PhD Thesis

Computational Representation and Discovery of Transcription Factor Binding Sites

Dissertation submitted for the degree of
Doctor of Philosophy in Biomedical Engineering

PhD Student: Joan Maynou Fernàndez
PhD Advisor: Dr. Alexandre Perera i Lluna

Universitat Politècnica de Catalunya
Programa de Doctorat en Enginyeria Biomèdica

Barcelona, 2015

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola de Doctorat

Acta de qualificació de tesi doctoral

Curs acadèmic:

Nom i cognoms

Programa de doctorat

Unitat estructural responsable del programa

Resolució del Tribunal

Reunit el Tribunal designat a l'efecte, el doctorand / la doctoranda exposa el tema de la seva tesi doctoral titulada

_____

_____.

Acabada la lectura i després de donar resposta a les qüestions formulades pels membres titulars del tribunal, aquest atorga la qualificació:

☐ NO APTE ☐ APROVAT ☐ NOTABLE ☐ EXCEL·LENT

(Nom, cognoms i signatura) President/a	(Nom, cognoms i signatura) Secretari/ària

(Nom, cognoms i signatura) Vocal	(Nom, cognoms i signatura) Vocal	(Nom, cognoms i signatura) Vocal

_____, _____ d'/de _____ de _____

El resultat de l'escrutini dels vots emesos pels membres titulars del tribunal, efectuat per l'Escola de Doctorat, a instància de la Comissió de Doctorat de la UPC, atorga la MENCIÓ CUM LAUDE:

☐ SÍ ☐ NO

(Nom, cognoms i signatura) President de la Comissió Permanent de l'Escola de Doctorat	(Nom, cognoms i signatura) Secretari de la Comissió Permanent de l'Escola de Doctorat

Barcelona, _____ d'/de _____ de _____

A la meva mare,

L'Univers no és cap idea meva.
La meva idea de l'Univers sí que és una idea meva.
La nit no cau pels meus ulls,
La meva idea de la nit és el que cau pels meus ulls.
Fora del fet que jo pensi i que quasevol tingui pensaments,
la nit cau concretament
i el fulgor de les estrelles existeix com si tingués pes.

Alberto Caeiro

Contents

List of Tables

List of Figures

Agraïments

La realització d'una tesi doctoral és un camí llarg i ple d'obstacles, però amb bons moments. Aquest camí hauria sigut difícilment assolible si no fos per l'ajuda de persones i institucions, tant en l'àmbit acadèmic, econòmic i personal.

El primer agraïment és pel meu director de tesi, Alexandre Perera. Per confiar en mi des del principi i ensenyar-me que el veritable art d'investigar és qüestionar-s'ho tot. El meu reconeixement més sincer per la seva tasca.

En Pere Caminal, Montserrat Vallverdú, Joan-Josep Gallardo-Chacón, Erola Pairó i Santiago Marcos, per les aportacions realitzades en els articles presents en aquesta tesi. Un especial agraïment a Erola Pairó pel gran esforç realitzat en el desenvolupament de MEET i per les hores de confidències compartides.

A cadascun els professors del departament pels seus comentaris i consells. I a Susana Vinga, pel seu suport durant la meva estada a Lisboa i pel seu interès en el meu treball.

A tots els companys de doctorat que van formar part del laboratori d'Enginyeria Biomèdica però en especial a: Ainara, Andrés, Andrey, Bere, Christian, Francesc, José Valencia, Lna, Leo, Puy, Raimon i Rudys....per tots els bons moments que vàrem compartir junts dins i fora de la universitat. Però sobretot, per fer que cada moment de la tesi fos genial. Sou els millors!

Agrair a totes les institucions que m'han ajudat econòmicament i que han permès que pogués desenvolupar la meva recerca. En especial a la Universitat Politècnica de Catalunya, Institut de Biotecnologia de Catalunya i el CIBER-bbn (Bioingenería, Biomateriales y Nanomedicina).

A la meva família....als meus pares, en Martí, a l'Eva, en Francesc, a la Sònia, en Josep i a la Mònica... per ensenyar-me els valors del treball, la constància, el

sacrifici i per escoltar-me, comprendre'm i animar-me quan les coses no sortien. I agrair especialment a l'Eva pel disseny de la portada de la tesi. I pels nous vinguts Jan, Laia, Guillem, Pau, Gerard i Txell per fer que mai falti un somriure a casa. A l'Ana i en Luís per cuidar-me com un fill més. A tots els meus amics de sempre, però en especial en Pol, l'Aran i la Laura... pel seu suport incondicional.

A la meva dona, Patrícia, per estar al meu costat en tot moment, en els bons i mals moments i ajudar-me a continuar quan jo era el primer que ja no creia en les meves possibilitats. Per tot això i més, moltes gràcies.

Finalment, m'agradaria acabar recordant una persona molt especial, un clar exemple de lluita i que, desgraciadament, ja no és entre nosaltres. Mare, et brindo la tesi.

Abstract

The information about how, when, and where are produced the proteins has been one of the major challenge in molecular biology. The studies about the control of the gene expression are essential in order to have a better knowledge about the protein synthesis.

The gene regulation is a highly controlled process that starts with the DNA transcription. This process operates at the gene level, hereditary basic units, which will be copied into primary ribonucleic acid (RNA). This first step is controlled by the binding of specific proteins, called as Transcription Factors (TF), with a sequence of the DNA (Deoxyribonucleic Acid) in the regulatory region of the gene. These DNA sequences are known as binding sites (BS).

The binding sites motifs are usually very short (5 to 20 bp long) and highly degenerate. These sequences are expected to occur at random every few hundred base pairs. Besides, a TF can bind among different sites. Due to its highly variability, it is difficult to establish a consensus sequence. The study and identification binding sites is important to clarify the control of the gene expression.

Due to the importance of identifying binding sites sequences, projects such as ENCODE (Encyclopedia of DNA Elements), have dedicated efforts to map binding sites for large set of transcription factor to identify regulatory regions. The genome sequence availability and in gene expression analysis technologies have also allowed the development of computational methods for motif detection. Thanks for these advances, in the last years a large number of algorithms have been applied to different motif models. Most of these algorithms are developed to resolve sequences of prokaryotic organisms or simple eukaryotic organisms like yeast. In general, the false positive rate (FPR) is high true positive rate (TFR). To study higher organisms, with a higher complexity in their genomes, it is necessary of more sensitive methods.

In this thesis, we have approached the problem of the binding site detection from another angle. We have developed a set of toolkit for motif binding detection based on linear and non-linear models. First of all, we have been able to characterize binding sites using different approaches. The first one is based on

the information that there is in each binding sites position. The second one is based on the covariance model of an aligned set of binding sites sequences.

From these motif characterizations, we have proposed a new set of computational methods to detect binding sites. First, it was developed a new method based on parametric uncertainty measurement (Rényi entropy). This detection algorithm evaluates the variation on the total Rényi entropy of a set of sequences when a candidate sequence is assumed to be a true binding site belonging to the set. This method was found to perform especially well on transcription factors that the correlation among binding sites was null.

The correlation among binding sites positions was considered through linear, Q-residuals, and non-linear models, α-Divergence and SIGMA. Q-residuals is a novel motif finding method which constructs a subspace based on the covariance of numerical DNA sequences. When the number of available sequences was small, The Q-residuals performance was significantly better and faster than all the others methodologies.

α-Divergence was based on the variation of the total parametric divergence in a set of aligned sequenced with binding evidence when a candidate sequence is added. Given an optimal q-value, the α-Divergence performance had a better behavior than the others methodologies in most of the studied transcription factor binding sites. And finally, a new computational tool, SIGMA, was developed as a trade-off between the good generalisation properties of pure entropy methods and the ability of position-dependency metrics to improve detection power. In approximately 70% of the cases considered, SIGMA exhibited better performance properties, at comparable levels of computational resources, than the methods which it was compared.

This set of toolkits and the models for the detection of a set of transcription factor binding sites (TFBS) has been included in an R-package called *MEET*.

Resum

La informació sobre com, quan i on es produeixen les proteïnes ha estat un dels majors reptes en la biologia molecular. Els estudis sobre el control de l'expressió gènica són essencials per conèixer millor el procés de síntesis d'una proteïna.

La regulació gènica és un procés altament controlat que s'inicia amb la transcripció de l'ADN. En aquest procés, els gens, unitat bàsica d'herència, són copiats a àcid ribonucleic (RNA). El primer pas és controlat per la unió de proteïnes, anomenades factors de transcripció (TF), amb una seqüència d'ADN (àcid desoxiribonucleic) en la regió reguladora del gen. Aquestes seqüències s'anomenen punts d'unió i són específiques de cada proteïna. La unió dels factors de transcripció amb el seu corresponent punt d'unió és l'inici de la transcripció.

Els punts d'unió són seqüències molt curtes (5 a 20 parells de bases de llargada) i altament degenerades. Aquestes seqüències poden succeir de forma aleatòria cada centenar de parells de bases. A més a més, un factor de transcripció pot unir-se a diferents punts. A conseqüència de l'alta variabilitat, és difícil establir una seqüència consensus. Pert tant, l'estudi i la identificació del punts d'unió és important per entrendre el control de l'expressió gènica.

La importància d'identificar seqüències reguladores ha portat a projectes com l'ENCODE (Encyclopedia of DNA Elements) a dedicar grans esforços a mapejar les seqüències d'unió d'un gran conjunt de factors de transcripció per identificar regions reguladores. L'accés a seqüències genòmiques i els avanços en les tecnologies d'anàlisi de l'expressió gènica han permés també el desenvolupament dels mètodes computacionals per la recerca de motius. Gràcies aquests avenços, en els últims anys, un gran nombre de algorismes han sigut aplicats en la recerca de motius en organismes procariotes i eucariotes simples. Tot i la simplicitat dels organismes, l'índex de falsos positius és alt respecte als veritables positius. Per tant, per estudiar organimes més complexes és necessari mètodes amb més sensibilitat.

En aquesta tesi ens hem apropat al problema de la detecció dels punts d'unió des de diferents angles. Concretament, hem desenvolupat un conjunt d'eines per

la detecció de motius basats en models lineals i no-lineals. Les seqüències d'unió dels factors de transcripció han sigut caracteritzades mitjançant dues aproximacions. La primera està basada en la informació inherent continguda en cada posició de les seqüencies d'unió. En canvi, la segona aproximació caracteritza la seqüència d'unió mitjançant un model de covariància.

A partir d'ambdues caracteritzacions, hem proposat un nou conjunt de mètodes computacionals per la detecció de seqüències d'unió. Primer, es va desenvolupar un nou mètode basat en la mesura paramètrica de la incertesa (entropia de Rényi). Aquest algorisme de detecció avalua la variació total de l'entropia de Rényi d'un conjunt de seqüències d'unió quan una seqüència candidata és afegida al conjunt. Aquest mètode va obtenir un bon rendiment per aquells seqüències d'unió amb poca o nul·la correlació entre posicions.

La correlació entre posicions fou considerada a través d'un model lineal, Q-residuals, i dos models no-lineals, α-Divergence i SIGMA. Q-residuals és una nova metodologia per la recerca de motius basada en la construcció d'un subespai a partir de la covariància de les seqüències d'ADN numèriques. Quan el nombre de seqüències disponible és petit, el rendiment de Q-residuals fou significant millor i més ràpid que en les metodologies comparades.

α-Divergence avalua la variació total de la divergència paramètrica en un conjunt de seqüències d'unió quan una seqüència candidata és afegida. Donat un q-valor òptim, α-Divergence va tenir un millor rendiment que les metodologies comparades en la majoria de seqüències d'unió dels factors de transcripció considerats. Finalment, un nou mètode computacional, SIGMA, va ser desenvolupat per tal millorar la potència de detecció considerant les bones propietats de generalització dels mètodes d'entropia purs i les mètriques que consideren la dependència entre posicions. En un 70% dels casos considerats, SIGMA va obtenir millor rendiment que els mètodes amb els quals va ser comparat.

Aquest conjunt d'eines més els models per la detecció d'un conjunt de seqüències d'unió dels factors de transcripció (TFBS) han sigut inclosos en un paquet en R anomenat *MEET*.

Resumen

La información sobre cómo, cuándo y dónde se producen las proteínas ha sido uno de los mayores retos en la biología molecular. Los estudios sobre el control de la expresión génica son esenciales para conocer mejor el proceso de síntesis de una proteína.

La regulación génica es un proceso altamente controlado que se inicia con la transcripción del ADN. En este proceso, los genes, unidad básica de herencia, son copiados a ácido ribonucleico (RNA). El primer paso es controlado por la unión de unas proteínas, conocidas como factores de transcripción (TF), con una secuencia del ácido desoxirribonucleico (ADN) en la región regulatoria del gen. Estas secuencias se conocen como secuencias de unión y son específicas de cada proteína. Estas proteínas son los factores de transcripción (TF). La unión de los factores de transcripción con su correspondiente secuencia de unión es el inicio de dicho proceso.

Los puntos de unión son secuencias muy cortas (5 a 20 pares de bases de longitud) y altamente degeneradas. Estas secuencias pueden suceder de forma aleatoria cada centenar de pares de bases. Además, un factor de transcripción puede unirse a diferentes puntos de unión. A consecuencia de la alta variabilidad, es difícil establecer una secuencia consensus. Por lo tanto, el estudio y la identificación de los puntos de unión es importante para entender el control de la expresión génica.

La importancia de identificar secuencias reguladoras ha provocado que proyectos como ENCODE (Encyclopedia of DNA Elements) a dedicar grandes esfuerzos a mapear las secuencias de unión de un gran conjunto de factores de transcripción para identificar regiones reguladoras. El acceso a secuencias genómicas y los adelantos en las tecnologías de análisis de la expresión génica ha permitido también el desarrollo de los métodos computacionales para la búsqueda de motivos. Dichos algoritmos han sido aplicados en la búsqueda de motivos en organismos procariotas y eucariotas simples. Aunque la complejidad de los organismos estudiados no es alta, el índice de falsos positivos alto con respecto a los verdaderos positivos. Por lo tanto, para estudiar organismos más complejos son necesarios

métodos con más sensibilidad.

En esta tesis, hemos planteado el problema de la detección de los puntos des de diferentes ángulos. Se ha desarrollado un conjunto de herramientas para la detección de motivos basados en modelos lineales y no-lineales. Las secuencias de unión de los factores de transcripción han sido caracterizadas mediante dos aproximaciones. La primera aproximación está basada en la información inherente contenida en cada posición de las secuencias de unión. En cambio, la segunda aproximación caracteriza el motivo mediante un modelo de covarianza.

A partir de ambas caracterizaciones, proponemos un nuevo conjunto de métodos computacionales para la detección de las secuencias de unión. Primero, se desarrolló una metodología basada en la medida paramétrica de la incertidumbre (entropía Réenyi). Este algoritmo de detección evalúa la variación total de la entropía de Réenyi de un conjunto de secuencias de unión cuando una secuencia candidata es añadida al conjunto. Este método obtuvo un buen rendimiento para esas secuencias de unión con poca o nula correlación entre posiciones.

La correlación entre posiciones fue considerada a través de un modelo lineal, Q-residuals, y dos modelos no-lineales, α-Divergence y SIGMA. Q-residuals es una nueva metodología para la búsqueda de motivos basada en la construcción de un subespacio a partir de la covarianza de las secuencias de ADN numéricas. Cuando el número de secuencias disponible es pequeño, el rendimiento de Q-residuals fue significativamente mejor y más rápido que en las metodologías comparadas.

α-Divergence evalúa la variación total de la divergencia paramétrica en un conjunto de secuencias de unión cuando una secuencia candidata es añadida. Dado un q-valor óptimo, α-Divergence tuvo un mejor rendimiento que los métodos comparados en la mayoría de puntos de unión considerados. Finalmente, un nuevo método computacional, SIGMA, se desarrolló para mejorar la potencia de detección considerando las buenas propiedades de generalización de los métodos de entropía puros y las métricas que consideran la dependencia entre posiciones. En un 70% de los casos considerados, SIGMA obtuvo mejores resultados que los métodos con los cuales fue comparado.

Dicho conjunto de herramientas más los modelos para la detección de un conjunto de secuencias de unión de los factores de transcripción (TFBS) han sido incluidos en un paquete en R llamado *MEET*.

Preface

This thesis is about *novo* motif detection, particularly applied to detection Transcription Factor Binding Sites (TFBS). The core of this thesis is the development of a new family of computational methods for the characterization and detection of protein binding sequences through linear and nonlinear measurements.
The work performed about the motif detection has been raised since different approximations. Linear model is based on covariance model, which considers the correlation between binding sites positions. The nonlinear model is based on information gain model which allow to consider the depedence or independence among binding sites positions according to the TFBS characteristics. Both models, linear and nonlinear model, have been shown to be useful in the characterization and detection protein binding sites.

Specific Aims

The research in this PhD has been oriented to the development of new computational methods for the characterization and the detection of Transcription Factor Binding Sites. The following objectives were established for this thesis:

1. To study the biological problem of the control of gene expression: transcription factors, DNA-protein binding sites, Regulatory regions in DNA, architecture of DNA-binding domains. To gain knowledge on the process of binding of the transcription factor with the gene promoter.

2. Characterization of protein binding sequences through linear and nonlinear models.

3. Estimation of the positions of the site involved in the binding process by means of linear and nonlinear models.

4. To detect binding sites of transcription factors considering a nonlinear measurement called Rényi entropy which assumes independency among binding sites positions.

5. To study dependency among positions by means of linear and nonlinear models.

6. To develop an optimization algorithm to calculate the correlated binding sites positions.

7. To detect TFBS considering the correlation between binding sites positions through linear and nonlinear model. This methodology has been applied genome sequence.

8. Public distribution of the R-package produced during this thesis to scientific community. This package implements most of the methods present above.

9. To integrate both approximation in a detector based on nonlinear measurement.

Contributions

The main contribution of this thesis is the development of a new family of methodology for motif finding based on linear and nonlinear measurements. This knowledge lets us to know better the process of binding of the transcription factor with the gene promoter. Moreover, these methodologies allow to detect sequences with a higher efficiency suitable to work with more complex organisms, such as the eukaryotics organisms.
A part from that, an R library has been developed on motif detection. *MEET* is an R-package that integrates a set of tools, internal and external algorithms, and TF-models library ($\sim$ 500 models) for transcription factor binding sites detection.

Part I

Framework

Chapter 1

Genomic Background

1.1 Introduction

There are different types of essential molecules for the life: small molecules, proteins and nucleic acids (DNA and RNA). Small molecules carry energy, transmit signals and are liked into macromolecules (e.g. simple sugars, amino acids, water,...). On the other hand, the proteins are the chief actors within the cell, said to be carrying out the duties specified by the information encoded in genes. Proteins play a main role in catalyzing chemical reactions (enzymes), cell signalling and signal transduction (e.g. transcription factors) and structural functions.

The information to produce each kind of protein is carried in the genetic material. Nucleic Acids carry genetic information or form structures within cells. The most common nucleic acids are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). DNA acts as the permanent repository of genetic information in most cells while RNA plays several important roles in the processes of transcribing genetic information from deoxyribonucleic acid (DNA) into proteins [1].

The ability of cells to maintain a high degree of order in a chaotic universe depends on the genetic information that is expressed, maintained, replicated, and occasionally improved by the basic genetic processes RNA and protein synthesis, DNA repair, DNA replication, and genetic recombination. In these processes, which produce and maintain the proteins and nucleic acids of a cell, the information in a linear sequence of nucleotides is used to specify either another linear chain of nucleotides (a DNA or an RNA molecule) or a linear chain of amino acids (a protein molecule) [1].

1.2 Genetic Information

Nucleic Acids contains the information for determining the amino acid sequence and hence the structure and function of all the cells proteins. Moreover, nucleic acids are part of the cellular structures that select and align amino acids in the correct order as a polypeptide chain is being synthesized, and they catalyze a number of fundamental chemical reactions in cells, including formation of peptide bonds between amino acids during protein synthesis [41].

Deoxyribonucleic acid (DNA) is a molecule that contains the genetic instructions used in the development and functioning of all known living organisms. A DNA molecule consists of two long chains composed by the combination of four nucleotides - adenine (A), thymine (T), cytosine (C) and guanine (G)- joined by phosphodiester bonds [1]. Moreover, according to biochemical properties, nitrogenous bases allow to arrange in three classes [27]:

1. Molecular structure: A and G are purines (R), while C and T are pyrimidines (Y).

2. Strength of links: bases A and T are linked by two hydrogen bonds (W-weak bond), while C and G liked by three hydrogen bonds (S-strong bond).

3. Radical content: A and C contain the amino (NH_3) group in the large groove (M class), while T and G contain the keto ($C = O$) group (K class)

The information stored in DNA is arranged in hereditary units, now known as genes, which control identifiable traits of an organism. Hence, a gene is a unit of DNA that contains the information to specify synthesis of a single polypeptide chain or functional RNA [41]. RNA, like DNA, it is composed of a linear sequence of nucleotides, but it has two chemical differences: the sugar-phosphate backbone of RNA contains ribose instead of a deoxyribose sugar and the base thymine (T) is replaced by uracil (U), a very closely related base that likewise pairs with A [1]. A gene is a portion of DNA that contains coding sequences known as exons and a regulatory region known as the promoter. Moreover, a eukaryotic gene contains also noncoding sequences known as introns. The set of genes in an organism is known as its genome, which may be stored in one or more chromosomes, as shown in the Figure 1.1[1]. The region of the chromosome at which a particular gene is located is called its locus. A chromosome consists of a single, very long DNA helix on which thousands of genes are encoded (e.g. the genome of *Homo sapiens* has approximately 3.000.000.0000 base pairs).

[1] http://www.genome.gov/. Last consulted 2012

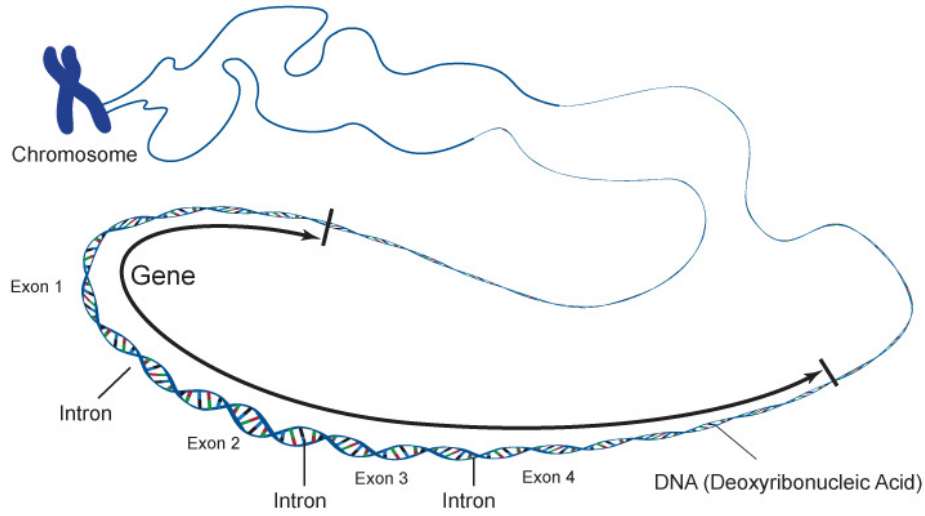


Figure 1.1: Schema of the information stored in DNA from chromosome to gene. Adapted from http://www.genome.gov/.

1.3 Regulation of Gene Expression

Each gene encodes the information to synthesize a particular protein or a particular functional ribonucleic acid (RNA) [70]. When a gene is active, the coding and noncoding, in eukaryotic cells, sequences are copied in a process called transcription, producing an RNA copy of the gene's information. The molecules resulting from gene expression, whether RNA or protein, are known as gene products, and are responsible for the development and functioning of all living things, as shown in the Figure 1.2.

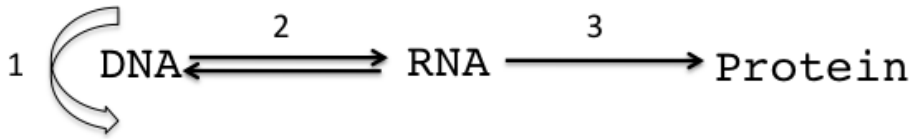


Figure 1.2: Schema of the information flow in cells. Step 1 is the DNA replication. Step 2 is the Transcription. And finally, step 3 is the Translation.

The information flow could be described as follows:

1. **DNA replication**. A DNA sequence is copied to yield a molecule nearly identical to the starting molecules.
2. **Transcription**. A portion of DNA sequence is converted to the corresponding RNA sequence.
3. **Translation**. The polypeptide sequence corresponding to the mRNA sequence is synthesized.

Transcription

The machinery for cell functions are proteins and its synthesis is regulated through gene expression [80]. Gene regulation is a highly controlled process. This process operates at the level of transcription by selecting those genes that will be copied into the primary RNA transcript, as shown in the Figure 1.3 [2]. The transcription is the central point of regulatory mechanisms and is controlled by the presence of short DNA sequences within gene regulatory regions. These regions are the binding sites for specific proteins that are known as Transcription Factors (TF) [63]. The function of these proteins is to bind themselves to specific sequences within the gene regulatory region. Then, these proteins interact with one another and with the RNA polymerase (RNAP) enzyme itself in order to regulate the rate of transcription. TFs recognize DNA sequences 6-8 bp long, suggesting that gene regulation is carried out for a large number of TF.

Translation

In the eukaryotic cells, transcription and translation stages are not directly connected, as the nuclear membrane physically separates the process, as shown in the Figure 1.4 [3]. The mRNA obtained must be modified to leave the nucleus using the processes of 7-metilguanosina, polyadenylation and splicing. After the mRNA has been processed, it is translated into an amino acids sequence, process known as translation. Here the ribosome, an enormously complex molecular machine composed of both RNA and protein, carries out the second process, called translation. During translation, the ribosome assembles and links together amino acids in the precise order dictated by the mRNA sequence according to the nearly universal genetic code [41]. These polypeptides or proteins form structural proteins and enzymes that control the metabolic processes in cells [80].

[2] Adapted from http://www.genome.gov/. Last consulted 2012
[3] Adapted from http://www.genome.gov/. Last consulted 2012

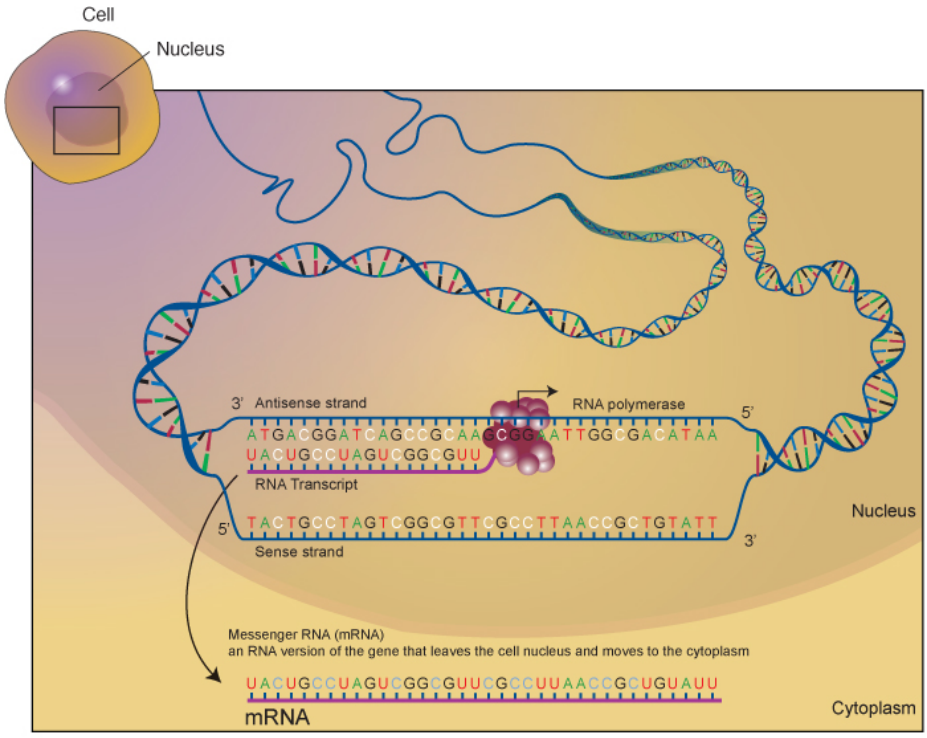


Figure 1.3: Gene is transcribed from DNA to mRNA by means of Transcription process. Adapted from http://www.genome.gov/.

1.4 Regulatory regions in DNA

DNA sequence contains several regions that regulate the transcription. These regions are fundamentals for the positioning of the basic transcriptional machinery or for the regulation. In eukaryotic organisms, these regions are the following: core, proximal and distal promoter, as shown Figure 1.5 [4].

Core Promoter

The core promoter is a short of DNA (300-500bp) that contains several DNA elements that facilitate the binding of regulatory elements, as shown Figure 1.6

[4] Adapted from [55]

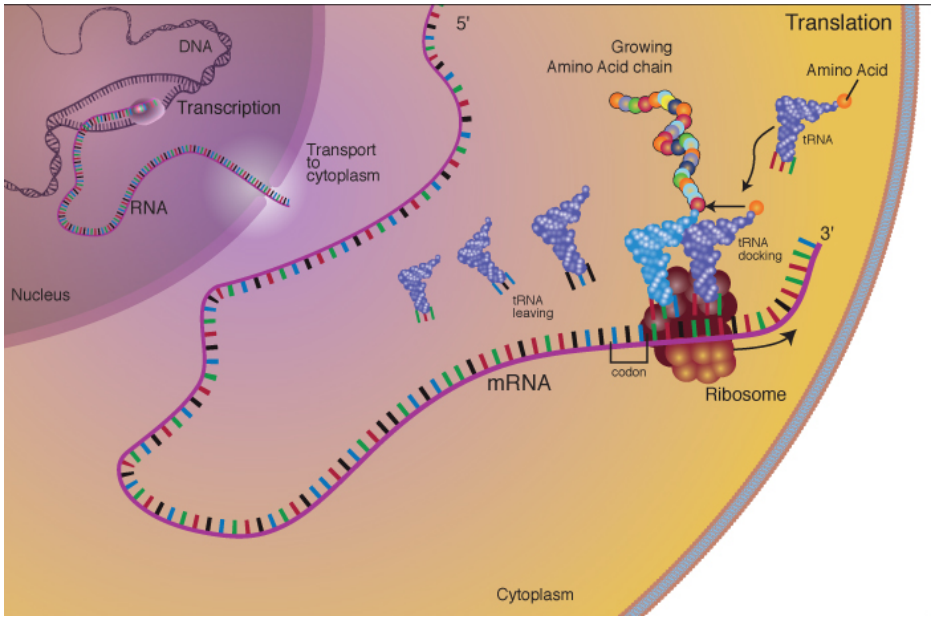


Figure 1.4: mRNA is translated to a sequence of amino acids through translation process. Adapted from http://www.genome.gov/.

[5]. Specifically, the core promoter on the DNA is capable of initiating basal transcription and it is used to position the RNA polymerase (RNAP). The main element of the core promoter is the TATA-box, an AT-rich sequence acts as a binding site for TATA-binding protein (TBP). TBP together with TATA-associated factors (TAFs) forms the complex TFIID, first step on the transcriptional complex. More core promoter elements are the following [60]:

1. **Inr**. Initition (Inr) motif contains Transcription Start Site (TSS).
2. **BRE**. BRE is present in a subset of the TATA-containing core promoters. According to the position regarding to the TATA-box, BRE can be define upstream or downstream. Moreover, BRE can act in both a negative and a positive manner.
3. **DPE**. DPE (downstream promoter element) is another motif important for transcriptional activity. DCE is formed by three subunits.

[5]Adapted from [60]

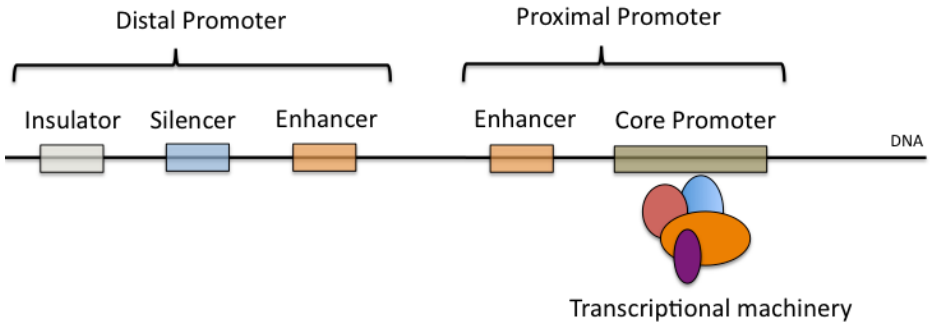


Figure 1.5: Schema of the regulatory regions in DNA [55].

4. **MTE**. Motif ten element.
5. **DCE**. Downstream core element.
6. **XCP1**. X core promoter element 1.

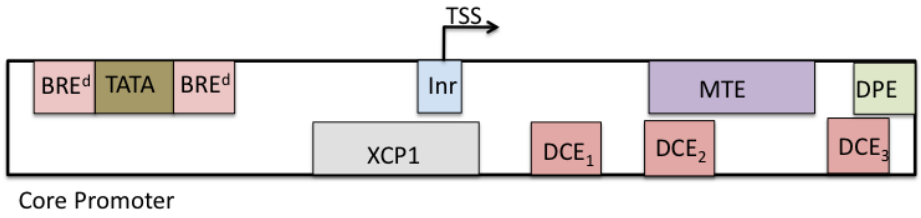


Figure 1.6: Schema of the core promoter structure [60].

Proximal and Distal Promoter Elements

In general, the proximal and distal promoters are sequences upstream of the gene. Specifically, distal promoter contains primary regulatory elements. On the other hand, proximal promoter is composed of additional regulatory elements, often with a weaker influence than proximal promoter. Proximal and distal promoter elements are the following: enhancers, silencers and insulators.

Enhancers

An enhancer is a short region of DNA that brings out transcription levels of genes. These enhancer regions are non-coding sequences that are strongly conserved and can be found up and downstream of the TSS. Its main characteristic is that regulates gene expression in highly specific tissues, developmental stages, or combinational of these [107].

Silencers

A silencers is a short region capable of binding transcription regulation factors termed repressors. Two distinct classes of silencers exist: position-independent silencers and position-dependence silencers or negative regulatory elements (NREs). Position-Independent motifs are related to pre-initiation complex (PIC) and are normally found upstream of the TSS. On the other hand, NREs function is prevent the binding of TFs to their respective cis-regulatory motif and can be found both up and downstream of the TSS [60].

Insulators

An insulator is DNA sequence that can block enhancers and silencers interactions. Two distinct types of insulators exist: enhancer-blocking insulators and barrier insulators. The enhancer-blocking insulators control gene activation by enhancers and interfere with the enhancer-promoter interaction only if insulator is between the enhancer and the promoter. Whereas, barrier insulator safeguards against the spread of heterochromatin.

1.5 Architecture of DNA-binding domains

To understand better the transcriptional regulatory processes, it is essential to know the structure of the DNA-binding domains. A DNA-binding domain is the specific region of a DNA-binding protein that allows to bind to TFs with DNA. The main DNA-binding domain structures are the following: *Helix-turn-Helix (HTH)*, *Zinc-Finger* and *Domains with α-helix*

Helix-turn-Helix

Helix-turn-Helix (HTH) is made up of two α-helix linked by a sharp β-turn. There are some domains that contain this structure. These domains are Homeodomain, Myb domain, Forkhead or winged helix and ETS-type DNA-binding domains, as shown in the Figure 1.8.

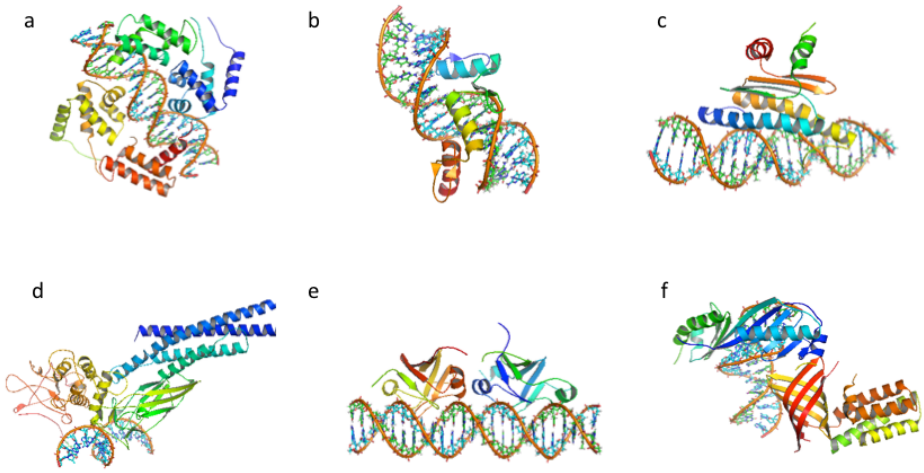


Figure 1.7: Different DNA-binding domains.**a** Helix-turn-Helix(pdb code 1IC8). **b** Zinc-Finger(pdb code 2KMK). **c** Domains with α-helix (pdb code 1C7U), **d** β-barrel (pdb code 2KIN), **e** β-sandwich (pdb code 1BG1) and **f** Domains with β-strand (pdb code 1NH2). All figures were produced with PyMol 1.3.

Homeodomain and Myb domain are characterized by the conserved pattern of hydrophobic amino acids that form a hydrophobic core between three α-helices (Figure 1.8 (a and b)) [70]. On the other hand, Forkhead or winged helix and ETS-type DNA-binding domains have a structure of the hepatocyte nuclear factor (HNF3). Forkhead DNA-binding domain consist of three helices containing a HTH motif, a twisted, antiparallel three β-sheet and C-terminal random coil (Figure 1.8 (c)). The ETS is composed of three helices containing a HTH motif and a four-stranded antiparallel β-sheet (Figure 1.8 (d)).

Zinc-finger

Zinc-finger consists of a two-stranded β-sheet and a α-helix that is held together by a zinc ion ligated to two cysteine and two histidine residues. Other proteins show different zinc-finger modules. These domains are Nuclear hormone receptors, GATA factors and Zn_2Cys_6 binuclear cluster, as shown in the Figure 1.9.

Nuclear hormone receptors are characterized by eight cysteine residues that nucleate two zinc-binding clusters. The main feature of this structural domain is an arrangement of two helices that are oriented perpendicularly to each other and

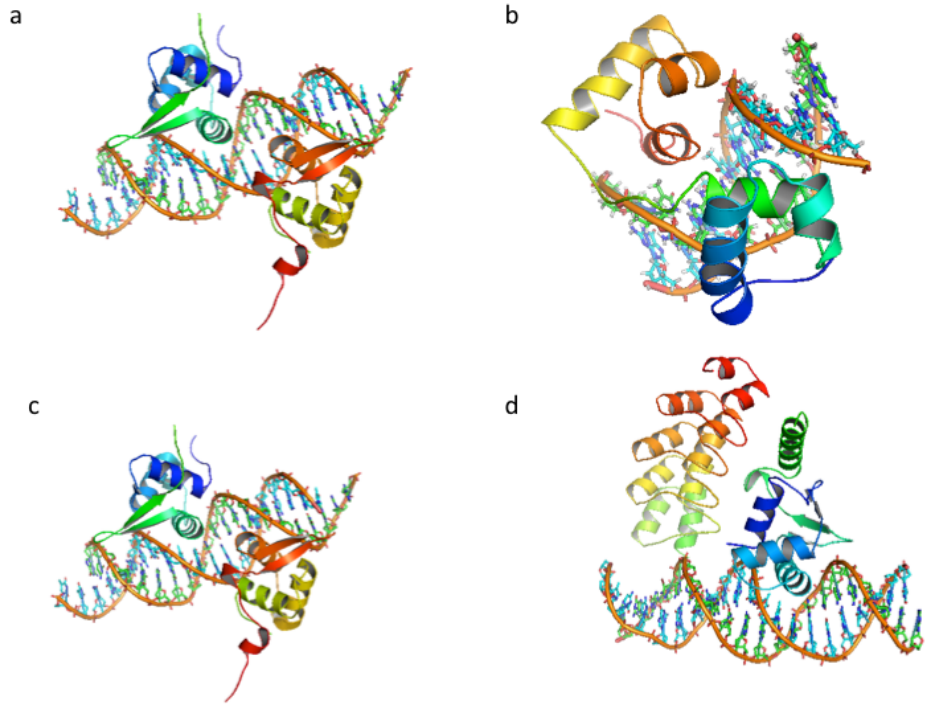


Figure 1.8: Different Helix-turn-Helix domains,**a** Homeo (pdb code 1K78), **b** Myb (pdb code 1MSE), **c** Forkhead (pdb code 3G73) and **d** ETS type DNA-binding domain (pdb code 1AWC). All figures were produced with PyMol.

crossing at their mid-points (Figure 1.9 (a)). GATA factors contain a Cys_2Cys_2 motif. This structure is characterized by zinc ion bound to its four cysteine ligands form the core of the domain. The core is composed of two small, irregular, two-stranded antiparallel β-sheets and an α-helix followed by a long loop that leads into the carboxyl terminal tail [70] (Figure 1.9 (b)). And finally, Zn_2Cys^6 has six cysteines that ligate two zinc ions in a single cluster. Moreover, it has two short α-helices each capped at their N-terminus by a pair of cysteine ligands, as shown in the Figure 1.9 (c).

α-helical structure

A part of HTH and Zinc-finger, there is a set of TF that are characterized for its α-helical structure. These TF are classified in different patterns according

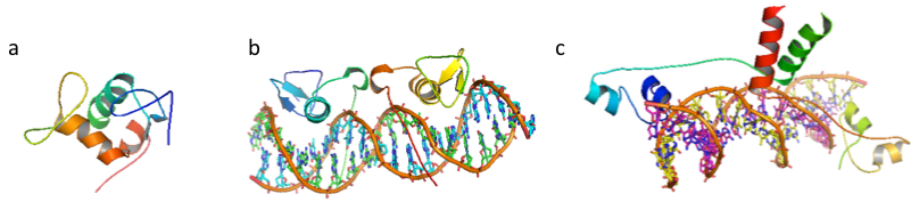


Figure 1.9: Different Zinc-finger modules.**a** Nuclear Hormone Receptor (pdb code 2EBL), **b** GATA factors (pdb code 3DFU) and **c** binuclear cluster (pdb code 1D66). All figures were produced with PyMol.

to the residues as following: MADS, basic Leucine Zipper (bZip motif), basic Helix-loop-Helix (bHLH) and basic Helix-loop-Helix Zipper (bHLHZ) motifs and High mobility group domain (HMG domain), as shown Figure 1.10.

MADS is characterized by its three layers. N-terminus is an antiparallel coiled of two amphipathic α-helices , one from each monomer of the dimmer. The coiled-coil is oriented parallel to the minor groove of the binding site and the DNA is bent around the protein with each α-helix binding in adjacent major grooves. The other layer is a four stranded antiparallel β-sheet (Figure 1.10 (a)). Basic Leucine Zipper (bZip motif) consists of a basic region that binds to DBA and a leucine-rich region which is involved in the dimerization of bZip proteins (Figure 1.10 (b)). Basic Helix-loop-Helix (bHLH) is formed by a α-helix that interacts in the major grove of the DNA (Figure 1.10(c)). On the other hand, basic Helix-loop-Helix Zipper (bHLHZ) is formed by four-helix bundle with the DNA formed a coiled-coil leucine zipper (Figure 1.10 (d)). And finally, High mobility group domain (HMG domain) is formed by three α-helices held together by two hydrophobic cores (Figure 1.10 (e)).

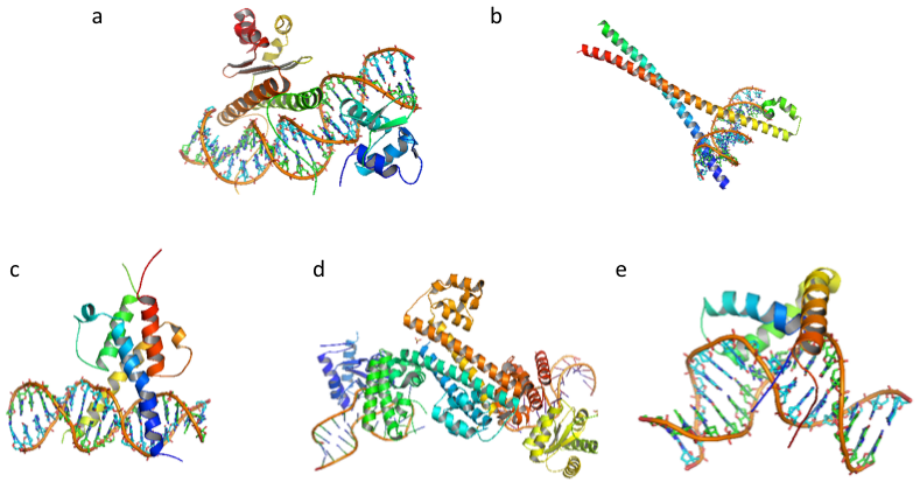


Figure 1.10: Different α-helical structure.**a** MADS (pdb code 1k60). **b** Basic Leucine Zipper (pdb code 2WT7). **c** basic Helix-loop-Helix (pdb code 1A0A). **d** basic Helix-loop-Helix Zipper (pdb code 3SIU). **e** High Mobility Group (HMG) (pdb code 3U2B). All figures were produced with PyMol.

Chapter 2

State of Art

2.1 Motif Finding Problem

Regulatory sequence detection is a critical facet for understanding the cell mechanisms in order to coordinate the response to stimuli. Protein synthesis involves the binding of a transcription factor (TF) to specific sequences, called transcription factor binding sites (TFBS), in a process related to the gene expression initiation.

The problem is that to find this kind of sequence is a complex task for several reasons. First of all, TFBS are usually short ($5-20bp$) and highly degenerate. As a consequence of this, it is difficult to establish a consensus sequences. Besides, the spatial locations of TFBS relative to a gene are highly variable. TFBS can be found close to gene or until hundred of thousands of nucleotides. Moreover, one TF may bind to more than one site. Then, these sites are degenerate DNA sequence. And finally, regulatory sequences can bind to work together with the consideration that the limits of the modules are not clear.

Due to its characteristics, great efforts have been devoted to find TFBS through experimental and computational methods. The experimental methods are classified into footprinting methods [12] and on high-throughput methods. In general, all of these methods are costly, time consuming, low resolution, significant background noise, and the considerable systematic bias [60]. As consequence of this, several computational method for the finding of protein binding sequences have been developed to decrease labour intensive, the cost and, mainly, to reduce the false positive rate and to increase the sensitivity [123].

2.2 DNA Model

To describe DNA sequences, motifs or candidate sequences, is necessary to use a model. The model provides a satisfying description of the phenomenon under study [115]. These models can be classified into motif model and sequence model according to biological sequence.

Motif Model

Many computational approaches for TFBS identification problem have been developed over the last decades. The main goal is mutual among all computational approaches, which is to find protein binding sequence. The difference among them is according to DNA motif model considered [16]. To consider a good model is essential to carry out a good TFBS detection [39]. DNA motif model is divided into three types: pattern models, profile models and, finally, numerical models.

Pattern Models

Pattern models are considered the first method to represent DNA motif which are simply strings over the 4-alphabets $\{A, C, T$ and $G\}$ than form DNA [39]. Given a set of aligned sequences with binding evidence, a consensus nucleotide sequence is assigned to each position using the IUPAC (International Union of Pure and Applied Chemistry) from the frequency of nucleotides in the positions of the aligned sequences. The nucleotide with the highest frequency is taken as the representative nucleotide. If there is ambiguity among nucleotides, IUPAC code allows represent it through another letters, i. e. R, Y,...., to see Figure 2.1.

Figure 2.1: Left to right: Consensus sequence and IUPAC code

Through consensus sequence, a simple regular expression search is enough for TFBS detection. This model has two main limitations. First of all, consensus sequence assumes that the positions are independent and, moreover, the majority binding sites are not represented with a consensus sequence [100].

Profile Models

A complex model was introduced to represent the majority binding sites through profile model called Position Weight Matrix (PWM). Given a set of objects (letters), this model represents these objects on profile from its empirical frequency. This model allows to capture information on the variability of a set of binding sites in a quantitative manner, which is not possible through pattern model [39]. According to quality of the set of objects, there are different kinds of approximation based on profile models.

Position Independence Model The first approximation considers that the set of objects does not show dependencies among positions [117, 100]. This model is based on Position Count Matrix (PCM), which is estimated from a set of aligned sequences with binding evidence. PCM is a $4 * n$ matrix that contains the number of sequences with letter $\{A, C, T$ and $G\}$ in each position, to see Figure 2.2. Dividing PCM by the number total of sequences, N, obtains the position frequency matrix (PFM). Using the eq. (2.1), the frequency matrix is converted to a position weight matrix (PWM) or position specific scoring matrices (PSSMs).

$$W_{b,i} = log_2 \frac{p(b,i)}{p(b)} \tag{2.1}$$

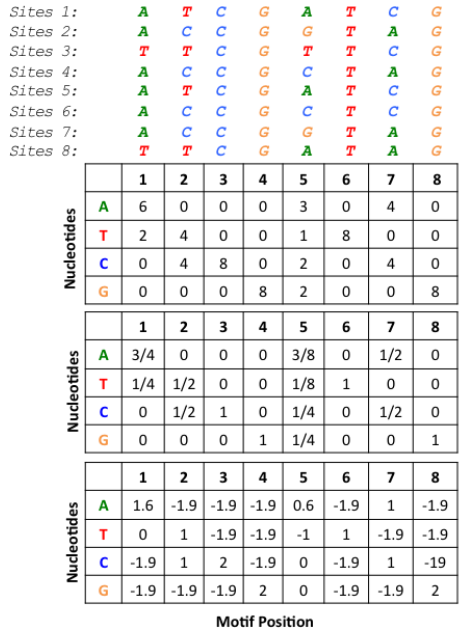
where $p(b)$ is the background probability of the base b, $p(b,i)$ is the experimental frequency of the base b in position i. The equation (2.1) converts normalized frequency to a log-scale [130], to see Figure 2.2. The null values are a problem for the log-conversion. There are different approximations to eliminate null values [56]. A sampling correction, known as *pseudocounts*, is added to each cell of the PFM [130]. This correction is defined as the square root of the number of sites, to see eq. (2.2)

$$p(b,i) = \frac{f_{b,i} + s(b)}{N + s(b)} \tag{2.2}$$

From position weight matrix, the score magnitude is defined as [116], to see equation 2.3

$$score = \sum_{i=1} w_{b,i} \tag{2.3}$$

The advantages of this model are that the scores are proportional to binding energies and the nucleotide probability values can determine the total the information content for each position. But, at the same time, this model assumes that the positions are independent which is its main limitation.



Sites 1:	A	T	C	G	A	T	C	G
Sites 2:	A	C	C	G	G	T	A	G
Sites 3:	T	T	C	G	T	T	C	G
Sites 4:	A	C	C	G	C	T	A	G
Sites 5:	A	T	C	G	A	T	C	G
Sites 6:	A	C	C	G	C	T	C	G
Sites 7:	A	C	C	G	G	T	A	G
Sites 8:	T	T	C	G	A	T	A	G

Nucleotides	1	2	3	4	5	6	7	8
A	6	0	0	0	3	0	4	0
T	2	4	0	0	1	8	0	0
C	0	4	8	0	2	0	4	0
G	0	0	0	8	2	0	0	8

Nucleotides	1	2	3	4	5	6	7	8
A	3/4	0	0	0	3/8	0	1/2	0
T	1/4	1/2	0	0	1/8	1	0	0
C	0	1/2	1	0	1/4	0	1/2	0
G	0	0	0	1	1/4	0	0	1

Nucleotides	1	2	3	4	5	6	7	8
A	1.6	-1.9	-1.9	-1.9	0.6	-1.9	1	-1.9
T	0	1	-1.9	-1.9	-1	1	-1.9	-1.9
C	-1.9	1	2	-1.9	0	-1.9	1	-19
G	-1.9	-1.9	-1.9	2	0	-1.9	-1.9	2

Motif Position

Figure 2.2: Up to down: A set of aligned binding sequences, Position Contant Matrix, Position Frequency Matrix and Position Weight Matrix.

PWM Extension Position Weight Matrix assumes two strong assumptions: independence between positions and that the variations of the TFBS come from the same consensus sequence. Biological experiments have shown the dependence among binding site positions [123] and that TFBS occur in clusters of functionally interacting TF in promoter regions, called transcriptional modules [13, 73]. A single factor may have different interaction partners for different genes [37]. To incorporate these characteristics, PWM have been generalized through mixture model and position dependencies model.

Position Dependencies Model PWM model considers to one mononucleotide for each column. In order to consider the dependence among adjacent

positions of the binding site, it is necessary to consider to multi-nucleotide at each position [102, 40]. This model fit better to the real binding sites, but a large number of cis-regulatory sequences are necessary to circumvent overfitting. Therefore, the mono-nucleotide models have only been rarely used [117]. To consider the dependence among all positions, another model it is necessary due to mono-nucleotide requires to many parameters. Bayesian networks [87, 9] , Markov models [139] and tree model are different models to consider general dependencies among positions.

Mixture Model As is well known, a single transcription factor might bind with different pattern of cis-regulatory sequences. Due to this characteristic, PFM does not fit correctly due to consider a single distribution for each position. The single distribution can be substituted by mixture distributions where each parameter represents one kind of binding [9]. The limitation of this model is that the number of parameter is high. However, many positions can be fitted with a single distribution. Therefore, there is model that combines both distributions [37].

Numerical Models

The conversion of genomic sequences from the symbolic, $\{A, T, C$ and $G\}$, into digital genomic signals allows the detection of protein binding sequences based on numerical methods. All kinds of mappings of symbolic genomic data have to be both truthful and unbiased. The mapping is truthful if all biologically relevant characteristics of the represented objects are expressed in corresponding mathematical properties of the samples in the resulting digital signal [27]. The methods of symbolic-to-digital conversion of genomic sequences are as follows:

Real Representation A simple way to transform a symbolic DNA sequence to a numerical vector is to assign arbitrarily four real numbers to represent the four nucleotides (e.g. $A = 1$, $T = 2$, $C = 3$, $G = 4$). This assignment may introduce an additional mathematical assumption such as $A < T < C < G$, which does not exist in symbolic sequences [136].

4-D Binary Indicators A symbolic DNA sequence can be represented by means of the presence or absence of four nucleotides on the position. This numerical representation is known as Voss's 4-dimensional binary indicator. For a symbolic DNA sequence, $s[0]$,...,$s[N-1]$, over the alphabet $\{A, T, C$ and $G\}$, four binary indicators sequences, $u_A(n)$, $u_T(n)$, $u_C(n)$, and $u_G(n)$ were proposed to identify the positions of the four symbols, that is $u_A[k] = 1$ if $s[k] = A$, and 0 otherwise, and similarly in the remaining three cases [129].

The advantage of the binary indicator representation of a DNA sequence is that it does not predefine any mathematical relationship among the symbols, only indicating the frequencies of the symbols. Thus it is widely utilized in the studies of detecting symbol distributions and periodicity features of a sequence [136]. However, this kind of representation is degenerated because two different DNA sequences may have the same power spectrum.

3-D Representation The 4-D vector representation can be reduced to three dimensions in which each of the four type nucleotides is represented as a 3-D vector having magnitude equal to 1 and pointing to the four directions from the centre to vertices of a regular tetrahedron [108]:

$$\begin{aligned}
\vec{a} &= +\vec{k}, \\
\vec{t} &= +\frac{2\sqrt{2}}{3}\vec{i} - \frac{1}{3}\vec{k}, \\
\vec{c} &= -\frac{2\sqrt{2}}{3}\vec{i} + \frac{\sqrt{6}}{3}\vec{j} - \frac{1}{3}\vec{k}, \\
\vec{g} &= -\frac{2\sqrt{2}}{3}\vec{i} - \frac{\sqrt{6}}{3}\vec{j} - \frac{1}{3}\vec{k},
\end{aligned} \tag{2.4}$$

The mathematical description of the code can be simplified by rotating the reference system [21]. It is also advantageous to give up the Euclidian normalization condition and to choose integer ± 1 coordinates for the vertices of the cube, including the points representing the bases, so that the base vectors in (2.5) take the simple form (2.5), as shown in Figure 2.3.

$$\begin{aligned}
\vec{a} &= \vec{i} + \vec{j} + \vec{k}, \\
\vec{t} &= \vec{i} - \vec{j} - \vec{k}, \\
\vec{c} &= -\vec{i} + \vec{j} - \vec{k}, \\
\vec{g} &= -\vec{i} - \vec{j} - \vec{k},
\end{aligned} \tag{2.5}$$

Each of the six edges corresponds to one class comprising a pair of nucleotides, as shown in Figure 2.3. The representation is three dimensional and the axes express the differences "weak minus strong bonds", "amino minus keto", and "purines minus oyrimidines".

Complex Representation Projecting the basic tetrahedron on a plane can reduce the dimensionality of the representation. Such planes can be chosen

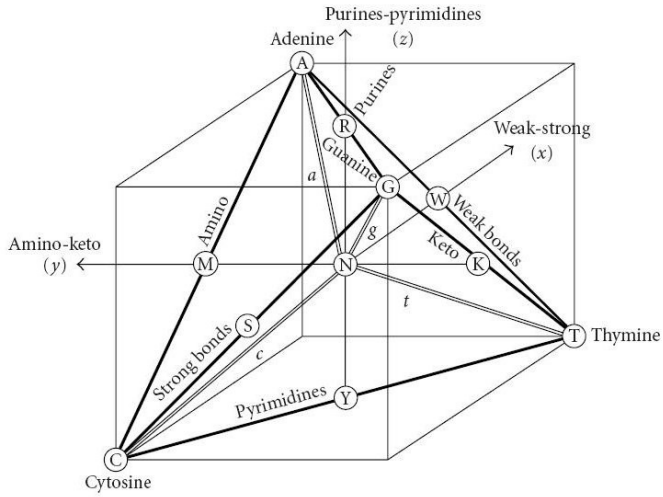


Figure 2.3: Nucleotide tetrahedron [21].

in various ways that conserve the symmetry of the representation and reflect biological properties in corresponding mathematical properties [21]. For instance, a pair of the Cartesian coordinate axes can define the planes. This representation is accompanied by some loss of visual information associated with crossing and overlapping of the resulting curve representing DNA by itself [113]. On the other hand, the planes can be put in correspondence with a complex plane. The complex representation has the advantage of better translating some of the features of the basis into mathematical properties.

Z curve Representation The Z curve consists of a series of nodes P_n ($n = 1, 2, ..., N$), whose coordinates are denoted by x_n, y_n and z_n:

$$\begin{aligned} x_n &= 2(A_n + G_n) - n, \\ y_n &= 2(A_n + C_n) - n, \\ z_n &= 2(A_n + T_n) - n, \end{aligned} \tag{2.6}$$

where A_n, C_n,G_n and T_n are the cumulative counts of nucleotides, A, C, G and T from beginning to the position n, respectively, $n = 0, 1, 2, ..., N$ and $A_0 = C_0 = G_0 = T_0 = 0$. Therefore, a DNA sequence can be decomposed into three series of digital signals. The connection of the nodes by line is called the Z curve of the DNA sequence [136].

DNA Sequences Model

Any computational method for motif detection needs a DNA sequences model to estimate the random expectation. DNA sequences model is applied to generate null model and servers as a reference to decide if motif might have specific biological function [115]. The main sequence model are permutation model, Bernoulli model and Markov model [115].

Permutation Model

The permutation model is first approximation of building random sequences. This model builds a set of all sequences, S, that have the same length, l, and containing the same number of $\{a, c, g \text{ and } t\}$ that the observed sequence S_{obs}. The characteristics of observed sequence are the frequencies of the nucleotides, dinucleotides,... The stochastic model assumes that each nucleotide is independent and that observed sequence has been randomly sampled from S with uniform probability [115]. The length of the observed sequence and the words with overlaps would make the calculation infeasible. This leads to the next approximation, Bernoulli Model.

Bernoulli Model

Bernoulli or Multinomial model considers that the sequence, S, with length l, is a succession of independent random residues $\{a, c, g \text{ and } t\}$ with probabilities $\mu(a)$, $\mu(c)$, $\mu(g)$ and $\mu(t)$ [115]. The simplest version of this model assumes that residues are equiprobable, but it is not a realistic case. Generally, Bernoulli model assumes that the probability is specific in each residue and constant on the whole sequence. In this case, the probability for a sequence, $\mu(w)$, is the product of the probability of the residues. The generalization of the Bernoulli model is Markov chain model, which assumes that the distribution of each residue depends of previous residues.

Markov Model

The Markov model is a probabilistic model that assumes that the residues are not necessarily independent and that the sequence is homogeneity. This means that the probability of the residues at the position i depends on the previous residues and that the sequence has the same probabilistic behaviour from its beginning to its end [115]. The order de Markov chain, m, represents the length of the memory in the sequence, e.g. when $m = 1$ each residues depends on its predecessor. For m is equal to 0, Markov model is Bernoulli model. The observed sequence allow estimate the transition probability. For each couple of residues

(a, c) from Υ, the transition probability $\mu(a, c)$ is the probability that X_i is a c, given X_{i-1} is an a, to see equation 2.7.

$$\mu(a, c) = \rho\{X_i = cX_{i-1} = a\} \tag{2.7}$$

The homogeneity assumption implies that the transition probability $\mu(a, c)$ does not depend on position i, $\sum_{c\in\Upsilon} \mu(a, c) = 1$. Then, the distribution of the residues X_i given the previous letter X_{i-1} is estimated from transition and conditional probabilities, equation 2.8.

$$\rho\{X_i = c\} = \sum_{a\in\Upsilon} \rho\{X_{i-1} = a\}\rho\{X_i = cX_{i-1} = a\} \tag{2.8}$$

$$= \sum_{a\in\Upsilon} \rho\{X_{i-1} = a\}\mu(a, c) \tag{2.9}$$

2.3 Motif Detection Methods

The identification of specific regulatory motifs or transcription factor binding sites in non-coding DNA sequences, which is essential to elucidate transcriptional regulatory networks, has emerged an obstacle for many researchers. Consequently, numerous motif discovery tools have been applied to solving this problem [132, 86, 48, 23]. The motif discovery *in silico* are classified into different classes according to the approach to the problem: deterministic, numerical, stochastic and phylogenetic.

Deterministic Algorithms

A word-based or deterministic algorithms search coincides with a pattern sequence, normally consensus sequences. Specifically, the pattern sequence is matched against candidate sequence. Each position is evaluated to a binary value indicating success or not [98]. Due to the characteristics of the cis-regulatory sequence, these algorithms can produce predictions with a low rate of false positives as well as a high rate of false negative. Deterministic algorithms are divided into different models according to model used. These models are: Oligo model, regular expression and mismatch expression.

Oligo model

One of the first contributions in computational methods for cis-regulatory sequences is based on oligo model. This model assigns 1 for a exact match, and 0 for all other sequences. One of the main algorithm based on oligo model is

Dyad Analysis [126]. This algorithm is based on a systematic counting of pairs of short words separated by a fixed distance, called spaced dyads, followed by a calculation of their statistical significance [126]. The results are admissible for short sequences. However, this method suffers important limitations: the motifs can not include spacers and, moreover, oligo model considers only exact matches.

Regular expression

The regular expression model is based on word counting methodology. In contrast to oligo model, the regular expression model is more flexible than oligo model. In this way, this model assigns 1 if the given substring is matched by an underlying regular expression. The models used in motif discovery are typically composed of exact symbols, ambiguous symbols, fixed gaps and/or flexible gaps [16]. Moreover, the motifs can include spacers. Sisha et al [111] proposed an algorithm based on regular expression model for the detection of transcription factor binding sites.

Mismatch expression

The last methodology based on deterministic approaches is mismatch expression. This methodology calculates the number of mistmatches, or Hamming distance, between the candidate sequence and the consensus. If the number of mistmatches is below a given threshold, the algorithm assigns 1 [124, 91]. From this idea, there are different variant of the model, e. g. the threshold can be considered as the sum of mistmatches between all motif occurrences and the underlying motif substring [66]. There are different algorithms that use mistmach expression for binding detection: Weeder [85] and Smile [74].

Numerical Algorithms

The conversion of genomic sequences from the symbolic into digital genomic signals allows using genomic signal processing for detection of protein binding sequences. A several number of methods have been employed, e.g: Support Vector Machines (SVM) [127]. The main algorithms are based on following techniques: Discrimiant analysis, Principal Component Analysis, Support Vector Machines and two-class Kernel Method.

Linear Discriminant Analysis and Quadratic Discriminant Analysis

Two classical statistical pattern-recognition methods are Linear discriminant analysis (LDA), [49] and Quadratic discriminant analysis (QDA). This method

has been used to categorize samples into two classes. Linear discriminant analysis (LDA) estimates an optimal plane surface that best separates points into two classes. Whereas, Quadratic discriminant analysis finds an optimal curved surface instead [138]. Simonis et al [109] applied this classifiers to predict gene co-regulation. Specifically, Simonis et al. established a method to discriminate co-regulated from non-co-regulated genes on the basis of counts of pattern occurrences in their non-coding sequences.

Support Vector Machines

Several algorithms based on Support Vector Machines (SVM) has been used by prediction regulatory motifs [118, 46, 52]. The goal of these algorithms is to classify the data in two class or categories: positive and negative binding sites [14]. In order to do this, the separator is chosen according to the maximum margin. The margin is the distance of the closest data point to the separator. Therefore, a SVM is a maximum margin classifier with an adjustable cost parameter.

The decision function, $D(x)$, is build using the training scores. Given a training set of number of scores of sequence $\{x_1, x_2, ..., x_n\}$ with know labels $\{y_1, y_2, ..., y_n\}$ where $y_i \in \{-1, +1\}$ such that, -1 are negative biding sites and $+1$ are positive binding sites, to see equation (2.10,2.11).

$$x \in class(+) when D(x) \geq +1;\ if y_i = +1 \tag{2.10}$$

$$x \in class(+) when D(x) \geq +1;\ if y_i = +1 \tag{2.11}$$

where $i \in \{1, 2, 3, 4, ..., n\}$ and x is on decision boundary when, $D(x) = 0$. Then, the discriminant function of hyperplane that divides the data points two classes is (eq. 2.12),

$$D(x) = w * x + b \tag{2.12}$$

where w is the weight vector, b is the bias value and $w * x$ is the dot product between the two vectors w and x. For correctly classification scores, it is necessary to optimize $min \frac{1}{2} w^2$. Several approach has been used to solve the optimization problem [11, 15].

The data may be non-linearly. In this situation, the linear classifier does not classify satisfactorily. In order to solve this problem, one approach is to map the data into a high dimensional feature space, called a feature space, including non-linear features and then use a linear classifier. In this new space, nonlinear decision boundaries have been estimated using kernel methods. The kernel functions widely used are the following:

- Polynomial kernel of degree d

$$k_{d,x}(x, x') = (x * x' + \delta)^d \tag{2.13}$$

- Gaussian kernel

$$k_\gamma(x, x') = exp(-\frac{1}{\gamma}x - x'^2) \tag{2.14}$$

  where γ is the width of the Gaussian Kernel.

- Sigmoid kernel

$$k_\alpha(x, x') = tanh(\alpha(x * x' + \delta)) \tag{2.15}$$

One-class SVM To apply SVM is necessary to know the positive and negative samples. Sometimes, it is no possible to determine experimentally where transcription factor will certainly not bind (a negative set of sequences) [47]. In this sense, recognition of TFBSs only may be characterized as a one-class classification problem [52]. There are several different approach for the one-class problem [52, 103, 119]. Jian et al [52] estimate the support of probability of distribution of known TFBS through one-class SVM and incorporates multiple factors to aid the recognition of TFBS. Schölkopf et al [103] proposed hyperplane method and, finally, Tax et al [119] used a method based on Support Vector Data Description (SVDD). This method creates outlier uniformly in and around target class. One-class SVM is based on hyperplane method. Given a set of training examples, $\{x_1, x_2, ..., x_n\}$ for a class X where, $X \in R_n$. If the mapping function is such as $\phi : X \to H$. Where H is the feature space. The optimisation problem is defined as, to see equation 2.16.

$$min\{\frac{1}{2}w^2 + \frac{1}{nv}\sum_{i=1}^{n}\xi_i - \rho\} \quad w * \phi(x_i) \geq \rho - \xi, i = 1, 2, 3, .., n \quad \xi_i \geq 0. \tag{2.16}$$

The decision function is,

$$D(x) = sign(w * \phi(x) - \rho) \tag{2.17}$$

where w is the weight vector, υ is the upper bond of fraction on the outliers and lower bound on the fraction of the support vector and $0 < \upsilon \leq 1$, ξ is the slack variable to penalize misclassification, ρ is the bias and n is the number of examples.

Stochastic Algorithms

Computational methods for motif detection based on probabilistic models avoid any numerical representation of the nucleotides. The main approaches based on statistical used to detect TFBS are Expectation maximization (EM) and Gibbs's sampling (GS). Both approaches is based on PWM, whereas EM is deterministic and GS is stochastic. Given the same set of initial parameter, EM will always converge to the same solution. Instead, GS may give different solutions.

Expectation maximization

Expectation maximization (EM) algorithm is a deterministic approach for TFBS detection. This algorithm considers that all sequence is composed for two parts statistically different: background genomic sequence and the binding sites. Each part is modeled differently. The binding sites are modeled as a PFM where each $P(b,i)$ is the probability of observing a specific based $b \in \{A,C,G,T\}$ and i is the position. Instead, the background is modeled as an overall probability for each of the four bases. The overall probability is $P(b,0)$ where 0 is any position except binding site positions. In the simplest case, the overall probability is equiprobable.

The motif finding problem consists of determining the PFM for TF, binding site locations and background probability. To do this, EM takes as an input a set of unaligned sequences and a motif length and returns probabilistic model of the motif. EM algorithm assumes that each sequence of the dataset contains a motif whose position is unknown. The motif has been generated by a sequence of independent and multinomial random variables. As the set of sequences in the dataset are unaligned, it is necessary to determinate an offset. EM considers a initial step, $P_0(b,i)$ and $P(b,0)$, where each sequence and each position are equally likely to be true binding sites. This measure is used to estimate $P(b,i)$. EM algorithm recalculates successively $P_0(b,i)$ and $P(b,i)$ until is minor than error, ϵ. The likelihood [5] of the model is just the probability of the data given the model, to see equation 2.18.

$$log(likelihood) = N \sum_{j=1}^{W} \sum_{b \in \{A,C,G,T\}} P(b,j) * log(P(b,j)) \tag{2.18}$$

$$+ N(L-W) * \sum_{b \in \{A,C,G,T\}} P(b,0) * log(P(b,0)) \tag{2.19}$$

$$+ Nlog(\frac{1}{L-W+1}) \tag{2.20}$$

where N is the number of the sequences in the dataset, L is the length of the sequences, W is the length of the shared motif. EM algorithm determines a local maximum for the likelihood of the model parameters [25]. Although, EM has some limitations [30], which are associated with data input. First, the algorithm assumes that the user knows the length of the binding site, but usually this is not the case. Second, EM assumes that in each sequence in dataset contains exactly one motif. This means that these sequences with multiple appearances will under-contribute, and the sequences with no motif will be over-represented [5]. Therefore, the algorithm's sensitive depends of the set of initial parameters [5]. The main computational method based on EM is MEME (Multiple expectation–maximization for Motif Elicitation) [5, 7]. (MEME) algorithm is employing a maximum-likelihood ratio heuristic for determining the best number of model free parameters, found by help of an EM-based approach on a two-component mixture model. The algorithm is multi-initialized for searching over several possible motif widths and a greedy algorithm seeks multiple motifs [5, 7]. For each motif discovered, MEME reports the occurrences (sites), consensus sequence, and the level of conservation (measured as the information content) at each position in the pattern. The MAST (Motif Alignment and Search Tool) sequence homology search algorithm uses the QFAST algorithm to calculate the statistical significance of the found matches. MAST compares a group of motifs to each sequence in a database of sequences. For each motif, it finds the position in the sequence that best matches it, calculates the -value of the match (position value)[6]. This value is normalized for the length of the sequence (sequence value). The significance of the combined match is finally obtained from the product of values for all the motifs. MEME/MAST may encounter local-maximum problems when dealing with large data sets and requires of multiple runs to ensure meaningful finding

Gibbs' sampling

The Gibbs' sampling is a stochastic approach for TFBS detection. This algorithm uses a random sampling step. It means, indeed, that with same initial parameters, the solutions may be different. Therefore, Gibbs' sampling is more likely to find global optimum than EM, if run enough. Gibb's sampling take as an input a set of sequences and returns a probabilistic model from Bayesian theory [64, 30]. The main characteristic of Gibb's sampling is that the algorithm requires no prior knowledge about binding site to build the optimal motif profile. Gibb's algorithm iterates in various steps: predictive update and sampling step. First of all, a sequence is chosen randomly from a set of sequences. From this set of sites aligned, a probabilistic profile and background model are generated. From these measurements, sampling step, the likelihood ratio is calculated for each possible subsequence in the selected sequence. By means of likelihood ra-

tion, a new motif start position is estimated. The stochastic process ensure that the model is the global optimum. One a optimal motif profile has been generated, the algorithm is re-run for allowing predictions for multiple binding motifs. Some algorithms based on Gibb's sampling are[125], GLAM [36], Motif Sampler[120], AlignACE [50]....GLAM is an algorithm to estimate the width of the aligned motif automatically [36] and the statistical significance alignment. MotifSampler [120] uses Gibbs sampling to model the background through Markov model. And finally, AlignACE [50] takes as an input a set of sequences and returns these motifs that are overrepresented.

Phylogenetic Motif Model

Phylogenetic Motif Models (PMMs) considers the idea of the evolutionary conservation [44] which is an extension of the PWM. Given a motif, PWM is based on an estimation of the probability in a single genome. Whereas, PMM takes into consideration the probability of an ungapped region in a multiple alignment sequences that evolved independently from a mutual ancestral.

A class of scanning algorithm, which are an extension of simple PWM scanning algorithms, has used this model. The main difference between these algorithms is that PWM scans a single sequence, whereas PMM algorithm scans a multiple alignment of orthologous sequences. This approach needs an explicit model nucleic acid substitution and a phylogenetic tree. These model describes the relationship and evolutionary distance among the species [44]. A motif detection based on PMM is Monkey algorithm [79]. MONKEY[79] is an algorithm of identification of TFBS in multispecies alignments. MONKEY algorithm uses a PMM [42] to calculate the likelihood of conserved sites and statistical significance to each hit.

Part II

Binding Sites Detection

Chapter 3

Computational Detection of Transcription Factor Binding Sites through Differential Rényi Entropy

This chapter is an exact copy of the paper:

- **Computational Detection of Transcription Factor Binding Sites through Differential Rényi Entropy**. J.Maynou, M. Vallverdú, F. Clarià, J.J. Gallardo-Chacón, P. Caminal and A. Perera. IEEE Trans. Information Theory, vol. 56, no. 2, pp: 734-741, Feb. 2010.

3.1 Abstract

Regulatory sequence detection is a critical facet for understanding the cell mechanisms in order to coordinate the response to stimuli. Protein synthesis involves the binding of a transcription factor to specific sequences in a process related to the gene expression initiation. A characteristic of this binding process is that the same factor binds with different sequences placed along all genome. Thus, any computational approach shows many difficulties related with this variability observed from the binding sequences. This paper proposes the detection of transcription factor binding sites based on a parametric uncertainty measurement (Rényi entropy). This detection algorithm evaluates the variation on the total Rényi entropy of a set of sequences when a candidate sequence is assumed

to be a true binding site belonging to the set. The efficiency of the method is measured in form of Receiver Operating Characteristic curves on different transcription factors from *Saccharomyces cerevisiae* organism. The results are compared with other known motif detection algorithms such as Motif Discovery scan (MDscan) and Multiple EM for Motif Elicitation (MEME).

Binding Sites, Gene Regulation, Motif Detection, Rényi Entropy, Sequence Analysis, Transcription Factor.

3.2 Background

Deoxyribonucleic acid (DNA) is a molecule that contains the genetic instructions used in the development and functioning of all known living organisms. A DNA molecule consists of two long chains composed by the combination of four nucleotides (adenine (A), thymine (T), cytosine (C) and guanine (G)) joined by phosphodiester bonds [1]. Watson and Crick proposed in 1953 a theoretical structure based in the X-ray diffraction data analysis consisting in a double helix [131]. Their proposal consisted of a molecule with the nitrogenous bases in the inner part and the pentose phosphates on the outer part. The structure is stabilized thanks to the hydrogen bounds between the so called complementary bases (A with T and G with C). This simple three-dimensional structure contains the necessary material not only for the synthesis of all the necessary molecules in a living organism and its replication but also for cell self regulation. Several motifs in the DNA sequence are used by cells as the labels for the different functions such as replication spots and chromosome segregation during cell division, or methylation points for genes or even chromosomes inactivation [1]. The machinery for these cell functions is composed mainly by proteins and their synthesis start with a process called gene transcription [80]. Accurate gene temporal and spatial regulation allows the diversity in cell behaviour, which is necessary to maintain life. One first step in the control of the transcription process is the association between specific proteins with their target binding sites in the DNA sequence [45]. In addition, they also bind other modulation factors and the RNA polymerase enzyme [62]. These proteins, that carry out their function in gene regulatory regions, are known as transcription factors. Transcription factors recognise specific motifs in the DNA sequence, however they do not open the double helix in order to interact directly with nitrogenous bases but the interaction is with the domains generated by the nucleotide residues in the helix surface structure. As a result, a distinctive pattern of non covalent interactions is produced such as *Van der Waals* attraction, hydrogen and ionic bonds or hydrophobic interactions. All these bindings are weak but they have an additive effect in the association, which provides the structure with stability, specificity and flexibility [63]. Besides, due to this intrinsic complexity in the relationships

between acid nucleic and proteins, it is difficult to establish a specific sequence for binding detection [80]. Any method aiming the detection of binding sites within DNA sequences must consider the variability shown on these. Moreover, there is a connection dependence between binding site positions (along the sequence) which is related with interaction stability between the transcription factor and its binding sites.

Information theory has been applied in genetics aiming for different problems: from the visualization of the information of a sequence set to its characterization with entropies [97, 101, 100, 117]. Nevertheless, main efforts for the detection of binding sites have employed alternative ways to detect variable binding motifs, not fully based on information theory. Most relevant are the probabilistic methods, where the most representative models are based on Position Weight Matrices (PWM) or Position-Specific Weight Matrices (PSWM). A PWM is a matrix of score values corresponding to the symbols of the alphabet and its relationship to each position in a standard pattern [98]. There are several types of PWMs: frequency matrices contain the absolute frequency of a nucleotide at each motif position, weight matrices contain the relative frequency of a nucleotide at a motif position as an estimation of the probability of this fact, and finally, log-odds matrices contain the log of the quotient between the probability of finding a particular nucleotide at a certain position in sequences containing the real motif and the background frequency of the letter at the same position [130]. In particular, publicly available detectors have been published like MDScan [69] and MEME/MAST [4]. MDScan is based on the combination of word enumeration and position-specific weight matrix. This method constructs a frequency matrix from the occurrences collected for each consensus to explore the regulatory regions of the most over expressed genes [86]. Each matrix is evaluated according to the approximate Maximum a Posteriori (MAP) scoring function [69] against a background model of the $3rd$ order. The score matrices are optimized by means of a larger set of highly expressed genes. MEME/MAST is a tool for discovering motifs, sequence patterns that occurs repeatedly in a group of related DNA sequences, and for searching sequence databases using obtained motifs. Given a set of unaligned sequences, the MEME algorithm is employing a maximum likelihood ratio heuristic for determining the best number of model free parameters, found by help of an Expectation Minimization (EM) based approach on a two-component mixture model. The algorithm is multi-initialized for searching over several possible motif widths and a greedy algorithm seeks multiple motifs [4, 5]. For each motif discovered, MEME reports the occurrences (sites), consensus sequence and the level of conservation (measured as the information content) at each position in the pattern. The MAST (Motif Alignment and Search Tool) sequence homology search algorithm uses the QFAST algorithm to calculate the statistical significance of the found matches. MAST compares

a group of motifs to each sequence in a database of sequences. For each motif, it finds the position in the sequence that best matches it, calculates the p-value of the match (position p-value) [6]. This p-value is normalized for the length of the sequence (sequence p-value). The significance of the combined match is finally obtained from the product of p-values for all the motifs. MEME/MAST may encounter local-maximum problems when dealing with large data sets and requires of multiple runs to ensure meaningful findings [69].

In this paper, we propose the detection of transcription factor binding sites using a differential measure based on a parametric entropy. The method evaluates a total parametric entropy contained in an aligned set of sequences with known binding and analyses the total information change when the candidate sequence is included in the set. The performance of the parametric entropy measure based detector is compared against MDScan and MEME/MAST.

3.3 Method

Given the assumption that the total information content can be computed from a set of similar objects, the variation on this total information when a new object is added to this set will depend on the similitude of the new object to the set. The total information will not vary if the new object does not add variability on the previous set. On the other hand, the total information will increase if this new object is dissimilar to the set.

This paper employs this rational for constructing a detector based on the total variation of Rényi parametric entropy of a set of sequences when a new sequences is assumed that belongs to the set.

From this principle, the proposed method starts with a matrix of aligned sequences with binding evidence. The transcription factor binding sites, TFBS, are detected in a candidate sequence by means of the total parametric entropy of the aligned sequences summing for each specific position [89], Fig. 3.1.

Any new candidate sequence added to the training matrix will cause a variation on the order or the information of the set of aligned sequences. The detection of an active site depends on the change of this measure from the aligned sequences if the candidate sequence is assumed to belong to the set of aligned sequences. For random sequences, the disorder observed from the set will increase. For a true binding site, the candidate sequence is not expected to modify in a significant way the total information on the aligned sequence set.

A classical uncertainty measure is the Shannon entropy. This paper employs Rényi entropy for this measurement [57]. Rényi is a parametric entropy which depends on q, namely the order in the Rényi entropy. This parameter modulates the probability of occurrence of each symbol, emphasizing or suppressing this value as q decreases or increases, respectively. This measurement allows us to

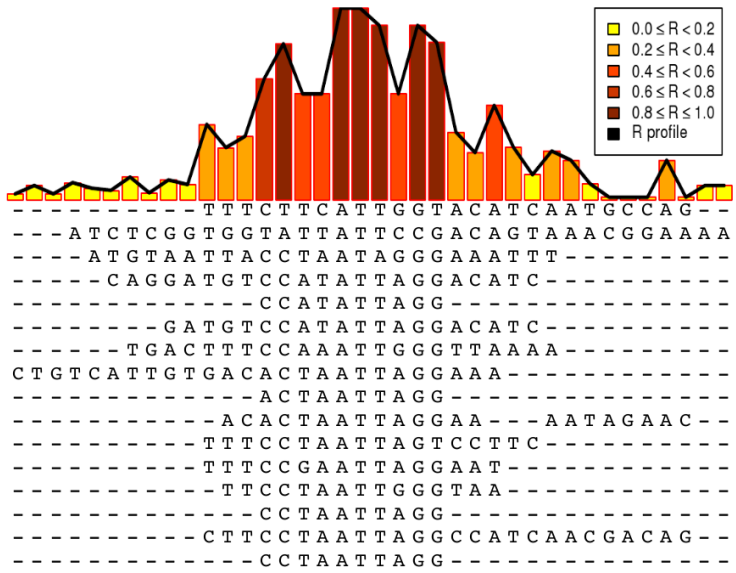


Figure 3.1: Information content in a matrix of aligned sequences as a Redundancy profile.

build a parametric detector with variable sensibility modulated by the Rényi order, q. We define a set of functions in order to measure the variation on the total Rényi entropy from the set of aligned sequences.

Given the small number of sequences available for each transcription factor (see Table 5.1), the detector has been characterized through a leave-one-out cross-validation. Each individual sequence is used as a test sequence of a training classifier with the rest of $n-1$ sequences, where n is the number of sequences. The results have been obtained in contrast with randomly generated candidate sequences. These random sequences have been generated considering the nucleotide frequency statistics of the transcription factor organism. Each random sequence contained 1000 nucleotides. That is done successively for each sequence within the training matrix.

3.4 Information content measures

Shannon [104] defined the entropy of a system as a measure of uncertainty of its structure. Shannon's entropy is based on the concept that the information

Table 3.1: Summary of The Transcription Factors Analyzed

Organism	TF	Bases	Sequences	Binding domain
S. cerevisiae	$MCM1$	38	16	MADs box
S. cerevisiae	$ROX1$	12	20	HMG-box
S. cerevisiae	$ABF1$	37	22	Zinc finger

Table 3.2: Statistics of $H_q^{n_b}$ for no equiprobable genomic composition

n	$E(H_{0.1}^{n_b})$	$sd_{0.1}$	$E(H_{0.5}^{n_b})$	$sd_{0.5}$	$E(H_1^{n_b})$	sd_1	$E(H_2^{n_b})$	sd_2
2	0.730	0.444	0.730	0.444	0.730	0.444	0.730	0.444
3	1.120	0.425	1.101	0.432	1.078	0.439	1.037	0.455
4	1.359	0.389	1.324	0.396	1.282	0.405	1.208	0.417
5	1.517	0.352	1.470	0.359	1.414	0.367	1.319	0.379
10	1.847	0.215	1.775	0.219	1.694	0.234	1.572	0.269
25	1.978	0.050	1.917	0.074	1.852	0.110	1.752	0.161
50	1.989	0.009	1.946	0.037	1.898	0.067	1.819	0.109

gain from an event is inversely related to its probability of occurrence [80]. The Rényi entropy [96] is a parametric entropy measure that can be considered as a generalization of Shannon entropy. The Rényi entropy of a random variable X with N possible states $(X_1, X_2, \cdots, X_i, \cdots, X_N)$, where the probability for each state i, given by p_i such that $\sum_{i=1}^{N} p_i = 1$, is defined as,

$$H_q = \frac{1}{1-q} log_2 \sum_{i=1}^{N} p_i^q \tag{3.1}$$

where, variable X are nucleotides A, T, C and G in each DNA sequence position and the Rényi order q is a positive real number different than 1 (also known as α parameter in [96]). The Rényi entropy is a nonnegative measurement for all $q \geq 0$ and converges to Shannon entropy when q tends to 1.

$$\lim_{q \to 1} H_q = -\sum_{i=1}^{N} p_i log_2 p_i \tag{3.2}$$

A normalized redundancy R can be defined as,

$$R_q = 1 - \frac{H_q}{H_q \mid_{max}} \tag{3.3}$$

where the redundancy is normalized depending on the maximum entropy. This quantity covaries with the information content and is normalized between

0 and 1. Redundancy nulls when all the four bases are having similar representation. In contrast, redundancy takes 1 when there is a complete conservation of a base at that position. For a group of aligned sequences, the measurement of the redundancy gives information about the complexity of the nucleotides distribution in the conserved sequence.

3.5 Database Description

The algorithm requires a group of aligned nucleotide sequences with binding evidence. These sequences are obtained from the organism *Saccharomyces cerevisiae* which was the first eukaryotic organism with its genome completely sequenced. This organism contains around sixteen million of nucleotides distributed among sixteen chromosomes. The following transcription factors have been considered: $MCM1$, $ROX1$ and $ABF1$. A brief summary of the data for each transcription factor (TF) is shown in Table 5.1, where Bases is number of bases in the alignment, Sequences is number of aligned sequences and, finally, the binding domain is the binding structure with binding sites. Each transcription factor shows different structural strategies to interact with binding sites. $MCM1$ presents a quaternary structure acting as a dimer. It is also able to interact with other proteins acting as repressor or activator depending on the complementary elements in the complex [93]. On the other hand, $ROX1$ binds to DNA by means of a high-mobility-group motif (HMG) acting as a repressor of hypoxic genes under normoxic conditions. Many genes are repressed coordinately when oxygen is present allowing the aerobic metabolism. Furthermore, these genes will be activated together if there is a decrease in oxygen concentration allowing fast adaptive response [8, 68]. Finally, $ABF1$, which is directly involved in regulation of genes related with chromatin stability and accessibility (http://www.yeastgenome.org), shows a zinc finger motif. This DNA binding element presents a Zinc atom in the polypeptide chain in order to maintain a tertiary structure able to interact with DNA.

The dataset has been obtained from the TRANSFAC database, version 7.0 Public 2005, [135], `http://www.genregulation.com/pub/databases.html`. An *R* library has been developed for automatic sequence extraction from the database given a transcription factor name [95]. These sequences have been aligned by means of MUSCLE [31], MUltiple Sequence Comparison by Log-Expectation, to obtain an aligned matrix of the sequences and each nucleotide involved in each position. MUSCLE is based on a progressive alignment method split in two stages: a first stage in which MUSCLE performs a pairwise alignment, and a second step in which the multiple alignment is built by adding the sequences sequentially to the growing alignment according to the pairwise alignment. Further information on sequence alignment problems can be found in Morgenstern

et al. [78].

3.6 Correction of the Finite Sample Size Effect

Every aligned matrix of sequences is formed by a finite number of samples. The probability estimation error using the nucleotide frequency causes a bias on the uncertainty measurement [102]. To correct for this, we have precomputed the exact Rényi entropy correction factor for small samples, $n < 50$.

The probability of obtaining a particular combination of n bases, n_b, can be found by means of a multinomial distribution. The information, $H_q^{n_b}$, is calculated and weighted by the probability of the obtained combination.

If n_A, n_T, n_C and n_C are the number of A, T, C and G in a concrete position, and P_A, P_T, P_C and P_G are the frequencies of each base within the genome. Then, the probability to obtain a particular combination of n_a to n_t, called n_b, can be estimated by means of [102]:

$$P^{n_b} = \frac{n!}{n_A! n_T! n_C! n_G!} P_A^{n_A} P_t^{n_T} P_C^{n_C} P_G^{n_G} \tag{3.4}$$

where $n = n_A + n_T + n_C + n_G$. $P_A^{n_A}$, $P_T^{n_T}$, $P_C^{n_C}$ and $P_G^{n_G}$ are the probabilities corresponding to the frequency of each nucleotide in the *Saccharomyces cerevisiae* with values taken from literature [39]. The factorial computes the number of possibilities associated to each combination. From the uncertainty Hn_b, defined in [102], the probability, P^{n_b} of obtaining the parametric uncertainty, $H_q^{n_b}$, can be defined as:

$$H_q^{n_b} = \frac{1}{1-q} log_2 \sum_{b=A}^{G} \left(\frac{n_b}{n}\right)^q \tag{3.5}$$

$$E(H_q^{n_b}) = \sum_{all n_b} P^{n_b} H_q^{n_b} \tag{3.6}$$

The correction for the standard deviation finite sample is then:

$$Var = [\sum_{all n_b} P^{n_b} (H_q^{n_b})^2 - E(H_q^{n_b})^2] \tag{3.7}$$

$$sd = \left[rVar(H_q^{n_b})\right]^{1/2} \tag{3.8}$$

where r is the length of the binding site. If we consider $r = 1$, the finite sample correction for different parameters of Rényi entropy, are shown in the Fig. 3.2 and also in Table 3.2.

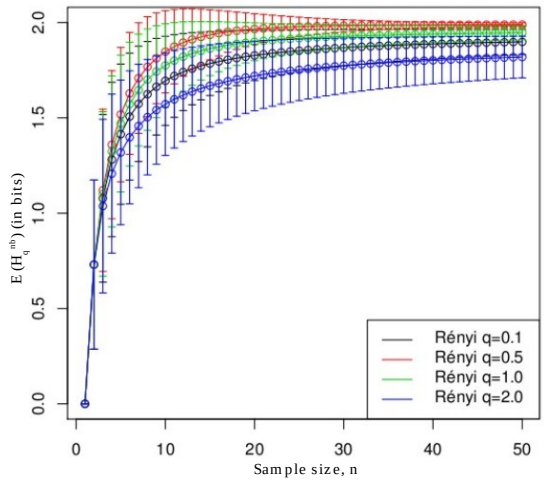


Figure 3.2: $E(H_q^{n_b})$ regarding number of sites, n.

3.7 Motif Detection

Using the matrix of aligned sequences, Rényi entropy is computed for each position on the binding site following the schema shown in Fig. 3.3. The values of the redundancy for each variable position (nonconserved among all sequences) are close to 0. On highly ordered positions, the redundancy has values close to the unity.

The measurement of the variation of the total redundancy when the candidate sequence is added to the set has been computed by using two functions. These functions consider normalized and nonnormalized forms as in eq. (3.9) and (3.10),

$$\omega(q,i,\theta) = \mid \sum_{i=1}^{L} R_q^m \gamma \mid^{-1} \tag{3.9}$$

$$\rho(q,i,\theta) = \left[\frac{\sum_{i=1}^{L} R_q^m \gamma}{\beta} \right]^{-1} \tag{3.10}$$

where , γ and β are

$$\gamma(q,i,\theta) = \mid R_q^m - R_q^s \mid \tag{3.11}$$

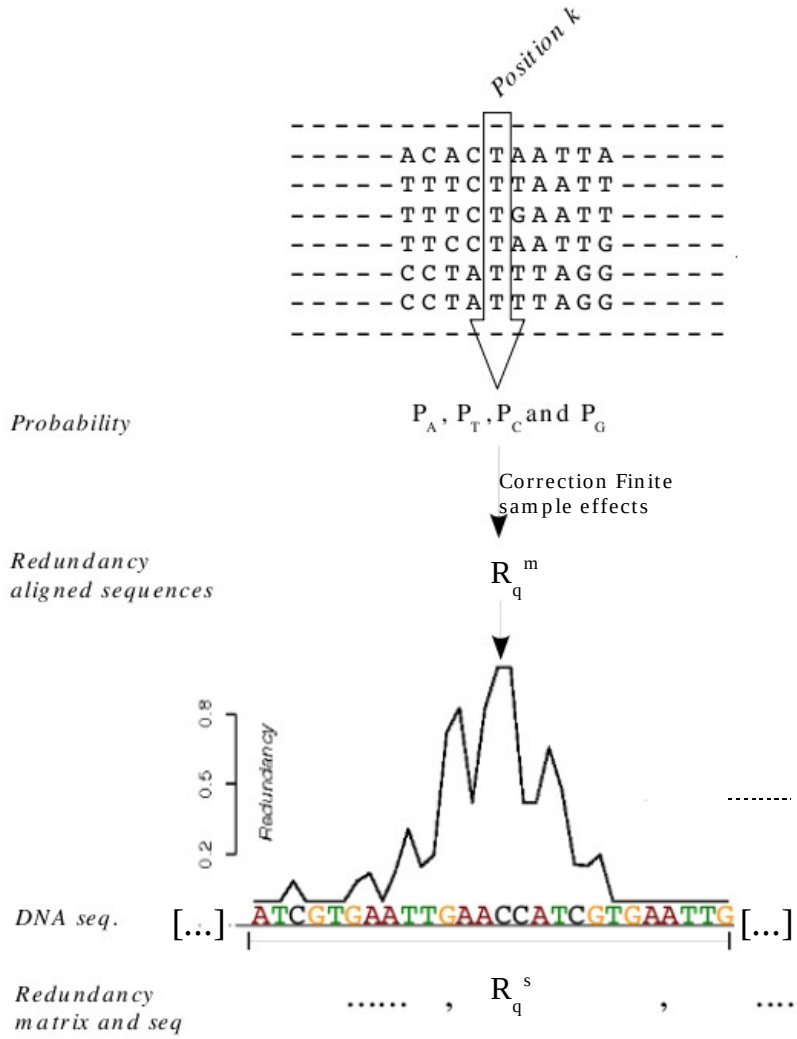


Figure 3.3: Schematic representation of the developed method for the Transcription Factor binding sites detection in random DNA sequence.

$$\beta(q, i, \theta) = R_q^m + R_q^s \tag{3.12}$$

where i is a specific position of the binding site and θ is the aligned set of sequences. The redundancy profile is an L-dimensional vector, where L is the total number of positions of the binding site. The R_q^m measurement determines the Rényi entropy for a position on the set of aligned sequences whereas the R_q^s will contain the equivalent parametric entropy when the candidate sequence is assumed to belong to the set. The variation in the parametric entropy of the aligned sequences is considered by means of the difference between redundancies, γ. For a random sequence, the order of the system decreases, increasing the value of γ. Both expressions (3.9) and (3.10) define indexes which allow for the discrimination between a random sequence and a sequence that belongs to a binding site. The developed method, Fig. 3.3, based on the criterion defined previously, is as follows:

1. For each position within the matrix of aligned sequences, the probability corresponding to each nucleotide is estimated. Missing values are imputed to a multistate nucleotide with probabilities corresponding to the frequency of each nucleotide in the corresponding organism, with the corresponding statistics found in Thakurta et al. [39] .

2. The redundancy profile is calculated from the PWM, correcting finite sample effects as in Schneider [102].

3. Steps 1 an 2 are repeated, adding the candidate sequence to the set.

4. For each redundancy profile variation obtained from the studied sequences, a scalar quantity is computed using the different functions defined in (3.9) and (3.10).

3.8 Results

The redundancy measurement provides information about the symbolic variance observed in a position of the set of aligned sequences. The lower is the symbolic variance, the higher values of redundancy are obtained. In fact, the redundancy gives information on how much a particular position has been conserved on the sequences. In Figure 3.4, we can visualize the variability of each position of the $MCM1$, $ROX1$ and $ABF1$ transcription factor by means of the correspondent redundancy profile for different q-values. The dependence of the entropic profiles with q is also shown. As q increases, the noise in the redundancy also increases. On the other hand, with low q values the redundancy signal also decreases.

Table 3.3: Area Under Convex Surface

	MCM1		ABF1		ROX1	
q	ω	ρ	ω	ρ	ω	ρ
0.1	0.96526	0.97136	0.97582	0.97102	0.99892	0.99934
0.2	0.96737	0.96988	0.97673	0.97031	0.99897	0.99949
0.3	0.97095	0.97375	0.97751	0.97205	0.99899	0.99951
0.4	0.97428	0.98038	0.98032	0.97331	0.99906	0.99954
0.5	0.97781	0.98323	0.98273	0.97556	0.99914	0.99959
0.6	0.98042	0.98478	0.98467	0.98073	0.99917	0.99961
0.7	0.98312	0.98557	0.98658	0.98119	0.99928	0.99965
0.8	0.98591	0.99307	0.98807	**0.98212**	0.99935	0.99968
0.9	0.98785	0.99238	0.98972	0.98085	0.99942	0.99971
1.0	0.98975	0.99289	0.99091	0.97817	0.99945	0.99975
1.1	0.99148	**0.99319**	0.99213	0.97840	0.99949	0.99976
1.2	0.99295	0.99265	0.99282	0.97426	0.99952	0.99977
1.3	0.99415	0.99086	0.99305	0.96304	0.99954	**0.99977**
1.4	0.99503	0.98792	0.99351	0.96071	0.99955	0.99977
1.5	0.99573	0.98335	**0.99354**	0.95524	0.99958	0.99976
1.6	**0.99636**	0.97676	0.99342	0.94968	0.99959	0.99976
1.7	0.85695	0.97236	0.99318	0.94382	**0.99960**	0.99976
1.8	0.87206	0.97283	0.96513	0.88819	0.68167	0.92628
1.9	0.88576	0.94642	0.96466	0.87922	0.68665	0.92806
2.0	0.89408	0.95321	0.96413	0.87017	0.69188	0.92965
MEME	0.99723		0.99898		0.99921	
MDscan	0.97929		0.97547		0.96758	

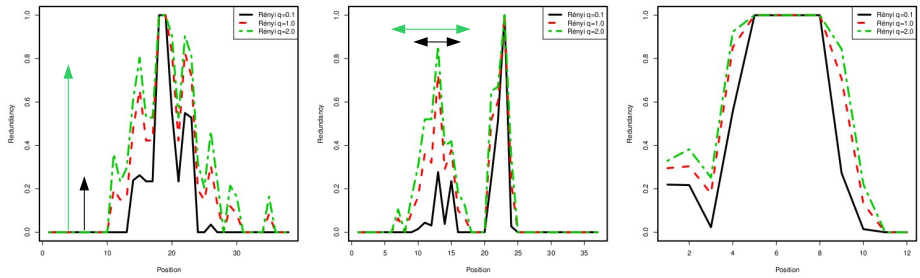


Figure 3.4: Left to right: Redundancy profile for different q-values for the recognizers $MCM1$, $ABF1$ and $ROX1$ of the *Saccharomyces cerevisiae*.

Therefore, the redundancy profile of the transcription factor depends on the Rényi order. An optimal q-value is suggested as a trade-off between the noise included in the redundancy signal and the attenuation of the same one. In summary, the Rényi order modulates the amplitude, Fig. 3.4 (left), and the number of positions that belong to a binding site, Fig. 3.4 (right). This is also interesting in order to evaluate the positions involved to the binding sites once the best q value is found.

The detector proposed in this paper evaluates the perturbation into the total Rényi value to check whether the information is destroyed with addition of the candidate sequence to the set of aligned sequences.The performance of the detector in the case of $ABF1$, $ROX1$ and $MCM1$ as a Receiver Operating Characteristic (ROC) for different functions and given for different values of q is shown on Fig. 3.5, respectively. The best learning system will be the one which produces a larger area under the convex surface, AUC. The performance of the Rényi based detector is compared against two publicly available detectors: MDscan.2004 [69] and MEME [4], version 4.1.0. The default parameters have been used by these algorithms except the witdh of motifs.

In Table 3.3, it can be observed that the detector has a different behaviour depending on the q-value and the function used. Specifically, the number of true and false positives depends on the q-value and the considered function. For example, given a number of true positives, the number of false positives changes according to the functional and the q-value. Considering any of the two functions employed, if q decreases, the number of positions of the transcription factor that we consider decreases, but the number of true positives and false positives increases.

The Rényi order q does depend on the transcription factor binding sites cha-

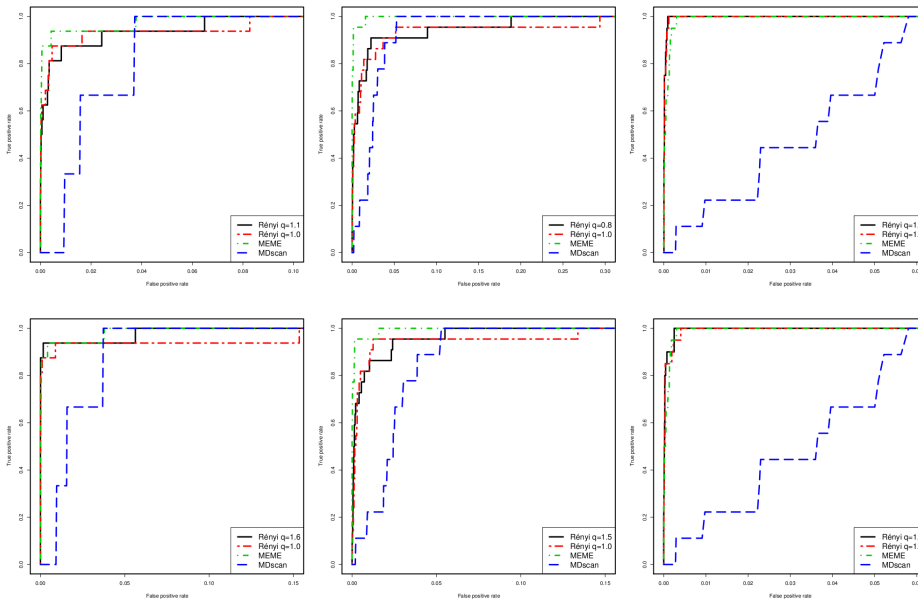


Figure 3.5: Left to right: ROC curve for the different detector in $MCM1$, $ABF1$ and $ROX1$ for ρ (up) and ω (down).

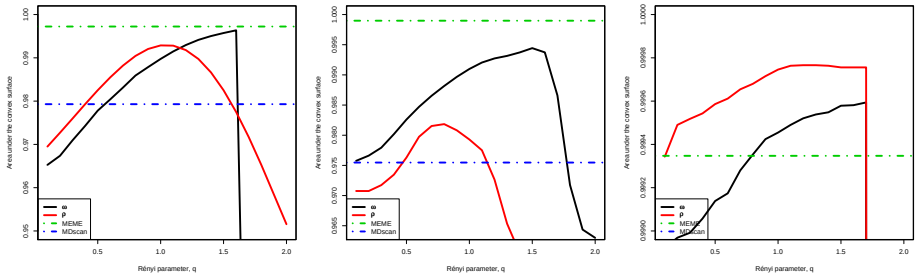


Figure 3.6: Left to right:Area under convex surface versus Rényi parameter for $MCM1$, $ABF1$ and $ROX1$. On the right figure, performance of MDscan falls below the axis.

racteristics. Therefore, a q-value should be adjusted for each training sequence set and for each function considered.

Given a training sequence set and a function as defined in eq. (6.5), the optimal q-value has been estimated by means of leave-one-out cross-validation from q-value set. For each q-value, ROC curve has been calculated. From this results, the optimal q-value has been considered according to the area under convex surface maximum.

Generally, low q values will depress the Redundancy profile, turning ω and ρ more selective, whereas large q values will promote redundancy values. Large q values will show large number of true positives at the cost of introducing additional noise in the ω and ρ, increasing false positives. Therefore, an optimal q-value is the result of a balance between the noise and the attenuation of the redundancy signal and it is obtained using on the cost criteria established, and considering the AUC maximum.

In Fig. 3.6, the area under convex surface is shown as function and the q-value. Establishing an optimal function and q-value allows us not only to parametrize and optimize the detector, but also to define the effective positions in the binding site that play a role in the binding process. The best functional and optimal q-value can be selected for the application given the cost criterion established for miss classifications of True Positives or the area under convex surface maximum.

The performance of MDscan and MEME has also been tested with the same data and the same validation conditions. The Area under convex surface for the different methods is compared in Table 3.3, where the maximum AUC for each transcription factor and function in shown in bold.

The AUC of a classifier is equivalent to the probability that the classifier

will rank a randomly chosen positive instance higher than a randomly chosen negative instance one [34]. This statistic will always be between 0 and 1.0. AUC takes 1.0 when the computational method classify all the sequences correctly and 0.5 when the classifier is random. In Table 3.3, it can be observed that the AUC for parametric entropy measure method is larger than MDscan and that the performance of the proposed detector is similar to the one showed by MEME for this data. In the case of $ROX1$ transcription factor, the proposed detector improves MEME probably because the later is not able to fit a proper statistical model from the poor binding profile shown by $ROX1$.

In summary, the proposed method improves the performance given by MDscan in all cases when the proper q-value has been adjusted through leave-one-out cross-validation. The performance of the detector in comparison to MEME depends on the binding site structure. For binding sites that show no correlation between the positions involved in the binding, the entropy based method outperforms MEME. This is seen in the case of ROX1. On the contrary, $ABF1$ is known to show certain correlation between the positions on the binding. MEME is able to model this correlation between positions, which could explain the difference in performance between MEME and our method in this case. An additional benefit of the proposed method is that with the process of optimizing the q-value we obtain the positions in the profile that are involved in the binding process. Figure 3.4 shows that for the range of the optimal q-value for $MCM1$ binding site positions ($q = 1.6$), more positions around the main peak are included into consideration by the algorithm. The nucleotides included in the binding profile are specific for each transcription factor due to the binding mechanism.

3.9 Conclusions

In this contribution, we have presented a methodology to detect the transcription factor binding sites. This method is based on the variation of the total parametric Rényi entropy in a set of aligned sequences with binding evidence when the candidate sequence is assumed to belong to the aligned set of sequences. The detector employs a parametric entropy, yielding to a parametric detector that depends on the order of the Rényi entropy q. This parametrization provides two main advantages. First, it leverages the strong and weak symbol probabilities when computing the total entropy of the binding sequences, obtaining a detector with variable sensibility. Secondly, through the optimization of the Rényi parameter, an estimation of the positions of the site involved in the binding process is determined. The detector has to carefully consider the finite sample effects for computing the entropies.

This algorithm has been applied on the detection of $ABF1$, $MCM1$ and $ROX1$ recognizers from a random sequence. The obtained results improve bin-

ding site detection based on Shannon entropy. The parametric uncertainty measure gives additional information related to binding site than Shannon entropy. The Rényi order q depends on the transcription factor binding sites characteristics. This parameter is adjusted for each sequence set and for each function considered by means of a cross validation.

The proposed method has shown better performance than MDscan, which is a combined word enumeration and position specific weight matrix in the case of binding site discrimination against random generated sequences. Moreover, the obtained results are comparable with the results of MEME, which is based on the technique of expectation maximization to fit a two-component finite mixture model. Specifically, our method improves the results of MEME for $ROX1$ transcription factor and is comparable with the rest of factors despite assuming independency between positions in the binding sequence.

Chapter 4

A Subspace Method for the Detection of Transcription Factor Binding sites

This chapter is an exact copy of the paper:

- **A subspace method for the detection of transcription factor binding sites**. E. Pairo, J. Maynou, S. Marco, A. Perera: Bioinformatics 28(10):1328-1335(2012).

4.1 Abstract

The identification of the sites at which transcription factors (TF) bind to DNA is an important problem in molecular biology. Many computational methods have been developed for motif finding, most of them based on position-specific scoring matrices (PSSM) which assume the independence of positions within a binding site. However, some experimental and computational studies demonstrate that interdependences within the positions exist. In this paper, we introduce a novel motif finding method which constructs a subspace based on the covariance of numerical DNA sequences. When a candidate sequence is projected into the modelled subspace, a threshold in the Q-residuals confidence allows us to predict whether this sequence is a binding site. Using the TRANSFAC and JASPAR databases, we compared our Q-residuals detector with existing PSSM methods. In most of the studied transcription factor binding sites, the Q-residuals detector performs significantly better and faster than MATCH and MAST. As compared to Motifscan, a method which takes into account interdependences, the

performance of the Q-residuals detector is better when the number of available sequences is small.

4.2 Background

Deoxyribonucleic acid (DNA) sequence motifs are short sequence patterns with biological function. In the gene promoter region, there are DNA sequence motifs which hint at the interaction between the gene regulation machinery and the nucleic acids. They are involved in several DNA and ribonucleic acid (RNA) processes, such as the binding of some proteins to DNA, the ribosome binding to mRNA, and mRNA processing [26]. Protein biosynthesis starts with a transcription process. This process, for example in eukaryotes, is led by several types of RNA polymerase which require special DNA sequences in promoters and a set of transcription factor proteins.

Due to the importance of gene regulation, a major problem in molecular biology is to discover the location of the transcription factor binding sites (TFBS) within the genome. But the fact that most transcription factors bind to short, degenerate sequences makes it difficult to find sequence patterns to model the binding sites [130]. Many algorithms try to characterize these patterns, and such algorithms may be classified into consensus-based methods or alignment-based methods [86].

Most of the algorithms developed target the location of transcription factor binding sites. These follow one of two strategies: (1) to discover common binding sites into a set of unaligned sequences of corregulated genes and (2) to make use of the previous knowledge of sequences to search for a motif within a genome [33, 23, 98, 43].

The algorithms which use the previous knowledge of the binding site sequences are mostly based on Position Specific Scoring Matrices (PSSM) [117]. PSSM are matrices of frequencies of each nucleotide in each position of the binding site. Some examples of these algorithms are MATCH [54], which uses information at each position to construct a PSSM; MAST [6], based on the QFAST algorithm and part of MEME suite [7]; rVISTA [71] which uses evolutionary data; and ITEME [75] which calculates the information loss of the binding sites. These models assume that the positions in binding sites are statistically independent. However, experimental evidence shows that TFBS have interdependences between positions [18] and some computational studies suggest the same [123]. These findings have motivated the development of new strategies which take into account position interdependences. Models based on Markov chains,

such as WAM [137], are restricted to modelling interdependences between adjacent positions. Other algorithms estimate non-adjacent interdependences using permuted Markov models [139]; Bayesian networks [9]; variable order Bayesian networks [10, 19]; or graphs [81]. Detectors constructed using these techniques have higher accuracy, but require the tuning of many parameters for optimal operation which typically requires a large number of binding site instances. Additionally, most of these algorithms are computationally intensive.
On the other hand, a large body of knowledge exists for specific event detection in numerical sequences (signals), and the conversion of symbolical DNA sequences into numerical DNA sequences has been widely used in genomic signal processing to extract relevant biological information from DNA sequences. For example, numerical conversions have been used to identify protein coding regions by studying their periodicity [2, 22, 106].
In this paper, we propose a detector based on the Q-residuals of a numerical sequences covariance model. This contribution aims to study to what extent the covariance can capture information on position interdependences between binding sites. Our hypothesis is that, when projected into the subspace defined by the covariance, sequences belonging to the modelled TFBS should have smaller Q-residuals than chromosomic or random sequences, consequently Q-residuals could be used to detect binding sites. The proposed detector was compared to the PSSM based methods, MAST and MATCH, using real genomic data. It was also compared to the Motifscan method which calculates interdependences between positions.

4.3 Data

TFBS sequences were extracted from the TRANSFAC 7.0 2005 public database [135] and from JASPAR 2010 [92]. For the JASPAR database, the motifs with 10 or more sequences were extracted. To carry out the study, we selected 43 motifs corresponding to *Homo sapiens*, 25 from *Mus musculus*, 11 from *Rattus norvegicus*; a further 10 were randomly chosen from all the TFBS available for *Drosophila melanogaster*. For the TRANSFAC database, the 108 motifs with more than 10 sequences were chosen. These motifs were multiple-aligned using the CLUSTALW2 algorithm [61] with default parameters. The alignment was performed N times, where N is the number of sequences for each motif, using a leave-one-out cross validation (L.O.O.) procedure. The 23 TFBS motifs having a core with more than 5 consecutive positions without gaps at each step of the L.O.O. procedure were used to compare our method to the existing PSSM algorithms. These binding sites correspond to eukaryotic organisms of different level of complexity, ranging from *Saccharomyces cerevisiae* to *Homo sapiens* and

including *Drosophila melanogaster*, *Rattus norvegicus*, *Mus musculus* and *Gallus gallus*. The number of selected sequences from JASPAR totalled 89 motifs. The relation of the 89 JASPAR motifs and the 23 TRANSFAC motifs is given in the supplementary material 2, and a summary of the TF used for each organism can be seen in table 4.1.
All promoter sequences from the organisms used, with the exception of *Saccharomyces cerevisiae*, were extracted from the Eukaryotic promoter database (EPD) sequences [99], using the EPD version based on EMBL release 105 (September, 2010). The sequences located at the positions from -1000 to 500 relative to the Transcription Start Site (TSS) were used to construct the background model, consisting of the nucleotide frequencies for the promoters of each organism. In *Saccharomyces cerevisiae*, the extracted sequences correspond to promoter sequences in chromosome 1 and 16 of the EMBL chromosome database [53], release 94 (March, 2008).
In each organism, we randomly chose two promoter sequences of length 1501 nucleotides for use as background sequences. In *Drosophila melanogaster*, we used the sequences from -1000 to 500 relative to TSS of FAF gene as background 1 and the same range of nucleotides from gene $CG12170$ as background 2. In *Mus musculus*, the same range of nucleotides was set and the $Igk'T$ gene was used as background 1, while gene $Igk'MPC11$ was used as background 2. In *Rattus norvegicus*, background 1 was extracted from the myosin $LC3_fP2$ gene and background 2 from $PSBPC2$. For *Homo sapiens*, the promoter corresponding to background 1 was in the region of the gene $RPS9P2+$ while the promoter corresponding to background 2 was relative to $PSMA2$ TSS. In the study of *Gallus gallus*, background 1 was relative to $apoVLDLII$ TSS and background 2 relative to $a'A-globin$ TSS. Finally, in *Saccharomyces cerevisiae*, the background 1 sequence generally corresponded to positions 44730-46230 in chromosome 1. However, an exception was made for ABF1 binding sites, since ABF1 binding sites are present in that promoter; for ABF1 background 1, the sequence used corresponded to the positions 678930-680430 in this organism's chromosome 16 while, for background 2, the positions from 11410 to 12910 in chromosome 1 were used in all the organism's studied binding sites.

4.4 Preprocessing

The aligned matrix of DNA sequences had to be converted to a rectangular matrix of numerical sequences.

The first step was to translate symbolic DNA to numerical sequences using the conversion process proposed by Silverman et al [108], where each nucleotide

Table 4.1: Information about motifs used for each organism

Organism	JASPAR	TRANSFAC	Total
Saccharomyces cerevisiae	0	7	7
Drosophila melanogaster	10	3	13
Mus musculus	25	4	29
Rattus norvegicus	11	4	15
Homo sapiens	43	4	47
Gallus gallus	0	1	1
TOTAL	89	23	112

is placed at the vertex of a regular tetrahedron as in equation (4.1):

$$\begin{aligned} A &\equiv (0,0,1) \\ C &\equiv (-\frac{\sqrt{2}}{3}, \frac{\sqrt{6}}{3}, -\frac{1}{3}) \\ G &\equiv (-\frac{\sqrt{2}}{3}, -\frac{\sqrt{6}}{3}, -\frac{1}{3}) \\ T &\equiv (2\frac{\sqrt{2}}{3}, 0, -\frac{1}{3}) \end{aligned} \tag{4.1}$$

where A, C, G and T are points in 3-D Euclidian space corresponding to the a, c, g and t nucleotides respectively. This conversion was chosen because it is symmetric for all nucleotides and is widely used in genomic signal processing [67].
After conversion, each DNA sequence of length M became a sequence of length $3 \times M$, concatenating numerical vectors corresponding to each nucleotide. Then, the N sequences belonging to the same transcription factor were arranged in matrix format. The result was an $N \times (3M)$ matrix of numerical DNA.
Where gaps were produced during the alignment process, we imputed the numerical value of these gaps into the mean of the chromosome, taking into account the nucleotide probability distribution of the background organism and the conversion process. The location of the gaps within the tetrahedron is thus given by equation (4.2)

$$GAP = P(a)A + P(c)C + P(g)G + P(t)T \tag{4.2}$$

In this equation, GAP is a three-element vector corresponding to the position of the gap within the tetrahedron; A, C, G, T are the positions of a, c, g,

t nucleotides in the vertexes of the tetrahedron; $P(a), P(c), P(g), P(t)$ are the nucleotides probabilities in the promoter of the organism. Only those positions where the information was available for at least half of the sequences were imputed, the others were neglected.

4.5 Definition of the Subspace Method

A covariance subspace model was computed for each binding motif using a Principal Component Analysis (PCA) of the numerical DNA sequence representation [88]. To carry out the PCA, first the covariance of the numerical DNA matrix was calculated, then the data projected into the subspace where the covariance matrix is diagonal. In this subspace, relatively few components explain most of the covariance, thus reducing the dimensionality of the problem. This yields a bilinear decomposition of the set of aligned DNA sequences as defined in equation (4.3):

$$X = AB^T + E \tag{4.3}$$

where X is a $N \times (3M)$ TFBS numerical matrix, with N being the number of TFBS sequences and M the number of TFBS positions. A is the projected data, consisting of an $N \times nPCS$ matrix called scores, where $nPCS$ is the number of principal components chosen to construct the subspace. B is the $(3M) \times nPCS$ loading matrix which defines the subspace into which data is projected, and E is the $N \times (3M)$ error matrix.

The covariance is a $3M \times 3M$ matrix which captures the covariances between the numerical positions. When it is diagonal, no interdependences exist between positions of a specific binding site. This information is, in our model, explained in those loadings which are almost zero when a position is conserved, and which differ from zero (either in a positive or negative sense) when a position varies. In the supplementary material 1, an example of the covariance matrix and the loadings for the DL binding sites, where covariances exists, is presented.

The detector was built using the Q-residuals, which are the square of the Euclidean distance from a sequence to the subspace generated by the Principal Components model. Given a candidate sequence, the Q-residuals can be calculated using equation (4.4):

$$Q = EE^T \tag{4.4}$$

where E is the $3M$ error vector obtained from projecting the sequence into the Principal Components subspace, and Q is the Q-residual of the candidate sequence.

The model should explain most of the variance and, as outlined above, sequences belonging to the studied TF should have smaller Q-residuals than the other sequences. Defining a threshold in Q-residuals should be sufficient to allow distinguishing between TFBS and sequences not belonging to the modelled TFBS. The threshold chosen is based on the Q-residuals statistics [51], resulting in a confidence interval for a sequence belonging to our model. The Q-residuals distribution corresponding to the modelled TFBS sequences are first converted into a new N(0,1) quantity C (i.e. C is normally distributed with mean $\mu = 0$ and variance $\sigma = 1$). The quantile with the desired confidence interval can be then calculated from this normal distribution. The constructed detector depends on the number of principal components chosen.

4.6 Comparison to PSSM algorithms

To compare our detector to existing PSSM methods, the MEET R package (available in the R-forge project `http://r-forge.r-project.org/projects/meet`), was developed [84]. This R package allows us to combine several alignment methods with different algorithms to search for TFBS within a large sequence. The package can be configured to call external alignment methods including CLUSTALW2, MUSCLE [31], and MEME which has as an internal multiple alignment method. The proposed Q-residuals method is compared both with MAST and with an implementation of the MATCH algorithm which takes into account the probability distribution of the nucleotides in the promoter sequences of each organism.

To implement MATCH, the algorithm explained in [54] was used, however the background nucleotide probability distribution specific for each organism was also used. To detect a motif, first the PSSM matrix was calculated. Then, using this matrix, the information of each position was calculated as in equation (4.5).

$$I(i) = \sum_{B=A,C,G,T} f_{i,B} ln(\frac{f_{i,B}}{P_B}) \tag{4.5}$$

where $I(i)$ is the information of position i, $f_{i,B}$ is the frequency of the B nucleotide in this position and P_B is the background probability of the B nucleotide. The Score of a sequence of length M was calculated as in equation

(4.6).

$$Score = \sum_{i=1}^{M} I(i) f_{i,b_i} \tag{4.6}$$

where f_{i,b_i} is the frequency of the corresponding b_i nucleotide for the sequence in position i and $I(i)$ is the information in the same position. Finally a Similarity Score for the sequence and the core (first five consecutive more conserved positions), as explained in equation (4.7) was used to discriminate between TFBS and other sequences as in the MATCH program (publicly available in TRANSFAC 7.0).

$$SimilarityScore = \frac{Score - Min}{Max - Min} \tag{4.7}$$

Max and Min being the maximum and minimum possible scores for a candidate sequence.
Comparison with the MAST algorithm was done using the downloadable MEME 4.4.0 source available at the MEME suite - this allowed us to combine different alignment algorithms to construct the PSSM and then use the PSSM as an input to MAST. To calculate the PCA model and the Q-residuals in R, the pcaMethods R package was used [114].
CLUSTALW2 with the default parameters, $gapextend = 0.2$, $gapopen = 10$ was used to align the sequences in all the methods compared in TRANSFAC.

4.7 Validation

The MEET R package performs a double L.O.O to calculate the ROC curves, the Area under ROC curve (AUC), and the errors associated with them. Given a motif of N sequences, first a sequence A is removed and inserted into the background sequence. Then, the remaining $N - 1$ sequences of the same motif are used for a standard L.O.O to construct models with $N - 2$ sequences. These $N - 2$ sequences are first aligned and the chosen algorithm is applied to build a model. Finally, each one of the $N - 1$ models of the L.O.O. is used to detect the sequence A within the known position of the background. After that, sequence A is again inserted into the group and another sequence B is used, this whole process being repeated N times. As the location of the true positives is known, the threshold of the detectors can be varied in order to generate the N different ROC curves and AUCs. Thresholding is detector-specific; for Q-residuals it is the residuals statistics of the PCA model, for MATCH it is the sequence similarity and for MAST it is the p-value. Once the N ROC curves are generated, the

standard deviation is used to estimate the variability of the ROC curve points and the AUC.
In the case of the Q-residuals detector, AUC was calculated for from 1 to 10 principal components; in the case of MATCH, the varying parameter was the Core Similarity, ranging from 0.5 to 0.95 in increments of 0.05. Only one set of ROC curves and AUCs was calculated in MAST because the length of the sequence (the parameter to optimize in MEME) is defined by the number of positions of the PSSM constructed using the aligned sequences.
The mean and the variance of AUC for the studied range of principal components were calculated for each motif. Models built using different numbers of principal components can have an equivalent performance when the AUC mean and the AUC variance are taken into account. Between these models, the one with smallest AUC variance averaged between backgrounds 1 and 2 was chosen as the best model. The same criterion was used to choose the threshold of Core Similarity in the MATCH algorithm.

As the number of negative examples greatly exceeded the number of positive examples in this study, it was also convenient to compare the algorithms using Precision-Recall (PR) curves. There exists a unique correspondence between the PR curves and the ROC curves, and when an algorithm dominates in the ROC spaces it also dominates in the PR space, however optimizing the AUC under the two different methods is not the same thing [24]. To show that the PR curves confirm the results obtained with the ROC curves, we calculated the curves with the optimal parameters for each detector (supplementary material 3). The ROC curves, the AUC and the PR curves were calculated using the ROCR package [110].

4.8 Interdependences between positions

The improvement in detection of Q-residuals should be linked to the interdependences between positions in each binding site. To study this relation, the mutual information $MI_{i,j}$ between positions i and j of the binding sites was calculated using equation (4.8):

$$MI_{i,j} = \sum_{b_i,b_j} P_{b_i,b_j,i,j} log_2 \frac{P_{b_i,b_j,i,j}}{P_{b_i,i}, P_{b_j,j}} \tag{4.8}$$

where b_i and b_j correspond to the nucleotides in the studied positions i, j and P_{b_i} is the probability of the b_i nucleotide in the position i. The joint probability of having nucleotide b_i in position i and b_j in position j is described by

P_{b_i,b_j}. The Bayes Factor (BF) described in equation (4.9) was used to test the Null hypothesis, H_0, of independence between positions i and j against H_1, the alternative hypothesis of dependence, in order to determine the significance of the dependencies found:

$$BF(H_0;\ H_1) = \frac{\Gamma(\sum_{b_i,b_j} \alpha_{b_i,b_j}}{\Gamma(M + \sum_{b_i,b_j} \alpha_{b_i,b_j}} \prod_{b_i} \frac{\Gamma(N(b_i,i) + \alpha_{b_i})}{\Gamma(\alpha_{b_i})}$$
$$\prod_{b_j} \frac{\Gamma(N(b_j,j) + \alpha_{b_j})}{\Gamma(\alpha_{b_j})} \prod_{b_i,b_j} \frac{\Gamma(\alpha_{b_i,b_j})}{\Gamma(N(b_i,b_j,i,j) + \alpha_{b_i,b_j})} \tag{4.9}$$

where M is the size of the bindings sites sequences, $N_{b_i,i}$ is the number of b_i nucleotides in position i, and α refers to the parameter of the Dirichlet prior distribution. This measure was used in previous studies to show which positions of a transcription factor have interdependences [140, 123]. When $\alpha_{b_i,b_j} = 1$ and $\alpha_{b_i} = \sum_{b_j} \alpha_{b_i,b_j}$ the Bayes Factor is related to the mutual information as shown in equation (4.10) [77].

$$log_2(BF(H_0;\ H_1)) \approx -MMI_{i,j} \tag{4.10}$$

Formula (4.10), where $MI_{i,j}$ is the mutual information and M the size of the binding sites, was used to calculate the Bayes Factor, $BF(H_0;\ H_1)$. And as in [123], a threshold of $BF < 0.1$ was set to indicate strong evidence of interdependences between positions. For each motif, the percentage of positions showing interdependences, I_{dep}, was calculated.

4.9 Comparison to Motifscan

Naughton et al [81], used 94 JASPAR (2006) motifs to compare Motifscan, a graph-based method which takes into account interdependences, to PSSM methods. To do the comparison, they calculated the ROC_N curves, where N is the number of sequences for the selected motif, and its AUC.
Using the same methodology and 93 of the 94 motifs of the old JASPAR version (the old version of the remaining one was not available), the AUC of the ROC_N curves was calculated for the Q-residuals detector, and the results were used to compare the detectors. The comparison between Motifscan and Q-residuals using the 93 JASPAR motifs is available as supplementary material 4.

4.10 Results

In this section, we first present the results of the comparison between the Q-residuals detector, MATCH and MAST using the 112 motifs presented above and two different backgrounds for each organism. Then, we describe in more detail the comparison between MAST and Q-residuals, and show a study of the interdependences. We present an analysis of the computational time needed for each one of the studied detection algorithms, and finally we compare the Q-residuals detector to the results obtained with Motifscan.

One example of detection can be seen in the cMyB motif in figure 4.1, a set of transcription factor binding sites for *Homo sapiens*. The ROC curves show the performance of the three algorithms using the first background for *Homo sapiens*. A significant improvement is observed when the Q-residuals detector is used in place of MAST or MATCH.

To visualize the performance of the three different detectors in all the studied transcription factors, Table 4.2 summarizes the results for Q-residuals, MATCH and MAST for the two different backgrounds in each organism for TRANSFAC. The best number of components (usually between 1 and 4) is shown, together with the mean AUC for each background and method. The results for all the studied transcription factors are available as supplementary material 2.

To quantify the differences in performance between the Q-residuals detector and the other algorithms, a Wilcoxon rank-test [134] was performed on the AUC distributions, using the null hypothesis that the two distributions are the same versus the alternative hypothesis that AUC using Q-residuals is closer to 1 than when MAST or MATCH are used. In table 4.2 and the supplementary material 2, the increment in AUC and the significance of the test are displayed, and it can be seen that Q-residuals performs significantly better than Match in 57 of the 112 motifs studied and significantly better than MAST in 63 of them, with $p - value < 0.05$.

To better visualize the detectors, we present the AUC box plots in figure 4.2. These box plots represent the AUC and its variation when the leave-one-out cross validation is applied. Figure 4.2 shows the box-plots for the first background and the JASPAR motifs corresponding to *Mus musculus*. In most cases, not only is the mean AUC closer to one in Q-residuals, but the variance is also smaller, which suggests that the Q-residuals algorithm behaves more robustly

An average of the PR curves obtained in each leave-one-out iteration is also

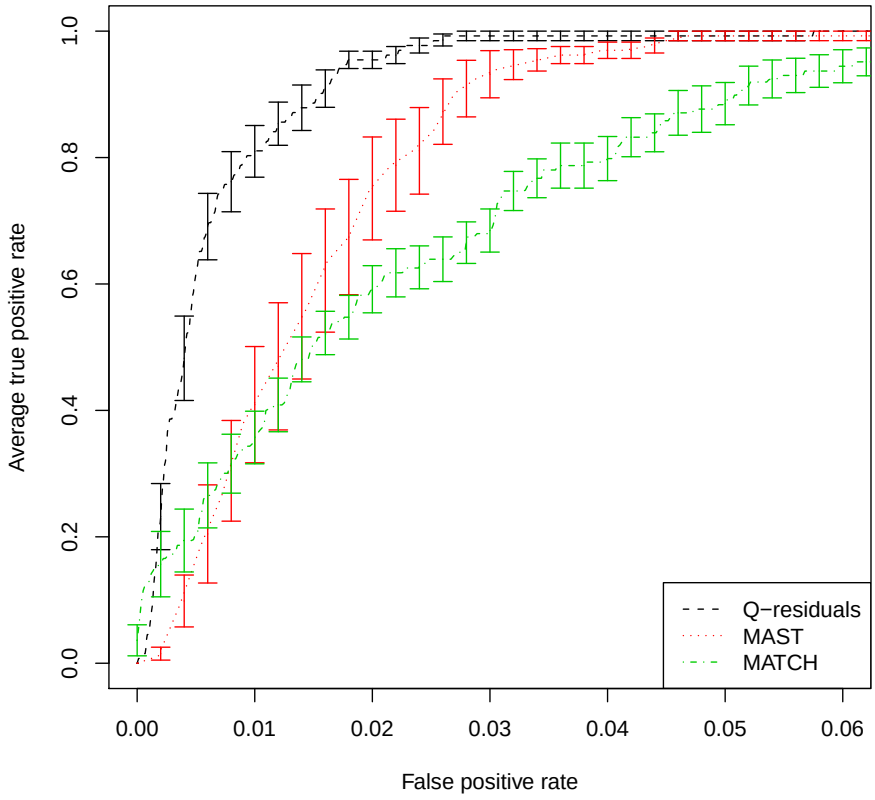


Figure 4.1: ROC curve for Q-residuals in black, MAST in red and MATCH in green using the cMyB transcription factor and the Homo Sapiens background 1. The ideal number of components and the ideal MATCH Core Similarity were used to compute the ROC curve. The error bars correspond to the variation in detection using the L.O.O cross validation. The figure shows the improvement of detection using Q-residuals

presented as supplementary material 3, showing that, when these curves are used, the Q-residuals detector also performs better than the PSSM algorithms in most of the cases.

The percentage of positions showing interdependences, *Idep*, varies among

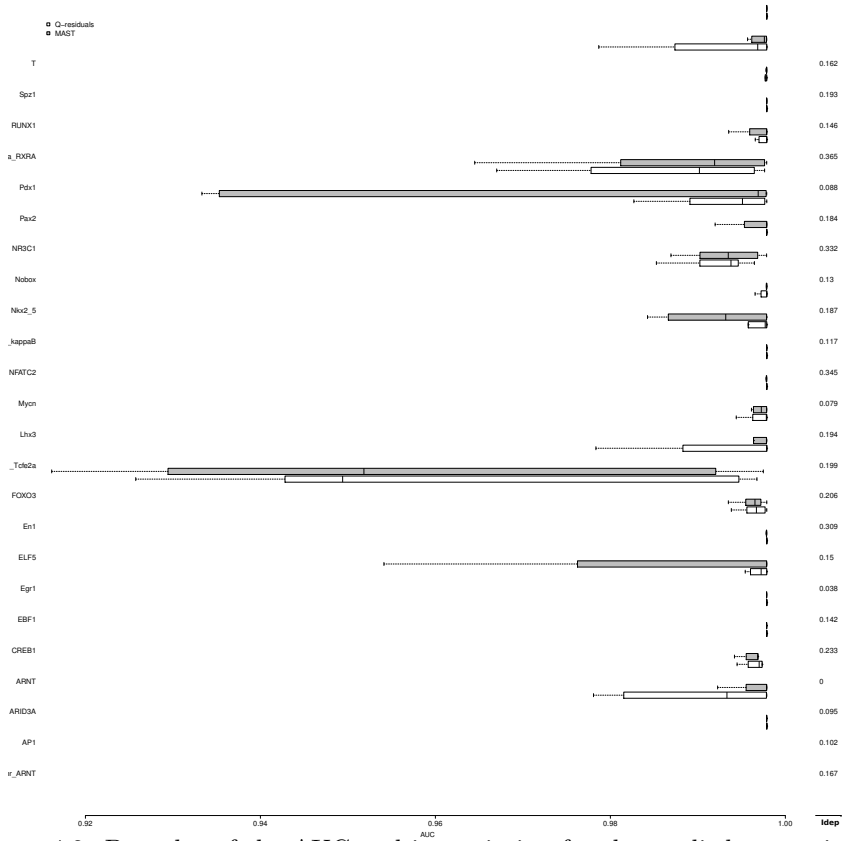


Figure 4.2: Box plot of the AUC and its variation for the studied transcription factors, comparing the Q-residuals detector with the chosen number of components (in white) to MAST (in gray). The results correspond to the background 1 of each organism. Idep corresponds to the rate of positions within a binding site which have significant interdependences.

the studied binding sites as can be observed in figure 4.2. A correlation test was performed between the *Idep* and the improvement in binding site detection when Q-residuals detector was compared to MAST. The improvement in binding site detection was derived by subtracting the mean AUC for each binding site calculated using each method. Results show a significant correlation between the number of strong interdependent sites within a binding locus and the amount of improvement of the Q-residuals detector over MAST (as measured in terms of AUC). Performing the test on the results for JASPAR database gave a $p - value = 0.004$; the corresponding result for the TRANSFAC database was a

$p - value = 0.04$.

The computational times of the Q-residuals detector, of MAST and of our R implementation of MATCH were compared for the detection of TFBS within promoter sequences. To compare the three algorithms, the MAST algorithm (MEME version 4.4.0), the C code for Q-residuals using the ideal number of components, and our implementation of MATCH algorithm in R with the ideal Core Similarity were used. The background corresponded to background 1 for each organism - this consisted of 1500 nucleotides. The threshold for each method was set in such a way that the number of positives was similar. In the case of MAST a p-value of p=0.001 was chosen, for Q-residuals a confidence interval of C=0.95 was set, and for MATCH the Similarity was set to S=0.85. The time was calculated for 100 iterations of the program. The average computational time in detection for the TRANSFAC database motifs are $0.003 \pm 0.001s$ using the Q-residuals detector, $0.0191 \pm 0.001s$ using MAST and $0.33 \pm 0.03s$ for the R implementation of MATCH. The results show that the Q-residuals detector is faster than MAST and the R implementation of MATCH in all the studied binding sites.

The Q-residuals detector was also compared to Motifscan, an algorithm which takes into account interdependences. Using the same criteria as [81], a 5% increase in the ROC_N AUC was required for an improvement to be considered significant. The results showed that in 34 of the 93 studied motifs Motifscan performs better than either the Q-residuals detector or the PSSMs methods, that Q-residuals is the best detector in 25 of the 93 motifs while PSSM is best in just 1 of them. The three detectors perform equally well in 16 motifs; Q-residuals and Motifscan are equally good and better than PSSM in 16 motifs; Q-residuals and PSSM are better than Motifscan in 3 motifs; and Motifscan and PSSM are better than Q-residuals in 9 of the 93 motifs. The AUC performance is shown in the supplementary material 4. A visualization of the results in figure 4.3 shows that the performance of Q-residuals is more sensitive to the number of positions. When the sequences are short, the number of false positives using the Q-residuals detector increases, leading to a smaller AUC. Motifscan performs better in this situation but, on the other hand, it needs more training sequences, so when the number of sequences is small, Q-residuals performs better than Motifscan. Focussing on the 37 motifs which have less than 20 sequences available, in 43.24% of the cases the AUC of Q-residuals shows it to be significantly the best algorithm, while Motifscan is best in just 27.02% of the instances.

In most cases, even if Motifscan is significantly better than Q-residuals, the Q-residuals algorithm performs better than PSSM methods for this comparison also.

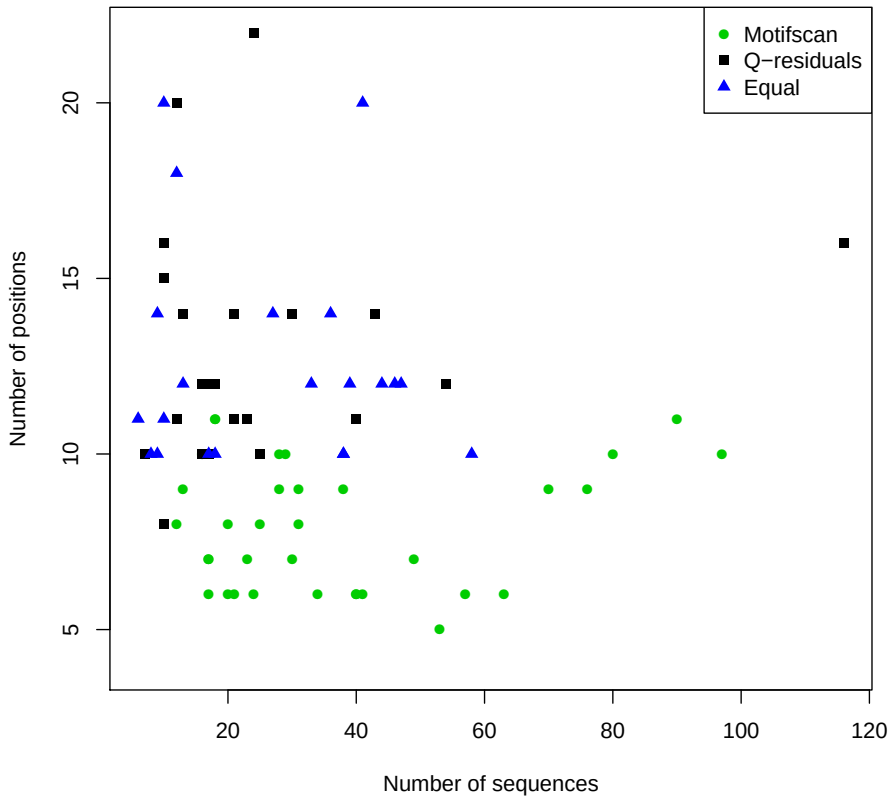


Figure 4.3: Number of positions and number of sequences of the motifs where Motifscan was the best algorithm, green point; where Q-residuals was the best algorithm, black box; or where both perform equally well (less than 5% difference in AUC) in blue triangle. Q-residuals performs better for small number of sequences, but performs worse when the number of position per sequence is small.

4.11 Conclusions

Calculating the residuals of the covariance model of the numerical TFBS has been demonstrated to be an effective method of detecting TFBS within real data, with better performance than existing MEME and MATCH methods.
The results show that, when there are no interdependences, our method is at least as good as the PSSM methods we used for comparison, but we also found a correlation between the improvement in AUC and the percentage of positions showing interdependences in a transcription factor. This result proves that covariance can capture position interdependences in TFBS, and that a covariance-based model can be useful in detecting TFBS within large databases.

When we compared the computational time of the Q-residuals detector and PSSM based methods, we found that Q-residuals is faster; in contrast, other methods which take into account interdependences usually carry a high computational cost. Another advantage of the Q-residuals detector, as compared to methods which take into account position interdependences, is that Q-residuals does not need a large amount of data in order to build a reliable detector.

The ideal number of components was chosen following a robustness criterion, biasing sequence background independence. Usually the number of components which satisfies the above condition is small, models having between 1 and 4 components explain most of the variance of the motif. Differences in detection using a range of components are not always significant.
As compared to a method which takes into account interdependences, Q-residuals shows a significant performance improvement when the number of sequences is small, but it also shows a larger sensitivity to the number of positions. Q-residuals needs more positions than Motifscan or PSSM to decrease the number of false positives.

Table 4.2: Results for Q-residuals detector compared to MATCH and MAST algorithms, corresponding to the 2 backgrounds of each organism in TRANSFAC. The AUC shown for each method is the mean of the areas using the cross-validation method and the number of principal components for Q-residuals is chosen as the number of components with less variance in the AUC. The ΔAUC is the mean AUC improvement of Q-residuals versus MATCH and MAST, respectively. The level of significance corresponds to the p-value calculated when a Wilcoxon-rank test is performed, with the null hypothesis being that the AUC distributions using Q-residuals detector and the other algorithm are the same and the alternative hypothesis being that the AUC distributions calculated with the Q-residuals detector is closer to one. A description of the 89 JASPAR motifs and 23 TRANSFAC motifs can be found in the supplementary material 2.

TF	nPCs	Q-residuals 1	Q-residuals 2	Match 1	Match 2	ΔAUC Match [1]	MAST 1	MAST 2	ΔAUC MAST [1]
ABF1	4	0.9991	0.9975	0.9902	0.9964	$5 \cdot 10^{-3}$ ***	0.9957	0.9986	$1.14 \cdot 10^{-3}$
BCD	3	0.9961	0.9952	0.9912	0.9884	$5.85 \cdot 10^{-3}$***	0.9913	0.9947	$2.68 \cdot 10^{-3}$*
CAT8	3	0.9998	0.9995	0.9971	0.9978	$2.21 \cdot 10^{-3}$***	0.9999	0.9992	$9.02 \cdot 10^{-5}$
CEBP β 35	3	0.9931	0.9965	0.9863	0.9878	$7.75 \cdot 10^{-3}$ **	0.9936	0.9946	$6.66 \cdot 10^{-4}$
cJun	1	0.9868	0.9915	0.9700	0.9813	$1.35 \cdot 10^{-2}$ **	0.9575	0.9880	$1.64 \cdot 10^{-2}$*
cMyB	1	0.9905	0.9907	0.9714	0.9714	$1.92 \cdot 10^{-2}$***	0.9818	0.9869	$6.21 \cdot 10^{-3}$*
DL	1	0.9982	0.9962	0.9835	0.9864	$1.23 \cdot 10^{-2}$ ***	0.9682	0.9917	$1.73 \cdot 10^{-2}$*
E2F	4	0.9997	0.9998	0.9991	0.9998	$3.00 \cdot 10^{-4}$ *	0.9988	0.9995	$5.26 \cdot 10^{-4}$
GAL4	1	0.9998	0.9999	0.9742	0.9759	$2.48 \cdot 10^{-2}$ ***	0.9875	0.9653	$2.34 \cdot 10^{-2}$*
GCN4	1	0.9988	0.9997	0.9936	0.9937	$5.68 \cdot 10^{-3}$ ***	0.9951	0.9935	$5.06 \cdot 10^{-3}$***
HNF1 α	9	0.9945	0.9940	0.9807	0.9850	$1.14 \cdot 10^{-2}$ *	0.9943	0.9921	$2.1 \cdot 10^{-3}$
HNF4 α	4	0.9957	0.9972	0.9870	0.9938	$6.05 \cdot 10^{-3}$ *	0.9937	0.9957	$1.79 \cdot 10^{-3}$
HNF6 α	1	0.9977	0.9996	0.9961	0.99358	$3.81 \cdot 10^{-3}$***	0.9838	0.9949	$9.37 \cdot 10^{-3}$*
IRF1	2	0.9992	0.9994	0.9727	0.9912	$1.74 \cdot 10^{-2}$**	0.9970	0.9992	$1.22 \cdot 10^{-3}$
IRF8	3	0.9991	0.9981	0.9926	0.9791	$1.28 \cdot 10^{-2}$ ***	0.9928	0.9967	$3.86 \cdot 10^{-3}$**
KR	3	0.9923	0.9965	0.9933	0.9838	$5.85 \cdot 10^{-3}$ *	0.9926	0.9929	$1.69 \cdot 10^{-3}$
LyF1	3	0.9952	0.9958	0.9689	0.9823	$1.99 \cdot 10^{-2}$***	0.9903	0.9853	$7.68 \cdot 10^{-3}$**
MIG1	1	0.9986	0.9954	0.9766	0.9475	$3.49 \cdot 10^{-2}$ ***	0.9895	0.9896	$7.49 \cdot 10^{-3}$*
NF κ B	2	0.9998	0.9999	0.9995	0.9999	$3.08 \cdot 10^{-4}$*	0.9991	0.9998	$4.38 \cdot 10^{-4}$ ***
p50	2	0.9996	0.9999	0.9995	0.9999	$4.86 \cdot 10^{-5}$	0.9994	0.9998	$1.72 \cdot 10^{-4}$ *
RFX1	7	0.9921	0.9969	0.9721	0.9867	$1.51 \cdot 10^{-2}$ ***	0.9871	0.9837	$9.09 \cdot 10^{-3}$*
ROX1	8	0.9998	0.9985	0.9997	0.9993	$-3.5 \cdot 10^{-4}$	0.9996	0.9980	$3.40 \cdot 10^{-3}$*
T3R α	6	0.9923	0.9919	0.9754	0.9852	$1.18 \cdot 10^{-2}$***	0.9854	0.9757	$1.15 \cdot 10^{-2}$**

Chapter 5

Computational Detection on cis-regulatory sequences through α-Divergence Analysis

5.1 Background

The information theory has been applied in genetics for the visualization of the information of a set of sequences and its characterization with entropies [101, 75]. Previous contributions have explored the use of an information gain method in order to detect binding sequences [75]. This contribution assumes no correlation between binding site positions. We propose an information theoretic methodology to binding site detection that measures the correlation among binding sites base positions through α-Divergence. The performance of the parametric divergence measure based detector is compared against MEME/MAST and Rényi algorithm (first order).

5.2 Method

The method is based on the idea that total information content in a set of objects can be computed by means of divergence measurements. When a new object is added to set, the total information will change according to similitude between of the new object to the set.If the new object is similar to the set, the total

information variation is not significant. On the other hand, if this new object is different to the set, the total information will increase.

The set of objects is a matrix of aligned sequences with binding evidence. Hence, we construct a detector based on the total information variation of a set of sequences when a candidate sequences is added to the set. The information of a set of sequences is measured by means of α-Divergence which considers dependence between binding site positions. Any candidate sequence added to the training matrix will cause a variation on the information and the correlation between binding sites of the set of aligned sequences. For random sequences, the correlation between binding site positions in the system will decrease. For a true binding site, the variation on the correlation between binding site positions will be not significant on the aligned sequence set.

A classical divergence measure is the Kullback-Leibler divergence [58]. This algortihm employs Rényi Divergence, known as α-Divergence, for this measurement. Rényi Divergence is a parametric divergence which depends on q (or q), namely the order in the Rényi entropy. The joint probability occurrence of each couple of symbols is modulated, emphasizing or suppresing this value, according to q-value[75]. If q-value decreases, the probability of occurence of each couple of symbols increases. On the contrary, If q-value increases, the probability decreases. Hence, a parametric detector that considers based on K_2 dependency model, can be built through this measurement. Moreover, detector's sensibility is moduled by q Rényi order.

For each transcription factor (see Table 5.1), the number of sequences available in the dataset is small. Hence, the detector has been characterized by means of leave one out cross validation (LOOCV). Each individual sequence is used as a test sequence of a training classifier with the rest $n-1$ sequences, where n is the number of sequences. Each new set of training sequences is relined up with Multiple Sequence Alignment (MSA). The results have been calculated with genomic sequence of the organism eukariotics considered (see Table 5.2).

5.3 Information content measures

The Kullback-Leibler (KL) divergence is a measure in statistics that quantifies, in bits, the proximity of two probability distributions P and Q [58]. The Rényi divergence is a parametric divergence measure that can be considered as a generalization of Kullback-Leibler divergence. This divergence is also known as the α-Divergence. Rényi divergence of order q for two discrete variables, X and Y, with N possible states $(X_1, X_2, \cdots, X_i, \cdots, X_N)$ and $(Y_1, Y_2, \cdots, Y_i, \cdots, Y_N)$, is defined as,

$$D_q(X;\ Y) = \frac{1}{q-1} log_2 \sum_{i=1}^{N} P_i^q Q_i^{1-q} \tag{5.1}$$

where, variables X and Y are nucleotides in two different positions. The Rényi divergences are non-negative for all $q > 0$ and converges to Kullback-Leibler divergence when q tends to 1.

$$\lim_{q \to 1} D_q(X;\ Y) = \sum_{N} \sum_{N} P_i log_2 \left(\frac{P_i}{Q_i} \right) \tag{5.2}$$

In our case, $P = p(X, Y)$ and $Q = p(X)p(Y)$, the Kullback-Leibler divergence is the mutual information which is a quantity that measures the mutual dependence of two variables,

$$I(X;\ Y) = \sum_{N} \sum_{N} p(X, Y) log_2 \left(\frac{p(X, Y)}{p(X)p(Y)} \right) \tag{5.3}$$

$$= H(X) + H(Y) - H(X, Y) \tag{5.4}$$

where $H(X)$ and $H(Y)$ are the marginal entropies, and $H(X, Y)$ is the joint entropy of X and Y. The mutual information measure is symmetric and non-negative. $I(X;\ Y) = 0$ holds if and only if two variables (X, Y) are statiscally independent under no finite sample effect.

5.4 Database Description

A set of aligned nucletotides sequences with binding evidence are required by α-Divergence algorithm. These sequences come from different organisms eukariotic (see Table 5.1). One transcription factor binding site has been considered for each organism. Each transcription factor is characterized by its structure and its strategies to interact with cis-regulatory sequences, Table 5.1. The dataset has been obtained from Jaspar [128], `http://jaspar.genereg.net/`. The results have been computed from genomic sequences which have been obtained from Eukaryotic Promoter Database (EPD) [99], Table 5.2.

5.5 Motif Detection

The measurement of the correlation between binding site positions is computed by means of α-Divergence from the matrix of aligned sequences. The studied

Table 5.1: Summary of The recognizers Analyzed

Organism	Recognized	Base	Aligned Sequences
Mus musculus	*Mycn*	31	6
Rattus norvegicus	*CREB1*	12	16
Drosophila melanogaster	*VIS*	34	6
Homo sapiens	*ELK1*	28	16

Table 5.2: Summary of Background Sequences

Recognized	Background Sequences	Range
Mycn	$EP07119(+)MmIgk'MPC11$	$(-1000, 500)$
CREB1	$EP24038(+)RnmyosinLC3_fP2$	$(-1000, 500)$
VIS	$EP17014(+)DmsnRNAU1$	$(-1000, 500)$
ELK1	$EP74078(+)HsRPS9P2+$	$(-1000, 500)$

has been done from q values between 0 and 2. When q tends to 1, α-Divergence converges to mutual information. In this case, for high correlation between positions, the values of mutual information are close to $H_i($ site position $i)$. On the other hand, the mutual information close to zero when the positions are non-correlated. When q tends to 2, α-Divergence converges to X^2-Divergence [59]. From this rational, a function has been used to evaluation of the variation of the α-Divergence between the training matrix and the training matrix when the candidate sequence is added to the set. This function is defined as,

$$\eta = [\gamma * (R * R_t)^{1/2}]^{-1};\ \gamma =\mid D_q - D_q^s \mid \tag{5.5}$$

where, D_q is the α-Divergence matrix of the set of aligned sequences, D_q^s is the α-Divergence matrix considering the training matrix with the candidate sequence, R is the redundancy of the set of aligned sequence and R_t is the transpose redundancy. The variation of the information matrix when adding the candidate sequence has been considered by the variation produced in the total cross-site α-Divergence, Fig 5.1. For a true binding site, the dependence between site positions will be equal. Hence, γ will be about 0 because the binding sequence does not modify the aligned sequence set information. On the other hand, when the sequence is a random sequence, the dependence between positions will decrease and γ will increase. We can define a detector that allows for the discrimination between a random sequence and a sequence that belongs to a binding site. Hence, the developed algorithm, based on the criterion defined previously, is as follows:

1. A preliminary study about the significant dependencies between binding

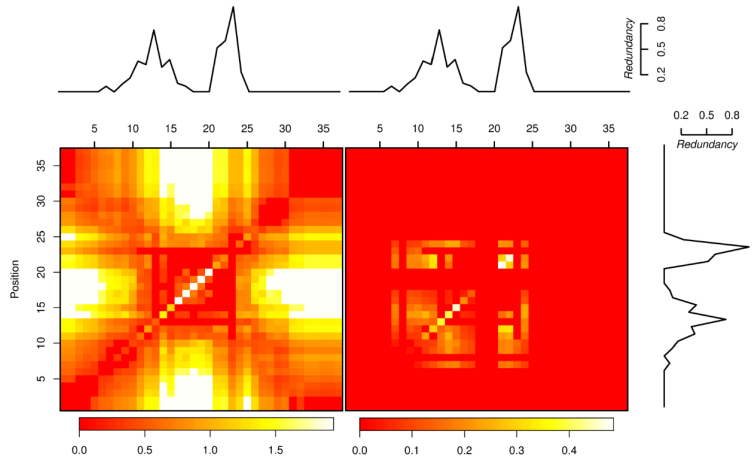


Figure 5.1: (left) Mutual Information heatmap between binding site positions for ABF1. Redundancy is plotted on top; (right) Product between mutual Information matrix weighted by the exterior product of the redundancy profile.

site positions is computed by means of α-Divergence.

2. Significant dependencies between binding site positions have been chosen from error finite sample effects [38, 59].

3. Considering all the significant position, the joint probability for each possible state of two symbols have been calculated. Each value has been saved in a matrix 4x4 where each row and each column corresponds a symbol of $\{A, C, G \text{ and } T\}$.

4. Considering the training matrix with the new sequence added, we read of the symbol $\{A, C, G \text{ and } T\}$ in the new sequence only for the significant positions in the training matrix. From the symbols, we look for the joint probability on the matrix saved.

5. α-Divergence is calculated from the joint probability.

6. A scalar quantity has to be computed using the function defined in equation (5.5) from the α-Divergence.

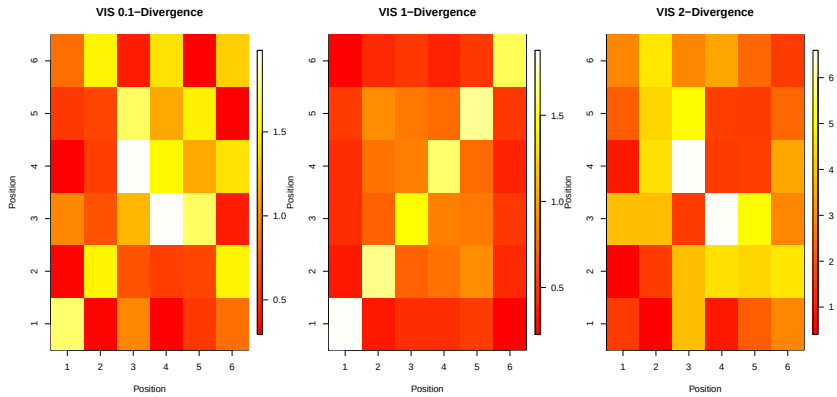


Figure 5.2: Left ot right: α-Divergence heatmap between binding site positions for VIS for q equal to 0.1,1 and 2.

5.6 Results

The α-Divergence measurement determines the correlation through binding site position. When this measurement is zero, we consider that doesn't exist dependence between site positions. On the other hand, if this measurement is positive, there is correlation between site positions. According to the amplitude value, the dependence between binding site positions will be high or low. The number binding sites correlated and its amplitude can be modulated by q parameter, Figure 5.2.

Generally, low q-value will depress the divergence matrix, whereas large q-value will promote divergence values. Therefore, large q-value will show large number of binding site dependence at the cost of introducing additional noise.

The performance of the detector is shown as a Receiver Operating Characteristic (ROC) for different cis-regulatory sequences and q-values, Fig 5.3. The performance of the α-Divergence detector has been compared against a MEME/-MAST [4] and Rényi algorithm (first order)[75]. The detector that produces a bigger area under the convex surface (AUC) will the best learning system.

In Table 5.3, it can be observed that the α-Divergence has a better behaviour than the other detectors. Therefore, assuming position dependence modulated by q-value helps to improve over Entropy method and MEME/ MAST. Moreover, given one Transcription Factor Binding Site, we can be observed how the number of true positives and false positives depends on the q-value, Figure 5.3. The best q-value can be chosen for the detection according to the cost criterion established

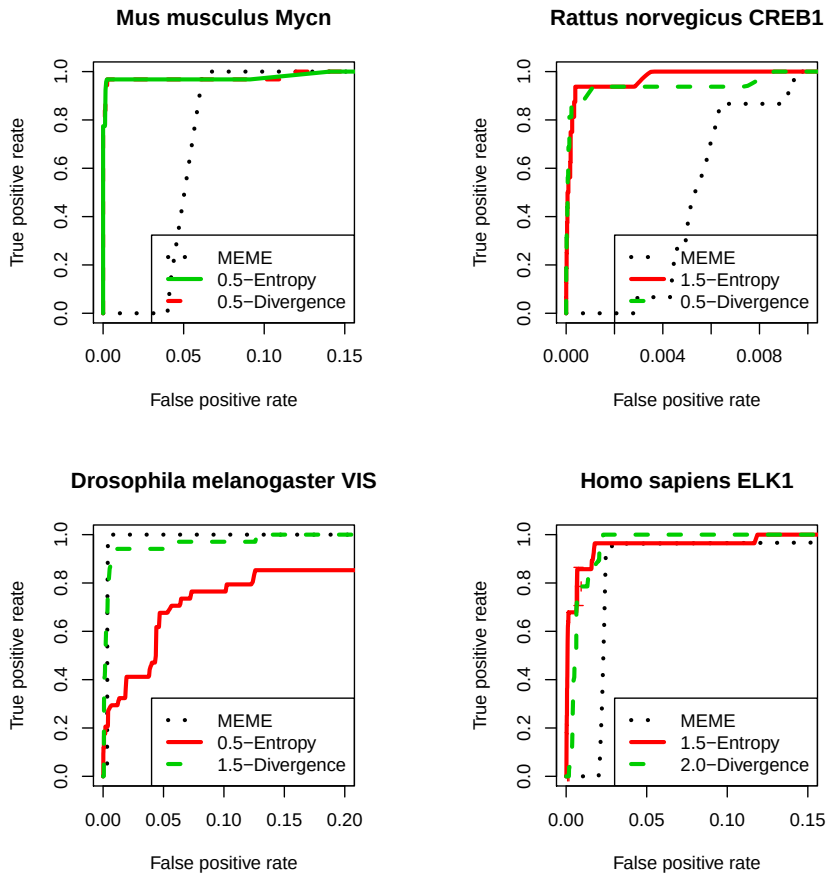


Figure 5.3: Left ot right: Area Under Convex Surface for differents TFBS and organisms.

Table 5.3: Area Under Convex Surface

	Entropy			Divergence			MEME/MAST	
TFBS	q	AUC	Error	q	AUC	Error	AUC	Error
Mycn	0.5	0.99817	0.00862	0.5	**0.99933**	0.00345	0.99872	0.00905
CREB1	1.5	0.99971	0.00084	0.5	**0.99987**	0.00036	0.99952	0.00142
VIS	0.5	0.93448	0.0849	1.5	**0.99532**	0.01168	0.97874	0.04769
ELK1	1.5	0.99341	0.01941	2.0	**0.99398**	0.00358	0.98849	0.02085

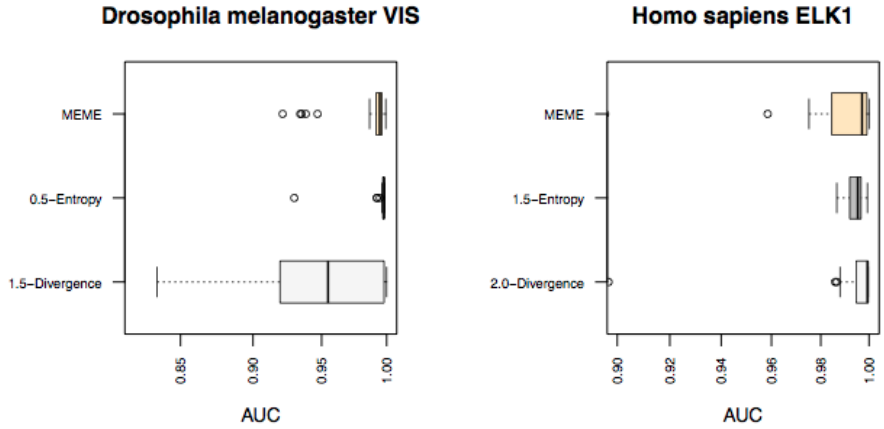


Figure 5.4: Left ot right: Boxplot AUC for differents TFBS and organisms.

between True Positive and False Positive and the area under convex surface maximum.

Given the optimal q-value, the difference between populations for each TFBS and method is shown in the Figure 5.4. It is observed how the populations are different according to method used and the TFBS. Basically, the degree of dispersion and skewness in the data depends on the degree of conservation of the binding site positions. Given biding positions conserved, the degree of dispersion in the data is low and skewness is high. As binding positions conserved decreases, the degree of dispersion increases and skewness decreases.

5.7 Conclusions

We have presented a methodology to detect the cis-regulatory sequences. This method is based on the variation of the total parametric divergence in a set of aligned sequenced when a candidate sequence is added. From this measurement,

the correlation between binding site positions have been considered. The detector employs a parametric divergence. This parameterization allow modulation of the number binding sites correlated and its amplitude. The parametric divergence measurement gives additional information related to binding site than mutual information. The q-value depends on the transcription factor binding sites characteristics. This parameter should be adjusted for each cis-regulatory sequence set by means of a cross validation. This algorithm has been applied on the detection of Mycn, CREB1, VIS and ELK1 recognizers from genomic sequences. The obtained results improve cis-regulatory sequences detection based on Rényi and MEME/MAST. This method and Rényi algorithm are included in the R-package MEET with the name ITEME (Information Theory Elements for Motif Estimation).

Chapter 6

Sequence Information Gain based on Motif Analysis

This chapter is an exact copy of the paper:

- **Sequence Information Gain based on Motif Analysis**. J. Maynou, E. Pairó, S. Marco and A. Perera. BMC Bioinformatics 2015, 16:377 (9 November 2015).

6.1 Abstract

The detection of regulatory regions in candidate sequences is essential for the understanding of the regulation of a particular gene and the mechanisms involved. This paper proposes a novel methodology based on information theoretic metrics for finding regulatory sequences in promoter regions.

This methodology (SIGMA) has been tested on genomic sequence data for *Homo sapiens* and *Mus musculus*. SIGMA has been compared with different publicly available alternatives for motif detection, such as MEME/MAST, Biostrings (Bioconductor package), MotifRegressor, and previous work such Qresiduals projections or information theoretic based detectors. Comparative results, in the form of Receiver Operating Characteristic curves, show how, in 70% of the studied Transcription Factor Binding Sites, the SIGMA detector has a better performance and behaves more robustly than the methods compared, while having a similar computational time. The performance of SIGMA can be explained by its parametric simplicity in the modelling of the non-linear co-variability in the binding motif positions.

Sequence Information Gain based Motif Analysis is a generalisation of a non-linear model of the cis-regulatory sequences detection based on Information Theory. This generalisation allows us to detect transcription factor binding sites with maximum performance disregarding the covariability observed in the positions of the training set of sequences. SIGMA is freely available to the public at http://b2slab.upc.edu.

6.2 Background

The information encoded in genetic sequences is expressed by means of a gene regulation process, which begins with a gene transcription step. The binding between specific proteins and their target sites in DNA is a key step in the control of the transcription process. These proteins – transcription factors (TF) – recognise specific motifs in DNA known as Transcription Factor Binding Sites (TFBS) or cis-regulatory sequences. The prediction, identification and detection of cis-regulatory sequences is a key factor in understanding gene regulation and in inferring regulatory networks [132, 133]. TFBS are usually very short (5 to 20 base pairs long) and highly degenerate, which gives rise to an extremely difficult identification problem due to low statistical power, as short sequences are expected to occur at random every few hundred base pairs. Due to their high variability, a consensus sequence approach for detection is insufficient. There is also evidence that this variability exhibits correlation between positions among the regulatory sequence [105, 123], and that this correlation could contain information which would help reduce the false positive rate and increase the sensitivity of a detector [83].

Due to the importance of identifying cis-regulatory sequences, much effort has been devoted to mapping the binding sites for a large set of transcription factors. An important recent project is the ENCODE (Encyclopedia of DNA Elements) project, which has been able to map 4 million regulatory regions in the human genome, opening new possibilities for computational methods [29]. Motif detection methods may be classified in different ways, depending on the approach adopted. Some reviews focus on the biology of motif discovery in regulatory regions [130, 98], whereas other publications focus more on the representation of the motifs: consensus-based methods and alignment-based methods [86]: consensus-based methods use word algorithms which consider binary hit/no-hit values [16, 111], and alignment-based methods use a set of alignment sequences with binding evidence to assign putative motifs to a candidate sequence. These latter methods could be classified as either numerical or stochastic models: numerical models are based on a mathematical representation of the nucleotides, whereas stochastic models, which are probably the most popular methods, are based on Position Weight Matrices (PWM) or Position Specific Weight Matrices

(PSWM) [117]. A PWM is a matrix of scores corresponding to the frequency of the sequence symbols for each binding site position. The PWMs allow the capture of the variability over a sequence of nucleotides from a set of binding site positions [39], although there is the implicit assumption of independence between the residues of the aligned sequence matrix. PWM representations have been used in several algorithms to discover over-represented patterns from candidate sequences [112].

As noted above, statistical studies have shown the dependence among binding site positions variability. The common strategies for incorporating these dependencies within motif detectors include the extension of the PSSM approach to include pairs of correlated positions [81, 56], m^{th} order Markov chains (HMM) [139, 32] and Bayesian Networks[10, 94, 140, 9]. HMM can model the position interdependencies as long as high order HMMs, or a Bayesian approach are used but, in order to train any of both methods model sufficiently well, a huge training set of sequences would be required ($\pm$ 1000 or more sequences per model).

A popular method, based on some of the previous work, is MEME/MAST, which provides an improved detection performance [3]. MAST is part of the MEME suite and uses a Q-FAST algorithm for finding motifs. Although these strategies may perform well in some datasets, they have shown certain limitations in the number of dependencies which may be considered between positions, in their ability to model dependencies between more distant positions, and in the large number of parameters which need to be adjusted in the models [105].

Previous work by our group proposed a parametric detector using the Rényi Entropy for binding site detection [75]. This measurement allowed us to build variable-sensitivity detectors modulated by the Rényi order – this assumed independence between binding site positions. A first approximation for modelling the correlation among binding site positions, known as Qresiduals, used a linear embedding to represent the set of binding site sequences [83] and employed a residuals-based approach as the detection statistic. Other non-related work modelled the pure correlation between binding site positions through non-linear correlations based on the variation of mutual information [76].

Statistical pattern recognition has also been applied to identification of sequence motif. Luo et al [72] propose to use discriminant analysis for the prediction of Transcription Start Sites (TSS). From non-parametric measure, similar to Shannon information, Luo et al [72] provide information about the variance observed in the dataset. This strategy has good performance for the binding motif detection when the motif positions are not correlated among them. But, this measurement does not allow modelling the dependencies among motif positions.

In this paper, we propose a generalisation of a non-linear model based on Information Theory, which allows modeling DNA contact by the protein and the biological interaction among binding sites using a small training set of sequences

(5-50 sequences model). This new approach aims at a trade-off between the good generalisation properties of pure entropy methods and the ability of position-dependency metrics to improve detection power.

The performance of the proposed detector method, named SIGMA (Sequence Information Gain based Motif Analysis), is compared with different computational methods for binding site detection: MEME/MAST [3], Biostrings [82], MotifRegressor [20], Qresiduals [83] and a previously published set of algorithms based on information theory [75, 76].

6.3 Information Gain Space

The information gain has been measured for each TFBS by means of two parametric uncertainty estimators. The rationale is based on the idea that the total information gain of a set of true TFBS aligned sequences will change according to the similitude of the new candidate sequence to that set (Figure 6.1). The first estimator measures the total amount of information change produced by assuming position independence, whereas the second estimator measures the total amount of change of per-position mutual information (capturing pure correlation among binding site positions). Both estimators are computed by a parametric uncertainty measurement.

Let us consider a set of I aligned sequences (s_i) with binding evidence $M = \{s_i, i = 1, \cdots, I\}$, and the same set including a candidate sequence s_c, $S = s_c \cup M$.

Following Figure 6.1, let a be the coordinate corresponding to the set M, with axes determined by the two measures previously mentioned. When a new candidate sequence is considered in S, both measures will vary to b or g depending on the nature of the candidate sequence. When the candidate sequence is a binding site sequence, $(b,)$ the variation on the information will be not significant. However, when the candidate sequence is a genomic sequence, (g), the amount of information will vary significantly. With a sufficient training set, this information gain space can be split in two regions, genomic and binding, by means of a simple discriminant analysis which will define a decision boundary, as highlighted as a dashed line in Figure 6.1. The decision boundary shape is the result of applying non-linear function.

6.4 Information content measures

We have employed as parametric uncertainty measurements the Rényi entropy and Rényi Divergence (also called α-Divergence) [96], which are defined as:

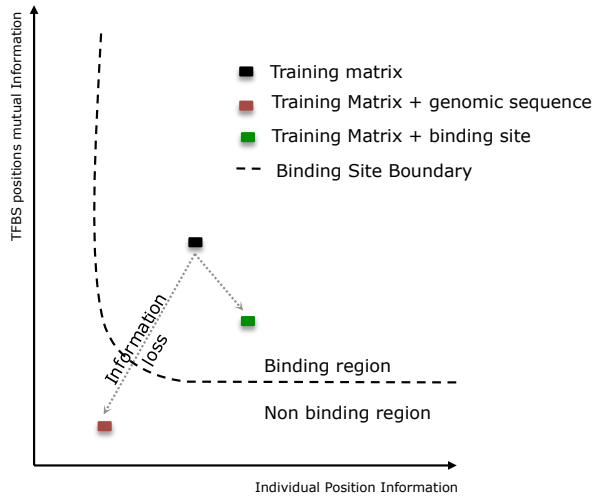


Figure 6.1: Information gain space defined by means of the variation on the information. X-axis on the graph shows the total amount of information change produced by assuming position independence. Y-axis shows the total amount of information change produced by assuming the correlation among positions. Black box Training matrix, red box Training matrix with genomic sequence, green box Training matrix with binding sites sequence. The broken line is the decision boundary.

$$H_q(X) = \frac{1}{1-q} log_2 \sum_{i=1}^{4} p(X_i)^q \tag{6.1}$$

$$D_q(X;\, Y) = \frac{1}{q-1} log_2 \sum_{i=1}^{4} \sum_{j=1}^{4} P(X_i, Y_j)^q Q(X_i, Y_j)^{1-q} \tag{6.2}$$

where X_i and Y_j are the nucleotides $\{A,\ T,\ C$ and $G\}$ at different DNA sequence positions, $P(X,Y) = p(X,Y)$, $Q(X,Y) = p(X) * p(Y)$ and the q is the Rényi order which modulates the probability of occurrence of each symbol. $p(X,Y)$ is the joint probability of X and Y, $p(X)$ and $p(Y)$ are the marginal probability. Both measurements ($H_q(X)$ and $D_q(X;\, Y)$) depend on q which is

a positive real number ($q \neq 1$)and both are non-negative for all $q \geq 0$. This parametrisation allows the building of a variable-sensitivity detector exploiting the statistical properties of the Redundancy, R, where R is defined as [75].

The measurement of the variation when the candidate sequence is added to the set has been computed using two heuristic functions, see (eq.6.3 and eq.6.4). These functions depend on two parameters, γ and ω, which measure the difference between redundancies, eq.6.5, and divergence, eq.6.6, between the set of aligned sequences without the candidate sequences, s_i, and with candidate sequence, M. These are estimated as described in Maynou et al [75].

$$\rho(q, M) = \mid \sum_{i=1}^{L} R_q^{M_i} \gamma_i \mid^{-1} \tag{6.3}$$

$$\eta(q, M) = \mid \sum_{i=1}^{L} \mid R_q^{M_i} \mid \omega_i \mid^{-1} \tag{6.4}$$

where, γ_i and ω_i are

$$\gamma_i = \mid R_q^{M_i} - R_q^{S_i} \mid \tag{6.5}$$

$$\omega_i = \mid D_q^{M_i} - D_q^{S_i} \mid \tag{6.6}$$

where L is the number of nucleotides in the binding region, M is the aligned set of sequences with binding evidence and i is a specific column of M. R_q^M is the redundancy, normalized depending on the maximum entropy on the set of aligned sequences, whereas R_q^S contains the equivalent parametric entropy when the candidate sequence is assumed to belong to the set. The redundancy profile is a L-dimensional vector, where L is the total number of positions of the binding site. D_q^M is the divergence matrix of the set of aligned sequences and D_q^S is the divergence matrix considering the training matrix with the candidate sequence. The main diagonal is set to zero in each of these matrices, D_q^M and D_q^S. The variation in the information is therefore calculated by means of γ and ω and q-values are optimised at the validation stage within the range $(0, 2]$. As q increases, the noise included in the redundancy signal also increases [75]. From q-values higher than 2, signal-to-noise ratio is not optimal.

For a genomic sequence, the order of the system will decrease the values of γ and ω, whereas for a binding sequence the order of the system will not be altered substantially. Each candidate sequence will therefore be characterised by the pair $(x = (\rho, \eta))$ and classified as genomic or binding by means of a Quadratic Discriminant Analysis (QDA), as shown in Figure 6.6. The decision boundary,

$H(y)$, is defined from the distribution of the variation on the information, x, for each class, genomic or binding, in the information gain space.

Binding site detection by means of the SIGMA algorithm can be summarized as follows, see Figure 6.2:

1. Given a set of aligned sequences with binding evidence M, estimate the redundancy profile R_q^M and the Rényi Divergence D_q^M (eq. 6.1) and (eq. 6.2).

2. Given a new candidate sequence, re-estimate both values assuming the candidate sequence belongs to M, R_q^S and D_q^S,.

3. Compute the variation on the information $x = (\rho, \eta)$ as defined in eq. (6.3) and eq. (6.4).

4. Quadratic Discriminant Analysis is applied to the information gain space from the set of computed features.

5. Steps 3 and 4 are iterated over for each candidate sequence.

Additionally, for characterisation of the results we define a heuristic magnitude C, related to the *Complexity* of M, in order to characterise the degree of pure correlation between the variability of binding site positions in M, see (eq. 6.7). C computes element by element the ratio between divergence value, where $D_q\mid_{i,j}$ is the element of D_q at row i and column j, and maximun entropy, H_{max} without to considerer the main diagonal. The average of the ratios define the complexity of M.

$$C = \frac{\sum_{i,j=1}^{N} D_q\mid_{i,j}}{N * (N-1) * H_{max}} \quad ; \qquad i \neq j \tag{6.7}$$

where D is the parametric uncertainty measurement considered, N is the size of the binding sites, q is the Rényi order and H_{max} is the maximum entropy for the set of probabilities $p(X)$ and $p(y)$, see section 2.2.. C is a value between 0 and 1. When C is close to 1, the degree of correlation among binding site positions is high.

6.5 Database Description

Data has been obtained from the Jaspar database [128], `http://jaspar.genereg.net/` (see Table 6.1 and Table 6.2).

The *JASPAR Core* provides non-redundancy and high-quality alignment matrices for each transcription factor [128]. Results have been computed with

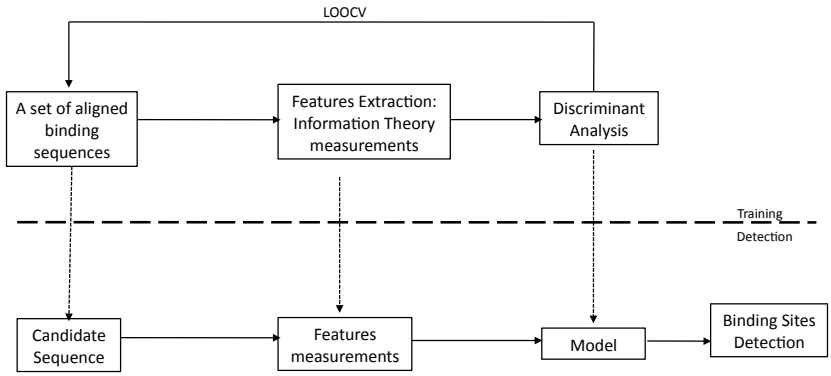


Figure 6.2: The essential steps in the training and detection process are shown for the SIGMA algorithm.

background genomic sequences from the Eukaryotic Promoter Database (EPD) [99], using the EPD version based on the EMBL release 105 (sept 2010). The background loci chosen were $EP74078(+)HsRPS9P2+$ for *Homo sapiens* and $EP07119(+)MmIgk0\ MPC11$ for *Mus musculus*.

6.6 Optimization

To apply SIGMA methodology to TFBS detection over genomic sequence, we should calculate the variation of the information, eq.6.4, as many times as the length of the sequence I (typically millions nucleotides). Given a sequence position, we must calculate the divergence between the binding site positions. This means that we must compute $\frac{L*(L-1)}{2}$ times the joint probability for each training matrix, where L is the total number of binding site positions in M. The running time of the algorithm depends on the length of the candidate sequence and on the number of binding site positions.The run time is therefore linear in the length

Table 6.1: Summary of the Transcription Factors Analysed for the *Homo sapiens* organism obtained from Jaspar database.

TF	Family	Base	Sequences
ELK4	Ets	9	20
ETS1	Ets	6	40
NFATC2	REL	7	26
MYCMAX	bHLH	12	21
E2F1	E2F	8	10
MAX	bHLH	12	17
NFIL3	bZIP	11	23
NFE2L2	bZIP	11	20
INSM1	Zinc finger	12	24
CREB1	bZIP	12	16
Irf2	IRF	18	12
FOXO3	Forkhe	8	13
HLF	bZIP	12	18
NFKappaB	REL	10	38
MZF114	Zinc finger	6	20
ESR1	HNR	9	18
FOXD1	Forkhe	8	20
MZF1513	Zinc finger	10	16
Ap1	bZIP	7	18

Table 6.2: Summary of the Transcription Factors Analysed for the *Mus musculus* organism from Jaspar database.

TF	Family	Base	Sequences
Pax2	Homeo	8	31
FOXO3	Forkhe	8	13
NFkappaB	REL	10	38
ARID3A	ARID	6	27
EBF1	bHLH	25	10
En1	Homeo	11	10
NR3C1	HNR	18	9
Egr1	Zinc finger	11	15
Ap1	bZIP	7	18
Runx1	Runt	11	26
CREB1	bZIP	12	16
AhrARNT	bHLH	6	24
Pdx1	Homeo	6	31
NFATC2	REL	7	26
Lhx3	Homeo	13	20
ARNT	bHLH	6	20
ELF5	ETS	9	44

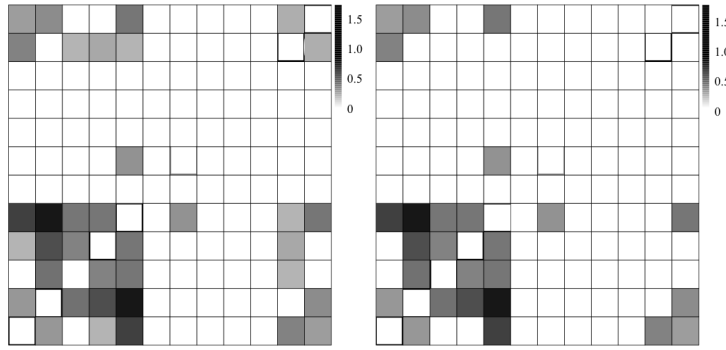


Figure 6.3: Left: Rényi Divergence, $D^M_{q=1}$, considering all possibles correlations between binding site positions. Right: $D^M_{q=1}$ considering only significant dependences between binding site positions after applying the error finite sample correction. Black boxes mean maximum correlation and white boxes mean zero correlation between binding site positions.

of the input sequence and quadratic in the length of the binding site L.

$$T(L) \in O(L^2) \tag{6.8}$$

The optimization algorithm is based on considering only the correlated binding site positions. The η function has been calculated considering only the Rényi-divergence of the correlated binding site positions (showing positive correlations) through a screening on the possible positive dependencies between these positions.

Any two binding site positions are considered to be correlated if the Rényi divergence score is bigger than the error finite sample. This error yields to a bias on the uncertainty parametric measurement caused by estimating the probability using the nucleotide frequencies [75]. After the screening, we only compute based on the correlated positions of the training matrix as shown in Figure 6.3.

For each pair of positions (i, j) in M where $i, j = \{1, ..., L\}$, the joint probability for all the possible combinations of (x_i, x_j)=$\{A, C, G$ and $T\}$ are precomputed and stored in a 4×4 matrix. We construct a library (B_{i,j,x_i,x_j}) of sixteen 4×4 matrices containing all the possible joint probability values for each pair of positions i and j (as illustrated in Figure 6.4).

For each new candidate sequence, we have to consider only the symbols matching correlated positions and read the joint probability value from the lookup table B_{i,j,x_i,x_j}. The Rényi divergence and the discrimination function, η

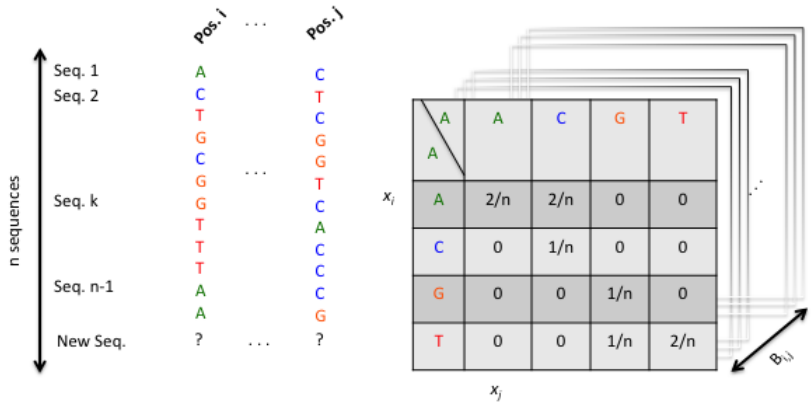


Figure 6.4: For each pair of positions i and j we calculate a joint probability matrix, B_{i,j,x_i,x_j}, using all possible combinations of $\{A, C, G \text{ and } T\}$.

are then computed from these values. The estimated total number of significant transcription factor site dependencies in *Homo sapiens* and *Rattus novergicus* is approximately 50% and 37% [122]. In this way, the computing time can be reduced by approximately an order of magnitude.

6.7 Validation

In order to build a model for each set of binding site sequences, the SIGMA detector has been characterized by means of leave-one-out cross validation (loo-cv). Each method has its own characteristic parameter. The range of the parameter used is different for each detector, see Table 6.3. The detector performance depends on the value of these parameters which have been selected employing loo-cv. Taking as a criteria a heuristic magnitude, ν_{auc}. This parameter has been computed from the mean and variance of the area under the N ROC curve (AUC_N) [83], which will be maximised for all methods.

$$\nu_{auc} = \mu_{auc} * (1 - \sigma_{auc}) \tag{6.9}$$

where μ_{auc} and σ_{auc} are the mean and the variance of AUC_N. ν_{auc} is a value between 0 and 1. When ν_{auc} is close to 1, the mean is close to 1 and the variance is close to 0.

Table 6.3: Summary of the characteristic parameters and the range considered for the validation of each computational method used.

Method	Parameter	Range
SIGMA	Rényi Order	(0,2]
MEME/MAST	Length Motif (L)	[1,L]
Qresiduals	Principal Components	[1,10]
Entropy	Rényi Order	(0,2]
Divergence	Rényi Order	(0,2]
Biostrings	Not Applicable	Not Applicable
MotifRegressor	Length Motif (L)	[1,L]

From the performance data, we have calculated the mean and standard deviation of the AUC for each transcription factor and method by means of the outer loo-cv. This process has been repeated for all the TFs listed in Table 6.1 and Table 6.2.

6.8 Results and Discussion

We first show a characterisation of how the performance of the individual algorithms based on Entropy and Divergence depends on the complexity properties of the training matrix (M) C, (eq. 6.7), see Figure 6.5. The performance of these algorithms will vary on C depending on the design of each algorithm and the true correlation between positions found for each set of binding sequences. As one would expect, the total Entropy algorithm has a better behaviour with low values of C, whereas a Divergence based approach improves its performance when C is large. The SIGMA approach is partially based on both measurements and aims at finding a trade-off between both approximations in order to maximise the performance over the full dynamic range of C.

Figure 6.6 shows an example of real case where each input sequence is represented as a point in (ρ, η) coordinates. This set of samples includes genomic or binding sequences as shown in the figure. It is clear from the figure that both variables are contributing to the separation of the true binding site sequences.

The performance of SIGMA, MEME/MAST, Qresiduals, Entropy, Divergence, Biostrings and MotifRegressor has been compared against the same set of TFs under the same validation conditions described in the previous section. In Figure 6.8, it can be observed that the mean and standard deviation depend both on the Transcription Factor and on the method considered. The performance among all the methods has been compared by means of the ν_{auc} parameter

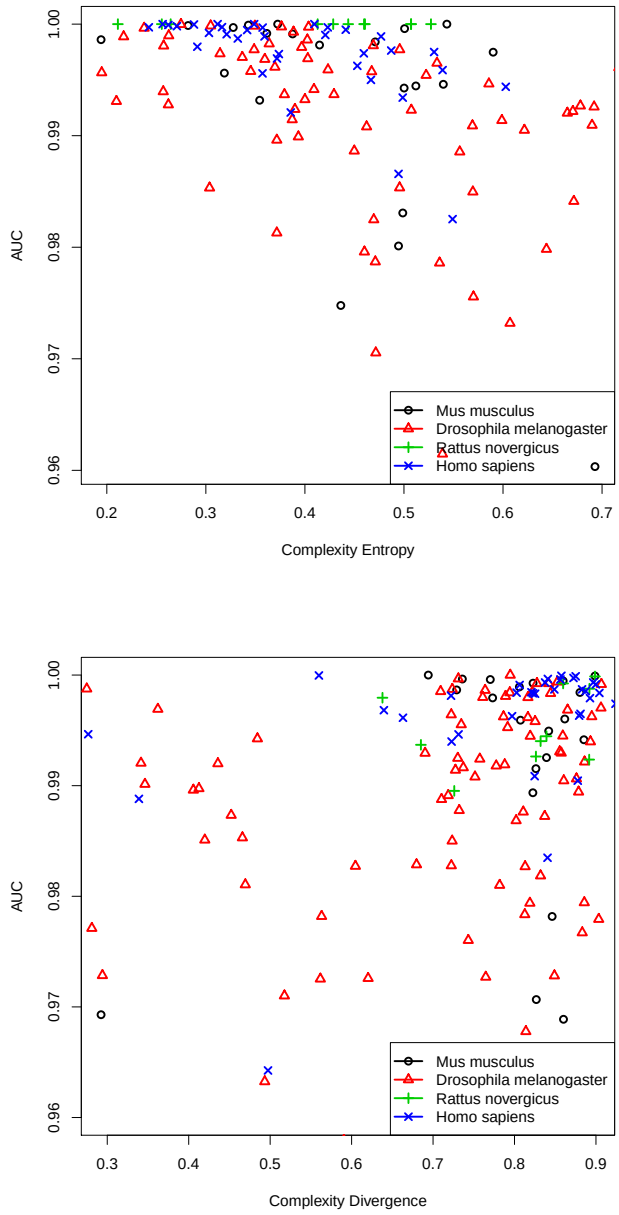


Figure 6.5: Up to down: Entropy and Divergence performances against *Complexity* (degree of correlation between binding site positions) for a set TF of different organisms (× *Homo sapiens* , △ *Drosophila melanogaster*, + *Rattus norvegicus*, *o* *Mus musculus*) . Entropy performs better for low *Complexity*. On the contrary, Divergence performs better for large *Complexity*.

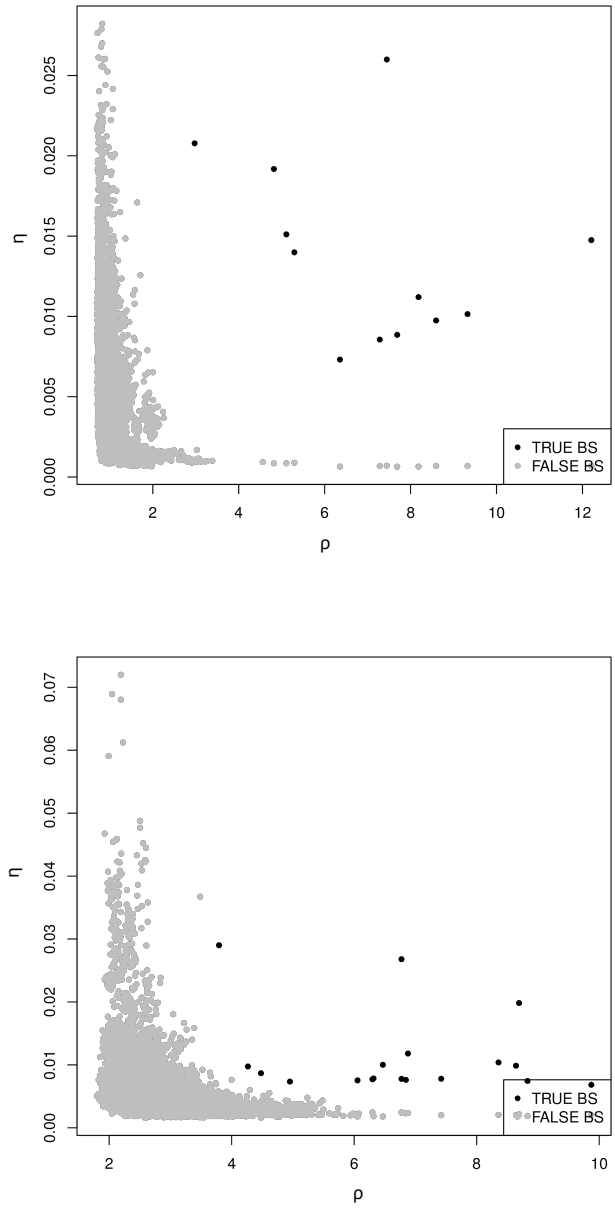


Figure 6.6: Empirical representation of the concept depicted in Figure 6.1. Up to down: Information Gain when candidate sequences are inserted in the Transcription Factor Binding Sites *Irf2* and *HLF* for the *Homo sapiens* organism. Black points correspond to candidate sequences which are true binding site sequences. Grey points correspond to candidate sequences which are false binding site sequences.

Table 6.4: Per CPU, the total run time was calculated on a $2.3GHz$ Intel Core 2 Duo $P8600$ computer with $4GB$ RAM.

Method	Run time (s)	sd (s)
SIGMA	0.132	0.007
Qresiduals	0.119	0.006
Entropy	0.051	0.003
Divergence	0.081	0.004
MEME/MAST	0.019	0.001
Biostrings	0.004	0.0001
MotifRegressor	0.144	0.02

described in eq. (6.9).

In Figure 6.7, the ν_{auc} parameter is shown for each transcription factor and method. Based on the ν_{auc} values, in approximately 70% of the TFBS under study, SIGMA shows better performance than the other methods. In 20% of the TFs, the performance of the others methods is better than that of SIGMA. In the remaining cases, the SIGMA performance is similar to one or several of the computational methods considered. In most cases, the mean AUC is close to one and the variance is approximately zero, which suggests that SIGMA also behaves more robustly than other methods, as seen in Table 6.5 and Table 6.6.

We computed a Wilcoxon rank-test [134] in order to estimate whether the improvement in performance is statistically significant. The null hypothesis was that the AUC distributions between SIGMA and other methods were the same and the alternative hypothesis was that the AUC distributions were different. The level of significance is represented by $-log_{10}(p_{value})$. Any $p_{value} > 0.05$ is shown in bold, see Table 6.7 and Table 6.8). In most cases, it can be observed that the difference between the AUC distributions is significant.

The computational time of SIGMA was compared with the set of computational methods considered. The C code for Qresiduals, Entropy and Divergence using the model obtained in validation and MEME/MAST (Version 4.4.0) was used and has been made publicly available. The run time was obtained in comparison with randomly generated candidate sequences of 1500 nucleotides. The total time has been calculated from 100 iterations of each algorithm. The averages of the computational times in detection for the set of TF considered of *Homo sapiens* (Table 6.5 and Table 6.6) are shown in Table 6.4.

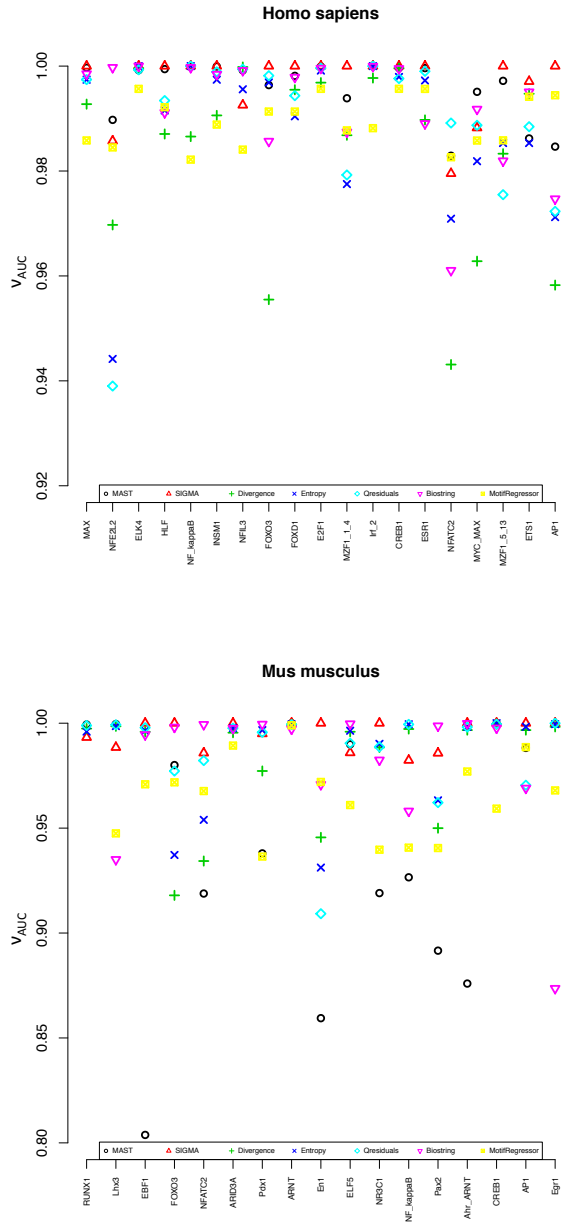


Figure 6.7: Top to bottom: Performance of each algorithm (o MAST,$\triangle$ SIGMA, $+$ Divergence, $\times$ Entropy,$\diamond$ Qresiduals, $\bigtriangledown$ Biostring, $\diamond$ MotifRegressor) is shown through ν_{auc}, (eq. 6.9), for a set of TFBS for the *Mus musculus* and *Homo sapiens* organisms. When ν_{auc} is close to 1, the mean is close to 1 and the variance is close to zero. For each TF, the best computational method will be that for which ν_{auc} is closest to 1.

6.9 Conclusions

A new methodology based on a discriminant analysis of two information theoretic measures has been proposed for binding site detection.

The variation on the information has been measured through two parametric uncertainty measurements (the Rényi entropy and Rényi divergence). The method focusses on the variation in these information measures when a new sequence is assumed to belong to a training set of sequences with known binding properties.

This methodology allows us to detect cis-regulatory sequences with maximum performance disregarding the co-variability observed in the positions of the training set of sequences. SIGMA has been characterised on the detection problem for a large set of transcription factors and compared with different motif detection algorithms. AUC distributions have been calculated which show that there is a statistically significant difference between SIGMA performance and the performance of the other methods. In approximately 70% of the cases considered, SIGMA has exhibited better performance properties, at comparable levels of computational resources, than the methods with which it was compared.

As you can see through the heuristic parameter, SIGMA method is more robust than the other methods. A model based on both parametric uncertainty measurements can be useful to detect cis-regulatory sequences. But when the number of the positions involved in the binding sites process is small, the SIGMA performance is comparable with the rest of the computational methods.

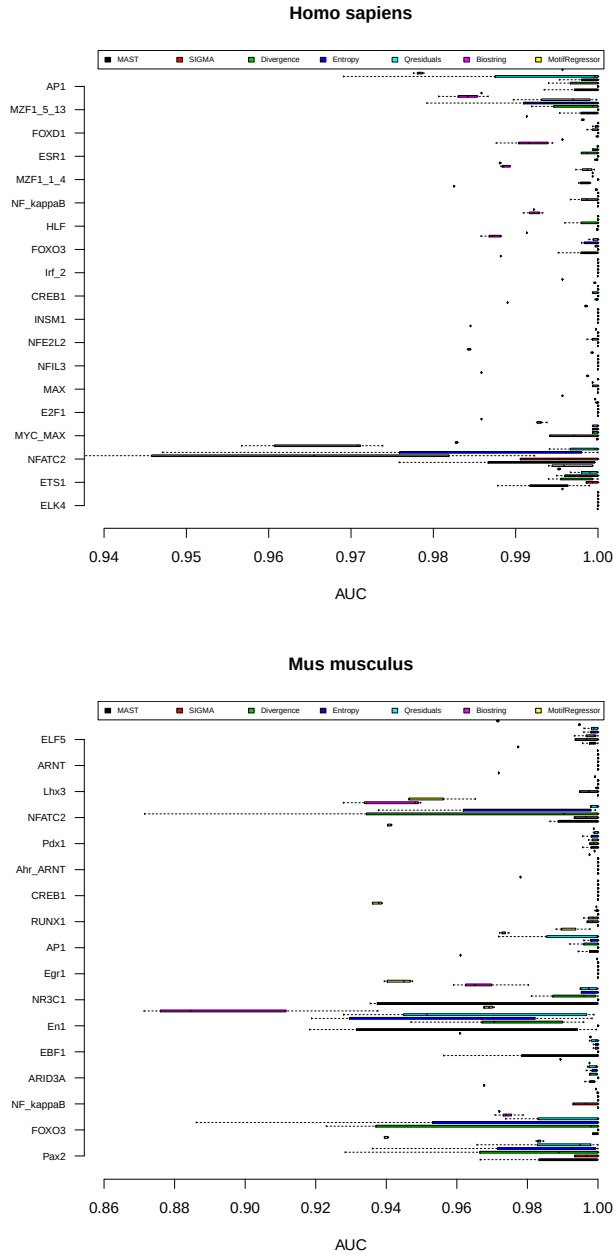


Figure 6.8: Top to bottom: Box plot of the AUC and its variation for the studied transcription factors for the *Homo sapiens* and *Mus musculus* organisms using different computational methods: black MAST, red SIGMA, green Divergence, blue Entropy, cyan Qresiduals, pink Biostring and yellow MotifRegressor. The background sequences used have been $EP74078(+)HsRPS9P2+$ for the *Homo sapiens* and $EP07119(+)MmIgk0\ MPC11$ for the *Mus musculus*.

Table 6.5: Results for the set of computational methods considered for each TF of the *Homo sapiens* organism. The ν_{auc} is defined through the mean and variance of the AUC_N using a cross-validation method. Given a TF and method, ν_{auc} is chosen with maximum mean and lower variance in the AUC_N.

	ν_{AUC}						
TFBS	MEME/MAST	Qresiduals	SIGMA	Entropy	Divergence	Biostrings	MotifRegressor
ELK4	0.99923	0.99993	1	1	0.99961	1	0.99566
ETS1	0.98621	0.98845	0.99707	0.98533	0.99473	0.99508	0.99415
NFATC2	0.98291	0.98915	0.97952	0.97091	0.94311	0.98284	0.98263
MYCMAX	0.9951	0.98872	0.98823	0.98187	0.96281	0.99178	0.98581
E2F1	0.99991	0.99963	1	0.99915	0.99685	0.99958	0.99566
MAX	0.99968	0.99743	1	0.99741	0.99275	0.99852	0.98583
NFIL3	0.9992	0.9994	0.99256	0.99558	0.999823	0.99917	0.98408
NFE2L2	0.98975	0.93901	0.98573	0.94418	0.96973	0.99974	0.9845
INSM1	0.99993	0.99891	1	0.99741	0.9906	0.99842	0.98885
CREB1	0.99965	0.99763	1	0.99793	0.99962	0.99953	0.99567
Irf2	0.99995	1	1	1	0.99773	0.99995	0.98817
FOXO3	0.99638	0.99817	1	0.99688	0.95549	0.98567	0.9915
HLF	0.99943	0.99343	1	0.99155	0.98706	0.99113	0.99216
NFkappaB	0.99987	1	1	1	0.98657	0.98256	0.98217
MZF114	0.99387	0.97925	1	0.97751	0.98682	0.98743	0.98775
ESR1	0.99962	0.99901	1	0.99725	0.98974	0.98903	0.9957
FOXD1	0.99814	0.99436	1	0.99043	0.99549	0.99787	0.99133
MZF1513	0.99719	0.97549	1	0.98534	0.9833	0.98193	0.98585
Ap1	0.98465	0.97231	1	0.97121	0.95825	0.97469	0.99445

Table 6.6: Results for the set of computational methods considered for each TF of the *Mus musculus* organism. The ν_{auc} is defined through the mean and variance of the AUC_N using a cross-validation method. Given a TF and method, the ν_{auc} is chosen with maximum mean and lower variance in the AUC_N.

	ν_{AUC}						
TFBS	MEME/MAST	Qresiduals	SIGMA	Entropy	Divergence	Biostrings	MotifRegressor
Pax2	0.89161	0.96215	0.98572	0.96323	0.94998	0.98245	0.93971
FOXO3	0.98005	0.97719	1	0.93721	0.91796	0.97079	0.972
NFkappaB	0.92656	0.99944	0.982322	0.99949	0.99723	0.99939	0.96767
ARID3A	0.99757	0.99764	1	0.99771	0.99548	0.99753	0.98933
EBF1	0.80379	0.99787	1	0.9964	0.99593	0.99769	0.95929
En1	0.85943	0.90921	1	0.93119	0.94558	0.8736	0.96797
NR3C1	0.91904	0.98873	1	0.99017	0.98844	0.95811	0.94069
Egr1	0.99983	0.99996	1	0.99956	0.99826	0.99969	0.961
Ap1	0.98823	0.97044	1	0.99828	0.99672	0.96902	0.98861
Runx1	0.99937	0.99891	0.99323	0.99601	0.99743	0.99951	0.93645
CREB1	0.99997	0.99953	1	1	0.99958	0.99987	0.97698
AhrARNT	0.87593	0.99816	1	0.99828	0.99672	0.99721	0.99901
Pdx1	0.93796	0.99565	0.99499	0.99669	0.97722	0.99871	0.94051
NFATC2	0.91883	0.98219	0.98581	0.95394	0.934316	0.93503	0.9475
Lhx3	0.99961	0.99924	0.98846	0.99862	0.99852	0.9981	0.97183
ARNT	0.99998	0.99935	1	0.99945	0.99945	0.9999	0.9999
ELF5	0.98992	0.99045	0.98593	0.99641	0.99593	0.99453	0.97089

Table 6.7: The level of significance corresponding to $-log_{10}(p_{value})$ calculated using the Wilcoxon-rank test for the *Homo sapiens* organism. The null hypothesis is that the AUC distributions between SIGMA and the other computational methods are the same and the alternative hypothesis is that the AUC distributions are different. $p_{value} > 0.05$ is in shown in bold.

	$-log_{10}(p_{value})$					
TFBS	Qresiduals	MEME/MAST	Entropy	Divergence	Biostrings	MotifRegressor
ELK4	1.58	1.46	5.80	9.41	9.48	9.60
ETS1	3.48	7.55	7.96	7.52	7.51	7.85
NFATC2	**0.71**	7.61	2.81	5.21	9.48	9.59
MYCMAX	2.25	7.59	2.31	7.83	7.55	9.60
E2F1	1.58	7.12	2.33	3.12	7.56	9.6
MAX	3.73	4.16	2.66	5.13	5.10	6.46
NFIL3	**1.20**	6.10	**1.19**	6.05	6.21	7.82
NFE2L2	**1.20**	4.10	**0.80**	2.98	4.35	5.11
INSM1	2.33	8.63	**1.20**	2.08	8.95	10.11
CREB1	2.31	8.47	**1.20**	**1.20**	8.47	8.68
Irf2	**0.80**	6.79	3.37	6.14	6.78	6.89
FOXO3	2.31	6.11	5.63	5.20	6.48	8.26
HLF	3.38	4.45	**0.80**	**1.20**	2.08	6.02
NFkappaB	**1.20**	6.87	3.40	6.50	6.83	6.96
MZF114	7.52	13.95	10.99	3.90	14.11	9.65
ESR1	1.95	6.10	3.74	5.43	6.11	7.81
FOXD1	1.95	1.32	**1.20**	**1.09**	7.11	8.22
MZF1513	6.10	3.72	3.41	3.78	3.71	4.32
Ap1	4.75	13.51	2.67	3.03	13.5	17.14

Table 6.8: The level of significance corresponding to $-log_{10}(p_{value})$ calculated using the Wilcoxon-rank test for the *Mus musculus* organism. The null hypothesis is that the AUC distributions between SIGMA and the other computational methods are the same and the alternative hypothesis is that the AUC distributions are different. $p_{value} > 0.05$ is in shown in bold.

	$-log_{10}(p_{value})$					
TFBS	Qresiduals	MEME/MAST	Entropy	Divergence	Biostrings	MotifRegressor
Pax2	3.40	10.11	**0.81**	**1.20**	9.89	11.37
FOXO3	2.66	4.06	4.06	4.06	4.06	4.13
NFkappaB	7.14	8.80	5.65	4.88	9.13	11.08
ARID3A	10.05	2.68	**0.17**	**0.17**	2.68	9.5
EBF1	6.78	3.09	3.52	5.61	3.73	14.27
En1	4.06	4.82	2.66	5.10	5.10	6.47
NR3C1	3.37	5.79	**0.80**	**1.20**	4.53	7.14
Egr1	**1.20**	2.15	2.43	**2.14**	2.15	7.89
Ap1	4.75	4.76	2.66	4.76	4.76	4.89
Runx1	4.75	10.65	10.21	10.21	10.23	12.7
CREB1	1.57	3.71	3.01	2.66	3.71	3.72
AhrARNT	**1.19**	3.80	6.35	11.04	11.13	11.36
Pdx1	2.06	9.15	**0.80**	**0.80**	9.15	9.59
NFATC2	**0.21**	**0.66**	3.67	**0.05**	4.25	15.46
Lhx3	4.47	5.78	**0.80**	**0.80**	5.47	7.36
ARNT	**0.80**	**0.48**	**0.45**	1.78	**0.45**	11.28
ELF5	2.37	2.20	6.15	9.48	9.48	9.57

Chapter 7

An R library for the detection of TFBS

7.1 Introduction

MEET 5.1 is an R-package including 523 models carefully built and a set of tools for TFBS detection. The models allow the detection of cis-regulatory sequences in different organisms and the different tools allow a directly comparison among algorithms on the same dataset. The parametric space can be independently explored for each one of the included algorithms. The MEET 5.1 R-package is available as a contributed package from the Comprehensive R Archive Network (CRAN). A web interface of for the MEET package is also available.

Internal algorithms

The package includes three algorithms, ITEME(Entropy [75] and Divergence [76]) and Q-residuals [83]. ITEME calculate the information of an aligned set of binding sites, and then the variation of this information when a candidate sequence is added to the model. The assumption made is that, when the new sequence is a binding site, the information gain will be near zero, because the sequence will be similar to the previous ones, but when the sequence is not a binding site the information added will be larger. To calculate the variation of the information, two approaches can be taken: to consider that the position within the binding sites are independent using a Rényi entropy [96] or to take into account position interdependences using the divergence.

The Q-residuals detector is based on a principal components analysis (PCA) of the numerical representation of DNA sequences. The first step is to convert

the aligned binding sites sequences into a matrix of numerical DNA sequences through the conversion proposed by [108]. Each nucleotide is placed at the vertex of a regular tetrahedron. Using this conversion, the M aligned sequences of length L became a $M \times 3 \cdot L$ matrix of numerical sequences. Then a PCA is applied to the numerical sequences matrix.
The error E of the principal components model is a matrix of $M \times (3L)$ dimensions that is used to calculate the residuals used to detect the binding sites.

In a similar manner the hypothesis of the method is that, when a new sequence is projected into the principal components subspace, binding sites will have smaller Q-residuals than genomic sequences.

External algorithms

The package allows the use of MDscan and MEME/MAST [4, 6] if these programs are detected as available on the installation system. The package also includes a custom implementation of the MATCH algorithm [54] in R, as there no exists any public version. MATCH uses the information per position in order to define a similarity score between the motif and a candidate sequence. where Max and Min are the maximum and minimum possible scores for a candidate sequence. This score is calculated for the sequence and the core (which is defined as the 5 consecutive positions with more information), then a threshold is set in these scores in order to differentiate binding sites. The main difference between the included implementation and the original algorithm in MATCH is that, in order to calculate the information, the nucleotide probabilities in the background of the organism have been considered.
MEME/MAST can be downloaded from the MEME suite [7] and MDscan from the MDscan web page [69]. The current version of MEET 5.1 is prepared to work with MEME version 4.4.0. and MDscan (2004).

7.2 Architecture of MEET 5.1

The package includes a library of 523 optimized models from 181 motifs extracted from the JASPAR (2010) core. This consists on the Q-residuals, the Divergence and the Entropy models of the TFBS that have more than 10 available sequences in the JASPAR (2010) database and correspond to the organisms: *Drosophila melanogaster*, *Rattus norvegicus*, *Mus musculus* and *Homo sapiens*. A relation with the number of models for each organism and algorithm can be seen in Table 7.1. The toolkit includes means for html output reports and a web service is available exposing the detection mode with the constructed motifs.

Organism	Entropy	Divergence	Qresiduals	TOTAL
Drosophila melanogaster	92	92	102	286
Homo sapiens	43	43	43	129
Rattus norvegicus	11	11	11	33
Mus musculus	25	25	25	75
TOTAL	171	171	181	523

Table 7.1: Summary of the models included for each organism and method to the models library of the MEET 5.1 R-package.

In order to build new models MEET 5.1 offers a training mode. In this mode, a leave-one-out (l.o.o.) cross validation studies the performance of the algorithms dependending on their parameters. It calculates the Receiving Operating Characteristics (ROC) curve and its Area Under Curve (AUC) to choose the parameters that provide with a high AUC with small variance in each step of the l.o.o. (see Table 7.2 and Table 7.3).

The training mode is also included in the MEET 5.1 package to allow the construction of models for motifs not included in the library. The constructed models can be directly used in the detection mode.
The architecture of MEET 5.1 can be seen in the Figure 7.1. All the implemented algorithms can be combined using the same input and output parameters.

Homo sapiens				
TF	**Sequences**	$C_{Qresiduals}$	$C_{Entropy}$	$C_{Divergence}$
ELK1	28	0.9988	0.9741	0.9904
RREB1	11	1	1	1
SRY	28	0.9988	0.9892	0.9932
TLX1-NFIC	16	1	1	1
RXRAVDR	10	1	1	0.9987
RORA-1	25	1	1	0.9998
ETS1	40	0.9953	0.9854	0.9947
E2F1	10	1	0.9992	0.9969
NKX31	20	0.9998	0.9948	0.9991
Irf-2	12	1	1	0.9977
Mus musculus				
TF	**Sequences**	$C_{Qresiduals}$	$C_{Entropy}$	$C_{Divergence}$
PPARgamma-RXRA	31	1	1	1
ARID3A	27	0.9977	0.9977	0.9956
ARNT	20	1	0.9995	0.9995
T	40	1	1	1
ELF5	44	0.9966	0.9965	0.9959
CREB1	16	1	1	0.9996
Hand1-Tcfe2a	29	0.9995	0.9957	0.9964
RUNX1	26	0.9999	0.9960	0.9974
NFkappaB	38	0.9999	0.9995	0.9972
Mycn	31	0.9994	0.9963	0.9964

Table 7.2: List of the first 10 TF for *Homo sapiens* and *Mus musculus* with the performance of each of the algorithms present in the MEET 5.1 models library, according to Equation 7.1.

Rattus norvegicus				
TF	**Sequences**	$C_{Qresiduals}$	$C_{Entropy}$	$C_{Divergence}$
AP1	18	0.9959	0.9678	0.9794
CREB1	16	1	0.9990	0.9995
FEV	13	1	1	0.9945
Foxd3	47	1	0.9977	0.9720
Foxq1	18	1	0.9994	0.9899
Mafb	15	0.9986	0.9811	0.9761
NFkappaB	38	1	0.9998	0.9952
NR3C1	9	0.9994	0.9881	0.9836
SP1	8	0.9996	0.9928	0.9875
NFATC2	26	0.9974	0.9556	0.9882
Drosophila melagonaster				
TF	**Sequences**	$C_{Qresiduals}$	$C_{Entropy}$	$C_{Divergence}$
Abd-B	21	0.9975	0.9931	0.9954
BH-2	22	0.9980	0.9907	0.9942
CG15696	33	0.9975	0.9796	0.9908
dl-2	23	0.9989	1	0.9994
Dr	21	1	1	1
opa	21	0.9998	0.9984	0.9970
ro	23	0.9999	0.9975	0.9984
Six4	22	0.9966	0.9802	0.9950
slp1	41	0.9983	0.9960	0.9944
ttk	22	1	0.9974	1

Table 7.3: List of the first 10 TF for *Rattus norvegicus* and *Drosophila melagonaster* with the performance of each of the algorithms present in the MEET 5.1 models library, according to Equation 7.1.

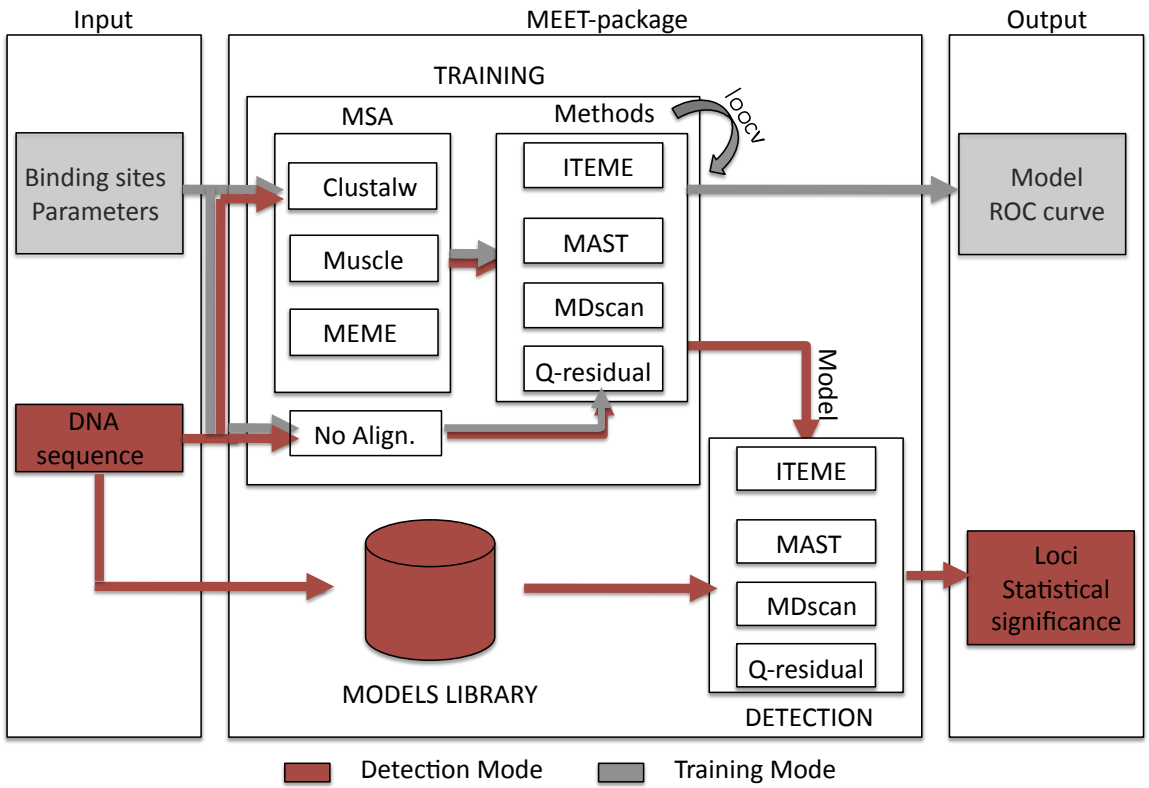


Figure 7.1: Description of the MEET 5.1 architecture including the functionalities and the included internal and external algorithms.

7.3 Training mode

To allow the use of motifs not included in the library, MEET 5.1 includes a training mode. Using this mode it is possible to optimize the parameters of a detector and to output a model compatible with the detection mode. The inputs of the training mode are the binding sequences and also a background sequence in fasta format (a DNAmeet.afa background sequence of 1500 nulceotides is included in the package). If the binding sequences to construct the model are not aligned, MEET 5.1 allows the use of two alignment algorithms MUSCLE [31] and CLUSTALW [121] when they are installed in the computer. MEME [4] can also be used as a motif discovery algorithm if the input are fasta sequences of corregulated genes.

The optimization of the parameters is performed using a double l.o.o..This procedure can be performed for a range of parameters, and the best model is chosen in a heuristic procedure using Equation 7.1 [83] which takes into account the highest AUC and also the stability of the detection.

$$C = \mu(AUC)(1 - \sigma(AUC)), \tag{7.1}$$

where $\mu(AUC)$ is the mean of the AUCs for all the steps of the l.o.o. and $\sigma(AUC)$ is the variance.
The function Construct model calls one of the algorithms to perform the double l.o.o. Then the ROC curve and the AUC are computed and these results are used to create the best model. The output is the best model, the AUC and the ROC curve corresponding to the best parameters.

The next R code is an example of how to run MEET 5.1 in training mode, using the Q-residuals detector and AP1 binding sites of *Homo sapiens*. In the code, *TF* are the AP1 binding site sequences in fasta format needed to construct the model, *seqin* is a DNA background sequence needed for the training mode, *alg* is the desired alignment, in this case the sequences are previously aligned, *mode* refers to the training or detection, *org* refers to the background organism (*Homo sapiens*), *method* is the algorithm used and finally *vector* indicates the parameters that we want to evaluate. In the example a number of components Q-residuals from 1 to 8 is studied to look for the optimal model.

```
library("MEET")
pathMEET <<- system.file("sequences", package = "MEET")
TrainingResult <<- MEET(TF = paste(pathMEET, "AP1.fa"),
                        seqin = paste(pathMEET, "DNAhomo.fa"),
```

```
                        alg = "NONE",
                        mode = "training",
                        org = "Homo sapiens",
                        method = "Qresiduals",
                        vector = c(1:8))
```

The output is a list that can be divided in three parts: two generic parts which have the consensus sequence of the motif and the input parameters of the MEET function (organism, algorithm, etc.) and the third part that contains the results.

```
print(TrainingResult$$Consensus)

 [1] "w" "b" "h" "n" "k" "v" "r"

print(TrainingResult$Results$model)
```

The results part is also a list which incorporates the chosen model, the AUC for the range of parameters studied and the ROC curve of the chosen model.

```
model
Importance of component(s):
                  PC1    PC2    PC3    PC4     PC5    PC6    PC7
R2             0.2844 0.1813 0.1520 0.1218 0.08648 0.0764 0.0432
Cumulative R2 0.2844 0.4657 0.6178 0.7396 0.82608 0.9025 0.9457
21  Variables
18  Samples
0  NAs ( 0 %)
7  Calculated component(s)
Scores structure:
[1] 18  7
Loadings structure:
[1] 21  7
Numerical Matrix
      [,1]     [,2]    [,3]
A  0.00000  0.00000  1.0000
C -0.47140  0.81650 -0.3333
G -0.47140 -0.81650 -0.3333
T  0.94281  0.00000 -0.3333
- -0.04243  0.07348  0.0600
a  0.00000  0.00000  1.0000
c -0.47140  0.81650 -0.3333
```

```
g -0.47140 -0.81650 -0.3333
t  0.94281  0.00000 -0.3333
JacksonPars$h0
0.2955
JacksonPars$x1
1.54
JacksonPars$x2
-0.925
JacksonPars$x3
3.987
ncolTFBS
7
```

The AUC and the ROC curve can be used to compare the performance of different detectors, and also to compare the AUC of the studied detector in the range of studied parameters. This allows the user to have another criteria to choose the optimal model and to build a custom motif detector.
The chosen model can be easily recovered from the MEET results. If the user prefers to visualize how the performance of the detector changes as the main parameter is changed, a simple boxplot of the AUC can be helpful to visualize the mean and the variance of the AUC using each one of the parameters. In the example above, with the Q-residuals detector and the AP1 motif from *Homo sapiens*, the following text will recover the model and plot the AUC for the number of principal components going from 1 to 10 as it can be seen in the equation 7.2.

As an example, the training mode of MEET 5.1 R-package can be used to compare the performance of the different searching algorithms in 9 transcription factors from the JASPAR database [17]: AP1, E2F1, ETS1, HLF, NFLI3 from *Homo sapiens*, ARNT, FOXO3, NFκB, SPZ1 from *Mus musculus* and ROX1 from the organism *Saccharomyces cerevesiae* from TRANSFAC 7.0 (2005) database [135].

As a background, a promoter sequence of each one of the organisms has been chosen randomly. For human and mouse the background has been extracted from Eukaryotic Promoter Database (EPD) [99], and the EPD version based on the EMBL release 105 (sept 2010) has been used. The range of nucleotides goes from the positions -1000 to 500 relative to the transcription start site (TSS) from $Igk'T$ gene in mouse and $RPS9P2+$ gene in humans. In *Saccharomyces cerevesiae* the nucleotides corresponding to the positions 44730-46230 in chromosome 1 were used, and they can be found in EMBL chromosome database [53], release 94 (march 2008).

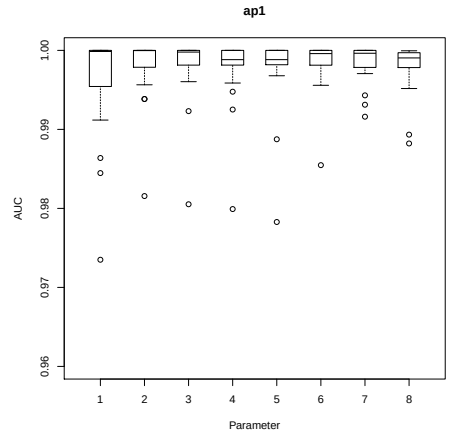


Figure 7.2: Boxplot of the AUC for Q-residuals detector and the AP1 motif from *Homo sapiens*.

The range of parameters to study in Q-residuals is a set from 1 to 10 principal components. In MATCH, the Core Similarity has been explored from 0 to 0.95 each 0.05, and, finally, in both Divergence and Entropy from ITEME the values $q = 0.1, 0.2, 0.5, 1, 2$ have been studied. The results can be observed in Table 7.4 where the mean AUC for all the algorithms is shown.

7.4 Detection mode

The detection mode of the MEET R-package can be used to look for binding sites within genomic sequences. The input can be (1) one of the models included in the library (2) one model constructed using the training mode (3) the parameters needed to construct one model. As in the case of the training mode, the generic function `Detection()` calls a specific function for one of the algorithms. It can be directly a prediction function which looks for binding sites or, in the case the inputed values are the parameters, first a model function. When the prediction function has looked for binding sites within the inputed problem sequence, the ouput given is: the sequences of the binding sites found, their p-value and their position within the larger sequence. If the searched binding sites belong to the models included in the MEET library the found sequences can also be visualized with a generated HTML file, using the function `writeResultsHTML()`.

Comparison Results					
TF	**Qresiduals**	**Entropy**	**Divergence**	**MATCH**	**MAST**
AP1	0.9893	0.9921	0.979	0.9868	**0.9925**
E2F1	0.9998	0.9979	0.9992	0.9995	**0.9999**
ETS1	0.9965	0.9956	**0.9972**	0.9922	0.9931
HLF	**0.9985**	0.9974	0.9965	0.9953	0.9688
NFLI3	0.9993	0.9992	0.9997	0.9980	**0.9999**
ARNT	0.9998	0.9998	0.9998	**1**	0.9999
FOXO3	0.9914	0.9747	0.9663	0.9765	**0.9947**
NFκB	**0.9998**	0.9747	0.9663	0.9765	0.9865
SPZ1	0.9944	0.9931	**0.9960**	0.9910	0.9913
ROX1	**0.9999**	0.9992	0.9941	0.9997	0.9937

Table 7.4: Table with the comparison of the performance of the detectors included in MEET 5.1 using 10 sets of transcription factor binding sites in JASPAR and TRANSFAC database and backgrounds corresponding to promoters of each organism (human, mouse and yeast). The result shown is the mean of the AUC for each TFBS and each method. The best method depends on the binding sites.

In the next example, the model obtained with the training method and the Qresiduals algorithm shown above is used for the detection of the AP1 binding sites in an *Homo sapiens* promoter. As the output of the training mode is directly used as a model for the detection mode there is no need to include the parameters of the algorithm. In the example, *seqin* is a DNA sequence with unknown binding sites, *mode* is detection, *model* refers to the built model using the training mode in the example above, *threshold* is the desired p-value threshold and *method* is the desired algorithm, in this case Q-residuals.

```
testAP1 <<- MEET(TF = paste(pathMEET, "AP1.fa", sep = "/"),
                 seqin = paste(pathMEET, "DNAmeet.afa"),
                 mode = "detection",
                 alg = "NONE",
                 model = FinalModel,
                 threshold = 0.1,
                 method = "Qresiduals")

print(testAP1$$Results)

    Position    Value       Direction    Sequence
```

```
1     "76"  "0.089673"     "f"       "TGAGTAAA"
2    "797"  "0.089673"     "f"       "TGAGTAAG"
```

In the next example we used the detection mode of the MEET 5.1 package in order to find annotated some TFBS in the UCSC Genome Browser on Human [28] database, to see Table 7.7. For each TF, we used candidate sequences from −1000 to 1000 relative to TFBS position, to see Table 7.5. Each background sequence has been extrated from Ensembl database version 2012 [35]. A brief summary of the candidate sequence for each transcription factor (TF) is shown in Table 7.5.

TF	Reference Genome	Chr	Start	End
PPARG	*GRCh37*	1	1.533.773	1.535.795
SRF	*GRCh37*	1	1.321.188	1.323.205
FOXO3	*GRCh37*	1	1.629.932	1.631.945
Pax6	*GRCh37*	1	1.607.019	1.609.039

Table 7.5: Candidate Sequence description.

```
  Position Value       Direction    Sequence
1  "1003"   "0"           "f"       "TGTAAACAT"
2  "291"    "0.0004983" "f"         "AGTTCACAC"
3  "686"    "0.0009965" "f"         "AGTAACCAG"
```

In Table 7.6, we have compared the results using some methods included in the library and the results obtained using another motif detection toolkits such as the rtfbs package [90] and TFBS::Site perl module [65]. The results show that MEET-package is comparable with rtfbs package and TFBS::Site perl modul in all the binding sites studied. The Detection Rank and the Run Time change according to TF and method considered, but in all the cases have the same order of magnitude.

A web service of the detection mode is publicly available through `http://sisbio.recerca.upc.edu:8080/`. This platform is mainly based on the Python language and is developed using a web framework named *web.py* (*http://webpy.org/*). In order to access R from Python in a simple and robust way it is used the RPy^2 package. The web pages are created in *HyperText Markup Language* (HTML) and, to make the user interface dynamic and user friendly, it is used JavaScript, *Asynchronous JavaScript And XML* (*AJAX*) and *JQuery* (*http://jquery.com/*), are employed to make the result similar to a dynamic online application rather

TF	Method	Score	Detection Rank	Run Time (s)
Pax6	MEET(Entropy)	$3.0 * 10^{-3}$	7	0.23
	rtfbs	21.46	6	0.75
	TFBS::Site	7.36	2	0.48
PPARG	MEET(Qresiduals)	0.96	10	0.32
	rtfbs	29.58	7	0.15
	TFBS::Site	4.16	6	0.49
SRF	MEET (Divergence)	$1.0 * 10^{-3}$	3	0.49
	rtfbs	19.59	2	0.15
	TFBS::Site	6.8	4	0.52
FOXO3	MEET(Entropy)	0	1	0.18
	rtbs	16.83	1	0.17
	TFBS::Site	12.15	1	0.49

Table 7.6: TFBS detection through MEET 5.1, rtfbs and TFBS::Site perl module. Detection Rank is the order that have been found the TFBS according to each method.

than a static Web site. The Figure 7.3 shows the configuration step where the user needs to upload or paste a DNA sequence in FASTA format, select one or more models provided by the application (*Transcription Factors*), select the detection algorithm (*Method*) and select the p-value used as the threshold in detection (*Threshold*). The models provided by the application are grouped by organism and each organism contains a set of TF that can be selected.

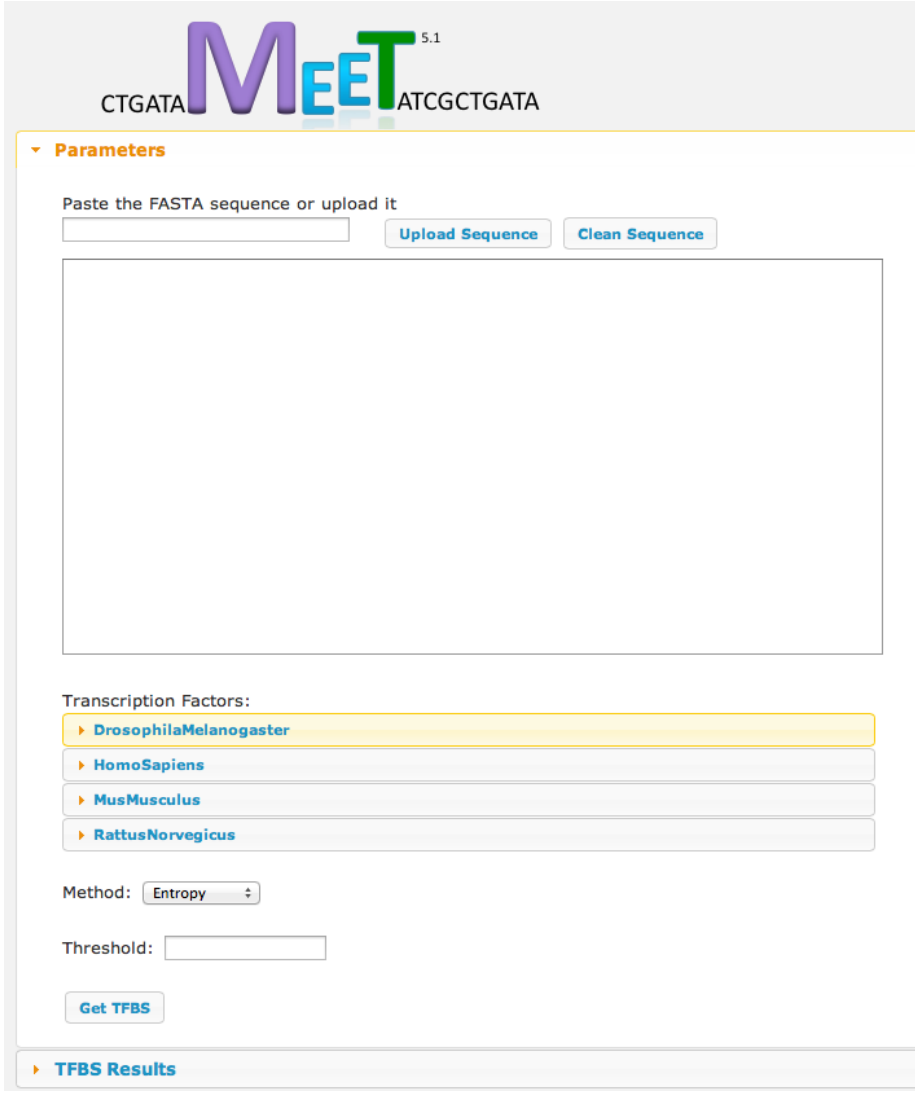


Figure 7.3: Initial view of the web of the MEET 5.1 R-package. The user can choose several motifs for each organism, paste or upload a sequence in .fasta format and then then the package will look for binding sites within the sequence.

TF	Chr	Organism	Position TFBS	Sequence
PPARG	1	*Homo sapiens*	1.858.134	ATGTAGGCCACCAGCAGGCA
SRF	1	*Homo sapiens*	1.322.191	ACCTAATATAG
FOXO3	1	*Homo sapiens*	1.630.932	TGTAAACA
Pax6	1	*Homo sapiens*	1.858.134	TGTAAACA

Table 7.7: HMR conserved Transcription Factor Binding Sites from UCSC Genome Browser on Human Feb. 2009.

Chapter 8

Results and Conclusions

8.1 Summary of Results

The characterization of the TFBS through the uncertainty measurement provides information about the variance of the set of nucleotide $\{A, C, G \text{ and } T\}$ in each binding sites position. Classically, this uncertainty has been measured by means of Shannon entropy. A generalisation of Shannon entropy has been proposed to characterize the transcription factor binding sites called Rényi entropy. This measurement depends on the Rényi order (q). From this measurement, we have estimated the redundancy profile. Biologically, the redundancy profile gives information on how much a particular position has been conserved on the sequences. As we have visualized, the redundancy profile depends on the Rényi order. As q increases, the noise in the redundancy also increases. With low q values the redundancy signal also decreases. Hence, the redundancy profile of the transcription factor depends on the Rényi order. We have just concluded that the optimal q-value is suggested as a tradeoff between the noise included in the redundancy signal and the attenuation of the same one. The Rényi order allows to modulate the amplitude and the number position that belong to a binding site. The nucleotides included in the binding profile are specific for each transcription factor due to the binding mechanism.

A first approximation, we have presented a non-linear method to detect the transcription factor binding sites assuming independency between positions in the binding sequence. The algorithm evaluates the variation on the total Rényi entropy of a set of sequences when a candidate sequence is assumed to be a true binding site belonging to the set. The parametrization provides two main advantages. First, it leverages the strong and weak symbol probabilities when computing the total entropy of the binding sequences, obtaining a detector with

variable sensibility. Secondly, the optimization of Rényi order allow to determinate the positions of the site involved in the binding process. Hence, the Rényi order depends on the TFBS characteristics and has to be adjusted for each TFBS. The results obtained has shown better performance than MDscan, which is a method based on deterministic models. The performance of the detector in comparison MEME depends on the binding site structure.

To consider the correlation among binding sites position, we have applied a linear and non-linear model to represent the set of binding sequences. The linear model is based on the residuals of the covariance of the numerical TFBS. This approximation has been demonstrated to be an effective method to detect TFBS. The performance has been compared against external algorithms. The results shown that when there are not correlation between positions, this method are comparable with results of PSSM methods. As the number of correlation between binding sites positions increases, the Q-residuals performance improves the results obtained for the other methods when the number of sequences is small, but it shows a larger sensitivity to the number of positions. These results prove that covariance can capture position interdependences in TFBS, and that a covariance-based model can be useful to detect TFBS within large databases. In general, the number of principal components that explain all the variance are between 1 and 4. This methodology shows a larger sensitivity to the number of positions, but it shows a significantly improvement on the performance with when the number of sequences is small. Q-residuals need more positions than Motifscan or PSSM to decrease the number of false positives. Instead, the computational time of the Q-residuals detector and PSSM based methods we found that Q-residuals is faster, in contrast with other methods that take into account interdependences which usually have a high computational cost.

The non-linear is based on the idea that total information content in a set of objects can be computed by means of divergence measurements. The information of a set of sequences is measured by means of α-Divergence which considers dependence among binding site positions. We have observed that the number binding sites correlated and its amplitude can be modulated by q-value (or α parameter). Large q-value will show large number of binding site dependence at the cost of introducing additional noise. The performance of the α-Divergence detector has been compared against a MEME/MAST and Rényi entropy. The results shown α-Divergence has a better behavior than the other detectors. Therefore, assuming position dependence modulated by q-value helps to improve over Entropy method and MEME/MAST. Given one Transcription Factor Binding Site, we can be observed how the number of true positives and false positives depends on the q-value. The best q-value can be chosen for the detection according to the cost criterion established between True Positive and False Positive and the area under convex surface maximum.

Both non-linear models, Rényi entropy and α-Divergence, have been integrated in one detector called SIGMA (Sequence Information Gain based on Motif Analysis). This new methodology allows us to detect cis-regulatory sequences with maximum performance regarding the conservation and the co-variability observed in the positions of the training set of sequences. The method focusses on the variation in these information measures when a new sequence is assumed to belong to a training set of sequences with known binding properties. SIGMA has been characterised on the detection problem for a large set of transcription factors and compared with different motif detection algorithms. AUC distributions have been calculated which show that there is a statistically significant difference between SIGMA performance and the performance of the other methods. In approximately 70% of the cases considered, SIGMA has exhibited better performance properties, at comparable levels of computational resources, than the methods with which it was compared.

All these tools for the detection of cis-regulatory sequences have been published into a R-packages. The MEET 5.1 library consists in 523 motif models from four different organisms *Drosophila melagonaster*, *Rattus norvegicus*, *Mus musculus* and *Homo sapiens*. The models have been built using three different algorithms, Q-residuals, Rényi Divergence and Rényi Entropy. This package includes three detectors and can access to some external detectors that can be executed and controlled directly from MEET 5.1. The package also includes an interface to external alignment algorithms. All the internal and external algorithms can be combined in order to optimize the best detection possible. The training mode of MEET 5.1 allows the direct comparison between algorithms, which can be carried out using not only the ROC curves and the AUC but also the error associated to them. MEET 5.1 has as an output the optimal model, which can be used in the detection mode in order to find unknown binding sites within a large sequence. The output of MEET 5.1 also allows the user to choose any model to run the detection mode. The package is documented and freely available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-projct.org., and A web interface of for the MEET package is also available

8.2 Conclusions

The research presented here has contributed to improve the computational representation and discovery of transcription factor binding sites in the following aspects.

- We have applied parametric uncertainty measurement, Rényi Entropy and Rényi Divergence, to evaluate the complexity of the nucleotides distribution

in the conserved sequence. To evaluate the variance of the set of nucleotide in each binding sites position, we have corrected the effect of finite sample size. The Rényi order modulates the amplitude and the number of positions involved to the binding sites. We have suggested an optimal q-value as a trade-off between the noise included in the redundancy signal and the attenuation of the same one.

- We have proposed a parametric detector using the Rényi Entropy for finding regulatory sequences in promoter regions. This measurement allowed us to build variable-sensitivity detectors modulated by the Rényi order. This methodology assumed the independence among binding sites. This approximation has shown better performance than others algorithms when the correlation between binding sites is null.

- The correlation among binding sites positions have been considered through linear and non-linear model.

- Q-residuals used a linear embedding to represent the set of binding site sequences and employed a residuals-based approach as the detection statistic. Q-residuals detector performs significantly better and faster than MATCH and MAST in most of the studied transcription factor binding sites. Compared to Motifscan, a method that take into account interdependences, the performance of the Q-residuals detector is better when the number of available sequences is small.

- α-Divergence is a non-linear model based on the variation on the correlation among binding site position when a new sequence is added to the set. We have observed that the number binding sites correlated and its amplitude can be modulated by q-value (or α parameter). The best q-value can be chosen for the detection according to the cost criterion established between True Positive and False Positive and the area under convex surface maximum. Given an optimal q-value, the performance α-Divergence has a better behavior than Rény algorithm and MAST.

- Both non-linear models based on Information Theory has been integrated in one detector called SIGMA. This new approach aims at a trade-off between the good generalisation properties of pure entropy methods and the ability of position-dependency metrics to improve detection power. In approximately 70% of the cases considered, SIGMA has exhibited better performance properties, at comparable levels of computational resources, than the methods with which it was compared.

- A set of tools was coded as an R package named MEET 5.1. The core of the package relies in 523 models carefully built and a set of tools for

cis-regulatory detection. The models allow the detection of cis-regulatory sequences in different organisms and different tools allow a directly comparison among algorithms on the same dataset. The package is documented and freely available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-projct.org.

Chapter 9

Publications

9.1 Indexed Journal Papers

- J. Maynou, E. Pairó and A. Perera.**Sequence Information Gain based on Motif Analysis**.BMC Bioinformatics 2015, 16 :377 (9 November 2015)
- E. Pairó, J. Maynou, S. Marco and A. Perera. **A subspace method for the detection of transcription factor binding sites**. Bioinformatics (Oxford, England), 28(10):1328–35, May 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts147.
- J.Maynou, M. Vallverdú, F. Clarià, J.J. Gallardo-Chacón, P. Caminal and A. Perera. **Computational detection of Transcription Factor Binding Site using a parametric entropy measure**. IEEE Trans. Information Theory, vol. 56, no. 2, pp: 734-741, Feb. 2010.

9.2 International Conference

International Conferences

- E. Pairó, J. Maynou, M. Vallverdú, P. Caminal, S. Marco and A. Perera, **MEET: Motif Elements Estimation Toolkit**. 33^{st} Annual International Conference of the IEEE Engineering in Medicine and Biology Society.
- J. Maynou, M. Vallverdú, F. Clarià, J.J. Gallardo-Chacón, P. Caminal and A. Perera. **Transcription Factor Binding Site Detection through**

Position Cross-Mutual Information variability analysis. 31^{st} Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

- J. Maynou, M. Vallverdú, F. Clarià, A. Perera and P. Caminal. **Detection of transcripction factor binding sites using Rényi Entropy**. 8th IEEE International Conference on BioInformatics and BioEngineering, 2008.

9.3 National Conferences

- J. Maynou, E. Pairó, R. Massanet, M. Vallverdú, P. Caminal y A. Perera. **Caracterización y análisis de las interacciones de regulación entre los factores de transcripción y los genes**, XXIX Congreso Anual de la Sociedad Española de Ingeniería Biomédica.
- J. Maynou, M. Vallverdú, P. Caminal y A. Perera. **Algoritmo de búsqueda de secuencias cis-regulatorias basado en el análisis del incremento de la información mediante la divergencia de Rényi**, XXIX Congreso Anual de la Sociedad Española de Ingeniería Biomédica.
- E. Pairó, J. Maynou*, M. Vallverdú, J.J. Gallardo-Chacón, P. Caminal, S. Marco and A. Perera. **MEET: Motif Elements Estimation Toolkit**, XXVIII Congreso Anual de la Sociedad Española de Ingeniería Biomédica.
- J. Maynou, M. Vallverdú, F. Clarià, J.J. Gallardo-Chacón, P. Caminal and A. Perera. **Detección de los puntos de unión de los factores de transcripción mediante el análisis de la variabilidad de la información mutua cruzada**. XXVII Congreso Anual de la Sociedad Española de Ingeniería Biomédica.
- J. Maynou, M. Vallverdú, F. Clarià, A. Perera and P. Caminal, **Detección de los puntos de unión de los factores de transcripción mediante la medida de la entropía de Rényi** . XXVI Congreso Anual de la Sociedad Española de Ingeniería Biomédica.

9.4 Software published on open source license

- MEET R package containing several motif model and algorithms to detect transcription factor binding sites (Chapter 6). The package is available as a contributed package from the Comprehensive R Archive Network (CRAN). A web interface of for the MEET package is also available.

Part III

Appendix

Chapter 10

Appendix A: Database Description

10.1 Transcription Factor Databases

The transcription factor databases store information about transcriptional regulation. These databases are classified according to the kind of stored information and the considered organism. A brief summary of the different databases is shown in table 10.1 [1].

The dataset used in this thesis comes from TRANSFAC and Jaspar databases. TRANSFAC[2] is a database about eukaryotic transcription regulating DNA sequence elements and the transcription factors binding to and acting through them [135]. All information about transcription factors is classified in six tables: SITES, GENE, FACTOR, CELL, CLASS and MATRIX. SITE gives information on (regulatory) transcription factor binding sites within eukaryotic genes. GENE gives a short explanation of the gene where a site (or group of sites) belongs to. FACTOR describes the proteins binding to these sites. CELL gives brief information about the cellular source of proteins that have been shown to interact with the sites. CLASS contains some background information about the transcription factor classes, while the MATRIX table gives nucleotide distribution matrices for the binding sites of transcription factors, as shown in table 10.2.

On the other hand, JASPAR[3] is open-access collection of transcription factor binding site (TFBS) matrices [128]. JASPAR contains different collections of

[1] http://bioinformatics.wikidot.com/transcription-factor-databases. Last consulted 2009

[2] http://www.gene-regulation.com.

[3] http://jaspar.binf.ku.dk/.

Table 10.1: Transcription Factor Databases

Database	Organism	Comment	Access
Transfac	*eukaryotic*	*Cis-acting regulatory DNA elements and transacting factors*	*Registration*
Jaspar	*eukaryotic*	*Trasnscription Factors modelled as matrices*	*Free*
TRRD	*Eukaryotic*	*Transcription Regulatory Regions Database*	*Free*
TRED	*Human, mouse, rat*	*Transcriptional Regulatory Element Database*	*Free*
Protein Lounge	*eukaryotic*	*Database of Transcription factors of humans and others organisms*	*Registration*
PLACE	*Plants*	*Database of Plant Cis-acting Regulatory DNA elements*	*Free*
SCPD	*Yeast (S.cerevisiae)*	*The Promoter Database of Saccharomyces cerevisiae*	*Free*
EPD Eukaryotic promoter database	*eykaryotic*	*Non-redundant collection of eukaryotic POL II promoters*	*Free*
RegulonDB	*Prokaryotic (E. coli)*	*Database of Transcription factor of Escherichia coli*	*Free*

Table 10.2: Information contents of TRANSFAC release 7.0 (2005).

Tables	Entries
SITE	7915
GENE	2397
FACTOR	6133
CELL	1307
CLASS	50
MATRIX	398

TFBS models derived by diverse approaches, but only three subsets is applying for genome analysis.

1. The JASPAR CORE database contains a curated, non-redundant set of profiles, derived from published collections of experimentally defined transcription factor binding sites for different organisms.

2. The JASPAR FAM partition houses familial 15 binding profiles (also referred to as consensus profiles) for 11 major structural classes of factors. The collection facilitates prediction of TF binding domain structures based on profile information alone.

3. JASPAR phyloFACTS. This new subset of the database contains a set of matrices that are derived from evolutionarily conserved sequences in the regulatory regions of mammalian genes.

Chapter 11

Appendix B: Glossary

1. **Adenine** is one of four chemical bases in DNA which encode the cell's genetic instructions. Adenine forms chemical bonds with thymine (T).
2. **Amino Acids** are a set of 20 different molecules used to build proteins.
3. **Binding site** is a region on a protein, DNA, or RNA to which to specific other molecules and ions form a chemical bond.
4. **bp** abbreviation of nucleotides or base pairs.
5. **Cis-regulatory element** is a region of DNA or RNA that regulates the expression genes. A cis-element may be located in several regions of the DNA: upstream or downstream of the gene's coding sequence and introns. Generally, a cis-regulatory elements are binding sites for several transcription factors.
6. **Chromosome** is an organized structure of DNA located in the nucleus of the cell.
7. **Cytosine** is one of for chemical bases in DNA which encode the cell's genetic instructions. Cytosine forms chemical bonds with guanine (G).
8. **DNA** Deoxyribonucleic acid. It is the molecule that carries genetic instructions in all living organisms.
9. **DNA-binding domain** is a protein domain that may contain one or several motif that recognize DNA.
10. **DNA replication** is the process by which DNA is duplicated.

11. **Enhancer** is a region of DNA located in the promoter domain that enhances the transcription levels of genes.
12. **Exon** is the portion of a gene that codes for amino acids.
13. **Gene** is the basic unit of inheritance.
14. **Gene expression** is the process by which the information encoded in a gene is used for the protein synthesis.
15. **Gene regulation** is the process of control of the gene information.
16. **Genome** is the set of hereditary material in a living organism.
17. **Genotype** is an individual's collection of genes.
18. **Guanine** is one of for chemical bases in DNA which encode the cell's genetic instructions. Guanine forms chemical bonds with cytosine (c).
19. **Insulator** is a region of DNA located in the promoter domain that its main function is enhancer-blocking.
20. **Intron** is the portion of a gene that does not code for amino acids.
21. **Nucleic acids** are macromolecules that its main functions are storage and expression of genetic information.
22. **Nucleotide** is the basic element of nucleic acids (RNA or DNA). Nucleotide is formed by a sugar molecule attached to a phosphate group and a nitrogen-containing base.
23. **Promoter** is a sequence of DNA located near the genes. Generally, promoter is upstream of the gene that regulates. Promoter is an essential element in the transcription process.
24. **Protein** is a amino acids sequence essential for all living organisms. Proteins have to do several functions in the cell: structural, mechanical, biochemical and cell signaling.
25. **RNA** Ribonucleic acid.
26. **Silencer** is a region of DNA located in the promoter domain that is capable of binding transcription regulation factors termed repressors.
27. **Thymine** is one of for chemical bases in DNA which encode the cell's genetic instructions. Thymine forms chemical bonds with adenine (A).
28. **Transcription** is the process of making an RNA copy of a gene sequence.

29. **Transcription Factors** are proteins that bind to specific DNA sequences to regulate transcription.

30. **Transcription Factor Binding Sites** is a specific region of the DNA that binds with a transcription factor.

31. **Translation** is the process of translating messenger RNA to an amino acids sequence.

Bibliography

[1] B Alberts, A Johnson, J Lewis, M Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell, Fourth Edition*. Garland, 2002.

[2] D Anastassiou. Genomic signal processing. *Signal Processing Magazine, IEEE*, 18(4):8–20, 2001.

[3] T L Bailey, M Boden, F A Buske, M Frith, C E Grant, L Clementi, J Ren, W W Li, and W S Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37(Web Server issue):W202–8, July 2009.

[4] T L Bailey and C Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36. Department of Computer Science and Engineering, and University of california at San Diego, and La Jolla, and California 92093-0114, AAAI Press, 1994.

[5] T L Bailey and C Elkan. The value of prior knowledge in discovering motifs with MEME. In *Proc. Int Conf, Intell. Syst. MOl. Biol.*, volume 3, pages 21–29. AAAI Press, 1995.

[6] T L Bailey and M Gribskov. Combining evidence using p-values: Application to sequence homoly searches. *Bioinformatics*, 14:48–54, 1998.

[7] T L Bailey, N Williams, C Misleh, and W W Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34(Web Server issue):W369–73, July 2006.

[8] B Balasubramanian, C V Lowry, and R S Zitomer. The Rox1 repressor of the Saccharomyces cerevisiae hypoxic genes is a specific DNA-binding protein with a high-mobility-group motif. *Mol Cell Biol*, 13(10):6071–6078, October 1993.

[9] Y Barash, G Elidean, N Friedman, and T Kaplan. Modeling dependencies in protein-dna binding sites. In *RECOMB*, 2003.

[10] I Ben-Gal, A Shani, A Gohr, J Grau, S Arviv, A Shmilovici, S Posch, and I Grosse. Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics (Oxford, England)*, 21(11):2657–66, June 2005.

[11] A Ben-Hur, C S Ong, S Sonnenburg, B Scholkopf, and G Ratsch. Support vector machines and kernels for computational biology. *PLoS computational biology*, 2008.

[12] M Blanchette and S Sinha. Separating real motifs from their artifacts. *Bioinformatics (Oxford, England)*, 17 Suppl 1:S30–8, January 2001.

[13] H Bolouri and E H Davidson. Modeling DNA sequence-based cis-regulatory gene networks. *Developmental biology*, 246(1):2–13, June 2002.

[14] E Boser, N Vapnik, Isabelle M Guyon, and T Bell Laboratories. Training Algorithm Margin for Optimal Classifiers. *Perception*, pages 144–152, 1992.

[15] S Boyd and L Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[16] A Brazma, I Jonassen, I Eidhammer, and D Gilbert. Approaches to the automatic discovery of patterns in biosequences. *J. Comput. Biol.*, 5(2):279–305, 1998.

[17] J C Bryne, E Valen, M E Tang, T Marstrand, O Winther, I da Piedade, A Krogh, B Lenhard, and A Sandelin. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Research*, 36(Database issue):D102–6, January 2008.

[18] M L Bulyk, P L F Johnson, and G M Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research*, 30(5):1255–61, March 2002.

[19] R Castelo and R Guigó. Splice site identification by idlBNs. *Bioinformatics (Oxford, England)*, 20 Suppl 1:i69–76, August 2004.

[20] E M Conlon, X S Liu, J D Lieb, and J S Liu. Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 100(6):3339–44, March 2003.

[21] P D Cristea. Conversion of nucleotides sequences into genomic signals. *Journal of cellular and molecular medicine*, 6(2):279–303, 2002.

[22] P D Cristea. *Genomic Signal processing and statistics*, chapter Representation and analysis of DNA sequences. Hindawi Publishing Corporation, 2005.

[23] M K Das and H K Dai. A survey of DNA motif finding algorithms. *BMC bioinformatics*, 8 Suppl 7:S21, January 2007.

[24] J Davis and M Goadrich. The Relationship Between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine learning*, Pittsburg, PA, 2006.

[25] A P Dempster, N M Laird, D B Rubin, Statistical Society, and Series B Methodological. Maximum Likelihood from Incomplete Data via the EM Algorithm. 39(1):1–38, 1977.

[26] P D'haeseleer. How does DNA sequence motif discovery work? *Nature biotechnology*, 24(8):959–61, August 2006.

[27] E R Doughterty, I Shmulevich, J Chen, and Z J Wang. Genomic Signal Processing and Statistics. *EURASIP Book Series on Signal Processing and Comunications*, 2005.

[28] T R Dreszer, D Karolchik, A S Zweig, A S Hinrichs, B J Raney, R M Kuhn, L R Meyer, M Wong, C A Sloan, K R Rosenbloom, G Roe, B Rhead, A Pohl, V S Malladi, C H Li, K Learned, V Kirkup, F Hsu, R a Harte, L Guruvadoo, M Goldman, B M Giardine, P a Fujita, M Diekhans, M S Cline, H Clawson, G P Barber, D Haussler, and W James Kent. The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Research*, 40(Database issue):D918–23, January 2012.

[29] I Dunham and et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[30] R. Durbin, S. R Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambrigde University Press, 1998.

[31] R C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–7, January 2004.

[32] K Ellrott, C Yang, F M Sladek, and T Jiang. Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics (Oxford, England)*, 18 Suppl 2:S100–9, January 2002.

[33] L Elnitski, V X Jin, P J Farnham, and S J M Jones. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome research*, (12):1455–64, December 2006.

[34] T Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, (27):861–874, 2006.

[35] P Flicek and et al. Ensembl 2012. *Nucleic Acids Research*, 40(Database issue):D84–90, January 2012.

[36] M C Frith, Ulla Hansen, John L Spouge, and Zhiping Weng. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Research*, 32(1):189–200, January 2004.

[37] B Georgi and A Schliep. Context-specific independence mixture modeling for positional weight matrices. *Bioinformatics (Oxford, England)*, 22(14):e166–73, July 2006.

[38] B Goebel, Z Dawy, J Hagenauer, and J C Mueller. An approximation to the distribution of finite sample size mutual infomation Estimates, 2005.

[39] D GuhaThakurta. Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Research*, 34(12):3585–98, January 2006.

[40] S Gunewardena and Z Zhang. A hybrid model for robust detection of transcription factor binding sites. *Bioinformatics (Oxford, England)*, 24(4):484–91, February 2008.

[41] H Lodish, A Berk, S L Zipursky, P Matsudaira, D Baltimore and J Darnell. *Molecular Cell Biology*. 2000.

[42] A L Halpern and W J Bruno. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular biology and evolution*, 15(7):910–7, July 1998.

[43] S Hannenhalli. Eukaryotic transcription factor binding sites-modeling and integrative search methods. *Bioinformatics*, 24(11):1325–1331, 2008.

[44] J Hawkins, C Grant, W S Noble, and T L Bailey. Assessing phylogenetic motif models for predicting transcription factor binding sites. *Bioinformatics (Oxford, England)*, 25(12):i339–47, June 2009.

[45] D T Holloway, M Kon, and C Deisi. Classifying transcription factor targets and discovering relevant biological features. *Biol Direct*, 3:22, 2008.

[46] D T Holloway, M Kon, and C Delisi. Integrating genomic data to predict transcription factor binding. *Genome informatics. International Conference on Genome Informatics*, 16(1):83–94, January 2005.

[47] P Hong, X S Liu, Q Zhou, X Lu, J S Liu, and W H Wong. A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics (Oxford, England)*, 21(11):2636–43, June 2005.

[48] J Hu, B Li, and D Kihara. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Research*, 33(15):4899–913, January 2005.

[49] C J Huberty. *Applied Discriminant Analysis*. John Wiley & sons, New York, USA, wiley seri edition, 1994.

[50] J D Hughes, P W Estep, S Tavazoie, and G M Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*, 296(5):1205–1214, 2000.

[51] J E Jackson. *A user's guide to Principal Components*, chapter 2, pages 36–40. John Wiley & Sons, Inc., 2004.

[52] B Jiang, M Q Zhang, and X Zhang. OSCAR: One-class SVM for Accurate Recogni- tion of Cis-elements. *Bioinformatics*, 0(0):1–7, 2007.

[53] C Kanz and et al. The embl nucleotide sequence database. *Nucleic Acids Research*, 33(suppl 1):D29–D33, 2005.

[54] A E Kel, E Gossling, I Reuter, E Cheremushkin, O V Kel-Margoulis, and E Wingender. MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*, 31(13):3576–3579, 2003.

[55] Tae-Min Kim and P J Park. Advances in analysis of transcriptional regulatory networks. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(1):21–35, 2011.

[56] O D King. A non-parametric model for transcription factor binding sites. *Nucleic Acids Research*, 31(19):116e–116, October 2003.

[57] A Krishnamachari, V moy Mandal, and Karmeshu. Study of DNA binding sites using the Rényi parametric entropy measure. *Journal of theoretical biology*, 227(3):429–36, April 2004.

[58] S Kullback and R A Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22:79–86, 1951.

[59] P Kumar. Generalized relative J-Divergence measure and properties. *Int. J. Contemp. Math. Sci*, 13:597–609, 2006.

[60] I Ladunga. *Computational Biology of Transcription Factor Binding*, volume 17. March 2010.

[61] M A Larkin, G Blackshields, N P Brown, R Chenna, P A McGettigan, H McWillian, F Valentin, I M Wallace, A Wilm, R Lopez, J D Thompson, T J Gibson, and D G Higgins. ClustalW and ClustalX version 2. *Bioinformatics*, 23(21):2947–2948, 2007.

[62] D S Latchman. *Principles of Medical Biology*, volume 5. JAI Press Inc., 1996.

[63] D S Latchman. *Eukaryotic Transcription Factors*. Academic Press, 5 edition, 2007.

[64] C E Lawrence, S F Altschul, M S Boguski, J S Liu, a F Neuwald, and J C Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science (New York, N.Y.)*, 262(5131):208–14, October 1993.

[65] B Lenhard and W W Wasserman. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics (Oxford, England)*, 18(8):1135–6, August 2002.

[66] N Li and M Tompa. Analysis of computational approaches for motif discovery. *Algorithms for molecular biology : AMB*, 1:8, January 2006.

[67] A Wee-Chung Liew, H Yan, and M Yang. Pattern recognition techniques for the emerging field of bioinformatics: A review. *Pattern Recognition*, 38(11):2055 – 2073, 2005.

[68] José J M Ter Linde and H Yde Steensma. A microarray-assisted screen for potential Hap1 and Rox1 target genes in Saccharomyces cerevisiae. *Yeast*, 19(10):825–840, July 2002.

[69] X Shirley Liu, Douglas L Brutlag, and Jun S Liu. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol*, 20(8):835–839, 2002.

[70] J. Locker. *Transcription Factors*. San Diego, bios, academic press edition, 2001.

[71] G Loots and I Ovcharenko. rVista 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Research*, 32:W217–W221, 2004.

[72] J Lu and L Luo. Prediction for human transcription start site using diversity measure with quadratic discriminant. pages 316–321, 2008.

[73] M Z Ludwig, N H Patel, and M Kreitman. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development (Cambridge, England)*, 125(5):949–58, March 1998.

[74] L Marsan. Extracting structured motifs using a suffix tree - Algorithms and application to promoter consensus identification. pages 210–219, 2000.

[75] J Maynou, JJ Gallardo-Chacon, M Vallverdú, P Caminal, and A Perera. Computational detection of transcription factor binding sites through differential rényi entropy. *Information Theory, IEEE Transactions on*, 56(2):734 –741, feb. 2010.

[76] J Maynou, M Vallverdú, F Clarià, JJ Gallardo-Chacon, P Caminal, and A Perera. Transcription factor binding site detection through position cross-mutual information variability analysis. In *31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society*, volume 2009.

[77] T Minka. Bayesian inference, entropy and the multinomial distribution. Technical report, Microsoft Research, 2003.

[78] B Morgenstern, S Goel1, A Sczyrba, and A Dress. AltAVisT: Comparing alternative multiple sequence alignments. *Bioinformatics*, 19(3):425–426, 2003.

[79] A M Moses, D Y Chiang, D A Pollard, V N Iyer, and M B Eisen. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome biology*, 5(12):R98, January 2004.

[80] R Mutihac, A Cicuttin, and R C Mutihac. Entropic approach to information coding in DNA molecules. *Materials Science and Engineering: C*, 18(1-2):51–60, December 2001.

[81] B T Naughton, E Fratkin, S Batzoglou, and D L Brutlag. A graph-based motif detection algorithm models complex nucleotide dependencies in transcription factor binding sites. *Nucleic Acids Research*, 34(20):5730–9, January 2006.

[82] H Pages, P Aboyoun, R Gentleman, and S DebRoy. *Biostrings: String objects representing biological sequences, and matching algorithms*, 2015. R package version 2.26.3.

[83] E Pairó, J Maynou, S Marco, and A Perera. A subspace method for the detection of transcription factor binding sites. *Bioinformatics (Oxford, England)*, 28(10):1328–35, May 2012.

[84] E Pairó, J Maynou, M Vallverdú, P Caminal, S Marco, and A Perera. MEET: motif elements estimation toolkit. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, volume 2011, pages 6483–6, January 2011.

[85] G Pavesi, G Mauri, and G Pesole. Methods for pattern discovery in unaligned biological sequences. 2(4):1–14, 2001.

[86] G Pavesi, G Mauri, and G Pesole. In silico representation and discovery of transcription factor binding sites. *Briefings in bioinformatics*, 5(3):217–36, September 2004.

[87] J Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers, 1988.

[88] K Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[89] A Perera, M Vallverdú, F Clarià, J M Soria, and P Caminal. DNA binding site characterization by means of Rényi entropy measures on nucleotide transitions. *IEEE Trans Nanobioscience*, 7(2):133–141, June 2008.

[90] N Peterson, A Martins, M Hubisz, and A Siepel. *rtfbs: R Transcription Factor Binding Site identification tool*, 2012. R package version 0.2.

[91] P Pevzner and S Sze. Combinatorial approaches to finding subtle signals in DNA sequences. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 269–278, 2000.

[92] E Portales-Casamar and et al. Jaspar 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 38(suppl 1):D105–D110, 2010.

[93] M Primig, H Winkler, and G Ammerer. The DNA binding and oligomerization domain of MCM1 is sufficient for its interaction with other regulatory proteins. *EMBO J*, 10(13):4209–4218, 1991.

[94] R Pudimat, EG Schukat-Talamazzini, and R Backofen. A multiple-feature framework for modelling and predicting transcription factor binding sites. *Bioinformatics (Oxford, England)*, 21(14):3082–8, July 2005.

[95] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[96] A Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Simposium on Mathematics, Statistics and Probability*, pages 547–561, 1961.

[97] A Gatenby Robert and B Roy Frieden. Information Theory in Living Systems, Methods, applications and Challenges. *Bulletin of Mathematical Biology*, 69:635–657, 2007.

[98] G K Sandve and F Drabløø. A survey of motif discovery methods in an integrated framework. *Biology direct*, 1:11, January 2006.

[99] C D Schmid, R Perier, and P Bucher. EDP in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Research*, 34:D82–85, 2006.

[100] T D Schneider. Information content of individual genetic sequences. *Journal of theoretical biology*, 189(4):427–41, December 1997.

[101] T D Schneider and R M Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, October 1990.

[102] T D Schneider, G D Stormo, L Gold, and A Ehrenfeuch. The Information Content of Binding Sites on Nucleotide Sequences. *J. Mol. Biol*, 188:415–131, 1986.

[103] B Scholkopf and A J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press Cambridge, 2001.

[104] C E Shannon. A mathematical theroy of communication. *The Bell System Technical journal*, 27:379–423 and 623–656, July and October 1948.

[105] E Sharon, S Lubliner, and E Segal. A feature-based approach to modeling protein-DNA interactions. *PLoS computational biology*, 4(8):e1000154, January 2008.

[106] I Shmulevich and E R Dougherty. *Genomic Signal Processing (Princeton Series in Applied Mathematics)*. Princeton University Press, Princeton, NJ, USA, 2007.

[107] A Siepel and et al. Evolutionarily conserved elements in vertebrate , insect , worm , and yeast genomes. pages 1034–1050, 2005.

[108] B D Silverman and R Linske. A measure of DNA periodicity. *Journal of Theoretical Biology*, 118:295–300, 1986.

[109] N Simonis, S J Wodak, G N Cohen, and J van Helden. Combining pattern discovery and discriminant analysis to predict gene co-regulation. *Bioinformatics (Oxford, England)*, 20(15):2370–9, October 2004.

[110] T Sing, O Sander, N Beerenwinkel, and T Lengauer. ROCR: visualizing classifier performance in R. *Bioinformatics (Oxford, England)*, 21(20):3940–1, October 2005.

[111] S Sinha and M Tompa. A statistical method for finding transcription factor binding sites. In *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, volume 8, pages 344–354, 2000.

[112] S Sinha and M Tompa. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research*, 30(24):5549–60, December 2002.

[113] J Song and H Tang. A new 2-D graphical representation of DNA sequences and their numerical characterization. *Journal of biochemical and biophysical methods*, 63(3):228–39, June 2005.

[114] W Stacklies, H Redestig, M Scholz, D Walther, and J Selbig. pcaMethods a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23(9):1164–1167, 2007.

[115] R Stéphane, R François, and S Schbath. *DNA, words and models.* Cambridge University Press, 2005.

[116] G D Stormo. Consensus patterns in DNA. *Methods Enzymol*, 21:183–211, 1990.

[117] G D Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[118] Y Sun, M Robinson, R Adams, R Boekhorst, A G Rust, and D Davey. Using feature selection filtering methods for binding site prediction. *IEEE International Conference on Cognitive Informatics*, 2006.

[119] DMJ Tax and RPW Duin. Support Vector Data Description. *Machine Learning*, 54:45–66, 2004.

[120] G Thijs, M Lescot, K Marchal, S Rombauts, B De Moor, P Rouzé, and Y Moreau. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics (Oxford, England)*, 17(12):1113–22, December 2001.

[121] J D Thompson, D G Higgins, and T J Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.

[122] A Tomovic. *Computational analysis of promoters and DNA-protein interactions*. PhD Thesis, University of Basel, Faculty of Science, 2009.

[123] A Tomovic and E J Oakeley. Position dependencies in transcription factor binding sites. *Bioinformatics (Oxford, England)*, 23(8):933–41, April 2007.

[124] M Tompa. An Exact Method for Finding Short Motifs in Sequences, with Application to the Ribosome Binding Site Problem. *Conf. Intelligent Systems for Molecular Biology*, 1999.

[125] M Tompa and et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, 23(1):137–44, January 2005.

[126] J van Helden, A F Rios, and J Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research*, 28(8):1808–18, April 2000.

[127] J P Vert and W S Noble. Kernels for gene regulatory regions. *Spectrum*, 2005.

[128] D Vlieghe, A Sandelin, P J De Bleser, K Vleminckx, W W Wasserman, F van Roy, and B Lenhard. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Research*, 34(Database issue):D95–7, January 2006.

[129] R F Voss. Evolution of long-range fractal correlations and 1/f noise in dna base sequences. *Physical Review letters*, 68(25):3805–3808, 1992.

[130] W W Wasserman and A Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature reviews. Genetics*, 5(4):276–87, April 2004.

[131] J D Watson and F H Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[132] W Wei and X-D Yu. Comparative analysis of regulatory motif discovery tools for transcription factor binding sites. *Genomics, proteomics & bioinformatics / Beijing Genomics Institute*, 5(2):131–42, May 2007.

[133] T W Whitfield, J Wang, P J Collins, E C Partridge, S F Aldred, N D Trinklein, R M Myers, and Z Weng. Functional analysis of transcription factor binding sites in human promoters. *Genome biology*, 13(9):R50, September 2012.

[134] F Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83, 1945.

[135] E Wingender, X Chen, R Hehl, H Karas, I Liebich, V Matys, T Meinhardt, M Prüss, I Reuter, and F Schacherer. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*, 28(1):316–319, 2000.

[136] C Yin. Numerical Representation of DNA Sequences Based on Genetic Code Context and Its Applications in Periodicity Analysis of Genomes. *DNA Sequence*, (2):5–9, 2008.

[137] M O Zhang and T G Marr. A weight array method for splicing signal analysis. *Comput. Appl. Biosci.*, 9(5):499–509, 1993.

[138] M Q Zhang. Computational prediction of eukaryotic protein-coding genes. *Nature reviews. Genetics*, 3(9):698–709, September 2002.

[139] X Zhao, H Huang, and T P Speed. Finding short DNA motifs using permuted Markov models. *Journal of computational biology a journal of computational molecular cell biology*, 12(6):894–906, 2005.

[140] Q Zhou and J S Liu. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics (Oxford, England)*, 20(6):909–16, April 2004.