EPFL

Thèse n° 8784

Landscapes of DNA Mechanics and Genomes

Présentée le 3 février 2023

Faculté des sciences de base
Chaire d’analyse appliquée
Programme doctoral en mathématiques

pour l’obtention du grade de Docteur ès Sciences

par

Thomas Antonin ZWAHLEN

Acceptée sur proposition du jury

Prof. F. Eisenbrand, président du jury
Prof. J. Maddocks, directeur de thèse
Prof. M. Meila, rapporteuse
Prof. R. Rohs, rapporteur
Prof. E. Abbé, rapporteur

École
polytechnique
fédérale
de Lausanne

2023

to life

Acknowledgements

First of all, I want to thank my supervisor John Maddocks, who led like a captain even in the worst weather. This work would not have been possible without the friendly support of the past and present members of the LCVMM group: Alex, Alessandro, Jarek, Pauline, Alastair, Thomas, Lennart, Giulio, Rahul, Jannes, Raushan, Harmeet, Carine, Marina; very special thanks to all my beloved friends: Aurelien, Quentin, Marc, Eugénie, Dan, Claudia, Boris, Natacha, Léa, Gmür, Kuno, Adrien, Constant, Matthieu, Andres, Lilith, Raphaël, Camille, Luis, Camille, Romain, Athéna, and all the ones I will realise I have forgotten to include in that list; to my family, for their love; to Jean-Philippe and Chris Martin, for all the emotional support. Last but not least, thanks to Cassandre, for all the joy and laughter we share together.

This work was supported by the Swiss National Science Foundation, project numbers 200020_163324 and 200020_182184.

Lausanne, November 10, 2022 T. Z.

Abstract

DNA is of fundamental interest in science, of course playing a central role in biology and medicine, but also being a subject of great interest in chemistry, physics, and nanotechnology. This thesis is a contribution towards bridging the different length scales and level of detail in descriptions inherent to multiple different perspectives on DNA, from fully atomistic Molecular Dynamics computer simulations, through the statistical mechanics of (hetero-)polymers, to genomics and bioinformatics.

The local physical properties – such as intrinsic shape and flexibility – of the DNA double-helix are today widely believed to be influenced by its specific base sequence in a highly significant way. Furthermore, there is strong evidence that these properties play a role in many important cellular processes, from chromatin compaction to protein binding and nucleosome positioning. In order to address such biologically pertinent problems, our aim is to develop mathematical and computational tools based on the sequence-dependent mechanical properties of the double-helix so as to be able to both predict and identify exceptional sequences and sites within genome-length data sets.

For this, we build on the previously developed cgDNA+ coarse-grain model, which provides a detailed sequence-dependent description of the statistical mechanics of DNA. The cgDNA+ model is trained on Molecular Dynamics (MD) simulation data, and has been shown to reproduce statistics drawn from MD time series to a remarkably high degree of accuracy for both training set and test simulations. For a given sequence of any finite length, it predicts a multivariate Gaussian distribution on a set of configuration coordinates corresponding to standard degrees of freedom of DNA (tilt, roll, twist, etc.), with the particular property that the inverse covariance (or *stiffness*) matrix is *banded*, with an overlapping squares sparsity pattern. A particular feature of the cgDNA+ model is that its mean, or ground state, has a significant, non-local dependence on the DNA sequence, as is known to be the case in reality.

The first main result of this thesis concerns the marginals of such Gaussian distributions. Namely, we present a pure linear algebra derivation of the fact that marginals of Gaussians with inverse covariances with an overlapping square sparsity pattern also have banded inverse covariance matrices, with the same sparsity pattern. Our proof of this elementary, but for us fundamental, basic result further provides a highly efficient procedure for computing such marginal inverse covariance matrices. This marginalisation procedure is then implemented in the specific context of the cgDNA+ model as a computational tool which we have named cgDNAloc. The computational efficiency of cgDNAloc allows large (e.g. with millions of elements) ensembles of marginal (but still quite high-dimensional) Gaussians to be computed by sliding a window along genomic length-scale sequences. cgDNAloc also allows for an averaged marginalisation to be computed for a given sequence fragment embedded in all possible flanking sequences, which, due to the challenge of the non-local sequence-dependence of DNA mechanics, is necessary to assign particular statistical mechanics properties to specific short sequence fragments in a large variety of sequence contexts.

As a first application of cgDNAloc, we introduce and use some dimensionality reduction methods to visualise and cluster cgDNA+ predictions on exhaustive ensemble of k-mers, embedded in flanking sequences. In particular, we apply a form of Fisher-Information weighted Principal Component Analysis to ensembles of cgDNA+ Gaussians of k-mers, with the desirable property to be invariant under a linear change of coordinates. This methods yields a striking clustering of all k-mer sequences based only on their base pair content in the purine/pyrimidine alphabet. This clustering is much less clear cut in a standard PCA analysis. As a second, illustrative application, we present a method inspired from information theory techniques to scan the genome of S. cerevisiae (brewer's yeast) in search of mechanically exceptional sequences, thus using cgDNAloc to bridge the gap in scales between atomistic models and genomics.

The body of the thesis concerns sequence-dependence in the standard $\{A, T, C, G\}$ base alphabet. However within biology it is now known that base modifications such as methylation play an important role. In a discussion of future work in the Conclusions section some preliminary results are presented which indicate that methylation has a very strong impact on the statistical mechanics of the DNA double helix.

Keywords: DNA mechanics, banded model, genomics, outlier detection, information theory, dimensionality reduction.

Résumé

L'ADN présente un intérêt fondamental pour la science, jouant bien sûr un rôle central en biologie et en médecine, mais étant également un sujet de grand intéret en chimie, en physique et en nanotechnologie. Cette thèse est une contribution à l'établissement d'un pont entre les différentes échelles de longueur et le niveau de détail des descriptions inhérentes à de multiples perspectives différentes sur l'ADN, depuis les simulations informatiques de dynamique moléculaire entièrement atomistiques jusqu'à la génomique et la bioinformatique, en passant par la mécanique statistique des (hétéro)polymères.
Les propriétés physiques locales - telles que la forme et la flexibilité intrinsèques - de la double hélice de l'ADN sont aujourd'hui largement considérées comme étant influencées de manière très significative par sa séquence de bases spécifique. En outre, il existe des preuves solides que ces propriétés jouent un rôle dans de nombreux processus cellulaires importants, de la compaction de la chromatine à la liaison des protéines et au positionnement des nucléosomes. Afin d'aborder ces problèmes biologiquement pertinents, notre objectif est de développer des outils mathématiques et informatiques basés sur les propriétés mécaniques de la double hélice dépendant de la séquence, afin de pouvoir à la fois prédire et identifier des séquences et des sites exceptionnels dans des ensembles de données de longueur génomique.

Pour ce faire, nous nous appuyons sur le modèle à gros grains cgDNA+ précédemment développé, qui fournit une description détaillée de la mécanique statistique de l'ADN en fonction de la séquence. Le modèle cgDNA+ est entrainé sur des données de simulation de dynamique moléculaire. Pour une séquence donnée de longueur finie, il prédit une distribution gaussienne multivariée sur un ensemble de coordonnées de configuration correspondant aux degrés de liberté standard de l'ADN (inclinaison, roulis, torsion, etc.), avec la propriété particulière que la matrice de covariance inverse (ou matrice de rigidité) est creuse, avec un modèle de sparsité à blocs carrés superposés. Une caractéristique particulière du modèle cgDNA+ est que sa moyenne, ou état fondamental, a une dépendance significative et non locale sur la séquence d'ADN, comme c'est le cas dans la réalité.

Le premier résultat principal de cette thèse concerne les marginales de telles distributions gaussiennes. Plus précisément, nous présentons une dérivation purement à base d'algèbre linéaire que les marginales des gaussiennes dont les covariances inverses ont un motif de sparsité en blocs carrés superposés ont également des matrices de covariance inverse en bandes, avec le meme motif de sparsité. Notre preuve de ce résultat de base élémentaire, mais pour nous fondamental, fournit en outre une procédure très efficace pour calculer ces matrices de covariance inverse marginales. Cette procédure de marginalisation est ensuite implémentée dans le contexte spécifique du modèle cgDNA+ en tant qu'outil de calcul que nous avons nommé cgDNAloc. L'efficacité de calcul de cgDNAloc permet de calculer de grands ensembles (par exemple, des millions d'éléments) de distributions marginales gaussiennes (mais toujours de très haute dimension) en faisant glisser une fenêtre le long de séquences d'échelles de longueur génomiques. cgDNAloc permet également de calculer une marginalisation moyenne pour un fragment de séquence donné intégré dans toutes les séquences flanquantes possibles, ce qui, en raison du défi posé par la dépendance de la séquence non locale de la mécanique de l'ADN, est nécessaire pour attribuer des propriétés de mécanique statistique particulières à de courts fragments de séquence spécifiques.

Les résultats de l'analyse de l'ADN peuvent être utilisés dans une grande variété de contextes de séquence.

Comme première application de cgDNAloc, nous introduisons et utilisons certaines méthodes de réduction de la dimensionnalité pour visualiser et partitionner les prédictions de cgDNA+ sur un ensemble exhaustif de k-mers, intégrés dans des séquences flanquantes. En particulier, nous appliquons une forme d'analyse en composantes principales pondérée par l'information de Fisher à des ensembles de distributions cgDNA+ de k-mers.

De manière frappante, cette méthode permet de regrouper toutes les séquences de k-mer en se basant uniquement sur leur contenu en paires de bases dans l'alphabet purine/pyrimidine. Cette partition est beaucoup moins nette dans une analyse PCA standard. Comme seconde application illustrative, nous présentons une méthode inspirée des techniques de la théorie de l'information pour scanner le génome de S. cerevisiae (levure de bière) à la recherche de séquences mécaniquement exceptionnelles, utilisant ainsi cgDNAloc pour combler le fossé des échelles entre les modèles atomistiques et la génomique.

Le corps de la thèse concerne la dépendance des séquences dans l'alphabet de base standard $\{A, T, C, G\}$. Cependant, en biologie, on sait maintenant que les modifications des bases telles que la méthylation jouent un role important. Dans une discussion sur les travaux futurs dans la section Conclusion, quelques résultats préliminaires sont présentés, qui indiquent que la méthylation a un impact très important sur la mécanique statistique de la double hélice de l'ADN.

Contents

Acknowledgements **v**

Abstract (English/Francais) **vii**

Introduction **3**

PART I: Background **11**

1 cgDNA+: a sequence dependent coarse-grained model of DNA mechanics . . . 12
  1.1 Protein Binding Sites and DNA mechanics . . . 19
2 Elements on Gaussian distributions . . . 24
  2.1 Some Facts on Gaussians . . . 24
  2.2 Divergences, distances and metrics on Multivariate Gaussians . 26
  2.3 A note on scale invariance . . . 30
  2.4 Averaging Gaussians . . . 31
3 Metric PCA: a coordinate-invariant Principal Component Analysis . . . 34
  3.1 Principal Component Analysis . . . 34
  3.2 A Fisher metric PCA . . . 35

PART II: cgDNAloc and Applications **41**

4 cgDNAloc: dealing with non-locality . . . 42
  4.1 A prior algorithm for maximum entropy fitting of banded covariances with overlapping squares . . . 43
  4.2 Marginals of banded inverse covariances are banded . . . 54
  4.3 Computing the inverse of marginal covariance with banded inverse: a fast algorithm . . . 58
  4.4 Non-locality and cgDNA+ marginals . . . 60
5 Visualising and Clustering cgDNA+ predictions . . . 68
  5.1 An exhaustive study of short sequence fragments . . . 68
  5.2 Comparison with the $MDNA$ parameter set . . . 86
6 Scanning genomes for outlier sequences . . . 90
  6.1 A toolbox to compare cgDNA+ distributions . . . 91

6.2 Typical and atypical sequences, from tens to hundreds of base pairs . . . 92

Conclusion and Discussion **109**

Bibliography **118**

Introduction

It is arguable that DNA is one of the most famous molecules in history. Since the description of its structure by Franklin, Wilkins, Watson and Crick in 1953, the double helix has become a fundamental cornerstone in biology, and remains today at the centre of our understanding of life. Beyond this role in the life sciences and medicine, the DNA molecule is also of great interest to chemists, physicists, materials scientists (for nanotechnology) and computer scientists (for its unique structural and dynamical features that allow it to store information in both a stable and flexible way[1]).

In its most common form DNA molecules form right-handed double helices of two chains of bases or nucleotides. Briefly, the fundamental units of DNA are nucleotides; they are formed by a nucleobase coupled to a pentose sugar (deoxyribose) and a phosphate group. The base in each nucleotide is of four possible types: adenine (A), cytosine (C), guanine (G), and thymine (T), and the type of base encodes the sequence of the DNA. The four bases are divided into two types of heterocyclic aromatic compounds: adenine and guanine have two rings and are called purines (denoted by R), while cytosine and thymine have a single ring and are called pyrimidines (and are denoted by Y). In particular purines are approximately twice as big as pyrimidines. The base in each nucleotide binds to a base in another complementary nucleotide through hydrogen bonds, always (in standard DNA) a purine with a pyrimidine, and more specifically with A always binding to T with two hydrogen bonds, and C binding to G with three hydrogen bonds. The resulting structures are called base pairs, sometimes meaning just the two bases, and sometimes meaning the full two nucleotides, i.e. with the sugars and phosphates included. The base pairs are then covalently (i.e. very strongly compared to hydrogen bonds) linked to each adjacent neighbour through the sugar-phosphate backbones, which are formed because the phosphate in each nucleotide is covalently bonded to the sugar ring of an adjacent nucleotide. (The bases in adjacent nucleotides can also interact directly through various mechanisms, e.g. mutual avoidance, and other so-called stacking interactions.) The detailed chemistry of how the phosphate group attaches to the adjacent sugar implies a direc-

---

[1] For example, it has recently been used to encode entire pieces of music [1].

tionality or orientation of the sugar-phosphate backbone via a conventional numbering of the carbon atoms inside the closed, deoxyribose sugar ring. By convention the base sequence in the nucleotides forming a single backbone is always read and written from the 5' (or phosphoryl) end to the 3' (or hydroxyl group) end, or, more briefly, in the 5'-3' direction, with the sequence in the $\{A, T, C, G\}$ alphabet. A significant point for this work is that the two backbones in the double helix have opposite orientations, so that the double helix is formed by two *antiparallel* DNA backbones, or Watson and Crick strands, meaning that the 5' end of one strand, say Watson, always corresponds to the 3' end of the other, say Crick, strand. In particular after one particular backbone is picked as the reading, or Watson, strand, the sequence, or primary structure, of a double helical fragment of DNA is just an arbitrary string of letters from the four element alphabet. In bioinformatics this string is routinely of millions of letters in length, while the physical properties of specific sequences are known to vary significantly between fragments at the length scale of only 10 base pairs. And in the four letter alphabet there are already more than a million different possible 10-mer sequences.

A significant point in what follows is that if the choice of reading strand is switched from the Watson to Crick strands, then the AT and GC base pairing rules combined with the 5' to 3' sequence reading convention and the anti-parallel backbones implies a transformation of sequence. For example for a four base pair tetramer (or 4-mer) the sequence (5')ATCG(3') is the same physical eight nucleotide double helical fragment as the sequence (5')CGAT(3'), where here we have explicitly kept the 5' and 3' end labels, which are typically suppressed. Thus the number of independent sequences for a fragment of given length is approximately halved, although the presence of palindromic sequences complicates the count slightly. A palindromic sequence is one where the Watson and Crick sequence reads are identical, for example (5')AGCT(3'). Palindromic sequences are known to be important in molecular biology (for example the P in the famous CRISPR gene editing acronym stands for palindromic), and palindromy will play a significant role in the sequence clustering results presented here.

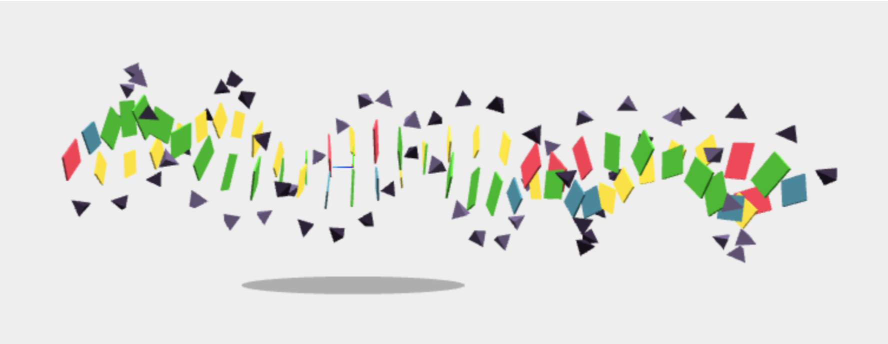


Figure 1: An example of a DNA average or ground state shape reconstruction obtained from cgDNAweb https://cgdnaweb.epfl.ch/ which is the online implementation of the cgDNA+ model that is the starting point for the work presented in this thesis.

The starting point for this thesis is the cgDNA+ model described in detail in [2], which is itself an evolution from the prior cgDNA model [3]. The cgDNA+ model is state of the art within the class of models which predict for a double helical fragment of DNA of given sequence S an equilibrium distribution, or probability density function (pdf), expressed in a set of coordinates describing the configuration of the DNA at a certain level of resolution. Different levels of resolution correspond to the modeling decision of which level of coarse-graining to adopt, and the specific choice of a set of model coordinates w. Such models take the form of a mathematical formula for the equilibrium statistical mechanical distribution ρ expressed as function of the chosen model coordinates w

$$\rho(w;\mathrm{S}) = \frac{1}{Z(\mathrm{S})} e^{-U(w;\mathrm{S})},$$

with in the Gaussian or multivariate normal approximation

$$U(w;\mathrm{S}) = \frac{1}{2}(w-\mu(\mathrm{S}))\cdot \mathbf{K}(\mathrm{S})(w-\mu(\mathrm{S}))$$

Here the minimal energy configuration, or *groundstate* vector $\mu(\mathrm{S})$ is regarded as a prediction of average or expected shape, while rigidity is encoded in the inverse covariance or precision or, for us , *stiffness matrix* $\mathbf{K}(\mathrm{S})$. The constant, or partition function, $Z(\mathrm{S})$ simply normalises the distribution. The existence of such an equilibrium distribution pdf is not guaranteed in any mathematical sense, but it is expected to arise on physical grounds due to a fluctuation-dissipation hypothesis applied to the interaction between the DNA itself and a surrounding solvent heatbath. The accuracy of the Gaussian approximation to the equilibrium distribution is also open to question. One motivation for making the Gaussian approximation is that it is typically very difficult to check model predictions of statistics against any observed experimental data statistics beyond second order, so that a Gaussian distribution is automatically the maximum entropy approximation to the available data.

From this point of view, many early coarse grain models of DNA assumed that each base pair was a single rigid body, or equivalently that the degrees of freedom between the two bases in a base pair were only treated implicitly, with the focus being on understanding a distribution of inter base pair parameters (or coordinates). In this case, for a fragment of n base pairs the configuration coordinate $w \in \mathbb{R}^{6(n-1)}$. Moreover, two distinct nearest-neighbour locality assumptions were typically made at this level of coarse graining, namely that the statistics of each subset of 6 inter base pair coordinates, or junction variables, in w, were independent one from another, and that the statistics of a given junction, depended only on the sequence composition of the two flanking base pairs. In the Gaussian approximation these assumptions translate into the statements that the stiffness matrix $\mathbf{K}(\mathrm{S})$ is 6×6 block diagonal, with each block depending only on the local sequence (and indeed many authors further assumed that the 6×6 blocks were themselves diagonal, which is now known to be a very poor approximation). Similarly the subvectors of the groundstate vector $\mu(\mathrm{S})$ have only a local, dimer, dependence on sequence.

The cgDNA model introduced an explicit treatment of flexibility between the two bases in a base pair by including an explicit treatment of *intra* base pair parameters, and allowing coupling between *inter* and *intra* coordinates, but still with the degrees of freedom to the phosphate group only being treated implicitly within each nucleotide. At this level of coarse graining, we have that $w \in \mathbb{R}^{(12n-6)}$ for an n base pair fragment. One of the unique features of the $cgDNA$ Gaussian model is that it was shown that assuming a physical nearest neighbour-interaction between bases with a local sequence dependence corresponds to assuming that the free energy U is a sum of localised quadratic energies along the DNA chain. However, because of the double chain topology, this leads to a banded (not block diagonal) stiffness matrix $\mathbf{K}(\mathrm{S})$ with 18×18 blocks with 6×6 overlaps, but with each 18×18 block only having localised (dimer) sequence dependence. For such banded stiffness matrices the associated covariance matrix is dense, with nonlocal sequence dependence of its entries throughout. And similarly it was shown that the groundstate vector $\mu(\mathrm{S})$ also has (often quite strong) nonlocal sequence dependence. The parameter sets for the cgDNA model were extracted by fitting model predicted statistics for a small set of training oligomers (of moderate length) to comparable statistics drawn directly from long duration, fully atomistic molecular dynamics (or MD) simulations of the same fragments. The model was verified by testing the locality assumptions made during the parameter estimation stage to statistics drawn from both training set and independent, test MD simulations, with quite good fits. (The accuracy between MD simulations and physical reality is another, distinct issue.)

Finally the cgDNA+ model considered here, added an explicit treatment of the degrees of freedom between the phosphate group and base in each nucleotide. There is no new mathematical structure introduced in passing from the cgDNA to cgDNA+ models, and all the remarks concerning nonlocality of covariace and groundstate carry over. In cgDNA+ the configuration coordinate is now $w \in \mathbb{R}^{(24n-18)}$ for a n base pair fragment, and the stiffness matrix $\mathbf{K}(\mathrm{S})$ now is banded with 42×42 blocks with 18×18 overlaps. The main difference in passing from cgDNA and cgDNA+ is the observation that the fit to statistics taken from MD simulation is now remarkably good [2]. The doubling of the dimension of the degrees of freedom in the Gaussian leads to an order of magnitude decrease in fitting error (measured in Kullback-Leibler divergence per degree of freedom). With its particularly high level of detail, a price to pay for the cgDNA+ model accuracy is the relatively high numbers of parameters involved, as well as the high dimensionality of the predicted Gaussian. For example, for a sequence of only 11bp in the length $w \in \Re^{246}$, and the coordinate groundstate vector and (symmetric) stiffness matrix represent together a total of $30,627$ independent numbers. In this sense the cgDNA+ model has the flavour of a machine learning model, where no one model parameter is of great interest, rather it is understanding ensembles of model predictions that is the main goal.

The goals of this thesis are to introduce and exploit mathematical frameworks and tools to tackle the challenges of understanding the large ensembles of high-dimensional Gaussians generated by the cgDNA+ model. First, based on a geometrical approach, we discuss a variety of ways to describe a space of multivariate Gaussian distributions. Second, we use appropriate unsupervised statistical learning techniques for dimensionality reduction on data sets generated by the cgDNA+ model. This provides general mappings of DNA mechanical properties, that allow for visualisation, and clustering. Finally, We construct a generic outlier detection method and show how to apply it to scan genomes in the search for sequences with exceptional mechanical properties.

The thesis is structured in two parts: Part I (Chapter 1-3) is dedicated to a summary of the necessary theoretical background that will be used, namely the cgDNA+ model, the basic theory of Gaussian distributions, and presentation of some statistical learning techniques, all of which are used in Part II. Most of the content of Part I can be regarded as standard, or at least previously known, results.

Chapter 1 gives a basic introduction to the cgDNA+ model.

Chapter 2 starts with a reminder about multivariate Gaussian distributions and their basic properties, then goes on with exposing different possible choices of divergences, distances and metrics to compare Gaussian distributions, with an emphasis on scale invariance. It then ends with a discussion on the notion of average of an ensemble of Gaussians.

Chapter 3 is dedicated to theoretical background around dimensionality reduction techniques. It presents the various methods used throughout this work, and in particular for the analysis of the cgDNAloc data sets studied in Part II. We then illustrate how these methods can be applied to ensembles of Gaussian distributions. Close attention is paid to the invariance of the projection to lower dimension by linear change of coordinates on the space on which the Gaussian distributions in the ensemble are defined. In particular, a simple but appropriate ”stiffness-weighted” *metric* PCA is introduced.

In Part II, we start (Chapter 4) with an original proof of a simple result concerning marginalisation of Gaussians with banded stiffness matrices which yields the computationally efficient marginalisation tool cgDNAloc. We then (Chapter 5-6) proceed to the applications of the different methods and tools of Part I and discuss the results.

In Chapter 4, we extend a previous result by Glowacki [4] on the maximum entropy fit of an inverse covariance matrix with a prescribed sparsity pattern consisting of overlapping diagonal square subblocks, to show that (square) marginals of such matrices inherit the same banded structure. Moreover, we provide an algorithm to compute these marginal inverse covariances explicitly in a direct and highly efficient way. Although the result stands in more generality, the given algorithm is motivated by and is particularly useful for applications of the cgDNA+ model to short sequences (sites, or *loci*) embedded in large DNA fragments - typically, genomic material). For it allows us to address the crucial non-locality of the cgDNA+ predictions, and thus can capture the effects of nearby base pair content in the flanking regions of a site. We call cgDNAloc the marginalisation algorithm, as well as the resulting Gaussian marginal probability density.

Chapter 5 deploys the tools of Chapter 3 and 4 in an attempt to map, and classify, the cgDNA+ predictions for exhaustive lists of short sequences (up to 10 bp). This is done through unsupervised dimensionality reduction methods, starting with simple Principal Component Analysis. In particular, we show how the *metric* PCA introduced in Chapter 3 leads, for any sequence length N, to a clear clustering of the 4^N possible sequences after projection. This clustering is entirely determined by the purine/pyrimidine content of the sequences, with the number of clusters consistently equal to 2^N.

In the same spirit as Chapter 5, Chapter 6 is devoted to tackle the following question: can we characterise sub-sequences with “exceptional” mechanical properties? In the context of the cgDNA+ model, we develop and apply an outlier detection procedure to detect sequences whose cgDNA+ distribution is either particularly far from, or close to an *average* distribution - as described in Chapter 2. Together with the cgDNAloc tool, this procedure is then applied on exhaustive ensembles of sequences as in Chapter 5. Then, it is also used to scan the chromosomes of *S. Cerevisiae*, for sub-sequences of length from dozens to 100 bps. In both cases, we find that sequences that can be considered as exceptionally far from average, are sequences with high A/T content, and more specifically even sequences with high AA/ TT dimer step content.

The thesis closes with a Conclusion section including a discussion of possible directions for future work. In particular we present some preliminary results for the cgDNA+ model, indicating that when the standard sequence alphabet is extended to include some epigenetically modified bases, then some new and strikingly strong variations in the ensembles of equilibrium distributions can be observed.

PART I: Background

Inside eukaryotic cells, DNA is tightly packed in the nucleus and for most of the cell cycle, chromosomes form a fuzzy structure called *chromatin*. In humans, each single cell contains the equivalent of about 2 meters of DNA. In such a dense environment, the accessibility of genomic regions to the various factors interacting with DNA in the nucleus plays a crucial role to gene regulation and other nucleic processes.

In terms of scales, the range between single base pair and chromosomes is filled with several intermediate structures, whose description and understanding is still an active field of research. Worth to mention is the fundamental subunit of chromatin, the *nucleosome*. It consists of 147 bp of DNA wrapped around a core of eight *histone* proteins. Each human cell contains about 30 million nucleosomes [5], thus covering major part of the genome.

Nucleosomes, as well as higher order structures in the chromatin, strongly rely on the local chemical, but also mechanical properties of the DNA molecule of length of a few tens of base pairs. Indeed beyond the paradigmatic straight double helix model of Watson and Crick, it is now known that DNA fragments come in a wide variety of conformations. More precisely, one distinguishes between two phenomenon: first, depending on the underlying basepair sequence, some fragments are not straight, but can exhibit strong intrinsic bending [6]. This is referred to as DNA *shape*. Second, placed in a thermal bath (as it is inside the nucleus), fluctuations affect the DNA conformation in time as a function of its local *flexibility* - or, equivalently, its local *rigidity*, or *stiffness*. It should be emphasised that DNA shape and stiffness do not have a universal definition, and their precise meaning can vary by context.

When it comes to physical models of the DNA shape and flexibility, it is all about scale. The most detailed models are provided by Molecular Dynamics (MD). These full atomistic *in silico* simulations allow to reproduce the thermodynamics of a molecule in solution for periods of time corresponding to (up to) tens of microseconds. Despite their impressive descriptive power, the computational cost of these models and their certain lack of flexibility (one DNA fragment per simulation) does not make them directly applicable to genomic contexts, for example. At the other side of the scale spectrum, rod-like models can approximate mechanical features of DNA pieces at the scale of chromosomes, but do not generally account for sequence content.

At short length scales, popular *coarse grained* models are of two major types: the so called *rigid base pair* models [7] and the *rigid base* models. In rigid base pair models, all the atoms forming a base are approximated as a single rigid body. The conformation of an entire DNA chain is then described through the relative rigid body motion between each neighboring base pair, with 6 degrees of freedom: 3 rotational (tilt, roll, twist), and 3 translational (slide, shift, stagger). In rigid base models, 6 additional degrees of freedom are added to describe the relative displacement between the two bases inside each base pair.

In this work, we will focus on one of these coarse grain models, the cgDNA+ model. This model is based on the previous rigid base cgDNA model, which was developed by Maddocks, Gonzalez, Petkeviciute et al. [8] [3]. In addition to the base pair degrees of freedom, the cgDNA+ upgrade [2] also includes an explicit description of the positions and rotations of *phosphate groups* along the chain.

While parametrisation of coarse grain models can be done in different manners, the cgDNA model family is parametrised based purely on MD simulation data. The idea is to extract essential features - namely, shape and stiffness - of the dynamics from a ensemble of MD simulations, and then being able to predict these features for any given DNA sequence.

1 cgDNA+: a sequence dependent coarse-grained model of DNA mechanics

In this section, we briefly present the cgDNA+ model, which will be the central tool of this thesis. The original fully detailed description of it is available in [2].

The cgDNA+ model is a sequence-dependent coarse-grain model of double stranded B-form DNA with an explicit treatment of bases and phosphate groups. The goal of this model is to predict the sequence-dependent groundstate as well as the flexibility of double-stranded B-form DNA. There is a now a variety of different sets of parameters for cgDNA+ (see [9], [2], [10]). All the analysis and computations performed for this thesis use the DNA parameter set from [10], which is still currently the most recent update.

We start by shortly describing a prior version of the model, called cgDNA, that shares common essential features with cgDNA+. It is a sequence–dependent, rigid–base, coarse–grain model of B–form DNA, without explicit treatment of phosphate groups, and was introduced in [3, 11]. This model is also parametrised from full atomistic molecular dynamics (MD) simulations of a set of sequences of short length. Given a parameter set $\mathcal{P}$ and an arbitrary DNA sequence $\mathcal{S}$, cgDNA predicts a Gaussian equilibrium probability density function in configuration space (see below), by reconstructing the mean $\mu \equiv \mu(\mathcal{P}, \mathcal{S}) \in \mathbb{R}^N$, or ground state, and the precision matrix $K \equiv K(\mathcal{P}, \mathcal{S}) \in \mathbb{R}^{N \times N}$, or stiffness matrix:

$$\rho(w; \mathcal{P}, \mathcal{S}) = \frac{1}{Z} \exp\left\{-\frac{1}{2}(w-\mu) \cdot K(w-\mu)\right\}. \tag{1}$$

The coordinate vector $w \in \mathbb{R}^N$ encodes the configuration in the following way: coarse-grained at the base level, a molecule of double stranded DNA can be interpreted as a double chain of rigid bodies. More details can be found in [12].

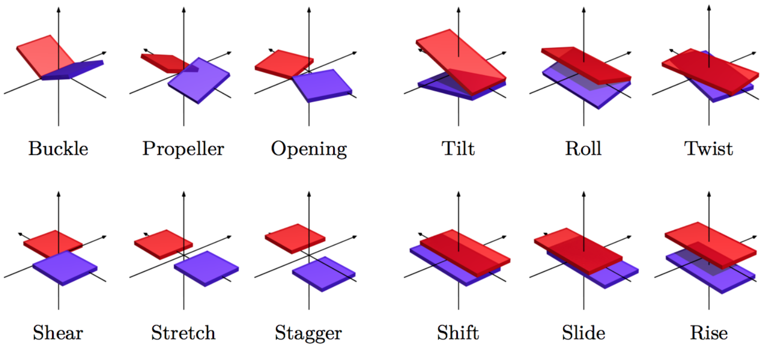


Figure 2: The twelve (6 rotational, 6 translational) standard degrees of freedom in a rigid-base model of the DNA molecule like the cgDNA, model, and the cgDNA+ model - which also includes explicit degrees of freedom for phosphate groups of the DNA backbone.

A set of internal coordinates for the double chain configuration is then introduced, which is divided into *inter* coordinates describing the 3 translational and 3 rotational degrees of freedom between consecutive base pairs, and *intra* coordinates describing these degrees of freedom between the two bases inside a base pair. Those take standard names: Tilt, Roll, Twist, etc. (see Figure 2). For a DNA fragment of N bp, there are a total of $6(N-1)$ inter coordinates plus $6N$ coordinates, thus $w \in \mathbb{R}^{12N-6}$. The Curves+ software [13] is used in the fitting procedures. When considering a double representation of a n base–pair long DNA sequence $\mathcal{S}$, the internal coordinates $w(\mathcal{S}) \in \mathbb{R}^{12n-6}$ satisfy the following physical property, which reflects Crick-Watson symmetry,

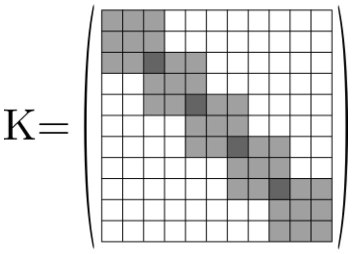
$$w(\mathcal{S}) = E_{2n-1} w(\overline{\mathcal{S}}), \tag{2}$$

where $\overline{\mathcal{S}}$ is the complementary sequence to $\mathcal{S}$ and $E_{2n-1} \in R^{(12n-6)\times(12n-6)}$ is a block, trailing diagonal matrix composed by $2n-1$ copies of $E = \text{diag}(-1,1,1,-1,1,1) \in \mathbb{R}^{6\times 6}$

$$E_{2n-1} = \begin{bmatrix} & & & E \\ & & E & \\ & \iddots & & \\ E & & & \end{bmatrix}, \tag{3}$$

where $E_{2n-1} = E_{2n-1}^T = E_{2n-1}^{-1}$.

Another assumption concerns the stiffness matrix $K(\mathcal{S}) \in \mathbb{R}^{(12n-6)\times(12n-6)}$: based on observation of statistics estimated from MD trajectories, we assume that the stiffness matrix K is banded, i.e is a sparse matrix in which non–zeros entries are all close to a diagonal band. More precisely the sparsity pattern of K is 18×18 block diagonal with 6×6 overlaps:



Now we turn our attention to the updated cgDNA+ model, which is described in [2]. Similarly to cgDNA, given a parameter set $\mathcal{P}$ and an arbitrary DNA sequence $\mathcal{S}$, cgDNA+ predicts a Gaussian equilibrium probability density function in the configuration space by reconstructing the ground state $\mu \equiv \mu(\mathcal{P},\mathcal{S}) \in \mathbb{R}^M$ and the stiffness matrix $K \equiv K(\mathcal{P},\mathcal{S}) \in \mathbb{R}^{M\times M}$:

$$\rho(w;\mathcal{S},\mathcal{P}) = \frac{1}{Z}\exp\left\{-\frac{1}{2}(w-\mu)\cdot K(w-\mu)\right\}, \tag{4}$$

where this time $M = 24N - 18$, with N the length in basepairs of the sequence $\mathcal{S}$. Any configuration $w \in \mathbb{R}^{24N-18}$ can be divided into $N-1$ sets of 6 inter–base–pair internal coordinates, N sets of 6 intra–base–pair internal coordinates, and $2(N-1)$ sets of 6 base–to–phosphate internal coordinates. As a single internal is represented by a six dimensional vector the total number of components in the configuration w are $24N - 18$. The inter– and intra–base–pairs coordinates havealready been introduced in chapter 2 of [2] while the base–to–phosphate internal coordinates have been introduced in chapter 8 of [2] and are further discussed in the next section.

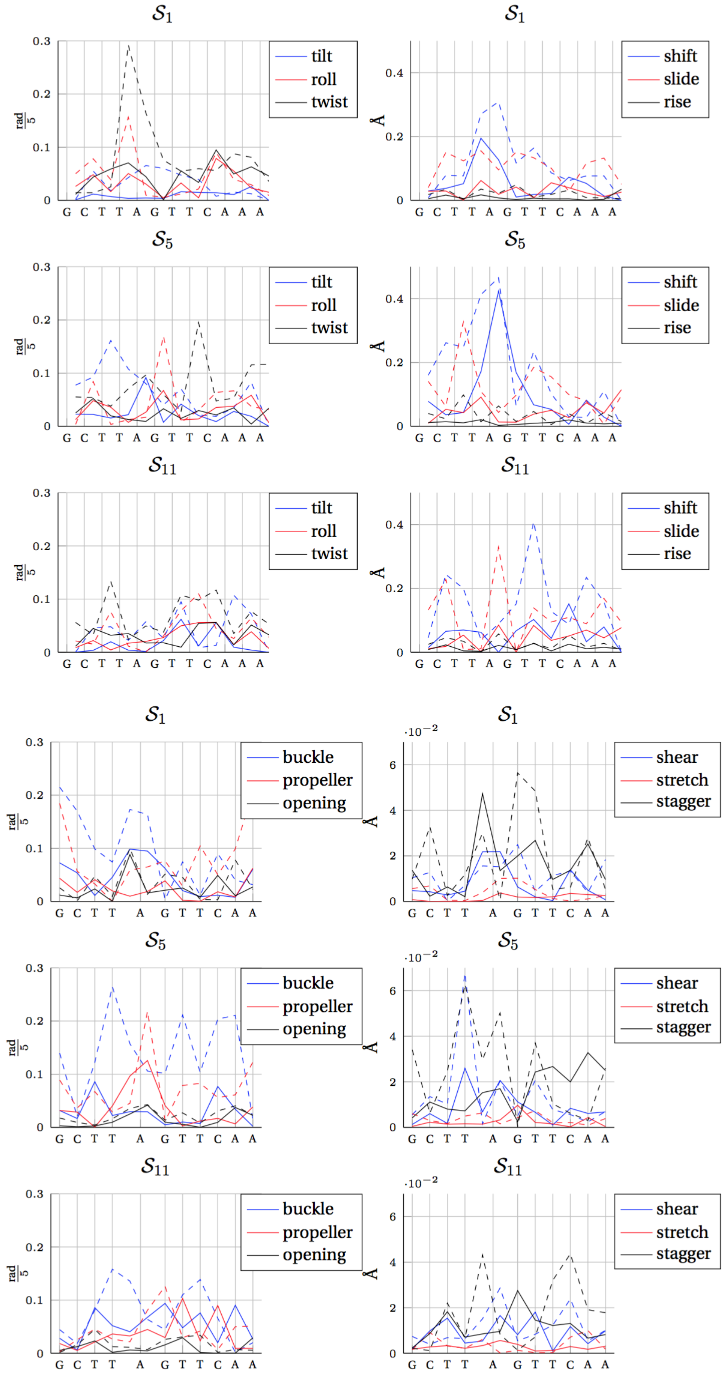


Figure 3: (Courtesy of Alessandro Patelli's thesis [2]) Absolute error between model predicted interbasepair (top) and intrabasepair (bottom) degrees of freedom and MD observation. In solid, the error obtained by the cgDNA+ model and in dashed the error obtained by the cgDNA model. The sequences considered are part of the training dataset.

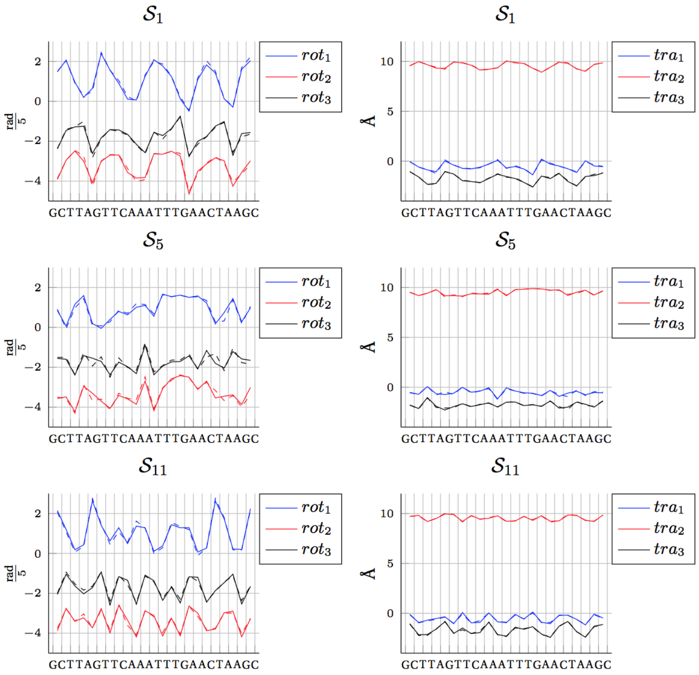


Figure 4: (Courtesy of Alessandro Patelli's thesis [2]) Comparison of base-to-phosphate degrees of freedom, on the reading strand, between cgDNA+ predictions (solid line) and MD observation (dashed line) for the sequences (1,5,11) of the Palindromic Library. In the first column we show the rotational coordinates, while in the second the translations.

The parameter set format for the cgDNA+ model is a natural extension of the cgDNA one, but with the particular property that the end sigma vector and stiffness matrices will be of different dimension than the interior ones. In detail the cgDNA+ parameter set is defined by

$$\mathcal{P} = \left\{ \sigma^{5'\alpha\beta}, \sigma^{\alpha\beta}, K^{5'\alpha\beta}, K^{\alpha\beta} \right\}_{\alpha\beta\in D', 5'\alpha\beta\in D} \subset \mathbb{P}_{\text{tot}}, \tag{5}$$

where $\mathbb{P}_{\text{tot}} = [\mathbb{R}^{36}]^{16} \times [\mathbb{R}^{42}]^{10} \times [\mathbb{S}^{36}]^{16} \times [\mathbb{S}^{42}]^{10}$, and $\mathbb{S}^N$ is the set of $N \times N$ symmetric matrices. The end sigma vectors are of dimension 36 while the interior ones are of dimension 42. Equivalently, the stiffness end blocks are of dimension 36×36 while the interior ones are of dimension 42×42. The difference in dimension between interior and end blocks is due to the fact that in the MD simulations the first phosphate group on both strands is absent. Consequently, the first and last base–pair levels are composed of only an intra–base–pair degree of freedom and a single base–to–phosphate set of internal coordinates.

Let $\mathcal{P}$ be a cgDNA+ parameter set of the form (5) and let $\mathcal{S}$ be a N base–pair long DNA sequence. We can define the reconstruction rule for the stiffness matrix $K(\mathcal{P}, \mathcal{S})$ and the weighted shape vector $\sigma(\mathcal{P},\mathcal{S})$ in the following way:

$$K(\mathcal{P},\mathcal{S}) = P_d^T K_d P_d, \tag{6}$$

$$\sigma(\mathcal{P},\mathcal{S}) = P_d^T \sigma_d, \tag{7}$$

$$\mu(\mathcal{P},\mathcal{S}) = K(\mathcal{P},\mathcal{S})^{-1}\sigma(\mathcal{P},\mathcal{S}), \tag{8}$$

where

$$\begin{aligned} K_d &= \text{diag}(K^{5'X_1X_2},\ldots,K^{X_iX_{i+1}},\ldots,K^{X_{n-1}X_n3'}),\\ \sigma_d &= (\sigma^{5'X_1X_2},\ldots,\sigma^{X_iX_{i+1}},\ldots,\sigma^{X_{n-1}X_n3'}),\end{aligned}$$

and the matrix $P_d \in \mathbb{R}^{42N-12\times 24N-18}$ reads

$$P_d = \begin{bmatrix} I_{18} & & & & \cdots & \\ & I_{18} & & & & \\ & I_{18} & & & & \\ & & I_6 & & & \\ & & & I_{18} & & \\ & & & I_{18} & & \\ & & & & I_6 & \\ \vdots & & & & \ddots & \vdots \\ & & & & & I_{18} \end{bmatrix}, \tag{9}$$

where we use the notation I_n for the n–dimensional identity matrix. We recall that the $3'$ end blocks for both stiffness and weighted shape can be computed using Crick–Watson symmetry. More precisely,

$$K^{\overline{\alpha\beta}3'} = E^{5'} K^{5'\alpha\beta} E^{5'}, \tag{10}$$

where $E^{5'}$ is defined by

$$E^{5'} = \begin{bmatrix} & & & & & E \\ & & & & I_6 & \\ & & & E & & \\ & & I_6 & & & \\ & E & & & & \\ I_6 & & & & & \end{bmatrix}, \tag{11}$$

with $E = \text{diag}(-1,1,1,-1,1,1) \in \mathbb{R}^6$, and $E^{5'} \in \mathbb{R}^{36\times 36}$ satisfy $[E^{5'}]^{-1} = [E^{5'}]^T =$

$E^{3'} \in \mathbb{R}^{36\times 36}$ with

$$E^{3'} = \begin{bmatrix} & & & & & I_6 \\ & & & & E & \\ & & & I_6 & & \\ & & E & & & \\ & I_6 & & & & \\ E & & & & & \end{bmatrix}. \tag{12}$$

For the interior blocks the Crick–Watson symmetry rule is given by

$$K^{\overline{\alpha\beta}} = E^{\text{int}} K^{\alpha\beta} E^{\text{int}}, \tag{13}$$

with

$$E^{\text{int}} = \begin{bmatrix} & & & & & & I_6 \\ & & & & & E & \\ & & & & I_6 & & \\ & & & E & & & \\ & & I_6 & & & & \\ & E & & & & & \\ I_6 & & & & & & \end{bmatrix}, \tag{14}$$

which satisfy $E^{\text{int}} = [E^{\text{int}}]^T = [E^{\text{int}}]^{-1}$. Finally we have that the complementary sequence $\overline{\mathcal{S}}$ of $\mathcal{S}$ must satisfy

$$\begin{aligned} \mu(\mathcal{P},\mathcal{S}) &= E_N^+ w(\mathcal{P},\overline{\mathcal{S}}), \\ K(\mathcal{P},\mathcal{S}) &= E_N^+ K(\mathcal{P},\overline{\mathcal{S}}) E_N^+, \end{aligned}$$

where

$$E_N^+ = \begin{bmatrix} & & & E^{5'} \\ & & E^{\text{int}} & \\ & E^{\text{int}} & & \\ \iddots & & & \\ E^{3'} & & & \end{bmatrix}. \tag{15}$$

1.1 Protein Binding Sites and DNA mechanics

We conclude this chapter with a short detour on the role of DNA mechanics in the understanding of protein binding sites. There are two reasons to do so: the first one is to provide the context around the research question that is the topic of Chapter 6, that is, can we identify specific short sites in longer sequences as “mechanical outliers” - in a sense that will be made more precise in the following sections of this chapter. The second one is that it is an opportunity to present the notion of sequence logo, which is both a common and very useful way of visualising lists of sequences in a compact way, and the most widespread basic probabilistic model for the prediction of protein binding sites in genomic data. We make extensive use of this visualisation tool in Chapter 5 and 6.

A challenging topic in the field of today's genomic studies concerns protein-DNA interactions and the understanding, finding and prediction of so called Transcription Factor Binding Sites (TFBS). Transcription factors (TF) are proteins that bind to DNA, often upstream from a gene transcription start site. TF are known to play an important role in gene regulation and expression. As opposed to the well determined way genes encode information for the formation of proteins, there is no clearly identified mechanism by which TF find their binding sites in the genome. That is, transcription factor binding site recognition, although clearly not a purely random process, does not obey simple sequence-based coding rules [14]. It has thus been an active field of research for several decades to understand the means by which these recognitions occur. Alternatively, when it comes to prediction of binding sites, statistical approaches that did not make use of a description of the binding mechanism at the physical level have demonstrated some level of efficiency. These approaches encompass a large variety of methods, ranging from basic probabilistic models to sophisticated machine learning algorithms such as neural networks [15] [16] [17] [18].

One of the earliest type of probabilistic model is based on so called position weight matrices (PWM). It was proposed by Stormo and al. [19] and is now still widely used. Briefly, a site to be modeled is seen to have a fixed length of L nucleotides. One then associates to it a matrix $W \in \mathbb{R}^{4\times L}$, that is as table of scores for each base in $\{A, C, G, T\}$ and each position of the site. For any given DNA sequence $S = \mathrm{X}_1 \cdots \mathrm{X}_L$ of length L, a total score for that sequence is computed from the matrix W as follows: first S can be represented as a $4 \times L$ matrix $\{s_{ib}\}$ of 1 and 0, where each column is filled with zero except at the row corresponding to the base X_i of S. The score of S with respect to W is then given by the matrix scalar product:

$$S \cdot W = \sum_{i=1}^{L} \sum_{b=1}^{4} w_{ib} s_{ib},$$

where $W = \{w_{ib}\}$. Usually, PWM are built using some available experimental data consisting of a collection of sequences $\{S_j\}$ of length L that have been found to be binding sites for the protein studied. One often chose to define the entries of W to be related to the frequencies f_{ib} of the base b and at position i computed from the collection $\{S_j\}$. Most commonly,

$$w_{bi} = \log_2(f_{bi}).$$

This choice allows to interpret the score of a sequence as its probability to be drawn from a product of L independent discrete random variables on $\{A, C, G, T\}$, with weights calculated from the dataset $\{S_j\}$.

Despite their simplicity, PWM models have demonstrated some good performance for many different types of TFBS prediction problems [20] [21]. However, their scope is fundamentally limited by the assumption of independence between individual bases. This caveat have already been addressed by extending to 2, 3 or even k-mer models, which could capture these non local features, but at the cost of increasing model complexity [22]. Furthermore, even these more elaborate models do not account explicitly for the chemical or physical processes at stake in each specific protein-DNA binding.

A useful aspect of the PWM approach is that it comes with a simple way to visualise sequences associated to a site. This visual representation is commonly referred to as a *sequence logo* (Fig 5). Sequences logos can just show base frequencies (frequency logos) at each site position, but more often these frequencies are weighted by *information content* (information content logos), defined as

$$\mathrm{IC}(i) = \sum_{b=1}^{4} f_{ib} \log_2\left(\frac{f_{ib}}{0.25}\right). \tag{16}$$

Sequence logos are very useful for representing collections of sequences in a compact way. As a drawback, it only presents base pair content, all correlation between frequencies of neighbouring base pairs are lost. As an attempt to overcome this issue, one can try to represent frequencies of (e.g.) dimers. In order to do so, we have built simple plots that will be referred to as *dimer logos*. See Figure 6 for an example of this type of logo.

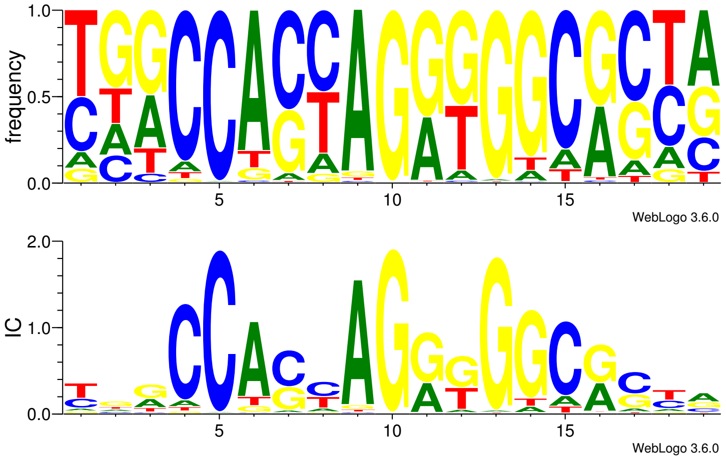


Figure 5: Sequence logos of CTCF zinc finger protein binding sites [23], showing *top*: base frequencies at each site position, *bottom*: the same frequencies weighted by information content IC at each position, so that zero amplitude implies equal frequencies.

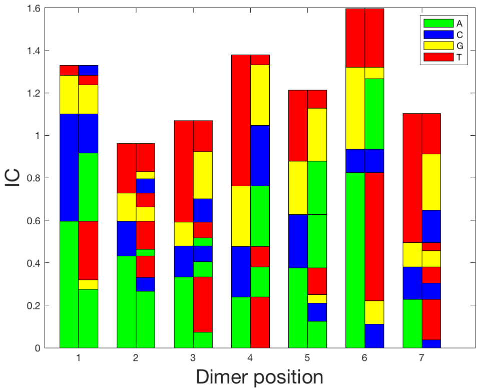


Figure 6: An example of a dimer logo for a collection of sequences of length 8bp. Each double column represents frequencies of all the successive dinucleotides X_iX_{i+1} at position $i = 1, \dots, 7$. The left part of the ith double column shows the frequencies of the X_i's. These frequencies are themselves split on the right column to show the frequencies of the following bp X_{i+1}. The total frequencies of say an A at base pair three appear in both the second column of the second step and the first column of the third step, but reordered and the information content of the two steps need not to be the same

The fact that transcription factors favor certain sequences in their DNA binding, i.e. their binding *specificity*, is understood at the physical level to be the result of two different phenomena. The first one involves binding energies associated with chemical interactions (e.g. formation of hydrogen bonds) between the protein and specific bases of the DNA chain. This type of recognition depends only on the chemical properties of each base type, and thus is referred to as *direct readout*. On the other hand, the formation of the DNA-protein complex also involves an *elastic energy* contribution, by requiring to deform both the DNA and the protein into their bound configuration. In this case, DNA sequence, having an impact on the local structure (shape and flexibility) of the double helix, can influence the specificity of the binding *even at base positions where no binding occurs* [24]. For this reason, that sort of recognition mechanism is called *indirect readout*.

The respective contribution of direct and indirect readout in TF binding has been examined, with varying outcomes depending on the protein class studied [25] [26] [27]. Nevertheless, it is now accepted that sequence-dependent structural features of DNA can play a significative role in site recognition [28] [29] [30] [31] [32] [33].

2 Elements on Gaussian distributions

Here we gather some standard results on Gaussian distributions, divergences, distances and metrics on probability distributions, and make a useful remark on their invariance by linear change of scale.

2.1 Some Facts on Gaussians

We start by recalling some basic facts about Gaussian distributions on $\mathbb{R}^n$.

Definition 1. *Let $n \geq 1$ be an integer. A continuous random variable $\mathbf{x}$ on $\mathbb{R}^n$ is said to be* normal*, or* Gaussian*, of dimension n if its probability density function* (pdf) $\rho : \mathbb{R}^n \to [0, 1]$ *takes the form*

$$\rho(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{Z} e^{-(\mathbf{x}-\mu)\cdot\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}.$$

The parameters are the mean *vector $\mu \in \mathbb{R}^n$, and the* covariance *matrix $\Sigma \in \mathbb{R}^{n\times n}$ which is symmetric and positive definite. The normalisation constant*

$$Z = Z(\Sigma) = \int_{\mathbb{R}^n} e^{-(\mathbf{x}-\mu)\cdot\mathbf{\Sigma}(\mathbf{x}-\mu)} d\mathbf{x} = \det(2\pi\Sigma)^{1/2}$$

ensures that

$$\int_{\mathbb{R}^n} \rho(\mathbf{x}; \mu, \mathbf{\Sigma}) d\mathbf{x} = 1.$$

Remark 1. *In this work, we will use the words (probability) density, or density function, or pdf, or (probability) distribution, interchangeably. In particular, we will never use the word distribution to refer to the general theory of distributions as linear operators. The measured spaces will always be finite dimensional (usually $\mathbb{R}^n$).*

Remark 2. *(notation) A general probability distribution will usually be denoted by the letter ρ. When referring to a Gaussian distribution, we will usually write $\rho(\mathbf{x}; \mu, \Sigma)$, or simply $\rho(\cdot; \mu, \Sigma)$.*

Remark 3. *(1st and 2nd moment of a Gaussian)*

The parameters (μ, Σ) of a Gaussian distribution arise as its first and second moment:

$$\mu = \int_{\mathbb{R}^n} \mathbf{x}\, \rho(\mathbf{x}; \mu, \Sigma) d\mathbf{x},$$

$$\Sigma = \int_{\mathbb{R}^n} (\mathbf{x} - \mu)(\mathbf{x} - \mu)^T \rho(\mathbf{x}; \mu, \Sigma) d\mathbf{x}.$$

Definition 2. *Let us denote the set of Gaussian distributions of dimension* n *by*

$$\mathcal{G} = \{\rho(\cdot; \mu, \mathbf{\Sigma}) : \mu \in \mathbb{R}^n, \Sigma \in \mathsf{S}_n^+\},$$

where S_n^+ *denotes the set of symmetric, positive-definite matrix of size* n*.*

Remark 4. *The canonical mapping* $\mathbb{R}^n \times \mathsf{S}_n^+ \to \mathcal{G}$ *endows* $\mathcal{G}$ *with a differential structure, turning* $\mathcal{G}$ *into a smooth manifold:*

$$\mathcal{G} \cong \mathbb{R}^n \times \mathsf{S}_n^+.$$

This implies that $\mathcal{G}$ *is an open, convex (thus connected) manifold of dimension* $n + \frac{n(n+1)}{2}$*.*

Of particular importance to us is the inverse of the covariance matrix.

Definition 3. *The* inverse covariance *matrix, or* precision *matrix, or* stiffness *matrix, of a covariance matrix* $\mathbf{\Sigma}$ *is simply its matrix inverse:*

$$\mathbf{K} = \mathbf{\Sigma}^{-1}.$$

Any set of smooth coordinates $\in \mathbb{R}^{n+\frac{n(n+1)}{2}}$ on $\mathbb{R}^n \times \mathsf{S}_n^+$ provides a set of smooth coordinates on $\mathcal{G}$. By abuse of notation, we will often use $\boldsymbol{\theta} = (\mu, \mathbf{\Sigma})$ as coordinates, where it is implicit that only the lower (or upper) triangular entries of $\mathbf{\Sigma}$ are needed. Importantly, we will also sometimes use $\boldsymbol{\theta} = (\mu, \mathbf{K})$ to parametrise $\mathcal{G}$.

Definition 4. *Let* $\rho(\cdot; \mu, \mathbf{\Sigma})$ *Gaussian pdf on* $\mathbb{R}^n$*. Its entropy* S *is*

$$S(\rho) = -\int_{\mathbb{R}^n} \rho(w) \ln(\rho(w)) dw = \frac{1}{2} \ln\left((2\pi e)^n \det(\mathbf{\Sigma})\right)$$

For a n-dimensional normally distributed variable $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$, with mean parameter μ and covariance matrix Σ, it is a standard and well-known result that the marginal distribution of a subset $(x_{i_1}, \ldots x_{i_k})$ of coordinates of $\mathbf{x}$ is itself normal, with mean parameter vector

$$\widetilde{\mu} = \begin{bmatrix} \mu_{i_1} \\ \vdots \\ \mu_{i_k} \end{bmatrix}$$

and covariance matrix $\widetilde{\boldsymbol{\Sigma}} = (\sigma_{pq})_{p,q=1}^{k}$, with $\sigma_{pq} = \Sigma_{i_p i_q}$, $p, q = 1, \dots, k$.

2.2 Divergences, distances and metrics on Multivariate Gaussians

There are many ways that we can compare probability distributions in general, and Gaussian distributions in particular. Here we give a non-exhaustive overview of different metrics (in a generic sense) that can be used to express some notion of proximity between pdfs. Some tools that are central to information theory, such as the Kullback-Leibler divergence, and its connection to the Fisher information, are introduced. The latter plays a historically fundamental role in the field of *information geometry*, which regards it as a Riemmanian metric on the manifold of parametric probability densities.

In the wide landscape of different tools available to compare objects, points in space, or in our case distributions, a geometry-oriented mind will have a preference for *distances*. As the meaning of the word can greatly vary depending on the context, let us clarify here that throughout this work, we will use it only in the precise sense of metric spaces. The word *metric*, which is also used to designate a large variety of different notions, will for us specifically refer to Riemannian metrics; that is, a smoothly varying inner product defined on a smooth manifold.

For statisticians, who often do not adopt a geometrical view on probability densities, requiring a proximity measure to satisfy triangle inequality, or even symmetry, is in many cases too restrictive. They will rather use the weaker notion of *divergence* (see below). As a matter of fact, one of the central tool in statistics and information theory is the famous divergence called *Kullback-Leibler*, whose definition we give now.

Definition 5. *Let* ρ_1, ρ_2 *be any probability distribution on* $\mathbb{R}^n$ *with respect to the standard (flat) measure* dw, $n \geq 1$. *The* Kullback-Leibler divergence, *or* relative entropy *between* ρ_1 *and* ρ_2 *is*

$$\mathrm{KL}(\rho_1, \rho_2) = \int_{\mathbb{R}^m} \rho_1(w) \ln\left(\frac{\rho_1(w)}{\rho_2(w)}\right) dw. \tag{17}$$

There are of course technical conditions to ensure that KL is well-defined, but since we will use it exclusively for Gaussian distributions (for which (17) is explicit provided that), we will not discuss them. We limit ourselves to describing some basic but useful properties.

Properties 1.

1. $\mathrm{KL}(\rho_1, \rho_2) \geq 0$*;*
2. $\mathrm{KL}(\rho_1, \rho_2) = 0 \Leftrightarrow \rho_1 = \rho_2$*;*
3. *In general,* $\mathrm{KL}(\rho_1, \rho_2) \neq \mathrm{KL}(\rho_2, \rho_1)$*;*
4. *In general,* KL *does not satisfy the triangle inequality (therefore, it is not a distance).*
5. (Additivity for independent variables) *Suppose* ρ_1 *and* ρ_2 *split as* $\rho_1(x_1, \dots, x_n) = \rho_1^1(x_1) \dots \rho_1^n(x_n)$ *and similarly* $\rho_2(x_1, \dots, x_n) = \rho_2^1(x_1) \dots \rho_2^n(x_n)$*. Then*

$$\mathrm{KL}(\rho_1, \rho_2) = \sum_{i=1}^{n} \mathrm{KL}(\rho_1^i, \rho_2^i).$$

The first two properties above characterise a statistical *divergence*. The fact that KL lacks symmetry in general motivates the use of its symmetrised version (in fact, the original definition [34] by Kullback and Leibler has this form)

$$\mathrm{KL}^{sym}(\rho_1, \rho_2) := \frac{1}{2}(\mathrm{KL}(\rho_1, \rho_2) + \mathrm{KL}(\rho_2, \rho_1)).$$

For multivariate Gaussians $\rho_i = \rho(\cdot; \mu_i, \mathbf{K}_i^{-1})$, $i = 1, 2$, the Kullback-Leibler divergence can be computed explicitly:

$$\mathrm{KL}(\rho_1, \rho_2) = \frac{1}{2}\Big[(\mu_1 - \mu_2) \cdot \mathbf{K}_2(\mu_1 - \mu_2) + \mathrm{tr}(\mathbf{K_2}\mathbf{K_1}^{-1}) + \ln(\det(\mathbf{K}_2^{-1}\mathbf{K_1}) - n\Big]. \tag{18}$$

The first term on the left-hand side of (18) is called the *squared Mahalanobis distance*:

$$\mathrm{MH}(\mu_1, \mathbf{K}_1, \mu_2, \mathbf{K}_2) = \frac{1}{2}(\mu_1 - \mu_2) \cdot \mathbf{K}_2(\mu_1 - \mu_2)$$

Although MH has the structure of a weighted inner product between the mean vectors μ_1 and μ_2, it should be noted that MH is *not* a squared distance between the pairs $(\mu_1, \mathbf{K}_1)$ and $(\mu_2, \mathbf{K}_2)$. Nevertheless, it is still a useful quantity to compare the mean vectors of two Gaussians, as it enjoys the same properties listed in Properties 1 as KL. A symmetrised version

$$\mathrm{MH}^{sym}(\mu_1, \mathbf{K}_1, \mu_2, \mathbf{K}_2) = \frac{1}{4}(\mu_1 - \mu_2) \cdot (\mathbf{K}_1 + \mathbf{K}_2)(\mu_1 - \mu_2)$$

is also often used.

Closely connected to the Kullback-Leibler divergence is the *Fisher information.* For a one dimensional parametric pdf $p(x;\theta)$, it is defined as

$$I^{\mathrm{Fisher}}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log(p(x;\theta))|\theta\right] \tag{19}$$

In the case where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ is multidimensional, then the Fisher information becomes a matrix, whose entries read

$$\mathbf{I}^{\mathrm{Fisher}}_{ij}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\theta_j}\log(p(x;\boldsymbol{\theta}))|\boldsymbol{\theta}\right] \tag{20}$$

for $i, j = 1, \ldots, n$.

The matrix $\mathbf{I}^{\mathrm{Fisher}}(\boldsymbol{\theta})$ endows the manifold $\mathcal{M}$ of the parametric family $\{\rho(x,\theta)\}$ with a Riemmanian metric, turning $\mathcal{M}$ into a *statistical manifold.*

The connection between the Fisher information matrix and the Kullback-Leibler arises when expanding the latter with respect the parameter in the second argument:

$$\mathrm{KL}(\rho(x;\boldsymbol{\theta}),\rho(x,\boldsymbol{\theta}')) = \mathrm{KL}(\rho(x;\boldsymbol{\theta}),\rho(x;\boldsymbol{\theta}')) + \frac{\partial}{\partial\boldsymbol{\theta}'}\mathrm{KL}(\rho(x;\boldsymbol{\theta}),\rho(x;\boldsymbol{\theta}'))|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}(\boldsymbol{\theta}'-\boldsymbol{\theta}) + \frac{1}{2}(\boldsymbol{\theta}'-\boldsymbol{\theta})\cdot\ \mathbf{H}_{\mathrm{KL}}(\boldsymbol{\theta})(\boldsymbol{\theta}'-\boldsymbol{\theta}) + \mathcal{O}(|\boldsymbol{\theta}'-\boldsymbol{\theta}|^3), \tag{21}$$

where we used $\frac{\partial}{\partial\boldsymbol{\theta}'}$ to denote the gradient operator with respect to $\boldsymbol{\theta}'$, and $\mathbf{H}_{\mathrm{KL}}(\boldsymbol{\theta})$ is the Hessian matrix with entries

$$\mathbf{H}_{\mathrm{KL}}(\boldsymbol{\theta})_{ij} = \frac{\partial^2}{\partial\theta_i'\partial\theta_j'}\mathrm{KL}(\rho(x;\boldsymbol{\theta}),\rho(x;\boldsymbol{\theta}'))|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}.$$

On the other hand, a direct computations shows that

$$\mathbf{I}_{ij}^{\mathrm{Fisher}}(\boldsymbol{\theta}) = -\int_{\mathbb{R}^n}\rho(x;\boldsymbol{\theta})\frac{\partial^2}{\partial\theta_i\theta_j}\log(\rho(x;\boldsymbol{\theta})) = \frac{\partial^2}{\partial\theta_i'\partial\theta_j'}\mathrm{KL}(\rho(x;\boldsymbol{\theta}),\rho(x;\boldsymbol{\theta}'))|_{\boldsymbol{\theta}'=\boldsymbol{\theta}} = \mathbf{H}_{\mathrm{KL}}(\boldsymbol{\theta})_{ij},$$

meaning that the Fisher information matrix arises as the Hessian matrix of the Kullback-Leibler divergence. Moreover, the first two Properties 1 imply that the first two terms in (21) vanish, so that

$$\mathrm{KL}(\rho(x;\boldsymbol{\theta}),\rho(x,\boldsymbol{\theta}')) = \frac{1}{2}(\boldsymbol{\theta}'-\boldsymbol{\theta})\cdot\ \mathbf{I}^{\mathrm{Fisher}}(\boldsymbol{\theta})(\boldsymbol{\theta}'-\boldsymbol{\theta}) + \mathcal{O}(|\boldsymbol{\theta}'-\boldsymbol{\theta}|^3).$$

In other words, the Fisher information matrix is a second-order approximation of the KL divergence. Remarkably, it can also be shown that the same approximation holds when expanding KL with respect to the parameter of the probability density in the first argument of the divergence.

As $\mathbf{I}^{\mathrm{Fisher}}(\boldsymbol{\theta})$ is symmetric and positive definite, it defines a *Riemannian metric* on the manifold $\mathcal{M}$, turning it into a Riemmanian manifold. In turn, it induces a distance function $d_F(\rho(x;\boldsymbol{\theta}_1),\rho(x;\boldsymbol{\theta}_2))$ on $\mathcal{M}$ via the standard formula

$$d_F(\rho(x;\boldsymbol{\theta}_1),\rho(x;\boldsymbol{\theta}_2)) = \inf_{\boldsymbol{\theta}(t_1)=\boldsymbol{\theta}_1,\boldsymbol{\theta}(t_2)=\boldsymbol{\theta}_2}\int_{t_1}^{t_2}\sqrt{\dot{\boldsymbol{\theta}}(t)\cdot\mathbf{I}^{\mathrm{Fisher}}\ (\boldsymbol{\theta}(t))\dot{\boldsymbol{\theta}}(t)}dt$$

.

In the specific case when $\mathcal{M}=\mathcal{G}$, the manifold of Gaussian distributions $\rho(x;\mu,\mathbf{K}^{-1})$, parametrised by $\boldsymbol{\theta} = (\mu, \mathbf{K})$, the Fisher information matrix takes the form

$$\mathbf{I}^{\text{Fisher}}(\mu, \mathbf{K}) = \begin{bmatrix} \mathbf{K} & \mathbf{0}_{n,n\times n} \\ \mathbf{0}_{n\times n,n} & \frac{1}{2}\mathbf{K}^{-1} \otimes \mathbf{K}^{-1} \end{bmatrix} \tag{22}$$

where $\otimes$ denotes the tensor product, or just Kronecker product, between matrices. Concretely, for any pair $(\Delta\theta_1, \Delta\theta_2) = ((\Delta\mu_1, \Delta\mathbf{K}_1), (\Delta\mu_2, \Delta\mathbf{K}_2))$ sitting on the tangent space $\mathcal{T}_{(\mu,\mathbf{K})}\mathcal{M}$, the inner product $\langle\Delta\theta_1, \Delta\theta_2\rangle_{\text{Fisher}}$ associated to $\mathbf{I}^{\text{Fisher}}$ reads:

$$\begin{aligned} \langle\Delta\theta_1, \Delta\theta_2\rangle_{\text{Fisher}} &= \Delta\mu_1\mathbf{K}\Delta\mu_2 + \frac{1}{2}\Delta\mathbf{K}_1(\mathbf{K}^{-1} \otimes \mathbf{K}^{-1})\Delta\mathbf{K}_2 \\ &= \Delta\mu_1\mathbf{K}\Delta\mu_2 + \frac{1}{2}\operatorname{tr}(\Delta\mathbf{K}_1\mathbf{K}^{-1}\Delta\mathbf{K}_2^T\mathbf{K}^{-1}). \end{aligned}$$

In particular, the first term in this inner product takes the form of a squared Mahalonobis distance. This will be of importance in Chapter 5, where we apply the Fisher information inner product to represent ensembles of vectors μ's.
The corresponding distance function d_F is called the *Fisher Rao.* To the knowledge of the author, no closed form general formula for d_F is available, but some cases can be computed explicitly. For example, when the covariance matrix $\boldsymbol{\Sigma}$ is diagonal, then it can be shown that in the space with $(\mu, \boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n))$ coordinates, d_F is, up to constant, equivalent to a hyperbolic metric (see [35] for details on this beautiful fact).

It also can be shown that in the limit where $\rho_1 \to \rho_2$, we have

$$\sqrt{2\,\mathrm{KL}(\rho_1, \rho_2)} \to d_F(\rho_1, \rho_2),$$

which shows that the Kullback-Leibler divergence can be seen as an approximation of the Fisher-Rao distance (or vice versa).

2.3 A note on scale invariance

A desirable feature when comparing probability distributions, and Gaussians in particular, is to satisfy some invariance under change of scale of the base space - here $\mathbb{R}^n$. It is indeed the case that when the change is linear, and the pdfs are Gaussians, then the different divergences and distances presented above do have this property.

More precisely, let $\mathbf{A}$ be a linear transformation on $\mathbb{R}^n$. Then under the linear change of coordinates $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ any Gaussian distribution $\rho = \rho(\cdot; \mu, \boldsymbol{\Sigma})$ maps on the corresponding Gaussian $\rho^{\mathbf{A}} = \rho(;\mathbf{A}\mu, \mathbf{A\Sigma A})$. If ρ_1, ρ_2 are two Gaussians, then a direct computation yields

$$\mathrm{KL}(\rho_1^{\mathbf{A}}, \rho_2^{\mathbf{A}}) = \mathrm{KL}(\rho_1, \rho_2).$$

Similarly,

$$\mathrm{KL}^{sym}(\rho_1^{\mathbf{A}}, \rho_2^{\mathbf{A}}) = \mathrm{KL}^{sym}(\rho_1, \rho_2)$$

and

$$\mathrm{MH}^{sym}(\rho_1^{\mathbf{A}}, \rho_2^{\mathbf{A}}) = \mathrm{MH}(\rho_1, \rho_2)$$

2.4 Averaging Gaussians

Suppose we are given an ensemble of Gaussian probability distributions $\{\rho_i\}_{i=1}^M$

with means and covariances $\{\mu_i, \mathbf{K}_i^{-1}\}$, and consider an associated collection of positive weights $\{\alpha_i\}$ summing up to 1. We aim at constructing an *average* Gaussian pdf

$$\rho_{av} = \rho(\cdot; \mu_{av}, \mathbf{K}_{av}^{-1})$$

of the enesmble. A first simple approach to this is based on sampling: given, for each $i = 1, \dots, M$, a collection $\{\mathbf{w}_i^j\}_{j=1}^{N_i}$ sampled from ρ_i, then μ_i and $\mathbf{K}_i^{-1}$ can be approximated by the usual formula

$$\mu_i^{samp} = \frac{1}{N_i} \sum_{j=1}^{N_i} w_i^j, \quad \mathbf{K}_i^{-1} = \Sigma_i^{samp} = \frac{1}{N_i} \sum_{j=1}^{N_i} (w_i^j - \mu_i^{samp}) \otimes (w_i^j - \mu_i^{samp}) \tag{23}$$

Notice that Σ_i^{samp} is not the common unbiased covariance estimator, since $\frac{1}{N_i - 1}$ has been replaced by $\frac{1}{N_i}$ for convenience., and for us N_i is typically very large.

Then by direct computation, it follows that the sample mean and covariance matrix associated to the ensemble

$$\bigcup_{i=1}^{M} \{\mathbf{w}_i^j\}_{j=1}^{N_i}$$

is

$$\overline{\mu} = \sum_{i=1}^{M} \alpha_i \mu_i, \quad \overline{\boldsymbol{\Sigma}} = \sum_{i=1}^{M} \alpha_i \left(\mathbf{K}_i^{-1} + \mu_i \otimes \mu_i \right) - \overline{\mu} \otimes \overline{\mu}, \tag{24}$$

where we have defined the weights

$$\alpha_i = \frac{N_i}{\sum_i N_i},$$

satisfying $\sum_{i=1}^{M} \alpha_i = 1$.

Accordingly, we can define a *sampling average* Gaussian distribution as

$$\rho_{av}^{samp} = \rho(\cdot; \overline{\mu}, \overline{\boldsymbol{\Sigma}}).$$

Another approach, or rather class of approaches, arises from solving a minimisation principle of the general form

$$\rho_{av}^{D} = \operatorname*{argmin}_{\rho \in \mathcal{C}} \sum_i \alpha_i \, \mathrm{D}(\rho_i, \rho). \tag{25}$$

Here $\mathcal{C}$ is a class of probability distributions, and D is a generic divergence (in certain cases, a distance). An elementary computation yields

$$\sum_i \alpha_i \, \mathrm{KL}(\rho_i, \rho) = \sum_i \alpha_i \, \mathrm{KL}(\rho_i, \overline{\rho}) + \mathrm{KL}(\overline{\rho}, \rho), \tag{26}$$

where $\overline{\rho} = \sum_i \alpha_i \rho_i$ is the mixture distribution the ρ_i.

From relation (26), that some authors [36] [37] refer to as the *compensation identity*, it immediately follows that

$$\operatorname*{argmin}_{\rho \in \mathcal{C}} \sum_i \mathrm{KL}(\rho_i, \rho) = \operatorname*{argmin}_{\rho \in \mathcal{C}} \mathrm{KL}(\overline{\rho}, \rho), \tag{27}$$

for $\mathcal{C}$ any set of probability distributions. In particular, setting $\mathcal{C}$ to be a set of multivariate Gaussian with banded stiffness matrix (according to a cgDNA overlapping diagonal blocks sparsity pattern), then the solution ρ_{av} to the minimisation (27) is determined by its mean and covariance

$$\mu_{av} = \overline{\mu}, \quad [[K_{av}^{-1} = \overline{\Sigma}]],$$

where the brackets restrict the matrix equality to the inside of the stencil of K_{av}. An algorithm introduced by J. Glowacki in his thesis [4] for such kind of maximum entropy fitting problem allows to compute K_{av} under the above conditions (see also Chapter 4 of the present thesis).

In what follows, the $\{\rho_i\}$ will often be cgDNA probability distribution of a set of sequences $\{S_i\}$. In this case, we will refer to ρ_{av} as the *cgDNA average* distribution for the collection of sequences $\{S_i\}$. This averaging procedure has the advantage of being consistent with the use of the KL-divergence as proximity measure between distributions, and also is easy to compute even for a large collection $\{S_i\}$.

We end this chapter with an observation that links $\mathbf{K}_{av}$ to the covariance matrix $\mathbf{C}_{\mathbf{\Sigma}}$ of the means μ_i's, defined as

$$\mathbf{C}_{\mathbf{\Sigma}} = \frac{1}{N} \sum_i (\mu_i - \overline{\mu}) \otimes (\mu_i - \overline{\mu}).$$

Indeed, from equation (24) we immediately get

$$\mathbf{K}_{av}^{-1} = \mathbf{K}_h^{-1} + \mathbf{C}_{\mathbf{\Sigma}}, \tag{28}$$

where

$$\mathbf{K}_h = \left[\frac{1}{N} \sum_i (\mathbf{K}_i^{-1}) \right]^{-1}$$

is the harmonic average of the $\mathbf{K}_i$.

3 Metric PCA: a coordinate-invariant Principal Component Analysis

The goal of this chapter is to introduce the tools that will be used in Chapter 5 for visualisation and clustering of ensembles of cgDNA+ Gaussian distributions. In particular, we address the issue of the high dimensionality of those distributions in the context of analysing large DNA sequence datasets, such as those produced either by exhaustively generating short k-mers or as the outcome of high throughput sequencing. Dimensionality reduction for data analysis is a vast and active field of study, and we certainly do not aim to provide an overview of it here. Instead, we present a simple but useful variation of Principal Component Analysis (PCA) that we call *Fisher PCA* on the space of multivariate Gaussian distributions, that is essentially a linear version of Kernel PCA with a Fisher information matrix as an alternative metric tensor. This technique exhibits the elegant property of being invariant under a linear change of coordinate of the underlying Euclidean space. It is also related to Kullback-Leibler divergence (which would be a natural choice of proximity measure in graph-based approaches such as Multidimensional Scaling or Laplacian Eigenmaps), while staying computationally very efficient due to its linear nature.

3.1 Principal Component Analysis

Here we recall what Principal Component Analysis is. It is probably the most popular dimensionality reduction method, and one of the most basic ones, but remains a very powerful tool for data analysis. There is abundant literature on the topic (see for example [38] [39]) and our goal is not to give a comprehensive overview on this method and its various applications. Rather, we will describe its fundamental principles, approached from a linear algebra perspective: in the end PCA boils down to diagonalising a sample covariance matrix, and projecting the data onto its eigenvectors.

More precisely, let $\mathcal{E} = \{\mathbf{x}_i\}_{i=1}^{p} \subset \mathbb{R}^n$ be a set of observations, or a *data ensemble.* We associate to it the $p \times n$ data matrix

$$\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_p]^{\mathrm{T}}.$$

Then the sample mean of $\mathcal{E}$ is

$$\overline{\mathbf{x}} = \frac{1}{p} \sum_{i=1} \mathbf{x}_i.$$

In what follows, we will often assume without loss of generality that the data matrix $\mathbf{X}$ is *centered*, meaning that the sample mean $\overline{\mathbf{x}} = \mathbf{0}$.

The sample covariance matrix, or *data covariance* matrix is defined as

$$\mathbf{C}_{\mathbf{X}} = \frac{1}{p}\sum_i (\mathbf{x}_i - \overline{\mathbf{x}}) \otimes (\mathbf{x}_i - \overline{\mathbf{x}}).$$

Note that for convenience, we chose the denominator to be $\frac{1}{p}$ and not $\frac{1}{p-1}$, thus opting for the biased maximum likelihood covariance estimator. For large p, the difference is negligible.

The projection $\mathbf{T}_{PCA}$ of the data matrix $\mathbf{X}$ onto the eigenvectors $\mathbf{V} := [\mathbf{v}_1 \cdots \mathbf{v}_n]$ of $\mathbf{C}_{\mathbf{X}}$ is given by

$$\mathbf{T}_{PCA} := \mathbf{X}\mathbf{V}.$$

Dimensionality reduction from n to L is obtained by reordering the columns of $\mathbf{V}$ according to the magnitude of the corresponding eigenvalue of $\mathbf{C}_{\mathbf{X}}$, and retaining only the first L eigenvectors, so that

$$\mathbf{T}^L_{PCA} := \mathbf{X}\mathbf{V}^L,$$

with $\mathbf{V}^L = [\mathbf{v}_1 \cdots \mathbf{v}_L]$ and we have assumed that the $\mathbf{v}_i$s are already arranged in descending order of their corresponding eigenvalue.

3.2 A Fisher metric PCA

In this section, we describe a variation of Principal Component Analysis based on a change of inner product - or metric - on the data space. For this reason, we will refer to it as *metric* PCA. This idea and terminology are not new (see [40], [41] and references therein). This method can be recognised as a particularly simple special case of Kernel PCA (see [38] for a detailed description), where the kernel is linear.

We then present a particular application of this method to the case where the data ensemble is a set of Gaussian distributions, and the metric is given by the Fisher information evaluated at the average (in the sense presented in Chapter 2) of the ensemble. We will show why in that case, the method has the property of being invariant under linear change of coordinates on the underlying space of the Gaussian distributions.

Let thus $\mathbf{X}$ be $p \times n$ a data matrix, $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_p]^{\mathrm{T}}$, and let $\mathbf{B}$ be a $n \times n$ positive definite matrix. The Cholesky decomposition $\mathbf{M}$ of $\mathbf{B}$ writes as

$$\mathbf{B} = \mathbf{M}^{\mathrm{T}}\mathbf{M},$$

with $\mathbf{M}$ an upper triangular square matrix. We define a new data matrix $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_p]^{\mathrm{T}}$ by

$$\mathbf{Y} = (\mathbf{M}\mathbf{X}^{\mathrm{T}})^{\mathrm{T}},$$

so that $\mathbf{Y}$ satisfies

$$\mathbf{y}_i \cdot \mathbf{y}_j = \mathbf{y}_i^{\mathrm{T}}\mathbf{y}_j = (\mathbf{M}\mathbf{x}_i)^{\mathrm{T}}(\mathbf{M}\mathbf{x}_j)^{\mathrm{T}} = \mathbf{x}_i \cdot \mathbf{B}\mathbf{x}_j.$$

In other words, the linear transformation $\mathbf{x} \mapsto \mathbf{B}\mathbf{x}$ sends the standard euclidean inner product to the new weighted inner product

$$< \cdot, >_{\mathbf{B}}: (\mathbf{w}, \mathbf{z}) \mapsto \mathbf{x} \cdot \mathbf{B}\mathbf{z},$$

In the language of Riemannian manifolds - at the risk of an certain excess of pedantry, considering the fact that the manifold here is simply $\mathbb{R}^n$ - the matrix $\mathbf{B}$ plays the role of a metric tensor.

In particular, the eigendecomposition of the data covariance matrix of $\mathbf{Y}$

$$\mathbf{C}_{\mathbf{Y}} = \frac{1}{p}\sum_i (\mathbf{y}_i - \overline{\mathbf{y}}) \otimes (\mathbf{y}_i - \overline{\mathbf{y}})$$

can be seen as a *generalised* eigenvalue problem of the form

$$\mathbf{C}_{\mathbf{X}}\mathbf{z} = \lambda \mathbf{B}^{-1}\mathbf{z} \tag{29}$$

where $\mathbf{C}_{\mathbf{X}}$ is the data covariance matrix of $\mathbf{X}$.

The generalised eigenvalues λ_i of this generalised eigenproblem are the same as the eigenvalues of the eigenproblem

$$\mathbf{C}_{\mathbf{Y}}\mathbf{v} = \lambda\mathbf{v} \tag{30}$$

of the data covariance matrix $\mathbf{C}_{\mathbf{Y}}$, and there exists (see Theorem 15.3.3 in [42]) a set of generalised eigenvectors $\{\mathbf{z}_i\}$ for (29) satisfying

$$\mathbf{z}_i \cdot \mathbf{B}^{-1}\mathbf{z}_j.$$

Since $\mathbf{B}$ matrix is non singular, there is a one-to-one correspondance

$$\mathbf{z}_i \leftrightarrow \mathbf{M}^{\mathrm{T}}\mathbf{v}_i$$

between the eigenvectors $\mathbf{z}_i$ of (29) and the eigenvectors $\mathbf{v}_i$ of (30).

Definition 6. *Let* $\mathbf{X} \in \mathbb{R}^{p\times n}$ *be a centered data matrix in* $\mathbb{R}^n$ *such that the associated covariance matrix* $\mathbf{C}_{\mathbf{X}}$ *is non singular, and* $\mathbf{B}$ *a square* $n \times n$ *positive definite matrix. Let* $\mathbf{B} = \mathbf{M}^T\mathbf{M}$ *be the upper Cholesky decomposition of* $\mathbf{B}$. *We call* metric PCA *the outcome of the projection* $\mathbf{T} = \mathbf{T}_{mPCA}$ *of* $\mathbf{Y} = (\mathbf{M}\mathbf{X}^{\mathrm{T}})^{\mathrm{T}}$ *onto the basis of eigenvectors* $\{\mathbf{v}_i\}$ *of the covariance* $\mathbf{C}_{\mathbf{Y}}$ *of* $\mathbf{Y}$*:*

$$\mathbf{T}_{mPCA} := \mathbf{Y}\mathbf{V},$$

with $\mathbf{V} = [\mathbf{v}_1 \cdots \mathbf{v}_n]$. *The* $\mathbf{v}_i$*'s are called the* metric principal components.

Remark 5.

1. *The metric PCA procedure of Definition 6 is equivalent to projecting the original data* $\mathbf{X}$ *onto the generalised eigenvectors. Namely, if* $\mathbf{Z} = [\mathbf{z}_1 \cdots \mathbf{z}_n]$*, then we have*

$$\mathbf{T}_{mPCA} = \mathbf{X}\mathbf{Z}.$$

2. *Metric PCA as defined in defintion 6 can be seen as a particular case of the well-known* kernel PCA *algorithm [43], where the kernel* $k(\mathbf{x}, \mathbf{y})$ *is linear, and given explicitly by*

$$k(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{B}\mathbf{y}.$$

Definition 7. *Let* $\mathbf{\Phi}_{\mathbf{A}} : \mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ *be a linear change of coordinate (thus* $\mathbf{A}$ *is invertible) on* $\mathbb{R}^n$*. We call a* $n \times n$ *matrix* $\mathbf{C}$ covariance-like *with respect to* $\mathbf{\Phi}_{\mathbf{A}}$ *if* $\mathbf{C}$ *transforms as* $\mathbf{A}\mathbf{C}\mathbf{A}^T$ *under the change of coordinate* $\mathbf{\Phi}_{\mathbf{A}}$*.*
Similarly, a $n \times n$ *matrix* $\mathbf{B}$ *is said to be* stiffness-like *with respect to* $\mathbf{\Phi}_{\mathbf{A}}$ *if it transforms as* $\mathbf{A}^{-T}\mathbf{B}\mathbf{A}^{-1}$ *under* $\mathbf{\Phi}_{\mathbf{A}}$*.*

Proposition 1. *Let* $\mathbf{B} \in \mathsf{S}_n^+$ *be stiffness-like. Then for any data ensemble* $\{\mathbf{x}_i\}_{i=1}^p$ *in* $\mathbb{R}^n$*, the outcome of metric PCA is invariant under any linear change of coordinate* $\mathbf{\Phi}_{\mathbf{A}}$*. That is,* $\mathbf{T}_{mPCA}$ *is identical when the data ensemble* $\{\mathbf{x}_i\}_{i=1}^p$ *is transformed into* $\{\mathbf{\Phi}_{\mathbf{A}}(\mathbf{x}_i)\}_{i=1}^p$*.*

Proof. The result follows immediately from Remark 2, and the fact that kernel PCA only depends on the values $k(\mathbf{x}_i, \mathbf{x}_j)$; because $\mathbf{B}$ is stiffness-like, one has

$$k(\mathbf{\Phi}_{\mathbf{A}}(\mathbf{x}_i), \mathbf{\Phi}_{\mathbf{A}}(\mathbf{x}_j)) = (\mathbf{A}\mathbf{x}_i)^T(\mathbf{A}^{-T}\mathbf{B}\mathbf{A}^{-1})\mathbf{A}\mathbf{x}_j = \mathbf{x}_i^T\mathbf{x}_j = k(\mathbf{x}_i, \mathbf{x}_j),$$

so that those values are invariant under the change of coordinate $\mathbf{\Phi}_{\mathbf{A}}$. ☐

Now we turn our focus on the specific case where the data ensemble $\{\mathbf{x}_i\}_{i=1}^n$ is a set of Gaussian distribution parameters $\boldsymbol{\theta}_i = (\mu_i, \mathbf{K}_i)$ sitting on the manifold $\mathcal{G}$ of Chapter 2. Strictly speaking, the $\boldsymbol{\theta}_i$'s do not sit in $\mathbf{R}^n$, but the ambient space is finite dimensional and can easily be identified with $\mathbf{R}^{n+\frac{n(n+1)}{2}}$.

For any Fisher information matrix

$$\mathbf{I}^{\text{Fisher}}(\mu, \mathbf{K}) = \begin{bmatrix} \mathbf{K} & \mathbf{0}_{n,n\times n} \\ \mathbf{0}_{n\times n,n} & \frac{1}{2}\mathbf{K}^{-1} \otimes \mathbf{K}^{-1} \end{bmatrix} \tag{31}$$

where $(\mu, \mathbf{K}) \in \mathcal{G}$ as introduced in Chapter 2 is symmetric, positive definite.

Definition 8. *We will call* Fisher PCA *the particular kind of metric PCA that is applied to a data ensemble* $\boldsymbol{\theta}_i = (\mu_i, \mathbf{K}_i)$ *of Gaussian distribution parameters by setting* $\mathbf{B} = \mathbf{I}^{\text{Fisher}}(\overline{\mu}, \overline{\mathbf{\Sigma}}^{-1})$*, where*

$$\overline{\mu} = \sum_{i=1}^{M} \alpha_i \mu_i, \quad \overline{\mathbf{\Sigma}} = \sum_{i=1}^{M} \alpha_i \left(\mathbf{K}_i^{-1} + \mu_i \otimes \mu_i\right) - \overline{\mu} \otimes \overline{\mu},$$

as defined in equation (24) of Chapter 2.

Remark 6. *The* B *matrix of Definition 8 is stiffness-like, thus Fisher PCA satisfies the invariance property of Proposition 1.*

PART II: cgDNAloc and Applications

4 cgDNAloc: dealing with non-locality

This chapter is dedicated to the development of the central mathematical tool used in this thesis. We show that marginals of banded inverse covariances with a specific overlapping block sparsity pattern are also banded with the same inherited sparsity pattern. Furthermore, we show that these marginal inverse covariances have very few entries differing from those of the corresponding submatrix of the original inverse covariance, and that the marginal stiffness matrix can be obtained from the original stiffness matrix by truncation combined with small localised modifications (as opposed to a full inversion of a matrix). This procedure is given explicitly, and allows for efficient computation of marginals of our class of banded Gaussians.

The results presented in this chapter all follow from an elegant algorithm by Glowacki [4, 44] for computing maximum entropy fitting of banded covariance matrices. That algorithm is briefly presented in the next section.

We mention that probability distributions with structured inverse covariances have been extensively studied in the context of what is now known as *graphical models* [45, 46]. These models play an important role in modern machine learning, but describing their full framework extends well beyond the scope of this thesis. Very briefly, in these models a joint (discrete or continuous) probability distribution is associated to a graph, where the nodes represent random variables, and connecting edges must satisfy some properties related to conditional dependence of the variables. The graphical model representation of a multivariate normal joint distribution, called a *Gaussian graphical model,* is an undirected graph whose edges identify with the non-zero entries in the stiffness matrix $\mathbf{K}$, (see section 5.3 in [45]). Moreover, Gaussians with banded stiffness matrices can be viewed as a (rather simple) case of a *decomposable* Gaussian graphical model. For a banded $\mathbf{K}$, it is easy to see that the *junction tree* associated to this graph ([45], section 7.1.2) is tractable, with maximum *clique-size* equal to the bandwidth. Therefore, the *Junction Tree (JT) Algorithm* ([45], section 7.1.2) can be used to tractably obtain any marginal. As already remarked in [4] [44], Glowacki's algorithm, which is our starting point, arises as a particular case of decomposition of Gaussian graphical model. By considering only the important, but particular, case of banded matrices Glowacki was able to present an explicit and simply described algorithm, with a self-contained proof involving only elementary linear algebra. The situation is analogous for the results on marginals of banded Gaussians presented in this chapter.

More precisely, marginalisation in the *JT* algorithm is performed in several steps, one of which being the factorisation of the model joint distribution ρ into a ratio of product of potentials ϕ, indexed over the *cliques* and *separator* in the associated graph. Then, an iterative *belief propagation* procedure is applied on a particular hypergraph, the junction tree, whose nodes are cliques of the model graph. In the particular case of a Gaussian joint distribution with an overlapping square banded stiffness matrix $\mathbf{K}$, the cliques and separators are associated to corresponding square sub-blocks and overlaps of the covariance matrix $\mathbf{K}^{-1}$ of ρ. The resulting junction tree in this case has the very simple form of a path graph and the decomposition of $\mathbf{K}$ yields a factorisation of ρ where the potentials ϕ take the form of multivariate Gaussians with sub-blocks and overlaps of $\mathbf{K}^{-1}$ as covariance matrices. As a result, Theorem 2 in this chapter can be seen as a particular case of Proposition 2 in [47]. Despite being less general, the proof presented in this chapter only makes use of elementary linear algebra, relying on the explicit procedure of Glowacki for computing maximum entropy fit on banded stiffness matrices. Moreover, by restricting to the particular case of banded matrices, we are able to state and prove a very simple and explicit marginalisation algorithm (see Corollary 3), which is all that is directly applied in the rest of the thesis for fast marginalisation of cgDNA+ Gaussian distributions.

4.1 A prior algorithm for maximum entropy fitting of banded covariances with overlapping squares

We start by briefly presenting Glowacki's algorithm [4, 44] for maximum entropy fitting of covariances with an overlapping squares sparsity pattern. The idea is the following: given a covariance matrix $\mathbf{C}$, together with a sparsity pattern $\mathcal{N}$ consisting of the union of square diagonal sub-blocks of $\mathbf{C}$, we want to find a completion $\widetilde{\mathbf{C}}$ of $\mathbf{C}$ such that $\widetilde{\mathbf{C}}$ coincides with $\mathbf{C}$ inside the pattern $\mathcal{N}$, and that $(\widetilde{\mathbf{C}})^{-1}$ vanishes outside the pattern $\mathcal{N}$. As previously showed in [45], such a $\widetilde{\mathbf{C}}$ exists and is unique, but Glowacki's algorithm provides a direct procedure to build the banded inverse of $\widetilde{\mathbf{C}}$.

For consistency, we will use the same notation as in [4]. In particular, for a matrix in $\mathbb{R}^{n\times n}$ we introduce an *index set* - usually denoted by $\mathcal{N}$ - as a collection of indices $\mathcal{N} \subset \{(i,j) : 1 \le i,j \le n\}$. The index set of all other indices will be denoted by $\mathcal{N}^c$, the complement to $\mathcal{N}$. The subset of indices of $\mathcal{N}$ obtained by retaining only indices less than p will be denoted by $\mathcal{N}_p$. We will put double brackets $[[\cdot]]_{\mathcal{N}}$ around a matrix to mean that we only consider entries with indices belonging to the index set $\mathcal{N}$. For example, expressions such as

$$[[\mathbf{A}]]_{\mathcal{N}} = [[\mathbf{B}]]_{\mathcal{N}}, \ \ [[\mathbf{A}]]_{\mathcal{N}^c} = [[\mathbf{B}]]_{\mathcal{N}^c}. \ \ [[\mathbf{A}]]_{\mathcal{N}_p} = [[\mathbf{B}]]_{\mathcal{N}_p}$$

are valid if and only if $\mathbf{A}$ and $\mathbf{B}$ agree in the entries associated to the index set $\mathcal{N}$, $\mathcal{N}^c$,

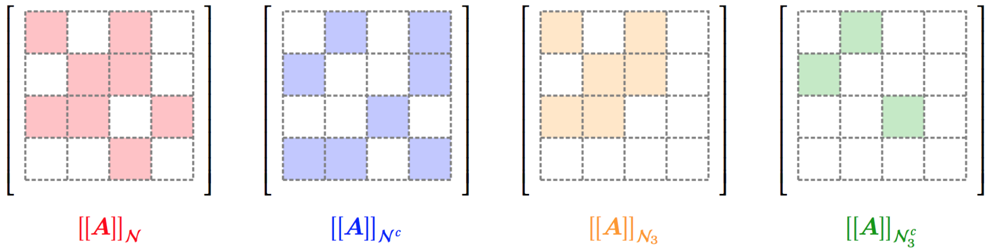

$[[\mathbf{A}]]_{\mathcal{N}}$
$[[\mathbf{A}]]_{\mathcal{N}^c}$
$[[\mathbf{A}]]_{\mathcal{N}_3}$
$[[\mathbf{A}]]_{\mathcal{N}_3^c}$

Figure 7: (Courtesy of J. Glowacki) Entries corresponding to an index set $\mathcal{N}$ on a 4×4 matrix **A**. Here $\mathcal{N} = \{(1,1),(1,3)(3,1),(2,2),(2,3),(3,2),(3,4),(4,3)\}$, $\mathcal{N}^c = \{(1,2),(2,1),(1,4),(4,1),(2,4),(4,2),(3,3),(4,4)\}$, $\mathcal{N}_3 = \{(1,1),(1,3)(3,1),(2,2),(2,3),(3,2)\}$, $\mathcal{N}_3^c = \{(1,2),(2,1),(3,3)\}$.

or $\mathcal{N}_p$ respectively (see Figure 7). We will sometimes abuse this notation, writing for instance

$$[[\mathbf{A}]]_{\mathcal{N}} = \mathbf{0}$$

to mean that (the entries of) **A** (whose indices lie) inside of $\mathcal{N}$ must vanish. Such a property being particularly useful in the context of Gaussian model fitting, where sparse (banded) matrices are often aimed for, we will sometimes refer to an index set as a *sparsity pattern*.

All square matrices are assumed to be symmetric.

For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, and integers $1 \le i \le j \le m$, $1 \le k \le l \le n$, we denote by

$$\mathbf{A}_{(i,j),(k,l)} = \begin{bmatrix} a_{i,k} & a_{i,k+1} & \cdots & a_{i,l} \\ a_{i+1,k} & a_{i+1,k+1} & \cdots & a_{i+1,l} \\ \vdots & \vdots & \ddots & \vdots \\ a_{j,k} & a_{j,k+1} & \cdots & a_{j,l} \end{bmatrix} \tag{32}$$

the block submatrix of $\mathbf{A} = (a_{p,q})_{1\le p\le m, 1\le q\le n}$ obtained by keeping only entries from row i to j and from column k to l.

In what follows, we concentrate on a specific class of index sets formed by the (possibly multiple) overlap of square diagonal sub-blocks of the matrix at hand. These index sets are completely determined by the set of top-right corners of the square sub-blocks - see Figure 8 for an illustration. Formally, we introduce the following

Definition 9. *Given an integer* $n > 1$, *we call a* ***corner set*** *(of dimension* n*) a collection* *of index pairs* $\{(i_s, j_s)\}_{s=1}^k$*, satisfying*

$$\begin{aligned} i_1 &= 1, & i_s &< i_{s+1} \leq j_s \\ j_k &= n,, & j_s &< j_{s+1}. \end{aligned}$$

Definition 10. *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *and let* $\{(i_s, j_s)\}_{s=1}^k$ *be a corner set of dimension* n*. Then* $\{(i_s, j_s)\}_{s=1}^k$ *induces an associated sequence* $\{\mathbf{A}_{[s]}\}_{s=1}^k$ *of square diagonal sub-blocks of* $\mathbf{A}$

$$\mathbf{A}_{[s]} = \mathbf{A}_{(i_s,j_s),(i_s,j_s)}.$$

In turn, the collection $\{\mathbf{A}_{[s]}\}_{s=1}^k$ *induces an* ***overlapping square index set*** $\mathcal{N}$*, consisting of all the indices of entries of* $\mathbf{A}$ *lying inside one (or several) of the sub-blocks* $\mathbf{A}_{[s]}$*.*
For each $1 < s \leq k$*, the overlap between* $\mathbf{A}_{[s-1]}$ *and* $\mathbf{A}_{[s]}$ *yields a partitioning of* $\mathbf{A}_{[s]}$ *into four sub-blocks as follows:*

$$\mathbf{A}_{[s]} = \begin{bmatrix} \mathbf{A}_{[s]_{1,1}} & \mathbf{A}_{[s]_{1,2}} \\ \mathbf{A}_{[s]_{2,1}} & \mathbf{A}_{[s]_{2,2}} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{(i_s,j_{s-1}),(i_s,j_{s-1})} & \mathbf{A}_{(i_s,j_{s-1}),(j_{s-1}+1,j_s)} \\ \mathbf{A}_{(j_{s-1}+1,j_s),(i_s,j_{s-1})} & \mathbf{A}_{(j_{s-1}+1,j_s),(j_{s-1}+1,j_s)} \end{bmatrix} \tag{33}$$

Remark 7. *Provided that all the* $\mathbf{A}_{[s]}$ *are invertible, we can write*

$$\mathbf{A}_{[s]} = \begin{bmatrix} \mathbf{A}_{[s]_{1,1}} & \mathbf{A}_{[s]_{1,2}} \\ \mathbf{A}_{[s]_{2,1}} & \mathbf{A}_{[s]_{2,2}} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{\Psi}_s(\mathbf{A}) & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{[s]_{1,1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_s(\mathbf{A}) \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{\Omega}_s(\mathbf{A}) \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{34}$$

where we have defined

$$\mathbf{\Omega}_s(\mathbf{A}) = (\mathbf{A}_{[s]_{1,1}})^{-1}\mathbf{A}_{[s]_{1,2}}, \tag{35}$$

$$\mathbf{\Psi}_s(\mathbf{A}) = \mathbf{A}_{[s]_{2,1}}(\mathbf{A}_{[s]_{1,1}})^{-1}, \tag{36}$$

$$\mathbf{H}_s(\mathbf{A}) = \mathbf{A}_{[s]_{2,2}} - \mathbf{A}_{[s]_{2,1}}(\mathbf{A}_{[s]_{1,1}})^{-1}\mathbf{A}_{[s]_{1,2}}, \tag{37}$$

or $\mathbf{\Omega}_s = \mathbf{\Omega}_s(\mathbf{A})$*,* $\mathbf{\Psi}_s = \mathbf{\Psi}_s(\mathbf{A})$*,* $\mathbf{H}_s = \mathbf{H}_s(\mathbf{A})$ *for short. In particular, we have*

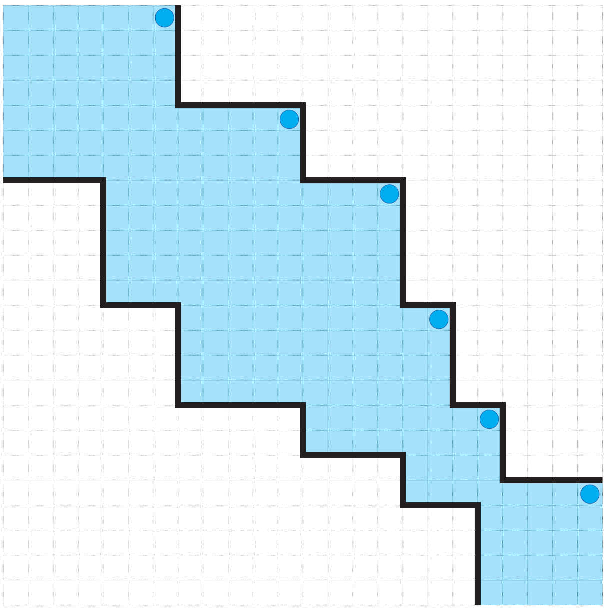


Figure 8: An example of an overlapping square index set $\mathcal{N}$ (in blue) on a 24×24 matrix formed a collection of $k = 6$ square overlapping sub-blocks $\{\mathbf{C}_{[s]}\}_{s=1}^{k}$ of various sizes, with associated cornerset $\{(1,7),(5,12),(8,16),(13,18),(17,20),(20,24)\}$. Note that, although it is not the case in this example, there can be multiple overlaps between the $\mathbf{C}_{[s]}$.

$$(\mathbf{A}_{[s]})^{-1} = \begin{bmatrix} \mathbf{I} & -\boldsymbol{\Omega}_s(\mathbf{A}) \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} (\mathbf{A}_{[s]_{1,1}})^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{H}_s(\mathbf{A}))^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\boldsymbol{\Psi}_s(\mathbf{A}) & \mathbf{I} \end{bmatrix} \tag{38}$$

$$= \begin{bmatrix} (\mathbf{A}_{[s]_{1,1}})^{-1} + \boldsymbol{\Omega}_s \mathbf{H}_s^{-1} \boldsymbol{\Psi}_s & -\boldsymbol{\Omega}_s \mathbf{H}_s^{-1} \\ -\mathbf{H}_s^{-1} \boldsymbol{\Psi}_s & \mathbf{H}_s^{-1} \end{bmatrix} \tag{39}$$

so that

$$(\mathbf{A}_{[s]})^{-1} - \begin{bmatrix} (\mathbf{A}_{[s]_{1,1}})^{-1} & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} +\boldsymbol{\Omega}_s \mathbf{H}_s^{-1} \boldsymbol{\Psi}_s & -\boldsymbol{\Omega}_s \mathbf{H}_s^{-1} \\ -\mathbf{H}_s^{-1} \boldsymbol{\Psi}_s & \mathbf{H}_s^{-1} \end{bmatrix} \tag{40}$$

Theorem 1. *Let* $\mathbf{C}$ *be a symmetric, positive definite matrix and* $\mathcal{N}$ *be an index set containing the diagonal. Let* $\mathsf{C}_{\mathcal{N}}(\mathbf{C})$ *be the set of all symmetric, positive definite matrices coinciding with* $\mathbf{C}$ *inside* $\mathcal{N}$*, that is*

$$\mathsf{C}_{\mathcal{N}}(\mathbf{C}) = \{\mathbf{B} : \mathbf{B} = \mathbf{B}^T,\ \mathbf{B} > 0,\ [[\mathbf{B}]]_{\mathcal{N}} = [[\mathbf{C}]]_{\mathcal{N}}\}.$$

Then there exists a unique matrix $\widetilde{\mathbf{C}}$ *in* $\mathsf{C}_{\mathcal{N}}(\mathbf{C})$ *whose inverse vanishes outside* $\mathcal{N}$*:*

$$\left[\left[\widetilde{\mathbf{C}}^{-1}\right]\right]_{\mathcal{N}} = \mathbf{0}.$$

Furthermore, $\widetilde{\mathbf{C}}$ *has maximum determinant amongst matrices in* $\mathsf{C}_{\mathcal{N}}(\mathbf{C})$*:*

$$\widetilde{\mathbf{C}} = \max_{\mathbf{C} \in \mathsf{C}_{\mathcal{N}}(\mathbf{C})} \det(\mathbf{C}).$$

Recalling that for a Gaussian pdf $\rho(\cdot; \mu, \mathbf{C})$ on $\mathbb{R}^N$, its entropy S is

$$S(\rho) = -\int_{\mathbb{R}^N} \rho \ln(\rho) = \frac{1}{2} \ln\left((2\pi e)^N \det(\mathbf{C})\right)$$

leads to the following definition:

Definition 11. *Given a symmetric, positive definite matrix* $\mathbf{C}$ *with sparsity pattern given by the index set* $\mathcal{N}$ *(containing all the diagonal entries). The matrix* $\widetilde{\mathbf{C}}$ *of Theorem 1 is referred to as the* maximum entropy fit *of* $\mathbf{C}$ *with respect to* $\mathcal{N}$*, and is denoted by*

$$\widetilde{\mathbf{C}} = \mathsf{Maxentf}_{\mathcal{N}}(\mathbf{C}).$$

In the particular case of an overlapping squares index set, Glowacki [4] provides a completely explicit, simple recursive procedure to compute the maximum entropy fit (in the sense of Definition 11), as well as its inverse. This implies defining two sequences of nested matrices - one for the maximum entropy fit $\widetilde{\mathbf{C}}$ and one for its inverse.

Definition 12. *Let* $\mathbf{C}$ *be a symmetric matrix,* $\{\mathbf{C}_{[s]}\}_{s=1}^{k}$ *a collection of diagonal subblocks induced by some cornerset, with associated overlapping squares index set* $\mathcal{N}$*.*
Assume that all the $\mathbf{C}_{[s]}$ *and all the overlaps* $\mathbf{C}_{[s]_{1,1}}$ *are invertible (which is typically the case e.g. if* $\mathbf{C}$ *is a covariance matrix).*
We recursively define two sequences $\{\mathbf{\Phi}_{\langle s \rangle}\}_{s=1}^{k}$, $\{\mathbf{F}_{\langle s \rangle}\}_{s=1}^{k}$ *of nested matrices as follows:*

(i) For $s = 1$,

$$\mathbf{\Phi}_{\langle 1\rangle} = (\mathbf{C}_{[1]})^{-1}, \tag{41}$$

$$\mathbf{F}_{\langle 1\rangle} = \mathbf{C}_{[1]}; \tag{42}$$

(ii) for $s > 1$,

$$\mathbf{\Phi}_{\langle s\rangle} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & -\mathbf{\Omega_s} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_{\langle s-1\rangle} & & \mathbf{0} \\ & & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{H_s}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{\Psi_s} & \mathbf{I} \end{bmatrix}, \tag{43}$$

$$\mathbf{F}_{\langle s\rangle} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Psi_s} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{F}_{\langle s-1\rangle} & & \mathbf{0} \\ & & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{H_s} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{\Omega_s} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{44}$$

where $\mathbf{\Omega}_s = \mathbf{\Omega}_s(\mathbf{C})$, $\mathbf{\Psi}_s = \mathbf{\Psi}_s(\mathbf{C})$, $\mathbf{H}_s = \mathbf{H}_s(\mathbf{C})$ are defined as in Remark 7. Note that the sequences are nested in the sense that $\mathbf{F}_{\langle s-1\rangle}$ is always a top left corner sub-block of $\mathbf{F}_{\langle s\rangle}$ (and similarily for the $\mathbf{\Phi}_{\langle s\rangle}$)

Verifying that $\{\mathbf{\Phi}_{\langle s\rangle}\}_{s=1}^k$ and $\{\mathbf{F}_{\langle s\rangle}\}_{s=1}^k$ yield the desired maximum entropy fit only involves some basic matrix computations. However, it is much easier to grasp on a concrete example, as depicted in Figure 9.

Lemma 1. *Let $\mathbf{C} \in \mathbb{R}^n$ be a symmetric matrix, and let $\{(i_s, j_s)\}_{s=1}^k$ be a cornerset inducing a collection $\{\mathbf{C}_{[s]}\}_{s=1}^k$ and an index set $\mathcal{N}$ as in Definition 12. Let $\{\mathbf{\Phi}_{\langle s\rangle}\}_{s=1}^k$, $\{\mathbf{F}_{\langle s\rangle}\}_{s=1}^k$ be the sequences of matrices of Definition 12. Then for each $s = 1, \ldots, n$, we have*

$$\mathbf{\Phi}_{\langle s\rangle} = (\mathbf{F}_{\langle s\rangle})^{-1}. \tag{45}$$

Furthermore

$$[[\mathbf{F}_{\langle s\rangle}]]_{\mathcal{N}_{j_s}} = [[\mathbf{C}]]_{\mathcal{N}_{j_s}}, \tag{46}$$

$$\left[\left[\mathbf{\Phi}_{\langle s\rangle}\right]\right]_{\mathcal{N}_{j_s}} = \mathbf{0}. \tag{47}$$

Proof. We only give a sketch here, the details can be found in [4], Lemma P1.1.1.

The proof is by induction. Property (45) follows by direct computation with the definitions of $\boldsymbol{\Phi}_{\langle s\rangle}$ and $\mathbf{F}_{\langle s\rangle}$.

For $s=1$, properties (46) and (47) are easily showed. For the induction step, observe that assuming $[[\mathbf{F}_{\langle s-1\rangle}]]_{\mathcal{N}_{j_{s-1}}} = [[\mathbf{C}]]_{\mathcal{N}_{j_{s-1}}}$ implies that $\mathbf{F}_{\langle s-1\rangle}$ can be partitioned as follows:

$$\mathbf{F}_{\langle s-1\rangle} = \begin{bmatrix} \mathbf{F}_{\langle s-1\rangle_{1,1}} & \mathbf{F}_{\langle s-1\rangle_{1,2}} \\ \mathbf{F}_{\langle s-1\rangle_{2,1}} & \mathbf{C}_{[s]_{1,1}} \end{bmatrix}. \tag{48}$$

Replacing that expression in the definition (44) of $\mathbf{F}_{\langle s\rangle}$ yields

$$\begin{aligned} \mathbf{F}_{\langle s\rangle} &= \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi}_{\mathbf{s}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{F}_{\langle s-1\rangle_{1,1}} & \mathbf{F}_{\langle s-1\rangle_{1,2}} & \mathbf{0} \\ \mathbf{F}_{\langle s-1\rangle_{2,1}} & \mathbf{C}_{[s]_{1,1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{H}_{\mathbf{s}} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \boldsymbol{\Omega}_{\mathbf{s}} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \\ &= \left[\begin{array}{ccc} \mathbf{F}_{\langle s-1\rangle_{1,1}} & \mathbf{F}_{\langle s-1\rangle_{1,2}} & \mathbf{F}_{\langle s-1\rangle_{1,2}}\boldsymbol{\Omega}_s \\ \mathbf{F}_{\langle s-1\rangle_{2,1}} & \multicolumn{2}{c}{\multirow{2}{*}{$\mathbf{C}_{[s]}$}} \\ \boldsymbol{\Psi}_s\mathbf{F}_{\langle s-1\rangle_{2,1}} & & \end{array}\right]. \end{aligned} \tag{49}$$

where $\boldsymbol{\Omega}_s = \boldsymbol{\Omega}_s(\mathbf{C})$, $\boldsymbol{\Psi}_s = \boldsymbol{\Psi}_s(\mathbf{C})$, $\mathbf{H}_s = \mathbf{H}_s(\mathbf{C})$ are the matrices of Remark 7.

Equation (49) shows that $[[\mathbf{F}_{\langle s\rangle}]]_{\mathcal{N}_{j_s}} = [[\mathbf{C}]]_{\mathcal{N}_{j_s}}$. Importantly, it also shows outside the pattern $\mathcal{N}_{j_s}$, $\mathbf{F}_{\langle s\rangle}$ that the rectangular sub-blocks of $\mathbf{C}$ above and to the left of $\mathbf{C}_{[s]}$ are given by the expressions $\mathbf{F}_{\langle s-1\rangle_{1,2}}\boldsymbol{\Omega}_s$, and $\boldsymbol{\Psi}_s\mathbf{F}_{\langle s-1\rangle_{2,1}}$ respectively.

On the other hand, a partition similar to (48) for $\boldsymbol{\Phi}_{\langle s-1\rangle}$

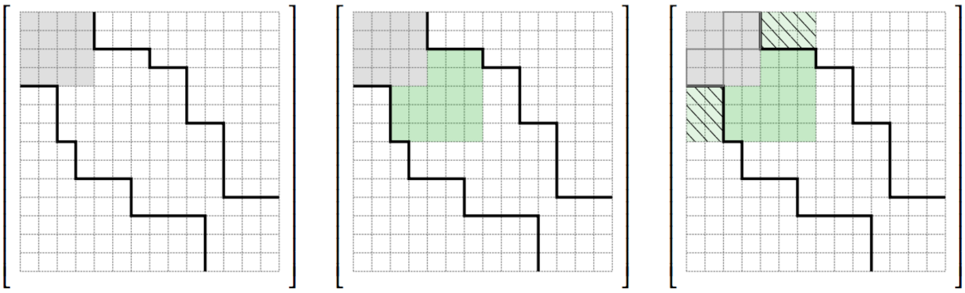
$$\boldsymbol{\Phi}_{\langle s-1\rangle} = \begin{bmatrix} \boldsymbol{\Phi}_{\langle s-1\rangle_{1,1}} & \boldsymbol{\Phi}_{\langle s-1\rangle_{1,2}} \\ \boldsymbol{\Phi}_{\langle s-1\rangle_{2,1}} & \boldsymbol{\Phi}_{\langle s-1\rangle_{2,2}} \end{bmatrix}, \tag{50}$$

inserted in the definition (43) of $\boldsymbol{\Phi}_{\langle s\rangle}$ yields

$$
\begin{aligned}
\mathbf{\Phi}_{\langle s\rangle} &= \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & -\mathbf{\Omega_s} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_{\langle s-1\rangle_{1,1}} & \mathbf{\Phi}_{\langle s-1\rangle_{1,2}} & \mathbf{0} \\ \mathbf{\Phi}_{\langle s-1\rangle_{2,1}} & \mathbf{\Phi}_{\langle s-1\rangle_{2,2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{H_s}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{\Psi_s} & \mathbf{I} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{\Phi}_{\langle s-1\rangle_{1,1}} & \mathbf{\Phi}_{\langle s-1\rangle_{1,2}} & \mathbf{0} \\ \mathbf{\Phi}_{\langle s-1\rangle_{2,1}} & \mathbf{\Phi}_{\langle s-1\rangle_{2,2}} + \mathbf{\Omega}_s \mathbf{H}_s^{-1} \mathbf{\Psi}_s & -\mathbf{\Omega}_s \mathbf{H}_s^{-1} \\ \mathbf{0} & -\mathbf{H}_s^{-1} \mathbf{\Psi}_s & \mathbf{H}_s^{-1} \end{bmatrix}.
\end{aligned}
\tag{51}
$$

The fact that $\mathbf{F}_{\langle s-1\rangle_{2,2}} + \mathbf{\Omega}_s \mathbf{H}_s^{-1} \mathbf{\Psi}_s = \mathbf{C}_{[s]_{1,1}}$, together with the induction assumption $\left[\left[\mathbf{\Phi}_{\langle s-1\rangle}\right]\right]_{\mathcal{N}_{j_{s-1}}} = \mathbf{0}$, then implies that $\left[\left[\mathbf{\Phi}_{\langle s\rangle}\right]\right]_{\mathcal{N}_{j_s}} = \mathbf{0}$ as well.

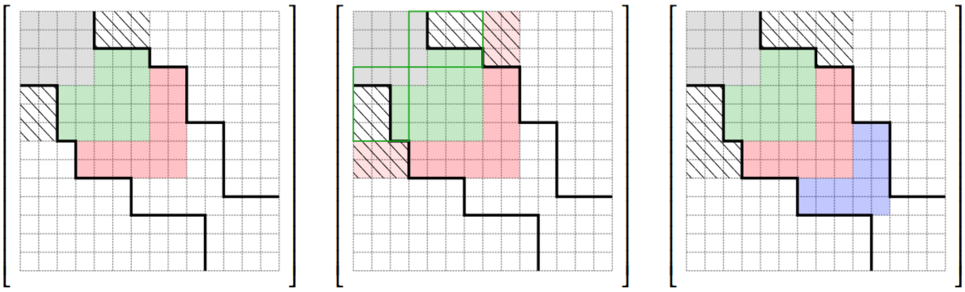
$\square$



(a) fill $[[\boldsymbol{F}_{\langle 1\rangle}]]_{\mathcal{N}_1}$ with $[[C]]_{\mathcal{N}_1}$

(b) fill the rest of $[[\boldsymbol{F}_{\langle 2\rangle}]]_{\mathcal{N}_2}$ with $[[C]]_{\mathcal{N}_2}$

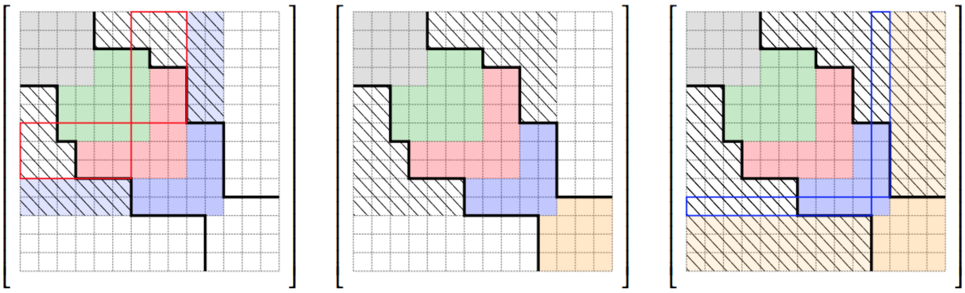
(c) fill $[[\boldsymbol{F}_{\langle 2\rangle}]]_{\mathcal{E}_2}$ with $\boldsymbol{F}_{\langle 1\rangle_{1,2}}\boldsymbol{\Omega}_2$ and its transpose



(d) fill the rest of $[[\boldsymbol{F}_{\langle 3\rangle}]]_{\mathcal{N}_3}$ with $[[C]]_{\mathcal{N}_3}$

(e) fill $[[\boldsymbol{F}_{\langle 3\rangle}]]_{\mathcal{E}_3}$ with $\boldsymbol{F}_{\langle 2\rangle_{1,2}}\boldsymbol{\Omega}_3$ and its transpose

(f) fill the rest of $[[\boldsymbol{F}_{\langle 4\rangle}]]_{\mathcal{N}_4}$ with $[[C]]_{\mathcal{N}_4}$

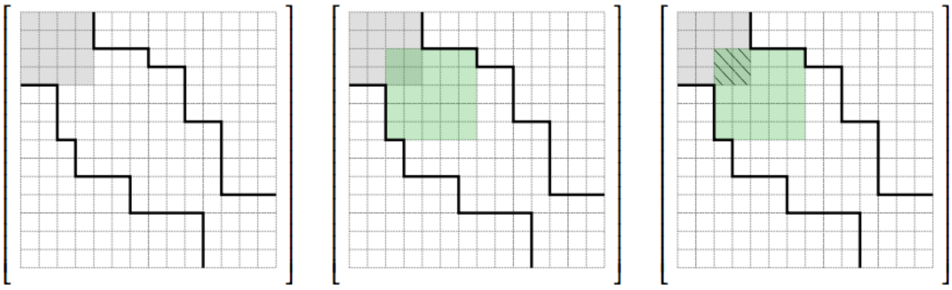


(g) fill $[[\boldsymbol{F}_{\langle 4\rangle}]]_{\mathcal{E}_4}$ with $\boldsymbol{F}_{\langle 3\rangle_{1,2}}\boldsymbol{\Omega}_4$ and its transpose

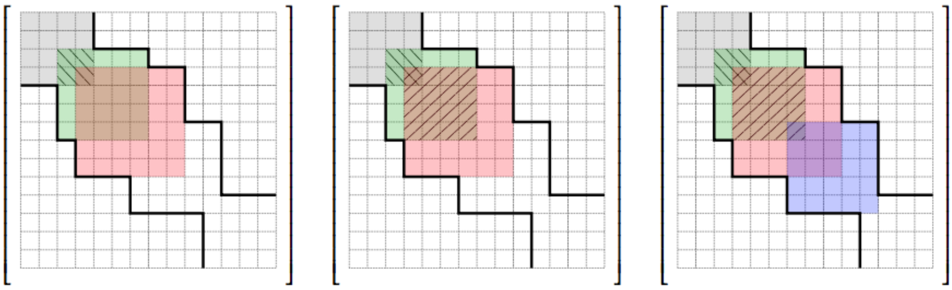
(h) fill the rest of $[[\boldsymbol{F}_{\langle 5\rangle}]]_{\mathcal{N}_5}$ with $[[C]]_{\mathcal{N}_5}$

(i) fill $[[\boldsymbol{F}_{\langle 5\rangle}]]_{\mathcal{E}_5}$ with $\boldsymbol{F}_{\langle 4\rangle_{1,2}}\boldsymbol{\Omega}_5$ and its transpose

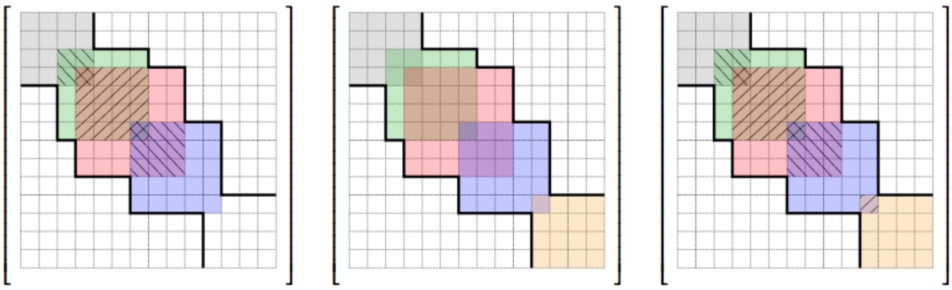
Figure 9: (Courtesy of Jarek Glowacki's thesis [4]) An illustrative example of the construction of the maximum entropy fit $\widetilde{\mathbf{C}}$ of a symmetric matrix $\mathbf{C}$. Here the overlapping squares index set $\mathcal{N}$ has associated cornerset $\{(1,4),(3,7),(4,9),(7,11),(11,14)\}$. At each step $s = 1, \ldots k = 5$, entries outside $\mathcal{N}$ in areas corresponding to $\mathbf{F}_{\langle s\rangle}$ are filled with $\mathbf{F}_{\langle s-1\rangle_{1,2}}\boldsymbol{\Omega}_s$ and its transpose, while entries inside $\mathcal{N}$ are left unchanged.



(a) add $\left(C_{[1]}\right)^{-1}$ (b) add $\left(C_{[2]}\right)^{-1}$ (c) subtract $\left(C_{[2]_{1,1}}\right)^{-1}$



(d) add $\left(C_{[3]}\right)^{-1}$ (e) subtract $\left(C_{[3]_{1,1}}\right)^{-1}$ (f) add $\left(C_{[4]}\right)^{-1}$



(g) subtract $\left(C_{[4]_{1,1}}\right)^{-1}$ (h) add $\left(C_{[5]}\right)^{-1}$ (i) subtract $\left(C_{[5]_{1,1}}\right)^{-1}$

Figure 10: (Courtesy of Jarek Glowacki's thesis [4]) An illustrative example of the construction of the *inverse* $\widetilde{\mathbf{C}}^{-1}$ of the maximum entropy fit $\widetilde{\mathbf{C}}$ of the same symmetric matrix $\mathbf{C}$ as in example of Fig 9.

Corollary 1. *Let* $\mathbf{C} \in \mathbb{R}^{n\times n}$ *be a symmetric, positive definite matrix, and let* $\{(i_s, j_s)\}_{s=1}^{k}$ *be a cornerset of dimension* n *inducing an index set* $\mathcal{N}$*. Set* $\{\mathbf{\Phi}_{\langle s\rangle}\}_{s=1}^{k}$, $\{\mathbf{F}_{\langle s\rangle}\}_{s=1}^{k}$ *as in Definition 12. Then*

(i)

$$\mathbf{F}_{\langle k\rangle} = \mathsf{Maxentf}_{\mathcal{N}}(\mathbf{C}),$$

(ii) For $s = 2, \ldots, n$*, the matrices* $\mathbf{\Phi}_{\langle s\rangle}$ *can be constructed iteratively by adding succes-*

sively the inverses of the square sub-blocks $\mathbf{C}_{[s]}$*, and substracting the inverses of the overlaps* $\mathbf{C}_{[s]_{1,1}}$*. Precisely, for* $s = 2, \dots, n$*, we have the recursive relation*

$$\boldsymbol{\Phi}_{\langle s\rangle} = \begin{bmatrix} \boldsymbol{\Phi}_{\langle s-1\rangle} & & \mathbf{0} \\ & & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{C}_{[s]})^{-1} & \\ \mathbf{0} & & \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{C}_{[s]_{1,1}})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{52}$$

(iii) Each step of the construction of $\mathbf{F}_{\langle k\rangle} = \mathsf{Maxentf}_{\mathcal{N}}(\mathbf{C})$ *leaves elements of* $\mathbf{C}$ *inside* $\mathcal{N}$ *unchanged (in the sense that* $[[\mathbf{F}_{\langle s\rangle}]]_{\mathcal{N}_{j_s}} = [[\mathbf{C}]]_{\mathcal{N}_{j_s}}$ *but replaces rectangular sub-blocks of* $[[\mathbf{C}]]_{\mathcal{N}_{j_s}}$ *directly above and to the left of* $\mathbf{C}_{[s]}$ *respectively by* $\mathbf{F}_{\langle s-1\rangle_{1,2}}\boldsymbol{\Omega}_s$*, and* $\boldsymbol{\Psi}_s\mathbf{F}_{\langle s-1\rangle_{2,1}}$*. These blocks only depends on elements of* $\mathbf{C}$ *inside of* $\mathcal{N}$*.*

Proof. (i) follows directly from relations (46) and (47) of Lemma 1 when $s = k$, together with the uniqueness of the maximum entropy fit. For (ii), equation (40) in Remark 7 allows to rewrite (51) as

$$\begin{aligned} \boldsymbol{\Phi}_{\langle s\rangle} &= \begin{bmatrix} \mathbf{F}_{\langle s-1\rangle_{1,1}} & \mathbf{F}_{\langle s-1\rangle_{1,2}} & \mathbf{0} \\ \mathbf{F}_{\langle s-1\rangle_{2,1}} & \mathbf{F}_{\langle s-1\rangle_{2,2}} + \boldsymbol{\Omega}_s\mathbf{H}_s^{-1}\boldsymbol{\Psi}_s & -\boldsymbol{\Omega}_s\mathbf{H}_s^{-1} \\ \mathbf{0} & -\mathbf{H}_s^{-1}\boldsymbol{\Psi}_s & \mathbf{H}_s^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{\Phi}_{\langle s-1\rangle} & & \mathbf{0} \\ & & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_s\mathbf{H}_s^{-1}\boldsymbol{\Psi}_s & -\boldsymbol{\Omega}_s\mathbf{H}_s^{-1} \\ \mathbf{0} & -\mathbf{H}_s^{-1}\boldsymbol{\Psi}_s & \mathbf{H}_s^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{\Phi}_{\langle s-1\rangle} & & \mathbf{0} \\ & & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{C}_{[s]})^{-1} & \\ \mathbf{0} & & \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{C}_{[s]_{1,1}})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \end{aligned}$$

as desired. □

It is relation (52) in the previous corollary that gives rise to the remarkably simple and elegant procedure to compute the inverse of the maximum entropy fit $\mathsf{Maxentf}_{\mathcal{N}}(\mathbf{C})$ only by inverting the sub-blocks $\mathbf{C}_{[s]}$ and their overlaps. This yields to significant reduction in computation complexity and time. An illustration of the procedure of an example in given in Figure 10.

4.2 Marginals of banded inverse covariances are banded

We now turn our focus towards a specific class of square, symmetric, positive definite matrices: covariance matrices.

In particular, for a n-dimensional normally distributed variable $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$, with mean parameter μ and covariance matrix Σ, it is a standard and well-known result that the marginal distribution of a subset $(x_{i_1}, \dots x_{i_k})$ of coordinates of $\mathbf{x}$ is itself normal, with mean parameter vector.

$$\widetilde{\mathbf{x}} = \begin{bmatrix} \mu_{i_1} \\ \vdots \\ \mu_{i_k} \end{bmatrix}$$

and covariance matrix $\widetilde{\boldsymbol{\Sigma}} = (\sigma_{pq})_{p,q=1}^{k}$, with $\sigma_{pq} = \Sigma_{i_p i_q}$, $p, q = 1, \dots, k$.

This motivates the following definition, which considers the case when $i_1, \dots, i_k$ are consecutive indices.

Definition 13. *Let* $\mathbf{C} \in \mathbb{R}^{n \times n}$ *be a covariance matrix. We call a square diagonal sub-block* $\mathbf{C}'$ *of* $\mathbf{C}$ *a* square marginal *of* $\mathbf{C}$. *A* square marginal pattern $\mathcal{M}$ *is an index set of the form*

$$\mathcal{M} = \{(i,j)\ : a \leq i, j \leq b \mid 1 \leq a \leq b \leq n\}.$$

The indices a *and* b *are respectively called the* starting *and* ending *index of* $\mathcal{M}$. *For such* $\mathcal{M}$, *we denote by*

$$\mathsf{Marg}_{\mathcal{M}}(\mathbf{C}) = \mathbf{C}_{(a,b),(a,b)}. \tag{53}$$

the corresponding square marginal of $\mathbf{C}$.

Introducing the notation (53) only for matrix diagonal square sub-blocks might seem unnecessarily specific (or heavy), but it helps to emphasise the applications of results below to covariances of multivariate Gaussian distributions.

We want to study the interaction between marginalisation and sparsity pattern. We therefore introduce the following definition, illustrated in Figure 12.

Definition 14. *Let $\mathcal{N}$ be any index set of dimension $n \in \mathbb{N}$, and let $\mathcal{M}$ be a square marginal pattern with starting and ending index $1 \leq a, b \leq n$. We define $\mathcal{N}^{\mathcal{M}}$, the $(b-a)$-dimensional index set* induced *by $\mathcal{M}$ on $\mathcal{N}$ as*

$$\mathcal{N}^{\mathcal{M}} = \{(i-a+1, j-a+1) : \ (i,j) \in \mathcal{N}, a \leq i, j \leq b\}.$$

Then

Theorem 2. *Let $\mathbf{C}$ be a symmetric, positive definite matrix, together with $\mathcal{N}$ an overlapping squares index set. Then for any square marginal pattern $\mathcal{M}$, we have*

$$\mathsf{Marg}_{\mathcal{M}}\left(\mathsf{Maxentf}_{\mathcal{N}}(\mathbf{C})\right) = \mathsf{Maxentf}_{\mathcal{N}^{\mathcal{M}}}\left(\mathsf{Marg}_{\mathcal{M}}(\mathbf{C})\right),$$

where $\mathcal{N}^{\mathcal{M}}$ denotes the overlapping squares index set on $\mathsf{Marg}_{\mathcal{M}}(\mathbf{C})$ induced by $\mathcal{N}$.

Proof. Let a and b be the starting, respectively ending indices of $\mathcal{M}$.

Notice first that without loss of generality, one can assume $a = 1$. For if the theorem is true in the cases when $a = 1$, then it true as well when $b = n$ (and a is arbitrary): just renumber the elements of $\mathbf{C}$ from its bottom right corner the its top left corner.

Now, for general a and b, applying this weaker version of the theorem twice successively - first with a as starting index, and $b' = n$ as ending index, then with $a' = 1$ as starting index, and b as ending index - yields the desired result. Therefore, in what follows we assume that $a = 1$.

Let $\{(i_s, j_s)\}_{s=1}^{k}$ be the cornerset of $\mathcal{N}$, and $\{\mathbf{C}_{[s]}\}_{s=1}^{k}$ the associated collection of overlapping sub-blocks of $\mathbf{C}$. Notice that

$$[[\mathbf{C}]]_{\mathcal{M}} = \mathsf{Marg}_{\mathcal{M}}(\mathbf{C}).$$

Thus, by Definition 14 of the induced index set, there exists $\widetilde{k} \leq b \leq k$ such that the cornerset of $\mathcal{N}^{\mathcal{M}}$ is given by $\{(i_s, j_s)\}_{s=1}^{\widetilde{k}-1} \cup \{(i_{\widetilde{k}}, b)\}$. The associated sequence $\{\widetilde{\mathbf{C}}_{[\mathbf{s}]}\}_{s=1}^{\widetilde{k}}$ of sub-blocks of $\mathsf{Marg}_{\mathcal{M}}(\mathbf{C})$ satisfies

$$\widetilde{\mathbf{C}}_{[s]} = \mathbf{C}_{[s]}, \quad s = 1, \dots, \widetilde{k} - 1. \tag{54}$$

Furthermore, $\widetilde{\mathbf{C}}_{[\widetilde{k}]}$ appears as a (top left) sub-block of $\mathbf{C}_{[\widetilde{k}]}$. Combining this with the overlapping squares splitting (Definition 10) of $\widetilde{\mathbf{C}}_{[\widetilde{k}]}$, namely

$$\widetilde{\mathbf{C}}_{[\widetilde{k}]} = \begin{bmatrix} \widetilde{\mathbf{C}}_{[\widetilde{k}]_{1,1}} & \widetilde{\mathbf{C}}_{[\widetilde{k}]_{1,2}} \\ \widetilde{\mathbf{C}}_{[\widetilde{k}]_{2,1}} & \widetilde{\mathbf{C}}_{[\widetilde{k}]_{2,2}} \end{bmatrix}, \tag{55}$$

we can write $\mathbf{C}_{[\widetilde{k}]}$ as

$$\mathbf{C}_{[\widetilde{k}]} = \begin{bmatrix} \widetilde{\mathbf{C}}_{[\widetilde{k}]_{1,1}} & \widetilde{\mathbf{C}}_{[\widetilde{k}]_{1,2}} & \mathbf{\Gamma}_{[\widetilde{k}]_{1,3}} \\ \widetilde{\mathbf{C}}_{[\widetilde{k}]_{2,1}} & \widetilde{\mathbf{C}}_{[\widetilde{k}]_{2,2}} & \mathbf{\Gamma}_{[\widetilde{k}]_{2,3}} \\ \mathbf{\Gamma}_{[\widetilde{k}]_{3,1}} & \mathbf{\Gamma}_{[\widetilde{k}]_{3,2}} & \mathbf{\Gamma}_{[\widetilde{k}]_{3,3}} \end{bmatrix}. \tag{56}$$

On the other hand, $\mathbf{C}_{[\widetilde{k}]}$ also splits as in Definition 10

$$\mathbf{C}_{[\widetilde{k}]} = \begin{bmatrix} \mathbf{C}_{[\widetilde{k}]_{1,1}} & \mathbf{C}_{[\widetilde{k}]_{1,2}} \\ \mathbf{C}_{[\widetilde{k}]_{2,1}} & \mathbf{C}_{[\widetilde{k}]_{2,2}} \end{bmatrix}, \tag{57}$$

thus, since $\mathbf{C}_{[\widetilde{k}]_{1,1}} = \widetilde{\mathbf{C}}_{[\widetilde{k}]_{1,1}}$, we have

$$\mathbf{C}_{[\widetilde{k}]_{1,2}} = \begin{bmatrix} \widetilde{\mathbf{C}}_{[\widetilde{k}]_{1,2}} & \mathbf{\Gamma}_{[\widetilde{k}]_{1,3}} \end{bmatrix}, \tag{58}$$

$$\mathbf{C}_{[\widetilde{k}]_{2,1}} = \begin{bmatrix} \widetilde{\mathbf{C}}_{[\widetilde{k}]_{2,1}} \\ \mathbf{\Gamma}_{[\widetilde{k}]_{3,1}} \end{bmatrix} \tag{59}$$

Similarily, let $\mathbf{F}_{\langle s \rangle}$, and $\widetilde{\mathbf{F}}_{\langle s \rangle}$ be the sequences of Definition 12 for $\mathbf{C}$, and $\mathsf{Marg}_{\mathcal{M}}(\mathbf{C})$, respectively. Then we have

$$\mathbf{F}_{\langle s \rangle} = \widetilde{\mathbf{F}}_{\langle s \rangle} \quad s = 1, \dots, \widetilde{k} - 1.$$

Now for $s = \widetilde{k}$, we have by equation (49)

$$
\begin{aligned}
\widetilde{\mathbf{F}}_{\left\langle \widetilde{k} \right\rangle} &= \begin{bmatrix} \widetilde{\mathbf{F}}_{\left\langle \widetilde{k}-1 \right\rangle_{1,1}} & \widetilde{\mathbf{F}}_{\left\langle \widetilde{k}-1 \right\rangle_{1,2}} & \widetilde{\mathbf{F}}_{\left\langle \widetilde{k}-1 \right\rangle_{1,2}} \widetilde{\boldsymbol{\Omega}}_{\widetilde{k}} \\ \widetilde{\mathbf{F}}_{\left\langle \widetilde{k}-1 \right\rangle_{2,1}} & & \\ \widetilde{\boldsymbol{\Psi}}_{\widetilde{k}} \widetilde{\mathbf{F}}_{\left\langle \widetilde{k}-1 \right\rangle_{2,1}} & \multicolumn{2}{c}{\widetilde{\mathbf{C}}_{[\widetilde{k}]}} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{F}_{\left\langle \widetilde{k}-1 \right\rangle_{1,1}} & \mathbf{F}_{\left\langle \widetilde{k}-1 \right\rangle_{1,2}} & \mathbf{F}_{\left\langle \widetilde{k}-1 \right\rangle_{1,2}} \widetilde{\boldsymbol{\Omega}}_{\widetilde{k}} \\ \mathbf{F}_{\left\langle \widetilde{k}-1 \right\rangle_{2,1}} & & \\ \widetilde{\boldsymbol{\Psi}}_{\widetilde{k}} \mathbf{F}_{\left\langle \widetilde{k}-1 \right\rangle_{2,1}} & \multicolumn{2}{c}{\widetilde{\mathbf{C}}_{[\widetilde{k}]}} \end{bmatrix},
\end{aligned}
\tag{60}
$$

where $\widetilde{\boldsymbol{\Omega}}_{\widetilde{k}} = \boldsymbol{\Omega}_{\widetilde{k}}(\widetilde{\mathbf{C}}) = (\widetilde{\mathbf{C}}_{[\widetilde{k}]_{1,1}})^{-1} \widetilde{\mathbf{C}}_{[\widetilde{k}]_{1,2}}$, and $\widetilde{\boldsymbol{\Psi}}_{\widetilde{k}} = \boldsymbol{\Psi}_{\widetilde{k}}(\widetilde{\mathbf{C}}) = \boldsymbol{\Psi}_{\widetilde{k}}(\widetilde{\mathbf{C}}) = \widetilde{\mathbf{C}}_{[\widetilde{k}]_{2,1}} (\widetilde{\mathbf{C}}_{[\widetilde{k}]_{1,1}})^{-1}$ as in Remark 7.

With still the same convention $\boldsymbol{\Omega}_{\widetilde{k}} = \boldsymbol{\Omega}_{\widetilde{k}}(\mathbf{C}) = (\mathbf{C}_{[\widetilde{k}]_{1,1}})^{-1} \mathbf{C}_{[\widetilde{k}]_{1,2}}$, we can decompose the top right corner block in (60) as follows:

$$
\begin{aligned}
\mathbf{F}_{\left\langle \widetilde{k}-1 \right\rangle_{1,2}} \boldsymbol{\Omega}_{\widetilde{k}} &= \mathbf{F}_{\left\langle \widetilde{k}-1 \right\rangle_{1,2}} \boldsymbol{\Omega}_{\widetilde{k}} \\
&= \mathbf{F}_{\left\langle \widetilde{k}-1 \right\rangle_{1,2}} (\mathbf{C}_{[\widetilde{k}]_{1,1}})^{-1} \mathbf{C}_{[\widetilde{k}]_{1,2}} \\
&= \mathbf{F}_{\left\langle \widetilde{k}-1 \right\rangle_{1,2}} (\widetilde{\mathbf{C}}_{[\widetilde{k}]_{1,1}})^{-1} \begin{bmatrix} \widetilde{\mathbf{C}}_{[\widetilde{k}]_{1,2}} & \boldsymbol{\Gamma}_{[\widetilde{k}]_{1,3}} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{F}_{\left\langle \widetilde{k}-1 \right\rangle_{1,2}} \widetilde{\boldsymbol{\Omega}}_{\widetilde{k}} & \mathbf{F}_{\left\langle \widetilde{k}-1 \right\rangle_{1,2}} (\mathbf{C}_{[\widetilde{k}]_{1,1}})^{-1} \boldsymbol{\Gamma}_{[\widetilde{k}]_{1,3}} \end{bmatrix}.
\end{aligned}
\tag{61}
$$

An analogous computation for the bottom left corner block of (60) gives

$$
\boldsymbol{\Psi}_{\widetilde{k}} \mathbf{F}_{\left\langle \widetilde{k}-1 \right\rangle_{2,1}} = \begin{bmatrix} \widetilde{\boldsymbol{\Psi}}_{\widetilde{k}} \mathbf{F}_{\left\langle \widetilde{k}-1 \right\rangle_{2,1}} \\ \boldsymbol{\Gamma}_{[\widetilde{k}]_{3,1}} (\widetilde{\mathbf{C}}_{[\widetilde{k}]_{1,1}})^{-1} \mathbf{F}_{\left\langle \widetilde{k}-1 \right\rangle_{2,1}} \end{bmatrix}.
\tag{62}
$$

A careful inspection of (60), as well as the two splittings (61) and (62) shows that $\widetilde{\mathbf{F}}_{\left\langle \widetilde{k} \right\rangle}$ does in fact coincide with $\mathbf{F}_{\left\langle \widetilde{k} \right\rangle}$ inside the square marginal pattern $\mathcal{M}$. That is,

$$
\mathsf{Marg}_{\mathcal{M}} \left(\mathbf{F}_{\left\langle \widetilde{k} \right\rangle} \right) = \widetilde{\mathbf{F}}_{\left\langle \widetilde{k} \right\rangle}.
$$

Since $\mathbf{F}_{\langle\widetilde{k}\rangle}$ is a top left sub-block of $\mathbf{F}_{\langle k\rangle}$, this implies

$$\mathsf{Marg}_{\mathcal{M}}(\mathbf{F}_{\langle k\rangle}) = \widetilde{\mathbf{F}}_{\langle\widetilde{k}\rangle}. \tag{63}$$

But Corollary 1 tells that the last element in the sequence $\{\mathbf{C}_{[s]}\}_{s=1}^{k}$ (resp. $\{\widetilde{\mathbf{C}}_{[s]}\}_{s=1}^{\widetilde{k}}$) is the maximum entropy fit of $\mathbf{C}$ (resp. $\mathsf{Marg}_{\mathcal{M}}$) with respect to $\mathcal{N}$ (resp. $\mathcal{N}^{\mathcal{M}}$):

$$\mathbf{F}_{\langle k\rangle} = \mathsf{Maxentf}_{\mathcal{N}}(\mathbf{C}),\;\; \widetilde{\mathbf{F}}_{\langle\widetilde{k}\rangle} = \mathsf{Maxentf}_{\mathcal{N}^{\mathcal{M}}}(\mathsf{Marg}_{\mathcal{M}}(\mathbf{C})).$$

Thus, (63) is the desired result, and the proof is complete.

□

Corollary 2. *Let* $\mathbf{C}$ *be a covariance matrix such that* $\mathbf{C}^{-1}$ *is banded with sparsity pattern given by an overlapping squares index set* $\mathcal{N}$*, i.e*

$$[[\mathbf{C}^{-1}]]_{\mathcal{N}} = \mathbf{0}.$$

Let $\mathcal{M}$ *be a square marginal pattern for* $\mathbf{C}$*. Then the inverse of the square marginal of* $\mathbf{C}$ *associated to* $\mathcal{M}$ *vanishes outside* $\mathcal{N}^{\mathcal{M}}$*:*

$$[[\mathsf{Marg}_{\mathcal{M}}(\mathbf{C})^{-1}]]_{\mathcal{N}^{\mathcal{M}}} = \mathbf{0}.$$

4.3 Computing the inverse of marginal covariance with banded inverse: a fast algorithm

In the previous section, we have recalled that computing the marginal probability distribution of a multivariate Gaussian $\rho(\cdot;\mu,\Sigma)$ simply boils down to cropping μ and Σ, eliminating the marginalised coordinates. In many applications, one might be interested in explicitly computing the inverse

$$\mathbf{K}_{\mathcal{M}} := (\mathsf{Marg}_{\mathcal{M}}(\mathbf{C}))^{-1}$$

of the symmetric, positive definite matrix $\mathbf{C}$.

In general, there is no reason why $\mathbf{K}_{\mathcal{M}}$ should be equal, nor actually related to $\mathbf{C}^{-1}$. In the case where $\mathbf{C}$ is the maximum entropy fit with respect to an overlapping square index set $\mathcal{N}$, Corollary 2 shows that $\mathbf{K}_{\mathcal{M}}$ inherits the same bandedness as the full inverse $\mathbf{C}^{-1}$. It actually turns out that although

$$\mathbf{K}_{\mathcal{M}} \neq \mathsf{Marg}_{\mathcal{M}}(\mathbf{C}^{-1}),$$

only some small correction is needed on $\mathsf{Marg}_{\mathcal{M}}(\mathbf{C}^{-1})$ to obtain $\mathbf{K}_{\mathcal{M}}$.

Corollary 3. *Let* $\mathbf{C}$ *be a symmetric, positive definite matrix, together with* $\mathcal{N}$ *an overlapping squares index set, with associated sequence of sub-blocks* $\{\mathbf{C}_{[s]}\}_{s=1}^{k}$. *Let* $\mathcal{M}$ *be a square marginal pattern for* $\mathbf{C}$, *and* $\{\widetilde{\mathbf{C}}_{[s]}\}_{s=1}^{\widetilde{k}}$ *the sequence of sub-blocks of* $\mathsf{Marg}_{\mathcal{M}}(\mathbf{C})$ *associated to* $\mathcal{N}^{\mathcal{M}}$.
Then $\mathbf{K}_{\mathcal{M}}$ *and* $\mathsf{Marg}_{\mathcal{M}}(\mathbf{C}^{-1})$ *coincide everywhere, except inside* $\widetilde{\mathbf{C}}_{[1]}$ *and* $\widetilde{\mathbf{C}}_{[\widetilde{k}]}$, *the first and last sub-blocks of the index set* $\mathcal{N}^{\mathcal{M}}$.

Specifically, let $\widetilde{k}_1$, *be the index such that*

$$\{\widetilde{\mathbf{C}}_{[s]}\}_{s=1}^{\widetilde{k}} = \{\widetilde{\mathbf{C}}_{[1]}, \mathbf{C}_{[\widetilde{k}_1+1]}, \dots, \mathbf{C}_{[\widetilde{k}_1+\widetilde{k}-1]}, \widetilde{\mathbf{C}}_{[\widetilde{k}]}\}$$

Then the marginal pattern $\mathcal{M}$ *induces a splitting on the inverses of* $\mathbf{C}_{[\widetilde{k}_1]}, \mathbf{C}_{[\widetilde{k}_1+\widetilde{k}]}$ *:*

$$(\mathbf{C}_{[\widetilde{k}_1]})^{-1} = \begin{bmatrix} \mathbf{B}_{[\widetilde{k}_1]_{1,1}} & \mathbf{B}_{[\widetilde{k}_1]_{1,2}} \\ \mathbf{B}_{[\widetilde{k}_1]_{2,1}} & \mathbf{B}_{[\widetilde{k}_1]_{2,2}} \end{bmatrix}, \quad (\mathbf{C}_{[\widetilde{k}_1+\widetilde{k}]})^{-1} = \begin{bmatrix} \mathbf{B}_{[\widetilde{k}_1+\widetilde{k}]_{1,1}} & \mathbf{B}_{[\widetilde{k}_1+\widetilde{k}]_{1,2}} \\ \mathbf{B}_{[\widetilde{k}_1+\widetilde{k}]_{2,1}} & \mathbf{B}_{[\widetilde{k}_1+\widetilde{k}]_{2,2}} \end{bmatrix}$$

where $\mathbf{B}_{[\widetilde{k}_1]_{2,2}}$ *have the same size as* $\widetilde{\mathbf{C}}_{[1]}$, *and* $\mathbf{B}_{[\widetilde{k}_1+\widetilde{k}]_{1,1}}$ *have the same size as* $\widetilde{\mathbf{C}}_{[\widetilde{k}]}$.

Then we have

$$\mathbf{K}_{\mathcal{M}} = \mathsf{Marg}_{\mathcal{M}}(\mathbf{C}^{-1}) + \begin{bmatrix} \widetilde{\mathbf{C}}_{[1]}^{-1} - \mathbf{B}_{[\widetilde{k}_1]_{2,2}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \widetilde{\mathbf{C}}_{[\widetilde{k}]}^{-1} - \mathbf{B}_{[\widetilde{k}_1+\widetilde{k}]_{1,1}} \end{bmatrix} \tag{64}$$

Proof. This directly follows from Theorem 2 and the iterative procedure (52) to compute the inverse of the maximum entropy fit of $\mathbf{C}$ (see also Figure 10). □

Equation (64) above means that in order to compute $\mathbf{K}_{\mathcal{M}}$ from $\mathsf{Marg}_{\mathcal{M}}(\mathbf{C}^{-1})$, it suffices to replace the contribution of the $(\mathbf{C}_{[\widetilde{k}_1]})^{-1}$ by $\widetilde{\mathbf{C}}^{-1}_{[1]}$ in the top left corner of $\mathbf{K}_{\mathcal{M}}$ (corresponding to the first square sub-block of the induced sparsity pattern $\mathcal{N}^{\mathcal{M}}$) , and the contribution of $(\mathbf{C}_{[\widetilde{k}_1+\widetilde{k}]})^{-1}$ by $\widetilde{\mathbf{C}}^{-1}_{[\widetilde{k}]}$ in the bottom right corner (corresponding to the last square sub-block of the induced sparsity pattern $\mathcal{N}^{\mathcal{M}}$) of $\mathbf{K}_{\mathcal{M}}$.

In particular, (64) provides a simple algorithm to compute $\mathbf{K}_{\mathcal{M}}$ from $\mathbf{C}^{-1}$ that requires full inversion to recover $\mathbf{C}$, but no further full inversion of $\mathsf{Marg}_{\mathcal{M}}(\mathbf{C})$. The latter is replaced with the inversion of the four smaller sub-blocks $\mathbf{C}_{[\widetilde{k}_1]}$, $\mathbf{C}_{[\widetilde{k}_1+\widetilde{k}]}$, $\widetilde{\mathbf{C}}_{[1]}$ and $\widetilde{\mathbf{C}}_{[\widetilde{k}]}$.

4.4 Non-locality and cgDNA+ marginals

Given a sequence $\mathcal{S} = \mathrm{X}_1 \cdots \mathrm{X}_N$, $\mathrm{X_i} \in \Theta = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, $i = 1, \ldots, N$, the cgDNA+ model provides a $(24N-18)$-dimensional Gaussian probability distribution

$$\rho_{cgDNA} = \rho(\cdot; \mu(\mathcal{S}), \mathbf{K}(\mathcal{S})^{-1}).$$

Here the inverse covariance, or *stiffness* matrix $\mathbf{K}(\mathcal{S})$ is a banded matrix, and is obtained by summing the $N-1$ overlapping square blocks

$$\{\mathbf{K}^{5'}_{\mathrm{X}_1\mathrm{X}_2}, \mathbf{K}_{\mathrm{X}_2\mathrm{X}_3}, \ldots, \mathbf{K}_{\mathrm{X}_{N-2}\mathrm{X}_{N-1}}, \mathbf{K}^{3'}_{\mathrm{X}_{N-1}\mathrm{X}_N}\}$$

from the cgDNA+ parameter set on its diagonal. The first and last blocks, called endblocks, are of dimension 36×36, while the $N-3$ interior blocks are of size 42×42. The overlaps are of of size 18×18. This structure gives rise to a particular sparsity pattern $\mathcal{N}_N$ of overlapping square sub-blocks. An example for $N = 5$ is shown in Figure 14

By construction, we have

$$[[\mathbf{K}(\mathcal{S})]]_{\mathcal{N}_N} = \mathbf{0},$$

thus, since $\mathbf{K}(\mathcal{S})$ is symmetric positive definite, we have that

$$\mathbf{K}(\mathcal{S})^{-1} = \mathsf{Maxentf}_{\mathcal{N}_N}(\mathbf{K}(\mathcal{S})^{-1})$$

Furthermore, any local change in the base sequence $\mathcal{S}$ - replacing X_i with $\mathrm{X}'_i \neq \mathrm{X}_i$ -

would also result in a local change in $\mathbf{K}(\mathcal{S})$, as it only affects (at most) two corresponding sub-blocks $\mathbf{K}_{\mathrm{X}_{i-1}\mathrm{X}_i}$ and $\mathbf{K}_{\mathrm{X}_i\mathrm{X}_{i+1}}$. In contrast, the associated cgDNA+ covariance matrix

$$\mathbf{C}(\mathcal{S}) := \mathbf{K}(\mathcal{S})^{-1}$$

exhibits a completely non-local sequence dependence on $\mathcal{S}$.

This property is of particular importance in any situation where $\mathcal{S}$ is embedded as a subsequence of a sequence

$$\mathcal{S}' = \mathrm{Y}_1 \cdots \mathrm{Y}_{N_l} \mathrm{X}_1 \cdots \mathrm{X}_N \mathrm{Z}_1 \cdots \mathrm{Z}_{N_r} \tag{65}$$

with

$$\mathcal{S}_l := \mathrm{Y}_1 \cdots \mathrm{Y}_{N_l} \;\; \mathcal{S}_r = \mathrm{Z}_1 \cdots \mathrm{Z}_{N_r}$$

the left and right *flanking sequences* of $\mathcal{S}$, of length $N_l, N_r \geq 1$ [2] respectively. To address this kind of context, we introduce a simple tool that will be the core of all applications of the cgDNA+ model throughout this thesis.

Definition 15. *Let* $\mathcal{S} = \mathrm{X}_1 \cdots \mathrm{X}_N$ *be a sequence,* $\mathrm{X}_i \in \Theta = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ *embedded in a longer sequence* $\mathcal{S}'$ *as in (65). Let*

$$\mu(\mathcal{S}') = \begin{bmatrix} \mu'_1 \\ \vdots \\ \mu'_n \end{bmatrix}, \quad \mathbf{K}(\mathcal{S}') = (k'_{ij})_{i,j=1}^n, \;\; \mathbf{C}(\mathcal{S}') = \mathbf{K}(\mathcal{S}')^{-1} = (c'_{ij})_{i,j=1}^n$$

be the cgDNA+ parameters associated to $\mathcal{S}'$ *with* $n = 24(N + N_l + N_r) - 18$. *We define*

$$\rho_{loc}(\cdot; \mathcal{S}, \mathcal{S}') = \rho_{loc}(\cdot; \mu(\mathcal{S}, \mathcal{S}'), \mathbf{C}(\mathcal{S}, \mathcal{S}'))$$

to be the marginal distribution of $\rho_{cgDNA}(\cdot; \mathcal{S}')$ *over all the coordinates associated to the flanking sequences* $\mathcal{S}_l$ *and* $\mathcal{S}_r$ *in* $\mathcal{S}'$. *Namely*

---

[2] In practice, for genomic applications, the order of N_l and N_r can be up to $10^6 - 10^9$.

$$\mu(\mathcal{S},\mathcal{S}') := \begin{bmatrix} \mu'_{n_1} \\ \vdots \\ \mu'_{n_2} \end{bmatrix}, \quad \mathbf{C}(\mathcal{S},\mathcal{S}')) := (c'_{ij})_{i,j=n_1}^{n_2}, \quad \mathbf{K}(\mathcal{S},\mathcal{S}')) := \mathbf{C}(\mathcal{S},\mathcal{S}'))^{-1}$$

with $n_1 = 24N_l + 1$, $n_2 = 24(N_l + N) - 18$.

We will refer to the particular marginal probability density $\rho_{loc}(\cdot;\mathcal{S},\mathcal{S}')$ of $\rho_{cgDNA}(\cdot;\mathcal{S}')$ as the *cgDNAloc* pdf associated to $\mathcal{S} \subset \mathcal{S}'$. The sequence $\mathcal{S}$ is referred to as the *core* sequence.

By construction, both the *cgDNAloc* groundstate vector $\mu(\mathcal{S},\mathcal{S}')$ and covariance matrix $\mathbf{C}(\mathcal{S},\mathcal{S}'))$ exhibit a completely non-local sequence dependence with respect to any change in $\mathcal{S}$. Importantly, differences in the flanking sequences $\mathcal{S}_l$ and $\mathcal{S}_r$ surrounding the *same* core sequence $\mathcal{S}$ lead to global changes in $\mu(\mathcal{S},\mathcal{S}')$ and $\mathbf{C}(\mathcal{S},\mathcal{S}'))$. In practice, the amplitude of these changes decrease exponentially as the modification in $\mathcal{S}'$ is introduced further away from the core sequence $\mathcal{S}$. For example the influence of the flanking sequences on $\mu(\mathcal{S},\mathcal{S}')$ becomes negligible beyond 5bp. See Figure 11.

On the other hand, it is a direct consequence of Corollary 3 that the *cgDNAloc* stiffness matrix $\mathbf{K}(\mathcal{S},\mathcal{S}'))$ is not globally affected by changes in $\mathcal{S}'$ outside $\mathcal{S}$, but only locally at the first and last blocks of its sparsity pattern. See Figure 14 for an example of this property.

Gains in computational time obtained via the cgDNAloc marginalisation algorithm compared to a naive marginalisation of stiffness matrix through full inversion of a sub-block of the covariance matrix can be quantified by running both algorithms on a randomly generated set of sequences. See Table 1 and 2. The obtained gain is expected to increase polynomially with sequence length: cgDNAloc only requires inversion of local sub-blocks of $\mathbf{C}(\mathcal{S},\mathcal{S}'))$ with fixed size. In contrast, full matrix inversion has polynomial complexity in matrix size, which also corresponds to polynomial complexity in sequence length.

Sequence length	10bp	20bp	50bp	100 bp
full inversion	2.6	8.3	60.9	383.1
cgDNAloc	2.4	5.2	23.8	95.8
Time increase	8%	37 %	61 %	75 %

Table 1: Running time (in seconds) comparison between naive marginalisation of cgDNA+ stiffness matrices of 1000 randomly generated DNA sequences of varying core length (10, 20, 50, 100 bps) through full inversion of covariance sub-block and marginalisation via the cgDNAloc algorithm. The computation is performed on a regular laptop with MATLAB. Flanking sequences are always of a fixed length of 5 bp. Time increase (in percent) when using full inversion compared to fast cgDNAloc algorithm is indicated for each window size.

Sequence length	10bp	20bp	50bp
full inversion	3.9	1.4	69.0
cgDNAloc	2.2	2.2	1.8
Time increase	74 %	426 %	3778 %

Table 2: Running time (in seconds) comparison for scanning a 1000 bp randomly generated sequence and constructing cgDNA+ marginal stiffness matrices for various window sizes (10, 20, 50). The marginal stiffness matrices are obtained by computing the covariance matrices (by full inversion) corresponding to overlapping chunks of 100 bp, and then either by full inversion of the window sub-blocks or via the cgDNAloc algorithm. The computation is performed on a regular laptop with MATLAB. Time increase (in percent) when using full inversion compared to fast cgDNAloc algorithm is indicated for each core length.

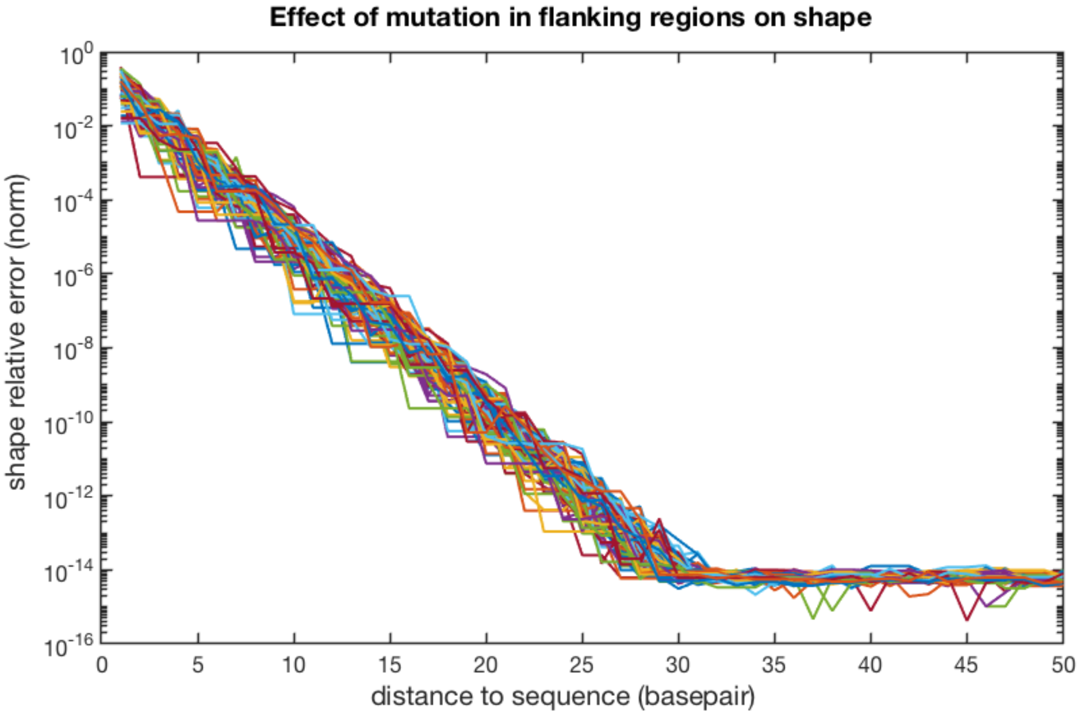


Figure 11: Relative change (in norm) of the ground state vector μ of 100 random 10bp sequences to which random flanking 50bp sequences where added and successively randomly mutated, starting furthest away at the ends and getting closer to the core sequence.

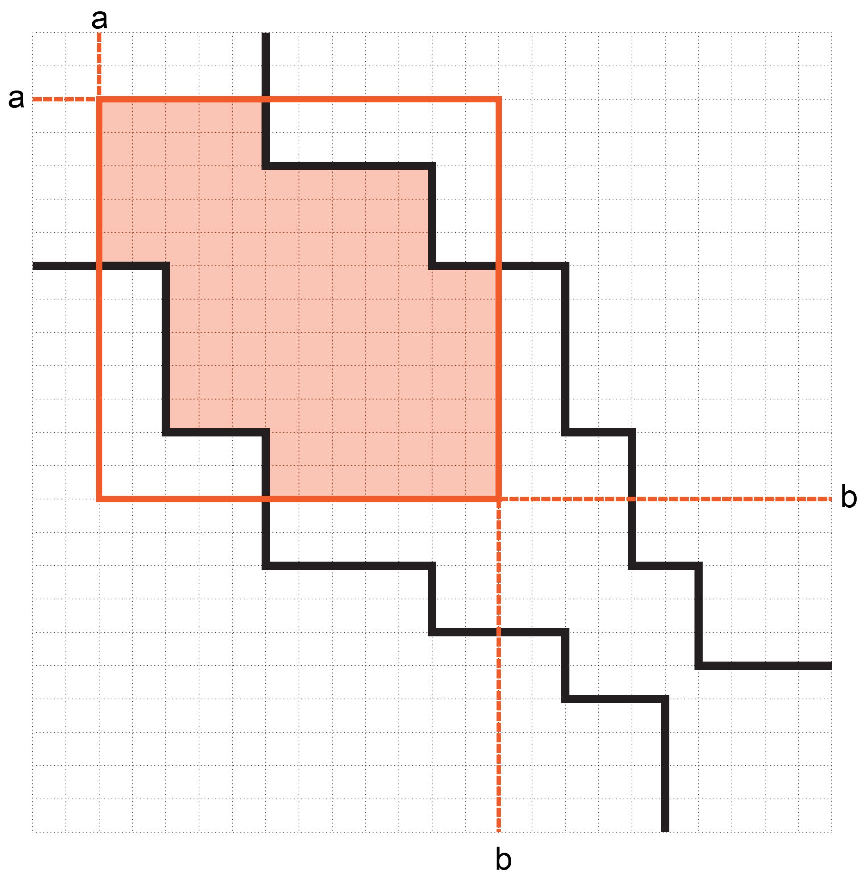


Figure 12: For the marginal $\mathsf{Marg}_{\mathcal{M}}(\mathbf{C})$ of a matrix $\mathbf{C}$, the index set $\mathcal{N}^{\mathcal{M}}$ induced by the marginal pattern $\mathcal{M}$ (the orange square) with starting index a and ending index b on an overlapping square index set $\mathcal{N}$ (in black) simply corresponds to the entries where the patterns $\mathcal{N}$ and $\mathcal{M}$ overlap.

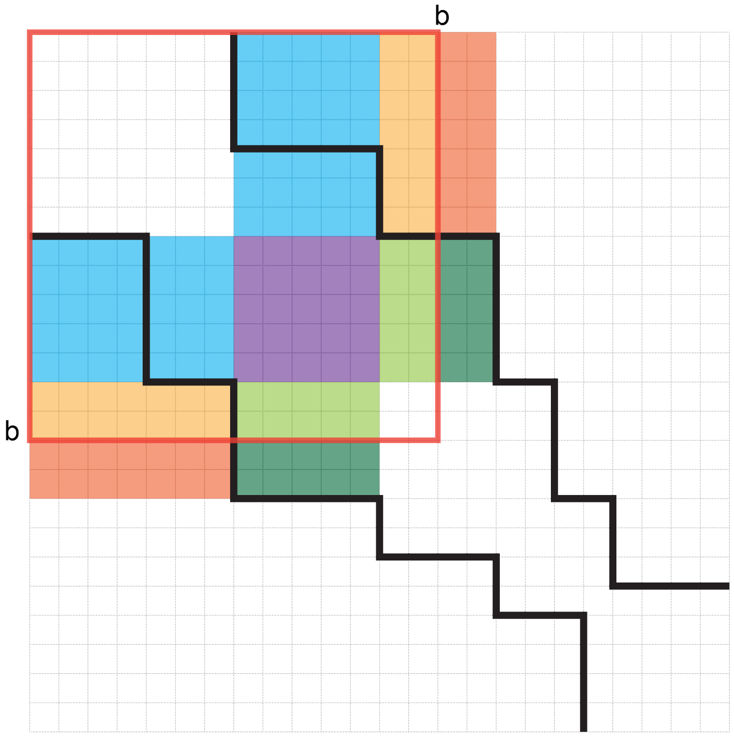


Figure 13: A schematisation of the key step in the proof of Theorem 2. The marginal pattern $\mathcal{M}$ is depicted as a red frame, with $a = 1$ and $b = 14$. This implies $\widetilde{k} = 3$, i.e. the induced sparsity pattern $\mathcal{N}^{\mathcal{M}}$ on $\mathsf{Marg}_{\mathcal{M}}(\mathbf{C})$ is formed by the 3 overlaping square blocks $\{\widetilde{\mathbf{C}}_{[s]}\}_{s=1}^{3}$, while the sparsity pattern $\mathcal{N}$ on the full matrix $\mathbf{C}$ is formed of six sub-blocks $\{\mathbf{C}_{[s]}\}_{s=1}^{6}$. Since $a = 1$, we have $\widetilde{\mathbf{C}}_{[1]} = \mathbf{C}_{[1]}$, $\widetilde{\mathbf{C}}_{[2]} = \mathbf{C}_{[2]}$, and $\widetilde{\mathbf{C}}_{[3]}$ is a top left sub-block of $\mathbf{C}_{[3]}$. The top right and bottom left corner (in light and dark orange) of $\mathbf{F}_{\langle 3\rangle}$ are given by $\mathbf{F}_{{\langle 2\rangle}_{1,2}}\Omega_3 = \mathbf{F}_{{\langle 2\rangle}_{1,2}}(\mathbf{C}_{{[3]}_{1,1}})^{-1}\mathbf{C}_{{[2]}_{1,2}}$ and its transpose. But $\mathbf{C}_{{[2]}_{1,2}}$ splits as $[\widetilde{\mathbf{C}}_{{[3]}_{1,2}} \quad \mathbf{\Gamma}_{{[3]}_{1,3}}]$, which implies that the top right and bottom left corner of $\widetilde{\mathbf{F}}_{\langle 3\rangle}$ (in light orange) have the same corresponding entries as $\mathbf{F}_{\langle 3\rangle}$. As a result, the square marginal $\mathsf{Marg}_{\mathcal{M}}(\mathbf{C})$ also have the maximum entropy property with respect to its sparsity pattern $\mathcal{N}^{\mathcal{M}}$.

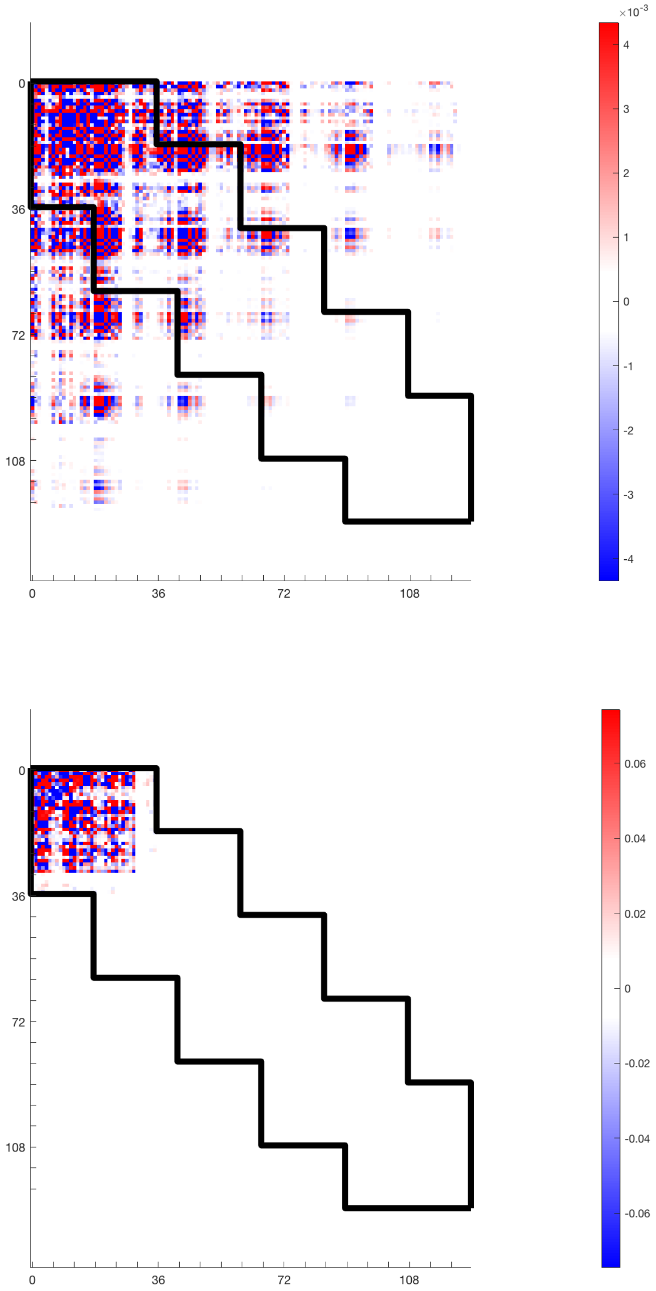


Figure 14: Effect on the cgDNA+ marginal covariance and stiffness matrices of a single bp change in the flanking sequences. Here the original sequence is $\mathcal{S} = \mathrm{AAAAA}$, embedded in $\mathcal{S}_1' = \mathrm{AAAAA}\mathcal{S}\mathrm{AAAAA} = \mathsf{poly}_{15}(\mathrm{A})$ and $\mathcal{S}_2' = \mathrm{AAAAC}\mathcal{S}\mathrm{AAAAA}$. (Top) The difference of the marginal covariances $\mathbf{C}(\mathcal{S}, \mathcal{S}_1') - \mathbf{C}(\mathcal{S}, \mathcal{S}_1')$ is non-zero both inside and outside the sparsity pattern; (Bottom) in contrast, the difference of the stiffness matrices $\mathbf{K}(\mathcal{S}, \mathcal{S}_2') - \mathbf{K}(\mathcal{S}, \mathcal{S}_1')$ is localised in the first square sub-block of the pattern.

5 Visualising and Clustering cgDNA+ predictions

5.1 An exhaustive study of short sequence fragments

This chapter is dedicated to the study of cgDNA+ predictions for exhaustive ensembles of short DNA sequences. In particular, we present visualisations after projections of shape vectors, first via simple Principal Component Analysis (PCA), then by applying a weighted version of PCA that we have named Fisher PCA, as described in Chapter 3. For both methods, properties of corresponding eigenvalues spectra are highlighted, and dimensionality discussed. Based on these projections, the presence or absence of natural clustering of the ensemble data is considered.

The computation presented in this chapter were originally done using the first parameter set $ps1$ introduced by Patelli in his PhD thesis [2]. They were then replicated with a more recent parameter set $MDNA$ computed by Sharma [10]. We discuss the resulting differences in the last section of this chapter.

Setting a scale for k-mer comparison

The systematic comparison of short sequences of length k bp, also called k-mers, requires a reference scale to be set. For this aim, motivated by Property 1 in Chapter 2, we use a normalised version of the Kullback-Leibler divergence:

$$\mathrm{KL}(\rho_1,\rho_2)^{sym}_{deg} = \frac{1}{n}\,\mathrm{KL}^{sym}(\rho_1,\rho_2), \tag{66}$$

where ρ_1, ρ_2 are Gaussian distribution on $\mathbb{R}^n$. In the case of cgDNA+ distributions of k-mer fragments, we have $n = 24k - 18$.

We refer to table 6.1 in [10], which shows an average error between MD simulated distributions and cgDNA+ reconstructed pdf of the order of $\mathcal{E}_{KL_{sym}} = 0.03$. This number will serve as a baseline reference scale, under which divergence between two different cgDNA+ or cgDNAloc pdfs will be considered negligible.

To illustrate this scale, we show examples of 1D Gaussian distribution in Figure 15.

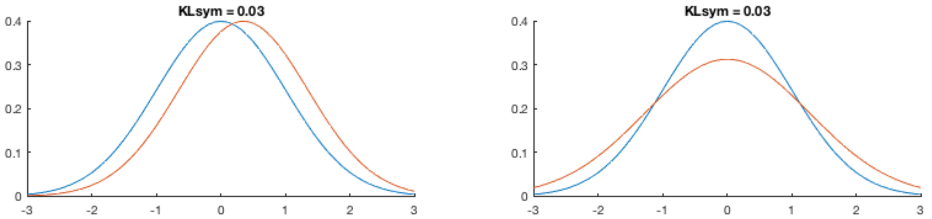


Figure 15: Two examples of 1-dimensional Gaussian distributions ρ (in red) $\mathcal{N}(\mu, \sigma)$ such that $\mathrm{KL}_{sym}(\rho, \rho_0) = 0.03$, with ρ_0 (in blue) a standard normal distribution $\mathcal{N}(0, 1)$. On the left: $\sigma = 1$; on the right: $\mu = 0$.

Methods

To generate the sequence ensembles, we proceed as follows: for $k = 2, \ldots, 10$, every 4^k possible k-mers, i.e. sequences $\mathrm{S}_i = \mathrm{X}_{i,1} \cdots \mathrm{X}_{i,k}$ of length k are produced to form the ensemble $\mathcal{S} = \{S_i\}_{i=1}^{4^k}$. According to the range (5 bps) of non locality of sequence dependence observed in Chapter 4, each S_i is then padded by random flanking sequences $\mathrm{X}_{i,l} = \mathrm{Y}_{i,1} \cdots \mathrm{Y}_{i,5}$ and $\mathrm{X}_{i,r} = \mathrm{Z}_{i,1} \cdots \mathrm{Z}_{i,5}$ of length 5 on each side of S_i, yielding a unique sequence

$$\mathrm{S}_i' = Y_{i,1} \cdots \mathrm{Y}_{i,5}\ \mathrm{X}_1 \cdots \mathrm{X}_k\ \mathrm{Z}_{i,1} \cdots \mathrm{Z}_{i,5}.$$

The cgDNA+ output shape vector and stiffness matrix of each S_i' are computed. After marginalisation over the flanking sequences $\mathrm{X}_{i,l}$ and $\mathrm{X}_{i,r}$ on each side of S_i, implemented by the procedure described in Chapter 4, this yields the two collections of ensembles of marginal shape vectors $\mathcal{M}_k = \{\mu_i\}_{i=1}^{4^k} \subset \mathbb{R}^d$, $d = 24k - 18$, and of marginal stiffness matrices $\mathcal{F}_k = \{\mathbf{K}_i\}_{i=1}^{4^k} \subset \mathbb{R}^{d^2}$. Although these ensembles and their properties a priori depend on the choices of flanking sequences $\mathrm{X}_{i,l}$ and $\mathrm{X}_{i,r}$, none of the observations described below are affected by this choice in any significant way. Indeed, ensembles of cgDNAloc pdfs were also generated and averaged over some exhaustive or non exhaustive sets of flanking sequences of a fixed length. These pdfs were on average very close to the ones computed with random flanking, with a mean error in terms of symmetric Kullback Leibler divergence of $\mathcal{E} = 0.007$, well below the scale introduced in the previous paragraph. However, replacing averaged pdfs with randomly flanked ones implies that the ensemble $\mathcal{M}_k$ do not necessarily satisfy the palindromic condition $\mathbf{E} \cdot \mathcal{M}_k = \mathcal{M}_k$, where $\mathbf{E}$ is the involution introduced in Chapter 1. To remedy this problem, with purposely symmetrise the ensemble $\mathcal{M}_k$ by adding all the symmetrical ground states $\mathbf{E} \cdot \mathcal{M}_k$ to it.

A wide variety of approaches for data clustering are available to use [48]. For example, hierarchical clustering techniques are very common methods, and are based on iterative cluster agglomeration or cluster division via a default (usually euclidean) or ad hoc distance between data points - although it does not need in general to satisfy the axioms of a distance. But this requires computation of pairwise distances, which becomes quadratically expensive with the size of dataset. In the particular case at hand of exhaustive ensembles of k-mer cgDNA+ shape vectors, the ensemble size grows exponentially as 4^k, thus rendering the former type of approach costly to implement. Moreover, because some scale-invariance property is desired, the use of the symmetric Mahalanobis distance (see Chapter 2) appears as a natural choice, increasing the computational cost of hierarchical clustering methods even more. As a workaround to these issues, we start by reducing the dimensionality of the data. For this aim, there is again a great variety of different approaches and methods, ranging from simple feature selection techniques to manifold learning [49] [50] [51] [52]. Although we have applied Multidimensional Scaling (MDS) [53] and Laplacian eigenmaps [49] on the given dataset, these two methods also suffer from the cost of taking a distance matrix as an input, whose computational cost again becomes too high as k increases. Therefore, we will not discuss the outcome of these methods here. Instead, we focus on the outcome of PCA and the Fisher PCA method described in Chapter 3. The former, being a very standard and widely used algorithm, will serve as a reference, while the latter enjoys both the scale-invariance property (see Remark 6) and being a linear dimensionality reduction method, thus scaling well with dataset size. It is also possible to visualize the data in dimension 2 or 3 by using dimensionality reduction, which by inspection can reveal clusters or other interesting characteristics.

Notice that on shapes, PCA assumes a Euclidean metric to compare shape vectors, whereas Fisher PCA (only the shape component) corresponds to replacing the euclidean metric by a version of the Mahalanobis distance, where the weight $\mathbf{K}_2$ is replaced by the constant matrix $\mathbf{K}_{av}$. As such, we can regard Fisher PCA on shapes as an efficient way of computing an approximation of the pairwise Mahalonobis distance matrix of the dataset through the averaging of all stiffness matrices $\mathbf{K}_i$.

In both standard PCA and Fisher PCA, projection on a lower dimensional subspace can be performed as a way of reducing size of the data by eliminating modes of low variance. There is no absolute principles to determine the choice of a cutoff for the dimension p of the subspace, but there are rules of thumb - e.g. thresholds on cumulative variance (cumulative sum of eigenvalues must represent 90, 95, or 99% of total variance). We rather focus on inspection of the spectra, and look for a spectral gap. More precisely, for an eigenvalue spectrum $\{\lambda_i\}_{i=1}^{d}$, the dimension p to be retained will be selected as

$$p = \underset{i=1,\ldots,d-1}{\operatorname{argmax}} \ \lambda_{i+1} - \lambda_i. \tag{67}$$

The existence of a spectral gap at low dimension will allow for visualisation, and potentially lead to sensible clustering. Moreover, a small dimension p allows for efficient use of standard clustering algorithms in $\mathbb{R}^p$. Here we have chosen K-means clustering, a standard, well-established algorithm [54]. The algorithm requires the number of desired clusters K to be provided as an input. This hyper-parameter K is chosen according to the silhouette criterion [55], in the range $[2, 2^{k+1}]$.

Standard PCA

For brevity, and since the results are qualitatively very similar, we focus on the outcome of PCA for the ensemble of 4-mers. Figure 16 shows the spectrum of the shape covariance matrix associated with the ensemble. That is, the spectrum of $\mathbf{X}^T\mathbf{X}$, where $\mathbf{X} = [\mu_1 \cdots \mu_p]^T \in \mathbb{R}^{p\times n}$ is the centered data matrix of cgDNAloc ground states of dimension $p = 24 \times 4 - 6 = 90$.

A first remark is that the eigenvalues λ of this spectrum are all simple,even in the low part of the spectrum in Figure 16, with eigenvalues appearing all close to zero, while in reality they are all distinct and range from 10^{-4} to 10^{-7}. This fact has an interesting consequence on the structure of the set of eigenvectors, that we now describe.

Suppose $\mathcal{M} = \{\mu_i\}_{i=1}^p$ is any ensemble of vectors in $\mathbb{R}^n$ with the property to be $\mathbf{E}$-invariant, with $\mathbf{E} \in \mathbb{R}^{n\times n}$ any involution, i.e. $\mathbf{EE} = \mathbf{I}_n$. That is, we have

$$\mathbf{E} \cdot \mathcal{M} = \mathcal{M}. \tag{68}$$

We define the *shape covariacne* as the sample covariance matrix of the ensemble $\mathcal{M}$, which can be written as

$$\mathbf{C}_{\mathcal{M}} = \frac{1}{M} \sum_i (\mu_i - \overline{\mu}) \otimes (\mu_i - \overline{\mu}),$$

where

$$\overline{\mu} = \frac{1}{M} \sum_i \mu_i$$

is the sample mean, which satisfies

$$\mathbf{E}\,\mathbf{C}_{\mathcal{M}}\,\mathbf{E} = \mathbf{C}_{\mathcal{M}}.$$

In particular, since $\mathbf{E}$ is an involution, we have

$$\mathbf{C}_{\mathcal{M}}\,\mathbf{E} = \mathbf{E}\,\mathbf{C}_{\mathcal{M}}.$$

This immediately implies the following property: if λ is an eigenvalue of $\mathbf{C}_{\mathcal{M}}$, associated to the unit eigenvector v, then either λ has multiplicity 2, or $\mathbf{E}v = \pm v$.

Furthermore, the fact that $\mathbf{E} \cdot \mathcal{M} = \mathcal{M}$ implies that the data matrix $\mathbf{X}$ satisfies

$$\mathbf{E}\mathbf{X} = \tilde{\mathbf{X}}$$

where $\tilde{\mathbf{X}} = [\mu_{\sigma(1)} \cdots \mu_{\sigma(p)}]^T$, and σ is a permutation of $(1, \ldots, p)$ (Note that in case of cgDNA+ groundstates, σ fixes the μ's associated to palindromic sequences).

Put together, the facts above yield the following remark:
Remark 8. *Suppose that all eigenvalues of* $\mathbf{C}_{\mathcal{M}}$ *are simple. Let* $\mathbf{T}_{PCA} = [\mathbf{t}_1, \ldots, \mathbf{t}_n]^T$, *then for each* $i = 1, \ldots, p$, *we have*

$$\mathbf{t}_i = \mathbf{D}\mathbf{t}_{\sigma(i)},$$

where σ *is the permutation of* $(1, \ldots, p)$ *such that* $\mathbf{X}^T = \mathbf{E}\tilde{\mathbf{X}}^T$, *and* $\tilde{\mathbf{X}} = [\mu_{\sigma(1)} \cdots \mu_{\sigma(p)}]^T$ *and* $\mathbf{D} = diag(\alpha_1, \ldots, \alpha_n)$, *with* $\alpha_i \in \{-1, +1\}$ *the eigenvalues of* $\mathbf{E}$.

The eigenvalue spectrum of the shape covariance matrix for the exhaustive 4-mer ensemble is shown in Figure 16. All the eigenvalues are simple, even though they appear in approximate pairs (1 & 2, 3 & 4, etc.). Thus the symmetry of eigenvectors described in Remark 8 does apply.

The two main gaps that can be observed are between the second and the third eigenvalues, and between the 4th and the 5th. In the case of the particular ensemble of 4-mers used here (keeping in mind that the random flanking can lead to some variability in that regard), the maximum gap in the eigenvalue spectrum appears after the fourth eigenvalue, which yield a dimension for projection $p = 4$. We then show all possible 2D projections in Figure 17. These projections exhibit various axial symmetries, that are directly related to Remark 8.

For reasons to which we will return in the next section, we then colour each data point according to the unique translation of the corresponding DNA sequence to the purine/ pyrimidine alphabet. See also Figure 18 for the easier case of 2-mers, where the sequences are labeled. In both cases, one can observe some proximity between sequences that translate to the same representative in the R/Y alphabet. However, no presence of obvious clustering can be inferred from these scatter plots. In general, it does not appear that PCA reveals a clear structure (e.g. clustering) on ensembles of cgDNAloc ground states of 4-mers. This fact will be made more precise in the next section, where a comparison will be drawn with the outcome of clustering on ground states projected via Fisher PCA. We mention that this observation is not specific to the choice of $k = 4$, but also holds for all the higher values of k studied (up to $k = 9$) - with perhaps the exception of $k = 2$, for which translation of the sequences in the R/Y alphabet appears to yield separate clusters (see Figure 18).

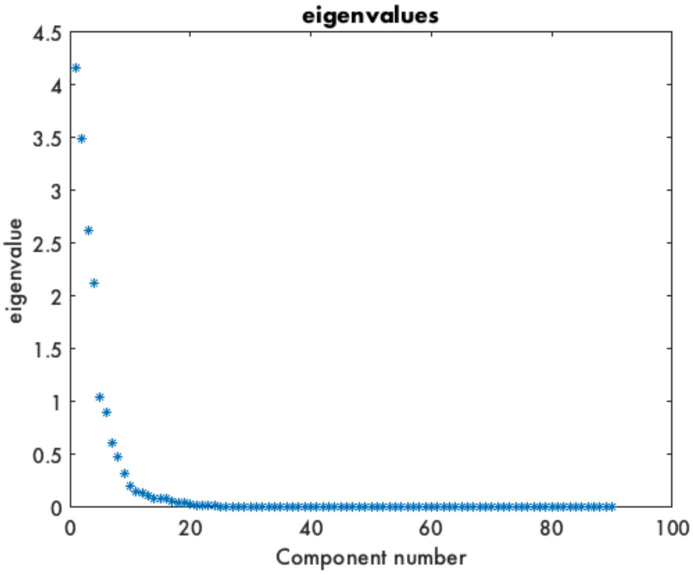


Figure 16: Standard PCA: eigenvalue spectrum of $\mathbf{C}_X$ the shape covariance matrix of exhaustive ensemble of cgDNAloc ground state (shape) vectors for 4-mer sequences. The two main gaps appear after the second and fourth eigenvalues. Despite being very close to zero, eigenvalues in the lower, approximately continuous part of the spectrum are all positive and distinct.

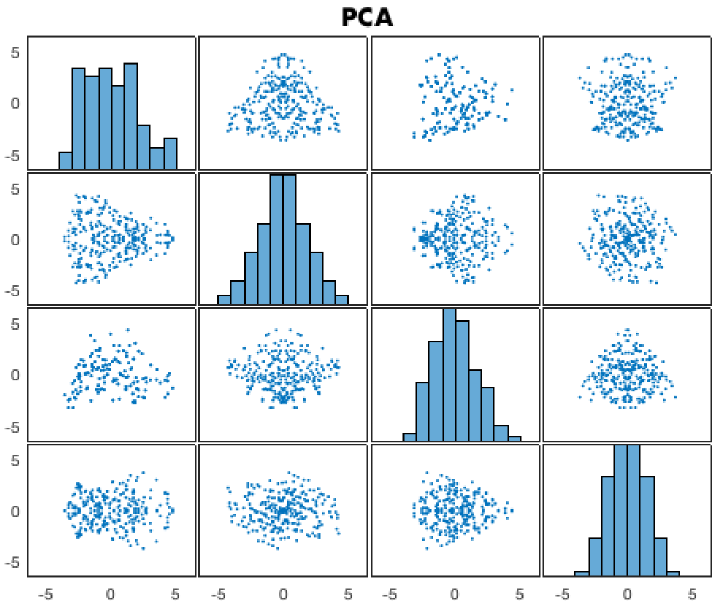


Figure 17: All possible 2-dimensional projections on the 4 principal components of PCA, from PC1 (top and left) to PC4 (bottom and right) for exhaustive ensemble of cgDNAloc ground state (shape) vectors for 4-mers. The subplot on ith row and jth column shows the projection on the 2-dimensional subspace corresponding to PCi and PCj (thus pairs of subplots placed symmetrically with respect to the diagonal are identical up to symmetry). Subplots on the diagonal show histograms of the distribution of the shape vector one-dimensional projection onto PC1 (top left) to PC4 (bottom right). Symmetries (e.g. subplot for projection on PC1 and PC2) can be observed that reflect the enforcement of Watson-Crick symmetry on the ensemble of shape vectors - see equation (68).

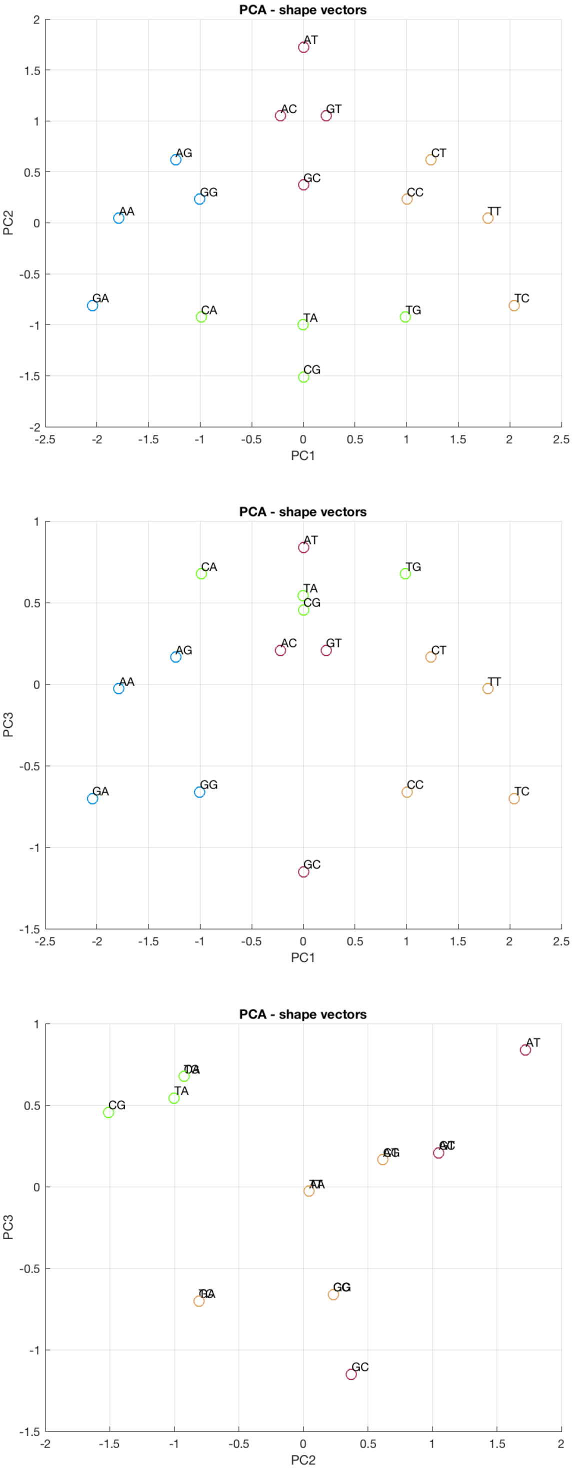


Figure 18: 2-dimensional projections on the 3 principal components PC1, PC2, and PCA3 of standard PCA for exhaustive ensemble of cgDNAloc groundstate (shape) vectors for 2-mers. Points with the same colour correspond to identical sequences in the purine/pyrimidine alphabet.

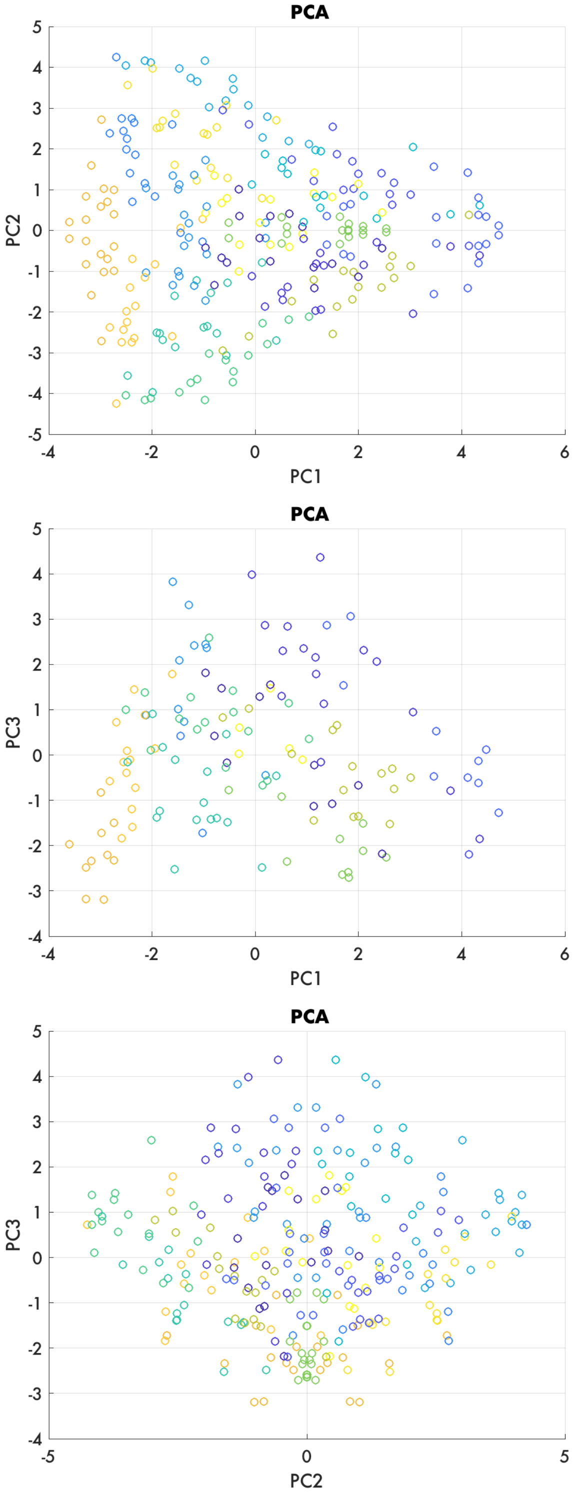


Figure 19: 2-dimensional projections on the 3 principal components PC1, PC2, and PC3 of standard PCA of exhaustive ensemble of cgDNAloc ground state (shape) vectors for 4-mers. Points with the same colour correspond to identical sequences in the purine/pyrimidine alphabet. There is no visually obvious clustering.

Fisher PCA yields a Purine/Pyrimidine Clustering

We now present the outcome of replacing the standard PCA with Fisher PCA (see Definition 8 in Chapter 3) for the study of exhaustive k-mer ensembles. Again, we focus on the case when $k = 4$, but all the results presented here also apply to the other values of $k = 3, \ldots, 9$. Recall that the Fisher PCA relies on finding the generalised eigenvalues of the generalised eigenproblem

$$\mathbf{C_X v} = \lambda \mathbf{K}_{av}^{-1} \mathbf{v}, \tag{69}$$

as introduced in Chapter 3. The presence of the average inverse stiffness matrix on the right-hand side of equation (69) ensures non-dimensionality of the problem, thus leading to an embedding that is invariant under change of scale on the shape vectors.

The spectrum of the generalised eigenvalues for the case $k = 4$ is shown in Figure 20. A clear gap is present between the 4th and 5th eigenvalues. Figures 21 and 22 show generalised eigenvalue spectra for all the other values of $k = 3, \ldots, 9$. Again, the systematic presence of a gap after the kth eigenvalue can be observed. Furthermore, the top eigenvalues appear to come closer to identical pairs (first and second, third and fourth) as k increases. This phenomenon might be related to the symmetry (68) of the ensemble of shape vectors, however the mechanism responsible for it remains to be precisely understood.

As in the PCA case, all eigenvalues are simple, even on the lower part of the spectrum. One oberves a particularly striking resemblance of the spectra shown in Figures 21 and 22 for varying k, both in terms of range of the eigenvalues and in terms of overall structure. While similar ranges can easily be attributed to the non-dimensionality induced by the weighting of the eigendecomposition by $\mathbf{K}_{av}^{-1}$, similarities in structure suggest an underlying hierarchy of mixed modes, independent of length, governing the behavior of cgDNA+ ground states. Careful inspection of the generalised eigenvectors did not provide any particular indication on the precise nature of these modes, so it remains to be further investigated. Importantly, those features were observed to be very stable when changing the randomly generated flanking sequences of the k-mers in the ensembles. This strongly suggests that the dominant differences in cgDNA+ ground state predictions reside in k distinct modes.

To investigate the reason behind this gap, we now turn to the visualisation of the projection of the ensemble via Fisher PCA. As implied by the criterion (67) and the systematically observed gap in eigenvalue spectrum, the dimension p of the space projected onto is equal to k for each ensemble of k-mer shape vectors. Outcomes of the projections onto the first k generalised eigenvectors are shown in Figure 23. As in the PCA case, projections exhibit axial symmetries. These can be explained with an argument, *mutatis mutandis*, identical to the one of Remark 8. In contrast to the PCA case, visual inspection suggests the presence of clusters, particularly in those 2-dimensional projections involving the first principal component 1 (PC1).

To investigate the nature of these clusters, we perform a standard K-means algorithm on the k-dimensional projection of the data matrix $\mathbf{X}$, following the procedure described at the beginning of the present chapter. The silhouette criterion [55] always lead to select an optimal number for clusters of $K = 2^k$, each containing approximately (but very close to exactly) 2^k data points. Sequence logos of the corresponding sequences for each cluster are shown in Figure 27. They show a particularly clear classification of the sequences according to their translation into the purine/pyrimidine alphabet. This is confirmed when coloring data points of the scatter plots of the projection according to the R/Y content of the corresponding k-mers. Indeed, as can be seen in Figure 26, sequences sharing the same R/Y content not only gather together, but tend to belong to the same clusters that are visible in Figure 23. In contrast, K-means clustering applied to the k-dimensional standard PCA projection of the same ground state ensembles did not lead to an optimal number of $K = 2^k$ clusters, nor to a clear classification of the sequences (see Figure 28). The difference is also particularly striking on 3D projection onto the first three PC; views of 3D scatter plots of those 3D projections for PCA and for Fisher PCA are shown on Figure 24 and 24. These results do not depend on the sequence length k, nor on random initiation of the K-means algorithm in any significant way.

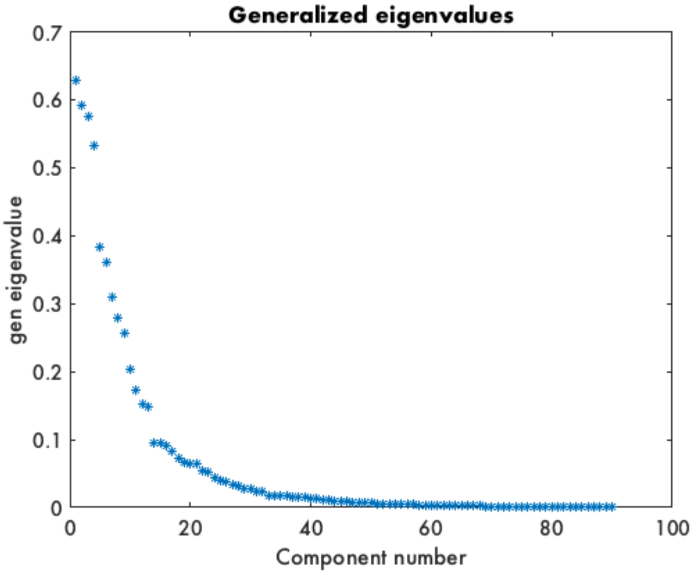


Figure 20: Fisher PCA: generalised eigenvalue spectrum of the generalised eigenvalue problem (69) of exhaustive ensemble of cgDNAloc ground state (shape) vectors for 4-mer sequences. A clear gap appears after the kth eigenvalue, as is also typical for other values of k.

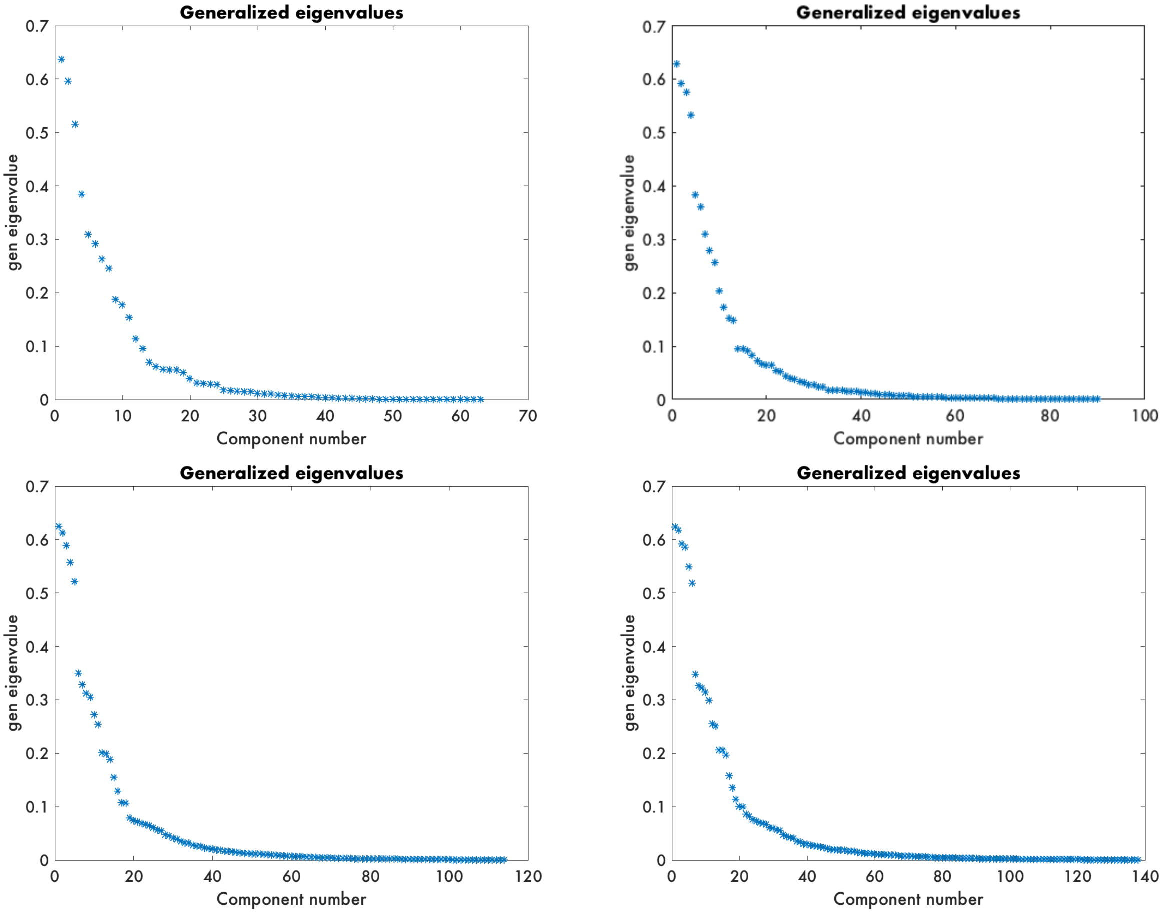


Figure 21: Fisher PCA: eigenvalue spectra of the generalised eigenvalue problem (69) for exhaustive ensemble $\mathcal{M}_k$ of cgDNAloc ground state (shape) vectors for k-mer sequences. Top: $k = 3, 4$. Bottom: $k = 5, 6$. Scale invariance of the Fisher PCA results in similar ranges (10^{-7} to around 0.6) of the eigenvalues, independent of k. A gap appears systematically after the kth eigenvalue, leading to a dimensionality reduction of ensemble of vectors to $p = k$. The overall shape of the spectrum appears very stable, and has been observed not to vary significantly with the random flanking sequences generated for building $\mathcal{M}_k$.

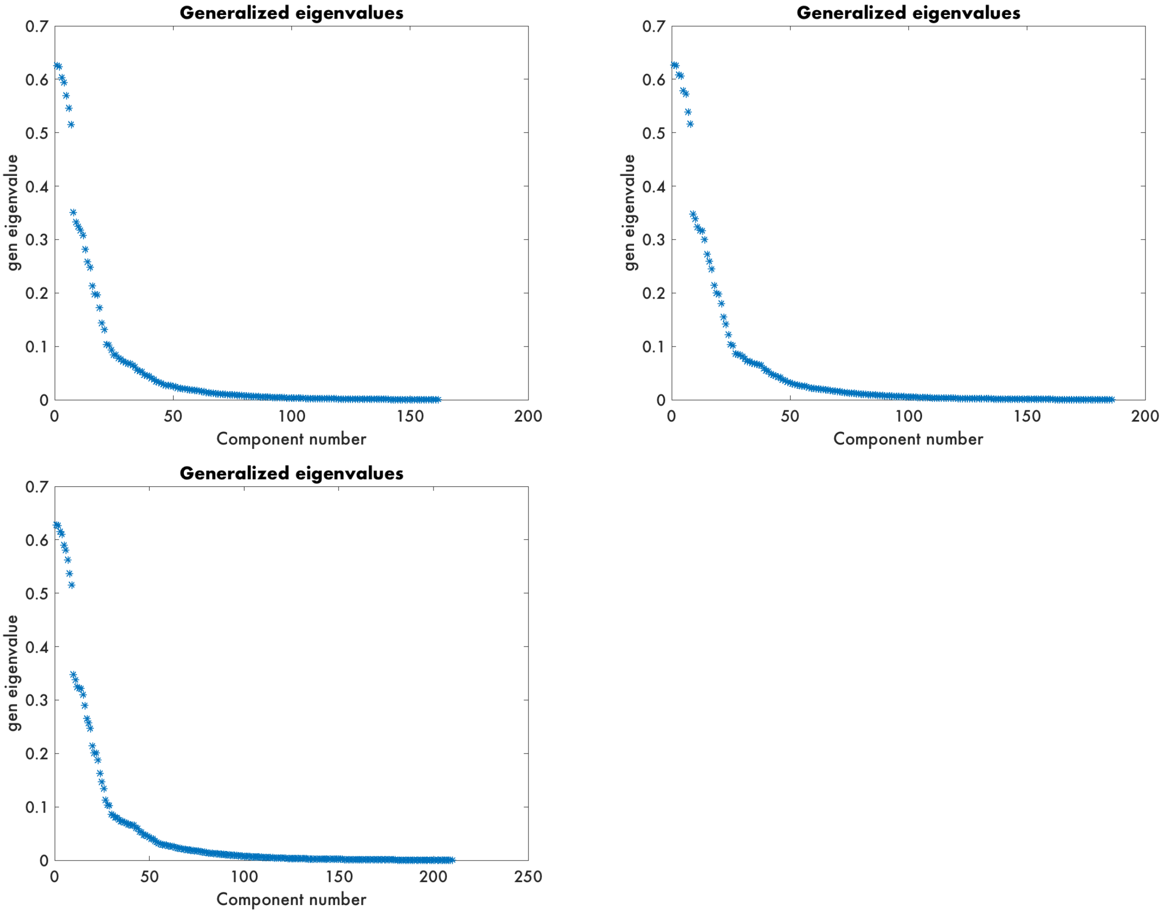


Figure 22: See caption of Figure 21, now for the cases $k = 7, 8$ (top) and $k = 9$ (bottom).

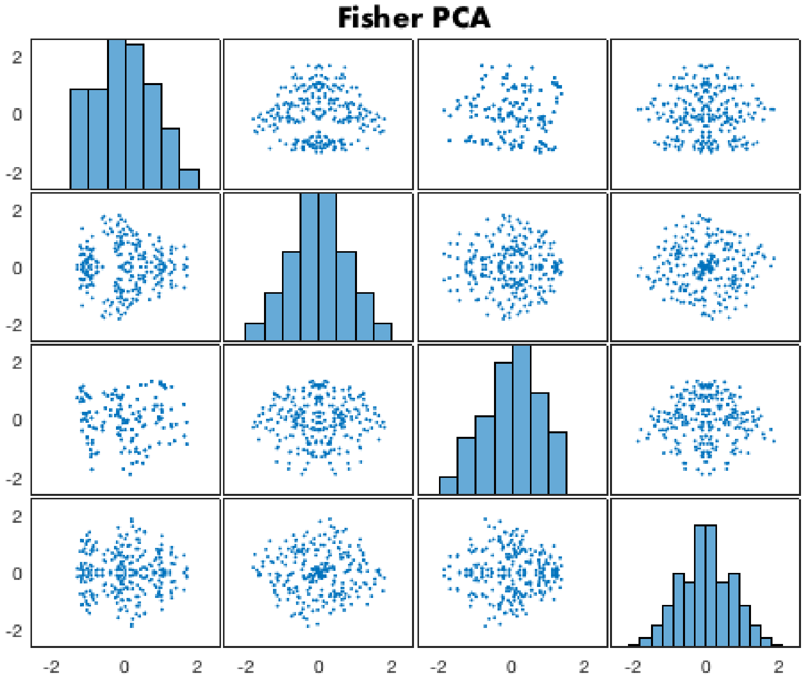


Figure 23: All possible 2-dimensional projections on the 4 principal components of Fisher PCA, from (generalised) PC1 (top and left) to (generalised) PC4 (bottom and right) for exhaustive ensemble of cgDNAloc ground state (shape) vectors for 4-mers. Here principal components are to be interpreted as generalised, in the sense that they correspond to the generalised eigenpair for the generalised eignvalue problem (69). The subplot on ith row and jth column shows the projection on the 2-dimensional subspace corresponding to PCi and PCj (thus pairs of subplots placed symmetrically with respect to the diagonal are identical up to symmetry). Subplots on the diagonal show histograms of the distribution of the shape vector one-dimensional projection onto PC1 (top left) to PC4 (bottom right). As in the PCA, symmetries can be observed that reflect the enforcement of Watson-Crick symmetry on the ensemble of shape vectors. Subplots corresponding to PC1 in particular suggest the presence of clear clusters, which turns out to be confirmed later on by applying the K-means clustering algorithm (see below).

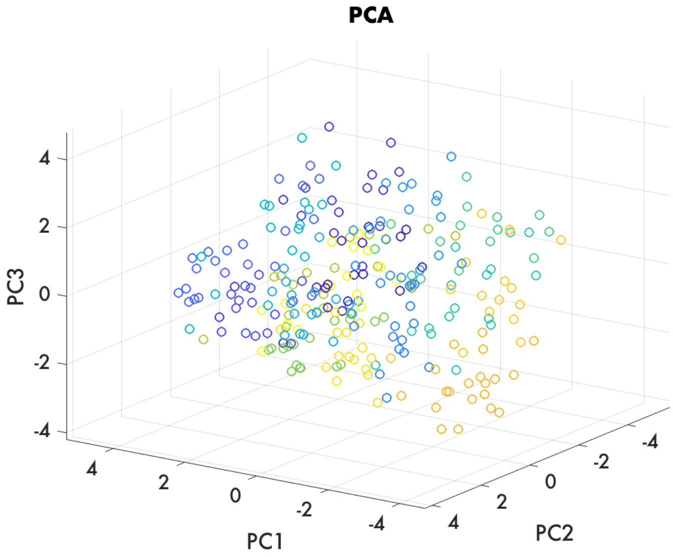


Figure 24: 3-dimensional projection on the 3 principal components PC1, PC2, and PC3 of standard PCA of exhaustive ensemble of cgDNAloc ground state (shape) vectors for 4-mers. Points with the same colour correspond to similar sequences in the purine/pyrimidine alphabet.

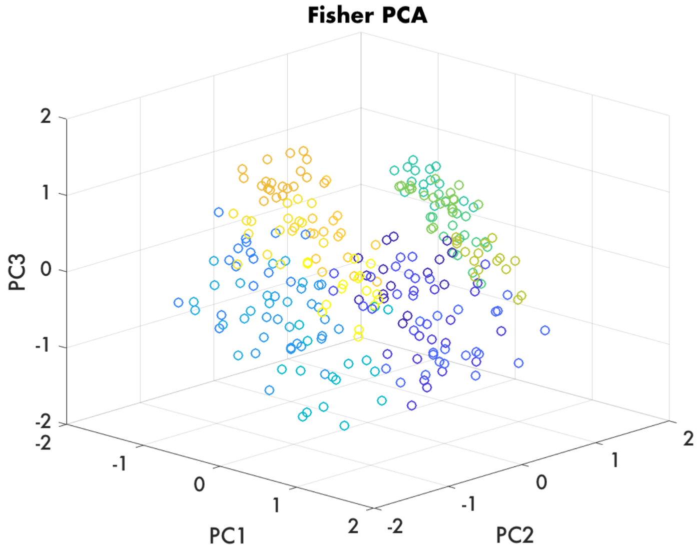


Figure 25: 3-dimensional projection on the 3 principal components PC1, PC2, and PC3 of (metric) Fisher PCA on exhaustive ensemble of cgDNAloc ground state (shape) vectors for 4-mers. Points with the same colour correspond to identical sequences in the purine/pyrimidine alphabet. There is a now a strong clustering evidence, as opposed to the standard PCA case.

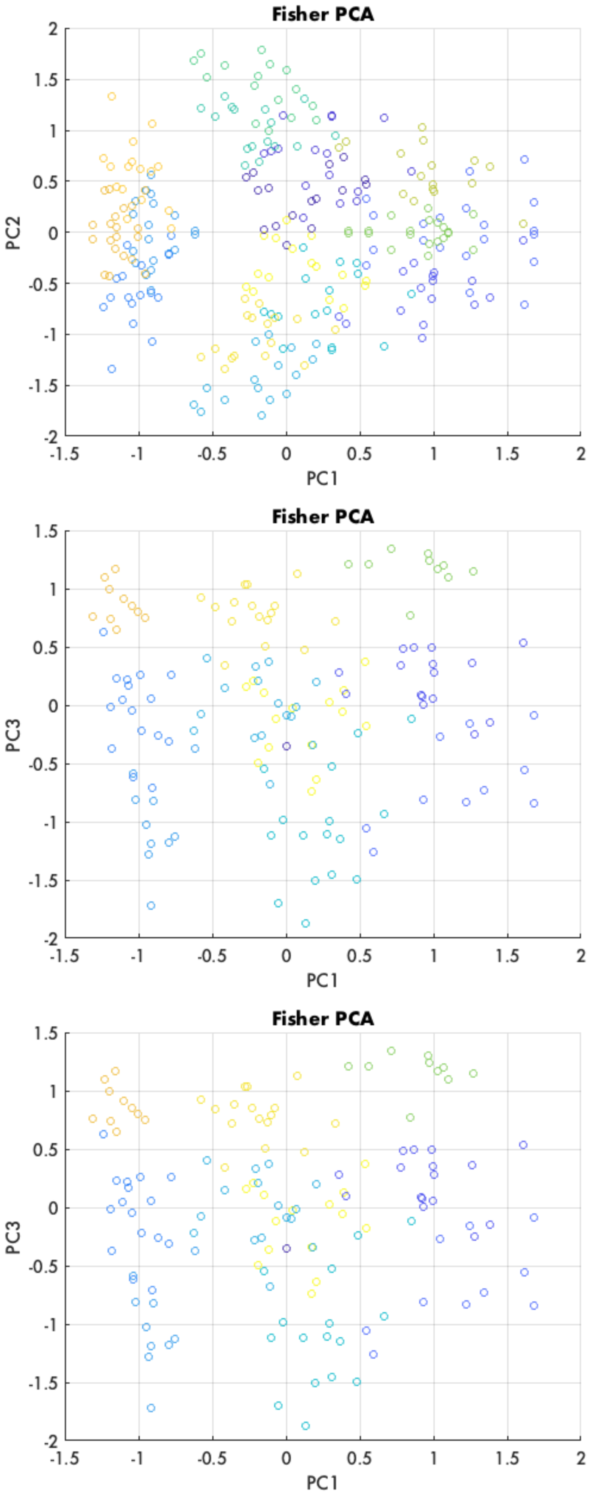


Figure 26: 2-dimensional projection on the 3 principal components PC1, PC2, and PC3 of Fisher PCA of exhaustive ensemble of cgDNAloc ground state (shape) vectors for 4-mers. Points with the same colour correspond to identical sequences in the purine/pyrimidine alphabet.

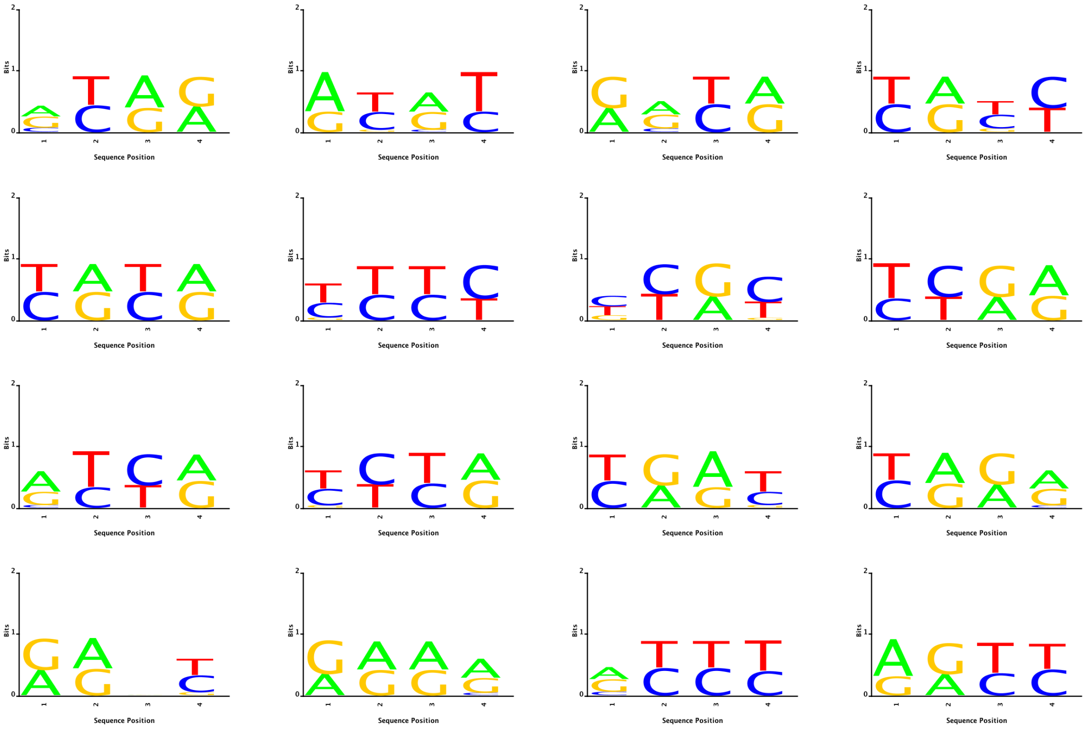


Figure 27: Sequence logos representing 4-mers of groundstate vectors in each cluster obtained by Fisher PCA projection of those vectors onto the first 4 generalised eigenvectors. Clustering is obtained via simple K-means, with $K = 2^4 = 16$ selected as optimal via the silhouette criterion. Each cluster contains approximately 2^4 vectors.

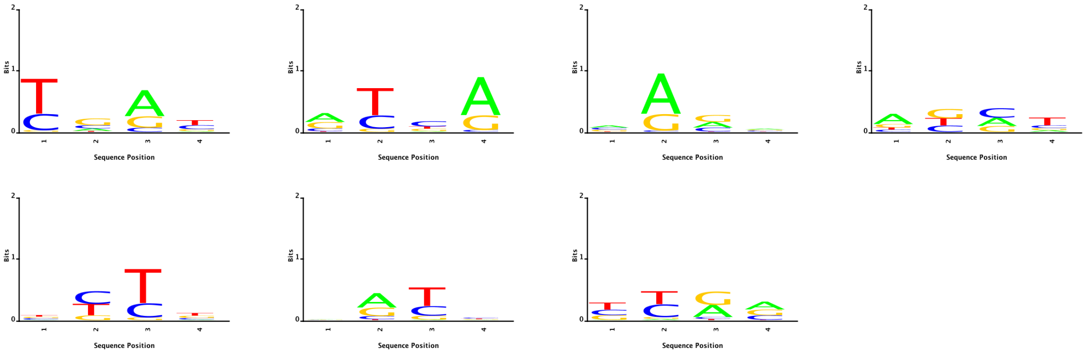


Figure 28: Sequence logos representing 4-mers of groundstate vectors in each cluster obtained by standard PCA projection of those vectors onto the first 4 eigenvector. Clustering is obtained via simple K-means, with $K = 7$ selected as optimal via the silhouette criterion. The signal is much weaker than for the analogous metric PCA clusters.

5.2 Comparison with the $MDNA$ parameter set

We finish this chapter by brief comments on the results obtained with the same procedure as in the last sections, but replicated with the most recent $MDNA$ parameter set, introduced by R. Sharma in his PhD thesis [10]. The purpose of this comparison is to test the sensitivity of the cgDNA+ model predictions to different Molecular Dynamics simulations conditions: the $MDNA$ and the $cgDNA + ps1$ parameter sets differ in the type of water model and the ions model used in the MD simulations from which the parameters are extracted. Precisely, $cgDNA + ps1$ makes use of the SPC/E [56] water model, with ions parameters introduced by Dang [57], while $MDNA$ uses the TIP3P water model [58] and Joung and Cheatham ion [59].

The main and striking difference when using the $MDNA$ parameter set appear in the generalised eigenvalue spectra of the Fisher PCA. Figure 30 and 31 show analogous plots to the ones shown in Figure 21 and 22, with which they can be compared to. As can be observed, the presence of a gap after the kth eigenvalue is now much less obvious, and the overall shape of the spectra, while staying consistent across the different lengths k, slightly differ from the ones of Figure 21 and 22, with a small second gap appearing after the lth eigenvalue, with $l = 3k + 2$. In general however, it is still true that the dimension p selected yb the largest gap criterion 67 is equal to k. Furthermore, clustering of projected shapes via K-means still lead to clusters of k-mers with identical purine/pyrimidine content (see Figure 29).

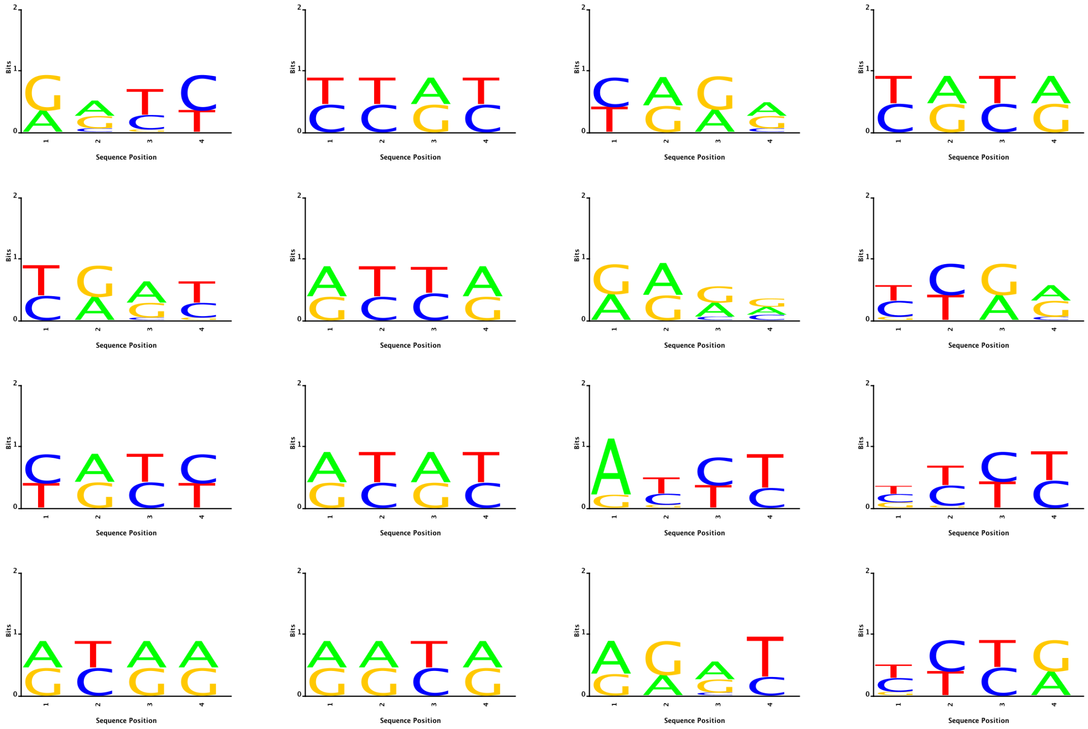


Figure 29: Sequence logos representing every sets of 4-mers associated to ground state vectors belonging to the same cluster. Clustering is obtained via the K-means algorithm after projecting down to $p = k$ principal components, generated by generalised eigenvectors. The silhouette criterion [55] selected $K = 2^4 = 16$ as the optimal number of clusters. Each cluster contains close to exactly 2^4 shape vectors. The data used is the same as the one shown in Figure 27, but with the $ps1$ parameter set replaced by $MDNA$

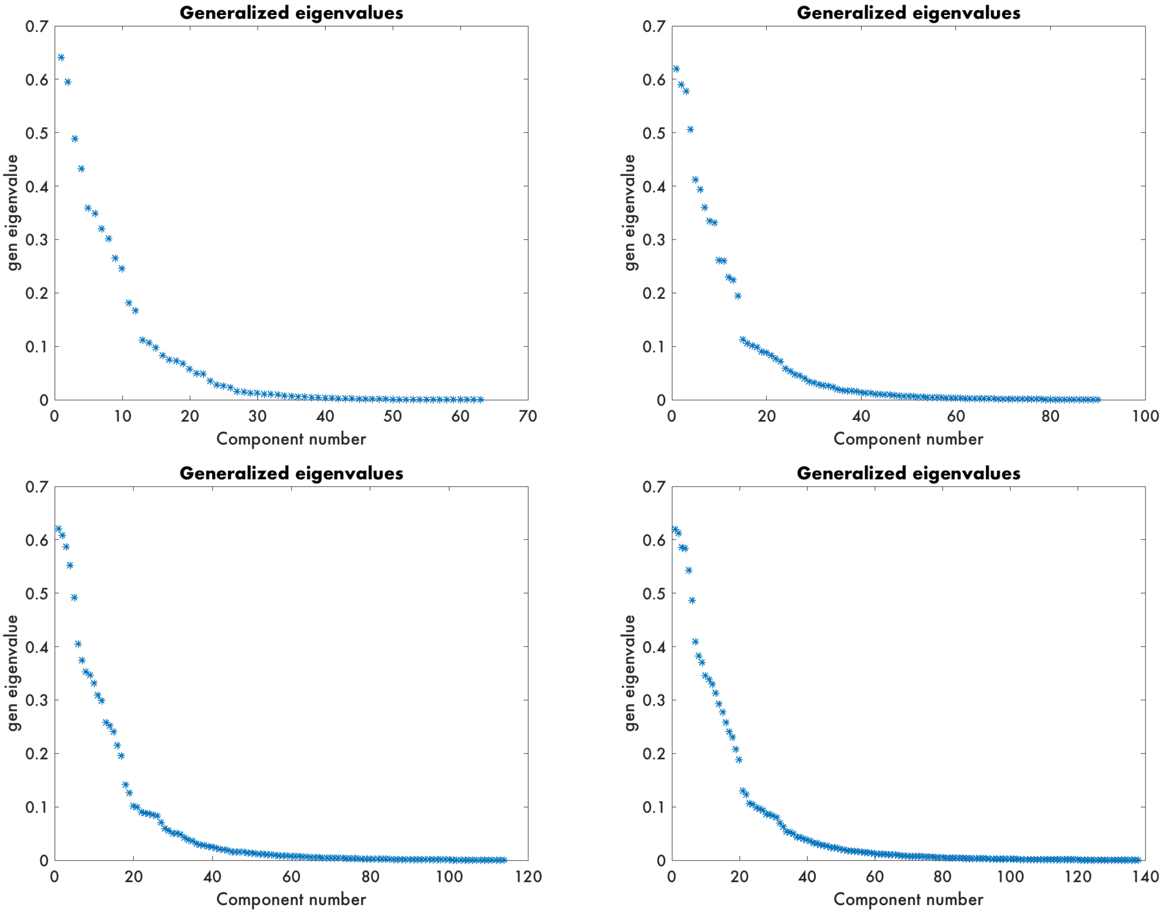


Figure 30: Fisher PCA computed on data generated with $MDNA$ parameter instead of $ps1$ parameter set: generalised eigenvalue spectrum of the generalised eigenvalue problem (69) of exhaustive ensemble $\mathcal{M}_k$ of cgDNAloc ground state (shape) vectors for k-mer sequences. Top: $k = 3, 4$. Bottom: $k = 5, 6$. Compared to the corresponding projections computed using the $ps1$ parameter set, the gap after the kth eigenvalue is now reduced, and a second small gap is present after $l = 3k + 2$ eigenvalues.

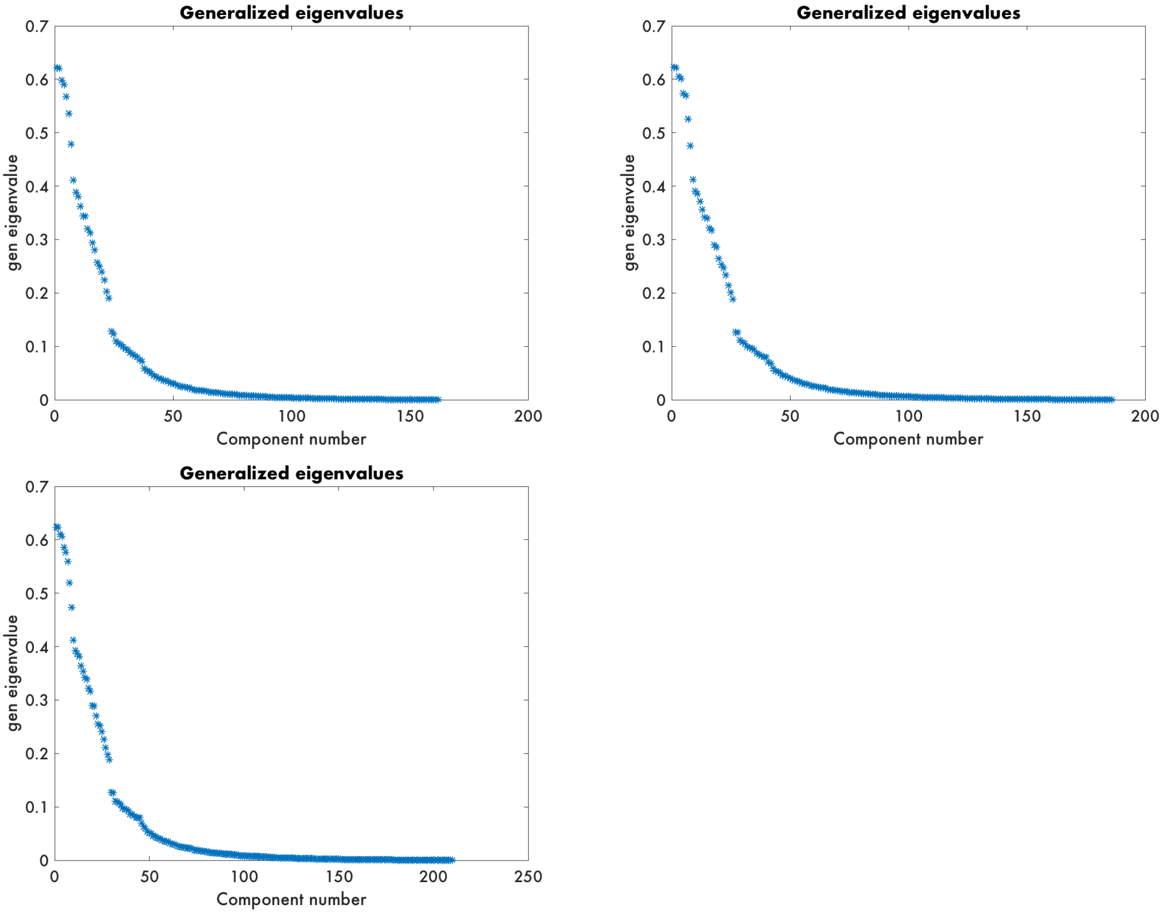


Figure 31: Fisher PCA computed on data generated with $MDNA$ parameter instead of $ps1$ parameter set: generalised eigenvalue spectrum of the generalised eigenvalue problem (69) of exhaustive ensemble $\mathcal{M}_k$ of cgDNAloc ground state (shape) vectors for k-mer sequences. Top: $k = 7, 8$. Bottom: $k = 9$. Compared to the corresponding projections computed using the $ps1$ parameter set, the gap after the kth eigenvalue is now reduced, and a second small gap is present after $l = 3k + 2$ eigenvalues.

6 Scanning genomes for outlier sequences

The cgDNA model has so far been applied to predict mechanical properties such as persistence length [60], J-factors [61], or to study minicircles [4] [12]. All of these applications involved DNA length scales of a few hundred base pairs, and correspond to the length scale for which the model was targetted with training on 10-20 bp fragments.

Although the general chromosomic behaviour inside a cell is certainly the product of a large variety of complex interacting physical and chemical mechanisms that are operating at different scales, it is widely believed that the sequence-dependent mechanical properties of the DNA molecule at the scale of tens to thousands of bp can play a role in certain problems that are relevant to biology at these longer scales.

In particular, with its sequence-dependent, non-local description of DNA shape and stiffness, the cgDNA+ model offers an interesting framework to study the following question: Can we identify in (potentially any) genomic data short sequences and/or sites that exhibit 'exceptional' mechanical properties,in some sense to be made precise, and what are those properties?

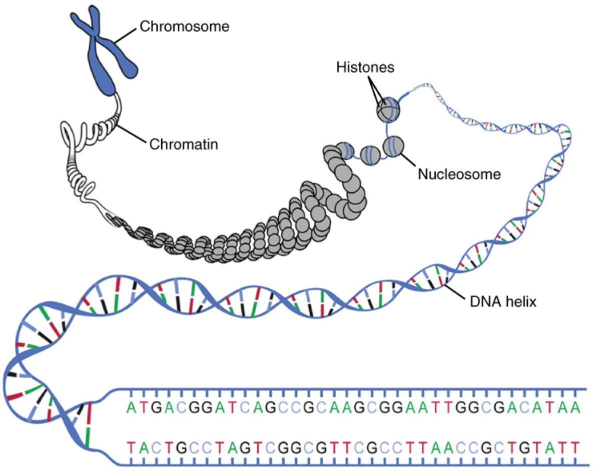

Chromosome
Histones
Chromatin
Nucleosome
DNA helix
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

Figure 32: A schematic representation of the different length scales of DNA source: https://cnx.org/contents/9TxHOD3O@4/The-Nucleus-and-DNA-Replicatio

As a first step toward the investigation of these questions with the cgDNA+ model, we have introduced some mathematical tools that allow to apply it to sequences without ends, or short sequences embedded in long ones. Namely, in Chapter 4, we have described a procedure for fast computation of marginals of Gaussian distributions whose stiffness matrix exhibits an overlapping squares sparsity pattern. Chapter 2 introduced a variety of proximity measures to compare probability distributions, which will constitute a toolbox for the numerical experiments of this chapter. Specifically, we describe the methodology we developed for this aim, as well as some first applications in relation with the two questions formulated above.

6.1 A toolbox to compare cgDNA+ distributions

We now come back to these tools mentioned above in the context of comparisons of cgDNA+ distributions, with the aim to explore the mechanical properties of DNA sequences of large datasets.

For example, if ρ_1, ρ_2 are the cgDNA+ distributions with mean μ_i and stiffness matrix $\mathbf{K}_i$, $i = 1, 2$, associated with two sequences (of the same length) $\mathrm{S}_1, \mathrm{S}_2$, then $\mathrm{KL}(\rho_1, \rho_2)$ is a number which we interpret as the proximity of their statistical mechanical structure, as described by the cgDNA+ model. As already mentioned, this quantity can be computed explicitly as

$$\mathrm{KL}(\rho_1, \rho_2) = \frac{1}{2}\left[\mathrm{tr}\left(\mathbf{K}_2\mathbf{K}_1^{-1}\right) - \ln\frac{\det\mathbf{K}_2}{\det\mathbf{K}_1} - n\right] + \frac{1}{2}(\mu_1 - \mu_2)\cdot\mathbf{K}_2(\mu_1 - \mu_2). \tag{70}$$

In this context, the KL divergence can be thought as an analogue to Shannon entropy for continuous distributions, and in particular is used as a standard way to compare them (although it should again be pointed out that KL does not satisfy the axioms of a distance). Incidentally, the integral form

$$\mathrm{KL}(\rho_1, \rho_2) = \int \rho_1 \ln\frac{\rho_1}{\rho_2}$$

of the KL divergence is formally very close to the information content $\mathrm{IC}(i)$ defined in (16), which is precisely a discrete relative Shannon entropy with respect to a uniform distribution on the possible bases $\{A, C, G, T\}$ at a position i.

It should also be mentioned that KL divergence is used extensively in the machinery of cgDNA+ as an objective function, both for truncation of the stiffness matrix into a banded version (to satisfy the nearest neighbour assumption), and for parameter optimisation [44]. In our setting, it will serve to compare a given sequence (e.g.a putative binding site) to a reference distribution, the form of which could vary depending on the precise problem at hand.

For the purpose of comparison, in what follows we will make use of various versions of KL, namely: $\mathrm{KL}(\rho_1, \rho_2)$, $\mathrm{KL}(\rho_2, \rho_1)$, $\mathrm{KL}^{sym}(\rho_1, \rho_2)$, $\mathrm{MH}^{sym}(\rho_1, \rho_2)$ where $\rho_i = \rho(\cdot; \mu_i, \mathbf{K}_i)$. Note that all the quantities are normalised by number of degrees of freedom, i.e. the dimension n of the underlying space $\mathbb{R}^n$ of ρ_i (see equation (66) in Chapter 5).

6.2 Typical and atypical sequences, from tens to hundreds of base pairs

Here we address the question of detecting sequences with ‘exceptional mechanical properties’, in the following sense: let $\mathcal{S}' = \{\mathrm{S}'_i\}$ be an ensemble of sequences and $\mathcal{S} = \{\mathrm{S}_i\}$ an ensemble of subsequences of $\mathcal{S}'$, that is we have $\mathrm{S}'_i \subset \mathrm{S}_i$ for each i. Then this defines an ensemble of $\{\rho_i = \rho_{loc}(\cdot; \mathrm{S}_i, \mathrm{S}'_i)\}$ of marginal cgDNA+ distributions, as introduced in Chapter 4. In turn, the *average* cgDNA+ distribution ρ_{av} of the ρ_i is defined as in Chapter 2. Given any proximity measure d between probability distributions, we consider the *outlier* sequences S_j satisfying

$$d_j = d(\rho_j, \rho_{av}) \geq t,$$

where t is a threshold to be specified. In practice, we will express the parameter t as a multiple of the empirical standard deviation σ of the 1-dimensional distribution of the d_j as the sequence varies:

$$t = p\,\sigma,$$

where p is an integer.

Note that in the spirit of Property 5, all these proximity measures d_j are normalised by number of degrees of freedom. This choice will also be justified a posteriori by the observation that the observed values of d_j ranges in practice over similar values for different dimensions of cgDNA+ pdfs (induced by different choices of sequence lengths).

The generic procedure above can be applied to *scan* long (e.g. genomic) sequences in search for outliers sites. These sites are of fixed length l, a free parameter that typically we take to range from tens to hundreds of bps. Longer sequences are not considered, as they go beyond the scales for which the cgDNA+ model was envisioned where sequence dependent elasticity is expected to play a significant role.

Hence, given a sequence S, e.g. a chromosome , the ensemble $\{\mathrm{S}_i'\}$ is extracted via a sliding window approach, with window length $l' = l + 2f$, where f is the length of the regions flanking S_i inside S_i'. See Fig 33 for a schematic representation of the procedure.

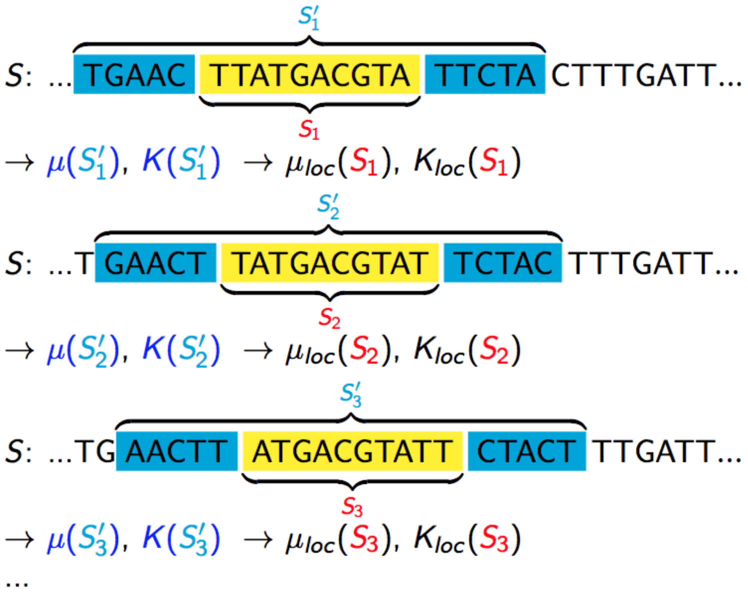

S_1'
S: ...TGAAC TTATGACGTA TTCTA CTTTGATT...
S_1
$\to \mu(S_1'), K(S_1') \to \mu_{loc}(S_1), K_{loc}(S_1)$
S_2'
S: ...T GAACT TATGACGTAT TCTAC TTTGATT...
S_2
$\to \mu(S_2'), K(S_2') \to \mu_{loc}(S_2), K_{loc}(S_2)$
S_3'
S: ...TG AACTT ATGACGTATT CTACT TTGATT...
S_3
$\to \mu(S_3'), K(S_3') \to \mu_{loc}(S_3), K_{loc}(S_3)$
...

Figure 33: Illustration of the sliding-window scanning of a DNA sequence S with cgDNAloc. Here cgDNA+ ground states and stiffness matrices $\mu(\mathrm{S}_i'), \mathbf{K}(\mathrm{S}_i')$ are built for all sites $\mathrm{S}_i' \subset \mathrm{S}$ of length 20bp. Then, marginal cgDNAloc parameters $\mu_{loc}(\mathrm{S}_i), \mathbf{K}_{loc}(\mathrm{S}_i)$ are extracted, with $\mathrm{S}_i \subset \mathrm{S}_i'$ a core site of length 10bp in the example shown here.

In *S. cerevisiae* chromosomes

In this section, we present the outcome of the genome scanning procedure applied to the genome of *S. cerevisiae* (brewer's or baker's yeast). All the computations presented were performed on chromosomes I to VIII, with essentially the same outcome. Thus, for brevity, we only present and discuss the case of chr I. This chromosome has a length of 230,208 bp. Here we show and discuss the signals d_j for the cases where d is given by one of the four quantities $\mathrm{KL}(\rho_j, \rho_{av})$, $\mathrm{KL}(\rho_{av}, \rho_j)$, $\mathrm{KL}^{sym}(\rho_j, \rho_{av})$, $\mathrm{MH}^{sym}(\rho_j, \rho_{av})$, that we all assume to be normalised by number of degrees of freedom equal to $24l - 6$ with different choices of window size $l =$ 10, 20 or 50 bp, and a fixed flanking length of 5 bp. As will be shown in what follows, the properties of the signals d_j are remarkably insensitive to the choice of l. As a matter of fact, these properties are also valid for l = 11, 100, or even 147 bp (the length of DNA wrapped around a nucleosome - see Figure 46.

Figure 34 and 35 show all the d_j signals along chr I for a window size of l =10bp, together with a histogram of their cumulative distributions. A first observation to be made is that the *range* of all the signals d_j is much higher than with the reference scale defined for KL^{sym} in Chapter 5. This fact reflects the high sequence dependence of DNA mechanics, as modeled by the cgDNA+ model. The case with the least variation, and also the lowest values of d_j is when $d_j = \mathrm{KL}_{(}\,\rho_j, \rho_{av})$. This might be related to the choice of ρ_{av}, which by construction minimizes the sum of the d_j's, but a complete explanation of this difference remains to be elucidated.

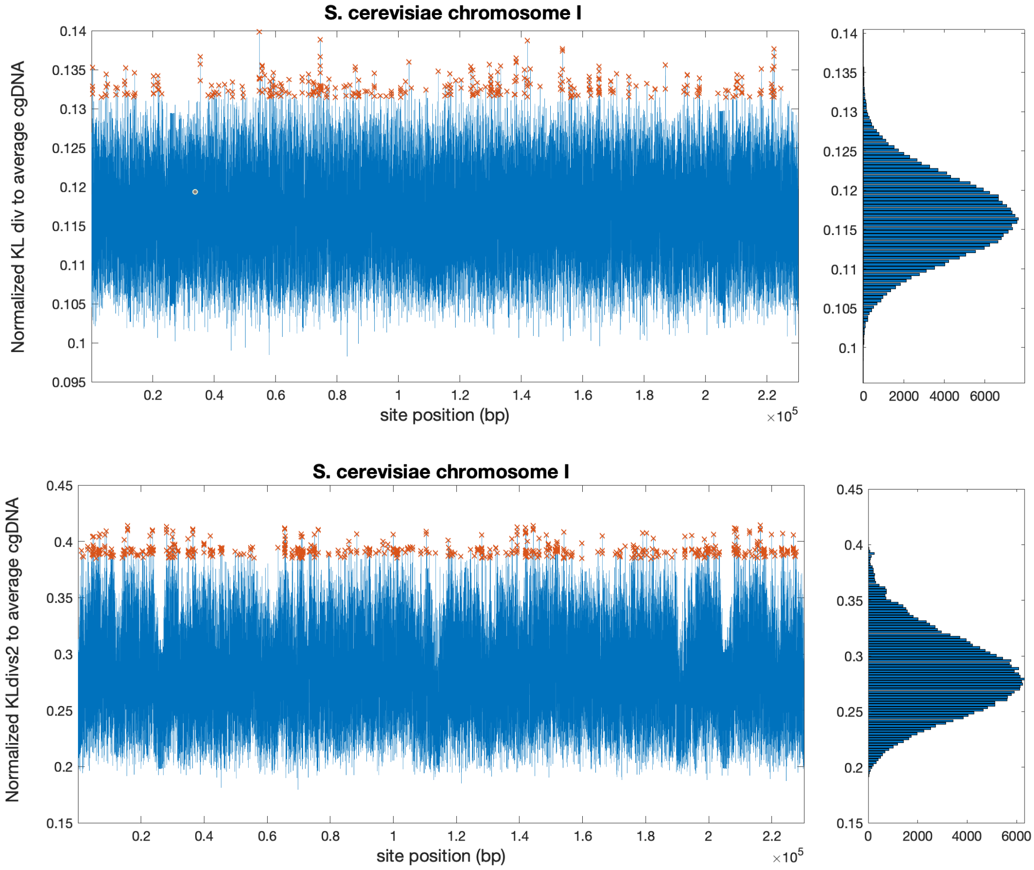


Figure 34: Chromosome-wide KL-divergence (with both order of the arguments) to cgDNA+ average Gaussian, for 10 bp sliding window in *S.cerevisiae* chr I. Top shows values of $\mathrm{KL}(\rho_j, \rho_{av})$, while bottom shows values of $\mathrm{KL}(\rho_{av}, \rho_j)$. Outliers are marked with red crosses. Note the very different scales between the two signals, reflecting the fundamental non-symmetric nature of the Kullback-Leibler divergence, and directly related to the definition of ρ_{av} as a minimizer of $\mathrm{KL}(\rho_j, \rho_{av})$. Qualitatively, $\mathrm{KL}(\rho_j, \rho_{av})$ (top) also appears more homogeneous along chrI, and its cumulative distribution is closer to a smooth normal distribution. In contrast, $\mathrm{KL}(\rho_{av}, \rho_j)$ (bottom) shows regions with only low values, and its cumulative distribution is less regular in the high value, with a very small mode around the higher end of the distribution, suggesting the presence of a population a sequences with exceptionally far from average statistical mechanical properties.

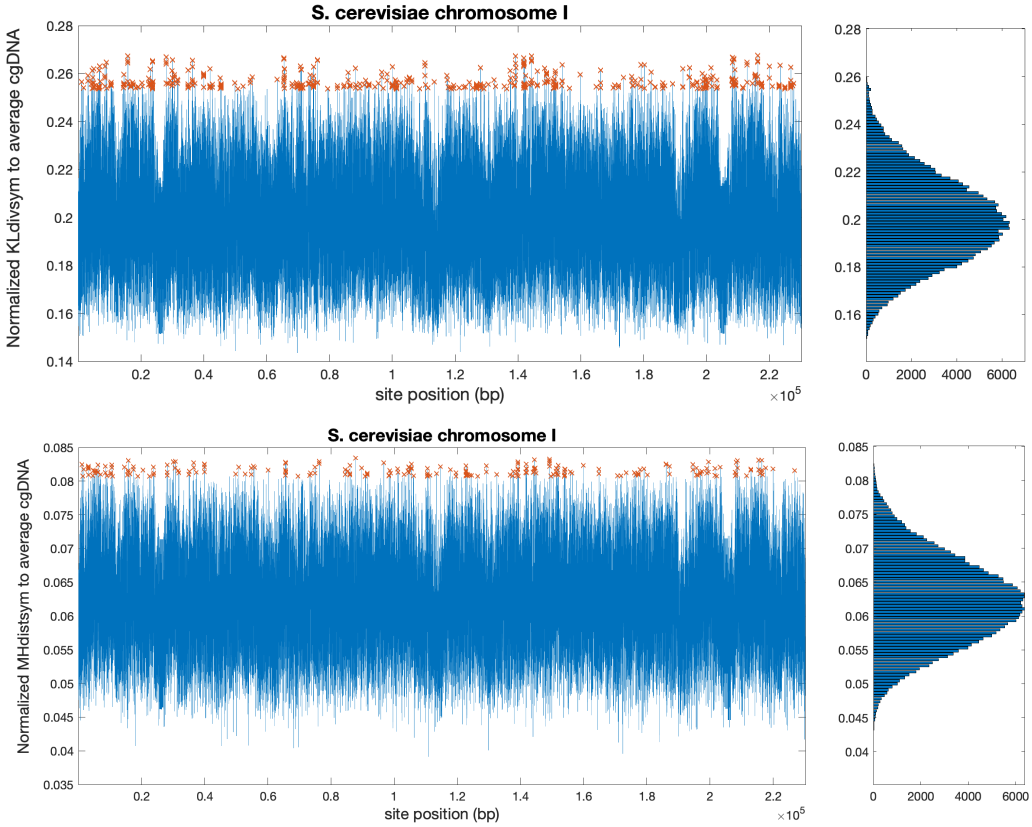


Figure 35: Chromosome-wide KL-divergence symmetrised (top), and Mahalanobis distance symmetrised (bottom) to cgDNA+ average signal, for 10 bps sites in *S. cerevisiae* chr I. Outliers are marked with red crosses. Notice the difference in scales and ranges of the two signals (recall that KL^{sym} decomposes as the sum of MH^{sym} plus an extra term involving only stiffness matrices). Nevertheless, the signals appear to follow the same patterns, particularly in the regions with no or few high values. This observation tends to indicate that it is in fact the stiffness matrix $\mathbf{K}_i$ that has the most impact on $\mathrm{MH}^{sym}(\rho_{av}, \rho_j)$ - because if it where the ground state vector μ_i, the two signals would be uncorrelated.

A second observation to be made is that for all the signals d_j, the minimum values across the entire chromosome is far from zero, which means that the distributions ρ_i are all far from the average distribution ρ_{av}. Moreover, the typical minimum distance d_j exhibits an order magnitude comparable to typical range of the distribution of the d_j (see for example the top signal in Figure 35). In other words, no particular sequence appears close to average DNA (as we defined it) in terms of its statistical mechanical properties.

Figure 36 shows sequence logos obtained from the procedure described in the previous section, with $t = 3$, a window length of 10 bp, with marginalisation over 5bp flanking sequences, and the various choices of d. Most of these logos (with the exception of the first one) indicate that the selected outlier sequences exhibit a relatively very a high A/T content. This preference is stable across the various possible positions of the base pair inside the sequence, with little more variety in bp frequency at both the first and last position of the logos. This variation could be attributed to the previously discussed non-local sequence dependence of the cgDNA+ model, with bp content of both flanking sequences influencing the score d_j of the sequence. The precise mechanism remains unclear however. Together with the observation made in Figure 35 on the role of the stiffness matrix $\mathbf{K}_i$, the fact that the outliers for the signal d_j for the symmetrised Mahalanobis distance also yields a strong preference for A/T base pair suggests that this preference is influenced by the stiffness $\mathbf{K}_i$. It is still unclear to us why the first signal $d_j = \mathrm{KL}(\rho_j, \rho_{av})$ does not yield outlier sequences with particular base pair preferences. The conjecture is that it is related to the nature of the average ρ_{av}, which by definition minimises the sum of the $\mathrm{KL}(\rho_j, \rho_{av})$.

Further analysis of these outlier sequences can be carried out using dimer logos (see Figure 37). With again the exception of the case when $d_j = \mathrm{KL}(\rho_j, \rho_{av})$, all ensembles of outlier sequences obtained exhibit a high content of AA and TT dimers. Again, this preference is stable across the various possible positions of the dimers inside the sequence. This fact implies that these sequences are characterized by runs of A or runs of T. Although the proportions of A and T are not exactly identical, as can be observed from the logos, they are very close. That fact is not surprising, in regard to the built-in Watson-Crick symmetry of the parameters of the cgDNA+ Gaussian distributions (see Chapter 1).

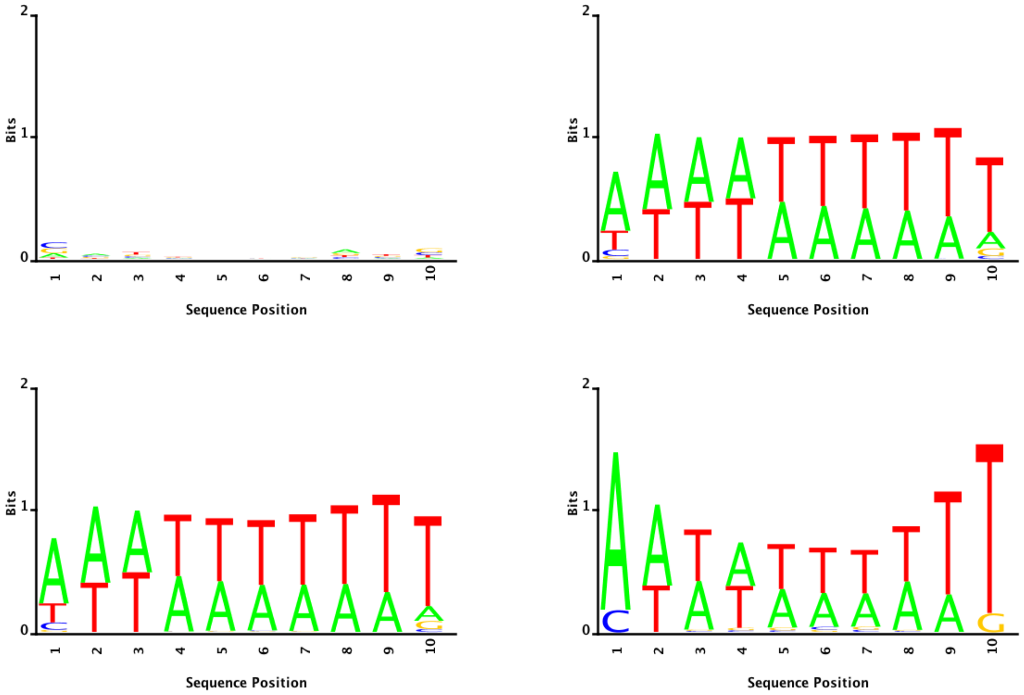


Figure 36: Sequence logos built from the high outliers (3 standard deviations above the mean) in chromosome-wide $\mathrm{KL}(\rho_j, \rho_{av})$, $\mathrm{KL}(\rho_{av}, \rho_j)$, $\mathrm{KL}^{sym}(\rho_j, \rho_{av})$, and $\mathrm{MH}^{sym}(\rho_j, \rho_{av})$ to cgDNA+ average signal, for 10 bps sites in *S. cerevisiae* chr I. All but the first logo show a very strong preference for A/T base pair content. That preference appears overall independent from the position of the base pair inside the sequences, with some slight variation at the first and last (or two first and two last in the case of the Mahalanobis distance) position.

As no value of the d_j is close to zero (compared to the range of their distribution), it is also meaningful and of interest to look at sequences S_j satisfying

$$d_j = d(\rho_j, \rho_{av}) \leq t,$$

which we will refer to as *low outliers*.

Logos of low outlier sequences for a window size of 10 bp are represented in Figure 38. Despite being less striking than in the case of high outliers, with a lower information content reflecting a larger variety in monomer frequencies, one can still observe a preference for sequences with high C/G content.

This preference is also present on the corresponding dimer logos (see Figure 39) which shows that low outliers are rich in CC and GG dimers.

With the purpose of studying the effect of increasing the window size l in the scanning procedure, we also compute analogous d_j signals along *S. cerevisiae* with $l = 20$bp and $l = 50$bp. Analogous d_j signals to the ones on Figure 34 and 35 are shown on Figure 40, 41 (20bp) Figure 43 and 44 (50bp). In both cases it is quite remarkable that the scale of the different d_j are very similar to the case of 10bp. This fact a posteriori motivates the choice of normalising all the d proximity measures by number of degrees of freedom. Furthermore, all properties previously highlighted in the case $l = 10$bp still hold for these larger window sizes: close to normal cumulative distribution, qualitative similarity and matching areas of fewer sequences with high d_j values, with the exception of the case $d_j = \mathrm{KL}(\rho_j, \rho_{av})$. The main differences appear to be a relative decrease in noisiness of the signals, which is presumably due only to a smoothing effect related to the ratio between the window size and the shift of 1bp that is used in the scanning. This also explains why outlier sequences also appear increasingly more clustered together as l increases.

Sequence logos of ensembles of high outlier sequences for the cases 20bp and 50bp are shown on Figures 42 and 45. As in the case $l = 10$bp, a rather homogenous preference for A/T base pair can be observed, even though the amplitude of the information content in these logos decrease slightly as the window size l increases. This preference is even still apparent for the case [3] when $l = 147$bp,which is the length of DNA wrapped around a nucleosome - see Figure 46.

---

[3] Note that this signal was obtained with the prior cgDNA model, with which the experiments presented in this chapter were originally conducted

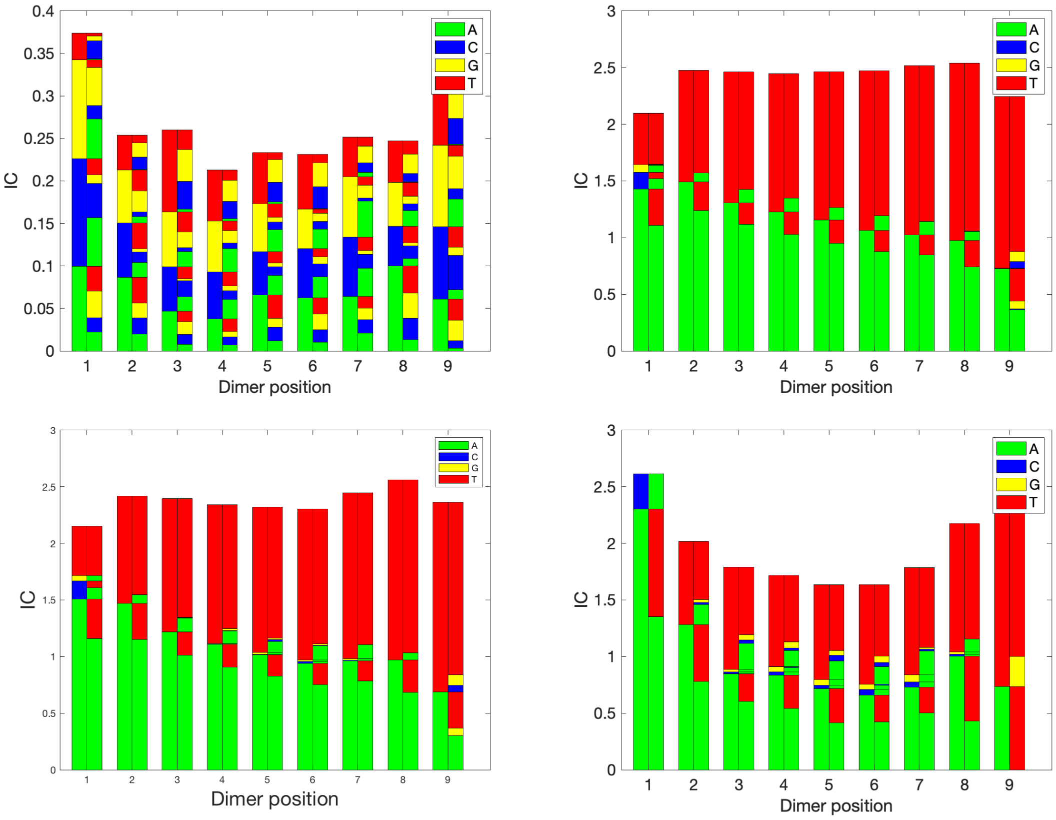


Figure 37: Dimer logos built from the high outliers (3 standard deviations above the mean) in chromosome-wide $\mathrm{KL}(\rho_j, \rho_{av})$, $\mathrm{KL}(\rho_{av}, \rho_j)$, $\mathrm{KL}^{sym}(\rho_j, \rho_{av})$, and $\mathrm{MH}^{sym}(\rho_j, \rho_{av})$ to cgDNA+ average signal, for 10 bps sites in *S. cerevisiae* chr I. With the exception of the one on the first panel, each of these logos shows a very strong preference for AA and TT dimer content, in a way that is mostly independent from the position of the dimer inside the sequence, meaning that the corresponding outlier sequences are mostly formed by runs of A and runs of T.

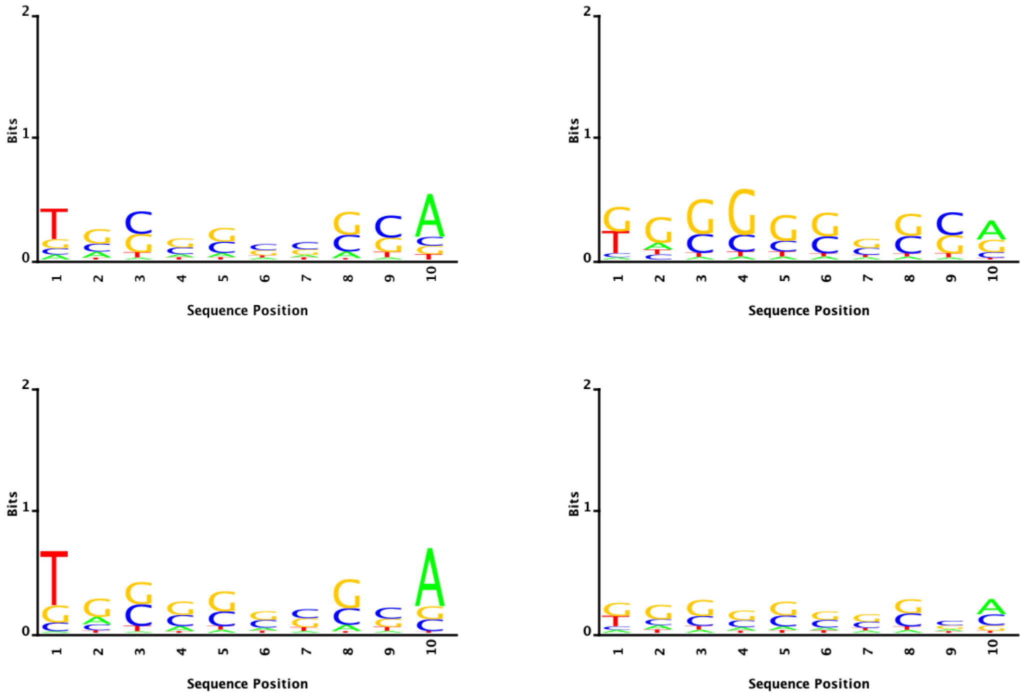


Figure 38: Sequence logos built from the low outliers (3 standard deviations above the mean) in chromosome-wide $\mathrm{KL}(\rho_j, \rho_{av})$, $\mathrm{KL}(\rho_{av}, \rho_j)$, $\mathrm{KL}^{sym}(\rho_j, \rho_{av})$, and $\mathrm{MH}^{sym}(\rho_j, \rho_{av})$ to cgDNA+ average signal, for 10 bps sites in *S. cerevisiae* chr I. Despite being slightly weaker than in the case of high outliers, these logos show some preference for G and C content, particularly on the logo corresponding to the $\mathrm{KL}(\rho_{av}, \rho_j)$ signal (top right).

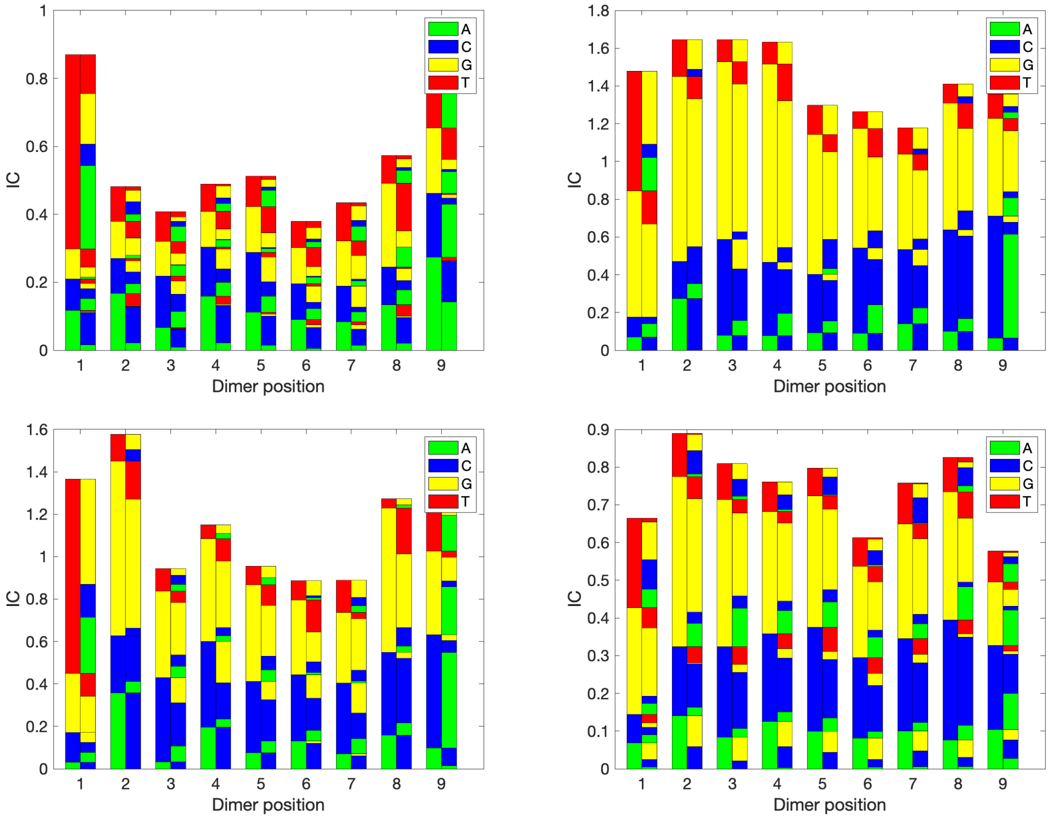


Figure 39: Dimer logos built from the low outliers (3 standard deviations below the mean) in chromosome-wide $\mathrm{KL}(\rho_j, \rho_{av})$, $\mathrm{KL}(\rho_{av}, \rho_j)$, $\mathrm{KL}^{sym}(\rho_j, \rho_{av})$, and $\mathrm{MH}^{sym}(\rho_j, \rho_{av})$ to cgDNA+ average signal, for 10 bps sites in *S. cerevisiae* chr I. As was observed for the case of high outliers, the C/G preference observed in the monomer sequence logos turns out to be a preference in CC/GG dimers. This feature is particularly striking in the case $d_j = \mathrm{KL}(\rho_{av}, \rho_j)$ (top right) and $\mathrm{MH}^{sym}(\rho_j, \rho_{av})$ (bottom right).

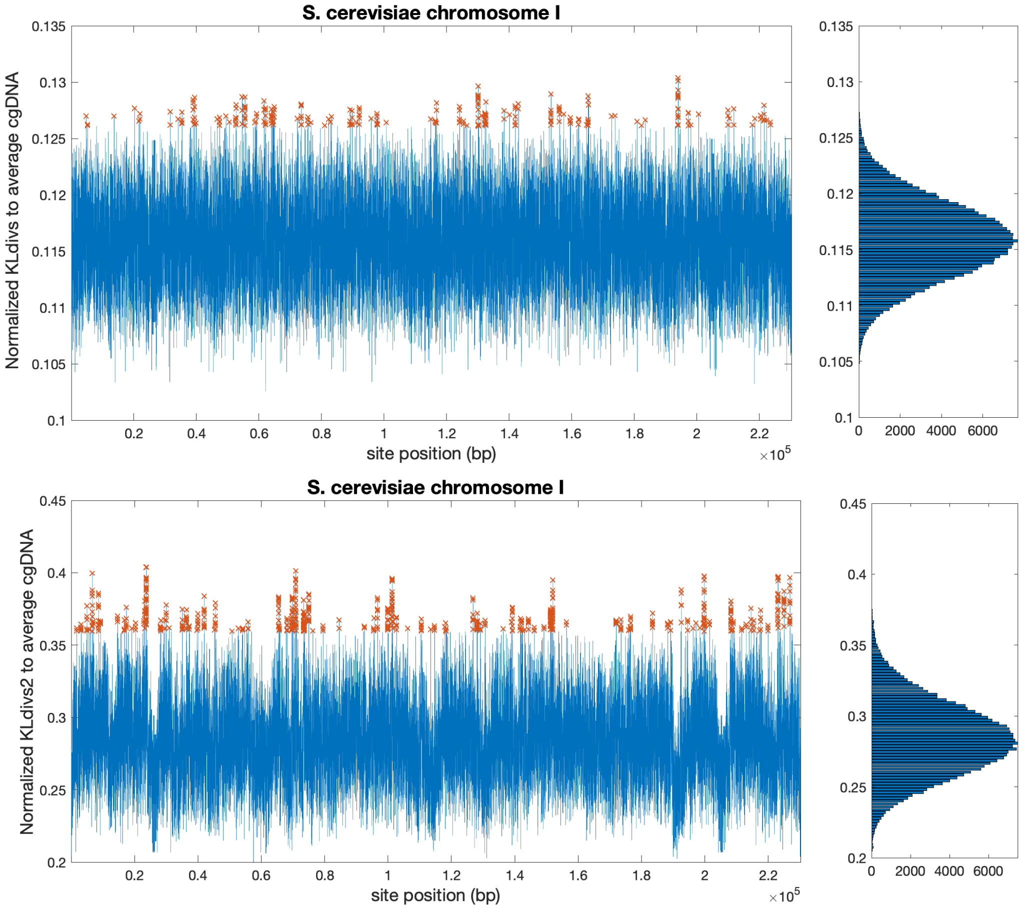


Figure 40: Chromosome-wide KL-divergence (with both order of the arguments) to cgDNA+ average signal, for 20 bp sliding window in *S.cerevisiae* chr I. Top shows values of $\mathrm{KL}(\rho_j, \rho_{av})$, while bottom shows values of $\mathrm{KL}(\rho_{av}, \rho_j)$. Outliers are marked with red crosses. We observe again a lower scale and smaller range for the values $d_j = \mathrm{KL}(\rho_j, \rho_{av})$, related to the definition of the average ρ_{av} as a minimiser of the sum of $\mathrm{KL}(\rho_j, \rho_{av})$. That signal also appear noisier than $d_j = \mathrm{KL}(\rho_{av}, \rho_j)$, as was observed when scanning with a window size of 10bp.

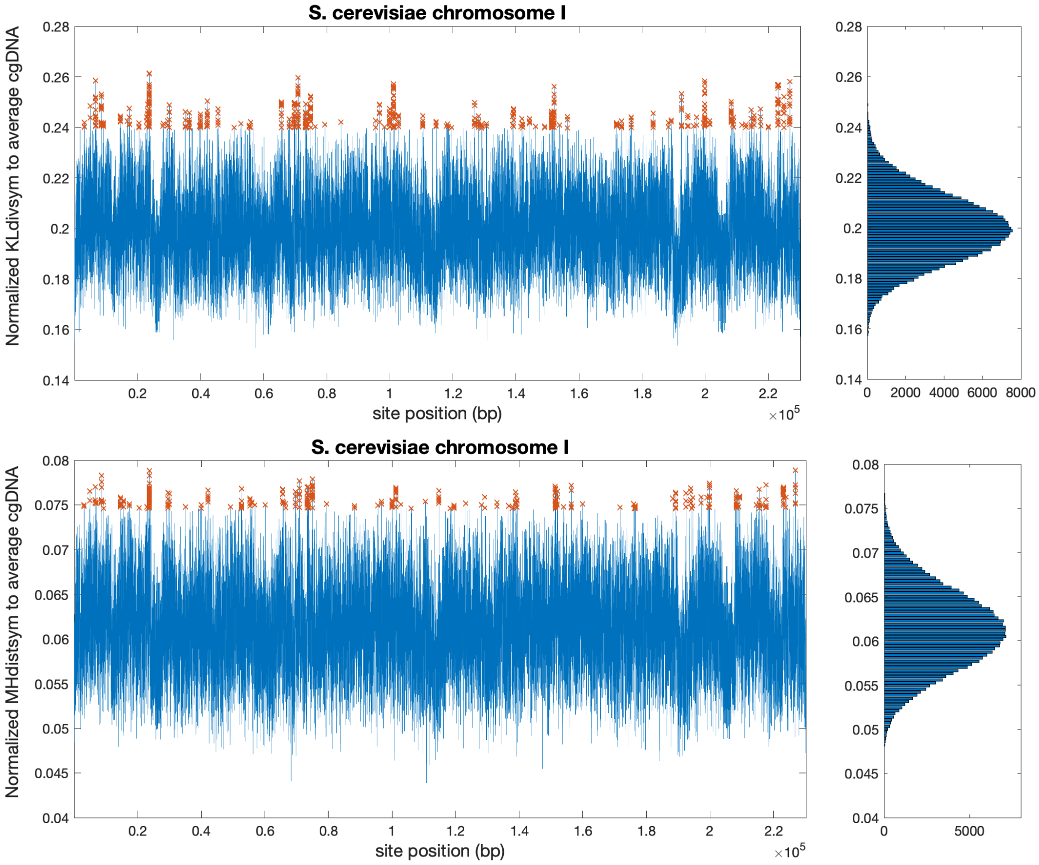


Figure 41: Chromosome-wide KL-divergence symmetrised (top), and Mahalanobis distance symmetrised (bottom) to cgDNA+ average signal, for 20 bps sites in *S. cerevisiae* chr I. Outliers are marked with red crosses. Scales of both signals highly resemble the scales of the analogous signals d_j computed with a window size of 10bp, which a posteriori makes the choice of normalisation by number of degrees of freedom pertinent for comparison. It is also the case that these two signals exhibit similar patterns - e.g., areas with low number of high values, indicating a predominant role of the stiffness matrix $\mathbf{K_i}$ in the behavior of MH_{sym}.

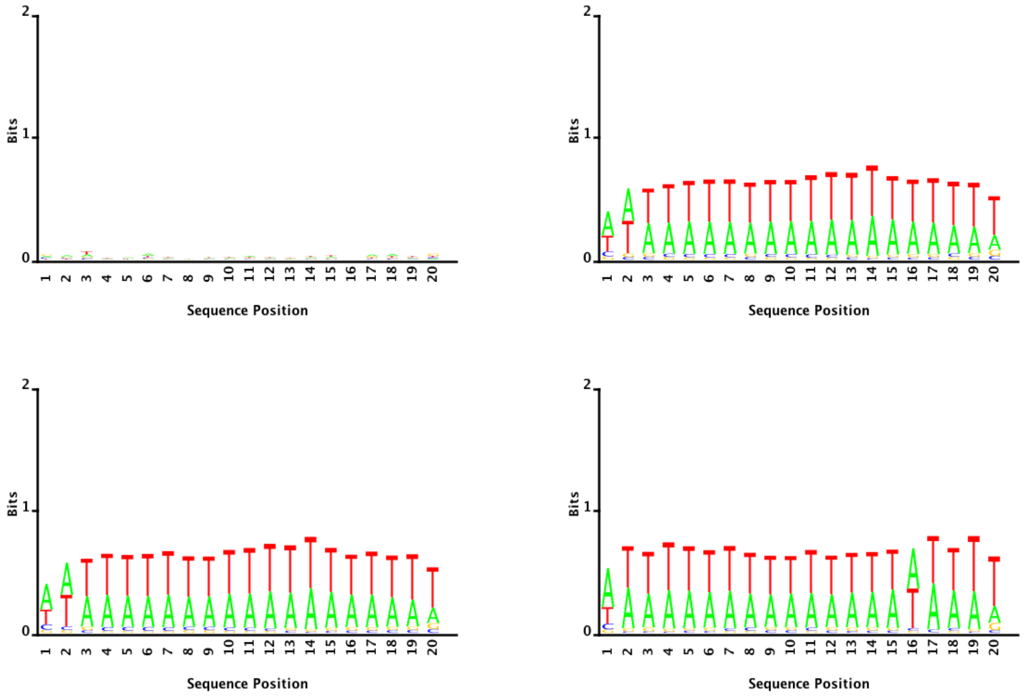


Figure 42: Sequence logos built from the high outliers (3 standard deviations above the mean) in chromosome-wide $\mathrm{KL}(\rho_j, \rho_{av})$, $\mathrm{KL}(\rho_{av}, \rho_j)$, $\mathrm{KL}^{sym}(\rho_j, \rho_{av})$, and $\mathrm{MH}^{sym}(\rho_j, \rho_{av})$ to cgDNA+ average signal, for 20 bps sites in *S.cerevisiae* chr I. Similarly to the corresponding 10 bp logos (Figure 36), all but the first logo show a very strong preference for A/T base pair content. That preference appears overall independent from the position of the base pair inside the sequences, with some slight variation at the first and last position.

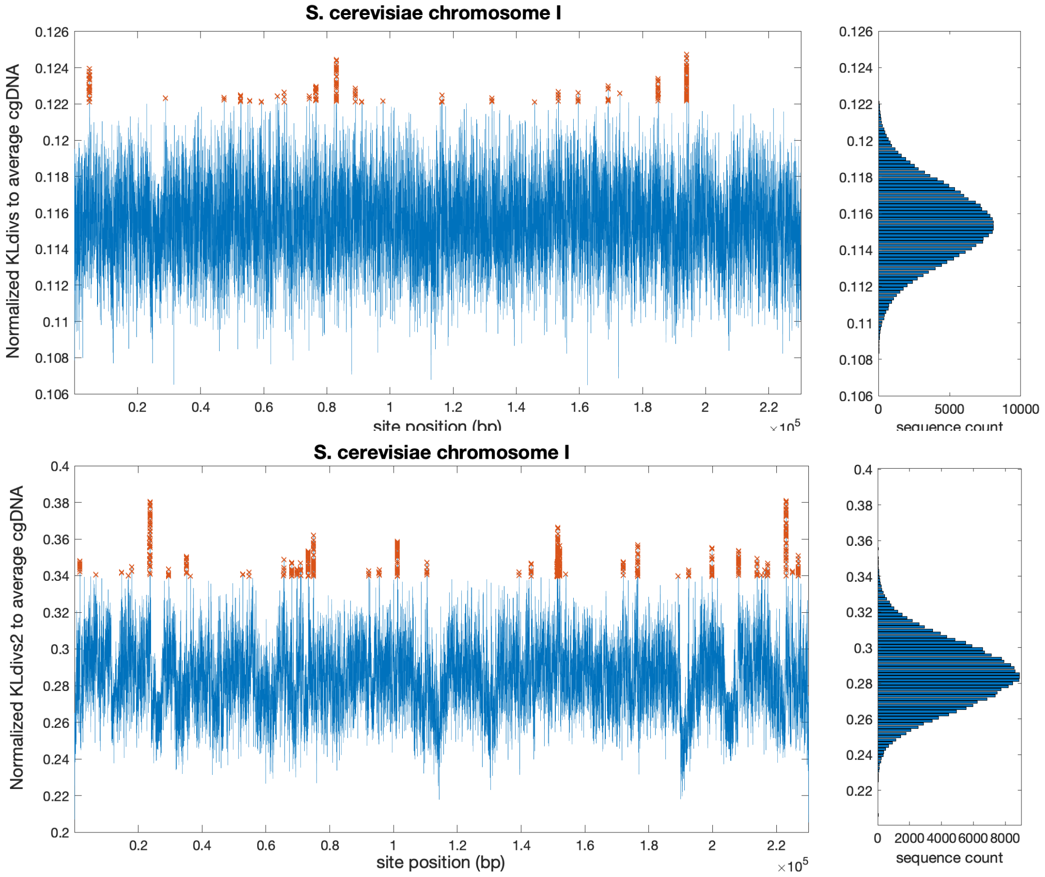


Figure 43: Chromosome-wide KL-divergence (with both order of the arguments) to cgDNA+ average signal, for 50 bp sliding window in *S.cerevisiae* chr I. Top shows values of $\mathrm{KL}(\rho_j, \rho_{av})$, while bottom shows values of $\mathrm{KL}(\rho_{av}, \rho_j)$. Note the different scales of the two signals. Outliers are marked with red crosses. With a 5-fold increase in scanning window size compared to the 10 bp case, but still scanning shift of 1 bp, a window smoothing can be observed in particular when $d_j = \mathrm{KL}(\rho_{av}, \rho_j)$ (bottom).

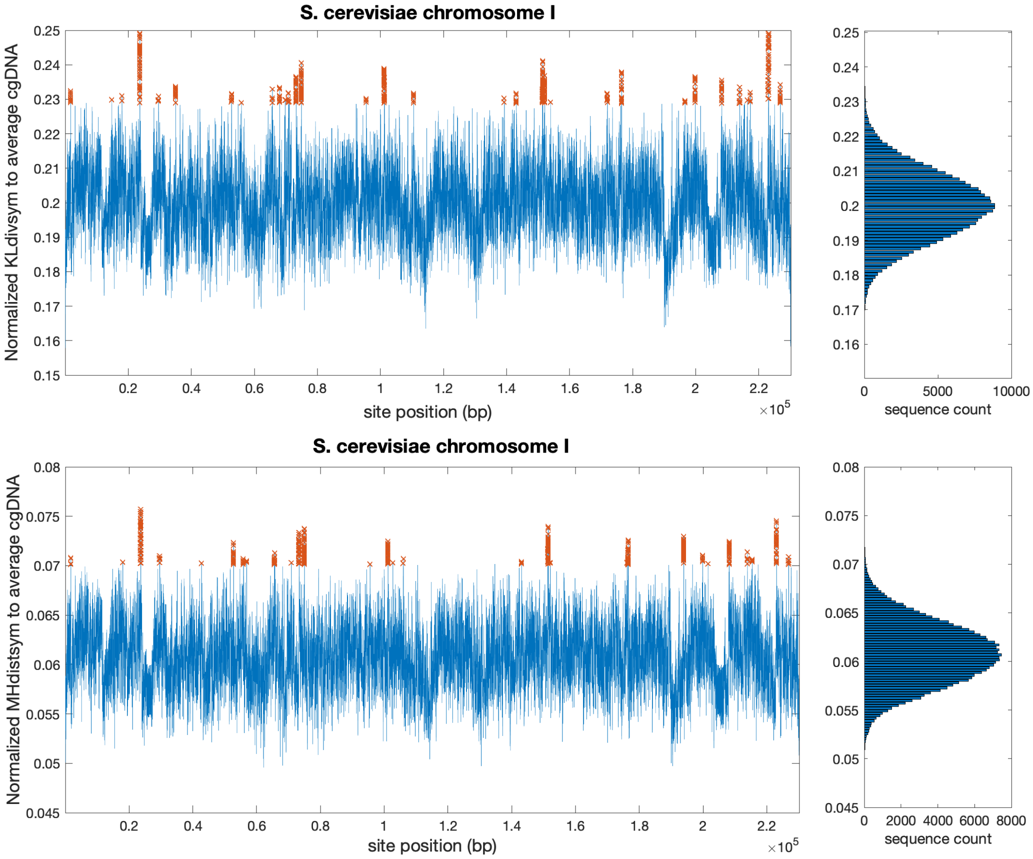


Figure 44: Chromosome-wide KL symmetrised (top), and Mahalanobis symmetrised (bottom) to cgDNA+ average signal, for 50 bps sites in *S. cerevisiae* chr I. Outliers are marked with red crosses. With a 5-fold increase in scanning window size compared to the 10 bp case, but still scanning shift of 1 bp, a window smoothing can be observed.

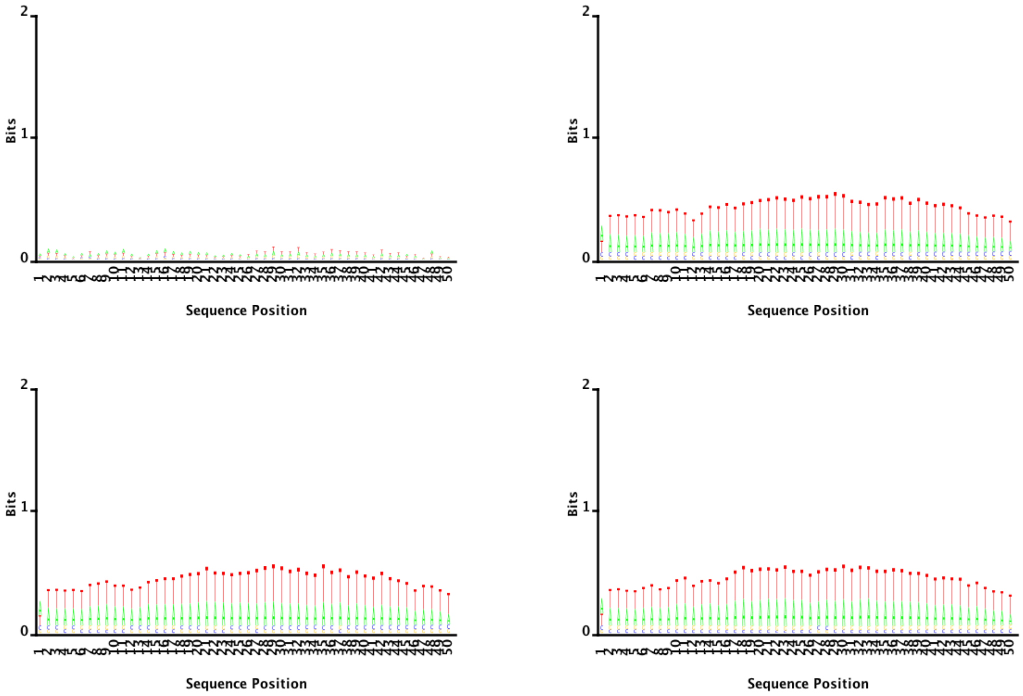


Figure 45: Sequence logos built from the high outliers (3 standard deviations above the mean) in chromosome-wide $\mathrm{KL}(\rho_j, \rho_{av})$, $\mathrm{KL}(\rho_{av}, \rho_j)$, $\mathrm{KL}^{sym}(\rho_j, \rho_{av})$, and $\mathrm{MH}^{sym}(\rho_j, \rho_{av})$ to cgDNA+ average signal, for 50bp sites in *S.cerevisiae* chr I. Similarly to the corresponding 10 bp logos (Figure 36), all but the first logo show a very strong preference for A/T base pair content. As window size increases, information content in the sequence logos of outlier decreases.

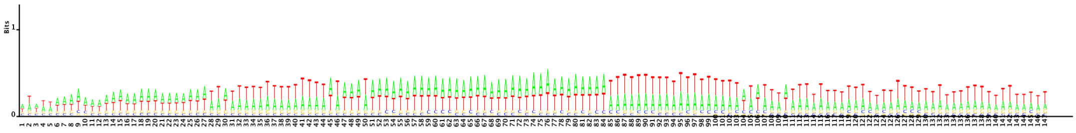


Figure 46: Sequence logos built from the high outliers (3 standard deviations above the mean) in chromosome-wide $\mathrm{KL}^{sym}(\rho_j, \rho_{av})$, to cgDNA average signal, for 147bp sites in *S. cerevisiae* chr I. A weaker but still noticeable preference for A/T content can be observed.

Conclusion and Discussion

We start by briefly summarising the main results of the thesis.

In the framework of the sequence-dependent statistical mechanics cgDNA+ coarse grain model, that was introduced in [2] building on the precursor cgDNA model [62] [8], we have developed a tool that we have named cgDNAloc, that allows scanning of genomic length-scale sequences using a sliding-window approach with short test sequence fragment. This tool permits proper account to be taken for the significant non-local effect of flanking sequence on any groundstate vector predicted by cgDNA+.

In order to be useful, the cgDNAloc algorithm has to be extremely efficient in order to be able to treat the vast number of (still rather high dimensional) marginals that need to be computed in realistic applications. The large number of sequences to be treated arise in two different ways: first because of the inherent exponential growth of the size of exhaustive ensembles of possible DNA sequences of length k, or k-mers. Secondly, for scanning genomes, the sequence data is so long that the computation of many marginals is still required just because there are so many window locations. The efficiency of the cgDNAloc algorithm is possible due to the main mathematical cornerstone of this thesis, namely: marginals of Gaussians with banded (i.e. overlapping squares sparsity pattern) inverse covariances also have banded inverse covariance matrices (with the same overlapping squares sparsity pattern), and in marginalisation the only entries that change from the sub-block of the original inverse covariance are in the first and last overlap region. This result is not obvious, and has to our knowledge only been described previously in analogous versions expressed in the language of graphical models (see [45, 47]) a literature that is unlikely to be accessible to many molecular biologists. This may be because marginalisation of Gaussians naturally involves considering sub-blocks of the covariance matrix itself, which for banded inverse covariances is still dense, albeit with nontrivial, hidden, dependencies between covariance sub-blocks. Our observation leads to our efficient algorithm for computing marginals, with direct application to specific cgDNA+ marginals implemented in cgDNAloc. We make two further remarks. First in the case where the overlaps in the block sparsity are absent, so that a block diagonal sparsity arises for the inverse covariance (and in this case also for the covariance itself), then our results reduce to the well-known, essentially trivial result for marginalising block diagonal inverse stiffness matrices. In particular the marginal stiffness matrix is itself block diagonal, with the only blocks that may not be sub-blocks of the original stiffness matrix, being the first and last ones. Second, one possible reason that our result is not widely known, is that just to formulate the hypothesis of an overlapping block sparsity pattern the order of the components in the pdf variable w must be fixed and specified. In many applications in statistics this is not such a natural hypothesis. But in our polymer chain model of DNA the ordering leading to overlapping block sparsity is very natural, and arises due to nearest neighbour interaction along the polymer.

Similarly the block sparsity pattern is only preserved in the particular marginalisation where consecutive runs of an initial and final range of variables are to be eliminated. Other marginals are also interesting. For example marginalising over the base-phosphate degrees of freedom in the cgDNA+ model leaves precisely a Gaussian in the inter and intra variables appearing in the cgDNA model. However the cgDNA+ marginal stiffness matrices for inter and intra variables are dense, albeit with rapidly decaying entries far from the diagonal. Specifically the marginal stiffnesses are not limited to the banded sparsity pattern assumed in cgDNA corresponding to locality in those degrees of freedom. This observation is probably why the cgDNA+ model predictions are considerably closer to statistics observed directly from MD simulation. It is an example where computing with a higher dimensional, finer grain, Gaussian model with a structured stiffness matrix can be better than assuming a lower dimensional model without a structured stiffness matrix. In fact an analogous observation was previously made in [4] in computing DNA persistence lengths, which only depend on the distribution of the inter variables. Persistence lengths could therefore be computed with a lower dimensional rigid base pair marginal of the cgDNA rigid base model, in which the intra variables have been eliminated. But the inter variable marginal stiffness of the banded cgDNA stiffness matrix is not itself banded, and as a consequence it was much more efficient to run simulations using the original higher dimension, finer grain, but structured Gaussian distribution on both intra and inter variables.

We described and introduced some dimensionality reduction methods to visualise and cluster ensembles of multivariate Gaussians. The methods are fast, invariant under a linear change of coordinates, and theoretically grounded in information geometry (via the Fisher metric). We applied the methods to ensembles of cgDNAloc predictions of Gaussians associated to exhaustive evaluation of all k-mer sequences. This process yielded a very clear cut clustering of 2^k clusters, with the clusters grouping pdfs for k-mers according to their sequence in the reduced purine/pyrimidine alphabet. In particular, the clustering we obtained is only revealed using the Fisher metric, rather than after a more standard PCA projection.

In the Chapter 6, we introduced methodology to use cgDNAloc to scan genomic-length sequences in search for outlier sites - in the sense of understanding the sequences that are furthest in their sequence-dependent statistical mechanical properties (as predicted by the cgDNA+ model) from the sequence independent, averaged pdf. We applied this method to scan parts of the *S. cerevisiae* genome (chr I-VIII) with various window sizes. We observed a strong and consistent signal, independent of window length, that outliers had a significantly enhanced A/T content preference as expressed in standard monomer-based sequence logos. The outliers even had a significant preference for AA/TT dimer steps as expressed in their less standard, dimer-based sequence logos.

We conclude this discussion with an outlook on potential further applications of the techniques introduced in this thesis combined with the cgDNAloc marginalisation procedure. The preliminary example data shown below has been computed with marginals of the cgDNA model, and for reasons of available time has not yet been extended to the analogous, but more accurate cgDNA+ computations.

As we described in section 1.1, one very important application in molecular biology is to be able to identify binding sites along the DNA sequence for Transcription Factor Binding Proteins (or TFBPs). The belief is that TFBPs bind to DNA via so-called indirect read out, i.e. the protein is believed to recognise certain statistical mechanics properties of the binding sites, rather than directly recognising the DNA sequence itself. (Such direct read out does happen for other proteins.) Typical experimental data for a specific TFBP is a library of sequences known to contain a binding site at a specific location. And the problem is to be able to scan genomes to identify all other likely binding sites for that specific TFBP.

To illustrate the potential of the cgDNAloc machinery to be applied to the study indirect readout, we briefly discuss a method that is analogous to the scanning of genomic sequence by frequency matrices, but with replacing the score matrix representation of a site with an average *consensus* cgDNAloc distribution. The procedure goes as follows: from the known experimental data a sequence logo over all known binding sites can be computed. Then a *consensus* cgDNAloc marginal distribution for binding sites for that particular protein can be computed by averaging over an ensemble of sub-sequences respecting the known binding site sequence logo. Once the consensus binding site marginal is known, it can be slid along the genome, in an analogous methodology to the one presented in Chapter 6, but now to identify any window locations where there is a particularly good agreement between the binding site consensus marginal, and the actual marginal at that location.

Fig 47 shows data for the particular example of the CTCF TFBP, where the binding site is known to be of length 19 bp. The *consensus* binding marginal distribution is slid along a 201 bp fragment containing precisely one known CTCF binding site. In this example, the minimal normalised KL divergence to the consensus binding cgDNAloc marginal distribution is obtained exactly at the site. This strongly suggests a role of DNA mechanics in binding site recognition, as already suggested for example by Rohs et al in [31]. However there are also many 'false positive' dips in the signal profile. One problem with using KL divergence as the similarity measure is that it may well be too stringent a criterion for similarity. More specifically it seems likely that a TFBP will bind to DNA locations where some of the lowest energy mode deformations of that particular sequence are a good match for the physical properties, shape and stiffness, of the protein. However for KL divergence between two marginals to be small, not only do the low energy modes have to be close, rather all modes have to be close, and it is perhaps not so likely that closeness of high energy modes will be a physically pertinent criterion for binding affinity.

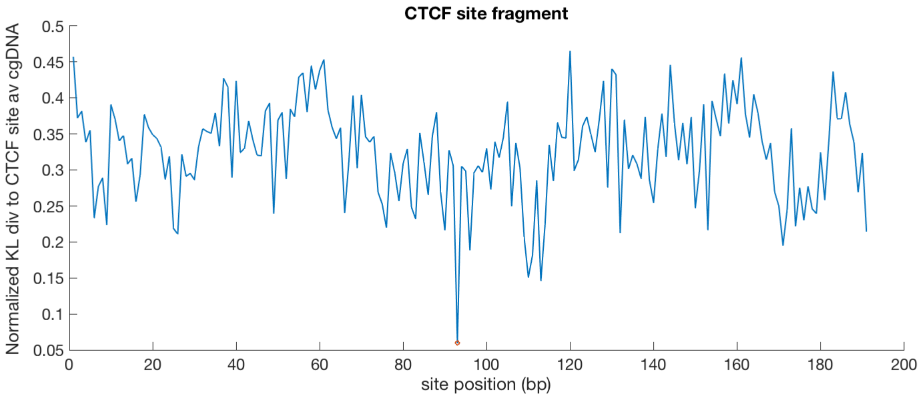


Figure 47: KL-divergence (normalised by number of degrees of freedom) between the 19bp cgDNAloc consensus binding site marginal distribution for the CTCF TFBP plotted against window location along a 201 bp DNA fragment known to contain one CTCF binding site, located at the red circle, which is also the global minimum of the signal.

Another important contemporary subject in the molecular biology of DNA is to understand the consequences of epigenetic base modifications. It is widely believed that such modifications affect the statistical mechanics of the double helix. One way to quantify that belief is to study the effects of epigenetic modifications on the spectra over an ensemble of sequences S of the sequence dependent *apparent* $\ell_p(S)$ and *dynamic* $\ell_d(S)$ persistence lengths of DNA fragments. As is more fully explained in [8] the apparent persistence length ℓ_p is a quite standard measure in polymer physics of a length scale on which tangent-tangent correlation decays along a polymer. It can be shown for Gaussian distributions such as cgDNA that the value of $\ell_p(S)$ depends on *both* the ground state $\mu(s)$ and the stiffness $K(S)$. Generally the spectrum over sequence of $\ell_p(S)$ can be very broad, with some very low values for some sequences. In fact the linear fit that arises in defining $\ell_p(S)$ is itself very bad when the ground state $\mu(s)$ is highly bent, which is invariably the case for low values of $\ell_p(S)$. For that reason the concept of sequence-dependent *dynamic* persistence length $\ell_d(S)$ was introduced in [8] by factoring out a ground state dependent part of the tangent-tangent correlation decay. Then the remaining part is fit to provide the dynamic persistence length $\ell_d(S)$ which to a good approximation depends only on the stiffness matrix $K(S)$. The spectrum over sequence of $\ell_d(S)$ still has variations of up to 50% between different sequences, but it is much more compact than the spectrum for $\ell_p(S)$. Thus the dynamic persistence length $\ell_d(S)$ is a good scalar proxy indicating an overall stiffness of the particular sequence, with larger values corresponding to stiffer sequences. And if there is a large difference between dynamic and apparent persistence lengths $\ell_d(S) - \ell_p(S) \gg 1$ for a specific sequence, then that is a single scalar proxy for the sequence being highly bent.

Two of the most biologically important epigenetic base modifications are 5' methylation or hydroxymethylation of cytosines C appearing in a CpG dimer step. From the point of view of the cgDNA coarse grain model the possibility of such modifications is just an extension of the sequence alphabet beyond the standard four member $\{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}$ alphabet to allow three versions of CpG parameter set junction blocks, depending on whether the cytosines in that junction step are unmodified, or methylated or hydroxymethylated. (In fact there is also the possibility of e.g. hemimethylation, where only the Watson strand cytosine is modified, and the Crick strand cytosine is not modified, etc. But here we will only present data for the more common fully modified case, where either both or neither backbone cytosines are modified.) A preliminary cgDNA parameter set in the extended sequence alphabet has been estimated from an appropriate library of MD simulations, where as before the accuracy of the cgDNA parameter set depends on the accuracy of the underlying MD simulation and its potentials.

With extended cgDNA parameter sets (see [10]) in hand for unmodified, methylated, and hydroxymethylated cases, the statistical mechanics effects of epigenetic modifications can be examined as follows. We first generated a library of 10K random 220bp long sequences with equal frequencies of each of the four standard bases at each location. For each of the 10K sequences we then generated three versions, an unmodified one, one where every CpG step was methylated, and one where every CpG step was hydroxymethylated. Then both apparent ℓ_p and dynamic ℓ_d persistence lengths were computed for each each of the three variants for each of the 10K sequences. Figure 48 provides histograms of the three possible differences between the apparent persistence lengths for each of the three versions of each sequence. It can be observed that none of the histograms of differences are sign definite, so that a change of certain type may either increase or decrease the apparent persistence length depending on the particular sequence. In general there is the trend that methylated and hydroxymethylated apparent persistent lengths are higher than for the analogous unmodified sequences, but it is only a trend that is not true for all sequences. In contrast the signal in Fig.49 for the analogous differences in dynamic persistence lengths $\ell_d(S)$ is more striking. The histogram of differences in ℓ_d between methylated and hydroxymethylated sequences is again sign indefinite, but is now quite narrow and centered close to zero. It can therefore be concluded that methylated versions of a sequence are sometimes slightly stiffer and sometimes slightly softer than the hydroxymethylated version of the same sequence, but any difference is always quite small. In contrast the differences in ℓ_d between either methylated or hydroxymethylated sequences and its unmodified version is always positive and is usually quite large. Some differences will always be quite small simply because there can be comparatively few CpG steps in some of the 10K 220bp sequences, so the modifications between the full sequences will be comparatively small. But the data on the dynamic persistence lengths ℓ_d reveals the clear signal that either methylation or hydroxymethylation base modifications stiffens any sequence compared to the unmodified values, and significantly stiffens most sequences compared to the unmodified values. When this dynamic persistence length signal is combined with the broader and indefinite distributions for the apparent persistence length differences, it can also be concluded that both methylation and hydroxymethylation have significant effects on the ground state shape $\mu(S)$. However there is no obvious pattern to indicate whether the epigenetic changes lead to an overall straightening or bending of the ground state. This is in fact a quite subtle question as it can depend on the phasings of the modified junctions.

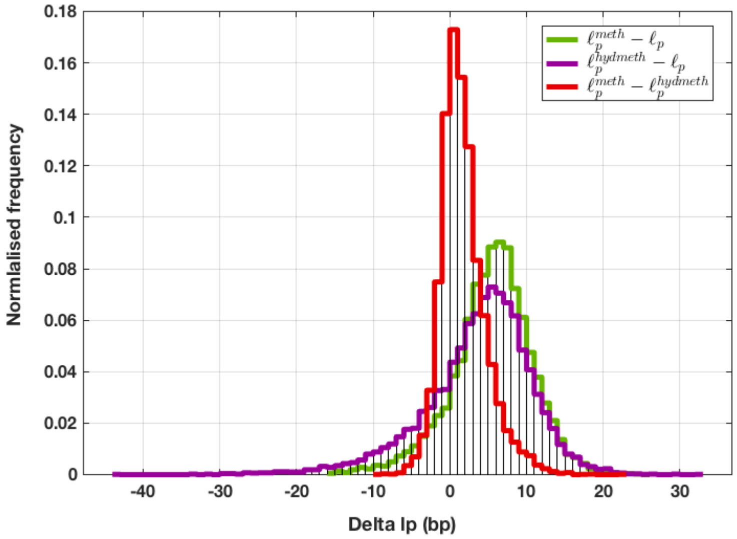


Figure 48: Normalized histograms of differences in apparent persistence lengths for 10,000 random (uniform discrete distribution on {A,C,G,T}) sequences of length 220 bp. Each curve shows the variation between 2 among the 3 different versions (standard, fully methylated, fully hydroxymethylated) of the sequences.

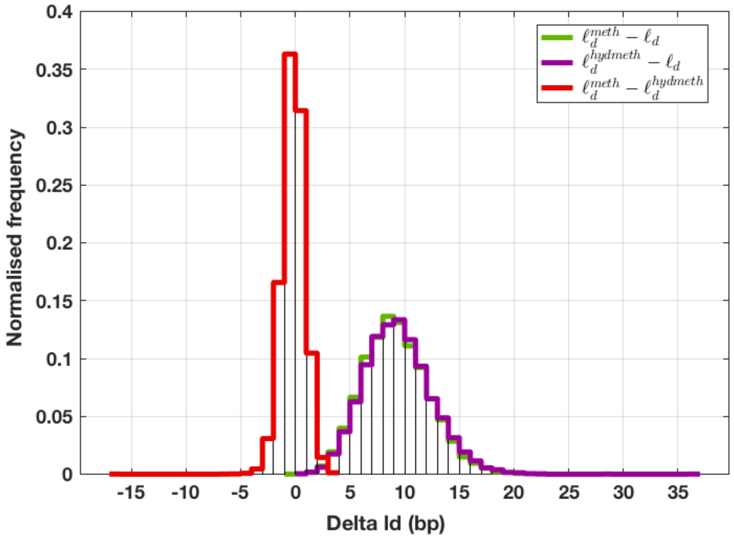


Figure 49: Normalized histograms of differences in dynamic persistence lengths for 10,000 random (uniform discrete distribution on {A,C,G,T}) sequences of length 220 bp. Each curve shows the variation between 2 among the 3 different versions (standard, fully methylated, fully hydroxymethylated) of the sequences.

Another more striking observation concerning the consequences of methylation sequence modifications can be seen in a version of the scanning procedure presented in

Chapter 6. We retain the unmodified sequence averaged mean distribution. But now we scan with a sliding window in which any CpG dimers are methylated. As illustrated in Fig 50, the signal is now strikingly different from the data presented in Chapter 6, where the outliers were all AA/TT rich sequences, and the cumulative distribution was single peaked. Now the outliers are much further from the mean, and the cumulative distribution of distances from the mean is itself multi-peaked. A closer analysis then reveals that each peak corresponds to the number of CpG steps lying inside the window (where for this window size, the maximum possible number of CpG steps is four). Thus in contrast to the standard alphabet case, the outliers with methylation are the CpG rich sequences, rather than AA/TT rich sequences. Moreover the changes in KL distance to the sequence averaged pdf induced by even a single methylated CpG step is much larger than any variation with sequence in the standard alphabet.

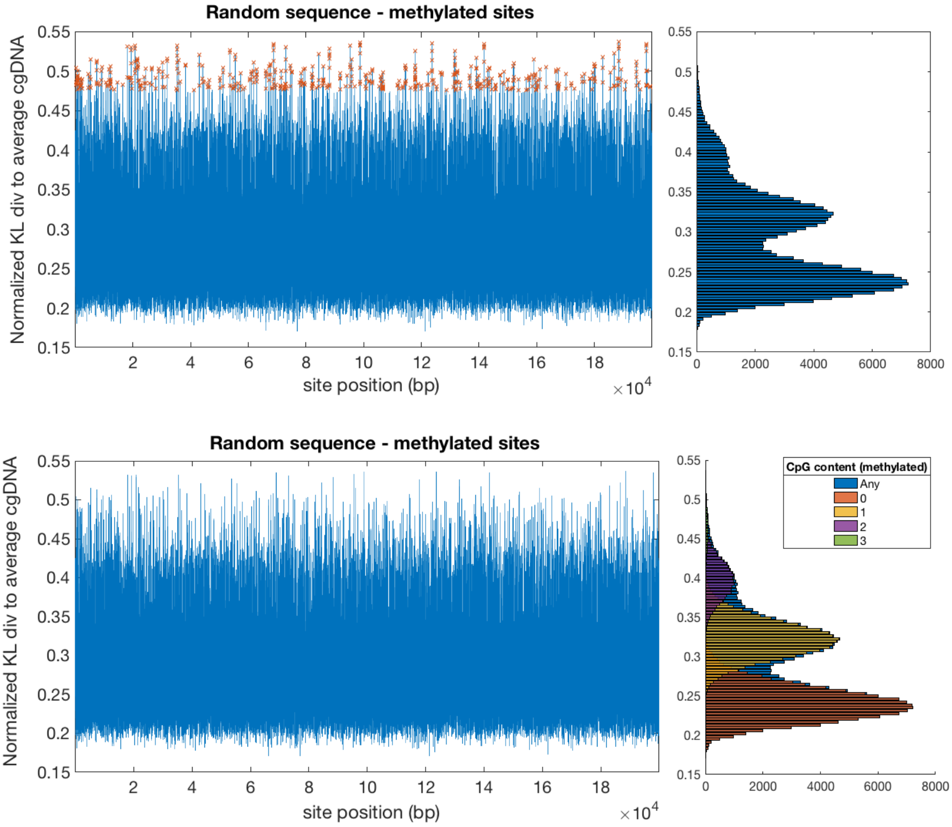


Figure 50: (top) Chromosome-scale KL to average cgDNA signal. (3 standard deviations from the mean), for 10 bps sites in a uniformly random assembled chromosomic length sequence, with all CpG steps fully methylated. Cumulative distribution is shown on the right. (Bottom): the same signal, with the cumulative distribution separately coloured by actual number of methylated CpG contained in the site.

This last observation highlights the drastic effect of CpG methylation on the statistical mechanical properties of DNA as modeled by the cgDNA+ model on the basis of Molecular dynamics simulation, and opens the door to future investigations in that regard, in the context of a growing interest for the role of epigenetic modifications in TF binding (see e.g. [63]) and in genomics in general.

Bibliography

[1] M. Antonini, L. Cruz, E. da Silva, D. Melpomeni, T. Ebrahimi, S. Foessel, E. Gil, S. Antonio, G. Menegaz, F. Pereira, *et al.*, “DNA-based media storage: State-of-the-art, challenges, use cases and requirements,” 2021.

[2] A. Patelli, *A sequence-dependent coarse-grain model of B-DNA with explicit description of bases and phosphate groups parametrised from large scale Molecular Dynamics simulations.* PhD thesis, EPFL, 9552, 2019.

[3] O. Gonzalez, D. Petkeviciute, and J. H. Maddocks, “A sequence-dependent rigid-base model of DNA,” *J. Chem. Phys.*, vol. 138, no. 5, 2013.

[4] J. Głowacki, *Computation and Visualization in Multiscale Modelling of DNA Mechanics.* PhD thesis, EPFL,7062, 2016.

[5] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, “Chromosomal DNA and its packaging in the chromatin fiber,” in *Molecular Biology of the Cell. 4th edition*, Garland Science, 2002.

[6] E. Segal and J. Widom, “Poly (dA: dT) tracts: major determinants of nucleosome organization,” *Current opinion in structural biology*, vol. 19, no. 1, pp. 65–71, 2009.

[7] W. K. Olson and V. B. Zhurkin, “Modeling DNA deformations,” *Current opinion in structural biology*, vol. 10, no. 3, pp. 286–297, 2000.

[8] D. Petkevičiūtė, M. Pasi, O. Gonzalez, and J. Maddocks, “cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA,” *Nucleic acids research*, vol. 42, no. 20, pp. e153–e153, 2014.

[9] L. De Bruin and J. H. Maddocks, “cgDNAweb: a web interface to the cgDNA sequence-dependent coarse-grain model of double-stranded DNA,” *Nucleic acids research*, vol. 46, no. W1, pp. W5–W10, 2018.

[10] R. Sharma, *cgNA+: A sequence-dependent coarse-grained model of double-*

stranded nucleic acids. PhD thesis, EPFL, 9792, 2022.

[11] D. Petkeviciute, *A DNA Coarse-Grain Rigid Base Model and Parameter Estimation from Molecular Dynamics Simulations.* PhD thesis, EPFL, 5520, 2019.

[12] A. E. Grandchamp, *On the statistical physics of chains and rods, with application to multi-scale sequence-dependent DNA modelling.* PhD thesis, EPFL, 6977, 2016.

[13] R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, and K. Zakrzewska, "Conformational analysis of nucleic acids revisited: Curves+," *Nucleic acids research,* vol. 37, no. 17, pp. 5917–5929, 2009.

[14] M. Slattery, T. Zhou, L. Yang, A. C. Dantas Machado, R. Gordân, and R. Rohs, "Absence of a simple code: How transcription factors read the genome," *Trends Biochem. Sci.*, vol. 39, no. 9, pp. 381–399, 2014.

[15] Q. Zhou and J. S. Liu, "Extracting sequence features to predict protein–DNA interactions: a comparative study," *Nucleic acids research,* vol. 36, no. 12, pp. 4137–4148, 2008.

[16] S. Inukai, K. H. Kock, and M. L. Bulyk, "Transcription factor-DNA binding: beyond binding site motifs," *Curr. Opin. Genet. Dev.*, vol. 43, pp. 110–119, 2017.

[17] R. Siddharthan, "Dinucleotide weight matrices for predicting transcription factor binding sites: Generalizing the position weight matrix," *PLoS One,* vol. 5, no. 3, 2010.

[18] I. Erill and M. C. O'Neill, "A reexamination of information theory-based methods for DNA-binding site identification," *BMC Bioinformatics,* vol. 10, no. 1, p. 57, 2009.

[19] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht, "Use of the perceptron algorithm to distinguish translational initiation sites in E. coli," *Nucleic acids research,* vol. 10, no. 9, pp. 2997–3011, 1982.

[20] G. D. Stormo, "Modeling the specificity of protein-DNA interactions.," *Quant. Biol.*, vol. 1, no. 2, pp. 115–130, 2013.

[21] Y. Zhao, S. Ruan, M. Pandey, and G. D. Stormo, "Improved models for transcription factor binding site identification using nonindependent interactions," *Genetics,* vol. 191, no. 3, pp. 781–790, 2012.

[22] J. Kähärä and H. Lähdesmäki, "Evaluating a linear k-mer model for protein-DNA interactions using high-throughput SELEX data," *BMC Bioinformatics,* vol. 14, no. SUPPL10, p. S2, 2013.

[23] A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C. Y. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, and W. W. Wasserman, "JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles," *Nucleic Acids Res.*, vol. 42, no. D1, pp. 142–147, 2014.

[24] N. B. Becker, *Sequence dependent elasticity of DNA.* PhD thesis, Technische Universitt Dresden, 2007.

[25] M. Michael Gromiha, J. G. Siebers, S. Selvaraj, H. Kono, and A. Sarai, "Intermolecular and intramolecular readout mechanisms in protein-DNA recognition," *J. Mol. Biol.*, vol. 337, no. 2, pp. 285–294, 2004.

[26] C. Kauffman and G. Karypis, "Computational tools for protein-DNA interactions," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 1, pp. 14–28, 2012.

[27] A. Mathelier, B. Xin, T. P. Chiu, L. Yang, R. Rohs, and W. W. Wasserman, "DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo," *Cell Syst.*, vol. 3, no. 3, pp. 278–286.e4, 2016.

[28] L. Yang, Y. Orenstein, A. Jolma, Y. Yin, J. Taipale, R. Shamir, and R. Rohs, "Transcription factor family-specific DNA shape readout revealed by quantitative specificity models," *Molecular systems biology*, vol. 13, no. 2, p. 910, 2017.

[29] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig, "The role of DNA shape in protein-DNA recognition.," *Nature*, vol. 461, no. 7268, pp. 1248–1253, 2009.

[30] W. Ma, L. Yang, R. Rohs, and W. S. Noble, "DNA sequence+shape kernel enables alignment-free modeling of transcription factor binding," *Bioinformatics*, vol. 33, no. May, pp. 3003–3010, 2017.

[31] J. Li, J. M. Sagendorf, T.-P. Chiu, M. Pasi, A. Perez, and R. Rohs, "Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding," *Nucleic acids research*, vol. 45, no. 22, pp. 12877–12887, 2017.

[32] S. Ruan and G. D. Stormo, "Comparison of discriminative motif optimization using matrix and DNA shape-based models," pp. 1–8, 2018.

[33] J. Yang and S. A. Ramsey, "A DNA shape-based regulatory score improves position-weight matrix-based recognition of transcription factor binding sites," *Bioinformatics*, vol. 31, no. 21, pp. 3445–3450, 2015.

[34] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of*

mathematical statistics, vol. 22, no. 1, pp. 79–86, 1951.

[35] S. I. Costa, S. A. Santos, and J. E. Strapasson, “Fisher information distance: A geometrical reading,” *Discrete Applied Mathematics*, vol. 197, pp. 59–69, 2015.

[36] B. Fuglede and F. Topsoe, “Jensen-Shannon divergence and Hilbert space embedding,” in *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, p. 31, IEEE, 2004.

[37] F. Topsøe, “Basic concepts, identities and inequalities-the toolkit of information theory,” *Entropy*, vol. 3, no. 3, pp. 162–190, 2001.

[38] D. Barber, *Bayesian reasoning and machine learning.* Cambridge University Press, 2012.

[39] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning.* No. 4, Springer, 2006.

[40] W. Thacker, “Metric-based principal components: data uncertainties,” *Tellus A*, vol. 48, no. 4, pp. 584–592, 1996.

[41] I. T. Jolliffe, *Principal Component Analysis, Second edition.* Springer, 2002.

[42] B. N. Parlett, *The symmetric eigenvalue problem.* SIAM, 1998.

[43] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[44] O. Gonzalez, M. Pasi, D. Petkeviciute, J. Glowacki, and J. Maddocks, “Absolute versus relative entropy parameter estimation in a coarse-grain model of DNA,” *Multiscale Modeling & Simulation*, vol. 15, no. 3, pp. 1073–1107, 2017.

[45] S. L. Lauritzen, *Graphical models.* Clarendon Press, 1996.

[46] M. I. Jordan, *Learning in graphical models.* MIT press, 1999.

[47] F. Jensen, S. Lauritzen, and K. Olesen, “Bayesian updating in causal probabilistic networks by local computations,” *Computational Statistics Quarterly*, vol. 4, pp. 269–282, 1990.

[48] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis.* John Wiley & Sons, 2009.

[49] J. McQueen, M. Meila, J. VanderPlas, and Z. Zhang, “megaman: Manifold learning with millions of points,” *arXiv preprint arXiv:1603.02763*, 2016.

[50] R. R. Coifman and S. Lafon, “Diffusion maps,” *Applied and computational harmonic analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[51] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint arXiv:1802.03426*, 2018.

[52] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction.* Springer, 2007.

[53] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications.* Springer Science & Business Media, 2005.

[54] G. A. Seber, *Multivariate observations.* John Wiley & Sons, 2009.

[55] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[56] H. Berendsen, J. Grigera, and T. Straatsma, “The missing term in effective pair potentials,” *Journal of Physical Chemistry*, vol. 91, no. 24, pp. 6269–6271, 1987.

[57] L. X. Dang, J. E. Rice, J. Caldwell, and P. A. Kollman, “Ion solvation in polarizable water: molecular dynamics simulations,” *Journal of the American Chemical Society*, vol. 113, no. 7, pp. 2481–2486, 1991.

[58] E. Neria, S. Fischer, and M. Karplus, “Simulation of activation free energies in molecular systems,” *The Journal of chemical physics*, vol. 105, no. 5, pp. 1902–1921, 1996.

[59] I. S. Joung and T. E. Cheatham III, “Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters,” *The Journal of Physical Chemistry B*, vol. 113, no. 40, pp. 13279–13290, 2009.

[60] J. S. Mitchell, J. Glowacki, A. E. Grandchamp, R. S. Manning, and J. H. Maddocks, “Sequence-Dependent Persistence Lengths of DNA,” *J. Chem. Theory Comput.*, vol. 13, no. 4, pp. 1539–1555, 2017.

[61] Y. Tong and R. S. Manning, “Quantifying the impact of simple DNA parameters on the cyclization J-factor for single-basepair-addition families,” *Sci. Rep.*, vol. 8, no. 1, p. 4882, 2018.

[62] O. Gonzalez and J. H. Maddocks, “Extracting parameters for base-pair level models of DNA from molecular dynamics simulations,” *Theor. Chem. Acc.*, vol. 106, no. 1-2, pp. 76–82, 2001.

[63] J. F. Kribelbauer, X.-J. Lu, R. Rohs, R. S. Mann, and H. J. Bussemaker, "Toward a mechanistic understanding of dna methylation readout by transcription factors," *Journal of molecular biology*, vol. 432, no. 6, pp. 1801–1815, 2020.

[64] T. Zhou, L. Yang, Y. Lu, I. Dror, A. C. Dantas Machado, T. Ghane, R. Di Felice, and R. Rohs, "DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale," *Nucleic acids research*, vol. 41, no. W1, pp. W56–W62, 2013.

[65] Y. Weiss and W. Freeman, "Correctness of belief propagation in gaussian graphical models of arbitrary topology," *Advances in neural information processing systems*, vol. 12, 1999.

[66] A. Mathelier, B. Xin, T.-P. Chiu, L. Yang, R. Rohs, and W. W. Wasserman, "DNA shape features improve transcription factor binding site predictions in vivo," *Cell systems*, vol. 3, no. 3, pp. 278–286, 2016.

[67] K. Carter, R. Raich, and A. Hero, "Learning on manifolds for clustering and visualization," in *Proc. of Allerton Conference*, 2007.

[68] B. Fuglede and F. Topsoe, "Jensen-Shannon divergence and Hilbert space embedding," *Int. Symp. onInformation Theory, 2004. ISIT 2004. Proceedings.*, pp. 30–30, 2004.

[69] M. Zuiddam, R. Everaers, and H. Schiessel, "Physics behind the mechanical nucleosome positioning code," *Phys. Rev. E*, vol. 96, no. 5, pp. 1–15, 2017.

[70] J. Walter, O. Gonzalez, and J. Maddocks, "On the stochastic modeling of rigid body systems with application to polymer dynamics," *Multiscale Model. Simul.*, vol. 8, no. 3, pp. 1018–1053, 2010.

[71] B. Deplancke, D. Alpern, and V. Gardeux, "The Genetics of Transcription Factor DNA Binding Variation," *Cell*, vol. 166, no. 3, pp. 538–554, 2016.

[72] F. Mordelet, J. Horton, A. J. Hartemink, B. E. Engelhardt, and R. Gordon, "Stability selection for regression-based models of transcription factor-DNA binding specificity," *Bioinformatics*, vol. 29, no. 13, pp. 117–125, 2013.

[73] V. Arsigny, X. Pennec, and N. Ayache, "Bi-Invariant Means in Lie Groups. Application to Left-Invariant Polyaffine Transformations," 2006.

[74] A. Höglund and O. Kohlbacher, "From sequence to structure and back again: approaches for predicting protein-DNA binding.," *Proteome Sci.*, vol. 2, no. 1, p. 3, 2004.

[75] T. Dršata, N. Špačková, P. Jurečka, M. Zgarbová, J. Šponer, and F. Lankaš, “Mechanical properties of symmetric and asymmetric DNA A-tracts: Implications for looping and nucleosome positioning,” *Nucleic Acids Res.*, vol. 42, no. 11, pp. 7383–7394, 2014.

[76] M. Annala, K. Laurila, H. Lähdesmäki, and M. Nykter, “A linear model for transcription factor binding affinity prediction in protein binding microarrays,” *PLoS One*, vol. 6, no. 5, pp. 1–13, 2011.

[77] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent, “The UCSC Genome Browser Database,” *Nucleic Acids Res.*, vol. 31, no. 1, pp. 51–54, 2003.

[78] L. Yang, T. Zhou, I. Dror, A. Mathelier, W. W. Wasserman, R. Gordân, and R. Rohs, “TFBSshape: A motif database for DNA shape features of transcription factor binding sites,” *Nucleic Acids Res.*, vol. 42, no. D1, 2014.

[79] D. Nigatu, W. Henkel, P. Sobetzko, and G. Muskhelishvili, “Relationship between digital information and thermodynamic stability in bacterial genomes,” *Eurasip J. Bioinforma. Syst. Biol.*, vol. 2016, no. 1, pp. 1–13, 2016.

[80] H. T. Rube, C. Rastogi, J. F. Kribelbauer, and H. J. Bussemaker, “A unified approach for quantifying and interpreting DNA shape readout by transcription factors,” *Mol. Syst. Biol.*, vol. 14, no. 2, p. e7902, 2018.

[81] E. T. Jaynes, “Information theory and statistical mechanics,” *Physical review*, vol. 106, no. 4, p. 620, 1957.

[82] Y. Yin, E. Morgunova, A. Jolma, E. Kaasinen, B. Sahu, S. Khund-Sayeed, P. K. Das, T. Kivioja, K. Dave, F. Zhong, *et al.*, “Impact of cytosine methylation on dna binding specificities of human transcription factors,” *Science*, vol. 356, no. 6337, p. eaaj2239, 2017.

[83] P. Fletcher, S. Joshi, C. Lu, and S. Pizer, “Gaussian Distributions on Lie Groups and Their Application to Statistical Shape Analysis,” vol. 2732, pp. 450–462, 2003.

[84] S. Rao, T. P. Chiu, J. F. Kribelbauer, R. S. Mann, H. J. Bussemaker, and R. Rohs, “Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein-DNA binding,” *Epigenetics and Chromatin*, vol. 11, no. 1, pp. 1–11, 2018.

Thomas Zwahlen

Curriculum Vitae

Education

2016	**MSc in Mathematics**, *École Polytechnique Fédérale de Lausanne*, GPA 5.52. **Minor in Space Technologies**
2013	**BSc in Mathematics**, *École Polytechnique Fédérale de Lausanne*, GPA 5.12 .
2011-2012	**Exchange Year**, *University of Waterloo, ON (Canada)*.
2008	**Swiss Federal Maturity**, *Gymnase Auguste Piccard (Lausanne)*. Awards for Mathematics, Philosophy, Latin, Best Final Results, Best Final Exams.

Thesis and Projects

PhD Thesis	*Landscapes of DNA Mechanics and Genomes*.
Master Thesis	*Maximal Radius of Hyperbolic Manifolds and Other Global Invariants*.
Minor Project	*Simulation of Radiation Pressure Impact on Asteroid Trajectories*.

Experience

2016-present	**Research Scientist**, EPFL, *Laboratory for Computation and Visualisation of Mathematics and Mechanics*.
2012-2021	**Teaching Assistant**, EPFL. Advanced Analysis I-IV for Maths, Physics and Engineering, Linear Algebra, Mathematical Modelling of DNA, Mechanical Biology, Differential Geomety of Framed Curves.
2012-2015	**Personal Tutor**. Analysis and Calculus, Topology, Physics, Geometry.
2013	Participation to Emahp (Mathematical and Humanitarian Project) in South Africa.