Jarosław GIL [0000-0001-9228-9937]*, *Andrzej POLAŃSKI* [0000-0002-1793-9546]*

APPLICATION OF GILLESPIE ALGORITHM FOR SIMULATING EVOLUTION OF FITNESS OF MICROBIAL POPULATION

Abstract

In this study we present simulation system based on Gillespie algorithm for generating evolutionary events in the evolution scenario of microbial population. We present Gillespie simulation system adjusted to reproducing experimental data obtained in barcoding studies – experimental techniques in microbiology allowing tracing microbial populations with very high resolution. Gillespie simulation engine is constructed by defining its state vector and rules for its modifications. In order to efficiently simulate barcoded experiment by using Gillespie algorithm we provide modification – binning cells by lineages. Different bins define components of state in the Gillespie algorithm. The elaborated simulation model captures events in microbial population growth including death, division and mutations of cells. The obtained simulation results reflect population behavior, mutation wave and mutation distribution along generations. The elaborated methodology is confronted against literature data of experimental evolution of yeast tracking clones sub-generations. Simulation model was fitted to measurements in experimental data leading to good agreement.

1. INTRODUCTION

Experimental evolution techniques for populations of microbial organisms is a fast-developing area of scientific research, which provides important measurement data concerning scenarios and parameters of adaptive haploid evolution of microbes under different conditions. It has wide areas of applications both in cellular biology where it leads to advances in understanding adaptation mechanisms in microbial populations, and in applications of evolutionary genetics to other research fields such as epidemiology, virology or cancer research. Fast advances in molecular biology give impulses to extending volumes and resolutions of available measurement data in experimental microbial evolution. Implementing techniques, such as high throughput sequencing, labelling and barcoding in evolving microbial cultures allow obtaining detailed data on kinetics and dynamics of their evolution.

---

* Department of Computer Graphics, Vision and Digital Systems, Silesian University of Technology, Gliwice, Poland, jaroslaw.gil@polsl.pl

Recent publications in the field of experimental evolution of microbes (Bush et al., 2020; Kinnersley et al., 2021; Blundell et al., 2019; Nguyen Ba et al., 2019; Levy et al., 2015) provide datasets of significant size and accuracy enabling the study of genomic mechanisms and genetic forces behind their adaptation mechanisms to environmental conditions. Among the referenced experimental techniques an approach based on barcoding of microbial DNA (Blundell et al., 2019; Nguyen Ba et al., 2019) is of special interest due to possibility of tracing the large number of sub-populations/clones of microbial population over the whole experiment time range.

Important aspects of the study of experimental microbial evolution is characterization of roles and strengths of genetic forces, replication, mutation, genetic drift and selection (Neher, 2013) in scenarios of their adaptation to external conditions (Beckman & Loeb, 2005; Kvitek & Sherlock, 2013). In order to understand strengths, inter-plays between genetic forces and their dynamics, mathematical models of adaptive evolution of haploid populations are developed. Mathematics and computer modeling behind adaptive evolution lead to expressing the increase of adaptive potential of a microbial population as the result of propagation of the fitness wave, driven by emergence of favorable mutations in the DNA replication process and selection against less fitted species (Desai & Fisher, 2007). Mathematical modeling of dynamics of adaptive evolution is a wide area of research including approaches such as deterministic modeling using partial differential equations (Fisher-Kolmogorov equations) e.g., (Wang et al., 2019), or stochastic modeling of processes of cellular replication and death and occurrence of (driver) mutations, by using Markov birth - death processes (Baar et al., 2016; Castillo & Virgilio, 2015) branching processes (Wild, 2011; Yakovlev, Stoimenova & Yanev, 2008; Castillo & Virgilio, 2015) or multitype branching processes (Bozic et al., 2010; McFarland, Mirny & Korolev, 2014; Nguyen Ba et al., 2019; Levy et al., 2015).

When laws describing probabilities of evolutionary events become complicated the approach to modeling by using stochastic simulations can be very useful (Foo, Leder & Michor, 2011). Stochastic simulations are very flexible and can capture arbitrary scenarios of adaptation evolution. The basic approach to (forward) stochastic simulation of cellular evolution is by using Gillespie's algorithm (Gillespie, 2001; Marchetti, Priami & Thanh, 2017) where the state of the simulated process is defined by mutation profiles of cells and possible events are drawn randomly on the basis of probability distributions dependent on the state of the process.

In this study we present the implementation of the Gillespie's algorithm for evolutionary simulations tailored to the scenario of experimental yeast evolution with adaptation to external conditions, quantified by using the barcoding technique (Nguyen Ba et al., 2019; Levy et al., 2015). Novelty of our approach is adjusting (fitting) Gillespie simulation algorithm to the barcoded evolution experimental scenario by suitable changes of its construction. We describe the approach to fitting evolutionary parameters, intensities of cellular/microbial birth and deaths, intensities of mutations and mutations evolutionary advantage/fitness parameter to observational data. We show results of simulation algorithm versus published data on baker yeast (saccharomyces cerevisiae) growth (Levy et al., 2015).

2. MODEL DESCRIPTION

Elaborating simulation system for microbial evolution relies on defining stochastic models of events in the cellular lifetime: cell death, cell division or mutation, which occur during cell division. Methodology for obtaining timing and orderings of events relies on the use of the Gillespie algorithm, which provides a way to obtain suitable probability distributions of times of events on the basis of the state of the processes. Basic approach is to simulate events cell by cell, iterating through whole (Marchetti, Priami & Thanh, 2017). First improvement is considering all cell events with occurring time less than tau (tau leap algorithm (Nguyen Ba et al., 2019)) (Cao, Gillespie & Petzold, 2006). However, tau leap modification still does not significantly improve simulation time for larger populations. The binned version of Gillespie algorithm, elaborated in this study, makes possible to consider microbial population size above one million of cells/microbes/viruses/bacteria with reasonable simulation times.

2.1. Binned Gillespie algorithm

Tab. 1. Simulation parameters assumed as (Levy et al., 2015) experiment conditions. Whole population is divided into groups which represents barcoded cells

Model parameters	
mutation effect	2.5 – 15% by Poisson distribution
initial population	50'000 × 100 cells
population capacity	$5 \cdot 10^6$
mutation probability	20%
simulation cycles	1000

The basic Gillespie's algorithm modification by binning means grouping microbial cells by one independent characteristic which can describe a group of cells. For that purpose, number of mutations in the cell was chosen. Simulation begins with large population of cells with no mutation. In each cycle some of them can obtain new mutation which cause cells differentiation. Every cycle of simulation consists of three steps: generating random number of dying and dividing cells based on accurate probabilities and tau step, calculating number of mutating cells and population actualization. Probabilities are given by exponential distribution (1), (2).

$$P(death) = 1 - exp(-tau * sum(population)/capacity) \qquad (1)$$

$$P(divide) = 1 - exp\big(-tau/(1 - fitness)\big) \qquad (2)$$

Each mutation provides a change in fitness coefficient which has impact on the intensity of cellular division process. Calculation of fitness for single cell is easy – mutation provides percentage change in cell structure. To calculate it for one bin we assume that one mutation has impact on mean bin fitness (3). Mutating cells have the same fitness in the bin, new mutations have additional effect. Change of cells adjustment is obtained on the basis of probabilities in the Poisson distribution.

$$fitness(t + 1) = (fitness(t) * n_{cells} + mutations \\ * mutation\ effect)/n_{cells}$$

(3)

2.2. Simulating barcoding experimental data

In barcoding experiments DNA strands acquire markers allowing for their unique identification. At present, barcoding technique allows introducing hundreds of thousands of markers to populations of microorganisms to trace their evolution. Barcoding is particularly useful in studies of clonal evolution. To track separate clones lineages, it is assumed that population is grouped by clone reference barcodes introduced to the microbial population. In (Levy et al., 2015) experiment was performed based on tracking specific barcodes firstly inserted into microbial cell genomes.

Except bin sizes fitness parameter based on (3) and mutation wave in the form of mean mutation number were calculated. In figure 1 basic information about population evolution in time is presented. In the left side single lineage sizes (a) and sizes distributions (c) are presented. In the right-side mutation (d) and fitness wave (b) (the track of population adjustment) are presented. Model parameters are presented in table 1. For simulation parameters described in (Levy et al., 2015) were implemented.
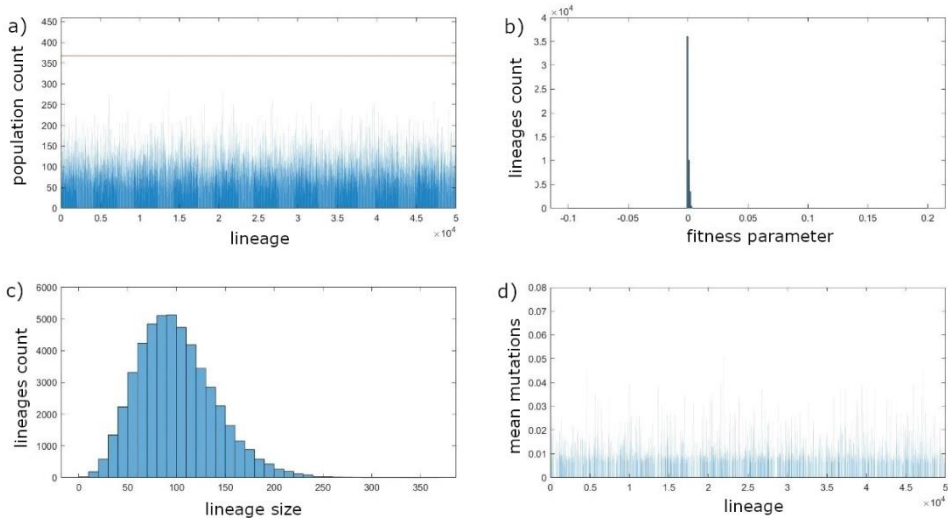

Fig. 1. Lineage binned algorithm result according to one generation, a) linages sizes with highlighted maximum value, b) whole population fitness wave, c) linage sizes distribution, d) mean numbers of mutations along population lineages.

Simulation time is greater for that modification than binned by mutation number because of greater number of separate bins. The complexity of algorithm is about O(n) because number of iterations depends on bin number. The basic algorithm complexity was about O(n³) cause of multiple iterations and comparisons through whole population cell by cell. For binning by mutation number simulation time is flexible because of changing number of bins. For binning by lineage characteristic number of bins are steady - one cycle time is approximately same.

Two separate simulations were made. First one with assumed parameters in table 1, second one with different initial population and population capacity. Lineages size were providing as experimental data from (Levy et al., 2015) as like simulation parameters were based on.

Pseudocode of our algorithm which shows its basic construction is listed below.

```
SET initial parameters
    {
    SET initial population size as 5'000'000
    SET capacity as 5'000'000
    SET mutation probability as 0.2
    SET tau step as 0.005
    SET step number as 1000
    }

CREATE initial population
    {
    CREATE population array as 50'000 groups of 100 cells
    CREATE mutation array as zeros for 50'000 cell groups
    CREATE fitness array as ones for 50'000 cell groups
    SET mutation effect for each group from 0.025 to 0.15
    }

WHILE step is less than step number
    {
    IF step is multiplier of 10
        {
        SHOW population plot
        SHOW mutation plot
        SHOW fitness plot
        }

    UPDATE population
        {
        CALCULATE death probability (1-exp (-tau*(cells number / capacity))) for whole
population

            FOR group in population
            {
                CALCULATE division probability (1-exp(-tau/(1-fitness))) for group
                CALCULATE number of dying and dividing cells
                SET mutating cells as 0.2 dividing cells

                SUBSTRACT number of dying cells from group
                ADD number of dividing cells to group

                CALCULATE new mean of mutation number for group ((mutation number * cells + new
mutations)/cells)
                CALCULATE new fitness for group ((fitness * cells + new mutations * mutation
effect)/cells)
            }
        }
    }
```

Listing 1. Binned Gillespie algorithm pseudocode

3. RESULTS AND DISCUSSION

Two simulation experiments were made:
1. Simulation basing on 50'000 lineages consisting of 100 clones and parameters described in table 1 – similar to (Levy et al., 2015),
2. Simulation basing on data provided by (Levy et al., 2015) – 500'000 lineages with different initial clones number.

3.1. 1st Experiment

To estimate model accuracy several simulations were pursued and the best suited parameters where chosen. In figure 2 simulated trajectories of lineages evolution are presented. Red, blue and white colors correspond to calculated fitness of single bins. From 50'000 introduced bins about 10% became extinct and few grew to much larger groups. For adaptive lineages the growth rate was almost about 1 magnitude in small time step – about 100 seconds. In theory cell fitness should provide more division events – as results show. Because of no initial differentiation between lineages, small amount of them adapt gaining sufficient value of fitness coefficient.
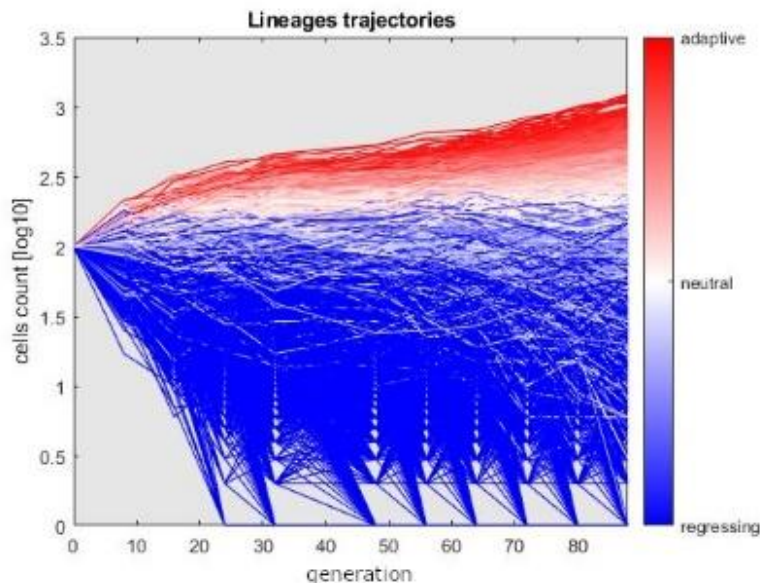

Fig. 2. Simulated lineages trajectories (for lineages greater than 100 cells): red color – lineages with good adaptation, blue color – bad fitness lineages; white color explains neutral fitness effect lineages

Growth of adaptive lineages causes increase in cell death probability. Cells with lower fitness mostly die, which results in lineages extinction. Some clones dominate population imposing smaller genetic variety inside the whole population. In figure 3 linage size distributions, with fitness factor of lineage marked by blue/orange colors, are presented. Adaptive lineages are growing as shown in the figure - lineages distribution changes its mean value from 10^2 to about $10^{2.5}$. The high value bars on figure near zero value represents extinct lineages.

Population differentiation proceeds as assumed. Few lineages gathered beneficial mutation (adaptive lineages) which results in higher fitness factor and rise in lineage size. Some of clones mutations had small effect on its fitness causing lineage extinction. Mostly mutation provided group adjustment establishing lineage size at steady level.

Further simulations should cause more lineage extinctions because of fast growth of adapted ones. Death probability, which depends, by logistic relation, on the whole population size, will be shaped mostly by the biggest groups. Extended simulation time could provide information about dominant clones. Fast lineage growth would cause extinction of small clones. Genetic variety should be very small and population should be composed of genetic clones.
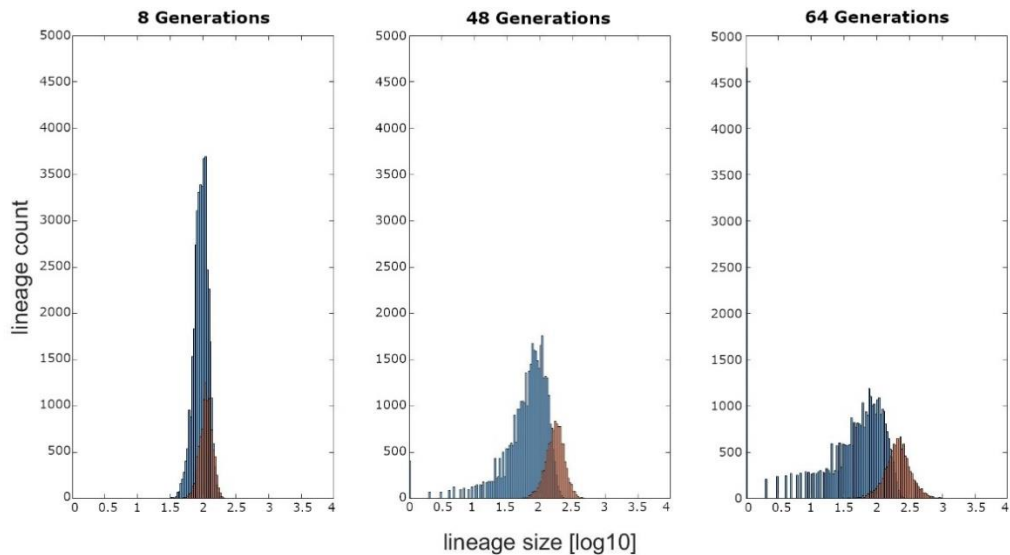

Fig. 3. Lineage size distribution in simulated data: blue part means neutral and regressing lineages, red part represents adaptive lineages

3.2. 2nd Experiment

Basing on data provided by (Levy et al., 2015) simulation was performed assuming that initial lineages are varied. Initial lineage size and its fitness was set according to experimental data. Mutation effect on group fitness was same as in 1st experiment – 2.5-15% described by Poisson distribution. Provided data is not reflection of simulation. Research group was gathering information about cells by selecting random number from population. Our goal was to obtain similar characteristic of population growth. In figure 4 simulated lineages trajectories are presented. Differences between figure 2 and figure 4 follow from different initial model state. When groups are assumed to be equal at simulation beginning fitness of whole population is distributed as mutation factor. At simulation begin both division and death probability are dependent on whole population – in further cycles dependency changes. When initial size of lineages is randomized death probability is mostly dependent on high size lineages. More cells also provide more division and mutation events which causes lineage fitness increase. Big groups are growing, gaining good adaptation and causing extinction of smaller or less adapted lineages.
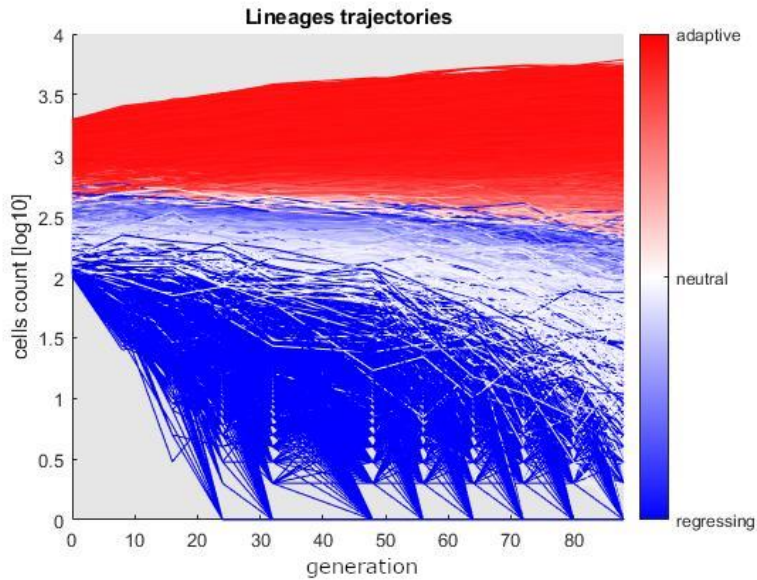
Fig. 4. Simulated lineages trajectories based on experimental data (for lineages greater than 100 cells): red color – lineages with good adaptation, blue color – bad fitness lineages; white color explains neutral fitness effect lineages
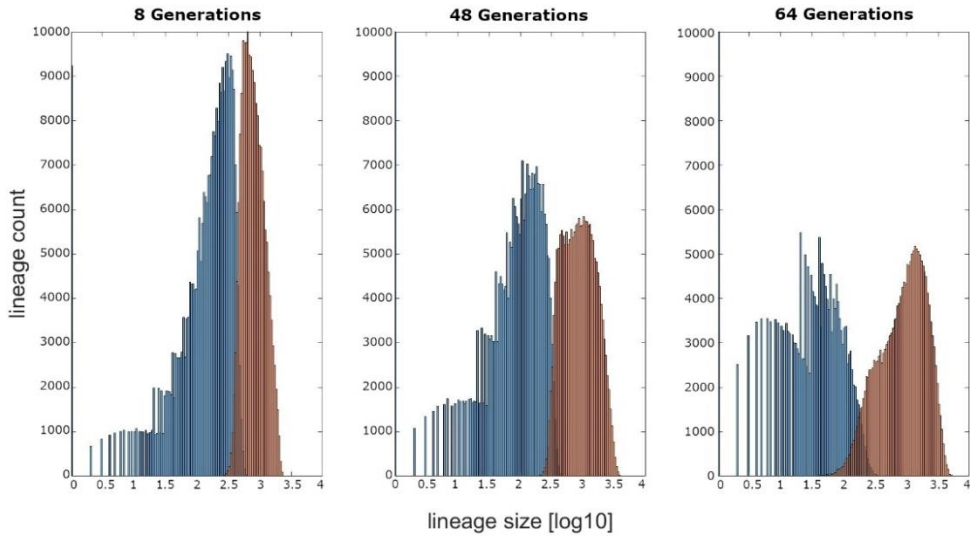

Fig. 5. Lineage size distribution in experimental data: blue part means neutral and regressing lineages, red part represents adaptive lineages

In figure 5 lineage size distribution and on figure 6 its fitness factor is presented to demonstrate that at initial stage of simulation the groups were different from each other. Real experiment shows that adaptive lineages should grow at exponential rate when gaining appropriate size and fitness factor. Further cycles of population evolution should show that. At the right-hand side of the picture the change of lineage sizes is shown. It creates some kind of evolution front so it is possible that population would rise much faster in next steps.
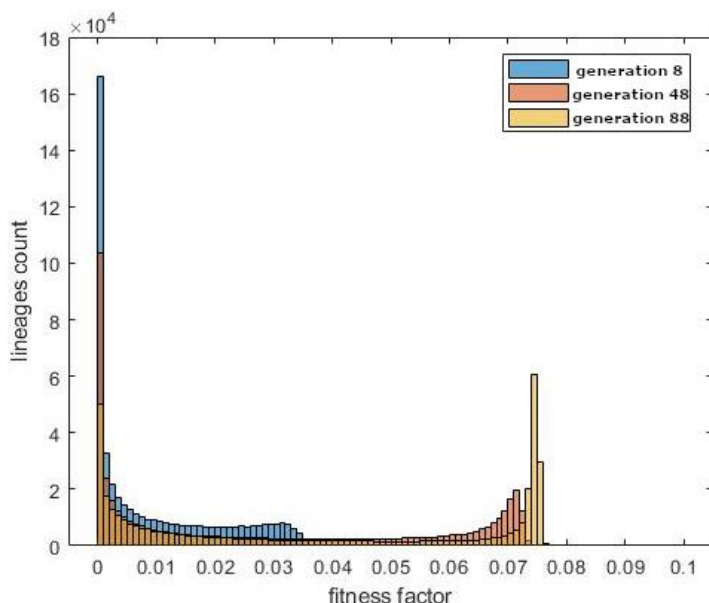
Fig. 6. Fitness change

Changing in lineage size and in its fitness is caused by multiple mutations. Some of them could provide good adaptation and increase in fitness factor. Figure 6 describes how lineages adapt in simulation time. High fitness factor can be explained by gaining beneficial mutation by a lineage. Low value of fitness mostly occurs in lineages for which mutation effects were neutral or have no mutations at all. Differentiation inside population is visible and caused by multiple mutations in big groups and small number of mutations in small ones.

3.3. Summary

Computational experiments give insights into evolution of fitness of microbial population. Microbial populations are adapting to environmental conditions by the process of asexual evolution. Due to the lack of recombination force, the whole microbial population can be partitioned into clones – subpopulation of identical or similar genetic profiles. Adaptation is fast because of cell replications and advantageous mutation.

The evolutionary process is studied by the clonal evolution theory. In this paper clonal evolution is numerically simulated by appropriately defined Gillespie simulation engine presented in subsection 2.2. Stochastic simulations reproduce clonal evolution scenarios with fast adaptation as observed in many biological systems, such as population of bacteria, some fungi, or in the processes of mitotic evolution of cellular subpopulations of organisms, importantly in cancer cellular populations development. Arising many genetic duplicates provides very fast, nearly chaotic, population growth with lack of genetic variety. Cell mutation causes differentiation between subpopulations. Mutations can cause positive, negative and neutral effect. Positive effect is observed when mutation persist in many cells, negative causes cell death. For good interpretation mutation effect valiant allelic frequency (VAF) coefficient is needed to be defined. Positive effect mutation is observed for high values of VAF, neutral and negative for low.

Population behavior is similar in both experiments pursued in this study – lineages sizes increase, some of them extinct and some stay on steady level. Gaining by lineage beneficial mutation causes good adaptation of clone as a result of fitness factor increase. Simulation parameters are dependent mostly on big groups of clones which are replacing smaller ones. Death probability increases while population growth, division probability changes because of mutations. Big clone groups easily compensate cells death because of good adaptation while smaller ones mostly regress and extinct. The beneficial mutations are visible because adaptive lineage size distribution moves to higher values.

4. CONCLUSION

The main conclusion of this study is that microbial experimental evolution traced with high resolution by using barcoding technique can be efficiently reproduced by using Gillespie simulation engine. In contrast to branching process from (Levy et al., 2015) our model contains direct mechanism of whole microbial population growth or decline. Our simulation algorithm is computationally efficient, it scales well with large population sizes. Our computations prove that results of simulations can easily reconstruct observations obtained in the microbial experimental evolution scenario. Results of our simulations algorithms shown in Figures 2-5 show good agreement with experimental data. Moreover, Gillespie simulation technique is very flexible, so it will easily cover possible variations in measurement scenarios in the future versions of experimental evolution researches.

Acknowledgements

REFERENCES

Baar, M., Coquille, L., Mayer, H., Hölzel, M., Rogava, M., Tüting, T., & Bovier, A. (2016). A stochastic model for immunotherapy of cancer. *Scientific Reports*, *6*(1), 24169. https://doi.org/10.1038/srep24169

Beckman, R. A., & Loeb, L. A. (2005). Negative Clonal Selection in Tumor Evolution. *Genetics*, *171*(4), 2123–2131. https://doi.org/10.1534/genetics.105.040840

Blundell, J. R., Schwartz, K., Francois, D., Fisher, D. S., Sherlock, G., & Levy, S. F. (2019). The dynamics of adaptive genetic diversity during the early stages of clonal evolution. *Nature Ecology & Evolution*, *3*(2), 293–301. https://doi.org/10.1038/s41559-018-0758-1

Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K. W., Vogelstein, B., & Nowak, M. A. (2010). Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, *107*(43), 18545–18550. https://doi.org/10.1073/pnas.1010978107

Bush, S. J., Foster, D., Eyre, D. W., Clark, E. L., De Maio, N., Shaw, L. P., Stoesser, N., Peto, T. E. A., Crook, D. W., & Walker, A. S. (2020). Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines. *GigaScience*, *9*(2), giaa007. https://doi.org/10.1093/gigascience/giaa007

Cao, Y., Gillespie, D. T., & Petzold, L. R. (2006). Efficient step size selection for the tau-leaping simulation method. *The Journal of Chemical Physics*, *124*(4), 044109. https://doi.org/10.1063/1.2159468

Castillo, F., & Virgilio, N. (2015). *Stochastic Modeling of Cancer Tumors using Moran Models and an Application to Cancer Genetics* [Thesis, Rice University]. https://scholarship.rice.edu/handle/1911/87795

Desai, M. M., & Fisher, D. S. (2007). Beneficial Mutation–Selection Balance and the Effect of Linkage on Positive Selection. *Genetics*, *176*(3), 1759–1798. https://doi.org/10.1534/genetics.106.067678

Foo, J., Leder, K., & Michor, F. (2011). Stochastic dynamics of cancer initiation. *Physical Biology*, *8*(1), 015002. https://doi.org/10.1088/1478-3975/8/1/015002

Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, *115*(4), 1716–1733. https://doi.org/10.1063/1.1378322

Kinnersley, M., Schwartz, K., Yang, D.-D., Sherlock, G., & Rosenzweig, F. (2021). Evolutionary dynamics and structural consequences of de novo beneficial mutations and mutant lineages arising in a constant environment. *BMC Biology*, *19*(1), 20. https://doi.org/10.1186/s12915-021-00954-0

Kvitek, D. J., & Sherlock, G. (2013). Whole Genome, Whole Population Sequencing Reveals That Loss of Signaling Networks Is the Major Adaptive Strategy in a Constant Environment. *PLOS Genetics*, *9*(11), e1003972. https://doi.org/10.1371/journal.pgen.1003972

Levy, S. F., Blundell, J. R., Venkataram, S., Petrov, D. A., Fisher, D. S., & Sherlock, G. (2015). Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*, *519*(7542), 181–186. https://doi.org/10.1038/nature14279

Marchetti, L., Priami, C., & Thanh, V. H. (2017). *Simulation Algorithms for Computational Systems Biology*. Springer International Publishing. https://doi.org/10.1007/978-3-319-63113-4

McFarland, C. D., Mirny, L. A., & Korolev, K. S. (2014). Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. *Proceedings of the National Academy of Sciences*, *111*(42), 15138–15143. https://doi.org/10.1073/pnas.1404341111

Neher, R. A. (2013). Genetic draft, selective interference, and population genetics of rapid adaptation. *Annual Review of Ecology, Evolution, and Systematics*, *44*(1), 195–215. https://doi.org/10.1146/annurev-ecolsys-110512-135920

Nguyen Ba, A. N., Cvijović, I., Rojas Echenique, J. I., Lawrence, K. R., Rego-Costa, A., Liu, X., Levy, S. F., & Desai, M. M. (2019). High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast. *Nature*, *575*(7783), 494–499. https://doi.org/10.1038/s41586-019-1749-3

Wang, C.-H., Matin, S., George, A. B., & Korolev, K. S. (2019). Pinned, locked, pushed, and pulled traveling waves in structured environments. *Theoretical Population Biology*, *127*, 102–119. https://doi.org/10.1016/j.tpb.2019.04.003

Wild, G. (2011). Inclusive Fitness from Multitype Branching Processes. *Bulletin of Mathematical Biology*, *73*(5), 1028–1051. https://doi.org/10.1007/s11538-010-9551-2

Yakovlev, A. Y., Stoimenova, V. K., & Yanev, N. M. (2008). Branching Processes as Models of Progenitor Cell Populations and Estimation of the Offspring Distributions. *Journal of the American Statistical Association*, *103*(484), 1357–1366. https://doi.org/10.1198/016214508000000913