The Physics of DNA and Chromosomes

Davide Marenduzzo describes a selection of topics where physicists have contributed to our current understanding of DNA and chromosomes.

physicsworld Discovery

The Physics of DNA and Chromosomes

The Physics of DNA and Chromosomes

Davide Marenduzzo
School of Physics and Astronomy, University of Edinburgh, UK

IOP Publishing, Bristol, UK

Multimedia content is available at http://iopscience.iop.org/book/978-0-7503-1602-6.

ISBN 978-0-7503-1602-6 (ebook)

DOI 10.1088/978-0-7503-1602-6

Version: 20181001

Physics World Discovery
ISSN 2399-2891 (online)

British Library Cataloguing-in-Publication Data: A catalogue record for this book is available from the British Library.

Published by IOP Publishing, wholly owned by The Institute of Physics, London

IOP Publishing, Temple Circus, Temple Way, Bristol, BS1 6HG, UK

US Office: IOP Publishing, Inc., 190 North Independence Mall West, Suite 601, Philadelphia, PA 19106, USA

For Ile, Cucco, Boogie and Bollie

Contents

Abstract

Each of the tens of trillions of cells making up your body contains about two metres of DNA, which need to fit within the 10 microns container that is its nucleus—roughly a tenth of the diameter of a human hair. How is the DNA arranged in such a tight spot? A liver and a brain cell contain exactly the same genetic material, as they come from the same egg cell, yet they work very differently, because the patterns of genes that are on and off in the two is completely distinct. How is this at all possible? Biophysicists have found general principles that are beginning to answer these and similar questions. In this ebook we explore some of these principles, and describe a selection of topics where physicists have contributed to our current understanding of DNA and chromosomes.

Acknowledgements

I would like to very warmly thank my collaborators Chris A Brackley, Peter R Cook and Davide Michieletto, with whom I have worked on and discussed many of the topics discussed in this book. I also thank Chris and Davide for several of the figures and videos which I presented. Additionally, I thank Jim Allan, Alessandro Bentivoglio, Nick Gilbert, Giuseppe Gonnella, James Johnson, David Keszenman-Pereyra, Alex Morozov and Enzo Orlandini for helpful discussion and collaboration pertaining to some of the topics discussed here. My research on DNA and chromosome biophysics is supported by the European Research Council (ERC consolidator grant THREEDCELLPHYSICS, grant agreement number 648050).

Author biography

Davide Marenduzzo

Davide Marenduzzo is professor of computational biophysics at the School of Physics and Astronomy at the University of Edinburgh, and currently a European Research Council fellow. His current interests include the biophysics of DNA, chromatin and chromosomes, but also cell motility, active matter and liquid crystal physics. Davide has coauthored about 200 articles on these and related topics in peer-reviewed journals, and he is author of a couple of popular physics articles in *Physics World*. Additionally, he coordinates outreach activities at the Higgs Centre for Theoretical Physics. Davide was awarded the 2010 IUPAP Young Scientist Prize in Statistical Physics 'for his contributions to the statistical mechanics of biomolecular and soft matter systems, in particular his groundbreaking simulations of DNA organization, of liquid crystals, and of active gels'.

The Physics of DNA and Chromosomes

1 Introduction

The Dunn School of Physiology is a beautiful building in the north of Oxford. It looks a bit like a large British townhouse, or perhaps a comfortable bed-and-breakfast, and sits just beside the University Parks, a lovely green area full of enthusiastic youngsters looking to play a game of cricket, rugby or football (typically, in decreasing order of ability). Nowadays, the school's research is flourishing, however, its heyday was during the Second World War, when Florey, the Australian head of department at the time, together with his collaborators, understood how penicillin could be used as an antibiotic (yes, that's right—it was not Fleming, who first found the chemical in his lab!).

Back in the spring of 2004, one of the current members of the school, a cell biologist working on chromosome transcription, was scratching his head over some hard-core polymer physics paper. Although this might seem an unlikely occupation for a biologist, it turns out that for some reasons many more life scientists have a bent for maths and physics than you might expect (an example is the founder of the Dunn School, a pathologist who was also a keen mathematician). The puzzled biologist was Peter Cook, who is now well-known for his pioneering work in the 1990s about transcription factories—we shall discuss these in some detail in section 3. Not managing to get through the algebra and physics in the paper quickly enough for his liking, Cook decided to look for help. Browsing a few related physics papers, one caught his eye. A co-author was employed as a postdoc in Keble Road at the time, less than half a mile from him. Cook e-mailed him to ask if they could have a chat about some aspects of his research, where he suspected a bit of physics might have helped.

I was that physics postdoc, and that meeting with Peter was one of the most scientifically influential I've had! I had already made a big U-turn in my career when I abandoned particle physics (I worked on magnetic monopoles and Dirac strings for my MPhys thesis) to do a PhD and postdoc in soft condensed matter and

doi:10.1088/978-0-7503-1602-6ch1

polymer physics. However, my involvement in biophysics was still relatively superficial at that time. After our chat that spring of several years ago, this all changed in a big way. I had stumbled on a problem which I still love, the biophysics of mammalian chromosomes, for which I have had to learn a lot more biology than I previously imagined I would ever need. Just as importantly, I had started a great collaboration, as Peter and I have stuck together ever since. In case you are curious, the chat we had was about a physical mechanism for the emergence of transcription factories: it took us some 10 years to find a satisfactory answer. This, we believe, is the 'bridging-induced attraction' discussed in section 3, although I should warn you that not everyone would agree with us that the problem is solved! In passing, keep in mind this is true of many of the things I will write: while I have strived to provide a balanced view, many of these topics are relatively new research-wise and some are still pretty controversial.

2 Background

A few definitions, the double helix, and chromatin

To begin with, we should introduce some of the words and concepts that we'll be using time and again in what follows. This is the aim of this first section—it can therefore be used as a glossary to go back to when needed.

DNA is the genetic material that is inside all living organisms: it contains all the information to make us into what we are. DNA is a very long, double-stranded polymer, where each strand contains a sequence of nucleotides. There are three components in a nucleotide: sugar, phosphate and base. Whilst the sugar and the phosphate are the same for each nucleotide, there are four possible bases in the DNA alphabet: cytosine (C), guanine (G), adenine (A) and thymine (T). The two strands in the molecule are bound by hydrogen bonds between the bases: these form between C and G, or between A and T. The two strands are therefore 'complementary': if you know the sequence of one strand, you can figure out the sequence of the other, as the bases will be those which form hydrogen bonds with the bases in the first strand. Each pair of bases—one in each strand—is called a 'base pair', often denoted 'bp'. Distances along the DNA molecules are measured by counting how many base pairs we must travel to go from one point to another: typical distances we shall use are kilobase-pairs or Mega-basepairs (kbp or Mbp, respectively).

Now, time for some geometry! DNA is a narrow polymer, with a thickness of about 2 nm, similar to the width of a base pair. The physiological form of DNA within our bodies is called 'B-DNA', the iconic right-handed double helix first discovered by Watson, Crick and Franklin (figure 1(*a*)). The story goes that Rosalind Franklin isolated two types of DNA fibres from natural sources: B-DNA, when she kept the fibres wet, and A-DNA when she kept them dry (figure 1(*b*)). James Watson and Francis Crick interpreted the B-form, the simpler of the two, in their landmark 1953 paper on DNA structure.

Why DNA forms a double helix is beautifully illustrated in the book *Understanding DNA* by Chris Calladine and Horace Drew (the later editions are also co-authored by Ben Luisi and Andrew Travers). The key issue is that, while

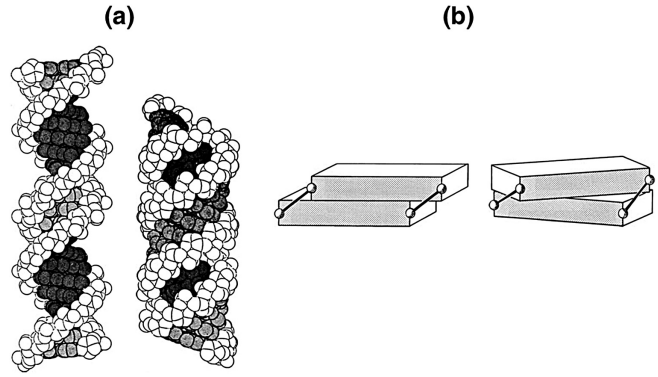


Figure 1. DNA structures. (*a*) Structural representation of B-DNA (left) and A-DNA (right). (*b*) A simple representation of base pairs as parallelepipeds and phosphates as spheres. As the distance between neighbouring phosphates is fixed, they cannot stack exactly on top of each other. In order to tuck away as much surface of the hydrophobic bases as possible, the bases may arrange in a skewed latter (left), or helical (right) fashion: in 3D the most favourable configuration is the latter, and this is the reason why DNA is a double helix. Reproduced from Calladine C R, Drew H R, Luisi B F and Travers A A 2004 *Understanding DNA: The Molecule and How it Works* (London: Elsevier/Academic Press).

phosphate and sugars are soluble in water, the bases are not (a bit like fat). Our cells and nuclei are full of water; hence the base pairs try to bury as much of their surface area as possible in regions unreachable by water molecules. A useful geometric model for a base pair is a brick of width 2 nm (the diameter of DNA), depth about 1 nm, and height 0.34 nm. The phosphates may be considered little spheres on the side of each brick (figure 1(*b*)), separated by 0.6 nm. Because the phosphates are tethered, it is not possible to change this distance much. As a result, it is not possible to stack the bases on top of each other to bury their widest face, because the phosphates will push the bases apart, allowing water in. One possible solution is the skewed ladder structure (figure 1(*b*) left); however, Nature favours the helical structure in the right panel of figure 1, as this leads to a little more shielding of the fatty bases from water.

A few more numbers and definitions are now required. While double-stranded DNA is the genetic material of most living organisms, it comes in different sizes and shapes. In prokaryotes such as bacteria, DNA exists as a closed loop of a few Mbp. Eukaryotes confine their DNA inside a nucleus, which is separated by the rest of the cell (the cytoplasm) by a nuclear membrane. Eukaryotic organisms range from bakers' yeast, through to worms, flies and insects, all the way to mammals such as us. Usually there are several strands of DNA in eukaryotic nuclei. For instance, most human cells are diploid, meaning they contain 23 pairs of chromosomes (one of each pair from the mother, one from the father). Each chromosome is a linear piece of DNA, normally much longer than in bacteria. In humans, chromosomes are typically about 100 Mbp.

In eukaryotes, DNA is always associated with proteins, called histones (most are octamers, i.e., complexes of eight proteins) to form the 'chromatin fibre' that makes

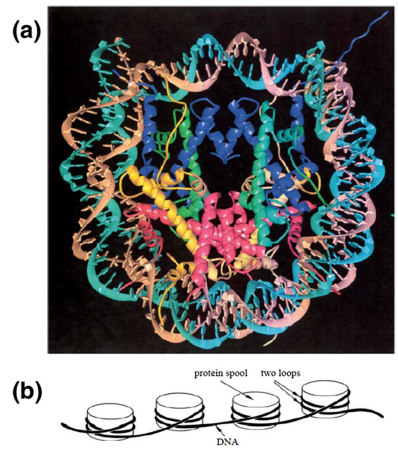


Figure 2. Chromatin. (*a*) Crystal structure of the nucleosome, showing the DNA and histone proteins. Colour-coding of proteins: H2A, orange; H2B, red; H3, blue and H4, green. Two copies each of histones H2A, H2B, H3 and H4 self-assemble into a histone octamer. (*b*) A chromatin fibre is made up of many nucleosomes, each with a left-handed DNA wrapping. There are many possible structures for a chromatin fibre, the one pictorially shown here corresponds to open chromatin, presumably corresponding to active chromatin: it is also referred to as 10 nm fibre, as its thickness coincides with that of a histone octamer. Inactive chromatin regions are thought to fold up into more compact structures. Reproduced from Calladine C R, Drew H R, Luisi B F and Travers A A 2004 *Understanding DNA: The Molecule and How it Works* (London: Elsevier/Academic Press).

up our chromosomes. The basic unit of chromatin is the nucleosome core particle (figure 2(*a*)). A nucleosome core particle normally consists of 146 base pairs of double-helical DNA tightly wrapped around a histone octamer for just under two turns. The wrapping is itself helical: intriguingly, this helix is left-handed (with very few exceptions) as opposed to the right-handed helix of DNA itself. The interaction between histone octamers and DNA is due to simple electrostatics, as histone octamers are highly positively charged, whereas the DNA is negatively charged due to the presence of phosphates. It is not completely clear why DNA exists as chromatin in eukaryotes, but not in bacteria as the evolutionary advantage is at first sight unclear. One possible reason is that tightly wrapped DNA can fit into the nucleus more easily (although several other layers of compaction are needed, see the section 'Basics of DNA organisation' and section 3). Additionally, the wrapping probably mechanically protects DNA, but also renders our genes much less active than say, those of bacteria, which might be a good thing for more complex organisms allowing more precise control over gene expression.

Successive nucleosome core particles are joined by small stretches called linker DNA, whose size varies (a possible arrangement of the resulting chromatin fibre is sketched in figure 2(*b*)). In mammals, linker DNA is typically about 50 base pairs, while in yeast it is much shorter (about 20 base pairs or even a bit less). Slightly puzzlingly, although nucleosome core particles are more tightly packed, the genome of yeast is more active than that of mammals. In all cases, the linker DNA is small

enough to look stiff (its length is smaller than the persistence length, which we discuss in more detail below). The physics of chromatin is discussed in an excellent review article by Helmut Schiessel in *Journal of Physics: Condensed Matter*, which is recommended for the reader wanting to know more on this topic.

Finally, a couple more facts about the way the information stored in DNA is retrieved by the cellular machinery will help us in what follows. First, the 'central dogma' of biology states that DNA makes RNA, and that RNA makes proteins, the molecular machines doing most of the work required inside a cell. The former process is known as transcription, which we will focus on within this book; the latter is called translation. DNA transcription is performed by a molecular motor called RNA polymerase. An RNA polymerase typically transcribes about 100 bp s^{-1} in bacteria; whereas in humans, nucleosomes create a barrier to its progress, and the rate is about 25 bp s^{-1}, or about 1 kbp min^{-1}. Not all genes in DNA molecules are turned on: the pattern of active genes changes from cell to cell (an eye, a liver and a white blood cell all have very different sets of active genes), and over time (when we catch the flu, a specific set of genes is switched on to deal with the inflammation). The pattern of gene activation depends on a complex interplay between gene position, helper proteins (called transcription factors), and molecular marks on DNA and histones which are 'on top' of the genetic code and facilitate or hinder transcription. A second process which we will touch upon is DNA replication. Before the cell divides, its DNA must be replicated (so that just prior to mitosis there are 92 chromosomes in the nucleus). Replication is performed by another polymerase, known as DNA polymerase—this is a bit faster that its RNA cousin, and moves at up to 5 kbp min^{-1} in mammals (1000 bp s^{-1} in bacteria).

The linking number

As DNA is a double helix, we can ask what its 'pitch' (the height of one complete turn) is. In B-DNA, this is about 10.5 base pairs, or 3.6 nm. The linking number is another important factor for describing the 3D structure of a DNA molecule, which we describe in this section. We shall make use of this concept in section 3, when we discuss the physics of DNA supercoiling.

The linking number is defined for any two curves, either closed, or kept fixed at their extremities: it measures the number of times that the two curves wind around each other. We can also define a linking number for a ribbon, by considering the winding of its two edges: in this case, this quantity is simply equal to the number of times that we twist the ribbon. If we close the ribbon, we need to make a whole number of 360 degree turns, so that the top edge on one end of the ribbon matches the top edge on the other end (video 1).

An important feature of the linking number of a closed ribbon, such as a rubber band, is that is remains constant over time, whereas this is not true for a linear ribbon (such as those used for wrapping presents). Why? Firstly, consider the open ribbon. If you twist up the ends, you can clearly change its linking number at will. Once you seal the two ends, though, for instance, when you glue the ends of a paper ribbon together, then the linking number is trapped, there is no way to change it.

Video 1. This video shows a simple experiment demonstrating the concept of linking number, by using a paper ribbon, glue, and a pair of scissors. Available at http://iopscience.iop.org/book/978-0-7503-1602-6.

Because the two strands of a DNA molecule in its relaxed state are kept close and equidistant by base pairing through hydrogen bonding, we can represent them with the two edges of a ribbon of a given width, and we will use this analogy as follows. A DNA-like ribbon is highly twisted. As the double helix of B-DNA makes a full turn every 10.5 base pairs (see above), the linking number of a B-DNA loop (called a plasmid) of 1050 base pairs is 100, whereas the linking number of a typical human chromosome is about 10 million!

Two closed curves with a linking number of n are linked at n points. You can convince yourself of this with a simple but neat experiment (video 1). First, take two ribbons, colouring them on one side to be able to see the twist clearly. Then, take the two ends of the first ribbon and glue them together, without introducing any twist. After this, take the second ribbon, and now twist it by 360 degrees before gluing the ends, taking care that the coloured faces match each other. Now take some scissors, and cut both ribbons through the middle, along their long axes. The untwisted ribbon is easily cleaved to yield two unlinked smaller ribbons. The case of the twisted ribbon is not as trivial: the two resulting halves are linked at one point, forming what is known in maths as a Hopf link (video 1). By the way, if you liked this experiment, you may try and see what happens if the ribbon instead has half or one and a half twist: the results are quite interesting!

In this simple experiment, if we think of the ribbon as a DNA molecule, the scissors represent the helicase protein, which cuts the two strands of DNA during replication. Being linked is no good for the newly replicated molecules: they are desperate to go their separate ways to give rise to two different cells. Indeed, if the link is not undone quickly, the cell will die (more accurately, commit suicide, called

'apoptosis') instead of replicating successfully. This topological danger lurking behind DNA replication is the reason why cells contain a vast amount of 'topological enzymes', known collectively as topoisomerases. Type II topoisomerase is a molecular machine capable of cutting and then rejoining double-stranded DNA, to undo a knot or a link. The DNA of most bacteria is a single giant loop, hence our model experiment simulates directly what happens to it during replication. Therefore, a type II topoisomerase has to act at each of the ~400 000 linking points generated at each round of replication of an *Escherichia coli* bug, which occurs approximately once every 30 min in rapidly growing bacteria in the lab. Although human chromosomes are linear, they are not safe from the linking trap either. Indeed, most eukaryotic chromosomes contain an extensive number of genomic loops, which are held together by DNA-binding proteins (more on this in section 3), and these loops are often present during replication, leading to entanglement if topoisomerase II is inactive.

Quite remarkably, the linking trap is exploited by many anticancer drugs, which are based on different molecular pathways to inhibit the action of topoisomerases. The idea is that cancer cells divide faster than others, hence they encounter the linking problem more often. Because inhibiting topoisomerase leads to linking and cell death, these drugs are statistically more likely to kill cancer cells, which are more likely to be dividing at any given moment. This is also the reason why standard anticancer drugs have unpleasant side effects such as hair loss: they target all rapidly dividing cells—hence hair cells are destroyed, as well as cancer cells.

Basics of genome organisation

To understand how daunting the problem of genome organisation within the cell is, we can start from a simple back-of-the-envelope estimate. The human genome of a diploid nucleus contains about 6 billion base pairs in total. We know that a base pair in B-DNA is about 0.34 nm in size, therefore, if we were to stretch out all the DNA inside a single cell in our body, it would be about 2 m long, yet it must fit within a nucleus, which is only about 10 microns across. To capture the enormity of the problem, try and calculate the whole amount of DNA in meters that each of our bodies contain. This is a truly astronomical amount: it would be enough to reach Pluto from the Earth! The reason why such an incredible amount of DNA can be stored in cells, or organisms, is that it is so thin. Even so, a sizable fraction of our mass is DNA, and packaging it all is a highly nontrivial problem for polymer physics.

To discuss how living organisms solve this problem, let us start with the simplest lifeforms. Bacteriophages, or phages, are viruses which attack and kill bacteria—they are, therefore, in principle useful for eukaryotes like us, and might even provide a viable, if a touch scary, alternative to antibiotics should we run out of them in the future (unlike drugs, these viruses evolve with the bacteria, so can keep killing them as they mutate). Phages are essentially spheres (the capsid) packed with a polymer, double-stranded DNA, at almost crystalline densities. Genome organisation does not get simpler than this—nevertheless, it took a long time to identify the structure of DNA within the capsid!

One of the issues that makes the problem non-trivial is a competition between scales. One is the size of the capsid, which for most phages is about 50–100 nm. Another is the persistence length of DNA, which is the length over which the DNA bends due to thermal fluctuations. The persistence length is also about 50 nm, so fluctuations are significant at this scale, which muddles things up a bit. The current consensus is that electrostatic and steric DNA–DNA interactions are the key players at these densities, and lead to a spool-like arrangement of the DNA, a bit like a sailor's anchor neatly stowed away. If you are interested in the physics of phages, Charles Knobler and Bill Gelbart wrote a neat popular physics article in *Physics Today* in 2008, which is well worth reading.

Understanding the basics of DNA arrangement within bacteriophages is useful, especially because it shows that simple polymer physics models and ideas are very useful in the field. However, genome organisation in bacteria and eukaryotes is much more complicated! This is because the physical parameters differ, and because there are proteins associated with such genomes, affecting their physics. We shall discuss the importance of proteins for genome organisation in eukaryotes in section 3. Here, we review some basic polymer physics features of the organisation, which are worth keeping in mind.

In a bacterium such as *Escherichia coli*, there are about 4 million base pairs (which would be about 1 mm if stretched out) contained in a cell size of about 1–2 microns. A simple calculation shows that the density is much lower than in the phage, as the DNA itself makes up only about 1% of the total cell volume. bacterial DNA also differs from phage DNA in that it is constantly slightly unwound by some of the bug's proteins —as a result the DNA is 'supercoiled'. We shall see a few consequences of this in section 3.

In humans (or other eukaryotes), DNA exists as chromatin. The volume occupied by our chromosomes is about 10%–20% of the volume of the nucleus where they are confined. While the problem of genome organisation is a very challenging one, we do know a few things about chromosome morphology thanks to microscopy (actually, these facts were known long before Watson and Crick!). Firstly, chromosomes condense and become individually visible under a microscope during mitosis—when the cell divides—forming the well-known X-shaped cylindrical structures shown in biology textbooks. When the cell is not dividing (during 'interphase') chromosomes are decondensed, and much more similar to swollen polymers.

We also know that during interphase chromosomes form 'territories'. By 'painting' each of them a different colour, we can see that each chromosome occupies its own volume, and there is little intermingling between different chromosomes. The physical basis of territory formation was studied by Angelo Rosa and Ralf Everaers a few years back. They modelled human chromosomes as a melt of polymers interacting only via volume exclusion. Under such assumptions, after the mitotic cylinders decondense at the onset of interphase, if we could wait long enough all chromosomes would intermingle (because this is the equilibrium state of a melt of linear polymers). However, due to sheer size, the dynamics of human chromosomes are so sluggish that they remain separate forever in practice, as calculations suggest that the intermingling time is several decades! Rosa and

Everaers' simulations also suggest that within simpler eukaryotes such as yeast, the intermingling is much quicker, and territories should not be observed there, in good agreement with experimental observations.

3 Current directions

All twisted up: supercoiling, transcription, and nucleosome assembly

Coiled phone cords and supercoiling

DNA supercoiling is a fascinating phenomenon that was first discovered in the 1960s by scientists studying some puzzling behaviour of tumour virus DNA. While it originated as an abstract mathematical concept from the field of topology, supercoiling has several far-reaching and striking consequences for the biophysics of the cell. But what is supercoiling in practice? To visualise it, think of one of those old telephone sets from figure 3(*a*)—the one shown is that of my office in Edinburgh. After several phone calls, I find that the cord invariably gets 'all twisted up' (coiling up as in figure 3(*a*)): why this happens will become clear in what follows. While the cord is entangled, there are usually no knots. Therefore, I can get rid of the tangles simply by lifting the phone in the air, and letting the headset hang so that it can freely swivel around its axis to unwind the coils. The state of the tangled phone cord resembles supercoiled DNA, and is not unlike the coiled structure attained by the genomes of many bacteria.

The linking number theorem

To understand more about what supercoiling is and what its important consequences are, we need to make a little detour, and go back to the definition of linking number (see section 2). Recall that the linking number for a ribbon (which is topologically equivalent to a DNA molecule) measures the number of times that the edges of the ribbon wind around each other.

Although the linking number of a closed DNA molecule, or ribbon, is conserved, there are many possible conformations that are compatible with this constraint. It is

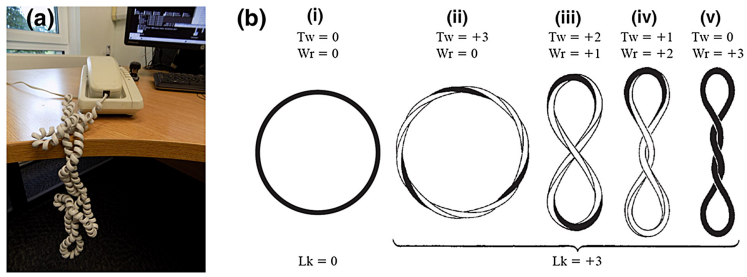


Figure 3. Telephone cords and rubber tubes. (*a*) A coiled-up phone cord. The structure resembles that of supercoiled DNA. As an exercise, you can try and determine the handedness of the writhe in the coils (see text). (*b*) Some structures of a rubber tube (see text) with linking number equal to 0 (i) or +3 (ii–v). Panel (*b*) is reproduced from Calladine C R, Drew H R, Luisi B F and Travers A A 2004 *Understanding DNA: The Molecule and How it Works* (London: Elsevier/Academic Press).

useful to describe such conformations via two geometric properties: the twist (Tw), and the writhe (Wr). These are demonstrated in figure 3(b) for the case of a rubber rod with square cross section (topologically identical to a ribbon). One of the faces of the rod has been coloured black in the figure to help us trace its twist. If you run through the loops in figure 3(b), the twist counts how many times the colour switches, for instance, from black to white (but not white to black): more precisely, it measures the winding of the rubber rod around its centreline. The writhe, instead, counts the number of times that the rod crosses over itself in 3D. Therefore, there are one, two and three units of writhe in the three rightmost configurations of the rod in figure 3(b).

Both twist and writhe have a sign. The convention is that a right-handed twist is positive. The sign convention for the writhe requires orienting the loop (which way does not matter) and analysing the rotation required to superimpose the top segment onto the bottom one (where the shortest angle counts). If the rotation is anticlockwise, then the writhe is positive, or right-handed; if it is clockwise, the writhe is negative and corresponds to a left-handed crossing. All the writhe-related crossings in figure 3(b) are right-handed, or positive.

An important result, known as White's theorem, connects the linking number (or changes thereof) to the sum of the twist and the writhe. White's theorem states simply that the linking number is the sum of the twist and the writhe,

$$\mathrm{L}k \;=\; \mathrm{T}w + \mathrm{W}r. \tag{1}$$

While the proof requires some sophisticated maths, we can test this relationship in practice. For instance, consider the four rightmost configurations in figure 3(b). These all have linking number +3. The configuration in (ii) is obtained by holding the tube in a plane—it has no crossing, hence zero writhe, whereas there are three right-handed units of twist (as the pattern goes black–white–black–white–black–white, i.e., black–white three times). Rubber tubes and paper ribbons, like DNA, do not like having to twist, hence this configuration is not energetically stable, and the tube writhes. The three configurations in (iii, iv, v) all have non-zero writhe, however the sum of the twist plus the writhe is always +3, which is the linking number. As we know, this is conserved in a closed tube (see section 2). The linking number theorem (1) is central for the physics of DNA supercoiling.

As a little exercise in applying this bit of maths, have another look at the phone cord in figure 3(a). The coiling is all down to writhe, but can you identify the sign of the crossing? If you are truly ambitious, you may try and guess if the phone-cord owner is right-handed or left-handed, and how many phone calls he or she has received since last getting rid of the tangles in the cord. (Hint: you will need a couple more assumptions on the room layout, but after that you can answer the question!)

Supercoiling, gel electrophoresis and knotting

After this little detour, we are ready for a more precise definition of DNA supercoiling. In B-DNA in its relaxed form, we have seen (see section 2) that the two strands of the genome wind around each other once every 10.5 base pairs, and that the resulting double helix is right-handed. Twisting DNA away from this relaxed state introduces supercoiling. If we over-twist the double helix, so that the

DNA strands wind around each other more tightly, we say that there is positive supercoiling; if the DNA unwinds, we refer to negative supercoiling.

Consider again the example of our DNA loop of 1050 base pairs used earlier on. If it is made of ideal B-DNA, in thermodynamic equilibrium the linking number would be 100 (1 for each 10.5 base pairs). Imagine now adding a right-handed turn to the helix, so that the linking number is 101: we say that the DNA is positively supercoiled. If, on the other hand, the DNA helix is under-twisted, we say it is negatively supercoiled. The supercoiling level is defined as the ratio between the excess or deficit in Lk and the equilibrium value of Lk—therefore the plasmid with 101 helical turn has a supercoiling of +0.01. Because twist is generally more expensive (free-energy-wise) than bend for a DNA molecule, positive or negative supercoiling normally results in writhe, or in more writhe than twist—at least for naked DNA not associated with proteins.

The writhing is useful, because it gives supercoiled DNA the characteristic shape of coiled-up phone cords, which can be spotted by experimentalists in electron microscopy images. Remarkably, it is also possible to detect DNA supercoiling even without imaging. How can this be achieved? The idea is rather simple, and it exploits writhing once again. Because DNA molecules are negatively charged, we can use an electric field to drive them through a medium. It turns out that agarose gel is a good medium, a jelly-like material filled with microscopic holes (a bit like a squishy bath sponge). The experiment also works best if the field is weak, to avoid effects associated with nonlinear response. The DNA molecules will pass through the gel more quickly if they are smaller, as they are more likely to squeeze through the pores constituting the gel. Therefore, highly writhed supercoiled DNA loops will move faster than loops with no writhe, and can be separated quite easily in practice.

While gel electrophoresis is a simple and neat idea to detect supercoiling, you may worry at this point that positively and negatively supercoiled DNA may have a similar writhe, either right- or left-handed, hence they would be more or less the same size and move at the same speed through the gel. Does this mean we cannot use electrophoresis to tell the handedness of supercoiled DNA? Actually, no. There is a clever trick that the experiments play to obtain this information: it relies on the use of DNA intercalators. An intercalator is a small molecule that can be inserted into the DNA double helix and change its preferred twist, or helical pitch. For instance, the dye ethidium bromide, EtBr, reduces the equilibrium twist between neighbouring base pairs from just over 34 degrees, corresponding to the canonical 10.5 bp/turn of B-DNA, to about 26. In other words, when EtBr is used in the experiment, the DNA molecule untwists. As the total linking number of the two strands needs to stay the same, the DNA molecule must gain some positive writhe, to compensate for the loss of twist. Therefore, the intercalator breaks the symmetry between positive and negative supercoiling, and the two can now be distinguished via gel electrophoresis.

Gel electrophoresis experiments are also useful for detecting DNA knots. The principle is the same: more complicated knots tend to be smaller than simpler knots with the same overall DNA length, hence they move faster under a weak electric field. The relationship between knotting and supercoiling is pretty interesting. Both

may be viewed as superficially similar DNA entanglements, but the cell has very distinct ways of dealing with each. DNA knots are pretty rare in cells, from bacteria to human. This is presumably because DNA knotting interferes with transcription, replication, and normal DNA function. Topoisomerases need to get rid of these tangles pretty quickly if they arise in a living cell. Supercoiling, on the other hand, is found routinely. Not only is it tolerated, it appears to be sought after: for instance, there is a topological enzyme in bacteria, called gyrase, whose main function is to introduce negative supercoiling in the bug's DNA. Negative supercoiling may be beneficial because it is associated with a local untwisting of the double helix. The double helix needs to be opened by several essential protein machines, such as the polymerases which transcribe or replicate DNA. Therefore, these machines have a bit of their work done if the DNA is slightly unwound to begin with.

Research has shown that there are also further, subtler relationships between knotting and supercoiling. Andrzej Stasiak of the University of Lausanne and his collaborators found that supercoiled DNA loops gets knotted less frequently than torsionally relaxed DNA with no writhe. These results, found via computer simulations, are at first sight surprising, as supercoiling leads to writhing and one may imagine that writhed configurations may be more likely to be knotted with respect to un-writhed loops. Still, the findings were in pleasing agreement with experiments reporting that bacterial cells without gyrase (the enzymes that creates supercoiling) exhibit much more knotting that wild-type cells with gyrase.

But supercoiling may provide even more help avoiding knots. Research again led by Stasiak showed that a DNA knot is subtly different in the presence or absence of supercoiling. With no supercoiling, the knot is often diffuse, and spread over a large portion of the molecule: a topoisomerase needing to find this quickly to undo it will thus have a hard time. With supercoiling, the knot localises dramatically—a bit like the knot in a rope when we pull it taut. The localisation would allow topoisomerase to act much more efficiently, as the knot can be found simply by sensing the local DNA density which is greater close to the knot.

The physics of DNA knots is a fascinating topic in itself; here we do not dwell on it further, as our concern is only its relation to DNA supercoiling. Readers interested in knowing more about knotty DNAs and their physics may wish to read the feature article entitled 'The knotted strands of life', published in the April 2013 issue of *Physics World*.

Supercoiling and transcription

Gene transcription is an example from biological physics where supercoiling is important. Transcription is the process through which a protein, RNA polymerase, opens the DNA helix to read its base pair sequence and transcribes it into RNA (see section 2).

It turns out that there are far-reaching topological (and geometric) consequences of transcription. The RNA polymerases must first open the DNA, and then move along it to read the DNA sequence. If rotation of the RNA polymerase and its associated transcription machinery is hindered, as is likely the case in the very crowded intracellular environment, then gene transcription inevitably leads to the

creation of positive supercoiling ahead of the tracking polymerase, and negative supercoiling in its wake (figure 4). This is the basis of the 'twin supercoiling domain' model, which refers to the formation of twin domains in the DNA where the supercoiling is of opposite signs. The simplified model in figure 4 has the polymerase tracking, whereas some scientists (among these Peter Cook in Oxford) have shown a great deal of evidence that in some cases at least the polymerase is fixed and reels in the DNA—this does not matter, the twin supercoiled domains would form anyway provided there is relative motion.

The extent of supercoiling generated by transcription is quite extraordinary: for every 10 bp or so that are transcribed, the linking number changes by a maximum of +1 units ahead of the polymerase and −1 behind. An average gene in bacteria is 1000 base pairs, so there might be as many as 100 units of positive and negative linking number created for each transcription event: mammalian genes are about 10-fold larger, so the values there may be much greater.

What happens to all the supercoiling generated by transcription? It turns out that this is a pretty interesting physics question! When the gene is 'off', supercoiling can redistribute over the whole DNA molecule, dissipating via diffusion. Just as in a gas

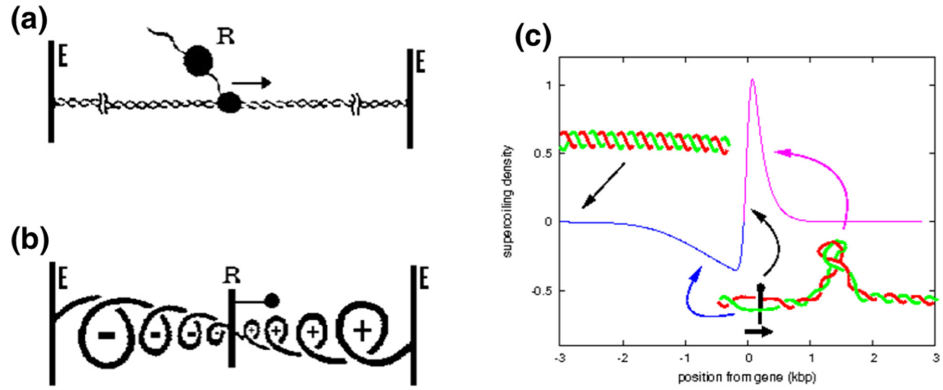


Figure 4. The twin supercoiling domain model. (*a*) and (*b*) A cartoon of the twin supercoiling domain model. In (*a*), the diagram shows the transcriptional machinery of the cell (R) tracking along linear double-helical DNA. The E bars at the ends of the DNA represent fixed points at which the DNA/chromosome is stuck, due to interaction with other proteins or the cell membrane. The same conclusions are, however, reached if E represent long DNA tails with high viscosity, which would similarly resist the natural relaxation of supercoiling. As the transcriptional assembly travels along the DNA, it must unwind the helix; otherwise the transcriptional machinery would need to travel around the DNA as it progressed, which would tangle the transcript around the DNA. The unwinding generates positive supercoils in front of the transcription site and negative supercoils behind it (see (b)). The plot in (*c*) shows how the supercoiling density may vary as a function of distance from the promoter (the values come from a stochastic model of supercoiling in transcription proposed recently within our group). Panels (*a*) and (*b*) are reproduced from Liu L F and Wang J C 1987 Supercoiling of the DNA template during transcription *Proc. Natl. Acad. Sci. USA* **84** 7024. Copyright (1987) National Academy of Sciences. Panel (*c*) is reproduced from Brackley C A, Johnson J, Bentivoglio A, Gilbert N, Gonnella G and Marenduzzo D 2016 Stochastic model of supercoiling-dependent transcription *Phys. Rev. Lett.* **117** 011101.

there is a flux of particles from a denser to a less dense region, in a DNA molecule there is a flux of supercoiling from a supercoiled region to a torsionally relaxed one. If we could wait long enough, we would end up with a DNA molecule where the supercoiling is either zero, or back to the value before the gene was transcribed. The typical time, t_D, needed for supercoiling to diffuse over a length comparable to the size λ of a gene will be given by

$$t_D = \lambda^2/D \tag{2}$$

where D is the diffusion coefficient for supercoiling measured in $\mathrm{bp}^2\ \mathrm{s}^{-1}$, and ignoring possible numerical prefactors in (2). The key question is then whether the supercoiling diffuses fast enough, before the next transcription event begins. If it does not, then an intriguing positive feedback may occur between supercoiling and transcription. First, transcription generates negative supercoiling in its wake. This region includes the promoter (start) of the gene, which is where the polymerase needs to bind to initiate transcription. But, as we previously mentioned, negative supercoiling leaves the DNA slightly unwound, and the polymerase will be more likely to initiate transcription again, creating further negative supercoiling, so the cycle can repeat and reinforce itself. This possibility is exciting for physicists who study dynamic systems, as they know that positive feedback can lead to interesting and highly non-trivial results.

To understand whether supercoil diffusion is slow enough to allow this type of feedback, we need to compare the diffusive timescale in (2) with the typical lag time between two transcription events—let us call it t_l. It is common to measure the transcription rate for a given gene, k_t, in number of transcripts per second—this gives the lagtime between two transcriptions as $t_l = 1/k_t$. The ratio between t_D and t_l gives a dimensionless number, which provides the key to the answer to our question. If t_D/t_l is large, this means that diffusion is slower than transcription, and the feedback we previously highlighted can be set up. This ratio is

$$\theta = t_D/t_t = k_t\lambda^2/D, \tag{3}$$

and it turns out that this is also a good approximation for the average degree of negative supercoiling at the promoter. For bacterial DNA, genes are about 1 kbp long, and many have transcription rates of about one per minute. While D is much harder to estimate, experiments suggested that D can be as low as 0.1 $\mathrm{kbp}^2\ \mathrm{s}^{-1}$ for naked DNA in the lab. Combining these numbers we get supercoiling at the promoter of about −0.2 (two fewer helical windings every 10). This is a large value, and experiments suggest that it is enough to affect polymerase binding. A similar conclusion is reached by considering human genes, which are about 10 kbp on average, and are transcribed less often (once per hour on average), by assuming a similar value for supercoil diffusion—this leads to a supercoiling of −0.3 at the promoter. However, in human cells DNA is wrapped around histones to form chromatin (see section 2), so the value of D may well be different from that of naked DNA.

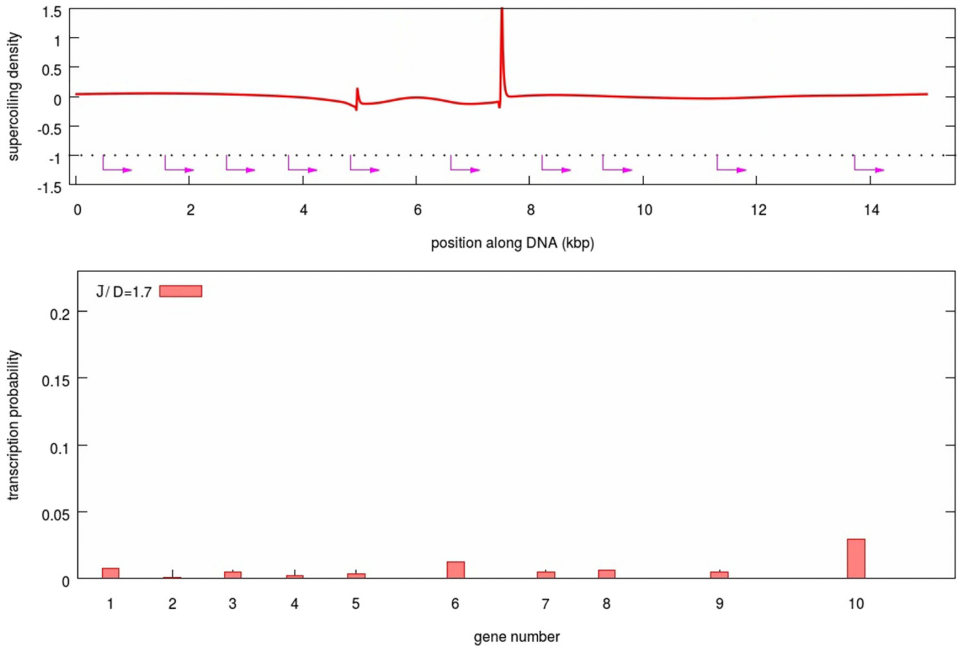
The positive feedback between supercoiling and transcription may therefore be strong enough to lead to detectable effects. In 2016, in a collaboration between

Edinburgh University and the University of Bari, we proposed a simple stochastic dynamic model to investigate such effects. We found that the model may explain quite a few previous experimental observations of transcription dynamics. First, the feedback can work with as few as one gene: in other words, a gene can exploit the supercoiling generated by its transcription to upregulate its own expression! This phenomenon may explain the observation of transcriptional 'bursts', where some genes are typically inactive, but at times get transcribed much more frequently. To see how supercoiling can yield bursting transcription, imagine the binding rate of polymerase to a gene is low. Imagine that at some time, by chance, the gene fires (i.e., gets transcribed): the feedback now promotes further firing, so that the gene is now more active. If, however, a fluctuation prevents polymerase from firing in quick succession, supercoiling diffuses away and we are back to the quiescent state. This results in bursts of activity separated by inactive spells. Second, the model predicts that gene orientation should affect transcriptional rates, hence expression. Suppose that two genes are transcribed in opposite, convergent, directions. The supercoiling produced by either gene will switch off the other one, leading to negative feedback of one gene on the neighbour. Imagine instead the genes are divergent: now the supercoiling produced by one gene will stimulate transcription of the other. Therefore, the feedback we have identified boosts the expression of divergent genes. In line with these predictions, experiments show that activating one gene in bacteria facilitates the expression of its neighbour if it is pointing the other way. Even more interestingly, human DNA is full of 'divergent' genes, much more than would be expected if gene orientation were random: one may speculate that this is to take advantage of the supercoiling-transcription feedback.

The model also correctly reproduces the effect of topoisomerases on transcription. There is at least one intriguing prediction that has no current experimental counterpart. Imagine an arrangement of genes with the same orientation on a DNA loop. Let us consider any of these genes, say gene A. When gene A is transcribed, say to the right, it sets up a flux of negative supercoiling towards gene B, its left neighbour, and a flux of positive supercoiling towards C, its right neighbour. Therefore, gene A activates B and represses C. Later on, transcription of gene B will activate its left neighbour, and repress its right neighbour, gene A. This will go on indefinitely if the genes are all parallel (and the supercoiling-transcription feedback strong enough), so that a transcription wave is travelling anticlockwise on the loop (video 2). While it is unknown whether such waves occur *in vivo*, it may now be possible to build a genetic circuit to recreate one such pattern in the lab.

Supercoiling domains and chromatin self-assembly

Within our cells, DNA wraps around histone proteins to form the chromatin fibre. As there are about 1.75 turns of DNA per nucleosome and the path of the DNA around a histone is a left-handed helix, there are 1.75 units of negative writhe per nucleosome. Even without any under or over-twist, this wrapping introduces negative supercoiling in human chromosomes. The deficit in linking number has been found to be about one for each nucleosome, not 1.75. This result was puzzling for some time (and became known as the 'linking number paradox'), until it was



Video 2. This video shows the result from a simulation of the coupled dynamics of supercoiling and transcription in a DNA loop. The top panel shows the time evolution of the supercoiling density as a function of the position along the DNA loop (in kilo-base pairs, or kbp). The position of genes is shown by magenta arrows. Transcription of a gene is accompanied by a spike in the supercoiling. The bottom panel shows a set of histograms which measure how many times each of the genes has been transcribed up to now; it is synchronised with the dynamics shown in the top panel. The dynamics is sped up in the middle of the video to render the transcription wave more visible. For more details, including the meaning of the parameters J and D (supercoiling transcription-associated flux and supercoiling diffusion coefficient, respectively) see Brackley C A *et al* 2016 Stochastic model of supercoiling-dependent transcription *Phys. Rev. Lett.* **117** 011101. Video available at http://iopscience.iop.org/book/978-0-7503-1602-6.

observed that the wrapped DNA is actually a bit over-twisted, contributing the missing 0.75, meaning the linking number theorem still applies. As the loss in linking number is 1 per nucleosome, and as a nucleosome contains about 200 base pairs, or just under 20 helical turns, the negative supercoiling in mammalian DNA is about −0.05: strikingly, this is very much like the case in bacteria, although as far as I know a deep reason why this is the case has yet to be found.

Some negative supercoiling may well be beneficial for eukaryotic cells such as ours. If supercoiling were mostly stored as writhe, rather than twist, it would not favour polymerase binding, hence initiation of transcription. However, it would still help once transcription starts. We know the polymerase will generate positive supercoiling in front, and negative supercoiling behind. The positive (right-handed) supercoiling will destabilise the left-handed wrapping of the DNA in front of the polymerase, which is probably beneficial as tighly-wrapped up nucleosomes constitute an obstacle for transcription. Simultaneously, the negative supercoiling will favour the reassembly of a left-handed nucleosome on the DNA in the wake of the

transcribing machinery. Although this model is appealing as it highlights a possible role of supercoiling in our cells, there is little experimental evidence for it, as it is so difficult to monitor the dynamics of single molecules in a cell in real time.

Experimentalists can now measure the average supercoiling in a population of cells, as a function of position along the DNA. To do so, they exploit the properties of the psoralen molecule, another intercalator (which is also a potent cancer-provoking chemical). As psoralen binds preferentially to locally unwound DNA, it can be used to construct a map of supercoiling, or more precisely of local DNA twist, in human chromosomes. These maps, put together in Nick Gilbert's and David Levens's labs, show an interesting property: supercoiling is not uniform or randomly distributed over the whole chromosomes, but organised into domains where the DNA is on average either over or underwound. As predicted, transcription is an important factor. Inhibiting transcription leads to massive changes in the maps. Type I or type II topoisomerase can alter supercoiling, either by cutting one strand and letting it swivel around the other (topo I) or by mediating a crossing in double-stranded DNA segments to change the writhe from positive to negative or vice versa (topo II). We would therefore expect that these maps should also be dependent on the action of topoisomerases, and again they are: tampering with topo I or topo II activity also leads to sizable changes in the supercoiling domains. Experiments reveal domains about 100 kilo-base pairs in size, and smaller foci, about 1–2 kbp long, centred around transcription start sites. Domains are also present in bacteria, but about 10-fold smaller. Beyond the key role of transcription and topoisomerase, the mechanisms underlying the emergence of these domains are not yet understood. Possibly, supercoiling generated at the promoter diffuses until it finds a yet-unidentified barrier.

In a recent computer simulation study, we stumbled upon another subtle consequence of supercoiling, this time affecting the self-assembly, or 'reconstitution', of a chromatin fibre from DNA and a set of proteins. Chromatin reconstitution is a well understood process which experimentalists use to make chromatin *in vitro* outside the cell. As histone proteins are positively charged, they are naturally attracted to DNA, which is negatively charged. We therefore reasoned that the only requirements for a simulation of chromatin reconstitution would be to start with a long DNA molecule and histone proteins, and then wait for electrostatic interactions to assemble a chromatin fibre, as in the lab. Unfortunately (or interestingly), it turned out that things were not so simple! What we observed was that the simulated reconstituted chromatin fibre was full of defects, often resulting in nucleosome clustering (figure 5(*a*)). Such clustering is absent in 'native' chromatin fibres extracted from eukaryotic nuclei. The simplest structure formed by such fibres is the so-called 10-nm fibre, which has a typical beads-on-a-string morphology (figure 5(*b*)). The beads are well isolated nucleosomes, and the string is that DNA in between them.

What was happening in the simulations, which led to defect formation? A first clue was that the reconstitution worked much better when we allowed DNA segments to cross through each other—a process mimicking the action of topo II (figure 5(*b*), video 3). A second clue came from a detailed analysis of the binding dynamics between histones and DNA. While some of the nucleosomes formed

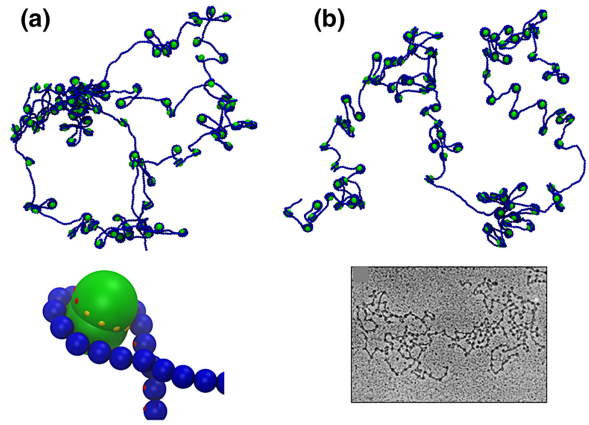
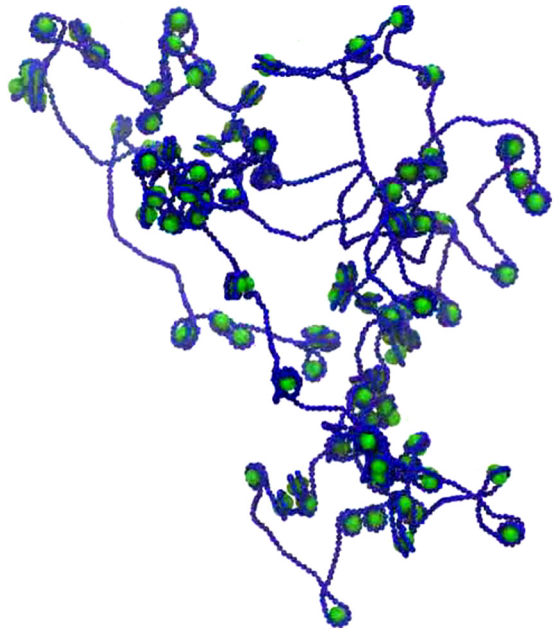


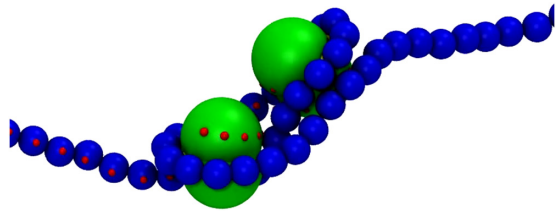
Figure 5. Supercoiling and chromatin self-assembly. (*a*) The top snapshot results from a chromatin self-assembly simulation, with no topo II (strands cannot pass through each other). There are clusters of nucleosomes, which arise due to defects created early in the simulation. By far the most common one is the partially wrapped nucleosome shown in the bottom. Colour codes: green = histone octamer; blue = DNA beads; patches on histone octamers = location of positive charges; patches on DNA = location of negative charges. (*b*) The top snapshot now shows a simulation snapshot obtained *with* topo II. The chromatin fibre is much more regular, there is almost no clustering, and the structure is similar to the 10-nm fibre (or bead-on-a-string structure) shown in the bottom (electron microscopy image). Reproduced from Brackley C A, Allan J, Keszenman-Pereyra D and Marenduzzo D 2015 Topological constraints strongly affect chromatin reconstitution *in silico Nucl. Acids Res.* **43** 63.

properly, in some others the DNA only partially wrapped its histone octamer, typically with one right-handed turn (recall this is the wrong handedness for nucleosomes). This partially wrapped structure formed early on (figure 5(*a*)) and initiated more defects later in the simulation (see video 4). As you may by now have guessed, the issue is one of supercoiling! The partially wrapped structures have a writhe of +1 (and no twist), whereas we need to convert these into conformations with negative writhe (ultimately −1.75) to get properly wrapped left-handed nucleosomes. How can we do this? One way is to use topo II: a simple double-strand crossing immediately changes a right-handed crossing into a left-handed one. Another option is to use equation (1). Although the DNA molecules we considered in the simulations were linear, their linking numbers can only change very slowly as supercoiling needs to diffuse through the whole molecule and dissipate at the open ends (which are free to rotate). Therefore, the partially wrapped nucleosome in figure 5(*a*), which has Wr = +1 and Tw = 0, hence Lk = +1, is equivalent to a left-handed wrapping with Wr = −1 and Tw = +2 (see the ribbons in figure 6). This is normally an unlikely structure, as the DNA between nucleosomes does not like twisting away from its equilibrium angle. Topo I may change this though: if present, it allows us to get rid of the expensive twist so the transition between right-handed and left-handed wrapping takes place. Indeed, our simulations showed that chromatin assembled well when we simulated either topo II or topo I action (videos 3 and 5).

Therefore, supercoiling creates a generic problem for chromatin reconstitution in the lab, and presumably for chromatin assembly *in vivo* as well. Indeed,



Video 3. This video shows the initial and final configuration of a simulated experiment where chromatin is reconstituted from histone octamers and DNA; histone octamers have a left-handed helical binding path for DNA on their surface, to simulate localised charges. There are two cases considered: one where strand crossing in the DNA is disallowed, and another one where it is permitted (for a short period of time) to mimic the action of the topoisomerase II enzyme. The chromatin fibre which self-assembles is more ordered in the second case. Two-step reconstitution (referred to in the video) refers to the fact that the patches on the histone octamer surface are activated in two steps (different colours correspond to patches activated immediately or in the second step), to model the process of chromatin reconstitution in the lab, where an octamer self-assembles from two tetramers, and the first tetramer may interact with DNA before the second joins to form the octamer. For more details, see Brackley C A, Allan J, Keszenman-Pereyra D, Marenduzzo D 2015 Topological constraints strongly affect chromatin reconstitution *in silico Nucl. Acids Res.* **43** 63. Video available at http://iopscience.iop.org/book/978-0-7503-1602-6.



Video 4. This movie shows close-up views of the defects in the chromatin fibre reconstituted in computer simulations. Red patches on the DNA correspond to locations of charged phosphates (interacting with the patches on the histone octamers), and allow one to reconstruct the twist in the modelled molecule. For more details, see Brackley C A, Allan J, Keszenman-Pereyra D, Marenduzzo D 2015 Topological constraints strongly affect chromatin reconstitution *in silico Nucl. Acids Res.* **43** 63. Video available at http://iopscience.iop.org/book/978-0-7503-1602-6.

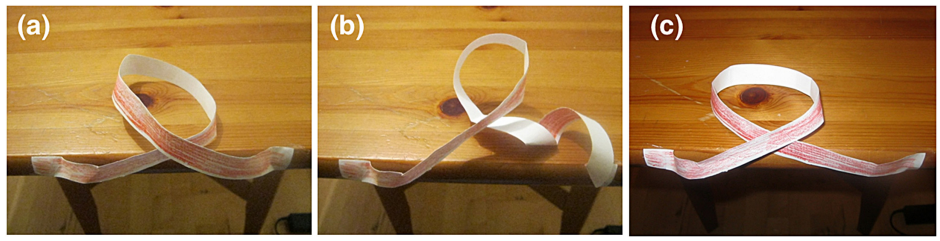


Figure 6. Twist and writhe in ribbons. A possible pathway from positive to negative writhe in a ribbon: the partially wrapped structure in figure 5(a) needs to undergo this transition if a proper left-handed nucleosome is to be assembled. The ribbon in (a) has positive writhe +1 and no twist, it is equivalent (via the linking number theorem) to the ribbon in (b), which has writhe −1 and two units of right-handed twist. If topo I acts on (b), it can remove the two units of twist to turn (b) into the ribbon in (c).

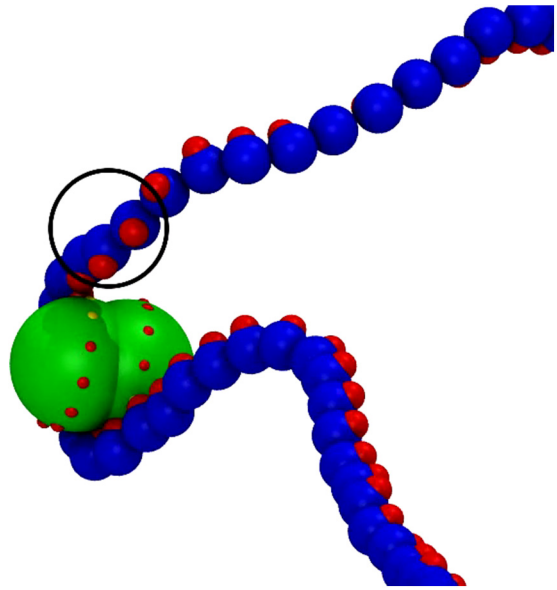
reconstitution in the test tube requires either a very long time, or the presence of cell extracts containing a number of auxiliary proteins, including topo I, topo II, and other chaperone proteins which inhibit nucleosome clustering. The simulation results also explain experiments with yeast 'minichromosomes', consisting of small DNA loops associated with histone proteins. When the loops are positively supercoiled, a proper chromatin fibre can only be assembled in the presence of topo I or topo II, but without these, the positive supercoiling leads to defective nucleosomes in the yeast minichromosomes, just as in figure 5(*a*).

Packing it all up: the 3D spatial organisation of the nucleus

3D organisation of chromosomes and proteins

The conformations adopted by human chromosomes in 3D are important because they determine the pattern of genes that are transcriptionally active or silent. These 3D conformations change significantly during aging, when cells are sick (in cancer, for instance), and during development. To see a simple but striking example of the importance of 3D structure, think of two different cells in your body, for instance one making up your eye and another the skin on your hand. Quite remarkably, both cells have exactly the same genetic material to play with, as they differentiated from the same egg cell. Yet, the genes that are active, hence the proteins that are produced in each of these, are very different! This is possible, in large part, due to the different 3D structure of the chromosomes. If a gene needs to be silenced in the eye, a simple way to achieve this is to tightly fold up the associated chromatin so that polymerases cannot access the promoter. Compact chromatin structures are called 'heterochromatin', in contrast to 'euchromatin', which are open chromatin fibres more likely to bind the transcription machinery.

Polygenic traits are a more complex example of the potential importance of 3D genome organisation. Examples of these are susceptibility to type II diabetes, or simply human height. Such features are not determined by a single gene: i.e., there is no single gene which will determine how tall we will be as adults or how likely it is that we will develop type II diabetes. Instead, scientists have found hundreds of regulatory sequences in our genes that correlate with these traits. The correlation is



Video 5. This video shows how modelling the action of topological enzymes—topo-II and topo-I—can resolve a defect consisting in a partially wrapped nucleosome. The first part of the video show simulated topo-II action, the second part shows simulated topo-I action. For more details, see Brackley C A, Allan J, Keszenman-Pereyra D, Marenduzzo D 2015 Topological constraints strongly affect chromatin reconstitution *in silico Nucl. Acids Res.* **43** 63. Video available at http://iopscience.iop.org/book/978-0-7503-1602-6.

weak, so more are constantly being discovered. It is likely that these traits are not only determined by DNA sequence, but also by the overall pattern of active genes and regulatory elements, which in turn depends on 3D genome structure. This reasoning raises intriguing possibilities for future medicine, as many diseases that are not linked to mutations in the 1D sequences of base pairs in human DNA might be caused by inappropriate folding in 3D. This is the driver for the dramatic surge in interest in finding the 3D structure of chromosomes *in vivo*, which led to the development of 'chromosome conformation capture techniques' to determine it experimentally, as we will discuss later in this section.

Some basic aspects of genome organisation are discussed in section 2; however, we are still very much in the dark when it comes to chromosome organisation in eukaryotes such as ourselves, as well as much simpler ones such as the fruitfly, or even yeast.

What *is* clear is that proteins associating with the chromosomes play a very important role in their organisation. This is because many such proteins can bind to more than one DNA region, creating long-range genomic contact or chromatin loops which shape the 3D structure of chromosomes. As we shall see, the organisation of these proteins inside the nucleus is also highly non-random, and intimately related to the organisation of chromatin.

Nuclear bodies, transcription factories and the bridging-induced attraction

Scientists looking at cell nuclei often find that some proteins inside it (visualised, for instance, via fluorescence) are not uniformly distributed, but rather cluster to form clearly visible spots of variable size. These clusters are now collectively known as nuclear bodies, and have a notable history. They are important for many reasons, and have an intriguing relationship with chromosome organisation.

Believe it or not, nuclear bodies have been known since the middle of the 19th century, more than a century before the discovery of the DNA double helix by Watson and Crick. The first nuclear body to be discovered was the biggest: the nucleolus, which much later turned out to be an assembly site for nascent ribosomes (the machines which translate RNA into proteins). At the beginning of the 20th century, Ramon y Cajal (who now give the name to a prestigious postdoctoral Spanish fellowship) found more nuclear bodies; the 'Cajal bodies', clusters containing a high concentration of a protein known as coilin, which are most visible in rapidly dividing cells. Both the nucleolus and Cajal bodies are unlike other organelles that are found in cells, as they are not enclosed by a membrane. It is therefore likely that they self-assemble from proteins, and possibly other components.

There are several other examples of nuclear bodies. Promyeolocytic, or PML, bodies were discovered in the 1960s and are named after PML protein, the main component of the clusters, and a tumour suppressor. PML bodies are disrupted in cells of patients with leukaemia, and reassemble following treatment with anticancer drugs. More recently, it was found that polymerases and transcription factors also form clusters in the nucleus—a discovery originally made in Peter Cook's group in Oxford in the 1990s. Factories may boost transcription, as they contain a high local concentration of all the required proteins. Interestingly, evidence suggests that there are also specialised transcription factories which produce proteins related to each other, for instance by having similar functions.

You may think at this point that this all seems like hard-core biology—can physics really help us shed any light on such structures? One clue comes from the sheer number of proteins that form nuclear bodies. If so many different proteins tend to cluster in the nucleus, there may well be a generic reason for this, and this is where physics usually helps! Let us start from something really basic, and see where this takes us. Consider a simple toy model for chomatin interacting with some associating proteins. The toy model is schematically described in figure 7(*a*). A chromatin fibre viewed as a bead-and-spring polymer, interacts with some spheres, representing proteins. A key feature of the model is that the red spheres representing proteins in figure 7(*a*) are multivalent, so that they can form 'molecular bridges' that stabilise loops. As we previously mentioned, there are several chromatin-binding proteins that can do this. This simple setup defines the 'strings-and-binders' model introduced by Mario Nicodemi and collaborators at the University of Naples.

We were working with this model in Edinburgh a few years ago to study something unrelated, the traffic of transcription factors on chromatin. What we found, to our surprise, was that even when we assumed no interaction between the proteins apart from steric repulsion, if the attraction between chromatin and

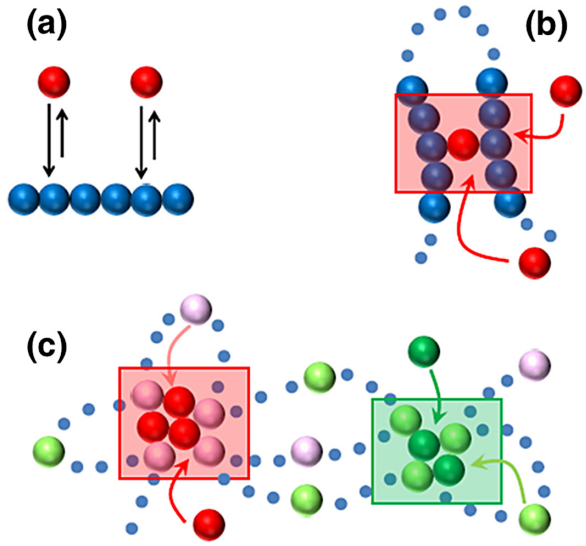


Figure 7. The bridging-induced attraction. Cartoon of the toy model discussed in the text. (*a*) and (*b*) Bridging-induced attraction leads to protein clustering, here with non-specific interactions. Colour codes: blue = chromatin fibre; red = proteins. Arrows indicate (non-specific) binding. In (*b*) the shaded area highlights the increase in local chromatin density due to bridging, which is the key to the feedback described in the text. Panel (*c*) shows an analogous cartoon, but now with specific binding. Colour codes: red and green spheres = red and green proteins; pink and light-green spheres: specific binding sites for red and green proteins respectively; blue = chromatin (binds to red and proteins non-specifically). The boxes indicate separate feedbacks leading to red and green factories/clusters.

proteins was large enough to allow multivalent binding (i.e., bridging), then proteins spontaneously clustered. After some head scratching, we realised that the mechanism behind this clustering is rather simple. The idea is that clustering is due to a thermodynamic positive feedback loop which works as follows (figure 7(*b*)). First, proteins bind to chromatin, and as they are at least bivalent, form a molecular bridge between two different DNA segments. This bridging brings distant parts of the chromosome together to increase the local chromatin concentration, which makes it more likely that additional proteins in the soluble pool will bind as they diffuse by. Once they have bound, these proteins will form additional molecular bridges that increase the chromatin concentration further. The cycle repeats, triggering the formation of stable protein clusters. This is the 'bridging-induced attraction', which we first discussed in a paper in the *Proceedings of the National Academy of Sciences of USA* in 2013. Note we are assuming that protein concentration is low enough that, even if all proteins bind to chromatin, they cannot uniformly cover the fibre. If protein concentration is very high, bridging induces macroscopic collapse of the whole fibre, which was studied in the Nicodemi group.

Now back to nuclear bodies and transcription factories: these structures are very much like the clusters in figure 7. The bridging-induced attraction provides a simple and appealing mechanism to explain the formation of nuclear bodies, as it applies to all multivalent proteins which interact with DNA. Importantly, in all the structures from

the nucleolus to transcription factories, at least some of the proteins in the bodies associate with DNA. Additionally, this physical picture also explains why there is no need of an enclosing membrane: clusters self-assemble due to the bridging-induced attraction which acts between different proteins, much in the same way as a liquid droplet of water forms due to the attractive interactions between water molecules. And just as in the liquid droplet, the underlying attractive interactions provide protein clusters with a surface tension determining their shape. It is remarkable that these complex cellular structures assemble through simple thermodynamic processes. As we are about to see, the parallel between biology and this model runs deeper.

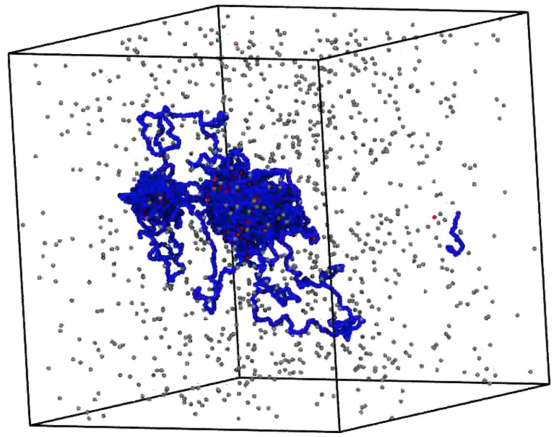
Cluster sizes, dynamics and switching proteins

It is now time to ask some harder questions of our simple model. To begin with, what sized clusters are formed through the bridging-induced attraction, and how do they compare with the size of nuclear bodies which are normally between 100 nm and 1 micron?

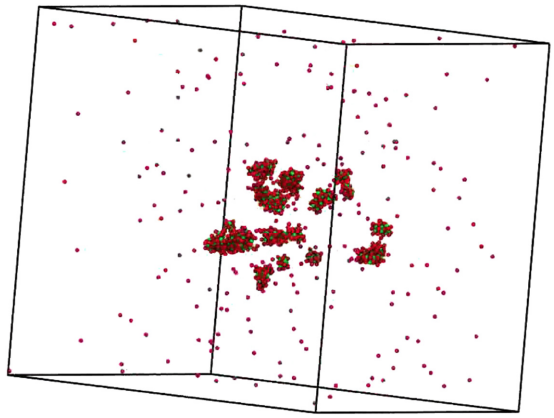
In the simple case in which the bridging proteins only bind 'non-specifically', which means to any blue bead in figure 7(*b*), then the feedback we described yields clusters that continue to grow indefinitely. If we waited long enough, we would get one single giant cluster. This growth is driven by surface tension: the principle is the same for merging oil droplets in water (droplets merge until the oil phase separates fully from the water phase). Clearly, this is not realistic for nuclear bodies!

However, most DNA-binding proteins also bind specifically, as well as non-specifically. A simple modification of the toy model includes a stronger specific binding (e.g. red proteins to pink chromatin beads in figure 7(*c*)). Clusters still form via the bridging-induced attraction (figure 7(*c*)), but now they no longer grow indefinitely; instead, they reach a self-limiting size. This is because clustering of specifically-bound beads creates rosettes, or other structures with many chromatin loops, and bringing these together is entropically costly (i.e., there are very few possible conformations where loops are together due to steric clashes between them). Crucially, one can show that the entropic cost rises super-linearly with loop number (i.e., creating n loops costs more than n times the cost of one loop). On the other hand, the energy gain can scale only linearly (putting together more factors increases the total energy as each binds a few chromatin sites). Therefore, at some point the entropic cost becomes too large, and outweighs the energetic gain and arrests cluster growth. Simulations show that this size depends on the pattern of specific binding sites and on interaction strength. Choosing parameters relevant for a polymerase or transcription factor associated with gene activation gives cluster size of 10–20 units, in good agreement with size of transcription factories. If you wish to see the results of simulations for yourself, have a look at videos 6 and 7. The first corresponds to the case of non-specific interaction (figure 7(*b*)), and demonstrates that clusters coarsen; the second corresponds to the case of specific and non-specific interactions (figure 7(*c*)), and shows that cluster growth stops when they reach a reasonably well defined size.

This is good—but what about the existence of specialised factories: can the bridging-induced attraction explain why they form? To answer this question, imagine that two different types of transcription factor (i.e., 'red' and 'green')



Video 6. This video shows a simulation of cluster formation in the case where proteins (red spheres) bind to chromatin (blue) non-specifically. Clusters coarsen indefinitely. Chromatin is made invisible at times to render protein clusters more easily visible. Available at http://iopscience.iop.org/book/978-0-7503-1602-6.



Video 7. This video shows a simulation of cluster formation in the case where proteins bind to chromatin both specifically and non-specifically. Sites of specific interactions are coloured in green. Clusters form, but now reach a self-limiting size, they do not coarsen indefinitely this time. Chromatin is made invisible at times to render protein clusters more easily visible. Available at http://iopscience.iop.org/book/978-0-7503-1602-6.

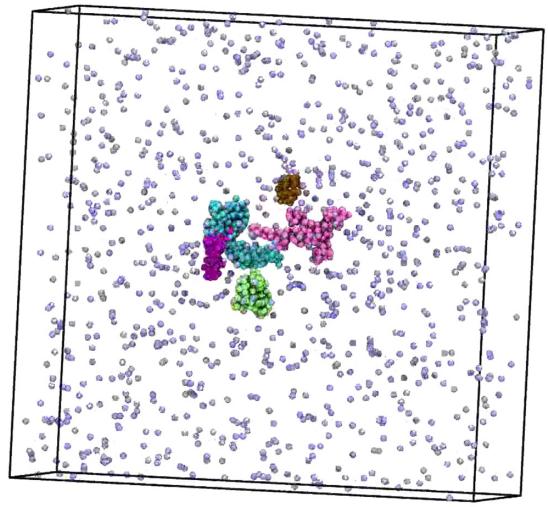
bind specifically to different beads on the fibre (i.e., pink and light green; figure 7(*c*)). Then, the bridging-induced attraction works for the red and green factors separately. For instance, red factors increase the local concentration of pink chromatin binding sites, which recruits more red proteins, and so on. Consequently, the clusters that emerge tend to contain either red factors and pink beads, or green factors and light-green beads. Therefore, specialised factories can be simply explained if different types of transcription factors, or polymerases, bind to different regions of the genome. The most striking example of specialised factories is that of RNA polymerase II and III, which form distinct foci. These two proteins bind to different sequences; hence they are well represented by the red and green beads in figure 7(*c*).

There is another important property of nuclear bodies, which is subtler as it is associated with their dynamics, rather than their morphology. One may expect that the protein clusters comprising nuclear bodies, once formed, should be relatively static, as they are held together by protein–protein and DNA–protein interactions. Surprisingly, this is not the case. Fluorescence microscopy reveals that the components of nuclear bodies exchange rapidly with their surroundings. If we were to photograph a given cluster at two separate times, the protein building blocks would not be the same. Essentially all proteins in it will have been recycled. Even more puzzlingly, the shape of the cluster remains approximately the same over time! So, a nuclear body is a bit like a Japanese temple made of bamboo sticks: this looks the same after a thousand years, after each piece of wood in the structure has been replaced many times.

The dynamic nature of nuclear bodies appears to be out of reach for our model. This is because up to now we have imagined that the interactions between DNA and proteins in the clusters are governed by thermodynamics. In thermodynamics, if you need to make a structure such as a nuclear body, which is stable for a long time in the cell, you would prescribe a strong attraction between the components. But if you do that, the structure tends to become static, as each component is stuck to the others with a strong bond.

However, there is no reason for the interaction between DNA and proteins to be purely thermodynamic in nature. Indeed, ATP hydrolysis continuously provides biochemical energy to the components of the cell, and drives them out of thermodynamic equilibrium all the time. (Often, biologists say that thermodynamic equilibrium means death for the cell.) One use of the energy from ATP is to 'modify' the proteins in the nucleus or cytoplasm. These modifications are chemical reactions which change the physico-chemical properties of the protein. These are known as *post-translational modifications*, because they occur once the protein is fully translated. Many of these modifications, such as acetylation and phosphorylation, change the charge of the protein, and therefore affect its ability to bind DNA. Inspired by these biochemical processes, we recently studied a simple version of our model, where proteins can switch between an 'on' state, which binds to chromatin, and an 'off' state, which does not. This 'switching' is a very simple model for a post-translational modification, and leads to significant differences in the dynamics of the clusters arising from the bridging-induced attraction.

Without switching, proteins can only unbind from a cluster thermodynamically, which requires a long time so proteins rarely exchange between clusters. In video 8, we show results from a simulation without switching. Here, for clarity of visualisation, we have removed chromatin beads (this had been simulated, but is simply not shown), and coloured beads according to the cluster they belong to. The video shows that clusters keep their colour because the proteins are stuck in their clusters due to strong thermodynamic attraction. In contrast, switching proteins form clusters which exchange their components, leading to the multicoloured clusters later in the simulations (see video 9). Non-switching proteins therefore form static clusters; switching proteins, instead, yield dynamic clusters which retain memory of their shape—just like nuclear bodies! The memory comes from the underlying



Video 8. This video follows the evolution of some protein clusters, which have already formed by the beginning of the simulation shown, through the briding-induced attraction elucidated in the text and in video 7. Chromatin is not shown for clarity. The proteins are coloured according to the cluster they belong to when the movie starts, while proteins not in any clusters at that time are gray. Proteins are not switching. It can be seen that clusters do not change their colour: this means that the proteins making up each clusters do not exchange with the diffusing pool (or with proteins in other clusters). For more details, see Brackley C A *et al* 2017 Ephemeral protein binding to DNA shapes stable nuclear bodies and chromatin domains *Biophys. J.* **112** 1085. Video available at http://iopscience.iop.org/book/978-0-7503-1602-6.

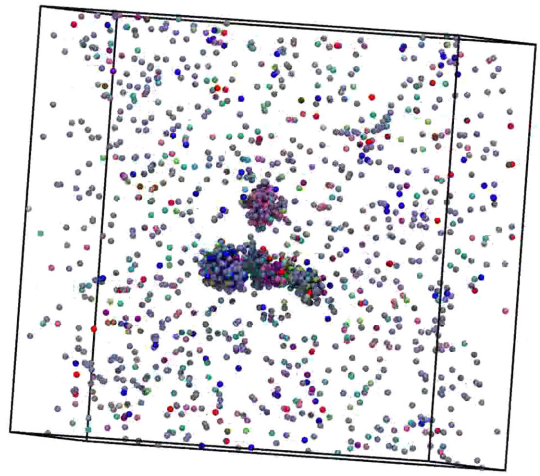
chromatin scaffold, which remains relatively unchanged while the proteins bound to it keep changing.

Taken together, these results strongly support the conjecture that nuclear bodies can be explained by simple physical models. Bridging-induced attraction provides the glue keeping clusters together, whereas protein switching endows the clusters with their dynamic nature, typical of nuclear bodies and transcription factories. The model also yields a prediction for the timescale of recycling—the switching time, or the typical time over which a protein undergoes a modification. Fluorescence microscopy reveals turnover times of the order of minutes, which is compatible with our predictions.

Chromosome domains

Up to now, all we have spoken about has been proteins. But what about the main players (at least in this book), the DNA and chromosomes? What is their 3D structure, which we argued is so important to determine transcriptional activity of the genes? As we shall see, the two problems are very much linked.

Experimentalists have developed a neat technique to address these questions, and determine which regions of a chromosome are close to one another in space. To do so, they first freeze the 3D structure of chromatin by using a molecular glue, normally a crosslinker such as formaldehyde. They then use ‘molecular scissors’ (so-called restriction enzymes) to cut the DNA into a myriad of tiny pieces (normally each



Video 9. This video is the analog of video 8, but with switching proteins. By the end of the simulations all clusters have mixed colours, which means that the proteins in the clusters are not the same throughout and exchange with the diffusing pool (or with proteins in other clusters). It can be seen that the shapes of the protein clusters remain approximately the same at all times, even if the constituent proteins are continuously recycled. For more details, see Brackley C A *et al* 2017 Ephemeral protein binding to DNA shapes stable nuclear bodies and chromatin domains *Biophys. J.* **112** 1085. Video available at http://iopscience.iop.org/book/978-0-7503-1602-6.

piece is a few hundred base pairs long). Because of the formaldehyde, DNA fragments which were together before cutting are often still glued together. At this point, the fragments must be identified. To do so, there are various protocols, but they all essentially aim to fish out as many pairs of glued fragments as possible and sequence them, so that their original position in the genome can be retraced. These techniques, known collectively as 'chromosome conformation capture', allow scientists to build 'contact maps', describing the frequency of contacts between any two points in the genome. To get accurate statistics for the likeliness of contacts, an enormous sequencing effort is required, and usually researchers need to harvest contacts from a whole population of cells. These (pretty expensive!) experiments are known as high-throughput chromosome conformation captures, or in short Hi-C, and they have given us unprecedented insight into the 3D organisation of chromosomes.

Contact maps obtained using Hi-C show that each chromosome folds into distinct domains. For historical reasons, these were originally named topologically-associating domains, or TADs. Research proved this to be a bit of a misnomer, as it is now doubtful whether topology is the main mechanism involved in domain formation (in hindsight, this name would have been instead a perfect fit for the supercoiling domains discussed previously). Domain size is variable, with higher-resolution studies typically uncovering smaller TADs in the range between 100 kbp and 2 Mbp. Initially found in mouse and human cells, this organisation is actually conserved, and also found in budding yeast and even in bacteria, where they are called 'chromosomal interaction domains'.

Bioinformatic analysis suggests that eukaryotic TADs tend to be associated with the local chromatin state: active and inactive regions typically form separate domains. At a higher order, active/inactive domains contact other active/inactive domains, to form A/B compartments (A is active, B inactive). An interesting question is whether there is anything peculiar at inter-domain boundaries. Analysis of Hi-C results showed that that these boundaries are often active genes. A protein which is also found in between domains is the CCCTC-binding transcription factor, or CTCF—whose role is subtle and has been the subject of huge controversies in the recent past. CTCF is associated with chromosome loops, which we will revisit in the next section. One of the reasons why CTCF has attracted a lot of interest is that it appears to play lesser roles in simpler eukaryotes, such as yeast or flies, whereas it is much more important for the regulation of mammalian genomes, such as those of mice and men.

Now back to polymer physics. In the models for figures 7(*a*)–(*c*), we have seen that the bridging-induced attraction leads to the formation of protein clusters. Each of these is associated with a cluster of polymer beads, which can be called a 'chromatin domain'. Do such domains resemble the chromosome domains found experimentally? To answer this question, in 2016 we studied an extension of the strings-and-binders model, including two types of proteins (i.e., binders). The first type binds active chromatin regions, such as transcribed genes, promoters or enhancers (the latter are DNA sequences which, when bound by some activator proteins, can increase transcription of particular genes). Examples of this are the protein complexes which make up transcription factories. The second type of proteins binds inactive chromatin, or heterochromatin. For instance, heterochromatin protein 1 (HP1) and the polycomb repressor complex are multivalent proteins which bind to inactive chromatin—and are required for silencing. We applied this model to study the folding of human chromosome 19, which stands out as one of the most gene-rich. To set up our simulations, we identified beads as active or inactive according to bioinformatic maps. The multivalent proteins fold the chromosome, and we can directly compare contact maps found *in vivo* with those found *in silico*. The agreement is good, especially considering there is no fitting involved in the model (figure 8). As we have seen for the toy model (figure 7(*c*)), active and inactive factors (and their cognate binding sites) cluster separately, so the model naturally yields the A (active) and B (inactive) compartments seen in Hi-C contact maps. Additionally, the model correctly predicts many of the TADs and their boundaries—85% of boundaries are correctly identified to within 100 kbp. These results suggest a possible use of this fitting-free model predicting the contact map of genomic regions for which we do not have Hi-C (for instance, for different cell types in humans, or for different organisms). All that is needed is a 1D map of active and inactive chromatin.

There are other intriguing features of 3D chromosome organisation which escape the Hi-C analysis of intrachromosomal contact maps. Perhaps the most important is that chromosomes are not positioned randomly within the nucleus. For instance, in human lymphocytes gene-rich chromosomes (such as chromosome 19) are positioned internally, and gene-poor ones (such as chromosome 18, which is similar in size) are peripheral. In general, heterochromatin tends to be on the outside of the

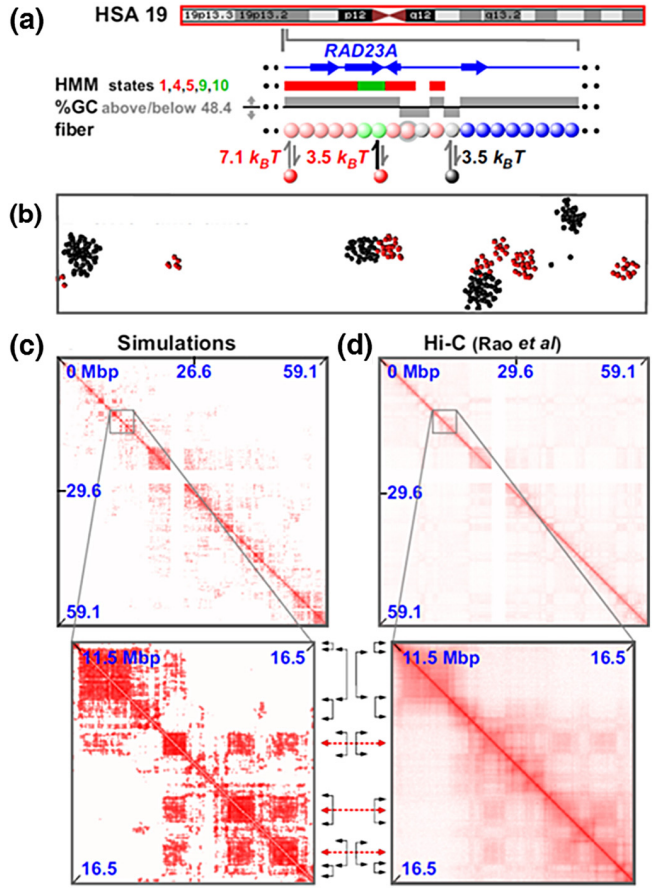


Figure 8. Chromosome domains and Hi-C. The simple 'strings-and-binders' model of figure 7 is generalised here to study whole chromosomes. (*a*) Red and black proteins represent active and inactive transcription factors. The former bind active regions (nonspecifically to green regions, which are transcribed genes, and strongly, or specifically, to pink regions, which are promoters or enhancers). The latter bind inactive regions (nonspecifically to gray beads). (*b*) The bridging-induced attraction creates separate red and black clusters. (*c*) and (*d*) Correspondingly, chromatin domains are visible in the contact maps, in simulations (*c*) and experiments (*d*). Reproduced from Brackley C A, Johnson J, Kelly S, Cook P R and Marenduzzo D 2016 Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains *Nucl. Acids Res.* **44** 3503.

nucleus, where it forms 'lamina-associated domains' (the lamina is the membrane separating the nucleus from the cytoplasm). While specific interactions between lamina and nucleosomes are important to determine this organisation, other factors have been shown to play a role. For instance, the entropy of flexible open chromatin, which is gene-rich, leads to increased repulsion from a wall, such as the lamina. During development, genes which are silenced often go to the periphery, whilst those that are activated migrate to the centre of the nucleus. All this strongly suggests that the position of chromosomes is an important factor for cellular control. However,

there are very few rules in biology without exceptions! The rod retinal cells of nocturnal mammals provide a striking counterexample, where the heterochromatin is central and active genes are close to the lamina. Quite remarkably, this inverted organisation turns the rod cells into efficient collecting lenses, which helps their owners see clearly at night. The pathway through which inversion occurs remains elusive, but it appears that the cell might have found a way to fight the biophysical forces leading to the standard organisation to gain an evolutionary advantage.

A puzzle: CTCF and cohesin, loop extrusion or diffusion

The model in figure 8 does a pretty good job at describing the contact maps observed experimentally. However, there's a catch! When the Lieberman–Aidan lab published a higher resolution Hi-C study in 2014, they uncovered a pretty puzzling fact. They identified brighter spots in their maps, which correspond to long-lived chromatin loops. These are important, as they are known to regulate gene expression. The classic loop motif, reported in biology textbooks, brings together promoters and enhancers to boost the transcription of some genes. However, the Aidan lab focussed on another kind of loop, which involved the protein CTCF. CTCF has been given many names, from molecular fastener, to chromatin insulator or even master organiser; the truth is that we do not yet know exactly how it regulates gene expression, although there is clear evidence that it does so, at least in humans.

What we do know is that the binding sites of CTCF, like those of many other transcription factors, are not symmetrical, so that each binding site has an orientation and can be visualised as an arrow on the chromatin fibre. Therefore, there are four possible orientations for a loop with two CTCF proteins at its anchor. The two CTCF binding sites could be divergent, convergent or in one of two parallel orientations. If CTCF loops were determined by some thermodynamic combination of protein–DNA and protein–protein interactions, there would be no way to discriminate between a convergent and a divergent arrangement. This is because if we keep the 3D structure of chromatin exactly the same and just flip the CTCF arrows we can turn a convergent loop into a divergent one. Strikingly, though, the Lieberman–Aidan lab found that over 90% of CTCF loops involve convergent binding sites, a few percent were in a parallel arrangement, and virtually no divergent loops were found.

How can the bias for convergent loops come about? One possible answer lies with the observation that CTCF in loops is bound to cohesin. Cohesin is a ring-like protein complex thought to bind chromatin (or DNA) by embracing it. There are two models for how cohesin may do this in practice; either as molecular 'hand-cuffs' made by two rings, each of which embraces one chromatin fibre (figure 9(*a*)), or as a single ring that embraces two chromatin fibres at once. In both cases, cohesin behaves as a sliding link between two chromatin fibres. In polymer physics, such structures are known as 'slip links', and we can think of them as carabiners, which clip on a climber's rope. The main role of cohesin (hence its name) is to regulate cohesion between the two copies of each replicated chromosome. However, recent evidence points to an important role for this complex in gene regulation, where it acts in concert with CTCF. The importance of cohesin is clear to doctors, as we now know that mutation in the gene coding from NIBPL, a 'loading' protein that is

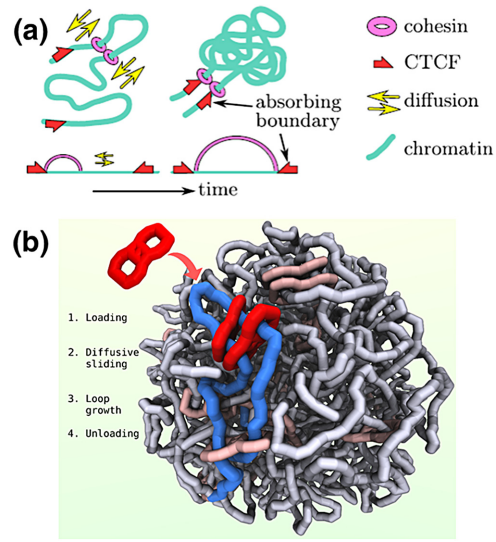


Figure 9. Cohesin, CTCF and genomic loops. (*a*) Sketch of a simple model for the formation of convergent loops. Cohesin here is seen as a molecular hand-cuff (in polymer physics, this is called a slip link), another view is that it could be a single ring embracing two chromatin fibres: the results would be unchanged in the simple model. Here we assume that the two rings in the handcuff diffuse along the chromatin fibre until each finds a CTCF protein pointing towards it, at which point they get stuck. Another model postulates that the two rings have a motor activity and move unidirectionally: this is the loop extrusion model. (*b*) A snapshot of a chromatin fibre where loops are formed by diffusing slip links. Red handcuffs are about to bind or have just bound at a loading site (arrow); pink handcuffs have bound previously. The blue loop is the largest cohesin-mediated loop close to the highlighted binding site. The enlargment of this loop is due to the ratchet effect described in the text. Panel (*a*) reproduced from Brackley C A, Johnson J, Michieletto D, Morozov A N, Nicodemi M, Cook P R and Marenduzzo D 2017 Nonequilibrium chromosome looping via molecular slip links *Phys. Rev. Lett.* **119** 138101.

required for cohesin to bind to DNA, leads to a rare genetic disorder called Cornelia de Lange syndrome after the Dutch paediatrician who first described it. This condition is linked to developmental diseases, and to inappropriate regulation of some responsive genes. These are genes which are normally silent and need to be pushed into action in response to a stimulus (e.g. inflammation). This is all consistent with the idea that cohesin plays a key role in gene regulation.

Intriguingly, cohesin interacts with CTCF in a way which depends on CTCF orientation: the two proteins only stick together when the CTCF binding motif faces cohesin. Inspired by the biology of cohesin and CTCF, Leonid Mirny and colleagues at MIT proposed in 2015 an appealing model that could crack the convergent CTCF loop riddle. They introduced the ‘loop extrusion model’, which was based on an earlier statistical physics model by John Marko, a well-known DNA biophysicist now at Northwestern University. Mirny’s version of the model postulates that the two monomers in a cohesin dimer can actively create loops of 100–1000 kbps by travelling in opposite directions along the chromosome. A possible alternative is that

a molecular machine pushes the two monomers from behind, leading to the same outcome. If the two monomers start on adjacent positions along the chromatin fibre, they can distinguish between the two sites of CTCF, just as a person walking along a street can tell whether the friend they are about to meet is looking towards them or not. Mirny's team then assumed that cohesin will only bind CTCF when it 'faces' it, and will bounce back or knock CTCF off the chromatin track when it 'faces' the other way.

Whilst appealing, the loop extrusion model may not be the definitive answer to the cohesion–CTCF Hi-C puzzle. The main issue with the model is that it requires each cohesin to determine and maintain the correct direction in order to extrude (rather than shrink) a loop, and that the cohesin speed must be large enough to extrude a loop of 100–1000 kbp while it remains bound to DNA. As experiments showed that this residence time, τ, is about 20 min, a relative speed of 5 kbp min^{-1} between the two monomers is required. This is large in molecular terms, as the RNA polymerase, a very efficient molecular motor active during interphase, only moves at about 1 kbp min^{-1} on chromatin. Additionally, while very recent experiments have shown that condensin (which is structurally related to cohesin) moves on naked DNA, experiments on cohesin have never once found directed motion, but only diffusion, on either DNA or chromatin fibres.

In recent work from our group, we suggested that there might be another explanation for convergent loops. Imagine that cohesin performs random diffusion whilst holding on to a sliding loop, until it finds CTCF (figure 9(*a*)). As the two meet along a one-dimensional track, cohesin can distinguish between the two sides of CTCF, just as in the loop extrusion model. Then, if diffusion is fast enough, stable loops will only form between convergent arrows, without the need for any motor. Estimates for cohesin diffusion on chromatin vary between 0.1 $\text{kbp}^2\ \text{s}^{-1}$ to about 100 $\text{kbp}^2\ \text{s}^{-1}$, according to the conformation of cohesin and the presence or absence of ATP. Taking an intermediate value of 10 $\text{kbp}^2\ \text{s}^{-1}$, dimensional analysis (or random walk theory) suggests that the typical distance travelled by cohesin before detaching is given by

$$l \sim (D\,\tau)^{1/2} \tag{4}$$

which is just over 100 kbp. You may worry that this calculation is based on estimates from *in vitro* experiments. Is diffusion going to be fast enough in the crowded nucleus, where the greater viscosity will slow down most biophysical processes? Our largescale computer simulations of confined chromatin fibre (figure 9(*b*)) suggest that this is still the case. It is also possible to favour loop enlargement over shrinkage by exploiting a subtle ratchet effect. Imagine a first cohesin molecule binds to a chromosome region. Cohesin does not bind randomly, but at specific loading sites. By embracing the fibre at two adjacent points, cohesin can fold it into a mini-loop, and then travel up and down the fibre by diffusion, as before. If at this point a second cohesin is loaded, it limits movement of the first back towards the loading site. If we imagine several cohesins are loaded while the first is still diffusing, they will exert an osmotic pressure biasing the diffusion of the first, and leading to larger loop formation. Computer simulations show that this subtle effect can lead to a boost

in the size of the loops formed. In summary, the simulations suggest the simple diffusive model provides a viable alternative to loop extrusion. At the very least, it cannot yet be discarded without additional experiments.

4 Outlook

This ebook contains a selection of topics where I feel that physics has made a significant contribution to our understanding of DNA and chromosomes. However, there are many more open problems in the field to be explored. The number of DNA-related questions that biophysicists are addressing has increased over the last few years, and all signs point to this trend continuing, if not intensifying, in the future.

The large amount of data being gathered on chromosome organisation will inevitably raise more and more questions. It was due to new data in 2014 that the problem of convergent CTCF looping described above came to the fore. Other experiments and theories, not covered in this book, have begun to unravel the fascinating mystery of how chromosomes transform from their interphase self, where they are organised into domains and compartments, to the compact X-shaped structures seen during cell division. No doubt, new data will lead to fascinating new insights on this old problem. At a more fundamental level, we hope to discover more about the three-dimensional biophysics of DNA replication and transcription, where supercoiling, knotting, and other aspects DNA topology touched upon are likely to play a key role.

One further imagines that with the development of increasingly advanced experimental tools we'll soon be able to probe the microscopic and mechanistic underpinning of many other DNA-related structures, such as the mesmerising dance of chromosomes during meiosis, in sexual reproduction, or the inner workings of DNA repair. Furthermore, our understanding of chromosome structures during development, aging and diseases such as cancer may facilitate the development of novel treatments for diseases. While experiments will always be fundamental for this type of science going forward, biophysics and theory are equally necessary components, as these are needed to unearth the deep reasons underlying the phenomena we observe. One day, some even believe we will be able to use predictive computer simulations to model the organisation of chromosomes in a patient, with a view to delivering personalised care for genetic diseases (including cancer). This may well be over-optimistic, but the very fact that some believe it to be possible is indicative of the future impact of biophysics!

If you like this stuff, then I bet there will be a lot of work for you in the future! Although some of my colleagues feel that these topics are too far from the traditional realm of physics, I have never once regretting entering this field—I only wish I had done so sooner! But then, I believe that my previous experience with soft matter, polymers and liquid crystals has come in handy so many times—and this is true also of my particle physics upbringing, although I do not admit this bit so often.

Additional resources

This list contains some additional reading and references discussed in the text.

Background

Calladine C R, Drew H R, Luisi B F and Travers A A 2004 *Understanding DNA: The Molecule and How it Works* (London: Elsevier Academic Press)

Watson J D and Crick F H C 1953 A structure for deoxyribose nucleic acids *Nature* **4356** 737

Schiessel H 2003 The physics of chromatin *J. Phys.: Cond. Matt.* **15** R699

Knobler C M and Gelbart W M 2008 The physics of phages *Phys. Today* **61** 42

Current directions

All twisted up: supercoiling, transcription and nucleosome assembly

Bates A D and Maxwell A 2005 *DNA Topology* (New York: Oxford University Press)

Burnier Y, Dorier J and Stasiak A 2008 DNA supercoiling inhibits DNA knotting *Nucl. Acids Res.* **36** 4956

Witz G, Dietler G and Stasiak A 2011 Tightening of DNA knots by supercoiling facilitates their unknotting by type II topoisomerases *Proc. Natl. Acad. Sci. USA* **108**

Liu L F and Wang J C 1987 Supercoiling of the DNA template during transcription *Proc. Natl. Acad. Sci. USA* **84** 7024

Brackley C A, Johnson J, Bentivoglio A, Gilbert N, Gonnella G and Marenduzzo D 2016 Stochastic model of supercoiling-dependent transcription *Phys. Rev. Lett.* **117** 011101

Brackley C A, Allan J, Keszenman-Pereyra D and Marenduzzo D 2015 Topological constraints strongly affect chromatin reconstitution *in silico Nucl Acids Res.* **43** 63

Gilbert N and Allan J 2014 Supercoiling in DNA and chromatin *Curr. Opin. Genet. Dev.* **25** 15

Packing it all up: the 3D spatial organisation of the nucleus

Cook P R 2001 *Principles of Nuclear Structure and Function* (New York: Wiley)

Barbieri M, Chotalia M, Fraser J, Lavitas L M, Dostie J, Pombo A and Nicodemi M 2012 *Proc. Natl. Acad. Sci. USA* **109** 16173

Brackley C A, Taylor S, Papantonis A, Cook P R and Marenduzzo D 2013 Nonspecific briding-induced attraction drives clustering of DNA-binding proteins and genome organization *Proc. Natl. Acad. Sci. USA* **110** E3605

Brackley C A, Johnson J, Kelly S, Cook P R and Marenduzzo D 2016 Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains *Nucl. Acids Res.* **44** 3503

Rao S S *et al* 2014 A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping *Cell* **159** 1665

Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N and Mirny L A 2016 Formation of chromosomal domains by loop extrusion *Cell Rep.* **15** 2038

Brackley C A, Johnson J, Michieletto D, Morozov A N, Nicodemi M, Cook P R and Marenduzzo D 2017 Nonequilibrium chromosome looping via molecular slip links *Phys. Rev. Lett.* **119** 138101