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

Proposta para o Trabalho de Conclusão de Curso

Aluno: Gustavo Rodrigues Cayres Silva

Orientador: Marcelo da Silva Reis

Instituto de Matemática e Estatística
Universidade de São Paulo, São Paulo, 27 de abril 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. Esta proposta é uma atualização de uma proposta, enviada para a FAPESP e já aceita, contemplando o projeto de Iniciação Científica (processo 2016/17775-3) que será realizado sinergicamente a este Trabalho de Conclusão de Curso.

Sumário

1 Introdução **1**

2 Objetivos **2**

3 Metodologia **3**

3.1 Modelagem do sistema dinâmico estudado 3

3.2 Descrição matemática da dinâmica do modelo utilizando cadeias de Markov . 6

3.3 Implementação, simulação e testes do modelo dinâmico 7

4 Resultados Parciais **8**

5 Plano de Trabalho **10**

5.1 Descrição das atividades 10

6 Disseminação dos Resultados **11**

Referências **12**

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". Todavia, a frequência e o período de disparo de cada origem de replicação não são homogêneos. Além disso, há variações na eficiência da maquinaria de replicação em duplicar DNA a partir de cada uma dessas origens, o que, em muitos casos, envolve colisões dessa maquinaria com a de transcrição. A dinâmica de colisão entre maquinarias de replicação e de transcrição deve variar entre diferentes espécies, levantando a hipótese de que ela esteja por trás das diferenças de arquitetura genômica observadas entre diferentes espécies protozoários do grupo de kinetoplastídeos (*TriTryps*).

Um esforço para investigar essa questão foi reportado em 2015: 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 e de espécies do gênero *Leishmania*, todas elas do grupo dos kinetoplastídeos [1]. 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, nosso grupo fez 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, tanto para *T. cruzi* como para os demais kinetoplastídeos, permanece como problema em aberto.

2 Objetivos

Esta proposta de projeto contempla a construção de um modelo computacional estocástico a fim de descrever a dinâmica de replicação do DNA de *Trypanosoma cruzi*. Esperamos, com isso, entender diversos aspectos dessa dinâmica, em particular como as colisões entre as maquinarias de replicação e de transcrição influenciam o processo de replicação de DNA como um todo. Entender essas propriedades emergentes da dinâmica de replicação em *T. cruzi*, e em kinetoplastídeos em geral, é de grande relevância tanto para o projeto temático mencionado na introdução quanto para um dos subprojetos do CeTICS, o "Controle de ciclo celular e replicação de DNA: chaves moleculares", também coordenado pela Dra. Maria Carolina Elias e co-coordenado pelo Prof. Dr. Hugo Armelin.

O desenho do modelo dinâmico deverá levar em consideração a distribuição das origens de replicação, assim como as velocidades médias de replicação e de transcrição, os sítios de início de transcrição e outras informações a priori, como as extraídas de modelos cinéticos de redes de sinalização molecular. Inicialmente propomos modelar a dinâmica desse sistema utilizando cadeias de Markov.

Já a implementação do modelo será realizada utilizando a linguagem de programação *Python* e seguirá boas práticas de programação, das quais destacamos o uso de controladores de versão e de testes automatizados de unidade e de integração. Numa segunda etapa, pretendemos estender o uso desse modelo para outros kinetoplastídeos, em particular para *Trypanosoma brucei* e também para espécies do gênero *Leishmania*.

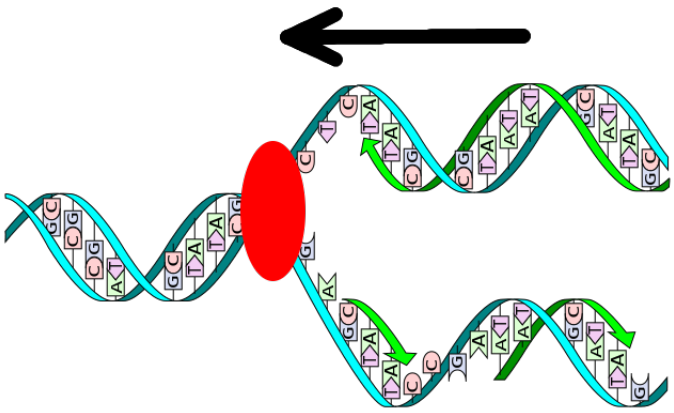


Figura 1: Forquilha de replicação de DNA. Conforme o replissoma (elipse em vermelho) avança no processo de replicação, no sentido indicado pela seta, a dupla hélice é desenrolada e cada fita simples atua como um molde (em azul) para a fita complementar que será sintetizada no processo de replicação (em verde). Figura adaptada da Wikipédia.

3 Metodologia

3.1 Modelagem do sistema dinâmico estudado

Conforme apresentamos na Introdução, a duplicação do material genômico de uma célula ocorre a partir do acoplamento de maquinarias de replicação de DNA (também chamadas de replissomas) a coordenadas específicas de cada cromossomo, denominadas origens de replicação; a partir destas últimas, pode ocorrer o fenômeno conhecido por *disparo de origem de replicação*: dois replissomas associam-se ao DNA e iniciam duas forquilhas de replicação (figura 1). Esses dois replissomas seguem em direções opostas, tomando como referência a coordenada no cromossomo da origem de replicação a partir da qual eles partem.

Todavia, ao longo da fase S do ciclo celular, a replicação não é o único fenômeno que envolve desenrolamento da dupla fita de DNA: também está presente o processo de transcrição, isto é, a produção de RNA mensageiro (mRNA). Em kinetoplastídeos a transcrição é policistrônica, ou seja, uma única região de transcrição inclui diversos genes. Durante esse processo, a maquinaria de transcrição, cujo componente mais importante é a RNA polime-

rase (RNAP), liga-se a uma região específica denominada região promotora; em seguida, é desenrolada a dupla fita de DNA, e a RNAP percorre a fita molde para codificar a molécula de mRNA (figura 2).

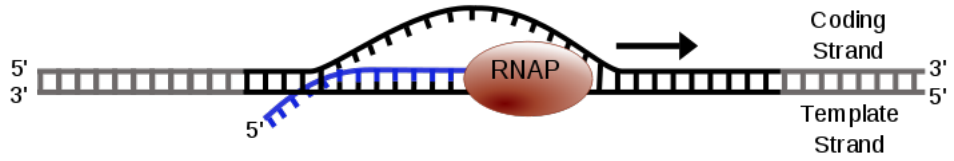


Figura 2: Esquema do processo de transcrição. Após o desenrolamento da dupla fita de DNA, a RNAP percorre a fita molde (*template strand*) no sentido 3' → 5', produzindo assim uma molécula de RNA contendo a sequência nucléica descrita pela fita codificadora (*coding strand*). Figura adaptada da Wikipédia.

Como disparos de origem de replicação ocorrem simultaneamente a diversas transcrições que são iniciadas ao longo de todo o genoma, há um risco constante de choque entre as maquinarias desses dois processos. Tais choques podem ser traseiros (*head-to-tail*) ou frontais (*head-to-head*). Nos choques traseiros, uma RNAP e um replissoma avançam sobre o cromossomo no mesmo sentido, com aquela adiante desta (figura 3(A)); a resolução dessa colisão é imediata, com o replissoma deslocando a RNAP do DNA (figura 3(B)) ou então simplesmente avançando sobre a mesma (figura 3(C)). Já nos choques dianteiros, uma RNAP e um replissoma avançam sobre o cromossomo em sentidos opostos, convergindo para uma mesma coordenada na qual ocorre o choque (figura 3(D)); a resolução deste tipo de conflito pode ocorrer com o deslocamento ou o *bypass* da RNAP após uma longa pausa (figura 3(E)), ou então com o colapso do replissoma (figura 3(F)), sendo que neste último caso é necessário que ocorra um novo disparo de origem de replicação para completar a replicação desse trecho do cromossomo [2].

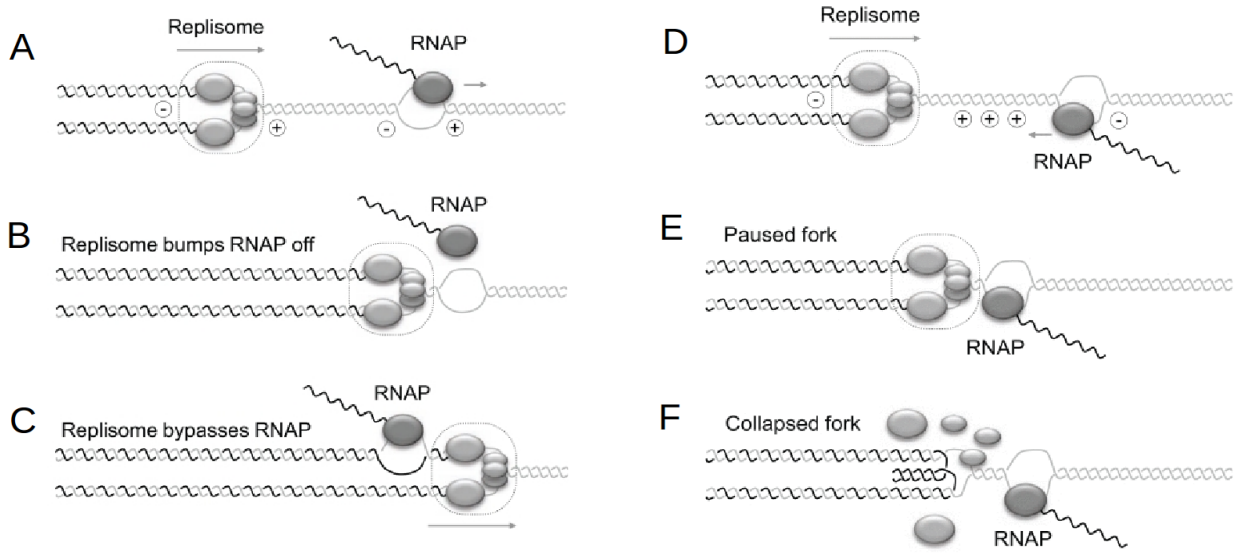


Figura 3: Os dois tipos de colisão entre as maquinarias de replicação e de transcrição: colisão *head-to-tail* (**A**) e suas duas possíveis resoluções (**B** e **C**); colisão *head-to-head* (**D**) e suas duas possíveis resoluções (**E** e **F**). Figuras adaptadas do artigo de Lin e colegas (2012) [3].

Além do mapeamento por MFA-Seq das origens de replicação o_i^j de cada cromossomo i em *T. cruzi*, outros parâmetros da dinâmica de replicação de DNA já foram determinados para esse kinetoplastídeo; por exemplo, em 2011, nosso grupo utilizou uma técnica conhecida por *Single Molecule Analysis of Replicated DNA* (SMARD) para calcular a velocidade média do replissoma em *T. brucei* [2]; mais recentemente, nosso grupo realizou esse mesmo procedimento para *T. cruzi*. Dessa forma, para compormos um modelo da dinâmica da transcrição em *T. cruzi*, podemos combinar essas informações com outras disponíveis em bancos de dados públicos, como o tamanho N_i do cromossomo i, associado a pares de coordenadas $\langle s_i^j, e_i^j \rangle$, que descrevem a coordenada s_i^j (e_i^j) do início (final) da região policistrônica j do cromossomo i. Na figura 4 apresentamos uma instância desse modelo.

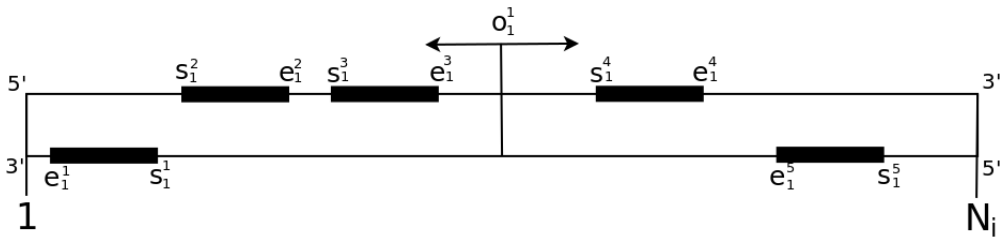


Figura 4: Exemplo de aplicação do modelo simplificado da dinâmica de replicação de DNA em *T. cruzi*. Esta instância tem um cromossomo ($j = 1$), uma origem de replicação (o_1^1) e cinco pares de coordenadas que descrevem regiões do genoma onde ocorrem transcrições. Observe que se $s_1^j < e_1^j$ ($s_1^j > e_1^j$) então a RNAP transcreve na direção da coordenada N_1 (1) do cromossomo.

3.2 Descrição matemática da dinâmica do modelo utilizando cadeias de Markov

Definida a estrutura do modelo dinâmico (figura 4), precisamos adotar uma descrição matemática da dinâmica em si do sistema. Para este fim, discretizaremos o tempo de simulação em um nível que seja suficiente para capturar todos os eventos importantes da dinâmica, em particular as colisões e suas resoluções. Além disso, a descrição matemática da dinâmica precisará contemplar as seguintes decisões:

1. Escolha do conjunto de origens de replicação a serem disparadas em cada simulação, que inclui origens constitutivas (disparadas mais frequentemente e apontadas pelo MFA-Seq) e as origens flexíveis (disparadas menos frequentemente e decididas aleatoriamente);

2. Escolha da frequência de disparo das transcrições;

3. Escolha do tempo máximo de simulação, o qual serve de indicador para o empacamento das forquilhas;

4. Escolha do tipo de resolução das colisões *head-to-head*, assim como o tempo necessário para tal (para colisões *head-to-tail*, num primeiro momento aproximaremos a resolução

como instantânea).

Como não podemos ter certeza da influência de estados passados do processo, é necessário utilizarmos cadeias de Markov para que possamos realizar as estimativas com vários graus de dependência passada. Mais especificamente, para um genoma com m regiões de transcrição e n origens de replicação, seja $\mathbf{x}_t \in \mathbb{R}^{m+2n}$ um vetor que indica, para um instante de tempo t, a coordenada no cromossomo onde se localiza a RNAP de cada região de transcrição e as coordenadas de cada par de forquilhas (potencialmente) disparadas de origens de replicação. Seja também ϕ_k uma função de transição que leva $k \geq 2$ vetores a vetores. Definimos a cadeia de Markov de memória k que governa a dinâmica de replicação como:

$$\mathbf{x}_t = \phi_k(\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-k}), \tag{1}$$

para qualquer $t \geq k$. Portanto, o desafio aqui é estimar ϕ_k para diferentes valores de k, utilizando para este fim as informações biológicas descritas anteriormente.

3.3 Implementação, simulação e testes do modelo dinâmico

Como mencionamos na seção 2, a implementação do modelo será realizada em *Python*, fazendo uso de Programação Orientada a Objetos (POO) para aproximar o modelo computacional do problema real. Na figura 5 mostramos uma proposta de como o modelo dinâmico pode ser descrito utilizando POO.

A simulação terá como entrada as informações biológicas mencionadas anteriormente, em particular as produzidas por nosso grupo e as extraídas de bancos de dados públicos tal como o TriTryDB (http://tritrypdb.org/tritrypdb/) [4]. A saída das simulações será constituída de todo o histórico da evolução da dinâmica de replicação, em particular o registro das colisões, o que inclui informações sobre o tipo, a coordenada onde a mesma ocorreu e o tempo de resolução que foi necessário.

A manutenção do bom funcionamento do modelo será realizada através da implementação de testes automatizados, tanto unitários quanto de integração, do uso do gerenciador de controle de versão *git*, e da aplicação geral de boas técnicas de programação.

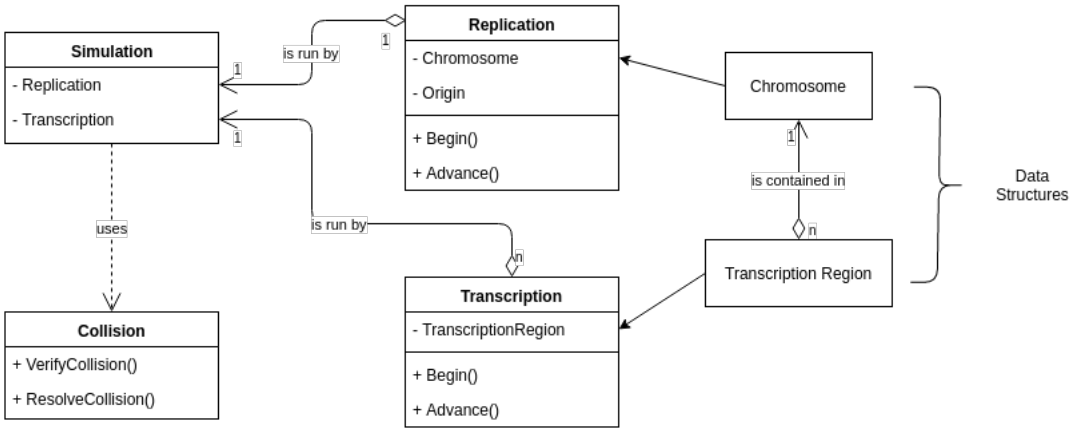


Figura 5: Diagrama de classes contendo uma proposta da descrição do modelo dinâmico utilizando Programação Orientada a Objetos (POO).

4 Resultados Parciais

O modelo foi inicialmente testado sobre o genoma nuclear de Trypanosoma cruzi (CL-Brener), baixado do repositório TriTrypDB. Através de experimentos iniciais com esse modelo aplicado a T. cruzi, conseguimos apontar algumas propriedades emergentes do sistema:

- Há uma correlação positiva (coeficiente de Pearson $\rho = 0{,}90$) entre o tempo total de simulação (i.e. a duração da fase S) e o tempo de reparo da replicação (figura 6).
- Notamos também uma correlação negativa ($\rho =$ -0,74) entre a duração total do processo e o intervalo entre a ativação de transcrições (figura 7).
- Com base nos itens anteriores, não foi surpreendente encontramos uma forte correlação ($\rho = 0{,}91$) entre o tempo da fase S e o número de colisões frontais (figura 8).

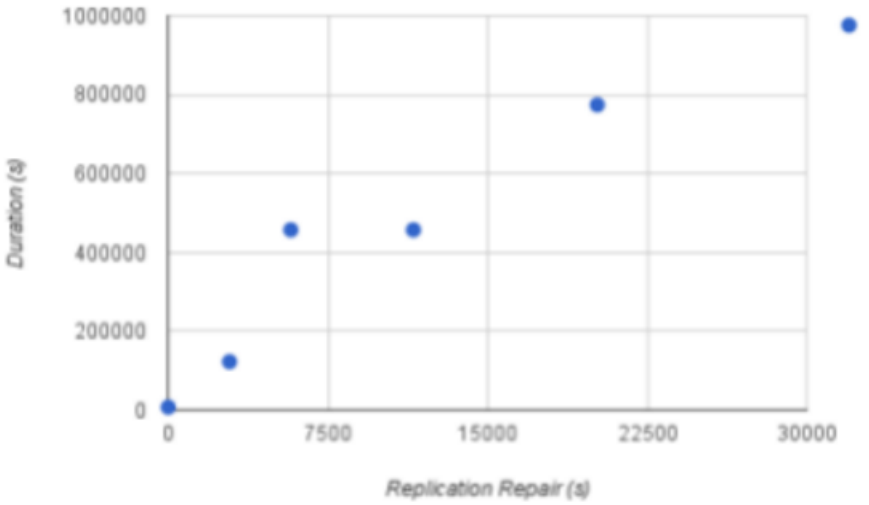


Figura 6: Duração da fase S em função do tempo de reparo das colisões frontais.

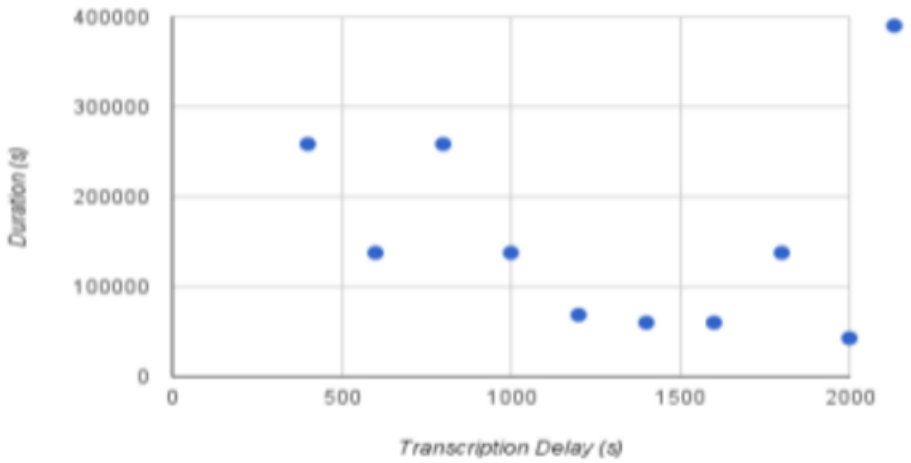


Figura 7: Duração da fase S em função do intervalo entre o disparo de transcrições.

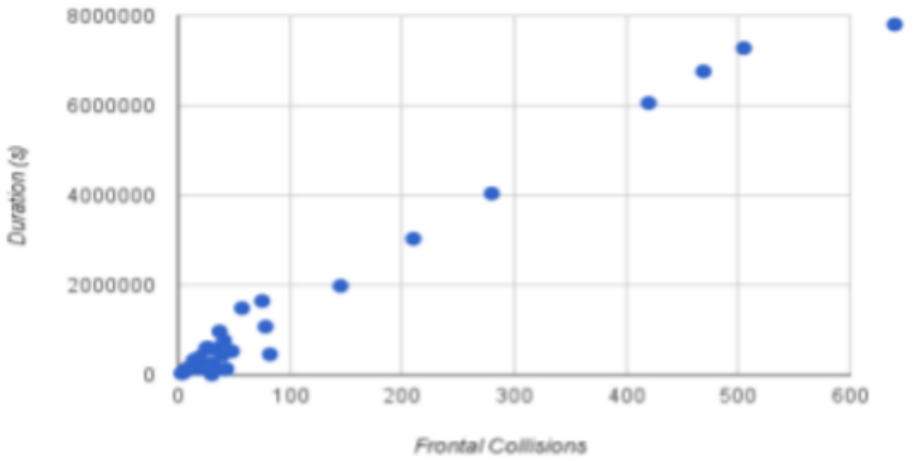


Figura 8: Duração da fase S em função do número de colisões frontais.

5 Plano de Trabalho

A execução do projeto proposto exigirá, entre outras coisas: organização em um banco de dados relacional das informações biológicas referentes à replicação de DNA em kinetoplastídeos, em particular das de *T. cruzi*; desenho, implementação, teste e validação do modelo computacional estocástico; uso do modelo validado para estudar a dinâmica das colisões; extrapolação da metodologia para outros kinetoplastídeos e, como *benchmarking*, também para *Saccharomyces cerevisiae* (levedura), um importante organismo-modelo eucarioto. Na seção 5.1 mostramos uma lista dessas atividades, suas respectivas descrições e o cronograma esperado.

5.1 Descrição das atividades

Essas atividades serão realizadas durante o primeiro semestre deste ano. As atividades para o segundo semestre ainda serão decididas conforme os resultados obtidos nos próximos meses.

Atividade 1. Realização de simulações com a versão atual do modelo e análise dos resultados.

Atividade 2. Conversão dos resultados em planilhas e gráficas para melhor interpretação e difusão das informações obtidas.

Atividade 3. Análise da sensibilidade dos parâmetros do modelo, a fim de apontarmos regiões desses parâmetros que melhor descrevem os fenômenos biológicos.

Atividade 4. Expansão do modelo a partir dos dados obtidos das simulações anteriores.

Atividade 5. Desenvolvimento do relatório parcial de avanços para a FAPESP.

Atividade/mês	Abr.17	Mai.17	Jun.17	Jul.17
Atividade 1	**x**	-	-	-
Atividade 2	**x**	**x**	-	-
Atividade 3	-	**x**	**x**	-
Atividade 4	-	-	**x**	-
Atividade 5	-	-	**x**	**x**

Tabela 1: cronograma de atividades previstas para este semestre, listadas acima

6 Disseminação dos Resultados

Além da monografia e do poster que serão produzidos para este Trabalho de Conclusão de Curso, o projeto e seus resultados também serão expostos através de:

(a) Disciplina no IME. Os primeiros resultados do trabalho foram expostos através da disciplina MAC215 – Atividade Curricular em Pesquisa, ao longo da qual foram elaborados relatórios de progresso do projeto e também foi apresentado um pôster contendo resultados parciais.

(b) Atividades no Instituto Butantan. O aluno deverá participar da Reunião Científica Anual do Butantan, mostrando seus resultados em um pôster.

(c) Atividades no cenário acadêmico nacional. Espera-se que os resultados sejam apresentados em alguma conferência nacional de Bioinformática (por exemplo, no X-Meeting) e, posteriormente, com as devidas validações biológicas, sejam reportados em um artigo científico.

(d) Distribuição do código O código-fonte final do projeto será disponibilizado para a comunidade acadêmica, sob a licença GNU GLP, no repositório web GitHub.

Referências

[1] Catarina A Marques, Nicholas J Dickens, Daniel Paape, Samantha J Campbell, and Richard McCulloch. Genome-wide mapping reveals single-origin chromosome replication in *Leishmania*, a eukaryotic microbe. *Genome biology*, 16(1):1, 2015.

[2] Simone Guedes Calderano, William C Drosopoulos, Marina Mônaco Quaresma, Catarina A Marques, Settapong Kosiyatrakul, Richard McCulloch, Carl L Schildkraut, and Maria Carolina Elias. Single molecule analysis of *Trypanosoma brucei* DNA replication dynamics. *Nucleic acids research*, page gku1389, 2015.

[3] Yea-Lih Lin and Philippe Pasero. Interference between DNA replication and transcription as a cause of genomic instability. *Current genomics*, 13(1):65–73, 2012.

[4] 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*. *Nucleic acids research*, 38(suppl 1):D457–D462, 2010.