Desenho e simulação de modelos computacionais da dinâmica de replicação de DNA em kinetoplastídeos

Monografia apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
aprovação em MAC499 – Trabalho
de
Conclusão de Curso

Aluno: Gustavo Rodrigues Cayres Silva

Orientador: Marcelo da Silva Reis

Centro de Toxinas, Resposta-imune e Sinalização Celular
(CeTICS)

Laboratório Especial de Ciclo Celular, Instituto Butantan

São Paulo, 15 de Dezembro de 2017

Resumo

A replicação de DNA de células eucariotas é um processo que ocorre essencialmente de forma paralelizada, iniciando-se diversas vezes ao longo da fase S do ciclo celular em sítios denominados "origens de replicação". Todavia, a dinâmica desse processo é sujeita a fatores tais como a frequência de disparo de cada origem e eventuais colisões entre as maquinarias de replicação e de transcrição. Existe a hipótese de que essa dinâmica esteja por trás das diferenças de arquitetura genômica entre *Trypanosoma cruzi* e outros kinetoplastídeos. Para investigar essa questão, neste projeto propomos a construção de um modelo computacional estocástico para estudar as propriedades dinâmicas da replicação em *T. cruzi*, em particular aspectos das colisões que ocorrem durante esse processo. Esse modelo computacional será desenhado utilizando cadeias de Markov, implementado em Python e ajustado utilizando informações biológicas da literatura e também as produzidas em nosso laboratório. Esperamos, ao simular o modelo ajustado com diferentes condições, descobrir propriedades emergentes da dinâmica de replicação de DNA em *T. cruzi* e, num segundo momento, também em outros kinetoplastídeos e em levedura.

Palavras-chave: Ciclo celular, modelo computacional, replicação do DNA, transcrição do DNA, kinetoplastídeos

Agradecimentos

Agradeço às seguintes pessoas, cujo auxílio e suporte foram indispensáveis para o desenvolvimento deste projeto:

- Marcelo da Silva Reis, meu orientador e colega durante esses vários meses, responsável por me guiar, auxiliar e, claro, cobrar resultados dos quais ele sabia que eu era capaz, tendo em vista sempre o meu melhor.
- Marcelo Santos da Silva, colega do Instituto Butantan que contribuiu imensamente com as informações necessárias para o avanço e desenvolvimento do projeto.
- Gustavo Estrela de Matos, colega do IME-USP responsável por me indicar a este projeto e a quem estarei sempre agradecido.
- Duílio Felix Oliveira da Silva, ex-administrador da Rede Linux e estimado amigo que me ajudou repetidas vezes com várias dúvidas de programação.
- Todos meus colegados do Instituto Butantan e do IME-USP, cujo suporte foi essencial para meu sucesso acadêmico.
- Meus pais, Alcíone de Oliveira Santos Cayres e Miltom Rodrigues da Silva, cujo suporte emocional e financeiro me permitiu focar exclusivamente nos estudos.
- Vinícius Carlos Brando, meu parceiro que esteve do meu lado com grande atenção, mesmo quando era impossível retribuí-la.

Muito obrigado e espero que meus esforços neste trabalho só tenham nos aproximado e aumentado o apreço que temos uns pelos outros.

Conteúdo

1 Introdução **1**

1.1 Organização do trabalho 2

2 Conceitos Fundamentais **3**

2.1 O DNA, o genoma e o cromossomo 3

2.2 O ciclo celular 5

2.2.1 A fase S 5

2.3 As origens de replicação 6

2.3.1 Origens constitutivas, flexíveis e dormentes 6

2.4 O RNA e o RNA mensageiro 8

2.5 A transcrição 8

2.5.1 Transcrições policistrônica e constitutiva 9

2.6 Conflitos replicação-transcrição 9

3 Um modelo da dinâmica de replicação de *T. brucei* TREU927 **10**

3.1 Representação da estrutura do genoma 10

3.1.1 Maquinarias de replicação 11

3.1.2 Maquinarias de transcrição 11

3.2 Representação da dinâmica de replicação de DNA 11

3.3 Definição da distribuição de probabilidade de disparo de origens de replicação 13

4 Implementação de um simulador do modelo dinâmico **16**

4.1 Organização de informações *a priori* em banco de dados relacional 16

4.2 Especificação do simulador 17

4.3 Escolha da linguagem de programação 18

4.4 Execução das simulações 19

5 Resultados em experimentos computacionais **22**

5.1 Simulação determinística somente com origens constitutivas 22

5.2 Simulações de Monte Carlo sem transcrição 24

5.2.1 Construção da distribuição de probabilidades de origens constitutivas e flexíveis 24

5.2.2 Resultados das simulações 24
5.3 Simulações de Monte Carlo com transcrição 27
5.4 Simulações de Monte Carlo com transcrição e origens dormentes 27
5.4.1 Produção da distribuição de probabilidades de origens dormentes . 27

6 Conclusão **28**
6.1 Recapitulação do projeto e das contribuições do mesmo 28
6.2 Visão geral do aluno sobre o projeto 28
6.3 Histórico do projeto 28
6.4 Desafios superados ao longo da execução do projeto 29

Bibliografia **30**

Capítulo 1

Introdução

Em organismos eucariotos, a replicação de DNA consiste na duplicação dos cromossomos que compõem o material genético de uma dada célula. Tal processo ocorre essencialmente de forma paralelizada, iniciando-se diversas vezes ao longo da fase S do ciclo celular em sítios denominados "origens de replicação". Também durante o ciclo celular acontece o processo de transcrição de RNA ao longo do cromossomo. Propriedades como a probabilidade de ativação de cada origem de replicação e o momento da fase S em que elas são ativadas não são homogêneas, assim como não o é a atividade de transcrição em cada região do DNA. Há variações, ainda, na eficiência da maquinaria de replicação em duplicar o DNA a partir de cada uma dessas origens; o que, em muitos casos, é devido a colisões dessa maquinaria com a de transcrição. A dinâmica de colisão entre maquinarias de replicação e de transcrição deve variar entre espécies distintas, levantando a hipótese de que ela esteja por trás das diferenças na arquitetura genômica observadas entre as várias espécies de protozoários do grupo dos kinetoplastídeos (*TriTryps*) [Sil+17].

Vários esforços para investigar essa questão foram reportados: em 2012, utilizando *next-generation marker frequency analysis* (MFA-Seq), uma técnica ômica de larga escala, foi mapeada a distribuição das origens de replicação nos cromossomos de *Trypanosoma brucei* [Tie+12] TREU927 e, em 2015, de espécies do gênero *Leishmania* [Mar+15], todas elas do grupo dos kinetoplastídeos. Mais recentemente, sob a coordenação da Dra. Maria Carolina Elias, foi iniciado o Projeto Temático "*How do common and diverged features of the replicative stress response shape the biology of TriTryp parasites?*", com participação do Dr. Richard McCulloch (Universidade de Glasgow, Reino Unido). Através dessa colaboração, foi feito o mapeamento com MFA-Seq das origens de replicação em *Trypanosoma cruzi*, o kinetoplastídeo causador da Doença de Chagas. Apesar desse ensaio mostrar de forma imediata a frequência de disparo de cada origem de replicação, a dinâmica de colisão entre maquinarias de replicação e de transcrição, para *T. cruzi*, *T. brucei* e demais kinetoplastídeos, permanece como problema em aberto.

Este trabalho propõe-se a abordar esse problema através da construção de um modelo computacional capaz de simular a fase S do ciclo celular de kinetoplastídeos, para isso

utilizando dados biológicos como a distribuição das origens de replicação, as velocidades médias de replicação e de transcrição e os sítios de início de transcrição, permitindo a adequação do modelo aos resultados observados *in vitro*. Como estudo de caso, focamos na modelagem e simulação da replicação de *T. brucei* TREU927, o parasita responsável pela doença do sono. Com este trabalho, esperamos contribuir para um maior entendimento sobre a influência das colisões sobre a replicação de DNA como um todo, ao utilizar o modelo ajustado para prever o comportamento do sistema em diferentes situações.

1.1 Organização do trabalho

O restante desta monografia está organizado da seguinte maneira: no capítulo 2, explicaremos alguns conceitos biológicos importantes para o entendimento deste trabalho. No capítulo 3, apresentaremos a estrutura do modelo proposto e as regras que governam a sua dinâmica. No capítulo 4, percorreremos os detalhes computacionais do desenvolvimento e da implementação do modelo. Já no capítulo 5, analisaremos os resultados obtidos, comparando-os com resultados obtidos por experimentos *in vitro*; destacaremos ainda a contribuição desses resultados para o avanço da área. Finalmente, no capítulo 6 faremos uma avaliação do trabalho realizado durante este ano e das dificuldades para a criação do projeto e desta monografia.

Capítulo 2

Conceitos Fundamentais

O entendimento desta monografia envolve alguns conhecimentos essenciais graduação na área de Ciências Exatas; portanto, neste capítulo, abordaremos alguns desses conceitos a fim sobre biologia celular, que normalmente não são trabalhos em uma de facilitarmos a leitura dos capítulos seguintes para o público-alvo deste trabalho. Conceitos adicionais também serão apresentados nos capítulos seguintes, conforme forem sendo necessários.

2.1 O DNA, o genoma e o cromossomo

O DNA (abreviação de *deoxyribonucleic acid*) é a molécula que carrega as instruções do material genético de um organismo vivo. Suas informações controlam toda a dinâmica celular, incluindo seu crescimento, proliferação e resposta à estímulos externos. Grande parte dos recursos e da arquitetura de uma célula são voltados para a manutenção e a replicação de seu conjunto de moléculas de DNA, conhecido como genoma (figura 2.1).

O genoma de uma célula é formado por duas moléculas de DNA complementares entre si, formando um cromossomo. Cada uma dessas moléculas de DNA, também chamadas de fitas, é constituída de uma sequência de nucleotídeos (moléculas pequenas que funcionam como "tijolos" na montagem da molécula de DNA), como representado na figura 2.2. Nucleotídeos podem apresentar diferenças químicas entre si, o que nos permite classificá-los em 4 grupos, representados comumente pelas letras A, T, C e G. É sabido que, a fim de unir as fitas do DNA, ocorrem ligações, através de pontes de hidrogênio, entre grupos específicos desses nucleotídeos. Especificamente, nucleotídeos do grupo A ligam-se a nucleotídeos do grupo T, enquanto os do grupo C ligam-se aos do grupo G.

Além da estrutura e composição química do DNA, costumamos adotar um sistema de orientação para as molécula: as extremidades do cromossomo são marcadas com os valores 3' ou 5', de forma que sequencia do DNA seja lida sempre na direção $3 \rightarrow 5$ (figura 2.3).

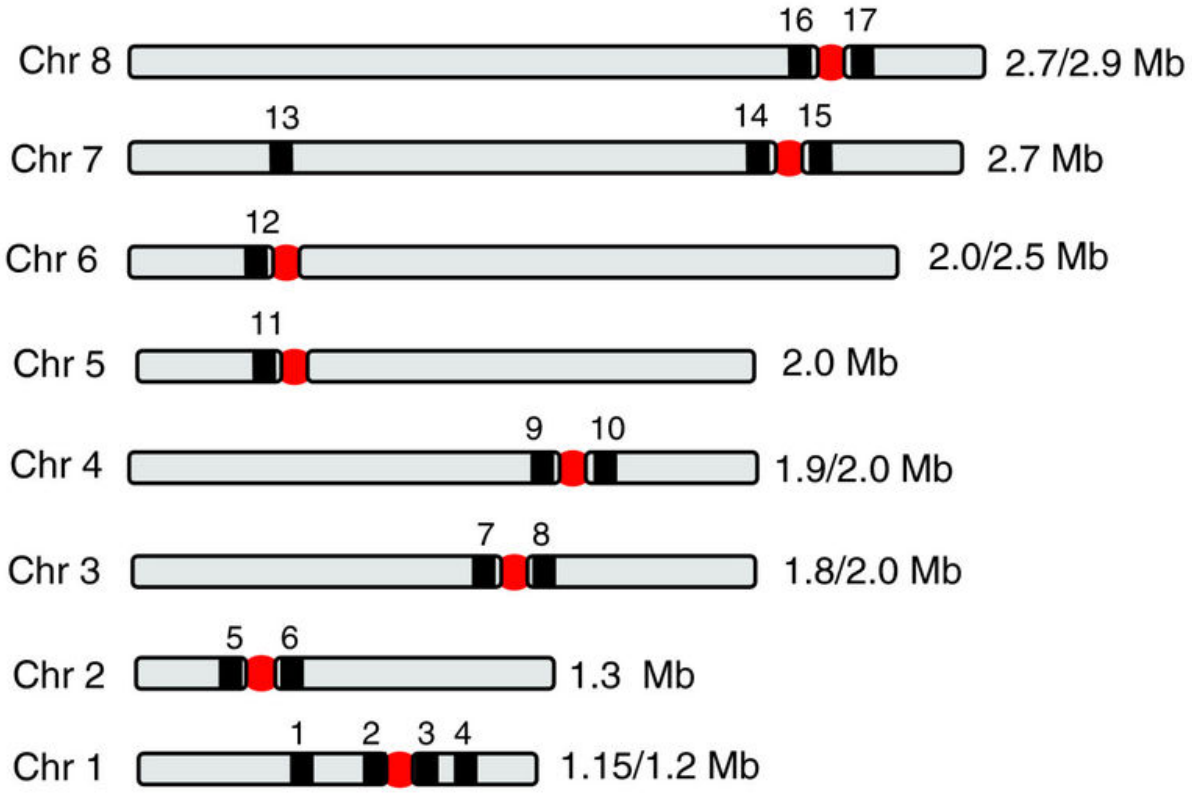


Figura 2.1: **Cariótipo do *Trypanosoma brucei*.** Número e estrutura dos cromossomos que formam o genoma do *Trypanosoma brucei*. Ao lado de cada cromossomo é apresentado seu comprimento, em milhares de bases. Figura extraída de Obado et al. [Oba+07].

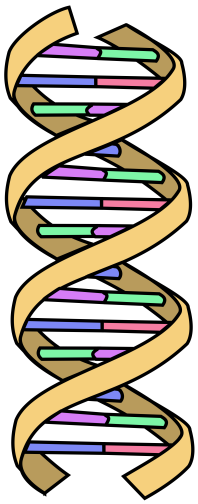


Figura 2.2: **Molécula de DNA.** Note sua estrutura de dupla-fita organizada em formato de hélice. Figura extraída de domínio público.

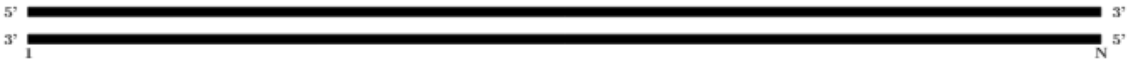


Figura 2.3: **Representação do cromossomo.** Representação da estrutura de cromossomo utilizada no modelo, com N pares de bases.

2.2 O ciclo celular

O ciclo celular é uma série de eventos que ocorrem em cada célula, culminando na duplicação de seu material genético e em sua divisão celular. A divisão celular é realizada por todos os organismos vivos, sendo essencial tanto para seu crescimento (no caso de organismos multicelulares) quanto para sua reprodução. Em organismos eucariotos (aqueles que possuem membranas envolvendo o material genético e outras partes internas da célula), como os kinetoplastídeos estudados neste trabalho, as fases que compõem o ciclo celular (figura 2.4) são:

1. *Gap 1* (**G1**): A célula cresce e se prepara para a duplicação do genoma;

2. *Synthesis* (**S**): A célula duplica seu material genético, sintetizando novas moléculas de DNA, idênticas à originais;

3. *Gap 2* (**G2**): A célula cresce e se prepara a a divisão celular;

4. *Mitosis* (**M**): A célula se divide em duas células filhas, com cada uma delas recebendo uma das duas cópias do material genético e entrando novamente em G1.

Quando estão em G1, células também podem entrar em um estado quiescente (G0), no qual ela não cresce nem se replica. Caso elas eventualmente saiam de G0, as mesmas retornam para G1 (figura 2.4). Nesta pesquisa, estamos interessados em estudar os processos que ocorrem durante a fase S do ciclo celular e como esses processos interagem entre si.

2.2.1 A fase S

Na fase S do ciclo celular ocorre a duplicação do material genético da célula. Durante esta fase, conjuntos de proteínas se ligam ao DNA, formando "maquinarias de replicação", também conhecidas como replissomas, as quais serão responsáveis por percorrer a dupla-fita de DNA, duplicando-a (figura 2.5). No fim da fase S, a célula terá o dobro do material genético inicial, quantidade necessária para a divisão celular que ocorrerá em seguida.

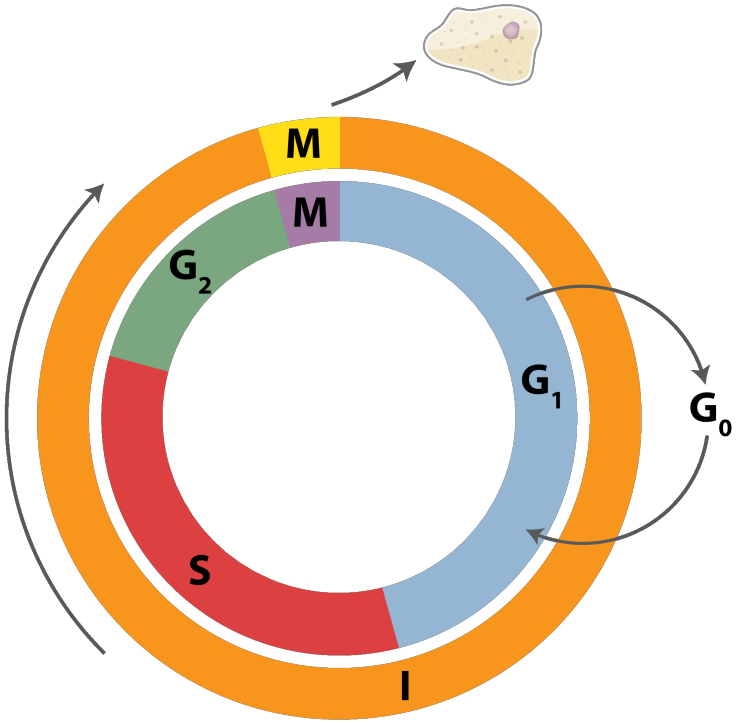


Figura 2.4: **Representação das fases do ciclo celular.** Durante o ciclo, a célula percorre as fases G1, S, G2 e M, nesta ordem. As fases G1, S e G2 são consideradas subdivisões da intérfase (segmento laranja). Figura extraída de domínio público.

2.3 As origens de replicação

Já foi mostrado que, na fase S da grande maioria dos organismos, replissomas não se ligam a qualquer região do DNA, mas sim a pontos específicos, cuja localização depende tanto de fatores filogenéticos (como a espécie do organismo), quanto de fatores mais complexos como o estado da molécula de DNA durante a replicação, incluindo a presença de danos sobre o cromossomo e a porcentagem da molécula já duplicada no momento da ligação [Shi+98]. A esses pontos damos o nome de **origens de replicação**, conceito que será amplamente abordado nos próximos capítulos.

2.3.1 Origens constitutivas, flexíveis e dormentes

Dependendo da frequência e das circunstâncias de disparo das origens replicativas, podemos agrupá-las em 3 categorias distintas:

1. **Origens constitutivas** disparam consistentemente nos primeiros momentos da fase S e são as mais visíveis em experimentos de mapeamento de disparo das origens.

2. **Origens flexíveis** são regiões do DNA com menor probabilidade de ligação de maquinarias de replicação, mas cujo disparo ainda é necessário para que a fase S

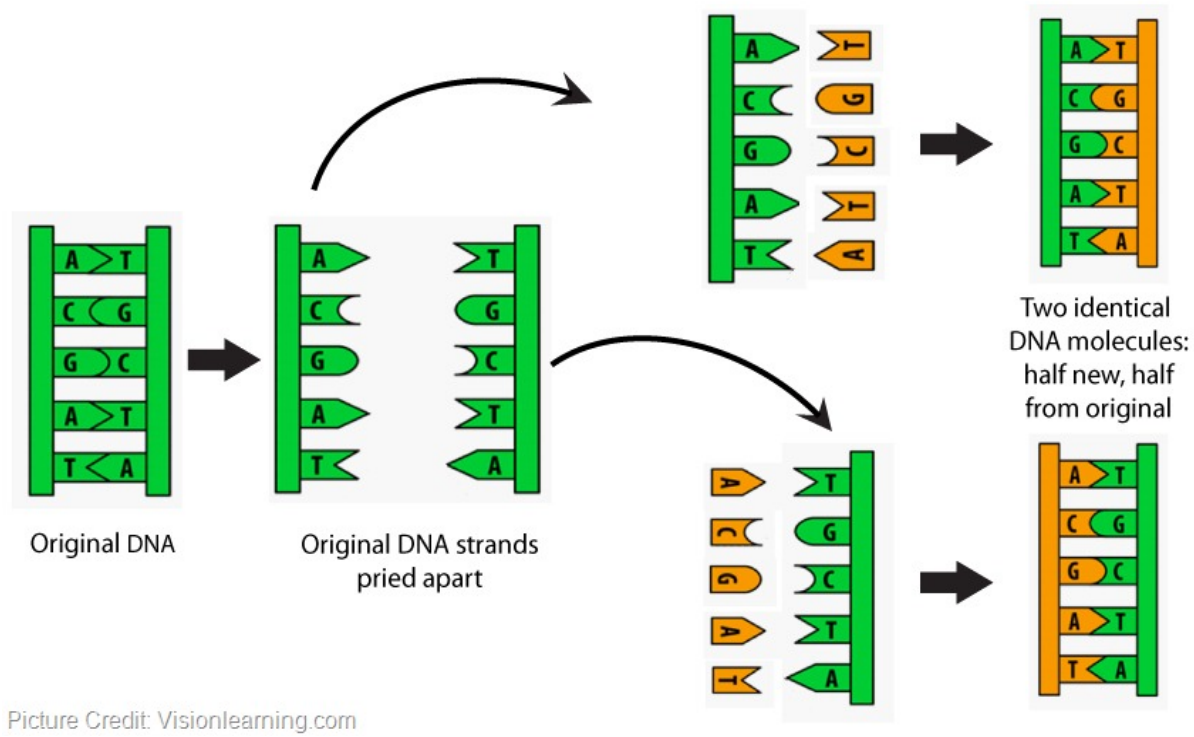


Figura 2.5: **Duplicação do material genético durante a fase S.** Proteínas se ligam à molécula de DNA e são responsáveis por separar suas duas fitas e por ligar os nucleotídeos que formarão a molécula nova.Durante o ciclo, a célula percorre as fases G1, S, G2 e M, nesta ordem. As fases G1, S e G2 são consideradas subdivisões da interfase (segmento alaranjado). Figura extraída de domínio público.

seja completa em tempo hábil. A denominação *flexível* vem exatamente do falta de consistência na posição de disparo destas origens.

3. **Origens dormentes** são aquelas que não disparam naturalmente durante a fase S, mas sim são induzidas à ativação na presença de falhas e danos ocorridos no processo de replicação.

No capítulo seguinte, ficará claro como nosso modelo trata cada uma destas categorias.

2.4 O RNA e o RNA mensageiro

Outra molécula de extrema importância para a célula, o RNA (*ribonucleic acid*) é responsável, entre outras funções, por expressar segmentos específicos do DNA, os genes. Tal expressão resulta em moléculas de RNA conhecidas como RNA mensageiros (mRNAs) contêm a informação genética do DNA em um formato que pode ser interpretado e utilizado pelos vários mecanismos das células. Por exemplo, mRNAs são utilizados por uma maquinaria macromolecular conhecida como ribossomo para a produção de proteínas, traduzindo a informação contida no mRNA em sequências específicas de aminoácidos.

2.5 A transcrição

Moléculas de mRNA são sintetizadas principalmente por uma proteína chamada RNA polimerase (RNAP). Uma RNAP percorre um dado gene, sintetizando "on-the-fly"a molécula de mRNA, utilizando para isso a sequência de nucleotídeos do gene como molde (figura 2.6).

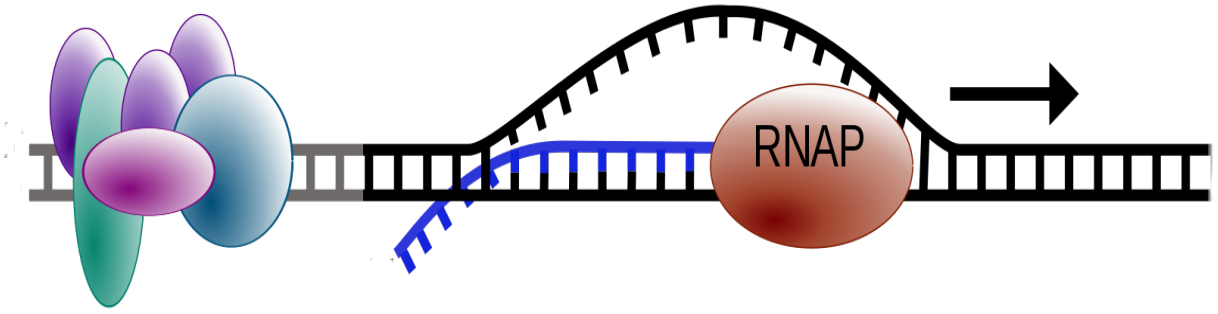


Figura 2.6: **Processo de transcrição.** RNAP (sigla para RNA polimerase, representada em laranja) é uma das proteínas responsáveis pela transcrição. Note a cadeia de RNA (azul) sendo criada *sobre* a cadeia de DNA (preto). As proteínas à esquerda estão realizando outras atividades sobre a molécula de DNA. Figura extraída de domínio público.

2.5.1 Transcrições policistrônica e constitutiva

Durante o desenvolvimento deste trabalho, fizemos duas assunções importantes sobre a transcrição, com base em informações sobre o comportamento desse processo no *Trypanosoma brucei*.

- A transcrição é **policistrônica**, ou seja, cada RNA mensageiro pode ser transcrito a partir de múltiplos genes consecutivos. Portanto estamos mais interessados na análise das regiões de transcrição que na localização de cada gene individual.
- A transcrição é **constitutiva**, ou seja, ela ocorre durante todo o ciclo celular com uma frequência constante (a frequência da transcrição representa o número de vezes que maquinarias de transcrição ligam-se às regiões policistrônicas, por minuto).

2.6 Conflitos replicação-transcrição

Tendo em vista os processos estudados anteriormente, supondo que a replicação e a transcrição ocorram de forma concomitante ao longo da fase S do ciclo celular, é evidente que esses processos podem entrar em conflito entre si. De fato, quando as maquinarias de replicação (replissoma) e transcrição (RNAP) se aproximam uma da outra sobre uma fita de DNA, elas se influenciam negativamente, levando a problemas (algumas vezes irreparáveis) na duplicação do material genético.

No decorrer desta monografia, daremos o nome de "colisão" a essa influência negativa entre as maquinarias. Em especial, se as maquinarias estiverem se movendo em sentidos opostos sobre o DNA no momento do conflito, a colisão será "frontal", ou "head-to-head", já se as maquinarias estiverem se movendo no mesmo sentido, a colisão será "não-frontal", ou "head-to-tail". As colisões frontais são as maiores responsáveis pelo dano causado aos cromossomos durante a fase S e, portanto, será mais estudada neste trabalho.

Capítulo 3

Um modelo da dinâmica de replicação de *T. brucei* TREU927

Neste capítulo, abordaremos o modelo proposto para simular a dinâmica de replicação de *T. brucei* TREU927. Na criação desse modelo, selecionamos aspectos da estrutura do genoma e da dinâmica do ciclo celular necessários para o desenho de um sistema dinâmico que capturasse aspectos essenciais da replicação de DNA desse parasita. Com isso, esperávamos aproximar nosso modelo às medidas experimentais obtidas em ensaios *in vitro* e *in vivo* e, assim, validar os resultados obtidos.

3.1 Representação da estrutura do genoma

Para a realização deste trabalho, nos restringimos ao processo de replicação do genoma que ocorre ao longo da fase S do ciclo celular. Mais especificamente, focamos da descrição das estruturas e propriedades que dizem respeito à replicação e transcrição do material genético, que serão apresentadas a seguir.

Conforme apresentamos no capítulo 2, o genoma é o conjunto de todos os cromossomos que compõem o material genético de uma dada célula. Cada simulação de nosso modelo é aplicada sobre um instância do genoma do *T. brucei* TREU927 e, dessa forma, todos os resultados obtidos são relacionados ao genoma como um todo.

O cromossomo é a unidade formadora do genoma e sobre sua estrutura ocorrem todos os processos relevantes para nosso estudo. Em nosso modelo, uma vez que podemos guardar a informação sobre em qual fita pertence uma dada região policistrônica, podemos representar cada cromossomo por um único vetor, cujo comprimento é o número de pares de bases que o formam (figura 2.3).

Durante uma simulação, cada cromossomo é percorrido em ambos os sentidos por vários elementos dinâmicos que pertencem a um dos dois tipos: as maquinarias de replicação e de transcrição.

3.1.1 Maquinarias de replicação

A fim de realizar a replicação do DNA, várias maquinarias conhecidas por replissomas se ligam ao cromossomo (o par de bases em que ocorreu a ligação é conhecido como "origem de replicação") e percorrem a fita de DNA em ambos os sentidos, formando duas forquilhas replicativas (figura 3.1). Dessa forma, em uma simulação representamos a replicação de um par de bases simplesmente atualizando o vetor com um valor que indica que houve replicação desse nucleotídeo naquele instante de tempo desde o início da simulação; é interessante fazer dessa forma, pois assim podemos recuperar facilmente o histórico da dinâmica de replicação do cromossomo. Além disso, ao longo de uma simulação, também precisamos guardar a localização corrente de cada um dos replissomas. Em nossas simulações, trabalhos com um número máximo de replissomas que podem estar disparados ao mesmo tempo, de maneira similar à feita por Gindin e colegas [Gin+14]; com isso, emulamos a restrição da disponibilidade desse recurso no interior da célula.



Figura 3.1: **Representação do processo de replicação.** Neste exemplo, colocamos no cromossomo, de forma arbitrária, duas origens de replicação, θ_1 e θ_2. Note como as forquilhas replicativas (ovais azuis) partem das origens e percorrem o cromossomo nos dois sentidos, duplicando-o (segmento verde).

3.1.2 Maquinarias de transcrição

Essas maquinarias, cujo principal representante é a RNAP, também se ligam ao cromossomo; entretanto, os pontos do cromossomo em que ocorrerão ligações da RNAP são regiões imediatamente anteriores ao início de regiões policistrônicas. Como a transcrição de *T. brucei* TREU927 é constitutiva, a frequência de disparo de RNAPs em cada uma das regiões policistrônicas é constante (figura 3.2). Ao longo de uma simulação, precisamos guardar a localização corrente de cada uma das RNAPs. Quando uma RNAP chega ao final da região policistrônica, ela é removida da simulação. Em nosso modelo, não colocamos um limite máximo de RNAPs operando simultaneamente.

3.2 Representação da dinâmica de replicação de DNA

Devido a nosso interesse nos detalhes do processo de replicação do DNA, limitamos este estudo à fase S do ciclo celular, com as simulações capturando dados desde o momento da agregação da primeira maquinaria de replicação (em outras palavras, a ativação da primeira origem) até a duplicação completa do material genético.



Figura 3.2: **Representação do processo de transcrição.** As RNAPs, ou seja, as maquinarias de transcrição (círculos vermelhos), estão contidas às regiões de transcrição, ou seja, às regiões policistrônicas (setas amarelas). RNAPs se ligam no início da região e se movem num único sentido (indicado pela seta) para seu fim. Note que múltiplas RNAPs podem percorrer a mesma região de transcrição simultaneamente, pois cada região policistrônica dispara RNAPs a uma frequência fixa (i.e., elas têm transcrição constitutiva).

A partir do momento em que uma maquinaria de replicação se liga ao DNA, duas forquilhas partem do par de base da ligação e começam a percorrer o cromossomo em sentidos opostos, duplicando a fita numa velocidade média de aproximadamente 64 pares de bases por segundo [Cal+15]. Simultaneamente, maquinarias de transcrição percorrem as regiões de transcrição do cromossomo (cada região de transcrição é percorrida em um sentido específico) a uma velocidade aproximada de 40 pares de bases por segundo [Cal+15].

O disparo das origens de replicação está limitado à disponibilidade de recursos da célula: o número máximo de forquilhas percorrendo o DNA em um determinado instante é um dos parâmetros de grande importância de nosso modelo (utilizamos desde uma quantidade pequena de recursos, por volta de 5 forquilhas simultâneas, até quantidades maiores do que o esperado para o genoma humano, aproximadamente 60 forquilhas [Gin+14]). Analogamente, o processo de transcrição também está sujeito aos recursos celulares. Essa restrição foi traduzida, no modelo, através do controle da frequência com que maquinarias iniciam o processo de transcrição em cada região (variamos a frequência entre 0 e 1 disparos por minuto, de acordo com fontes de nosso laboratório).

Embora duas forquilhas replicativas possam colidir entre si, esse processo é esperado durante a fase S: ele representa o fim da duplicação de um segmento do DNA e é seguido pelo desprendimento das forquilhas, sem dano ao DNA (figura 3.3). As regiões de transcrição em uma fita do DNA são disjuntas, portanto todas maquinarias de transcrição percorrendo a mesma região se movem na mesma direção, garantindo que não haja colisão entre essas maquinarias.

Todavia, como discutimos nos capítulos anteriores, um dos pontos de estudo centrais de nosso modelo são os conflitos existentes entre os processos de transcrição e replicação que ocorrem durante a fase S, principalmente devido a colisões entre as várias maquinarias percorrendo a fita de DNA concomitantemente. No modelo, as possíveis naturezas dos conflitos e seus respectivos tratamentos, tendo como base dados biológicos obtidos em nosso laboratório, são:

- **Colisão Frontal (*head-to-head*).** Ocorre quando uma forquilha replicativa encontra uma maquinaria de transcrição se movendo em sentido oposto (figura 3.4).

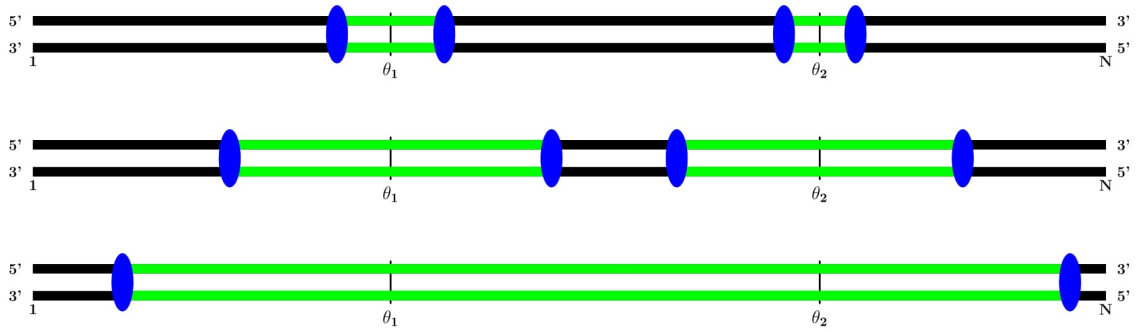


Figura 3.3: **Encontro de forquilhas replicativas.** Ao se encontrarem, as forquilhas se desligam do DNA e podem ser usadas futuramente no disparo de outras origens.

A colisão frontal leva a um grande estresse sobre a estrutura do DNA, em geral causando danos a sua estrutura. Neste caso, a maquinaria de transcrição é retirada da fita de DNA, levando à falha desta transcrição. Da mesma forma, a forquilha replicativa é incapaz de continuar a duplicação do DNA e se desliga, ficando disponível para ser dispara em outro ponto do genoma (no mesmo ou em outro cromossomo). A duplicação do DNA só será completa caso forquilhas provindas de outra origem de replicação atinjam a região do conflito; a vinda de tais forquilhas pode ou não ser induzida em resposta ao dano deste tipo de colisão: testamos essas duas possibilidades nos experimentos computacionais que serão apresentados no capítulo 5.

- **Colisão Não-Frontal (*head-to-tail*).** Ocorre quando uma forquilha replicativa encontra uma maquinaria de transcrição se movendo no mesmo sentido (esta colisão é possível pois a velocidade média de transcrição é menor que a de replicação, como mostrado na figura 3.5). Em conflitos desta natureza, a maquinaria de transcrição também é retirada do DNA, porém o menor dano ao DNA permite que a forquilha replicativa continue o processo de duplicação do DNA normalmente.

3.3 Definição da distribuição de probabilidade de disparo de origens de replicação

Durante o desenvolvimento do modelo, foi necessário estipularmos uma algoritmo não-arbitrário de decisão para as regiões do cromossomos onde devem ocorrer a ligação das maquinarias replicativas (origens de replicação). Num primeiro momento, focamos nos resultados de experimentos biológicos de *MFA Sequencing* [Tie+12], a partir do qual apontamos pontos específicos do DNA como origens constitutivas (como abordado no capítulo anterior, são origens que consistentemente disparam nos primeiros momentos da fase S).

Utilizando somente esses resultados, entretanto, não fomos capazes de adequar o tempo

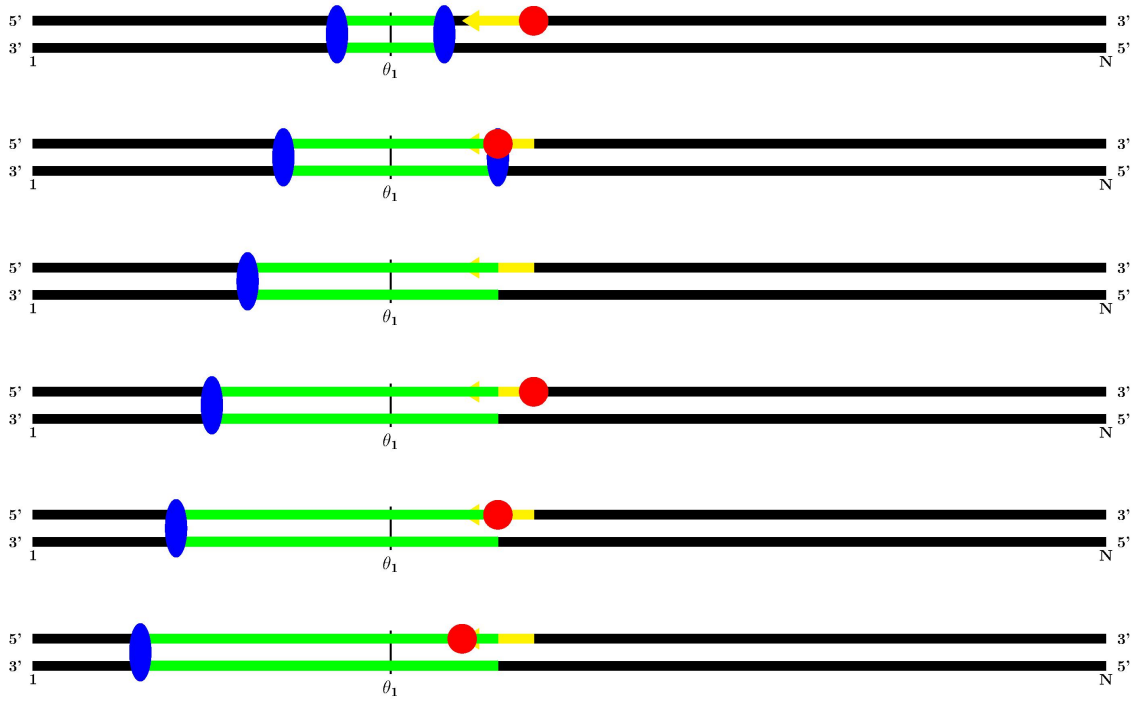


Figura 3.4: **Representação da colisão frontal entre maquinarias.** A forquilha replicativa movendo-se para a direita colide frontalmente com a maquinaria de transcrição. Isso faz com que ambas sejam retiradas da fita de DNA. Outras maquinarias de transcrição podem surgir e transcrever a região onde houve o conflito.

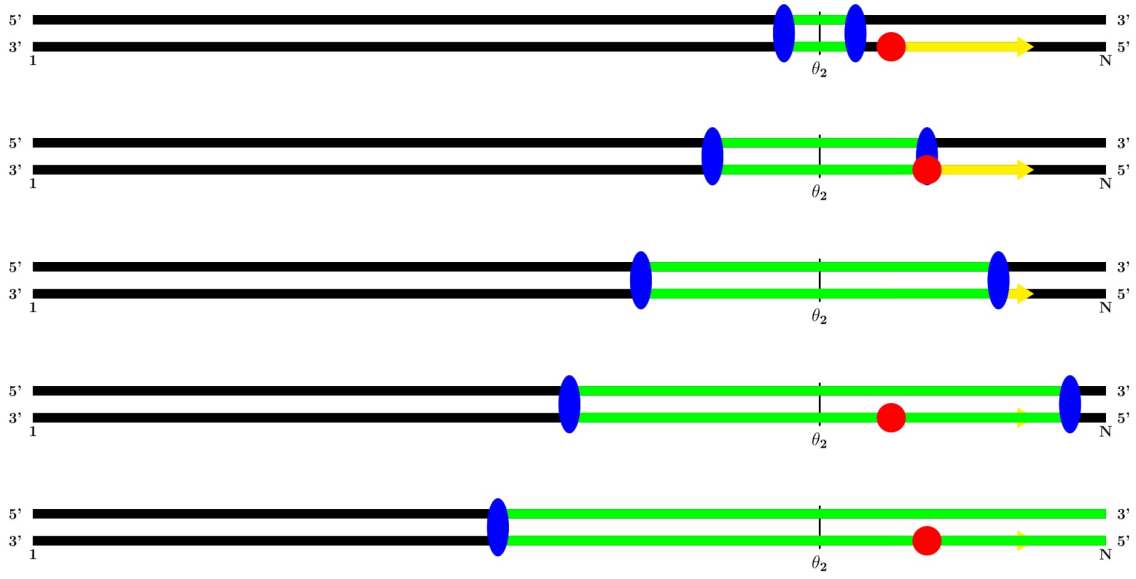
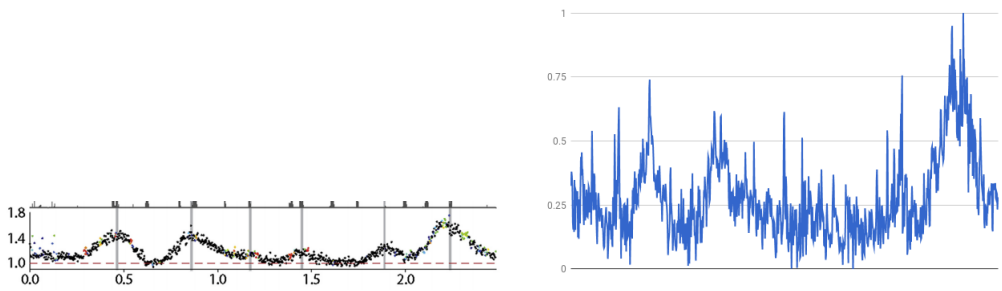


Figura 3.5: **Representação da colisão não-frontal entre maquinarias.** A forquilha replicativa movendo-se para a direita colide com a maquinaria de transcrição movendo-se no mesmo sentido. Isso faz com que a maquinaria de transcrição seja retirada da fita de DNA, mas a forquilha replicativa procede normalmente. Novas maquinarias de transcrição podem surgir sobre a região já duplicada do cromossomo.



(a) (b)

Figura 3.6: **(a)** Exemplo de resultados do experimento de MFA-Seq para um cromossomo (cromossomo 8 do *T. brucei REU927*). **(b)** Distribuição de probabilidade obtida através de uma transformação linear sobre os resultados do experimento. Note a acentuação dos picos e vales, feita para a melhor tradução do modelo. Em ambas as figuras, o eixo horizontal representa as bases da fita de DNA.

de fase S ao obtido nos experimentos *in vitro*. Com isso em mente, decidimos aprimorar o modelo através da aplicação de conceitos utilizados no modelo de Gindin *et al.* [Gin+14]. Ao invés de forçar o disparo de origens em algumas regiões, estipulamos duas propriedades para o genoma:

1. O número máximo N de forquilhas replicativas que podem estar ativas simultaneamente numa mesma célula.

2. Uma distribuição de probabilidade P que define, para cada base de um cromossomo, sua probabilidade de ativação em cada instante de simulação da fase S.

Em conjunto, essas propriedades efetivamente transformam nosso modelo em uma simulação de Monte Carlo, na qual as origens disparam majoritariamente ao redor dos picos de P. A limitação do número de forquilhas traduz de forma coerente o comportamento biológico da replicação, no qual a disponibilidade de material proteico é uma limitação para o avanço da replicação do material genético. Essa distribuição de probabilidade também foi obtida através de *MFA Sequencing*; entretanto, ao invés de garantir que os picos do experimento serão regiões constitutivas, a distribuição P simplesmente traduz os resultados do experimento de maneira estocástica.

Utilizando esta forma de modelagem, percebemos que a separação entre origens constitutivas e flexíveis é fluída: uma "origem constitutiva" é simplesmente uma origem cuja probabilidade de disparo ultrapassa um limiar estipulado, não havendo outras diferenças entre esses tipos de origens. As origens dormentes, por outro lado, estão fora da distribuição de probabilidade, sendo que seu disparo está condicionado exclusivamente a presença de colisões frontais ocorridas sobre o cromossomo. De fato, a ativação destas origens tem como objetivo a duplicação de regiões que ficariam privadas de maquinarias replicativas devido à danos causados ao DNA.

Capítulo 4

Implementação de um simulador do modelo dinâmico

Neste capítulo, apresentaremos os trabalhos visando a implementação de um simulador para o modelo dinâmico proposto no capítulo 3. Após o estudo extensivo de diversos outros trabalhos na área, organizamos a implementação em quatro etapas: i) coleta de informações *a priori* relevantes em repositórios públicos e na literatura, organizando-os em um banco de dados relacional; ii) especificação do simulador; iii) escolha de uma linguagem de programação apropriada; iv) implementação dos métodos e das estruturas de dados necessários para gerenciar a progressão de uma simulação. As decisões tomadas em cada uma dessas etapas foram pontos continuamente revisados durante o desenvolvimento deste trabalho.

4.1 Organização de informações *a priori* em banco de dados relacional

Para tornar o acesso das informações eficientes e, mais importante, facilitar o compartilhamento dos dados obtidos com outros modelos, optamos por utilizar um banco de dados relacional. Para o gerenciador desse banco, optamos pelo SQLite (figura 4.1), cujo ponto forte é a facilidade de acesso pela linguagem Python.

Uma vez definido o esquema do banco, tratamos de populá-lo com as informações necessárias para a simulação da dinâmica de replicação de *T. brucei* TREU927. Os dados para a modelagem da estrutura do genoma, incluindo o comprimento de cada cromossomo e as regiões de transcrição, foram obtidos do *TryTripDB*, um banco de recursos sobre kinetoplastídeos. As velocidades das forquilhas replicativas e de transcrição (RNAP) provêm de resultados obtidos por Calderano *et al.* [Cal+15], enquanto a duração média da fase S partiu de experimentos biológicos realizados em nosso laboratório por Santos *et al.* A distribuição de probabilidades e a distância interorigem médias foi obtida a partir

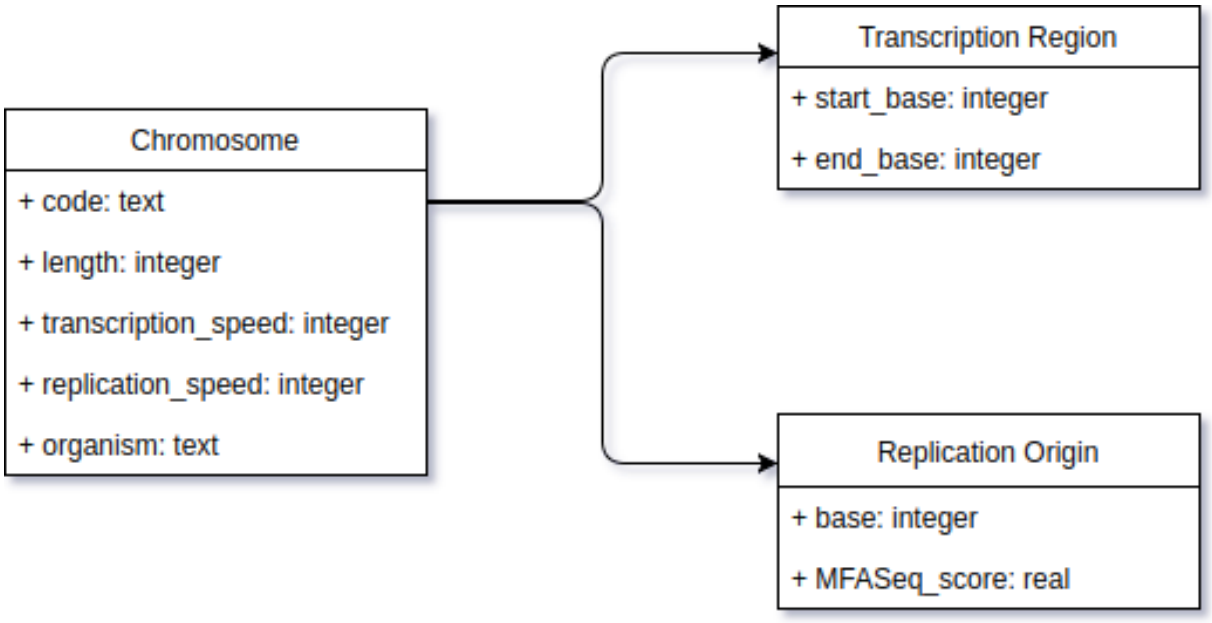


Figura 4.1: **Diagrama das Entidades do Banco.** Note que não é necessário modelar a entidade "genoma", uma vez que ele é o conjunto de todos os cromossomos com o mesmo organismo.

de dados de MFA-Seq de Tiengwe *et al.* [Tie+12].

4.2 Especificação do simulador

Na modelagem de processos biológicos, o uso de orientação a objetos permite estudarmos cada estrutura isoladamente e, num momento seguinte, estudar o comportamento do sistema quando essas estruturas interagem entre si. Dessa forma, mapeamos as estruturas definidas no capítulo 3 em objetos (figura 4.2). Faremos agora uma descrição das classes especificadas.

Classe *Genome*. Agrupa os cromossomos e é responsável por verificar o fim da simulação, que ocorre quando todos os cromossomos foram replicados.

Classe *Chromosome*. Responsável por armazenar e atualizar o estado dos pares de bases: quando uma forquilha replicativa percorre uma região de um cromossomo, este a duplica, marcando as bases que a compõem como duplicadas.

Classe *Fork*. Percorre um cromossomo, podendo ser instâncias das classes *Transcription Fork* ou *Replication Fork*, ambas derivadas de *Fork*. Como dito anteriormente, o movimento das forquilhas de replicação é registrado através da duplicação dos pares de bases, enquanto que o movimento da forquilha de transcrição não fica armazenado, contudo ele é imperativo para podermos avaliar as colisões entre as duas maquinarias.

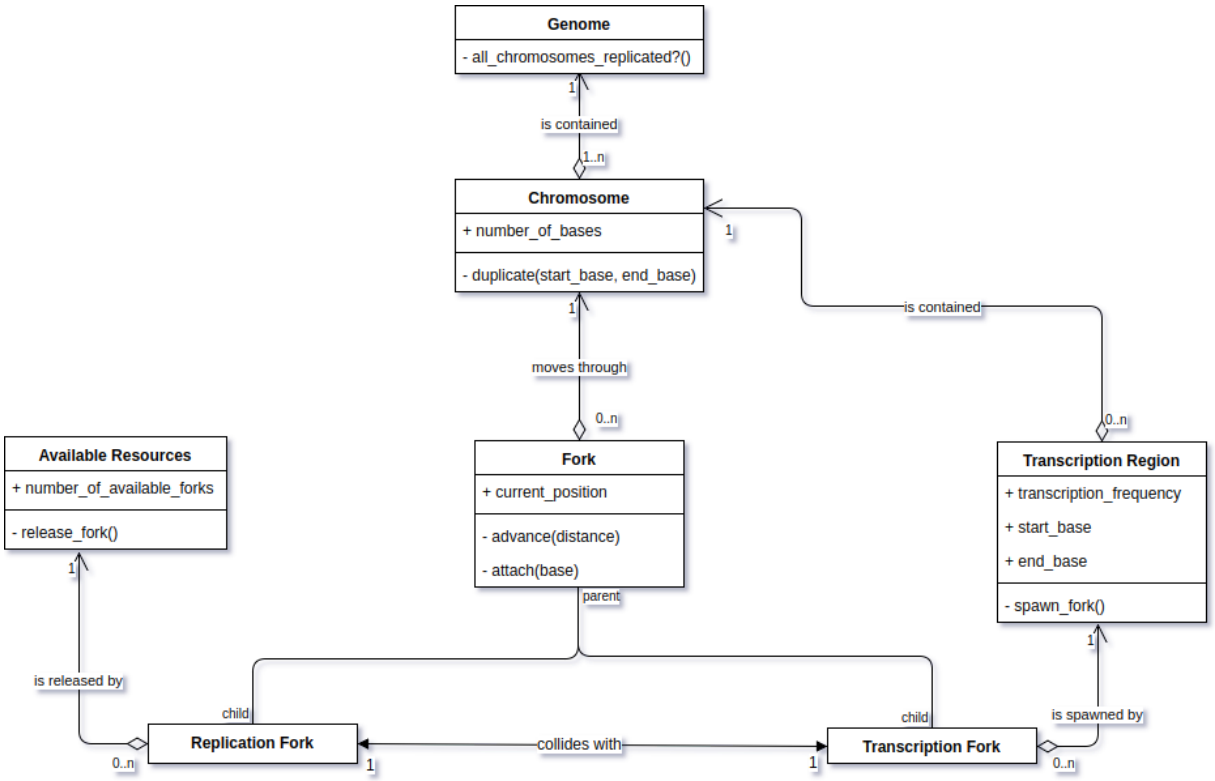


Figura 4.2: **Diagrama UML dos Classes Relevantes do Sistema.**

Além das classes acima, existem classes de gerenciamento, responsáveis por ligar forquilhas ao cromossomo no momento oportuno:

Classe *Transcription Region*. Representa cada região de transcrição de um cromossomo e, com uma certa frequência pré-definida, cria novas maquinarias de transcrição que percorrem a região (desde o par de bases inicial da região até o final).

Classe *Available Resources*. Libera forquilhas para se ligarem aos cromossomos. Note que cada região de transcrição é própria de um cromossomo (de fato, ela representa um segmento de bases daquele cromossomo); os recursos para forquilhas replicativas, por outro lado, são compartilhados por todos os cromossomos de um genoma. Dessa forma traduzimos mais precisamente a "competição" por replissomas que é característica da fase S do ciclo celular [Gin+14].

4.3 Escolha da linguagem de programação

Para escrevermos o código do projeto, escolhemos a linguagem de programação Python. A escolha foi feita principalmente pelo fato de Python permitir atender a quatro necessidades da implementação, a saber:

1. Desenvolver *scripts* para a análise e integração dos dados obtidos na literatura a nossos próprios. A variedade de formatos em que essas informações são encontradas requereu uma linguagem capaz de fazer o tratamentos de *strings* de forma rápida e natural;

2. Organizar as estruturas biológicas com orientação a objetos, de forma a isolar as características de cada entidade e limitar a interação entre diferentes estruturas, permitindo uma melhor adequação dessas interações ao comportamento observado experimentalmente;

3. Aproveitar a eficiência de poderosas bibliotecas matemáticas, como o *NumPy* e o *SciPy*, permitindo cálculos mais precisos e eficientes, principalmente no tratamento de colisões;

4. Paralelizar a execução das simulações, aumentando a eficiência e a escalabilidade do modelo.

Em alguns raros momentos do desenvolvimento, utilizamos as linguagens Perl, para analisar rapidamente alguns dos resultados das simulações, e SQL, para gerenciar diretamente o nosso banco de dados.

4.4 Execução das simulações

A execução do modelo envolve repetidas iterações sobre os processos de replicação e de transcrição, cada uma dessas iterações representando uma fração do tempo da fase S. Após a preparação dos cromossomos, o modelo imediatamente entre na fase S do ciclo celular, no qual dois processos principais acontecem: o avanço das maquinarias ligadas aos cromossomos e a ligação de novas maquinarias a esses.

O avanço das maquinarias, como exposto no capítulo 3, está sujeito a colisões de diversos tipos. Nosso sistema utiliza *Discrete Event Simulation*, ou seja, seu avanço ocorre em intervalos de tempo suficientemente pequenos, de forma que seja mínima a chance de ocorrência de múltiplos conflitos simultâneos. Assim, as colisões são tratadas logo após sua ocorrência, seguindo o comportamento estipulado para colisões frontais e não-frontais discutido no capítulo anterior (figura 4.3). Devemos considerar, ainda, que cada maquinaria avança somente até seus limites. Forquilhas replicativas movem-se até encontrarem uma das pontas do cromossomo, outra forquilha seguindo em direção oposta ou uma base já duplicada do cromossomo. Já maquinarias de transcrição estão limitadas às suas respectivas regiões de transcrição.

A ligação de novas forquilhas replicativas (caixa azul na figura 4.3) foi um dos pontos de maior discussão deste trabalho; a maneira definitiva que escolhemos para esse procedimento, similar à feita por Gindin e colegas [Gin+14], é apresentada na figura 4.4. A

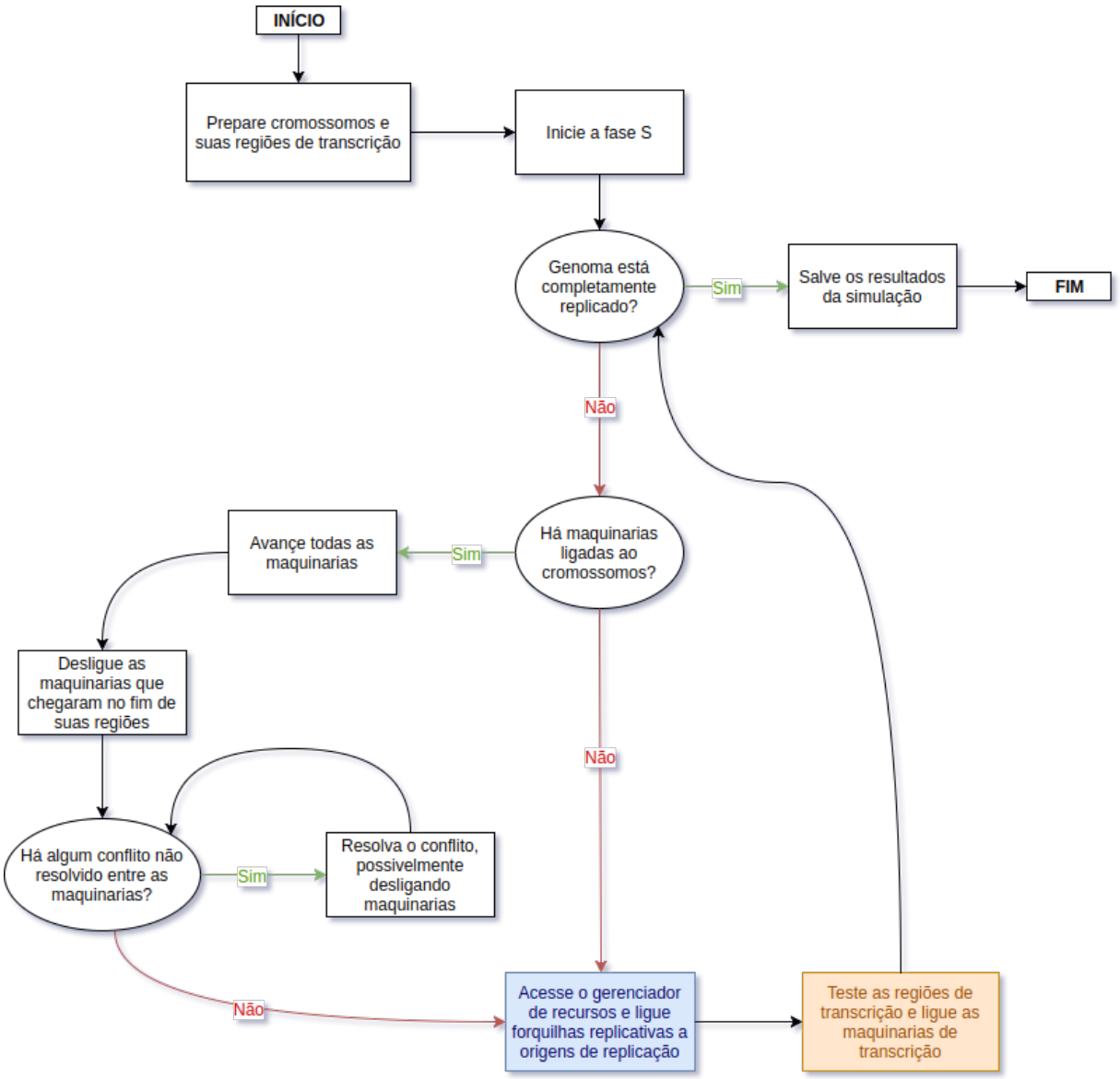


Figura 4.3: **Fluxograma Principal da Execução do Programa.** As caixas retangulares representam processos da execução, enquanto as ovais mostram pontos de testes. As caixas em azul e em laranja são explicadas em mais detalhes nesta seção.

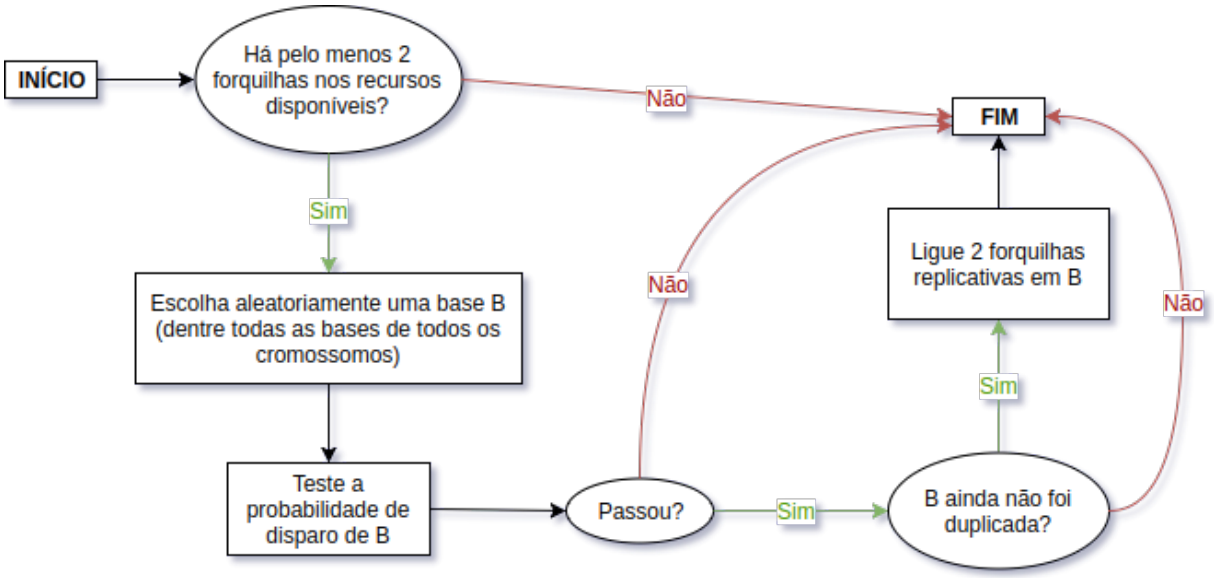


Figura 4.4: **Fluxograma do Processo de Ligação de Forquilhas Replicativas.** As caixas retangulares representam processos da execução, enquanto as ovais mostram pontos de testes.

cada intervalo de tempo pré-definido, o programa escolha uma base dentre todas as do genoma para ser disparada (em outras palavras, para receber forquilhas replicativas e se tornar uma origem de replicação). Para toda base do genoma está associada uma probabilidade de ativação que será testada quando aquela base for escolhida para o disparo. Ser escolhida e passar no teste de disparo, entretanto, não são suficientes para escolher a base como uma origem: somente bases ainda não duplicadas podem ser escolhidas como origens (caso contrário, forquilhas replicativas seriam ligadas a regiões já duplicadas, o que é biologicamente impossível).

A ligação de maquinarias de transcrição (caixa laranja na figura 4.3) é mais simples: a cada intervalo de tempo pré-definido, cada região de transcrição dispara uma maquinaria em sua posição inicial. Essa maquinaria percorre a região e é retirada ao atingir o fim daquela (ou possivelmente antes, no caso de algum conflito).

Capítulo 5

Resultados em experimentos computacionais

5.1 Simulação determinística somente com origens constitutivas

Num primeiro momento, restringimos nossa simulação às origens constitutivas apontadas pelo ensaio de MFA-Seq [Tie+12], as quais foram disparadas simultaneamente no início da simulação. Sob essas condições, mesmo com a ausência de transcrição, não fomos capazes de simular toda a replicação do material genético dentro do tempo de fase S medido experimentalmente. Esses resultados foram sumarizados na figura 5.1.

É importante notar que, embora alguns cromossomos tenham completado sua replicação dentro do tempo experimental, o tempo de duração da fase S é definido como o intervalo de tempo entre o primeiro disparo de origem até a replicação completada de todos os cromossomos. No experimento realizado, por exemplo, o tempo de duração da fase S seria de 28448 segundos (tempo de replicação máximo entre os cromossomos), aproximadamente 400% maior que a duração experimental.

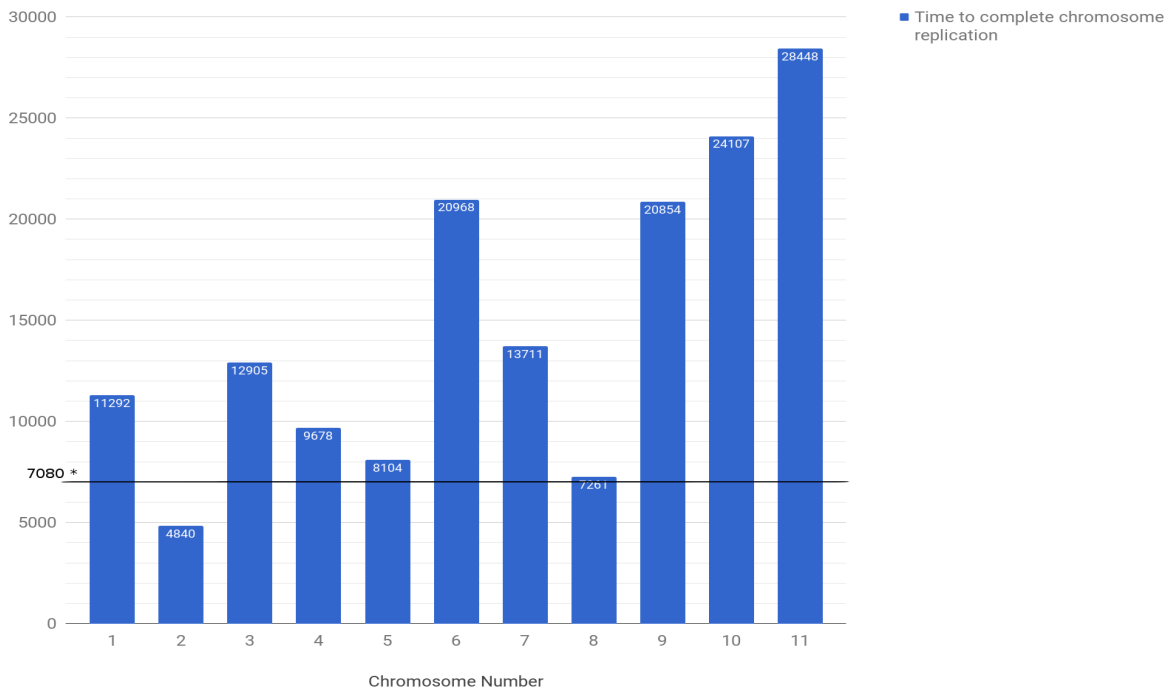


Figura 5.1: **Tempo da fase S simulada para cada um dos 11 cromossomos do *T. brucei*.** A linha em preto representa o tempo de fase S obtido experimentalmente para o organismo. *Novos experimentos em nosso laboratório corrigiram o tempo médio da fase S de 7080 segundos para 8300 segundos, o que ainda está muito aquém do obtido nesta simulação.

5.2 Simulações de Monte Carlo sem transcrição

A partir dos resultados anteriores, ficou evidente a necessidade de um sistema não-determinístico de escolha das origens de replicação. Com esse objetivo, partimos dos dados quantitativos do MFA-Seq para a criação de uma distribuição de probabilidade capaz de reger o sistema de disparo.

Estipulamos ainda, com base nos resultados obtidos por Gindin *et al.* [Gin+14], um limite para o número máximo de forquilhas que podem atuar simultaneamente sobre o genoma. Perceba que este limite aplica-se ao genoma como um todo, e não há cada cromossomo, já que o processo de replicação ocorre sobre todo o genoma da célula e há a necessidade de compartilhamento de recursos entre os cromossomos.

5.2.1 Construção da distribuição de probabilidades de origens constitutivas e flexíveis

Como discutimos anteriormente, os conceitos de origens constitutivas e flexíveis deixaram de ser definidos à força a partir deste ponto do modelo. A ligação das forquilhas replicativas ao DNA simplesmente segue uma distribuição de probabilidade obtida através dos experimentos de MFA-Seq (figura 5.2, linhas azuis). A medida que a simulação avança, diferentes pontos do cromossomo são escolhidos como possíveis regiões de disparo; nesse momento, a probabilidade definida pela distribuição é testada e, em caso de sucesso, aquele ponto torna-se uma origem de replicação, a partir do qual partem forquilhas replicativas nos dois sentidos da molécula de DNA.

5.2.2 Resultados das simulações

Ao realizarmos as simulações utilizando a distribuição de probabilidade construída, foi necessária a validação dos resultados, verificando quão bem foi feita a tradução desta distribuição para a efetiva replicação dos cromossomos.

Com esse intuito, registramos o instante da fase S simulada em que cada par de bases dos cromossomos foi replicada (figura 5.3). Dessa forma, fomos capazes de constatar que regiões com alta probabilidade de disparo realmente foram duplicadas mais cedo durante a fase S, em relação a regiões com baixa probabilidade (exemplo mostrado na figura 5.4).

Após esta validação, obtivemos resultados sobre 2 propriedades de interesse: o tempo médio de duração da fase S (TM) e a distância média entre origens consecutivas para o genoma (distância interorigens, DM). Através destes resultados, expostos na tabela 5.1, não fomos capazes de encontrar uma única quantidade de recursos N que satisfizesse tanto a duração da fase S (aproximadamente 8300 segundos) e a distância interorigens (aproximadamente 260 mil bases), obtidos *in vitro*.

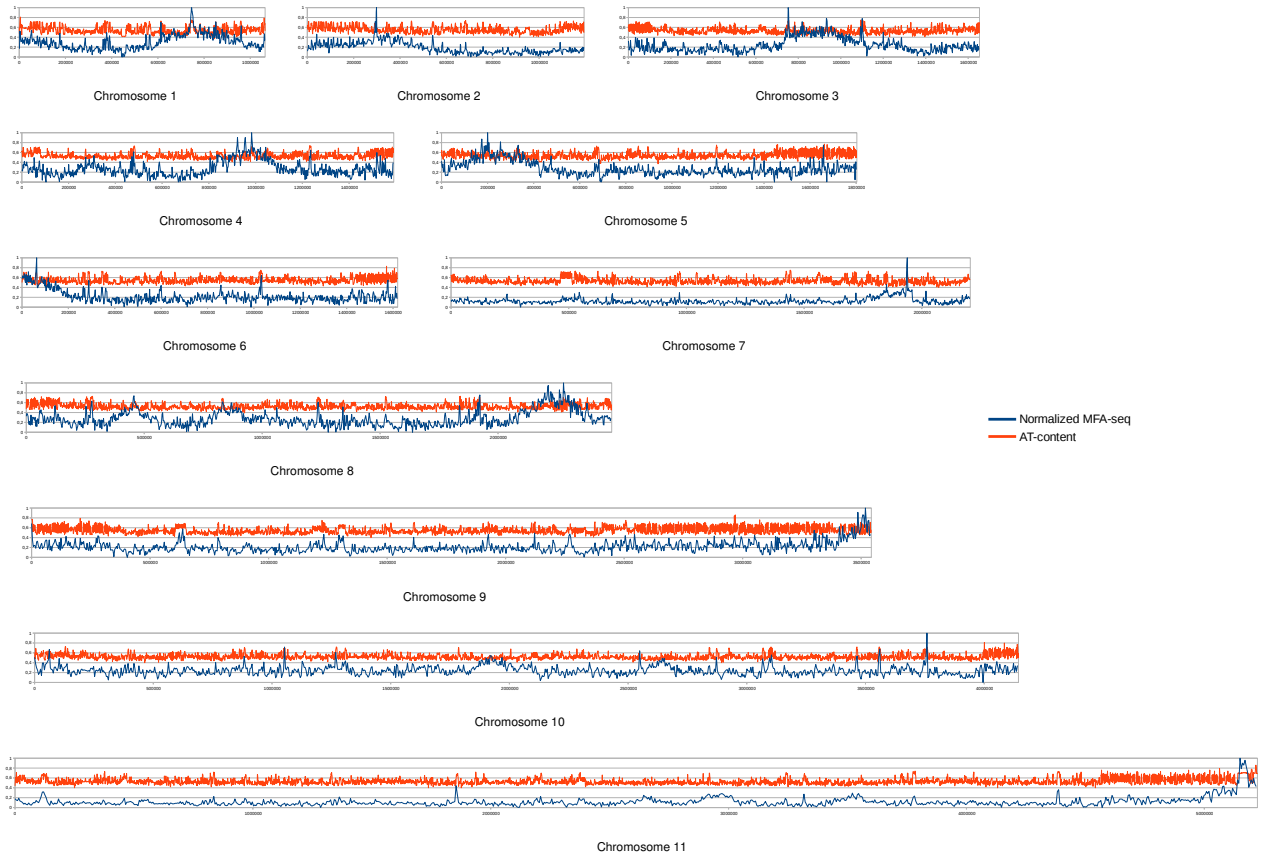


Figura 5.2: **Distribuição de probabilidade e conteúdo AT dos cromossomos.** Em azul, está representada a distribuição de probabilidade obtida a partir de dados de MFA-Seq para os cromossomos do *T. brucei*. Em laranja, está o conteúdo AT de cada região desses cromossomos.

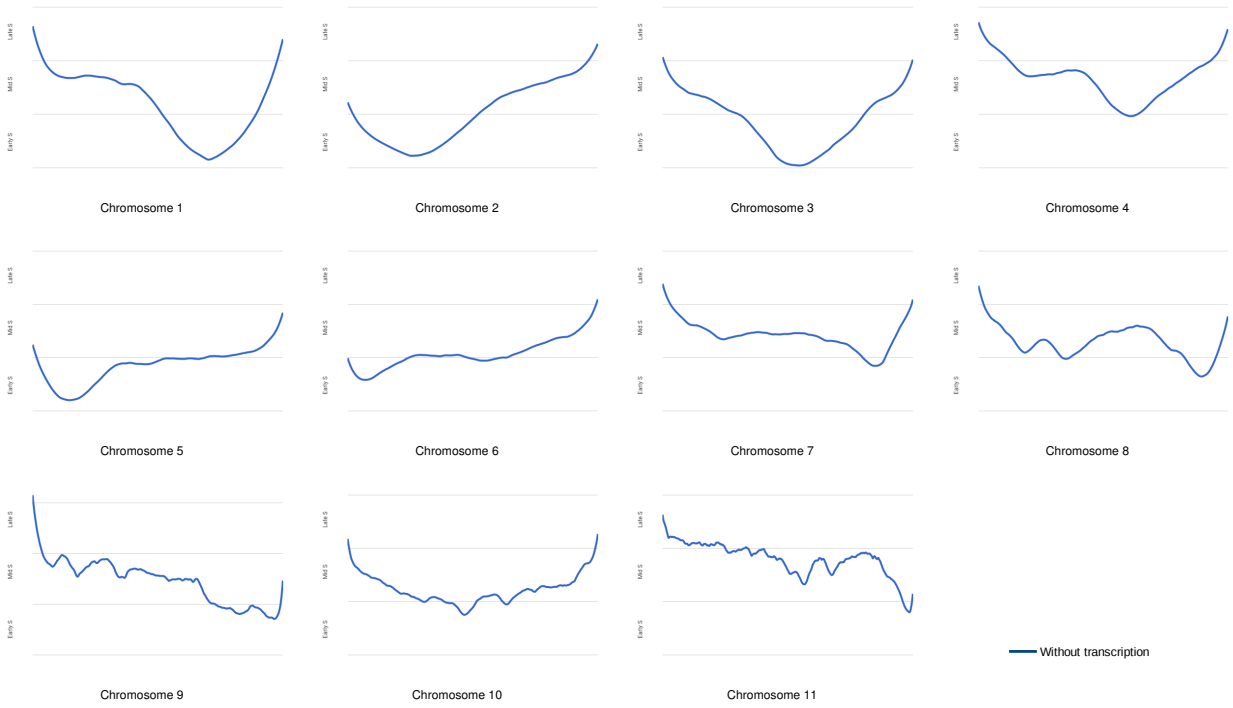
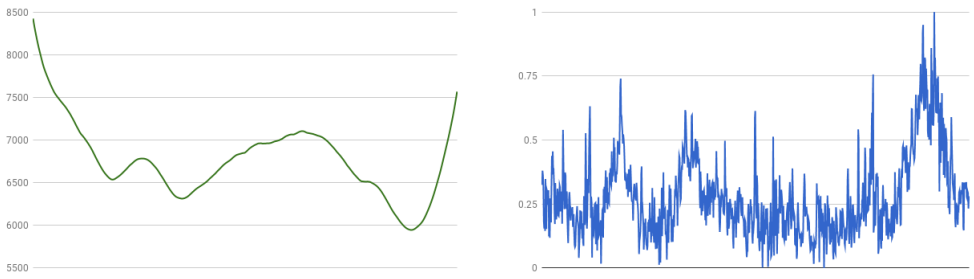


Figura 5.3: **Instante de replicação das bases dos cromossomos durante a fase S, na ausência de transcrição.**



(a) (b)

Figura 5.4: **(a)** Tempo de replicação de cada base do cromossomo 8, na ausência de transcrição. **(b)** Distribuição de probabilidade para as bases do cromossomo 8. Note que as regiões com alta possibilidade de disparo (picos em **(b)**) correspondem às regiões que replicaram mais cedo durante a fase S (vales em **(a)**).

Tabela 5.1: **Resultados das simulações na ausência de transcrição, após 1000 simulações.**

N	TM (segundos)	DM (bases)
10	**45227.4 ± 479.8**	**281796.7 ± 33630.6**
20	22074.4 ± 120.7	154564.9 ± 15092.5
30	14634.8 ± 62.1	106760.9 ± 8466.0
40	10955.3 ± 41.4	81950.3 ± 5803.1
50	**8765.9 ± 32.9**	**66746.3 ± 4150.1**
60	7307.5 ± 29.4	56411.1 ± 3386.2
70	6269.6 ± 27.0	49071.9 ± 2805.7
80	5492.9 ± 25.2	43324.0 ± 2263.2
90	4886.1 ± 23.2	38762.6 ± 1854.4
100	4405.9 ± 23.2	35180.6 ± 1605.8

Entretanto conseguimos estipular valores para N que respeitassem pelo menos um desses valores. Esses valores servirão como base para as simulações seguintes, na presença de transcrição.

5.3 Simulações de Monte Carlo com transcrição

5.4 Simulações de Monte Carlo com transcrição e origens dormentes

5.4.1 Produção da distribuição de probabilidades de origens dormentes

Capítulo 6

Conclusão

Neste capítulo recapitularemos os resultados obtidos, assim como listaremos os desafios que deverão ser enfrentados em uma eventual continuidade desta linha de pesquisa.

6.1 Recapitulação do projeto e das contribuições do mesmo

A fazer.

6.2 Visão geral do aluno sobre o projeto

O desenvolvimento deste trabalho foi, sem dúvida, minha melhor realização acadêmica. Através dele tive a oportunidade de estudar conteúdos em áreas distantes das que eu esperava abordar quando estava iniciando este curso de computação, além de poder me inserir no âmbito da pesquisa, experienciando a dinâmica das carreiras fora do mercado de trabalho. Este trabalho também serviu como uma ótima oportunidade para aprimorar minhas habilidades em desenvolvimento de software, apresentação e divulgação de projetos, organização e realização de atividades interdisciplinares.

6.3 Histórico do projeto

A preparação para o desenvolvimento deste projeto começou em agosto de 2016, quando, por indicação de um colega do curso, conheci o Dr. Marcelo da Silva Reis e começamos a estudar o tema da replicação celular. Abordamos desde os pontos mais gerais até as especificidades que seriam importantes para o desenvolvimento de nosso modelo.

Começamos, então, o trabalho árduo de desenvolvimento deste projeto, intercalado frequentemente por momentos de estudo e aprofundamento na literatura, com enfoque

tanto nos conceitos biológicos quanto na resolução dos problemas computacionais que surgiram ao longo do desenvolvimento.

No final de 2016, o trabalho foi apresentando com um pôster no IME-USP, como finalização da disciplina Apoio à Pesquisa. Já em outubro de 2017, outro pôster foi apresentado na *X-Meeting*, conferência internacional de bioinformática e, em novembro de 2017, mais uma apresentação foi feita na Reunião Científica Anual do Instituto Butantan. Todas essas apresentações reforçam a importância do trabalho e permitiram meu crescimento como aluno e pesquisador.

6.4 Desafios superados ao longo da execução do projeto

Ao longo da criação do modelo surgiram inúmeros desafios, principalmente sobre conhecimentos biológicos dos quais eu tinha pouca ou nenhuma base no início do desenvolvimento. As várias exposições do trabalho ao público científico, incluindo toda a preparação até a eloquência durante as apresentações, também foram dificuldades vencidas durante esses últimos meses.

Bibliografia

[Asl+10] Martin Aslett, Cristina Aurrecoechea, Matthew Berriman, John Brestelli, Brian P Brunk, Mark Carrington, Daniel P Depledge, Steve Fischer, Bindu Gajria, Xin Gao et al. «TriTrypDB: a functional genomic resource for the *Trypanosomatidae*». Em: *Nucleic acids research* 38.suppl 1 (2010), pp. D457–D462.

[Cal+15] Simone Guedes Calderano, William C Drosopoulos, Marina Mônaco Quaresma, Catarina A Marques, Settapong Kosiyatrakul, Richard McCulloch, Carl L Schildkraut e Maria Carolina Elias. «Single molecule analysis of *Trypanosoma brucei* DNA replication dynamics». Em: *Nucleic acids research* (2015), gku1389.

[Gin+14] Yevgeniy Gindin, Manuel S Valenzuela, Mirit I Aladjem, Paul S Meltzer e Sven Bilke. «A chromatin structure-based model accurately predicts DNA replication timing in human cells». Em: *Molecular systems biology* 10.3 (2014), p. 722.

[HF+04] Christiane Hertz-Fowler, Chris S Peacock, Valerie Wood, Martin Aslett, Arnaud Kerhornou, Paul Mooney, Adrian Tivey, Matthew Berriman, Neil Hall, Kim Rutherford et al. «GeneDB: a resource for prokaryotic and eukaryotic organisms». Em: *Nucleic acids research* 32.suppl 1 (2004), pp. D339–D343.

[LP12] Yea-Lih Lin e Philippe Pasero. «Interference between DNA replication and transcription as a cause of genomic instability». Em: *Current genomics* 13.1 (2012), pp. 65–73.

[Mar+15] Catarina A Marques, Nicholas J Dickens, Daniel Paape, Samantha J Campbell e Richard McCulloch. «Genome-wide mapping reveals single-origin chromosome replication in *Leishmania*, a eukaryotic microbe». Em: *Genome biology* 16.1 (2015), p. 1.

[Oba+07] Samson O Obado, Christopher Bot, Daniel Nilsson, Bjorn Andersson e John M Kelly. «Repetitive DNA is associated with centromeric domains in Trypanosoma brucei but not Trypanosoma cruzi». Em: *Genome biology* 8.3 (2007), R37.

[Shi+98] Katsuhiko Shirahige, Yuji Hori, Katsuya Shiraishi, Minoru Yamashita, Keiko Takahashi, Chikashi Obuse, Toshiki Tsurimoto e Hiroshi Yoshikawa. «Regulation of DNA-replication origins during cell-cycle progression». Em: *Nature* 395.6702 (1998), pp. 618–621.

[Sil+17] Marcelo Santos Silva, Paula Andrea Marin Muñoz, Hugo Aguirre Armelin e Maria Carolina Elias. «Differences in the Detection of BrdU/EdU Incorporation Assays Alter the Calculation for G1, S, and G2 Phases of the Cell Cycle in Trypanosomatids». Em: *Journal of Eukaryotic Microbiology* (2017).

[Tie+12] Calvin Tiengwe, Lucio Marcello, Helen Farr, Nicholas Dickens, Steven Kelly, Michal Swiderski, Diane Vaughan, Keith Gull, J David Barry, Stephen D Bell et al. «Genome-wide analysis reveals extensive functional interaction between DNA replication initiation and transcription in the genome of Trypanosoma brucei». Em: *Cell reports* 2.1 (2012), pp. 185–197.