UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRONÓMICA, ALIMENTARIA Y DE BIOSISTEMAS

MÁSTER EN BIOLOGÍA COMPUTACIONAL

DEPARTAMENTO DE BIOTECNOLOGÍA-BIOLOGÍA VEGETAL

DEPARTAMENTO DE EVOLUCIÓN MOLECULAR

Modeling the dynamics of CRISPR arrays in microbial communities

TRABAJO FIN DE MÁSTER

Autor: Jimena Martín Reina

Tutor Externo: Jaime Iranzo Sanz

Tutor Académico: Carlos Pérez Cantalapiedra

June, 2025

MODELING THE DYNAMICS OF CRISPR ARRAYS IN MICROBIAL COMMUNITIES

Memoria presentada por Jimena Martín Reina para la obtención del título de Graduada en el máster en Biología Computacional por la Universidad Politécnica de Madrid

Fdo: Jimena Martín Reina

VºBº Tutor UPM

Don Carlos Pérez Cantalapiedra
Dpto. de Biotecnología-Biología Vegetal
Centro de Biotecnología y Genómica de Plantas (CBGP) - Universidad Politécnica de Madrid (UPM)

VºBº Tutor externo

Don Jaime Ignacio Iranzo Sanz
Dpto. de Evolución molecular
Centro de Astrobiología (CAB) – Consejo Superior de Investigaciones Científicas (CSIC)

Madrid, julio 2025

Agradecimientos

Tengo la inmensa suerte de haber escrito este Trabajo de Fin de Máster. Es una inmensa suerte, porque la serie de condiciones que han tenido que cumplirse para que yo haya acabado redactando este trabajo no son para nada triviales.

He nacido en una familia que me ha apoyado e impulsado a perseguir mis objetivos académicos.
Me he criado rodeada de libros accesibles para una mente curiosa.
Mis profesores en el colegio y en el instituto me han motivado a seguir el camino del estudio.
El momento de entrar a los estudios superiores se ha dado en un tiempo y lugar en el que no se me ha prohibido ir a la universidad por mi condición.
La música, la literatura, el teatro, la pintura… me han consolado cuando lo he necesitado.
Una red de amistades ha hecho que las lágrimas derramadas merecieran la pena.
Cierto guitarrista ha evitado que muchas veces me fuera al *traste*.
Vivo desde siempre en una pequeña *burbuja acorazada* sin que me hayan ocurrido la guerra, pobreza, hambre, explotación, ni desastres naturales.
Se me han ofrecido infinidad de becas y ayudas que han hecho posible mi camino.
No podría haber recibido más amabilidad de las personas que he encontrado durante mi estancia en el CAB, especialmente de mis compañeros del grupo.
La casualidad ha hecho que conociera y contara como tutor con un científico encomiable.

La lista no termina ahí, pero con estas líneas trato de aludir a aquellos que han contribuido durante mi vida a que yo escriba estos mismos agradecimientos. Necesitaría muchas vidas más para compensar lo que habéis hecho por mí.

Index

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF TABLES

LIST OF ABBREVIATIONS

- CRISPR-Cas: Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated genes
- DNA: Deoxyribonucleic acid
- BLAST: Basic Local Alignment Search Tool
- tracrRNA: trans-activating CRISPR ribonucleic acid
- RNA: ribonucleic acid
- crRNA: CRISPR ribonucleic acid
- cas: CRISPR-associated
- RNase: ribonuclease
- PAM: protospacer adjacent motif
- MGE: mobile genetic element
- Nsp: number of spacers
- a: alpha
- g: gamma
- b: beta
- rnum: random number

ABSTRACT

CRISPR-Cas systems provide adaptive immunity in prokaryotes by incorporating short sequences of viral DNA, known as spacers, into the CRISPR arrays in the host genome. These spacers enable defense against future viral infections, which can be persistent in the long-term (endemic) or transitory (epidemic). The finite size of CRISPR arrays imposes a trade-off between remaining updated and retaining long-term memory in complex environments which is still subject to understanding.

This thesis presents a mathematical and computational model that simulates the evolution of CRISPR arrays in microbial communities, distinguishing between spacers targeting endemic versus epidemic viruses. The model explores a continuous spectrum of viral scenarios—from exclusively endemic to purely epidemic compositions—while incorporating parameters that modulate selection pressure, crossover point between viral incidences, and the relative abundance of endemic viruses. Simulations reveal that these parameters govern whether CRISPR arrays prioritize short-term or long-term immune memory.

Results indicate that long-term immunity emerges under low selection pressure and stable viromes, conditions consistent with observations in environments such as the human gut microbiome. In contrast, environments with high viral turnover, such as hydrothermal vents, are expected to favor rapidly updating arrays with short-term specialization. The model successfully reproduces key patterns observed in empirical studies and offers testable predictions about array structure and ecological adaptation.

Although simplified, the model identifies fundamental variables influencing CRISPR array dynamics and provides a foundation for further empirical integration. It offers a conceptual framework for understanding immune memory in prokaryotes and guiding future studies on microbial evolution in diverse viral ecosystems.

CHAPTER 1: INTRODUCTION

1.1. State of the art: CRISPR system and spacer dynamics

In the words of Eric S. Lander: *"It's hard to recall a revolution that has swept biology more swiftly than CRISPR"* [1].

The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) system has gained public attention due to its unprecedented applications in genome editing. While it is now a cornerstone in molecular biology and biotechnology, public and even scientific recognition often centers more on its utility as a tool than on its biological origins and natural dynamics [2]. However, this is understandable considering the incredible advances of this technology, like helping a carbamoyl-phosphate synthetase 1 deficient baby [3].

CRISPR research history did not start as a quest to edit the human genome, but it was born out of the curiosity of Francisco Mojica.

When he began his doctoral studies in 1989 at the University of Alicante, he joined a laboratory working on an archaeal microbe with extreme salt tolerance that had been isolated from Santa Pola's marshes. When examining the first DNA fragments of this species, Mojica found a curious structure—multiple copies of a near-perfect, roughly palindromic, repeated sequence of 30 bases, separated by spacers of roughly 36 bases—that did not resemble any family of repeats known in microbes [1].

Figure 1. Natural Park of the Santa Pola Salt Flats. Image from natural sea salt producers Bras del Port's webpage. (https://www.brasdelport.com/parque-natural-de-las-salinas-de-santa-pola/)

Similar repeats found in other species made Mojica think it must be related to an important function in prokaryotes. The confirmation came three years later.

He extracted the spacers from this structure and searched for similarity with any other known DNA sequence. Two-thirds turned out to match viruses or conjugative plasmids related to the microbe carrying the spacer. Mojica then realized that CRISPR loci must encode the instructions for an adaptive immune system that protected microbes against specific infections [1].

Besides Mojica, the contributions of researchers such as Gilles Vergnaud, Alexander Bolotin, Philippe Horvath, John van der Oost, Eugene Koonin, Luciano Marraffini, Emmanuelle Charpentier, Jennifer Doudna and many others were essential to discover everything we know today about CRISPR-Cas system [1].

Thanks to their efforts, we know how the general CRISPR-Cas mechanism works. Following, we explain the mechanism corresponding to the Type II system, which has been the basis for genetic editing and is represented in Figure 2.

(A) The locus contains a CRISPR array, four protein-coding genes and the tracrRNA. The CRISPR array contains repeat regions (black diamonds) separated by spacer regions (colored rectangles) derived from previous infections. The cas9 gene encodes a nuclease that confers immunity by cutting invading DNA, while the cas1, cas2, and cns2 genes encode proteins that function in the acquisition of new spacers from invading DNA.

(B) The CRISPR array and the tracrRNA are transcribed, giving rise to a long pre-crRNA and a tracrRNA.

(C) These two RNAs hybridize via complementary sequences and are processed to shorter forms by Cas9 and RNase III.

(D) The resulting complex (Cas9 + tracrRNA + crRNA) then begins searching for DNA sequences that match the spacer sequence (shown in red). Binding to the target site also requires the presence of the protospacer adjacent motif (PAM).
(E) Once Cas9 binds to a target site with a match between the crRNA and the target DNA, it cleaves the DNA [1].

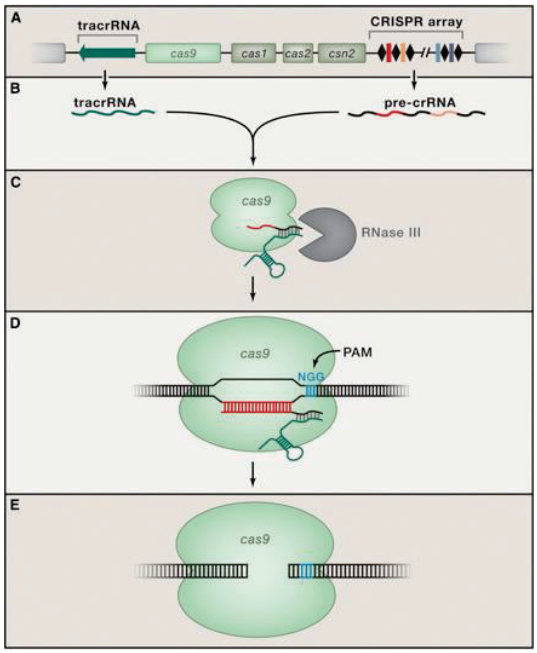

Figure 2. Class2, Type II CRISPR-Cas9 System from Streptococcus thermophilus. Figure from Eric S. Lander (2015).

However, there are six CRISPR types divided into two classes, differentiated by their effector. The classification, representative members, and typical characteristics of each CRISPR-Cas system are summarized in Table 1 from Annex I [4].

In the 2000s, experiments showed that new spacers are added to the array in a process (adaptation) wherein two proteins, Cas1 and Cas2, integrate DNA fragments [5]. As integration typically occurs at the end of the array called "leader" [6,7], it is polarized: the leader end contains diversity in spacers while the distal end is conserved [5]. This pattern supports that adaptation is sequential: newest spacers are found in the leader end, while the oldest ones are in the distal end. Moreover, in some studies it was shown that the spacer's position is a time stamp for the moment the CRISPR-Cas captured the virus' genetic material [8].
Studies have shown that adaptation often occurs immediately after viral DNA injection, and that spacers acquired from early-injected genomic regions provide better immunity than spacers acquired from late-injected regions [9].
Apart from adaptation, processes like duplication and deletion of spacer-repeat units occur randomly in the array and can be selected in the population if they contribute to the species' fitness [10,11]. Spacers that confer immunity against prevalent phages are often retained, especially in environments with consistent viral pressures [12].

1.2. Ecological relevance of CRISPR

Microbial communities, and therefore natural environments, are very influenced by bacteria-phage interactions. Phages significantly impact bacterial populations by causing cell death, thereby influencing nutrient cycles; in response, bacteria have evolved the CRISPR-Cas adaptive immune system, which drives phage genome rearrangements and escape mutations—evidenced by the high frequency of mutations in regions of MGE targeted by CRISPR spacers [13].
Just as Prometheus chained to the rock, eternally condemned to regenerate his liver after Zeus's eagle devours it each day, bacteria find themselves in an unending coevolutionary entanglement with their viral predators. This is a classic example of an evolutionary arms race.
Therefore, geographically distinct microbial populations exhibit unique CRISPR spacer profiles, reflecting the hosts' CRISPR adaptation to local viral populations [12].
Some recent studies have shown that in long-term coexistence of MGE and hosts, we find polymorphic hosts populations, including immune and sensitive lineages [12]. However, the lack of studies regarding how CRISPR and microbial evolution behave in the complex environments found in living hosts, as humans, makes it difficult to draw conclusions. Some studies show immunodominance of a single host lineage in the human gut and lung [12], so it is

not clear that the previous statements can apply to this environment.

Moreover, whether distal spacers provide long-term immune memory is under discussion. While the fact that they match conserved regions of viral genomes may be a sign of maintaining these spacers for long-term immunity [14]; the low transcription rates and the broad conservation of this region [5] could lead us to the conclusion that they do not play a relevant role in defense against local MGE. This should be further explored in future research, as it is used in mathematical models as a general assumption. These models conclude there is a balance between having a big immune repertoire and the molecular machinery limitations [15]; between updating CRISPR arrays and losing immune memory too fast [16]; and between viral defense and autoimmunity [17].

Another aspect that requires additional investigation is the differential impact of acute or epidemic versus chronic or endemic viral infections on CRISPR array spacer dynamics. It has been hypothesized that the former may lead to rapid spacer turnover, while the latter to retention of spacers in the long-term [12]. In this work, we will focus in exploring this effect, together with how the position of the spacer in the array can inform us about its effectiveness for the host's immunity.

We aim to connect the viral context to which a bacterial lineage is exposed with certain statistical properties of its CRISPR system. This would allow us to know relevant aspects about the way in which phages affect different bacterial populations from the analysis of metagenomics data, with possible applications to the development of microbiome control strategies as phage therapy.

1.3. Objectives of the study

The aim of this study is to gain understanding of CRISPR-Cas immune systems, in particular, the dynamics of spacers gain and loss and the trade-off between update and long-term immunity in the context of the virome. To achieve this, we explore the CRISPR array properties in different viral environments to gain significant patterns or behaviors about spacers turnover and distribution in the array and extrapolate them if possible.

To reach the main goal, we define the following specific objectives:
1. Create a mathematical and computational model that simulates gain and loss of spacers in the CRISPR array, distinguishing from endemic and epidemic-targeting spacers.
2. Provide the model with parameters that can represent different viromes, selection pressures and dominance states between acute and chronic infections.
3. Reproduce experimental results with the constructed model.

CHAPTER 2: MATERIALS AND METHODS

2.1. Mathematical formulation of the model

2.1.1. Assumptions

The following assumptions form the basis of the model:

- The model simulates a **single CRISPR array** as a representative for a bacterial community. In nature, multiple arrays may coexist within a single genome or across different individuals in a population [2,18,19], but, for simplicity, we represent only one.
- The **size** of the CRISPR array is held **constant** throughout the simulation. In reality, CRISPR array lengths can vary between individual bacteria, within a single cell, and over time [2,18,19]. We did not model the phenomena of duplication and deletion that occur in the middle of the array [5].
- Reflecting **natural selection**, the model favors spacers that contribute more to bacterial fitness without strict optimization: spacer loss is probabilistic, with lower-fitness-contribution spacers more likely to be removed.
- The model assumes that **bacterial survival** depends exclusively on CRISPR-based immunity. Consequently, the array simulated is interpreted as the one most likely to persist under selection. In nature, however, bacterial fitness is shaped by many other factors, and CRISPR immunity is only one among them.
- Viral dynamics are modeled with **exponential decay**, which is a simplification. Actual phage incidence patterns may follow different, more complex trajectories depending on ecological and host-related factors.
- The **ratio of endemic to epidemic viruses** is fixed during a simulation. However, empirical studies show that this ratio fluctuates over time in natural environments, responding to changes in host populations, environmental conditions, and virus-host coevolution [20].
- We did not consider that the last spacer-repeat unit rarely participates in rearrangements—according to literature, potentially because of polymorphisms [5].

2.1.2. Definition of variables and parameters

To simulate the dynamics of a CRISPR array, we define the following variables:

- **Incidence (*W*):** Abundance of a specific virus in the simulated environment, measured in arbitrary units.
- **Fitness contribution:** Probability that a spacer contributes to host survival during a single time step, by providing immunity against its corresponding phage.
- **Age (*t*):** Number of generations (since acquisition) a spacer has remained in the array.
- ***Beta* (*b*):** Decay rate of viral incidence over time. In the case of strict endemic viruses, ***b=1*** (we used 0.9999 for practical reasons). Values <1 map the full continuum from fast epidemics (***b*** close to 0) to long persisting viruses (***b*** close to 1).

We also distinguish:

- **Epidemic virus:** High initial incidence, rapidly decline, and disappear.
- **Endemic virus:** Lower initial incidence, slowly decline, and persist longer.
- Although we use a discrete **classification of viruses** as endemic or epidemic, their behavior represents a spectrum.
- **Proximal end / Distal end:** The proximal end of the array is where new spacers are added; the distal end contains the oldest spacers.

The basic parameters of the model are:

- ***Alpha* (*a*):** Controls the rate of new virus influx. It can be interpreted as a selection parameter. Defined on the interval (0, 1].
- ***Gamma* (*g*):** Time point at which incidence curves of endemic and epidemic viruses intersect.
- ***pEndemic*:** Proportion of new spacers that target endemic viruses. Value between 0 and 1.

2.1.3. Equations governing spacer dynamics

Incidence over time is defined as:

Equation 1

$$W(t) = log\left(\frac{1}{a}\right) \cdot b^{-g} \cdot b^{t}$$

Since $b \in (0,1)$, b^t ensures that incidence exponentially decays over time. Endemic viruses (***b*** close to 1) decay slowly; epidemic viruses (low ***b***) decay quickly.
The term b^{-g} shifts the incidence curve for epidemic targets (for endemic the term will be close to 1). High incidences will be reached for low ***b*** and high ***g***.
As ***a*** approaches 0, ***log(1/a)*** increases, raising the overall incidence. For ***a=1***, incidence would be 0.
These effects are shown in Figure 3.

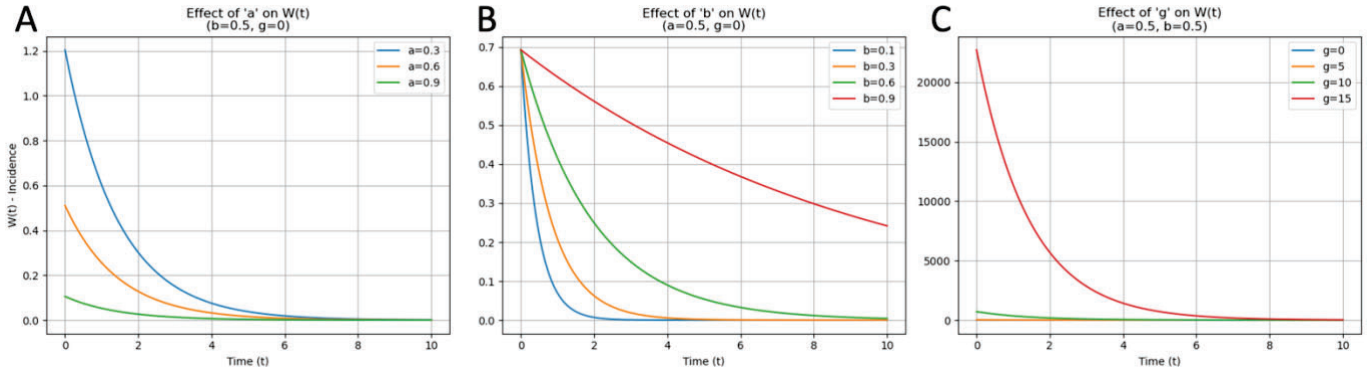

Figure 3. Incidence (W) over time for different values of parameters a (Figure A), b (Figure B) and g (Figure C).

Fitness contribution is derived as:
Equation 2

$$fitness = 1 - e^{-W(t)}$$

Higher incidence implies greater utility of a spacer, thus increasing its likelihood of retention.
Fitness can also be interpreted as the probability of encountering a virus with incidence ***W*** in a time step, based on the probability of observing at least one event in a Poisson process with incidence rate ***W***.

Therefore, the probability of removing a certain spacer is called ***Pdisp*** (Equation 3) and is obtained from the term ***Qi*** divided by the summation of Q_i for all the spacers. Replacing ***W*** in the fitness expression and operating with the logarithms, we get the following equation.

Equation 3

$$P_{disp\ i} = \frac{Q_i}{\sum Q_i};\ Q_i = 1 - fitness = a^{b^{(t-g)}}$$

2.2. Computational implementation

The source code (in C++ and Python) and datasets used in this study are available at my GitHub repository.

2.2.1. Transition from equations to algorithm

The model simulates CRISPR array evolution and virus dynamics in discrete time steps. At each step, the CRISPR array undergoes one deletion and one insertion event:

- A spacer is removed from the array with a probability ***Pdisp*** (Equation 3) —spacers that provide less protection are more likely to be lost.
- A new spacer is inserted at the proximal (leader) end of the array, shifting existing spacers toward the distal end.

The newly inserted spacer targets either an endemic or an epidemic virus. With probability ***pEndemic***, it targets an endemic virus and is assigned a decay rate parameter ***b = 0.9999***. With probability ***1 – pEndemic***, it targets an epidemic virus, and ***b*** is drawn from a uniform distribution between 0 and 1.

Throughout the simulation, each spacer's properties—such as age, fitness, incidence, and ***b***—are updated to reflect changes in viral abundance and array composition.

Apart from the parameters mentioned in Section 2.1.2, the model needs the following auxiliary parameters to create the simulated environment:

- **Array size (*Nsp*):** Total number of spacers in the CRISPR array. Fixed during the simulation. We use 10 and 40 as representative values based on empirical data [2,18,19, 21].
- **Numer of generations to simulate:** Total number of time steps the simulation takes.
- **Number of replicates:** Total number of parallel simulations done when running the model code.

Output Parameters:

- **Beta (*b*), Age, Fitness, Incidence (*W*):** Also recorded as output variables during simulation.
- **endemicBeta, epidemicBeta:** Mean ***b*** values for spacers targeting endemic or epidemic viruses, respectively, at each array position.

2.2.2. Initial "toy model" in C++

The initial version of the model was implemented in C++, based on a foundational codebase provided by my advisor, which I later fixed and developed. C++ was chosen primarily for its computational efficiency and suitability for running iterative simulations.

The code uses fixed values for parameters such as array size, efficacy thresholds, and simulation length.
Random number generation in this implementation relies on a time-based seed, introducing stochasticity tied to the system clock. This ensures variation across runs without affecting reproducibility, as identical outcomes can be guaranteed by manually fixing the seed, which appears in output files.

Several programming tasks in C++—such as generating random numbers or managing data arrays—required custom implementations or verbose code.

The libraries used in the C++ scripts are part of the standard C library [22].

2.2.3. Translation to Python

After validating the initial model in C++, I translated the core functionality into Python. This translation served primarily as an exercise to ensure I fully understood the model's structure and logic. Python was chosen for this step because it is the language I am most familiar with, and the translation was encouraged by my advisor for learning and verification purposes.

The Python version successfully replicated the behavior of the C++ model using equivalent logic. The translation benefited from Python's built-in functions and high-level syntax, allowing for a more compact and readable implementation.

However, the Python model was considerably slower in execution time, especially when running large-scale simulations or multiple iterations. As performance was critical for this study, all subsequent development and analysis were carried out exclusively using the C++ version. The Python translation thus served as a one-time validation step and was not used for further simulations or model extensions.

The code uses random [23] and numpy [24] libraries.

2.2.4. Model modifications and extensions

Throughout the development process, several modifications were made to the original toy model to correct

implementation issues and enhance functionality.

Adaptive simulation time.
To ensure the system reached a steady state without unnecessarily long simulations, a dynamic stopping condition was implemented. The simulation starts with a default number of generations (5,000), and if the last spacer's age is not less than one-fifth of the total time, the simulation length ***Tmax*** is doubled and rerun. This condition ensures the spacers have enough time to stabilize, while avoiding wasteful computation.
Python-based tests showed that the system's behavior at T=200,000 was nearly indistinguishable from that at much shorter times (e.g., T=447), indicating that replicate count had more influence on variability than simulation time (Section 3.3.).

The parameters used in these simulations were arbitrarily set to:
alpha = 0.1000, gamma = 0.0000, pEndemic = 0.2000, Nsp = 40, Tmax = 3, Num. replicates = 1000

To illustrate convergence behavior, we represent the resulting curves from running the model at ten different times. The final time point was set to 200,000 generations, where the system is known to have reached a stationary state. The intermediate time points were chosen using a logarithmic scale:

Equation 4

$$unit = \frac{\log_{10}(200000)}{10}$$

Each time then will be defined as:

Equation 5

$$t_i = 10^{i \cdot unit} \; for \; i = 1, 2, \ldots, 9$$

Handling extreme values.
For extreme parameter combinations (e.g. ***g=60*** and ***Nsp=40***), efficacy values could become very large, causing numerical overflow. To prevent this, efficacy values were stored in log scale, and the fitness function was computed as
Equation 6

$$fitness = 1 - e^{-e^{\log(W)}}$$

only when efficacy was below a threshold (***log(W) < 40***). Above that, fitness was set to 1 directly, since the exponential of very large values is functionally indistinguishable from 1 in this context.

We compared the results of the model with and without this implementation, and they were qualitatively the same.

2.3. Simulation setup

2.3.1. Parameter selection and computational experiments

To explore the model's behavior across a wide range of conditions, we developed a Python script that automatically generates combinations of parameter values. The parameters included in these combinations were ***a***, ***g*** and ***pEndemic*** (Section 2.1.1).

Values for ***a*** were generated logarithmically between 0.1 and 1. Values of 0 and 1 were excluded to avoid undefined or trivial results.
For ***g***, we initially considered 5 equidistant values from 0 and 20. For ***pEndemic***, values ranged from 0.0 to 0.9 in increments of 0.1. The extreme value ***pEndemic = 1*** was excluded since it is trivial.
Values were exported in a txt format used later to run the simulations (Section 2.3.2.).

Simulations were conducted using two array lengths: an array of 10 spacers, and another of 40 spacers. Initially, each array configuration was tested using the same 450 parameter combinations, covering the ranges described above. However, results revealed interesting dynamics for ***Nsp = 10*** with ***g > 10***, motivating further exploration. To ensure comparability for the case where ***g > Nsp***, we extended the parameter set for ***Nsp = 40*** to include ***g = 60***, resulting in 90 additional simulations (for a total of 540 simulations for ***Nsp = 40***). These extended runs led to challenges with overflow, discussed previously in Section 2.2.5.

Finally, 990 simulations were done in total, a vast exploration of different conditions that ensures a robust sampling of model behavior.

2.3.2. Execution on high-performance computing clusters

All simulations were executed on the JWST cluster at the Astrobiology Center.
Although the original goal was to perform 10,000 replicates per parameter set, memory access restrictions on the cluster prevented this. As a workaround, simulations were run in two rounds of 5,000 replicates each. The results from both runs were then averaged using a custom Bash script to obtain the final replicate means.

Simulations were launched using a Bash script that automated the process: it read parameter sets from the parameters.txt file (Section 2.3.1), executed the compiled C++ model for each line, and saved the resulting outputs. Jobs were distributed across five parallel background processes. For the case with ***g = 60***, all simulations were run serially in a single process.

2.3.3. Development of data visualization and analysis tools

All simulation outputs were analyzed in Python Jupyter notebooks. Each simulation produced a ***.txt*** file, which was batch-loaded using ***glob*** and ***os*** [23] and processed with pandas [25].
Initially, output variables were visualized as line plots across spacer positions for all simulations to gain an overview; in some cases, a color code based on one input parameter was applied to distinguish between simulations.

For more targeted analysis, specific curve features were extracted.
For fitness curves: two flanking maxima, minimum, ratio between maxima (left/right) and Euclidean distance between the midpoints of each flank with the minimum. The Euclidean distance served as a proxy for how "open" or "closed" the curve is: larger distances indicate wider, more spread-out curves, while smaller distances reflect tighter, more compact patterns.
For beta curves: initial and final value (as reference), minimum position and value, and the position where 99.8% of the maximum is reached.
For age curves: maximum of the curve.
These values were saved into summary ***.csv*** files using a custom script.

Further analysis included histograms, boxplots, and heatmaps, with input parameters (***a***, ***g***, ***pEndemic***) and output features independently sorted to improve pattern detection. This dual-sorting approach in heatmaps helped highlight trends and groupings across simulations. Due to the varying ranges of feature values, heatmaps represented scaled data using MinMaxScaler [25], which scaled each one to a [0,1] range. Color-coded plots for these features were generated with one input parameter on the x-axis and the other two represented by color and transparency. When needed, encoding combinations were adjusted to reveal trends better.
The Python libraries used for this part were glob, math and os as part of Python's standard library [23], numpy [24], pandas [25], scikit-learn [26], seaborn [27], scipy [28], and matplotlib [29].

2.4. Preliminary analytical study

2.4.1. Calculation of average values

To start with a preliminary analytical study, we first compute the expected value of the parameter ***b***.
Each new spacer's ***b*** value is drawn from a mixture distribution. With probability ***pEndemic***, the spacer targets an endemic phage, and its ***b*** value is deterministically 0.9999 (Dirac delta). With the remaining probability ***1−pendemic***, ***b*** is uniformly distributed in the interval [0,1]. Therefore, the expected ***b*** value of a new spacer is:

Equation 7

$$E[b] = pEndemic * 0.9999 + (1 - pEndemic) * 0.5$$

As spacers are subject to selection, those with higher match probabilities are more likely to persist in the array, affecting the average ***b*** value in a certain position. Therefore, this formula gives the expected ***beta*** value for the first spacer in the array, which remains unaffected by selection.

2.4.2. Selection dynamics: dominance conditions

If we apply logarithms to Equation 1, the nested exponentials are removed, and we obtain a linear expression:

Equation 8

$$\log\big(W(t)\big) = \log\left(log\left(\frac{1}{a}\right)\right) + \log(b) * (t - g)$$

This coincides with a typical straight-line form:

Equation 9

$$f(x) = C + m * x$$

Here, ***log (log (1/a))*** is a constant term (specific to each simulation), and ***log(b)·(t−g)*** is the variable term, with ***log(b)*** acting as the slope and ***t−g*** as the shifted time variable.
When the constant term dominates, ***log(W(t))*** is largely unaffected by time or the spacer's ***b*** value, meaning that all spacers behave similarly regardless of their viral target type. In contrast, if the variable term dominates, differences in ***b*** and time become crucial in shaping the efficacy dynamics, and selection for endemic-target spacers will occur.

The constant term dominates the behavior of ***log(W(t))*** when the variable term remains small in absolute value. This occurs under the following conditions:

- ***b*** is close to 1 (endemic-target spacers), so ***log(b)≈0***,
- ***t*** is close to ***g***, minimizing time-dependent effects,
- ***a*** is small (close to 0), making the constant term large in magnitude.

In contrast, the variable term dominates when:

- ***b*** is close to 0 (epidemic-target spacers), making ***log(b)*** strongly negative,
- spacers are old (***t≫g***), so the product ***log(b)·(t−g)*** grows in magnitude,
- ***a*** is close to 1, reducing the constant term.

Dominance here refers to relative magnitude, not sign: the dominant term is the one contributing most significantly to the value of ***log(W(t))***.

Therefore, in the region where ***b→1*** and ***t≈g***, the system behaves neutrally, with ***log(W(t))*** nearly constant and independent of viral dynamics. Elsewhere, the spacer efficacy becomes strongly dependent on both time and ***b***, and selection pressures vary accordingly.

Beyond this mathematical dominance, the model also exhibits a **dominance shift** between epidemic- and endemic-targeting spacers. By setting the incidence (***W***) expressions equal for the two spacer types (according to Equation 1 and the value of ***b*** for endemic-targeting spacers ***(b=1)***):

Equation 10

$$\log\frac{1}{a} = \log\frac{1}{a} * b^{t-g}$$

Since ***b<1*** for epidemic spacers, ***log(b)<0***. Therefore:

Equation 11

$$1 = b^{t-g} \rightarrow (t-g) * \log(b) = 0 \rightarrow t = g$$

Thus, at ***t=g***, the incidences of epidemic- and endemic-targeting spacers intersect.

- Before ***t=g***, epidemic spacers have higher incidence → epidemic dominance,
- After ***t=g***, endemic spacers have higher incidence → endemic dominance.

This marks a **transition point** in the selection dynamics: early in a spacer's lifespan, epidemic spacers are favored due to their initially high incidence. Later, as their incidence decays, endemic spacers—whose efficacy decays more slowly—become more beneficial. The parameter ***g*** therefore sets the **timing of this shift**.

CHAPTER 3: RESULTS

3.1. Morphological patterns in simulation outcomes

3.1.1. Fitness contribution of spacers

Across the initial set of 450 simulations, we identified **four distinct morphological patterns** in the spatial distribution of spacer fitness within the CRISPR array. Here, fitness refers to the contribution of each spacer to host immunity: a low fitness value indicates that the virus targeted by that spacer is rarely encountered, while a high value suggests the virus is common and the spacer plays a central role in immune defense. These patterns were defined by the presence or absence of **left and right plateaus** in the fitness profile and were the same for both array sizes (Section 3.4), though we present Nsp=40 results here.

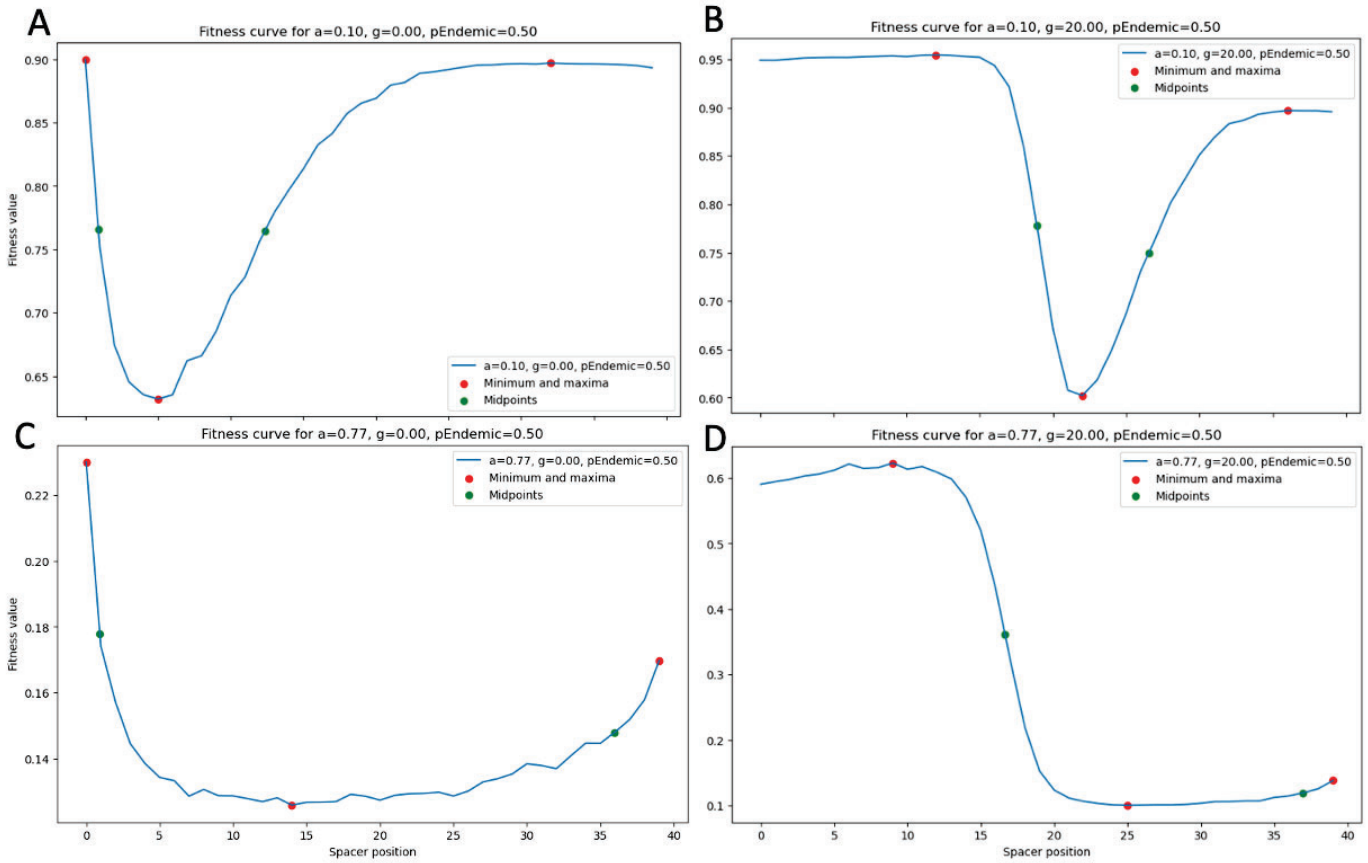

Figure 4. Typical morphologies of fitness contribution found in 40 spacers-long array simulations. The four curves were obtained for pEndemic=0.50 as a fixed value. Each curve differs in values of parameter a and g.

The **left plateau** is associated with the parameter ***g*** (gamma). Gamma defines the critical time at which the incidence curves for all viral types intersect. Before this point, epidemic-targeting spacers have higher incidence than endemic-targeting ones; after this point, the reverse is true. Thus, ***g*** determines **how long epidemic viruses dominate the selective landscape**.
When ***g*** is greater than zero, it creates a temporal "window" in which epidemic-targeting spacers—despite their short-lived nature—provide the highest immediate fitness benefit due to their initially high incidence. This dominance is reflected in the fitness profile as this **extended initial plateau** at the beginning of the array, present in Figure 3B and 3D. As time progresses, epidemic incidences decay, leading to the characteristic drop following the plateau.
In this way, larger ***g*** values result in broader initial plateaus before endemic spacers begin to dominate.

The **right plateau** depends on parameter ***a*** (alpha). Lower ***a*** values (i.e., higher incidence) increase selection pressure for endemic spacers. As a result, the lowest-fitness spacers are more likely to be removed, driving the system closer to an optimal solution. In contrast, when ***a*** is large (approaching 1), the fitness values of all spacers become more similar—identical in the limit ***a = 1***—reducing the efficacy of selection. Therefore, parameter ***a*** effectively controls the **selection intensity** in the system, interpolating between a neutral regime ($a \approx 1$) and strong selection ($a \approx 0$).

The parameter ***pEndemic*** only substantially affects the morphology in its extremes. When ***pEndemic=0***, all spacers target epidemic viruses, and selection only acts on close to endemic spacers (***b*** randomly close to 1). This eliminates the right plateau, because few spacers will be endemic by chance (in a uniform distribution from 0 to 1). However, for all other values of ***pEndemic***, the overall morphological patterns remain qualitatively stable, with only minor variations in shape.

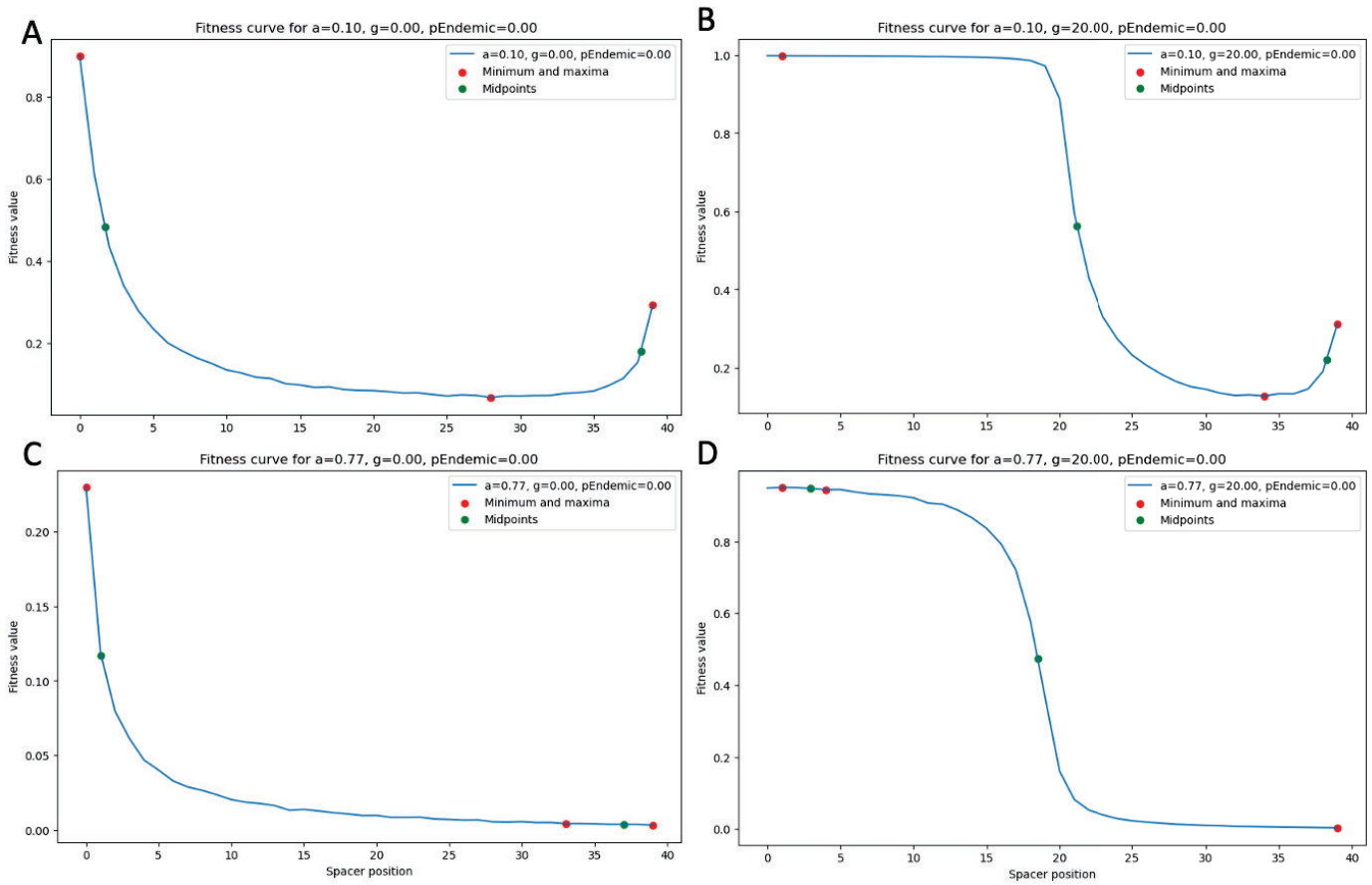

Figure 5. Corresponding morphologies showed in Figure 3, but with pEndemic=0, showing the effect of this extreme value. Figure D is the most similar morphology to pEndemic different than 0 (Figure 3D).

The variations in shape due to ***pEndemic*** are shown below. Higher values (close to 1) increase the proportion of endemic-targeting spacers, enhancing their representation in the array over time. Since these spacers are responsible for the long-term recovery of fitness, their abundance results in a **more extended right plateau**. In contrast, when ***pEndemic*** is low (approaching 0), endemic spacers are scarce and rarely selected. Because epidemic-targeting spacers do not sustain fitness over time, the resulting profiles often show only a **short or absent plateau**, as recovery is limited.

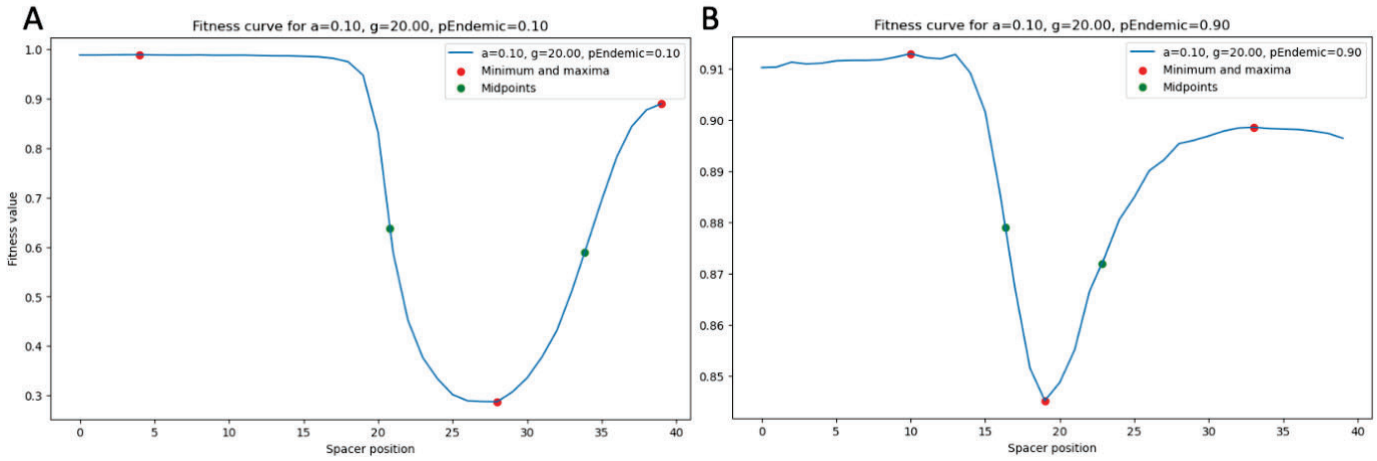

Figure 6. Comparative between extreme values of pEndemic. We see the morphology is similar, but the right plateau is lacked in A.

A particularly interesting behavior emerged when ***g*** exceeded the array length (***Nsp***). In these cases, the crossover point where the incidences of epidemic- and endemic-targeting spacers are expected to converge (***t=g***) is never reached during the lifespan of the spacers in the array.

As a result, **epidemic-targeting spacers dominate early selection**, leading to the rapid removal of endemic-targeting spacers—despite their long-term advantage. However, **epidemic spacers are short-lived**, as their incidence drops quickly with time. Positions in the array are taken by epidemic-targeting spacers, so **endemic spacers do not survive long enough to reach the point where their incidence would eventually surpass that of epidemics**. In other words, neither spacer type persists long enough to reach the shift in dominance at ***t=g***, so endemic spacers never "benefit" from this change.

This dynamic creates a characteristic fitness profile: an initial slight rise, as low-incidence endemics are removed, followed by a sharp decline when the short-lived epidemics lose efficacy.

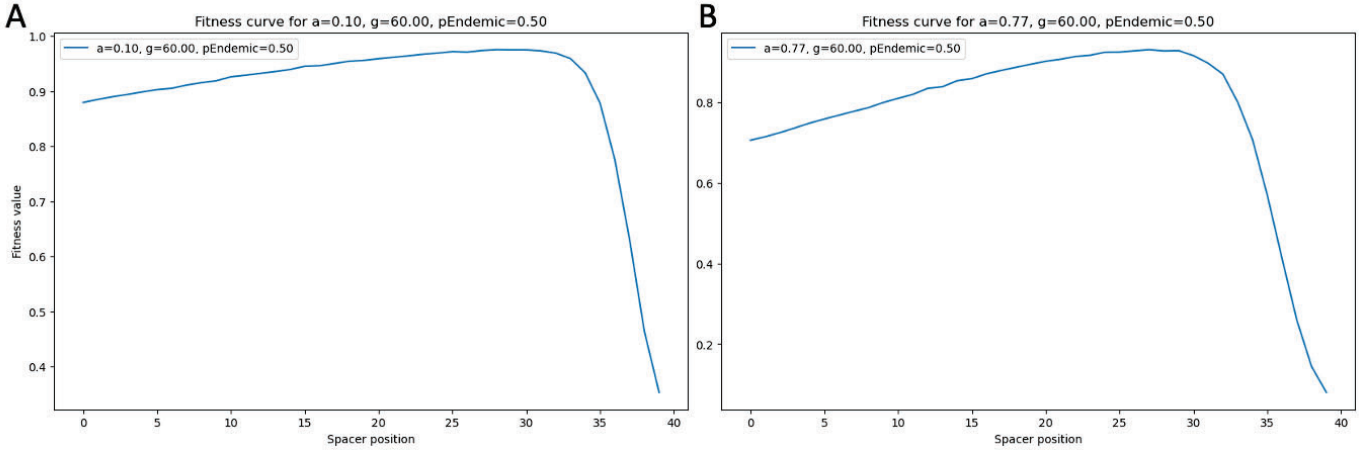

Figure 7. Fitness contribution morphology for g>Nsp, in this case g=60 and Nsp=40. Variations in parameter a do not modify this concrete morphology.

In this case (***g=60***), parameter ***a*** is not so determining in the morphology of fitness contribution along the array. On the contrary, parameter ***pEndemic*** does have a significant effect.
If the value of ***pEndemic*** is high enough (close to 1), some endemic spacers will survive and will stay in the array. If they stay in the long term, they will pass the cross-point and reach the state in which they dominate. In these cases (***pEndemic >= 0.7***), we have a morphology like the previously mentioned, but with a recovery of fitness at the tail of the array due to the selected endemic spacers.

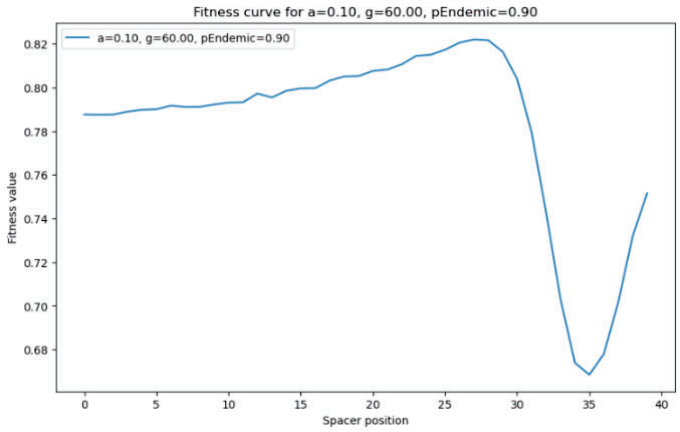

Figure 8. Fitness contribution morphology for g>Nsp, but a high pEndemic, meaning that some endemic-targeting spacers remain in the array and generate the final increase in fitness value.

For low values of ***pEndemic*** (close to 0), we almost have a constant (between 0.98 and 1) value of fitness contribution along the array. In this case, the few endemic-target spacers will not survive the first steps of the simulation, leading to an epidemic-target spacers' hegemony. As we can see in Figure 8, the initial increase is due to the few endemic-targeting spacers that will take the first position with a probability of 0.10; while the decrease in the last positions is due to the decline of viruses incidences with time.
We see this morphology for values of ***pEndemic <= 0.3***.

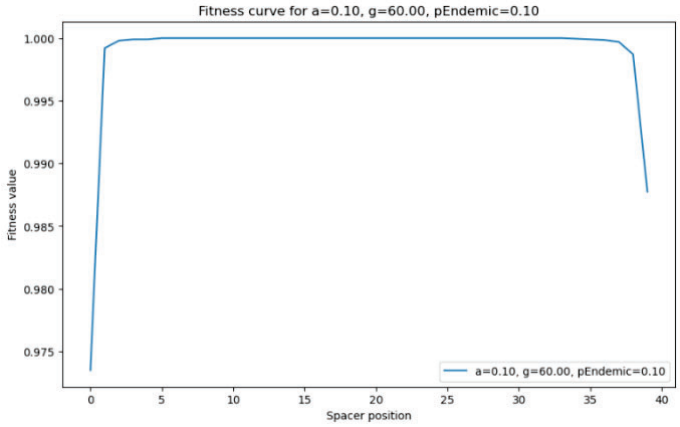

Figure 9. Fitness contribution morphology for g>Nsp but pEndemic close to 0. In this case, there is not selection for endemic-targeting spacers.

3.1.2. Beta values of spacers

The previous conditions of input parameters are represented now regarding the values of parameter beta, which refers to the type of virus targeted by each spacer. The higher the beta, the more specialized the spacer will be for endemic viruses.

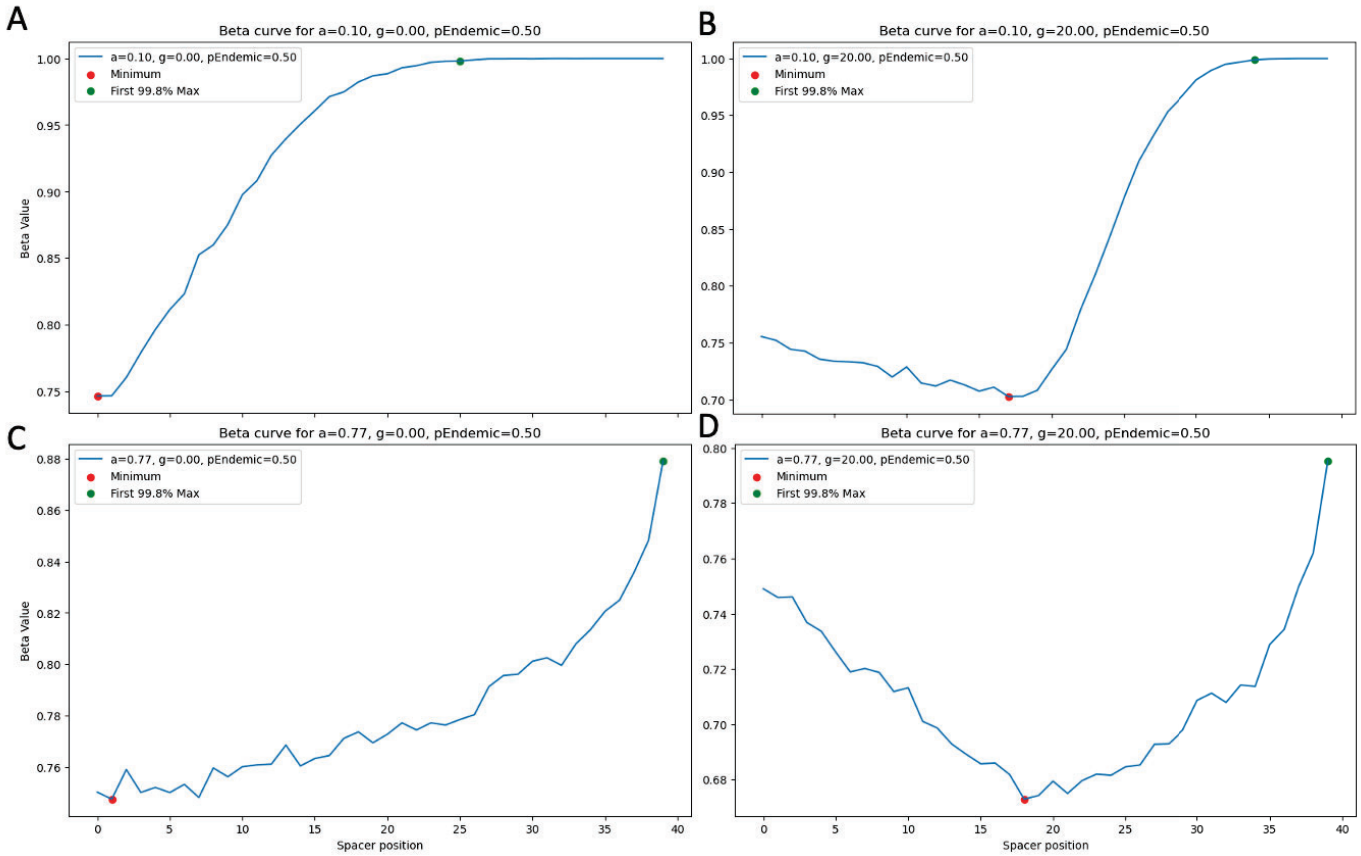

Figure 10. Beta morphologies along the array for the same parameter conditions as in fitness contribution profiles in Figure 3.

These results confirm what we already know from the fitness contribution profiles. Parameter ***a*** restraints endemic-targeting spacers' selection (when close to 1), causing the lack of the right plateau and the value reached in Figure 9C and 9D, which is not the maximum (1) as in Figure 9A and 9B. The final ***beta*** value is between 0.8 and 0.9, reflecting that some epidemic-targeting spacers occasionally persisted to the end of the array, while some endemic spacers failed to do so in certain simulations.

The horizontal shift of parameter ***g*** is also evident in these figures. When simulations directly begin in the endemic-dominance (***g=0***), beta values increase from the start. However, for ***g=20***, the curve initially descends, reaching a minimum shortly before position 20, to then grow. This inflection marks the point where epidemic-targeting spacers—initially dominant—begin to decline, and endemic-targeting spacers start to accumulate.

This decline is elongated in the array for ***g=60***, where the shift of dominance is not reached within the size array. The curves are characterized by a decrease due to the removal of endemic-targeting spacers. Only a small increase in ***beta*** values can be observed in the final positions (36-40), being higher for low values of parameter ***a*** (Figures 10A and 10B).

For a ***pEndemic*** value of 0.9 (Figure 10C), the final position reaches a value of ***beta*** that indicates it contains endemic-targeting spacers (***b=1***), and the increase starts in position 31. (Notice the proportion endemic-epidemic is 90%-10%).

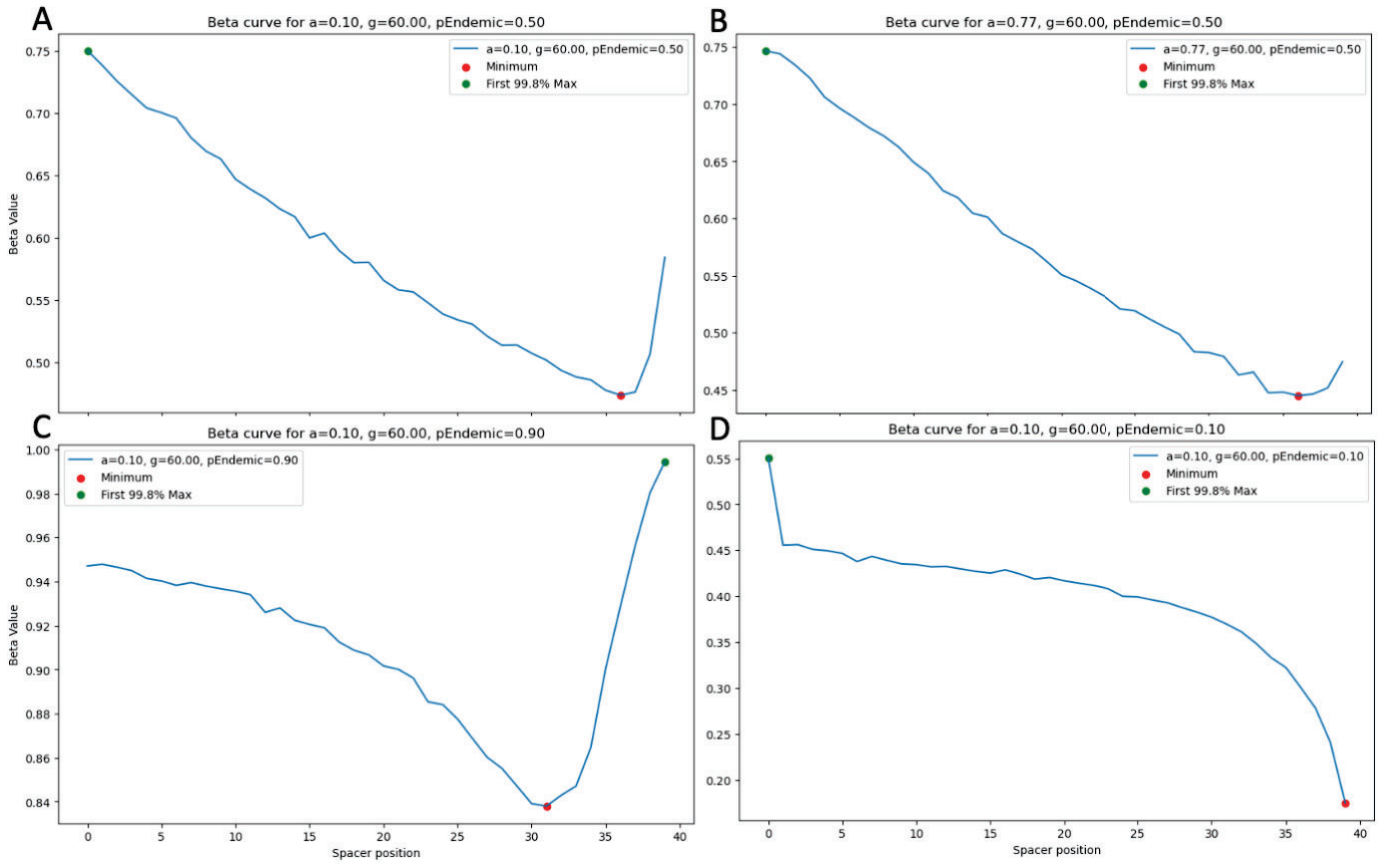

Figure 11. Beta morphologies for g>Nsp and different a and pEndemic values.

Under the parameter combination ***a=0.10*** and ***g=60***, we can observe the influence of ***pEndemic*** on the ***b*** profile (Figure 10D). When ***pEndemic = 0.5*** (Figures 10A and 10B), the curves descend only to approximately ***b = 0.45***—slightly below the expected mean of 0.5 for a uniform distribution of epidemic-targeting spacers. In contrast, when ***pEndemic = 0.10*** (Figure 10D), the curve continues descending, reaching values as low as 0.2 or below. This is the effect of reduced endemic representation: selection favors **highly epidemic-targeting spacers**, which initially have extremely high incidence values.

In this scenario (***g=60***), spacer age is expected to remain low. Epidemic-targeting spacers rapidly lose incidence over time, leading to a decline in their fitness contribution and causing them to be replaced more frequently. Therefore, the array fails to retain older spacers, particularly endemic ones, which never reach the point where their long-term advantage would become relevant.

3.1.3. Age values of spacers

For variable age, there is only one morphology found, a consistent **monotonic increase** in spacer age.
The first half of the array shows a close to linear increase, while the second half, particularly the final few positions, displays a **sharp upward curve**, indicating the accumulation and persistence of older spacers. This steep growth is characteristic of the **selection process favoring long-lived spacers**.

The dashed line represents a reference for comparison. As just reasoned, for ***g=60*** the overall ages of the spacers are very reduced as compared with ***g*** values between 0 ad 20 (figure x).

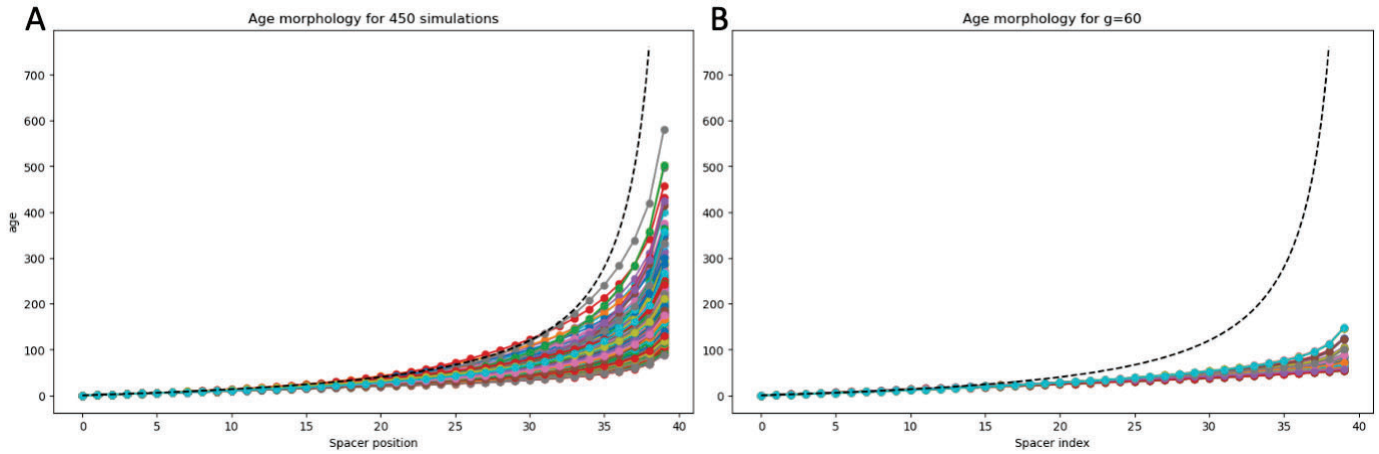

Figure 12. Age distribution across the g<Nsp 450 simulations (A) and the g>Nsp 90 simulations (B), plotted by spacer position.

3.2. Sensitivity analysis of input parameters

3.2.1. Fitness contribution of spacers

We analyzed the association between input parameters and output using heatmaps, focusing here on how parameters affect spacer fitness along the array.

The heatmaps confirm that the existence of the left plateau, which is due to a shift in the initial peak along the first positions of the array, does depend on the parameter ***g***, as noted in Section 3.1. As shown in Figure 12A, parameter ***g*** is correlated with the position of the left maximum. For the maximum ***g*** value (20 in these simulations, represented as 1 because of the data scaling), we get the maximum positions of the left maximum.
However, although the existence and the length of the left plateau depends on the parameter ***g***, the height the plateau can reach depends on the three input parameters.
As parameter ***a*** affects the incidence range (higher values of ***a*** cause lower overall incidences), it also affects the fitness values. Parameter ***pEndemic***, as we mentioned in Section 3.1, has an influence in its extremes. Having no (or few) endemic-targeting spacers means that the initial peak (if elongated, plateau) can reach its maximum potential value.

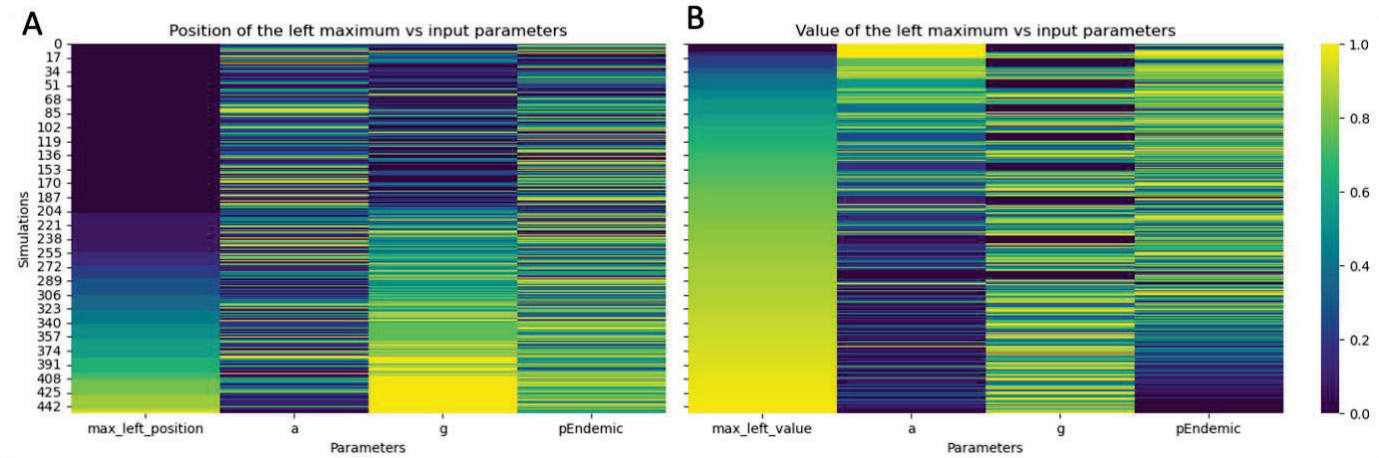

Figure 13. Heatmaps representing on the left column a measure of fitness contribution profile: the position of the left maximum (Figure A) and the value of that maximum (Figure B), next to input parameter values. Each column data was scaled from 0 to 1.

We also mentioned in Section 3.1 that the right plateau depends on parameter ***a***. We confirm with the following heatmaps that this is true, and in this case the length and the height of the right plateau depend on this parameter.
As seen in Figure 13, when ***a*** is close to 0, the right plateau begins earlier and reaches a higher value, because selection is more plausible. Higher ***a*** values (close to 1) produce a restraint in the selection for endemic-targeting spacers, meaning the plateau is not present (Figure 13A, the position of the right maximum appears in yellow, meaning the

maximum is in the final spacer) and the value is the minimum obtained in all the simulations, which is zero (Figure 13B).

We observe parameter ***g*** is not clearly associated with this right maximum and plateau except for the position. In this case, the highest values of ***g*** (20 in this set of simulations) also prevent the right plateau from being formed. This is due to the shift in the incidences crossing point. For higher ***g*** values, the moment when endemic-targeting spacers start to dominate will come later in time, so the maximum values of fitness contribution are also shifted.

Regarding ***pEndemic*** parameter, when 0, as we have already mentioned, we find the minimum possible value of right maximum. However, by chance some new spacers get to be close to endemic, making the maximum lower than when endemics are present, but not zero. This is also why we see a correlation with the position of the maximum. The right plateau (represented when max_right_position measure appears in purple) is longer for greater ***pEndemic*** values (because the endemic-targeting spacers are the ones who make the fitness recover). Nevertheless, the association is not as clear as with parameter ***a***. The curious pattern we can see in Figure 13B, is made from slight changes that ***pEndemic*** values produce in right maximum values.

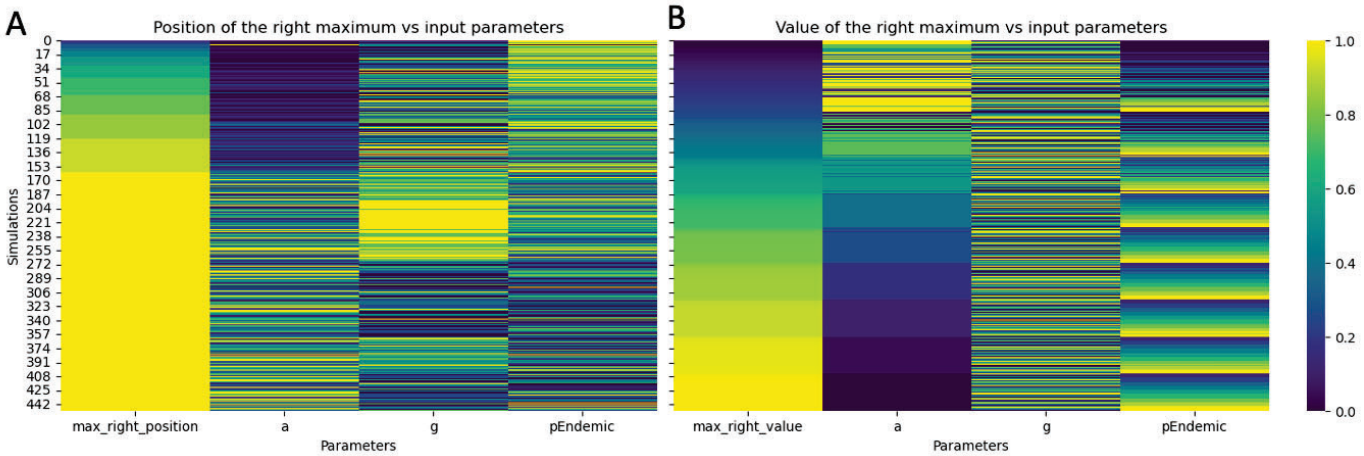

Figure 14. Heatmaps representing on the left column a measure of fitness contribution profile: the position of the right maximum (Figure A) and the value of that maximum (Figure B), next to input parameter values. Each column data was scaled from 0 to 1.

Euclidean distance (section 2.3.3) is a measure that depends very much on the length of the two plateaus. Longer plateaus lead to smaller Euclidean distances. As shown in the figure below, the existence of both plateaus (higher ***g*** and ***pEndemic*** values and lower ***a*** values) is associated with the minimum Euclidean distance found in all 450 simulations. The inverse situation is also confirmed here.

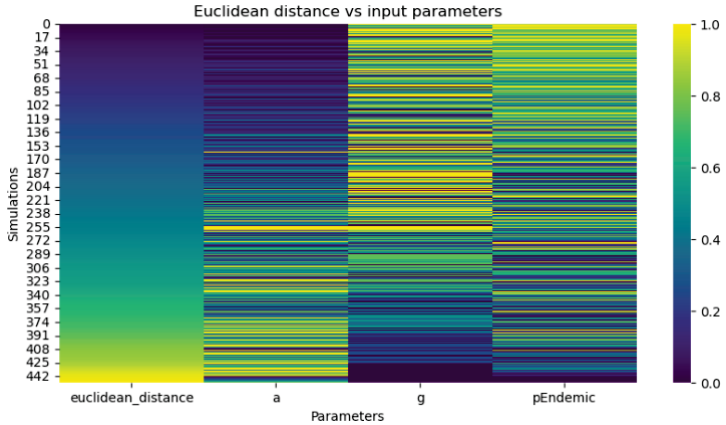

Figure 15. Heatmap representing on the left column a measure of Euclidean distance between midpoints in the fitness contribution profile, next to input parameter values. Each column data was scaled from 0 to 1.

Overall, we have found that the left plateau is exclusively dependent on parameter ***g***, while the right plateau, although having an important dependence on parameter ***a***, also depends on ***pEndemic*** and ***g***.

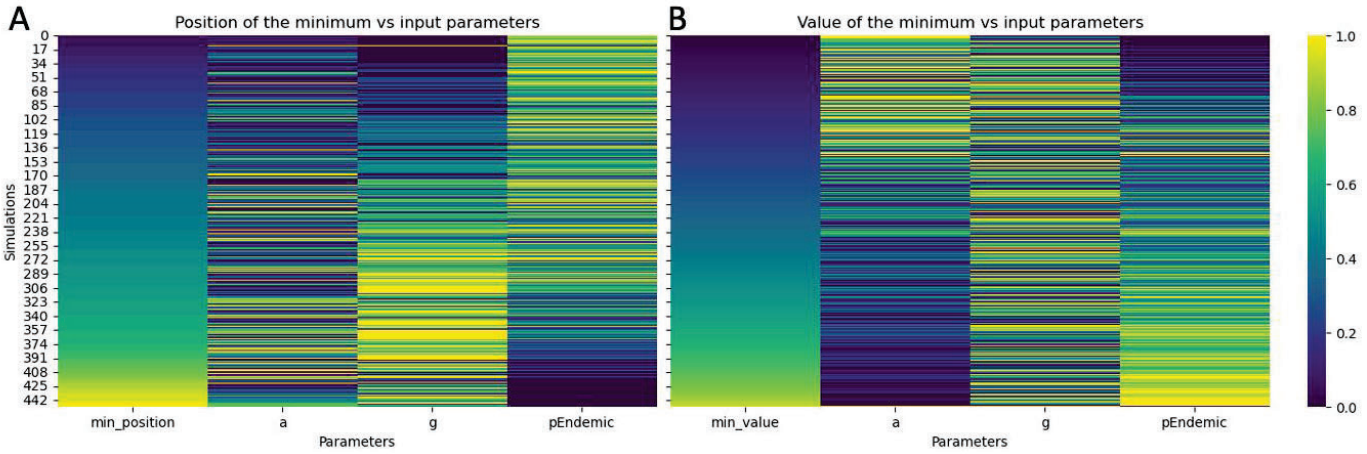

Figure 16. Heatmaps representing on the left column a measure of fitness contribution profile: the position of the minimum (Figure A) and the value of that minimum (Figure B), next to input parameter values. Each column data was scaled from 0 to 1.

Heatmaps comparing input parameters with the position and value of the fitness curve minimum (Figure 15), confirm earlier observations. Parameter ***g*** shifts the minimum horizontally (as with the left maximum), while parameter ***a*** affects its heigh. Parameter ***pEndemic*** affects both, especially in its extreme values. For ***pEndemic=0*** (i.e., simulations with no pure endemic-targeting spacers) the minimum shifts rightward and reaches the lowest observed value—zero. As seen in Figure 4B, the absence of endemic-targeting spacers produces a prolonged and deeper decline in fitness contribution, until the residual presence of a few endemic-targeting spacers eventually initiates a small recovery. In contrast, early presence of these spacers leads to a limited decrease followed by a significant recovery (Figure 3B).

The ratio of the peak's distribution (section 2.3.3) shows high outliers (yellow stripe in Figure 16) when the right maximum value approaches zero, a case found for ***pEndemic=0*** and high values of parameter ***a*** (0.77—Figures 4C and 4D. Here it is confirmed that parameter ***a*** and ***pEndemic*** have a major influence on the right maximum value (present in the right plateau). The highest recoveries of fitness (that is, the smallest ratio of peaks values) are associated with more endemic-targeting spacers selection (low values of parameter ***a***), in addition to more endemic-targeting spacers abundance (high ***pEndemic*** values) and endemic-targeting spacers dominance from the beginning (***g=0***).
With high ***pEndemic*** values we mean close to 1, but not 1. This is due to the existing competition between endemic-targeting spacers when there are too many of them in the simulation. The maximum recovery of fitness contribution at the end of the CRISPR-array was found with ***pEndemic*** values of 0.8, because this competition restraints the fitness contribution recovery.

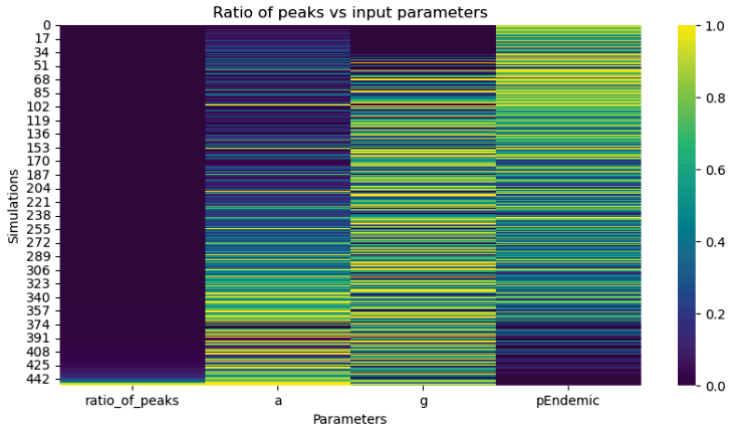

Figure 17. Heatmap representing on the left column a measure of fitness contribution profile: the ratio of left peak divided by right peak, next to input parameter values. Each column data was scaled from 0 to 1.

A recapitulation of the relationships between the morphological measures of fitness contribution and input parameters of the simulations can be found in the following heatmap.

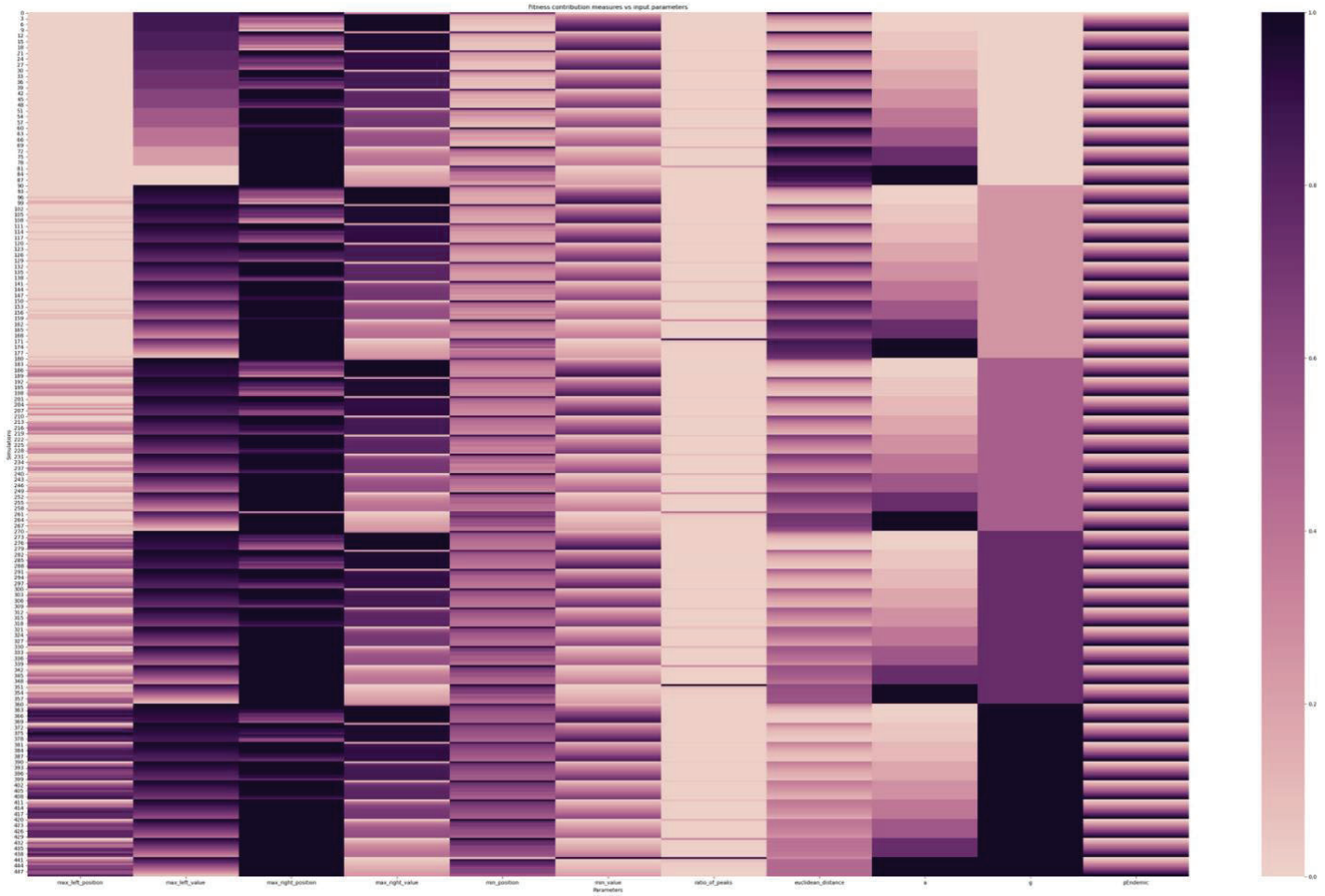

Figure 18. Overall heatmap of fitness contribution measures vs input parameters values. Each column was scaled from 0 to 1.

3.2.2. Beta values of spacers

The curve properties extracted from beta results were the initial and minimum value, the position of the minimum and where the 99.8% of the maximum is reached and the final value (section 2.3.3).

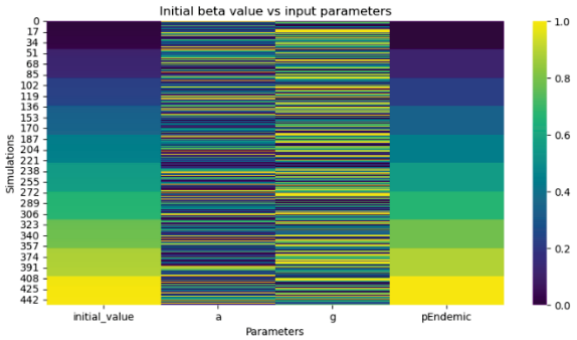

Figure 19. Heatmap representing on the left column the initial value of beta parameter in the CRISPR array, next to input parameter values. Each column data was scaled from 0 to 1.

The initial beta values were the expected considering the mixture distribution (equation 6, section 2.4.1), so ***pEndemic***-dependant, and not affected by ***a*** and ***g*** parameters (Figure 18).

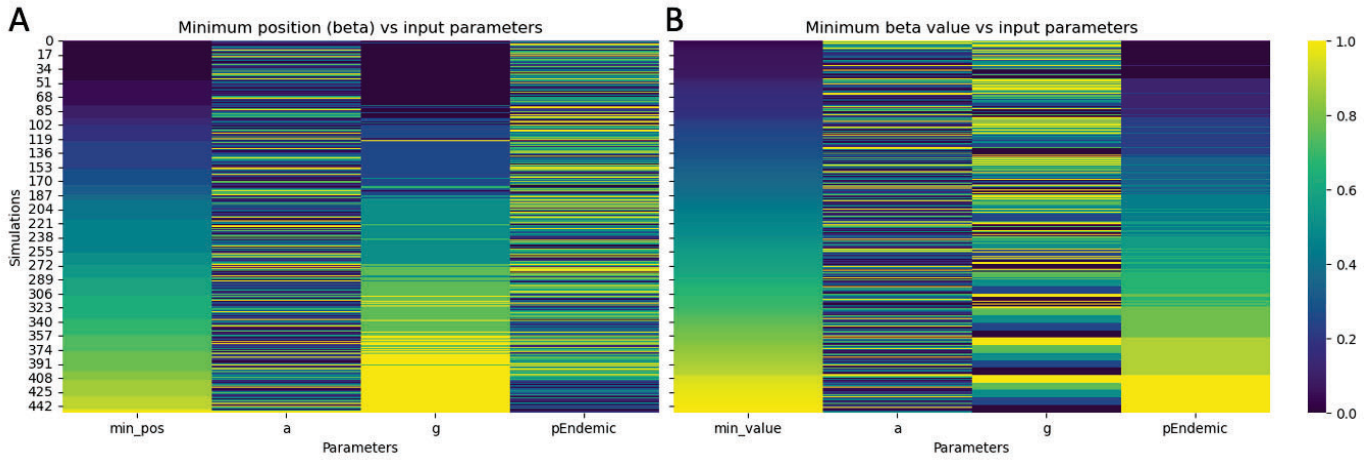

Figure 20. Heatmap representing on the left column the position of the minimum in the beta profile (A) and the value of that minimum (B), next to input parameter values. Each column data was scaled from 0 to 1.

Regarding the position and the value of the minimum, it is confirmed what was explained in section 3.1: the position is dependent on the shift of dominances caused by parameter ***g*** and the value is associated with parameter ***pEndemic***.

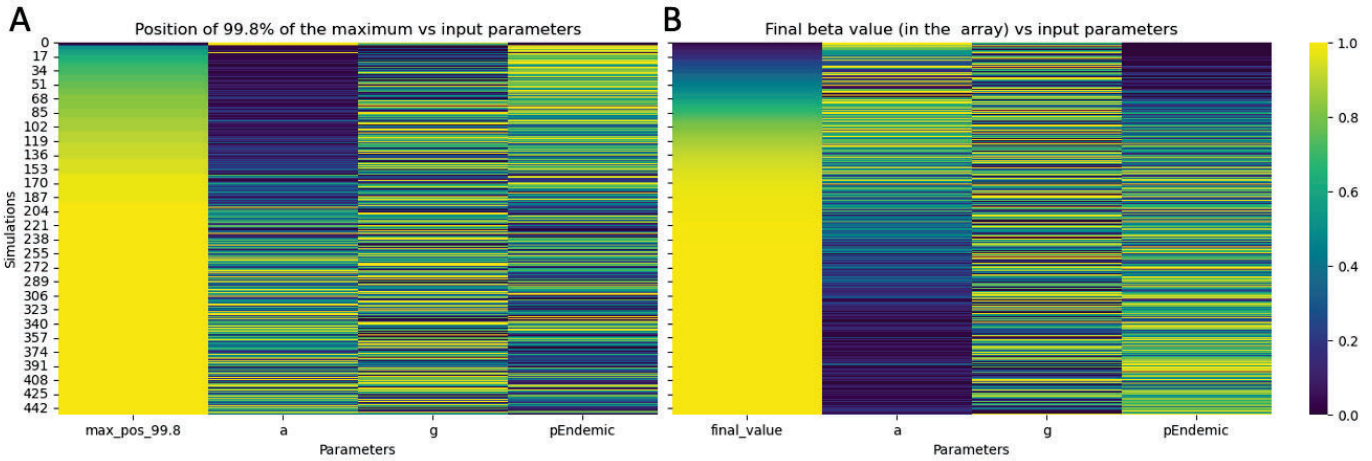

Figure 21. Heatmap representing on the left column the position where the 99.8% of the maximum beta value is reached (A) and the value of the last spacer (B), next to input parameter values. Each column data was scaled from 0 to 1.

The position where the 99.8% of the maximum is reached (a way to measure the length of the right plateau) depends strongly on parameter ***a.***
In Figure 20A, we can see the left column — representing the position where the 99.8% of the maximum beta value is reached — is zero (purple) for the highest ***a*** and ***g*** value and lowest ***pEndemic***. In these cases, the beta value did not recover, so the maximum value is found in the first spacer.
The final beta value depends on both ***a*** and ***pEndemic***. Beta values of 1 (Figure 20B, in yellow) are found for low ***a*** (so endemic selection is allowed) and high ***pEndemic*** (endemics need to be present to be selected).

3.3.3. Age values of spacers

The only distinction between simulations regarding age is the maximum they reach.

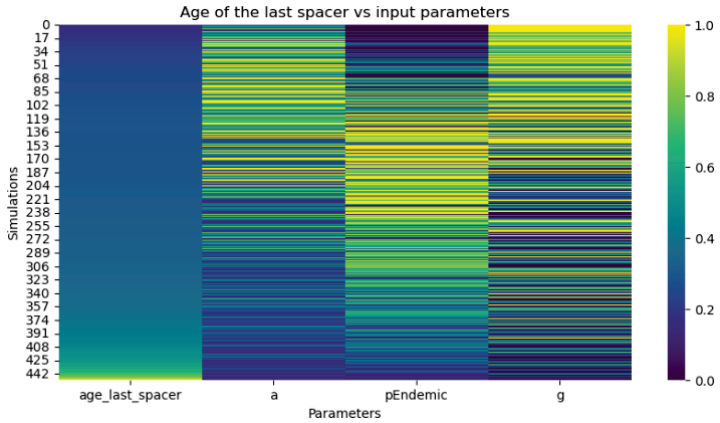

Figure 22. Heatmap representing on the left column the age of the last spacer, next to input parameter values. Each column data was scaled from 0 to 1.

As expected, the maximum age the system can reach depends on the three parameters. The oldest spacers are found in a endemics-selective environment (low ***a***), with endemic's presence but without competition (low ***pEndemic***, not zero) and with an endemic-dominance situation from the beginning (***g=0***).

3.3. Impact of generation number on observed dynamics

The resulting curves for the principal output parameters after running simulations with different total time were the following:

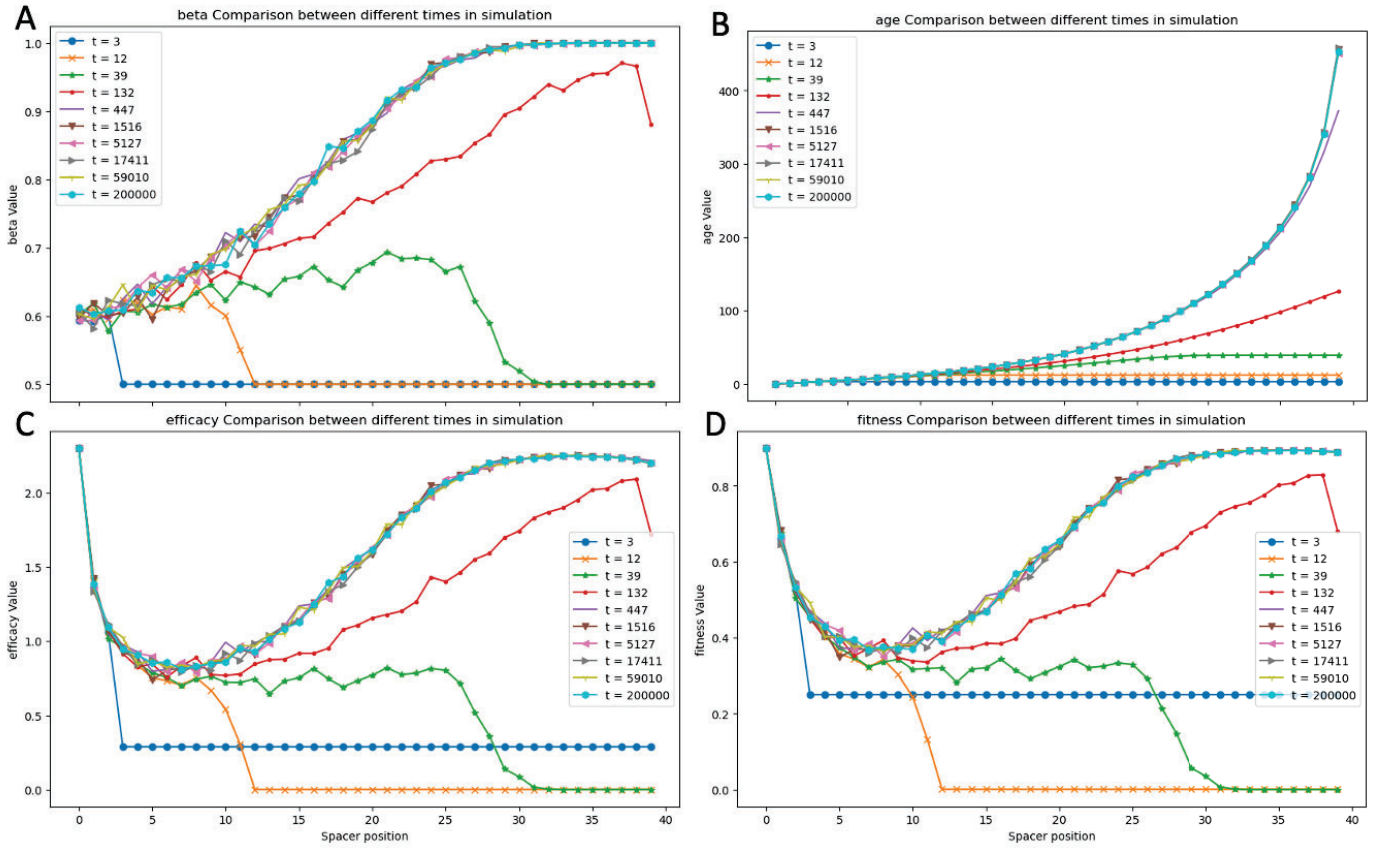

Figure 23. Analysis of output parameters across different simulation times to determine the minimum time required to reach a stationary state.

From t = 447 on, the beta, age, efficacy and fitness values are almost the same as the ones obtained with t = 200000. The betas converge to 0.9999 at the same time; the same occurs to efficacy and fitness values.
The variability we observe is due to the number of replicates. The way to make results more reproducible would be to increase the number of replicates, regardless of the time of simulation.
Regarding the final spacer's age for t = 447, it does not reach the final value of the rest because simulation time is

lower than the maximum age of equilibrium.

The system converges to a stationary state in around 500 generations. Once it has reached it, the distributions do not change.

3.4. Influence of CRISPR-array size on system behavior

To assess the influence of CRISPR array length on model outcomes, simulations were conducted as mentioned, using both short arrays (***Nsp=10***) and long arrays (***Nsp=40***). The results showed that the overall behaviors and trends remained qualitatively consistent across both array lengths.

The simulation outputs for both configurations are available in the project repository. Additionally, comparative visualizations are provided in Annex II.

CHAPTER 4: DISCUSSION

The results previously presented highlight the effects of the model's core parameters in shaping CRISPR array dynamics. As we have concluded in chapter 3, the parameter ***a*** controls selection strength, that is, the retention of endemic-targeting spacers. The parameter ***g*** defines the temporal point of crossover between epidemic and endemic dominance, directly influencing the spatial structure of fitness and beta profiles along the array. Meanwhile, ***pEndemic***, which controls the relative frequency of endemic versus epidemic viruses, shows a strong impact at its extremes: simulations with ***pEndemic=0*** or ***pEndemic=1*** produce markedly different behaviors. However, across intermediate values (roughly between 0.2 and 0.8), the morphological patterns remain qualitatively stable.

According to the model, the evolution of the CRISPR array may lead to prioritizing short-term or long-term protection. In our results, we have seen that long-term immunity may be selected when parameter ***g*** is lower than the size of the array and, therefore, spacers can reach the point where endemic viruses have higher incidences than epidemic viruses. The other condition to select long-term immunity is to have a low value of parameter ***a***.
For ***g<Nsp*** and ***a*** close to 1, only the last spacer is endemic targeting. This applies for any ***pEndemic*** value except zero, because endemic-targeting spacers are needed to select them.
On the other hand, ***g>Nsp*** is associated with short-term memory. Only when ***pEndemic*** is very high (0.9) and ***a*** very low (0.10) the last spacer is endemic targeting, but that is still a short-term specialized array.
We can biologically interpret parameter ***a*** as an inverse of the rate of exposure to viruses. For epidemic viruses, which disappear quickly, a high exposure rate corresponds to frequent encounters with new viral types; while for endemic, a high rate may reflect encounters with viral types already present in the environment.
In that case, smaller flows lead to long-term memory. An example in nature is the human gut microbiome. In this environment, CRISPR immunity updates are rare because the virome is stable and cells are not often in contact with new viruses [30, 31].

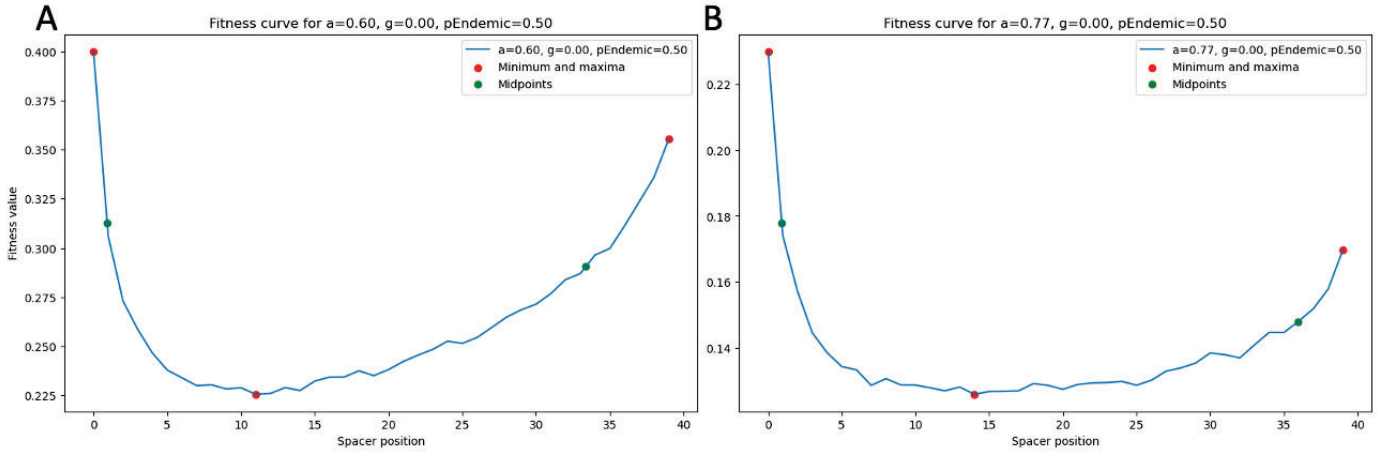

Figure 24. Fitness profile for pEndemic=0.50, g=0.00 and a=0.60 (A) or 0.77 (B). These were the closest results to López-Beltrán et al. work.

López-Beltrán, Botelho, and Iranzo [12] investigated the local adaptation of CRISPR spacers in the human gut. Their empirical curves resemble our theoretical dynamics under parameters that allow selection for endemic-targeting spacers (Figure 23) — when ***g=0*** (dominance of endemic viruses) and ***pEndemic=0.5***, capturing the persistent presence of gut endemic phages such as crAss-like viruses [32]. The closest match in our model was obtained with ***a=0.60*** and ***a=0.77***, which does not correspond with the low frequency of new viruses that we have previously mentioned.
The qualitative effect of ***a*** in the model is a selective pressure effect that can be interpreted as the inverse of the rate of exposure to viruses. However, this effect could be also qualitatively interpreted as how relevant CRISPR immunity is in a host's survival. If there are too many factors outside CRISPR that determine host's survival, being CRISPR defense just a small additional contribution, then the selective pressure in CRISPR will be small (no selection for endemic-targeting spacers, high values of ***a***). Following this interpretation, ***a=1*** would represent the situations where a new virus appears in the population and there is no change in host population's composition (balance between immune versus sensitive hosts), meaning that CRISPR is irrelevant.

The medium-to-high value of parameter ***a*** in the curves that best match the results of López-Beltrán et al. suggests that

selection for endemic-targeting spacers exists but is weak. This implies that, in the gut microbiome, the survival of bacterial lineages depends largely on factors beyond CRISPR-Cas-mediated immunity.
This interpretation is coherent with the findings of López-Beltrán et al., who observed evidence of selective sweeps in the gut microbiome, but these were weak. When viruses targeted by existing spacers appeared, a slight decrease in host population diversity was detected. However, not all hosts without the corresponding spacer disappeared. In other words, CRISPR provokes a change in survival, but not a radical one.

The opposite situation is found in nature in the deep-sea hydrothermal vent ecosystem. Here, the high viral abundances force robust defense mechanisms, which are constantly evolving.
Studies in *Vibrio* species showed that CRISPR-Cas are predominantly present on its mobile genetic elements [33]. This could be seen as a strategy to rapidly change their CRISPR immunity depending on the unstable environment.
If experimental data from this environment were available, we would expect our model to align best with parameters combination involving a significant value of ***pEndemic*** and a low value of ***a***.
This expectation is based on observations that viruses were largely endemic to individual vent sites [34], indicating restricted dispersal, and in some cases, viral assemblages persisted over time. Since viruses are a major source of microbial mortality in marine systems [34] — being especially abundant at deep-sea hydrothermal vents — host survival likely depends on effective CRISPR-mediated defenses. The model's selection for endemic-targeting spacers (reflected by low ***a***) would be essential for bacterial persistence.

Most of the models related to this field study the effect of CRISPR-Cas on the bacteriophage dynamics at the populational level [11,35,36,37] or spacer content evolution [38]. However, they track the whole CRISPR array without modeling positional structure inside, and without distinguishing between endemic- and epidemic-targeting spacers.
In contrast, our model offers understanding of the interplay between array position, viral types (including their proportion and their magnitudes) and selective pressure.
It also generates testable predictions, that could be validated with experimental results. For instance, in an environment rich in endemic viruses and were CRISPR-mediated immunity is essential for survival, we expect to see CRISPR arrays with old and conserved spacers. On the other hand, CRISPR arrays focused on short-term memory are predicted in settings with few endemic viruses or where other defense systems are more important to host's survival.

The assumptions we mentioned in section 2.1.1 illustrate the simplicity of the model. However, the fact that it does not represent accurately population dynamics in nature does not diminish its value.
The effects observed — such as the influences of dominances and proportion between endemic and epidemic viruses and selective pressure on array composition — align with biological intuition. It is reasonable to expect that these trends would also come up in more complex models and experiments. Its value relies as well in identifying the key variables and interactions that deserve closer attention, working as an exploratory model to interpret CRISPR dynamics and guide future experimental validations.

The possibilities of development are limitless, from reframing the assumptions — bringing the model closer to reality — to combining empirical time-series CRISPR data with the model results.

CHAPTER 5: CONCLUSIONS

In this thesis, we developed a mathematical and computational model to study the dynamics of CRISPR-Cas spacers in microbial communities, with particular focus on the distinction between endemic- and epidemic-targeting spacers. The model allowed us to explore how different viral environments and selection pressures shape the dominance and persistence of both kinds of spacers in the CRISPR arrays.

In the model, key parameters—namely ***a***, ***g***, and ***pEndemic***—play critical roles in determining the array's immunological profile. Therefore, they help us simulate a broad spectrum of viral compositions, from fully endemic to fully epidemic. The model provides a conceptual framework to understand the trade-off between short- and long-term CRISPR immunity in these different environments.

Moreover, this work differentiates from previous models by incorporating positional structure within the CRISPR array and explicitly differentiating between viral types. This approach helped replicate the qualitative dynamics of the spacer's contribution to the host's defense of bacteriophages obtained in López-Beltrán et al. It also generates testable predictions regarding memory specialization and immune strategy adaptation across viral contexts.
Spacer turnover is exclusively assessed at the distal positions of the CRISPR array—bearing in mind that in each time step of the model, one spacer is inserted, and one is removed—, as proximal positions reflect recent insertions and selection effects require time to manifest. High final spacer ages indicate greater array stability, whereas low final ages suggest a more dynamic or unstable CRISPR array.

While the model is intentionally simple and does not represent the full complexity of real-world CRISPR array. dynamics, it highlights essential mechanisms and variables that influence CRISPR immunity. As such, it serves as an exploratory tool to guide future experimental research and model refinement.

Future directions include incorporating more biologically realistic assumptions, integrating empirical time-series data, and extending the framework to consider population-level and co-evolutionary processes. These developments would be crucial to deepen our understanding of CRISPR array evolution and its role in microbial ecosystem resilience.

BIBLIOGRAPHY

1. Lander, E. S. (2016). The heroes of CRISPR. *Cell*, *164*(1–2), 18–28. https://doi.org/10.1016/j.cell.2015.12.041
2. Pavlova, Y. S., Paez-Espino, D., Morozov, A. Yu., & Belalov, I. S. (2021). Searching for fat tails in CRISPR-Cas systems: Data analysis and mathematical modeling. *PLOS Computational Biology*, *17*(3), e1008841. https://doi.org/10.1371/journal.pcbi.1008841
3. Musunuru, K., Grandinette, S. A., Wang, X., Hudson, T. R., Briseno, K., Berry, A. M., Hacker, J. L., Hsu, A., Silverstein, R. A., Hille, L. T., Ogul, A. N., Robinson-Garvin, N. A., Small, J. C., McCague, S., Burke, S. M., Wright, C. M., Bick, S., Indurthi, V., Sharma, S., … Ahrens-Nicklas, R. C. (n.d.). Patient-specific in vivo gene editing to treat a rare genetic disease. *New England Journal of Medicine*, 0(0). https://doi.org/10.1056/NEJMoa2504747
4. Xu, Y., & Li, Z. (2020). CRISPR-Cas systems: Overview, innovations and applications in human disease research and gene therapy. *Computational and Structural Biotechnology Journal*, *18*, 2401–2415. https://doi.org/10.1016/j.csbj.2020.08.031
5. Garrett, S. C. (2021). Pruning and tending immune memories: Spacer dynamics in the CRISPR array. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.664299
6. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, *315*(5819), 1709–1712. https://doi.org/10.1126/science.1138140
7. Jansen, R., van Embden, J. D. A., Gaastra, W., & Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology*, *43*(6), 1565–1575. https://doi.org/10.1046/j.1365-2958.2002.02839.x
8. Weinberger, A. D., Sun, C. L., Pluciński, M. M., Denef, V. J., Thomas, B. C., Horvath, P., Barrangou, R., Gilmore, M. S., Getz, W. M., & Banfield, J. F. (2012). Persisting viral sequences shape microbial CRISPR-based immunity. *PLOS Computational Biology*, *8*(4), e1002475. https://doi.org/10.1371/journal.pcbi.1002475
9. Modell, J. W., Jiang, W., & Marraffini, L. A. (2017). CRISPR–Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. *Nature*, *544*(7648), 101–104. https://doi.org/10.1038/nature21719
10. Jiang, W., Maniv, I., Arain, F., Wang, Y., Levin, B. R., & Marraffini, L. A. (2013). Dealing with the evolutionary downside of CRISPR immunity: Bacteria and beneficial plasmids. *PLOS Genetics*, *9*(9), e1003844. https://doi.org/10.1371/journal.pgen.1003844
11. Childs, L. M., Held, N. L., Young, M. J., Whitaker, R. J., & Weitz, J. S. (2012). Multiscale model of CRISPR-induced coevolutionary dynamics: Diversification at the interface of Lamarck and Darwin. *Evolution*, *66*(7), 2015–2029. https://doi.org/10.1111/j.1558-5646.2012.01595.x
12. López-Beltrán, A., Botelho, J., & Iranzo, J. (2024). Dynamics of CRISPR-mediated virus–host interactions in the human gut microbiome. *The ISME Journal*, *18*(1), wrae134. https://doi.org/10.1093/ismejo/wrae134
13. Paez-Espino, D., Sharon, I., Morovic, W., Stahl, B., Thomas, B. C., Barrangou, R., & Banfield, J. F. (2015). CRISPR immunity drives rapid phage genome evolution in *Streptococcus thermophilus*. *mBio*, *6*(2), e00262-15. https://doi.org/10.1128/mBio.00262-15
14. Stern, A., Mick, E., Tirosh, I., Sagy, O., & Sorek, R. (2012). CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Research*, *22*(10), 1985–1994. https://doi.org/10.1101/gr.138297.112
15. The size of the immune repertoire of bacteria. (n.d.). *PNAS*. Retrieved June 4, 2025, from https://www.pnas.org/doi/full/10.1073/pnas.1903666117
16. Selective maintenance of multiple CRISPR arrays across prokaryotes. (n.d.). *The CRISPR Journal*. Retrieved June 4, 2025, from https://www.liebertpub.com/doi/10.1089/crispr.2018.0034
17. A scaling law in CRISPR repertoire sizes arises from the avoidance of autoimmunity. (n.d.). *Current Biology*. Retrieved June 4, 2025, from https://www.sciencedirect.com/science/article/pii/S0960982222007801
18. Toms, A., & Barrangou, R. (2017). On the global CRISPR array behavior in class I systems. *Biology Direct*, *12*. https://doi.org/10.1186/s13062-017-0193-2
19. Crawley, A. B., Henriksen, E. D., Stout, E., Brandt, K., & Barrangou, R. (2018). Characterizing the activity of abundant, diverse and active CRISPR-Cas systems in lactobacilli. *Scientific Reports*, *8*(1), 11544. https://doi.org/10.1038/s41598-018-29746-3

20. Roy, K., Ghosh, D., DeBruyn, J. M., Dasgupta, T., Wommack, K. E., Liang, X., Wagner, R. E., & Radosevich, M. (2020). Temporal dynamics of soil virus and bacterial populations in agricultural and early plant successional soils. *Frontiers in Microbiology*, *11*, 1494. https://doi.org/10.3389/fmicb.2020.01494
21. Shabbir, M. A. B., Hao, H., Shabbir, M. Z., Hussain, H. I., Iqbal, Z., Ahmed, S., Sattar, A., Iqbal, M., Li, J., & Yuan, Z. (2016). Survival and evolution of CRISPR–Cas system in prokaryotes and its applications. *Frontiers in Immunology*, *7*. https://doi.org/10.3389/fimmu.2016.00375
22. ISO/IEC. (2014). ISO international standard ISO/IEC 14882:2014(E) – Programming language C++ [Working draft]. Geneva, Switzerland: International Organization for Standardization. https://isocpp.org/std/the-standard
23. Python Software Foundation. (2023). Python language reference (Version 3.11) [Computer software]. https://www.python.org
24. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2
25. The pandas development team. (2020). pandas-dev/pandas: Pandas (Version 1.0.5) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3509134
26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html
27. Waskom, M. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021
28. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2
29. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55
30. Zhang, A.-N., Gaston, J. M., Cárdenas, P., Zhao, S., Gu, X., & Alm, E. J. (2025). CRISPR-Cas spacer acquisition is a rare event in human gut microbiome. *Cell Genomics*, *5*(1), 100725. https://doi.org/10.1016/j.xgen.2024.100725
31. Shkoporov, A. N., Clooney, A. G., Sutton, T. D. S., Ryan, F. J., Daly, K. M., Nolan, J. A., McDonnell, S. A., Khokhlova, E. V., Draper, L. A., Forde, A., Guerin, E., Velayudhan, V., Ross, R. P., & Hill, C. (2019). The human gut virome is highly diverse, stable, and individual specific. *Cell Host & Microbe*, *26*(4), 527–541.e5. https://doi.org/10.1016/j.chom.2019.09.009
32. Shkoporov, A. N., Khokhlova, E. V., Stephens, N., Hueston, C., Seymour, S., Hryckowian, A. J., Scholz, D., Ross, R. P., & Hill, C. (2021). Long-term persistence of crAss-like phage crAss001 is associated with phase variation in *Bacteroides intestinalis*. *BMC Biology*, *19*(1), 163. https://doi.org/10.1186/s12915-021-01084-3
33. McDonald, N. D., Regmi, A., Morreale, D. P., Borowski, J. D., & Boyd, E. F. (2019). CRISPR-Cas systems are present predominantly on mobile genetic elements in *Vibrio* species. *BMC Genomics*, *20*(1), 105. https://doi.org/10.1186/s12864-019-5439-1
34. Thomas, E., Anderson, R. E., Li, V., Rogan, L. J., & Huber, J. A. (2021). Diverse viruses in deep-sea hydrothermal vent fluids have restricted dispersal across ocean basins. *mSystems*, *6*(3), e00068-21. https://doi.org/10.1128/mSystems.00068-21
35. Shu, M., Fu, R., & Wang, W. (2017). A bacteriophage model based on CRISPR/Cas immune system in a chemostat. *Mathematical Biosciences & Engineering*, *14*(5 & 6), 1361–1377. https://doi.org/10.3934/mbe.2017070
36. Fs, B., Yi, W., Ev, K., & Gp, K. (2014). Pseudo-chaotic oscillations in CRISPR-virus coevolution predicted by bifurcation analysis. *Biology Direct*, *9*, 13. https://doi.org/10.1186/1745-6150-9-13
37. Han, P.,Niestemski, L. R., Barrick, J. E., & Deem, M. W. (2013). Physical model of the immune response of bacteria against bacteriophage through the adaptive CRISPR-Cas immune system. *Physical Biology*, *10*(2), 025004. https://doi.org/10.1088/1478-3975/10/2/025004
38. Kupczok, A., & Bollback, J. P. (2013). Probabilistic models for CRISPR spacer content evolution. *BMC Evolutionary Biology*, *13*(1), 54. https://doi.org/10.1186/1471-2148-13-54
39. Xu, Y., & Li, Z. (2020). CRISPR-Cas systems: Overview, innovations and applications in human disease research and gene therapy. *Computational and Structural Biotechnology Journal*, *18*, 2401–2415. https://doi.org/10.1016/j.csbj.2020.08.031

ANNEX I

Table 1. Summary of CRISPR-Cas systems. Extracted from Xu y Li, «CRISPR-Cas systems» [39].

Class	Type	Subtype	Effector	Target	Nuclease domains	TracrRNA requirement	PAM/PFS
1 (multi-Cas proteins)	I	A, B, C, D, E, F, U	Cascade	dsDNA	HD fused to Cas3	No	–
1	III	A, B, C, D	Cascade	ssRNA	HD fused to Cas10	No	–
1	IV	A, B	Cascade	dsDNA	unknown	No	–
2 (single-Cas protein)	II	A	SpCas9	dsDNA	RuvC, HNH	Yes	NGG
2	II	A	SaCas9	dsDNA	RuvC, HNH	Yes	NNGRRT
2	II	B	FnCas9	dsDNA/ssRNA	RuvC, HNH	Yes	NGG
2	II	C	NmCas9	dsDNA	RuvC, HNH	Yes	NNNNGATT
2	V	A	Cas12a (Cpf1)	dsDNA	RuvC, Nuc	No	5′ AT-rich PAM
2	V	B	Cas12b (C2c1)	dsDNA	RuvC	Yes	5′ AT-rich PAM
2	V	C	Cas12c (C2c3)	dsDNA	RuvC	Yes	5′ AT-rich PAM
2	VI	A	Cas13a (C2c2)	ssRNA	2xHEPN	No	3′PFS: non-G
2	VI	B	Cas13b (C2c4)	ssRNA	2xHEPN	No	5′PFS: non-C; 3′PFS:NAN/NNA
2	VI	C	Cas13c (C2c7)	ssRNA	2xHEPN	No	–
2	VI	D	Cas13d	ssRNA	2xHEPN	No	–

ANNEX II

In this complementary annex, we include some figures comparing the results obtain for the long (40 spacer) and the short (10 spacers) arrays. Both arrays experiment a qualitatively similar evolution, so the array size does not influence the morphology of the system.

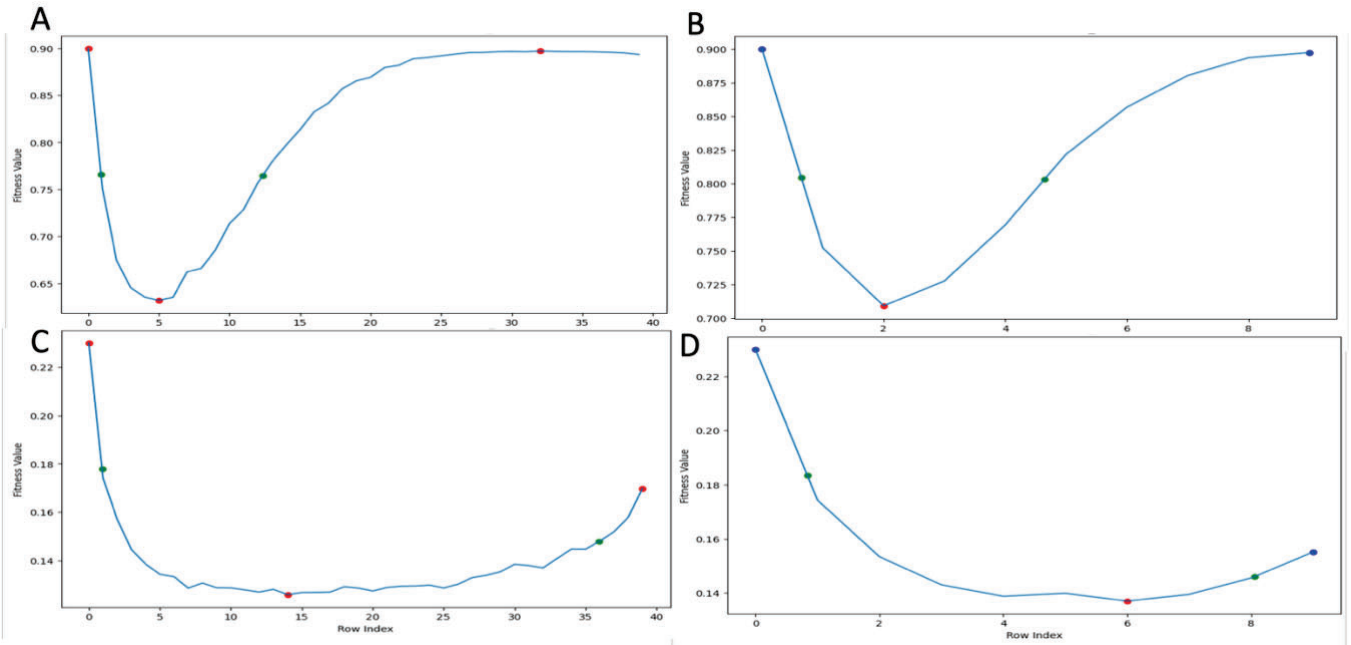

Figure 25. Fitness contribution profiles for Nsp=40 (Figures A and C) and Nsp=10 (Figures B and D). Parameters g=0 and pEndemic=0.50. Parameter a=0.10 (Figures A and B) and 0.77 (Figures C and D).

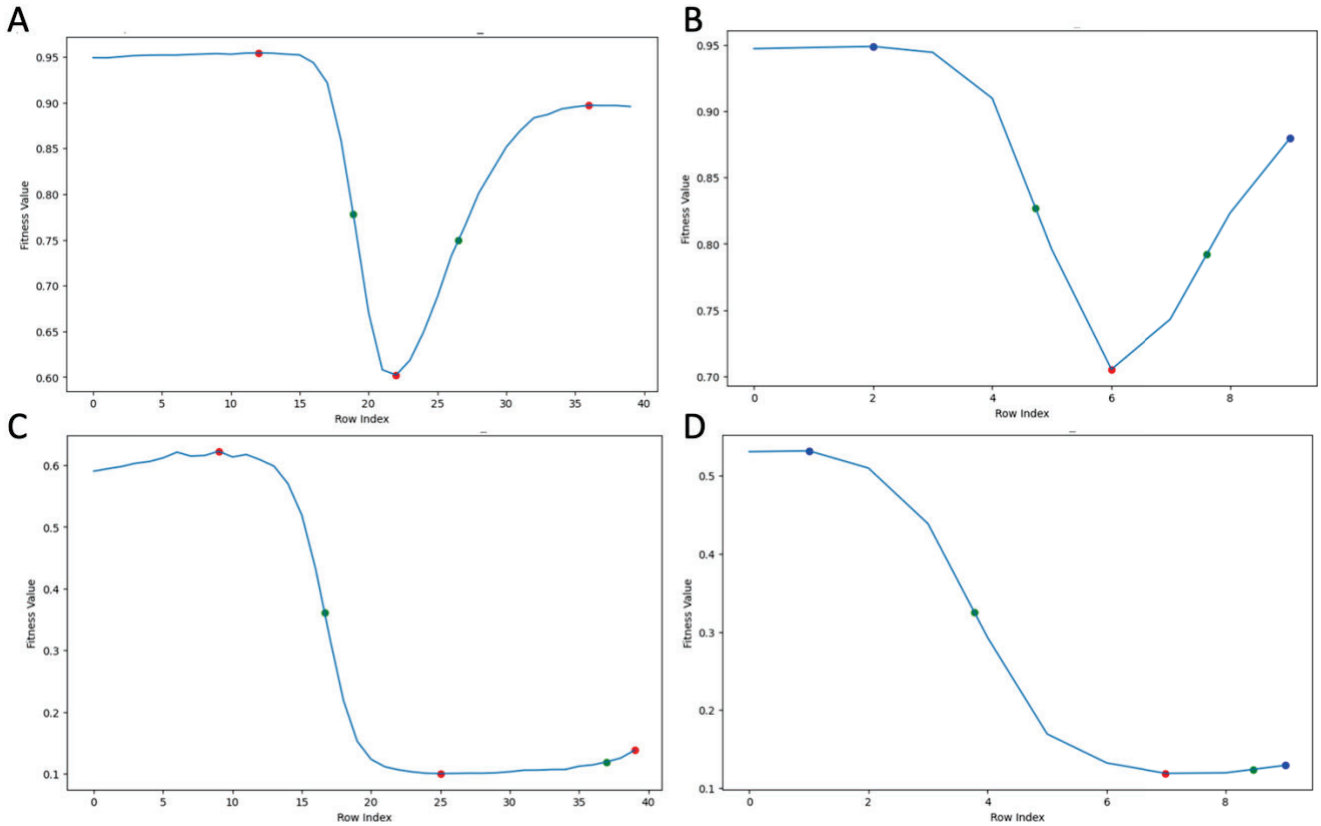

Figure 26. Fitness contribution profiles for Nsp=40 (Figures A and C) and Nsp=10 (Figures B and D). Parameter pEndemic=0.50. Parameter a=0.10 (Figures A and B) and 0.77 (Figures C and D). Parameter g=5 (Figures B and D) and 20 (Figures A and C).