

DOCTORAL THESIS of
UNIVERSITÉ PIERRE ET MARIE CURIE and
UNIVERSITY OF MANIPAL

École Doctorale de Physique EDPIF (Paris),
Biologie Computationnelle et Quantitative (Paris) and
National Centre for Biological Sciences (Bangalore)

Physics of bacterial nucleoid organization and large-scale gene expression

Doctoral thesis in Physics
of Vittore Ferdinando Scolari

Under the supervision of
M. Marco Cosentino Lagomarsino and
M. Aswin Sai Narain Seshasayee

Defended on 15 October 2014

in front of a jury composed of :

M. Marco Cosentino Lagomarsino	Directeur de thèse
M. Aswin Sai Narain Seshasayee	Directeur de thèse
M. Enrico Carlon	Rapporteur
M. Andrea Parmeggiani	Rapporteur
M. Jean-François Joanny	Examinateur
M. Julien Mozziconacci	Examinateur
M. Bianca Sclavi	Examinateur

© 2014
Vittore Ferdinando Scolari
ALL RIGHTS RESERVED

Version purged of copyrighted material not owned by authors
For all figures and logos: `http://www.lcqb.upmc.fr/scolari/thesis_figures.zip`

Printed in date
September 29, 2014

Abstract

The genomic DNA of bacteria exists in a complex and dynamic protein-rich state, which is highly organized at various length scales. This has implications for modulating (when not directly enabling) the core biological processes of replication, transcription and segregation. This thesis describes a work of physical modeling and data analysis focused on the E. coli genome organization, in the form of a nucleo-protein structure called "nucleoid", and on how nucleoid organization affects gene expression.

The first part of the work is a review of the recent experimental and theoretical advances quantifying the physical organization (compaction and geometry) of the bacterial chromosome with its complement of proteins (the nucleoid). In particular, we highlight the role that statistical and soft condensed matter physics play in describing this system of fundamental biological importance, specifically recalling classic and more modern tools from the theory of polymers.

A second part of the work discusses a simple polymer physics model inspired by two main features of the nucleoid: osmotic self-adhesion and protein bridging. This idealized situation is considered from the theoretical and simulation point of view. The results are summarized by a qualitative characterization of the phase diagram of this model which shows the general feature that distinct domains may form without the need for intra-specific interactions.

The thesis also covers several data analysis approaches to test possible connections between the physical organization of the nucleoid with gene expression (RNA-Seq) and protein binding (ChIP-Seq) datasets. This latter part contains a description of the NuST webtool, which consists of a database which collect datasets from past experiments and an implementation of simple multi scale statistical analysis tools. Additionally, starting from an accurate description of the pipeline for the analysis of raw gene expression data, we introduce a correlation study of a large number (about 300) of genome-wide expression data-sets, also compared to the outcome to the published genome interaction map (Hi-C) data. Finally, we discuss how the qualitative results of the polymer model presented in the previous part of the thesis may be compared with experimental data.

Résumé

L'ADN génomique des bactéries existe dans un complexe dynamique riche en protéines, qui est très bien organisé à différentes échelles de longueur. Cela a des implications pour la modulation (ou l'activation) des processus biologiques fondamentaux de la réplication, la transcription et la

ségrégation. Cette thèse décrit un travail de modélisation et d'analyse des données physiques en misant l'accent sur l'organisation du génome de la bactérie *E. coli*, sous la forme d'une structure forme de ADN et protéines appelée "nucléoïde", et sur la façon dont l'organisation du nucléoïde affecte l'expression des gènes.

La première partie du travail est une revue des progrès récents expérimentaux et théoriques quantifiant l'organisation physique (la géométrie et le compactage) du chromosome bactérien avec son complément de protéines (nucléoïde). En particulier, nous soulignons le rôle que la physique de la matière molle et la physique statistique jouent dans la description de ce système d'une importance fondamentale en biologie, en rappelant en particulier les outils classiques et plus modernes de la théorie des polymères.

Une deuxième partie de l'ouvrage traite d'un modèle simple de la physique des polymères inspiré par deux caractéristiques principales du nucléoïde: auto-adhérence osmotique et effet des protéines de pontage. Cette situation idéalisée est considéré du point de vue théorique et de la simulation sur ordinateur. Les résultats sont résumés par une caractérisation qualitative du diagramme de phase de ce modèle, qui montre la caractéristique générale que les nucléoïdes forment des domaines distincts sans avoir besoin d'interactions intra-spécifiques.

La thèse couvre également plusieurs approches d'analyse de données pour tester les connexions possibles entre l'organisation physique du nucléoïde avec l'expression des gènes (ARN-Seq) et des protéines (ChIP-Seq). Cette dernière partie contient une description de l'outil web NuST, qui se compose d'une base de données qui collecte des ensembles de données provenant d'expériences passées et une mise en œuvre d'un outil Web qui permet d'effectuer de simples analyses statistiques sur de multiples échelles. En outre, à partir d'une description précise de la procédure pour l'analyse de données d'expression génique, nous présentons une étude de corrélation d'un grand nombre (environ 300) de mesures d'expression de l'ensemble du génome dans différentes conditions de croissance, et nous le comparons avec les cartes d'interaction (Hi-C) spatiale entre le chromosome. Enfin, nous discutons de la façon dont les résultats qualitatifs du modèle de polymère présenté dans la partie précédente de la thèse peuvent être comparés avec les données expérimentales.

I dedicate this thesis to all the person which were on my side during this latest three years of doctoral studies.

I also want to thank some friends: Andrea Cavallone, Bogdan Mirauta, Daniele Auvin, Elisa Brambilla, Emanuele Bagnaschi, Giulia Malaguti, Lucilla Taddei, Maria Rita Fumagalli, Matteo Osella, Mina Zarei, Orso Maria Romano, Roberto Bondesan, Stefano Vezzoli, for your good company, the fun and interesting discussions during my life in Paris; Aalap Mogre, Avantika Lal, Deepti Chandraprakash, Manoj Kumar Sumathi, Oliver Muthmann, Parul Singh, Rahul Somavanshi, Rajalakshmi Srinivasan, Revathy Krishnamurthi, Saurabh Mahajan, Savita Chib, Supriya Khedkar, and all the other hosts in NCBS, Bangalore, for making my permanence in your country an amazing and insightful experience; finally a big thank to the whole of my family Angela, Pier Vittorio, Nonna Pina, Giuseppe, Tommaso, Luigi, Ambra, Emma, Rosi and Ettore for your immense patience, and to Elizaveta for your support and comprehension.

Thanks also to the organizers of the PDI–MSC, IRD France, which granted me the doctoral contract and to all my fellow students, and most importantly to my supervisors Marco and Aswin because you've been a great guide.

Author list

- The following authors contributed to Chapter 2: Vincenzo G. Benza, Bruno Bassetti, Kevin D. Dorfman, Krystyna Bromek, Pietro Cicuta and Marco Cosentino Lagomarsino
- The following authors contributed to Chapter 3: Marco Cosentino Lagomarsino
- The following authors contributed to Chapter 4: Marco Cosentino Lagomarsino, Mina Zarei, Matteo Osella, Rajalakshmi Srinivasan, Deepti Chandraprakash, Revathy Krishnamurthi, Parul Singh, Sandeep Krishna and Aswin Sai Narain Seshasayee

Contents

Preface i

Abstract . . . i

Dedication . . . iii

Author list . . . iv

1 Introduction 1

2 Physical descriptions of the bacterial nucleoid at large scales 3

2.1 Background . . . 3

2.2 Measurements . . . 6

2.3 Models . . . 19

2.4 Discussion. Hypotheses and paradoxes concerning nucleoid geometry and dynamics. . . . 22

2.5 Conclusions. . . . 26

3 Combined collapse by bridging and self-adhesion in a prototypical polymer model inspired by the bacterial nucleoid. 29

3.1 Introduction . . . 29

3.1.1 Methods . . . 31

3.2 Model . . . 36

3.3 Results . . . 37

3.4 Discussion and Conclusions. . . . 45

3.5 Appendix . . . 47

3.5.1 Role of loop interactions in the collapse due to sparse bridging. . . . 47

3.5.2 Scaling argument for the entropy of a star polymer. . . . 49

4 Interplay of global transcriptional regulation and chromosome organization of bacteria 51

4.1 Introduction . . . 51

4.1.1 Experimental methods, conceptual description . . . 52

4.2 NuST (Nucleoid Survey Tools) . . . 57
4.2.1 Availability and Implementation: . . . 57
4.2.2 Introduction . . . 57
4.2.3 Detailed description . . . 58
4.3 Epistatic effects of the homologous H-NS and StpA genes knock out . . . 62
4.3.1 Genomic analysis reveals epistatic silencing of "expensive" genes in *Escherichia coli* K-12 . . . 63
4.3.2 Converging evolutionary strategies redress global gene expression imbalances caused by the loss of a global gene silencing system in *E. coli* . . . 79
4.4 Hi-C data analysis . . . 95
4.4.1 Alignment and filtering . . . 96
4.4.2 Hi-C interaction map from available *E. coli* data . . . 97
4.4.3 Analysis on Caulobacter . . . 100
4.4.4 Comparison with large scale expression data . . . 102

5 Conclusions and outlook 107

Chapter 1

Introduction

The demarcation line between a "physical" and a "biological" system is rapidly becoming anachronistic, to the point that it possibly hinders research in both disciplines. This is particularly true in the context of gene expression, which is a fundamental process in biology at the molecular level, common to all life on earth: the genetic code is read out ("transcribed") from DNA and written into RNA, which, in the case of messenger RNA, is then translated into proteins. In order for the right number of proteins to be produced in response to changes in environment, internal states, and stimuli, all cells are capable of tightly regulating this sequence of events [1]. The physico-chemical implementation of this gene regulation process takes place primarily through specific DNA-binding interactions of transcription factors, which repress or promote transcription.

It is becoming very clear that the genome's conformational properties as a polymer come into play in the processes involved in the regulatory fine-tuning of gene expression, in particular its topological, chemical, geometric and mechanical properties. These properties can influence the activity of the transcription factors and can play a role in the coordination of a large scale cellular response. Evidence for this level of regulation has been found in the different kingdoms of life. We focus here on the efforts to describe the genome's physical state in the case of bacteria, where it is (perhaps surprisingly) less explored than for higher life forms.

Understanding this problem requires stringent, quantitative experiments with the standards of physics, together with up-to-date physical models and arguments. Since considerable knowledge has been developed in polymer science over the last 50 years, it is very tempting to try to apply this knowledge to understand the energy and time scales involved in maintaining the DNA at the same time compact and accessible inside a cell, and the role of its geometry, structure and compaction in gene regulation.

BOX 1: Main factors affecting nucleoid organization.

- **Supercoiling by topoisomerases.** These enzymes affect the winding of DNA, and play a role in the creation of the observed branched structure of plectonemic loops along the genome.
- **Nucleoid associated proteins (NAPs).** These proteins bind to DNA with different specific modes, each responsible for a different aspect of organization, ranging from double-strand bridging to nucleoprotein filament formation.
- **Confinement.** Millimeters of genomic DNA are confined within the small cell volume of a bacterium.
- **Molecular crowding.** There is a high concentration of macromolecules present in the cytoplasm, This factor could affect nucleoid organization at different levels, for example creating a general effective self-attraction favoring collapse, and strong depletion attraction between large objects, such as ribosomes.
- **Replication and transcription.** These are the non-equilibrium processes of DNA and mRNA production that continuously take place in dividing cells. The first affects the sheer amount of genome present and supercoiling and makes both highly nonsteady, the second can create uneven ribosome concentrations. Both contribute to non-thermal force and displacement fluctuations of the chromosome.

The first part of the work is a review of the recent experimental and theoretical advances quantifying the physical organization (compaction and geometry) of the bacterial chromosome with its complement of proteins (the nucleoid). In particular, we highlight the role that statistical and soft condensed matter physics play in describing this system of fundamental biological importance, specifically recalling classic and more modern tools from the theory of polymers.

A second part of the work discusses a simple polymer physics model inspired by two main features of the nucleoid: osmotic self-adhesion and protein bridging. This idealized situation is considered from the theoretical and simulation point of view. The results are summarized by a qualitative characterization of the phase diagram of this model which shows the general feature that distinct domains may form without the need for intra-specific interactions.

The thesis also covers several data analysis approaches to test possible connections between the physical organization of the nucleoid with gene expression (RNA-Seq) and protein binding (ChIP-Seq) datasets. This latter part contains a description of the NuST webtool, which consists of a database which collect datasets from past experiments and an implementation of simple multi scale statistical analysis tools. Additionally, starting from an accurate description of the pipeline for the analysis of raw gene expression data, we introduce a correlation study of a large number (about 200) of genome-wide expression data-sets, also compared to the outcome to the published genome interaction map (HI-C) data. Finally, we discuss how the qualitative results of the polymer model presented in the previous part of the thesis may be compared with experimental data.

Chapter 2

Physical descriptions of the bacterial nucleoid at large scales

This chapter discusses challenges that arise in the biological arena, in which mature experimental and theoretical tools from the physical sciences might now allow significant progress. We will not explain the physical models in great technical detail, the references given here are a starting point and a "compass" in order to evaluate different modeling approaches.

2.1 BACKGROUND

Bacteria are single cell organisms of fundamental importance in nature and to mankind. They are "prokaryotes", in the etymological sense that they lack a nucleus as a compartment enclosed by a membrane (in contrast to the "eukaryotic" cells that make up animals and plants). Indeed, all the DNA, RNA, and proteins in the bacterial cell are always present together in the same single compartment. While the classical picture of bacteria (still implicitly adopted by many theoretical and experimental investigators) views them as little more than a bag (or "well-stirred reactor") of proteins and DNA, it is now clear that this tenet is flawed on many different levels. Despite their lack of membrane-bound organelles, bacteria have a high degree of intracellular spatial organization, related to most cellular processes. Perhaps the highest organized structure of the bacterial cell is the genome itself.

In most bacteria, the chromosome is a single circular DNA molecule confined by the cell membrane. In *E. coli* (the best studied bacterial system), the chromosome consists of about 4.7 million base pairs (bp) and has a total length of 1.5 mm [2, 3]. Bacterial DNA is organized into a specific structure called the nucleoid, which is composed of DNA, RNA and proteins, and occupies a well-defined region of the cell [4]. The nucleoid is organized by a set of nucleoid associated proteins, or "NAPs" (such as Dps and transcription factors Fis, H-NS, IHF, HU), which can modify the shape of the DNA both at local and global levels [5,6]. Since the linear size of the genomic DNA is orders of magnitude larger than the length of the cell, it must be packaged and organized in such a way

that the resulting structure is compact, while still allowing the primary information processing, genome replication and gene expression [7]. In fact, recent evidence strongly suggests that changes in chromosome architecture can directly affect the accessibility and activity of the regulatory proteins at the local level as well as at larger scales. In addition, the genome can efficiently control gene expression by changing the way DNA-binding regulatory proteins can access their target sites via the chromosome architecture [8]. Historically, the nucleoid has been visualized by transmission electron microscopy (TEM), phase-contrast microscopy and confocal scanning light microscopy; an overview of early microscopic visualization of the nucleoid can be found in [9]. These studies showed that the nucleoid mostly occupies a separate subcellular region, without being bound by a membrane, and that thin DNA threads extrude from this region.

The double-stranded genomic DNA is generally torsionally constrained in bacteria, typically in such a way that its linking number (the number of times each single strand of DNA winds around the other) is lower than in the relaxed configuration. In biological words this is referred to as "negatively supercoiled", and the specific difference in linking number is called superhelical density. The name is due to the fact that supercoiled DNA develops nonzero writhe, or "supercoils", *i.e.* it is wrapped around itself in the manner of a twisted telephone cord. This torsional constraint has important consequences for gene expression, as the mechanical stress carried by a negatively supercoiled configuration can locally weaken the interaction between the two strands. The resultant breaking of base-pair bonds between the two helices is required for the initial steps of transcription, as well as DNA replication and recombination. Since most DNA-binding proteins bind to DNA sensitively to the arrangement of the two strands, and in particular to their average distance, negative supercoiling usually also facilitates protein binding [8]. The level of supercoiling is tightly regulated by the cell, and it can be changed by the action of specific enzymes such as topoisomerases and gyrases.

The average supercoiling is generally negative in bacterial cells (except for thermophilic bacteria [10]). In the case of *E. coli*, the supercoiling superhelical density is maintained around the value $\sigma = -0.025$, where the supercoil density σ is defined as the relative change in linking number due to the winding (or unwinding) of the double helix, where negative values of σ correspond to an unwound helix [3]. This value is set by the constraining action of DNA-binding proteins and the combined activity of topoisomerases. Deviations larger than 20% to either side of σ are detrimental to cell growth. For example, overwinding causes the formation of DNA structures that impede transcription and replication, while excessive underwinding leads to poor chromosome segregation [3].

The most important *features linking nucleoid organization and cell physiology* are summarized here, and discussed further in the chapter. See also BOX 1 on page 2 for a brief description of the main factors affecting nucleoid organization.

(i) At large scales, it is seen from *in vitro* experiments that the nucleoid is composed of topologically unlinked dynamic domain structures; these are due to supercoiling forming plec-

tonemes and toroids [11], and stabilized by nucleoid-associated proteins, for example Fis. This combination of effects gives the chromosome the shape of a branched tree-like polymer visible from TEM [12–14]. Topological domains are thought to be packaged during replication, otherwise it appears that their boundaries are "fluid" and randomly distributed [12,15].

(ii) Strong compaction is experimentally observed *in vivo*, and possibly arises from confinement within the cell boundaries, but also from various factors such as molecular crowding [16] and supercoiling [17]. The degree of compaction changes with the cell's growth conditions and in response to specific kinds of stress. The *E. coli* chromosome, with a linear size of 1.5mm, occupies a volume of 0.1-0.2 μm^3 (the bare DNA volume is about a factor 20-30 smaller), which brings the need to study the folding geometry of the nucleoid [3];

(iii) Supercoiling and nucleoid organization play an important role in gene expression [8]. For example, during rapid growth, while several chromosome equivalents are present in the cell due to multiple ongoing replication cycles, an increased production of some nucleoid-associated proteins is observed, compacting the chromosome and probably giving rise to specific transcription patterns in a way that is not yet fully characterized.

(iv) RNA polymerase, the nucleoid-bound enzyme responsible for gene transcription, is concentrated into transcription foci or "factories", which can affect the nucleoid structure by bringing together distant loci [18,19].

(v) For cells that are not replicating the genome, the position of genetic loci along the chromosome is linearly correlated with their position in the cell [20–22]. The exact subcellular positioning of different loci varies in different bacteria [23].

(vi) During replication, daughter chromosomes demix and segregate before cell division. In *E. coli*, the two arms of the chromosome are segregated in an organized <left-right-left-right> asymmetric fashion from the center of the cell [23–25]. Replicated chromosomal loci are thought to be immediately recondensed, as they appear to preserve the linear arrangement while they are moved in opposite directions to assume their final position in the incipient daughter cell [7,15].

In short, the nucleoid's physical organization plays a major role in the most important cellular processes, such as cell division, DNA replication, and gene expression. Explaining these links is a long-standing open problem in microbiology. Today, it can be revisited with new quantitative experiments, including both the "-omics" approaches and more sophisticated and controlled experimental techniques allowing the analysis of nucleoids both *in vitro* and *in vivo*. This is paralleled by a renewed interest in the quantitative characterization of bacterial physiology [26–28], of which the nucleoid constitutes a fundamental part. Also note that the relative chromosomal positioning and the orientation of genes are subject to natural selection on evolutionary time-scales as shown by the comparison of the genetic maps of different bacteria [29,30].

Consequently, the field is blossoming with a new wave of studies, hypotheses and findings. While many new pieces of evidence are available, the challenge of building coherent pictures for physical nucleoid organization and its role in the different cell processes remains open. Our scope here is to present the main hypotheses and the experimental and modeling tools that have been put forward in order to understand the physical aspects of nucleoid organization, and give the interested reader an ordered account of the known facts.

2.2 MEASUREMENTS

We start by reviewing some of the salient experimental findings at the scale of whole-nucleoids, with a particular emphasis on the more recent results. The evidence that we will discuss emerges from a combination of experimental biology, biophysics, high throughput biology and bioinformatic approaches (see BOX 2 on page 6). Most of the studies cited here are based on *E. coli* or *Caulobacter*.

BOX 2: Several experimental techniques can probe the nucleoid at large scales.



Advanced microscopy together with cell-biological techniques yield information about structure and dynamics of the nucleoid (Right). **High-resolution tracking** of tagged loci allows measurement of the local viscoelastic properties of the nucleoid. **Static configurations** of chromosomal loci in fixed cells allow determination of their spatial arrangement within nondividing cells and following cell division. Dynamic tracking at long time scales gives information on the chromosome's **"choreography"** followed over a cell cycle and on the macrodomain subcellular arrangement. Fu-

sions of NAPs with fluorescent proteins also enable evaluation of their **localization** within the nucleoid.

Nucleoids can be purified and manipulated outside of a cell in order to access more directly their biophysical properties (Left). This procedure implies release of confinement and crowding, and dilution of binding proteins. As a result, purified nucleoids are several times larger in radius than the size of a cell. Traditionally, purified nucleoids were imaged by electron microscopy, showing a ramified plectonemic structures. More recently, investigators have concentrated on characterizing their organization as polymers **tracking of labeled loci**, fluorescence correlation spectroscopy(**FCS**) and probing them mechanically (**AFM**).

In addition, information on nucleoid organization can be obtained from high-throughput experimental datasets (Bottom). **Transcriptomics** (using sequencing or microarrays) can be used to probe transcriptional response to nucleoid perturbations, such as NAP deletions, changes in the average level of supercoiling, or local release of a plectonemic loop. Another important source of data is **protein occupancy**, for example by NAPs, and its correlation with gene expression. This information is obtained both by microarray (CHiP-chip) and by next-generation sequencing techniques (ChIP-seq). **Recombination** has been used to define macrodomains, as compartments within which recombination between chromosomal segments was more likely than recombination with segments laying outside of the compartment. Finally, **Chromosome Conformation Capture** (3C) techniques probe the spatial vicinity of pairs of chromosomal loci in the (average) cell. They can be combined with sequencing in order to produce high-throughput data sets (Hi-C).

Chromosome spatial arrangement and compartmentalization.

Strikingly, the intracellular localization of a given chromosomal locus in a cell is remarkably deterministic, as revealed by fluorescent tagging of chromosomal loci on *E. coli* and *Caulobacter crescentus* [20, 22, 24, 31–34]. The series of chromosome segments is localized along the long axis of the cell in the same order as their positions along the chromosome map, with the interlocus distance typically linearly proportional to (arclength) genomic distance. The precise location of individual loci varies in the known bacteria, and might depend on DNA-membrane tethering interactions [23]. A recent microscopy study on *E. coli* [35] considered the positioning of the chromosome in the short-axis section of the cell. They found that the Ter region occupies the periphery of the nucleoid, at a larger distance from the longitudinal axis with respect to the rest of the chromosome. In newborn or non-replicating cells, the two chromosome arms are spatially arranged such that loci on the left arm of the chromosome lie in one half of the cell and loci on the right arm lie in the opposite half, with the replication origin between them. It is tempting to interpret the resulting sausage-shaped structure as a chromosomal "fiber." In a recent study [22], the cell-to-cell variability of loci positioning in non-replicating cells was used to estimate an internal elasticity, which, perhaps not surprisingly, appears to be much higher than expected from a naive estimate

for a linear polymer.

Even more recently [36], high-throughput Chromosome Conformation Capture (3C) techniques have been used in combination with live-cell fluorescent tagging of loci, in order to determine the global folding architecture of the *Caulobacter crescentus* swarmer cell genome. These data indicate that a chromosomal fiber exists also in this case, spanning the whole chromosomal ring. Additionally to the linear spatial arrangement of loci according to their chromosomal coordinate, loci of the left chromosomal arm tend to be very proximal to symmetric loci on the right arm. The resulting structure is a compressed ring-like fiber, which, the authors argue, typically takes an eight shape, free to roll around the long cell axis. They also find that the symmetry in the cross-chromosomal arm interactions is determined by the protein-dense attachment point to the cell membrane at the old pole of the cell, triggered by the binding of the ParB protein to its target parS binding sites. Moving the parS pole-anchoring site by 400 Kb along the chromosome (but not the replication origin) determines a sliding of the whole interaction structure, as in a tank crawler. This sliding is slightly asymmetric, suggesting the presence of supplementary attachment points between chromosomal arms or between the chromosome and the cell body.

An improved

Another important recent discovery, consistent with the mentioned correlation between chromosome arms and cell halfs, is the existence of "macrodomains" [37, 39, 40] often described as chromosomal isolated compartments. The first evidence in this direction [37] came from measurements of the recombination frequency between loci, see figure 2.1A. All else being equal, this should be proportional to the probability that the two chromosomal segments come into contact within the cell. For a well mixed polymer, the recombination frequency should be uniform. However, experiments show a highly non uniform pattern, compatible with a compartmentalized structure with clear boundaries. Four macrodomains of a few hundred Kb in size have been identified, corresponding to regions surrounding the replication origin and terminus, and to two symmetric regions at the edges of the Ter macrodomain, see box and figure 2.1C,D,E. The remaining "non-structured regions" appear to have different physical properties. Subsequent studies have confirmed the presence of macrodomains and measured their dynamics using fluorescently labeled loci [38, 40, 41], see figure 2.1B,D. While the molecular mechanisms responsible for this level of organization are not yet clear, the same authors also found that the Ter macrodomain appears to be condensed by a single DNA-binding protein (MatP, figure 2.2) with a small set of specific binding sites [42]. Other proteins with macrodomain-specific DNA-binding properties have recently been identified [43–46], and appear to be conserved in related bacteria. In view of the results from this thread of work, macrodomains might be seen as a process of microphase separation triggered by specific protein binding. Interestingly, as mentioned above macrodomain-like regions also emerge from independent large-scale genomic data [47–49].

During replication, the chromosomes segregate following a well-defined "choreography", which has been the subject of multiple studies [23, 50, 51]. While segregation is not the main focus here,

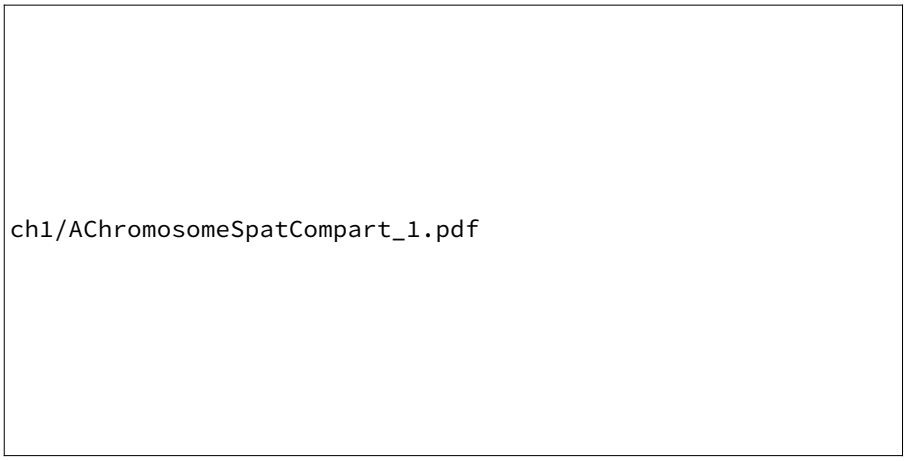

Figure 2.1: *Chromosome compartmentalization and spatial arrangement of genes.* (A) An example of the data that was used for the genetic definition of Right and Ter macrodomains, reprinted by permission from Macmillan Publishers Ltd. from [37], © 2004. The plot shows the frequency of recombination events between the genetic position indicated by the arrow and other positions probed along the genome (The x-axis indicates the base pair coordinates of the chromosome). This should be uniform for a well-mixed polymer. On the contrary, experiments show a highly non-uniform pattern, compatible with a compartmentalized structure with clear boundaries. Dashed vertical lines indicate the delimitation of macrodomains. Colored bars show the extent of the defined macrodomains. (B) Effective diffusion constants obtained from loci tracking experiments, reprinted by permission from John Wiley and Sons Ltd. [38], © 2008 . The plot shows clear differences between the behavior of loci placed within macrodomains and in non structured regions (NSR1-4,NSL-2). (C) After [37] (reprinted by permission from Macmillan Publishers Ltd. © 2004), graphical representation of the macrodomains and their boundaries within the *E coli* genome. Structured macrodomains are indicated as colored arcs, black dashed arcs indicate non-structured regions. (D) From [38] (reprinted by permission from John Wiley and Sons Ltd. © 2008), the genetic insulation of the macrodomains correlates with spatial insulation in subcellular territories. The positions of 10 foci were superimposed according to the barycenter of their trajectory during 30 intervals of 10 s at the home position. Foci from tags at OriC, NSR-3, Right-2 and Ter-6 were plotted. (E) Schematic of macrodomains' localization within the cell. Ter is localized at the periphery of the nucleoid, towards the cell membrane [35] (reprinted by permission from Biomed Central, © 2011).

it is useful to discuss it briefly, as the existing approaches to this problem (both experimental and theoretical) are intimately linked with chromosome organization and will be mentioned in the following. Specifically, spatial reorganization of the segregating chromosome arms appears to preserve qualitatively the relationship between loci distance along the chromosome and in the cell. Moreover, reorganization ensures that the two replication forks remain in opposite halves of the cell during replication and that the relative orientation of the two reorganized nucleoids in predivision cells is not random. Quite interestingly, the spatial separation of sister chromosomes is not a continuous process, but has been observed to proceed through "snaps" [38, 52], suggesting the existence of energetic or entropic barriers for separation, possibly overcome by active processes.

There is debate on what is the main driver for chromosome segregation: Entropic repulsion forces due to strong confinement into a box of linear polymers have been proposed to explain this behavior at least in part [51]; in experiments on replicating *B. subtilis* the chromosome compaction and spatial organization have been hypothesized to result from non-equilibrium dynamics [50], rather than from an entropic repulsion process; in *Caulobacter crescentus,* a contribution from bidirectional extrusion of the newly synthesized DNA from the transcription complex has been pos-

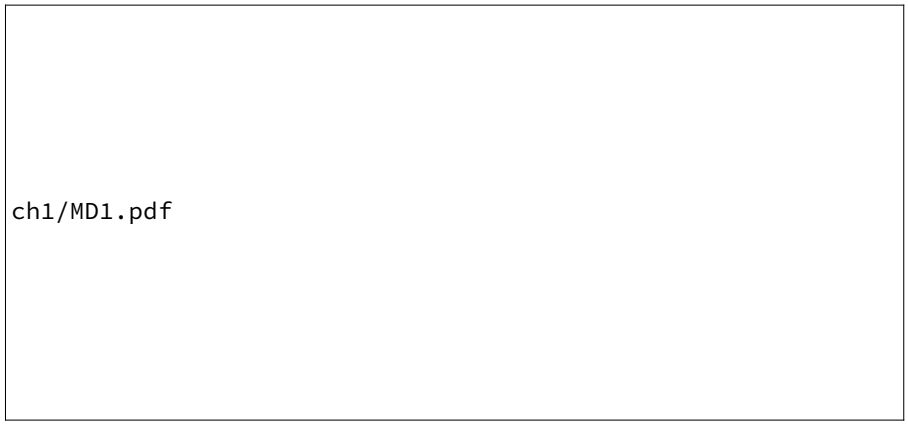

Figure 2.2: *The protein MatP organizes the Ter macrodomain by specific binding.* The circular chromosome of *E. coli* divided into 4 macrodomains and 2 non structured zones (see figure 2.1). (A) From [42], ChIP-chip assays for MatP binding show a specific affinity of this protein with Ter macrodomains, indicating that the macrodomain is condensed by MatP, and suggesting that other macrodomains might be condensed by dedicated proteins with a small set of specific binding sites. (B) From [42], the localization bias of MatP binding sites is conserved among enterobacteria, (*S. typhimurium LT2* with genome size 4683 Kb, *E. carotovora SCRI1043* - 5064 Kb), Vibrio (*V. cholerae Cl 2* - 961 kbp), and Pasteurella (*Y. pestis KIM 4* - 600 kbp) species. (C) MatP controls DNA compaction in the Ter macrodomain. The histogram represents the proportion of cells with given interfocal distances between foci of two Ter MD markers in WT and matP deletion mutant cells. (D) Fluorescence microscopy images from the two experiments (scale bars 2μm). These results indicate that the effect of MatP on foci colocalization is associated with compaction. (A-D) reprinted from [42], © 2008, with permission from Elsevier.

tulated to contribute to chromosome segregation [23, 53]. Experiments using inhibition of protein synthesis by chloramphenicol show that this produces nucleoids with a more rounded shape and induces fusion of separate nucleoid bodies [54]. Upon release from protein synthesis inhibition, the two nucleoids reoccupy the DNA-free cell independently of cell elongation [55]. The control of the segregation mechanism by protein synthesis processes could be indirect: ongoing protein synthesis can affect nucleoid compactness and segregation at multiple levels, including the decrease in the amount of enzymes modifying the topology of the DNA or carrying out transcription and DNA replication. The idea that membrane protein synthesis activates nucleoid segregation directly has also been proposed [56] (in bacteria the presence of cotranscriptional translation creates a direct physical link between the genome and the membrane [57]) Also note that current arguments based on segregation by entropy and confinement might turn out to be inconsistent with interpreting macrodomains as the result of a microphase separation, since a variety of interactions, and specifically protein-DNA binding processes could play an important role in defining the free energy of the nucleoid. In a recent work [58] the choreography of the large-scale nucleoid organization during the time of replication was observed by means of labeling the Ori, Ter, Left and right with different colored fluorescent labels. This allowed to observe three predominant longitudinal patterns for the organization of the chromosome during the cell cycle: ter–left–ori–right in the G1 phase, ori1–left–ter–right–ori2 in the S phase and left–ori–right–ter–ter–left–ori–right in the G2 phase. By using monte-carlo polymer simulations the authors were able to verify that segregation by entropy is able to account for the observed choreography only by constraining the origin and

terminus regions to specific positions of the cell volume and by adding specific interactions which folds the polymer into macrodomains.

Supercoil domains and nucleoid associated proteins.

At smaller scales, the circular chromosome of *E. coli* is organized in plectonemic loops, or "supercoil domains." Those regions are separated by topological barriers formed by nucleoid associated proteins (NAPs) such as Fis and H-NS, see figure 2.3. These proteins bridge two strands of DNA by binding to both, and prevent the propagation of torsional energy. As a result, they also prevent the spreading of uncontrolled effects on gene expression in case of accidental DNA breaks or mechanical strain, caused for example by advancing replication forks [3, 12, 13].

Multiple NAPs have been identified, each with its specific binding properties (reviewed in [5]). Besides bridging double strands, they can change the local shape of DNA inducing bends or hinges and form nucleoprotein filaments. Additionally, NAPs often have multiple DNA binding modes which might be dependent on physiological factors. A good example of this behavior is H-NS. Its binding results in DNA—H-NS—DNA bridges, but also forms a rigid nucleoprotein filament which could act as zipper *in vitro* [59–61]. It seems that the nucleoprotein filament formation may be important, as it has been found to be a structure shared by other NAPs including HU [62] and StpA [63]. Biologically, this could hypothetically provide a mechanism for environmental sensing by NAPs. It is thus possible that our current understanding of NAP binding, being based on the limited number of conditions tested, is incomplete and generalization of a DNA binding property to other solution conditions may be dangerous.

The structure of supercoil domains was studied in *E. coli* by Postow and coworkers, through analysis of the supercoiling-sensitive transcription of more than 300 genes following relaxation by restriction enzymes *in vivo*, and by electron microscopy [12]. They concluded that domain barriers may vary dynamically and/or across a population, but they follow an exponential length distribution. The average domain size is $\simeq 10 - 15$ Kb, implying the existence of about $200 - 400$ domains [3, 12]. Branches of the same typical length are visible directly from electron micrographs of purified nucleoids [12, 14]. Thus, the genome topology may be visualized as a branched structure with supercoiled domains that are subject to modulation by nucleoid-associated proteins, and active processes such as DNA transcription and DNA replication.

The factors responsible for establishing the boundaries of supercoiled domains and the determinants of domain size and number are still largely unknown. While H-NS and Fis, along with MukB, the analogue of the eukaryotic SMC (Structural Maintenance of Chromosomes) protein, could be involved in stabilizing plectonemic conformations, their precise roles and importance in this context have not yet been established in a definitive fashion, either *in vitro* or *in vivo* [8, 13, 65, 66].

By contrast, a consistent amount of information on the binding of NAPs in different condi-

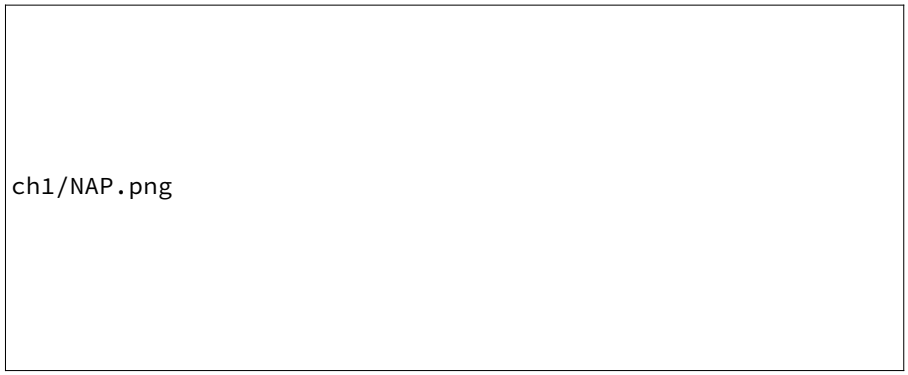

Figure 2.3: *Branched plectonemic conformation of the genomic DNA molecule in E. coli and connection with nucleoid associated proteins.* (A) From [12] (Reprinted by permission from CSHL Press, © 2004), electron micrograph of a purified *E. coli* chromosome. The branched plectonemic structure is visible. From image analysis, Postow and coworkers measured a roughly exponential distribution for the length of supercoiling loops, with the average around 10-15Kb. Note that the extraction and purification process might interfere with many of the properties of the conditions in which the nucleoid structure is naturally found. (B) Electron micrograph of Fis binding to plasmid DNA (from [64]). This dimeric NAP stabilizes plectonemic branches of supercoiled DNA (schematized in the right panel). In particular, Schneider and coworkers measured a Fis to DNA ratio of 1 dimer per 325 bp DNA. (C) From [64], electron micrograph of H-NS binding to plasmid DNA, elongated complexes (polymers) are formed (schematized in the right panel), presumably containing two DNA duplexes. Schneider and coworkers measured a ratio of H-NS to DNA of 1 molecule per 10 bp. (B-C) reprinted from [64] by permission of Oxford University Press.

tions and its effects on the cell state is available from high-throughput experimental techniques (figure 2.4.) Nucleoid associated proteins can modulate the nucleoid conformation structure in response to changes in environmental conditions [5]. This can result in large-scale changes in gene expression [8]. The local mechanical action of NAPs on DNA is often well-characterized by single molecule experiments [5], which also lead to the observation of chromatin-like nucleoprotein "fibers" [70]. Large-scale NAP binding data in specific growth conditions was obtained from high-throughput experiments involving microarrays (CHiP-chip) or sequencing (CHiP- seq) [66,67,71–74]. Furthermore, transcriptomics studies profiled the changes in gene expression upon different nucleoid perturbations, such as NAP deletion and/or altered supercoiling [48,69,75,76]. Many of these data sets show linear regions of dense binding that often correspond to macrodomain boundaries, and associate with global or NAP-dependent transcriptional response and its correlation with codon bias [47, 49, 77].

The physical origin of this preference in binding and the gene expression changes at the boundaries of macrodomains is not precisely clear. One possibility is that macrodomain boundaries might be co-localized by NAP structures. A recent interesting experimental study [68] looked at protein occupancy along the genome regardless of protein identity. This work uncovered extended polymer-like domains rich in bound proteins (including NAPs) with an average length of 1.6Kb, associated with transcriptionally silent or transcriptionally enhanced regions (and also with high intrinsic DNA curvature).

When cells enter the stationary phase a radical, global condensation of the nucleoid occurs. It is believed that this is a mechanism via which the cell can protect its DNA in harsh conditions [70].

ch1/Omics.pdf

Figure 2.4: *Examples of high-throughput data an bioinformatic analysis concerning nucleoid organization* (A) Left panel: ChIP-seq binding profile of H-NS, plotted using data from [67]; right panel: sketch representing the extended protein occupancy domains detected by [68]. These are polymer-like nucleoprotein complexes, presumably related to nucleoid organization. (B) Chromosomal sectors emerge from the local correlation (plotted here) between expression levels for wild type in log-phase growth and a codon bias index, related to translation pressure (from [47], reprinted by permission from Macmillan Publishers Ltd: Molecular Systems Biology, © 2010). These sectors correlate well with macrodomain organization. (C) Top panel: effective transcriptional regulatory network (red links) and areas of influence of Fis and H-NS (colored circles) obtained from transcriptomics experiments combining NAP mutants and perturbations in supercoiling background (from [69], Reprinted by permission from BioMed Central, © 2008). Bottom panel: clusters of transcriptional response to nucleoid perturbations (left, data from [69] and similar experiments) and NAP binding (right) correlate with the organization of macrodomains (outer arcs) and chromosomal segments (inner arcs) of the genome (from [49], reproduced by permission of The Royal Society of Chemistry).

AFM studies have shown that the structure of the DNA differs at the supercoiling level [70] and that action of Dps and CbpA, the NAPs that replace Fis in this growth phase, is quite different. The Dps and CbpA proteins produce compact aggregates (which can protect DNA from degradation by nucleases) rather than binding to distributed sites as Fis does [78]. Interestingly, the action of Fis counters Dps-induced compaction through a transcriptional response affecting the expression of topoisomerase and gyrase [79].

Compaction by molecular crowding, specific proteins, transcription factories, and confinement.

A different question concerns identifying the main factors contributing to nucleoid compaction and organization. The main candidates are macromolecular crowding, electrostatic self-attraction, supercoiling and nucleoid proteins.

Macromolecular crowding, or the high concentration of macromolecules present in the cytoplasm, is generally believed to be an important determinant on the basis of theoretical arguments [16, 80], which predict a possible phase separation mechanism between nucleoid and cytoplasm. Note that the generic term "macromolecular crowding" might include depletion interactions (see below), together with a number of additional effects of entropic and energetic nature.

DNA condensation by crowders can be observed *in vitro* under very controlled and well-understood conditions, for example by experiments using dextran or PEG, demonstrating that naked DNA can

be directly condensed by these crowders [81–84] Similar (but less controlled) behavior, is shown by purified nucleoids [85]. At the same time, experimental studies on isolated nucleoids obtained from mutants lacking various NAPs [86] suggest that the effects of crowding on compaction are substantial and independent of the NAP composite background. It has also been suggested that the action of NAPs could be aimed at antagonizing compaction rather than compacting the nucleoid [87]. However, this must be a complicated, combined effect involving forces of different nature. In the absence of crowding and confinement, it is obvious that the radius of gyration of the genome would be smaller if it were organized, e.g. in a branched structure of plectonemic loops stabilized by DNA-bridging NAPs [2, 12].

As a particular case of crowding effect, it has been proposed that the (attractive) depletion interactions, well known in colloid science, might play an important role in chromosome organization [88] This force is due to reduction in total solvent excluded volume upon formation of a molecular complex. Depletion interactions are consequential when large molecular assemblies are formed in presence of smaller particles. In fast-growth conditions, genome-bound RNA polymerase is localized into a few transcriptionally active foci or "transcription factories" [18,89], analogous to the eukaryotic case [88,90]. Depletion interactions were suggested to explain the formation of these macromolecular assemblies [90]. Interestingly, the formation of these foci has been associated with the presence of the NAP protein HU [48]. We add that a similar argument might hold for the local compaction of the surroundings of OriC, rich in ribosomal RNA producing regions, in a macrodomain-microphase. In other words, ribosome-rich ribosomal RNA transcripts, attached to the genome through RNA polymerase, could help compact the Ori region. If this should be the case, the compaction properties of this region would change with the number of ribosomes being synthesized, i.e. with growth rate and translation efficiency [26] (One can also speculate that, for the same reason, newly-replicated DNA could be sequentially aggregated during replication.) On one hand, experiments show that transcription of ribosomal RNA operons (which are generally located in the Ori macrodomain) is related to the compaction of nucleoids observed upon inhibition of translation [89]. On the other hand, these processes must be intersected with the (non entropic) binding of NAPs, given the evidence connecting specific NAPs to the compaction of the Ter macrodomain (MatP) and indirectly to the effect of the HU protein on organization of the Ori macrodomain.

A very recent study [91] systematically addressed the chromosomal localization and role in spatial organization of the nucleoid of five major NAPs (HU, Fis, IHF, StpA, H-NS) using fluorescent protein fusions and super-resolution fluorescence microscopy. In the growth conditions tested all the proteins showed scattered distributions in the nucleoid, except for H-NS, which seemed to form (on average) two distinct foci per chromosome copy, bringing together different (even distant) H-NS targets, see figure 2.5. Thus, H-NS should be added to the list of NAPs with a compacting action on the nucleoid associated with the formation of specific foci. The long-range interactions between H-NS binding targets were validated by 3C, and show no apparent coherence with



Figure 2.5: *Effect of H-NS on chromosome compaction.* From [91], Reprinted with permission from AAAS. H-NS was shown by Wang and coworkers to form a small set of foci in the cell, which bring together distant H-NS targets. Together with MatP (figure 2.2), this evidence proves a role of drives of energetic origin in chromosome compaction. (A) Super-resolution fluorescence imaging of fluorescent protein fusion of H-NS showing compact H-NS clusters in the nucleoid. The *E. coli* cells are shown in the bright-field image (left). The z coordinate of each localization is color-coded (top bar). In comparison, a conventional fluorescence image of the same cells is shown (right). (B) Scattered distribution of HU in the nucleoid. (Left) Bright-field image; (right) super-resolution image. Similar diffuse distributions were observed for Fis, IHF, and StpA. (C) Effect of deleting H-NS on the subcellular distribution of H-NS target genes (hdeA, hchA) versus non-targets (lacZ). The 2D histograms of the relative hdeA, hchA, and lacZ locus positions normalized to the cell dimensions are shown, the genes were fluorescently labeled. The first quartile of the cell plots the (color-coded) probabilities of finding the fluorescent locus in the particular position, the rest of the cell is filled with mirror images to aid the eye (grid size 100 to 200 nm). In each case, 2000 to 5000 gene locus positions were analyzed. H-NS deletion has little effect on lacZ distribution, despite this locus is only a few Kb apart from the closest H-NS binding site.

the macrodomain structure; loci pairs that are near to H-NS targets but are not targets show no 3C signal and wider distributions of subcellular distances (evaluated with microscopy). Another study [92] used both a precisely timed steerable optical trap and a high throughput microcapillary Coulter counter to probe the physical properties and size of purified nucleoids, at the individual nucleoid level. The purified nucleoids extracted from cells without the H-NS gene are approximately five-fold larger in stationary phase than the ones in WT cells. This suggests that H-NS plays an important role in regulating the nucleoid structure in the stationary phase. In general, DNA associated with many NAPs has a much larger surface, which should enhance the depletion interactions. For instance, DNA coated with an H-NS nucleoprotein filament will have a diameter of about 20 nm instead of the 2 nm of naked DNA. H-NS nucleoprotein filament formation could strongly enhance the depletion attractions with respect to other NAPs not forming filaments, and filaments formed at remote locations on the contour of the chromosomal DNA could find each other in a crowded environment. This may explain the results from super-resolution imaging mentioned above [91]. We note again that this is one of the counter-intuitive aspects of molecular crowding, since one might think that it would be easier to encounter another filament if there

were no "obstacles". The confusion can come from thinking about diffusion, which is hindered in a crowded environment due to the higher effective viscosity, versus thermodynamics, where the depletion interactions increase the probability of observing two nucleofilaments in contact at equilibrium. Adopting a simple Kramers-type model for the kinetics, it is not obvious whether the rate of association would go down due to the prefactor or go up due to the lower free energy minimum. Implicit in the entire discussion implicating crowding in aggregation in this cellular context is the idea that the distribution of nucleoprotein filaments is at near-equilibrium even though the cell is dividing rapidly.

Finally, depending on the degree of autonomous compaction, the confinement exerted by the cell wall might play a relevant role in nucleoid organization and segregation [51]. This is expected to be particularly significant in fast-growth conditions, where the genome needs to be highly accessible for transcription (and thus will not be condensed) and the amount of genome per cell is higher due to overlapping replication rounds [25].

To summarize, the degree of condensation and the geometry of the nucleoid are strongly dependent on the growth phase and growth rate of the bacteria. Such changes may be modeled by means of equilibrium states, slowly evolving in accordance to "external" control parameters (e.g. the concentration of the various NAPs). However, it seems likely that nonequilibrium processes are also important, and relevant aspects of the problem might be lost in attempting to describe the nucleoid condensation process purely in terms of an approach to thermodynamic equilibrium. For example, on the other side of the spectrum in terms of biological complexity, the recently found scale-invariant structure of the human genome [93, 94] was suggested not to be the result of an equilibrium state, but similar in nature to the so-called "crumpled globule", or "fractal globule" [95], since loci that are near along the genome arclength coordinate are also physically proximal in three-dimensional space. This proximity contrasts with what happens in an (equilibrium) collapsed polymer, where the linear structure is completely mixed in the globule.

Let us quote a few experimental findings, some of which are very recent, supporting the role of nonequilibrium processes, specifically for bacteria. The initial conditions, i.e. the choreography of DNA replication, appear to play a central role in defining the final structure of the nucleoid [96]. In artificial *E. coli* strains with two distant origins instead of just one [97], the two origins initiate replication synchronously at the expected separate positions of the genetic loci associated with them. Replication forks move independently, indicating that replication does not occur in a single replication factory and that the replication machinery is recruited to origins rather than vice versa. Most importantly, in these experiments progression of replication plays a major role in determining the space-time pattern of locus segregation. Nucleoids growing in rich medium at different stages of the cell cycle were visualized in fixed cells using direct stochastic optical reconstruction microscopy (dSTORM) [98]. The transition from one to two well-separated nucleoids begins with one nucleoid located at mid-cell (2.7μm cell length), progresses as sisters separate into the two cell halves (3.0 and 3.5 μm). Nucleoids decondense after reaching their final positions at cell quar-

ters (3.7 and 3.9 μm). The same authors also observed an asymmetry between sister chromosomes, which accumulates for ongoing replication rounds. The large scale structure of the *B. subtilis* nucleoid [50] has been observed at various stages of the replication process. The newborn portions of the chain are compacted and sequentially conveyed towards the poles, resulting in an ordered, spiraling structure. A strong correlation between space coordinate and genomic coordinate is preserved, similar to the linear behavior observed in *E. coli* [22]. A choreography of this sort is found also in *Caulobacter* [53]. Finally, the large scale spiral structure of the nucleoid of *Bdellovibrio bacteriovorous* [99] also suggests a metastable steady state, sustained by cooperative motion and/or energy exchanges. It seems difficult to disregard the deterministic replication-segregation dynamics in describing such phenomenology [21, 23].

Viscoelasticity and structural units.

Tracking studies of fluorescently labeled chromosomal loci have evaluated *in vivo* dynamic properties of the nucleoid with fairly high time resolution, measuring for example the mean-square displacement (MSD) of the loci or the time autocorrelation function [35, 38, 100]. In general, these measurements give information on the local relaxation time scales of the nucleoid and its viscoelastic behavior, see figure 2.6. On one hand, for large time-scales [38], loci mobility correlates well with macrodomain structure. In particular, the MSD saturates at the spatial scale of the macrodomain size. On the other hand, especially for smaller time scales, the mean square displacement of a locus is seen to follow a power law: $\boldsymbol{MSD} \equiv \langle(\vec{x}(s, t + \Delta t) - \vec{x}(s, t))^2\rangle \approx c \cdot (\Delta t)^{\alpha}$ (where s is its arclength genomic coordinate) and the exponent α seems to be universally close to 0.4, independent of s [100]. Perhaps surprisingly, an extra-chromosomal RK-2 plasmid showed the same behavior, while smaller RNA particles had a higher subdiffusive exponent. The authors interpret the subdiffusive behavior as an intrinsic effect of a viscoelastic cytoplasm diffusion medium.

It must be mentioned that the localization of RK-2 plasmids appears to be highly regulated [101, 102]. In a more recent work [103] the same exponent α was measured at short time-scales (0.1–10 s) using fluorescently labeled chromosomal loci in *E. coli*. A dependence of the prefactor of the loci's apparent diffusion on both their subcellular localization and chromosomal coordinate was observed along the whole nucleoid. This result indicates that either non-equilibrium fluctuations due to enzyme activity or the organization of the genome as a polymer–protein complex vary as a function of the distance from the origin of replication. The underlying viscoelasticity of the bacterial cytoplasm surrounding the nucleoid is still poorly understood. Some characterization has been approached *in vivo* via FRAP measurements of diffusing GFP [104].

The observed anomalous diffusion has been modeled using phenomenological approaches. There are a variety of dynamical models exhibiting anomalous diffusion, such as Langevin equations with time dependent viscosity (also equivalent to fractional Langevin equations), continuous time random walks, and random walks over a fractal object [106–108]. Weber et al. [100, 109]

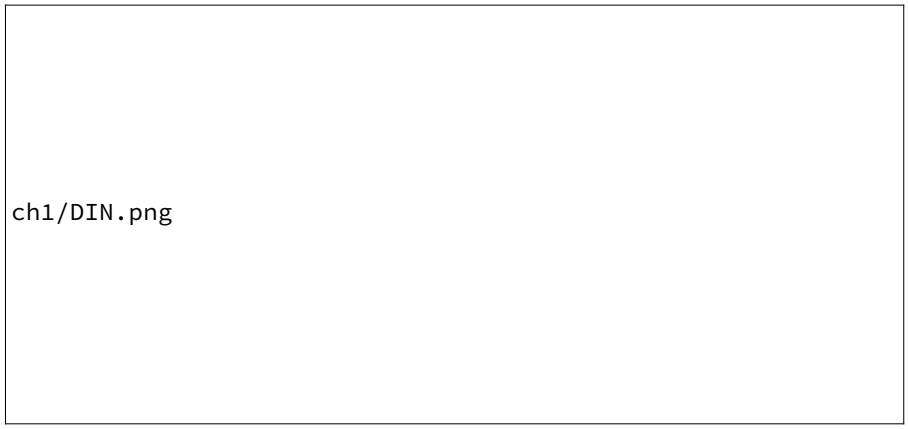

Figure 2.6: *Optical techniques in nucleoid dynamic visualization.* (A) From [35] (reprinted by permission from Biomed Central, © 2011), using phase contrast for cell visualization, membrane staining (FM 4-64), DNA staining (DAPI), and YFP-ParB for foci dynamic measurement, it is possible to produce "informed maps" of the bacteria by overlaying different micrographs. (B) Movies can be used to to follow foci dynamics. The plot, from [38] (reprinted by permission from John Wiley and Sons Ltd. © 2008), shows a typical trajectory of a fluorescent focus, from which dynamical properties of the nucleoid can be inferred. (C) From [38], reprinted by permission from John Wiley and Sons Ltd. © 2008. The trajectories of fluorescent dots define territories in which genetic loci are placed within the cell. (D) From [100] (http://prl.aps.org/abstract/PRL/v104/i23/e238102), reprinted, © 2010 by the American Physical Society. A measurement of the ensemble-averaged MSD for live and fixed cells using fluorescent loci. The authors showed that the dynamical exponent of the foci is universal (and close to 0.4). The observed anomalous diffusion has been explained phenomenologically as a consequence of the viscoelastic behavior of the nucleoid. We hypothesize that this could be a consequence of a fractal organization of the nucleoid similar to a "fractal globule", *see Discussion*. (E, F) Measurements on purified nucleoids, from [105], reprinted with permission from Elsevier, © 2007. Using Fluorescence Correlation Spectroscopy, nucleoid volume and amplitude of FCS correlation functions have been measured at varying supercoiling (induced by different concentrations of chloroquine drug). The plots show that the nucleoid reach a maximum volume and minimum FCS correlation at full supercoiling relaxation. The FCS correlation signal can be used to deduce the size of structural units of the polymer.

compared their data with the results obtained on the basis of the first two approaches, and the fractional Langevin equation gave a more satisfactory agreement. The above method, used as a null model for for subdiffusive chromosomal dynamics, uncovers the existence of rare but ubiquitous "rapid movements" of chromosomal loci exhibiting near-ballistic dynamics [110]. Additionally, rapid movements are in some cases coupled with known transitions of chromosomal segregation and show a preference for longitudinal motion. This result supports an emerging picture of the bacterial chromosome as off-equilibrium active matter. A recent work [111], which focused on global visualization of living *E. coli* nucleoids dynamics defined by HupA-mCherry fluorescence observed longitudinal density waves, which flux back and forth along the nucleoid.

The anomalous diffusion exponent *per se* does not give quantitative information on the geometry of the nucleoid. Physical measurements, possible on purified nucleoids [105, 112, 113], can enrich the scenario. In particular, Romantsov and coworkers [105] have obtained, using fluorescence correlation spectroscopy, the time-dependent coarse-grained density distribution of purified nucleoids. They conclude that the polymer appears to be composed of a set of "structural units" defined by a measurable correlation length. The measurements were performed at varying degrees

of supercoiling, induced by different concentrations of the chloroquine drug. The size of the structural units was found to vary from 50-100Kb in high (positive or negative) supercoiling to 3Kb at zero supercoiling. The diameter of the purified nucleoid varied from 2.5μm in high supercoiling to 3.5μm in low supercoiling. The authors also estimated the typical diameter of the structural units from the diffusion constant (obtained from the decay of the fluorescence autocorrelation function) and Stokes-Einstein's relation. Perhaps surprisingly, the resulting size of structural units was near 70-80 nm regardless of supercoiling. Thus the emerging picture for the unconfined genome is that of a string of $\sim$ 100 highly dense "beads", each containing $\sim$ 100 Kuhn lengths (effective independently jointed elementary polymer segments) of DNA each. These values apply in the presence of supercoiling but in the absence of crowding and confinement effects, as for purified nucleoids most of the cytoplasmic (and probably a considerable part of the DNA-binding) proteins are probably diluted away.

2.3 MODELS

We will now review a few modeling approaches put forward in recent years, and point to some more classic work in polymer physics that we believe could be relevant in this context. As the reader will have observed, the wealth of existing experimental results is appealing on one hand, but, on the other hand, does not offer any clear grasp on a small set of relevant ingredients necessary for building coherent physical descriptions of the nucleoid. Consequently, the approaches adopted in the literature are highly diverse and heterogeneous in terms of premises, methods, ingredients, and points of view.

The notable efforts to understand entropic aspects of chromosome segregation (reviewed in [51]) have revived the studies on entropic forces of single and multiple confined polymers dating back to Edwards and DeGennes [114–117]. As we have pointed out above, the role of bound proteins on nucleoid entropy is typically disregarded in these arguments.

A comprehensive review of the literature on confined DNA in different contexts is provided in [118]. Considering this approach, there remains the open issue that the observed segregation and compaction times might be too short to be compatible with an entropic process. For linear polymers this time has been evaluated by Molecular Dynamics simulations to scale like the square of the number of monomers, N^2, which, for sufficiently large N, will be smaller than the $O(N^3)$ chain diffusion time [119]. However, it is not straightforward to use these scaling relations for empirically relevant estimates. Moreover, these estimates will depend on the model used to represent the structure of the nucleoid and its correlation with the replication process, linking this problem to other unanswered questions. Overall, it seems likely that entropy is only part of the story, and active / nonequilibrium processes of different kinds might play a role in chromosome segregation.

We will now turn our attention to work concerning nucleoid organization and cellular arrangement. Buenemann and Lenz have attempted to understand the linear arrangement of chromoso-

mal loci in terms of a purely geometrical model [120], where a linear polymer in the form of a string of blobs, is confined within a cylinder and locked at one or at a few loci. This constrained geometry obviously provides an ordering mechanism, as long as the blobs are large enough with respect to the cylinder's diameter. This model makes predictions on the spatial arrangement of the chromosome in mutants of *C. crescentus* [20] and on the cell-cycle dependent ordering in *E. coli*. A very recent simulation study [121] explains the linear ordering observed in *E. coli* as a product of confinement and entropic repulsion of a string of linearly arranged chromosomal loops. In order to show this, they represent the chromosome as a confined circular self-avoiding chain under the constraint that consecutive loops, identical in size, are distributed along the arclength coordinate, while the Ter region does not contain such loops. Their simulations show both linear ordering along the cell axis and Ter region occupying the outer periphery of the nucleoid [35], as properties of the equilibrium states. Intriguingly, the same study suggests that this linear-loop ordering could be a consequence of the transcription network organization. To support this point, they simulate a polymer where transcription factor-target pairs are coupled by attractive harmonic interactions, and show that the linear ordering is recovered. It is well-known that transcription factor-target distances have a statistical tendency to be short along the chromosome [122].

Vettorel and coworkers [123] performed an abstract study motivated by the possible nature of the compartmentalization and the structural units of a generic (eukaryotic or prokaryotic) chromosome forming a crumpled or fractal globule mentioned in section 2.2 (figure 2.7). Specifically, they explore metastable collapsed states of polymers, where the total size scales as $N^{1/3}$ as in an equilibrium compacted globule, but a much higher degree of compartmentalization is present. At odds with the intrinsic disorder of the equilibrium globule, in a fractal globule, for any pair of loci separated by a chain length s, the distance $R(s)$ has the scaling behavior $R(s) \approx s^{1/3}$ [93]. In other words, both the equilibrium globule (i.e. the equilibrium collapsed structure emerging from polymer self-attraction or unfavorable entropy of mixing) and the fractal globule have mass fractal dimension $D_f = 3$, but a generic volume (and in particular a blob in the DeGennes sense) inside the equilibrium globule can include non-sequential segments, while in the fractal globule it contains a single sequential segment [95]. This structure can be understood as resulting from a process where condensation sequentially involves larger and larger scales in s, so that the genomic proximity is preserved in a scale invariant fashion. As a consequence, distant portions of the chromosome will occupy different compartments within the globule.

In order to obtain a fractal globule in a simulation, a constraint preventing entanglement must exist [123]. To obtain this condition, Vettorel and coworkers considered a semi-dilute or concentrated solution of mutually disentangled (unconcatenated) rings. Also note that because of the constraint preventing entanglement during collapse, a fractal globule is generally larger in size than an equilibrium globule of equal chain length (while for both the total size scales as $R \approx N^{1/3}$). The fractal globule configuration has proven to be relevant for the description of genomic DNA organization in non-dividing eukaryotic (human) cells [93,94]. In particular, it has been found that the

human chromosomes display this structure from $\Delta s \approx 500$Kb to $\Delta s \approx 7$Mb. This range of sizes might be relevant to bacteria that have genomes spanning a few Mb (also because the lower cutoff might be related to the fiber organization of chromatin, which in bacteria is different). Thus, as the authors speculate, the fractal globule description might be useful for the nucleoid as well. We will further discuss this point in the light of the available data in section 2.4.

We have already mentioned above the theoretical work on the role of macromolecular crowding on compaction [16, 80], and of depletion interactions on loop formation (and possibly on macrodomain organization [90]). A recent simulation study [124] has concentrated on the chromosome-shaping role of transcription factories. Considering a self-avoiding worm-like chain with a fixed hard-core repulsion radius, and short-ranged bridging protein complexes, they show that this system can take "micro-structured" collapsed globule configurations, where bridging complexes cluster and regions of high and low densities of interacting sites coexist in a microphase separated thermodynamic state. More abstract analytical studies could provide a useful context for understanding the role of NAP binding [125, 126].

Finally, some attention has been devoted recently to the role played by the branched plectonemic structure of the nucleoid [127, 128], which is believed to have relevant implications for transcription [8, 12]. Provided the correct questions are formulated, this topic has the advantage of being placed in a strong framework developed in the past 30 years, building on the classical calculations of Zimm and Stockmayer [129] for considering arbitrarily ramified ghost chains as Gaussian networks, and obtaining their equilibrium properties [130–132]. For example, some recent work has focused on the induction of loops involving multiple polymer segments [133]. In this framework, reliable estimates for the self-avoiding case can be obtained with Flory-like arguments, taking into account the fact that branched polymers have higher internal topological complexity, which makes the repulsive interactions stronger than between linear or circular chains. The same approach is also possibly relevant for studying dynamic aspects of loci mobility [134, 135], which can be compared with simulations and more phenomenological models [109]. Polymer modeling of the *Caulobacter crescentus* chromosome propose as an explaination for the presence of multiple and largely independent spatial domains a polymer structure which is given by a composition of supercoiled plectonemes arrayed into a bottle brush–like fiber [136]. Finally, a number of very refined results based on renormalization group / field theoretical methods have been obtained in a more abstract context using the "randomly branched polymers" model (or "lattice animals"), typically defined as the ensemble of all the clusters of connected sites (monomers) on a regular lattice. While this ensemble is probably too general, it is possible that these results have implications for questions related to the structure of the nucleoid. For example, recent work based on Langevin dynamics [137] analyzes their collapsed regime, obtaining a fractal dimension $D_f = 2.52$, intermediate between the swollen chain ($D_f = 2$) and the fully compacted globular state ($D_f = 3$).

2.4 DISCUSSION. HYPOTHESES AND PARADOXES CONCERNING NUCLEOID GEOMETRY AND DYNAMICS.

We would like to discuss here some speculations on the possible links between the experimental and theoretical results discussed above.

Let us start with an analogy to the geometrical organization of eukaryotic chromatin, where different geometrical features are observed at different scales, ranging from the known fiber organization up to the arrangement in preferred "chromosome territories" within the nucleus. In this case, experimental techniques such as Chromosome Conformation Capture (3C) and its high-throughput variants [94], or FISH (fluorescence in situ hybridization, where fluorescent tags are attached to pairs of DNA loci) allow, for instance, to measure the mean absolute distance, $\boldsymbol{R}$, and the contact probability, $\boldsymbol{P_c}$, of two genomic loci at arclength distance s. For bacteria, these techniques entail a number of specific difficulties, but the data are starting to be available, as already mentioned [36]. However, the study by Umbarger and coworkers focuses on the large-scale nucleoid 3D architecture, rather than on more detailed properties of the interaction map. Apart from saturation at large genomic distances, $\boldsymbol{R}(s)$ is typically found to increase with s [138] with an approximate power-law behavior, $\boldsymbol{R}(s) \sim s^{\nu}$. Here, ν is a scaling exponent, interpretable as $1/\boldsymbol{D}_f$, which empirically varies with scale of observation and cell type in eukaryotes. If a stretch of length s of a polymer spans a region of size $\boldsymbol{R}$, in $\boldsymbol{D}$ dimensions, it can occupy a volume $V \sim \boldsymbol{R}^{\boldsymbol{D}} \sim s^{\nu \boldsymbol{D}}$. Thus, heuristically, one expects that the probability of one end of the stretch meeting the other scales as $\boldsymbol{P_c} \sim 1/V \sim s^{-\boldsymbol{D}\nu}$. Experimentally, for loci on the same chromosome, $\boldsymbol{P_c}(s)$ decreases as a power-law $\boldsymbol{P_c}(s) \sim s^{-1}$, for a set of length scales in the approximate interval 0.5-7Mb for s [94], which provides evidence for a fractal globule-like organization, $\nu = 1/3$. This is confirmed by FISH data for $\boldsymbol{R}(s)$ on a smaller range of genomic lengths [138].

However, experimental data on $\boldsymbol{R}(s)$ for chromatin are complex, as the ν exponent is cell-type specific and varies with genomic length, reflecting different degrees and modes of chromatin compaction. At short genomic distances ν is found in the range $0.2 - 0.6$ at short distances, and $\boldsymbol{R}(s)$ reaches a plateau (i.e., $\nu \sim 0$) at order 10 Mb genomic distances, because of chromosome territories [138–140]. For the *E. coli* nucleoid, the first available quantitative data [22] was indicating that $\boldsymbol{R}(s)$ might scale like s for non-replicating chromosomes at scales of 0.3-2 Mb. A more recent result [34] which measure distances between fluorescent-labeled DNA loci of various interloci contour lengths reveled a crossover between a scaling of $\boldsymbol{R}(s)$ as $s^{0.22}$ for DNA segments less than about 300 kb to a linear scaling in s for DNA segments greater then about 300 kb; the authors claim that the exponent 0.22 is connected to the plectonemic supercoiled structure of the chromosome.

In principle, a link between nucleoid geometry and the measured nucleoid local dynamics is expected. The anomalous diffusion of chromosomal loci in *E .coli* has been modeled in terms of

fractional Langevin equations [109]; such an approach correctly reproduces the temporal behavior of the loci, but disregards the geometry of the structure containing them. Very likely, for any polymer, this structure has a fractal character, and one would like to understand how this influences the motion of the loci. A minimal model would consider a relaxation equation over the fractal. For a purely self-avoiding polymer one has $D_f \approx 1.7$. In the case of a self-attracting polymer where attraction is screened by self-avoidance, i.e. at the θ point [115,141], it is reasonable to assume a mass fractal dimension $D_f = 2$, as for a ghost chain. When the tendency towards compaction increases, one expects that D_f will increase accordingly.

In order to illustrate this point with an example, we can consider protein structures. Data from the Protein Data Bank [142] for 200 proteins with a number of amino acids ranging from $N \approx 100$ to $N \approx 10000$ give values of D_f from 2.3 up to 2.6 [143]. On a larger scale, high resolution X-ray spectroscopy has resolved the ribosome structure at the atomic level. Such data indicate [144] that the heavier 50S unit is fully compacted, with $D_f = 3$, while the lighter 30S unit has $D_f = 2.8$; it has been argued that the sparser structure of 30S is compatible with a dynamic geometry, as required in the translation process. Folded proteins are generally described as an harmonic network, by the so-called Gaussian Network Model [145], and we can try to apply a similar reasoning to the nucleoid. Close to the fully collapsed regime of a very long polymer such as a bacterial chromosome, where one expects $D_f \approx 3$, the harmonic approximation strictly does not apply, because hard-core repulsion acts against chain compression. The total energy can then be written as an harmonic "Rouse" term, describing waves propagating along the chain, plus a self-attraction term and a self-avoidance term. The relaxation dynamics in such conditions (neglecting hydrodynamic interactions) has been studied within a continuum model approach by Pitard and Orland [146]. They find that the relaxation time τ of the globule scales with the polymer length N as $\tau \approx N^{5/3}$. Taking into account that the globule size R scales as $R \approx N^{1/3}$ (i.e. it fills space), this implies that $R^2 \approx (\tau^{3/5})^{2/3} \approx (\tau)^{2/5}$ (since $N \approx \tau^{3/5}$). This scaling appears to coincide with the experimental value for the anomalous diffusion exponent $\alpha = 0.4$ measured by Weber and coworkers [100], previously quoted in the text. In other words, the Rouse subdiffusive dynamics of a collapsing (and thus off-equilibrium) globule follows the same scaling law as the observed local dynamics of nucleoid loci within a range of time scales. While this might be simply a coincidence, it leads us to speculate that the measured dynamic exponent for the mean-square displacement might be the consequence of a fractal-globule-like nature of the nucleoid, at least within a range of length scales.

This argument can be recast in more generic terms. Rouse polymer relaxation dynamics was originally explored by De Gennes [147], who obtained the scaling relation $z = 2 + D_f$, where $z \equiv 2/\alpha$ is the so-called dynamical exponent. At the θ point [115, 141] where one has $D_f = 2$, the DeGennes's relation gives $\alpha = 1/2$, the Rouse result for non-interacting chains. As the polymer dimension D_f increases, a smaller value of α is to be expected; in the compacted configuration, where $D_f = 3$, the relation gives $z = 5$, which is the result reported above. DeGennes's work is based on scaling arguments, but is confirmed by field-theoretical methods [148] for two-body

interactions.

To our knowledge, in the collapsed regime the relation has been proved only at the level of mean field [146] by modeling, in the spirit of a virial expansion, the effective interaction with an attractive two-body term and a repulsive three-body term.

In conclusion, if the nucleoid behaves like a fractal globule (i.e. an off-equilibrium polymer collapsing because of self attraction, but where entanglement is prevented by topological constraints), or more in general if it has fractal dimension $D_f = 3$, from mean field theory one expects the subdiffusion exponent $\alpha = 0.4$. Conversely, if the DeGennes relation is valid, the experimental result $\alpha \approx 0.4$, observed in *E. coli* and in large plasmids [100], implies that the nucleoid fractal dimension could be $D_f \approx 3$.

To our knowledge the available experimental result that comes closer to a direct measurement of D_f deals with the mean square displacement of fluorescently tagged replisomes [149]. In the approximation of constant replication fork velocity along the mother DNA, replication time in this experiment and genomic arclength distance have equal scaling. Hence, neglecting the global movements due to chromosomal segregation, the replisome's anomalous diffusion exponent α_R measures the effective fractal dimension $D_{f,R}$ of the replicating DNA. Specifically, the scaling $MSD(t) \sim t^{\alpha_R}$ implies $\alpha_R \sim 2/D_{f,R}$. Obviously, this measured $D_{f,R}$ in principle contains errors due to fluctuations of a stationary background as well as large scale effects associated with coherent restructuring of the nucleoid. The latter processes are relevant for segregation, but in the initial phases of replication one can assume that they can be disregarded, as we are in the presence of a "weak perturbation" of the stationary (non-replicating) structure. In such a case, the time fluctuations of the fork velocity and the steady state anomalous diffusion of genetic loci would be the main corrections to be taken into account in order to estimate D_f from the measured $D_{f,R}$. Quite interestingly the experimental estimates of the exponent α_R from [149] are $\alpha_R \sim 0.66$ and 0.58 for experiments with 30s and 5min time lapses respectively. If these numbers were confirmed by further measurements and accurate data analysis, they would support the hypothesis that $D_f \sim 3$ for a range of chromosomal scales, independently on the reasoning presented above, based on the DeGennes scaling relation $z = 2 + D_f$. In the framework of alternative interpretations for the subdiffusive behaviour which understand it as a result of viscoelastic relaxation of the cytoplasm [100], the observed segregation dynamics has been explained by the Rouse dynamics resulting from a constant pulling force of 0.13pN [150]

Clearly, for the nucleoid one expects a structure with a range of fractal dimensions in different scale regimes, as suggested by the case of chromatin and the ribosome. The existence of a range of fractal dimensions is also supported by the microscopy results for fluorescent pairs of loci reporting a linear correlation between loci distance in the cell and along the chromosome [22]. The linear correlation could result from any sequentially ordered segregation process generating a uniform mass density. For example, one might consider a periodic winding of the chromosome [47, 151], that could be produced by a segregation choreography of the type observed in *B. subtilis* as well

as in *E. coli*. However, this needs to be reconciled with the observed subdiffusion. As for the case of chromatin, a hierarchical structure is very logical, since some chromosomal functions, such as transcription, replication and DNA repair, require a certain degree of plasticity, and are not compatible with full compaction at all scales and at all times. Thus, the linear correlation of loci subcellular position and genomic distance can be consistent with $D_f \approx 3$, but further work is needed to determine the range of spatial scales where these properties apply.

The main objection against the fractal-globule as a long-lasting transient state supported by topological constraints is the ubiquitous presence of topoisomerases, DNA enzymes able to cut and paste strands and thus easily resolve these constraints. An interesting theoretical study has focused on the entanglement of tethered rings [152]; it is argued that entanglements would "condense", i.e. aggregate in space, in physically relevant situations, which, in presence of enzymes, would facilitate further the resolution of topological constraints. This kind of objection holds for both the eukaryotic and the prokaryotic case. It cannot be excluded that the non-equilibrium constraints leading to the fractal structure observed in Hi-C experiments are caused by something else, or more in general, and more plausibly, that a different physical process than simple topological constraints leads to the observed phenomenology. However, the generic reasoning presented above for connecting D_f and α might be robust with respect to these considerations.

Other approaches to chromatin organization aim at reproducing the interlocus distance $R(s)$ and the distribution of interacting loci $P_c(s)$ with alternative polymer models, such as a collapsing self-avoiding walk in a solution of organizing proteins which can bind and act as discrete self-attraction points, representing organizing proteins [153]. The spirit of this kind of study is to go beyond a dominant role of entropy, and take more seriously the "energetic part" of the free energy, and in particular the organizing proteins. The consequences of this hypothesis are explored by analyzing the resulting equilibrium structures for the polymer. Nicodemi and coworkers have recently found that such a model polymer could be close the θ point for empirically relevant protein concentrations, and small variation of the concentration of binding proteins around this state could recapitulate a considerable part of the observed phenomenology of nuclear eukaryotic DNA [140].

Finally, as mentioned above, it is also worthwhile to consider whether equilibrium statistical mechanics is even the proper starting point to understand the structure of the bacteria nucleoid. Cells expend a considerable amount of energy maintaining steady-state, non-equilibrium environments. A classic example is the membrane potential, whose existence requires an elaborate mechanism for pumping protons through the membrane. Given that the genomic information is arguably the most important part of the cell, containing both the instructions for the current cell and the inheritable information for the next generation, it seems unlikely that bacteria have evolved such that the structure of the nucleoid is resigned to equilibrium.

2.5 CONCLUSIONS.

To conclude, we briefly review some of the main features of the partial and emerging picture of the nucleoid, from the physics viewpoint. All these problems still need to be understood in a quantitative framework for bacterial physiology [26], and in particular for varying growth rates (and subcellular compositions) and during adaptation to different growth conditions [154].

A first problem that can be isolated is the explanation of its compaction/condensation properties. Likely mechanisms that can influence (positively or negatively) nucleoid condensation (and more than one can be at play) include (i) supercoiling, bending and looping, in interplay with binding of NAPs, which can cause punctual or polymer-like links building aggregation foci (H-NS, MatP) and stabilize a ramified plectonemic loop structure (Fis), (ii) consequences of molecular crowding, in the form of both phase separation and depletion interactions, and (iii) (nonequilibrium) segregation after replication, which could be induced by different physical processes.

A second, related problem is the *geometry* of the nucleoid, which requires understanding how the subunits are arranged in the cell at different scales and times. It is likely that the organization principles in a given cell state (determined e.g. by growth rate and growth phase) are different at different scales. At the micron scale, the experiments seem to converge towards a linearly-arranged sausage-shaped structure, sometimes wrapped by the Ter region, and the main outstanding questions seem to relate to the physical mechanisms behind segregation and its choreography. Below this scale, the existence of macrodomains and transcription foci still elude a physical explanation, which could be microphase separation stabilized by short-ranged attractions of chemical (organizing proteins such as MatP) or of entropic (ribosome-induced depletion interactions) origin. At an even smaller scale, an organization in blobs or fibers seems to be equally elusive, despite the existence of numerous pieces of evidence for different aspects of NAP binding and plectonemic loop formation and stabilization.

Finally, it is important to point out that within the layered information given here there lies more than one unresolved question. For example, if macrodomains are microphases structured by protein binding, then certainly these proteins must play an important role in the configurational entropy of the nucleoid, which is not considered in the arguments concerning entropy-driven chromosome segregation. Also, if the genome is compacted (at least in a range of scales) in a fractal or conventional globule configuration by attractive interactions of entropic or energetic origin, this will greatly affect its entropy, and thus its mechanical properties, loci dynamics and the interactions between segregating chromosomes. Equally important, the supercoiling-independent size of structural units measured for purified nucleoids (whose size varies with supercoiling) appears challenging for theoretical explanations. While we are certainly far from a coherent and consistent physical description of the nucleoid, there is a clear abundance of existing data and many ongoing experiments merging quantitative biophysics and high-throughput molecular biology. These emerging results, together with the fragmented but partially successful modeling approaches, make

us believe that we might be on the verge of resolving at least some of the existing issues regarding the physics of the bacterial nucleoid.

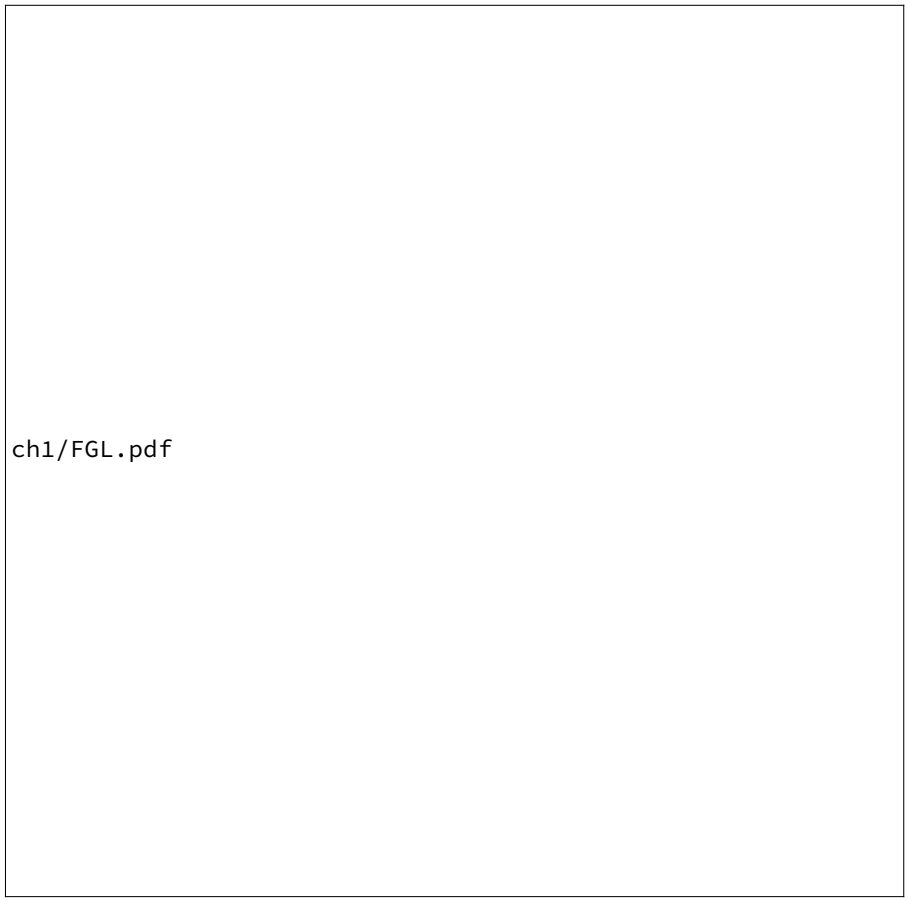

Figure 2.7: *Main universal features of the "fractal globule" model.* (A) From [93] (with kind permission from Springer Science+Business Media), root-mean squared end-to-end distance $R(s)$ and probability of a contact $P(s)$ as a function of the genomic distance s between the ends of a subchain of $N = 32000$ monomers, from simulations of a fractal globule and a collapsed (equilibrium) globule. Notice the different scalings in the two situations. All random collapsed configurations of a polymer in two dimensions behave as a fractal globule, as topological entanglement is forbidden. This does not hold in three dimensions, as a large number of knots can be generated. In a kinetically constrained situation where knots are not present, this property is restored, and the resulting metastable configuration has the property that segments that are close in arclength distance, are also close in space. It has been argued that this configuration might be relevant for eukaryotic chromatin within a range of length and time scales. (B) From [94] (reprinted with permission from AAAS), contact probability as a function of genomic distance averaged across the genome (blue) of an eukaryotic cell obtained by the Hi-C technique. It shows a power law scaling between 500 kb and 7 Mb (shaded region) with a slope of –1.08 (fit shown in cyan). This analysis indicates that the eukaryotic DNA might be organized similarly to a fractal globule.

Chapter 3

Combined collapse by bridging and self-adhesion in a prototypical polymer model inspired by the bacterial nucleoid.

3.1 INTRODUCTION

Focusing on *E. coli*, the chromosome is a single circular molecule of about 4.7 million base pairs (Mbp) ($\approx$ 1.5 mm) [2,3]. Nucleoid associated proteins, or "NAPs" (such as Dps and transcription factors Fis, H-NS, IHF, HU, and condensin MukBEF), can modify the shape of the DNA both at local and global levels [5,6,8]. Of particular interest are bridging interactions [61] (possible at least from Fis, H-NS, and MukBEF), which can in principle induce looped domain formation, through mechanisms that are believed to be important also for eukaryotic chromatin [124,155,156]. For example, a study combining super-resolution microscopy with genetic "chromosome-conformation capture" (3C) techniques on the NAP H-NS explicitly reported it to form a small set of foci in the cell, bringing together distant binding sites [91]. Additionally H-NS reduces the size of purified nucleoids [92]. RNA polymerase, the DNA-binding enzyme responsible for gene transcription, might also concentrate into transcription foci or "factories," affecting the nucleoid structure by bringing together distant loci [18,19].

The *E. coli* nucleoid, with a linear size of 1.5mm, occupies a well-defined region of the cell, with a volume of 0.1-0.2 μm^3 (the bare DNA volume is about a factor 20-30 smaller) [3]. Strong nucleoid compaction into a structure that does not fill the volume of the cell is experimentally observed *in vivo* [87, 157]. The degree of compaction is modulated by the cell's growth conditions and in response to specific external cues. Rather than confinement from the cell boundaries, the dominant force for this compaction is likely to come from self-attraction due to molecular crowding and forces of entropic origin effectively causing a short-ranged self-attraction [16, 80]. This self-adherent polymer organization is consistent with both *in vivo* observations [111, 157] and *in vitro* experiments [158] with purified nucleoids. Note that compaction from bridging alone is not

likely to be responsible of this behavior, as cytoplasm-free nucleoids are larger than cells, even if some NAPs are reported to stay bound [92, 158, 159]. Finally, a sub-Rouse viscoelastic dynamics of individual loci, whose mean apparent diffusion varies with chromosomal coordinates [100, 103] suggests that (i) a simple polymer model is not likely to fully capture nucleoid organization (ii) the organization and dynamics of inter-loci tethering might also be complex.

Importantly, the state of the nucleoid is far from being an amorphous mass, randomly organized, such as e.g., one expects from a classical equilibrium collapsed globule [93]. On the contrary, the picture of a folded object with persistent (but dynamic) mesoscopic features, including a linear ordering of loci within the cell an overall coiled shape should be more realistic [22, 24, 111, 157, 158]. In these respects, one important reported feature of the *E. coli* chromosomes are the so-called "macrodomains" [37, 39, 40]. Often described as isolated compartments, such domains are roughly replichore-symmetric, i.e., mirror the order of replication of the *E. coli* genome, from the replication origin locus, oriC, to the terminus region, Ter). The first evidence for macrodomains [37] came from a non uniform pattern in the recombination frequency between chromosomal loci (which should be proportional to the population-averaged probability that the two chromosomal segments come into contact within the cell). Four macrodomains of a few hundred Kb in size have been identified, which divide the chromosome into six contiguous regions [43, 160]. Subsequent studies have confirmed the presence of macrodomains using fluorescently labeled loci [38, 40, 41]. The Ter macrodomain appears to be condensed by a single DNA-binding NAP, MatP, which has a small set of specific binding sites in the Ter region [42]. A recent modeling study has implicated the differential condensation levels by macrodomains, together with the targeting of the Ori and Ter regions to specific subcellular positions, to the generation of the chromosome segregation pattern observed *in vivo* [58].

Additionally, nucleoids are composed of topologically unlinked dynamic domain structures, due to supercoiling (torsional constraints generated by active processes and frozen by bridging) forming plectonemes and toroids [2], and stabilized by nucleoid-associated proteins, such as Fis and H-NS. This combination of effects gives the chromosome a looped shape [12–14], where the loops form a tree of plectonemes. Supercoiling and nucleoid organization affect gene expression [8, 12, 21]. The level of supercoiling is tightly regulated by the cell, and it can be changed by the action of specific enzymes such as topoisomerases and gyrases.

While the bridging and loop-forming interactions are sometimes incorporated in polymer models of the bacterial chromosome [58,124,161–164], the standard approach is to neglect a possible self-adhesion and consider a confined polymer within the cell volume. In this chapter, we set out to investigate how the combination of bridging and homogeneous collapse in a generic polymer physics (equilibrium) framework, using computer simulations and with the help of theoretical mean-field and scaling arguments. Rather than intending this as an explicit model for the chromosome, our intent is to generically explore the consequences of these two basic ingredients, in order to help more realistic model development and to link the phenomenology with the vast existing knowl-

edge in polymer physics, and with chromosome models available in the literature [124, 160, 161]. In this spirit, we deliberately ignore the role of supercoiling, and we do not consider confinement. We also neglect segregation dynamics and other non-equilibrium drives (see the Discussion). Our main results are a qualitative characterization of the state diagram of this model, where we observe a sharp collapse to a rosette-like state where bridging interactions overcome entropy loss by loops. This collapse can be described by classical Flory-like mean-field arguments only when the density of bridges is sufficiently high, and homogeneous self-attraction modulates the collapse by affecting the interactions between rosette arms. Finally we find that bridging can form multiple rosette domains, which are stable under a wide set of conditions and whose number is set by the model parameters.

3.1.1 Methods

The process of modeling of biological matter involves the understanding of the mechanics of their constituents. We outline the foundations of polymer physics in order to give an introduction of the basic properties which characterize a polymer like DNA, we further introduce the main methods which we used to extract results from the models.

The freely jointed chain

The primary challenge in modeling individual polymeric molecules such as DNA is the very large number of degree of freedom and the many different length scales inherent in their structure that need to be taken account. A discussion about the choice of the scale of interest regarding a particular physical system is central in any simulation work: the primary reason for this being not the economical constraints, which set the power of the fastest computers available, or computational complexity of the simulations at lower scales but proven ability of simulation at macroscopic scales to produce the results accurate enough to be applied to the understanding of physical properties of system even without knowing the microscopical details.

In biophysics, the most commonly used coarse-grained models of polymer chains is the *freely jointed chain* (FJC) and the *worm like chain* (WLC) [141, 165, 166]. The FJC may be described as a linear series of a large number $N + 1$ of connected rigid rods $\vec{r}_1, \vec{r}_2, ..., \vec{r}_N$ of fixed length b interlinking the sequence of points $\vec{R}_0, \vec{R}_1, ..., \vec{R}_N$, so that the i-th bond $\vec{r}_i = \vec{R}_i - \vec{R}_{i-1}$. Each rod is free to rotate in any direction.

Since the bond vectors $\vec{r}_n$ are independent of each other, the distribution for the polymer conformation can be factorized:

$$\Psi(\{\vec{r}_n\}) = \prod_{n=1}^{N} \psi(\vec{r}_n) \tag{3.1}$$

where $\psi(\vec{r})$ denotes the random distribution of a vector of constant length:

$$\psi(\vec{r}) = \frac{1}{4\pi b^2}\delta(|\vec{r}| - b). \tag{3.2}$$

To characterize the size of a polymer, we consider the end-to-end vector $\vec{R}$ of the chain,

$$\vec{R} = \vec{R}_N - \vec{R}_0 = \sum_{n=1}^{N} \vec{r}_N. \tag{3.3}$$

for symmetry reasons $\langle \vec{R} \rangle$ is zero, but $\langle \vec{R}^2 \rangle$ has finite value which is one of the important characteristic length of the chain. Let R_0 be defined by:

$$R_0 = \sqrt{\langle \vec{R}^2 \rangle} \tag{3.4}$$

it is easy to show that

$$R_0 = b\sqrt{N}. \tag{3.5}$$

The FJC is often used to represent highly flexible polymers for there is no correlation in orientation between any of the rods, the step length b and the number of steps N can set the end-to-end distance R_0 of the chain if chosen appropriately. Although it is intuitive that the number of steps N is directly connected to the contour length of the polymer L, an introduction of the WLC model is useful to connect the parameter b with an intuitive idealization of real polymers.

The worm like chain

Another polymer model is the worm-like chain. The WLC better describe the mechanics of stiff polymers. It is specified by a continuously differentiable curve $\vec{r}(s)$ of length L parametrized by the arc length parameter s. The model is described by an Hamiltonian that quantifies the cost of bending the curve:

$$H[\vec{r}(s)] = \frac{\kappa}{2}\int_0^L \mathrm{d}s \left(\frac{\mathrm{d}^2\vec{r}(s)}{\mathrm{d}s^2}\right)^2, \tag{3.6}$$

where κ is the bending modulus. In this model, which neglect longitudinal fluctuations, an important length-scale emerges which is the *persistence length* governing the decay of tangent-tangent correlations and provides a quantitative measure of polymer's flexibility. The persistence length is defined through the relation

$$\langle \vec{t}(s) \cdot \vec{t}(s + \Delta s) \rangle = \exp(-\Delta s/P) = \exp(-\Delta s k_B T/\kappa), \tag{3.7}$$

and through this equation we can obtain the average and to end square distance:

$$\langle R_0^2 \rangle = \left\langle \left(\int_0^L \mathrm{d}s\vec{t}(s) \right)^2 \right\rangle = \int_0^L \mathrm{d}s \int_0^L \mathrm{d}s' \langle \vec{t}(s) \cdot \vec{t}(s') \rangle = 2P^2 \left(\frac{L}{P} + e^{-\frac{L}{P}} - 1 \right). \quad (3.8)$$

It follows that in the limit of flexible WLC where $L \gg P$ we have

$$R_0 = 2\sqrt{P \cdot L}. \quad (3.9)$$

A comparison of equations 3.5 and 3.9 leads to the identification of b with double the persistence length $2P$, for sufficiently flexible or long chains, while the FJC model result inadequate to describe the complete phenomenology of the WLC for the $P \simeq L$ scenario.

Due to the central limit theorem, both the FJC model in the long chain limit $N \gg 1$ and the WLC in the flexible limit have the property of Gaussian measure for the end-to-end distance $\vec{R}$:

$$\Phi(\vec{R})\ \mathrm{d}\vec{R} = \left(\frac{3}{2\pi R_0^2} \right)^{3/2} \exp\left[\frac{3}{2} \frac{\vec{R}^2}{R_0^2} \right] \mathrm{d}\vec{R}. \quad (3.10)$$

In the case of our interest, the nucleoid, the persistence length of bare DNA has been measured to be equal to $\sim$ 210nm(616bp) and a polymer arc-length of $\sim$ 1.577mm(4639kbp). Thus, at length scales greater than the mean size of a gene ($s \sim$ 1000bp), the effect of the persistence length P and of the contour length L on the end-to-end distance are not distinguishable. As we will describe in the next section, our polymer is inspired by a random flight model similar to a FJC (but we use a different $\psi(\vec{r})$ in eq. 3.2 which is ininfluential for the behavior of the model), we leave the parameter b free in order to simulate the model at different levels of discretization. Both the FJC and the WLC are ideal polymer models, the addition of interaction impose a probability bias for different polymer configurations. In the section 3.2 we describe our model and the different configurations of interactions which we took into consideration. These forces are able to cause a transition between a coiled state and a collapsed state, the most simple form of polymer collapse which is due to the effect given by the balance between repulsive steric forces and global attractions has been described in textbooks and is called the Theta transition.

The Theta transition

In real polymers, segment distant along the chain do interact if they come close to each other in space. DNA-DNA interactions in the bacterial cell, due to screening effects of the complex solvent (the cytoplasm) in which the polymer exists, are of short range nature. Those forces can be divided in two classes: attractive forces and repulsive. Thus the interaction between the polymer segments n and m can be expressed by a potential energy $u(r)$ which depends only on their separation $r = |\vec{R}_n - \vec{R}_m|$.

An obvious interaction is the steric effect: since the segment has finite volume, other segments cannot come into his own region. Because of this interaction the chain will swells, this is what is called the volume excluded interaction. The divergence of the potential energy for superimposing segments can be modeled by a infinite potential wall in $u(r)$ for small distance. Other, attractive forces, can be instead modeled at first order of approximation by the two parameters of force energy-strength and range.

The depletion of the molecules in the cytoplasm due to the presence of the long DNA macromolecule results in an effective attraction between segments. This effect have been recently shown in DNA and PET solutions [167] and has been advanced as a possible driving force of chromosomal organization [88]. In the cell, this force can be estimated to have a maximum range of ($\sim$ 5nm), which is the diameter of a typical crowding protein. The theory connecting protein crowding to an effective interaction energy has been developed [168,169], the interaction energy has been estimated for rigid spheres to be proportional to the density of the macromolecule solution. Experimental results and mean-field estimates indicate a total free energy from depletion interaction in the range of $\sim 10^3 - 10^4 k_B T$ per nucleoid [167].

As far as for the property of large length scale is concerned, the details of the interaction (i.e. the shape of $u(r)$) will not matter. The statistical properties of the chain in the case of mostly repulsive interaction has been described by Flory [170]. In this theory, the size of the polymer is the result of the balance of two effect: a repulsive excluded volume interaction which tend to swell the polymer and the elastic energy arising from the chain connectivity which tend to shrink the polymer. Consider the free energy of an ideal chain whose end-to-end vector is fixed at $\vec{R}$. This is given by

$$\beta F(\vec{R}) = -\ln \Phi(\vec{R}) + \text{terms independent of } \vec{R} \tag{3.11}$$

In the absence of interactions, the free energy is obtained from eq. 3.10 as

$$\beta F(\vec{R}) = 3\alpha^2/2 + \text{terms independent of } \vec{R}, \tag{3.12}$$

where $\alpha = R/R_0$ is called the swelling ratio. To estimate the effect of the excluded volume interactions we disregard the connectivity of the chain and calculate the interaction energy of a "segment" gas confined in a volume R^3. Since the concentration of the gas is $\rho \simeq N/R^3$, we obtain:

$$\beta F = 3\alpha^2/2 + N/2[\upsilon\rho + B_3\rho^2] \tag{3.13}$$

where υ is the second order virial expansion coefficient and B_3 is the third order. The interactions are mostly repulsive only if the parameter υ is greater than zero (how to calculate the parameter υ from the potential energy $u(r)$ is explained in the next paragraph), in this case the third-order term is not important and the average size of the polymer can be estimated from the value of α which



Figure 3.1: Illustration of the model. A: The interparticle potential comprises a hard-core repulsion between all discretization "beads" of the polymer (whose range is set by the parameter σ), and a short-ranged attraction potential of range a and depth ϵ_u for the homogeneous self-attraction (acting on all beads) and ϵ_l for the sparse bridging interactions. Additionally, consecutive beads are subject to a maximum separation hard constraint of length λ. B: Parametrization of the position of bridging interactions. A total of p bridging beads are placed across p equally spaced regions of length $\simeq N/p$. The beads in the interspersing regions only feel the weaker self-attraction of energy ϵ_u.

minimizes F. From $\partial F/\partial\alpha = 0$, we have

$$\alpha \simeq \left(\frac{N\upsilon}{R_0^3}\right)^{1/5} \quad \text{or otherwise} \quad R \propto N^{3/5} \tag{3.14}$$

which estimate the scaling of the size of the polymer in the swelled phase to the number of segments.

The coefficients υ and B_3 can be calculated form $u(r)$ using the Mayer cluster integrals, usually $u(r)$ consists of a hard core potential $u_{\text{hard}}(r)$ and a weak attractive potential $u_{\text{attr}}(r)$, in such case υ is estimated as:

$$\upsilon = \int \mathrm{d}r \left[1 - \exp\left(-\beta u_{\text{hard}}(r)\right)\left(1 - \beta u_{\text{hard}}(r)\right)\right] = A - \beta B \tag{3.15}$$

where A and B are constant independent of the temperature (but dependent of the range and energy strength of interactions). For this model, the Θ temperature is defined as $B/(k_B A)$. When the temperature is equal to Θ, the second virial coefficient is zero and the behavior for the mean end-to-end distance is the same as the ideal polymer (i.e. $\alpha = 1$), for lower temperatures ($\upsilon < 0$) the polymer collapses and the volumes occupied by the collapsed volume is set by value of the third virial coefficient B_3.

3.2 MODEL

The basic ingredients of the model are illustrated in Fig. 3.1. Our simulation uses a simple off-lattice Monte Carlo algorithm with Metropolis rejection rule. The algorithm is a variant of the "bead-spring" polymer model used in ref. [171]. The polymer is represented as a linear string of N spherical "beads" of diameter σ, connected by bonds of maximal extension $\lambda = 2.36\sigma$ We simulated polymers composed of up to 512 beads. All monomers interact via a hard-core repulsion potential

$$U_r(r_{ij}) = \begin{cases} 0, & \text{if } r_{ij} > \sigma \\ \infty, & \text{if } r_{ij} \leq \sigma \, . \end{cases}$$

Additionally, consecutive monomers feel the nearest-neighbor bonds as

$$U_b(r_{i,i+1}) = \begin{cases} \infty, & \text{if } r_{i,i+1} > \lambda \\ 0, & \text{if } r_{i,i+1} \leq \lambda \, . \end{cases}$$

The short-ranged attraction, applied between all monomers, is modeled as a negative square well between the two bounds imposed by the two above potentials, and within a maximum range of $a = 1.44\sigma$ (Fig 3.1A). The depth of the attractive potential is ϵ_u for all beads, modeling a generic short-ranged attraction due to depletion effects / molecular crowding [172]. Bridging interactions are modeled as sparsely chosen beads with additional square-well attractive potentials of the same range, but acting only on other beads of the same class with interaction energy ϵ_l.

BOX 3: Forces considered in the model

1. **The hard core repulsion**: the DNA molecule self-interact through van der Waals repulsive forces at short range. The polymer is represented as a linear string of N spherical "beads" of diameter σ which cannot overlap.
2. **The chain links force**: in the DNA molecule monomers are covalently bounded together to form a polymeric chain. The beads are connected by bonds of maximal extension $\lambda = 2.36\sigma$.
3. **Cytoplasm protein crowding**: The depletion of the macromolecules in the cytoplasm due to the presence of the long DNA macromolecule results in an effective attraction between couples of DNA double Helix segments, the range of this force is comparable to the diameter of a typical crowding protein. The short-ranged attraction, applied between all monomers, is modeled as a negative square well between the two bounds imposed by the two above potentials, and within a maximum range of $a = 1.44\sigma$.
4. **Bridging interactions**: Bridging interactions are modeled as sparsely chosen beads with additional square-well attractive potentials of the same range, but acting only on other beads of the same class with interaction energy ϵ_l.

Fig. 3.1B illustrates the criteria for placing the bridging interactions and their parametrization.



Figure 3.2: Collapse under homogeneous short-ranged attraction and Flory mean-field theory. A: Collapse curve of the swelling parameter α, plotted as a function of the attraction energy ϵ_u, for different values of N. B and C: Comparison of collapse curves with mean-field theory for polymers with different excluded volume (varying σ) and $N = 320$. Solid lines are solutions of the modified mean-field theory, Eq. (3.16), while dashed lines are solutions of the Flory mean-field theory. Panel C shows that for small values of σ the inflexion point of the collapse curve becomes very steep, and moves towards larger values of ϵ_u energy. This feature is qualitatively captured by the modified Flory mean-field theory. The numerical collapse points matched the reentrant inflexion points of the theoretical curves, indicated by stars on the solid lines in the plot.

We considered a situation where bridging beads are equally spaced, in number p; $\eta = p/N$ is the total fraction of beads occupied by the bridging regions. We are especially interested in the regime where p is fairly small (note that in this regime the value of p may be explicitly relevant, and consequently we will not refer to η alone as the control parameter).

3.3 RESULTS

Collapse from homogeneous self-attraction.

As a test scenario, we first considered the limit case of collapse of a simple polymer with homogeneous self-attraction (a homopolymer). In this case one expects to find the conventional theta transition, and this is the case in our simulations. To show this, we considered the swelling parameter α, defined here as the ratio of the mean end-to-end distance of the polymer $R_e = \langle |\mathbf{r}_N - \mathbf{r}_1| \rangle$

in a given condition and its value $R_0 = \lambda N^{1/2}$ at the theta point. Fig. 3.2A shows a plot of this quantity as a function of the homogeneous attraction energy per bead ϵ_u, for polymers with increasing N. The theta point is located where the swelling parameter equals one. We verified that its position corresponds well to the prediction of the Flory mean-field theory, which defines the theta point from the balancing, in the second virial coefficient, of the excluded volume interaction term with the attraction term. In our case both terms can be estimated from the Mayer function, respectively as $v_1 = \frac{4}{3}\pi\sigma^3$ (repulsion) and $v_2 = \frac{4}{3}\pi\beta\epsilon_u(a^3 - \sigma^3)$ (attraction), with $\beta = 1/(k_B T)$.

In our simulations, the shape of the collapse curves of the swelling parameter changes with varying excluded volume (Fig. 3.2), i.e. varying σ. As σ increases, the theta point correctly shifts towards larger energy values (Fig. 3.2B). However, for increasingly "thin" polymers, while the theta point still follows the predicted behavior, the inflexion point of the swelling ratio plotted as a function of ϵ_u becomes very steep, and radically moves towards increasingly larger values of ϵ_u instead of smaller ones. These values, and not the theta point, correspond to where the "collapse", intended as a major jump in the swelling ratio, effectively takes place. This phenomenology has been reported previously [173], and may correspond to a first-order transition related to polymer crystallization [174].

The commonly used way to find the globule size and reproduce the collapse curve by a Flory-like mean-field theory is to counterbalance the two-body interaction term described above with a three-body excluded volume term. We find improved agreement with the following variant [173]

$$\beta F = 3[\alpha^2/2 - log\alpha] + N/2[v\rho + B_3\rho^2]\,, \tag{3.16}$$

where $\rho = N/R^3$ is the concentration, $v = v_1 - v_2$ plays the role of an effective "excluded volume" parameter and $B_3 = (v_1)^2/6$. Note the logarithmic entropic term, which can be understood as coming from the prefactor of the radial distribution of a freely-jointed chain. This term becomes relevant for $\alpha < 1$ and its addition captures rather efficiently the simulation behavior described above (Fig. 3.2B). For extremely small excluded volumes, the reentrant inflexion points of the theoretical collapse curve appear to match rather well (Fig. 3.2C) the collapse points observed in our simulations. To avoid confusion, in the following we will restrict ourselves to parameter values where this sharp collapse phenomenology *does not* occur.

Collapse in presence of bridging interactions

We now discuss the case of collapse driven by sparse bridging interactions only ($\epsilon_u = 0$). This case is sometimes presented as analogous to the classical collapse transition of a homopolymer [124,161]. While we confirm this analogy, we also find that there are important physical differences between the two situations. The conventional theta point and collapse are determined essentially by a balance between excluded volume and attractive interaction. In presence of sparse bridging points, the relevant contribution to the partition function balanced by attractive energy is not monomer

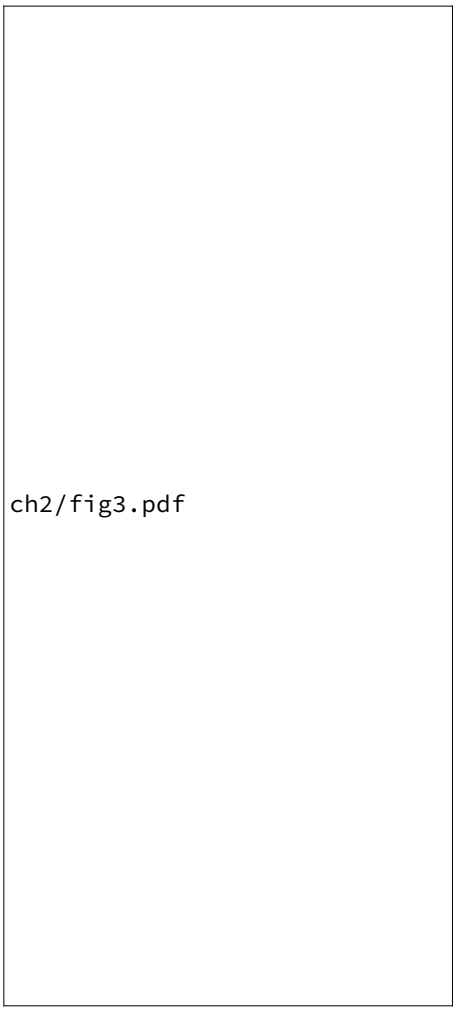

Figure 3.3: Collapse due to bridging proteins. A: Switch-like transition to a collapsed rosette-like state for polymers for small numbers of bridging interactions p, solid lines (—) show the predictions of the theoretical estimates assuming two states (Eq.3.19), while red dots (···) correspond to the full estimate for $p = 4$ (Eq. 3.18), both taking into account the measured scaling $P_{(\text{loop})} \sim g^{-2.27}$. The star polymer contribution has been neglected (see Fig. 3.7). B: Two-state dynamics of the number of active bridging interactions (contacts) for $p = 4$. The interaction energy $\epsilon_l = 6.52 k_B T$ is set close to the critical value measured in panel A. The plot shows switching between a compact and a swollen state as a function of Monte Carlo time, compatibly with a first-order phase transition showing phase coexistence between a completely collapsed state (with six contacts) and the swollen state. C: Collapse curves (swelling parameter vs ϵ_l) plotted for large p (and η); the transition energy moves to lower values for decreasing p (coded by color and symbols, see legend), the case $p/N = 1$ corresponds to the case of uniform collapse. The parameters have been chosen so that the phenomenology of Fig. 3.2C is not present. The solid lines are fits from the Flory-like theory (Eq. 3.16) on the parameters β and B_3, showing that the theory works for large $\eta = p/N$ only. D: Transition energy plotted as a function of an effective concentration of bridging proteins, estimated from p, N, by a Langmuir model at equilibrium. We assumed different effective dissociation constants K_d for the binders with a range of values typical for transcription factors [175] (1 to 500 nM). Dashed lines are guides to the eye. The region of low binder concentration is expected to produce a switch-like transition to a rosette state, while the high-concentration regime is expected to follow Flory-like mean-field behavior. An increase of K_d corresponds to a rescaling of the transition curve. Simulations were performed using 256 beads and varying $p, \epsilon_u, \epsilon_l$, with typical thermalization times of $1.2 \cdot 10^7$ Monte Carlo sweeps.

excluded volume, but rather the entropy reduction for closing loops between bridging points [90] (this kind of switch-like looping transition has also been studied in the context of transcriptional regulation [176]). In order to show this, we have considered explicitly the following simplified partition function [90, 176],

$$Z = \sum_C d_C e^{\beta n_b^{(C)} \epsilon_l - \Delta S(C)} , \tag{3.17}$$

where C runs over the possible bridging states, with degeneracy d_C and n_b is the number of bridging interactions in a given configuration. For small p this can be estimated as the number of pairs of interacting bridging monomers. For example, for $p = 4$, $n_b^{(\min)} = 0$, and $n_b^{(\max)} = 6$ (i.e. $n_b^{(\max)} = p(p-1)/2$). For large cores of bridging beads, the effects of hard core repulsion and finite-length attractive interactions make this term linear in p (proportional to the core volume). We suppose that there is only one "fully collapsed" state C^* where the bridging interactions are concentrated in a spherical shaped core (we neglect for the moment surface effects in the core, see below). The probability for such state is

$$P(C^*) = Z^{-1} e^{\beta n_b^{(\max)} \epsilon_l - \Delta S(C^*)} . \tag{3.18}$$

To estimate the entropy loss contributions for each loop, we use the first return probability for the polymer of length $g = (N-p)/p$. This can be estimated for large g as the return probability of a random walk. i.e. $P_{(\text{loop})} \simeq (a^3/\lambda^3) g^{-3/2}$; this approximation is correct for ghost polymers while the scaling for the swollen state is $P_{(\text{loop})} \sim g^{-2.27}$ [90]. While the entropy of many configurations can be estimated as a combination of loops of varying length, complex constrained configurations can emerge that are not reducible to simple loops, but the entropy loss terms can be computed case by case in a straightforward way (for example for $p = 4$ only one such configuration emerges, connecting bridging points 1-3 and 2-4). Fig. 3.3A shows a comparison of the calculation carried out for $p = 4$, and direct simulation. The agreement between the estimated and measured $P(C^*)$ is satisfactory, indicating that indeed a loop entropy reduction switch is relevant.

Additionally, we find that for p sufficiently small, the dominant contribution to the estimate is the collapsed state C^*, and the transition can be captured by a simplified two-state partition function keeping into account only the fully unfolded and the fully bridged configurations (Fig. 3.3A),

$$Z_R = (p-1)! + P_{\text{loop}}(g)^{p-1} e^{\beta \epsilon_l n_b^{(\max)}} , \tag{3.19}$$

obtained from equation 3.17 and 3.18 with $\Delta S(C^*) = (p-1) \log \left[P_{\text{loop}}(g) \right]$. The prediction looses quantitative precision for increasing p, due to the fact that the entropy reduction due to interactions between loops [177] is neglected. We verified that the estimate is accurate for a ghost chain, and produced a refined estimate keeping into account the star-polymer contributions at scaling level (see Appendix 3.5.1 and Fig. 3.7AB). Altogether, this evidence supports the picture of a switch-like transition to a rosette state, driven by the competition between bridging attraction and looping

entropy, and somewhat reminiscent of microphase separation in diblock copolymers [178].

Fig 3.3C shows the observed collapse curves for the swelling ratio for different numbers of bridging interactions p at fixed N. We noted that, for small p, the transition points show better agreement if one uses the total energy ($p\epsilon_l$) as control parameter, rather than using $1/\eta = p/N$ as rescaling factor for ϵ_l, confirming that for small p, η and p are not equivalent (in the thermodynamic limit, p has to be linear in N in order to observe collapse [179,180]). Close to the critical point, the dynamics of the swelling parameter as a function of Monte Carlo time shows switching between two well-defined states (Fig. 3.3B), a swollen one and a compact rosette-like one. This suggests that the switch-like transition is likely first order (similarly to the transitions observed in diblock copolymers).

Importantly, there is a crossover between this likely first-order and the Flory-like (second order) homopolymer collapse behavior (as shown in Fig. 3.2), when η and p are sufficiently large. Camacho and Schanke [180], working on a very similar system, suggested the possibility of a first-order collapse transition for sparse bridging, changing to second order for $\eta \simeq 0.6$. We were not able in this study to explore systematically this crossover, which resembles a standard second-order collapse for values of $\eta = p/N$ greater than $0.3 - 0.4$. However, we are not aware of existing studies fully characterizing the crossover between the switch-like transition driven by loop entropy and the Flory-like collapse. A less exotic possibility is that (as in a liquid-gas phase transition) in the thermodynamic limit the transition is always first-order, unless $\eta = 1$.

Finally, Fig. 3.3D makes a parallel between the case of sparse bridging interactions with no homogeneous self-attraction, and the "strings and binders switch" (SBS) model used in the context of eukaryotic chromatin [156,164]. The only difference between the two situations is that in the SBS model a pre-defined set of sparse bridging interactions can be occupied or not by a gas of bridging binders, while the set of bridging locations is fixed in the model considered here. To produce an estimate mapping the two situations, we considered a Langmuir process, and assumed it to be in chemical equilibrium. This procedure gives a mean value of p corresponding to a binder concentration, depending on the dissociation constant K_d of the binders (we assumed K_d within the range of values typical of transcription factors [175], around 1 to 500 nM). Specifically, $p \simeq N/(1 + K_d/c)$, where c is the concentration of binders. We find that this rough estimate leads to a state diagram that reproduces qualitatively the features of the SBS model (Fig. 3.3D).

Combination of sparse and homogeneous interactions

We now consider the case of collapse driven by both sparse bridging interactions and homogeneous self-attraction ($\epsilon_u > 0$). The configurations of the polymer under these conditions are subject to at least two different regimes, corresponding to the conditions in which homogeneous interactions are large or small with respect to the collapse energy for the homopolymer alone (we assume in both cases that the bridging interactions are stronger $\epsilon_l > \epsilon_u$). Fig. 3.4B shows different collapse curves for the swelling ratio of the polymer in the two regimes, for $p/N = 8/256$. In the regime where

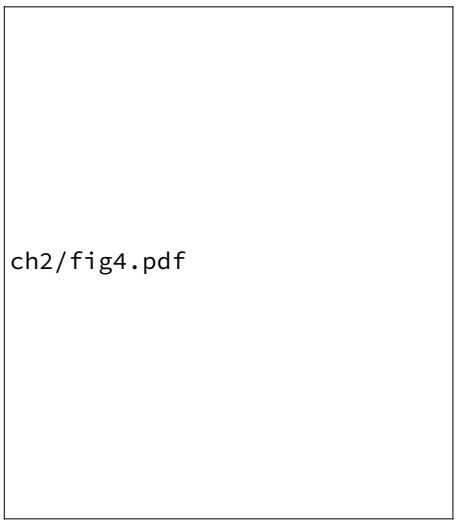

Figure 3.4: Combined effects of bridging interactions and homogeneous self-attraction. A: Snapshots of typical simulated collapsed configurations with bridging interaction stronger than uniform self-attraction. Left: the collapse is driven by bridging, while the rest of the polymer forms closed swollen loops. Right: if ϵ_u is sufficiently large, both core of bridging beads and looped arms are collapsed. B: collapse curves (swelling ratio α vs ϵ_l) shown for different values of ϵ_u (coded by color and symbols, see legend) and $p/N = 8/256$. For larger values of homogeneous attraction energy ϵ_u which put the homogeneous part of the polymer below the theta point, the formation of a core of the bridging proteins still occurs inside the formed globule, but the point at which this occurs ${\epsilon_l}^*\big|_{\epsilon_u > \epsilon_u^*} \simeq 2k_BT$ appears to depend only weakly on ϵ_u. Simulations were performed with $N = 256$ and $p = 8$ for $7.5 \cdot 10^8$ Monte Carlo sweeps, and with varying values of ϵ_u and ϵ_l (see legend).

the loops of the rosette are swollen ($v > 0$) the collapse is driven by the homogeneous interactions while in the regime where the arms are collapsed ($v < 0$) this does not happen.

For the swollen arms case, increasing uniform self-attraction (i.e. increasing ϵ_u) also shifts the observed transition point ${\epsilon_l}^*\big|_{\epsilon_> \epsilon_u^*}$ towards lower values (see Fig. 3.4B). The largest contribution to this shift comes from the reduced excluded volume interactions between the looping chains (lower v). This shift reflects partially the effect of adding the contribution of the uniform interactions directly on the bridging beads, but we verified that this contribution is considerably smaller. As previously observed [181], the collapse curve for increasing ϵ_u in the first regime resembles the collapse of a homopolymer even in the presence of a small fraction of monomers which interact more strongly. Fig. 3.4B shows that the effect of bridging interactions on the swelling ratio is negligible before a critical energy ${\epsilon_l}^*\big|_{\epsilon_u > \epsilon_u^*} \simeq 2k_BT$, which depends on p, where the bridging interactions form a core at the center of the globule. This effect modifies the swelling ratio (defined from the end-to-end vector) since the end beads are bridging.

The dependency of the free energy in the looped chains from the excluded volume coefficient v can be estimated by the analogy of the rosette-like collapsed conformation to a star polymer with $f \simeq p$ arms. Following Daoud and Cotton [182] (see Appendix 3.5.2), we estimate this entropic

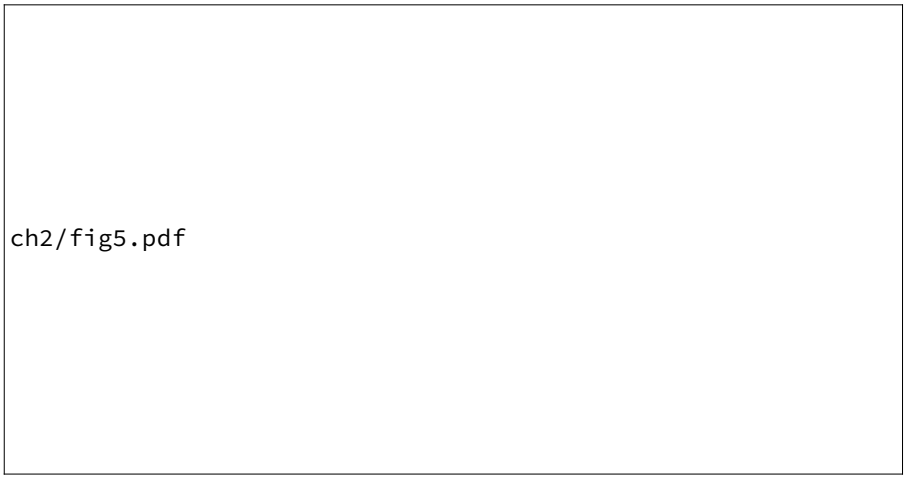

Figure 3.5: Self-assembled micro-segregation into domains and shape anisotropy. A: Snapshots and colormaps of mean-square distance between polymer segments (in units of R_0^2) shown for typical two-domain (left, $\epsilon_l + \epsilon_u = 2.4$, $\epsilon_u = 0.15$, 10^8 Monte Carlo sweeps) and three-domain (right, $\epsilon_l + \epsilon_u = 2.4$, $\epsilon_u = 0.005$, 10^7 Monte Carlo sweeps) stable configurations; multiple domains are stable due the entropic repulsion of the loops arranged in a star-like configuration. B: Stability of multi-domain configurations over single-domain collapsed state. The interaction maps of panel A are shown for increasing Monte Carlo times. The initial configuration of the simulation prepared as a single domain (left, as in Fig. 3.4A), after $9.3 \cdot 10^6$ and $3.7 \cdot 10^7$ Monte Carlo sweeps the system (parameters as in panel A) shows a disintegration of the single-domain configuration and a progressive reordering of the stable configuration in distinct clusters (middle), (right) at 10^8 sweeps a two-domain configuration is stable. Simulations are performed with $N = 512$, $\sigma = 0.027$ $\lambda = 0.057$, $p = 32$.

contribution to the free energy to scale as

$$\beta F_{(\text{corona})} \sim f^{3/2} \log \left(N f^{7/6} v^{1/3} \left(\frac{b}{r_0} \right)^{5/3} \right) , \tag{3.20}$$

for $v > 0$ where b is the monomer size and r_0 the size of the core. Inserting Eq. (3.20) as additional term in the mean-field theory leads to predict a logarithmic shift of the transition point from $v = 0$, which is consistent with simulated data (see Fig. 3.4B and Fig. 3.7C).

Compartmentalization

Intriguingly, we find that the switch-like collapse, in the presence of sparse interactions can lead to long-lived states where *multiple* collapsed micelle-like domains coexist. Fig. 3.5A shows snapshots of such configurations. Multiple domains are also visible from interaction maps, i.e. tables where the indexes are the discrete arc-length coordinates of the polymer beads and the entries are proportional to the mean-square distances between beads with given coordinates (Fig. 3.5AB). States with multiple domains were stable in our simulations for as long as we could measure. Additionally, states prepared with the initial conditions of a single domain would switch to a two- or three-domain state, which was then observed to be stable (Fig. 3.5B). This observation leads us to believe that the multi-domain configurations might not be metastable, but true equilibrium states. Addi-



Figure 3.6: The shape anisotropy due to multiple domains (measured as the ratio between the first and the third eigenvalue of the inertia matrix) shows a transition when the attractive energy ϵ_u crosses a threshold. The transition is due to coalescence of multiple domains. Simulations are performed with $N = 512$, $\sigma = 0.027$ $\lambda = 0.057$, $p = 32$.

tionally, the states can be still observed in presence of homogeneous self-attraction, i.e. for $\epsilon_u > 0$, and the existence of such states affects the size and the shape anisotropy of the collapsed globules, measured as the ratio between the first and the third eigenvalue of the polymer's inertia matrix (Fig. 3.6). This analysis suggests that the elongated multi-domain configurations can be stable only if ϵ_u is such that the polymer arms are above their theta point.

In order to argue that multi-domain states could be stable, we rely on the following scaling argument, related to the case of polysoaps [183, 184]. We consider a configuration with q domains, made of a "core" of p/q bridging monomers and a "corona" of $p - 1 \simeq p/q$ loops. Treating the loops of each domain as a star polymer [183] with $f = p/q$ "arms", and considering the leading contribution to the entropy scaling $f^{3/2}$ [182] leads to estimate the free energy change by subdivision into q domains as

$$\Delta F_{(\text{corona})} \sim p^{3/2} q^{-1/2} \tag{3.21}$$

The remaining relevant terms in the free energy are energetic, and contain a core volume term (proportional to ϵ_l) which is not affected by the partitioning the polymer into domains, and the surface tension term leading to the change

$$\Delta F_{(\text{core})} \sim \epsilon_l (p)^{2/3} q^{1/3} \,. \tag{3.22}$$

Minimizing the two contributions with respect to q (assumed continuous) leads to the expected equilibrium number of domains

$$q_{\text{eq}} \sim p \epsilon_l^{-6/5} \,, \tag{3.23}$$

indicating that the number of stable domains should increase for larger p and decrease if ϵ_l is too large.

This behavior is observed in our simulations. However, this argument can be regarded as only qualitative. For example, the estimate in Eq. (3.20), since $r_0 \sim q^{1/3}$ implies $\Delta F_{(\text{corona})} \sim$

$p^{3/2}q^{-1/2}\log\left(p^{5/3}q^{-2/3}\right)$, which (even neglecting prefactors) affects the predicted scaling. Additionally, the argument for the entropy is consistent only when the arms are above the theta point ($v > 0$). Note that, differently from the case of polysoaps, the multi-domain states in this picture are due to the trade-off between the energetic cost of the core surface when making multiple domains and the entropy increase due to making star-like configurations with an increasing number of arms. Hence, according to the argument, a multi-domain state is stable because partitioning $(p-1)$ arms into more than one domain is less costly compared to the energy cost of partitioning the bridging proteins in multiple domains. Finally, for large cores, an additional contribution to the entropy can be expected from the non-bridging beads that are "buried" in the core because of their vicinity to bridging beads. The most naive estimate for this additional cost for core size assumes that if a bridging bead is at position $r < r_0$ within the core, it brings in $(r_0 - r)^2$ beads (i.e. buried beads will behave like a theta-point polymer within the core). since there are order r^2 beads in each shell of the core, integrating along the radial coordinate will make the cost scale as r_0^5 (i.e. $p^{5/3}$) explicitly favoring fragmentation into multiple domains. While this effect might only need to quantitative corrections in our simulations (since typically p is small), it might be important in realistic situations.

3.4 DISCUSSION AND CONCLUSIONS.

We have analyzed a generic model of polymer collapse driven by a combination of homogeneous and sparse attractive interactions. Quantitative scaling arguments and simulation in parallel allowed us to access some basic aspects of the equilibrium behavior of this system. There are two main results. First, we find a crossover between Flory-like collapse and a switch-like, presumably first-order compaction where bridging counterbalances loop formation. Both phenomenologies often feature as ingredients of simple physical models for chromatin [90, 93, 120, 155, 156], but the interplay between the two is relatively inexplored. Second, states with multiple micelle-like domains can exist, in a manner that is reminiscent of micellar polysoaps [184]. We find that such multi-domain states are stable: similar conclusions were reached by a recent study of a polymer model with sparse bridging interactions, motivated by transcription factories [124], and motivated with a Flory-like theory for a macroscopic extended network of foci. To explain these structures, our work takes the complementary approach of considering the stability of star-like rosettes.

Both of our main results are backed a series of simple mean-field/scaling arguments. While little of this knowledge can be called radically new, our combined simulation and analytical approach helps linking the results with generic knowledge in different sectors of the polymer physics literature, including mean-field and micelle-like collapse [90, 173], polysoaps [184], star polymers [177, 182]). Overall, we believe that there is an importance in elucidating these links, especially because they are not always kept into account in the current landscape of models of (bacterial) chromatin [160, 161]. Interestingly, the difference between "homogeneous" (Flory-like) and "heterogeneous" (switch-

like) collapse has been explored in the 1990s with the motivation of protein folding [179, 180, 185]. Dynamically, they are related to the difference between "downhill" folding, where large gains in stabilizing energy and loss in conformational entropy are balanced in a way that a large range of structures can be observed at the same time, and "two-state" folding, where intermediate structures do not matter. Biologically, one can imagine that the collapse and swelling of genomic regions may be tuned to be switch-like or second order, in order to be differentially controlled externally by the cell. Additionally, theoretical arguments are useful to elucidate the main ingredients causing a specific effect. In particular, they lead us to speculate that the stability of multi-micelle configurations is conferred by the competition between surface tension of the core and the entropic cost of the corona, which is similar to a star polymer. Our simulation results appear to be in line with this hypothesis. A further entropic cost due to burial of non-bridging bonds in the core is speculated to play a role for large cores. In analogy with diblock copolymers, one can speculate the existence of more complex micellar phases, for example with cylindrical symmetry, which might be exploited biologically. Overall, compared to previous literature, our work provides a more precise analysis of the collapse transitions by comparing simulations in different regimes and discusses in some more detail the orders of transitions. Additionally, we address the role of homogeneous self-attraction, which is usually disregarded in studies motivated by genome organization.

The model explored here is solely based on self-adhesion and bridging, as motivated by recent observations on bacterial chromatin. Other possibly important factors were voluntarily left out, in order to obtain a cleaner description of theoretical consequences of these two ingredients. A possibly very important feature of bacterial chromosomes disregarded by this model is the role of supercoiling, and the effectively branched structure of plectonemes. Modeling work on *Caulobacter* [136] supports the hypothesis that supercoil loops induced by transcription, using an effective numerical description of supercoil-induced branching. A more detailed model indicates that supercoiling facilitates the probability of protein-induced bridging [186]. Regarding the specific findings reported here, on the light of these studies we believe that a branched structure induced by supercoiling could modulate both the loop formation entropy and the loop-loop interaction entropy, thus affecting both the contact map and the polymer size at fixed conditions. Due to limitations of our simulation technique, we also left out from this modeling framework topological constraints, which have been implicated for eukaryotes [93, 187]. Such constraints lead to long-lived metastable states in a collapsing polymer or a melt of rings, characterized by a dense fractal-like organization (and in contrast with the dense but non-fractal organization of an equilibrium globule). We have previously hypothesized [160] that the peculiar sub-diffusive dynamics of *E. coli* chromosomal loci [100, 103] might be connected to this kind of organization.

It would be premature to draw any clearcut biological conclusions based on this simple model regarding phenomena occurring in real chromosomes. The main biological insights of this work are the generic notions that the features and even nature of the collapse transition may be tunable by modulating the sparsity of bridging and the homogeneous self-adhesion, and that the tendency

to form domains could be intrinsic of the bridging (and tuned by the osmotic self-adhesion), and require in principle little or no inter-specificity of domains [164]. This kind of spontaneous sorting mechanism might play a role in the observed correlation between the position of genetic loci along the chromosome and their position in the cell [22, 42], and possibly also in the resolution of the identity of segregating sister chromosomes [58, 188].

3.5 APPENDIX

3.5.1 Role of loop interactions in the collapse due to sparse bridging.

This Appendix briefly addresses how accounting for loop interactions can improve the estimates for the switch-like collapse presented in Fig. 3.3A. The scaling of the probability of collapse for swollen arms depends on the excluded volume interaction parameter v through two effects. Firstly, as already mentioned, and accounted in the estimates, the swelling of the arms due to the self-interactions changes the scaling for the contact probability from $P_{\text{loop}} \sim g^{-3/2}$ to $P_{\text{loop}} \sim g^{-2.27}$ [90]. Secondly, and more importantly, there is an interaction, so far neglected between the arms due to the peculiar star-like configuration of the collapsed polymer. The estimates of the switch-like transition from Eqs. 3.18 and 3.19 do not show perfect agreement with simulations (Fig. 3.3A) with increasing p because this entropy reduction due to the excluded volume interactions between loops is neglected. This is confirmed by the fact that the approximation of Eq. 3.19 (valid for sufficiently small p) is very accurate for ghost chains (Fig. 3.7A).

A loop interaction term must depends on the arm length g as well as from the interaction parameter v and the number of arms f. The entropy reduction for a star polymer with f arms of length N/f has been studied in the thermodynamic limit ($N \to \infty$) [189]. One can define a set of scaling exponents γ_f which define the scaling between the number of configurations of a star polymer made of f arms each of length N/f as

$$Z = \mu^{-N} (N/f)^{\gamma_f - 1} \ . \tag{3.24}$$

Due to the effect of this scaling, the collapse energy of the polymer with bridging interactions is shifted to higher values for increasing p by an entropic contribution as [190]

$$\Delta S_{star} = \sigma_{2p} \log \frac{N}{p-1}, \text{ with } \sigma_{2p} = \gamma_p - \frac{\gamma_1 + 1}{2} \ . \tag{3.25}$$

Numerical calculations of γ_f for selected values of f have been carried out using field-theoretical methods [189] and lattice polymer simulations [177, 191–195].

In a complementary way, the problem of the star polymer has been approached by scaling ar-

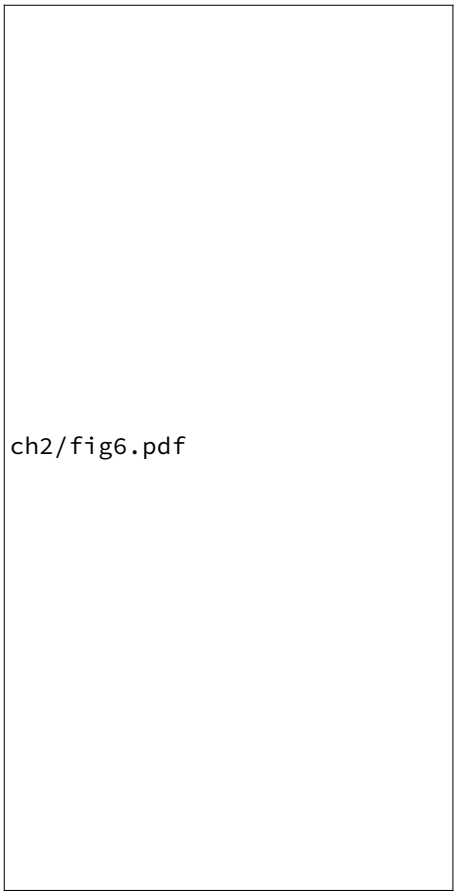

Figure 3.7: Role of entropy loss due to loop excluded volume interactions in the collapse due to sparse bridging. A: Collapse of a ghost polymer with bridging interactions. In this case, the inter-arms interactions are absent, and the plot shows good agreement of the predictions of Eqs. 3.18 and 3.19 to simulation data. The formulas include a scaling of $P_{\text{loop}} \sim g^{-3/2}$ measured from simulations. B: improved theoretical predictions by a shift in the transition energy of value $\Delta S_{star} = a_0(2p-3/2)^{1.68}\log[256/p]$ (see Eq. 3.26) where $a_0 = 0.00258$ is a fitted parameter. C: Agreement of the scaling of the transition point between the simulations for the polymer with bridging interactions ($p/N = 8/256$) in the swollen arm regime and the predictions of Eq. 3.20.

guments based on polymer blobs [182]. This approach shows that the scaling of the free energy of the star polymer is proportional to $\sim f^{3/2}$ [196], setting a scaling for γ_f. The simulations used to compute γ_f numerically gave [177] the relation

$$\gamma_f - 1 \sim -(f - 3/2)^{1.68} \tag{3.26}$$

which is in agreement with the scaling estimate.

The simulations and scaling estimates discussed so far do not account for the excluded volume strength. In order to include the role of excluded volume, we propose a scaling for the entropic cost of the star polymer term as in Eq. 3.20 (see Appendix 3.5.2), which includes also the logarithmic correction factor depending on the homogeneous interaction term (and containing the parameter v measuring interaction strength). Fig. 3.7B shows that adding such a term (scaling as Eq. 3.20) in Eq. 3.19 sensibly improves the agreement of the theoretical estimate with the simulated switch-like collapse. Fig. 3.7C specifically tests the role of the logarithmic dependency of the transition point from v, finding satisfactory qualitative agreement.

3.5.2 Scaling argument for the entropy of a star polymer.

These notes sketch the calculation of the number of blobs of a star polymer with f arms, each of length N, following Daoud and Cotton [182]. The number of blobs are then used as a proxy for the entropy (blob ansatz). The polymer is described as a series of concentric shells, each of which by definition contains f blobs (one for each arm). The size of the blobs ξ depends on the radial coordinate of the shell (called r). Since the shell at r has surface $\sim r^2$, the size of each blob is $\xi^2 = r^2/f$, which means that $\xi \sim r/f^{1/2}$, i.e. $\xi = ar$, with $a = Cf^{-1/2}$. ξ is also the thickness of the shell at coordinate r.

Starting form a core of size r_0 we now imagine to stack (iteratively) discrete blobs of the proper size on each shell. Each stacked shell will determine the coordinate of the following one, and hence its blob size.

$$r_i = a \sum_{k=1}^{i-1} r_k + a\frac{r_i}{2} + ar_0 \,,$$

with

$$\xi_i = \xi(r_i) = ar_i \,,$$

leading to the expressions

$$r_n \left(1 - \frac{a}{2}\right) = r_0 a(1+a)^{n-1} \,,$$

and

$$\xi_n = r_0 a^2 (1+a)^{n-1} \,.$$

We note now that the relationship $\xi_n \sim g_n^{3/5} v^{1/5} b$, also has to hold (each blob is a swollen

polymer of bond length b), where $v = 1/2 - \chi$, and g is the number of monomers in a blob. Hence,

$$g_n = Ae^{\frac{5}{3}\left(n-\frac{1}{2}\right)a} ,$$

where

$$A = \frac{(a^2 r_0)^{5/3}}{v^{1/3} b^{5/3}}.$$

We now follow (on one branch of the star) all the monomers in all the blobs, and impose that their total has to be N. Inverting this relationship gives an estimate for $\beta F \simeq f N_{\mathbf{blob}}$ (since the argument involves the blobs of one arm). One has

$$\sum_{n=1}^{N_{\mathbf{blob}}} g_n = N ,$$

hence

$$N \simeq \int_1^{N_{\mathbf{blob}}} \mathrm{d}n A e^{\frac{5}{3}na} ,$$

from which

$$N_{\mathbf{blob}} \sim \frac{1}{a} \log Na/A$$

where we neglected additive and multiplicative numerical constants, leading to the expression

$$\beta F \sim f^{3/2} \log \left(N f^{7/6} v^{1/3} \left(\frac{b}{r_0} \right)^{5/3} \right) ,$$

which is Eq. (3.20).

Chapter 4

Interplay of global transcriptional regulation and chromosome organization of bacteria

4.1 INTRODUCTION

Different experimental results suggest the presence of an interplay between global transcriptional regulation and chromosome spatial organization in bacteria [5,8,197,198]. The success of a cell's survival under different growth conditions depends on its ability to regulate a coordinated transcriptional response to specific environmental changes or stresses, involving large groups of genes [1, 199]. In bacteria, transcriptional regulation of genes depends on the binding of proteins to DNA, but also on the physical configurations of the resulting mesoscopic protein/DNA complex, the nucleoid [8, 18, 200, 201].

Nucleoid organization itself remains to be fully characterized. Most of NAPs have been identified [5], and in some cases their local action is well known: for example, DNA bridging by H-NS, StpA, or Fis is believed to be important for DNA loop stabilization [202]. Such supercoiled plectonemic loops can be topologically isolated from the rest of the genome, modulating protein binding and transcription rates [200].

The relation between nucleoid state and gene expression has been the subject of several recent studies. Experimentally, this question has been addressed using transcriptomics and chromatin immunoprecipitation combined with microarrays (ChIP-chip, see sec. 4.1.1). Experiments have characterized RNA-polymerase (RNAP) and specific NAP binding by ChIP-chip [66,73,203]. By monitoring generic protein binding throughout the chromosome [68], Vora *et Al.* have found extended protein binding regions (called "extended protein occupancy domains" or EPODs) connected to either transcriptionally silent or highly expressed clusters of genes. Finally, transcriptomics has been applied to the study of the effects of the knockout of specific bridging nucleoid protein and to changes in the level of negative supercoiling [69, 75, 76].

Computationally, attempts have been made to characterize the binding specificity of NAPs [204, 205] and to interrogate the one-dimensional arrangement of genes of different categories for signals possibly related to nucleoid structure [122, 206]. In particular, a thread of studies [47, 151, 207, 208] on possible "periodicity" patterns has found a number of interesting spatial regularities in the arrangement of genes belonging to different categories. For example, the correlation of gene expression with gene codon bias, has recently been related to large contiguous "sectors" along the chromosome [47].

In this context, we present several data analysis approaches which we developed in order to test possible connections between the physical organization of the nucleoid with gene expression (RNA-Seq, see sec. 4.1.1) and protein binding (ChIP-Seq, see sec. 4.1.1) datasets. This chapter contains a description of the NuST webtool, which consists of a database of data from published experimental and computational studies and an implementation of simple multi scale statistical analysis tools. The chapter continues with a study of the epistatic effect of the homologous H-NS and StpA genes knock out in *E. coli*, those genes, if deleted independently, do not show strong effect on phenotype, while double mutant show a loss of fitness. Epistatic effects are evaluated from the trascriptomic point of view using RNA-seq technology: we show that H-NS-repressed genes belong to two types, termed "epistatic" and "unilateral". In the absence of H-NS, the expression of "epistatically controlled genes" is repressed by StpA, whereas that of "unilaterally controlled genes" is not. We evidence that there is a bi-layered gene silencing system - involving the homologous H-NS and StpA - operating on horizontally-acquired genes among others, furthermore the expression of these genes caused by H-NS and StpA deletion, perturb the gene expression homeostatic equilibrium and thus represses the expression of more important genes. Using laboratory evolution we show that two global gene regulatory mutations – (a) point mutations inactivating the stress-responsive sigma factor RpoS or σ_{38} and (b) an amplification of $\sim 40\%$ of the chromosome centered around the origin of replication – converge in partially reversing the global gene expression imbalance. In the last part, we introduce a correlation study of a large number (about 300) of genome-wide expression data-sets, also compared to the outcome to the published genome interaction map (HI-C, see sec. 4.1.1) data. Finally, we discuss how the qualitative results of the polymer model presented in the previous part of the thesis may be compared with experimental data.

4.1.1 Experimental methods, conceptual description

Sequencing is the process by which it is possible to determine the precise order of nucleotides within a DNA molecule. Recent advances in DNA sequencing have revolutionized the field of genomics, making it possible for single research groups to extract large amounts of sequence data. Owing to the tremendous progress in sequencing technology, new techniques have been elaborated with the scope of implementing basic research on protein-DNA interactions, gene transcrip-

tion and chromosome conformation [209, 210]. The following paragraph we will briefly describe those new techniques.

BOX 4: Glossary of sequencing applications

The following protocols has been widely applied to the study of chromosome conformation and transcriptional control:

- *ChIP-seq*: A method that allows the binding sites for a specific protein to be identified *in vivo* throughout a genome based upon crosslinking, immunoprecipitation, and DNA sequencing. *ChIP-chip* refers to the same technology where the sequencing step is replaced by a microarray essay.
- *RNA-seq*: An approach to transcriptome profiling that uses deep-sequencing technologies. RNA is converted in a library of cDNA, sequenced and mapped to the reference genome.
- *Chromosome conformation capture*: A technique that identifies physical interactions between pairs of genome segments. Macromolecules are chemically crosslinked in living cells, and then the DNA is extracted, digested with a restriction enzyme and subjected to intramolecular ligation. The use of sequencing , microarrays or PCR (Polymerase Chain Reaction) permit to map the interactions to the reference genome.

ChIP-Seq

Genome-wide mapping of protein–DNA interactions and epigenetic marks is essential for a full understanding of transcriptional regulation. The main tool for investigating these mechanisms is Chromatin immunoprecipitation (ChIP), which is a technique for assaying protein–DNA binding in vivo [211, 212]. Using chromatin immunoprecipitation (ChIP) followed by massively parallel sequencing, the specific DNA sites that interact with transcription factors or other chromatin-associated proteins can be profiled. The ChIP process enriches the crosslinked proteins or modified nucleosomes of interest using an antibody specific to the protein or to a flag tag attached to it. Purified DNA can be sequenced on any of the next-generation platforms. A schematic representation of the pipeline is reported in Fig. 4.1.

Reads from sequencing can be mapped to the *E. coli* K12 MG1655 genome using publicly available software packages such as BWA [213]. Gene annotations can be obtained from different on-line databases such as Ecocyc [214]. Due to the presence of multiple replication forks, which permits *E. coli* growth rate to be faster than the replication rate, the aligned data typically present an Ori-TER gradient of coverage which is proportional to the gene copy number. This trend can be subtracted by normalizing the data with a Mock IP (a ChIP-seq experiment without antibody). In order to define binding regions, a cutoff read count must be defined, and consecutive stretch of DNA with read count greater than or equal to the cutoff should be flagged, pairs of adjacent regions so obtained need to be merged to give a single binding region. The PeakSeq algorithm [215] can give a probabilistic estimation for the significance of the binding in those regions [67].



Figure 4.1: Schematic representation of a Chromatin Immuno Precipitation protocol: DNA is cross-linked to proteins *in vivo* in order to enhance the binding, the cells are lysate and DNA is fragmented by sonication. Interactions with DNA are isolated by a specific antibody. After reversing the cross-link, the DNA is sequenced and the binding profile can be obtained by alignment to a reference genome.

RNA-Seq

The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific physiological condition [216]. RNA-Seq uses recently developed deep-sequencing technologies. In general, a population of RNA is converted to a library of cDNA fragments. Each molecule, with or without amplification, is then sequenced in a high-throughput manner to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing).The reads are typically 30–400 bp, depending on the DNA-sequencing technology used. (See Fig. 4.2)

Most transcriptome studies are interested in comparing the expression levels of the same gene across multiple conditions. Therefore, methods which are directed at preparing RNA-seq data for computing differential expression of the same gene across multiple conditions have been developed. The data from different conditions need to be normalized for comparison. Two statistical approaches - EdgeR [217] and DESeq [218] are readily available as packages within the Bioconductor suite.

Chromosome conformation capture

Physical interactions between elements can be detected with the Chromosome Conformation Capture (3C) method [219]. 3C uses formaldehyde cross-linking to covalently trap interacting chromatin segments throughout the genome. Interacting elements are then restriction-enzyme-digested and intramolecularly ligated (Fig. 4.3). The frequency with which two restriction fragments become ligated is a measure of the frequency by which they interact in the nucleus.

3C uses PCR (Polymerase Chain Reaction) to detect individual chromatin interactions, which is particularly suited for relatively small-scale studies focused on the analysis of interactions between a set of candidate elements. PCR is a technology to amplify specific sequences of DNA, identified by opportunely synthesize primers, after extraction by replication *in vitro*. After PCR the amount

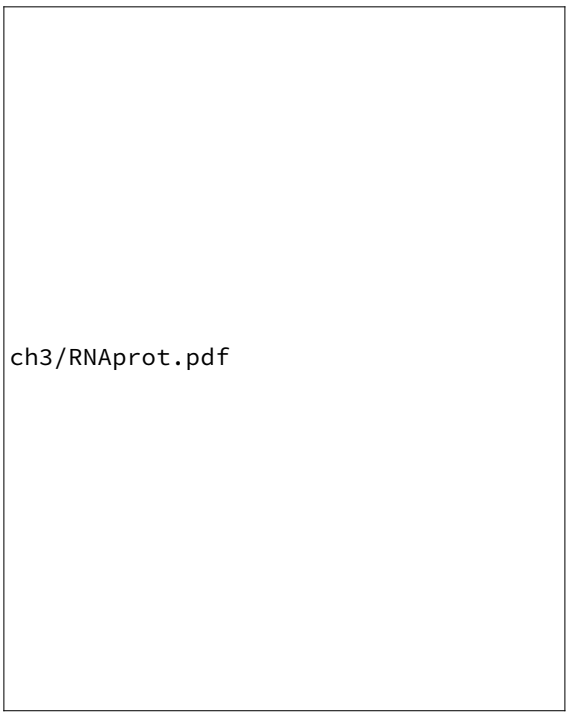

Figure 4.2: Schematic representation of a RNA sequencing protocol: after RNA extraction from the cell, a ribosomal RNA cleanup is needed in order to obtain a greater coverage. The RNA library is then copied on cDNA in order to be sequenced. A comparison of the read counts per genes in different bacterial strands results in an estimation of the differential expression.

of amplified signal can be quantified using fluorescent dyes. However, simply PCR detection is not conducive to *ab initio* and large-scale mapping of chromatin interactions. To overcome this problem, 3C libraries need to be analyzed using a high-throughput detection method such as microarrays or DNA sequencing.

After crosslinking and fragmentation using restriction enzymes, the segments spanning a ligation site needs to be isolated for identification, this fundamental step is done by using a carbon copy in the 5C variant of 3C [220], or by using biotinylated nucleotides during ligation and isolating it using streptavidin as in Hi-C [94, 136, 221].

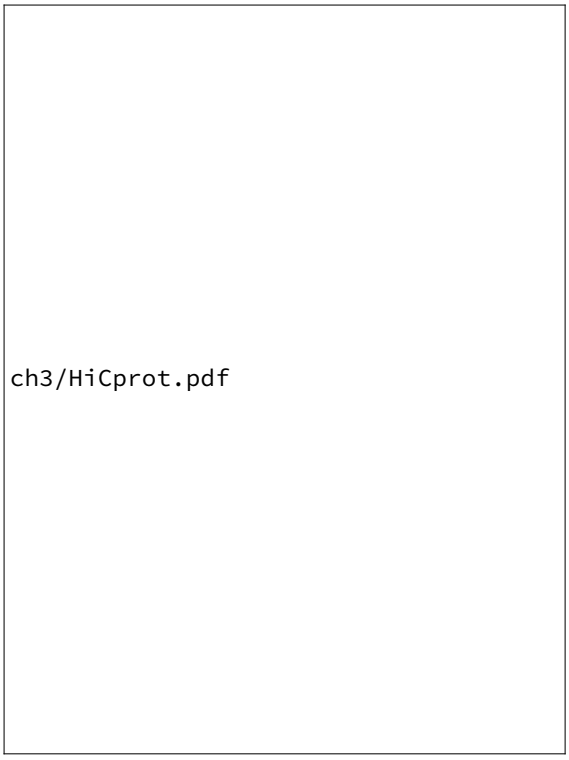

Figure 4.3: Schematic representation of a chromosome conformation capture protocol: DNA is cross-linked in the cell so that regions that may be distal in primary sequence but adjacent in space are cross-linked together. A restriction enzyme is used to fragment the cross-linked DNA and the ends of fragments ligated together. This ensures that spatially proximal chromosomal regions are now covalently linked to be adjacent in primary sequence. After reversing the cross-links, the DNA is fragmented and those spanning a ligation site specifically isolated (using a carbon copy as in 5C, or by using biotinylated nucleotides during ligation and isolating it using streptavidin as in Hi-C). Paired-end sequencing of DNA on either side of the ligated site helps link loci together and generate contact frequency/probability maps.

4.2 NUST (NUCLEOID SURVEY TOOLS)

This section describes NuST (Nucleoid Survey tools); NuST is a web tool, based on the *E. coli* genome, which gives the non-expert the possibility to analyze the aggregation of genes or loci sets along the genome coordinate, at different scales of observation.

The identification and clear visualization of spatial clusters of contiguous genes targeted by specific DNA-binding proteins or sensitive to nucleoid perturbations can elucidate links between nucleoid structure and gene expression patterns. Similarly, statistical analysis to assess correlations between results from independent experiments can provide the integrated analysis needed in this line of research. It is useful to discover correlations between different sources of data (e.g. expression, binding, or genomic data) and genome organization. A user can use it on data sets in the form of gene lists coming from his/her own experiments or bioinformatic analysis, but also make use of the internal database, which collects data from many published studies.

4.2.1 Availability and Implementation:

NuST (Nucleoid Survey Tools) is a web server (available at `http://www.lgm.upmc.fr/nust/`). The website is implemented in PHP, SQLite and Ajax, with all major browsers supported, while the core algorithms are optimized and implemented in C. NuST has an extensive help page, and provides a direct visualization of results as well as different downloadable file formats. A template Perl code for automated access to the web-server can be downloaded at `http://www.lgm.upmc.fr/nust/downloads/`, in order to allow the users to use NuST in systematic bioinformatic analysis.

4.2.2 Introduction

The complex interplay between chromosome organization and gene expression requires integration of data from different high- and low-throughput experiments with statistical analysis at multiple scales. The web server described here is an effort to fill this gap. Since many of the nucleoid structural features (e.g. supercoil domains and macrodomains) involve contiguous genomic regions, a typical hallmark of nucleoid-mediated regulation is aggregation along the chromosome of genes having specific properties [47, 49, 222]. The main tool is able to identify significant linear aggregation clusters of a gene set, considering different observation scales [49].

The clusters are represented graphically on-line to give an immediate grasp on the results and their significance and a direct comparison with known chromosomal sectors and macrodomains. In order to integrate data sets from independent sources, different sliding window histograms representing gene positioning can be overlayed, and compared using the local Pearson correlation coefficient.

A collection of many relevant experimental results is stored in an internal database, to allow the user to compare/integrate with multiple published data sources. The already fairly rich internal database will be progressively extended, and experts in the field are invited to contact us to upload their results.

The web server is currently based on *E. coli*, for which detailed information on both gene expression and chromosomal organization is available. We plan an extension to other bacteria, in order to allow multi-species comparisons in the future. An extensive step-by-step documentation introduces analysis that can be performed using the web server and is divided into an introductory help page, and a "learn by example" page guiding the user through the interpretation of the results and the choice of the parameters.

Finally, a template Perl script can be downloaded from the website. This script allows to query the server and download results in different formats. The script is extensively commented and can be extended to produce general codes that systematically access the web server to upload data sets, perform analysis and download analyzed data.

Some of the algorithms implemented in NuST were previously used to show the existence of contiguous clusters of nucleoid-sensitive genes and to compare their position to the macrodomain positions [49]. The results suggest the possibility of a nucleoid-mediated switch between a motile and a biofilm lifestyle, proving that the analysis available in NuST can unveil relevant biological information and potentially motivate further experimental research.

4.2.3 Detailed description

The input data sets are single column text files with one gene ID for each row. Standard gene IDs are Regulon DB database [223] (for different gene IDs the server proposes synonyms). Sample data sets as well as the complete list of accepted synonyms can be obtained from the "Download" page. Loaded data sets are stored in the "Personal" part of the internal database and can be accessed for further analysis. They are deleted at the end of each anonymous session. A login (obtained sending an email to the administrator) allows to keep personal data for multiple sessions. The "Common" part of the database contains data sets from the literature, organized by type of data and experimental technique. The information about the data sources, is shown by clicking on a data set name. A new data set can be created merging or intersecting existing data. As the intersection between the two selected lists is taken, the result of a hypergeometric test assessing its statistical significance is printed. The list of "Common" dataset with description is summarized in table 4.1.

The main features of the server are described in Figure 4.4. The "linear aggregation" analysis (Figure 4.4A) detects significant aggregation along the genomic coordinate, by considering the density of genes at different scales on the genome using grids of different bin sizes, and comparing with 10000 realizations of a random shuffling null model [49]. The output is directly visualized on the web site with two bitmap pictures and a table (Figure 4.4B). A "cluster plot" shows the sig-

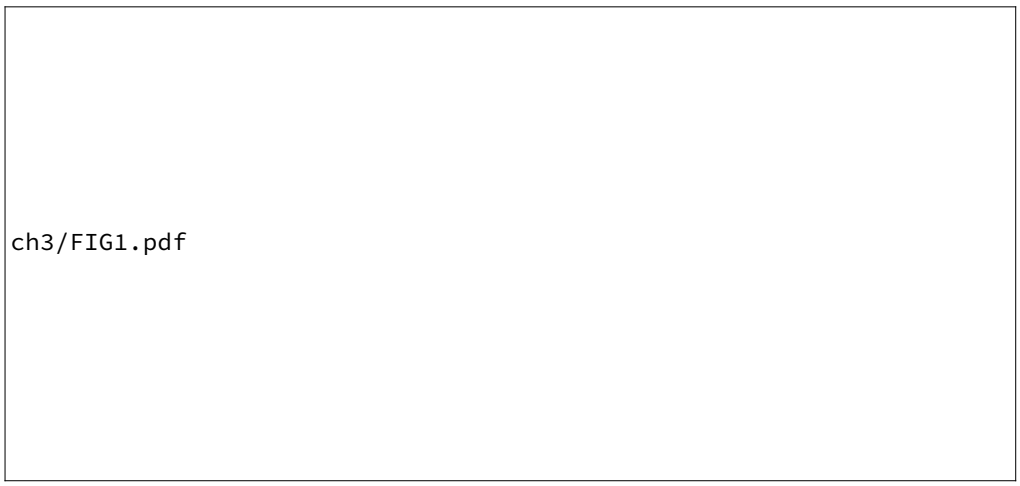

Figure 4.4: **Main features of NuST.** (A) One-dimensional aggregation of gene set. Starting from a list of genes, their density is evaluated using a sliding window of different size and compared to a shuffling null model to extract the significant clusters. (B) Diagram of the statistically significant linear aggregation clusters The box indicates the position of the peak at a given scale of analysis (x axis) while the whisker the maximal extension of the cluster, with color-coded P-values. The right panel reports the positions of chromosomal macrodomains [37, 43], chromosomal sectors defined in [47], and the names of some important genes. (C) Circular representation of the most significant clusters found for the same data set of Figure 2. The outer colored ring represents macrodomains [37, 43] while the inner ring represents chromosomal sectors defined by Mathelier and Carbone [47]. The table reports the cluster ID, the scale of observation, the cluster coordinates in megabases, and the P-value associated to each cluster. (D) Overlaid histogram of two data sets in the common database. The input files belonging to the common data sets can be read in the upper part of the plot. The normalized number of genes located inside a window centered in each chromosomal position is shown. In this example the window size is fixed to $L/16$, where L is the total length of the genome. (E) Local Pearson correlation function between sliding windows histograms. The (customizable) window size is $L/32$, where L is the total length of the genome. The plot legend reports the total Pearson correlation coefficient. In this case, while globally there is no correlation, specific regions can present co-occurrence or mutual exclusion.

nificant clusters for the different observation scales, with a color code representing their P-value. The x axis identifies the given scale (identified by the bin size) and the y axis draws the clusters as boxes, with color coded P-values, comparing them to macrodomains [37, 230] and chromosomal sectors defined in ref. [47]. An alternative graphical representation shows the statistically significant clusters as colored wedges whose transparency increases with size (Figure 4.4C). Finally, the table reports the main properties of each cluster.

Sliding window histograms (Figure 4.4D) allow the comparison of the gene density of different data sets, at a user-defined window size (step size 500 bp). Finally, a tool evaluates the local contribution to the Pearson correlation coefficient along the chromosome between the gene densities of two input lists (Figure 4.4E), by computing the quantity $(x_i - \langle x \rangle)(y_i - \langle y \rangle)/(\sigma_x \sigma_y)$ for each window position i, where x_i and y_i are the signals (gene densities) of the x and y data sets, within the window at position i, $\langle \cdot \rangle$ stands for average and σ for standard deviation. Since averages and standard deviations are calculated across the whole genome, the global Pearson correlation coefficient is a number between -1 (linear anticorrelation) and $+1$ (linear correlation). The local product does not have this constraint, but the number still represents a measure of positive or neg-

List of "Common" datasets in NuST
Genes linked to specific phenotypes (essential genes and genes responsive to external conditions)
Conditionally essential genes due to auxotrophy [224]
Conditionally essential genes in rich media [224]
Genes whose expression is affected by growth conditions [223]
Genes whose expression is affected by the carbon source [223]
List of *E. coli* essential genes [225]
List of *E. coli* essential genes [226]
Multi-stress responsive genes [224]
Genes next to binding sites of a regulator from binding data (ChIP-chip, ChIP-seq, etc)
Genes overlapping with heEPODs (highly expressed extended protein occupancy domains) [68]
Genes overlapping with tsEPODs (transcriptionally silent extended protein occupancy domains) [68]
Putative CRP target genes from ChIP-chip experiments [19]
Putative FIS target genes from ChIP-chip experiments [203]
Putative FIS target genes from ChIP-seq experiments [67]
Putative FIS target genes in early-exponential phase from ChIP-seq experiments [67]
Putative FIS target genes in mid-exponential phase from ChIP-seq experiments [67]
Putative FIS target genes in mid-log phase from ChIP-chip experiments [66]
Putative FNR target genes in midlog phase from ChIP-chip experiments [74]
Putative FNR target genes in midlog phase in the presence of oxygen from ChIP-chip experiments [74]
Putative FNR target genes in stationary phase from ChIP-chip experiments [74]
Putative H-NS target genes from ChIP-chip experiments [72]
Putative H-NS target genes from ChIP-seq experiments [67]
Putative H-NS target genes in early-exponential phase from ChIP-seq experiments [67]
Putative H-NS target genes in mid-exponential phase from ChIP-seq experiments [67]
Putative H-NS target genes in mid-log phase from ChIP-chip experiments [66]
Putative H-NS target genes in stationary phase from ChIP-seq experiments [67]
Putative H-NS target genes in transition-to-stationary phase from ChIP-seq experiments [67]
Putative IHF target genes in mid-log phase from ChIP-chip experiments [66]
Putative RNA-polymerase target genes during rapid growth from ChIP-chip experiments [19]
Genes sensitive to nucleoid perturbations from transcriptomics experiments
FIS knockout sensitive genes in all stages of growth [76]
FIS knockout sensitive genes in the early-exponential growth phase [76]
FIS knockout sensitive genes in the late-exponential growth phase [76]
FIS knockout sensitive genes in the mid-exponential growth phase [76]
FIS knockout sensitive genes in the stationary phase [76]
FIS-knockout sensitive genes under negative-supercoiling [75]
FIS-knockout sensitive genes under positive-supercoiling [75]
H-NS knockout sensitive genes under negative-supercoiling [75]
H-NS knockout sensitive genes under positive-supercoiling [75]
Supercoiling sensitive genes [75]
Supercoiling sensitive genes under FIS knockout [75]
Supercoiling sensitive genes under H-NS knockout [75]
Genes with specific annotations (according to external databases)
Genes related to small RNAs [223]
Genes whose protein products have transmembrane domains [227]
Genes with paralogs [227]
Putative horizontally transferred genes [228]
Mutated genes in driven evolution experiments
List of mutated genes in *E. coli* long-term evolution experiment through 20,000 generations [229]
List of mutated genes in *E. coli* long-term evolution experiment through 40,000 generations [229]
Target genes of a regulator (according to external databases)
Genes regulated by small RNAs [223]
Genes regulated by the sigma factor 24 [223]
Genes regulated by the sigma factor 28 [223]
Genes regulated by the sigma factor 32 [223]
Genes regulated by the sigma factor 38 [223]
Genes regulated by the sigma factor 54 [223]
Genes regulated by the sigma factor 70 [223]

Table 4.1: Sources: [224] Nichols *et al* 2011, [223] Regulon DB, [225] Baba *et al* 2006, [226] Gerdes *et al* 2003, [68] Vora *et al* 2009, [19] Grainger *et al* 2005, [66] Grainger *et al* 2006, [74] Grainger *et al* 2007, [203] Cho *et al* 2008, [67] Kahramanoglou *et al* 2011, [72] Oshima *et al* 2006, [76] Bradley *et al* 2007, [75] Blot *et al* 2006, [227] Ensembl bacteria, [228] HGT-DB, [229] Barrick *et al* 2009.

ative correlation. Note that the global Pearson correlation coefficient can change with observation scale, as the sliding window histograms may become different.

To the best of our knowledge the NuST web server is the first freely available computational tool performing multi-scale analysis of the linear aggregation of specific gene sets along the chromosome. It is designed for the investigation of links between bacterial chromosome organization and the global gene expression program, a problem in which such multi-scale approach and comparison/integration of different kinds of data can be highly informative.

4.3 EPISTATIC EFFECTS OF THE HOMOLOGOUS H-NS AND STPA GENES KNOCK OUT

In this section we describe a recent research work carried principally by our group in NCBS, Bangalore: the Bugbears group. The work focus on the epistatic effect of the homologous H-NS and StpA genes knock out in *E. coli* and it is divided in two parts: in the first part the epistatic effect are measured and analyzed, in the second part we describe an experiment of laboratory evolution with the aim of finding and understanding the suppressor mutations which reverse the effects of *hns* and *stpA* deletions.

Horizontal gene transfer is a major force in bacterial evolution. It introduces drastic changes in an organism's gene content by transferring entire functional pathways. This allows the host organism to rapidly explore new niches and phenotypes, including pathogenesis. The effects of a horizontal gene transfer event extend beyond mere addition of new genes; it introduces new selective forces that might lead to a chain reaction of further modifications to the core or the conserved host genome. These might include, for example, changes in the metabolic or the regulatory circuitry of the cell, [231–233] deletions of core genes, [234] and fundamental changes in genomic base composition as imposed by acquired restriction-modification systems. [235, 236] Further, there might be costs associated with expressing such acquired genes. [237] Therefore, it is not surprising that horizontally acquired genes are carefully regulated, not least at the level of gene expression. [238, 239]

Gene expression in versatile bacteria such as the model organism *Escherichia coli* is regulated at the transcriptional level by a complex network of interactions mediated by sigma factors, which are components of the initiating RNA polymerase holoenzyme, and transcription factors. Transcription factors may be global or local in scope depending on various parameters including the numbers of genes that they regulate, their propensity to be involved in combinatorial control, and the numbers of environmental conditions in which they are active. [240] One of the global regulators of transcription in *E. coli* is the protein called H-NS, whose regulatory scope extends to 20% of the genes encoded by the genome. [66, 67, 72] H-NS is a small protein of 137 amino acids, which recognizes DNA that is A+T-rich in sequence [67, 241, 242] and/or specific DNA geometries such as intrinsic bending [243–245] and Holliday junctions. [246] It is also involved in transcription termination, [247] and can bind and stabilize mRNA. [248] As a transcriptional regulator, H-NS silences gene expression. It oligomerises on the DNA and forms DNA–protein–DNA bridges or rod-like structures, which might be structural motifs for strong transcriptional suppression. [60, 65, 249] Recent evidence has suggested that the 500 binding regions of H-NS collapse together spatially into two foci, [91] which might act as the bacterial heterochromatin. In particular, H-NS is a silencer of horizontally acquired genes, [72, 242, 250] which in enterobacteria display a tendency to be A+T-rich. [67, 251] Despite being a regulator of horizontally acquired genes, H-NS itself is well conserved in enterobacteria. Functional homologs of H-NS have been detected even in distant bacterial genera such as *Bacillus* and *Mycobacteria*. [72] Taken together,

H-NS is a powerful transcriptional repressor and a prominent link between the conserved, or the core genome of an organism and the accessory, or the recently acquired genome.

H-NS does not act alone. Its activity is enhanced or diminished by other proteins including StpA, Hha and YdgT (Cnu). [252] (a) StpA is a paralogue of H-NS, the two proteins sharing nearly 70% sequence similarity. [253, 254] A recent in vitro study has shown that StpA could also form DNA filaments similar to those formed by H-NS and thus silence gene expression. [63] Recent super-resolution microscopy showed different patterns of localization for H-NS and StpA. [91] This is in curious contrast to ChIP-chip studies, which demonstrated that the genome-wide binding profile of StpA reflects that of H-NS. [255] H-NS and StpA repress each other transcriptionally. [253] The absence of H-NS not only abolishes StpA binding to two-thirds of its sites, [255] but also makes StpA subject to rapid degradation by the Lon protease. [256] Thus, the expression level of StpA – which contributes to its activity – is controlled by a balance between its synthesis and degradation. Prior studies, encompassing different strains of *E. coli* have shown that the Δ*stpA-hns* double mutant has a severe growth defect. [257, 258] However, a mutation in *hns* has only a mild growth defect, except under certain conditions in stationary phase. [259] But a Δ*stpA* deletion does not appear to adversely affect growth as measured by optical density. (b) Hha is a small protein that can functionally replace the N-terminal oligomerization domain of H-NS. [260] It also has a role in H-NS distinguishing between horizontally-acquired and core regions of the genome. [261] (c) YdgT is homologous to Hha and might share some of its functions. [262]

Taken together, the global gene regulatory system around H-NS appears to be a multi-layered silencing system that operates in regions of the genome with high A+T content, many of which are likely products of recent horizontal gene transfer. What is (are) the selective force(s) associated with the ability of H-NS to silence gene expression either on its own or in collaboration with other proteins? In sections 4.3.1 and 4.3.2 we present two experimental studies that attempts to answer this question using *E. coli* as a model.

4.3.1 Genomic analysis reveals epistatic silencing of "expensive" genes in *Escherichia coli* K-12

Results:

Growth characteristics

We first assessed the contributions of co-regulators acting around H-NS – namely, StpA, Hha and YdgT – to the growth of *E. coli*. We made single deletions of *hns*, *stpA*, *hha* and *ydgT* (Δ*hns*, Δ*stpA*, Δ*hha* and Δ*ydgT* respectively) in *E. coli* K12 MG1655. In addition, we made double deletions of *hns* with each of the other three co-regulators (Δ*stpA-hns*, Δ*hha-hns* and Δ*ydgT-hns*). Among the four single mutants, only Δ*hns* has a slight growth defect in LB medium (Fig. 4.5a). Among the double mutants (Fig. 4.5b), the Δ*stpA-hns* mutant appears to be deficient in growth, an observa-

tion not explained by the growth curves of Δ*hns* and Δ*stpA* taken in isolation. This strain shows a somewhat pronounced lag phase and a considerably smaller maximal growth rate than any of the other mutants. This is in agreement with earlier reports investigating the Δ*stpA-hns* double mutant. [257, 258] But, the other double mutants – Δ*hha-hns* and Δ*ydgT-hns* – grow similarly to Δ*hns*.



Figure 4.5: (a) Growth curves of the wildtype and the various single mutant strains constructed in this study; (b) growth curves of the wildtype, Δ*hns*, and the various double mutant strains involving *hns*. The curves track the median across 6 replicates (2 biological and 3 technical), with the error bars marking the standard error. (c) Heatmap showing gene expression profiles of a selected set of genes in the various strains subjected to RNA-seq experiments. The cladogram was drawn using treeview and the heatmap using the matrix2png web server. Each column represents a strain, with the number in parenthesis identifying the two replicates. The wildtype, Δ*hns*, Δ*stpA*, and Δ*stpA-hns* strains are marked in bold text. Each row represents a gene. The colour in each cell shows the gene expression level of the corresponding gene in a strain. Red corresponds to high expression and white to low expression, with the range in-between shown by the intensity of the red colour. The region in the top, marked by the thick-bordered rectangle, shows genes which are specifically up-regulated in Δ*stpA-hns* when compared to Δ*hns*. The expression measure was computed as follows. The number of reads mapping to a gene was divided by the length of the gene. Then, this value was divided by the mode of the distribution of expression values across all genes in our dataset. This is a robust method of normalizing for the variation in the total number of reads obtained for different samples. The expression measure is represented on a $\log_2$ scale.

"Epistatic" control of gene expression by H-NS and StpA

We studied the control of gene expression by H-NS, either alone or in combination with one of StpA, Hha and YdgT. For each of the strains used here (wildtype, Δ*hns*, Δ*stpA*, Δ*hha*, Δ*ydgT*,

Δ*stpA-hns*, Δ*hha-hns* and Δ*ydgT-hns*), we performed transcriptome experiments – of cells grown to the mid-exponential phase in rich LB medium – using RNA-seq.

We made a gene expression matrix in which each mRNA-encoding gene was represented by a vector of processed read counts (see Methods). Hierarchical clustering of this matrix using uncentered correlation revealed the presence of two major clusters of samples (Fig. 4.5c). The first cluster contained the Δ*hns* single mutant and all the three double mutants. The second cluster comprised the wildtype and the single mutants of the three co-regulators. Thus, among the single mutants, Δ*hns* has the strongest effect on gene expression under the condition tested here, with most ($> 75\%$) responding genes being up-regulated in the mutant. Δ*stpA*, Δ*hha* and Δ*ydgT* have little to no effect. Among the double mutants, Δ*hha-hns* and Δ*ydgT-hns* are similar to Δ*hns*. However, Δ*stpA-hns* – forming a separate branch within the first cluster – has many genes that are upregulated relative to Δ*hns* as well as the wildtype, despite Δ*stpA* being similar to the wildtype. Based on the above results, we restricted further analysis to H-NS and the homologous StpA. The deletion of Δ*stpA* in a Δ*hns* background has a global impact on gene expression, whereas it does not in the wildtype. This suggests that certain genes are regulated by both the proteins in an epistatic manner. Herein, the word epistasis means that the transcriptional effect of Δ*stpA-hns* is not explained by the sum of the effects of the component single deletions. We classified genes into (a) those that are regulated epistatically by H-NS and StpA (for convenience referred to as "epistatic genes"); (b) those that are regulated unilaterally by H-NS ("unilateral genes"); (c) control genes, which are not regulated by H-NS or StpA, on the basis of our transcriptome data. Epistatically regulated genes were defined as those that are up-regulated in Δ*stpA-hns* relative to both the wildtype and Δ*hns*. About half of these epistatic genes are first up-regulated in Δ*hns* relative to the wildtype, and further up-regulated in Δ*stpA-hns*; the remaining do not respond to Δ*hns* but are up-regulated only in Δ*stpA-hns*. Unilateral genes are those which are up-regulated in Δ*hns* relative to the wildtype, but show no response to a further deletion of *stpA*. In total, we assembled a list of 360 epistatic and 610 unilateral genes (Fig. 4.6).

H-NS and StpA binding to epistatically and unilaterally controlled genes

Using data from a previous ChIP-seq study of H-NS [67], we find that much of the transcriptional up-regulation (in mutant compared with the wildtype) observed in our study can be explained by the direct effect of H-NS binding proximally to the gene ($> 70\%$ of up-regulated genes are bound by H-NS). There is little difference between unilateral and epistatic genes in this respect. Kahramanoglou *et al.* [67] reported that the length of an H-NS binding region on the chromosome is correlated with the degree of repression of the target gene. We observe that epistatic genes are bound by longer H-NS binding tracts than unilateral genes (Fig. 4.7a; $P = 10^{-3}$; Wilcoxon test). This, tentatively, may be consistent with previously reported *in vitro* data that the heteromeric association of H-NS and StpA is more stable than the self-association of either protein. [263]

ChIP-chip experiments for StpA in wildtype and Δ*hns* E. coli K1232 found that the binding

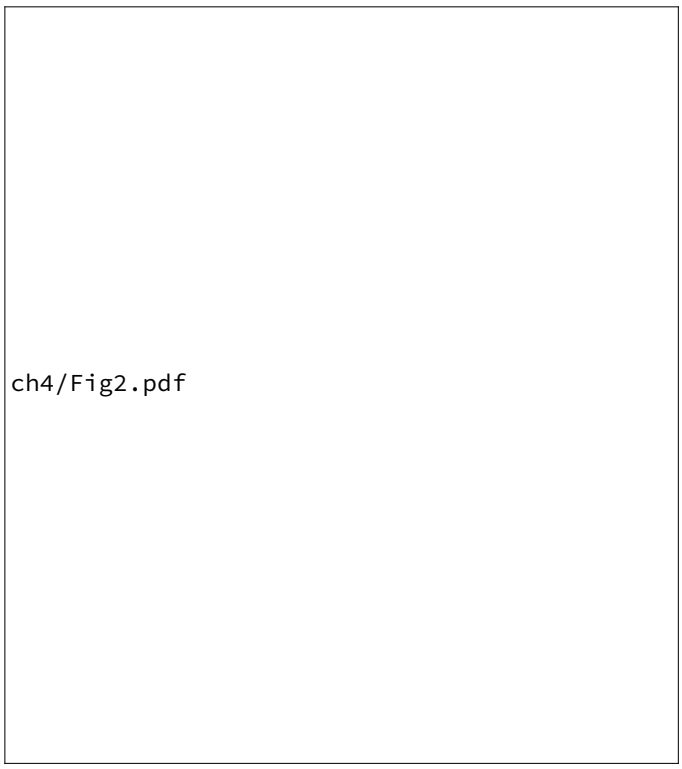

Figure 4.6: The pie chart shows the number of genes belonging to each class of genes studied here. The graphs show the inter-quartile range of the distribution of fold changes on a $\log_2$ scale (relative to the wildtype, as calculated by the program EdgeR), across various sets of genes, in Δ*hns* and Δ*stpA-hns* (red: epistatic genes; blue: unilateral genes; gray: control genes).



Figure 4.7: (a) Distributions of the length of H-NS binding tracts for epistatic (right) and unilateral (left) genes. These numbers were obtained from ChIP-seq data published by Kahramanoglou and colleagues [67]. (b) The number of epistatic (right) and unilateral (left) genes which are bound by StpA in the wildtype and in Δ*hns*. These data were obtained from ChIP-chip experiments reported by Uyar *et al.* [255] (c) Distribution of the density of high affinity H-NS binding oligonucleotides [14] for epistatic and unilateral genes. In all the boxplots used in this paper, the whiskers are drawn at 1.5× IQR above and below the third and the first quartile respectively.

profile of StpA is virtually identical to that of H-NS in the wildtype. However in Δ*hns*, StpA binding is abolished from two-thirds of its wildtype binding sites. By integrating these results with our transcriptome data, we show that epistatic genes tend to be bound by StpA irrespective of whether H-NS is present or not (Fig. 4.7b). However, for the unilateral genes, StpA binding is lost in the absence of H-NS ($P = 10^{-10}$; Fisher's exact test).

What properties make the StpA target a specific subset of its binding sites in the absence of H-NS? Towards answering this, we assembled a set of DNA sequence motifs with high affinity for H-NS, based on publicly available *in vitro* affinity data for H-NS to 8-mer oligonucleotides. [241] For each epistatic and unilateral gene, we calculated the density of these high affinity binding motifs within the corresponding ChIP-seq-based H-NS binding region. We find that epistatic genes have a higher density of high-affinity H-NS binding motifs than unilateral genes (Fig. 4.7c; $P = 10^{-}10$; Wilcoxon test). This suggests that StpA, in the absence of H-NS, binds only to those target sites to which it has very high binding affinities.

To summarize, epistatically silenced genes tend to lie at sites where (a) H-NS binds in particularly long tracts and; (b) StpA binding is independent of H-NS availability. Thus epistatic control of gene expression by H-NS and StpA is explained by the genome-wide binding profiles of the two regulators.

Enrichment of horizontally-acquired and non-essential genes among epistatic genes

Is there a selective force that determines which genes are controlled unilaterally and which ones are regulated epistatically? Towards answering this question, we introduced various published genome-scale experimental data and computational predictions, and analyzed our transcriptome data further.

H-NS is a well-known silencer of horizontally-acquired genes, a finding that is reflected in our data as well. We predicted 600 horizontally-acquired genes using the program Alien Hunter, [264] which defines regions of abnormal higher-order oligonucleotide usage as horizontally-acquired. For a nearly 50% A+T genome like that of E. coli, one would expect these predictions to pick genes with very high, or very low A+T contents. However, the data show that most horizontally-acquired genes thus predicted tend to be A+T-rich (see for example, ref. [67]), i.e. are asymmetric around the median A+T-content for all genes across the genome. Using this dataset, we find that a significantly higher proportion (30%) of epistatic genes are horizontally-acquired when compared to unilateral genes (21%; $P = 9.9 \cdot 10^{-4}$; Fisher's exact test).

Previous studies have made large-scale deletions of tracts of genes in the *E. coli* genome. These differ from single-gene deletion libraries such as the KEIO collection in the sense that hundreds of genes had been deleted in a single strain. Of the three such studies that we investigated, two had made deletions (removing 7-–15% of the genes encoded) that did not affect the growth of *E. coli*, [265,266] whereas the third had resulted in cells with growth defects and aberrant morphology and nucleoid structure. [267] Thus, the groups of genes identified in each of the first two studies are

non-essential to the growth of *E. coli* under standard laboratory conditions. The first study (Posfai dataset) used a comparative genomic survey to identify tracts of the genome to be deleted, [266] whereas the second study (Yu dataset) used a random gene disruption strategy. [265] The chromosomal deletions in the Posfai dataset covered 750 genes, whereas those in the Yu dataset covered 470 genes in total, with a maximum of 285 deleted in a single strain (cumulative Yu dataset). By overlaying these gene lists on our data, we find that epistatic genes are statistically enriched for nonessential genes, when compared to unilateral genes (Posfai dataset: 30% of epistatic *vs.* 20% of unilateral genes, $P = 7.2 = 10^{-4}$, Fisher's exact test; Yu dataset: 26% of epistatic *vs.* 16% of unilateral genes, $P = 2.8 = 10^{-4}$; cumulative Yu dataset: 18% of epistatic *vs.* 11% of unilateral genes, $P = 10^{-3}$; Fisher's exact test). We note that the list of horizontally acquired genes, identified by the Alien Hunter program, overlaps with those presented by Posfai and Yu. However, the statistical enrichments stated above are independent of these overlaps. Finally, among the genes unique to the third study referred here, [267] where the deletions produced *E. coli* with growth defects, there is little enrichment for unilateral or epistatic genes (9% of each of unilateral, epistatic and control genes).

From silenced to de-repressed expression of epistatically and unilaterally controlled genes

Next, we used our transcriptome data to obtain the expression levels of epistatic and unilateral genes in the fully controlled wildtype cells. The expression levels of both sets of genes are low, as expected from the role of H-NS as a transcriptional silencer. Nevertheless, genes under epistatic control are expressed at significantly lower levels than those under unilateral regulation (Fig. 4.8a; $P = 10^{-10}$, for all relevant comparisons, Wilcoxon test). In a previous study, Vora and colleagues had used microarrays to define genes which fall within transcriptionally silent protein-bound regions of the chromosome. [68] In agreement with our transcriptome, epistatic genes are more likely to fall in such regions than unilateral genes, which in turn show a much higher tendency to be silent than control genes (27% for epistatic, compared to 17% for unilateral; $P = 1.1 \cdot 10^{-4}$, Fisher's exact test; 3% for control). These are consistent with the previously stated observation that epistatic genes are bound by longer H-NS binding tracts than unilateral genes.

Next, we compared the expression levels of epistatic and unilateral genes in the de-repressed state represented by Δ*stpA-hns*. We find a trend that is opposite to that in the wildtype cells. Both epistatic and unilateral genes are expressed at very high levels compared to the negative reference (Fig. 4.8b; $P = 10^{-10}$, for both comparisons, Wilcoxon test). In particular, epistatic genes are expressed at significantly higher levels than unilateral genes ($P = 2.1 \cdot 10^{-7}$, Wilcoxon test). Consequently, the fold change in expression between the de-repressed double mutant and the fully controlled wildtype is higher for epistatic than unilateral genes.

In summary, epistatic genes display properties that are at the extreme end of H-NS regulated genes, in terms of (a) high binding affinity to H-NS; (b) tendency to have been recently acquired and/or non-essential; (c) the degree of silencing imposed by H-NS (and StpA); (d) high expression



Figure 4.8: (a) Distributions of expression levels, calculated as described in the legend to Fig. 4.5, for epistatic (right), unilateral (middle) and control (left) genes in the wildtype transcriptome data. (b) As in (a), except that these data are from the transcriptome of Δ*stpA-hns*. The y-axis is on the same scale as in (a). (c) Distributions of gene expression levels for genes ranked in the top 25% (labeled "high", right), in terms of the density of high-affinity H-NS binding sites, and those in the bottom 25% (labeled "low", left). (d) Distributions of intrinsic gene expression cost for epistatic (right), unilateral (middle) and control (left) genes.

levels in the absence of gene silencing.

De-repressed gene expression and H-NS target site selection

Does the expression level select DNA sites bound by H-NS? To address this question, we tested for association between the de-repressed expression level of a H-NS-regulated gene (both unilateral and epistatic) and the density of high affinity H-NS binding sites. We find that genes with the top 25% of H-NS binding affinities have a significantly higher de-repressed expression levels than those in the bottom 25% ($\boldsymbol{P} = 2.1 \cdot 10^{-7}$, Wilcoxon test; Fig. 4.8c). In fact, there is a weak, but significant, correlation between the de-repressed expression level of a gene and the density of high affinity H-NS binding sites targeting them (Spearman rank correlation coefficient = 0.23; $\boldsymbol{P} = 4.7 \cdot 10^{-8}$). This is not entirely explained by the separation of H-NS-regulated genes into unilateral and epistatic, which differ from each other in their binding affinities to H-NS as well as their de-repressed expression levels. Even within the set of unilateral genes, the above correlation is maintained (Spearman rank correlation coefficient = 0.21; $\boldsymbol{P} = 8.3 \cdot 10^{-4}$). This is not true for epistatic genes for which the binding affinities to H-NS and the expression levels are already so high that the trend might be saturated. Thus we speculate that selection might have favored H-NS binding affinities to be more directed towards genes that are highly expressed in the absence of a silencing system.

A+T-content and gene expression cost: a hypothesis

We investigated whether de-repressed expression of epistatic genes would be expensive to the cell in terms of the metabolic burden it imposes. In the present context, we are interested in the cost of transcription, as our study does not access protein levels. In general, gene expression cost is measured by the cost of breaking activated P bonds during both transcription and translation, in addition to the metabolic processes that synthesize the nucleotide and amino acid building blocks. [268] Here we suggest that transcription of an A-rich gene at high levels may impose a higher metabolic cost than that of a G-rich gene. This is based on the following arguments: (a) the incorporation of an "A" in the mRNA abstracts ATP – both phosphates and the adenosine – from metabolic reactions. An inspection of 2000 enzymatic reactions listed in the Ecocyc database [214] shows that 15% of these utilize ATP. Though the synthesis of GTP from GMP requires the conversion of ATP to ADP, the adenosine itself is conserved. The loss of a GTP itself by incorporation in an RNA molecule is less expensive as it is utilized in only 1% of enzymatic reactions in Ecocyc. (b) A recent study by Raghavan and colleagues showed that high-level expression of GFP has a greater adverse effect on fitness, when the gene is A+T-rich. [269] This is despite the fact that GFP itself is not toxic to the cell, and that the expression level of GFP is not affected by the A+T-content of the gene encoding it. This work noted however that this effect was dependent on translation. Nevertheless, some studies have shown that translation has a positive effect on mRNA stability (reviewed by ref. [270]). Therefore, in the absence of translation, the mRNA could be degraded rapidly, allowing recycling of nucleotides. The Raghavan study in fact reports that high-A+T mRNA has smaller half-life than high G+C-mRNA, consistent with the sequence preferences of RNAse E. [271]

Based on (a) above, we costed each copy of a gene as the number of "A" multiplied by 0.15 plus the number of "G" multiplied by 0.01. This is speculatively representative of the intrinsic cost of transcribing a gene, based only on its sequence characteristics. As expected from the high A+T content of H-NS-bound genes, both unilateral and epistatic genes have higher intrinsic cost than control genes, with epistatic genes only slightly costlier than unilateral genes (Fig. 4.8d; $P = 10^{-10}$ for epistatic and $4.2 \cdot 10^{-7}$ for unilateral genes, compared to the negative control; $P = 1.6 \cdot 10^{-3}$, comparing epistatic with unilateral genes). The slightly higher intrinsic cost for epistatic genes may be because of the tendency of these genes to be a bit longer than unilateral genes (mean gene length of 1072 nt for epistatic and 975 nt for unilateral genes), though there is little difference between the two sets of genes in their A+T-contents (mean A+T-content of 52.2% and 52.4% across the ORF for epistatic and unilateral genes respectively). The cost of expressing a gene also depends on the level to which it is expressed. Higher gene expression implies higher cost. Thus, under any given condition, the product of the intrinsic cost of a gene and its expression level can be thought to represent its transcriptional cost. The fact that epistatic genes transition from being silent in the wildtype to being extremely highly expressed in Δ*stpA-hns* mean that their transcriptional cost will be very high in the double mutant. This will be more muted for unilateral genes. Given that epistatic and unilateral genes differ only slightly in their intrinsic cost, any difference in the total cost

is likely because of the higher de-repressed expression of epistatic genes. Whether this difference between the two sets of H-NS-regulated genes makes any contribution to the growth defect of the double mutant is difficult to comment on.

Summary of results

Our study suggests that gene silencing by H-NS in *E. coli* is directed more towards genes that are expressed at extremely high levels in the absence of this repressive system (Fig. 4.9). This operates at two, partially overlapping, levels. Highly transcribable (i.e. in the absence of gene silencing) genes are first enriched for high-affinity H-NS binding sites. Moreover, these genes also recruit StpA as a molecular backup for H-NS, thus building a global epistatic gene silencing network.

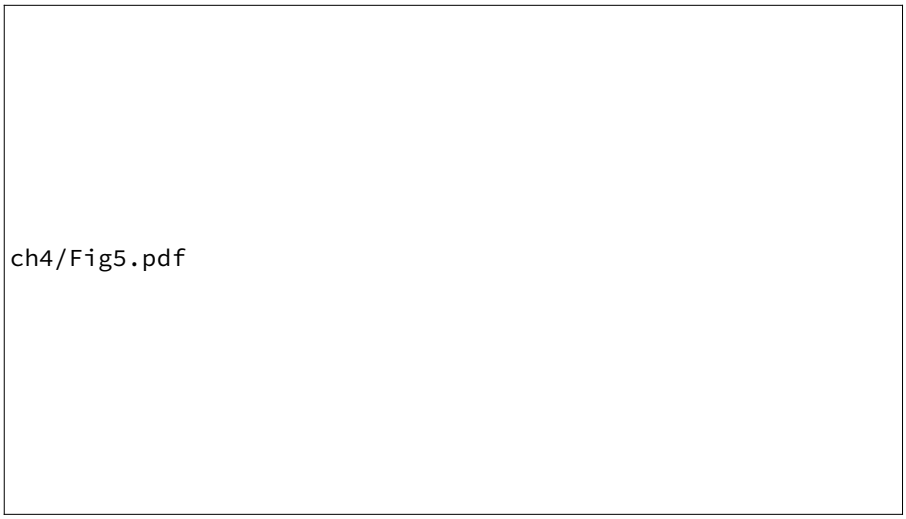

Figure 4.9: A graphical summary of our results, comparing epistatic (red box) with unilateral (blue box) genes. The DNA sequence motif for H-NS is from Kahramanoglou et al.'s ChIP-seq study. [67] The structure of H-NS (within the green hexagon) is from the PDB ID 1HNS; this is also used to represent StpA (within the magenta hexagon) for which we could not find any structure.

Discussion

Gene expression cost and targeted gene silencing

While horizontal gene transfer is a major force in bacterial evolution, there are many barriers to it. First, it interferes with a well-evolved cellular machinery. This interference might be because of the biochemical activities of the horizontally-acquired gene products. Or it might stem from high costs of gene expression. [237] Thus, the establishment of a horizontally acquired gene represents a balance between fitness benefit or powerful parasitism (e.g. toxin–antitoxin systems) and the costs it imposes.

The genome of *E. coli* K12 MG1655 has 600 horizontally-acquired genes, predicted based on their oligonucleotide content. [264] Most of these have higher A+T-content than other genes.

The gene silencing system around H-NS is directed towards repressing many highly transcribable genes, which have high A+T content. Here, the word transcribable is used to refer to the expression of a gene in the de-repressed Δ*stpA-hns*. The selection for silencing might operate at multiple levels. First, highly transcribable genes contain a higher density of high affinity H-NS binding motifs. This might imply that H-NS binding specificity has evolved to target highly transcribable DNA. A previous ChIP-chip study of H-NS has shown that higher expression of a gene might attract H-NS to it, at least for selected promoters. [66] This might be a conserved feature in many bacterial genomes, based on previous studies reporting highly similar sequence specificity for the silencing system in Mycobacteria, [241] which has only a low sequence similarity with H-NS. Second, StpA is recruited to highly transcribable genes in the absence of H-NS, thus presenting a second, backup layer of transcriptional repression. This selectivity of StpA to a subset of all its possible targets is likely to emerge from its relatively low expression levels. [254] Additionally, the formation of filaments of strong H-NS–StpA heterodimers at loci with high densities of high affinity H-NS binding motifs may also impose powerful transcriptional silencing. We also note here that certain gene functions might also select for epistatic gene silencing. For example, many genes involved in the biosynthesis of the O-antigen and capsular polysaccharides are epistatically regulated, as are several genes also under FNR control. We note that gene silencing can also be influenced by additional and unrelated players, such as the transcription terminator Rho, [77, 272] and a study of their functional and/or physical interactions with H-NS is likely to emerge as an important area for future work.

Assuming that highly transcribable horizontally-acquired genes do provide their host with a selective advantage under certain conditions, their establishment in the host genome is facilitated by their silencing by H-NS (and StpA). An important question that emerges is the following. Is there a selective need for *E. coli* to maintain these sequences in a highly transcribable state? If not, one might surmise that mutations that decrease their transcription could have occurred. There might be several answers to this. One is that the acquisition of these genes occurred so recently that they have not had enough time to mutate sufficiently and get fixed in the population. Another possibility is that their expression is determined by the local geometry or topology of the DNA, which makes them highly permissible to transcription; this is in contrast to a few bases in the promoter region that permit high recruitment of RNA polymerase. Such a situation is likely to be more difficult to correct with point mutations. Alternatively, transcriptional silencing by H-NS might have ensured that there is no selection against their current sequence composition. A final hypothesis is that such a sequence composition is essential for these sequences to be functional. For example, it is believed that these sequences have important functional roles under certain stress conditions, including the stationary phase. [259, 273] Under these circumstances, overall transcription is low, at least in part due to less negative supercoiling. Here, transcription may be favored from highly transcribable DNA; in fact, a microarray study has shown that many genes that are up-regulated in response to a loss of negative supercoiling are A+T-rich. [274] Finally, the larger chromosomal

context might play a role in gene expression patterns. Zarei, Scolari and colleagues [49, 77] have observed that there is a statistical enrichment for H-NS regulated genes to be around the terminus of replication. Recent studies by Sobetzko et al. [222, 275] observed that during exponential phase of growth, genes around the origin are more transcribed than those around the terminus. This is consistent with an inferred decrease in superhelical density from the origin to the terminus. However, under the stationary phase, the trend is reversed with higher expression observed around the terminus.

We adopted the methods of Zarei, Scolari and colleagues(Section 4.2 and references [49, 77, 276]) to compare the genomic coordinates of epistatic and unilateral genes in the context of chromosomal macrodomains, within which intrachromosomal recombination events occur preferentially. [37] Epistatic genes are preferentially located at the boundaries of these macrodomains – especially between the TER and the LEFT macrodomains and an edge of the RIGHT macrodomain (Fig. 4.10). On the other hand, many unilateral genes are spread around the Ter macrodomain. The exact roles of global regulatory proteins including H-NS in chromosomal location and topology dependent gene expression programs remain to be understood in detail, although a recent study has studied the impact of local supercoiling on H-NS control of transcription from a specific bacterial promoter. [277]

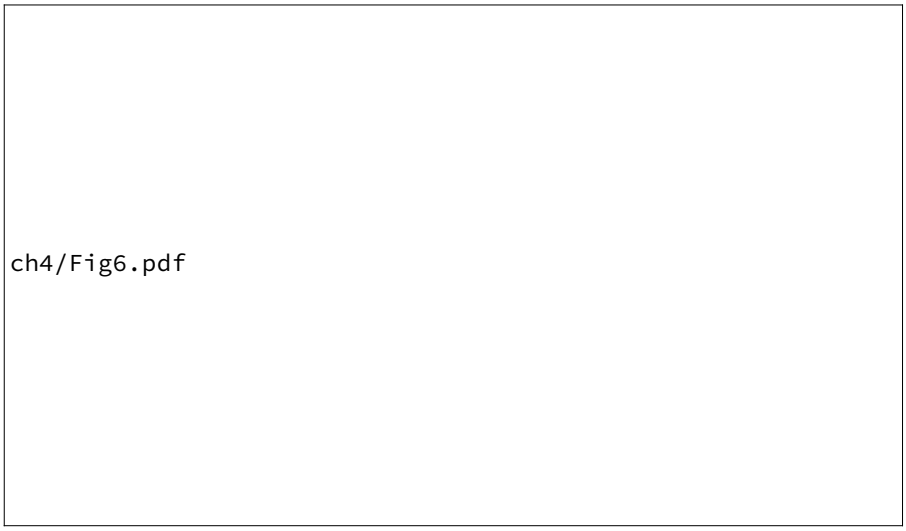

Figure 4.10: The figure shows statistically significant clusters of epistatic and unilateral genes in relation to their chromosomal positions. These figures were drawn using the NuST web server (see Sec 4.2). The clusters are represented by the red coloured wedges, with the intensity of the colour increasing with the size of the cluster. The outer circle shows the macrodomains described by Valens and colleagues [37] (green: ORI macrodomain; red: RIGHT macrodomain; cyan: TER macrodomain; blue: LEFT macrodomain; white: non-structured regions). The inner circle shows the various chromosomal sectors, computationally defined as regions showing a positive correlation between transcriptional level and a measure of codon adaptation index [47]. Epistatic genes are clustered at certain macrodomain boundaries, primarily those between the TER and the LEFT macrodomains, and between the RIGHT and its proximal non-structured region. Unilateral genes are spread through the body of the TER macrodomain.

An epistatic relationship between hns and stpA

Various studies, together straddling several strains of E. coli, have shown that a double mutant lacking both StpA and H-NS displays a growth and viability defect. [257, 258] However, this effect cannot be predicted from the growth characteristics of the two single mutants alone. Though this appears to suggest a form of molecular backup between two homologous proteins, the backup is unequal: whereas the Δ*stpA* mutant hardly affects gene expression, a Δ*hns* mutant leads to up-regulation of at least 500 genes as agreed upon by several studies. [67, 72, 242, 257] Whereas the absence of *stpA* does not visibly affect in vivo DNA binding of H-NS on a genomic scale, that of hns abolishes binding of StpA to about two-thirds of its wildtype binding regions. [255] Though both H-NS and StpA repress each others transcription, leading to the up-regulation of one in the absence of the other, the degree of up-regulation conferred on StpA by the absence of H-NS may not be sufficient to compensate for the lack of H-NS. [254] This in part might be due to the higher susceptibility of StpA to degradation by Lon in the absence of H-NS. [256] A complete backup of H-NS activity by StpA might not be desirable as there are conditions, such as a late stationary phase, in which an attenuated H-NS is selected for ref. [259].

The close relative of *E. coli*, *Salmonella enterica Typhimurium*, encodes three H-NS-like proteins. In this organism, the single Δ*stpA* mutant affects the expression of 5% of genes, [278] in contrast to what we, in this paper, and others, have reported for *E. coli*. The relationship of H-NS with StpA is in contrast to that with Sfh, another homolog of H-NS studied in *Salmonella*. Sfh binds to a subset of H-NS binding sites in wildtype cells, but expands its reach in Δ*hns*. [279] Sfh is different from StpA however, in the sense that it is a plasmid encoded protein, which serves the role of a stealth agent, as opposed to StpA which is well-ensconced in the chromosome. At the other end of the scale is *Yersinia*, also an enterobacteriaceae, which does not encode a H-NS homolog, and in which Δ*hns* is lethal. [280] Therefore, the function of H-NS in concert with homologous regulators is taxa-dependent, offering scope for more detailed evolutionary analysis.

A molecular backup between two global transcriptional repressors

The use of homologous proteins (H-NS and StpA) acting together to silence expensive genes might confer robustness in the event of a mutation adversely affecting the function of either. Alternatively, during rapid exponential growth, the presence of multiple chromosomal copies might select for a second protein (read StpA) to target loci which may not be sufficiently bound by H-NS. [255, 281] Under this model, one might expect selection to favor the positioning of epistatic genes in chromosomal regions of higher ploidy during exponential growth, i.e. closer to the origin than the terminus. We tested this hypothesis, using previously generated ChIP-seq mock-IP data for mid-exponential phase E. coli, [67, 282] which by reflecting a quantitative genome sequencing experiment, identifies loci present in higher copies (Fig. 4.10). We do not find any enrichment for epistatic genes to be present near the origin, or in any other way, in higher copy numbers. In fact,

previous studies and our results presented in Fig. 4.10 report a tendency for H-NS-regulated genes to be localized around the edges of the chromosomal macrodomain comprising the terminus. [49] Therefore, the more parsimonious interpretation of our data might be in favor of a molecular backup in anticipation of a mutational event. Whether this might confer an evolutionary advantage especially in the face of large population sizes is debatable.

It has been previously suggested that there may be selection against target overlap between homologous regulators, on the basis of limited data available in the RegulonDB database. [283] However, this may not be valid for conditions and systems which are benefited by an epistatic circuit as described here.

Savageau's demand theory of gene regulation posits that the repressive mode of gene expression control is favored over activation under situations where the target gene is rarely required, [284] which may well be true of the horizontally-acquired targets of H-NS and StpA. This is because the mutational loss of a repressor is expected to be selected against when a target gene is expressed constitutively in a low-demand regime. The fact that H-NS is a global transcriptional repressor, and that its loss leads to extremely high expression of many of its targets, might lead to an amplified selection against its loss or inactivation; this could select for backup by StpA.

Examples of molecular backups involving transcriptional regulators – not necessarily repressors – have been described on a genomic scale in the yeast *Saccharomyces cerevisiae*, where a whole genome duplication event generated a paralog of each gene, many of which have been retained. [285, 286]

*The growth defect of the Δ*stpA-hns* mutant: multiple causes?*

The fact that a Δ*stpA* mutant in a Δ*hns* background has an effect on growth suggests that there are unique molecular features associated with the Δ*stpA-hns* double mutant, presumably at the transcriptional level. A study by the Uhlin group had demonstrated that the global regulator CRP is down-regulated by as much as five-fold in a Δ*stpA-hns* strain. [287] We generated a Δ*crp* mutant and observed that it shows only a small growth defect in LB medium; however a Δ*crp-hns* double mutant is very slow in growth (data not shown), suggesting a genetic interaction between the two. The above-mentioned work from the Uhlin group also showed that the double mutant might be experiencing a stringent response-like state, which is induced by the alarmone ppGpp. The down-regulation of CRP could be reversed in a mutant that was unable to synthesize ppGpp. This mutant, as well as overexpression of CRP, were reported to be partial suppressors of the growth defect seen in the Δ*stpA-hns* double mutant. The above study concluded that changes in the supercoiling state of the DNA in the Δ*stpA-hns* double mutant was the ultimate cause of these effects.

Our transcriptome data revealed a 3.5-fold down-regulation of the *crp* gene in the Δ*stpA-hns* strain. We performed an RNA-seq experiment with our Δ*crp* mutant and found that many genes that are down-regulated in Δ*stpA-hns* (when compared to Δ*hns*) are also differentially expressed in Δ*crp* (62% of genes down-regulated in Δ*stpA-hns* compared to Δ*hns* are differentially expressed in

Δ*crp*). Most (92%) of these genes are also down-regulated in the Δ*crp* mutant, emphasizing that the transcriptional effects are consistent between CRP and the H-NS–StpA pair. Half of these are not bound by H-NS, in contrast to the up-regulated genes where nearly three-fourths are bound by H-NS. This demonstrates that the down-regulation of a significant proportion of these genes is due to secondary effects, distal from H-NS and StpA. Note however that 50% of these genes are in fact bound by H-NS and it is not clear if H-NS – in some combination with CRP – can act as an activator at these loci. However, unlike in the study by the Uhlin group, we do not observe signatures of a stringent response transcriptome among the down-regulated genes, on the basis of microarray data generated by Traxler and colleagues. [288] The Uhlin study had also shown that changes in supercoiling effected by H-NS and StpA could be the primary cause of these effects. But, an overlap analysis performed using the NuST web server (see Sec. 4.2), does not show any enrichment of supercoiling-sensitive genes [75] among those down-regulated in Δ*stpA-hns*.

Thus, a component of the growth defect observed in the double mutant might be explained by the down-regulation of CRP. Though this could indirectly emerge from the effect of H-NS (and StpA) on supercoiling, which in turn affects the stringent response, [287] we are unable to see signatures of these possible causes in our transcriptome. It may however be worth noting that the strain of *E. coli* used in the Uhlin study did not encode an active RelA (ppGpp synthase), whereas our strain has both RelA and SpoT, with the former accounting for much of the transcriptional effects of a Δ*relA-spoT* double mutant. [237]

Our analysis of transcriptome data generated here suggests that de-repressed expression of metabolically expensive genes might play an additional – admittedly minor – role in causing the growth defect of the Δ*stpA-hns* double mutant. And that this might be an additive effect spread across hundreds of genes. This may be supported, albeit tenuously at this point, by a three-fold upregulation of the expression of genes encoding Tsx, a transporter of (deoxy)nucleosides, [289] and the transketolase TktB, which is involved in nucleoside biosynthesis not only by being a minor player in the production of ribose-5-phosphate, [290] but also by being part of the RNA degradosome. [291] In this context, it is also worth noting that TktA, a homolog of TktB, influences chromosome topology; [292] its role in the Δ*stpA-hns* or the Δ*hns* backgrounds is not known.

The extent to which gene expression cost might affect growth in the present context can only be speculated upon. More so since a rigorous mathematical definition of sequence-dependent gene expression cost that can be applied to functional genomic data is not available at present. However, cost of gene expression, including transcription and translation, is expressed in terms of growth of bacterial strains expressing specific non-selective genes such as GFP at high levels. Using such a definition of gene expression cost, a previous study had proposed that the cost of expressing unneeded proteins is at its highest during early stages of growth. [293] This, in the Δ*stpA-hns* context, may at best be reflected in the long lag phase and a slow early exponential growth. In fact, this may contribute to the small growth defect of the Δ*hns* single mutant as well.

Two additional gene expression-based hypotheses for the growth defect of Δ*stpA-hns*, namely

the downregulation of an essential prophage-encoded gene racR and upregulation of a locus for capsular biosynthesis, were evaluated and tentatively invalidated. These are described in Fig. 4.11.

ch4/Fig7.pdf

Figure 4.11: The above graph shows the growth curve of triple mutants Δ*cps-stpA-hns* (top) and Δ*rac-stpA-hns* (bottom) in comparison with Δ*stpA-hns* and wildtype *E. coli* MG1655. No rescue of the Δ*stpA-hns* growth phenotype was observed on deletion of the rac prophage or the capsular polysaccharide (*cps*) genes in Δ*stpA-hns* background.

The slow growth phenotype of the double mutant may also derive contributions from non-transcriptional sources. For example, H-NS and StpA may be involved in repair of double strand DNA breaks, [294] with the former shown to bind to Holliday junctions in vitro. [246] The possible occurrence of such events especially as part of homologous recombination occurring during rapid exponential growth, might also contribute to the growth defect.

Methods

Strains and general growth conditions

The *Escherichia coli* variants used in the work are following: *E. coli* K-12 MG1655 wildtype (CGSC # 6300); MG1655 Δ*hns* (Δ*hns::kan*^{r}); MG1655 Δ*stpA* (Δ*stpA::kan*^{r}); MG1655 Δ*hha* (Δ*hha::kan*^{r}); MG1655 Δ*ydgT* (Δ*ydgT::kan*^{r}); MG1655 Δ*stpA-hns* (Δ*hns-stpA::kan*^{r}); Δ*hha-hns* (Δ*hns-hha::kan*^{r}); Δ*ydgT-hns* (Δ*hns-ydgT::kan*^{r}). Luria broth was used for normal growth. 50 µg/mL of Ampicillin or Kanamycin was used as per requirement, primarily during the gene deletion process.

Construction of MG1655 knock-outs

All gene deletions were achieved by the λ Red recombination system, described by Datsenko and Wanner [295] using plasmids pKD46 and pKD4 or pKD13. Knockout strains generated by this method were selected on LB Kanamycin (50 µg/mL) plates and deletion was confirmed by PCR using specific primers. Double knockouts were also generated by the same method (Datsenko and Wanner) after removing the kanamycin cassette using pCP20.

Growth curves

Overnight grown culture was inoculated in fresh LB to 1 : 100 ratio and the growth of cells monitored by measuring the optical density at 600 nm. All these growth experiments were performed in 96 well plates, incubated at 37° in a plate reader (Tecan®, infinites F200 PRO) with constant

shaking at 87 rpm. OD_{600} was measured every $\sim$ 16 minutes. We note that this could cause issues with aeration; nevertheless, similar growth curves were observed in flasks rotating at 200 rpm in shakers.

RNA extraction and mRNA enrichment

For RNA extraction the overnight cultures were inoculated in 100 µL of fresh LB to bring the initial OD of the fresh culture to 0.03 and the flasks were incubated at 37° with shaking at 200 rpm. Two biological replicates were performed for each sample. Samples were collected at the mid-exponential phase ($OD_{600} \sim 0.5$ for Δ*stpA-hns* and ~ 0.9 for all other strains) of growth. Slight modifications of previously published protocols, [67] based on those recommended by the manufacturer for the TRIzol (Invitrogen) bacterial RNA isolation kit, were used. Briefly, cells were pelleted down at 8000 rpm for 10 minutes at 4 °C. Pellets were washed with autoclaved DEPC-treated RNAse-free water and cells were snap-frozen in liquid nitrogen and stored at -80 °C until required. Cells were homogenized with sterile pestle followed by RNA extraction using TRIzol reagent according to manufacturer's instruction until the 70% ethanol wash step. Total RNA was treated with DNAse I (Invitrogen Cat No. 18068-015) according to the manufacturer's instruction. Further precipitation of RNA and ribosomal RNA cleanup was achieved by MICROBExpress bacterial mRNA purification Kit. Total RNA was suspended in $10\mu L$ of RNAse free water (Ambions®), Cat No. AM9932). Concentration and the quality of the RNA were determined on a Nanodrop and by visualization under agarose gel respectively.

Sequencing libraries were prepared using TruSeq RNA sample preparation kit v2 (Illumina, Catalog No. RS-122-2001) according to the manufacturer's guidelines. Prepared libraries were checked for quality on an Agilent Bioanalyzer, and sequenced for 50 cycles – from one end – on an Illumina HiSeq1000 platform. Raw data have been deposited with GEO under the accession number GSE40313.

Data analysis

The 50-mer single-end sequence reads were mapped to the *E. coli* K12 MG1655 genome using BWA. [213] Gene annotations were obtained from the Ecocyc database. [214] The number of reads falling within each gene was calculated based on the base position to which the first nucleotide of the read was mapped. A matrix of read counts was generated with 16 columns, one per strain (n = 8) each in duplicate, and 4245 rows, one for each mRNA gene. This matrix was fed into the Bioconductor (`http://www.bioconductor.org`) package EdgeR [217] for analysis of differential expressions. To calculate relative expression levels of a gene within a sample, the read count for each gene was first divided by the length of the gene, and then by the mode of the distribution of expression measures thus obtained across all genes for that sample. These values are expressed on a $\log_2$ scale. The fold changes computed by EdgeR correlates well with those obtained by the above method (Pearson correlation coefficient = 0.94). The normalized gene expression matrix was subjected to single-

linkage hierarchical clustering using the program Cluster, and the result viewed using Treeview and the Matrix2png web server (`http://www.chibi.ubc.ca/matrix2png/`). Third party data were obtained from various sources, all listed in the Results. Statistical analyzes were performed in R (`http://r-project.org`).

Analysis of H-NS binding affinity to DNA sites
In vitro affinity data for H-NS to 8-mer oligonucleotides were generated by Gordon and colleagues [241] using a protein binding microarray in which various oligonucleotides were spotted and their affinities to a protein (H-NS in this case) applied to the array measured. The binding affinity towards each oligo was expressed in the form of a z-score (Z). H-NS and StpA bind to similar target sites; therefore the above-described semi-quantitative and relative measure of affinity of different DNA sites to H-NS can be readily applied to StpA. We first defined high affinity H-NS/StpA binding sites as those with $Z \geq 2$, which includes $\sim 5\%$ of all the oligonucleotides tested. These sites include a majority of sequences containing the sites ATAAA (56%) and AATAA (68%), which form part of the high-affinity H-NS binding motifs identified by computational analysis by Lang et al. [204] and Kahramanoglou et al. [67] respectively. These sites should be clearly contrasted with the much larger binding regions identified by ChIP-seq or ChIP-chip analysis; in fact most ChIP-seq/ChIP-chip binding regions would contain one or more binding motifs. We scored each ChIP-seq-defined H-NS binding region by the density of high affinity sites contained in them, i.e. the proportion of all 8-mers within the binding region (using a sliding window of 1 nt) that were classified as having high affinity for H-NS. This value was then transferred to the gene associated with the binding region.

4.3.2 Converging evolutionary strategies redress global gene expression imbalances caused by the loss of a global gene silencing system in *E. coli*

In the previous section we described how the loss of *hns* and *stpA* gens dramatically upsets the gene expression balance of the cell, by causing otherwise silent loci to be expressed at exceptionally high levels, at least in part due to pervasive transcription from A+T-rich sequence stretches to which H-NS and StpA bind. The Δ*hns-stpA* double mutant of E. coli K12 also suffers from a severe reduction in growth rate (see Fig. 4.5b, at least a part of which might emerge from the "cost" associated with the uncontrolled expression of genes which are silent in wildtype cells).

We exploited these characteristics of the Δ*hns-stpA* double mutant to investigate the evolutionary trajectories that help E. coli redress the global disruption of its gene expression state, using in-vitro laboratory evolution experiments, analyzed using deep sequencing of genomes and transcriptomes.

Results and Discussion

In-vitro laboratory evolution of Δhns-stpA

We asked whether and how the bacterium would adapt to such a global upset of its gene expression state. Towards this, we performed in-vitro laboratory evolution experiments (see Fig. 4.12) in which two parallel populations of Δ*hns-stpA* – derived from the same colony, isolated and frozen soon after the experimental deletion of *hns* and *stpA* – were grown in LB medium in 24-well plates under batch conditions. Subculturing into fresh medium was performed towards the end of exponential growth or early in stationary phase. This ensures that selection will be for increasing growth rate, and not be influenced by stationary phase survival. Frozen stocks were made at regular intervals, so that evolution could be "replayed" from such intermediates in the event of a contamination at a later stage, or if required otherwise.



Figure 4.12: This figure shows a schematic representation of our laboratory evolution experiment. Green coloured boxes represent samples taken for population genomic DNA sequencing; cyan boxes show clones isolated for single colony paired-end genomic DNA sequencing; strains selected for transcriptome experiments are in gray.

We observed a rapid, but steady increase in growth rate over the evolution experiment which together covered only a period of ∼ 250 hours, spread across ∼ 20 serial batch cultures (Figure 4.13a). For further analysis we studied the characteristics of the genomes of two frames, one from the 100-hour population and the other from the 250-hour population (HS100 for the ∼ 100 growth-hour frame; HS250 for the ∼ 250 growth-hour frame; Figure 4.13b), and compared them with those of the parent (Δ*hns-stpA*). The discussion in this section will be largely restricted to our analysis of HS100, an early time-point in the evolution of Δ*hns-stpA* towards higher fitness.



Figure 4.13: Suppression of the growth defect of Δ*hns-stpA* during the course of a laboratory evolution experiment. (a) This figure shows a heatmap representation of growth curves of bacterial populations obtained after the specified hours of evolution. Each row represents a snapshot of the evolution experiment, and each column a time point during a single batch culture. The snapshots labeled HS100 ($\sim$ 108 hours, spread across 9 serial batch cultures) and HS250 ($\sim$ 252 hours, across 21 serial batch cultures) were selected for population genome sequencing. The wildtype (bottom row) and the Δ*hns-stpA* parent (top row) are also indicated. The colour in each cell is indicative of OD60, with cyan representing lowest values and orange the highest. (b) This figures shows the growth curves of HS100 (black), HS250 (red), the wildtype (green) and the Δ*hns-stpA* parent (blue). The error bars show the standard error across two biological and six technical replicates.

Growth recovery by σ_{38} inactivation

We performed deep sequencing (50-mer from a single end, Illumina HiSeq-1000) of populations – one for the parental population Δ*hns-stpA*, and one for each of the two evolved lines for HS100 and HS250, giving $\sim 250-380$ fold coverage of the haploid genome per sample. We first performed an analysis of point mutations and short indels in HS100 in comparison to Δ*hns-stpA*. This showed that the population is genetically heterogeneous, with differences between the two populations in the mutations accrued. All mutations detected were present only in a sub-population, indicating the presence of multiple trajectories. We do not discuss each of these mutations here, but present selected observations. About 40% of the sequence reads from one of the two populations

displayed a polymorphism in the *rpoS* gene, encoding the σ_{38} factor (RpoS) responsible for the general stress response of *E. coli* [296]. We performed paired-end sequencing (100-mer from each end) of the genomes of 16 colonies, all derived from the HS100 population carrying the mutation in *rpoS*. Bacteria from each of the 16 colonies showed higher growth rates in liquid LB medium when compared with Δ*hns-stpA*. Analysis of single-nucleotide variations in these data, using the BRESEQ pipeline [297] or otherwise from first principles, revealed the presence of polymorphisms in *rpoS* in 8 colonies. Several mutants carried a STOP codon at position E43 or Y61 of the σ_{38} amino acid sequence, thus indicating an inactivating mutation; another clone encoded a I154N mutation. That all these mutations lead to inactivation of the σ38 regulon was verified by a simple test for catalase activity, in which a patch of cells with active σ_{38} effervesce rapidly on addition of H2O2, whereas those lacking functional σ_{38} do not (data not shown); this emerges from the dependence of the expression of catalase gene expression on σ_{38}. To further validate the role of σ_{38} inactivation in increasing the growth rate of Δ*hns-stpA*, we constructed a Δ*rpoS-hns-stpA* triple deletion mutant and find that it shows a higher growth rate than Δ*hns-stpA*, to a level similar to that of the HS100 clones.

A large amplification and growth recovery

We performed a detailed analysis of the read coverage distribution of the HS100 population and compared it with that of Δ*hns-stpA*. The genome was divided into 200-nt bins, and the number of reads mapping within each bin calculated, and normalized to enable comparisons across samples if required. In the Δ*hns-stpA* population, we see a nearly uniform read coverage across the genome (Figure 4.14a), with exceptions including the rac prophage region, which appears to have been deleted at least in a subset of the population, and various other repetitive elements where unique read mapping is not possible.

The HS100 population however showed a dramatically different pattern. An ∼ 2 Mb region around the origin of replication had an ∼ 1.2-fold higher coverage than the remaining ∼ 2.5 Mb centered around the terminus (Figure 4.143b-d; this ratio will be called $\boldsymbol{R}_c$ below). This is not a gradual decrease in coverage from the origin to the terminus, which can be explained by the origin firing multiple times per cell division in rapidly-dividing exponential phase cells. Instead, the two domains are separated by sharp boundaries. To further validate the inferred presence of these breakpoints in sequencing depth, we performed the read coverage distribution analysis for the 16 HS100 colonies that we had sequenced. We noticed the presence of the same pattern, only clearer than in the population, in three of these colonies (20% of the colonies; Figure 4.14e-g). None of these colonies carried a mutation in *rpoS*, though they had a mis-sense mutation in the gene fusA encoding the translation elongation factor EF-G. Consistently across the data from the two HS100 populations and the three individual HS100 clones, the two boundaries are located at IS elements, one boundary being at *insC1* and the other at *insC5*. *insC1* and *insC5* are identical in sequence and are part of the *IS2* transposon, which is present in multiple copies in many *E. coli* strains. The

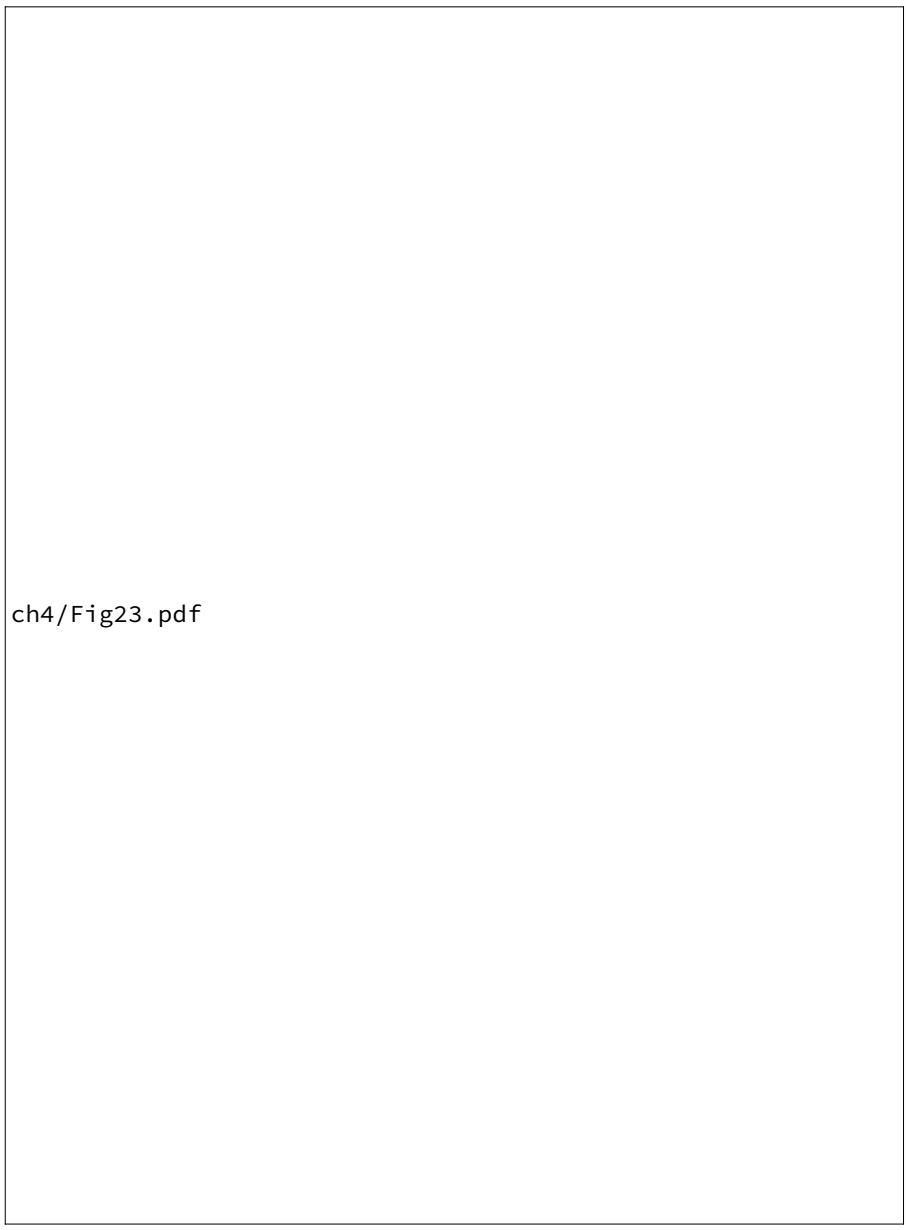

Figure 4.14: Amplification of $\sim 40\%$ of the E. coli chromosome around the origin of replication. The read coverage, on $\log_2$ scale, as a function of genome coordinate is plotted for (a) Δ*hns-stpA* and for (b) the evolved population HS100. (c) This panel shows the distribution of read coverage across the chromosome for HS100, directing the reader to its bimodal shape. (d) This panel shows the correlation in coverage between the two HS100 populations sequenced. (e-f) These sections show the read coverage – as a function of genome coordinates (left) and as a distribution (right) – for three clones carrying the duplication. (h) This figure shows a circular plot of the chromosome, indicating the following key elements: *insC1* and *insC5*, which are the end-points of the duplication; the origin (Ori) and the terminus (Ter) of replication; major structural domains of the chromosome as defined by the elegant definition macrodomains (outer circle) by Valens and co-workers [37], where the unstructured elements are in white. The plot was drawn using the NuST web server [276].

difference in coverage between the two regions of the chromosome could be validated by qPCR experiments with primers directed to specific loci encoded in the two segments.

Naively speaking, the data could be explained by both a deletion of the 2.5 Mb region around the terminus (in a sub-population within the colony), or an amplification of the 2 Mb domain around the origin. The former is unlikely, given the presence of the terminus of replication in the region, as well as the presence of a large number of genes essential for *E. coli* survival. On the other hand, amplification can be promoted by repetitive elements, including rRNA loci and transposable elements such as *IS2*, by unequal recombination [298]. In the single clone sequencing data, $\boldsymbol{R}_c$ is 1.5 – 1.8. That the fold difference does not approach two in these data may suggest rapid development of heterogeneity in the population during growth in liquid culture, possibly because of the instability of large duplications. We then used simulations of read coverage from genomic DNA populations wherein the proposed duplication is prevalent to different extents (Figure 4.15) to systematically show the following. $\boldsymbol{R}_c$ values in the 1.2 range, as observed in the read coverage plots for the HS100 population, can be explained by the presence of an amplification in $\sim$ 20% of the population; this is consistent with the number of HS100 clones (3 out of 16) showing this coverage pattern. Similarly, $\boldsymbol{R}_c$ values in the 1.5-1.8 range – as observed in the single-clone sequencing data – can be explained by an amplification in 60-80% of the population (Figure 4.16).

Reversal of the Δhns-stpA gene expression imbalance by σ_{38} inactivation

Next we selected, from the several fully-sequenced HS100 clones, one with an inactivating mutation in σ_{38} (*rpoS*^{mut}, carrying a Y61* mutation), and one with the large chromosomal amplification around the origin of replication (*ori*^{2}). These clones, alongside a Δ*rpoS-hns-stpA* strain were subjected to RNA-seq experiments to assess the transcriptional changes occurring in these mutants relative to the parental Δ*hns-stpA*. In *rpoS*^{mut}, $\sim$ 320 genes were up-regulated relative to Δ*hns-stpA*, and $\sim$ 420 down-regulated. A large proportion of genes ($\sim$ 80%; $\boldsymbol{P} < 10^{-10}$; Fisher's Exact Test) differentially expressed in *rpoS*^{mut} could be recapitulated in Δ*rpoS-hns-stpA*; further, many genes that were down-regulated in a *rpoS*^{mut} are known targets of σ_{38} ($\sim$ 30%; $\boldsymbol{P} < 10^{-10}$; Fisher's Exact Test) reported by the Hengge lab [299] and / or present in the RegulonDB database [223] ($\sim$ 400 genes in this collection of σ_{38} targets). These suggest that the transcriptional effects of *rpoS*^{mut} are indeed a result of σ_{38} inactivation.

Remarkably, a significant proportion of genes down-regulated in *rpoS*^{mut} when compared to Δ*hns-stpA* had previously been up-regulated in Δ*hns-stpA* relative to the wildtype ($\sim$ 35%; $\boldsymbol{P} \sim 10^{-7}$; Fisher's Exact Test; Figure 4.17a). Similarly, many genes up-regulated in the former had been down-regulated in the latter comparison ($\sim$ 50%; $\boldsymbol{P} < 10^{-10}$; Fisher's Exact Test; Figure 4.17a); up-regulation of gene expression resulting from the inactivation of a σ-factor is likely to be an indirect effect. The chance that genes which are differentially expressed in Δ*hns-stpA* relative to the wildtype would further change in expression in the same direction in *rpoS*^{mut} in comparison to Δ*hns-stpA* is significantly below random ($\boldsymbol{P} < 10^{-5}$ for up- and down-regulated genes; Fisher's Exact Test).

Previous research had demonstrated a direct link between H-NS and σ_{38}. Inactivation of H-

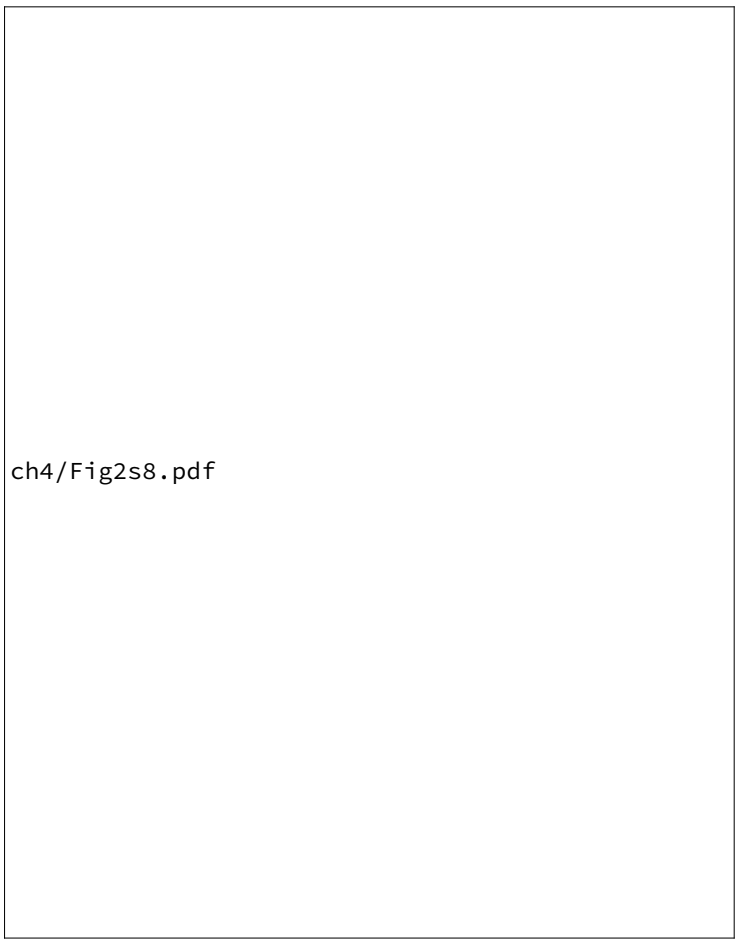

Figure 4.15: The above figure shows simulated read coverage coverage as a function of genome coordinate for different values of p in the deletion and the duplication scenarios. After generating 8000000 of random fragments for a given population, we plot the histogram of the fragments along the genome after binning it in segments 200 base pair long. The data is plotting for $p = 0.05, 0.1, 0.5$ in the two different scenarios

NS increases σ_{38} levels by stabilizing it through a cascade of regulatory interactions [300]. It is notable that in certain Salmonella, obtaining a Δ*hns* mutant is contingent on having a certain *rpoS* background [250]. In E. coli, partial inactivations of H-NS and σ_{38} interact at a genetic level to produce different physiological outcomes during distinct stages of growth [259]. Despite this, only a small proportion of genes known to be σ_{38} targets (60 genes out of 400 σ_{38} targets in our database) in wildtype E. coli are affected by Δ*hns-stpA* in comparison to the wildtype. Therefore, a small proportion (15%) of the $rpoS^{mut}$-mediated reversal of the transcriptional up-regulation seen in the parental Δ*hns-stpA* over the wildtype can be explained by previously known targets of σ_{38}. Given evidence that H-NS might block access of A+T-rich DNA to other proteins [301], including RNA polymerase [302], it might well be the case that σ_{38} has a direct role to play in transcribing H-NS targets, but only in the absence of H-NS; this would have been missed in the current datasets of σ_{38} targets, but may be provided by the transcriptome of Δ*rpoS-hns-stpA* provided by this study, which can be further developed by ChIP-chip / ChIP-seq studies of σ_{38} in a Δ*hns-stpA* background.

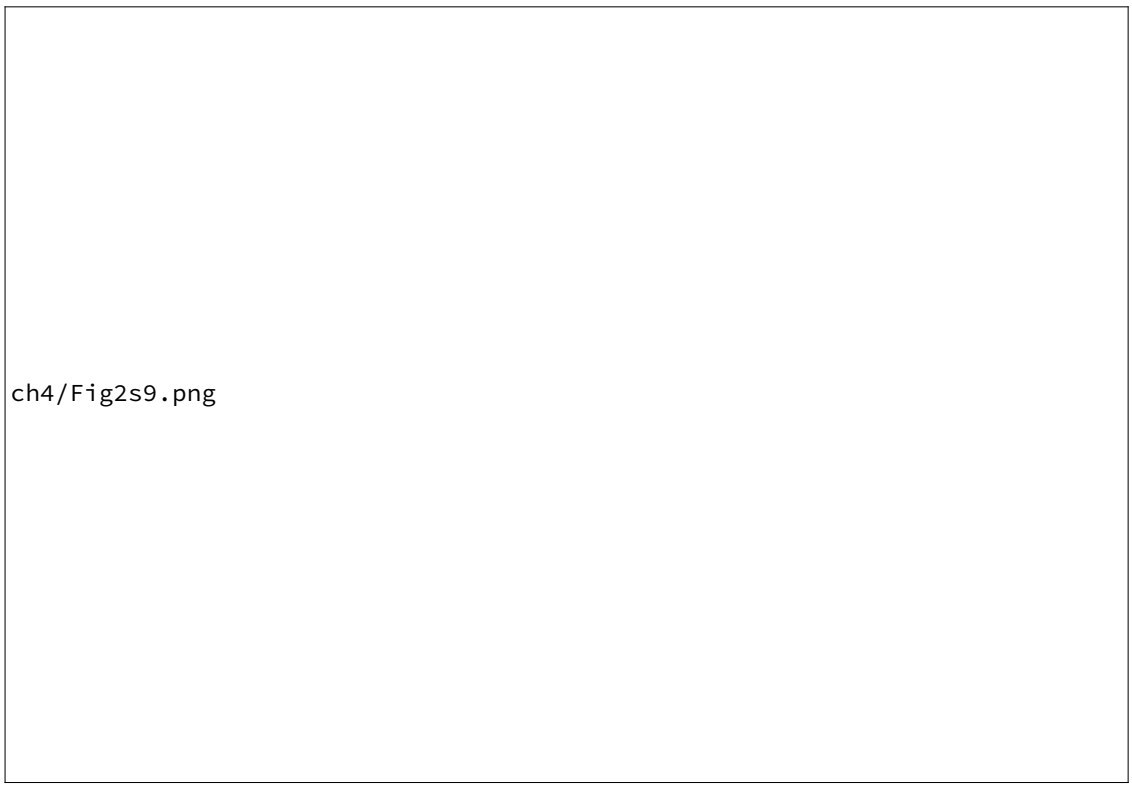

Figure 4.16: The two scatter plots on the left show the density of the randomized reads mapped to *E. coli* K12 chromosome for a chosen values of p. The dependence of the mean coverage as a function of the mutated nucleoid fraction p is different in the duplication and deletion scenario. It is possible to quantify the dependence of the fold change in mean coverage in function of p in the two scenarios (see right graphs). In the deletion scenario the fold change between the two regions diverges more as a function of p than in the duplication scenario.

In summary, σ_{38} inactivation in a Δ*hns-stpA* background results in a decrease in expression of several of its targets, while partially reversing the disruption of the gene expression imbalance experienced by Δ*hns-stpA*.

Converging transcriptional outputs of the large chromosomal amplification and σ_{38} inactivation

Towards analyzing the interplay between the large chromosomal amplification around the origin of replication and gene expression, we first tested for differential distribution of genes up- and down-regulated in Δ*hns-stpA*, between the amplified and the non-amplified segments of the chromosome. Comparing genes which are down-regulated in Δ*hns-stpA* relative to the wildtype with those that are up-regulated, the former are slightly more likely to be localized within the amplified segment of the chromosome than the latter ($P \sim 5 \cdot 10^{-3}$; Fisher's Exact Test). This is consistent with the observation that H-NS-repressed genes are more likely to be localized around the terminus of replication than the origin. The chromosome of the reference *E. coli* used here encodes six repeated *insC* elements, two of which form the boundaries of the amplification. Amplifications promoted by pairs of such elements located around the origin of replication could result

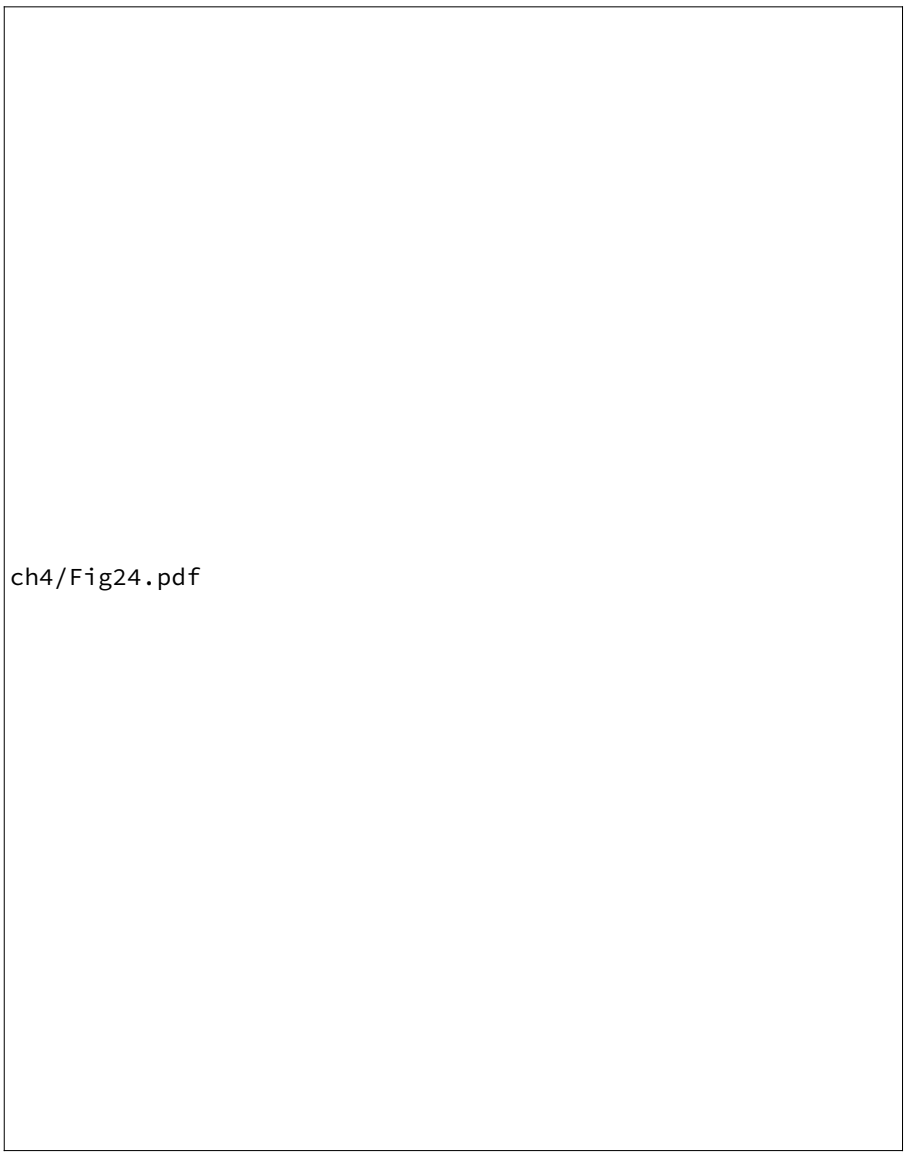

Figure 4.17: Converging transcriptional outputs of distinct evolutionary strategies. (a) This panel shows a heatmap of differential expression in three comparisons: Δ*hns-stpA* v. wildtype, *rpoS*^{mut} v. Δ*hns-stpA* and *ori*^{2} v. Δ*hns-stpA*. Red indicates up-regulation and blue, down-regulation. This shows that many of the transcriptional changes observed in the Δ*hns-stpA* v. wildtype comparison are reversed in the latter two comparisons. Note that this representation includes only those genes for which differential expression in the Δ*hns-stpA* v. wildtype comparison is reversed in a at least one of the other two comparisons shown. (b) This panel places the above result in the context of the chromosome organization, showing that in all the three comparisons presented, up- and down-regulation are concentrated in distinct chromosomal regions. In these graphs, the pie slices represent chromosomal regions which are statistically enriched ($P < 0.01$) for a certain set of genes, defined here by the set of genes differentially-expressed (up-regulated in red; down-regulated in blue) in a stated comparison. The plot considers the density of genes at different scales on the genome using grids of different bin sizes, and compares empirical data with results from a random null model (shuffled gene positions). The colours in the circle indicate chromosomal macrodomains (green: ori domain; blue: left domain; cyan: ter domain; red: right domain; white: unstructured elements).

in a similar distribution of H-NS-regulated genes between the amplified and the non-amplified part of the chromosome. However, we note that the pair of *insC* elements that we observe in our

study is unique in being located symmetrically around the origin of replication (concurrent replication forks), as well as being in or close to the non-structured regions of the chromosome (Figure 4.14h) [49], which are more likely to be involved in recombination with distal loci. These features might make the reported amplification a more accessible evolutionary strategy than recombination events between other pairs of repeats.

Next, we compared the transcriptome of the evolved *ori*^{2} strain against the parental Δ*hns-stpA*. For this, to account for possible artifacts that could be introduced in data analysis by the large duplication, we calculated the normalization factor for the RNA-seq data using reads from the non-amplified segment and then used this factor to normalize the whole dataset. We find that as many as $\sim$ 920 genes are up-regulated in *ori*^{2} relative to Δ*hns-stpA*, whereas $\sim$ 210 are down-regulated. As expected, a large majority of the up-regulated genes (66%; $P < 10^{-10}$; Fisher's Exact Test) are localized to the amplified region of the genome.

Despite the large difference in the number of differentially expressed genes between the *ori*^{2} and *rpoS*^{mut} evolved strains, they show convergent patterns (Figure 4.17a). Like the *rpoS*^{mut} mutant, the *ori*^{2} strain also partly reverses the gene expression state disruption in the Δ*hns-stpA* mutant, with *ori*^{2} up-regulating the expression levels of a majority of genes (54%; $P < 10^{-10}$; Fisher's Exact Test) that are down-regulated in Δ*hns-stpA* compared to the wildtype. Interestingly, *ori*^{2} also results in the down-regulation of several genes up-regulated in the Δ*hns-stpA* parent, with $\sim$ 42% ($P = 4 \cdot 10^{-10}$; Fisher's Exact Test) of genes down-regulated in *ori*^{2} having been up-regulated by Δ*hns-stpA* relative to the wildtype. Overall, a large proportion of genes up-regulated in *rpoS*^{mut} are also similarly affected in *ori*^{2} (62%; $P < 10^{-10}$; Fisher's Exact Test). Despite the similarities between the two evolved strains, the nature of the large amplification in *ori*^{2} results in a significantly larger number of genes being up-regulated, resulting in a greater reversal of the transcriptional down-regulation experienced by Δ*hns-stpA* relative to the wildtype.

Gene expression patterns in the context of chromosomal architecture

We tested for the presence of significant enrichments of differentially expressed genes in specific regions of the chromosome using the NuST web server (Figure 4.17b). This analysis shows the following: The direct role of H-NS and StpA in silencing gene expression by binding to the chromosome acts around the terminus of replication [49, 77]; the indirect consequence of the Δ*hns-stpA* mutant in down-regulating many highly expressed genes shows a slight preference to be localized in the origin-half of the chromosome, primarily in the unstructured region located on the left chromosomal arm (referred to as NS-L below). In contrast, the direct effect of *rpoS*^{mut} is the down-regulation of many genes around the terminus; this is consistent with previously published statistical analysis suggesting that stationary phase-induced genes tend to be encoded around the terminus [222]. However, this has the indirect consequence of up-regulating several genes in NS-L, which had been down-regulated by the Δ*hns-stpA* mutation. The *ori*^{2} mutant, while directly

up-regulating genes in the amplified region around the origin including many within NS-L, also results in the down-regulation of several H-NS-regulated genes located around the terminus. In these correlations, the effect – direct or indirect – around the terminus is very prominent, whereas that around the origin is less so. These together suggest a novel three-way link among H-NS-dependent gene silencing, σ_{38} function and chromosome organization in E. coli, with particular emphasis on the role of gene organization around the terminus of replication.

Towards investigating the opposing character of gene expression states between two orthogonal segments of the chromosome further, we performed a computational meta-analysis of publicly-available gene expression data across $\sim$ 300 different conditions. In this analysis, we calculated the Pearson correlation coefficients ($\boldsymbol{PCC}$; Figure 4.18a) between the gene expression vectors of all pairs of genes, and defined correlated ($\boldsymbol{PCC} \geq 0.5$) and anti-correlated ($\boldsymbol{PCC} \leq -0.5$) gene pairs. Analysis of anti-correlated gene pairs in the context of their position on the chromosome suggests the presence of two large chromosomal domains: genes localized around the terminus of replication tend to show anti-correlated gene expression vectors to those located on the origin half of the chromosome including the NS-L region (Figure 4.18b). Therefore, the opposing gene expression pattern between the terminus and the origin of replication that we observe in the mutants in the present work appears to be a general trend observed across a large number of genome-wide gene expression studies. This is in contrast to the chromosomal positioning of co-expressed gene pairs, which are more likely to be present in the same half of the chromosome (Figure 4.18c). However, the mechanisms underlying the persistent opposing gene expression patterns of distinct chromosomal regions – in particular the prospect of chromosome structure influencing these - remain to be addressed.

Materials and Methods

Strains and general growth conditions

The *Escherichia coli* variants used in the work are following: *E. coli* K-12 MG1655 wild type (CG SC # 6300); MG1655 Δ*hns-stpA* (Δ*hns-stpA::kanr*); Δ*hns-stpA-rpoS* (Δ*stpA-hns-rpoS::kanr*). HS100 refers to the population of Δ*hns-stpA* mutant evolved for 100 hours under laboratory conditions and HS250 to those evolved for 250 hours. *ori*^{2} refers to a single colony isolated from HS100, showing a duplication of $\sim$ 2 Mb region around the origin, and *rpoS*^{mut} to a single colony from HS100, with an inactivating point mutation in rpoS. Luria broth was used for normal growth, 50 mg/mL of ampicillin or kanamycin was used as per requirement Construction of E. coli K12 MG1655 knockouts All gene deletions were achieved by the Lambda Red recombination system, described by Datsenko and Wanner [295] using primers pKD46 and pKD4 or pKD13. Knockout strains generated by this method were selected on LB Kanamycin (50mg/mL) plates and deletion was confirmed by PCR using specific primers. Double knockouts were also generated by the same method after removing the kanamycin cassette using pCP20.



Figure 4.18: Gene expression correlations in the context of chromosome architecture: (a) This figure shows a histogram of the Pearson correlation coefficients between the expression vectors of all pairs of genes in E. coli. (b, c) In these heatmaps, the horizontal and the vertical indicate chromosomal bins of size 100kb each. The colour in each cell represents the number of gene pairs - with (b) anti-correlated or (c) correlated gene expression vectors - encoded in the pair of bins represented by it. White indicates low numbers and black high numbers. Below (b), numbers of genes differentially expressed across multiple conditions marked in the figure are shown as a function of genomic coordinate, aligned against the positions indicated by the heatmap above.

Laboratory evolution experiments

The laboratory evolution experiment was carried out for two parallel populations of an overnight grown culture, revived from the frozen stock made from a single colony of Δ*hns-stpA*. Cells were grown in 24-well plates, shaking at 200 rpm, until late exponential phase and diluted by a factor of 1:50 into fresh LB broth. Glycerol stocks were made at the end of every passage. HS100 and HS250 – populations evolved for $\sim$ 100 hrs and $\sim$ 250 hrs respectively – were selected for genome sequencing. For single colony genome sequencing, a population of HS100 or HS250 cells was plated on LB agar, and single colonies picked; glycerol stocks were made for all the single colonies isolated. This procedure is represented by a schematic in figure 4.12.

Growth curves

Overnight grown culture was inoculated in fresh LB to a 1:100 ratio and the growth of cells monitored by measuring the optical density at 600 nm. All these growth experiments were performed in 96 well plates, incubated at 37 °C in a plate reader (Tecan, infinite®F200 PRO) with constant shaking. OD600 was measured every $\sim$ 16 minutes. Each growth curve trial comprised two biological and six technical replicates.

Whole Genome Sequencing

Genomic DNA was isolated from Δ*hns-stpA* and evolved strains using SIGMA GenElute^{TM} Bacterial Genomic DNA Kit (Cat.No. NA2120) using the manufacturer's protocol. Sequencing library was prepared for Illumina sequencing and the quality of the libraries checked using Agilent Bioan-

alyzer at the genomics facility, Centre for Cellular and Molecular Platforms (C-CAMP), Bangalore. DNA isolated from HS100 and HS250 populations was sequenced for 50 cycles from one end and that from the single colonies sequenced for 100 cycles from both ends on an Illumina HiSeq1000 platform at C-CAMP. Raw data have been deposited with NCBI-SRA under the accession number SRP043310.

RNA extraction and mRNA enrichment

For RNA extraction, the overnight cultures were inoculated in 100 mL of fresh LB to bring the initial OD of the fresh culture to 0.03 and the flasks were incubated at 37°C with shaking at 200 rpm. Two biological replicates were performed for each sample. Samples were collected at the mid-exponential phase ($OD600 \sim 0.5$ for Δ*hns-stpA*; $0.7 \sim 0.8$ for $rpoS^{mut}$, Δ*hns-stpA-rpoS* and *ori*^{2}; ~ 0.9 for the wildtype). Protocols, based on those recommended by the manufacturer for the TRIzol (Invitrogen) bacterial RNA isolation kit, were used for RNA isolation as described previously [303]. Libraries were prepared for RNA-seq and sequenced for 50 cycles from one end on an Illumina HiSeq 1000 platform, following manufacturer's recommendations, at C-CAMP. Raw data have been deposited with NCBI-SRA under the accession number SRP043518.

Mutation calling

50-mer single end reads obtained from HS100 and HS250 populations were mapped to *E. coli* K12 MG1655 genome (NC000913, v2) using BWA [304] using the -q 20 parameter, sorted and indexed with SAMtools [305]. SNPs and indels were detected using VarScan [306]. These mutations were not detected in the parental genome of Δ*hns-stpA*. SNPs and indels were identified from the single colony genome sequencing data using the BRESEQ pipeline [297], which uses Bowtie [307] for sequence alignment; these results can be browsed at `http://bugbears.ncbs.res.in/hns_evol`. SNPs identified using this analysis were validated by Sanger sequencing using specific primers. Read coverage across the genome was calculated for non-overlapping windows of 200nt each using custom scripts, and normalized by the mode of the distribution across these bins as described earlier [303]. Coverage is represented on the $\log_2$ scale.

Transcriptome data analysis

50-mer single-end sequence reads were mapped to the E. coli K12 MG1655 genome using BWA [304]. Gene annotations were obtained from the Ecocyc database [214]. The number of reads falling within each gene was calculated based on the base position to which first nucleotide of the read was mapped. A matrix containing the read count for each gene across the sequenced samples was fed into the Bioconductor (`http://www.bioconductor.org`) package EdgeR [217] for analysis of differential expression. The genes which are differentially expressed by at least two-fold with a p-value of 0.00001 were considered for further analysis.

Differential expression in the *ori*^{2} strain was calculated differently. Here the mode of the distribution of per-gene read counts [303] across only the non-amplified portion of the genome was

used as a normalization factor. This was performed for *ori*^{2} as well as for the parental Δ*hns-stpA* and the wildtype. The normalized read counts were processed for differential expression using the LIMMA pipeline [308].

Publicly-available gene expression (transcriptome) data for E. coli were downloaded from the M3D database (`http://m3d.mssm.edu`; [309]). This comprises RMA-normalized gene expression measurements from Affymetrix microarrays across ∼ 300 conditions. Pearson correlation coefficients between pairs of gene expression vectors were calculated using a combination of PERL and R scripts. The chromosome was binned into segments of 100 kb each. For each pair of bins, the number of genes with correlated ($PCC \geq 0.5$) or anti-correlated ($PCC \leq -0.5$) gene expression vectors was computed and plotted as a matrix using the matrix2png web server (`www.chibi.ubc.ca/matrix2png`).

Validation of RNA sequencing data using RT-PCR

Selected RNA sequencing results - from comparisons involving *rpoS* mutants - were validated using RT-PCR. mRNA was isolated from the cells collected at the same time points as for RNA sequencing using TRIZOL, followed by DNAse treatment. mRNA after DNAse treatment was precipitated using ammonium chloride as described in the MICROBExpress RNA purification kit and used for the RT-PCR. RT-PCR reactions were carried out using the reagents provided with the Takara one step RT-PCR kit using manufacturer's instructions. 16S RNA gene, *rrsA* was used as an internal control.

Catalase Assay

10 µl of overnight grown culture was spotted on an LB agar plate and incubated at 37 °C for 24 hours, after which 10µL of H_2O_2 was added. The cells with active σ_{38} effervesce almost immediately, whereas those with inactive σ_{38} show delayed effervescence. Approximate time taken for the effervescence to start was measured and used as a proxy for σ_{38} activity.

Conclusions

Gene expression in bacteria is affected by a variety of regulators, including but not limited to nucleoid-associated, chromosome-shaping global transcription factors. Among these is the gene silencing system centered around the protein H-NS, which emerges as a tractable model system for studying the above-mentioned triangle. H-NS binds to A+T-rich sequence tracts, many of which are horizontally-acquired, and keeps them transcriptionally silent [67, 72, 242, 250, 310]. We had previously shown, using genomic-scale analysis, that the A+T-rich sequences bound by H-NS are intrinsically capable of high gene expression, which, as a cumulative increase in gene expression over ∼ 20% of the genome, could impose a high metabolic cost on the organism [303]. Recent research from David Grainger's laboratory has further shown that high expression from these loci

emerge from pervasive transcription from many promoter-like elements found in A+T-rich sequences [311]. We have also shown that the recruitment of a backup system for H-NS-mediated gene silencing – comprising the H-NS-homolog StpA – is more likely to be diverted to genes with high intrinsic transcription [303]. Therefore, the gene silencing function of H-NS and its partners is directed towards the silencing of highly-transcribable genes. An indirect consequence of the disruption of the H-NS-centered gene silencing mechanism is the down-regulation of transcription from a large number (10-12% of all genes) of otherwise highly-expressed genes. Thus, de-silencing horizontally-acquired genes results in a global disruption of the gene expression state of the cell. How does the cell adapt to such a circumstance, short of re-acquiring the silencing system?

The present work has shown that two distinct evolutionary strategies – inactivation of σ_{38} (RpoS), the σ-factor for general stress response, as well as a well-structured duplication of $\sim 40\%$ of the genome centered around the origin of replication – converge in partially redressing the transcriptional imbalance of a strain lacking the H-NS/StpA gene silencing system. The direct effects and the indirect consequences of these mutations appear to target orthogonal halves of the chromosome: stress-responsive / stationary phase-associated [222] and horizontally-acquired [77, 312] genes are encoded around the terminus of replication, and the consequence of their up-regulation being manifested in the down-regulation of many genes located closer to the origin of replication, and vice-versa. This presents an intriguing connection between the contrasting direct and indirect effects of two distinct global regulatory systems and chromosome architecture. These also indicate a certain determinism in where the indirect effects of a mutation in a global regulator are concentrated, and the mechanisms behind this will be of interest to future investigations. Meta-analysis of microarray data for $\sim$ 300 conditions show the presence of a certain anti-correlation in expression profiles between genes around the terminus and those around the origin of replication. Further, this study also indicates a role for σ_{38} in transcription of many H-NS targets. Polymorphisms in rpoS are common in both laboratory and environmental / pathogenic isolates of E. coli [313]. Studies in chemostats have also shown that σ_{38} inactivation leads to shorter doubling times in low nutrient environments, and that such environments select for loss-of-function mutations in *rpoS* [314]. In many genomes, transient amplification of DNA mediated by repetitive rRNA loci is common [315]. Further, duplication of large segments of the chromosome can be selected in low nutrient environments [316, 317]. Transient chromosomal aneuploidy has also been observed in yeast in response to stress [318]. Preliminary experiments performed by us do suggest that the Δ*hns-stpA* culture might in fact be experiencing a "low-nutrient" state; whether this is the ultimate selective force behind the emergence of the two mutations remains to be tested.

Nevertheless, the brute-force approach of knocking-out the general stress response – thus affecting the expression of a large number of genes – may not necessarily be wise, and large amplifications are unstable. In line with this, these variations exist transiently [318], as we do not see evidence for these in either of the two HS250 populations, or in the four HS250 clones picked for sequencing. For example, the HS250 populations harbor mutations that might affect the acid

stress response regulon of GadE [319], or the anaerobic-metabolic targets of AppY [320], both of which are influenced by H-NS and StpA. The HS250 clones also carry mutations in RecC, which is involved in recombination, and MutL, which acts in the mismatch repair pathway. These complement the previously reported property of CRP up-regulation to partially suppress the growth defect of Δ*hns-stpA* [287]. Further examination of the roles, if any, of these mutations in suppressing the growth defect of Δ*hns-stpA* is beyond the scope of this work. These specific affects confer a selective advantage over an en-masse inactivation of the σ_{38} regulon or unstable large amplifications. However, we suggest that the early increase in growth rates, as a result of two easily accessible mutational strategies – the global effects of the large chromosomal amplification (promoted easily by repetitive elements) and σ_{38} inactivation (a highly mutable locus) - might serve to increase the substrate pool for finer-level selection to act.

4.4 HI-C DATA ANALYSIS

The previous section describes the connection between the function of H-NS and StpA, both of which are NAPs, to gene expression. This analysis also showed how the set of genes whose expression is related to these two important global regulators are found in specific clusters along the chromosome coordinate, which are identifiable with large scale statistical tests (e.g. with NuST). The same question about the links between nucleoid organization and gene expression can be posed more in general. This latest section will introduces a preliminary analysis linking public available data of recent Chromosome Conformation Capture (3C) experiments linking public available data with large scale gene expression patterns. We gave a brief explanation of the technique of Chromosome conformation capture 3C in previous parts of the manuscript: an introduction to the experimental techniques can be found in previous parts of the manuscript (seq. 4.1.1) and a review of the main published results from the biophysical viewpoint is presented in chapter 2.

High-throughput 3C techniques in bacteria have been first used to determine the global folding architecture of the *C. crescentus* swarmer cell genome [36, 136], which is easier to access experimentally than *E. coli*, due to the well-characterized polar tethering of the chromosome and the experimental simplicity of cell synchronization. These data support a chromosomal fiber-like organization, linearly ordered in a compressed ring-like fiber, and taking a twisted toroidal shape, similar to the number eight inside the cell. Higher-resolution data [136] also show spatial domains of interactions, which exhibit a hierarchical nested organization over a range of length scales ($\sim 50 - 200$ Kb). These domains are stable throughout the cell cycle and are reestablished concomitantly with DNA replication. Additionally, domain boundaries co-occur with highly-expressed genes, and the domain organization is enabled by transcription. Such domains are hypothesized to be composed of transcription-induced supercoiled plectonemes arrayed into a "bottle brush" fiber. Regarding *E. coli*, the only currently available contact map has been published by the O'Sullivan group [321], while finer-scale experiments are expected to appear soon.

Figure 4.18 shows the general trend, observed across a large number of genome-wide gene expression studies, of an opposing gene expression pattern between the terminus and the origin of replication and a chromosomal positioning of co-expressed gene pairs, that are more likely to be localized around the Ori and Ter regions. The mechanism underlying those patterns still remains to be addressed. Here we ask the question of whether this prominent pattern of gene co-expression can be explained, at least partially, by the simple spatial proximity of the expressed genes in the bacterial cell. The most detailed maps of spatial proximity are nowadays obtained by the sequencing technique of 3C. The following section contains a possible outlook for a future work on a comparison of increasingly available 3C bacterial data with the co-expression matrix.

The next subsections introduce the data for both *E. coli* and *Caulobacter* and shows the result of the pipeline proposed recently by the Dekker group to obtain a contact probability matrix [221]. The data published by for *E. coli*, from ref. [321], is organized in 8 files downloadable from the SRA

database at NCBI, which correspond to the following experiments:

- `SRR554454.sra` Exponential interactions code "`L1A`",
- `SRR554455.sra` Exponential interactions code "`L1B`",
- `SRR554456.sra` Exponential interactions code "`L1C`",
- `SRR554457.sra` Exponential interactions code "`L2`",
- `SRR554458.sra` SHX treated interactions code "`S1A`",
- `SRR554459.sra` SHX treated interactions code "`S1B`",
- `SRR554460.sra` SHX treated interactions code "`S1C`",
- `SRR557710.sra` SHX treated interactions code "`S2`";

where the numbers in codes stand for the biological replicate while the letters correspond to different sequencing lanes; SHX stands for serine hydroxamate which is a drug that induces stringent response in *E. coli*. In the following section we present the analysis on the data relative to the experiment `L2`, since it is the dataset with highest coverage within the non-SHX treated replicates. Nevertheless, the same analysis performed on the other data set did not yield any significant differences on the results presented.

4.4.1 Alignment and filtering

The data consist in paired end reads, whose length is 100 sequenced bases per end. A paired end read is the result of DNA sequencing of a single fragment performed on both sides of the fragment. The use of the information that the read ends belong to the same fragment is essential in order to identify pairs of interacting 3C fragments. Read ends have been aligned using the publicly available software package Bowtie 2 [322] on the *E. coli* K12 reference genome. In order to guarantee the maximum number of aligned read ends, we employed an iterative alignment procedure which consist in aligning truncated read ends at 20bp length, then increasing the truncation by steps of 20bp only for unaligned read ends until only read ends truncated to the full 100 bp length are left unaligned. The number of reads that have been aligned at the end of this procedure is reported in table 4.2.

The pool of reads aligned to the reference genome on both ends is the only useful set of reads for the data analysis since each read end defines one of the two elements of the chromosome segment pair-interaction. The data analysis protocol includes two additional filtering steps before the construction of the interaction matrix. The first step corresponds to the filtering of duplicates, which are assumed to be artifacts of PCR amplification. The duplicates are paired reads which

Set	Total reads	Unaligned	Single	Both	No dup.	Filtered
L1A	23267246	362089	954467	21950690	6133305	1363227
L1B	34998975	577303	1454752	32966920	7124728	1656343
L1C	30187162	580814	1268580	28337768	7192510	1560418
L2	45661884	6561474	1919874	43085863	12943960	5507976
S1A	29372162	533989	1447731	27390442	7067945	2698946
S1B	26150212	394755	934048	24821409	7062768	2464849
S1C	33538443	530177	1360490	31647776	9131956	3602954
S2	47117446	790627	2303277	44023542	12913488	5547465

Table 4.2: Number of aligned reads in our analysis of the Hi-C data from ref. [321]: the number of reads per experiment is reported in this table. The first column (Total reads) shows the size (in paired-end reads) of the raw data. The second column reports how the reads are partitioned regarding to alignment, "unaligned" are the reads that cannot align on the reference genome, "single" are the reads that had only one end aligned, "both" are the reads that aligned on both ends; these are the only data that can be used in the subsequent steps of the data analysis. The third column reports the number of reads which are left after removal of duplicates, which are considered amplification artifacts. The fourth column reports the reads selected by the filtering out nearest-neighbors restriction fragment and dangling ends located at lower distance than the estimated read length.

align to exactly the same right and left base pair. The removal of duplicates reduces the number of aligned reads to about 27% (See table 4.2 column "No dup.").

The second filter excludes from the analysis reads that clearly do not belong to physical interactions but are instead by-products of the sequencing experiment. For instance pairs of read ends belonging to the same or the nearest-neighbors restriction fragment have a high probability of being read-ends of a single DNA segment, and the same holds for reads that are dangling-ends and are found at a lower distance than the estimated read-length (see Fig. 4.19). This filtering step corresponds to a cut-off on the interactions detected at short distances and thus is supposed to not alter the long-distance quantitative properties of the interaction matrix. The reads that are not filtered by this and the previous steps of the analysis are only about 10% of the original aligned reads (See table 4.2 column "Filtered").

4.4.2 Hi-C interaction map from available *E. coli* data

We plotted the distribution of read ends mapped along the chromosome (see Fig. 4.20a). In absence of sequencing bias, this plot would represent the specific propensity of interaction of each restriction fragment. Indeed, as expected, the general pattern follows the typical Ori-TER copy number variation, which appears in any DNA sequencing experiment in *E. coli* (see section 4.1.1).

Another important observable is the histogram of the distance between aligned read ends as a function of the inter loci arc-distance along the chromosome coordinate. In human datasets, this observable has been connected to the inter-loci contact probability [94]. The contact probability in human has been shown to depend on the distance as a power law ($p(s) \sim s^{-\alpha}$), the exponent α can be a tool to discern between realistic polymer models to explain the data [323, 324] (see chapters 2



Figure 4.19: Histogram of the distances between dangling aligned ends in all datasets. From the histogram the read length can be estimated to be in most cases of $\sim$ 800bp. In order to apply the low-distance cut-off, the mean read length must be estimated. The read length can be measured either by gel electrophoresis or from the distribution of distances between dangling end aligned pairs, both give consistent results of around $\sim$ 400bp in the original work on human [221].

and 3 of the thesis). The results of this histogram for the L2 dataset (other datasets being similar) is a quite uniform distribution with a sensible reduction of interactions for distances greater than 1Mb. The reason for such a flat distribution for lower distances, which does not fit any available polymer model, may be an artifact due either to a slow execution of the experiment or to an inefficient isolation of ligated interactions from other random sequences; the reduction of interactions for distances greater than 1Mb instead may be due to the copy number difference between Ori and the rest of the chromosome, although this effect is trivial, the conformation of the chromosome is known to depend on the velocity of replication [58].

Using the raw data of aligned read pairs, it is possible to construct a matrix. This is done by dividing the genome in bins of a suitable size (which depends on the coverage of interactions, we chose a bin-size of 40000bp), and, for each bin-bin pair, counting the number of reads having one end in each of the two considered bins. Figure 4.21a shows this matrix as a heatmap. In a matrix constructed this way the gene copy number information dominates over the interactions.

The task of distinguishing the sequencing bias from the propensity of making interactions (excluding the gene copy number pattern) is a difficult problem to solve in general. The protocol instead requires to normalize the data with a specific algorithm in order to subtract the bias and obtain a contact probability matrix. This matrix would have sums over row and columns equal to 1, the probability does not depend on the propensity of specific fragments to interact but only on the pair wise propensity of interacting between pairs of fragments from different chromosomal loci. An iterative correction algorithm [221] has been developed in order to subtract those biases. The algorithm assumes factorizable biases, which means that the raw matrix element, E_{ij}, for every pair of regions, (i, j), can be written as:

$$E_{ij} = B_i T_{ij} B_j, \tag{4.1}$$

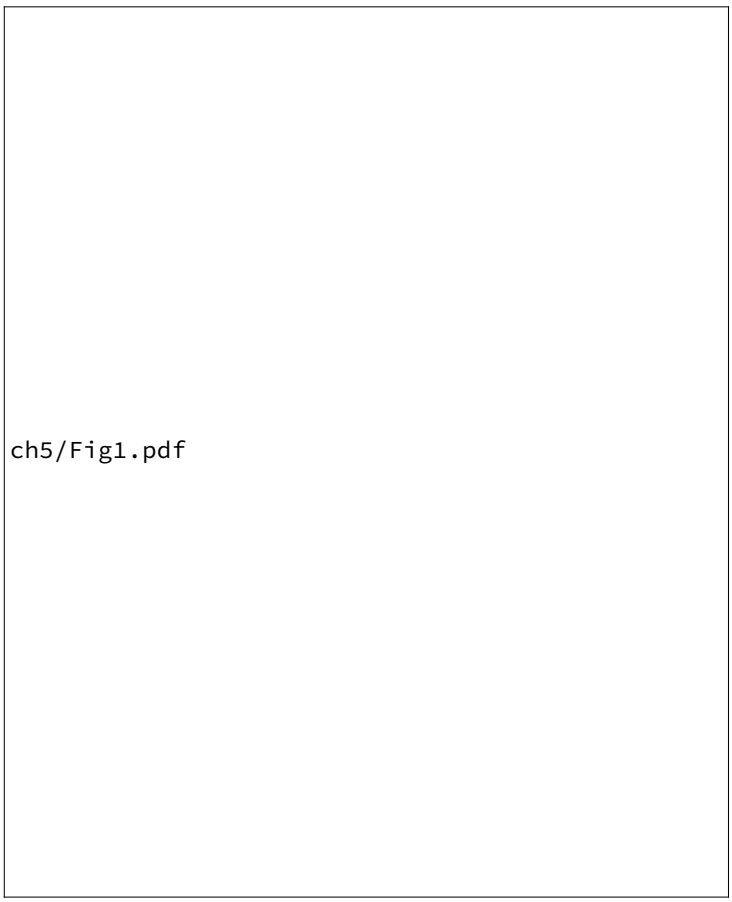

Figure 4.20: Interaction propensity and bias for *E. coli* Hi-C, L2 dataset. (a) The distribution of aligned read ends, at a first glance the plot follows the typical Ori-TER copy number variation along the genome, other variations along the pattern should be attributed either to sequencing bias or to specific propensity differences between individual restriction fragments to make interactions (e.g. trough binding of bridging proteins). (b) Hi-C measured contacts histogram as a function of genomic arc-length distance s, the graph shows a quite uniform distribution with a sensible reduction of interactions for distances greater than 1Mb. (c) Normalization bias, which is subtracted to the raw data after application of the iterative correction algorithm [221].

where $\boldsymbol{B}_i$ and $\boldsymbol{B}_j$ are the biases and $\boldsymbol{T}_{ij}$ is the sought matrix of relative contact probabilities, normalized as $\sum^j \boldsymbol{T}_{ij} = 1$. This normalization ensures a uniform coverage profile, equal visibility of each region in an iteratively corrected contact map and can reveal specific interactions otherwise buried by visibility-induced biases.

The factorizable bias as well as the real interaction matrix can be iteratively extracted from the data by repeating the following algebraic operations

$$\begin{aligned} &\text{(a)} \qquad \boldsymbol{Z}_j = \sum^i \boldsymbol{T}_{ij} \,/\, \left(\text{Mean}^j \left[\sum^i \boldsymbol{T}_{ij}\right]\right)^{1/2}; \\ &\text{(b)} \qquad \boldsymbol{T}_{ij} = \boldsymbol{T}_{ij} \,/\, \left(\boldsymbol{Z}_i \boldsymbol{Z}_j\right); \\ &\text{(c)} \qquad \boldsymbol{B}_j = \boldsymbol{B}_j \boldsymbol{Z}_j; \end{aligned} \tag{4.2}$$

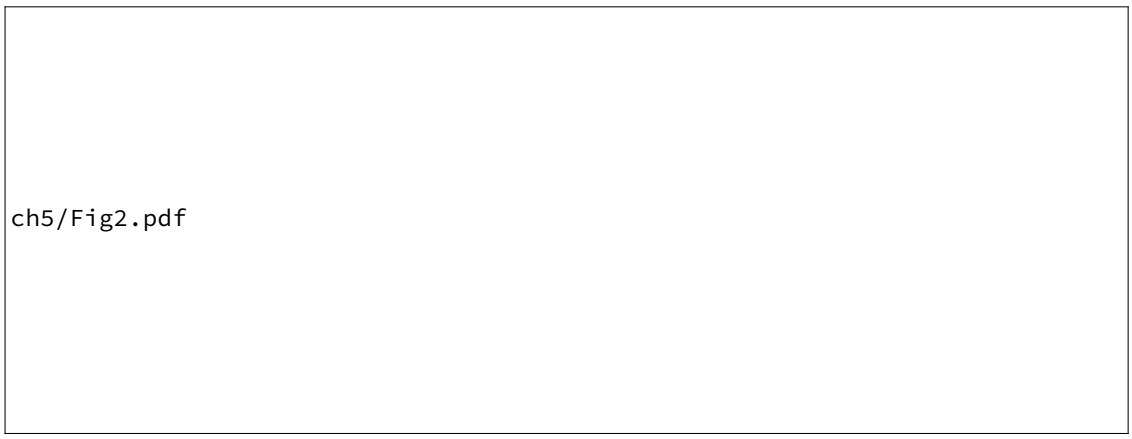

Figure 4.21: Heatmap of interaction matrices for *E. coli* Hi-C, L2 dataset. (a) Hi-C contact matrix, obtained by dividing the genome in bins of 40000bp size and counting the number of reads which ends align to each bin-bin pair, the gene copy number information dominates over interactions. (b) Contact probability matrix after the application of the iterative correction algorithm, note that this algorithm makes this matrix almost diagonal, this indicatess that the experiment did not produce information about genome interactions which is not factorizable into loci-specific biases.

starting with $\boldsymbol{B}_j = 1$ and $\boldsymbol{T}_{ij} = \boldsymbol{E}_{ij}$, until the variance of the column-sum vector is very small: $\mathrm{Var}^j \left[\sum^i \boldsymbol{T}_{ij}\right] < \epsilon$, where $\epsilon = 10^{-12}$ in our analysis. At the end of this process the matrix $\boldsymbol{T}_{ij}$ approximates the contact probability matrix and $\boldsymbol{B}_i$ is equal to the bias vector. Any information about interaction between fragment pairs, in matrix form, is not supposed to be factorizable and thus is left intact in the process if its signal is relevant in the dataset. After applying the algorithm to the L2 raw data matrix, we obtain the bias vector $\boldsymbol{B}_i$ reported in fig. 4.20c and the interaction probability matrix reported as an heatmap in fig. 4.21b; The plots refer to the L2 dataset, other datasets show similar results. The resultant bias vector is, as can be expected, highly correlated with the coverage, while the output interaction probability matrix is almost diagonal. This indicates that the experiment did not produce information about genome interactions that are not factorizable.

From this analysis we conclude that the currently available dataset of *E. coli* Hi-C, at the scales of observations considered by us, does not contain useful information for capturing the chromosome conformation because the data is mostly composed by factorizable bias. A detailed analysis of sequencing artifacts which could compose the bias the human Hi-C map is contained in reference [325] and indicates that GC content near the ligated fragment end is an important effect; we observe in *E. coli* that the main source bias at large scales is the gene copy number. The fact that we could not obtain non-factorizable interactions may be due to technical problems in the execution of the experiment.

4.4.3 Analysis on Caulobacter

Since the analysis on *E. coli* does not give us any conclusive information, we present in this subsection the result on another experiment done on *Caulobacter*. There are currently two available chromosome conformation capture datasets for *Caulobacter*: one published in ref. [36] and an-

other one with better coverage as well as higher resolution published in ref. [136]. We analyzed the data from the first of the two studies; unlike the one related to *E. coli*, the measurement of inter-loci interactions was not obtained through sequencing but through a combination of carbon copy and microarrays (5C). A matrix of interaction between restriction sites is publicly available for three different replicates (`REP1`, `REP2` and `REP3`). Thus, the data does not need alignment and filtering, but the last step of the previous analysis (the iterative correction algorithm) is still applicable. The total number of reported interactions for the three replicates is reported in table 4.3, in order to increase the coverage, we performed the analysis by merging the data of the three replicates (`SUM`).

Set	Total reported interactions
`REP1`	23267246
`REP2`	34998975
`REP3`	30187162
`SUM`	88453383

Table 4.3: Number of reported interactions for *Caulobacter*. The chromosome conformation capture data for this experiment [36] do not use the technique of sequencing to count interaction, nevertheless a quantitative estimate of the interactions between restriction fragments are reported in matrix form. The table reports the total number of interactions contained in this matrix.

Because of the low coverage and the non-homogeneity of the restriction sites, we had to bin the matrix in bins 80000bp wide, producing a coarse-grained matrix, this step is necessary to avoid large parts of the matrix with zero interactions and is assumed to not interfere with the large-scale interaction quantitative behavior. In the same way we did for the *E. coli* dataset, we plot the density of interactions as a function of the loci arc-length coordinate along the chromosome (Fig. 4.22a). As in the case of *E. coli*, the profile shows a *parS*-TER gradient dependent on the copy number and other biases.

We then plotted the histogram of the interactions as a function of the genomic distance (Fig. 4.22b). The result is very different from *E. coli*: in this case, for distances greater than 100000bp, the distribution follows a power law behavior $p(s) \sim s^{-\alpha}$ with parameter $\alpha = 1$. Surprisingly this result is similar to the observations in human genome [94], and in that case a fractal globular polymer has been proposed as an explanatory model [93]. However those trends are still subject of scientific debate [323, 324]. Notably, the same plot on the higher-resolution dataset [136] does not appear to follow the exact same trend, and more a detailed analysis is need to compare the two sources of data.

Figure 4.23a plots the heatmap of the contact map of the provided raw data matrix after coarse graining. Figure 4.22c plots the bias extracted from the application of the iterative correction algorithm and finally figure 4.23b is the heatmap of the corrected matrix (which is the 3C contact probability independent of bias). The plots show a structure with slowly decaying interactions along the chromosome backbone, with a symmetric crossed interaction structure between the left and the right arms. This tertiary structure reflects the fact that in *Caulobacter* the *parS* site is physically

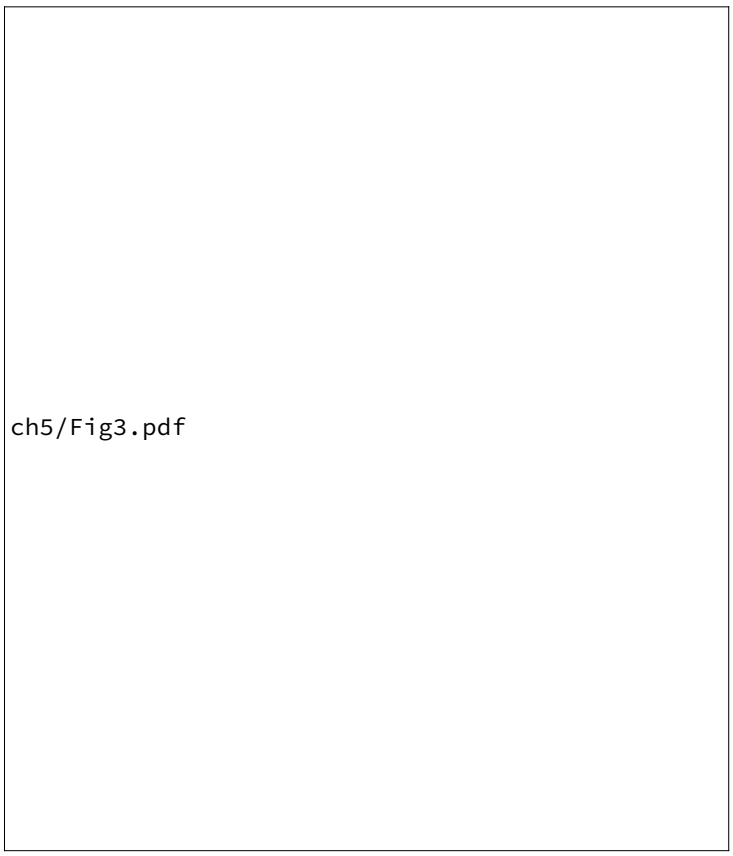

Figure 4.22: Interaction propensity and bias for *Caulobacter* 3C, SUM dataset. (a) The distribution of interactions, after rebinning to 80000bp, follows the typical *parS*-TER copy number variation along the genome. (b) 3C measured contacts histogram, as a function of genomic distance s, the plot shows a power law behavior $p(s) \sim s^{-\alpha}$ with parameter $\alpha = 1$. (c) Normalization bias, which is subtracted to the raw data after application of the iterative correction algorithm [221].

anchored to one cell pole, and the left and right arms fill the cell cytoplasm linearly and symmetrically along the cell axis toward the other pole, where Ter is located. In this case, the 3C contact probability (corrected data) shows a clear improvement over the raw data matrix.

4.4.4 Comparison with large scale expression data

The contact probability matrix of *Caulobacter* gives some understanding of the large scale conformation of the chromosome, which is uncomparable to any other available experimental technique. Unfortunately, available data available related to *E. coli* are affected by technical problems but it is nevertheless envisionable that another experiment could soon provide something of similar quality. This section reports a preliminary analysis of large scale genomic data about gene expression which could possibly be compared with nucleoid conformation. This comparison is an important testing ground for the hypothesis linking chromosome state and gene expression discussed in chapter 2. Additionally, it can leverage on abountant data, since collections of gene expression data

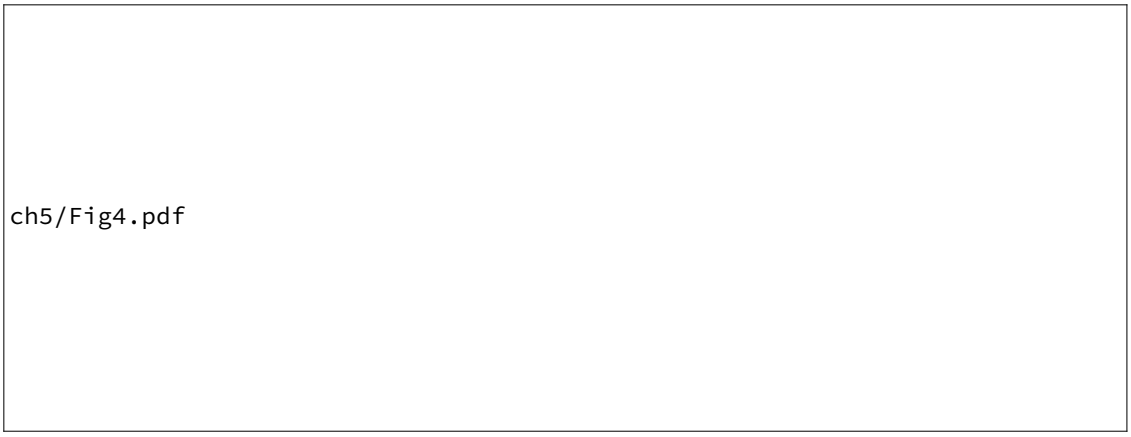

Figure 4.23: Heatmap of interaction matrices for *Caulobacter* 3C, SUM dataset. (a) 3C contact matrix, obtained by dividing the genome in bins of 80000bp size and counting the number of reads which ends align to each bin-bin pair. (b) Contact probability matrix after the application of the iterative correction algorithm, note that for *Caulobacter* the 3C contact probability (corrected data) shows a clear improvement over the raw data matrix.

available from past experiments are massively available.

We downloaded publicly-available gene expression (transcriptome) data for *E. coli* from the M3D database (http://m3d.mssm.edu; [309]). This comprises RMA-normalized gene expression measurements from Affymetrix microarrays across $\sim$ 300 conditions. In this dataset, high positive correlations between gene expression vectors is an indication that the expression of those genes respond in a coherent way (e.g. get both up-regulated or down-regulated) to multiple variations of living conditions.

Pearson correlation coefficients between pairs of gene expression vectors were calculated and the chromosome was binned into segments of 100Kbp each. For each pair of bins, the number of gene pairs with higly correlated expression (Pearson correlation coefficient – $PCC \geq 0.6$) was computed and stored in a matrix. The matrix plotted as an heatmap is reported in figure 4.25a. From the matrix we extracted observables that can be compared with 3C data. We extracted the distribution of correlated genes on the genome by summing over the matrix columns, which is equivalent to summing the number of correlated genes per bin (fig. 4.24a). This distribution showed a less marked Ori-TER pattern of copy number gradient along the chromosome compared to the *E. coli* Hi-C data.

The histogram of correlated genes plotted as a function of their genomic distance can be compared to the Hi-C measured contact histogram. We plot this histogram in figure 4.24b, the graph shows a power law behavior $p(s) \sim s^{-\alpha}$ with parameter $\alpha = -1.3$, more in accordance with results from *Caulobacter* 3C than *E. coli* Hi-C. Finally, we applied a cutoff for correlations at distances comparable to operon size ($s \leq 10000$bp) and the iterative correction algorithm on the correlation matrix, as for the 3C datasets, and obtained the heatmap of figure 4.25b.

We observe the the presence of two well-defined clusters of correlation around Ori and Ter, having comparable size to *E. coli* macrodomains. Also, differently from *E. coli* Hi-C but similarly

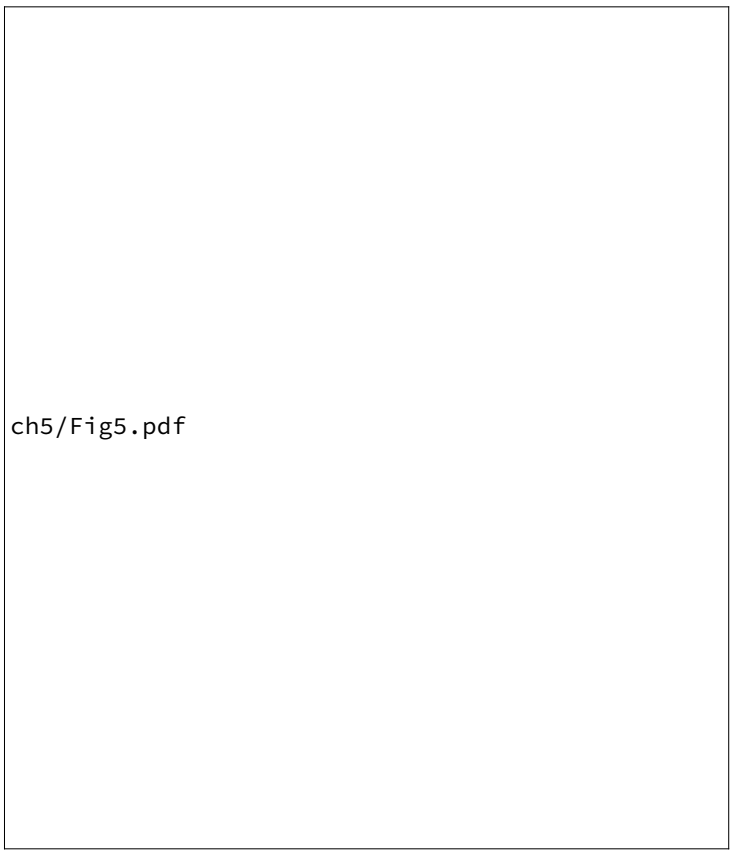

Figure 4.24: Correlation propensity and bias for *E. coli* – large scale collection of expression experiments. (a) The distribution of correlated genes, after rebinning to 100000bp wide bins, the Ori-TER pattern of copy number gradient along the chromosome is not important in comparison to the result of Hi-C (fig. 4.20a). (b) Histogram of correlated genes as a function of genomic arc-length distance s. The plot follows a power law behavior $p(s) \sim s^{-\alpha}$ with parameter $\alpha = -1.3$, more similar with results from *Caulobacter* 3C than *E. coli* Hi-C. (c) Normalization bias, which is subtracted to the raw data after application of the iterative correction algorithm [221].

to *Caulobacter* Hi-C, we observe a strong decaying correlation at small genomic distances along the genome backbone, which could be connected to the polymeric structure of the nucleoid. We also observed a symmetric crossed interaction structure between the left and the right arms around the Ori macrodomain.

At the present stage of the analysis and due to the conditions of existing experimental datasets, it has not been possible to compare the gene expression correlation matrix with the chromosome conformation capture of *E. coli*. However, some similarities appear between this matrix and the *Caulobacter* 3C contacts: the decay of the contacts with chromosome arc-length is power law with exponent $\alpha = 1$ while the decay of contacts in this latest analysis is still a power law with a slightly higher exponent $\alpha = 1.3$. The measure of this exponent demonstrates a long range structure of gene expression patterns which represents a novel discovery. Also, if the same exponent would be found in *E. coli* 3C, it may be a distinctive signature of *E. coli* with respect to *Caulobacter*. We asked the question of whether gene expression is linked to gene co-localization in the cytoplasm [88]; in

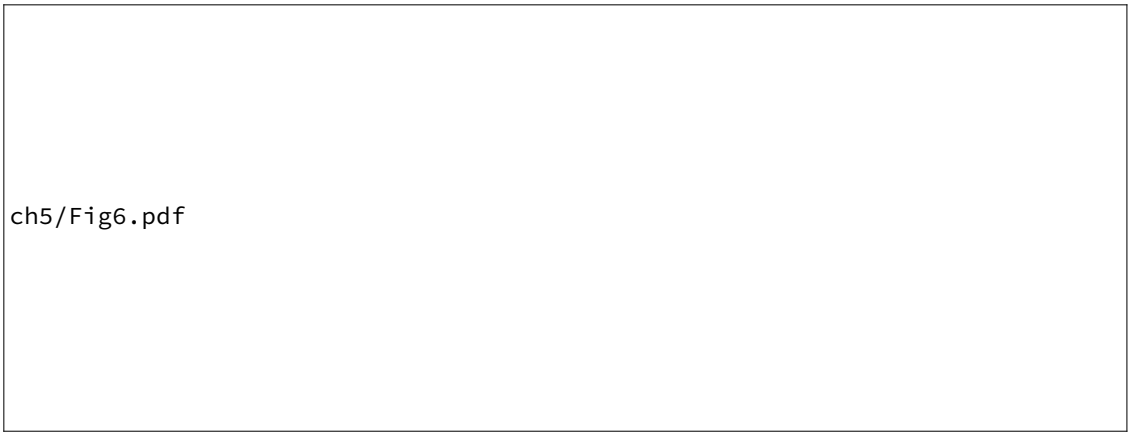

Figure 4.25: Heatmap of correlation matrices for *E. coli* – large scale collection of expression experiments. (a) correlation matrix, it is the same matrix reported in Fig. 4.18: the color in each cell represent the number of gene pairs with correlated gene expression vector. (b) Co-expression matrix after the application of a cutoff for correlations at small distances ($s \leq 10000$bp) and the iterative correction algorithm, note that the result is much more similar to the result of 3C in *Caulobacter* than the result in *E. coli*.

that regard, the presence of Ori and TER cluster in expression may be an effect of the macrodomain structure in those regions in *E. coli*. Also, if the output of a noise free *E. coli* Hi-C experiment would be similar to the measured *Caulobacter* contact matrix, then the crossed interaction structure between left and right arms around the Ori macrodomain which is present in both the expression matrix and in *Caulobacter* Hi-C would be a non-trivial signature that genes which are co-localized in the cytoplasm also are also co-expressed under different growth conditions.

Chapter 5
Conclusions and outlook

This thesis has described a set of studies using physical modeling and data analysis focused on the *E. coli* nucleoid organization, and on the links between nucleoid organization and gene expression.

Chapter 2 is a review of recent and less-recent results concerning chromosome organization, focusing in particular on the role of soft condensed matter physics in describing this system of fundamental biological importance. We made the effort of specifically recalling classic and more modern tools from the theory of polymers, as well as recent advances in experimental technologies. Our modeling effort started by the recent experimental evidence suggesting that protein-induced bridging and osmotic self-adhesion play an important role in determining the compaction of bacterial chromosomes.

The main results of chapter 3, which describes a polymer physics study based on these ingredients, are a crossover between Flory-like collapse, for dense bridging, and a switch-like collapse for sparse bridging, where bridging interactions compete with entropy loss terms from the looped arms grouped in a conformation resembling a star-like rosette. Furthermore, our model describes qualitatively the formation of multiple domains without intra-specificity of interactions. Our results lead us to speculate that, in real chromosomes, the features of the collapse transition and the formation of domains may be tunable by modulating the sparsity of bridging and the homogeneous self-adhesion, with a possible role in the observed relation between the position of genetic loci along the chromosome and their position in the bacterial cell and possibly also in the resolution of the identity of segregating sister chromosomes.

Finally chapter 4 presents a series of data-analysis studies investigating the links between bacterial chromosome organization and the global gene expression program. We described the NuST web server (`http://www.lgm.upmc.fr/nust/`) , which is a freely available computational tool developed to perform multi-scale analysis of the linear aggregation of specific gene sets along the chromosome. Our study of the epistatic effects of the homologous H-NS and StpA genes knock out shows that the global gene regulatory system around H-NS appears to be a multi-layered silencing system that operates in regions of the genome with high A+T content, many of which are likely products of recent horizontal gene transfer. The Δ*hns-stpA* double mutant of *E. coli* K12 also suffers from a severe reduction in growth rate, suppressor mutations (one of which was a global

duplication) were mostly aimed at directly up-regulating genes around the origin including many within non-structured-left macrodomain, also resulting in the down-regulation of several H-NS regulated genes located around the terminus. Finally, our preliminary analysis of large scale gene expression and chromosome conformation capture datasets shows important similarities (and differences) between patterns of coherent expression under different growth conditions in *E. coli* and colocalization in the cytoplasm in *Caulobacter*.

Future outlooks

The speculations stemming from the results of chapter 3 need to be addressed both by the development of an improved model and by experimental observation of the predicted behaviors. There are some important ingredients that are missing in the model but are possibly relevant from the biophysical point of view. (1) Bridging protein binding are not uniformly distributed along the chromosome, (2) there is configurational disorder due to the low binding affinity of nucleoid associated proteins, (3) active processes such as transcription and replication play a role in keeping the nucleoid conformation out of equilibrium, (4) the role of supercoiling changes the microscopical properties of DNA. In vitro techniques of non-invasive and label-free manipulation of purified nucleoids [92, 158] could be used to test the phase diagram of the model and the relevance of the additional ingredients. Further, extensive use of Hi-C [36, 94, 136, 321], or two color fluorescent tagging [22, 103, 110, 188, 326] *in vivo* could quantify organization properties of DNA loci under different conditions with a qualitative possible comparison with the model behavior.

The results of chapter 4 could take enormous advantage from high resolution data of the chromosome conformation capture of *E. coli* in multiple conditions. While we know that the gene expression patterns related to NAPs knock-out and supercoil perturbation [49, 75] are clustered and structured in a non-trivial way along the chromosome the evolutionary pressure at the origin of this pattern remains unknown. Whether it is the regulation of transcription, mediated by gene copy number and by the co-expression of genes closely located in the cytoplasm [18, 88], or some other forces which prevent a mixing of the gene order (which has been shown to be highly conserved in enterobacteria [30, 312]) is still an open question to be addressed. Hopefully addressing all these problems will contribute building a quantitative framework for bacterial physiology. The tools to prove how expression is related to chromosome conformation and gene spatial localization will shortly be available. We need an appropriate and generous application of cell biology and sequencing tecniques both to cell population as well as at the individual cell level, and the development of statistical tools of data analysis.

LIST OF PUBLICATIONS:

Submitted/in preparation:

6. V.F. Scolari, M. Cosentino Lagomarsino. *Combined collapse by bridging and self-adhesion in a prototypical polymer model inspired by the bacterial nucleoid*.
5. R. Srinivasan, V.F. Scolari, A.S.N Seshasayee. *Converging evolutionary strategies redress global gene expression imbalances caused by the loss of a global gene silencing system in E. coli*.

Published/accepted:

4. R. Srinivasan, D. Chandraprakash, R. Krishnamurthi, P. Singh, V.F. Scolari, S. Krishna, A.S.N. Seshasayee. *Genomic analysis reveals epistatic silencing of "expensive" genes in Escherichia coli*. (**Mol. BioSyst., 2013, 9,** 2021-2033)
3. V.G. Benza, K.D. Dorfman, V.F. Scolari, K. Bromek, P. Cicuta, M. Cosentino Lagomarsino. *Physical descriptions of the bacterial nucleoid at large scales, and their biological implications*. (2012 **Rep. Prog. Phys.** *75* 076602)
2. V.F. Scolari, B. Bassetti, B. Sclavi, M. Cosentino Lagomarsino. *Nucleoid Survey Tools - NuST*, Web-tool http://www.lgm.upmc.fr/nust/ (**Bioinformatics** (2012) **28** (12): 1643-1644)
1. V.F. Scolari, B. Bassetti, B. Sclavi, M. Cosentino Lagomarsino. *Gene clusters reflecting macrodomain structure respond to nucleoid perturbations*. (**Mol. BioSyst., 2011, 7,** 878-888)

Bibliography

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. Watson, *Molecular Biology of the Cell* (Garland, 2008), 5th ed.

[2] N. J. Trun and J. F. Marko, "Architecture of a bacterial chromosome (review)," American Society for Microbiology News **64**, 276 (1998).

[3] J. Stavans and A. Oppenheim, "Dna-protein interactions and bacterial chromosome architecture," Physical Biology **3**, R1– (2006).

[4] D. J. Sherratt, "Bacterial Chromosome Dynamics," Science **301**, 780–785 (2003).

[5] M. S. Luijsterburg, M. C. Noom, G. J. Wuite, and R. T. Dame, "The architectural role of nucleoid-associated proteins in the organization of bacterial chromatin: a molecular perspective," J Struct Biol **156**, 262–72 (2006).

[6] R. L. Ohniwa, Y. Ushijima, S. Saito, and K. Morikawa, "Proteomic analyses of nucleoid-associated proteins in escherichia coli, pseudomonas aeruginosa, bacillus subtilis, and staphylococcus aureus." PLoS One **6**, e19172 (2011).

[7] M. Thanbichler, S. C. Wang, and L. Shapiro, "The bacterial nucleoid: a highly organized and dynamic structure," J Cell Biochem **96**, 506–521 (2005).

[8] S. C. Dillon and C. J. Dorman, "Bacterial nucleoid-associated proteins, nucleoid structure and gene expression." Nat Rev Microbiol **8**, 185–195 (2010).

[9] C. Robinow and E. Kellenberger, "The bacterial nucleoid revisited." Microbiol Rev **58**, 211–232 (1994).

[10] F. Confalonieri, C. Elie, M. Nadal, C. de La Tour, P. Forterre, and M. Duguet, "Reverse gyrase: a helicase-like domain and a type i topoisomerase in the same polypeptide." Proc Natl Acad Sci U S A **90**, 4753–4757 (1993).

[11] J. F. Marko and S. Cocco, "The micromechanics of dna," Phys. world **March**, 37–41 (2003).

[12] L. Postow, C. Hardy, J. Arsuaga, and N. Cozzarelli, "Topological domain structure of the escherichia coli chromosome." Genes Dev **18**, 1766–79 (2004).

[13] D. Skoko, D. Yoo, H. Bai, B. Schnurr, J. Yan, S. M. McLeod, J. F. Marko, and R. C. Johnson, "Mechanism of chromosome compaction and looping by the escherichia coli nucleoid protein fis," Journal of Molecular Biology 364, 777–798 (2006).

[14] R. Kavenoff and B. C. Bowen, "Electron microscopy of membrane-free folded chromosomes from escherichia coli." Chromosoma 59, 89–101 (1976).

[15] M. Thanbichler and L. Shapiro, "Chromosome organization and segregation in bacteria." J Struct Biol 156, 292–303 (2006).

[16] R. de Vries, "Dna condensation in bacteria: Interplay between macromolecular crowding and nucleoid proteins." Biochimie 92, 1715–1721 (2010).

[17] R. Stuger, C. L. Woldringh, C. C. van der Weijden, N. O. E. Vischer, B. M. Bakker, R. J. van Spanning, J. L. Snoep, and H. V. Weterhoff, "Dna supercoiling by gyrase is linked to nucleoid compaction," Molecular Biology Reports 29, 79–82 (2002).

[18] D. J. Jin and J. E. Cabrera, "Coupling the distribution of rna polymerase to global gene regulation and the dynamic structure of the bacterial nucleoid in escherichia coli," J Struct Biol 156, 284–91 (2006).

[19] D. C. Grainger, D. Hurd, M. Harrison, J. Holdstock, and S. J. Busby, "Studies of the distribution of escherichia coli camp-receptor protein and rna polymerase along the e. coli chromosome," Proc Natl Acad Sci U S A 102, 17693–8 (2005).

[20] P. H. Viollier, M. Thanbichler, P. T. McGrath, L. West, M. Meewan, H. H. McAdams, and L. Shapiro, "Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial dna replication." Proc Natl Acad Sci U S A 101, 9257–9262 (2004).

[21] A. M. Breier and N. R. Cozzarelli, "Linear ordering and dynamic segregation of the bacterial chromosome." Proc Natl Acad Sci U S A 101, 9175–9176 (2004).

[22] P. A. Wiggins, K. C. Cheveralls, J. S. Martin, R. Lintner, and J. Kondev, "Strong intranucleoid interactions organize the Escherichia coli chromosome into a nucleoid filament," Proceedings of the National Academy of Sciences 107, 4991–4995 (2010).

[23] E. Toro and L. Shapiro, "Bacterial chromosome organization and segregation." Cold Spring Harb Perspect Biol 2, a000349 (2010).

[24] X. Wang, X. Liu, C. Possoz, and D. J. Sherratt, "The two escherichia coli chromosome arms locate to separate cell halves," Genes Dev 20, 1727–31 (2006).

[25] H. J. Nielsen, B. Youngren, F. G. Hansen, and S. Austin, "Dynamics of escherichia coli chromosome segregation during multifork replication," J Bacteriol 189, 8660–6 (2007).

[26] M. Scott and T. Hwa, "Bacterial growth laws and their applications." Curr Opin Biotechnol 22, 559–565 (2011).

[27] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa, "Interdependence of cell growth and gene expression: origins and consequences." Science 330, 1099–1102 (2010).

[28] A. Zaslaver, S. Kaplan, A. Bren, A. Jinich, A. Mayo, E. Dekel, U. Alon, and S. Itzkovitz, "Invariant distribution of promoter activities in escherichia coli." PLoS Comput Biol 5, e1000545 (2009).

[29] H. Ochman and E. A. Groisman, "The origin and evolution of species differences in escherichia coli and salmonella typhimurium." EXS 69, 479–493 (1994).

[30] E. P. C. Rocha, "The organization of the bacterial genome." Annu Rev Genet 42, 211–233 (2008).

[31] H. Niki, Y. Yamaichi, and S. Hiraga, "Dynamic organization of chromosomal dna in escherichia coli," Genes Dev 14, 212–23 (2000).

[32] X. Liu, X. Wang, R. Reyes-Lamothe, and D. Sherratt, "Replication-directed sister chromosome alignment in escherichia coli." Mol Microbiol 75, 1090–1097 (2010).

[33] H. J. Nielsen, J. R. Ottesen, B. Youngren, S. J. Austin, and F. G. Hansen, "The escherichia coli chromosome is organized with the left and right chromosome arms in separate cell halves," Mol Microbiol 62, 331–8 (2006).

[34] S.-H. Hong, E. Toro, K. I. Mortensen, M. A. D. de la Rosa, S. Doniach, L. Shapiro, A. J. Spakowitz, and H. H. McAdams, "Caulobacter chromosome in vivo configuration matches model predictions for a supercoiled polymer in a cell-like confinement," Proceedings of the National Academy of Sciences 110, 1674–1679 (2013).

[35] J.-C. Meile, R. Mercier, M. Stouf, C. Pages, J.-Y. Bouet, and F. Cornet, "The terminal region of the e. coli chromosome localises at the periphery of the nucleoid." BMC Microbiol 11, 28 (2011).

[36] M. A. Umbarger, E. Toro, M. A. Wright, G. J. Porreca, D. Baù, S.-H. Hong, M. J. Fero, L. J. Zhu, M. A. Marti-Renom, H. H. McAdams, L. Shapiro, J. Dekker, and G. M. Church, "The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation." Mol Cell 44, 252–264 (2011).

[37] M. Valens, S. Penaud, M. Rossignol, F. Cornet, and F. Boccard, "Macrodomain organization of the Escherichia coli chromosome." EMBO J 23, 4330–4341 (2004).

[38] O. Espéli, R. Mercier, and F. Boccard, "Dna dynamics vary according to macrodomain topography in the e. coli chromosome." Mol Microbiol 68, 1418–27 (2008).

[39] L. Moulin, A. Rahmouni, and F. Boccard, "Topological insulators inhibit diffusion of transcription-induced positive supercoils in the chromosome of escherichia coli." Mol Microbiol 55, 601–10 (2005).

[40] O. Espéli and F. Boccard, "Organization of the escherichia coli chromosome into macrodomains and its possible functional implications." J Struct Biol 156, 304–10 (2006).

[41] C. Lesterlin, R. Mercier, F. Boccard, F. Barre, and F. Cornet, "Roles for replichores and macrodomains in segregation of the escherichia coli chromosome." EMBO Rep 6, 557–62 (2005).

[42] R. Mercier, M.-A. Petit, S. Schbath, S. Robin, M. E. Karoui, F. Boccard, and O. Espéli, "The matp/mats site-specific system organizes the terminus region of the e. coli chromosome into a macrodomain." Cell 135, 475–485 (2008).

[43] R. T. Dame, O. J. Kalmykowa, and D. C. Grainger, "Chromosomal macrodomains and associated proteins: implications for dna organization and replication in gram negative bacteria." PLoS Genet 7, e1002123 (2011).

[44] M. A. Sánchez-Romero, S. J. W. Busby, N. P. Dyer, S. Ott, A. D. Millard, and D. C. Grainger, "Dynamic distribution of seqa protein across the chromosome of escherichia coli k-12." MBio 1 (2010).

[45] N. K. Tonthat, S. T. Arold, B. F. Pickering, M. W. V. Dyke, S. Liang, Y. Lu, T. K. Beuria, W. Margolin, and M. A. Schumacher, "Molecular mechanism by which the nucleoid occlusion factor, slma, keeps cytokinesis in check." EMBO J 30, 154–164 (2011).

[46] H. Cho, H. R. McManus, S. L. Dove, and T. G. Bernhardt, "Nucleoid occlusion factor slma is a dna-activated ftsz polymerization antagonist." Proc Natl Acad Sci U S A 108, 3773–3778 (2011).

[47] A. Mathelier and A. Carbone, "Chromosomal periodicity and positional networks of genes in escherichia coli." Mol Syst Biol 6, 366 (2010).

[48] M. Berger, A. Farcas, M. Geertz, P. Zhelyazkova, K. Brix, A. Travers, and G. Muskhelishvili, "Coordination of genomic structure and transcription by the main bacterial nucleoid-associated protein hu." EMBO Rep 11, 59–64 (2010).

[49] V. F. Scolari, B. Bassetti, B. Sclavi, and M. C. Lagomarsino, "Gene clusters reflecting macrodomain structure respond to nucleoid perturbations." Mol Biosyst 7, 878–888 (2011).

[50] I. A. Berlatzky, A. Rouvinski, and S. Ben-Yehuda, "Spatial organization of a replicating bacterial chromosome." Proc Natl Acad Sci U S A 105, 14136–14140 (2008).

[51] S. Jun and A. Wright, "Entropy as the driver of chromosome segregation," Nat Rev Micro 8, 600–607 (2010).

[52] M. C. Joshi, A. Bourniquel, J. Fisher, B. T. Ho, D. Magnan, N. Kleckner, and D. Bates, "Escherichia coli sister chromosome separation includes an abrupt global transition with concomitant release of late-splitting intersister snaps." Proc Natl Acad Sci U S A (2011).

[53] R. B. Jensen, S. C. Wang, and L. Shapiro, "A moving dna replication factory in caulobacter crescentus." EMBO J 20, 4952–4963 (2001).

[54] J. M. van Helvoort, J. Kool, and C. L. Woldringh, "Chloramphenicol causes fusion of separated nucleoids in escherichia coli k-12 cells and filaments." J Bacteriol 178, 4289–4293 (1996).

[55] J. M. V. Helvoort, P. G. Huls, N. O. Vischer, and C. L. Woldringh, "Fused nucleoids resegregate faster than cell elongation in escherichia coli pbpb(ts) filaments after release from chloramphenicol inhibition." Microbiology 144 (Pt 5), 1309–1317 (1998).

[56] V. Norris, "Hypothesis: chromosome separation in escherichia coli involves autocatalytic gene expression, transertion and membrane-domain formation." Mol Microbiol 16, 1051–1057 (1995).

[57] C. L. Woldringh and N. Nanninga, "Structural and physical aspects of bacterial chromosome segregation," J Struct Biol 156, 273–83 (2006).

[58] I. Junier, F. Boccard, and O. Espéli, "Polymer modeling of the e. coli genome reveals the involvement of locus positioning and macrodomain structuring for the control of chromosome conformation and segregation." Nucleic Acids Res (2013).

[59] R. Amit, A. B. Oppenheim, and J. Stavans, "Increased bending rigidity of single dna molecules by h-ns, a temperature and osmolarity sensor." Biophys J 84, 2467–2473 (2003).

[60] R. T. Dame, M. C. Noom, and G. J. L. Wuite, "Bacterial chromatin organization by h-ns protein unravelled using dual dna manipulation." Nature 444, 387–390 (2006).

[61] P. A. Wiggins, R. T. Dame, M. C. Noom, and G. J. L. Wuite, "Protein-mediated molecular bridging: a key mechanism in biopolymer organization." Biophys J 97, 1997–2003 (2009).

[62] J. van Noort, S. Verbrugge, N. Goosen, C. Dekker, and R. T. Dame, "Dual architectural roles of hu: formation of flexible hinges and rigid filaments," Proc Natl Acad Sci U S A 101, 6969–74 (2004).

[63] C. J. Lim, Y. R. Whang, L. J. Kenney, and J. Yan, "Gene silencing h-ns paralogue stpa forms a rigid protein filament along dna that blocks dna accessibility." Nucleic Acids Res (2011).

[64] R. Schneider, R. Lurz, G. Lüder, C. Tolksdorf, A. Travers, and G. Muskhelishvili, "An architectural role of the escherichia coli chromatin protein fis in organising dna." Nucleic Acids Res 29, 5107–5114 (2001).

[65] S. Maurer, J. Fritz, and G. Muskhelishvili, "A systematic in vitro study of nucleoprotein complexes formed by bacterial nucleoid-associated proteins revealing novel types of dna organization," Journal of Molecular Biology 387, 1261–1276 (2009).

[66] D. C. Grainger, D. Hurd, M. D. Goldberg, and S. J. W. Busby, "Association of nucleoid proteins with coding and non-coding segments of the escherichia coli genome," Nucleic Acids Research 34, 4642–4652 (2006).

[67] C. Kahramanoglou, A. S. N. Seshasayee, A. I. Prieto, D. Ibberson, S. Schmidt, J. Zimmermann, V. Benes, G. M. Fraser, and N. M. Luscombe, "Direct and indirect effects of h-ns and fis on global gene expression control in escherichia coli." Nucleic Acids Res 39, 2073–2091 (2011).

[68] T. Vora, A. K. Hottes, and S. Tavazoie, "Protein occupancy landscape of a bacterial genome," Molecular Cell 35, 247–253 (2009).

[69] C. Marr, M. Geertz, M. Hütt, and G. Muskhelishvili, "Dissecting the logical types of network control in gene expression profiles." BMC Syst Biol 2, 18 (2008).

[70] J. Kim, S. H. Yoshimura, K. Hizume, R. L. Ohniwa, A. Ishihama, and K. Takeyasu, "Fundamental structural units of the escherichia coli nucleoid revealed by atomic force microscopy," Nucleic Acids Res 32, 1982–92 (2004).

[71] D. C. Grainger, M. D. Goldberg, D. J. Lee, and S. J. Busby, "Selective repression by fis and h-ns at the escherichia coli dps promoter," Mol Microbiol 68, 1366–77 (2008).

[72] T. Oshima, S. Ishikawa, K. Kurokawa, H. Aiba, and N. Ogasawara, "Escherichia coli histone-like protein H-NS preferentially binds to horizontally acquired DNA in association with RNA polymerase." DNA Res 13, 141–53 (2006).

[73] J. T. Wade, K. Struhl, S. J. Busby, and D. C. Grainger, "Genomic analysis of protein-dna interactions in bacteria: insights into transcription and chromosome organization," Mol Microbiol 65, 21–6 (2007).

[74] D. C. Grainger, H. Aiba, D. Hurd, D. F. Browning, and S. J. Busby, "Transcription factor distribution in escherichia coli: studies with fnr protein," Nucleic acids research 35, 269–278 (2007).

[75] N. Blot, R. Mavathur, M. Geertz, A. Travers, and G. Muskhelishvili, "Homeostatic regulation of supercoiling sensitivity coordinates transcription of the bacterial genome," Embo Rep. 7, 710–715 (2006).

[76] M. D. Bradley, M. B. Beach, A. P. J. de Koning, T. S. Pratt, and R. Osuna, "Effects of fis on escherichia coli gene expression during different growth stages," Microbiology 153, 2922–2940 (2007).

[77] M. Zarei, B. Sclavi, and M. C. Lagomarsino, "Gene silencing and large-scale domain structure of the e. coli genome," Molecular BioSystems 9, 758–767 (2013).

[78] S. Cosgriff, K. Chintakayala, Y. T. A. Chim, X. Chen, S. Allen, A. L. Lovering, and D. C. Grainger, "Dimerization and dna-dependent aggregation of the escherichia coli nucleoid protein and chaperone cbpa." Mol Microbiol 77, 1289–1300 (2010).

[79] R. L. Ohniwa, K. Morikawa, J. Kim, T. Ohta, A. Ishihama, C. Wada, and K. Takeyasu, "Dynamic state of dna topology is essential for genome condensation in bacteria," EMBO J 25, 5591–602 (2006).

[80] T. Odijk, "Osmotic compaction of supercoiled dna into a bacterial nucleoid," Biophys Chem 73, 23–9 (1998).

[81] C. Zhang, P. G. Shao, J. A. van Kan, and J. R. C. van der Maarel, "Macromolecular crowding induced elongation and compaction of single dna molecules confined in a nanochannel." Proc Natl Acad Sci U S A 106, 16651–16656 (2009).

[82] A. Estevez-Torres and D. Baigl, "Dna compaction: fundamentals and applications," Soft Matter 7, 6746–6756 (2011).

[83] W.-H. Huang, A. A. Zinchenko, C. Pawlak, Y. Chen, and D. Baigl, "Dynamic conformational behavior and molecular interaction discrimination of dna/binder complexes by single-chain stretching in a microdevice." Chembiochem 8, 1771–1774 (2007).

[84] W. Xu and S. J. Muller, "Polymer-monovalent salt-induced dna compaction studied via single-molecule microfluidic trapping." Lab Chip 12, 647–651 (2012).

[85] S. B. Zimmerman, "Studies on the compaction of isolated nucleoids from escherichia coli," J Struct Biol 147, 146–58 (2004).

[86] S. B. Zimmerman, "Cooperative transitions of isolated escherichia coli nucleoids: implications for the nucleoid as a cellular phase," J Struct Biol 153, 160–75 (2006).

[87] S. B. Zimmerman, "Shape and compaction of escherichia coli nucleoids," J Struct Biol 156, 255–61 (2006).

[88] D. Marenduzzo, K. Finan, and P. R. Cook, "The depletion attraction: an underappreciated force driving cellular organization." J Cell Biol 175, 681–686 (2006).

[89] J. E. Cabrera, C. Cagliero, S. Quan, C. L. Squires, and D. J. Jin, "Active transcription of rrna operons condenses the nucleoid in escherichia coli: examining the effect of transcription on nucleoid structure in the absence of transertion." J Bacteriol 191, 4180–4185 (2009).

[90] D. Marenduzzo, C. Micheletti, and P. R. Cook, "Entropy-driven genome organization." Biophys J 90, 3712–3721 (2006).

[91] W. Wang, G.-W. Li, C. Chen, X. S. Xie, and X. Zhuang, "Chromosome organization by a nucleoid-associated protein in live bacteria." Science 333, 1445–1449 (2011).

[92] V. V. Thacker, K. Bromek, B. Meijer, J. Kotar, B. Sclavi, M. C. Lagomarsino, U. F. Keyser, and P. Cicuta, "Bacterial nucleoid structure probed by active drag and resistive pulse sensing." Integr Biol (Camb) (2013).

[93] L. A. Mirny, "The fractal globule as a model of chromatin architecture in the cell." Chromosome Res 19, 37–51 (2011).

[94] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." Science 326, 289–293 (2009).

[95] Grosberg, Rabin, Havlin, and Neer, "Crumpled globule model of the three-dimensional structure of dna." Europhys. Lett. 23 (5), 373–378 (1993).

[96] S. S. Daube, D. Bracha, A. Buxboim, and R. H. Bar-Ziv, "Compartmentalization by directional gene expression." Proc Natl Acad Sci U S A 107, 2836–2841 (2010).

[97] X. Wang, C. Lesterlin, R. Reyes-Lamothe, G. Ball, and D. J. Sherratt, "Replication and segregation of an escherichia coli chromosome with two replication origins." Proc Natl Acad Sci U S A 108, E243–E250 (2011).

[98] C. Spahn, U. Endesfelder, and M. Heilemann, "Super-resolution imaging of escherichia coli nucleoids reveals highly structured and asymmetric segregation during fast growth," Journal of Structural Biology 185, 243 – 249 (2014).

[99] C. Butan, L. M. Hartnell, A. K. Fenton, D. Bliss, R. E. Sockett, S. Subramaniam, and J. L. S. Milne, "Spiral architecture of the nucleoid in bdellovibrio bacteriovorus." J Bacteriol 193, 1341–1350 (2011).

[100] S. C. Weber, A. J. Spakowitz, and J. A. Theriot, "Bacterial chromosomal loci move subdiffusively through a viscoelastic cytoplasm," Phys. Rev. Lett. 104, 238102 (2010).

[101] K. Kolatka, M. Witosinska, M. Pierechod, and I. Konieczny, "Bacterial partitioning proteins affect the subcellular location of broad-host-range plasmid rk2." Microbiology 154, 2847–2856 (2008).

[102] A. I. Derman, G. Lim-Fong, and J. Pogliano, "Intracellular mobility of plasmid dna is limited by the para family of partitioning systems." Mol Microbiol 67, 935–946 (2008).

[103] A. Javer, Z. Long, E. Nugent, M. Grisi, K. Siriwatwetchakul, K. D. Dorfman, P. Cicuta, and M. Cosentino Lagomarsino, "Short-time movement of e. coli chromosomal loci depends on coordinate and subcellular localization." Nat Commun 4, 3003 (2013).

[104] M. C. Konopka, I. A. Shkel, S. Cayley, M. T. Record, and J. C. Weisshaar, "Crowding and confinement effects on protein diffusion in vivo." J Bacteriol 188, 6115–6123 (2006).

[105] T. Romantsov, I. Fishov, and O. Krichevsky, "Internal structure and dynamics of isolated escherichia coli nucleoids assessed by fluorescence correlation spectroscopy," Biophys J 92, 2875–84 (2007).

[106] S. Burov and E. Barkai, "Fractional langevin equation: overdamped, underdamped, and critical behaviors." Phys Rev E Stat Nonlin Soft Matter Phys 78, 031112 (2008).

[107] Y. He, S. Burov, R. Metzler, and E. Barkai, "Random time-scale invariant diffusion and transport coefficients." Phys Rev Lett 101, 058101 (2008).

[108] S. Condamin, V. Tejedor, R. Voituriez, O. Bénichou, and J. Klafter, "Probing microscopic origins of confined subdiffusion by first-passage observables." Proc Natl Acad Sci U S A 105, 5675–5680 (2008).

[109] S. C. Weber, J. A. Theriot, and A. J. Spakowitz, "Subdiffusive motion of a polymer composed of subdiffusive monomers." Phys Rev E Stat Nonlin Soft Matter Phys 82, 011913 (2010).

[110] A. Javer, N. J. Kuwada, Z. Long, V. G. Benza, K. D. Dorfman, P. A. Wiggins, P. Cicuta, and M. C. Lagomarsino, "Persistent super-diffusive motion of escherichia coli chromosomal loci," Nature communications 5 (2014).

[111] J. K. Fisher, A. Bourniquel, G. Witz, B. Weiner, M. Prentiss, and N. Kleckner, "Four-dimensional imaging of e. coli nucleoid organization and dynamics in living cells." Cell 153, 882–895 (2013).

[112] S. Cunha, T. Odijk, E. Suleymanoglu, and C. L. Woldringh, "Isolation of the escherichia coli nucleoid," Biochimie 83, 149–54 (2001).

[113] S. Cunha, C. L. Woldringh, and T. Odijk, "Restricted diffusion of dna segments within the isolated escherichia coli nucleoid," J Struct Biol 150, 226–32 (2005).

[114] S. Jun and B. Mulder, "Entropy-driven spatial organization of highly confined polymers: lessons for the bacterial chromosome." Proc Natl Acad Sci U S A 103, 12388–93 (2006).

[115] P. G. DeGennes, *Scaling concepts in Polymer Physics* (Cornell University Press, 1979).

[116] S. F. Edwards and K. F. Freed, "The entropy of a confined polymer. i," Journal of Physics A: General Physics 2, 145 (1969).

[117] Y. Jung and B.-Y. Ha, "Overlapping two self-avoiding polymers in a closed cylindrical pore: Implications for chromosome segregation in a bacterial cell." Phys Rev E 82 (5),, 51926–51931 (2010).

[118] D. Marenduzzo, C. Micheletti, and E. Orlandini, "Biopolymer organization upon confinement." J Phys Condens Matter 22, 283102 (2010).

[119] A. Arnold and S. Jun, "Time scale of entropic segregation of flexible polymers in confinement: implications for chromosome segregation in filamentous bacteria." Phys Rev E Stat Nonlin Soft Matter Phys 76, 031901 (2007).

[120] M. Buenemann and P. Lenz, "A geometrical model for dna organization in bacteria." PLoS One 5, e13806 (2010).

[121] M. Fritsche, S. Li, D. W. Heermann, and P. A. Wiggins, "A model for escherichia coli chromosome packaging supports transcription factor-induced dna domain formation." Nucleic Acids Res (2011).

[122] P. B. Warren and P. R. ten Wolde, "Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of escherichia coli." J Mol Biol 342, 1379–1390 (2004).

[123] T. Vettorel, A. Y. Grosberg, and K. Kremer, "Statistics of polymer rings in the melt: a numerical simulation study." Phys Biol 6, 025013 (2009).

[124] I. Junier, O. Martin, and F. Képès, "Spatial and topological organization of dna chains induced by gene co-localization," PLoS Comput Biol 6, e1000678 (2010).

[125] H. Diamant and D. Andelman, "Binding of molecules to dna and other semiflexible polymers." Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 61, 6740–6749 (2000).

[126] Kantor and Kardar, "Conformations of randomly linked polymers." Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 54, 5263–5267 (1996).

[127] T. Odijk and J. Ubbink, "Dimensions of a plectonemic dna supercoil under fairly general perturbations," Physica A: Statistical and Theoretical Physics 252, 61–66 (1998).

[128] J. Ubbink and T. Odijk, "Electrostatic-undulatory theory of plectonemically supercoiled dna." Biophysical journal 76, 2502–2519 (1999).

[129] B. H. Zimm and W. H. Stockmayer, "The dimensions of chain molecules containing branches and rings," The Journal of Chemical Physics 17, 1301+ (1949).

[130] J. U. Sommer and A. Blumen, "On the statistics of generalized gaussian structures: collapse and random external fields," Journal of Physics A: Mathematical and General 28, 6669 (1995).

[131] O. Farago and Y. Kantor, "Elasticity of gaussian and nearly gaussian phantom networks," Phys. Rev. E 62, 6094– (2000).

[132] W. W. Graessley, "Linear viscoelasticity in gaussian networks," Macromolecules 13, 372–376 (1980).

[133] Sumedha and M. Weigt, "A thermodynamic model for the agglomeration of dna-looping proteins," Journal of Statistical Mechanics: Theory and Experiment 2008, P11005 (2008).

[134] F. Jasch, C. von Ferber, and A. Blumen, "Dynamics of randomly branched polymers: configuration averages and solvable models." Phys Rev E Stat Nonlin Soft Matter Phys 68, 051106 (2003).

[135] M. Dolgushev and A. Blumen, "Dynamics of semiflexible treelike polymeric networks." J Chem Phys 131, 044905 (2009).

[136] T. B. K. Le, M. V. Imakaev, L. A. Mirny, and M. T. Laub, "High-resolution mapping of the spatial organization of a bacterial chromosome." Science 342, 731–734 (2013).

[137] H.-K. Janssen and O. Stenull, "Collapse transition of randomly branched polymers: Renormalized field theory." Phys Rev E Stat Nonlin Soft Matter Phys 83, 051126 (2011).

[138] J. Mateos-Langerak, M. Bohn, W. de Leeuw, O. Giromus, E. M. M. Manders, P. J. Verschure, M. H. G. Indemans, H. J. Gierman, D. W. Heermann, R. van Driel, and S. Goetze, "Spatially confined folding of chromatin in the interphase nucleus." Proc Natl Acad Sci U S A 106, 3812–3817 (2009).

[139] L. S. Shopland, C. R. Lynch, K. A. Peterson, K. Thornton, N. Kepper, J. von Hase, S. Stein, S. Vincent, K. R. Molloy, G. Kreth, C. Cremer, C. J. Bult, and T. P. O'Brien, "Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence." J Cell Biol 174, 27–38 (2006).

[140] M. Barbieri, M. Chotalia, L.-M. Lavitas, J. Fraser, J. Dostie, A. Pombo, and M. Nicodemi, "Complexity of chromatin folding: the strings and binders switch model." submitted (2011).

[141] A. Grosberg and A. Khokhlov, *Statistical Physics of Macromolecules* (AIP, NY, 1994).

[142] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank." Nucleic Acids Res 28, 235–242 (2000).

[143] M. B. Enright and D. M. Leitner, "Mass fractal dimension and the compactness of proteins." Phys Rev E Stat Nonlin Soft Matter Phys 71, 011912 (2005).

[144] C.-Y. Lee, "Mass fractal dimension of the ribosome and implication of its dynamic characteristics." Phys Rev E Stat Nonlin Soft Matter Phys 73, 042901 (2006).

[145] S. Reuveni, R. Granek, and J. Klafter, "Anomalies in the vibrational dynamics of proteins are a consequence of fractal-like structure." Proc Natl Acad Sci U S A 107, 13696–13700 (2010).

[146] E. Pitard and H. Orland, "Dynamics of the swelling or collapse of a homopolymer," Europhys. Lett. 41(4), 467–472 (1998).

[147] P. DeGennes, "Dynamics of entangled polymer solutions i. the rouse model," Macromolecules 9, 587 (1976).

[148] K. Wiese, "Dynamics of selfavoiding tethered membranes i model a dynamics (rouse model)," Eur. Phys. J. B1, 269–272 (1998).

[149] R. Reyes-Lamothe, C. Possoz, O. Danilova, and D. Sherratt, "Independent positioning and action of escherichia coli replisomes in live cells." Cell 133, 90–102 (2008).

[150] T. J. Lampo, N. J. Kuwada, P. A. Wiggins, and A. J. Spakowitz, "Physical modeling of chromosome segregation in e. coli reveals impact of force and dna relaxation," Biophysical Journal 106, 440a (2014).

[151] F. Képès and C. Valliant, "Transcription-based solenoidal model of chromosome," ComPlexUs 1, 171–180 (2003).

[152] J. F. Marko, "Linking topology of tethered polymer rings with applications to chromosome segregation and estimation of the knotting length." Phys Rev E Stat Nonlin Soft Matter Phys 79, 051905 (2009).

[153] M. Nicodemi and A. Prisco, "Thermodynamic pathways to genome spatial organization in the cell nucleus." Biophys J 96, 2168–2177 (2009).

[154] G. Muskhelishvili, P. Sobetzko, M. Geertz, and M. Berger, "General organisational principles of the transcriptional regulation system: a tree or a circle?" Mol Biosyst 6, 662–676 (2010).

[155] C. A. Brackley, S. Taylor, A. Papantonis, P. R. Cook, and D. Marenduzzo, "Nonspecific bridging-induced attraction drives clustering of dna-binding proteins and genome organization." Proc Natl Acad Sci U S A 110, E3605–E3611 (2013).

[156] M. Barbieri, M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo, and M. Nicodemi, "A model of the large-scale organization of chromatin." Biochem Soc Trans 41, 508–512 (2013).

[157] N. Hadizadeh Yazdi, C. C. Guet, R. C. Johnson, and J. F. Marko, "Variation of the folding and dynamics of the escherichia coli chromosome with growth conditions." Mol Microbiol 86, 1318–1333 (2012).

[158] J. Pelletier, K. Halvorsen, B.-Y. Ha, R. Paparcone, S. J. Sandler, C. L. Woldringh, W. P. Wong, and S. Jun, "Physical manipulation of the escherichia coli chromosome reveals its soft nature." Proc Natl Acad Sci U S A 109, E2649–E2656 (2012).

[159] A. S. Wegner, S. Alexeeva, T. Odijk, and C. L. Woldringh, "Characterization of escherichia coli nucleoids released by osmotic shock." J Struct Biol 178, 260–269 (2012).

[160] V. G. Benza, B. Bassetti, K. D. Dorfman, V. F. Scolari, K. Bromek, P. Cicuta, and M. C. Lagomarsino, "Physical descriptions of the bacterial nucleoid at large scales, and their biological implications." Rep Prog Phys 75, 076602 (2012).

[161] M. Barbieri, A. Scialdone, A. Piccolo, A. M. Chiariello, C. di Lanno, A. Prisco, A. Pombo, and M. Nicodemi, "Polymer models of chromatin organization." Front Genet 4, 113 (2013).

[162] M. Fritsche, S. Li, D. W. Heermann, and P. A. Wiggins, "A model for escherichia coli chromosome packaging supports transcription factor-induced dna domain formation." Nucleic Acids Res 40, 972–980 (2012).

[163] D. W. Heermann, H. Jerabek, L. Liu, and Y. Li, "A model for the 3d chromatin architecture of pro and eukaryotes." Methods 58, 307–314 (2012).

[164] M. Barbieri, M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo, and M. Nicodemi, "Complexity of chromatin folding is captured by the strings and binders switch model." Proc Natl Acad Sci U S A 109, 16173–16178 (2012).

[165] J. F. Marko and E. D. Siggia, "Stretching dna," Macromolecules 28, 8759–8770 (1995).

[166] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics* (Oxford, 1988).

[167] M. Kojima, K. Kubo, and K. Yoshikawa, "Elongation/compaction of giant dna caused by depletion interaction with a flexible polymer." J Chem Phys 124, 024902 (2006).

[168] S. Asakura and F. Oosawa, "Interaction between particles suspended in solutions of macromolecules," Journal of Polymer Science 33, 183–192 (1958).

[169] A. Y. Grosberg, I. Y. Erukhimovitch, and E. I. Shakhnovitch, "On the theory of ψ-condensation," Biopolymers 21, 2413–2432 (1982).

[170] P. J. Flory, "The configuration of real polymer chains," The Journal of Chemical Physics 17, 303–310 (1949).

[171] A. Cacciuto and E. Luijten, "Self-avoiding flexible polymers under spherical confinement," Nano Lett. 6, 901–905 (2006).

[172] M. G. Noro and D. Frenkel, "Extended corresponding-states behavior for particles with variable range attractions," The Journal of Chemical Physics 113, 2941–2944 (2000).

[173] P. De Gennes, "Collapse of a polymer chain in poor solvents," J. Phyique Lett. 36, 55–57 (1975).

[174] M. P. Taylor, W. Paul, and K. Binder, "All-or-none proteinlike folding transition of a flexible homopolymer chain," Phys. Rev. E 79, 050801 (2009).

[175] N. E. Buchler, U. Gerland, and T. Hwa, "On schemes of combinatorial transcription logic," Proceedings of the National Academy of Sciences 100, 5136–5141 (2003).

[176] L. Saiz and J. M. G. Vilar, "Dna looping: the consequences and its control." Curr Opin Struct Biol 16, 344–350 (2006).

[177] H.-P. Hsu, W. Nadler, and P. Grassberger, "Scaling of star polymers with 1-80 arms," Macromolecules 37, 4658–4663 (2004).

[178] L. Leibler, "Theory of microphase separation in block copolymers," Macromolecules 13, 1602–1617 (1980).

[179] Kantor and Kardar, "Collapse of randomly linked polymers." Phys Rev Lett 77, 4275 (1996).

[180] C. J. Camacho and T. Schanke, "From collapse to freezing in random heteropolymers," EPL (Europhysics Letters) 37, 603 (1997).

[181] A. K. Dasmahapatra, G. Kumaraswamy, and H. Nanavati, "Collapse transition in random copolymer solutions," Macromolecules 39, 9621–9629 (2006).

[182] Daoud, M. and Cotton, J.P., "Star shaped polymers : a model for the conformation and its concentration dependence," J. Phys. France 43, 531–538 (1982).

[183] O. V. Borisov and A. Halperin, "Polysoaps: The signatures of intrachain self assembly in θ solvents," Macromolecular Symposia 117, 99–107 (1997).

[184] O. V. Borisov and A. Halperin, "Micelles of polysoaps: the role of bridging interactions," Macromolecules 29, 2612–2617 (1996).

[185] Bryngelson and Thirumalai, "Internal constraints induce localization in an isolated polymer molecule." Phys Rev Lett 76, 542–545 (1996).

[186] F. Benedetti, J. Dorier, Y. Burnier, and A. Sasiak, "Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes." Nucleic Acids Res 42, 2848–2855 (2014).

[187] J. D. Halverson, J. Smrek, K. Kremer, and A. Y. Grosberg, "From a melt of rings to chromosome territories: the role of topological constraints in genome folding." Rep Prog Phys 77, 022601 (2014).

[188] C. Lesterlin, E. Gigant, F. Boccard, and O. Espeli, "Sister chromatid interactions in bacteria revealed by a site-specific recombination assay," The EMBO journal 31, 3468–3479 (2012).

[189] L. Schäfer, C. von Ferber, U. Lehr, and B. Duplantier, "Renormalization of polymer networks and stars," Nuclear Physics B 374, 473–495 (1992).

[190] A. Hanke and R. Metzler, "Entropy loss in long-distance dna looping," Biophysical journal 85, 167–173 (2003).

[191] J. Batoulis and K. Kremer, "Thermodynamic properties of star polymers: good solvents," Macromolecules 22, 4277–4285 (1989).

[192] G. S. Grest, "Structure of many-arm star polymers in solvents of varying quality: A molecular dynamics study," Macromolecules 27, 3493–3500 (1994).

[193] S. Caracciolo, M. S. Causo, and A. Pelissetto, "Monte carlo results for three-dimensional self-avoiding walks," Nuclear Physics B-Proceedings Supplements 63, 652–654 (1998).

[194] G. Zifferer, "Monte carlo simulation studies of the size and shape of linear and star-branched polymers embedded in the tetrahedral lattice," Macromolecular theory and simulations **8**, 433–462 (1999).

[195] K. Shida, K. Ohno, M. Kimura, and Y. Kawazoe, "Monte carlo study of the second virial coefficient and statistical exponent of star polymers with large numbers of branches," Macromolecules **33**, 7655–7662 (2000).

[196] T. Witten and P. Pincus, "Colloid stabilization by long grafted polymers," Macromolecules **19**, 2509–2513 (1986).

[197] R. T. Dame, "The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin," Mol Microbiol **56**, 858–70 (2005).

[198] G. Muskhelishvili and A. Travers, "Integration of syntactic and semantic properties of the dna code reveals chromosomes as thermodynamic machines converting energy into information," Cellular and Molecular Life Sciences **70**, 4555–4567 (2013).

[199] I. Cases, V. de Lorenzo, and C. A. Ouzounis, "Transcription regulation and environmental adaptation in bacteria." Trends Microbiol **11**, 248–253 (2003).

[200] A. Travers and G. Muskhelishvili, "DNA supercoiling - a global transcriptional regulator for enterobacterial growth?" Nat Rev Microbiol **3**, 157–169 (2005).

[201] S. Rimsky and A. Travers, "Pervasive regulation of nucleoid structure and function by nucleoid-associated proteins." Curr Opin Microbiol **14**, 136–141 (2011).

[202] R. T. Dame, M. S. Luijsterburg, E. Krin, P. N. Bertin, R. Wagner, and G. J. L. Wuite, "DNA bridging: a property shared among H-NS-like proteins." J Bacteriol **187**, 1845–1848 (2005).

[203] B.-K. Cho, E. M. Knight, C. L. Barrett, and B. Ø. Palsson, "Genome-wide analysis of fis binding in escherichia coli indicates a causative role for a-/at-tracts," Genome research **18**, 900–910 (2008).

[204] B. Lang, N. Blot, E. Bouffartigues, M. Buckle, M. Geertz, C. O. Gualerzi, R. Mavathur, G. Muskhelishvili, C. L. Pon, S. Rimsky *et al.*, "High-affinity dna binding sites for h-ns provide a molecular basis for selective silencing within proteobacterial genomes," Nucleic acids research **35**, 6330–6337 (2007).

[205] P. N. Hengen, S. L. Bartram, L. E. Stewart, and T. D. Schneider, "Information analysis of fis binding sites," Nucleic acids research **25**, 4994–5002 (1997).

[206] N. Sonnenschein, M.-T. Hütt, H. Stoyan, and D. Stoyan, "Ranges of control in the transcriptional regulation of escherichia coli," BMC systems biology **3**, 119 (2009).

[207] M. A. Wright, P. Kharchenko, G. M. Church, and D. Segre, "Chromosomal periodicity of evolutionarily conserved gene pairs," Proc Natl Acad Sci U S A **104**, 10559–64 (2007).

[208] F. Képès, "Periodic transcriptional organization of the e. coli genome," Journal of molecular biology **340**, 957–964 (2004).

[209] C. J. Dorman, "Genome architecture and global gene regulation in bacteria: making progress towards a unified model?" Nature Reviews Microbiology 11, 349–355 (2013).

[210] A. Lal and A. S. N. Seshasayee, "The impact of next-generation sequencing technology on bacterial genomics," in "A Systems Theoretic Approach to Systems and Synthetic Biology II: Analysis and Design of Cellular Systems," (Springer, 2014), pp. 31–58.

[211] M. J. Solomon, P. L. Larsen, and A. Varshavsky, "Mapping proteindna interactions in vivo with formaldehyde: evidence that histone h4 is retained on a highly transcribed gene," Cell 53, 937–947 (1988).

[212] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney *et al.*, "Genome-wide profiles of statı dna association using chromatin immunoprecipitation and massively parallel sequencing," Nature methods 4, 651–657 (2007).

[213] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," Bioinformatics 25, 1754–1760 (2009).

[214] I. M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muñiz-Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman *et al.*, "Ecocyc: a comprehensive database of escherichia coli biology," Nucleic acids research 39, D583–D590 (2011).

[215] J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein, "Peakseq enables systematic scoring of chip-seq experiments relative to controls," Nature biotechnology 27, 66–75 (2009).

[216] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," Nature Reviews Genetics 10, 57–63 (2009).

[217] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," Bioinformatics 26, 139–140 (2010).

[218] S. Anders and W. Huber, "Differential expression analysis for sequence count data," Genome biol 11, R106 (2010).

[219] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, "Capturing chromosome conformation," science 295, 1306–1311 (2002).

[220] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum *et al.*, "Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements," Genome research 16, 1299–1309 (2006).

[221] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny, "Iterative correction of hi-c data reveals hallmarks of chromosome organization," Nat Meth 9, 999–1003 (2012).

[222] P. Sobetzko, A. Travers, and G. Muskhelishvili, "Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle." Proc Natl Acad Sci U S A **109**, E42–E50 (2012).

[223] S. Gama-Castro *et al.*, "RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units)." Nucleic Acids Res **39**, D98–105 (2011).

[224] R. J. Nichols, S. Sen, Y. J. Choo, P. Beltrao, M. Zietek, R. Chaba, S. Lee, K. M. Kazmierczak, K. J. Lee, A. Wong *et al.*, "Phenotypic landscape of a bacterial cell," Cell **144**, 143–156 (2011).

[225] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori, "Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection," Molecular systems biology **2** (2006).

[226] S. Gerdes, M. Scholle, J. Campbell, G. Balazsi, E. Ravasz, M. Daugherty, A. Somera, N. Kyrpides, I. Anderson, M. Gelfand *et al.*, "Experimental determination and system level analysis of essential genes in escherichia coli mg1655," Journal of bacteriology **185**, 5673–5684 (2003).

[227] P. J. Kersey, D. Lawson, E. Birney, P. Derwent, M. Haimel, J. Herrero, S. Keenan, A. Kerhornou, G. Koscielny, A. Kähäri *et al.*, "Ensembl genomes: extending ensembl across the taxonomic space," Nucleic acids research **38**, D563–D569 (2010).

[228] S. Garcia-Vallve, E. Guzmán, M. Montero, and A. Romeu, "Hgt-db: a database of putative horizontally transferred genes in prokaryotic complete genomes," Nucleic acids research **31**, 187–189 (2003).

[229] J. E. Barrick, D. S. Yu, S. H. Yoon, H. Jeong, T. K. Oh, D. Schneider, R. E. Lenski, and J. F. Kim, "Genome evolution and adaptation in a long-term experiment with escherichia coli," Nature **461**, 1243–1247 (2009).

[230] O. Espéli and F. Boccard, "Organization of the Escherichia coli chromosome into macrodomains and its possible functional implications." J Struct Biol **156**, 304–310 (2006).

[231] W. Davids and Z. Zhang, "The impact of horizontal gene transfer in shaping operons and protein interaction networks–direct evidence of preferential attachment," BMC evolutionary biology **8**, 23 (2008).

[232] M. J. Lercher and C. Pál, "Integration of horizontally transferred genes into regulatory interaction networks takes many million years," Molecular biology and evolution **25**, 559–567 (2008).

[233] C. Pál, B. Papp, and M. J. Lercher, "Adaptive evolution of bacterial metabolic networks by horizontal gene transfer," Nature genetics **37**, 1372–1375 (2005).

[234] A. T. Maurelli, R. E. Fernández, C. A. Bloch, C. K. Rode, and A. Fasano, ""black holes" and bacterial pathogenicity: a large genomic deletion tha t enhances the virulence of shigella spp. and enteroinvasive escherichia coli," Proceedings of the National Academy of Sciences **95**, 3943–3948 (1998).

[235] E. P. Rocha, A. Danchin, and A. Viari, "Evolutionary role of restriction/modification systems as revealed by comparative genome analysis," Genome research 11, 946–958 (2001).

[236] A. S. N. Seshasayee, P. Singh, and S. Krishna, "Context-dependent conservation of dna methyltransferases in bacteria," Nucleic acids research 40, 7066–7073 (2012).

[237] C. Park and J. Zhang, "High expression hampers horizontal gene transfer," Genome biology and evolution 4, 523–532 (2012).

[238] W. W. Navarre, M. McClelland, S. J. Libby, and F. C. Fang, "Silencing of xenogeneic dna by h-ns—facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign dna," Genes & Development 21, 1456–1471 (2007).

[239] C. J. Dorman, "H-ns, the genome sentinel." Nat Rev Microbiol 5, 157–161 (2007).

[240] A. Martınez-Antonio and J. Collado-Vides, "Identifying global regulators in transcriptional regulatory networks in bacteria," Current opinion in microbiology 6, 482–489 (2003).

[241] B. R. Gordon, Y. Li, A. Cote, M. T. Weirauch, P. Ding, T. R. Hughes, W. W. Navarre, B. Xia, and J. Liu, "Structural basis for recognition of at-rich dna by unrelated xenogeneic silencing proteins," Proceedings of the National Academy of Sciences 108, 10690–10695 (2011).

[242] S. Lucchini, G. Rowley, M. D. Goldberg, D. Hurd, M. Harrison, and J. C. Hinton, "H-ns mediates the silencing of laterally acquired genes in bacteria," PLoS Pathogens 2, e81 (2006).

[243] R. Thei Dame, C. Wyman, and N. Goosen, "Structural basis for preferential binding of h-ns to curved dna," Biochimie 83, 231–234 (2001).

[244] T. A. Owen-Hughes, G. D. Pavitt, D. S. Santos, J. M. Sidebotham, C. S. Hulton, J. C. Hinton, and C. F. Higgins, "The chromatin-associated protein h-ns interacts with curved dna to influence dna topology and gene expression," Cell 71, 255–265 (1992).

[245] R. Spurio, M. Falconi, A. Brandi, C. L. Pon, and C. O. Gualerzi, "The oligomeric structure of nucleoid protein h-ns is necessary for recognition of intrinsically curved dna and for dna bending," The EMBO journal 16, 1795–1805 (1997).

[246] N. Sharadamma, Y. Harshavardhana, P. Singh, and K. Muniyappa, "Mycobacterium tuberculosis nucleoid-associated dna-binding protein h-ns binds with high-affinity to the holliday junction and inhibits strand exchange promoted by reca protein," Nucleic acids research 38, 3555–3569 (2010).

[247] S. Saxena and J. Gowrishankar, "Modulation of rho-dependent transcription termination in escherichia coli by the h-ns family of proteins," Journal of bacteriology 193, 3832–3841 (2011).

[248] C. C. Brescia, M. K. Kaw, and D. D. Sledjeski, "The dna binding protein h-ns binds to and alters the stability of rna in vitro and in vivo," Journal of molecular biology 339, 505–514 (2004).

[249] Y. Liu, H. Chen, L. J. Kenney, and J. Yan, "A divalent switch drives h-ns/dna-binding conformations between stiffening and bridging modes," Genes & development **24**, 339–344 (2010).

[250] W. W. Navarre, S. Porwollik, Y. Wang, M. McClelland, H. Rosen, S. J. Libby, and F. C. Fang, "Selective silencing of foreign dna with low gc content by the h-ns protein in salmonella," Science **313**, 236–238 (2006).

[251] C. J. Cardinale, R. S. Washburn, V. R. Tadigotla, L. M. Brown, M. E. Gottesman, and E. Nudler, "Termination factor rho and its cofactors nusa and nusg silence foreign dna in e. coli," Science **320**, 935–938 (2008).

[252] D. M. Stoebel, A. Free, and C. J. Dorman, "Anti-silencing: overcoming h-ns-mediated repression of transcription in gram-negative enteric bacteria," Microbiology **154**, 2533–2545 (2008).

[253] A. Zhang, S. Rimsky, M. Reaban, H. Buc, and M. Belfort, "Escherichia coli protein analogs stpa and h-ns: regulatory loops, similar and disparate effects on nucleic acid dynamics." The EMBO journal **15**, 1340 (1996).

[254] J. Sonnenfield, C. Burns, C. Higgins, and J. Hinton, "The nucleoid-associated protein stpa binds curved dna, has a greater dna-binding affinity than h-ns and is present in significant levels in hns mutants," Biochimie **83**, 243–249 (2001).

[255] E. Uyar, K. Kurokawa, M. Yoshimura, S. Ishikawa, N. Ogasawara, and T. Oshima, "Differential binding profiles of stpa in wild-type and hns mutant cells: a comparative analysis of cooperative partners by chromatin immunoprecipitation-microarray analysis," Journal of bacteriology **191**, 2388–2391 (2009).

[256] J. Johansson and B. E. Uhlin, "Differential protease-mediated turnover of h-ns and stpa revealed by a mutation altering protein stability and stationary-phase survival of escherichia coli," Proceedings of the National Academy of Sciences **96**, 10776–10781 (1999).

[257] C. M. Müller, U. Dobrindt, G. Nagy, L. Emödy, B. E. Uhlin, and J. Hacker, "Role of histone-like proteins h-ns and stpa in expression of virulence determinants of uropathogenic escherichia coli," Journal of bacteriology **188**, 5428–5438 (2006).

[258] B. Sonden and B. Uhlin, "Coordinated and differential expression of histone-like proteins in escherichia coli: regulation and function of the h-ns analog stpa." The EMBO journal **15**, 4970 (1996).

[259] S. Chib and S. Mahadevan, "Involvement of the global regulator h-ns in the survival of escherichia coli in stationary phase," Journal of bacteriology **194**, 5285–5293 (2012).

[260] S. Rodríguez, J. M. Nieto, C. Madrid, and A. Juárez, "Functional replacement of the oligomerization domain of h-ns by the hha protein of escherichia coli," Journal of bacteriology **187**, 5452–5459 (2005).

[261] R. C. Baños, A. Vivero, S. Aznar, J. García, M. Pons, C. Madrid, and A. Juárez, "Differential regulation of horizontally acquired and core genome genes by the bacterial modulator h-ns," PLoS genetics **5**, e1000513 (2009).

[262] A. Vivero, R. C. Baños, J. F. Mariscotti, J. C. Oliveros, F. García-del Portillo, A. Juárez, and C. Madrid, "Modulation of horizontally acquired genes by the hha-ydgt proteins in salmonella enterica serovar typhimurium," Journal of bacteriology **190**, 1152–1156 (2008).

[263] P. G. Leonard, S. Ono, J. Gor, S. J. Perkins, and J. E. Ladbury, "Investigation of the self-association and hetero-association interactions of h-ns and stpa from enterobacteria," Molecular microbiology 73, 165–179 (2009).

[264] G. S. Vernikos and J. Parkhill, "Interpolated variable order motifs for identification of horizontally acquired dna: revisiting the salmonella pathogenicity islands," Bioinformatics 22, 2196–2203 (2006).

[265] B. J. Yu, B. H. Sung, M. D. Koob, C. H. Lee, J. H. Lee, W. S. Lee, M. S. Kim, and S. C. Kim, "Minimization of the escherichia coli genome using a tn5-targeted cre/loxp excision system," Nature biotechnology 20, 1018–1023 (2002).

[266] G. Pósfai, G. Plunkett, T. Fehér, D. Frisch, G. M. Keil, K. Umenhoffer, V. Kolisnychenko, B. Stahl, S. S. Sharma, M. De Arruda *et al.*, "Emergent properties of reduced-genome escherichia coli," Science 312, 1044–1046 (2006).

[267] M. Hashimoto, T. Ichimura, H. Mizoguchi, K. Tanaka, K. Fujimitsu, K. Keyamura, T. Ote, T. Yamakawa, Y. Yamazaki, H. Mori *et al.*, "Cell size and nucleoid organization of engineered escherichia coli cells with a reduced genome," Molecular microbiology 55, 137–149 (2005).

[268] A. Wagner, "Energy constraints on the evolution of gene expression," Molecular biology and evolution 22, 1365–1374 (2005).

[269] R. Raghavan, Y. D. Kelkar, and H. Ochman, "A selective force favoring increased g+ c content in bacterial genes," Proceedings of the National Academy of Sciences **109**, 14504–14507 (2012).

[270] A. Deana and J. G. Belasco, "Lost in translation: the influence of ribosomes on bacterial mrna decay," Genes & development **19**, 2526–2533 (2005).

[271] K. J. McDowall, S. Lin-Chao, and S. N. Cohen, "A+ u content rather than a particular nucleotide order determines the specificity of rnase e cleavage." Journal of Biological Chemistry **269**, 10790–10796 (1994).

[272] J. M. Peters, R. A. Mooney, J. A. Grass, E. D. Jessen, F. Tran, and R. Landick, "Rho and nusg suppress pervasive antisense transcription in escherichia coli," Genes & development 26, 2621–2633 (2012).

[273] X. Wang, Y. Kim, Q. Ma, S. H. Hong, K. Pokusaeva, J. M. Sturino, and T. K. Wood, "Cryptic prophages help bacteria cope with adverse environments," Nature communications 1, 147 (2010).

[274] B. J. Peter, J. Arsuaga, A. M. Breier, A. B. Khodursky, P. O. Brown, and N. R. Cozzarelli, "Genomic transcriptional response to loss of chromosomal supercoiling in escherichia coli." Genome Biol 5(11), R87 (2004).

[275] P. Sobetzko, M. Glinkowska, A. Travers, and G. Muskhelishvili, "Dna thermodynamic stability and supercoil dynamics determine the gene expression program during the bacterial growth cycle," Molecular BioSystems 9, 1643–1651 (2013).

[276] V. F. Scolari, M. Zarei, M. Osella, and M. C. Lagomarsino, "Nust: analysis of the interplay between nucleoid organization and gene expression," Bioinformatics 28, 1643–1644 (2012).

[277] Z.-A. Ouafa, S. Reverchon, T. Lautier, G. Muskhelishvili, and W. Nasser, "The nucleoid-associated proteins h-ns and fis modulate the dna supercoiling response of the pel genes, the major virulence factors in the plant pathogen bacterium dickeya dadantii," Nucleic acids research 40, 4306–4319 (2012).

[278] S. Lucchini, P. McDermott, A. Thompson, and J. C. Hinton, "The h-ns-like protein stpa represses the rpos (σ38) regulon during exponential growth of salmonella typhimurium," Molecular microbiology 74, 1169–1186 (2009).

[279] S. C. Dillon, A. D. Cameron, K. Hokamp, S. Lucchini, J. C. Hinton, and C. J. Dorman, "Genome-wide analysis of the h-ns and sfh regulatory networks in salmonella typhimurium identifies a plasmid-encoded transcription silencing mechanism," Molecular microbiology 76, 1250–1265 (2010).

[280] R. C. Baños, J. I. Pons, C. Madrid, and A. Juárez, "A global modulatory role for the yersinia enterocolitica h-ns protein," Microbiology 154, 1281–1289 (2008).

[281] M. C. Noom, W. W. Navarre, T. Oshima, G. J. Wuite, and R. T. Dame, "H-ns promotes looped domain formation in the bacterial chromosome," Current Biology 17, R913–R914 (2007).

[282] A. I. Prieto, C. Kahramanoglou, R. M. Ali, G. M. Fraser, A. S. Seshasayee, and N. M. Luscombe, "Genomic analysis of dna binding and gene regulation by homologous nucleoid-associated proteins ihf and hu in escherichia coli k12," Nucleic acids research p. gkr1236 (2011).

[283] M. Cosentino Lagomarsino, P. Jona, B. Bassetti, and H. Isambert, "Hierarchy and feedback in the evolution of the escherichia coli transcription network," Proc Natl Acad Sci U S A 104, 5516–20 (2007).

[284] M. A. Savageau, "Comparison of classical and autogenous systems of regulation in inducible operons," Nature 252, 546–549 (1974).

[285] A. Gitter, Z. Siegfried, M. Klutstein, O. Fornes, B. Oliva, I. Simon, and Z. Bar-Joseph, "Backup in gene regulatory networks explains differences between binding and knockout results," Molecular systems biology 5 (2009).

[286] J. Ihmels, S. R. Collins, M. Schuldiner, N. J. Krogan, and J. S. Weissman, "Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss," Molecular systems biology 3 (2007).

[287] J. Johansson, C. Balsalobre, S.-Y. Wang, J. Urbonaviciene, D. J. Jin, B. Sondén, and B. E. Uhlin, "Nucleoid proteins stimulate stringently controlled bacterial promoters: a link between the camp-crp and the (p) ppgpp regulons in escherichia coli," Cell 102, 475–485 (2000).

[288] M. F. Traxler, S. M. Summers, H.-T. Nguyen, V. M. Zacharia, G. A. Hightower, J. T. Smith, and T. Conway, "The global, ppgpp-mediated stringent response to amino acid starvation in escherichia coli," Molecular microbiology **68**, 1128–1148 (2008).

[289] R. Benz, A. Schmid, C. Maier, and E. Bremer, "Characterization of the nucleoside-binding site inside the tsx channel of escherichia coli outer membrane reconstitution experiments with lipid bilayer membranes," European Journal of Biochemistry **176**, 699–705 (1988).

[290] A. Iida, S. Teshiba, and K. Mizobuchi, "Identification and characterization of the tktb gene encoding a second transketolase in escherichia coli k-12." Journal of bacteriology **175**, 5375–5383 (1993).

[291] M. E. Regonesi, M. Del Favero, F. Basilico, F. Briani, L. Benazzi, P. Tortora, P. Mauri, and G. Deho, "Analysis of the escherichia coli rna degradosome composition by a proteomic approach," Biochimie **88**, 151–161 (2006).

[292] C. D. Hardy and N. R. Cozzarelli, "A genetic selection for supercoiling mutants of escherichia coli reveals proteins implicated in chromosome structure," Molecular microbiology **57**, 1636–1652 (2005).

[293] I. Shachrai, A. Zaslaver, U. Alon, and E. Dekel, "Cost of unneeded proteins in e. coli is reduced after several generations in exponential growth," Molecular cell **38**, 758–767 (2010).

[294] K. Shiraishi, Y. Ogata, K. Hanada, Y. Kano, and H. Ikeda, "Roles of the dna binding proteins h-ns and stpa in homologous recombination and repair of bleomycin-induced damage in escherichia coli," Genes & genetic systems **82**, 433–439 (2007).

[295] K. A. Datsenko and B. L. Wanner, "One-step inactivation of chromosomal genes in escherichia coli k-12 using pcr products," Proceedings of the National Academy of Sciences **97**, 6640–6645 (2000).

[296] R. Hengge-Aronis *et al.*, "Regulation of gene expression during entry into stationary phase," Escherichia coli and Salmonella: cellular and molecular biology, 2nd ed. ASM Press, Washington, DC pp. 1497–1512 (1996).

[297] D. E. Deatherage and J. E. Barrick, "Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq," in "Engineering and Analyzing Multicellular Systems," (Springer, 2014), pp. 165–188.

[298] D. I. Andersson and D. Hughes, "Gene amplification and adaptive evolution in bacteria," Annual review of genetics **43**, 167–195 (2009).

[299] H. Weber, T. Polen, J. Heuveling, V. F. Wendisch, and R. Hengge, "Genome-wide analysis of the general stress response network in escherichia coli: σs-dependent genes, promoters, and sigma factor selectivity," Journal of Bacteriology **187**, 1591–1603 (2005).

[300] A. Battesti, Y. Tsegaye, D. Packer, N. Majdalani, and S. Gottesman, "H-ns regulation of irad and iram antiadaptors for control of rpos degradation," Journal of bacteriology **194**, 2470–2478 (2012).

[301] K. S. Myers, H. Yan, I. M. Ong, D. Chung, K. Liang, F. Tran, S. Keleş, R. Landick, and P. J. Kiley, "Genome-scale analysis of escherichia coli fnr reveals complex features of transcription factor binding," PLoS genetics **9**, e1003565 (2013).

[302] S. S. Singh and D. C. Grainger, "H-ns can facilitate specific dna-binding by rna polymerase in at-rich gene regulatory regions," PLoS Genet **9**, e1003589 (2013).

[303] R. Srinivasan, D. Chandraprakash, R. Krishnamurthi, P. Singh, V. F. Scolari, S. Krishna, and A. S. N. Seshasayee, "Genomic analysis reveals epistatic silencing of "expensive" genes in escherichia coli k-12," Mol. BioSyst. **9**, 2021–2033 (2013).

[304] H. Li and R. Durbin, "Fast and accurate long-read alignment with burrows–wheeler transform," Bioinformatics **26**, 589–595 (2010).

[305] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin *et al.*, "The sequence alignment/map format and samtools," Bioinformatics **25**, 2078–2079 (2009).

[306] D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, R. K. Wilson, and L. Ding, "Varscan: variant detection in massively parallel sequencing of individual and pooled samples," Bioinformatics **25**, 2283–2285 (2009).

[307] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg *et al.*, "Ultrafast and memory-efficient alignment of short dna sequences to the human genome," Genome Biol **10**, R25 (2009).

[308] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," Statistical applications in genetics and molecular biology **3** (2004).

[309] J. J. Faith, M. E. Driscoll, V. A. Fusaro, E. J. Cosgrove, B. Hayete, F. S. Juhn, S. J. Schneider, and T. S. Gardner, "Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata," Nucleic acids research **36**, D866–D870 (2008).

[310] C. J. Dorman, "Probing bacterial nucleoid structure with optical tweezers." Bioessays **29**, 212–216 (2007).

[311] S. S. Singh, N. Singh, R. P. Bonocora, D. M. Fitzgerald, J. T. Wade, and D. C. Grainger, "Widespread suppression of intragenic transcription initiation by h-ns," Genes & development **28**, 214–219 (2014).

[312] E. P. Rocha, "The replication-related organization of bacterial genomes," Microbiology **150**, 1609–1627 (2004).

[313] T. Ferenci, "What is driving the acquisition of muts and rpos polymorphisms in escherichia coli?" Trends in microbiology **11**, 457–461 (2003).

[314] G. Chen, C. L. Patten, and H. E. Schellhorn, "Positive selection for loss of rpos function in< i> escherichia coli</i>," Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis **554**, 193–203 (2004).

[315] P. Anderson and J. Roth, "Spontaneous tandem genetic duplications in salmonella typhimurium arise by unequal recombination between rrna (rrn) cistrons," Proceedings of the National Academy of Sciences 78, 3113–3117 (1981).

[316] R. V. Sonti and J. R. Roth, "Role of gene duplications in the adaptation of salmonella typhimurium to growth on limiting carbon sources." Genetics 123, 19–28 (1989).

[317] D. S. Straus and G. R. Hoffmann, "Selection for a large genetic duplication in salmonella typhimurium," Genetics 80, 227–237 (1975).

[318] A. H. Yona, Y. S. Manor, R. H. Herbst, G. H. Romano, A. Mitchell, M. Kupiec, Y. Pilpel, and O. Dahan, "Chromosomal duplication is a transient evolutionary solution to stress," Proceedings of the National Academy of Sciences 109, 21010–21015 (2012).

[319] F. Hommais, E. Krin, J.-Y. Coppee, C. Lacroix, E. Yeramian, A. Danchin, and P. Bertin, "Gade (yhie): a novel activator involved in the response to acid environment in escherichia coli," Microbiology 150, 61–72 (2004).

[320] T. Atlung and L. Brøndsted, "Role of the transcriptional activator appy in regulation of the cyx appa operon of escherichia coli by anaerobiosis, phosphate starvation, and growth phase." Journal of bacteriology 176, 5414–5422 (1994).

[321] C. Cagliero, R. S. Grand, M. B. Jones, D. J. Jin, and J. M. O'Sullivan, "Genome conformation capture reveals that the escherichia coli chromosome is organized by replication and transcription." Nucleic Acids Res 41, 6058–6071 (2013).

[322] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," Nature methods 9, 357–359 (2012).

[323] L. Mirny, "The genome folding problem," Annual Review of Biophysics 44 (2014).

[324] M. Nicodemi and A. Pombo, "Models of chromosome structure," Current opinion in cell biology 28, 90–95 (2014).

[325] E. Yaffe and A. Tanay, "Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture," Nature genetics 43, 1059–1065 (2011).

[326] N. J. Kuwada, K. C. Cheveralls, B. Traxler, and P. A. Wiggins, "Mapping the driving forces of chromosome structure and segregation in escherichia coli." Nucleic Acids Res 41, 7370–7377 (2013).