Implementation of a simple to use Gene Transcription Simulator

***Eduardo da Veiga Beltrame** and **Jane Kondev.** Biophysics Department, Brandeis University.*

Introduction

There is currently no user friendly tool for modeling gene expression at the level of individual polymerases elongating through the gene. There is a large diversity of published models which share many common features, but are generally inaccessible to newcomers. We have implemented a stochastic transcription simulator that is simple to use and customize, available at *www.genesim.org*.

This platform should be useful for individuals who are not interested in developing models from scratch either analytically or computationally, but would still like to investigate how gene expression is affected by the characteristics of a genetic circuit, such as length, binding or elongation rate. It can also yield predictions for systems too complicated to be treated analytically, whose dynamics would not be intuitively apparent.

Modeling Framework

We are interested in developing a quantitative model for transcription, the first step in gene expression, where a particular segment of DNA is read and translated into RNA by the enzyme RNA polymerase. Polymerase binds to DNA, and slides forward, reading the DNA sequence and synthetizing the correspondent RNA at every step.

Genes are modelled as unidimensional lattices of length L where each locus in the lattice correspond to a stretch of DNA the size of the polymerase footprint. Each position may be either empty or be occupied by one polymerase.

Every locus has 5 rates defining how many "actions per unit of time" happen on average on that locus. They are all treated as simple biochemical reaction steps, and have an exponential probability distribution over time.

Elongation: If the locus is occupied, the polymerase can take a step forward, freeing its location and occupying the next one.

Backtrack: If the locus is occupied, the polymerase it can take a step back, freeing its location and occupying the previous one.

Binding: If the locus is free, a new polymerase may bind to it.

Termination: If the locus is occupied, it can become unoccupied, and an RNA transcript is produced.

Premature termination: If the locus is occupied, it can become unoccupied, but no RNA transcript is produced.

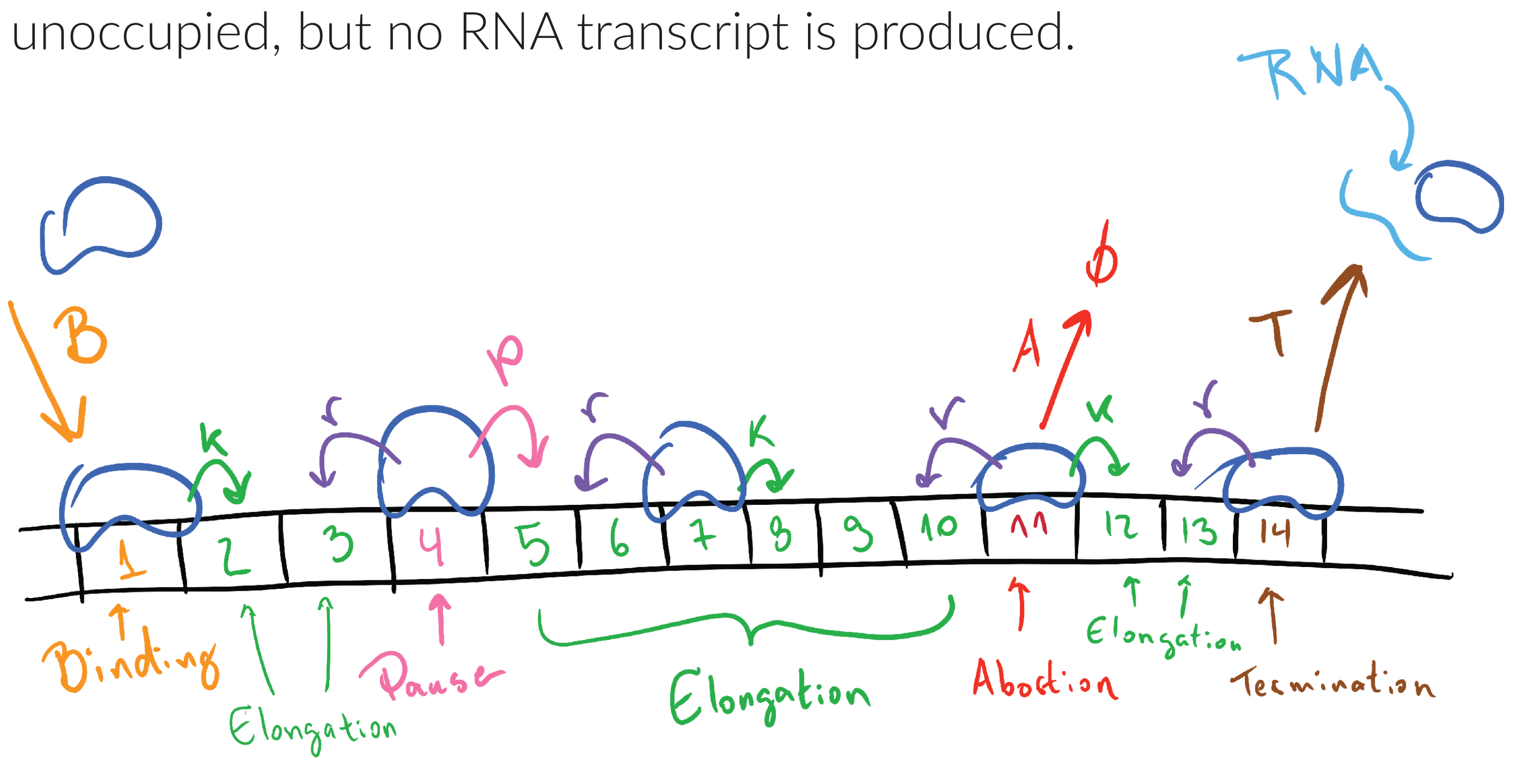

***Figure 1:** An example gene model.*

*Locus 1 is a binding element, where polymerase can bind with rate **B**, and elongate with rate **k**, with all other rates being zero.*

*On loci 2-3, 5-10 and 12-13, elongation rate is **k**, and backtrack rate is **r**.*

*Locus 4 is also an elongation element with backtrack rate **r**, but elongation rate **p**. If we consider **p < k**, then locus 4 would be a pausing element, where polymerase takes a longer time than normal to advance.*

*On locus 11, in addition to normal elongation rates, the polymerase may abort with rate **A**. On locus 14 it can either backtrack with rate **r**, or terminate with rate **T** and produce an RNA transcript.*

Actual Data!

Algorithm

The stochastic simulation was implemented as a Python script using the Gillespie algorithm, which loops through the following steps and updates the gene state until the end time is reached:

1) The rates for all possible actions A_i are summed into $Z = \Sigma_i A_i$

2) An action A_i with rate r_i is chosen with probability $P_i = r_i / Z$

3) The time the action A_i takes to happen is drawn from $p(\Delta t) = e^{-t \cdot r_i}$

4) The time is increased by Δt and the gene state is updated

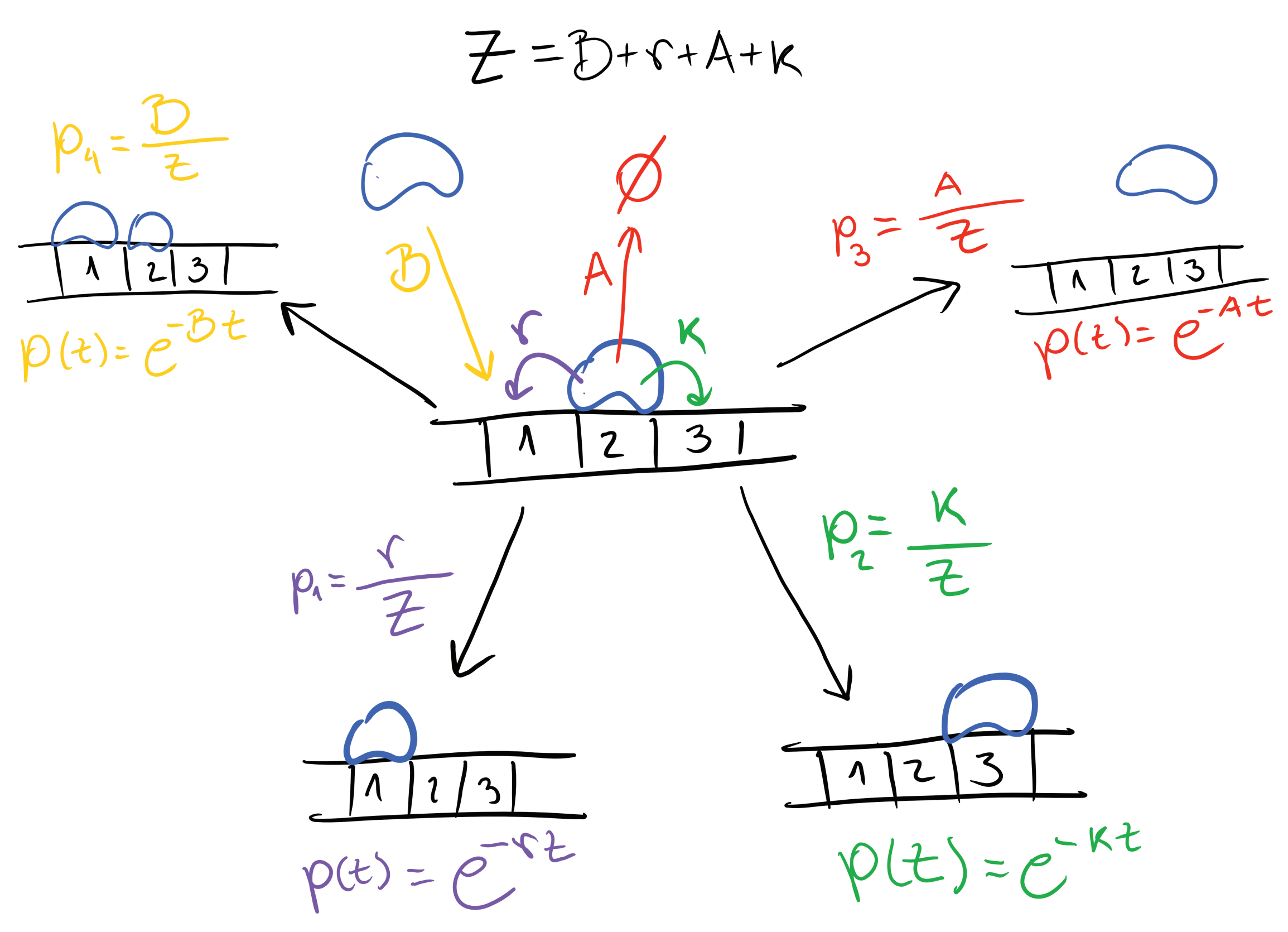

***Figure 2:** Illustration of the algorithm for a 3 loci gene. In the current state there is only one polymerase bound, on locus 2. Four things can happen:*

*1) The polymerase can backtrack at rate **r***

*2) It can elongate, taking a step forward at rate **k***

*3) It can abort, leaving the location free at rate **A***

*4) A new polymerase can bind on locus 1 with rate **B***

In this case $Z = B + r + A + k$. The probability of a given action A_i ocurring depends on the corresponding rate r_i normalized by Z, so $p_i = r_i / Z$.

The time interval is then drawed from the distribution $p(t) = e^{-r_i t}$

Results

The simulator output is shown below for the example gene of figure 1. Four other gene occupancy plots are also shown on the borders of this poster, depicting a gene with no pause, long pause, high abortion date and slow termination (can you deduce which is which?). We are currently working on original investigations using the simulator and expanding it's capabilities. We encourage the interested reader to visit www.genesim.org to experiment with the simulator.

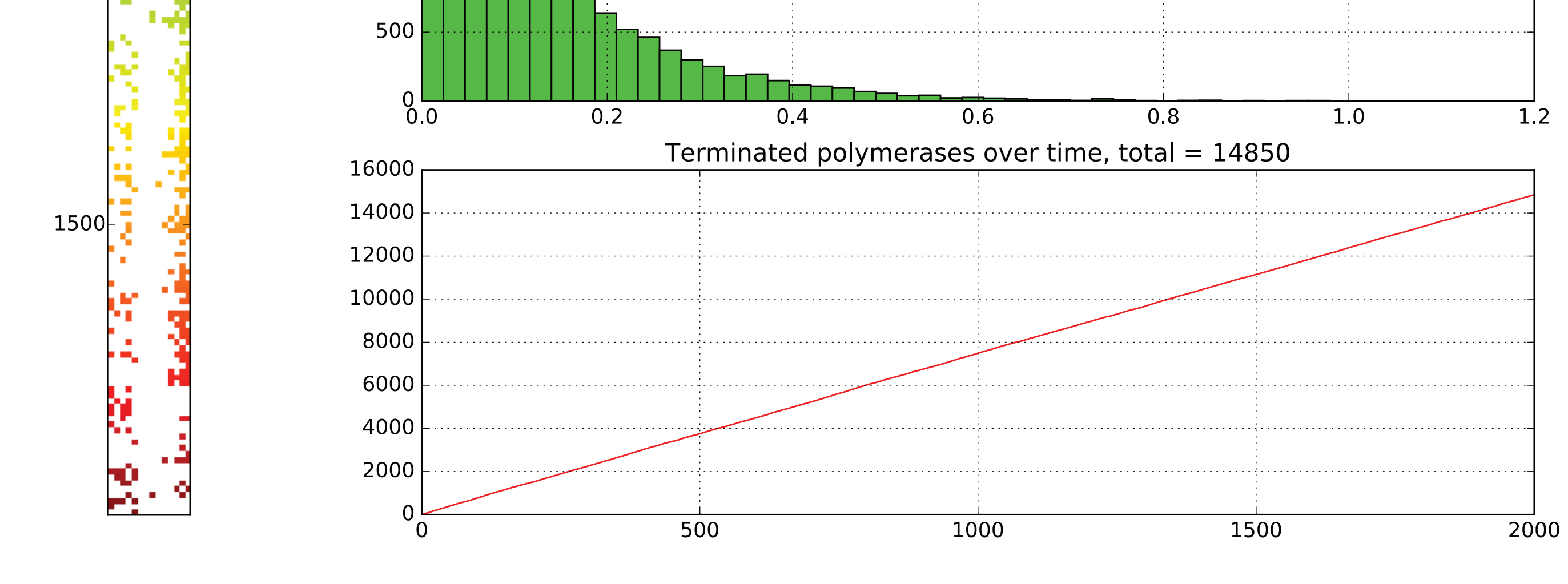

***Figure 3:** Example of the simulator output for thegene discussed on figure 1, with L = 14, k = 1000, r = 50, p = 20, A = 1, T = 10 and total time = 2000.*

***a)** A gene occupancy plot: each line corresponds to a snapshot of the gene in a moment of time, white pixels are empty locations and pixels representing polymerases colored in the order which they were bound.*

***b)** A histogram of the times each polymerase took to transcribe the gene*

***c)** A histogram of the intervals between two consecutive terminations.*

***d)** The number of terminated polymerases over time.*