A Genome-Wide Evolutionary Simulation of the Transcription-Supercoiling Coupling

Théotime Grohens, Sam Meyer, Guillaume Beslon

A Genome-Wide Evolutionary Simulation of the Transcription-Supercoiling Coupling

Théotime Grohens[1], Sam Meyer[2] and Guillaume Beslon[1]

[1]INRIA Beagle Team, Artificial Evolution and Computational Biology,
Université de Lyon, Inria, ECL, INSA Lyon, Université Claude Bernard Lyon 1, Université Lumière Lyon 2,
CNRS, LIRIS UMR 5205, F-69622, France
[2]Université de Lyon, INSA Lyon, Université Claude Bernard Lyon 1, CNRS, UMR5240 MAP, F-69622, France
theotime.grohens@insa-lyon.fr

Abstract

DNA supercoiling (SC), the level of under- or overwinding of the DNA polymer around itself, is widely recognized as an ancestral regulation mechanism of gene expression in bacteria. Higher negative SC levels facilitate the opening of the DNA double helix at gene promoters, and increase the associated expression levels. Different levels of SC have been measured in bacteria exposed to different environments, leading to the hypothesis that SC variation can be an environmental response. Moreover, DNA transcription has been shown to generate local variations in the SC level, and therefore to impact the transcription of neighboring genes.

In this work, we study the coupled dynamics of DNA supercoiling and transcription at the genome scale. We implement a genome-wide model of gene expression based on the transcription-supercoiling coupling (TSC). We show that, in this model, a simple change in global DNA SC is sufficient to trigger differentiated responses in gene expression levels via the TSC. Then, studying our model in the light of evolution, we demonstrate that this SC-mediated non-linear response to environmental change can serve as the basis for the evolution of specialized phenotypes, through the selection of a specific genomic architecture.

Introduction

The DNA molecule is a double-stranded polymer of nucleotides that plays a fundamental role in life. It is shaped as a double helix which rotates around itself at a rate of around one turn per 10.5 base pairs (Krogh et al., 2018). However, when subject to physical forces, it can become overwound or underwound, or writhe around itself, in a process known as DNA supercoiling (SC); SC level σ is measured as the density of extra turns (or coils) per base pair. In bacterial cells, DNA is usually slightly underwound (Lal et al., 2016), with $\sigma < 0$, a typical value being $\sigma_0 \approx -0.066$ in *E. coli* (Crozat et al., 2005). The SC level is tightly regulated by a class of enzymes called topoisomerases. The main topoisomerases are topoisomerase I and gyrase: gyrase uses ATP to maintain DNA in a negative SC state by adding negative coils, while topoisomerase I relaxes SC and does not need ATP (Martis B. et al., 2019). DNA SC furthermore plays an important role in bacterial cells as an ancestral regulator of gene activity (Dorman and Dorman, 2016). Indeed, as shown in figure 1A and 1B, low SC levels ($\sigma < \sigma_0$) favor higher expression of bacterial genes, as the thermodynamic reaction of promoter opening required to begin transcription is facilitated (Meyer and Beslon, 2014). Conversely, high SC levels ($\sigma > \sigma_0$) reduce gene expression; such variations in DNA SC have been shown to affect 7% of *E. coli* genes (Peter et al., 2004). In some bacterial species, such as *Buchnera aphidicola*, an obligate aphid endosymbiont with a greatly shrunk genome, gene regulation nevertheless takes place in the near-total absence of transcription factors; in these species, DNA SC is suspected to be the main, if not the sole, regulatory mechanism (Brinza et al., 2013).

Dynamic properties of DNA supercoiling DNA SC is under the influence of both internal and external constraints. It varies both in time, during the lifecycle of the bacterium, which alternates between growth and stationary phases (Krogh et al., 2018), and in space, as different regions of the chromosome experience different SC levels (Lal et al., 2016; Junier and Rivoire, 2016). In bacteria, the SC homeostasis is mainly regulated by topoisomerases, especially topoisomerase I and gyrase, but nucleoid-associated proteins (NAPs) such as *H-NS* also play a topological role by preventing the relaxation of superhelical stress at their fixation points in the genome, resulting in topological domains that have different SC levels (Krogh et al., 2018). Moreover, the global SC level can change as a genome-wide response to environmental changes such as pH (Martis B. et al., 2019).

Crucially, the transcription process of DNA to RNA itself plays a role in local SC level variations. Indeed, when an RNA polymerase (RNAP) transcribes a gene, it follows the helical twist of the DNA template, but its rotation is hampered by frictional drag (Ma and Wang, 2016). This generates an accumulation of positive SC through overwinding of the DNA molecule downstream of the transcribed region, and of negative SC through underwinding upstream of the transcribed region (Liu and Wang, 1987), as shown in figure 1C. As transcription is itself regulated by SC (figures 1A and B), this results in an dynamic interaction between the

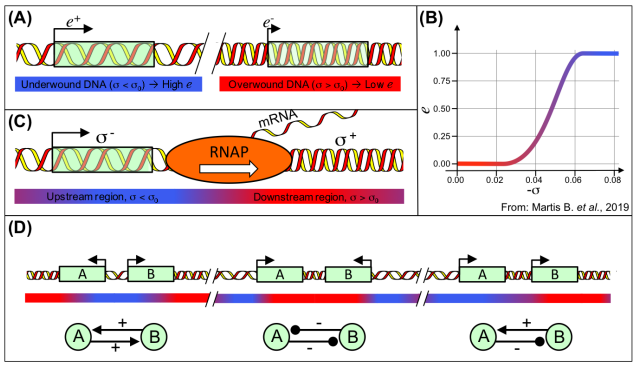

Figure 1: **A**. When DNA is underwound (lower σ, left), gene transcription rates are higher than when DNA is overwound (higher σ, right). **B**. Promoter activity e increases with the level of negative SC $-\sigma$. **C**. The transcription of a gene by RNA polymerase (RNAP) generates a decrease in SC upstream of the transcribed gene, and an increase downstream. **D**. Transcription-supercoiling coupling: the sign of the interaction between neighboring genes depends on their relative orientation.

transcription levels of neighboring genes, which has been termed the transcription-supercoiling coupling (TSC).

The Transcription-Supercoiling Coupling The existence of a coupling between transcription and supercoiling, or transcription-supercoiling coupling (TSC) has been experimentally shown through measurements of the expression level of neighboring genes in both prokaryotic and eukaryotic organisms (Meyer and Beslon, 2014). Its influence on gene expression is represented in figure 1D: when two neighboring genes stand in diverging orientations (figure 1D, left), the transcription of each gene generates a local increase in negative SC around the other gene, thereby increasing the other gene's transcription level; both genes therefore reinforce each other's transcription. Conversely, when two neighboring genes face each other in converging orientations (figure 1D, center), each gene is situated downstream of the RNA polymerase during the other gene's transcription, leading to a decrease in negative SC and therefore a lower transcription level; both genes inhibit each other. Finally, if two genes are in a colinear orientation (figure 1D, right), the downstream gene up-regulates the upstream gene, and the upstream gene down-regulates its downstream neighbor. The typical distance at which this interaction operates is 2,500 bp on each side of the transcribed gene (Meyer and Beslon, 2014).

Several models have been proposed to describe the TSC. In Meyer and Beslon (2014), a quantitative model of the SC level at a locus of interest is proposed. DNA transcription is regulated by the opening energy of DNA around gene promoters, which directly depends on the SC level. In this model, the reciprocal influence of neighboring genes can be obtained by computing the difference in transcription levels due to SC and subsequent SC variation, and iterating this system until a fixed point is reached. El Houdaigui et al. (2019) describe a more detailed stochastic model of DNA transcription involving explicit RNA polymerases and topoisomerases. The transcription level of a genomic region of interest is simulated using discrete time steps, during which RNA polymerases attach to the DNA template, progress along the transcribed region while generating positive SC downstream and negative SC upstream, and detach from the DNA, relaxing SC constraints.

These models however limit themselves to mechanistic descriptions of the local interaction between genes, but do not try to generalize to the whole-genome scale nor to the evolutionary level. Yet, the dense gene content of bacteria suggests that the TSC can generate a global transcriptional interaction network through the propagation of local SC variations, as opposed to the much more isolated eukaryotic genes. Indeed, in bacteria, distances between neighboring genes are classically around 1,000 bp (Blattner, 1997), low enough to connect multiple genes through the TSC. Moreover, global gene regulation through SC does evolve in nature, as exemplified in *Buchnera* (Brinza et al., 2013), an endosymbiotic bacteria with streamlined genomes in which SC has evolved as one of the main regulation mechanisms, and in experimental evolution, where SC has been shown to drive the evolutionary response of *E. coli* strains.

Evolution of DNA supercoiling regulation in bacteria In the Long-Term Evolution Experiment (LTEE) (Lenski et al., 1991), 12 populations of *E. coli* bacteria have been evolving for over 80,000 generations in a glucose-limited environment. Not only have parallel increases in the level of SC been measured in 10 of those populations (Crozat et al., 2010), but mutations in two genes regulating DNA SC, *topA* and *fis*, have been identified as the genetic basis for this phenotypic change, and have been verified to confer a fitness advantage (Crozat et al., 2005). These results suggest a strong selection pressure to tune SC to the new environment of the LTEE.

Evolution of SC has also been observed in the wild: in *Dickeya dadantii*, a plant pathogenic bacteria, different genomic regions exhibit markedly different responses to changes in SC (Muskhelishvili et al., 2019), allowing the expression of certain genes only in specific environments. This suggests that specific chromosomal organizations can evolve as a way to trigger the activity of certain pools of genes depending on the change in SC level caused by different environments.

Both the importance of SC regulation and the detailed mechanisms of the TSC at the local scale have been well studied, but a thorough analysis of the genome-wide effect

of the TSC on gene expression, and of its possible evolutionary use by natural selection, remains missing. Here, we describe a new model which incorporates a high-level model of global SC regulation and of the TSC within an *in silico* experimental evolution setting. Using this model, we first investigate the non-linear variation in gene transcription levels at the whole-genome scale in response to variations of the global SC level. Then, we study the evolutionary trajectory of genomic organization (the relative orientations and positions of genes on the genome) under the influence of the TSC.

We show that in our model, a genome-scale gene interaction network emerges from local interactions, allowing complex phenotypic modifications in response to the change of a single parameter, the global SC level. Moreover, we demonstrate that, using genomic inversions as the sole mutation operator, evolution can select genomes displaying qualitatively different phenotypes under different basal SC conditions.

A Genome-Wide TSC Model

Overview We present a model aiming at studying the genome-wide effect of the transcription-supercoiling coupling in bacteria, expanding upon previous work which only considered local interactions. The model, written in Python, consists in an individual-based simulation[1]. An individual possesses a circular genome, comprising a fixed number of genes, that is represented using the string-of-pearls paradigm: the genome is simply a ring of genes. Each gene is described by the following characteristics: its position on the genome, its orientation, and its initial expression level.

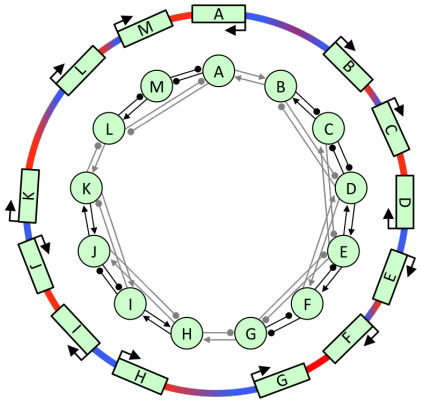

Figure 2: Genes on a genome and local SC variations (outer ring), and associated interaction network (inner ring). Outer ring color signals locally high ($\sigma > \sigma_0$, red) or low ($\sigma < \sigma_0$, blue) SC levels. Closer genes interact more strongly (black arrows) than farther apart genes (gray arrows).

The Transcription-Supercoiling Coupling (TSC) Our model aims at modeling the interaction between the transcriptional activities of genes at the whole-genome scale and the local level of SC. When a gene is transcribed, the RNA polymerase that performs the transcription acts as a topological barrier, and prevents the relaxation of SC on either side of itself (El Houdaigui et al., 2019). As a consequence, DNA upstream of the transcribed gene is underwound and the negative SC level increases during transcription, while DNA downstream of the transcribed gene is overwound and the negative SC level decreases (figure 1C). This affects the transcription of neighboring genes, as a higher level of negative SC around a gene's promoter increases transcriptional activity. Therefore, a highly expressed gene can increase the expression level of its upstream genes, and decrease the expression level of its downstream genes, as has been measured in vivo (El Houdaigui et al., 2019), and as has been already modeled (Meyer and Beslon, 2014). In gene-dense genomes such as prokaryotic ones, this is likely to create a genome-wide transcriptional network, the topology of which depends on the relative gene orientation. Figure 2 shows an example genome, including the local SC variations due to gene transcription, and the resulting gene interaction network. Genes not only interact with their closest neighbors, but also (albeit more weakly) with more distant genes.

Description of the model In order to simulate the influence of the TSC on gene expression levels, we model the temporal change of these expression levels over a number of time steps, called the lifecycle of an individual.

For an individual with n genes, we first compute the local variation in supercoiling at the locus of each gene, which is due to the expression of other genes, in the form of a gene interaction matrix. Its coefficients are given by the following equation, describing the influence of gene j on gene i:

$$\frac{\partial \sigma_i}{\partial e_j} = \varepsilon \max(1 - \frac{d(i,j)}{d_{max}}, 0) \quad (1)$$

More precisely, the interaction level depends on the relative orientation of the two genes, as the transcription of a gene has a positive effect on upstream genes and a negative effect on downstream genes. We choose $\varepsilon = 1$ if the gene i is upstream of the gene j (or if $i = j$) and $\varepsilon = -1$ otherwise. It also depends on gene distance, as genes that are further apart on the genome interact less, so the strength of the interaction linearly decreases with the intergenic distance $d(i,j)$, reaching 0 when $d(i,j) = d_{max}$, the maximum distance above which the interaction vanishes.

Using this interaction matrix, we can now compute the SC variation at the locus of each gene at each time step, which depends on the expression level of all the other genes:

$$\Delta\sigma_i(t) = \sum_{j=1}^{n} \frac{\partial \sigma_i}{\partial e_j} e_j(t) \quad (2)$$

[1] Source code available here, and archived here.

Finally, we obtain the expression level of the focal gene at $t+1$, which depends on both the local SC variation $\Delta\sigma_i(t)$ and the variation in global SC level β:

$$e_i(t+1) = \phi(\Delta\sigma_i(t), \beta) \tag{3}$$

Where ϕ is an non-linear saturating activation function that ensures that the expression levels of genes do not diverge over time (figure 1B). Promoters are therefore constitutive in this model, and their activity only depends on the local SC level.

In the model, the variation in global SC level $\beta = (\sigma - \sigma_0)/\sigma_0$ represents the result of topoisomerase activity, which causes a relative variation in SC from σ_0 to σ. As a parameter of the activation function ϕ used in computing the temporal evolution of gene activation levels, β acts as a bias, and its effect depends on the precise shape of this activation function. Using a classical linear activation function, we would obtain: $\phi(\Delta\sigma, \beta) = \max(0, \min(c\Delta\sigma + \beta, 2))$, meaning that, depending on the sign of β, the expression level of an isolated gene would go to 0 or 2 independently of its initial value (here, c is a scaling factor representing the change in gene activity per variation in supercoiling). Another natural choice would be the hyperbolic tangent function, used in neural networks: $\phi(\Delta\sigma, \beta) = \tanh((c\Delta\sigma - 1) + \beta) + 1$. However, this function has only one fixed point, and this fixed point is attractive and stable, meaning that whatever the initial expression level of an isolated gene, its expression level will converge towards this fixed point.

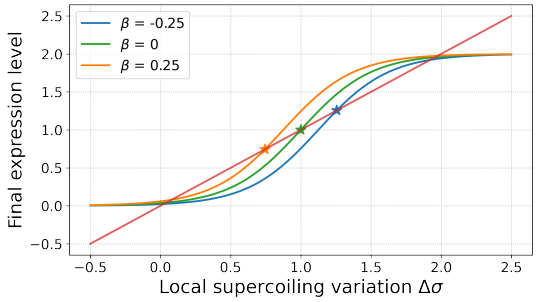

Figure 3: Activation functions ϕ for $\alpha = 2$ and 3 values of β. Stars indicate the values of e^* (saddle point between the basins of attraction of the activated and inhibited states) for the 3 values of β.

Adding a gain parameter $\alpha > 1$ to the hyperbolic tangent yields $\phi(e, \beta) = \tanh(\alpha(e-1) + \beta) + 1$. The activation function now has three different fixed points e^-, e^*, and e^+, as represented in figure 3. e^- and e^+ are both attractive stable fixed points, and e^* is the saddle point between their basins of attraction: if an isolated gene's expression level e is lower (resp. higher) than e^*, it will converge to e^- (resp e^+).

In order to keep the expression levels between 0 and 2, we use a rescaled hyperbolic tangent function, using a gain value of $\alpha = 2$ and $c = 1$:

$$\phi(\Delta\sigma, \beta) = \tanh(2(\Delta\sigma - 1) + \beta) + 1 \tag{4}$$

Finally, we iterate this procedure for a defined number of time steps, and obtain a time series of gene expression levels.

The bistability of the activation function ϕ allows us to define *activated* and *inhibited* states for each gene, depending on the expression level that the gene converges to; note that due to the non-linear asymmetric interactions between neighboring genes, more complex behaviors than simple convergence to the fixed points can emerge, as exemplified in figure 8.

Influence of the Global Supercoiling Level

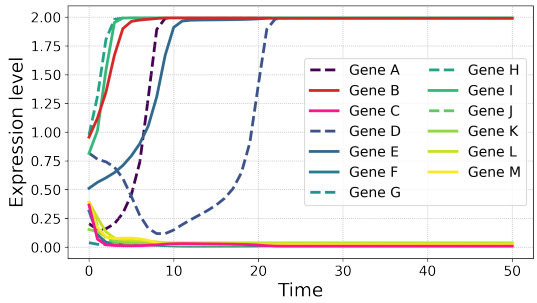

Figure 4: Lifecycle of the individual presented in figure 2, under relative global SC $\beta = 15\%$. Solid lines represent forward genes, and dashed lines reverse genes.

Figure 4 shows the lifecycle of an example individual with a genome of 13,000 bp and $n = 13$ genes, over $t_{end} = 51$ time steps. Initial gene expression levels are randomly chosen between 0 and 1. The non-linear effect of the interaction between neighboring genes can be clearly seen in this individual. Six genes (A, B, D, E, H, and I) reach the activated state by the end of the individual's lifecycle, while the others reach the inhibited state. These activated genes can be grouped into 3 pairs (A and B, D and E, H and I), all of which are pairs of adjacent genes in divergent orientations. Even though gene A starts at an expression level well below the activation threshold for an isolated gene, it reaches the activated status thank to the positive interaction with gene B. Moreover, we can see that this dynamic system shows complex behavior, as the activation level of gene D begins by decreasing until it is activated by the increasing expression of gene E.

Figure 5 captures the influence of β on the repartition of genes between the activated and inhibited states, in an example individual with 10 genes. From left to right: at a low value of $\beta = -15\%$, meaning that DNA is overwound compared to normal, only one gene is activated by the end of

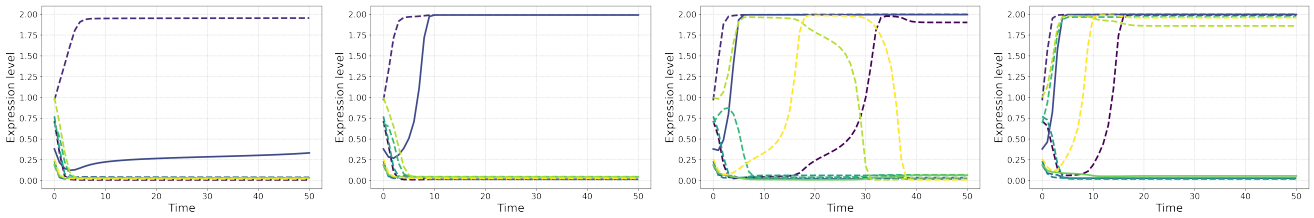

Figure 5: Influence of the relative global SC $\beta = (\sigma - \sigma_0)/\sigma_0$ on the lifecycle of an individual with 10 genes. From left to right: $\beta = -15\%$, 1 genes on; $\beta = 0$, 2 genes on; $\beta = 15\%$, 3 genes on; $\beta = 30\%$, 6 genes on. Higher values of β result in the activation of more genes, reflecting the *in vivo* effect of higher negative SC.

the lifecycle. As the relative change in SC β increases to 0, corresponding to normal relaxation of DNA, the positive interaction of the second activated gene with the first gene makes it go over the activation threshold, and it reaches the activated state before the end of the lifecycle. At an even higher value of $\beta = 15\%$, even more genes reach activation; however, some of them later go back to the inhibited state, demonstrating the complex genome-scale interactions produced by the TSC. Finally, for $\beta = 30\%$, the activation threshold becomes so low that most genes are activated, and stay strongly activated until the end of the lifecycle.

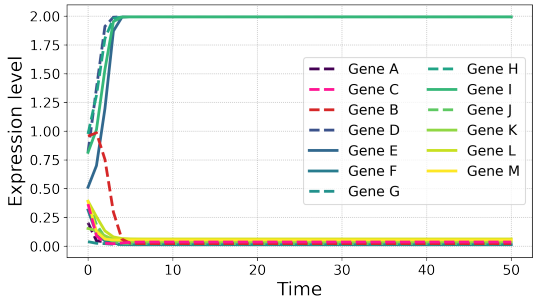

Figure 6: Individual differing from the individual in figure 4 by a segment inversion affecting genes B and C (in shades of red); see figure 7 for details of the inversion.

Figure 6 again shows the lifecycle of the individual in figure 4, after a slight change in its genomic architecture due to a segment inversion. The start point of this inversion falls between genes A and B and its end point between genes C and D; this results in the reversal of segment [BC] relative to the rest of the genome. Here, we can see that the diverging orientation that existed between genes A and B has vanished, replaced by a set of genes in colinear orientation, from A to D. This genomic reorganization results in the loss of the activation of genes A and B, even though the initial expression level of gene B is above the required threshold for the activation of an isolated gene, as the two genes no longer interact positively; only the pairs of genes D and E, and H and I, remain activated.

Based on these observations, we can confirm that in our model, the transcription-supercoiling coupling generates complex networks of genome-wide interactions between genes, and that these networks are directly dependent on the architecture of the genome.

Evolutionary Genome-Wide TSC Model

Having shown that transcriptional activity depends on the organization of the genome, we now question to which extent evolution can leverage genome organization to adapt gene regulatory activity to different environments.

In this section, we expand our model into an evolutionary simulation. At each generation of the simulation, each individual is evaluated and its fitness value is computed, based on its transcriptional activity. Then, the individuals of the new generation are chosen by picking their ancestor from the current generation, with a probability proportional to the ancestor's fitness. The model is panmictic, meaning that any individual in the population can be chosen as the ancestor of any new individual. Finally, the genome of each new individual stochastically undergoes a number of mutations, before the new individual is evaluated again; these mutations importantly do not impact genes themselves, but only the spatial organization of the genome.

Fitness In order to compute the fitness of an individual, we define an optimal phenotype $\tilde{e}$ and compute the *gap*, or average L^2 distance of the individual's gene expression levels e to the optimal levels $\tilde{e}$ (2 for genes to be activated, and 0 for genes to be inhibited), averaged over the Δt last time steps, with the expression:

$$g(e) = \frac{1}{\Delta t} \sum_{t=t_{end}-\Delta t+1}^{t_{end}} \frac{1}{n} \sum_{i=1}^{n} (e_i^t - \tilde{e}_i)^2 \qquad (5)$$

Then, we rescale this fitness by applying an exponential scaling: $fitness(e) = e^{-kg(e)}$, where k is a scaling factor representing the selection pressure. A higher value of k means that more-adapted individuals, those which have a smaller gap, will have an even higher fitness value compared to other individuals; we typically use $k = 50$, meaning that

a small decrease in the gap compared to other individuals yields a large reproductive advantage.

Mutational operator: genomic inversions We introduce only one kind of mutation in our model: genomic inversions. In an inversion, two points are chosen randomly on the genome, and the genomic content between these points is reversed: genes are reinserted in the genome in the opposite orientation and order, taking care to update all intergenic distances appropriately. An inversion has no effect if both its endpoints fall in-between the same two genes, and can impact any number of genes otherwise.

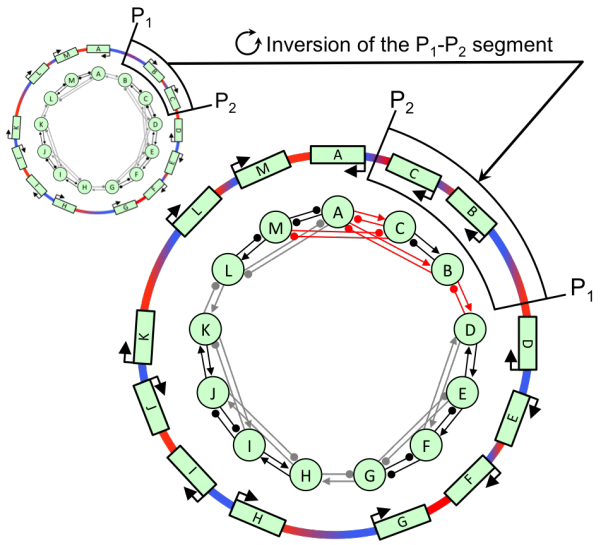

Figure 7: Result of the inversion of a genomic segment containing genes B and C from the individual presented in figure 2. The gene interactions which changed due to the inversion are drawn in red.

Figure 7 presents a genome obtained by performing an inversion on the genome in figure 2. As a result of this inversion, genes B and C have been switched from the forward to the backward orientation, and the intergenic distances between A and C on the one hand, and B and D on the other hand, have been modified; however, the relative orientation of B and C, and hence their interaction subnetwork, remain unchanged. This results in changes to the gene interaction network: instead of mutual activation between genes A and B and mutual inhibition between genes C and D, all four of those genes now lie in colinear orientations, in which each of these genes activates its upstream neighbor but represses its downstream neighbor.

When mutating an individual, we draw the number of inversions k to perform from a Poisson law with parameter $\lambda = 2$, giving an average of 2 inversions between an individual and its ancestor; the probability of not undergoing any mutations is $P(k = 0) = e^{-\lambda} \approx 0.136$.

Evolution of Gene Expression Levels in Different Environments

In this section, we describe an experiment aimed at determining whether, in our simple mathematical model, different phenotypes can evolve as a response to different SC levels induced by the environment, as has been observed in *Dickeya dadantii* (Muskhelishvili et al., 2019).

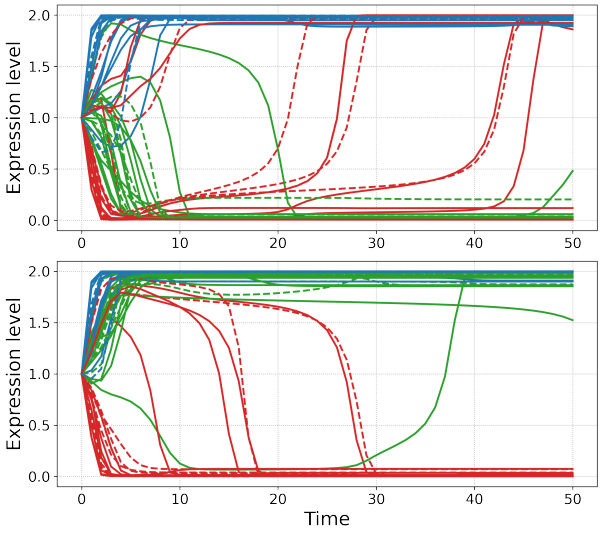

Figure 8: Lifecycle of the best individual in replicate 3 after 200,000 generations. Environment A is on top and B at the bottom. Blue: *AB* genes, red: *A* genes, green: *B* genes.

Experimental setup We model the evolution of populations of individuals facing two different environments, named A and B. Each environment is defined by its value of β, respectively β_A and β_B, which represent the relative change in the SC level σ due to the environment. In order to have environments with relatively high number of activated genes, we chose $\beta_A = 15\%$ and $\beta_B = 30\%$; these correspond to the two rightmost subfigures of figure 5.

We separate genes in three classes, based on the environments in which they must be activated: either both A and B (*AB* genes), only A (*A* genes), or only B (*B* genes). These classes allow us to define optimal phenotypes for both environments: in environment A, both *AB* and *A* genes should be activated, whereas *B* genes should be inhibited. Conversely, in environment B, only *AB* and *B* genes should be activated, but not *A* genes; see figure 8 for the lifecycle of an individual in both environments.

We define the phenotype E_A, E_B of an individual as the pair of temporal series of its gene expression levels in each environment, which is computed as in the model description. Then, we compute its average gaps g_A and g_B in each environment over the last $\Delta t = 5$ time steps of its lifecycle (5), and finally its fitness as $f = e^{-k(g_A+g_B)}$.

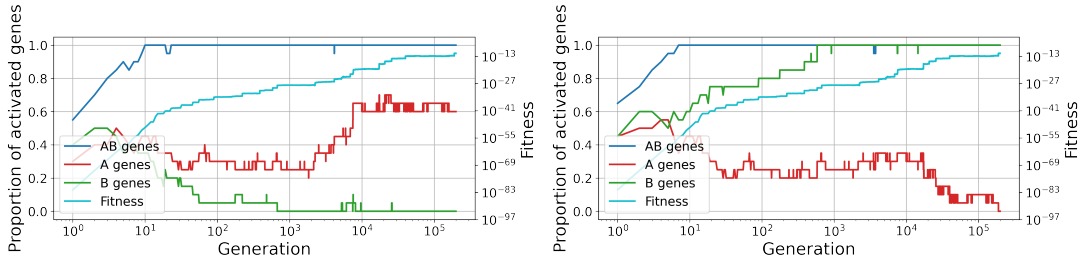

Figure 9: Gene activation levels and fitness of the best individual at every generation of replicate 3, with a population size of $N = 100$, for 200,000 generations. The proportion of active *AB* genes increases until it reaches 1, in both environment A (left) and B (right). The proportion of active *A* (resp. *B*) genes increases in environment A (resp. B) and decreases in environment B (resp. A) over time. Fitness keeps increasing until the end of the run, suggesting that fitter phenotypes remain reachable.

We initialize the simulation with a clonal population of $N = 100$ copies of an initial individual with the following genome: 60 genes in random orientations, uniformly distributed along a 60,000 bp genome, and equally divided between the *AB*, *A* and *B* classes. We choose a maximum interaction distance of $d_{max} = 2500$, meaning that each gene initially interacts with its 2 closest neighbors in each direction through the TSC. Note that as inversions may change intergenic distances, genes can become closer or further apart during evolution.

Finally, we evolved 15 different populations for 200,000 generations; this lasted approximately 24h on a computer with an Intel Xeon E5-2640 v3 @ 2.60GHz CPU, using around 1GB of RAM.

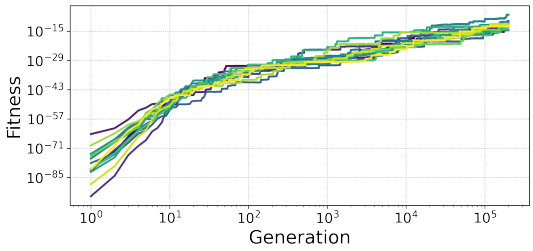

Figure 10: Evolution of the fitness of the best individual of each replicate at every generation.

Adaptation of gene expression levels to different environments The evolution of one of the 15 replicate populations is shown in figure 9. We can first see that the proportion of activated *AB* genes quickly rises to 1 in both environments A and B; this shows that evolving a phenotype that is resistant to environmental perturbations is easy in the model. For *A* genes and *B* genes, we observe an asymmetric tendency towards activation in the target environment, and inhibition in the opposite environment; the difference in the proportion of activated *B* genes between environments A and B is much higher than that of *A* genes. This asymmetry can be understood from the model, as the environments are themselves not symmetric ($\beta_A = 15\%$, $\beta_B = 30\%$), and gene activation is easier in environment B than in environment A; this means that an isolated gene that is activated at $\beta = \beta_A$ is necessarily activated at $\beta = \beta_B$, while the opposite is not true.

Figure 8 focuses on the lifecycle of the best individual of the last generation of replicate 3, in both environments. The phenotypes displayed in each environment present clearly distinct gene expression patterns, showing that a specific gene expression pattern for each environment can evolve through natural selection in a single genome.

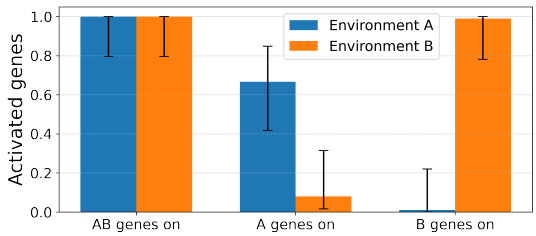

Figure 11: Mean proportion and 95% confidence intervals (Wilson score for binomial proportions) of activated genes in each environment at the end of the lifecycle of the final best individual. For *A* and *B* genes, activation levels differ: p-value 1.07×10^{-15} for *A* genes, and p-value 7.38×10^{-21} for *B* genes (Student's t-test for dependent samples).

The evolution of the fitness of the best individual in each replicate is presented in figure 10. In all 15 cases, fitness gradually increases, albeit at a progressively slower rhythm (note the logarithmic time scale in figure 10). All runs progress during the 200,000 generations, suggesting that adaptation would continue were the experiment to run longer. Finally, figure 11 summarizes these differences in gene expression levels, between environments A and B, for each of the three sets of genes, averaged over the 15 repetitions.

These results show that, in a gene transcription model that is structured around the transcription-supercoiling cou-

pling, complex gene interaction networks, sensitive to environmental variation, can arise as an adaptation to different environments, mediated by the influence of a single parameter: the variation in global SC level β.

Discussion and Perspectives

DNA supercoiling plays a fundamental role in the regulation of gene transcription in bacteria, a significant proportion of which is mediated by the transcription-supercoiling coupling. While the role of basal SC (Lal et al., 2016; Ma and Wang, 2016; Dorman and Dorman, 2016; Martis B. et al., 2019), its evolutionary importance (Crozat et al., 2005, 2010) and the mechanistic details of the TSC (Meyer and Beslon, 2014; El Houdaigui et al., 2019) have all already been well studied, no existing work did to our knowledge answer the question of the possible role of the TSC at both the whole-genome and evolutionary levels.

In this work, we have developed a genome-wide model of the influence of DNA SC on gene transcription, incorporating both a response to global SC level and the transcription-supercoiling coupling. We have shown that, in our model, the effect of the TSC is not limited to the local scale, but globally affects gene expression levels and enables a non-linear response to DNA SC changes due to environmental influences, via the selective activation or inhibition of specific genes. Furthermore, we have shown, using an *in silico* experimental evolution approach, that natural selection can leverage this biophysical mechanism to finely tune the expression levels of several pools of genes, in order to exhibit two qualitatively different phenotypes when exposed to different environments, as has been observed *in vivo* in pathogenic bacteria (Martis B. et al., 2019).

Our model voluntarily stays very simple, but it incorporates the most salient feature of the TSC, the non-linear interaction between neighboring genes; in our evolutionary simulations, complex phenotypes emerge despite the simplicity of the mutation operator, which does not affect genome length or basal gene expression levels. Integrating more dimensions to the model, such as promoter-specific responses to the SC level, inducible promoters, or more mutation operators such as gene deletions, duplications, or translocations, would doubtlessly yield interesting results. In order to bring this model closer to biology, a valuable approach would be to incorporate it into a larger framework, such as the Aevol *in silico* experimental evolution platform (Batut et al., 2013), in order to leverage the power of a well-understood digital model organism.

References

Batut, B., Parsons, D. P., Fischer, S., Beslon, G., and Knibbe, C. (2013). In silico experimental evolution: A tool to test evolutionary scenarios. *BMC Bioinformatics*, 14(Suppl 15):S11.

Blattner, F. R. (1997). The Complete Genome Sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462.

Brinza, L., Calevro, F., and Charles, H. (2013). Genomic analysis of the regulatory elements and links with intrinsic DNA structural properties in the shrunken genome of Buchnera. *BMC Genomics*, 14(1):73.

Crozat, E., Philippe, N., Lenski, R. E., Geiselmann, J., and Schneider, D. (2005). Long-Term Experimental Evolution in Escherichia coli . XII. DNA Topology as a Key Target of Selection. *Genetics*, 169(2):523–532.

Crozat, E., Winkworth, C., Gaffe, J., Hallin, P. F., Riley, M. A., Lenski, R. E., and Schneider, D. (2010). Parallel Genetic and Phenotypic Evolution of DNA Superhelicity in Experimental Populations of Escherichia coli. *Molecular Biology and Evolution*, 27(9):2113–2128.

Dorman, C. J. and Dorman, M. J. (2016). DNA supercoiling is a fundamental regulatory principle in the control of bacterial gene expression. *Biophysical Reviews*, 8(3):209–220.

El Houdaigui, B., Forquet, R., Hindré, T., Schneider, D., Nasser, W., Reverchon, S., and Meyer, S. (2019). Bacterial genome architecture shapes global transcriptional regulation by DNA supercoiling. *Nucleic Acids Research*, 47(11):5648–5657.

Junier, I. and Rivoire, O. (2016). Conserved Units of Co-Expression in Bacterial Genomes: An Evolutionary Insight into Transcriptional Regulation. *PLOS ONE*, 11(5):e0155740.

Krogh, T. J., Møller-Jensen, J., and Kaleta, C. (2018). Impact of Chromosomal Architecture on the Function and Evolution of Bacterial Genomes. *Frontiers in Microbiology*, 9:2019.

Lal, A., Dhar, A., Trostel, A., Kouzine, F., Seshasayee, A. S. N., and Adhya, S. (2016). Genome scale patterns of supercoiling in a bacterial chromosome. *Nature Communications*, 7(1):11055.

Lenski, R. E., Rose, M. R., Simpson, S. C., and Tadler, S. C. (1991). Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. *The American Naturalist*, 138(6):1315–1341.

Liu, L. F. and Wang, J. C. (1987). Supercoiling of the DNA template during transcription. *Proceedings of the National Academy of Sciences*, 84(20):7024–7027.

Ma, J. and Wang, M. D. (2016). DNA supercoiling during transcription. *Biophysical Reviews*, 8(S1):75–87.

Martis B., S., Forquet, R., Reverchon, S., Nasser, W., and Meyer, S. (2019). DNA Supercoiling: An Ancestral Regulator of Gene Expression in Pathogenic Bacteria? *Computational and Structural Biotechnology Journal*, 17:1047–1055.

Meyer, S. and Beslon, G. (2014). Torsion-Mediated Interaction between Adjacent Genes. *PLoS Computational Biology*, 10(9):e1003785.

Muskhelishvili, G., Forquet, R., Reverchon, S., Meyer, S., and Nasser, W. (2019). Coherent Domains of Transcription Coordinate Gene Expression During Bacterial Growth and Adaptation. *Microorganisms*, 7(12):694.

Peter, B. J., Arsuaga, J., Breier, A. M., Khodursky, A. B., Brown, P. O., and Cozzarelli, N. R. (2004). Genomic transcriptional response to loss of chromosomal supercoiling in Escherichia coli. *Genome Biology*, page 16.