8

Three-Dimensional Structure of DNA

Chemically, DNA is a long polymer. This polymer is packed inside the nucleus of the cell by binding to sets of specific proteins called nucleosomes. Chromatin refers to the combined structure of DNA and these DNA binding proteins. One can visualize the three-dimensional structure of DNA as a long polymer winding around at different scales. The three-dimensional structure of chromatin plays a crucial role in a large variety of fundamental biological processes, including replication and expression. For instance, to be transcribed (i.e., to generate RNA from this DNA), the local structure of DNA has to be accessible to different proteins that bind to specific locations. Distant genomic locations can be brought together and coregulated by the same transcriptionary machinery. In this way, the structure of chromatin regulates expression of RNA and influences the expression of proteins by different cells. Thus the three-dimensional structure of DNA determines why two cells containing the same genome can function in very different ways. For instance, a motor neuron and a B-cell share the same genome but their behavior, function, and the proteins they express are vastly different.

In an eukaryotic cell there are well-studied structures that organize chromatin (Figure 8.1). The two DNA strands are 2.5 nanometers wide and coil in the form of a helix of 10.4 base pairs per turn. 146 bases of DNA can wrap around nucleosomes, a structure of four proteins (histones). Histones pack into 30 nanometer filaments in a highly compact way (heterochromatin). At even larger scales, on the order of a million bases, these structures are assembled into different territories, as topological associated domains or TADs. In [330] Lieberman-Aiden et al. suggested specific large conformations, called *fractal globules*. At still larger scales, it has been postulated that different chromosomes are located at specific three-dimensional chromosomal territories. The location of genomic regions within these territories is associated to transcriptionally active domains [375]. Due to the nature of the relevant biological processes, including DNA replication, repair, and transcription, DNA presents a highly dynamical nature.

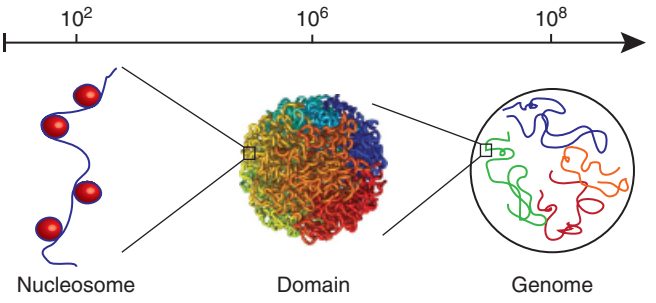

Figure 8.1 Summary of some structures found in chromatin organization. Nucleosomes are sets of proteins that bind DNA at 150 base pair scale. At much larger scales, on the order of megabases, chromatins organize into different territories bound by specific proteins. At even larger scales different structures have been proposed, including the so-called *fractal globule* structure. At still larger scales chromosomes can be found in separate chromosomal territories. Source: [330]. From Erez Lieberman-Aiden, et al., Comprehensive mapping of long-range interactions reveals folding principles of the human genome, Science 326.5950 (2009): 289–293. © 2009 Reprinted with permission from AAAS.

In the previous chapters, we have explored the use of high-throughput sequencing technologies to read the genome of organisms. There are preliminary indications that these techniques can also be used to infer three-dimensional properties of DNA across different genomic scales. We will describe here genome wide chromatin conformation techniques. Data generated in this way provides information about genomic locations that are in close proximity in three dimensions. We will follow reference [163] to describe how topological techniques can be used to infer and quantify three-dimensional structural properties of DNA. In particular, persistent homology provides a natural framework for summarizing these properties. We will first demonstrate the efficacy of persistent homology techniques in data from simulated polymers, and then in the circular bacterial genome of *C. crescentus* and a human lymphoblastoid cell line. We believe that this will be a fertile area for the application of topological data analysis in the future.

8.1 Background

In the last few years there have been extraordinary developments in providing genome wide information on the three-dimensional structure of DNA [25, 138, 330]. Chromosome conformation capture technologies give a variety of methods to study the three-dimensional organization of chromatin inside the nucleus of a cell. These methods identify genomic locations that are in very close physical proximity (in space).

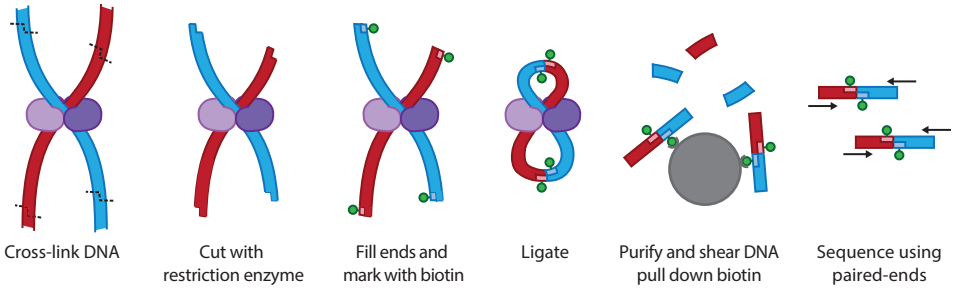

Figure 8.2 Hi-C protocol to map the three-dimensional chromatin structure. First, nearby regions in DNA are cross-linked in formaldehyde. Then DNA is fragmented using a restriction enzyme. A biotinylated residue is incorporated and ends are ligated. The DNA is sheared and the junctions are pulled down with streptavidin beads that bind strongly to biotin. The purified fragments are then sequenced, leading to information on close proximity DNA fragments.

Chromosome conformation capture technologies vary depending on the extent of the regions interrogated. Hi-C protocols use high-throughput sequencing techniques to provide genome wide maps of DNA interaction. A common Hi-C protocol has been summarized in Figure 8.2: DNA fragments in close proximity are cross-linked in formaldehyde, DNA is fragmented, nearby fragments are ligated to form close loops. These loops are then sequenced. Mapping sequence reads into a reference genome, one can identify specific genomic locations in close proximity. The final result is summarized in a contact data matrix that quantifies the genomic locations that are in close three-dimensional proximity [138]. Specifically, if there are n locations, the contact matrix C is an $n \times n$ matrix such that the entry $C_{ij} = C_{ji}$ encodes the proximity between locations i and j. Further processing involves denoising and normalization of the raw data matrix [25].

There are many caveats with the use of this procedure. For instance, we are assuming that different cells present the same three-dimensional structure. Contact matrices represent the average over many different cell configurations, and might not represent any specific configuration. Recently chromosome conformation capture techniques have been applied to single cells [374, 375, 486]. These studies have found that despite a large degree of variability, megabase scale domains are well maintained across individual cells.

8.2 TDA and Chromatin Structure

We are interested in learning global properties of the three-dimensional structure of DNA: how DNA folds at different scales. Hi-C data provides the means to

capture information about the three-dimensional information of DNA as off diagonal elements in the contact matrix. Specifically, once correctly normalized, one assesses the proximity between different genomic regions using the contact matrix. Small loops in DNA can be detected as close to the diagonal non-zero elements in the contact matrix.

Topological data analysis provides a natural language to identify and quantify loops in the three-dimensional structure. One can apply persistent homology to the similarity matrices derived from Hi-C data to try to detect the shape of the DNA. In the following discussion, we focus on PH_1, the one-dimensional persistent homology barcodes, which represent one-dimensional physical loops in DNA.

Recall that each element of the barcode is an interval $[b_i, d_i)$, where b_i is the smallest scale and d_i is the largest scale where the class is present. Following the work of MacPherson and Schweinhart [339], one can define the size of a particular persistent homology class as the mean of the birth and death:

$$x_i = \frac{b_i + d_i}{2}.$$

The values of x_i represent a sample of the distribution of different folding scales inferred from the Hi-C data. Of particular interest in these analyses are large scale interactions (larger than 100 kilobases) that can represent different biologically meaningful structures. For instance, it has been observed that transcription happens in transcription factories, specific three-dimensional locations in the nucleus that accrue large protein complexes involved in transcription [265]. Typically transcription factories encompass tens of RNA polymerases together with a variety of proteins involved in RNA processing, such as helicases, transcription factors, splicing factors, etc. Transcription factories can be physically observed by electron microscopy. Other types of structure can be observed where promoters (the region in a gene associated with transcriptional initiation of a gene, located near the transcription start site of the gene) interact with enhancers (a region in the genome that enhances the transcription of a particular gene) at long distances, sometimes at megabase scales. Long range interactions in chromatin are known to occur in DNA repair and replication among many other processes.

Different biologically distinct structures represent different looping structures in DNA. For instance, a one-jump loop can be observed on promoter-enhancer interactions, while multi-jump loops (multiple loops) from diverse regions can be associated to a transcription factory (Figure 8.3). So the scale and structure of loops can provide biologically relevant information.

While persistent homology provides interesting information on the size and number of cycles, one may be interested in specific details concerning a particular

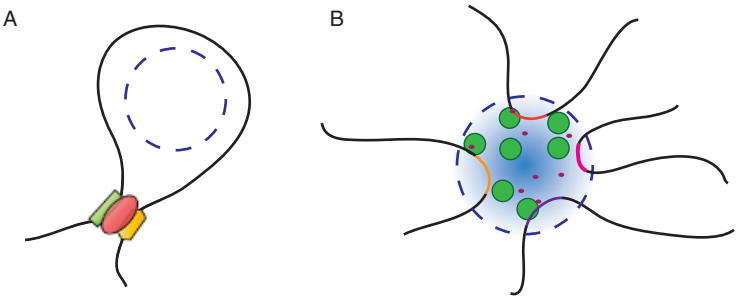

Figure 8.3 The scale and structure of loops can provide biologically relevant information. Here are two examples of known structures associated to transcriptional regulation. (A) Enhancers are genomic loci that regulate the expression of genes that could be located in different genomic locations. Loops in DNA can put enhancers and promoters in close proximity. (B) Other examples can be found in transcription factories containing RNA polymerases, splicing proteins, and other proteins involved in transcription and RNA processing. These are regions that can be linked to distant genomic locations. Source: [163]. Reprinted with permission: © EAI European Alliance for Innovation 2016.

loop in the class. This is important to identify genomic regions of interest and to link structure to function, for instance, linking specific structures to regulation of specific genes. If we identify a class, how can we select a particular member? One obvious criterion is to identify the cycle in the homology class that has minimal size. In the context of Hi-C data and contact maps, Emmett and colleagues proposed [163] to identify minimal cycles as corresponding to the shortest length along the genome being homologically independent to other classes born at smallest scales [449]. However, it has been shown that finding the minimal cycle is an NP-complete problem [113], and so approximation techniques are necessary.

8.3 Simulations

The molecule of DNA can be treated as a polymer with specific biophysical properties. As such it can be modeled as a long homogeneous flexible fiber with interactions within specific sites (see, for instance, [148]). Simulations allow evaluation of the inference procedures used from Hi-C and other types of chromosomal conformation data.

In [163] a 50 megabase chromatin polymer was simulated by considering a polymer formed by 1000 smaller monomers of a few nucleosomes each (Figure 8.4). Specific interactions representing protein-mediated interactions were incorporated by hand at ten random positions in the polymer. A contact map was constructed

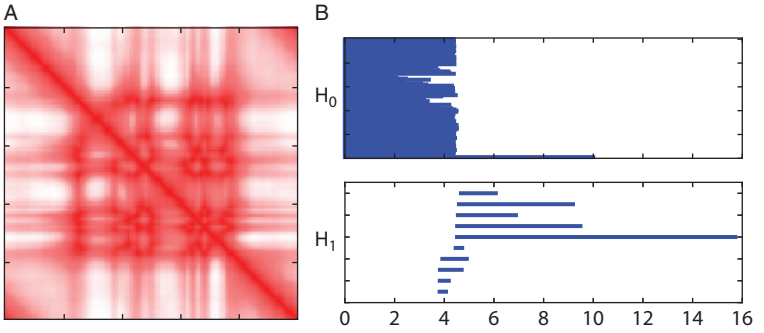

Figure 8.4 Simulations of DNA as a polymer. DNA can be simulated as a long polymer consisting of a large number of monomeric units interacting at specific places. Here, we show the data of a 50 Mb polymer with 10 fixed loops at random positions in the genome consisting of 1000 monomeric units. (A) The average of 5000 simulations allows construction of a contact map. (B) Using persistent homology in a similarity matrix derived from the contact map one can clearly identify the ten loops as ten long bars in dimension one persistent classes. Source: [163]. Reprinted with permission: © EAI European Alliance for Innovation 2016.

using 5000 conformations (Figure 8.4). The contact map was transformed into a similarity matrix $d = 1 - \rho$, where ρ is the Pearson correlation between two genomic positions. The one-dimensional homology groups clearly identify ten long bars, corresponding to the interacting position in the polymer. Polymer simulations provide a nice way to optimize the identification of interactions from contact maps.

8.4 The Topology of Bacterial DNA

We now explore the persistent barcode diagrams derived from real data in two very different systems: a bacterium and an eukaryotic cell.

The typical size of a bacterial genome is a few megabases, which if linearly stretched would reach more than 1 mm. However, bacteria are only a few micrometers long, meaning that the genome has to be compacted 1000 fold. Although bacteria do not have a proper nuclear membrane, there is an irregular shaped region, called the nucleoid, that aggregates most of the genomic material. There are several mechanisms of packing the bacterial DNA genome into the cell. The first mechanism is negative DNA supercoiling. The DNA is a double helix that twists every 10.5 bases. If a segment of DNA of length L (in bases) is circularized by pasting the two ends, there will be a number of turns expected that create a relaxed DNA $\sim L/10.5$. DNA supercoiling occurs when there is an over (positive) or under (negative) winding of DNA (see top panel of Figure 8.5). In general most DNA presents a negative supercoiling. In bacteria, negative supercoiling generates plectonemic

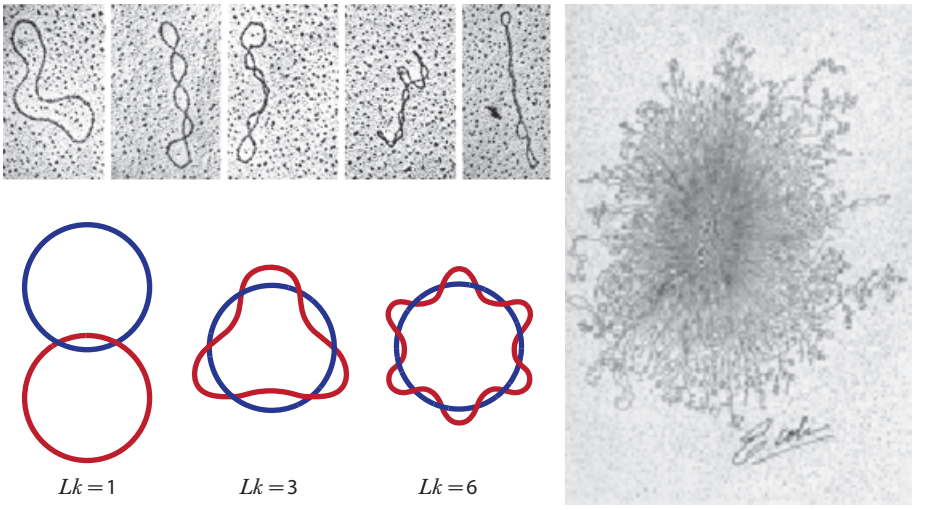

Figure 8.5 Left: Diverse levels of twisting generate supercoiling in circular DNA. Source: [306]. From Kornberg A. 1980 DNA replication, p. 29. San Francisco, CA: W. H. Freeman. Supercoiling can be quantified by the linking number between the two strands of DNA if ends are joined. In a relaxed DNA configuration one should expect a turn every 10.5 bases. So in a length L DNA fragment, the expected linking number is $L/10.5$; deviations from this linking number lead to different levels of supercoiling. Right: Plectoneme emanating from an *E. coli* nucleoid core. Source: [211]. © Designergenes Posters Ltd; in memory of Dr Ruth Kavenoff 1944–1999.

loops. A simple way of characterizing the level of supercoiling is by the linking number of the two strands of DNA (see bottom panel of Figure 8.5). Nice work relating the topology of links and knots to the structure of DNA has been carried out by different groups [20, 80, 126, 178, 308, 415].

The second mechanism of compacting bacterial DNA is by topological domains, supercoiled domains insulated from each other. Topological domains vary in size but are of order 10 kb, indicating that a typical bacterium genome contains hundreds of topological domains. The structure of each domain is kept independent by protein and RNA complexes that work as boundary elements. Of special significance is the structural maintenance of chromosome (SMC) condensin complexes and topoisomerases. SMC proteins form part of a highly conserved complex from bacteria to human that bridge different chromosomal loci working as a high level scaffold. Topoisomerases work as enzymes that can cut the DNA strains, modifying the DNA topology, changing the winding and unlinking DNA loops.

Macrodomains are much larger structures (of almost a megabase) that encompass many topological domains and suggest highly structured regions with reduced

DNA mobility. Understanding the structure of bacterial DNA, the boundary elements, the macrodomain structure, the functional characterization, transcription, and replication remains a very active area of research. For a nice review on bacterial chromosomal organization see [523].

Caulobacter crescentus is a gram negative bacterium ubiquitously found in water. The genome has a length of 4 megabases in a circular chromosome coding for near four thousand genes. Hi-C interaction data from *Caulobacter crescentus* was examined as published in [317]. This paper found several structures, including chromatin interaction domains at scales of 100 kilobases and smaller plectonemes. A simple model of nesting these chromatin structures was proposed where plectonemes were arranged in a brush-like fashion and topological domains encompassed several plectonemes (see panel A of Figure 8.6). The contact matrix binned at a ten kilobase resolution from a wildtype *Caulobacter* cell is represented in panel B of Figure 8.6. Panel C shows the barcode diagram in dimensions 0 and 1 computed from the associated similarity matrix. The one-dimensional barcode shows a very interesting bimodal structure corresponding to a large number of smaller loops and a few larger ones, as shown in the bimodal distribution of panel D of Figure 8.6.

To specify genomic locations associated to particular one-dimensional persistent homology classes, one can identify small cycles within each class. These representative cycles are depicted in Figure 8.7. The bimodal distribution shown in panel D of Figure 8.6 shows two types of loops. The smaller loops are located close to the diagonal (as expected) and could be related to structural maintenance complexes [523]. More interestingly, the larger ones (of approximately 100 kb) connect extremely distant genomic locations in particular locations, suggesting specific genomic locations associated to large range interactions in *Caulobacter crescentus*.

8.5 The Topology of Human DNA

If stretched end to end, the roughly 6 billion bases of the human genome would stretch nearly two meters, yet they are able to occupy a volume of only a few cubic micrometers within the cell nucleus. Even more surprising (or perhaps not surprising at all), this million-fold compression is highly non-random, and exhibits a complex hierarchical structure which impacts genome function intimately by regulating gene expression. This multiscale pattern exhibits increasing levels of complexity: from nucleosomes every 150 bases, to interactions between promoters at the megabase scale, to topologically associated domains at the 10 megabase scale, and finally, to the organization of the chromosomes [138]. Chromatin conformation is dynamic, and changes continuously throughout the cell

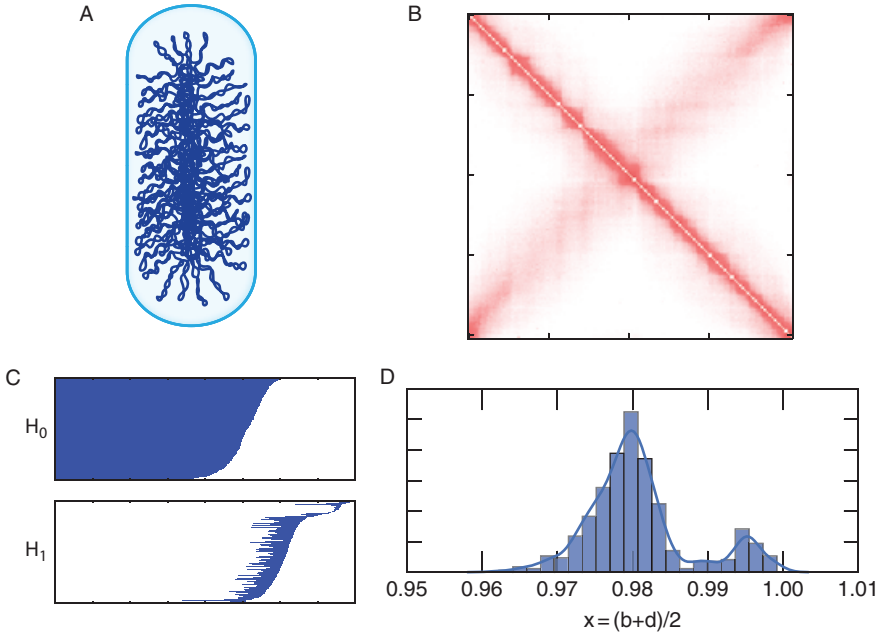

Figure 8.6 Persistent homology study of bacterial DNA structure. (A) Cartoon model of the structure of DNA in *Caulobacter crescentus*. The genome is contained in a large circular chromosome. At smaller scales there are compact domains and plectonemes, supercoiled DNA loops emanating from a central circular fiber. (B) The contact map reveals an off-diagonal structure reflecting the circular nature of the bacterial chromosome. Persistent homology maps (C) indicate a more refined structure shown as distribution of H_1 bar sizes showing a bimodal distribution of DNA folding patterns (D). Source: [163]. Reprinted with permission: © EAI European Alliance for Innovation 2016.

cycle under the influence of a diverse range of chromatin remodeling proteins. Chromatin architecture can also be impacted by post-translational modifications of histones, including but not limited to methylation and acetylation of specific residues on histone tails. Many data sets of human cell Hi-C data have been published [274, 330, 420]. In [163], Emmett *et al.* applied the persistent homology pipeline to study the three-dimensional chromatin structure of a healthy human lymphoblastoid cell line published in [330]. Figure 8.8 shows the contact map of chromosome 1 binned at 1 megabase resolution. On the right of the same figure, the barcode diagrams for dimensions zero, one and two are given. The one-dimensional persistence diagram reveals an interesting pattern of short and long range interactions, as shown in the bimodality of sizes in Figure 8.9. These results support previous observations [330] regarding topological associated domains of size 10 megabases.

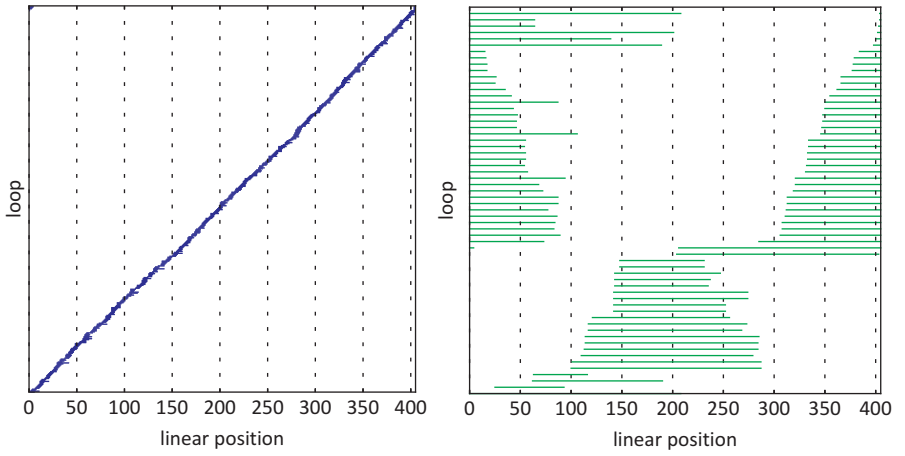

Figure 8.7 Genomic position of minimal cycles associated to persistent homology classes in dimension one. As previously shown, loops can be divided into two types. On the left, smaller loops distribute uniformly across the genome represented as the diagonal. The right panel shows the genomic positions associated to larger loops which clearly show two large interacting domains. Source: [163]. Reprinted with permission: © EAI European Alliance for Innovation 2016.

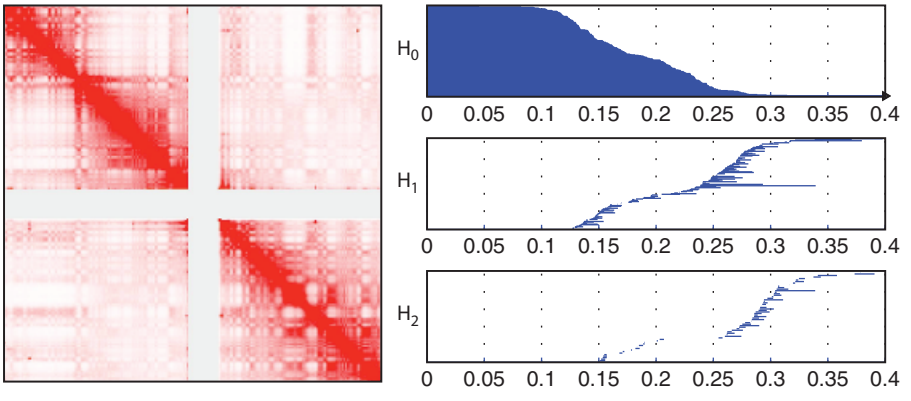

Figure 8.8 Hi-C data for chromosome 1 from a human lymphoblastoid cell line. On the left we show the contact map representation. The white band in the middle represents the centrosome where information is not available. On the right, persistent homology barcode diagrams in dimensions zero, one, and two reveal long-range interaction patterns. Source: [163]. Reprinted with permission: © EAI European Alliance for Innovation 2016.

8.6 Summary

The application of topological approaches to studying the three-dimensional structure of DNA is still in the early stages. We expect that the technology, methods, and many of the ideas reviewed here will be evolving very rapidly in the next few years.

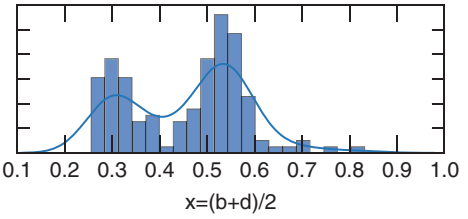

Figure 8.9 The one-dimensional persistent homology barcode of a human lymphoblastoid cell line shows a clear bimodal distribution related to topological associated domains of an approximate size of 10 megabases.

- Cells with the same genome can have vastly different form and function. The three-dimensional architecture plays an important role regulating important biological processes.
- Chromatin structure in the nucleus of cells presents structure at different scales, from hundreds of bases (nucleosomes) to topological associated domains at megabase scale, to chromosomal territories.
- These structures are associated to biologically functional processes, such as RNA transcription.
- Recently chromosomal conformation capture techniques are generating genome wide contact maps reporting large scale interations.
- Topological data analysis techniques, in particular persistent homology, are a natural language to study contact maps and infer the size and number of loop structures in the genome.

8.7 Suggestions for Databases and Software

- The Mirny group has produced software for polymer models `http://bitbucket.org/mirnylab/openmm-polymer` that allow one to perform simulations, and recreate contact maps.
- The data from the work of Lieberman-Aiden et al. [330] can be found at `http://hic.umassmed.edu/welcome/welcome.php`. The 3D Genome Browser at Penn State allows one to browse existing Hi-C data sets and to visualize user-generated Hi-C data sets `http://promoter.bx.psu.edu/hi-c/`.
- Large collections of Hi-C data sets can be found at GEP DataSets `www.ncbi.nlm.nih.gov/gds/?term=hi-c`.