A Study on Gradient Boosting Classifiers

Juliano Garcia de Oliveira
Advisor: Prof. Dr. Roberto Hirata Jr.

Department of Computer Science – Institute of Mathematics and Statistics
University of São Paulo

Introduction

Trypanosomatida is a family of pathogenic protozoa, whose one of most prominent members is *Trypanosoma brucei*, the parasite behind the African sleeping sickness. Given the public health significance of this protozoan, its biology is int , in a parallelized fashion, on several specific DNA locations called replication origins. There are three types of origins: constitutive, flexible and dormant; the latter is only initiated as a response to DNA replication failure. Recently, putative constitutive origins were mapped for *T. brucei* through MFA-seq assays [1]. Moreover, a DNA replication dynamic model was implemented in a Python-based simulator called ReDyMo [2]. Simulations of this model highlighted the relevance of replication-transcription conflicts on the origin firing distribution along the genome [2]. However, whether this parasite has or not dormant origins remained as open question.

Gradient Boosting Machines

- **General objective:** to computationally test the hypothesis of the presence of dormant origins in *T. brucei* strain TREU927.
  1. it will require the designing of a multiscale model
  2. this model will be used to simulate the dynamics of a parasite population in exponential growth
- **Specific goal:** to port the ReDyMo simulator to C++ in order to:

Study Methodology

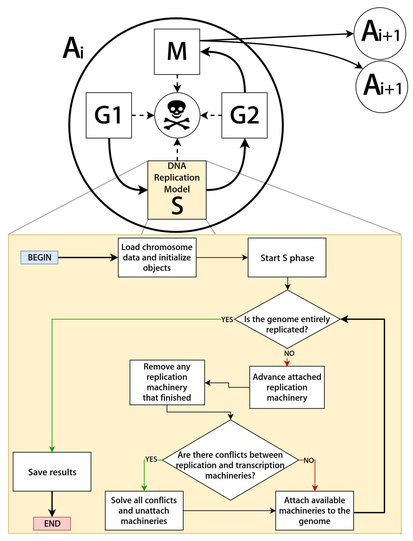

Fig. 1: Diagram of the proposed agent-based multiscale model. Each cell (agent) has an outer and an inner layers. In the outer layer it is simulated all phases of cell cycle (G1, S, G2/M). Each cell either finishes its cycle successfully (so it spawns two daughter cells) or dies (and is removed from simulation). The S-phase duration and success is given by the DNA replication dynamic model contained in the inner layer (orange box).

References

[1] Tiengwe et al. Cell Reports (2012). DOI: 10.1016/j.celrep.2012.06.007
[2] da Silva et al. Submitted (2018). DOI: 10.1101/398016

Statistical Analysis

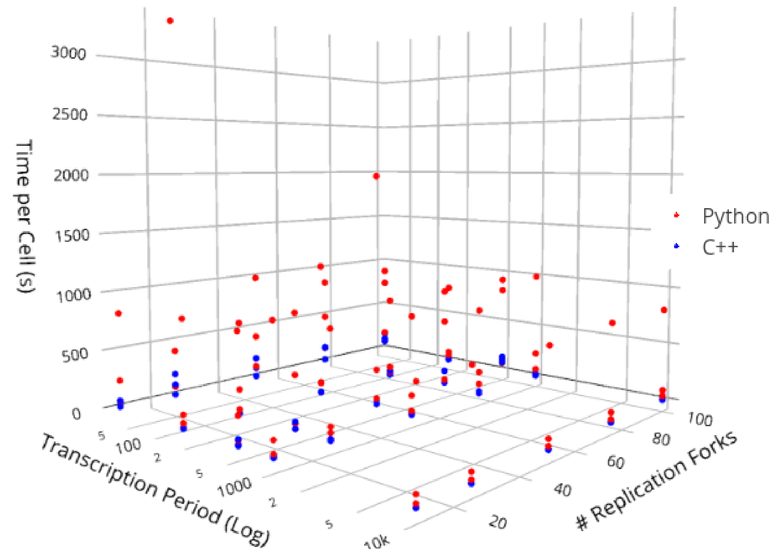

Fig. 2: Performance comparison between C++ and Python implementations. In this assay, we compared the Python version of the ReDyMo simulator (red dots) with its C++ port (blue dots). We tested both simulations with different values for transcription frequency and number of replication forks. For each pair of these parameters, we carried out batches of 1, 10, and 100 simulations, one batch at a time. In the case of the 100-simulations batch, up to 40 simulations were executed at the same time. On average, C++ simulations required only 8 % of the Python computational time.

Results and Single-Factor Models

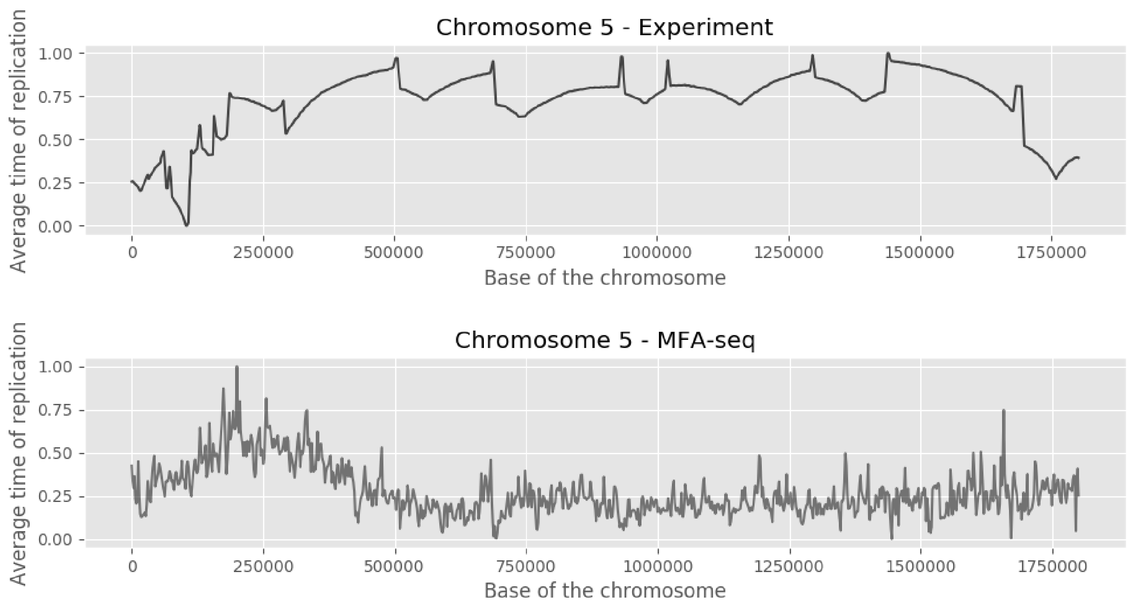

Fig. 3: Simulations with uniform probability of origin firing along the genome. In this experiment, we carried out 3,000 independent simulations, each one with transcription period equal to 10,000 simulation iterations and 25 replication forks. Then we split each chromosome into 1,000 equally-sized bins. For each simulation and for each bin, we computed the mean iteration that each base was replicated. Finally, for each

Conclusions

- We implemented and tested a C++ port of ReDyMo, which is much more efficient than the Python-based version
- We used this new version to show preliminary results that suggest an influence of transcription activity (i.e., its conflicts with replication forks) on the origin firing positioning
- Next step is the implementation of the outer layer of the multiscale model within the FLAME framework.

juliano.garcia.oliveira@usp.br