UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

Pedro Rafael Costa

Modelo cinético estocástico para a transcrição considerando colisões entre moléculas de RNA Polimerase

Botucatu – SP

2012

Pedro Rafael Costa

Modelo cinético estocástico para a transcrição considerando colisões entre moléculas de RNA Polimerase

Dissertação apresentada ao Instituto de Biociências da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de Botucatu, para a obtenção do título de Mestre em Ciências Biológicas - Genética.

Orientador:
Prof. Dr. Ney Lemke

Mestrado em Ciências Biológicas - Genética
Departamento de Física e Biofísica
Instituto de Biociências
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Campus de Botucatu

Botucatu – SP

2012

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO DE AQUIS. E TRAT. DA INFORMAÇÃO
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ***ROSEMEIRE APARECIDA VICENTE***

Costa, Pedro Rafael.
Modelo cinético estocástico para a transcrição considerando colisões entre as moléculas de RNA polimerase / Pedro Rafael Costa. – Botucatu : [s.n.], 2012

Dissertação (mestrado) – Universidade Estadual Paulista, Instituto de Biociências de Botucatu
Orientador: Ney Lemke
Capes: 20202008

1. Análise estocástica. 2. Ácido ribonucleico. 3. Genética – Expressão. 4. Células – Mecanismo de controle.

Palavras-chave: Alongamento transcricional; Colisões entre moléculas; Modelo cinético estocástico; Pausas transcricionais; Transcrição múltipla.

Para meus pais, Cláudio e Rose.

Agradecimentos

Meus mais sinceros agradecimentos à todos que me ajudaram na elaboração desse trabalho, em especial:

– Ao Professor Doutor Ney Lemke, pela 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 ao colega Doutor Marcio Luis Acencio, pelos momentos de discussão e descontração;

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

– À minha namorada, Gabriela Vitti Stenico, por estar sempre por perto para dividir as alegrias e tristezas da pós-graduação;

– À Fundação de Amparo à Pesquisa do Estado de São Paulo, pelos dois anos de apoio financeiro.

“ ‘There’s no use trying,’ she said ‘one can’t believe impossible things.’
‘I daresay you haven’t had much practice,’ said the Queen.”
Lewis Carroll, Through the Looking-Glass.

Resumo

A transcrição realizada pela RNA polimerase (RNAP) é um processo cuidadosamente controlado no desenvolvimento e na manutenção das funções vitais dos organismos. O desenvolvimento de novas técnicas e equipamentos para seu estudo, como as técnicas de pinça ótica ou magnética, a microscopia de força atômica e a fluorescência de molécula única, complementaram os resultados dos estudos bioquímicos tradicionais e nos levaram a um maior entendimento do processo. A ocorrência de "pausas" em sítios específicos durante o alongamento, já observadas na década de 80, passou a ser estudada com maior interesse devido a sua importância biológica: acredita-se que essas pausas assegurem que a transcrição e a tradução ocorram simultaneamente em bactérias, permitam o dobramento correto das estruturas secundárias e terciárias do RNA, facilitem a ligação de reguladores de alongamento e precedam a etapa de terminação transcricional. Modelos teóricos baseados na estabilidade termodinâmica do complexo de alongamento transcricional foram bem sucedidos na previsão da cinética do alongamento. Seus resultados indicaram que a RNAP pode ser vista como um motor molecular e sua motilidade possui características do modelo de catraca browniana. Entretanto, esses modelos consideram a presença de apenas uma polimerase realizando a transcrição. Experimentos recentes mostraram que a ocorrência de colisões entre essas enzimas durante a transcrição múltipla de um mesmo gene altera seu comportamento. Baseados nesses resultados, propomos a generalização de um dos modelos estocásticos que consideram a sequência molde para o estudo desse fenômeno. Em nossa aproximação, colisões entre as moléculas RNAP modificam a taxa de ocorrência da transcrição. A implementação do modelo foi realizada em ambiente Mathematica e é baseada no algoritmo de Gillespie. Realizamos simulações para transcrição única e para transcrição múltipla, e comparamos os resultados entre elas e entre o modelo original, além de confrontá-los com os resultados experimentais. Nossa aproximação para transcrição múltipla recuperou os resultados experimentais e obteve um poder de predição de pausas superior. Além disso, os resultados mostraram que a colisão entre as enzimas colabora no aumento de velocidade de trancrição: obtivemos um aumento médio de 48% para as sequências estudadas.

Palavras-chave: Modelo Cinético Estocástico, Alongamento Transcricional, Transcrição Múltipla, Pausas Transcricionais, Colisões entre moléculas.

Abstract

The transcription of the information encoded within the DNA to the RNA molecule is exquisitely controlled during the development of the organisms and to its vital functions and has as the protagonist the RNA polymerase enzyme (RNAP). The development of single-molecule techniques, such as the magnetic and optical tweezers, atomic-force microscopy and single-molecule fluorescence, increased our understanding of the process, complementing traditional biochemical studies. The non-homogeneity of the RNAP movement due to the occurrence of "pauses" at specific sites during elongation was revealed using electrophoresis gels. It is believed that these pauses ensure concurrency between transcription and translation in bacteria, allow the correct folding of RNA secondary and tertiary structures, facilitate the binding of regulating factors during elongation and preceding the transcriptional termination step. Theoretical models have been proposed to explain and predict the RNAP kinetics during the polymerization. Models based on the thermodynamic stability of the transcription elongation complex recover much of the kinetics and indicate that its movement has a Brownian ratchet mechanism. However, experiments showed that if more than one RNAP molecule initiate from the same promoter, their behavior changes and new phenomenona are observed. We proposed and implemented a theoretical model that considers collisions between RNAP molecules and predicts their cooperative behavior during multi-round transcription. The model generalizes a stochastic sequence-dependent model. In our approach, collisions between elongating enzymes modify their transcription rate values. We performed the simulations in Mathematica and compared the results of the single and the multiple-molecule transcription with experimental results and other theoretical models. Our multi-round approach could recover several expected behaviors, and achieved a better predictive power, than the single-round one. Our findings show that the collisions between the enzymes collaborate to the transcription, carrying it out 48% faster on average for the studied sequences.

Key Words: Multi-round transcription, RNA polymerase, Transcription Elongation, Stochastic Kinetic Model, Molecules collision.

Sumário

Lista de Figuras

Lista de Tabelas

Prefácio

Os conceitos biológicos, bioquímicos, termodinâmicos e computacionais aplicados neste trabalho estão apresentados no primeiro Capítulo, no esforço de tornar esta dissertação auto-contida. Ela segue os passos necessários para a criação de um modelo matemático a partir de um fenômeno biológico, apresentando o problema e quais as ferramentas serão utilizadas para resolvê-lo. Os tópicos abordados são os seguintes:

1. **A vida como resultado da expressão gênica:** Conceitos pertinentes à biologia molecular, sobretudo ao processo de transcrição do DNA. Outros temas são tratados superficialmente pois, apesar de sua importância, não são alvos desse trabalho.

2. **Simulações estocásticas:** Definição de estocasticidade e apresentação dos dois métodos de Gillespie para simulação estocástica de reações químicas.

3. **Cinética enzimática:** Conceitos físicos e químicos utilizados nos modelos implementados foram discutidos aqui. É dada ênfase na aproximação de Michaelis-Menten e cálculo da Energia Livre de Gibbs dos ácidos nucléicos.

No Capítulo Estado da Arte, o modelo cinético estocástico dependente da sequência de Bai *et al.* é apresentado em detalhes. Em seguida, no Capítulo 3, são apresentados os escopos desse trabalho. Os métodos empregados na implementação do modelo, com ênfase no algoritmo produzido, são apresentados no Capítulo 4. A apresentação dos resultados obtidos e sua discussão estão unidos no Capítulo 5. Finalmente, apresentam-se as conclusões do trabalho. Anexo à dissertação estão as sequências utilizadas nas simulações e os eventos que o aluno participou com suas respectivas produções, além da versão inicial do artigo que será submetido à revista *PLoS Computational Biology* (Conceito Qualis-CAPES Ciências Biológicas I: A1; Fator de Impacto: 5.76).

1 Introdução

1.1 A vida como resultado da expressão gênica

Como diferenciar os seres vivos de outras formas de organização da matéria? Seres vivos são sistemas abertos, capazes de reduzir sua entropia interna através de substâncias externas e da energia livre tomada do ambiente, posteriormente liberadas em uma forma degradada, ou seja, são capazes de manter sua homeostase e sua organização interna através de seu metabolismo e da capacidade de resposta aos estímulos externos, aumentando a entropia do ambiente à sua volta (SCHRÖDINGER, 1944). A unidade básica da vida é a célula, formada por diferentes estruturas responsáveis por manter essas condições. Os componentes químicos mais importantes do ponto de vista estrutural da célula são as *proteínas*, polímeros de moléculas de aminoácidos alfa unidos através de ligações peptídicas.

Toda a informação necessária para a homeostasia fica armazenada em polímeros com capacidade de replicação. Esses polímeros sofrem alterações que garantem a diversidade entre os seres vivos mesmo dentro de uma mesma espécie, e, devido às pressões ambientais, essa diversidade pode gerar uma herdabilidade diferencial. Os polímeros responsáveis pelo armazenamento das informações são os compostos orgânicos de ácidos desoxirribonucléicos (*ADN*, ou *DNA*) e de ácidos ribonucléicos (*ARN* ou *RNA*). Ambas moléculas possuem *nucleotídeos* como unidade básica, compostos formados por uma base nitrogenada, uma pentose e um grupo fosfato, unidos através ligações fosfodiéster.

1.1.1 O Dogma Central da Biologia Molecular

O Dogma Central da Biologia Molecular, articulado em 1958 por Francis Crick e representado na Figura 1, apresenta um quadro de como ocorre a transferência de informação entre os biopolímeros citados: a informação presente no DNA pode ser copiada para uma nova molécula de DNA (*replicação*), ou para uma molécula de RNA (*transcrição*). Por sua vez, a informação presente no RNA pode ser lida e convertida em proteína (*tradução*).

Em 1970, Crick viu-se obrigado a atualizar seu Dogma, pois estudos mostraram que em condições especiais pode-se observar 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, 1970).

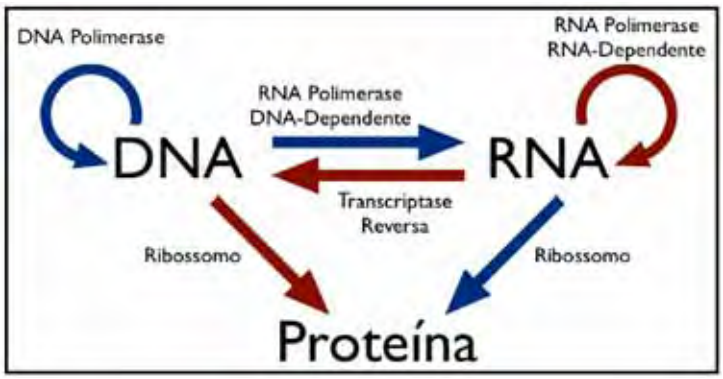

Figura 1: Dogma central da biologia molecular. Cada seta representa uma reação, com sua respectiva enzima indicada. Em azul, as transferências usuais; em vermelho, reações que podem ocorrer em condições especiais (CRICK, 1970).

Este trabalho enfatiza a etapa de *transcrição* do DNA. Para entrarmos em detalhes a respeito do processo em si, é necessário estudarmos antes os constituintes dessa reação: *DNA*, *RNA* e *RNA polimerase*.

1.1.2 A molécula de DNA

O médico suíço Friedrich Miescher foi o primeiro pesquisador a constatar a presença de uma molécula no núcleo celular composta por hidrogênio, carbono, oxigênio, nitrogênio e fósforo, e denominou-a de *nucleína*. Seu trabalho foi publicado em 1871 e visto com descrença pela comunidade científica (DAHM, 2005). Após a confirmação da presença do composto e de seu caráter ácido por Richard Altmann, que sugeriu a denomiação *ácido nucléico*, além da posterior determinação de sua composição química, que determinou que o açúcar presente nessa molécula é a desoxirribose (KARP, 2005), Frederick Griffith conduziu um importante experimento com cepas de bactérias *Streptococcus pneumoniae* em 1928. Ele mostrou que destroços de bactérias mortas de uma determinada cepa eram capazes de *transformar* as células bacterianas vivas da outra, que passavam a possuir as características da primeira (GRIFFITH, 1928). Em 1944, Oswald Avery e seus colegas conduziram experimentos para determinar qual era o "agente transformador" do experimento de Griffith. Os resultados indicaram fortemente que o DNA, e não as proteínas, como se esperava até então, carregavam o material genético (AVERY; MACLEOD; MCCARTY, 1944). A Experiência de Hershey-Chase confirmou o fato irrefutavelmente em 1952 (HERSHEY;

CHASE, 1952). No ano seguinte, James Watson, Francis Crick e Maurice Wilkins, com base no trabalho de Rosalind Franklin, determinaram a estrutura química da molécula de DNA. Segundo o modelo por eles apresentado, o DNA organiza-se em duas fitas anti-paralelas (dupla-hélice), onde as bases dos nucleotídeos emparelham-se com as respectivas bases complementares: Adenina (A) com Timina (T) e Citosina (C) com Guanina (G), como pode ser visto na Figura 2.

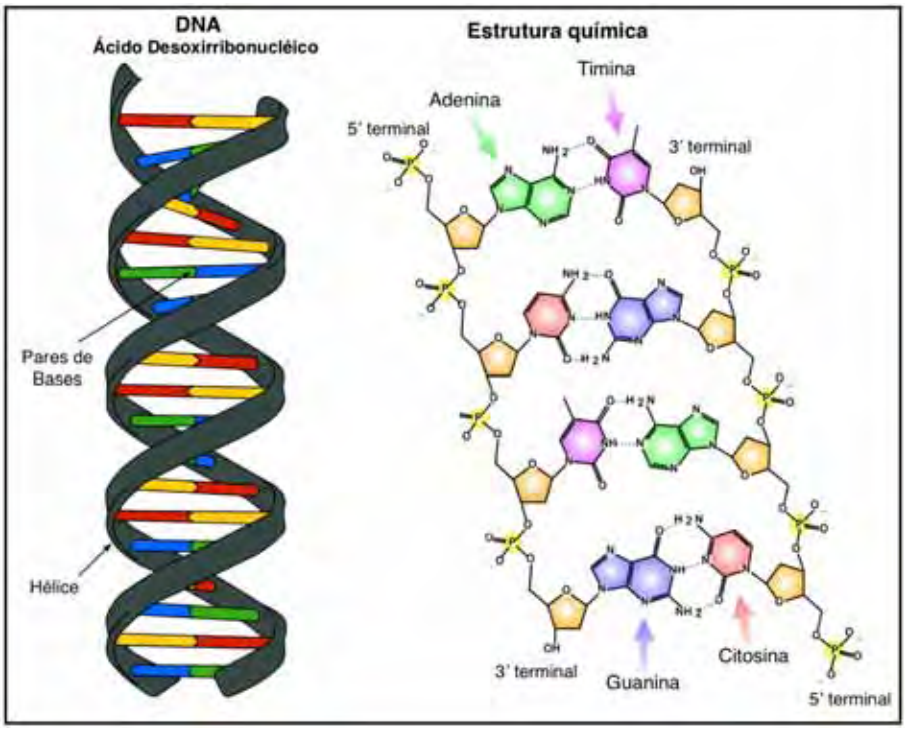

Figura 2: Gravura representando a estrutura do DNA. O DNA é formado por duas fitas complementares seguindo o pareamento Watson-Crick: A-T forma duas pontes de hidrogênio, enquanto C-G forma três ligações. Desta forma a interação entre CG é mais forte que AT. Observando sua estrutura química, podemos notar 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 $5'$, estão presentes em extremidades opostas. Toma-se essa referência quando queremos citar uma determinada sequência de nucleotídos: $5'-\mathrm{AATTCCGG}-3'$, por exemplo. Fonte:Wikimedia Commons.

Denomina-se *genoma* o conjunto das informações hereditárias presentes nos organismos, resultado da sequência das bases constituintes de seu material genético. Qualquer região localizável no genoma que possa ser transcrita e associada à pelo menos uma região regulatória é chamada de *gene* (PEARSON, 2006). Como exemplos de regiões regulatórias, podemos citar os *promotores*, regiões que atuam como sítio de ligação da RNAP, e os *acentuassomos*, sequências nas quais proteínas podem se ligar e aumentar os níveis de transcrição (GRIFFITHS et al., 2006).

1.1.3 A molécula de RNA

Diferentemente do DNA, o RNA apresenta uma estrutura de fita única, capaz de se dobrar em estruturas terciárias devido a complementaridade entre suas bases. As bases nitrogenadas presentes são as mesmas do DNA, com exceção da Timina, substituída pela Uracila (U). A pentose presente em sua composição é denominada *ribose*. Sua estrutura pode ser vista na Figura 3.

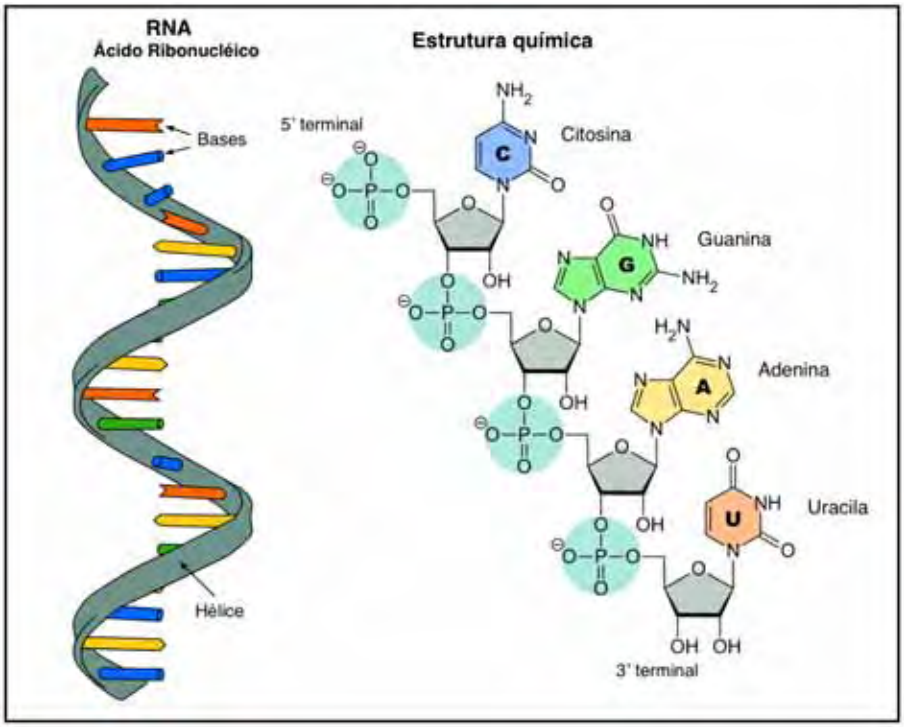

Figura 3: Estrutura do RNA. O filamento de RNA é bastante semelhante ao DNA, entretanto trata-se de uma fita única, com a base Uracila substituindo a base Timina. As ligações açúcar-fosfato ocorrem nas posições 5′ e 3′ do açúcar, como no DNA. Assim, uma cadeia de RNA possui uma extremidade 5′ e outra 3′. Fonte: Wikimedia Commons.

O estudo das moléculas de RNA iniciou-se juntamente com o estudo do DNA. Trata-se de uma molécula bastante versátil, presente em importantes reações biológicas. Os tipos clássicos de RNA são os seguintes:

1. **RNA mensageiro (mRNA):** Resultado da transcrição de um gene, carrega a informação necessária para a produção de proteínas.

2. **RNA transportador (tRNA):** Responsável pelo transporte das moléculas de aminoácidos até os ribossomos, onde serão unidos para a formação de proteínas.

3. **RNA ribossômico (rRNA):** São os principais componentes dos ribossomos, macromoleculas que guiam a montagem da cadeia de aminoácidos.

Podemos citar ainda os pequenos RNAs nucleares (snRNA) e nucleolares (snoRNA), envolvidos na modificação pós-transcricional do mRNA, pequenos RNAs de interferência (siRNA) e os microRNAs (miRNA), capazes de regular a transcrição dos genes, além do RNA genômico viral. As moléculas de RNA que possuem função catalítica são chamadas de *ribozimas* (GRIFFITHS et al., 2006).

1.1.4 A enzima RNA polimerase

Assim que o Dogma Central da Biologia Molecular foi estabelecido, uma busca pela maquinaria responsável pelos fenômenos foi iniciada. Então, no final dos anos 50 e início dos anos 60, a atividade da RNA polimerase DNA-dependente foi identificada (BURGESS, 1971). A RNA polimerase (RNAP, Figura 4) é a enzima responsável pelo processo conhecido como *transcrição*. Essa enzima é considerada um *motor molecular* (SCHLIWA, 2003), capaz de converter energia química em movimento, como a miosina e a cinesina. Quanto à velocidade de seu deslocamento, a RNAP chega a ser 10 vezes mais lenta que os demais motores moleculares, mas sua capacidade de carga é consideravelmente maior, chegando a 25 pN contra 5 pN dos demais motores (WANG et al., 1998).

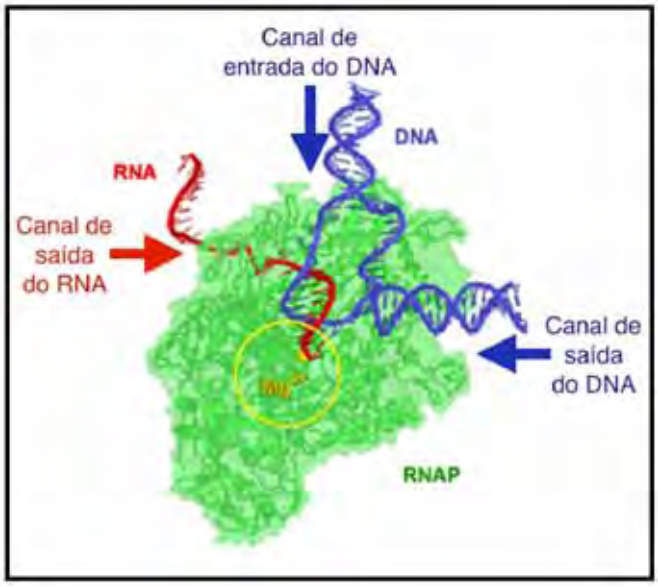

Figura 4: Representação da RNAP. Podemos observar o canal principal, onde o híbrido RNA/DNA se localiza durante a transcrição, o canal de saída do RNA e os canais de entrada e saída do DNA. Em destaque o sítio ativo da enzima, onde o íon Mg^{2+} está presente. Fonte: Wikimedia Commons.

Em procariotos, encontramos apenas uma variante da RNAP. Já nos eucariotos, existem diferentes variantes dessa molécula, classificadas de acordo com o tipo de RNA que sintetizam. Além disso, existem outros tipos de RNAP presentes apenas nas mitocôndrias ou nos cloroplastos por exemplo. As três comuns a todos os eucariotos são:

- RNAPI: responsável pela síntese de um pré-RNA ribossômico (rRNA) que quando maduro forma a maior região de RNA presente no ribossomo;
- RNAPII: sintetiza os precursores do RNA mensageiro (mRNA), além de grande parte dos pequenos RNAs nucleares (snRNA) e dos microRNAs.
- RNAPIII: síntese de RNAs transportadores (tRNA), responsáveis pelo transporte dos aminoácidos corretos até os ribossomos, e de outros pequenos RNAs.

O pesquisador e prêmio Nobel, Jacques Monod, considerado um dos fundadores da Biologia Molecular, afirmou em 1954: "O que for dado como verdadeiro para a bactéria *Eschericia Coli* deverá ser verdadeiro para elefantes" (FRIEDMANN, 2004)). Apesar das diferenças estruturais e da ausência de parentesco entre as RNAP de eucariotos e de procariotos, ensaios bioquímicos e dados estruturais mostram que as diferentes famílias compartilham várias características, o que justifica a busca por mecanismos enzimáticos comuns (SOUZA, 1996; CRAMER, 2002). Por exemplo, todas as polimerases convergiram para o mesmo mecanismo químico de catálise, promovido pelos dois íons de Mg^{2+} localizados em seu sítio ativo (STEITZ, 1998, 1999).

1.1.5 O processo de transcrição

A primeira etapa do Dogma Central da Biologia Molecular trata da transferência das informações contidas na molécula de DNA para a molécula de RNA. Esse processo é conhecido como *transcrição* e possui a RNAP como enzima responsável pela sua execução. Durante esse processo, ela forma juntamente com a fita-dupla de DNA e a fita de RNA nascente o *Complexo de Alongamento Transcricional* (CAT). A transcrição deve ser controlada cuidadosamente durante o desenvolvimento dos seres vivos e para a manutenção de suas funções vitais, pois controla todo o metabolismo celular. Após inciado, o processo pode ser dividido em três fases principais, apresentados na Figura 5: *iniciação*, *alongamento* e *terminação*. 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 nossa compreensão desses fenômenos, complementando os estudos bioquímicos tradicionais (HERBERT; GREENLEAF; BLOCK, 2008).

Para o início da transcrição, é necessária a presença de uma sequência promotora na fita de DNA. A RNAP é capaz de se ligar nessa região na presença de fatores de transcrição, passando então para a fase de *iniciação*. Durante essa fase, a RNAP abre a dupla hélice do DNA e expõe a fita molde que será transcrita, formando uma estrutura

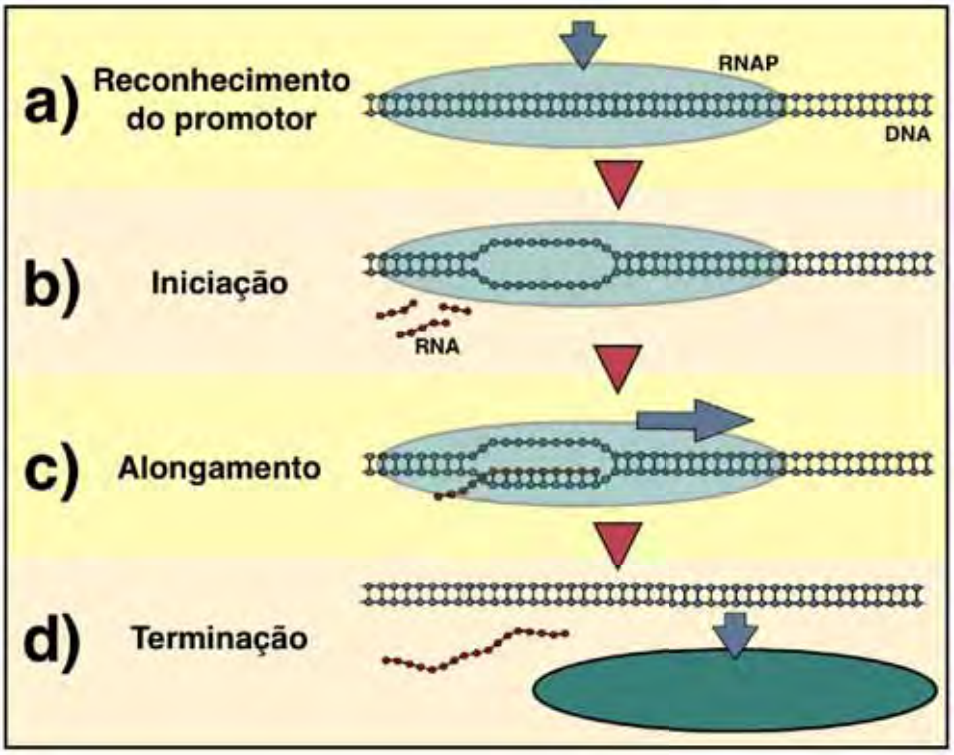

Figura 5: Etapas da transcrição. a) Reconhecimento do promotor: a RNAP liga-se ao DNA dupla fita quando encontra os sítios promotores, expondo a fita de DNA que servirá de modelo. b) Iniciação: produção de pequenos filamentos de RNA durante a iniciação, liberando-os em seguida, mas sem abandonar o promotor. c) Fase de alongamento: cadeias de RNA serão polimerizadas, incorporando-se um ribonucleotídeo complementar ao desoxirribonucleotídeo presente no seu sítio ativo. Após a incorporação, a enzima move-se um nucleotídeo no sentido $5' \rightarrow 3'$ na fita molde e reinicia o processo. d) Terminação: reconhecimento do sítio de terminação, desestruturação do CAT, liberação do transcrito produzido e abandono d a fita de DNA (BUC; STRICK, 2009).

denominada de *bolha de transcrição*. Para abandonar essa região, ocorre um processo conhecido como *iniciação abortiva*, onde pequenas fitas de RNA (2-15 nucleotídeos) são produzidas e liberadas (MCCLURE; CECH; JOHNSTON, 1978). Experimentos mostraram que, durante esse processo, a RNAP acomoda dentro de sua estrutura um nucleotídeo da fita de DNA a cada ligação fosfodiéster formada. Na liberação da RNA abortada, a RNAP expulsa o DNA acumulado, que retorna à conformação original (KAPANIDIS et al., 2006). Esse processo é conhecido com *scrunching*, representado na Figura 6.

Quando a RNAP finalmente abandona a região promotora, a transcrição dos primeiros 8-12 nucleotídeos leva à formação do CAT, constituído pela enzima, pelo DNA e pela fita de RNA nascente. A fase de *alongamento* se inicia quando o CAT desliza ao longo da fita de DNA, incluindo novos ribonucleotídeos à estrutura do RNA durante o processo. Esse novo ribonucleotídeo dependerá do desoxirribonucleotídeo presente no sítio ativo da RNAP e será complementar ao mesmo, seguindo o pareamento de Watson-Crick. O sentido da transcrição é na fita de DNA $3' \rightarrow 5'$ e o resultado será uma fita de RNA $5' \rightarrow 3'$,

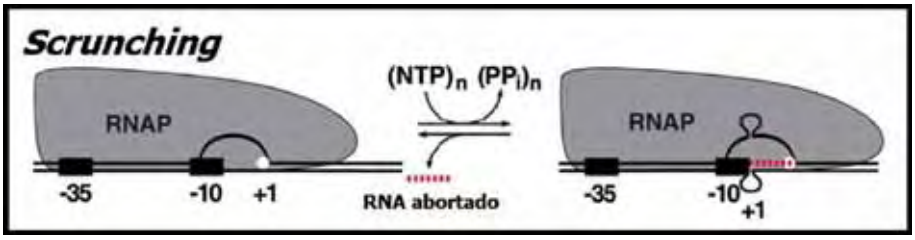

Figura 6: Processo de *scrunching*. A RNAP "puxa" uma base da fita molde para seu interior, realizando a transcrição sem abandonar o promotor. Após certo comprimento, a enzima cliva a fita transcrita, repetindo o processo indeterminadamente, levando à produção de pequenos RNAs (KAPANIDIS et al., 2006). Na figura, temos representado um promotor consenso de *E. coli* (lacCONS), com as caixas pretas mostrando a posiçao de seus elementos -10 e -35. *NTP*: Nucleotídeo Trifosfatado, *PPi*: Pirofosfato.

devido à conformação das ligações entre as bases. Técnicas para determinação da velocidade do alongamento, tais como *run-on assays* (O'BRIEN; LIS, 1993), reação em cadeia de polimerase por transcriptase reversa (RT-PCR) (TENNYSON; KLAMUT; WORTON, 1995) e hibridização com fluorecência *in situ* (FISH) (FEMINO et al., 2003) chegaram a valores que variaram de 1,1 à 2,5 quilobases por minuto, enquanto dados obtidos *in vivo* indicaram 4,3 quilobases por minuto (DARZACQ et al., 2007).

Para explicar o movimento da RNAP na fita de DNA, diversos modelos foram propostos. O modelo da catraca browniana, onde flutuações térmicas entre os estados pré e pós-translocado desta enzima são mecanicamente retificados pela ligação do NTP e pela sua hidrólise, resultando em um movimento unidirecional, é o mais aceito (HERBERT; GREENLEAF; BLOCK, 2008).

Processos alternativos que podem ocorrer durante o alongamento estão representados na Figura 7. Pausas durante essa fase são frequentes e podem não somente 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). Pausas breves, com duração menor que 25 segundos e presentes em média a cada 100 pares de bases, são conhecidas como *pausas ubíquas* e representam cerca de 95% das pausas detectadas (NEUMAN et al., 2003). Pausas de maior duração, ditas estáveis, são conhecidas pelo seu papel regulador e geralmente estão ligadas à formação de estruturas terciárias da molécula de RNA, denominadas *grampos*, ou devido ao processo conhecido como *backtracking*, quando a RNAP recua alguns pares de bases. A probabilidade do CAT entrar em um desses processos alternativos depende de interações entre o complexo e sequências específicas do DNA ou do RNA nascente. Fatores de transcrição reguladores também podem ser responsáveis por estes fenômenos (HERBERT et al., 2006).

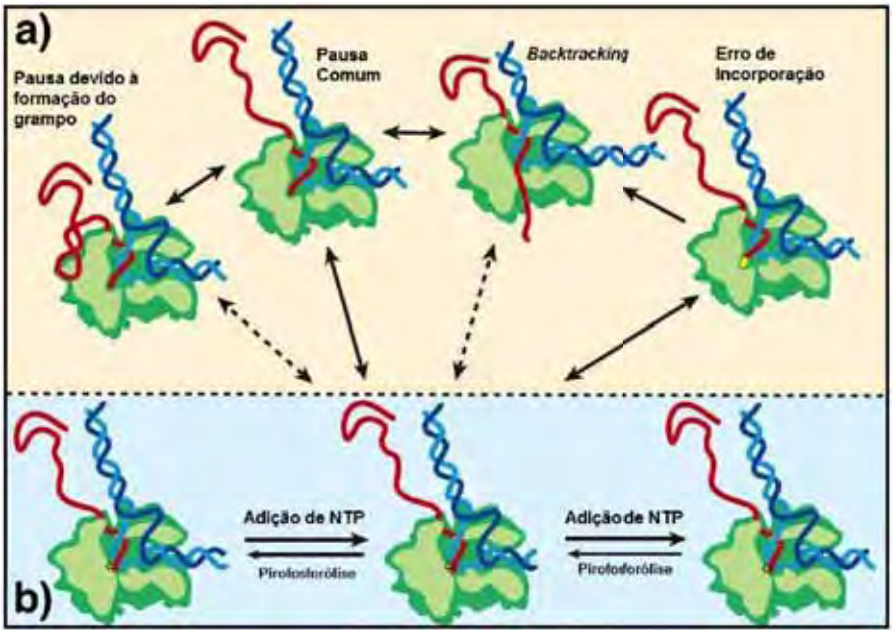

Figura 7: Processos alternativos que podem ocorrer durante o alongamento. Setas contínuas representam processos com maior probabilidade de ocorrência em relação às tracejadas. a) Processos alternativos. b) Alongamento propriamente dito (HERBERT et al., 2006).

Caso não se desprenda da fita de DNA, o CAT seguirá com o alongamento até atingir o *sítio de terminação* da fita molde. Durante a *terminação* ocorre o desenrolamento do híbrido RNA-DNA, a liberação da molécula de RNA pela RNAP e o colapso da bolha de transcrição, formando novamente a estrutura estável do DNA em fita dupla (GREIVE; HIPPEL, 2005). Esse processo pode ocorrer de diferentes maneiras, ainda não totalmente esclarecidas. Um dos mecanismos conhecidos ocorre numa região denominada *sítio de terminação intrínseca*, que após ser transcrita gera uma estrutura em forma de grampo, seguida por uma sequência de aproximadamente sete uridinas, formando um híbrido DNA-RNA particularmente fraco com as desoxiadenosinas complementares (MARTIN; TINOCO, 1980). O CAT é induzido a pausar nessa região e impedido de recuar devido a interações que envolvem o grampo de RNA formado e a polimerase. O tempo de parada do CAT nesse sítio aumenta dessa forma, permitindo que o completo dobramento do grampo e as alterações estruturais na RNAP desestabilizem a bolha de transcrição (HERBERT; GREENLEAF; BLOCK, 2008).

1.1.6 Transcrição múltipla

Não é raro os fatores de transcrição continuarem ligados a região promotora, recrutanto novas moléculas de RNAP para a transcrição dessa mesma região. Denominaremos esse fenômeno de *transcrição múltipla*. A bactéria *Eschericia coli*, por exemplo, durante

seu crescimento exponencial, possui até 13.000 moléculas de RNAP transcrevendo ativamente, concentradas em apenas 90 unidades transcricionais (GRIGOROVA et al., 2006). Genes ribossomais são conhecidos por serem altamente transcritos e famosos pela sua organição em forma de "árvore de Natal". Na micrografia presente na Figura 8, podemos notar o motivo da comparação: a fita molde de DNA seria o "tronco" da árvore enquanto os RNAs nascentes de comprimento crescente seriam os "galhos" desta árvore. Nestas micrografias, podemos ver até centenas de galhos, indicando uma alta concentração de RNAPs nesses genes (FOE, 1977; HAMMING et al., 1981).

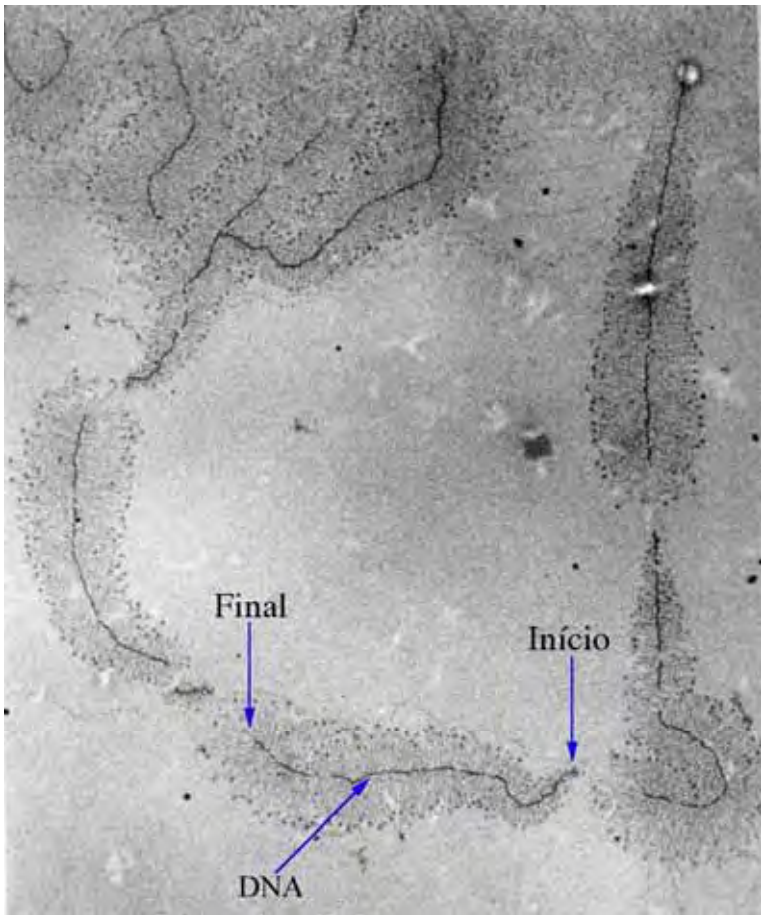

Figura 8: Micrografia mostrando a transcrição múltipla em genes de RNA ribossômico. Note sua estrutura em forma de "árvore de Natal". Fonte: Wikimedia Commons.

Existem vários artigos mostrando que o comportamento transcricional é alterado nesses casos. Epshtein and Nudler (EPSHTEIN; NUDLER, 2003) mostraram que caso exista um sítio de pausa muito intenso na fita molde, a transcrição completa dessa sequência será eficiente apenas na presença de múltiplas moléculas de RNAP. Em outro trabalho do mesmo laboratório (EPSHTEIN et al., 2003), foi mostrado que na presença de obstáculos protéicos ligados à fita de DNA, como ocorre *in vivo*, a transcrição dessa região dependerá da transcrição múltipla: as moléculas de RNAP removem os obstáculos presentes na

fita quando se chocam com eles, mas a intensidade desse choque dependerá do número de enzimas participantes. O modelo para tráfego intenso de RNAPs apresentado por Klumpp (KLUMPP, 2011) mostrou que a ocorrência de *backtracking* fica restrita pois as enzimas acabam funcionando como obstáculos para fenômeno. Rajala e colegas (RAJALA et al., 2010) concluíram através de seu modelo que a presença de múltiplas enzimas e de pausas na sequência molde acarreta em pequenos surtos de produção de RNA.

1.2 Simulações estocásticas

1.2.1 Reações químicas

Reações químicas são a linguagem canônica para modelos biológicos. Expressam processos químicos complexos tanto qualitativamente como quantitativamente. Especificá-las corretamente é fundamental para o sucesso do modelo. Genericamente,

$$n_a A + n_b B \xrightarrow{k} n_c C + n_d D$$

representa que n_a moléculas do tipo A reagem com n_b do tipo B para produzirem n_c moléculas do tipo C e n_d de D. A letra k acima da seta, que indica o sentido da reação, representa a taxa de ocorrência da mesma, sendo equivalente à probabilidade da colisão entre n_a moléculas de A com n_b moléculas de B ocorrer com energia suficiente para que reajam entre si e gerem seus produtos. Utiliza-se a notação $[A]$, por exemplo, para se referir a concentração de moléculas de A durante uma reação.

1.2.2 Conceito de estado

O estado de um determinado sistema mostra sua condição para um instante determinado e contém informações suficientes para prever seu comportamento no futuro. Intuitivamente, trata-se do conjunto de variáveis que precisam ser acompanhadas em um determinado modelo. Na modelagem estocástica, o estado depende de uma lista com número atual de moléculas de cada tipo, chamada de *configuração do sistema*, podendo ser a própria configuração, ou a distribuição de probabilidades atual da configuração.

Por exemplo, 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, essa mudança de estado ocorre da seguinte forma:

uma única molécula de A é convertida numa molécula de A'; 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.

1.2.3 O algoritmo de Gillespie

Nesse algoritmo, para cada período de tempo o sistema está em um único estado exato. Uma transição consiste em executar uma reação r, sendo n o número máximo de possíveis transições a partir de um estado. Utilizando-se geradores de números aleatórios com a distribuição de probabilidade correta, o algoritmo determina qual reação será executada.

Para um sistema em um determinado estado, o método direto retorna qual será a próxima reação a ocorrer e quando ela irá ocorrer. Estas questões podem ser respondidas probabilisticamente especificando a probabilidade de densidade $P(r,t)$ para a próxima reação r que ocorre no tempo t (GIBSON; BRUCK, 2000). Pode ser mostrado que

$$P(r,t)dt = a_r \mathrm{Exp}\left(-t\sum_{j=1}^{n} a_j\right)dt$$

onde a_r é proporcional a probabilidade de uma certa reação r ocorrer. O Algoritmo 1 apresenta os passos para calcular qual será a próxima reação e quando ela irá ocorrer usando as probabilidades de densidade das reações. O método da primeira reação, apresentado no Algoritmo 2, calcula um tempo estimado t_i para todas as reações possíveis e seleciona qual é o menor dentre eles, atribuindo a r essa reação e a t o seu tempo.

Algoritmo 1: Método Direto

1. Inicialize, determine o tempo (T) e número de moléculas de cada composto inicial.
2. Calcule a função probabilidade de densidade a_i, para todas as reações.
3. Escolha r de acordo com a distribuição:

$$P_r(r) = \frac{a_r}{\sum_{j=1}^{n} a_j}$$

4. Escolha t de acordo com uma distribuição exponencial:

$$P(t) = \left(\sum_{j=1}^{n} a_j\right)\mathrm{Exp}\left(-t\sum_{j=1}^{n} a_j\right)dt$$

5. Altere o número de moléculas para refletir a execução da reação r e o tempo para $T = T + t$.

6. Volte para o passo 2.

Algoritmo 2: Método da primeira reação

1. Inicialize, determine o tempo (T) e número de moléculas de cada composto inicial.

2. Calcule a função probabilidade de densidade a_i para todas as reações

3. Calcule o tempo estimado t_i para cada reação de acordo com uma distribuição exponencial com parâmetro a_i:

$$t_i = \frac{1}{a_i} \log\left(\frac{1}{\text{rand}}\right)$$

   onde *rand* é um número aleatório determinado a partir de uma distribuição uniforme.

4. Atribua a r a reação com o menor t_i.

5. Atribua a t o valor de t_r.

1.3 Cinética enzimática

A cinética enzimática estuda as reações químicas catalizadas pelas *enzimas*, em especial o estudo das taxas de ocorrência dessas reações.

As enzimas geralmente são proteínas capazes de manipular outras moléculas, denominadas *substratos*. Essa manipulação leva a *catálise* da reação química entre seus substratos, ou seja, aumenta sua velocidade de ocorrência, sem o consumo da enzima no processo. O resultado da interação da enzima com seus substratos é chamado de *produto*. Grande parte das reações que ocorrem em ambiente celular possui uma enzima responsável para tornar sua velocidade compatível com a vida. Basicamente, o conjunto de enzimas disponíveis numa célula determina seu metabolismo. Cerca de 3700 enzimas já foram caracterizadas (BAIROCH, 2000).

A cinética enzimática pode ser afetada de várias formas. As chamadas moléculas *inibidoras* diminuem a ativadade enzimática, enquanto as moléculas *ativadoras* são capazes de aumentá-la. Temperatura, pressão, pH e concentração do substrato também são fatores que interferem da velocidade da reação.

Toda reação química é determinada pelas leis da Termodinâmica. Tratamos uma reação como *espontânea* quando não há necessidade de absorção de energia para sua ocorrência ou como *não espontânea*, quando sua ocorrência depende da absorção de energia na forma de calor, luz ou eletricidade, por exemplo. Essa energia é chamada de *energia de ativação*.

A catalização de uma reação leva à diminuição de sua energia de ativação, como podemos ver na Figura 9. Como a energia necessária para a ocorrência da reação diminui, a catalização leva a sua aceleração. Algumas enzimas chegam a acelerar milhões de vezes uma reação (RADZICKA; WOLFENDEN, 1995).

Figura 9: Gráfico representativo do caminho de uma reação. A energia de ativação representa a barreira energética que necessita ser ultrapassada para sua ocorrência. Note a alteração de seu valor na presença de um catalisador. Fonte: Wikimedia Commons.

1.3.1 Cinética de Michaelis-Menten

Em 1913, o bioquímico alemão Leonor Michaelis e o físico canadense Maud Menten propuseram um modelo para a reação de hidrólise da sacarose em glicose e frutose, catalisada pela enzima *invertase* (MENTEN; MICHAELIS, 1913). Neste modelo, a enzima E liga-se ao substrato S e forma um complexo ES, convertido num produto P mais a enzima. Podemos representar esta reação da seguinte forma:

$$E + S \underset{k_{-1}}{\overset{k_1}{\leftrightarrows}} ES \xrightarrow{k_2} E + P$$

onde k_1, k_{-1} e k_2 representam as taxas das reações e as setas representam seu sentido, indicando que a ligação do substrato à enzima é um processo reversível. Se assumirmos que a concentração enzimática é muito menor que a concentração de substrato, que a variação

da concentração do complexo ES pode ser tomada como zero e que a concentração total da enzima não varia com o tempo, a taxa de formação do produto, ou velocidade global da reação, é dada por

$$v = \frac{d[P]}{dt} = \frac{V_{\max}[S]}{K_m + [S]}$$

Logo, a taxa dessa reação aumenta assintoticamente com o aumento da concentração do substrato S até o valor máximo de $V_{\max}$, obtido quando todo o substrato estiver ligado a uma enzima. A constante de Michaelis-Menten, K_m, é dada pelo inverso da afinidade do substrato com a enzima, ou seja,

$$K_m = \frac{k_{-1} + k_2}{k_1}$$

Seu valor representa a concentração de substrato necessária para que a reação ocorra a uma velocidade igual a metade da velocidade máxima (LENINGER, 2006). Os valores de K_m e de $V_{\max}$ podem ser determinados a partir de uma série de ensaios enzimáticos, realizados variando-se a concentração dos substratos.

1.3.2 Reação enzimática na presença de inibidor

Existem dois tipos de reações enzimáticas na presença de inibidor: as *reversíveis* e as *irreversíveis* (LENINGER, 2006). No último caso, o inibidor possui uma constante de dissociação tão baixa que inativa a enzima permanentemente. Os inibidores reversíveis podem ser classificados em 4 tipos:

1. **Competitivo:** Neste caso, o inibidor e o substrato não podem se ligar simultaneamente na mesma enzima, competindo pelo seu sítio ativo. Este tipo de inibidor não altera o valor de $V_{\max}$, mas aumenta o valor de K_m aparente, K_m^{app}.

2. **Incompetitivo:** Quando o inibidor se liga apenas ao complexo enzima-substrato. O valor de $V_{\max}$ diminui, assim como o K_m^{app}.

3. **Misto:** O inibidor misto pode se ligar tanto a enzima como ao complexo enzima-substrato. Com isso, o valor de $V_{\max}$ diminui e o valor de K_m aparente aumenta.

4. **Não competitivo:** A inibição não competitiva é uma forma de inibição mista. Neste caso, se o inibidor estiver ligado à enzima, sua atividade decai, mas sem afetar a taxa de ligação do substrato a mesma. Como resultado, o valor de $V_{\max}$ diminui enquanto o K_m^{app} se mantém.

2 Estado da Arte

2.1 Modelo cinético estocástico para o alongamento transcricional

Utilizando os conceitos apresentados, Bai *et al.* (BAI; SHUNDROVSKY; WANG, 2004) criaram um modelo para o alongamento transcricional, combinando a essência dos trabalhos de Yager e von Hippel (YAGER; VONHIPPEL, 1991) e de Guajardo e Souza (GUAJARDO; SOUSA, 1997). Denominaremos este modelo de *Modelo B*. O *Modelo B* é baseado no modelo de catraca termal para motores moleculares e considera cada passo do alongamento como uma reação em três etapas:

$$\begin{aligned} \mathrm{CAT}_{(\mathrm{n,pré})} &\overset{K_\mathrm{i}(\mathrm{n})}{\longleftrightarrow} \mathrm{CAT}_{(\mathrm{n,pós})} \\ \mathrm{CAT}_{(\mathrm{n,pós})} + \mathrm{NTP} &\overset{K_\mathrm{d}^\mathrm{NTP}}{\longleftrightarrow} \mathrm{CAT}_{(\mathrm{n,pós})} \times \mathrm{NTP} \\ \mathrm{CAT}_{(\mathrm{n,pós})} \times \mathrm{NTP} &\overset{k_\mathrm{max}^\mathrm{NTP}}{\longrightarrow} \mathrm{CAT}_{(\mathrm{n+1,pré})} + \mathrm{PPi} \end{aligned} \tag{2.1}$$

Esta equação inclui o deslocamento do CAT, a ligação de um novo nucleotídeo (NTP) e sua catálise química. Os estados *pré* e *pós* para o CAT representam a posição do RNA nascente em relação ao sítio ativo da enzima, enquanto n é o comprimento desse RNA. A Figura 10 ilustra os estados possíveis para CAT, além de seus componentes e estrutura. $K_\mathrm{d}^\mathrm{NTP}$ e $k_\mathrm{max}^\mathrm{NTP}$ possuem valores experimentalmente estabelecidos e dependem do nucleotídeo que será incorporado.

O valor para $K_\mathrm{i}(\mathrm{n})$ é dado por

$$K_\mathrm{i}(\mathrm{n}) = \exp[(\Delta G_{(n,\mathrm{pós})} - \Delta G_{(n,\mathrm{pré})} - Fd)/k_B T] \tag{2.2}$$

onde $\Delta G_{(n,m)}$ é a energia livre de Gibbs para determinada conformação do CAT e F representa uma força externa aplicada na RNAP. A distância entre dois pares de base na fita de DNA, representada aqui por d, é de aproximadamente 0.34 nm (BAI; FULBRIGHT; WANG, 2007).

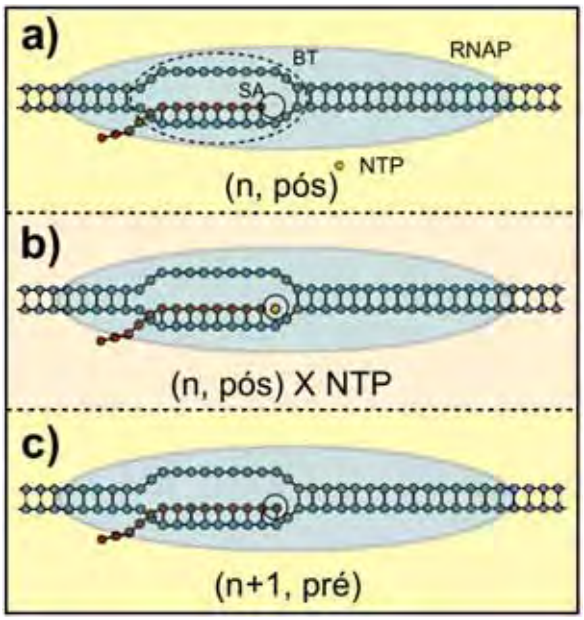

Figura 10: Esquema dos estados e da estrutura do CAT. *SA*: Sítio Ativo, *BT*: Bolha de Transcrição. O CAT será dado pelo comprimento n do RNA e pela posição de seu sítio ativo em relação a extremidade 5' do RNA nascente, m. É equivalente usar $m = 0$ ou $m =$ pré e $m = 1$ ou $m =$ pós. a) Pós-translocado: o sítio ativo está livre. Aqui podemos ver a estrutura do CAT, formado por um híbrido RNA-DNA de 8 pares de nucleotídeos, uma bolha de transcrição de 14 pares de bases (12 separados mais 1 presente nos limites entre a bolha e o DNA fita dupla), e a RNAP englobando 32 pares de bases 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 nt. O CAT se movimentará um nucletído para frente, retornando para *a*.

A reação presente na Equação 2.1 corresponde à cinética enzimática de Michaelis-Menten na presença de um inibidor competitivo. O esquema típico da reação de Michaelis-Menten é modificado para incluir a ligação do inibidor I na enzima livre:

$$EI + S \underset{k_{-3}}{\overset{k_3}{\longleftrightarrow}} E + S + I \underset{k_{-1}}{\overset{k_1}{\longleftrightarrow}} ES + I \overset{k_2}{\longrightarrow} E + P + I$$

Como o inibidor não se liga ao complexo ES nem ao substrato, o valor aparente da constante de Michaelis-Menten, K_m^{app} , aumenta e passa a ser dado por $K_m^{app} = K_m(1 + [I]/K_i)$, sendo K_i a constante de dissociação do inibidor e $[I]$ sua concentração. Tem-se então como expressão da velocidade da reação:

$$v = \frac{d[P]}{dt} = \frac{V_{\max}[S]}{K_m^{app} + [S]}$$

Logo, realizando as substituições equivalentes, a taxa de ocorrência da reação global da Equação 2.1 é dada por:

$$k_{\text{main}}(\text{n}) = \frac{k_{\max}^{\text{NTP}}\,[\text{NTP}]}{K_{\text{d}}^{\text{NTP}}\{1 + K_{\text{i}}(\text{n})\} + [\text{NTP}]} \tag{2.3}$$

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

$$k_{\mathrm{n,m}\to\mathrm{m}\pm1} = k_0 \exp[(\Delta G^{\dagger} - \Delta G_{(n,m)} + Fd(F))/k_B T] \tag{2.4}$$

onde $k_0 = 1.0 \times 10^9\ s^{-1}$ é um pré-fator constante e $\Delta G^{\dagger} = 41,2\ k_B T$ representa a barreira energética para ocorrência desse fenômeno. Esses valores não foram medidos experimentalmente: foram ajustados por Bai *et al.* a partir de seus resultados.

2.2 Determinação da energia livre de Gibbs, ΔG, para ácidos nucléicos

Os valores da energia livre de Gibbs para os ácidos nucléicos são determinados a partir da temperatura de desnaturação das fitas duplas, T_m, na qual metade das fitas duplas em solução encontram-se separadas. Considere a reação:

$$S_1S_2 \leftrightarrow S_1 + S_2$$

A constante de equilíbrio desta reação será dada por $K = [S_1][S_2]/[S_1S_2]$. Como a relação entre ΔG e K é dada por $\Delta G = -RT \ln K$, onde T é a temperatura e R a constante dos gases ideais, teremos:

$$\Delta G = -RT \ln \frac{[S_1][S_2]}{[S_1S_2]}$$

No caso da reação de desnaturação do DNA, S_1S_2 representa a dupla-fita de DNA e S_1 e S_2 as respectivas fitas separadas. Quando atingida a T_m, a relação $[S_1][S_2]/[S_1S_2]$ será dada por C_T/x, onde C_T é a concentração total de fitas simples na reação, com $x = 4$ para fitas não auto-complementares e com $x = 1$ para fitas auto-complementares. Como

$$\Delta G_T = \Delta H - T\Delta S \tag{2.5}$$

teremos:

$$T_m = \frac{-\Delta G}{R \ln(C_T/x)} = \frac{-\Delta H}{\Delta S - R \ln(C_T/x)} \tag{2.6}$$

Determinando-se os valores experimentais de T_m para diferentes concentrações, podemos ajustar a curva de $1/T_m$ por $\ln C_T/x$ e determinar os valores de ΔH e de ΔS para uma determinada fita dupla de DNA. De maneira semelhante, pode-se determinar esses valores para um híbrido DNA-RNA (SANTALUCIA; ALLAWI; SENEVIRATNE, 1996).

3 Objetivos

1. Implementação do modelo dependente da sequência de Bai *et al.* para alongamento transcricional, atualizando os parâmetros utilizados;

2. Criação de um algoritmo para generalização do modelo, passando a contemplar a transcrição múltipla, incluindo choques entre as RNAP;

3. Comparação dos resultados obtidos a partir das simulações de Monte Carlo com os experimentais e com os obtidos pelo modelo original de Bai *et al.*.

4 Métodos

4.1 Implementação do *Modelo B*

Para a implementação do *Modelo B*, foram criadas funções para cada uma de suas etapas utilizando o software Mathematica. Como parâmetros importantes para sua simulação foram atualizados, denominamos essa implementação de *Aproximação para Transcrição Única*, *ATU*. Para simplificação, utilizaremos k_{back} como nomenclatura equivalente à $k_{\text{n,m}\to\text{m}\pm1}$. Os valores dos parâmetros presentes na Equação 2.1 estão na Tabela 1.

Tabela 1: Valores para os parâmetros da Equação 2.1. Tais valores variam conforme o nucleotídeo em questão e foram determinados experimentalmente por Bai *et al.* (BAI; FULBRIGHT; WANG, 2007).

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

4.1.1 Cálculo das energias livres de Gibbs, ΔG, para o CAT

Os valores de $\Delta G_{(n,m)}$ para o CAT foram descritos primeiramente por Yager e von Hippel (YAGER; VONHIPPEL, 1991) e medem sua estabilidade. Para cada conformação, este valor é obtido através de um somatório de três energias distintas:

$$\Delta G_{(n,m)} = \Delta G_{(n,m;\text{bolhaDNA})} + \Delta G_{(n,m;\text{RNA}-\text{DNA})} + \Delta G_{(n,m;\text{intRNAP})} \tag{4.1}$$

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 nucléicos.

O método dos vizinhos mais próximos é utilizado para determinação das energias livres de Gibbs das sequências de ácidos nucléicos (SANTALUCIA; ALLAWI; SENEVIRATNE,

1996). Esse método baseia-se no fato que as interações entre as diferentes fitas depende não somente das bases que formam as pontes de hidrogênio em si, mas também das bases vizinhas a essa ligação. Nas Tabelas 2 e 3 são apresentados os valores utilizados neste trabalho.

Para determinação da energia livre de uma sequência qualquer, somam-se os valores para cada par, além de outros parâmetros que podem alterar a estabilidade das fitas. Por exemplo, a energia livre de Gibbs liberada na quebra da sequência $5'-\mathrm{CGTTGA}-3'$ a 37 °C será dada por

$$\begin{aligned}
\Delta G(\text{total}) &= \Delta G_{\text{ini}} + \Delta G_{\text{simetria}} + \sum \Delta G_{\text{duplas}} + \Delta G_{\text{ATterminal}} \\
5' - \mathrm{CGTTGA} - 3' &= \Delta G_{\text{ini}} + \Delta G_{\text{simetria}} + \\
3' - \mathrm{GCAACT} - 5' \quad &+ CG + GT + TT + TG + GA + AT_{\text{terminal}} \\
&\quad\; GC \quad CA \quad AA \quad AC \quad CT \\
\Delta G(\text{total}) &= +1,96 + 0 - 2,17 - 1,44 - 1,00 - 1,45 - 1,30 + 0,05 \\
\Delta G(\text{total}) &= -5,35 \text{kcal/mol}
\end{aligned}$$

Note que não há penalidade para simetria nessa sequência, pois $5' \rightarrow 3' \neq 3' \rightarrow 5'$.

Nos cálculos, o terceiro termo da Eq. 4.1 foi tomado como zero para simplificação. Como os parâmetros ajustados para o *Modelo B* consideraram que a transcrição ocorre a 24 °C, utilizamos a Eq. 2.5 para a determinação dos valores de ΔG_{24}.

4.1.2 Taxas de ocorrência das reações

O algoritmo base para simulação da *ATU* considera que a RNAP pode tomar dois caminhos distintos: enzima pode estar em alongamento normal, ou em *backtracking*. Quando em alongamento normal, calcula-se o valor de k_{main} utilizando a Equação 2.3 considerando-se um híbrido RNA-DNA com 9 nt antes do sítio ativo da enzima ser desocupado, e uma bolha de transcrição com 15nt. Determina-se também um valor para k_{back} a partir da Equação 2.4, para a mesma conformação do CAT. Para ambos os casos, $F = 0$ pN. Se a RNAP entrou em *backtracking*, calcula-se apenas um valor, k_{back}, tanto para o movimento no sentido $5' \rightarrow 3'$, como para o movimento no sentido oposto.

Tabela 2: Valores para cálculo da estabilidade da fita dupla de DNA. Dados retirados de (SANTALUCIA; HICKS, 2004).

Sequência	ΔH (kcal/mol)	ΔS (e.u.)
AA/TT	-7,6	-21,3
AT/TA	-7,2	-20,4
TA/AT	-7,2	-21,3
CA/GT	-8,5	-22,7
GT/CA	-8,4	-22,4
CT/GA	-7,8	-21,0
GA/CT	-8,2	-22,2
CG/GC	-10,6	-27,2
GC/CG	-9,8	-24,4
GG/CC	-8,0	-19,9
Iniciação	+0,2	-5,7
Penalidade por AT terminal	+2,2	+6,9
Correção devido a simetria	0,0	-1,4

Tabela 3: Valores para cálculo da estabilidade da fita híbrida de RNA-DNA. Dados retirados de (SUGIMOTO et al., 1995).

Sequência	ΔH (kcal/mol)	ΔS (e.u.)
rAA/dTT	-7,8	-21,9
rAC/dTG	-5,9	-12,3
rAG/dTC	-9,1	-23,5
rAU/dTA	-8,3	-23,9
rCA/dGT	-9,0	-26,1
rCC/dGG	-9,3	-23,2
rCG/dGC	-16,3	-47,1
rCU/dGA	-7,0	-19,7
rGA/dCT	-5,5	-13,5
rGC/dCG	-8,0	-17,1
rGG/dCC	-12,8	-31,9
rGU/dCA	-7,8	-21,6
rUA/dAT	-7,8	-23,2
rUC/dAG	-8,6	-22,9
rUG/dAC	-10,4	-28,4
rUU/dAA	-11,5	-36,4
Iniciação	+1,9	-3,9

4.1.3 Simulações Estocásticas

O movimento de uma única molécula de RNAP pela fita de DNA é intrinsecamente estocástico, isto é, sua evolução temporal é analisável em termos de probabilidades. 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. Nosso modelo será, portanto, constituído de uma série de reações de mudança entre os estados possíveis da polimerase.

Instalamos um pacote para o Mathematica com as funções necessárias para o algoritmo de Gillespie (SHAPIRO et al., 2003). Inserimos, então, as reações apresentadas na Eq 2.1. Utilizando o *método da primeira reação*, armazenamos o tempo e a posição da RNAP na fita de DNA a cada ciclo completado. A cada novo passo, verifica-se se a RNAP 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}. Durante o *backtracking*, a RNAP possui a mesma probabilidade de se movimentar para frente ou para trás, sem alterar o tamanho do transcrito produzido.

Realizamos simulações para as seguintes sequências: deleções D104, D111, D112, D123, D167 e D387 da região inicial do genoma do bacteriófago T7 (LEVIN; CHAMBERLIN, 1987) ([NTP]:10 μM e 25 μM) e seq10 até seq13, presentes no traballho de Tadigotla *et al.* (TADIGOTLA et al., 2006) ([NTP]: 40 μM para seq10, 30 μM para as outras). Todas as simulações foram feitas considerando-se uma temperatura de 24 °C e força externa nula. No total, foram realizadas 4800 simulações para cada sequência.

4.1.4 Funções para simulação do *Modelo B*

As principais funções implementadas no pacote para Mathematica para simulação do *Modelo B* são apresentadas a seguir:

- $\boldsymbol{\Delta H_{dd}(par)}$, $\boldsymbol{\Delta S_{dd}(par)}$, $\boldsymbol{\Delta H_{dr}(par)}$, $\boldsymbol{\Delta S_{dr}(par)}$: Retorna os respectivos valores de entalpia e entropia para o par de nucleotídeos vizinhos. *dd*: DNA-DNA; *dr*: DNA-RNA.

- $\boldsymbol{\Delta G_{dd}(\mathrm{par}, T)}$, $\boldsymbol{\Delta G_{dr}(\mathrm{par}, T)}$: Utiliza as funções anteriores para determinação da energia livre de Gibbs para um determinado par de vizinhos na temperatura T, utilizando a Equação 4.1.

- $\boldsymbol{k_{max}(nt),\ K_d(nt)}$: Retorna os valores de k_{max}^{NTP} e de K_d^{NTP} para o próximo nucleotídeo *nt* que será incluído na fita de RNA.

- $\boldsymbol{P(seq)}$: Cria uma lista com todos os pares necessários para cálculo de ΔG para a sequência de entrada.

- $\boldsymbol{C(seq)}$: Verifica se a sequência *seq* possui AT terminal e/ou simetria, retornando os respectivos valores presentes na Tabela 2.

- $\boldsymbol{R(seq, conc, T, F)}$: Utiliza as funções anteriores para determinar k_{main} utilizando a Equação 2.3 e k_{back} utilizando a Equação 2.4, para uma determinada temperatura T sob ação de uma força externa F. *seq* deverá possuir exatamente 15 nucleotídeos: 14 nucleotídeos da bolha de transcrição no estado pré-translocado e 14 nucleotídeos da bolha de transcrição no estado pós-translocado (ver Figura 10).

- $\boldsymbol{N(seq, conc, T, F)}$: Simula a reação global da Equação 2.1 e a reação de entrada em *backtracking*, escolhendo aquela cujo tempo de ocorrência for menor e retornando esses dados, utilizando as funções anteriormente apresentadas. *seq*: sequência que compõe a bolha de transcrição nos estados pré e pós-translocados; *conc*: concentração (em micromolares) dos nucleotídeos presentes na solução; T: temperatura da reação; F: valor da força externa que estará sendo aplicada na RNAP.

- $\boldsymbol{B(seq, T, F)}$: Simula a RNAP quando em *backtracking*. Neste caso, ela poderá se movimentar no sentido contrário ao *backtracking*, retornando futuramente à transcrição normal, ou continuar a se movimentar no sentido contrário ao da transcrição. Essa função retorna qual ocorreu e quanto tempo foi necessário para tal. As taxas de ocorrência das reações foram determinadas utilizando a função R. As entradas são idênticas às apresentadas na função S_{norm}, mas no caso utiliza-se apenas os 14 nucleotídeos da bolha de transcrição no estado pré-translocado.

- $\boldsymbol{E(seq, conc, T, F)}$: Simula o alongamento completo de uma sequência *seq*, presente numa solução com *conc* micromolares de nucleotídeos na temperatura de T Kelvin, sob ação de uma força externa de F pN. Gera um vetor TPT formado por subvetores com 3 entradas cada: tempo decorrido, posição atual e tamanho do transcrito produzido. A cada nova simulação realizada, um novo subvetor é adicionado ao vetor TPT.

4.2 Aproximação para transcrição múltipla

O algoritmo simplificado para a *Aproximação para Transcrição Múltipla*, *ATM*, está representado na Figura 11. Determina-se, durante a simulação, uma equação para a posição de cada polimerase em função do tempo, utilizando-se uma aproximação linear. Maiores detalhes sobre como as colisões são resolvidas serão apresentados no próximo tópico. As etapas do algoritmo implementado são as seguintes:

(1) Inicialização: Definir a sequência molde (ex.: "AACCTTGG..."), a concentração de nucleotídeos e a temperatura durante a reação. Definir a força que o RNAP irá aplicar sobre a outra em caso de colisão, a posição e o tamanho da bolha de transcrição e do híbrido de DNA-RNA, além da região que a RNAP engloba. Definir o número de enzimas que irão realizar a transcrição, o tempo necessário para a iniciação e do tempo de alongamento máximo. Defina o tempo de alongamento, $t_{\mathrm{a}} = 0$ s.

(2) Ligação/Desligamento da RNAP: Se qualquer molécula de RNAP atingir o fim da cadeia de DNA, remova-a da simulação. Se o número de moléculas de RNAP adicionado até agora é menor do que o número máximo permitido e há espaço suficiente para outra RNAP ligar-se na fita de DNA, inclua uma nova enzima na primeira posição permitida e armazene a sua posição como zero e seu tempo como t_{a} mais o tempo atribuído para a iniciação. Determine o número de polimerase ligadas ao DNA, n.

(3) Critério de Parada: Se não houver mais moléculas de RNAP na fita de DNA ou t_{a} for maior do que o tempo de alongamento máximo estabelecido, pare.

(4) Simulação: Armazene a posição atual de todas as moléculas de RNAP, $S0_{\mathrm{i}}$, e para cada uma que ocupa uma posição que seja um número inteiro, armazene t_{a}, $S0$ e o comprimento da fita de RNA produzida. Gere uma nova simulação para essa enzima. Se ela estiver no alongamento normal, use a reação global de Equação 2.1. Para o cálculo das taxas de reação, utilize as Equações 2.3 e 2.4, com $F = 0$ pN. Se a molécula estiver em *backtracking*, gere uma simulação para a reação de recuo/avanço, calculando sua taxa usando a Equação 2.4, com $F = 0$ pN. Armazene a posição esperada após o movimento, $S\mathrm{f}$ e o tempo necessário para completar o seu passo, $t\mathrm{f}$.

(5) Equação da Posição: Utilizando uma aproximação linear, determine e armazene a equação da posição em função do tempo para todas as enzimas, isto é,

$$S(t) = S0 + ((S\mathrm{f} - S0)/t\mathrm{f})t$$

Se $n = 1$, faça $t_\mathrm{a} = t\mathrm{f}$ e vá para a etapa 7.

(6) Colisão: Resolva $S(t)_\mathrm{i} = S(t)_{\mathrm{i}+1}$, para "i" de 1 até $n - 1$, determinando o tempo necessário para que a colisão aconteça, $t\mathrm{c}_{\mathrm{i,i}+1}$, além de sua posição, $S\mathrm{c}_{\mathrm{i,i}+1}$. Determine se alguma colisão realmente ocorreu, isto é, $S\mathrm{c}_{\mathrm{i,i}+1}$ está entre $S0$ e $S\mathrm{f}$ para as enzimas "i" e "i + 1", e se $t\mathrm{c}_{\mathrm{i,i}+1} < t\mathrm{f}_\mathrm{i}$ e $t\mathrm{c}_{\mathrm{i,i}+1} < t\mathrm{f}_{\mathrm{i}+1}$. Se sim, armazene o tempo em que ocorre. Selecione o tempo mínimo armazenado entre todos os $t\mathrm{c}$ e $t\mathrm{f}$, t_min. Se t_min for um $t\mathrm{f}$, vá para a etapa 7; se não, realize uma nova simulação para as enzimas que colidiram entre si, utilizando um processo semelhante a etapa 4, mas alterando os valores das taxas de reação de acordo com o tipo de choque que ocorreu. Determine um novo $S_\mathrm{i}(t)$ para ambas, com $S0_\mathrm{i} = S\mathrm{c}$.

(7) Reposicionamento: Calcule $S(t_\mathrm{min})$ para cada molécula, e faça $S0 = S(t_\mathrm{min})$. Faça $t_\mathrm{a} = t_\mathrm{a} + t_\mathrm{min}$. Vá para a etapa 2.

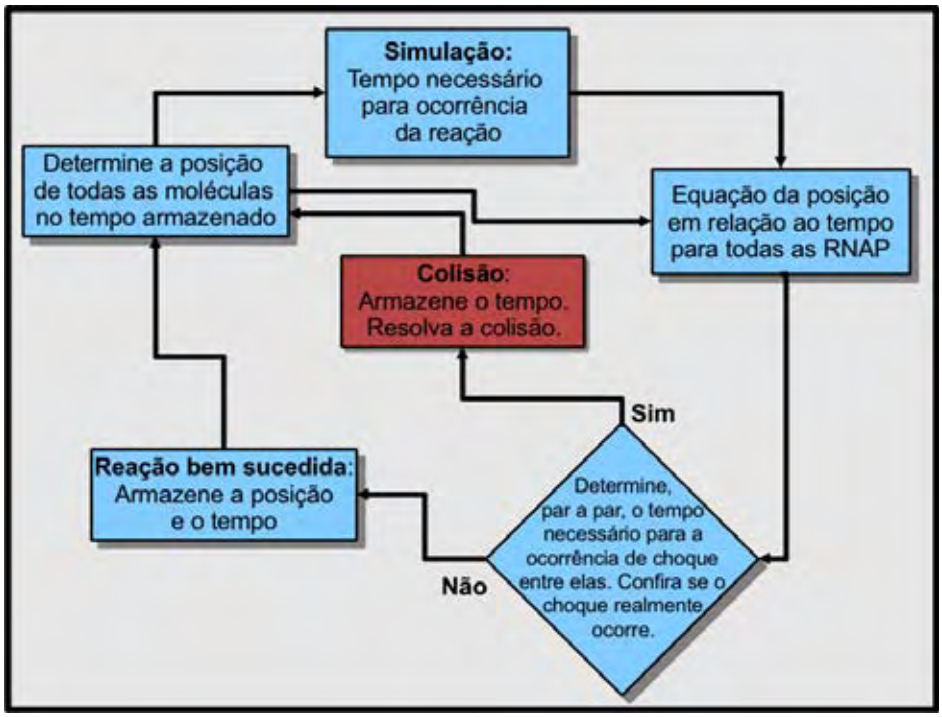

Figura 11: Resumo do algoritmo criado. Os passos em azul representam as reações normais, equivalente à *ATU*. Quando uma colisão é detectada, ela é resolvida alterando os valores das taxas de ocorrência das enzimas participantes. Detalhes dessa etapa são apresentados na próxima seção.

4.2.1 Colisões

Durante a simulação, tanto a molécula de RNAP traseira (T) e como a dianteira (D) podem estar em alongamento Normal (N), ou em *backtracking*, movendo-se tanto para trás (Bt) como para a frente (Bf). Devido a isto, podemos categorizar as colisões entre as moléculas de RNAP em seis diferentes casos. Basicamente, se T está realizando o alongamento normal, ela aplica 25 pN em D. Como resultado, D aplica 25 pN em T na direção oposta. Este valor foi escolhido pois é equivalente a máxima força de resistência que uma molécula RNAP que está transcrevendo é capaz de lidar (WANG et al., 1998). Se D está recuando, e após a colisão T movimenta-se para frente, T a empurra para frente. Se ambas estiverem recuando, tratamos a colisão como sendo perfeitamente elástica.

Resumindo, tratamos as colisões da seguinte forma:

1) *TN-DN*: Calcule $K_{\mathrm{i}}(n)$ utilizando $F = +25$ pN para D e $F = -25$ pN para T.

2) *TN-DBt* ou *TN-DBf*: O valor de F no cálculo de $k_{\mathrm{n,m}\rightarrow\mathrm{m+1}}$ para D será de +25 pN e no cálculo de $K_{\mathrm{i}}(n)$ para T será de -25 pN. Se após a simulação, T movimentar-se para frente, D também se movimentará para frente e o tempo máximo necessário para isto será igual ao tempo de T.

3) *TBf-DBt* ou *TBf-DBf* ou *TBt-DBt*: Permute as velocidades de D e T em suas equações $S(t)$. Não há necessidade de uma nova simulação. Mude os valores de $S\mathrm{f}$ em suas respectivas equações apropriadamente.

Para todos os casos, após a determinação do novo tempo para ocorrência das reações, t', determinamos $t\mathrm{f}$ como sendo $t\mathrm{f} = (1 - |S\mathrm{f} - S\mathrm{c}|)t'$. Para que não ocorram ultrapassagens, o valor de $t\mathrm{f}$ da enzima que tenta realizar a ultrapassagem é limitado: seu tempo sempre é maior ou igual ao tempo da outra enzima.

4.2.2 Funções para simulação da transcrição múltipla

Várias das funções presentes no Capítulo 4.1.4 foram reutilizadas na simulação da transcrição múltipla. As novas funções implementadas no pacote para Mathematica estão diretamente relacionadas com as etapas citadas no Algoritmo do Capítulo 4.2. As principais são as seguintes:

- $\boldsymbol{P(t_1, s_1, t_2, s_2, t)}$: Gera a equação da posição em relação ao tempo, $S(t)$, para as moléculas de RNAP que estão realizando a transcrição (Etapa 5, Capítulo 4.1.4).

Sua entrada consiste do tempo e da posição na qual a enzima inicia seu movimento, e do tempo e posição esperados para seu término.

- $\boldsymbol{A(seq, conc, T, F, n, p)}$: Função para o alongamento múltiplo. Une todas as funções apresentadas, apresentando sub-rotinas para resolução das colisões citadas no Capítulo 4.2.1. Simula o alongamento completo da sequência *seq* presente numa solução com *conc* micromolares dos nucleotídeos a uma temperatura de T Kelvin, conforme explicitado no Algoritmo do Capítulo 4.2. Esse alongamento será executado por n moléculas de RNAP que iniciam a transcrição p segundos após a disponibilização da região promotora. Essas enzimas poderão exercer uma força de intensidade máxima igual a F pN.

4.2.3 Simulações

Realizamos simulações para as mesmas sequências utilizadas na *ATU*, e para as mesmas condições. Para cada sequência, o número de moléculas de RNAP permitido variou de 2 até 10. Para cada caso e para cada sequência foram realizadas 4800 simulações.

4.3 Critérios para análise

4.3.1 Critério para sítios de pausa

O critério utilizado para determinação dos sítios de pausa é o mesmo escolhido por Bai e Wang (BAI; WANG, 2010). Esse critério depende da determinação dos tempos necessários para inclusão de cada nucleotídeo na fita de RNA, e varia de sequência para sequência. Para cada comprimento n do RNA produzido, determinou-se seu respectivo tempo para inclusão de um novo nucleotídeo, $\tau(n)$. Um sítio de pausa fica definido quando

$$\tau(n) > (1/\eta)\text{Min}\{\tau(n)\} \tag{4.2}$$

O termo $\text{Min}\{\tau(n)\}$ representa o menor tempo entre todos os $\tau(n)$ para uma dada sequência, temperatura e concentração de NTP disponível. O parâmetro η é ajustado para otimizar a previsão das pausas, considerando todas as sequências estudadas. Utilizando o mesmo preceito de Tadigotla *et al.* (TADIGOTLA et al., 2006), nós optamos por minimizar a proporção do número de previsões incorretas pelo número das previsões corretas, ou seja, o número de falsos positivos e falsos negativos dividido pelo número de verdadeiros positivos. Os valores utilizados para $\tau(n)$ equivalem ao valor do terceiro quar-

til da distribuição dos tempos obtidos nas simulações que consideram 24 °C e os menores valores de concetração de NTP, reproduzindo exatamente o experimento que determinou a posição desses sítios (LEVIN; CHAMBERLIN, 1987). O valor de η encontrado por Bai *et al.* para o *Model B* foi de 0,05. Entretanto, para a *ATU*, nós encontramos $\eta = 0,64$ e para a *ATM*, $\eta = 0,38$. Esses valores equivalem a 0,8 s para a *ATU* e 1,4 s para a *ATM*.

4.3.2 Géis teóricos

Levin e Chamberlin foram os primeiros a identificar os sítios de pausas transcricionais (LEVIN; CHAMBERLIN, 1987). Utilizando nucleotídeos radioativos como substrato para o alongamento transcricional, é possível observar a distribuição dos comprimentos dos transcritos produzidos após certo intervalo de tempo em um gel de eletroforese. Marcas mais acentuadas e que persistem nesses géis deixam claros os sítios de pausa e sua intensidade. Empregando o mesmo princípio, podemos gerar "géis de eletroforese teóricos", que, quando comparados com o original, ajudam a validar qualitativamente o modelo. Comparamos os resultados obtidos pela *ATU* e pela *ATM*. Como os parâmetros utilizados foram ajustados para 24 °C, e os géis originais foram gerados a partir da reação a 30 °C, os valores para as concentrações de NTP foram aumentadas nas simulações: de 10 μM para 25 μM (BAI; SHUNDROVSKY; WANG, 2004). Para a criação dos géis, foi necessária a determinação empírica dos limiares para seu escurecimento, ou seja, qual a concentração mínima de transcritos necessária para iniciar seu escurecimento, e qual a concentração de saturação do gel. Além disso, para o gel da *ATM*, consideramos que apenas a primeira RNAP deixará suas marcas, pois a concentração de nucleotídeos radioativos utilizada nos experimentos originais foi relativamente baixa.

5 Resultados e Discussão

A Figura 12 ilustra um exemplo de resultado obtido durante a simulação da ATM, ou seja, a posição em relação ao tempo para cada uma das moléculas de RNAP que iniciaram a transcrição durante uma simulação. As regiões coloridas representam o espaço ocupado pelas enzimas durante a reação. O choque entre as moléculas ocorre nos pontos onde regiões de diferentes tonalidades se tocam, e não ocorrem ultrapasagens. Ficam claras as regiões onde as moléculas necessitam de um intervalo maior de tempo para continuar a transcrição, além de regiões onde se observa o fenômeno de *backtracking*. Apesar do caráter ilustrativo da figura, nota-se que os principais pontos de choque entre as moléculas ocorrem justamente nos candidatos à sítios de pausa: quando o choque não ocorre, as enzimas tendem a passar mais tempo "presas" àquela posição.

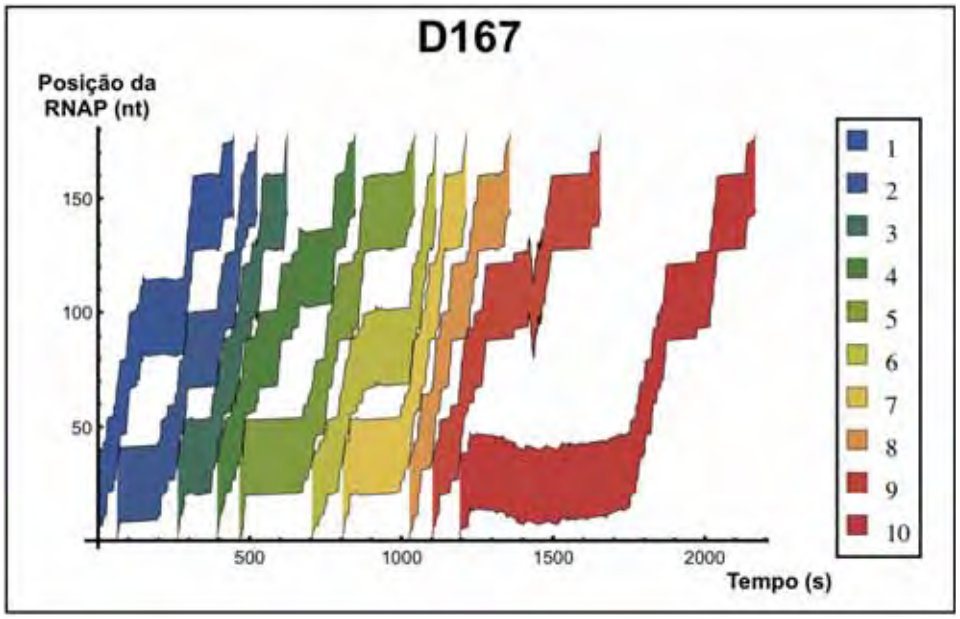

Figura 12: Exemplo do resultado obtido pelas simulações. Observa-se o comportamento cinético das moléculas de RNAP durante a ATM para a sequência D167. No eixo das abscissas temos o tempo total da reação e no eixo das ordenadas a posição da molécula no segmento de DNA. Cada região representa o espaço ocupado pela enzima durante a transcrição, sendo que as cores indicam a ordem de aconcoramento à fita de DNA. Note que não há sobreposições entre as regiões, apenas pontos onde as mesmas se encontram, indicando a ocorrência de uma colisão entre as enzimas. Além disso, ficam claras as regiões candidatas a sítio de pausa e a ocorrência de *backtracking*.

Inicialmente, iremos comparar o comportamento cinético da transcrição entre a *ATU* e a *ATM*, representado pelos resultados obtidos para as sequências D167 e D387. Na Figura 13a temos os tempos necessários para a inclusão do próximo nucleotídeo na fita de RNA, ou Tempo Entre Incorporações, *TEI*, pelo seu comprimento atual, com a *ATU* em verde e a *ATM* (primeira RNAP) em azul. O valor escolhido para os tempos representa sempre o terceiro quartil da distribuição dos mesmos para aquela posição. Já a Figura 13b mostra esses tempos obtidos para a *ATU* ordenados e os compara com os respectivos tempos para a *ATM*. Podemos notar, observando a Figura 13a, que o alongamento costuma ocorrer muito rapidamente na maioria das posições. Entretanto, algumas inclusões necessitam de um tempo muito maior: estas são posições candidatas a sítio de pausa. A diferença entre as aproximações é extremamente acentuada nesses sítios, o que podemos confirmar observando a Figura 13b. Portanto, é desta forma que a *ATM* acelera o processo de transcrição: a presença de várias RNAP durante o alongamento transcricional aumenta a probabilidade dessas moléculas passarem por sítios de pausa mais intensos, ou seja, diminui a probabilidade dessas enzimas ficarem presas a esses sítios, como confirmado por Epshtein e Nudler (EPSHTEIN; NUDLER, 2003).

Nota-se na Figura 13b que esse efeito é bastante significativo para a sequência D167. Mas para a sequência D387 ele se torna menos evidente. A diminuição do tempo de parada não ocorre de forma tão intensa ou simplesmente não ocorre, pois a enzima trazeira pode ficar presa a um local de pausa anterior, chegando demasiadamente tarde para auxiliar adequadamente a molécula à frente. Ou, como na seqüência D387, os sítios de pausa são pouco intensos, e as enzimas não possuem tempo suficiente para cooperar entre elas. O local da pausa mais intensa neste caso é logo no início da sequência e durante a transcrição não há outra enzima transcrevendo uma posição anterior. Mesmo assim, a *ATM* não obteve nenhum *TEI* significantemente superiores aos obtidos pela *ATU* para os comprimentos apresentados no gráfico.

Apesar da variedade na intensidade desse efeito para as diferentes sequências simuladas, a *ATM* acelerou significantemente a passagem das moléculas de RNAP pelos sítios de pausa. Na *ATU*, considetando apenas os sítios onde os tempos entre incorporações são mais significativos, ou seja, maiores que o nonagésimo quinto percentil da distribuição do *TEI* para cada uma das sequências, obtemos cerca de 44% do tempo necessário para o alongamento completo dessas fitas. Esse valor foi reduzido para 27% nas simulações considerando a *ATM*.

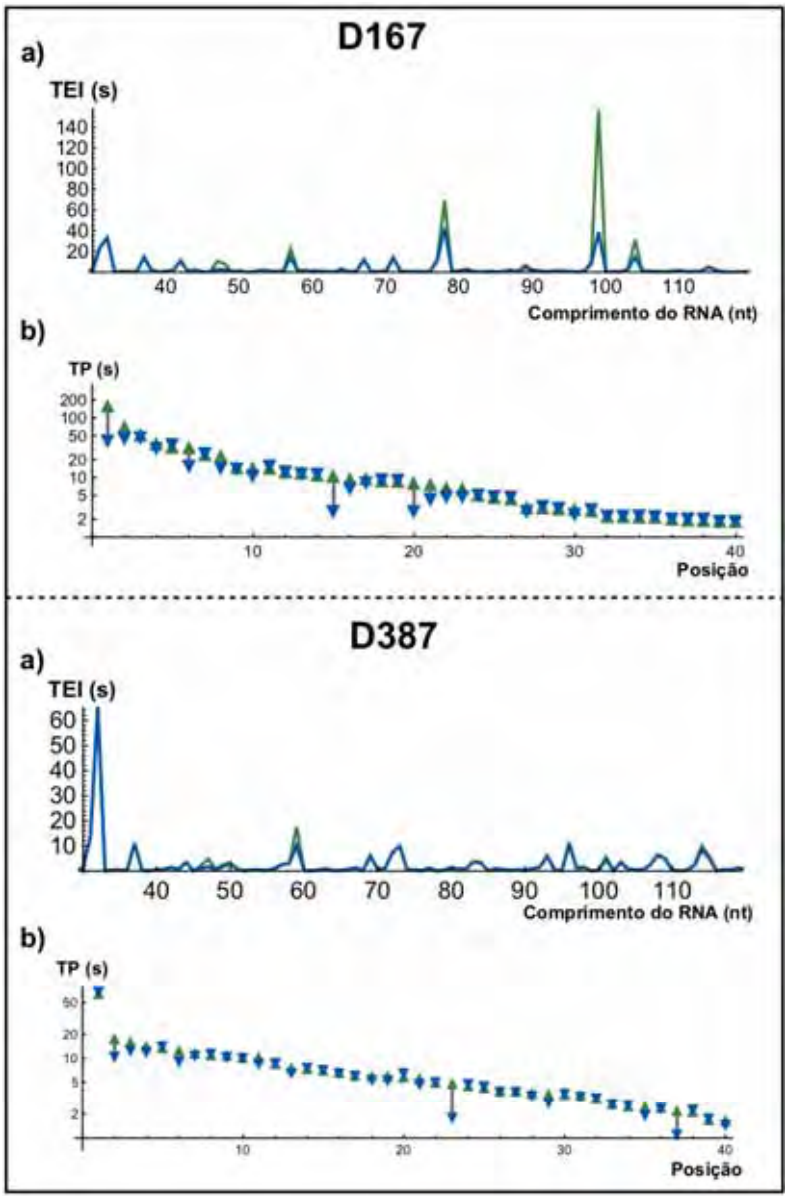

Figura 13: Comportamento cinético do alongamento para as sequências D167 e D387. a) Tempos Entre Incorporações (*TEI*) dos nucleotídeos pelo comprimento da fita de RNA. Em verde temos os *TEI* para a *ATU* e em cinza para a *ATM*. As curvas estão praticamente sobrepostas: diferenças significativas entre elas surgem apenas nas posições que possuem maiores valores de *TEI*. b) 40 maiores valores de *TEI* para a *ATU* (triângulo voltado para cima, em verde) e respectivo valor para a *ATM* (triângulo voltado para baixo, em azul). O gráfico está em escala logarítmica. Para essas posições, a *ATM* nunca obteve valores de *TEI* siginificantemente superiores aos obtidos pela *ATU*.

A supressão do *backtracking*, notada por Epshtein *et al.* (EPSHTEIN et al., 2003) e reproduzida pelo modelo de Klumpp (KLUMPP, 2011), também ocorre em nossa aproximação. O gráfico da distribuição das distâncias percorridas pelas RNAP durante o *backtracking* considerando todas as sequências estudadas está representado na Figura 14. Os valores das medianas dessas distribuições estão muito próximos, mas nas simulações considerando a *ATU*, as enzimas movimentam-se por distâncias maiores: o valor obtido no terceiro quartil dessa distribuição é 2,6 vezes maior do que no mesmo quartil da *ATM*. Essa supressão ocorre pois as enzimas em posições anteriores as que estão em recuo funcionam como barreiras físicas para esse movimento e facilitam o retorno das moléculas de RNAP em *backtracking* para a posição na qual poderá continuar o alongamento.

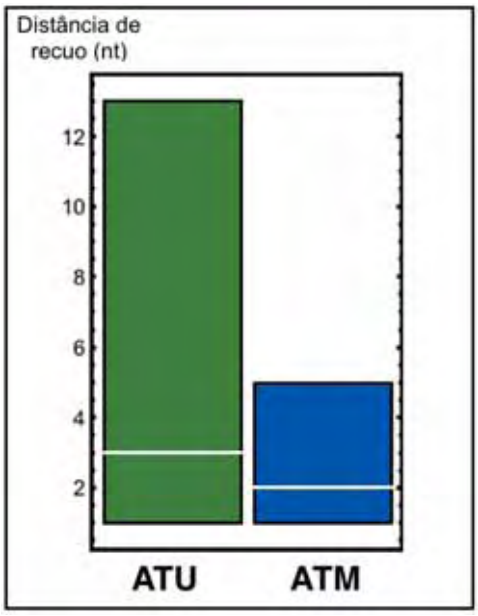

Figura 14: Distribuição das distâncias percorridas durante o *backtracking*. O limite inferior e superior das barras equivalem respectivamente ao primeiro e ao terceiro quartis. A mediana está destacada em branco. Apesar de valores medianos próximos, as enzimas movimentam-se significantemente mais longe na *ATU*.

Para avaliarmos os efeitos globais da transcrição múltipla, determinamos a distribuição dos tempos necessários para o alongamento completo (TAC) de uma determinada região de DNA para cada uma das enzimas que participa do processo. O resultado está representado na Figura 15, onde os limites inferior e superior de cada barra indica, respectivamente, a posição do primeiro e do terceiro quartil dessas distribuições, com a mediana destacada em branco. Nota-se que as primeiras moléculas de RNAP transcrevem a sequência mais rapidamente que as demais, enquanto as últimas enzimas são sempre as mais lentas. Este comportamento persiste para todas as sequências estudadas. O principal fator que torna as últimas enzimas mais lentas é justamente o fato de se manterem

presas aos sítios de pausa por não haver nenhuma outra molécula que possa facilitar sua passagem pelo sítio através da colisão. Isso torna seu movimento semelhante ao observado pela ATU, mas o gráfico mostra que seus tempos chegam a ser superiores aos obtidos por essa aproximação. Isso evidencia que as colisões prejudicam a reação de alongamento da enzima que choca-se com a molécula de RNAP imediatamente à sua frente, ou seja, nem todas as moléculas são beneficiadas durante a transcrição múltipla.

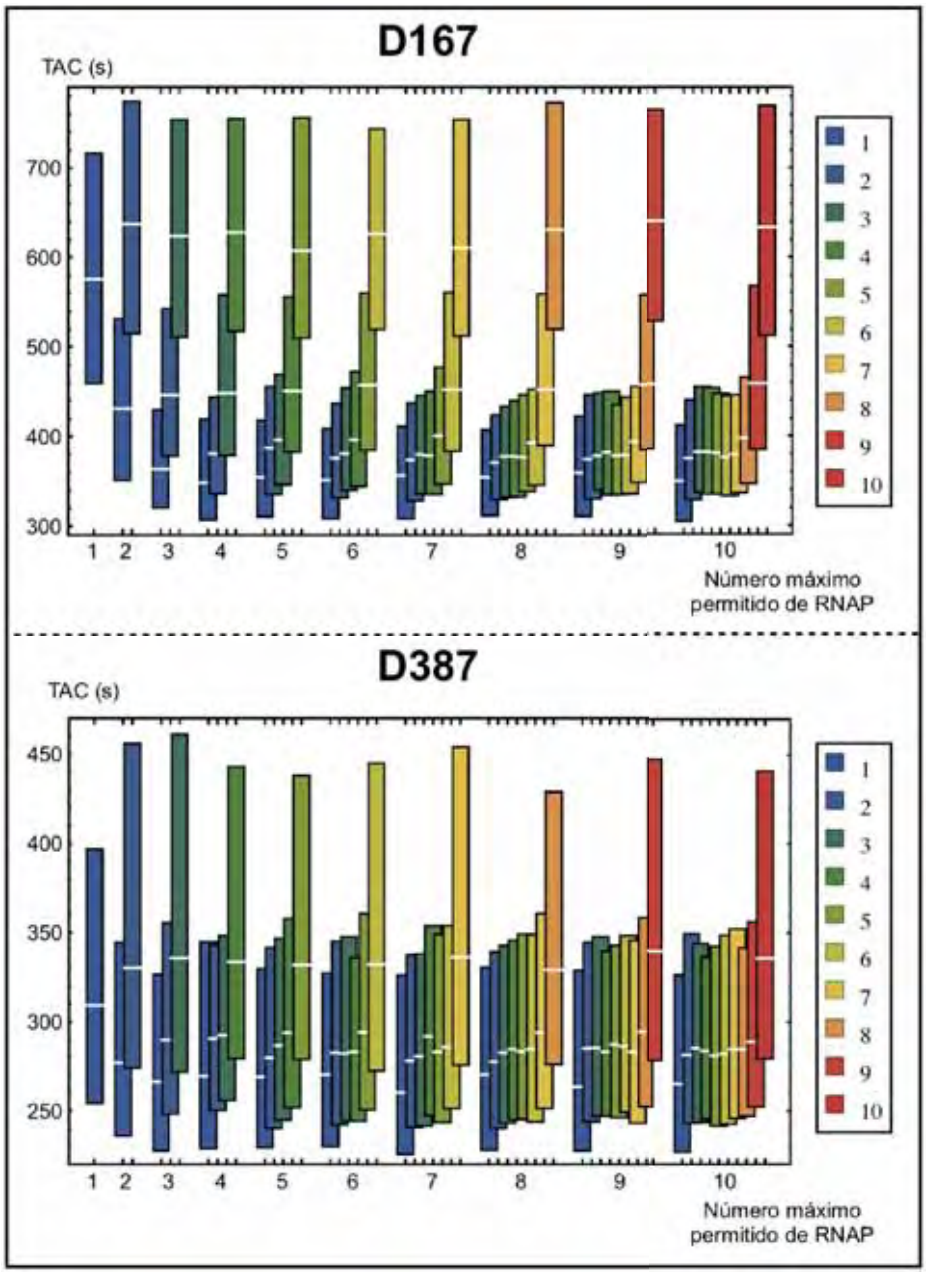

Figura 15: Distribuição dos tempos para o alongamento das sequências para cada RNAP. Tempo para Alongamento Completo (TAC) das sequências estudadas pelo número de moléculas de RNAP que iniciaram a transcrição. As cores indicam a ordem em que as enzimas foram incluídas na fita de DNA. O limite inferior e superior das barras equivalem respectivamente ao primeiro e ao terceiro quartis. A mediana está destacada em branco.

Afim de mensurarmos a eficiência da transcrição múltipla, definimos como tempo médio para a transcrição, TMT, a soma de todas as medianas dos TAC para cada uma das enzimas que participou da reação durante uma determinada simulação dividido pelo número de enzimas presentes naquela simulação. A partir desses valores, determinamos a relação entre as velocidades de transcrição média para as aproximações, dividindo o TMT resultante da ATU pelos valores obtidos nas ATM. Esse resultado está apresentado na Figura 16. A inclusão de moléculas de RNAP que poderão se ancorar a fita de DNA molde acarreta num aumento da velocidade média da transcrição. Entretanto, a intensidade desse fenômeno é dependente da sequência, conforme observado anteriormente. Por exemplo, obtivemos um aumento de apenas 3% na velocidade para a sequência D112 e de 48% para a sequência Seq13. A sequência D112 foi a única a apresentar uma diminuição na velocidade de transcrição na ATM, mas somente para a simulação que permitiu a participação de apenas 2 enzimas durante a reação. Note que as curvas possuem um comportamento assintótico: isso deve-se ao fato da fita molde ser finita, e portanto possuir um limite de enzimas ativas realizando sua transcrição. Além disso, o número, intensidade e posição das pausas também é um fator limitante: por exemplo, pausas intensas localizadas no inicio da sequência, como no caso da sequência D387, restringem a presença de várias moléculas de RNAP ativas na mesma.

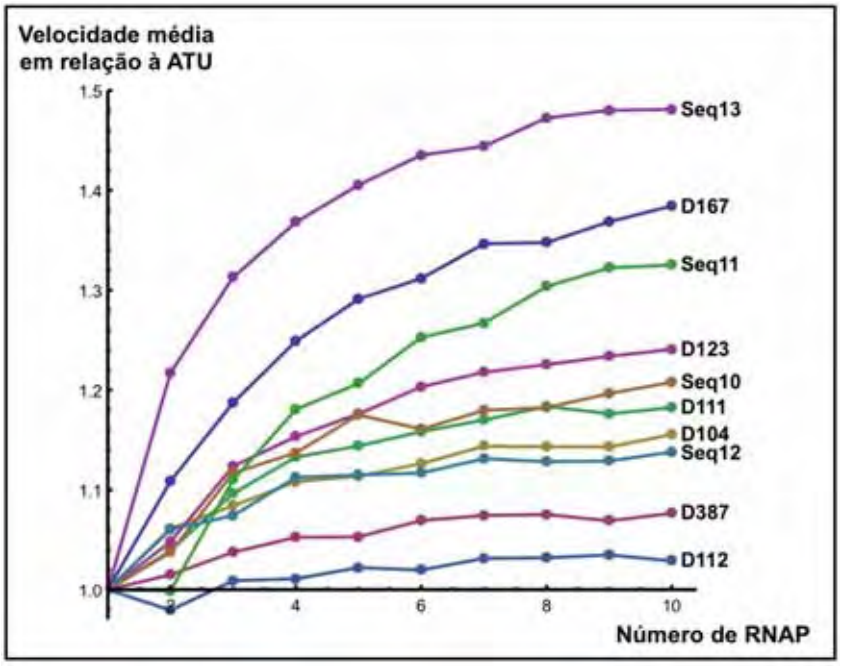

Figura 16: Relação entre as velocidades médias de transcrição das aproximações. Cada cor reprensenta uma sequência, nomeada no final da respectiva curva. A relação da velocidade média está apresentada em função do número de enzimas permitido durante a simulação. Note o comportamento assintótico das curvas e a disparidade dos resultados entre as sequências.

Utilizando o critério apresentado na Seção 4.3.1, determinamos os sítios de pausa para ambas as aproximações para compararmos o poder de predição dos modelos. Usando o mesmo tratamento de Tadigotla *et al.* (TADIGOTLA et al., 2006), consideramos que pausas localizadas a menos de 3 pares de base constituem um único aglomeramento de pausas. A Figura 17 compara a proporção global do número de previsões incorretas pelo número de previsões corretas entre as diferentes abordagens. Seus valores e as barras de erro, que representam o desvio padrão, foram determinados utilizando a técnica de *bootstrapping*: substituí-se os resultados de uma ou mais sequências do conjunto por outras do mesmo conjunto e determina-se o novo valor dessa relação, refazendo quantas vezes forem necessárias para os valores da média e do desvio se estabilizarem. Comparando os valores obtidos pelo *Modelo B*, nota-se que os resultados foram sutilmente aprimorados utilizando os parâmetros atualizados (ATU) e ainda melhores para a ATM.

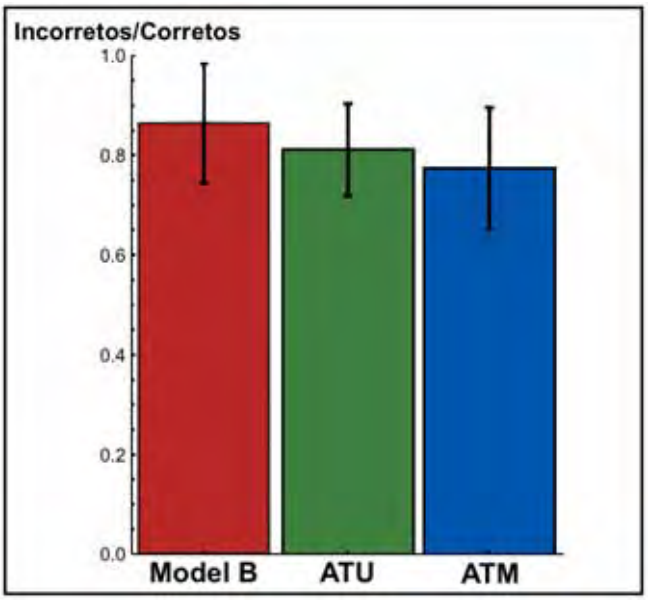

Figura 17: Relação entre o número de previsões incorretas pelo número de acertos de cada modelo. Valores determinados utilizando a técnica de *bootstrapping*. As barras de erro representam o desvio padrão obtido.

A Figura 18 compara os géis teóricos obtidos para tempos determinados com os resultados experimentais. Essa figura compara os resultados dos ensaios bioquímicos com os obtidos teoricamente. Temos os resultados para a aproximação para transcrição única, aproximação para transcrição múltipla e uma representação dos resultados experimentais. Os comprimentos dos transcritos podem ser vistos eixo vertical, partindo, de baixo para cima, do menor para o maior produto da transcrição. A intensidade das bandas indica a concentração de transcritos naquela região. Esses géis confirmam a afirmação de Bai *et al.* (BAI; SHUNDROVSKY; WANG, 2004) de que essa modelagem pode recuperar adequada-

mente a cinética da produção de RNA. Para ambas as sequências, os resultados obtidos pelas simulações estão qualitativamente satisfatórios. A diferença entre as aproximações estão em algumas bandas que aparecem com intensidades diferentes e durante intervalos diferentes. Pode-se notar que a *ATM* está mais próxima do resultado esperado, pois basicamente reduz o número de falsos positivos, sem prejudicar os verdadeiros positivos, justificando o resultado apresentado pela Figura 17.

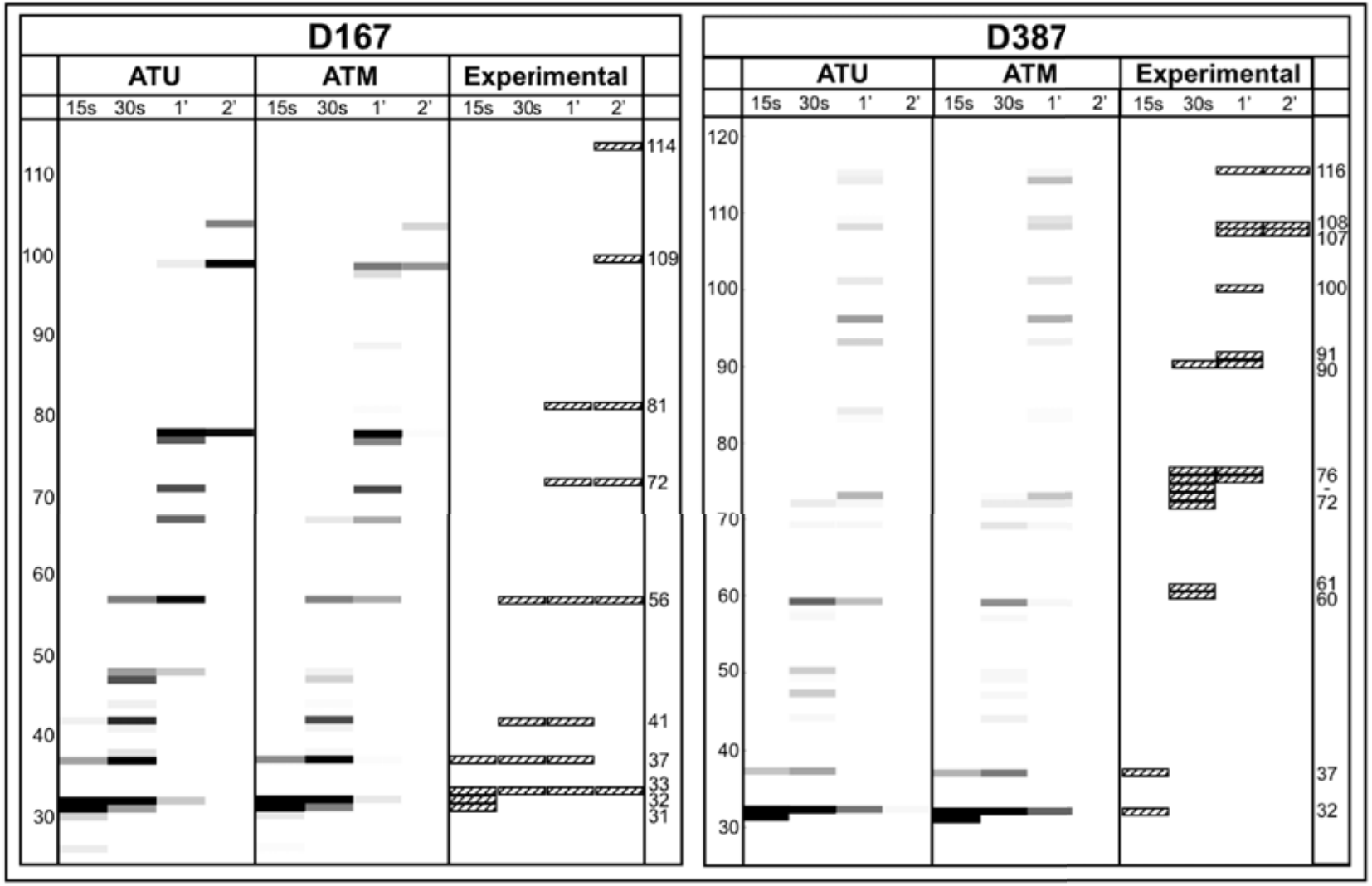

Figura 18: Comparação entre os géis teóricos e o resultado experimental. Comparação entre os resultados obtidos utilizando a *ATU* e a *ATM*, para diferentes tempos de transcrição. Cada coluna representa o tempo de transcrição permitido durante aquela simulação. A intensidade das bandas é proporcional a quantidade de transcritos com determinado comprimento produzidos pelas enzimas até o tempo permitido. Os resultados experimentais representados aqui são reproduções dos géis experimentais apresentados no trabalho de Bai e Wang (BAI; WANG, 2010). Note que alguns falsos positivos presentes na *ATU* surgem com menor intensidade na *ATM*.

6 Conclusões

Atualização do Modelo B

O modelo cinético estocástico para o alongamento transcricional apresentado por Bai *et al.* foi implementado em linguagem Mathematica e seus parâmetros foram atualizados com os dados disponíveis na literatura. Denominamos essa modelagem de *Aproximação para Transcrição Única*, afim de diferenciar os resultados originais de Bai *et al.* dos produzidos neste trabalho.

Generalização do modelo: transcrição múltipla

Foi criado um algoritmo capaz de simular o processo de transcrição múltipla a partir do Modelo B, denominado no trabalho de *Aproximação para Transcrição Múltipla*. O código foi implementado em linguagem Mathematica, e lida com os seis tipos de colisões possíveis entre as moléculas de RNAP durante o alongamento transcricional, utilizando princípios termodinâmicos e mecânicos para tal.

Resultados obtidos pelas aproximações

Os resultados estão de acordo com a literatura, mostrando que a presença de múltiplas moléculas de RNAP pode aprimorar a transcrição, reduzindo a duração das pausas e suprimindo o *backtracking*. Este fenômeno pode ser interpretado como uma forma de regulação intrínseca: apesar de algumas pausas serem importantes durante o alongamento, podem ser muito longas e a transcrição múltipla é uma maneira eficiente para atenuá-las em genes que necessitam ser altamente transcritos, por exemplo. Entretanto, como esperado, esse aumento não é indefinido: a taxa de alongamento atinge um valor máximo, definido pelo número de moléculas de RNAP que a fita molde comporta. Este número varia de sequência para sequência e está ligado diretamente ao comprimento da sequência, a concentração de NTP e do número, posição e intensidades dos sítios de pausa. A cinética da transcrição e consequentemente o poder de predição de sítios de pausa na *ATM* é

ligeiramente melhor que na *ATU*, mas o número de falsos positivos ainda é elevado. As prováveis fontes de erro residem nas considerações e nos parâmetros ajustados por Bai *et al.*: por exemplo, assumiu-se que a barreira energética para o *backtracking* é constante durante todo o alongamento, e que a energia livre de Gibbs para os pares de base presentes no CAT são comparáveis aos valores medidos em solução. O modelo ainda carece da inclusão de fatores externos à estrutura da bolha de trancrição, como o dobramento e a interação da RNAP com os polímeros de ácidos nucléicos, que podem interferir de forma considerável no alongamento transcricional.

Perspectivas futuras

A contribuição da interação do RNA nascente com a RNAP não foi incorporada em nossa modelagem. Sabe-se que o RNA é capaz de se dobrar e formar estruturas tridimensionais: pausas devido a formação de grampos de RNA, conhecidas pelo seu papel regulatório, já foram discutidas na literatura (HERBERT; GREENLEAF; BLOCK, 2008) e não podem ser previstas pela modelo apresentado. Além disso, as estruturas formadas pelo RNA possuem as mais diferentes funções nos organismos. Prever e calcular a contribuição dessas estruturas formadas por moléculas de RNA poderá não só aprimorar o modelo, como contribuir no entendimento de suas interações intracelulares.

Referências

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.

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. **Journal of Experimental Medicine**, v. 79, n. 2, p. 137–158, 1944.

BAI, L.; FULBRIGHT, R. M.; WANG, M. D. Mechanochemical kinetics of transcription elongation. **Physical Review Letters**, v. 98, n. 6, 2007.

BAI, L.; SHUNDROVSKY, A.; WANG, M. D. Sequence-dependent kinetic model for transcription elongation by rna polymerase. **Journal of Molecular Biology**, v. 344, n. 2, p. 335–349, 2004.

BAI, L.; WANG, M. D. Comparison of pause predictions of two sequence-dependent transcription models. **Journal of Statistical Mechanics-Theory and Experiment**, n. 12, 2010.

BAIROCH, A. The enzyme database in 2000. **Nucleic Acids Research**, v. 28, n. 1, p. 304–305, 2000.

BUC, H.; STRICK, T. **RNA Polymerases as Molecular Motors**. 1ª. ed. Cambridge: Royal Society of Chemistry, 2009.

BURGESS, R. R. Rna polymerase. **Annual Review of Biochemistry**, v. 40, p. 711–740, 1971.

CRAMER, P. Common structural features of nucleic acid polymerases. **Bioessays**, v. 24, n. 8, p. 724–729, 2002.

CRICK, F. Central dogma of molecular biology. **Nature**, v. 227, n. 5258, p. 561–563, 1970.

DAHM, R. Friedrich miescher and the discovery of dna. **Developmental Biology**, v. 278, n. 2, p. 274–288, 2005.

DARZACQ, X. et al. In vivo dynamics of rna polymerase ii transcription. **Nature Structural and Molecular Biology**, v. 14, n. 9, p. 796–806, 2007.

EPSHTEIN, V.; NUDLER, E. Cooperation between rna polymerase molecules in transcription elongation. **Science**, v. 300, n. 5620, p. 801–805, 2003.

EPSHTEIN, V. et al. Transcription through the roadblocks: the role of rna polymerase cooperation. **EMBO Journal**, v. 22, n. 18, p. 4719–4727, 2003.

FEMINO, A. M. et al. Visualization of single molecules of mrna in situ. **Methods in Enzymology**, v. 361, p. 245–304, 2003.

FOE, V. E. Modulation of ribosomal rna synthesis in *Oncopeltus fasciatus*: An electron microscopic study of the relationship between changes in chromatin structure and transcriptional activity. **Cold Spring Harbor Symposia on Quantitative Biology**, v. 42, p. 723–740, 1977.

FRIEDMANN, H. C. From butyribacterium to e. coli : An essay on unity. **Biochemistry Perspectives in Biology and Medicine**, v. 47, p. 47–66, 2004.

GIBSON, M. A.; BRUCK, J. Efficient exact stochastic simulation of chemical systems with many species and many channels. **Journal Physical Chemistry**, v. 104, n. 9, p. 1876–1889, 2000.

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

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.

GRIFFITH, F. The significance of pneumococcal types. **Journal of Hygiene**, v. 27, n. 2, p. 113–159, 1928.

GRIFFITHS, A. J. F. et al. **Introduçao à Genética**. 8ª. ed. São Paulo: Guanabara Koogan, 2006.

GRIGOROVA, I. L. et al. Insights into transcriptional regulation and sigma competition from an equilibrium model of rna polymerase binding to dna. **Proceedings of the National Academy of Sciences of the United States of America**, v. 130, n. 14, p. 5332–5337, 2006.

GUAJARDO, R.; SOUSA, R. A model for the mechanism of polymerase translocation. **Journal of Molecular Biology**, v. 265, n. 1, p. 8–19, 1997.

HAMMING, J. et al. Electron microscopic analysis of transcription of a ribosomal rna operon of *E. coli*. **Nucleic Acids Research**, v. 9, n. 6, p. 1339–1350, 1981.

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.

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

HERSHEY, A. D.; CHASE, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. **The Journal of Cell Biology**, v. 36, p. 39–56, 1952.

KAPANIDIS, A. N. et al. Initial transcription by rna polymerase proceeds through a dna-scrunching mechanism. **Science**, v. 314, n. 5802, p. 1144–1147, 2006.

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

KLUMPP, S. Pausing and backtracking in transcription under dense traffic conditions. **Journal of Statistical Physics**, v. 142, n. 6, p. 1252–1267, 2011.

LENINGER, A. **Princípios de Bioquímica**. 4ª. ed. São Paulo: Artmed, 2006.

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.

MARTIN, F. H.; TINOCO, I. J. Dna-rna hybrid duplexes containing oligo(da:ru) sequences are exceptionally unstable and may facilitate termination of transcription. **Nucleic Acids Research**, v. 6, n. 10, p. 2295–2299, 1980.

MCCLURE, W. R.; CECH, C.; JOHNSTON, D. E. A steady state assay for the rna polymerase initiation reaction. **The Journal of Biological Chemistry**, v. 253, n. 24, p. 8941–8948, 1978.

MENTEN, L.; MICHAELIS, M. I. Die kinetik der invertinwirkung. **Biochem. Z**, v. 49, p. 333–369, 1913.

NEUMAN, K. C. et al. Ubiquitous transcriptional pausing is independent of rna polymerase backtracking. **Cell**, v. 115, n. 11, p. 437–447, 2003.

O'BRIEN, T.; LIS, J. T. Rapid changes in drosophila transcription after an instantaneous heat shock. **Molecular and Cellular Biology**, v. 314, n. 6, p. 3456–3463, 1993.

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

RADZICKA, A.; WOLFENDEN, R. A proficient enzyme. **Science**, v. 267, n. 5194, p. 90–93, 1995.

RAJALA, T. et al. Effects of transcriptional pausing on gene expression dynamics. **PLoS Computational Biology**, v. 6, n. 3, 2010.

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.

SAEKI, H.; SVEJSTRUP, J. Q. Stability, flexibility, and dynamic interactions of colliding rna polymerase ii elongation complexes. **Molecular Cell**, v. 35, n. 2, p. 191–205, 2009.

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.

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

SCHLIWA, M. Molecular motors. **Nature**, v. 422, n. 4, p. 759–765, 2003.

SCHRÖDINGER, E. **What Is Life?** 1ª. ed. New York: Cambridge University Press, 1944.

SHAPIRO, B. E. et al. Cellerator: extending a computer algebra system to include biochemical arrows for signal transduction simulations. **Bioinformatics**, v. 19, n. 5, p. 677–678, 2003.

SOUZA, R. Structural and mechanistic relationships between nucleic acid polymerases. **Trends in Biochemical Sciences**, v. 21, n. 5, p. 186–190, 1996.

STEITZ, T. A. Structural biology: A mechanism for all polymerases. **Nature**, v. 391, p. 231–232, 1998.

STEITZ, T. A. Dna polymerases: Structural diversity and common mechanisms. **Journal of Biological Chemistry**, v. 274, n. 6, p. 17395–17398, 1999.

SUGIMOTO, N. et al. Thermodynamic parameters to predict stability of rna/dna hybrid duplexes. **Biochemistry**, v. 34, n. 35, p. 11211–11216, 1995.

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.

TENNYSON, C. N.; KLAMUT, H. J.; WORTON, R. G. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. **Nature Genetics**, v. 9, n. 2, p. 184–190, 1995.

WANG, M. D. et al. Force and velocity measured for single molecules of rna polymerase. **Science**, v. 282, n. 5390, p. 902–907, 1998.

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.

Anexo 1

Sequências utilizadas nas simulações

Seq10: AUC GAG AGG GAC ACG GCG AAU AGC CAU CCC AAU CGA CAC CGG GGU CCG GGA UCU GGA UCU GGA UCG CUA AUA ACA UUU UUA UUU GGA UCC CCG GGU ACC GAG CUC GAA UUC ACU GGC CGU CGU UUU ACA ACG UCG UGA CUG GGA AAA CCC UGG CG

Seq11: AUC GAG AGG GAC ACG GCG AAU AGC CAU CCC AAU CCG ACA CCG GGG CAU CGA GUG GGA CAC GGC GAA UAG CCA UCC CAA UCG ACA CCG GGG UCC GGG AUC UGG AUC UGG AUC GCU AAU AAC AGG CCU GCU GGU AAU CGC AGG CCU UUU UAU UUG GAU CCC CGG GUA

Seq12: AUC GAG AGG GCC ACG GCG AAC AGC CAA CCC AAU CGA ACA GGC CUG CUG GUA AUC GCA GGC CUU UUU AUU UGG AUC CCC GGG UA

Seq13: AUC GAG AGG GCC ACG GCG AAC AGC CAA CCC AAU CCG AAC AGC CAU CAU CCU CAG UAU UCA GGU AGC UGU UGA GCC UGG GGC GGU AGC GUG CUU UUU UCG AAU UCA CUU AAU GGU AAU CUC G

D104: AUC GAG AGG GAC ACG GCG AAU AGC CAU CCC AAU CGA CAC CGG GGU CAA CCG GAU AAG UAG ACA GCC UGA UAA GUC GCA CGA CAG AAA GAA AUU GAC CGC GCU AAG GCC CGU AAA GAA CGU CAC GAG GGG CGC UUA GAG GCA CGC AGA UUC AAA CGU CGC A

D111: AUC GAG AGG GAC ACG GCG AAU AGC CAU CCC AAU CCA CAC GUC CAA CGG GGC AAC CGU AUG UAC ACC UGA UGG GUU CGC AAU GAA ACA ACG AAU CGA ACG CCU UAA GCG UGA ACU CCG CAU UAA CCG CAA GAU UAA CAA GAU AGG UUC CGG CUA UGA CAG A

D112: AUC GAG AGG GAC ACG GCG AAU AGC CAU CCC AAU CGA CAC CGG GGU CAA CCG GAU AAG UAG ACA GCC UGA UAA GUC GCA CUA GAA CAG GCA CUA GCC AAC ACA CUG AAC GAU AUC UCA UAA CGA AGA UAA AGG ACA CAA UGC AAU GAA CAU UAC CGA CAU C

D123: AUC GAG AGG GAC ACG GCG AAU AGU GAG AAC UUG GCG AGA GAA CAA CCU CGA ACG CCG CAA GGC ACA AGA GAG GGC GGC GUG GCA UAG ACG AAA GGA AAA GGU UAA AGC CAA GAA ACU CGC CGC ACU UGA ACA GGC ACU AGC CAA CAC ACU GAA CGC UAU C

D167: AUC GAG AGG GAC ACG GCG AAU AGC CAU CCC UAA CGU CUA CGA UGU ACA GCG CCA CGC UGG AUG CUA UAC GGU GGU ACU UGA CGC ACU UAA GGA UUG CGA GCG UUU CAA CAA UGA UGC CCA UUA UAA AUA CGC UGA GAU UGC AAG CGA CAU CAU UGA UUG C

D387: AUC GAG AGG GAC ACG GCG AAU AGC CAU CCC AAU CGA CAC CGG GGU CAA CCG GAU AAG UAG ACA GCC UGA UAA GUC GCA CGA AAA ACA GGU AUU GAC AAG CGU CAA GGU AUG CUU AUC GAC UUA CUG GUC GAG AUG GUC AAC AGC GAG ACG UGU GAU GGC G

Anexo 2

Trabalhos baseados nos resultados obtidos

1. Apresentação Oral: COSTA, P. R. ; LEMKE, N. . Cooperative behavior on gene expression: a stochastic sequence-dependent model for transcription elongation. XXXV Encontro Nacional de Física da Matéria Condensada. Águas de Lindóia-SP, 2012.

2. Apresentação de pôster: COSTA, P. R. ; LEMKE, N. . Modelo cinético estocástico da etapa de alongamento transcricional na presença de múltiplas polimerases. VII EVFITA - Encontro de Verão de Física do ITA. São José dos Campos-SP, 2012. Menção Honrosa.

3. Apresentação Oral: COSTA, P. R. ; LEMKE, N. . Comportamento das pausas transcricionais em um modelo cinético estocástico para mais de uma molécula de RNA polimerase na fita de DNA. VII Confiam - Congresso de Física Aplicada à Medicina. Botucatu-SP, 2011.

4. Apresentação de pôster: COSTA, P. R. ; LEMKE, N. . Determinação teórica da sequência ótima de códons para a transcrição de uma dada sequência de peptídeos. VII Confiam - Congresso de Física Aplicada à Medicina. Botucatu-SP, 2011.

5. Apresentação de pôster: COSTA, P. R. ; LEMKE, N. . Simulating RNAP Transcription Pauses using Gillespie Algorithm. Encontro de Física. Foz do Iguaçu-PR, 2011.

6. Apresentação de pôster: COSTA, P. R. ; LEMKE, N. . Comportamento cinético esperado de múltiplas polimerases durante o alongamento de uma mesma região de DNA. X Workshop da Pós-graduação. Botucatu-SP, 2011.

7. Apresentação de pôster: COSTA, P. R. ; CASTRIOTA, M. ; LEMKE, N. . Simulating RNA polymerase mobility using a stochastic kinetic model and comparison with agarose gel electrophoresis. 6th International Conference of the Brazilian Association for Bioinformatics and computational Biology. Ouro Preto-MG, 2010.

8. Apresentação de pôster: COSTA, P. R. ; LEMKE, N. . Stochastic Kinetic Model for Eukaryotic Transcription considering collisions among RNA Polymerases II. Symposium on Systems Biology. Natal-RN, 2010.

Cooperative RNA Polymerase Molecules Behavior on a Stochastic Sequence-Dependent Model for Transcription Elongation

Pedro Rafael Costa, Ney Lemke
Departamento de Física e Biofísica, Instituto de Biociências de Botucatu,
UNESP - Univ Estadual Paulista, Distrito de Rubião Jr. s/n, Botucatu, São Paulo, 18618-970, Brasil

Abstract

The transcription of the information encoded within the DNA to an RNA molecule is exquisitely controlled during the development of the organisms and to its vital functions and has as the protagonist the RNA polymerase enzyme (RNAP). The development of single-molecule techniques, as magnetic and optical tweezers, atomic-force microscopy and single-molecule fluorescence, increased our understanding of the process, complementing traditional biochemical studies. Theoretical models have been proposed to explain and predict the RNAP kinetics during the polymerization. Models based on the thermodynamic stability of the transcription elongation complex recover much of the kinetics and indicate that its movement has a Brownian ratchet mechanism. However, experiments showed that if more than one RNAP molecule initiates from the same promoter, their behavior slightly change and new phenomenona are observed. We proposed and implemented a theoretical model that considers collisions between RNAP and predict their cooperative behavior during multi-round transcription. The model generalizes Bai *et al*. stochastic sequence-dependent model. In our approach, collisions between elongating enzymes modify their transcription rate values. We perform the simulations in Mathematica and compared the results of the single and the multiple-molecule transcription with experimental results and other theoretical models. Our multi-round approach could recover several expected behaviors, and achieved a better predictive power, when compared with the single-round one. Our findings show that the collisions between the enzymes collaborate to the transcription, carrying it out 48% faster on average for the studied sequences.

Author Summary

Since the arises of molecular biology there is an interest in the study of transcription of the information within DNA to an RNA molecule. With the advent of more advanced techniques, we observe in more details the behavior of the enzyme responsible for the process, the RNA polymerase (RNAP). The traditional biochemical studies combined with the results of techniques capable of tracking the movement of a single RNAP molecule stimulated the development of theoretical models for explaining and predicting its kinetic behavior. However, we need to consider the interactions that could occur between these enzymes when they transcript simultaneously the same DNA region. We present an approach that considers the interactions between these enzymes. Using large scale simulations we concluded that several RNAP cooperate in the transcription, as observed experimentally. Our results showed an increase of the overall transcription rate since collisions reduces significantly pause lengths.

Introduction

The first step of the Central Dogma of Molecular Biology concerns the transport of information within DNA to an RNA molecule. This process, known as *transcription*, must be exquisitely controlled during the development and maintenance of living beings. It can be divided into three phases, *initiation*, *elongation* and *termination*, and has as the protagonist the RNA polymerase enzyme (RNAP), regarded as a molecular motor.

RNAP scan the duplex DNA to find the sites for transcription initiation, known as *promoters*, bind to them, and expose the DNA template. Therefore its active site is in the correct position, the Transcriptional Elongation Complex (TEC) starts the elongation phase. During this phase, the RNAP polymerizes RNA chains, incorporating a ribonucleoside complementary to the nucleotide present in its active site. After incorporation, it moves along one nucleotide in the template strand and then restarts the process. When it recognizes the site for termination, the TEC is disassembled and the RNAP releases the transcript and disengages from DNA. Throughout this process, RNAP is able to recruit accessory proteins for several of these activities.

The development of single-molecule techniques, such as magnetic and optical tweezers, atomic-force microscopy and single-molecule fluorescence, increased our understanding of the phenomenon, complementing traditional biochemical studies (for review, see Herbert *et al*. [1]). Kinetics studies showed that the transcription does not occur at a uniform rate: the RNAP could get stuck at specific sites. These events are referred to as *transcriptional pauses*. These pauses may be due to the formation of RNA hairpins that interact with RNAP, due to the movement of the polymerase in the opposite direction of transcription, a phenomenon known as *backtracking*, or it may be caused by sequence dependent