delta G = \Delta H - T\Delta S \tag{5.1}$$

Em uma reação química ou uma mudança na configuração da estrutura do RNA, ΔG quantifica a espontaneidade da reação. Se ΔG de um determinado processo for negativa, o produto desse processo é favorecido e ele pode prosseguir espontaneamente. Por outro lado, uma ΔG positiva favorece os reagentes, impedindo o processo de prosseguir espontaneamente. Quando o equilíbrio é alcançado, $\Delta G = 0$ e a energia livre é transformada em calor ou aumenta a quantidade de entropia [19].

No caso da predição da estrutura secundária do RNA, ΔG é usada para quantificar a espontaneidade da molécula na dobradura em configurações específicas da estrutura secundária. A contribuição de energia de cada par de bases é determinada pelo número de pontes de hidrogênio da ligação. Como o par C–G é mantido por três pontes de hidrogênio, ele possui uma energia $\Delta G = -3$ Kcal/mol; já o par A–U, que é mantido por duas pontes, tem energia $\Delta G = -2$ Kcal/mol, considerando uma temperatura de 37^{o}C [19].

Nos algoritmos de predição da estrutura ótima a partir da minimização da energia livre é necessário atribuir uma energia a cada possível estrutura construída e comparar as estabilidades termodinâmicas relativas de todas as possíveis estruturas para uma determinada sequência. A energia global livre de uma estrutura molecular completa é, geralmente, estimada pela soma dos termos de energia livre independentes das diferentes partes de uma estrutura secundária [14][20]. A equação 5.2 mostra o cálculo da energia global livre E para uma estrutura S, sendo $e(r_i, r_j)$ a contribuição de energia ΔG entre as bases emparelhadas i e j pertencentes à sequência [19][30].

$$E(S) = \sum_{i,j \in S} e(r_i, r_j) \tag{5.2}$$

5.2 O modelo do vizinho mais próximo

Quando uma molécula de RNA dobra, sua energia global livre é reduzida à medida que são formados pares de bases. Isso, por sua vez, aumenta a estabilidade global da molécula [19]. Entretanto, da mesma forma que dois pares de bases adjacentes formam uma pilha de pares e estabilizam a estrutura, bases desemparelhadas em laços *hairpin*, laços salientes, laços interiores e laços de multi-ramificações desestabilizam a conformação [16].

Para uma cadeia de biopolímero, a energia livre de Gibbs pode ser calculada pelo modelo aditivo do vizinho mais próximo. Este é chamado de "aditivo" porque a energia global é dada pela soma das energias livres de seus elementos estruturais individuais. O termo "vizinho mais próximo" significa que a energia livre de cada estrutura depende somente das subestruturas que ela possui e dos pares de bases adjacentes [6].

Sendo assim, em vez de assumir que os pares de bases são completamente independentes, como no algoritmo de Nussinov-Jacobson, a energia livre das hélices é baseada na contribuição das energias de empilhamento de pares de bases e na contribuição da energia desestabilizadora dos laços. Assim, a energia global livre de uma estrutura é, grosseiramente, a soma das energias das subestruturas (laços) da estrutura secundária. Relações de recorrência capturam o tamanho e tipo de cada laço e retornam a estrutura com energia global mínima [19].

Os algoritmos de minimização de energia apresentam praticamente a mesma mecânica dos algoritmos de maximização no número de pares de bases, porém são mais complexos, uma vez que distinguem diferentes tamanhos e tipos de laços e atribuem a pontuação dos emparelhamentos de acordo com pares de bases adjacentes do vizinho mais próximo [14]. O algoritmo de Zuker, implementado nos programas *Mfold* [46][48][49][50] e *ViennaRNA* [33] é um exemplo de algoritmo de programação dinâmica para identificar a energia global mínima para uma determinada sequência de RNA [46][47]. Outro exemplo é o algoritmo de McCaskill [27], que calcula probabilidades para o emparelhamento de bases. Tais algoritmos apresentam complexidade computacional $O(n^3)$, sendo n o tamanho da sequência [20].

5.3 Introdução do modelo termodinâmico no algoritmo de Nussinov-Jacobson

Neste trabalho, para a predição da estrutura secundária do RNA considerando os efeitos de temperatura, foi introduzido um tratamento termodinâmico para o algoritmo de Nussinov-Jacobson, que é um algoritmo de maximização de pares de bases. Assim, o novo algoritmo gerado maximiza o número de pares de bases para um valor de temperatura predefinido. Enquanto no algoritmo original os pares de bases complementares C–G e A–U sempre emparelham-se, nessa versão do algoritmo a formação do par ocorre com uma probabilidade que depende da temperatura da solução em que a molécula encontra-se. A implementação do algoritmo modificado foi feita na linguagem de programação Java, utilizando a versão encontrada em [35].

Para o cálculo da probabilidade utiliza-se o peso de Boltzmann (já apresentado no Capítulo 3), que descreve a distribuição de equilíbrio para um conjunto (*ensemble*) de moléculas a uma dada temperatura. Considerando $\beta = \dfrac{1}{K_B T}$, a equação 5.3 representa a expressão matemática do peso de Boltzmann, sendo ε a energia da ligação e $\varepsilon_T = K_B T$ a energia térmica (K_B é uma constante chamada constante de Boltzmann e T é a temperatura do sistema) [6][21].

$$p = e^{\frac{-\varepsilon}{\varepsilon_T}} = e^{-\beta\varepsilon(i,j)} \tag{5.3}$$

Desta forma, bases complementares poderão ou não emparelhar-se, de acordo com a probabilidade p calculada para cada par de bases. A soma das energias de todos os pares será a energia total da molécula. Em relação ao algoritmo original, o cálculo de $\delta(i,j) = 1(0)$ se (i,j) emparelhadas (desemparelhadas) é alterado de:

$$\delta\left(i,j\right) = \begin{cases} 1 & \text{se } (i,j) \text{ complementares} \\ 0 & \text{se } (i,j) \text{ não complementares} \end{cases} \tag{5.4}$$

para as regras da equação 5.5:

$$\delta\left(i,j\right) = \begin{cases} 0 & \text{se } (i,j) \text{ não complementares} \\ \xi = \begin{cases} 1 & \text{com } p = e^{-\beta\varepsilon} \\ 0 & \text{com } p = 1 - e^{-\beta\varepsilon} \end{cases} & \text{se } (i,j) \text{ complementares} \end{cases} \tag{5.5}$$

As regras de emparelhamento mostradas na equação 5.5 foram inseridas no algoritmo de Nussinov-Jacobson da maneira descrita a seguir. Para cada possível par de bases, calcula-se

a probabilidade de emparelhamento p, com base em sua energia correspondente. É sorteado um número aleatório r entre 0 e 1. Realiza-se uma comparação: se o número sorteado for maior ou igual à probabilidade calculada, então as bases emparelham-se, caso contrário elas ficam desemparelhadas, como mostra o conjunto de regras da equação 5.6. Esse procedimento é similar ao algoritmo de Metropolis, amplamente utilizado em Mecânica Estatística [3].

$$\delta(i,j) = \begin{cases} 0 & \text{se } (i,j) \text{ não complementares} \\ \xi = \left\{ \begin{array}{ll} 1 & \text{se } r \geqslant p \\ 0 & \text{se } r < p \end{array} \right\} & \text{se } (i,j) \text{ complementares} \end{cases} \tag{5.6}$$

5.4 Caracterização das moléculas

Com a alteração descrita na seção anterior, o algoritmo de Nussinov-Jacobson torna-se estatístico, ou seja, as grandezas físicas devem ser obtidas de acordo com suas distribuições estatísticas e seus valores médios estudados de acordo com tais distribuições. Para descrever as conformações obtidas aplicando o algoritmo de Nussinov-Jacobson modificado, foram calculadas as seguintes grandezas:

- a energia média das moléculas $\langle E \rangle$, onde $\langle\rangle$ significa a média estatística sobre um *ensemble* de moléculas;
- a distribuição de energia das moléculas no *ensemble*;
- o calor específico do sistema C_V.

Em Mecânica Estatística, o valor médio de uma grandeza de interesse é calculado como:

$$\langle A \rangle = \frac{\sum_r A_r e^{-\beta E_r}}{\sum_r e^{-\beta E_r}} = \frac{\sum_r A_r e^{-\beta E_r}}{Z} \tag{5.7}$$

onde A indica a grandeza em estudo e Z é a função de partição, definida no capítulo 3.

Assim, a energia média de um sistema é dada por:

$$\langle E \rangle = \frac{\sum_r E_r e^{-\beta E_r}}{Z} \tag{5.8}$$

No entanto, percebendo que $\sum_r E_r e^{-\beta E_r} = -\frac{\partial}{\partial \beta}\left(\sum_r e^{-\beta E_r}\right)$, pode-se reescrever a equação 5.8 como:

$$\langle E \rangle = -\frac{1}{Z}\frac{\partial Z}{\partial \beta} \tag{5.9}$$

Da mesma forma, a energia quadrática média é dada pela equação:

$$\left\langle E^2 \right\rangle = \frac{\sum_r E_r^2 e^{-\beta E_r}}{Z} = -\frac{1}{Z}\frac{\partial^2 Z}{\partial \beta^2} \tag{5.10}$$

Na implementação do algoritmo de Nussinov-Jacobson, a energia média pode ser calculada diretamente, simplesmente somando as energias de cada par de bases formado i, ao invés de efetuar todas as derivadas mostradas, como mostra a equação 5.11.

$$\langle E \rangle = \frac{\sum_i E_i}{\textit{Número de moléculas}} \tag{5.11}$$

Já o calor específico de um sistema, calculado por meio da Mecânica Estatística, é dado por:

$$C_V = \frac{\beta}{T}\frac{\partial^2 lnZ}{\partial \beta^2} \tag{5.12}$$

onde Z é a função de partição.

No entanto, não existe uma forma de calcular as derivadas da função de partição nesse tipo de sistema. Para contornar esse problema, reescreve-se a equação 5.12 como:

$$C_V = \frac{\beta}{T}\frac{\partial}{\partial \beta}\frac{\partial lnZ}{\partial \beta} \tag{5.13}$$

Desenvolvendo a equação 5.13, obtém-se:

$$C_V = \frac{\beta}{T}\frac{\partial}{\partial \beta}\left(\frac{1}{Z}\frac{\partial Z}{\partial \beta}\right) = \frac{\beta}{T}\left[\frac{1}{Z}\frac{\partial^2 Z}{\partial \beta^2} - \frac{1}{Z^2}\left(\frac{\partial Z}{\partial \beta}\right)^2\right] \tag{5.14}$$

Como, na Mecânica Estatística, $\langle E \rangle = \dfrac{1}{Z}\dfrac{\partial Z}{\partial \beta}$, a equação do calor específico fica escrita como:

$$C_V = \frac{\beta}{T}\left[\left\langle E^2 \right\rangle - \langle E \rangle^2\right] \quad \text{ou} \quad \frac{C_V}{K_B} = \beta^2\left[\left\langle E^2 \right\rangle - \langle E \rangle^2\right] \tag{5.15}$$

Esse resultado é uma consequência de um teorema chamado “Teorema de Flutuação-Dissipação” e relaciona o calor absorvido por um sistema com as flutuações na energia do sistema, sujeito a uma variação de temperatura [3].

5.5 Resultados obtidos

Após a implementação do algoritmo de Nussinov-Jacobson com tratamento termodinâmico, foram realizados diversos testes a fim de avaliar o comportamento do algoritmo modificado. Como citado no Capítulo 4, o número de configurações possíveis cresce exponencialmente conforme o tamanho da sequência, portanto, para realizar os testes, foram escolhidas sequências de RNA curtas, com 121 bases. A fim de comparar resultados, foram simuladas moléculas de RNA de grupos diferentes. As sequências investigadas foram retiradas de um banco de dados público disponível na Internet [39] e referem-se a três grupos de reinos de bactérias: *Archaea*, *Eubacteria* e *Eukaryota* [44]. Dentro de cada grupo, escolheu-se duas moléculas diferentes, para efeito de comparação. A Tabela 5.5 apresenta as moléculas analisadas, com suas respectivas sequências de RNA.

Tabela 5.1: Moléculas simuladas com o algoritmo de Nussinov-Jacobson com tratamento termodinâmico

GRUPO	ESPÉCIE	SEQUÊNCIA DE RNA
Archae	*Nanoarchaeum equitans*	GUUUCGUGGGAGGGCCAUAGCGGCCCGGGAA CCACCCGUACCCAUCUCGAACACGGAAGUUA AGCCGGGCCGCGUCCCGAGUGGUACUGCCCC GCGAAGGGGUGGGAAGCUCGGGAUGCCC
	Natronococcus occultus	UAAGGCGGCCAUAGCGGCGGGGUUCCUCCCG UACCCAUCCCGAACACGGAAGAUAAGCCCGC CUGCGUAUUGGUGAGUACUGGAGUGGGAGAC CCUCUGGGAGAGCUGAUUCGCUGCUUUA
Eubacteria	*Actinomadura madurae*	CGUUCGGUGGUUUUGGCGAGGGGGAAACACC CGGUCCCAUUCCGAACCCGGAAGUUAAGCCU CUCAGCGCCGAUGGUACUGCAUGGGAGACUG UGUGGGAGAGUAGGACACCGCCGGACUU
	Halorhodospira halophila	UGCCUGGCGACCAUAGCGAGCGGGAACCACC CGAUCCGAUGCCGAACUCGGCAGUGAAACCG CUCAGCGCCGAUGGUAGUGCGACCACGCUGU CGUGCGAGAGUAGGUCAUCGCCAGGCCC
Eukaryotaa	*Carpopeltis crispata*	ACAUUCGGCCAUACCAGGACGACAAAUACCC CAUCCCAUCUCGAACUGGGCAGUUAAGUCUC CUCGGGCGCGCUUAGUACUGAGGUCAGGGAU GACUCGGGAAUCGCGCGUGCUGAAUGUU
	Cryptomonas paramecium	UUCUGUACGGUCAUACCUGGUUGGAAACGGC GGAUCCCGUCCGAUCUCCGAAGCUAAGCAAC CAUGGGCGUGUCUAGUACUCAGGUGGGGGAC CACUGGGGAAGCGCACGUACUGUACAGC

Para tornar possível a análise estatística dos dados, o programa computacional foi executado 5000 vezes para cada sequência. O parâmetro de entrada foi o inverso da temperatura $\beta = \frac{1}{K_B T}$, que variou de 0 até o valor em que a molécula alcançou o número máximo de pares de bases. Para cada β, foram gerados arquivos contendo os dados de energia e calor específico.

Comparando os dados dos diferentes grupos de moléculas avaliados, conclui-se que não há uma diferença significativa entre eles, ou seja, os três grupos (*Archaea*, *Eubacteria* e *Eukaryota*) apresentam comportamentos similares quando investigados com o algoritmo de Nussinov-Jacobson com tratamento termodinâmico. O número máximo de pares de bases para as sequências variou entre 46 e 48 pares.

O programa computacional gera, para cada molécula, todas as configurações possíveis a

uma dada temperatura. A Figura 5.1 mostra uma das configurações ótimas geradas para uma molécula *Nanoarchaeum equitans*, do grupo *Archaea*, com $\beta = 10$. Para esse valor de β a molécula forma o número máximo de pares possíveis que, neste caso, é 46. As regiões emparelhadas correspondem às duplas hélices da cadeia, enquanto as regiões desemparelhadas correspondem às porções lineares.

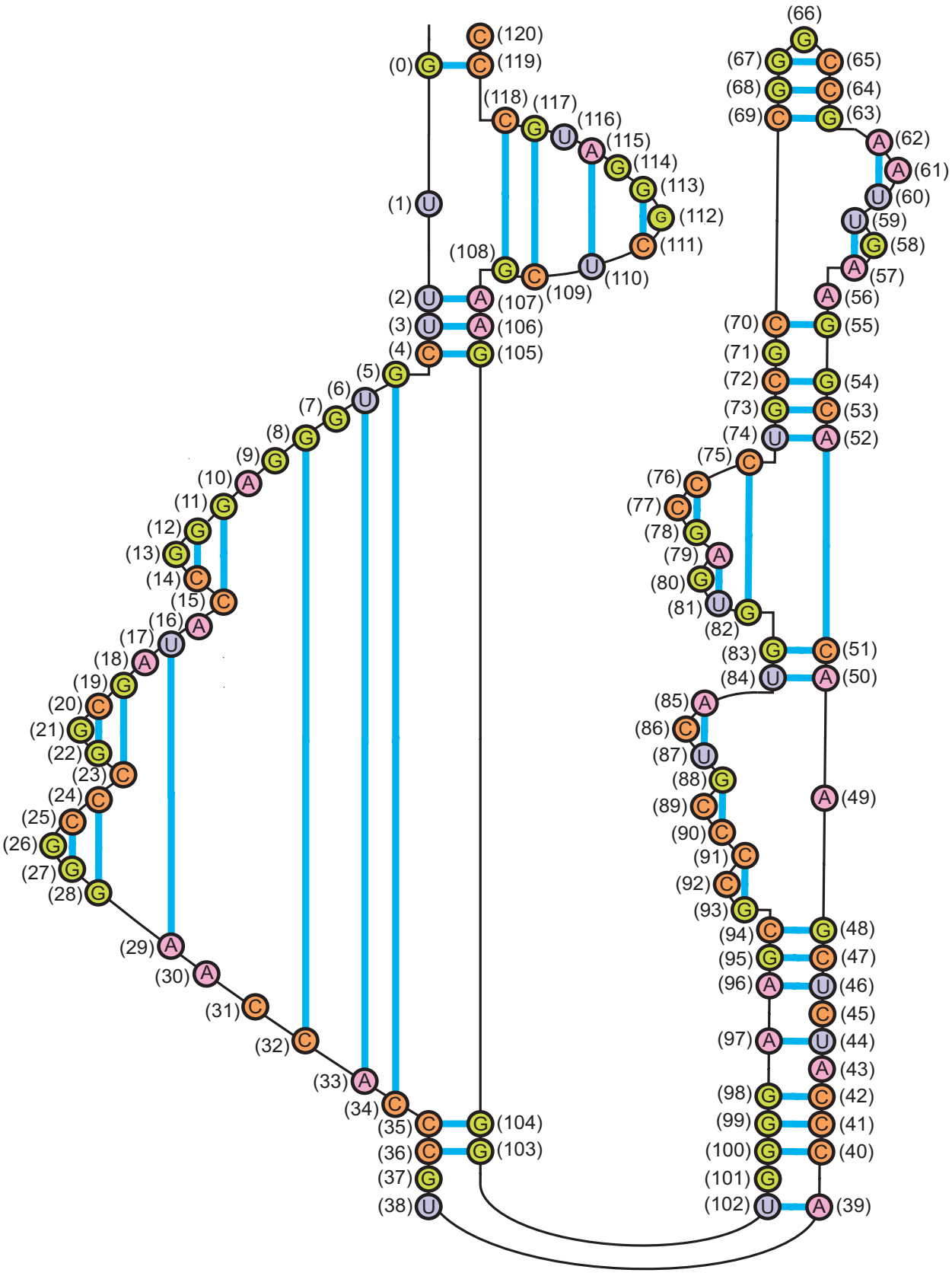

Figura 5.1: Amostra de configuração ótima para uma molécula *Nanoarchaeum equitans* para $\beta = 10$

A Figura 5.2 apresenta o gráfico da energia média $\langle E \rangle$ obtida como função de β para a molécula *Carpopeltis crispata*, do grupo *Eukaryota*. Pode-se observar que, para β baixo, isto é, para temperatura alta, as moléculas formam poucos pares de bases. Conforme β aumenta (e a temperatura diminui) são formados mais emparelhamentos, até que a energia alcance um valor de saturação, correspondente ao número máximo de pares possíveis para aquela molécula que, para o exemplo em questão, é 48. Poderia-se esperar um número máximo aproximado de 60 pares, no caso da molécula ser um homopolímero, porém isso não ocorre devido à heterogeneidade da sequência. Sendo assim, o número máximo de pares para uma sequência heterogênea, geralmente, é inferior ao número máximo para um homopolímero.

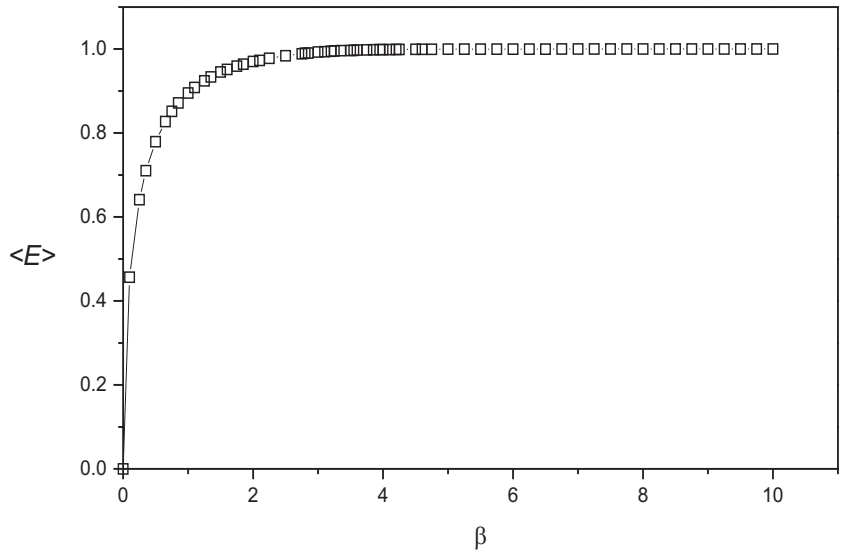

Figura 5.2: Gráfico da energia média como função de β para uma molécula *Carpopeltis crispata*

Devido aos efeitos de temperatura, nem todas as moléculas terão o mesmo número de pares emparelhados para um determinado β. Sendo assim, pode-se fazer uma estimativa da distribuição das energias das moléculas para as 5000 configurações geradas. A Figura 5.3 apresenta o gráfico da distribuição das energias como função de β. Observa-se que, para β baixo, os valores de energia variam em um intervalo mais largo do que para β mais alto. À medida que β aumenta (e a temperatura diminui), o intervalo de energias atingidas vai ficando mais estreito até que, a uma dada temperatura, todas as moléculas possuem a mesma energia, limitadas pelo número máximo de pares que podem formar. Esse resultado concorda com o gráfico da energia média (Figura 5.2), que mostra que as energias atingem um valor de saturação a uma

determinada temperatura.

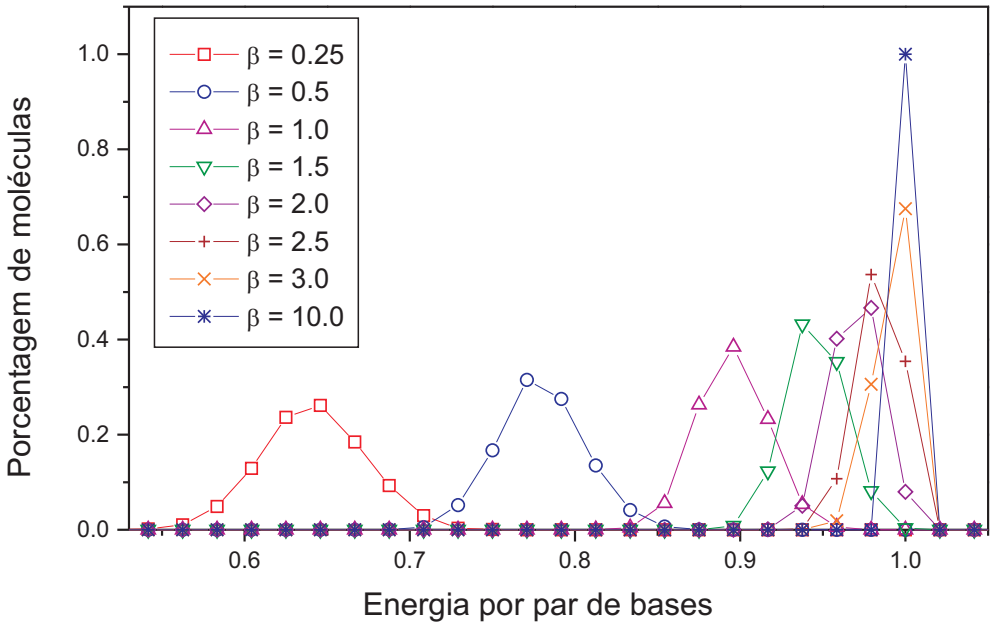

Figura 5.3: Gráfico da distribuição das energias como função de β para uma molécula *Carpopeltis crispata*

Os resultados para a energia mostram que os biopolímeros podem ser encontrados em duas fases: uma fase com baixa densidade de pares de bases, que será chamada de *Fase 1*, e outra fase com alta densidade de pares de bases, que será chamada de *Fase 2*. A *Fase 1* ocorre para β baixo (alta temperatura) e a *Fase 2* ocorre para β alto (baixa temperatura). O polímero pode sofrer transições entre as fases citadas acima mediadas pela variação de temperatura do ambiente onde ele se encontra. A temperatura na qual o polímero muda de fase é chamada de "ponto crítico".

A Figura 5.4 apresenta o gráfico do calor específico $\frac{C_V}{K_B}$ como função de β, também para a molécula *Carpopeltis crispata*. A temperatura para a qual o calor específico é máximo corresponde ao ponto crítico do sistema.

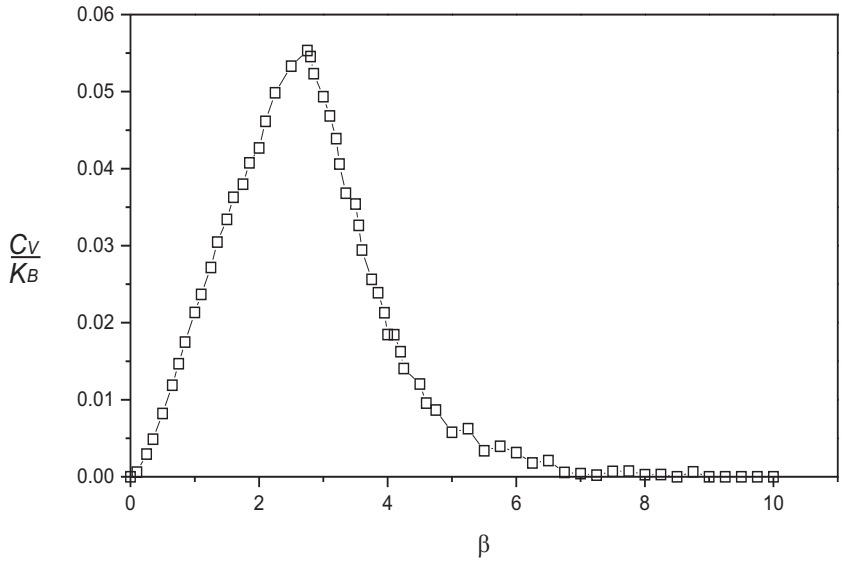

Figura 5.4: Gráfico do calor específico como função de β para uma molécula *Carpopeltis crispata*

O algoritmo de Nussinov-Jacobson com tratamento termodinâmico mostrou que é possível mapear a formação de pares de bases variando a temperatura. Ele forneceu o cálculo da energia média e do calor específico das moléculas e detectou a existência de duas fases que o polímero pode assumir mediante a temperatura na qual se encontra. Os resultados obtidos com o programa computacional foram satisfatórios, e condizem com os resultados de experimentos reais e modelos de simulação.

Capítulo 6

Considerações Finais

Este trabalho de conclusão de curso estudou o algoritmo de Nussinov-Jacobson para a predição da estrutura secundária de biopolímeros do tipo RNA. Foram apresentados conceitos sobre biologia molecular e o problema de *RNA-folding*, a fim de fornecer ao leitor um embasamento teórico da parte biológica. Foram descritos os cálculos para a determinação da estrutura secundária a partir de uma função de partição iterativa. Esses cálculos constituíram a base para o desenvolvimento do algoritmo estudado.

Foram apresentados os métodos físicos e computacionais para a predição da estrutura do RNA e justificou-se a utilização do algoritmo de Nussinov-Jacobson, o qual utiliza técnicas de programação dinâmica para obter a melhor estrutura. Explicou-se o funcionamento do algoritmo, que fornece a estrutura ótima a partir da maximização do número de pares de bases.

Após a análise e implementação do algoritmo, ele foi melhorado, recebendo um tratamento termodinâmico que permitiu avaliar a conformação da molécula de RNA como função da temperatura do ambiente onde ela se encontra. Para efeito de estatística, foram realizados diversos testes com o algoritmo modificado, os quais forneceram os valores da energia média e do calor específico das diferentes conformações moleculares.

Na análise da energia média das moléculas, os resultados obtidos mostraram que, à alta temperatura, são formados poucos pares de bases e, conforme a temperatura diminui, aumenta o número de emparelhamentos, até que a energia atinge um valor de saturação, limitado pelo número máximo de pares possíveis para a molécula. A distribuição das energias como função da temperatura mostrou que, para baixa temperatura, os valores de energia variam em um intervalo mais estreito do que para alta temperatura, até que todas as moléculas atinjam a mesma energia, condizendo com o valor de saturação obtido na avaliação da energia média.

Os resultados obtidos para o calor específico detectaram a existência de duas fases que podem ser assumidas pelo polímero, sendo uma fase com baixa densidade de pares de bases (para alta temperatura) e outra com alta densidade de pares (para baixa temperatura). O valor máximo assumido pelo calor específico fornece a temperatura em que ocorre a transição entre as fases, chamada de ponto crítico do sistema.

Conclui-se que os resultados foram satisfatórios e os objetivos do trabalho foram alcançados. Com a implementação e análise do algoritmo de Nussinov-Jacobson com tratamento termodinâmico foi possível avaliar a formação de pares de bases mediante a variação da temperatura e encontrar o ponto crítico do sistema.

Como um trabalho futuro, poderia ser feita a implementação da influência das energias de empilhamento (e não apenas das energias de emparelhamento) dos pares de bases formados, considerando os efeitos de rigidez da molécula.

Referências Bibliográficas

[1] BAO, F. S. **Python Implementation of Nussinov Folding Algorithm for RNA Secondary Structure Prediction**. Disponível em <http://narnia.cs.ttu.edu/drupal/node/123>. Acesso em: Novembro, 2008. Texas Tech University - Dept. of Computer Science.

[2] BELLMAN, R. **Dynamic Programming**. Princeton - New Jersey: Princeton University Press, 1957.

[3] BINDER, K.; HEERMANN, D. W. **Monte Carlo Simulation in Statistical Physics: An Introduction**. 2. ed. Berlin: Springer-Verlag, 1992.

[4] CECH, T. R. Conserved sequences and structures of group i introns: building an active site for RNA catalysis - a review. **Gene**, Boulder, v.73, p.259–271, December, 1988.

[5] CIBIV. **Structure Prediction**. Disponível em <http://www.cibiv.at>. Acesso em: Julho, 2009. CIBIV - Center for Integrative Bioinformatics Vienna / MFPL - Max F. Perutz Laboratories.

[6] CLC bio, Denmark - EUA. **Bioinformatics Explained**, 2008.

[7] COLNAGO, L. A.; ALMEIDA, F. C. L.; VALENTE, A. P. Espectrometria de massa e RMN multidimensional e multinuclear: Revolução no estudo de macromoléculas biológicas. **Química Nova na Escola**, [S.l.], v.16, p.32–37, Outubro, 2002.

[8] CRICK, F. On protein synteshis. In: SYMPOSIA OF THE SOCIETY FOR EXPERIMENTAL BIOLOGY, 1958. **Proceedings...** Cambridge: [s.n.], 1958. p.138–163.

[9] CRICK, F. Central dogma of molecular biology. **Nature**, Cambridge, v.227, p.561–563, August, 1970.

[10] DESCHÊNES, A. **A genetic algorithm for RNA secondary structure prediction using stacking energy thermodynamic models**. Burnaby - Canada: Simon Fraser University, April, 2005. Dissertação.

[11] DE SOUTO, M. C. P. **Conceitos Básicos de Biologia Molecular**. Disponível em <http://www.dimap.ufrn.br/∼marcilio/BIOINFORMATICA/course-Bioinformatica.htm>. Acesso em: Julho, 2009. DIMap - Departamento de Informática e Matemática Aplicada /UFRN - Universidade Federal do Rio Grande do Norte.

[12] DE SOUSA, F. L. **Otimização extrema generalizada: um novo algoritmo estocástico para o projeto ótimo**. São José dos Campos: INPE - Instituto Nacional de Pesquisas Espaciais, Setembro, 2002. Tese.

[13] EDDY, S. R. et al. **Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids**, chapter: RNA Structure Analysis, p.260–298. Cambridge University Press, Cambridge, 1. ed., 1999.

[14] EDDY, S. R. How do RNA folding algorithms work? **Nature Biotechnology**, New York, v.22, n.11, p.1457–1458, November, 2004.

[15] FLAMM, C. et al. RNA folding at elementary step resolution. **RNA**, USA, v.6, p.325–338, 2000.

[16] FREYHULT, E. **A Study in RNA Bioinformatics: Identification, Prediction and Analysis**. Sweden: Uppsala University, December, 2007. Tese.

[17] GATTASS, R. et al. **Fundamentos da Ressonância Magnética Funcional**. Disponível em <http://www.cerebromente.org.br/n13/tecnologia/ressonancia.htm>. Acesso em: Julho, 2009. Cérebro e Mente.

[18] GOMES, R.; ZILHãO, R. **Before Molecular Biology...** Disponível em <http://bmg.fc.ul.pt/Disciplinas/GBM/aulas/1DogCentrEstrutMolBiolMol.pdf>. Acesso em: Julho, 2009. BMG - Biologia Molecular e Genética / Faculdade de Ciências - Universidade de Lisboa.

[19] HENDRIKS, A. **A parallel evolutionary algorithm for RNA secondary structure prediction**. Burnaby - Canada: Simon Fraser University, July, 2005. Dissertação.

[20] HIGGS, P. G. RNA secondary structure: physical and computational aspects. **Quarterly Reviews of Biophysics**, Manchester, v.33, n.3, p.199–253, August, 2000.

[21] HUANG, K. **Statistical Mechanics**. 2. ed. New York: John Wiley & Sons, 1987.

[22] KIM, H. **RNA Folding with Nossinov-Jacobsen Algorithm**. Disponível em <http://www.ibluemojo.com/school/rna_folding.html>. Acesso em: Novembro, 2008. University of Washington.

[23] LE, S. Y.; ZUKER, M. Common structures of the 5' non-coding RNA in enteroviruses: Thermodynamical stability and statistical significance. **Journal of Molecular Biology**, Ottawa, v.216, p.729–741, December, 1990.

[24] MATHEWS, D. et al. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. **Journal of Molecular Biology**, New York, v.288, p.911–940, March, 1999.

[25] MATHEWS, D. H.; TURNER, D. H. Experimentally derived nearest-beighbor parameters for the stability of RNA three- and four-way multibranch loops. **Biochemistry**, New York, v.41, p.869–880, January, 2002.

[26] MATHEWS, D. H.; TURNER, D. H. Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. **Journal of Molecular Biology**, New York, v.317, p.191–203, May, 2002.

[27] MCCASKILL, J. S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. **Biopolymers**, [S.l.], v.29, n.6-7, p.1105–1119, May-June, 1990.

[28] NUSSINOV, R. et al. Algorithms for loop matchings. **SIAM Journal an Applied Mathematics**, [S.l.], v.35, p.68–82, July, 1978.

[29] PACHECO, M. A. C. **Algoritmos Genéticos: Princípios e Aplicações**. Disponível em <http://www.ica.ele.puc-rio.br/DownloadsAcesso em: Julho, 2009. ICA - Laboratório de Inteligência Computacional Aplicada / Pontifícia Universidade Católica do Rio de Janeiro.

[30] POONIAN, J. **DPA-RNAPredict: A Dynamic Programming Algorithm for RNA Secondary Structure Prediction**. Burnaby - Canada: Simon Fraser University, May, 2007. Dissertação.

[31] RAMALHO, J.; TAVARES, L.; REAL, S. **Replicação e Síntese Protéica**. Disponível em <http://www.esec-odivelas.rcts.pt/BioGeo/ficha_rep.htm>. Acesso em: Julho, 2009. Escola Secundária de Odivelas.

[32] ROSA, A. **Statistical Mechanics of Polymer Stretching**. Trieste: Scuola Internazionale Superiore di Studi Avanzati, October, 2003. Tese.

[33] SCHUSTER, P. et al. From sequences to shapes and back: a case study in RNA secondary structures. **Proc. Biol. Sci.**, [S.l.], v.255, n.1344, p.279–284, March, 1994.

[34] SCHUSTER, P. Prediction of RNA secondary structures: from theory to models and real molecules. **Reports on Progress in Physics**, Bristol, v.69, p.1419–1477, April, 2006.

[35] SENGER, C. **NutRussian - Nussinov Implementation**. Disponível em <http://page.mi.fu-berlin.de/csenger/alg/NutRussian.html>. Acesso em: Novembro, 2008. Freie Universität Berlin.

[36] SETUBAL, J. C.; MEIDANIS, J. **Introduction to Computational Molecular Biology**. 1. ed. Boston: PWS Publishing, 1996.

[37] SHUBEITA, F. M.; NAVAUX, P. O. A. **Computação Evolutiva e Lógica Fuzzy**. Disponível em <http://www.inf.ufrgs.br/procpar/disc/cmp135/trabs/fauzi/t1/CompEvolutivaLogicaFuzzy.doc>. Acesso em: Julho, 2009. PPGC - Programa de Pós-graduação em Computação / UFRGS - Universidade Federal do Rio Grande do Sul.

[38] SOUZA, M. J. F. **Metaheurísticas**. Disponível em <http://www.decom.ufop.br/prof/marcone/Disciplinas/InteligenciaComputacional/TranspSimulatedAnnealing.pdf>. Acesso em: Julho, 2009. DECOM - Departamento de Computação / UFOP - Universidade Federal de Ouro Preto.

[39] SZYMANSKI, M. et al. **5S Ribosomal RNA Database**. Disponível em <http://www.man.poznan.pl/5SData/>. Acesso em: Outubro, 2009.

[40] TINOCO, I. J.; BUSTAMANTE, C. How RNA folds. **Journal of Molecular Biology**, Berkeley, v.293, p.271–281, October, 1999.

[41] TSANG, H. H.; WIESE, K. C. Sarna-predict: A simulated annealing algorithm for RNA secondary structure prediction. In: 2006 IEEE SYMPOSIUM ON COMPUTATIONAL INTELLIGENCE IN BIOINFORMATICS AND COMPUTATIONAL BIOLOGY, 2006. **Proceedings...** Toronto - Canada: [s.n.], 2006. p.466–475.

[42] TSANG, H. H.; WIESE, K. C. Sarna-predict: A study of RNA secondary structure prediction using different annealing schedules. In: 2007 IEEE SYMPOSIUM ON COMPUTATIONAL INTELLIGENCE IN BIOINFORMATICS AND COMPUTATIONAL BIOLOGY, 2007. **Proceedings...** Honolulu - Hawaii: [s.n.], 2007. p.239–246.

[43] WATERMAN, M. S.; SMITH, T. F. RNA secondary structure: A complete mathematical analysis. **Mathematical Biosciences**, USA, v.42, n.3-4, p.257–266, December, 1978.

[44] WOESE, C. R.; KANDLER, O.; WHEELIS, M. L. Towards a natural system of organisms: Proposal for the domains archea, bacteria, and eucarya. **Proc. Natl. Acad. Sci. - PNAS**, USA, v.87, n.12, p.4576–4579, June, 1990.

[45] XIA, T. et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with watson-crick base pairs. **Biochemistry**, New York, v.37, p.14716–14735, October, 1998.

[46] ZUKER, M.; STIEGLER, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. **Nucleid Acids Research**, USA, v.9, n.1, p.133–148, January, 1981.

[47] ZUKER, M.; SANKOFF, D. RNA secondary structures and their prediction. **Bulletin of Mathematical Biology**, New York, v.46, n.4, p.591–621, July, 1984.

[48] ZUKER, M. Computer prediction of RNA structure. **Methods in Enzymology**, United States, v.180, p.262–288, 1989.

[49] ZUKER, M. On finding all suboptimal foldings of an RNA molecule. **Science**, New York, v.244, n.4900, p.48–52, April, 1989.

[50] ZUKER, M.; MATHEW, D. H.; TURNER, D. H. Algorithms and thermodynamics for RNA secondary structure prediction: a pratical guide. In: RNA BIOCHEMISTRY AND BIOTECHNOLOGY (Barciszewski, J. & Clark, B. F. C., eds.). NATO ASI SERIES. KLUWER ACADEMIC PUBLISHERS, 1999. **Proceedings...** Boston: [s.n.], 1999. p.11–43.