Decifrando a dinâmica de replicação de DNA em *Trypanosoma cruzi* através de modelagem computacional

Candidato: Victor Seiji Hariki

Pesquisador Responsável: Marcelo da Silva Reis

Centro de Toxinas, Imuno-resposta e Sinalização Celular (CeTICS)

Laboratório de Ciclo Celular (LCC)

Instituto Butantan, São Paulo, 27 de março de 2020.

Resumo

Tripanossomatídeos são protozoários endoparasitas cujos genes estão organizados em policistrons e têm transcrição constitutiva ao longo de todo ciclo celular. Esses fatos sugerem que conflitos entre as maquinarias de replicação de DNA e de transcrição levam a um aumento do disparo de origens de replicação na fase S do ciclo celular. Para investigar essa hipótese, desenvolvemos um modelo de replicação de DNA para *Trypanosoma brucei* que foi calibrado com dados de MFA-seq e utilizado para prever e validar que um aumento dos níveis de transcrição constitutiva de fato leva a um aumento do número de origens disparadas. Recentemente, foi concluído o mapeamento das origens de *T. cruzi* através de ensaios de MFA-seq, que revelou que diversas origens estão localizadas em regiões codificadoras de *dispersed gene family 1* (DGF-1), uma família de genes que é relevante para o ciclo de vida do parasita. Uma vez que genes DFG-1 têm uma grande variabilidade genética, uma possibilidade é que a mesma seja uma consequência de conflitos entre as maquinarias de replicação de DNA e de transcrição, o que implicaria que a distribuição de origens de replicação em *T. cruzi* é condicionada pela organização genômica desse parasita. Neste projeto, propomos investigar essa questão através da modelagem computacional da dinâmica de replicação de DNA de *T. cruzi*. Para isso, adaptaríamos para *T. cruzi* o modelo dinâmico da programação de replicação de DNA desenvolvido para *T. brucei*, calibraríamos o mesmo com dados de MFA-seq, e faríamos experimentos computacionais para diferentes conjuntos de parâmetros. Executaríamos os mesmos experimentos também para outras distribuições (sintéticas) de policistrons, o que permitiria a comparação com a organização genômica real. Dessa forma, esperamos avaliar o impacto da organização genômica e de conflitos replicação-transcrição na programação do disparo de origens de replicação em *T. cruzi*, em particular em regiões de genes DFG-1, assim ajudando a elucidar um dos mecanismos que garantem o sucesso do parasita na invasão e sobrevivência em seu hospedeiro.

Deciphering the DNA replication dynamics in *Trypanosoma cruzi* through computational modeling

Applicant: Victor Seiji Hariki

Advisor: Marcelo da Silva Reis

Center of Toxins, Immune-response and Cell Signaling (CeTICS)

Laboratório de Ciclo Celular (LCC)

Instituto Butantan, São Paulo, March 27, 2020.

Abstract

Trypanosomatids are endoparasitic protozoan whose genes are organized into polycistrons and have constitutive transcription along the whole cell cycle. Those facts suggest that conflicts between DNA replication and transcription machineries yield increased replication origin firing in the cell cycle S phase. To investigate that hypothesis, we developed a DNA replication model for *Trypanosoma brucei* that was calibrated with MFA-seq data and used it to predict and validate that increasing constitutive transcription levels indeed increase the number of fired origins. Recently, it was concluded the DNA origins mapping for *T. cruzi* through MFA-seq assays, which unveiled that several origins are located in coding regions of the dispersed gene family 1 (DFG-1), a family of genes that is relevant for the parasite life cycle. Once DFG-1 genes have high genetic variability, one possibility is that conflicts between DNA replication and transcription machineries are responsible for such variability, which implies that the origin firing distribution in *T. cruzi* is conditioned by the genomic organization of that parasite. In this project, we propose to investigate such question through the computational modeling of the DNA replication dynamics of *T. cruzi*. We would adapt for *T. cruzi* the dynamic model of DNA replication programming originally developed for *T. brucei*, calibrate it with MFA-seq data, and carry out a number of computational experiments for different sets of parameter values. We would also repeat those assays for other (synthetic) distribution of polycistrons, thus allowing the comparison with the actual genomic organization. Therefore, we expect to be able to assess the impact of the interplay between genomic organization and replication-transcription conflicts on the DNA origin firing programming in *T. cruzi*, in particular in DFG-1 genomic regions, thus helping to understand one of the underlying mechanisms that guarantees the successful invasion and survival of the parasite in its host.

Sumário

1 Introdução

A família dos tripanossomatídeos é composta por protozoários endoparasitas obrigatórios, alguns deles de relevância biomédica, tais como o *Trypanosoma cruzi* (o agente etiológico da doença de Chagas) e o *T. brucei* (causador da doença do sono). Como as moléstias causadas por espécies dessa família de protozoários são consideradas pela Organização Mundial da Saúde como doenças negligenciadas [1], a biologia desses organismos é intensamente estudada em busca de alvos moleculares, visando intervenções farmacológicas para prevenção e tratamento de infecções. Tais estudos já revelaram que tripanossomatídeos têm características biológicas peculiares; entre elas, o fato de seu genoma ter seus genes organizados em policistrons (i.e., sequências lineares de genes que são transcritos juntos em uma única molécula de RNA) e também a presença de uma transcrição constitutiva (i.e., que ocorre o tempo todo, sem regulação por fatores de transcrição em sua região promotora) ao longo de todo o ciclo celular. Por conta disso, a dinâmica de replicação de DNA desses organismos também conta com propriedades bastante específicas.

A replicação de DNA em eucariotos, o que inclui os protozoários, é iniciada em sítios genômicos denominados origens de replicação. Em cada um desses sítios, um par de maquinarias de replicação denominadas replissomos se liga ao mesmo e dá início à replicação do DNA, com os dois replissomos partindo em sentidos opostos; chamamos esse fenômeno de *disparo de origem de replicação*. A transcrição, por sua vez, ocorre com o auxílio de uma enzima chamada RNA polimerase (RNAP). A organização genômica dos tripanossomatídeos leva a ocorrência de conflitos entre as maquinarias de replicação de DNA e de transcrição. Existem dois tipos desses conflitos: colisões traseiras (*head-to-tail*) e frontais (*head-to-head*) [2]. Colisões traseiras são resolvidas com o replissomo desalojando a RNAP da fita de DNA; já as colisões frontais em geral não são solucionadas facilmente, e suas ocorrências podem levar a um colapso do replissomo ou mesmo à quebra da fita de DNA [2]. Visando estudar o impacto dos conflitos entre as maquinarias de replicação de DNA e de transcrição na dinâmica de replicação de DNA em *Trypanosoma brucei*, desenvolvemos um modelo computacional estocástico que simula a programação da fase S desse organismo (Figura 1). Esse modelo foi baseado em um modelo dinâmico de replicação de DNA reportado anteriormente [3], e calibrado com dados experimentais disponíveis na literatura, tais como a distribuição dos sítios putativos de origens de replicação fornecidos por ensaios de MFA-seq [4], assim como outras propriedades do processo tais como velocidade estimada do replissomo e duração da fase S [5, 6]. Simulações desse modelo com diferentes conjuntos de parâmetros mostraram que níveis crescentes de transcrição constitutiva levam a um aumento do número de conflitos, que por sua vez induzem um incremento no número de origens de replicação disparadas; todavia, tal

aumento não produz alterações significativas no tempo necessário para replicar todo o DNA desse parasita, uma previsão que foi confirmada experimentalmente [7].

Mais recentemente, em um estudo liderado pela Dra. Maria Carolina Elias (Laboratório de Ciclo Celular, Instituto Butantan), foram concluídos ensaios de MFA-seq para outro tripanossomatídeo, o *T. cruzi* (Araujo e colegas, manuscrito enviado para publicação). Ao comparar o perfil de replicação de DNA por cromossomo fornecido por esses ensaios de MFA-seq com as localizações dos policistrons nesse parasita, foi constatado um grande número de origens de replicação putativas em regiões codificadoras de genes da família *dispersed gene family* (DFG-1). Genes DGF-1 são fundamentais no ciclo de vida de *T. cruzi*, pois os mesmos codificam proteínas de superfície celular importantes no sucesso da infecção e também da sobrevivência do parasita no hospedeiro. Como os genes DGF-1 possuem grande variabilidade genética, uma possibilidade é que conflitos entre as maquinarias de replicação e de transcrição desempenhem um papel fundamental nessa propriedade, uma questão em aberto e que poderia ser investigada com modelos computacionais similares aos que empregamos com sucesso em *T. brucei*.

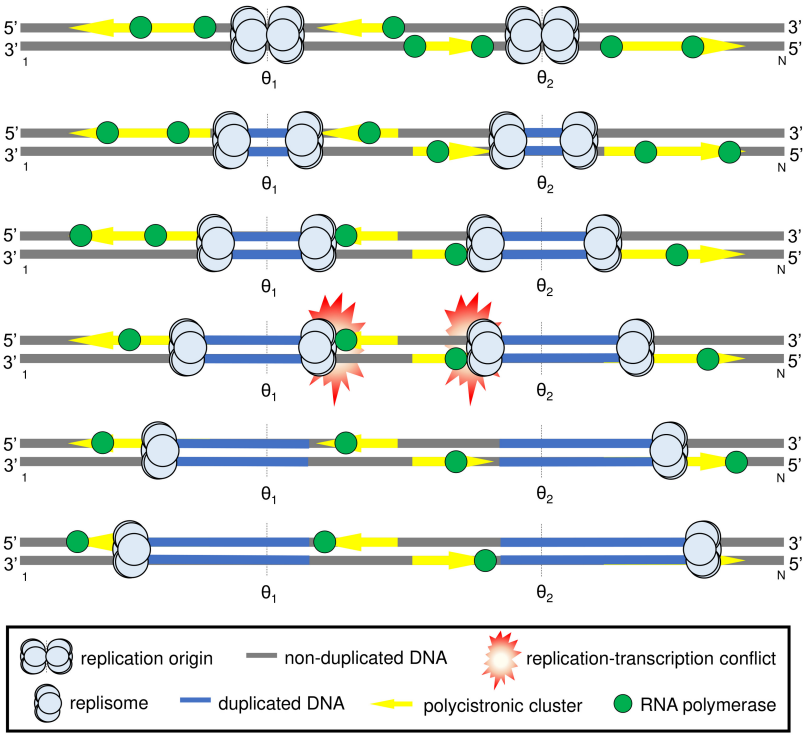

Figura 1: Modelo dinâmico da replicação de DNA em *T. brucei*. No exemplo, temos um segmento de um cromossomo com duas origens de replicação (θ_1 e θ_2). Quando simulamos replicação de DNA e transcrição ao mesmo tempo, se temos uma colisão frontal entre maquinarias de replicação de DNA (replissomo) e de transcrição (RNAP), então ambas são desalojadas do DNA, deixando a replicação inacabada. No caso deste exemplo, uma terceira origem de replicação, localizada entre θ_1 e θ_2, precisaria ser disparada para completar a replicação interrompida por tal conflito. Figura extraída de da Silva e colegas [7].

2 Objetivos

O objetivo geral desta proposta é a utilização de modelagem computacional para investigar a programação da replicação de DNA de *Trypanosoma cruzi*, o protozoário causador da doença de Chagas. Para este fim, adaptaríamos para *T. cruzi* o modelo dinâmico desenvolvido originalmente para *T. brucei* e implementado em linguagem de programação C++, utilizando para ajuste desse modelo dados disponíveis na literatura e também o MFA-seq de *T. cruzi* recém-concluído.

Como objetivo específico, propomos testar a hipótese de que a organização genômica do parasita combinada com conflitos entre as maquinarias de replicação de DNA e de transcrição são importantes tanto para a programação da replicação de DNA em *T. cruzi* quanto para a variabilidade genética verificada nos genes da família DFG-1. Para isso, executaríamos experimentos computacionais, nos quais simularíamos o modelo a ser construído um grande número de vezes para cada conjunto de valores de parâmetros. Inicialmente, utilizaríamos nessas simulações a organização genômica real do parasita, que seria obtida de repositórios públicos de informações biológicas; num segundo momento, repetiríamos esses ensaios com topologias aleatorizadas (i.e., sintéticas), o que nos permitiria testar, ao comparar as programações de replicação de DNA obtidas com a topologia genômica real contra as obtidas simulando o modelo com topologias sintéticas, a hipótese mencionada acima.

3 Metodologia

A metodologia desta proposta de projeto de iniciação científica consistiria na execução de três etapas sequenciais: 3.1) Coleta e organização de informações da replicação de DNA de *T. cruzi* disponíveis, tanto da literatura e/ou repositórios públicos quanto os produzidos em nosso laboratório e ainda não publicados; 3.2) Execução dos experimentos computacionais; 3.3) Análise dos resultados obtidos.

3.1 Coleta e organização de informações biológicas

Informações biológicas relevantes para a construção do modelo dinâmico deverão ser coletadas de repositórios públicos tais como o TriTrypDB e o GeneDB [8, 9]. Entre outras informações relevantes, precisamos coletar:

- Tamanho de cada cromossomo, isto é, o número de pares de bases;
- Distribuição da programação de replicação de DNA (obtida com MFA-seq);

- Localização, tamanho e sentido de transcrição dos policistrons;
- Duração de todo o processo de replicação de DNA;
- Velocidade estimada do replissomo.

Para organizar todas essas informações, cogitamos utilizar o gerenciador de banco de dados SQLite, já empregado nas primeiras implementações do simulador do modelo de replicação de DNA de *T. brucei* reportado anteriormente [7]. Todavia, caso seja necessário adotar um gerenciador mais robusto, uma alternativa seria utilizarmos o MySQL.

3.2 Experimentos de simulação do modelo de replicação de DNA de *T. cruzi*

Uma vez organizados todos os dados mencionados na seção anterior, a próxima etapa seria realizar os experimentos computacionais propriamente ditos. Para esse fim, adaptaríamos o simulador de modelo de replicação de DNA de *T. brucei* que foi utilizando no trabalho anterior [7], disponível publicamente em:

github.com/msreis/ReDyMo-CPP.

Esse simulador foi codificado em C++, o que garante um bom desempenho do ponto de vista de consumo de tempo computacional. Além disso, utilizaríamos para realizar lotes de simulações uma servidora recém-adquirida pela Diretoria Científica do Instituto Butantan, que conta com uma configuração adequada para computação científica de alto desempenho:

- 4 processadores Intel Xeon Platinum, 2.1 GHz, cada um com 28 núcleos (56 threads) Turbo e 38 MB de cache;
- 1 TB de memória RAM, RDIMM;
- Placa NVIDIA Tesla P40 24GB Passive GPU (3.840 núcleos).

Como essa servidora conta com uma placa de CPUs, caso seja necessário poderíamos melhorar o escalonamento das simulações fazendo uso dos 3.840 núcleos de GPUs disponíveis, adaptando o código-fonte para esse fim.

Simulações seriam feitas utilizando todas as informações biológicas armazenadas no banco de dados; as exceções seriam os experimentos com localizações e tamanhos aleatórios de policistrons, que utilizariam os policistrons reais apenas para calcular o tamanho somado de todas as regiões codificadoras, que por sua vez seria empregado como restrição na aleatorização (i.e., isso equivale a colocar como restrição que a topologia sintética precisa transcrever os mesmos genes que a topologia real).

3.3 Análise dos resultados

Os resultados produzidos nas simulações serão organizados, com o auxílio de *scripts* em Python ou outra linguagem de programação, em tabelas e planilhas, sobre as quais calcularemos as principais estatísticas da dinâmica de replicação de DNA de *T. cruzi*; entre elas, analisaremos:

- Distância entre origens de replicação (média, mediana e desvio padrão);
- Distribuição da programação de replicação de DNA por cromossomo (média e desvio padrão);
- Duração da fase S (média, mediana e desvio padrão);
- Número de colisões frontais (média, mediana e desvio padrão).

Além disso, analisaremos especificamente os resultados referentes a regiões de genes da família DFG-1, para avaliar em que situações (i.e., conjuntos de parâmetros) conflitos entre maquinarias de replicação de DNA e de transcrição são observados com alta frequência nas regiões codificadoras desses genes.

4 Plano de trabalho e cronograma de execução

Para a execução deste projeto proposto, listamos abaixo as principais atividades previstas; o diagrama de Gantt com o cronograma é apresentado na tabela 1. Este plano de atividades supõe que o Candidato inicie suas atividades em 1 de abril de 2020, assim como que as mesmas sejam realizadas no contexto da disciplina MAC499 – Trabalho de Formatura Supervisionado. Portanto, ao final deste ano, o aluno deverá entregar uma monografia que sintetizaria todas as atividades desenvolvidas ao longo desta proposta.

Atividade 1: Estudo dirigido de artigos científicos pertinentes ao projeto proposto;

Atividade 2: Coleta das informações biológicas necessárias para desenhar o modelo de replicação de DNA; organização dos dados em um banco de dados;

Atividade 3: Refatoração do código do simulador, adaptando o mesmo para *T. cruzi*;

Atividade 4: Simulações com o simulador refatorado na servidora do Instituto Butantan;

Atividade 5: Análise dos resultados iniciais;

Atividade 6: Adaptação do código do simulador para que o mesmo faça uso de GPUs (caso seja necessário);

Atividade 7: Nova rodada de simulações, incluindo pontos que eventualmente forem observados na primeira bateria de simulações;

Atividade 8: Análise da segunda rodada de resultados;

Atividade 9: Escrita e entrega da monografia;

Atividade 10: Apresentação de pôster na disciplina MAC499 e também na Reunião Científica Anual do Butantan.

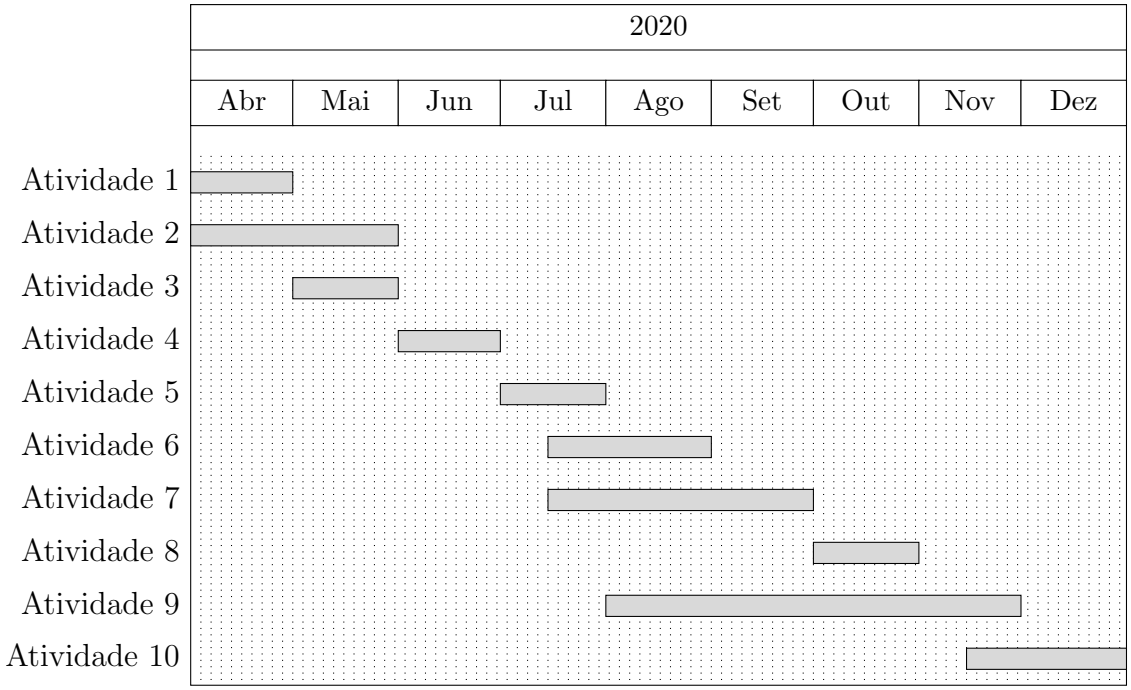

Tabela 1: Diagrama de Gantt contendo o cronograma de execução deste projeto proposto.

5 Formas de avaliação e disseminação dos resultados

Os resultados deste projeto proposto serão avaliados em dois aspectos: do ponto de vista de desempenho computacional, espera-se que seja possível realizar simulações em uma escala adequada para responder às questões levantadas neste projeto. Caso seja necessário, faremos uso de GPUs para aumentar o desempenho computacional de nossas simulações. Do ponto de vista semântico, espera-se que com o modelo a ser desenvolvido

e ajustado seja possível testar a hipótese levantada, seja a resposta para a mesma positiva ou não.

Os resultados serão disseminados inicialmente através da monografia do Candidato e também com os pôsteres que o mesmo deverá apresentar tanto na disciplina de Trabalho de Conclusão de Curso quanto na Reunião Científica Anual do Butantan (caso esta última não seja cancelada por conta da pandemia de COVID-19). No médio prazo, esperamos reportar os resultados deste projeto proposto em uma publicação científica em um periódico de bom nível, tal qual já fizemos em da Silva e colegas [7].

Referências

[1] Peter J. Hotez, David H. Molyneux, Alan Fenwick, Jacob Kumaresan, Sonia Ehrlich Sachs, Jeffrey D. Sachs, and Lorenzo Savioli. Control of neglected tropical diseases. *New England Journal of Medicine*, 357(10):1018–1027, 2007.

[2] Tatiana García-Muse and Andrés Aguilera. Transcription-replication conflicts: how they occur and how they are resolved. *Nature Reviews Molecular Cell Biology*, 17(9):553–563, 2016.

[3] Yevgeniy Gindin, Manuel S. Valenzuela, Mirit I. Aladjem, Paul S. Meltzer, and Sven Bilke. A chromatin structure-based model accurately predicts dna replication timing in human cells. *Molecular Systems Biology*, 10(3):722, 2014.

[4] 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*. *Cell Reports*, 2(1):185–197, 2012.

[5] Simone G. 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.

[6] Marcelo S. da Silva, Paula Andrea Marin Muñoz, Hugo A. Armelin, and 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. *Journal of Eukaryotic Microbiology*, 2017.

[7] Marcelo S. da Silva, Gustavo R. Cayres-Silva, Marcela O. Vitarelli, Paula A. Marin, Priscila M. Hiraiwa, Christiane B. Araújo, Bruno B. Scholl, Andrea R. Ávila, Richard McCulloch, Marcelo S. Reis, and Maria Carolina Elias. Transcription activity contributes to the firing of non-constitutive origins in african trypanosomes helping to maintain robustness in S-phase duration. *Scientific Reports*, 9(1):18512, 2019.

[8] 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. *Nucleic Acids Research*, 32(suppl 1):D339–D343, 2004.

[9] 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.