博士論文

Doctoral thesis

In-silico identification of transcription-factor binding sites and structural features of transcriptional regulatory regions

(転写因子結合部位と転写制御領域の構造的特徴のイン・シリコ同定)

バシャ　グティエレス　ユセフ

Abstract

Understanding of transcriptional regulation have always been one of the most fundamental problems in molecular biology research. One of the biggest challenges for scientists in order to reach that goal is the discovery of sequence elements that are bound by DNA-binding proteins, called transcription factor binding sites (TFBS).

These TFBS, also known as regulatory motifs, are considerably difficult to identify, given the fact that they are very often short in length and can show sequence variation. To facilitate the task of identifying these elements, a wide range of different computational strategies have been developed over the years, relying on the overrepresentation and/or the evolutionary conservation that TFBS usually show. In the recent years, the rapid development of the machine learning field has brought a wide range of new techniques that can aid scientists in the discovery of binding sites. One of these new algorithms, which had never been applied to motif finding, is known as *topic models*.

Topic models are statistical algorithms which mix concepts of natural language processing and machine learning, and whose purpose is to find the topics contained in a set of documents and define the structure of their contents by hierarchical Bayesian analysis. Similarly, a set of biological sequences can be thought as combinations of k-mers (words), and motifs as clusters of words with a certain meaning (topics), from which it can be inferred that motif discovery can be seen a problem equivalent to finding the topics contained in a set of documents.

Taking this as a hypothesis, a new method in which motifs are identified by constructing a topic model from a set of sequences was developed in this study. And, additionally, by using the perplexity, which is a measure of the accuracy with which the distribution of a topic model predicts a sample, an algorithm based on a genetic algorithm (GA) developed in a previous study was improved by reducing the number of false positives.

In a similar way to the progress in the machine learning field, in the most recent years, there have been many noteworthy advances in our understanding of the mechanisms of control of gene expression. An important example is the discovery that control of transcriptional regulation is very often carried out by a very small fraction of the hundreds of transcription factors that are present in the cell. The way this regulation is controlled is by following a process in which this small fraction of key transcription factors bind cooperatively to individual enhancer elements and targeting genes by the use of cofactors and RNA polymerase II. Enhancers are thus a fundamental resource for identifying genetic variations that might alter binding of TFBS and lead to disease. In fact, recent studies suggest that an important portion of these disease-associated variations occur precisely in enhancer regions. Therefore, the identification of the few key transcription factors responsible of the control of the transcriptional regulation of a given tissue along with the identification of the enhancers involved in gene expression becomes fundamental for unravelling the mechanisms responsible of tissue-specific diseases and syndromes.

Along with the motif finding methods previously presented, this project proposes a workflow to determine the elements involved in tissue-specific transcriptional regulation which, in a first step, the binding sites of the key transcription factors are predicted in a data set formed by promoters and enhancers, to then study their structural features and evolutionary conservation. The goal is to narrow down the study of the TFs, promoters and enhancers involved in transcriptional regulation to a few significant candidates.

Both of the two motif finding methods here presented were tested and their performances compared to another 14 motif finding tools by the scores obtained in seven different statistical coefficients.

From the results obtained, it can be inferred that the CTM method shows remarkable levels in Sensitivity (both at nucleotide level, *nSn*, and at site level, *sSn*), and only the other method here proposed, the Statistical GA, is able to outperform it. The Average Site Performance (*sASP*) is considerably high as well, and the rest of the statistics show levels comparable to most of the other methods. The Statistical GA method, after the addition of the perplexity measure to reduce the number of false positives, practically doubled all of the average statistics. From these results, arises the conclusion that topic models are a perfectly valid method to develop motif finding algorithms.

The workflow to determine the elements of tissue-specific transcriptional regulation was tested for cerebellar tissue, and 6 candidate key TFBS have been identified, from which 3 of them are known motifs (Striatum-HSF1, HepG2-HSF1, MCF7-FOXM1). Additionally, 43 promoters and 29 enhancers were identified as candidates to be involved in cerebellar transcriptional regulation. The whole set is considerably robust given the occurrence near peaks, the statistical significance of motif co-occurrence, and the high evolutionary conservation of binding sites in both promoters and enhancers.

Hopefully this would help in the future to improve our understanding of tissue-specific transcriptional regulation in the cerebellum and to serve as the basis of different studies focused on the analysis of transcriptional regulation in specific tissues.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Prof. Kenta Nakai, who offered me his advice, guided me and supported me in every possible way during my Master and my PhD studies, and even from the time of the application before knowing if I would be able to join his laboratory.

Besides my advisor, I would also like to thank the rest of the members of the "Laboratory of Functional Analysis "*in silico*", for creating a good atmosphere for research and offering good advice for the project, and also for being always willing to lend a hand during the toughest times in the life outside the lab.

I would also like to thank the staff in the department of Computational Biology and Medical Sciences, of the Graduate School of Frontier Sciences of the University of Tokyo, for the institutional help, and the Ministry of Education, Culture, Sports, Science and Technology (MEXT), for awarding me the scholarship that allowed this project to be carried out.

Last but not least, I would like to thank my family, my friends and all the people dear to me, who gave me the necessary emotional support during this whole process and will continue to do so after it.

Table of contents

Index of Tables

Index of Figures

Introduction

1. Introduction

1.1. Overview

This study tries to delve into the process of transcriptional regulation and shed some light on certain issues that have been targeted by scientists for decades in order to improve our understanding of the most fundamental biological processes.

The structure of the thesis that follows includes a main chapter and an additional chapter, corresponding each one to different methods and approaches, but both aiming to contribute to the study of transcriptional regulation.

The main chapter focuses on the study of new *in-silico* methods for motif discovery and the design of novel algorithms that can help to the development of the field in the future. More specifically, the focus is on the application of the so-called *topic models*, which are statistical algorithms belonging to the artificial intelligence field, and which combine machine learning with natural language processing techniques in order to discover the structure of the different topics contained in a set of documents. For this first chapter, we developed two algorithms: one of them completely relying in topic models in order to find motifs in a set of biological sequences, and a second one relying mainly in the combination of statistical coefficients with a genetic algorithm structure but then mixed with topic models to improve its accuracy.

The additional chapter proposes a simple *in-silico* workflow to determine which ones are the main transcription factors involved in tissue-specific transcriptional regulation, how they cooperate with each other, how their binding sites are distributed around the TSS in the corresponding promoters and to which enhancers they bind. By using the cerebellum as the tissue of study, this chapter tries to simplify a way to identify the elements involved in the cerebellar transcriptional regulation.

Before plunging into the study in depth, this section will offer some background information about the basic topics on which the research here described stand: transcriptional regulation, motif finding, genetic algorithms and topic models.

1.2. Background

1.2.1. Transcriptional regulation

The regulation of transcription is a biological process through which a cell can control the conversion of DNA to RNA, also known as transcription. Thanks to this mechanism, the cell can regulate gene activity in order to respond to a specific signal both from outside or inside the cell.

Transcriptional regulation is one of the most essential mechanisms of life in any given organism, and therefore its understanding is a fundamental problem in molecular biology. Even though there is still a very long way until we reach that level of understanding, we have already taken some important steps towards that goal [1].

We know that transcriptional regulation is controlled by a type of proteins called **transcription factors (TF)**, which collaborate with other proteins (usually those known as cofactors) to bind the corresponding promoters and enhancer elements and modulate the activity of RNA polymerase (**Fig. 1**).

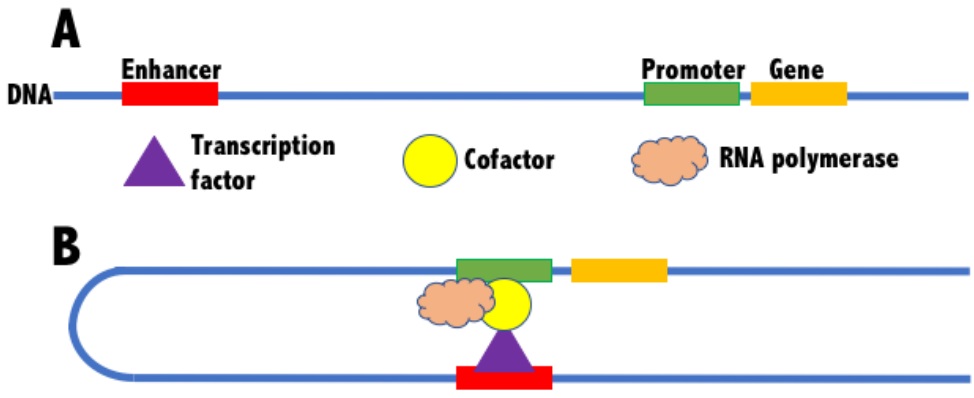

Figure 1. Transcriptional regulation with enhancer and cofactor
A. Elements involved in the transcription mechanism
B. The transcription factor, by using a cofactor and RNA polymerase, binds to the enhancer and allows it to connect with the corresponding promoter

Figure 1: A typical scenario of transcriptional regulation in eukaryotes

Promoters are portions of DNA in which the transcription of a specific gene is initiated. They are always located around and on the same strand and same direction of the transcription start site (TSS) of the gene, and their length is usually between 50 and 1500 base pairs long.

Enhancers, similarly to promoters, are bound by transcription factors in order to initiate transcriptional regulation, and are also of a similar length. The main difference between promoters and enhancers is that enhancers are located far from the TSS (up to 1 million base pairs away) and can be upstream or downstream of it, and in any direction. They are thus more difficult to locate and study than promoters.

Cofactors, also known as transcription coregulators, are another type of protein, whose function is to act in concert with TFs to either activate (in this case, they are known as coactivators) or repress (corepressors) transcription in a particular gene.

RNA polymerase is an enzyme whose main function is to produce the transcript DNA in the transcriptional regulatory process.

1.2.2. Motif finding

1.2.2.1. Overview

Sequence motifs are patterns formed by nucleotides or amino-acids, typically short in length and repeated very often in DNA or protein sequences, which have some sort of biological significance. In the case of DNA motifs, their usual function is to act as a binding site for proteins.

When these proteins are involved in transcriptional regulation, they are called transcription factors, and the instances of the motifs that act as binding sites for them are known as transcription factor binding sites (TFBS).

Since discovering these TFBS is fundamental to understand transcriptional regulation, a wide variety of motif finding methods have been developed over the years [2][3].

Historically, the first methods able to determine the locations of binding sites were experimental methods such as DNAase footprinting [4] and gel-shift assays [5]. But their high cost and the rapid growth of computational performance favoured the (still ongoing) development of a wide range of different computational techniques for motif discovery in which scientists rely nowadays for their research.

Given the importance of TFBS in transcriptional regulation and the existence of both databases of known motifs and tools for the discovery of *de novo* motifs aplenty, motif finding has always been one of the fundamental fields within biological research in the past and present, and presumably in the near future.

1.2.2.2. Representation of motifs

There are several ways of representing motifs, which are vital to know to understand how computational motif finding methods work.

A DNA **consensus sequence** is the sequence of the most common nucleotides found at each position in a sequence alignment, showing which nucleotides are constant among the different sequences in the alignment, and which ones vary from each other (either limited to some specific nucleotides or unconstrained). **Table 1** shows how each one of the possible nucleotides or combinations of nucleotides are represented in a consensus sequence.

Consensus sequences are simple and clear, but they have an important drawback, which is that they are less informative than other representations, due to its inability to show actual frequencies of the appearances of the specific residues in a specific location in the alignment. To overcome this limitation, matrices and sequence logos are usually used instead.

Table 1: Summary of single-letter IUPAC symbols

Symbol	Meaning	Description
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	G or A	puRine
Y	T or C	pYrimidine
M	A or C	aMino
K	G or T	Keto
S	G or C	Strong
W	A or T	Weak
H	A or C or T	not-G, H follows G
B	G or T or C	not-A, B follows A
V	G or C or A	not-T (not-U), V follows U
D	G or A or T	not-C, D follows C
N	G or A or T or C	aNy

Matrices are more accurate than sequence logos thanks to the fact that they keep scores for each nucleotide at each position in the alignment. There are two basic types of matrices for motif representation:

- **Position Frequency Matrix (PFM)**: contains a raw value for the frequency of each nucleotide at a given position.
- **Position Weight Matrix (PWM)**: contains relative frequencies after the normalization of the corresponding scores of the PFM and the calculation of the log likelihoods of the elements.

Position Frequency Matrix (PFM)

>MCF7-FOXM1

0.022	0.024	0.033	0.921
0.478	0.001	0.520	0.001
0.001	0.053	0.001	0.945
0.001	0.001	0.100	0.898
0.001	0.001	0.241	0.757
0.617	0.001	0.381	0.001
0.015	0.793	0.022	0.170
0.132	0.296	0.001	0.571
0.027	0.296	0.129	0.548
0.517	0.036	0.043	0.404

Motif logo

Consensus sequence

TRTTTACTTW

Figure 2: Motif representation examples

Finally, we have **sequence logos**. Sequence logos do not actually differ from a PWM in the amount of information shown, but they sacrifice the accuracy of the numbers of the matrix in exchange of a clearer graphical representation which depicts a consensus sequences in which we can actually have an idea of the relative frequencies with which each nucleotide appear at a given position.

1.2.2.3. Motif finding methods

Motif discovery algorithms can be classified into three major types [3]: the first of these three types and the most important of them, which consists of methods based on promoter sequences of coregulated genes from a single genome can, in turn, be classified into word-based algorithms and probabilistic algorithms. The other two major types of algorithms are, methods based on phylogenetic footprinting (orthologous promoter sequences of a single gene from multiple related species) and methods based on promoter sequences of coregulated genes and phylogenetic footprinting.

Algorithms based on promoter sequences of coregulated genes

As stated previously, this is the most common way of searching for binding sites. It focuses on the overrepresentation of *k-mers* in a set of promoter sequences of coregulated genes. They can be classified in two main subtypes depending on their combinatorial approach.

Word-based algorithms (or string-based algorithms), the first of these subtypes, are based on the frequencies with which a given motif occurs in the set of sequences. They are especially appropriate when we are searching for motifs which are short and with very similar occurrences, which is the case of eukaryotic genomes, but theoretically they are not that accurate discovering TFBS, given that these are very often weakly constrained (their instances have many differences with each other). Apart from this, another drawback is that they produce a high number of false positives.

Probabilistic algorithms, the other major subtype of methods based on promoter sequences of coregulated genes, use probabilistic likelihood to predict the location of sites, using PWMs for the representation of motifs. The main advantage of this approach is that it is very often faster and more lightweight. In contrast with word-based algorithms, they are more suited for prokaryotic genomes, due to the fact that they are able to find longer motifs. Their biggest weak point, conceptually speaking, is that they do not guarantee finding globally optimal solutions, since they lack a local search mechanism. Two of the most widely used algorithms belong to this kind: expectation maximization (EM) and Gibbs sampling.

The **expectation-maximization algorithm (EM)** is an iterative method which assumes that there is an instance of the motif in each one of the input sequences to then try to find maximum likelihood estimates by, in each iteration, performing first an expectation (E) step and then a Maximization (M) step (hence the name of the approach) to maximize the expected likelihood. This approach allows both identifying sites and characterizing motifs concurrently.

Gibbs sampling methods **[6]** are based on a Markov-chain Monte Carlo (MCMC) that tries to obtain a sequence of observations from a specified multivariate probability distribution. It identifies the pattern that is the most statistically significant by finding the alignment that can maximize the ratio of the probability of the mentioned pattern in contrast with a background probability. This approach is a randomized algorithm (it produces different results each time it is run) and even though it is more than twenty years old, it is still used in many motif finding methods.

Algorithms based on phylogenetic footprinting

Phylogenetic footprinting is an approach that consists of the aligning of non-coding regions of DNA of orthologous sequences of closely-related species in order to determine if there are conserved motifs. The clear advantage that they offer over the algorithms based on coregulated genes is that a motif can be identified even if it is specific of a single gene, without the need of finding coregulated genes. These methods are growing thanks to the increasingly bigger collection of sequenced genomic regions of a wide variety of organisms.

Algorithms based on promoter sequences of coregulated genes and phylogenetic footprinting

This approach basically tries to integrate the advantages of both of the types previously presented, and for that purpose they combine the overrepresentation and conservation of motifs in order to avoid the drawbacks of both of the types as much as possible.

1.2.3. Genetic Algorithms

1.2.3.1. Overview

A genetic algorithm (GA) is an artificial intelligence method that, inspired by the process of natural selection, tries to find optimum, often approximate, solutions to any kind of optimization or search problem.

These kinds of methods start with a typically random population of candidate solutions and, by evaluating these solutions iteratively by a fitness function, they evolve them by imitating the mechanisms of crossover and mutation through a number of generations until an optimal solution is found.

In spite of their name, genetic algorithms are not only used in bioinformatics, but also in a variety of fields, from economics to automated design. The main challenge when designing a genetic algorithm is defining the fitness function, which should be able to successfully evaluate the individual solutions to decide which ones of them will survive to every generation, and additionally determining the correct representation of the population as well as the mutation and crossover functions.

1.2.3.2. Methodology

The way in which a GA evolve its collection of candidate solutions is by the use of an iterative process which goes through each one of the following steps:

Initialization

The initial set of candidate solutions (population) is usually generated at random, since, if the rest of the functions are properly designed, the algorithm will work with any kind of initial set of solutions. It is advised, thus, to avoid putting an excessive effort on creating the initial population and simply start by a random set.

Evaluation

The first step of every iteration of the algorithm is evaluating the fitness of each one of the candidate solutions (individuals). This step is the most crucial, and the overall quality of the algorithm will depend on how well designed the fitness function is. The fitness function can be of any type imaginable and it always depends on the type of problem we are solving. There are certain methods in which the population is very large and only a random portion of its individuals are evaluated in each iteration.

Selection

In this step, the fittest individuals (the ones who obtained the highest scores in the fitness function) are selected to survive, and the less fit individuals eliminated from the population. There are several techniques to do this, such as using a specific threshold to decide which ones survive and which ones do not (this threshold can be either constant or relative, for example, by making that only 50% of the population survive) and the elitist approach, in which the individuals are taken randomly in groups of two, making the fittest of the individuals of each pair survive and the other die.

Crossover

In order to add new solutions to the population, individuals are selected randomly in pairs that are recombinated to create new candidate solutions to add to the population. Since these new solutions preserve some of the characteristics of its parents (fit individuals), it is more likely that they will survive to subsequent generations.

There are several types of crossover techniques, from which the following are the most commonly used:

- One-point crossover: In this approach, the content of both parents is swapped from a randomly picked crossover point.
- Two-point crossover: The same idea as the one-point crossover but selecting two points instead of one.
- "Cut and splice": This approach is a variant of the one-point crossover, but in this case the crossover points are at different positions in each one of the parents, so the resulting offspring will have different lengths.
- Uniform Crossover and Half -Uniform Crossover: In this case, a fixed mixing ratio is used, so that the offspring will have a given ratio of content from each parent.
- Three parent crossover: This is not really a type a crossover, but a variation in which three parents are used instead of two.

Mutation

Mutation is necessary to guarantee diversity among the population, since, without it, it is possible that the population converge to a unique kind of individual, which would deprive the algorithm of finding an optimum solution. It works by happening randomly, according to a defined probability, affecting random individuals, which are slightly modified. The most crucial point of a good mutation function is finely tuning the probability with which it appears. If it is excessively high, it might lead to the disappearance of good candidates from the population, and if it is too low, on the other hand, the lack of diversity previously explained may happen.

Termination

Finally, a termination condition must be defined in order to set the most appropriate moment to stop. The conditions most commonly used are the moment in which a global optimum is found (if the fitness function is able to determine this optimum), the moment in which a certain number of generations is reached, and the moment in which the fitness scores get stuck at a given value (local optimum).

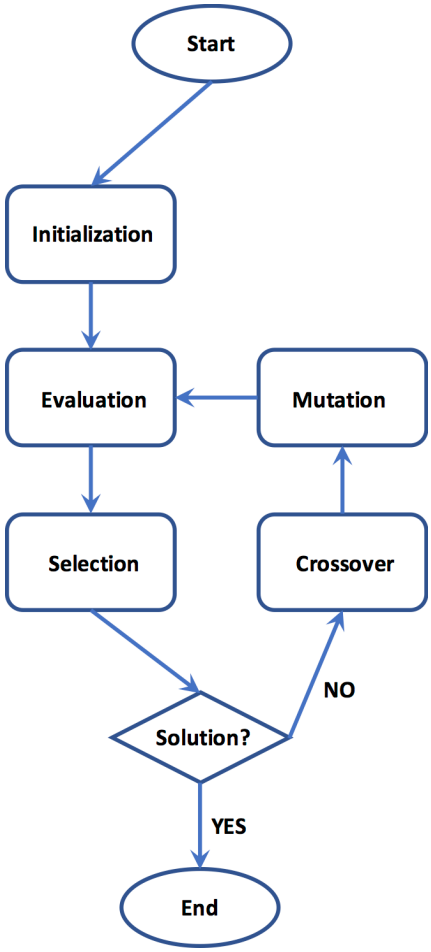

Figure 3: A typical genetic algorithm structure

1.2.4. Topic Models

1.2.4.1. Overview

Topic models are statistical algorithms which mix concepts of natural language processing and machine learning, and whose purpose is to find the topics contained in a set of documents and define the structure of their contents by hierarchical Bayesian analysis [7].

Topic models are able to determine the existing topics (defined as clusters of related words) in a set of documents and how they are distributed among them by examining the statistical properties of each of the words of a given vocabulary. Their main advantage is that they are able to give us insights of large collections of documents without actually having to read each one of them, which is considerably useful in an era in which huge amounts of information need to be analysed in the shortest period of time possible.

Even though they were originally created with the only purpose of analysing ordinary documents, their structure makes them potentially interesting to be applied in other fields.

1.2.4.2. Methodology

Computationally speaking, topic models are formally defined as a multinomial distribution over a given vocabulary in a set of documents. This could be explained in a clearer way by stating that topic models consider a fictional scenario in which the way in which we write consists of picking words from different boxes that contain them and mixing them into a document. For example, if we desire to write a document about climate change in Greenland, and we want it to be three fourths related to climate change and one fourth to Greenland, we would take two boxes with the tags "climate change" and "Greenland", each one of them containing words related to that topic, and we would combine them with each other, taking 75% of the words from the former box and 25% from the latter, and with other words common to any topic (which topic models discard from their vocabularies) to form our document. In this scenario, which is a conceptualization which seems far from the actual process but it is quite plausible from a computational point of view, the boxes that contain the words are the topics in the topic models. The question that topic models try to answer is then, *can we reverse this process and determine the boxes from the statistical properties of the distribution of the words in the documents?*

The intuition behind topic modelling is that the distribution of the words shared by a set of documents will give us the proportion with which each one of the documents treat a given topic.

For that purpose, we would have the following inputs for the algorithm:

- A vocabulary formed by N words
- A collection of M documents ($d_1, ..., d_M$)
- The estimated number of topics S that will be found in the documents

And, generally speaking, the basic steps to follow for the topic model to determine the topics would be as follows:

- For each document d_i in the input set:
    1. Select a random distribution R of d_i over the K topics (t_1, ..., t_S).
    2. For each word w_j in d_i:
        a. Pick a topic t_x randomly from the distribution R and assign w_j to it.
        b. For each one of the topics t_k in R:
            i. Calculate the proportion of words assigned to the topic t_k in the document d_i at the current moment, $P(t_k|d_i)$
            ii. Compute the proportion of assignments originating from the word w_j to the topic t_k over the whole set of documents, $P(w_j|t_k)$
            iii. Reassign w_j to that topic which has the highest probability, computed by the product of both of the probabilities previously calculated, $P(t_k|d_i) * P(w_j|t_k)$.

All of these steps would be repeated iteratively until a stable set of topic assignments is obtained.

In practice, topic models adopt a structure more complex than the basic workflow explained above, and there are several types of algorithms depending on the different approaches.

1.2.4.3. Types of topic models

Latent Dirichlet Allocation (LDA)

This is the most common and simplest type of topic model and it works by assuming that the topic distribution has a sparse Dirichlet prior, that is, that the documents contain only a small set of topics which, in turn, are represented by a small set of meaningful words. Its major limitation is its inability to model correlations between topics [8].

Pachinko Allocation (PAM)

This algorithm provides more flexibility and greater expressive power than LDA by modelling correlation between topics in addition to the characteristics of LDA [9].

Correlated Topic Model (CTM)

As well as the PAM algorithm, CTM also models correlation between topics [10]. The main characteristic of the algorithm is the logistic normal distribution, which, through the transformation of a multivariate normal random variable, allows for a general pattern of variability between the components of the distribution [11].

Main Chapter: A study on the application of topic models to motif finding algorithms

2. A study on the application of topic models to motif finding algorithms

2.1. Introduction

Topic models consider that from a set of sequences and a vocabulary of important words contain in them, a distribution of topics among the documents can be obtained. The question that this chapter tries to answer is, if we can consider DNA sequences as documents in which the words are k-mers formed by nucleotides, can we infer the motif structure of a set of sequences by defining a vocabulary of meaningful k-mers?

To give an answer to that question, two methods were developed for this study.

CTM Method

The first method here presented tries to directly answer the proposed question by defining a motif finding algorithm based on topic models **[12]**. A diagram representing a typical topic model with all of its elements and their corresponding counterparts for motif finding is shown in **Fig. 4**.

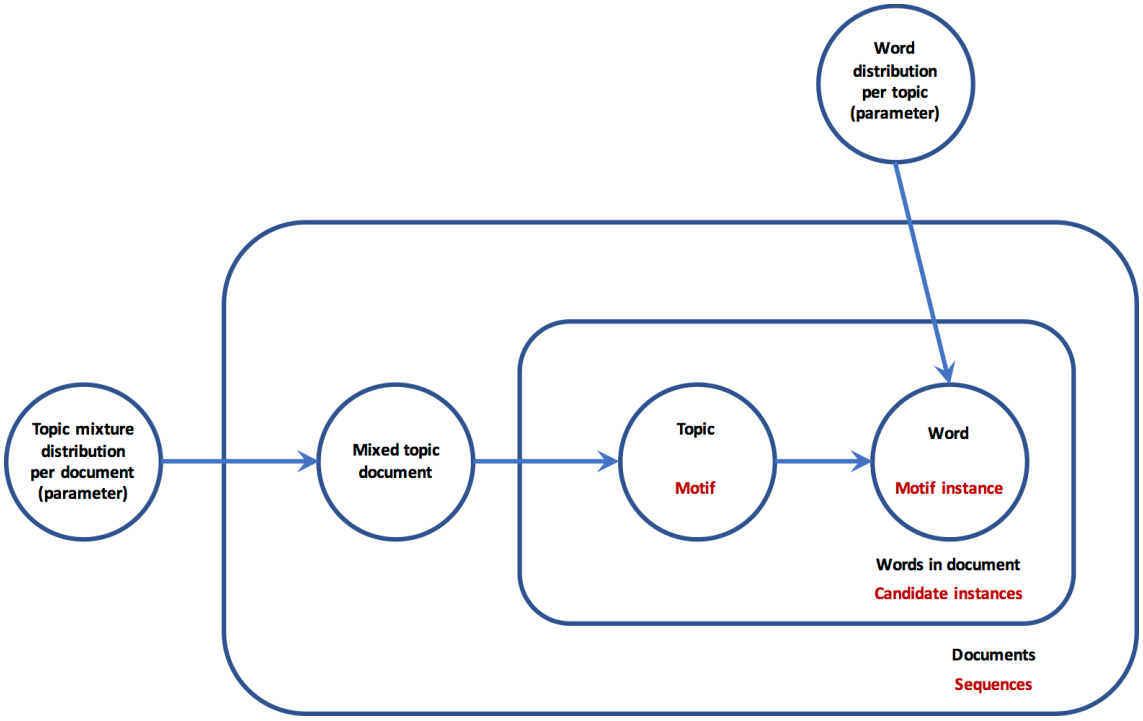

Figure 4: Adaptation of topic models to the motif finding problem (Adapted from Gutierrez and Nakai [12])

This algorithm, thus, works as a topic model in order to analyse a set of input sequences and find the distribution of a collection of k-mers in clusters which would represent motifs.

Statistical GA

Additionally to the method just described, in this chapter we also examine if topic models can be used in combination with other approaches. The motivation behind this is that, as all of the basic types of motif finding methods have important drawbacks, it seems obvious that the best option should be a combination of several methods that can overcome those issues.

Prior to this study, we designed another motif finding method, which was based on a GA and the use of several statistical coefficients in a complex fitness function subdivided in three steps [13].

This algorithm showed a remarkable performance in comparison with other methods. Nonetheless, it had an important drawback, which, as it is common in other word-based methods as well, was a high rate of false positives. One of the main reason for this was that it was unable to measure the confidence with which it reported a solution. In other words, it always reported at least a motif per data set, since it lacked of a mechanism through which to know if the best solution possible was actually good enough for the problem.

In addition to the CTM Method previously presented, in this chapter we will also study how the addition of topic models can dramatically reduce the number of false positives reported by this algorithm, to which we will refer from now on as the Statistical GA method.

2.2. Creating a motif finding algorithm based on topic models from scratch: The CTM Method

2.2.1. Strategy

Once we have adapted a topic model to motif finding, the first difficulty we come up against when we start to design the algorithm is the absence of actual words to form a vocabulary. In contrast with a document, which has obvious words clearly separated by spaces, a sequence is just a very long string in which finding meaningful words is not an easy task, and that is precisely one of the main reasons why motif finding is such a challenging problem.

A topic model, however, deals with a similar issue when defining the vocabulary. Even though the words are clearly identified, most of these words, which do not add information about the topics, are actually useless for the algorithm and would only add noise to it. Articles, prepositions or adverbs are examples of these "noisy" words.

Therefore, the strategy in our case would be similar, and instead of eliminating meaningless words, the process to follow consists of searching for words which are repeated with certain frequency and add them to the vocabulary.

However, we find a new issue at this point, which is the impossibility of finding all of the possible meaningful words for our vocabulary without having to spend a considerable amount of time an effort.

To overcome this problem, we decided to use the structure of a GA for the problem in order to select the optimum vocabulary for our topic model.

As for the type of motif finding algorithm, it is a word-based algorithm. Even though most word-based algorithms assume that every sequence contains at least one instance of each motif, in the case of the CTM Method, there is no specific expectation about the number of occurrences and any possibility is considered. It is stochastic, since it most likely reports a different set of solutions every time it is run, and given that these solutions are approximate, we can classify it as well as a heuristic algorithm. As the vocabulary is formed by simple words, the method predicts only ungapped sites. Motifs which contain gaps, however, can be found split into its different parts and it would correspond to the researcher to locate them and link these parts.

2.2.2. Structure

As mentioned previously, the structure of the CTM Method is that of a GA, with the following definition for each one of its parts:

Initialization

In order to initialize the population, it is very important to define how this population will be represented. As our goal is to find an optimum vocabulary for our topic model, each individual of the population must be a candidate vocabulary, and, as it usually happens in a GA, these individuals are initialized randomly.

More formally, an individual or candidate vocabulary is defined as a collection of n k-mers, being n a fixed number predefined as a parameter. The length k of each k-mer, on the other hand, is variable and randomly chosen at this step between a minimum and a maximum, both also defined as parameters. The population size is also predefined as a parameter. We can now define the initialization process as follows:

For a data set formed by M sequences $S_0,..,S_M$ of any length, a minimum k-mer length k_{min}, a maximum k-mer length k_{max}, a population size N, a number of words per candidate vocabulary n, and a minimum number of occurrences c_{min} for every k-mer in the whole set of S sequences:

- For each individual N_i:
    1. Pick a length k randomly within the range $k_{min} : k_{max}$.
    2. Pick a sequence S_i randomly.
    3. Pick an index j in S_i, being j a random integer between 0 and $length(S_i)$.
    4. Obtain the word w, from the k-mer $S_i[j:j+k-1]$.
    5. Compute $c(w)$: the number of instances of w in the whole set of sequences $S_0,..,S_M$, allowing for a 25% of mismatches (each instance needs to overlap 75% of the nucleotides of w).
    6. Shuffle the nucleotides contained in w to form a new word $w_{shuffled}$, and compute $c(w_{shuffled})$.
    7. If $c(w) - c(w_{shuffled}) >= c_{min}$, add w (represented as a tuple with the number of the sequence in which it was found, the index in which it starts in the sequence, and its length: (i, j, k)) to N_i.
- Repeat the above steps until $size(N_i) = n$.

In contrast with a typical topic model, repetition of words is allowed, since they will be represented by different pairs sequence-position, and motifs are expected to contain several exact instances of the same word.

Evaluation

In the evaluation step is when the power of topic models come into play. But before going further, it is crucial to select the most appropriate type of topic model.

Since, from a biological point of view, motifs are correlated to each other (they collaborate with each other in the transcriptional regulatory process), it seems logical that the best option is a topic model that considers correlation between topics. Given that a CTM is more advanced and versatile than a PAM, we decided to use a CTM in our algorithm **[10]**. In particular, we applied the CTM defined in the *topicmodels* R library **[14]**.

The evaluation step works, therefore, by, for each individual, building a CTM to which we feed the corresponding vocabulary (collection of k-mers of the candidate individual) and the set of sequences (in which each sequence is represented as a list of the words of the vocabulary along with their corresponding number of occurrences in the given sequence).

However, once the CTM is built, it is still required to decide how the fitness function will measure the quality of the current model. For that purpose, topic models count on the **perplexity** measure.

So, in summary, the fitness function consists of creating a CTM with the vocabulary of the given individual and the original set of documents and computing the perplexity of this model.

Perplexity

The perplexity is a measure used in natural language processing to evaluate how accurately a sample is predicted by a probabilistic model. It is computed by separating the data set into a training set and a test set and calculating the log-likelihood of the samples in the test set after training the model with the training set. The lower the perplexity, the more accurate the probabilistic model is. The following is the corresponding formula used to calculate the perplexity in our algorithm, which is the one provided by Hornik and Grün for their CTM implementation in R **[14]**:

$$Perp(\omega) = exp\left\{-\frac{\log(p(\omega))}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)}}\right\}$$

where $n^{(jd)}$ is the frequency with which the *j*th word (k-mer) appears in the document (sequence) *d*.

Selection

We chose an elitist approach for the selection step. That is, the individuals of the population are randomly picked in pairs, to then, for each pair, keep the one with the lowest perplexity in the population and discarding the other. At the end of this step, half of the population remains in our set of individuals (*N/2*).

Crossover

To fill up the population until we have *N* individuals again, the crossover steps follows a one-point crossover strategy. Again, individuals are selected in random pairs, to then pick a random index *i* (crossover point) between *0* and the individual vocabulary size *n*, and swap the words from both parents around that index. Two new children individuals will be born: one formed with the words from *0* to *i* from the first parent and the words from *i+1* to *n* from the second parent, and another one with the words from *0* to *i* from the second parent and the words from *i+1* to *n* from the first parent.

Mutation

The mutation step happens only at certain moments depending on a frequency defined as an initial parameter. When mutation happens, an individual is selected randomly from the population and a random number of the words in its vocabulary are slightly shifted by a random number of positions *r*: If the k-mer *(i, j, k)* (found in the sequence *i*, at the index *j*, and with length *k*) is affected, it becomes *(i, j+r, k)*, being $j+r <= length(S_i)-k$.

Termination

The GA is terminated when a given number of generations (passed as a parameter) is reached. When the iterations are finished, the *x* surviving individuals with the highest perplexity are selected (being *x* a parameter representing how many candidate solutions the researcher wants to further study), for each one of them a new CTM is built, and the resulting topics reported as motifs of the set of sequences.

2.3. Improving a previously developed algorithm by the use of the perplexity measurement: The Statistical GA

2.3.1. Description of the original algorithm

The original Statistical GA algorithm [13], as its name suggests, is based on the structure of a GA, which in this case is modified so that the evaluation and selection steps are merged and divided in three sub-steps, each one of them using different statistical coefficients to calculate a different fitness value and discard unfit solutions. The three coefficients that ultimately evaluate the fitness of the individuals of the population are the Fluffiness Coefficient [15], the Thinness Coefficient [16], and, the most important of the three, the Mann-Whitney U-Test [17].

The other addition of this algorithm is that, instead of using the input sequences as they are, it joins them in a random order into a supersequence to then divide it into subsequences of a given length passed as a parameter (but keeping the information about the position in which the original sequences begin and end in the supersequence in order to avoid broken k-mers), so that, in each iteration, the fitness is calculated against a given subsequence and the individuals underrepresented in it are swiftly discarded, avoiding unnecessary calculations.

Each individual of the population is a word represented by a position in the supersequence and a fixed length (parameter). The algorithm is typically run several times for different lengths of the k-mers, and then the results combined.

Fig. 5 shows graphically the structure and flow of the algorithm, highlighting the parts which differ from a traditional GA.

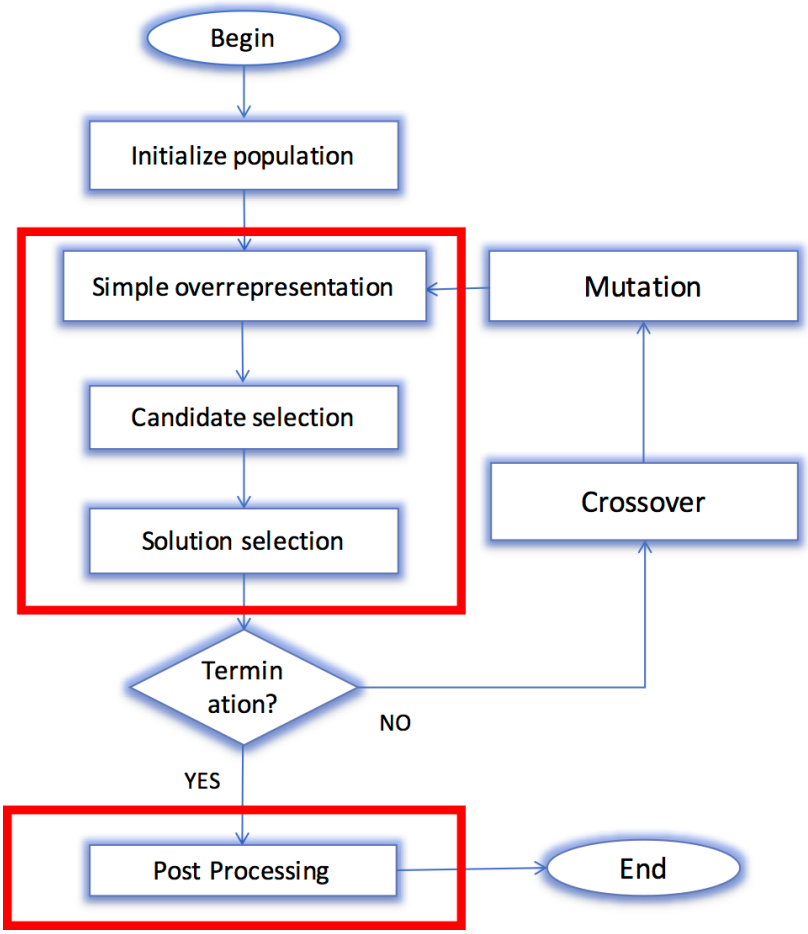

Figure 5: Structure and flow of the Statistical GA. In red, the additions in relation to a classical GA (Adapted from Gutierrez and Nakai [12])

2.3.2. Addition of topic models

The Statistical GA proved to be a promising method, showing a noteworthy performance in comparison with other popular motif finding tools. However, it had a clear disadvantage, which was the excessive number of false positives reported. There were several reasons for that, but one of the most important reasons was its inability to measure the confidence of the results reported. In other words, it always predicted the best candidates, even if those candidates were still not good enough for the data set.

To overcome this drawback, the solution was to use a CTM in a similar way as it was used in the CTM Method, but in this case only in the post-processing stage, in which originally a clustering process is already carried out. The idea is to take the set of final solutions returned by the GA and build a CTM using them as words of a vocabulary. If the CTM is unable to cluster them as topics (motifs) with confidence, then the perplexity score would be too high. The number of topics is set at 10, as an arbitrary number of motifs expected to be found in a set of sequences higher than 1 but not especially high. Taking advantage of this, our Statistical GA now uses the perplexity as the way to measure the confidence with which a solution is reported as a motif of a set of sequences. If this perplexity is lower than 100, the motif is considered a final solution. If not, it is discarded. **Fig. 6** shows how the new Statistical GA works after combining it with topic models.

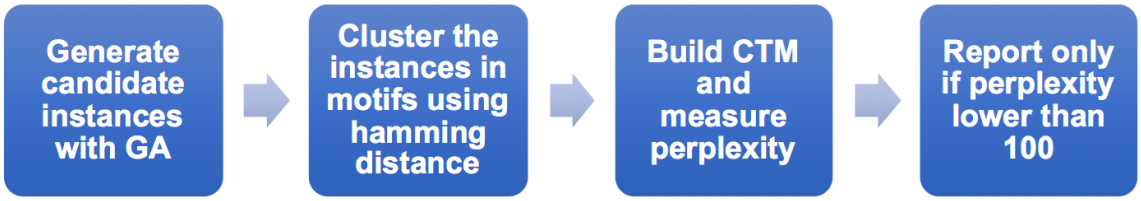

Figure 6: Flow of the Statistical GA after the addition of topic models (adapted from Gutierrez and Nakai [12])

2.3.3. Assessment

In order to test the performance of the two methods proposed, we decided to use the assessment designed by Tompa *et al.* [2]. This assessment tried to set a benchmark to measure the performance of motif finding methods, by focusing on how accurately they predict sites, and leaving computational performance and execution times on the side.

The benchmark is formed by 52 data sets of four different organisms: human, mouse, fly and yeast. These 52 data sets are, in turn, classified into three different kinds. Some of them are actual promoter sequences containing actual sites (Type Real), others are synthetic random sequences generated from the corresponding genome with planted binding sites (Type Generic), and finally there are synthetic sequences with embedded sites generated by a Markov chain (Type Markov). In addition to the 52 data sets, there are 4 negative controls containing no sites, making a total of 56 data sets.

The original study evaluated the performance of 14 methods, to which we added our two proposed methods, for a total of 16 methods. **Table 2** shows the list of these methods, each one accompanied by a brief statement of their methodologies and the corresponding reference.

Table 2: List of the motif finding methods compared in the assessment

Tool	*Methodology*	*Reference*
AlignACE	Based on Gibbs Sampling. Measures the overrepresentation by log likelihood.	[18]
ANN-Spec	Determines DNA binding specificity by PWM.	[19]
Consensus	PWM for the representation to then try to find the most informative matrix.	[20]
GLAM	Based on Gibbs Sampling. Automatically optimizes the motif width.	[21]
The Improbizer	Uses EM to find statistically improbable PWMs.	[22]
MEME	Uses EM to optimize E-value.	[23]
MEME3	Improves MEME's accuracy by using a correction factor in the objective function.	[23]
MITRA	Hypergeometric score of the occurrences of each candidate motif against a set of background sequences.	[24]
MotifSampler	Mix of Gibbs Sampling and a Markov model.	[25]
Oligo/dyad-analysis	Counts occurrences of each k-mer against the expectation of a negative binomial distribution.	[26], [27]
QuickScore	Exhaustive searching and a background Markov model.	[28]
SeSiMCMC	Modification of Gibbs sampling in combination with a Markov model.	[29]
Weeder	Exhaustive oligo frequency analysis and a consensus-based algorithm.	[30]
YMF	Exhaustive search and z-score for selection of motifs.	[31]
Statistical GA	A modified GA and a combination of three statistical coefficients.	[13]
CTM Method	Uses CTMs with the structure of a GA.	[12]

The tests were performed by allowing each method to report no more than one motif per data set (in other words, either zero or one motif). The format for the solutions reported is as a list of the sites that form the motif, each one along with the corresponding sequence and position in which they appear.

For the evaluation of the accuracy of the motifs reported, the assessment uses the eight statistics described in **Table 3**.

Table 3: Statistics to measure the performance of the methods in the assessment

	Coefficient	*Formula*
Nucleotide level	**nSn**	$nSn = \frac{nTP}{nTP + nFN}$
	nPPV	$nPPV = \frac{nTP}{nTP + nFP}$
	nSP	$nSp = \frac{nTN}{nTN + nFP}$
	nPC [32]	$nPC = \frac{nTP}{nTP + nFN + nFP}$
	nCC [33]	$nCC = \frac{nTP \times nTN + nFN \times nFP}{\sqrt{(nTP + nFN)(nTN + nFP)(nTP + nFP)(nTN + nFN)}}$
Site level	**sSn**	$sSn = \frac{sTP}{sTP + sFN}$
	sPPV	$sPPV = \frac{sTP}{sTP + sFP}$
	sASP [33]	$sASP = \frac{sSn + sPPV}{2}$

TP refers to the number of true positives, TN to the true negatives, FP to the false positives, and FN to the false negatives. At site level, a predicted site is considered to match the corresponding site in the sequences if it overlaps at least 25% of it.

2.4. Results

To test our proposed methods, we used the following parameters:

- CTM Method
    - Maximum number of iterations for the GA: 90
    - Number of candidate solutions (individuals in the population): 50
    - Number of words per candidate vocabulary: 1000
    - Minimum word length: 6 bp
    - Maximum word length: 30 bp
    - Number of expected topics in each CTM: 10
    - Mutation rate: 0.1
    - Maximum number of reported solutions (vocabularies): 10
- Statistical GA
    - 3 runs with different k-mer lengths:
        - Length 8 bp, allowing for 2 mismatches
        - Length 10 bp, allowing for 3 mismatches
        - Length 12 bp, allowing for 4 mismatches
    - Maximum number of iterations for the GA: 100
    - Number of candidate solutions (individuals in the population): 200
    - Length of the subsequences: 500 bp
    - Mutation rate: 0.1
    - Maximum number of reported solutions (instances): 100
    - Maximum similarity (for clustering): 0.7
    - Number of expected topics in each CTM: 10
    - Maximum perplexity: 100

And then, for each one of them, selected the solution with the best score (fitness value) as the reported motif (or none if the perplexity is lower than 100).

In order to study the results, we computed the average scores for each one of the statistics following the same method as in the original assessment. This method consists of the following:

1. For each *nTP*, *nTN*, *nFP*, *nFN*, *sTP*, *sTN*, *sFP*, *sFN*, calculate the sum of the given value among all of the corresponding data sets.
2. Once we have the sums of the eight basic values, compute the scores of the eight statistics using the corresponding sums. Example:

$$avg(sSn) = \frac{\sum sTP}{\sum sTP + \sum sFN}$$

The figure **Fig. 7** shows the average scores of the main seven statistics (all of them except the specifity, nSp, which is always near 100 and makes the charts confusing) for all 56 data sets, and the figures **Fig. 8**, **Fig. 9**, **Fig. 10** and **Fig. 11** show the average scores divided by organisms. **Fig. 12** shows the improvement in each one of the statistics of the Statistical GA after the addition of the perplexity measure to evaluate the confidence with which a solution is reported.

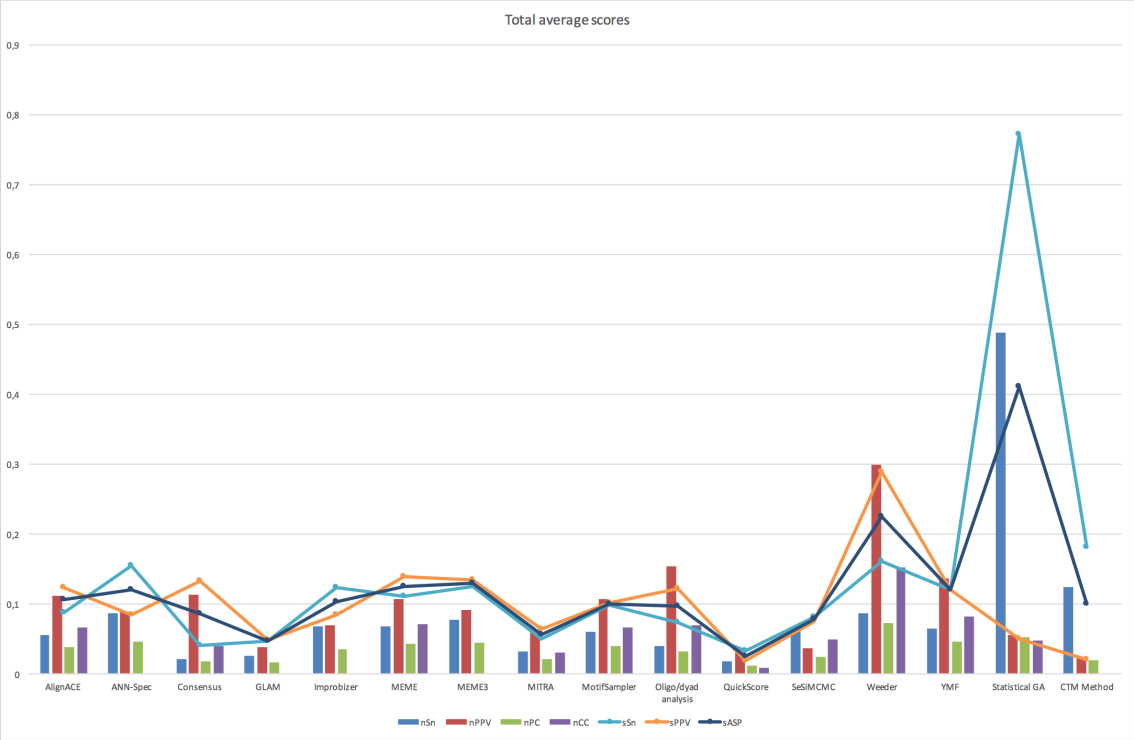

Figure 7: Total average scores of the 16 methods in the 56 data sets (adapted from Gutierrez and Nakai [12])

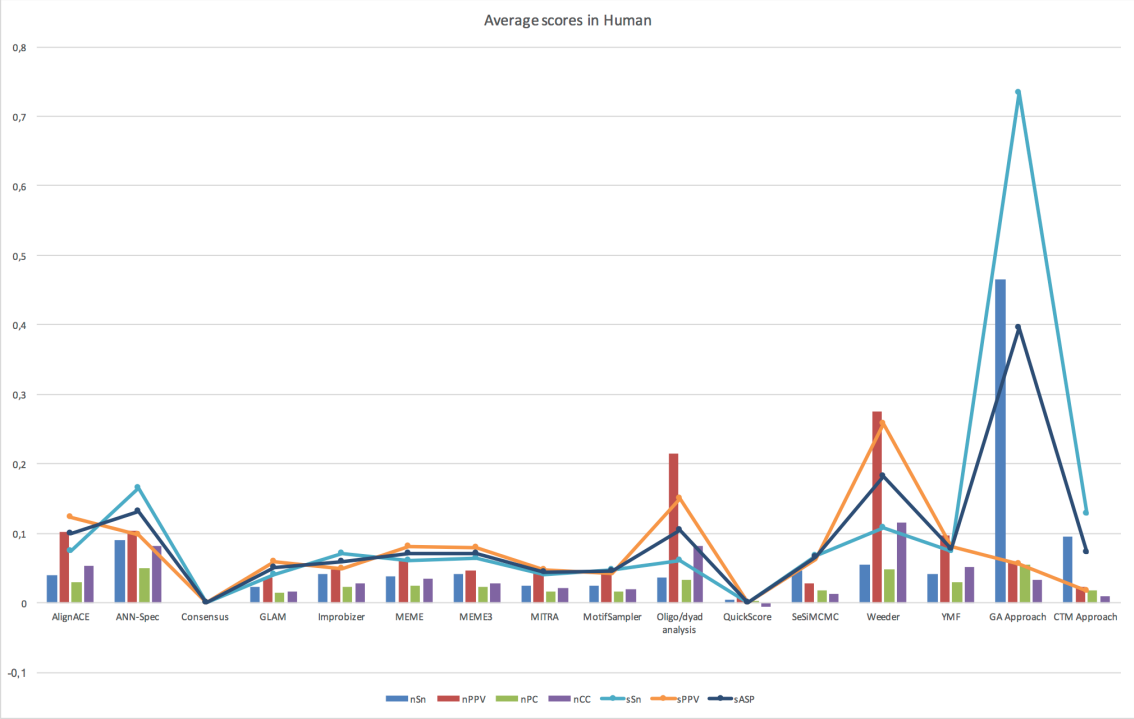

Figure 8: Average scores of the 16 methods in the Human data sets (adapted from Gutierrez and Nakai [12])

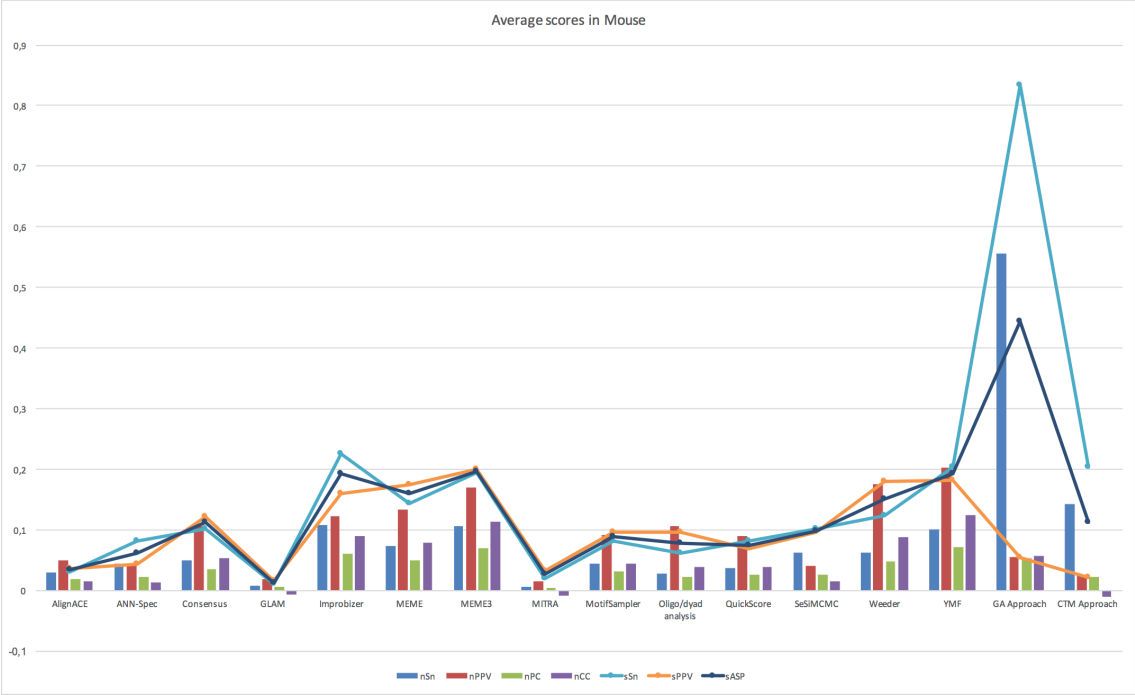

Figure 9: Average scores of the 16 methods in the Mouse data sets (adapted from Gutierrez and Nakai [12])

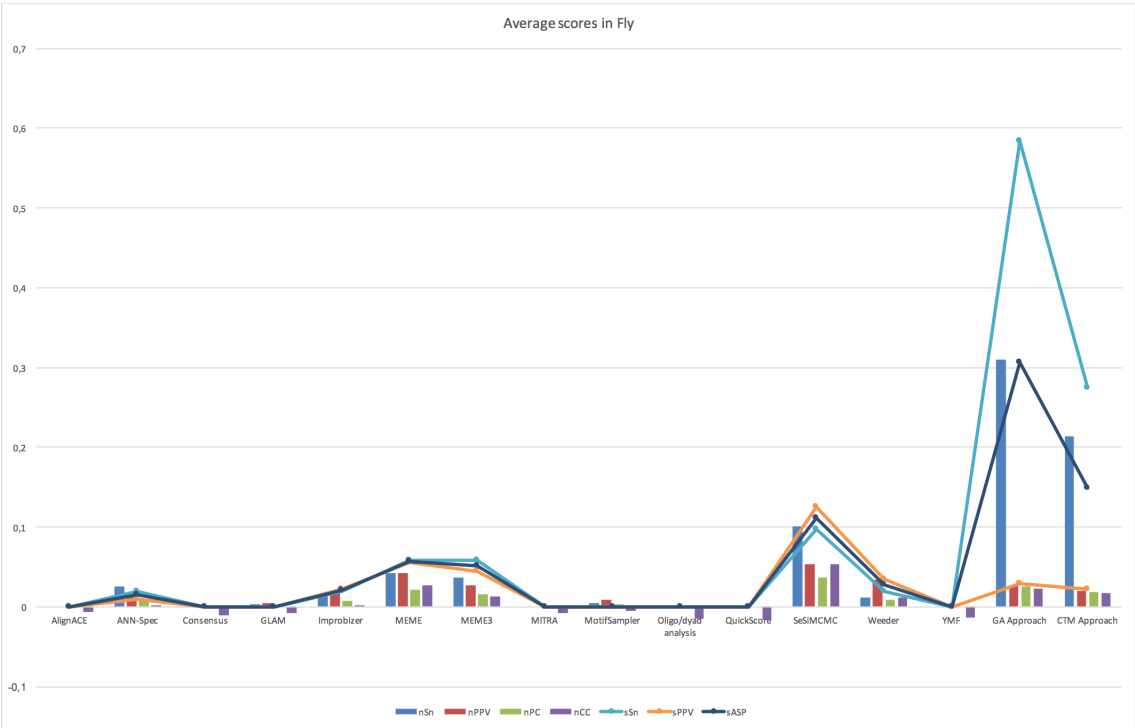

Figure 10: Average scores of the 16 methods in the Fly data sets (adapted from Gutierrez and Nakai [12])

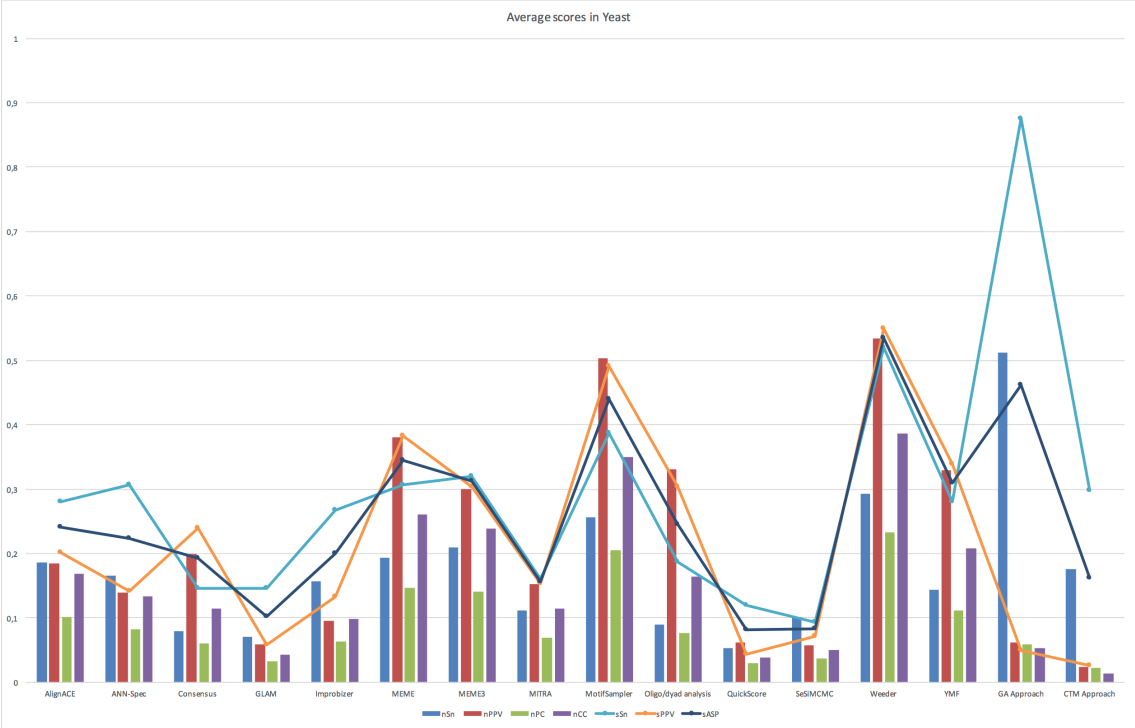

Figure 11: Average scores of the 16 methods in the Yeast data sets (adapted from Gutierrez and Nakai [12])

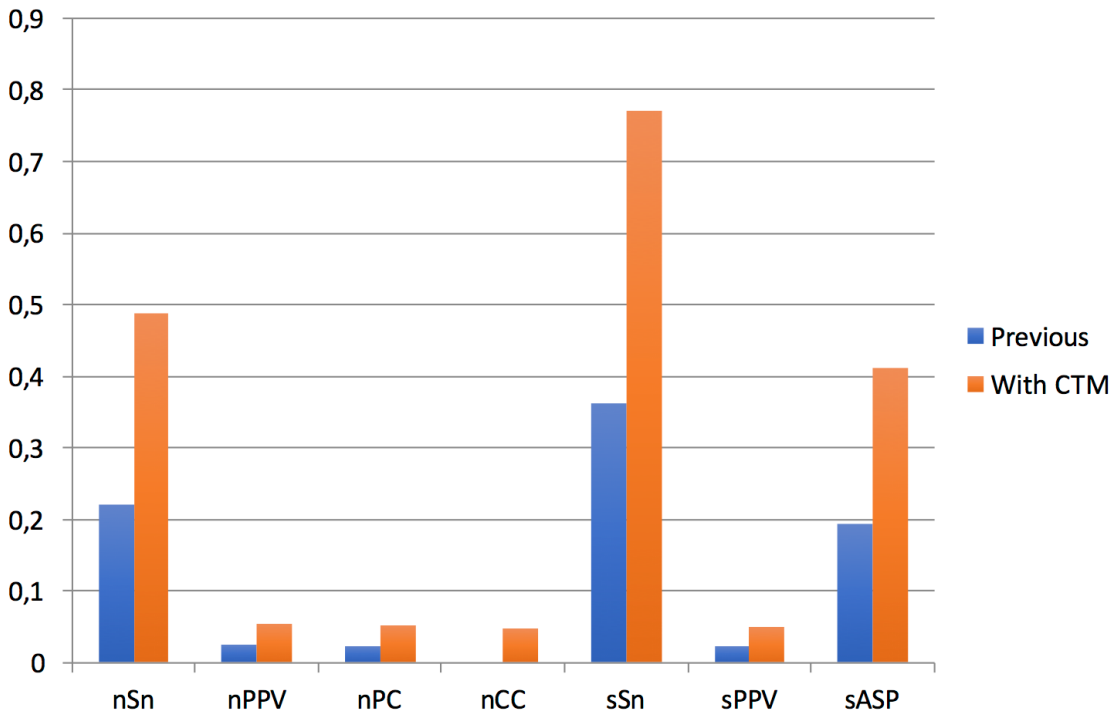

Figure 12: Total average scores of the Statistical GA before and after the addition of topic models (adapted from Gutierrez and Nakai [12])

2.5. Discussion

When it comes to evaluating the performance of motif finding methods, given the fact that the motif finding problem itself has no mathematical perfect solution and the underlying biology is yet to be completely understood, it is extremely challenging to design a benchmark able to take into consideration the whole variety of advantages, disadvantages, organisms, types of DNA sequences, etc., and give an indisputable claim about which method is the best and which one is the worst.

Considering that we cannot take the benchmark as a source of the absolute truth about the quality of the methods studied, we can still get a good grasp about their performances if we keep in mind the following points that could make the results favour some methods and be detrimental to others when computing their accuracies [2]:

- All of the binding sites considered in the assessment were obtained from TRANSFAC [34]. These binding sites are supposed to be experimentally validated, but there is always the possibility of an error in the data base, or an error of the researchers when obtaining the information.
- Most of the methods analysed, including both of our proposed tools, work with motif lengths no longer than 30 bp. However, some of the sites included in the benchmark are longer than that.
- Each of the methods studied are required to report either one or none motifs, even though some of the data sets (especially those of Type Real) might contain more several different motifs.
- The fact of limiting the reporting to just zero or one motifs affects also the behaviour of the tools, since they would usually report several motifs and there is the possibility of the appearance of some arbitrariness when picking just one of them.

- At the moment of calculating the average scores, we stated that a specific method was used for it. However, other different ways to make these calculations could have been used, providing different outcomes. In particular, the strategy used tends to favour those methods which report no motifs in most of the cases rather than those which report more correct motifs but make some mistakes in their predictions as well. This explains why the Statistical GA improved so dramatically after the addition of topic models to avoid wrong predictions.

That being said, if we take a look at the average scores (**Fig. 7**, **Fig. 8**, **Fig. 9**, **Fig. 10**, **Fig. 11**), we can verify that the CTM Method is second in both of the Sensitivity scores (at nucleotide level and at site level), being the only tool able to outperform it our Statistical GA. It is also considerably accurate at Average Site Performance. In PPV, on the other hand, it gives somewhat poor scores. The rest of the scores are satisfying enough though not significantly high.

From this we can infer that it is able to report many true positives but needs to improve the number of false positives reported, which is high in contrast with other methods. This was expectable, given its word-based nature and the fact that, in a similar way as it happened with the Statistical GA prior to the addition of topic models, it lacks of a mechanism to measure the confidence of the results reported. It is likely that this would improve if used in combination with other approaches, as in the case of the Statistical GA.

In the case of the mentioned Statistical GA, in **Fig. 12** we can see how all of its scores, roughly speaking, were doubled after the addition of the CTM to measure the confidence of the results.

Now this tool is visibly the one with the best performance in the assessment, especially in Sensitivity (at site level it is near 80%, which in theory would mean that almost 80% of the sites were successfully predicted by the tool) and Average Site Performance. It still requires some improvement in the number of false positives reported, since, even though it now report motifs likely to be correct, it still reports an excessive number of occurrences for each motif. Once a mechanism to detect the correct instances of each motif is added, it could become a very accurate tool.

Also, it is worth mentioning that, if the format of the solutions required by the assessment had been matrices, both of our methods would have improved their scores in Positive Predicted Value, due to the fact that the reason why they report many false positives is the mentioned excessive number of instances per motif, but the high rate of true positives proves that the motifs, represented as a matrix or a consensus sequence are correct in a high percentage.

As for the computational times, which are also very important for scientists, as mentioned before, they are not taken into account by the assessment, so we avoided this comparison as well. Nevertheless, some facts can be mentioned about the computational performance of our methods: The Statistical GA seems to be remarkably fast with very large data sets, but slower than other methods in the case of small data sets. This can be avoided, though, by tuning the parameters according to the size of the data set (in this study we used the same parameters for every data set). The CTM Method, on the other hand, is considerably slower, but it clearly improves when the number of sequences is three or less. So, this could be improved in the future by, for example, following a similar strategy as the Statistical GA and joining the input sequences in supersequences so that the number of them is never higher than three.

In brief, considering the goal of this study, which was proving the suitability of topic models to the motif finding problem, we believe the results obtained are satisfying enough to assert that the initial hypothesis was true. In addition, similarly to what happens with most studies about the performance of motif finding tools [3], we also reach the conclusion that the best approach for motif discovery is always using several methods in combination instead of relying on just one. The success of the Statistical GA by combining the original version of the algorithm with topic models further proves this statement and provides us with the idea that, when it comes to designing motif finding algorithms, it is fundamental as well to mix ideas of several techniques, in order to get the advantages of all of them and avoid their drawbacks as much as possible.

Additional Chapter: *In-silico* analysis of structural features of transcriptional regulatory regions

3. *In-silico* analysis of structural features of transcriptional regulatory regions

3.1. Introduction

3.1.1. The mechanisms of transcriptional regulation in humans

Transcriptional regulation in humans is a process that involves a wide range of transcription factors, cofactors and chromatin regulators acting cooperatively. If during any of these steps there is a misregulation of gene expression, many different diseases and syndromes can be caused by it. Therefore, understanding the whole regulatory process as well as all the elements and mechanisms involved in it, might lead to the discovery of how those disorders appear, and ultimately how to prevent them and treat them. This problem, given its complexity and the considerable number of elements and mechanisms involved, has always supposed a considerably challenging task for the scientific community.

In the most recent years, fortunately, there have been many noteworthy advances in our understanding of the mechanisms of control of gene expression which situate us one step closer to the final goal. One of the most important examples is the discovery that control of transcriptional regulation is very often carried out by only a very small fraction of the hundreds of transcription factors that are present in the cell [35]. The way this regulation is controlled is by following a process in which the aforementioned fraction of key transcription factors bind cooperatively to individual enhancer elements and targeting genes by the use of cofactors and RNA polymerase II.

Enhancers are thus a fundamental resource for identifying genetic variations that might alter binding of TFBS and lead to disease. In fact, recent studies suggest that a considerably important portion of these disease-associated variations occur precisely in enhancer regions [36].

Regarding cofactors and chromatin regulators, although they are usually expressed in most cell types, when mutations occur in these genes, it generally leads to tissue-specific phenotypes.

Therefore, to fully understand the mechanisms responsible of tissue-specific diseases, it is fundamental to improve our knowledge of the interactions between transcription factors and both cofactors and chromatin regulators.

In summary, unravelling the mechanisms responsible of tissue-specific diseases and syndromes is a difficult task, but recent developments in the field showed that focusing our attention on the following crucial points might make the task more approachable to scientists:

1. Identification of the few **key transcription factors** responsible of the control of the transcriptional regulation of the given tissue.
2. Identification of the **enhancers** involved in gene expression.
3. Identification of the **cofactors and chromatin regulators** that interact with the corresponding key transcription factors.

3.1.2. Cerebellar transcriptional regulation

In this study, we chose the cerebellum as the tissue to study the workflow proposed for the analysis of transcriptional regulatory regions by their structural features. Initially, the main motivation to pick the cerebellum as the target of the study was merely a matter of availability: There was enough data available to obtain a full data set of cerebellum-specific promoters and another one of cerebellum-specific enhancers. However, there are other reasons that make this choice potentially interesting for future research.

The function of the cerebellum in humans has always been considered to be exclusively motor coordination. Nonetheless, some recent studies are suggesting that its function could actually be not only limited to that, but, in opposition to traditional belief, it seems possible that it could also be implicated in cognition and emotion. If this theory was true, then genetic variations in the cerebellar pathways might

lead to diseases unrelated to movement [37].

For that reason, identifying the tissue-specific key transcription factors, enhancers, cofactors and chromatin regulators responsible for transcription regulation in the cerebellum could provide a few more clues to answer to the question of whether there are other functions in which the cerebellum is involved apart from motor coordination. And, in addition to this, it could contribute to the development of efficient gene therapy for cerebellar genetic disorders.

This chapter contains an *in-silico* analysis of cerebellum-specific transcriptional regulation in which, in a first step, the binding sites of the key transcription factors are predicted in two data sets formed by cerebellum-specific promoters and cerebellum-specific enhancers respectively, to then study their structural features and evolutionary conservation, with the ultimate goal of obtaining a narrow set of transcription factors, promoters and enhancers presumably involved in the cerebellar transcriptional regulation.

3.2. Methodology

3.2.1. Outline

The method starts by obtaining reliable data sets of cerebellum-specific promoters and cerebellum-specific enhancers. Then, to determine which are the key transcription factors in cerebellar transcriptional regulation, motif finding was performed with several tools on both data sets. Both known and de novo motifs are considered, trying to add information that previous studies on other tissues did not take into consideration [38]. Then, a set of candidate key TFBS is built by selecting those binding sites which are either overrepresented in the promoters and present near peaks in enhancers, or overrepresented in enhancers and present near peaks in the promoters. After these candidates are selected, their positional and structural characteristics in the promoters are analysed, given that previous studies proved that these features might as well be important for transcriptional regulation [39][40]. Additionally, some studies determined that transcription factor binding locations associated to tissue-specific biological pathways and disease loci usually occur in combination and are evolutionary conserved and deeply shared among species [41], so in this analysis evolutionary conservation, as well as the previously mentioned co-occurrence of binding sites, are also taken into account. After studying their structural features and their evolutionary conservation, the key transcription factors in cerebellar transcriptional regulation would be identified and the promoter structure of the cerebellum would be outlined. Fig. 13 shows graphically which would be the flow of the method proposed.

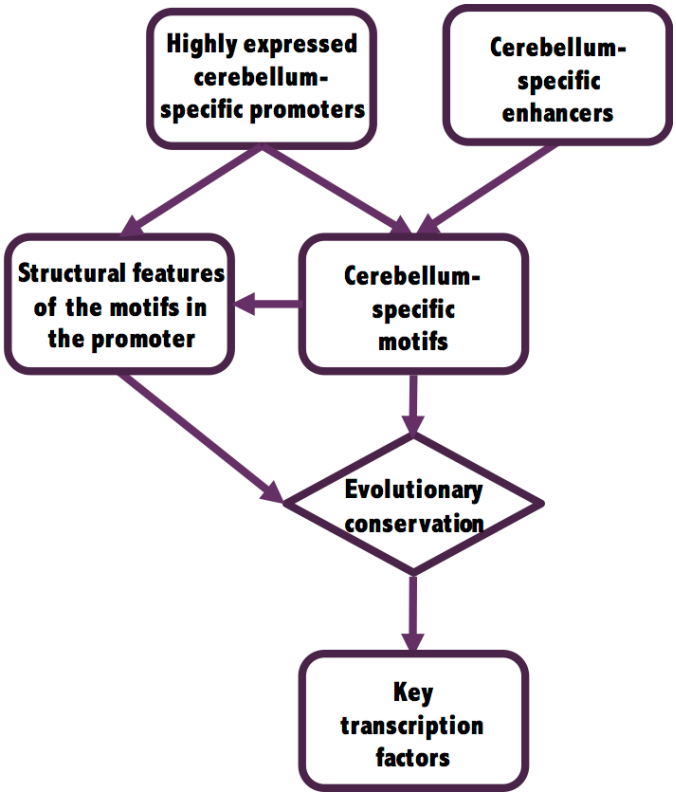

Figure 13: Flowchart of the method proposed for studying tissue-specific transcriptional regulation

3.2.2. Obtaining the data

The data set of cerebellum-specific promoters was obtained from the Expression Atlas database [42] by searching for genes with expression levels over 50 FPKM in the cerebellum and under-expressed (near 0 FPKM) in other parts of the brain. A data set of **226 genes** was obtained, and their promoters extracted with a length of 1500 bp (-1000 bp to 500 bp around the TSS). Additionally, a control data set of 228 genes was obtained by searching for genes with expression levels over 50 FPKM in other human tissues and under-expressed in the cerebellum.

The data set of enhancers was obtained from VISTA Enhancer Browser [43]. It is important to clarify that the data is not technically "cerebellum-specific" but more specifically "hindbrain-specific", given that the VISTA Enhancer Browser (and any other source) does not reach that level of detail in its specificity and only provides data of a higher-level organ (the hindbrain contains the cerebellum, apart from the medulla and the pons). A list of **38 enhancers** which are expressed in hindbrain but not in other tissues was obtained.

3.2.3. Motif finding

For the motif finding, the tool HOMER [44] was used, given its accuracy and, especially, its set of functionalities, such as peak annotation and the ability to inform the distance to the nearest TSS, which considerably simplify our work.

The process followed to find the list of candidates for being binding sites of the key TF in the cerebellum is the following:

1. Perform motif finding on the promoter data set
2. Perform motif finding on the enhancer data set
3. Perform motif finding on the control data set
4. Using the tool TFdiff [45], select only those motifs from the promoter and enhancer data sets which are not overrepresented in the control data set as well
5. Perform peak annotation on the enhancer data set
6. Reduce the list of motifs found in the promoters to those which are present near active peaks in the enhancers
7. Perform peak annotation on the promoter data set
8. Reduce the list of motifs found in the enhancers to those which are present near active peaks in the promoters
9. Unify both lists of motifs into one, removing redundancy

3.2.4. Analysis of structural features and evolutionary conservation in promoters

3.2.4.1. Overview

Our approach to determine if a candidate motif is part of the key TFBS mix concepts of two different studies. Vandenbon *et al.* [46] established three structural and positional rules for modelling the promoter structure that were later extended and further investigated in other studies [39][40]. The original set of rules were "presence", "absolute positioning" and "pairwise positioning".

In another research, Ballester *et al.* [41] studied the evolutionary conservation of liver-specific motifs among five species (*Homo sapiens*, *Macaca mulatta*, *Mus musculus*, *Rattus norvegicus* and *Canis familiaris*) and concluded that binding sites related to liver-specific biological pathways are shared by at least three of the species considered, and that this scenario is likely to occur in other tissues.

In our approach, we borrow some concepts from both studies and consider the following factors to determine whether a TF is key in cerebellum-specific transcriptional regulation:

- **Presence in enhancers**: the given motif has binding sites in hindbrain-specific enhancers
- **Presence near active peaks**: the given motif has binding sites in the near proximity of active peaks of either an enhancer or a promoter
- **Pairwise co-occurrence**: the given motif is likely to have binding sites in promoters that contain binding sites of other key TFs
- **Evolutionary conservation**: the given motif is evolutionary conserved in at least two other species apart from human among macaque, mouse, rat and dog (or an equivalent set of species)

3.2.4.2. Presence in enhancers

The presence in enhancers is already considered in the motif finding phase. Therefore, there is no need for further investigation, since all of the motifs under study at this point are present in hindbrain-specific enhancers.

3.2.4.3. Presence near active peaks

Similarly to the presence in enhancers, the presence near active peaks is already taken into account in the motif finding step.

This step is again performed by the tool HOMER [44], which contains a program called annotatePeaks, able to associate peaks with nearby genes and find motif occurrences in peaks, among other functions.

The criteria used in this study is considering the binding sites that start no further than 100 bp away from a peak in either a cerebellum-specific promoter or a hindbrain-specific enhancer.

3.2.4.4. Pairwise co-occurrence

To study the pairwise co-occurrence of binding sites, the functionality provided by HOMER [44] to obtain motif statistics around the TSS is used. This functionality allows for the generation of a file showing a table with the pairwise co-occurrence enrichment expressed by the corresponding log-P value of each pair of motifs.

We set the threshold for the logP value at **-2**, to then remove all of the logP values from the table which are over that threshold. Then, the motifs which have no more scores present at the table are discarded from the candidate set, and only those which have a significance pairwise co-occurrence (logP value under -2 with any other candidate) are kept and considered for the evolutionary conservation.

At this point, the promoter data set is as well narrowed down to only those promoters which contain sites of at least two of the resulting candidate motifs (only promoters with motif co-occurrence).

3.2.4.5. Evolutionary conservation

This is the last step of the workflow. The approach is similar to the one proposed by Ballester *et al.* [41], but simplifying it to investigate only which of the candidate motifs are deeply shared among species.

The EPO alignments of the Ensembl Compara tool [53] are the source of the evolutionary conservation study. Apart from human, another 17 species are considered, consisting of 7 primates and 10 other mammals:

- Primates: Gorilla, Chimpanzee, Orangutan, Macaque, Olive baboon, Vervet-AGM, Marmoset
- Other mammals: Mouse, Mouse SPRET/EiJ, Rat, Rabbit, Horse, Cat, Dog, Pig, Cow, Sheep

A given site is considered to be conserved if it overlaps at least 50% of the corresponding human binding site.

A given site is considered to be deeply shared among species if it is shared by at least 8 species in total (including human) from which at least 4 are primates and 1 is another mammal.

A given promoter is considered to be deeply shared if it contains at least two deeply shared binding sites from two different motifs.

3.3. Results

After performing the motif finding step as previously explained, the list of candidate motifs to be binding sites for the key TF in the cerebellum is the shown in Fig. 14 (known motifs) and Fig. 15 (putative *de novo* motifs).

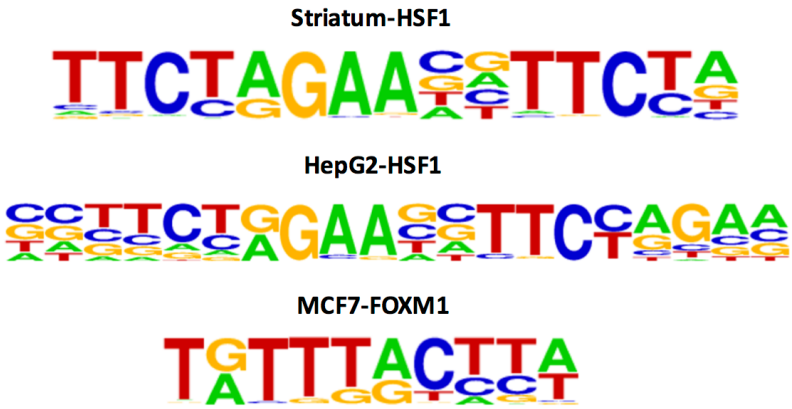

Figure 14: Cerebellum-specific known motifs

Figure 15: Cerebellum-specific putative *de novo* motifs

Of these 14 motifs, 3 are known and 11 *de novo*, and 7 of them have binding sites near active peaks in enhancers, and the other 7 near active peaks in promoters.

The table of pair-wise co-occurrence of the 14 candidate motifs is shown in **Table 4**.

Table 4: Pairwise motif co-occurrence in cerebellum-specific promoters (the scores are the corresponding logP values for each pair of motifs)

MOTIF NAME (# SITES)	STRIATUM-HSF1 (25)	HEPG2-HSF1 (24)	MCF7-FOXM1 (58)	16-TCTCTTTCTA (18)	17-CCCCCAGGGG (21)	20-GATGGAATAG (13)	3-TTCTATCC (86)	4-TTCAGCGA (68)	2-GTGGCCTAAA (2)	5+4+5-GARGGAAGGAAGG (37)	1-GGGTCCTG (43)	2-ATTGTCAC (4)	3-CACACTTA (116)	5-GAAGGGCG (156)
STRIATUM-HSF1 (25)	-4.62	-5.94	-3.04	-2.71	0.62	0.62	-1.88	0.84	1.00	1.04	-1.36	-0.46	2.23	-0.90
HEPG2-HSF1 (24)	-5.94	-2.25	-2.01	0.42	1.19	0.58	-1.25	-0.97	1.00	-0.87	-3.58	-0.46	1.53	-0.64
MCF7-FOXM1 (58)	-3.04	-2.01	-1.90	-2.04	0.75	-0.72	-1.85	-1.09	1.00	0.60	1.05	-1.19	-2.83	0.79
16-TCTCTTTCTA (18)	-2.71	0.42	-2.04	-3.73	-1.30	0.37	-0.77	-0.99	1.00	0.97	-1.16	-1.33	-0.88	0.65
17-CCCCCAGGG (21)	0.62	1.19	0.75	-1.30	-1.99	-0.98	1.44	1.69	1.00	-1.15	-1.93	-0.46	1.32	0.70
20-GATGGAATAG (13)	0.62	0.58	-0.72	0.37	-0.98	-0.46	1.53	1.59	-0.46	-0.90	1.38	1.00	0.88	1.02
3-TTCTATCC (86)	-1.88	-1.25	-1.85	-0.77	1.44	1.53	0.85	0.75	1.00	-1.04	1.20	-0.71	-0.81	-0.74
4-TTCAGCGA (68)	0.84	-0.97	-1.09	-0.99	1.69	1.59	0.75	0.69	1.00	0.56	1.04	-1.95	2.04	0.72
2-GTGGCCTAAA (2)	1.00	1.00	1.00	1.00	1.00	-0.46	1.00	1.00	1.00	-0.46	1.00	1.00	0.36	-0.82
5+4+5-GARGGAAGGAAGG (37)	1.04	-0.87	0.60	0.97	-1.15	-0.90	-1.04	0.56	-0.46	-1.07	-0.71	-0.46	-1.43	-1.56
1-GGGTCCTG (43)	-1.36	-3.58	1.05	-1.16	-1.93	1.38	1.20	1.04	1.00	-0.71	-1.00	-0.46	-1.04	-1.15
2-ATTGTCAC (4)	-0.46	-0.46	-1.19	-1.33	-0.46	1.00	-0.71	-1.95	1.00	-0.46	-0.46	1.00	1.06	-0.55
3-CACACTTA (116)	2.23	1.53	-2.83	-0.88	1.32	0.88	-0.81	2.04	0.36	-1.43	-1.04	1.06	-2.80	-1.08
5-GAAGGGCG (156)	-0.90	-0.64	0.79	0.65	0.70	1.02	-0.74	0.72	-0.82	-1.56	-1.15	-0.55	-1.08	-3.46

After removing the motifs that have no pairwise co-occurrence with a logP value under -2, we obtained a new set of 6 candidate motifs, from which 3 are known and 3 *de novo*, and 3 are active in promoters and the other 3 in enhancers. The list of candidate motifs after this step is shown in **Fig. 16**.

At this step, we reduced the number of promoters from the initial 209 to 75. It is worth mentioning that the promoters which only contained Striatum-HSF1 and HepG2-HSF1 with their sites overlapped were also removed, in order to consider only pairwise occurrence of motifs which have no similarities in their nucleotide composition.

Figure 16: Candidate motifs after filtering to include only those with significant co-occurrence

After the evolutionary conservation step, the same **6 motifs** are still part of the set, and **29 enhancers** out of the initial 38 contain deeply shared binding sites. As for the promoters, after this final step, the list has been reduced to **43 promoters** which contain deeply shared binding sites with pairwise co-occurrence. Fig. 17 and Fig. 18 show statistics about the site conservation in promoters and enhancers respectively.

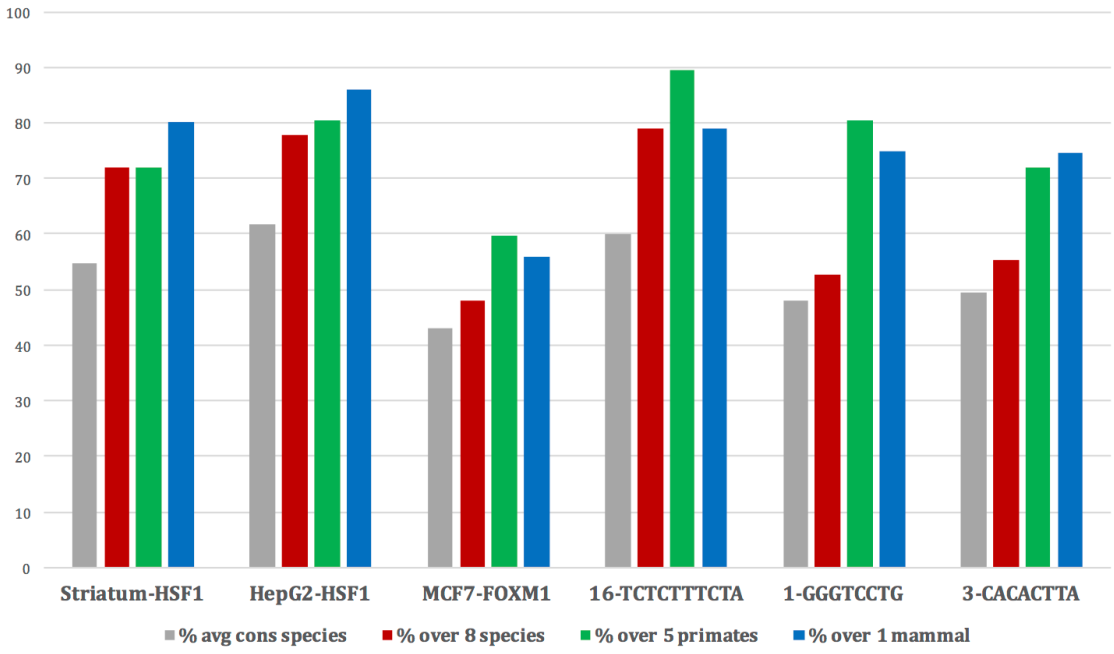

Figure 17: Conservation statistics for sites in promoters

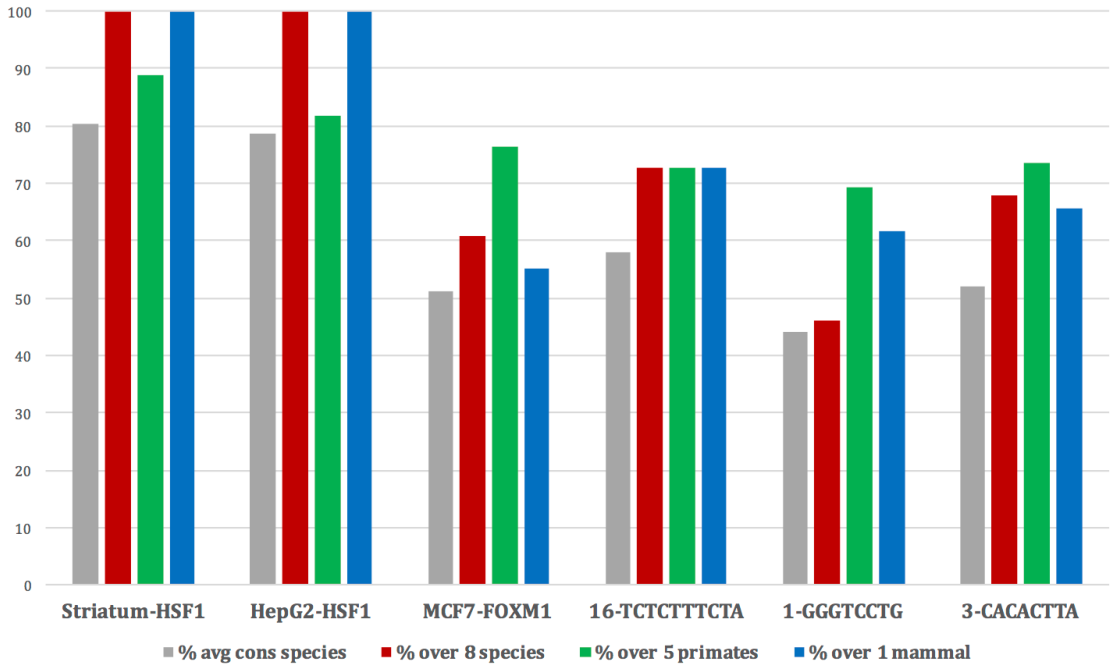

Figure 18: Conservation statistics for sites in enhancers

3.4. Discussion

We have designed a simple and straightforward *in-silico* method to determine the transcription factors, promoters and enhancers involved in tissue-specific transcriptional regulation. The results obtained seem to be quite robust, given that they are characterized by co-occurring and deeply shared TFBS and their corresponding promoters and enhancers. However, experimental validation will be needed in order to confirm the results and the accuracy of the method, which at the moment can simply be used as a tool to narrow down the objects of analysis in tissue-specific transcriptional regulation.

Probably the most potentially important finding of the study would be the determination of which enhancers are related to transcriptional regulation in the cerebellum, since it is the first time that enhancers are considered in such kind of study.

Due to that, it will be interesting also to check how much the inclusion of the enhancer data set to our research conditioned the results and confirming if it is actually fundamental to consider them in similar studies.

Apart from the enhancers, we could as well obtain a set of promoters which include co-occurring TFBS and have a considerable evolutionary conservation. These promoters could be highly involved in cerebellar transcriptional regulation.

In addition to this, we obtained 3 putative *de novo* motifs which are not significantly similar to any known motifs and which are candidates to be key in the cerebellar transcriptional regulation. If they can be confirmed as actual motifs after the experimental validation, this could become an interesting finding.

Finally, looking at the list of 3 known motifs that resulted from the analysis, the following observations can be inferred:

The transcription factor **HSF1** (Heat Shock Factor 1) is very often highly conserved in eukaryotes. This fact is consistent with our results, which include two different types of HSF1. The main function of this protein is usually to respond to cellular stress. Given that there is no evidence of cellular stress in the cerebellum, it would be interested to study this hypothetical situation.

Some other studies relate the transcription factor HSF1 to growth and aging **[47][48]**, which would be consistent with other studies that suggest a relation between the function of the cerebellum and aging **[49]**. It is therefore worth investigating which is its specific function within the cerebellum and its relationship with disease.

The transcription factor **FOXM1** plays a crucial role in cell cycle progression, which relates it to several types of cancers, including brain cancer **[50][51]**, and also to aging. Again, it makes sense to find it present in the cerebellum, and in case of being confirmed as one of the key TFs involved in transcriptional regulation, its function in the cerebellum must be further investigated.

Bringing together the analysis of the known TFs obtained as candidates for being key regulators of the cerebellar transcription, it is noteworthy that all of them have been linked to aging in several studies. Traditionally it has been believed that the function of the cerebellum is motor control, due to the observations of how damage to the cerebellum affects the response of the body, but there are no conclusive studies about its actual function [37]. In fact, some research clearly links motor performance as well as cognitive functions with aging [51]. And several other studies have suggested a cognitive function of the cerebellum related to aging [52], but so far there is no evidence proving that hypothesis. In addition, there are some studies that concluded that the cerebellar expression and methylation patterns across different ages are somehow odd and considerably different from other brain tissues [54]. All these suggest the possibility that the function of the cerebellum could be actually cognitive and aging-related, and motor control would be affected by these, as they are closely related.

Though those ideas about a possible function of the cerebellum related to aging are at this point mere speculation, our results and those of other studies suggest that the possibility would be worth further studying.

As for our future plan after experimental validation, one interesting path would be to take this study to a more granular level and produce an analysis of the transcriptional regulation in each cerebellar cell type. Also, performing the same kind of study in other tissues is within our future goals.

Conclusions

4. Conclusions and closing remarks

In the two chapters presented in this thesis we have delved into the process of transcriptional regulation in two different studies that try to offer new insights in this complex field.

Even though, not only in transcriptional regulation as a whole, but also in these two minute studies in comparison, there is still a long way to follow, the findings, conclusions and prospective results here shown offer some answers that can contribute for future advances in regulatory analysis.

The task of motif discovery has been one of the hardest challenges in molecular biology for decades and the approaches proposed in both chapters of this study will not change that fact. However, they offer an interesting addition that had never been considered before and that might help researchers to produce more accurate tools for motif finding: the use of topic models. Additionally, the Statistical GA here presented shows a very promising performance that, after a few adjustments (such as reducing the number of false positives and creating a strategy for the fine tuning of the parameters depending on the data set) could make it one of the most useful existing tools for motif discovery.

The workflow proposed for the study on the tissue-specific transcriptional regulation, on another note, needs experimental validation before we can throw definitive conclusions. Nonetheless, the results are statistically robust and eminently promising and, in case of being confirmed, they could help in the future to define how transcriptional regulation works in the cerebellum. The study could determine a few putative *de novo* motifs which could also be cerebellum specific, as well as point out which are the promoters and enhancers more prone to be related with cerebellar transcriptional regulation. And finally, it could serve as a basis to simplify research in case the validity of the method is confirmed by experimental validation. The *in-silico* workflow proposed would be very useful to be applied in the future to other tissues in a similar kind of research.

In conclusion, this study presents several findings, ideas and approaches that might be of great assistance in future research to try to fathom the complex process of transcriptional regulation in the cell.

References

References

1. Phillips, Theresa. "Regulation of transcription and gene expression in eukaryotes." Nature Education 1.1 (2008): 199.
2. Tompa, Martin, et al. "Assessing computational tools for the discovery of transcription factor binding sites." Nature biotechnology 23.1 (2005): 137-144.
3. Das, Modan K., and Ho-Kwok Dai. "A survey of DNA motif finding algorithms." BMC bioinformatics 8.7 (2007): S21.
4. Galas, David J., and Albert Schmitz. "DNAase footprinting a simple method for the detection of protein-DNA binding specificity." Nucleic acids research 5.9 (1978): 3157-3170.
5. Garner, Mark M., and Arnold Revzin. "A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system." *Nucleic acids research* 9.13 (1981): 3047-3060.
6. Lawrence, Charles E., et al. "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment." *SCIENCE-NEW YORK THEN WASHINGTON*- 262 (1993): 208-208.
7. Blei, David M. "Probabilistic topic models." *Communications of the ACM* 55.4 (2012): 77-84.
8. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.
9. Li, Wei, and Andrew McCallum. "Pachinko allocation: DAG-structured mixture models of topic correlations." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.

10. Lafferty, John D., and David M. Blei. "Correlated topic models." *Advances in neural information processing systems.* 2006.

11. Aitchison, John. "The statistical analysis of compositional data." Journal of the Royal Statistical Society, Series B, 44(2) (1982) (1986):139–177.

12. Gutierrez, Josep Basha, and Kenta Nakai. "A study on the application of topic models to motif finding algorithms." *BMC Bioinformatics* 17.19 (2016): 129.

13. Gutierrez, Josep Basha, Martin Frith, and Kenta Nakai. "A genetic algorithm for motif finding based on statistical significance." *International Conference on Bioinformatics and Biomedical Engineering*. Springer International Publishing, 2015.

14. Hornik, Kurt, and Bettina Grün. "topicmodels: An R package for fitting topic models." *Journal of Statistical Software* 40.13 (2011): 1-30.

15. Abnizova, Irina, et al. "Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the Drosophila genome: the fluffy-tail test." *BMC bioinformatics* 6.1 (2005): 109.

16. Shu, Jian-Jun, and L. I. Yajing. "A statistical thin-tail test of predicting regulatory regions in the Drosophila genome." *Theoretical Biology and Medical Modelling* 10.1 (2013): 11.

17. Mann, Henry B., and Donald R. Whitney. "On a test of whether one of two random variables is stochastically larger than the other." *The annals of mathematical statistics* (1947): 50-60.

18. Hughes, Jason D., et al. "Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae." *Journal of molecular biology* 296.5 (2000): 1205-1214.

19. Workman, C. T., and G. D. Stormo. "ANN-Spec: a method for discovering transcription factor binding sites with improved specificity." *Pac Symp Biocomput*. Vol. 5. 2000.

20. Hertz, Gerald Z., and Gary D. Stormo. "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences." *Bioinformatics* 15.7 (1999): 563-577.

21. Frith, Martin C., et al. "Finding functional sequence elements by multiple local alignment." *Nucleic acids research* 32.1 (2004): 189-200.

22. Ao, Wanyuan, et al. "Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR." *Science* 305.5691 (2004): 1743-1746.

23. Bailey, Timothy L., and Charles Elkan. "The value of prior knowledge in discovering motifs with MEME." *Ismb*. Vol. 3. 1995.

24. Eskin, Eleazar, and Pavel A. Pevzner. "Finding composite regulatory patterns in DNA sequences." *Bioinformatics* 18.suppl 1 (2002): S354-S363.

25. Thijs, Gert, et al. "A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling." *Bioinformatics* 17.12 (2001): 1113-1122.

26. Helden, Jacques van, Bruno André, and Julio Collado-Vides. "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies." *Journal of molecular biology* 281.5 (1998): 827-842.

27. Helden, Jacques van, Alma Rios, and Julio Collado-Vides. "Discovering regulatory elements in non-coding sequences by analysis of spaced dyads." *Nucleic acids research* 28.8 (2000): 1808-1818.

28. Régnier, Mireille, and Alain Denise. "Rare events and conditional events on random strings." *Discrete Mathematics & Theoretical Computer Science* 6.2 (2004): 191-214.

29. Favorov, A. V., et al. "Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length and its validation on the ArcA binding sites." *Proc. of BGRS* 2004 (2004): 269-272.

30. Pavesi, Giulio, et al. "Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes." *Nucleic acids research* 32.suppl 2 (2004): W199-W203.

31. Sinha, Saurabh, and Martin Tompa. "YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation." *Nucleic acids research* 31.13 (2003): 3586-3588.

32. Pevzner, Pavel A., and Sing-Hoi Sze. "Combinatorial approaches to finding subtle signals in DNA sequences." *ISMB*. Vol. 8. 2000.

33. Burset, Moises, and Roderic Guigo. "Evaluation of gene structure prediction programs." *genomics* 34.3 (1996): 353-367.

34. Wingender, Edgar, et al. "TRANSFAC: a database on transcription factors and their DNA binding sites." *Nucleic acids research* 24.1 (1996): 238-241.

35. Lee, Tong Ihn, and Richard A. Young. "Transcriptional regulation and its misregulation in disease." *Cell* 152.6 (2013): 1237-1251.

36. Maurano, Matthew T., et al. "Systematic localization of common disease-associated variation in regulatory DNA." *Science* 337.6099 (2012): 1190-1195.

37. Reeber, Stacey L., Tom S. Otis, and Roy V. Sillitoe. "New roles for the cerebellum in health and disease." *Frontiers in systems neuroscience* 7 (2013).

38. Rincon, Melvin Y., et al. "Genome-wide computational analysis reveals cardiomyocyte-specific transcriptional cis-regulatory motifs that enable efficient cardiac gene therapy." *Molecular Therapy* 23.1 (2015): 43-52.

39. Vandenbon, Alexis, and Kenta Nakai. "Modeling tissue-specific structural patterns in human and mouse promoters." *Nucleic acids research* 38.1 (2010): 17-25.

40. López, Yosvany, Alexis Vandenbon, and Kenta Nakai. "A set of structural features defines the cis-regulatory modules of antenna-expressed genes in Drosophila melanogaster." *PloS one* 9.8 (2014): e104342.

41. Ballester, Benoit, et al. "Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways." *Elife* 3 (2014): e02626.

42. Petryszak, Robert, et al. "Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants." *Nucleic acids research* 44.D1 (2016): D746-D752.

43. Visel, Axel, et al. "VISTA Enhancer Browser—a database of tissue-specific human enhancers." *Nucleic acids research* 35.suppl 1 (2007): D88-D92.

44. Heinz, Sven, et al. "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities." *Molecular cell* 38.4 (2010): 576-589.

45. Hooghe, Bart. *In silico approaches to studying transcriptional gene regulation: prediction of transcription factor binding sites and applications thereof*. Diss. Ghent University, 2011.

46. Vandenbon, Alexis, and Kenta Nakai. "Using simple rules on presence and positioning of motifs for promoter structure modeling and tissue-specific expression prediction." *Genome Informatics* 21 (2008): 188-199.

47. Xiao, XianZhong, et al. "HSF1 is required for extra-embryonic development, postnatal growth and protection during inflammatory responses in mice." *The EMBO journal* 18.21 (1999): 5943-5952.

48. Anckar, Julius, and Lea Sistonen. "Regulation of HSF1 function in the heat stress response: implications in aging and disease." *Annual review of biochemistry* 80 (2011): 1089-1115.

49. Seidler, Rachael D., et al. "Motor control and aging: links to age-related brain structural, functional, and biochemical effects." *Neuroscience & Biobehavioral Reviews* 34.5 (2010): 721-733.

50. Laoukili, Jamila, Marie Stahl, and René H. Medema. "FoxM1: at the crossroads of ageing and cancer." *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1775.1 (2007): 92-102.

51. Koo, Chuay-Yeng, Kyle W. Muir, and Eric W-F. Lam. "FOXM1: From cancer initiation to progression and treatment." *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1819.1 (2012): 28-37.

52. Rapoport, Mark, Robert van Reekum, and Helen Mayberg. "The role of the cerebellum in cognition and behavior: a selective review." *The Journal of neuropsychiatry and clinical neurosciences* 12.2 (2000): 193-198.

53. Aken, Bronwen L., et al. "Ensembl 2017." *Nucleic acids research* 45.D1 (2016): D635-D642.

54. Hernandez, Dena G., et al. "Distinct DNA methylation changes highly correlated with chronological age in the human brain." *Human molecular genetics* 20.6 (2011): 1164-1172.