Ride the supercoiling : Evolution of supercoiling-mediated gene regulatory networks through genomic inversions

Théotime Grohens

N°d'ordre NNT : 2022ISAL0126

THÈSE de DOCTORAT DE L'INSA LYON,
membre de l'Université de Lyon

ED 512
École Doctorale InfoMaths

Informatique et Applications

Soutenue publiquement le 14/12/2022, par :
Théotime Grohens

Ride the Supercoiling: Evolution of Supercoiling-Mediated Gene Regulatory Networks through Genomic Inversions

---

Devant le jury composé de :

Achaz, Guillaume	Professeur des Universités	Président	
Scornavacca, Céline	Directrice de recherche	CNRS	Rapportrice
Junier, Ivan	Chargé de Recherche HDR	CNRS	Rapporteur
Varoquaux, Nelle	Chargée de Recherche	CNRS	Examinatrice
Nasser, William	Directeur de Recherche	CNRS	Examinateur
Achaz, Guillaume	Professeur des Universités	Université Paris-Cité	Examinateur
Meyer, Sam	Maître de Conférences	INSA Lyon	Examinateur
Beslon, Guillaume	Professeur des Universités	INSA Lyon	Directeur de thèse

Département FEDORA – INSA Lyon - Ecoles Doctorales

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **https://www.edchimie-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage secretariat@edchimie-lyon.fr	**M. Stéphane DANIELE** C2P2-CPE LYON-UMR 5265 Bâtiment F308, BP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne **directeur@edchimie-lyon.fr**
E.E.A.	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** **https://edeea.universite-lyon.fr** Sec. : Stéphanie CAUVIN Bâtiment Direction INSA Lyon Tél : 04.72.43.71.70 secretariat.edeea@insa-lyon.fr	**M. Philippe DELACHARTRE** INSA LYON Laboratoire CREATIS Bâtiment Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne CEDEX Tél : 04.72.43.88.63 **philippe.delachartre@insa-lyon.fr**
E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** **http://e2m2.universite-lyon.fr** Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.e2m2@univ-lyon1.fr	**Mme Sandrine CHARLES** Université Claude Bernard Lyon 1 UFR Biosciences Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69622 Villeurbanne CEDEX **sandrine.charles@univ-lyon1.fr**
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** **http://ediss.universite-lyon.fr** Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.ediss@univ-lyon1.fr	**Mme Sylvie RICARD-BLUM** Institut de Chimie et Biochimie Moléculaires et Supramoléculaires (ICBMS) - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tél : +33(0)4 72 44 82 32 **sylvie.ricard-blum@univ-lyon1.fr**
INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** **http://edinfomaths.universite-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 infomaths@univ-lyon1.fr	**M. Hamamache KHEDDOUCI** Université Claude Bernard Lyon 1 Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tél : 04.72.44.83.69 **hamamache.kheddouci@univ-lyon1.fr**
Matériaux	**MATÉRIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec. : Yann DE ORDENANA Tél : 04.72.18.62.44 yann.de-ordenana@ec-lyon.fr	**M. Stéphane BENAYOUN** Ecole Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX Tél : 04.72.18.64.37 **stephane.benayoun@ec-lyon.fr**
MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** **http://edmega.universite-lyon.fr** Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bâtiment Direction INSA Lyon mega@insa-lyon.fr	**M. Jocelyn BONJOUR** INSA Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69621 Villeurbanne CEDEX **jocelyn.bonjour@insa-lyon.fr**
ScSo	**ScSo*** **https://edsciencessociales.universite-lyon.fr** Sec. : Mélina FAVETON INSA : J.Y. TOUSSAINT Tél : 04.78.69.77.79 melina.faveton@univ-lyon2.fr	**M. Bruno MILLY** Université Lumière Lyon 2 86 Rue Pasteur 69365 Lyon CEDEX 07 **bruno.milly@univ-lyon2.fr**

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Ride the Supercoiling: Evolution of Supercoiling-Mediated Gene Regulatory Networks through Genomic Inversions

Théotime Grohens

Théotime Grohens. *Epistasis, Supercoiling and Genome Structure*. 2021.
Generated using the Wombo DREAM text-to-image model,
using its title as a prompt.

Dis
C'est à quoi que ça sert?
Dis
C'est pourquoi qu'on s'en sert?
[...]
Si c'est à ça que ça sert
Si c'est à ça que ça sert
Si c'est comme ça qu'on s'en sert
Si c'est comme ça qu'on s'en sert
Je vais le faire et le refaire

Faire Et Refaire
Ascendant Vierge

Résumé en français

L'évolution des êtres vivants par sélection naturelle est souvent présentée comme un processus impossible à prédire, car elle trouve sa source dans les mutations aléatoires qui affectent le cœur du vivant, c'est-à-dire la molécule d'ADN, dont la séquence est le support principal de l'information biologique. Pourtant, s'il n'est pas possible d'identifier avec certitude quelles mutations vont survenir en réponse à une pression de sélection donnée, ni lesquelles parmi celles-ci vont être fixées, de nombreuses expériences laissent penser que le chemin que suit l'évolution n'est pas entièrement dû au hasard. Cette observation n'est pas nouvelle à l'échelle des organismes macroscopiques : Darwin la faisait déjà dans l'*Origine des Espèces* (Darwin, 1859). Il l'exposa ensuite plus en détail dans la *Variation des animaux et des plantes sous l'action de la domestication* (Darwin, 1868), en décrivant de nombreuses plantes et animaux sélectionnés et domestiqués par l'espèce humaine depuis des dizaines de milliers d'années. Ce caractère répétable est également observable à l'échelle des micro-organismes, avec l'apparition sans cesse renouvelée de résistances aux traitements d'infections bactériennes ou virales (Levy and Marshall, 2004). Ce n'est toutefois que depuis la fin du XXe siècle, avec le développement du séquençage ADN, que l'on est capable d'essayer de comprendre les soubassements de cette répétabilité, et donc de cette prédictibilité phénotypique, à l'échelle moléculaire. En effet, on peut désormais observer que c'est parfois le même gène, voire le même nucléotide à l'intérieur d'un gène, qui est touché par des mutations lorsqu'on répète une expérience de sélection pour une caractéristique donnée (Wortel et al., 2021). Dans ce cas, l'évolution ne semble ainsi plus pouvoir suivre une multitude de chemins différents pour parvenir au même résultat visible, mais semble au contraire contrainte de s'en tenir à un itinéraire bien défini.

L'un des mécanismes qui peuvent expliquer ce caractère répétable de l'évolution est l'épistasie, ou le rôle que joue le contexte génétique sur l'effet d'une mutation donnée. En effet, il est possible qu'une mutation ait un effet favorable en présence d'une autre mutation, mais un effet défavorable en l'absence de celle-ci. Ces relations épistatiques peuvent ainsi contraindre les options qui se présentent à l'évolution, en imposant qu'une mutation dans un gène donné survienne avant une autre dans un second gène, afin que la seconde soit favorable. Dans un contexte de compétition entre souches différentes au sein d'une même population (par exemple, de bactéries pathogènes), mieux comprendre ces relations épistatiques permettrait alors par exemple de prédire plus finement la fixation ou non de futures mutations, et par là la souche victorieuse, offrant la possibilité d'orienter plus finement un traitement. Le type de relations épistatiques le plus souvent étudié est celui des interactions entre mutations ponctuelles (c'est-à-dire entre mutations n'affectant qu'un seul nucléotide,

ou parfois quelques nucléotides contigus) à l'intérieur d'un même gène, car ce sont les mutations les plus faciles à détecter. Dans ce cas, le changement d'un acide aminé présent à un certain endroit de la protéine codée par le gène peut voir son effet modulé par le changement d'un autre acide aminé de la protéine. Ces relations épistatiques sont de mieux en mieux comprises, par exemple en mesurant exhaustivement la valeur sélective des 2^N mutants possibles pour un groupe de N nucléotides d'intérêt (voir Achaz et al. (2014) pour une vue d'ensemble de telles expériences). D'autres types d'interactions épistatiques, plus complexes et moins bien étudiés, existent toutefois. En particulier, il peut y avoir des interactions épistatiques entre différents types de mutations, comme entre mutations locales et réarrangements chromosomiques, ou entre gènes jouant des rôles de natures différentes, par exemple entre un gène codant pour une protéine régulatrice et un autre gène dont l'expression est régulée par cette protéine. Par exemple, la duplication d'une séquence à l'intérieur d'un gène donné peut être suivie d'une divergence et d'une spécialisation ultérieures de chacune des parties répétées, comme dans la famille des spectrines (Thomas et al., 1997), protéines qui jouent un rôle important dans la structure des cellules eucaryotes. Il y a alors épistasie entre un réarrangement chromosomique – la duplication d'une partie d'un gène – et les mutations ponctuelles qui la suivent : en l'absence de cette duplication, les mutations qui rendent possible la spécialisation des parties dupliquées du gène seraient en effet délétères.

Un cas particulier d'interactions épistatiques est celui des interactions entre les mutations dans les gènes régulant la superhélicité de l'ADN (que j'appellerai mutations de superhélicité) et les mutations dans les gènes eux-mêmes régulés par la superhélicité. La superhélicité de l'ADN, c'est-à-dire le niveau d'enroulement de l'ADN autour de lui-même, joue en effet un rôle important dans la régulation de la transcription des gènes chez les bactéries, car le niveau de transcription des gènes dépend directement de la superhélicité au niveau de leur promoteur. L'intérêt évolutif des mutations de superhélicité, ainsi que leur caractère répétable, ont été particulièrement mis en exergue grâce à la *Long Term Evolution Experiment* (*LTEE*) menée dans le laboratoire de Richard Lenski (Lenski et al., 1991). Dans cette expérience, 12 souches de la bactérie *Escherichia coli* évoluent depuis 1988 dans un environnement de laboratoire et 11 des 12 souches de l'expérience ont vu leur niveau de superhélicité augmenter très tôt au cours de celle-ci, grâce à des mutations touchant un faible nombre de gènes bien identifiés (Crozat et al., 2010). Comme le niveau de superhélicité est finement régulé par l'activité de plusieurs enzymes (appelées topoisomérases) et par la fixation de protéines sur l'ADN, une mutation dans un gène codant pour l'une ou l'autre de ces protéines peut en effet engendrer un changement de l'activité transcriptionnelle à l'échelle du génome entier. Les mutations répétées de superhélicité apparaissant dans cette expérience pourraient donc être le signe d'un paysage d'interactions épistatiques biaisé, qui augmenterait la proportion de mutations favorables pouvant apparaître dans les génomes qui les contiennent, rendant par là plus probable le futur succès évolutif de leur lignée.

Le questionnement majeur sous-tendant les travaux que j'ai menés pendant ma thèse a donc été de déterminer à quel point la présence ou non de mutations de superhélicité dans une lignée permet de prédire le futur succès de celle-ci, afin de comprendre plus généralement l'influence des biais épistatiques dans la répétabilité et la prédictibilité de l'évolution. Pour cela, j'ai étudié le rôle évolutif des mutations de superhélicité au cours de l'adaptation à un nouvel environnement, en employant une approche d'évolution expérimentale *in silico*, qui

s'inscrit dans le cadre plus large de la biologie évolutive des systèmes (Beslon et al., 2021). J'ai commencé par intégrer un modèle d'expression des gènes prenant en compte le niveau de superhélicité à l'échelle du chromosome dans un logiciel de simulation d'évolution existant au sein de mon équipe de thèse, le logiciel *Aevol*. Ce modèle et les premiers résultats obtenus à l'aide de celui-ci sont présentés dans le chapitre 3 de la thèse. Les expériences menées dans ce cadre ont permis d'obtenir un résultat évolutif qualitativement semblable à celui de la *LTEE*, la superhélicité étant la cible de nombreuses mutations au début de l'évolution. Toutefois, celle-ci se stabilise rapidement alors même que le reste du génome des individus continue d'évoluer, ne permettant pas de conclure sur de possibles interactions épistatiques.

Or, le rôle que tient le niveau de superhélicité de l'ADN dans la régulation de l'expression des gènes bactériens provient en réalité de son caractère extrêmement dynamique (Martis B. et al., 2019). Ce caractère dynamique de la superhélicité, tant dans le temps que le long du génome, est en particulier dû à la transcription elle-même des gènes (Visser et al., 2022). En effet, d'après un modèle initialement proposé par Liu and Wang (1987), lorsqu'un gène est en cours de transcription par une ARN polymérase, l'encombrant complexe qui en résulte ne peut pivoter autour de l'ADN aussi vite que l'ADN s'enroule autour de lui-même. Le couple ainsi exercé sur l'ADN provoque alors une accumulation de superhélicité en avant du gène transcrit et un déficit de superhélicité en arrière de celui-ci. La transcription d'un gène donné peut donc influer – par l'intermédiaire des changements locaux de superhélicité qu'elle engendre – sur la transcription des gènes à proximité de celui-ci et ainsi créer un réseau d'interactions entre les niveaux d'expressions de gènes proches sur le génome. Une modélisation de la superhélicité prenant en compte les variations locales de celle-ci dues à la transcription semble donc pertinente pour l'étude du rôle évolutif des mutations de superhélicité. Une approche aussi précise s'étant révélée délicate à mettre en place dans *Aevol*, j'ai opté pour la création d'un nouveau modèle représentant plus abstraitement le génome, mais décrivant plus fidèlement le couplage entre superhélicité et transcription. J'ai implémenté ce nouveau modèle – appelé *EvoTSC* – en Python. Le code de celui-ci est disponible à l'adresse suivante : `https://gitlab.inria.fr/tgrohens/evotsc`.

À l'aide d'*EvoTSC*, j'ai dans un premier temps montré que, dans un modèle où le seul mécanisme de régulation de l'activité des gènes est le couplage médié par la superhélicité entre les niveaux de transcription de gènes proches, et où les seules mutations possibles sont les inversions chromosomiques (qui réorganisent les positions relatives des gènes), il est possible d'obtenir par sélection naturelle des individus dont les gènes atteignent avec précision des niveaux d'expression optimaux dépendant de l'environnement. En particulier, il est possible d'obtenir des gènes activés par une relaxation globale de l'ADN, alors que leurs promoteurs sont intrinsèquement inhibés par la relaxation de l'ADN. Ces premiers résultats sont présentés dans le chapitre 4. Ils démontrent que la superhélicité peut jouer un rôle majeur dans la régulation de l'activité des gènes bactériens en permettant l'existence de réseaux de régulation génétique même en l'absence de facteurs de transcription. Ils ont été publiés, d'abord sous forme d'article dans la conférence *ALIFE 2021* (Grohens et al., 2021), puis dans une version étendue dans le journal associé, *Artificial Life* (Grohens et al., 2022b).

Dans un second temps, j'ai cherché à caractériser plus en détail l'impact évolutif de la superhélicité sur la structure des génomes et des réseaux de régulations bactériens. Toujours en utilisant le modèle *EvoTSC*, j'ai montré qu'au niveau le plus local, des paires convergentes

ou divergentes de gènes voisins se forment, conformément aux prédictions théoriques du couplage entre superhélicité et transcription. J'ai montré que cette organisation à l'échelle locale du génome n'était toutefois pas entièrement suffisante pour expliquer les niveaux d'expression des gènes observés dans le génome complet, mais que des sous-réseaux impliquant jusqu'à plusieurs dizaines de gènes peuvent au contraire être nécessaires. Enfin, en utilisant une approche par knock-out de gène, j'ai montré que, dans le génome des individus évolués, c'est sous la forme d'un réseau unique et s'étendant à l'échelle du génome entier que s'organise la régulation de l'expression des gènes dans le modèle *EvoTSC*. Ce second ensemble de résultats est présenté dans le chapitre 5 et a été mis en forme dans une prépublication qui sera prochainement soumise à relecture par les pairs (Grohens et al., 2022a).

Dans le chapitre 6, je présente ensuite un ensemble d'expériences complémentaires qui montrent la robustesse des résultats du modèle *EvoTSC* présentés dans les chapitres précédents, en réponse à des variations des principaux paramètres du modèle visant à représenter la diversité des génomes bactériens et des changements environnementaux. J'ai finalement incorporé dans *EvoTSC* un modèle d'évolution du niveau de superhélicité globale, afin de pouvoir caractériser, de la même manière que dans les expériences menées avec *Aevol*, les possibles relations épistatiques entre mutations de superhélicité et réarrangements chromosomiques. Je me suis en particulier intéressé à l'étude des paysages adaptatifs (*fitness landscapes*) résultant des mutations de superhélicité. Ces résultats sont présentés dans le chapitre 7 et ouvrent la voie à la conclusion de ce manuscrit au chapitre 8.

L'annexe A présente les contributions logicielles que j'ai réalisées tout au long de ma thèse. J'ai d'abord participé au développement d'*Aevol* et d'outils associés pour gérer des simulations et analyser les résultats obtenus avec celles-ci. Ensuite, j'ai développé le modèle *EvoTSC*, ainsi qu'un ensemble d'outils pour visualiser et analyser les données produites par le modèle.

Pour finir, l'irruption de la pandémie de Covid-19 en France au printemps 2020 a perturbé le cours de ma thèse d'une manière particulière. Je me suis en effet porté volontaire pour participer à une collaboration entre l'Assistance Publique-Hôpitaux de Paris (AP-HP) et un groupe de chercheur·ses et d'ingénieur·es Inria constitué à cet effet. Avec l'accord de mon directeur de thèse, j'ai alors interrompu mes travaux sur la superhélicité pour me consacrer pleinement à cet effort pendant plusieurs semaines en avril et mai 2020. Dans ce cadre, j'ai participé à la construction d'un modèle de l'épidémie de Covid-19 dans l'agglomération parisienne, visant à aider les équipes de l'AP-HP à suivre en temps réel – et essayer de prédire – l'évolution de cette épidémie à l'aide de données de régulation médicale. Ces travaux ont par la suite mené à une publication (Gaubert et al., 2020), présentée dans l'annexe B.

Table of Contents

Résumé en français i

Table of Contents v

List of Figures vii

List of Tables x

1 Introduction 1

2 Background 5
- 2.1 DNA Supercoiling in Bacteria . . . 5
  - 2.1.1 Gene Regulation by DNA Supercoiling . . . 6
  - 2.1.2 A Dynamic DNA Supercoiling Level . . . 7
  - 2.1.3 Supercoiling and Evolution . . . 7
- 2.2 The Transcription-Supercoiling Coupling . . . 8
- 2.3 Existing Models of the Transcription-Supercoiling Coupling . . . 9
- 2.4 An Evolutionary Systems Biology Approach . . . 10
- 2.5 Conclusion . . . 12

3 Looking for Supercoiling Epistasis in *Aevol* 13
- 3.1 Introduction . . . 13
- 3.2 The *Aevol* model . . . 14
  - 3.2.1 Overview . . . 14
  - 3.2.2 The Genotype-Phenotype Map in *Aevol* . . . 16
  - 3.2.3 Fitness . . . 17
  - 3.2.4 Mutational Operators . . . 17
- 3.3 Modeling DNA Supercoiling in *Aevol* . . . 17
  - 3.3.1 Level of DNA Supercoiling . . . 17
  - 3.3.2 Gene Expression . . . 18
  - 3.3.3 Mutational Operator . . . 19
- 3.4 Results . . . 19
  - 3.4.1 Experimental Setup . . . 19
  - 3.4.2 Studying Lineages . . . 20
  - 3.4.3 Evolution of the Fitness Level . . . 20

3.4.4 Evolution of the Supercoiling Level . . . 21
3.4.5 Looking for Epistasis . . . 23
3.5 Conclusion . . . 25

4 Evolution of Environmental Sensing through DNA Supercoiling 27
4.1 A Genome-Wide Model of the Transcription-Supercoiling Coupling . . . 28
4.1.1 Mathematical Description of the Model . . . 29
4.1.2 Effect of the Environmental Supercoiling on Gene Activation Levels . 32
4.1.3 Influence of Relative Gene Positions on Gene Activation Levels . . . 33
4.2 An Evolutionary Genome-Wide Model of the Transcription-Supercoiling Coupling . . . 34
4.2.1 Evolutionary Model: Evolution in Two Separate Environments . . . 34
4.2.2 Fitness . . . 34
4.2.3 Mutational Operator: Genomic Inversions . . . 35
4.2.4 Experimental Setup and Parameter Values . . . 36
4.2.5 Adaptation of Gene Expression Levels to Different Environments . . 37
4.2.6 Robustness of Gene Network Evolution . . . 41
4.3 Discussion and Perspectives . . . 43

5 Structure of Supercoiling-Mediated Gene Regulatory Networks 47
5.1 An Evolutionary Model of the Transcription-Supercoiling Coupling . . . 48
5.1.1 Individual-Level Model . . . 48
5.1.2 Evolutionary Model . . . 50
5.2 Results . . . 52
5.2.1 Experimental Setup . . . 52
5.2.2 Evolution of Regulation by the Transcription-Supercoiling Coupling 53
5.2.3 Evolution of Local Genome Organization . . . 57
5.2.4 Local Interactions Do Not Recapitulate the Regulatory Network . . . 59
5.2.5 A Whole-Genome Gene Regulatory Network . . . 62
5.3 Discussion and Perspectives . . . 66
5.4 Conclusion . . . 68

6 Evaluating the Robustness of the *EvoTSC* Model 69
6.1 Interaction Distance . . . 70
6.2 Mean Intergenic Size . . . 73
6.3 Environmental Shift in Supercoiling . . . 78
6.4 Number of Genes . . . 81
6.5 Introducing Indels . . . 85
6.6 Discussion . . . 88

7 Looking for Supercoiling Epistasis in *EvoTSC* 91
7.1 Experimental Framework . . . 92
7.1.1 Introducing Supercoiling Mutations . . . 92
7.1.2 Environmental Shock . . . 94

7.1.3 Experimental Protocol . . . 95
7.2 Results . . . 96
7.2.1 Evolution after an Environmental Shock . . . 96
7.2.2 Supercoiling Fitness Landscapes . . . 98
7.2.3 Evolution with Supercoiling Mutations Only . . . 102
7.3 Discussion . . . 105

8 Conclusion **107**
8.1 Summary . . . 107
8.2 Perspectives . . . 109

A Software Contributions **111**
A.1 *Aevol* . . . 111
A.1.1 DNA Supercoiling in *Aevol* . . . 111
A.1.2 Tooling for *Aevol* . . . 111
A.2 *EvoTSC* . . . 112
A.2.1 Technical Description . . . 112
A.2.2 Tooling . . . 113
A.2.3 Use . . . 113

B Covid-19 Task Force **115**

Bibliography **147**

List of Figures

2.1 Role of supercoiling in transcription and description of the transcription-supercoiling coupling . . . 6
2.2 Template of an evolutionary systems biology simulation . . . 11

3.1 Overview of the *Aevol* model . . . 15
3.2 Effect of supercoiling on the phenotype of an individual in *Aevol* . . . 18
3.3 Evolution of the fitness of the control and experimental runs in *Aevol* . . . 21
3.4 Evolution of the supercoiling level of the experimental runs in *Aevol* . . . 22
3.5 Measuring epistasis with the average times before and after mutations . . . 24

4.1 Hand-drawn genome and local interactions resulting from the TSC . . . 28
4.2 Example individual in the proof-of-concept model . . . 30
4.3 Influence of environmental supercoiling on the phenotype of the example individual in Figure 4.2 . . . 32

4.4 Effect of a genomic inversion on the example individual in Figure 4.2 33
4.5 Effect of an inversion on the hand-drawn genome in Figure 4.1 36
4.6 Fraction of activated genes of each type at the end of evolution in the proof-of-concept model 38
4.7 Fitness of every replicate during evolution in the proof-of-concept model 38
4.8 Number of activated genes during evolution in one of the replicates in the proof-of-concept model 39
4.9 Best individual at the end of evolution in one of the replicates in the proof-of-concept model, evaluated in both environments 40
4.10 Parameter exploration in the proof-of-concept model: varying ε 41
4.11 Parameter exploration in the proof-of-concept model: varying c 42
4.12 Parameter exploration in the proof-of-concept model: varying σ_A and σ_B 43

5.1 Example individual in the advanced model, evaluated in both environments . . . 48
5.2 Average fitness during evolution in the advanced model 53
5.3 Best individual at the end of evolution in one of the replicates in the advanced model, in both environments 54
5.4 Average number of activated genes during evolution in the advanced model . . 55
5.5 Average gene expression as a function of background supercoiling at the end of evolution in the advanced model 56
5.6 Number of gene pairs and supercoiling effect per type of gene pair 58
5.7 Example minimal subnetworks needed for gene inhibition in an evolved individual 60
5.8 Minimal subnetwork size needed to obtain the correct activation state per gene type 61
5.9 Example evolved individual with a knocked-out gene, evaluated in both environments 63
5.10 Effective interaction graph of an evolved individual, and distribution of effective interaction graph WCCs in evolved and random individuals 64
5.11 Average in- and out-degree of effective interaction graph nodes for evolved and random individuals 65

6.1 Evolution of the number of activated genes in each environment, with an interaction distance of 25 kb 70
6.2 Average fitness during evolution, with an interaction distance of 25 kb 71
6.3 Average gene expression as a function of background supercoiling, with an interaction distance of 25 kb 71
6.4 Average gene expression as a function of background supercoiling, with increasing interaction distances, in random genomes 72
6.5 Average in- and out-degree of effective interaction graph nodes, with an interaction distance of 25 kb 73
6.6 Evolution of the number of activated genes in each environment, with increasing mean intergenic distances 74
6.7 Average fitness during evolution, with increasing mean intergenic distances . . 75

6.8 Average gene expression as a function of background supercoiling, with an intergenic distance of 10 kb . . . 76
6.9 Average gene expression as a function of background supercoiling, with increasing mean intergenic distances, in random genomes . . . 77
6.10 Evolution of the number of activated genes in each environment, with decreasing environmental supercoiling shifts . . . 78
6.11 Average fitness during evolution, with decreasing environmental supercoiling shifts . . . 79
6.12 Average gene expression as a function of background supercoiling, with an absolute environmental supercoiling shift of 0.001 . . . 80
6.13 Average gene expression as a function of background supercoiling, with an absolute environmental supercoiling shift of 0.0001 . . . 81
6.14 Evolution of the number of activated genes in each environment, with a 300-gene genome . . . 82
6.15 Average fitness during evolution, with a 300-gene genome . . . 83
6.16 Average gene expression as a function of background supercoiling, with a 300-gene genome . . . 84
6.17 Average intergenic size during evolution, with indels . . . 86
6.18 Evolution of the number of activated genes in each environment, with indels . . . 86
6.19 Average fitness during evolution, with indels . . . 87
6.20 Average gene expression as a function of background supercoiling, with indels . . . 87

7.1 Average basal supercoiling and fitness during evolution of the wild-types, with basal supercoiling level mutations . . . 93
7.2 Evolved wild-type individual before and after an environmental shock . . . 94
7.3 Evolution of the number of activated genes in each environment, with a . . . 96
7.4 Average fitness relative to the ancestor and basal supercoiling, during evolution after an environmental shock . . . 97
7.5 Supercoiling fitness landscapes for the wild-type individuals evolved with and without supercoiling mutations . . . 99
7.6 Supercoiling fitness landscapes after an environmental shock, with and without supercoiling mutations . . . 100
7.7 Supercoiling fitness landscapes after evolution after an environmental shock, with and without supercoiling mutations . . . 101
7.8 Average relative fitness during evolution after an environmental shock, with only supercoiling mutations . . . 103
7.9 Evolution of the basal supercoiling level in shocked individuals with supercoiling mutations only . . . 103
7.10 Fitness landscapes with only supercoiling mutations . . . 104

List of Tables

3.1 Table of parameter values for the *Aevol* runs 20

5.1 Table of parameter values used for the advanced *EvoTSC* runs 53

6.1 Table of parameter values explored in additional *EvoTSC* simulations 69

Chapter 1

Introduction

Note: this chapter is an English translation of the French summary at the start of the thesis.

The evolution of living organisms by natural selection is often presented as a process that is impossible to predict, because it finds its source in the random mutations that affect the heart of living beings: the DNA molecule, whose sequence is the main carrier of biological information. However, if it is not possible to identify with certainty which mutations will occur in response to a given selection pressure, nor which of these mutations will be fixed, many experiments suggest that the course followed by evolution is not entirely random. This observation is not new at the scale of macroscopic organisms, as Darwin already made it in the *Origin of Species* (Darwin, 1859). He then exposed it in more detail in the *Variation of Animals and Plants under Domestication* (Darwin, 1868), in which he describes numerous plant and animal species selected and domesticated by humans over the last tens of thousands of years. This repeatability is also observable at the scale of microorganisms, with the constantly renewed appearance of resistance to treatments of bacterial or viral infections (Levy and Marshall, 2004). However, it is only since the end of the 20th century, with the development of DNA sequencing, that we have been able to try and understand the basis of this repeatability – and therefore of this phenotypic predictability – at the molecular level. Indeed, we can now observe that it is sometimes the same gene, or even the same nucleotide within a gene, that is affected by mutations when a selection experiment for a given trait is repeated. In this case, evolution no longer seems to be able to follow a multitude of different paths to the same visible result, but instead seems to be forced to stick to a well-defined path.

One mechanism that may explain this repeatability of evolution is epistasis, or the role that the genetic context plays on the effect of a given mutation. Indeed, it is possible for a mutation to have a favorable effect in the presence of another mutation, but an unfavorable effect in its absence. These epistatic relationships can thus constrain the options that are available to evolution, by requiring that a mutation in one gene occurs before another mutation in a second gene, so that the second mutation is favorable. In a context of competition between different strains within the same population (for example, of pathogenic bacteria), a better understanding of these epistatic relationships would allow, for example, to predict more accurately the fixation or not of future mutations, and thus the winning strain, offering the possibility to guide treatments more accurately. The most often studied type of epistatic

relationships is that of interactions between point mutations (i.e., mutations affecting only one nucleotide, or sometimes a few contiguous nucleotides) within the same gene, as these mutations are the easiest to detect. In this case, changing an amino acid present at a certain position in the protein coded by the gene can modulate the effect of changing another amino acid in the protein. These epistatic relationships are now better and better understood, for example by exhaustively measuring the selective value of the 2^N possible mutants for a group of N nucleotides of interest (see Achaz et al. (2014) for an overview of such experiments). Other more complex and less well-studied types of epistatic interactions however exist. In particular, there may be epistatic interactions between different types of mutations, such as between local mutations and chromosomal rearrangements, or between genes playing different kinds of roles, for example between a gene encoding a regulatory protein and another gene whose expression is regulated by that protein. For example, the duplication of a sequence within a given gene can be followed by a subsequent divergence and specialization of each of the repeated parts, such as in the spectrin family of proteins (Thomas et al., 1997), which play an important role in the structure of eukaryotic cells. In that case, there is epistasis between a chromosomal rearrangement – the duplication of a part of a gene – and the point mutations which follow this duplication: in the absence of this duplication, the mutations which make possible the specialization of the duplicated parts of the gene would have been deleterious.

A particular case of epistatic interactions is the interactions between mutations in genes which regulate DNA supercoiling (which I will call supercoiling mutations throughout this manuscript) and mutations in genes themselves regulated by supercoiling. DNA supercoiling, i.e. the level of twisting and writhing of DNA around itself, indeed plays an important role in the regulation of gene transcription in bacteria, because the transcription level of a gene depends directly on the level of supercoiling at its promoter. The evolutionary interest of supercoiling mutations, as well as their repeatability, has been particularly highlighted in the *Long Term Evolution Experiment* (*LTEE*) conducted in the laboratory of Richard Lenski (Lenski et al., 1991). In this experiment, 12 strains of the bacterium *Escherichia coli* have been evolving since 1988 in a laboratory environment, and 11 of the 12 strains in the experiment have seen their level of supercoiling increase very early in the experiment, thanks to mutations affecting a small number of well identified genes (Crozat et al., 2010). Since the level of supercoiling is finely regulated by the activity of several enzymes (called topoisomerases) and by the binding of nucleoid-associated proteins to DNA, a mutation in a gene coding for any of these proteins might lead to a genome-wide change in transcriptional activity. The repeated supercoiling mutations that appear in this experiment could therefore indicate a biased epistatic interaction landscape, which would increase the proportion of favorable mutations that can appear in the genomes that contain these supercoiling mutations. Such biased epistatic relationships could thereby make the future evolutionary success of the lineages that bear these supercoiling mutations more likely.

The major question underlying my PhD work was therefore to determine to which extent the presence or absence of supercoiling mutations in a lineage can predict its future evolutive success, in order to understand more broadly the influence of epistatic biases in the repeatability and predictability of evolution. To this end, I investigated the evolutionary role of supercoiling mutations during adaptation to a new environment through an *in silico*

experimental evolutionary approach, situated within the broader framework of evolutionary systems biology (Beslon et al., 2021). I started by integrating a model of gene expression that describes the level of supercoiling at the whole-chromosome scale into *Aevol*, an artificial evolution software platform that has been developed in my research team. This model and the first results obtained with it are presented in Chapter 3 of the thesis. The experiment carried out in this framework led to results qualitatively similar to that of the *LTEE*, in that the supercoiling level was the target of many mutations at the beginning of the evolution. However, the supercoiling rapidly stabilized while the rest of the genome of the individuals continued to evolve, making it impossible to conclude on possible epistatic interactions within these experiments.

The role that the level of DNA supercoiling plays in regulating bacterial gene expression indeed actually stems from its extremely dynamic character (Martis B. et al., 2019). The dynamic character of supercoiling, both in time and along the genome, is in particular due to gene transcription itself (Visser et al., 2022). Indeed, according to a model originally proposed by Liu and Wang (1987), when a gene is being transcribed by an RNA polymerase, the bulky complex that results cannot rotate around DNA as fast as DNA wraps around itself. The torque exerted by this complex on DNA thus causes an accumulation of supercoiling in front of the transcribed gene, and a deficit of supercoiling behind the gene. The transcription of a given gene can therefore influence – through the local changes in supercoiling that it generates – the transcription of genes near that gene, and thus create a network of interactions between the expression levels of nearby genes in the genome. Modeling supercoiling in order to take into account the local variations of supercoiling caused by transcription therefore seems particularly relevant for the study of the evolutionary role of supercoiling mutations. As such a precise approach proved to be difficult to implement in *Aevol*, I opted for the creation of a new model representing the genome more abstractly, but describing the coupling between supercoiling and transcription more faithfully. I implemented this new model – called *EvoTSC* – in Python, and its code is available at the following address: `https://gitlab.inria.fr/tgrohens/evotsc`.

Using EvoTSC, I first showed that, in a model where the only mechanism for regulating gene activity is the supercoiling-mediated coupling between the transcription levels of nearby genes, and where the only possible mutations are chromosomal inversions (which rearrange the relative positions of genes on the genome), it is possible to obtain through natural selection individuals whose genes precisely reach environment-dependent target expression levels. In particular, it is possible to obtain genes that are activated by global DNA relaxation, even though their promoters are intrinsically inhibited by DNA relaxation. These first results are presented in Chapter 4. They demonstrate that supercoiling can play a major role in the regulation of bacterial gene activity, by enabling the emergence of gene regulatory networks even in the absence of transcription factors. These results were first published as a paper in the *ALIFE 2021* conference (Grohens et al., 2021), and then in an extended version in the associated journal *Artificial Life* (Grohens et al., 2022b).

In a second step, I sought to characterize in more detail the evolutionary impact of supercoiling on the structure of bacterial genomes and regulatory networks. Still using the *EvoTSC* model, I showed that at the most local level, convergent or divergent pairs of neighboring genes are formed, in accordance with the theoretical predictions of the transcription-

supercoiling coupling. I showed that this organization at the local genome scale is however not entirely sufficient to explain the gene expression levels observed in the whole genome, but that sub-networks involving up to dozens of genes may instead be required. Finally, using a gene knockout approach, I showed that in the genome of evolved individuals, the regulation of gene expression in the EvoTSC model is organized as a single genome-wide network. This second set of results is presented in Chapter 5 and has been written up in a pre-publication that will be submitted for peer-review (Grohens et al., 2022a).

In Chapter 6, I then present a set of complementary experiments that underline the robustness of the results that I described in the previous chapters in response to variations in the main parameters of the model, which aim at representing the diversity of bacterial genomes and possible environmental perturbations. Finally, I incorporated into *EvoTSC* a model of the evolution of the global supercoiling level, in order to be able to characterize, in the same way as in the experiments carried out with *Aevol*, the possible epistatic relationships between supercoiling mutations and chromosomal rearrangements. In particular, I studied the fitness landscapes that stem from supercoiling mutations in the model. These results are presented in Chapter 7 and set the stage for the conclusion of this manuscript in Chapter 8.

Appendix A presents the software contributions that I made throughout my PhD. I first participated in the development of the *Aevol* framework and of associated tools that manage simulations and analyze their results. I then developed the *EvoTSC* model, as well as a set of tools to visualize and analyze the data produced by the model.

Finally, the outbreak of the Covid-19 pandemic in France in the spring of 2020 disrupted the course of my PhD in a particular way. At that time, I volunteered to participate in a collaboration between the Assistance Publique-Hôpitaux de Paris (AP-HP) and a group of Inria researchers and engineers formed for this purpose. With the agreement of my supervisor, I interrupted my work on supercoiling to fully dedicate myself to this effort for several weeks in April and May 2020. In this context, I participated in the construction of a model of the Covid-19 epidemic in the Paris area, which aimed at helping the AP-HP teams to follow in real time – and try to predict – the evolution of this epidemic using medical regulation data. This work subsequently led to a peer-reviewed publication (Gaubert et al., 2020), which is presented in Appendix B.

Chapter 2

Background

In this chapter, I introduce the biological concepts and methods that will be used throughout this manuscript. I first present DNA supercoiling and its regulation in bacteria. I outline its role in gene transcription, and the reciprocal effect of transcription on supercoiling, which jointly result in what is called the transcription-supercoiling coupling (TSC). Then, I discuss a few cases studies in which supercoiling might have played an important evolutionary role, and that illustrate the interest of studying DNA supercoiling through the lens of evolution. Finally, I briefly present the general method with which I tackle the questions raised in Chapter 1 throughout the manuscript.

2.1 DNA Supercoiling in Bacteria

DNA is the material basis of genetic information. It is a flexible polymer that comprises two strands of nucleotides that coil around each other, at a rate of 10.5 base pairs per turn in the absence of external constraints. When subjected to torsional stress, DNA can either writhe and form 3-dimensional loops, or twist around itself more or less tightly than in its relaxed state (Travers and Muskhelishvili, 2005); both writhing and twisting are referred to as DNA supercoiling. The level of supercoiling is measured as the relative density σ of supercoils in over- or under-wound DNA, as compared to relaxed DNA. DNA is positively supercoiled ($\sigma > 0$) when it is overwound, and negatively supercoiled ($\sigma < 0$) when it is underwound. In bacteria, DNA is normally maintained in a moderately negatively supercoiled state, with a reference value of $\sigma_{basal} = -0.06$ in *Escherichia coli* (Travers and Muskhelishvili, 2005). In these organisms, the supercoiling level is an important regulator of gene transcription (Dorman and Dorman, 2016). Moreover, as transcription itself impacts DNA supercoiling (Liu and Wang, 1987), this results in a coupling between these two processes, which Figure 2.1 presents an overview of. As a general rule, genes are transcribed at a higher rate when DNA is more negatively supercoiled (A), following a sigmoidal response curve (B). Transcription generates positive and negative supercoiling downstream and upstream of transcribed genes respectively (C), resulting in a coupling between the expression levels of neighboring genes that depends on their relative orientations (D).

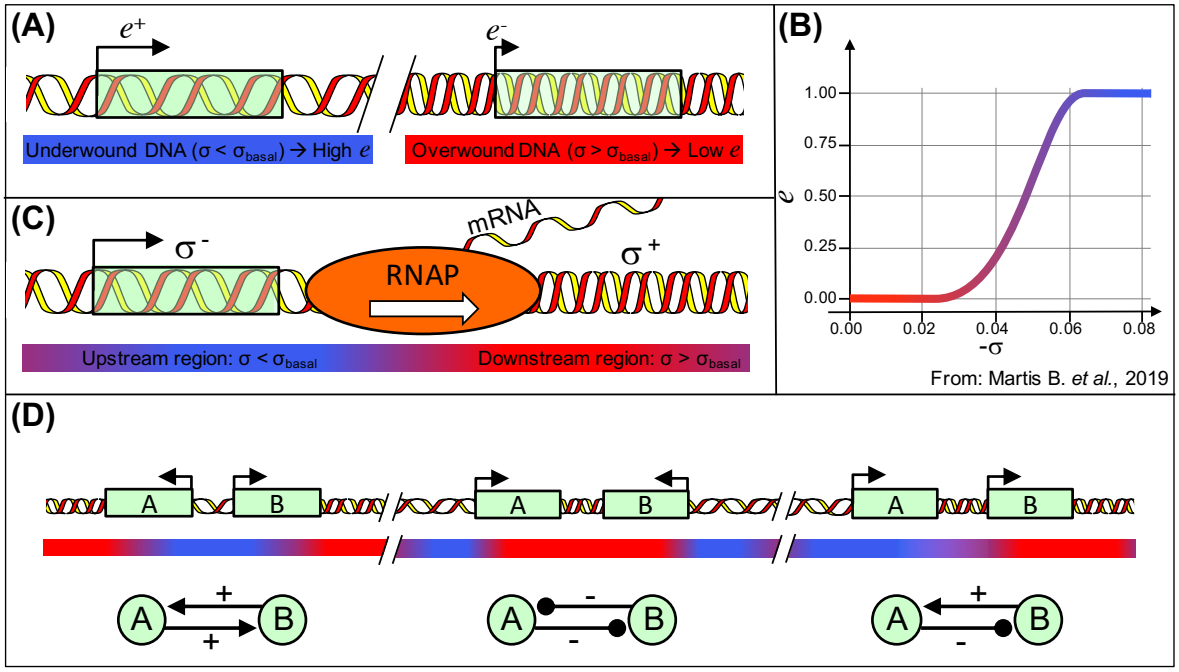


Figure 2.1: **A**. When DNA is underwound ($\sigma < \sigma_{basal}$, left), gene transcription rates are higher than when DNA is overwound ($\sigma > \sigma_{basal}$, right). **B**. Promoter activity (equivalently, transcription level) e increases with the level of negative supercoiling $-\sigma$. **C**. The transcription of a gene by RNA polymerase (RNAP) generates a decrease in supercoiling upstream of the transcribed gene, and an increase downstream of the transcribed gene. **D**. Transcription-supercoiling coupling: the sign of the interaction between neighboring genes depends on their relative orientation. Figure reproduced from Grohens et al. (2021).

2.1.1 Gene Regulation by DNA Supercoiling

The level of DNA supercoiling influences gene expression, as more negatively supercoiled DNA facilitates the initiation of transcription (Figure 2.1 A). The thermodynamical reaction of opening the DNA double strand, which is the initial step of gene transcription, is indeed favored in more negatively supercoiled DNA (El Houdaigui et al., 2019), resulting in a sigmoidal response curve of gene expression to DNA supercoiling (Figure 2.1 B).

Due to this effect, supercoiling has experimentally been shown to act as a broad regulator of gene expression in several bacteria. In *E. coli*, Peter et al. (2004) showed that 7% of genes were sensitive to a relaxation of chromosomal DNA, of which one third were up-regulated by relaxation and two thirds down-regulated. Similar results were obtained for *S. enterica*, in which 10% of genes were sensitive to DNA relaxation (Webber et al., 2013), and for *S. pneumoniae*, in which around 13% of genes were sensitive to relaxation (Ferrandiz et al., 2010). When instead inducing extreme negative supercoiling in *D. dadantii*, 13% of the genes in the exponential phase and 7% in the stationary phase were affected (Pineau et al., 2022).

In *D. dadantii*, different genomic regions moreover exhibit markedly different responses to changes in supercoiling (Muskhelishvili et al., 2019), allowing the expression of pathogenic

genes only in stressful environments. Finally, DNA supercoiling might be an especially important regulator of gene activity in bacteria with reduced genomes, such as the obligate aphid endosymbiotic bacterium *Buchnera aphidicola*. *B. aphidicola* is nearly devoid of transcription factors, and supercoiling is therefore thought to be one of the sole regulation mechanisms available in this bacteria (Brinza et al., 2013).

2.1.2 A Dynamic DNA Supercoiling Level

The level of DNA supercoiling in bacteria is primarily controlled by topoisomerases, enzymes that alter DNA supercoiling by cutting and rotating the DNA strands (Duprey and Groisman, 2021). The two main topoisomerases are gyrase, which dissipates positive supercoiling by introducing negative supercoils at an ATP-dependent rate, and topoisomerase I, which oppositely relaxes negative supercoiling (Martis B. et al., 2019). But numerous other processes also impact the level of DNA supercoiling, either by generating new supercoils or by constraining their diffusion.

In particular, according to the *twin-domain* model of supercoiling (Liu and Wang, 1987), the transcription of a gene by RNA polymerase generates both positive and negative supercoils. As a consequence of the drag that hampers the rotation of the RNA polymerase complex around the DNA sequence during transcription, positive supercoiling builds up upstream of the transcribed gene, and negative supercoiling downstream of the transcribed gene (Visser et al., 2022). This phenomenon is pictured in subfigure C of Figure 2.1. Moreover, while the intrinsic flexibility of the DNA polymer would in principle allow supercoils to propagate freely along the chromosome, many nucleoid-associated proteins such as FIS, H-NS or HU bind to bacterial DNA (Krogh et al., 2018), in addition to RNA polymerases. These DNA-bound proteins create barriers that block the diffusion of supercoils, resulting in what have been named topological domains of supercoiling (Postow et al., 2004).

The level of DNA supercoiling can furthermore be affected by numerous environmental stresses in bacteria. Salt shock transiently increases negative DNA supercoiling in *E. coli* (Hsieh et al., 1991); the acidic intracellular environment relaxes DNA in the facultative pathogen *Salmonella enterica* var. Typhimurium (Marshall et al., 2000); and higher temperatures relax DNA in the plant pathogen *Dickeya dadantii* (Hérault et al., 2014). These constraints overall paint the picture of a very dynamic DNA "supercoiling landscape" in bacteria (Visser et al., 2022), with a supercoiling level that varies in both time and space during the bacterial lifecycle and along the chromosome.

2.1.3 Supercoiling and Evolution

Gene regulation by DNA supercoiling can itself be subject to evolution by natural selection, as a mechanism through which gene expression levels can be adapted to new environments. In the *Long-Term Evolution Experiment* (*LTEE*) (Lenski et al., 1991), 12 populations of *E. coli* have been maintained for over 80,000 generations, evolving and adapting to a glucose-limited environment. In 11 of the 12 populations in the experiment, an increase in fitness was linked to mutations in genes which participate directly or indirectly in the regulation of the supercoiling level, such as *topA*, *fis*, or *dusB* (Crozat et al., 2010). When inserted into the ancestral

strain, the mutant *topA* and *fis* alleles increased the level of negative supercoiling as well as the bacterial growth rate, demonstrating that supercoiling mutations can play a role in the adaptation to new environments through their broad regulatory effect (Crozat et al., 2005). From an epistasis perspective, the repeated fixation of supercoiling mutations in the *LTEE* suggests that these mutations could confer an evolutionary advantage on the lineages in which they appear by favoring the apparition of compensatory mutations in supercoiling-regulated genes; but this possible epistatic role should nonetheless be disentangled from their direct fitness effect in order to draw clear conclusions.

The regulation of gene expression by DNA supercoiling could moreover be a force that participates in shaping the evolution of the organization itself of bacterial genomes. Indeed, supercoiling-sensitive genes tend to group in up- or down-regulated clusters in *E. coli* (Peter et al., 2004), *S. enterica* (Webber et al., 2013) and *S. pneumoniae* (Ferrandiz et al., 2010). This suggests the possibility of a phenotypic role in the co-localization of genes in these clusters, through a common regulation of their transcription (Sobetzko, 2016). Synteny segments, or clusters of neighboring genes that show correlated expression patterns, are indeed evolutionarily conserved across *E. coli* and the distantly related *Bacillus subtilis*, strengthening the hypothesis that these domains could play an important role in the regulation of bacterial gene expression through supercoiling-mediated interactions (Junier and Rivoire, 2016).

2.2 The Transcription-Supercoiling Coupling

As shown in Figure 2.1C, the transcription of a given gene by an RNA polymerase generates an accumulation of positive supercoiling downstream of that gene, and of negative supercoiling upstream of that gene, because of the hindered movement of the polymerase (Liu and Wang, 1987; Visser et al., 2022). If a second gene is located closely enough to this first gene on the genome, the change in supercoiling at the location of the promoter of the second gene will impact the transcription rate of that gene, as negative supercoiling usually facilitates gene transcription (Forquet et al., 2021). In turn, the transcription of the second gene will also generate a local change in supercoiling that affects the first gene, resulting in an interaction between the transcription levels of these two genes, which has been called the transcription-supercoiling coupling (Meyer and Beslon, 2014). Depending on the relative orientation of these genes, the coupling can take several forms. Divergent genes increase their respective transcription level in a positive feedback loop; convergent genes inhibit the transcription of one another; and in tandem genes, the transcription of the downstream gene increases the transcription of the upstream gene, while the transcription of the upstream gene decreases the transcription of the downstream gene.

This supercoiling-mediated interaction between neighboring genes has been documented in several bacterial genetic systems. In the *E. coli*-related pathogen *Shigella flexneri*, the *virB* promoter is normally only active at high temperatures, but can be activated at low temperatures by the insertion of a phage promoter in divergent orientation (Tobe et al., 1995). Similarly, the expression of the *leu-500* promoter in *S. enterica* can be increased or decreased by the insertion of upstream transcriptionally active promoters, depending on their orientation relative to *leu-500* (El Hanafi and Bossi, 2000). The magnitude of the effect of the

transcription-supercoiling coupling has also been explored in a synthetic construct, in which the inducible *ilvY* and *ilvC E. coli* promoters have been inserted on a plasmid in divergent orientations. In this system, a decrease in the activity of *ilvY* is associated with a decrease in *ilvC* activity, and an increase in *ilvY* activity with an increase in *ilvC* activity as well (Rhee et al., 1999).

There are, however, hints that the biological relevance of the transcription-supercoiling coupling might not be confined to these few specific instances. Indeed, in *E. coli*, the typical size of topological domains – inside which the positive and negative supercoils generated by gene transcription can propagate – is usually estimated to measure around 10 kb (Postow et al., 2004), while transcription-generated supercoiling could propagate up to 25 kb in each direction around a transcribed gene (Visser et al., 2022). As genes measure on average 1 kb and intergenic distances 120 bp in *E. coli* (Blattner, 1997), any single topological domain on the *E. coli* chromosome therefore encompasses multiple genes that can potentially interact via the transcription-supercoiling coupling. A statistical analysis of the relative position of neighboring genes on the *E. coli* chromosome indeed shows that genes that are up-regulated by negative supercoiling have more neighbors in divergent orientations, while genes that are down-regulated by negative supercoiling have more neighbors in converging orientations (Sobetzko, 2016), further suggesting that the transcription-supercoiling coupling plays a role in regulating the activity of genes located in the same topological domain.

2.3 Existing Models of the Transcription-Supercoiling Coupling

Several mathematical and computational models have been proposed to describe the effect of the transcription-supercoiling coupling on the expression level of neighboring genes. In Meyer and Beslon (2014), a quantitative model of the supercoiling level at a locus of interest is proposed, in order to study the transcription-supercoiling coupling in a pair of adjacent genes. In that model, DNA transcription is regulated by the opening free energy of DNA around gene promoters, which directly depends on the supercoiling level. The reciprocal influence of neighboring genes is then obtained by computing the difference in transcription levels due to supercoiling and the subsequent variation in supercoiling, and iterating this system until a fixed point is reached. A more detailed stochastic model is presented in El Houdaigui et al. (2019). This model aims at making quantitative predictions of gene expression levels, and introduces explicit RNA polymerases and topoisomerases that delineate dynamic supercoiling domains inside which supercoils immediately propagate. In that model, 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 supercoiling in the downstream domain and negative supercoiling in the upstream domain, and finally detach from DNA, merging the two domains separated by the polymerase.

Another family of biophysical models aims at describing the movement of RNA polymerases along the genome during gene transcription and therefore model the level of DNA

supercoiling, as supercoils impact the speed at which polymerases can progress forward. In Brackley et al. (2016), a stochastic model of the transcription of co-oriented genes is proposed in order to study transcriptional bursts. This model is qualitatively different from the models presented above, as it explicitly models the level of supercoiling as a function of time and position along DNA, whereas the former models consider supercoiling to be constant in the intervals delimited by polymerases or nucleoid-associated proteins. A similar model is introduced in Sevier and Levine (2017), in order to study the possible stalling of DNA polymerases due to excessive transcription-generated supercoiling in a single gene. This second model has then been extended to accommodate the supercoiling-mediated interaction of neighboring genes in Sevier and Levine (2018), making qualitatively similar predictions of gene transcription rates as the first set of models presented above. This model has finally been used to propose a toggle switch (Gardner et al., 2000) in which gene regulation by transcription factors is replaced by regulation by transcription-generated supercoiling (Sevier and Hormoz, 2021).

The common limit to all models described above is that these models focus on mechanistic descriptions of the supercoiling-mediated local interaction between neighboring genes, but do not try to generalize to the whole-genome scale nor to an evolutionary time frame. Exploring the role of the transcription-supercoiling coupling at the scale of complete bacterial genomes, and the reaction of supercoiling-mediated interactions to the changes in relative gene positions that can be caused by genomic rearrangement, however seems necessary in order to decipher the evolutionary role of supercoiling mutations.

2.4 An Evolutionary Systems Biology Approach

The models of the transcription-supercoiling coupling presented above demonstrate that the system that emerges from the coupled transcription of neighboring genes on a genome is complex, in the sense that it presents behaviors that cannot be explained by modeling each gene in isolation. Moreover, studying the epistatic interactions between mutations in genes that regulate supercoiling and in genes that are regulated by supercoiling requires the addition of another layer of complexity, as the effect of supercoiling mutations or genomic rearrangements on gene transcription levels cannot be directly predicted from the mutations themselves. In order to tackle this problem, I followed the methodological approach of evolutionary systems biology, which adapts the tools of systems biology to study not only complex systems themselves but also their evolution – in the Darwinian sense – over time, with the help of computer simulations (Beslon et al., 2021).

The core of this approach is represented in Figure 2.2, which describes the evolution of a population of complex systems, or "digital organisms" (Adami, 2006). Each complex system in the population is represented by means of a data structure DS_i, which represents an instantiation of underlying system with a particular set of parameter values. Using this data structure, we can evaluate the systemic properties of interest S_i of this individual, and its fitness value f_i. A new population of complex systems can then be created by selecting reproducers based on fitness, and making their underlying data structure undergo stochastic mutations that can affect both their systemic properties and their fitness. This cycle can then

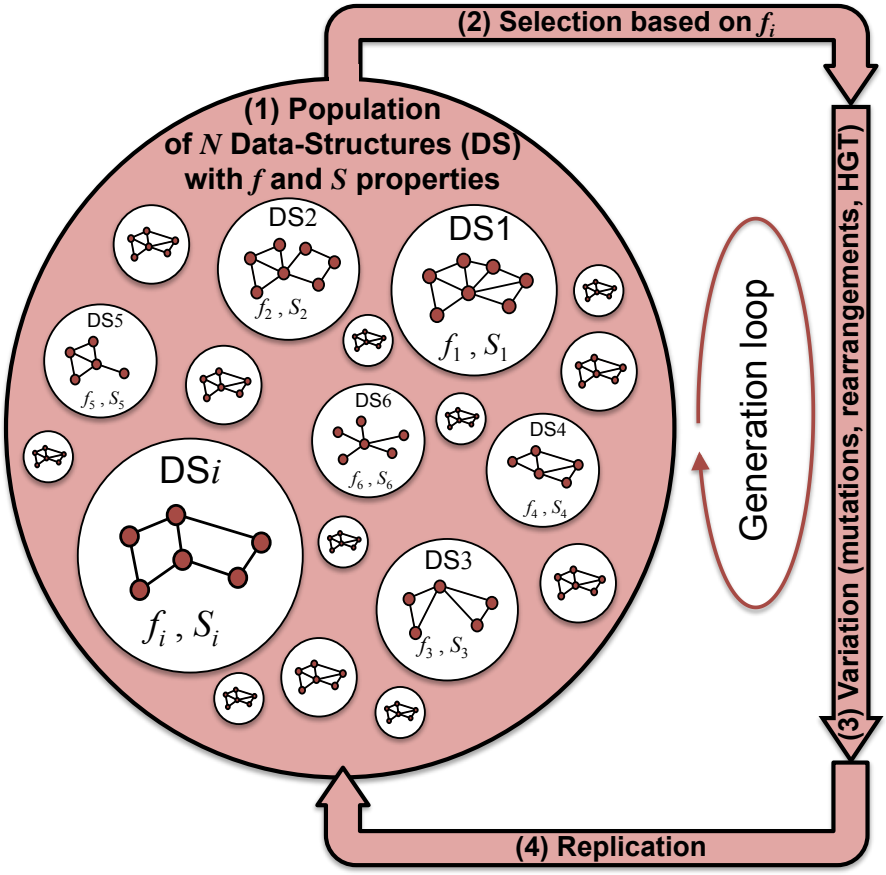


Figure 2.2: Template of an evolutionary systems biology simulation. In a population of complex systems, each individual i is represented by an inner data structure DS_i. Systemic properties of interest S_i emerge from this data structure, which is also used to compute a fitness level f_i. The population can then evolve by following an evaluation-selection-variation-replication loop. Figure reproduced with permission from (Beslon et al., 2021).

be repeated for a given number of generations, resulting in an experimental framework called "*in silico* experimental evolution", as it adapts the traditional *in vivo* experimental evolution methodology to the computational study of the evolution of arbitrary complex systems.

Both *Aevol* (introduced in Chapter 3) and *EvoTSC* (introduced in Chapter 4), the *in silico* artificial evolution platforms that I developed and used during my PhD, follow this methodology. In both platforms, each complex system represents an individual that is described by its genome; the models diverge in the data structure with which they represent genomes, as *Aevol* is a nucleotide-level model, whereas *EvoTSC* is a "string-of-pearls" model, in which genomes are represented by a series of genes separated by non-coding sections (see Hindré et al. (2012) for an overview of these formalisms). The fitness of individuals is obtained in both platforms by evaluating their gene transcription levels, and comparing these transcription levels to an implicit (in *Aevol*) or explicit (in *EvoTSC*) target. Finally, the systemic properties of individuals differ in each model, according to their choice of underlying data structure: *Aevol* can be used to study properties such as genome size or the proportion of

coding bases, whereas *EvoTSC* can be used to study the arrangement of genes on the genome.

2.5 Conclusion

The level of DNA supercoiling is an interesting property of bacterial genomes, standing at the crossroads of many processes: it is finely regulated by the joint action of topoisomerases and nucleoid-associated proteins, but remains sensitive to the external influence of environmental stress, and to the internal influence of gene transcription. The repeated mutations targeting the regulation of the supercoiling level in the *LTEE* demonstrate the role that supercoiling can play – through its central position in genome biology – in the adaptation of bacterial populations to new environments, and make it an ideal example to study the role of epistatic interactions in guiding evolutionary trajectories. Moreover, as gene transcription itself depends on DNA supercoiling, the resulting interplay between transcription and supercoiling generates a complex web of interactions between neighboring genes in the dense bacterial genomes: the transcription-supercoiling coupling. While the effect of this coupling has already well been studied at the scale of a few neighboring genes, expanding this analysis to the whole-genome scale seems necessary in order to have a qualitative understanding of the phenotypic consequences of supercoiling mutations, and hence of their possible epistatic interactions.

Finally, an *in silico* experimental evolution approach seems to be the most promising way to tackle the study of the evolutionary role of these mutations, as this methodology enables the combination of a complex model of the transcription-supercoiling coupling – an integral part of gene regulation by DNA supercoiling – with an evolutionary model that allows for the emergence of complex epistatic interactions that can influence evolutionary trajectories.

Chapter 3

Looking for Supercoiling Epistasis in *Aevol*

In this chapter, I present the first main line of work of that I undertook during my Ph.D. In order to understand the role of epistasis in the prediction of evolution, I focused on the study of the specific case of mutations affecting the DNA supercoiling level in bacteria, which were shown to be repeatable at the phenotypic level, and partially repeatable at the molecular level, in an experimental evolution setting. In order to replicate these results in the easier to study *in silico* artificial evolution setting offered by the *Aevol* software platform, I implemented a model of supercoiling in *Aevol*, and tried to detect epistatic interactions in the evolutionary trajectories of populations that evolved in the model.

3.1 Introduction

In the *Long-Term Evolution Experiment* (LTEE), started by Richard Lenski in 1988 (Lenski et al., 1991), 12 populations of *E. coli* cells, originating from the same ancestral strain, were placed to evolve in a new environment, an Erlenmeyer flask containing a glucose-limited medium. Every day since the beginning of the experiment, which has reached over 75,000 generations of bacteria and is still running, a sample from each population has been propagated into fresh medium, and samples have been cryogenically conserved every 500 generations, resulting in the longest-running evolution experiment in the lab. The LTEE demonstrated that fitness can keep on increasing for much longer than originally expected in a constant environment (Good et al., 2017). As sequencing capacity and synthetic biology subsequently developed in the late 1990s and early 2000s, identifying the precise DNA mutations underpinning these increases in fitness became possible. When sequencing the conserved *E. coli* lineages in the LTEE, beneficial mutations – mutations that confer on their bearer a higher growth rate than the ancestral strain in the conditions of the experiment – were in particular found in the *topA* gene and the *fis* gene, in one of the twelve lineages (Crozat et al., 2005).

The genes affected by these mutations are involved in the regulation of DNA supercoiling: Topoisomerase I (encoded by *topA*) directly modifies the supercoiling level by introducing

supercoils, and FIS (encoded by *fis*) is a nucleoid-associated protein which helps regulate supercoiling by binding to DNA. This makes these mutations extremely interesting in two regards. First, there is no direct phenotypic link between the supercoiling level of the chromosome and the growth rate of the bacteria, and yet these mutations, when inserted into the genetic background of the ancestral strain, still confer a fitness advantage. Second, mutations affecting supercoiling-regulating genes, especially in *gyrA* and *fis*, were subsequently found in 11 of the 12 replicates of the experiment after 20,000 generations of evolution, a rate that is much higher than for randomly chosen genes (Crozat et al., 2010). A possible interpretation of this repeated mutational targeting of supercoiling-regulating genes is that, by globally altering the transcriptional landscape of the bacteria (as the level of supercoiling directly affects gene transcription), these mutations enable the exploration of new evolutionary pathways that would have been deleterious in the ancestral strain, and enable the lineages that bear these mutations to evolve faster than the competing strains. In other words, there could be positive epistatic relationships between these mutations and the subsequent adaptive mutations that they enable through rewiring the fitness landscape of their bearer lineages.

In this chapter, I describe how I leveraged the *Aevol in silico* experimental evolution platform in order to test this evolutionary hypothesis in the simpler, more controlled setting of artificial evolution. *Aevol* is a model that is particularly well-suited to this problem, for several reasons. First, the genome biology of individuals is modeled very precisely in *Aevol*. The genome is described at the nucleotide level, and the transcription and translation stages that constitute the core of biological gene expression are accurately represented in the model. Second, *Aevol* incorporates a rich variety of mutational operators. It includes both genomic rearrangements such as inversions and translocations or duplications and deletions, and local mutations such as indels and switches. The richness of the genome-level description of *Aevol* therefore makes it an ideal tool for the study of epistatic relationships.

The chapter starts with a brief overview of the *Aevol* model; then, I present the model of supercoiling and its effect on transcription that I incorporated into the model, and describe the experiment that I performed in order to test the presence of epistasis between supercoiling mutations and other kinds of mutations.

3.2 The *Aevol* model

3.2.1 Overview

The *Aevol* platform, developed in the Inria Beagle team (Rutten et al., 2019), is a software suite designed to run artificial evolution experiments on a computer, rather than at the bench. It was originally created to investigate the influence of classical population genetics parameters such as population size, mutation rate, or selection pressure on genomes themselves, seen as an integral part of the phenotype and not only as the source of genetic information. In *Aevol*, individuals have a very abstract phenotype, in exchange for a genome that is modeled down to the nucleotide level, and follows the "central dogma of molecular biology" (Crick, 1958), with an accurate representation of RNA transcription and gene translation. This approach

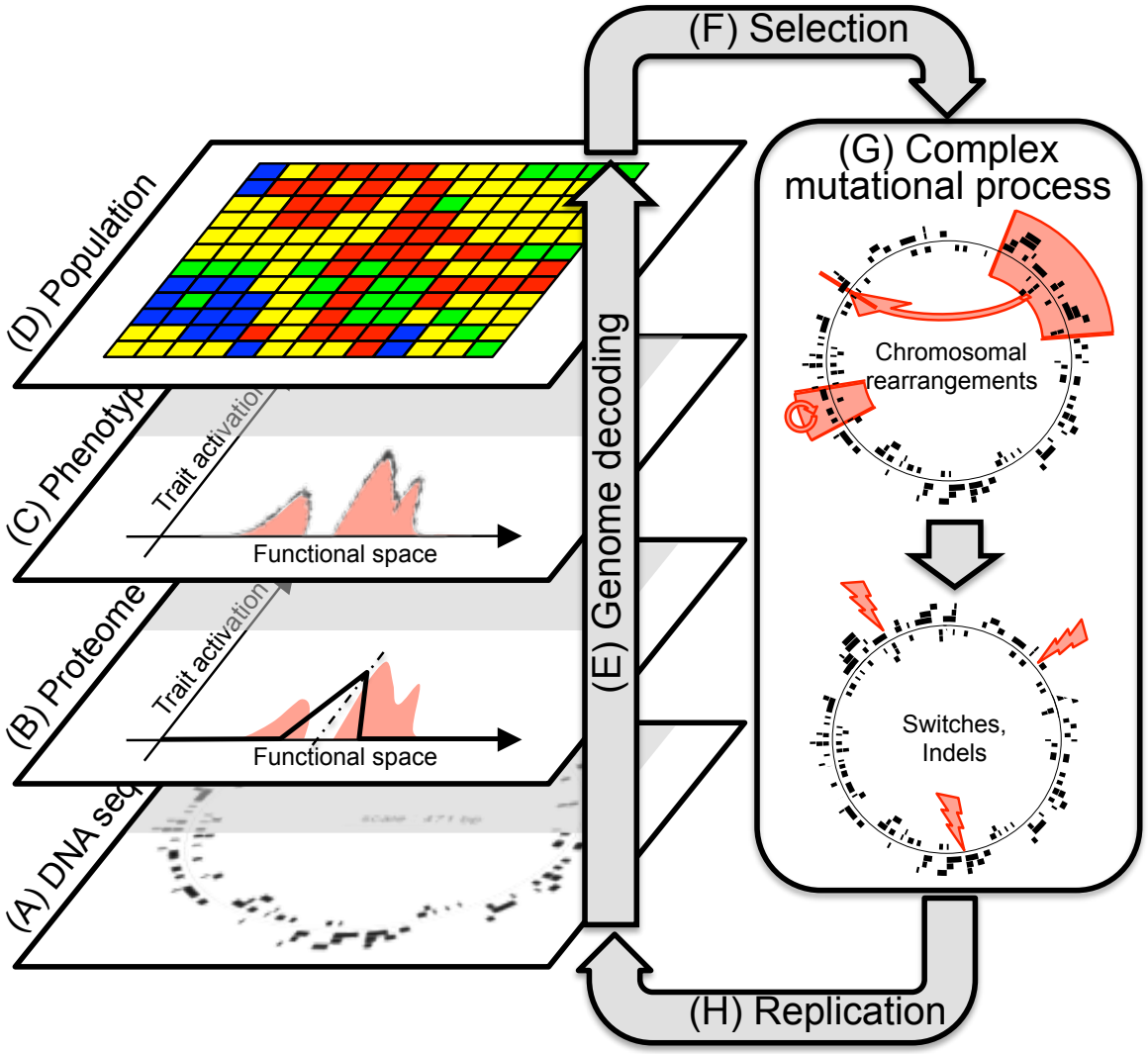


Figure 3.1: Broad overview of the *Aevol* model. In *Aevol*, an evaluation-selection-replication evolutionary loop is applied to a population of individuals defined by their genome, encoded as a circular string of nucleotides (A, central ring), on which RNA sequences and genes are decoded (A, black segments). The resulting proteins (B, in black) are mapped to an abstract phenotypic space, and summed in order to obtain the phenotype of the individual (C, in black), which is compared to an optimal phenotype (that implicitly represents the environment – B and C, in pink), in order to compute its fitness. In the model, the population is laid out on a square grid (D), with one individual per cell. In order to produce a new generation, the ancestor of the new individual in each cell is chosen at random among the neighboring individuals, proportionally to their fitness (F). Once this ancestor is chosen, its genome undergoes a series of random mutations, including rearrangements and local mutations (G), in order to obtain the genome of the new individual in the cell at the next generation (H).

contrasts with other artificial evolution platforms such as *Avida* (Adami and Brown, 1994; Ofria and Wilke, 2004), which aim at studying the evolutionary process itself, rather than its impact on biological organisms. For instance, *Aevol* has been used to study the effect of mutation rate on genome size (Knibbe et al., 2005), or of the selection pressure on the percentage of non-coding bases and number of genes on the genome (Batut et al., 2013). As an excellent and very thorough description of *Aevol* (in French) can be found in Liard (2020), the following presentation of the model will be kept short and focused on the aspects relevant to this research. Figure 3.1 provides a comprehensive overview of the evolutionary algorithm at the core of *Aevol*.

3.2.2 The Genotype-Phenotype Map in *Aevol*

A genome or genotype in *Aevol* consists in a sequence of binary characters (0 or 1), which represents a double-stranded circular sequence of DNA. The genome sequence explicitly describes the first (forward) strand of DNA, while the second (reverse) strand is obtained by complementing the sequence, replacing 0 by 1 and vice-versa. In order to turn this genotype into a phenotype (Figure 3.1 C), the decoding algorithm starts by looking for sequences that code for RNAs, reading the forward strand left-to-right and the reverse strand right-to-left. An RNA starts with a promoter sequence, which has to match a consensus sequence with up to d_{max} errors, and ends with a hairpin-like terminator. Then, each RNA is scanned for genes, which start with a ribosome binding site followed by a 3-nucleotide start codon, which defines the reading frame. Reading continues until a stop codon is found in the same frame, and the resulting string of codons is then translated into a protein, or discarded if no stop codon is found in frame before the end of the RNA sequence. An RNA can thus contain zero, one, or several protein-coding genes.

As the genetic alphabet is binary in *Aevol*, there are 8 different 3-nucleotide codons, and 6 codons can therefore be used to encode protein data, in addition to the start and stop codons. These codons are grouped into three pairs, each respectively encoding the width w, height h, and mean position m of a triangle kernel function from $[0, 1]$ to $[0, 1]$ (as represented in Figure 3.1 B). The mean m represents the main function that the protein fulfills in the abstract phenotypic space, the height h the intensity with which it does, and the width w the pleiotropic ability of the protein to fulfill neighboring phenotypic functions.

In order to obtain the final contribution of the protein to the phenotype, the constitutive height h of the gene is weighted by the expression level e of the RNA that carries the gene, which depends on the activity of the promoter of that RNA. In the model, the promoter activity decreases linearly with the difference d between its sequence and the consensus sequence, and vanishes when $d > d_{max}$. The expression level of the RNA is then given by the following equation: $e = 1 - \frac{d}{1+d_{max}}$. Finally, in order to compute the complete phenotype of the individual from the set of its proteins, the kernel functions representing each gene are summed, resulting in a piecewise-linear phenotype function. As the maximum degree to which each phenotypic function can be fulfilled is bounded by 1, the phenotype function is finally capped using Łukasiewicz operators in order to keep within this limit.

3.2.3 Fitness

Once the phenotype of an individual has been decoded from its genome, we can compute its fitness. As the environment is indirectly specified by an optimal phenotype, we first compute a phenotypic gap as the integral of the absolute value of the difference between the phenotype of the individual and the optimal phenotype, taken over the range of phenotypic values (the L^1 distance between the functions). Then, we compute the fitness as the inverse exponential of the phenotypic gap, multiplied by a selection coefficient: the higher the coefficient, the larger the difference in fitness between individuals with the same difference in gap.

3.2.4 Mutational Operators

Once the ancestor of a new individual has been chosen, a set of random mutations are applied to its genome to obtain the new genome. These mutations are split into two classes, depending on the proportion of the genome that they can affect: genomic rearrangements, and local mutations.

Genomic rearrangements can affect up to the whole genome, and comprise four kinds of structural changes: duplications, deletions, inversions, and translocations. In each of these rearrangements, the two endpoints of the affected segment are first drawn randomly on the genome. In a large duplication, an additional insertion point is randomly selected on the genome, and the genetic content located between the endpoints is copied at the insertion point. In a large deletion, the genetic content that was present between the endpoints is simply discarded. In an inversion, the segment is reinserted left-to-right between the endpoints, reversing the orientation of every gene located on the inversion. Finally, in a translocation, the genetic content is removed, turned into a circular plasmid, cut at a random point in the plasmid, and reinserted at another insertion point in the genome.

Local mutations, on the contrary, comprise small insertions, small deletions (collectively known as indels), and switches. In a small insertion or deletion, up to 6 contiguous bases are either inserted (choosing each base at random) or deleted at a random point in the genome. In a switch, the value of a random nucleotide is switched, from 1 to 0 or vice-versa.

3.3 Modeling DNA Supercoiling in *Aevol*

In order to model the effect of supercoiling on gene transcription in *Aevol*, I chose to start with very simple approximations, concerning the level of supercoiling itself and its effect on transcription, as the *Aevol* model is already quite complex on its own.

3.3.1 Level of DNA Supercoiling

First, I consider the supercoiling level as constant along the genome and over time, which can be interpreted as taking the spatial and temporal average of the (actually dynamic) supercoiling level. To implement this model inside *Aevol*, I changed the genotype of individuals by adding, alongside the string-of-nucleotides genome, a single parameter γ which represents

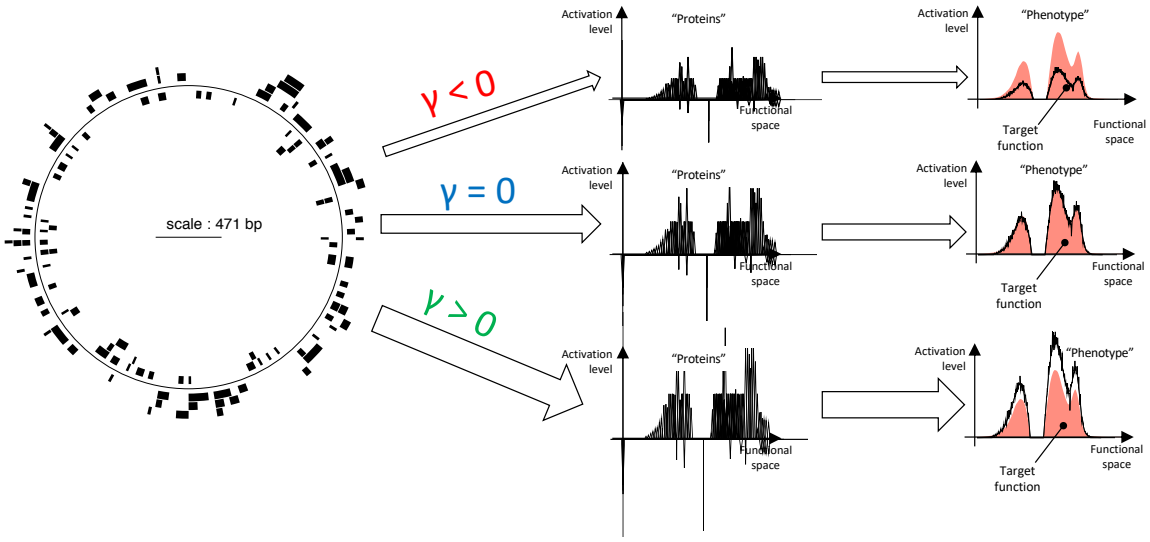


Figure 3.2: Effect of supercoiling on the phenotype of an *Aevol* individual. Left: genome (central ring) and genes (black rectangles) of the individual. Middle: kernel function encoded by every gene in the phenotypic space, affected by an excess of positive supercoiling (top), no extra supercoiling (middle), or an excess of negative supercoiling (bottom). Right: phenotype of the individual in each situation, compared to the optimal phenotype (in pink).

the relative variation in the supercoiling level σ of this individual compared to a reference supercoiling level σ_0: $\gamma = \frac{\sigma - \sigma_0}{\sigma_0}$.

3.3.2 Gene Expression

To keep the model as simple as possible, I also chose to model the effect of supercoiling on transcription as having the same linear effect on the transcription rate of every RNA on the genome. I therefore updated the computation of the gene expression e to take supercoiling into account, in addition to promoter activity:

$$e = (1 - \frac{d}{1 + d_{max}}) \cdot (1 + \gamma) \tag{3.1}$$

The effect of supercoiling on the phenotype of an example (pre-evolved) individual in the model is presented in Figure 3.2. When $\gamma < 0$ (top row) – when there is an excess of positive supercoiling compared to the baseline – the expression of every gene is decreased. When γ is equal to 0 (middle row) – when the supercoiling level is equal to the baseline – there is no change to gene expression levels, which replicates the behavior of the original *Aevol* model. Finally, when $\gamma > 0$ (bottom row) – when there is more negative supercoiling than in the baseline – the expression of every gene is increased.

3.3.3 Mutational Operator

In biological organisms, the supercoiling level is not only a direct property of the DNA molecule, but is also controlled by topoisomerases and nucleoid-associated proteins that are not modeled in *Aevol* (as the phenotypic space is completely abstract), and changes in the supercoiling level come from mutations affecting the genes that encode these proteins, such as *gyrA* or *fis* (Crozat et al., 2005). In order to model mutations in the supercoiling level in *Aevol*, I chose a continuous model, in which a small variation in γ indirectly reflects the effect of a mutation in one of the supercoiling-controlling genes.

When an individual reproduces, we first use a Bernoulli trial, with a probability p that represents the probability that a supercoiling-protein gene undergoes a non-synonymous mutation, to decide whether to change the supercoiling level. Then, if the supercoiling level should change, we draw a variation in relative supercoiling $\Delta\gamma$ according to a normal distribution $\mathcal{N}(0, s^2)$, and finally set the relative supercoiling level γ' of the offspring to $\gamma' = \gamma + \Delta\gamma$. The parameters of these laws are parameters of the simulation, and their values are given in Table 3.1. Throughout this chapter, I will for the sake of clarity refer to the usual DNA-affecting mutations presented in 3.2 as *genomic mutations*, and to the mutations in the supercoiling level presented here as *supercoiling mutations*.

3.4 Results

As presented in the introduction of this chapter, the goal of implementing a supercoiling model in *Aevol* was twofold. The first aim was to see to which extent adding a new dimension to the phenotypic space, and a new mutational operator to explore this new dimension, would allow populations to evolve faster than allowed by the original model, thanks to the wide jumps in the phenotypic landscape that are made possible by the supercoiling mutations. The second aim was to disentangle the possible epistatic effects between supercoiling mutations and genomic mutations in *Aevol*. In this section, I first present the experimental setup that I used in order to answer these questions. Then, I show that adding regulation by supercoiling did not measurably increase the rate of adaptation of populations compared to the control, and that supercoiling indeed follows a very constrained evolutionary trajectory in these experiments. Finally, I conclude that I could not find any observable epistasis between supercoiling and other mutations, when using the supercoiling model presented in Section 3.2.

3.4.1 Experimental Setup

In order to tackle these questions, I ran two sets of simulations: the experimental runs using the supercoiling model, and the control runs using the vanilla version of *Aevol*. Each set of runs comprises 5 replicate populations, which were evolved for 1,000,000 generations, each starting from a clonal population. Each of the initial individuals was obtained by randomly drawing 5,000 bp-long genomes, until a genome with a non-zero fitness (i.e., at least one protein-coding gene partially matching the phenotypic target) was found. The simulations were run on a 24-core Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz server with 128 GB of

Parameter	Symbol	Value
Population size	N	1,024 (32x32 grid)
Initial genome size	g_0	5,000 bp
Local mutation rate	μ_{loc}	10^{-7} bp^{-1}.gen^{-1}
Rearrangement rate	μ_{rear}	10^{-6} bp^{-1}.gen^{-1}
Initial supercoiling level	γ_0	0
Supercoiling mutation probability	p	10^{-1}
Supercoiling mutation variance	s^2	10^{-2}
Generations	T	1,000,000
Number of replicates	n	5

Table 3.1: Table of parameter values used in the *Aevol* evolutionary runs. The top part describes parameters common to the experimental and control set of rules, the middle part the supercoiling-related parameters introduced in the supercoiling model, and the bottom part simulation-specific parameters.

RAM, and lasted approximately a week for each set of replicates. The limited number of replicates for each set of simulations was chosen to balance their energy expenditure with the preliminary character of the work, which alleviates the need for statistical strength in the resulting data. All the data from this experiment is available online on the Zenodo platform.

3.4.2 Studying Lineages

The data that is presented in the rest of this section was obtained by reconstructing the lineage, starting from the initial generation, of a random individual at the last generation of each replicate. Studying a given lineage, rather than the best individual at every generation (which need not sire one another), allows us to reconstruct the precise set of mutations that happened throughout the evolutionary history of this lineage, and therefore gives us information about the possible causal link between these mutations, and hence about their possible epistatic relationships.

As a theoretical haploid Wright-Fisher population with N individuals coalesces on average in $2N$ generations without mutation or selection (Felsenstein, 2019), we chose to analyze the data from generation 0 to 990,000 of every replicate (excluding the last 10,000 generations), ensuring that the last individual in each lineage is indeed ancestral to the whole population of the last generation of that replicate.

3.4.3 Evolution of the Fitness Level

Figure 3.3 presents, on the left-hand side, the fitness of the individual at every generation of the lineage of the final population, or lineage individual, in each replicate. In each case, fitness follows a broadly sigmoid shape (noting that both axes are logarithmic): the fitness of

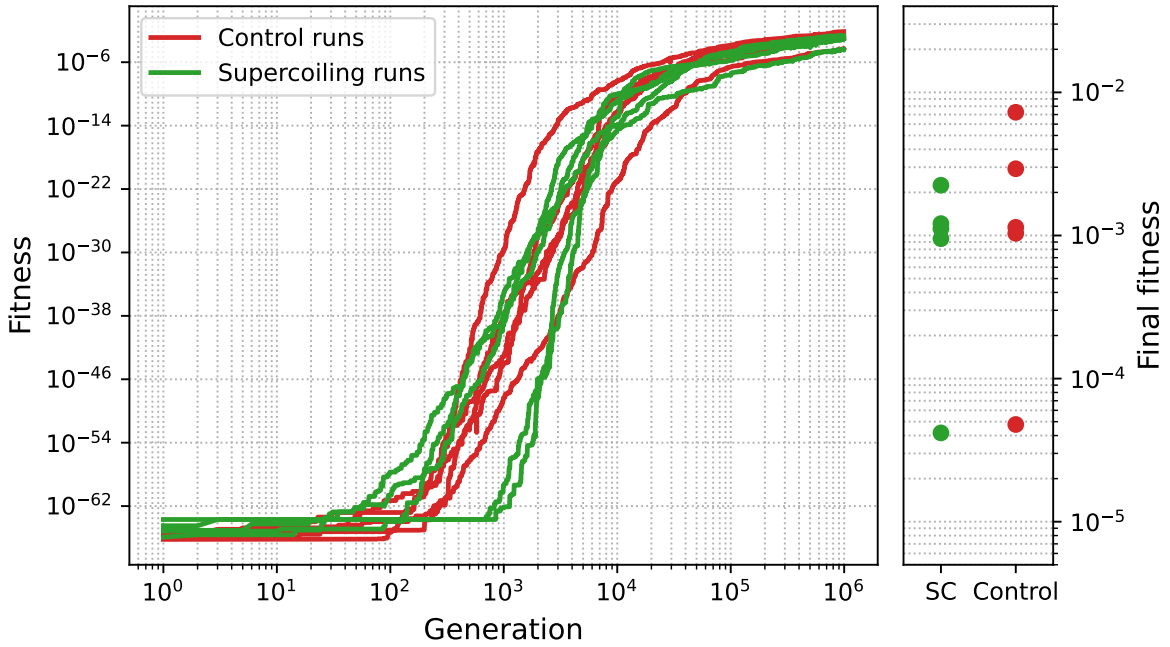


Figure 3.3: Left: Evolution of the fitness at every generation throughout the lineage of the final population of each replicate of the experimental (green) and control (red) runs. Both axes follow logarithmic scales. Right: Fitness of the lineage individual at the 990,000th generation of each run, separated between supercoiling (green) and control (red) runs.

each run quickly increases from generation 100 up to generation 100,000, then slows down for the remaining 900,000 generations, but never completely ceases to progress, mirroring in *Aevol* the open-ended evolution observed in the LTEE. The right-hand side of Figure 3.3 shows the fitness of the lineage individual at the 990,000th generation of each run.

With the limited number of replicates of each run, there is no discernible difference in fitness between the two experimental conditions, with and without mutations in the supercoiling level. Adding the new phenotypic dimension of the supercoiling level, and the associated supercoiling mutational operator, therefore does not seem to play an important role in the rate of evolution of the populations modeled in *Aevol*.

3.4.4 Evolution of the Supercoiling Level

Figure 3.4 shows the evolution of the supercoiling level throughout the lineage in each of the 5 replicates. In every run, the supercoiling level evolves only at the very beginning of the run, stabilizing in a few tens of thousands of generations, and remains essentially constant afterwards. This is in strong contrast to the fitness of the runs (presented in Figure 3.4), which keeps increasing until the end of the runs.

It therefore seems that the supercoiling level might play a role in the early evolution of the runs, but not in their long-term fitness improvement. This result is in a sense slightly disappointing but was not entirely unexpected. Indeed, in the early evolution of individ-

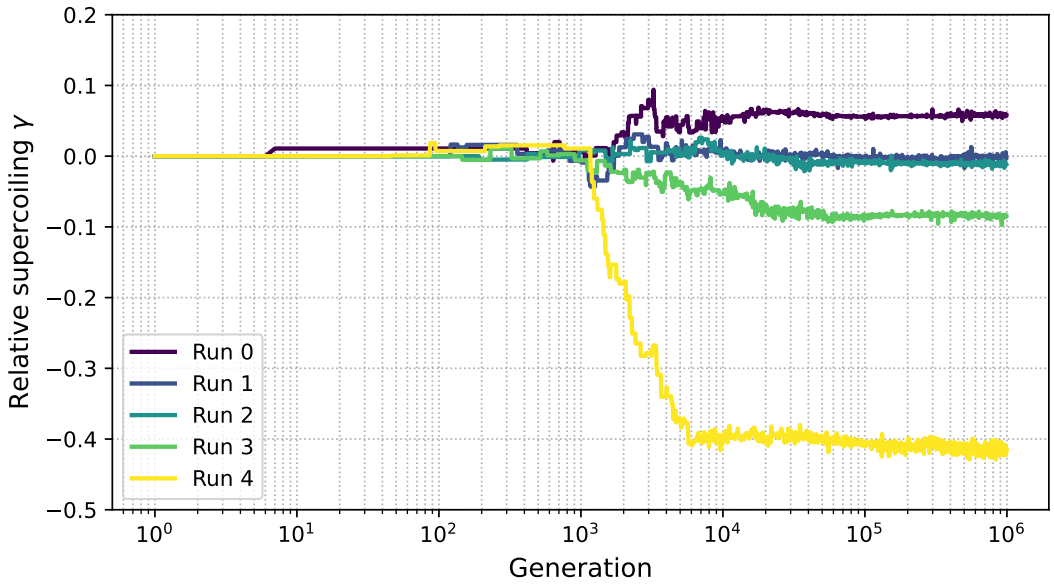


Figure 3.4: Evolution of the relative level of supercoiling at every generation of the lineage of each of the experimental runs.

uals in *Aevol*, the phenotypic target is only very imperfectly approached by the proteins expressed by the individual. At that stage, mutations in the supercoiling level, which affect the expression level of every protein equally, could indeed have a positive effect by bringing the whole phenotype closer to the optimum, in a very broad stroke. Indeed, as individuals in *Aevol* evolve from an ancestor with a single good gene, phenotypic functions are often under-performed by individuals at the early stages of evolution in the model, as the width and height of the kernel functions of their genes can be small, and as the expression level of their RNAs can be quite low if their promoters contain too many errors. The different supercoiling values towards which each of the replicates tends to converge in Figure 3.4 could therefore be interpreted as a founding effect coming from the genome of the original individual in that run, which could be confirmed by a more detailed analysis of the series of mutations that happened in each lineage.

However, as evolution progresses, and as the optimal phenotype is more and more closely matched by the individual, changing the whole expression profile at once becomes less and less susceptible to be favorable. This case is represented in Figure 3.2: any change in supercoiling, be it positive or negative, will decrease the fitness of the individual, and supercoiling mutations are therefore less and less susceptible to be picked up in the lineage.

These results tend to show that the model in which supercoiling has a global, linear effect on gene expression levels is too simplistic in order to produce phenotypic effects that are variable enough to have a chance to be picked up by selection; and therefore that this model is insufficient to study the interplay between supercoiling mutations and genomic mutations in *Aevol*.

3.4.5 Looking for Epistasis

Waiting Intervals Before and After Mutations In order to detect signs of positive or negative epistasis between the different kinds of mutations, I used the following approach, which considers the waiting intervals before and after mutations happen: if, for a given mutation type, the average interval until a new favorable mutation fixes in the lineage after a mutation of that type is smaller than the average interval since the last favorable mutation before that mutation, this could be interpreted as a sign that the mutation has increased the probability of a favorable mutation happening; in other terms, as a broadening of the evolutionary paths available to the genome, or a sign of positive epistasis between that kind of mutation and other kinds of mutations. On the contrary, if it takes longer for a new favorable mutation to fix in the lineage after that mutation, it would be a sign that the evolutionary paths have been constrained by the inversion: a sign of negative epistasis.

The data obtained following this approach is presented in Figure 3.5. For each mutation type, it shows the average number of mutations of that type that fixed in the lineage of each replicate, as well as the average time after which a mutation of that types fixes after a non-neutral mutation (left), and before a non-neutral mutation fixes after a mutation of that type (right), in the control runs (top) and in the experimental runs (bottom).

Epistasis of Duplications and Deletions In the control, a faint pattern seems to be discernible for large-scale inversions and deletions: the average time to a new mutation after a deletion is slightly higher than the time before a deletion, hinting that deletions could present a negative epistasis with other mutations. Conversely, the time to a new mutation after a duplication is slightly lower than the time before the mutation, hinting that duplications could on the contrary present positive epistasis with other mutations. Local mutations, as well as rearrangements and inversions, do however not seem to swing one way or the other.

In the experimental runs, no such pattern is visible at first sight, including for the supercoiling mutations, and the global average waiting intervals are smaller than in the control, which is consistent with the introduction of a new mutation type. There therefore seems to be no sign of epistasis between supercoiling mutations and genomic mutations, when following the approach explained above.

Role of the Genome Size A hypothesis that could explain the pattern visible in the control runs for large deletions and duplications is that the difference in waiting intervals – positive or negative epistasis – is simply due to the change in the genome size caused by these mutations. All mutation rates in the model are indeed proportional to the genome size, and the expected number of mutations at each generation therefore increases and decreases with the genome size (assuming that there is no fitness effect or selection). However, in the experimental runs, the probability of a supercoiling mutation does importantly not depend on the genome size. A hypothesis to explain the disappearance of the signal (possibly) there in the control for duplications and deletions could therefore be that supercoiling mutations tick according to their own clock, which depends on their parameters p and s^2, but not on the state of the genome itself. For example, if a certain beneficial indel is allowed to happen because of a duplication and indeed happens sometime after that duplication in the lineage,

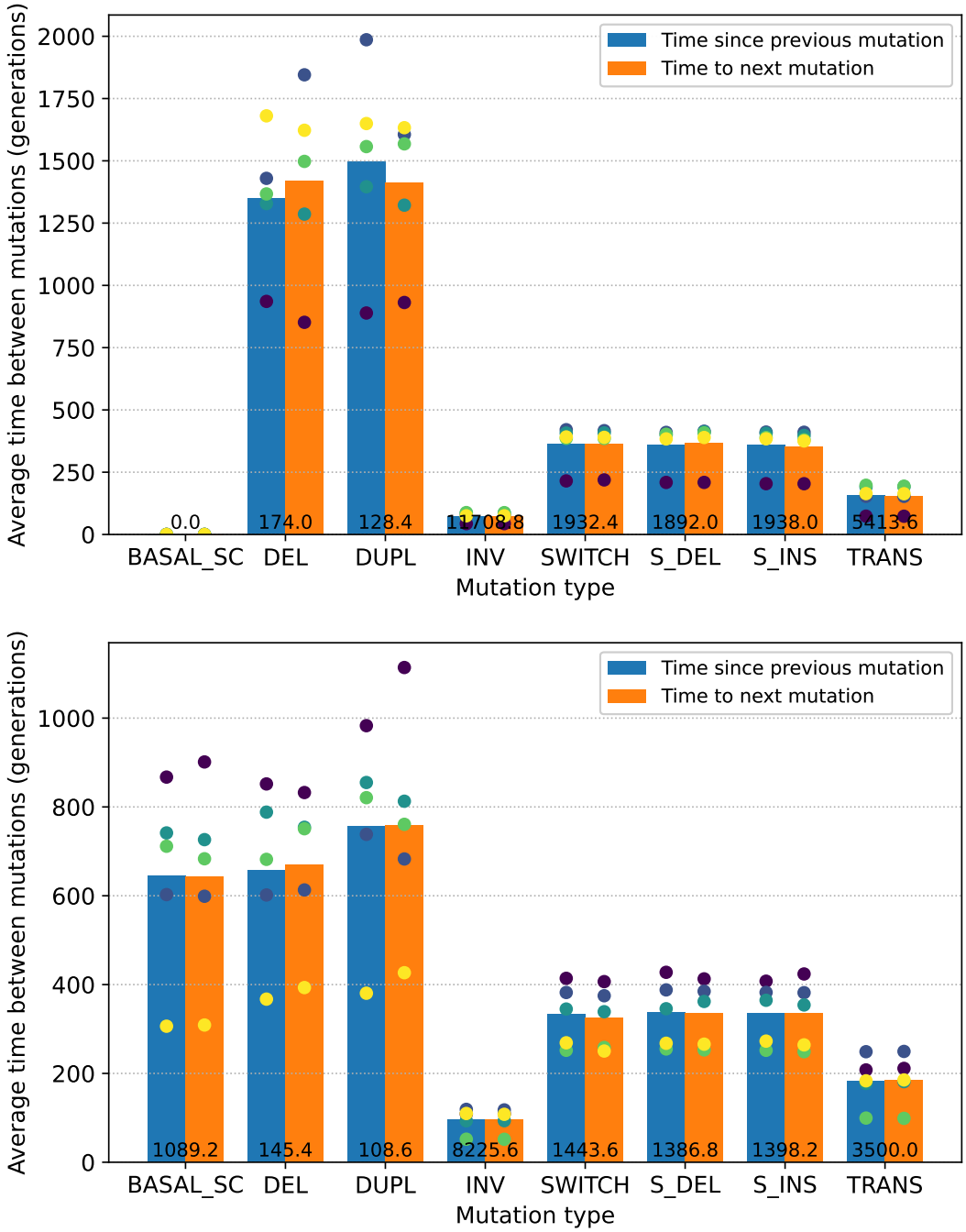


Figure 3.5: Average time before and after a mutation of each kind, in the control runs (top) and the experimental runs (bottom). For each kind of mutation, the times presented show the wait time until a neutral mutation of that kind after a non-neutral mutation of any kind, and the time until the next non-neutral mutation of any kind after a mutation of that kind. The bars show the average over the five replicates, and the colored dots show the value for every replicate. The average number of neutral mutations of each type is displayed at the bottom of the corresponding bar.

but a supercoiling mutation has happened between the duplication and the indel, then the signal from this particular epistatic relationship will have been hidden by that supercoiling mutation.

3.5 Conclusion

The goal of this initial work was to study how supercoiling mutations affect the fitness landscape of individuals in *Aevol*, that is the possible epistatic interactions between supercoiling mutations and other kinds of mutations. In order to tackle this question, I implemented a model of the effect of the supercoiling level on gene expression, as well as a model of mutations in the supercoiling level, in *Aevol*. Using this version of the model, I ran evolutionary experiments, in which I compared the evolution of populations with supercoiling with control populations by analyzing the fitness, supercoiling level, and mutations fixed in the lineage of individuals that leads to the final population of every replicate.

With the limited data that was available, I could not find a difference in the evolution rates of each set of experiments, and deduced that supercoiling does not seem to play an important evolutionary role in this model; this result was substantiated by the fact that the supercoiling level converges very quickly to a fixed level in the evolutionary history of each population. I then tried to detect signals of positive or negative epistasis between the different kinds of mutations, by looking at the waiting intervals between each kind of mutation. While this approach did not lead to meaningful results in the experimental runs, it did hint at a possible epistatic link between duplications or deletions and the other mutation kinds, due to their effect on genome size, in the control runs, which seems promising for further investigation.

The verdict of these preliminary experiments was that the model of supercoiling that I implemented in *Aevol*, in which supercoiling is kept constant along the genome and affects the expression level of all genes equally, was probably too simplistic to obtain meaningful results. Rather than pursuing this avenue of research further by implementing a more precise model in *Aevol*, I chose instead to go in a different direction. In order to decouple the complexity of the *Aevol* model from the study of the evolutionary role of supercoiling, I decided to simplify the individual model, genotype-phenotype map, and mutational operators as much as possible, in order to model the effect of supercoiling on gene expression more precisely while keeping the overall complexity of the model in check. The results of this renewed approach are presented in the following chapters.

Chapter 4

Evolution of Environmental Sensing through DNA Supercoiling

This chapter presents the proof-of-concept version of the *EvoTSC* model, and the first results that I obtained with that version of the model: I show that the evolution of differentiated expression levels in different environments is possible when gene expression is only regulated by the transcription-supercoiling coupling. The text of the chapter is an edited version of an article published in the *Artificial Life* journal (Grohens et al., 2022b).

Both the importance of gene regulation via supercoiling and the detailed mechanisms of the transcription-supercoiling coupling, at the local scale, have already been studied extensively in the literature (see Section 2.1). However, a thorough analysis of the effect of the transcription-supercoiling coupling on gene expression at the whole-genome scale – and of its possible evolutionary use by natural selection – remains lacking, in particular in the dense prokaryotic genomes, in which large groups of genes are likely to interact through this coupling. In this work, we describe a new model which incorporates a high-level model of global supercoiling regulation and of the transcription-supercoiling coupling 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 in the global supercoiling level. Then, we study the evolutionary trajectory of gene activation patterns in individuals subjected to different environments.

We show that in our model, a genome-scale gene interaction network emerges from local supercoiling-mediated interactions, and creates a reaction norm in response to the change of a single parameter, the global supercoiling level, caused by different environments. Moreover, we demonstrate that, using genomic inversions as the only mutation operator, and therefore only changing the relative positions and orientations of genes on the genome, evolution can select genomes displaying qualitatively different phenotypes in different environments characterized by different global supercoiling conditions.

4.1 A Genome-Wide Model of the Transcription-Supercoiling Coupling

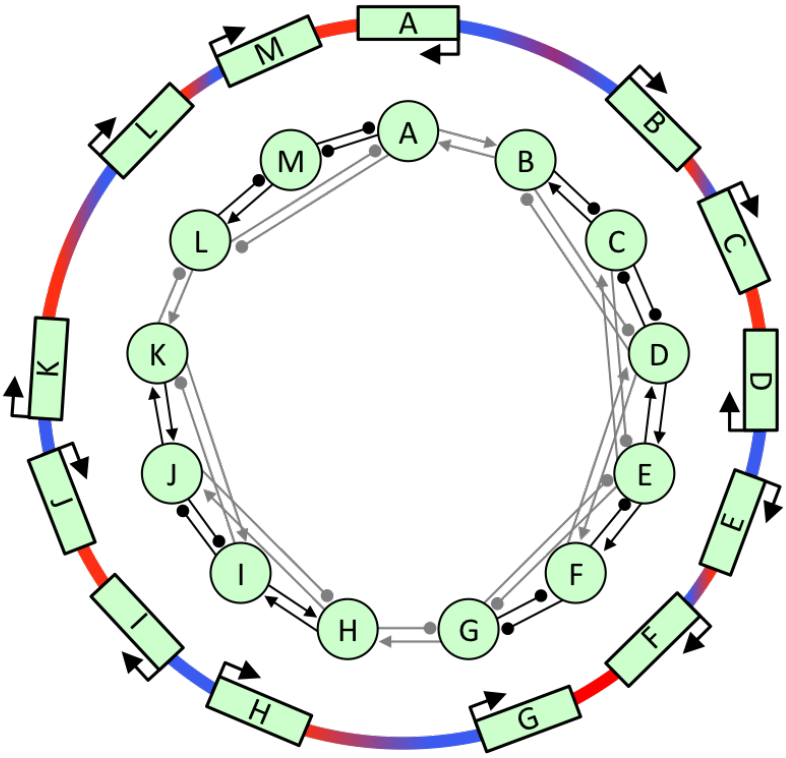


Figure 4.1: Genes along an example genome and local variations in supercoiling (outer ring), and the associated gene interaction network (inner ring). The outer ring color shows locally high ($\sigma > \sigma_0$, red) or low ($\sigma < \sigma_0$, blue) supercoiling levels due to gene transcription. In the inner ring, closer genes interact more strongly (black arrows) than genes that are farther apart (gray arrows), either positively (pointed arrows) or negatively (rounded arrows) depending on their relative orientations.

Our model consists in an individual-based simulation, written in Python. Its source code is available at `https://gitlab.inria.fr/tgrohens/evotsc/-/tree/alife-journal`. It is also preserved for long-term archival using the Software Heritage online archive (Di Cosmo, 2020). An individual in the model is represented by a circular genome (representative of most bacterial genomes), comprising a fixed number of genes, separated by non-coding intergenic regions. Each gene is described by the following characteristics: its locus on the genome, its orientation, and its basal transcription (or expression) level. As we are mainly interested in the interplay between supercoiling and transcription, we voluntarily do not make the difference between gene expression levels, understood as mRNA or protein concentrations, and transcription levels, the immediate rate of mRNA production. Indeed, assuming a separation of timescales between the fast equilibrium of the transcription-supercoiling coupling, and

the slow degradation of mRNAs, the concentration of a given mRNA is directly proportional to the transcription rate of its source gene.

Figure 4.1 illustrates the role played by the transcription-supercoiling coupling in an example genome. It includes the local supercoiling variations due to gene transcription, and the resulting gene interaction network, with each gene possibly activating or inhibiting its neighbors, depending on their relative orientations. Importantly to our approach, here genes do not interact only with their closest neighbors, but also with more distant genes, as is likely to be the case in the gene-rich bacterial genomes: *E. coli* gene promoters are around one thousand base pairs apart (Peter et al., 2004), and the transcription-generated supercoiling propagates around a few thousand base pairs on each side of the transcription site (Postow et al., 2004).

4.1.1 Mathematical Description of the Model

We model the transcription-supercoiling coupling between an individual's genes as a system of equations, which relate the supercoiling level at the locus of each gene σ_i (for i ranging from 1 to n, the number of genes of the individual), and the expression level of every gene e_i. The parameters of the system are described by the genome of the individual, as will be detailed below.

In our model, the supercoiling at a given locus on the genome depends on three factors: the individual's basal supercoiling level σ_{basal}, the variation in supercoiling due to environmental conditions σ_{env}, and the variation in supercoiling due to the transcription of the neighboring genes. We compute this local variation in supercoiling at the locus of each gene with the help of a gene interaction matrix, whose coefficient at position (i, j) describes the influence of gene j on gene i. The coefficients are given by the following equation:

$$\frac{\partial \sigma_i}{\partial e_j} = \eta \cdot c \cdot \max(1 - \frac{d(i,j)}{d_{max}}, 0) \tag{4.1}$$

More precisely, the interaction level between two genes depends on the relative orientation of the genes, as the transcription of a gene increases supercoiling at the locus of downstream genes and decreases supercoiling at the locus of upstream genes (remember that an increase in supercoiling means a decrease in transcription). Therefore, we choose $\eta = 1$ if gene i is downstream of gene j and $\eta = -1$ otherwise (if $i = j$, $\eta = 0$ as a gene does not interact with itself). The interaction level also depends on gene distance, as genes that are further apart on the genome interact less strongly, so the strength of the interaction linearly decreases with the intergenic distance $d(i, j)$, and reaches 0 when $d(i, j) = d_{max}$, the maximum distance above which the interaction vanishes. Finally, an interaction coefficient c is applied to adjust the strength of the coupling.

Using this interaction matrix, we compute the level of supercoiling σ_i at the locus of every gene, which depends on the transcription level of all the other genes, on the basal supercoiling level, and on the environmental supercoiling level:

$$\sigma_i = \sigma_{basal} + \sigma_{env} + \sum_{j=1}^{n} \frac{\partial \sigma_i}{\partial e_j} e_j \tag{4.2}$$

The transcription level e_i of every gene as a function of total supercoiling is then modeled with a sigmoidal activation curve, following El Houdaigui et al. (2019). The equation is given below:

$$e_i = \frac{1}{1 + e^{(\sigma_i - \sigma_0)/\varepsilon}} \tag{4.3}$$

In this equation, σ_0 is a parameter that represents the inflexion point of the sigmoid, that is the supercoiling level at which the gene is at half its maximum transcription rate, and ε a scaling factor that represents the strength of the dependence of the transcription level on the supercoiling level.

Finally, in order to obtain the phenotype of an individual, we numerically compute a solution to the system of equations 4.2 and 4.3, using a fixed point algorithm. This solution represents the state (of gene expression and supercoiling at every locus) towards which the individual would converge over time. Let $f(e_i)$ be the function that computes new supercoiling levels σ'_i from e_i using equation 4.2, and then computes new expression levels e'_i from the new σ'_i using equation 4.3, and finally returns e'_i. In order to compute a fixed point of f, that is a set of transcription levels e^*_i such that $f(e^*_i) = e^*_i$, we start with basal transcription levels e^0_i (that are a property of each gene), and iterate the sequence $e^{t+1}_i = \frac{1}{2}(e^t_i + f(e^t_i))$, until the difference between two successive iterations is below a given threshold. In our setting, this algorithm has empirically always converged to a solution that is a stable fixed point of the function, and that is therefore interpretable from a biological perspective.

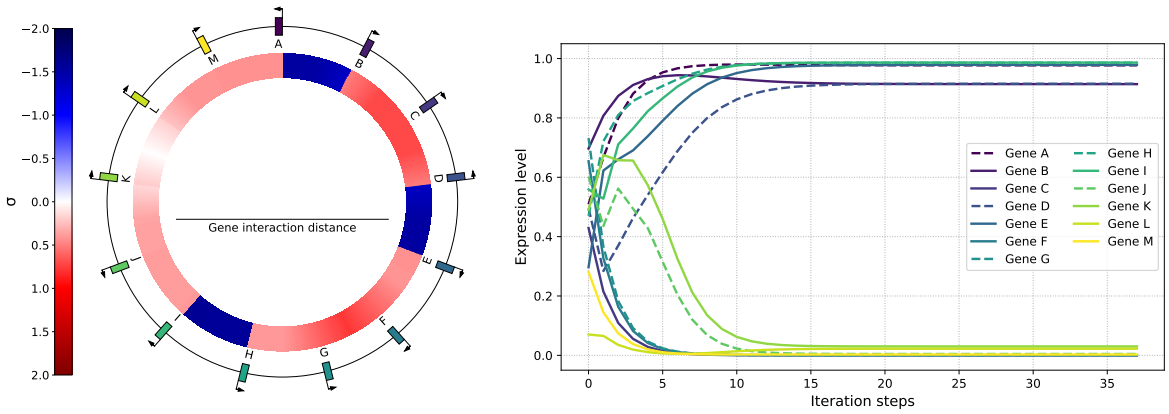


Figure 4.2: Left: genome (outer ring) and stable state level of supercoiling σ (inner ring) of an example individual with 13 genes in the model. Right: transcription levels of the individual's genes during the iterations of the fixed point computation, in an environment given by $\sigma_{env} = 0.05$. Solid lines represent genes in forward orientation, and dashed lines represent genes in reverse orientation.

Figure 4.2 shows the genome (left, outer ring) of an example individual with a genome of 13,000 bp and $n = 13$ genes evenly spaced along the genome, and with a basal supercoiling of $\sigma_{basal} = -0.06$. The basal transcription level of each gene is randomly chosen between 0 and

1, and all the iterations of the fixed point algorithm that result in the final gene transcription levels are shown on the right. In this individual, the non-linear effect of the interaction between neighboring genes is clearly visible. Indeed, six genes (A, B, D, E, H, and I) end up at a high transcription rate at the fixed point (or solution) of the system, while the others end up at low transcription rates. 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 D has a low (around 0.3) basal transcription rate, it eventually reaches a high transcription state because of its positive interaction with gene E. Conversely, genes F and G start with a high transcription rate, but are repressed by their neighbors H and E, and are therefore silenced as the system converges. We can also observe complex behaviors in the model, as the gene expression levels pass through very different states during convergence to the solution. Indeed, the transcription level of gene K initially increases due to its interaction with gene J, but both genes end up in a low transcription state, as they are inhibited by the very active gene I. The final supercoiling level along the genome (left, inner ring) moreover demonstrates the effect of the transcription-supercoiling coupling on local supercoiling. Highly transcribed genes, such as A and B, generate a large variation in the supercoiling level on their upstream and downstream sides, and the positive feedback loop between genes in divergent pairs is made clear by the very high negative value of the supercoiling level between each of the genes in these two pairs.

4.1.2 Effect of the Environmental Supercoiling on Gene Activation Levels

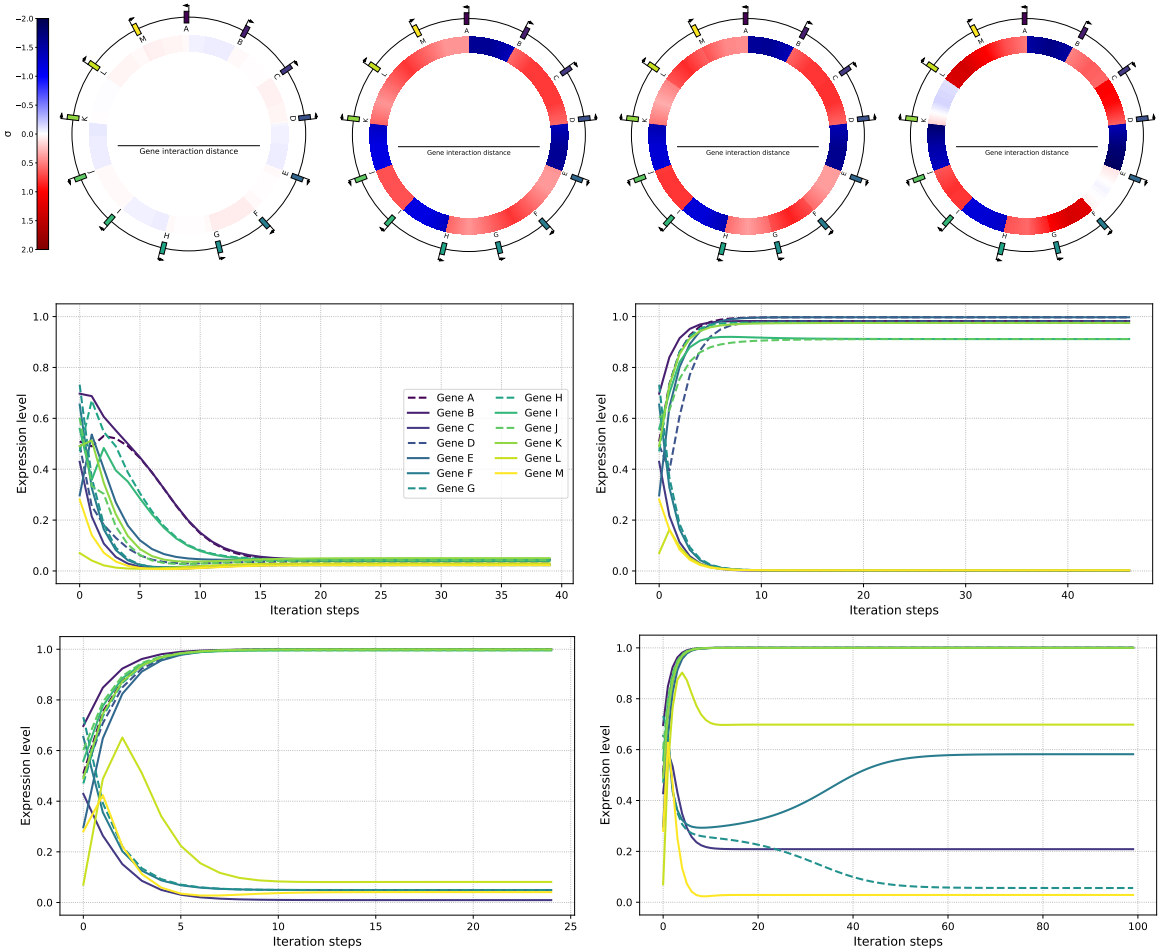


Figure 4.3: Influence of the environment supercoiling σ_{env} on the stable state local supercoiling level (top row) and gene transcription levels (bottom rows) of the example individual. From left to right and top to bottom: at $\sigma_{env} = 0.1$, no genes are activated ($e > 0.5$); at $\sigma_{env} = 0.0$ and at $\sigma_{env} = -0.1$, 8 genes are activated; at $\sigma_{env} = -0.2$, 10 genes are activated. Lower values of σ_{env} result in the activation of more genes, reflecting the *in vivo* effect of higher negative supercoiling.

Figure 4.3 captures the influence of the environmental change in supercoiling σ_{env} on the local supercoiling level due to the transcription-supercoiling coupling (top row) and on the repartition of genes between the activated and inhibited states (bottom rows), again using the example individual already shown in Figure 4.2. From left to right and top to bottom: at a high value of $\sigma_{env} = 0.1$, meaning that DNA is severely overwound compared to normal, no gene is activated (with an expression level $e > 0.5$) at all. As the external influence of the environment on supercoiling decreases to $\sigma_{env} = 0$, corresponding to normal relaxation

of DNA, and then to $\sigma_{env} = -0.1$, 8 out of the 13 genes of the individual reach an activated state. Finally, for $\sigma_{env} = -0.2$, there is a strong environmental pressure towards high gene transcription levels, and most genes are indeed activated; however, even at this level of σ_{env}, some genes remain shut down, because of the high amount of positive supercoiling (in red) generated by the transcription of their neighbors.

4.1.3 Influence of Relative Gene Positions on Gene Activation Levels

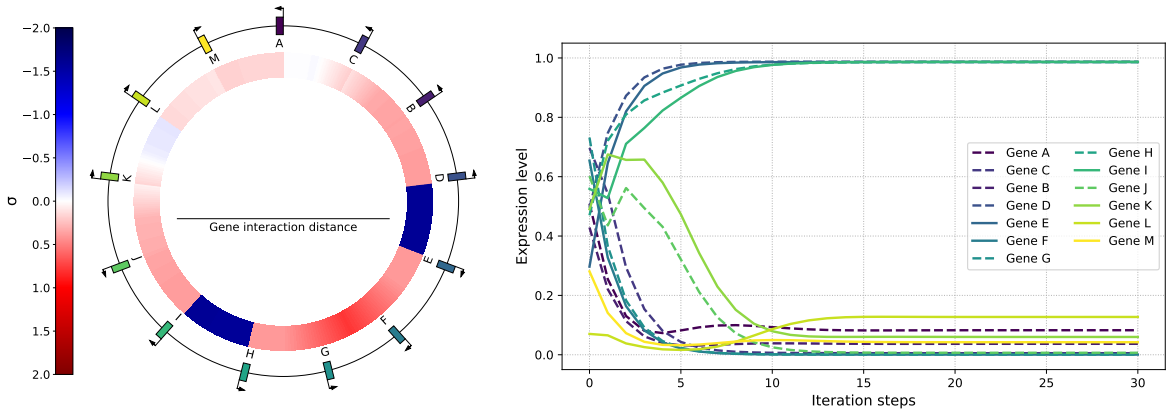


Figure 4.4: Genome, local supercoiling and gene expression levels of a new individual obtained from the individual in Figure 4.2 by switching the positions and orientations of genes B and C.

Figure 4.4 again shows the local supercoiling and gene expression levels of the individual in Figure 4.2, after reversing the positions and orientations of genes B and C. This is an example of a genomic inversion, which will be presented in further detail in section 4.2.3. The starting 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 was present 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, as gene B is now more strongly inhibited by gene D due to its closer genomic location, and as genes A and B are not in a positive feedback loop – due to diverging orientations – any longer; 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 directly depend on the architecture of the genome.

4.2 An Evolutionary Genome-Wide Model of the Transcription-Supercoiling Coupling

After evidencing that transcriptional activity depends on the organization of the genome, we now question to which extent evolution can simultaneously leverage the organization of the genome and the transcription-supercoiling coupling in order to adapt gene regulatory activity to different environments. Indeed, as has been observed in *Dickeya dadantii* (Muskhelishvili et al., 2019), different phenotypes can evolve as a response to different supercoiling levels induced by the environment, and the transcription-supercoiling coupling could play a role in enabling the existence of this reaction norm.

In this section, we expand our model into an evolutionary simulation. At each generation of the simulation, all individuals are evaluated and their fitness values are computed, based on their gene transcription levels. 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, during replication, the genome of each new individual stochastically undergoes a number of mutations, before the new individual is evaluated again; importantly, these mutations do not impact genes themselves, but only the spatial organization of the genome: gene orientations, syntenies, and intergenic distances.

4.2.1 Evolutionary Model: Evolution in Two Separate Environments

We model the evolution of populations of individuals that experience two different environments, named A and B. Each environment is defined by its value of σ_{env}, respectively σ_A and σ_B, which represent the change in the supercoiling level due to the environment (Dorman and Dorman, 2016). In order to have environments with distinct effects, we choose a value of $\sigma_A = 0.1$, for which isolated genes are effectively inhibited (as in the top-left panel of Figure 4.3), and a value of $\sigma_B = -0.1$, for which some but not all genes are activated (bottom-left panel).

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

4.2.2 Fitness

In order to compute the fitness of an individual, we define an optimal phenotype $\tilde{e}^A$ (resp. $\tilde{e}^B$), corresponding to the vector of the expected expression level $\tilde{e}_i^A$ for each gene i in environment A (resp. environment B). We choose an expected expression level of $\tilde{e} = 1$ for genes that should be activated, which corresponds to the maximum possible expression level

of a gene in our model. Similarly, we choose $\tilde{e} = 0$ for genes that should be inhibited, which is the minimum expression level that is attainable. Then, in each environment, we compute the gap g_A (resp. g_B), or average square distance of the individual's gene transcription levels e^A (the vector constituted by the transcription level e_i^A of each gene i) to the optimal levels $\tilde{e}^A$ (resp. e^B and $\tilde{e}^B$). The gap g_A is computed as follows:

$$g_A(e^A) = \frac{1}{n}\sum_{i=1}^{n}(e_i^A - \tilde{e}_i^A)^2 \tag{4.4}$$

The gap g_B is computed in the same way. Finally, we compute the fitness of the individual by summing the gap in each environment, and applying an exponential scaling: $f = e^{-k(g_A+g_B)}$, where k is a scaling factor representing the selection pressure. A higher value of k means that well-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.

4.2.3 Mutational Operator: Genomic Inversions

We introduce only one kind of mutation in our model, which is genomic inversions: we choose two breakpoints randomly on the genome, and reverse the genomic content between these points. Genes are then reinserted in the genome in the opposite orientation and order, taking care to update all intergenic distances appropriately. Note that in our model, genes have a length of zero and the breakpoints can therefore not fall inside a gene. Moreover, an inversion has no effect if both breakpoints fall between two neighboring genes (as only an intergenic region would be affected), but can impact any number of genes otherwise. Genomic inversions hence affect gene syntenies and orientations, and therefore affect gene expression levels as presented in subsection 4.1.3. When mutating a genome during reproduction, 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 $P(k = 0) = e^{-\lambda} \approx 0.136$.

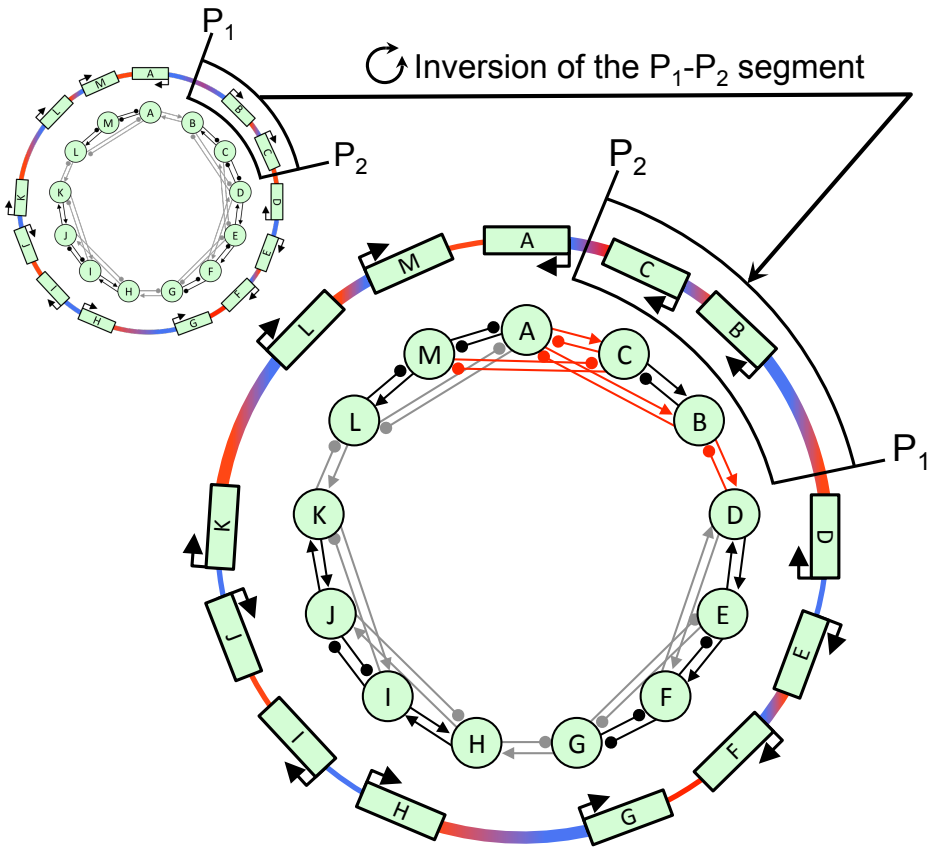


Figure 4.5: Result of the inversion of a genomic segment containing genes B and C from the individual presented in Figure 4.1. The gene interactions which have changed due to the inversion are drawn in red. This illustration genome corresponds to the actual individual in our model presented in Figure 4.4.

Figure 4.5 presents a genome obtained by performing an inversion on the genome shown in Figure 4.1. 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 genes now lie in colinear orientations, in which each of these genes activates its upstream neighbor but represses its downstream neighbor.

4.2.4 Experimental Setup and Parameter Values

We initialized 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 chose a maximum interaction distance of $d_{max} = 2500$, meaning that each gene initially interacts with its

2 closest neighbors in each direction through the transcription-supercoiling coupling. Note that as inversions may change intergenic distances, genes can move closer or further apart during evolution. We set the basal supercoiling level σ_{basal} to the average supercoiling level in *E. coli* of -0.06 (Crozat et al., 2005), and σ_0 to -0.06 as well, so that in the absence of other sources of supercoiling (either environmental or through the coupling), the default activity level of a gene is 0.5. Finally, we set $c = 0.3$, in order to have comparable values for the variations in supercoiling due to the environment and due to the transcription-supercoiling coupling, and $\varepsilon = 0.03$, so that the variations in supercoiling have a qualitatively mild effect on gene expression.

In order to run the simulations, we evolved 15 different populations for 250,000 generations; the simulation lasted for approximately 48h on a computer with Intel Xeon E5-2640 v3 @ 2.60GHz CPUs, using around 100 MB of RAM per replicate. All the data from the experiment is available online on the Zenodo platform.

4.2.5 Adaptation of Gene Expression Levels to Different Environments

Figure 4.6 summarizes the differences in the proportion of activated genes for each of the three sets of genes, between environments A and B, averaged over the 15 repetitions. In the figure, we consider a gene to be activated if its activity at the end of the lifecycle is over 0.5, and we look at the average proportion of activated genes in the best individual of every replica. Let us recall that the evolutionary target for *AB* genes is an expression level of 1 in both environments, for *A* genes an expression level of 1 in environment *A* and 0 in *B*, and vice-versa for *B* genes. After 250,000 generations of evolution, individuals have acquired genomes that allow all *AB* genes to be activated in both environments, and that allow all *B* genes to be activated in environment B and inhibited in environment A. On average, over 60% of *A* genes are activated in environment A, which imposes a positive change in supercoiling ($\sigma_A = 0.1$) and makes gene activation harder. Conversely, less than 5% of *A* genes are activated in environment B, in which gene activation is easier ($\sigma_B = -0.1$). The final expression levels of *A* genes therefore show that specific sets of genes can be activated by the transcription-supercoiling coupling despite environmental hurdles.

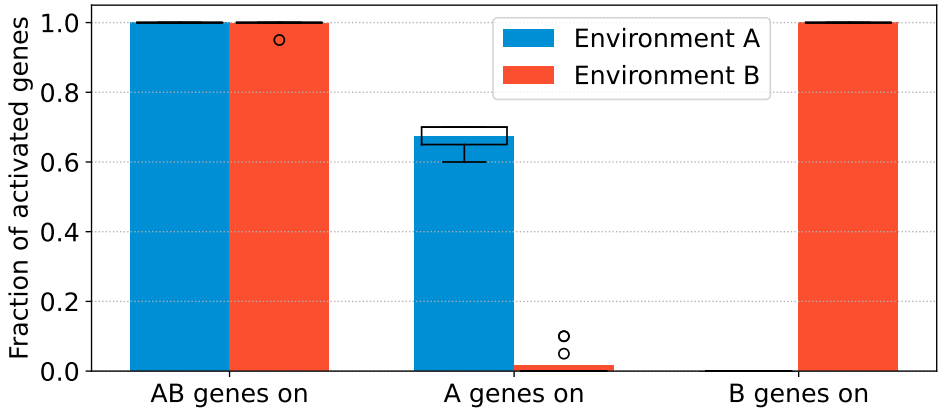


Figure 4.6: Fraction of activated genes of each type in each environment at the end of the life-cycle, averaged over the best individuals in the last generation of each replica. The boxplots represent the median and quartiles, and the dots flier data points. For A genes and B genes, activation levels differ depending on the environment: p-value 2.40×10^{-17} for A genes, and p-value $< 1 \times 10^{-25}$ for B genes (Student's t-test for dependent samples).

Furthermore, in each of the 15 replicates, the fitness of the best individual in the population increases continuously over the course of evolution, as shown in Figure 4.7. As their respective fitness keeps increasing until the end of the simulation, this suggests that fitter phenotypes remain reachable through further evolution by genomic rearrangements. The rhythm of evolution is however progressively slower and slower (note the logarithmic time scale in the figure), as the pool of available favorable mutations decreases.

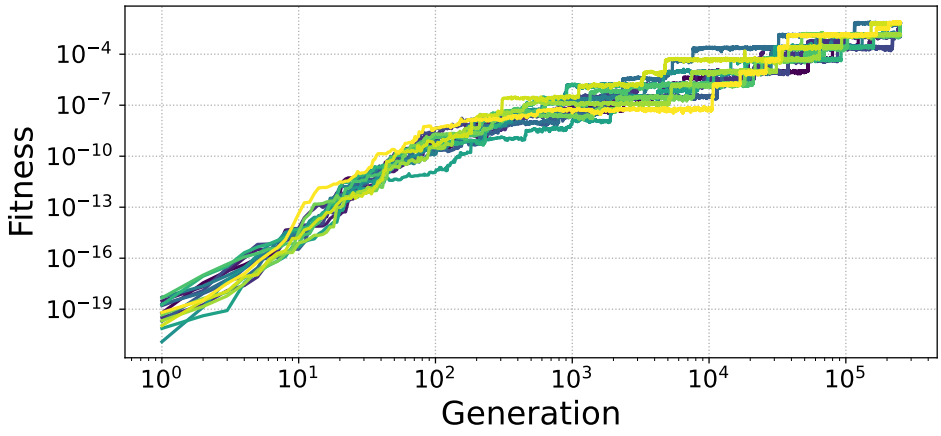


Figure 4.7: Evolution of the fitness of the best individual of each replicate at every generation.

Finally, details of the evolution of one of the 15 replicate populations are shown in Figure 4.8. We can first see that the number of activated AB genes of the best individual at each generation quickly rises to 20 (out of 20 genes of that type) in both environment A and environment B; this shows that evolving a phenotype that is resistant to environmental perturbations, having genes that are always activated, is easy in the model. For A genes

and B genes, we observe an asymmetric tendency during the course of evolution towards activation in the target environment, and inhibition in the opposite environment. However, the difference in the number of activated B genes between environment A and environment B is much higher than for A genes. As already mentioned above, this asymmetry comes from the different requirements expected of A genes and B genes: gene activation is easier in environment B than in environment A, as it is easier for a gene to become activated in an environment with a lower overall supercoiling level. A genes therefore have to be activated in a harder environment, and inhibited in a simpler environment, whereas B genes have to do the opposite.

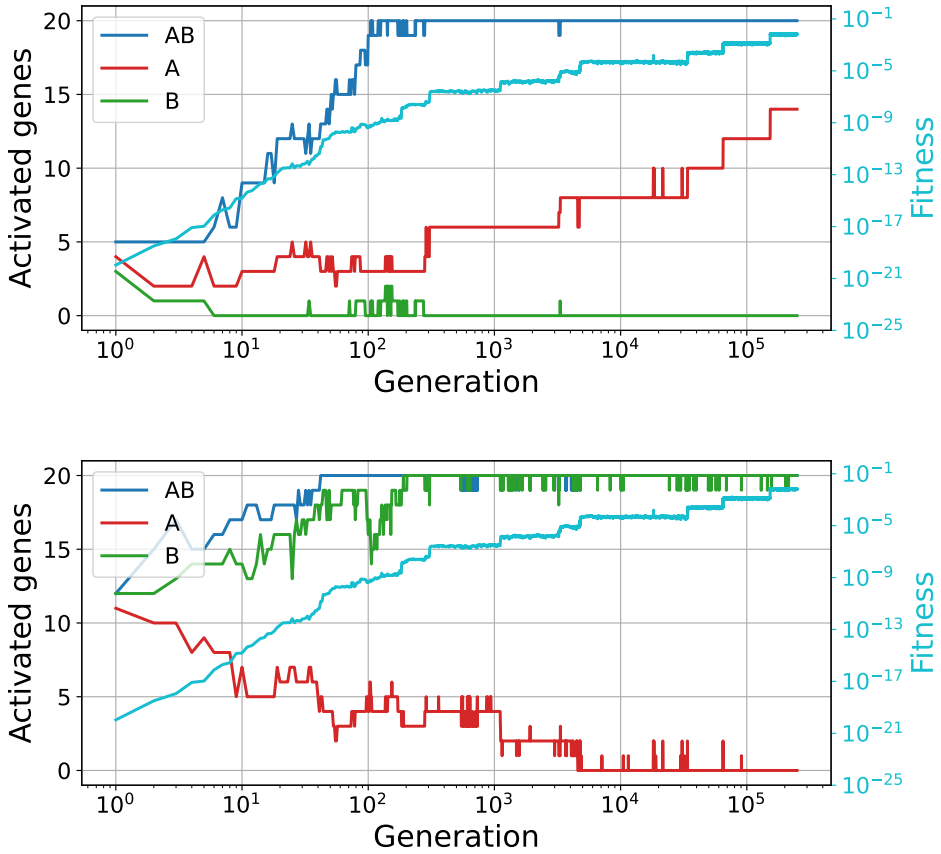


Figure 4.8: Number of activated genes of each type and fitness of the best individual at every generation of replicate 13, with a population size of $N = 100$, for 250,000 generations. The number of active AB genes increases until it reaches 20, in both environment A (top) and environment B (bottom). The number of active A (resp. B) genes increases in environment A (resp. B) and decreases in environment B (resp. A) over time, thus converging towards their evolutionary target.

This is shown in more detail in Figure 4.9, which shows the supercoiling level and gene activation levels of the best individual of the last generation of replicate 13, in both environments. The phenotypes displayed in each environment present clearly distinct gene ex-

pression patterns. In environment A (top), nearly all genes converge directly towards their final state, whereas in environment B (bottom), most *A* genes (in red) and some *B* genes (in green) show a complex trajectory of activation levels before reaching their stable state. Moreover, genomic domains with markedly different supercoiling levels emerge through the transcription-supercoiling coupling, with both very overwound and very underwound zones. These domains also show qualitatively different responses to different environments: in some domains, the supercoiling level is very similar (around gene 0, gene 15 or gene 55 for example), while in others supercoiling is completely different in each environment (between genes 20 and 35). This shows the plasticity of the response to environmental change at the local supercoiling level.

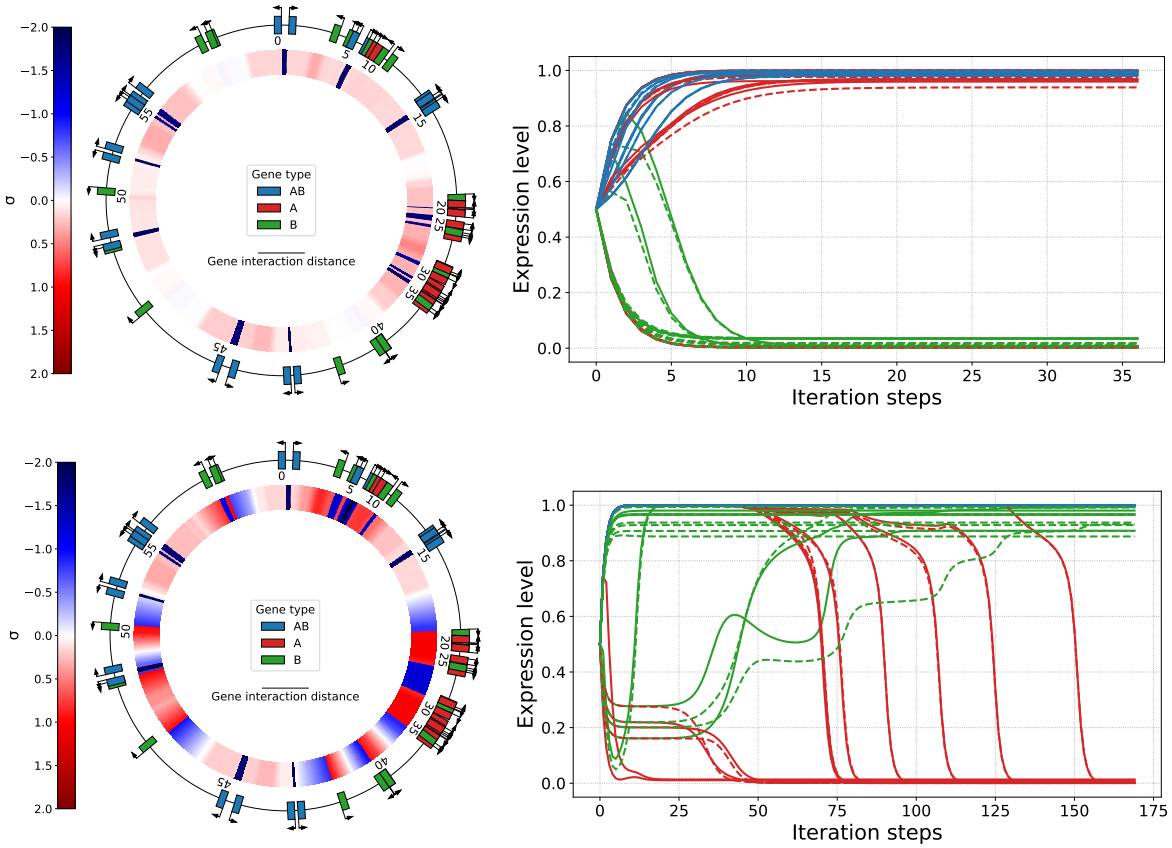


Figure 4.9: Local supercoiling along the genome and gene transcription levels of the best individual in replicate 13 after 250,000 generations. Environment A is on top and environment B at the bottom. *AB* genes are colored blue, *A* genes colored red, and *B* genes colored green.

Our experimental results show that, in a model of gene transcription that is structured around the transcription-supercoiling coupling, complex gene interaction networks can in fact evolve. These gene interaction networks are sensitive to environmental variations, which are mediated in our model by a single parameter: σ_{env}, the amount of global supercoiling that is due to the environment.

4.2.6 Robustness of Gene Network Evolution

In order to ensure that our results remain experimentally valid over a broad range of parameter values, we ran additional sets of simulations. We changed respectively the sensitivity of gene promoters to supercoiling changes (ε in equation 4.3), the interaction coefficient used in computing the local supercoiling due to the transcription-supercoiling coupling (c in equation 4.1), and the strength of the change in supercoiling imposed by the environment (σ_A and σ_B). We chose sets of logarithmically-spaced values for each parameter, and ran 5 replicates of the evolution experiment for 250,000 generations for each parameter value. Note that, for extreme parameter values, gene expression levels did in some cases not converge to stable states by the maximum number of computation steps. In this situation, we chose to retain the gene expression levels at the last step as the phenotype of the affected individuals.

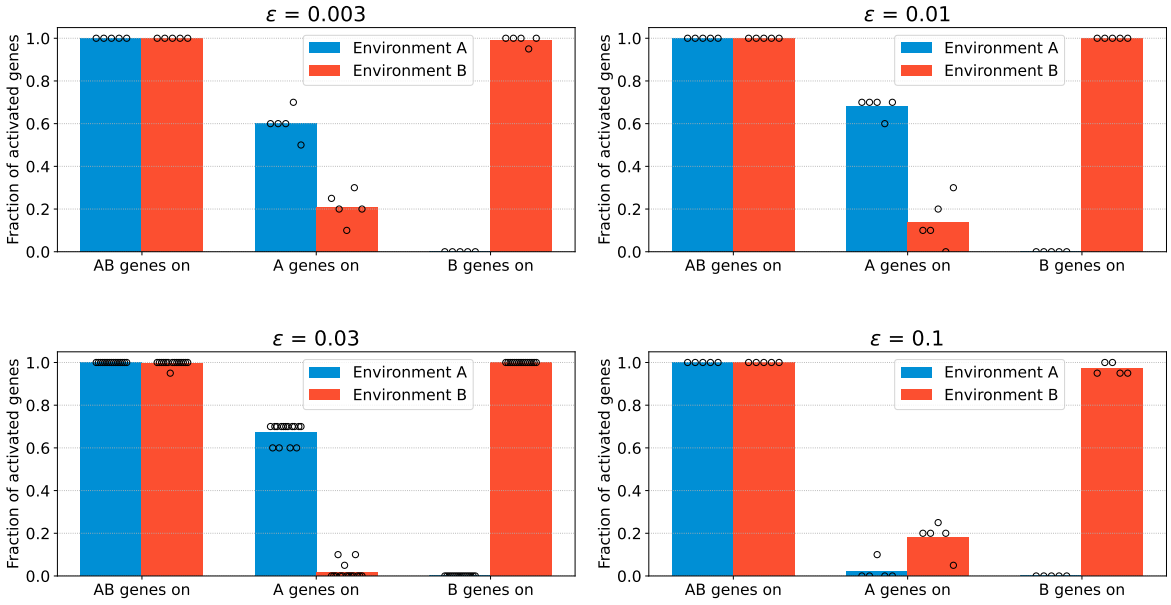


Figure 4.10: Average fraction of activated genes in each environment at the end of evolution, for increasing values of ε, from top to bottom and left to right. Every replicate is shown as a dot, and the bottom-left panel ($\varepsilon = 0.03$) recalls data from the main run (which has 15 replicates) for comparison. For all values of ε except 0.1, the behavior from the main run is qualitatively replicated.

The results of these additional simulations are presented in figures 4.10, 4.11 and 4.12. For ε, we chose values of $\varepsilon = 0.003$, $\varepsilon = 0.01$, and $\varepsilon = 0.1$, compared to an initial value of $\varepsilon = 0.03$, and the results are shown in Figure 4.10. For the values of ε lower than the default (top row), representing a higher sensitivity of promoters to supercoiling, we observe the evolution of differentiated gene expression levels as in the main run (bottom-left panel), whereas for the higher value of ε (bottom-right panel), A genes are still not expressed in environment A by the end of evolution. In this case, promoters are not sensitive enough to

the supercoiling variations caused by the transcription-supercoiling coupling, and genes are unable to overcome the highly positive supercoiling of environment A.

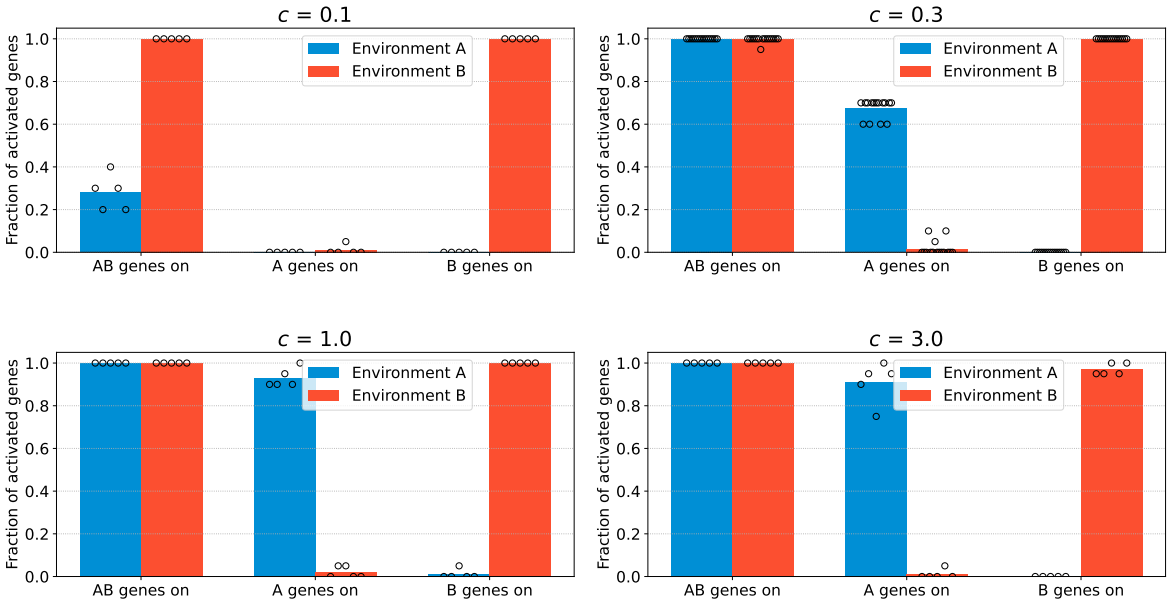


Figure 4.11: Average fraction of activated genes in each environment at the end of evolution, for increasing values of c, from top to bottom and left to right. Every replicate is shown as a dot, and the top-right panel ($c = 0.3$) recalls data from the main run for comparison. For all values of c except 0.1, the behavior from the main run is qualitatively replicated.

For c, we chose values of $c = 0.1$, $c = 1.0$, and $c = 3.0$, for an initial value of $c = 0.3$, and the results are shown in Figure 4.11. Similarly to ε, when c is too low (top-left panel), genes do not interact strongly enough for a differentiated phenotype to evolve as a function of the environment, whereas higher values of c (bottom row) show the same evolutionary behavior as the main run (top-right panel).

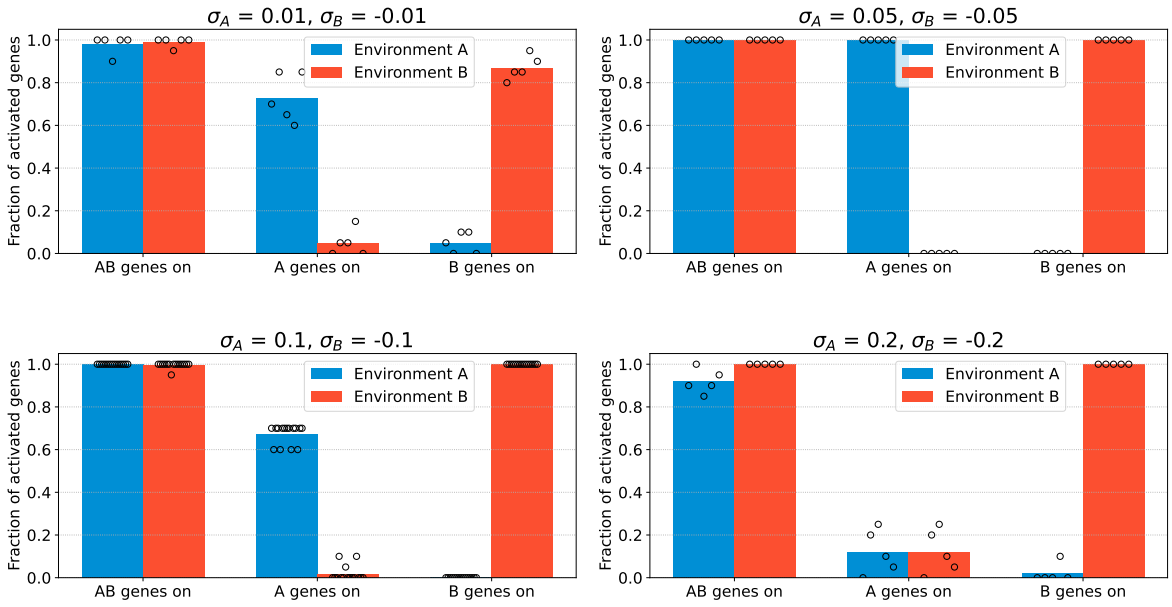


Figure 4.12: Average fraction of activated genes in each environment at the end of evolution, for more and more distinct environments σ_A and σ_B, from top to bottom and left to right. Every replicate is shown as a dot, and the bottom-left panel ($\sigma_A = 0.1$, $\sigma_B = -0.1$) recalls data from the main run for comparison. For all values except $\sigma_A = 0.2$ and $\sigma_B = -0.2$, the behavior from the main run is qualitatively replicated.

Finally, we also investigate different amplitudes in the difference in supercoiling level between the two environments, by choosing values of $\sigma_A = 0.01$, $\sigma_A = 0.05$ and $\sigma_A = 0.2$, and $\sigma_B = -\sigma_A$ respectively in each case (for an initial value of $\sigma_A = 0.1$ and $\sigma_B = -0.1$). We observe that, when $\sigma_A = 0.2$ (bottom-right panel), the environmental supercoiling constraint is too high and A genes are not activated in environment A by the end of the runs. However, for environments closer to each other than the default (top row), evolution is able to leverage the differences in supercoiling between these environments to evolve differentiated phenotypes, as in the main run (bottom-left panel), showing that our model remains sensitive to small changes in environmental supercoiling.

To conclude, in our model, the gene interaction network is therefore able to respond to different environments and can evolve an efficient regulation of gene expression under a broad range of parameter values, reinforcing the hypothesis that a supercoiling-mediated coupling between gene expression levels could indeed play a functional role in biological organisms.

4.3 Discussion and Perspectives

DNA supercoiling plays a fundamental role in the regulation of gene transcription in bacteria, and an important part of this role could be mediated by the local variations in supercoiling that are caused by the transcription-supercoiling coupling. While the influence of the global

supercoiling level on gene transcription (Lal et al., 2016; Ma and Wang, 2016; Dorman and Dorman, 2016; Martis B. et al., 2019), the evolutionary importance of supercoiling regulation (Crozat et al., 2005, 2010; Duprey and Groisman, 2021) and the mechanistic details of the transcription-supercoiling coupling (Meyer and Beslon, 2014; El Houdaigui et al., 2019) have all already been studied, no existing work did to our knowledge tackle the question of the possible role of the transcription-supercoiling coupling both at the whole-genome scale and on an evolutionary time scale.

In this work, we have developed a genome-wide model of the influence of DNA supercoiling on gene transcription, incorporating both the global influence of the environment and the local variations in the supercoiling level that are due to the transcription-supercoiling coupling. We have shown that, in our model, complex interactions implicating several genes emerge from the coupling between supercoiling and transcription. Indeed, *A* genes display an activation pattern that would not be obtainable without the network of interactions that results from the coupling. Thanks to this network, *A* genes are activated in an environment where isolated genes would be inhibited, and inhibited in an environment where isolated genes would be activated. The transcription-supercoiling coupling therefore enables the selective activation or inhibition of specific sets of genes, providing a non-monotonic response to environmental variations through changes in the level of DNA supercoiling. Furthermore, we have shown, using an *in silico* experimental evolution approach, that natural selection can leverage this biophysical mechanism to selectively turn on or off several pools of genes, using only the very simple mutation operator of genomic inversions, that affect the relative positions and orientations of genes on the genome but do not change genome length or basal gene transcription rates, and that this behavior is able to evolve under a wide range of parameter values. This response of gene transcription levels to DNA supercoiling reflects a phenomenon which has been observed *in vivo* in the expression of pathogenicity-related genes in specific environments, such as the normally lethal inside of the macrophage for the mammalian pathogen *S. enterica* (Cameron et al., 2013), or plant tissue for *D. dadantii* (Hérault et al., 2014).

Our model voluntarily stays very simple, only incorporating the most important feature of the transcription-supercoiling coupling, which is the non-linear interaction between the expression levels of neighboring genes. This simplicity therefore hints at the possible pervasiveness of this regulation mechanism throughout the prokaryotic realm. Nonetheless, in order to go further and represent more accurately the diversity of gene behaviors found in real life, several more dimensions could be integrated to the model. At present, the target for genes in our model is bistability, meaning that genes should end up fully activated or fully inhibited. A more biologically plausible approach would be to relax this restriction and give genes arbitrary expression targets, in order to determine to which extent the transcription-supercoiling coupling is able to finely regulate gene expression. Furthermore, unlike in our model (in which all genes have the same response curve to DNA supercoiling), the genes of biological organisms can show different responses to the supercoiling level. These differences are partly caused by the GC content at the gene promoter (Forquet et al., 2021), and some genes can even respond in the opposite direction to DNA relaxation, that is to say be activated rather than inhibited by less negatively supercoiled DNA. This behavior is for instance present in the *gyrA* and *gyrB* genes that encode the gyrase subunits in *E. coli* (Peter

et al., 2004). Moreover, while our model studies its role in an abstract transcription model, supercoiling intervenes during different parts of the initiation and termination of transcription, as well as in in transcript elongation (Martis B. et al., 2019). Incorporating such precise mechanistic processes into our model could give more accurate information on the link between the position of genes on the genome and their transcription rate. Similarly, increasing the number of genes of individuals in our model to match bacterial gene numbers might provide more fine-grained results, but is computationally intractable in the current implementation of the model. Furthermore, investigating the behaviors of individuals when they are placed successively in different environments, rather than evaluated separately in each environment, would also bring more information on the plasticity of the network of gene interaction levels that emerges from the transcription-supercoiling coupling. Finally, another valuable approach in order to bring this model closer to biology would be to incorporate it into a larger existing framework, such as the Aevol *in silico* experimental evolution platform (Rutten et al., 2019), which models the bacterial genome in much more detail, in order to leverage the power of a well-understood digital organism model.

Chapter 5

Structure of Supercoiling-Mediated Gene Regulatory Networks

This chapter presents the second version of the *EvoTSC* model and the results obtained with that version. The second version of *EvoTSC* builds upon the proof-of-concept presented in the previous chapter by using a more precise model of promoter sensitivity to supercoiling, and by using experimentally obtained parameter values in order to obtain more biologically meaningful results. Using this overhauled model, I study the structure of the gene regulatory networks that enable the transcriptional response of evolved genomes to different environments. The text of the chapter is an edited version of the *bioRxiv* preprint (Grohens et al., 2022a).

In previous work, we demonstrated the theoretical possibility of the evolution of conditional gene activation or inhibition in different environments in a simple model in which the sole regulatory mechanism is the local level of DNA supercoiling, and in which the only mutational operator is genomic inversions. In this work, we focus on exploring the range of genomic organizations that can be generated by selection to regulate gene expression levels in different environments through the transcription-supercoiling coupling. To that end, we start by presenting a version of the model that represents bacterial genomes more closely, as this mechanism has been proposed to be an important factor that shapes the organization of bacterial genomes in particular. Using this more precise model, we observe the emergence of complex environment-driven patterns of gene expression, and characterize the spatial organization of genes along the genome that underlie these patterns. We first show that genes are locally organized in convergent or divergent pairs that leverage the transcription-supercoiling coupling for mutual activation or inhibition, and observe the emergence of relaxation-activated genes, as described in bacterial genomes. Then, we show that this local organization is not entirely sufficient to fully account for the complex gene expression patterns that we observe in the model, but that gene inhibition in particular requires the interaction of a large number of genes. Finally, we show that, in our model, genes form a dense genome-wide regulatory network, providing insight into the regulatory role of the organization of bacterial genomes.

5.1 An Evolutionary Model of the Transcription-Supercoiling Coupling

5.1.1 Individual-Level Model

We define the genotype of an individual in our model as a single circular chromosome that is representative of a bacterial chromosome. The chromosome consists in a fixed number of protein-coding genes, which are separated by non-coding intergenic segments of varying sizes, and has a basal supercoiling level σ_{basal}. Each gene on the chromosome is characterized by its starting position (genes cannot overlap in our model), its orientation (on the forward strand or on the reverse strand), its length, and its basal expression level. We define an environment by the shift $\delta\sigma_{env}$ that it imposes to the supercoiling level of the chromosome. We then define the phenotype of an individual in a given environment as the gene expression levels that are solution of the system given by the interaction of its genes through the transcription-supercoiling coupling (described below), on a chromosome with a background supercoiling level of $\sigma_{basal} + \delta\sigma_{env}$.

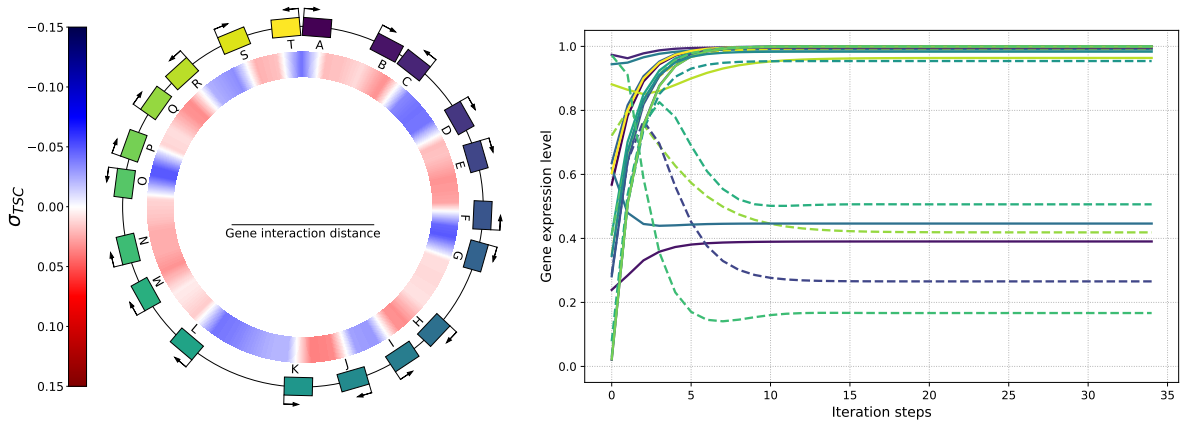


Figure 5.1: Left: genome (outer ring) and level of supercoiling generated by transcription (σ_{TSC}, inner ring) of an example genome with 20 genes placed at random positions and orientations and colored by position, with a gene length and average intergenic distance of 1 kb each, and a basal supercoiling level of $\sigma_{basal} = -0.066$. The individual is evaluated in an environment in which $\delta\sigma_{env} = 0$. Right: evolution of the expression level of each gene of the individual during the computation of the solution to the system given by equations 5.2, 5.3, and 5.4, starting from random initial values. Solid lines represent genes on the forward strand, dashed lines genes on the reverse strand, and gene colors are the same as on the genome.

The genome of an example individual with 20 genes and a basal supercoiling $\sigma_{basal} = -0.066$ is shown on the left-hand side of Figure 5.1. Inside the genome is the resulting local level of DNA supercoiling when this individual is evaluated in an environment with a supercoiling shift of $\delta\sigma_{env} = 0$. As expected given the twin-domain model of supercoiling, we can

observe a buildup in negative supercoiling (blue) between genes in divergent orientations, such as genes C and D or F and G, and a buildup in positive supercoiling (red) between genes in convergent orientations, such as genes J and K or Q and R. The right-hand side panel of Figure 5.1 shows the computation of the stable state of gene expression levels for this individual. Note that, in this model and throughout this chapter, we conflate gene transcription rates with mRNA concentrations, as we assume that mRNAs degrade at a constant rate, and as transcription rates in our model are only affected by the effect of supercoiling on transcription. We furthermore conflate transcription rates with expression levels, as we again assume proteins to be translated at a rate proportional to the associated mRNA concentrations and to degrade at a constant rate.

Effect of Transcription on Supercoiling For an individual with a genome containing n genes, each expressed at a level e_i, we model the influence of the transcription of each gene on the level of supercoiling at the promoter of every other gene in the form of an n-by-n interaction matrix. The coefficient $\frac{\partial \sigma_i}{\partial e_j}$ at indices (i, j) in this matrix represents the variation in DNA supercoiling at the promoter of gene i due to the transcription of gene j. The value of this coefficient is given by the following formula:

$$\frac{\partial \sigma_i}{\partial e_j} = \eta \cdot c \cdot \max(1 - \frac{d(i,j)}{d_{max}}, 0) \tag{5.1}$$

η represents the sign of the interaction, which depends on the position and orientation of gene j relative to gene i, according to the twin-domain model. If gene j is upstream of gene i, and if it is on the same strand as (or points towards) gene i, then its transcription generates a buildup in positive supercoiling at gene i ($\eta = 1$). Conversely, if gene j is upstream of gene i but on the other strand than (or points away from) gene i, it generates a buildup in negative supercoiling at gene i ($\eta = -1$). If gene j is instead located downstream of gene i, the sign of the interaction in each case is switched: $\eta = 1$ if the genes are on the same strand, and $\eta = -1$ otherwise.

We then apply a torsional drag coefficient c, which represents the intensity to which the transcription-generated torsion of DNA affects the local supercoiling level through drag. Finally, the strength of the interaction decreases linearly with the distance $d(i, j)$ between the promoter of gene i, which is the position where the local level of supercoiling affects the probability that an RNA polymerase binds to the DNA and starts transcribing gene i, and the middle of gene j, which is the average location of the RNA polymerases that transcribe gene j, assuming that DNA is transcribed at a constant speed. When this distance reaches a threshold of d_{max}, the two genes are considered to be too far away to interact and the effect vanishes; in other words, d_{max} represents the maximum interaction distance on either side of the gene.

Effect of Supercoiling on Transcription In order to compute the transcription level of a given gene, we first compute the opening free energy of its promoter, which depends on the local supercoiling level, following a sigmoidal curve that increases with negative supercoiling until a saturation threshold is reached (Forquet et al., 2021). In order to model this effect, we

adapted the equations and parameter values presented in El Houdaigui et al. (2019), which are based on the *in vitro* analysis of the transcription of model bacterial promoters. We first compute the local level of supercoiling σ_i at the promoter of gene i, which is the sum of the background supercoiling level $\sigma_{basal} + \delta\sigma_{env}$ (which is constant along the genome for any given individual), and of the local variation in supercoiling caused by the transcription of every other gene (represented in Figure 5.1 as σ_{TSC}):

$$\sigma_i = \sigma_{basal} + \delta\sigma_{env} + \sum_{j=1}^{n} \frac{\partial \sigma_i}{\partial e_j} e_j \tag{5.2}$$

We compute the expression level of the gene using a thermodynamic model of transcription. First, we compute the opening free energy U_i of the promoter of gene i, which depends on σ_i, the level of supercoiling at the promoter and on σ_0, the level of supercoiling at which the opening free energy is at half its maximum level, according to the following sigmoidal function:

$$U_i = \frac{1}{1 + e^{(\sigma_i - \sigma_0)/\varepsilon}} \tag{5.3}$$

The, we compute the expression level e_i of gene i using the promoter opening free energy, with a scaling constant m:

$$e_i = e^{m(U_i - 1)} \tag{5.4}$$

The transcription level of a gene is therefore expressed in arbitrary units between e^{-m}, the minimum expression level when the promoter is most hindered by supercoiling (when U_i = 0), and 1, the maximum expression level, when the promoter is most activated by supercoiling (when U_i = 1). Throughout this chapter, we will describe a gene as activated if its transcription level is above the mean of these two values $e_{1/2} = \frac{1}{2}(e^{-m} + 1)$, and inhibited otherwise.

Computation of Gene Expression Levels We define the phenotype of an individual in an environment (described by $\delta\sigma_{env}$) as the set of gene expression levels that is solution to the system given by equations 5.2, 5.3 and 5.4, in that environment. In order to compute this phenotype, we numerically compute a solution to the system of equations using a fixed-point iteration algorithm, and starting from an initial state in which all genes are expressed at $e_{1/2}$. A representative example of this computation can be found in Figure 5.1: After an initially unstable phase, the algorithm quickly converges to a fixed point of expression levels.

5.1.2 Evolutionary Model

Equipped with a model of the coupling between DNA supercoiling and gene transcription at the whole-genome scale, we now extend it into an evolutionary framework. In order to study the transcriptional response of individuals placed in different environments, we model the evolution of a population of individuals, each behaving as described in subsection 5.1.1, in two distinct environments named A and B. Environment A is a DNA relaxation-inducing

environment, with a supercoiling shift of $\delta\sigma_{env} = \delta\sigma_A > 0$, and environment B is a DNA hypercoiling-inducing environment, with a supercoiling shift of $\delta\sigma_{env} = \delta\sigma_B < 0$. We then define three classes of genes with environment-specific target expression levels: *AB* genes should be expressed in both environments, akin to housekeeping genes; *A* genes should be expressed in environment A but not in environment B; and, conversely, *B* genes should be expressed in environment B but not in environment A; both classes represent environment-specific genes such as the pathogenic genes of *S. enterica* or *D. dadantii* (Cameron and Dorman, 2012; Hérault et al., 2014).

Fitness Let (e_A^A, e_B^A, e_{AB}^A) be the average gene expression level per gene type of an individual with n genes in environment A, (e_A^B, e_B^B, e_{AB}^B) the average gene expression per type in environment B, and $(\tilde{e}_A^A, \tilde{e}_B^A, \tilde{e}_{AB}^A)$ and $(\tilde{e}_A^B, \tilde{e}_B^B, \tilde{e}_{AB}^B)$ be target expression values for each gene type in each environment. For environment A, we choose to set $\tilde{e}_A^A = \tilde{e}_{AB}^A = 1$, and $\tilde{e}_B^A = e^{-m}$, which are respectively the maximal and minimal attainable gene expression levels in the model. Similarly, for environment B, we set $\tilde{e}_B^B = \tilde{e}_{AB}^B = 1$, and $\tilde{e}_A^B = e^{-m}$. We then compute the sum g of the squared error between the mean and targeted expression levels for each gene type in each environment:

$$g = \sum_{i \in \{A,B,AB\}} \left(e_i^A - \tilde{e}_i^A\right)^2 + \sum_{i \in \{A,B,AB\}} \left(e_i^B - \tilde{e}_i^B\right)^2 \tag{5.5}$$

Finally, we define the fitness of the individual as $f = \exp(-k \cdot g)$, where k is a scaling factor representing the intensity of selection: as k increases, the difference in fitness, and hence in reproductive success, between individuals with different values of g also increases.

Generational Evolutionary Algorithm At each generation, we compute the fitness of each individual, by computing their gene transcription levels in each environment, as previously described. In order to create the next generation, we choose a parent from the current population for each individual in the new population, with a probability proportional to the fitness of the parent. Then, we create the genome of the new individual by stochastically applying mutations to the genome of its parent.

Mutational Operator: Genomic Inversions As this work aims at studying genome organization, we chose to use genomic inversions as the only mutational operator, so that genes can be reordered on the chromosome through evolutionary time; note that other genomic rearrangements, such as translocations, can be modeled as a series of well-chosen consecutive inversions, and are therefore implicitly present in our model.

In order to perform a genomic inversion, we choose a start point and an end point uniformly at random, in the non-coding intergenic sections. This ensures that genes cannot be broken apart by inversions, as we assume that gene losses are lethal and therefore never conserved. Having chosen the ends of the inversion, we extract the DNA segment between these ends and insert it in the reverse orientation. The inversion therefore switches the orientation of every gene inside the segment, but conserves the relative positions and distances

of these genes. Note that, contrary to the intergenic sections that are inside of the inversion, the intergenic sections that are at its boundaries change according to the position of the start and end points of the inversion, allowing the distances between genes to change over evolutionary time.

When mutating an individual, we first draw a number of inversions to perform from a Poisson law of parameter $\lambda = 2$, meaning that the offspring of the individual will on average undergo two inversions, and then perform each inversion in succession to obtain the final mutated offspring.

5.2 Results

In this section, we first show that, as in the proof-of-concept model presented in (Grohens et al., 2021), populations of individuals in the model presented in Section 5.1 evolve gene expression levels that match their targets in each environment. Then, we show that, consistently with the theoretical expectations of the twin-domain model, the genomes of evolved individuals are enriched in pairs of divergent or convergent genes that leverage the transcription-supercoiling coupling to regulate gene expression. Finally, we show that the gene regulatory network generated by the transcription-supercoiling coupling cannot simply be recapitulated by these local interactions, but rather encompasses the whole genome.

5.2.1 Experimental Setup

We evolved 30 populations of 100 individuals, each starting from clones of a random individual with 60 genes (20 of each type), for 1,000,000 generations. The parameter values that we used are given in Table 5.1, and can be broadly grouped into genome-level parameters (gene length, intergenic distance, basal supercoiling level and supercoiling transmission distance) and promoter-level parameters (promoter opening threshold and energy, crossover width). Both the genome-level parameters that describe the chromosome and the promoter-level parameters used to compute the transcriptional response to supercoiling were taken from experimental values measured in *E. coli*. In our model, we introduced the torsional drag coefficient as a new parameter that represents the influence of torsional drag on the local level of supercoiling, and empirically chose its value so that this effect is of the same magnitude as that of the other sources of supercoiling variations.

Parameter	Symbol	Value	Reference
Gene length	l	1,000 bp	Blattner (1997)
Initial intergenic distance	d_0	125 bp	Blattner (1997)
Supercoiling transmission distance	d_{max}	5,000 bp	Postow et al. (2004)
Basal supercoiling level	σ_{basal}	-0.066	Crozat et al. (2005)
Torsional drag coefficient	c	0.03	
Promoter opening threshold	σ_{opt}	-0.042	El Houdaigui et al. (2019)
Inverse promoter opening energy	m	2.5	El Houdaigui et al. (2019)
Crossover width	ε	0.005	El Houdaigui et al. (2019)

Table 5.1: Table of parameter values used in the evolutionary runs. The upper set of parameters is the genome-level parameters, the lower set the promoter-level parameters, both taken from the *E. coli* literature; the middle parameter is a new addition from our model.

The simulation was implemented in Python, with computationally heavy parts optimized using the `numba` package (Lam et al., 2015). The source code for the simulation, as well as the data analysis code, are available online at `https://gitlab.inria.fr/tgrohens/evotsc`. Running the complete simulation took around 36 hours of computation on a server using a 24-core Intel Xeon E5-2620 v3 @ 2.40GHz CPU, with each replicate running on a single core and using approximately 300 MB of RAM. All the data from this experiment is available online on the Zenodo platform.

5.2.2 Evolution of Regulation by the Transcription-Supercoiling Coupling

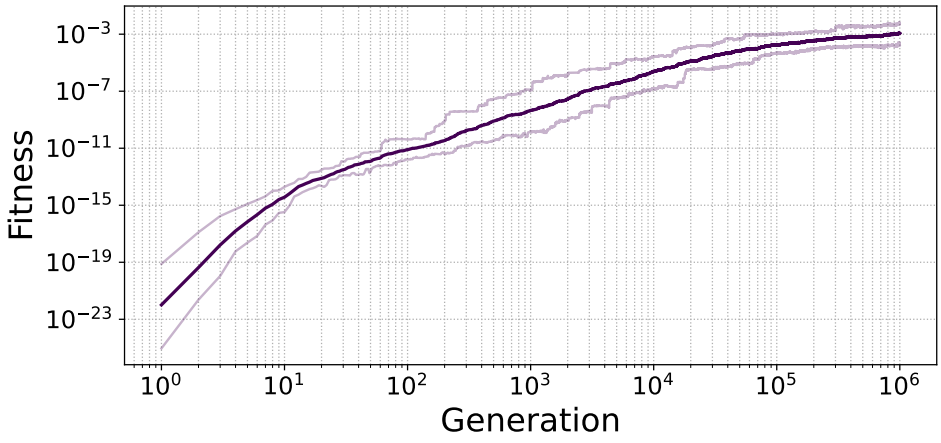


Figure 5.2: Geometric average of the fitness of the best individual in each of the 30 replicates, at every generation. Lighter lines represent the first and last decile of the data.

In our simulations, the fitness of the best individual in each population increases over evolutionary time, as shown in Figure 5.2, meaning that evolution is able to select phenotypes that are closer and closer to the target. More precisely, the expression levels of the genes of each type in individuals in our model therefore evolve towards their respective targets, as previously defined in Section 5.1.2.

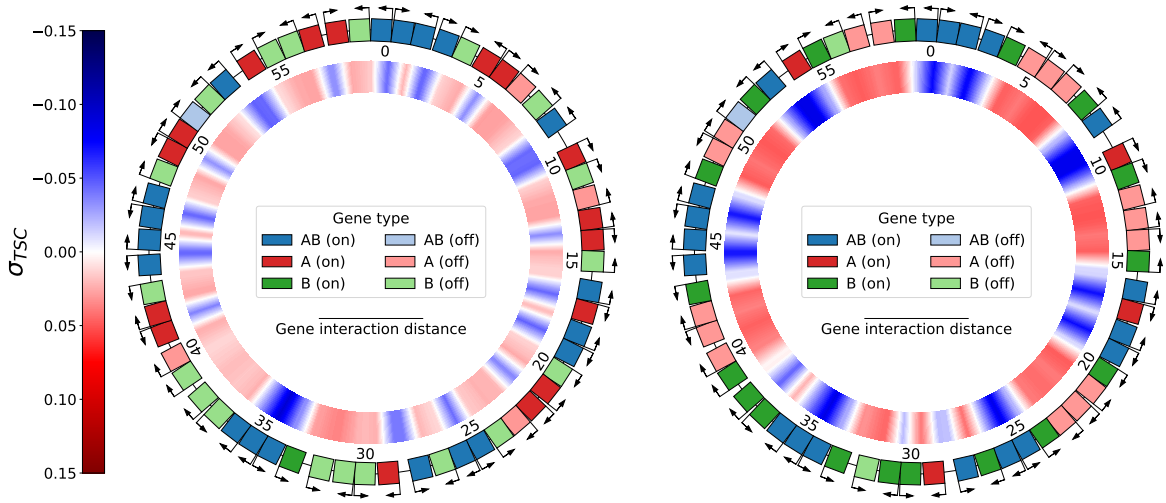


Figure 5.3: Genome of the best individual at the last generation of replicate 21, evaluated in environments A (left) and B (right). In addition to Figure 5.1, the outer ring shows the state of each gene: dark color, activated – light color, inhibited. The inner ring shows the level of transcription-generated DNA supercoiling at every position on the genome: Shades of blue represent negative supercoiling, and shades of red positive supercoiling.

The genome of an example evolved individual at the end of the simulation is depicted in Figure 5.3, along with its level of local supercoiling and gene activity in each environment. Different activation patterns for each gene class are clearly visible on the genome of this individual. Indeed, all *AB* genes except one are activated (dark blue) in each environment, whereas 19 out of 20 *B* genes are correctly inhibited (light green) in environment A (left) and 18 correctly activated (dark green) in environment B (right). Conversely, 16 *A* genes are activated (dark red) in environment A, and 16 inhibited (light red) in environment B.

The transcription-generated supercoiling that is represented in the inner ring furthermore changes consistently with the gene activation patterns between the two environments: red zones, where DNA is positively supercoiled, contain inhibited genes, whereas blue zones, where DNA is negatively supercoiled, contain activated genes. This individual therefore shows that it is possible for evolution to adjust the gene expression levels of an individual in our model to an environment-dependent target, by relying only on the transcription-supercoiling coupling and on the relative positions of genes on the genome.

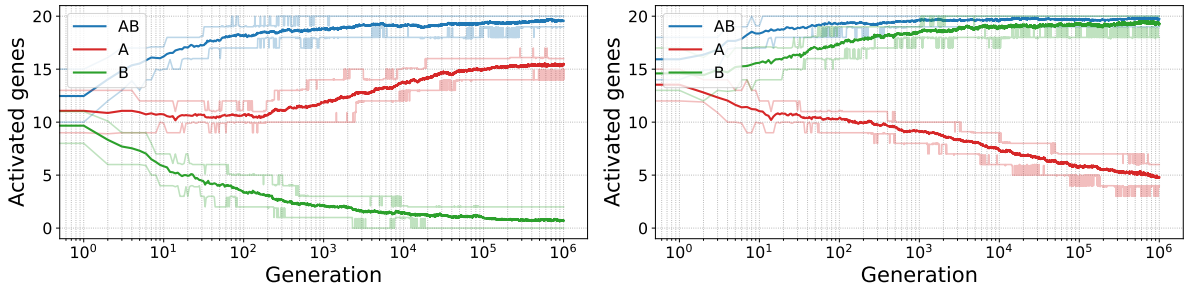


Figure 5.4: Average number of activated genes (with an expression level above $e_{1/2}$) of each type, out of 20, in the best individual at every generation, averaged over the 30 replicates, in environments A (left) and B (right). Lighter lines represent the first and last decile of the data.

Evolution of Class-Specific Gene Expression Levels These results are however not specific to this particular individual. Figure 5.4 shows that, averaging over all replicates, the number of activated genes in each class evolves towards their respective target. In each environment, the average number of activated *AB* genes quickly reaches nearly 20, its maximum value, as expected from their target; *B* genes follow the same behavior, evolving towards nearly full activation in environment B and nearly full inhibition in environment A. *A* genes follow a slightly different course, as the number of activated *A* genes seems to converge to approximatively 15 out of the expected 20 in environment A, but continues to decrease towards the expected 0 in environment B by the end of the simulations.

The incomplete match to their target of *A* genes does however not come as a complete surprise. Environment A is indeed characterized by a positive supercoiling shift $\delta\sigma_A > 0$, while environment B is characterized by a negative supercoiling shift $\delta\sigma_B < 0$. As positive supercoiling hinders promoter opening, it is more difficult for a gene to have a high transcription rate in environment A than in environment B. *A* genes must therefore complete the more difficult task of being activated in the "hard" environment A, while being inhibited in the "easy" environment B. Differentiated expression levels nonetheless evolve in our model for each type of gene, as a result of the different supercoiling levels imposed by the environmental conditions.

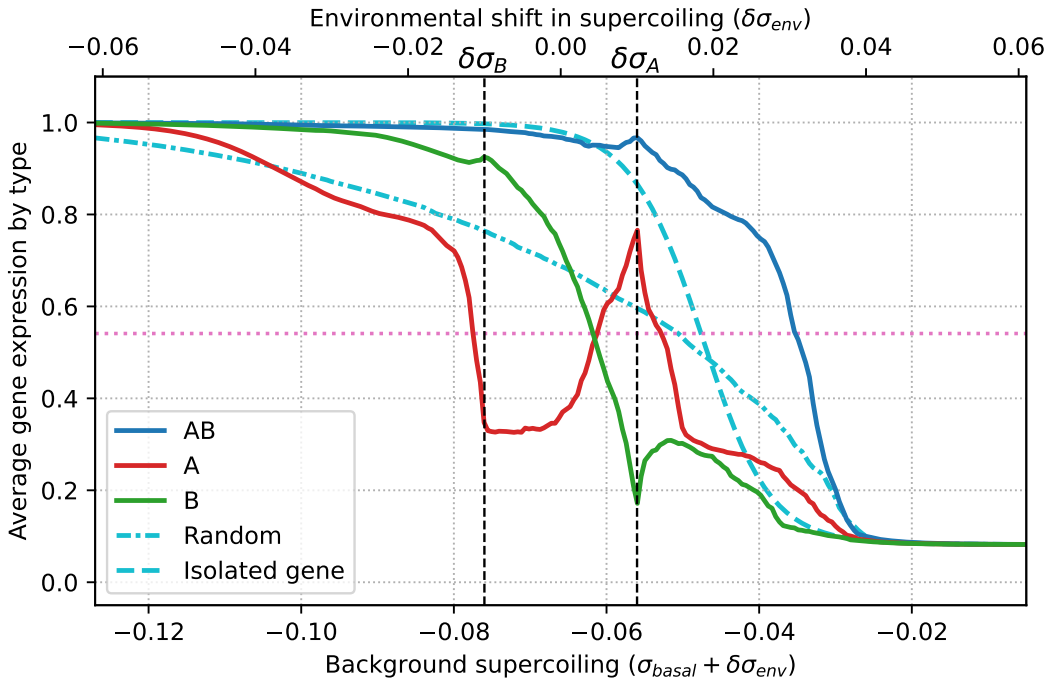


Figure 5.5: Average gene expression level for each class of gene (*A*, *B*, and *AB*), as a function of the background supercoiling level $\sigma_{basal} + \delta\sigma_{env}$, averaged over the best individual of each of the 30 replicates. The dash-dotted light blue line represents the average expression level of genes on a random genome, and the dashed light blue line represents the expression level of a single neighbor-less gene. The black vertical lines represent environments A and B, in which individuals evolve during the simulation, and the pink horizontal line marks $e_{1/2}$, the threshold above which a gene is considered active.

Evolution of Relaxation-Activated Genes In our model, the expression level of a gene increases exponentially with the opening free energy of its promoter, which itself increases as a sigmoidal function of negative supercoiling. When measuring the response of an individual's genes to variation in the background supercoiling $\sigma_{basal} + \delta\sigma_{env}$, one could therefore expect a qualitatively similar response.

Figure 5.5 shows the responses of genes of different types to the background supercoiling level (as explained in equation 5.2), and highlights striking differences between the expression of evolved, random, or isolated genes, as well as between the different gene types in evolved genomes. The light blue lines in the figure serve as a reference point, showing the response of an isolated, non-interacting gene to environmental supercoiling (dashed line), and the average response (dash-dotted line) of genes on 30 random genomes, generated using the parameters from Table 5.1. While *AB* and *B* genes (blue and green curves) display an expression level that decreases with the level of negative supercoiling, and that remains qualitatively similar to the behavior of random genes (dash-dotted line), *A* genes display a completely different behavior. These genes show a non-monotonic response to environmental supercoiling, as their expression level decreases until a local minimum in expression at

$\delta\sigma_B$, then increases again even though negative supercoiling decreases until a local maximum at $\delta\sigma_A$, before decreasing again like the other kinds of genes. In other words, A genes present a phenotype of activation by environmental relaxation of DNA, for values between $\delta\sigma_B$ and $\delta\sigma_A$, even though the promoter activity of an isolated A gene decreases with DNA relaxation.

The transcription-supercoiling coupling therefore provides a regulatory layer that mediates the transcriptional response to the global variation in DNA supercoiling caused by different environments. Indeed, it remarkably allows in our model for the evolution of a response that is opposite not only to the response displayed by a non-interacting, neighborless gene, but also to the response of genes placed at random on a similar genome, demonstrating the importance of the relative position of genes on the genome.

5.2.3 Evolution of Local Genome Organization

Having characterized the different patterns of gene transcription that evolved in our simulations in response to the two different environmental conditions, we sought to determine the genome organization that necessarily underlies these patterns, since the only difference between individuals in our model is the relative position and orientation of the genes on their genome.

We started by studying genome organization at the local level, and measured the relative abundance of pairs of neighboring genes in every relative orientation: convergent, divergent, or in tandem. The relative orientation between neighboring genes determines the mode of interaction between these genes, by applying the twin-domain model of transcription-generated supercoiling to the promoter of each gene: mutual activation for divergent genes, mutual inhibition for convergent genes, and activation (resp. inhibition) of the upstream (resp. downstream) gene by the downstream (resp. upstream) gene.

As the different gene types must evolve different activation patterns in each environment to have a high fitness in the model, we separated the pair counts by the type of each gene in the pair, resulting in 9 kinds of pairs. Finally, in order to quantify the actual strength of the coupling between the genes in a given type of pair, we summed the total level of positive and negative supercoiling generated by the transcription of each gene in the pair at the promoter of the other gene for all relative orientations. The results are presented in Figure 5.6, with the left-hand side panel showing the number of pairs of each kind, and the right-hand side panel the corresponding transcription-generated supercoiling levels. Several patterns markedly emerge from the data.

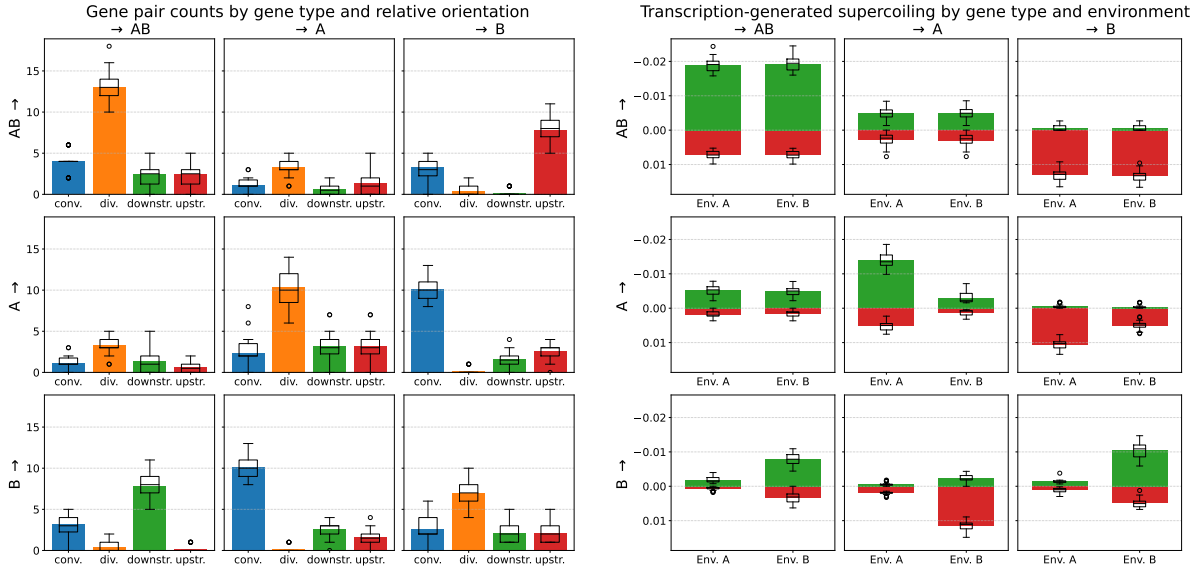


Figure 5.6: Interactions between pairs of neighboring genes. The left-hand side panel shows the number of pairs of each kind, split by the type of the first gene (sub-row) and of the second gene (sub-column) in the pair, and by relative orientation (bars in each sub-panel: convergent, divergent, upstream, or downstream). For instance, the top-right panel shows the influence of AB genes on B genes, and the bottom-left panel the influence of B genes on AB genes (in the same pairs). In that case, there are on average 7.8 *AB* genes directly upstream of a *B* gene (in red), or 7.8 *B* genes directly downstream of an *AB* gene (in green) on an evolved genome. The right-hand side panel shows, for each kind of pair, the total amount of positive (red) and negative (green) transcription-generated supercoiling due to each gene type (sub-row) measured at the promoter of each gene type (sub-column), summing over all orientations, but in each environment. All data is averaged over the best individual of each of the 30 replicates, and box plots indicate the median and dispersion between the replicates.

Genomes Are Enriched in Divergent *AB*/*AB* Gene Pairs The most frequently found kind of gene pair in the evolved genomes is divergently oriented *AB*/*AB* pairs. 13 such pairs are found on average (see the *AB*/*AB* sub-panel on the left-hand side of Figure 5.6), out of a possible maximum of 20 (since any given gene can only be part of a single divergent pair), meaning that two-thirds of *AB* genes are part of a divergent pair with another *AB* gene. The mostly divergent *ABAB* gene pairs generate an average negative supercoiling of around -0.012 at their promoters, in both environments (summing the positive and negative bars in the *AB*/*AB* sub-panel on the right-hand side of Figure 5.6). This value is comparable in magnitude to but has the opposite sign than the shift in supercoiling caused by environment A ($\delta\sigma_A = 0.01$), showing that the interaction between neighboring genes can locally counteract the global shift in supercoiling caused by this environment in order to maintain high gene expression levels.

Genomes also contain divergent *A*/*A* and *B*/*B* gene pairs, although less frequently than divergent *AB*/*AB* pairs. As both *A* genes and *B* genes must be conditionally expressed or inhibited depending on the environment, the unconditionally positive feedback loop resulting

from a divergent orientation seems less evolutionarily favorable for A/A or B/B pairs than for AB/AB pairs. Divergent A/A and B/B pairs moreover result in slightly weaker interactions (middle and bottom-right sub-panel of the right-hand side of Figure 5.6), in the environment in which these genes are active. On the contrary, divergent A/B gene pairs are almost never found, and this is consistent with theoretical expectation, since A and B genes must no be expressed in the same environment.

The local organization of the genome in divergent AB/AB gene pairs therefore seems to be favored by evolution, as this pattern allows for a high expression of these genes in both environments, while divergent A/B gene pairs, which would lead to a lower fitness, are oppositely very rarely found in evolved genomes.

Genomes are Enriched in Convergent *A/B* Gene Pairs The pattern in which B genes appear most frequently, and A genes very frequently (just after divergent A/A pairs), is in convergent A/B gene pairs. In this case, each gene in the pair should theoretically inhibit the expression of the other gene. In environment A, A genes indeed generate an average positive supercoiling variation of 0.01 at the promoter of convergently oriented B genes (the effect of B genes on convergent A genes in environment B is similar), decreasing their expression with a strength that is again comparable to the environmental change in supercoiling, whiles B genes are mostly inhibited and therefore do not impact A genes. In environment B, it is instead B genes that strongly inhibit A genes through the generation of positive supercoiling.

Convergently oriented A/B gene pairs therefore form toggle switches, or bistable gene regulatory circuits, in which the expression of one gene represses the expression of the other gene (Gardner et al., 2000). In accordance with the targeted expression patterns of A and B genes, we therefore observe that the local organization of the genome into toggle switches, like the divergent A/B pairs, is favored by evolution in order to produce environment-dependent differentiated expression levels.

5.2.4 Local Interactions Do Not Recapitulate the Regulatory Network

In order to understand the extent to which the gene regulatory network generated by the transcription-supercoiling coupling can be reduced to the local organization into pairs of genes described above, we expanded our scope to study the behavior of subnetworks of neighboring genes of increasing sizes.

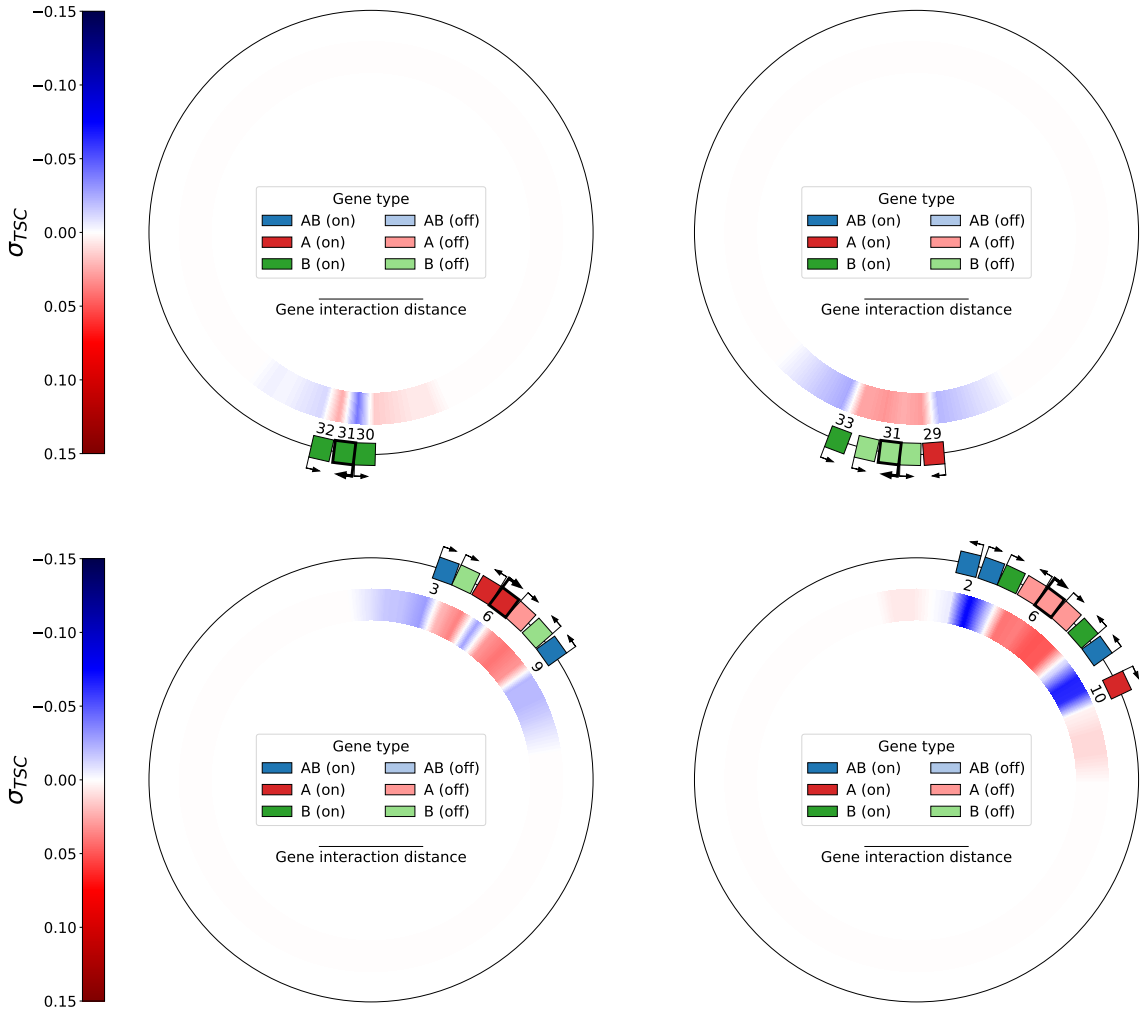


Figure 5.7: Top: subnetworks of size 3 (left) and 5 (right) centered around gene 31 (of type *B*, in bold) of the best individual at the end of replicate 21, evaluated in environment A. Bottom: subnetworks of size 7 (left) and 9 (right), centered around gene 6 (of type *A*, in bold) of the same individual, evaluated in environment B.

For every odd subnetwork size k between 1 and the genome size, and for every gene on the genome, we extracted the subnetwork of k consecutive genes centered around that gene, and computed the expression level of every gene in this subnetwork, in the same way as for a complete genome, in each environment. This allowed us to compute the minimum subnetwork size at which a gene has the same activation state as in the complete genome, which we interpret as an indicator of the complexity of the interaction network necessary to produce the activation state of that gene in the complete genome. Two representative examples are presented in Figure 5.7, and the complete results are then shown in Figure 5.8.

Figure 5.7 depicts the subnetworks that are needed in order to obtain the inhibition of a representative gene of type A in environment B, and of a representative gene of type B in

environment A, taken from the genome of an evolved individual. The *B* gene is not inhibited by a subnetwork of size 3, but needs a subnetwork of size 5 to be inhibited, and similarly, the *A* gene is not inhibited by a subnetwork of size 7, but needs a subnetwork of size 9 to be inhibited. In each case, increasing the size of the subnetwork by two (one gene on each side) completely changes the resulting gene expression levels, alongside with the associated level of transcription-generated supercoiling. Indeed, in the top example, all 3 genes in the small subnetwork switch states when evaluated inside the larger subnetwork, and in the bottom example, the two *B* genes and two out of the three central *A* genes switch activation states when moving from the small to the large subnetwork. In these examples, the activity of a gene is therefore not only dependent on its closest neighbors, but on a quite larger section of the genome.

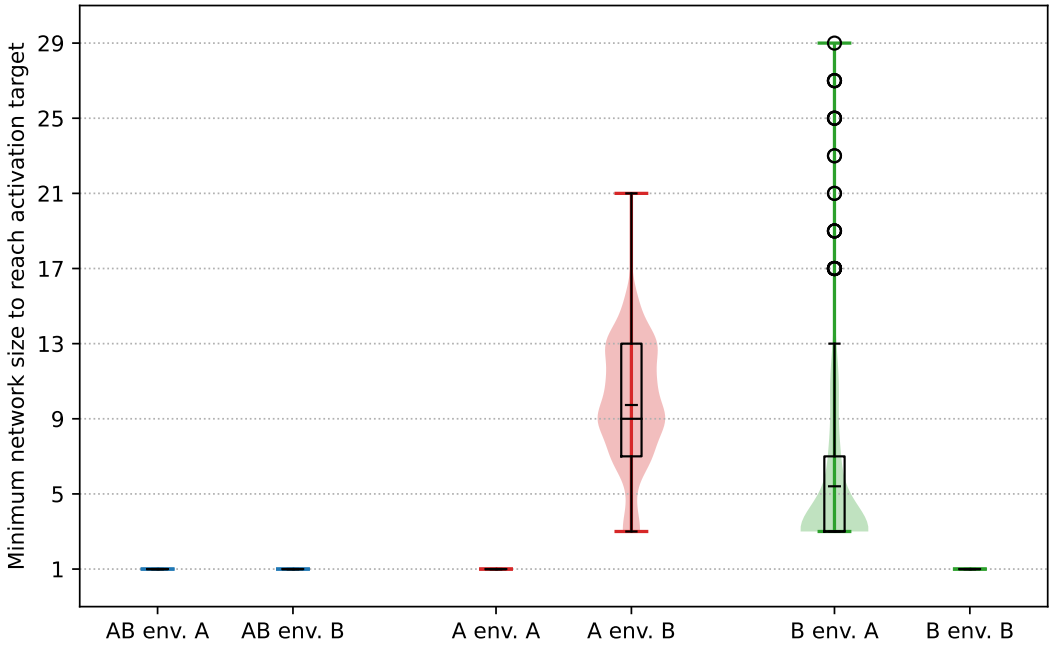


Figure 5.8: Minimal contiguous subnetwork size needed for the central gene in the subnetwork to have the same activation state as in the complete genome, for each gene type, and in each environment, for every gene of the best individual at the end of each replicate. In each case, a box plot showing quartiles and fliers is overlaid on a violin plot representing the whole distribution, and the mean is represented by a smaller tick. The data is computed only for genes which present the correct activation state in both environments, which represents 97,7% of *AB* genes, 92,7% of *B* genes and 53,2 % of *A* genes.

We averaged this data over every gene that presents the correct activation state in each environment, in the best individual of every replicate, and very different patterns once more appear, depending on whether the targeted behavior for the gene is activation or inhibition, as depicted in Figure 5.8. For *AB* genes in both environments, as well as for *A* genes in environment A and *B* genes in environment B, the experimentally obtained minimum subnetwork size is 1, which is consistent with the expression profile of an isolated gene, shown

in Figure 5.5: With a basal supercoiling value of $\sigma_{basal} = -0.06$, an isolated gene already experiences a high expression level in both environments, even without interactions.

When the evolutionary target of the gene is inhibition, that is for *A* genes in environment B and for *B* genes in environment A, the picture is however quite different. In this case, a significantly larger subnetwork is needed in order to obtain inhibition of the central gene: The median subnetwork size is 9 (4 genes on each side) for *A* genes. For *B* genes, the median size is smaller than for *A* genes, but higher than when the target is activation: Genes always need at least a subnetwork of size 3 (1 gene on each side), and several outliers need a subnetwork of more than 20 genes.

The gene regulatory networks evolved through the transcription-supercoiling coupling therefore exhibit a structure that cannot always be summarized by the pairwise interactions between neighboring genes, but that can on the contrary require the participation of a significantly larger number of genes in order to make genes display the same activation state as in the full genome.

5.2.5 A Whole-Genome Gene Regulatory Network

The effect of the transcription of every gene on the local supercoiling at every other gene (which decreases linearly with distance) provides a natural graph to represent the interactions between the genes in the genome of an individual. However, as the effective impact of a gene on the expression of other genes depends on the transcription level of that gene, this theoretical graph provides an inaccurate picture of the gene interactions that actually take place, as every gene ends up expressed at a different level. In order to characterize more finely the gene regulatory networks that evolve in our experiments, we therefore constructed a different graph, which we call the effective interaction graph, using transcriptional gene knockouts.

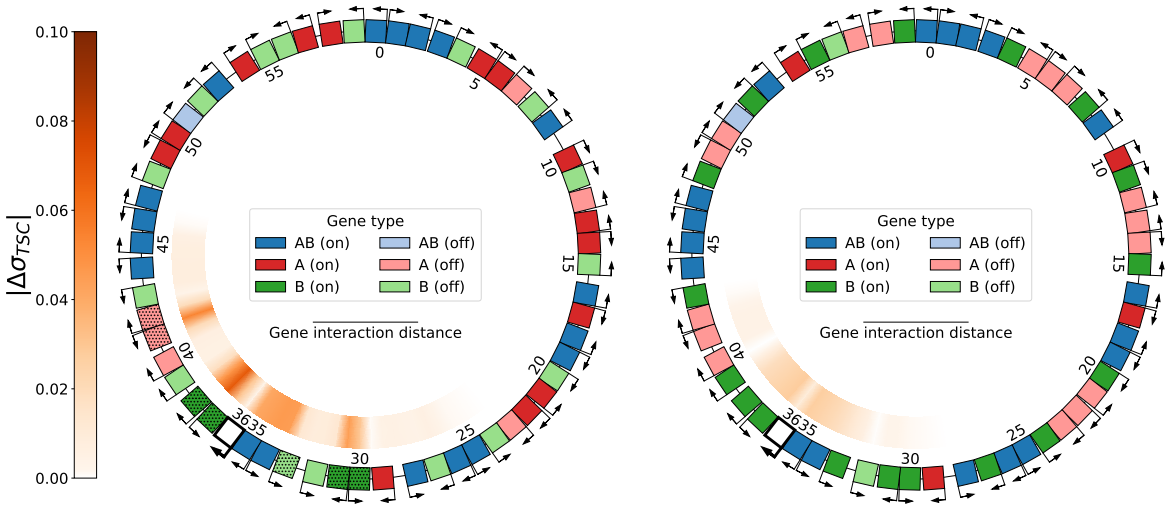


Figure 5.9: Knockout of gene 36 (of type *AB*, in bold, colored white) of the best individual at the end of replicate 21, evaluated in environments A (left) and B (right). Hatched genes represent genes whose activation state was switched by the knockout compared to their state in the original genome. The inner ring represents the absolute difference in the level of local supercoiling $|\Delta\sigma_{TSC}|$ between the knockout genome and the original genome (in Figure 5.3).

Transcriptional Gene Knockouts A transcriptional gene knockout (or simply a gene knockout in this chapter) completely suppresses the transcription of a gene, adapting the principle of gene knockouts to the transcription rather than the translation step of gene expression. In order to knock out a gene in an individual in our model, we simply set the transcription rate of that gene to zero during every step of the computation of the gene expression levels of that individual. This virtually removes the knocked-out gene from the genome, while keeping the intergenic distance between its upstream and downstream neighbors unchanged, and mimics a loss of function its the promoter. The result of such a knockout on the genome of an evolved individual is shown in Figure 5.9. The knocked-out gene is gene 36, which is of type *AB* and originally activated in both environments (see Figure 5.3 for the original genome). We can see that, in environment A, knocking out this gene results in a switch of the activation state for 7 genes (hatched in the left-hand side of Figure 5.9), that are not all contiguously located, and in local supercoiling changes that propagate to the bottom left third of the genome, to a distance that is larger than the gene interaction distance. In environment B, knocking out this gene results in milder supercoiling changes that do not lead to the switch of any gene. In this example, knocking out even a single gene can therefore substantially affect gene expression levels, significantly switching the activation state of other genes on the genome, even when they are out of reach of direct interaction with the knocked-out gene.

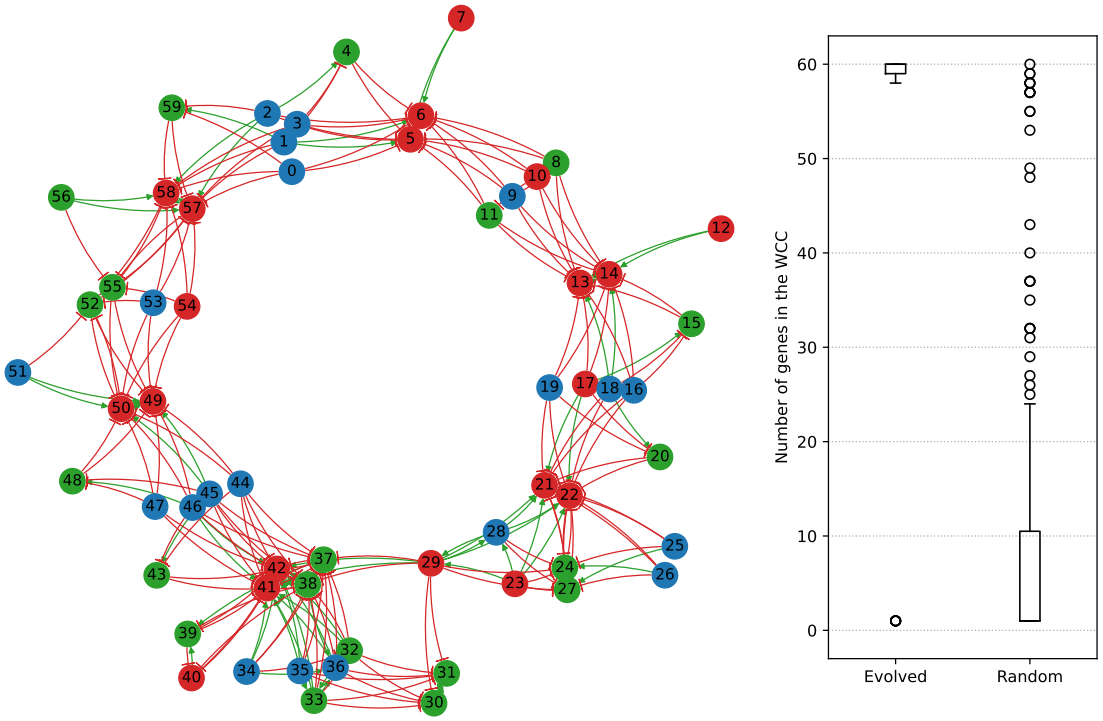


Figure 5.10: Left: effective interaction graph of the best individual at the last generation of replicate 21, obtained by knocking out each gene and measuring the resulting gene switches in each environment. Activation edges are drawn in green, and inhibition edges in red. The numbering of the genes is the same as in Figures 5.3, 5.8 and 5.9. Right: distribution of weakly connected component (WCC) sizes in the effective interaction graphs of the evolved individuals (left) and the random individuals (right).

Constructing Effective Interaction Graphs In order to construct the effective interaction graph introduced above, we simply add an edge from a gene to every other gene whose activation state is switched by knocking out that gene, in one environment or the other. If the knockout switches off a gene that was activated in the complete genome, we mark the edge as an activation edge, meaning that the knocked-out gene was necessary to activate the switched gene. If the knockout switches on a gene that was inhibited in the complete genome, we conversely mark the edge as an inhibition edge. If knocking out a gene switches the same gene in the two environments, we only add the edge once (we do not build a multigraph). The effective interaction graph of our example individual is presented on the left-hand side of Figure 5.10. In the case of this individual, there is a single weakly connected component (WCC), meaning that all genes interact as part of a single whole-genome regulatory network; this is the case in the best individual of 26 out of the 30 replicates.

Structure of the Effective Interaction Graphs We computed the effective interaction graph of the best individual in each replicate, and compared these graphs with the effective interaction graphs of 30 random individuals drawn using the same genome parameters (in Table 5.1). The results are presented on the right-hand side of Figure 5.10. The effective interaction graphs of evolved individuals are clearly different from the interaction graphs of random individuals. We can see that the evolved genomes have WCC sizes of 58 to 60 genes, comprising nearly every to every gene on the genome, along with very few single-gene WCCs (left). On the other hand, WCC sizes in the random genomes span the whole range from single-gene to whole-genome WCCs, with most of the connected components counting less than 10 genes (right).

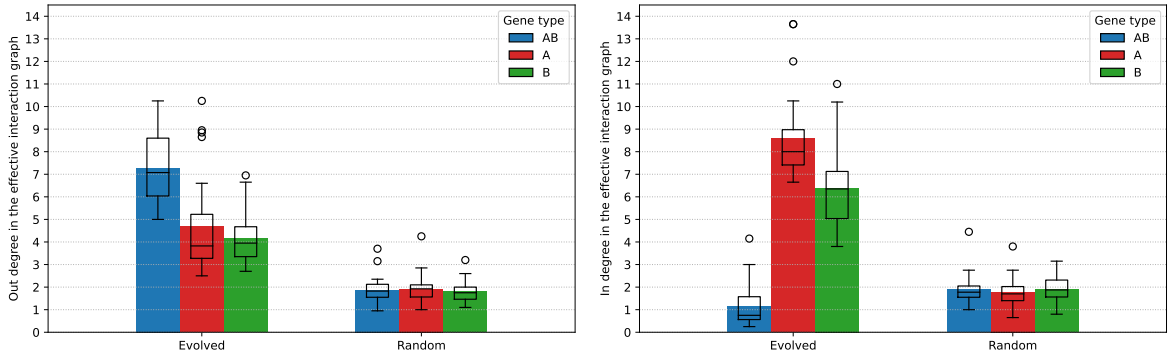


Figure 5.11: Left: average out-degree (number of genes switched by knocking out a given gene) of the nodes in the effective interaction graph, separated by gene type, for evolved and random individuals. Right: average in-degree (number of genes whose knockout switches a given gene) of the nodes in the effective interaction graph, separated by gene type, for evolved and random individuals.

The evolved genomes are indeed much more connected than the random genomes, as we can see in Figure 5.11, which presents the out- and in-degree of genes (averaged by gene type) in the effective interaction graphs of the genomes. The left-hand side of Figure 5.11 shows the average out-degree of each gene type, or the number of genes that are switched by knocking out a gene of that type. While knocking out a gene in a random genome switches the state of just under 2 other genes on average, the figure is much higher in the evolved genomes. Knocking out *A* and *B* genes switches 4 other genes on average, and knocking out *AB* genes up to 7 other genes; *AB* genes therefore play a quantitatively more important regulatory role than *A* genes or *B* genes, which can be explained by the fact that *AB* genes are activated in both environments, while most *A* and *B* genes are inhibited in one environment or the other.

When looking at the in-degree of the genes, or the number of genes whose knockout will make a given gene switch activation states, we can see that the evolved genomes are again much more connected than the random genomes, and that the in-degree depends on the type of the gene. Indeed, *AB* genes are only switched by one other gene on average, meaning that their activation state is robust to perturbations in the regulatory network. The robustness of *AB* gene state is expected, as these genes must have the same activation state in both environments. On the contrary, *A* genes and *B* genes have an in-degree that is much

higher, meaning that their activation state relies on the regulatory action of a large number of other genes, making them more sensitive to the variations between the two environments.

The evolution of the the relative positions of genes on the genome, by leveraging the feedback loop between the transcription of neighboring genes that is mediated by DNA supercoiling, therefore results in our model in the emergence of gene regulatory networks that connect the whole genome into a single entity, rather than a juxtaposition of independent subnetworks. The network structure that evolves furthermore allows genes to dampen, or amplify, the result of the environmental shift in supercoiling on their activation states, as required by their evolutionary targets.

5.3 Discussion and Perspectives

DNA supercoiling, through its effect on promoter activation (Forquet et al., 2021), is an important actor of the regulatory response of bacteria to changing environmental conditions (Martis B. et al., 2019). But supercoiling itself is in return impacted by transcription, as presented in the twin-domain model of Liu and Wang (1987). Indeed, transcription has been shown to play a major role in shaping the bacterial DNA supercoiling landscape (Visser et al., 2022). Taken together, these observations raise the question of the extent to which the position itself of genes on the genome can regulate their activity, via the coupling of the transcription levels of neighboring genes through local changes in DNA supercoiling.

In order to assess the theoretical possibility of the evolution of such a gene regulatory network, and to determine the potential consequences of the evolution of such a network on the organization of the genome, we developed in this work an evolutionary model of the transcription-supercoiling coupling (expanding upon a proof-of-concept presented in Grohens et al. (2021)), in which populations of individuals must evolve differentiated gene expression levels in response to different environmental conditions, with the transcription-supercoiling coupling as the only regulatory mechanism and inversions as the only mutational operator. As a the dynamic supercoiling level between actively transcribed genes would be very difficult to model quantitatively, our model voluntarily stays very simple in this regard, and focuses instead on providing a qualitative overview of the range and complexity of the regulatory interactions between neighboring genes that can be mediated by the transcription-supercoiling coupling.

We showed that, in this model, gene regulation by DNA supercoiling is indeed a sufficient mechanism to evolve environment- and gene-specific patterns of activation and inhibition. In particular, we observed the emergence of genes that are more expressed in a relaxation-inducing environment (or relaxation-activated genes), even though this behavior goes against the facilitated opening of the -10 promoter element by RNA polymerase during the initiation of transcription (Forquet et al., 2021). This property has been analyzed in detail *in vivo* in the classical example of the *gyrA* promoter, and was shown to result from the unusual sequence of that promoter (Menzel and Gellert, 1987), but for many other genes, this property is less firmly established and depends on the experimental conditions, with experiments finding a proportion of relaxation-activated genes varying between 27% and 70% in *S. enterica* (Pineau et al., 2022). Our results demonstrate that this behavior can re-

sult not only from the specific sequence of the promoter (as for the *gyrA* promoter), or the length of its spacer (Forquet et al., 2022), but also from the local genomic organization (as suggested in El Houdaigui et al. (2019)), and confirm the importance of this additional mode of regulation for the first time in an evolutionary simulation.

We found that evolved genomes in the model are enriched in divergent pairs of always-active genes, as well as in convergent pairs that act as bistable toggle switches (Gardner et al., 2000; Johnstone and Galloway, 2022); the evolution of such systems substantiates the theoretical predictions made by models that explicitly describe the movement of RNA polymerases during gene transcription, such as Sevier and Hormoz (2021). Then, we showed that the local organization of the genome into convergent or divergent pairs of genes is not sufficient to explain the transcriptional response of individuals to different environments, but that larger subnetworks can be required to selectively inhibit genes in specific environments. Such regulation of gene expression through interaction with groups of neighboring genes could help explain the evolutionary persistence of synteny groups between *E. coli* and *S. enterica* (Junier and Rivoire, 2016), as well as through the evolutionary history of *B. aphidicola* (Brinza et al., 2013). Indeed, we show that local interactions can play a role in regulating the expression of neighboring genes, and genomic rearrangements might disrupt these local interactions. Finally, we used transcriptional knockouts, adapting the classical tool of gene knockouts (Baba et al., 2006) to our transcription-centric model, in order to characterize the evolved gene regulatory networks in further detail. We first showed that these regulatory networks integrate the entire genome of evolved individuals into a single connected unit, in opposition to the sparser, disconnected regulatory networks displayed by randomly generated individuals. Then, we showed that the structure of these networks leverages the transcription-supercoiling coupling to increase or decrease the sensitivity of genes to perturbations in the regulatory network, strengthening the differentiated expression patterns that are the evolutionary target for each gene type.

All in all, our simulations demonstrate that the transcription-supercoiling coupling provides a regulatory mechanism that is precise enough for the evolution of complex regulation patterns that only depend on the arrangement of genes on the genome.

Several work directions still remain open to investigation. From an evolutionary perspective, the experimental framework in which we tested our model is at present very simple, and could be extended. The desired gene expression levels in our model are binary, targeting maximal or minimal transcription, but could be replaced by an arbitrary level between these values for each gene, in order to see whether the local organization into pairs as well as the whole-genome regulatory network that we described are preserved under these less constrained conditions. Similarly, we could refine the environmental challenge faced by individuals by evaluating them in each environment in succession, rather than separately, or by continuously changing the environment over evolutionary time. From a theoretical perspective, a range of mechanistic biophysical models of the transcription-supercoiling coupling have been put forward, with different hypotheses underpinning the coupling: Brackley et al. (2016) shows a phase transition in the transcription regime as the number of transcribing RNA polymerases increases; Sevier and Hormoz (2021) shows that bursty transcription can emerge from the transcription-supercoiling coupling; and Meyer and Beslon (2014) and El Houdaigui et al. (2019) try to predict gene expression levels quantitatively from the

local DNA supercoiling level. An important vindication of these theoretical approaches to the interplay between supercoiling and transcription would therefore be to verify the extent to which these models, including ours, conform to one another as the level of abstraction changes. Moreover, integrating a model of gene regulation by DNA supercoiling into a more comprehensive evolutionary model of the genome that allows for classical gene regulation via transcription factors, such as the model presented in Crombach and Hogeweg (2008), would help shed light on the coevolution between the different modes of gene regulation that are available to bacterial genomes. Finally, from an experimental perspective, a better understanding of the regulatory interactions caused by the transcription-supercoiling coupling could help design more reliable synthetic genetic constructs, as explored in Johnstone and Galloway (2022).

5.4 Conclusion

To the best of our knowledge, our work is the first to model the regulatory role of supercoiling on transcription at a many-gene scale, using evolutionary simulations. It demonstrates the importance of the direct interactions between genes that are mediated by local changes in DNA supercoiling on their transcription rates, as well as the precision and versatility of the regulatory activity stemming from these interactions. For experimentalists, it provides an underlying theory that could help explain the heterogeneous transcriptomic response (with both up- and down-regulation of multiple genes) observed in bacteria confronted to supercoiling variations, due among others to virulence-inducing environments (Dorman, 2019) or to gyrase-inhibiting antibiotics (de la Campa et al., 2017). For evolutionists, it provides a plausible evolutionary rationale for the observed conservation of local gene order between closely related bacteria (Junier and Rivoire, 2016) and along evolutionary histories (Brinza et al., 2013). Finally, for synthetic biologists, it provides a theory to help predict in finer detail the gene transcription levels that can be expected from a given gene syntax (Johnstone and Galloway, 2022), which could help design more robust genetic circuits.

Chapter 6

Evaluating the Robustness of the *EvoTSC* Model

In this chapter, I explore the robustness of the characteristics of evolving populations to variations in the parameters of the *EvoTSC* model. To this aim, I present several sets of additional evolutionary simulations. In these simulations, I first measure whether populations are able to evolve differentiated gene expression patterns as a response to different environments in the model variants. I then compare the speed of evolution of these populations with the main runs presented in Chapter 5. I first explore the sensitivity of the model to the genome-level parameters (see Table 6.1 below): the maximum interaction distance for the transcription-supercoiling coupling, the mean intergenic distance, and the strength of the environment-caused shift in background supercoiling. I then investigate simulating a higher number of genes on the genome. Finally, I discuss the evolutionary effect of allowing intergenic distances to mutate, by introducing indels in intergenic sections as a new mutational operator in the model. The data from this experiment is available online on the Zenodo platform.

Parameter	**Symbol**	**Value**	**# of Replicates**
Interaction distance	d_{max}	**5 kb**	**30**
(Section 6.1)		25 kb	15
		10 bp	15
Mean intergenic size	d_{mean}	**125 bp**	**30**
(Section 6.2)		1,000 bp	15
		10,000 bp	15
		0.0001	15
Environment supercoiling shift	$\delta\sigma_{A/B}$	0.001	15
(Section 6.3)		**0.01**	**30**

Table 6.1: Table of the parameters and associated values explored in additional experiments (separated by horizontal lines). For each experiment, the row in bold font corresponds to the parameter values used in the main run described in Chapter 5, and is shown for reference.

6.1 Interaction Distance

The size of the topological domains of bacterial genomes, inside which DNA supercoils can freely propagate, has historically been estimated to be on the order of a few thousand base pairs (El Hanafi and Bossi, 2000; Postow et al., 2004; Kouzine et al., 2013). Recent work has however suggested that the size of transcription-generated twin domains could be much larger than this, and reach up to 25 kb on either side of the transcribed gene (Visser et al., 2022). As the size of the topological domains sets a limit to the number of genes that can interact through the transcription-supercoiling coupling, it is likely to play an important role in the structure of the supercoiling-mediated gene regulatory networks. By increasing the number of genes a given gene is coupled with, a larger interaction distance could make genomic inversions more deleterious by making them disrupt a larger part of the regulatory networks at their boundaries, but it could also allow more robust regulatory networks to evolve through a higher connectivity. In this section, I present simulations run with an interaction distance of 25 kb, a five-fold increase from the value used in the main runs, in order to test these hypotheses.

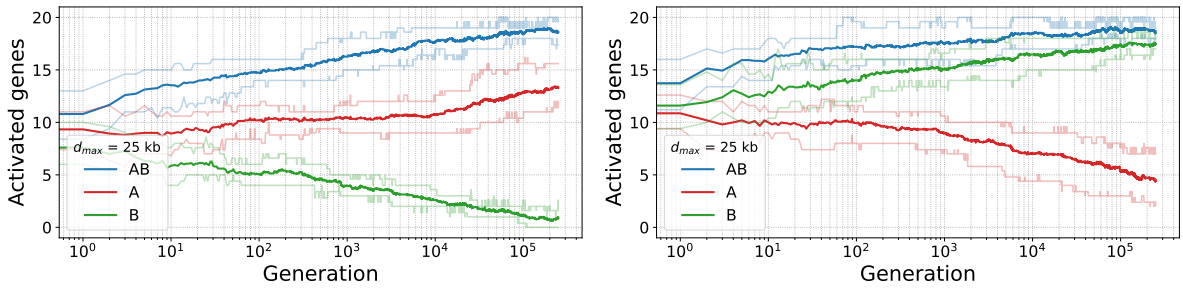


Figure 6.1: Evolution of the number of activated genes in environment A (left) and environment B (right), with an interaction distance of 25 kb. Lighter lines represent the first and last decile of the data.

Figure 6.1 shows the evolution of the number of activated genes of each type in the simulations with the interaction distance of 25 kb, over 250,000 generations (to be compared to Figure 5.4 in Chapter 5). As in the main run, the number of activated genes of each type evolves towards their respective target. Differentiated activation patterns can therefore still evolve even when the topological domains are 5 times larger.

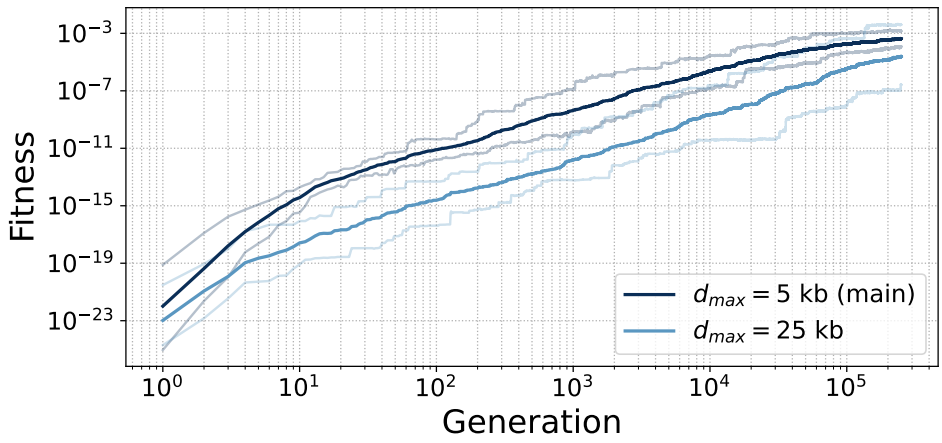


Figure 6.2: Average fitness during evolution with an interaction distance of 25 kb (light blue), and in the main run (dark blue). Lighter lines represent the first and last decile of the data.

Figure 6.2 shows the evolution of the average fitness of the best individual in each replicate of the simulation with a larger interaction distance, compared with the evolution of fitness in the main run, over 250,000 generations. While fitness is systematically lower throughout evolution with the larger interaction distance, it nonetheless follows a qualitatively similar curve, and keeps on increasing until the end of the runs, suggesting that it could eventually reach the same value as in the main run (presented in full in Figure 5.2).

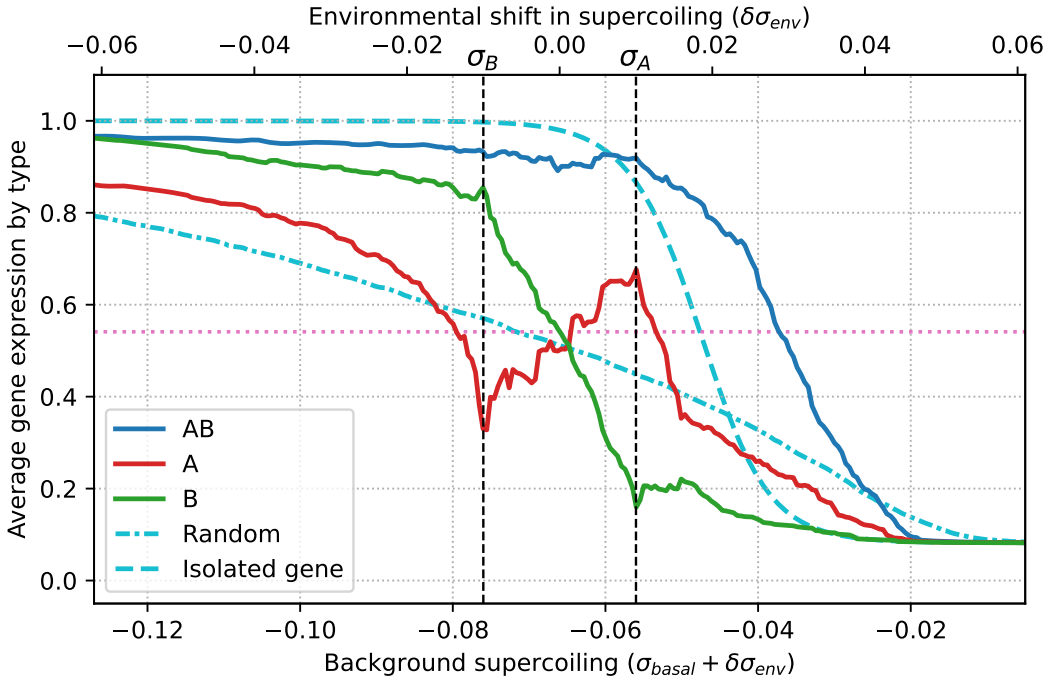


Figure 6.3: Average gene expression by type as a function of background supercoiling, with an interaction distance of 25 kb. The dash-dotted line represents the average expression of genes on a random genome with an interaction distance of 25 kb.

Figure 6.3 shows the average gene expression by type of evolved individuals, as a function of the background supercoiling level. As in the main run, *A* genes display a relaxation-activated phenotype. *AB* and *B* genes are relaxation-inhibited, but nonetheless display quite different behaviors than the (dash-dotted light blue) curve for genes on a random genome with the same parameters, which present a much flatter response curve to background supercoiling than with the default interaction distance of 5 kb (shown in Figure 5.5).

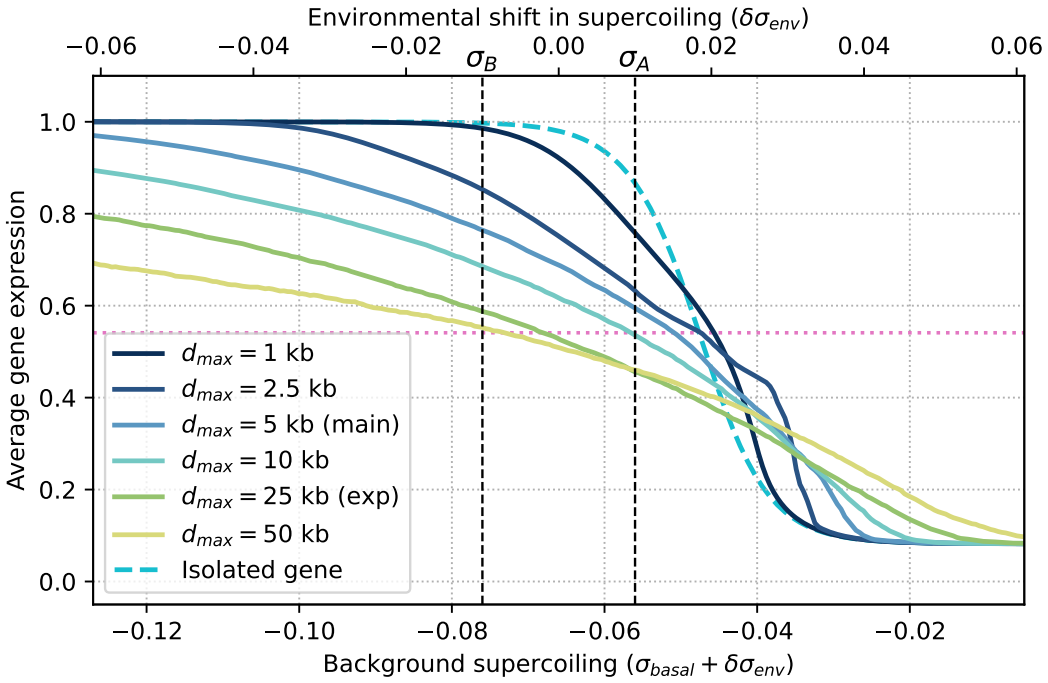


Figure 6.4: Average gene expression as a function of background supercoiling in random genomes with increasing interaction distances (full lines), and theoretical expression level of an isolated gene (dashed line).

Comparing the effect of background supercoiling variation on the transcriptional activity of genes on random genomes with an interaction distance of 5 kb (Figure 5.5) or 25 kb (Figure 6.3) seems to indicate that a larger interaction distance buffers the effect of the background supercoiling on gene expression in the model. In order to test this hypothesis, I measured the average gene expression as a function of background supercoiling in random genomes with maximum interaction distances ranging from 1 kb to 50 kb, with an average intergenic distance remaining constant at 125 bp. For each interaction distance, I generated 100 genomes, and made each genome undergo 100 replication events to shuffle the genes on the genome via genomic inversions. Figure 6.4 shows the result of this experiment. As the interaction distance increases, the response curve of genes to a changing background supercoiling level indeed becomes flatter and flatter, which can be interpreted as a buffering of the effect of the environmental perturbation on gene expression by the supercoiling-mediated interaction of a higher and higher number of genes. Conversely, as the interaction distance gets closer to zero, the curve is closer and closer to that of an isolated, non-interacting gene.

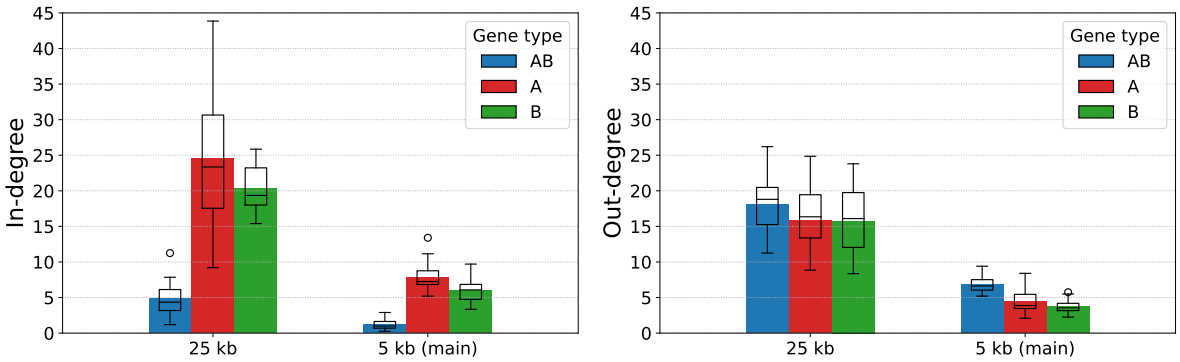


Figure 6.5: Average out-degree (left) and in-degree (right) of the genes in the effective interaction graph, separated by gene type, for individuals with an interaction distance of 5 kb (main run) or 25 kb.

Figure 6.5 finally presents the average in- and out-degree of the genes in the effective interaction graph obtained with gene knockouts, compared to the same data from the main run. As could be expected, the gene regulatory networks are much more connected with the larger interaction distance, but the behavior for each gene type remains qualitatively the same as in the main run.

Overall, the evolution of supercoiling-mediated gene regulatory networks that are able to show environment-specific activation patterns, and in particular relaxation-activated genes, therefore seems robust in our model to a larger interaction distance. Even when the networks are more densely connected, buffering gene expression, the reorganization of the genome via genomic inversion allows for specific behaviors for each gene type.

6.2 Mean Intergenic Size

In the main run, I set the initial intergenic distance between every gene to 125 bp, or a gene density of 88%, representing the average *E. coli* intergenic distance (Postow et al., 2004). Bacteria actually present a wider range of gene densities, from 51% in *Sodalis glossinidius*, a bacteria that is undergoing massive pseudogenization after recently adopting an endosymbiotic lifestyle (Toh et al., 2006), up to 95% in *Thermotoga maritima*, an extremophile bacteria living in heated marine sediment (Nelson et al., 1999). Like the interaction distance, the mean intergenic distance plays a role in the connectivity of the gene regulatory networks that stem from the transcription-supercoiling coupling: when the mean intergenic distance is small, more genes can on average be affected by the transcription of any given gene. Conversely, when the mean intergenic distance is large compared to the interaction distance, the emergence of regulatory subnetworks isolated from each other by large intergenic regions becomes possible. In this section, I test the robustness of the results with regard to the range of gene densities found across living organisms. I present simulations with mean intergenic distances increasing logarithmically from 10 bp (or a gene density of 99%), comparable to the median intergenic distance of 3 bp found in *Pelagibacter ubique*, a free-living marine bacte-

ria (Giovannoni et al., 2005), up to 10 kb (or a gene density of 10%), akin to the gene-scarce eukaryotic genomes (Dávila López et al., 2010).

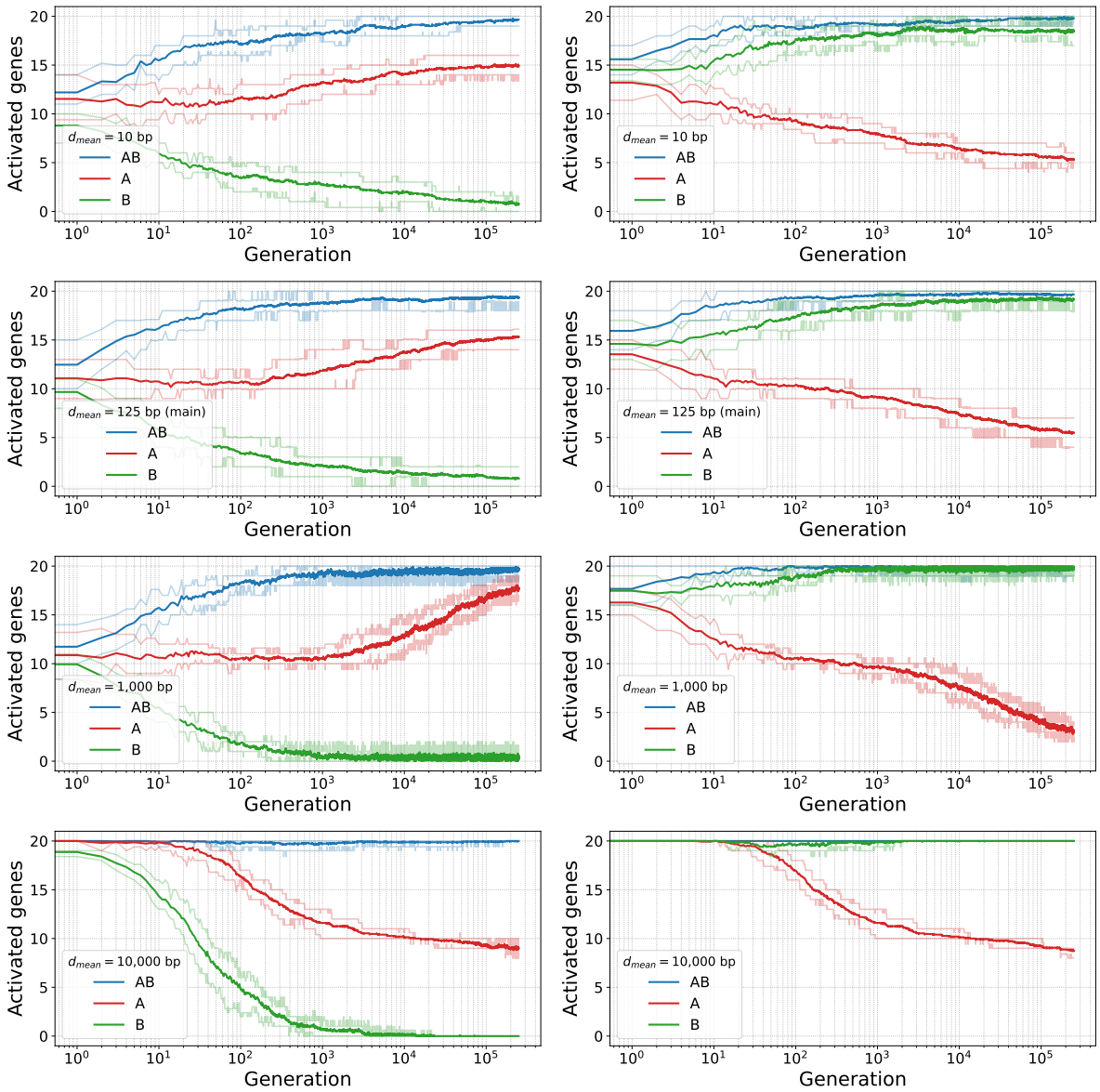


Figure 6.6: Average number of activated genes per gene type in environment A (left) and B (right) during evolution, for average intergenic sizes from top to bottom of 10 bp, 125 bp (main run), 1 kb, and 10 kb. Lighter lines represent the first and last decile of the data.

Figure 6.6 shows the evolution of the number of activated genes for each gene type in each environment, with mean intergenic distances increasing from top to bottom, for 250,000 generations. For intergenic distances of 10 bp and 1 kb, the number of activated genes converges towards the target in each environment, as in the main run (which has a mean intergenic distance of 125 bp). For a mean intergenic distance of 10 kb, the behavior is however qualitatively different. While *AB* genes and *B* genes evolve towards the correct activation

state in each environment, the proportion of activated *A* genes stays close to 50% in each environment, showing that *A* genes do not evolve a well-differentiated expression pattern depending on the environment.

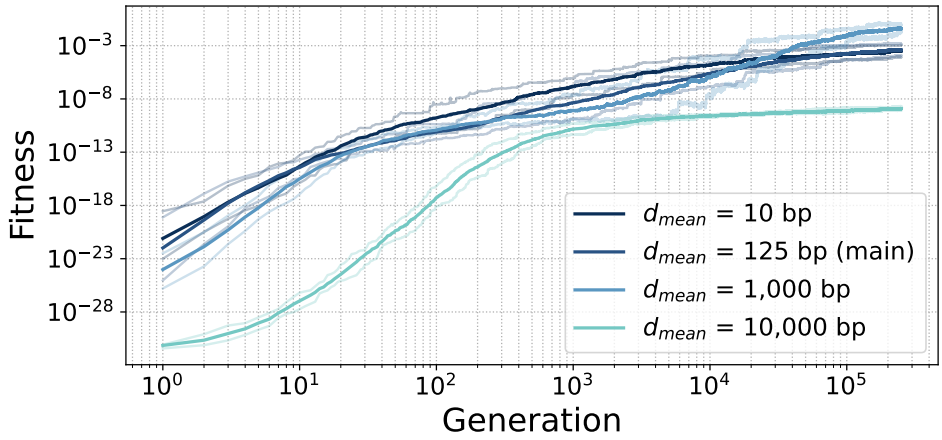


Figure 6.7: Average fitness during evolution for average intergenic distances of 10 bp, 125 bp (main run), 1 kb, and 10 kb. Lighter lines represent the first and last decile of the data.

Figure 6.7 shows the evolution of the average fitness of the best individual in each replicate of the simulations, for each value of the mean intergenic distance, including the main run for comparison. It confirms the results seen in the previous figure: for intergenic distances from 10 bp to 1 kb, populations evolve successfully towards differentiated gene activation patterns. For an intergenic size of 10 kb (in light blue), however, fitness increases much more slowly, and seems to converge towards a much smaller value.

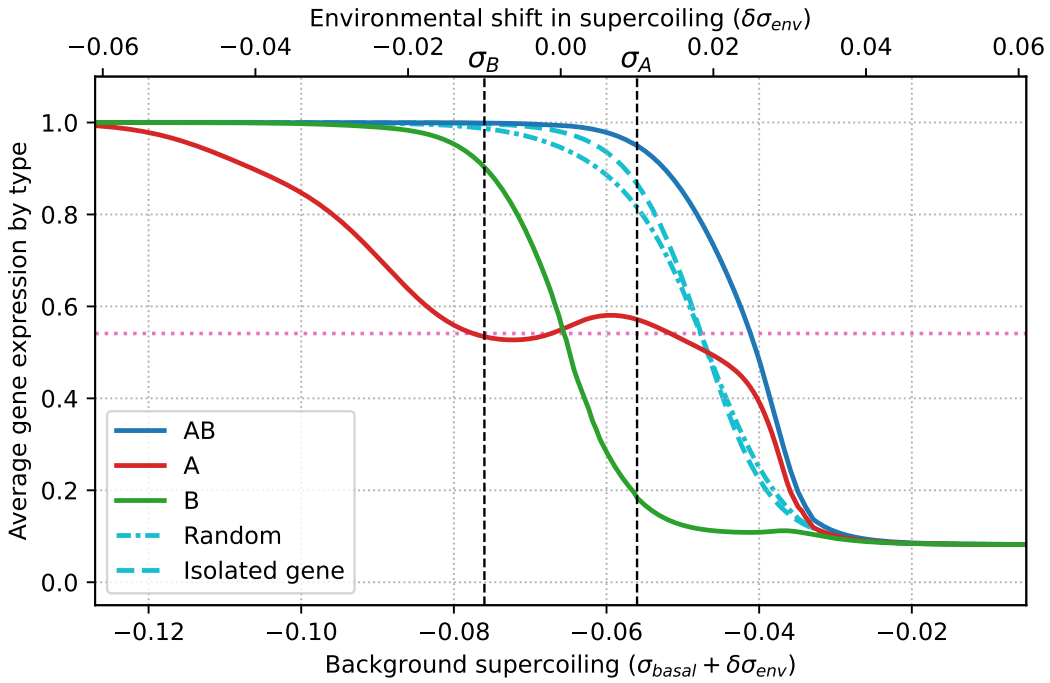


Figure 6.8: Average gene expression as a function of background supercoiling, with an intergenic distance of 10 kb.

Figure 6.8 shows the average gene activity as a function of background supercoiling, for the best individual in each of the replicates with a mean intergenic distance of 10 kb. In this case, we do not observe the clear relaxation-activated phenotype for *A* genes that was present in the main simulation. Instead, the average expression level of *A* genes is indeed slightly lower in environment B than in environment A, but remains close to half expression for background supercoiling values between -0.08 and -0.05; on the contrary, both *B* and *AB* genes display expression curves that closely match their respective targets. Note also that, with an intergenic distance of 10 kb, the average activity of genes on random genomes is almost identical to that of an isolated gene (as could be expected).

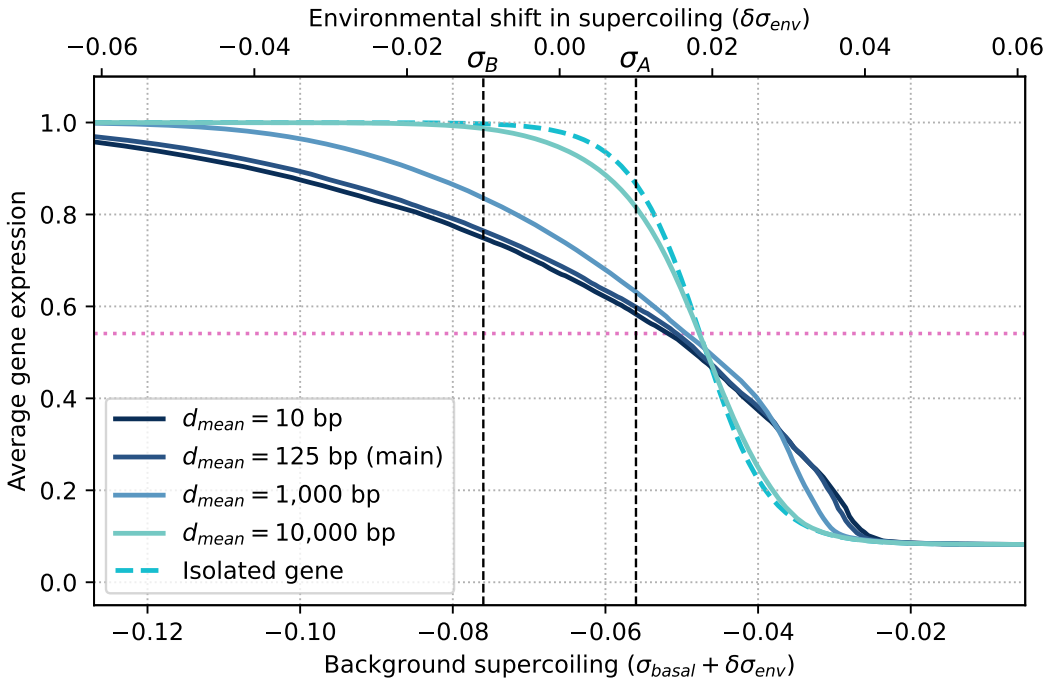


Figure 6.9: Average gene expression as a function of background supercoiling in random genomes with increasing mean intergenic distances (full lines), and theoretical expression level of an isolated gene (dashed line).

Figure 6.9 shows the expression level as a function of background supercoiling for genes on random genomes with increasing mean intergenic distances. For distances from 10 bp to 1 kb, the curves are qualitatively similar to one another, and quite different to the expression of an isolated, non-interacting genes. In particular, gene expression levels are sensibly lower than the maximum in both environment A and environment B. On the other hand, for a mean intergenic distance of 10 kb (in light blue), genes behave very closely to an isolated gene, and are all almost fully activated in both environments. This behavior is also visible in Figure 6.6 (bottom row) at the beginning of evolution, and explains the initially much lower fitness at an intergenic distance of 10 kb compared to the other values in Figure 6.7.

Several hypotheses could explain the incomplete fulfillment of the evolutionary target in simulations with a mean intergenic distance of 10kb. First, individuals begin evolution at a much lower fitness with a mean intergenic distance of 10 kb. This initial impediment however does not explain why the number of activated A genes remains similar in both environments throughout evolution, contrary to the other runs. Another possible explanation could come from the mutational operator – genomic inversions – used during evolution. As the endpoints of the genomic inversions are chosen by picking two bases uniformly at random in the intergenic regions, the number of fully neutral inversions increases with the size of these regions. Indeed, if the two endpoint of an inversion fall in regions that are not within interaction distance of any gene, the inverted region does not interact with the rest of the genome, and the inversion is therefore completely neutral. The increased proportion of neutral mutations when the intergenic distance is too large could make the exploration

of the fitness landscape more difficult for populations, by creating hard to cross fitness valleys or plateaus between fitness peaks. There is indeed no theoretical reason why genomes with large intergenic distances could not reach comparable fitnesses to the other intergenic distances in the mode, as a genome in which most of the intergenic content is compacted into a single intergenic region would have a qualitatively similar behavior, except for the few bordering genes on each side, to the same genome with a short intergenic region in the same position.

Evolution of gene regulation by supercoiling is therefore overall resilient to the range of gene densities seen in bacteria, but the evolution of relaxation-activated genes breaks down at the lower gene densities that are more characteristic of eukaryotes.

6.3 Environmental Shift in Supercoiling

In the model, the environmental shifts in supercoiling σ_A and σ_B represent the effect of external stresses, such as salt shock, or pH or temperature changes, on the DNA supercoiling level, as they affect for example topoisomerase activity. In this section, I tested the robustness of the evolution of differentiated expression levels when the shift in supercoiling caused by the environment σ_A and σ_B is 10 times and 100 times smaller than in the main run, i.e. when the stress is of a lower intensity. The approach is similar to the one presented in Section 4.2.6, which however uses the proof-of-concept version of the model.

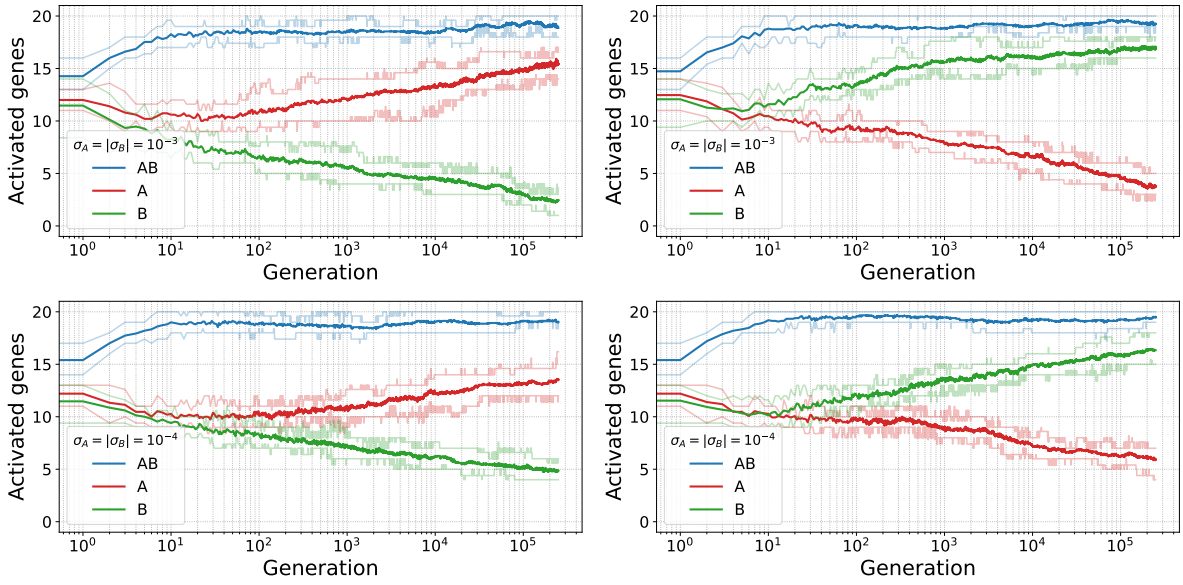


Figure 6.10: Evolution of the number of activated genes in environment A (left) and environment B (right), with environmental supercoiling shifts $\sigma_A = 10^{-3}$ and $\sigma_B = -10^{-3}$ (top) and $\sigma_A = 10^{-4}$ and $\sigma_B = -10^{-4}$ (bottom). Lighter lines represent the first and last decile of the data.

Figure 6.10 shows the evolution of the number of activated genes for each gene type in each environment, with environmental shifts in supercoiling 10 times smaller than the

main run (top) and 100 times smaller (bottom), for 250,000 generations. In both cases, differentiated gene expression patterns evolve, although evolution seems to be slower when σ_A and σ_B are 100 times smaller than in the main run. This shows that evolution of different gene expression levels as a response to externally-induced perturbations in the supercoiling level can therefore take place even when these perturbations are very minute compared to translation-generated supercoiling, and reinforces the plausibility of the hypothesis that DNA supercoiling can be used as an sensory device for the regulation of gene expression in response to environmental stress.

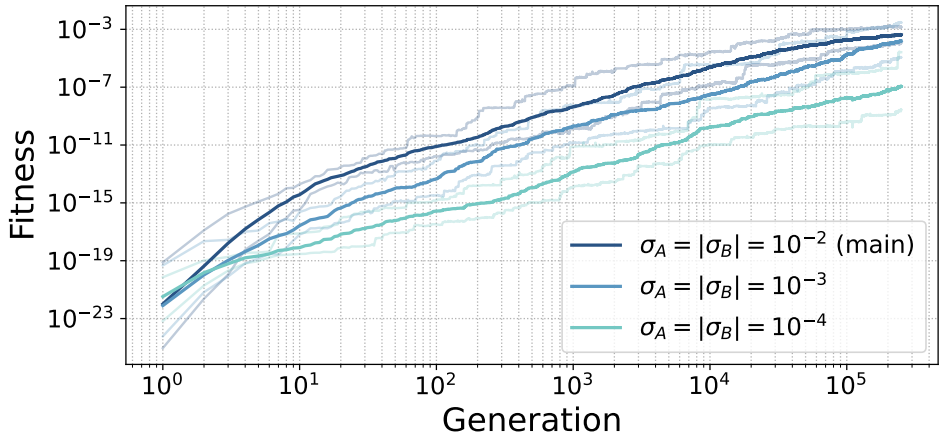


Figure 6.11: Average fitness during evolution, with environmental shifts in supercoiling logarithmically decreasing in absolute value: $\sigma_A = 10^{-2}$ and $\sigma_B = -10^{-2}$ (main run), $\sigma_A = 10^{-3}$ and $\sigma_B = -10^{-3}$ (10 times smaller than in the main run) and $\sigma_A = 10^{-4}$ and $\sigma_B = -10^{-4}$ (100 times smaller than the main run). Lighter lines represent the first and last decile of the data.

Figure 6.11 shows the evolution of the average fitness of the best individual in each replicate, for each pair of environmental supercoiling values. It confirms the results seen in the previous figure: fitness keeps increasing throughout the simulation in all cases, but more slowly when the environmental perturbation is the smallest.

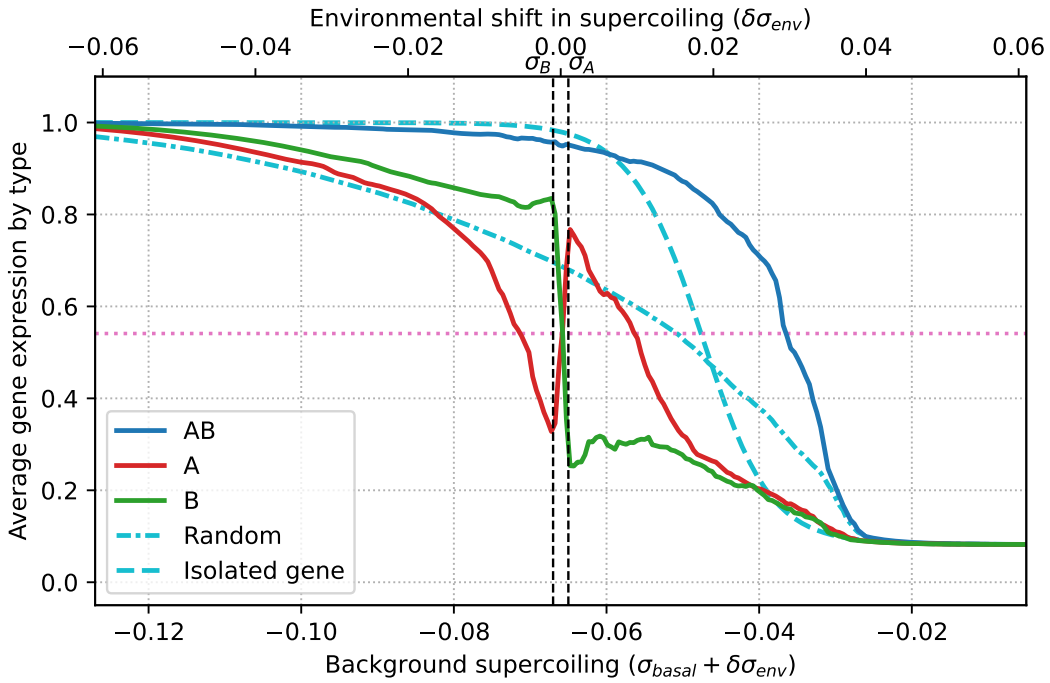


Figure 6.12: Average gene expression as a function of background supercoiling, with environmental supercoiling shifts $\sigma_A = 10^{-3}$ and $\sigma_B = -10^{-3}$.

Figure 6.12 shows the average gene expression per gene type as a function of background supercoiling, for the best individuals at the end of evolution with environmental shifts in supercoiling of $\sigma_A = 10^{-3}$ and $\sigma_B = -10^{-3}$. Even though the difference between the two environments is 10 times smaller than in the main run, *A* genes are still able to evolve a relaxation-activated phenotype, and *B* genes are still able to quickly transition from high activation to high inhibition in a much shorter range of supercoiling values.

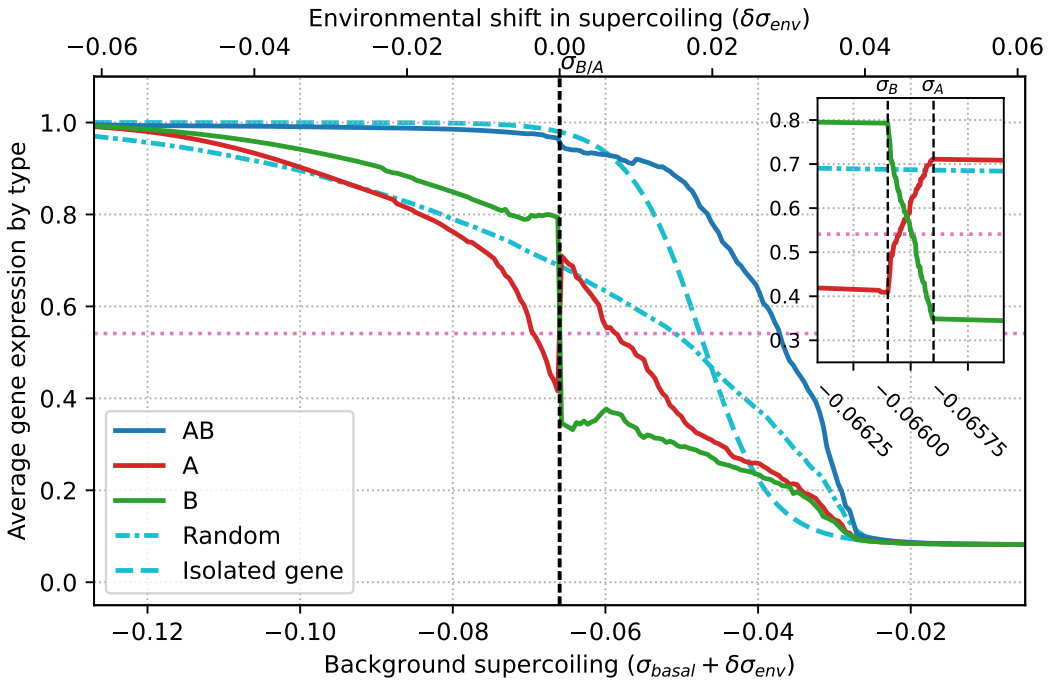


Figure 6.13: Average gene expression as a function of background supercoiling, with environmental supercoiling shifts $\sigma_A = 10^{-4}$ and $\sigma_B = -10^{-4}$. The inset at the top right of the figure shows a 150x zoom on supercoiling shift values near zero.

Figure 6.13 similarly shows average gene expression per gene type as a function of background supercoiling, this time with environmental supercoiling shifts $\sigma_A = 10^{-4}$ and $\sigma_B = -10^{-4}$. Strikingly, *A* genes and *B* genes are still able to evolve different expression levels in the two environments, even though they are 100 times closer than in the original experiment (see the inset for the precise expression levels between σ_B and σ_A). Even when the environments have an effect that is around 100 times smaller than the transcription-generated supercoiling (see Figure 5.3 for an example genome and the associated transcription-generated supercoiling values), the gene regulatory networks that evolve in the simulations are still able to separate the different environments and lead gene expression levels to very different states.

These results further confirm the results of the parametric exploration of the proof-of-concept version of *EvoTSC*, at the end of Chapter 4. They show that perturbations that have little to no effect either on an isolated gene, or on a random genome (averaging over every gene), can be picked up and amplified by supercoiling-mediated gene regulatory networks, and result in clearly differentiated gene expression levels.

6.4 Number of Genes

All the simulations presented up to now were run with $n = 60$ genes on the genomes of individuals. Although genes in our model correspond to transcriptional units, and could describe

operons that contain multiple genes, this number remains much lower than the real number of genes in bacteria, which ranges from the 482 protein-coding genes found in *Mycoplasma genitalium*, the bacteria with the smallest-known genome (Glass et al., 2006), up to over 9,000 predicted genes in *Sorangium cellulosum* (Schneiker et al., 2007). At first sight, increasing the total number of genes in the model should not qualitatively change the simulation results, as with a genome size of 60, genes already only interact via the transcription-supercoiling coupling with a small proportion of the genome. In order to verify this hypothesis, I ran simulations with $n = 300$ genes, with maximum supercoiling interaction distances of $d_{max} =$ 5 kb (the value used in the main runs) and $d_{max} = 25$ kb. As the algorithmic complexity of the *EvoTSC* scales quadratically with the number of genes, the simulations were run for 100,000 generations only to keep their execution time manageable, but the simulations already show qualitative results by that time.

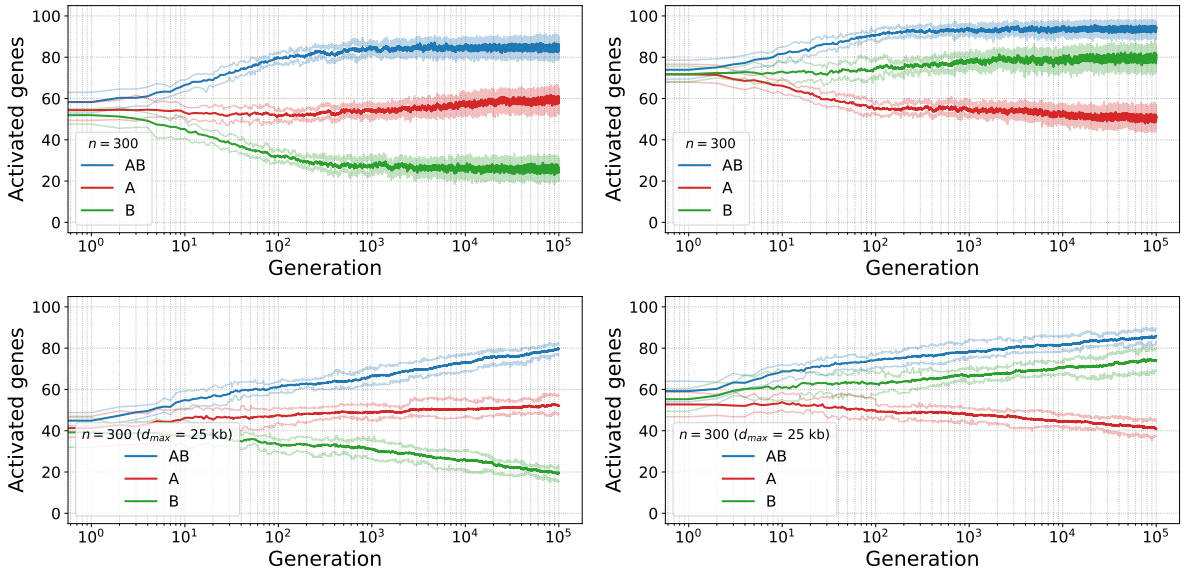


Figure 6.14: Evolution of the number of activated genes in environment A (left) and environment B (right), with a genome containing 300 genes, and an interaction distance of 5 kb (top) and 25 kb (bottom). Lighter lines represent the first and last decile of the data.

Figure 6.14 shows the evolution of the number of activated genes of each type in each environment, for populations of individuals with a 300-gene genome, for 100,000 generations. While these simulations lasted only for 100,000 simulations, the evolutionary trajectories are already different from the ones taken by the main run (in Figure 5.4). Different numbers of activated genes evolve in each case as a function of the environment supercoiling, but the patterns differ depending on the interaction distance. With the shorter interaction distance, the number of activated genes of each type seems to plateau after about 10,000 generations (top). On the contrary, evolution seems to continue by the end of the runs for the wider interaction distance (bottom).

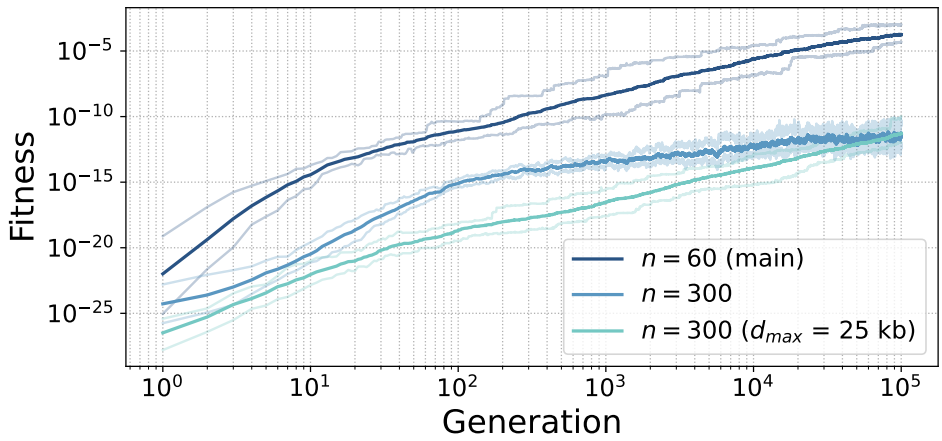


Figure 6.15: Average fitness during evolution, with a 300-gene genome, and an interaction distance of 5 kb or 25 kb. Lighter lines represent the first and last decile of the data.

Figure 6.15 shows the evolution of the average fitness of the best individual in each replicate of the simulations with 300 genes per individual, compared to the main run. As expected given the high proportion of A genes with an incorrect activation state in each environment that is visible in both sets of simulations in Figure 6.14, the fitness of the runs with 300 genes remains much lower than the fitness of the main runs throughout evolution. Moreover, the fitness trajectory in the runs with 300 genes also depends on the interaction distance, in accordance with the evolution of the number of activated genes in each environment presented above. For an interaction distance of 5 kb, fitness indeed seems to plateau at a much lower value than the main run by the end of the 100,000 generations, while it keeps on steadily increasing for an interaction distance of 25 kb.

Figure 6.16 shows the average expression of genes for each gene type as a function of the background supercoiling level, for interaction distances of 5 kb (top) and 25 kb (bottom). In both cases, and similarly to the simulation with a 10 kb intergenic distance, A genes do not seem to meaningfully evolve a relaxation-activated phenotype by the end of the simulations. With 300 genes, the response of each gene type seems to diverge less from the behavior of genes on a random genome than in the other simulations presented above. In particular, with an interaction distance of 5 kb, the average activity of A genes is above the activation threshold at σ_B (i.e., in environment B), in contrast to all the other experimental settings except when the mean intergenic distance is 10 kb (Figure 6.8). With an interaction distance of 25 kb, the average expression of A genes on the contrary goes below the activation threshold at σ_B and above the threshold at σ_A, which correlates with the higher fitness observed for this value of the intergenic distance.

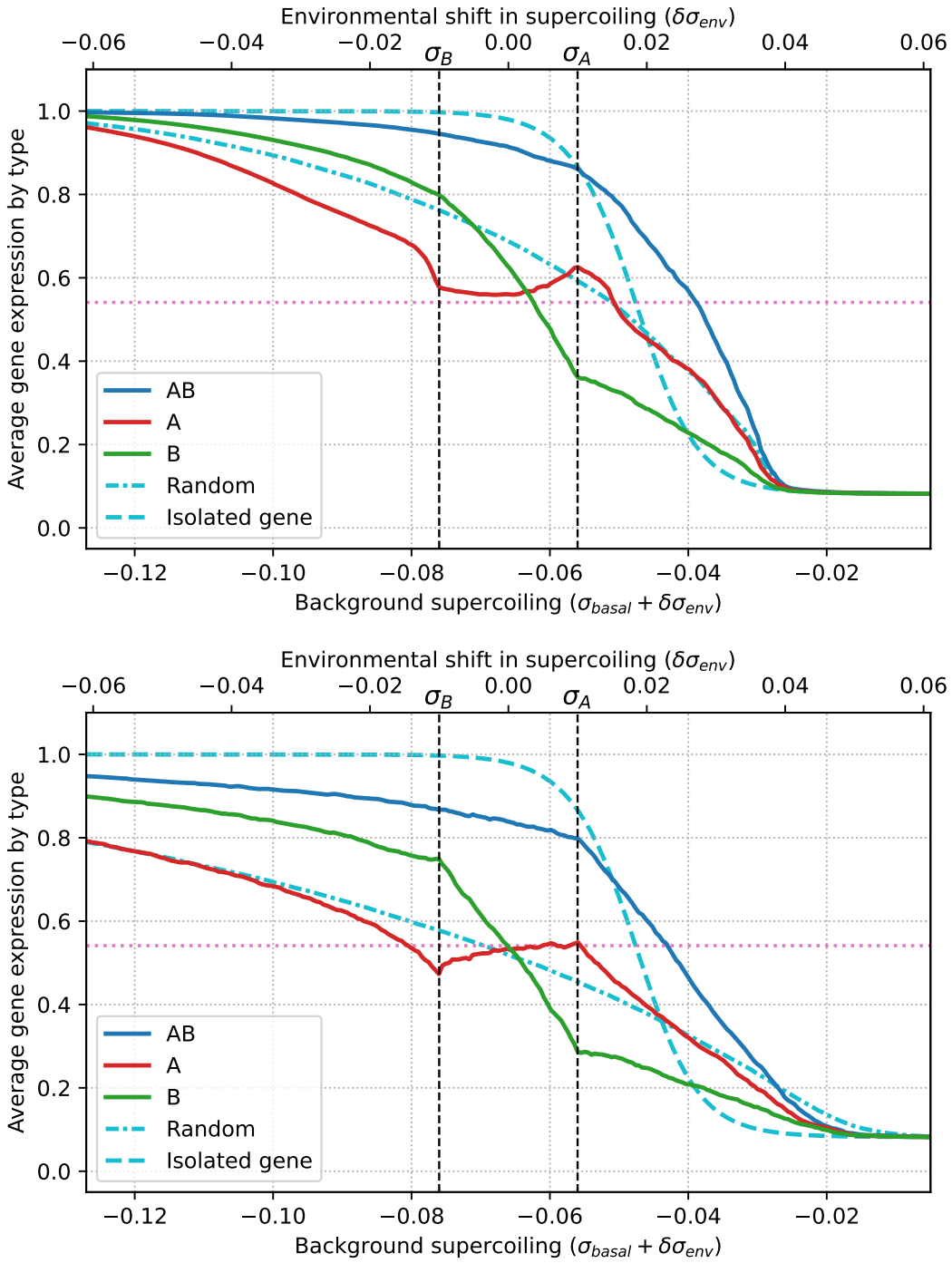


Figure 6.16: Average gene expression as a function of background supercoiling, with a 300-gene genome and an interaction distance of 5 kb (top) and 25 kb (bottom).

A possible hypothesis for the slower evolution of the runs with $n = 300$ genes lies in the much larger size of the fitness landscape to explore during evolution through genomic inversions, and in the combinatorics of these inversions. As an inversion is defined by two

points chosen on the genome, the number of different inversions grows with the square of its size; if we disregard the role of the intergenic sequences (which have an average length of 125 bp in all the runs in this section), this number grows with the square of the number of genes. The number of inversions to explore in order to find beneficial relative gene positions could therefore be much higher in larger genomes; as the number of inversions does not depend on genome size, larger genomes could take more time to explore the fitness landscape. A second hypothesis also depends on the number of genes affected by inversions, but in a slightly different way. As the end points of genomic inversions are chosen at random, the average size of an inversion is of half the genome with all parameter values, but the absolute distance between the end points of the inversion changes with the number of genes on the genome. In larger genomes, the genes located near one of the end points of an inversion are therefore further apart from the genes located near the other end point of the inversion, and could therefore be more rarely part of the same gene regulatory network than in smaller genomes. In particular, inversions affecting only a few genes could be rarer, making it more difficult to locally adjust gene regulatory networks, even though the relative orientation of neighbors and local networks play an important role (see Chapter 5). This would in particular explain why the runs with 300 genes but a larger interaction distance evolve better, as the breadth of the regulatory networks increases with the interaction distance.

A further investigation of the behavior of the model with larger number of genes could therefore seem warranted, but the overall qualitative evolution of different gene activation levels in response to environmental perturbations remains present when increasing the number of genes on the genome in the model, at least when the interaction distance is large.

6.5 Introducing Indels

In Section 6.2, we saw that the evolution of differentiated gene expression patterns depends on the mean intergenic distance: at low to intermediate values (akin to bacterial genomes), genes present environment-specific activity levels, while A genes fail to do so at the largest tested intergenic distance (akin to eukaryotic genomes). In order to understand more finely the role of intergenic distances in the evolution of the regulatory networks underpinning these expression patterns, I introduced a new mutational operator which allows these distances to evolve, through the addition or deletion of a small number of bases between genes (indels). The last set of simulations presented in this chapter tackle the exploration of the model with this additional mutational operator.

In order to perform an indel in the model, we first pick a number of base pairs to add or delete, by drawing from a normal distribution $\mathcal{N}(0, s^2)$, with $s^2 = 10$. Then, we draw uniformly at random a gene, and add or remove the corresponding number of bases from the intergenic section starting immediately after that gene, in the forward direction. If the intergenic section is too small to delete the chosen number of base pairs, we try drawing another gene at random, before giving up after a certain number of tries. For these simulations, I ran 15 replicates for 1,000,000 generations (as in the main runs in Chapter 5), in order to allow for the mean intergenic distance to converge. The initial value of the mean intergenic distance was set to 125 bp, in order to match the main set of simulation.

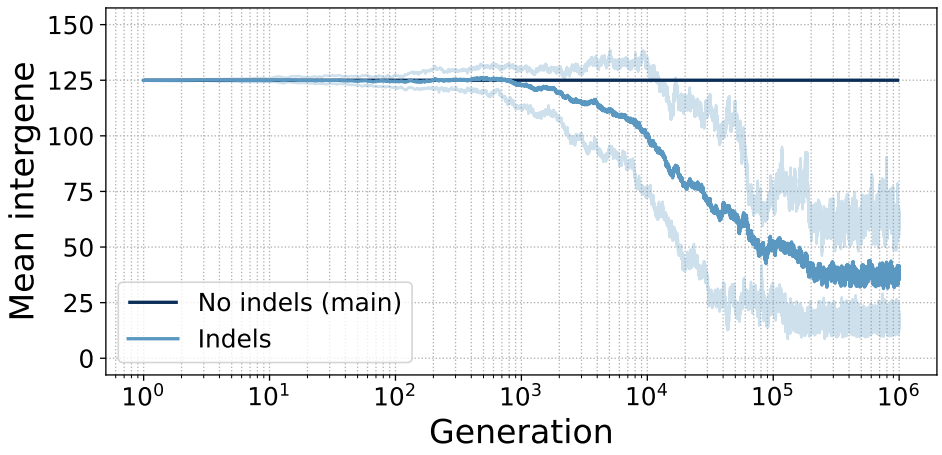


Figure 6.17: Average intergenic distance during evolution with indels, compared with the constant average intergenic distance in the main run. Lighter lines represent the first and last decile of the data.

Figure 6.17 show the evolution of the mean intergenic distance, averaged over the best individual of each replicate, at every generation, compared to the initial value of 125 bp. The average intergenic distance actually decreases during evolution, and seems to converge to a value of around 40 bp. This result is opposite to what could have been expected, as the highest fitness when varying intergenic distances is actually reached for a value of 1 kb (Figure 6.7). There therefore seems to be a selection pressure towards reducing intergenic distances through indels, even though this does not lead to the highest attainable fitness.

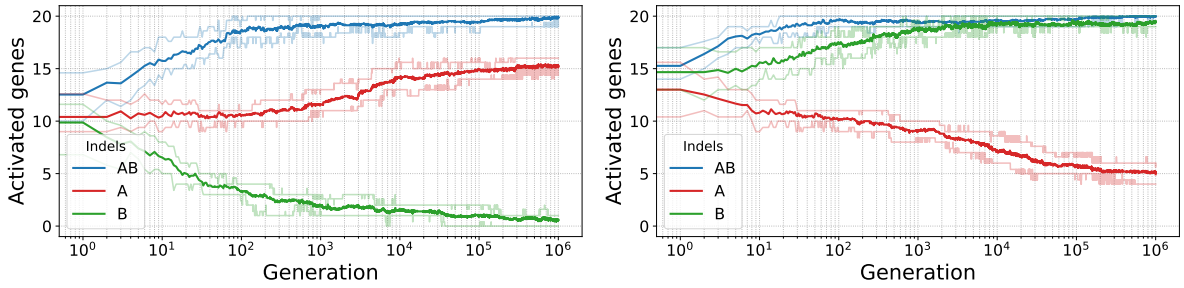


Figure 6.18: Evolution of the number of activated genes per gene type in environment A (top) and environment B (bottom), with indels. Lighter lines represent the first and last decile of the data.

Figure 6.18 shows the evolution of the number of activated genes of each type, in each environment. Similarly to the main run, differentiated expression levels evolve in the two environments, with in particular around 75% of *A* genes activated in environment A and 75% of *A* genes inhibited in environment B.

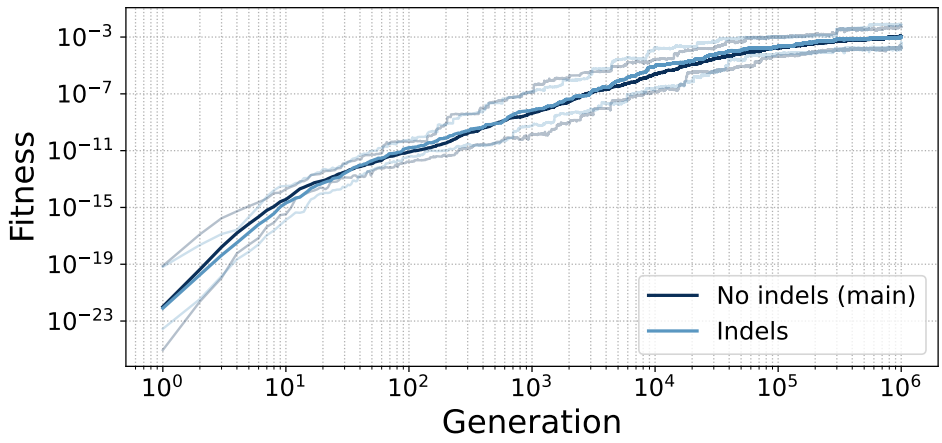


Figure 6.19: Average fitness during evolution, with indels and without indels (main run). Lighter lines represent the first and last decile of the data.

Figure 6.19 shows the average fitness during evolution of the individuals in the simulations with intergenic distance mutations, compared to the main simulations. As in the runs with an intergenic distance of 10 bp (in Figure 6.7), evolution progresses quite similarly to the main run. Consistently with the observed evolution of the mean intergenic distance, but still surprisingly as evolution towards higher intergenic distances could be possible, populations with indels attain a lower fitness after 1,000,000 generations ($1.12 \cdot 10^{-3}$) than populations with a constant mean intergenic distance of 1 kb after 250,000 generations ($3.78 \cdot 10^{-2}$).

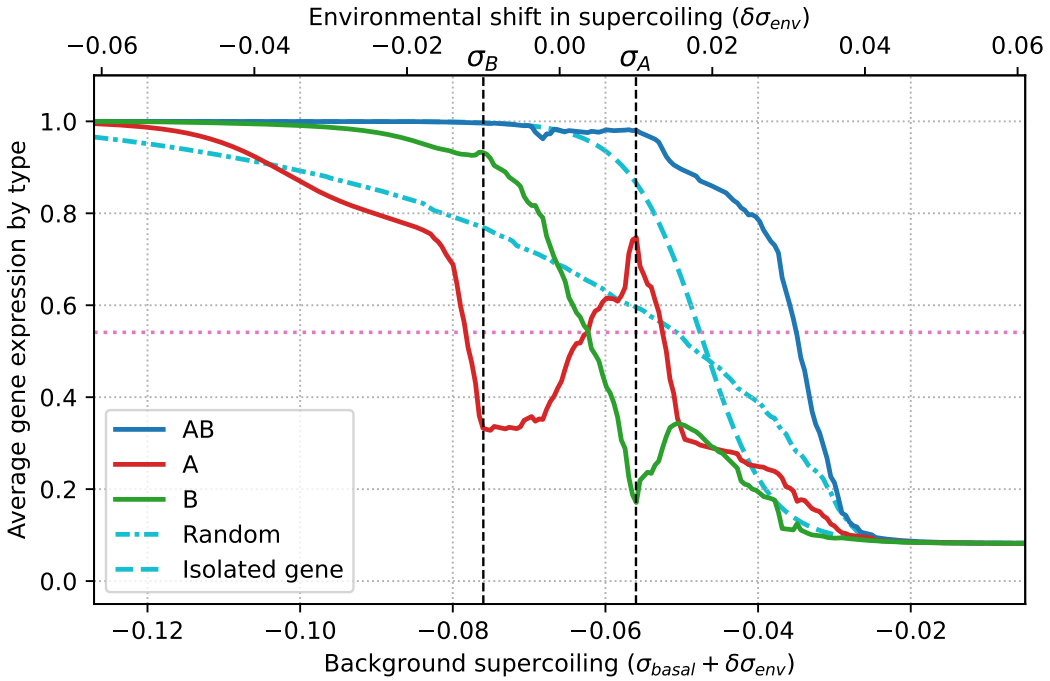


Figure 6.20: Average gene expression for each gene type type as a function of background supercoiling, with indels.

Finally, Figure 6.20 shows the average gene activity for each gene type as a function of background supercoiling, for the simulations with indels. Consistently with the results obtained in Section 6.2, we can see that *A* genes display a relaxation-activated phenotype.

These results seem to suggest that, even though evolving towards higher intergenic distances would allow populations to reach greater fitnesses, there is a stronger selection pressure that keeps intergenic distances at a low value. As in section 6.4, this selection pressure could be related to the combinatorics of genomic inversions. As the number of different inversions grows quadratically with the number of intergenic bases, reducing the mean intergenic distance diminishes the size of the evolutionary space that is available through genomic inversions. If beneficial inversions happen more often in individuals with reduced genomes than with larger genomes, that is if the evolvability of reduced genomes is higher than that of larger genomes, then a small mean intergenic distance could provide a short-term evolutionary advantage even though it leads to a lower fitness peak in the long term. Although this hypothesis warrants further investigation, these results could be interpreted as an indirect conflict between direct selection favoring a higher-dimensional fitness landscape with higher peaks, and indirect selection favoring a lower-dimensional fitness landscape with more easily reachable peaks.

6.6 Discussion

The simulations carried out with the different parameter values presented above show that the results obtained with the model are overall quite robust, as differentiated gene expression by type and environment evolve over a wide range of biologically relevant values. In particular, these results are robust with regard to the size of topological domains, which corresponds to the maximum interaction distance for the transcription-supercoiling coupling and has been given different experimental values; to the mean intergenic distance, in a range that is representative of the bacterial species; and to the intensity of changes in the background supercoiling that represent environmental sources of stress. The gene regulatory networks that evolve in the model can indeed differentiate between environments that create much smaller levels of supercoiling than generated by transcription itself.

Overall, changing the value of each parameter in the model has two main consequences. The first – direct – consequence is to affect the phenotype of individuals, by modulating the effect of the transcription-supercoiling coupling on gene expression (for example, by increasing the number of neighbors a gene interacts with). This direct effect is illustrated in Figures 6.4 and 6.9. The second – indirect – consequence is to change the structure of the fitness landscape that underlies the possible evolutionary trajectories, by possibly affecting its dimension, its ruggedness, or the number and proportion of beneficial, neutral, or deleterious mutations. This indirect effect is illustrated in Figures 6.7 and 6.15. The fact that genomes with a higher number of genes or intergenic distances (above 10 kb) are not able to evolve a fitness as high as that of the smaller genomes exemplifies the strength of this second effect. Indeed, as the number of available mutations using the mutational operator of genomic inversions scales quadratically with the number of base pairs on the genome (as described above), the proportion of possible genotypes that can be explored over a given

number of generations with a constant number of individuals diminishes with as genome size increases. In order to quantify this effect, an interesting experiment would be to run additional simulations in which every length-related parameter (maximum interaction distance, mean intergenic distance, and gene length) is scaled down by the same factor. This would allow us to control the number of available mutations and therefore the speed of evolution, but should in principle not affect the biological relevance of the model.

Finally, introducing indels as an additional mutational operator reveals further characteristics of the fitness landscape in the model. Indeed, it is in simulations with an intergenic distance of 1 kb that the highest fitness was reached in the model, but the intergenic distances in the simulations with indels evolved in the opposite direction, resulting in a comparatively lower fitness. As the simulations with indels started with an intergenic distance of 125 bp, it would be interesting to explore different initial values of this parameter to see whether populations always converge towards lower intergenic distances, or if there is a threshold above which the higher fitness peak observed at high intergenic distances is reachable. In particular, a possible hypothesis to explain the attraction to low mean intergenic distances in simulations with indels could again be related to the genomic inversions, through a second-order selection for robustness or evolvability. If larger genomes have descendants that more frequently have lower fitness because of deleterious inversions than smaller genomes (or respectively less frequently have higher fitness), there could be a short-term indirect selection pressure towards smaller genomes, even though the fitness peak that is reachable in the long-term is lower than with larger genomes. In order to quantify the effect of these second-order selection pressures, both the robustness and evolvability of individuals in the model could be estimated experimentally, by measuring the average fitness of a large number of descendants of individuals taken from populations evolved with different parameter values.

The structure of the fitness landscape, as it emerges from the interplay between the biological parameters that control the supercoiling-mediated interaction between neighboring genes and the mutational operators that generate new genomes, therefore seems to play a fundamental role in determining the evolutionary trajectories that are available to evolving populations. Moreover, the structure itself of the fitness landscape changes in response to mutation of key parameters (such as genome size), resulting in a dynamic fitness seascape. For this reason, a better understanding of the evolution of this fitness seascape, or equivalently of the role that supercoiling mutations play in determining the evolutionary trajectories that are available to populations, would shed light on the epistatic interactions between mutations in the *EvoTSC* model, and therefore be an important research direction to pursue.

Chapter 7

Looking for Supercoiling Epistasis in *EvoTSC*

The results obtained with the *EvoTSC* model that I have presented up until now tackle the role that DNA supercoiling plays in the evolution of the structure of bacterial genomes, via the transcription-supercoiling coupling. In this chapter, I take the *EvoTSC* model in another direction, and go back to the idea of epistasis between mutations in the supercoiling level and other mutations, which was the question at the root of the research agenda of my PhD. In the experiment conducted with *Aevol* and presented in Chapter 3, the main hypothesis to explain why I was not able to detect a signal of epistasis between supercoiling mutations and other kinds of mutations is that the model of supercoiling that I implemented could have been too simplistic. That model might indeed not leave room for supercoiling mutations to open up evolutionary paths in the fitness landscape, and allow the lineages that bear these mutations to evolve faster than other lineages. In *EvoTSC*, supercoiling is on the contrary sufficiently finely modeled to allow the evolution of regulatory networks based on local variations in the supercoiling level (as shown in the previous chapters), which indicates that supercoiling mutations could present such an evolutionary role in this model.

In this chapter, I present an experiment – inspired by the *LTEE* – in which I measure whether previously evolved individuals can adapt faster to new environmental conditions with or without supercoiling mutations, in order to verify this hypothesis using the *EvoTSC* model. In the *LTEE*, supercoiling mutations have indeed been shown to evolve repeatedly, and to confer direct fitness benefits (Crozat et al., 2005, 2010). In order to let the supercoiling level of individuals in *EvoTSC* evolve, I introduce a mutational operator similar to the one used in the *Aevol* experiment presented in Chapter 3. Unlike in the *Aevol* experiment, the non-linear effect of the basal supercoiling level on gene expression in the *EvoTSC* model could allow populations with supercoiling mutations to follow qualitatively different evolutionary trajectories than populations with a constant supercoiling level in this model. In this chapter, I first present the methodology of this experiment, including the new mutational operator for the evolution of the supercoiling level. Then, I show that in this experiment, populations indeed adapt faster to new environments with supercoiling mutations than without supercoiling mutations. Finally, in order to understand this evolutionary advantage, I investigate the fitness landscapes that result from supercoiling mutations in the *EvoTSC* model.

7.1 Experimental Framework

Performing exactly the same experiment as the *Aevol* experiment described in Chapter 3 by measuring the waiting intervals before and after supercoiling mutations is not possible in *EvoTSC* (at the time of writing), as the ancestry tree of the population throughout generations and the precise set of mutations at each reproduction event are not recorded. Studying the lineage of the final population in order to study the properties of the mutations that fixed in the lineage is therefore not possible in *EvoTSC*. In order to evaluate the possible epistatic interactions between supercoiling mutations and other mutations in *EvoTSC*, I instead devised another experiment, which reproduces the setup of the *LTEE* in an *in silico* setting.

This experiment consists in two successive evolutionary runs. First, I let two sets of populations – with and without supercoiling mutations – evolve for 1,000,000 generations, and extracted *wild-type* individuals from these evolved populations. To obtain these wild-types, I reused the 30 populations that already evolved without supercoiling mutations that I presented in Chapter 5, and let 10 fresh populations evolve with supercoiling mutations. Then, I subjected these wild-types to new environmental conditions, replicating the beginning of the *LTEE*. In order to model these new conditions within the *EvoTSC* model, I reassigned at random the types (*A*, *B* or *AB*) of every gene in the genome of the wild-type individuals, while keeping constant the number of genes of each type in the genomes. As the environment in *EvoTSC* is represented by a pair of environments (A and B) with different gene expression targets, this corresponds to replacing the original environments with new environments A' and B', in which different subsets of genes (*A'*, *B'*, or *AB'*) must be activated or inhibited. Throughout this chapter, I will refer to this change of environment as an *environmental shock*, and to the individuals with shuffled gene types as *shocked* individuals. For the sake of clarity, I also refer to the new environments as A and B, and to the new gene types as *A*, *B* and *AB*. Note that, as there are three gene types, any given gene has a one in three chance of keeping the same type (and expression target) after the shock, and a two in three chance of having a new type. I then used these shocked individuals to create new populations, and let these populations evolve again in the new environments in order to study their re-adaptation.

7.1.1 Introducing Supercoiling Mutations

The mutational operator that I used for the mutations in the basal supercoiling level σ_{basal} of individuals in *EvoTSC* is similar to the one that I implemented in *Aevol* and that is presented in Section 3.3.3. When mutating an individual, we first decide whether to mutate its basal supercoiling level with a probability p, and draw a small change $\delta\sigma_{basal}$ to be added to the supercoiling level according to a normal law $\mathcal{N}(0, s^2)$. In this experiment, $p = 0.1$ and $s^2 = 0.0001$. Then, exactly as in the main experiment, the individual can undergo a series of genomic inversions which rearrange the relative position of genes on its genome.

Figure 7.1 presents the average fitness (top) and basal supercoiling level (bottom) of the best individual in each replicate during the evolution of the wild-type populations, with and without supercoiling mutations. We can first see that, in the wild-types that evolve with supercoiling mutations (light blue), fitness evolves in a qualitatively similar fashion to the main run, and is slightly higher by the end of evolution than without supercoiling mutations (dark

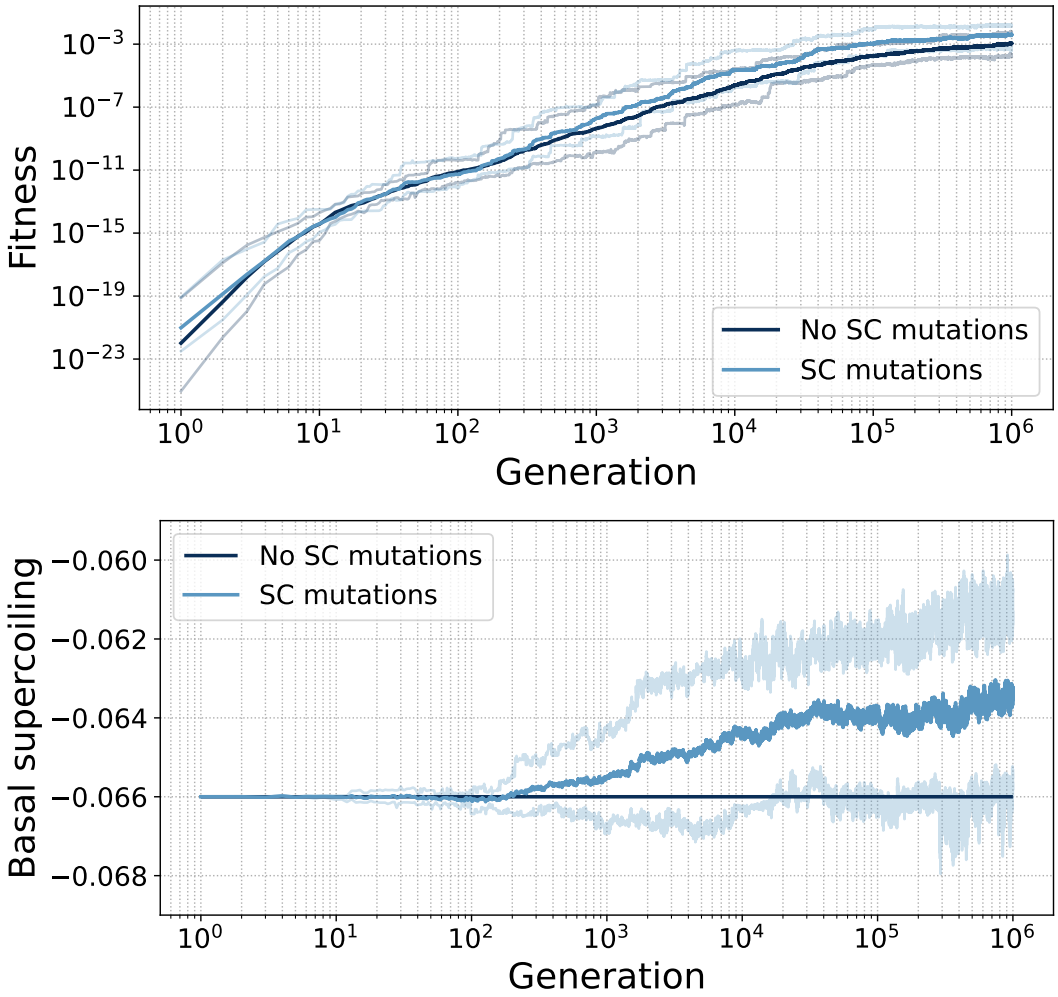


Figure 7.1: Top: average fitness of the best individual in every replicate during evolution, for the 10 wild-types with (light blue) and the 30 wild-types without (dark blue) supercoiling mutations. At the last generation, the fitness of populations with supercoiling mutations is significantly higher than the fitness of populations without supercoiling mutations ($p = 7.3 \cdot 10^{-4}$, Student's t-test for independent samples). Bottom: average basal supercoiling level of the best individual in every replicate during evolution of the wild-types with (light blue) and without (dark blue) supercoiling mutations. Lighter lines represent the first and last decile of the data.

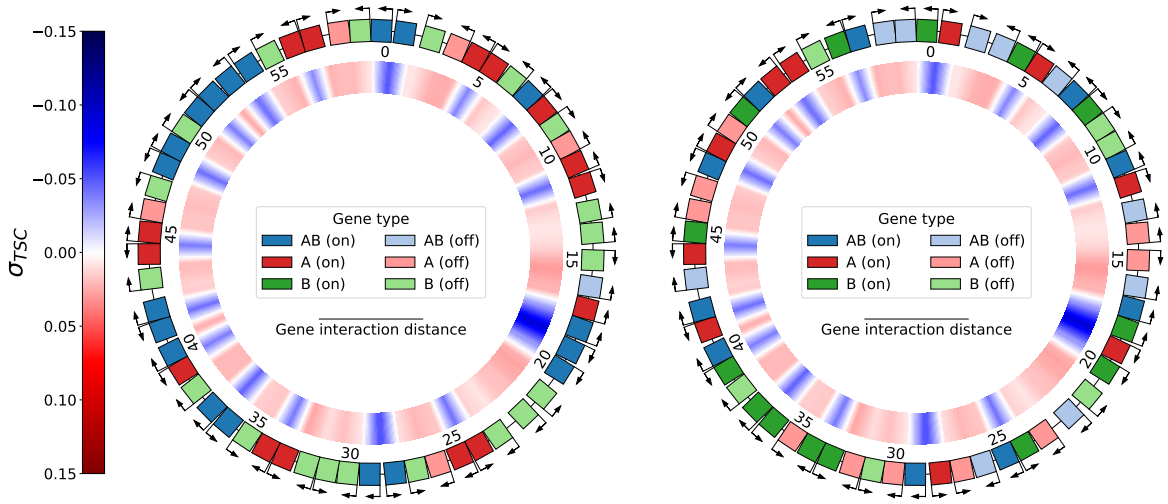


Figure 7.2: Genome of one of the wild-types that evolved with supercoiling mutations (left), and shocked individual created from that individual (right), both evaluated in environment A. The gene type (color) and activity (light or dark) of two-thirds of the genes changes, but not the local supercoiling level, as relative gene positions and gene expression levels remain constant.

blue, see the statistical test in the caption of the figure). Then, looking at the basal supercoiling level, we can see that the average level of negative supercoiling decreases over time during evolution. This indicates that the supercoiling level can indeed be targeted by selection in the model, and that there is a clear selection pressure towards reducing the amount of negative supercoiling in the specific context of this experiment. A possible hypothesis to explain both the higher fitness and lower negative supercoiling level of wild-types with supercoiling mutations comes from recalling that, with the initial basal supercoiling level of $\sigma_{basal} = -0.066$, genes tend to have a high expression level in both environments (see the dash-dotted curve of Figure 5.5). As a consequence, decreasing the level of negative supercoiling of the genome corresponds to shifting the background supercoiling in both environments to a less negative value. This decreases the bias towards high gene expression that is present in both environments, and therefore facilitates gene inhibition when required by the environment (data not shown).

7.1.2 Environmental Shock

In order to simulate the effect of an environmental shock on a given individual, that is to say of replacing environments A and B by new environments A' and B', we assign a new type at random to every gene on the genome of this individual, ensuring that the number of genes of each type remains constant. As there are 3 gene types (*A*, *B*, and *AB*), each gene has one in three chances of effectively staying of the same type, and two in three chances of effectively changing types. This represents the fact that some genes that had to be activated (resp. inhibited) in environment A or B must now be inhibited (resp. activated) in environment A'

or B'. Note that the only element of the environments that changes is the subset of genes that must be activated in each environment, but not the shift in supercoiling σ_A or σ_B that is caused by either environment.

A representative example of an environmental shock is depicted in Figure 7.2. On the left-hand side is the genome of a wild-type individual that evolved with supercoiling mutations, and on the right-hand side is the result of applying an environmental shock to this individual. The type (color) of two third of the genes changes, but not the local supercoiling level along the genome, as the gene positions themselves – and hence their expression level, as determined by the transcription-supercoiling coupling – remain unchanged. As a result, a large number of genes end up wrongly activated or inhibited, which opens the door to future compensatory mutations in the re-adaptation to this new environment: while the fitness of the wild-type was $1.34 \cdot 10^{-2}$, the fitness of the shocked individual is only $3.08 \cdot 10^{-29}$.

7.1.3 Experimental Protocol

The populations that I used for the wild-types without supercoiling mutations are the main populations already presented in detail in Chapters 5 and 6. For the wild-types with supercoiling mutations, I evolved 10 new populations, for the same number of generations as the main runs, with all other parameters kept exactly the same. I then chose 5 representative wild-types at random from each set of simulations. From each of these wild-type individuals that evolved with or without supercoiling mutations, I created 5 different shocked individuals with shuffled genes, resulting in a total of 25 shocked individuals. For each shocked individual, I then created 5 populations, each initialized with clones of that individual but using different seeds, and let each population evolve for 50,000 generations, in order to recover from the environmental shock. This allowed me to compare the speed of the initial evolution after the shock in 125 populations with supercoiling mutations, and 125 populations without supercoiling mutations. All the data from this experiment is available online on the Zenodo platform.

7.2 Results

7.2.1 Evolution after an Environmental Shock

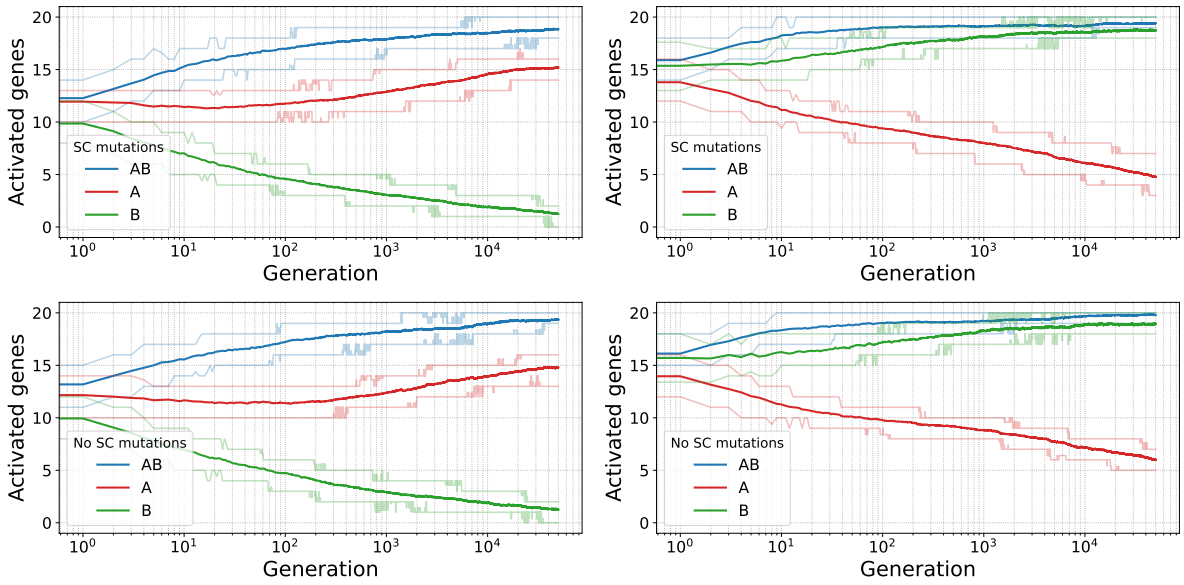


Figure 7.3: Average number of activated genes of each type in environment A (left) and B (right) during evolution, with (top) and without (bottom) supercoiling mutations. Lighter lines represent the first and last decile of the data.

Figure 7.3 shows the evolution of the average number of activated genes of each type in each environment after the environmental shocks, averaged over the 125 simulations without supercoiling mutations (top) and the 125 simulations with supercoiling mutations (bottom). As could be expected after the shock – as it does not affect gene positions and orientations – and given the example individual in Figure 7.2, the initial number of activated genes is initially very similar for each type (note that the first shown generation is the first non-clonal generation after the shock, and that one round of mutation and selection has therefore already taken place). However, the number of activated genes of each type then quickly evolves towards their respective targets, as in the previous simulations conducted with the model. Starting from a genome in which genes have been positioned (by selection) to form a regulatory network adapted to the environments before the shock therefore does not seem to hinder the evolution of a regulatory network adapted to new environments after the shock.

In the *LTEE*, the repeated fixation of supercoiling mutations in 11 out of the 12 replicates shows that, in each of these replicates, the lineages that bear these mutations were able to outcompete the other lineages present in the replicate. A similar pattern can be observed in this experiment in the evolution of populations with supercoiling mutations, compared to the evolution of populations without supercoiling mutations, after an environmental shock. Figure 7.4 shows the evolution of the average relative fitness of the best individual of each population compared to the fitness (before the environmental shock) of the wild-type individual that the population originates from (top), and the evolution of the basal supercoiling

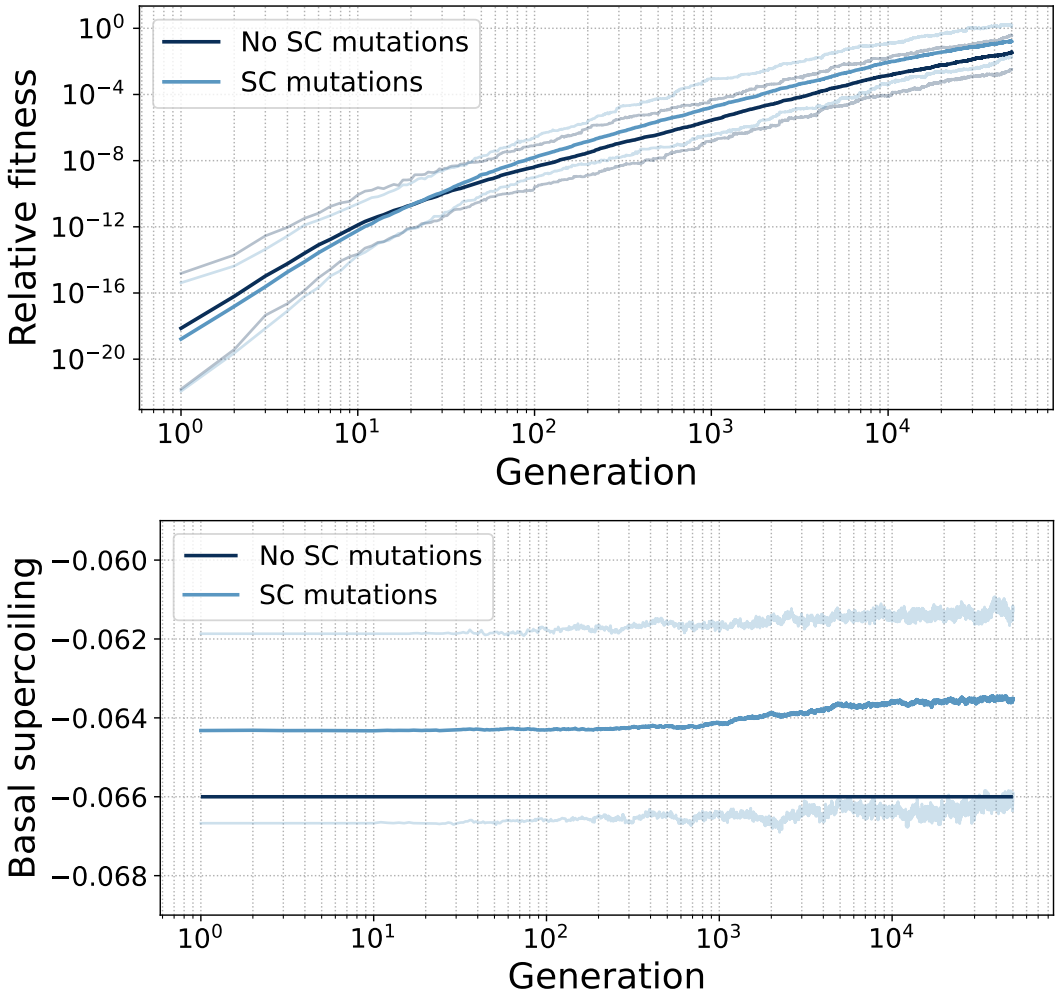


Figure 7.4: Top: Evolution of the average fitness relative to the wild-type before the environmental shock, for the populations with (light blue) and without (dark blue) supercoiling mutations. At the last generation, the relative fitness of populations with supercoiling mutations is significantly higher than the relative fitness of populations without supercoiling mutations ($p = 5.8 \cdot 10^{-6}$, Student's t-test for independent samples). Bottom: Evolution of the average basal supercoiling level of the populations with supercoiling mutations (light blue), compared to the basal supercoiling level of populations without supercoiling mutations (dark blue). Lighter lines represent the first and last decile of the data.

level in populations with supercoiling mutations. Similarly to the evolution of the wild-types, it seems that populations in which supercoiling can evolve perform better over time than populations in which it cannot. In particular, populations with supercoiling mutations end up with a higher average fitness than populations without supercoiling mutations, after only 50,000 generations of evolution (see the statistical test in the caption of the figure). Averaging over all 125 populations with supercoiling mutations, the supercoiling level (bottom) seems to evolve towards a slightly less negative level than just after the shock. However, as the supercoiling level had not clearly converged by the end of the evolution of the wild-types (see Figure 7.1), this trend could equally well be the same as in the wild-types or a sign of adaptation to new environments after the shock.

7.2.2 Supercoiling Fitness Landscapes

In the *LTEE*, two main hypotheses have been put forward to explain the repeated fixation of supercoiling mutations that has been observed in the lineages of the experiment. These mutations could indeed provide an evolutionary advantage to the lineages in which they appear, by increasing their evolvability through epistatic interactions with supercoiling-regulated genes. However, some of these mutations have been shown to be directly advantageous, in that they already confer a fitness benefit when inserted into the ancestral strain (Crozat et al., 2005). It is therefore possible that these mutations were simply selected for their immediate benefit, and did not play a particular role in shaping evolutionary trajectories in the fitness landscape through epistatic interactions. As these hypotheses also apply to the results of this experiment, I decided to further study the direct fitness effect of supercoiling mutations in the *EvoTSC* model, by examining the associated fitness landscapes. As a first step, I computed the empirical fitness landscapes for supercoiling mutations of the wild-type individuals before and after the environmental shock, and after re-adaptation to their new environments, in order to see to which extent these fitness landscapes are explored in practice during evolution in the model.

The fitness landscape represents fitness as a function of the genotype. As we are interested in the fitness effect of supercoiling mutations, we consider the genotype of individuals as consisting only in their basal supercoiling level, while considering their genomic organization – and the associated gene regulatory networks – constant. The fitness landscape is hence one-dimensional, and can be easily explored and represented. The fitness landscapes of the wild-type individuals are presented in Figure 7.5. The 5 wild-types that evolved with supercoiling mutations are shown in the top panel, and the 5 wild-types that evolved without supercoiling mutations are shown in the bottom panel. In each landscape, the star represents the basal supercoiling level of the individual itself. For the wild-types that evolved without supercoiling mutations, all wild-types have a basal supercoiling level $\sigma_{basal} = -0.066$, but this is not the case for the wild-types that evolved with supercoiling mutations (see Figure 7.1 (bottom) for the evolution of the average of their supercoiling level). All fitness landscapes have a roughly pyramidal shape, with a well-defined main fitness peak, surrounded by descending slopes that contain small local peaks. All wild-types, which are the result of 1,000,000 generations of evolution, are located at the global peak of their respective fitness landscape. This indicates that, for these individuals, no higher fitness is reachable through

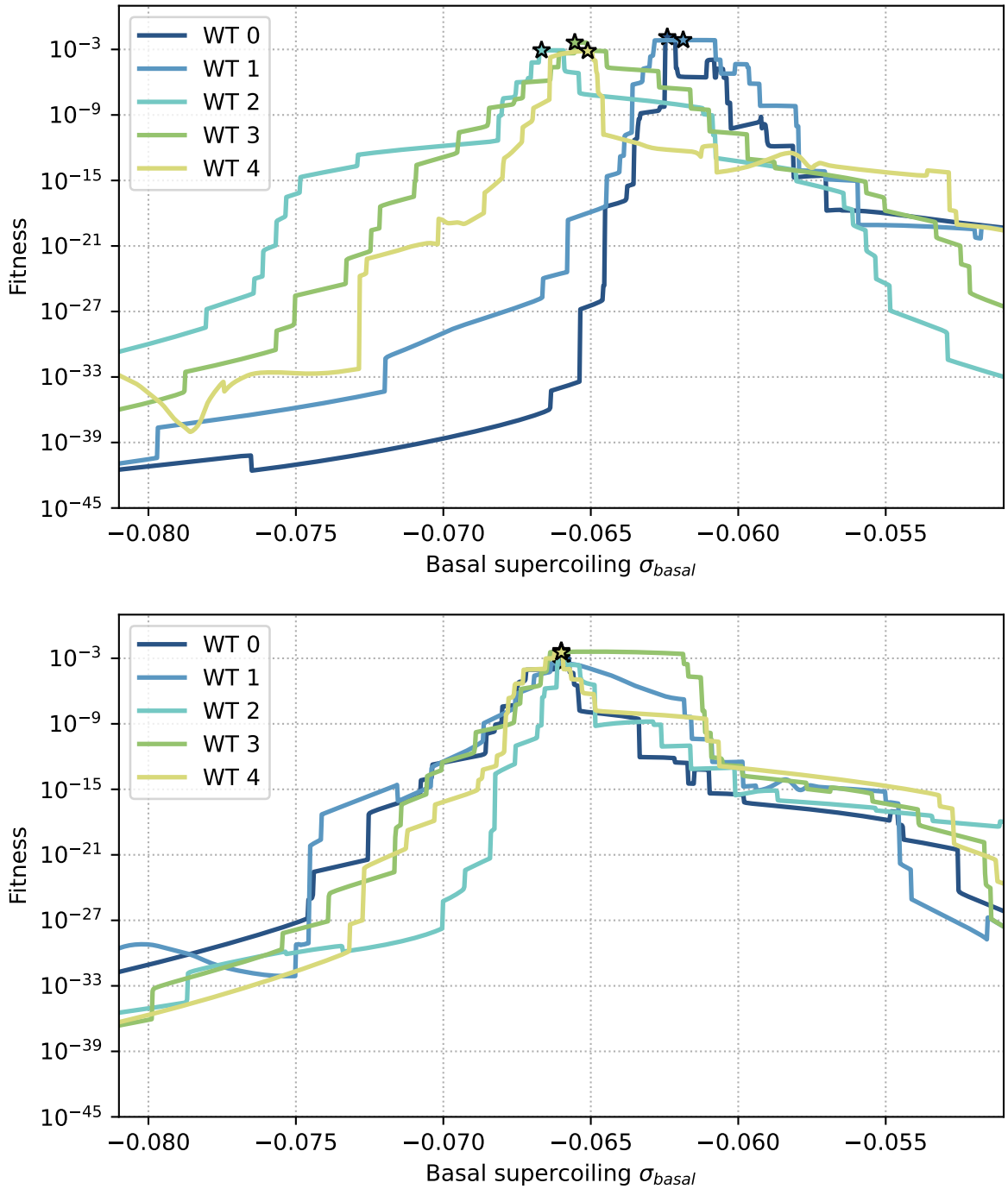


Figure 7.5: Fitness as a function of the basal supercoiling level, for the wild-type individuals evolved with (top) and without (bottom) supercoiling mutations. The stars represent the basal supercoiling level and fitness of each wild-type.

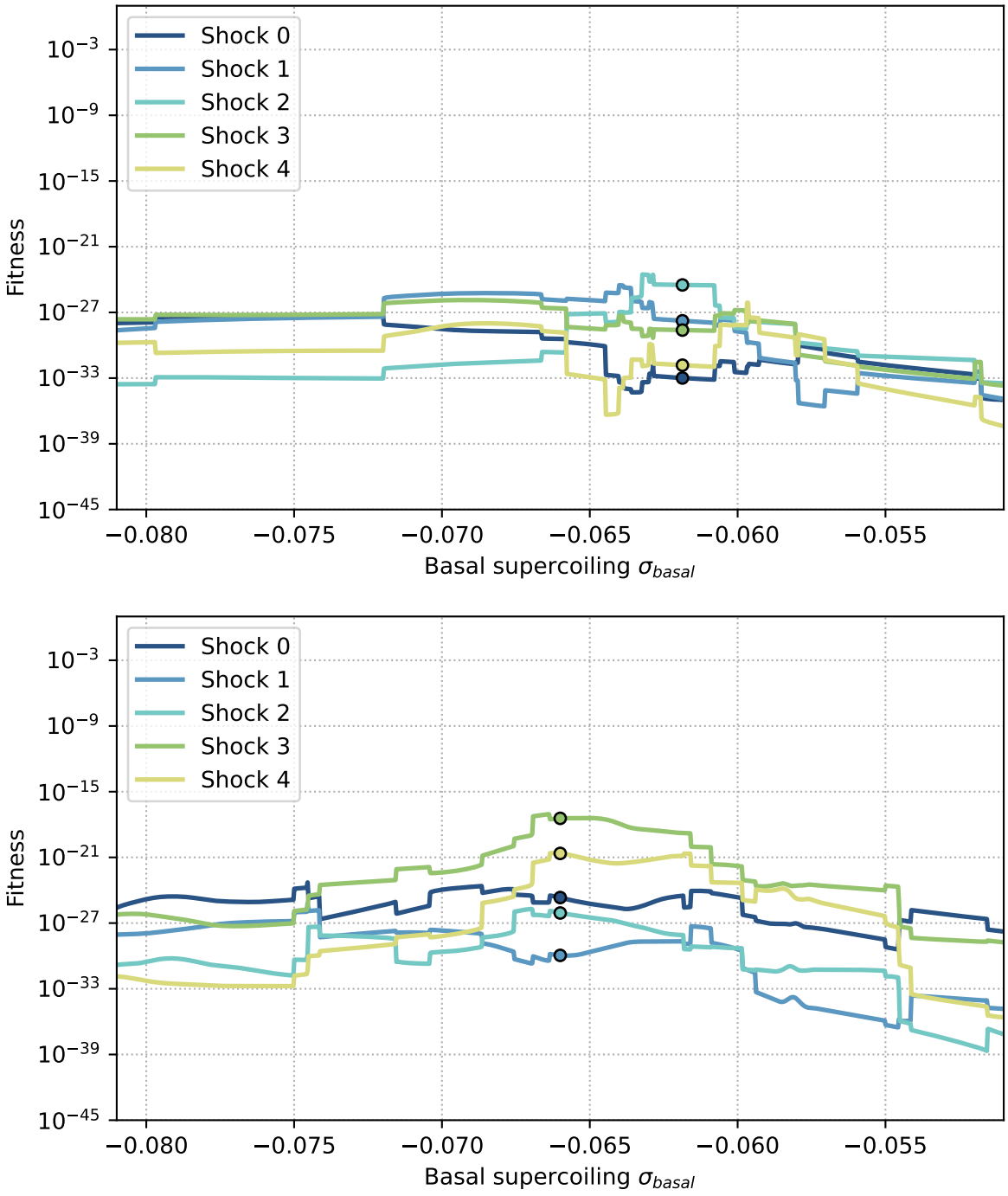


Figure 7.6: Fitness as a function of the basal supercoiling level, for five shocked individuals created from one of the wild-types, with (top) and without (bottom) supercoiling mutations. The circles represent the basal supercoiling level and fitness of each shocked individual.

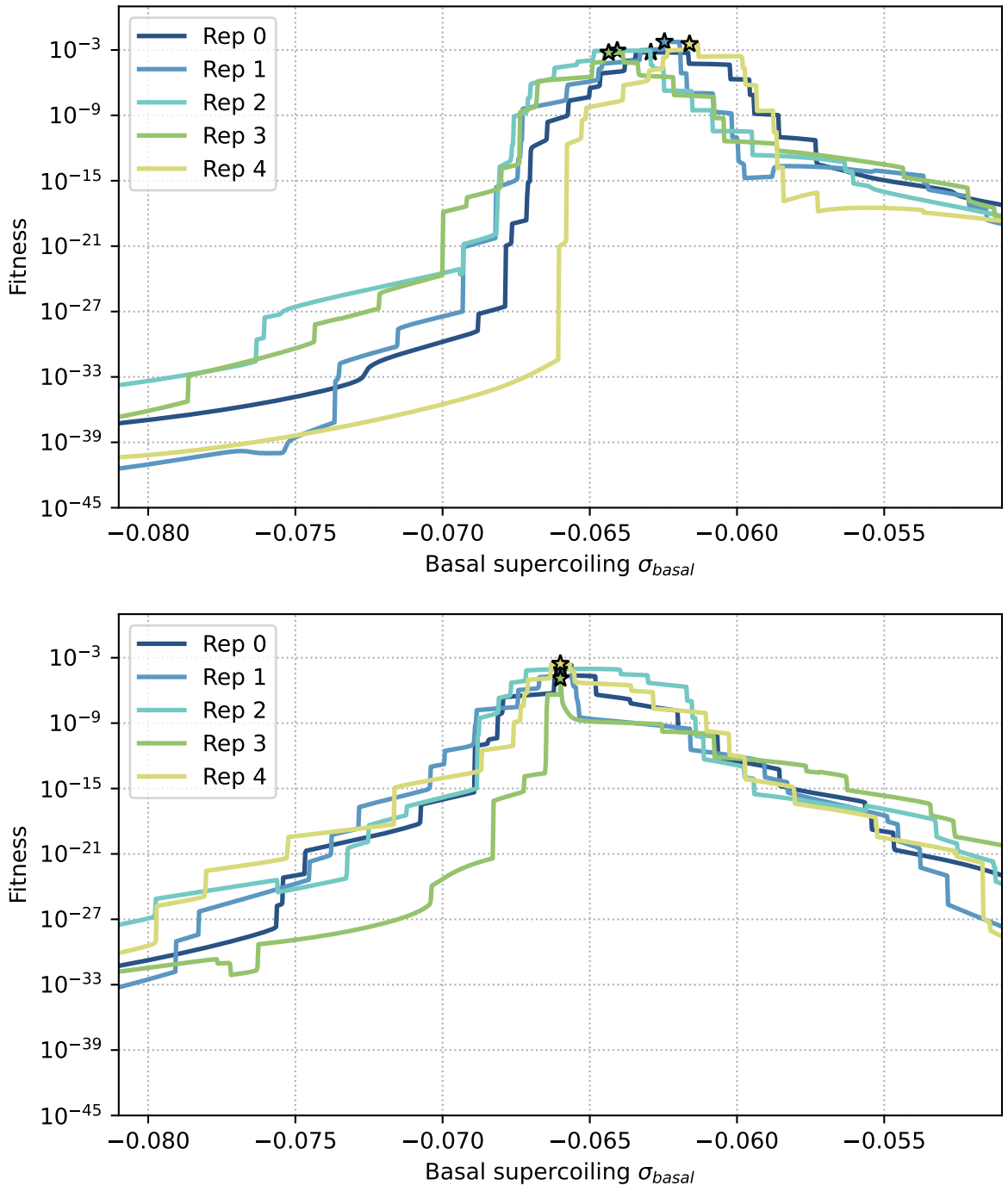


Figure 7.7: Fitness as a function of the basal supercoiling level, for the five replicates of one of the shocked wild-types after 50,000 generations of evolution with (top) and without (bottom) supercoiling mutations. The stars represent the basal supercoiling level and fitness of the best individual in each replicate.

supercoiling mutations only, meaning that when supercoiling mutations are available, both the supercoiling level and the genomic organization coevolve in order to reach the summit of the fitness landscape. Even in the absence of supercoiling mutations, the fitness landscape that emerges from genomic rearrangements is nonetheless shaped in such a way that the supercoiling level of the individual stands on a fitness peak.

Figure 7.6 shows the fitness landscapes of 5 shocked individuals obtained from one of the wild-types that evolved with (top) or without (bottom) supercoiling mutations, and the circle on each curve represents the fitness of each shocked individual. First, we can see that the fitness peaks on these landscapes are much lower than in the wild-type individuals in Figure 7.5, which shows that environmental shocks greatly affect the fitness landscape. Moreover, none of the shocked individuals stand on a fitness peak any longer, which shows that even with the same genomic organization, the optimal supercoiling level is different after an environmental shock for these individuals.

Finally, Figure 7.7 shows for comparison the fitness landscapes at the end of the 50,000 generations of evolution for 5 replicates of one of the shocked wild-types, originating from a population that evolved with (top) or without (bottom) supercoiling mutations. In both cases, after only 50,000 generations, the fitness landscape already has a comparable shape to that of the wild-type individuals, with a single fitness peak at which the evolved individual is located. Even after an environmental shock, and no matter whether supercoiling mutations are available to evolution or not, the best individual at the end of evolution is therefore at the global peak of the supercoiling fitness landscape that emerges through the evolution of its genomic organization.

7.2.3 Evolution with Supercoiling Mutations Only

In order to understand to which extent the exploration of these supercoiling fitness landscapes is actually driven by supercoiling mutations, rather than by the genomic inversions which alter these fitness landscapes, I re-ran the tape of evolution after the environmental shock. This time, I let only the supercoiling level of individuals evolve, but not their genomic organization, and ran these new simulations only for the 5 wild-types which had already evolved with supercoiling mutations.

The evolution of the average fitness in these simulations is shown in Figure 7.8. During the 50,000 generations of evolution, the average fitness of each population with only supercoiling mutations (light blue) does increase, but to a much smaller extent than when genomic rearrangements are allowed (dark blue). For each wild-type, the basal supercoiling level evolves over time, indicating that selection is taking place, but the fitness increase resulting from these mutations nonetheless remains considerably smaller than what is possible with genomic inversions.

Figure 7.9 shows the evolution of the basal supercoiling level of all replicates of the 5 shocked individuals created from one of the wild-types that evolved with supercoiling mutations, when only supercoiling mutations are allowed. For each shocked individual, the supercoiling level of each of their replicates seems to converge to nearly identical values by the end of evolution. This shows that, when only supercoiling mutations are available, evolution seems to be fully repeatable. In simulations in which only the supercoiling level

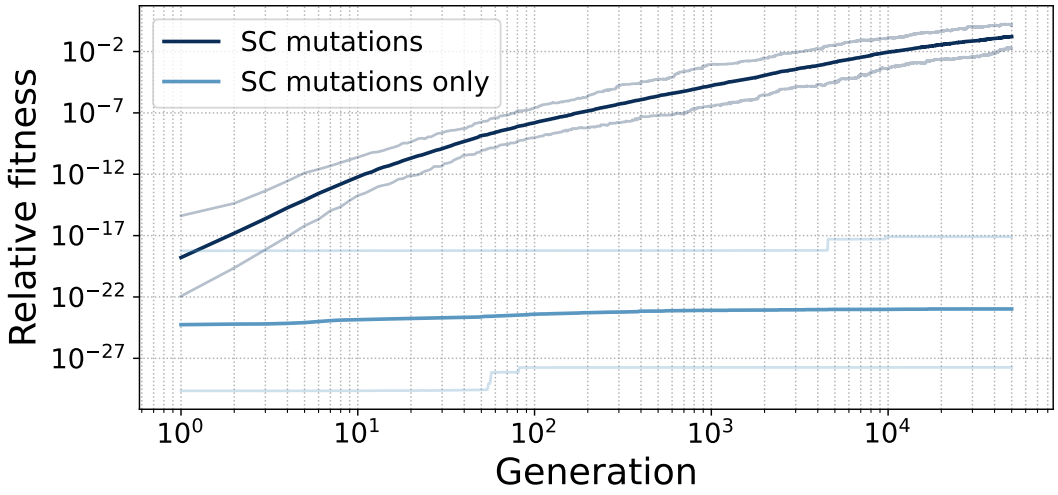


Figure 7.8: Evolution of the average fitness relative to the wild-type before an environmental shock, for populations with only supercoiling mutations (light blue) and populations with both supercoiling mutations and genomic rearrangements (dark blue). Lighter lines represent the first and last decile of the data.

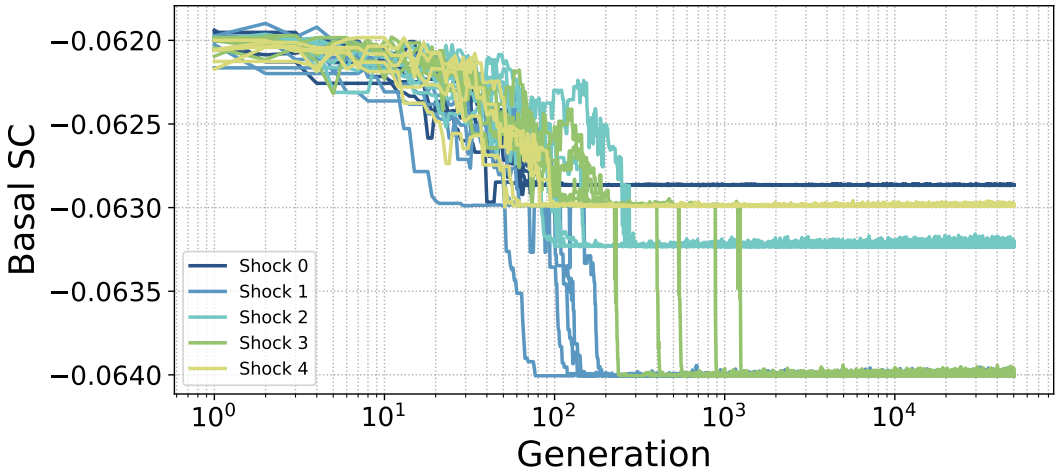


Figure 7.9: Evolution of the basal supercoiling level of each replicate of all the shocked individuals created from the wild-type presented in Figure 7.2. For each shocked individual, the 5 replicates are drawn in the same color to highlight the repeatability of the final basal supercoiling level.

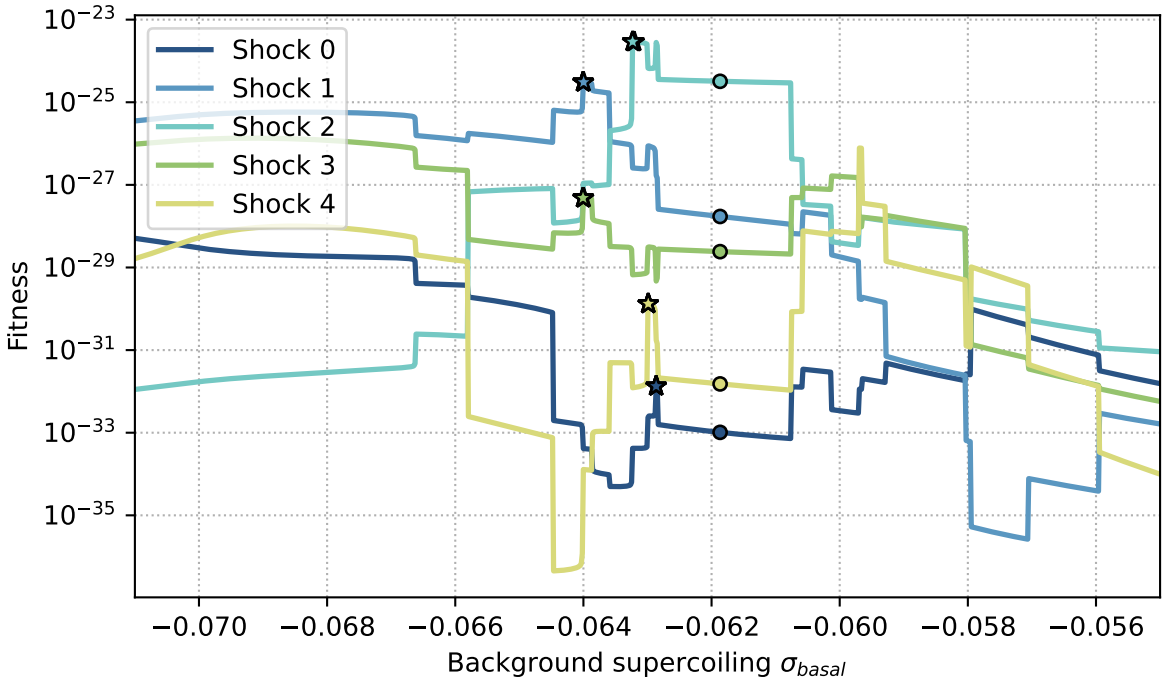


Figure 7.10: Fitness landscapes of the 5 shocked individuals obtained from one of the wild-types that evolved with supercoiling mutations. The circles represent the initial basal supercoiling level of each shocked individual (the same as in the top panel of Figure 7.6), and the stars represent the basal supercoiling level of each of the 5 replicates of each shocked individual at the end of evolution. Note that for all shocked individuals, the 5 stars of each of their evolved replicates are superimposed.

can evolve, the shape of the fitness landscape does indeed not change, as it depends on the (constant) organization of genes on the genome. In these simulations, it is therefore possible to compare the evolved individuals with the original shocked individuals directly on their fitness landscape. The fitness landscapes of the same shocked individuals are presented in Figure 7.10, along with the original supercoiling level of each shocked individual (circles) and the evolved supercoiling level of each replicate (stars).

We can first observe that the basal supercoiling level (circle) of each of shocked individual does not stand on a peak of their respective fitness landscape anymore, after these landscapes have been altered by an environmental shock. At the end of evolution, the basal supercoiling level of each replicate (stars) of these shocked individuals however always reaches a peak of the altered fitness landscape, but not necessarily a global one. For each shocked individual, the 5 evolved replicates of that individual seem to reach the same fitness peak, as the 5 stars on each landscape are virtually stacked at the same location. In this case, when the only mutations allowed are supercoiling mutations, the evolutionary process therefore seems to be completely reproducible (as is also shown in Figure 7.9). Moreover, while the evolved populations all reach local fitness peaks, they do not all cross the fitness valleys that separate their local fitness peaks from higher, but further located, peaks. Indeed, while populations

generated from shocked individuals 1 and 2 in particular were able to cross fitness valleys and reach the global peak of their respective fitness landscape, populations 3, 4 and 5 were not. In particular, the basal supercoiling level of population 4 evolved towards a local fitness peak that is located in the opposite direction to the global peak of the fitness landscape of that population.

Overall, the evolutionary trajectories of populations in which only supercoiling mutations are allowed show us that in the *EvoTSC* model, supercoiling mutations do have an evolutionary effect, but that this effect is very weak compared to that of genomic rearrangements. In this model, supercoiling mutations seem to play a role in finding local optimizations on the fitness landscape that stems from the genomic organization of individuals, rather than in opening new evolutionary paths through a broad regulatory effect on gene activity.

7.3 Discussion

In this chapter, I presented a new experiment that I conducted with the *EvoTSC* model, with the aim of evaluating the role of supercoiling mutations in evolution after a change in environmental conditions, following the example of the *LTEE*. In the *LTEE*, such supercoiling mutations were indeed found in all but one of the lineages, and the repeated character of these mutations therefore seems to indicate that they played an important role in adapting to the new conditions of the experiment. However, as these mutations were shown to be directly beneficial in the ancestral strain, the extent to which these mutations could additionally have been indirectly selected because of their epistatic interactions remains unclear.

In order to study this question in the *in silico* setting of the *EvoTSC* model, I first evolved wild-type populations with supercoiling mutations, and showed that these populations were able to reach higher fitness than populations that evolved without supercoiling mutations. I then subjected wild-type individuals extracted from both populations to environmental shocks and let populations seeded with these individuals evolve again. I showed that populations in which the supercoiling can evolve re-evolved higher fitness – when compared to their wild-type ancestor before the shock – than populations in which it cannot, after 50,000 generations of evolution. As in the *LTEE*, supercoiling mutations in *EvoTSC* therefore seem to provide a relative advantage to the lineages in which they appear.

In order to have a more quantitative idea of the evolutionary possibilities afforded by these supercoiling mutations, I then computed empirical fitness landscapes as a function of supercoiling in the wild-type and re-evolved individuals. I showed that, in the wild-types that evolved both with or without supercoiling, the supercoiling level of evolved individuals corresponds to the global peak of the one-dimensional supercoiling fitness landscape. As these fitness landscapes are rooted in the organization of the genes on the genome (via the transcription-supercoiling coupling), these results show that supercoiling mutations are in fact not necessary to reach the peak of the fitness landscape, even though the peaks are on average higher with supercoiling mutations (as shown during the evolution of the wild-types themselves). I then ran another set of simulations in which only the supercoiling level is allowed to evolve, in order to evaluate precisely the extent to which supercoiling mutations enable the exploration of the supercoiling fitness landscape. I showed that, while supercoil-

ing mutations do allow populations to repeatably reach local fitness peaks (as could naively be expected), they are by themselves not sufficient to allow populations to reach the global peak of their respective fitness landscape.

Taken all together, these observations seem to indicate that, in the *EvoTSC* model, supercoiling mutations allow for a local exploration of the fitness landscape generated by genomic rearrangements rather than for the opening of new evolutionary paths that would otherwise remain inaccessible. These mutations could nonetheless provide a significant enough fitness advantage to the lineages in which they appear, by finding a locally optimal level of supercoiling, for these lineages to be repeatedly selected. The perfect repeatability of the discovery of local fitness peaks through supercoiling mutations in the context of this experiment echoes the repeated fixation of supercoiling mutations in the *LTEE*, suggesting that these mutations could indeed have been selected for their direct fitness effect rather than for their indirect effect on evolvability or robustness. A further answer to this question could be obtained with the help of the complete lineages and mutation histories of populations which evolve in the *EvoTSC* model. This data would indeed allow us to reconstruct the succession of fitness landscapes – known as the fitness seascape – generated by successive genomic inversions, and to assess the extent to which supercoiling mutations indeed allow evolving populations to reach the successive peaks of this fitness seascape during evolution.

Chapter 8

Conclusion

8.1 Summary

The overarching scientific question at the root of the work I conducted during my PhD is the following: how can evolution, an inherently random process, sometimes be reproducible? In order to refine this far-reaching problem into a more actionable research program, I focused on the study of the epistatic interactions between different kinds of mutations, and asked the following question: can an improved understanding of the beneficial or deleterious nature of epistatic interactions help us predict which strain within a given population has the highest potential to harbor future favorable mutations and thus outcompete other strains, given its current genetic background? Inspired by results obtained in the *Long-Term Evolution Experiment* (Lenski et al., 1991), I singled out DNA supercoiling as a prime example with which to try to corroborate this suggestion (Chapter 2). The level of DNA supercoiling is indeed at the same time intricately governed by many cellular processes – that are subject to possible mutations – and a fundamental actor in the regulation of gene transcription and expression. The central role of DNA supercoiling, which bridges the physical structure of the genome with the molecular phenotype of the cell, makes mutations that influence its level prime suspects in the shaping of evolutionary trajectories by epistatic interactions.

In order to tackle this problem, I first implemented a simple model of the level of supercoiling and of its effect on transcription in the *Aevol in silico* experimental evolution platform, but was not able to establish that supercoiling mutations played a measurable role in speeding up evolution after they occur (Chapter 3). As I estimated this lack of results to be due to the inner complexity of the *Aevol* model and the resulting overly simplistic representation of supercoiling that I could incorporate into the model, I designed and implemented a new multi-scale model, called *EvoTSC*. This new model trades off the precise genome description of *Aevol* for a much more detailed modeling of the coupling between transcription and supercoiling. I first validated the relevance of this new model by showing that its description of supercoiling is rich enough to allow for the evolution of differentiated expression patterns in response to environmental perturbations (Chapter 4), even when the only available mutations are genomic inversions, and observed the emergence of relaxation-activated genes in the model (Chapter 5). I then characterized the gene regulatory networks, mediated by

the transcription-supercoiling coupling, that are at the root of these differentiated transcriptional responses. In particular, I showed that these regulatory networks are based on the relative positions of genes on the genome, and that they require the interaction of multiple genes to function, spanning wide swathes of the genome.

In order to reinforce the degree of confidence that we can have in the conclusions hitherto drawn from the *EvoTSC* model, I then ran additional sets of simulations with the same model but using different parameter values (Chapter 6). I showed that the results presented above are robust with respect to the size of the topological domains (the distance at which supercoiling propagates), to the size of the intergenic regions (within the range of observed bacterial values), and to the intensity of the environmental perturbations (in particular when the perturbations are very small compared to the other sources of supercoiling). Overall, I was able to make two main conclusions with the help of the *EvoTSC* model regarding the transcription-supercoiling coupling. First, I showed that this coupling provides a material basis for gene regulation even in the absence of transcription factors, as well as a mechanism that could explain the activation of certain genes by DNA relaxation. Second, I showed that this coupling also plays a plausible role in the shaping of the organization of bacterial genomes over evolutionary time, via the supercoiling-mediated gene regulatory networks that evolve through successive genomic rearrangements.

These results, even though they provide some insight into the evolution of bacterial genomes, do however not yet answer the original question of the epistatic interactions between supercoiling mutations and other mutations, which was left inconclusively answered in Chapter 3. I therefore addressed this question again, this time using the *EvoTSC* model, by introducing supercoiling mutations alongside the original genomic inversions and studying the associated supercoiling fitness landscapes (Chapter 7). I first showed that populations in which supercoiling can mutate along genomic inversions evolve a slightly higher fitness after the same number of generations than populations in which it cannot. Then, replicating the environmental shock present at the beginning of the *LTEE* into the *EvoTSC* model, I showed that populations with supercoiling mutations additionally recover faster from an environmental shock than populations without supercoiling mutations, echoing the experimental results from the *LTEE* (Crozat et al., 2010). Finally, I showed that on their own, supercoiling mutations are only able to provide a local exploration of the supercoiling fitness landscape that stems from the organization of genes on the genomes, but that they do so in a repeatable manner despite their inherent randomness.

In this work, I used an evolutionary systems biology approach to study the evolution of gene regulation by DNA supercoiling in bacteria. Using both the *Aevol* and *EvoTSC in silico* experimental evolution platforms, I showed that supercoiling mutations seem to be selected for their direct fitness benefits, rather than for the increased evolvability that they could provide to their bearer through biased epistatic interactions. These results indicate that the repeated supercoiling mutations observed in the *LTEE* could similarly have been selected as a result of their direct benefits.

The results that I obtained with the *EvoTSC* model nonetheless underline the significant role that supercoiling could play in gene regulation in bacteria. Indeed, these results show

that, far from being limited to the direct biophysical effect of DNA supercoiling on transcription, gene regulatory networks that are expressive enough to activate or inhibit subsets of genes can evolve when the only regulator of gene expression is the supercoiling generated by gene transcription. In particular, regulation by supercoiling is sufficient to generate relaxation-activated genes, such as the ones found in *E. coli*, *S. pneumoniae* or *D. dadantii* (Peter et al., 2004; Ferrandiz et al., 2010; Pineau et al., 2022). It could moreover explain the evolutionary conservation of the relative positions of groups of neighboring genes, or syntenies, that has been for example observed between *E. coli* and *S. enterica*, as these groups of genes could coordinate their expression levels through local changes in supercoiling (Junier and Rivoire, 2016). Finally, these results reinforce the hypothesis that DNA supercoiling could play an important regulatory role in bacteria with streamlined genomes such as *B. aphidicola*, which is nearly devoid of traditional transcription factors (Brinza et al., 2013).

8.2 Perspectives

The work presented in this manuscript could be furthered along at least two main directions. The first direction would be to continue the investigation of supercoiling as a representative example of the role that epistatic interactions play in the structure of fitness landscapes, in order to better understand the repeatability of the apparition and fixation of such mutations in the *LTEE*. A first step in this direction would be to pursue in more detail the analysis of evolutionary trajectories in the *EvoTSC* model, by recording individual lineages during evolution and explicitly reconstructing the associated mutational histories. This would allow the same experiment as the one presented in *Aevol* (in Chapter 3) to be performed in a model in which supercoiling mutations have an effect on the fitness landscape – and hence on evolutionary trajectories – that is less independent from the rest of the genotype, and in particular from genome structure. Studying the fixation times of mutations in the *EvoTSC* model would in particular provide data that could prove easier to interpret than in *Aevol*, as there are only two possible kinds of mutations (genomic inversions and supercoiling mutations) in *EvoTSC*, compared to the 8 kinds of local and global mutations in *Aevol*. Another natural, but slightly more difficult to carry out, step in that direction would be to instead implement the more precise model of supercoiling used in *EvoTSC* in *Aevol*. Replicating the results already obtained in *EvoTSC* in a model in which genomes can evolve in additional ways in response to supercoiling mutations would provide an additional degree of confidence in the results presented in this work.

The common thread at the root of both these directions is to build models in which supercoiling mutations are sufficiently finely modeled to allow these mutations to cause jumps in the fitness landscape that are substantial enough to guide evolution towards different paths, conditionally to their occurrence in a lineage. Such models indeed seem necessary in order to definitely answer the question of the nature of the evolutionary role of supercoiling mutations from a theoretical perspective. Whether in *EvoTSC* or in *Aevol*, it seems promising to explore this direction more quantitatively by computing the distribution of fitness effects of supercoiling mutations in either model, or the fitness of double mutants (coupling a genomic inversion with a supercoiling mutation), instead of relying on the random occurrence

of these mutations in evolving populations. This more local – but more detailed – analysis of the fitness landscape could shed further light on the possible bias of supercoiling mutations towards beneficial epistatic interactions with other mutations.

The second main direction in which to pursue the work presented in this manuscript would be to move away from questioning the evolutionary role of supercoiling and instead towards a quantitative description of the regulatory role of the transcription-supercoiling coupling, at the mesoscale of bacterial topological domains. As supercoiling has been hypothesized to play an important role in bacterial gene regulation (El Houdaigui et al., 2019), it would be very interesting to obtain a model that explicitly accounts for the supercoiling level when predicting the expression levels of genes in a given – possibly synthetic – genetic system. In this regard, one of the main drawbacks of the model of the interaction between transcription and supercoiling that is used in *EvoTSC* is its lack of an explicit time scale. As such, *EvoTSC* therefore models the expected average behavior of genes subject to this interaction, but cannot adequately picture the dynamic nature of gene transcription by RNA polymerases. Yet, the stochasticity of this process plays an important role in the accurate modeling of gene transcription dynamics, as exemplified in Sevier and Hormoz (2021). Incorporating an explicit description of the movement of polymerases along DNA, of the resolution of supercoils by topoisomerases, and of the formation of supercoiling barriers by nucleoid-associated proteins into *EvoTSC* would provide us with a model able to disentangle the contribution of these processes to gene transcription and to provide quantitative predictions of gene expression levels. Another recently evidenced component of the effect of DNA supercoiling on bacterial gene transcription is the effect of promoter discriminator sequence and spacer length on transcription initiation (Forquet et al., 2021, 2022; Pineau et al., 2022). Incorporating this neglected effect into the model would further reinforce the quality of the prediction of gene expression levels by the model. Within the evolutionary framework of *EvoTSC*, such a quantitative model would moreover allow the study of the transcriptional response of genetic systems to perturbations caused by mutations in each of their components, be it gene order, promoter sequence, or topoisomerase activity, and hence the prevision of possible future evolutionary trajectories.

Appendix A

Software Contributions

A.1 *Aevol*

For the first year of my Ph.D., I mainly used the *Aevol in silico* experimental evolution platform, which is available here: https://gitlab.inria.fr/aevol/aevol, to run experiments. *Aevol* has been in development in the Inria BEAGLE team for over 17 years, and contains over 90,000 lines of C++ code. It has been used to run experiments that resulted in numerous publications, among which Knibbe et al. (2005), Batut et al. (2013), or Rutten et al. (2019). While I was using *Aevol*, I took an active part in the maintenance of its complex code base, and particularly focused on fixing memory leaks, and on improving the overall quality of the code. The associated commits can be found here: https://gitlab.inria.fr/aevol/aevol/-/commits/main?author=Théotime%20Grohens.

A.1.1 DNA Supercoiling in *Aevol*

As presented in Section 3.3, I developed during my Ph.D. a version of *Aevol* that incorporates the effect of supercoiling on gene transcription. This version can be found here: https://gitlab.inria.fr/tgrohens/aevol/-/tree/rebased-supercoiling. Developing this extension of the model required extensive changes in the evaluation of individuals, in order to take supercoiling into account. I had to update both the code handling the mutations – as I added a new kind of mutations to the model – and the *Aevol* post-treatments, the code that analyses experimental data after the main simulation is complete.

A.1.2 Tooling for *Aevol*

While using *Aevol*, I developed a set of Python tools, *aevol-utilities* (available here: https://gitlab.inria.fr/tgrohens/aevol-utilities), in order to make it easier to carry out full-fledged experiments. The default *Aevol* executables indeed do not provide the most accessible interface to an untrained user, and do not reflect the way the platform is used to run experiments at present. The typical use case of *Aevol* consists in the following steps, for each replicate of a given experiment:

1. Creating an initial random individual and an initial population using this individual, with a different seed for each replicate.
2. Running the proper evolution experiment for a given number of generations.
3. Picking an individual in the final population, and tracing its lineage since the original population.
4. Computing a series of statistics over this lineage, by recovering every individual in the lineage.
5. Analyzing and plotting the resulting data.

The *aevol-utilities* repository comprises a Python library (named `aevol.py`), and a series of Jupyter notebooks that are built upon this library. The `Runner aevol` notebook automates the first four steps of the pipeline above. It takes care of creating properly initialized repositories in order to minimize the risk of data loss, of starting concurrents *Aevol* jobs with a parametrizable level of parallelism (cores per job and number of parallel jobs), can handle restarting interrupted simulations, and can run *Aevol* post-treatments once the main simulation is finished. The package also contains Jupyter notebooks that give easily approachable and modifiable examples of how to analyse *Aevol* data. Finally, the *aevol-utilities* tools are in use by other members of the team, and could be part of a larger move towards a new high-level Python interface for *Aevol*.

A.2 *EvoTSC*

In the second part of my Ph.D., I developed *EvoTSC*, a Python simulation that implements an individual-level model of the transcription-supercoiling coupling and a population-level *in silico* artificial evolution platform. The software is accessible here: `https://gitlab.inria.fr/tgrohens/evotsc`. *EvoTSC* has been used in several publications. Chapter 4, based on Grohens et al. (2021) and on Grohens et al. (2022b), presents a first version of the transcription-supercoiling coupling model and the results obtained using a corresponding version of *EvoTSC*. Chapter 5 presents a more realistic version of the individual-level model and the associated results, which have been published as a preprint (Grohens et al., 2022a) and will be submitted to peer review. The corresponding versions of the code, and accompanying data analysis notebooks, can be found in the `alife-model` (Grohens et al., 2021) and `alife-journal` (Grohens et al., 2022b) branches of the git repository.

The current version of *EvoTSC* used in Chapters 6 and 7 implements the more detailed genome model presented in Section 5.1, and is available on the `phd` branch of the repository.

A.2.1 Technical Description

EvoTSC is written in Python, and measures around 1,800 lines of Python code. It comes as an installable `pip` package, and is licensed under a 3-clause BSD license. The `evotsc.py`

file contains the main classes (`Gene`, `Individual` and `Population`) that are used in the simulations. The `core.py` files contains the computationally heavy parts of the code, which are accelerated using `numba` (Lam et al., 2015). The `run.py` contains the code responsible for initializing a simulation, and data input and output. The `lib.py` contains miscellaneous functions used throughout the notebooks. Finally, the `plot.py` file contains the code responsible for plotting, using the `matplotlib` library (Hunter, 2007). This file contains in particular the `plot_genome_and_tsc` function used to create all the circular genome plots in the manuscript.

A.2.2 Tooling

In order to analyze the results of *EvoTSC* simulations more easily, I developed an extensive set of notebooks for data analysis and visualisation. The notebooks are available in a separate repository: `https://gitlab.inria.fr/tgrohens/evotsc-notebooks`, and are also licensed under the BSD 3-clause. In particular, these notebooks contain the code responsible for the computation of gene expression as a function of supercoiling (such as in Figure 5.5), for the generation of contiguous gene subnetworks (Figure 5.8), and for the gene knock-out analysis (Figure 5.9, 5.10 and 5.10).

A.2.3 Use

I have up until now been the only user of *EvoTSC*. However, the code has been written to be readily reusable and extendable. It is in particular documented with a `README.md` file, and extensive comments throughout the code.

Appendix B

Covid-19 Task Force

At the start of the Covid-19 pandemic (March-April 2020), I took part in an Inria task force that was created to provide scientific expertise to external stakeholders. In particular, we worked with the Assistance Publique-Hôpitaux de Paris (AP-HP), in order to help the AP-HP better understand and model the progression of the Covid-19 epidemic in the four central departments of the Paris metropolitan area. We proposed a model that used emergency call regulation data, such as the total number of calls, the number of calls resulting in the dispatch of an emergency vehicle, the number of patients hospitalized with Covid-19, and the number of patients in Intensive Care Units (ICUs), as input data to estimate the progression of the epidemic. Using this model, we were able to show that there were strong discrepancies between the different departments, and that it was possible to predict the evolution of the number of cases from the emergency call regulation data.

The rest of this appendix comprises the journal article (Gaubert et al., 2020), published in the *Comptes-Rendus Mathématique* of the French Academy of Sciences, that describes this work in detail.

Understanding and monitoring the evolution of the Covid-19 epidemic from medical emergency calls: the example of the Paris area

Stéphane Gaubert[1,2], Marianne Akian[1,2], Xavier Allamigeon[1,2], Marin Boyet[1,2]
Baptiste Colin[1,2], Théotime Grohens[1,3], Laurent Massoulié[1,4,5], David P. Parsons[1]
Frédéric Adnet[6,7], Érick Chanzy[6], Laurent Goix[6], Frédéric Lapostolle[6,7]
Éric Lecarpentier[6], Christophe Leroy[6], Thomas Loeb[6], Jean-Sébastien Marx[6]
Caroline Télion[6], Laurent Tréluyer[6] and Pierre Carli[6,8]

[1] INRIA
[2] CMAP, École polytechnique, IP Paris, CNRS
[3] Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205
[4] ENS, CNRS, PSL University
[5] Microsoft Research-INRIA Joint Centre
[6] AP-HP
[7] Université Paris XIII, Bobigny
[8] Université Paris-Descartes, Paris

Emails: [1]: Prenom.Nom@inria.fr [6]: Prenom.Nom@aphp.fr

June 10, 2020

Abstract. We portray the evolution of the Covid-19 epidemic during the crisis of March-April 2020 in the Paris area, by analyzing the medical emergency calls received by the EMS of the four central departments of this area (Centre 15 of SAMU 75, 92, 93 and 94). Our study reveals strong dissimilarities between these departments. We show that the logarithm of each epidemic observable can be approximated by a piecewise linear function of time. This allows us to distinguish the different phases of the epidemic, and to identify the delay between sanitary measures and their influence on the load of EMS. This also leads to an algorithm, allowing one to detect epidemic resurgences. We rely on a transport PDE epidemiological model, and we use methods from Perron-Frobenius theory and tropical geometry.

Comprendre et surveiller l'évolution de l'épidémie de Covid-19 à partir des appels au numéro 15: l'exemple de l'agglomération parisienne

Résumé. Nous décrivons l'évolution de l'épidémie de Covid-19 dans l'agglomération parisienne, pendant la crise de Mars-Avril 2020, en analysant les appels d'urgence au numéro 15 traités par les SAMU des quatre départements centraux de l'agglomération (75, 92, 93 et 94). Notre étude révèle de fortes disparités entres ces départements. Nous montrons que le logarithme de toute observable épidémique peut être approché par

une fonction du temps linéaire par morceaux. Cela nous permet d'identifier les différentes phases d'évolution de l'épidémie, et aussi d'évaluer le délai entre la prise de mesures sanitaires et leur effet sur la sollicitation de l'aide médicale urgente. Nous en déduisons un algorithme permettant de détecter une resurgence éventuelle de l'épidémie. Notre approche s'appuie sur un modèle d'EDP de transport de l'évolution épidémique, ainsi que sur des méthodes de théorie de Perron-Frobenius et de géométrie tropicale.

1 Introduction

The outbreak of Covid-19 in France has put the national Emergency Medical System (EMS), the *SAMU*, in the front line. In the *Île-de-France* region, one most affected by the epidemic, the SAMU centers of Paris and its inner suburbs experienced a major increase in the number of calls received and of the number of ambulance dispatches for Covid-19 patients.

We show that indicators based on EMS calls and vehicle dispatches allow to analyze the evolution of the epidemic. In particular, we show that EMS calls are early signals, allowing one to anticipate vehicle dispatch. We provide a method of short term prediction of the evolution of the epidemic, based on mathematical modeling. This leads to *early detection and early alarm mechanisms* allowing one either to confirm that certain sanitary measures are strong enough to contain the epidemic, or to detect its resurgence. These mechanisms rely on simple data generally available in EMS: numbers of patient records tagged as Covid-19, and among these, numbers of records resulting in medical advice, ambulance dispatch, or Mobile Intensive Care Unit dispatch. We also provide a comparative description of the evolution of the epidemic in the four central departments of the Paris area, showing spatial dissimilarities, including a strong variation of the doubling time, depending on the department.

Our approach relies on several mathematical tools in an essential way. Indeed, the Covid-19 epidemic has unprecedented characteristics, and, given the lack of experience of similar epidemics, one needs to rely on mathematical models. We use transport PDE to represent the dynamics of Covid-19 epidemic. Transport PDE capture epidemics with a significant time interval between contamination and the start of the infectious phase (in contrast, ODE models without time delays allow instantaneous transitions from contamination to the infectious phase). In the early stage of the epidemic, in which the majority of the population is susceptible, this dynamics becomes approximately linear and order preserving. Then, it can be analyzed by methods of Perron–Frobenius theory. Our main theoretical result shows that the logarithm of epidemic observables can be approximated by a piecewise linear map, with as many pieces as there are phases of the epidemic (i.e., periods with different contamination conditions), see Theorem 1. This methods allows us to identify, the phases of the epidemic evolution, and also to evaluate the time interval between sanitary measures and their impact on epidemic observables, like vehicle dispatch. The idea of piecewise linear approximation and of "log glasses", a key ingredient of the present approach, arises from tropical geometry.

The present work started on March 13^{th}, and led to the algorithm presented here. A preliminary version of this algorithm was used, on March 20^{th}, to forecast the epidemic wave, anticipating that the peak load of SAMU (which occurred around March 27^{th}) would be different depending on the department of the Paris area. We subsequently applied our method to provide Assistance Publique – Hôpitaux de Paris (AP-HP), on April 5^{th}, with an early report, quantifying the efficiency of the lockdown measures from the estimation of the contraction rate of the epidemic in the different departments. This algorithm is now deployed operationally in the four SAMU of AP-HP. This work may be quickly reproduced in any EMS.

Although it was developed for Covid-19 and for EMS calls, the present monitoring method is generic. It may also apply to other medical indicators, see Section 3.4, and to other epidemics, for instance, influenza.

This paper is a crisis report, giving a unified picture of a work done jointly by a team of physicians of the SAMU of AP-HP and applied mathematicians from INRIA and École polytechnique. Medical, epidemiological, and mathematical aspects are intricated in this work. We received help from several physicians, researchers and engineers, not listed as authors, and also help from several organizations. They are thanked in the acknowledgments section.

This paper should be understood as an announce. The results will be subsequently developed in several papers, with different subsets of coauthors. It is intended to be read both by a medical and a mathematical

audience. The first part of the paper, up to Section 6 included, and the conclusion, are intended to a broad audience. Mathematical tools are presented in Section 7, Section 8, Section 9 and in the appendix.

The present work shows the epidemiological significance of the calls received by the EMS, it focuses on the mathematical modeling aspects, on the description of the evolution of the epidemy in the Paris area, and on prediction algorithms. The current work[1] with an intersecting set of authors, is coordinated with the present one. It focuses on medical aspects. It makes a case study of the Covid-19 crisis of March-April 2020, in Paris, considering the EMS and the hospital services in a unified perspective. It shows that the calls received by SAMU are early predictors of the future load on ICU.

2 Context

The mission of the SAMU centers is to provide an appropriate response to calls to the number 15, the French toll-free phone number dedicated to medical emergencies. This service is based on the medical regulation of emergency calls, in the sense that for each patient, a physician decides which response is most appropriate. Thus, depending on the evaluation over the phone of the severity of the case and the circumstances, the response may be a medical advice, a home visit by a general practitioner, the dispatch of a team of EMTs (Emergency Medical Technicians) of either a first aid association or the Fire brigade, or an ambulance of a private company. A Mobile Intensive Care Unit (MICU), staffed by a physician, a nurse and an EMT, is sent to the scene as a second or a first tier, when a life threatening problem is suspected. The role of the SAMU in the management of disasters or mass casualties has been described elsewhere [18, 4]. The city of Paris and its inner suburbs are covered by 4 departmental SAMU Center-15 : Paris (75), Hauts-de-Seine (92), Seine Saint-Denis (93), and Val-de-Marne (94), see the map on Figure 3. They serve a population of 6.77 million inhabitants. These four Center-15 are part of the public hospital administration, AP-HP (Assistance Publique – Hôpitaux de Paris). They operate identically and use the same computerized call management system. Since the outbreak of the Covid-19 epidemic, the French government instructed the public that anyone with signs of respiratory infection or fever should not go directly to the hospital emergency room to limit overcrowding, but should call number 15 for orientation. To comply with the recommendations of the health care authorities, the four Center-15 applied the same procedures: after medical call regulation, only patients with signs of severity or significant risk factors were transported by EMTs and ambulances to hospitals, either to Emergency Room (ER) or newly created Covid-19 Units. The cases presenting a life-threatening emergency, mostly respiratory distress, were managed by an MICU team and then admitted directly in Intensive Care Unit (ICU). All other cases were advised to stay at home and isolate themselves. When necessary, these patients were also eligible for a home visit by a general practitioner or a consultation appointment the following days.

In order to maintain a rapid response when a major increase in the number of calls was observed, the four Center-15 implemented specific procedures. Switchboard operators and medical staff was reinforced, and for calls related to Covid-19 an interactive voice server —triaging the calls to dedicated computer stations— was developed. Patient evaluation and management were improved by introducing video consultation, sending of instruction using SMS, giving the patient the option to be called back. Prehospital EMT teams were also significantly reinforced by first aid volunteers, and additional MICU were created. Since January 20$^{\text{th}}$ 2020 all calls and patient records related to Covid-19 were flagged in the information system of Center-15 and a daily automated activity report was produced.

3 Methods

In this section, we describe the methods used in this work, in a way adapted to a general audience. Mathematical developments appear in Sections 7 to 9 and in the appendix.

[1]COVID19 APHP-Universities-INRIA-INSERM, Emergency calls are early indicators of ICU bed requirement during the COVID-19 epidemic, medRxiv:2020.06.02.20117499, June 2020.

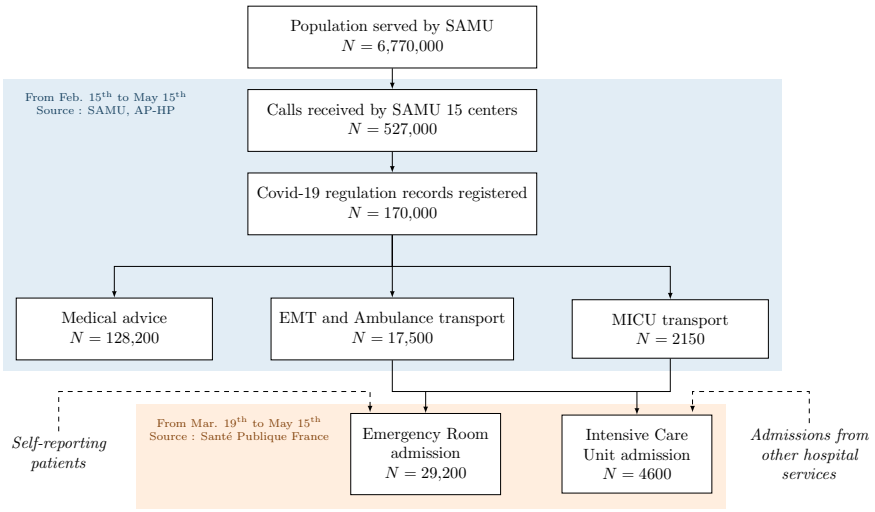


Figure 1: Flowchart: from calls to Center 15 to admission in hospital units. The numbers are summed over the departments 75, 92, 93 and 94 of the Paris area.

3.1 Classification of calls

In order to develop a mathematical analysis of the evolution of the epidemic, we classified the calls tagged as Covid-19 in three categories, according to the decision taken:

Class 1: calls resulting in the dispatch of a Mobile Intensive Care Unit;

Class 2: calls resulting in the dispatch of an ambulance staffed with EMT;

Class 3: calls resulting in no dispatch decision. Such calls correspond to different forms of medical advice (recommendation to consult a GP, specific instructions to the patient, etc.).

We shall denote by $Y_{\mathrm{MICU}}(t)$ (resp. $Y_{\mathrm{EMT}}(t)$ and $Y_{\mathrm{adv}}(t)$) the number of MICU transports (resp. the number of ambulances transport and the number of medical advices) on day t, for patients tagged with suspicion of Covid-19. We shall call these functions of time the *observables*, in contrast with $C(t)$, the actual number of new contaminations on day t, which cannot be measured. We developed a piece of software that computes these observables by analyzing the medical decisions associated with the patient records, made accessible daily by AP-HP.

3.2 Mathematical properties of the observables

To analyze these observables, we shall rely on a mathematical model.

A standard approach represents the evolution of an epidemic by an ordinary differential equation (SEIR ODE), representing the evolution of the population in four compartments: "susceptible" (S), "exposed" but not yet infectious (E), "infectious" (I), and finally, "removed" from the contamination chain (R), either by recovery or death. A refinement of the SEIR model splits the S and E compartments in sub-compartments corresponding to different age classes. It includes a contact matrix, providing differentiated age-dependent infectiosity rates [27]. Another refinement includes additional compartments, representing, for instance, patients at hospital [13].

In contrast with such ODE models, we use a partial differential equation (PDE), i.e., an infinite dimensional dynamical system, described in Section 7. Our approach is inspired by the PDE model of Kermack

and McKendrick [22]. We use PDE, rather than ODE, to take into account the presence of *delays* in the contamination process: the median incubation period of Covid-19 is estimated of 5.1 days, with a 95% confidence interval of 4.5-5.8 and an heavy tail [25], in line with other human coronaviruses having also long incubation times, like SARS [35] and MERS [34]. (This may be compared with a median incubation time of 1.4 days [95% CI, 1.3–1.6] for the toxigenic Cholera [2], or with an interval of 36 hours between infection by pneumonic plague and first symptoms in Brown Norway rats, with rapid letality, 2-4 days after infection [1].) ODE models assume exponentially distributed transitions times from one compartment from another. This entails that the interval elapsed between contamination and the time an individual becomes infectious can be arbitrarily small, so ODE models are more adapted to epidemics with a short incubation time. Using PDE, as is done in Section 7, takes delays into account.

We shall limit here our analysis to the early stage of the epidemic, assuming that the population that has been infected is much smaller than the susceptible population. This approximation is reasonable at least in the initial part of the epidemic, according to the study [32] which gives an estimate of 5.7% for the proportion of the population in France that has been infected prior to May 11^{th}, 2020. Then, the dynamics becomes linear and order-preserving. The latter property entails that the observables are an increasing function of the size of the initial population that is either exposed or infected.

Results of Perron–Frobenius and of Krein-Rutman theory, which we recall in Section 7, entail that, *if the sanitary measures stay unchanged*, there is a rate λ, such that the number of newly contaminated individuals at day t grows as $C(t) \simeq K_C \exp(\lambda t)$, as $t \to \infty$, where K_C is a positive constant.

The number $\delta := (\log 2)/\lambda$, when it is positive, represents the *doubling time*: every δ days, the number of new contaminations per day doubles. When δ is negative, the epidemic is in a phase of exponential decay. Then, the opposite of δ yields the time after which the number of new contaminations per day is cut by half. For the analysis which follows, it is essential to consider, instead of $C(t)$, its logarithm, $\log C(t) \simeq \log K_C + \lambda t$. The exponential growth or decay of $C(t)$ corresponds to a linear growth or decay of the logarithm.

We shall also see in Section 7 that all the epidemiological observables evolve with the same rate. E.g., assuming that all the patients transported by MICU were contaminated τ_{MICU} days before the transport, and that a proportion π_{MICU} of the contaminated individuals will require MICU transport, we arrive at $Y_{\text{MICU}}(t) = \pi_{\text{MICU}} C(t - \tau_{\text{MICU}})$, and so $\log Y_{\text{MICU}}(t) \simeq \log \pi_{\text{MICU}} + \log K_C + \lambda(t - \tau_{\text{MICU}})$. Similar formulæ apply to Y_{EMT} and Y_{adv}, and to other observables based for instance on ICU admissions or deceases. A finer model of observables, taking into account a distribution of times τ_{MICU}, instead of a single value, is presented in Section 7.3.

For the analysis which follows, we shall keep in mind that *the logarithm of all the observables is asymptotically linear as* $t \to \infty$, and that *the rate,* λ*, is independent of the observable.*

3.3 Piecewise linear approximation of the logarithm of the observables

When the sanitary measures change, for instance, when lockdown is established, the rate λ changes. So, the logarithm of the observables cannot be approximated any more by a linear function. However, a general result, stated as Theorem 1 below, shows that this logarithm can be approximated by a *piecewise linear function* with as many linear pieces as there are phases of sanitary policy. This result stems from the order preserving and linear character of the epidemiological dynamics, and so, it holds for a broad class of epidemiological models; several examples of such models are discussed in Section 7.

In the Paris area, there are three relevant sanitary phases to consider from February to May, 2020: initial growth (no restrictions); "stade 2" (stage 2) starting on Feb. 29^{th} (prevention measures), and then lockdown from March 17^{th} to May 11^{th}. Sanitary phases are further described in Section 5.1.

Since the number ν of sanitary phases is known (here $\nu = 3$), we can infer the different values of λ attached to each of these phases, by computing the best piecewise linear approximation, $\mathcal{L}(t)$ with at most ν pieces of the logarithm of an observable $Y(t)$. To compute a robust approximation, we minimize the ℓ_1 norm, $\sum_t |\mathcal{L}(t) - \log Y(t)|$, where the sum is taken over the days t in which the data are available. Finding the best approximation $\mathcal{L}$ is a difficult optimization problem, for the objective function is both non-smooth and non-convex. Methods to solve this problem are discussed in Appendix A.

3.4 Epidemic alarms based on doubling times

To construct epidemic alarms, we shall compute a linear fit, $\mathcal{L}(t) = \alpha + \beta t$, to the variables $\log Y(t)$, where Y is an epidemic observable. The principle is to trigger an alarm when the doubling time becomes positive, or equivalently, when the slope β becomes positive.

Assuming that values of $Y(t)$ are known over a temporal window, there are simple ready-to-use methods for computing estimates $\hat{\beta}$ for the slope β. We can also determine the probability p^+ that the slope is positive. These methods are detailed in Section 9. On their basis, we propose the following **alarm raising mechanism, allowing one to deploy a gradual response**.

This mechanism relies on the two following observables, Y_{adv}, the number of calls resulting in medical advice, and $Y_{\text{disp}} := Y_{\text{EMT}} + Y_{\text{MICU}}$, the number of dispatched vehicles. The consolidation of the observables Y_{EMT} and Y_{MICU} is justified, because the two time series both correspond to the stage of aggravation, albeit with different degrees, and so they evolve more or less at the same time.

First define a temporal window of days t over which the linear fit $\mathcal{L}_{\text{adv}}(t) = \alpha_{\text{adv}} + \beta_{\text{adv}} t$ to $\log Y_{\text{adv}}(t)$ is made. By default we consider the last ten days prior to the current day. Similarly, we compute a linear fit $\mathcal{L}_{\text{disp}}(t) = \alpha_{\text{disp}} + \beta_{\text{disp}} t$ to $\log Y_{\text{disp}}(t)$ over the same time window.

Our algorithm will generate both a *warning* and *alarms*. A warning is a mere incentive to be careful. An unjustified warning is bothersome but generally harmless, so we accept a high probability of false positive for warnings. An alarm may imply some actions, so we wish to avoid false alarms. For this reason, we shall consider two different probability thresholds, ϑ_{alarm} and ϑ_{warn}, say $\vartheta_{\text{alarm}} = 75\%$ and $\vartheta_{\text{warn}} = 25\%$. With this setting, we will be warned as soon as the probability of the undesirable event is $\geqslant 25\%$, and we will be alarmed when the same probability becomes $\geqslant 75\%$. Of course, these thresholds can be changed, depending on the risk level deemed to be acceptable. We shall denote by p^+_{adv} the probability that the slope β_{adv} is positive, and by p^+_{disp} the probability that β_{disp} is positive. These probabilities are evaluated on the basis of statistical assumptions detailed in Section 9.

1. A **warning** is provided when $p^+_{\text{adv}} \geqslant \vartheta_{\text{warn}}$, meaning that the probability that the slope β_{adv} of the curve of the logarithm of the *calls for medical advice* over the corresponding time window be positive is at least ϑ_{warn}. This should be interpreted as a mere warning of epidemic risk: choosing ϑ_{warn} as above, the odds are at least 25% that the epidemic is growing.

2. This warning is subsequently transformed into an **alarm** when $p^+_{\text{adv}} \geqslant \vartheta_{\text{alarm}}$. Choosing ϑ_{alarm} as above, the odds that the epidemic is growing are now at least 75%.

3. Such an alarm is then subsequently transformed into a **confirmed alarm** if we still have $p^+_{\text{adv}} \geqslant \vartheta_{\text{alarm}}$, and if, in addition, $p^+_{\text{disp}} \geqslant \vartheta_{\text{alarm}}$, meaning that the probability that the slope of the logarithm of the curve of ambulances and MICU dispatches be positive is now above ϑ_{alarm}. Again, this estimate is defined in terms of a time window over which β_{disp} is estimated. We use the same default values of ten days and ϑ_{alarm} as above.

As shown in Section 4, the indicators based on vehicle dispatch are by far less noisy than the indicators based on calls for medical advices, but their evolution is delayed. This is the rationale for using medical advice for an early warning and early alarm, and then vehicle dispatch for confirmation.

Instead of considering the probability p^+, we could consider the upper and lower bounds of a confidence interval $[\beta^-_\epsilon, \beta^+_\epsilon]$ for the estimated slope β, with a probability threshold ϵ. Then we may, trigger a warning when $\beta^+_\epsilon \geqslant 0$, and an alarm when $\beta^-_\epsilon \geqslant 0$. This leads to an essentially equivalent mechanism. We prefer the algorithm above as it allows to interpret the thresholds in terms of false positives and false negatives.

Given the severity of the risk implied by Covid-19, it may be desirable to complete the previous alarm, based only on tail probabilities of the slope, by a different type of alarm, based on a threshold of doubling time, D. The alarm will be triggered if the odds that the doubling time be positive and smaller than D are at least one half. An indicative value of D might be 14 days: a doubling of the number of arrivals of Covid-19 patients in hospital services every 14 days may be quite challenging, justifying an alarm, and the slope corresponding to this doubling time seems significant enough to avoid false alarms. Again, the value of

D can be changed arbitrarily depending on the acceptable level of risk. Moreover, this other type of alarm can still be implemented in two stages: early alarm, with the medical advice signal, and then confirmed alarm, with the vehicle dispatch signal.

In addition, Section 9 provides more sophisticated ready-to-use methods for obtaining sharper confidence intervals or probabilities for the slope β, resulting in more precise alarm mechanisms, when different time series are available. We require, however, that these series correspond to events occurring approximately at the same stage in the pathology unfolding. Here, we used the trivial aggregator, $Y_{\text{disp}} = Y_{\text{EMT}} + Y_{\text{MICU}}$. There is an optimal way to mix different series to minimize the variance of the composite estimator, explained in Section 9.

This methodology is generic. It could thus also apply to obtain a sharper confidence interval for the early indicator by combining its estimate $\hat{\beta}_{\text{adv}}$ with that of other time series associated with signals that correspond to the same stage in pathology unfolding. Specifically, the count $Y_{\text{GP}}(t)$ of patients consulting general practitioners for recently developed Covid-19 symptoms, if available, provides such a signal. A linear fit to $\log Y_{\text{GP}}(t)$ would then yield an estimate $\hat{\beta}_{\text{GP}}$ which can be combined with $\hat{\beta}_{\text{adv}}$ to refine the corresponding confidence interval. In this way, we can mix several early but noisy indicators to get an early but less noisy consolidated indicator.

4 Results – data analysis

4.1 Key figures and graphs

From February 15$^{\text{th}}$ to May 15$^{\text{th}}$, we counted a total of 170,166 patient files tagged with a suspicion of Covid-19, distributed as follows in the different departments: 53,646 in Dep. 75; 36,721 in Dep. 92; 49,703 in Dep. 93; and 30,096 in Dep. 94.

The flow of calls to the SAMU of the Paris area, and its impact on ER and ICU, is shown on Figure 1. The data concerning the ER and the ICU are taken from the governmental website SPF (Santé Publique France) [15], it is available only from March 19$^{\text{th}}$.

On Figure 2, we represent, in logarithmic ordinates, the numbers of events of different types, summed over the four departments of the Paris area (75, 92, 93 and 94): (i) the number of patients calling the SAMU (including patients not calling for Covid-19 suspicion); (ii) the number of calls tagged as Covid-19 not resulting in a vehicle dispatch (i.e., as discussed in §3.1, all kinds of medical advices); (iii) the number of calls tagged as Covid-19 resulting in an ambulance or MICU dispatch,

We obtained the data (i) by analyzing the phone operator log files. Since a patient may call the Center 15 several times, we eliminated multiple calls to count unique patients. To compute data (ii) and (iii), we developed a software to analyze the "medical decision" field of the regulation records.

Using logarithmic ordinates is essential on Figure 2, as it allows to visualize on the same graph signals of different orders of magnitude (e.g, there is a ratio of 20 between the peak number of patients calling and the peak number of vehicles dispatched).

The evolution of the number of vehicles dispatched (MICU and ambulances) is shown on Figure 3, for each department of the Paris area (still with logarithmic ordinates).

We provide in Table 1 the doubling times of the number of vehicles dispatched (ambulances and MICU), for the different departments, measured in days (abbreviation "d").

We now draw several conclusions from the previous analysis.

4.2 The increase in the number of calls for medical advice provides an early, but noisy, indicator of the epidemic growth

As shown in Figure 2, the peak of the number of calls for medical advice was on March 13$^{\text{th}}$. However, this date, four days before the lockdown (March 17$^{\text{th}}$), is not consistent with epidemiological modeling. This peak seems rather to be caused by announcements to the population, see the discussion in Section 6.1.

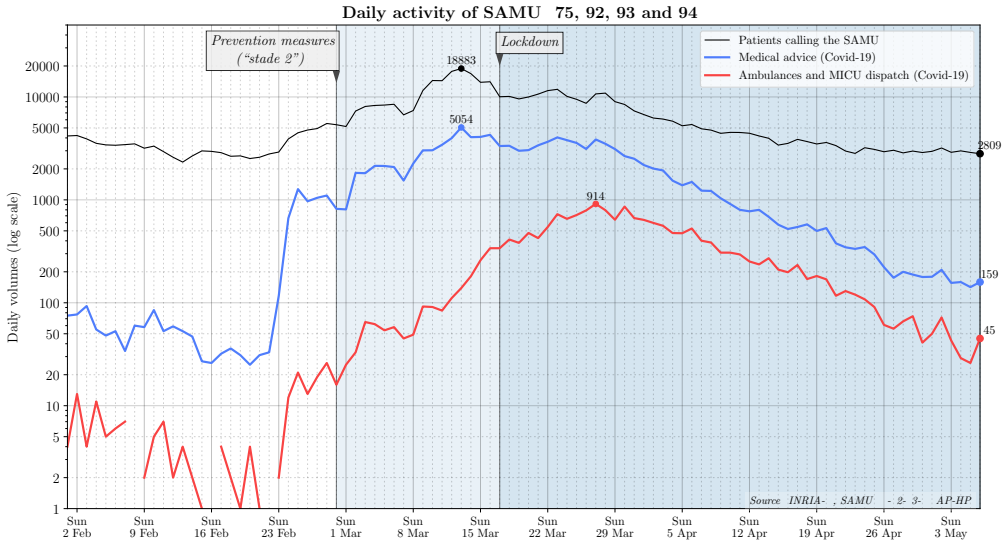


Figure 2: Number of patients calling Center-15, of MICU and ambulances dispatch for Covid-19 suspicion in the Paris area (departments 75, 92, 93 and 94)

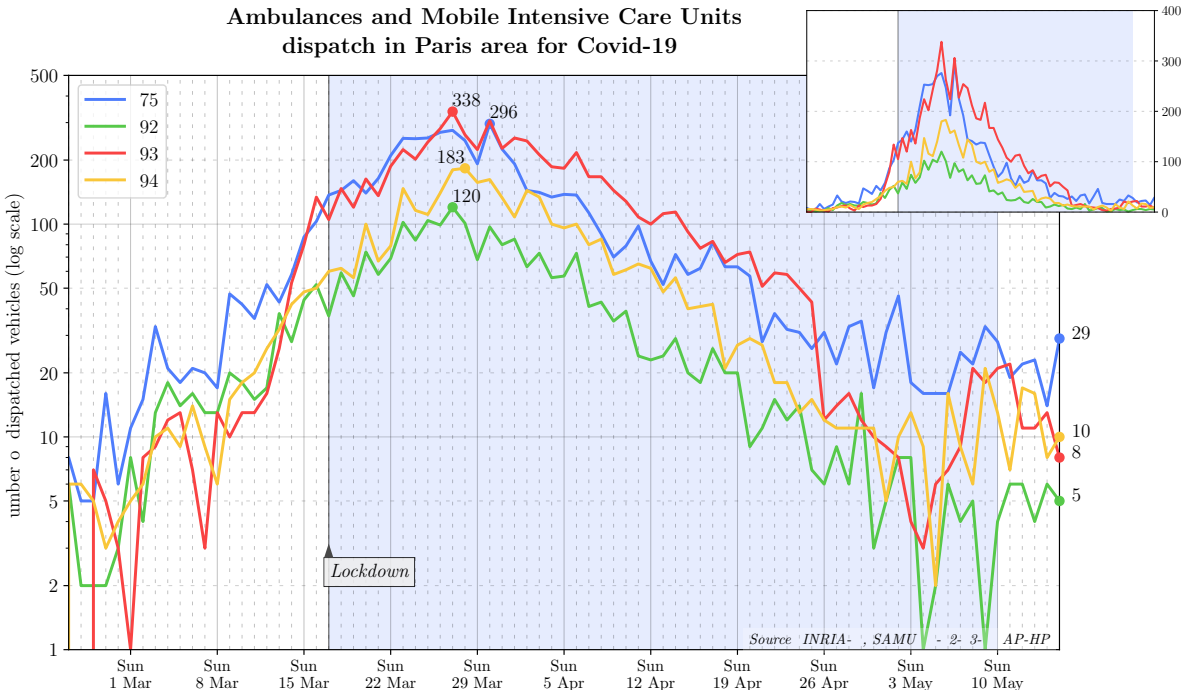
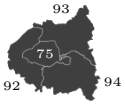


Figure 3: Comparison of the evolution of the epidemic in the different departments of the Paris area: numbers of vehicles dispatch by department. The figure inset displays the same curves in usual linear ordinates to keep in mind the different magnitudes at stake. A map of the Paris area, showing the departments 75, 92, 93, 94, is at the bottom right of the figure.



	Feb. 28$^{\text{th}}$– Mar. 15$^{\text{th}}$	Mar. 15$^{\text{th}}$– Mar. 29$^{\text{th}}$	Mar. 29$^{\text{th}}$– April 24$^{\text{th}}$
75	5.9 d	9.8 d	-9.4 d
92	4.9 d	10.6 d	-8.3 d
93	4.2 d	8.5 d	-10.2 d
94	4.6 d	6.9 d	-7.7 d

Table 1: Doubling time of the number of MICU and ambulances dispatched, for different periods, for each department, obtained by a least squares approximation of the logarithm of this number. The opposite of a negative doubling time yields the halving time.

4.3 The epidemic kinetics vary strongly across neighboring departments

In the initial phase of the epidemic (Feb. 28$^{\text{th}}$–March 15$^{\text{th}}$), the doubling time was significantly shorter in the 93 department (4.2 d) than in central Paris (5.9 d). The 93 department, with 1.6M inhabitants, is less populated than central Paris (2.1M inhabitants). Another difference between the departments concerns mobility. Movement from the population from central Paris to smaller towns and cities or to countryside were observed, after March 12$^{\text{th}}$, the date of the first presidential address concerning the Covid-19 crisis.

In order to quantify this mobility, we requested information from Enedis, the company in charge of the electricity distribution network in France, and also from Orange and SFR, two operators of mobile phone networks.

Enedis provided us with an estimation of the departure rates of households, based on a variation of the volume of electricity consumed, aggregated at the level of departments and districts (i.e., *arrondissements*).

SFR provided us with estimates of daily flows from the Paris area to other regions, again aggregated at the scale of the departments or districts, based on mobile phone activity, confirming this decrease of population.

Orange Flux Vision provided us with daily population estimates, at the scale of department, based on mobile phone activity. By March 30$^{\text{th}}$, the population, during the night, was estimated to be 1.6M inhabitants in central Paris, versus 1.35M in the 93.

However, the epidemic peak was higher in the 93 than in the 75 (338 dispatches versus 296). The contraction rate in the period after the peak (March 29$^{\text{th}}$–Apr. 24$^{\text{th}}$) was also smaller in the 93, with a halving time of 10.2 days, to be compared with 9.4 days in the 75. Possible explanations for these strong spatial discrepancies are discussed in Section 6.5.

5 Results – mathematical modeling

5.1 Delay between implementation of sanitary policies and its effect on hospital admissions

We explained in Section 3.3, based on Theorem 1 below, that the logarithm of an epidemic observable can be approached by a piecewise linear map with as many pieces as there are stages of sanitary measures.

So, we look for the best approximation, in the ℓ_1 norm, of the logarithm of the number of vehicles dispatched (ambulances and MICU), by a piecewise linear map with at most three pieces. This best approximation is shown on Figure 4. It is computed by the method of Appendix A.

In order to evaluate the influence of a sanitary measure on the growth of the epidemic, an approach is to compare the date of the measure with the date of the change of slope of the logarithmic curve, consecutive to the measure. This method is expected to be more robust than, for instance, a comparison of peak values, because the best piecewise-linear approximation is obtained by an optimization procedure *taking the whole sequence into account*. Indeed, a local corruption of data will not change significantly the date of change of slope, if the problem is *well conditioned*. This is the case in particular if the difference between consecutive

slopes is sufficiently important. In other words, we can identify in a more robust manner the time of effect of a strong measure than of a mild one.

Let us recall the main changes of sanitary measures in the Paris area, between February and May 2020. We may distinguish the following phases:

- *Initial development of the epidemic*, no general sanitary measures in the Paris area, until Feb 29^{th}, first day of so-called "stade 2" by the authorities (following "stade 1" in which measures intended to prevent the introduction of the virus in France – like quarantine in specific cases– were taken).
- *"Stade 2" (stage 2) measures*: general instructions of social distancing given to the population (e.g., not shaking hands), ban on large gatherings. Moreover, some large companies created crisis committees, and decided to take more restrictive measures than the ones required by the authorities, including for instance banning meetings with more of 10 people, and banning business travels. Restrictive measures in companies were deployed gradually during the work week from March 2^{nd} to March 6^{th}.
- *School closure* on March 16^{th}.
- *Lockdown* on March 17^{th}. The lockdown ended on May 11^{th}, throughout the country.

Hence, we may interpret the variations in the slope in the piecewise linear approximation of the logarithm of the number of ambulances and MICU dispatched, shown on Figure 4, as the effect of sanitary measures. The dates where the slope changes are represented in the figure by dotted lines. Thus, the latest breakpoint of the piecewise linear approximation of the 75 curve (in blue) arises on March 26^{th}, to be compared with March 30^{th} in the 93 (red curve). The dates of breakpoints in the 92 and 94 are intermediate. Given the first strong measure (closing of schools) was taken on March 16^{th}, we may evaluate the delay between a sanitary measure and its effect on the ambulances and MICU dispatch to be between 10 and 14 days. This corresponds to a delay between contamination and occurrence of severe symptoms.

5.2 Construction of statistical indicators of epidemic resurgence based on emergency calls

We implemented the alarm mechanism based on the inference of doubling times described in Section 3.4 and further explained in Section 9. The method is illustrated on Figure 5. Given a time period where the data is known, we perform a linear regression on the number of medical advices and vehicles dispatched.

The light shaded, tubular areas around the curves are based on confidence intervals for the fluctuations of the observed log-counts. The dark-shaded, trapezoidal areas prolongate the tubular areas with straight lines, the slopes of which correspond to confidence intervals for the slope of the linear regression. We display such confidence domains for the last known data in May, performing a 6-day forecast based on the last ten days. In order to validate the method, we also display these domains for older data in March and April, performing for each a 6-day forecast based on a number of past days. For these time-periods, the short-term confidence domains are seen to satisfactorily contain the data of the following days. Observe how the shape of the confidence trapezoids depends on the number of points and the variance of the data used to compute them.

The needle-shaped (or clock hand) indicators depicted in each trapezoidal confidence domain illustrate the alarm mechanism of Section 3.4. For each slope inference, there is a $\vartheta_{\text{warn}} = 25\%$ probability that the real-value of the slope we estimated using recent data is greater than the slope of the thin needle on the picture. Likewise there is a $\vartheta_{\text{alarm}} = 75\%$ probability that the real slope is greater than the slope of the fat hand. As a result, a warning (resp. an alarm) on the dynamics of the medical advice curve should be triggered as soon as the thin needle (resp. the fat needle) has a positive angle with respect to the horizontal. We have depicted the horizontal with dashed lines to enhance readability.

On May 25^{nd}, no warning nor alarm is triggered, since all the needle-indicators are below the positivity threshold, indicating with at least a 75% confidence level that based on the last ten days, the two observed signals are on a decreasing trend. Note that due to the relative stagnation of the medical advice curve in the

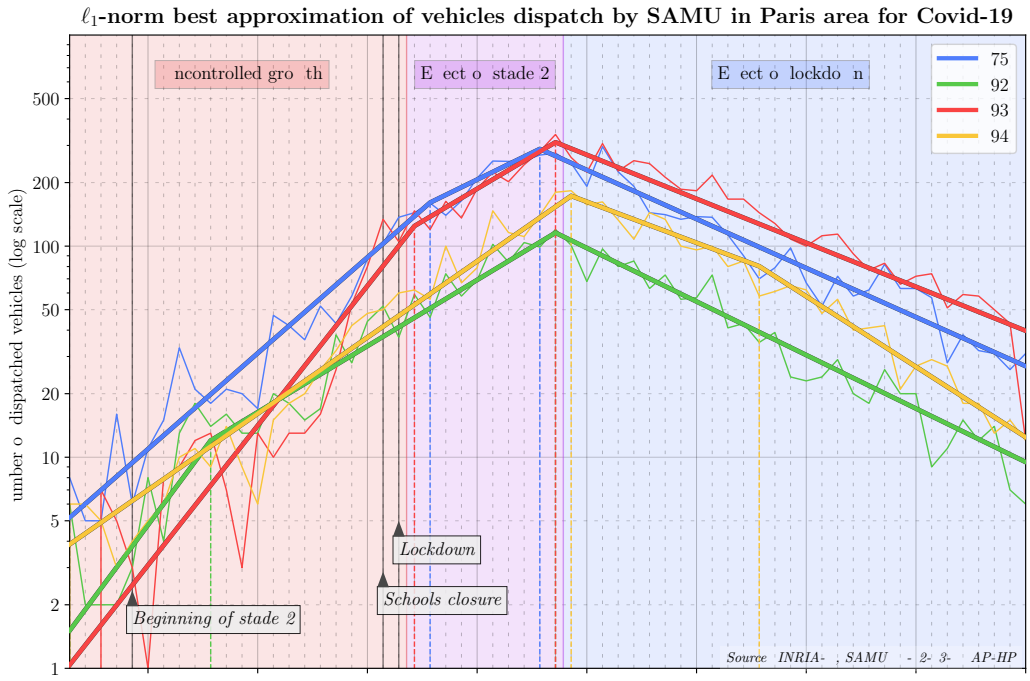


Figure 4: Logarithm of the number of ambulances dispatched: the effect of the successive sanitary measures

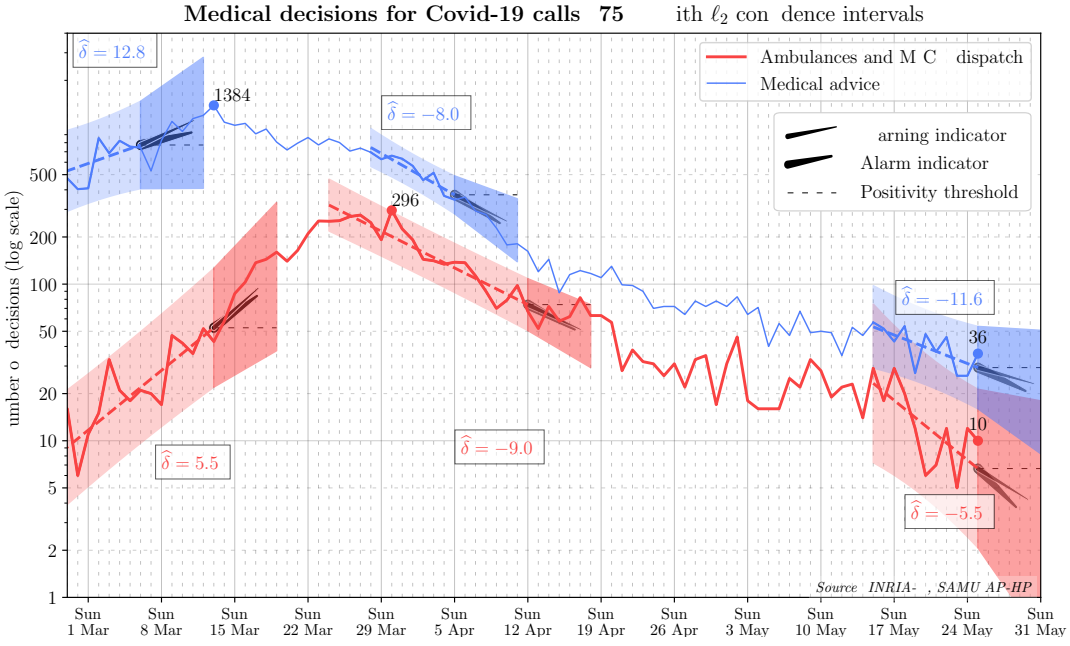


Figure 5: Short term predictor for SAMU 75, with confidence regions and warning and alarms indicators

two first weeks of May, computing our indicators a few days earlier (such as May 22$^{\text{nd}}$) would have raised an warning to arouse vigilance due to the uncertainty on the future trend, but no alarm.

The numbers ϑ_{warn} and ϑ_{alarm} need to be carefully calibrated, and for this, additional data for the forecoming weeks may be helpful.

6 Discussion

6.1 Calls resulting in medical advice are highly influenced by the instructions given to the population

The blue curve on Figures 2 and 5 counts the number of times medical advices was given, of all kinds (calls resulting in recommendations given to the patient but no vehicle dispatched). It is generally associated with early events in the unfolding of the pathology, and in particular occurrence of the first symptoms. An estimate of 5.1 days between the date of contamination and the date of the first symptoms is given in [25], so we may assume that calls for medical advice are made by patients 5-8 days after contamination. We observed in Section 4.2 that the peak in the number of calls resulting in medical advice was on March 13$^{\text{th}}$. Hence, assuming the peak of contaminations was just before lockdown, the peak of the curve of new symptom occurrences should occur *only several days after the lockdown time* (March 17$^{\text{th}}$). This indicates that the curve of calls resulting in a medical advice did not give a reliable picture of the epidemic growth around March 13$^{\text{th}}$. Indeed, this curve is very sensitive to changes in the instructions given to the population and to political announcements, notable examples of which include the following: recommendation to patients to call emergency number 15, instead of going directly to emergency departments (to avoid contamination and overcrowding); – the presidential announcement on March 12$^{\text{th}}$ of more restrictive measures to be deployed from March 16$^{\text{th}}$, making the population more aware of the growth of the epidemic.

6.2 The indicators of medical advice given and ambulances and MICU dispatch can be used to monitor the epidemic

Setting aside perturbations due to political announcements or changes in the policy for calling SAMU, the curve of calls for medical advice should be a reliable and early estimate of the curve of ambulances dispatched, which is triggered at a later stage in the unfolding of the pathology when symptom severity increases. Thus, it gives an early signal allowing both SAMU and hospitals to anticipate by several days an increase in load.

We can give a rough estimate of this delay by considering the peak dates in Figure 2. Epidemiological modeling indicates that the number of new contaminations grows exponentially until a sanitary measure that is strong enough to contain the epidemic is taken. In the present case, the candidates for such strong measures are the school closing (March 16$^{\text{th}}$) or the lockdown (March 17$^{\text{th}}$). Considering the last day previous to the measures, we may assume that the peak date for new contaminations was on March 15$^{\text{th}}$ or March 16$^{\text{th}}$. As mentioned above, according to [25], the time between contamination and first symptoms is estimated to be 5.1 days. Hence, if calls for medical advice were representative of first symptoms, the peak value for these calls would have been between around March 20th or March 21$^{\text{st}}$. The peak of the number of dispatches of ambulances and MICU was on March 27$^{\text{th}}$. This leads to an estimate of 6-7 days for the delay between the curve of the true need for medical advice and the curve of vehicles dispatch.

The alarm mechanism was developed during the crisis, after March 20$^{\text{th}}$. Had it been available before, in the 75, the early warning would have occurred by February 24$^{\text{th}}$ (for both Covid-19 indicators on medical advice and vehicle dispatch), alarm would have been been triggered on February 25$^{\text{th}}$ based on medical advice, and a confirmation alarm would have occurred on February 27$^{\text{th}}$ based on ambulance and MICU dispatch.

Tracking the same signal at a finer spatial resolution (a neighborhood rather than a department) may enable epidemic surveillance during the period following the lifting of measures such as travel bans. Indeed in view of the spatial differentiation in doubling times at the department level, it appears plausible that disparities may also be present at much finer spatial granularities, with resurgences localized to towns or

neighborhoods. Deployment of alarm mechanisms constructed from event counts at finer spatial granularities could be used to identify clusters of resurgence early on, and guide subsequent action.

Seasonal influenza, whose early symptoms can be mistaken for those of Covid-19, can also trigger a linear growth of the logarithms of the number of calls, or vehicle dispatched. However the slope of this logarithmic curve is expected to be shallower, owing to the lower contagiosity of influenza. One could thus distinguish, on the basis of the observed slope, whether one is confronted with seasonal flu or with an outbreak of Covid-19.

6.3 Jumps of the curves of the number of calls may be caused by large clusters or influenced by neighboring countries

The curves of the number of calls for medical advice, and of vehicles dispatched (Figure 2), both jump between February 23^{rd} and 25^{th}. Epidemiogical models are unlikely to produce shocks of this type in the absence of exceptional factors. Several significant epidemic events occurred at nearby dates, including:

1. The development of the Covid-19 epidemic in the north of Italy, a region closely tied to France (the first lockdowns occurred around February 21^{st} in the province of Lodi). The school vacation ended on February 23^{th}in the Paris area. A significant number of parisians went back from Italy during the week-end of February 22^{nd}-23^{th}(end of school vacation).

2. A large Evangelist meeting (Semaine de Carême de l'Eglise La Porte Ouverte Chrétienne de Bourtzwiller, Haut-Rhin), from February 17^{th} to February 21^{st}, identified by Agence Régionale de Santé Grand Est as the source of a cluster phenomenon [12].

The potential influence of the north Italy epidemic was pointed out by Paul-Georges Reuter (private communication). At this stage, the essential factors are not yet known. The influence of mobility on the development of the epidemic in the Paris area will be studied in a further work.

6.4 Patients from different areas tend to call the SAMU at different stages of the pathology

Considering the piecewise linear curves in Figure 4, we note that in the 93, the date of the break point is shifted of 3 days, by comparison with the 75, suggesting that by this time, the patients of 93 were calling Center 15 at a later stage of the evolution of the disease. This hypothesis is confirmed by an examination of the ratio of the number of MICU dispatched over the number of ambulances dispatched. For instance, on March 30^{th}, there were 9 MICU dispatches and 276 other dispatches in the 75, to be compared with 25 MICU dispatches and 262 other dispatches in the 93, i.e., ratios of 3.3% in the 75 and 9.5% in the 93.

In the same way, the first breakpoints of the curves give an indication of the times at which "stade 2" measures influence the epidemic growth. These dates range from March 15^{th} (for 92) to March 22^{nd} (for 75). It may be the case that these dates are dispersed because the change of slope is relatively small, meaning that the effect of stade 2 is mild. Indeed, the milder the slope change, the more the estimation of the corresponding date is sensitive to noise. Another effect which may have perturbed the curves is the important mobility of the population in the 4 departments, between March 12^{th} and March 16^{th}.

6.5 The strong spatial heterogeneity of the evolution of the epidemic may be explained by local conditions

We observed in Section 4.3 that in the initial phase of the epidemic (Feb. 28^{th}–March 15^{th}), the doubling time was significantly shorter in the 93 department, whereas in the contraction phase, the halving time was significantly higher.

One may speculate that the contraction rate in the lockdown phase is influenced by intra-familial contaminations. In this respect, according to a survey of INSEE, the national institute of statistics, the average size of a household is of 2.6 in the 93, versus 1.9 in the 75 (values in 2016 [19]).

We also remark that just after the peak on March 27, and up to April 6, the curve of the department 93 on Figure 3 has the shape of a high, oscillating plateau, decaying more slowly than the curve of the department 75. This may be caused by changes in the nature of the dominant mode of contaminations, intra-familial contamination becoming an essential part of the kinetics during lockdown.

One may also speculate that the blowup rate in the initial phase is higher when the population is more dependent on public transport, or working in jobs with more contamination risk. These aspects will be further studied elsewhere.

After Stade 2 was announced, during the period from March 2$^{\text{nd}}$ to March 6$^{\text{th}}$, a number of large companies took specific measures (e.g., forbidding avoidable small group meetings, enforcing travel restrictions, restricting office access), in addition to the general measures (not shaking hands, forbidding large meetings) enforced by the authorities. This may have led to a decrease of the number of contamination on the workplace, and one explanation for the increase of the doubling time.

7 Epidemiological model based on transport PDE

7.1 Taking delays into account: a transport PDE SEIR model

We now introduce a multi-compartment transport PDE model, representing the dynamics of Covid-19. As explained in Section 3.2, in contrast to ODE models, that assume that the transition time from a compartment to the next one has an exponential distribution, PDE models capture *transition delays* bounded away from zero, an essential feature of Covid-19. An interest of this PDE model also lies in its unifying character: it includes as special cases, or as variations, SEIR ODE models that have been considered [27, 13].

We shall keep the traditional decomposition of individuals in compartments, "susceptible" (S), "exposed" (E), "infectious" (I), and "removed" from the contamination chain (R), as explained in Section 3.2, but the state variables attached to the E and I compartments will take the time elapsed in the compartment into account, and thus, will be infinite dimensional.

For all $t \geqslant 0$, we denote by $n_E(x,t)$ the density of the number of individuals that were contaminated x time units before time t, and that are not yet infectious at time t, i.e., the number of exposed individuals that began to be exposed at time $t-x$. Then, the size of the exposed population at time t is given by

$$E(t) = \int_0^\infty n_E(x,t)\,\mathrm{d}x \ . \tag{1}$$

Similarly, we denote by $n_I(x,t)$ the density of the number of individuals that became infectious x time units before time t, and that are not yet removed from the contamination chain at that time. Then, the size of the infectious population at time t is given by

$$I(t) = \int_0^\infty n_I(x,t)\,\mathrm{d}x \ . \tag{2}$$

Finally, we denote by $S(t)$ the number of susceptible individuals at time t, and by $R(t)$ the number of individuals that have been removed from the contamination chain before time t.

The total population at time t is given by

$$N(t) := S(t) + E(t) + I(t) + R(t) \ .$$

We consider the following system of PDE and ODE, with integral terms in the boundary conditions:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{S(t)}{N(t)} \int_0^\infty K_{I\to E}(x,t) n_I(x,t)\,\mathrm{d}x \ , \tag{3a}$$

$$n_E(0,t) = \frac{S(t)}{N(t)} \int_0^\infty K_{I\to E}(x,t) n_I(x,t)\,\mathrm{d}x \ , \qquad \frac{\partial n_E}{\partial t}(x,t) + \frac{\partial n_E}{\partial x}(x,t) + K_{E\to I}(x,t) n_E(x,t) = 0 \ , \tag{3b}$$

$$n_I(0,t) = \int_0^\infty K_{E\to I}(x,t) n_E(x,t)\,\mathrm{d}x \ , \qquad \frac{\partial n_I}{\partial t}(x,t) + \frac{\partial n_I}{\partial x}(x,t) + K_{I\to R}(x,t) n_I(x,t) = 0 \ , \tag{3c}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \int_0^\infty K_{I\to R}(x,t) n_I(x,t)\,\mathrm{d}x \ . \tag{3d}$$

We assume that an initial condition at time 0, $S(0)$, $n_E(\cdot,0)$, $n_I(\cdot,0)$ and $R(0)$ is given.

This is inspired by the so called "age structured models" considered in population dynamics. Kermack and McKendrick developed the first model of this kind to analyze the Plague epidemy of Dec. 1905 – July 1906 in Mumbai [22]. Von Forster [37] studied a similar model. Nowadays, these models are used as a general tool in population dynamics, with applications to biology and ecology [31, 28],

In these models, "age" refers to the age elapsed in a compartment – each transition to a new compartment resets to zero the "age" of an individual. In contrast, in the classical SEIR literature based on ODE, the standard notion of age (time elapsed since birth) is taken into account, via a contact matrix tabulating age-dependent infectiosity rates [27]. These two notions of age should not be confused. In the sequel, we shall use quotes, as in "age", to denote the age in a compartment, and will omit quotes to denote the ordinary age (since birth).

We suppose that $K_{I\to E}$, $K_{E\to I}$ and $K_{I\to R}$ are given *nonnegative* functions. The value $K_{E\to I}(x,t)$ gives the departure rate from the compartment E to the compartment I, for individuals of "age" x in the compartment E, at time t. Similarly, $K_{I\to R}(x,t)$ gives the departure rate from the compartment I to the compartment R. As in the classical SEIR model, the departure term from the susceptible compartment, i.e., the right-hand-side of (3a) is bilinear in the number $S(t)$ of susceptible individuals and in the population of infectious individuals $n_I(\cdot,t)$, and we normalize by the size of the population $N(t)$. The term $K_{I\to E}(x,t)$ can interpreted as an infection rate.

Differentiating $N(t)$ with respect to time, using the system above, and assuming that for all $t \geqslant 0$, $n_E(x,t)$ and $n_I(x,t)$ vanish when x tends to infinity, we verify that the total population $N(t)$ is independent of time.

When the functions $K_{I\to E}, K_{E\to I}$ and $K_{I\to R}$ are constant, taking into account (1) and (2), we recover the classical SEIR model from the dynamics (3):

$$\dot{S} = -\frac{S}{N} K_{I\to E} I \ , \tag{4a}$$

$$\dot{E} = \frac{S}{N} K_{I\to E} I - K_{E\to I} E \ , \tag{4b}$$

$$\dot{I} = K_{E\to I} E - K_{I\to R} I \ , \tag{4c}$$

$$\dot{R} = K_{I\to R} I \ . \tag{4d}$$

In the sequel, we shall consider (3) instead of (4), and we shall assume that the rates $K_{E\to I}(x,t) = K_{E\to I}(x)$ and $K_{I\to R}(x,t) = K_{I\to R}(x)$ are functions of x, independent of time. The rate $K_{I\to E}$ will have the product form

$$K_{I\to E}(x,t) = \mu(t)\psi(x) \ .$$

The function $\psi(\cdot)$ is fixed, it is nonnegative and not *a.e.* zero. In this way, the infectiosity of an individual depends on his "age" in the infectious phase, whereas the term $\mu(t)$ represents the control of the epidemic by sanitary measures (social distancing, wearing masks, closing schools, lockdown, etc.). We shall assume that the infectiosity rate $K_{I\to E}(x,t)$ is the only parameter which can be controlled, hence, $\mu(\cdot)$ is a decision

variable. A variant of the ODE model (4), in which $K_{I\to E}$ depends on time, but not on x, is considered in [9].

For epidemics in their early stages, i.e., when the number of individuals in the exposed, infectious, or removed compartments is negligible with respect to the number of susceptible individuals, the classical SEIR model is well-approximated by a linear system (see e.g. [22, 3]) tracking only the populations in the (E) and (I) compartments. As noted in Section 3.2, the fraction of the French population exposed prior to May 11 is estimated of 5.7% (see [32]), which justifies reliance on this linear approximation in our context. The same approximation applies to the present PDE model. This is translated to the assumption $S(t)/N(t) \simeq 1$, and we are reduced to the following system:

$$n_E(0,t) = \int_0^\infty \mu(t)\psi(x) n_I(x,t)\,\mathrm{d}x \ , \qquad \frac{\partial n_E}{\partial t}(x,t) + \frac{\partial n_E}{\partial x}(x,t) + K_{E\to I}(x) n_E(x,t) = 0\,, \quad \text{for } x>0\,, \tag{5a}$$

$$n_I(0,t) = \int_0^\infty K_{E\to I}(x) n_E(x,t)\,\mathrm{d}x\,, \qquad \frac{\partial n_I}{\partial t}(x,t) + \frac{\partial n_I}{\partial x}(x,t) + K_{I\to R}(x) n_I(x,t) = 0 \ , \quad \text{for } x>0 \ . \tag{5b}$$

This is a two-compartment generalization of the renewal equation, studied in Chapter 3 of [31].

In the sequel, we shall assume that there is a maximal "age" x_E^* of an individual in the exposed state. Similarly, we shall assume that is a maximal "age" x_I^* of an individual in the infectious state. These assumptions, which are consistent with epidemiological observations [25], will be incorporated in our model by forcing all remaining exposed individuals of "age" x_E^* to become infectious, with "age" 0. Similarly, all the remaining infectious individuals are removed when reaching "age" x_I^*. So, the function n_E is now only defined on the interval $[0, x_E^*]$, and similarly, n_I is only defined on $[0, x_I^*]$. This leads to the following system:

$$n_E(0,t) = \int_0^{x_I^*} \mu(t)\psi(x) n_I(x,t)\,\mathrm{d}x\,, \quad \frac{\partial n_E}{\partial t}(x,t) + \frac{\partial n_E}{\partial x}(x,t) + K_{E\to I}(x) n_E(x,t) = 0\,, \quad \text{for } 0<x<x_E^*\,, \tag{6a}$$

$$n_I(0,t) = \int_0^{x_E^*} K_{E\to I}(x) n_E(x,t)\,\mathrm{d}x + n_E(x_E^*,t)\,, \tag{6b}$$

$$\frac{\partial n_I}{\partial t}(x,t) + \frac{\partial n_I}{\partial x}(x,t) + K_{I\to R}(x) n_I(x,t) = 0\,, \quad \text{for } 0<x<x_I^*\,, \tag{6c}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t}(t) = \int_0^{x_I^*} K_{I\to R}(x) n_I(x,t)\mathrm{d}x + n_I(x_I^*) \ . \tag{6d}$$

This system may be obtained as a specialization of (5), in which $K_{E\to I}(x)$ is replaced by $K_{E\to I}(x)\mathbb{1}_{[0,x_E^*]}(x) + \delta_{x_E^*}(x)$, where $\mathbb{1}$ denotes the indicator function of a set, and δ denotes Dirac's delta function.

We shall assume, in the sequel, that the following assumption holds.

Asssumption 1. *The functions $K_{E\to I}(\cdot)$, defined on $[0, x_E^*]$, and $\psi(\cdot)$ and $K_{I\to R}(\cdot)$, defined on $[0, x_I^*]$, are nonnegative, measurable and bounded. Moreover, the function ψ does not vanish* a.e. *and the point x_I^* is the maximum of the essential support of the function ψ.*

Indeed, considering the boundary condition in (6a), we see that a population of "age" $x > \max \operatorname{ess\,supp} \psi$ in the infected (I) compartment will not participate any more to the contamination chain. Hence, the last part of Assumption 1 is needed to interpret R has the number of *all* the removed individuals.

Systems of PDE of this nature have been studied in particular by Michel, Mischler and Perthame, see [28, 31], and also, with an abstract semigroup perspective, in the work by Mischler and Scher [29].

Then, using the boundedness of the coefficients (Assumption 1), and arguing as in the proof of Theorem 3.1 of [31] – which concerns the case of a single compartment – one can show that the system (6) admits a unique solution in the distribution sense $n := (n_E, n_I)$ with $n_E \in \mathcal{C}(\mathbb{R}_{\geqslant 0}, L^1([0, x_E^*]))$ and $n_I \in \mathcal{C}(\mathbb{R}_{\geqslant 0}, L^1([0, x_I^*]))$. Hence, we can associate to the PDE (5) a well defined family of time evolution linear operators $(T_{s,t})_{t\geqslant s\geqslant 0}$, acting on the space $L^1([0, x_E^*]) \times L^1([0, x_I^*])$. The operator $T_{s,t}$ maps an initial condition at time $s \geqslant 0$, that is a couple of functions $n(\cdot, s) := (n_E(\cdot, s), n_I(\cdot, s))$, to the couple of functions

$n(\cdot,t) := (n_E(\cdot,t), n_I(\cdot,t))$ at $t \geqslant s$. These operators are order preserving, meaning that, if $n^1(\cdot,s)$ and $n^2(\cdot,s)$ are two initial conditions such that $n^1_E(x,s) \leqslant n^2_E(x,s)$ and $n^1_I(x,s) \leqslant n^2_I(x,s)$ for all $x \geqslant 0$, then the inequalities $n^1_E(x,t) \leqslant n^2_E(x,t)$ and $n^1_I(x,t) \leqslant n^2_I(x,t)$ hold for all $x \geqslant 0$ and for all $t \geqslant s$.

An alternative modeling, more in the spirit of [22], would be to consider a single compartment, describing the evolution of the density $n(x,t)$ of individuals that were contaminated at time $t-x$ by the system:

$$n(0,t) = \int_0^\infty \mu(t)\psi(x)n(x,t)\,\mathrm{d}x\,, \quad \frac{\partial n}{\partial t}(x,t) + \frac{\partial n}{\partial x}(x,t) + K(x)n(x,t) = 0\,, \quad \text{for } 0 < x < x^*\,, \tag{7}$$

where $x^* > x^*_E$ is fixed, and $\psi(x) = 0$ for $x < x^*_E$. Then, $E(t) = \int_0^{x^*_E} n(x,t)\mathrm{d}x$ yields the size of the exposed compartment. However, we prefer the model (6) as it allows us to represent variable incubation times.

The system (6d) can be extended to represent infectiosity rates that depends on the ages (time elapsed since birth) of individuals, with infectiosity rates given by a contact matrix, as in [27]. It suffices to split each compartment in sub-compartments, corresponding to different age groups. This will be detailed in a further work.

7.2 A Perron-Frobenius Eigenproblem for Transport PDE

When the control $\mu(t)$ is constant and positive, the family of time evolution operators $(T_{s,t})_{t\geqslant s\geqslant 0}$ is determined by the semigroup $(S_t = T_{0,t})_{t\geqslant 0}$, and the long term evolution of the dynamical system (3) can be studied by means of the *Perron–Frobenius eigenproblem*

$$\bar{n}_E(0) = \int_0^{x^*_I} \mu\psi(x)\bar{n}_I(x)\,\mathrm{d}x\ , \quad \frac{\mathrm{d}\bar{n}_E}{\mathrm{d}x}(x) + (\lambda + K_{E\to I}(x))\bar{n}_E(x) = 0 \quad \text{for } 0 < x < x^*_E\,, \tag{8a}$$

$$\bar{n}_I(0) = \int_0^{x^*_E} K_{E\to I}(x)\bar{n}_E(x)\,\mathrm{d}x + n_E(x^*_E)\ , \quad \frac{\mathrm{d}\bar{n}_I}{\mathrm{d}x}(x) + (\lambda + K_{I\to R}(x))\bar{n}_I(x) = 0 \quad \text{for } 0 < x < x^*_I\,, \tag{8b}$$

where $\bar{n} := (\bar{n}_I(\cdot), \bar{n}_E(\cdot))$ is a nonnegative eigenvector, and λ is the eigenvalue.

Since the functions $K_{E\to I}$ and $K_{I\to R}$ are independent of time, the existence of an eigenvector is an elementary result:

Proposition 1. *Suppose that Assumption 1 holds. Then, the eigenproblem* (8) *has a solution* $(\bar{n}, \lambda)$*, where* $\bar{n} = (\bar{n}_E, \bar{n}_I)$*, the functions* $\bar{n}_E$ *and* $\bar{n}_I$ *are continuous and positive, and* $\lambda \in \mathbb{R}$*. Moreover, the eigenvalue* λ *is unique, and the eigenvector* $\bar{n}$ *satisfying the latter conditions is unique up to a multiplicative constant.*

The proof of this proposition exploits a classical argument in renewal theory, see Lemma 3.1 p. 57 of [31]. We give the proof, leading to a semi-explicit representation of the eigenvector, which we shall need in Section 8.

We note first that the uniqueness of λ can be deduced from a general observation, of independent interest:

Lemma 1. *Let* $w = (w_E, w_I)$*, with* $w_E \in L^1([0, x^*_E])$ *and* $w_I \in L^1([0, x^*_I])$*, be such that*

$$\alpha\bar{n} \leqslant w \leqslant \beta\bar{n} \tag{9}$$

for some $\alpha, \beta > 0$*. Then,*

$$\alpha\exp(\lambda t)\bar{n} \leqslant S_t w \leqslant \beta\exp(\lambda t)\bar{n}, \qquad \textit{for all } t \geqslant 0\ . \tag{10}$$

Proof. This follows from the order preserving and linear character of the semigroup S_t, together with $S_t\bar{n} = \exp(\lambda t)\bar{n}$. □

Therefore, Lemma 1 shows that λ is the growth rate of $n(\cdot,t)$ as $t \to \infty$, for all initial conditions $w = n(\cdot,0)$ satisfying (9). In particular, λ is unique.

We next provide a semi-explicit formula for the eigenvector. We set

$$F^{\lambda}_{E\to I}(x) := \int_0^x (\lambda + K_{E\to I}(z))\mathrm{d}z\,,\ \text{for } 0 \leqslant x \leqslant x^*_E \quad F^{\lambda}_{I\to R}(x) := \int_0^x (\lambda + K_{I\to R}(z))\mathrm{d}z\,,\ \text{for } 0 \leqslant x \leqslant x^*_I\,.$$

Integrating the differential equations (8), we see that a nonnegative eigenvector $\bar{n} = (\bar{n}_E, \bar{n}_I)$ necessarily satisfies:

$$\bar{n}_E(x) = \exp\big(-F^{\lambda}_{E\to I}(x)\big)\bar{n}_E(0)\,,\ \text{for } 0 \leqslant x \leqslant x^*_E \quad \bar{n}_I(y) = \exp\big(-F^{\lambda}_{I\to R}(y)\big)\bar{n}_I(0)\,,\ \text{for } 0 \leqslant y \leqslant x^*_I\,. \tag{11}$$

We deduce from the above relations that $\bar{n}_E(\cdot)$ is continuous on $[0, x^*_E]$, and that $\bar{n}_I(\cdot)$ is continuous on $[0, x^*_I]$. Using the boundary condition in (8b), we deduce that if $\bar{n}_E(0) = 0$, then $\bar{n}$ is identically 0, a contradiction. Hence, $\bar{n}$ is everywhere positive. Moreover, (11) and (8b) entail that the eigenvector $\bar{n}$ is unique, up to a scalar multiple.

Using the boundary conditions in (8), and specializing (11) to $x = x^*_E$ and $y = x^*_I$, we deduce that $\mu G^{\lambda}\bar{n}_E(0) = \bar{n}_E(0)$, where

$$G^{\lambda} = \Big(\int_0^{x^*_I} \psi(x)\exp\big(-F^{\lambda}_{I\to R}(x)\big)\mathrm{d}x\Big)\Big(\int_0^{x^*_E} K_{E\to I}(y)\exp\big(-F^{\lambda}_{E\to I}(y)\big)\mathrm{d}y + \exp\big(-F^{\lambda}_{E\to I}(x^*_E)\big)\Big)\,.$$

Therefore, to find an eigenvector, we must solve the equation $\mu G^{\lambda} = 1$ (the so-called "characteristic equation" in renewal theory). Since the functions $K_{E\to I}$, $K_{I\to R}$ and ψ are integrable, and ψ is nonzero on a set of positive measure, we deduce that $\lim_{\lambda\to-\infty} G^{\lambda} = +\infty$. We also have $\lim_{\lambda\to+\infty} G^{\lambda} = 0$. Moreover, the map $\lambda \mapsto G^{\lambda}$ is continuous. Since $\mu > 0$, by the intermediate value theorem, we can find λ such that $\mu G^{\lambda} = 1$, and this λ is the eigenvalue. This concludes the proof of Proposition 1. □

The asymptotic bound (10) can be reinforced, by showing that, for all positive initial conditions w,

$$S_t w = C_1(w)\bar{n}\exp(\lambda t) + O(\exp(\lambda_2 t))\ ,\qquad \text{as } t\to\infty\ , \tag{12}$$

for some positive constant $C_1(w)$, and $\lambda_2 < \lambda$. This result, with an explicit control of λ_2, can be obtained as follows. We make a diagonal scaling, using the positive eigenvector, and we normalize the semigroup to make the Perron eigenvalue λ equal to zero. This leads to the semigroup

$$(\tilde{S}_t w)(x) := \exp(-\lambda t)\bar{n}^{-1}(x)[S_t(w\bar{n})](x)\,.$$

In potential theory, a version of this scaling is known as *Doob's h-transform* (see e.g. [14]). The semigroup $\tilde{S}_t$ obtained in this way is associated with a Markov process, and, so, the spectral gap of this semigroup can be bounded in terms of Doeblin's ergodicity coefficient [16, 5], leading to (12). These aspects will be detailed elsewhere. Alternatively, the relative entropy inequality technique of [28] allows one to establish the convergence of $n(\cdot, t)$ to the eigenvector, modulo multiplicative constants, as t tends to infinity.

7.3 Universality of the log-rate of epidemic observables

Epidemic observables are obtained by applying a continuous linear form to the state variable. Supposing that $n_I(\cdot, t)$ is a continuous function, an epidemic observable will be of the form

$$Y_{\kappa}(t) = \varphi(n(\cdot, t)) := \int_0^{x^*_I} n_I(x, t)\,\mathrm{d}\kappa(x)\ , \tag{13}$$

where $\mathrm{d}\kappa(x)$ is a nonnegative nonzero Borel measure. Epidemic events anterior to the infectious phase, like contamination, are by nature hard to detect, so the observable depends only on n_I.

Proposition 2. *Suppose that Assumption 1 holds, let $(\lambda, \bar{n})$ denote the solution of the Perron-Frobenius eigenproblem (8), and suppose that for some $T > 0$, there exist positive constants α, β such that $\alpha\bar{n} \leqslant n(\cdot, T) \leqslant \beta\bar{n}$. Then, for all epidemic observables of the form (13), the map $t \mapsto \log Y_{\kappa}(t) - \lambda t$ is bounded. A fortiori,*

$$\lim_{t\to\infty} \frac{1}{t}\log Y_{\kappa}(t) = \lambda\ .$$

Proof. Taking $w = n(\cdot, T)$ in Lemma 1, we deduce that

$$\alpha \exp(\lambda t)\bar{n} \leqslant S_{T+t} n(\cdot, 0) = n(\cdot, T+t) \leqslant \beta \exp(\lambda t)\bar{n}\,, \text{ for } t > 0\ .$$

It follows that $\log \alpha + \lambda t + \log \varphi(\bar{n}) \leqslant \log Y_\kappa(T+t) \leqslant \log \beta + \lambda t + \log \varphi(\bar{n})$. □

A simple example of observable, discussed in Section 3.2, consists of *pure delays*. For instance, we assumed that the number of dispatches of MICU is given by $Y_{\mathrm{MICU}}(t) = \pi_{\mathrm{MICU}} C(t - \tau_{\mathrm{MICU}})$ where $C(t)$ is the number of contaminations at time t, π_{MICU} the proportion of contaminated patients who will need a MICU transport, and τ_{MICU} a fixed delay. This can be obtained as a special case of (13), taking $K_{E\to I} \equiv 0$, so that the transition tom E to I occurs always at time x_E^*, and $\mathrm{d}\kappa := \pi_{\mathrm{MICU}} \delta_{\tau_{\mathrm{MICU}} - x_E^*}$, where δ is the Dirac δ function.

Other events can be considered: medical advice, EMT dispatch, admission to ICU, or decease. These events corresponds to different values of the proportion π and of the delay τ. By Proposition 2, the rate $\lim_{t\to\infty} t^{-1} \log Y(t)$ will be the same for all the corresponding observables, although the convergence of the function $t^{-1} \log Y(t)$ to its limit will be observed in a delayed manner, for observables corresponding to the latest stages of the pathology.

7.4 Discrete versions of the epidemiological model

The reader interested in ODE model of epidemics might wish to note that the previous analysis applies to such finite dimensional models. Instead of the transport PDE (5), we may consider an ODE of the form

$$\dot{v} = Mv \tag{14}$$

where $v(t) \in \mathbb{R}^n$ and M is a $n \times n$ matrix with non-negative off-diagonal terms, a so-called *Metzler matrix*. In the original SEIR model [3], the matrix M, obtained by considering the (E, I)-block equations (4b), (4c), with $S/N \simeq 1$, is of dimension 2. In the generalizations of the SEIR model considered in [27, 13], the dimension n is increased to account for other compartments. One can also discretize the PDE system (5) using a monotone (upwind) finite difference scheme, and this leads to a system of the form (14).

In all these finite dimensional models, the matrix M is Metzler and irreducible. Then, the Perron–Frobenius theorem for linear, order-preserving semigroups (see [7]) implies that M admits a unique eigenvalue λ of maximal real part. Furthermore λ is algebraically simple and real, and its associated eigenvector u has strictly positive coordinates. Then, it follows from the spectral theorem that

$$v(t) = \exp(\lambda t)u + o(\exp(\lambda_2 t))$$

as $t \to \infty$, where λ_2 is the maximal real part of an eigenvalue of M distinct from λ. Again, in this discrete model, an epidemic observable $Y(t)$ is obtained by applying a nonnegative linear form to the vector $v(t)$, i.e, $Y(t) = \varphi^\top v(t)$, for some nonnegative column vector φ.

8 Tropicalization of the logarithm of nonnegative observables of switched Perron–Frobenius dynamics

8.1 Hilbert's geometry applied to piecewise linear approximation

We introduce an abstract setting, which captures epidemiological models in which most individuals are susceptible. This setting applies, in particular, to the transport PDE model of (5), when the transition functions are supported by compact intervals, and to the general finite dimensional Metzler model (14).

We consider $(V, \leqslant)$, a partially ordered Banach space, with topological dual V'. We denote by $V_{\geqslant 0} := \{v \in V \mid v \geqslant 0\}$ the set of nonnegative elements of V, which is a convex cone. This cone must be pointed (i.e., $V_{\geqslant 0} \cap (-V_{\geqslant 0}) = \{0\}$), since the relation $\leqslant$ is a partial order. We require this cone to be closed.

We consider a sequence of m semigroups $S^i = (S^i_t)_{t\geqslant 0}$, for $i \in [m]$, where $[m] := \{1, \dots, m\}$. We assume that for all $i \in [m]$, and for all $t \geqslant 0$, S^i_t is a bounded linear operator from V to itself, and that the semigroup property holds, i.e., $S^i_{t+s} = S^i_t \circ S^i_s$. We shall say that the semigroup S^i is *order preserving* if, for all $v \in V_{\geqslant 0}$, and for all $t \geqslant 0$, $S^i_t v \in V_{\geqslant 0}$.

We shall consider commutation instants, $t_0 := 0 < t_1 < \cdots < t_{m-1}$, These instants will correspond to significant epidemiological dates, for instance, dates at which sanitary measures are taken. We set $t_m := +\infty$.

We select an initial condition $v_0 \in V_{\geqslant 0}$, and consider the abstract dynamical system obtained by switching between the evolutions determined by the semigroups $S^1, \dots, S^m$, at the successive times $t_1, \dots, t_{m-1}$. The state of this dynamical system, at time $t \in [t_j, t_{j+1})$, is given by

$$v_t := S^{j+1}_{t-t_j} \circ S^j_{t_j - t_{j-1}} \circ \cdots \circ S^1_{t_1 - t_0}(v_0) \ . \tag{15}$$

Recall that a *part* of the closed convex cone $V_{\geqslant 0}$ is an equivalence class for the relation $\sim$ such that, for $v, w \in V_{\geqslant 0}$, we have $v \sim w$ if and only if there exists two positive constants α and β such that $\alpha v \leqslant w \leqslant \beta v$. A part is *trivial* if it is reduced to the equivalence class of the zero vector. *Hilbert's projective metric* d_H is defined on every non-trivial part of $V_{\geqslant 0}$ by the following formula

$$d_H(v, w) = \log \inf \left\{ \frac{\beta}{\alpha} : \alpha, \beta > 0,\ \alpha v \leqslant w \leqslant \beta v \right\} \ .$$

The infimum is achieved, since $V_{\geqslant 0}$ is closed. The map d_H is nonnegative, it satisfies the triangular inequality, and $d_H(v, w)$ vanishes if, and only if, v and w are proportional – this justifies the term "projective metric". This metric plays a fundamental role in Perron–Frobenius theory and in metric geometry, and also in tropical geometry, see [26, 30, 10] for background.

When $V_{\geqslant 0} = (\mathbb{R}_{\geqslant 0})^n$ is the standard orthant, and when all the entries of the vectors v and w are positive, we have

$$d_H(v, w) = \max_{k\in[n]} (\log v_k - \log w_k) - \min_{k\in[n]} (\log v_k - \log w_k) \ .$$

Denoting by e the unit vector of $\mathbb{R}^n$, we observe that

$$d_H(v, w) = \| \log v - \log w \|_H$$

where the notation $\log v$ is understood entrywise, and

$$\|z\|_H = 2 \min_{c\in\mathbb{R}} \|x - ce\|_\infty \ .$$

In other words, up to a logarithmic change of variables, d_H arises by modding out the normed space $(\mathbb{R}^n, \| \cdot \|_\infty)$ by the one-dimensional space $\mathbb{R}e$.

We shall suppose that every semigroup S^i has an eigenvector $u^i \geqslant 0$, with eigenvalue λ^i, meaning that

$$S^i_t u^i = \exp(\lambda^i t) u^i, \quad \forall t \geqslant 0 \ .$$

Since S^i_t preserves $V_{\geqslant 0}$, this entails that λ^i is real.

We choose a linear form $\varphi \in V'$ which we require to take nonnegative values on $V_{\geqslant 0}$. We shall think of V has the *state space* and φ as an *observable*. We consider the following scalar observation of the dynamics

$$Y_t := \varphi(v_t) \ .$$

We shall assume, in addition, that φ does not vanish on v_t, for all $t \geqslant 0$. Then, we can define the image of the observation by the logarithmic map

$$y_t := \log Y_t, \qquad \forall t \geqslant 0 \ .$$

The following result shows that the logarithm of the observation stays at finite distance from a piecewise linear map.

Theorem 1. *Suppose that the semigroups $S^1, \dots, S^m$ are order preserving. Suppose in addition that the initial condition v_0 and the eigenvectors $u^1, \dots, u^m$ all lie in the same non-trivial part of $V_{\geqslant 0}$, and that the linear form φ takes positive values on this part. Then, there exists a constant C such that the piecewise linear map $t \mapsto y_t^{\mathrm{trop}}$ defined, for $t \in [t_j, t_{j+1})$, by*

$$y_t^{\mathrm{trop}} := \lambda_{j+1}(t - t_j) + \lambda_j(t_j - t_{j-1}) + \cdots + \lambda_1(t_1 - t_0) + C \ ,$$

satisfies

$$|y_t - y_t^{\mathrm{trop}}| \leqslant \frac{\Delta}{2}, \qquad \forall t \geqslant 0 \ ,$$

where

$$\Delta = d_H(v_0, u^1) + d_H(u^1, u^2) + \cdots + d_H(u^{m-1}, u^m) \ .$$

Proof. By definition of Hilbert's projective metric, we can find positive constants α_0, β_0, such that $\alpha_0 u^1 \leqslant v_0 \leqslant \beta_0 u^1$ and $d_H(v_0, u^1) = \log(\beta_0/\alpha_0)$. Similarly, for all $i \in [m-1]$, we can find positive constants α_i, β_i, such that $\alpha_i u^{i+1} \leqslant u^i \leqslant \beta_i u^{i+1}$, and $d_H(u^i, u^{i+1}) = \log(\beta_i/\alpha_i)$. For all $j \geqslant 0$, with $j \leqslant m-1$, and for all $t \in [t_j, t_{j+1})$, we set

$$z_t := \lambda_{j+1}(t - t_j) + \lambda_j(t_j - t_{j-1}) + \cdots + \lambda_1(t_1 - t_0) \ .$$

Since the semigroups S^i are linear and order preserving, we prove by induction

$$\exp(z_t)\alpha_j \dots \alpha_0 u^{j+1} \leqslant v_t \leqslant \exp(z_t)\beta_j \dots \beta_0 u^{j+1} \ .$$

We observe that

$$\alpha_{m-1} \dots \alpha_{j+1} u^m \leqslant u^{j+1} \leqslant \beta_{m-1} \dots \beta_{j+1} u^m$$

and so

$$\exp(z_t)\alpha_{m-1} \dots \alpha_0 u^m \leqslant v_t \leqslant \exp(z_t)\beta_{m-1} \dots \beta_0 u^m \ .$$

Applying the linear form φ to latter inequalities, taking the image by the log map, and setting

$$C := \log \varphi(u^m) + \frac{1}{2} \sum_{j=0}^{m-1} \log(\beta_j \alpha_j) \ ,$$

we arrive at the bound of the theorem. □

A general principle from tropical geometry states that using "logarithmic glasses" reveals a piecewise linear structure [36, 20]. Theorem 1 is inspired by this principle. This motivates the notation y^{trop}, for the "tropicalization" of the logarithm of the observable Y.

Theorem 1 carries over to discrete time systems in a straightforward manner.

8.2 Application to the transport PDE model

Theorem 1 applies in particular to the transport model (6). Then, as noted above, the evolution operator of the system (5) preserves the space $V = L^1([0, x_E^*]) \times L^1([0, x_I^*])$. Moreover, when the epidemiological control term $\mu(t)$ is constant, Proposition 1 shows that the eigenproblem (8) has a positive and continuous solution $\bar{n}$, with a real eigenvalue λ. Different stages of sanitary policies correspond to successive values $\mu^1, \dots, \mu^m$ of $\mu(t)$, leading to different semigroups S^i, $i \in [m]$. Then, the solution $v_t := n(\cdot, t)$ of (5) is determined as in (15). Each semigroup S^i yields a continuous and positive eigenvector $u^i := \bar{n}^i$ satisfying (8) associated with a real eigenvalue λ^i of S^i. Two continuous and positive functions defined on a compact interval are always in the same part of the cone of nonnegative functions of V, so Theorem 1 applies to this model.

We next give an explicit estimate for the Hilbert projective distances between eigenvectors, arising in Theorem 1.

Proposition 3. *Suppose that Assumption 1 holds, and that for $i = 1, 2$, $(\lambda^i, \bar{n}^i)$ is the solution $(\lambda, \bar{n})$ of the Perron-Frobenius eigenproblem (8) when $\mu = \mu^i$. Then, we have*

$$d_H(\bar{n}^1, \bar{n}^2) \leqslant |\lambda_1 - \lambda_2|(x_E^* + x_I^*) \ .$$

The term $x_E^* + x_I^*$ is the maximal time elapsed between contamination and the end of infectiosity.

Proof of Proposition 3. Suppose, without loss of generality, that $\bar{n}_E^i(0) = 1$ for $i = 1, 2$. Let $\phi_{E\to I}(x) := -\int_0^x K_{E\to I}(z)\mathrm{d}z$ and $\phi_{I\to R}(x) := -\int_0^x K_{I\to R}(z)\mathrm{d}z$. Then, (11) yields

$$\bar{n}_E^i(x) = \exp(-\lambda^i x)\phi_{E\to I}(x)\,, \tag{16}$$

$$\bar{n}_I^i(0) = \int_0^{x_E^*} K_{E\to I}(x)\phi_{E\to I}(x)\exp(-\lambda^i x)\mathrm{d}x + \phi_{E\to I}(x_E^*)\exp(-\lambda^i x_E^*)\,, \tag{17}$$

$$1 = \bar{n}_E^i(0) = \mu^i\Big(\int_0^{x_I^*} \psi(x)\exp(-\lambda^i x)\mathrm{d}x\Big)\bar{n}_I^i(0) \ . \tag{18}$$

Let $j \in \{1, 2\}$ be distinct from i. Then, setting $t^+ := \max(t, 0)$, bounding $\exp(-\lambda^i x)$ by $\exp(-\lambda^j x)\exp\big((\lambda^j - \lambda^i)^+ x_E^*\big)$ in (17), and applying a similar bound in (18), we obtain:

$$\bar{n}_E^i(0) \leqslant \frac{\mu^i}{\mu^j}\exp\big((\lambda^j - \lambda^i)^+(x_I^* + x_E^*)\big)\bar{n}_E^j(0)\,.$$

which yields $\mu^j/\mu^i \leqslant \exp\big((\lambda^j - \lambda^i)^+\big)(x_I^* + x_E^*)$. We deduce that $\bar{n}_I^i(x) \leqslant \bar{n}_I^j(x)\exp\big((\lambda^j - \lambda^i)^+\big)(x_I^* + x_E^*)$. So,

$$d_H((\bar{n}_E^i, \bar{n}_I^i), (\bar{n}_E^j, \bar{n}_I^j)) \leqslant (\lambda^j - \lambda^i)^+(x_I^* + x_E^*) + (\lambda^i - \lambda^j)^+(x_I^* + x_E^*) = |\lambda^i - \lambda^j|(x_I^* + x_E^*)\,. \qquad \square$$

The bound of Theorem 1 may be refined. This is left for further work.

9 Short term predictions

We now describe the basic methodology we propose to build confidence intervals for future occurrences of medical events related to epidemic progression, and raise alarms about its potential resurgence. We first consider a single time series of numbers of event occurrences. We then describe how to consolidate several time series corresponding to distinct medical events in order to construct improved alarm criteria. The simpler case of least squares fitting is considered first, the more robust ℓ_1 alternative is described next.

Time series for a single type of events: Let $X(1), \ldots, X(n)$ be indices of days, and we aim to do a forecast based on observations made on these days. Typically, on day d_0, we may select $n = 7$ and let $X(1) = d_0 - n, \ldots, X(n) = d_0 - 1$ to perform a forecast on the basis of the last seven days. Let $Y(t)$ denote the count of medical events (for instance, dispatches of ambulances) on day $X(t)$, and let $Z(t) = \log Y(t)$. Based on the previous discussion (epidemiological modeling) we assume that for all $t = 1, \ldots, n$,

$$Z(t) = \alpha + \beta X(t) + \epsilon_t$$

for constants α, β, where ϵ_t denotes some random noise. For simplicity we assume here i.i.d. noise sequence $\epsilon_1, \ldots, \epsilon_n$, and that each ϵ_t admits a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ with zero mean and variance σ^2.

Least-square estimates for the parameters α, β are then provided by

$$\hat{\beta} = \frac{\sum_{t=1}^n (X(t) - \bar{X})(Z(t) - \bar{Z})}{\sum_{t=1}^n (X(t) - \bar{X})^2}, \quad \hat{\alpha} = \bar{Z} - \hat{\beta}\bar{X}, \tag{19}$$

where

$$\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X(t), \quad \bar{Z} = \frac{1}{n}\sum_{t=1}^{n} Z(t). \tag{20}$$

The variance σ^2 can be estimated as

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{t=1}^{n}(Z(t) - \hat{Z}(t))^2, \tag{21}$$

where

$$\hat{Z}(t) := \hat{\alpha} + \hat{\beta}X(t). \tag{22}$$

Under the assumptions of i.i.d. Gaussian errors ϵ_t, we have that, for each t corresponding to a future day day $X(t)$ (in particular, $t \notin \{1, \dots, n\}$), the three following variables:

$$\frac{\hat{\alpha} - \alpha}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{t'=1}^{n}(X(t')-\bar{X})^2}}}, \quad \frac{\hat{\beta} - \beta}{\hat{\sigma}\sqrt{\frac{1}{\sum_{t'=1}^{n}(X(t')-\bar{X})^2}}}, \quad , \frac{Z(t) - \hat{Z}(t)}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X(t)-\bar{X})^2}{\sum_{t'=1}^{n}(X(t')-\bar{X})^2}}},$$

all admit a bilateral Student distribution with $n-2$ degrees of freedom (see [21, Ch. 28] or [11]). Denote by t_γ^{n-2} the γ-th quantile of this distribution. For $\epsilon \in [0, 1]$, this provides us with the following confidence intervals with confidence $1 - \epsilon$:

$$\begin{array}{lcl}
\alpha & \in & \left[\hat{\alpha} - \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{t'=1}^{n}(X(t')-\bar{X})^2}}t_{1-\epsilon/2}^{n-2}, \hat{\alpha} + \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{t'=1}^{n}(X(t')-\bar{X})^2}}t_{1-\epsilon/2}^{n-2}\right], \\
\beta & \in & \left[\hat{\beta} - \hat{\sigma}\sqrt{\frac{1}{\sum_{t'=1}^{n}(X(t')-\bar{X})^2}}t_{1-\epsilon/2}^{n-2}, \hat{\beta} + \hat{\sigma}\sqrt{\frac{1}{\sum_{t'=1}^{n}(X(t')-\bar{X})^2}}t_{1-\epsilon/2}^{n-2}\right] \\
Z(t) & \in & \left[\hat{Z}(t) - \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X(t)-\bar{X})^2}{\sum_{t'=1}^{n}(X(t')-\bar{X})^2}}t_{1-\epsilon/2}^{n-2}, \hat{Z}(t) + \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X(t)-\bar{X})^2}{\sum_{t'=1}^{n}(X(t')-\bar{X})^2}}t_{1-\epsilon/2}^{n-2}\right]
\end{array} \tag{23}$$

As an illustration, for $n = 7$ and $\epsilon = 5\%$, we can plug in $t_{0.975}^5 = 2.571$ in the last interval, and thus obtain a 95%-confidence interval centered around $\hat{Z}(t)$ for $Z(t) = \log Y(t)$, the logarithm of the count $Y(t)$ on a future day $X(t)$, that is:

$$Z(t) \in \left[\hat{Y}(t) - 2.571 \times \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X(t)-\bar{X})^2}{\sum_{t'=1}^{n}(X(t')-\bar{X})^2}}, \hat{Z}(t) + 2.571 \times \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X(t)-\bar{X})^2}{\sum_{t'=1}^{n}(X(t')-\bar{X})^2}}\right] \tag{24}$$

Although we could extend this definition of the confidence interval for the short terms predictions of the value of $Z(t)$, we propose a more conservative confidence domain, in the shape of a trapezoid. It is obtained by extending the upper-bound $Z(t)^+$ (resp. the lower bound $Z(t)^-$ of the 95% confidence interval on $Z(t)$ by a line with slope equal to the upper-bound β^+ (resp. lower-bound β^-) of 95% confidence interval on β. For a given day t, the upper and lower envelopes of the trapezoid have ordinates

$$\left(\hat{\beta}(t - t_n) + \hat{Z}_n\right) \pm \left(\sqrt{\mathrm{Var}(\hat{\beta})}(t - t_n) + \sqrt{\widehat{\sigma}^2 + \mathrm{Var}(\hat{Z}_n)}\right) t_{1-\epsilon/2}^{n-2} \ .$$

If instead of the count $Y(t)$ on a particular day $X(t)$, we are interested in the trend of the epidemic, whether exploding or contracting, we should then consider the confidence interval for parameter β. Again for $n = 7$ and $\epsilon = 5\%$ this gives

$$\beta \in \left[\hat{\beta} - 2.571 \times \hat{\sigma}\sqrt{\frac{1}{\sum_{t'=1}^{n}(X(t')-\bar{X})^2}}, \hat{\beta} + 2.571 \times \hat{\sigma}\sqrt{\frac{1}{\sum_{t'=1}^{n}(X(t')-\bar{X})^2}}\right] \tag{25}$$

One-sided confidence intervals may also be provided, and are in fact more natural for the definition of alarm indicators.

For concreteness, assume we want to raise an alarm when the doubling time, $\delta = (\log 2)/\beta$, is δ^* days or less, where δ^* could be 10 for instance. This is equivalent to β exceeding $(\log 2)/\delta^*$. Thus δ is less than δ^* days with confidence $1-\epsilon$ when

$$\frac{\log 2}{\delta^*} < \hat{\beta} - t_{1-\epsilon}^{n-2}\sqrt{V},$$

where

$$V = \hat{\sigma}\sqrt{\frac{1}{\sum_{t'=1}^{n}(X(t') - \bar{X})^2}}.$$

Raising an alarm under this condition then amounts to calibrating the false positive probability at ϵ. For instance, for $\epsilon = 5\%$, and $n = 7$, we would plug in $t_{0.95}^{7} = 2.015$ in the above expression.

Alternatively, raising an alarm under the condition

$$\frac{\log 2}{\delta^*} < \hat{\beta} + t_{1-\epsilon}^{n-2}\sqrt{V},$$

corresponds to calibrating the false negative probability (probability of not raising an alarm while $\delta \leqslant 10$) at ϵ.

Our alarm indicators correspond to the first choice, i.e. calibration of a false positive rate, with δ^* set to $+\infty$.

Alarm indicators based on multiple types of events: Assume that several types j of events are available, and let J denote the corresponding set of events. For instance, we could distinguish between dispatches of ambulances bringing patients to Intensive Care Units as opposed to Non-intensive Care Units, thereby producing two distinct time series. Let $X_j(t)$, $t = 1, \ldots, n_j$ denote the days on which counts $Y_j(t)$ of type j event occurrences are to be used. Let $Z_j(t) = \log Y_j(t)$. We assume as before the linear regression model

$$Z_j(t) = \alpha_j + \beta_j X_j(t) + \epsilon_j(t),\ t = 1, \ldots, n^j\ .$$

Now for each of these times series, we can produce, based on the previous discussion, the estimator

$$\hat{\beta}_j := \frac{\sum_{t=1}^{n_j}(X_j(t) - \bar{X}_j)(Z_j(t) - \bar{Z}_j)}{\sum_{t=1}^{n_j}(X_j(t) - \bar{X}_j)^2},$$

where

$$\bar{X}_j = \frac{1}{n_j}\sum_{t=1}^{n_j} X_j(t), \quad \bar{Z}_j = \frac{1}{n_j}\sum_{t=1}^{n_j} Z_j(t).$$

Suppose in addition that the noise terms $\epsilon_j(t)$ are mutually independent, Gaussian, with zero mean and variance σ_j^2 for errors $\epsilon_j(t)$. Suppose finally that the exponents β_j all coincide with β, the exponent that is characteristic of the epidemic's progression. Denote by

$$V_j := \hat{\sigma}_j^2\sqrt{\frac{1}{\sum_{t=1}^{n_j}(X_j(t) - \bar{X}_j)^2}}, \tag{26}$$

where, reproducing the computations for a single time series, we let

$$\hat{\sigma}_j^2 := \frac{1}{n_j - 2}\sum_{t=1}^{n_j}(Z_j(t) - \hat{Z}_j(t))^2,$$

and

$$\hat{Z}_j(t) := \hat{\alpha}_j + \hat{\beta}_j X_j(t).$$

As previously, V_j is our estimate of the variance of estimate $\hat{\beta}_j$. We finally propose to combine the individual estimators $\hat{\beta}_j$ into

$$\hat{\beta} := \frac{\sum_{j \in J} \frac{1}{V_j} \hat{\beta}_j}{\sum_{j \in J} \frac{1}{V_j}}. \tag{27}$$

For the sake of simplicity, let us approximate the bilateral Student distribution with $n-2$ degrees of freedom by the standard distribution $\mathcal{N}(0,1)$. We then have the approximate distributions $\hat{\beta}_j \approx \mathcal{N}(\beta, V_j)$, and hence the approximate distribution $\hat{\beta} - \beta \approx \mathcal{N}(0, V)$,

$$V := \frac{1}{\sum_{j \in J} \frac{1}{V_j}}. \tag{28}$$

Weighing the individual estimators $\hat{\beta}_j$ by the reciprocal of their variances as just done minimizes the variance of the resulting estimator. The same approach as previously considered then leads to the following conditions for alarm raising:

To raise an alarm when the doubling time $\delta = (\log 2)/\beta$ exceeds δ^* days (e.g., $\delta^* = 10$), if we target a false alarm probability of ϵ, we are led to raise an alarm when Condition

$$\frac{\log 2}{\delta^*} < \hat{\beta} - g_{1-\epsilon}\sqrt{V}, \tag{29}$$

where $g_{1-\epsilon}$ is the $1-\epsilon$-quantile of the standard Gaussian distribution.

If instead we target a false negative probability (probability of not raising an alarm) at ϵ, we would then raise an alarm when

$$\frac{\log 2}{\delta^*} < \hat{\beta} + g_{1-\epsilon}\sqrt{V}, \tag{30}$$

More robust ℓ_1-based approach: The previous estimators and derived alarm conditions have the appeal of simplicity, but can be advantageously replaced by more robust versions, that are less sensitive to the presence of outliers.

A popular alternative is the following ℓ_1 criterion. We again consider $Z_j(t) := \log Y_j(t)$, where $Y_j(t)$ is the number of type j events on day $X_j(t)$. We then let $\hat{\alpha}_j$, $\hat{\beta}_j$ achieve the minimum of the criterion $\sum_{t=1}^{n^j} |\alpha + \beta X_t^j - Y_t^j|$. They are obtained by solving a linear program. Here we assume that observations $Z_j(t)$ are mutually independent and distributed according to density $f_{j,t}(z) = \frac{1}{2\lambda^j} \exp(-|z - \alpha_j - \beta_j X_j(t)|/\lambda_j)$. In other words this corresponds to adding a Laplace observation noise with density $\frac{1}{2\lambda^j} \exp(-|z|/\lambda_j)$ to the signal of interest $\alpha_j + X_j(t)\beta_j$. The above ℓ_1 minimization criterion corresponds to maximum likelihood estimation of α_j, β_j in this observational noise model, as its log-likelihood is given by

$$-|T| \log(2\lambda_j) - \sum_{t=1}^{n_j} \frac{|Z_j(t) - \alpha_j - \beta_j X_j(t)|}{\lambda_j}.$$

A rich theory for the performance of the resulting estimators is available, see for instance [23]. The latter work treats general i.i.d. errors, and do not restrict itself to e.g. Laplacian distribution of errors; recent work like [33] experiments techniques to obtain confidence intervals when distribution of errors is unknown. Here we make the choice of Laplace-distributed errors for sake of simplicity. In particular, the asymptotic theory in [23] suggests the approximation

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, V_j)$$

where

$$\hat{\lambda}_j := \frac{1}{n_j} \sum_{t=1}^{n_j} |Z_j(t) - \hat{\alpha}_j - \hat{\beta}_j X_j(t)|, \tag{31}$$

and

$$V_j := (\hat{\lambda}_j)^2 \frac{1}{\sum_{t=1}^{n_j} X_j(t)^2 - \frac{1}{n_j}(\sum_{t=1}^{n_j} X_j(t))^2}. \tag{32}$$

We again consider that multiple types $j \in J$ of time series are conjointly available, and that each β^j coincides with β, the parameter to be estimated. Assuming the $\hat{\beta}_j$ to be independent with $\hat{\beta} \sim \mathcal{N}(\beta_j, V_j)$, leads us to define the estimator

$$\hat{\beta} := \frac{\sum_{j\in J} \frac{\hat{\beta}_j}{(\hat{\lambda}_j)^2}}{\sum_{j\in J} \frac{1}{(\hat{\lambda}_j)^2}}, \tag{33}$$

whose distribution is then given by $\hat{\beta} \sim \mathcal{N}(\beta, V)$ where

$$V = \frac{1}{\sum_{j\in J} \frac{\sum_{t=1}^{n_j} X_j(t)^2 - (\sum_{t=1}^{n_j} X_j(t))^2/n_j}{(\hat{\lambda}_j)^2}}. \tag{34}$$

A symmetric $(1-\epsilon)$-confidence interval for β is then provided by

$$\beta \in I := \left[\hat{\beta} - g_{1-\epsilon/2}\sqrt{V}, \hat{\beta} + g_{1-\epsilon/2}\sqrt{V}\right]. \tag{35}$$

Similarly, $(1-\epsilon)$-confidence one-sided intervals for β are obtained by letting

$$\beta \in I' := [\hat{\beta} - g_{1-\epsilon}\sqrt{V}, +\infty),\ \beta \in I'' := (-\infty, \hat{\beta} + g_{1-\epsilon}\sqrt{V}]. \tag{36}$$

The doubling time δ is given by $(\log 2)/\beta$ if $\beta > 0$, and $+\infty$ otherwise. This gives the $1-\epsilon$-confidence conditions for δ:

$$\text{if } \hat{\beta} - g_{1-\epsilon}\sqrt{V} > 0,\ \delta \in I_1 = \left[0, \frac{\log 2}{\hat{\beta} - g_{1-\epsilon}\sqrt{V}}\right], \tag{37}$$

and

$$\delta \in I_2 = \left[\frac{\log 2}{\max(0, \hat{\beta} + g_{1-\epsilon}\sqrt{V})}, +\infty\right). \tag{38}$$

For concreteness assume we want to raise an alarm when δ is δ^* days or less, where δ^* could be 10. From the above consideration, δ is below δ^* days with confidence $1-\epsilon$ when

$$\frac{\log 2}{\delta^*} < \hat{\beta} - g_{1-\epsilon}\sqrt{V}.$$

Raising an alarm under this condition then amounts to calibrating the false positive probability at ϵ.

Alternatively, we may consider to raise an alarm under the condition

$$\frac{\log 2}{\delta^*} < \hat{\beta} + g_{1-\epsilon}\sqrt{V}.$$

This would correspond to calibrating the false negative probability (probability of not raising an alarm while $\delta \leqslant \delta^*$) at ϵ.

10 Conclusion

We have shown that monitoring of emergency calls to EMS allows to anticipate the evolution of an epidemic by providing several *early signals*, each with specific characteristics in terms of time lag and reliability.

Our study illustrates the spatially differentiated nature of the epidemic kinetics, with significant doubling time differences between neighboring departments.

Such spatial differentiation, if present at a granularity finer than that of departments considered here, could be exploited using the methods described in the present work in order to detect potential epidemic resurgences at the corresponding spatial granularity. This shows great promise in enabling detection of so-called epidemic clusters.

There is thus huge potential in the extension of this work and its application to finer spatial resolution.

Notwithstanding such extensions, monitoring epidemic kinetics through EMS calls at regional levels can already be exploited to define region-specific sanitary measures, such as lifting of travel bans, proportionate to the regional situation, and to allow early detection of epidemic resurgence. Importantly, we expect this finding to be applicable in full generality to EMS organizations worldwide. Thus the methods introduced here may be of wide applicability to combat Covid-19. Beyond Covid-19, EMS organizations have a unique role to play in early detection of sanitary crises.

11 Acknowledgments

We thank the operational team of DSI of AP-HP, who helped to extract information records, especially Stéphane Crézé, Laurent Fontaine, Pierre Cabot, François Planeix, Fabrice Tordjman, Grégory Terrell and Martine Spiegelmann.

We thank Pr. Renaud Piarroux for very helpful remarks. We thank Pr. Bruno Riou for his suggestion to include quantitative statistical estimates in the present article. We thank Pr. Frédéric Batteux for having provided epidemiological information. We thank Dr. François Braun (SAMU 57) and Dr. Vincent Bounes (SAMU 31) for providing comparison elements between their departments. We thank Dr. Nicolas Poirot for introducing us to SAMU 31. We thank Dr. Paul-Georges Reuter (SAMU 92) for useful comments on the interpretation of SAMU data relative to the Covid crisis.

We thank Ayoub Foussoul, for having developed a robust dynamic programming algorithm, allowing one to consolidate the results of this manuscript concerning the best piecewise linear approximation of the log of observables. We thank Jérôme Bolte, for providing insights on non-convex and non-smooth best-approximation problems.

We thank Tania Lasisz for her help in the administration of the project, and Guillermo Andrade Barroso, Thomas Calmant and Matthieu Simonin for their contribution to software development.

We thank NXO France Integrator of communication solutions team and SIS Centaure15 solution from GFI World team for the help they provided and their availability for the project.

We thank Orange Flux Vision (especially Jean-Michel Contet) for having provided daily population estimates, at the scale of the department, helping to calibrate our models.

We thank Enedis (especially Pierre Gotelaere and his team) for having provided an estimation of the departure rate of households, aggregated at the scale of departments and districts, helping us to refine our model.

We thank SFR Geostatistic Team (especially Loic Lelièvre) for having provided estimates of flows between Paris and province, aggregated at the scale of departments and districts, allowing us to incorporate mobility in our model.

Stéphane Gaubert thanks Nicolas Bacaër for a decisive help, concerning epidemiological and mathematical analysis, provided during the week of March 16th-20th. He thanks Cormac Walsh for improvements of the text. He also thanks Thomas Lepoutre for very helpful mathematical comments and suggestions concerning Section 7.

The INRIA–École polytechnique team thanks the Direction de Programme de la Plate Forme d'Appels d'Urgences – PFAU at Préfecture de Police, DOSTL (Régis Reboul), and Brigade de Sapeurs Pompiers de Paris (especially Gen. Jean-Marie Gontier and Capt. Denis Daviaud) for having provided precious elements of comparison concerning the calls received at the emergency numbers 17-18-112.

References

[1] D. M. Anderson, N. A. Ciletti, H. Lee-Lewis, D. Elli, J. Segal, K. L. DeBord, K. A. Overheim, M. Tretiakova, R. R. Brubaker, and O. Schneewind. Pneumonic plague pathogenesis and immunity in brown norway rats. *Am. J. Pathol.*, 174(3):910–921, 2009.

[2] Andrew S. Azman, Kara E. Rudolph, Derek A. T. Cummings, and Justin Lessler. The incubation period of cholera: A systematic review. *J. Infect.*, 66(5):432–438, 2013.

[3] N. Bacaer. Un modèle mathématique des débuts de l'épidémie de coronavirus en France. hal-02509142, 2020.

[4] D. J Baker, C. Télion, and P. Carli. Multiple casualty incidents: the prehospital role of the anesthesiologist in europe. *Anesthesiology clinics*, 25(1):179–188, 2007.

[5] V. Bansaye, B. Cloez, and P. Gabriel. Ergodic behavior of non-conservative semigroups via generalized doeblin's conditions. *Acta Applicandae Mathematicae*, 166:29–72, 2020.

[6] R. Bellman and R. Roth. Curve fitting by segmented straight lines. *J. Am. Stat. Assoc.*, 64(327):1079–1084, 1969.

[7] A. Berman and R.J. Plemmons. *Nonnegative matrices in the mathematical sciences*. Academic Press, 1979.

[8] G. C. Calafiore, S. Gaubert, and C. Possieri. Log-sum-exp neural networks and posynomial models for convex and log-log-convex data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(3):827–838, 2020.

[9] Yi-Cheng Chen, Ping-En Lu, Cheng-Shang Chang, and Tzu-Hsuan Liu. A time-dependent SIR model for COVID-19 with undetectable infected persons. `http://gibbs1.ee.nthu.edu.tw/A_TIME_DEPENDENT_SIR_MODEL_FOR_COVID_19.PDF`, 2020.

[10] G. Cohen, S. Gaubert, and J.-P. Quadrat. Duality and separation theorems in idempotent semimodules. *Linear Algebra and Appl.*, 379:395–422, 2004.

[11] H. Cramér. *Mathematical methods of statistics*, volume 43. Princeton university press, 1999.

[12] Agence Régionale de Santé Grand Est. Coronavirus Covid 19 en Grand Est: Point de situation. Press release of March 8$^{\text{th}}$, available from `https://www.grand-est.ars.sante.fr/system/files/2020-03/Covid19_point_GrandEst080320.pdf`, 2020.

[13] L. Di Domenico, G. Pullano, Ch. E. Sabbatini, P.-Y. Boëlle, and V. Colizza. Expected impact of lockdown in Île-de-France and possible exit strategies. Report #9 , www.epicx-lab.com/covid-19.html, 2020.

[14] E.B. Dynkin. Boundary theory of Markov processes (the discrete case). *Russian Math. Surveys*, 24(7):1–42, 1969.

[15] Santé Publique France. Données hospitalières relatives à l'épidémie de Covid-19, 2020. `https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/`, Retrieved on May 10th, 2020.

[16] S. Gaubert and Z. Qu. Dobrushin's ergodicity coefficient for Markov operators on cones. *Integral Equations and Operator Theory*, 81(1):127–150, 2015.

[17] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[18] M. Hirsch, P. Carli, R. Nizard, B. Riou, B. Baroudjian, Th. Baubet, V. Chhor, Ch. Chollet-Xemard, N. Dantchev, N. Fleury, J.-P. Fontaine, Y. Yordanov, M. Raphael, C. Paugam-Burtz, L. Lafont, and health professionals of AP-HP. The medical response to multisite terrorist attacks in paris. *The Lancet*, 386(10012):2535–2538, 2015.

[19] Insee. Ménages selon la taille en 2016. Comparaisons régionales et départementales, 2019. `https://www.insee.fr/fr/statistiques/2012714`.

[20] I. Itenberg, G. Mikhalkin, and E. Shustin. *Tropical algebraic geometry.* Oberwolfach seminars. Birkhäuser, 2007.

[21] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions.* Wiley, New York, 1994.

[22] W. O Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A*, 115:700–721, 1927.

[23] R. W Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[24] J. C. Lagarias, J. A. Reeds, M. H. Wright, , and P. E. Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9(1):112–147, 1998.

[25] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, and J. Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9):577–582, 2020. PMID: 32150748.

[26] B. Lemmens and R. Nussbaum. *Nonlinear Perron-Frobenius Theory*, volume 189 of *Cambridge Tracts in Mathematics*. Cambridge University Press, May 2012.

[27] C. Massonnaud, J. Roux, and P. Crépey. Covid-19: Forecasting short term hospital needs in France. Report available from Sfar.org, 2020.

[28] P. Michel, S. Mischler, and B. Perthame. General relative entropy inequality: an illustration on growth models. *J. Math. Pures et Appl.*, 84(9):1235–1260, May 11 2005.

[29] S. Mischler and J. Scher. Spectral analysis of semigroups and growth-fragmentation equations. *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, 33(3):849 – 898, 2016.

[30] A. Papadopoulos and M. Troyanov. Weak Finsler structures and the Funk weak metric. *Math. Proc. Cambridge Philos. Soc.*, 147(2):419–437, 2009.

[31] B. Perthame. *Transport equations in biology.* Birkhäuser, 2007.

[32] H. Salje, C. Tran Kiem, N. Lefrancq, N. Courtejoie, P. Bosetti, J. Paireau, A. Andronico, N. Hoze, J. Richet, C.-L. Dubost, Y. Le Strat, J. Lessler, D. Bruhl, A. Fontanet, L. Opatowski, P.-Y. Boëlle, and S. Cauchemez. Estimating the burden of SARS-CoV-2 in France. pasteur-02548181, 2020.

[33] G. Stangenhaus, S. C. Narula, and F. F. Pedro. Bootstrap confidence intervals for the minimum sum of absolute errors regression. *Journal of statistical computation and simulation*, 48(3-4):127–133, 1993.

[34] Virlogeux V., Fang V. J., Park M., Wu J. T., and Cowling B. J. Comparison of incubation period distribution of human infections with mers-cov in south korea and saudi arabia. *Sci. Rep.*, 6(35839), 2016.

[35] M. Varia, S. Wilson, Sh. Sarwal, A. McGeer, E. Gournis, Eleni Galanis, B. Henry, and Hospital Outbreak Investigation Team. Investigation of a nosocomial outbreak of severe acute respiratory syndrome (sars) in toronto, canada. *CMAJ*, 169(4):285–292, 2003.

[36] O. Viro. Dequantization of real algebraic geometry on logarithmic paper. In *European Congress of Mathematics, Vol. I (Barcelona, 2000)*, volume 201 of *Progr. Math.*, pages 135–146. Birkhäuser, Basel, 2001.

[37] H. von Forster. Some remarks on changing populations. In Jr. F. Stohlman, editor, *The Kinetics of Cellular Proliferation*, pages 382–407. Grune & Stratton, New York, 1959.

A Appendix: algorithms to compute a best approximation of the logarithm of the number of events by a piecewise linear map

Given an epidemiologic observable $Y(t)$, we need to approximate $\log Y(t)$ by a function

$$\mathcal{L}(t) := \min_{1 \leqslant j \leqslant \nu} (\lambda_j t + c_j)\,,$$

where ν is the number of phases with constant sanitary policy during the considered time period. The parameters λ_j, c_j are assumed without loss of generality to satisfy $\lambda_1 \leqslant \lambda_2 \cdots \leqslant \lambda_\nu$. The concavity constraint imposed on the approximating function $\mathcal{L}(t)$ makes the problem different from standard function approximation problems, and contributes to the robustness of the fitting procedure by reducing the amount of overfitting.

The two most natural criteria for fitting function $\mathcal{L}(t)$ to observations $\log Y(t)$ are to minimize either a least squares, or ℓ_2 loss function $\sum_{t\in\mathcal{T}} |\mathcal{L}(t) - \log Y(t)|^2$, or an ℓ_1 loss function $\sum_{t\in\mathcal{T}} |\mathcal{L}(t) - \log Y(t)|$, where $\mathcal{T}$ is a finite set of time instants at which observations have been made. As discussed in Section 9, the ℓ_1 formulation is more robust in being less sensitive to outliers, and is the one used on Figure 4.

The corresponding optimization problem over parameters λ_i, c_i is non-convex as soon as $\nu \geqslant 2$. A straightforward option is to use a derivative free procedure, like the Nelder-Mead [24] algorithm. Depending on the initial point, this algorithm may converge to a local minimum, which may not be epidemiologically significant. So, a possibility is to guide the algorithm by providing it a initial guess of the optimal solution. To do, we start by an a priori selection of the time periods over which function $\mathcal{L}(t)$ is linear (which could be obtained by prior knowledge of delay parameters τ and times of policy changes, or found by brute force search). We then determine a minimum cost linear fit of target function $\log Y(t)$ over each such period, and use the concave envelope of the resulting function as our initial condition for local search. This is how we initially obtained the best ℓ_1 approximation shown on Figure 4. We also used CMA-ES for comparison [17]. Both Nelder-Mead and CMA-ES algorithms appear to be sensitive to the initial conditions. Notice in this respect that the objective function is linear on the cells of a polyhedral complex and that it can be constant on certain unbounded cells of this complex, so a local search algorithm may be trapped in a cell in which the function is constant. Another perspective is to observe that this best approximation problem is equivalent to a learning problem, looking for the parameters of a neural networks with a single hidden layer and min-type activation functions, see [8]. This allows one to apply (nonsmooth) optimization algorithms used in learning, still leading in general to a local optimum. An approach leading to the global optimum is dynamic programming, originating from Bellman [6]. Ayoub Foussoul (École polytechnique) provided us with a dynamic programming solver, implementing several refinements, and allowing us to certify the global optimality of the approximation shown in Figure 4, up to a fixed precision.

Bibliography

Achaz, G., Rodriguez-Verdugo, A., Gaut, B. S., and Tenaillon, O. (2014). The Reproducibility of Adaptation in the Light of Experimental Evolution with Whole Genome Sequencing. In Landry, C. R. and Aubin-Horth, N., editors, *Ecological Genomics*, volume 781, pages 211–231. Springer Netherlands, Dordrecht. (cited on pages ii and 2)

Adami, C. (2006). Digital genetics: Unravelling the genetic basis of evolution. *Nature Reviews Genetics*, 7(2):109–118. (cited on page 10)

Adami, C. and Brown, C. T. (1994). Evolutionary Learning in the 2D Artificial Life System "Avida". https://arxiv.org/abs/adap-org/9405003. (cited on page 16)

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., and Mori, H. (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: The Keio collection. *Molecular Systems Biology*, 2(1). (cited on page 67)

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. (cited on pages 16 and 111)

Beslon, G., Liard, V., Parsons, D. P., and Rouzaud-Cornabas, J. (2021). Of Evolution, Systems and Complexity. In Crombach, A., editor, *Evolutionary Systems Biology: Advances, Questions, and Opportunities*, pages 1–18. Springer International Publishing, Cham. (cited on pages iii, 3, 10, and 11)

Blattner, F. R. (1997). The Complete Genome Sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462. (cited on pages 9 and 53)

Brackley, C. A., Johnson, J., Bentivoglio, A., Corless, S., Gilbert, N., Gonnella, G., and Marenduzzo, D. (2016). Stochastic Model of Supercoiling-Dependent Transcription. *Physical Review Letters*, 117(1):018101. (cited on pages 10 and 67)

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. (cited on pages 7, 67, 68, and 109)

Cameron, A. D. S. and Dorman, C. J. (2012). A Fundamental Regulatory Mechanism Operating through OmpR and DNA Topology Controls Expression of Salmonella Pathogenicity Islands SPI-1 and SPI-2. *PLoS Genetics*, 8(3):e1002615. (cited on page 51)

Cameron, A. D. S., Kröger, C., Quinn, H. J., Scally, I. K., Daly, A. J., Kary, S. C., and Dorman, C. J. (2013). Transmission of an Oxygen Availability Signal at the Salmonella enterica Serovar Typhimurium fis Promoter. *PLoS ONE*, 8(12):e84382. (cited on page 44)

Crick, F. H. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163. (cited on page 14)

Crombach, A. and Hogeweg, P. (2008). Evolution of Evolvability in Gene Regulatory Networks. *PLoS Computational Biology*, 4(7):e1000112. (cited on page 68)

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. (cited on pages 8, 13, 19, 37, 44, 53, 91, and 98)

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. (cited on pages ii, 2, 7, 14, 44, 91, and 108)

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray. (cited on pages i and 1)

Darwin, C. (1868). *The Variation of Animals and Plants under Domestication*. John Murray. (cited on pages i and 1)

Dávila López, M., Martínez Guerra, J. J., and Samuelsson, T. (2010). Analysis of Gene Order Conservation in Eukaryotes Identifies Transcriptionally and Functionally Linked Genes. *PLoS ONE*, 5(5):e10654. (cited on page 74)

de la Campa, A. G., Ferrándiz, M. J., Martín-Galiano, A. J., García, M. T., and Tirado-Vélez, J. M. (2017). The Transcriptome of Streptococcus pneumoniae Induced by Local and Global Changes in Supercoiling. *Frontiers in Microbiology*, 8:1447. (cited on page 68)

Di Cosmo, R. (2020). Archiving and Referencing Source Code with Software Heritage. In Bigatti, A. M., Carette, J., Davenport, J. H., Joswig, M., and de Wolff, T., editors, *Mathematical Software – ICMS 2020*, volume 12097, pages 362–373. Springer International Publishing, Cham. (cited on page 28)

Dorman, C. J. (2019). DNA supercoiling and transcription in bacteria: A two-way street. *BMC Molecular and Cell Biology*, 20(1):26. (cited on page 68)

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. (cited on pages 5, 34, and 44)

Duprey, A. and Groisman, E. A. (2021). The regulation of DNA supercoiling across evolution. *Protein Science*, page pro.4171. (cited on pages 7 and 44)

El Hanafi, D. and Bossi, L. (2000). Activation and silencing of leu-500 promoter by transcription-induced DNA supercoiling in the Salmonella chromosome: Transcription-dependent modulation of leu-500 promoter in topA mutants. *Molecular Microbiology*, 37(3):583–594. (cited on pages 8 and 70)

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. (cited on pages 6, 9, 30, 44, 50, 53, 67, and 110)

Felsenstein, J. (2019). *Theoretical Evolutionary Genetics*. https://evolution.genetics.washington.edu/pgbook/pgbook.html. (cited on page 20)

Ferrandiz, M.-J., Martin-Galiano, A. J., Schvartzman, J. B., and de la Campa, A. G. (2010). The genome of Streptococcus pneumoniae is organized in topology-reacting gene clusters. *Nucleic Acids Research*, 38(11):3570–3581. (cited on pages 6, 8, and 109)

Forquet, R., Nasser, W., Reverchon, S., and Meyer, S. (2022). Quantitative contribution of the spacer length in the supercoiling-sensitivity of bacterial promoters. *Nucleic Acids Research*, 50(13):7287–7297. (cited on pages 67 and 110)

Forquet, R., Pineau, M., Nasser, W., Reverchon, S., and Meyer, S. (2021). Role of the Discriminator Sequence in the Supercoiling Sensitivity of Bacterial Promoters. *mSystems*, 6(4). (cited on pages 8, 44, 49, 66, and 110)

Gardner, T. S., Cantor, C. R., and Collins, J. J. (2000). Construction of a genetic toggle switch in Escherichia coli. *Nature*, 403(6767):339–342. (cited on pages 10, 59, and 67)

Gaubert, S., Akian, M., Allamigeon, X., Boyet, M., Colin, B., Grohens, T., Massoulié, L., Parsons, D. P., Adnet, F., Chanzy, É., Goix, L., Lapostolle, F., Lecarpentier, É., Leroy, C., Loeb, T., Marx, J.-S., Télion, C., Tréluyer, L., and Carli, P. (2020). Understanding and monitoring the evolution of the Covid-19 epidemic from medical emergency calls: The example of the Paris area. *Comptes Rendus. Mathématique*, 358(7):843–875. (cited on pages iv, 4, and 115)

Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., Bibbs, L., Eads, J., Richardson, T. H., Noordewier, M., Rappé, M. S., Short, J. M., Carrington, J. C., and Mathur, E. J. (2005). Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science*, 309(5738):1242–1245. (cited on page 74)

Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M. R., Maruf, M., Hutchison, C. A., Smith, H. O., and Venter, J. C. (2006). Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences*, 103(2):425–430. (cited on page 82)

Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E., and Desai, M. M. (2017). The dynamics of molecular evolution over 60,000 generations. *Nature*, 551(7678):45–50. (cited on page 13)

Grohens, T., Meyer, S., and Beslon, G. (2021). A Genome-Wide Evolutionary Simulation of the Transcription-Supercoiling Coupling. In *The 2021 Conference on Artificial Life*. MIT Press. (cited on pages iii, 3, 6, 52, 66, and 112)

Grohens, T., Meyer, S., and Beslon, G. (2022a). Emergence of Supercoiling-Mediated Regulatory Networks through Bacterial Chromosome Rearrangements. https://www.biorxiv.org/content/10.1101/2022.09.23.509185v1. (cited on pages iv, 4, 47, and 112)

Grohens, T., Meyer, S., and Beslon, G. (2022b). A Genome-Wide Evolutionary Simulation of the Transcription-Supercoiling Coupling. *Artificial Life*, pages 1–18. (cited on pages iii, 3, 27, and 112)

Hérault, E., Reverchon, S., and Nasser, W. (2014). Role of the LysR-type transcriptional regulator PecT and DNA supercoiling in the thermoregulation of *pel* genes, the major virulence factors in *Dickeya dadantii*: *Dickeya dadantii* PecT protein and virulence thermoregulation. *Environmental Microbiology*, 16(3):734–745. (cited on pages 7, 44, and 51)

Hindré, T., Knibbe, C., Beslon, G., and Schneider, D. (2012). New insights into bacterial adaptation through in vivo and in silico experimental evolution. *Nature Reviews Microbiology*, 10(5):352–365. (cited on page 11)

Hsieh, L. S., Rouviere-Yaniv, J., and Drlica, K. (1991). Bacterial DNA supercoiling and [ATP]/[ADP] ratio: Changes associated with salt shock. *Journal of Bacteriology*, 173(12):3914–3917. (cited on page 7)

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95. (cited on page 113)

Johnstone, C. P. and Galloway, K. E. (2022). Supercoiling-mediated feedback rapidly couples and tunes transcription. https://www.biorxiv.org/content/10.1101/2022.04.20.488937v1. (cited on pages 67 and 68)

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. (cited on pages 8, 67, 68, and 109)

Knibbe, C., Beslon, G., Lefort, V., Chaudier, F., and Fayard, J. M. (2005). Self-adaptation of Genome Size in Artificial Organisms. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Capcarrère, M. S., Freitas, A. A., Bentley, P. J., Johnson, C. G., and Timmis, J., editors, *Advances in Artificial Life*, volume 3630, pages 423–432. Springer Berlin Heidelberg, Berlin, Heidelberg. (cited on pages 16 and 111)

Kouzine, F., Gupta, A., Baranello, L., Wojtowicz, D., Ben-Aissa, K., Liu, J., Przytycka, T. M., and Levens, D. (2013). Transcription-dependent dynamic supercoiling is a short-range genomic force. *Nature Structural & Molecular Biology*, 20(3):396–403. (cited on page 70)

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. (cited on page 7)

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. (cited on page 44)

Lam, S. K., Pitrou, A., and Seibert, S. (2015). Numba: A LLVM-based Python JIT compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC - LLVM '15*, pages 1–6, Austin, Texas. ACM Press. (cited on pages 53 and 113)

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. (cited on pages ii, 2, 7, 13, and 107)

Levy, S. B. and Marshall, B. (2004). Antibacterial resistance worldwide: Causes, challenges and responses. *Nature Medicine*, 10(S12):S122–S129. (cited on pages i and 1)

Liard, V. (2020). *Origine Évolutive de La Complexité Des Systèmes Biologiques : Une Étude Par Évolution Expérimentale in Silico*. Thèses, Université de Lyon. (cited on page 16)

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. (cited on pages iii, 3, 5, 7, 8, and 66)

Ma, J. and Wang, M. D. (2016). DNA supercoiling during transcription. *Biophysical Reviews*, 8(S1):75–87. (cited on page 44)

Marshall, D. G., Bowe, F., Hale, C., Dougan, G., and Dorman, C. J. (2000). DNA topology and adaptation of Salmonella Typhimurium to an intracellular environment. *Phil. Trans. R. Soc. Lond. B*, page 10. (cited on page 7)

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. (cited on pages iii, 3, 7, 44, 45, and 66)

Menzel, R. and Gellert, M. (1987). Modulation of transcription by DNA supercoiling: A deletion analysis of the Escherichia coli gyrA and gyrB promoters. *Proceedings of the National Academy of Sciences*, 84(12):4185–4189. (cited on page 66)

Meyer, S. and Beslon, G. (2014). Torsion-Mediated Interaction between Adjacent Genes. *PLoS Computational Biology*, 10(9):e1003785. (cited on pages 8, 9, 44, and 67)

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. (cited on pages 6 and 34)

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., McDonald, L., Utterback, T. R., Malek, J. A., Linher, K. D., Garrett, M. M., Stewart, A. M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D., Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., White, O., Salzberg, S. L., Smith, H. O., Venter, J. C., and Fraser, C. M. (1999). Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima. *Nature*, 399(6734):323–329. (cited on page 73)

Ofria, C. and Wilke, C. O. (2004). Avida: A Software Platform for Research in Computational Evolutionary Biology. *Artificial Life*, 10(2):191–229. (cited on page 16)

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. (cited on pages 6, 8, 29, 44, and 109)

Pineau, M., Martis B., S., Forquet, R., Baude, J., Villard, C., Grand, L., Popowycz, F., Soulère, L., Hommais, F., Nasser, W., Reverchon, S., and Meyer, S. (2022). What is a supercoiling-sensitive gene? Insights from topoisomerase I inhibition in the Gram-negative bacterium *Dickeya dadantii*. *Nucleic Acids Research*, 50(16):9149–9161. (cited on pages 6, 66, 109, and 110)

Postow, L., Hardy, C. D., Arsuaga, J., and Cozzarelli, N. R. (2004). Topological domain structure of the Escherichia coli chromosome. *Genes & Development*, 18(14):1766–1779. (cited on pages 7, 9, 29, 53, 70, and 73)

Rhee, K. Y., Opel, M., Ito, E., Hung, S.-p., Arfin, S. M., and Hatfield, G. W. (1999). Transcriptional coupling between the divergent promoters of a prototypic LysR-type regulatory system, the ilvYC operon of Escherichia coli. *Proceedings of the National Academy of Sciences*, 96(25):14294–14299. (cited on page 9)

Rutten, J. P., Hogeweg, P., and Beslon, G. (2019). Adapting the engine to the fuel: Mutator populations can reduce the mutational load by reorganizing their genome structure. *BMC Evolutionary Biology*, 19(1):191. (cited on pages 14, 45, and 111)

Schneiker, S., Perlova, O., Kaiser, O., Gerth, K., Alici, A., Altmeyer, M. O., Bartels, D., Bekel, T., Beyer, S., Bode, E., Bode, H. B., Bolten, C. J., Choudhuri, J. V., Doss, S., Elnakady, Y. A., Frank, B., Gaigalat, L., Goesmann, A., Groeger, C., Gross, F., Jelsbak, L., Jelsbak, L., Kalinowski, J., Kegler, C., Knauber, T., Konietzny, S., Kopp, M., Krause, L., Krug, D., Linke, B., Mahmud, T., Martinez-Arias, R., McHardy, A. C., Merai, M., Meyer, F., Mormann, S., Muñoz-Dorado, J., Perez, J., Pradella, S., Rachid, S., Raddatz, G., Rosenau, F., Rückert, C., Sasse, F., Scharfe, M., Schuster, S. C., Suen, G., Treuner-Lange, A., Velicer, G. J., Vorhölter, F.-J., Weissman, K. J., Welch, R. D., Wenzel, S. C., Whitworth, D. E., Wilhelm, S.,

Wittmann, C., Blöcker, H., Pühler, A., and Müller, R. (2007). Complete genome sequence of the myxobacterium Sorangium cellulosum. *Nature Biotechnology*, 25(11):1281–1289. (cited on page 82)

Sevier, S. A. and Hormoz, S. (2021). Collective polymerase dynamics emerge from DNA supercoiling during transcription. https://www.biorxiv.org/content/10.1101/2021.11.24.469850v3. (cited on pages 10, 67, and 110)

Sevier, S. A. and Levine, H. (2017). Mechanical Properties of Transcription. *Physical Review Letters*, 118(26):268101. (cited on page 10)

Sevier, S. A. and Levine, H. (2018). Properties of gene expression and chromatin structure with mechanically regulated elongation. *Nucleic Acids Research*, 46(12):5924–5934. (cited on page 10)

Sobetzko, P. (2016). Transcription-coupled DNA supercoiling dictates the chromosomal arrangement of bacterial genes. *Nucleic Acids Research*, 44(4):1514–1524. (cited on pages 8 and 9)

Thomas, G. H., Newbern, E. C., Korte, C. C., Bales, M. A., Muse, S. V., Clark, A. G., and Kiehart, D. P. (1997). Intragenic duplication and divergence in the spectrin superfamily of proteins. *Molecular Biology and Evolution*, 14(12):1285–1295. (cited on pages ii and 2)

Tobe, T., Yoshikawa, M., and Sasakawa, C. (1995). Thermoregulation of virB transcription in Shigella flexneri by sensing of changes in local DNA superhelicity. *Journal of Bacteriology*, 177(4):1094–1097. (cited on page 8)

Toh, H., Weiss, B. L., Perkin, S. A., Yamashita, A., Oshima, K., Hattori, M., and Aksoy, S. (2006). Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Research*, 16(2):149–156. (cited on page 73)

Travers, A. and Muskhelishvili, G. (2005). DNA supercoiling — a global transcriptional regulator for enterobacterial growth? *Nature Reviews Microbiology*, 3(2):157–169. (cited on page 5)

Visser, B. J., Sharma, S., Chen, P. J., McMullin, A. B., Bates, M. L., and Bates, D. (2022). Psoralen mapping reveals a bacterial genome supercoiling landscape dominated by transcription. *Nucleic Acids Research*, 50(8):4436–4449. (cited on pages iii, 3, 7, 8, 9, 66, and 70)

Webber, M. A., Ricci, V., Whitehead, R., Patel, M., Fookes, M., Ivens, A., and Piddock, L. J. V. (2013). Clinically Relevant Mutant DNA Gyrase Alters Supercoiling, Changes the Transcriptome, and Confers Multidrug Resistance. *mBio*, 4(4). (cited on pages 6 and 8)

Wortel, M. T., Agashe, D., Bailey, S. F., Bank, C., Bisschop, K., Blankers, T., Cairns, J., Colizzi, E. S., Cusseddu, D., Desai, M. M., van Dijk, B., Egas, M., Ellers, J., Groot, A. T., Hackel, D. G., Johnson, M. L., Kraaijeveld, K., Krug, J., Laan, L., Laessig, M., Lind, P. A., Meijer, J., Noble, L. M., Okasha, S., Rainey, P. B., Rozen, D. E., Shitut, S., Tans, S. J., Tenaillon, O., Teotonio,

H., de Visser, J. A. G. M., Visser, M. E., Vroomans, R. M. A., Werner, G. D. A., Wertheim, B., and Pennings, P. S. (2021). Towards evolutionary predictions: Current promises and challenges. https://ecoevorxiv.org/4u3mg/. (cited on page i)

FOLIO ADMINISTRATIF

THÈSE DE L'INSA LYON, MEMBRE DE L'UNIVERSITÉ DE LYON

NOM : Grohens

DATE DE SOUTENANCE : 14/12/2022

Prénom : Théotime

TITRE : M.

NATURE : Doctorat

Numéro d'ordre : 2022ISAL0126

École Doctorale : InfoMaths (ED 512)

Spécialité : Informatique et Applications

RÉSUMÉ : Evolution is often considered an unpredictable process, as genetic mutations happen at random. But the fixation of mutations is not completely arbitrary, as mutations need to pass the sieve of natural selection to be retained. In particular, the beneficial or deleterious character of a mutation can depend on the genetic background in which it happens, an effect called epistasis. In this work, I study a particular kind of epistatic interactions in bacteria: the interplay between mutations in the mechanisms regulating DNA supercoiling -- the level of over- or under- winding of DNA -- and genomic rearrangements.

I present *EvoTSC*, a mathematical and computational model of DNA supercoiling tailored to study the mutual interaction between gene transcription and DNA supercoiling (the transcription-supercoiling coupling or TSC), and integrated into a full-fledged evolutionary simulation. I first validate the model by showing that evolution can leverage this coupling to evolve gene regulatory networks that are able to tune gene expression levels in response to environmental perturbations, by changing only the relative positions of the genes through genomic inversions. I then show that, in *EvoTSC* as well as in the evolutionary simulation platform *Aevol*, introducing supercoiling mutations does not seem to speed up evolution, indicating that the evolutionary relevance of epistatic interactions might be not as important as initially thought. Using *EvoTSC*, I additionally show that the TSC can lead some genes to be activated by an excess of positive supercoiling, providing a plausible mechanism to explain the similar behavior observed in many bacterial genes. Finally, I characterize the structure of these supercoiling-mediated gene regulatory networks, showing that they cannot be reduced to local pairwise interactions. Interaction with many neighboring genes can indeed be needed to regulate gene expression through supercoiling, providing a possible explanation to the evolutionary conservation of gene syntenies.

MOTS-CLÉS : evolution, epistasis, DNA supercoiling, gene transcription

Laboratoire de recherche : Laboratoire d'informatique en image et systèmes d'information (LIRIS)

Directeur de thèse: Guillaume Beslon

Président de jury : xxx

Composition du jury :

Scornavacca, Céline	Directrice de recherche	CNRS	Rapportrice
Junier, Ivan	Chargé de Recherche HDR	CNRS	Rapporteur
Varoquaux, Nelle	Chargée de Recherche	CNRS	Examinatrice
Nasser, William	Directeur de Recherche	CNRS	Examinateur
Achaz, Guillaume	Professeur des Universités	Université Paris-Cité	Examinateur
Meyer, Sam	Maître de Conférences	INSA Lyon	Examinateur
Beslon, Guillaume	Professeur des Universités	INSA Lyon	Directeur de thèse