UNIVERSITÀ DEGLI STUDI DI PADOVA
Dipartimento di Fisica e Astronomia "G. Galilei"
Corso di Laurea magistrale in Fisica
Percorso teorico e modellistico

Matteo Abbate

Interpretation of chromosome conformation capture data in terms of graphs.

Tesi finale di Laurea magistrale

SUPERVISOR ERASMUS:
Prof. Dieter W. Heermann,
Institute for Theoretical Physics, Ruprecht-Karls Universitaet Heidelberg
INTERNAL SUPERVISOR:
Prof. Flavio Seno

Academic Year 2015/2016

Ai miei genitori.

Contents

1 Introduction

The cell is the fundamental unit that characterizes any living being. From the simplest organisms, the *prokaryotes*, which have only one cell, through the most complex as the human organisms that contain around 100 thousands billions cells in their bodies, the concept of *life* of the cell is always strictly connected to its capacity to duplicate itself.
Duplicating consists of generating another cell that is able to build by itself the same molecules that allow the mother cell to lead all its biochemical processes. All the information necessary to generate these molecules are contained in the *genome* of a cell. It consists of chains of molecules called DNA. During the duplication process, the genome is copied and transferred to the daughter cell. According to the amount of information that the genome can carry, its chains can achieve huge lengths and generally they must be compacted by nearly three orders of magnitude to fit within the limited volume of a cell. Hence, they are organized in compacted structures called *chromosomes*.
Therefore, understanding how chromosomes duplicate and how they are expressed, i.e. how they generate proteins, is fundamental in order to discover how life works. Furthermore, investigating gene expression can provide instruments that are useful to understanding diseases due to genetic mutations. Many studies confirmed that local structural properties of the chromosomes can influence gene expression, DNA replication and repairing [1] [57] [15] [13]. Understanding how chromosomes fold can provide insight into the relationship between the chromatine structure and its functional activity.
Despite its importance, discovering higher order structural features is still complicated due to technical limitations. Some traditional high resolution measurements like electron microscopy or fluorescence are not easily applicable or don't provide information about the entire genome. Other techniques such as light microscopy don't achieve sufficient resolution instead.
Recently a new methodology has been proposed [17]. It provides high-resolution topological information for an entire genome. In fact, the Capturing Chromosome Conformation (or 3C) technique consists of counting how often different loci of a genome interact with each other. The resolution of the experiment has been recently increased and the most recent measurements at higher resolution are referred to as Hi-C method [37].
Hence, this technique yields a set of relative frequencies of contact between different parts of the same genome. These data, usually presented in matrix

form, can be used to analyze the overall spatial organization of the chromosomes. In particular, it is reasonable to assume that the contact frequency between two loci is somehow related to their distance: the bigger the frequency, the smaller the expected distance.
The goal of this work is to obtain a consistent method to reconstruct the three dimensional structure of the chromosome by means of the Hi-C data. In particular, this thesis is inspired by a previous technique that interprets the Hi-C data in terms of graph theory. Graph theory is the theory of mathematical structures called *graphs* which consist of a set of vertexes connected by edges. Its applications are widespread: from physics to psychology, from sociometry to linguistic. In this case, we will apply it to a biological problem. We will consider each locus of the genome as a vertex of the graph and each contact frequency between two loci as a weight to associate to the edge connecting them.
The problem of drawing a graph in a 3 dimensional space is also known as the so-called *embedding problem*. A solution of this problem is the Multidimensional scaling method (MDS) [59] [54]. Given a set of distances between the vertexes of a graph, this method returns an Euclidean set of coordinates. Hence, we will propose different ways to define distances between the vertexes of a graph and we will investigate which are the conditions that it must satisfy in order to solve the embedding problem. In particular, we will stress the use of a matrix called *Laplacian matrix* of the graph, demonstrating that its spectrum and its eigenvectors have some interesting properties. We will show that a distance called *effective resistance* yields to the most consistent reconstruction.
Some simulations of Hi-C data will be performed in order to verify the theoretical statements, focusing on polymers with a linear, circular or rosette conformation.
The thesis work is organized as follows: in the first chapter we will present a theoretical approach to the problem remarking some knowledge of graph theory and proposing some definitions for distances. Then the embedding problem and the multidimensional scaling method are described both generally and in the case of the previously defined distances. In the second chapter some results are presented, showing the consistence of the MDS for simulated polymers and applying it to a set of real Hi-C data. Finally, in the last chapter we will propose some further applications of the MDS method. In particular, we will see how to combine Hi-C data with fluorescence measurements. The latter provide a set of directly measured distances between just

few loci. We will present two different methods, showing their advantages and their limits.

2 Theory

2.1 Chromosome Conformation Capture

The biological unit of any living organism is the cell. It contains all the molecules that take part to the main biological processes, but, most importantly, it is the smallest unit able to replicate independently. This is the basic concept of life. In fact, the cell is able to divide itself in two parts, duplicating some particular molecules that contain all the biological information, i.e. that can then produce all the other useful biomolecules.
The molecules that carries all these genetic instructions are the deoxidribonucleic acid (DNA). They consist of long chains with the shape of a double helix. They are polymers composed by repeating units called nucleotides. There are 4 different kinds of nucleotides and their particular sequence in the chain decodes for the production of specific biomolecules.
Since the DNA carries all the information necessary to the growth of the cell, it contains a huge amount of nucleotides. The length of a DNA sequence is usually measured in base pair (bp), that is a unit consisting of two coupled nucleotides. Each filament contains several millions nucleotides and each cell may contain many filaments of DNA. This means that the length of a DNA chain can be very high and it must be compacted nearly by three orders of magnitude to fit the limited volume of a cell. This is the reason why DNA is not usually found on its own, but it is organized in packaged structures called *nucleoids* in the bacteria [45] and *chromosomes* in the eukaryotes, where they are confined inside the nucleus of the cell. The importance of chromosomes and nucleoids is hence fundamental, because they are the instrument that every living being uses to transmit life.
The number of chromosomes per cell is specific of the organism. The human cell has got 46 chromosomes, the dog's cell 78 and the prokaryotes' cell, monocellular organisms, just one. The shape of the chromosomes is highly dynamic: it is different in each type of organism and changes even during the different phases of the life of the cell. For example, in the prokaryotes the chromosome is circular, while in the human cell it can have an X shape during the metaphase of the cell. In this case the chromosome is said to be acrocentric.
The structural properties and the spatial conformations of chromosomes have been linked with important chromosomal activities [57]. It has been proven that even local high order structural features such as loops, axes, interchro-

mosomal connections have important roles in the genetic expressions, i.e. in the way the DNA is decoded in order to produce biomolecules. For instance, the time of replication has been linked to the spatial disposition of different regions of the chromosomes [1] [15] [13].

Despite its importance, the investigation of the chromosome spatial conformation is limited by technical limitations. In fact, light microscopy doesn't reach a sufficient resolution and electron microscopy, which would allow to have high resolution, is not easily applicable to study specific loci of the chromosome. The fluorescence technique consists in fusing some fluorescent protein with the chromosome and permits the visualization of individual loci, but only few positions can be examined simultaneously. The FISH (Fluorescence In Situ Hybridization) technique permits to visualize multiple loci, but it requires several treatments that may affect the chromosome conformation. A significant contribution to the exploration of the structure of the chromosomes has been given by Dekker at al. in 2002 [17]. They proposed a methodology called Capturing Chromosome Conformation (3C) to characterize some overall physical properties at high resolution.
It consists in isolating the part of the cell where the chromosome are and subject them to a process called formaldehyde fixation (Fig. 1). This process creates cross-links between different parts of the DNA through proteins. A cross-link determines a contact between different segments of the DNA. Then each contact is counted and the relative frequency with which different *loci* have become cross-linked is registered, by means of a reaction called quantitative PCR (Polymerase chain reaction).

In principle, the 3C method required to choose a set of target loci and count the cross links among them. A technique to obtain an unbiased genomewide analysis, i.e. through the entire DNA sequences, was proposed in 2009 [37]. Since it improved the resolution of the 3C measure, it has been called Hi-C. In the following years, the resolution has been further increased [19] [44] and today we may have a precision of 1 kilobase for the human genome.
The results of the Hi-C measure are usually represented in a matrix form. If N is the number of loci analyzed, an $N \times N$ matrix W is used. The entries w_{ij} represent the relative frequency of contacts between the loci i and j. This can be visualized in an heat map (Fig. 2 and 3). The use of these topological information has permitted to individuate some regions of the chromosomes

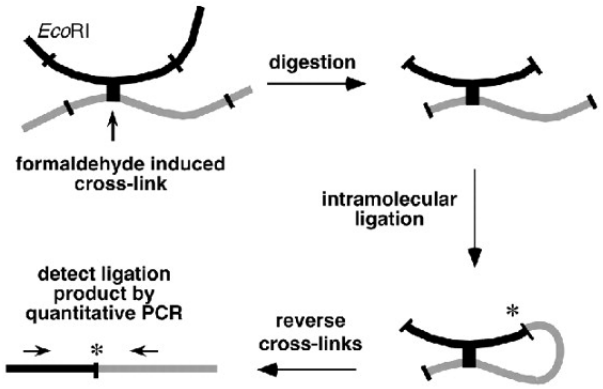

Figure 1: Capturing Chromosome Conformation process. [Adapted from: [17]] Schematic representation of the assay: first the cross-link is created using the formaldehyde. Then, after the EcoRI molecule digestion and intramolecular ligation, the PCR mediates the detection of the ligation products, after reversal of the cross link.

that are closer in space through the analysis of particular patterns of the matrix.

Despite these results, an exact three dimensional chromosomes' conformation has not been given yet. A possible use of these information to arise a 3D reconstruction of the structure has been proposed by Lesne et al. [35]. In this case, the matrix W has been interpreted as the adjacency matrix of a weighted graph. This interpretation may be extended and raised to other applications. Because of this, it will be investigated deeply in the next sections, beginning from some remarks of the general graph theory.

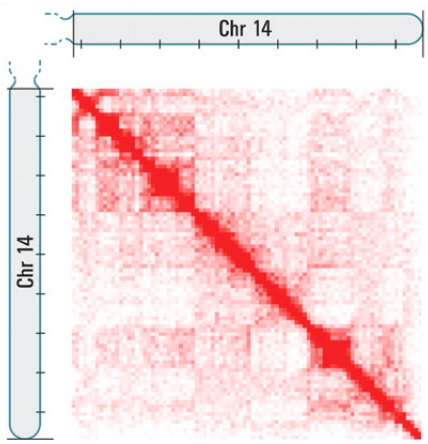

Figure 2: Example of Hi-C output. [Adapted from: [37]] Hi-C produces a genomewide contact matrix. The matrix shows the intrachromosomal interactions on chromosome 14. It is acocentric; the short arm is not shown. The dimension of each pixel is 1Mb locus. Intensity corresponds to the total number of reads (0 to 50).

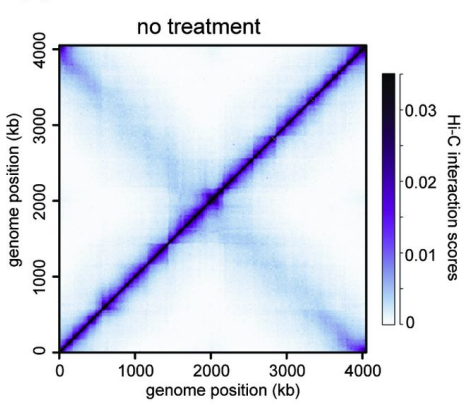

Figure 3: Example of Hi-C output. [Adapted from: [34]] A normalized Hi-C contact map of a bacterial chromosome is presented. The *Caulobacter crescentus* chromosome is circular which can be noted in the off-diagonal intensity. It consists of multiple, largely indipendent spatial domains likely comprised.

2.2 Spectral Graph Theory

A graph $G = (V, E)$ is a set of points V called *vertexes* connected by links called *edges* represented by E.
The graph order is the number n of its vertexes. The graph size is the number m of its edges. The vertex degree is the number of the edges departing from that vertex.
A graph is *undirected* if the edges have not orientation and *directed* or digraph if they have an orientation and may be represented by an arrow (Fig. 4 a-b). A loop is an edge that connects a vertex with itself.
A multiple graph is a graph in which two vertexes can be connected by two or more edges. A simple graph is an undirected graph which does not contain multiple edges or loops. A path between two vertexes i and j is an ordered sequence of consecutive edges starting from i and ending in j.

The structure of a graph is usually represented by means of a matrix called adjacency matrix A whose entries a_{ij} are 0 if vertexes i,j are not connected and 1 otherwise. Obviously, the matrix of an undirected graph is symmetric. Let's give now some additional definitions.
A *complete* graph is a graph in which each pair of vertexes is connected by an edge (Fig. 4 c).
A *weighted* graph is a graph in which each edge is equipped with a number called weight. In this case the adjacency matrix entries w_{ij} correspond to the weight of the ij edge and the vertex degree is defined as the sum of the weights of all its vertexes. The degree matrix K of a graph is a diagonal matrix, whose diagonal consists of the degrees of the vertexes of the graph.
A graph is *connected* if for each pair of vertexes at least one path exists between them. It is *disconnected* otherwise (Fig. 4 d). This means that it is always possible to reach a vertex from another one. A connected component of a graph is a subgraph in which two vertexes are always connected by a path and which is not connected to any other vertex of the graph. A connected graph has got one connected component and a complete graph is always connected.
A *tree* is an acyclic connected undirected graph, i.e. a graph where any two vertexes are connected by exactly one path. A forest is a disjoint union of trees.

Let's now consider a simple complete weighted graph. We will give an

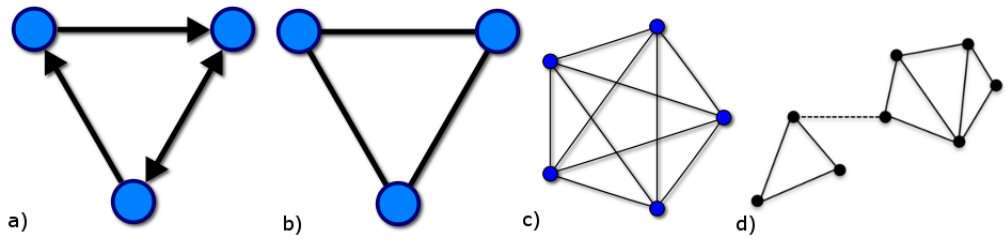

Figure 4: Examples of graphs. a) Directed graph **b)** Undirected graph **c)** Complete undirected graph **d)** Connected graph that becomes disconnected when the thedashed edge is removed.

important definition that is adaptable in the case of unweighted graphs simply substituting each nonzero weight with 1.
The *Laplacian Matrix* of a graph is defined as $L = K - A$:

$$L = \begin{pmatrix} k_1 & -w_{12} & ... & -w_{1n} \\ -w_{12} & k_2 & ... & -w_{2n} \\ ... & ... & ... & ... \\ -w_{n1} & -wn2 & ... & k_n \end{pmatrix}$$

or for each component $l_{ij} = \delta_{ij} k_{ij} - w_{ij}$.
This matrix is important because it allows us to unveil some topological properties of the graph.
The name "*Laplacian*" derives from the fact that the i-th row of the matrix gives the value of the discrete Laplacian operator on the vertex i in N dimensions [5]. In fact, a Laplacian operator (for the sake of simplicity here represented in 3 dimensions and for an unweighted graph) applied to a function $\phi(x, y, z)$ is defined as:

$$\nabla^2 \phi(x, y, z) = \frac{\partial^2 \phi(x, y, z)}{\partial x} + \frac{\partial^2 \phi(x, y, z)}{\partial y} + \frac{\partial^2 \phi(x, y, z)}{\partial z}$$

When the function is defined only for discrete values, the derivative is substituted by the finite differences:

$$\phi(x + 1, y, z) - \phi(x, y, z) = \phi(x, y, z) - \phi(x - 1, y, z)$$

Hence, the Laplacian becomes:

$$\nabla^2 \phi(x, y, z) = \sum_{\alpha,\beta,\gamma} \phi(\alpha, \beta, \gamma) - k\phi(x, y, z)$$

where α, β, γ indicate the coordinates of the neighbors to x, y, z. The relationship between the Laplacian operator and the matrix is clear comparing this result with any sum of the entries of a row of the Laplacian matrix (with a minus sign).
As it will be clear later, the Laplacian matrix is also analogous to the *Laplace-Beltrami* operator on manifolds.

The study of the graph Laplacian eigenvalues is known as spectral graph theory [12] [51]. It has been developed in the past decades in order to to deduce the principal properties and structure of a graph from its Laplacian spectrum, showing that there is an interesting analogy with spectral Riemann differential geometry.
Furthermore, it has been proven that the Laplacian spectra is strictly influenced by topological factors, as clustering, symmetries and degree distribution [39]. We define now the normalized Laplacian as

$$\mathcal{L} = T^{-1/2} L T^{-1/2}$$

where T denotes the diagonal matrix with i-th entry $\sum_j w_{ij}$.

We can view it as an operator in the space of the continuous functions $g : V(G) \to \mathbb{R}$ which satisfies [12]

$$\mathcal{L}g(u) = \frac{1}{\sqrt{k_u}} \sum_{v,u\sim v} \left(\frac{g(u)}{\sqrt{k_u}} - \frac{g(v)}{\sqrt{k_v}} \right) w_{uv}$$

Since $\mathcal{L}$ is symmetric with non negative entries, its eigenvalues are all real and non-negative. By construction, the Laplacian matrix kernel has at least dimension 1, i.e. it has a 0 eigenvalue whose corresponding eigenvector is the constant eigenvector (with all entries equal). We can calculate the other eigenvalues in terms of the Rayleigh quotient of $\mathcal{L}$ [41]. Let g be an arbitrary function that assigns a value $g(v)$ to each vertex v of the graph. Using the notation $\langle , \rangle$ to indicate the standard inner product in $\mathbb{R}^n$, the Rayleigh quotient of $\mathcal{L}$ is then

$$\frac{\langle g, \mathcal{L}g \rangle}{\langle g, g \rangle} = \frac{\langle f, Lf \rangle}{\langle T^{1/2}f, T^{1/2}f \rangle} = \frac{\sum_{u\sim v} w_{uv}(f(u) - f(v))^2}{\sum_v (f(v))^2 k_v}$$

where $g = T^{1/2}f$ and the $\sum_{u\sim v}$ sums up all the pairs of adjacent vertexes. The minimum value of the Rayleigh quotient corresponds to the minimum

eigenvalue of $\mathcal{L}$. It is easy to see it is 0, letting f be a trivial function that assigns a constant value to each vertex. The first smallest eigenvalue after 0 is the the smallest value of the Rayleigh quotient, when f is a function orthogonal to the trivial one.
In the non trivial case the functions f are usually called *harmonic eigenfunctions* of $\mathcal{L}$. For their particular properties, these functions often find many applications, e.g. in video imaging. In fact, in order to solve PDE problems, they can be used as eigenfunctions for a mesh refinement process and they respect the symmetries of the surface of the grid [36].
The above formulation for the non trivial eigenvalue corresponds in a natural way to the eigenvalues of the Laplacian-Beltrami operator [3] that we will define now in the unweighted case. Considering a smooth m-dimensional manifold M embedded in $\mathbb{R}^k$, the Riemann structure on the manifold is induced by the standard Riemann structure on $\mathbb{R}^k$. Let now f be a map $f : M \to \mathbb{R}$, then the Laplace-Beltrami operator is defined as $\mathcal{L} = \nabla \cdot \nabla f$. For the Stokes theorem then:

$$\int_M \|\nabla f\|^2 = \int_M \mathcal{L}(f) f$$

We see that $\mathcal{L}$ is positive semidefinite and the f that minimizes $\int_M \|\nabla f\|^2$ has to be an eigenfunction of $\mathcal{L}$. We further notice then that the gradient ∇f is a vector field on the manifold, such that for small δx:

$$|f(x + \delta x) - f(x)| \approx |\langle \nabla f(x), \delta x \rangle| \leq \|\nabla f\| \|\delta x\|$$

The eigenvalues for the Laplace-Beltrami operator are thus:

$$\lambda_M = \inf \frac{\int_M \|\nabla f\|^2}{\int_M f^2}$$

For a general k-th eigenvalue of the Laplacian matrix, we have

$$\lambda_k = \inf_{f \perp TP_{k-1}} \frac{\sum_{u \sim v} w_{uv}(f(u) - f(v))^2}{\sum_v f(v)^2 k_v}$$

where P_{k-1} denotes the subspace generated by the harmonic eigenfunctions corresponding to the eigenvalues smaller than λ_k. Let's now show the following [12]:

Theorem. If $\lambda_i = 0$ and $\lambda_{i+1} \neq 0$ then G has exactly $i+1$ connected components.

Proof. If G is connected, the eigenvalue 0 has multiplicity 1 since any harmonic eigenfunction with eigenvalue 0 assumes the same value at each vertex. The proof of the theorem follows from the fact that the union of two disjoint graphs has as its spectrum the union of the spectrum of the original graphs.

The energy of a graph is defined as $E = \sum_i^N \lambda_i$ where λ_i is the i-th eigenvalue of the Laplacian [26].
This definition yields to a natural analogy between the Laplacian and the Hamiltonian operator, e.g. in the Schroedinger equation. Each eigenvalue can represent a different energy level and the correspondent eigenmode is analogous to a state function. A similar approach is typical for the quantum graph theory [33]. This suggests that the study of the Laplacian spectrum can be interpreted in terms of quantum repulsion between energy levels [46].

The eigenvalue problem can be formulated also in a useful matrix formalism as a constrained minimization problem. Note that for any vector $\mathbf{y}$ we have:

$$\frac{1}{2}\sum_{i,j}(y_i - y_j)^2 w_{ij} = \mathbf{y}^T L \mathbf{y}$$

In fact, since W is symmetric, we can write:

$$\sum_{i,j}(y_i^2 + y_j^2 - 2y_i y_j)w_{ij} = \sum_i y_i^2 K_{ii} + \sum_j y_j^2 K_{jj} - 2\sum_{i,j} y_i y_j w_{ij} = 2\mathbf{y}^T L \mathbf{y}$$

The problem is formulated in order to find the stationary points of this function under the constraint $\mathbf{y}^T M \mathbf{y} = 1$ for a generic matrix M. The constraint removes an arbitrary scaling factor in the embedding and the matrix M provides a natural measure on the Laplacian eigenvectors of the graph. Let λ be a lagrange multiplier, then the solution of the minimization problem is:

$$L\mathbf{y} = \lambda M \mathbf{y}$$

Let's now analyze what happens with different choices of the metric constraint M:

- choosing $M = \mathbb{1}$ leads to $L\mathbf{y} = \lambda\mathbf{y}$ with the constraint $\mathbf{y}^T\mathbf{y} = \mathbb{1}$, hence it is equivalent to solve the eigenvalues problem for the Laplacian;

- choosing M as a general diagonal matrix whose eigenvectors are $\mathbf{y}$ leads to $L\mathbf{y} = \lambda\lambda_M\mathbf{y} = \lambda_L\mathbf{y}$, hence the lagrange multiplier is equal to $\lambda = \frac{\lambda_L}{\lambda_M}$. In this particular example we can see how $\sqrt{\lambda_M}$ gives naturally a metric to each eigenvector of the Laplacian. In fact, since $L\mathbf{y} = \lambda_L\mathbf{y}$, then we can simply solve the minimization problem as

$$\lambda_M\mathbf{y}^T L\mathbf{y} = \mathbf{y'}^T L\mathbf{y'} = \frac{1}{2}\sum_{i,j}(y'_i - y'_j)^2 w_{ij}$$

  with $\mathbf{y'} = \sqrt{\lambda_M}\mathbf{y}$;

- for some reasons that will be clear later, let's show a particular case in which M is a general matrix symilar to a diagonal matrix, i.e. such that $M = \mathbf{x}^T\Lambda\mathbf{x} = \mathbf{x'}^T\mathbf{x'}$ where $\mathbf{x'} = \Lambda^{1/2}\mathbf{x}$ with the condition $\mathbf{x}^T\mathbb{1}\mathbf{x} = \mathbb{1}$. Then $(\mathbf{xy})^T\Lambda(\mathbf{xy}) = \mathbb{1}$. Suppose now that for some particular eigenvectors $\mathbf{y} = \mathbf{x}/\sqrt{\lambda_M}$ holds, i.e. the eigenvectors of M and L are parallel. In this case the constraint condition is always true and the minimization problems is solved by $L\mathbf{y} = \frac{\lambda}{\sqrt{\lambda_M}}M\mathbf{x} = \lambda\lambda_M\mathbf{y} = \lambda_L\mathbf{y}$ i.e. again $\lambda = \lambda_L/\lambda_M$.
  In the trivial case in which all the eigenvalues of M are equal to 1, we obtain the generalization of the previous case when M is symilar to the identity matrix.

Let's note that, according to the definition of energy of a graph given above and considering the case in which $M = \mathbb{1}$, we have:

$$E = \sum_i \lambda_i^L = \sum_i \mathbf{y}_i^T L\mathbf{y}_i$$

Hence we can redefine the energy in one of the previous general cases as:

$$E = \sum_i \mathbf{y'}_i^T L\mathbf{y'}_i = \sum_i \lambda_i^M\lambda_i^L$$

That can be interpreted as an energy in which the eigenvalues λ_M give a weight to each energetic eigenvalue of the Laplacian, according to the chosen metric.

The eigenvectors will represent the eigenmodes of the system. The first one, corresponding to the mode of null energy, is the constant vector and represents the trivial equilibrium case in which all the points have the same coordinates.

2.3 Distances in graph theory

We will now remark some basic concepts of general topology, beginning with the definition of a distance, in order to apply it later in the graph theory.

Definition. A *metric* or distance on a set X is a function $d : X \times X \to [0, \infty)$ such that satisfies the following conditions:

1. non negativity: $d(x, y) \geq 0$ for any x,y and $d(x, y) = 0$ iff $x = y$;
2. symmetry: $d(x, y) = d(y, x)$;
3. triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$ for any x,y,$z \in X$.

Definition. Given a vector space V a *norm* on V is a function $\| \cdot \| : V \to \mathbb{R}$ such that:

1. $\|av\| = a\|v\|$ for $a \in \mathbb{R}$
2. $\|(v + w)\| \leq \|v\| + \|w\|$
3. $\|v\| = 0$ iff v is the zero vector.

It follows that $\|v\| \geq 0$ for any $v \in V$.

Definition. Two norms are equivalent if $\alpha\|x\|_1 < \|x\|_2 < \beta\|x\|_1$ for some real numbers $\alpha, \beta \geq 0$.

Theorem. All norms are equivalent in a finite dimensional space.

It is possible to define a distance induced by a norm as $d(x, y) = \|(x-y)\|$.

Definition. Given a set X a topology $\mathcal{T}$ is a subset of the partition set $\mathcal{P}(X)$ such that:

1. $0 \in \mathcal{T}$, $X \in \mathcal{T}$
2. The union of elements of $\mathcal{T}$ still belongs to $\mathcal{T}$
3. The intersection of a finite number of elements of $\mathcal{T}$ belongs to $\mathcal{T}$

The elements of $\mathcal{T}$ are called *open* sets.

A metric can naturally induce a topology, simply defining an open set as the set of all elements y such that $d(x, y) < r$ where r is called radius of the open set.

Definition. Two metrics d and D on X are equivalent if all open subsets of X are equal with respect of d and D. They are strongly equivalent if $\alpha d < D < \beta d$ for some real numbers $\alpha, \beta > 0$.

These definitions leads to an important theorem [40].

Theorem. In $\mathbb{R}^n$ all the metrics induced by a norm are equivalent, i.e. they induce the same topology.

This is generally not true if the distance is not induced by a norm. We will now introduce a general way to define a distance not induced by a norm in a discrete space X, hence applicable in the graph space. This method is based on the following definition of proximity matrix. Proximity measures for the vertexes of directed and undirected graphs arise in a wide range of applications, from cristallography to mathematical sociology (applied, for instance, to the social networks).

We will introduce it in the discrete case, since this case will be useful for our applications [10].

Definition. Let X be a discrete set of dimension n. A proximity (or accessibility or connectedness) measure is a function $p : X \times X \to [0, \infty)$

that can be represented by a $n \times n$ matrix P and that satisfies the following conditions for any multigraph:

1. non negativity: $p_{ij} \geq 0$ for any $i, j = 1, \ldots n$;
2. symmetry: the matrix P is symmetric;
3. reversal property: if the graph is directed, the reversal of all its arcs results in the transposition of the proximity matrix;
4. diagonal maximality: for any $i, j = 1, \ldots n$ such that $i \neq j$, $p_{ii} > p_{ij}$ and $p_{ii} > p_{ji}$ holds;
5. triangle inequality for proximities: for any $i, j, k = 1, \ldots n$ $p_{ij} + p_{ik} - p_{jk} \leq p_{ii}$ holds. If then $j = k$ and $i \neq j$ the inequality is strict.

The definition of this matrix will be very useful because, as we will see, sometimes it is easier to define a map that satisfies the above triangle inequality, rather than the distance one. Nevertheless, a simple theorem allows us to construct a metric through the use of a proximity measure.

Theorem. Consider the quantity $d_{ij} = p_{ii} + p_{jj} - p_{ij} - p_{ji}$ with $i, j = 1, \ldots n$. This defines a distance, i.e. it satisfies the axioms of a metric.

In fact, provided that p satisfies the conditions and the triangle inequality of a proximity measure, then d satisfies the triangle inequality for a distance.

In the matrices formalism the distance matrix can be written as:

$$D = \mathbf{1} diag P^T + diag P \mathbf{1}^T - 2P$$

We are now ready to give some definitions of distances for graphs, using both traditional definitions or definition through the use of proximities. In particular we will focus on: shortest path distance, resistance distance, connectedness distances, Laplacian distance, natural proximity distance, entropy distance and potential proximity distance, Katz proximity distance.

Shortest path distance. Let's define the *length* of an edge as the inverse of its weight and the length of a path as the sum of lengths of all the edges belonging to that path. Considering all possible paths between two

```
for k from 0 to N
    for j from 0 to N
        for i from 0 to N
            if dist[i][j]>dist[i][k]+dist[k][j]
                dist[i][j]=dist[i][k]+dist[k][j]
            end if
```

Table 1: Floyd-Marshall algorithm. The algorithm has to be repeated until its convergence, i.e. the triangular disequation is always respected.

vertexes i and j, the shortest path distance is defined as the length of the shortest path connecting i and j. This is the most common distance between vertexes defined in graph theory. Many algorithms have been developed in order to calculate it, as Williams [56], Pettie and Ramachdran [43] or Johnson Dijkstra for directed graph [28].
Nevertheless, the algorithms that will concern this thesis is the Floyd-Warshall algorithm shown in Table 1 [22]. Its time complexity is of order $O(N^3)$. Let's note that this method defines a true distance. The shortest path distance is obviously symmetric, positive and it is 0 if and only if the points are identical. Furthermore, by construction the minimal path length going from i to j is always smaller or equal to any other path length connecting i and j passing through any other vertex k.

We should finally mention that even if the definition and properties of the shortest path distance do not depend on the definition of the length of an arch, this doesn't hold for the particular value it will assume. Hence, it is always possible to find another definition of length of an edge, according to the real meaning of the network and of the weights. For example, another possible choice for the length may be $1/w^a$ where a is a generic real number (bigger or smaller than one).
Expanding with a Taylor series each length, it is straightforward to see that it always respects the distances equivalence inequality, hence the topology induced by these distances is the same. Another interesting example of length definition is known as chemical shortest path distance and has been proposed by [4]. In this case the length of an edge is defined as $l_{ij} = 1 - e^{w_{ij}}$. This approach is based on the fact that, if w_{ij} represents the probability of interac-

tion between i and j, the interaction is governed by a long range percolation.

Resistance distance. [30]. Let's define the resistance of an edge again as the inverse of its weight. The resistance distance between two vertexes i and j is then defined as the equivalent resistance between the points i and j, obtained by considering the graph as an electric circuit. That is, if two resistances are connected in series, the equivalent resistance is the sum of their resistances, if two resistance are connected in parallel, the inverse of the equivalent resistance is equal to the sum of the inverses of their resistances. We can note that this definition is equivalent to the definition of an equivalent spring constant in a system of springs whose elastic constant is equal to the weight of the edges. In this case the rules for the calculus of the equivalent spring constant in series or parallel have to be inverted.
According to this definition, it has been proved [9] that the resistance distance can be calculated using the Moore-Penrose generalized inverse of the Laplacian matrix as a proximity measure [55] [42]:

$$d_{ij}^r = l_{ii}^+ + l_{jj}^+ - 2l_{ij}^+$$

The use of the pseudoinverse of the Laplacian is due to the fact that the Laplacian matrix is never invertible since its kernel has at least dimension 1 in the case of a connected graph. In the only case of *connected* graphs, the pseudoinverse of the Laplacian can be calculated as [10] [52]

$$L^+ = (L + J)^{-1} - J$$

where J is the $n \times n$ matrix with all entries $1/n$.

Let's now note some remarkable facts about the spectrum of the pseudoinverse, using the following theorem [32]:

Theorem. Let A and B be two $n \times n$ hermitian matrices with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_n$ respectively and let $\mu_1 \geq \cdots \geq \mu_n$ denote the ordered eigenvalues of the sum of matrices $A + B$. The following properties hold:

1. $\sum \mu_i = \sum \nu_i + \sum \lambda_i$
2. $\nu_1 \leq \mu_1 + \lambda_1$

3. $\nu_{i+j+1} \leq \lambda_{i+1} + \mu_{j+1}$ with $0 \leq i, j, i+j < n$

Since the eigenvalues of J are 0 with multiplicity n-1 and 1 with multiplicity 1, using the third property of the previous theorem we find that the eigenvalues ν of the matrix $(L+J)$ the inequality $\nu_i \leq \min\{\lambda_i+1, \lambda_{i-1}\}$ must hold. Let O_i denote the positive difference $\lambda_{i-1}-\nu_i$. From the conservation of the trace in the previous theorem, O_i is null if $\min\{\lambda_i+1, \lambda_{i-1}\} = \lambda_{i-1}$, consequently $\nu_1 = \lambda_n+1$. In the general case $\nu_1 = \lambda_n+1+\sum_i(\lambda_{i+1}-\nu_i) = 1+O_{tot}$. That means that the 0 eigenvalue of the Laplacian has been replaced with a positive quantity bigger than 1. Taking the inverse of this matrix and repeating the same calculus for the matrix $(L+J)^{-1} - J$ we find that the eigenvalues are $\lambda_i^+ \leq 1/\nu_i$ with $i \neq n$ and $\lambda_{min}^+ = \frac{1}{\lambda_{min}+1+O_{tot}} - 1 + O'_{tot}$. For small O_{tot} and O'_{tot} the non-zero eigenvalues of the pseudoinverse are equal to the inverse of the non-zero eigenvalues of the Laplacian and the zero eigenvalue remains zero.

In our specific case, we can further use the following theorem, usually common in quantum physics.

Theorem. If two hermitian matrices commute, they share a common eigenvectors basis.

The matrices we consider are real and symmetric. The matrix J in particular commutes with any matrix, hence they share a common basis of eigenvectors. Then it is easy to show that if two matrices share the same eigenvectors, their sum will have the same eigenvectors too and its eigenvalues will be the sum of the corresponding eigenvalues. In our case $\nu_1 = \lambda_{min} + 1$ and $\nu_i = \lambda_i$, because λ_{min} and 1 are the eigenvalues of the matrices L and J, respectively, that correspond to the constant eigenvector.
Hence, the Laplacian and its pseudoinverse have the same eigenvectors basis and the corresponding non zero eigenvalues are one the inverse of the other. The eigenvector corresponding to the 0 eigenvalue is the constant vector for both the matrices.

Theorem. [30] For any pair of vertexes i, j in G, $d_{short} \leq d_{res}$ with equalities true iff there is only a single path between i and j.

Proof. Let π be the shortest path between i and j. Increasing a resistance for any edge e not in π, the resistance distance strictly decrease. Letting all the resistances $r_e \to \infty$ for all e not in π, then $d_{res} \to d_{short}$. However, then $d_{res} \geq d_{short}$ with equality only if there is not any edge e out of π.

Corollary. The shortest path and the resistance distances are the same between every pair of vertexes of a connected graph iff the graph is a tree.

As the above theorem suggests, it is possible to prove the existence of a relationship between these two distances. In particular, both of them belong to a generalized class of distances called forest distance d_α [7]. It is a one-parametric family of distances which reduce to the shortest path and the resistance distances at the limiting values of the parameter α.
Let $Q_\alpha = (\mathbb{1} + L_\alpha)^{-1}$, where α is a real parameter and L_α is the Laplacian of the graph obtained from G, applying a certain transformation of the edges' weights depending on α. For instance, a possible choice is $L_\alpha = \alpha L$ where each weight has been multiplied by α. Let γ be a positive factor. We define the matrix H_α as $H_\alpha = \gamma(\alpha - 1)Q_\alpha^*$ where Q_α^* is the matrix obtained from Q_α taking the logarithm to base α of each entry of Q_α.

Theorem. For any connected multigraph G and any α, $\gamma > 0$, the matrix D_α whose entries are $d_{ij} = 1/2(h_{ii} + h_{jj}) - h_{ij}$, is a matrix of distances on $V(G)$.

We are interested in some consequences of this theorem in a particular case. In fact, consider the edge transformation $w_{ij}(\alpha) = \alpha w_{ij} e^{(-1/(w_{ij}\alpha))}$ and a positive scaling factor γ such that $\lim_{\alpha\to 0^+} \gamma(\alpha) = 1$ and $\lim_{\alpha\to\infty} \gamma(\alpha) = 2/n$ (e.g. $\gamma(\alpha) = (2/n\alpha + \beta)/(\alpha + \beta)$ where $\beta > 0$ is a parameter). Then, the following theorem for *connected* graphs is true.

Theorem. For any connected multigraph G and any $i, j = 1, \ldots n$, the distance $d_\alpha(i, j)$ defined above converges to the shortest path distance as $\alpha \to 0^+$ and to the resistance distance as $\alpha \to \infty$.

Hence, both the shortest path distance and the resistance distance fit, as limiting cases, into the framework of generalized logarithmic forest distances. The proof of this theorem is based on the fact that in these limits H_α converges to a matrix of spanning rooted forests of the graph that can

be considered as a proximity measure [8].

The resistance distance can also be directly related to the eigenvalues and eigenvectors of the Laplacian matrix [58]. In fact, let u_{ik} denote the i-th entry of the k-th eigenvector of L. Expressing the pseudoinverse Laplacian entries as $l^+_{ij} = \sum_k \frac{1}{\lambda_k} u_{ik} u_{jk}$ we find:

$$d^r_{ij} = l^+_{ii} + l^+_{jj} - 2l^+_{ij} = \sum_k \frac{1}{\lambda_k}(u_{ik} - u_{jk})^2$$

Connectivity-based distances. [18]. The strength s of a a path π of n edges is defined as the minimum weight among all the weights of the edges of the path: $s = \min w(e_i)$. The strength of connectedness between two vertexes i and j is defined as the maximum strength among all possible paths between i and j:

$$CONN_G(ij) = \max\{s(\pi) : \pi \text{ is a path between } i \text{ and } j\}$$

The calculus of the strength connectedness is easily recognizable as a maximum flow problem. That is, it is possible to consider each weight of the edge as the maximum amount of a flow that can pass through the edge and the strength of connectedness is thus the maximum amount of flow that can pass from a source i to j or viceversa. We can define some distances based on the strength of connectedness.
The *strongest connectivity distance* is defined as $d_{ss}(ij) = 1/CONN_G(ij)$ and $d(ii) = 0$. If i and j are not connected by any edge, then $CONN_G(ij) = 0$ and $d = \infty$. By this definition it is straightforward to see that d_{ss} is positive, symmetric and equal to 0 iff $i = j$. Note that for any three vertexes i, j, k the inequality $CONN(i, j) \geq \min\{CONN(i, k), CONN(k, j)\}$ holds since $CONN(i, k)$ can not be smaller than the strength of any path between i and j. This yields:

$$\frac{1}{CONN_G(ij)} \leq \frac{1}{\min\{CONN_G(ik), CONN_G(kj)\}} \leq \frac{1}{CONN_G(ik)} + \frac{1}{CONN_G(kj)}$$

That is $d_{ss}(i, j) \leq d_{ss}(i, k) + d_{ss}(k, j)$.
The δ-distance is defined as $d_\delta(ij) = 1 + \Delta - CONN(i, j)$ where Δ is the maximum weight of all arcs and $d_\delta(ii) = 0$ for any i.

The distance d_δ is symmetric and since $CONN \leq \Delta$ holds always, then $d_\delta(ij) \geq 0$ for any i, j and it is 0 iff $i = j$. Then $1 + \Delta - CONN < 1+\Delta-\min\{CONN(ik), CONN(kj)\}$ and therefore $d_\delta(ij) < d_\delta(ik)+d_\delta(kj)$. The disadvantage of the use of these distances is that the computational cost in terms of time of the algorithms proposed to solve the maximum flow problem for *undirected* graphs is high. For instance, it is necessary to apply the Ford-Fulkerson algorithm for any possible pair of vertexes of the graph [23] [48] .

Laplacian distance. According to the definition of Laplacian of a graph given above, one may want to define a distance consistently to the Laplacian minimization problem. Provided that the problem is solved and the eigenvectors matrix is known, the distance between two points i and j will be $d_{ij} = \sqrt{\sum_r^n (x_i^r - x_j^r)^2}$.
We notice that, considering the case in which $M = \mathbb{1}$, from the orthogonalization constraint condition $X^T X = 1$ we obtain

$$d_{ij} = \sqrt{\sum_k^r (x_{ri}^2 + x_{rj}^2 - 2x_i^r x_j^r)} = \sqrt{2}\delta_{ij}$$

. This is a trivial case in which all distances are equals (hence they still respect all the metric conditions). Another choice of M, for instance a non trivial diagonal matrix, can lead to other results. A useful definition of M should in any case give a consistent description of the system. This problem is tautological: as we will see in the next section, a consistent metric matrix M can be constructed if a distance matrix is already previously defined.
With the so-called *algebraic distance* a similar definition has been proposed [11]. In particular, it has been shown that it is possible to calculate how fast the distances converge to their constant value using a computation algorithm.

Natural proximity distance. A natural definition of a proximity measure is the following: $p_{ij} = w_{ij}$ and $p_{ii} = \sum_j w_{ij}$. In this case, in fact $p_{ij} + p_{ik} - p_{jk} \leq p_{ii} = \sum_{i'} p_{ii'}$ since among all $p_{ii'}$ we consider also p_{ij} and p_{ik}. It is clearly a proximity measure and hence it will be used to compute a distance matrix of the graph.

Entropy distance. Let W be a set of probabilities of mutually exclusive events. The Shannon entropy is defined as a function that describes the

uncertainty associated to this probability distribution [49]. In particular:

$$S(w_1, \dots w_n) = -\sum_k w_k log w_k$$

We can apply it to the graph theory if we consider each entry of the adjacency matrix as one of the probability measures of W:

$$S = -\sum_i \sum_{j>i} w_{ij} log w_{ij}$$

The entropy proximity measure p_{ij} is defined as the increase of the Shannon entropy of the graph when an edge of weight w_{ij} is added: $p_{ij} = \Delta S_{ij} = -w_{ij} log w_{ij}$. Analogously the entropy proximity measure of a vertex is the increase of entropy obtained adding that vertex to the graph, i.e. adding all the edges linked to that vertex: $p_{ii} = \sum_j (-w_{ij} log w_{ij})$. Since $w_{ij} \leq 1$, this proximity measure is nonnegative and symmetric. By construction, it respects the diagonal maximality and it is easy to prove that it respects also the triangle inequality for proximities. Hence, it is possible to define a distance using the transformation between proximities and distances.
The main characteristic of Shannon entropy is to represent the amount of disorder of a system due to a lack of information. That is, it is maximum when the probability that an event occurs is 0.5 but it is null when we are able to say for sure that an event is impossible ($w_{ij} = 0$) or certain ($w_{ij} = 1$). The shape of this function is illustrated in Fig. 5a. Hence, the interpretation of the Shannon entropy as a proximity measure may not be physically meaningful. In fact, an additional condition that is rather natural to require to the proximity measure, provided that it respects the other ones, is the monotonicity:

Definition. *Monotonicity.* If the weight w_{kt} of some edges in a multigraph increases or a new edge is added from k to t, then:

1. $\Delta p_{kt} > 0$ and for any $i, j = 1 \dots n$ $i, j \neq k, t$ implies $\Delta p_{kt} > \Delta p_{ij}$;

2. for any i if there is a path from i to k and each path from i to t includes k, then $\Delta p_{it} > \Delta p_{ik}$;

3. for any i_1, i_2, if i_1 and i_2 can be substituted for i in the hypothesis of item 2, then $p_{i_1 i_2}$ does not increase. Thus, the proximity between two

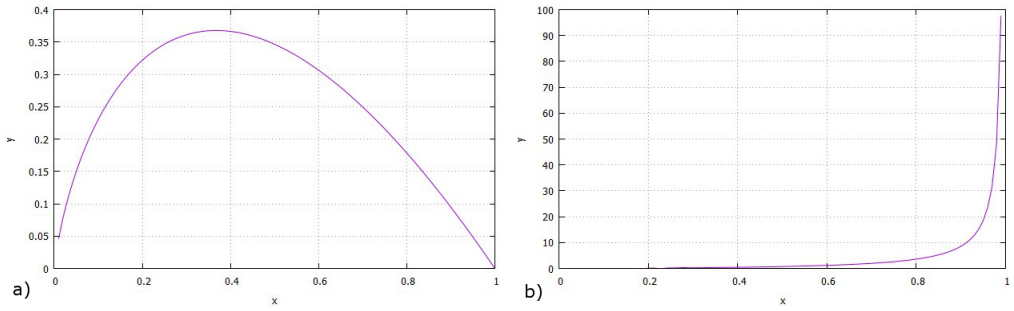

Figure 5: Curves of proximities. The curves show how the proximity changes with the weights, in the case of the Shannon entropy proximity (**a**) and in the potential proximity (**b**) in which the monotonicity condition is satisfied.

vertexes does not increase whenever the bond that is added or increased is extraneous for the connection of these two vertexes.

This requirement may be interpreted as the request that the proximity measure is maximum when the probability that an event occurs is 1 and minimum when the event is impossible.
This condition is clearly not respected by the entropy proximity measure and we will modify it in order to let it be monotonic.

Potential proximity measure. With respect to the classical thermodynamics, we can define a function $U : P \to \mathbb{R}$ called *potential*, such that $\Delta S = -U/T$ where T is the temperature of the system that will be set to 1 for the sake of simplicity. We can interpret the probability w_{ij} as a specific mass that drives the attraction between i and j and therefore it is an intrinsic property of both the vertexes i and j. Then the potential U can be defined as a central potential: $U_{ij} = -w_{ij}^2/p_{ij}$. It yields: $\Delta S_{ij} = w_{ij}^2/p_{ij}$ and using the definition of entropy introduced in the case of the entropy distance, we obtain:

$$p_{ij} = \frac{w_{ij}}{log(1/w_{ij})}$$

.
This proximity measure called *potential proximity measure* is positive, symmetric and respects the triangular inequality and the diagonal maximality.
Furthermore, as shown in Fig. 5b, this proximity measure satisfies the reasonable request of monotonicity, diverging in the case in which an event is certain.

It is important to notice that if we had defined directly $U_{ij} = w_{ij}^2/d_{ij}$, we would have obtained a definition of distance that *a priori* was not said to respect the triangular inequality.

Katz distance. This distance is defined using the Katz similarity matrix Q of the graph as proximity measure. This matrix has been proposed in the social sciences field or sociometric, f.i. to compute the popularity of a person [29] [21] [14]. In an unweighted graph, the entries of the Katz matrix represent the number of possible paths connecting the two correspondent vertexes.
The number of paths of length 1 is represented by the adjacency matrix A. The elements of the n-th power of A indicate the numbers of paths with length n. For instance the entry of A^2 are $a_{ij}^{(2)} = \sum_r a_{ir}a_{rj}$ and each component $a_{ir}a_{rj}$ is equal to 1 only if i is linked to r and r is connected to j. Hence, the matrix of the number of all possible paths connecting two vertexes is:

$$Q = A + A^2 + A^3 + ... = \sum_k A^k = (\mathbb{1} - A)^{-1} - \mathbb{1}$$

In the case of a weighted graph, this matrix represents the probability of connection between the vertexes. The proximity matrix is obtained by taking Q and adding a diagonal matrix whose entries are the sums of the correspondent rows of Q.

2.4 The embedding problem

In the previous section we have seen how it is possible to define distances between the vertexes of a graph. In this section we will explore the ways and the conditions to obtain a set of Cartesian coordinates in a generic r dimensional space from a complete set of distances among N points, with $r \leq N$. This problem is known as *bound embedding problem* and in the graph theory it has been used to draw a graphical representation of the network. Further in this thesis, we will stress the use of the distances defined above in order to reproduce a three dimensional configuration of the graph.

Solutions to the the bound embedding problem have been proposed since 1938 [60] and are often used even in mathematical psychology, marketing,

sociology, political science to obtain a geometrical representation of the similarities among objects. In psychology, for example, the set of data consists in similarities of human judgments and the more similar they are the closer their correspondent points will be in the embedding multidimensional space. This the reason why the method is often referred to as Multidimensional Scaling (MDS) [59].

Let X be a set of N points of unitary mass and D be an $N \times N$ matrix of distances among them. D is symmetric, its diagonal elements all equal zero and for any $i, j, k = 1 \dots N$ the inequality $D_{ik} \leq D_{ij} + D_{jk}$ holds. We aim to find an n dimensional set of Cartesian coordinates for the N points.

Theorem. [16] The distance between the barycenter or center of mass, 0, of each point in any Euclidean space in terms of the remaining distances is given by:

$$d_{0i}^2 = \frac{1}{N} \sum_j D_{ij}^2 - \frac{1}{N^2} \sum_{k>j=1} D_{jk}^2$$

Proof. Let $\mathbf{r}_{lk}$ denote the vector from point l to k. From the definition of the center of mass $\sum_j \mathbf{r}_{0j} = 0$ and since $\mathbf{r}_{0j} = \mathbf{r}_{0i} + \mathbf{r}_{ij}$ for any i we have $\sum_j (\mathbf{r}_{0i} + \mathbf{r}_{ij}) = 0$. Hence $\mathbf{r}_{0i} = -\sum_j \mathbf{r}_{ij}/N$.

$$D_{0i}^2 = \mathbf{r}_{0i} \cdot \mathbf{r}_{0i} = \frac{1}{N^2} \sum_j \sum_k \mathbf{r}_{ij} \cdot \mathbf{r}_{ik}$$

By the law of cosines:

$$\begin{aligned} D_{0i}^2 &= \frac{1}{2N^2} \sum_j \sum_k (D_{ij}^2 + D_{ik}^2 - D_{jk}^2)^2 = \\ &= \frac{1}{2N^2} \left[2(N-1) \sum_{j=2} D_{ij}^2 - 2 \sum \sum_{j<k} D_{jk}^2 \right] = \\ &= \frac{N-1}{N^2} \sum_{j=2} D_{ji}^2 - \frac{1}{N^2} \sum \sum_{j<k} D_{jk}^2 = \\ &= \frac{1}{N} \sum_j D_{ji}^2 - \frac{1}{N^2} \sum \sum_{k>j=1} D_{jk}^2 \end{aligned}$$

Let now X be the matrix $n \times N$ of the n Cartesian principal axes coordinates of the N points and $\lambda_k = \sum_{i=0}^{N}(x_{ik})^2$ be the moment along the k-th coordinate axis. Define the diagonal matrix Λ whose entries are the moments and the matrix $n \times N$

$$Y = \Lambda^{-1/2} X$$

Let's extend Λ to an $N \times N$ matrix adding $N - n$ rows and columns of zeros and consequently extend also Y to a $N \times N$ matrix adding $N - n$ rows derived from the first n rows by Gram-Schmidt orthogonalization. By construction, Y is unitary: $Y^T Y = 1$. Let's finally define the Gram matrix $G = X^T X = Y^T \Lambda Y$. This matrix is hence positive semidefinite of rank n and its eigenvalues are the moments of the distribution. Using the low of cosines as done in the proof of the previous theorem we can write

$$X^T X = \frac{1}{2}[d_{0i}^2 + d_{0j}^2 - D_{ij}^2] = M$$

This is also called *metric matrix* and coincides with the Gram matrix of the N points [27]. It is possible to write the matrix M also in the form:

$$M = \frac{1}{2} H' D^2 H$$

where H is the matrix $\mathbb{1} - 1/NJ$, $H' = -H$ and J is the matrix with all elements equal to 1. The entries of H are $h_{ij} = \delta_{ij} - 1/N$. In fact:

$$\begin{aligned} m_{ij} &= \frac{1}{2} \sum_{i'j'} \left(-\delta_{ii'} + \frac{1}{N} \right) D_{i'j'}^2 \left(\delta_{j'j} - \frac{1}{N} \right) = \\ &= -\frac{1}{2} D_{ij}^2 + \frac{1}{2} \sum_{j'} \frac{D_{ij'}^2}{N} + \frac{1}{2} \sum_{i'} \frac{D_{i'j}^2}{N} - \frac{1}{2} \sum_{i'j'} \frac{D_{i'j'}^2}{N^2} = \\ &= \frac{1}{2}[d_{0i}^2 + d_{0j}^2 - D_{ij}^2] \end{aligned}$$

Theorem. The sum of the rows or the columns elements of the metric matrix M is zero.

Proof.

$$\sum_j m_{ij} = \sum_j \left(-\frac{1}{2} D_{ij}^2 + \frac{1}{2} \sum_{j'} \frac{D_{ij'}^2}{N} + \frac{1}{2} \sum_{i'} \frac{D_{i'j}^2}{N} - \frac{1}{2} \sum_{i'j'} \frac{D_{i'j'}^2}{N^2} \right) = 0$$

Corollary. The metric matrix kernel dimension is at least 1 and it is not invertible.

Proof. From the previous theorem it is shown that the dimension of the row or column space of M is maximum N-1. So does the rank.

The following theorem shows which are the conditions for the D_{ij} to be the distances of an n-simplex lying in a Euclidean space $\mathcal{R}^r$ of dimension r but not in a Euclidean space $\mathcal{R}^{r-1}$ of dimension $r-1$.

Theorem. [47] A necessary and sufficient condition that the D_{ij} are the lengths of the edges of an n-"simplex" lying in $\mathcal{R}^r$ but not in $\mathcal{R}^{r-1}$ is that the quadratic form

$$F(x_1, \dots x_n) = \frac{1}{2} \sum_{i,j} (d_{0i}^2 + d_{0j}^2 - D_{ij}^2) x_i x_j$$

is positive (i.e. always ≥ 0) and of rank r. That is, the associate matrix M has to be positive semidefinite of rank r.

Proof. See Appendix A.

The definition of the metric matrix M given above is actually a particular case of a more general case in which an isometric transformation of the coordinates is taken into account, i.e. a translation of the origin or a rotation of the axes. The consequence of such a transformation for the metric matrix is that it won't be anymore calculated relatively to the barycenter of the system but to another point. In particular, it implies the following theorem, that we won't demonstrate for the sake of synthesis.

Theorem. [24] Given $M = (\mathbb{1} - \mathbf{1}\mathbf{s}^T) D^2 (\mathbb{1} - \mathbf{s}\mathbf{1}T)$ with $\mathbf{s}$ a N-dimensional vector, the distance D is Euclidean iff M is semidefinite positive for any $\mathbf{s}$ such that $\mathbf{s}^T\mathbf{1} = 1$ and $\mathbf{s}^T D^2 \neq 0$.

If we chose $\mathbf{s} = 1/N\mathbf{1}$ we obtain the case described before. In this case H is also called *geometric centering matrix*.

Hence, the rank of the metric matrix allows us to know the minimum dimension in which it is possible to embed the graph. If the set of distance

would be measured directly in a three dimensional space then we should expect to find a metric matrix with three positive eigenvalues which coincide with the three principal moments of the inertia tensor and N-3 eigenvalues equal to zero. Nevertheless, it should be considered that measure or computational errors may cause some eigenvalues to be not exactly zero or even negative [53]. If we used a set of distances defined in the N dimensional space X of the vertexes of the graph, we could obtain even a rank of dimension N-1. The i-th coordinate of the k-th point is then $X_{ki} = \sqrt{\lambda_k} Y_{ki}$ where Y_k is the k-th eigenvector of M. It represents the first principal axis of the set of points.

As mentioned, because of some calculus errors some eigenvalues may result to be negative. This would not allow to take the root squared of them. The literature on the solution of this problem is divided. Some [16] use to order the eigenvalues in decreasing order of magnitude and hence take the root squared of their absolute values. Others [35] simply set to zero all the negative eigenvalues, implicitly reducing the dimension of the space. In this thesis, for the future calculus we will consider the latter choice in order to avoid to consider some negative eigenvalues (due to approximation errors) as positive and maybe influential.

The Euclidean distance between two points will be: $d_E = \sqrt{\sum_k^r \lambda_r (x_r^i - x_r^j)^2}$. Hence, it strictly depends on the amount of nonzero eigenvalues. In the future we will use the notation $O(\lambda) = d_N - d_3$ to indicate the difference between the distance in the N dimensional space and distance obtained by using only the first three coordinates.

Therefore, the role of the eigenvalues is to engage each direction of this Euclidean space of a natural metric measure proportional to the corresponding momentum of inertia in that direction. In terms of graph theory, it seems natural to choose this metric as a metric constraint for the Laplacian minimization problem.

It becomes evident that a matricial form to express the Euclidean distance is [20]

$$D = \mathbf{1} diag X^T X + diag X T X \mathbf{1} - 2X^T X$$

That means that if M is semidefinite positive, it is naturally a proximity measure for the Euclidean distance. The viceversa is not always true. A proximity measure, in fact, is not always an Euclidean metric, in the sense that it is not always possible to write $P = X^T X$ for some Cartesian coordinates X. In any case, it is always possible to construct a distance D from P,

verify that it is euclidean constructing M and check that it is semidefinite positive and deduce the coordinates from $M = X^T X$ with the multidimensional scaling method.
In fact, the equation $d_{ij} = p_{ii} + p_{jj} - 2p_{ij} = m_{ii} + m_{jj} - 2m_{ij}$ is a necessary but not sufficient condition to have $M = P$.

Nevertheless, M and P can share some properties, under certain conditions on P.

Theorem. M and P share a common set of eigenvectors iff

$$\sum_r \left[\sum_{l,k} p_{lr} - p_{lk} \right] p_{ri} = c$$

with c a real constant for any $i = 1, \dots N$.

Proof. We will show the conditions to let M and P commute. Let $P*$ be the matrix $P^* = diagP\mathbf{1}^T + \mathbf{1}diagP^T$ whose entries are $p^*_{ij} = p_{ii} + p_{jj}$.Then:

$$M = -\frac{1}{2} H D^2 H = -\frac{1}{2}(\mathbb{1} - \frac{1}{N} J)(P^* - 2P)(\mathbb{1} - \frac{1}{N} J)$$

Hence each element of M is:

$$\begin{aligned} m_{ij} &= -\sum_{i'j'} (\delta_{ii'} - \frac{1}{N*})(p_{i'i'} + p_{j'j'} - 2p_{i'j'})(\delta_{j'j} - \frac{1}{N}) = \\ &= -2 \left(\frac{\sum_l (p_{lj} + p_{il})}{N} - \frac{\sum_{lk} p_{lk}}{N^2} - p_{ij} \right) \end{aligned}$$

Let's now calculate the element c_{ij} of the matrix MP and the element c^*_{ij} of the matrix PM.

$$\begin{aligned} c_{ij} &= -\frac{2}{N} \sum_r \left[\sum_l (p_{il} + p_{rl}) - \frac{\sum_{lk} p_{lk}}{N} - p_{ir} \right] p_{rj} = \\ &= -\frac{2}{N} \left[\left[\sum_r p_{rj} \right] \left(\frac{p_{il} - p_{lk}}{N} \right) + \sum_{lr} p_{lr} p_{rj} - \sum_r p_{ir} p_{rj} \right] \end{aligned}$$

$$c^*_{ij} = -\frac{2}{N}\sum_r p_{ir}\left[\sum_l (p_{rl}+p_{jl}) - \frac{\sum_{lk} p_{lk}}{N} - p_{jr}\right] =$$
$$= -\frac{2}{N}\left[\left[\sum_r p_{ir}\right]\left(\frac{p_{jl}-p_{lk}}{N}\right) + \sum_{lr} p_{lr}p_{ir} - \sum_r p_{ir}p_{rj}\right]$$

It is easy to verify that the condition $c_{ij} = c^*_{ij}$ holds iff the condition expressed by the theorem is satisfied.

Corollary. If the sum of the rows of the proximity matrix P is constant, M and P share the same eigenvectors basis.

This means that this additional condition given to the proximity matrix P assures that it can be a metric matrix. In particular, the requirement that the sum of its rows is constant is equivalent to require that one of its eigenvector is the constant eigenvector $\mathbf{1}$. In fact, it is known from elementary algebra that:

Theorem. Given a matrix A, the sum of the elements of any row is constant iff one of the eigenvectors of A is the constant vector.

Finally, let's note that even if the same distance can be calculated using the matrix P or the matrix M, it represents two different distances, because it is applied to two different spaces. The Euclidean distance is a metric induced by a norm in a finite r-dimensional space, hence topologically equivalent to any other metric induced by a norm in this space. The distance calculated by P instead is defined in the graph space. They are not said to be topologically equivalent, but it has to be opportunely demonstrated. To do so, since the dimensions of the spaces are different, we should consider the topology induced by the topology in the N-dimensional Euclidean space for the r-dimensional space .

Definition. Let U be a subset of $\mathbb{R}^r$ and $\mathbb{R}^r \in \mathbb{R}^N$ and let also $\mathcal{T}$ be a topology in $\mathbb{R}^N$. $\mathcal{T}_r$ is the topology induced by $\mathcal{T}$ in $\mathbb{R}^r$ and U is said to be an open set of $\mathcal{T}^r$ iff there exists an open subset V of $\mathbb{R}^N$ such that $V \cap \mathbb{R}^r = U$.

If there exists an embedding $i : \mathbb{R}^r \to \mathbb{R}^N$ an induced topology always exists. So it is true for projections.

2.5 Embedding of a graph

An Euclidean reconstruction of the coordinates for a set of N vertexes of an undirected weighted graph has already been performed in 2014 by Lesne et al. interpreting the matrix Hi-C data as the adjacency matrix of the graph [35]. In that case, they calculated the Gram matrix simply using the shortest-path distance. We will now focus on the description of the metric matrix obtained by the use of different distances in order to give an interpretation of the Euclidean structure built in this way. In particular, we will focus now on the shortest path, resistance, natural proximity and katz distance because in all these cases we can show that the metric matrix has the same eigenvectors of the Laplacian.

For the resistance distance the proximity measure matrix is given by the pseudoinverse of the Laplacian.

Theorem. The Metric matrix obtained by using the resistance distance shares an eigenvector basis with the Laplacian of the graph.

Proof. As we have already shown, the pseudoinverse of the Laplacian and the Laplacian itself have the same eigenvectors. The pseudoinverse of the Laplacian is the proximity measure for the resistance distance. Since one of its eigenvectors is the constant vector, the sum of its rows is constant and consequently it has the same eigenvectors of the metric matrix M.

Theorem. The Metric matrix obtained by using the shortest path distance shares an eigenvector basis with the Laplacian of the graph.

Proof. We will show that the Laplacian L and the matrix Q^*_α commute when $\alpha \to 0$. We can represent the matrix $Q_\alpha = (\mathbb{1} + L_\alpha)^{-1}$ as a series:

$$Q_\alpha = -[I + \alpha L + \alpha^2 L^2 + ...] = -[I + \alpha X]$$

where $X = [L + \alpha L^2 + ...]$.
Each element of the matrix Q^*_α is $q^*_{\alpha ij} = -\log_\alpha(\delta_{ij} + \alpha x)$. Then $q^*_{ij} = \lim_{\alpha\to 0} q^*_{\alpha ij} = \delta_{ij} - 1$. Hence $\sum_r l_{ir} q^*_{rj} = N l_{ij} - \sum_r l_{ir}$ and $\sum_r q^*_{ir} l_{rj} =$

$Nl_{ij} - \sum_r l_{rj}$. They are equal since the sum of the rows of the Laplacian matrix is constant.

Theorem. The Metric matrix obtained by using a natural proximity distance shares an eigenvector basis with the Laplacian of the graph.

Proof. The natural proximity measure matrix P commutes with L. In fact, denoted with c_{ij} the elements of PL and with c^*_{ij} the elements of LP, we note that they are equal:

$$c_{ij} = -\sum_{r \neq i,j} w_{ir} w_{rj} - \sum_l w_{il} w_{ij} + w_{ij} \sum_l w_{jl} = c^*_{ij}$$

Theorem. The Metric matrix obtained by using a Katz proximity distance shares an eigenvector basis with the Laplacian of the graph.

Proof. The Katz matrix $Q = (\mathbb{1} - W)^{-1} - \mathbb{1}$ has the same eigenvectors of W and W commutes with its Laplacian.

The multidimensional scaling reconstructs the coordinates, simply multiplying the eigenvectors of M for the squared root of the correspondent eigenvalue. This means that the N points will be distributed in the space along each axes as the principal axes of M. If these corresponds to the eigenvectors of L, the principal axes of M corresponds to the eigenmodes of the Laplacian of the graph. This equivalence is particularly interesting if we consider the system as a system of N points connected by springs with the weight of the correspondent edge as an elastic constant. Using M as the constraint metric in the Laplacian minimization problem we can then interpret the energy of the graph as the minimal energy of the springs system with the given distances as elongation. Furthermore, considering the energy of a rigid body

$$E = \frac{1}{2} I w$$

where I is the principal moment of inertia, in our case λ_M, and w is the frequency of the mode, this is analogous to the energy of the graph. In fact, in the case of the springs the frequency of the mode is equal to the elastic constant divided by the (unitary) mass.

3 Results

3.1 Introduction

In order to verify the correctness of the theory described in the previous chapter, we stressed the use of some simulations of Hi-C data for artificial polymers. In particular, the used polymers have been obtained with a lattice-bound Monte Carlo simulation of a chain in normal conditions, i.e. without any external potential. A simple excluded volume model interaction between the subunits was simulated, with a self-avoiding walk [6]. The length of each mesh of the lattice was fixed to 10 *nm*.
In particular, three different shapes of the polymer were simulated:

- *linear polymer* consisting of 300 monomers;
- *circular polymer* consisting of 400 monomers;
- *rosette polymer* consisting of 401 monomers and organized in 4 different petals with a common center.

For all of them around 15000 conformations have been simulated. For each single one the distance matrix between the loci has been computed. Any time a distance was lower than a fixed threshold, a contact between the points was counted. The threshold is the maximum possible bond length between two points and its mean is $\sqrt{10} \approx 2.7$. In this way a matrix of the relative frequencies of contacts has been constructed for every simulated shape. Therefore the Sinkhorn-Knopp algorithm [31] has been used in order to normalize them and obtain a doubly stochastic matrix that simulates the Hi-C data.
The algorithms used for the following calculus are listed in Appendix B.

3.2 Test of the Multidimensional scaling

In order to verify the validity of the 3D reconstruction in the case of an experimental measured or simulated distance matrix, the MDS has been applied for any polymer to the distance matrix obtained from one random single configuration over all the simulations. Then it has been applied to the distance matrix obtained by taking the average of the 15000 distance matrices of the simulated conformations for each polymer.

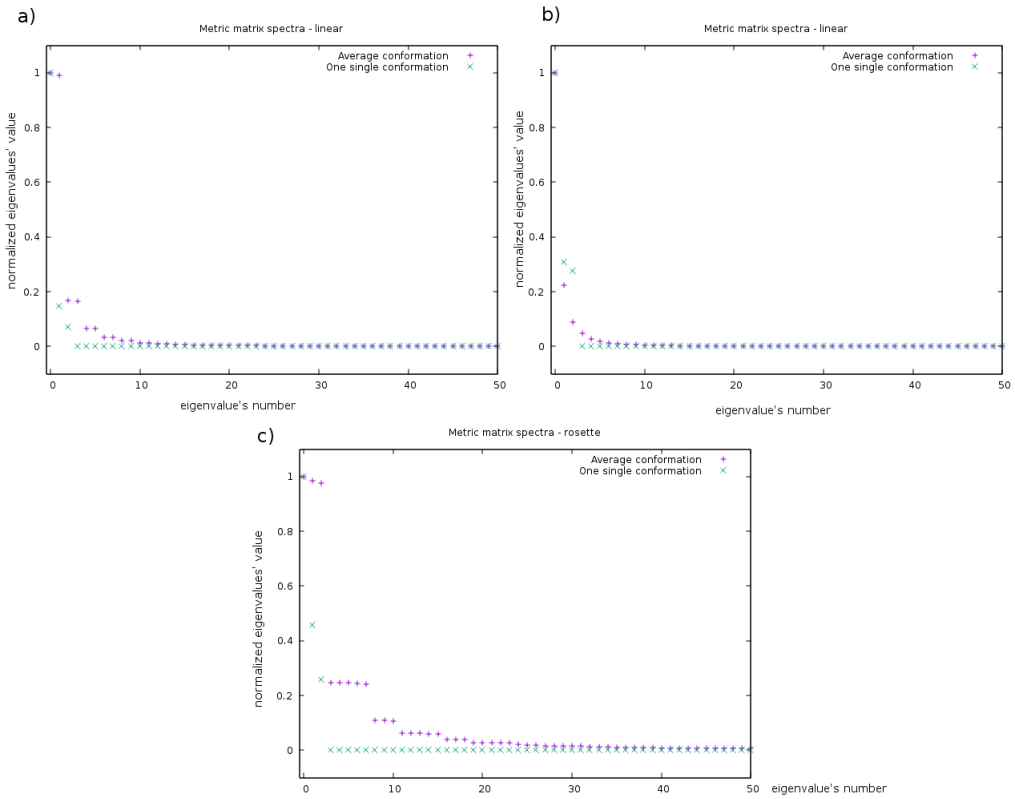

Figure 6: Spectra of the metric matrices. The metric matrices are obtained from the distances matrix computed with the simulation of a linear (a), circular (b) and rosette (c) polymers. Considering one single conformation among the 15000 simulations, the metric matrix has exactly rank 3 (green). Taking the average distances over all the conformations, the spectra is smoother and the rank is not exactly 3 (purple).

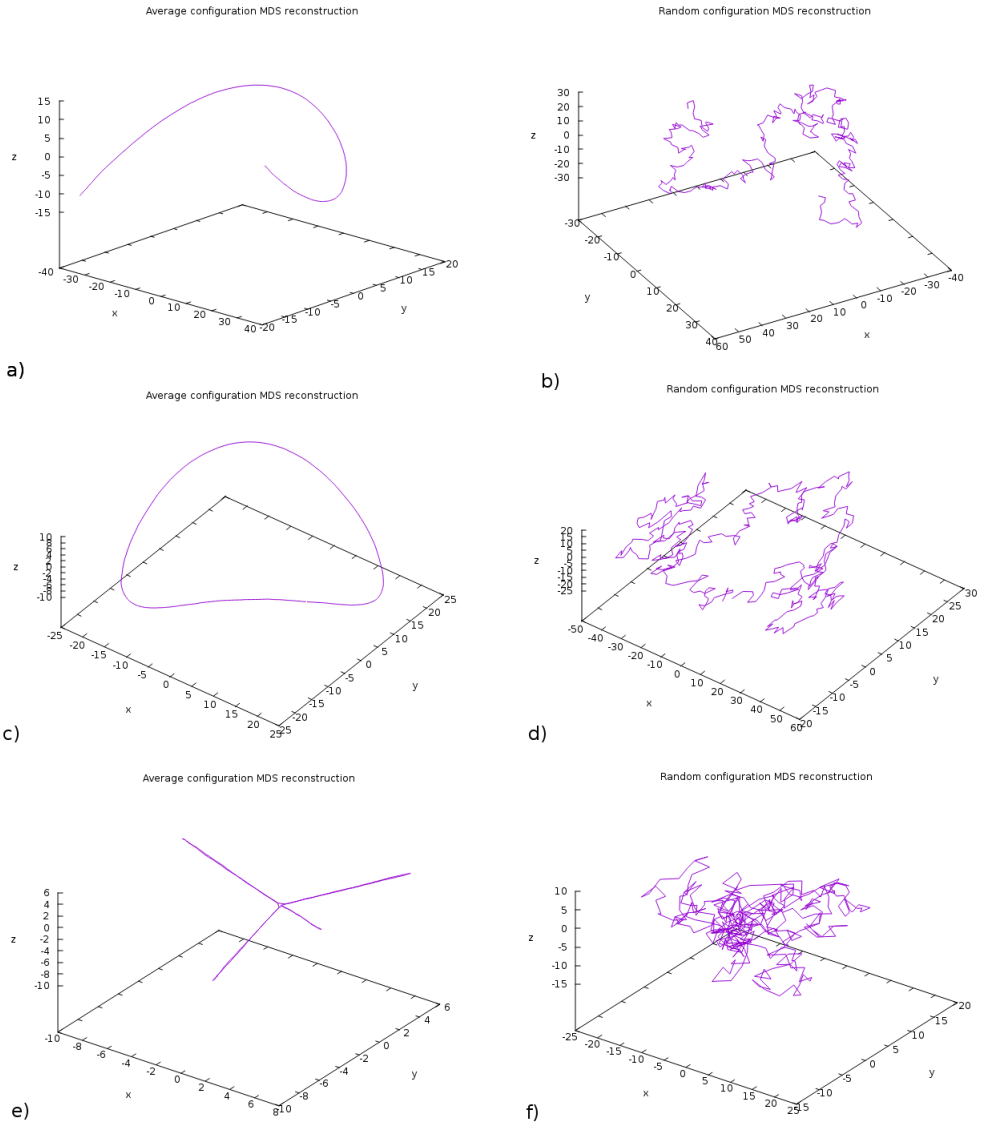

Figure 7: 3D reconstructions of the simulated polymers. The three dimensional reconstructions are obtained by using the MDS for the averaged distance matrices (a-c-e) and for one single conformation over all the simulations (b-d-f). The eigenvectors corresponding to the three highest eigenvalues have been used to obtain the coordinates.

Fig. 6 shows the spectrum of the metric matrix M obtained from each distance matrix. For one fixed single conformation the rank of the metric matrix is 3, hence the reconstruction in a three dimensional space is consistent. In order to verify this statement, the compatibility of each of the $N-3$ eigenvalues with the 0 value has been established to be sufficiently small. In the average case the spectrum of the metric matrices is not exactly 3 anymore. The eigenvalues tend to 0 slower. This is probably due to the fact that the distances are not measured directly. The average over a set of distances has introduced small errors on their definition. The three dimensional structures obtained from both the distances are shown in Fig. 7. The average structure respects the shape of the polymers and reasonably reflects their symmetry. Obviously, this symmetry is not kept taking only one conformation.

3.3 Hi-C simulations data

The Hi-C simulations data obtained as explained above have been interpreted as adjacency matrices of a weighted undirected graph. Their Laplacian spectra are shown in Fig. 8 .

Consequently, for each polymer the shortest path, resistance, natural proximity, entropy, potential and Katz distances have been calculated. The connectivity distances have not been calculated due to the long computation time of their algorithms. The eigenvalues distribution of the metric matrices are shown in Fig. 9.

These spectra show that some negative eigenvalues can arise from some distances, even if the distance matrices are well defined. This means that the set of distances is not embeddable in an Euclidean space. Nevertheless, we calculated the reconstruction by means of the three eigenvectors correspondent to only the first three positive eigenvalues.
The reconstructions obtained by these definitions of distances are shown in Fig. 10 (linear), 11 (circular), 12 (rosette).

Let's first note some qualitative properties of the reconstruction obtained:

- even if the spectra of these matrices are different pairwise, in some cases their reconstruction is the same. Therefore, even if in some cases

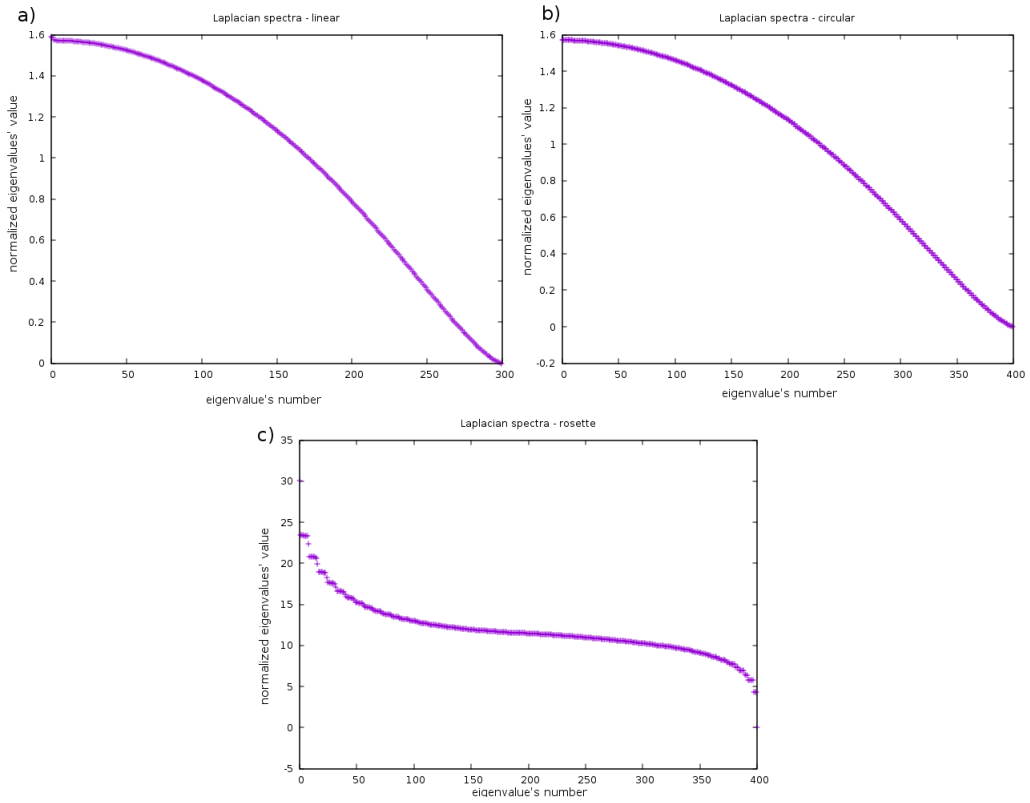

Figure 8: Laplacian spectra of the simulations. The Laplacian is obtained by considering the matrix of the relative frequencies of contacts as an adjacency matrix. Since all of the graphs are complete and connected, the Laplacian spectra have only one 0 eigenvalue in all the linear (a), circular (b) or rosette (c) cases.

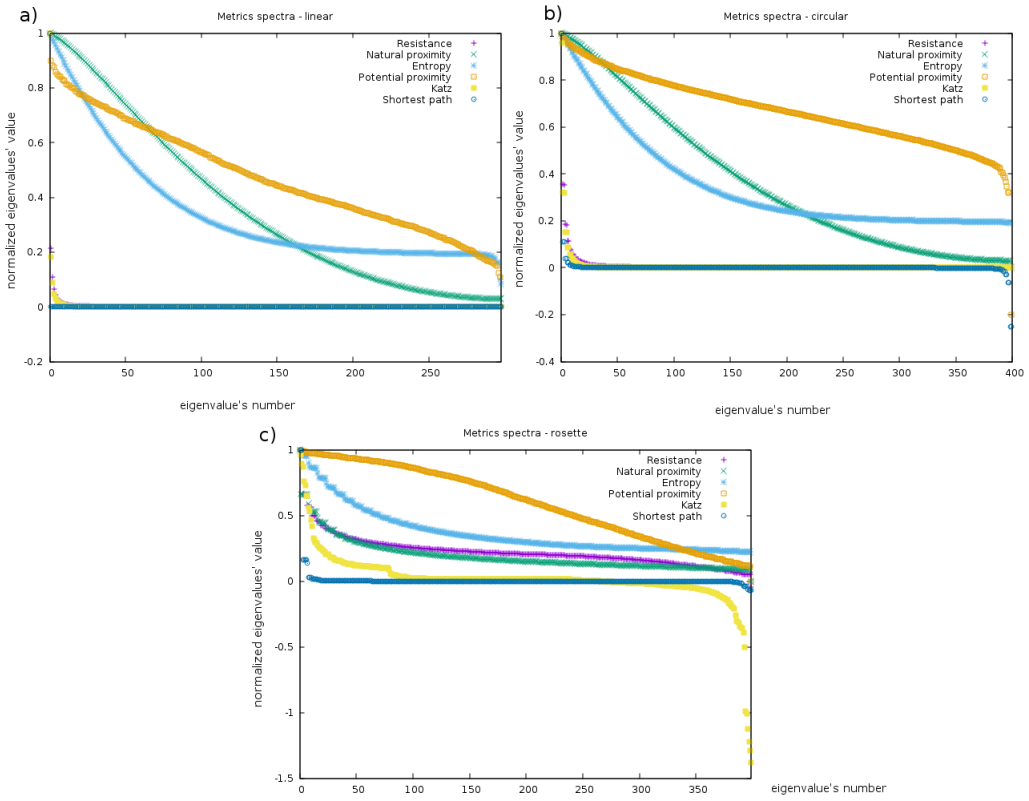

Figure 9: Spectra of the metric matrices. For the linear (a), circular (b) and rosette (c) cases, metric matrices have been computed using 6 different definitions of distances. Some of them (shortest path, potential, katz) show negative eigenvalues. The shortest path and resistance spectra tend always faster to 0.

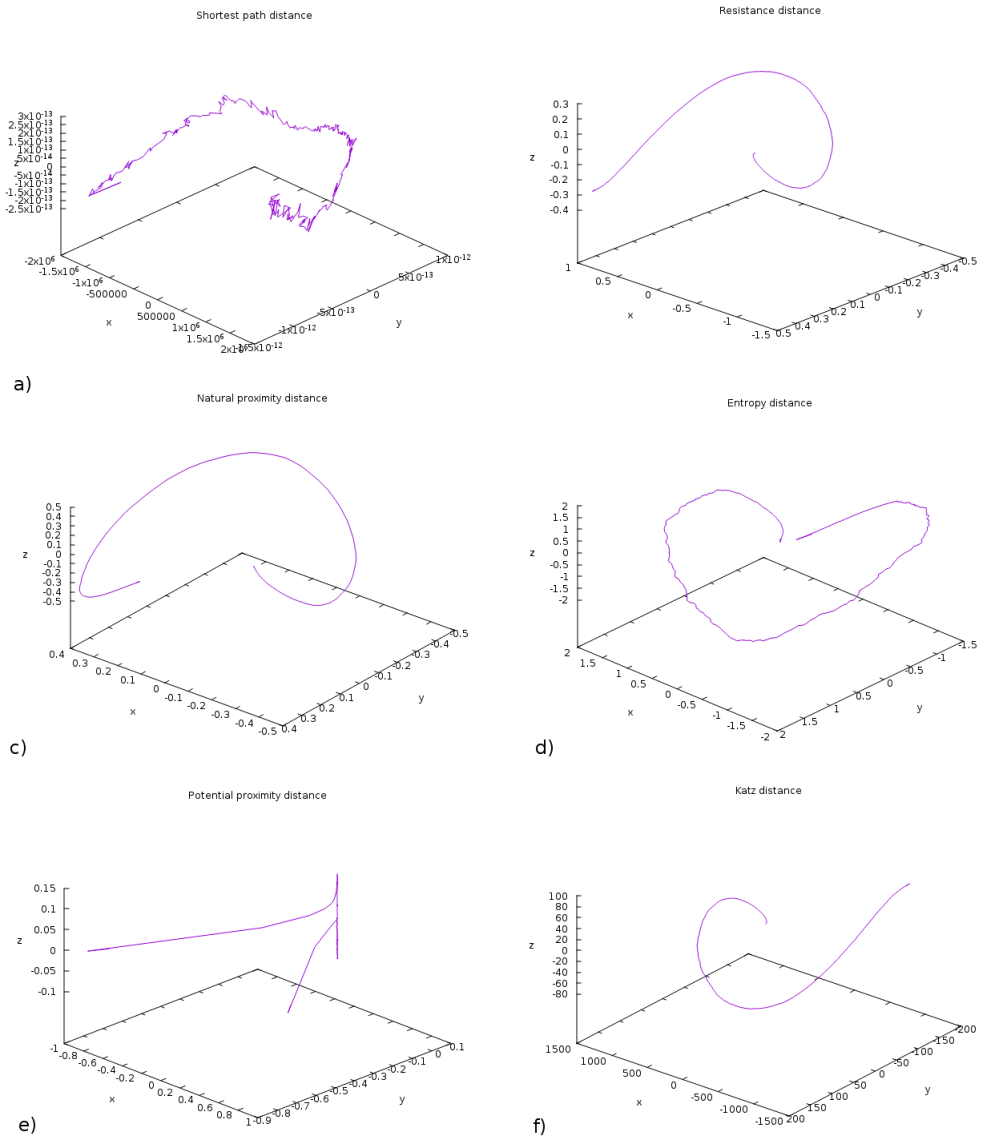

Figure 10: 3D Reconstruction of the simulated linear polymer. The graphs show the reconstruction obtained by using the MDS with the (a) shortest path, (b) resistance, (c) proximity, (d) entropy, (e) potential and (f) katz distances.

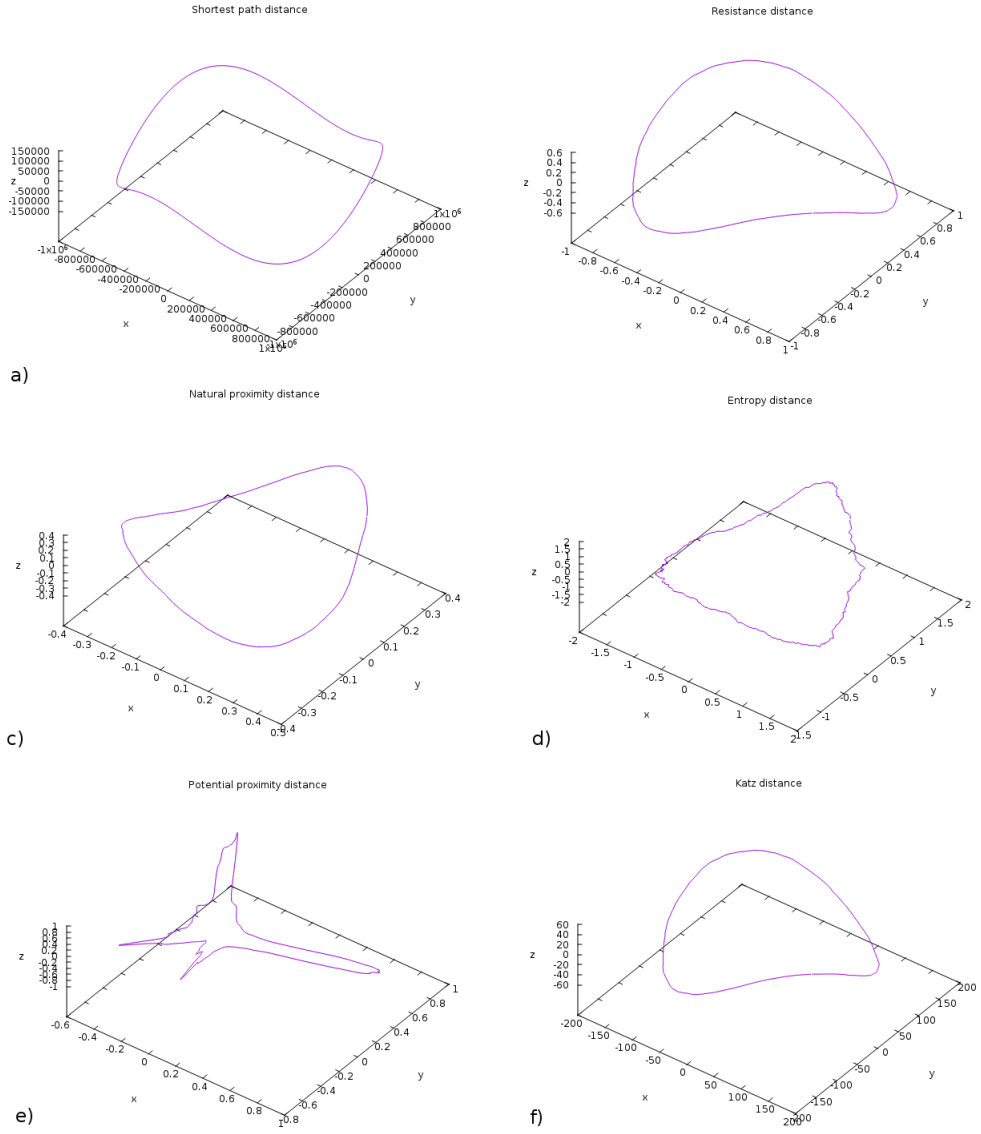

Figure 11: 3D Reconstruction of the simulated circular polymer. The graphs show the reconstruction obtained by using the MDS with the (a) shortest path, (b) resistance, (c) proximity, (d) entropy, (e) potential and (f) katz distances.

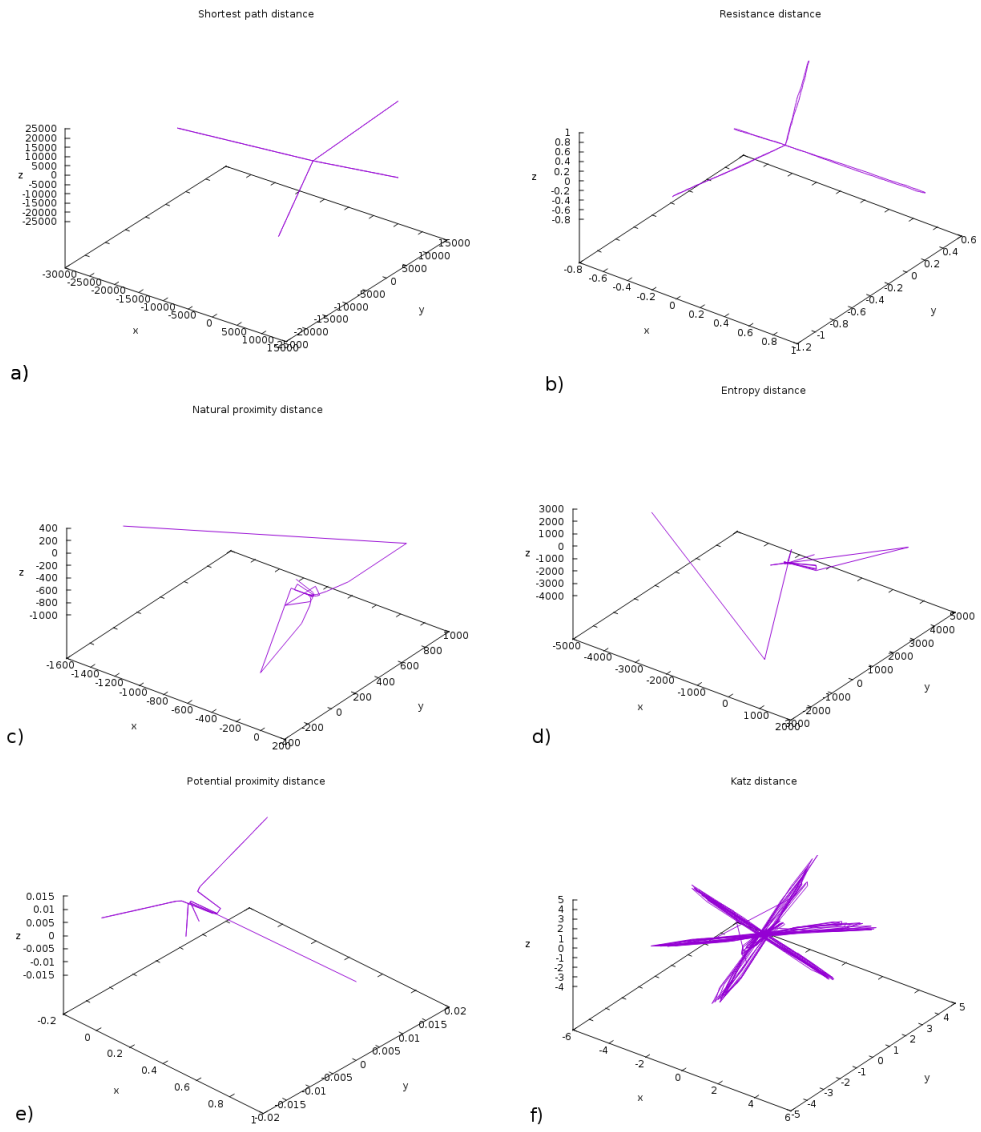

Figure 12: 3D Reconstruction of the simulated rosette polymer. The graphs show the reconstruction obtained by using the MDS with the (a) shortest path, (b) resistance, (c) proximity, (d) entropy, (e) potential and (f) katz distances.

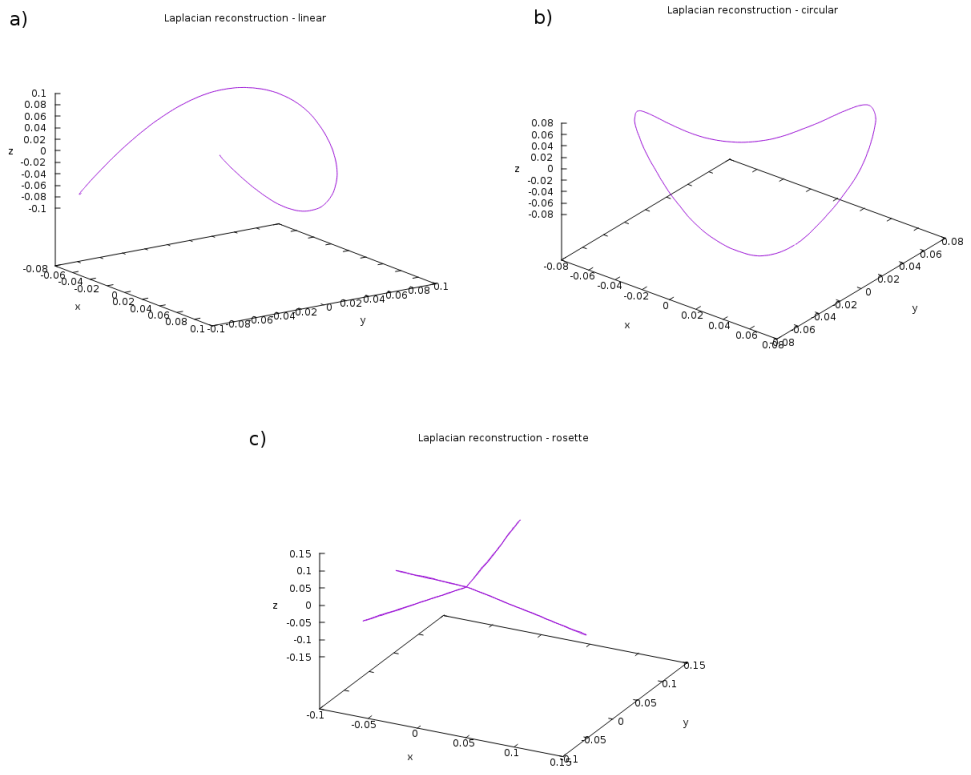

Figure 13: Laplacian reconstruction. Plots of the three eigenvectors of the Laplacian of the Hi-C matrices correspondent to its 3 smallest non-zero eigenvalues.

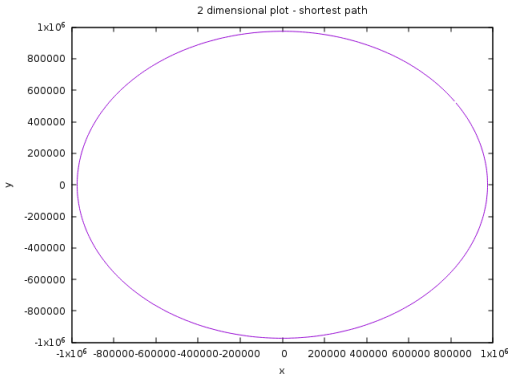

Figure 14: 2 dimensional plot of the circular simulation. The x and y axes of the shortest path reconstruction are plotted for the circular case.

the rank of the matrix is higher than in others, the projections of the conformation in a three dimensional space are still consistent with each other;

- the potential proximity distance reconstruction doesn't respect any property of the polymer and of the reconstruction obtained by using directly the simulated distance matrices;

- the entropy distance respects the symmetries but comparing it with the natural proximity distance small perturbations arise. This is reasonable, thinking that the entropy proximity measure is computed as a perturbation of the natural proximity one due to the multiplication of each entry of the matrix with its logarithm;

- in the rosette case, the Katz distance does not reproduce the same structure of the others. This doesn't happen in the linear and circular case where its spectrum was non-negative;

- equally, in the circular case, even if the two dimensional plot is consistent with the other reconstructions (Fig. 14), the third eigenvector, correspondent to the third axes, is different.

In order to quantify these statements, we considered the reconstructed coordinates vectors and we calculated the dot product with the corresponding coordinates previously obtained by using directly the simulated averaged distance matrix. The closer it is to 1 the more parallel the two vectors are,

hence the reconstruction is consistent with the expected structure.
The resistance reconstruction is always consistent with a precision higher than 93%. In the other cases the qualitative predictions of consistence are confirmed. Some examples are shown in Fig. 15.

Hence, not all the reconstructions are consistent among themselves and not all the reconstructions are consistent among the linear, circular and rosette simulations.
In order to explain this behavior, we show a reconstruction obtained by using the first three eigenvectors of the Laplacian of the graph instead (Fig. 13). Surprisingly, it is always consistent with the expected structure with a high precision. The average structure over a high number of simulations is exactly reproduced by the eigenmodes of the Laplacian of the contact-frequencies weighted graph relative to the three smallest nonzero energy states.
In the last years, the eigenvectors of the Laplacian have already been used as a set of coordinates for the so-called *synchronization dynamic* of networks [39].
From a theoretical point of view, it is possible to calculate the coordinates with the MDS only using the resistance distance among the distances proposed in this thesis. In fact, as explained above, the first 3 eigenvectors of its metric matrix are always the eigenvectors corresponding to the 3 smallest nonzero eigenvalues of the Laplacian. The resistance metric is the only one that assures that the eigenvectors are the same and the order is reversed. In the other cases, even if we demonstrated that the Laplacian and the metric matrix share the same set of eigenvectors, their order is not said to be respected. In fact, in these cases it is not possible to predict how the eigenvalues of the metric matrix depend from the Laplacian ones.

As an example, in the circular case the shortest path reconstruction didn't respect the order of the eigenvectors of the Laplacian. Nevertheless, substituting its third eigenvectors with its last one, the reconstruction would be fully consistent with the expected one (Fig. 16).

Finally let's note that according to the analogy of a system of springs connecting the vertexes of the graph, the constant eigenmode of the Laplacian represents the case in which all the elongations are null and all the springs assume their equilibrium length, that is 0.

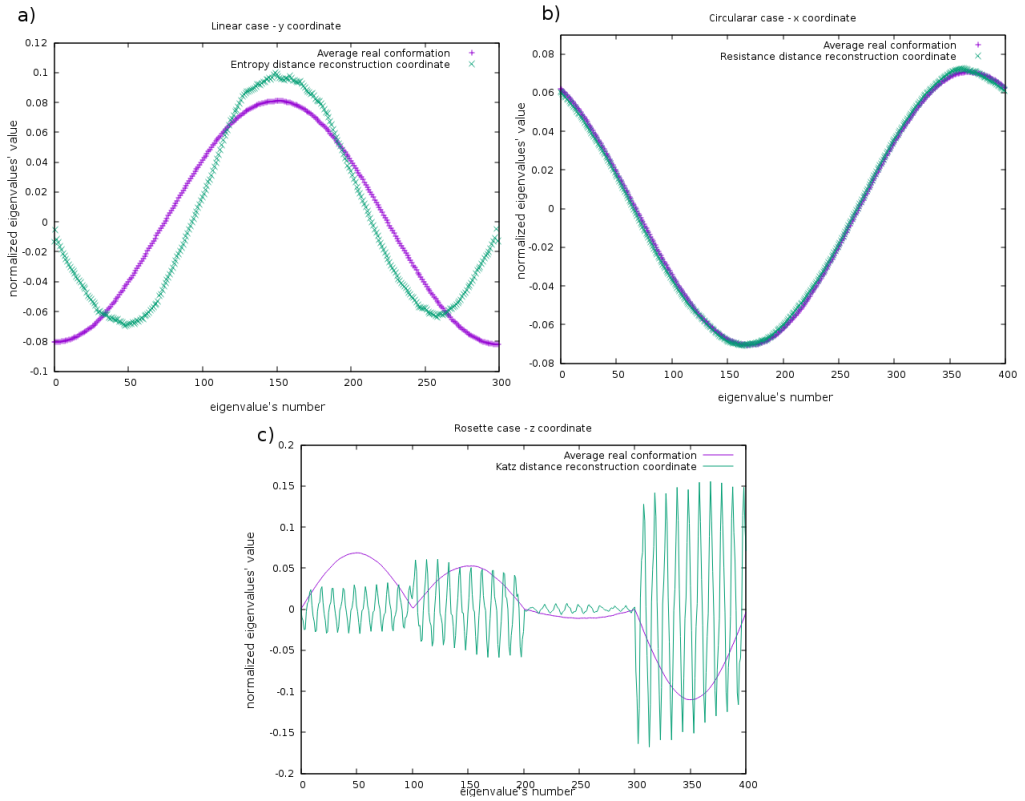

Figure 15: Examples of 1 dimensional comparison. (a) Linear case.The second eigenvector obtained by using the entropy distance is compared with the expected y coordinate. The resulting dot product is 0.878. (b) Circular case. The first eigenvector obtained by using the resistance distance is compared with the expected x coordinate. The resulting dot product is 0.999. (c) Rosette case. The third eigenvector obtained by using the Katz distance is compared with the expected z coordinate. Even if some macro-features are respected, the resulting dot product is 0.0001.

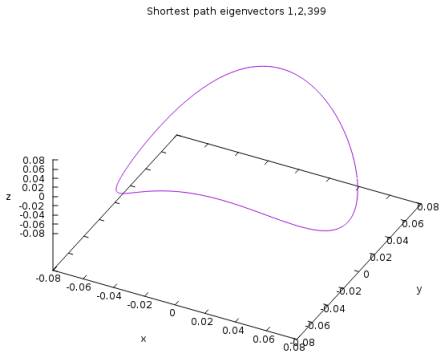

Figure 16: Example of different eigenvectors order. The plot shows the first, second and last eigenvector of the shortest path metric matrix in the circular case. The dot product between the last eigenvector and the expected z axes is 0.998, but this doesn't appear in the 3d reconstruction, since its correspondent eigenvalue is 0.

3.4 Real data

In this section we will apply the previously developed algorithms to a real set of Hi-C data. In fact, we verified that, simulating many times a self avoiding walk for different shapes of a polymer, the average structure is reproduced by the first eigenvectors of the Laplacian of the related contact graph. This reconstruction is obtained by using the resistance distance in the multidimensional scaling technique.

In the experimental Hi-C data, usually a certain amount of different cells is analyzed. They grow in common conditions and after a specific treatment they are subjected to formaldehyde fixation in order to create cross-links. The cross-links are then counted and finally the Hi-C matrix are created.
Here we will apply the MDS to Hi-C data of the *Bacillus subtilis* [38] (Fig. 17). Each analyzed cell contained one single and unreplicated chromosome. The genome has been divided into 1054 bins. The Hi-C data have been demonstrated to be reproducible and consistent with the data obtained previously.

Even in this case the resistance distance reconstruction has been demonstrated to be consistent with the first three eigenvectors of the Laplacian.
In the past, some applications which use the shortest path distance have already been proposed. MDS has been used in order to demonstrate the

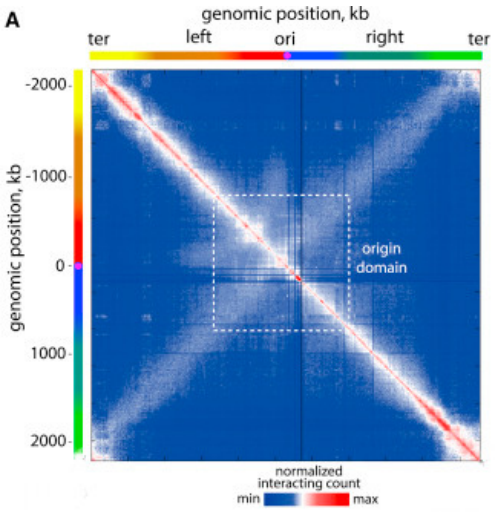

Figure 17: Hi-C matrix. [Adapted from [38]] Hi-C contact matrix of the *Bacillus subtilis*. The white dashed box represents the situ in which the replication begins.

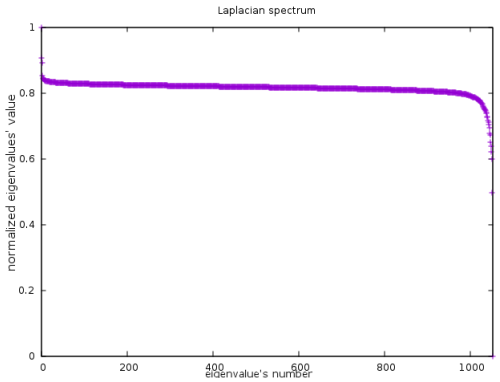

Figure 18: Laplacian spectrum. Spectrum of the Laplacian of the graph of the *Caulobacter crescentus*. Since there is only one 0 eigenvalue, the graph is connected.

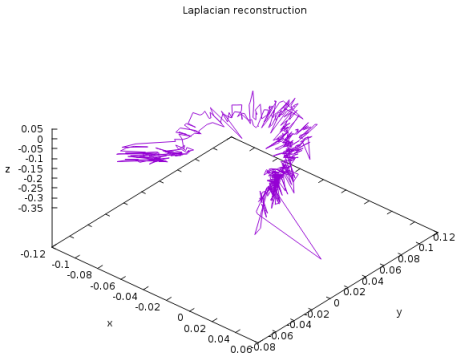

Figure 19: Laplacian reconstruction. Plots of the three eigenvectors of the Laplacian of the Hi-C matrices correspondent to its 3 smallest non-zero eigenvalues.

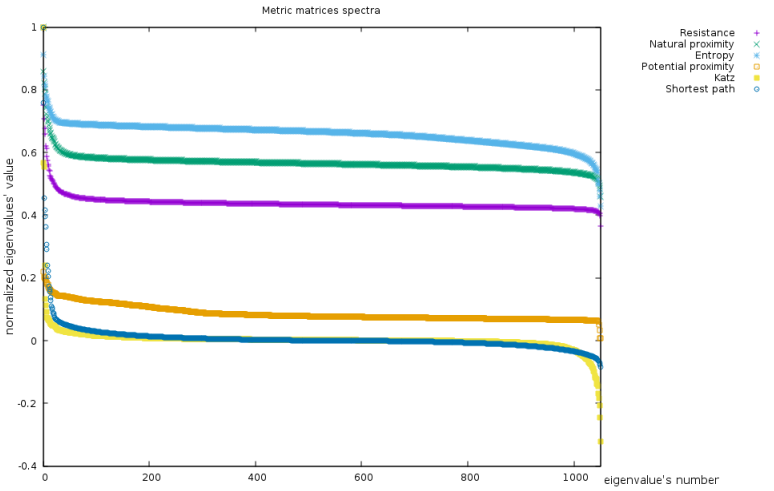

Figure 20: Spectra of the metric matrices. Metric matrices have been computed using 6 different definitions of distances. Some of them (shortest path, potential, katz) show negative eigenvalues.

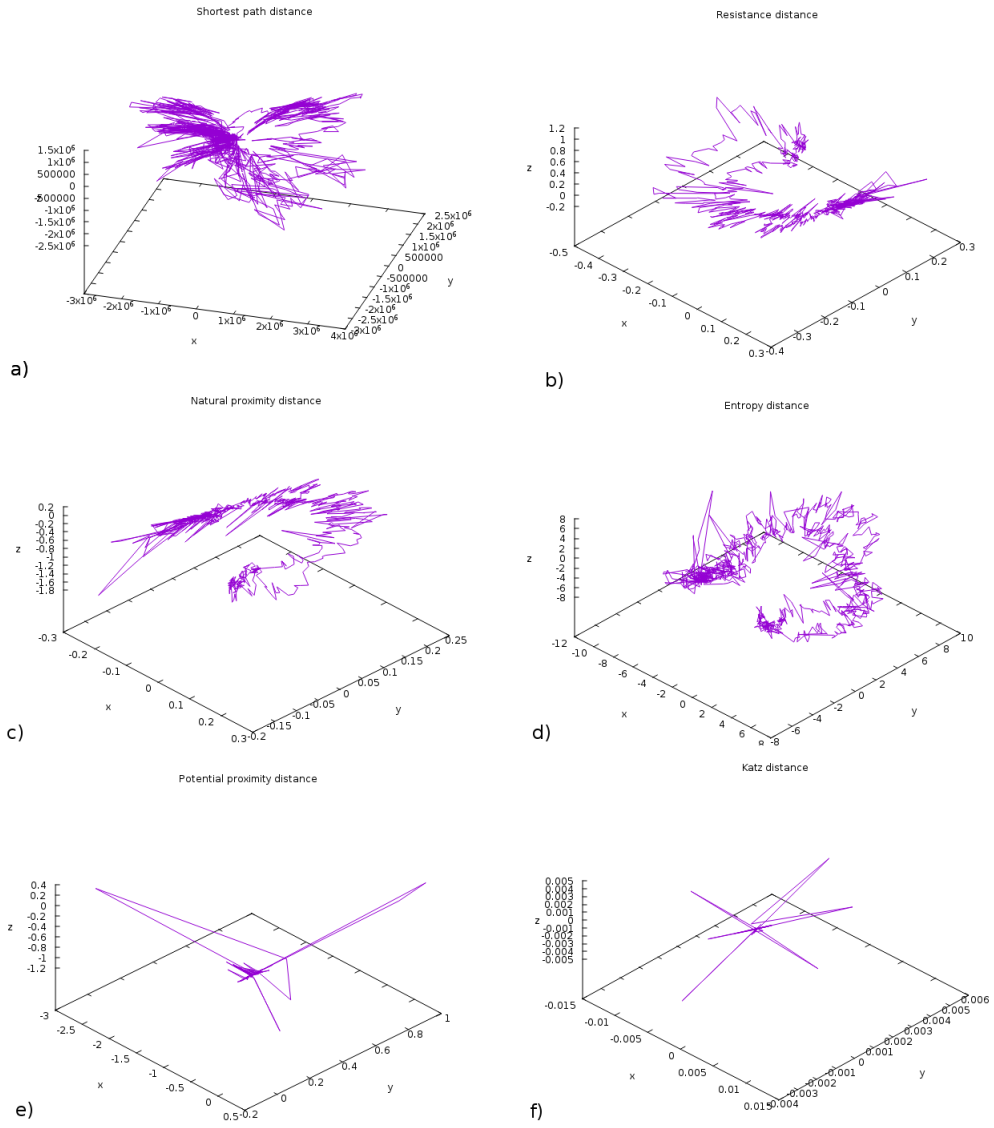

Figure 21: 3D Reconstruction of the *Caulobacter crescentus* chromosome. The graphs show the reconstruction obtained by using the MDS with the (a) shortest path, (b) resistance, (c) proximity, (d) entropy, (e) potential and (f) katz distances.

existence of particular domains and, in combination with super-resolution microscopy, to verify that the structure of the chromosome changes during the life cicle of the cell [38]. In particular, it has been used to unveil the factors responsible for its regulating folding. For further applications, the different reconstruction obtained by using the resistance distance may reveal other important features.

4 Other applications

4.1 Introduction

We have shown how it is possible to interpret the Hi-C data as an adjacency matrix of a complete weighted graph and consequently reconstructing a set of three dimensional Euclidean coordinates. We showed in which cases these coordinates correspond to the first three eigenvectors of the Laplacian of the graph. Nevertheless, this reconstruction unveils some overall properties of the chromosomes, carrying a loss of information about the local details of the structures, e.g. special patterns, loops or clusters.
The following chapter will investigate an application of the method shown above in the case in which distances between just a few pairs of points are known with a certain precision. The reason of this request arises from some fluorescence measures that permit to visualize multiple loci at the same time and hence to deduce their distances. Nevertheless, at the state of art it is not possible to measure directly a complete set of distances.
This thesis proposes two methods: the first one is a Bayesian method whose goal is to use the Hi-C data in order to complete the fluorescence distance matrix or, in other words, it uses these distances as a constraint in the reconstruction process; the second method asserts the uncertainty of the FISH distances and uses them to reconstruct a new adjacency matrix of contact probabilities that combine the two kinds of data.
In our simulations we will consider as sets of measured distances a collection of few distances chosen randomly from the previously used distance matrices, i.e. the one single conformation and the averaged distance matrix.

In both the methods we will stress the use of a probability density defined over some possible energy states that the graph can assume. The Boltzmann probability density of the states of the generalized energy E of a graph G is:

$$\rho(E) = e^{-E} = e^{\frac{1}{2}\sum_{ij} d_{ij}^2 w_{ij}} = \prod_{ij} e^{-\frac{1}{2} d_{ij}^2 w_{ij}}$$

Clearly, this probability density can change both with the frequencies w_{ij} and with the way the distances are defined. Some authors [2] have already used this definition of probability density for graphs focusing on the way the probability changes when the graph adjacency matrix is modified. In our cases instead we will be interested on considering the energy value over all

the possible length that the distances can assume. For this purpose, the following theorem on the number of independent distances will be useful.

Theorem. Given a set of N points in a d-dimensional space, the minimal number of distances between the points that can be defined freely is $m = \frac{d(d+1)}{2} + (N - d - 1)(d + 1)$ where $d \geq 2$ and $N \geq 2$.

Proof. The proof arises from symmetry considerations. Embedding $N = d + 1$ points needs necessarily $\sum_{i=0}^{d} n_i = \frac{d(d+1)}{2}$ lines. For any other point x added are then necessary $d + 1$ lines in order to not have any other degree of freedom. In fact, adding only d lines from x to other d points, then these points form a d-dimensional surface such that x can assume two different symmetric positions respect of the surface itself. Hence all the distances of x with the other points would change. Thus, it is necessary to add $(N-d-1)(d+1)$ distances.

By this demonstration it is clear that the choice of the m distances is not arbitrary. For instance, in the three dimensional case we must define for each point at least 4 distances with other points.
For a three dimensional space at least only $m = 4N - 10$ distances can be freely chosen. Fixing them the other ones will be certain. Let $\widetilde{d}$ denote the set of some m' independent distances, where $m' \leq m$. Then we define a partition function Z' of the graph as

$$\begin{aligned} Z' &= \int_0^\infty e^{-E(\widetilde{d})} d^{m'} \widetilde{d} = \\ &= \prod_{(ij)=1}^{m'} \int_0^\infty e^{-\frac{1}{2}\widetilde{d}^2_{(ij)} w_{(ij)}} d\widetilde{d}_{(ij)} = \\ &= \left(\sqrt{\frac{\pi}{2}}\right)^{m'} \prod_{(ij)=1}^{(m')} \frac{1}{\sqrt{w_{(ij)}}} \end{aligned}$$

where in the last step we used the Gaussian integral.

4.2 Bayesian method

The following method's goal is to modify the adjacency matrix W of the Hi-C data inserting an additional information. This means assuming that a set of M distances d^* is known from a statistical point of view. In fact, we can interpret the entries w_{ij} of W as the probability $p(ij)$ that there is a contact between the points i and j, i.e. their distance $d_{ij} = 0$.
Considering the additional information provided by d^* means calculating $p(ij|d^*)$. Using the Bayes theorem:

$$p(ij|d^*) = \frac{p(ij)p(d^*|ij)}{p(d^*)} = p(d^*|ij)w_{ij}$$

provided that $p(d^*) = 1$ and $d^* > 0$.
By definition, $p(d^*|ij) = 0$ whenever one of the distances of d^* is the distance between the points i and j. Let's assume the case in which $d_(ij) = 0 \notin d^*$ and let $\widetilde{d}$ denote all the independent distances not belonging to d^*. They will consequently be $m - M$. Then:

$$p(d^*|ij) = \frac{e^{-E(\widetilde{d})}e^{-E(d^*)}}{Z^*}$$

In this case, the partition function Z^* is calculated integrating the probability density not over all the independent distances, but only between the distances d^*. In fact, in this case we are interested on calculating it only over the values that the experimental measure could assume. Hence:

$$p(d^*|ij) = \frac{e^{-E(\widetilde{d})}e^{-E(d^*)}}{e^{-E(\widetilde{d})}\int_0^\infty e^{-E(d^*)}dd^*} = \sqrt{\frac{2}{\pi}}^M e^{-E(d^*)}\prod_M \sqrt{w_M}$$

Consequently:

$$w'_{ij} = p(ij|d^*) = w_{ij}\sqrt{\frac{2}{\pi}}^M e^{-E(d^*)}\prod_M \sqrt{w_M}$$

Hence, these assumptions yield to the multiplication of the adjacency matrix with a constant factor, always smaller than 1. The entries corresponding to the given set of distances are set to 0. Considering the set of N points as connected by springs with elastic constant equal to the contact frequency,

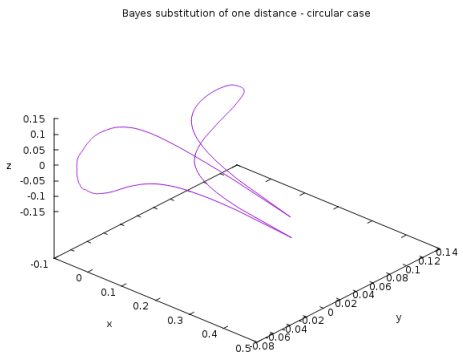

Figure 22: Substitution of one single distance. One defined and consistent, but rare distance has been imposed to the circular structure. The number of consistent distances is low and it doesn't show any interesting property.

then setting one of them to 0 means substituting the spring with a not deformable bar of fixed length.
The new adjacency matrix has to be newly normalized in order to let the rows and the columns sum again to 1.
Even if the method is theoretically interesting, it can not find immediate applications. Its principal defect is that the measured distance is applied during a second step of the algorithm and this may not be consistent. In fact, applying the reconstruction method to the new adjacency matrix does not assure that the distances between two points - whose distance is supposed to be d^* - will be exactly d^*. This condition should be imposed *a posteriori*, provided that d^* is consistent with the distances calculated, i.e. it satisfies the triangular inequality with any other third point. This case is really rare and, even if applied, it doesn't yield to any interesting result (see Fig. 22).

4.3 New frequency method

We showed some limitations of the previous problem. In particular, it is not possible to be sure *a priori* that the measured distance is consistent with the reconstruction. Furthermore, we didn't take into account the uncertainty of the measured distance. In fact, one may assert that the measured distance is not certain, but it is the most probable among all the possible values it can take. The following method arises from this assumption.

We will find a way to redefine the entries of the adjacency matrix. For this purpose, let's assume that each measured distance d^* is equal to the average distance that we would obtain by using the probability density $\rho(E)$. That is, consider

$$d^* = \sqrt{\langle d_1 \rangle^2 + \langle d_2 \rangle^2 + \cdots + \langle d_N \rangle^2}$$

where the average of each component of the distance is

$$\langle d_i \rangle = \frac{\int_0^\infty \lambda_i d_i e^{-E(d)} dd_1 \dots dd_{4N-10}}{\int_0^\infty e^{-E(d)} dd_1 \dots dd_{4N-10}} = \frac{\int_0^\infty d_i e^{-\frac{1}{2}\lambda_i d_i^2 w} dd_i}{\int_0^\infty e^{-\frac{1}{2}\lambda_i d_i^2 w} dd_i} = \sqrt{\frac{2}{\pi}} \frac{\sqrt{\lambda_i}}{\sqrt{w}}$$

Hence the distance is:

$$d^* = \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{w}} \sqrt{\sum_i \lambda_i}$$

Consequently, we can compute the related frequency of contact between two points i and j as:

$$w_{ij} = \frac{2}{\pi} \frac{1}{d_{ij}^{*2}} TrM$$

This definition can be interpreted as a constraint for the energy of the graph, since $w_{ij} d_{ij}^{*2}$ is dimensionally equivalent to the energy itself.
Let's note that, since the relative frequency can't be bigger than 1, then the measured distance should be rescaled in such a way that $d^* \geq \sqrt{\frac{2}{\pi TrM}}$. This means considering a rescaled distance: $d^*_{resc} = \frac{\sqrt{2/(\pi TrM)}}{d^*_{min}} d^*$ where d^*_{min} is the smallest value that experimentally d^* can assume. Alternatively, it is possible to define arbitrarily a threshold distance d^*_{min}, imposing that for any distance $d^*_{ij} \leq d^*_{min}$, then $w_{ij} = 1$. This means that the new probability w_{ij} does not depend on the constant factor $\frac{2TrM}{\pi}$, but only on the choice of the threshold distance.
After the substitution of the new few entries, it will be necessary to re-normalize the matrix with the Sinkhorn-Knopp algorithm [50], in order to let each row and column sum to 1. In fact, by construction the probabilities that we found are the probabilities of uncorrelated events: the sum of different probabilities in the same row may be bigger than 1. The normalization assures that the contacts are considered mutually exclusive events; the rescaling of the distances assures that the new frequency are in any case consistent with the dimensions of the structure of the polymer. The normalization lets

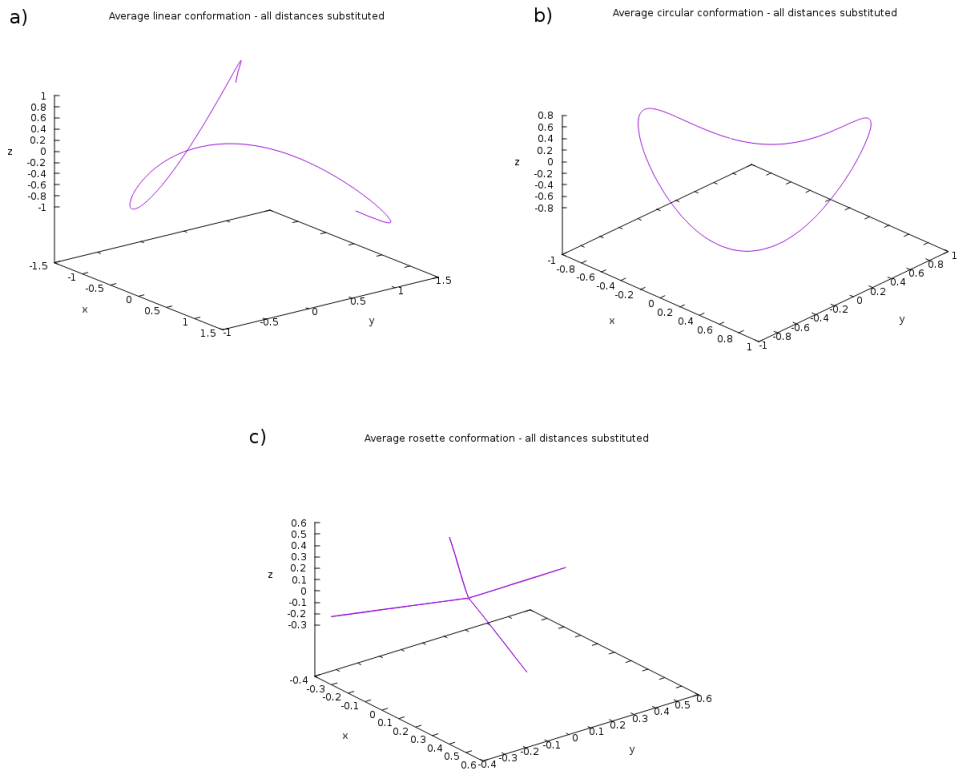

Figure 23: Reconstruction of the average structure. The algorithm of the new frequencies has been applied to the linear (a), circular (b) and rosette (c) structures using all the distances of the matrices of the averaged distances used also previously.

the new and the old data influence each other.
In order to verify the consistence of this construction we first reconstructed the conformation of the polymers, substituting *all* the entries of the adjacency matrices with the new frequencies, using the averaged distance that we have already used before (Fig. 7 a,c,e). Then, we applied the resistance distance MDS algorithm to obtain the three dimensional reconstruction. The threshold distance of the rescaling has been fixed as the maximum distance between two consecutive monomers in the simulated chain.
The results and the spectra of the metric matrices are shown in Fig. 24 and 23.

Computing the dot product of the original coordinates vectors with the

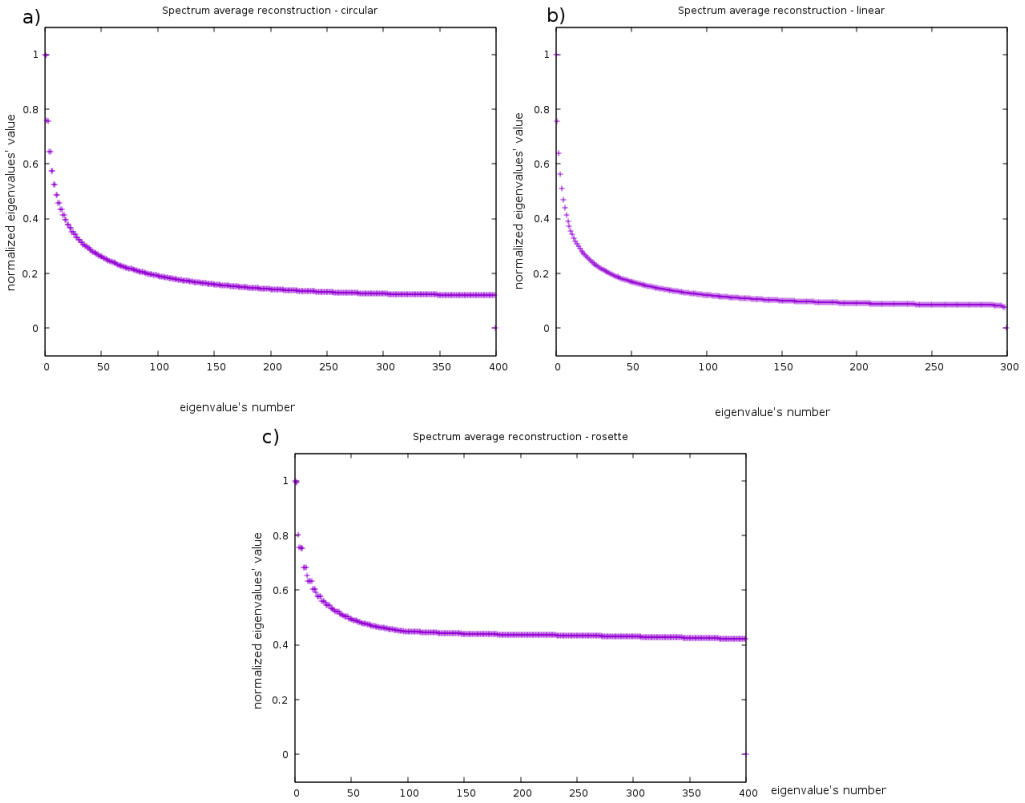

Figure 24: Spectra of the reconstruction of the average structure. The shown spectra are the spectra of the metric matrices of the reconstruction with the new frequencies applied to the linear (a), circular (b) and rosette (c) structures using all the distances of the matrices of the averaged distances used also previously.

correspondent reconstructed ones, they are verified to be exactly parallel. Consequently, they are parallel to the coordinates vector obtained by using the simulated Hi-C matrices. This result is reasonable. In fact the distances are averaged over a high number of simulation. The smaller the average distance, the higher the amount of small distances and the higher the frequency of contact too, according to the definition of frequencies used to calculate the Hi-C matrices.
Consequently, we substituted only 100 frequencies on the adjacency matrix, using only 100 random distances of the previously used set. The shape of the structure results in any case slightly perturbed, probably due to the renormalization process (Fig. 25).

Therefore, the algorithm has been applied not to the average distances, but to the distances among the points of the particular singular conformations shown in Fig. 7 b, d, f. Even in this case the threshold distance used to rescale the distances is the maximum distance between two consecutive monomers of the chain. The spectra and the reconstructions are shown in Fig. 27 and 26.

The reconstruction is now not exactly consistent with the original one. Table 2 shows the values of the dot products between the correspondent axes. The first axes result parallel with high precision. The precision is lower for the second axis, but still high, and finally it is not always acceptable for the third axis. Since many of the eigenvalues are different than 0, the reconstruction is a projection on the first three coordinates of a higher dimensional space. In the original reconstruction the rank of the metric matrix was instead exactly 3 since the set of distances was directly computed in a three dimensional space (Fig. 6). The inconsistency now arises from the errors introduced with the indirect computation of the distance through a contacts' frequency matrix.
Let's note that in the average structure's case the reconstruction is consistent since the spectrum of the metric matrix was already smoothed in the original reconstruction.

Finally, we substituted only 20, 100 and 1000 new frequencies to the adjacency matrix of the simulated Hi-C data corresponding to the average structure. This test was performed in order to show how the structure increasingly changes when new distances are added. The change of the struc-

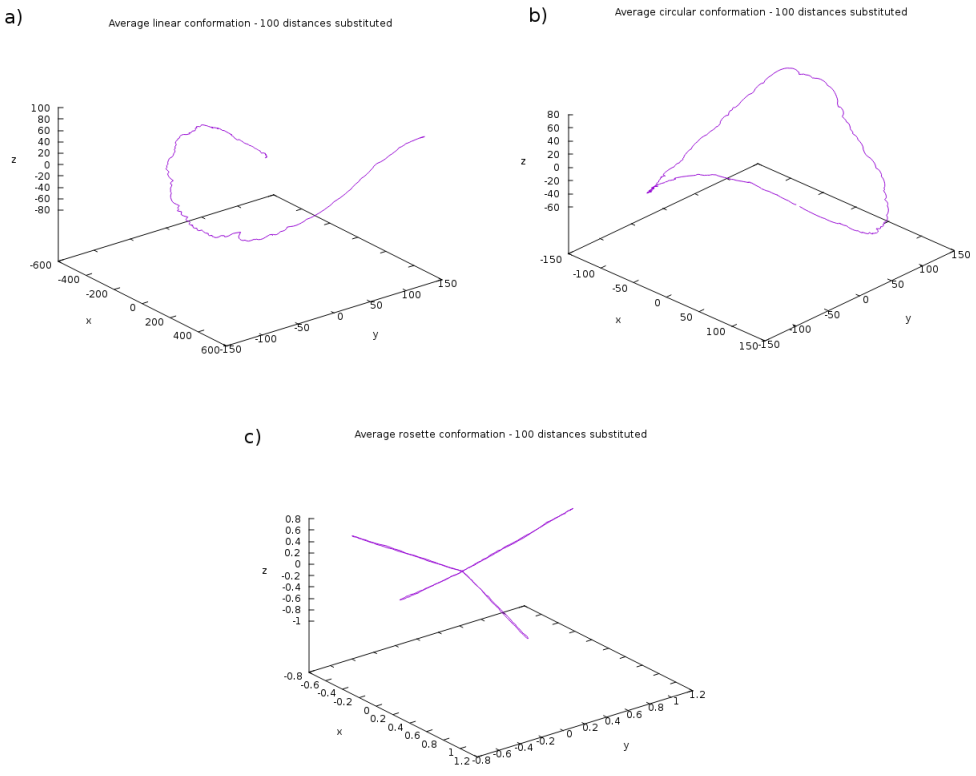

Figure 25: Reconstruction of the average structure with partial substitution. The algorithm of the new frequencies has been applied to the linear (a), circular (b) and rosette (c) structures using in each case only 100 distances of the matrices of the averaged distances used also previously.

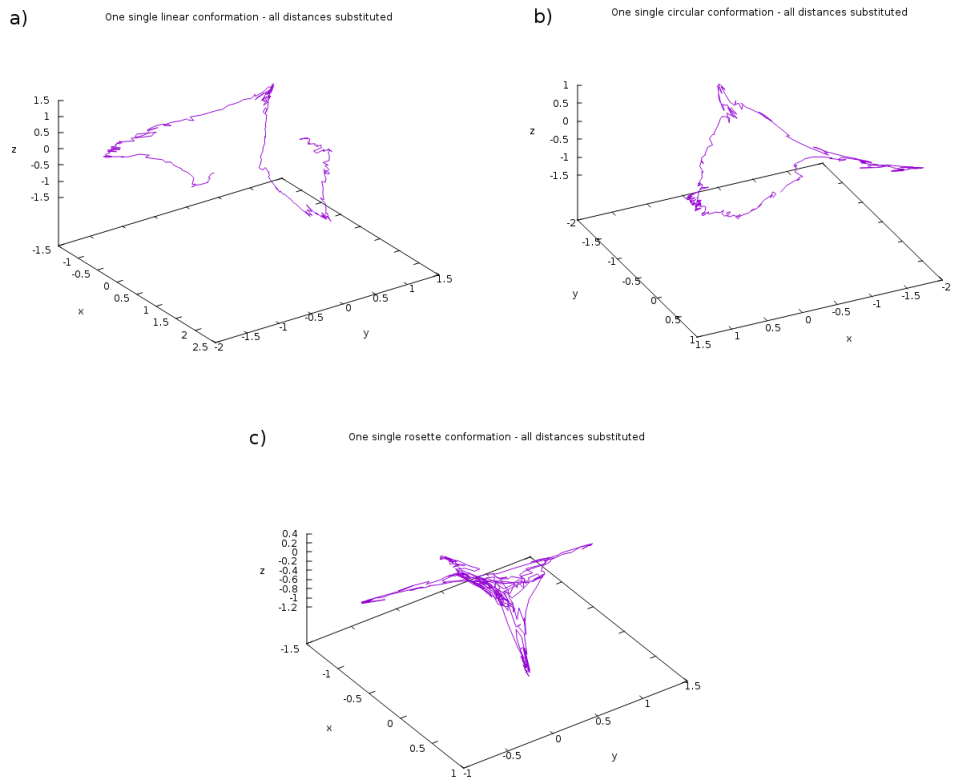

Figure 26: Reconstruction of one single structure. The algorithm of the new frequencies has been applied to the linear (a), circular (b) and rosette (c) structures using in each case all the distances of the matrices of the distances of one single conformation among all the simulated ones. In particular, the conformation used is the same shown in Fig. 7.

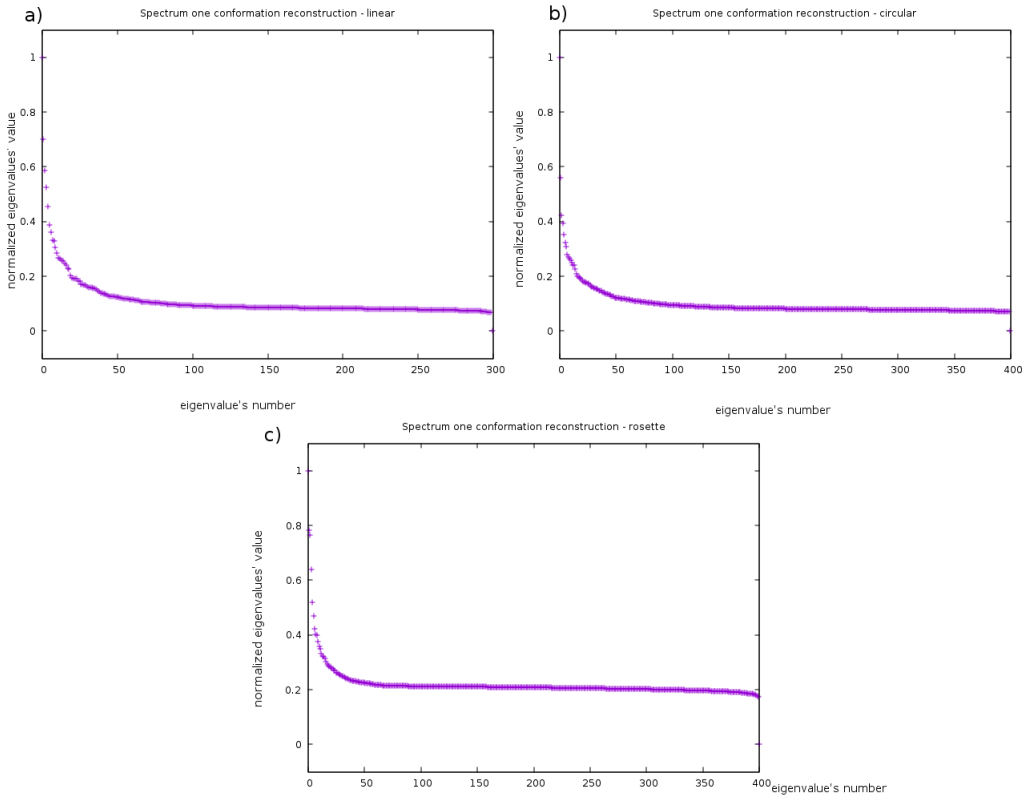

Figure 27: Spectra of the reconstruction of one single structure. The shown spectra are the spectra of the metric matrices of the reconstruction with the new frequencies applied to the linear (a), circular (b) and rosette (c) structures using all the distances of the matrices of the distances used also previously of one single conformation.

	x	y	z
Linear	0.973	0.874	0.812
Circular	0.992	0.705	0.551
Rosette	0.985	0.881	0.457

Table 2: Dot products between reconstructed and real coordinates. The theoretical value of the dot product for parallel vectors is 1. Here are shown the dot products between the real coordinates' vector and the ones reconstructed through the use of the new frequencies. The dot product decreases with the order of the axes.

ture is continuous and in each case deforms the average structure without breaking it, i.e. there are not points in isolated areas, not belonging to the structure (Fig. 28, 29, 30).

The method presented in this last chapter has been proposed in order to combine two different sets of data: one consisting of contact frequencies (Hi-C data) and one consisting of a set of few directly measured distances. We showed that if the set of measured distances is complete it is possible to reconstruct a second probabilities matrix analogous to the Hi-C matrix. The two matrices are consistent only if the set of distances coincides with the average distances over all the conformations analyzed by the Hi-C method. In the other cases, we obtained some reconstructions whose precision decreases with the order of the coordinates. Then we only changed some frequencies on the Hi-C matrix through the new algorithm showing that the structure slightly changes increasing the number of new frequencies, but always in a smooth way.
This technique is obviously not accurate enough, but it presents some interesting properties. Relationships between the spectra of the metric matrices and on the choice of a particular set of distances may be further analyzed in order to understand the behavior of the reconstruction.

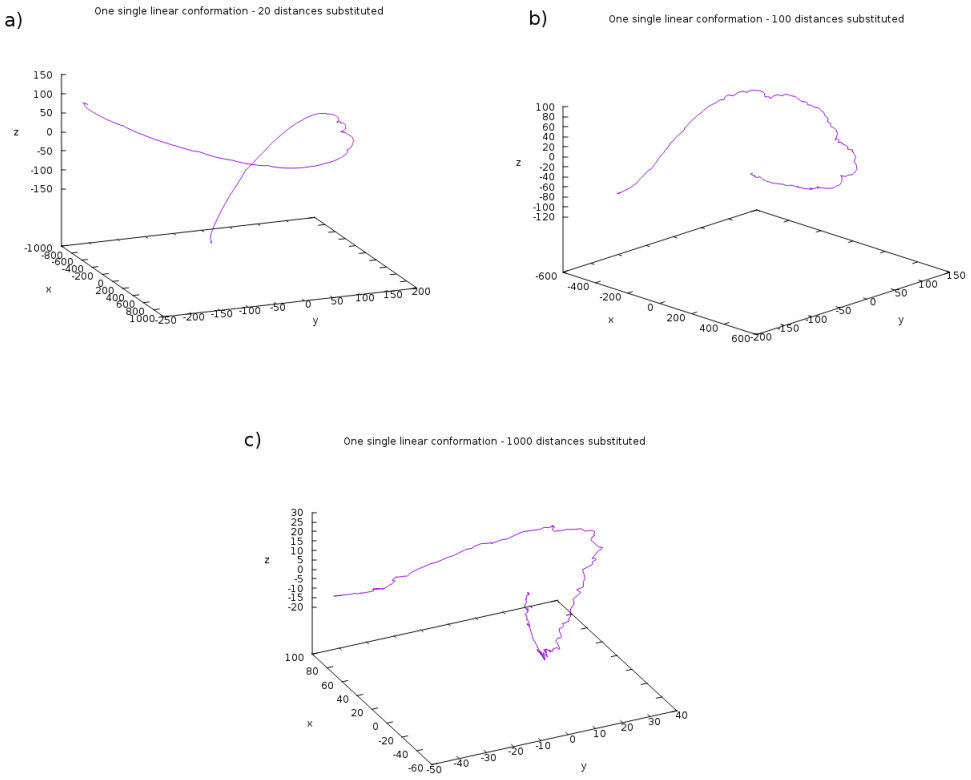

Figure 28: Reconstruction of the linear structure with partial substitution. The algorithm of the new frequencies has been applied to the linear structure using only (a) 20, (b) 100 or (c) 1000 distances of the matrix of the distances of one single conformation among all the simulated ones. In particular, the conformation used is the same shown in Fig. 7.

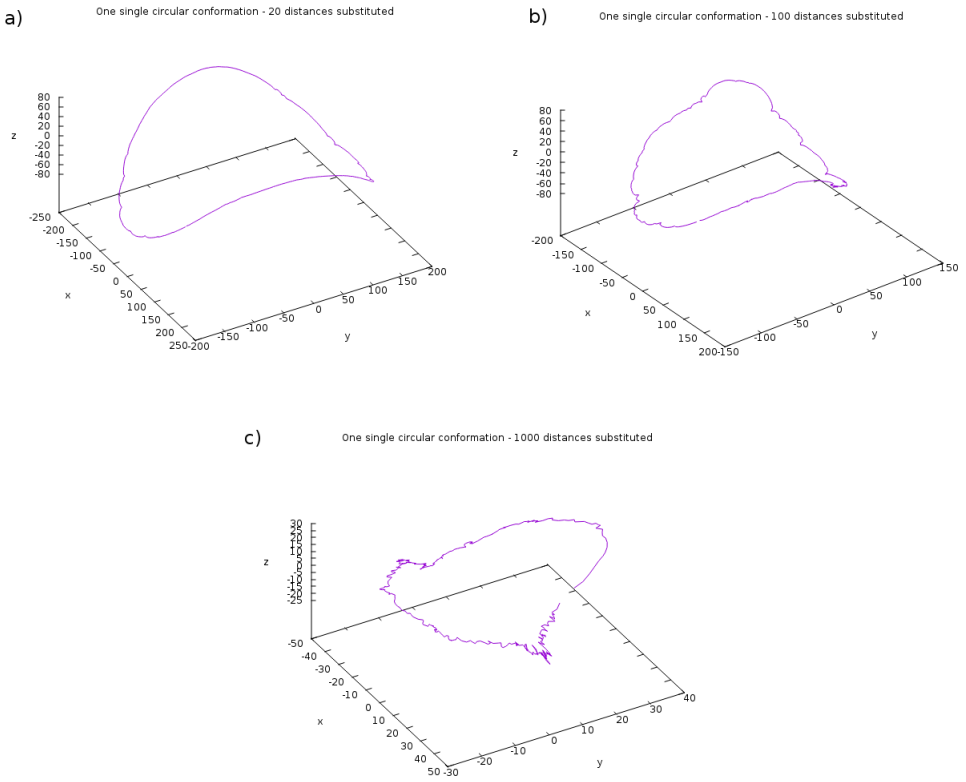

Figure 29: Reconstruction of the circular structure with partial substitution. The algorithm of the new frequencies has been applied to the circular structure using only (a) 20, (b) 100 or (c) 1000 distances of the matrix of the distances of one single conformation among all the simulated ones. In particular, the conformation used is the same shown in Fig. 7.

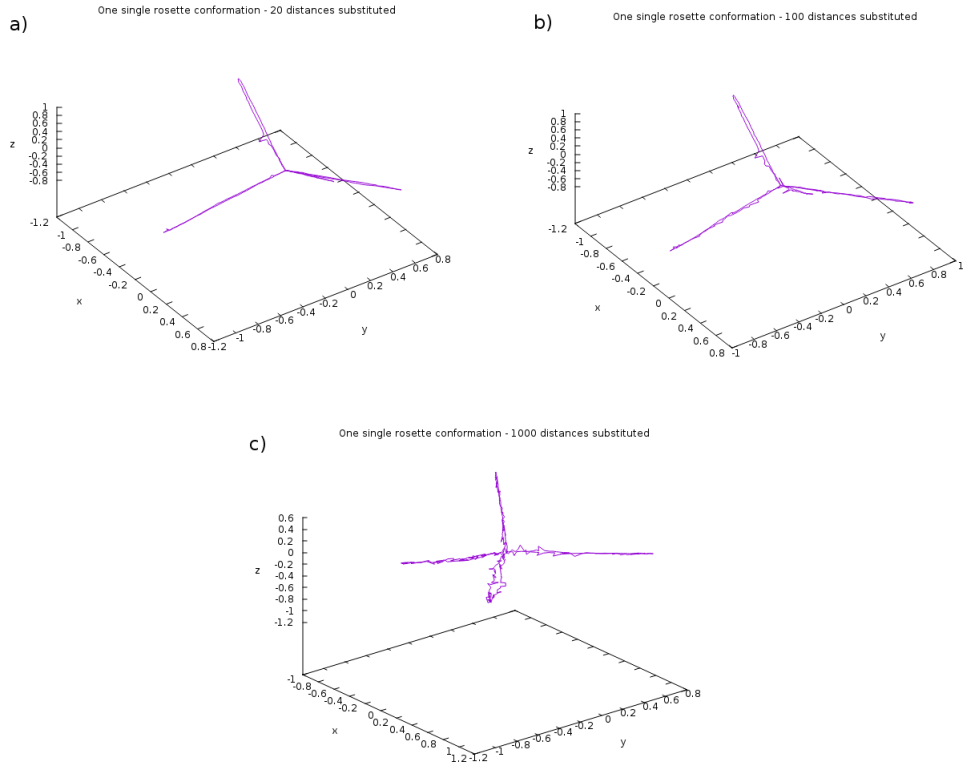

Figure 30: Reconstruction of the rosette structure with partial substitution. The algorithm of the new frequencies has been applied to the rosette structure using only (a) 20, (b) 100 or (c) 1000 distances of the matrix of the distances of one single conformation among all the simulated ones. In particular, the conformation used is the same shown in Fig. 7.

5 Conclusions

In this thesis, we investigated the structure of chromosomes, the molecules that carry all the genetic information of a living being. In fact, high order structural features may be connected with their function and with the gene expression. Despite its importance, at the state of the art there is not a technique to measure the coordinates or the distances between the loci of a chromosome directly and with high resolution. In the last years, Capturing Chromosome Conformation (3C) first and Hi-C later provided a method to obtain a genomewide topological information at high resolution.
In particular, the data consist in matrices of relative frequencies of contact between the loci. The goal of this thesis was to interpret this data in such a way to obtain a three dimensional reconstruction of the chromosome. This has been performed by means of the interpretation of the Hi-C matrix as the adjacency matrix of a weighted complete undirected graph. Therefore it was possible to define distances between the vertexes of the graph. We proposed different definitions of distances in graphs. In particular, we focused on the shortest path, resistance, natural proximity, entropy, potential proximity and Katz distances.
By defining these distances between the nodes of the graph - corresponding to the loci of the chromosome - we applied the Multidimensional scaling method in order to reconstruct the Euclidean coordinates.
Multidimensional scaling consists of generating a Gram or metric matrix M from the distance matrix. We showed that a necessary and sufficient condition to embed N points with the given distances in an Euclidean space of dimension r is M to be of rank r and semi-definite positive. The reconstructed coordinates' vector will be parallel to the eigenvectors of the metric matrix. If $r \geq 3$ we can reconstruct a projection in the first three coordinates.

Therefore, we simulated Hi-C data for three different polymers with a linear, circular and rosette shape. We calculated all the distance matrices for all of them and we applied the MDS for each of these distances. The reconstructed structures were not always equal each other and not always consistent with the expected structure. This has been calculated applying MDS to the average distances of all the simulated conformations.
The only distance whose reconstruction is always consistent is the resistance distance. Considering the Laplacian matrix L of the graph and a set of coordinates composed by its three eigenvectors corresponding to the lowest three

non zero eigenvalues, we obtain a reconstruction always consistent with the expected average structure of the polymers.
Any metric matrix that commutes with L, share with it a common eigenvector basis. The shortest path, resistance, proximity and Katz distances yield to this property. Among them, the resistance distance is the only one that provides the same reconstruction of the Laplacian eigenvectors. In fact, since its eigenvalues are the inverse of the Laplacian eigenvalues, the order of the eigenvectors is reversed.
We applied these algorithms to a set of real data too, obtaining further reconstructions.

Finally, we considered the problem of combining the Hi-C data with some distances directly measured with the fluorescence technique. For this purpose, we introduced a partition function on the energetic states that a graph can assume. Using it with a Bayesian approach, it is possible to modify the adjacency matrix and obtain a new reconstruction. Nevertheless, this method is not said to respect a priori the measured distances given as constraints.
Instead, forcing the measured distance to be the average distance obtained by means of the partition function, we found a quadratic inverse relationship between a measured distance and its corresponding contact frequency. As expected, using the complete set of average distances, this relationship is perfectly consistent with the expected average structure. Using the set of distances of one single conformation, the reconstruction is excellent for the x axis, but registers a lack of precision in the y and z axes. This is probably due to the average operation. We showed that this reconstruction modifies smoothly the shape of the polymer increasing the number of measured distances in the experiment. The insertion of only one new frequency influences even the other frequencies and the structure changes slightly. The reconstruction is not accurate enough, but it definitely suggests a basic method for further studies.

The goal of this thesis was to describe some theoretical basis on the multidimensional scaling process and on the application of this method in graphs. For this purpose we investigated the ways to define distances on graphs, their relationships with the Laplacian matrix and the conditions to embed the graph in an Euclidean space. During the work, many similarities with other physics fields have been underlined. Spectral graph theory is in

fact generally applicable to many areas. In this case the energy levels of the Laplacian have been interpreted as conditions for a three dimensional reconstruction of chromosomes. Further studies may continue to focus on this interpretation and on similar applications.
In fact, the study of chromosomes' structure is fundamental in biology and in medicine. It may unveil its relationship with chromosomes functionality or, in the worst case, its relationship with its dysfunctional role. In other words, it is important in order to understand the mechanisms of the transmission of life, that is finally the reason why we do research.

6 Appendix A

Given a complete set of distances between N points, the embedding problem consists in understanding when the given distances are Euclidean, i.e. when it is possible to find some Cartesian coordinates in a generic r-dimensional space ($r \leq N$). The following theorem establishes the minimum value of r and the necessary and sufficient conditions to embed the set of points.

Theorem. [47] A necessary and sufficient condition that the D_{ij} be the lengths of the edges of an n-"simplex" lying in $\mathcal{R}^r$ but not in $\mathcal{R}^{r-1}$ is that the quadratic form

$$F(x_1, \dots x_n) = \frac{1}{2} \sum_{i,j} (d_{0i}^2 + d_{0j}^2 - D_{ij}^2) x_i x_j$$

be positive (i.e. always ≥ 0) and of rank r. That is, the associate matrix M has to be positive semidefinite of rank r.

Proof. Necessity. Let $\mathbf{r}_0 = 0$ be the origin of an Euclidean space in which $\mathbf{r}_i$ are the coordinates of the i-th point. Considering a point P such that

$$\mathbf{r}_P = x_1 \mathbf{r}_1 + \dots + x_n \mathbf{r}_n$$

its distance from the barycenter is:

$$d_{0P}^2 = \sum_i x_i^2 \sum_\nu r_{i\nu}^2 + 2 \sum_{i<k} x_i x_k \sum_{\nu=1} r_{i\nu} r_{k\nu}$$

Since:

$$\sum_i x_i^2 \sum_\nu r_{i\nu}^2 = \sum_i x_i^2 r_{0i}^2 = 0$$

and:

$$2 \sum_\nu r_{i\nu} r_{k\nu} = \sum_\nu r_{i\nu}^2 + \sum_\nu r_{k\nu}^2 - \sum_\nu (r_{i\nu} - r_{k\nu})^2 = d_{0i}^2 + d_{0j}^2 - D_{ij}^2$$

we have $d_{0p}^2 = F(x_1 \dots x_n)$. Hence F is positive. Furthermore, $P = F = 0$ on a linear manifold of $n - r$ dimensions in the variables $x_1 \dots x_n$, hence F is of rank r.
Sufficiency. Let us assume F to be positive definite, i.e. $r = n$. By means

of a certain linear non singular transformation $y = H(x)$ we get $F = y_1^2 + \cdots + y_n^2$. Let's consider the Cartesian space of the variables $(y_1, \ldots y_n)$ and n correspondent points whose coordinates are:

$$(x_1, \ldots x_n) = (1, 0, \ldots 0), (0, 1, \ldots 0), \ldots, (0, 0, \ldots 1)$$

Their coordinates are found by $y = H(x)$, for their mutual distances we find:

$$\overline{x_0 x_i}^2 = F(0, \ldots 1 \ldots 0) = d_{0i}^2$$

$$\overline{x_i x_j}^2 = F(0, \ldots 1 \cdots -1 \ldots 0) = d_{0i}^2 + d_{0k}^2 - (d_{0i}^2 + d_{0k}^2 - d_{ik}^2) = d_{ik}^2$$

Hence, these are exactly the coordinates we are looking for. If $r < n$ then $F(x_1 \ldots x_n) = y_1^2 + y_2^2 + \cdots + y_r^2$ and the distances d_{0i}^2 and d_{ij}^2 are the squared lengths of the projections on the sub-space $(y_1 \ldots y_n)$, i.e. on a manifold $y_{r+1} = \cdots = y_n = 0$. By construction the n-simplex of points we found is contained in a $\mathcal{R}^r$ but not in $\mathcal{R}^{r-1}$.

7 Appendix B

Algorithms and codes.
The following algorithms are coded in language C++. We stressed the use of the template library *Eigen* [25], projected for linear algebra calculus. It allowed us to deal with matrices of high dimension and to compute quickly their spectral decomposition.
The input of any MDS algorithm is the Hi-C matrix. It is declared as `weight(N,N)` where `N` is the dimension of the matrix. In the following tables we show the algorithms to calculate the 6 distances matrices `dist(N,N)` used in the "Results" chapter from the `weight(N,N)` matrix.
We defined two Eigen types as follows:

```
typedef Matrix<long double, Dynamic, Dynamic> stdmat;
typedef Matrix<long double, 1, Dynamic> stdvec;
```

Shortest path.

```
long double eps=1e-20;
for (int i=0; i <N; i++) {
    for (int j=0; j <N; j++){
      if (abs(weight(i, j)) > eps)
   weight(i, j)=1./weight(i, j);
      if (abs(weight(i, j)) < eps)
   weight(i, j)=1e308;    }  } //theoretically infinity

for (int i=0; i<N; i++){
  for (int j=0; j<N; j++){
  dist(i, j)=weight(i, j);
    }  }

for(int i=0; i<N; i++){
   dist(i,i)=0.;
  }

int correct=1;
while(correct != 0){
//we iterate until all points respect the triangle inequality (correct=0)
  correct=0;
```

```
  for(int i=0; i<N;i++){
    for(int j=0; j<N;j++){
      for(int k=0;k<N;k++){
       if( ( dist(i, j)>( dist(i, k)+dist(k, j) )  )
        && (abs(dist(i, k))>eps)  && (abs(dist(k, j))>eps)  ) {
          dist(i, j)=dist(i, k)+dist(k, j);
          dist(j,i)=dist(i,j);
}   }   }  }
  for(int i=0; i<N;i++){
    for(int j=0; j<N;j++){
      for(int k=0;k<N;k++){
       if(  ( dist(i, j)>(dist(i, k)+dist(k, j)) )
        && (abs(dist(i, k))>eps)
        && (abs(dist(k, j))>eps)) {
          correct++;
}  }  }  }  } }
```

Resistance.

```
// Calculus of the Laplacian matrix
  stdvec rows(N);

  for (int i=0; i<N; i++){
    rows(i)=0.;
  }

  for (int i=0; i<N; i++){
    for (int j=0; j<N; j++){
      rows(i)=rows(i)+weight(i,j);
    } }

  for (int i=0; i<N; i++){
    for (int j=0; j<N; j++){
      weight(i,j)=-weight(i,j);
    }  }

  for(int i=0; i<N;i++){
    weight(i,i)=rows(i); }
```

```
  for (int i=0; i<N; i++){
  for (int j=0; j<N; j++){
    weight(i,j)=weight(i,j)+1./N;
  }
}

weight = weight.inverse();

for (int i=0; i<N; i++){
  for (int j=0; j<N; j++){
    weight(i,j)=weight(i,j)-1./N;
  }
}
for (int i=0; i<N; i++){
  for (int j=0; j<N; j++){
    dist(i, j)=weight(i, i)+weight(j,j)-2*weight(i,j);
  }  }
```

Natural proximity.

```
stdvec rows(N);

for (int i=0; i<N; i++){
  rows(i)=0.;
}

for (int i=0; i<N; i++){
  for (int j=0; j<N; j++){
    if( i!= j){
    rows(i)=rows(i)+weight(i,j);
  }}  }

stdmat prox(N,N); //proximity matrix
for(int i=0; i<N;i++){
  for(int j=0; j<N; j++){
    if (i != j){
      prox(i,j)=weight(i,j); }
    if (i == j){
      prox(i,j)=rows(i);}
```

```
   }  }

//Does it respect the proximity properties?

 int triangle =0;
  for (int i=0; i<N; i++) {
   for (int j=0; j<N; j++){
     if ( abs(prox(i,j) -prox(j,i))>eps) {
  cout << "Error symmetry " << i << "," << j << endl;}
     if ( prox(i,i) < prox(j,i) ) {
  cout << "Error maxim diagonal: " << i << "," << j << endl;}
     if ((prox(i,j))<0) {
  cout << "Error negative: " << i << "," << j << endl;}
 for(int k=0; k<N; k++) {
        if ((prox(i,j)+prox(i,k)-prox(j,k))>prox(i,i) && i != j && i!= k){
          cout << i << "  " << j << "  " << k << endl;
          triangle++;}}
      }    }
 cout << "Triangle = " << triangle << endl;

 stdmat dist(N,N);  // distance matrix
 for (int i=0; i<N; i++){
    for (int j=0; j<N; j++){
      dist(i, j)=prox(i,i)+prox(j,j)-2*prox(i,j);
     }    }
```

Entropy distance.

```
 for (int i=0; i<N; i++){
   for (int j=0; j<N; j++){
     if (abs(weight(i,j))> eps) {
      weight(i,j)= -weight(i,j)*log(weight(i,j));
     }
     if (abs(weight(i,j))< eps) { weight(i,j)=0.;}
   }  }

 stdvec rows(N);
 for (int i=0; i<N; i++){
   rows(i)=0.;
```

```
  }

  for (int i=0; i<N; i++){
    for (int j=0; j<N; j++){
      if( i!= j){
       rows(i)=rows(i)+weight(i,j);
    } }  }

  stdmat prox(N,N); //proximity matrix
  for(int i=0; i<N;i++){
    for(int j=0; j<N; j++){
      if (i != j){
       prox(i,j)=weight(i,j); }
      if (i == j){
       prox(i,j)=rows(i);}
    }  }
\\Check if it verifies the proximity properties as done before
  for (int i=0; i<N; i++){
    for (int j=0; j<N; j++){
      dist(i, j)=prox(i,i)+prox(j,j)-2*prox(i,j);
  }    }
```

The potential proximity distance is analogous.

Katz distance.

```
  for(int i=0;i<N;i++){
    for(int j=0;j<N;j++){
      weight(i,j)=-weight(i,j);
    }
  }
  for(int i=0;i<N;i++){
    weight(i,i)=1;
  }
  weight = weight.inverse();
  for(int i=0;i<N;i++){
    weight(i,i)=weight(i,i)-1;
  }
```

```
stdvec rows(N);

for (int i=0; i<N; i++){
  rows(i)=0.;
}

for (int i=0; i<N; i++){
  for (int j=0; j<N; j++){
    rows(i)=rows(i)+weight(i,j);
  }
}
stdmat dist(N,N);  //minimum path distance matrix
for (int i=0; i<N; i++){
  for (int j=0; j<N; j++){
    dist(i, j)=weight(i, i)+weight(j,j)-2*weight(i,j);
}  }
```

For each of these distance matrices we applied the MDS algorithm:

```
  stdvec somma(N);
  somma.setZero();// useful on the calculation of the barycenter distance
  long double somma2 = 0.;

  stdvec distbaryc(N);
  distbaryc.setZero();//distances from the barycenter
  for (int i=0; i<N; i++){
    for (int j=0; j<N; j++){
     somma(i)=somma(i)+dist(i, j)*dist(i, j);
    }
  }
  for (int j=0; j<N; j++){
    for (int k=0; k<j+1; k++){
     somma2=somma2+dist(j, k)*dist(j, k);
    }
  }

  for (int i=0; i < N; i++) {
    distbaryc(i)=somma(i)/N - somma2/(N*N);
```

```
    if (distbaryc(i) < 0)
     distbaryc(i)=0.;
  }

  stdmat gram(N,N); //Gram or metric matrix
  for (int i=0; i<N; i++){
    for (int j=0; j<N; j++){
      gram(i, j)=(distbaryc(i)+distbaryc(j)-dist(i, j)*dist(i, j))/2.;
  }   }

  EigenSolver<stdmat> es;
  es.compute(gram, true);

  stdvec eigenval = es.eigenvalues().real();
  stdmat eigenvec = es.eigenvectors().real();

  stdmat coord(N,N);
  long double eigenvalues[N];
  int neg=0;
  for (int i=0; i < N; i++) {
    eigenvalues[i]=eigenval(i); }

  //ordering the eigenvalues and the eigenvectors
  long double temp=10.;
  long double temp2=10.;
  int m=0;

  while(m<N-1){
   if( eigenvalues[m] < eigenvalues[m+1] ) {
     temp=eigenvalues[m];
        eigenvalues[m]=eigenvalues[m+1];
        eigenvalues[m+1]=temp;
        for (int j=0; j<N; j++){
          temp2=eigenvec(j, m);
          eigenvec(j, m)=eigenvec(j,m+1);
          eigenvec(j, m+1)=temp2;
}
 m=0;
```

```
}
else
     m++;
    }

   //calculate the coordinates
    for (int i=0; i<N; i++){
     if(eigenvalues[i]<0.){
        eigenvalues[i]=0.;} // set to zero the negative eigenvalues
     for (int j=0; j<N; j++){
       coord(i, j)=eigenvec(j, i)*((eigenvalues[i]));
     }    }
```

References

[1] Erik D Andrulis, Aaron M Neiman, David C Zappulla, and Rolf Sternglanz. Perinuclear localization of chromatin facilitates transcriptional silencing. *Nature*, 394(6693):592–595, 1998.

[2] Rolf Backofen, Markus Fricke, Manja Marz, Jing Qin, and Peter F Stadler. Distribution of graph-distances in boltzmann ensembles of rna secondary structures. In *International Workshop on Algorithms in Bioinformatics*, pages 112–125. Springer, 2013.

[3] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591, 2001.

[4] Marek Biskup. On the scaling of the chemical distance in long-range percolation models. *Annals of Probability*, pages 2938–2977, 2004.

[5] G. Caldarelli. *Scale-free networks: complex webs in nature and technology*. Oxford University press, 2007.

[6] I Carmesin and Kurt Kremer. The bond fluctuation method: a new effective algorithm for the dynamics of polymers in all spatial dimensions. *Macromolecules*, 21(9):2819–2823, 1988.

[7] Pavel Chebotarev. A class of graph-geodetic distances generalizing the shortest-path and the resistance distances. *Discrete Applied Mathematics*, 159(5):295–302, 2011.

[8] Pavel Chebotarev. The graph bottleneck identity. *Advances in Applied Mathematics*, 47(3):403–413, 2011.

[9] Pavel Chebotarev and Elena Shamis. Matrix-forest theorems. *arXiv preprint math/0602575*, 2006.

[10] Pavel Chebotarev and Elena Shamis. On proximity measures for graph vertices. *arXiv preprint math/0602073*, 2006.

[11] Jie Chen and Ilya Safro. Algebraic distance on graphs. *SIAM Journal on Scientific Computing*, 33(6):3468–3490, 2011.

[12] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[13] Daniel M Cimbora and Mark Groudine. The control of mammalian dna replication: a brief history of space and timing. *Cell*, 104(5):643–646, 2001.

[14] Ewan R Colman and Nathaniel Charlton. Separating temporal and topological effects in walk-based network centrality. *Physical Review E*, 94(1):012313, 2016.

[15] Thomas Cremer and Christoph Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews genetics*, 2(4):292–301, 2001.

[16] GORDON M Crippen and Timothy F Havel. Stable calculation of coordinates from distance information. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 34(2):282–284, 1978.

[17] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *science*, 295(5558):1306–1311, 2002.

[18] MV Dhanyamol and Sunil Mathew. Distances in weighted graphs. *Annals of Pure and Applied Mathematics*, 8(1):1–9, 2014.

[19] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[20] Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, 2015.

[21] Leon Festiger. The analysis of sociograms using matrix algebra. *Human relations*, 1949.

[22] Robert W Floyd. Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345, 1962.

[23] LR Ford Jr and DR Fulkerson. Maximal flow through a network. In *Classic papers in combinatorics*, pages 243–248. Springer, 2009.

[24] John Clifford Gower. Euclidean distance geometry. *Math. Sci*, 7(1):1–14, 1982.

[25] Gaël Guennebaud, Benoît Jacob, et al. Eigen v3. http://eigen.tuxfamily.org, 2010.

[26] Ivan Gutman. The energy of a graph: old and new results. In *Algebraic combinatorics and applications*, pages 196–211. Springer, 2001.

[27] Timothy F Havel, Irwin D Kuntz, and Gordon M Crippen. The theory and practice of distance geometry. *Bulletin of Mathematical Biology*, 45(5):665–720, 1983.

[28] Donald B Johnson. Efficient algorithms for shortest paths in sparse networks. *Journal of the ACM (JACM)*, 24(1):1–13, 1977.

[29] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[30] Douglas J Klein and Milan Randić. Resistance distance. *Journal of Mathematical Chemistry*, 12(1):81–95, 1993.

[31] Philip A Knight. The sinkhorn-knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

[32] Allen Knutson and Terence Tao. Honeycombs and sums of hermitian matrices. *Notices Amer. Math. Soc*, 48(2), 2001.

[33] Peter Kuchment. Quantum graphs: an introduction and a brief survey. *arXiv preprint arXiv:0802.3442*, 2008.

[34] Tung BK Le, Maxim V Imakaev, Leonid A Mirny, and Michael T Laub. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science*, 342(6159):731–734, 2013.

[35] Annick Lesne, Julien Riposo, Paul Roger, Axel Cournac, and Julien Mozziconacci. 3d genome reconstruction from chromosomal contacts. *Nature methods*, 11(11):1141–1143, 2014.

[36] Bruno Lévy. Laplace-beltrami eigenfunctions towards an algorithm that" understands" geometry. In *IEEE International Conference on Shape Modeling and Applications 2006 (SMI'06)*, pages 13–13. IEEE, 2006.

[37] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[38] Martial Marbouty, Antoine Le Gall, Diego I Cattoni, Axel Cournac, Alan Koh, Jean-Bernard Fiche, Julien Mozziconacci, Heath Murray, Romain Koszul, and Marcelo Nollmann. Condensin-and replication-mediated bacterial chromosome folding and origin condensation revealed by hi-c and super-resolution imaging. *Molecular cell*, 59(4):588–602, 2015.

[39] Patrick N McGraw and Michael Menzinger. Laplacian spectra as a diagnostic tool for network structure and dynamics. *Physical Review E*, 77(3):031102, 2008.

[40] Efe A. Ok. *Real Analysis with Economic Applications*. Princeton University press, 2011.

[41] Beresford N Parlett. *The symmetric eigenvalue problem*, volume 7. SIAM, 1980.

[42] Roger Penrose. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge Univ Press, 1955.

[43] Seth Pettie and Vijaya Ramachandran. Computing shortest paths with comparisons and additions. In *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 267–276. Society for Industrial and Applied Mathematics, 2002.

[44] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human

genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[45] Eduardo PC Rocha. The organization of the bacterial genome. *Annual review of genetics*, 42:211–233, 2008.

[46] Norbert Rosenzweig and Charles E Porter. " repulsion of energy levels" in complex atomic spectra. *Physical Review*, 120(5):1698, 1960.

[47] Isaac J Schoenberg. Remarks to maurice frechet's article"sur la definition axiomatique d'une classe d'espace distances vectoriellement applicable sur l'espace de hilbert. *Annals of Mathematics*, pages 724–732, 1935.

[48] Jonatan Schroeder, ALP Guedes, and Elias P Duarte Jr. Computing the minimum cut and maximum flow of undirected graphs. *RelatórioTécnico RT-DINF*, 3, 2004.

[49] Claude E Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.

[50] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[51] Daniel Spielman. Spectral graph theory. *Lecture Notes, Yale University*, pages 740–0776, 2009.

[52] Karen Stephenson and Marvin Zelen. Rethinking centrality: Methods and examples. *Social Networks*, 11(1):1–37, 1989.

[53] William R Taylor and András Aszódi. *Protein geometry, classification, topology and symmetry: A computational analysis of structure*. CRC Press, 2004.

[54] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[55] Eric W Weisstein. Moore-penrose matrix inverse. 2002.

[56] Ryan Williams. Faster all-pairs shortest paths via circuit complexity. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 664–673. ACM, 2014.

[57] Alan Wolffe. *Chromatin: structure and function*. Academic press, 1998.

[58] Wenjun Xiao and Ivan Gutman. Resistance distance and laplacian spectrum. *Theoretical Chemistry Accounts*, 110(4):284–289, 2003.

[59] Forrest W Young. *Multidimensional scaling: History, theory, and applications*. Psychology Press, 2013.

[60] Gale Young and Alston S Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3(1):19–22, 1938.

Acknowledgments

I owe my deepest gratitude to professor Dieter W. Heermann who offered to me to work on this thesis and warmly welcomed me in his students' team. My thanks are especially for his trust on my job, for his constant and professional presence and for his precious advices. I'm also really thankful to my desk-neighbour Andreas Hofmann who has always been really patient when I was lost in some points of the job and helped me every time I needed. It has been a pleasure working with both of you and be introduced to this interesting field of biophysics.
I definitely owe my thanks to professor Flavio Seno for his availability. He accepted immediately to be my supervisor, even if we would have not worked directly together, and he was always willing to help me as necessary.
I am in debt with Jan Kapar, for his widespread attention to my mistakes in English, and to my brother Bruno, for his good ideas and for listening to me as always when doubts arose during the work.

This thesis talks about distances and connections. Somehow, my last two years spoke to me about it as well. I must thank who was part of these connections during my studies.
There are persons that have been close to me wherever I was and whatever I was doing, they were in my heart: my family.
Then there are other persons that behaved like a family. Someone followed me "from the very first day until the last one", Giuliana, and someone else joined and supported me during this awesome journey. Even if distant for such a long time, my Sicilian friends have always been my landmark. In Padova, Federica, Michele (even because whenever he hosted me I was feeling at home), Pasquale, Laura, Matteo, Michele, Ciccio. Huge distances may even grew up among you, but I am sure that a strong connection will always exist. Then a "family" came in Padova from distant places and when then *I* went to distant places, as Boston or Heidelberg, I found special friends. Thanks, you simply made everything amazing, I won't forget it.
For me it is not easy to be close to everyone even when I am very far away. But such as in this thesis, behind certain distances there is always a solid structure. The beauty of life is hidden on reconstructing this structure together.
Finally, I say thanks even to my weaknesses, because when you are not afraid of them, they make you stronger. And happy.

Vorrei ringraziare prima di tutto il professore Dieter W. Heermann che mi ha offerto di lavorare a questa tesi e molto caldamente mi ha accolto nel suo team. Ho apprezzato molto l'immediata fiducia che mi ha dato, la sua costante e professionale presenza, i suoi preziosi consigli sono stati indispensabili. Ringrazio anche il mio vicino di scrivania, Andreas Hofmann, sempre paziente con me quando ero in difficoltà, aiutandomi ogni volta fosse necessario. E' stato davvero un piacere avere avuto l'opportunità di lavorare con voi e cimentarmi in questo interessante campo della biofisica.
Devo sicuramente dei ringraziamenti al professore Flavio Seno per la sua disponibilità e professionalità. Ha accettato immediatamente di seguirmi, nonostante non avremmo lavorato direttamente insieme, ed è sempre stato disposto ad aiutarmi quando necessario.
Sono in debito con Jan Kapar, per la sua capillare attenzione ai miei errori con l'inglese, e a mio fratello Bruno per le sue buone idee e per ascoltarmi, come sempre, quando a volte sorgeva qualche dubbio durante il mio lavoro.

Questa tesi parla di distanze, di connessioni e, in qualche modo, credo che anche questi ultimi due anni me ne abbiano parlato. Per questo credo di dovere dei ringraziamenti a chi di queste connessioni ha fatto parte, sostenendomi durante il mio percorso. In primo luogo a quelle persone che, non importa quanto distante io vada o che cosa io faccia, ci saranno sempre, al mio fianco, nel mio cuore: la mia famiglia.
C'è stato poi chi, sebbene non lo fosse, si è sempre comportato come parte di una famiglia durante questi anni di studi. Qualcuno c'è stato "dal primo fino all'ultimo giorno", Giuliana, altri invece si sono uniti durante il percorso. Nonostante la lontananza, i miei amici in Sicilia sono sempre rimasti un fermo punto di riferimento. A Padova, Federica, Michele (che ogni volta che mi ha ospitato mi ha fatto sentire a casa), Pasquale, Laura, Matteo, Michele, Ciccio. Potrebbero sorgere distanze enormi tra di voi, ma sono sicuro che una forte connessione esisterà sempre. Poi una "family" è venuta a Padova da molto lontano e quando invece sono stato io ad andare molto lontano, come a Boston o Heidelberg, ho trovato amici speciali. Grazie di cuore perché avete semplicemente reso tutto fantastico, non lo dimenticherò mai.
Non è sempre stato facile per me essere vicino a tutti in questi anni. Ma proprio come in questa tesi, sono convinto che dietro certe distanze ci sia sempre una struttura solida. Il bello sta proprio nel ricostrurirla, insieme.

Infine, ringrazio le mie debolezze, perché, quando non fanno più paura, ti rendono più forte. E felice.