Incorporation of Insertion Sequences in the artificial life software Aevol

Juliette Luiselli
team BEAGLE
under the supervision of Guillaume Beslon

7 weeks between Monday, 3rd of June and Friday, 26th July

Abstract

During my internship in the Beagle Team, I focused on the **modeling of Insertion Sequences (IS)**. Insertion Sequences are DNA sequences found in bacteria which are able to copy-paste (or **transpose**) themselves in the rest of the genome. They have therefore their own dynamic within the genome. However, this dynamic is not well understood and IS are not broadly studied. To better grasp what influences their dynamic, I **developed a model of their functioning and implemented it in the artificial life software Aevol**. Following this, I ran several **in silico experiments which revealed three types of dynamics**: either the IS are maintained with a low transposition rate, or they are strongly counter-selected when this rate is higher, or the genome size explodes when the transposition rate reaches even higher values. Moreover, I was able to shed light on the fact that a **changing environment helps IS to be maintained in the population**, even with a high transposition rate.

Preamble

The following internship has been undertaken in the context of a gap year in computer science during my biology formation. Since I was working on the representations of biological objects, many biological terms were used. In order to avoid making the text too cumbersome, they will be mostly defined in footnotes.

1 Insertion Sequences in biology and their dynamics

Insertion Sequences, usually called IS, are short DNA sequences found in bacteria known to act in a "selfish" way: they code for their own transposase, an enzyme allowing them to copy or cut and paste themselves all over the rest of the genome (see Fig.1). Although very short (800 to 2500 base pairs(bp)), IS can represent up to 2% of the genome[3], since there are up to several hundreds of them in some bacterial genomes[4].

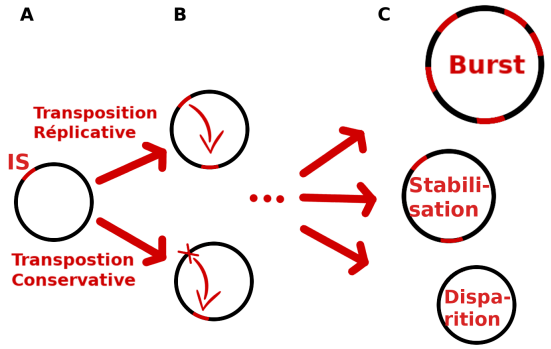

Figure 1: Insertion sequences dynamic. (A) Time 0 (B) One time step. (C) Possible long term dynamics

Due to their dynamic, the number of IS tends to explode under some circumstances[12], before they are "domesticated" by the host and reach a reasonable level again, but their exact dynamics is still not well understood[1],[9].

This domestication is thought to be due to the redundancy[1] of the genetic code[5]: As a matter of fact, it seems that some families of IS transpose onto some specific sequence of DNA[3],[14]. This target sequence could be mutated through silent mutations[2] (see Fig.2) when whithin a functional gene, which would prevent essential genes from being knocked-out and hence protect the host from most of the deleterious impacts of IS.

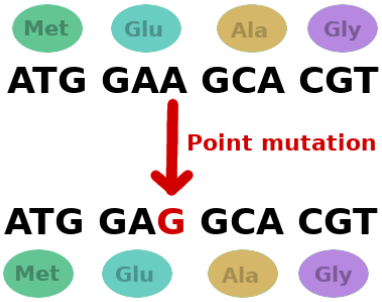

Figure 2: Proteic and DNA sequences before and after a silent mutation: the proteic sequence and hence the function of the gene is unchanged, while the DNA sequence is different from previously

However, there are many different IS famillies, and they do not all have specific target sequences. It has been suggested that some IS do not target specific sites but rather repetitive extragenic palindromic sequences[13], which can be found all over the genome, or have no target sequence at all. Some even seem to transpose preferentially at a certain distance from their original location, rather than into a specific type of DNA sequence[5].

But in any case, why are IS not always suppressed from the genome by mutating all possible sites of transposition, when a preference exists, and/or deleting the IS sequence itself?

Indeed, at first sight IS seem to be deleterious to their host since they require energy from it to do their own transposition, and they can insert into a functional gene, which generally makes it non-functional. Therefore, their existence itself has to be questioned: what mechanisms allow this selfish sequences to subsist without killing their host or being counter-selected? Two hypothesis compete here[10],[2].

On the one hand, they could have an occasional beneficial effect on the fitness[3], which could suffice

[1]The genetic code is said to be redundant because 64 codons (3 bases of DNA) code for 21 amino acids so there is more than one codon per amino acid.

[2]A silent mutation changes the DNA sequence but not the proteic sequence and hence not the function of the protein.

[3]The fitness is a value that describes the ability of an individual to survive and to reproduce itself. Although not directly measurable on real organisms, it is often approximated with the number of offsprings of the individual, and it depends on the environnement in which the individual is living.

for them to be selected. In fact, IS are strongly involved in many mutational processes since they cause duplication and deletion. Moreover, having copies of the same sequence all over the genome could favor strongly all kinds of genomic rearrangements[4,[5] since those occur preferentially within sequences similar to each other. As they are themselves transcripted, they also increase the transcription of nearby regions by recruiting enzymes for transcription[4]. Moreover, it has been observed that the exploding number of IS causes a hyper-mutability of the phenotype[12]. All these features could occasionally improve fitness when individuals are confronted to a fast changing environnement.

On the other hand, transposing fast enough to overcome natural selection and being transmitting horizontally[5] would be a different way for IS to maintain themselves, even with a negative impact on the fitness of the population.

This intership aims at answering a fundamental question about IS: **which conditions would allow the survival of IS in the genome, without them offering any direct beneficial impact on fitness?** To this end, we want to model the simplest rules that can currently be inferred from their observed dynamic. Since they strongly interact with the whole genome and the fitness of an individual as well as with other mutations, this model has to be included in a multiscale evolving simulation. IS modelisation will therefore be incorporated into AEVOL, an artificial life software developped by the Beagle Team in the INRIA laboratory of Lyon.

With this objective, I will first develop an algorithm modeling the dynamic of IS. Then, I will need to set up a simple version of AEVOL **with 4 bases**, to take the redundancy of the genetic code into account, since it does not exists in the software for now. After that, I will **integrate IS and their transposition into the software**. Once the software is running, I will calibrate the parameters to **search for the different possible dynamics**.

2 Aevol - modeling artificial life

To understand how to implement new functionalities in this software and how it allows to study evolution, it is essential to gain an insight into AEVOL itself[7]. AEVOL is a C++ sofware which has been developed by the Beagle team for almost 15 years, and is composed of over 70 000 lines of code split in many modules. To include my own code, I will use a dedicated branch on the INRIA gitlab (/beslon/aevol_ltisee/tree/juliette).

AEVOL (see Fig.3, from LIARD 2018[6]) emulates a population, composed of a fixed number a individuals. Each of them has its own binary genome. In the genome, some sequences are recognized as promoters[6] and mark the beginning of the transcripion from DNA to RNA. The transcription stops when a palindromic sequence is encountered (hairpin structure). RNAs are then translated to proteic sequences thanks to a system of recognized binding sequences and an artificial genetic code with 3 bases "codons". Each protein is then transformed into a mathematical function that corresponds to its functional contribution, represented by a triangle. Proteins are decrypted 3 bases at a time, that is why there is $2^3 = 8$ possible codons for a binary genome. One is associated with starting a protein, one with stopping it. Among the 6 others, 2 are for the width of the triangle (w_0 or w_1), 2 for the mean (m_0 or m_1) and 2 for the height (h_0 or h_1). We can thus define a binary number for each of the parameter and by normalizing it we get virtually an infinite number of possible triangles. The phenotype of an individual is the sum of its triangles, and we can define its fitness by computing the exponential of the difference between its phenotype and an optimal fitness function, the environment, which is usually a sum of Gaussians functions.

[4] A genomic rearrangement is a large scale mutation involving copying or cutting a sequence and pasting it elsewhere, possibly inversing it.

[5] Horizontal transmission occurs when non-parent organisms exchange genetic material.

[6] A promoter is a DNA sequence recognized by specific enzymes, which will begin the transcription at that localisation.

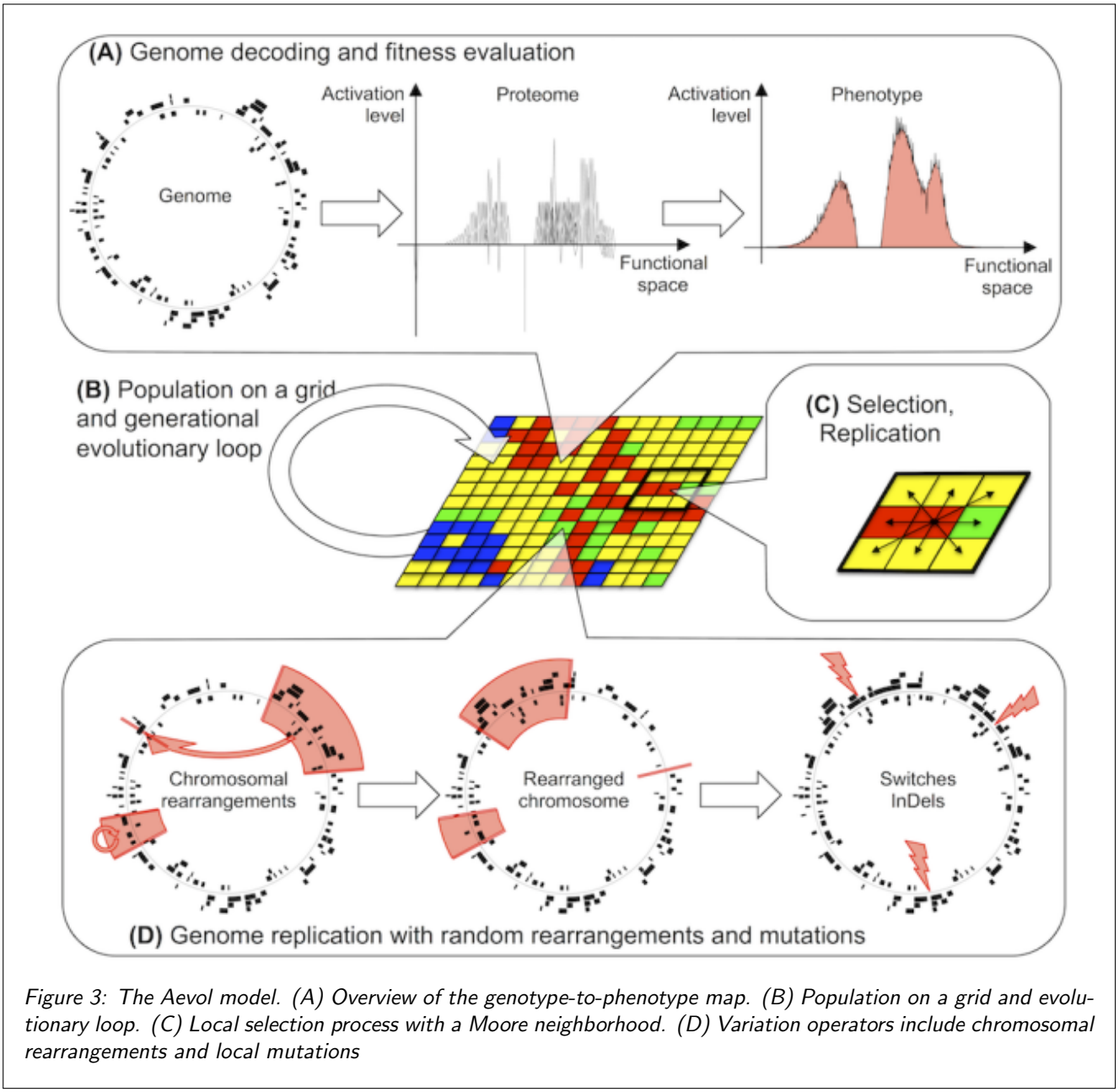

Figure 3: The Aevol model. (A) Overview of the genotype-to-phenotype map. (B) Population on a grid and evolutionary loop. (C) Local selection process with a Moore neighborhood. (D) Variation operators include chromosomal rearrangements and local mutations

Randomly selected with their fitness as weight, individuals reproduce asexually with slight chances of mutations, thus creating a new generation. At each time step, there is variation and selection, hence the emergence of evolution. This model allows to simulate evolution with only a few controllable parameters, such as the population size or the mutation rate, and to test various hypothesis on the causes of certain phenomenons. For example, it has been shown that complexity could arise in spite of natural selection, contrary to the common intuition that it exists thanks to natural selection[6]. The authors have shown that even when defining a very simple environnement (represented by a triangle optimal fitness function, hence possibly filled with a single gene and a single protein), several genes tend to arise in the early generations of the populations. This "complex" populations could not go back to a simple state (with fewer genes) unless the mutation rate was very high, and their final fitness was below the fitness of populations with fewer genes.

The mutations currently taken into account by AEVOL are point mutations, small or big insertions or deletions, and chromosomic rearrangements such as duplication, translocation or inversion (see Fig.4, from Paul BANSE).

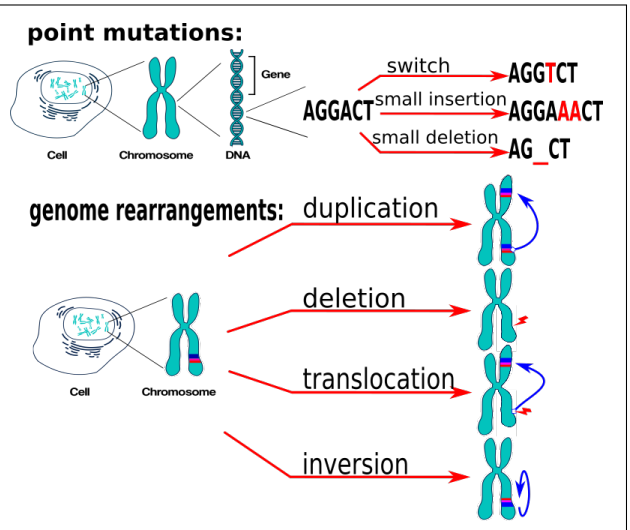

Figure 4: The different types of mutations taken into accound by AEVOL

Using this model, the aim of this study is to add IS in the AEVOL platform. IS will be defined by a given consensus sequence, from which they can differ by a certain Hamming distance to allow IS to resist mutations. Each IS will then be able to occasionaly transpose onto a target sequence, which is also defined previously.

3 Implementing transposition of IS

3.1 Redundancy of the genetic code

To include the possibility for the individual to defend itself against transposition by protecting its essential genes, we need to add the redundancy of the genetic code into AEVOL. In fact, in biology a codon is composed of 3 bases. Since there are 4 possibilities for each base (A, T, C or G), there are 64 different codons. Nevertheless, there are only 21 amino acids associated with them: several codons can code for the same amino acid and thus a point mutation can change the sequence without changing the final protein (see Fig2). So far, the genome in AEVOL was binary and each of the 8 possible codons coded for a different property. Hence, silent mutations were impossible. A simple way to introduce redundancy is to use 4 bases: 0, 1, 2 and 3 with 2 equivalent to 0 and 3 equivalent to 1 for the proteins and all preexistent mechanisms. Supposedly, this will not change the functioning of the software, except for the recognition of the IS target sites and IS sequences.

To implement redundancy, we first need to replace the initial creation of a random binary sequence by the generation of a random sequence of numbers between 0 and 3. Then, each time an action is determined by the nature of the base (0 or 1), we need to replace the test $x = 0$ by $x\%2 = 0$ and $x = 1$ by $x\%2 = 1$ to have the equivalences $1 \equiv 3$ and $2 \equiv 0$.

We need to take extra care when doing it because there are many occurences of these tests all over the code, and an unnoticed test could lead to a fully functional code while introducing biases in favour of one of the bases, or in favor of one of the strands. In fact, most often the comparison was done with an equality on the one strand and inequality on the other since "not being a 1" was equivalent to "being a 0", which is no longer true in a 4 bases model.

3.2 Algorithm of IS tranposition

In order to implement IS transposition, the naive way would be to scan the whole genome at each generation to recognize IS sequences and target sites and randomly transpose one or several IS onto a target site. However, this would be very costly to compute: a usual size for a generation is 1024 individuals, which have each a genome of several thousands base pairs so reading all genomes for over 100 000 generations, although not complex, is very costly. Another solution is thus to use the same kind of algorithm that the one currently used to model promoters.

As a matter of facts, promoters are stored in two lists, one for each strand. At the creation of the genome and each time a mutation occurs, a function checks whether this creates or destroys a promoter and updates the lists. Assuming n is the length of the genome, n_p the number of promoters, len_p the length of a promoter, and m the mean number of mutations at each generation this method is clearly advantageous while $n \gg n_p * len_p * m$.

Regarding the target sites, the chosen approach is to stick to biology: a transposed sequence can navigate through the cell, finding or not a target site before being destroyed. In the simulation, after an IS is chosen to transpose, we will try a given number of time n_{try} to find a target site by randomly choosing a site in the genome. If no target site is found in n_{try} trials, the transposition is cancelled.

We choose to implement the two types of transpositions: replicative and conservative. In a replicative transposition, the IS is copy-pasted onto a target site, while in a conservative transposition the IS is cut-pasted. In this way, if a conservative transposition fails because no target site is found, the IS will have been deleted, which occurs at various rates in biology too.

We can therefore propose this algorithm as leading idea for the code when passing from a generation to another. Naturally, the code itself will be inserted into different parts of AEVOL, and will not directly appear as such.

```
Do point mutations:
  for each point mutation:
    check if one of the IS sequence is
            concerned and update list
    check if a new IS is created

Do chromosomal rearrangement:
  for each rearrangement:
    check if one of the IS sequece is
            concerned and update list
      check if a new IS is created

Do IS transposition:
  compute the probability of transposition
            (function of the number of IS)
  compute the number of replicative and
            conservative transpositions

  for each conservative transposition:
    if there are still IS:
      remove the sequence
      try to find a target site
                  n_try times
        if found:
          transpose onto it
        update IS and promoters lists

  for each replicative transposition:
    if there are still IS:
      try to find a target site
                  n_try times
        if found:
          transpose onto it
          update IS and promoters lists
```

This code is executed only when computing the mutation, there is for now no reason for IS to affect another part of the code. They are believed to impact the transcription of nearby genes[4], but this will not be taken into account in this study since we focus on the dynamic without modeling any direct effect on fitness.

3.3 Parameters calibration: IS size, sequence and transposition rate

3.3.1 First calibration approach

Some decisions have to be made about the exact modeling of IS. First of all, their size and sequence. As a matter of fact, the usual size goes from 700bp up to 2500bp[3]. Nevertheless, AEVOL is a program, not a real bacteria. Its genome is broadly reshaped and typically much smaller than actual bacterial genomes. For example, promoters are comprised between 100 bp and 1000 bp in bacteria but are 22 bp in AEVOL (2 bases version). Moreover, we need IS to sometimes emerge spontaneously, but not too often, so as to have a significant percentage of the genome (higher than 0.1%) which is realistic (lower than 2%)

In order to calibrate this parameter, several cross tests have been launched: a size of 14, 16 or 18 bp, combined with a fixed maximum Hamming distance from the consensus of 4. Indeed, both the size of the consensus and the distance to it influence on the same parameter, the probabilty of spontaneous appearance and disapparearance, so there is no need to explore both. Each of these combinations is run 4 times with each transposition rates (10^{-4}, 10^{-5}, 10^{-6} transposition/IS/generation) and each other mutations rates (10^{-5}, 5.10^{-6}, 10^{-6} mutation/base/generation) as those parameters, the transposition rate in particular, are expected to play a huge role in the dynamics of the IS. The 108 calibration experiments were run for 100 000 generations on the Beagle cluster.

The target sequence length and sequence have to be fixed too, but the range of possibilities is much thinner since it is known to be very short in bacterias. In fact, in the literature it goes from 2 bp up to 20 bp[11]. In order to have a sequence long enough to be a little bit rare and so to potentially observe selection on it, and longer than a codon so that it is not selected as a codon, we decided to take it long (compared to the IS sequence length): 5 bp, with 1 difference allowed with respect to the consensus sequence.

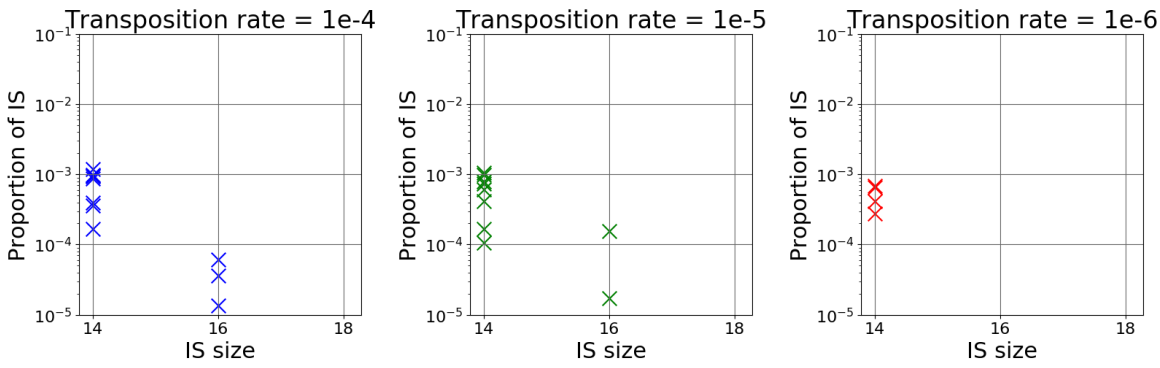

Figure 5: Proportions of IS (mean over the last 1000 generations) for the 3 IS sizes and 3 transposition rates tested. The mutation rate is 5.10^{-6} *mutation/base/generation.*
There is no data for size 18, and somes points are missing for size 16 too, because the proportion is 0 and the scale is logarithmic.

Results (see Fig.14) show that sizes of 16 and 18 do not allow to observe a significant proportion of IS, as this proportion falls to 0 in several experiments. Observing a dynamic in those conditions would be very difficult. We therefore decided to test smaller sizes (10 and 12bp) to compare with the values for a size of 14bp (see next subsection).

The transposition does not seem here to have a significant impact, contrary to what was expected. Thus, we suppose the proportion of IS observed here is nearly only due to the random presence of any defined sequence in the genome, so a higher transposition rate (10^{-3} transposition/IS/generation) will be tested.

Finally, the mutation rate had no impact on the proportion of IS and only little impact on the fitness (see Appendix 1). As it is expected to play an important role for the genome size (and so the computation time) but not on IS themselves, it was decided to take the middle value tested (5.10^{-6} mutation/base/generation).

3.3.2 Second calibration

Results of the second parameters calibration (see Fig.6) show that the proportion of IS reaches a new level when size decreases: up to 4% with a size of 10bp. This value is too high, in particular in comparison with maximal estimation of 2% announced in the introduction. To have almost 1% of IS in the genome, we choose a size of 12bp.

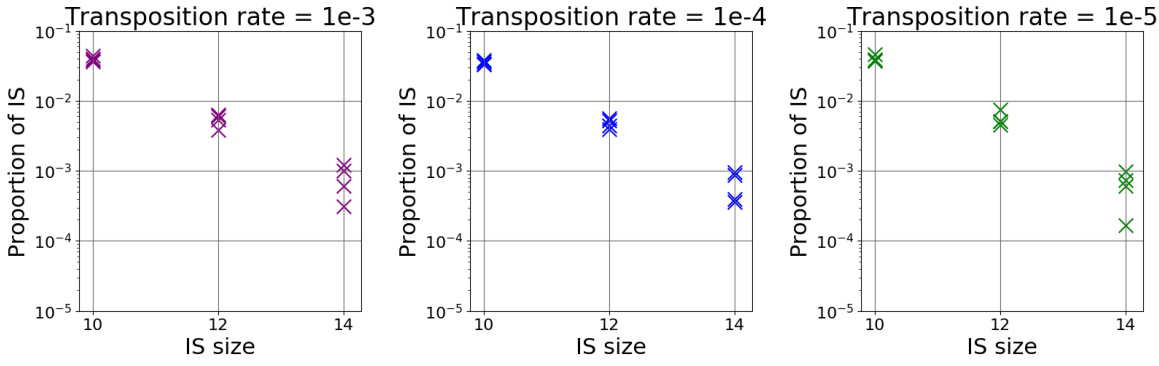

Figure 6: Proportions of IS (mean over the last 1000 generations) for the 3 sizes and 3 transposition rates tested. The mutation rate is 5.10^{-6} *mutation/base/generation.*

Once again, the transposition rate does not seem to have a crucial impact here. It will therefore be further tested with the chosen size (12bp) and mutation rate (5.10^{-6} mutation/base/generation) in order to look for different dynamics: once we do have IS in the genome at the expected proportion, its variation can be further studied.

Finally, these two calibration experiments also enable us to compare the fitness of the different populations, relatively to their proportion of IS (see Fig.7). We can note that the highest fitness are reached when there is no IS in the population, so there is indeed a deleterious effect of the presence of IS in a genome. However, as soon as IS are present, there is no distinguishable correlation between the fitness and the proportion. This suggests that their activity is low for any of those transposition rates.

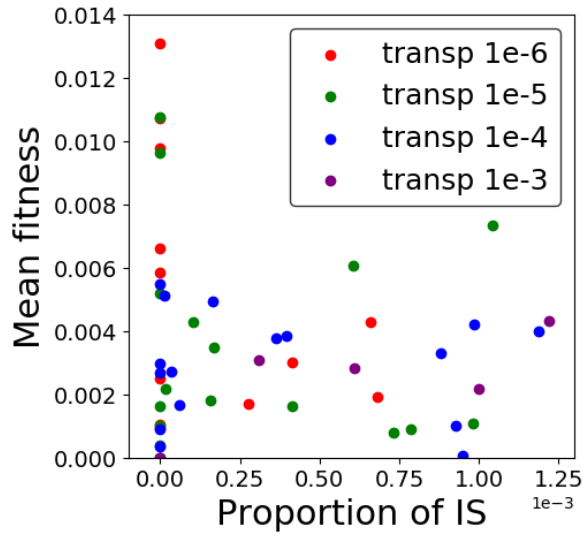

Figure 7: Mean fitness of the population plotted against the mean proportion of IS in the population (over the last 1000 generations) for the 5 sizes and 4 transposition rates tested.
No correlation is observed.

3.3.3 Does the functional form of the relation between the number of IS and the transposition rate have an major impact?

First of all, what is called the transposition rate has to be defined. Up to this point, we considered that the probability of an IS to transpose was constant at each generation regardless of the context. However, many different proposals can be found in the litterature about the functional form of the relation between the number of IS and the actual transposition rate. Some propose that the rate remains constant[5], as we did previously, but some that it decreases[9] because of a response from the cell to the accumulation of transposase, contrary to the classical hypothesis that it increases because of this accumulation, at least up to a threshold rate.

We therefore decided to test different functional forms and to observe the related dynamics. The aim here is to see whether the functional form influences the IS dynamics. Since we cannot explore all the parameter space at once, we explore this property after having calibrated the IS size, transposition rate and the mutation rate.

Five function forms were implemented into AEVOL during the internship: linearly increasing, linearly decreasing, constant, increasing hyperbolic and decreasing hyperbolic (see Fig.8). Due to time constraints, only the constant form and the linear positive form were thoroughly tested. In fact, other forms need more time to be calibrated since they contain more parameters.

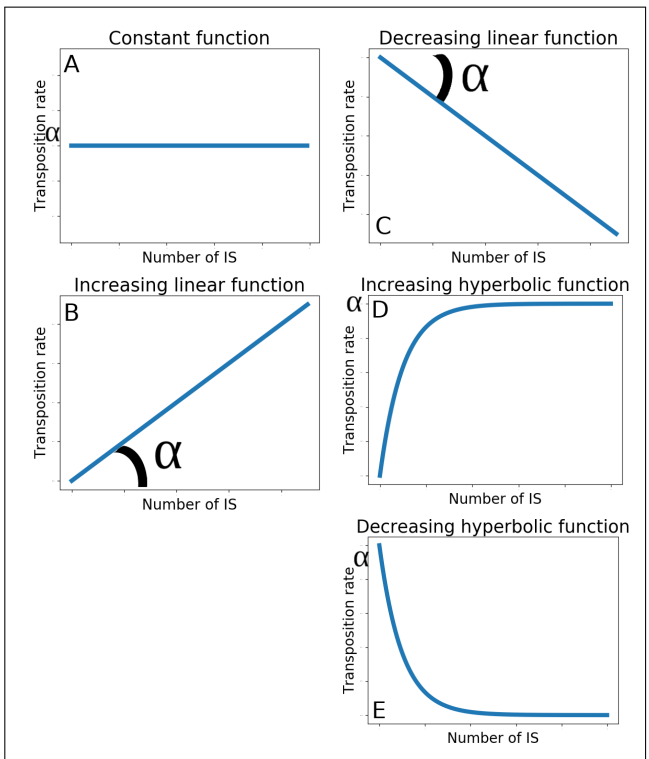

Figure 8: The different types of functional forms
(A) $R(n_{IS}) = cst$
(B) $R(n_{IS}) = n_{IS} \times \alpha$
(C) $R(n_{IS}) = -n_{IS} \times \alpha \times \beta + \alpha$
(D) $R(n_{IS}) = \alpha \times (1 - e^{-\beta \times n_{IS}})$
(E) $R(n_{IS}) = \alpha \times e^{-\beta \times n_{IS}}$

4 Testing the model

4.1 Has redundancy introduced new biases into the system?

To test whether redundancy introduced new biases into the system, I ran several tests.

4.1.1 Bases distribution

First of all, I tested whether the distribution between the 4 bases was uniform, since no bias is expected. To this end, I ran 10 experiments with different seeds but the same parameters[7] for 10 000 generations, and took the mean bases distribution of each population (which are almost clonal populations) to compute the confidence interval around the mean, with a 95% confidence level $I = [\mu - 2 \times \frac{\sigma}{\sqrt{n}}; \mu + 2 \times \frac{\sigma}{\sqrt{n}}]$. We obtained the following results :

Base	lower bound	upper bound
0	0.24583	0.25567
1	0.24688	0.25628
2	0.24546	0.25397
3	0.24281	0.25307

The intervals are quite broad due to the low number of "individuals" (each population is here counted as a single individual), but they are reasonable and all contain the unbiased value (0.25). Hence we cannot reject the hypothesis of the four rates being equal to 0.25 (with an error rate of 0.05)

In addition to that, we would like to verify that the distribution of the 2 bases version is originally uniform too, since that has not been measured previously. Therefore, we ran 10 experiments of the 2 bases AEVOL with the same seeds and parameters as previously and used them to compute the confidence interval around the mean, with a 95% confidence level. We obtained the following results :

Base	lower bound	upper bound
0	0.49680	0.50046
1	0.49953	0.50319

We cannot reject the hypothesis of the bases distribution being different from 0.5, which confirms our hypothesis: the four bases version of AEVOL did not introduce new biases in the bases distribution.

[7] population $= 1024$; mutation rate $= 10^{-5}$

4.1.2 Genes strand distribution

In most cases in AEVOL, we compare if a base is equal to 1 or 0 on the one strand and is different from that on the other one, which would lead to severe biases in the distribution of genes in case a test has been forgotten while changing for 4 bases. This is why I decided to test whether the distribution of genes between the two DNA strands is significatively different between the 4 bases and the 2 bases AEVOL. To run this test, I developed a new post-processing program to extract the number of genes on each stand for each individual of the population (`genes_count.cpp`).

I treated the data with a Mann-Whitney U test to verify whether the distribution is significatively different between the two versions and obtained a p-value of 0.40567.

There is thus no significant difference between the two versions of the software in the distribution of genes or bases.

4.2 Is redundancy necessary and sufficient to explain the dynamics of IS?

The 4 bases model was developped to enable target sites to be counter selected in essential genes. To verify whether this prediction is correct, we compare the proportion of target sites in genes for different transposition rates, regarding the total number of target sites and the proportion of bases of the genome included in genes. In fact, we could expect the selection against target sites in genes to be higher as the transposition rate increases. To this end, I developped another postprcessing program (`locate_target_IS.cpp`). To find and count target site we have to read the entire genome, which is very costly in time and cannot be done for all individuals at each generation. A post-processing program allows us to do this once the run is ended, or for each back-up of any experiment.

Here, we tested the proportion of target sites in genes for 384 populations with various ranges of parameters. In particular, 6 different transposition rates were tested with two different functional forms (see Fig.9).

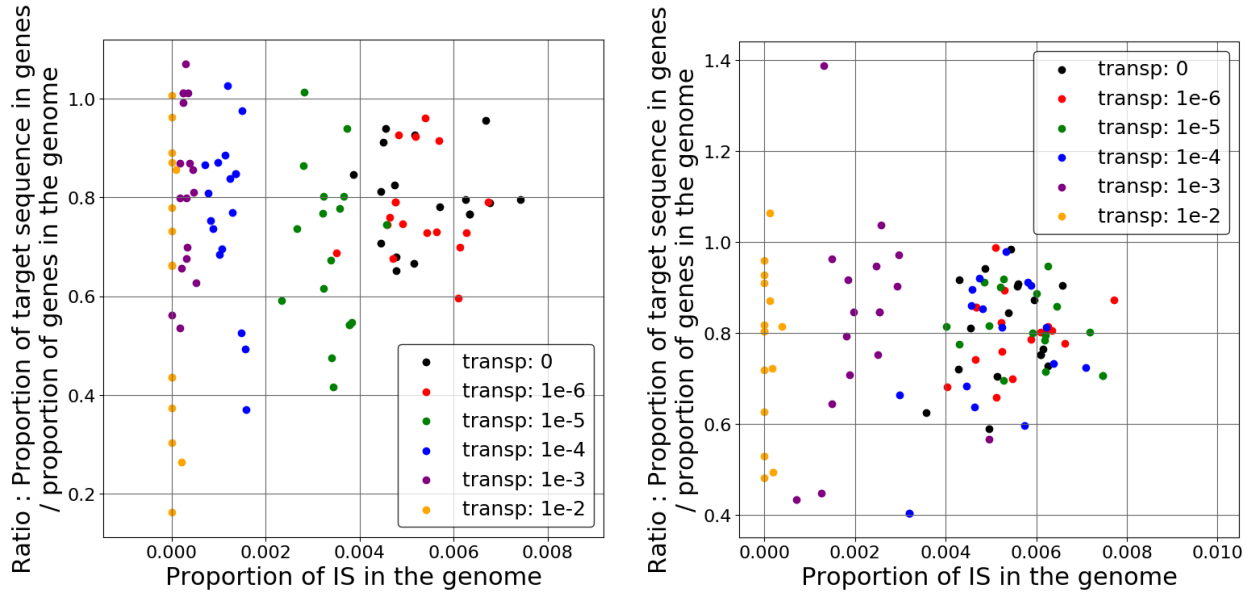

Figure 9: Ratio between the proportion of target sites in genes and the proportion of genes in the genome after 200 000 generations regarding the proportion of IS in the genome for the linear (left) and constant forme (right) experiments.

There is no significant correlation between the proportion of target sites in genes and the proportion of IS in the genome, but it appears that target sites or sometimes strongly counter-selected. This occurs only when the transposition rate is high (broader point dispersion). This suggests that it is not usefull to counter-selected target before IS become a real danger to the genome. However, once there are no more IS in the genome there is no need to counter-select target sites, which would explain why some ratio are so high even if a transposition rate of 10^{-2}.

We cannot conclude towards the dynamic of target sites counter selection, but it seems here that the redundancy of the genetic code does have an impact on it.

5 Exploring IS properties thanks to Aevol

The first 10 000 generations in Aevol are hardly comparable to what one can commonly observe in biology: the genome size explodes and the fitness increases rapidly. Then, the fitness continues to increase sporadically while the genome size decreases. To test the reaction of a population to a particular parameter, it is thus common to let evolve populations for a large number of generations before testing the wanted parameter. Here, we use 2 different populations that have previously evolved in the presence of IS of size 12, and mutation rate of 5.10^{-6} and a constant transposition rate of 10^{-6} (low enough for IS to not be selected against) for 100 000 generations. Such populations are called "wild types". At this stage, the fitness has reach a quasi stationary state around 10^{-2}.

5.1 Are IS strongly counter-selected?

At 100 000 generations, 8 different transposition rates were introduced: from no transposition at all up to a rate of 10^{-1}. Each rate was combined with a functional form (constant or positively linear) and each combination was repeated four times for each wild type.

The results (see Fig.15) clearly show that the proportion of IS decreases as the transposition rate increases, from 0 to 10^{-2}. This descent intervenes at a lower transposition rate for the linear form, which is logical since more transpositions occur under this condition.

This tends to prove that IS are strongly counter-selected as soon as they have a real impact. We can separate the curve into three domains. In the first one (transposition rate lower than 10^{-5} for the linear form and lower than 10^{-3} for the constant form), transposition seems to have no impact since the proportion of IS is similar to the one when there is no transposition at all.

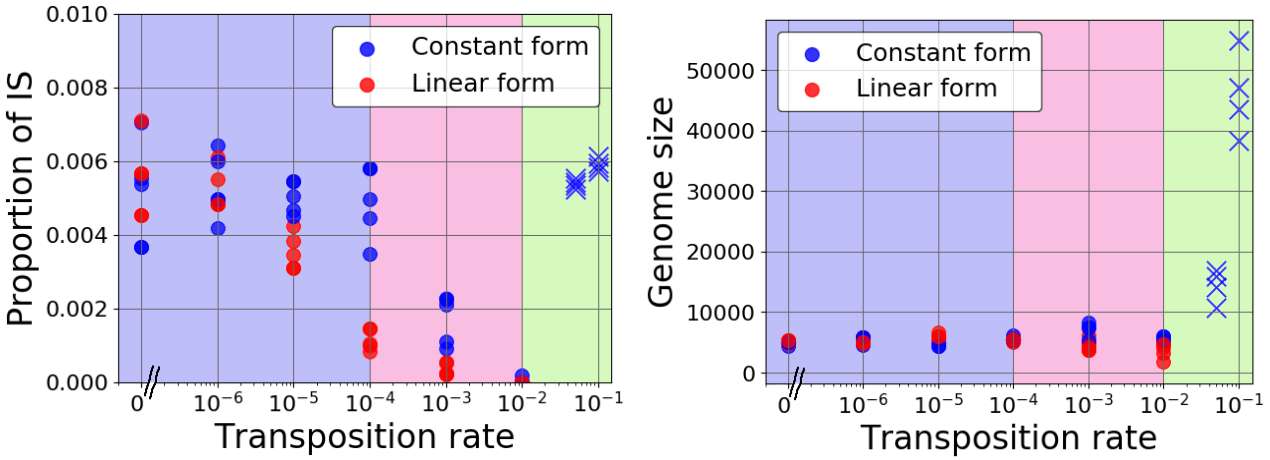

Figure 10: Proportion of IS, genome size as function of transposition rate. All experiments began with wild types. For rates from 0 to 10^{-2}, values are means over the 1000 last generations in run of 100 000 generations. For the transposition rate 5.10^{-2}, values are the mean over the last 10 generations in a run of 70 generations and for the transposition rate 10^{-1}, values are the mean over the last 10 generations in a run of 30 generations, due to the very high computing time as genome size explodes.

Once we reach a higher rate, IS are counter-selected and their activity is maintained at a low rate, their proportion decreasing as the transposition rate increases. This means that their activity is deleterious to the individuals. However, individuals are able to maintain their fitness high, meaning they are adapting themselves to resist transposition.

Finally, there is a threshold of transposition rate over which genomes are not resilient enough to resist IS invasion and we can observe that although the proportion of IS in the genome remains globally the same as with no transposition at all, the genome size explodes (making it very hard to compute, and thus experiments have to be stopped early), and the fitness decreases drastically (see Appendix 2). Thus, we can conclude that above a threshold transposition rate, individuals cannot control tranposition, which is highly deleterious.

We can conclude here that to be present in the genome of bacteria, IS must have a transposition rate low enough in order to not be too deleterious and thus not to be strongly counter selected, contrary to the intuition that a higher rate would allow them to maintain themselves in the genome. However, it is difficult to conclude regarding very high transposition rates, since we cannot observe long term dynamics due to computing time limitations.

5.2 Do we observe IS bursts?

In the above section, we conclude that when IS have an activity, they are strongly repressed by the cell and their proportion in the genome reaches almost 0 after 100 000 genertions. Since the experiments in which IS disappear are probably the ones in which they have a real activity (and probably a deleterious one), we decided to have a closer look on them.

By following their fitness, genome size, proportion of IS and number of transpositions at each generation across time, I spotted several tendencies: the fall in the IS proportion generally intervenes in the first 20 000 generations. Naturally, it is faster when the transposition rate is linear with the number of IS than when it is constant.

On some experiments, a detail was striking: temporary increases in the number of transpositions per generation in the population seem to be temporally correlated with a sudden increase (or decrease in 1 case) of the fitness (see Fig.11 for a representative example). To see whether one of the event occured before the other, I looked more precisely at what happened during the few thousands generations concerned (see Fig.12). It seems that the increase of transposition events occurs while the mean fitness increase after an outbreak in the fitness of the best individual, but it is hard to conclude on the mechanisms behind this

behaviour. By looking more precisely at the statistics (see Appendix 3), we can see that the first fitness increase is precisely correlated with the appearance of an IS in the best individual. However, analysing more precisely the relationship between IS dynamics and fitness increases would require the development of new post-processing programs (for example to know whether the best of a generation is issued from the best of the last generation) that are out of the scope of this internship.

5.3 Is a changing environnement beneficial for IS?

At 100 000 generations, the two wild types were also exposed to two different environnemental conditions: either they were kept in the same environnement as before, or they were put into a new one. Both experiments were repeated with the same parameters as in the previous subsection.

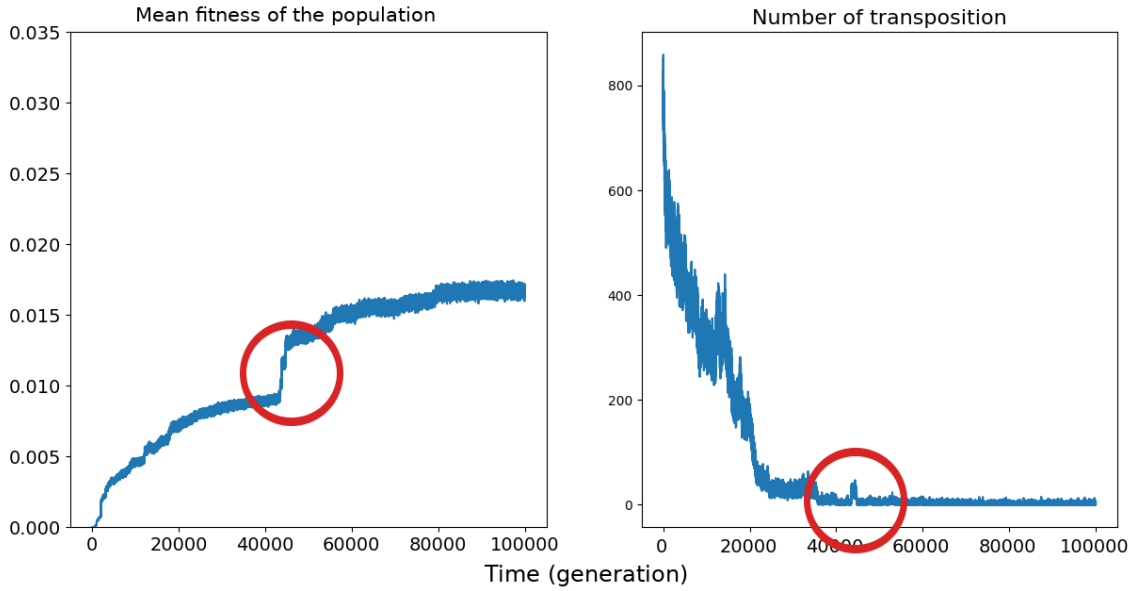

Figure 11: Representative example of a time correlation bewteen fitness increase and transpositions increase

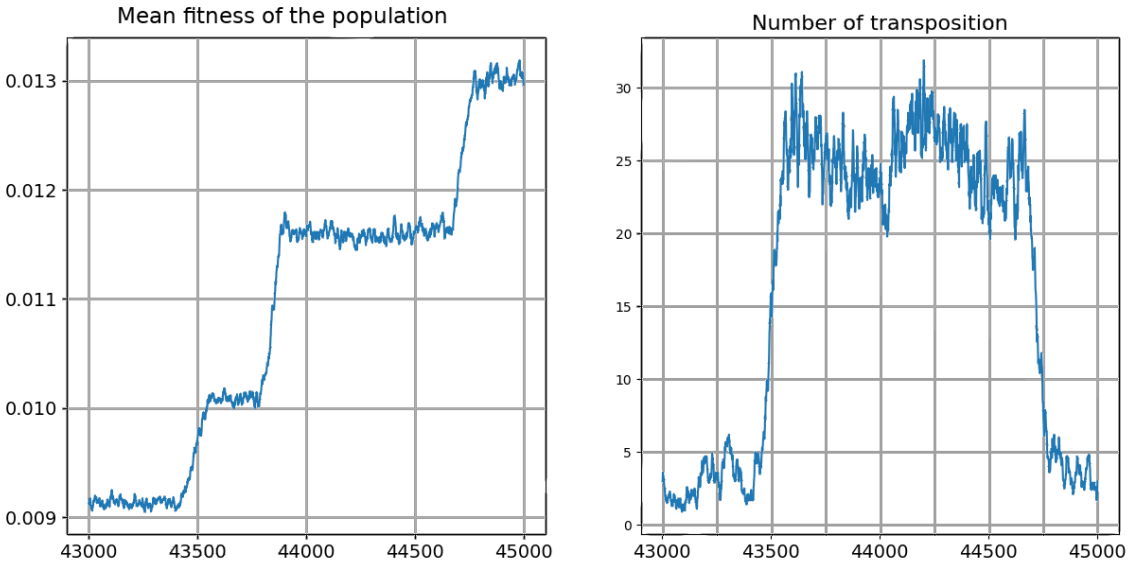

Figure 12: Zoom for generations 43000 to 45000 of the increase in transpositions and in fitness

It seemed that at high transposition rates, IS tended to be maintenained in the host genome for a longer time when the population was subject to an environnemental change, but the data did not suffice to verify that hypothesis. I therefore tested another feature of AEVOL, which randomly changes the environment across time. Due to a lack of time and some corrupted data, I cannot provide statistics on the observed differences, but through the example of Fig.13 it is clear that while IS are quickly repressed when the environment is stable, they are maintaiend at a relatively high proportion (around 0.5%) when the environment is regularly changing. This indicates that when the environment is changing, IS are not counter-selected, while this is the case in a constant environment. Further analysis are required to understand the cause of this activity and its possible advantage.

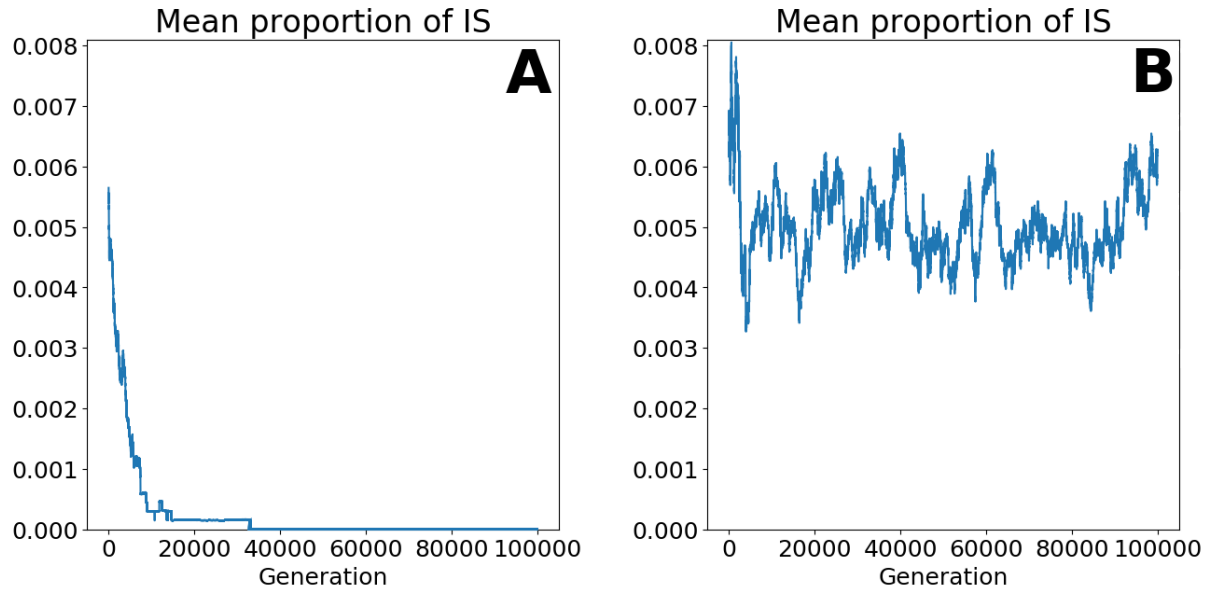

Figure 13: Representative example for the proportion of IS in a stable, where IS disappear, (A) or changing environment, where IS are maintained (B) with a constant transposition rate of 10^{-2} transposition/IS/generation

Conclusion and perspectives

Scientific conclusion

Throughout the internship, I successfully developed a prototype of IS management from the algorithmical modeling to the implementation in the artificial life software AEVOL and its testing. Different interesting dynamics have been observed, hence proving that this could be further studied thanks to *in silico* experiments.

For now, there remains a bug that corrupts experiments when they are stopped while running: the backup files do not behave correctly. Due to that, my experiments were run without interruption, but this remains to be improved in order for the IS modeling to be included in the full version of AEVOL and to run more demanding experiments.

We observe that there are two thresholds in the behaviour of IS, which change according to the functional form of the tranposition rate regarding the number of IS. Under the first threshold, IS have little to no activity and are therefore not strongly regulated by the individuals. Between the two threshold, IS are strongly counter-selected, their rate being maintained very low. This suggests that their activity is highly deleterious. Above the second transposition rate threshold, IS totally disrupt the genome, making it impossible to compute. We cannot therefort predict whether they would disrupt the genome until the death of the individuals, or whether a domestication would finally occur.

Perspectives

To complete this model, it would be crucial to take into account homologous sequences for genomic recombination. For now, genomic rearrangements occur at random positions in the genome, but it is known

that they mostly occur between homologous sequences in Biology[8]. There is a functional version of AEVOL taking that parameter into account[8], but for now it has not been tested with IS. Indeed, since IS are repeated sequences spread accross the genome, they might have a major impact on those kinds of genomic rearrangements.

Another interesting perspective would be to test the impact of IS and their prevalence in the population when the environment is changing across time. It was shown here that a changing environment could maintain IS in the population when there would be otherwise counter-selected, but to unveil a robust dynamic it would be necessery to test various intensities of environmental changes, with various transposition rates.

Last but not least, I was not able to decipher the causal effect of the correlation between fitness and transposition bursts, but there might be a fascinating dynamic to uncover here. To test this more precisely, it would be helpful to combine the version of AEVOL with IS with a version enabling us to follow the lineage of the individuals. In fact, knowing what mutations occured when to which lineage and what will be the fate of this lineage is crucial to study more precisely this dynamic.

Personal conclusion

This internship allowed me to follow a project from its very beginning, by picturing a model for IS transposition and bibliographical searches, to the end, by implementing it, testing it and even beginning to run experiments on it. Therefore, it was really fascinating and it gave me a quick overview of what a full research project can be.

Moreover, I had the opportunity to be a volunteer worker for the international conference "Mathematical Models in Ecology and Evolution" (MMEE) which took place in Lyon from 16th to 19th, July. This experience was also very instructive and it revealed to me another part of what scientific research is.

Therefore, I would like to thank sincerely and wholeheartedly the whole Beagle team for giving me time and entertaining discussions. In particular, I am gratefull to Guillaume BESLON for welcoming me in the team and always taking the time to discuss my results and to Jonathan ROUZAUD-CORNABAS for providing me help on C++ and experiments management.

[8] During cell division (mitosis), chromosomes are highly condensed and, through base complementary, homologous sequences tend to be stuck to each other. Strand breaks and repairs can occur at that moment

References

[1] M. Bichsel, A. D. Barbour & A. Wagner: *The early phase of a bacterial insertion sequence infection*. Theoretical Population Biology, vol. 78, no. 4:278–288, 2010. doi:10.1016/j.tpb.2010.08.003.

[2] M. Bichsel, A. D. Barbour & A. Wagner: *Estimating the fitness effect of an insertion sequence*. Journal of Mathematical Biology, vol. 66, no. 1-2:95–114, 2013. doi:10.1007/s00285-012-0504-2.

[3] M. Chandler & J. Mahillon: *Insertion Sequences Revisited*. In N. L. Craig, A. M. Lambowitz, R. Craigie & M. Gellert, editors, *Mobile DNA II*, pp. 305–366. American Society of Microbiology, 2002. doi:10.1128/9781555817954.ch15.

[4] J. Frey: *Insertion Sequence Analysis*. In R. Miles & R. Nicholas, editors, *Mycoplasma Protocols*, pp. 197–205. Humana Press, Totowa, NJ, 1998. doi:10.1385/0-89603-525-5:197.

[5] H. Lee, T. G. Doak, E. Popodi, P. L. Foster & H. Tang: *Insertion sequence-caused large-scale rearrangements in the genome of Escherichia coli*. Nucleic Acids Research, p. gkw647, 2016. doi:10.1093/nar/gkw647.

[6] V. Liard, D. Parsons, J. Rouzaud-Cornabas & G. Beslon: *The Complexity Ratchet: Stronger than selection, weaker than robustness*. In *The 2018 Conference on Artificial Life*. MIT Press, Tokyo, Japan, 2018. doi:10.1162/isal_a_00051.

[7] D. Parsons, C. Knibbe & G. Beslon: *Aevol : un modèle individu-centré pour l'étude de la structuration des génomes*. p. 8.

[8] D. P. Parsons, C. Knibbe & G. Beslon: *Homologous and nonhomologous rearrangements: Interactions and effects on evolvability*. In *in "European Conference on Artificial Life (ECAL*, pp. 622–629. MIT Press, 2011.

[9] S. A. Sawyer, D. E. Dykhuizen, R. F. DuBose, L. Green, T. Mutangadura-Mhlanga, D. F. Wolczyk & D. L. Hartl: *Distribution and Abundance of Insertion Sequences Among Natural Isolates of Escherichia coli*. p. 13, 1986.

[10] D. Schneider & R. E. Lenski: *Dynamics of insertion sequence elements during experimental evolution of bacteria*. Research in Microbiology, vol. 155, no. 5:319–327, 2004. doi:10.1016/j.resmic.2003.12.008.

[11] T. Sultana, A. Zamborlini, G. Cristofari & P. Lesage: *Integration site selection by retroviruses and transposable elements in eukaryotes*. Nature Reviews Genetics, vol. 18, no. 5:292–308, 2017. doi:10.1038/nrg.2017.7.

[12] O. Tenaillon, J. E. Barrick, N. Ribeck, D. E. Deatherage, J. L. Blanchard, A. Dasgupta, G. C. Wu, S. Wielgoss, S. Cruveiller, C. Médigue, D. Schneider & R. E. Lenski: *Tempo and mode of genome evolution in a 50,000-generation experiment*. p. 30, 2017.

[13] R. Tobes & E. Pareja: *Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements*. BMC Genomics, vol. 7, no. 1:62, 2006. doi:10.1186/1471-2164-7-62.

[14] M. Touchon, L.-M. Bobay & E. P. Rocha: *The chromosomal accommodation and domestication of mobile genetic elements*. Current Opinion in Microbiology, vol. 22:22–29, 2014. doi:10.1016/j.mib.2014.09.010.

6 Appendix

6.1 Parameters calibration

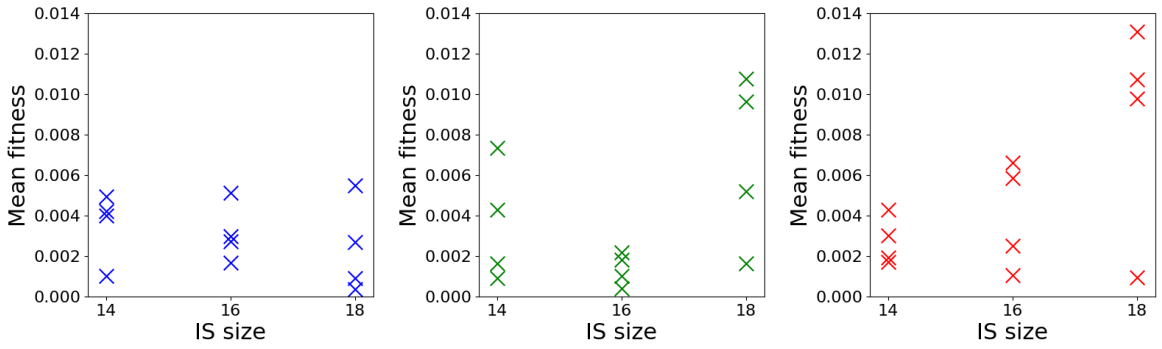

Figure 14: Fitness (mean over the last 1000 generations) for the 3 IS sizes and 3 transposition rates tested. The mutation rate is 5.10^{-6} *mutation/base/generation.*

6.2 Are IS strongly counter-selected?

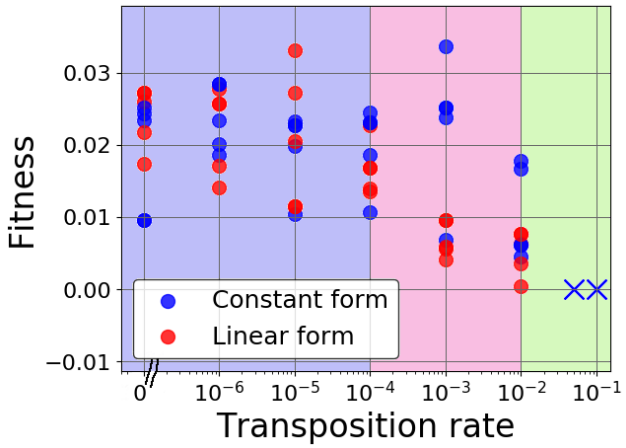

Figure 15: Fitness as function of transposition rate. All experiments began with wild types. For rates from 0 to 10^{-2}*, values are means over the 1000 last generations in run of 100 000 generations. For the transposition rate* 5.10^{-2}*, values are the mean over the last 10 generations in a run of 70 generations and for the transposition rate* 10^{-1}*, values are the mean over the last 10 generations in a run of 30 generations, due to the very high computing time as genome size explodes.*

6.3 Data around an IS bursts

Generation	Fitness (best)	number of IS (best)
43390	1.101498e-02	0
43391	1.101498e-02	0
43392	1.101498e-02	0
43393	1.101498e-02	0
43394	1.101498e-02	0
43395	1.101498e-02	0
43396	1.101498e-02	0
43397	1.101498e-02	0
43398	1.101498e-02	0
43399	1.101498e-02	0
43400	1.101498e-02	0
43401	1.101498e-02	0
43402	1.101498e-02	0
43403	1.101498e-02	0
43404	1.101498e-02	0
43405	1.216381e-02	1
43406	1.216381e-02	1
43407	1.216381e-02	1
43408	1.216381e-02	1
43409	1.216381e-02	1
43410	1.216381e-02	1
43411	1.216381e-02	1
43412	1.216381e-02	1
43413	1.216381e-02	1
43414	1.216381e-02	1
43415	1.216381e-02	1
43416	1.216381e-02	1
43417	1.216381e-02	1
43418	1.216381e-02	1
43419	1.216381e-02	1
43420	1.216381e-02	1

Statistics recorded for the shown experiment (constant functional form, transposition rate 1e-2, wild type 1, stable environment, seed 75158785)