Computational biology project II:

Evolution of bacteria on the example of E.Coli

MARKOVIĆ Milena (milena.markovic@grenoble-inp.org)
VONDRAČEK Dušan (dusan.vondracek@etu-univ.grenoble.alpes.fr)

December, 2020

1. Long-term evolution of bacterial populations

Following the experimental design of Richard Lenski, 8 populations of Escherichia coli were propagated in a lab for 50,000 generations, with 1:100 daily dilutions. At several time points, a clone from each of the 8 populations was sequenced. The file mutations_descendants.csv contains all mutations detected in each population at each timepoint. The file ancestor.csv lists the genes of the ancestor, which may be useful to interpret the mutations.

1.1 General questions

Question 1 — *If a mutation is found at a given time point in one of the populations, will it always also be found at later time points for the same population? Why?*

The answer can't be given in a simple way. We will try to explain the phenomenon of occurring mutations by using one of the pictures from the lecture slides (originally from Herron & Freeman[3]) presented on Figure 1.

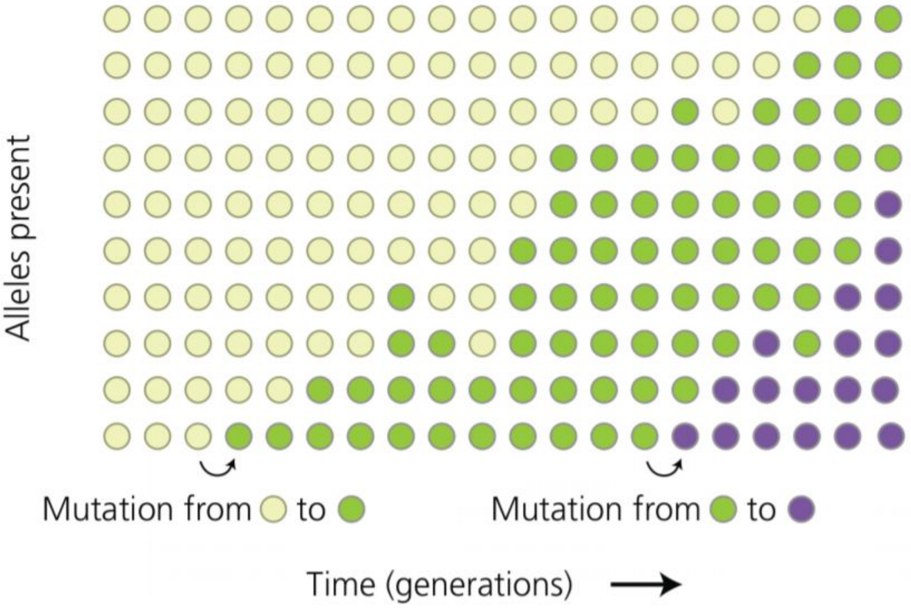


Figure 1. Mutation propagation through generations

All of the different genomes (individuals) that are in some population are colored according to one particular characteristic of interest. At generation zero, we have a very uniform population (all of the individuals have the same characteristic). At some point (in our picture, it is between generations 3 and 4) a mutation happens fully randomly. Moreso, if the mutation gives some advantage, we expect that over the next generation it's going to spread in the population (natural selection acting on the population). But if after some generations another mutation happens which makes the individuals even more fit, it will start spreading and eventually all of the individuals will mutate. The final generation we get is a good example of evolution by positive selection which drives genetic change over time.

Now, what does this tell us? Even though there was a beneficial mutation at a given point in time, eventually it was replaced by an even more beneficial mutation. Therefore, we cannot say with certainty that a mutation will always be found at later time points, for the same population.

On a side note, if we consider humans, mutations can be inheritable. Some have no special effects on the proper functioning of the organism, and some others make us more vulnerable to specific health conditions. Also, mutations may occur throughout our lives. They occur all the time and usually have absolutely no impact. Usually the body's defence systems pick them up and fix the mistakes or destroy the faulty cells. But sometimes the system breaks down and the genetic changes cause a health condition.

To conclude, mutations may stay present for a very long time in a population, but they most probably will not be always found at later time points.

Question 2 — *Can you find evidence of positive selection in this dataset?*

As Yang and Bielawski summed up in their work "Statistical methods for detecting molecular adaptation"[2], positive selection is a selection driving an increase in frequency of beneficial mutations. For this purpose, we will observe only SNP mutations. Traditionally, synonymous and nonsynonymous substitution rates are defined in the context of comparing two DNA sequences, with dS and dN as the numbers of synonymous and nonsynonymous substitutions per site, respectively. Thus, the ratio ω=dN/dS measures the difference between the two rates. If an amino acid change is neutral, it will be fixed at the same rate as a synonymous mutation, with ω=1. If the amino acid change is deleterious, purifying selection will reduce its fixation rate, thus ω<1. Only when the amino acid change offers a selective advantage is it fixed at a higher rate than a synonymous mutation, with ω>1. Therefore, an ω ratio significantly higher than one is convincing evidence for diversifying selection.

We will calculate this by counting the number of non-synonymous and synonymous substitutions within each gene that contains them at the end of last generation (time=50000). Then, we will check if the ratio of these two measures is ω>1 in order to find the ones that indicate positive selection. In this dataset (figure 2), evidence of positive selection can be found in one gene in population Pop1 and in 52 genes in Pop8. The method used to obtain these results can be found in the file *positive_selection.py*.

Population	Feat_id	dN	dS	Gene	Description	ω
Pop1	ECB_01306	2	1	ycjZ	predicted DNA-binding transcriptional regulator	2
Pop8	ECB_00138	2	1	htrE	predicted outer membrane usher protein	2
Pop8	ECB_00256	5	1	eaeH	attaching and effacing protein	5
Pop8	ECB_00288	2	1	prpD	2-methylcitrate dehydratase	2
Pop8	ECB_00522	2	1	nfrA	bacteriophage N4 receptor	2
Pop8	ECB_00554	2	1	fepE	regulator of length of O-antigen component of lipopolysaccharide chains	2
Pop8	ECB_00586	2	1	citC	citrate lyase synthetase	2
Pop8	ECB_00649	2	1	speF	ornithine decarboxylase isozyme	2
Pop8	ECB_00790	2	1	ybiW	predicted pyruvate formate lyase	2
Pop8	ECB_00801	2	1	yliF	predicted diguanylate cyclase	2
Pop8	ECB_00845	2	1	ECB_00845	hypothetical protein	2
Pop8	ECB_00906	2	1	pflA	pyruvate formate lyase activating enzyme 1	2
Pop8	ECB_00944	4	1	ycbS	predicted outer membrane usher protein	4
Pop8	ECB_01025	2	1	ycdR	predicted enzyme associated with biofilm formation	2
Pop8	ECB_01046	2	1	mdoH	glucosyltransferase MdoH	2
Pop8	ECB_01177	4	1	ycgV	predicted adhesin	4
Pop8	ECB_01215	2	1	adhE	fused acetaldehyde-CoA dehydrogenase/iron-dependent alcohol dehydrogenase/pyruvate-formate lyase deactivase	2
Pop8	ECB_01245	2	1	yciL	23S rRNA pseudouridylate synthase	2
Pop8	ECB_01362	2	1	ydbD	hypothetical protein	2
Pop8	ECB_01432	2	1	fdnG	formate dehydrogenase-N	2
Pop8	ECB_01444	2	1	yddR	D-ala-D-ala transporter subunit	2
Pop8	ECB_01510	2	1	ECB_01510	putative tail component of prophage	2
Pop8	ECB_01563	2	1	mlc	DNA-binding transcriptional repressor	2
Pop8	ECB_01620	2	1	nemA	N-ethylmaleimide reductase	2
Pop8	ECB_01843	3	1	torZ	trimethylamine N-oxide reductase system III	3
Pop8	ECB_01965	2	1	wcaA	predicted glycosyl transferase	2
Pop8	ECB_02159	2	1	yfaL	adhesin	2
Pop8	ECB_02217	2	1	yfbS	predicted transporter	2
Pop8	ECB_02366	2	1	ypfI	predicted hydrolase	2
Pop8	ECB_02442	2	1	yphH	predicted DNA-binding transcriptional regulator	2
Pop8	ECB_02512	2	1	ECB_02512	conserved hypothetical protein	2
Pop8	ECB_02573	2	1	hycC	NADH dehydrogenase subunit N	2
Pop8	ECB_02632	2	1	gudD	(D)-glucarate dehydratase 1	2
Pop8	ECB_02641	2	1	sdaC	predicted serine transporter	2
Pop8	ECB_02653	2	1	fucI	L-fucose isomerase	2
Pop8	ECB_02682	2	1	tas	predicted oxidoreductase	2

Pop8	ECB_02703	2	1	ygeW	hypothetical protein	2
Pop8	ECB_02751	2	1	ygfH	propionyl-CoA:succinate-CoA transferase	2
Pop8	ECB_03068	3	1	yhbH	predicted ribosome-associated	3
Pop8	ECB_03124	3	1	acrF	multidrug efflux system protein	3
Pop8	ECB_03176	3	1	gspD	general secretory pathway component	3
Pop8	ECB_03229	3	1	yhfT	predicted inner membrane protein	3
Pop8	ECB_03265	2	1	yhgH	gluconate periplasmic binding protein with phosphoribosyltransferase domain	2
Pop8	ECB_03270	3	1	malT	transcriptional regulator MalT	3
Pop8	ECB_03277	2	1	glpD	sn-glycerol-3-phosphate dehydrogenase	2
Pop8	ECB_03423	2	1	malS	periplasmic alpha-amylase precursor	2
Pop8	ECB_03459	2	1	ECB_03459	conserved hypothetical protein	2
Pop8	ECB_03507	3	1	spoT	bifunctional (p)ppGpp synthetase II/ guanosine-3'	3
Pop8	ECB_03719	2	1	ECB_03719	conserved hypothetical protein	2
Pop8	ECB_03980	3	2	yjdA	conserved protein with nucleoside triphosphate hydrolase domain	1.5
Pop8	ECB_04092	4	2	ytfN	hypothetical protein	2
Pop8	ECB_04255	3	1	yjjW	predicted pyruvate formate lyase activating enzyme	3
Pop8	ECB_04256	2	1	yjjI	hypothetical protein	2

Figure 2. Results of search for genes with the ratio dN/dS > 1

1.2 Mutation rate using synonymous substitutions

Among the different types of mutations, SNP (single nucleotide polymorphism) are the easiest to understand: a nucleotide is mutated into a different nucleotide. However such mutations do not always translate into amino-acids changes: synonymous SNP do not change the amino-acid encoded by the involved codons.

Question 3 — *What may explain the fixation of such synonymous mutations?*

A synonymous substitution (sometimes also called a silent substitution) is the evolutionary substitution of one base for another in a part of a gene coding for a protein, such that the produced amino acid sequence is not modified, but rather remains the same.

The explanation for this lies in the degeneracy of the genetic code. Degeneracy results because there are more codons than encodable amino acids. This means that some amino acids are coded for by more than one three base-paired codon.
Since some of the codons for a given amino acid differ by just one base pair from other codons which code for the same amino acid, a mutation that replaces the usual base by an alternative one will result in the same amino acid (and thus the name synonymous mutation - since the protein composed of the amino acids remains unchanged). Also, there is evidence that rates of nucleotide substitution are particularly high in the third position of a codon, where there is little functional constraint.

Assuming these synonymous SNP are neutral, we would like to use them to get a gross estimate of mutation rate. You can assume that 25% of the naturally occurring synonymous mutations are neutral, and that the ancestral genome contains 4Mbp of coding DNA.

Question 4 — *Could mutation rate vary between different time points or between the populations, and why?*

The results of research conducted by R. E. Lenski and M. Travisano summed up in [1] tell us that the answer to this question can be answered positively:
In the experiment, they observed the evolutionary change of 12 populations of Escherichia coli propagated for 10,000 generations in identical environments. What the results showed is that both morphology (cell size) and fitness (measured in competition with the ancestor) evolved rapidly for the first 2000 generations after the populations were introduced into the experimental environment, but both were nearly static for the last 5000 generations. In addition, the replicate populations diverged significantly one from another in both morphology and mean fitness, even though they were evolving in identical environments.

Both the nature of the gene and its environment can influence the mutation rate. In the above example, we suspect that the mutations occurred more rapidly because of the introduction of the populations to a new environment. But after around 2000 generations, the mutation rate lowered, so the most reasonable interpretation for the eventual stasis in the experimental populations was that the organisms have "run out of ways" to become much better adapted to their environment.

Because all populations have the same ancestor, the mutation rate at time point 0 was similar for all populations. If you find that at later time points the populations differ in mutation rates, find and indicate the time point at which the changes occur, and try to identify the genetic mechanism explaining this change.

If several populations seem to have similar mutation rates, you may pool the data for these populations to get a more precise estimate.

Question 5 — *Give an estimate of mutation rate with a confidence interval for each final population (or each group of final populations with similar mutation rates).*

We calculate the mutation rate μ with the following formula:

$$\mu = \frac{dS}{G*N} \quad (1)$$

where:
dS is the number of observed synonymous mutations
G is the number of generations
N is the effective number of synonymous target sites (=number of base pairs of the coding DNA)

The unit of measurement of the mutation rate is ***mutations / (generations * base pairs)***. In other words, number of mutations per base pair per generation.

These estimates do not take into account base composition or changes in genome size! In other words, we assume the insertions and deletions happen in a similar amount so that the genome length is viewed to be constant.

Now, let's use the previous formula to compute the mutation rates for each population. According to our assumption, dS (the number of synonymous mutations) takes up 25% of all SNP mutations whereas N (the effective number of synonymous target sites) is 4 million base pairs. We will compute the mutation rates in the generation of the final generation, so G (the number of generations) should equal 50000.

For that purpose, we computed the number of synonymous SNP mutations from every population. We did this in two ways:

A. By counting all SNP mutations in the last generation and approximating the number of synonymous mutations as 25% of this number
B. By counting SNP mutations categorized as 'snp_nonsynonymous'.

Remark: For the purpose of the report (and under the assumption made) we will focus on the count A! But, throughout this Question we will continue using both, simply for comparison. As for the confidence interval as well as the following questions, we will only observe the count A. The exact methods that we used for counting the mutations can be found in the accompanying file *mutation_rate.py*.

Population	SNP Mutations at generation 50000	Synonymous mutation count A	Synonymous mutation count B
1	1081	270.25	156
2	51	12.75	3
3	40	10	3
4	49	12.25	7
5	38	9.5	0
6	41	10.25	2
7	45	11.25	5
8	2585	646.25	294

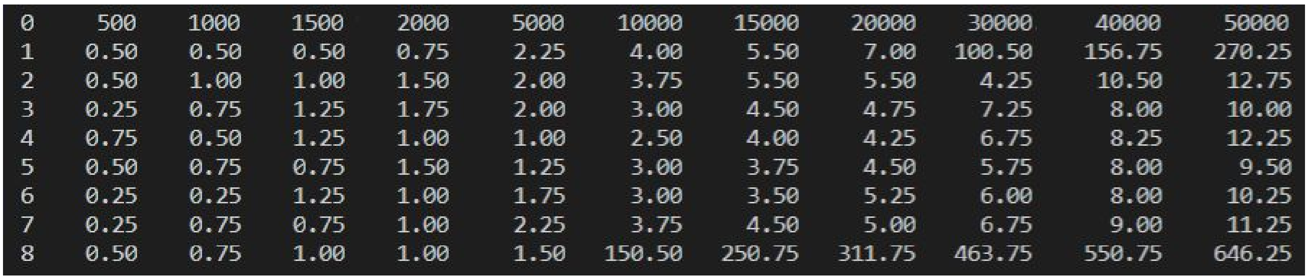
Figure 3. Results of counting SNP mutations in generation 50,000

Then, using the formula (1), we calculated the following mutation rates for each population at generation 50,000 (Figure 4)

Population	Mutation rate A ($*10^{-11}$)	Mutation rate B ($*10^{-11}$)
1	135.13	78.00
2	6.38	1.50
3	5.00	1.50
4	6.13	3.50
5	4.75	0
6	5.13	1.00
7	5.63	2.50
8	323.13	147.00

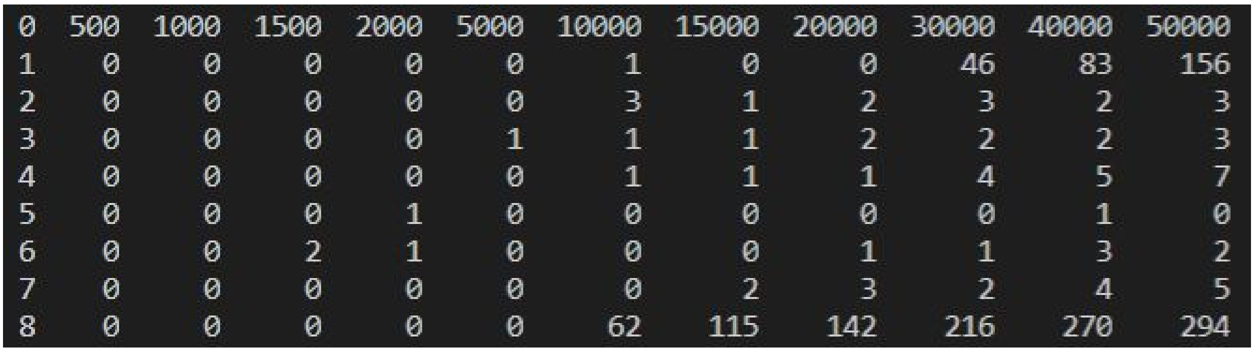
Figure 4. Calculated mutation rates for both

To expand our analysis, we computed the numbers of mutations after each of the generations given in the data. We did this so that we can follow the mutation rates throughout the evolution of the populations. The following tables show the total number of synonymous mutations per population in different generations (Figures 5 and 6)



```
0    500   1000   1500   2000    5000    10000    15000    20000    30000    40000    50000
1   0.50   0.50   0.50   0.75    2.25     4.00     5.50     7.00   100.50   156.75   270.25
2   0.50   1.00   1.00   1.50    2.00     3.75     5.50     5.50     4.25    10.50    12.75
3   0.25   0.75   1.25   1.75    2.00     3.00     4.50     4.75     7.25     8.00    10.00
4   0.75   0.50   1.25   1.00    1.00     2.50     4.00     4.25     6.75     8.25    12.25
5   0.50   0.75   0.75   1.50    1.25     3.00     3.75     4.50     5.75     8.00     9.50
6   0.25   0.25   1.25   1.00    1.75     3.00     3.50     5.25     6.00     8.00    10.25
7   0.25   0.75   0.75   1.00    2.25     3.75     4.50     5.00     6.75     9.00    11.25
8   0.50   0.75   1.00   1.00    1.50   150.50   250.75   311.75   463.75   550.75   646.25
```

Figure 5. Total number of synonymous mutations per generation, count A



```
0  500  1000  1500  2000  5000  10000  15000  20000  30000  40000  50000
1    0     0     0     0     0      1      0      0     46     83    156
2    0     0     0     0     0      3      1      2      3      2      3
3    0     0     0     0     1      1      1      2      2      2      3
4    0     0     0     0     0      1      1      1      4      5      7
5    0     0     0     1     0      0      0      0      0      1      0
6    0     0     2     1     0      0      0      1      1      3      2
7    0     0     0     0     0      0      2      3      2      4      5
8    0     0     0     0     0     62    115    142    216    270    294
```

Figure 6. Total number of synonymous mutations per generation, count B

Now we can plot the mutation rates for each population as a function of time (or generations):

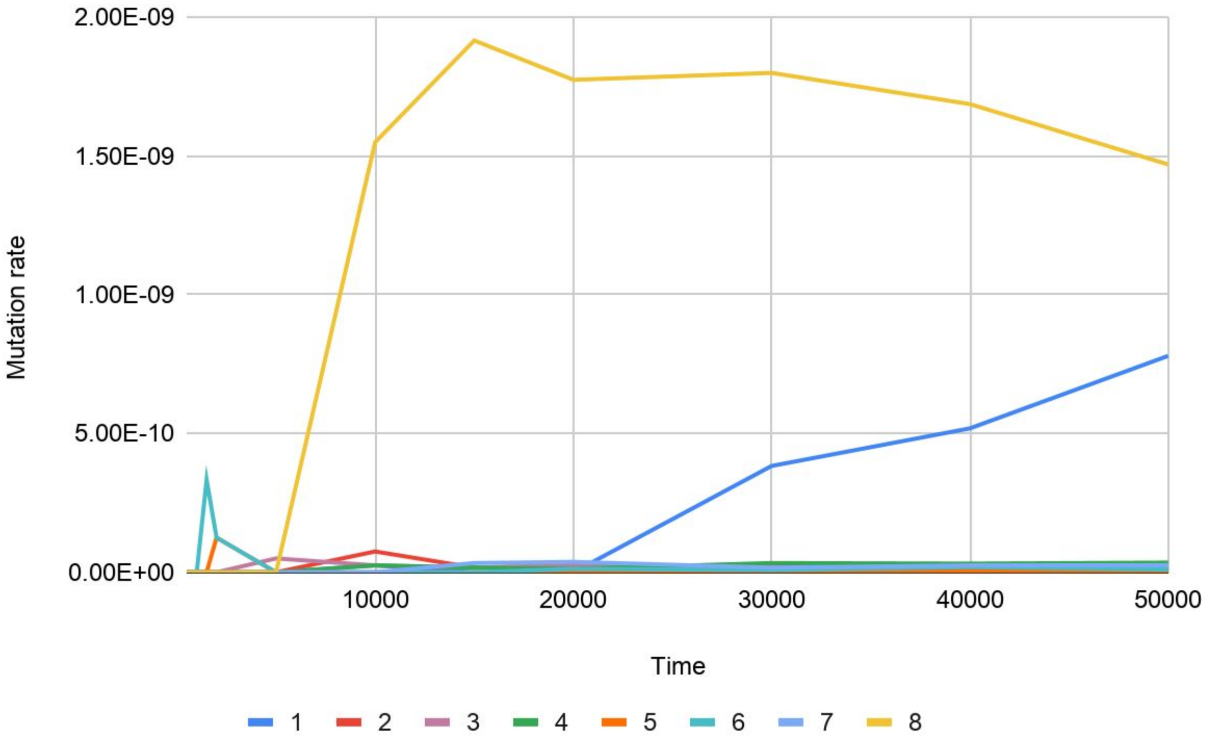


Figure 7. Mutation rates at generation T for each population, count A

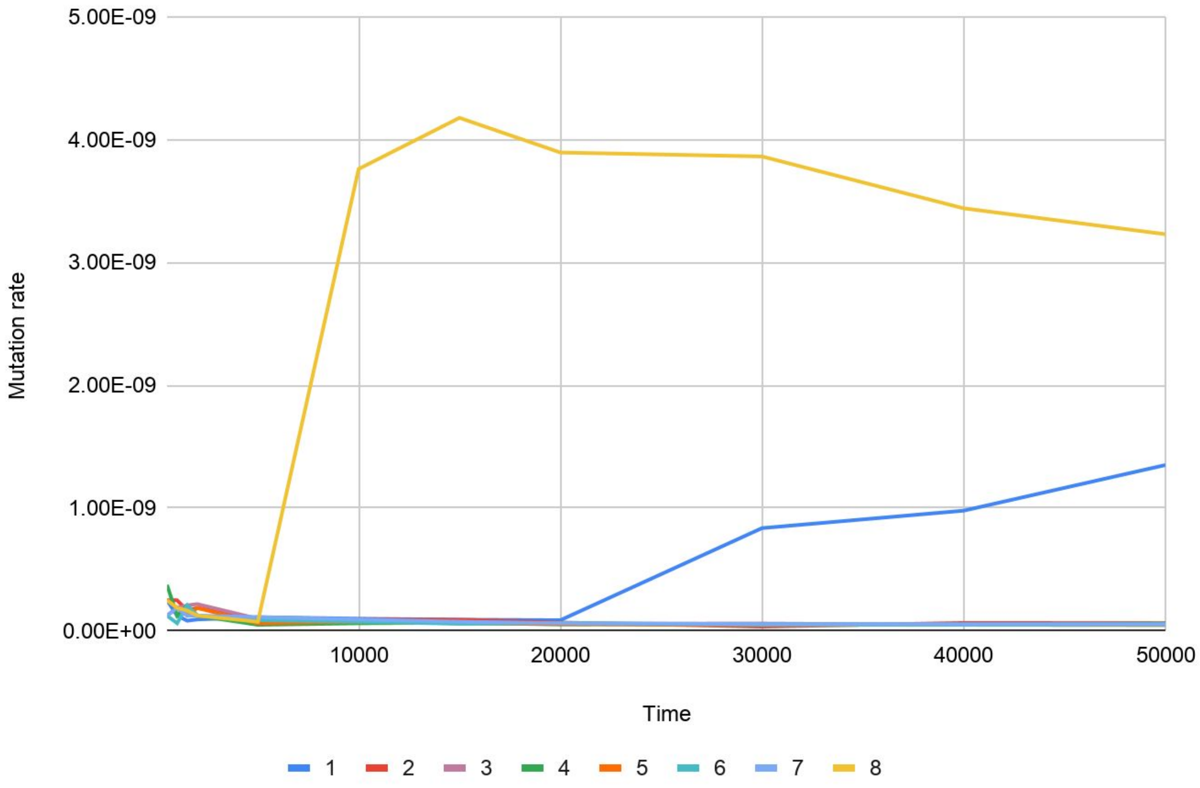


Figure 8. Mutation rates at generation T for each population, approximation B

As it can be seen from the table and the graphs, generations 1 and 8 have more significant mutation rates compared to populations [2, 7], in which the mutation rates are rather small. Since the mutation for populations [2, 7] are similar, we can group them and calculate only one confidence interval for them.

Next, we compute the 95% confidence intervals, estimated from the binomial distribution *B(N,p),* where *N* is the total number of base pairs of the codon, and *p* is the obtained value for the mutation rate(s). Since ***N* is very large and *p* is very small**, we will approximate the binomial distribution with the Poisson distribution. Then, we will view the mutation rates as the mean values of the separate Poisson distributions, and calculate the confidence intervals for the given values. This will be done by using the Normal approximation (since we have a large sample size). The formula for the confidence interval is as follows:

$$(\lambda - z * \sqrt{\frac{\lambda}{N}} \ , \ \lambda + z * \sqrt{\frac{\lambda}{N}})$$

where *z* is the "z-value" for the desired level of confidence (in our case, *z* = 1.96 for the 95% confidence). We compute three different confidence intervals for mutation rates, one for the 1st population's final generation, one for the 8th population's final generation, and one for the group of final populations [2, 7] which have similar mutation rates. For the group of populations we computed the average mutation rate as the value of *p*. The confidence intervals are given in table 9.

Population (or group)	Confidence interval limits $*10^{-11}$
1	$1.35*10^{-9} \pm 3.60*10^{-8}$
2-7	$5.50*10^{-11} \pm 7.26*10^{-9}$
8	$3.23*10^{-9} \pm 5.57*10^{-8}$

Figure 9. Confidence intervals for computed mutation rates

1.3 Parallel evolution

We expect that many neutral mutations will be fixed due to Genetic hitchhiking. To identify beneficial mutations and find which genes are under positive selection, we would like to analyse the repeatability of evolution in the 8 populations. When the same mutation is fixed in several populations, it is considerably more likely to be a beneficial mutation.

Question 6 — *Based on this reasoning, analyze the data to identify candidates with beneficial mutations. Compute the probability that a mutation is fixed in n or more populations under the null hypothesis that it is non-beneficial, and deduce a p-value for each candidate beneficial mutation. For each of these mutations that are SNP, indicate the name of the involved gene as well as the p-value.*

What we wish to find is a mutation that is fixed in more populations because, as written above, in that case it is more likely to be beneficial.

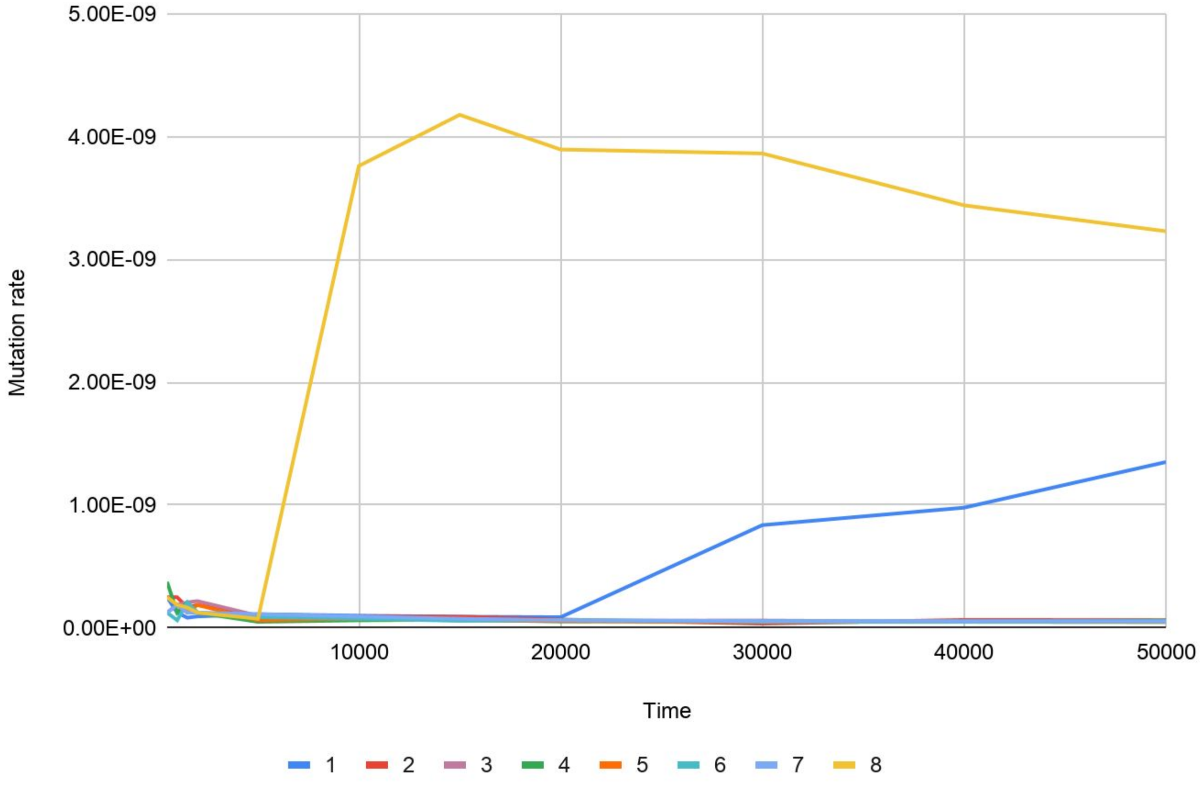


Figure 10. Mutation rates at generation T for each population, approximation B

Let's focus on the 1st and the 8th population for the time being. If we observe (figure X) we can notice that the mutations start to rise rapidly in different generations. To be more exact, the mutation rate for pop1 starts to rise between generations 20000 and 30000, whereas the mutation rate for pop8 starts to rise between generations 5000 and 10000.

We are interested to see whether a similar first mutation triggered these rises in mutation rates. Therefore, we observe all the mutations at generation 30000 for pop1, and all the mutations at generation 10000 for pop8. We wish to find mutations from each of the populations which change the same part of the genome.

To do so, we looked for all mutations that appear in both population 1 in generation 30.000 and population 8 in generation 10.000, as well as all mutations that are in the neighbourhood $\varepsilon = 10BP$ around the start position of their mutations - relevant code can be found in *parallel_evolution.py*. This resulted in the following list of mutations:

```
('Pop1', 500, 'DEL', 3894998, 3898057, 'Δ3,060 bp', 'large_deletion')
('Pop1', 30000, 'INS', 114034, 114034, '(C)6->7', 'small_indel')
('Pop2', 500, 'DEL', 3894997, 3900623, 'Δ5,627 bp', 'large_deletion')
('Pop3', 500, 'DEL', 3894997, 3902809, 'Δ7,813 bp', 'large_deletion')
('Pop4', 500, 'DEL', 3894997, 3901240, 'Δ6,244 bp', 'large_deletion')
('Pop5', 1500, 'DEL', 3894997, 3901410, 'Δ6,414 bp', 'large_deletion')
('Pop6', 1500, 'DEL', 3894997, 3898943, 'Δ3,947 bp', 'large_deletion’)
('Pop7', 1000, 'DEL', 3894997, 3895609, 'Δ613 bp', 'large_deletion')
('Pop8', 10000, 'DEL', 114034, 114034, '(C)6->5', 'small_indel')
('Pop8', 1000, 'DEL', 3894997, 3898178, 'Δ3,182 bp', 'large_deletion')
```

Here we can observe two mutations. The first one is a large deletion at start position 3894997 or 3894998, depending on the population, and ranging between 613 and 7,813 base pairs.

```
('Pop1', 500, 'DEL', 3894998, 3898057, 'Δ3,060 bp', 'large_deletion')
('Pop2', 500, 'DEL', 3894997, 3900623, 'Δ5,627 bp', 'large_deletion')
('Pop3', 500, 'DEL', 3894997, 3902809, 'Δ7,813 bp', 'large_deletion')
('Pop4', 500, 'DEL', 3894997, 3901240, 'Δ6,244 bp', 'large_deletion')
('Pop5', 1500, 'DEL', 3894997, 3901410, 'Δ6,414 bp', 'large_deletion')
('Pop6', 1500, 'DEL', 3894997, 3898943, 'Δ3,947 bp', 'large_deletion’)
('Pop7', 1000, 'DEL', 3894997, 3895609, 'Δ613 bp', 'large_deletion')
('Pop8', 1000, 'DEL', 3894997, 3898178, 'Δ3,182 bp', 'large_deletion')
```

This mutation appears in all populations, mostly in one of the first sequenced generations - both in the ones that have an increased mutation rate and the ones that do not. Some of these mutations are a deletion of an intron (Pop7) while others delete one or more protein and sugar genetic codes:

```
3895158, 3895577, 1, "ECB_03634", "rbsD", "predicted cytoplasmic sugar-binding protein"
3895585, 3897090, 1, "ECB_03635", "rbsA", "fused D-ribose transporter subunits of ABC"
3897095, 3898060, 1, "ECB_03636", "rbsC", "ribose ABC transporter permease protein"
3898085, 3898975, 1, "ECB_03637", "rbsB", "D-ribose transporter subunit"
3899101, 3900030, 1, "ECB_03638", "rbsK", "ribokinase"
3900034, 3901026, 1, "ECB_03639", "rbsR", "DNA-binding transcriptional repressor of ribose"
3900992, 3902419, -1, "ECB_03640", "yieO", "predicted multidrug or homocysteine efflux system"
3902442, 3903134, -1, "ECB_03641", "yieP", "predicted transcriptional regulator"
```

Even though these mutations happen on the same segment, they vary in length and effect greatly, so we cannot really consider them to be the same mutation. Because of this, along with the fact that none of the populations see any change in the mutation rate afterwards, it is safe to assume that this mutation is probably neutral.

The second mutation that we found is a small indel at 114034 that is observed only in populations 1 and 8 and it is found for the first time in the generations at which the mutation rate starts to rise

('Pop1', 30000, 'INS', 114034, 114034, '(C)6->7', 'small_indel')
('Pop8', 10000, 'DEL', 114034, 114034, '(C)6->5', 'small_indel')

The gene that is modified by this mutation is the following:

(113848, 114237, 1, 'ECB_00100', 'mutT', 'nucleoside triphosphate pyrophosphohydrolase', 114034)

The gene *mutT* encodes a protein that prevents AT->CG transversions during DNA replication and works along with other mutator genes to ensure high fidelity replication mechanisms which usually prevent such SNP mutations.[6] The reason why the mutations that we found on this gene could explain the increase in the mutation rate is that the mutations that happen changed it's genetic code, and as it is usually responsible for fixing A->C and T->G replication errors it now doesn't do what it was supposed to, so the increase in the mutations is mostly SNP mutations with those two transversions. To test out this hypothesis, we will check, out of all of the mutations, how many are actually the AT->CG transversions.

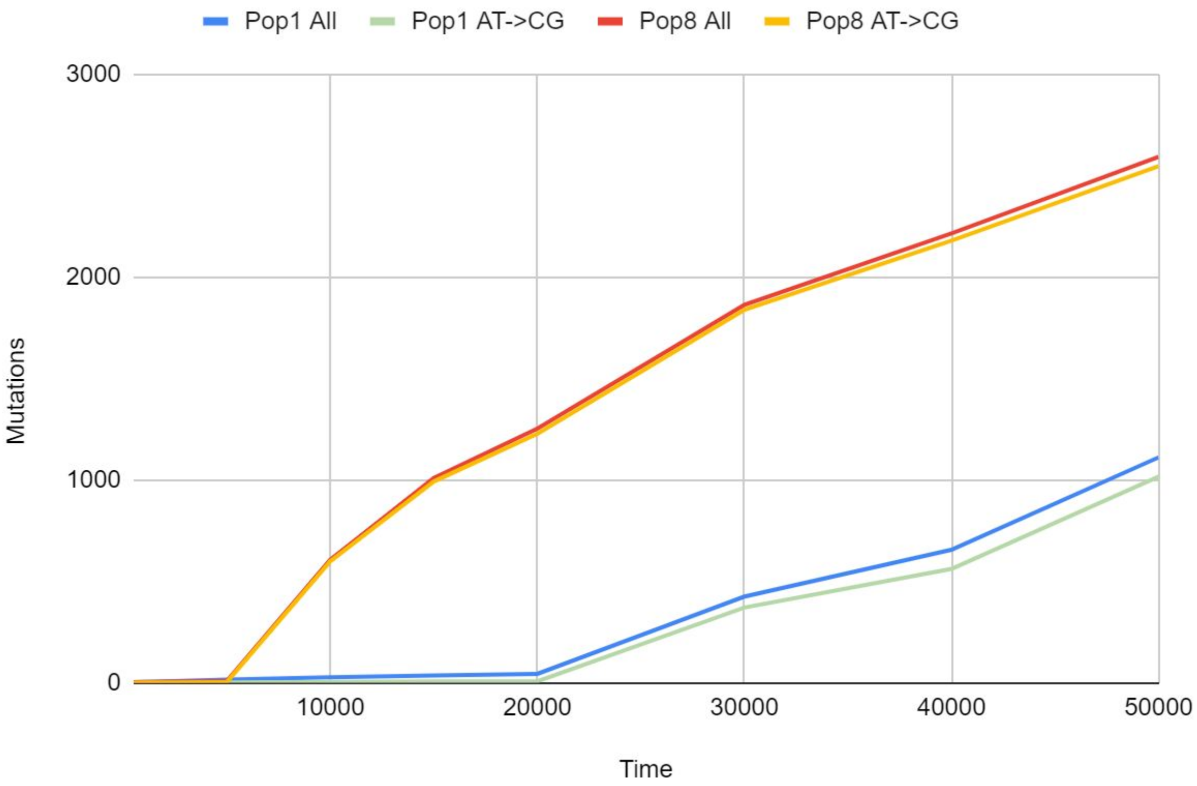


Figure 11. Comparison between the total number of mutations and AT->CG transversions

This is illustrated by the graph on figure 11: the darker color represents all mutations while the lighter one shows AT->CG. We can see that obviously, the overwhelming majority are, in fact, AT->CG: for population 1 in the final generation this is around 91.4% while for population 8 it is around 98.2%. For comparison, in the last generation before the mutation occurred, this was 15.9% for population 1 and 25% for population 8. To test out this hypothesis, we will check out of all of the mutations, how many are actually AT->CG mutations.

While this mutation is most likely not, on its own, particularly advantageous it can be beneficial. The evolutionary trajectory can go through a less optimal genotype before reaching a better one[7] which could be what is happening here as this mutation significantly increases the probability that random drift mutations might happen.

References

[1] Lenski, Richard E., and Michael Travisano. “Dynamics of adaptation and diversification: A 10,000-generation experiment with bacterial populations.” *PNAS*, vol. 91, no. 1, 1994, pp. 6808-6814. *pnas.org* [link]

[2] Yang, Ziheng, and Joseph P. Bielawski. “Statistical methods for detecting molecular adaptation.” *Trends in Ecology and Evolution*, vol. 15, no. 12, 2000, pp. 496-503. *cells*, [link]

[3] Freeman, Scott. *Evolutionary Analysis*. Prentice Hall, 1998

[4] Wielgoss S, Barrick JE, Tenaillon O, Cruveiller S, Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D. “Mutation Rate Inferred From Synonymous Substitutions in a Long-Term Evolution Experiment With Escherichia coli.” *G3* (Bethesda). 2011 Aug 1;1(3):183-186. doi: 10.1534/g3.111.000406. PMID: 22207905; PMCID: PMC3246271. [link]

[5] R.G. Fowler, R.M. Schaaper, “The role of the mutT gene of Escherichia coli in maintaining replication fidelity”, *FEMS Microbiology Reviews*, Volume 21, Issue 1, August 1997, Pages 43–54 [link]

[6] Bhatnagar SK, Bullions LC, Bessman MJ, “Characterization of the mutT nucleoside triphosphatase of Escherichia coli”. *J Biol Chem.* 1991 May 15;266(14):9050-4. PMID: 1851162. [link]

[7] Duret, L., “Neutral Theory: The Null Hypothesis of Molecular Evolution”, *Nature Education,* 2008 [link]

[8] Genetic code. (2020, December 8). In Wikipedia. [link]