Sequence-Based Prediction of Topologically Associated Domains Enhanced by Cross-Species Comparison

Amaury Leroy

School of Computer Science
McGill University
Montreal, Quebec, Canada
July 2020

Supervisor
Dr. Mathieu Blanchette

A thesis submitted to McGill University in partial fulfillment of

the requirements of the degree of Master of Computer Science

Abstract

The three dimensional conformation of DNA plays a crucial role in the mechanisms of cell dynamics like gene regulation. It is based on a hierarchical model where, in particular, chromatin is compartmentalised into structures known as Topologically Associated Domains (TADs). With a length of around 1Mb, they define regulatory landscapes where chromatin loci interact more frequently than with regions located in adjacent domains. In this thesis, we present a computational pipeline for sequence-based annotations of these functional units, enhanced by cross-species comparison. The motivation of this work stems from the small number of species for which TAD annotations are available, due to the cost of Hi-C experiments needed to detect them. Based on studies aiming to characterize TADs, we postulate that, first, their boundaries are enriched in specific DNA-binding proteins, whose sites are hardwired in the genome sequence. Second, they are mainly conserved across neighboring species, so that using phylogeny could improve TAD inference. The first step of the pipeline consists in scanning the raw genome sequence to detect transcription factor binding sites. Then, a recurrent neural network is trained based on these sites to infer TAD left and right boundaries. Finally, predictions from different species are compared to update them and improve the general performance of the task. We developed ten approaches, cross-validated on five mammals for liver cells, with final AUC scores ranging between 70% and 90% depending on the input required to train the models. We observe that DNA sequence features convey subsequent knowledge about TAD boundaries, but the performance is too low to expect an accurate genome-wide annotation. More strikingly, the cross-species comparison was still an unexplored approach that shows a consequent improvement in TAD boundary inference, providing a promising future for the understanding of these domains.

Abrégé

L'organisation tridimensionnelle de l'ADN joue un rôle clé dans les mécanismes de dynamique cellulaire comme la régulation génétique. Elle repose sur un modèle hiérarchique où, en particulier, la chromatine est compartimentée en structures appelées TADs. Avec une longueur d'environ 1Mb, ils définissent des unités de régulation au sein desquelles les loci de chromatine interagissent plus fréquemment. Dans cette thèse, nous présentons un pipeline de calcul pour l'annotation de ces unités fonctionnelles. Ce travail a été motivé par le fait que l'identification des TADs est restreinte à un faible nombre d'espèces, dû au coût important des expériences Hi-C, nécessaires à leur détection. A partir des nombreuses études sur ces structures, nous faisons l'hypothèse que les frontières des TADs sont enrichies en certaines protéines liées à l'ADN, détectables à partir de la séquence génomique. En outre, les TADs semblent globalement conservés entre espèces proches, laissant penser qu'une comparaison entre ces dernières pourraient être bénéfique à leur inférence. La première étape consiste en la détection des sites d'attache des facteurs de transcription à partir de la séquence ADN. Ensuite, un réseau récurrent est entraîné à prédire distinctement les frontières gauches et droites des TADs. Finalement, les prédictions des différentes espèces sont comparées pour être affinées. Nous avons développé 10 approches, validées sur 5 mammifères pour des cellules du foie. Elles atteignent des scores AUC compris entre 70% et 90%, selon la nature des composantes d'entrée du prédicteur. Nous observons que la séquence ADN porte des informations importantes sur les frontières des TADs, mais son seul apport donne des résultats insuffisants pour espérer une annotation complète et précise du génome. Plus remarquablement, une comparaison inter-espèces, approche encore inexplorée, montre une amélioration significative des prédictions, donnant des pistes prometteuses pour la compréhension des TADs.

Acknowledgements

I would first like to express my deepest gratitude to my supervisor, Dr. Mathieu Blanchette, for his support and guidance through my entire project. Thanks to his ongoing monitoring, his precious advice and his professional goodwill, he has been the cornerstone of my research work achievement.

My thanks also go to the examiner of my Master thesis.

I would finally like to thank every member of Dr. Blanchette's lab, who welcomed me and were always happy to answer my questions, no matter how seemingly inconsequential. I was highly impressed by the variety of topics covered, and they all have been a continuous source of motivation and intellectual stimulation.

Particular thought for Elliot Layne, Samy Coulombe, Christopher J.F. Cameron, Kaiwan Rahimian-Bajgiran, Dongjoon Lim, Zichao Yan and Alexandre Butyaev, who have been of great help during insightful conversations.

Table of Contents

List of Figures

List of Tables

Acronyms

3D Three-Dimensional

3C Chromosome Conformation Capture

bp base pair

ChIP-Seq Chromatin Immunoprecipitation Sequencing

CNN Convolutional Neural Network

CTCF CCCTC-Binding Factor

DL Deep Learning

DNA DeoxyriboNucleic Acid

FISH Fluorescent *In Situ* Hybridization

GRU Gated Recurrent Unit

LSTM Long Short-Term Memory

ML Machine Learning

NLP Natural Language Processing

PWM Position Weight Matrix

PCR Polymerase Chain Reaction

RNN Recurrent Neural Network

TAD Topologically Associated Domain

TFBS Transcription Factor Binding Site

Chapter 1

Introduction

The first draft sequence of the human genome was produced in 2001, thanks to innovative shotgun sequencing methods [1]. It constitutes the final piece of the 13-year-long Human Genome Project, opening countless perspectives in medicine and biology. Nevertheless, the insights provided by a linear sequence remain incomplete as they do not convey knowledge on both the spatial conformation and its dynamics inside the nucleus. The latter are now of a paramount interest to a new phase called the 4D nucleome project, aiming to map the genome both in space and time to better understand its functions and interactions [2, 3]. In this thesis, we aim at predicting a type of structure called Topologically Associated Domains (TADs) from Deoxyribonucleic Acid (DNA)-sequence features and phylogenetic data. This introductory chapter brings in the key concepts to understand the purpose of the study. We first explain in detail the three-dimensional (3D) organization of the genome and the experimental methods to capture it. Then, we tackle the dynamics of this structure across time, both at the scale of cell cycle and evolution of species. Finally, we review innovative computational tools applied to understanding 3D genomics, with a particular focus on the benefits that deep learning can provide, paving the way for such an approach to predict TADs.

1.1 3D Genomics

1.1.1 DNA Folding

Chromosome folding in the nucleus is critical in eukaryotes for very basic reasons. For example, the human genome consists of many molecules of DNA combining into a 2-meter-long structure which must fit inside each cell's nucleus. Therefore, a bio-mechanical compaction is needed. Equally, multiple evidences advocate for the crucial role of a non random DNA conformation on mechanisms of cell dynamics. Indeed, such functional architecture affects gene expression by bringing together genes and distal regulatory elements [4]. Some loops can also play a role in gene silencing and replication timing [5]. Finally, a complex interplay exists between spatial organization and mutation rates, with altered regions leading to disease like cancer [6]. Indeed, some Single Nucleotide Polymorphisms (SNPs) are associated with changes in chromatin accessibility and interactions linked with transcription between distal loci [7]. Therefore, the structure of DNA and its interactions with proteins to form a complex called chromatin, was until recently a key yet unheralded topic, but is now attracting growing interest. Many genomics approaches for mapping DNA in space allow us to understand it at unprecedented level, and will be largely developed in subsection 1.1.2. Thanks to them, we can bring to light different levels of packaging which are summarized in Figure 1.1. It is important to note that this multi-scale hierarchy is specific to mammals and may be broadened to other substructures, according to the precision of the definitions at stake.

First of all, DNA wraps twice around histone octamers to form nucleosomes of 147 base pairs (bp), reducing the length sevenfold. It can interact with several proteins thanks to binding sites all over the genome, creating chromatin loops (A) which may, for example, enhance or disrupt physical contacts between regulatory elements. This process is the cornerstone of gene expression. For some of these loops, bigger regions are bridged - on a scale of 40kbp to 3Mbp - called TADs (B), allowing recurring bonds in open chromatin [9]. At a higher scale and at least in mammals, TADs are partitioned into compart-

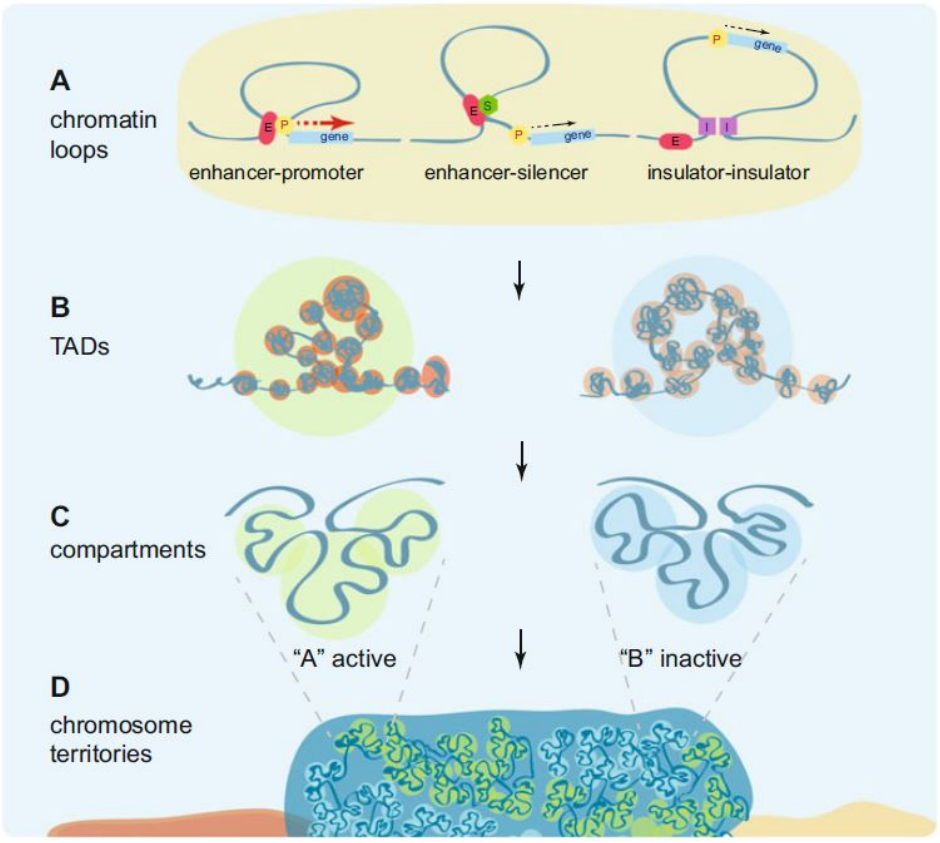


Figure 1.1: Hierarchy and organization of 3D mammalian genome reproduced from [8]

ments with megabase-scale interactions (C). ”A” compartments gather open and active chromatin while ”B” compartments are more compact and enriched in repressive chromatin. Finally, chromosomal territories (D) define even larger portions of the nucleus occupied by chromosomes with a functional role. Sexton et al. made an analogy with the organization of a protein reproduced in Figure 1.2, from the amino acid to a complex of proteins through functional units like helices or sheets [10]. Just like these structures influence the activity of the protein because of their shape, TADs are likely to be of paramount importance for processes inside the nucleus. We will focus on these units in this study.

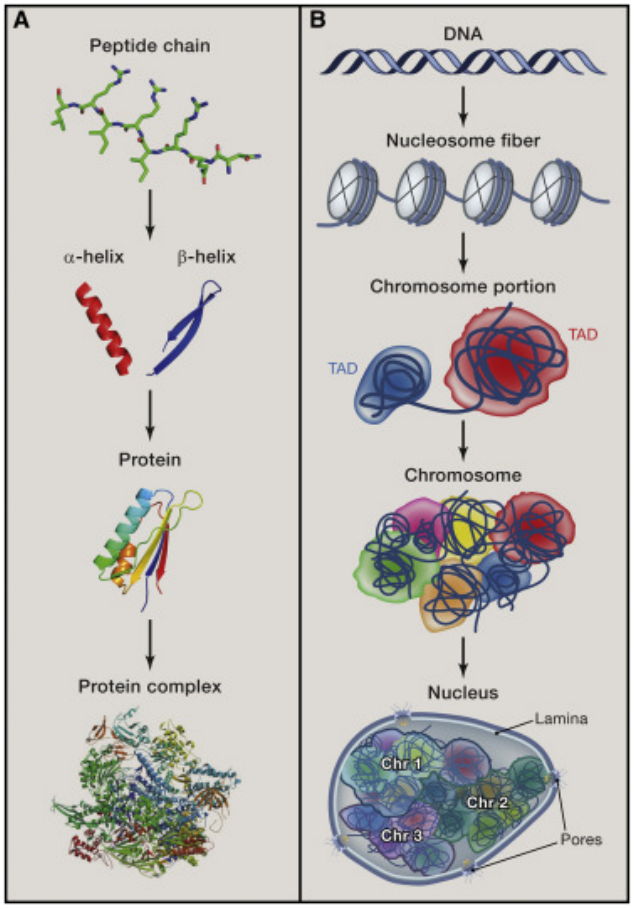


Figure 1.2: Analogy between DNA and protein structures, from [10]. 3D conformation is a highly organized process influencing theur functional role

1.1.2 Experimental Tools and 3C Technologies

The mapping of the genome organization was made possible by multiple innovative methods that stem from the so-called 3C-technologies - standing for Chromosome Conformation Capture - and that depend on the resolution desired and type of data available. Before that, the dynamics of DNA inside the nucleus was studied using microscopy-based techniques enabling direct and dynamic visualization of the structure - the fluorescent *in situ* hybridization (FISH) is a common process among many others [11]. Nevertheless, these approaches are limited to a small number of loci and do not provide a genome-wide

mapping. The 3C technologies retrieve distal bonds between loci based on the frequency at which the chromatin fragments interact [12]. Usual steps are described in Figure 1.3:

1. Formaldehyde cross-linking of cells to strengthen covalent bonds
2. Digestion by a restriction enzyme such as HidIII or MboI to capture interactions
3. ligation of DNA ends, reverse crosslinking, and quantification of ligation products with Polymerase Chain Reaction (PCR)
4. Labelling with biotin (B on the Figure 1.3, to identify true sites) of the products in the case of Hi-C, to create a library of DNA reads that is analyzed with massive parallel sequencing [13].

Each improvement on one of these steps - mainly in the approaches to identify interactions after reverse crosslinking, but also in the process of fragmentation like ChIA-PET that requires sonication - leads to the development of a new technique aiming at a particular resolution or type of interaction. Experiments on multiple regions of loci allow genome-wide results but represent a fixed superposition of snapshots from different genome structures rather than a unique source. Indeed, typical maps come from many different experiments that are performed at various possible states of a cell, resulting in an average estimation and limiting the insights on the real structure of a particular nucleus in a given state. New methods are now in development to study 3D genomics at single cell level, in particular for Hi-C [14].

Among 3C-technologies, Hi-C was introduced by Lieberman et al. in 2009, and is the cornerstone of the characterization of TADs [15]. The first two steps are similar to 3C but the technology then involves deep sequencing in order to generate a massive catalog of different ligations mapped to a reference genome - hundreds of millions of sequencing reads per experiment [8].

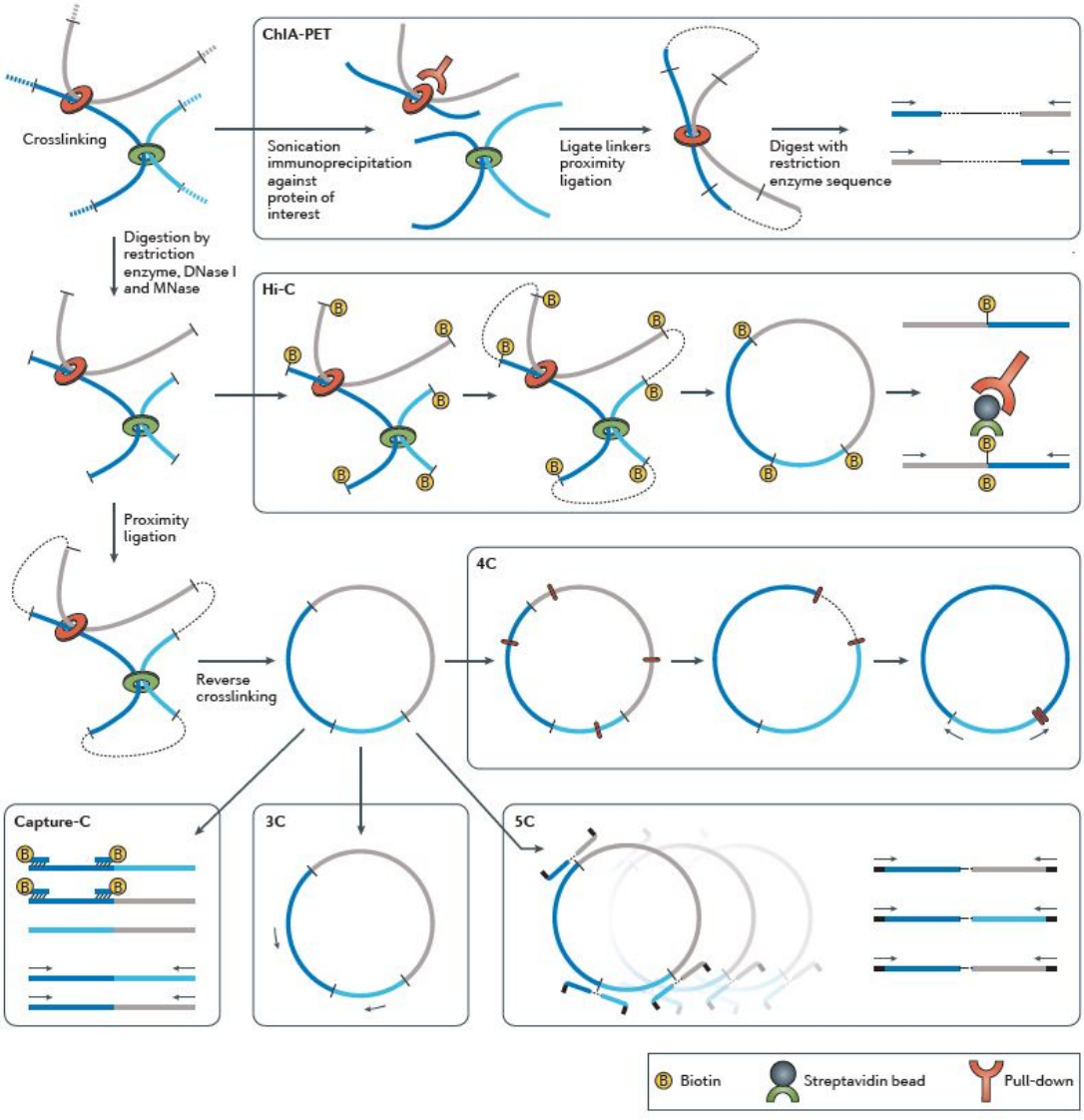


Figure 1.3: Fundamental steps of the 3C method and its variants, reproduced from [9]

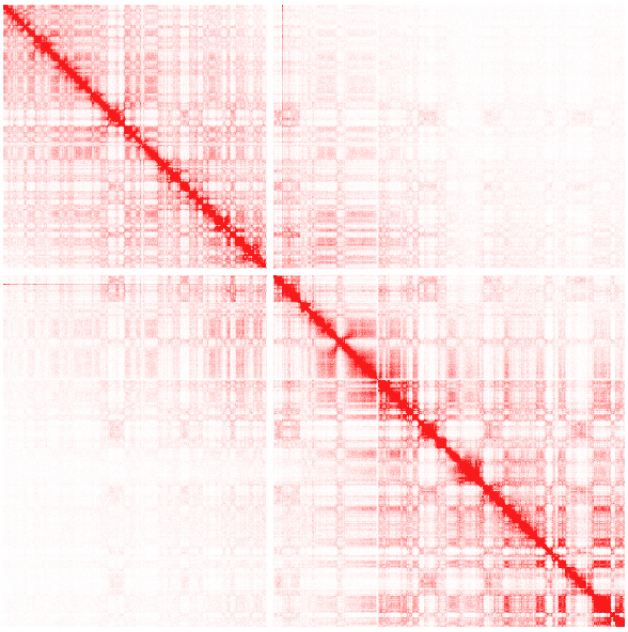


Figure 1.4: Example of Hi-C contact matrix for human chromosome 1. Interactions are more frequent on the diagonal, corresponding to loci that are close on the linear sequence. Bigger red squares highlight condensed loops, enriched in interactions, and the possible presence of TADs.

As shown in Figure 1.4, the typical output of an Hi-C experiment takes the form of a symmetric contact matrix, each cell representing the number of read pairs that have been found to physically interact. The resolution of a Hi-C experiment is the size of the bin and constitutes a very active research area. It first depends on the precision of the restriction enzyme used, more precisely the size of the resulting fragment - 400bp for MboI vs. 4kb for HindIII for instance. In addition, it is limited by the sequencing coverage to get a reasonable number of samples for each cell. Indeed, high-resolution coverage results in more precise but extremely sparse matrices - most pairs are observed zero times -, making the measurements unreliable and difficult to statistically analyze. Usually ranging from 40kb to 1Mb, recent studies achieved kb-resolutions [16], and even fragment-based resolution thanks to computational and mathematical correction of the matrix [17].

1.1.3 Structure and Formation of TADs

Thanks to Hi-C experiments performed on mammalian genomes, Topologically Associated Domains (TAD) were first identified by Dixon et al. in 2012 [18]. Following this study, TADs - often characterized by different mechanisms and size - were found in other vertebrates, *Drosophila* [19], *Caenorhabditis elegans* [20], and more or less similar structures have been identified in plants [21], yeast [22] or bacteria [23]. We will restrict our study

to mammals. The ubiquity of such domains in many species advocates for their importance as functional and conserved units of the genome, defined as regions inside which chromatin loci interact more frequently than with regions located in adjacent domains. Therefore, understanding their biological organization is crucial. We can see TADs as big condensed loops with a median size of 880kb, defining regulatory landscapes. However, the epigenetic state alone is no sufficient to create boundaries. Their formation lies in the joint action of several proteins binding to precise loci and interacting among themselves. TAD boundaries are then enriched for these proteins, but the latter can also be found inside a TAD to create less stable loops, and leading to some confusions on the exact delineation of these structures - TADs can often be divided into sub-TADs. The classification of TADs is also ambiguous and variable with studies as the mapping is highly dependent on Hi-C resolution.

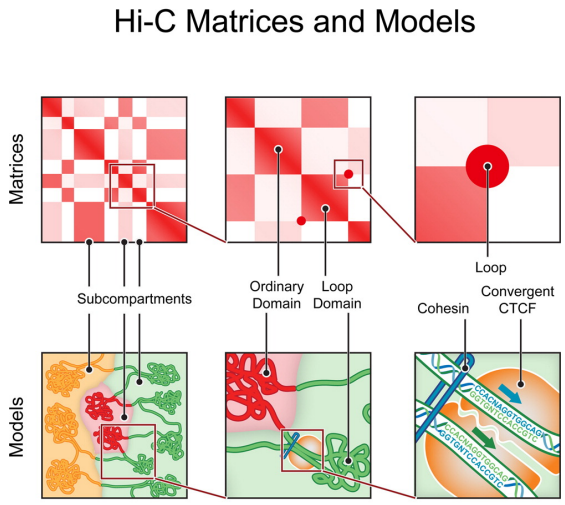
Many epigenomic marks have been studied to assess their influence on chromatin shape. Among them, the complex CTCF-cohesin has emerged as an essential piece for TAD formation [24]. Cohesin is a protein complex, previously known for its role as a structural mediator during cell division or DNA repair [25], but which has been found to also play a role in chromatin looping. More importantly, the CCCTC-binding factor - known as CTCF - is considered as the key player in anchoring TADs. Indeed, this zinc-finger protein is detected at 76% of boundaries [9]. It has always been a famous insulator protein, binding a nonpalindromic 20-bp DNA motif [24]. For 3D looping, two distal CTCF proteins, bound to DNA and surrounded by cohesin, can interact to gather loci and create a condensed loop. The orientation of the CTCF is crucial as they can only interact in a symmetric way, so that CTCF sites at loop anchors are associated with a convergent orientation [9]. The so-called loop extrusion model is described in Figure 1.5. Other putative elements have been found close to boundaries in high proportion. We can cite among others histone marks (like H3K4me3), housekeeping genes or DNA hypersensitive sites, although they have conflicting implications depending on the study. All these observations suggest that the assembly of domain structure is hardwired in the genome, and represent a useful source of data for the TAD prediction task. Nevertheless, a significant number of boundaries does not obey to the previously stated rules - 24% of boundaries are

independent from the CTCF-cohesin complex -, so that a precise sequence-based identification of TADs remains a difficult task. TAD conservation across species could represent a great additional source of information, and will be discussed in Subsection 1.2.2

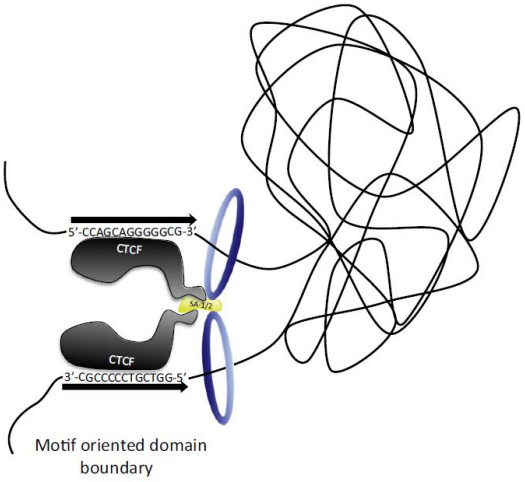
1.2 Dynamics of 3D Organization

1.2.1 TADs and cell differentiation

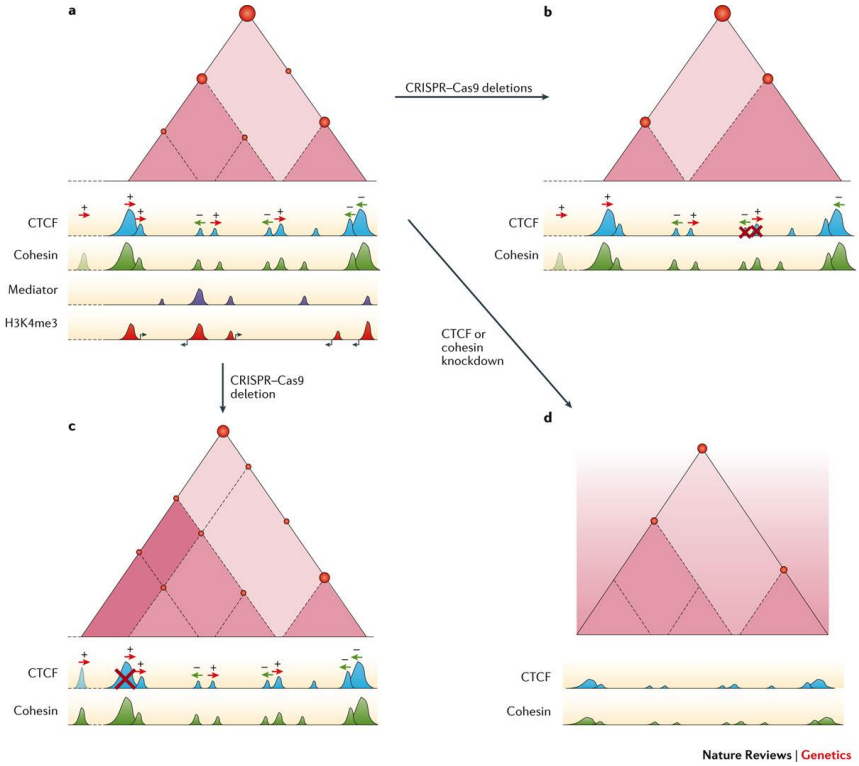
In addition to its complex spatial structure, chromatin has been found to evolve dynamically over time. In the light of our TAD inference problem, it represents a precious footprint to help us understand chromatin folding as a logical 4D process, consistent with its functional role in gene expression. First of all, interactions between distal loci may vary over the cell cycle. Bio-mechanical constraints of mitosis force the established organization to temporarily disrupt, but there is a striking global re-establishment of the 3D organization after each cell division [9], which proves that the folding is far from being random. Nevertheless, at the time scale of cell differentiation, some changes do occur. While the global structure of TADs remains stable, interactions inside TADs undergo major rearrangements which depend on the cell fate and the specific function to assume [26, 27]. Occasionally, the disruption of a TAD boundary can generate a new bigger one and allow completely new contacts and functions. In the same way, TADs can also switch from an active "A" compartment to an inactive "B" one, altering or fostering the TAD-TAD interactions [28]. These modifications lead to a new hierarchy in transcription processes and advocate for the essential role of chromatin folding, in particular for the stability of TADs as a functional unit of the genome across cell types.



(a) TADs are units represented as squares of high intensity in Hi-C interaction frequency data



(b) The convergent CTCF-cohesin complex is considered the most important for anchoring TADs.



(c) The orientation of the complex is crucial as some mutations or knockdown of binding sites can completely disrupt TADs or shape other ones

Figure 1.5: TAD formation as condensed loop visible from Hi-C matrix, thanks to a complex of proteins. Reproduced from [9,16,24]

1.2.2 TAD conservation across species

Because a pattern between genome organization across cell types emerges, it is natural to wonder what the link between 3D structure and species evolution is. Indeed, the functional role of TADs is an evidence of the negative selection applied to disruption of boundaries. A single mutation can break a boundary, reorganize TADs that create new ectopic contacts, gene misexpression and lead to the development of diseases like cancer [29]. Nevertheless, some rearrangement scenarios can provide new interesting functions and benefit the resulting phenotypes. For instance, in vertebrates, *HoxD* genes are organized in a specific conformation (at a boundary, flankes by two TADs), allowing a proper gene regulation. This structure is not conserved in the invertebrate Amphioxus, where the cluster is present within a single TAD. We may assume that the presence of two TADs is the consequence of a sequence of mutations, that is now stable for vertebrates as only a large deletion could lead to their fusion [30]. This specificity has created a new regulatory landscape in the vertebrate lineage. Recently, many studies tried to quantify the conservation of the 3D genome across evolution, mainly for vertebrates and in particular primates. It has emerged that many TADs are conserved between close species, the difference being linked to each phenotype's specificity. For example, half of TAD boundaries are shared between human and chimpanzee [31]. More precisely, as genomes mainly differ because of big rearrangement of syntenic blocks, it has been found that breakpoints are enriched at TAD boundaries and rare inside those domains, even between species with a more ancient common ancestor [32,33]. This observation is consistent with the molecular context of TADs, where regions inside TADs are condensed and less likely to mutate, while loci outside of TADs are associated with chromatin fragility.

The CTCF-cohesin complex is then an important stake to understand TAD conservation. First, because this protein complex plays different roles depending on the family of species, a common hypothesis states that it has mutated across evolution, sometimes disappearing - for *C. elegans*-, or becoming a central piece for insulation - for vertebrates [24]. This diversification go hand in hand with the increased rate of mutations at TAD bound-

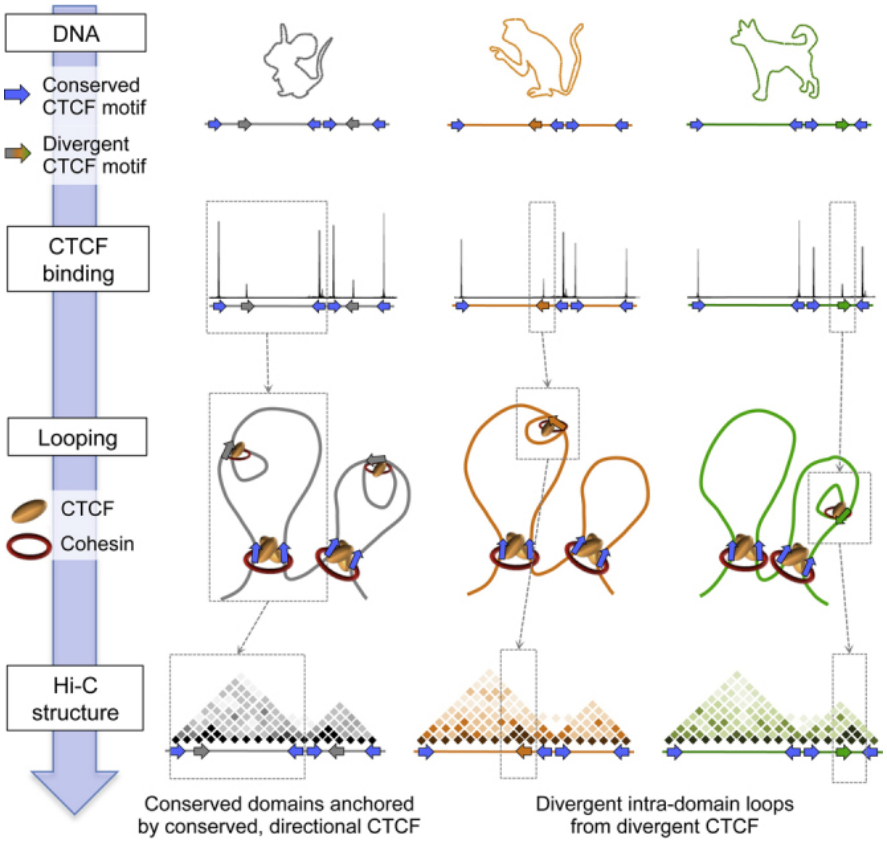


Figure 1.6: TAD conservation based on CTCF-cohesin complex across evolution, reproduced from [35]

aries, enriched with this complex. For vertebrates, conserved CTCF binding sites between species are highly correlated with convergence in their orientation, resulting in stable TADs. The *Six* homeobox gene cluster, crucial for their development, has then remained organized into two adjacent TADs [34]. That means that the orientation is a key for loop creation and gives the boundary the stability to last with evolution (see Figure 1.6). Conversely, non-conserved CTCF binding sites are enriched inside domains and more likely to mutate [35]. These results are strong evidence for the help of phylogeny for TAD prediction, and recent studies have brought out the conserved patterns of 3D structure with muli-species Hi-C data to map the Hi-C matrix as a mix between ancient, steady domains and new species-specific ones [36].

1.2.3 Character-based Phylogeny

In order to gather insights from this correlation between 3D organization and evolution, handling phylogenetic data is mandatory. The original idea is to reconstruct a phylogenetic tree based on a set of observed characters on multiple species on the principle of maximum parsimony. This was first introduced by Fitch in 1971 and Sankoff in 1975 with two algorithms tackling the problematic through different paradigms [37,38]. For a given bin of the genome, the character can be, for instance, one of the four nucleotides - each value is called a state -, or the presence or absence of a TAD boundary - binary states. Considering that the characters are observed at leaves, the goal is to label the internal nodes with maximum parsimony, in other words with the minimum number of changes along the fixed tree. The theory is based on the assumption that observed characters result from the fewest possible mutations, and the parsimony score is defined as the sum of all mutations found in the tree. Each mutation can be of different cost depending on the biological context, resulting in a weighted parsimony problem. Both algorithms are equivalent and involve dynamic programming. Sankoff's method computes the minimum score and the associated labelling, while Fitch's technique directly finds the same set of labels without the score. For Sankoff's algorithm, a forward step calculates the best score for every possible label from leaves to the root, and uses the fact that the score of a parent is based only on scores of its children. The state with smallest score at root is then the most parsimonious character. Then, a backward step travels down to the tree and assigns each vertex with the state that leads to this best parsimony. Fitch's algorithm is similar but reasons on ensembles of possible states rather than scores, the state of the parent being either the union or the intersection of the sets of its children. A tree and the summary of these two algorithms is highilghted in Figure 1.7

These solutions are efficient and widely-used, but ignore branch length on the phylogenetic tree. For instance, a mutation is more likely to occur between an ancestor and its child if the branch is long, witness to the slow evolution towards the new species. To overcome this issue, Felsenstein proposed a probabilistic model bringing into play Markov

Sankoff

$S_i(\text{parent}) =$
$\min_j(S_j(\text{left child}) + C_{ij})$
+
$\min_j(S_j(\text{right child}) + C_{ij})$

Fitch
T(n) the set of states for vertex n

T(parent)=
T(left child) ∩ T(right child)
if non-empty intersection

T(parent)=
T(left child) ∪ T(right child)
otherwise

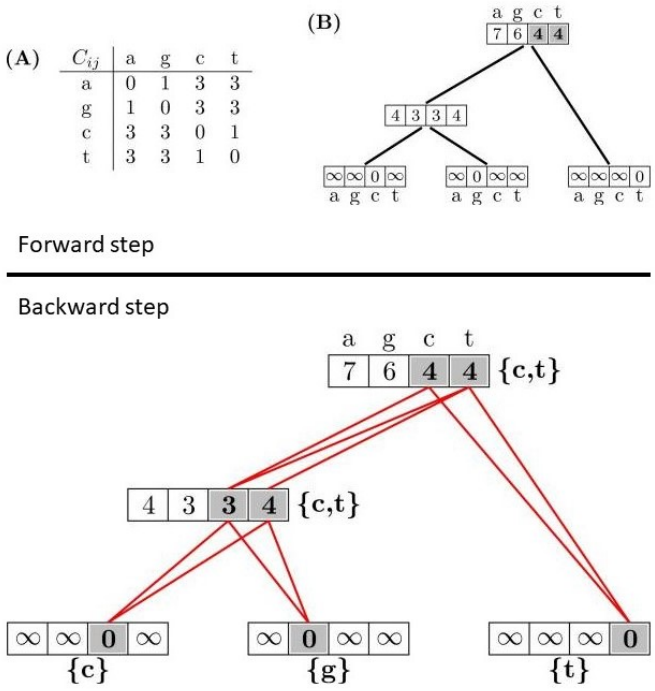


Figure 1.7: Sankoff's (top left) and Fitch's (bottom left) forward algorithms for character-based tree reconstruction with maximum parsimony. The backward step is common. Example with nucleotides (right) for Sankoff's method, reproduced from [39]

process [40]. The method also uses dynamic programming and a forward-backward pass through the tree. Given the topology T of a tree, label states (b, c) of two species (i, j), and some assumptions on independence, we can compute the likelihood of the subtree rooted at the common ancestor k given label state a as follows:

$$\Pr[T_k|a] = \sum_b \Pr[b|a, T_i] \Pr[T_i|b] \sum_c \Pr[c|a, T_j] \Pr[T_j|c]$$

Each sum corresponds to the information carried by a specific child, the first term being the probability of state transition, depending on branch length, the second one being the likelihood of the subtree rooted at this child. With dynamic programming, we are able to start from the leaves where the likelihood is known, to reach the root where the likelihood

is for the whole tree. The same traceback can finally be applied to reconstruct the best set of states, corresponding to the maximum likelihood. These kinds of methods are the cornerstone of a phylogenetic model to encapsulate all the knowledge on a tree, and link TAD predictions between species for a classification task.

1.3 Computational Tools for 3D Genomics

1.3.1 TAD callers using Hi-C data

In order to detect TADs, the main source of data comes from the results of experiments described in subsection 1.1.2. Hi-C contact maps are by far the most used data to infer 3D structure, but need a substantial downstream work. computational tools using Hi-C maps are the most common methods for TAD prediction. An impressive amount of software is available, and two published reviews summarize and compare their results [41, 42]. They differ in the input they require, the method they apply and the task they focus on. First of all, most of them take as input a Hi-C contact map, whether it be under .hic format file or raw interaction matrix. However, some tools also need external data like transcription marks from Chromatin Immunoprecipitation Sequencing (ChIP-Seq) experiments. Concerning the computational method, the common assumption is that chromatin contacts are more frequent within TADs than among them, with a peak at boundaries. Many approaches compute a linear score along a genome divided into bins. New ones involve statistical models or clustering on the contact matrix. Even more recently, graph theory has been useful to detect communities inside a graph whose adjacency matrix was the Hi-C map. Finally, the output of these models are specific and make their comparison sensitive. Some tools only return disjoint TADs, other accept overlapping or even nested domains, and the presence of inter-TAD gaps are also a point of divergence. A few of them only output the boundaries to avoid those confusions and the smoothing of boundaries to complete domains is left to the user. Beyond that, these differences highlight the fact

that the definition of a TAD is flexible and must be put in each study's context to draw insightful conclusions.

To compare TAD callers, both reviews tested the consistence between themselves and with commonly known biological behaviors - like manual TAD annotations also under serious debate, or CTCF enrichment around boundaries -, the robustness to data resolution and normalization, and the parameters required to get the best results. The common conclusion is that the predictions vary significantly between tools, but no clear winner emerges (see Figure 1.8). Indeed, some algorithms are very concordant with biological observations on validation set but struggle to adapt to high resolution and sparse matrices. Some tools better fit the detection of sub-TADs, while less shrewd but more robust ones focus on bigger steady domains. Overall, all these TAD callers are very satisfying tools but are task-specific, and the user should know the strengths and weaknesses of the chosen model to ensure a enlightened overview of the results.

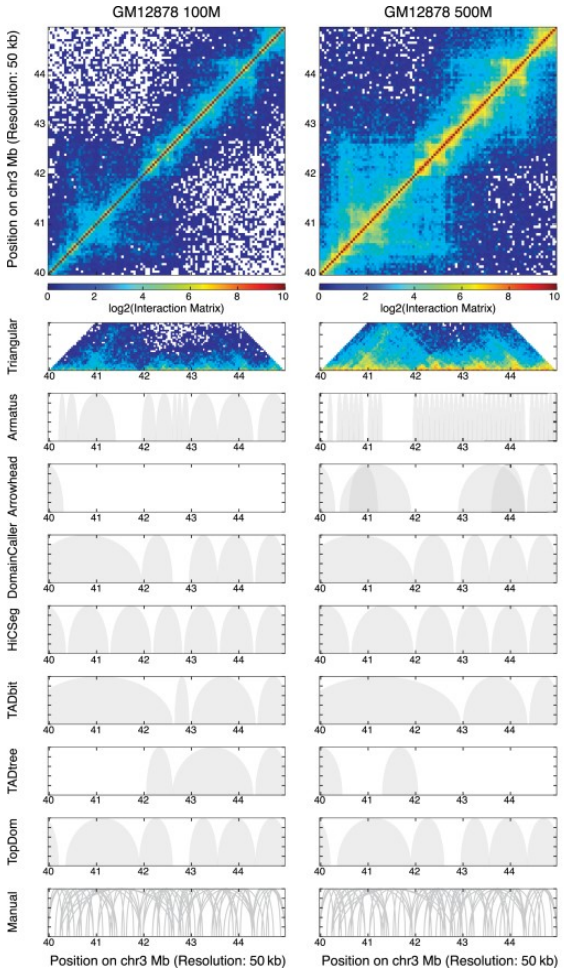


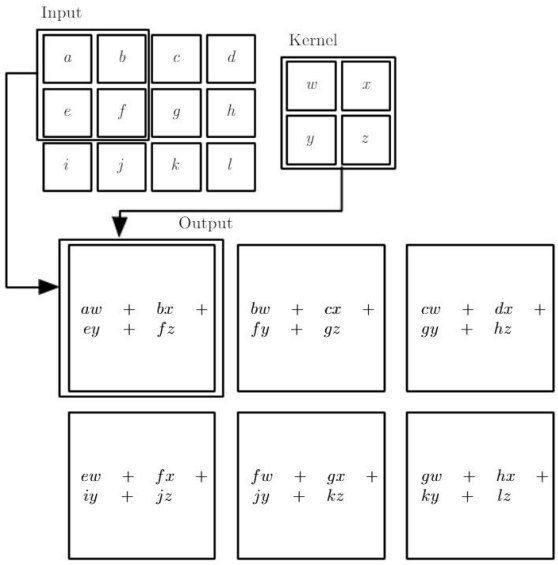
Figure 1.8: Comparison of seven TAD callers at two levels of sequencing depth, at 50 kb resolution, for chr3:40–45 Mb. Reproduced from [41]

1.3.2 Deep Learning background

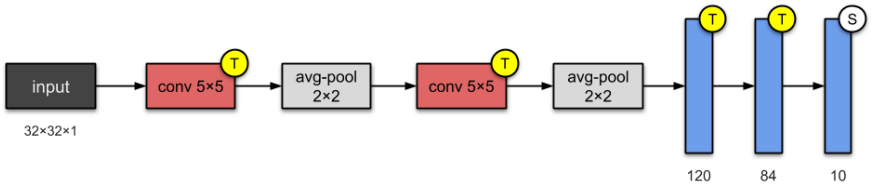
Bioinformatics has benefited from the tremendous progress in Machine Learning (ML) for several reasons. Indeed, the common issues encountered in biology and medicine often find an efficient solution thanks to computational methods involving ML. In the particular field of 3D genomics, Deep Learning (DL) - a subclass of ML - has become an inescapable tool for detecting complex patterns in very large datasets, like the sequence specificity of DNA-binding proteins, gene expression, methylation states, etc [43]. Here, we explain at a high level two very useful architectures for genomics applications: Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). While they were introduced for very different purposes - computer vision and Natural Language Processing, respectively -, they fit well to prediction tasks on genomics data. The main reference for this development is the book *Deep Learning* published in 2016 by Ian Goodfellow, Yoshua Bengio and Aaron Courville [44]. We consider that more classical methods like Random Forest, Support Vector Machine, Logistic Regression or Multi-Layer Perceptron are common knowledge.

On the one hand, CNNs [45] are neural networks performing impressively well on grid-like topology data like images. For genomics, the input can be seen as an image for a linear DNA sequence of length L transformed into a one-hot encoded vector with dimension 4 x L for the four nucleotides. It is thus very similar to the scanning of a sequence given a position weight matrix of a motif, with the difference that the network automatically learns the parameters of this matrix. Indeed, CNNs use convolution - while slightly different than the mathematical definition - instead of general linear matrix multiplication from perceptrons. The input is convolved with a kernel to catch relevant features with three crucial benefits. First, because the kernel has a smaller size than the input, the numbers of parameters to learn is reduced and efficiency increased. This is known as sparse weights improvement. Besides, the kernel is shared across multiple positions of the input, which drastically reduces memory storage requirements. Each kernel acquires a certain function of relevance to the broader classification task, like detecting edges or

contrast, which is relevant for all sets of pixels with the same number of parameters for each convolution. This invariance in space is both efficient and very useful to find insightful patterns in the image. An example of convolution operation is displayed on Figure 1.9a. After this, some activation functions like ReLU, or pooling transformations are performed, which help catching some variations in the position of an object between inputs. For example, if an edge at a particular position is detected for some input, we may want to link it to a similar edge translated by some pixels for an other input.



(a) Convolution operation on a 2D image



(b) LeNet architecture, standard template for more complex networks. Stacks of convolution/pooling layers and ending with fully-connected layers

Figure 1.9: CNN operation and architecture, reproduced from [44]

Stacking of multiple convolutional layers in various ways leads to very efficient architectures for detecting complex invariant patterns, widely used in image classification or voice recognition. Among others, we can cite famous networks such as LeNET-5 (by

LeCun et al. [46], for digit recognition, 1998, see Figure 1.9b), AlexNet or VGG (for image classification, 2012 and 2014), or ResNet-50 (first to use batch normalization and popularized skip connections, 2016).

On the other hand, RNNs [47] are designed for sequential data, when the input at time t can be dependent on that at time $t-1$ - or positions in the domain of genomics. Indeed, for usual neural networks, the main assumption is that the input is identically and independently sampled from an unknown distribution, which makes no sense for application such as natural language processing. Like CNNs, it uses parameter sharing across several time steps to generalize learning for input of various forms. But there is no more kernel; instead, the output of the network is a function of both the current input and previous blocks. There are many ways to connect different time steps, but they all include the definition of a hidden state which conveys the information across time. The simplest operation is as follows:

$$h_t = \tanh(W \times h_{t-1} + U \times x_t + b)$$

Here, h_t is the hidden state at time t, x_t is the input, W and U are shared weight matrices between hidden states and inputs respectively, and b is a bias term. The activation function is the hyperbolic tangent function. From this hidden state, an output can be computed through classical fully connected methods (with matrix V), while the state is stored for the next time-step. The computational graph and its unfolded representation is displayed on Figure 1.10a. The main drawback of basic RNNs is the difficulty to learn long-term dependencies [48]. It can be proven by computing the gradient with back-propagation that RNNs come across the delicate issue of vanishing or exploding gradients, corresponding to the behaviour of the latter through time-steps, preventing from efficiently learning. The most effective solution to this problem is the use of gated RNNs. A famous implementation is the Long Short-Term Memory network (LSTM) [49]. The idea is to create connections between time-step that have gradients that will never vanish or ex-

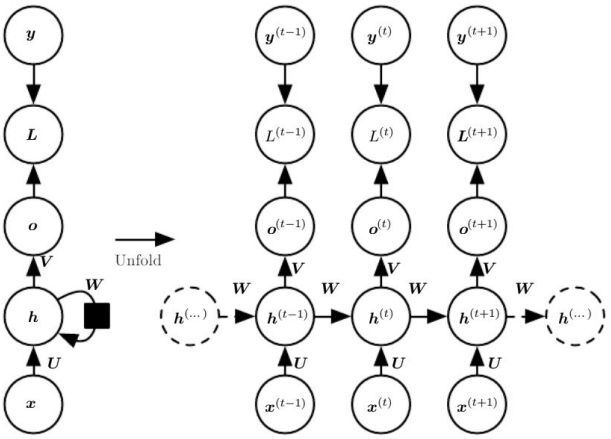
plode. For the original version, the folded module of the LSTM is no longer composed of a simple linear combination of hidden state and input, but has four different gates smartly interacting. The flow of information is controlled thanks to gates - input, forget, output - that allow or forbid it to be transferred. The equations, with the corresponding diagram on Figure 1.10b, are:

$$
\begin{aligned}
f_t &= \sigma(W_f \times [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \times [h_{t-1}, x_t] + b_i) \\
\hat{C}_t &= \tanh(W_C \times [h_{t-1}, x_t] + b_C) \\
C_t &= f_t \times C_{t-1} + i_t \times \hat{C}_t \\
o_t &= \sigma(W_o \times [h_{t-1}, x_t] + b_o) \\
h_t &= o_t \times \tanh(C_t)
\end{aligned}
$$

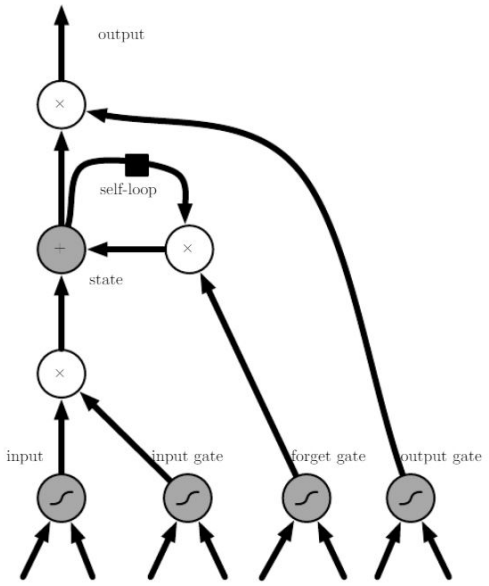
Here, the Ws are the shared weight matrices and bs the bias terms. f stands for the forget gate and regulates the information coming from previous states. Thanks to the sigmoid function, if f is close to 0, the information is not transferred, while all of it is taken into account if f is close to 1. The two next equations constitutes the input gate and controls the information from the current input. The result of these two gates is combined with the cell state C_t, and leads to an output o_t, but an output gate can shutdown it and is represented by the hyperbolic tangent in the last equation. At the end, the hidden state will undergo the same process at the next time-step. LSTMs are powerful tools to recall long term dependencies between inputs, which is particularly useful in the domain of 3D genomics if the genome is the input and we want to detect long range contacts for putative TADs.

1.3.3 Machine Learning for 3D Genomics

Because of the type and amount of data that bioinformaticians deal with, ML algorithms are increasingly used to understand the 3D organization of the genome. There are var-



(a) Computational graph of an usual RNN and its unfolded version. The hidden state h is conveying information across time-steps. x is the input, o the output, y the true label, L the loss function, and U, V, W are the shared weight matrices



(b) LSTM architecture, with more control on the information transferred. Two input gates corresponding to $i, \hat{C}$ in the equation, a forget gate f, an output gate o

Figure 1.10: RNN and LSTM diagrams, reproduced from [44]

ious sub-domains where ML finds an application, and we will focus on the prediction of protein-binding sites and TAD classification. For both of them, compared to classical computational tools, the real added value lies in the fact that the DNA sequence is often the input data, which is easy to find and deal with. Conversely, biological experiments like Hi-C, ChIP-Seq are very costly and their accessibility depends on the species. For the task of predicting protein-binding sites, CNNs achieve great results by one hot encoding the DNA sequence of length L to get an input of size $4 \times L$ that is read as a 1D image by the convolutional layers. The results are compared to the known binding motifs from experiments, but the trained paramaters of kernels can also be interpreted to detect new unknown motifs, enhancing both the performance and the understanding. Many remarkable studies have been carried out, but we can cite *DeepBind* from Alipanahi et al. in 2015, or *Basset* from Kelley et al. in 2016, which got more than 0.9 average AUC score over all DNA-binding proteins, for multiple line cells [50, 51]. Alternatively, Shen et al. used a type of gated RNN called Gated Recurrent Unit (GRU) with a sequence embedded

thanks the NLP tricks which transfroms a string into an array [52]. They had even better results with AUC score greater than 0.95 for the same data as *DeepBind*. Overall, these three tools outperformed classical algorithms scanning chunks of DNA sequences as k-mers and comparing them to the position weight matrix of the motif. Nevertheless, we should note that such scores remain quite poor in a setting were in a genome, 99% of the regions are negative examples. Thus the level of False-Discovery rate remains very high, which is a crucial feature for the interpretation of the results of our study.

An other kind of methods focus on genome-wide chromatin looping prediction, but depends on other features than DNA sequence. For example, Huang et al. developed in 2015 a model called Bayesian Additive Regression Tree (*BART*) to predict TAD boundaries thanks to histone marks from ChIP-Seq experiments, as there are a strong clue in favor of the formation of such loops [53]. The model consists in an ensemble method of regression trees that discriminates the continuous values of histone tracks, and achieved an AUC score of 0.774 on human IMR90 cells. Al Bkhetan and Plewczynski proposed *3DEpiloops* in 2018 to map all chromatin interactions at genome-wide level, based on many epigenomics data [54]. Once again, ensemble methods - here on classification task - with non deep learning models were implemented and reached 0.81 AUC score when compared against annotated Hi-C loops. A last interesting example introduces *EAGLE* by Gao and Qian in 2019, which has the particularity of focusing on a small number of processed features to predict enhancer-gene interactions [55]. These features are obtained by computing various scores from experimental data, and classical ML classifiers were trained. For instance, they defined the enhancer activity and gene expression profile correlation (Pearson correlation coefficient), based on RNA-seq measurement for gene expression, and on multiple high-throughput datasets for enhancer activity. This kind of human-built features helps the classifier to better perform and highlights the role of feature engineering.

Finally, a new paradigm tries to combine the two types of methods described above. Indeed, innovative algorithms try to detect patterns like chromatin looping based on DNA sequence only. The goal is to get rid of the dependency of Hi-C experiments, limited to a small number of species. The predictor would be able to map the 3D structure of the

genome given its sequence only, as it has been suggested that the assembly of domains is hardwired in the linear sequence - like CTCF for TADs [24]. Thus, because many species have their genome sequenced, the perspectives of understanding 3D structure for most of the phylogenetic tree become tangible. A probent published study higlighting this point comes from Henderson et al. in 2019, who focused on TAD boundaries prediction for fruit flies [56]. They tried different deep learning architectures, and the best performer includes a mix of CNN on one-hot DNA sequence, followed by a bidirectional LSTM (see Figure 1.11). They reached an AUC score of 0.9829 when compared with known annotations. They also put great effort in interpretability, by analyzing the kernels from the CNN to understand the motifs caught by the model. The comparison proves that deep learning can, at cost of sometimes complex architectures, retrieve characteristic footprints in DNA sequence, like the CTCF binding motif, and even detect new interesting transcription factors. Nevertheless, this study cannot really be compared to ours focusing on mammals, because of the different mechanisms involved in TAD formation. In mammals, TADs are larger and may be harder to detect, leading to a lower performance.

1.4 Thesis Outline

The background provided in last sections highlights that the 3D organization of chromatin can be inferred from sequence-level features. In this thesis, we propose a deep learning approach combining transcription factor binding sites and phylogenetic data to predict TAD boundaries for mammals. The objective of such a model is to get rid of the dependency on rare and costly Hi-C experiments to map the genome of species where this kind of data is not available. It takes inspiration from from the methods presented in sub-section 1.3.3 with ML algorithms based on specific processed features from DNA sequence, but differs from them since we take advantage of the links between species across the phylogenetic tree to retrieve TAD boundaries consistently with evolution rearrangement scenarios. We introduce various deep learning pipelines, each one of them reaching a particular scope of biological application depending on the type of data available to the user. We demonstrate

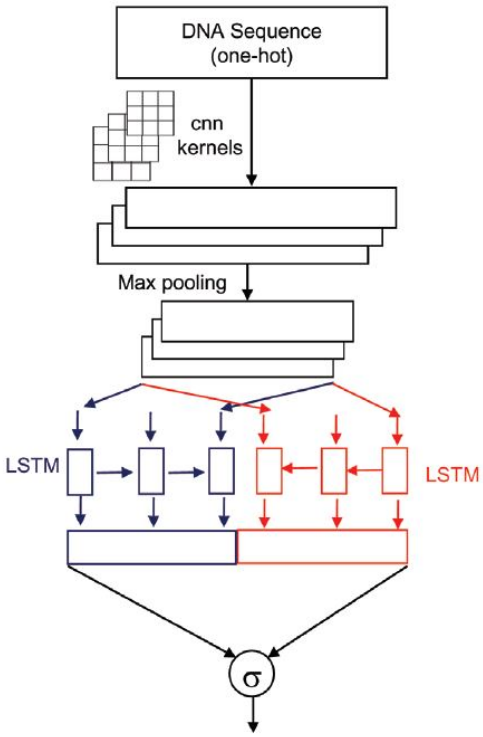


Figure 1.11: Best architecture to predict TAD boundaries from DNA sequence only. CNN followed by a bidirectional LSTM, before fully connected layers, reproduced from [56]

how sequence-based features are a precious yet not self-sufficient source of information to detect TADs. Our approaches achieve higher performance when combined with a cross-species comparison, proving the benefit of phylogeny for 3D genomics.

Chapter 2 of this thesis will be submitted shortly for publication in a bioinformatics journal. Amaury Leroy contributed to the design of the study, implemented the computational analyses, and wrote the manuscript. Mathieu Blanchette conceived the study, coordinated the computational analysis, and helped draft the manuscript.

Chapter 2

Sequence-Based Prediction of Topologically Associated Domains Enhanced by Cross-Species Comparison

Abstract

Background: Topologically Associated Domains (TADs) are functional units of condensed chromatin with a length of around 1Mb, crucial for gene regulation and disease development. We propose 10 machine learning pipelines composed of both a Long Short-Term Memory (LSTM) network and a cross-species comparison algorithm to predict TAD boundaries using features derived from genomic sequence and phylogeny.

Results: Our results are shown to be accurate across 5 mammals, even more when predictions are updated with the phylogenetic algorithm. This still unexplored approach is promising for a better understanding of TADs without Hi-C data.

Keywords: TADs, 3D genomics, Deep Learning, LSTM, Phylogeny

2.1 Introduction

Because of its crucial role in the mechanisms of cell dynamics, the three dimensional conformation of DNA is a key yet unheralded topic attracting growing interest in biology. First of all, chromosome folding of eukaryotic genomes is essential for bio-mechanical reasons, as an aggregated 2-meter-long molecule must fit into a micron-scale nucleus. In addition, the 3D structure is the result of a non-random process, intrinsically linked to essential biological functions like gene regulation [4]. Innovative approaches, gathered in the so-called 3C techniques, have uncovered general features of genome organization. In particular, Hi-C experiments, introduced by Lieberman et al. in 2009, revealed genome-wide cluster of DNA contacts, allowing us to better understand this relationship between architecture and function [15]. In mammalian genomes, DNA wraps twice around histon octamers to form nucleosomes of 147bp, reducing the length seven-fold. Some proteins can then bind to specific DNA loci to shape a complex called chromatin. Through this phenomenon, the emergence of loops can either enhance or disrupt physical contacts between regulatory elements, influencing the fate of gene expression. At a scale of hundreds of kilobases, packs of loops can bridge larger structures called TADs, allowing recurring bounds in open chromatin [9]. These domains are partitioned into compartments, classified as ”A” - for active chromatin - or ”B” - for repressive chromatin. Finally, territories define portions of the nucleus occupied by chromosomes with preferential and functional long-range interactions.

TADs were first identified by Dixon et al. in 2012, thanks to Hi-C experiments on human and mouse genomes [18]. Following this study, similar domains were found in various species, both in mammals or other phyla [30]. That ubiquity advocates for their importance as functional and conserved units of the genome, defined as regions inside which chromatin loci interact more frequently than with regions located in adjacent domains. With a median size of 880kb, they are made of condensed loops defining regulatory landscapes. Therefore, understanding their biological formation is crucial. It lies in the joint action of several chromatin marks binding to precise loci and among them-

selves, which are then enriched at TAD boundaries. In mammals, the complex CTCF-cohesin has emerged as an essential piece of this process, for which 76% of boundaries stem from [9,24]. In this so-called loop extrusion model, chromatin is extruded by cohesin until two distal and convergent CTCF proteins bind and create a border [57]. This results in stable domains, which are the cornerstone of essential functions such as transcription. As a consequence, a negative selection is applied to disruption of boundaries, going hand in hand with their conservation across species. The breakage of a boundary due to a mutation can reorganize the 3D structure, lead to gene misexpression and the development of diseases [29]. Nevertheless, some rearrangement scenarios can provide favourable functions and make new phenotypes arise. It has been found that evolutionary breakpoints are enriched at boundaries [32,33], and that TADs are mainly conserved across close species - more than half of TAD boundaries are shared between human and chimpanzee [31].

Hi-C contact maps are by far the most used source of data to infer TAD structure, with an impressive amount of available computational tools (for reviews, see [41, 42]). They all try to solve the same task but predictions vary greatly between methods, depending on the quantitative method or the resolution used. Nevertheless, the major drawback is that Hi-C data is only available for a small number of species because of the substantial cost of such experiments. In light of previously exposed arguments, advocating for an assembly of TAD structure hardwired in the DNA sequence through specific conserved chromatin marks, new methods have successfully started to detect 3D patterns based on features extracted from sequence genome only - a much accessible source of information. Some of them apply ensemble machine learning pipelines on epigenomics data to predict enhancer-gene interactions [55], chromatin contacts [54] or TAD boundaries [53]. In 2019, Henderson et al. even proposed a compelling approach based on DNA sequence only for prediction of TAD boundaries in fruit flies, with deep convolutional and recurrent neural networks [56]. Nevertheless, the precise mechanisms of TAD formation are still unclear, making their inference a difficult task. Even if the presence of complex CTCF-cohesin is a precious hint for the existence of boundaries, many of the latter obey to unknown rules involving various transcription marks.

An interesting and yet uninvestigated paradigm is to take advantage of from evolutionary data. Because they represent a functional unit of the genome, TADs are mainly conserved across species, and we can hope that divergent ones can even be inferred by the knowledge of rearrangement scenarios of syntenic blocks, in addition to sequence-features data. Initial predictions can then be enhanced by the comparison with close-species for which TADs have been either referenced or predicted, thanks to a probabilistic model inspired from the work of Fitch, Sankoff or Felsenstein [37, 38, 40]. That still unexplored approach is a main feature of the proposed study.

Here, we describe 10 machine learning models for the prediction of TAD boundaries in mammals, which are enhanced by cross-species comparison either during or at the end of the learning process. They differ in the input they require: either single genome sequence, multi-species genome sequences, or referenced TAD boundaries extracted from Hi-C data for known species. Their use depends on the data available to the user and type of study carried out. We demonstrate how sequence-based features are a precious yet not sufficient to detect TADs with high accuracy. Our approaches achieve higher performance when combined with a cross-species comparison, proving the benefit of comparative genomics for DNA conformation in 3D.

2.2 Material and Methods

TAD boundaries and sequence features acquisition

The scope of our study focuses on mammals, where TAD mechanisms have been relatively well characterized. As often as possible, we tried to start from raw data on which we performed *in silico* experiments to get rid of the dependency over processed biological marks, which means that we processed sequence genome to extract our features such as transcription factors or TAD labels. More precisely, we used data of liver cells from

five species: human (hg38), mouse (mm10), rabbit (oryCun2), macaque (rheMac2) and dog (canFam3). With the exception of hg38, raw Hi-C reads were extracted from Rudan et al. [35] to ensure consistency when comparing results. We then processed them with software from the Aiden lab to produce 5kb-resolution normalized Hi-C contact maps, using Juicer to create interaction frequency matrices [58], and Straw to process them [59]. For human, we did not find the same cell type but chose liver cancer cell line HepG2 instead. The already processed Hi-C data set was available thanks to the work of Dekker Lab as part of the 4D Nucleome Project [2]. These 5kb resolution Hi-C contact maps were the input for the TAD boundary caller RobusTAD [60], which was chosen because of its robust behaviour with respect to variations in coverage, resolution and noise level. With default parameters and for these five mammals, outputs of RobusTAD are scores for each bin, related to the likelihood to host a TAD boundary, which were used either as features or labels for our machine learning models. It is important to note that RobusTAD distinguishes left and right boundaries, which is very useful for TAD reconstruction - our task was also dual as outputs of the models were a pair of probabilities. For RobusTAD, such differentiation is made possible by a computation on the contact counts between the bin of interest and its antecedents - hint for a right boundary if this count is high, and vice versa between the bin of interest and the ones after for a left boundary. For sequence features acquisition, we scanned the genomes of each species with a Perl script and HOMER software [61], which accurately returns bins corresponding to a positive binding score of 335 transcription factors based on their Position Weight Matrices (PWM). We counted the predicted sites in each 5kb kin, consistent with the resolution of TAD boundaries, to produce features for each of the 335 transcription factors. The last step of the processing was the alignment of genome bins into consistent blocks between species. LiftOver was used with default parameters to map all the features and labels to the reference genome assembly mm10 [62]. This mapping is made possible thanks to whole-genome alignment chains, coming from UCSC Genome Browser (ref). Eventually, the full data set, processed in python [63] with tools from SciPy library [64], was made of 20 chromosomes of the multiple sequences genome partitioned into successive 5kb bins containing TAD boundaries (binary features) and Transcription Factor Binding Sites (TFBS) count (discrete values).

See Table 2.1 for accession references, and an example of two lines in the dataset in Figure 2.1.

Specie	Reference	Cell type	Restr. Enzyme	Accession Number
Human	4D Nucleome Project [2]	HepG2	DpnII	4DNFICSTCJQZ
Mouse	Rudan et al. [35]	liver	HindIII	GSE65126
Dog	Rudan et al. [35]	liver	HindIII	GSE65126
Rabbit	Rudan et al. [35]	liver	HindIII	GSE65126
Macaque	Rudan et al. [35]	liver	HindIII	GSE65126

Table 2.1: Hi-C data sets for five studied species

Machine Learning pipeline

The classification task involves various architectures, detailed in Results section 2.3. They depend on the data available to the user, but they all share the same building blocks, arranged in different orders. First of all, a classifier can take as input a feature vector made of the 335 TFBS counts from a given species on a specific bin and return the probability that this bin host a TAD boundary. For the whole study, we distinguished left and right boundaries so that the real output is a pair of probabilities. It is also important to note that the data set is highly imbalanced, as only 1% of the bins host a boundary. We will refer to this classifier under the same abbreviation f, even if each one of them takes input with different nature and dimension. In order to establish a baseline for our results, the classifier can naively take the form of a Random Forest, which is very simple to implement and optimize. We used the library Scikit-learn to perform our experiments with this pipeline [65]. Nevertheless, this kind of ensemble method consider the different bin examples as independent and identically distributed. That could limit the performance of the predictions because TADs have a characteristic size of L base pairs, so that a boundary is more likely to be found close to L base pairs after or before another one. To capture this idea, we can see the different bins as a sequence on which we can apply a recurrent neural network, taking into account the distal dependency between bins. More precisely, we chose to implement a Long Short-Term Memory (LSTM, [49]) model, as LSTM units have

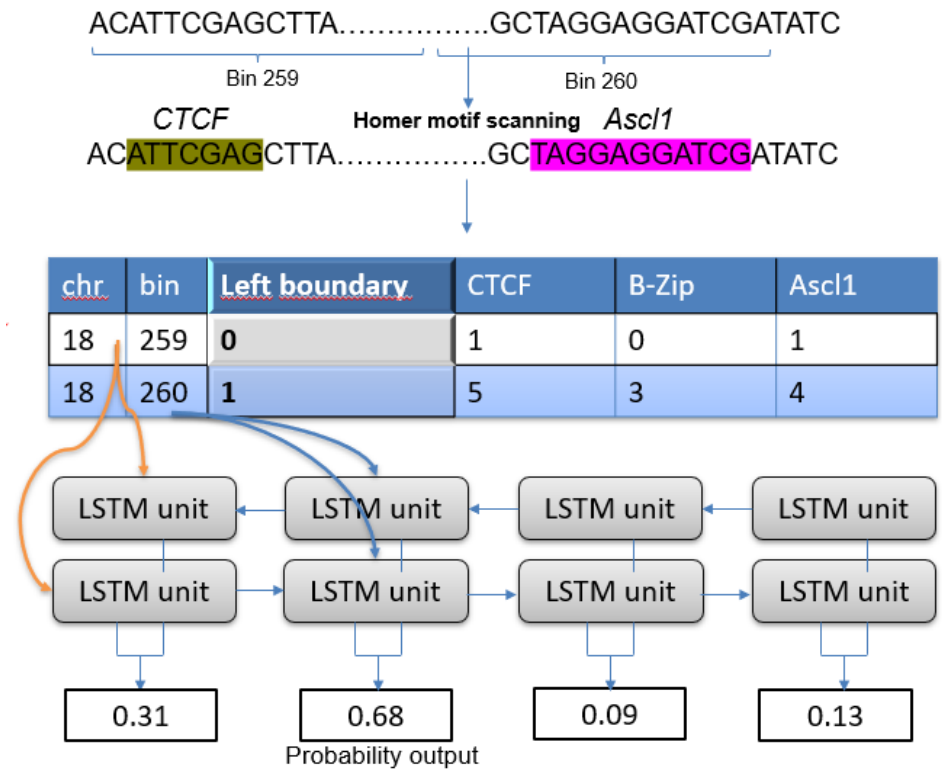


Figure 2.1: LSTM classifier pipeline with annotated genome as input. In this example, left boundary is the label, and we represent two typical rows corresponding to consecutive bins in chromosome 18, with predicted TFBS counts being fed to the bidirectional LSTM. The output is the probability for each bin to host a left boundary.

a great ability to recall distant states, and we made it bi-directional because the direction of scanning is independent from the presence of boundaries. Figure 2.1 summarizes the pipeline from raw DNA sequence, transformed into annotated genome, to LSTM predictor f. We set up LSTM units in the "many-to-many" way, inputs being a pool of 2000 ordered bins and outputs being the prediction to host a TAD boundary for each bin - left and right boundaries are separate outputs. It then deals with 10Mb-long sequences, which is an interesting trade-off between computational time and the necessity to have sequences long enough to embed several TAD boundaries and learn long range dependencies. Besides, considering a sequence longer than a chromosome would be deleterious as there is TADs cannot straddle multiple chromosomes.

We used a binary cross entropy loss function, with a larger weight ω on positive predictions - the inverse of the proportion of such labels - to overcome the imbalance of the data set. The loss function for predictions $f(x_i)$ and labels y_i, eluding the regularization term, is defined as

$$\frac{1}{N}\sum_i \omega y_i \log(\sigma(f(x_i))) + (1 - y_i \log(1 - \sigma(f(x_i))))$$

For cross validation, We chose to split the data set into entire chromosomes. Concretely, the first four chromosomes are for the training set, the fifth is for the validation set and the sixth is for the test set - and so forth for the remaining chromosomes. For both Random Forest and LSTM models, there are many hyper-parameters to fine-tune. For Random Forest, we can point the number of estimators, the depth of each tree, the criterion to assess the weight of a decision, and the minimal number of samples to split a node. For LSTM, we had to take care of the different dimensions of the architecture, especially the number of layers and the dimensions of hidden and cell state vectors, but also the classical parameters for a neural network, such as the batch size, learning rate, optimizer or regularization. We used the library Hyperopt [66] to carry out a Bayesian - not Gaussian - optimization, using Tree Parzen Estimator as a surrogate function. We justify this choice with both the ease of use of the library and the performance of the optimization algorithm, although a Gaussian method would have performed just as well. The most important hyper parameters had different optimized values depending on the final architecture. To find them, we ran Hyperopt with a 5-fold cross-validation pipeline to ensure robustness of the results obtained.

Phylogenetic algorithm - PhyloTAD

The second major building block is a phylogenetic algorithm that we call PhyloTAD. Its purpose is to update and enhance the accuracy of the initial predictions made by a predictor f on sequences from multiple species, to make them more consistent with the phylogenetic tree. Given the hypothesis that TAD breakage rarely occurs, sharing knowledge

about neighboring species can be of great help to infer the presence of a TAD on the target species. In order to compare predictions between nodes, the algorithm is inspired from Felsenstein's work [40] on the maximum likelihood of a tree for biological characters. For the sake of clarity, Figure 2.2 displays the general idea of this algorithm for a given bin, with the following notations:

- The features contained inside this bin can be embedded in a discrete vector of fixed size called C_u for species u.
- A classifier f is supposed to return the probability of hosting a TAD boundary given these features, defined as $\Pr[T_u = 1|C_u]$ with T_u the binary random variable representing the presence of a boundary for species u.
- The binary phylogenetic tree links these species across evolution, with the five modern mammals at leaves
- Some prior knowledge thanks to statistics on the data set, like the prior probability of hosting a TAD, i.e. $\Pr[T_u = 1]$

Like for Felsenstein's algorithm [40], the labels at leaves are first pooled to infer labels of ancestors up to the root during the forward step. Then, during the trace-back step, information goes from root to leaves taking into account the constraints inherent of the tree, in order to finally update the labels at leaves. The goal is to transform $\Pr[T_u = 1|C_u]$ into $\Pr[T_u = 1|(C_1, C_2, C_2, C_4, C_5)]$ which is a good representation of how evolution influences TAD structure.

The whole derivation of the model is available in the Supplementary Methods section 2.5. The idea is to find a dynamic programming method to pass information from the leaves to the internal nodes until the root, and an other method to make the reverse path. The quantity ruling the forward step is introduced by the function X_u defined at node

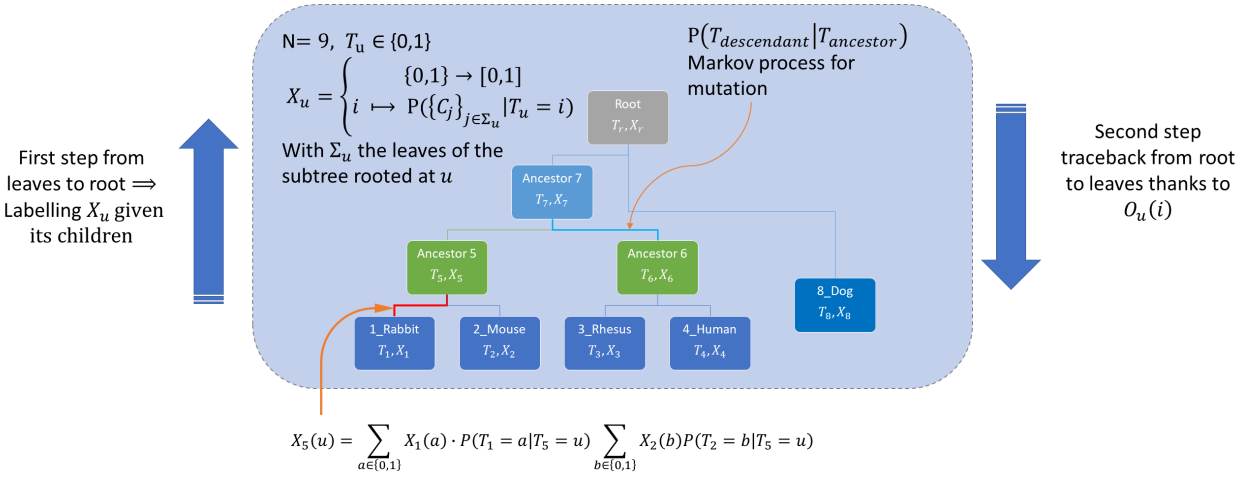


Figure 2.2: General architecture of node updating with PhyloTAD algorithm

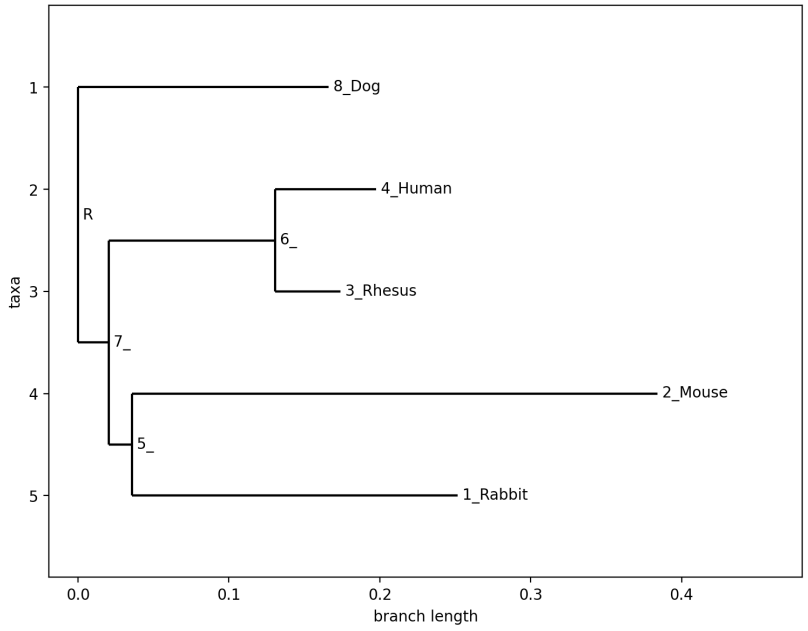


Figure 2.3: The phylogenetic tree used for this study, consisting of 5 mammals, from [67]

$u \in [1..N]$ ($N = 9$ is the number of nodes, or species in the entire tree) as:

$$X_u : i \in \{0, 1\} \mapsto \Pr[\{C_j\}\, j \in \Sigma_u \mid T_u = i]$$

for Σ_u the set of leaves in the sub-tree rooted at u. At leaves, this definition becomes:

$$X_u(i) = \Pr[C_u \mid T_u = i]$$

which is merely what the output of the predictor provides with Bayes' rules:

$$X_u(i) = \Pr[T_u = i|C_u]\frac{\Pr[C_u]}{\Pr[T_u = i]}$$

Given some assumptions on independence, if u is an internal node with children v - left - and w - right -, we find that:

$$\begin{aligned} X_u(i) = &\sum_{a\in\{0,1\}} X_v(a) \Pr[T_v = a \mid T_u = i] \\ &\times \sum_{b\in\{0,1\}} X_w(b) \Pr[T_w = b \mid T_u = i] \end{aligned}$$

In a more compact formulation, we can write $X_u(i) = L_u(i)R_u(i)$ with L_u (respectively R_u) for the nodes of the left (respectively right) branch respectively. The term $\Pr[T_v = a \mid T_u = i]$ constitutes the probability of evolution across a generation, following a Markov process that we represented by a transition probability matrix defined as $M = \exp(Q.t)$ with Q the mutation rate matrix - hyper-parameter - and t the branch length. A dynamic programming algorithm is then able to reach each ancestor given information from its children. When the root is reached, the algorithm retraces the path in the opposite direction to find for each leaf u : $\Pr[T_u = i \mid \{C_j\}_{j\in[1..5]}]$. To do so, we introduce a new set of functions:

$$O_u(i) = \Pr[\{C_j\}_j\, j \in\ \Omega_u \mid T_u = i]$$

for Ω_u the set of leaves not under the subtree rooted at u. In other words, $O_u(i)$ is the probability of having all the C_j which are strictly not under u, conditioned by the label at the node of interest. At the root, because there is a definition issue, we take the convention that $O_u(0) = O_u(1) = 1$, justified afterwards. At leaves, $O_u(i) = \Pr[\{C_j\}_{j \neq u} \mid T_u = i]$ which is very close to what we want as an output thanks to Bayes' rules:

$$\Pr[T_u = i \mid \{C_j\}_{j \in [1..k]}] = X_u(i) O_u(i) \frac{\Pr[T_u = i]}{\Pr[\{C_j\}_{j \in [1..k]}]}$$

This quantity is increasing as the node is closer to a leaf, thus we can dynamically compute every O_u from root to leaves given the previous calculation of all $X_u s$. The formula to pass from a node to its child, where u is the left child and $p(u)$ its parent, is as follows:

$$O_u(i) = \sum_{a \in \{0,1\}} \Pr[T_{p(u)} = a \mid T_u = i] R_{p(u)}(a) O_{p(u)}(a)$$

This formula is valid when the child is at the left position of its parent, and must be modified by switching from $R_{p(u)}$ to $L_{p(u)}$ when the child is at the right position - symmetric calculation. The convention of a probability equal to 1 at root ensures that the parent is not contributing to the calculation on its child when at root, because we only want to look at its brother which gathers all the leaves not under the running subtree. In that respect, all quantities are known thanks to the first part of the calculation - from leaves to root - and the current knowledge on O_u for parents. We extracted the phylogenetic tree linking the five mammals from the platform Interactive Tree of Life [67], and used the library Biopython to represent it internally [68]. The dynamic algorithm is implemented through a post-order traversal of the tree for the forward step, and a pre-order traversal for the backward step. For each bin, the overall complexity is linear in both space and time, ensuring a scalable application in our classification task with multiple bins.

2.3 Results

Here, we propose three global approaches - resulting in 10 models - for TAD boundary prediction, this partition corresponding to the input data required for the trained model: single genome sequence, multi-species genome sequence, or multi-species TADs.For all figures below, which describe the corresponding pipelines, we did not distinguish the baseline classifier - single-bin prediction with Random Forest - and the LSTM predictor - multi-bin prediction in the form of a sequence; they both refer to f. For the metrics, we chose AUC, which gives robust insights about the behaviour of the model, independent from any decision threshold. For all tables, AUCs are obtained by computing the average between left and right TAD boundary predictions. As LSTM always performed better than the baseline, we only displayed its results.

Single genome sequence as input

Many mammals were not studied for Hi-C experiments, and only the DNA sequence is available. In this case, we can still get partial information about TAD formation mechanisms since the data available to the cell is similar but much more complete and contextualized when the genome is folded. Then, a basic classifier f can learn from those marks for a single species to infer boundaries, without any knowledge on phylogeny. The first naive model, called "Transfer", is trained only on one species for which TADs are referenced, with computational predicted TFBS for 335 transcription factors as features. The labels are a pair of binary results, corresponding to the presence or absence of a left/right boundary on the bin of interest. We can then detect TADs for a new species by transferring the trained parameters for a set of features composed of the marks of a single target species, hoping that the mechanisms of TAD formation are similar between the species for training and for testing.

A more complex model, called "Pooling", is trained on multiple species where Hi-C data is available. The sets of {TFBS counts, label} pairs across species are pooled to create a bigger data set - the number of features is the same, but the number of examples increases 5-fold. The advantage of such a training is that, in addition to simulating a data augmentation, it implicitly takes into account the differences of TAD formation between mammals, preventing from overfitting and making the classifier more flexible to a new target species.

The last model taking as input a single genome only is species-specific: it is only applicable to species for which Hi-C data is available for training, therefore it does not represent a very useful tool for practical application. However, such an architecture may yield insights into the differences of mechanisms for TAD formation between species. Figure 2.4 highlights those three models, and results are summarized in Table 2.2.

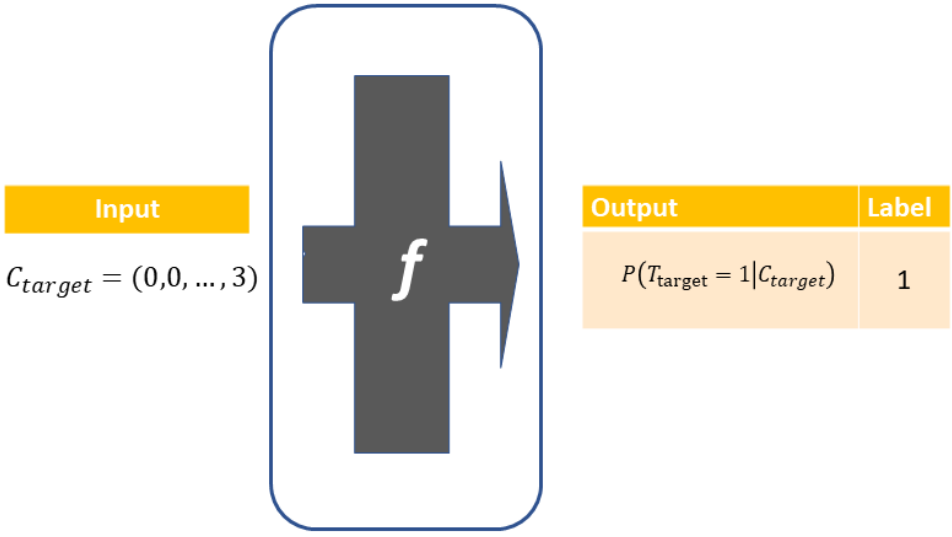


Figure 2.4: Pipeline for unified models with single genome sequence as input. The training is carried out either with a unique species (Transfer and Species-specific), or with the pooled set of all species for which Hi-C data is available (Pooling). TAD inference is then made possible by applying those models on a single annotated genome for the target species.

As expected, when pooling examples from different species, the overall results are improved compared to the "Transfer Model". The predictor f learns partial information about the sequence, with a AUC score around 0.7 for "Transfer Model", except for

Model	Mouse	Human	Dog	Macaque	Rabbit
Transfer	0.71	0.70	0.70	0.62	0.71
Pooling	0.73	0.73	0.74	0.63	0.73
Species-specific	0.71	0.71	0.72	0.63	0.72

Table 2.2: Area Under Curve (AUC) on testing set for the three first models and the five mammals. We trained the first model with TFBS from mouse training set, and transferred the trained parameters to other species

macaque, with disappointing results on many models, which is likely due to a poorer quality of data. For each species, "Pooling Model" has an accuracy around 2% higher, which is a small but comforting improvement to state that mammals have common mechanisms for TAD formation, and that a mutual training has to be prioritized. It is confirmed with the "Species-specific Model", which is better than the "Transfer Model" because TFBS from the target species have been learnt, but worse than the "Pooling model". It seems that increasing the number of examples, even if they do not correspond to the same specie, is valuable to the learning because of the common mechanisms of TAD formation. Nevertheless, the different AUC scores show that transcription factors only are not a sufficient source of data to accurately map TADs genome-wide.

Multi-species genome sequence as input

The second scenario is to consider that the we have DNA sequence from multiple species, including the target one. That is very plausible since genome sequence is available for a large number of species, and we only need to predict TFBS with Homer and align syntenic blocks with LiftOver - see Methods. It yields to the development of two models that could fairly be compared to the previous ones since only sequence information is used. The first approach consists in a concatenation, at each bin, of the TFBS feature vectors obtained at orthologous regions across all 5 species, allowing the predictor to learn by itself phylogenetic relations between them. The input dimension being fixed, it requires to keep the same species between training and testing, leading to issues when trying to

apply the model for practical use. The architecture of this approach is displayed on Figure 2.5.

A smarter approach is to take advantage of the insights from models with single genome sequence as input, in particular the pooling method which performs the best. We can apply those models for each species in our data sets, yielding to a multi-species prediction for each bin. To use evolutionary information, this set of predictions can be, on the one hand, the input of PhyloTAD. The goal is to update them, consistent with the relative distances between species. Species need to be phylogenetically placed in relation to others. On the other hand, the prediction outputs can be the input of a new classifier f', with a fixed input dimension corresponding to the number of target species. Then, the phylogeny is not explicitly mentioned and we can expect the network to implicitly learn it by itself, smoothing the independent results from f applied to distinct mammals. Compared to PhyloTAD which is deterministic, it has the advantage to need less knowledge on the relations between species, but requires a specific training for each new experiment, yielding a less flexible application in practice. The corresponding pipelines are schematized in Figure 2.6, and results are summarized in Table 2.3.

Model	Mouse	Human	Dog	Macaque	Rabbit
Concatenation	0.72	0.72	0.73	0.62	0.72
Pooling + PhyloTAD	0.77	0.79	0.78	0.71	0.78
Pooling + Smoothing f'	0.75	0.77	0.78	0.69	0.77

Table 2.3: AUC on testing set for multi-species models

Results from these three models are very informative about the benefits of phylogeny. First, the concatenation model performs very similarly to the pooling method with single genome sequence as input, with AUCs around 0.72 except for macaque. We can speculate that phylogenetic relationships are partially learnt by the predictor f without any knowledge required, which is more user-friendly as there is only one block in the pipeline. However, such a model is not really applicable to practical TAD inference as both the training and testing are needed on multi-species, while the pooling approach, once trained on known mammals, can be applied on the genome sequence of the target species only. The

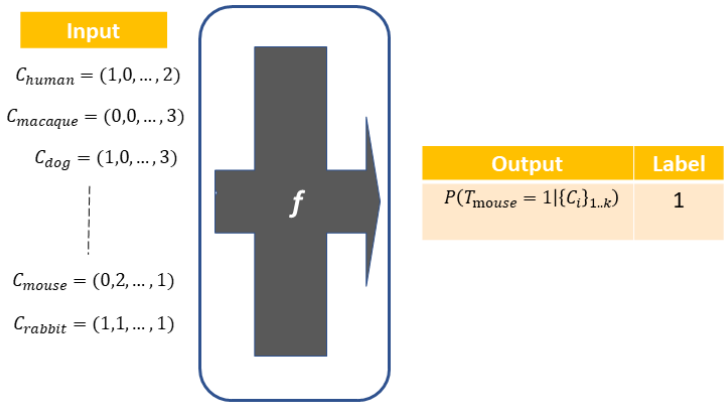


Figure 2.5: Concatenation model for 5 species, including the target species - mouse. The number of features is the number of TFBS times the number of species. Both training and testing must be carried out with the same sample of species to ensure consistent results

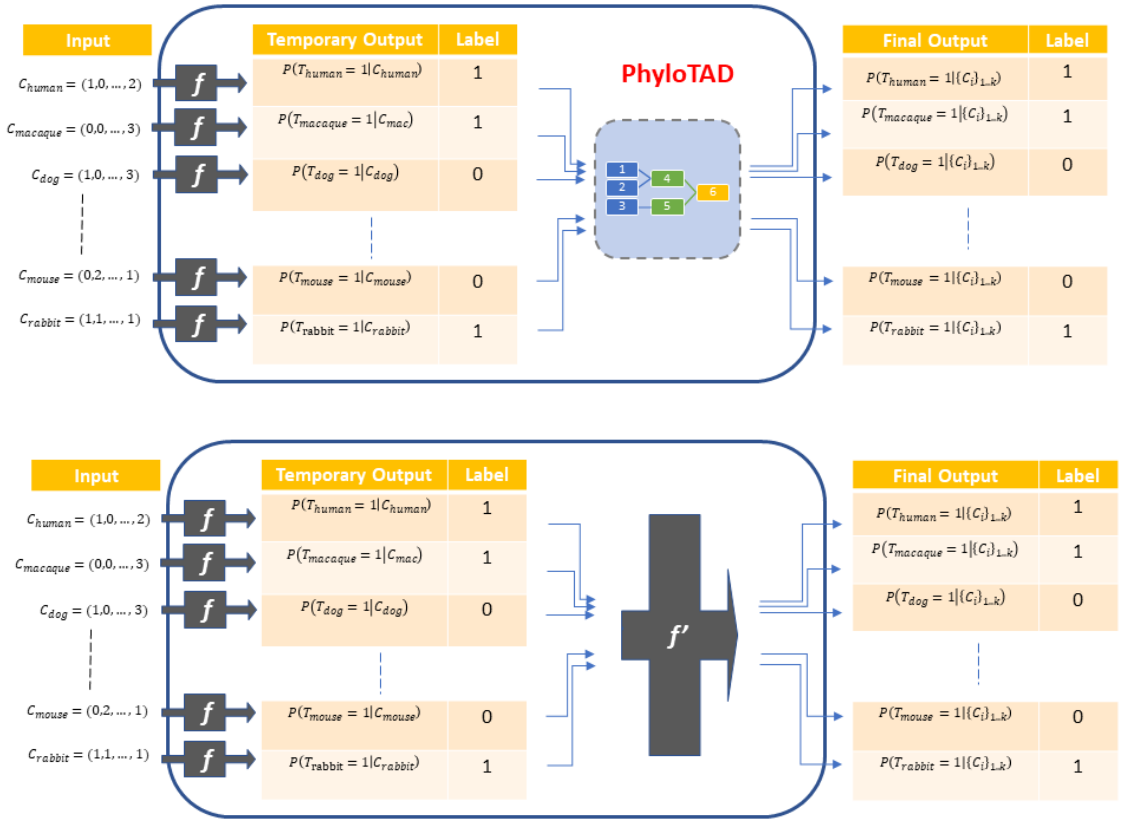


Figure 2.6: Pipelines combining general mammals predictor and phylogenetic tree to infer TAD boundaries in multiples species where no Hi-C data is available. The classifiers f are trained separately with specific genomes, and the results are merged with either PhyloTAD or a smoothing predictor f' to update temporary outputs

second method merging pooling from previous models with phylogeny is much more interesting, as it involves shared learning of various general trained classifiers for mammals, applied on new species. Here, the temporary predictions are updated either with a deterministic algorithm or with a new neural network. The first method is very efficient as it does not require any training, and generally performs better than the smoothing classifier f' - around 2% increase in AUC. Overall, for these two similar approaches, predictions are significantly improved compared to the pooling method before update. Indeed, we reach AUC close to 0.8 - an increase of more than 6%.

Multi-species TADs as input

A new paradigm consists in considering directly TAD annotations from Hi-C data from a certain subset of mammals to transfer them to a broader set of species where Hi-C data is unavailable. Very basically, our LSTM, with input dimensions fit to the input features, can infer boundaries from labels of neighboring species. This model requires to be trained on the species where Hi-C data is available. A second possible approach is to use PhyloTAD directly, which does not need training anymore. For each bin, the inputs are the labels for known species, and a prior probability for the target species - we chose the genome-wide frequency of hosting a TAD -, that will be updated. The task is less time consuming as the data needed to build the features is drastically lower, but it is important to note that we are limited by the small number of mammals with available Hi-C data, producing stereotyped and inflexible results, independent from the specificity of the target species. Both methods are displayed on Figure 2.7, with results in Table 2.4.

Model	Mouse	Human	Dog	Macaque	Rabbit
TAD concatenation	0.83	0.82	0.82	0.79	0.84
PhyloTAD only	0.81	0.82	0.80	0.78	0.81

Table 2.4: AUC on testing set for the models using TAD boundaries as input

The results presented above consider the different species where we know the true labels in order to compute the performance, but the final goal is to apply these models on

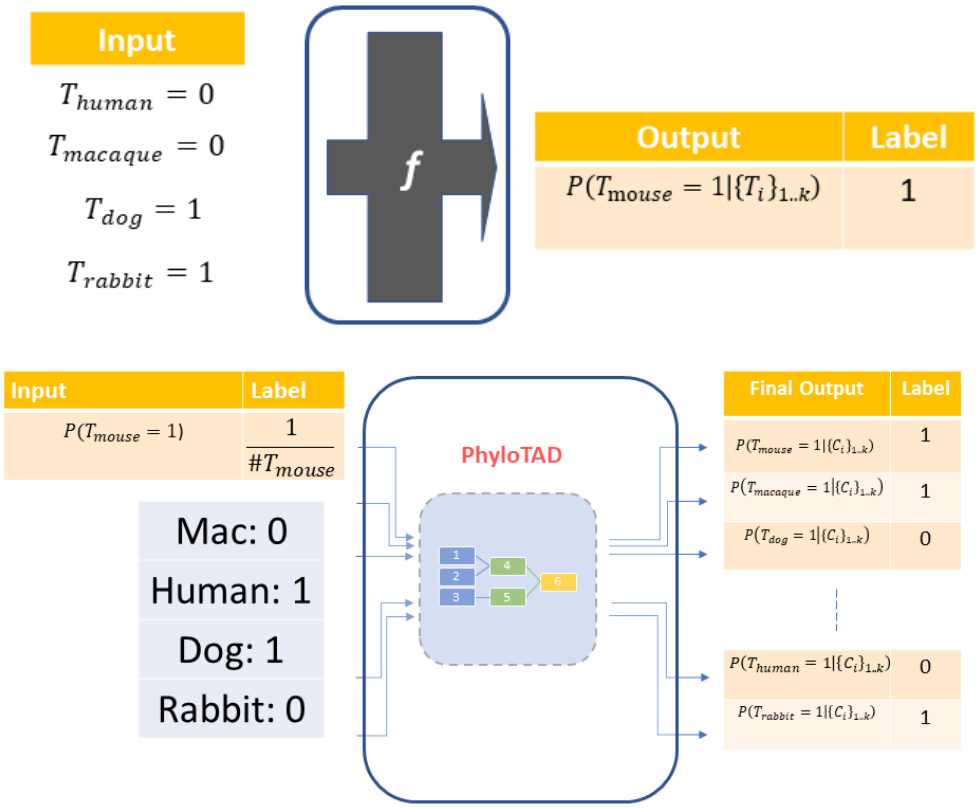


Figure 2.7: Models using TAD labels from known mammals to infer boundaries on target species, here mouse. It is done either with a deep learning classifier f - top - or by the deterministic algorithm PhyloTAD - bottom.

new target species. Unsurprisingly, TAD annotations from other species are very informative and the overall performance is better than in previous models. In addition, the gap in AUC between macaque and other mammals is partially bridged with this method, advocating for issues in genome sequence data rather than Hi-C contact matrices. Despite these satisfactory results, it is crucial to be aware that we would get the exact same results for distinct species, as the input would be fixed and independent from the target. Therefore, this solution is insightful on the role of phylogeny but hardly applicable in practice, at least on its own.

Multi-species genome sequence and TADs as input

The previous approaches advocate for the distinct and decisive contributions of both genome sequence and phylogeny for TAD inference. The final models merge both source of information to boost accuracy of predictions. The idea is to adapt the last model of Figure 2.7, replacing the naive prior by a temporary prediction from a classifier f based on sequence features. This classifier can be either species-specific, or general to mammals - the second option, "pooling", gave better performance when used alone. The other labels, fed as input for PhyloTAD, stem from mammals with known Hi-C data. Thus, we ran our experiments on a leave-one out basis, computing the performance on a target species with four other mammals as helpers to accuracy boosting. The architecture is highlighted on Figure 2.8, and results are summed up in Table 2.5, with a visualization of improvement for a portion of chromosome 3 in mouse experiment. Lastly, we tested a model - displayed on Figure 2.9 - consisting in a global concatenation of all data available, both TFBS and TAD labels for a subset of species. These input were used for a predictor f for which a high performance was expected, but with the same issues of inflexibility as the model "TAD concatenation".

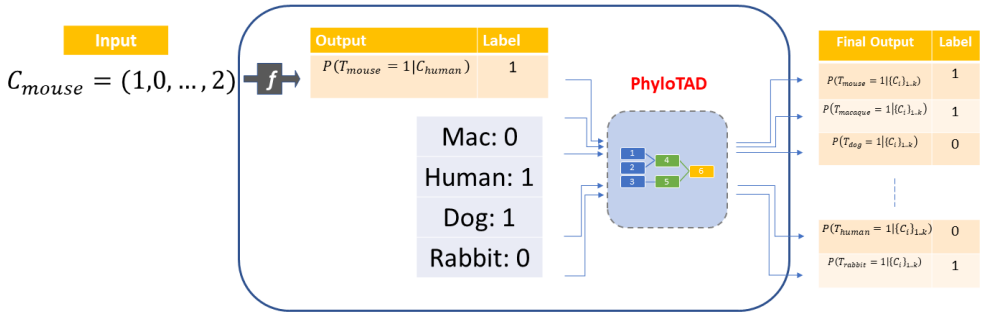


Figure 2.8: Pipeline merging inference from TFBS and update thanks to PhyloTAD for a target species, here mouse

The model for different species have better performance when combining sequence-based features with knowledge on evolutionary scenarios. The first model gives very satisfying results, as it only requires a general classifier trained on mammals taking as input a single genome sequence. In practice, PhyloTAD could take as input many more

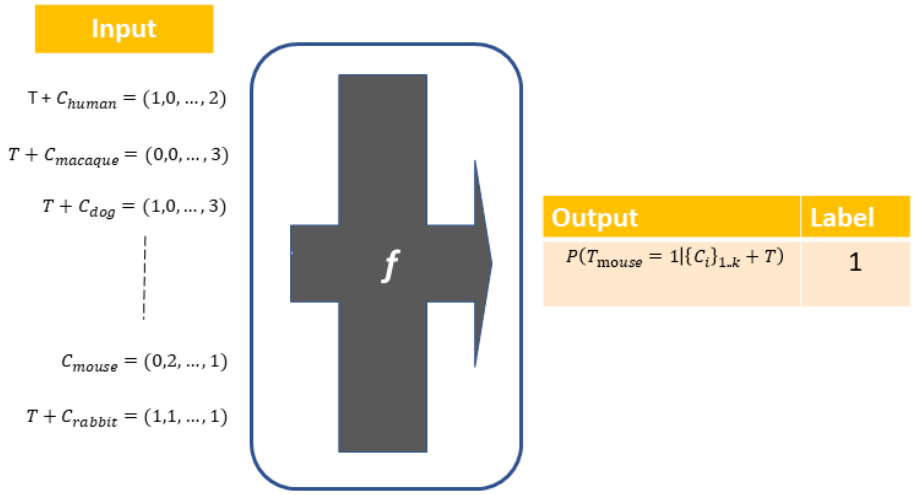


Figure 2.9: Global predictor taking all data available in a jumble

Model	Mouse	Human	Dog	Macaque	Rabbit
Pooling + PhyloTAD with labels	0.86	0.84	0.85	0.81	0.86
Global concatenation	0.89	0.89	0.90	0.85	0.88

Table 2.5: AUC - average of right/left labels - on testing set for the models using both genome sequence of the target mammals and TAD boundaries from other species as input

species, with a mix of true labels when Hi-C data is available - as experimented here, and of predictions from f when only the genome sequence is known - model called "Pooling + PhyloTAD". Therefore, it represents the most applicable approach to be easily used to infer TAD boundaries in a new species. The "Global Concatenation" model has even better AUC scores but heavier constraints on the input data.

Finally, we can make a qualitative assessment of our results for three typical models. Figure 2.10 shows the output probabilities compared to the true labels, for the mouse on a portion of chromosome 3. The top sub-figure concerns the pooling model before update, the middle one refers to the pooling model followed by PhyloTAD for all species, and the bottom one only differs because the input of PhyloTAD are true labels from related species except for the target ones. First, we notice that predicting TAD boundaries based on sequence-features only is a tough task, as the predicted TAD boundary probabilities of

the top sub-figure are very low, even if we often can recognize peaks close to true labels. After update with PhyloTAD, when sequence-based predictions from other species are also the results from f, we notice a significant improvement in the confidence of results. Peaks are getting sharper with a good accuracy, but sometimes are a few bins away from true labels, resulting in a still low AUC score. When inputs are true labels except for the target species, there is a striking increase in accuracy, peaks being very high for true labels and smoothed otherwise. Nevertheless, some true boundaries are still unrecognized by the different models, suggesting that other features, crucial for TAD formation, could be incorporated to improve the study. Indeed, these bins often correspond to sites with no CTCF site, the strongest hint for our predictor to assess the presence of a boundary.

2.4 Discussion and Conclusion

In this study, we have focused our interest on TAD boundary prediction when no Hi-C data is available for the species of interest. Such work has been motivated by the crucial importance of TADs in biological cell mechanisms, paradoxical with the poor knowledge of their locations in the genome of various species. Mapping the 3D structure of chromatin, in particular these big domains of high frequent contacts, would be a huge step in the understanding of gene regulatory networks. However, Hi-C experiments are quite expensive, thus limited to few species such as some mammals - human, macaque, dog, rabbit, mouse -, fruit flies or some plants. For mammals, many studies have converged to the establishment of an extrusion model in which a complex of proteins, including CTCF, could be the cornerstone of the formation of such large loops [24]. The hypothesis of an architecture hardwired in the genome was our first hypothesis. The second one stems from another observation, suggesting that these structures apply a strong negative selection because of their importance in gene regulation. Therefore, TADs are mainly conserved across species, giving us precious hints to recall TADs in known neighboring mammals.

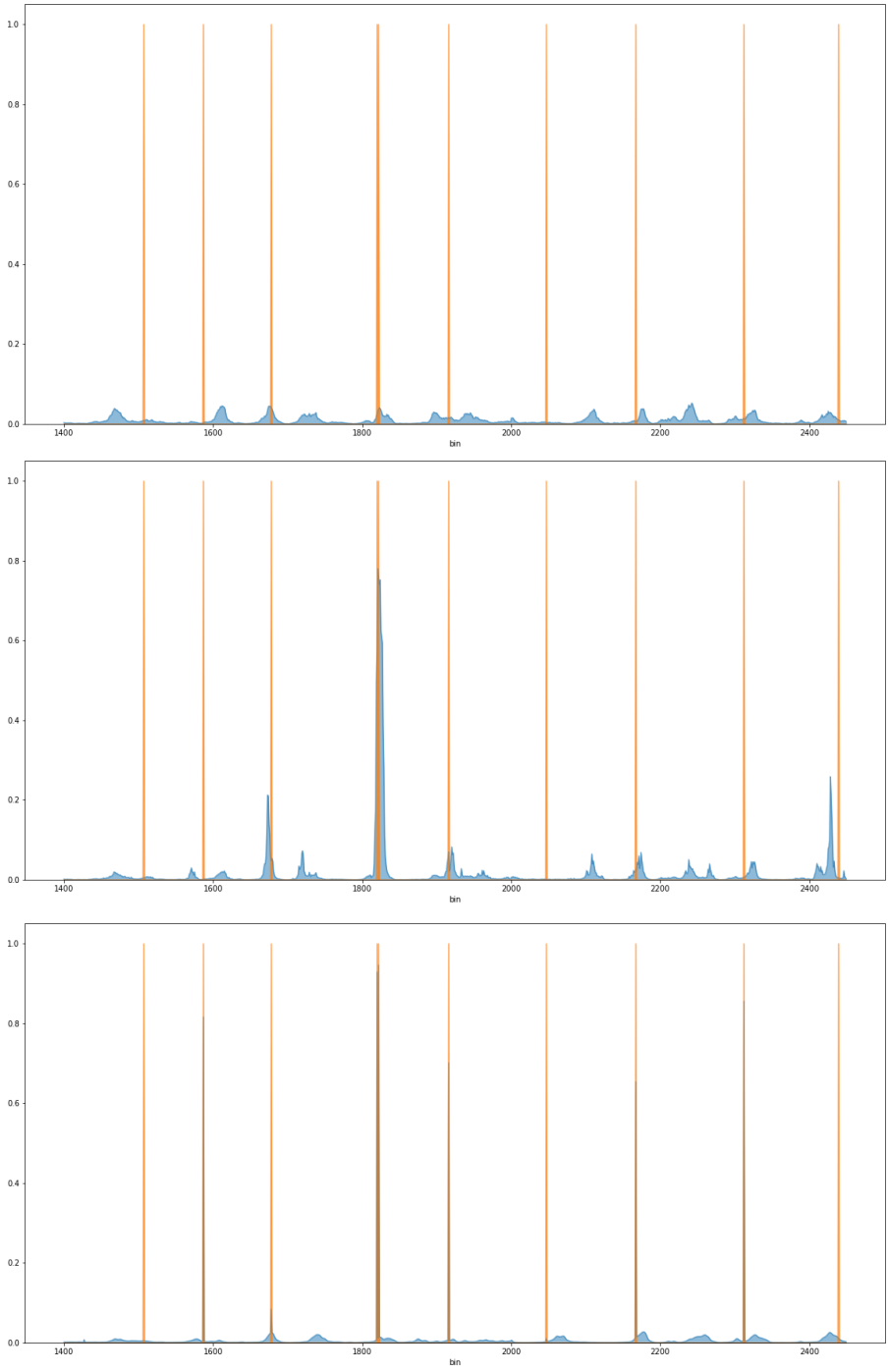


Figure 2.10: Probability outputs (blue) vs. true binary labels (orange) for chr3:7000000-12500000 in mm10 genome for three different scenarios: Pooling (top), Pooling + PhyloTAD (middle), Pooling + PhyloTAD with labels (bottom)

We developed different machine learning models to predict TAD boundaries in mammals, based on both sequence features and cross-species comparison. They differ in the input required to the user, but all include a general classifier made of either a Random Forest or a LSTM to scan the genome, and a phylogenetic algorithm to make the results more coherent with the tree of evolution. We focused on the five mammals for which Hi-C data is available, in order to assess the performance of our models, but the final goal is to apply them on species with unavailable Hi-C data to broaden knowledge on new species. A summary table 2.6 highlights the main results.

Model	Mouse	Human	Dog	Macaque	Rabbit
Transfer	0.71	0.70	0.70	0.62	0.71
Pooling	0.73	0.73	0.74	0.63	0.73
Species-specific	0.71	0.71	0.72	0.63	0.72
Concatenation	0.72	0.72	0.73	0.62	0.72
Pooling + PhyloTAD	0.77	0.79	0.78	0.71	0.78
Pooling + Smoothing f'	0.75	0.77	0.78	0.69	0.77
TAD concatenation	0.83	0.82	0.82	0.79	0.84
PhyloTAD only	0.81	0.82	0.80	0.78	0.81
Pooling + PhyloTAD with labels	0.86	0.84	0.85	0.81	0.86
Global concatenation	0.89	0.89	0.90	0.85	0.88

Table 2.6: Summary of AUC scores for all proposed models

Transfer, Pooling and Species-specific models suggest that sequence-based features do carry information about TAD formation mechanisms, but AUC scores are too low - around 0.72 - to rely exclusively on them for TAD identification. When they are followed by phylogenetic algorithms, either with PhyloTAD or with a new deep learning classifier f', there is a significant improvement – by 0.06 -, advocating for the subsequent benefit of such knowledge. All in all, adding true labels of known mammals with probability outputs from f for unknown mammals, a cross-species comparison enables a quite accurate prediction of both left and right boundaries on a target species, paving the way for a complete genome-wide identification of these structures. The different models are available as building blocks for an end-to-end software where the user could input their data set and apply the relevant trained model. An interesting question to discuss is the purpose of phylogenetic data for such task. Indeed, when a cell folds its genome, it does not have

access to phylogenetic information, but only to the sequence. Why does it systematically succeeds at producing the right fold, whereas our models do not? What information are we missing? One aspect is epigenetic data. Our predictor does not have DNA methylation and histone modification data, but the cell does. Still, these epigenetic modifications are ultimately sequence-specific. So all the information is the sequence, but presented in a more informative way, still hard to understand to our knowledge.

There are some aspects that can be improved to boost performance and possibilities. First of all, concerning sequence-based features, we were limited to 335 TFs with Homer database of motifs. We could have looked for other marks, such as histone marks, DNase I hypersensitive site, etc. However, as they cannot be accurately predicted directly from the sequence and require biological experiments, we would lose the initial purpose of sequence-based study. Some of them have putative effect on chromatin architecture and could have provide additional information to the model. In addition, Homer uses a deterministic algorithm for mapping sites. We could have used alternative methods, in particular innovative deep learning approaches such as Basset or DeepBind [50,51], which have excellent performance for predicting DNA-binding protein sites. To go even further, we could have replicated and incorporate their methods into our models, which introduce CNNs, to start from raw DNA sequence to detect particular patterns. This is a heavier but more independent and user-friendly method, because the user only need to input the sequence genome to have it annotated with TADs, which is an objective for future work. It has been done by Henderson et al. in fruit flies, thus the difference of families, in addition to the choice of bins exposed previously, makes the comparison irrelevant. We can also discuss about the choice of human Hi-C data. Indeed, it is the only species for which the source changes and the cell line is slightly different, with cancer cells HepG2. The issue with such cancer cells is the possible presence of trans-locations in the DNA sequence, hindering a proper learning. Nevertheless, this cell line was the closest to other species' lines we could find, and the percentage of difference should remain small, allowing the usage of such data with some reserve on the results. Finally, there are some ways of improvement in machine learning pipelines. We are currently working on a final pipeline where PhyloTAD is used inside the learning part - not after -, and included in the compu-

tation of gradients for backpropagation. We hope that this method will converge faster, balancing with the huge amount of calculus needed to compute gradients in each batch. The frequent dependency on the amount of data is also an interesting point to discuss. In this study, we were limited to 5 species, for which a consistent amount of Hi-C data was available. We can expect our model to perform better with the increase in Hi-C experiments, so that a virtuous cycle could be generated between *in vitro* experiments and *in silico* predictions, towards a better understanding of 3D genomics in mammals.

Competing Interests

The authors declare that they have no competing interests.

Author's Contribution

Amaury Leroy contributed to the design of the study, implemented the computational analyses, and wrote the manuscript. Mathieu Blanchette conceived the study, coordinated the computational anaysis, and helped draft the manuscript. All auhtors have reviewed and approved the final manuscript.

2.5 Supplementary Methods

Here, we derive the probabilistic model that leads to a dynamic programming perspective. $X_u(i)$ and $O_u(i)$ are the quantities we want to find with a recurrent formula. See the Methods section 2.2, in particular Figures 2.2 and 2.3 for the context. We assume that the probability of a vector embedding transcription marks only depends on the value of the

random variable T at the node of interest, and is independent of other vectors or the value of T on another node. In this derivation, u is the internal node of interest, with children v - left - and w - right -, $p(u)$ is the ancestor of u. Σ_u the set of leaves under the subtree rooted at u and Ω_u is the set composed of the node u and the leaves not under the subtree rooted at u.

$$
\begin{aligned}
X_u(i) &= \sum_{a,b\in\{0,1\}} \Pr[\{C_j\}_{j\in\Sigma_u} \mid T_u = i, T_v = a, T_w = b] \Pr[T_v = a \mid T_u = i] \Pr[T_w = b \mid T_u = i] \\
&= \sum_{a\in\{0,1\}} \sum_{b\in\{0,1\}} \Pr[\{C_j\}_{j\text{left}} \mid \{C_j\}_{j\text{right}}, T_u = i, T_v = a, T_w = b] \Pr[T_v = a \mid T_u = i] \\
&\times \Pr[\{C_j\}_{j\text{right}} \mid T_u = i, T_v = a, T_w = b] \Pr[T_w = b \mid T_u = i] \\
&= \sum_{a\in\{0,1\}} X_v(a) \Pr[T_v = a \mid T_u = i] \sum_{b\in\{0,1\}} X_w(b) \Pr[T_w = b \mid T_u = i] \\
&= L_u(i) R_u(i) \\
O_u(i) &= \sum_{a\in\{0,1\}} \Pr[\{C_j\}_{j\in\Omega_u}, T_{p(u)} = a \mid T_u = i] \\
&= \sum_{a\in\{0,1\}} \Pr[T_{p(u)} = a \mid T_u = i] \Pr[\{C_j\}_{j\in\text{right descendents of p(u)}} \mid T_{p(u)} = a, T_u = i] \\
&\times \Pr[\{C_j\}_{j\in\Omega_{p(u)}} \mid \{C_j\}_{j\in\text{right}}, T_{p(u)} = a, T_u = i] \\
&= \sum_{a\in\{0,1\}} \Pr[T_u = i \mid T_{p(u)} = a] \frac{\Pr[T_{p(u)} = a]}{\Pr[T_u = i]} R_{p(u)}(a) O_{p(u)}(a)
\end{aligned}
$$

Chapter 3

Conclusion

Spatial genomic organization plays a key role in gene regulation. The introduction of technologies such as Hi-C has allowed increasing insights into the 3D properties of genomes and cell types. In particular, a class of structure has been brought to light by Dixon et al. in 2012, which are now referred as Topologically Associated Domains (TADs) [18]. For mammals, they are considered as a crucial functional unit of chromatin, made of condensed loops of around 1Mb, inside which chromatin loci interact more frequently than with regions located in adjacent domains. TAD organization across genome sequence defines regulatory landscapes, as it enhances or disrupt physical contact between elements involved in gene expression. Therefore, understanding their biological architecture is of paramount importance, but still incomplete. This is a consequence of small number of species for which TAD annotations are available, due to the substantial cost of Hi-C experiments.

TADs are then a topic of high interest, and many studies have tried to characterize their mechanisms of formation. In mammals, the extrusion model, where CTCF and cohesin jointly interact with DNA sequence to create non-random loops, helps us better understand how TADs are created, and what are the patterns present at their boundaries [16].

The knowledge of DNA-binding protein sites can thus give precious hints to annotate the genome with TADs. In other words, the first hypothesis is that TAD boundaries are hardwired in DNA sequence. In addition, TADs are mainly conserved across mammals because of the negative selection they apply with gene regulation. Indeed, the disruption of a TAD can generate serious unsustainable change in the phenotype. Comparing neighboring mammals is then an interesting and still unexplored idea to infer TADs in unknwon species. The new flow of large scale data, in particular genomic sequences, enables many computational approaches to study and model TAD organization in the genome, including machine learning ones.

In this thesis, we show that TAD boundaries in mammals can be accurately predicted from the joint use of sequence-level features and evolutionary data. We implemented 10 machine learning models based on these two hypotheses, which differ in the input required to the user. Depending on the available data, the latter can apply a specific trained model to infer TAD right and left boundaries - separately - on the whole genome. The major building blocks are made of a general TAD predictor - either a Random Forest or a LSTM -, which scans the genome to give initial predictions based on sequence features only; and a phylogenetic algorithm comparing the predicted or referenced boundaries between close mammals to update the predictions and enhance the performance of boundary inference. To assess the robustness of our models, we extracted TAD annotations from five mammals and computed AUC scores for each one of them, before and after update.

The first insight is that DNA sequence features convey subsequent knowledge about TAD boundaries, but they are not sufficiently informative to accurately predict them for a practical use. Besides, we only used transcription factor binding sites from Homer, and we can expect that a broader library of regulatory marks - histone marks, DNase I hypersensitive sites, ... - would profit such an inference, even if they require experimental results. To go even further, it would be interesting to start from raw DNA sequence and build a more general predictor that would learn by itself such features, inspired from those algorithms. In this perspective, the user would only need to input raw genome sequence from the target species to get accurate feature-based predictions of right and left TAD bound-

aries. An interesting thought about the limitation in performance lies in the small number of training positive examples, which limit the ability to learn sophisticated models. As the training data is limited by the biology, no amount of additional efforts will be able to produce more data in species. That explains the idea to improve predictions, developed in next paragraph.

The second important lesson is that cross-species comparison represent a valuable improvement for TAD inference, as AUC scores are significantly enhanced after update through the phylogenetic tree. This new and still unprecedented idea - in the context of TAD prediction - have a high potential that need deeper investigations - on bigger databases, with larger trees - to be fully exploited. PhyloTAD is a user-friendly, deterministic and scalable probabilistic algorithm to propagate information through neighboring species. However, we only used it after the training of predictor f, and it could be interesting to incorporate it inside the learning. Computing the gradients would then be really heavy but we can expect the learning to converge faster as predictions would be more accurate for the very first batches. All in all, the future work will focus on the development of an end-to-end algorithm, taking as input raw DNA sequence. A machine learning pipeline would first learn relevant transcription mark counts inside each bin, that would be scanned thanks to a recurrent neural network which would output temporary left/right boundary predictions. Those outputs would then be compared to predictions from the same classifier applied to other species, or to referenced true boundaries from Hi-C experiments when available, in order to update them before back-propagation. Finally, after training, any mammals - and eventually other organisms - placed inside the phylogenetic tree could have its genome annotated with TAD boundaries, paving the way for a better understanding and new discoveries of 3D genomics.

We could eventually wonder if this type of cross-species boosting approach could be used for other bioinformatics tasks. Applied here in the context of TADs, the idea remains very general and could impact well beyond 3D genomics, for any sequence-based function prediction task, like TFBS or gene prediction.

Bibliography

[1] J. C. Venter, M. D. Adams, and et al., "The Sequence of the Human Genome," *THE HUMAN GENOME*, vol. 291, p. 51, 2001.

[2] J. Dekker, M. Guttman, V. O. Leshyk, J. T. Lis, S. Lomvardas, L. A. Mirny, C. C. O'Shea, P. J. Park, B. Ren, J. C. R. Politz, J. Shendure, and S. Zhong, "The 4D nucleome project," *Nature*, vol. 549, pp. 219–226, Sept. 2017.

[3] T. Cremer, M. Cremer, and C. Cremer, "The 4D Nucleome: Genome Compartmentalization in an Evolutionary Context," *Biochemistry (Moscow)*, vol. 83, pp. 313–325, Apr. 2018.

[4] D. Gorkin, D. Leung, and B. Ren, "The 3D Genome in Transcriptional Regulation and Pluripotency," *Cell Stem Cell*, vol. 14, pp. 762–775, June 2014.

[5] B. D. Pope, T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield, R. E. Thurman, Y. Cheng, G. Gülsoy, J. H. Dennis, M. P. Snyder, J. A. Stamatoyannopoulos, J. Taylor, R. C. Hardison, T. Kahveci, B. Ren, and D. M. Gilbert, "Topologically associating domains are stable units of replication-timing regulation," *Nature*, vol. 515, pp. 402–405, Nov. 2014.

[6] K. D. Makova and R. C. Hardison, "The effects of chromatin organization on variation in mutation rates in the genome," *Nature Reviews Genetics*, vol. 16, pp. 213–223, Apr. 2015.

[7] F. Grubert, J. Zaugg, M. Kasowski, O. Ursu, D. Spacek, A. Martin, P. Greenside, R. Srivas, D. Phanstiel, A. Pekowska, N. Heidari, G. Euskirchen, W. Huber, J. Pritchard, C. Bustamante, L. Steinmetz, A. Kundaje, and M. Snyder, "Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions," *Cell*, vol. 162, pp. 1051–1065, Aug. 2015.

[8] C. J. F. Cameron, J. Fraser, M. Blanchette, and J. Dostie, "Mapping and Visualizing Spatial Genome Organization," in *The Functional Nucleus* (D. P. Bazett-Jones and G. Dellaire, eds.), pp. 359–383, Cham: Springer International Publishing, 2016.

[9] B. Bonev and G. Cavalli, "Organization and function of the 3D genome," *Nature Reviews Genetics*, vol. 17, pp. 661–678, Nov. 2016.

[10] T. Sexton and G. Cavalli, "The Role of Chromosome Domains in Shaping the Functional Genome," *Cell*, vol. 160, pp. 1049–1059, Mar. 2015.

[11] Y. S. Fan, L. M. Davis, and T. B. Shows, "Mapping small DNA sequences by fluorescence in situ hybridization directly on banded metaphase chromosomes.," *Proceedings of the National Academy of Sciences*, vol. 87, pp. 6223–6227, Aug. 1990.

[12] J. Dekker, "Capturing Chromosome Conformation," *Science*, vol. 295, pp. 1306–1311, Feb. 2002.

[13] S. Sati and G. Cavalli, "Chromosome conformation capture technologies and their impact in understanding genome function," *Chromosoma*, vol. 126, pp. 33–44, Feb. 2017.

[14] T. Nagano, S. W. Wingett, and P. Fraser, "Capturing Three-Dimensional Genome Organization in Individual Cells by Single-Cell Hi-C," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 1654, pp. 79–97, 2017.

[15] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny,

E. S. Lander, and J. Dekker, "Comprehensive mapping of long range interactions reveals folding principles of the human genome," *Science (New York, N.Y.)*, vol. 326, pp. 289–293, Oct. 2009.

[16] S. Rao, M. Huntley, N. Durand, E. Stamenova, I. Bochkov, J. Robinson, A. Sanborn, I. Machol, A. Omer, E. Lander, and E. Aiden, "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping," *Cell*, vol. 159, pp. 1665–1680, Dec. 2014.

[17] C. J. Cameron, J. Dostie, and M. Blanchette, "HIFI: estimating DNA-DNA interaction frequency from Hi-C data at restriction-fragment resolution," *Genome Biology*, vol. 21, Dec. 2020.

[18] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, pp. 376–380, Apr. 2012.

[19] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli, "Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome," *Cell*, vol. 148, pp. 458–472, Feb. 2012.

[20] E. Crane, Q. Bian, R. P. McCord, B. R. Lajoie, B. S. Wheeler, E. J. Ralston, S. Uzawa, J. Dekker, and B. J. Meyer, "Condensin-driven remodelling of X chromosome topology during dosage compensation," *Nature*, vol. 523, pp. 240–244, July 2015.

[21] C. Wang, C. Liu, D. Roqueiro, D. Grimm, R. Schwab, C. Becker, C. Lanz, and D. Weigel, "Genome-wide analysis of local chromatin packing in Arabidopsis thaliana," *Genome Research*, vol. 25, pp. 246–256, Feb. 2015.

[22] T. Mizuguchi, G. Fudenberg, S. Mehta, J.-M. Belton, N. Taneja, H. D. Folco, P. FitzGerald, J. Dekker, L. Mirny, J. Barrowman, and S. I. S. Grewal, "Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe," *Nature*, vol. 516, pp. 432–435, Dec. 2014.

[23] T. B. K. Le, M. V. Imakaev, L. A. Mirny, and M. T. Laub, "High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome," *Science*, vol. 342, pp. 731–734, Nov. 2013.

[24] M. Vietri Rudan and S. Hadjur, "Genetic Tailors: CTCF and Cohesin Shape the Genome During Evolution," *Trends in Genetics*, vol. 31, pp. 651–660, Nov. 2015.

[25] K. Nasmyth and C. H. Haering, "Cohesin: Its Roles and Mechanisms," *Annual Review of Genetics*, vol. 43, pp. 525–558, Dec. 2009.

[26] B. Bonev, N. Mendelson Cohen, Q. Szabo, L. Fritsch, G. L. Papadopoulos, Y. Lubling, X. Xu, X. Lv, J.-P. Hugnot, A. Tanay, and G. Cavalli, "Multiscale 3D Genome Rewiring during Mouse Neural Development," *Cell*, vol. 171, pp. 557–572.e24, Oct. 2017.

[27] J. R. Dixon, I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie, Y. Diao, J. Liang, H. Zhao, V. V. Lobanenkov, J. R. Ecker, J. A. Thomson, and B. Ren, "Chromatin architecture reorganization during stem cell differentiation," *Nature*, vol. 518, pp. 331–336, Feb. 2015.

[28] J. Fraser, C. Ferrai, A. M. Chiariello, M. Schueler, T. Rito, G. Laudanno, M. Barbieri, B. L. Moore, D. C. Kraemer, S. Aitken, S. Q. Xie, K. J. Morris, M. Itoh, H. Kawaji, I. Jaeger, Y. Hayashizaki, P. Carninci, A. R. Forrest, The FANTOM Consortium, C. A. Semple, J. Dostie, A. Pombo, and M. Nicodemi, "Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation," *Molecular Systems Biology*, vol. 11, pp. 852–852, Dec. 2015.

[29] G. Fudenberg and K. S. Pollard, "Chromatin features constrain structural variation across evolutionary timescales," *Proceedings of the National Academy of Sciences*, vol. 116, pp. 2175–2180, Feb. 2019.

[30] Q. Szabo, F. Bantignies, and G. Cavalli, "Principles of genome folding into topologically associating domains," *Science Advances*, vol. 5, p. eaaw1668, Apr. 2019.

[31] I. E. Eres, K. Luo, C. J. Hsiao, L. E. Blake, and Y. Gilad, "Reorganization of 3D Genome Structure May Contribute to Gene Regulatory Evolution in Primates:," *bioRxiv*, Nov. 2018.

[32] J. Krefting, M. A. Andrade-Navarro, and J. Ibn-Salem, "Evolutionary stability of topologically associating domains is associated with conserved gene regulation," *BMC Biology*, vol. 16, p. 87, Aug. 2018.

[33] C. Berthelot, M. Muffato, J. Abecassis, and H. Roest Crollius, "The 3D Organization of Chromatin Explains Evolutionary Fragile Genomic Regions," *Cell Reports*, vol. 10, pp. 1913–1924, Mar. 2015.

[34] C. Gómez-Marín, J. J. Tena, R. D. Acemel, M. López-Mayorga, S. Naranjo, E. de la Calle-Mustienes, I. Maeso, L. Beccari, I. Aneas, E. Vielmas, P. Bovolenta, M. A. Nobrega, J. Carvajal, and J. L. Gómez-Skarmeta, "Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders," *Proceedings of the National Academy of Sciences*, vol. 112, pp. 7542–7547, June 2015.

[35] M. Vietri Rudan, C. Barrington, S. Henderson, C. Ernst, D. Odom, A. Tanay, and S. Hadjur, "Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture," *Cell Reports*, vol. 10, pp. 1297–1309, Mar. 2015.

[36] Y. Yang, Y. Zhang, B. Ren, J. R. Dixon, and J. Ma, "Comparing 3D Genome Organization in Multiple Species Using Phylo-HMRF," *Cell Systems*, vol. 8, pp. 494–505.e14, June 2019.

[37] D. Sankoff, "Minimal Mutation Trees of Sequences," *SIAM Journal on Applied Mathematics*, vol. 28, no. 1, pp. 35–42, 1975.

[38] W. M. Fitch, "Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology," *Systematic Zoology*, vol. 20, no. 4, pp. 406–416, 1971. Publisher: [Oxford University Press, Society of Systematic Biologists, Taylor & Francis, Ltd.].

[39] J. C. Clemente, K. Ikeo, G. Valiente, and T. Gojobori, "Optimized ancestral state reconstruction using Sankoff parsimony," *BMC Bioinformatics*, vol. 10, p. 51, Feb. 2009.

[40] J. Felsenstein, "MAXIMUM LIKELIHOOD AND MINIMUM-STEPS METHODS FOR ESTIMATING EVOLUTIONARY TREES FROM DATA ON DISCRETE CHARACTERS," *SYSTEMATIC ZOOLOGY*, p. 10, Mar. 1973.

[41] R. Dali and M. Blanchette, "A critical assessment of topologically associating domain prediction tools," *Nucleic Acids Research*, vol. 45, pp. 2994–3005, Apr. 2017.

[42] M. Zufferey, D. Tavernari, E. Oricchio, and G. Ciriello, "Comparison of computational methods for the identification of topologically associating domains," *Genome Biology*, vol. 19, Dec. 2018.

[43] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, "A primer on deep learning in genomics," *Nature Genetics*, vol. 51, pp. 12–18, Jan. 2019.

[44] I. G. a. Y. B. a. A. Courville, *Deep Learning*. MIT Press, 2016.

[45] Y. Lecun, "Generalization and network design strategies," *Connectionism in perspective*, 1989. Publisher: Elsevier.

[46] Y. LeCun, L. Bottou, Y. Bengio, and P. Ha, "Gradient-Based Learning Applied to Document Recognition," *IEEE*, p. 46, 1998.

[47] D. E. Rumelhart, G. E. Hintont, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, p. 4, 1986.

[48] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, pp. 157–166, Mar. 1994. Conference Name: IEEE Transactions on Neural Networks.

[49] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, pp. 1735–1780, Nov. 1997.

[50] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nature Biotechnology*, vol. 33, pp. 831–838, Aug. 2015.

[51] D. R. Kelley, J. Snoek, and J. Rinn, "Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks," *Genome Research*, p. gr.200535.115, May 2016.

[52] Z. Shen, W. Bao, and D.-S. Huang, "Recurrent Neural Network for Predicting Transcription Factor Binding Sites," *Scientific Reports*, vol. 8, Oct. 2018.

[53] J. Huang, E. Marco, L. Pinello, and G.-C. Yuan, "Predicting chromatin organization using histone marks," *Genome Biology*, vol. 16, Dec. 2015.

[54] Z. Al Bkhetan and D. Plewczynski, "Three-dimensional Epigenome Statistical Model: Genome-wide Chromatin Looping Prediction," *Scientific Reports*, vol. 8, Dec. 2018.

[55] T. Gao and J. Qian, "EAGLE: An algorithm that utilizes a small number of genomic features to predict tissue/cell type-specific enhancer-gene interactions," *PLOS Computational Biology*, p. 22, Oct. 2019.

[56] J. Henderson, V. Ly, S. Olichwier, P. Chainani, Y. Liu, and B. Soibam, "Accurate prediction of boundaries of high resolution topologically associated domains (TADs) in fruit flies using deep learning," *Nucleic Acids Research*, vol. 47, pp. e78–e78, July 2019.

[57] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. Mirny, "Formation of Chromosomal Domains by Loop Extrusion," *Cell Reports*, vol. 15, pp. 2038–2049, May 2016.

[58] N. C. Durand, M. S. Shamim, I. Machol, S. S. Rao, M. H. Huntley, E. S. Lander, and E. L. Aiden, "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments," *Cell Systems*, vol. 3, pp. 95–98, July 2016.

[59] N. C. Durand, J. T. Robinson, M. S. Shamim, I. Machol, J. P. Mesirov, E. S. Lander, and E. L. Aiden, "Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom," *Cell Systems*, vol. 3, pp. 99–101, July 2016.

[60] R. Dali, G. Bourque, and M. Blanchette, "RobusTAD: A Tool for Robust Annotation of Topologically Associating Domain Boundaries," preprint, Bioinformatics, Apr. 2018.

[61] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass, "Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities," *Molecular Cell*, vol. 38, pp. 576–589, May 2010.

[62] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, "The Human Genome Browser at UCSC," *Genome Research*, p. 12, 2002.

[63] G. van Rossum, "Python tutorial, Technical Report CS-R9526," *Centrum voor Wiskunde en Informatica (CWI), Amsterdam*, May 1995.

[64] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors, "SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, Mar. 2020. arXiv: 1907.10121.

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau, "Scikit-learn: Machine Learning in Python," *MACHINE LEARNING IN PYTHON*, p. 6, Oct. 2011.

[66] J. Bergstra, D. Yamins, and D. D. Cox, "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures," *Proceedings of the 30th ICML*, p. 9, 2013.

[67] I. Letunic and P. Bork, "Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation," *Bioinformatics*, vol. 23, pp. 127–128, Jan. 2007.

[68] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, pp. 1422–1423, June 2009.