Statistical study of the length of transcription units of Escerichia Coli and Bacillus Subtilis

Nicolas Omont

20th June 2003

Abstract In this study, we show evidence that the biases in distribution in length and orientation of transcription units on the chromosomes of Escherichia Coli and Bacillus Subtilis are statistically relevent. Then, we build the basis of a model useful to understand the origins of these biases. This model makes the hypothesis that the critical factor for the bias is the proportion of transcript interrupted for each transcription unit. Its validity is confirmed, among other things, by the estimation of the replication to transcription speed ratio in Escherichia Coli. Its main prediction is that a large replication to transcription speed ratio implies a low bias. The observations are consistent with a model in which codirectionnal collisions cost as much as head-on ones.

Résumé

1 Introduction

1.1 State of knowledge

Escherichia Coli has one circular chromosom. Its replication starts from a single locus called OriC [4][3]. The 2 replication forks formed by the replication complexes progress along the 2 half chromosomes and finish in the termination region, to which belong Ter locuses [1] [2] and more precisely at dif locus [5]. It is nearly diametrally opposed to OriC.

Bacillus Subtilis has also a single origin of replication and a termination region, even if the mechanisms are different from Escherichia Coli [1], however the half chromosomes defined are not of the same length.

During the process of replication, the transcription continues. In Escherichia Coli, the transcription rate in vivo is around 42 nucleotides per second [7] whereas the replication rate is estimated between 600 and 1000 nucleotides per second [8], i.e. 14 to 24 times faster than the transcription rate. In Saccharomyces cerevisiae, this ratio is estimated to be larger than 5 [6], and it is believed this is the case for the vast majority of organisms Therefore, there are collisions between transcriptases and replicases.

These collisions have been studied in vivo in Escherichia Coli by S. French in [10]. The experimental model consists in inserting a replication origin near a highly transcribed region of the chromosome of Escherichia Coli either upstream or downstream rrnB, a ribosomal RNA, probably one of the most active transcription unit of the chromosome. For codirectionnal collisions, it is observed that the replication fork goes through the transcription unit at a speed close from the one it would have without these collisions and that no transcription complexes are visible within 2000 to 3000 base pairs behind the fork on the DNA. For head-on collisions, it is observed that the progression of the replication fork is slowed in the transcription unit region and that no transcription complexes are visible behind the fork on the DNA. The interpretation is that the replication complex dislodges transcription complexes in both direction and that its speed is strongly reduced in case of head-on collisions.

Other studies on different models give different results.

M. Krasilnikova et al. studied the (in vivo) effects of $d(G)_n.d(C)_n$ repeats on a plasmid in Escherichia Coli [11]. They observe that these repeats block replication if they are in transcribed regions. The interpretation is that they block transcription complexes, and that these stalled transcription complexes block replication in either direction.

B. Liu et al. used an in vitro model based on bacteriophage T4 DNA replication apparatus with Escherichia Coli RNA polymerase [12] [13]. In [12], they observe that the codirectional collisions with with stalled or moving transcription

complexes do not slow replication and that transcription at least not always interrupted. In [13], they observe that head-on collisions take twice as much time to be solved as codirectional ones, that transcription is at least not always interrupted and that the transcription complex switches strand to use the newly synthesized strand as template. It also observed that DNA helicase is needed to solve head-on collisions. The interpretation is that there is a mecanism to solve all collisions.

M. Elias-Arnanz et al. studied (in vivo) the Bacillus Subtilis phage $\Phi 29$. In [14], they observe that, in case of codirectional collisions, a stalled transcription complex stops the replication fork, but when the transcription complex starts moving again, the replication fork does the same, but at a decreased speed. In [15], they observe that, in case of head-on collision, a stalled complex stops the replication fork, but when the transcription complex starts moving again, the replication fork does the same at the normal speed. The interpretation is that, given the fact that the chromosome is linear and that replication can initiate at either end, there is a mechanism to solve collisions between 2 replication forks, and that this mechanism allows to solve head-on collisions but not codirectional ones. Codirectional collision are solved by the normal end of the transcription of the unit.

A.M. Deshpande et al. studied in vivo on Saccharomyces Cerevisiae. In [6], they observed that there are sites that block the progression of the replication fork, that these sites correspond to transcription units transcribed head-on, and that the stalling of replication fork is visible only if transcription units are active. The interpretation is that head on collisions of transcription complexes with the replication fork slow this last one. They estimate that the pause is longer than 3 to 5 seconds.

This experiments show contradictory results, suggesting that collisions are not solved in the same way in the different models. In the last review on collision in Escherichia Coli [19], B.J. Brewer proposes that it has no mechanism to solve collisions.

He said that it was difficult to know whether, in case of codirectional collisions, the transcription was disrupted or the replication slowed. S. French suggests that the transcription is disrupted, but it is not clear in which proportion.

For head-on collisions, he proposes that the transcription complex collides with the helicase, i.e. the protein that unwinds and opens the DNA. He suggests that supercoils might have a role in signaling the transcription complex upstream, but nobody observed such a phenomenon since then, and if the transcription unit is very active, the replication fork will never go through. S. French suggests that the transcription complex is simply dislodged from the DNA, but that it takes more time than for codirectional collisions. As far as activities of transcription units are concerned, it was observed that genes coding for ribosomal proteins were transcribed codirectionally and that large chromosomal inversions are lethal.

This reinforced the idea that there is no mechanism to solve head-on collisions. However all studies are done with very active transcription units, so as to observe many collisions. To conclude, we will make the hypothesis that both codirectional

and head-on collisions are at least partially solved by disruption, but that head-on transcription slow more replication than codirectional ones.

1.2 Difficulty of solving each type of collision

However, we have to explain a point.

- Why don't we conclude that one type of collision easier to solve than the other?

Due to the relative speed of the two enzymes, head-on collisions are more likely to occur than codirectional ones. This may even be responsible for the fact that head-on transcriptions slow replication more than codirectional ones. It is easy to see that if the solving of a collision on given transcription unit takes t seconds and that the transcription initiation rate of this transcription unit is one every t seconds, the mechanism that solves head-on collisions becomes useless, which is probably the case for highly transcribed genes whose inversion is lethal to the bacteria. In a more general way, let's consider:

- t_c be the time needed for a collision to be solved
- r be the rate of transcription initiation (number of transcripts initiated per second)
- L the length of the transcription unit (transcribed head-on)
- V_h^{re} the speed of a replicase progressing freely
- C the number of collisions between transcriptases and a replicase progressing on the operon
- U_h^{re} the mean speed of the replicase progressing on the transcription unit.

We have the following equations:

$$\begin{cases} U_h^{re} &= L \,/\, (L \,/\, V_h^{re} + C t_c) \\ C &= r\, L \,/\, U_h^{re} \end{cases} \tag{1}$$

i.e.:

$$\frac{U_h^{re}}{V_h^{re}} = 1 - r t_c \tag{2}$$

The first remark is that this apparent speed is independent of the length of the operon transcribed. Then let's see if the effect is important. For exemple, a high initiation rate is 2 transcription initiation every second. If the time it takes to solve the collision is 0.25 second, it halves the speed of the replication. It it is less than 0.05 seconds, the slow down is less than 10%. A lower bound for the solving time is the time it takes to process one nucleotide, around 0.001 seconds, only 50 times less. In the absence of other evaluation, we should keep in mind that this effect may play a role in the fact that head-on collisions seem to slow more replication than codirectional ones. We also have to notice that the effect is only relevent when rt_c is near 1, which is the situation that has been experimented, and not the average situation of transcribing units. Therefore, one can make the hypothesis that only very active transcription units slows replication fork more if they are oriented head-on. For the majority of transcription units, this non-linear effect will not be visible.

1.3 Conclusion

As Brewer wrote, we can make the hypothesis that there is an evolutionnary advantage for transcription unit to be transcribed in the direction of the replication. But, with what we said, we have to go deeper in the understanding of this advantage:

- The replication rate and efficiency are higher. The equation 2 and experiments suggests the replication fail if a very active unit is transcribed head-on. However, this effect might not be the more important in terms of transcription unit distribution because it concerns very few units.

- The transcription efficiency is higher. In article to be published [20], E.P.C. Rocha suggests that essentiality more than activity is the leading factor to explain gene strand bias in Escherichia Coli and Bacillus Subtilis. It means that the most important advantage of being transcribed codirectionally is the fact that there will be less truncated RNA produced (proportionally to the total quantity of RNA produced for one gene). This make sense because truncated RNA make incomplete proteins that interfere in the original mechanisms, impeding all the cell functions. He suggests that even if this effect add with the first one, it is the leading one to explain transcription unit distribution.

Thus, we can refine hypothesis about organization of Escherichia Coli an Bacillus Subtilis genomes:

- There should be more transcription units coded in the direction of the replication
- This bias should be almost complete in favor of the codirectional direction for very active transcription units (cf. 2). However this bias should not increase with the activity for the large majority of moderately transcribed units.
- This bias should increase with the length of the transcription units , because the probability of interruption of one initiated transcription is proportional to the length of the transcription unit.

In the first part, the statistical physics model used to represent this 3 hypothesis is presented. In the last part, the model is correlated to Escherichia Coli and Bacillus Subtilis data available and its consistence evaluated.

2 Model

2.1 Variables

The chromosom is modelled by a circular line divided in two halves having potentially different characteristics. We use the index $h \in \{1, 2\}$ to indicate the half chromosome we want to refer to. For example, the first half of the chromosom is a line of length L_1^{tot} (in base pairs), the second of length L_2^{tot}:

$$L_h^{tot}\text{: length of half chromosome } h \tag{3}$$

On each half chromosome transcription and replication complexes travel. These complexes are modelled by points. They have an absciss, which is related to the position on which they are on the half chromosome. There is a collision when the two points have the same absciss. Each class of complex moves on each half chromosome at a given speed.

- V_{tr}^h for transcription complexes.
- V_h^{re} for replicaction complexes.

We will rather use the ratio of the rates:

$$\gamma_h = V_h^{re}/V_{tr}^h \tag{4}$$

Let's consider an operon of length L. We use the index $s \in \{+, -\}$ to indicate whether the variable is related to transcription units transcribed codirectionally or

head-on respectively. For instance, the interruption probabilities of an initiated transcription of operons of length L are noted $P^{int}_{+,h}(L)$ or $P^{int}_{-,h}(L)$.

Then, to model the distribution of transcription units, we make the approximation of a continuous distribution and note the presence probability $\mathrm{d}p_{s,h}(L)$. From the definition, we have:

$$\int_0^{+\infty} \sum_{s,h} \mathrm{d}p_{s,h}(L) = 1 \tag{5}$$

We note β_h the ratio of the fraction of transcription units transcribed codirectionally to the fraction of transcription units transcribed head-on :

$$\beta_h = \frac{\int_0^{+\infty} \mathrm{d}p_{+,h}(L)}{\int_0^{+\infty} \mathrm{d}p_{-,h}(L)} \tag{6}$$

We note η the ratio of the fraction of the transcription units transcribed on the first half chromosome to the fraction for the second one:

$$\eta = \frac{\int_0^{+\infty} \sum_s \mathrm{d}p_{s,1}(L)}{\int_0^{+\infty} \sum_s \mathrm{d}p_{s,2}(L)} \tag{7}$$

We will use the mean length of operons:

$$\theta_{s,h} = \int_0^{+\infty} L\, \mathrm{d}p_{s,h}(L) \tag{8}$$

Finally, to model the statistical equilibrium of operons, we will use the pseudo-energy C_h, which is to be correlated with the evolutionnary cost of a collision and the difficulty to solve it. This coefficient is unitless. In this part, we consider that it only depends on the complexes that collide but not on the direction of the collision.

2.2 Equations

Interruption probability

We consider an unit of length L transcribed head-on. Its transcription takes $t_{tr} = L/V^h_{tr}$. It will be interrupted if the replicase was either near enough from the end of the unit at the beginning of transcription (i.e. if the replicase was either within $L^{int} = V^{re}_h t_{tr}$ of the end of it) or on it at the beginning of replication. Therefore, the probability of interruption of an initiated transcription is proportionnal to the probability of presence of the replicase in this range. This leads to:

$$P^{int}_{-,h}(L) = \frac{L}{L^{tot}_h} \frac{V^{re}_h + V^h_{tr}}{V^h_{tr}} = \frac{L}{L^{tot}_h}(\gamma_h + 1) \tag{9}$$

For an operon transcribed in the direction of replication, the only difference is that, if the replicase is on the operon when the replication begins, the transcription cannot be interrupted because there will never be any collision. This leads to:

$$P^{int}_{+,h}(L) = \frac{L}{L^{tot}_h}\frac{V^{re}_h - V^h_{tr}}{V^h_{tr}} = \frac{L}{L^{tot}_h}(\gamma_h - 1) \tag{10}$$

Boltzmann-like equilibrium

We make the hypothesis that the equilibrium between all operons follows a Boltzmann-like law according to this pseudo-energy definition: L/L^{tot}_h as:

$$E_{s,h} \;=\; P^{int}_{s,h}(L)C_h \tag{11}$$

The mathematical formulation of the equilibrium is:

$$\mathrm{d}p_{s,h}(L) \;=\; Z^{-1}\exp\left(-E_{s,h}\right)\,\mathrm{d}L \tag{12}$$

In other words:

$$\begin{cases} \mathrm{d}p_{+,h}(L) &= Z^{-1}\exp\left(-C_h(\gamma_h-1)L/L^{tot}_h\right)\,\mathrm{d}L \\ \mathrm{d}p_{-,h}(L) &= Z^{-1}\exp\left(-C_h(\gamma_h+1)L/L^{tot}_h\right)\,\mathrm{d}L \end{cases} \tag{13}$$

As the parameter of an exponential law is the mean of the variable, we have:

$$\begin{cases} \theta_{+,h} &= L^{tot}_h\left(C_h(\gamma_h-1)\right)^{-1} \\ \theta_{-,h} &= L^{tot}_h\left(C_h(\gamma_h+1)\right)^{-1} \end{cases} \tag{14}$$

From 5, we have the normalisation constant Z:

$$\begin{aligned} Z &= \sum_{s,h}\theta_{s,h} \\ &= C_h\sum_h L^{tot}_h\frac{2\gamma_h}{{\gamma_h}^2-1} \end{aligned} \tag{15}$$

And, from 6:

$$R_h = \frac{\beta_h}{\theta_{+,h}/\theta_{-,h}} = 1 \tag{16}$$

We also have the usefull two-state boltzmann equilibrium for a pool of operons of the same length L:

$$\frac{\mathrm{d}p_{+,h}(L)}{\mathrm{d}p_{-,h}(L)} = \exp\left(\left(\frac{1}{\theta_{-,h}} - \frac{1}{\theta_{+,h}}\right)\frac{L}{L^{tot}_h}\right) = \exp\left(2\frac{C_h\,L}{L^{tot}_h}\right) \tag{17}$$

Finally, the definition 7 gives the bias in quantity η between the two half chromosomes:

$$\eta = \frac{L_1^{tot} \sum_s {\theta_{s,1}}^{-1}}{L_2^{tot} \sum_s {\theta_{s,2}}^{-1}} \tag{18}$$

We define the ratio associated:

$$Q = \eta \frac{L_2^{tot} \sum_s {\theta_{s,2}}^{-1}}{L_1^{tot} \sum_s {\theta_{s,1}}^{-1}} = 1 \tag{19}$$

2.3 Estimators

We define $N_{s,h}$:

$N_{s,h}$: Number of operons transcribed in direction s on half chromosome h (20)

Estimation of β_h

A simple estimator of β_h is:

$$\widehat{\beta_h} = \frac{N_{+,h}}{N_{-,h}} \tag{21}$$

We can easily draw a confidence interval for $\widehat{\beta_h}$ because $N_{+,h}/(N_{+,h}+N_{-,h})$ is the is the mean of a binomial variable. Therefore, the following distribution is asymptotically a centered reduced gaussian:

$$\sqrt{\sum_s N_{s,h}} \left(\left(\frac{\widehat{N_{+,h}}}{N_{+,h}+N_{-,h}} \right) - \frac{N_{+,h}}{N_{+,h}+N_{-,h}} \right) \rightarrow \mathcal{N}\left(0, \frac{N_{+,h}N_{-,h}}{(N_{+,h}+N_{-,h})^2}\right) \tag{22}$$

As

$$\frac{N_{+,h}}{N_{+,h}+N_{-,h}} = \frac{\beta_h}{\beta_h+1} \tag{23}$$

the δ-method gives:

$$\sqrt{\sum_s N_{s,h}} \left(\widehat{\beta_h} - \beta_h \right) \rightarrow \mathcal{N}(0, \beta_h) \tag{24}$$

Finally, we have approximately:

$$P\left(|\beta_h - \widehat{\beta_h}| < \frac{\widehat{\beta_h}}{\sqrt{\sum_s N_{s,h}}} \right) = 0.84 \tag{25}$$

$$\tag{26}$$

Estimator of η

A simple estimator of η is

$$\widehat{\eta} = \frac{\sum_s N_{s,1}}{\sum_s N_{s,2}} \tag{27}$$

$\sum_s N_{s,h}$ is also a binomial variable. Therefore:

$$P\left(|\eta - \widehat{\eta}| < \frac{\widehat{\eta}}{\sqrt{\sum_{s,h} N_{s,h}}}\right) = 0.84 \tag{28}$$

$$\tag{29}$$

Estimation of $\theta_{s,h}$

A simple estimator of $\theta_{+,h}$ is the mean of the length of operons transcribed in the direction s on the half chromosome h:

$$\widehat{\theta_{s,h}} = N_{s,h}^{-1} \sum L \tag{30}$$

The following distribution is asymptotically a centered reduced gaussian:

$$\sqrt{N_{s,h}}\left(\widehat{\theta_{s,h}} - \theta_{s,h}\right) \to \mathcal{N}(0, \theta_{s,h}^2) \tag{31}$$

Therefore, we approximately have:

$$P\left(|\theta_{s,h} - \widehat{\theta_{s,h}}| < \frac{\widehat{\theta_{s,h}}}{\sqrt{N_{s,h}}}\right) = 0.84 \tag{32}$$

$$\tag{33}$$

Secondary parameters estimators

These estimators are secondary because they rely on estimators already defined and not directly on data. From 14, we deduce the speed ratio and the evolutionnary cost.

$$\widehat{\gamma_h} = \frac{\widehat{\theta_{+,h}}/\widehat{\theta_{-,h}}}{\widehat{\theta_{+,h}}/\widehat{\theta_{-,h}} - 1} \tag{34}$$

$$\widehat{C_h} = \frac{1}{2}\left(\frac{1}{\widehat{\theta_{-,h}}} - \frac{1}{\widehat{\theta_{+,h}}}\right) \tag{35}$$

Finally, from 16 and 18, the two estimations of the conservation relations are:

$$\widehat{R_h} = \frac{\widehat{\beta_h}}{\widehat{\theta_{+,h}}/\widehat{\theta_{-,h}}} \tag{36}$$

$$\widehat{Q} = \widehat{\eta}\frac{L_2^{tot}\sum_s \widehat{\theta_{s,2}}^{-1}}{L_1^{tot}\sum_s \widehat{\theta_{s,1}}^{-1}} \tag{37}$$

Presence density estimator

We want to estimate the presence density $\mathrm{d}p_{s,h}(L)$. As it is a continuous variable, an approximation must be done. In fact, we will compute an estimator of:

$$\begin{aligned} a(L) &= \frac{1}{W}\int_L^{L+W}\frac{1}{\theta_{s,h}}\mathrm{d}p_{s,h}(L) \\ &= \exp\left(-L/\theta_{s,h}\right)\left(1-\exp\left(\frac{W}{\theta_{s,h}}\right)\right)\frac{1}{W} \\ &\approx \mathrm{d}p_{s,h}(L) \text{ if } W \ll \theta_{s,h} \end{aligned} \tag{38}$$

If i is a transcription unit, such an estimator is:

$$\widehat{\mathrm{d}p_{s,h}(L)} = \frac{1}{W\sum_{s_0,h_0}N_{s_0,h_0}}\sum_i \mathbb{1}_{L(o)\in[L,L+W[} \tag{39}$$

This estimator is also -to a constant- the estimator of the mean of binomial variable $\mathbb{1}_{L(i)\in[L,L+W[}$ which is of parameter $Wa(L)$. Therefore we have the following asymptotic distribution:

$$\sqrt{N_{s,h}}\left(\widehat{a(L)}-a(L)\right) \to \mathcal{N}\left(0, Wa(L)(1-Wa(L))\left(\frac{N_{s,h}}{\sum_{s_0,h_0}N_{s_0,h_0}W}\right)^2\right) \tag{40}$$

3 Results

3.1 Materials and methods

Escherichia Coli

The prediction of transcription units on the complete genome of E. Coli K12 comes from [17], itself referring to [16]. It gives predictions for 2 328 transcription units which is less than the 2 758 announced in the article. This transcription units account for 3 320 ORFs and there are 4 290 ORFs in the genome of Escherichia

Coli. It means that there are no prevision for 960 ORFs. This set of transcription units comprises 75% of known operons, but there are also errors coming from the litterature in the remaining 25%. With ad hoc scripts, we compute the length of transcription units as the distance between the end of the last ORF and the beginning of the first ORF. Due to gene names problems, we only have the length of $2\,283$ transcription units.

The origin of replication is OriC, in the ORF mioC near position $3\,923\,640$. We consider that the replication usually stops at dif near position $1\,586\,959$, between genes b1505 and b1506. Hopefully, the prediction is that they are not in the same transcription unit. The complete chromosom is $4\,639\,221$ base pair long. From this and from orientations of ORFs, we compute whether transcription units are transcribed codirectionally or head-on.

Bacillus Subtilis

To predict transcription units in Bacillus Subtilis complete genome, we use direction of transcription and transcription terminators from [18]. Two neighbouring ORFs are considered as being in different transcription units if they are transcribed in opposite direction or if they are separated by a transcription terminator. In the dataset we used, we have to choose a threshold to discriminate between real and false terminators. As we do not want to put the emphasize on selectivity rather than on sensitivity, we choose a threshold of 0.5, which maximizes accuracy.

To have an idea of the correctness of the prediction, we compare the distribution of length of transcription units for the two bacteria. The result is in graph 1. We see that the distributions are identical except that the mean length of transcription units in Bacillus Subtilis is 1.29 longuer than in Escherichia Coli.

The origin of replication in Bacillus Subtilis is near position 1 and the first half chromosom ends in $2\,023\,105$. The genome is $4\,214\,630$ base pair long. From this, we compute whether transcription units are transcribed codirectionally or head-on.

Model

All calculations and graphs are done with Scilab scripts.

3.2 Estimators

We call half chromosome 1 of Bacteria Subtilis the part of its chromosome going from $2\,023\,105$ to $4\,214\,630$ and half chromosome 2 the other part, which is shorter than the first one. We call half chromosome 1 of Escherichia Coli the part of its

chromosome going from $3\,923\,640$ to $1\,586\,959$ and half chromosome 2 the other part

The results are presented in table 1 to 4. The tables 1 and 2 present the parameters that characterize the distribution for each bacteria. We see that it is impossible to distinguish between the 2 half chromosomes of Escherichia Coli, therefore we will consider Escherichia Coli as one unique dataset so as to enhance quality of estimators.

The tables 3 and 4 present estimation of parameters that make sense only in the model. Apart from η, they are all computed from parameters of the first 2 tables.

The confidence intervalls in the tables 1 and 2 are computed for a deviation of 1.15σ, i.e. a confidence of 90%. In the last 2 tables, the confidence intervalls are based on the worst possible value of the initial parameters, i.e. a covariance of ± 1 between the initial parameters. A lower bound for the confidence is 0.90^n where n is the number of parameters that are used to compute the value.

3.3 Distributions

It is also important to look at the distribution to see if the model fits well the data. For this, we use two tools: First the quantiles, and then the approximation of $\mathrm{d}p_{s,h}(L)$ defined in 38. The quantiles are useful to see the tail of the distribution, while the approximation of $\mathrm{d}p_{s,h}(L)$ is useful to see the beginning of the distribution.

The graphs 2, 3, and 4 represent the quantile of the distribution function of the quantile of a normal exponential distribution for the 3 consistent set of data: Escherichia Coli complete chromosome and Bacillus Subtilis half chromosome 1 and 2. Each set of data is itself divided into transcription units transcribed codirectionally and head-on.

We chose to draw the quantiles function of an exponential distribution because, from 12 we expect the experimental distribution to be exponential, so that the experimental curb should be a line of slope $\theta_{s,h}$. The expected lines are also drawn on the graphs.

We see on the quantile graphs that it is difficult to see what happen near the origin, that is why we estimate the presence probability. The graphs present the lower bound and the upper bound for the presence probability for each of the consistent dataset. The confidence interval is 84%. The exponential drawn is the exponential expected from the model. The corresponding graphs are 5, 6, 7.

Finally, we compute the ratio between our two estimators, which shows the local bias in quantity, i.e. a ratio higher than 1 indicates that more than one half of

the operons of length between L and $L+W$ are transcribed codirectionally. The graph 8 is drawn with the following equation:

$$y = f(L) \quad = \quad \log_{10}\left(\frac{\widehat{\mathrm{d}p_{+,h}(L)}}{\widehat{\mathrm{d}p_{+,h}(L)}}\right) \tag{41}$$

From 17 we expect the experimental data to follow a line of slope $2C_h$. These expected lines are drawn on the graph. To give an idea of confidence intervalls for this graph, one can have a look at graphs 9, 10, and 11.

4 Discussion

The equation 11 concentrates the hypothesis of the model:

Hyp. 1 A transcription unit is seen as an unit subjected to evolutionnary pressure.

The subject of the pressure is traditionally the bacteria as a whole. The easiest measurement of this pressure is the differential of growth rate between two colonies. The survival of a transcription unit is linked to the survival of the bacteria and not necessarily to the fact that it is more easily transcribed. For example, if it prevents a more important gene from being transcribed, the gain is negative for the bacteria. However, we do not observe individual transcription units but only the overall distribution, which is globally linked to the gain for the bacteria of transcribing a given set of genes codirectionally. Therefore, the effect of this approximation are largely reduced.

Hyp. 2 There is a mechanism that allows a transcription unit to move from one half chromosom to the other.

This mechanism exists. Recombination can lead to the inversion of a part of the chromosome. However, unless two successive recombinations occurs, many genes are inverted and not only one. Therefore, the mechanism is quite undirect.

Hyp. 3 There is a mechanism that allows an inversion of the direction of transcription

This mechanism exists. The same recombination mechanism can lead to an inversion of the direction of transcription. It happens if Oric and Ter are not the same fragment recombined, else it leads to an inversion.

Hyp. 4 There is a mechanism that allows a transcription unit to change length

The splitting a transcription unit into several ones, or melting several ones in one. Besides, this is almost independent from the function because it

has been shown that the order of genes is more often conserved between two genomes than the belonging to a transcription unit. It suggests that functional coherence is guaranted by proximity rather than by belonging to the same transcription units, and that the coregulation through integration is not a strong advantage fro the bacteria.

Hyp. 5 The different distributions are exponential

We see on graphs 2, 7, and 7 that the distribution are not fully exponential. It is particularly visible on the half chromosome 2 of Bacillus Subtilis 7. There are generally too many short transcription units (< 1000 bp). However, the quality of the dataset is not sufficient to refine the model. In fact, the prediction of transcription units basically uses the fact that contiguous genes transcribed in opposite direction are in different transcription units. With this alone, the distribution is much closer from an exponential. When other criteria are introduced, as terminators or intergenic distance, long transcription units are split into short ones. The other extreme is the gene distribution, which is much more concentrated around its mean value. When transcription units are split, the bias in quantity β_h increases because these units are composed of more genes on the $+$ side than on the $-$ side, but the bias in mean length decreases, because the gene are almost equally long on either side.

Therefore, the distribution might change with better datasets but not in large proportion because there is no reason why there should be more false-positive than false negative and vice-versa.

So that we keep the exponential distribution model. The experimental data follow it sufficiently to be confident about the consistency of estimators. However, in the developpement of the model, it would be interesting to consider introducing a cost for too short transcription units, which would give a much more centered distribution than an exponential one.

Hyp. 6 The consistence estimators of the model are valid

The model gives estimators to evaluate its consistence. The first one is R_h that makes the ratio of the bias in quantity to the bias in mean length (cf 16). We see from table 4 that this estimator is undistinguishable from one for Escherichia Coli and Bacillus Subtilis half chromosome 1. For half chromosome 1, it is at least 1.5, which indicates that the bias in quantity β_h cannot be completly explained by the model. A difference in the cost of solving head-on or codirectional collisions would not explain this bias either. If we observe graph 10, we see that the exceed in bias is independent of the length of the transcription unit. It means that there is a bonus for

a transcription unit to be transcribed head-on independent of length. This bonus is about 1.3, which correspond to an evolutionnary cost of 0.26. This could be explained by the non-linear effect described in 2.

The second one is the half chromosome bias Q (cf 19). This estimator is of course around 1 for Escherichia Coli, but it is also for Bacillus Subtilis. It means that the bias in mean length between the 2 half chromosomes is correlated to the bias in quantity between these two chromosomes.

Hyp. 7 The cost of interruption is the same for head-on and codirectional collisions, i.e. a head-on collision is not more difficult to solve than a codirectional, it is just more frequent.

This hypothesis is to be checked with the speed ratios γ_h obtained. Other estimations exists for Escherichia Coli. They give $V_h^{re} = 600 - 1000\,\mathrm{b/s}$ [3] and $V_{tr}^h = 42\,\mathrm{b/s}$ [7], i.e. $\gamma_h = 14 - 24$. We found that this ratio was higher than 15 which confirms that we found the right magnitude for γ_h.

For Bacillus Subtilis, we have no estimation of the speeds, but, if we consider that it takes the same time to replicate the 2 half chromosomes and that transcription is equally fast on both of them, it roughly leads to:

$$\frac{\gamma_1}{\gamma_2} = 1.08 \tag{42}$$

From the estimation of the model, we expected more 1.5. We need more information to conclude on this point, because the time of replication may not be the same on both chromosomes. At least the direction of the bias is predicted.

Another solution to overcome such difficulties is not to make the hypothesis that head-on collisions and codirectional ones costs as much to the cell, it is easy to change the model. We replace 14 by:

$$\begin{cases} \theta_{+,h} &= L_h^{tot} \left(C_{+,h}(\gamma_h - 1)\right)^{-1} \\ \theta_{-,h} &= L_h^{tot} \left(C_{-,h}(\gamma_h + 1)\right)^{-1} \end{cases}$$

i.e.:

$$\frac{\widehat{C_{+,h}}}{C_{-,h}} = \frac{\widehat{\theta_{+,h}}}{\widehat{\theta_{-,h}}} \frac{\gamma_h + 1}{\gamma_h - 1} \tag{43}$$

Therefore, the mean length bias is decoupled from the speed ratio thanks to the introduction of a bias in the cost.

Hyp. 8 The evolutionnary cost is related to the proportion of transcripts aborted, and not at all on the quantity of transcripts aborted, i.e. we suppose that activities of genes does not play any role in the bias.

If the cost was related to the absolute number of transcripts aborted, the energy would be in L^2 (one time for the probability of one transcript to be interrupted, and one time for the number of transcripts that may be interrupted). This is not what we observe. Of course, it is always possible to find a model to cope with that fact, but it will be far more complex than this one.

Once the validity of the model is established, one can draw significative observations from it. First, the statistics confirmed the hypothesis developed in the introduction:

- More transcription units are transcribed codirectionally than head-on (cf table 4).
- The bias increases with the length of the transcription units (cf graph 8)

Then, in table 4, we see that:

- If the speed ratio is high, the bias in mean length and in quantity is low. In fact, the faster the replication is compared to the transcription, the lower is the difference of interruption probability between the two orientations.
- If the speed ratio is high, the cost of interruption is weak. Indeed, the bacteria compensates the fact that there are a lot of collisions by diminishing their costs. This assertion is conforted by the fact that the mean length of units transcribed head-on is nearly the same for the three subsets. It looks as if the cell could not afford to have a shorter mean lengthfor the units transcribed on one side.
- The difference between the two half chromosomes of Bacillus Subtilis is so large, that it might be that the two replication forks are different at the molecular level. In particular, one have to explain why there is an advantage of 0.26 for every gene transcribed codirectionally on half chromosome 1.

One should notice that the last 2 observations are not needless in the consistence of the model. Further experiments are needed, in order to confirm or infirm the hypothesis and the predictions of the model.

Bibliography

[1] Baker T.A., *Replication arrest*, Cell, Vol. 80, 521524, February 24, 1995.

[2] Hill T.M., *Arrest of bacterial replication*, Annu Rev Microbiol, 1992, 46:603-33

[3] J. L. Campbell and N. Kleckner, *E. coli oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork*, Cell, Vol. 62, Issue 5, 967-79, September 7, 1990.

[4] Hirose S., Hiraga S., Okazaki T., *Inititiation site of deoxyribonucleotide polymerization at the replication origin of the Escherichia Coli chromosome*, Mol Gen Genet (1983) 189:422-431

[5] Tecklenburg M., Naumer A., Nagappan O., Kuempel P. *The dif resolvase locus of the Escherichia coli chromosome can be replaced by a 33-bp sequence, but function depends on location*. Proc Natl Acad Sci U S A. 1995 Feb 28;92(5):1352-6.

[6] Deshpande A.M., Newlon C.S., *DNA replication fork pause site dependent on transcription*, Science, Vol 272, 17 May 1996, 1030-1033

[7] Gotta S.L., Miller O.L. Jr, French S.L., *rRNA transcription rate in Escherichia coli*, J Bacteriol, Vol. 173, Issue 20, 6647-9, October. 1991.

[8] M. Mok and K.J. Marians, *The Escherichia coli preprimosome and DNA B helicase can form replication forks that move at the same rate*, Biol. Chem., Vol. 262, Issue 34, 16644-54, December 5, 1987.

[9] Mol Microbiol 1995 Sep;17(5):825-31 The speed of the Escherichia coli fork in vivo depends on the DnaB:DnaC ratio. Skarstad K, Wold S.

[10] French S., *Consequences of replication fork movement through transcription units in vivo*, Science, Vol. 258, Issue 5086, 1362-5, 20 November 1992.

[11] The EMBO Journal Vol. 17,pp. 5095-5102, 1998, Transcription through a simple DNA repeat blocks replication elongation Maria M. Krasilnikova, George M. Samadashwily, Andrey S. Krasilnikov and Sergei M. Mirkin

[12] Liu B, Wong ML, Alberts B. Related Articles, *A transcribing RNA polymerase molecule survives DNA replication without aborting its growing RNA chain*, Proc Natl Acad Sci U S A. 1994 Oct 25;91(22):10660-4.

[13] Liu B., Alberts B.M., *Head on collision between a DNA replication apparatus and RNA polymerase transcription complex*, Science, Vol 267, Issue 5201, 1131-7, February 24, 1995.

[14] Elias-Arnanz M., Salas M., *Bacteriophage phi29 DNA replication arrest caused by codirectional collisions with the transcription machinery*, EMBO J, Vol. 16, Issue 18, 5575-83, September 15, 1997.

[15] Elias-Arnanz M., Salas M., *Resolution of head-on collisions between the transcription machinery and bacteriophage phi29 DNA polymerase is dependent on RNA polymerase translocation* , EMBO J, Vol. 18, 5675-5682, 1999.

[16] Blattner et al., The Complete Genome Sequence of Escherichia coli K-12, Science 1997 277: 1453-1462

[17] Salgado H., Moreno-Hagelsieb G., Smith T.F., and Collado-Vides J., *Operons in Escherichia coli: Genomic analyses and predictions*, Proc. Natl. Acad. Sci. USA, Vol. 97, Issue 12, 6652-6657, June 6, 2000.

[18] T. Vermat, Y. d'Aubenton-Carafa, Y. Vandenbrouck, A. Viari and C. Thermes. Evolution of Rho-independent transcription terminators in bacteria (not yet published)

[19] Brewer B.J. *When polymerases collide: replication and the transcriptional organization of the E. coli chromosome*, Cell. 1988 Jun 3;53(5):679-86.

[20] Rocha E.P.C., Danchin A., *Essentiality, not expressiveness, drives gene strand in bacteria*

5 Annexes

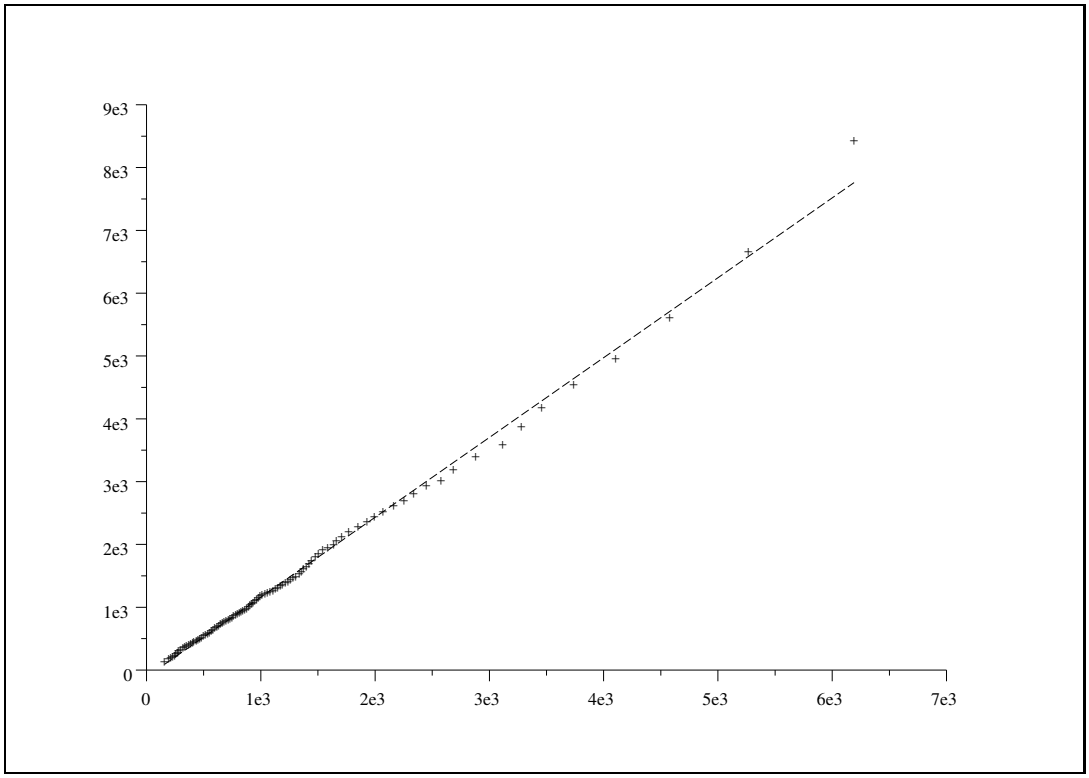

Figure 1: Quantile of length distribution of transcriptions units of Bacillus Subtilis function of this quantile for Escherichia Coli

	Escherichia Coli 1	Escherichia Coli 2	Escherichia Coli 1+2
L_h^{tot} (in bp, cf 3)	2.302.540	2.336.681	2.319.610
$N_{+,h}$ (cf 20)	581	593	1174
$N_{-,h}$ (cf 20)	556	553	1109
$N_{+,h}/N_{-,h}$	1.01-1.08	1.04-1.10	1.04-1.08
$\theta_{+,h}$ (in bp, cf 8)	1304-1434	1278-1405	1310-1401
$\theta_{-,h}$ (in bp, cf 8)	1241-1369	1163-1282	1220-1308

Table 1: Primary estimators for Escherichia Coli

	Bacillus Subtilis 1	Bacillus Subtilis 2	Bacillus Subtilis 1+2
L_h^{tot} (in bp, cf 3)	2.191.525	2.023.105	2.107.315
$N_{+,h}$ (cf 20)	923	705	1628
$N_{-,h}$(cf 20)	410	398	808
$N_{+,h}/N_{-,h}$	2.19-2.31	1.72-1.82	1.97-2.06
$\theta_{+,h}$ (in bp, cf 8)	1562-1685	1910-2083	1734-1836
$\theta_{-,h}$ (in bp, cf 8)	1130-1266	1059-1188	1114-1208

Table 2: Primary estimators for Bacillus Subtilis

η (cf 7)	1.21
$\eta \in$	1.18-1.23
Q (cf 19)	0.95-1.21

Table 3: Half chromosome bias for Bacillus Subtilis

	Bacillus Subtilis 1	Bacillus Subtilis 2	Escherichia Coli 1+2
L_h^{tot} (in bp, cf 3)	2.191.525	2.023.105	2.319.610
$N_{+,h}$ (cf 20)	923	705	1174
$N_{-,h}$ (cf 20)	410	398	1109
$N_{+,h}/N_{-,h}$	2.19-2.31	1.72-1.82	1.04-1.08
$\theta_{+,h}$ (in bp, cf 8)	1562-1685	1910-2083	1310-1401
$\theta_{-,h}$ (in bp, cf 8)	1130-1266	1059-1188	1220-1308
R_h (cf 16)	1.47-1.87	0.87-1.14	0.90-1.08
γ_h (cf 4)	5.84-9.52	3.90-4.29	15.0-$+\infty$
$10^4 C_h$ (cf 11)	0.751-1.46	1.59-2.32	0-0.528

Table 4: Estimators

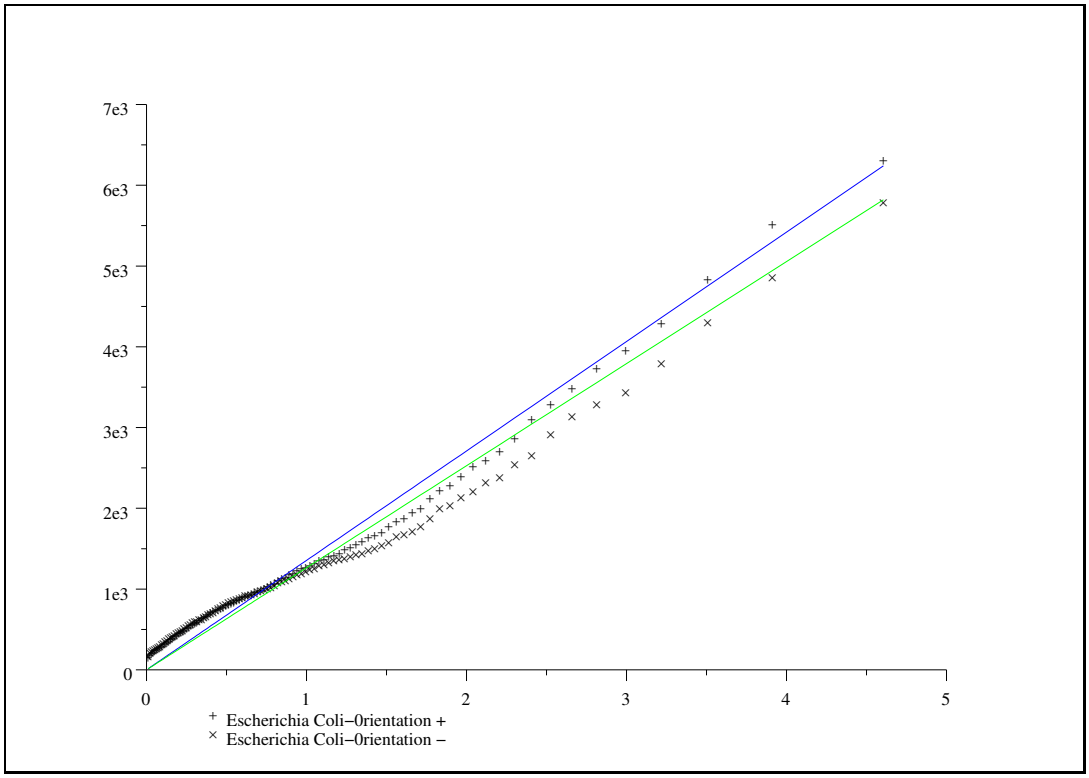

Figure 2: Quantiles of the length of transcription units in Escherichia Coli (in base pair) function of the quantile of a normal exponential distribution

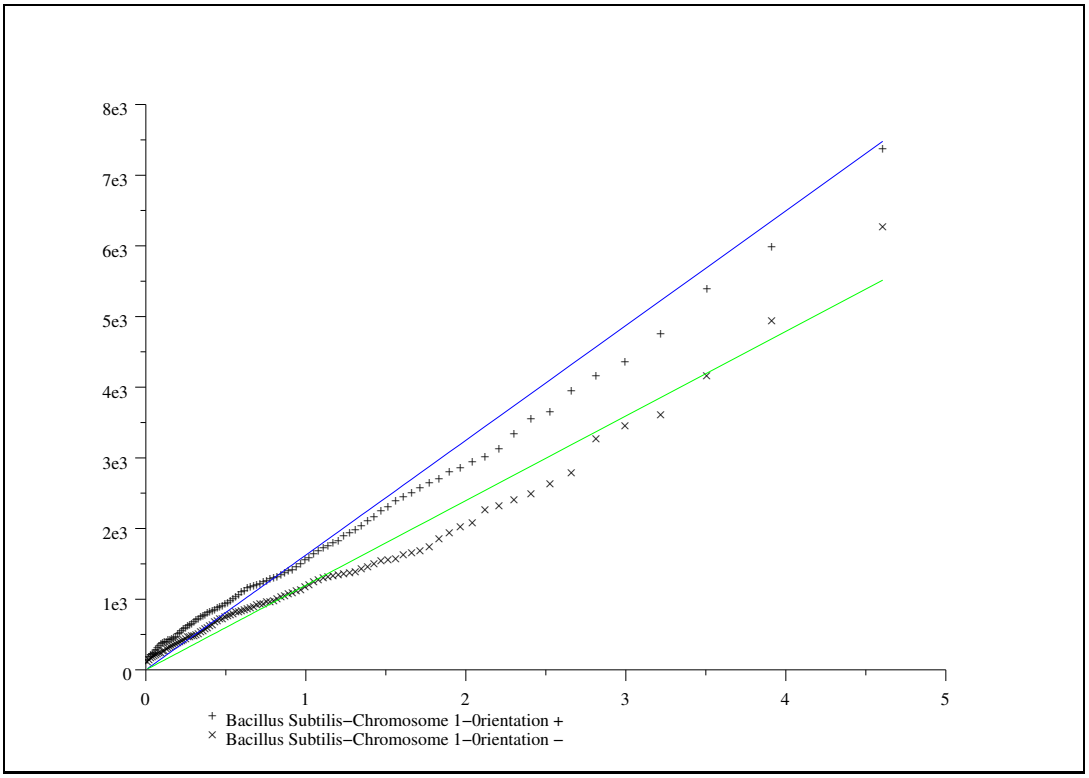

Figure 3: Quantiles of the length of transcription units in Bacillus Subtilis half chromosome 1 (in base pair) function of the quantile of a normal exponential distribution

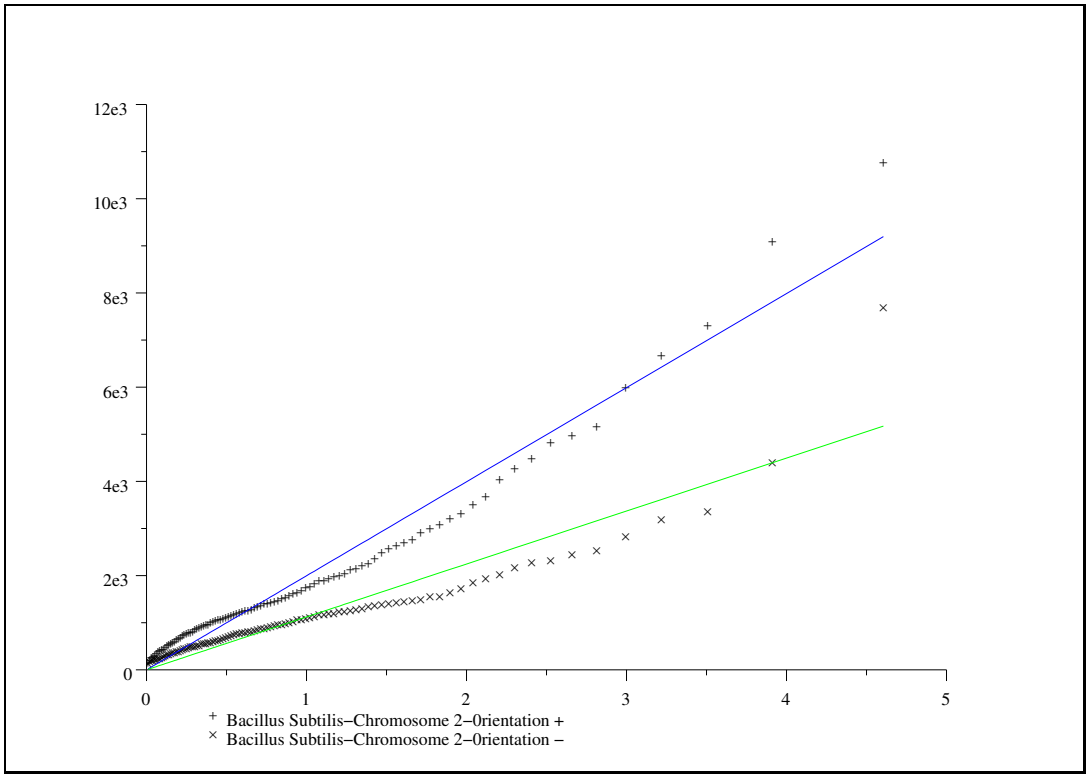

Figure 4: Quantiles of the length of transcription units in Bacillus Subtilis half chromosome 2 (in base pair) function of the quantile of a normal exponential distribution

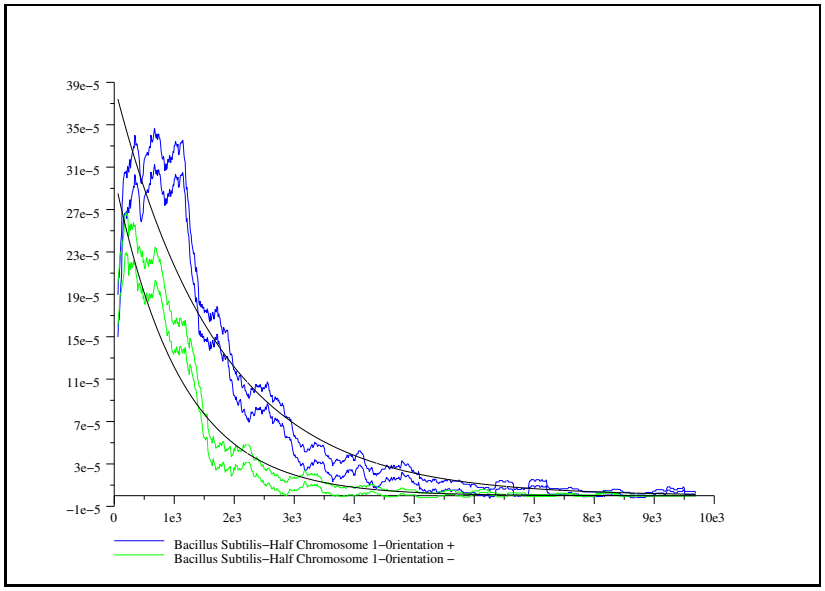

Figure 5: Presence density of trancription units in Escherichia Coli (window: 300 bp, confidence interval: 84%)

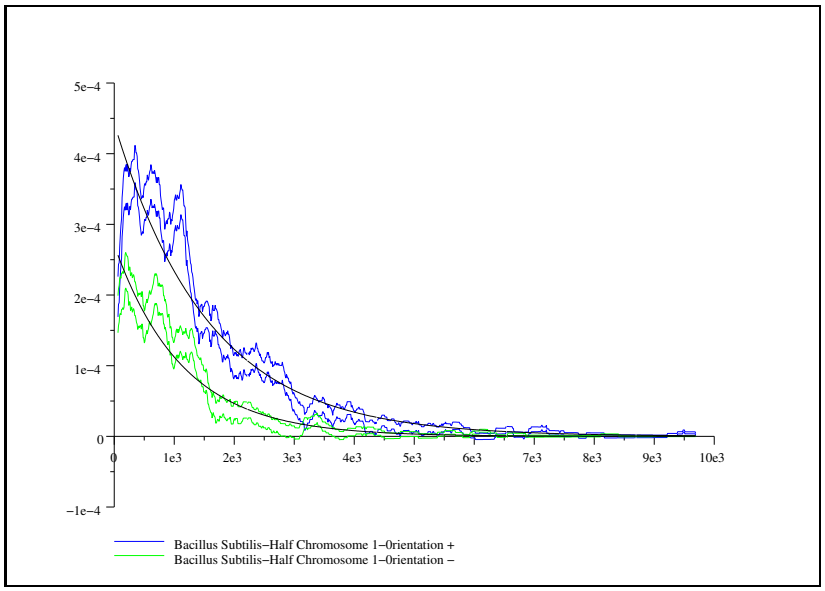

Figure 6: Presence density of operons in Bacillus Subtilis half chromosome 1 (window: 300 bp, confidence interval: 84%))

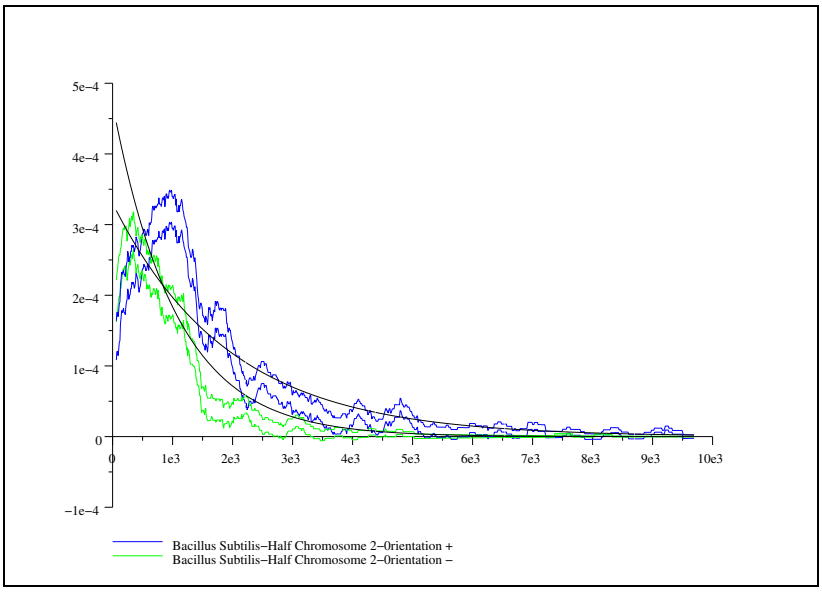

Figure 7: Presence density of operons in Bacillus Subtilis half chromosome 2 (window: 300 bp, confidence interval: 84%))

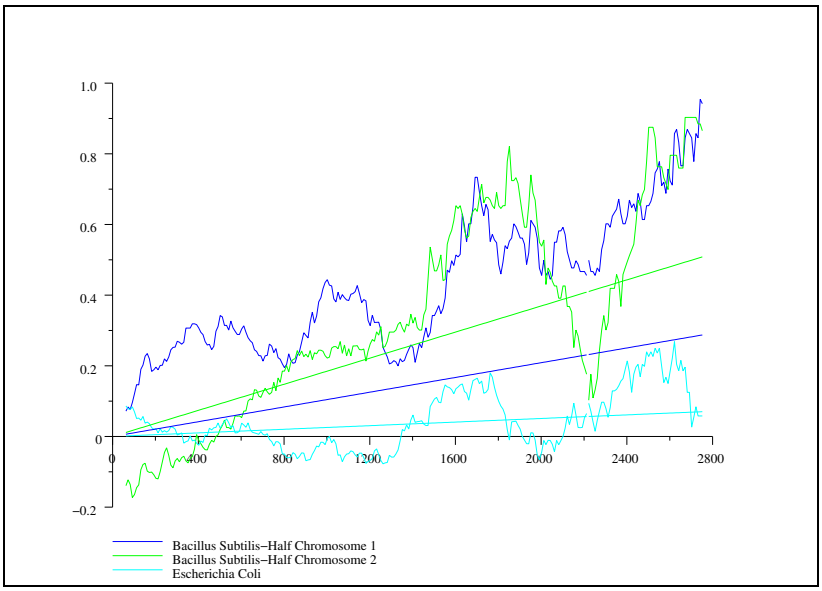

Figure 8: Presence density ratio (window: 300 bp)

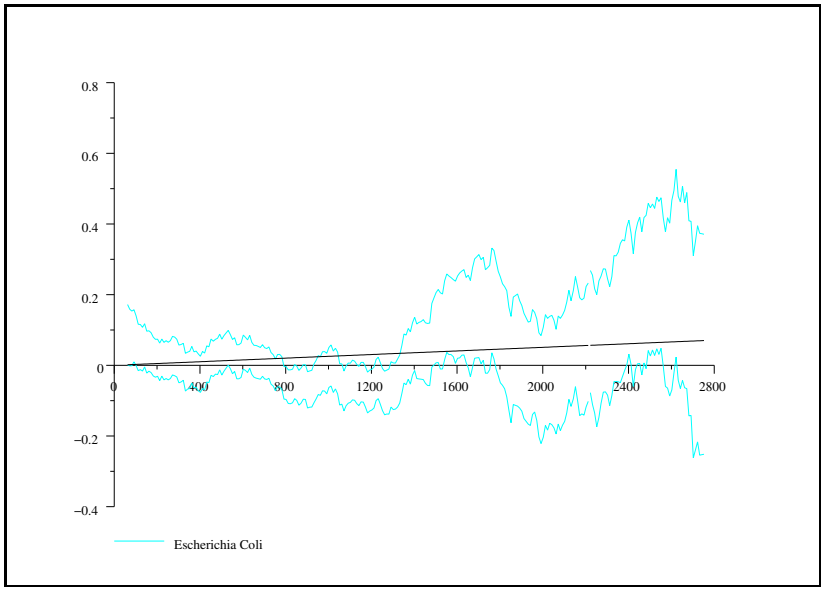

Figure 9: Presence density ratio in Escherichia Coli (window: 300 bp, confidence interval: 84%)

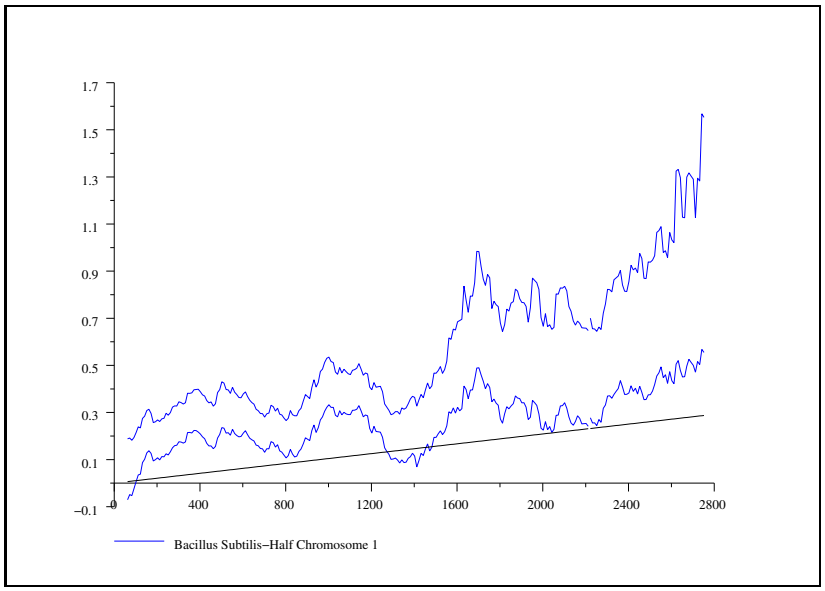

Figure 10: Presence density ratio in Bacteria Subtilis half chromosome 1 (window: 300 bp, confidence interval: 84%))

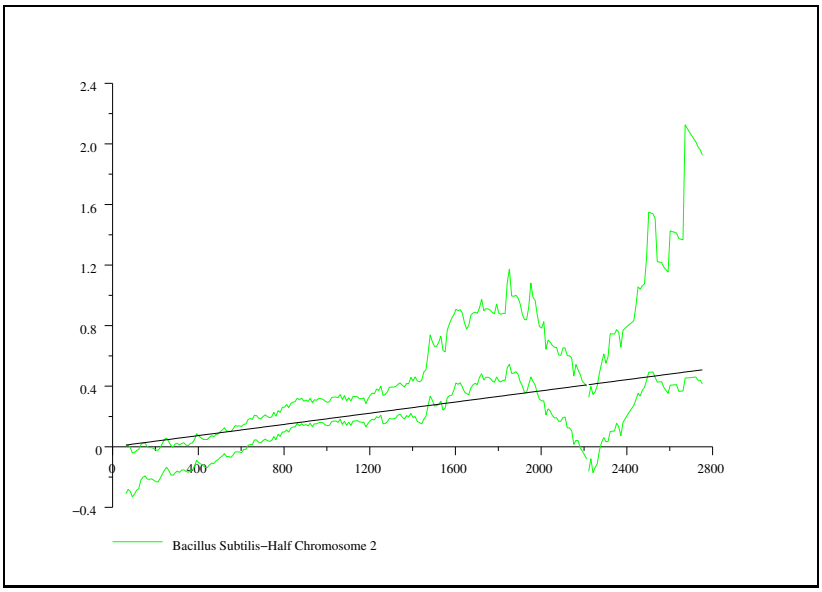

Figure 11: Presence density ratio in Bacteria Subtilis half chromosome 2 (window: 300 bp, confidence interval: 84%))