UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

Me. Pedro Rafael Costa

Influência da dinâmica transcricional no dobramento da molécula de RNA

Brasil

Agosto de 2016

Me. Pedro Rafael Costa

Influência da dinâmica transcricional no dobramento da molécula de RNA

Tese apresentada ao Instituto de Biociências da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de Botucatu, para obtenção do título de Doutor em Ciências Biológicas – Genética.

Universidade Estadual Paulista – UNESP

Departamento de Física e Biofísica

Programa de Pós-Graduação: Ciências Biológicas (Genética)

Orientador: Prof. Dr. Ney Lemke

Brasil

Agosto de 2016

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Costa, Pedro Rafael.
Influência da dinâmica transcricional no dobramento da molécula de RNA / Pedro Rafael Costa. - Botucatu, 2016

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Ney Lemke
Capes: 20804008

1. Acido ribonucleico. 2. Expressão gênica. 3. Biologia molecular. 4. Genética molecular.

Palavras-chave: Dobramento cotranscricional; Estrutura secundária; RNA.

Para meus pais, Cláudio e Rose.

Agradecimentos

Meus mais sinceros agradecimentos a todos que me ajudaram na elaboração deste trabalho, em especial:

– Ao Professor Doutor Ney Lemke, pelos 8 anos de orientação e incentivo;

– À equipe do Laboratório de Bioinformática e Biofísica Computacional do Departamento de Física e Biofísica do IBB-Unesp e agregados de Laboratórios vizinhos, em especial aos colegas Me. Rafael Takahiro Nakajima e Dr. Rafael Toledo de Souza, pelos momentos de discussão e descontração;

– À minha namorada e grande amiga Tahila Andrighetti, com quem dividi os bons e os maus momentos durante esse trabalho;

– Aos meus pais e à minha irmã, Cláudia, pelo apoio, carinho e atenção, sem os quais não estaria aqui hoje;

– Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – processo 152838/2012-0 – e à Fundação de Amparo à Pesquisa do Estado de São Paulo — processo 2012/19377-4 – pelos anos de apoio financeiro.

“Science is always wrong.

It never solves a problem without creating ten more.”

George Bernard Shaw

Resumo

Uma grande variedade de sequências de RNA presentes nos transcriptomas mas com funções ainda desconhecidas tem estimulado o desenvolvimento de técnicas experimentais e computacionais que colaborem na determinação do papel dessas moléculas. Entretanto, a compreensão dos mecanismos de ação de uma dada molécula de RNA envolve não somente a determinação de sua estrutura de mínima energia livre, mas também o estudo do comportamento de suas conformações metaestáveis. Os algoritmos existentes para predição da estrutura de moléculas de RNA ignoram armadilhas cinéticas que podem levar a formação de estruturas subótimas e utilizam modelos termodinâmicos incompletos, muitas vezes ignorando a formação de estruturas mais complexas, como os *pseudonós*. Nesse trabalho apresentamos um algoritmo para simulação do dobramento cotranscricional para o estudo dos efeitos da conformação espacial da molécula de RNA na cinética da transcrição. A partir da determinação das conformações permitidas, o algoritmo estabelece uma série de reações de transição entre esses estados, com valores das taxas de ocorrência ponderados através de uma distribuição de probabilidade de Boltzmann baseada na variação de energia livre de Gibbs desses estados. As energias livres das estruturas secundárias são determinadas segundo o *modelo dos primeiros vizinhos* e dois algoritmos foram implementados para o cálculo da energia livre dos pseudonós. Finalmente, simulações de Monte Carlo baseadas no algoritmo de Gillespie foram realizadas para determinação do caminho de dobramento da molécula. Exemplos de aplicações demonstram o potencial do programa desenvolvido.

Palavras-chaves: RNA. Estrutura secundária. Simulações de Monte Carlo. Dobramento cotranscricional. Biologia computacional.

Abstract

The discovery of a wide variety of RNA sequences present in transcriptomes with unknown function has stimulated the development of experimental and computational techniques to determine the function of these molecules. However, understanding the activities of a given RNA molecule involves not only finding its minimum free energy structure, but also studying its metastable structures. The methodologies available for predicting RNA structure ignore kinetic traps that can lead to formation of suboptimal structures and are based in over-simplified thermodynamic models. These methodologies usually can not predict RNA conformations with more complex topologies, such as *pseudoknots*. In this work we present a computational model for cotranscriptional folding that considers the influence of RNA structures during transcription elongation. After determining the allowed conformations for the sequence of interest, the algorithm established a series of allowed reactions between these structures. The reactions rates are weighted by the Boltzmann factor based on Gibbs free energy variation between the states. The free energies for secondary structures were estimated by the nearest-neighbor model, and two algorithms were implemented to calculate the free energy of pseudoknots. Finally, Monte Carlo simulations based on the Gillespie algorithm were performed to find out the RNA folding path. We show using examples the potential of the software developed.

Key Words: RNA. Secondary Structure. Monte Carlo simulations. Cotranscriptional folding. Computational biology.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

A, G	Adenina e Guanina: bases púricas presentes nos ácidos nucleicos.
C, T, U	Citosina, Timina e Uracila: bases pirimídicas presentes nos ácidos nucleicos.
CAT	Complexo de Alongamento Transcricional, constituído pela RNAP, pelo DNA e pela fita de RNA nascente.
DNA	*Deoxyribonucleic acid*, Ácido desoxirribonucleico, polímero composto por dexorribonucleotídeos unidos através de ligações fosfodiéster.
FP, FN	Métricas para classificadores. FP: *False positives*, falsos positivos; FN: *False negatives*, falsos negativos.
MFE	*Minimum Free Energy*, mínima energia livre. Relacionado à estrutura mais provável para uma dada molécula.
MRA	*Multiple Round Approach*, abordagem considerando múltiplas RNAPs durante a transcrição de um mesmo ORF.
NNDB	*Nearest Neighbor Database*, banco de dados com parâmetros e regras para determinação da energia livre de ácidos nucleicos a partir do princípio dos primeiros vizinhos.
NTP	Nucleotídeo trifosfato genérico.
ORF	*Open Reading Frame*, fase de leitura aberta: cada uma das sequências de DNA compreendidas entre um códon de início da tradução e um códon de terminação.
RNA	*Ribonucleic acid*, Ácido ribonucleico, polímero composto por ribonucleotídeos unidos através de ligações fosfodiéster.
RNAP	RNA polimerase DNA-dependente, enzima responsável pelo processo de transcrição do DNA em RNA.

SBR — Segmento Bifilamentar de RNA, componente primordial na formação de estruturas secundárias e pseudonós, composto por dois segmentos unifilamentares complementares inversos pareados de uma mesma molécula de RNA.

SUR — Segmento Unifilamentar de RNA em estruturas secundárias e pseudonós.

SRA — *Single Round Approach*, abordagem considerando apenas uma RNAP durante a transcrição.

TP, TN — Métricas para classificadores. TP: *True positives*, verdadeiros positivos; TN: *True negatives*, verdadeiros negativos.

Lista de símbolos

$[X]$	Concentração de X.
A, A'	Espécie química genérica.
A_1, A_2	Moléculas de ácido nucleico complementares.
C_T	Concentração total de filamentos numa reação de desnaturação.
$\mathbf{I}$	Vetor que representa a distribuição de segmentos unifilamentares e bifilamentares de uma dada conformação de uma molécula de RNA.
$\mathrm{Id}_{i,k}$	*Identificador* para a rede i, que se apresenta no modo k.
K	Constante de equilíbrio de uma reação química; notação também utilizada para a constante de rigidez da mola teórica entre SBRs para determinação da energia potencial devido à curvatura de um segmento unifilamentar presente num pseudonó.
k_0	Pré-fator constante para taxa de transição entre estados. Empiricamente, $k_0 = 0{,}1\ \mathrm{s}^{-1}$ para transição entre conformações de RNA.
k_B	Constante de Boltzmann: $1{,}38 \times 10^{-23}\ \mathrm{J{\cdot}K^{-1}}$.
$k_{i,j}$	Taxa de transição do estado i para o estado j.
l	Distância entre os nucleotídeos no modelo de contas e molas para determinação da energia potencial devido à curvatura de um segmento unifilamentar presente num pseudonó.
l_K	Comprimento de Kuhn para um polímero: $l_K = 2l_P$.
l_n	*KineFold*: Comprimento aparente de um SBR; *Vfold*: Número de nucleotídeos presentes no segmento unifilamentar do pseudonó.
l_P	Comprimento de persistência: relaciona a resistência ao dobramento da cadeia em relação à energia térmica.

$\mathbf{M}$	Matriz para identificação dos pareamentos possíveis para uma dada sequência de RNA.
$m_{i,j}$	Elemento localizado na linha i, coluna j da matriz $\mathbf{M}$.
nt	Símbolo para quantidade de nucleotídeos.
pb	Símbolo para quantidade de bases pareadas.
R	Constante dos gases ideais: 8,31 $\mathrm{J{\cdot}mol^{-1}{\cdot}K^{-1}}$.
ra_x, rf_y	*Rede aberta* do subtipo x e *rede fechada* do subtipo y.
s_n	*KineFold*: Número de nucleotídeos presentes num segmento unifilamentar. *Vfold*: Número de nucleotídeos presentes em um SBR do pseudonó.
T	Temperatura em Kelvin ou em graus Celsius.
T_m	*Melting temperature*, temperatura de desnaturação: temperatura na qual metade dos SBRs encontram-se separados em solução.
U_{curv}	Energia potencial devido à curvatura de um segmento unifilamentar presente em um pseudonó.
ΔG°_T	Variação da energia livre de Gibbs para uma dada temperatura T.
ΔH°	Variação da entalpia.
ΔS°	Variação da entropia.
θ	Ângulo entre as direções de hastes sucessivas, em radianos, no modelo de contas e molas para determinação da energia potencial devido à curvatura de um segmento unifilamentar presente num pseudonó.

Sumário

1 Introdução

Registros da busca por explicações capazes de justificar a transmissão das características parentais datam da Antiguidade Clássica. Filósofos como Teofrasto, Hipócrates e Aristóteles propuseram teorias a respeito da hereditariedade observada nos diferentes organismos.

A *Teoria da Pangênese* foi desenvolvida por Charles Darwin em 1868 para justificar os resultados observados em suas pesquisas envolvendo a seleção artificial de caraterísticas fenotípicas. De acordo com essa teoria, os organismos progenitores expostos a diferentes condições produziriam pequenas "gêmulas" durante sua vida, que poderiam ser transmitidas para suas células germinativas. O encontro promovido através da reprodução sexuada dessas partículas derivadas dos progenitores resultaria no surgimento de um novo organismo com características de ambos os participantes (DARWIN, 1868).

A Teoria da Pangênese de Darwin foi substituída pelas leis de herança de Mendel. A redescoberta de seus trabalhos, no início do século XX, uniu especialistas em diferentes campos, como virologia, microbiologia, bioquímica, física e química, na busca da compreensão dos aspectos fundamentais do conceito de "fatores de hereditariedade" propostos pelo monge agostiniano. Em 1905, o botânico dinamarquês Wilhelm Johannsen denominou esses fatores *gens*, do grego *genea*, "geração, raça, ascendência"; em português, *gene*.

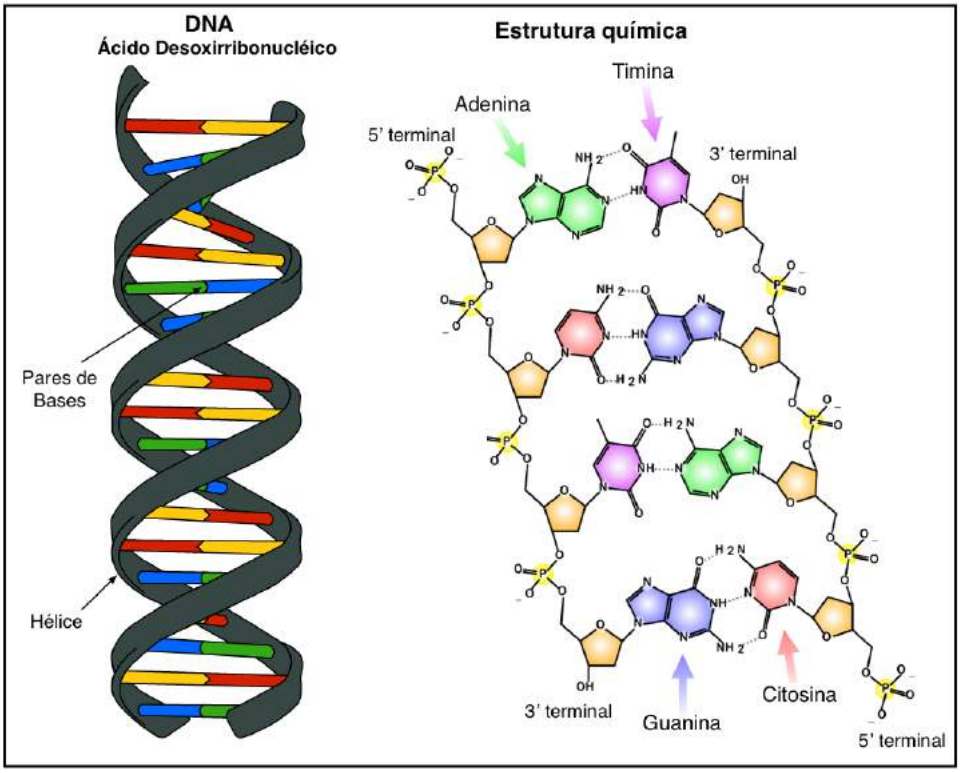
Inicialmente, as proteínas foram consideradas as moléculas portadoras do material genético. Em 1928, Frederick Griffith realizou experimentos com cepas de bactérias *Streptococcus pneumoniae*, nos quais destroços de bactérias mortas de uma determinada cepa eram capazes de *transformar* as células bacterianas vivas de outra, que passavam a possuir as características da primeira (GRIFFITHS et al., 2006). Oswald Avery e seus colegas pesquisadores do Instituto Rockefeller retomaram os trabalhos de Griffith e, em 1944, conduziram experimentos buscando determinar qual era, afinal, esse "agente transformador". A equipe descartou a possibilidade das proteínas atuarem como carregadoras do material genético diante dos resul-

tados obtidos, e concluíram que o verdadeiro responsável pela transmissão das características seria o polímero de *ácido desoxirribonucleico*, DNA (AVERY; MACLEOD; MCCARTY, 1944). Mais tarde, o experimento de Hershey e Chase (1952) confirmaria irrefutavelmente o DNA como a molécula responsável por carregar as informações hereditárias em determinados organismos.

O primeiro registro da existência do DNA data de 1871, quando o médico suíço Friedrich Miescher constatou a presença de uma molécula no núcleo celular composta por hidrogênio, carbono, oxigênio, nitrogênio e fósforo, denominando-a de *nucleína* (DAHM, 2005). Em 1889, seu aluno Richard Altmann confirmou a existência desse composto e determinou seu caráter ácido, passando a denominá-la *ácido nucleico*. Uma análise mais precisa de sua composição química, conduzida por Phoebus Levene, em 1919, revelou a presença de quatro bases nitrogenadas diferentes unidas através de seus grupos fosfato, e de um açúcar, denominado *desoxirribose* (KARP, 2005). Quase trinta e cinco anos depois, James Watson e Francis Crick, com base nos trabalhos de Rosalind Franklin e de Maurice Wilkins, determinaram a estrutura química da molécula de DNA, representada na Figura 1. No final de sua publicação, os autores notaram que seu modelo "... sugere de forma direta a existência de um possível mecanismo de replicação do material genético." (WATSON; CRICK et al., 1953).

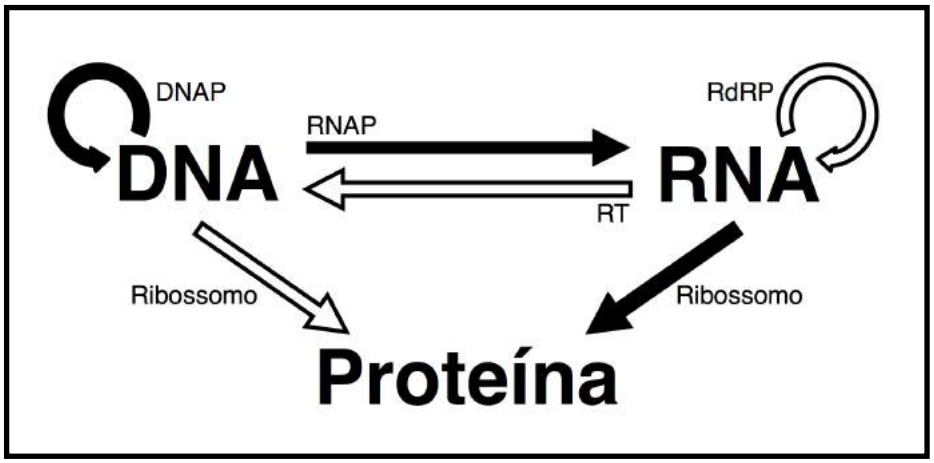
Francis Crick formulou em 1958 seu icônico *Dogma Central da Biologia Molecular*, sugerindo um quadro no qual as proteínas seriam resultado da transferência das informações inicialmente presentes na molécula de DNA através de uma molécula intermediária, o *ácido ribonucleico*, RNA. Durante a *transcrição*, parte da informação contida no DNA seria transmitida para uma molécula de RNA e em seguida essa informação seria transferida para uma cadeia de aminoácidos durante o processo de *tradução*. Além dessas transferências, a *replicação* da informação presente no DNA para uma nova molécula de DNA garantiria a manutenção das características hereditárias durante a divisão celular. Em 1970, Crick publicou uma versão atualizada de seu *Dogma*, incluindo novas transferências observadas em condições especiais: a replicação do RNA, a transferência da informação contida em moléculas de RNA para DNA (*transcrição reversa*) e até mesmo a tradução direta da informação contida no DNA (CRICK et al., 1970). A Figura 2 apresenta o quadro geral dessa versão.

Figura 1 – Esquema da estrutura do DNA



Legenda: O DNA é um polímero de *nucleotídeos*, estruturas compostas por uma base nitrogenada, uma pentose, e um grupo fosfato. A pentose presente em sua composição é denominada desoxirribose e suas bases nitrogenadas canônicas são a *citosina* (C), a *guanina* (G), a *adenina* (A) e a *timina* (T). Organiza-se em duas fitas complementares, de acordo com o pareamento proposto por Watson & Crick: Adenina e Timina interagem através de duas ligações de hidrogênio, e Citosina e Guanina por meio de três ligações de hidrogênio. Desta forma, a interação entre C/G é mais intensa que A/T. Destaca-se que a dupla fita é anti-paralela, ou seja, o radical OH, presente no carbono 3′ da pentose, e o grupo fosfato, presente no carbono 5′, localizam-se em extremidades opostas. Toma-se essa referência quando se cita uma determinada sequência de nucleotídeos: 5′–AATTCCGG–3′, por exemplo. Fonte: Modificado de User:Sponk / *Wikimedia Commons* / CC-SA-1.0 (2016) e de Madeleine Price Ball / *Wikimedia Commons* / CC-SA-1.0 (2016)

Figura 2 – Esquema do *Dogma central da biologia molecular*



Legenda: Conforme enunciado por Crick em 1970. Cada seta representa uma reação, com sua respectiva enzima indicada. Setas pretas representam as transferências usuais, enquanto as setas brancas indicam as reações que podem ocorrer em condições especiais (CRICK et al., 1970). DNAP: DNA polimerase DNA-dependente; RNAP: RNA polimerase DNA-dependente; RT: singla em inglês para transcriptase reversa, também chamada de DNA polimerase RNA-dependente; RdRP: sigla em inglês para RNA polimerase RNA-dependente. Fonte: Produzido pelo próprio autor

O DNA foi o principal ácido nucleico estudado até o início do século XXI, culminando com o *Projeto Genoma Humano*. Denomina-se *genoma* o conjunto das informações hereditárias presentes nos organismos, resultado da *sequência* das bases presentes em seu material genético. A definição inicial de *gene*, proposta no início do século XX, passou a ser considerada inadequada com o avanço da biologia molecular; atualmente classifica-se como gene qualquer região localizável no genoma que possa ser transcrita e associada a pelo menos uma região regulatória (PEARSON, 2006). Os *promotores* e os *acentuassomos* são exemplos de elementos reguladores, sendo que os primeiros atuam como sítio de ligação da enzima responsável pela transcrição e os últimos são sequências nas quais proteínas podem se ligar e aumentar os níveis de transcrição (GRIFFITHS et al., 2006).

O Projeto Genoma Humano foi concluído com sucesso em 2003 e, dentre outras questões, destacou as semelhanças entre genes presentes em *Homo sapiens* e os genes presentes em outras espécies. Apenas uma pequena porcentagem dos genes de camundongos, por exemplo, não possui uma óbvia contraparte humana.

Além disso, estima-se um número de pares de base e de genes bastante similar entre essas espécies (SCHLICK, 2006). Como essa similaridade gênica não pode ser desconsiderada, atribui-se as evidentes diferenças entre os organismos, pelo menos em parte, à *regulação* desses genes. Jacob e Monod (1961) sugeriram que o RNA poderia controlar a atividade dos genes, mas somente no início do século XXI foi possível estudar esse mecanismo.

1.1 A molécula de RNA

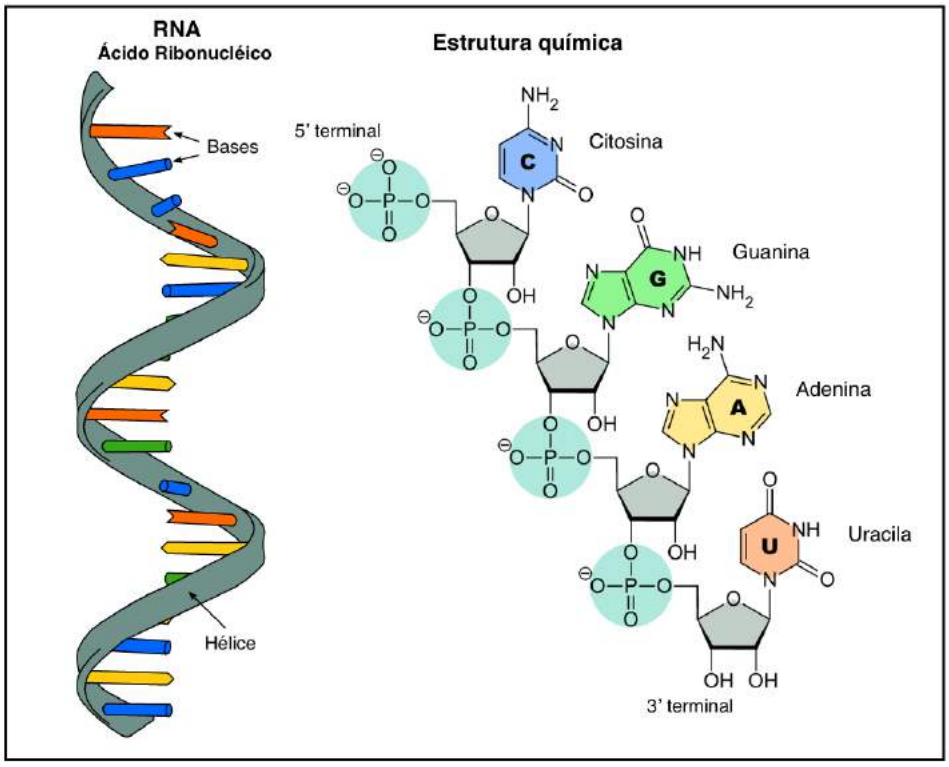
Assim como o DNA, o RNA, representado na Figura 3, é constituído por *nucleotídeos*. Nesse caso, a pentose presente é denominada *ribose*, e suas bases nitrogenadas canônicas são citosina (C), a guanina (G), a adenina (A) e a uracila (U). A função biológica da molécula de RNA é determinada pela sua sequência de nucleotídeos e por sua estrutura após a transcrição. A Tabela 1 apresenta alguns tipos de RNAs conhecidos e suas respectivas funções (GRIFFITHS et al., 2006). Agentes terapêuticos podem ser criados para explorar essas funções e até mesmo novas moléculas de RNA podem ser projetadas para interagir com alvos específicos.

Tabela 1 – Exemplos de moléculas de RNA.

Sequência	Função
RNA mensageiro (mRNA)	Resultado da transcrição de um gene, carrega a informação necessária para a produção de proteínas.
RNA transportador (tRNA)	Transporte dos aminoácidos até os ribossomos para a formação de proteínas.
RNA ribossômico (rRNA)	Principal componente do ribossomo.
Pequenos RNAs nucleolares (snoRNA)	Modificam o rRNA.
Micro RNA (miRNA)	Regulam a tradução das proteínas.
Pequenos RNAs nucleares (snRNA)	Modificação pós-transcricional do RNA (*splicing*).
Ribozimas	Apresentam atividade catalítica.

Fonte – Griffiths et al. (2006)

Figura 3 – Esquema da estrutura do RNA

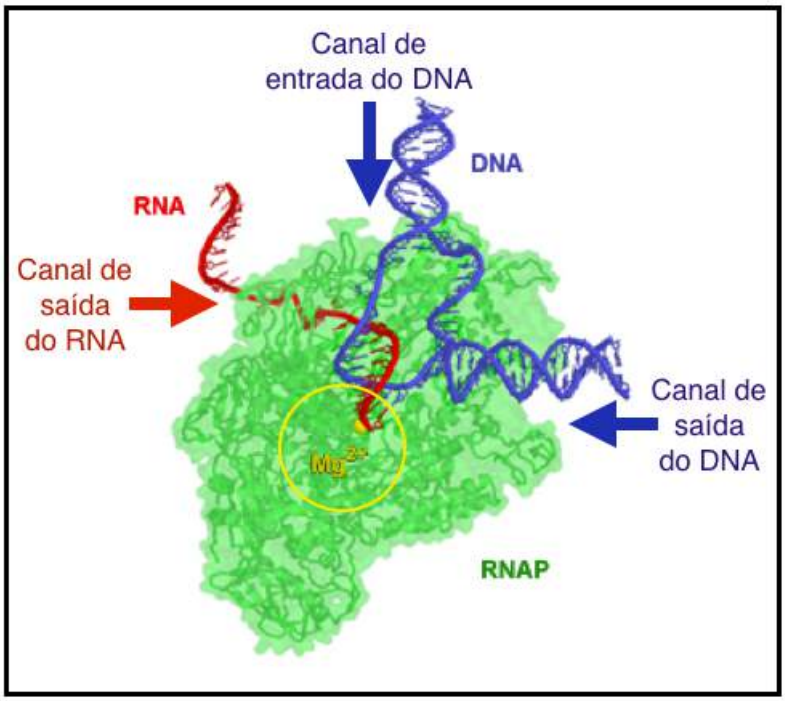


Legenda: O filamento de RNA é bastante semelhante ao DNA, com ligações fosfodiéster ocorrendo nas posições 5′ e 3′ do açúcar. Entretanto, apresenta-se em fita única e a base Uracila substitui a base Timina. Fonte: Modificado de User:Sponk / *Wikimedia Commons* / CC-SA-1.0 (2016)

1.2 O processo de transcrição

A primeira etapa do *Dogma Central* de Crick trata da transferência das informações contidas na molécula de DNA para a molécula de RNA. Esse processo é conhecido como *transcrição* e é vinculado à atividade RNA polimerase DNA-dependente (RNAP, representada na Figura 4). O desenvolvimento de técnicas de molécula única — como a utilização de pinças óticas e magnéticas, a microscopia de força atômica e a fluorescência de molécula única — aumentaram a compreensão de todas as etapas do processo, complementando estudos bioquímicos tradicionais (HERBERT; GREENLEAF; BLOCK, 2008). Suas características gerais são conservadas para todas as formas de vida celular, apresentando apenas diferenças de caráter estrutural na enzima responsável, além de diversas formas alternativas de regulação.

Figura 4 – Esquema da RNAP



Legenda: RNAP realizando a transcrição de um segmento de DNA. O híbrido RNA/DNA se acomoda no chamado *canal principal* durante o processo. Estão apontados os canais de entrada e saída do DNA e o canal de saída do RNA recém transcrito. O sítio ativo da enzima localiza-se na região circulada, onde se observa a presença de íons Mg^{2+}. A Figura apresenta a estrutura da enzima encontrada na bactéria *Thermus aquaticus*, porém a estrutura básica da RNAP conserva-se entre os organismos. Fonte: Modificado de User:Abbondanzieri / *Wikimedia Commons* / Domínio Público (2016)

Etapas da transcrição

O processo de transcrição tem início quando a RNAP identifica na fita de DNA a chamada *região promotora*. A polimerase ancora nessa região na presença de *fatores de transcrição*, caracterizando assim a primeira etapa da transcrição, denominada de *iniciação*. Nessa etapa, a enzima rompe as ligações entre as fitas de DNA, expondo a *fita molde* que será transcrita. A conformação resultante é denominada de *bolha de transcrição*.

Quando a RNAP finalmente abandona a região promotora, a transcrição dos primeiros nucleotídeos leva à formação do chamado *Complexo de Alongamento Transcricional*, CAT, constituído pela enzima, pelo DNA e pela fita de RNA nascente. O

movimento do CAT ao longo da fita de DNA enquanto novos ribonucleotídeos são incorporados à cadeia de RNA caracteriza a fase de *alongamento*. Idealmente, cada ribonucleotídeo (r) acrescentado dependerá do desoxirribonucleotídeo (d) presente no sítio ativo da RNAP, interagindo através de ligações de hidrogênio de acordo com os pareamentos de Watson-Crick: dA/rU, dT/rA, dC/rG e dG/rC. A transcrição segue no sentido $3' \rightarrow 5'$ da fita molde de DNA e resulta em uma fita de RNA $5' \rightarrow 3'$, devido à conformação das ligações entre as bases.

Caso não abandone prematuramente da fita de DNA, a RNAP seguirá com o alongamento até reconhecer o *sítio de terminação* na fita molde, iniciando a fase de *terminação*. Ocorre, então, a separação do híbrido RNA-DNA, a liberação da molécula de RNA pela RNAP e o colapso da bolha de transcrição, levando à formação da estrutura estável do DNA em fita dupla. Esse processo pode ocorrer de diferentes maneiras, ainda não totalmente esclarecidas (GREIVE; HIPPEL, 2005).

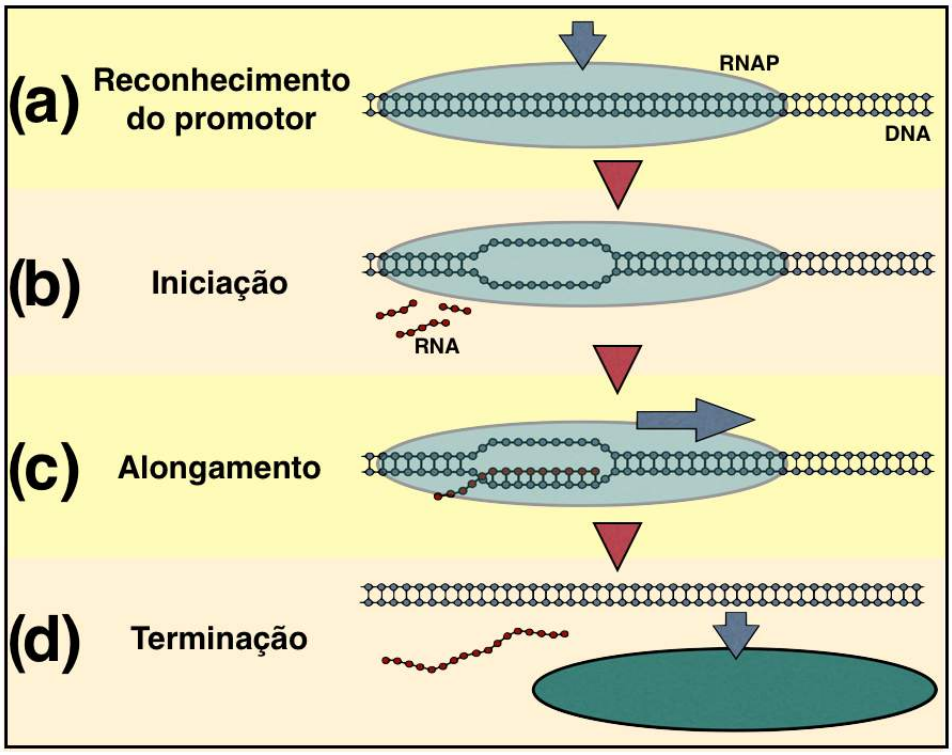
A Figura 5 apresenta um quadro geral das etapas apresentadas.

Pausas transcricionais

Devido à interações entre os elementos constitutivos do CAT, ou mesmo à interferência de uma molécula externa, a RNAP costuma apresentar um comportamento cinético irregular durante o processo de alongamento da sequência de RNA. *Pausas transcricionais* são eventos frequentes e podem não apenas diminuir a taxa de produção de RNA, como também permitir que fatores reguladores atuem no CAT, modificando a transcrição subsequente (ARTSIMOVITCH; LANDICK, 2002; RING; YARNELL, 1996).

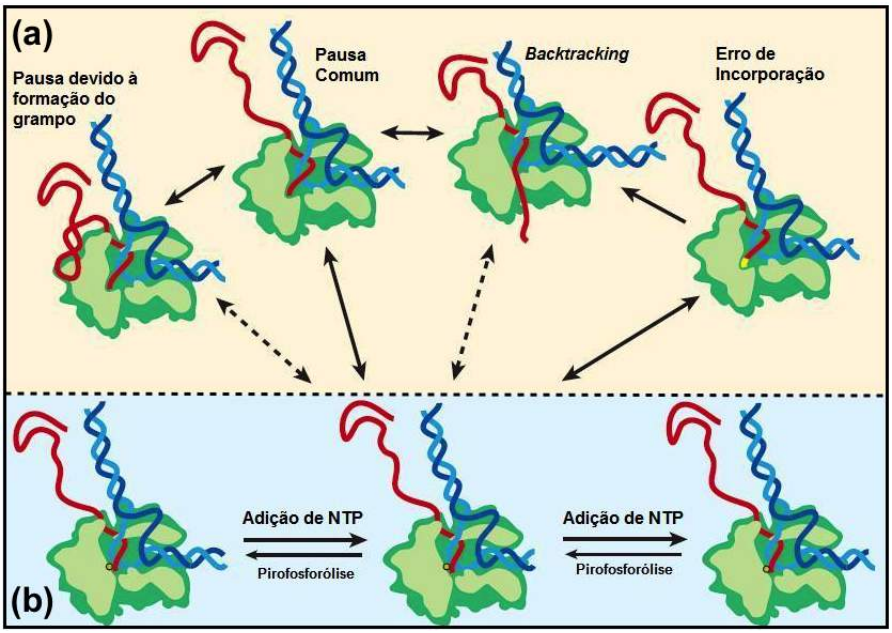
Essas pausas podem ser classificadas de acordo com o mecanismo envolvido e efeitos subsequentes à sua ocorrência. São três classes: as chamadas *pausas elementares*, *comuns*, ou do *tipo 1*, as pausas resultantes do *backtracking*, ou pausas do *tipo 2*, e pausas ligadas à formação de grampos (ZHANG; LANDICK, 2016). A Figura 6 ilustra esses eventos, apresentando-os como uma via alternativa à via principal da transcrição.

Figura 5 – Esquema das etapas da transcrição



Legenda: (a) Reconhecimento do promotor: Antes de iniciar o processo, a RNAP encontra um dos sítios promotores e se liga ao DNA em dupla fita. (b) Iniciação: após expor a fita de que servirá de modelo e abrir a *bolha de transcrição*, a RNAP produz pequenos filamentos de RNA, liberando-os em seguida, mas sem abandonar o promotor. (c) Fase de alongamento: cadeias de RNA são polimerizadas por incorporação de um ribonucleotídeo complementar ao desoxirribonucleotídeo presente no sítio ativo da enzima. Após a reação, a RNAP se move um nucleotídeo no sentido $3' \rightarrow 5'$ na fita molde, liberando o sítio ativo e permitindo uma nova polimerização. (d) Terminação: reconhecimento do sítio de terminação, desestruturação do CAT, liberação do transcrito produzido e abandono da fita de DNA. Fonte: Produzido pelo próprio autor

Figura 6 – Esquema dos processos alternativos da fase de alongamento



Legenda: (a) Processos alternativos. (b) Alongamento transcricional. Setas contínuas representam processos com maior probabilidade de ocorrência em relação às tracejadas. Fonte: Modificado de Herbert et al. (2006)

As *pausas elementares* são resultado de reorganizações conformacionais ainda não completamente compreendidas que interferem no sítio ativo da enzima, dificultando a incorporação do próximo ribonucleotídeo. A ocorrência de uma pausa elemental geralmente precede tanto a pausa devido ao *backtracking* como as pausas relacionadas aos grampos (ZHANG; LANDICK, 2016).

Backtracking é o movimento da RNAP no sentido $5' \rightarrow 3'$ da fita molde, *recuando* algumas bases. Além de pausar o alongamento da fita de RNA por longos intervalos de tempo, é considerado um mecanismo de verificação (*proofreading*) da transcrição. Durante esses eventos, a fita de RNA atravessa novamente os canais da RNAP e ribonucleotídeos incorporados incorretamente podem desestabilizar o CAT. Se identificado o erro, todo o segmento transcrito a partir de sua posição é clivado, e a transcrição reiniciada (HERBERT et al., 2006).

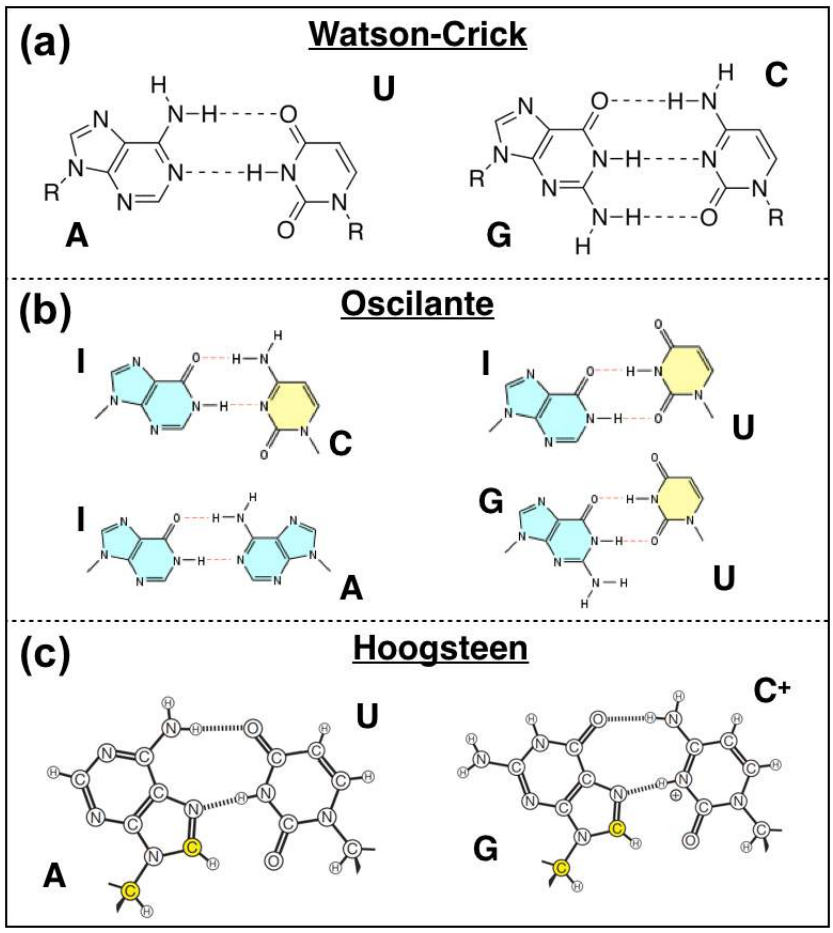
Finalmente, o RNA recém transcrito que abandonou a bolha de transcrição pode assumir uma conformação em grampo, estrutura resultante do pareamentos entre suas próprias bases nitrogenadas (mais detalhes na Seção seguinte).

Embora essas estruturas não interfiram obrigatoriamente na taxa de transcrição, um grampo recém formado pode interagir com a RNAP numa pausa elementar, alterando ainda mais a estrutura da enzima e estendendo o intervalo de tempo necessário para o CAT abandonar essa conformação, evento denominado de *pausa estabilizada por grampo*. Entretanto, um grampo pode também colaborar para reduzir intervalos de pausa do tipo 1 ou do tipo 2, "puxando" os ribonucleotídeos necessários para sua configuração mais estável e induzindo o movimento da RNAP no sentido da transcrição; grampos podem até mesmo inibir o *backtracking*, atuando como uma "barreira" que impede o movimento da enzima no sentido oposto ao da transcrição (ZHANG; LANDICK, 2016).

1.3 O dobramento da molécula de RNA

O RNA se apresenta em fita única, mas é capaz de se acomodar em estruturas estáveis através de interações entre suas próprias bases nitrogenadas. Esse potencial foi explorado durante toda a evolução, com destaque na estrutura do RNA transportador e do ribossomo. Na Figura 7 estão representados as principais interações observadas experimentalmente: (a) os pareamentos *canônicos de Watson-Crick*, (b) os pareamentos *oscilantes* e (c) as *interações de Hoogsteen*. No pareamento canônico, a adenina liga-se à uracila (A/U) através de duas ligações de hidrogênio, enquanto a guanina interage com a citosina (G/C) através de três ligações de hidrogênio. As interações entre guaninas e uracilas (G/U), tão comuns no RNA quanto o pareamento canônico G/C, são resultado do pareamento oscilante. Os pareamentos entre a base inosina (I), derivada de uma adenina, e três outras bases nitrogenadas (U, C, A) também ocorrem através desse tipo de interação. Por fim, as interações de Hoogsteen, que surgem devido à rotação das bases adenina e guanina, podem resultar em ligações envolvendo 3 e até mesmo 4 bases nitrogenadas.

Figura 7 – Esquema dos pareamentos entre bases nitrogenadas

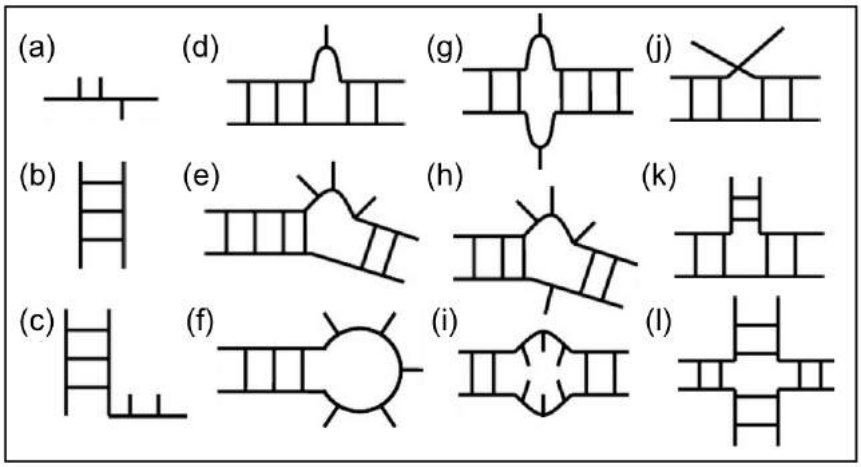


Legenda: (a) Pareamento Watson-Crick: A/U, C/G. (b) Pareamento Oscilante. (c) Pareamento Hoogsteen. Fonte: Modificado de User:Jypx3~commonswiki / *Wikimedia Commons* / Domínio público, de User:Fdardel / *Wikimedia Commons* / CC-SA-1.0 (2016) e de Nikolova et al. (2011)

Além da representação "gráfica" desses pareamentos, a notação de pontos-e-parênteses, em inglês *dot-bracket*, é frequentemente utilizada. Nessa notação, geralmente apresentada em fontes monoespaçadas, os pontos representam as posições não pareadas, enquanto os parênteses representam a posição dos pareamentos, identificados partindo-se do par mais interno para o mais externo. Por exemplo, a notação "`..(((...)))`" indica que há pareamento entre as bases das posições $5-9$, $4-10$ e $3-11$. Caso a utilização dos parênteses possa apresentar ambiguidade, outros símbolos com significância equivalente são utilizados, como os colchetes ("`[`" e "`]`").

O resultado direto dos pareamentos entre bases são as chamadas *estruturas secundárias*, apresentadas na Figura 8. Essas estruturas contribuem significantemente na estabilidade global da molécula de RNA (BRION; WESTHOF, 1997; TINOCO; BUSTAMANTE, 1999) e são definidas pela associação entre seus dois componentes fundamentais: os *Segmentos Unifilamentares de RNA*, SURs, e os *Segmentos Bifilamentares de RNA*, SBRs, resultantes do pareamento entre dois segmentos unifilamentares complementares inversos de uma mesma molécula de RNA.

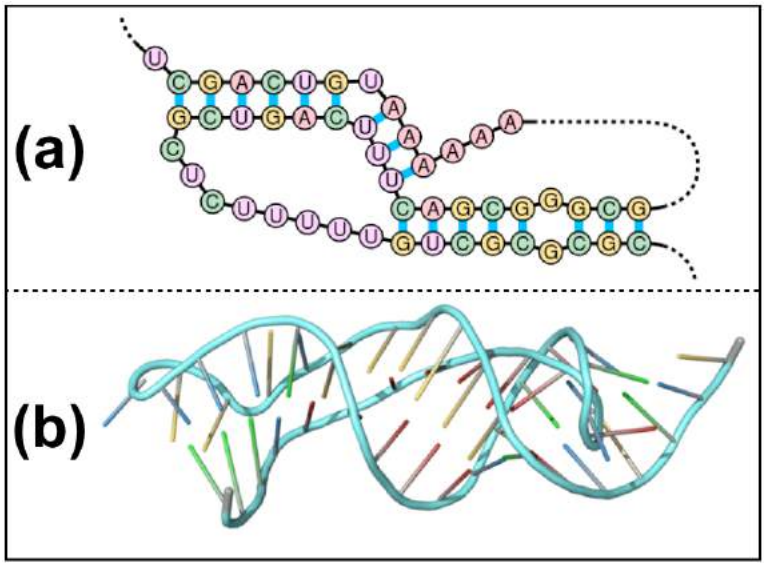
Figura 8 – Esquema das estruturas secundárias derivadas do pareamento entre as bases da molécula de RNA



Legenda: (a) Seguimento unifilamentar. (b) Seguimento bifilamentar. (c) Segmento bifilamentar com extensão terminal. (d) Protuberância composta por um nucleotídeo. (e) Protuberância composta por um SUR. (f) Grampo: SBR e alça composta por um SUR. (g) Par de bases incompatíveis interno. (h) SURs internos assimétricos. (i) SURs internos simétricos. (j) Entroncamento entre dois segmentos bifilamentares. (k) Entroncamento entre três segmentos bifilamentares. (l) Entroncamento entre quatro segmentos bifilamentares. Fonte: Blossey (2006)

As *estruturas terciárias* são resultado das interações entre duas ou mais estruturas secundárias. A conformação mais comum é o *pseudonó do tipo-H*, representado na Figura 9, resultado da interação entre a alça de um grampo e uma outra região unifilamentar da mesma molécula de RNA. Essa estrutura está presente nos mecanismos de replicação viral (BRIERLEY; PENNELL; GILBERT, 2007; BRIERLEY; GILBERT; PENNELL, 2008), em *riboswitches* sensíveis a metabólitos (KANG; PETERSON; FEIGON, 2009; KLEIN; EDWARDS; FERRÉ-D'AMARÉ, 2009) e é essencial para a atividade da RNA telomerase humana (CHEN; GREIDER, 2005; QIAO; CECH, 2008).

Figura 9 – Esquema da estrutura de um pseudonó do tipo-H



Legenda: Pseudonó do tipo-H presente na fração de RNA da telomerase humana. (a) Estrutura planificada. (b) Estrutura tridimensional. Fonte: Modificado de User:Sakurambo / *Wikimedia Commons* / CC-BY-SA-3.0 (2016) e de User:Fdardel / *Wikimedia Commons* / CC-BY-SA-3.0 (2016)

A determinação experimental das estruturas dos RNAs está restrita a dispendiosos métodos, como a cristalografia de raios X e ressonância magnética nuclear; além disso, limita-se a pequenas estruturas, uma vez que sistemas mais complexos possuem uma flexibilidade molecular significativa. O desenvolvimento de programas voltados para a simulação do dobramento das moléculas de RNA, amplamente difundidos na comunidade científica, permitiu a análise eficiente tanto de estruturas simples como de conformações complexas, partindo-se apenas de sua sequência de nucleotídeos.

1.4 Modelos computacionais para a molécula de RNA

A caracterização da conformação espacial de polímeros continua atraindo pesquisadores das mais diferentes áreas há várias décadas. O principal alvo dessa dedicação envolve a determinação das estruturas terciárias das *proteínas*, devido à sua aplicação direta no desenvolvimento de fármacos e para a compreensão dos

mecanismos inter e extracelulares. Porém, parte dessas equipes, além de novos pesquisadores, voltaram seus esforços para a compreensão do dobramento das moléculas de RNA. O desenvolvimento de novas ferramentas computacionais e a evolução na capacidade de processamento das máquinas permitiram que os estudos *in silico* dos ácidos nucleicos complementassem os dados experimentais obtidos nas últimas décadas. Simulações com precisão aceitável, tratáveis apenas em super computadores há mais de duas décadas, são realizadas em questão de horas em computadores pessoais modernos. Algoritmos e campos de força para várias escalas buscam assimilar os detalhes atômicos e as interações de longo alcance característicos de grandes estruturas moleculares.

Se compararmos uma molécula de RNA e uma proteína com mesmo número de resíduos, observaremos que o espaço de busca conformacional para o RNA será muito maior, uma vez que cada resíduo possui seis ângulos diedros, enquanto as proteínas possuem apenas dois. Porém, a variedade de resíduos favorece a simulação do RNA: são apenas quatro nucleotídeos com estrutura química bastante semelhante, ao passo que as proteínas apresentam 20 aminoácidos com grande diversidade de funções químicas. Além disso, o RNA possui regras de pareamento bastante claras, algo sem equivalente para o caso das proteínas. Essas interações muitas vezes correspondem à maior contribuição energética para essas moléculas, e consequentemente levam à formação *hierárquica* de estruturas. Assim, podemos desconsiderar as interações terciárias numa primeira aproximação: na verdade, uma previsão de alta qualidade dos elementos secundários costuma ser o primeiro passo para a modelagem tridimensional desses polímeros (TINOCO; BUSTAMANTE, 1999).

A dinâmica das moléculas de RNA pode ser classificada em dois tipos: (I) *flutuações em equilíbrio*, relacionadas com movimentos termais e correspondem aos "saltos" espontâneos que podem ocorrer entre as conformações possíveis na superfície de energia livre e (II) *transições conformacionais*, nas quais estímulos celulares, como o aumento de um determinado metabólito, criam estados de equilíbrio instáveis. Com a superfície de energia livre alterada, ocorre uma redistribuição dos estados conformacionais (DETHOFF et al., 2012).

A maior parte dos usuários hoje busca identificar a estrutura de mínima energia livre (*MFE*, do inglês *Minimum Free Energy*) para uma dada sequência. Diferentes estratégias podem ser utilizadas para esse fim: a Tabela 2 apresenta alguns exemplos de programas disponíveis, com uma breve descrição dos princípios utilizados em cada caso.

Tabela 2 – Programas para determinação da estrutura de mínima energia livre.

Programa	Descrição
RNAfold	Parte do pacote *ViennaRNA*, baseia-se em algoritmos de programação dinâmica desenvolvidos originalmente por Zuker e Stiegler (1981). O algoritmo para determinação da função partição é baseada no trabalho de McCaskill (1990). Desconsidera pseudonós. (GRUBER et al., 2008)
Sfold	Realiza uma amostragem estatística de todas as estruturas possíveis, com pesos determinados através de uma função partição de probabilidades. Desconsidera pseudonós. (DING; CHAN; LAWRENCE, 2004)
IPknot	Prevê a estrutura com máxima exatidão esperada utilizando *programação inteira* para o limiar de corte. Inclui pseudonós. (SATO et al., 2011)
Vfold	Modelos físicos baseados numa representação de baixa resolução da estrutura do RNA. Inclui pseudonós. (XU; CHEN, 2015)
ContextFold	Utiliza modelos baseados em aprendizagem computacional, com mais de 70 mil parâmetros livres. Desconsidera pseudonós. (ZAKOV et al., 2011)

Fonte – Produzido pelo próprio autor

Essas ferramentas computacionais possuem grande potencial e colaboram significativamente na compreensão das interações das moléculas de RNA no ambiente celular. Entretanto, determinar apenas a estrutura MFE pode não ser suficiente. A relevância biológica de uma molécula de RNA pode estar associada a uma "família" de estruturas subótimas. Além disso, pequenas discrepâncias nos parâmetros termodinâmicos experimentais utilizados também podem "mascarar" a estrutura MFE, efeito diretamente proporcional ao comprimento da sequência de estudo.

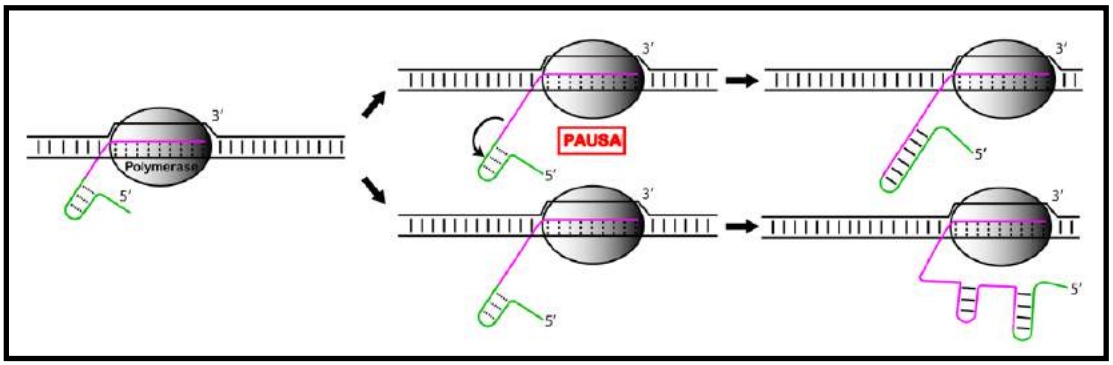
1.5 Proposta

A interação entre moléculas no interior de uma célula está longe de ser trivial. O RNA, hoje considerado ubíquo em vias essenciais, pode ser perturbado por íons, chaperonas, outros RNAs, enzimas *Dicer* e outras macromoléculas, mas eventos intrínsecos que ocorrem durante sua formação também são pertinentes.

Neste trabalho, propomos o desenvolvimento de um programa para simulação e análise das estruturas formadas durante o processo de alongamento transcricional. Os métodos citados na Seção anterior simulam apenas a "renaturação" da molécula em estudo, desconsiderando os efeitos da *cinética* da transcrição na conformação dessas moléculas.

A superfície de energia livre das estruturas de RNA é tão acidentada que caminhos distintos de dobramento podem levar a diferentes intermediários estruturais estáveis. O trabalho de Xayaphoummine et al. (2007) apresentou uma sequência de RNA que pode tomar diferentes conformações variando-se apenas o sentido de sua transcrição. A Figura 10 ilustra a competição entre estruturas alternativas, mediadas pela posição e pela intensidade das pausas transcricionais (ZHANG; LANDICK, 2016).

Figura 10 – Esquema do efeito do dobramento cotranscricional

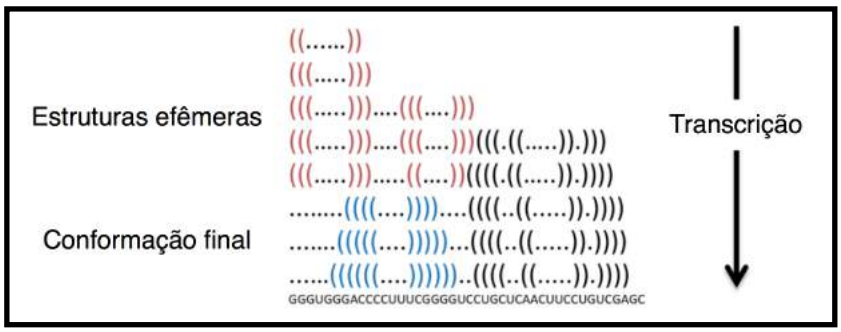


Legenda: Devido à variedade de conformações estáveis para a estrutura da molécula de RNA, seu dobramento pode ocorrer de forma distinta na presença de pausas transcricionais. Fonte: Modificado de Al-Hashimi e Walter (2008)

Essas conformações alternativas podem ser tão estáveis que se mantém ao final da transcrição ou se apresentam apenas como *estruturas efêmeras*, se desfazendo em favor de arquiteturas mais estáveis ao longo do processo, como representado

na Figura 11. Bacteriófagos do gênero Levivírus possuem um gene cuja transcrição resulta numa sequência de RNA que apresenta, em sua estrutura final, pareamentos envolvendo a região de ligação do ribossomo, tornando-a intraduzível. Porém, durante o alongamento, um pequeno SBR efêmero posterga a formação dessa estrutura inibitória, fornecendo uma janela de tempo na qual o ribossomo poderá acessar a sequência (MEERTEN; GIRARD; DUIN, 2001).

Figura 11 – Esquema da formação e transfiguração de estruturas efêmeras



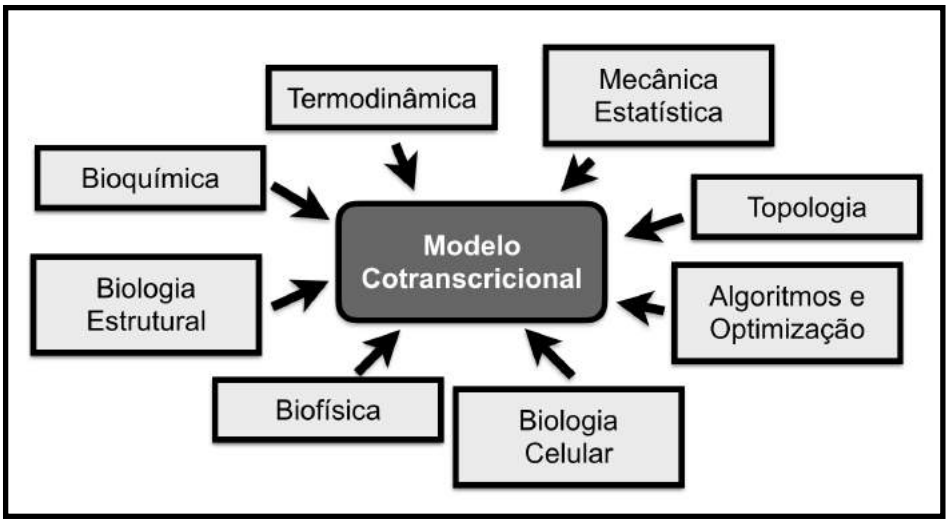
Legenda: Conforme a transcrição ocorre, novos pareamentos passam a ser permitidos. Caso sejam energeticamente favoráveis, as estruturas efêmeras se desfazem em favor dessas novas configurações. Fonte: Modificado de Geis et al. (2008)

Vários RNAs possuem sequências capazes de autoinduzir transições formadas dinamicamente durante a transcrição (NAGEL; PLEIJ, 2002). Mesmo as estruturas efêmeras podem interagir com outras moléculas ou mesmo com a RNAP e alterar o comportamento da transcrição na região onde se manifestam (ZHANG; LANDICK, 2016). O genoma do Vírus da Hepatite D é conhecido por apresentar uma ribozima com atividade catalítica autoclivadora fundamental para sua replicação. Chadalavada et al. (2000) identificaram que essa atividade pode ser moderada pela formação cotranscricional de estruturas alternativas na região que antecede o sítio de clivagem. A transcrição da região imediatamente anterior a esse sítio resultará numa conformação que suprime a autoclivagem; entretanto, se considerarmos a transcrição da região ainda mais preliminar, a formação de um grampo que contém o segmento que favoreceria o surgimento da estrutura inibidora do primeiro caso, garantirá a manifestação da estrutura com capacidade de autoclivagem dessa sequência. Também podemos citar o operon que codifica as proteínas ligadas à biossíntese do aminoácido triptofano em bactérias: a região inicial desse operon codifica

triptofanos *in tandem*, logo sua expressão estará ligada à abundância desse aminoácido. Durante a transcrição, duas estruturas alternativas, denominadas *terminadora* e *antiterminadora*, formam-se cotranscricionalmente, devido à presença dos ribossomos na fita de RNA recém transcrita. Quando há deficiência no fornecimento de triptofano, o ribossomo ocupa a região que os codifica, promovendo a formação da estrutura antiterminadora, permitindo que a maquinaria transcricional continue a transcrição desse operon; quando o triptofano apresenta-se em abundância, o ribossomo continua a tradução através da região inicial, promovendo a formação da estrutura terminadora. Essa estrutura desestabiliza o complexo de alongamento transcricional, abortando prematuramente o processo.

Destacamos a complexidade envolvida tanto na análise do processo de dobramento cotranscricional, como na sua modelagem e simulação. Como observa-se na Figura 12, trata-se de uma área transdisciplinar, envolvendo tanto conceitos teóricos como experimentais das ciências exatas e biológicas.

Figura 12 – Esquema das áreas envolvidas no estudo do processo de dobramento cotranscricional



Legenda: Conceitos e ferramentas de diferentes áreas de pesquisa são utilizados no desenvolvimento de um modelo dessa natureza. Fonte: Produzido pelo próprio autor

Em nossa aproximação, cada transição entre conformações do RNA nascente é modelada como uma reação química, com os valores para as taxas de ocorrência baseados na variação da energia livre de Gibbs entre os estados, ponderados através de uma distribuição de probabilidade de Boltzmann. O código inclui ainda a atualização de um algoritmo estocástico para a simulação do processo de alongamento transcricional, afim de realizar uma previsão inteiramente *in silico* dos principais eventos do processo.

Um modelo refinado para estudo do comportamento da molécula de RNA e do processo de transcrição pode contribuir no estudo de *riboswitches*, ribozimas e de novas moléculas de RNA não codificantes cada vez mais evidentes nos processos celulares descobertos nos últimos anos. Além disso, pode colaborar no desenvolvimento de técnicas de engenharia de moléculas de RNA, problema inverso complexo e de grande interesse e aplicabilidade nas ciências da saúde.

2 Objetivos

Os objetivos desse trabalho foram:

a) Implementar um algoritmo para simulação do dobramento cotranscricional da molécula de RNA;

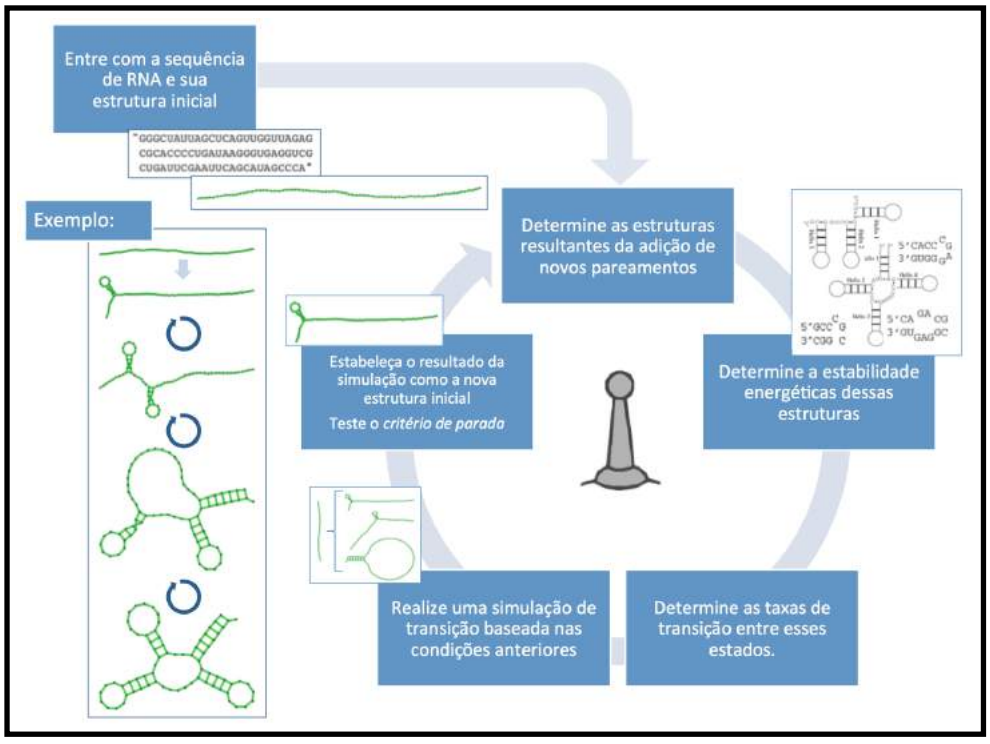
b) Estudar os efeitos da formação de estruturas efêmeras e permanentes no comportamento do alongamento transcricional utilizando o programa desenvolvido;

c) Estudar os efeitos da cinética da transcrição na formação de estruturas efêmeras e permanentes utilizando o programa desenvolvido.

3 Metodologia

Para compreender o desenvolvimento estrutural da molécula de RNA, elaboramos uma série de funções em *Wolfram Language* e organizadas no programa *Mathematica* 10.0. Resumidamente, essas funções discriminam quais são os pareamentos permitidos a partir de uma dada estrutura e determinam os valores de taxa de transição entre os estados resultantes da adição desses novos pareamentos e o estado atual, baseando-se em sua respectivas estabilidades energéticas. Então, aplicando conceitos de simulação estocástica, é possível determinar qual a próxima estrutura no "caminho" de dobramento da molécula de RNA em estudo. A Figura 13 apresenta o quadro simplificado do funcionamento do programa,batizado como *Bender*.

Figura 13 – Esquema simplificado do funcionamento do programa



Legenda: Um exemplo de resultado após quatro iterações está representado à esquerda: trata-se da clássica estrutura em "trevo" de um RNA transportador. No centro, o logotipo do programa. Fonte: Produzido pelo próprio autor

3.1 Pareamentos entre bases

Antes de tratarmos especificamente das estruturas, é necessário identificar sua unidade básica, os SBRs, derivados do pareamento entre as bases que constituem a sequência em estudo. Entretanto, estima-se que o número de pareamentos possíveis seja da ordem de $1{,}866^N$, onde N equivale ao comprimento da sequência (ZUKER; SANKOFF, 1984), relação que torna inviável uma implementação baseada simplesmente em busca exaustiva para identificação dos SBRs em sequências longas.

Algumas regras restritivas foram impostas para reduzir o espaço de possibilidades:

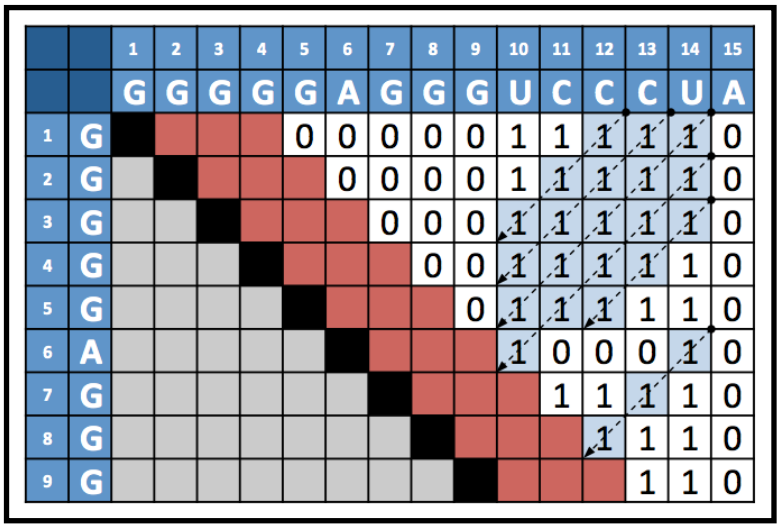
a) apenas o pareamento oscilante G/U e os pareamentos canônicos de Watson-Crick são permitidos;

b) os SBRs devem ser compostos por pelo menos três pares de bases;

c) os grampos constituídos por alças com menos de três nucleotídeos são desconsiderados, restrição válida devido aos efeitos estéricos nessas conformações.

Para a identificação desses pareamentos, implementamos uma abordagem baseada no método de programação dinâmica de Nussinov e Jacobson (1980): inicialmente, constrói-se uma matriz $\mathbf{M}$, na qual a primeira linha e a primeira coluna representam as bases encontradas na sequência em estudo. As entradas restantes recebem "1", caso o pareamento entre as bases correspondentes à suas coordenadas seja permitido, ou "0" caso contrário. Como a terceira restrição estabelece que o comprimento mínimo de um alça em um grampo é de três nucleotídeos, apenas a região acima dessa diagonal é preenchida, ou seja, somente os elementos $m_{i,j}$ com $j > i + 3$. Parte-se, então, do extremo superior direito da matriz, percorrendo-se as diagonais em busca de sequências formadas por pelo menos três "1", atendendo à segunda restrição. Quando encontradas, as coordenadas iniciais e finais dessa diagonal são identificadas. Caso seu comprimento seja maior que três, todos os SBRs contidos no segmento principal e compostos por pelo menos três elementos também são especificados. A Figura 14 apresenta um exemplo de construção dessa matriz.

Figura 14 – Esquema da matriz para determinação dos SBRs

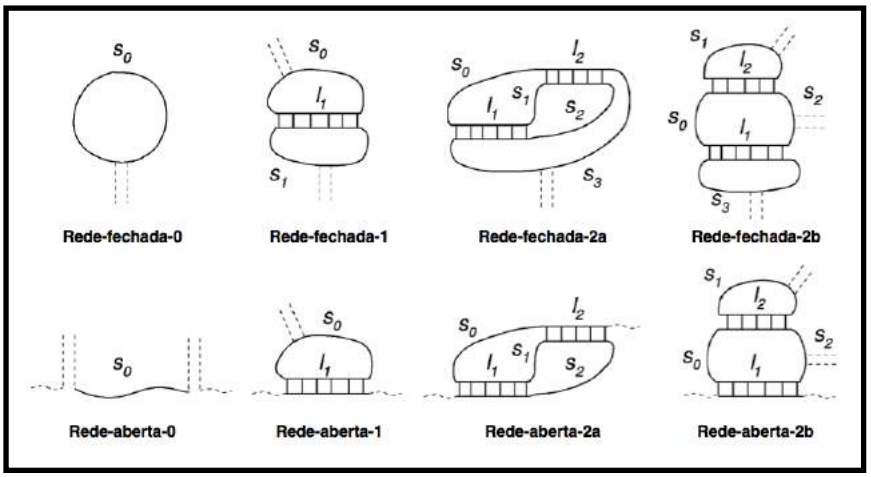


Legenda: Abordagem baseada no método de programação dinâmica de Nussinov. Temos a diagonal principal da matriz em preto. Os elementos abaixo dela não são preenchidos devido à evidente simetria da matriz $\mathbf{M}$. As três diagonais acima também são compostas por elementos nulos devido à restrição imposta em relação ao comprimento mínimo para alças de grampos. Os SBRs com pelo menos três pares de bases estão destacadas e atravessadas por setas que representam a direção em que as diagonais foram determinadas. Por exemplo, temos o segmento formado pelos pares $\{\{13,1\},\{12,2\},\{11,3\},\{10,4\}\}$ e que contém outros dois SBRs: $\{\{13,1\},\{12,2\},\{11,3\}\}$ e $\{\{12,2\},\{11,3\},\{10,4\}\}$. Fonte: Produzido pelo próprio autor

3.2 Identificação das subestruturas

Conforme apresentado na Introdução, Seção 1.3, as estruturas das moléculas de RNA são definidas de acordo com a distribuição de seus SBRs. Como a estabilidade de dada molécula está diretamente relacionada a essas conformações, a classificação e identificação das *subestruturas* resultantes é o primeiro passo para a determinação de seu respectivo valor de energia livre de Gibbs. Essas subestruturas serão denominadas de *redes*, seguindo as definições propostas por Isambert e Siggia (2000). Na Figura 15 estão representadas as *redes* que apresentam no máximo dois SBRs internos. Essa restrição permite cobrir de maneira bastante satisfatória o espaço de conformações presentes nos bancos de dados, sem sobrecarregar o algoritmo de identificação. Utilizaremos a notação ra_x para *redes abertas* do subtipo x e rf_y para *redes fechadas* do subtipo y. Desenvolvemos um algoritmo para identificação

Figura 15 – Esquema da arquitetura das *redes* com até dois SBRs internos



Legenda: SURs possuem s_n nucleotídeos e são "tensionados" por SBRs internos de comprimento aparente l_m. Fonte: Modificado de Isambert e Siggia (2000)

dessas redes, no qual a fita de RNA é "percorrida" do início ao fim, e um algarismo identificador é atribuído a cada nova transição entre um SUR e um SBR. SURs recebem valores ímpares únicos, enquanto valores pares únicos são atribuídos a ambos os segmentos que compõe um SBR. No final, teremos um vetor **I** composto por valores ímpares e pares intercalados. Cada uma das *redes* apresentadas na Figura 15 possui pelo menos uma sequência par-ímpar própria, denominada *Identificador* e representadas através dos vetores apresentados na Tabela 3. Nessa Tabela, p_j representa um elemento par, enquanto os espaços vazios, "_", indicam um elemento ímpar ou um sub-vetor de elementos ímpares. Algumas *redes* possuem mais de um *Identificador* pois diferentes SBRs iniciais alteram a ordem de leitura dos elementos subsequentes. No total, são 18 *Identificadores*, representados pela notação $\mathrm{Id}_{i,k}$, onde i indica a rede em questão e k a forma que se apresenta. O Algoritmo 3.1 apresenta como a busca desses *Identificadores* é realizada na estrutura de interesse. O código remove os elementos identificados do vetor **I** ao final de cada varredura, une os elementos ímpares sequenciais que surgem após essa remoção e reinicia a busca, até que nenhuma nova estrutura possa ser identificada. Se ainda restarem elementos pares em **I**, descartamos a respectiva conformação por se tratar de uma combinação de SBRs mais complexa do que o permitido. Um exemplo dessa abordagem está representado na Figura 16.

Tabela 3 – Tabela de *Identificadores* para redes

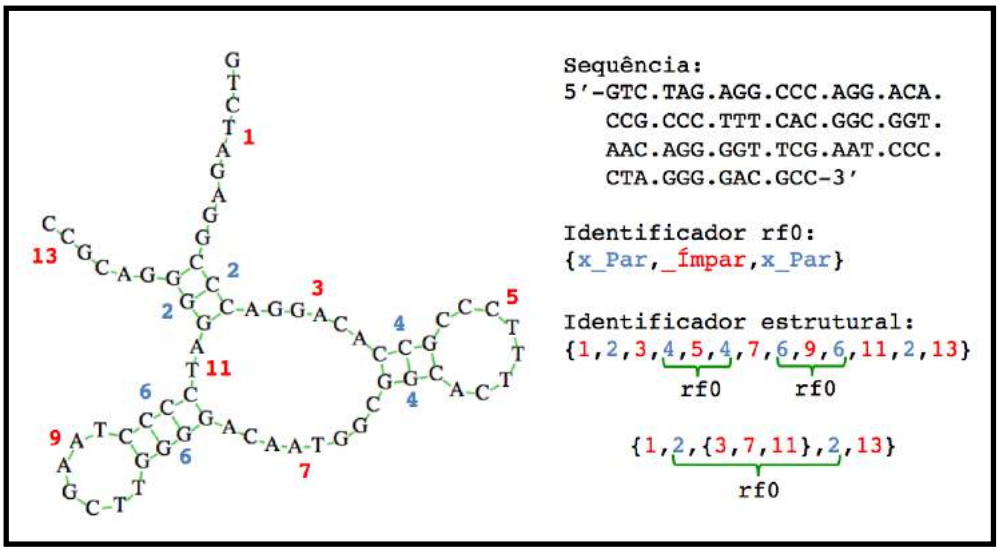
Rede	*Identificadores*, $\mathrm{Id}_{i,k}$
rf_0	$\mathrm{Id}_{1,1}$: $\{p_1, \_, p_1\}$
rf_1	$\mathrm{Id}_{2,1}$: $\{p_1, \_, p_2, \_, p_1, \_, p_3, \_, p_2, \_, p_3\}$
rf_{2a}	$\mathrm{Id}_{3,1}$: $\{p_1, \_, p_2, \_, p_3, \_, p_2, \_, p_3, \_, p_1\}$
rf_{2b}	$\mathrm{Id}_{4,1}$: $\{p_1, \_, p_2, \_, p_1, \_, p_3, \_, p_2, \_, p_4, \_, p_3, \_, p_5, \_, p_4, \_, p_5\}$, $\mathrm{Id}_{4,2}$: $\{p_3, \_, p_2, \_, p_4, \_, p_3, p_1, \_, p_2, \_, p_1, p_5, \_, p_4, \_, p_5\}$, $\mathrm{Id}_{4,3}$: $\{p_1, \_, p_2, \_, p_1, p_5, \_, p_4, \_, p_5, p_3, \_, p_2, \_, p_4, \_, p_3\}$
ra_{2b}	$\mathrm{Id}_{5,1}$: $\{p_1, \_\ p_2, \_, p_1, p_3, \_, p_2, \_, p_4, \_, p_3, p_4\}$, $\mathrm{Id}_{5,2}$: $\{p_1, \_, p_2, \_, p_1, p_4, p_3, \_, p_2, \_, p_4, \_, p_3\}$, $\mathrm{Id}_{5,3}$: $\{p_4, p_1, \_, p_2, \_, p_1, p_3, \_, p_2, \_, p_4, \_, p_3\}$, $\mathrm{Id}_{5,4}$: $\{p_4, p_3, \_, p_2, \_, p_4, \_, p_3, p_1, \_, p_2, \_, p_1\}$, $\mathrm{Id}_{5,5}$: $\{p_3, \_, p_2, \_, p_4, \_, p_3, p_4, p_1, \_, p_2, \_, p_1\}$, $\mathrm{Id}_{5,6}$: $\{p_3, \_, p_2, \_, p_4, \_, p_3, p_1, \_, p_2, \_, p_1, p_4\}$, $\mathrm{Id}_{5,7}$: $\{p_1, \_, p_2, \_, p_1, p_3, \_, p_4, \_, p_3, \_, p_4, p_2\}$, $\mathrm{Id}_{5,8}$: $\{p_3, \_, p_2, \_, p_4, \_, p_3, p_1, \_, p_2, \_, p_1, p_5, \_, p_4, \_, p_5\}$
ra_{2a}	$\mathrm{Id}_{6,1}$: $\{p_1, \_, p_2, \_, p_1, \_, p_2\}$
ra_1	$\mathrm{Id}_{7,1}$: $\{p_1, \_, p_2, \_, p_1, p_2\}$, $\{p_1, p_2, \_, p_1, \_, p_2\}$
ra_0	$\mathrm{Id}_{8,1}$: $\{\_\}$

Fonte – Produzido pelo próprio autor

Algoritmo 3.1: Busca de redes no vetor I

1. Inicialize.
2. Defina $i = 1$ e $k = 1$.
3. Percorra o vetor **I** em busca do padrão $\mathrm{Id}_{i,k}$.
4. Se $\mathrm{Id}_{i,k}$ for encontrado, remova-o do vetor **I**, e una os elementos ímpares em sequência em um único sub-vetor. Continue a busca por $\mathrm{Id}_{i,k}$ em **I**.
5. No final da varredura do vetor **I**, verifique se **I** foi modificado: se sim, volte para o passo 2; senão, verifique se k pode ser incrementado: se sim, incremente k e volte para o passo 3; senão, verifique se $i \neq i_{\max}$, dado $i_{\max} = 8$: se sim, incremente i, defina $k = 1$ e volte para o passo 3; senão, finalize.

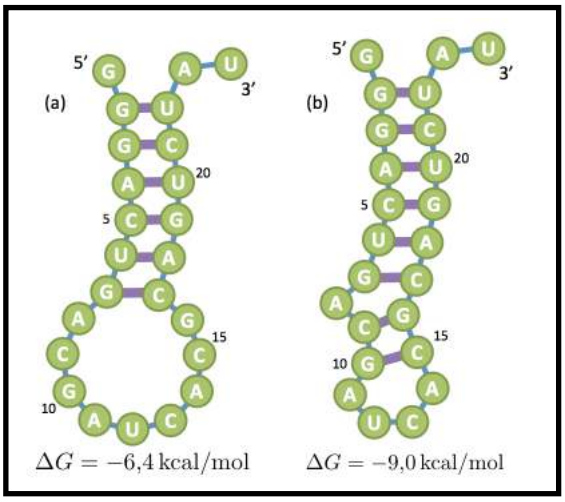
Figura 16 – Esquema de um exemplo de identificação das *redes* através da notação par-ímpar



Legenda: Após a identificação dos SURs, que recebem valores ímpares, e dos SBRs, que recebem valores pares, define-se o vetor **I** com os valores encontrados quando se percorre a fita no sentido $5' \rightarrow 3'$. Em seguida, realiza-se a busca pelos identificadores. O exemplo apresenta inicialmente duas redes-fechadas-0 (rf_0), mas a retirada dessas estruturas do vetor e posterior agrupamento dos SURs restantes permite encontrar mais uma rf_0 no vetor resultante. Fonte: Produzido pelo próprio autor

A última etapa para identificação das SBRs consiste em identificar posições nas quais essas estruturas podem ser "estendidas". Nos bancos de dados e analisando os resultados de outros programas de predição de estruturas, observam-se SBRs formadas por dois ou até mesmo por um único par de base, mas somente em casos onde esses pareamentos distam até um nucleotídeo de SBRs mais estáveis. Esse tipo de conformação pode contribuir para a estabilidade geral da molécula, pois além de sua contribuição energética direta, reduz o comprimento de SURs, como se observa na Figura 17. Após identificar os SBRs principais, o programa verifica nas vizinhanças a ocorrência desses eventos, uma vez que segmentos com menos de três pares de bases não são consideradas na etapa de busca inicial.

Figura 17 – Esquema representando o resultado da extensão de um SBR



Legenda: Comparação entre o valor da energia livre de uma estrutura após a extensão de um SBR. (a) Estrutura inicialmente determinada pelo algoritmo apresentado. (b) SBR estendido, contribuindo para a redução da energia livre final da estrutura, mesmo com a formação de uma protuberância (sítio 8). Sequência: PDB/NDB ID 1RHT (BORER et al., 1995). Fonte: Produzido pelo próprio autor

3.3 Variação da Energia livre de Gibbs para ácidos nucleicos

A próxima etapa do programa deve determinar a *Variação da Energia Livre de Gibbs* de formação de uma dada estrutura a partir da fita de RNA não pareada, ΔG°.

3.3.1 Estruturas secundárias

Classicamente, os valores da energia livre para ácidos nucleicos são estimados a partir da determinação da temperatura na qual metade dos segmentos bifilamentares encontram-se separados em solução, a chamada *temperatura de desnaturação*, T_m (do inglês, *melting temperature*). Para tal, parte-se da reação de desnaturação do segmento em questão,

$$A_1A_2 \rightleftharpoons A_1 + A_2,$$

na qual A_1A_2 representa as fitas ainda unidas, enquanto A_1 e A_2 representam as respectivas fitas separadas. A *constante de equilíbrio* dessa reação será dada por

$$K = \frac{[A_1][A_2]}{[A_1A_2]},$$

onde $[A_n]$ representa a concentração da fita A_n e $[A_1A_2]$ a concentração das fitas unidas. Como a relação entre ΔG° e K é dada por $\Delta G^\circ = -RT\ln K$, onde T é a temperatura e R a constante dos gases ideais, concluímos que

$$\Delta G^\circ = -RT\ln\frac{[A_1][A_2]}{[A_1A_2]}. \tag{3.1}$$

Na T_m, se as fitas não forem autocomplementares, ou seja, $A_1 \neq A_2$, teremos $[A_1A_2] = [A_1] = [A_2]$. Definindo C_T como a concentração total de fitas na reação, teremos $C_\mathrm{T} = 2[A_1A_2] + [A_1] + [A_2] = 4[A_1]$ e portanto

$$\frac{[A_1][A_2]}{[A_1A_2]} = \frac{[A_1]^2}{[A_1]} = [A_1] = \frac{C_\mathrm{T}}{4}.$$

Para fitas autocomplementares, $A_1 = A_2$, logo

$$\frac{[A_1][A_2]}{[A_1A_2]} = \frac{[A_1]^2}{[A_1A_1]}.$$

Na T_m, teremos $[A_1] = 2[A_1A_1]$ e portanto

$$C_\mathrm{T} = 2[A_1A_1] + [A_1] = 2[A_1] \quad \Rightarrow \quad \frac{[A_1]^2}{[A_1A_1]} = \frac{[A_1]^2}{[A_1]/2} = 2[A_1] = C_\mathrm{T}.$$

Concluí-se que a relação $[A_1][A_2]/[A_1A_2]$ será dada por C_T/x, com $x = 4$ para fitas não autocomplementares e com $x = 1$ para fitas autocomplementares. Isolando T na Equação (3.1) e substituindo-o por T_m, teremos

$$T_m = -\frac{\Delta G^\circ_{T_m}}{R\ln(C_\mathrm{T}/x)}.$$

Substituindo nessa equação a relação entre a variação de energia livre de Gibbs para uma temperatura constante $T = T_m$ em função da variação da Entropia, ΔS°, e da variação da Entalpia, ΔH°, dada por $\Delta G^\circ_{T_m} = \Delta H^\circ - T_m\Delta S^\circ$, e isolando T_m teremos

$$T_m = -\frac{\Delta G^\circ_{T_m}}{R\ln(C_\mathrm{T}/x)} = \frac{\Delta H^\circ}{\Delta S^\circ - R\ln(C_\mathrm{T}/x)}. \tag{3.2}$$

Se os valores experimentais de T_m para diferentes concentrações C_T forem determinados para um dado segmento bifilamentar de DNA, utilizando a Equação (3.2) podemos obter os valores para $\Delta G^\circ_{T_m}$ e, consequentemente, estimar os valores de ΔH° e de ΔS°. Os valores desses parâmetros para híbridos de DNA-RNA e para estruturas compostas apenas por RNA são determinados de maneira semelhante (SANTALUCIA; ALLAWI; SENEVIRATNE, 1996).

O banco de dados "*Nearest Neighbor Database*", NNDB (TURNER; MATHEWS, 2009), dispõe de um grande número de parâmetros obtidos a partir de experimentos dessa natureza, além de uma série de expressões e considerações para determinação do valor de ΔG° para estruturas secundárias. Implementamos todas as regras disponibilizadas segundo a compilação de 2004 do grupo de Turner, presente na Versão 1.02 (2011) do NNDB. Detalhes dos algoritmos estão presentes no Apêndice A.

Essa abordagem mostra-se bastante apropriada para a estimativa da energia livre de estruturas secundárias, mas insuficiente para estudos da energia livre de pseudonós. Além de uma enorme variedade de conformações, o que implicaria numa infinidade de experiências, interações de longo alcance que vão além das regras clássicas de pareamento poderão ser de difícil identificação durante os experimentos.

3.3.2 Pseudonós

Ainda não há um consenso para um modelo de determinação do valor da energia livre de Gibbs para estruturas com pseudonós. Optamos por implementar duas propostas distintas: o modelo de Isambert e Siggia (2000), denominado *KineFold*, e o modelo *Vfold*, proposto por Liu e Chen (2010).

KineFold

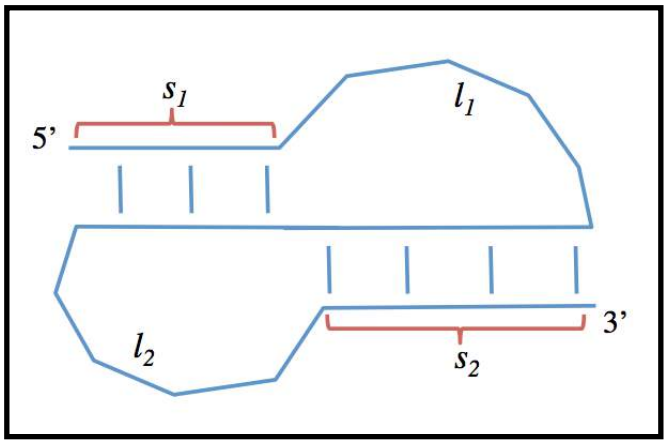
O método *KineFold* divide o problema em duas escalas. Primeiro determina subestruturas denominadas de *redes*, já apresentadas na Figura 15, de modo que o valor de sua entropia possa ser determinado analiticamente através de uma equação geral. Em seguida, trata a interação entre essas redes como um gel reticulado de granularidade grossa (*coarse-grained crosslinked-gel*), na qual cada rede é substituída por uma nodo único, conectados entre si por segmentos uni ou bifilamentares.

O valor da contribuição energética final varia de acordo com a classificação das redes e de suas dimensões. Esse modelo foi implementado seguindo estritamente as instruções disponibilizadas por Isambert e Siggia (2000), Xayaphoummine et al. (2007) e por Xayaphoummine, Bucher e Isambert (2005). Detalhes dessa abordagem estão no Apêndice B.1.

Vfold

O modelo conformacional baseado em ligações virtuais (*Virtual bond-based RNA conformational model*, *Vfold* (LIU; CHEN, 2010), trata com um nível de sofisticação adequado o tipo de pseudonó mais frequente em RNAs naturais: o *pseudonó tipo-H*. O modelo relaciona a entropia do SUR ao SBR que participa de sua formação, a partir da representação presente na Figura 18.

Figura 18 – Esquema do modelo *Vfold* para o pseudonó do tipo-H



Legenda: Representação do pseudonó segundo o modelo *Vfold*. Fonte: Modificado de Cao e Chen (2006)

Este modelo inclui etapas e considerações *a posteriori* para justificar as diferenças obtidas entre os resultados de seu modelo e os valores experimentais. Tais considerações geralmente envolvem a contribuição energética devido ao empilhamento coaxial. Além disso, importantes contribuições energéticas devido aos pareamentos localizados nas extremidades dos SBRs são desconsideradas. Então, propusemos algumas adaptações no algoritmo *Vfold* para contornar essas limitações.

De acordo com o modelo original, a variação da energia livre de Gibbs para um pseudonó seria dada por

$$\Delta G^{\circ}(s_1,s_2,l_1,l_2,l_3) = \Delta G^{\circ}_{s_1} + \Delta G^{\circ}_{s_2} - T\,\Delta S^{\circ}_{L_1} - T\,\Delta S^{\circ}_{L_2} + \Delta G^{\circ}_{\mathrm{EC}} + \Delta G^{\circ}_{\mathrm{conj}}, \qquad (3.3)$$

onde $\Delta S^{\circ}_{L_i}$ é a entropia para a componente L_i, $\Delta G^{\circ}_{s_i}$ é a energia livre para o SBR s_i segundo o modelo dos primeiros vizinhos, $\Delta G^{\circ}_{\mathrm{EC}}$ é a energia devido ao empilhamento coaxial entre os SBRs e $\Delta G^{\circ}_{\mathrm{conj}}$ = 1,3 kcal/mol é adicionado devido à variação da

entropia resultante da união das duas subunidades ($l_1 + s_2$ e $l_2 + s_1$) do pseudonó. O Apêndice B.2 apresenta em maiores detalhes essa aproximação. Para nossa adaptação, teremos

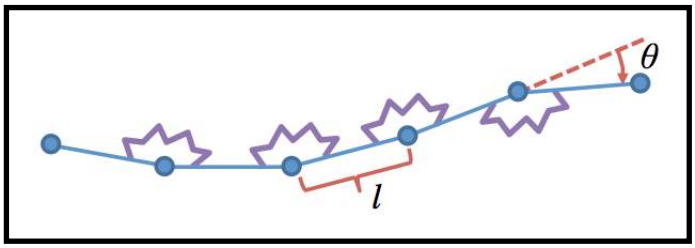
$$\Delta G^\circ = \Delta G^\circ_{s_1} + \Delta G^\circ_{s_2} - T(\Delta S^\circ_{L_1} + \Delta S^\circ_{L_2}) + \Delta G^\circ_{\text{conj}} + \Delta G^\circ_{\text{Ex}} + \Delta G^\circ_{\text{ini}}. \tag{3.4}$$

Os valores de $\Delta G^\circ_{s_1}$ e de $\Delta G^\circ_{s_2}$ são obtidos seguindo o modelo dos primeiros vizinhos. Os valores de $-T(\Delta S^\circ_{l_1} + \Delta S^\circ_{l_2})$ e $\Delta G^\circ_{\text{conj}} = +1{,}3$ kcal/mol foram mantidos conforme o trabalho de Liu e Chen (2010). Removemos da Equação (3.3) o termo $\Delta G^\circ_{\text{EC}}$, resultado do empilhamento coaxial entre os SBRs, isto é, resultado do alinhamento dos SBRs internos em um eixo comum, pois tal configuração implicaria em torções acentuadas, principalmente para pequenos pseudonós. Em seu lugar, incluímos a variação de energia devido à extensão terminal dos SBRs, $\Delta G^\circ_{\text{Ext}}$, segundo o modelo dos primeiros vizinhos. Finalmente, adicionamos uma energia inicial para a formação do pseudonó resultante da perda de entropia translacional e rotacional, que consideramos semelhante ao valor esperado devido à interação entre duas moléculas de RNA distintas: $\Delta G^\circ_{\text{ini}} = +4{,}1$ kcal/mol (TURNER; MATHEWS, 2009). Utilizando essa abordagem, obtivemos valores inferiores aos encontrados pelo modelo original. Determinamos, então, um novo termo energético, resultado da curvatura dos SURs.

Energia potencial devido à curvatura

Optamos por um modelo de granularidade grossa para analisarmos o comportamento desse termo num primeiro momento: na conformação conhecida como *cadeia discreta*, esferas rígidas, que correspondem aos nucleotídeos, estão ligadas entre si por hastes rígidas. Entre elas, inserimos *molas de torção*, responsáveis pelas propriedades elásticas da curvatura. A Figura 19 representa esse modelo.

Figura 19 – Esquema do modelo de contas e molas



Legenda: Cada nucleotídeo é substituído por uma esfera rígida. Nesse modelo, as esferas estão ligadas por hastes rígidas, e entre cada par de hastes subsequentes existe uma mola de torção. l representa a distância entre os nucleotídeos e θ o ângulo entre as direções de hastes sucessivas, em radianos. Fonte: Produzido pelo próprio autor

Seguindo a abordagem proposta por Underhill e Doyle (2005), analisando pequenos fragmentos dessa cadeia podemos, numa primeira aproximação, considerar que a força elástica será próxima da obtida para um *modelo de halteres*, pois seus comportamentos esperados nos limites serão os mesmos: para uma cadeia suficientemente longa, a energia devido a curvatura vai para zero; no limite inferior, a cadeia se comporta como uma haste rígida. De acordo com o modelo de Kratky-Porod (KRATKY; POROD, 1949), nessa configuração a energia potencial devido à curvatura será dada por

$$U_{curv}(\theta) = \frac{1}{2}\frac{K}{l}\theta^2 = \frac{1}{2}\frac{k_\mathrm{B} T\, l_P}{l}\theta^2, \tag{3.5}$$

onde K é a rigidez da mola entre as hastes, k_B é a constante de Boltzmann, T é a temperatura e l_P é o chamado de *comprimento de persistência*. Podemos ver l_P como um fator que relaciona a resistência ao dobramento da cadeia em relação à energia térmica, relacionado diretamente ao comprimento de Kuhn do polímero, l_K: $l_K = 2l_P$. Como para o RNA $l_K = 2{,}5\,l$ (ISAMBERT; SIGGIA, 2000), teremos

$$U_{curv}(\theta) = \frac{5}{8} k_\mathrm{B} T \theta^2. \tag{3.6}$$

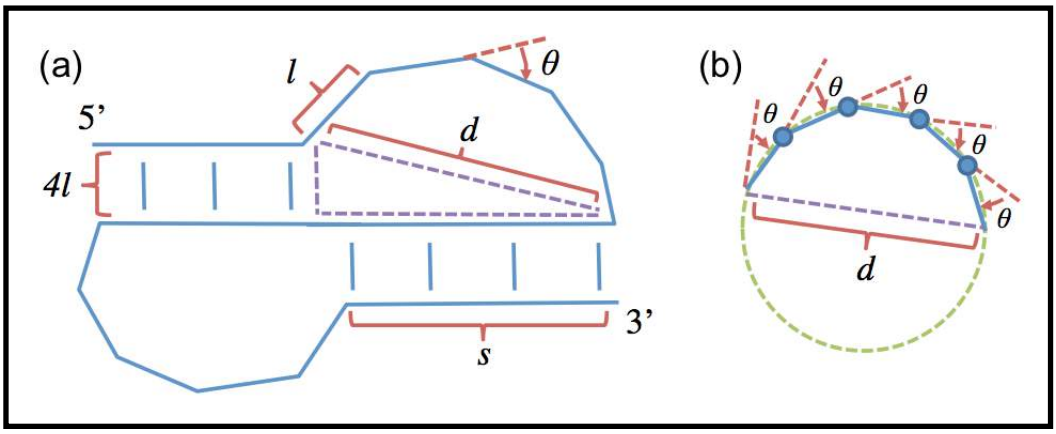
A próxima etapa consiste em determinar o valor de θ para os SURs presentes no pseudonó: para isso, aproximamos essas estruturas por uma linha poligonal simples, de modo que a distância entre seus extremos seja dada pela relação $d = \sqrt{(4l)^2 + s^2}$, onde s é igual ao número de pares de bases do SBR oposto ao

segmento em questão, n_s, multiplicado por l, e $4l$ representa a distância entre os segmentos que compõem os SBRs (ISAMBERT; SIGGIA, 2000). Essa linha poligonal é determinada de modo que todos os ângulos externos valham θ, e, portanto, corresponde às arestas de um polígono regular. A Figura 20 apresenta esse modelo. Seguindo esse princípio, teremos

$$\left| \sum_{k=1}^{n_l+1} \mathrm{e}^{i\,k\,\theta} \right| = \sqrt{16 + n_s^2}, \tag{3.7}$$

onde n_l representa o número de nucleotídeos que compõe o SUR. A Equação (3.7) não possui solução analítica; contudo, métodos numéricos são capazes de encontrar valores satisfatórios para θ.

Figura 20 – Esquema do modelo para determinação de θ



Legenda: (a) Representação do pseudonó. l: distância entre dois nucleotídeos subsequentes. A distância entre as extremidades do primeiro SUR é dada pela hipotenusa do triângulo retângulo em destaque, cujos catetos são obtidos a partir do comprimento s do SBR oposto, e pela distância entre os segmentos que constituem o SBR adjacente, $\approx 4l$. (b) Aproximação do SUR pela linha poligonal simples de um polígono regular, ou seja, as hastes correspondem às arestas desse polígono e portanto estão rotacionadas em θ radianos em relação à haste anterior. Figuras fora de escala. Fonte: Produzido pelo próprio autor

3.4 Taxas de ocorrência das reações

O valor para a taxa de ocorrência da reação que representa a transição entre dois estados permitidos para uma molécula de RNA é obtido a partir da diferença entre seus valores de energia livre, ponderados através do peso de Boltzmann, ou seja,

$$k_{i,j} = k_0 \,\mathrm{Exp}\left[\frac{\Delta G_j^\circ - \Delta G_i^\circ}{k_\mathrm{B} T}\right], \tag{3.8}$$

onde $k_{i,j}$ representa a taxa de transição do estado i para o estado j, ΔG_j° e ΔG_i° suas respectivas variações de energia livre de formação e k_0 é um pré-fator constante. Não existe base experimental para determinação exata do valor desse pré-fator, então optamos empiricamente por $k_0 = 0{,}1\ \mathrm{s}^{-1}$.

3.5 Simulações estocásticas: o algoritmo de Gillespie

O uso de uma abordagem baseada em reações químicas simuladas a partir do algoritmo de Gillespie (GILLESPIE, 1977) apresenta uma boa relação entre custo computacional, complexidade matemática e realismo biológico. Para tal, precisamos definir o conceito de *estado* nessa abordagem: o *estado* de um sistema representa sua condição em um determinado instante e contém informações suficientes para prever seu comportamento no futuro. Intuitivamente, trata-se do conjunto de variáveis que precisam ser acompanhadas nesse modelo.

Exemplificando, a reação

$$A \xrightarrow{k} A'$$

lida com duas espécies químicas, A e A', sendo A transformado em A' a uma taxa k. Utilizando um modelo estocástico, uma única molécula de A é convertida numa molécula de A', ou seja, o número total de elementos de A decresce em uma unidade e acrescenta-se uma unidade ao número de elementos de A'. A probabilidade desse evento ocorrer num intervalo de tempo diferencial dt é dada pelo produto entre a taxa k, o número de moléculas de A, e pelo tempo dt. Todo o sistema está em um único estado exato em cada período de tempo no algoritmo de Gillespie.

Essencialmente, uma transição consiste em executar uma reação R_j, com $1 < j < n$ e n igual ao número reações possíveis a partir do dado estado inicial. Através

de números aleatórios oriundos de uma distribuição de probabilidade uniforme, determina-se qual a próxima reação permitida e qual o tempo necessário para tal. A implementação foi realizada de acordo com o Algoritmo 3.2.

Algoritmo 3.2: Algoritmo de Gillespie - Método Direto

1. Inicialize: determine o tempo inicial, t_0, o tempo máximo da reação, $t_{\text{máx}}$, e número de moléculas iniciais de cada composto. Faça $t = t_0$.
2. Calcule as probabilidades a_j para cada reação, e faça $a_0 = \sum_{j=1}^{n} a_j$.
3. Determine dois números aleatórios oriundos de uma distribuição uniforme, r_1 e r_2.
4. Determine o tempo para a próxima reação, τ, dado que
$$\tau = \frac{1}{a_0} \ln\left(\frac{1}{r_1}\right).$$
5. Determine a próxima reação, R_μ, tal que $\mu \in \mathbb{N}^*$ e será o menor valor que satisfaça a relação
$$\sum_{j=1}^{\mu} a_j > r_2 a_0.$$
6. Altere o número de moléculas para refletir a execução da reação R_μ e faça $t = t + \tau$.
7. Verifique se o número de moléculas de algum reagente é zero ou se $t \geq t_{\max}$. Se sim, finalize. Senão, volte para o passo 2.

3.6 Programa desenvolvido

A partir dos conceitos expostos, foram elaboradas funções em *Wolfram Language* que:

a) Determinam quais são os SBRs que podem se formar a partir de uma dada estrutura, retornando a posição das ligações e seus respectivos participantes, utilizando os princípios apresentados na Seção 3.2. Para tal, substituem, na sequência de entrada, os nucleotídeos já pareados por elementos virtuais não pareáveis;

b) Encontram, a partir da matriz construída no item anterior, o conjunto de

todas as possíveis organizações resultantes da adição ou da remoção de um único SBR da conformação atual, incluindo as estruturas com pseudonós. Para o modelo *Vfold*, apenas os pseudonós do tipo-H são mantidos;

c) Determinam os valores de energia livre das estruturas de interesse, seguindo as instruções apresentadas na Seção 3.3;

d) Encontram o valor da taxa de transição entre os estados a partir dos valores de energia livre, utilizando a Equação (3.8);

e) Realizam uma simulação de Monte Carlo através do Método Direto do algoritmo de Gillespie, determinando qual a próxima estrutura no "caminho" de dobramento da molécula de RNA.

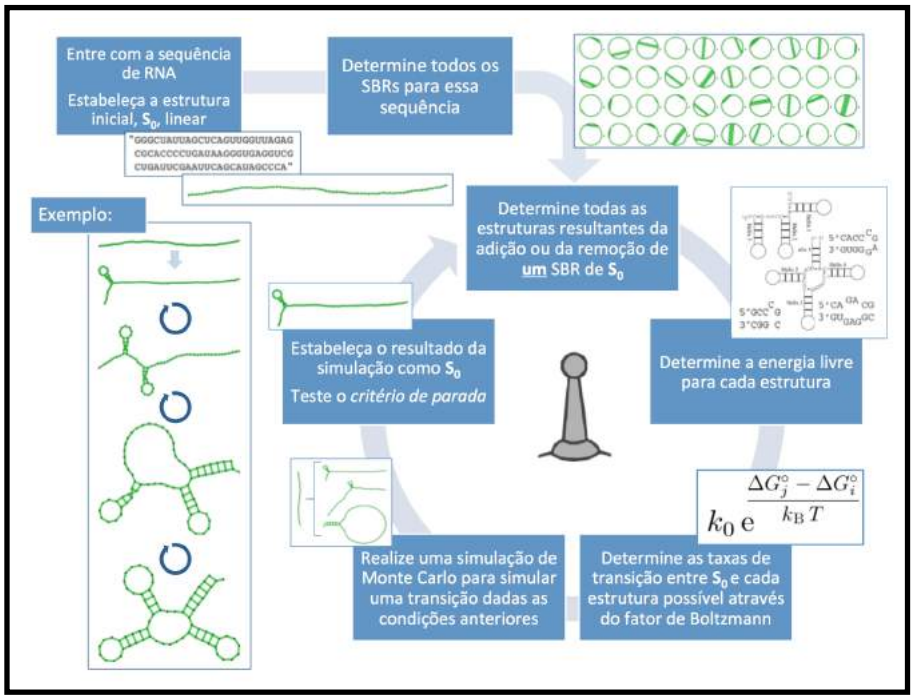
A Figura 21 apresenta um quadro do funcionamento do programa, semelhante a Figura 13, mas discriminando os conceitos desenvolvidos até aqui. Cada nova simulação pode representar um "caminho" distinto pela superfície de energia livre. Portanto, um número razoável de simulações deve ser realizado para que a saída represente um *conjunto* válido de estruturas subótimas possíveis.

3.7 Simulações cotranscricionais

Para as simulações cotranscricionais, determinamos as posições nas quais permitiremos que a fita de RNA recém transcrita assuma suas diferentes conformações, os chamados *sítios de pausa*. Consideramos que a distância entre o transcrito livre e a posição do sítio ativo da enzima é de 14 nt, pois além dos 9 nt que constituem o complexo RNA-DNA dentro da bolha de transcrição, 5 nt adicionais estarão protegidos no canal de saída da RNA polimerase (GESZVAIN; LANDICK, 2005).

Os sítios de pausa podem ser identificados através do algoritmo presente no Apêndice C, capaz de prever pausas relacionadas à estabilidade da bolha de transcrição no estado pré-translocado da RNAP em relação ao estado pós-translocado, que classificamos como *pausas do tipo 1* e pausas devido ao *backtracking*, classificadas como *pausas do tipo 2*, seguindo as definições apresentadas por Zhang e Landick (2016). Esse algoritmo também foi utilizado na elaboração de uma análise do com-

Figura 21 – Esquema do funcionamento do programa



Legenda: Apresentação esquemática das etapas envolvidas no programa desenvolvido. Fonte: Produzido pelo próprio autor

portamento da RNAP durante a transcrição de genes ribossomais de *Escherichia coli*, desenvolvida paralelamente a essa Tese. Maiores detalhes a respeito desse trabalho estão apresentados no Apêndice E.

O programa ainda permite que as posições de pausa sejam determinadas pelo usuário, seja para analisar o efeito de uma pausa forçada em um sítio específico ou caso os sítios de pausa já estejam experimentalmente estabelecidos.

Considerando a polimerização da fita de RNA até o primeiro sítio de pausa, o programa determina o comprimento da fita livre (posição da inclusão do novo nucleotídeo menos 14 nt) e então permite o dobramento desse segmento, numa janela temporal equivalente ao tempo estipulado para a pausa. O procedimento é repetido para os próximos sítios de pausa, até a fita de RNA estar completamente transcrita. No final, o programa realiza novas simulações até a estabilização da estrutura da molécula. Essa aproximação, além de biologicamente razoável, facilita a determinação das estruturas possíveis, pois reduz significantemente o espaço de possibilidades de pareamento entre as bases disponíveis.

3.8 Métricas

Para determinar e comparar as capacidades de nosso modelo, é necessário definir as métricas que serão utilizadas. Seguindo o critério de Puton et al. (2013), teremos:

TP: *True positives*, verdadeiros positivos, pares de bases corretamente preditos;

TN: *True negatives*, verdadeiros negativos, bases não-pareadas corretamente preditas. O valor de TN equivale ao número de pares de bases teóricos permitidos na sequência de referência e que também não estão presentes na estrutura predita. Para determinar tais pares, assume-se que a distância mínima entre pares de bases é de 1 nt;

FN: *False negatives*, falsos negativos, pares de bases presentes na estrutura de referência, mas não preditos;

FP: *False positives*, falsos positivos, pares de bases preditos que não estão presentes na estrutura de referência. São separados em três sub-categorias:

a) Inconsistentes: Pares de bases preditos em conflito com um par de base presente na estrutura de referência, ou seja, *i*/*j* é predito e *i*/*k* ou *h*/*j* estão pareados na referência, com $h \neq i$ e $j \neq k$;

b) Contraditórios: Pares de bases "não aninhados" com a estrutura de referência, afetando sua conformação espacial, ou seja, *i*/*j* é predito, mas existe um par de base *k*/*l* na estrutura de referência que satisfaz $k<i<l<j$.

c) Compatíveis: ϵ, representa os pares de bases que não foram classificados como TP, nem como contraditórios, nem como inconsistentes.

Com base nessas métricas, podemos definir os indicadores gerais, habitualmente utilizados para análises preditivas, tais como a *taxa de verdadeiros positivos*, TPR (*True positive rate*, em inglês), dada por

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

e o *valor preditivo positivo*, PPV (*Positive Predicted Value*, em inglês),

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + (\mathrm{FP} - \epsilon)}.$$

Além dessas medidas, o *Coeficiente de Correlação de Mathews*, MCC (*Matthews Correlation Coefficient*, em inglês), é utilizado para aferir a qualidade do modelo utilizado e facilitar comparações entre métodos, pois se trata de uma métrica que tem por objetivo englobar todas as anteriores. É dado pela expressão

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - (\mathrm{FP} - \epsilon) \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP} - \epsilon)(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP} - \epsilon)(\mathrm{TN} + \mathrm{FN})}}. \tag{3.9}$$

O valor máximo de MCC, $+1$, indica uma predição perfeita; seu intermediário, zero, indica que não existe diferença entre a abordagem utilizada e um preditor aleatório; seu mínimo, -1, representa o caso no qual a predição está totalmente em desacordo com o resultado observado (GORODKIN; STRICKLIN; STORMO, 2001).

3.9 Verificação da implementação

O processo de *verificação* de uma implementação visa apurar se os resultados obtidos por simulação representam de forma fiel a descrição e os conceitos do modelo no qual se baseia.

3.9.1 Estruturas secundárias

No primeiro teste de verificação, observamos se o programa retorna corretamente os valores de energia livre para as estruturas de exemplo presentes no NNDB (TURNER; MATHEWS, 2009). Além dos dados disponibilizados pelo próprio banco, comparamos nossos resultados com as saídas da ferramenta *RNAeval* do pacote *ViennaRNA* versão 2.2.5 e do *servidor web* para *KineFold* 2016. O pacote *ViennaRNA*, assim como o *KineFold*, baseia-se em métodos semelhantes aos de nossa implementação para determinação dos valores de energia livre de estruturas secundárias. A ferramenta *RNAeval* detalha os valores de energia livre obtidos para as diferentes subestruturas, permitindo comparações mais diretas. Não realizamos testes no servidor para o programa *Vfold* pois o mesmo estava inacessível no primeiro semestre de 2016.

Realizamos o mesmo procedimento para o exemplo proposto pelo *servidor web ViennaRNA*, uma estrutura em trevo clássica de um RNA transportador resultado do dobramento da sequência "5′-GGGCU AUUAG CUCAG UUGGU UAGAG CGCAC CCCUG AUAAG GGUGA GGUCG CUGAU UCGAA UUCAG CAUAG CCCA-3′". Nesse caso, além de verificar se os valores de energia livre para tal estrutura são compatíveis entre as implementações, realizamos também 20 réplicas com 1.000 simulações da renaturação para avaliar a distribuição das conformações estáveis permitidas pelo nosso programa.

Finalmente, verificamos com maiores detalhes os valores de energia livre de Gibbs para estruturas secundárias através do *servidor web KineFold*. As sequências, nesse caso, foram determinadas empiricamente, conforme a necessidade e tipo de estudo.

3.9.2 Pseudonós

Método *KineFold*

Verificamos a implementação do algoritmo de Isambert e Siggia (2000) comparando os resultados do cálculo de energia livre para pseudonós com diferentes níveis de complexidade obtidos via *Bender* com os valores obtidos através do *servidor web KineFold*.

Método *Vfold*

O segundo método implementado, *Vfold*, parte do valor da variação de entropia (ΔS) dos SURs presentes no sulco profundo e no sulco raso do pseudonó do tipo-H. Verificamos se nossos valores correspondiam aos publicados por Cao e Chen (2006), obtidos modelando os SURs como cadeias livres disposta de acordo com a simulação de um caminhante aleatório autoevitante numa *rede de diamante*.

Posteriormente, realizamos o cálculo da variação da energia livre para a formação do pseudonó *T4-35*, "5′-GAUUA UGCCA GCUAU GAGGU AAAGU GUCAU AGCAC-3′", exemplo apresentado no mesmo trabalho. Essa sequência também foi utilizada para a verificação da versão adaptada do *Vfold*, que inclui o termo extra resultante da energia potencial devido à curvatura dos SURs.

3.9.3 Tempo de execução do programa

Realizamos o dobramento cotranscricional *in silico* de sequências aleatórias com comprimentos múltiplos de 25, de 25 nt até 700 nt, segundo o modelo baseado no método *Vfold* para pseudonós, mensurando o tempo necessário para cada simulação. Através de regressão polinomial pelo método dos mínimos quadrados encontramos a função que melhor ajustaria as médias das distribuições dos tempos das 96 simulações realizadas para cada comprimento. A máquina utilizada durante os testes possuía processador de 2,10 GHz Intel® Xeon® E5-2620, módulos de memória de 1600 MHz DDR3 ECC Kingston®, sistema operacional Ubuntu 15.04 e *Mathematica* 10.0. Para efeito de comparação, instalamos localmente a versão de 2016 do programa *KineFold* (XAYAPHOUMMINE; BUCHER; ISAMBERT, 2005) e realizamos a simulação das mesmas sequências na mesma máquina, utilizando os parâmetros sugeridos pelos autores do programa.

3.10 Validação da implementação

O processo de *validação* de uma implementação visa apurar o nível de adequação dos resultados obtidos via simulação em relação aos dados apurados experimentalmente.

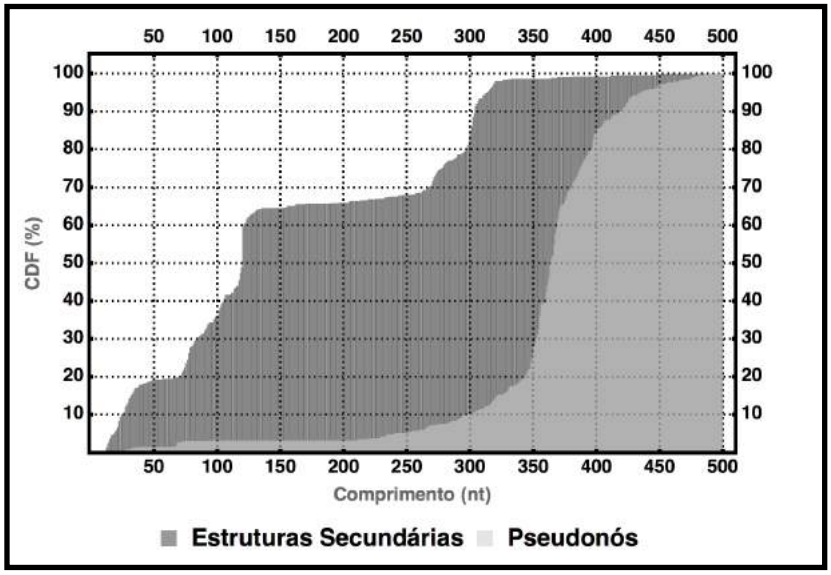
3.10.1 Cadeia de Markov para T4-35

A sequência *T4-35* é relativamente curta, apresentando 35 nucleotídeos. A fim de demonstrar algumas propriedades de nossa abordagem, exploramos a *Cadeia de Markov* para essa sequência. Para tal, encontramos todas as estruturas possíveis para a *T4-35*, que envolvam no máximo pseudonós do tipo-H, e determinamos o valor da energia livre de Gibbs para cada uma delas. Selecionamos apenas as energeticamente favoráveis, ou seja, aquelas com $\Delta G^\circ < 0$. Analisando as Cadeias de Markov podemos determinar a probabilidade do sistema atingir os chamados *estados absorventes*, conformações que não permitem a transição entre outros estados, e seriam equivalentes aos mínimos energéticos locais da sequência em estudo. Realizamos também 20 réplicas com 1.000 simulações cada da renaturação utilizando nosso programa para verificarmos quais as semelhanças entre os resultados de nossa abordagem e as saídas da Cadeia de Markov.

3.10.2 Banco de dados *RNAStrand*

Para avaliar o potencial preditor de nossa implementação, selecionamos sequências com comprimento máximo de 500 nt, formadas por somente uma molécula, descartando fragmentos e não redundantes (sem duplicatas) obtidas no Banco de dados *RNAStrand* 2.0. (ANDRONESCU et al., 2008). A Figura 22 apresenta a função distribuição acumulada (FDA ou CDF, do inglês *Cumulative distribution function*) do comprimentos das sequências. No total, obtivemos 637 estruturas secundárias e 604 estruturas com pseudonós.

Figura 22 – Gráfico da distribuição acumulada em função do comprimento da sequência



Legenda: A tonalidade mais escura representa a CDF para sequências compostas apenas por estruturas secundárias, enquanto a tonalidade mais clara representa a CDF para sequências com pseudonós. Observe que ≈65% das estruturas secundárias estudadas possuem menos que 150 nt (414 das 637 sequências), e que se atinge ≈100% com sequências de até ≈350 nt. O mesmo não se observa para estruturas com pseudonós, que possuem um crescimento acentuado em seu número a partir de ≈350 nt, com cerca de 20% delas abaixo desse comprimento (122 sequências).
Fonte: Produzido pelo próprio autor

A Figura 23 representa a porcentagem de nucleotídeos pareados nas estruturas: uma fração significativa apresenta entre 50 e 70% de suas bases pareadas, independente da presença de pseudonós. As estruturas com maior proporção de pareamento são sequências mais curtas e sem pseudonós: grande parte delas são

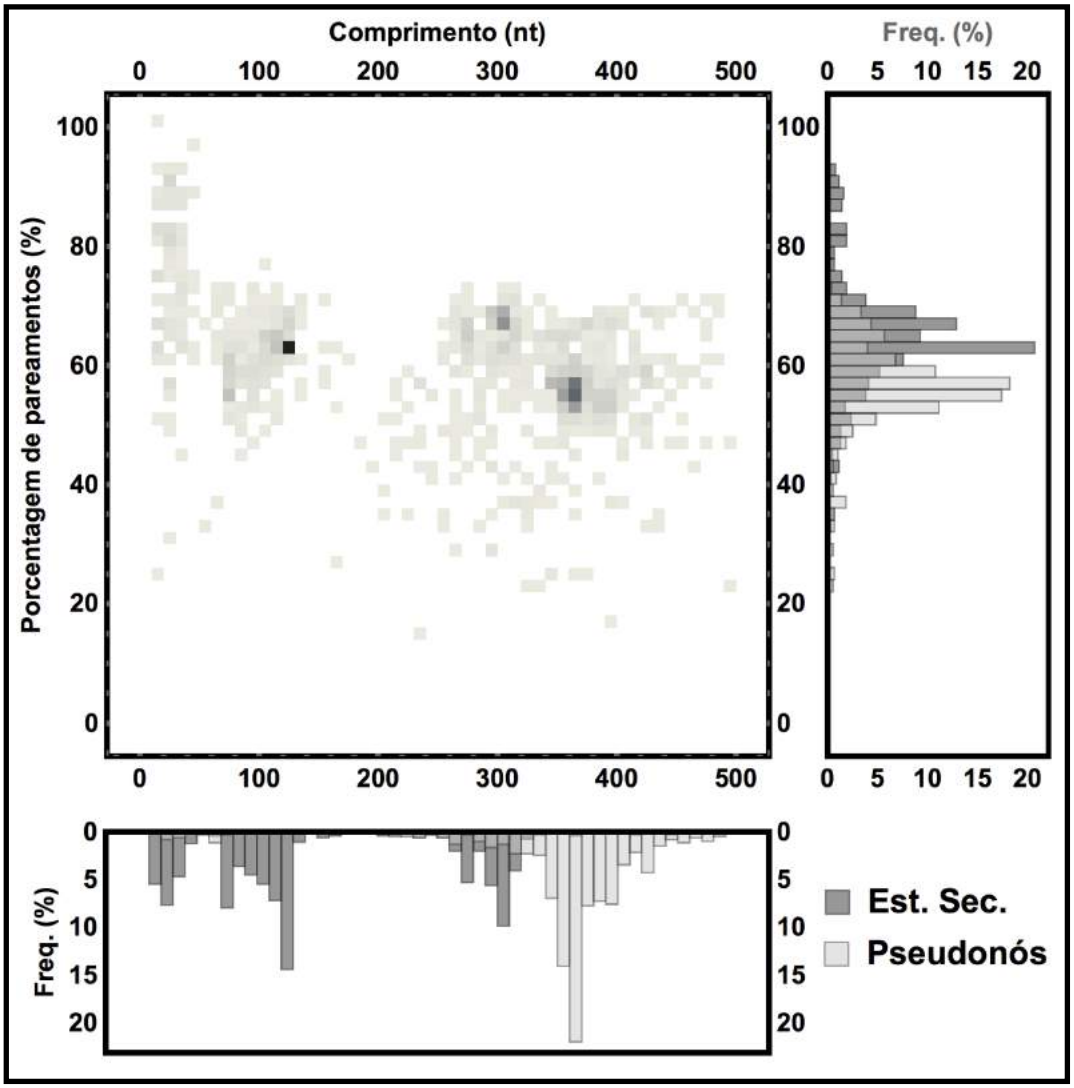
constituídas por um único grampo. Na Figura 24 temos os histogramas para a distribuição dos comprimentos dos SBRs e para o número de SBRs. Nota-se que, de acordo com a Figura 24a, os SBRs dificilmente possuem mais de 10 nt de comprimento, independentemente da presença ou ausência de pseudonós; já na Figura 24b, esse fator é relevante, pois as sequências com pseudonós possuem uma distribuição concentrada em torno de 23 SBRs, enquanto as estruturas secundárias se distribuem entre duas regiões: mais de 60% delas possuem menos de 10 SBRs, mas também ocorre um acúmulo na região onde se concentram os pseudonós.

Foram realizadas 48 simulações cotranscricionais do dobramento para cada uma dessas sequências, sem nenhum tipo de intervenção. Como entrada utilizamos somente a sequência de nucleotídeos da fita de RNA. O programa identificou os sítios de pausa teóricos para cada uma das sequências. Determinamos os valores de MCC para cada sequência e comparamos o desempenho de nosso programa qualitativamente e quantitativamente, através do teste não paramétrico para dados pareados de Wilcoxon, simulando as mesmas sequências utilizando o pacote *ViennaRNA* 2.2.5 (GRUBER et al., 2008), e o programa *KineFold* 2016 (XAYAPHOUMMINE; BUCHER; ISAMBERT, 2005). Ambos os programas foram instalados localmente.

3.10.3 Análise de sequências com estruturas competitivas

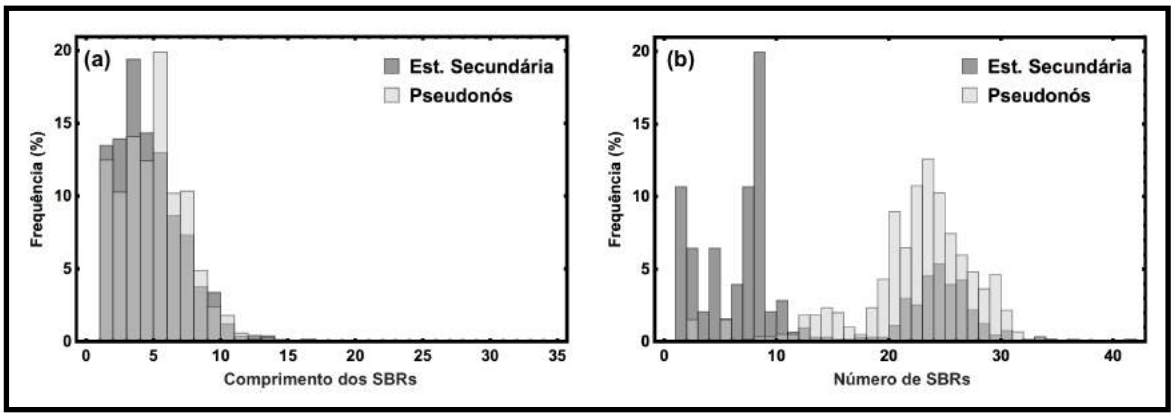
Xayaphoummine et al. (2007) propuseram uma sequência de RNA que se acomoda de tal forma que suas duas estruturas estáveis possíveis apresentam valores muito próximos de variação de energia livre de formação. Além disso, a sequência dada pelos nucleotídeos originais em ordem inversa apresenta o mesmo conjunto de *vizinhos mais próximos* para a determinação de sua energia livre, devido à simetria interna de seus SBRs. Tanto a sequência *direta* como a *inversa* podem assumir uma estrutura em grampo longo com protuberâncias ou uma estrutura composta por dois grampos independentes. Entretanto, devido ao *caminho de dobramento*, as sequências assumem configurações em proporções variáveis de acordo com as condições experimentais. A Figura 25 apresenta uma representação das estruturas discutidas. A sequência direta é dada por 5′-GGAAC CGUCU CCCUC UGCCA AAAGG UAGAG GGAGA UGGAG CAUCU CUCUC UACGA AGCAG AGAGA GACGA AGG-3′.

Figura 23 – Gráfico da distribuição de nucleotídeos pareados



Legenda: Sequências/estruturas selecionadas do banco *RNAStrand*. O gráfico da esquerda apresenta a distribuição das proporções de nucleotídeos pareados em função do comprimento da sequência, independente da presença ou ausência de pseudonós. Quanto mais escura a região, maior o número de sequências encontradas para aquela condição. O histograma à direita representa a distribuição das sequências com determinada porcentagem de nucleotídeos pareados. O histograma abaixo representa a distribuição dos comprimentos. Para os histogramas, a tonalidade mais escura representa sequências compostas apenas por estruturas secundárias, enquanto a tonalidade mais clara representa sequências com pseudonós. Fonte: Produzido pelo próprio autor

Figura 24 – Histograma da distribuição dos comprimentos e do número de SBRs nas estruturas

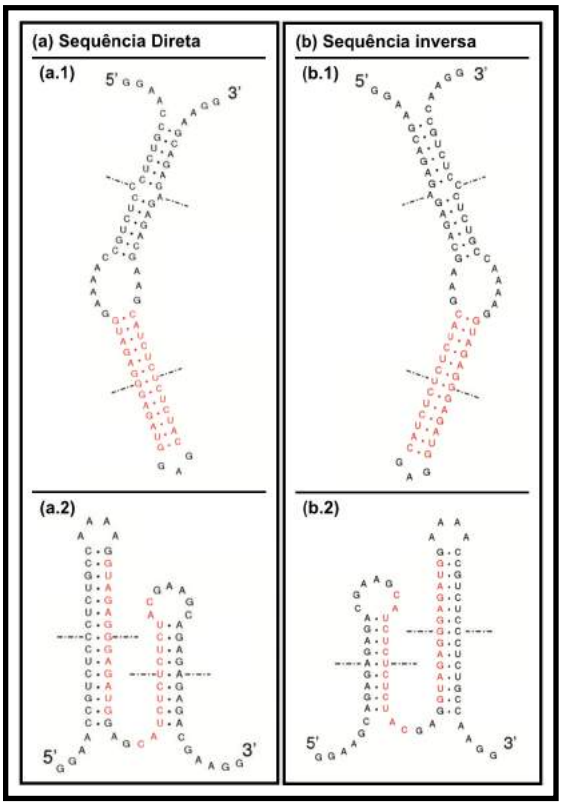


Legenda: Sequências/estruturas selecionadas do banco *RNAStrand*. (a) Distribuição dos comprimentos das SBRs. (b) Distribuição do número de SBRs. Fonte: Produzido pelo próprio autor

Xayaphoummine et al. (2007) sintetizaram *in vitro* a sequência direta a 37 °C e a sequência inversa em três situações: a 37 °C, a 25 °C e a 37 °C na presença do oligonucleotídeo 5′-CCTCTAC-3′. Em todos os casos, a enzima responsável pelo processo foi a RNAP do bacteriófago T7. Para a temperatura de 37 °C, estimaram que taxa de transcrição está entre 200 e 400 nt/s, enquanto que à 25 °C essa taxa é reduzida entre 3 e 4 vezes. A sequência direta assumiu apenas a estrutura composta por dois grampos independentes. A sequência inversa, por sua vez, assumiu a proporção de 9 estruturas em grampo longo para 1 estrutura composta por grampos independentes para 37 °C, apenas a estrutura em grampo longo para 25 °C e a proporção de 1 para 1 quando na presença do oligonucleotídeo.

Realizamos 20 réplicas com 1.000 simulações cada cotranscricionais considerando os sítios de pausa previstos para SRA, MRA e 20 réplicas com 1.000 simulações cada da renaturação, para ambas as sequências. Comparamos nossos resultados com os experimentais, analisando o efeito das pausas na distribuição de estruturas finais de nossas simulações. Realizamos mais dois conjuntos de simulações, com sítios de pausa determinados a partir dos resultados das simulações anteriores.

Figura 25 – Esquema das estruturas resultantes da versão direta e inversa de uma sequência de RNA



Legenda: As marcações indicam o centro de simetria dos SBRs. (a) Versão direta da sequência em estudo. (b) Versão inversa da sequência em estudo. Fonte: Modificado de Xayaphoummine et al. (2007).

3.10.4 Influência da estrutura do RNA nascente na cinética da RNAP

Simulamos o dobramento cotranscricional para sequências com sítios de pausa conhecidos, analisando se as possíveis estruturas efêmeras poderiam interferir na cinética da RNAP. Estudamos se inconsistências na predição de pausas utilizando o modelo para transcrição presente no Apêndice C poderiam ser resultado da formação dessas estruturas. No total, temos dados para dez sequências, sem estruturas experimentalmente documentadas: deleções D104, D111, D112, D123, D167 e D387 da região inicial do genoma do bacteriófago T7 (LEVIN; CHAMBERLIN, 1987) e Sequências 10, 11, 12 e 13, apresentadas no trabalho de Tadigotla et al. (2006). A Tabela 4 apresenta as sequências estudadas nesse trabalho. Para cada sequência, foram realizadas 20 réplicas com 240 simulações cotranscricionais cada, considerando as pausas experimentalmente verificadas e pausas teóricas do tipo 1 e 2.

Tabela 4 – Sequências com sítios de pausa experimentalmente verificados.

Nome	Sequência
Seq. 10	AUCGA GAGGG ACACG GCGAA UAGCC AUCCC AAUCG ACACC GGGGU CCGGG AUCUG GAUCU GGAUC GCUAA UAACA UUUUU AUUUG GAUCC CCGGG UACCG AGCUC GAAUU CACUG GCCGU CGUUU UACAA CGUCG UGACU GGGAA AACCC UGGCG
Seq. 11	AUCGA GAGGG ACACG GCGAA UAGCC AUCCC AAUCC GACAC CGGGG CAUCG AGUGG GACAC GGCGA AUAGC CAUCC CAAUC GACAC CGGGG UCCGG GAUCU GGAUC UGGAU CGCUA AUAAC AGGCC UGCUG GUAAU CGCAG GCCUU UUUAU UUGGA UCCCC GGGUA
Seq. 12	AUCGA GAGGG CCACG GCGAA CAGCC AACCC AAUCG AACAG GCCUG CUGGU AAUCG CAGGC CUUUU UAUUU GGAUC CCCGG GUA
Seq. 13	AUCGA GAGGG CCACG GCGAA CAGCC AACCC AAUCC GAACA GCCAU CAUCC UCAGU AUUCA GGUAG CUGUU GAGCC UGGGG CGGUA GCGUG CUUUU UUCGA AUUCA CUUAA UGGUA AUCUC G
D104	AUCGA GAGGG ACACG GCGAA UAGCC AUCCC AAUCG ACACC GGGGU CAACC GGAUA AGUAG ACAGC CUGAU AAGUC GCACG ACAGA AAGAA AUUGA CCGCG CUAAG GCCCG UAAAG AACGU CACGA GGGGC GCUUA GAGGC ACGCA GAUUC AAACG UCGCA
D111	AUCGA GAGGG ACACG GCGAA UAGCC AUCCC AAUCC ACACG UCCAA CGGGG CAACC GUAUG UACAC CUGAU GGGUU CGCAA UGAAA CAACG AAUCG AACGC CUUAA GCGUG AACUC CGCAU UAACC GCAAG AUUAA CAAGA UAGGU UCCGG CUAUG ACAGA
D112	AUCGA GAGGG ACACG GCGAA UAGCC AUCCC AAUCG ACACC GGGGU CAACC GGAUA AGUAG ACAGC CUGAU AAGUC GCACU AGAAC AGGCA CUAGC CAACA CACUG AACGA UAUCU CAUAA CGAAG AUAAA GGACA CAAUG CAAUG AACAU UACCG ACAUC
D123	AUCGA GAGGG ACACG GCGAA UAGUG AGAAC UUGGC GAGAG AACAA CCUCG AACGC CGCAA GGCAC AAGAG AGGGC GGCGU GGCAU AGACG AAAGG AAAAG GUUAA AGCCA AGAAA CUCGC CGCAC UUGAA CAGGC ACUAG CCAAC ACACU GAACG CUAUC
D167	AUCGA GAGGG ACACG GCGAA UAGCC AUCCC UAACG UCUAC GAUGU ACAGC GCCAC GCUGG AUGCU AUACG GUGGU ACUUG ACGCA CUUAA GGAUU GCGAG CGUUU CAACA AUGAU GCCCA UUAUA AAUAC GCUGA GAUUG CAAGC GACAU CAUUG AUUGC
D387	AUCGA GAGGG ACACG GCGAA UAGCC AUCCC AAUCG ACACC GGGGU CAACC GGAUA AGUAG ACAGC CUGAU AAGUC GCACG AAAAA CAGGU AUUGA CAAGC GUCAA GGUAU GCUUA UCGAC UUACU GGUCG AGAUG GUCAA CAGCG AGACG UGUGA UGGCG

Fonte – Levin e Chamberlin (1987) e Tadigotla et al. (2006)

4 Resultados e Discussão

4.1 Estruturas secundárias

Todos os valores de energia livre de Gibbs para as estruturas secundárias presentes no NNDB obtidos via nossa implementação foram idênticos aos presentes no banco. Alguns resultados para grampos simples estão apresentados na Figura 26, que inclui ainda os valores obtidos via *RNAeval* e via *KineFold*. Na figura, os grampos (c), (d) e (e) são considerados especiais pois seus valores experimentais de energia livre diferem do esperado segundo o modelo dos primeiros vizinhos. Nosso programa inclui uma lista com esses casos especiais e verifica-a sempre que necessário. Note que o *RNAeval* também inclui esses valores especiais, ao contrário do *KineFold*. No geral, os valores possuem diferenças de, no máximo, décimos de kcal/mol, o que indica que as outras implementações realizam algumas aproximações na determinação da estabilidade dessas moléculas.

Figura 26 – Esquema com os valores de energia livre para exemplos de grampos presentes no NNDB

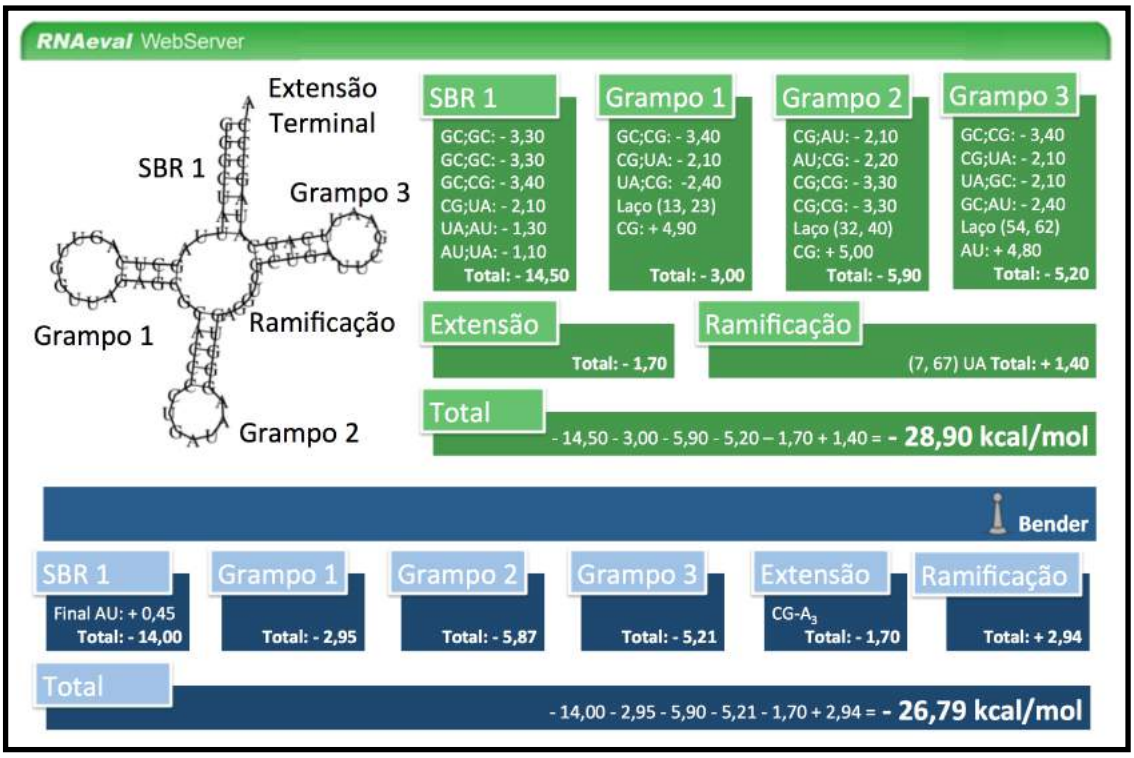
Estrutura	5'CACA$^{A}_{A}$A 3'GUGU$_{AA}$A (a)	5'CGGG$^{G}_{A}$ 3'GCCU$_{G}$A (b)	5'CACC^{C}G 3'GUGG$_{G}$A (c)	5'CACA$^{C}_{C}$C 3'GUGU$_{CC}$C (d)	5'CACA$^{G}_{A}$ 3'GUGU$_{G}$A (e)
NNDB/Bender	- 1,40 kcal/mol	- 1,90 kcal/mol	- 4,10 kcal/mol	-1,20 kcal/mol	- 1,80 kcal/mol
RNAeval	- 1,30 kcal/mol	- 2,60 kcal/mol	- 4,10 kcal/mol	- 1,20 kcal/mol	- 1,80 kcal/mol
KineFold	-1,70 kcal/mol	-2,60 kcal/mol	-3,30 kcal/mol	-1,60 kcal/mol	-2,00 kcal/mol

Legenda: Valores de energia livre para algumas estruturas de exemplo do NNDB. Os valores determinados através da nossa implementação, *Bender*, foram exatamente os mesmos disponíveis no NNDB. A ferramenta *RNAeval* inclui os grampos especiais (c), (d) e (e), mas aparentemente aproxima alguns resultados, assim como o *KineFold*. Fonte: Modificada de Turner e Mathews (2009)

A Figura 27 apresenta a descrição termodinâmica detalhada das subestruturas presentes no RNA transportador proposto pelo *ViennaRNA* segundo o servidor *RNAeval* e conforme nossa implementação. O valor da variação de energia livre

obtido não é o mesmo: aparentemente, o *RNAeval* realiza uma aproximação para determinar o valor de energia livre para múltiplas ramificações, desconsiderando a complexidade intrínseca nesses casos (veja Apêndice A.9). O valor obtido para essa estrutura via *KineFold* foi de −28,2 kcal/mol, resultado distinto das duas outras abordagens, mas o servidor não apresenta a descrição termodinâmica detalhada para comparação.

Figura 27 – Esquema da descrição termodinâmica detalhada da estrutura em trevo

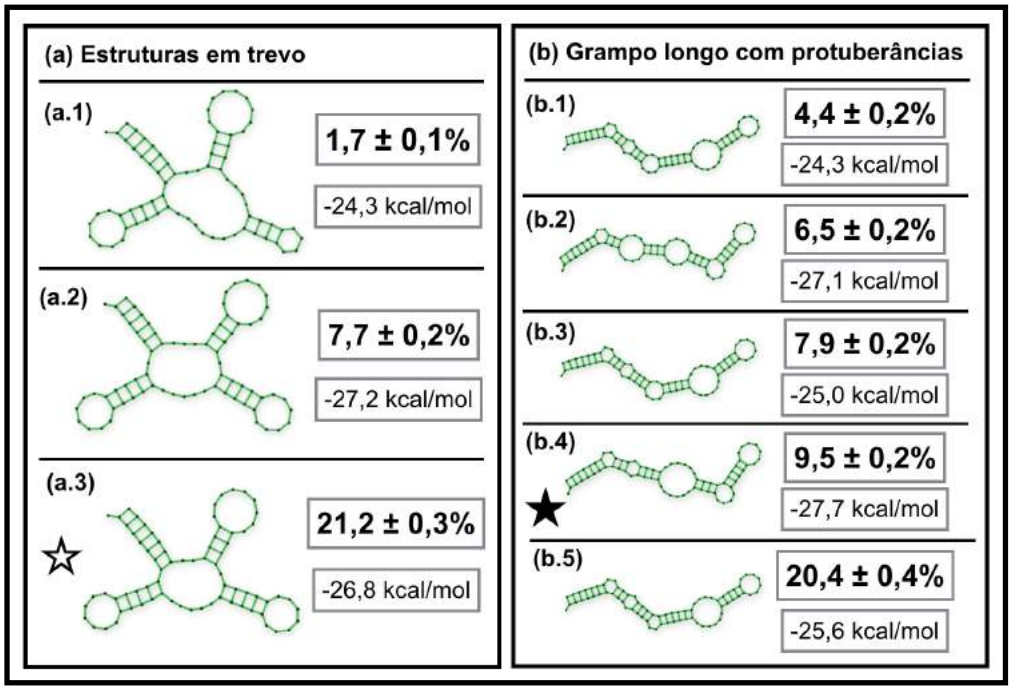


Legenda: Comparação entre os resultados dos programas *RNAeval* e *Bender* para a descrição termodinâmica detalhada das subestruturas para a estrutura em trevo. Não houve diferença significativa no cálculo da energia livre para os SBRs, apenas para a ramificação múltipla. Fonte: Produzido pelo próprio autor

Nas simulações de renaturação, nosso programa obteve a estrutura em trevo em ≈21% dos casos, maior porcentagem para uma única estrutura. Entretanto, nossa estrutura MFE corresponde a um grampo longo com protuberâncias, obtido em ≈10% das simulações. O valor da variação de energia livre para essa estrutura é de −27,7 kcal/mol, contra −26,8 kcal/mol para a estrutura de referência. Quando agrupamos as estruturas semelhantes, os trevos correspondem a ≈30% dos resultados, enquanto as hélices sequenciais representam ≈50% das saídas. A Figura 28

apresenta essas estruturas. A descrição termodinâmica dos SBRs sequenciais via *RNAeval* é equivalente à nossa implementação, indicando que a aproximação realizada pelo *ViennaRNA* para múltiplas ramificações é o que lhe garante a estrutura em trevo como estrutura MFE.

Figura 28 – Esquema das estruturas obtidas via *Bender*

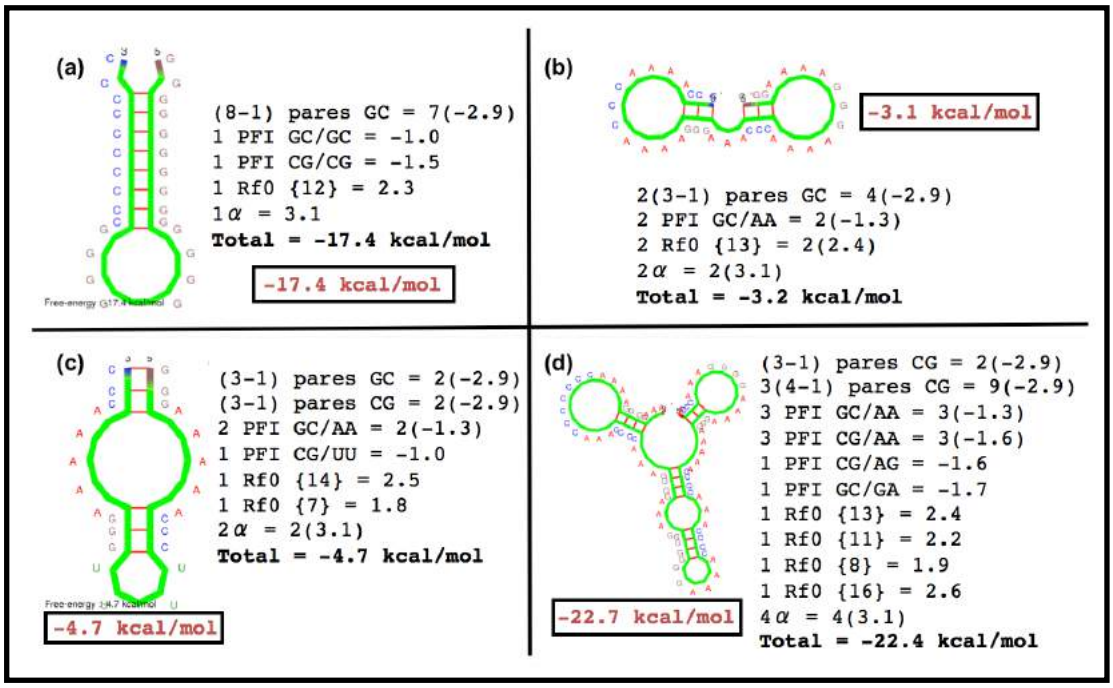


Legenda: Cada subfigura representa uma estrutura diferente, com a respectiva frequência nos resultados das simulações, o erro padrão e seu valor de variação de energia livre de Gibbs. A estrela branca indica a estrutura de referência, enquanto a preta indica a estrutura MFE obtida. (a) Estruturas em trevo; (b) Grampos longos com protuberâncias. Fonte: Produzido pelo próprio autor

Comparando os resultados do cálculo de energia livre de nossa implementação baseada no método *KineFold* com os resultados do próprio servidor, observamos uma diferença constante entre nossos resultados quando alongávamos SBRs ou SURs de sequências arbitrárias. Concluímos que essa diferença só se justificaria se os valores para *finais incompatíveis* utilizados pelo *KineFold*, único parâmetro constante durante essas alterações, fossem diferentes dos disponibilizados pelo NNDB. Para confirmar essa afirmação, os valores para esses pareamentos finais incompatíveis foram obtidos por engenharia reversa. Assim, os novos resultados para estruturas sem pseudonós tornaram-se equivalentes aos do servidor, salvo

pequenas diferenças devido, provavelmente, à aproximações numéricas realizadas pelos programas. A Figura 29 apresenta alguns exemplos de resultados obtidos a partir dessa abordagem.

Figura 29 – Esquema da descrição termodinâmica para *KineFold* sem pseudonós



Legenda: Descrições termodinâmicas detalhadas para cada estrutura. O valor de sua variação de energia livre de Gibbs encontra-se em negrito. O valor no quadro equivale ao obtido via *KineFold*. Todas as redes encontradas em estruturas secundárias sem pseudonós equivalem à rede-fechada-0 (rf_0). O valor entre chaves equivale ao comprimento do SUR. PFI: Pareamento final incompatível. (a) Grampo simples. (b) Grampo duplo. (c) Grampo com protuberância. (d) Combinação de protuberância com vários grampos. Fonte: Estruturas geradas pelo servidor web *KineFold* (XAYAPHOUMMINE; BUCHER; ISAMBERT, 2005) e cálculos organizados pelo próprio autor

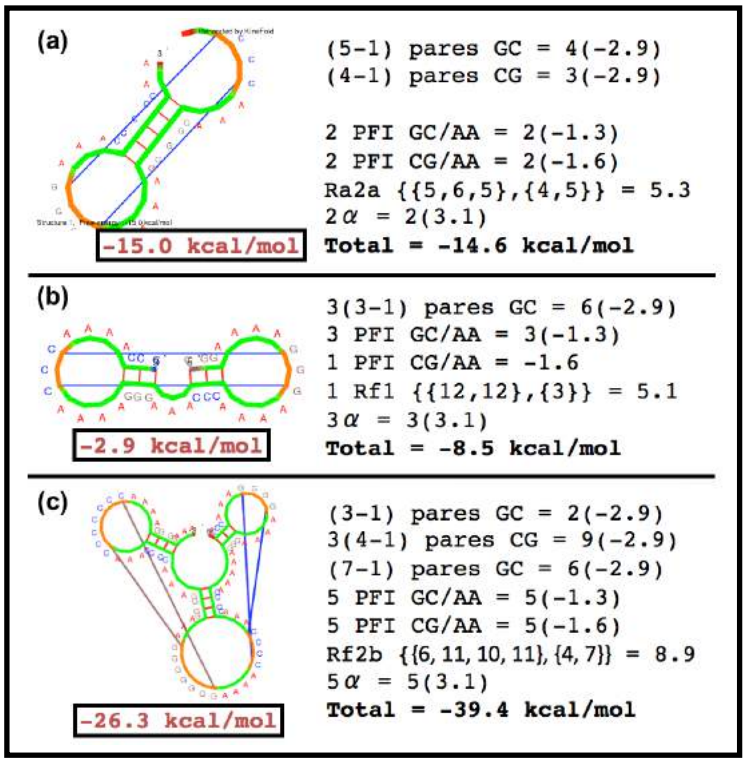
4.2 Pseudonós

KineFold

Comparando estruturas com pseudonós, nossa implementação passou a apresentar diferenças mais significativas em relação ao *KineFold*, como podemos observar na Figura 30. As diferenças se intensificavam com o aumento da complexidade

dos pseudonós, com resultados satisfatórios apenas para estruturas classificadas como redes-abertas-1. Concluímos que o termo referente a contribuição entrópica devido a estrutura do "gel reticulado de granularidade grossa" não estava corretamente implementado, pois o comportamento do restante das funções estava de acordo com Xayaphoummine, Bucher e Isambert (2005). Informações mais detalhadas sobre o comportamento desse "gel" não foram encontradas na literatura especializada, e também não há discussão do método nem mesmo na tese de um dos autores do método em questão (XAYAPHOUMMINE, 2004).

Figura 30 – Esquema da descrição termodinâmica para *KineFold* considerando pseudonós

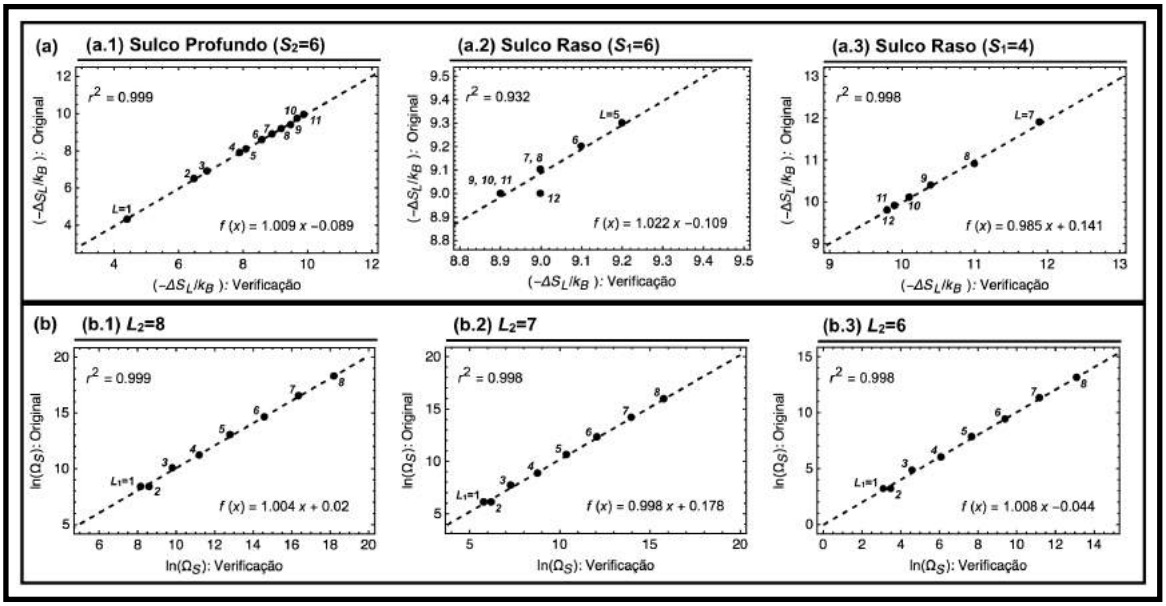


Legenda: Descrições termodinâmicas detalhadas para cada estrutura. O valor de ΔG° encontra-se em negrito. O valor no quadro equivale ao obtido via *KineFold*. Os valores presentes no primeiro subgrupo entre chaves após a classificação da rede equivale ao comprimento dos SURs, enquanto o segundo equivale ao comprimento dos SBRs. PFI: Pareamento final incompatível. (a) Rede-aberta-2a ou *pseudonó do tipo-H*. A diferença nos valores de energia livre para esse caso é pequena. (b) Rede-fechada-1, também chamada de *kissing hairpin*. Diferença moderada entre os valores encontrados. (c) Rede-fechada-2b. Diferença significativa entre as implementações. Fonte: Estruturas geradas pelo *servidor web KineFold* (XAYAPHOUMMINE; BUCHER; ISAMBERT, 2005) e cálculos organizados pelo próprio autor

Vfold

A Figura 31 apresenta os gráficos de correlação entre os valores das entropias obtidas a partir de nossas simulações para a disposição dos SURs em uma *rede de diamante* e os valores publicados por Cao e Chen (2006). A Figura 31a apresenta as entropias dos SURs para os diferentes sulcos do pseudonó, enquanto a Figura 31b o logaritmo natural do número de conformações para os SURs de um pseudonó composto por SBRs de 5 e 7 pb. Os resultados obtidos por nossa implementação apresentam coeficientes de determinação muito próximos de $r^2 = 1$, indicando que o algoritmo foi corretamente implementado.

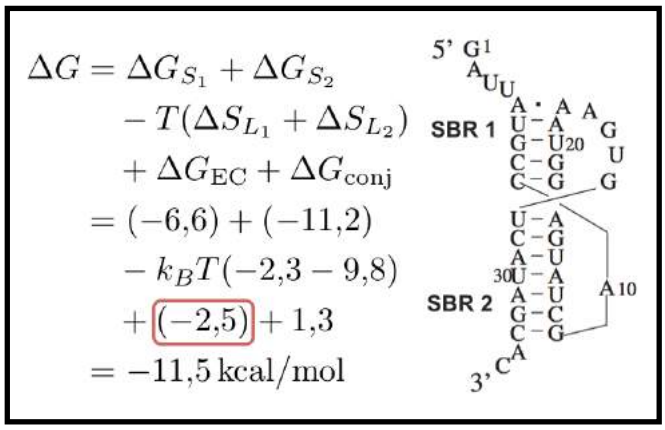
Figura 31 – Gráficos comparando os resultados obtidos nas simulações e os originais publicados para o *Vfold*



Legenda: Resultados originais pareados aos nossos resultados. No canto inferior direito de cada gráfico encontra-se a curva obtida por regressão linear, todas próximas de $f(x) = x$; no canto superior esquerdo temos o valor do coeficiente de determinação r^2, bastante próximo de $r^2 = 1$. (a) Variação da entropia para cada sulco de um pseudonó, considerando SBRs de comprimento S_2 e S_1, dado um SUR correspondente de comprimento L. (b) Resultados do cálculo de entropia conformacional para pseudonós constituídos por L_1 nt no SUR do sulco profundo e L_2 nt para o SUR do sulco raso, considerando SBRs de comprimento $S_1 = 5$ pb e $S_2 = 7$ pb, onde Ω_S é o resultado do produto entre o número de conformações encontradas para os SURs L_1 e L_2. Fonte: Dados originais obtidos de Cao e Chen (2006) e gráficos produzidos pelo próprio autor

A Figura 32 apresenta o cálculo da variação da energia livre para a formação do pseudonó *T4-35*, um dos exemplos publicado por Cao e Chen (2006). O valor obtido é inferior aos $-8{,}1$ kcal/mol encontrado experimentalmente na presença de 3 mM de Mg^{2+} e 50 mM de Na^{+}. Cao e Chen (2006) justificam essa diferença argumentando que a contribuição energética devido ao empilhamento coaxial é o fator problemático, indicando, portanto, que essa conformação não deve estar presente nessa estrutura. Considerações semelhantes foram realizadas para as outras sequências apresentadas no artigo, e esse fator era desprezado ou considerado quando pertinente.

Figura 32 – Esquema para cálculo da variação de energia livre segundo *Vfold*



Legenda: Sequência *T4-35*, utilizando a abordagem e os parâmetros de Cao e Chen (2006), que resultariam em uma variação de energia livre de $-11{,}5$ kcal/mol, dado $k_B T = 0{,}62$ kcal/mol. O valor experimental para essa estrutura é de $-8{,}1$ kcal/mol. Em destaque, a contribuição devido ao empilhamento coaxial, que, segundo os autores, não deve estar presente na estrutura experimental, indicando que esse fator deve ser removido. O valor sem essa contribuição passa a ser de $-9{,}0$ kcal/mol. Fonte: Modificado de Cao e Chen (2006)

Modelo adaptado

Para verificarmos o resultado de nossas adaptações, realizamos o cálculo da energia livre para o pseudonó T4-35. Na primeira aproximação, utilizando parâmetros atualizados, teremos

$$\Delta G^{\circ} = \Delta G^{\circ}_{S_1} + \Delta G^{\circ}_{S_2} - T(\Delta S_{L_1} + \Delta S_{L_2}) + \Delta G^{\circ}_{\mathrm{conj}} + \Delta G^{\circ}_{\mathrm{Ex}} + \Delta G^{\circ}_{\mathrm{ini}}. \tag{4.1}$$

Não incluímos a ligação A-A que Cao e Chen (2006) consideraram presente no primeiro SBR da estrutura, representada por um pequeno quadrado preenchido na Figura 32, o que resultou em valores ligeiramente diferentes para alguns fatores. Em nossos cálculos, $\Delta G^\circ_{S_1} = -6{,}3$ kcal/mol e $\Delta G^\circ_{S_2} = -11{,}9$ kcal/mol são as energias livres para os SBRs segundo o modelo dos vizinhos mais próximos enquanto $-T(\Delta S_{L_1} + \Delta S_{L_2}) = -k_B T(-2{,}3 - 9{,}2) = +7{,}1$ kcal/mol e $\Delta G^\circ_{\text{conj}} = +1{,}3$ kcal/mol são obtidos seguindo o modelo *Vfold*. Removemos o termo $\Delta G^\circ_{\text{EC}}$, e, em seu lugar, incluímos a variação de energia devido aos SRUs nas extremidades dos SBRs, nesse caso três finais pendentes e um pareamento final incompatível, $\Delta G^\circ_{\text{Ex}} = -0{,}8 - 1{,}7 - 0{,}2 - 1{,}7 = -4{,}4$ kcal/mol. Finalmente, adicionamos uma energia inicial para a formação do pseudonó, $\Delta G^\circ_{\text{ini}} = +4{,}1$ kcal/mol. O valor final de energia encontrado será de $\Delta G^\circ = -10{,}1$ kcal/mol, próximo do encontrado por Cao e Chen (2006). Obtivemos resultados semelhantes para as outras estruturas apresentadas no artigo de referência citado, ou seja, com valores inferiores ao esperado pelo modelo original. Incluímos, então, o novo termo energético, resultado da curvatura dos SURs.

Resolvendo as equações para θ apresentadas na Equação (3.7) e substituindo seus valores na Equação (3.5) do modelo para a energia armazenada devido à curvatura dos SURs, encontramos $U_{curv} \approx k_B T$ para os pseudonós do tipo-H de dimensões semelhantes ao T4-35. O valor dessa contribuição no cálculo da energia da estrutura presente na Figura 32 será de $U_{curv} = 0{,}8$ kcal/mol, o que resultará num ΔG° final de $-9{,}3$ kcal/mol, mais próximo do obtido pelo modelo original de Cao e Chen (2006). A inclusão desse termo também levou à valores de energia livre mais próximos para as outras estruturas apresentadas no mesmo trabalho.

O modelo Vfold *modificado foi escolhido como padrão para o programa*, devido à maior aderência de seus resultados obtidos nessa etapa de verificação. É possível alterar a forma que o programa lida com os pseudonós, mas os testes de validação apresentados a seguir foram realizados considerando as regras para esse modelo.

4.3 Tempo de execução do programa

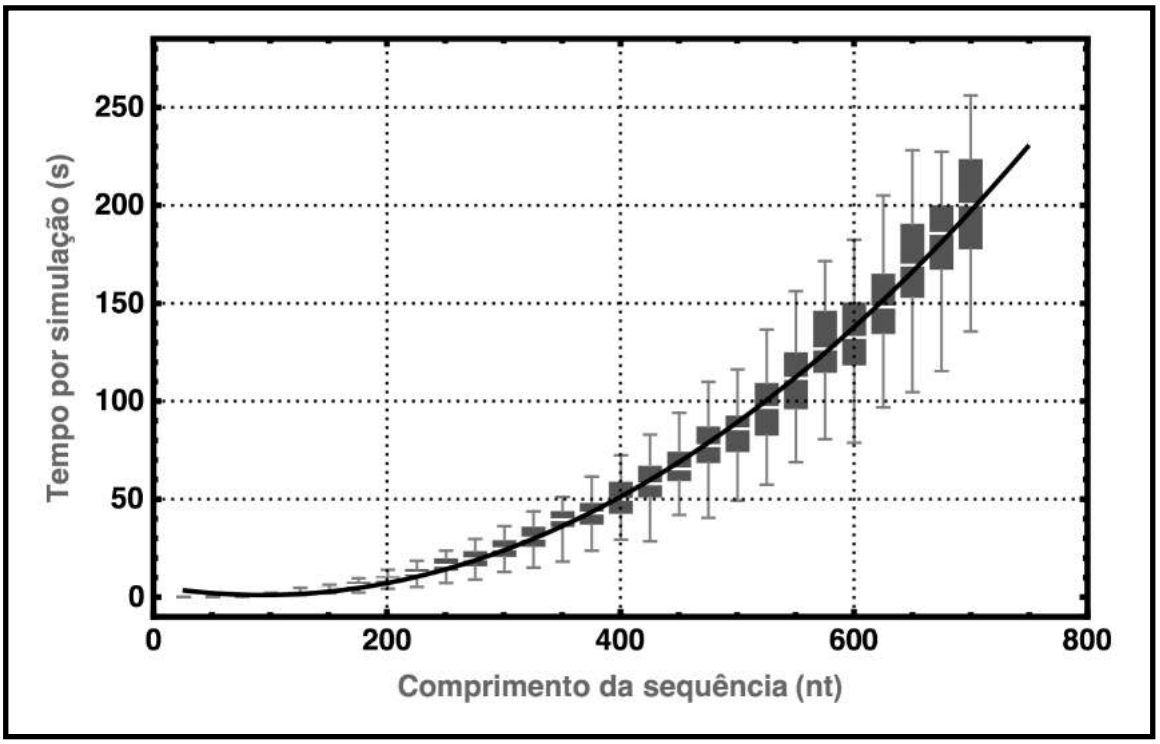
A Figura 33 representa o gráfico com as distribuições dos tempos de simulação de cada uma das 96 simulações de sequências aleatórias realizadas para

cada comprimento múltiplo de 25. Encontramos, através de regressão polinomial segundo o método dos mínimos quadrados, que o tempo médio necessário para simulação de uma sequência de comprimento l será dado por

$$T(l) = 5{,}3 \times 10^{-4} l^2 - 0{,}100 l + 5{,}72$$

. A curva $T(l)$ também está representada na Figura 33.

Figura 33 – Gráfico da distribuição dos tempos absolutos de processamento em função do comprimento da sequência

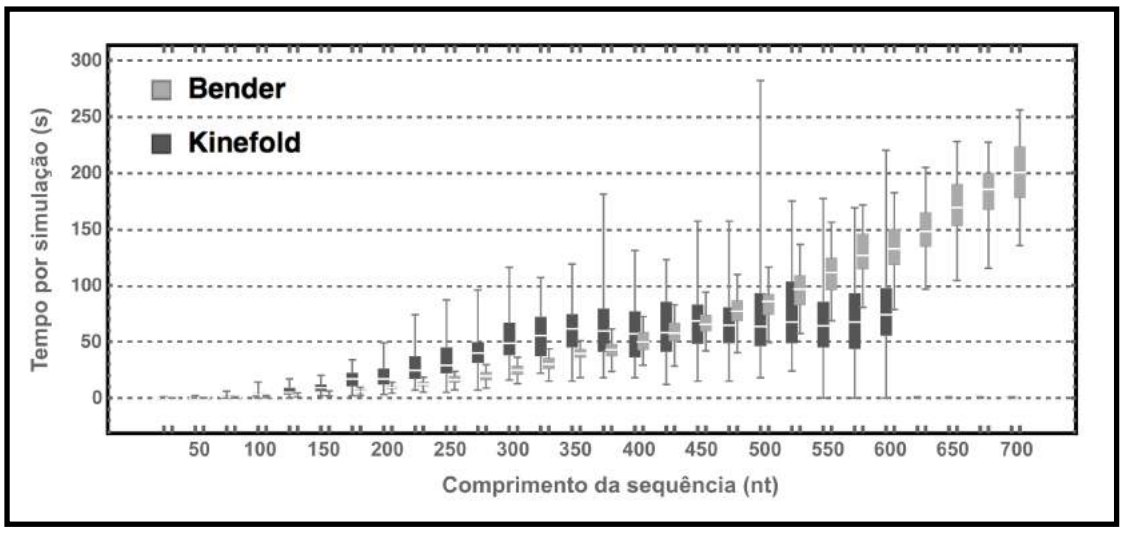


Legenda: Diagrama de caixa representando a distribuição dos tempos necessários para simulação de cada uma das 96 sequências aleatórias de comprimento l. A curva contínua é dada por $T(l) = 5{,}3 \times 10^{-4} l^2 - 0{,}100 l + 5{,}72$. Fonte: Produzido pelo próprio autor

Para compararmos o desempenho de nossa implementação, realizamos simulações cotranscricionais das mesmas sequências utilizando a versão de 2016 do programa *Kinefold*. As distribuições dos tempos necessários para ambas as abordagens estão apresentadas na Figura 34. *Kinefold* apresentou dois problemas: sequências a partir de 350 nt são truncadas, e apenas os primeiros 350 nt são simulados, o que justifica a assíntota observada para os valores das medianas das distribuições com comprimentos entre 350 e 600 nt; e sequências com mais de 600 nt

são abortadas. Ambas as medidas foram provavelmente adotadas devido ao fato que o número de estruturas possíveis cresce muito rapidamente em função do comprimento da sequência em estudo quando se considera a presença de pseudonós. Veremos na Seção 4.5 como a capacidade preditiva dos modelos também é reduzida substancialmente com o aumento do comprimento das sequências, o que justificaria as restrições adotadas pelo *Kinefold*. Entretanto, qualitativamente, nosso programa apresenta eficiência temporal superior à eficiência exibida pelo *Kinefold*, demonstrando que nossa implementação é competitiva segundo esse critério. Provavelmente, por considerar uma velocidade de transcrição constante durante o alongamento, o programa *Kinefold* realiza um número de simulações superior ao de nossa abordagem cotranscricional. Por incluirmos uma etapa de determinação dos sítios de pausa, e permitindo o dobramento apenas nessas regiões, reduzimos o número de iterações, assim como o espaço de conformações estruturais. Contudo, destacamos também que o programa *Kinefold* ainda possui como parâmetro o número de pseudonós máximo que a estrutura pode apresentar: seu padrão, utilizado nas simulações, é de apenas um pseudonó. Não restringimos esse fator para as simulações em nosso programa.

Figura 34 – Gráfico da distribuição dos tempos absolutos de processamento em função do comprimento da sequência para *Kinefold* e *Bender*



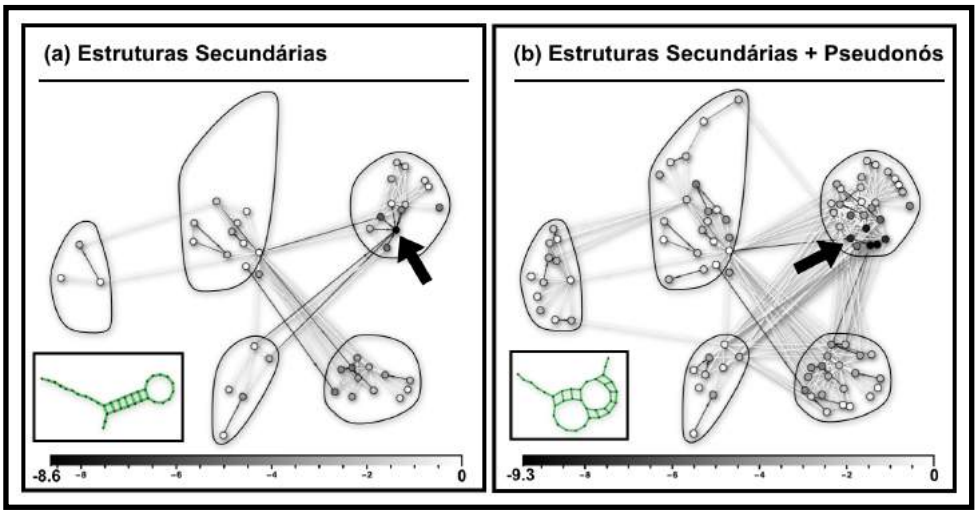
Legenda: Diagrama de caixa representando a distribuição dos tempos necessários para simulação de cada uma das 96 sequências aleatórias de comprimento l. Fonte: Produzido pelo próprio autor

4.4 Cadeia de Markov para T4-35

A sequência *T4-35* é composta por 35 nucleotídeos. São 193 pareamentos permitidos, com 52 possibilidades de SBRs com três ou mais pares de bases e 1.034 combinações entre eles. Dessas combinações, 635 envolvem no máximo pseudonós do tipo-H, das quais 91 possuem $\Delta G^\circ < 0$. A Figura 35 apresenta a cadeia de Markov para essa sequência, na qual cada ponto representa uma de suas estruturas energeticamente favoráveis. As setas entre eles indicam as transições possíveis entre as estruturas, com intensidade relacionada à probabilidade de ocorrência a partir da estrutura de origem. As regiões delimitadas determinam grupos de estruturas com características semelhantes. A estrutura MFE encontrada para cada caso está indicada. Nosso programa pode retornar as estruturas presentes na Figura 35b, que apresenta a estrutura da Figura 32 como estrutura MFE. A Figura 35a representa a cadeia de Markov obtida caso considerássemos apenas as estruturas secundárias. Destaca-se a ausência de várias estruturas e transições nesse grafo. Note que a estrutura MFE encontrada na ausência de pseudonós equivale ao SBR mais longo presente no pseudonó *T4-35*.

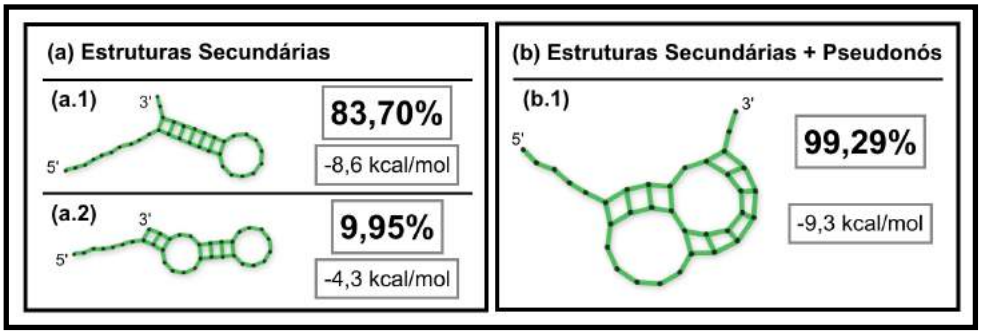
Os principais *estados absorventes* das cadeias estão representados na Figura 36, com suas respectivas probabilidades de absorção e energias livre de formação. A estrutura (a.1) equivale à estrutura MFE dentre as estruturas secundárias e como o estado absorvente mais provável dentre os 20 possíveis. No caso da cadeia de Markov que considera pseudonós, a absorção pelo estado (b.1), que equivale à estrutura de referência, é quase absoluta, de forma que os outros 36 estados absorventes podem ser desconsiderados. Podemos inferir, a partir desses resultados, que o caminho de dobramento em direção à estrutura MFE passa preferencialmente pelo grampo mais longo, e em seguida um novo SBR se forma a partir dos nucleotídeos não pareados presentes em sua alça, originando o pseudonó.

Figura 35 – Grafo da cadeia de Markov para *T4-35*



Legenda: Os pontos, cujas posições não variam entre as subfiguras, representam as estruturas com energia livre de formação com valores que variam de acordo com a escala presente. As ligações direcionadas representam as transições, com sua intensidade proporcional à sua respectiva probabilidade de ocorrência. As regiões delimitadas agrupam estruturas com características semelhantes. A posição da estrutura MFE na rede está indicada para cada caso e esquematizadas no canto de cada quadro. (a) Cadeia considerando apenas estruturas secundárias e (b) Cadeia incluindo pseudonós. Fonte: Produzido pelo próprio autor

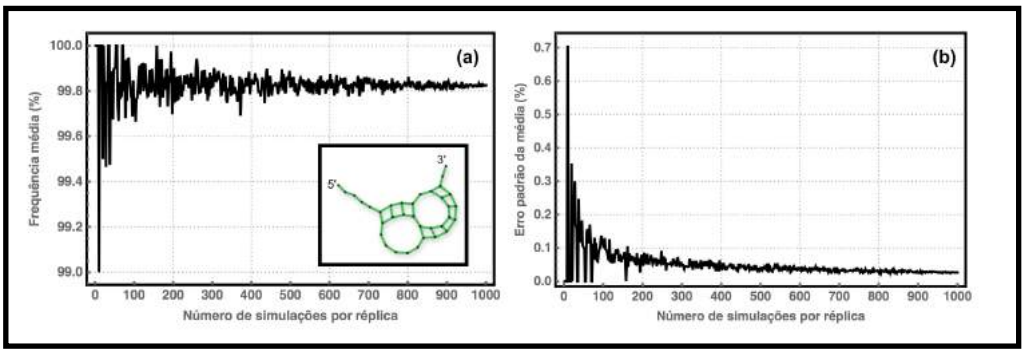
Figura 36 – Esquema das estruturas correspondentes aos estados absorventes da cadeia de Markov para T4-35



Legenda: Baseado nas *cadeias de Markov* da Figura 35. Ao lado de cada estrutura temos a probabilidade de absorção do sistema para o estado em questão, e sua respetiva variação de energia livre de formação. (a) Cadeia considerando apenas estruturas secundárias e (b) Cadeia incluindo pseudonós. Fonte: Produzido pelo próprio autor

Apesar das análises realizadas através da cadeia de Markov apresentarem resultados pertinentes, ela se torna inviável para sequências mais longas por basear-se na busca exaustiva dos estados permitidos. Nossa implementação constrói um subconjunto da Cadeia de Markov completa de uma sequência de RNA, pois considera apenas os primeiros vizinhos e as probabilidades de transição do estado atual a cada iteração. Se o número de iterações for suficiente, identificaremos um estado absorvente a cada simulação realizada. Realizando réplicas dessas simulações, a frequência de um determinado estado final tenderá ao valor de absorção daquele estado na cadeia de Markov completa, com custo computacional frequentemente inferior ao custo necessário para a determinação da cadeia de Markov completa. Os resultados de nossas simulações de renaturação para esse exemplo estão representados na Figura 37. Na Figura 37a observamos como a frequência média da estrutura representada na Figura 36b.1 varia de acordo com o número de simulações realizadas por réplica. A Figura 37b apresenta a variação do erro padrão dessas médias com o aumento do número de simulações por réplica. Os valores finais obtidos com 1.000 simulações por réplica foram de 99,8±0,03%, resultado compatível com a absorção de 99,29% obtida através da cadeia de Markov completa. Destacamos que o intervalo de variação da frequência média foi inferior a 1%, o que indicaria a alta tendência do sistema assumir o estado de MFE mesmo com um número reduzido de simulações.

Figura 37 – Gráfico da frequência média e erro padrão da média para a estrutura MFE da sequência T4-35

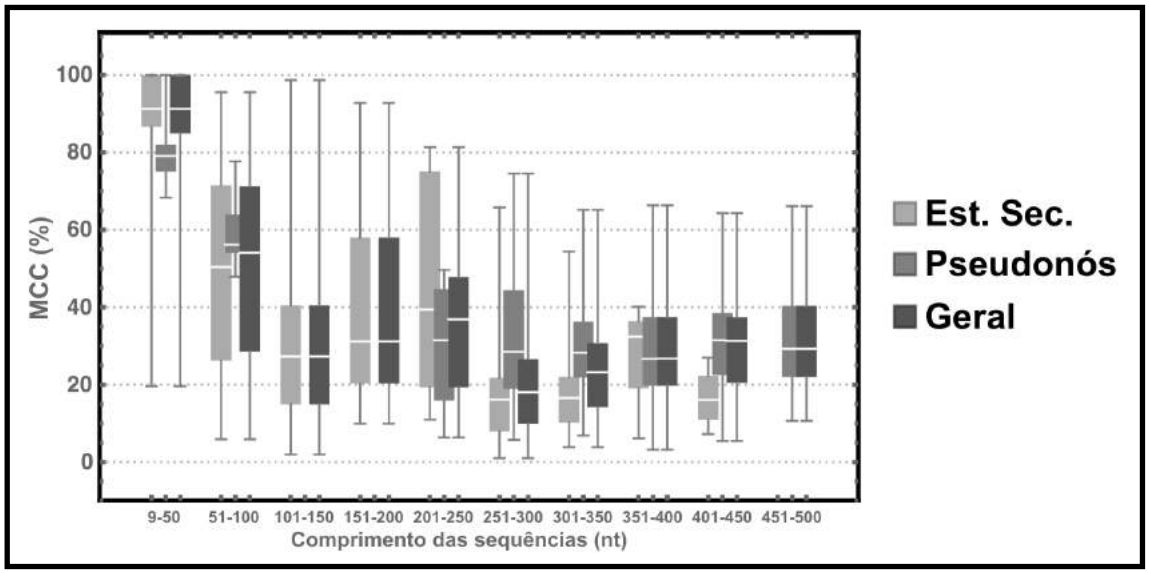


Legenda: Valores para a estrutura representada no quadro, equivalente à estrutura representada Figura 36b.1. (a) Frequência média para as 20 réplicas. (b) Erro padrão das médias apresentadas na Figura (a). Fonte: Produzido pelo próprio autor

4.5 Banco de dados *RNAStrand*

O diagrama de caixas da Figura 38 representa a distribuição dos valores de MCC calculado a partir da relação entre as previsões de nossas simulações cotranscricionais e a estrutura presente no banco de dados *RNAStrand*. Os dados estão agrupados de acordo com o comprimento das sequências, e representam apenas os casos nos quais alguma previsão pode ser realizada (PPV > 0). Conforme esperado, a dificuldade de previsão das estruturas aumenta conforme o número de nucleotídeos da sequência de estudo. Além disso, estruturas sem pseudonós apresentam um maior desafio para o algoritmo, uma vez que esse tipo de confromação é considerada em nossas simulações, aumentando o espaço de possibilidades e, consequentemente, o número de falsos positivos.

Figura 38 – Gráfico da distribuição dos MCCs para diferentes comprimentos de RNA obtidos via *Bender*



Legenda: Diagrama de caixa representando a distribuição dos valores de MCC para diferentes comprimentos de RNA, obtidos a partir da comparação entre os resultados de nossa implementação e as estruturas presentes no banco *RNAStrand*. Alguns intervalos estipulados não apresentaram sequências para análise. Fonte: Produzido pelo próprio autor

Na Tabela 5 temos a distribuição das sequências que apresentaram PPV > 0 dado seu intervalo de comprimento e tipo de estrutura. As sequências restantes possuem características que contrariam as regras restritivas impostas na etapa de

determinação de SBRs permitidos (Seção 3.1) ou que não guardam informação em relação ao seu caminho de dobramento. Entretanto, representam apenas ≈8% das 1241 sequências estudadas. As maiores dificuldades encontradas estão nos intervalos entre 100 e 150 nt, que engloba apenas estruturas secundárias (37 estruturas) e entre 250 e 300 nt (22 estruturas).

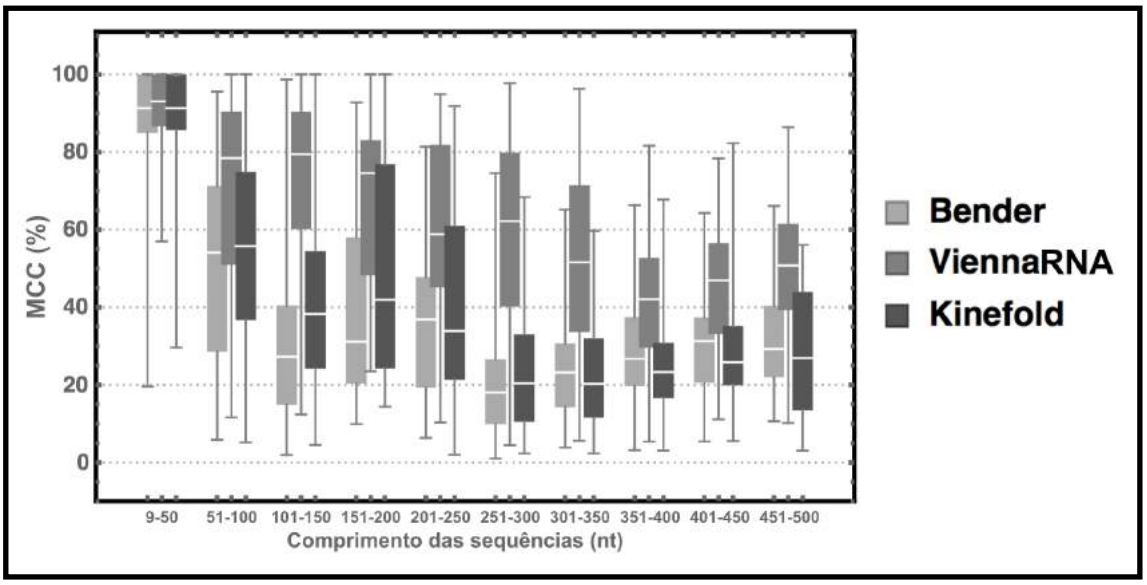
Tabela 5 – Distribuição das sequências que apresentaram PPV > 0 em relação ao comprimento e tipo de conformação

	9 – 50	51 – 50	101 – 150	151 – 200	201 – 250
Est. Secundárias	94% de 122	83% de 108	80% de 180	70% de 10	92% de 13
Pseudonós	100% de 9	90% de 10	0% de 0	0% de 0	100% de 12
Geral	95% de 131	84% de 118	80% de 180	70% de 10	96% de 25
	251 – 300	301 – 350	351 – 400	401 – 450	451 – 500
Est. Secundárias	80% de 108	96% de 87	100% de 4	100% de 4	0% de 1
Pseudonós	97% de 30	99% de 108	100% de 342	100% de 74	95% de 19
Geral	83% de 138	98% de 195	100% de 346	100% de 78	90% de 20

Fonte – Produzido pelo próprio autor

A Figura 39 apresenta um comparativo entre as distribuições dos MCCs gerais (estruturas secundárias e pseudonós) para diferentes intervalos de comprimento e diferentes abordagens. O mesmo critério da Figura 38: apenas previsões com PPV > 0 estão apresentadas. Assim, teremos representadas no gráfico para *Bender-ViennaRNA-Kinefold*: 545-625-611 estruturas secundárias (86%-98%-92% das 637 sequências), 600-600-590 pseudonós (99%-99%-97% das 604 sequências) e 1145-1225-1201 estruturas no geral (92%-98%-97% das 1241 sequências). A Figura 40 representa os MCCs apenas de acordo com a natureza da estruturas: o teste de Wilcoxon apresentou valores $p < 10^{-2}$ para todos os nove pareamentos entre as distribuições dessa Figura, o que atesta a hipótese alternativa de que as distribuições diferem em localização, ou seja, há diferença entre as medianas. Qualitativamente podemos notar que o que o algoritmo *ViennaRNA* apresenta maior poder de predição em relação aos outros dois, que, entretanto, apresentam comportamentos semelhantes. O índice MCC reduz significantemente com o aumento do comprimento da sequência estudada para todos os programas, mas o algoritmo *ViennaRNA* apresenta maior poder preditivo mesmo quando consideramos apenas pseudonós.

Figura 39 – Gráfico da distribuição dos MCCs gerais para diferentes comprimentos de RNA obtidos via *Bender*, *ViennaRNA* e *Kinefold*

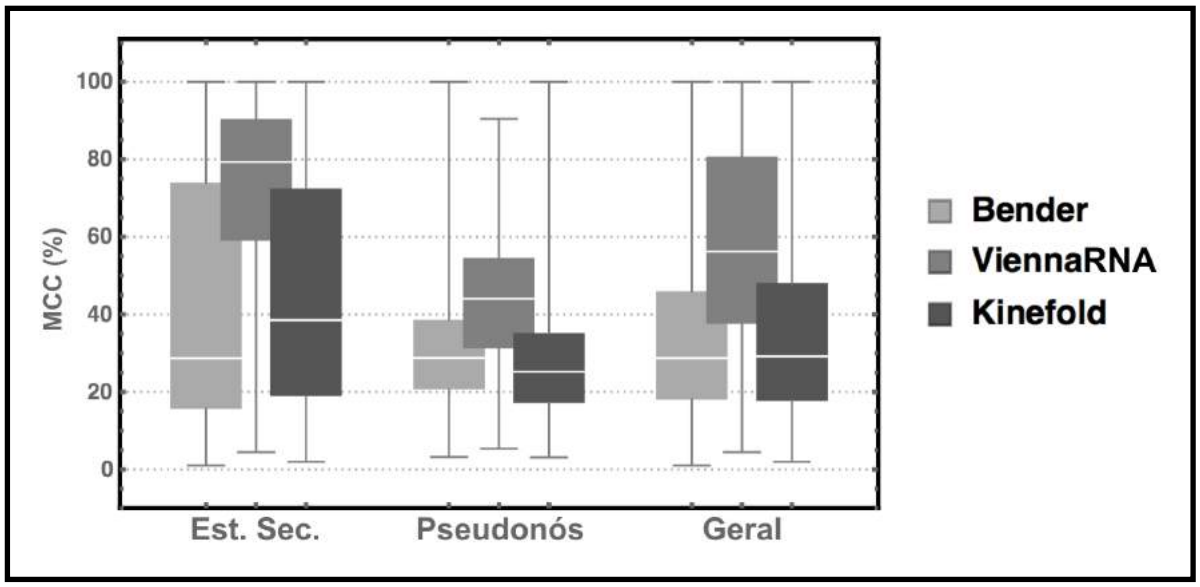


Legenda: Diagrama de caixa representando as distribuições dos valores de MCC para diferentes comprimentos de RNA, obtidas a partir da comparação entre os resultados de nossa implementação e as estruturas presentes no banco *RNAStrand*. Alguns intervalos estipulados não apresentaram sequências para determinadas estruturas.
Fonte: Produzido pelo próprio autor

A abordagem proposta pelo *ViennaRNA*, capaz de identificar a estrutura MFE através de programação dinâmica, é eficiente na determinação das estruturas finais das moléculas de RNA, o que indica que essas conformações finais guardam pouca informação em relação às suas *estruturas efêmeras*, provavelmente identificadas através das outras duas abordagens cotranscricionais. Além disso, apesar de não identificar os pseudonós integralmente, o programa ainda pode prever corretamente parte dessas estruturas, uma vez que os pseudonós são resultado da interação entre duas ou mais estruturas secundárias. O sucesso do programa *ViennaRNA* na predição de pseudonós é mais um argumento a favor da hipótese do dobramento hierárquico das moléculas de RNA (TINOCO; BUSTAMANTE, 1999). Também não podemos destacar o fato de que as sequências com pseudonós são mais longas, portanto é comum apresentarem *outras* subestruturas *além* dos pseudonós, mais facilmente identificáveis pelo algoritmo dinâmico desse programa. Os valores mais baixos de MCC obtidos por nosso método e pelo *KineFold* refletem o grande número

de falsos positivos obtidos, resultado da superfície de energia livre extremamente acidentada que as sequências apresentam. Restringindo o número de possibilidades, *ViennaRNA* obteve o maior número de verdadeiros positivos entre as abordagens.

Figura 40 – Gráfico da distribuição geral dos MCCs para diferentes tipos de estruturas de RNA presentes no *RNAStrand* via *Bender*, *ViennaRNA* e *Kinefold*

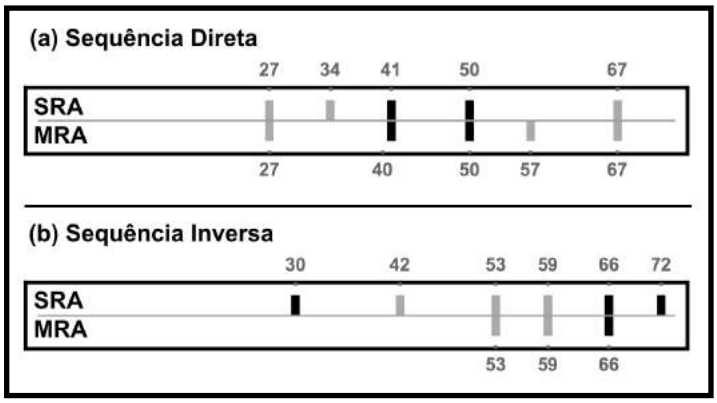


Legenda: Diagrama de caixa representando a distribuição geral dos valores de MCC para diferentes conformações, obtido a partir da comparação entre os resultados de nossa implementação e as estruturas presentes no banco *RNAStrand*. Fonte: Produzido pelo próprio autor

4.6 Análise de sequências com estruturas competitivas

A primeira etapa da simulação cotranscricional envolve determinar quais serão os sítios de pausa. Como não há registro experimental, os sítios de pausa teóricos encontrados pelo algoritmo estão apresentados na Figura 41, identificados como pausas do tipo 1, relacionadas à estabilidade da bolha de transcrição, ou do tipo 2, relacionadas ao *backtracking*.

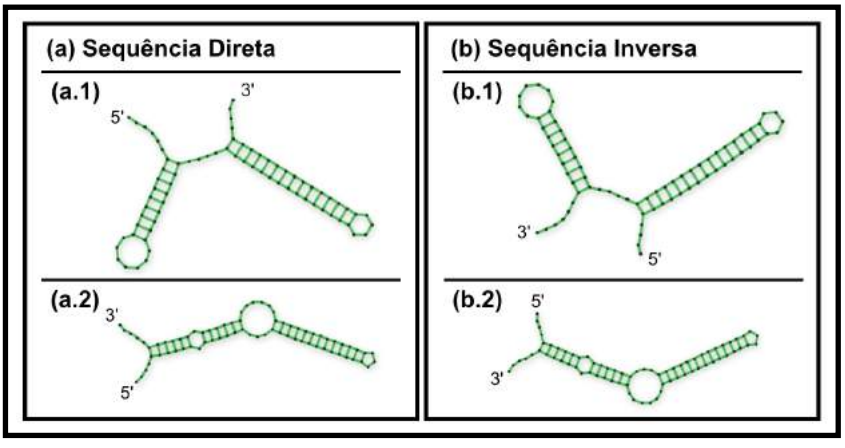
Figura 41 – Esquema dos sítios de pausa teóricos para seq. direta e inversa



Legenda: Comparação entre os sítios de pausa determinados via SRA e MRA. Em cinza temos as pausas do tipo 1 e em preto as pausas do tipo 2. (a) Versão direta da sequência em estudo. (b) Versão inversa da sequência em estudo. Fonte: Produzido pelo próprio autor

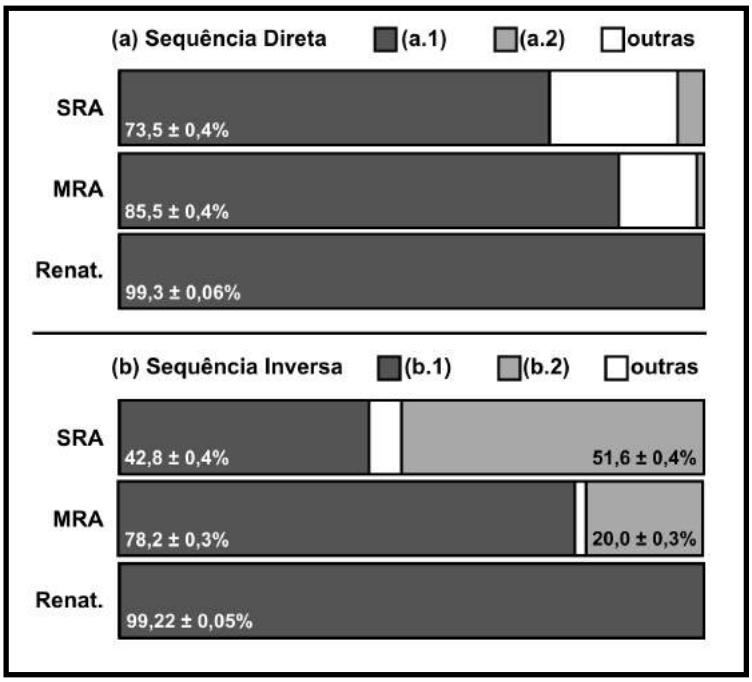
As estruturas mais frequentemente obtidas via simulações cotranscricionais equivalem as obtidas experimentalmente, e estão representadas na Figura 42. A Figura 43 apresenta as proporções resultantes para cada conjunto de simulações: considere a nomenclatura apresentada na Figura 42 como referência.

Figura 42 – Esquema das estruturas resultantes da simulação cotranscricional da sequência direta e inversa



Legenda: (a) Versão direta da sequência: (a.1) Estrutura MFE, $\Delta G^\circ = -41{,}8$ kcal/mol; (a.2) Estrutura alternativa, $\Delta G^\circ = -39{,}9$ kcal/mol. (b) Versão inversa da sequência em estudo: (b.1) Estrutura MFE, $\Delta G^\circ = -41{,}8$ kcal/mol; (b.2) Estrutura alternativa, $\Delta G^\circ = -39{,}4$ kcal/mol. Fonte: Produzido pelo próprio autor

Figura 43 – Esquema da frequência das estruturas obtidas pelas simulações da sequência direta e inversa em porcentagem



Legenda: Renat.: Renaturação. Nomenclatura conforme presente na Figura 42. (a) Versão direta da sequência em estudo. Valores para SRA: (a.1) $73{,}5 \pm 0{,}4\%$, (a.2) $4{,}5 \pm 0{,}2\%$; Valores para MRA: (a.1) $85{,}5 \pm 0{,}4\%$, (a.2) $< 1\%$. (b) Versão inversa da sequência em estudo: Valores para SRA: (b.1) $42{,}8 \pm 0{,}4\%$, (b.2) $51{,}6 \pm 0{,}4\%$; Valores para MRA: (b.1) $78{,}2 \pm 0{,}3\%$, (b.2) $20{,}0 \pm 0{,}3\%$. Fonte: Produzido pelo próprio autor

Analisando a Sequência Direta, observamos que sua estrutura MFE é a mais frequente em todos os casos estudados. A permutação da pausa no sítio 34 pela pausa no sítio 57 entre SRA e MRA resultou numa probabilidade maior de encontrarmos a estrutura MFE. Em ambas, a estrutura alternativa é pouco frequente. Entretanto, o resultado que mais se aproximou do experimentalmente observado por Xayaphoummine et al. (2007) foi o encontrado pela simulação de Renaturação. Isso indica que as pausas teóricas identificadas não são tão intensas como nossas simulações estimaram. As pausas 41 e 50 são classificadas como do tipo 2, e podem ser suprimidas pelo primeiro grampo, que pode agir como uma barreira física para o *backtracking* e favorecer a transcrição, mecanismo já observado experimentalmente (ZHANG; LANDICK, 2016). O grampo começa a ser energeticamente favorável quando 26 nucleotídeos foram liberados do canal de saída

da RNAP, o que corresponde em nosso modelo à quadragésima polimerização da sequência, correlacionando o mesmo com as posições de *backtrackong* previstas. Todavia, realizamos mais uma bateria de simulações baseadas nessas considerações e obtivemos um resultado bastante semelhante à simulação SRA, mas dessa vez favorecendo a estrutura alternativa: $72{,}2 \pm 0{,}3\%$ para a estrutura MFE e $15{,}3 \pm 0{,}2\%$ para a alternativa. Esse pode ser considerado um argumento a favor da velocidade média dessa polimerase estar entre 200 e 400 nt/s nas condições apresentadas, conforme estimado por Xayaphoummine et al. (2007), sem ocorrência de pausas significativas durante o percurso.

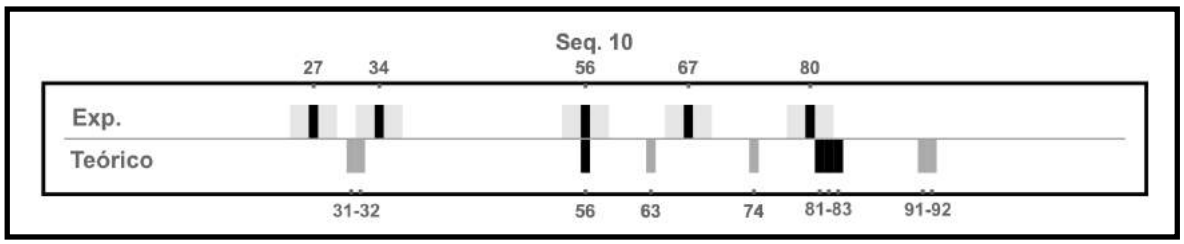
Já para a Sequência Inversa, os resultados teóricos destoaram dos experimentais logo no início. A saída para as pausas previstas via SRA se aproximou bastante do resultado experimental obtido quando o oligonucleotídeo estava presente. Esse oligonucleotídeo interfere no equilíbrio pois interage com a região central da sequência, favorecendo a formação primeiro grampo. A série de pausas previstas na primeira metade da sequência, poderia acarretar num comportamento semelhante; entretanto, observamos que a remoção dessas primeiras pausas, o que seria quase equivalente às pausas previstas via MRA, apenas deslocam o equilíbrio na direção da estrutura MFE. Portanto, o único fator que poderia corresponder à alteração do equilíbrio em favor da estrutura alternativa sem a influência de outra molécula seria a presença do sítio de pausa localizado na posição 72, no final da transcrição, presente no SRA e ausente no MRA. Nessa posição, teremos 58 nt livres para o dobramento. Realizamos mais uma série de simulações a partir desses resultados, e com a inserção forçada de apenas um sítio de pausa na posição 72, obtivemos a proporção de 9 grampos longos com protuberâncias para 1 estrutura MFE: $90{,}3 \pm 0{,}3\% : 8{,}9 \pm 0{,}2\%$. Esse resultado também condiz com a estimativa de velocidade da RNAP, sem ocorrência de pausas significativas durante quase todo o percurso, pois inserindo apenas uma pausa no término da transcrição conseguimos recuperar o resultado experimental publicado por Xayaphoummine et al. (2007).

4.7 Influência da estrutura do RNA nascente na cinética da RNAP

Como exemplo representativo das análises realizadas, os resultados obtidos para a Sequência 10 serão apresentado a seguir. Os demais resultados estão presentes no Apêndice D. As discussões envolvem os conceitos de *pausas do tipo 1*, relacionadas à estabilidade da bolha de transcrição no estado pré-translocado da RNAP em relação ao estado pós-translocado, e de *pausas do tipo 2*, que ocorrem devido ao movimento da RNAP no sentido oposto ao da transcrição (*backtracking*).

A Sequência 10 é composta por 155 ribonucleotídeos e sua transcrição apresenta cinco sítios de pausa experimentalmente determinados, obtidos através da análise de géis de poliacrilamida, com incerteza de ±3 nt. A Figura 44 apresenta a distribuição desses sítios, assim como os sítios de pausa identificados durante as simulações da transcrição dessa sequência através da SRA. As pausas previstas para os sítios 31-32, 63, 74 e 92-93 são do tipo 1, enquanto as pausas nos sítios 56 e 81-82-83 são do tipo 2.

Figura 44 – Esquema da distribuição dos sítios de pausa para a Sequência 10



Legenda: Sítios de pausa experimentais: 27, 34, 56, 67 e 80, com região de incerteza de ±3 nt apresentada. Pausas teóricas do tipo 1 estão representadas em cinza, enquanto pausas do tipo 2 estão representadas em preto. Fonte: Produzido pelo próprio autor

A Figura 45 representa o caminho de dobramento desse segmento segundo nossa implementação, representando a conformação da molécula nos sítios de pausas experimentalmente determinados. As estruturas presentes são obtidas a partir dos SBRs mais prováveis entre as réplicas. Entre os dois primeiros sítios de pausa, 27 e 34, não observamos SBRs. Logo, temos três posições de interesse:

a) Figura 45a: Observamos a formação da primeira estrutura quando a RNAP pausa no sítio 56. Trata-se de um grampo longo com protuberâncias, que

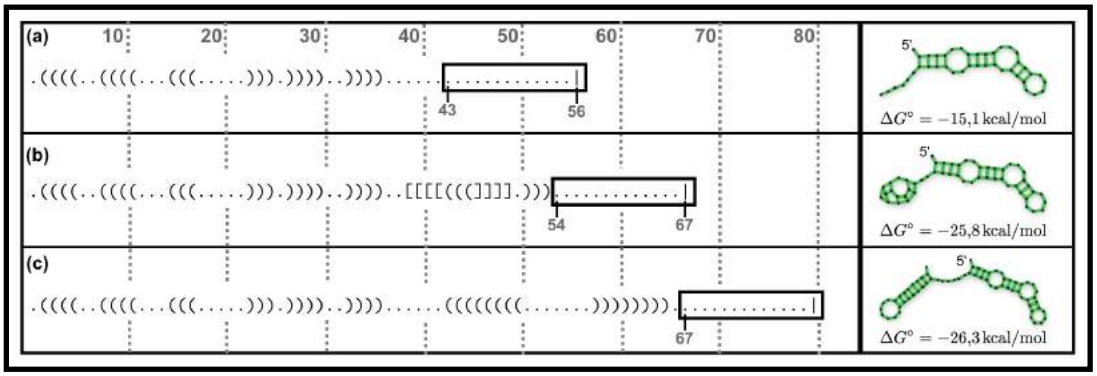
dista 6 nt do canal de saída da enzima naquele momento. Esse grampo torna-se energeticamente favorável quando o sítio ativo da RNAP está localizado no posição 39, ou seja, quando o transcrito livre compreender pelo menos 25 nt.

b) Figura 45b: Em seguida, ocorre a formação de um pequeno *pseudonó* efêmero, que surge imediatamente antes da transcrição do ribonucleotídeo 67, pausa experimentalmente verificada mas não identificada na transcrição *in silico*.

c) Figura 45c: Na última pausa, no sítio 80, observamos que o pseudonó se desfaz em favor da formação de um grampo longo, que poderia se apresentar depois da transcrição do nucleotídeo 72, pois observamos um SBR entre os SURs 43-50 e 58-65. Entretanto, dada a similaridade energética entre esse grampo e o pseudonó, essa nova conformação passa a ser energeticamente favorável apenas quando o transcrito livre compreender 65 nt, ou seja, apenas na região do sítio de pausa, 79-80.

Na Figura 45a a pausa do tipo 2 no sítio 56 aparenta não ser inibida pelo SBR formado, provavelmente devido ao SUR de 6 nt presente entre o SBR e canal de saída da enzima. Já a Figura 45b apresenta uma situação na qual podemos supor que a formação daquele pseudonó efêmero possa interferir no equilíbrio do Complexo de Alongamento Transcricional e esteja relacionada à pausa observada naquela região. Finalmente, nossas simulações apontaram pausas do tipo 2 na região 81-82-83: o grampo apresentado Figura 45c pode atuar como inibidor do *backtracking* e favorecer a transcrição nessa região. Entretanto, esse mecanismo de supressão poderá atuar apenas quando a RNAP recuar 2-3 nt, na região da pausa experimental observada no sítio 80.

Figura 45 – Esquema do caminho de dobramento para a Sequência 10



Legenda: Estruturas mais prováveis obtidas a partir das 20 réplicas realizadas em cada sítio de pausa experimental. Cada item apresenta as estruturas formadas na notação de pontos-e-parênteses e em um esquema de seu arranjo estrutural, com seu respectivo valor de variação da energia livre. O quadro sobre a representação de pontos-e-parênteses delimita a região "protegida" pela RNAP (híbrido RNA-DNA e canal de saída do RNA), com o caractere "|" indicando a posição do sítio ativo da enzima. (a) Pausa no sítio 56; a RNAP impede o pareamento dos nucleotídeos da posição 43 até a 56. (b) Pausa no sítio 67; a RNAP impede o pareamento dos nucleotídeos da posição 54 até a 67. (c) Pausa no sítio 80; a RNAP impede o pareamento dos nucleotídeos da posição 67 até a 80. Fonte: Produzido pelo próprio autor

5 Conclusões

Um modelo para predição de estruturas secundárias e pseudonós baseado em simulações estocásticas foi implementado em *Wolfram Language*. O programa compreende parâmetros atualizados para a variação de energia livre de Gibbs para pareamentos entre fitas de ácidos nucleicos, regras para determinação de estruturas secundárias baseadas nos dados do banco NNDB (TURNER; MATHEWS, 2009), e regras para determinação da entropia para pseudonós baseadas no modelo *Kinefold* e no modelo *Vfold*. O método de Isambert e Siggia (2000) considera pseudonós com vários níveis de complexidade, mas o modelo conformacional baseado em ligações virtuais *Vfold* (LIU; CHEN, 2010), mostrou-se mais apropriado por tratar com um nível de sofisticação adequado o tipo de pseudonó mais frequentemente observado em RNAs naturais, o *pseudonó do tipo-H*. Modificamos algumas etapas de seu algoritmo e incluímos um fator entrópico baseado na energia armazenada devido à curvatura do laço de RNA.

O modelo também permite a inclusão dos sítios de pausas transcricionais, e assim podemos realizar uma simulação *cotranscricional* do dobramento da molécula. Dessa forma, analisamos as conformações temporárias que as moléculas podem assumir antes de atingir sua estrutura final e os possíveis efeitos dessas estruturas metaestáveis. Discutimos as vantagens de uma análise baseada não somente na estrutura de mínima energia livre de uma molécula de RNA, mas também no conjunto de estruturas subótimas permitidas e no caminho pelo qual essas estruturas percorrem até atingir sua conformação de interesse biológico, além dos possíveis efeitos dessas estruturas metaestáveis.

Além disso, devido à natureza de nossa implementação, as análises não se baseiam apenas na estrutura de mínima energia livre da molécula de RNA, mas sim no conjunto de estruturas subótimas permitidas quando respeita-se a natureza estocástica do processo de dobramento desse polímero. Esse tipo de análise baseada em probabilidade reflete o resultado de um experimento de *Espalhamento de raios-x a baixo ângulo*, *SAXS*, que retorna os envoltórios das estruturas presentes em solução

ponderados pela sua concentração relativa. O programa possui, portanto, potencial para ser utilizado nas análises desse tipo de experimento, principalmente caso as moléculas de RNA apresentarem mais de uma conformação em seu estado de equilíbrio.

Também apresentamos os resultados da aplicação do programa desenvolvido nas sequências presentes no Banco de dados *RNAStrand* e comparamos os resultados obtidos por outros dois programas. Apesar de nosso modelo não ser voltado para a identificação das estruturas de mínima energia livre, quando utilizado para esse fim conquistou resultados comparáveis aos obtidos pelo *Kinefold*. Enfatizamos o potencial do programa para análise do comportamento de sequências com estruturas competitivas e como o mesmo pode ser utilizado no estudo da influência da estrutura do RNA nascente em sequências com pausas experimentalmente verificadas. Essa análise confirmou a sinergia entre a energia livre da bolha de transcrição e das estruturas presentes no RNA durante o alongamento para correta previsão da cinética transcricional, reduzindo o número de falsos positivos oriundos de pausas do tipo 2, reforçando verdadeiros positivos de mesma natureza e indicando pausas teóricas do tipo 1 que podem ser intensificadas devido à interação das estrturas nascentes com o CAT. Entretanto, ainda são observadas inconsistências entre o modelo e os resultados experimentais: técnicas mais criteriosas, como análises baseadas em dinâmica molecular, poderão colaborar na elucidação dos processos de interferência do RNA nascente na estabilidade do CAT para essas conformações.

Por se tratar da primeira versão desenvolvida desse programa, várias questões ainda podem ser analisadas. Por exemplo, quanto aos pseudonós: implementamos duas abordagens distintas, mas novos modelos podem ser estudados e desenvolvidos. Durante os testes realizados no Banco de Dados *RNAStrand*, notamos que facilmente obtínhamos mínimos locais para as estruturas devido à alta estabilidade dos SBRs: a implementação de uma etapa de arrefecimento simulado pode aumentar a cobertura do espaço de estados possíveis a um baixo custo computacional. Também podemos incluir novos parâmetros para o programa, como definir a energia mínima para aceitar um SBR, o número máximo de pseudonós que a sequência poderá admitir ou a presença de SBRs fixos durante a simulação. Finalmente, as saídas do programa podem ser aprimoradas: melhores opções de visualização, probabilidade

do pareamento entre bases ou mesmo o desenvolvimento de um algoritmo para predição de envoltórios para resultados de análises de SAXS.

As perspectivas para programas de análise estrutural de RNAs são prósperas. Interações entre a molécula de ácido ribonucleico e outras macromoléculas são sucessivamente identificadas com o avanço das técnicas experimentais. Essas moléculas podem alterar a conformação espacial do RNA diretamente ou alterando as características do processo de transcrição. RNAs *long non-coding* possuem papéis cada vez mais estudados, entretanto ainda apresentam um desafio aos métodos de predição atuais. Finalmente, estudos futuros podem analisar o efeito de alterações nos genomas, como no caso dos chamados *polimorfismos de nucleotídeo único*, na estrutura dessa polivalente molécula, ubíqua nos processos celulares.

Referências

AL-HASHIMI, H. M.; WALTER, N. G. RNA dynamics: it is about time. **Current opinion in structural biology**, Elsevier, v. 18, n. 3, p. 321–329, 2008. Citado na página 39.

ANDRONESCU, M. et al. RNA STRAND: the RNA secondary structure and statistical analysis database. **BMC bioinformatics**, BioMed Central Ltd, v. 9, n. 1, p. 340, 2008. Citado na página 66.

ARTSIMOVITCH, I.; LANDICK, R. The transcriptional regulator RfaH stimulates RNA chain synthesis after recruitment to elongation complexes by the exposed nontemplate DNA strand. **Cell**, v. 109, n. 2, p. 193–203, 2002. Citado na página 30.

AVERY, O. T.; MACLEOD, C. M.; MCCARTY, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. **The Journal of experimental medicine**, Rockefeller Univ Press, v. 79, n. 2, p. 137–158, 1944. Citado na página 24.

BAI, L.; FULBRIGHT, R. M.; WANG, M. D. Mechanochemical kinetics of transcription elongation. **Physical Review Letters**, v. 98, n. 6, 2007. Citado 4 vezes nas páginas 127, 129, 130 e 131.

BLOSSEY, R. **Computational biology: A statistical mechanics perspective**. [S.l.]: CRC Press, 2006. Citado na página 35.

BORER, P. N. et al. Proton NMR and structural features of a 24-nucleotide RNA hairpin. **Biochemistry**, ACS Publications, v. 34, n. 19, p. 6488–6503, 1995. Citado na página 51.

BRIERLEY, I.; GILBERT, R. C.; PENNELL, S. RNA pseudoknots and the regulation of protein synthesis. **Biochemical Society Transactions**, London: The Society, 1973-, v. 36, n. 4, p. 684–689, 2008. Citado na página 35.

BRIERLEY, I.; PENNELL, S.; GILBERT, R. J. Viral RNA pseudoknots: versatile motifs in gene expression and replication. **Nature Reviews Microbiology**, Nature Publishing Group, v. 5, n. 8, p. 598–610, 2007. Citado na página 35.

BRION, P.; WESTHOF, E. Hierarchy and dynamics of RNA folding. **Annual review of biophysics and biomolecular structure**, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 26, n. 1, p. 113–137, 1997. Citado na página 35.

CAO, S.; CHEN, S.-J. Predicting RNA pseudoknot folding thermodynamics. **Nucleic acids research**, Oxford Univ Press, v. 34, n. 9, p. 2634–2652, 2006. Citado 9 vezes nas páginas 54, 64, 78, 79, 80, 123, 124, 125 e 126.

CAO, S.; CHEN, S.-J. Predicting structures and stabilities for h-type pseudoknots with interhelix loops. **RNA**, Cold Spring Harbor Lab, v. 15, n. 4, p. 696–706, 2009. Citado 2 vezes nas páginas 123 e 125.

CHADALAVADA, D. M. et al. A role for upstream rna structure in facilitating the catalytic fold of the genomic hepatitis delta virus ribozyme. **Journal of molecular biology**, Elsevier, v. 301, n. 2, p. 349–367, 2000. Citado na página 40.

CHEN, J.-L.; GREIDER, C. W. Functional analysis of the pseudoknot structure in human telomerase RNA. **Proceedings of the National Academy of Sciences of the United States of America**, National Acad Sciences, v. 102, n. 23, p. 8080–8085, 2005. Citado na página 35.

COSTA, P. R.; ACENCIO, M. L.; LEMKE, N. Cooperative RNA polymerase molecules behavior on a stochastic sequence-dependent model for transcription elongation. **PloS one**, Public Library of Science, v. 8, n. 2, p. e57328, 2013. Citado 2 vezes nas páginas 127 e 131.

CRICK, F. et al. Central dogma of molecular biology. **Nature**, v. 227, n. 5258, p. 561–563, 1970. Citado 2 vezes nas páginas 24 e 26.

DAHM, R. Friedrich Miescher and the discovery of DNA. **Developmental biology**, Elsevier, v. 278, n. 2, p. 274–288, 2005. Citado na página 24.

DARWIN, C. **The variation of animals and plants under domestication**. [S.l.]: O. Judd, 1868. v. 2. Citado na página 23.

DETHOFF, E. A. et al. Functional complexity and regulation through RNA dynamics. **Nature**, Nature Publishing Group, v. 482, n. 7385, p. 322–330, 2012. Citado na página 37.

DING, Y.; CHAN, C. Y.; LAWRENCE, C. E. Sfold web server for statistical folding and rational design of nucleic acids. **Nucleic acids research**, Oxford Univ Press, v. 32, n. suppl 2, p. W135–W141, 2004. Citado na página 38.

GEIS, M. et al. Folding kinetics of large rnas. **Journal of molecular biology**, Elsevier, v. 379, n. 1, p. 160–173, 2008. Citado na página 40.

GESZVAIN, K.; LANDICK, R. The structure of bacterial RNA polymerase. **The bacterial chromosome. ASM, Washington, DC**, p. 283–296, 2005. Citado na página 60.

GILLESPIE, D. T. Exact stochastic simulation of coupled chemical reactions. **Journal Physical Chemistry**, v. 81, n. 25, p. 2340–2361, 1977. Citado na página 58.

GORODKIN, J.; STRICKLIN, S. L.; STORMO, G. D. Discovering common stem–loop motifs in unaligned RNA sequences. **Nucleic Acids Research**, v. 29, n. 10, p. 2135–2144, 2001. Citado na página 63.

GREIVE, S. J.; HIPPEL, P. H. von. Thinking quantitatively about transcriptional regulation. **Nature Reviews Molecular Cell Biology**, v. 6, n. 3, p. 221–232, 2005. Citado na página 30.

GRIFFITHS, A. J. F. et al. **Introduçao à Genética**. 8ª. ed. São Paulo: Guanabara Koogan, 2006. Citado 3 vezes nas páginas 23, 26 e 27.

GRUBER, A. R. et al. The vienna RNA websuite. **Nucleic acids research**, Oxford Univ Press, v. 36, n. suppl 2, p. W70–W74, 2008. Citado 2 vezes nas páginas 38 e 67.

GULTYAEV, A. P.; BATENBURG, F. van; PLEIJ, C. An approximation of loop free energy values of rna h-pseudoknots. **RNA**, Cold Spring Harbor Lab, v. 5, n. 5, p. 609–617, 1999. Citado na página 125.

HERBERT, K. M.; GREENLEAF, W. J.; BLOCK, S. M. Single-molecule studies of RNA polymerase: Motoring along. **Annual Review of Biochemistry**, v. 77, p. 149–176, 2008. Citado na página 28.

HERBERT, K. M. et al. Sequence-resolved detecton of pausing by single RNA polymerase molecules. **Cell**, v. 125, n. 6, p. 1083–1094, 2006. Citado na página 32.

HERSHEY, A. D.; CHASE, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. **The Journal of general physiology**, Rockefeller Univ Press, v. 36, n. 1, p. 39–56, 1952. Citado na página 24.

ISAMBERT, H.; SIGGIA, E. D. Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 97, n. 12, p. 6515–6520, 2000. Citado 10 vezes nas páginas 47, 48, 53, 56, 57, 64, 97, 121, 122 e 123.

JACOB, F.; MONOD, J. Genetic regulatory mechanisms in the synthesis of proteins. **Journal of molecular biology**, Elsevier, v. 3, n. 3, p. 318–356, 1961. Citado na página 27.

KANG, M.; PETERSON, R.; FEIGON, J. Structural Insights into riboswitch control of the biosynthesis of queuosine, a modified nucleotide found in the anticodon of tRNA. **Molecular cell**, Elsevier, v. 33, n. 6, p. 784–790, 2009. Citado na página 35.

KARP, G. **Biologia celular e molecular: conceitos e experimentos**. 1ª. ed. São Paulo: Editora Manole, 2005. Citado na página 24.

KLEIN, D. J.; EDWARDS, T. E.; FERRÉ-D'AMARÉ, A. R. Cocrystal structure of a class I preQ1 riboswitch reveals a pseudoknot recognizing an essential hypermodified nucleobase. **Nature structural & molecular biology**, Nature Publishing Group, v. 16, n. 3, p. 343–344, 2009. Citado na página 35.

KRATKY, O.; POROD, G. Röntgenuntersuchung gelöster fadenmoleküle. **Recueil des Travaux Chimiques des Pays-Bas**, Wiley Online Library, v. 68, n. 12, p. 1106–1122, 1949. Citado na página 56.

LEVIN, J. R.; CHAMBERLIN, M. J. Mapping and characterization of transcriptional pause sites in the early genetic region of bacteriophage T7. **Journal of Molecular Biology**, v. 196, n. 1, p. 61–84, 1987. Citado 3 vezes nas páginas 70, 71 e 133.

LIU, L.; CHEN, S.-J. Computing the conformational entropy for RNA folds. **The Journal of chemical physics**, AIP Publishing, v. 132, n. 23, p. 235104, 2010. Citado 6 vezes nas páginas 53, 54, 55, 97, 123 e 124.

MCCASKILL, J. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. **Biopolymers**, v. 29, p. 1105–1119, 1990. Citado na página 38.

MEERTEN, D. van; GIRARD, G.; DUIN, J. V. Translational control by delayed rna folding: identification of the kinetic trap. **Rna**, Cold Spring Harbor Lab, v. 7, n. 3, p. 483–494, 2001. Citado na página 40.

MENTEN, L.; MICHAELIS, M. I. Die Kinetik der Invertinwirkung. **Biochem. Z**, v. 49, p. 333–369, 1913. Citado na página 129.

NAGEL, J. H.; PLEIJ, C. W. Self-induced structural switches in RNA. **Biochimie**, Elsevier, v. 84, n. 9, p. 913–923, 2002. Citado na página 40.

NIKOLOVA, E. N. et al. Transient hoogsteen base pairs in canonical duplex dna. **Nature**, Nature Publishing Group, v. 470, n. 7335, p. 498–502, 2011. Citado na página 34.

NUSSINOV, R.; JACOBSON, A. B. Fast algorithm for predicting the secondary structure of single-stranded RNA. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 77, n. 11, p. 6309–6313, 1980. Citado na página 46.

PEARSON, H. What is a gene? **Nature**, v. 441, n. 7092, p. 398–401, 2006. Citado na página 26.

PLEIJ, C. W.; RIETVELD, K.; BOSCH, L. A new principle of rna folding based on pseudoknotting. **Nucleic Acids Research**, Oxford Univ Press, v. 13, n. 5, p. 1717–1731, 1985. Citado na página 125.

PUTON, T. et al. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. **Nucleic acids research**, Oxford Univ Press, v. 41, n. 7, p. 4307–4323, 2013. Citado na página 62.

QIAO, F.; CECH, T. R. Triple-helix structure in telomerase RNA contributes to catalysis. **Nature structural & molecular biology**, Nature Publishing Group, v. 15, n. 6, p. 634–640, 2008. Citado na página 35.

RING, B. Z.; YARNELL, W. S. Function of E-coli RNA polymerase sigma factor sigma(70) in promoter-proximal pausing. **Cell**, v. 86, n. 3, p. 485–493, 1996. Citado na página 30.

SANTALUCIA, J.; ALLAWI, H. T.; SENEVIRATNE, A. Improved nearest-neighbor parameters for predicting DNA duplex stability. **Biochemistry**, v. 35, n. 11, p. 3555–3562, 1996. Citado 2 vezes nas páginas 52 e 109.

SANTALUCIA, J.; HICKS, D. The thermodynamics of DNA structural motifs. **Annual Review Biophysics and Biomoleular Structure**, v. 33, p. 415–440, 2004. Citado na página 111.

SATO, K. et al. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. **Bioinformatics**, Oxford Univ Press, v. 27, n. 13, p. i85–i93, 2011. Citado na página 38.

SCHLICK, T. RNA: The cousin left behind becomes a star. In: **Computational Studies of RNA and DNA**. [S.l.]: Springer, 2006. p. 259–281. Citado na página 27.

SUGIMOTO, N. et al. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. **Biochemistry**, v. 34, n. 35, p. 11211–11216, 1995. Citado na página 112.

TADIGOTLA, V. R. et al. Thermodynamic and kinetic modeling of transcriptional pausing. **Proceedings of the National Academy of Sciences of U.S.A.**, v. 103, n. 12, p. 4439–4444, 2006. Citado 3 vezes nas páginas 70, 71 e 133.

TAKAHIRO, R. N. **Transcrição cooperativa de genes ribossomais em E. coli usando um modelo estocástico e dependente de sequência**. Dissertação (Mestrado em Ciências Biológicas - Genética) — Universidade Estadual Paulista Júlio de Mesquita Filho, Botucatu - São Paulo, 2015. Citado na página 130.

TINOCO, I.; BUSTAMANTE, C. How RNA folds. **Journal of molecular biology**, Elsevier, v. 293, n. 2, p. 271–281, 1999. Citado 3 vezes nas páginas 35, 37 e 88.

TURNER, D. H.; MATHEWS, D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. **Nucleic acids research**, Oxford Univ Press, p. gkp892, 2009. Citado 15 vezes nas páginas 53, 55, 63, 73, 97, 109, 110, 111, 112, 113, 114, 115, 116, 117 e 119.

UNDERHILL, P. T.; DOYLE, P. S. Development of bead-spring polymer models using the constant extension ensemble. **Journal of Rheology (1978-present)**, The Society of Rheology, v. 49, n. 5, p. 963–987, 2005. Citado na página 56.

WATSON, J. D.; CRICK, F. H. et al. Molecular structure of nucleic acids. **Nature**, v. 171, n. 4356, p. 737–738, 1953. Citado na página 24.

XAYAPHOUMMINE, A. **Simulations et expèriences sur le repliement de l'ARN : prédictions statistiques des pseudonœuds in silico et réalisation de commutateurs ARN par transcription in vitro**. Tese (Doctorat) — l'Université Louis Pateur Strasbourg I, 2004. Citado na página 77.

XAYAPHOUMMINE, A.; BUCHER, T.; ISAMBERT, H. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. **Nucleic acids research**, Oxford Univ Press, v. 33, n. suppl 2, p. W605–W610, 2005. Citado 5 vezes nas páginas 53, 65, 67, 76 e 77.

XAYAPHOUMMINE, A. et al. Encoding folding paths of RNA switches. **NUCLEIC ACIDS RESEARCH**, v. 35, n. 2, p. 614–622, JAN 2007. ISSN 0305-1048. Citado 7 vezes nas páginas 39, 53, 67, 69, 70, 91 e 92.

XU, X.; CHEN, S.-J. Physics-based rna structure prediction. **Biophysics Reports**, Springer, v. 1, n. 1, p. 2–13, 2015. Citado 2 vezes nas páginas 38 e 123.

YAGER, T. D.; VONHIPPEL, P. H. A thermodynamic analysis of RNA transcript elongation and termination in *Escherichia coli*. **Biochemistry**, v. 30, n. 4, p. 1097–1118, 1991. Citado na página 129.

ZAKOV, S. et al. Rich parameterization improves RNA structure prediction. **Journal of Computational Biology**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 18, n. 11, p. 1525–1542, 2011. Citado na página 38.

ZHANG, J.; LANDICK, R. A two-way street: Regulatory interplay between rna polymerase and nascent rna structure. **Trends in biochemical sciences**, Elsevier, v. 41, n. 4, p. 293–310, 2016. Citado 7 vezes nas páginas 30, 32, 33, 39, 40, 60 e 91.

ZUKER, M.; SANKOFF, D. RNA secondary structures and their prediction. **Bulletin of mathematical biology**, Springer, v. 46, n. 4, p. 591–621, 1984. Citado na página 46.

ZUKER, M.; STIEGLER, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. **Nucleic acids research**, Oxford Univ Press, v. 9, n. 1, p. 133–148, 1981. Citado na página 38.

Apêndices

APÊNDICE A – Estruturas secundárias

Esse Apêndice apresenta os conjuntos de regras implementadas em nosso programa para o cálculo da variação da energia livre de Gibbs para estruturas secundárias, segundo a compilação de 2004 do grupo de Turner, apresentada na Versão 1.02 (2011) do NNDB (TURNER; MATHEWS, 2009). Alguns parâmetros estão detalhados, mas optamos por omitir tabelas de parâmetros muito extensas.

A.1 Segmentos bifilamentares

Os experimentos realizados por SantaLucia, Allawi e Seneviratne (1996) mostraram que as interações entre as fitas podem ser modeladas de maneira satisfatória se considerarmos os efeitos da *vizinhança* dos pares de bases, ou seja, essas interações dependem não apenas das bases que formam as ligações de hidrogênio em si, mas também das bases vizinhas a essas ligações. Esse princípio deu origem ao *Modelo dos primeiros vizinhos* (*Nearest-neighbor model*) para determinação da energia livre de um segmento bifilamentar qualquer de comprimento n: somam-se os parâmetros determinados para os $(n-1)$ vizinhos presentes, além de outros parâmetros que podem alterar a estabilidade das fitas (SANTALUCIA; ALLAWI; SENEVIRATNE, 1996).

O *Modelo dos primeiros vizinhos* foi implementado e utilizado em duas situações: (I) para determinação da energia livre dos SBRs presentes nas estruturas secundárias da molécula de RNA e (II) para determinação da energia livre resultante das interações DNA-DNA e DNA-RNA presentes na bolha de transcrição e que influenciam na cinética da RNAP durante o alongamento. Para ilustrar como esses valores são determinados, a energia livre de Gibbs liberada na formação do segmento

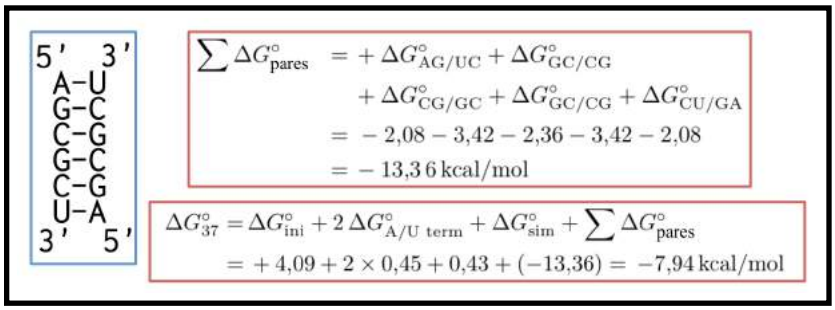
bifilamentar de DNA $5'$–CGTTGA–$3'$/$3'$–GCAACT–$5'$ a $37\,^{\circ}$C será dada por

$$\begin{aligned}\Delta G^{\circ}(\text{total}) &= \Delta G^{\circ}_{\text{ini}} + \Delta G^{\circ}_{\text{sim}} + \sum \Delta G^{\circ}_{\text{duplas}} + \Delta G^{\circ}_{\text{A-T term}} \\ &= \Delta G^{\circ}_{\text{ini}} + \Delta G^{\circ}_{\text{sim}} + \Delta G^{\circ}_{\text{CG}} + \Delta G^{\circ}_{\text{GT}} + \Delta G^{\circ}_{\text{TT}} + \Delta G^{\circ}_{\text{TG}} + \Delta G^{\circ}_{\text{GA}} + \Delta G^{\circ}_{\text{A-T}_{\text{term}}} \\ &= +1{,}96 + 0 - 2{,}17 - 1{,}44 - 1{,}00 - 1{,}45 - 1{,}30 + 0{,}05 \\ &= -5{,}35\,\text{kcal/mol}.\end{aligned}$$

Note que não há penalidade para simetria ($\Delta G^{\circ}_{\text{sim}}$) nessa sequência, pois $5' \rightarrow 3' \neq 3' \rightarrow 5'$. Os parâmetros utilizados em nossa implementação estão apresentados na Tabela A.1, na Tabela A.2 e na Tabela A.3 e foram determinados experimentalmente por diferentes laboratórios. Nas Tabelas, "XY/ZW" representa os vizinhos $5'$–XY–$3'$ ligados à $3'$–ZW–$5'$, com ΔH° e ΔG°_{37} dados em kcal/mol e ΔS° dado em cal·K^{-1}·mol^{-1}. Note que para o caso do RNA, o termo $\Delta G^{\circ}_{\text{ini}}$ estará presente somente em casos de interação entre duas moléculas distintas.

Um exemplo do cálculo da energia livre de Gibbs para um segmento bifilamentar de RNA, utilizando os parâmetros da Tabela A.2, está ilustrado na Figura A.1.

Figura A.1 – Esquema de cálculo para ΔG°_{37} de um segmento bifilamentar de RNA



Legenda: Representação da estrutura em estudo e as operações realizadas para determinação da respectiva variação de energia livre de Gibbs para sua formação. Como se trata da interação entre duas fitas distintas, o parâmetro de iniciação está presente. Em ambas extremidades da hélice, observamos um par A-U, portanto a penalidade será acrescentada duas vezes. Além disso, como a sequência $5' \rightarrow 3'$ é igual à $3' \rightarrow 5'$, adiciona-se a penalidade devido à simetria. Fonte: Modificado de Turner e Mathews (2009)

Tabela A.1 – Parâmetros para cálculo da estabilidade de um segmento bifilamentar de DNA

Sequência	ΔH°	ΔS°	ΔG°_{37}	Sequência	ΔH°	ΔS°	ΔG°_{37}
AA/TT	−7,6	−21,3	−1,0	AT/TA	−7,2	−20,4	−0,9
TA/AT	−7,2	−21,3	−0,6	CA/GT	−8,5	−22,7	−1,5
GT/CA	−8,4	−22,4	−1,4	CT/GA	−7,8	−21,0	−1,3
GA/CT	−8,2	−22,2	−1,3	CG/GC	−10,6	−27,2	−2,2
GC/CG	−9,8	−24,4	−2,2	GG/CC	−8,0	−19,9	−1,84
$\Delta G^\circ_{\text{ini}}$	+0,2	−5,7	+1,96	$\Delta G^\circ_{\text{A-T term}}$	+2,2	+6,9	+0,05
$\Delta G^\circ_{\text{sim}}$	0,0	−1,4	+0,4				

Fonte – SantaLucia e Hicks (2004)

Tabela A.2 – Parâmetros para cálculo da estabilidade de um segmento bifilamentar de RNA

Sequência	ΔG°_{37}	Sequência	ΔG°_{37}	Sequência	ΔG°_{37}
AA/UU	−0,93	UU/AA	−0,93	AU/UA	−1,10
UA/AU	−1,33	CU/GA	−2,08	AG/UC	−2,08
CA/GU	−2,11	UG/AC	−2,11	GU/CA	−2,24
AC/UG	−2,24	GA/CU	−2,35	UC/AG	−2,35
CG/GC	−2,36	GG/CC	−3,26	CC/GG	−3,26
GC/CG	−3,42	AG/UU	−0,55	UU/GA	−0,55
AU/UG	−1,36	GU/UA	−1,36	CG/GU	−1,41
UG/GC	−1,41	CU/GG	−2,11	GG/UC	−2,11
GG/CU	−1,53	UC/GG	−1,53	GU/CG	−2,51
GC/UG	−2,51	GA/UU	−1,27	UU/AG	−1,27
GG/UU	−0,50	UU/GG	−0,50	GU/UG	+1,29
UG/AU	−1,00	UA/GU	−1,00	UG/GU	+0,30
$\Delta G^\circ_{\text{ini}}$	+4,09	$\Delta G^\circ_{\text{A/U,G/U term}}$	+0,45	$\Delta G^\circ_{\text{sim}}$	+0,43

Fonte – Turner e Mathews (2009)

Tabela A.3 – Parâmetros para cálculo da estabilidade de uma fita híbrida de RNA-DNA

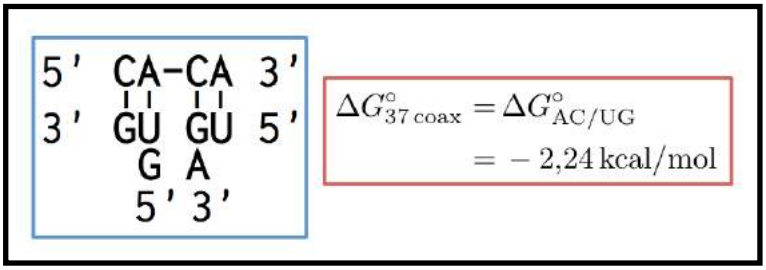
Sequência	ΔH°	ΔS°	ΔG°_{37}	Sequência	ΔH°	ΔS°	ΔG°_{37}
rAA/dTT	−7,8	−21,9	−1,0	rAC/dTG	−5,9	−12,3	−2,1
rAG/dTC	−9,1	−23,5	−1,8	rAU/dTA	−8,3	−23,9	−0,9
rCA/dGT	−9,0	−26,1	−0,9	rCC/dGG	−9,3	−23,2	−2,1
rCG/dGC	−16,3	−47,1	−1,7	rCU/dGA	−7,0	−19,7	−0,9
rGA/dCT	−5,5	−13,5	−1,3	rGC/dCG	−8,0	−17,1	−2,7
rGG/dCC	−12,8	−31,9	−2,9	rGU/dCA	−7,8	−21,6	−1,1
rUA/dAT	−7,8	−23,2	−0,6	rUC/dAG	−8,6	−22,9	−1,5
rUG/dAC	−10,4	−28,4	−1,6	rUU/dAA	−11,5	−36,4	−0,2
$\Delta G^\circ_{\text{ini}}$	+1,9	−3,9	+3,1				

Fonte – Sugimoto et al. (1995)

A.2 Empilhamento coaxial sequencial

Um caso especial ocorre quando dois segmentos bifilamentares independentes, mas subsequentes, se alinham no mesmo eixo. A energia livre de Gibbs dessa conformação é aproximada como se tratasse de um mesmo segmento, conforme apresentado na Figura A.2.

Figura A.2 – Esquema de cálculo para ΔG°_{37} de um empilhamento coaxial

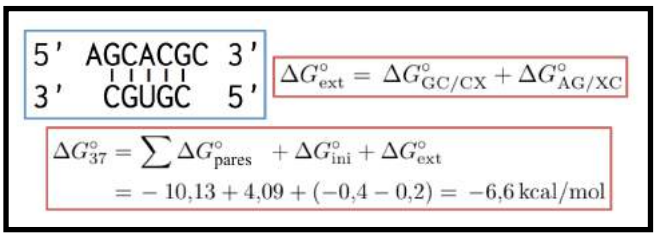


Legenda: Representação da estrutura em estudo e a operação realizada para determinação da respectiva variação de energia livre de Gibbs para sua formação. Apesar de tratada como um único segmento bifilamentar, a penalidade devido ao pareamento A/U no final do primeiro segmento ainda deverá ser aplicado no cálculo da estabilidade geral da molécula. Fonte: Modificado de Turner e Mathews (2009)

A.3 Extremidades pendentes

São nucleotídeos não pareados que se estabilizam no final de um segmento bifilamentar, em qualquer uma das duas extremidades, 3′ ou 5′. No RNA, extremidades 3′ costumam ser mais estáveis que as 5′. A Figura A.3 ilustra o cálculo da energia para esse tipo de estrutura, utilizando os parâmetros experimentalmente determinados e disponíveis no NNDB.

Figura A.3 – Esquema de cálculo para ΔG°_{37} de uma extremidade pendente

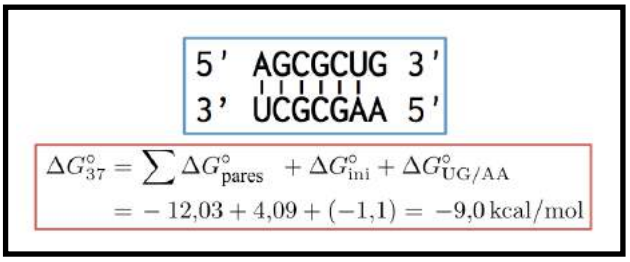


Legenda: Representação da estrutura em estudo e as operações realizadas para determinação da respectiva variação de energia livre de Gibbs para sua formação. “X” aqui representa a ausência do nucleotídeo. Fonte: Modificado de Turner e Mathews (2009)

A.4 Incompatibilidade terminal

Ocorre quando duas extremidades pendentes estão localizadas no final ou no início de um segmento bifilamentar, formando um par não canônico, conforme representado na Figura A.4. Seus valores estão experimentalmente determinados e disponíveis no NNDB.

Figura A.4 – Esquema de cálculo para ΔG°_{37} de uma incompatibilidade terminal

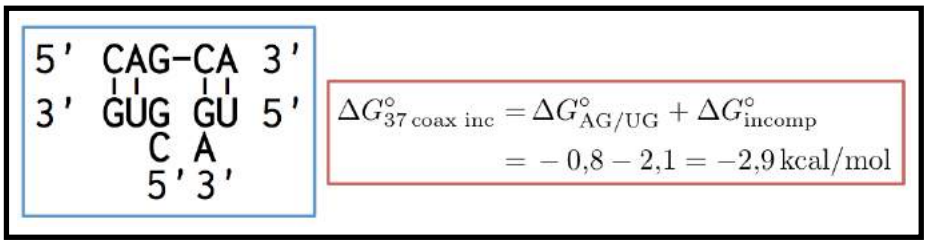


Legenda: Representação da estrutura em estudo e as operações realizadas para determinação da respectiva variação de energia livre de Gibbs para sua formação. Fonte: Modificado de Turner e Mathews (2009)

A.5 Empilhamento coaxial com incompatibilidade

Como no caso do empilhamento coaxial sequencial, dois segmentos bifilamentares se alinham, mas dessa vez um único pareamento incompatível, i.e., não canônico, está presente entre eles. Nessa situação, existem dois empilhamentos adjacentes: no primeiro, a incompatibilidade localiza-se adjacente ao segmento, e é tratada como uma incompatibilidade terminal; no segundo, adiciona-se um termo independente da sequência, $-2{,}1$ kcal/mol, conforme apresentado na Figura A.5, e se a "incompatibilidade" pudesse formar um pareamento de Watson-Crick ou um G/U, adiciona-se ainda $-0{,}4$ ou $-0{,}2$ kcal/mol, respectivamente, à variação da energia livre da estrutura. Os parâmetros para esse caso foram omitidos, mas estão disponíveis no NNDB.

Figura A.5 – Esquema de cálculo para ΔG°_{37} de um empilhamento coaxial com incompatibilidade



Legenda: Representação da estrutura em estudo e a operação realizada para determinação da respectiva variação de energia livre de Gibbs para sua formação.
Fonte: Modificado de Turner e Mathews (2009)

A.6 Grampo

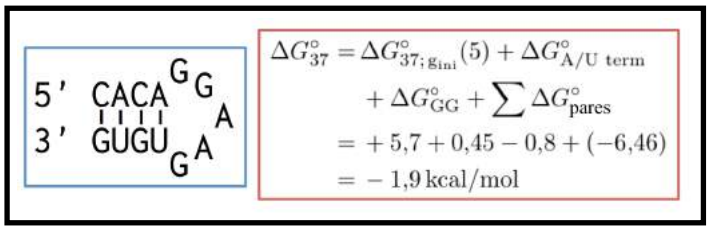
Algumas alças unifilamentares de comprimento 3, 4 e 6 nt apresentam pouca aderência ao modelo, mas possuem valores tabelados experimentalmente determinados, disponíveis no NNDB. Um exemplo do cálculo da variação da energia livre para alças genéricos está ilustrado na Figura A.6, com valores estimados a partir das seguintes equações:

$$\Delta G^\circ_{37;\text{g}}(n) = \begin{cases} \Delta G^\circ_{37;\text{g}_\text{ini}}(n) + \Delta G^\circ_{37;\text{g}_\text{C}}(n), & n = 3, \\ \Delta G^\circ_{37;\text{g}_\text{ini}}(n) + \Delta G^\circ_{37}(\text{term}) + \sum \Delta G^\circ_{37}(\text{bônus}) + \Delta G^\circ_{37;\text{g}_\text{C}}(n), & n > 3, \end{cases}$$

onde $\Delta G^\circ_{37;\mathrm{g_{ini}}}(n)$ representa a variação da energia livre mínima para formação dessas estruturas, com valores tabelados e disponíveis no NNDB, e ΔG°_{37}(term) é a energia da primeira incompatibilidade terminal. Caso a incompatibilidade seja U/U, G/A ou G/G, ou ainda se o último par do segmento bifilamentar presente antes da alça for G/U, adicionam-se as respectivas bonificações: $-0{,}9$, $-0{,}9$, $-0{,}8$ ou -2.2 kcal/mol. Alças compostas apenas por C recebem uma penalização extra, dada em função de seu comprimento por

$$\Delta G^\circ_{37;\mathrm{g_C}}(n) = \begin{cases} 1{,}5, & n = 3, \\ 0{,}3\,n + 1{,}6, & n > 3. \end{cases}$$

Figura A.6 – Esquema de cálculo para ΔG°_{37} de um grampo



Legenda: Representação da estrutura em estudo e as operações realizadas para determinação da respectiva variação de energia livre de Gibbs para sua formação. Temos como incompatibilidade terminal o par G/G, e a respectiva bonificação foi adicionada. Fonte: Modificado de Turner e Mathews (2009)

A.7 Protuberâncias

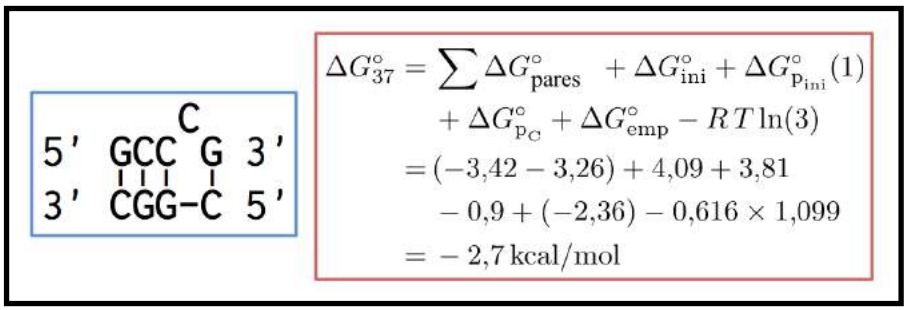
No caso das protuberâncias, o valor estimado da variação de energia livre em função do número de nucleotídeos n é dado por:

$$\Delta G^\circ_{37;\mathrm{prot}}(n) = \begin{cases} \Delta G^\circ_{37;\mathrm{p_{ini}}}(n) + \Delta G^\circ_{37;\mathrm{p_C}} + \Delta G^\circ_{37;\mathrm{emp}} - R\,T\,\ln(n_{\mathrm{est}}), & n = 1, \\ \Delta G^\circ_{37;\mathrm{p_{ini}}}(n), & 1 < n \leq 6, \\ \Delta G^\circ_{37;\mathrm{p_{ini}}}(6) + 1.75\,R\,T\,\ln(n/6), & n > 6, \end{cases}$$

onde n é o número de nucleotídeos não pareados e $\Delta G^\circ_{37;\mathrm{p_{ini}}}$ é a variação de energia inicial para essa conformação. Essa variação inicial está experimentalmente determinada e disponível no NNDB para $1 \leq n \leq 3$ e é linearmente extrapolada para $4 < n \leq 6$. Para $n = 1$, p_C representa uma protuberância especial onde um C pareado é

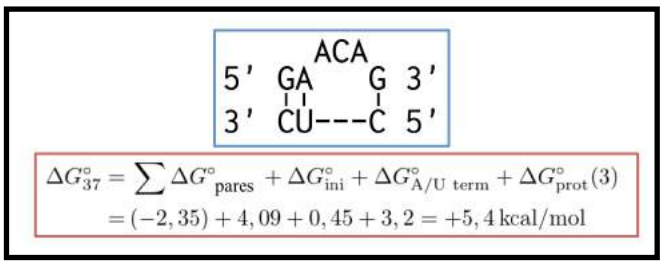
seguido de um C não pareado, $\Delta G^{\circ}_{37;\text{emp}}$ é a energia dos vizinhos que se formariam se não existisse a protuberância e n_{est} é o número de segmentos unifilamentares possíveis indiferenciáveis. Note que nessa situação, considera-se que os segmentos bifilamentares são ininterruptos, portanto não há penalização por A/U ou G/U nas extremidades desses segmentos. A Figura A.7 apresenta um exemplo de cálculo para $n = 1$ e a Figura A.8 para um caso com $n = 3$.

Figura A.7 – Esquema de cálculo para ΔG°_{37} de uma protuberância simples



Legenda: Representação da estrutura em estudo e as operações realizadas para determinação da respectiva variação de energia livre de Gibbs para sua formação. Trata-se de uma protuberância devido à presença de um único C não pareado, antecedido por um C pareado; a respectiva bonificação foi adicionada. Adicionamos também a contribuição caso a protuberância não existisse: no caso teríamos os vizinhos CG/GC. Note que a protuberância poderia ser resultado do não pareamento de qualquer uma das três citosinas; logo $n_{\text{est}} = 3$. Fonte: Modificado de Turner e Mathews (2009)

Figura A.8 – Esquema de cálculo para ΔG°_{37} de uma protuberância com 3 nt



Legenda: Representação da estrutura em estudo e as operações realizadas para determinação da respectiva variação de energia livre de Gibbs para sua formação. No caso, como $n = 3$, o parâmetro $\Delta G^{\circ}_{\text{p}_{\text{ini}}}$ encontra-se tabelado. Fonte: Modificado de Turner e Mathews (2009)

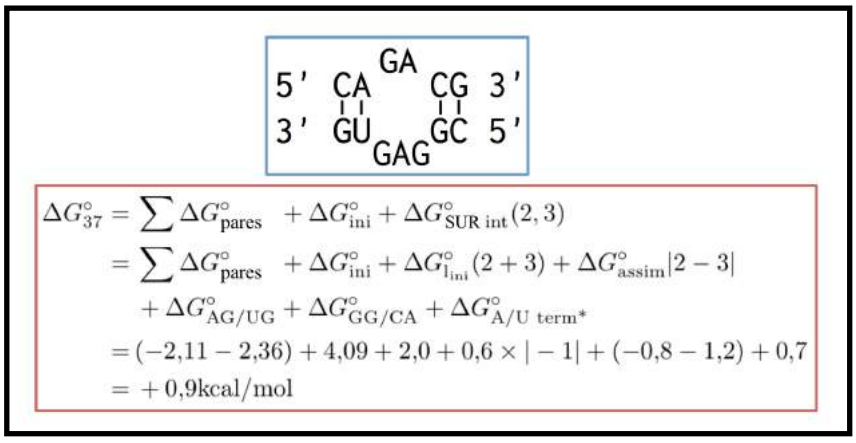
A.8 Laços internos

Pequenos laços internos, 1×1, 1×2, 2×2, possuem valores experimentalmente determinados, disponíveis no NNDB. A estabilidade de laços maiores é predita seguindo a equação

$$\Delta G^\circ_{37;\text{laço}}(n_1,n_2) = \Delta G^\circ_{37;\text{l}_\text{ini}}(n_1+n_2) + \Delta G^\circ_{37;\text{assim}}|n_1-n_2| + \sum \Delta G^\circ_{37;\text{inc}} + \Delta G^\circ_{\text{A/U,G/U term*}},$$

onde $\Delta G^\circ_{37;\text{l}_\text{ini}}(n_1+n_2)$ possui valores disponíveis para $n \le 6$, e é extrapolado para outros valores por $\Delta G^\circ_{37;\text{l}_\text{ini}}(n_1+n_2>6) = \Delta G^\circ_{37;\text{l}_\text{ini}}(6)+1.08\ln(n/6)$. Além disso, a constante $\Delta G^\circ_{37;\text{assim}} = 0{,}6$ kcal/mol surge em laços assimétricos e $\Delta G^\circ_{\text{A/U,G/U term*}} = 0{,}7$ kcal/mol substitui a penalidade $\Delta G^\circ_{\text{A/U,G/U term}}$ no cálculos da variação da energia das duplas hélices. As variações devido as incompatibilidades terminais, $\Delta G^\circ_{37;\text{inc}}$, variam em relação as utilizadas nas situações normais: quando os laços apresentam $1\times(n-1)$, $\Delta G^\circ_{37;\text{inc}} = 0$ kcal/mol e $\Delta G^\circ_{37;\text{incomp}}$ possui valores tabelados para laços 2×3 e para algumas outras incompatibilidades específicas; caso contrário, utilizam-se as regras apresentadas anteriormente. A Figura A.9 apresenta uma ilustração desse cálculo.

Figura A.9 – Esquema de cálculo para ΔG°_{37} de um laço interno



$$\begin{aligned}\Delta G^\circ_{37} &= \sum \Delta G^\circ_{\text{pares}} + \Delta G^\circ_{\text{ini}} + \Delta G^\circ_{\text{SUR int}}(2,3)\\ &= \sum \Delta G^\circ_{\text{pares}} + \Delta G^\circ_{\text{ini}} + \Delta G^\circ_{\text{l}_\text{ini}}(2+3) + \Delta G^\circ_{\text{assim}}|2-3|\\ &\quad + \Delta G^\circ_{\text{AG/UG}} + \Delta G^\circ_{\text{GG/CA}} + \Delta G^\circ_{\text{A/U term*}}\\ &= (-2{,}11-2{,}36)+4{,}09+2{,}0+0{,}6\times|-1|+(-0{,}8-1{,}2)+0{,}7\\ &= +0{,}9\text{kcal/mol}\end{aligned}$$

Legenda: Representação da estrutura em estudo e as operações realizadas para determinação da respectiva variação de energia livre de Gibbs para sua formação.
Fonte: Modificado de Turner e Mathews (2009)

A.9 Múltiplas ramificações

A variação da energia livre para regiões de fita simples resultantes da junção de mais de duas duplas hélices é dada por

$$\Delta G^{\circ}_{37;\mathrm{MR}} = \Delta G^{\circ}_{37;\mathrm{MR_{ini}}} + \mathrm{Min}\left\{\sum \Delta G^{\circ}_{37;\mathrm{emp}}\right\}.$$

A variação de energia para a iniciação segue

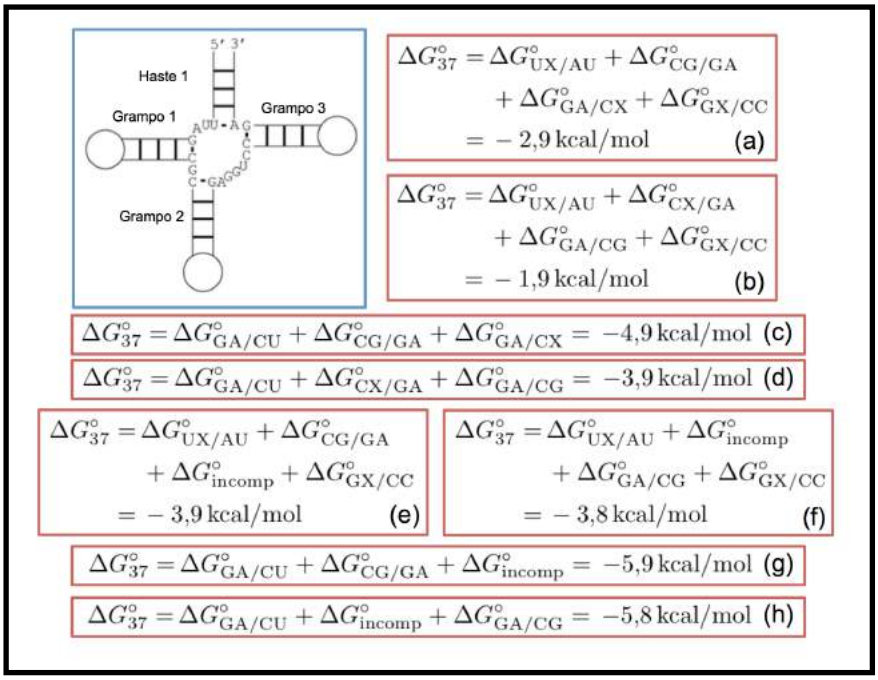
$$\Delta G^{\circ}_{37;\mathrm{MR_{ini}}} = 9{,}25 + 0{,}91 \times A - 0{,}63 \times B + \Delta G^{\circ}_{37;\mathrm{tensão}},$$

com $A = \mathrm{Min}\{2{,}0; n_{\mathrm{d}}\}$, onde n_{d} é o valor médio das diferenças entre o número de nucleotídeos não pareados ao lado de cada uma das duplas hélices presentes; B é o número de duplas hélices presentes e $\Delta G^{\circ}_{37;\mathrm{tensão}} = 3{,}14$ kcal/mol surge apenas em laços resultantes da junção de três duplas hélices com menos de dois nucleotídeos não pareados. O termo $\sum \Delta G^{\circ}_{37;\mathrm{emp}}$ é encontrado determinando-se todas as configurações possíveis para o laço e somando todas as combinações entre essas possíveis energias. A Figura A.10 ilustra o cálculo para o laço presente na clássica estrutura em "trevo de quatro folhas", presente no RNA transportador. Para esse caso, teremos

$$\Delta G^{\circ}_{37;\mathrm{MR_{ini}}} = 9{,}25 + 0{,}91 \times \mathrm{Min}\{2{,}0; (2+1+4+5)/4\} - 0{,}63 \times (4) = +8{,}6\,\mathrm{kcal/mol}$$

e, portanto, conforme a Figura A.10, $\Delta G^{\circ}_{37;\mathrm{MR}} = 8{,}6 - 5{,}9 = 3{,}6$ kcal/mol.

Figura A.10 – Esquema de cálculo para ΔG°_{37} de múltiplas ramificações



Legenda: Representação da estrutura em estudo e as operações realizadas para determinação da respectiva variação de energia livre de Gibbs para sua formação. São 4 SBRs e 8 configurações possíveis. A configuração (g) possui a menor variação de energia livre, –5,9 kcal/mol, logo será a configuração considerada nos cálculos. (a) Hélice 1 com U–3′ pendente, Hélice 2 com uma incompatibilidade terminal, Hélice 3 com A–3′ pendente e Hélice 4 com C–5′ pendente. (b) Hélice 1 com U–3′ pendente, Hélice 2 A–5′ pendente, Hélice 3 com incompatibilidade terminal, e Hélice 4 com C–5′ pendente. (c) Hélice 1 em empilhamento coaxial sequencial com a Hélice 4, Hélice 2 com incompatibilidade terminal, e Hélice 3 A–3′ pendente. (d) Hélice 1 em empilhamento coaxial sequencial com a Hélice 4, Hélice 2 A–5′ pendente, e Hélice 3 com incompatibilidade terminal. (e) Hélice 1 com U–3′ pendente, Hélice 2 em empilhamento coaxial com incompatibilidade G-A com Hélice 3, e Hélice 4 com C–5′ pendente. (f) Hélice 1 com U–3′ pendente, Hélice 2 em empilhamento coaxial com incompatibilidade A-G com Hélice 3, e Hélice 4 com C–5′ pendente. (g) Hélice 1 em empilhamento coaxial sequencial com Hélice 4 e Hélice 2 em empilhamento coaxial com incompatibilidade G-A com Hélice 3. (h) Hélice 1 em empilhamento coaxial sequencial com Hélice 4 e Hélice 2 em empilhamento coaxial com incompatibilidade A-G com Hélice 3. Fonte: Modificado de Turner e Mathews (2009)

APÊNDICE B – Pseudonós

Dois métodos foram implementados para o cálculo da variação da energia livre de Gibbs para pseudonós: *KineFold* e *Vfold*. O primeiro trabalha em duas escalas, analisando as redes individualmente e globalmente, enquanto o segundo baseia-se em observações estruturais, modelando a molécula de RNA através de ligações virtuais.

B.1 *KineFold*

Para a aproximação de Isambert e Siggia (2000), as duplas hélices são modeladas como "hastes", e regiões não-pareadas como cadeias gaussianas com comprimento de Kuhn $b = 1{,}5$ nm, ou 2,5 bases de comprimento $a = 6$ Å. A Figura 15 apresentou os oito tipos permitidos de redes, com no máximo duas duplas hélices internas. O valor da contribuição energética dessas estruturas varia de acordo com sua classificação. A equação geral para determinação da contribuição entrópica, S, é dada por

$$\mathrm{e}^{S/k} = \frac{\mathrm{e}^{-A_1 l_1^2 - A_2 l_2^2}}{D^{3/2}} \times \frac{\mathrm{e}^{2A_3 l_1 l_2} - \mathrm{e}^{-2A_3 l_1 l_2}}{4A_3 l_1 l_2}. \tag{B.1}$$

Um fator multiplicativo $\alpha = 0{,}0068$ é incluído para cada laço presente na estrutura. Os parâmetros presentes nesta equação estão apresentados na Tabela B.1. A classificação das redes segue a notação presente na Tabela 3. Nos cálculos, $\beta = 3/2 \cdot a \cdot b$ e o parâmetro l corresponde ao comprimento efetivo da dupla hélice presente na rede, dado por

$$l = \left[d^2 \mathrm{sen}^2 \left(\pi n_s / n_p \right) + h^2 \left(n_s / n_p \right)^2 \right]^{1/2}, \tag{B.2}$$

para duplas hélices curtas ($n_s \leq 6$), e

$$l = n_s/6 \left[d^2 \mathrm{sen}^2 \left(\pi 6/n_p \right) + h^2 \left(6/n_p \right)^2 \right]^{1/2}, \tag{B.3}$$

para duplas hélices longas, onde n_s é o número de pares de bases de RNA, $d = 3{,}3a$ é o diâmetro aparente da hélice, $n_p = 11$ é o número de pares de bases por volta completa e o comprimento da haste para cada volta da dupla hélice é dado por $h = 5a$.

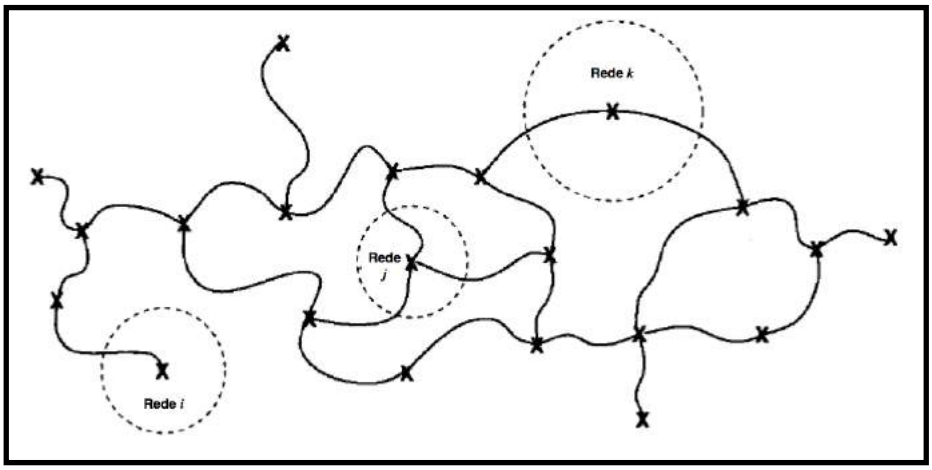
Tabela B.1 – Tabela de parâmetros para equação Equação (B.1)

Rede	D	A_1	A_2	A_3
ra_0	1	0	0	0
ra_1	s_0	β/D	0	0
ra_{2a}	$s_0s_1 + s_0s_2 + s_1s_2$	$\beta(s_1+s_2)/D$	$\beta(s_0+s_1)/D$	$\beta s_1/D$
ra_{2b}	$s_1(s_0+s_2)$	$\beta s_1/D$	$\beta(s_0+s_1+s_2)/D$	$\beta s_1/D$
rf_0	s_0	0	0	0
rf_1	s_0s_1	$\beta(s_0+s_1)/D$	0	0
rf_{2a}	$s_0s_3(s_1+s_2)+s_1s_2(s_0+s_3)$	$\beta(s_0+s_3)(s_1+s_2)/D$	$\beta(s_2+s_3)(s_0+s_1)/D$	$\beta(s_0s_2-s_1s_3)/D$
rf_{2b}	$s_1s_3(s_0+s_2)$	$\beta(s_0+s_2+s_3)s_1/D$	$\beta(s_0+s_1+s_2)s_3/D$	$\beta s_1s_3/D$

Fonte – Isambert e Siggia (2000)

Para a análise em larga escala, cada rede que constitui a estrutura é substituída por uma nodo único. Esses nodos estarão conectados entre si por regiões de fita simples ou regiões de fita dupla. A Figura B.1 ilustra o resultado dessa abordagem. A entropia conformacional resultante é determinada assumindo que esses nodos estão conectados através de "molas" Gaussianas, cujo comprimento médio quadrado é igual a distância média quadrada entre as redes conectadas em questão. De acordo com o trabalho de Isambert e Siggia (2000), a entropia desse "gel reticulado gaussiano" ("*Gaussian crosslinked gel*") é calculada através de $n-1$ integrações algébricas, onde n é o número de nodos e, consequentemente, de redes na estrutura. Quando dois nodos estão conectados por várias hastes, trata-se o problema como molas em paralelo, oscilando de acordo com a situação. Além disso, para melhor concordância com os resultados experimentais, Isambert e Siggia (2000) incluíram efeitos devido ao volume excluído, redefinindo o comprimento de equilíbrio das "molas": o expoente de volume excluído foi alterado de 0,5 para redes ideais para 0,65.

Figura B.1 – Esquema do "gel reticulado gaussiano"



Legenda: Análise em larga escala: cada nodo representa uma *rede*. As interações entre essas redes são modeladas através de "molas" gaussianas, cujo comprimento depende do número de nucleotídeos presentes nas regiões de fita simples ou duplas entre elas. Fonte: Modificado de Isambert e Siggia (2000)

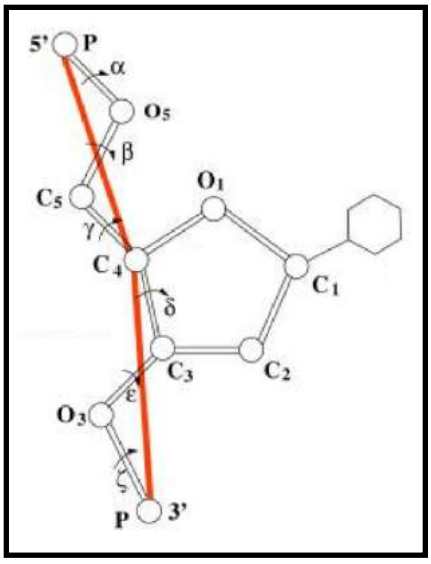
B.2 *Vfold*

O modelo *Vfold* sofreu alterações e inclusões ao longo dos anos, mas implementamos as regras e parâmetros apresentados por Cao e Chen (2006), Cao e Chen (2009) e por Liu e Chen (2010) por satisfazerem os objetivos desse trabalho. A publicação mais recente do grupo mostra que elementos estruturais mais complexos e outras subestruturas podem ser modelados através de sua abordagem (XU; CHEN, 2015).

Baseado em observações a respeito das estruturas conhecidas de RNA, o modelo aproxima a molécula através de ligações virtuais entre P-C_4, conforme mostra a Figura B.2, reduzindo drasticamente o espaço de conformações espaciais disponíveis. Destaca-se o fato que o ângulo dessa ligação virtual variar entre 90° e 120°, o que permite distribuir essas ligações numa *rede de diamante* numa primeira aproximação. Nessa rede, o ângulo de ligação é fixo e vale 109,5°. A distância entre as ligações é igual ao comprimento da ligação virtual, equivalente a 3,9 Å. Essa aproximação por granularidade grossa gera desvios médios quadrados (*RMSD*) da ordem de 2,2 Åquando duplas hélices de RNA com estruturas conhecidas são inseridas nessa rede. O modelo relaciona a entropia do laço à dupla hélice que

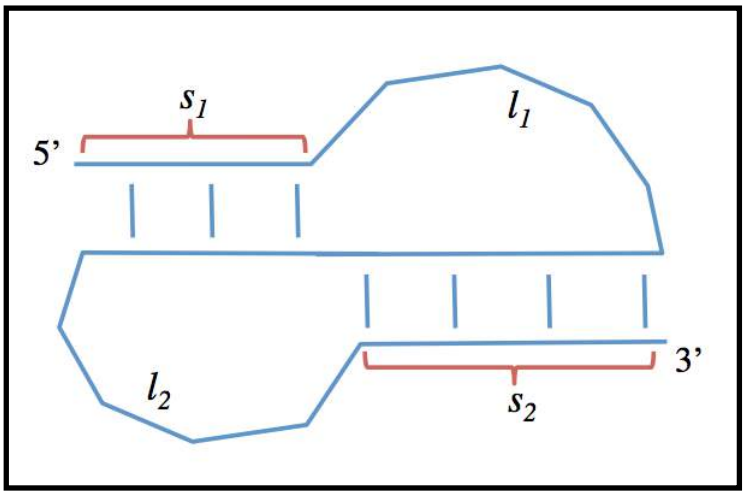
participa de sua formação, incluindo seu comprimento como um dos parâmetros necessários, e decompondo o pseudonó em duas subunidades, cada uma relacionada a um sulco, representados na Figura B.3.

Figura B.2 – Esquema para a aproximação da molécula de RNA por ligações virtuais



Legenda: Representação de uma base de RNA, apresentando os diferentes ângulos torcionais presentes, indicando a posição das ligações virtuais entre P-C_4. Fonte: Liu e Chen (2010)

Figura B.3 – Esquema para o pseudonó do tipo-H



Legenda: Representação indicando o sulco profundo, $l_1 + s_2$, e o sulco raso, $l_2 + s_1$. Fonte: Modificado de Cao e Chen (2006)

O número de conformações possíveis para cada subunidade é determinado através da enumeração exata de todos os caminhos auto-evitáveis permitidos na rede de diamante, com distância fixa entre as extremidades do laço. Esses valores serão denominados de ω_{L_i}, com i representando cada uma das subunidades do pseudonó. Um código para determinação dos valores de ω foi implementado. Desconsiderando a interação entre os dois laços, o número de conformações permitidas para o pseudonó será dada por

$$\Omega_S = \omega_{L_2} \cdot \omega_{L_2}. \tag{B.4}$$

A variação da entropia desses laços será dada por

$$\Delta S^\circ_{L_i} = -k_\mathrm{B} \ln \omega_{\mathrm{livre}(L_i)}/\omega_{L_i}, \tag{B.5}$$

onde $\omega_{\mathrm{livre}(L_i)}$ é dado pela enumeração do número de conformações permitidas para o laço livre, i.e., sem distância fixa entre suas extremidades. Determinamos localmente os valores para essas entropias, para laços de até 12 nt. Para laços mais longos, Cao e Chen (2006) sugerem a aproximação dos valores por

$$\ln \omega_{L_i} = a \ln(l - l_\mathrm{min} + 1) + b \ln(l - l_\mathrm{min} + 1) + c \quad \text{e} \tag{B.6}$$

$$\ln \omega_{\mathrm{livre}(L_i)} = 2{,}14\, l + 0{,}10, \tag{B.7}$$

onde l_min representa o menor laço permitido para o comprimento da dupla hélice da respectiva subunidade, e foram obtidos a partir de dados experimentais (PLEIJ; RIETVELD; BOSCH, 1985; GULTYAEV; BATENBURG; PLEIJ, 1999). Os valores para l_min, a, b e c em função do número de pares de bases presentes na respectiva hélice, s, estão presentes na Tabela B.2.

Cao e Chen (2009) estenderam o modelo, permitindo o cálculo da entropia para pseudonós do tipo-H com junções inter-hélices, ou seja, um terceiro laço, l_3, entre as duplas hélices s_1 e s_2. O valor da variação da entropia para cada subunidade passa a ser determinado através de

$$\Delta S^\circ_L / k_\mathrm{B} = a_1 \ln(l) + b_1 \tag{B.8}$$

Nesse caso, o número de possibilidades para os parâmetros a_1 e b_1 aumenta significantemente e serão omitidos aqui; seus valores estão presentes na material suplementar de Cao e Chen (2009).

Tabela B.2 – Tabela de parâmetros para equação Equação (B.6)

s_2	2	3	4	5	6	7	8	9	10	11	12
$l_{\min}$	4	2	1	1	1	1	2	2	2	5	5
a	0,12	0,39	−2,14	−2,22	−2,40	−2,61	−1,17	−1,66	−1,43	−0,14	0,77
b	1,96	1,92	2,15	2,11	2,18	2,21	2,03	2,09	2,09	2,06	1,84
c	0,52	−3,89	−2,09	−2,25	−2,33	−2,32	−1,96	−1,98	−2,93	0,15	−0,65
s_1	2	3	4	5	6	7	8	9	10	11	12
$l_{\min}$	4	2	3	4	4	5	5	6	6	9	9
a	0,95	0,32	1,77	3,99	7,73	8,38	4,52	9,05	4,77	2,74	4,69
b	1,84	1,92	1,82	1,55	1,29	1,16	1,61	1,15	1,68	2,05	1,80
c	−0,67	−3,90	−5,76	−5,86	−12,67	−11,45	−7,58	−11,45	−6,78	1,38	−1,11

Fonte – Cao e Chen (2006)

De posse dos valores das entropias dos laços, a variação da energia livre de Gibbs para um determinado pseudonó será dada por

$$\Delta G^{\circ}(s_1,s_2,l_1,l_2,l_3) = \Delta G^{\circ}_{s_1} + \Delta G^{\circ}_{s_2} - T\,\Delta S^{\circ}_{L_1} - T\,\Delta S^{\circ}_{L_2} + \Delta G^{\circ}_{\mathrm{EC}} + \Delta G^{\circ}_{\mathrm{conj}}, \tag{B.9}$$

onde $\Delta S^{\circ}_{L_i}$ é a entropia para o laço L_i, $\Delta G^{\circ}_{s_i}$ é a energia livre para a dupla hélice s_i segundo o modelo dos primeiros vizinhos, $\Delta G^{\circ}_{\mathrm{EC}}$ é a energia devido ao empilhamento coaxial entre as duplas hélices e $\Delta G^{\circ}_{\mathrm{conj}} = 1{,}3$ kcal/mol é adicionado devido à variação da entropia resultante da união das duas subunidades ($l_1 + s_2$ e $l_2 + s_1$) do pseudonó. Para todos os cálculos, consideramos $T = 310{,}15$ K.

APÊNDICE C – Sítios de pausa teóricos

A determinação dos sítios de pausa teóricos baseia-se no modelo descrito por Bai, Fulbright e Wang (2007) e atualizado por Costa, Acencio e Lemke (2013). Esse modelo foi novamente implementado, com algumas funções modificadas e outras optimizadas, de forma que simulações de sequências longas e com alta densidade de RNAP se tornassem possíveis. A Figura C.1 ilustra os estados possíveis para o CAT nesse modelo.

Cada nova incorporação é aproximada por uma reação química de três etapas, que representa os eventos apresentados na Figura C.1. Teremos:

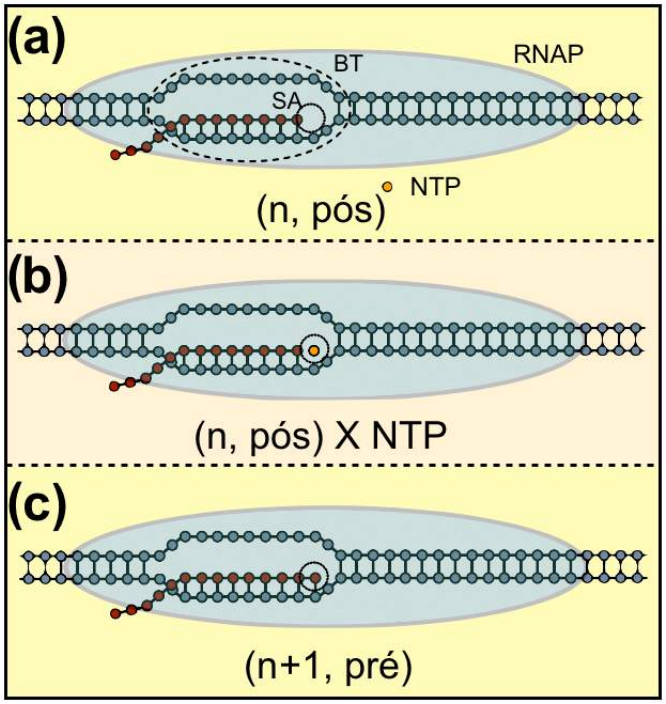
$$\begin{aligned} \mathrm{CAT}_{(n,\mathrm{pré})} &\xleftrightarrow{K_\mathrm{i}(n)} \mathrm{CAT}_{(n,\mathrm{pós})} \\ \mathrm{CAT}_{(n,\mathrm{pós})} + \mathrm{NTP} &\xleftrightarrow{K_\mathrm{d}^\mathrm{NTP}} \mathrm{CAT}_{(n,\mathrm{pós})} \times \mathrm{NTP} \\ \mathrm{CAT}_{(n,\mathrm{pós})} \times \mathrm{NTP} &\xrightarrow{k_\mathrm{max}^\mathrm{NTP}} \mathrm{CAT}_{(n+1,\mathrm{pré})} + \mathrm{PPi}. \end{aligned} \tag{C.1}$$

A primeira etapa dessa equação representa o deslocamento do CAT; a segunda trata a ligação de um novo nucleotídeo (NTP) através de uma catálise química. Os estados *pré* e *pós* para o CAT representam a posição do RNA nascente em relação ao sítio ativo da RNAP, enquanto n é o comprimento desse RNA. A Tabela C.1 apresenta os valores utilizados nesse trabalho para $K_\mathrm{d}^\mathrm{NTP}$ e $k_\mathrm{max}^\mathrm{NTP}$; tais valores foram experimentalmente estabelecidos e dependem do nucleotídeo que será incorporado. O valor para $K_\mathrm{i}(n)$ é dado por

$$K_\mathrm{i}(\mathrm{n}) = \exp[(\Delta G^\circ_{(n,\mathrm{pós})} - \Delta G^\circ_{(n,\mathrm{pré})} - F\,d)/k_\mathrm{B}T], \tag{C.2}$$

onde $\Delta G^\circ_{(n,m)}$ é a energia livre de Gibbs para determinada conformação do CAT e F representa uma força externa aplicada na RNAP: no caso, $F = 0$ pN. A distância entre dois pares de bases na fita de DNA, representada aqui por d, é de aproximadamente 0,34 nm.

Figura C.1 – Esquema dos estados e da estrutura do CAT



Legenda: SA: Sítio Ativo, BT: Bolha de Transcrição. O CAT será caracterizado pelo comprimento n do RNA e pela posição de seu sítio ativo em relação a extremidade 3' do RNA nascente, m. É equivalente usar $m = 0$ ou $m = \text{pré}$ e $m = 1$ ou $m = \text{pós}$. (a) Pós-translocado: o sítio ativo está livre. Temos a estrutura do CAT, formado por um híbrido RNA-DNA de 8 pb, uma bolha de transcrição de 14 pb, (12 estão separados mais 1 presente em cada limite entre a bolha e o DNA fita dupla), e a RNAP englobando 32 pb do DNA. (b) Pós-translocado, fase de incorporação: o sítio ativo foi recém ocupado, e o NTP será incorporado ao RNA. (c) Pré-translocado: sítio ativo ocupado e NTP efetivamente incorporado a fita de RNA. Nesse estado, o híbrido é formado por 9 pb. O CAT se movimentará um nucleotídeo para frente, retornando para uma conformação semelhante à apresentada em (a). Fonte: Produzido pelo próprio autor

Os valores de $\Delta G^{\circ}_{(n,m)}$ foram descritos primeiramente por Yager e vonHippel (1991) e medem sua estabilidade. Para cada conformação, este valor é obtido através de um somatório de três energias distintas:

$$\Delta G^{\circ}_{(n,m)} = \Delta G^{\circ}_{(n,m;\text{bolhaDNA})} + \Delta G^{\circ}_{(n,m;\text{RNA-DNA})} + \Delta G^{\circ}_{(n,m;\text{intRNAP})}. \tag{C.3}$$

O primeiro termo representa a energia liberada na quebra das pontes de hidrogênio entre os nucleotídeos complementares da fita de DNA para a formação da bolha de transcrição. O segundo termo é dado pela energia necessária para a formação do híbrido RNA-DNA. Ambos os termos são claramente dependentes da sequência que forma o CAT. Finalmente, o último termo representa a interação da RNAP com os ácidos nucleicos. O *Modelo dos primeiros vizinhos*, apresentado anteriormente, é utilizado para determinação das energias livres de Gibbs dessas fitas duplas. Nos cálculos, o terceiro termo da Equação (C.3) foi tomado como zero para simplificação.

Tabela C.1 – Valores para as constantes da reação química global da transcrição, presentes na Equação (C.1)

	ATP	UTP	GTP	CTP
$k^{\text{NTP}}_{\text{max}}$ (s^{-1})	50 ± 6	18 ± 1	36 ± 5	33 ± 6
$K^{\text{NTP}}_{\text{d}}$ (μM)	38 ± 7	24 ± 4	62 ± 18	7 ± 4

Fonte – Bai, Fulbright e Wang (2007)

A reação apresentada na Equação (C.1) corresponde à cinética enzimática de Michaelis-Menten na presença de um inibidor competitivo (MENTEN; MICHAELIS, 1913). O esquema típico da reação de Michaelis-Menten é modificado para incluir a ligação do inibidor na enzima livre e, realizando as devidas substituições, a taxa de ocorrência da reação global da Equação (C.1) será

$$k_{\text{main}}(\text{n}) = \frac{k^{\text{NTP}}_{\text{max}}[\text{NTP}]}{K^{\text{NTP}}_{\text{d}}\{1 + K_{\text{i}}(\text{n})\} + [\text{NTP}]}. \tag{C.4}$$

Além disso, esse modelo permite o movimento da RNAP na direção contrária à transcrição (*backtracking*). Sua taxa de ocorrência será dada por

$$k_{\text{n, m}\to\text{m}\pm1} = k_{\text{b}}\ \exp[(\Delta G^{\dagger} - \Delta G^{\circ}_{(n,m)} + F d(F))/k_{\text{B}}T], \tag{C.5}$$

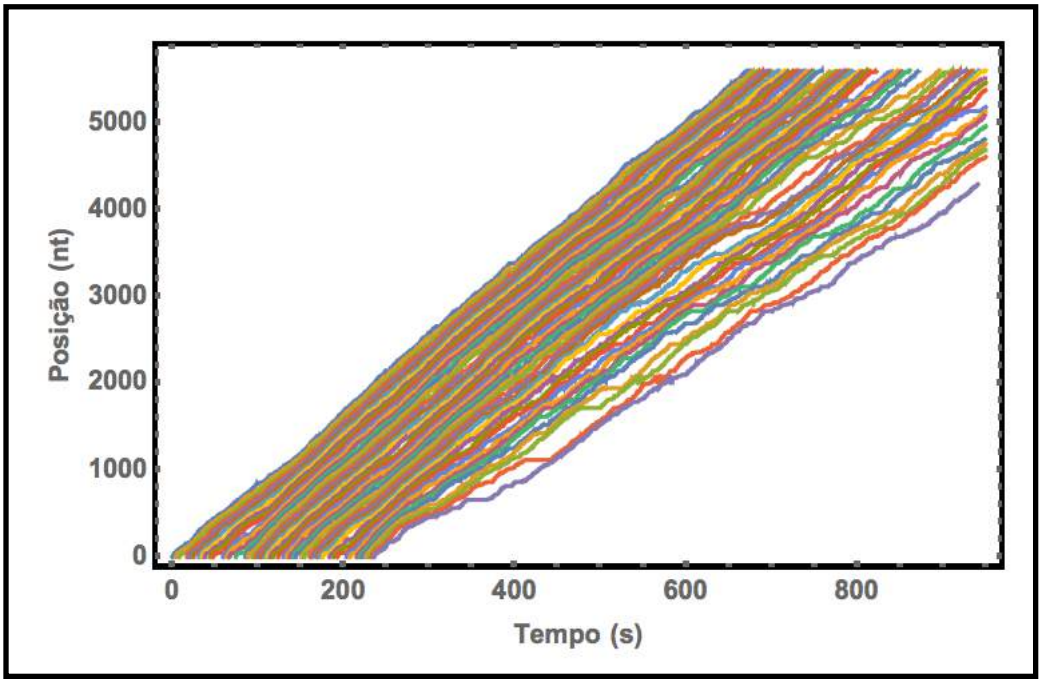
onde $k_{\text{b}} = 4{,}0 \times 10^{10}\ \text{s}^{-1}$ é um pré-fator constante e $\Delta G^{\dagger} = 41{,}2\ k_{\text{B}}T$ representa a barreira energética para ocorrência desse fenômeno. Esses valores não foram medidos

experimentalmente: foram ajustados por Bai, Fulbright e Wang (2007) a partir de seus resultados.

Inserimos, então, as reações apresentadas na Equação (C.1), realizando simulações estocásticas, utilizando o Método Direto do algoritmo de Gillespie. Armazena-se o tempo e a posição da RNAP na fita de DNA a cada ciclo e verifica-se o estado da RNAP: se ela está em alongamento normal ou em *backtracking*. Se a enzima entrar em *backtracking*, o algoritmo passa a simular a respectiva reação, utilizando o valor de k_{back} determinado pela Equação (C.5).

As novas funções implementadas para simulação do alongamento transcricionais são capazes de lidar com ≈50 RNAPs simultaneamente, em transcrição de sequências extremamente longas, com mais 5 mil nucleotídeos, como podemos ver na Figura C.2. Essa nova implementação tornou a etapa de identificação de sítios de pausa e de *backtracking* muito mais eficiente, e foi utilizada no programa para previsão dos sítios de pausa teóricos.

Figura C.2 – Gráfico para transcrição múltipla com 50 RNAPs



Legenda: Transcrição múltipla com 50 RNAPs na sequência do operon de RNA ribossômico *rrnB* de *E. coli*, com mais 5000 nt de comprimento. Fonte: Takahiro (2015)

Critérios para determinação dos sítios de pausa

Para sequências nas quais o comportamento cinético da RNAP é desconhecido, os sítios de pausa teóricos foram determinados através do seguinte procedimento: I) Simulações foram realizadas utilizando a abordagem da transcrição única (SRA, do inglês *Single Round Approach*) ou da transcrição múltipla (MRA, do inglês *Multiple Round Approach*); II) a partir da distribuição dos tempos de pausa para cada sítio, determinamos o valor de seu 5º percentil; III) selecionamos as n posições com tempos mais longos para esse percentil, com n determinado segundo o número de nucleotídeos que compõe a sequência dividido por 50; IV) voltamos para o passo II, mas dessa vez incrementando 5 percentis; V) no final, teremos listas com os sítios de pausa mais intensos a cada 5 percentis: selecionamos os termos presentes em pelo menos 70% dessas listas. Além disso, quando esses sítios de pausa distaram menos de 5 nucleotídeos entre si, reunimos esses valores em grupos compostos por até 3 membros, e consideramos a média desses conjuntos, arredondada para cima, como a posição da pausa.

O critério utilizado para determinação dos sítios de pausa para as sequências nas quais o comportamento cinético da RNAP é conhecido segue o proposto por Bai, Fulbright e Wang (2007):

$$\tau(n) > (1/\eta)\mathrm{Min}\{\tau(n)\}, \tag{C.6}$$

onde $\tau(n)$ representa o tempo de pausa no sítio n e o termo $\mathrm{Min}\{\tau(n)\}$ representa o menor tempo entre todos os valores de $\tau(n)$ para a dada sequência. Já η se trata de um parâmetro empírico: Costa, Acencio e Lemke (2013) definiram $0{,}38$ como o valor para transcrição múltipla e $0{,}64$ para transcrição única. Empiricamente optamos por $\eta = 0{,}5$, o que resulta em pausas da ordem de 1 s nas sequências de interesse.

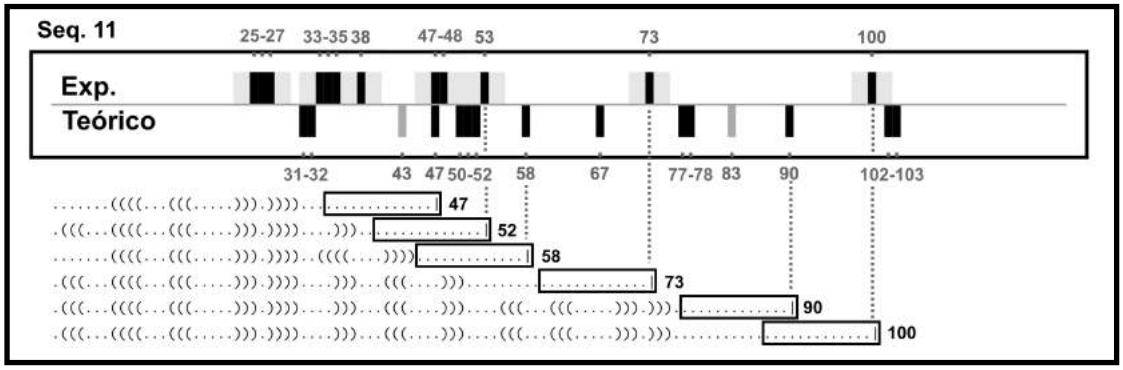
APÊNDICE D – Análise das estruturas nascentes

Os sítios de pausa experimentais e teóricos, assim como a conformação das esturuturas nascentes para as Sequências 11, 12 e 13 e para as deleções D104, D111, D112, D123, D167 e D387 da região inicial do genoma do bacteriófago T7 serão apresentadas a seguir. Os sítios de pausa experimentais para as Sequências 11, 12 e 13 foram obtidas por Tadigotla et al. (2006), enquanto para as deleções D104, D111, D112, D123, D167 e D387 foram publicadas por Levin e Chamberlin (1987). As Figuras seguem todas o mesmo padrão: representam o comportamento cinético experimental e teórico da RNAP na respectiva sequência, assim como o caminho de dobramento desse segmento segundo nossa implementação, representando a conformação da molécula nos sítios de interesse. A incerteza de ± 3 nt para os sítios de pausa experimentais está representada pelo sombreado cinza. Os sítios de pausa teóricos foram obtidos a partir da transcrição *in silico* segundo o modelo SRA. As discussões envolvem os conceitos de *pausas do tipo 1*, relacionadas à estabilidade da bolha de transcrição no estado pré-translocado da RNAP em relação ao estado pós-translocado, e de *pausas do tipo 2*, que se ocorrem devido ao movimento da RNAP no sentido oposto ao da transcrição (*backtracking*). Pausas teóricas do tipo 1 estão representadas em cinza, enquanto pausas do tipo 2 estão representadas em preto. As estruturas representadas logo abaixo do quadro de pausas em notação de pontos-e-perênteses foram obtidas a partir dos SBRs mais prováveis entre as réplicas das simulações, e estão alinhadas com as marcações do quadro de pausas. As linhas pontilhadas e o número à frente da representação da estrutura indicam a posição do sítio ativo da enzima. O pequeno quadro em torno de cada estrutura compreende os nucleotídeos que estão protegidos pelo canal interno da RNAP.

Sequência 11

Na Figura D.1 estão apresentados os resultados para a Sequência 11.

Figura D.1 – Esquema da distribuição dos sítios de pausa e respectivo caminho de dobramento para a Sequência 11



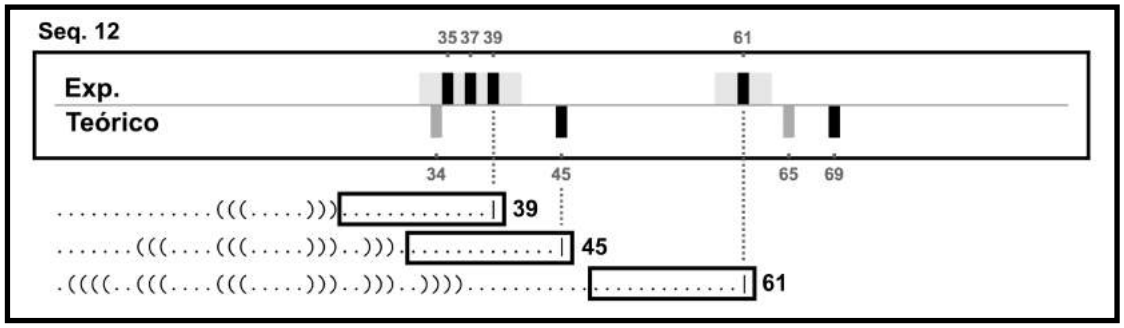
Legenda: Quadro de pausas experimentais e teóricas e respectiva estrutura predita para o RNA livre nos sítios mais relevantes para estudo. Fonte: Produzido pelo próprio autor

Não observamos SBRs até sítio de pausa na posição 38. Identificamos uma pausa do tipo 1 no sítio 43, que pode ser intensificada se o grampo formado imediatamente na saída do canal da RNAP interferir no equilíbrio enegético do CAT. O sítio 47 corresponde a uma pausa do tipo 2, observada experimentalmente e sem elementos estruturais que possam justificar sua supressão. Em seguida, observamos três sítios de pausa do tipo 2 agrupados, mas próximos de um SBR que poderá reduzir o potencial de recuo da RNAP, mantendo-a nesse agrupamento pausas observadas tanto teoricamente como experimentalmente. Já no sítio 58, o programa prevê uma pausa do tipo 2, mas não observada experimentalmente: a formação do grampo logo na saída do canal da enzima pode justificar essa supressão. Entre os sítios 67 e 83 não ocorre formação de SBRs, mas observamos falsos positivos e um verdadeiro negativo nessa região. Portanto, esse intervalo apresenta elementos suplementares não justificados por nossas aproximações. A pausa prevista no sítio 90 apresenta uma configuração semelhante a observada para o sítio 58: o SBR presente próximo a saída do canal pode justificar a supressão dessa pausa teória do tipo 2. Não se observam novos SBRs no transcrito nascente do sítio 90 até os últimos sítios de pausa teóricos, indicando a ausência de elementos restritivos para o *backtracking* previsto na região.

Sequência 12

Na Figura D.2 estão apresentados os resultados para a Sequência 12.

Figura D.2 – Esquema da distribuição dos sítios de pausa e respectivo caminho de dobramento para a Sequência 12



Legenda: Quadro de pausas experimentais e teóricas e respectiva estrutura predita para o RNA livre nos sítios mais relevantes para estudo. Fonte: Produzido pelo próprio autor

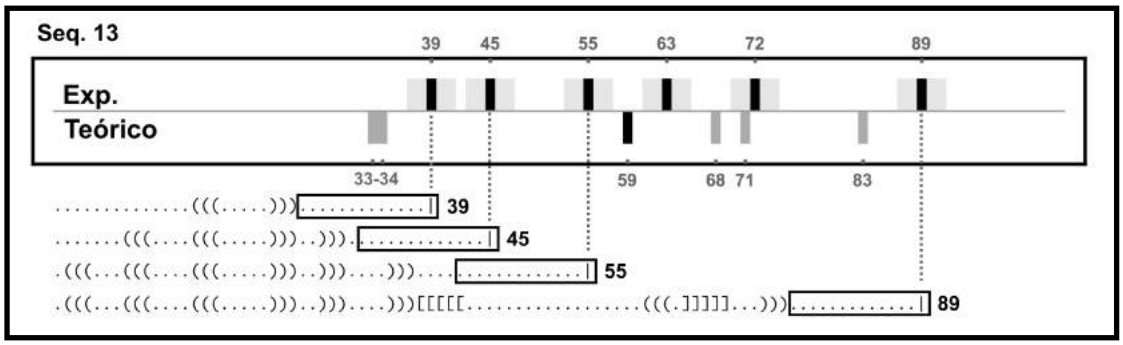
Não observamos SBRs antes do SUR apresentar pelo menos 25 nt. O grampo formado na saída do canal durante a 39ª incorporação apresenta energia livre de formação de −4,8 kcal/mol. Como há previsão de uma pausa do tipo 1 na região, o efeito combinado desses eventos pode ser responsável pelas pausas experimentalmente observadas. No sítio 45, o programa prevê um falso positivo do tipo 2, entretanto a presença do SBR pode suprimir o recuo da enzima nessa posição. Não observamos estruturas próximas à enzima além do SBR formado após a liberação do 36º ribonucleotídeo do canal de saída da RNAP. Logo, os eventos observados entre os sítios 61 e 69 não são corretamente identificados através de nossas aproximações.

Sequência 13

Na Figura D.3 estão apresentados os resultados para a Sequência 13.

As previsões dos sítios de pausa dessa sequência através do modelo SRA apresentam os mais baixos valores de TPR e PPV dentre os exemplos estudados. A tentativa de inclusão dos possíveis efeitos das estruturas formadas cotranscricionalmente não apresentou nenhum avanço nesse sentido.

Figura D.3 – Esquema da distribuição dos sítios de pausa e respectivo caminho de dobramento para a Sequência 13

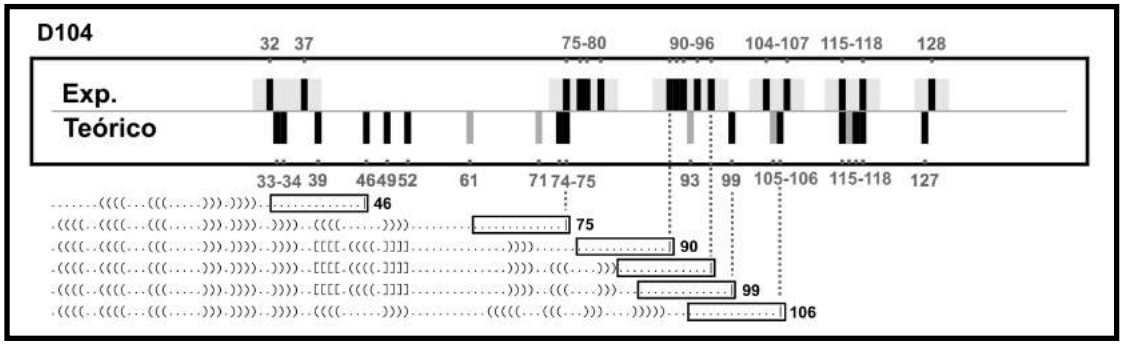


Legenda: Quadro de pausas experimentais e teóricas e respectiva estrutura predita para o RNA livre nos sítios mais relevantes para estudo. Fonte: Produzido pelo próprio autor

Deleção D104

Na Figura D.4 estão apresentados os resultados para a Deleção D104.

Figura D.4 – Esquema da distribuição dos sítios de pausa e respectivo caminho de dobramento para a Deleção D104



Legenda: Quadro de pausas experimentais e teóricas e respectiva estrutura predita para o RNA livre nos sítios mais relevantes para estudo. Fonte: Produzido pelo próprio autor

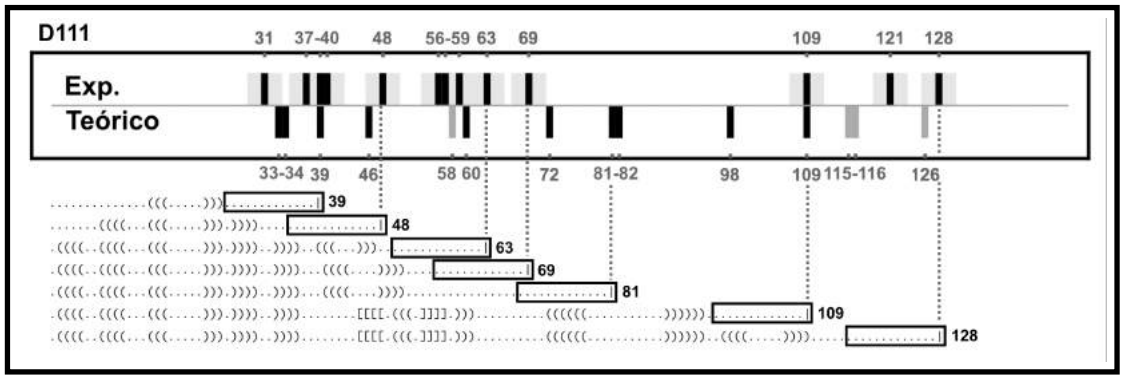
Novamente, a região inicial não apresentou SBRs. Observamos uma série de falsos positivos do tipo 2 e um falso positivo do tipo 1 sem elementos que possam justificar sua supressão. Já os sítios 74 e 75 são verdadeiros positivos do tipo 2 reforçados pelo mesmo motivo. O modelo SRA prevê a ocorrência de *backtracking* quando a RNAP atinge o sítio 99, incialmente identificado como um

falso positivo. Entretanto, dada sua proximidade em relação ao agrupamento de pausas experimentalmente verificadas, e a presença de um SBR na região *upstream*, esse recuo pode ser responsável pelo comportamento observado. A partir do sítio 106 não se observam novos SBRs no transcrito nascente até os últimos sítios de pausa teóricos, indicando a ausência de elementos restritivos para o *backtracking* previsto nessas regiões.

Deleção D111

Na Figura D.5 estão apresentados os resultados para a Deleção D111.

Figura D.5 – Esquema da distribuição dos sítios de pausa e respectivo caminho de dobramento para a Deleção 111



Legenda: Quadro de pausas experimentais e teóricas e respectiva estrutura predita para o RNA livre nos sítios mais relevantes para estudo. Fonte: Produzido pelo próprio autor

Não observamos SBRs antes do SUR apresentar pelo menos 25 nt. O grampo formado na saída do canal durante a 39ª incorporação pode suprimir o *backtracking* nessa posição. Porém, o resultado experimental refuta essa hipótese: o efeito de supressão esperado não é observado. Os primeiros 39 nt da Sequência 12 e da Deleção D111 apresentam alta similaridade (≈80%). Entretanto, o algoritmo não prevê *backtracking* no sítio 39 para a Sequência 12 pois as diferenças entre as sequências concentram-se justamente na bolha de transcrição dessa região. Ocorre também uma susbtituição de um nucleotídeo no SUR do grampo. Além disso, a Sequência 12 apresenta uam pausa do tipo 1 na região. Uma análise comparativa mais

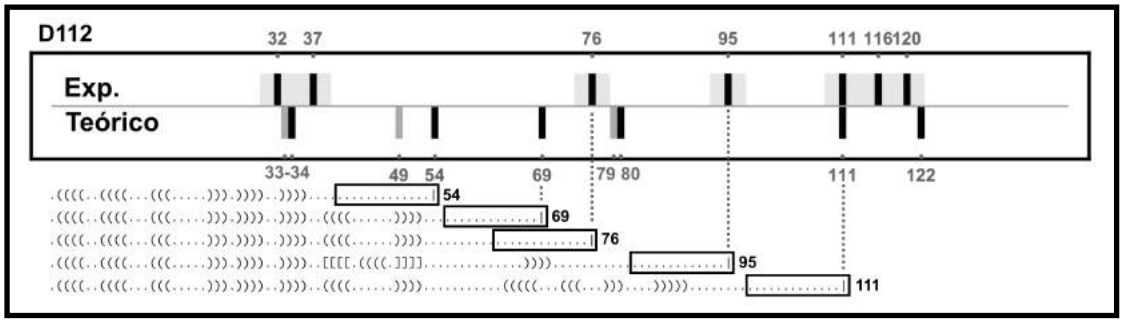
detalhada desses eventos, talvez por dinâmica molecular, poderá revelar maiores informações a respeito desse mecanismo de supressão.

O recuo previsto para a RNAP no sítio 46 fica restrito devido ao SBR 2 nt *upstream*. No sítio 60 ainda não houve a formação do grampo, que passa a ser energeticamente favorável no sítio seguinte. Logo não estão presentes elementos restritivos para o *backtracking* nessa posição. Também não se observam SBRs próximos do canal de saída da RNAP para a pausa do tipo 2 previstas nos sítios 72, 81-82 e 98. Desses, apenas o sítio 72 está próximo de uma pausa verificada no sítio 69. O recuo previsto para a RNAP no sítio 109 estará restrito devido ao SBR 1 nt *upstream*. A última estrutura observada corresponde ao grampo da região 97-109, que aparenta não afetar o comportamento da RNAP.

Deleção D112

Na Figura D.6 estão apresentados os resultados para a Deleção D112.

Figura D.6 – Esquema da distribuição dos sítios de pausa e respectivo caminho de dobramento para a Deleção D112



Legenda: Quadro de pausas experimentais e teóricas e respectiva estrutura predita para o RNA livre nos sítios mais relevantes para estudo. Fonte: Produzido pelo próprio autor

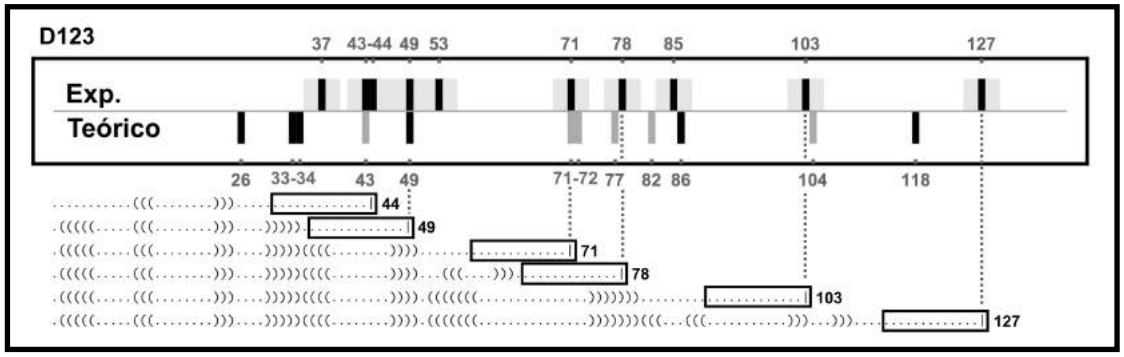
A região inicial não apresentou SBRs. Observamos um falso positivo do tipo 1 seguido por dois falsos positivos do tipo 2 sem elementos que possam justificar sua supressão. No sítio 80, também não se apresentam elementos supressores para essa pausa do tipo 2: entretanto, dada sua proximidade ao sítio de pausa experimentalmente verificado 76, o recuo nessa região pode ser responsável pelo

comportamento observado, reforçado pela pressença da pausa tipo 1 presente no sítio 79. Nem o modelo SRA, nem efeitos estruturais foram capazes de justificar a pausa experimental no sítio 95. Após a polimerização do 103º ribonucleotídeo, não se observam novos SBRs no transcrito nascente até os últimos sítios de pausa teóricos, indicando a ausência de elementos restritivos para o *backtracking* previsto nessas regiões.

Deleção D123

Na Figura D.7 estão apresentados os resultados para a Deleção D123.

Figura D.7 – Esquema da distribuição dos sítios de pausa e respectivo caminho de dobramento para a Deleção D123



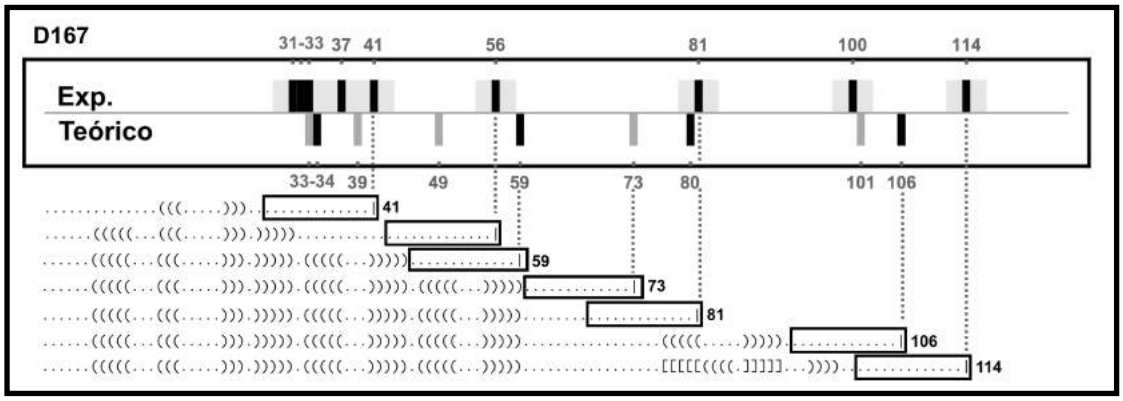
Legenda: Quadro de pausas experimentais e teóricas e respectiva estrutura predita para o RNA livre nos sítios mais relevantes para estudo. Fonte: Produzido pelo próprio autor

O primeiro grampo se forma na região 12-25 para essa sequência, sem interferir na região inicial da sequência. O recuo previsto para a RNAP no sítio 49 estará restrito devido ao SBR 1 nt *upstream*. Não se observa SBRs próximos ao canal da saída da RNAP nas pausas do tipo 1 localizadas no sítios 71-72; entretanto a pausa tipo 1 prevista no sítio 77 está relacionada à formação de um grampo efêmero logo na saída do canal da enzima. Nessa configuração, o grampo poderá interferir na estabilidade do CAT e aumentar a intensidade da pausa nessa posição. Apesar de novos grampos surgirem no restante do caminho de dobramento, nenhum deles possui potencial para interferir no comportamento cinético da RNAP.

Deleção D167

Na Figura D.8 estão apresentados os resultados para a Deleção D167.

Figura D.8 – Esquema da distribuição dos sítios de pausa e respectivo caminho de dobramento para a Deleção D167



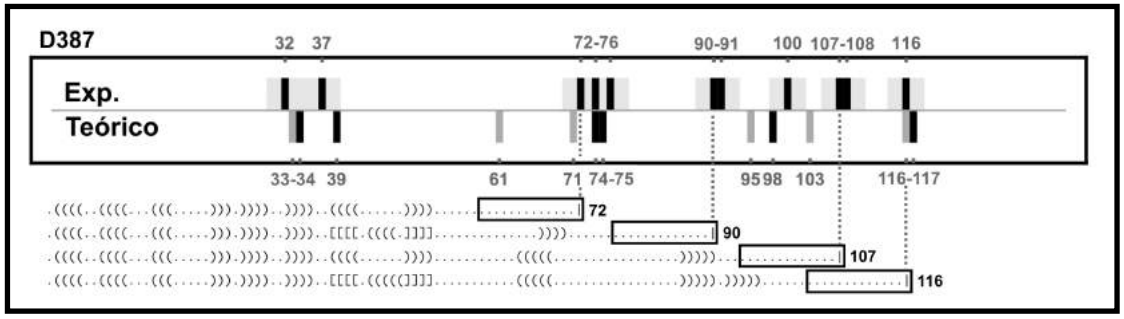
Legenda: Quadro de pausas experimentais e teóricas e respectiva estrutura predita para o RNA livre nos sítios mais relevantes para estudo. Fonte: Produzido pelo próprio autor

Logo na pausa do tipo 1 presente no sítio 39, já se observa a formação de grampo, o que poderá aumentar a intensidade da pausa nessa posição. O mesmo não ocorre no sítio 49. Já no sítio 59, observa-se uma configuração que tende a bloquear o recuo da RNAP, devido à formação do grampo logo no canal de saída da enzima. Em seguida temos uma pausa do tipo 1 na posição 73 apresenta a mesma configuração. Ambos os casos são inconsistentes com os resultados experimentais e pedem estudos mais detalhados. Adiante, não se observa elementos supressores para a pausa do tipo 2 localizada no sítio 80. Finalmente, o recuo previsto para a RNAP no sítio 106 estará restrito devido ao grampo 1 nt *upstream*, que se integrará o um pseudonó previsto após a transcrição do 111º nucleotídeo.

Deleção D387

Na Figura D.9 estão apresentados os resultados para a Deleção D387.

Figura D.9 – Esquema da distribuição dos sítios de pausa e respectivo caminho de dobramento para a Deleção D387



Legenda: Quadro de pausas experimentais e teóricas e respectiva estrutura predita para o RNA livre nos sítios mais relevantes para estudo. Fonte: Produzido pelo próprio autor

A única estrutura com potencial para interferir nas pausas previstas é o grampo localizado no intervalo entre o 64º e 90º nucleotídeo, que pode intensificar a pausa do tipo 1 localizada no sítio dessa sequência.

APÊNDICE E – Artigo

Paralelamente aos trabalhos desenvolvidos nessa Tese, o artigo intitulado "*Cooperative and Sequence-dependent Model for RNAP Dynamics: Application to Ribosomal Gene Transcription*" foi elaborado e submetido ao periódico de acesso livre *Scientific Reports*, com fator de impacto 5,23 (2015). Trata-se de uma análise do efeito colaborativo das RNAPs durante a transcrição de genes ribossomais de *Escherichia coli* segundo o modelo estocástico e dependente da sequência desenvolvido anteriormente pelo grupo. A versão revisada desse artigo encontra-se nas páginas seguintes. O trabalho passava por sua segunda revisão na data de conclusão dessa Tese.

Cooperative and Sequence-dependent Model for RNAP Dynamics: Application to Ribosomal Gene Transcription

Rafael Takahiro Nakajima[1,+], **Pedro Rafael Costa**[1,+,*], **and Ney Lemke**[1]

[1]Institute of Biosciences, UNESP - Univ Estadual Paulista, Department of Physics and Biophysics, Botucatu, 18618-689, Brazil
[*]costapr@ibb.unesp.br
[+]these authors contributed equally to this work

ABSTRACT

Escherichia coli ribosomal genes are a well-established experimental model used to investigate the transcription process. These genes are essential to cell physiology and therefore are strongly expressed. Multiple transcription units collaborate in rrn expression. Experiments involving electron microscopy showed non-uniform density of RNA polymerases transcribing these ribosomal operons. Here, we investigate RNAP collaborative transcription in *E. coli* ribosomal genes using a stochastic sequence-dependent model that included interactions among the RNAPs. We achieved results with better adherence to experimental data, considering a model with a different parametrization for genic and intergenic regions, as compared with previous attempts that used uniform parameters for genic and intergenic regions. Our model also showed that cooperative behaviour reduced the dwell times in pause sites predicted by single-round approach, but induced a new pausing event at a upstream position. We hope that this work might stimulate new experimental research and provide other scenarios to test our model predictions.

Introduction

Interest in the study of DNA transcription has increased since the rise of molecular biology. Transcription performed by the RNA polymerase enzyme (RNAP) can be divided into three phases: initiation, elongation and termination. RNAP scans duplex DNA to locate promoters sites for transcription initiation, followed by exposing of the DNA template and breaking of the hydrogen bounds between the strands. Once the RNAP active site is properly located, the now so-called Transcriptional Elongation Complex (TEC) enters the elongation phase and begins RNA polymerization by moving along the DNA strand. The TEC is then disassembled, and the RNAP releases the newly transcribed RNA and disengages from DNA during the termination phase. RNAP movement during the elongation phase is not uniformly progressive, but involves pauses and reverse movements (called backtracking). Transcriptional pauses can be structural: resulting from interactions between the RNAP and the nascent RNA or; sequence-specific: resulting from chemical interactions inside the transcription bubble. [1,2] These off-pathway states allow the recruitment of regulatory factors and play a major role in transcription termination.[3] Otherwise, collective behaviour can enhance transcription elongation rates, reducing the dwell time at several sites. Trailing RNAPs constitute physical barriers that prevent backtracking, and collisions between RNAPs push the leading molecule out of pausing sites, thereby inducing forward movement.[2,4]

Experimental evidence showed multiple-round transcription of some DNA regions. This behaviour have been observed and well characterized in the *Escherichia coli* bacteria ribosomal genes. For this organism several studies evaluated transcription rates and efficiency of the whole process. Ribosomal genes in *E. coli* are codified into seven distinctive operons: *rrnA*, *rrnB*, *rrnC*, *rrnD*, *rrnE*, *rrnF* and *rrnG*. Each operon is approximately 5500 bp long and contains a P1-regulated and a P2-constitutive promoter *in tandem*, followed by the 16S rRNA, 23S rRNA and 5S rRNA genes, along with transfer RNA (tRNA) gene sequences. The P1 promoter often shows more activity, resulting in P2 being physically blocked by the TEC. During the exponential growth, the *E. coli* cell needs a high ribosome concentration, and consequently, rRNA production is boosted. This process is known as growth rate-dependent control of ribosome synthesis.[5] The P1 initiation rate during steady-state growth is downregulated by interactions between the ppGpp molecule and the RNAP. The presence of the *Fis* protein bound upstream of the P1 promoter can enhance RNAP initiation rate. Meanwhile, the P2 promoter maintains basal levels of rRNA expression.[6] If an operon is deleted or inactivated, the intact operons are upregulated to supply the ribosome demand.[7] Ehrenberg, Dennis & Bremer inactivated four of the seven operons and noticed a 60% increase in the initiation rate at

other regions.[4] The elongation rate also increased from approximately 90 nt/s in the wild-type rrnB operon to an average of 180 nt/s.[5,7]

As these genes are highly transcribed, it is possible to observe hundreds of "branches" in electron micrographs, endorsing the high RNAP concentrations in these regions. Electron microscopy analysis also revealed that elongation rates strongly vary along these sequences, resulting in a non-uniform RNAP distribution along this operon.[5,8,9] Furthermore, traffic jams occur at 16S rRNA leader regions and between the 16S and the 23S rRNA genes.[5]

Fange *et al.*[7] conjectured that this RNAP density profile are caused by physical properties of the sequence, as the GC content. On the other hand, their simulations had shown that the initiation frequency was indirectly proportional to the the RNAP transient time in the *rrn* operon. They also proposed a DNA sequence-dependent thermodynamic model to predict the density of RNAP's during the transcription of rRNA in a wild-type bacteria and in bacteria with four of seven *rrn* operons inactivated by deletion.[4,5,9,10] Their model includes the initiation rate of transcription, translocation, and RNAP backward and forward tracking and it partially reproduced the observed transcript elongation rate variations along the *rrn* operon.[7]

In this study, we inspected the transcription of rRNA operons using an improved version of our multiple-round elongation approach,[2] a sequence-dependent thermodynamics-based model that includes collisions among the RNAPs. We simulated transcription of the first 5600 nt from the rrnB operon (Genbank U00096.3), obtained from Ecogene 3.0.[11]

We performed at least 300 simulations of the complete elongation under several conditions, herein called the *Nakajima-Costa-Lemke* (NCL) Models. The Models parameters were set and tunned based on *in vitro* and *in vivo* results, in particular the parameters of Bai *et al.*[12] and the RNAP-density profile along the segments of the *rrnB* operon.[9,10] The NCL Models embrace the *Standard Model* (SM), the *Heterogeneous Tuned Model* (TM[NTP]), the *Homogeneous Tuned Model* (HTM) and the *Intergenic Model* (IM). We analysed the kinetic behaviour, the operon transit times and the RNAP-density profile along 280 bp sections of the operon, comparing our output to the experimental results collected by electron-micrograph imaging from *in vivo* assays.[9]

Results

The multiple-round transcription model

The multiple-round approach (MRA) identifies and solves collisions between the RNAP molecules during multiple-round transcription and is based on the thermodynamic sequence-dependent model developed by Bai, Shundrovsky & Wang.[13] Each step of elongation was represented as a three-stage reaction that included the TEC translocation relative to the RNAP active site from pre-translocated to post-translocated, the ribonucleotide binding event and the chemical catalysis:

$$\begin{aligned} \mathrm{TEC}_{(n,\mathrm{pre})} &\overset{K_\mathrm{i}(n)}{\rightleftharpoons} \mathrm{TEC}_{(n,\mathrm{post})} \\ \mathrm{TEC}_{(n,\mathrm{post})} + \mathrm{NTP} &\overset{K_\mathrm{d}^\mathrm{NTP}}{\rightleftharpoons} \mathrm{TEC}_{(n,\mathrm{post})} \times \mathrm{NTP} \\ \mathrm{TEC}_{(n,\mathrm{post})} \times \mathrm{NTP} &\xrightarrow{k_\mathrm{max}^\mathrm{NTP}} \mathrm{TEC}_{(n+1,\mathrm{pre})} + \mathrm{PPi}, \end{aligned} \tag{1}$$

where n represents the current RNA length and $K_\mathrm{d}^\mathrm{NTP}$ and $k_\mathrm{max}^\mathrm{NTP}$ are NTP-specific (Table 1).

	ATP	UTP	GTP	CTP
$k_\mathrm{max}^\mathrm{NTP}$ (s^{-1})	50±6	18±1	36±5	33±6
$K_\mathrm{d}^\mathrm{NTP}$ (μM)	38±7	24±4	62±18	7±4

Table 1. NTP-dependent parameters for **Eq. 1**, experimentally determined by Bai, Fullbright & Wang.[12]

The overall rate for each ribonucleoside incorporation can be obtained considering a Michaelis-Menten kinetics:

$$k_\mathrm{main}(\mathrm{n}) = K_o \, \frac{k_\mathrm{max}^\mathrm{NTP}\,[\mathrm{NTP}]}{K_\mathrm{d}^\mathrm{NTP}\{1 + K_\mathrm{i}(\mathrm{n})\} + [\mathrm{NTP}]}, \tag{2}$$

where

$$K_\mathrm{i}(n) = \exp[(\Delta G_{(n,\mathrm{post})} - \Delta G_{(n,\mathrm{pre})} - Fd/k_B T]. \tag{3}$$

Here, $\Delta G_{(n,m)}$ is the standard Gibbs free energy for the TEC conformation, with n representing the length of the newly transcribed RNA, and m representing the position of the RNAP active site on the template. F is an external force applied to RNAP, with

$d \sim 0.34$ nm being the distance between adjacent nucleotides in the DNA strand.[12] There is also a backtracking/forward-tracking rate given by

$$k_{n,m\to m\pm1} = k_0 \exp[(\Delta G^{\dagger} - \Delta G_{(n,m)} + Fd(F))/k_B T]. \quad (4)$$

All the Gibbs free energy terms, ΔG, are related to the breaking of the hydrogen bonds between complementary nucleotides on the double-stranded DNA and the formation of the RNA-DNA-hybrid duplex, considering a transcription bubble 15 nt long and a RNA-DNA hybrid with 9 bp when the enzyme active site is occupied.

We used the Gillespie stochastic algorithm[14] to simulate each nucleotide incorporation during elongation. Each iteration of algorithm determined where each RNAP molecule was when these events occurred. If the algorithm predicted that two RNAPs would collide, considering that each one covered 40 bp on the DNA, the reaction was aborted and a new reaction was considered. If the trailing RNAP was performing normal elongation, it applied an external force of F on the leading RNAP. This would result in a change to the transient elongation rate, as seen in Equations 3 and 4. The transcript initiation relies only on the unobstructed promoter: as soon as there is no other RNAP occupying it, a new RNAP can start the transcription.

The NCL Models

Table 2 present details about the parameters employed in all NCL Models.

Parameter	Value	Details
F	15 pN	Force applied to the leading RNAP during a collision. The maximum load that a transcribing RNAP molecule can overcome is approximately 25 pN.[15]
$\Delta G^{\dagger}$	41.2 $k_B T$	Free-energy barrier for *backtracking*.
k_0	4.0×10^{10} s^{-1}	*Backtracking* constant pre-factor.
$[NTP]$	100 μM, 1000 μM	NTP concentration in μM.
$k_{\text{main}}(n < 280)$	300 nt/s	The Nus anti-termination factor speeds up the elongation on the early region, avoiding premature pausing and allowing continuous promoter regeneration.[10]

Table 2. Parameters details and their values for the NCL Models.

In our first Model, the *Standard Model*, SM, we used just the parameters presented in that Table 2. The high NTP concentration used in this Model ([NTP]=1000 μM) reflects the attempt to retrieve the experimental transcription rate. To improve the Model accuracy with a small increase in complexity, we explored the effects of K_o variation (Equation 2) on MRA. We started tuning K_o based on the RNAP-density profile and developed the *Heterogeneous Tuned Model* (TM[NTP]), and analyzed the effect of changing NTP concentration from 1000 μM to 100 μM. In these Models, each one of the 20 DNA segments had a specific fitting parameter, K_o, directly related to the elongation rate. Based on the results of the TM[NTP], we developed the *Homogeneous Tuned Model* (HTM) empirically setting an a uniform K_o value along the operon, which dramatically reduces the number of free-parameters in this approach as compared to TM[NTP]. We selected the lower NTP concentration when developing this Model, to enhance effects of the TEC conformation in Equation 2 and ensuring the density variation along the operon. Finally, empirically setting different K_o values for the intergenic and genic regions, we developed the *Intergenic Model* (IM). As a reference, we have also implemented the deterministic *Constant Model* (CM), where the RNAP elongation rate in each operon segment were obtained directly from the experimental data of Quan *et al.*[9]

Table 3 presents a summary about NCL Models and the number of free parameters (FPs) used during the simulations, along with the deterministic *Constant Model* (CM) and *Quan Average*,[7] which include as a parameter the initiation rate, k_I.

In our simulations, we also included a termination site on the 18$^{\text{th}}$ segment, where the RNAP has an 80% chance of dissociating from the DNA strand and aborting the transcription according to Fange *et al.*[10]

rRNA Operon Transcription *in silico*

In vivo, there are 53.4 active RNAPs molecules in average on the wild-type rrn operon and each one take about 61.5 s to transit along the operon.[5,9,10] Table 4 shows the average number of active RNAPs during the equilibrium state and the operon transit median times (T) for each Model and the corresponding ratio between these results and the experimental values.

Based on the results presented on Table 4, we choose TM1000, HTM and IM as representative Models for further discussion. The experimental profile of Quan *et al* were also used as reference for model adherence. Figure 1 compares the relative RNAP distribution along the operon for models TM1000, HTM and IM with the experimental data. Figure 2 shows the root-mean-square error (RMSE) between the experimental RNAP density data from Quan *et al.*[9] and the predict results plotted

Model	FPs	Details
SM	5	This model employs only the parameters listed on Table 2. [NTP]=1000 μM.
TM[NTP]	24	Tuned values for $K_{o;j}$ for each segment $2 \leq j \leq 20$. [NTP]=$\{100\ \mu$M, 1000 μM$\}$.
HTM	6	An overall $K_o = 12.5$ was empirically set aiming the best RNAP density fitness. [NTP]=100 μM.
IM	8	Overall $K_o = 12.5$, $K_o = 15$ for 16S-23S intergenic region and $K_o = 35$ for 5S-End region. [NTP]=100 μM.
CM	21	Average elongation rates from Quan *et al.*[9] for each segment and $k_I = 0.87$ s^{-1}.
Quan Av.	2	$k_I = 0.87$ and overall transcription rate of 91 nt/s.

Table 3. Number of free parameters (FPs) and details about the implemented models.

Model	RNAPs	T (s)
SM	105.2 (197%)	292.2 (475%)
TM100	33.6 (63%)	40.1 (65%)
TM1000	34.7 (65%)	51.8 (84%)
HTM	54.5 (102%)	80.1 (130%)
IM	50.7 (95%)	67.6 (110%)

Table 4. The average number of active RNAPs and the operon transit median time (T) for each model. The corresponding ratio between predicted and observed values (53.4 RNAPs and 61.5 s) are also presented.

against the Bayesian information criterion (BIC), which compares the performance among the models considering the number of free-parameters of each one (see Table 3). The results obtained by the models proposed by Fange *et al.* (FMDE) are also presented, keeping Fange *et al.* original nomenclature.[7] The result for the *Quan Average* is shown as well: it leads to a constant RNAP density (5%) between successive sections. The purple quadrant indicates the region where the predicted results RMSE and BIC are worse than *Quan average*, while the blue quadrant indicates the opposite: both RMSE and BIC are better than *Quan average*.

We continued to explore TM1000, HTM and IM, studying the RNAP behaviour for single-round and multiple-round approaches (SRA and MRA, respectively) for each case. Figure 3 shows the simplified box-plot (upper and lower quartiles) the operon transit times. We noticed a wider distribution for SRA, while the median MRA was lower. Biologically, this indicate that RNAP cooperation reduced overall transcription time and enhanced process uniformity.

Figure 4 shows an example of the RNAP kinetics for these Models and also for the deterministic *Constant Model* (CM), from the 500$^{\text{th}}$ base pair to the 2500$^{\text{th}}$ on the DNA strand spanning 30 s. In this representation, pausing events arise as horizontal segments. Comparing the time evolution of individual enzymes in the representative models TM1000, HTM and IM, and the CM, we perceive some the microscopic details of these Models. The RNAPs in CM presented a laminar-like flow behaviour that returned a uniform distribution in each segment, without collisions, pausing or backtracking. TM1000 maintained some uniformity in each segment, but traffic jams and pausing events eventually occurred. In HTM, we observed irregular behaviour presented as pausing, traffic jams and some RNAP backtracking. Finally, the RNAPs in IM reacted similarly to those in HTM, but slightly more erratic, with traffic jams, backtracking and even an evident pausing event occurring when the transcript was approximately 1750 nt-long.

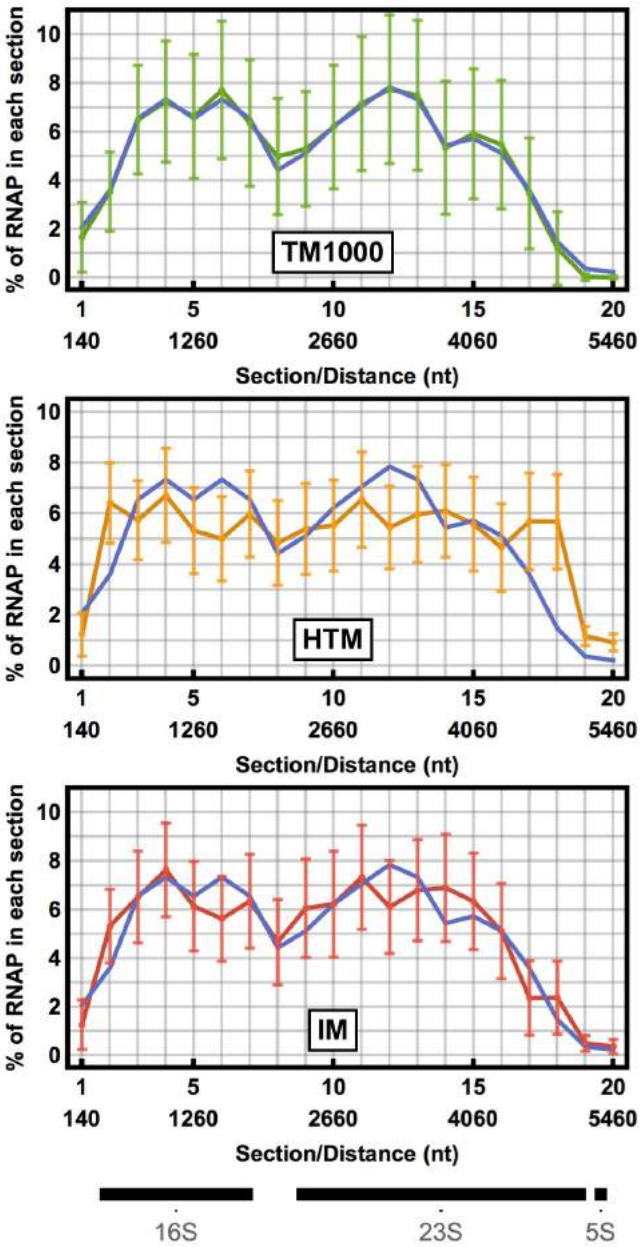


Figure 1. Relative numbers of active RNAPs in each section of the rrnB operon. The blue line represents experimental results from Quan *et al*.[9] The points represent the average number of RNAPs found during the simulations, and the error bars represent the standard deviation for each value.

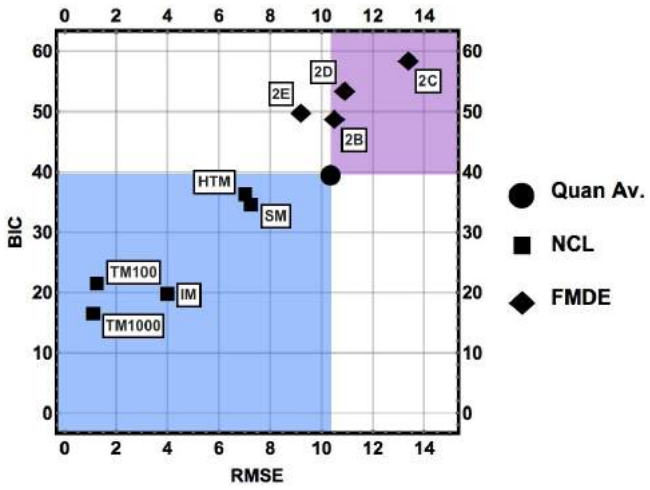


Figure 2. Adherence for each proposed model to experimental RNAP density profile. We characterized the simulated RNAP-density profiles by measuring the RMSE and BIC using experimental results[9] as reference. Lower values indicate better adherence for both scores.

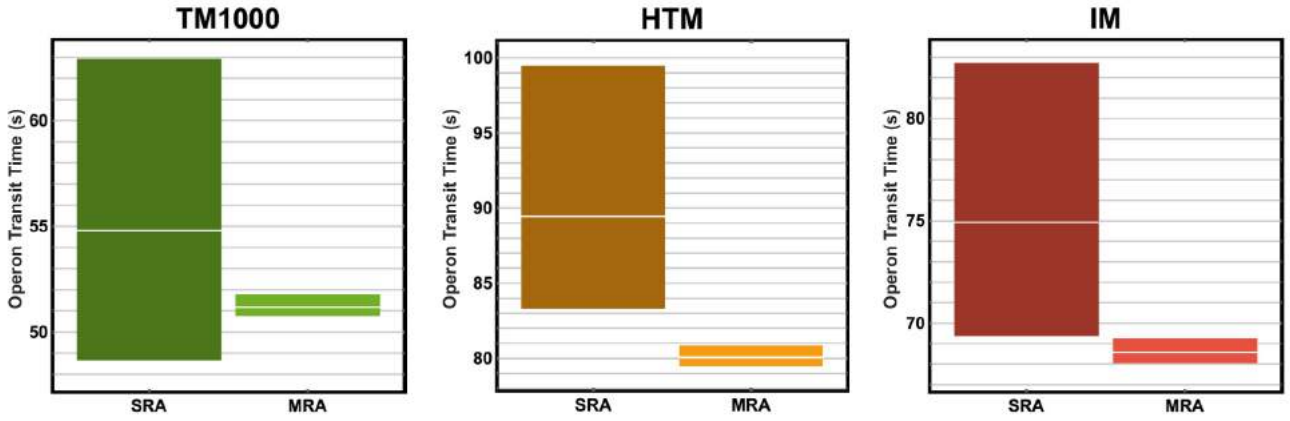


Figure 3. Comparison between the operon transit time distributions for SRA and MRA. The box represents the the upper and lower quartiles, while the white line represents the median.

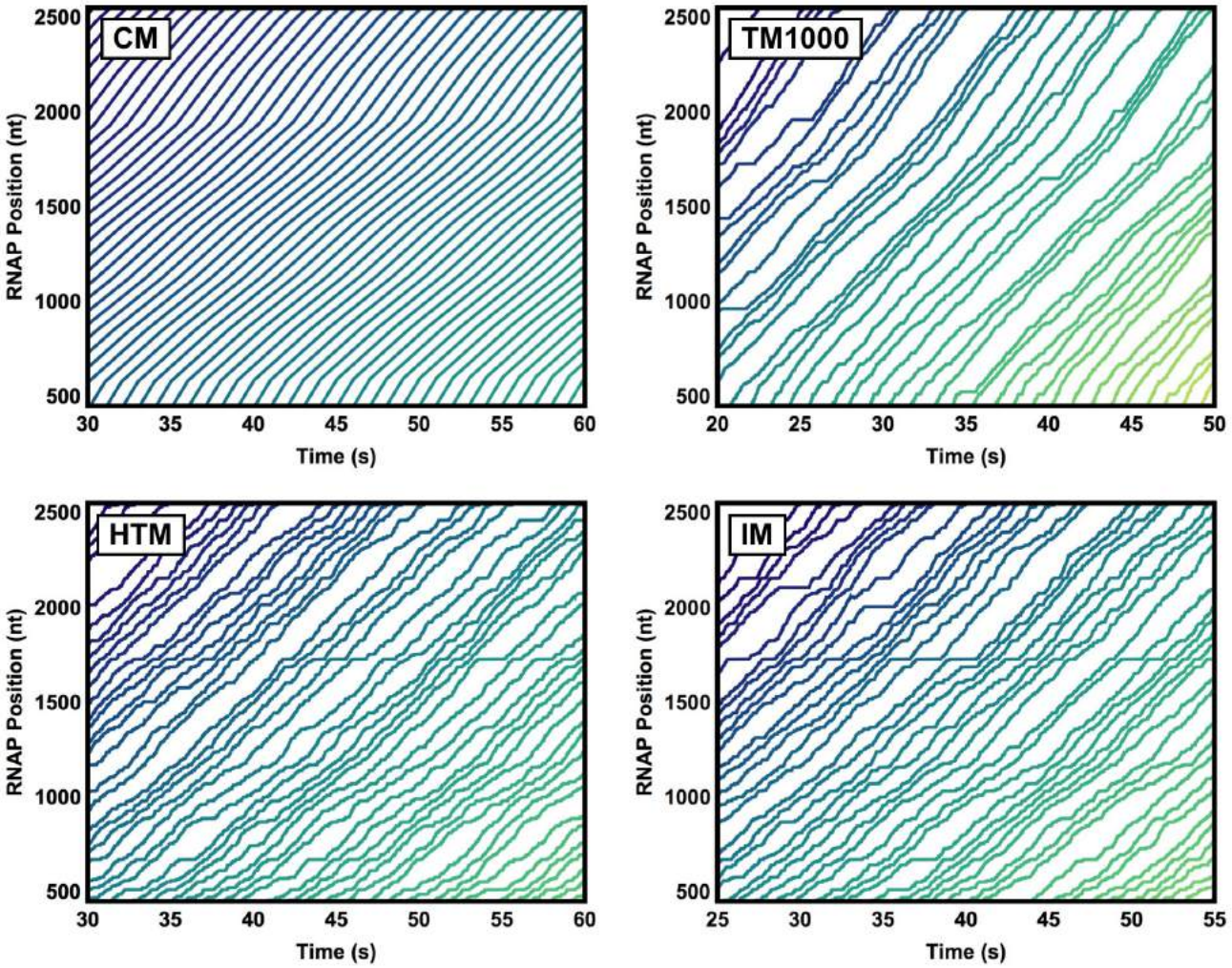


Figure 4. Examples of RNAP kinetic behaviour for different models. Each curve corresponds to a different RNAP, from the leaders (in blue tones) to the followers (green to yellow tones).

In summary the NCL Models performance:

SM: The SM results from Table 4 shows that this Model is inappropriate, exhibiting unrealistically low transcription rates, even at high NTP concentrations. That implies an uniform RNAP distribution along the operon, and, therefore, it is normal to find molecules flanking and being flanked by others due to the RNAP movement becoming very constrained. This was expected, nonetheless, since the original NTP-dependent parameters were obtained from *in vitro* experimental conditions.[12] The higher elongation rates associated with rRNA-operon transcription strongly suggested pausing suppression, indicating the presence of supplementary effects affecting this process *in vivo*.

TM1000: The RSME values for TM1000 were very small, i.e., the relative number of active RNAP-profile reassembles the experimental data, as well the operon transit time. Despite this, as can be seen in Table 4, the number of actively transcribing RNAP was approximately 65% of the 53.4 molecules found on average by Quan *et al.*[9]

HTM: The distribution densities of RNAPs in HTM exhibited an irregular profile and a higher RMSE as compared to later models, but the ratio between the number of active enzymes in this model and the experimental value was 102% (Table 4), which was slightly higher than the expected.

IM: Using IM, we obtained results that adhered more closely to experimental results relative to previous attempts, as shown in Figure 2. Even the number of active enzymes on the DNA strand and the operon transit time has agreed well with the data.

Based on the above results, we choose the IM as our best Model, as its output showed a good balance between the accuracy and the number of parameters used in simulations.

RNAP cooperative behaviour analysis

For further investigations about the differences between the SRA and the MRA, Figure 5A shows the third quartile of the distribution of dwell times for each NTP incorporation on the RNA strand for IM, comparing the SRA and MRA. We noticed that the SRA exhibited a wider distribution, but there was no clear spatial pattern between the approaches. When we analysed the differences between SRA and MRA times ($\Delta(n) = t_{\mathrm{SRA}}(n) - t_{\mathrm{MRA}}(n)$, Figure 5B) a clear pattern emerged. At strong pausing sites for SRA, $\Delta(n)$ was longer and $\Delta(n-40)$ exhibits negative values, indicating that $t_{\mathrm{MRA}} > t_{\mathrm{SRA}}$ at these positions. We emphasize that this 40 nt distance matched the minimum distance between the active sites of two subsequent RNAPs, because the enzymes enclosed 40 bp of the DNA strand in our model. We inferred that the trailing RNAP blocked and "pushed away" the leading RNAP from pausing sites, but this event would retard the trailing RNAP movement towards the upstream position. To test this hypothesis, Figure 5C depicts $\Delta(n) \times \Delta(n-40)$, where the fitted curve has a p-value < 0.01, indicating a strong correlation between these values. We also investigated correlations between $\Delta(n)$ and $\Delta(n-m)$ for $1 < m < 120$, but for any $m \neq 40$ we obtained p-values> 0.05, indicating that all of these correlations can be considered nonsignificant.

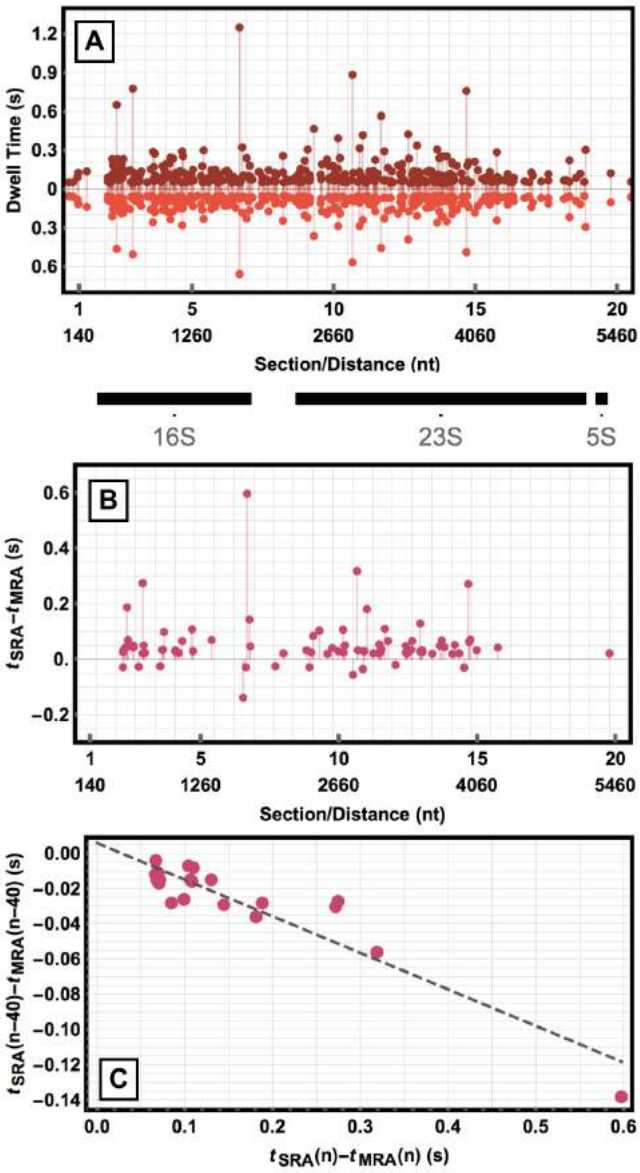


Figure 5. Dwell-time (t) analysis for IM. (A) Third quartile of the distribution of t for each position on the rrnB operon. The values for SRA are presented above, and the MRA profile is presented below. Values $\leq$0.05 s were excluded to enhance visualization; they corresponded for $\geq$90% occurrences in IM. (B) Differences between the t for SRA and MRA. (C) Correlation between the top 20 differences in the positive t distributions and the increase in t 40 nt upstream.

Discussion

We applied our multiple-round sequence-dependent transcription model to the rrnB operon to investigate the impact of cooperation among RNAPs on elongation dynamics. We recovered the experimental results by including details about the nature of this specific process. Therefore, the roughness of the RNAP-density profile might be explained by sequence-specific pausing sites and considerations about the genic and intergenic regions.

Despite their accuracy, the different Models allow us to to gain insight into the emergent properties of our multiple-round approach. The remarkable uniformity in operon transit times verified in Figure 3 can be explained by one isolated enzyme having a high probability of backtracking or entering pausing state along the entire DNA strand. In these cases, the molecules tend to abort transcription unless some other molecule assists them. Backtracking occurred in all proposed models, but it was strongly suppressed in MRA in relation to SRA due to the "roadblock" effect (for further details, see Klump & Hwa[6] and Costa, Acencio & Lemke[2]). MRA grants kinetic homogeneity and elongation solidity, which can be seen as intrinsic to transcription regulation. On the other hand, experimental and theoretical results indicate that backtracking correlates with transcriptional proofreading.[16,17] We can conjecture that backtracking suppression due to RNAP interactions will effect transcription accuracy. We believe that this point could be further investigated by including in our models misincorporations rates and by comparing the MRA transcripts output profile with RNAseq datasets.

Finally, we analyzed the transcriptional pausing profile in rrnB operon and RNAP cooperative behaviour according to our *Intergenic Model*. Transcription-pausing events can affect the regulation of other regions and allow recruitment of other molecules, such as transcription factors and inhibitors. The correlation between the most significant positive differences found in Figure 5B shown an increase on the dwell time 40 nt upstream in MRA. We stress that this upstream pausing-event induction could not be accurately predicted unless a model that explicitly considers interactions among RNAPs.

In summary, our extended model for multiple-round transcription was in accordance with the literature, adequately predicted the nature of RNAP transcription, and achieved improved results as compared with other models. Our multiple-round approach showed that cooperative behaviour among RNAPs reduced long pausing events, but the collisions induced trailing RNAPs to pausing at the corresponding sites. Furthermore, the experimental parameters for our simulations were obtained from *in vitro* experiments using small DNA segments[12] and we recovered the experimental results for the *rrnB* operon by changing the pace of transcription on genic and intergenic regions on a long DNA sequence. We expect that this model can be successfully applied in other contexts and we hope that our findings should stimulate further experimental assays to provide a richer dataset for to be used in testing our model predictions. We are currently extending this model to include the influence of the secondary structure on the elongation process. Tadigotla *et al*.,[18] analyzing sequences 170 nt long, allowed the entire RNA transcript beyond the exit channel to fold and included a free energy penalty associated with the breaking of base pairs of the resulting structure when RNAP backtracks. This approach can be applied on long sequences, such as the *rrnB* operon, but request careful analysis and further investigation on RNA cotranscriptional folding behaviour. Also, the tricky effects of these long secondary structure on RNAP itself still an open issue.

Methods

All computations were perform using in-house codes written Wolfram Language. The free-energy for the TEC are evaluated using a sequence-dependent nearest-neighbour model. We used data from SantaLucia, Allawi & Seneviratne [19] for the DNA-DNA energy and from Sugimoto *et al.*[20] for the RNA-DNA-hybrid energy. For further details and considerations on these calculations, see Bai, Shundrovsky & Wang.[13] The MRA simulations were performed until an equilibrium state was identified. Further information about the multiple-round approach can be found on Costa, Acencio & Lemke.[2]

References

1. Herbert, K. M. *et al.* Sequence-resolved detection of pausing by single rna polymerase molecules. *Cell* **125**, 1083–1094 (2006).
2. Costa, P. R., Acencio, M. L. & Lemke, N. Cooperative rna polymerase molecules behavior on a stochastic sequence-dependent model for transcription elongation. *PLOS ONE* **8**, e57328–11 (2013).
3. Landick, R. The regulatory roles and mechanism of transcriptional pausing. *Biochemical Society Transactions* **34**, 1062–1066 (2006).
4. Ehrenberg, M., Dennis, P. & Bremer, H. Maximum rrn promoter activity in escherichia coli at saturating concentrations of free rna polymerase. *Biochimie* **92**, 12–20 (2010).
5. Condon, C., French, S., Squires, C. & Squires, C. L. Depletion of functional ribosomal rna operons in escherichia coli causes increased expression of the remaining intact copies. *The EMBO journal* **12**, 4305 (1993).

6. Klumpp, S. & Hwa, T. Stochasticity and traffic jams in the transcription of ribosomal rna: Intriguing role of termination and antitermination. *Proceedings of the National Academy of Sciences* **105**, 18159–18164 (2008).
7. Fange, D., Mellenius, H., Dennis, P. P. & Ehrenberg, M. Thermodynamic modeling of variations in the rate of rna chain elongation of e. coli rrn operons. *Biophysical journal* **106**, 55–64 (2014).
8. Bremer, H. & Dennis, P. P. *Escherichia coli and Salmonella: Cellular and Molecular Biology*, chap. Modulation of chemical composition and other parameters of the cell by growth rate, 1553–1569 (ASM Press, Washington, D.C., 1996).
9. Quan, S., Zhang, N., French, S. & Squires, C. L. Transcriptional polarity in rrna operons of escherichia coli nusa and nusb mutant strains. *Journal of bacteriology* **187**, 1632–1638 (2005).
10. Dennis, P., Ehrenberg, M., Fange, D. & Bremer, H. Varying rate of rna chain elongation during rrn transcription in escherichia coli. *Journal of bacteriology* **191**, 3740–3746 (2009).
11. Zhou, J. & Rudd, K. E. Ecogene3.0. *Nucleic Acids Research* **41**, 613–624 (2013).
12. Bai, L., Fulbright, R. M. & Wang, M. D. Mechanochemical kinetics of transcription elongation. *Physical review letters* **98**, 068103 (2007).
13. Bai, L., Shundrovsky, A. & Wang, M. D. Sequence-dependent kinetic model for transcription elongation by rna polymerase. *Journal of molecular biology* **344**, 335–349 (2004).
14. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry* **81**, 2340–2361 (1977).
15. Wang, M. D. *et al.* Force and velocity measured for single molecules of rna polymerase. *Science* **282**, 902–907 (1998).
16. Shaevitz, J. W., Abbondanzieri, E. A., Landick, R. & Block, S. M. Backtracking by single rna polymerase molecules observed at near-base-pair resolution. *Nature* **426**, 684–687 (2003).
17. Mellenius, H. & Ehrenberg, M. Dna template dependent accuracy variation of nucleotide selection in transcription. *PloS one* **10**, e0119588 (2015).
18. Tadigotla, V. R. *et al.* Thermodynamic and kinetic modeling of transcriptional pausing. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 4439–4444 (2006).
19. SantaLucia, J., Allawi, H. T. & Seneviratne, P. A. Improved nearest-neighbor parameters for predicting dna duplex stability. *Biochemistry* **35**, 3555–3562 (1996).
20. Sugimoto, N. *et al.* Thermodynamic parameters to predict stability of rna/dna hybrid duplexes. *Biochemistry* **34**, 11211–11216 (1995).

Acknowledgements

The authors would like to thank the Brazilian agencies FAPESP (The State of São Paulo Research Foundation) for the financial support through the FAPESP research grants 2013/06683-2 and 2012/19377-4, CNPq (National Council for Scientific and Technological Development) research grant 152838/2012-0 and CAPES (Coordination of Improvement of Higher Education Personnel).

Author contributions statement

R.T. conceived the idea. P.C. developed the in-house codes (Wolfram Language). R.T. and P.C. conducted the simulations. N.L. supervised the entire project. All authors analysed the results, wrote and reviewed the manuscript.

Additional information

Competing financial interests: The authors declare no competing financial interests.