Master thesis proposal
Original strands in PCR

Krzysztof Bartoszek

October 8, 2025

Background — PCR

A polymerase chain reaction (PCR) is today the standard way of amplifying DNA data for any further analysis. You have seen it, on nearly every episode of CSI! It is that round thing, which spins after a number of test tubes have been put into it. The main idea is that one starts with a few DNA strands and the polymerase (similar to the one in living cells) in the test tube makes a copy of each strand. Then each strand is copied again and again leading to an exponential explosion in strands.

Biologists are of course interested how many such rounds of DNA strand copying should take place before they have enough material to work with. This is modelled by the so–called binary "branching processes". Put simply a branching process is a stochastic process that starts with a single (in general k) particle. This particles lives for a fixed (in general random) amount of time, then either dies or splits into 2 (in general a random number) particles. Then each particle behaves in the same way.

In the PCR case the DNA strands are the particles and they can die (DNA degrades naturally), they can split into two (polymerase succeeded) or nothing happens (polymerase failed). What makes the PCR case more interesting is that we do not start with a single strand but with an undisclosed collection of them (how do you know how much you collected from an old crime scene?). A possible question is given that a PCR has run for a number of rounds can we estimate the original number of strands [1]. This can be meaningful if we have DNA from multiple donors and want to compare the contribution of each one. Recent mathematical results [e.g. Thm. 2.1 1] show

that for this one does not need to consider a random process but approximate it with a deterministic dynamical system.

Thesis project

We first simulate a branching process, Z_n (number of DNA strands at time n) as follows. Let $Z_0 = n_0 \in \mathbb{N}$ (initial number of strands). Then, iteratively

$$Z_{k+1} = Z_k + \sum_{j=1}^{Z_k} \zeta_{k+1,j},$$

where $\zeta_{k+1,j} \in \{0, 1\}$ are binary random variables taking the value 1 with probability $p_{k+1} = (vK)/(k + Z_k)$. This probability corresponds to the carrying capacity, K, in a deterministic dynamical system [2]. You can explore other functions related to dynamical systems.

Define $X_k = Z_k/K$. A recent mathematical result [e.g. Thm. 2.1 1] is that if the initial number of DNA strands is comparable with K, i.e. $Z_0/K \sim x_0$ for large K then

$$X_k \approx f^{\circ k}(x_0),$$

where $f(x) = x + vx/(1 + x)$ (again explore other functions) and $f^{\circ k} = f \circ f \ldots \circ f$ is the function f applied k times. Illustrate this theorem with simulations.

For a given $k \in \mathbb{N}$ (k reasonable e.g. 10 or 15), estimate from simulations the probability $P(Z_0 = n|Z_k = m)$. In particular find $\mathrm{E}[Z_0|Z_k = m]$ and $\mathrm{Var}[Z_0|Z_k = m]$. Notice that here the initial state is a random one (as we should expect, the number of DNA strands on a crime scene is a random number). We can try different initial distributions. The question is to explore what can we say about its distribution based on the number of DNA strands after k iterations. How helpful is the deterministic approximation for finding Z_0? Notice that one may construct an estimator based on the discrete approximation.

Goals

The below general goals are for an "ideal" thesis. Depending on the student they will be made more specific in the direction of the student's interests.

In particular the focus of the work will not be on the mathematical models (these will be "provided") but on implementing and putting together software to do simulations, inference and explore the statistical aspects of the models.

- Become acquainted with branching processes, esp. simulating them [3].
- Become acquainted with simulating dynamical systems [2].
- Become acquainted with exploring, summarizing statistically a stochastic process and comparing it with a discrete dynamical system.

Data

In the scope of the project simulated data will be used.

References

[1] P. Chigansky, P. Jagers, and F. Klebaner. What can be observed in real time PCR and when does it show? *ArXiv e–prints*, 2016.

[2] G. de Vries, T. Hillen, M. Lewis, B. Schönfisch, and J. Muller. *A Course in Mathematical Biology: Quantitative Modeling with Mathematical and Computational*. SIAM, 2006.

[3] P. Haccou, P. Jagers, and V. A. Vatutin. *Branching Processes: Variation, Growth and Extinction of Populations*. Cambridge University Press, Cambridge, 2005.