In silico analysis of regulatory motifs in gene promoters

Nicolás Bellora Pereyra

PhD thesis

Barcelona, November 2009

UNIVERSITAT POMPEU FABRA
Departament de Ciències Experimentals i de la Salut

In silico analysis of regulatory motifs in gene promoters

Nicolás Bellora Pereyra

Memòria presentada per optar al grau de

Doctor en Biologia per la Universitat Pompeu Fabra.

Aquesta Tesi Doctoral ha estat realitzada sota la direcció de la

Dra. M. Mar Albà al Departament de Ciències Experimentals

i de la Salut de la Universitat Pompeu Fabra

M. Mar Albà **Nicolás Bellora Pereyra**

Barcelona, Novembre de 2009

The research in this thesis has been undertaken in the Evolutionary Genomics Group within the Grup de Recerca en Informàtica Biomèdica (GRIB) of the Parc de Recerca Biomèdica de Barcelona (PRBB), a consortium of the Institut Municipal d'Investigació Mèdica (IMIM), Universitat Pompeu Fabra (UPF) and Centre de Regulació Genòmica (CRG).

The research carried out in this thesis has been supported by a predoctoral fellowship from BBVA, a predoctoral contract from REEM, and research grants from the Spanish Government.

Abstract

Regulation of gene transcription is a complex process involving many different proteins, some of which bind in a sequence-specific manner to DNA motifs in the gene promoter. The need to maintain specific interactions between transcription factors and proteins involved in the RNA polymerase II complex is expected to impose constrains on the relative position and spacing of the interacting DNA motifs. The present work includes the development of a novel approach to identify motifs that show a preferential location in DNA sequences and the implementation of a public web application called PEAKS. We investigated if the arrangement and nature of the most common motifs depended on the breath of expression of the gene. We found differences that serve to illustrate that many key specific regulatory signals may be present in the proximal promoter region in mammalian genes. We also apply other methods for the identification of specific transcription factors (TFs) involved in the co-regulation of a set of genes. Data from experimentally-verified transcription factors binding sites (TFBSs) support the biological relevance of our findings.

Resum

La regulació de la transcripció dels gens és un procés complex que implica moltes proteïnes diferents, algunes de les quals s'unexien a motius específics d'ADN localitzats a la regió promotora dels gens. S'espera que la necessitat de mantenir les interaccions específiques entre els factors de transcripció i les proteïnes implicades en el complex de l'ARN polimerasa II imposi limitacions en la posició relativa i l'espaiat dels motius d' interacció amb l'ADN. La feina presentada en aquesta tesi inclou el desenvolupament d'un nou metode per l'identificació de motius que mostren una localització preferencial en seqüències d'ADN i l'implementació d'una aplicació web pública anomenada PEAKS. Hem investigat si la posició i la naturalesa de la majoria dels motius més frequents depèn del rang d'expresió del gen. Hem trobat diferències que serveixen per illustrar el fet que moltes senyals clau de regulació gènica poden estar presents en la regió proximal del promotor dels gens de mamífers. També hem aplicat altres mètodes per a l'identificació de factors de transcripció (TFs) específics involucrats en la co-regulació d'un grup de gens. Dades de llocs d'unió dels TFs (TFBSs) verificats experimentalment recolzen la rellevància biològica dels nostres resultats.

En ninguna estructura orgánica encontramos una forma geométrica pura, en ninguno de los ritmos de su vida una periodicidad exactamente calculable. Parece como si la Idea tuviera que sacrificar algo de su pureza y de su divinidad esenciales cada vez que quiere encarnarse en la naturaleza.

CARL GUSTAV CARUS (1789-1869)

Acknowledgements

En especial quiero agradecer a Mar por darme la oportunidad de empezar y de acabar este trabajo. A Roderic por su apoyo en los momentos difíciles. A Robert, Eduardo, Baldo, Jordi y Núria por su interés y su tiempo. A todos ellos por su paciencia. Tengo que agradecer a mucha gente que me apoyó durante estos últimos años. Desde los que me ayudaron a dar los primeros pasos: Pep, Enrique, Genis, Sergí, Charles... Hasta los que me guiaron en los últimos momentos: ...Domènec, Steve, Mireia, Macarena. A Xevi, Joan, Britta, Pedro, Thien, David y André por compartir cafés, pitis e inquietudes. A Òscar, Alfons, Miguel y Judith por mantener el tinglado funcionando y soportarme en "n" ocasiones. A la gente de Rode. A la Nurieta, Alice, Eneritz, Cristian, Alba y a Todos Los Compañeros con los que compartí alguna parte del camino. Y a los que me tuvieron que aguantar chapurreando inglés y me animaron a continuar, muchas gracias!

Contents

List of Tables

List of Figures

1

Introduction

Summary

In this chapter I introduce some biological concepts about transcriptional regulation. It also covers what is known about regulatory motifs present in eukaryotic promoter sequences. Finally, I also touch on some transcription factors features and the *in silico* prediction of their binding sites.

1.1 Overview

Gene expression varies in different cell types or in response to specific signals. According to the central dogma of biology, if genetic information that each individual inherits as DNA (the genotype) is to be converted into proteins, which is largely responsible for the characteristics of the individual (the phenotype), it must first be converted into an RNA product. RNA synthesis, or transcription, is the process of transcribing a particular region of a dsDNA sequence into a ssRNA sequence (see Figure 1.1). The mechanism of transcriptional regulation is orchestrated by proteins called transcription factors (TFs), which promote (as activators), or block (as repressors) the recruitment of the RNA polymerase II (Pol II) complex. Moreover, some RNA products have post-transcriptional regulatory properties. They include microRNAs (miRNAs) and small interfering RNA (siRNAs), the so-called non-coding RNA (ncRNAs) (Chen and Rajewsky, 2007). Therefore, regulation of gene transcription is the main control point in the regulation of gene expression and it depends on particular conditions and the cell type.

1.2 Complexity of transcriptional regulation

Context dependency

For a given locus there is a genomic regulatory context defined by several components and the resulting transcription is the output of many regulatory signals. The function of DNA-binding molecules, transcription factors and

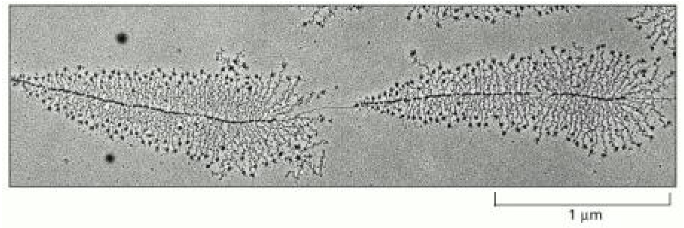


Figure 1.1: Transcription of two genes as observed under the electron microscope. Molecules of RNA polymerase are visible as a series of dots along the DNA with the newly synthesized transcripts (fine threads) attached to them. From the lengths of the newly synthesized transcripts, it can be deduced that the RNA polymerase molecules are transcribing from left to right. Adapted from (Miller and Beatty, 1969).

histones is always context dependent. They have specific affinities for DNA binding sites, and thus every sequence defines a unique affinity landscape with respect to each molecule. The molecules that interact with and bind to DNA result in a unique distribution of molecule-binding configurations at each sequence and lead to a transcriptional output (Segal and Widom, 2009).

Epigenetic control

Epigenetic mechanisms can ensure that differential expression patterns are stably inherited when cells divide. Basically, there are two epigentic mechanisms that are responsible of maintaining the heritable transcription states: chromatin remodeling and DNA methylation.

DNA is packaged into chromatin thereby constraining the size of the molecule that is approximately 2 meters of DNA per human cell. Chromatin represents a repeating unit of histones and DNA that form the nucleosome. Highly condensed chromatin (heterochromatin) is transcriptionally silent. Further specific patterns of histone tail modifications attract or repel regulatory proteins of the chromatin remodeling complex (Fischer et al., 2008). Gene-specific transcription factors bound to specific sites in the genome are known to recruit chromatin-remodeling factors and enzymes that covalently modify histones. This leads to the binding of other regulatory factors that act together with chromatin to create a permissive or non-permissive environment for gene expression (Hahn, 2008). During RNA synthesis, chromatin-modifying factors are associated with the elongation complex, ahead of RNA polymerase II, to generate a chromatin state that is permissive for transcription.

Methylation of DNA essentially leads to a repression of transcription by interfering with the binding sequence of transcription factors and through the binding of methyl-CpG binding proteins (MBD) (Wade, 2001). CpG islands (CGIs), short stretches of DNA that are often between 200 bp and 1 kb long, frequently contain unmethylated CpG dinucleotides in vertebrates. They are transcriptionally active genomic regions that contain multiple transcription start sites (TSSs) (Juven-Gershon et al., 2006).

Promoter regions

The RNA polymerase preinitiation complex, together with TFs, binds to regions upstream to the coding sequence, called gene promoter regions.

Transcription adapter proteins help recruiting basal transcription factors that are assembled in the pre-initiation complex (PIC) (see Figure 1.2). The RNA polymerase II complex is positioned at the transcription start site (TSS) by the PIC. The basal promoter (or core promoter) is the region where the RNA polymerase complex is recruited. The rest of the promoter corresponds to

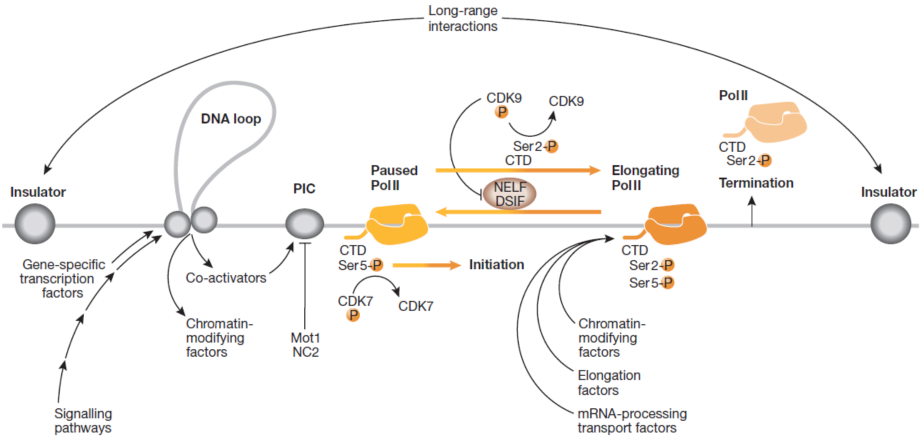


Figure 1.2: Transcriptional regulation and the pathway of RNA polymerase II transcription. Signaling pathways activate gene-specific transcription factors that bind to gene-regulatory regions. Distant factors can interact with more proximal factors through DNA looping. Transcription factors can recruit chromatin-modifying factors and transcriptional coactivator complexes. The co-activators function to recruit, and possibly to stimulate, the activity of the transcription pre-initiation complex (PIC). PIC assembly can be modulated by specific repressors (Mot1 and NC2). On initiation of transcription, the RNA polymerase II (Pol II) carboxy-terminal domain (CTD) is phosphorylated by a cyclin-dependent kinase (CDK7). Transcription elongation in higher eukaryotes is often blocked to generate a paused polymerase. The elongating Pol II with the CTD phosphorylated at specific serines recruits several factors that are involved in chromatin modification, transcription elongation, messenger RNA (mRNA) processing, mRNA transport and termination. Long-range chromatin interactions, for example between insulator elements, isolate chromatin domains to prevent the spread of regulatory signals. Taken from (Hahn, 2008).

transcription factor binding sites (TFBS) that confer specificity to transcription. In eukaryotes, there is no transcriptional activity from a promoter in absence of specific transcription factors. In consequence, by default the transcription is off (Wray et al., 2003). All promoters contain TFBSs for activators of the transcription, however only some contain TFBSs for repressors (Davidson, 2001).

Transcription factors can recruit chromatin-modifying factors, as well as transcriptional coactivator complexes, to the vicinity of the transcription start sites. The three-dimensional structure of DNA plays an important role in determining promoter regions. It has been proposed that highly flexible regions can position nucleosomes just downstream of the TSS (Pedersen et al., 1998). Recently, it has been shown that the nucleosome phasing relative to the TSSs is directly correlated with RNA polymerase II (Pol II) binding in humans (Schones et al., 2008).

Most core promoters do not have a single TSS but rather an array of closely located initiation sites. Basically, two positional distributions of TSSs are found in promoters, these with a single dominant peak, and these with a general broad distribution (Kawaji et al., 2006). Transcription from single peak promoters occurs from a single TSS or a localized cluster of TSSs in less than 10 bp. However, transcription from CpG islands initiates from multiple weak start sites that are often distributed over a region of about 100 bp (Juven-Gershon et al., 2006). This is conceptually different from alternative promoters, in which core promoters are separated by clear genomic space. Alternative promoters upstream from the coding sequence (CDS) can be more or less active under diverse cell conditions. Furthermore, it has been shown that differentially regulated alternative promoters are a common feature in protein-coding genes (Carninci et al., 2006).

Combinatorial control by transcription factors

For each gene, there are one or more control regions upstream or downstream of the transcription start site (TSS): promoters, enhancers, silencers and insulators (see Figure 1.2). Usually, enhancer regions are thousands of base pairs upstream or downstream from the promoters that they controls (Atchison, 1988). Transcription factors that bind to enhancers increase the rate of transcription of the gene. However, transcriptional enhancers located in introns can also affect both the Pol II elongation rate and alternative splicing (Kadener et al., 2002). Insulators prevent the actions of an enhancer from acting on the promoter of a downstream gene. When specific TFs bind to a given genomic region and block transcription or decrease the transcription rate, we use the term of silencer regions.

All these regulatory regions show a modular organization and each one is formed of one or more discrete regions called cis-regulatory modules (CRMs) (Yuh et al., 1998). CRMs produce the spatio-temporal expression patterns by 'reading-out' the concentrations of multiple TFs in specific cell conditions. Each CRM contains a cluster of transcription factor binding sites. We can consider these transcription factor binding sites (TFBSs) as the elementary units of this modular organization.

In many cases, TFs interacts in synergy, which means that their combined effect is larger than the sum of their individual effects. In other cases, antagonistic effects occur when TFs are able to bind to overlapping sites (Masquilier and Sassone-Corsi, 1992). Indeed, the regulatory competition between TFs occurs because in many cases they can recognize the same sites, albeit with different affinities. Gene transcription repressors can function by competing for DNA binding with activators, by masking the activation interface (Wang et al., 1997) or by direct interaction with general transcription factors (Song et al., 1995).

Down-regulation by ncRNAs and NMD

For many years the term ’gene’ has been synonymous for genomic regions encoding messenger RNA (mRNAs) that are translated into protein. However, recent genome-wide studies have shown that plant and animal genomes are pervasively transcribed and produces thousands of regulatory non-protein-coding RNAs (ncRNAs) (Yazaki et al., 2007). Short interfering or silencing RNAs (siRNAs) and microRNAs (miRNAs) are components of the RNA based mechanism of gene regulation, and can silence genes at the transcriptional and post-transcriptional levels. Mature miRNA molecules are either fully or partially complementary to one or more messenger RNA (mRNA) molecules, and their main function is to post-transcriptionally down-regulate gene expression. Recent studies indicate an alternative regulatory pathway is operative in diverse organisms, including plants and metazoans, where siRNAs have been shown to mediate transcriptional gene silencing (TGS) (Morris, 2008). TGS is archieved by the anti-sense strand of the siRNA targeting chromatin remodeling complexes to the specific promoter region. This siRNA targeting results in epigenetic modifications, rewriting of the local histone code and silent state chromatin marks (Pikaard, 2006).

Once in the cytoplasm, mRNAs are further subjected to translation-dependent surveillance. Here a process called nonsense-mediated mRNA decay (NMD) provides a way to degrade abnormal mRNAs that encode potentially deleterious truncated proteins. It has been estimated that one-third of naturally occurring alternatively spliced mRNAs are also targeted for NMD, potentially providing an additional mechanism to maintain correct levels of gene expression (Lejeune and Maquat, 2005).

The ENCyclopedia Of DNA Elements (ENCODE)

The ENCyclopedia Of DNA Elements (ENCODE) project is an international consortium that aims to identify all functional elements in the human genome sequence (ENCODE Project Consortium, 2004). During the year 2009, the ENCODE project released the first freeze (November 2008) of whole-genome experimental data produced for the production phase. They have included transcription factor binding sites of 12 transcription factors and RNA polymerase II in 7 cell types, histone modifications, DNA methylation, insulators and transcription maps.

We are only beginning to understand how to integrate all these regulatory signals into complex regulatory circuits. The results of ENCODE project will help us to understand the relations between cell conditions, cell pathways and genome-wide regulatory landscapes to characterize the regulatory circuits.

1.3 Transcription factors

Transcription factors (TFs) are proteins that bind to short DNA sequences (5-20 bp.). They can activate or repress the recruitment of Pol II to promoter regions, acting alone or as part of a protein complex. Usually, promoter regions contain 10 to 50 binding sites for 5 to 15 different transcription factors (Arnone and Davidson, 1997).

Cis and *trans* regulatory elements

A cis-regulatory element is a region of DNA involved in the regulation of gene expression. TFBSs are cis-regulatory elements mainly organized into CRMs. In contrast, transcription factors are trans-regulatory elements (or trans-activating factors) that interact with CRMs. The functionality of these cis-regulatory elements depends both on their accessibility and the relative amount of active transcription factors.

There are two basic mechanisms to regulate the potential activity of TFs: controlling the synthesis of the factor or regulating its activation. Synthesis and degradation are the basic mechanisms to control cell concentrations of TFs. In addition, alternative splicing (Shen et al., 1991) or translational regulation (Morris and Geballe, 2000) results in TF isoforms with distinct regulatory functions. However, protein synthesis is metabolically expensive and does not have the necessary rapid response time required for the regulation of inducible gene expression. Protein-ligand binding, protein-protein interactions and protein phosphorylation are cell mechanisms to rapidly regulate the function of TFs.

Phenotypic diversity between organisms may arise from changes in the regulation of gene expression, in addition to differences in the gene repertory (King and Wilson, 1975). A particular TF family may have different roles among eukaryotes, whereas others are specific to particular lineages. Comparative studies have shown that, for many TFs, the DNA-binding domain is highly conserved among eukaryotes while the remaining protein sequence is often very divergent (Riechmann et al., 2000). Protein-protein interaction and activation domains are usually located in the less conserved protein regions. These domains include amino acid (AA) tandem repeats, which can rapidly change their length due to the action of replication slippage (Fondon and Garner, 2004). In any case, most of the changes in regulatory networks are likely to occur in cis; changes in trans (transcription factors) may often have too strong effects (Wray et al., 2003).

Diversity of DNA binding domains

The genome of multicellular organisms contains a large number of diverse genes encoding transcription factors. Each of which has one or more DNA binding domains (DBDs) that define its sequence-specificity. They can be classified into families according on to structure of their DNA-binding domains, which can provide clues to their functions; for example, homeodomain containing TFs are often associated with developmental processes, and those in the interferon regulatory factor families are generally associated with triggering immune responses to viral infections (Luscombe et al., 2000).

Luscombe *et al*, classified eukaryotic and prokaryotic DNA-binding proteins into eight different structural groups, which further can be classified into 54 families (see Table 1.1). Into this classification they include TFs, histones, polymerases and enzymes. Many relevant families of vertebrate TFs were classified in 6 groups. The 'β-sheet' group only contains the TATA-box binding protein. The High movilty group (HMG) and MADS-box are members of the 'Other α-helix' group that also include histones. The 'Enzyme' group does not contain TF families and clusters diverse enzymes that do not have a common structural motif for DNA binding.

Table 1.1: Groups of DNA binding domains

DBD group	Fam	PP	EP	TFs family examples
1. Helix-turn-helix	16	32	28	homeodomain, ETS, TFIIB
2. Zinc-coordinating	4	0	23	C_2H_2 ZF, hormone receptor
3. Zipper-type	2	0	10	bZIP, helix-loop-helix
4. Other α-helix	7	1	5	HMG, MADS-box
5. β-sheet	1	0	8	TBP
6. β-hairpin/ribbon	6	10	1	T-box
7. Other	2	0	8	REL, Stat
8. Enzyme	16	43	68	

Fam = number of protein families in the DBD group, PP = number of prokaryotic proteins in the group, EP = number of eukaryotic proteins in the group, TFs family examples = examples of relevant vertebrate TFs families. Adapted from (Luscombe et al., 2000).

The proportion of TF encoding genes is different between organisms. Comparative studies of diverse archea, bacterial and eukaryotic genomes show that the number of TFs increases more rapidly relative to the total number of genes (van Nimwegen, 2003). A recent census estimates that there are around 1,400 transcription factors encoded in the human genome, with three families dominating 80% of the repertoir: C2H2 zinc-finger (675 TFs), homeodomain (275 TFs) and helix-loop-helix (87 TFs) (Vaquerizas et al., 2009).

Protein-DNA interactions

The structure of the DNA binding domain determines how it interacts with the DNA recognition motif. Moreover, other protein regions may also interact with the DNA molecule to stabilize the complex (Kurokawa et al., 2009). Hydrogen bridges and van der Waals force are the main determinants of the specificity of the amino acid sequence to a particular DNA sequence. However many TFs interact with the DNA as dimers or complexes, which in turn affects their protein structure and binding domain exposure.

Recent *in vitro* experiments were performed to determine the binding specificities of several mouse TFs. This technique is called protein binding microarray analysis (PBM) and is used to determine the binding affinities to all possible 8-mers in a unbiased manner. PBMs contain 60-mer probes in which 8-mers occur several times in different sequence context (Berger et al., 2006). The DNA binding affinities of 168 homeodomain TFs revealed sequence preferences and correlation between amino acid sequence of the DBD and the DNA binding profile. Using the same technique, Badis and colleagues found subtle preferences in transcription factor binding affinities, dependencies between positions, and alternative usage of protein binding-domains in 104 TFs that contained 22 DNA binding domain classes (Badis et al., 2009).

In general, different binding domain classes recognize different portions of the sequence space. Although many TFs from a binding domain class bound to the same highest-affinity 8-mers, they preferred different lower-affinity sites (see Figure 1.3). They also observed clear secondary DNA binding preferences for nearly half of the 104 mouse TFs. The secondary motifs could be

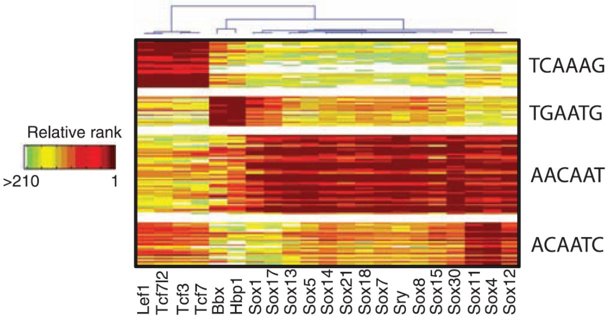


Figure 1.3: Clustergram of k-mers for the Sox family of transcription factors. 310 8-mers with significant scores for at least one of the 21 TFs. The rows corresponding to 8-mers were arranged to group those with shared sequence patterns. Adapted from (Badis et al., 2009)

classified into four different categories: variable spacer length, position interdependence, multiple effects and alternate recognition interfaces(see Figure 1.4). An example of variable spacer length was Jundm2, a member of the basic leucine zipper (bZIP) structural class, which can bind to alternative motifs with a spacer of 1 or 2 nucleotides. Previous reports indicated that different bZIP dimers bind to the CRE (TGACGTCA) and TRE (TGAGTCA) cis-regulatory elements with different affinities (Park et al., 1999). Binding motifs showing position interdependencies made up almost 19% of the TFs. Interestingly, such interdependences were not always in adjacent positions. Multiple effects consisted of a combination of position interdependencies and variable distances separating different parts of the motif. In other cases, TFs recognize their DNA binding sites through multiple, completely different interaction modes (alternate recognition interfaces). They can bind through alternate domains or by switching between alternative structural conformations.

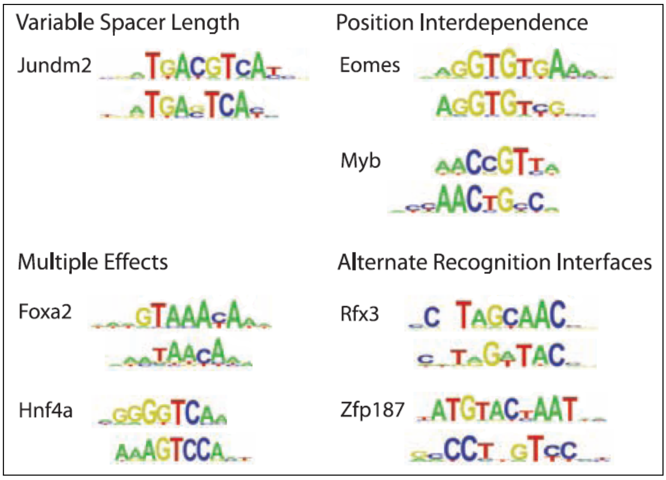


Figure 1.4: Examples of TFs with different class of secondary binding motifs. Adapted from (Badis et al., 2009)

1.4 Regulatory motifs in eukaryotic promoters

Two functional parts are always present in promoters of eukaryotic genes, but they are often difficult to recognize only from the information in the DNA sequence. One part is the basal promoter (or core promoter), where the RNA polymerase complex is recruited, and the other part corresponds to modules that confer specificity to transcription. The composition and organization of these modules and TFBSs vary hugely in different eukaryotic genes (Wray et al., 2003).

In mammals, sequence conservation upstream of the TSS is related with the function of the gene (Lee et al., 2005). Genes involved in complex processes such as development or cellular communication present a more conserved promoter, presumably because they contain more TFBSs. On the other hand, promoters of genes involved in basic processes, like ribosomal metabolism, present a limited conservation which indicates that they are simpler. Many of these genes are housekeeping, that is, expressed in all tissues, so they require less specific regulation (Farré et al., 2007).

Functional assays in cell cultures show that the region between -500 to +50 bp. relative to the TSS is sufficient to induce transcription of most human genes (Trinklein et al., 2003). However, mutations in the regulatory motifs can modify the binding affinity of the factors, and affect gene expression. For instance, one single mutation can lead to the acquisition of a new binding motif. The modular promoter organization implies that the expression in one tissue may evolve independently of the expression in another tissue (Wray et al., 2003).

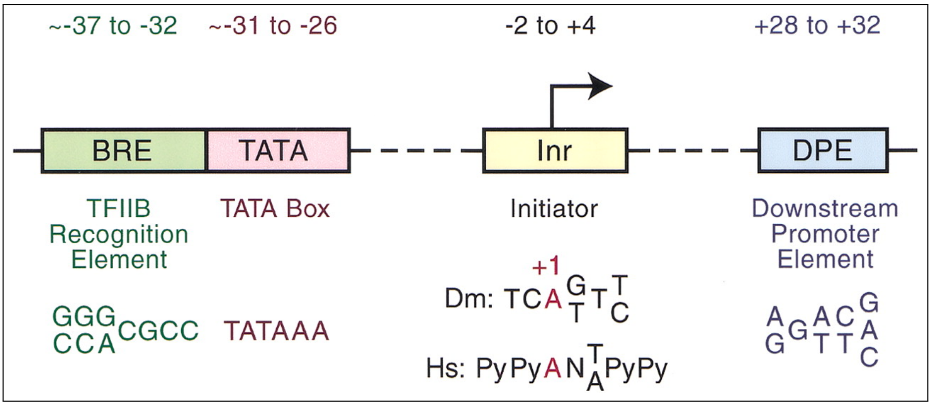


Figure 1.5: Core promoter elements Common cis-regulatory elements in core promoters are usually found at specific distances relative to the TSS. Any specific core promoter may contain some, all, or none of these motifs. The BRE is an upstream extension of a subset of TATA boxes. Different TFs bind upstream adjacent to the TATA-box in other genes. The DPE consensus was identified in *Drosophila* core promoters. The Inr consensus sequence is shown for both Drosophila (Dm) and humans (Hs). Taken from (Butler and Kadonaga, 2002).

There are several cis-regulatory elements which include diverse motifs, such as the TATA box, the initiator (Inr), the TFIIB recognition element (BRE), and the downstream core promoter element (DPE), that are commonly found in eukaryotic core promoters (see Figure 1.5). Each of these core promoter elements is found in some but not all core promoters. It appears that there are no universal core promoter elements (Butler and Kadonaga, 2002). Each of these motifs has specific functions related to the transcription initiation process. Distances between binding motifs result from requirements of proteins that interact with each other to regulate the transcription initiation complex.

The basal transcription factor TFIID is part of Pol II initiation complex. It comprises the TATA-box binding protein (TBP) and different associated factors (TAFs). A trimeric complex of TFIID with TBP, TAF250 and TAF150 bind to sequences that contain the TATA-box and the Inr regulatory elements at a given distance range. The conformation of the protein-DNA complex and binding motifs suggest that both structure and the primary sequence are involved in the sequence recognition (Chalkley and Verrijzer, 1999). Two different sequence-specific DNA-binding factors, TFII-I and YY1, have been found to interact with the Inr (YCANTY). TFII-I is a basic-helix-loop-helix (bHLH) protein that binds to Inr and the E-box (CACGTG) elements and stimulates transcription *in vitro*. In addition, other TFs such as E2F and USF can also stimulate transcription when bound to core promoters elements in specific genes (Smale, 1997).

CpG islands typically lack TATA-box motifs, but contain multiple GC-box motifs, which are bound by Sp1 and other related transcription factors (Butler and Kadonaga, 2002). In addition, transcription from CpG islands initiates from multiple weak start sites that are often distributed over a region of about 100 nt, in sharp contrast to transcription from TATA or DPE-dependent core promoters, which occurs from a single site or localized cluster (of less than 10 nt) of sites.

Regulatory motifs on promoters can be identified by virtue of their preferential location relative to the TSS. Analysis of occurrences of all possible 8-mers in 13,010 human promoter sequences found 9 significant motifs that cluster within 100 bp of the TSS (FitzGerald et al., 2004). Seven of these motifs corresponded to known binding sites for the TFs: Sp1, NF-Y, ETS, CREB, TBP, USF, and NRF-1 (see Figure 1.6).

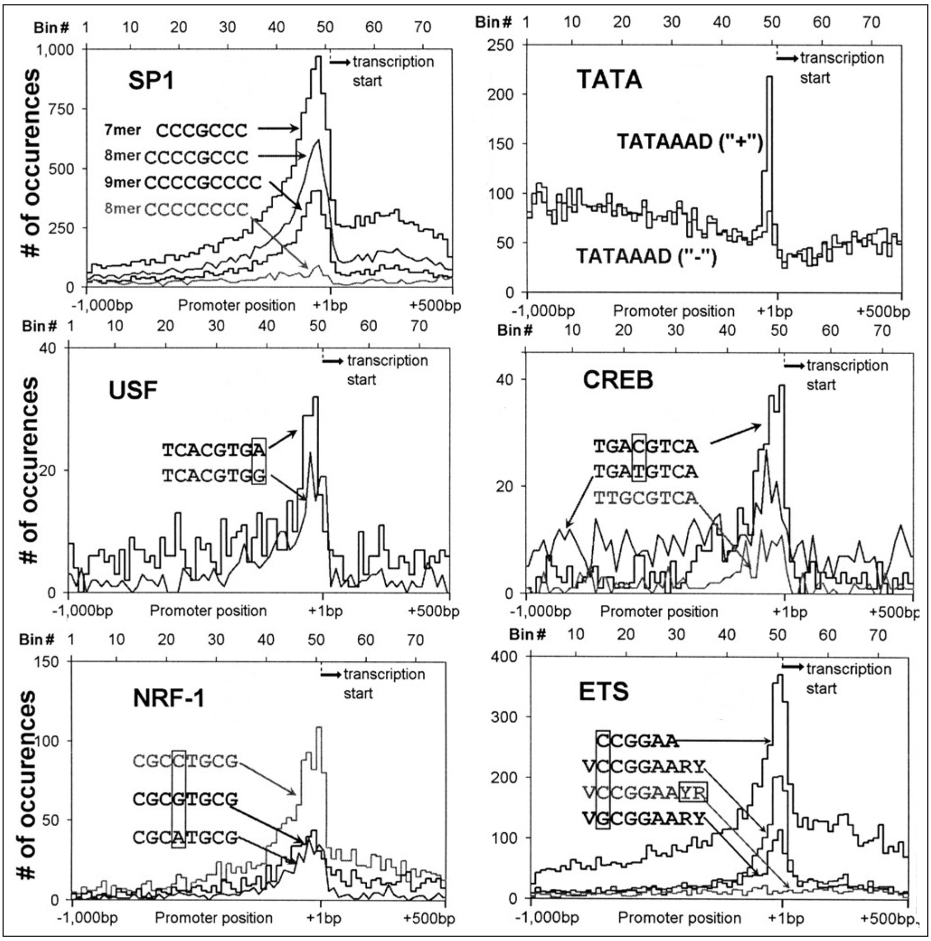


Figure 1.6: **Positional distribution of cis-regulatory motifs** Positional distribution of 8-mers and consensus sequences that correspond to 6 known TFs are show. The y-axis (bins) corresponds to the number of human promoters that contain the corresponding motif in this sequence range relative to the TSS (x-axis). **SP1** Three Sp1 binding sequences and a non peaking single base variation that does not correspond to a Sp1 binding site. **TATA** Strand-specific localization of the TATAAAD sequence. **USF** Two USF (TCACGTGG, TCACGTGA) sequences corresponding to E-boxes (CANNTG). **CREB** Three CRE-box (TGACGTCA) like sequences. **NRF1** Three NRF-1 binding sequences. **ETS** ETS core (CCGGAA), consensus sequence (VCCGGAARY), and a peaking (VGCGGAARY) and non-peaking VCCGGAAYR variant. Adapted from (FitzGerald et al., 2004).

Correlation of regulatory motif and gene expression data from 29 tissues indicate that motifs corresponding to ETS and NRF-1 factors are predominantly found in the promoters of housekeeping genes. On the contrary, TATA-boxes are abundant in the promoters of tissue-specific genes.

Regulatory motifs conservation

TBP is essential for transcription all three eukariotic RNA polymerases (I,II,III). Therefore the TBP and TBP-related factors are quite well conserved among eukariotes. Although bacteria lack TBP, archaea use a protein that is structurally quite similar to the eukaryotic TBP. The TATA-box is a cis-regulatory element found in all eukaryotes, but proportion of TATA promoters are different among them. Estimations using a small set of promoters show that 43% of core promoters in *Drosophila* (Kutach and Kadonaga, 2000) and 32% of human promoters contain a TATA box (Suzuki et al., 2001). However a recent study (Gershenzon and Ioshikhes, 2005) estimated that only around 5% of human promoters contain a TATA-box.

In organisms ranging from plants to mammals, the TATA box is typically located about 25-30 bp. upstream of the TSS. In contrast, in yeast its position is more variable. Although the consensus sequence for the TATA box is TATAAA, it has been observed that a wide range of sequences can function as a TATA box in yeast (Singer et al., 1990). The eukaryotic TATA box closely resembles the prokaryotic 10 bp. upstream Pribnow-box (TATAAT), which is recognized and bound by a subunit of RNA polymerase, but is located further upstream from the start site.

Distribution curves of 6-mers were evaluated to identify positional overrepresentation of DNA motifs with respect to the TSS or the translation start site (ATG). Resulting motifs from promoter datasets of diverse model species were compared. They included *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Saccharomyces cerevisiae*. In addition to the TATA-box, other cis-regulatory motifs were shared between eukaryotic promoters (Berendzen et al., 2006).

Phylogenetic conservation was used to identify putative cis-regulatory motifs, in mammalian genes, 5′ UTR and 3′UTR regions (Xie et al., 2005). Common conserved motifs of 5′ UTR and 3′UTR regions were found among vertebrates. This led to the identification of many known transcription-factor binding motifs, miRNA binding sequences and new putative regulatory motifs with unknown function (Xie et al., 2005). Many of the already known TF binding motifs clustered near the TSS, including the TATA-box, ETS/GABP, NRF-1, E-box (USF), CAAT-box, CRE-box, TRE-box, Sp1 and YY1. From the 3′-UTR analysis they estimated that miRNAs are involved in post-transcriptional regulation of at least 20% of human genes. Taken together these observations point to the existence of a eukaryotic promoter motif architecture that has been conserved throughout evolution.

1.5 In silico prediction of regulatory motifs

The annotation of genomes has been focused on identifying protein coding regions and predicting gene function, often leaving aside the prediction of regulatory elements in non-coding sequences. Regulatory regions play an essential role in gene function, but predicting them continues to be a challenge. Coding sequences have a regular and direct relationship with their immediate phenotype, which is a specific sequence of amino acids. On the contrary, the regulatory sequences have an indirect, not-linear, and context dependent relationship with their immediate phenotype: a particular profile of transcription (Wray et al., 2003).

In general, methods for the prediction of regulatory motifs are based on pattern matching or pattern discovery algorithms.

Pattern matching: A regular pattern is obtained from a group of TFBSs. Experimentally validated sequences are usually longer than the actual binding sequences and the first step is the alignment of the binding sites. A pattern or motif is often represented as a consensus sequence or as a Position Weight Matrix (PWM) also called Position Specific Weight Matrix (PSWM). A letter of the IUPAC code that represent the nucleotide composition of each column of the alignment is assigned for each position of the consensus sequence. This kind of representation loses information on the relative frequencies of nucleotides at each position. The PWM model better reflects the binding preferences at each position using the normalized frequency of the four possible nucleotides. Given a particular DNA sequence a quantitative score can be calculated by summing the values that correspond to the observed nucleotide at each position. For large and representative collections of binding sites, the scores are proportional to binding energies (Stormo, 2000). Libraries of PWMs of specific transcription factors of diverse organisms are available in several databases. JASPAR (Vlieghe et al., 2006) and TRANSFAC (Matys et al., 2006) are the most relevant for vertebrate TFs. Several programs that use PWMs have been developed to predict TFBS, some of which are available on-line and contain their own PWM libraries: Match[1] (Matys et al., 2006), ConSite[2] (Sandelin et al., 2004) or PROMO [3] (Farré et al., 2003). Other representations such as suffix trees have been used to predict TFBSs. Although the accuracy of PWM models has been questioned (Benos et al., 2002a) they are the most popular method for predicting binding sites.

Pattern discovery: Consist of detecting common motifs in a group of unaligned sequences. The objective functions for each method are similar, maximizing likelihoods or likelihood-ratios, but the methods for searching the space of possible alignments are very different. CONSENSUS is a program based on a greedy strategy that progressively adds sub-sequences to a set of alignments where each iteration extends a bounded number of partial alignments (Hertz et al., 1990). MEME is an expectation maximization (EM) method that considers all sites of the training data simultaneously and converges to a local maximum (Bailey and Elkan, 1994). The gibbs sampling algorithm is a stochastic variant of the EM method. Gibbs sampling and expectation maximization algorithms are broadly used for the discovery of regulatory motifs such as TFBS profiles. Most of the programs obtain motifs modeled as PWMs use for subsequent, by pattern matching programs.

A set of transcriptionally co-regulated genes under specific conditions is likely to be regulated by a common set of TFs. Many methods have been developed to predict relevant TFBSs in a set of sequences. They are based on word counting (k-mers), pattern matching (PWM) or pattern discovery algorithms. Most of them use phylogenetic footprinting or over-representation methods to predict common motifs. Although single PWM predictions may

[1] http://www.gene-regulation.com/pub/programs.html

[2] http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite

[3] http://alggen.lsi.upc.es

Table 1.2: In-silico methods to detect common cis-regulatory motifs

Strategy	Base motifs	Organisms	PB	Reference
CRMs	PWM	Hs		Wasserman and Fickett (1998)
global OR	7-mer	14 spp.		Tompa (1999)
neural networks	GS	Sc		Workman and Stormo (2000)
pattern discovery	EM	Dm	+	Ohler et al. (2002)
bayesian networks	GS	Ec		Qin et al. (2003)
global OR	8-mer	Hs	+	Marino-Ramírez et al. (2004)
pos. clustering	8-mer	Hs	+	FitzGerald et al. (2004)
paired motifs	PWM	Ec		Bulyk et al. (2004)
bayesian networks	GS	Sc, Ce		Beer and Tavazoie (2004)
PC	MCS	Hs		Kolbe et al. (2004)
PC	MCS	Hs	+	Xie et al. (2005)
PC/global OR	PWMs	Hs, Mm		Sui et al. (2005)
pos. disequilibrium	6-mer	Sc, Ce, Dm, At	+	Berendzen et al. (2006)
CRMs	PWM	Mm	+	Sharov et al. (2006)
pos. OR	PWM/6-mer	Mm	+	Bellora et al. (2007b)
meta-alignment	PWM	Hs, Mm		Blanco et al. (2007)

PB = positional bias of putative regulatory motifs were evaluated and detected (+). **base motifs** = origin of the evaluated motifs; **k-mers** = all possible k-mers; **PWM** = collections of Position Weight Matrices of known TFs; **GS** = Gibbs Sampling; **EM** = Expectation Maximization (MEME); **MCS** = Motif Conservation Score calculated using ortholog promoter multiple alignments; **PC** = phylogenetic conservation; **pos.** = positional; **OR** = over-representation; **CRMs** = Cis-Regulatory Modules stochastic models; **Organisms**: 14 spp. = 14 prokaryotic species, At = *Arabidopsis thaliana*, Mm = *Mus musculus*, Ce = *Caenorhabditis elegans*, Dm = *Drosophila melanogaster*, Ec = *Escherichia coli*, Hs = *Homo sapiens*, Mm = *Mus musculus*.

contain a high rate of false positives, over-represented predicted TFBSs are likely to be functionally relevant as the background noise is accounted for. However, one has to bear in mind that motif discovery algorithms and over-representation methods are highly dependent on the background model. Other methods that analyze the positional distribution of motifs to predict regulatory motifs, have been used on large promoter datasets. Sophisticated methods include those based on stochastic CRM definitions, bayesian networks, or meta-alignment of predicted TFBSs. Examples of 16 different methods applied to diverse sets of promoters are summarized in Table 1.2. In general, the specificity and sensitivity of such methods has not actually been estimated, implicitly because the partial coverage of the experimental evidence.

Although many regulatory motifs can be identified by virtue of their over representation or their conservation, others motifs might be difficult to identify. Indeed, low affinity binding sites, alternative recognition motifs and non-conserved functional binding sites (Odom et al., 2007) are common trends of TFBS in mammalian promoters.

2

Objectives

The work developed during the PhD was focused on improving our understanding of the organization of the regulatory signals present in the proximal promoter region relative to gene expression. The objectives of this thesis can be summarized as follows:

1. Development a novel computational method for the determination of positional dependent motifs on a set of DNA or RNA sequences.
2. Analyse relative location and spacing of cis-regulatory motifs in promoter of housekeeping and tissue-specific genes.
3. Identification of specific transcription factors involved in the co-regulation of a set of genes.

3

Results

Summary

In this chapter, I include 3 articles that are directly relevant to my thesis. Two of them are published. The first article is an application note describing a novel method and a web application to determine motifs with positional bias. Further research on mouse promoters using the method correspond to the second article. The third one describes a procedure to identify transcription factors involved in the regulation of a set of co-expressed genes.

3.1 PEAKS: identification of regulatory motifs by their position

Abstract

Many DNA functional motifs tend to accumulate or cluster at specific gene locations. These locations can be detected, in a group of gene sequences, as high frequency peaks with respect to a reference position, such as the transcription start site (TSS). We have developed a web tool for the identification of regions containing significant motif peaks. We show, by using different yeast gene datasets, that peak regions are strongly enriched in experimentally-validated motifs and contain potentially important novel motifs.

Availability: http://genomics.imim.es/peaks

Bellora N, Farré D, Mar Albà M.
PEAKS: identification of regulatory motifs by their position in DNA sequences.
Bioinformatics. 2007 Jan 15;23(2):243-4.

3.2 Positional bias of general and tissue-specific regulatory motifs

Abstract

Background: The arrangement of regulatory motifs in gene promoters, or promoter architecture, is the result of mutation and selection processes that have operated over many millions of years. In mammals, tissue-specific transcriptional regulation is related to the presence of specific protein-interacting DNA motifs in gene promoters. However, little is known about the relative location and spacing of these motifs. To fill this gap, we have performed a systematic search for motifs that show significant bias at specific promoter locations in a large collection of housekeeping and tissue-specific genes.

Results: We observe that promoters driving housekeeping gene expression are enriched in particular motifs with strong positional bias, such as YY1, which are of little relevance in promoters driving tissue-specific expression. We also identify a large number of motifs that show positional bias in genes expressed in a highly tissue-specific manner. They include well-known tissue-specific motifs, such as HNF1 and HNF4 motifs in liver, kidney and small intestine, or RFX motifs in testis, as well as many potentially novel regulatory motifs. Based on this analysis, we provide predictions for 559 tissue-specific motifs in mouse gene promoters.

Conclusion: The study shows that motif positional bias is an important feature of mammalian proximal promoters and that it affects both general and tissue-specific motifs. Motif positional constraints define very distinct promoter architectures depending on breadth of expression and type of tissue.

Bellora N, Farré D, Albà MM.
Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters.
BMC Genomics. 2007 Dec 13;8:459.

3.3 Identifying specific regulatory motifs in co-regulated genes

Abstract

Large scale transcriptomic experiments such as microarray-based gene expression profiling provide lists of genes that are co-regulated under particular conditions. In many cases such gene co-regulation results from transcriptional activation or repression by common transcription factors that interact with cis-regulatory motifs present in their promoters. Sequence specific transcription factors binding sites (TFBSs) can be predicted using position weighted matrices (PWMs). However, individual predictions have a high rate of false positives due to the promiscuity of binding sites and the model itself. A useful strategy to obtain a confident set of regulatory motifs is measuring their signal to background ratio and selecting those significantly overrepresented. We have set up one such method and we have used it to identify highly specific motifs in promoters of co-expressed genes in a set of genes of interest. In addition, the motifs can be selected in virtue of their level of conservation in vertebrate syntenic regions. The method can also be applied to detect other unknown cis-regulatory motifs by measuring the overrepresentation and conservation of DNA words or k-mers.

Identifying specific regulatory motifs in co-regulated genes

Method overview

We have developed a method, called MOR, to identify regulatory motifs that are highly specific of a group of genes under study. The statistical significance of the motifs is tested using as a background model the rest of promoter sequences, which filters out those motifs that are common in promoters in general, and leaves only motifs which are dataset-specific. Affymetrix probe IDs or ENSEMBL gene IDs are mapped to their corresponding refSeq IDs using Ensembl version 55 downloaded from BioMart (Smedley et al., 2009). Complete genome sequences and annotations of human, mouse, fly and yeast were downloaded from UCSC (Karolchik et al., 2008). We used refSeq experimental mRNA annotations to determine transcription start site (TSS) positions and extract their corresponding promoter sequences (see Table 1). In yeast we have used a genome wide full-length cDNA analysis (Miura et al., 2006) to determine the TSS position.

Promoter regions were extracted using two diverse length intervals, covering 700 bp. or 2000 bp.: from -1000 bp. to +1000 bp., and -600 bp.to +100 bp., relative to the TSS. The length can be chosen at the beginning of the analysis. Regions of alternative or bi-directional promoters often overlap. Such overlap interferes on the statistical analysis of sequences because occurrences of predicted TFBSs are counted several times. We select all annotated promoters for the input dataset, the

Table 1: **Available promoters per organism**

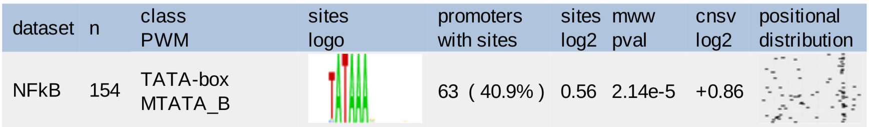
organism	genome version	promoters
Homo sapiens	hg18 (NCBI Build 36.1)	26280
Mus musculus	mm9 (NCBI Build 37)	21585
Drosophila melanogaster	dm3 (BDGP Release 5)	21158
Saccharomyces cerevisiae	sacCer1 (SGD)	3431

remaining promoters are kept as background dataset. Then overlapping regions are eliminated in both datasets, keeping only independent ones. The method can use available collections of PWMs, such as TRANSFAC (Matys et al., 2006) or JASPAR (Vlieghe et al., 2006), consensus motifs (CNS) or DNA words / short oligomers (k-mers).

In order to identify motifs that occur more frequently than expected and the proportion of genes with motifs, we compare the distributions of the input and background datasets. We use the non-parametric MannWhitney-Wilcoxon test (mww) (R Development Core Team, 2008) for assessing whether two independent samples of observations come from the same distribution. Significant motifs are selected based on mww p-value and the ratio between the mean number of genes with sites in the input and background datasets. A very large number of predictions for a given PWM often indicates poor matrix quality. With this test we also implicitely eliminate those low quality matrices.

Figure 1 shows the result output of an over-represented motif, the TATA-box, in the analysis with TRANSFAC vertebrate PWMs of a dataset of 154 human genes regulated by NF-kappaB on 700bp promoter regions. Columns correspond to dataset name (dataset), number of promoters (n), TF or motif class and representative

PWM (class/PWM), the sequence logo of PWM matches (sites logo), number and percentage of promoters with at least a site (promoter with sites), ratio between mean sites (sites log2), MannWhitney-Wilcoxon test p-value (mww pval), sites conservation (cnsv log2) (see next section) and sites positional distribution along promoters (positional distribution). Sequence logo represents the positional frequency of sites predicted in a given dataset.



dataset	n	class PWM	sites logo	promoters with sites	sites log2	mww pval	cnsv log2	positional distribution
NFkB	154	TATA-box MTATA_B		63 (40.9%)	0.56	2.14e-5	+0.86	

Figure 1: **TATA-box over-represented in genes regulated by NF-kappaB.**

There is redundancy between different matrices that correspond to the same TFs. Moreover, TFs with the same DNA binding domain class have similar binding profiles. In order to show non-redundant results we cluster those matrices with a similar binding profile, Most of them cluster TFs that belong to the same structural family. Other families with similar PWMs such as ATF, CRE and bZIP cluster together. We refer to these clusters as non-redundant motifs.

Conservation of coding regions and promoters

In order to provide more confident predicted binding sites for the set of candidate TFs we measured the sequence conservation in predicted sites and promoter regions. We used the conservation score between vertebrates genomic sequences called phastCons (Siepel et al., 2005), downloaded from the UCSC ftp site (Karolchik et al., 2008). We calculated the ratio between the observed mean conservation score in TFBS predictions and the expected mean conservation score in promoter sequences from the dataset. Distributions of conservation scores differ between gene regions. We considered four types of gene regions: coding exons, UTR exons, introns, and intergenic (non-available annotations, NA). We used refSeq annotations in the human and mouse genomes. Coding exons and UTR exons have higher scores than introns or non-coding regions. Conservation scores of experimental binding sites mapped to promoters regions are higher than UTR exons but lower than CDS exons. Promoter regions in our analysis overlap with diverse genomic elements, so we also calculate their specific score distribution. Transcription factor binding sites on promoters are more conserved than promoter average. We can expect more conservation for highly specific predictions when compared to their respective complete promoter sequences. For this reason, the conservation score ratio between predictions and promoters conservation (observed/expected) is given for each dataset and candidate TFs. In addition, we provide the conservation score of each single prediction in each gene promoter.

Evaluation of MOR in NF-kappaB regulated genes

Rel or NF-kappaB (NFkB) proteins comprise a family of eukaryotic transcription factors that are involved in the control of a large number of normal cellular and organismal processes, such as immune and inflammatory responses, developmental processes, cellular growth, and apoptosis. Rel/NF-kB transcription factors bind to 9-10 base pair DNA sites (called kB sites) as dimers. All vertebrate Rel proteins can form homodimers or heterodimers, except for RelB, which can only form heterodimers. This combinatorial diversity contributes to the regulation of distinct, but overlapping, sets of genes. In order to evaluate the method we retrieved a list of 167 genes regulated by NFkB (Pahl, 1999). Among those genes, 104 are known

NFkB targets in human, 25 are known targets in rodents and 38 correspond to putative targets. We extracted 154 non-overlapping human promoters, from -600 bp. to +100 bp. and -1000 bp. to +1000 bp. relative to the TSS.

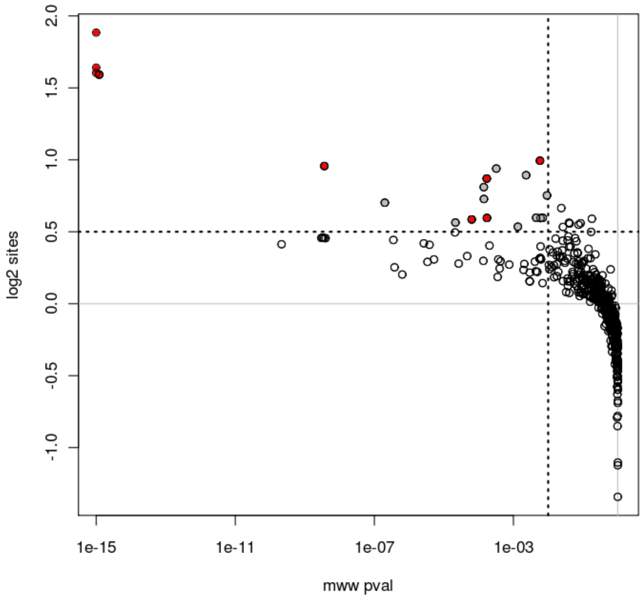


Figure 2: **Matrices selection in NF-kappaB regulated genes** Out of 638 vertebrate PWMs, 20 were overrepresented (sites log2 ratio ≥ 0.5, mww p-value ≤ 0.01, 10% of genes with sites) in 154 promoters of genes regulated by NF-kappaB. The promoter regions analized spanned from 600 bp. to +100 bp. relative to the TSS. In red 9 PWMs that correspond to NFkB and REL motif in the TRANSFAC vertebrate collection. In grey, other matrices that were also over-represented. sites log2 = ratio between mean sites, mww pval = MannWhitney-Wilcoxon test p-value.

We analyzed both datasets with a collection of 638 vertebrate PWMs from TRANSFAC 2009 (Matys et al., 2006). Sequence hits to a matrix were defined as those that showed an overall matrix relative similarity score ≥ 0.85 and core similarity score ≥ 0.99. This collection contains nine similar PWMs that correspond to NFkB and

REL binding motifs. To retrieve only over-represented TF candidates we selected PWMs using three criteria: sites distribution (mww p-value), mean number of sites per gene (sites log2) and a minimum of 10% of genes with predictions. Based on the distribution of scores and specific scores of 9 NFkB/Rel binding motifs, a sites log2 cutoff of 0.5 and a mww p-value cutoff of 0.01 were empirically selected (see Figure 2).

There were differences between the results obtained using different promoter sizes. In the -100 to +600 dataset we identified 20 significant PWMs. They correspond to 9 non-redundant binding motifs: NFkB/Rel, STAT, BACH, ZINC-FINGER, CEBP, TATA-box, HNF4, IRF and E-box (see Figure 4 A). On the other hand in the -1000 to +1000 dataset we identified 13 significant PWMs that correspond to 6 non-redundant binding motifs: NFkB/Rel, STAT, FORKHEAD, POU3F2, BACH and PAX/HNF3 (see Figure 4 B). Two binding motifs, NFkB and Stat, cluster most of the signifcant PWMs in both datasets. Six out of 9 NFkB/Rel PWMs on TRANSFAC, were significant in the longer dataset and STAT motif clusters 3 matrices in both datasets.

Applying restrictive cutoffs, sites log2 ratio ≥ 1, mww p-value $\leq$ 1e-3 and 25% of genes with sites, we identify only NFkB/Rel motifs in both datasets. Although we could loose important regulatory candidates the main expected regulator is found and results are consistent among datasets.

Comparison with alternative methods

We compared the results of MOR with those obtained by three methods available via web also based on PWM over-representation : (Bellora et al., 2007a) oPOSSUM (Sui et al., 2007), TFM-Explorer (Defrance and Touzet, 2006) and PEAKS (Bellora et al., 2007a). **oPOSSUM** searches over-represented JASPAR (Vlieghe et al., 2006) PWM predictions conserved between human and mouse promoters. It uses two different statistics to assess the overrepresentation of the number of sites and the proportion of genes with sites (Z-score and Fischer exact test respectively). Analyzed regions by default span from -5000 bp. to +5000 bp. relative to the TSS. **TFM-Explorer** initially identifies all potential TFBS using JASPAR or TRANSFAC PWMs in human and available mouse orthologous promoters. In the second step it uses an algorithm based on positional cumulative scoring scheme that is windows independent. High-scoring regions define the window size that is retained as candidate. Finally selected candidates are statistically evaluated using precomputed background models. **PEAKS** searches for overrepresented motifs with local over-representation on DNA sequences. Hits to diverse PWM libraries, including JASPAR and TRANSFAC, or k-mers of a given length are scored using a user defined sliding window. To determine the significance it uses a background model derived from the sequence dataset.

There are differences between methods: oPOSSUM and TF-Explorer use available orthologous promoters to calculate motif over-representation. TF-Explorer and PEAKS are based on local over-representation, that is, on identifying regions with a high-frecuency of predictions, while MOR or oPOSSUM obtain candidates using global-overrepresentation. Moreover, different PWM libraries are used by each method. We apply each method to the NFkappa-B regulated dataset of 154 genes, described above, that is enriched in experimental TFBS evidence. Because every method uses a different statistic we selected the default, non-restrictive, cut-offs defined by their authors (see Table 2).

Table 2: **Differences between methods**

method	PF	OR	library	PWMs	cut-offs	n
MOR	NO	global	TF-2009	638	Log2s ≥ 0.5 ; mww ≤ 0.01	20
oPOSSUM	YES	global	JASPAR	123	Z-scr ≥ 5; fisher ≤ 0.05	8
TFM-Explorer	YES	local	TF-7	508	p-value, top 20	20
PEAKS	NO	local	TF-7	508	p-value $\leq$ 1e-5	16

PF = phylogenetic footprinting (use of orthologous sequences), OR = over-representation class, library = vertebrate PWMs library (TF = TRANSFAC), PWMs = total number of matrices in the library, cut-offs = non-restrictive selection criteria, cut-offsn = number of significant matrices

In order to compare the TFBSs predicted by different methods we mapped diverse experimental binding sites of the TRANSFAC database (Matys et al., 2006) on 41 (27%) promoters of the dataset. The experimental binding sites comprised 117 different TFs or variants and 165 binding sites, as in some cases diverse TFs bind to the same site. We clustered similar transcription factors, belonging to the same family and having a similar binding profile, ie: NF-kappaB clusters RelA-p65, NF-kappaB, p50, c-Rel, p100, NF-kappaB2-p52, RelB, NF-kappaB(-like) and RBP-Jkappa(p50). The TFs corresponding to most of the significant motifs obtained by the four methods have experimental evidence (see Table3). Only 3 PWMs over-represented on TF-Exp analysis did not have any experimental evidence, they correspond to Pbx-1, Brn-2 and XFD-3 TFs. There are 14 TF groups that regulate more than one promoter, 13 can be successfully mapped to over-represented motifs. They are all detected by one or more methods, with the exception of the EGR binding motif. The results strongly indicate that those TFs, although only with experimental evidence in few promoters, are likely to regulate more genes

Table 3: **Identification of transcription factors with experimental evidence**

TF	DBD	genes	MOR	oPOSSUM	TFExp	PEAKS
NF-kappaB	REL	14	+	+	+	+
AP1/CRE	bZIP	14	+	+	+	+
C/EBP	bZIP	8	+		+	+
Sp1	ZF	8			+	+
NFAT	REL	4	+		+	
TBP	TBP	3	+		+	+
STAT	STAT	3	+			
HEN-1	bHLH	3	+			
Oct1/2	homeo	3			+	
EGR	ZF	3				
IRF	IRF	2	+			
AP2	bHSH	2				+
ELF	ETS	2		+		
PU.1	ETS	2		+		

Experimental TFBS that correspond to 14 transcription factors with experimental evidence in at least 2 promoters are listed. 38 transcription factors with sites in only one promoter were excluded, 3 of them mapped to over-represented PWMs. [+] The TF corresponds to one or more over-represented PWMs using a given method. TF = transcription factor, DBD = DNA binding domain family, genes = number of gene promoters with experimental TFBS.

Different PWMs of NF-kappaB and AP1/CRE of were over-represented with all methods and corresponds to the most abundant experimental TFBSs. AP1 (Fos/Jun heterodimer) is a basic leucine zipper (bZIP) domain TF that binds to TRE-box (TGASTCA) or a CRE-box (TGACGTCA) with different affinities (Hai and Curran, 1991). AP1 regulates gene expression in response to a variety of stimuli, including cytokines, growth factors, stress, and bacterial and viral infections. MOR detects BACH which is a similar bZIP with the same binding profile. Recently it has been characterized as a specific transcriptional repressor of the enzyme heme oxygenase-1 (HMOX1) (Reichard et al., 2008), one of the genes present on the dataset . Both transcription factors, NF-kappaB and AP1, have a well defined motif and a large number of PWM that model their binding sites.

Sp1 is a ubiquitous activator of numerous genes in the human genome and interacts with NF-kappaB (Parker et al., 1996). In virtue to its spatial distribution of their binding sites, this factor was found by the two methods that use local over-representation, but MOR or oPOSSUM did not predict Sp1 because it is not specific of this dataset. Another transcription factor found by 3 of the 4 methods was C/EBP, also called NF-IL6. This is a CCAAT enhancer binding protein that binds to a non-canonical CCAAT box and cooperate with NF-kappa in the regulation of many genes (Stein et al., 1993).

Other over-represented motifs only found by MOR, IRF and STAT, play a well known role in proximal enhansosomes in diverse promoters of NF-kappaB regulated genes interacting with AP1, C/EBP and Sp1 (Richmond, 2002). Those enhansosomes are present in promoters that contain a TATA-box which is consistent with results of MOR, TF-Explorer and PEAKS. Predictions of NFAT, NF-kappaB, STAT and IRF partially overlap, and all of them are involved in immune response in mammals. Ikaros-3 is a C_2H_2 zinc-finger (ZF) transcription factor over-represented in MOR analysis but their binding motif is similar to NFAT which is a REL factor. Although there is no experimental evidence in this dataset, the Ikaros family are important for the development of the immune system (John et al., 2009). In the oPOSSUM analysis PU.1 was over-represented, this is an ETS-domain transcription factor that activates gene expression during myeloid and B-lymphoid cell development. Other over-represented motifs that map to TFs with experimental evidence in only one promoter and found by different methods were: YY-1 (TF-Exp), HNF-4 (MOR) and MZF-1 (oPOSSUM).

Identifying specific TFBS in co-expressed gene clusters

Human genome-wide expression data were used to build a confident human gene co-expression network (Prieto et al., 2008). The network reveals a map of co-expression clusters organized in well defined functional sub-networks. The map shows that the larger clusters correspond to genes involved in mitochondrial (Mitochondrial-A and Mitochondrial-B), nuclear (Nuclear) and ribosomal related metabolism (Ribosomal). These clusters are enriched in housekeeping genes. Smaller clusters include: genes of the major histocompatibility complex (MHC) (Histocomp), genes that produce the cell surface CD antigens (CDAntigens), genes involved in metal ion homeostasis (Metalion), genes related to the extracellular matrix and cell adhesion (Adhesion) and genes related to the cytoskeleton (Cytoskeleton). We investigate their correlation with specific transcription factors. Clusters were analyzed using TRANSFAC PWMs and significant ones were selected (sites log2 ratio ≥ 1, mww p-value $\leq$ 1e-3, 25% of genes with sites). Considering all datasets, 60 PWMs were overrepresented. They correspond to 49 unique matrices because some of them were significant in more than one dataset, a total of 20 non-redundant motifs in 9 datasets (see Table 4 and Figures 6 and 5).

Table 4: **Significant motifs found in co-expressed gene clusters**

	dataset	seq	PWM	nrm	
1	Adhesion	14	5	1	ZINC FINGER
2	CDAntigens	35	7	4	GATA, ETS, IRF, ESTEROID RECEPTOR
3	Cytoskeleton	19	4	1	SRF
4	Histocomp	19	6	4	CAAT-box, IRF, SMAD3, P53
5	Metalion	7	4	4	ZINC FINGER, MTF1, X-box, CBF
6	Mitochondrial-A	57	7	6	E2F, CRE/ATF, YY, NRF1, ETS, MIZF
7	Mitochondrial-B	38	8	6	YY, E-box, E2F, NRF1, ETS, MIZF
8	Nuclear	130	8	3	CRE/ATF, STAF, E2F
9	Ribosome	40	11	5	ETS, YY, CRE/ATF, STAF, MAF

Some matrices were significant in multiple datasets. 49 unique PWMs that correspond to 20 non-redundant motifs (nrm) were over-represented in 9 co-expression clusters (dataset). seq=number of promoter sequences.

Housekeeping related transcription factors, such as YY, CREB/ATF and E2F (Bellora et al., 2007b), were the most over-represented in the largest clusters and absent in smallest ones. E2F, a TF related to cell cycle, was not significant in Ribosomal. NRF1, nuclear respiratory factor 1, was only significant in both Mitochondrial datasets. On the other hand STAF, activator of pol II and pol III promoters of several small RNA genes (Schaub et al., 1997), was overrepresented only in Nuclear and Ribosomal. Specific ETS matrices were significant in housekeeping related clusters such as SAP1A_01 (SAP-1a or Elk-4 TFs) or CETS1P54_03 (Ets-1 TF).

In contrast, another ETS factor, PU1_Q4 (SPI-1, a TF that activates gene expression during myeloid and B-lymphoid cell development) was only overrepresented in CDAntigens. In addition to ETS/PU1 other motifs related to immune response such as GATA, IRF (interferon response factor) and ESTEROID RECEPTOR (Tait et al., 2008) were overrepresented in CDAntigens.

In Histocompatibility, IRF, P53 (p53 TFs), SMAD3 (Smad3 TFs) and CCAAT (NF-Y) were overrepresented. The tumor suppressor p53 is an essential partner of Smads (Atfi and Baron, 2008) and mutations of Smad-3 were associated with cancer (Zhu et al., 1998). Although CCAAT is a common motif of vertebrate promoters, it was more abundant in this dataset. In Metalion, 4 PWMs were over-represented: MTF1, ZINC FINGER, X-BOX and CBF. MTF-1 is a specific TF that binds to metal responsive elements (Koizumi et al., 1999). Binding sites of specific ZINC FINGERS TFs were over-represented in Adhesion. All PWMs matched to the core sequence GAGGG. In Cytoskeleton, SRF motif was significant. Serum response factor (SRF) is an absolutely essential orchestrator of actin cytoskeleton and contractile homeostasis (Miano et al., 2007).

Discussion

MOR is based on the idea that co-expressed genes may share common cis-regulatory elements that correspond to TFBS of specific TFs. This method evaluate the sequence motifs enrichment using a simple statistitc test in a selection of non-overlaping promoters. In addition, our approach integrates information of experimental binding sites and phylogenetic conservation to identify highly reliable TFBS. Similar methods to predict common TFBSs that are based on motif over-representation or phylogenetic conservation were compared and MOR was used to identify regulatory motifs in co-expressed genes.

In order to evaluate the efficiency and the biological relevance we compare our results and the results of three alternative methods with the available experimental evidence. All TFs that correspond to motifs over-represented in MOR (n=20), oPOSSUM (n=8) and PEAKS (n=16) have experimental evidence, except 3 out of 20 motifs predicted by TF-Explorer. Same of them can be mapped to particular TFs, however many results are redundant because of the inherent redundancy of the PWM libraries. There are 14 TFs that regulate more than one promoter, only the motif of EGR was not detected by any method. NF-kappaB and AP1 were over-represented in all methods, other factors that interact with NF-kappaB, such as C/EBP or Sp1 were significant using different methods. TATA, AP1/CRE and Sp1, that correspond to motifs with positional bias (Bellora et al., 2007b), were detected by the 2 methods based on local over-representation. Those methods found motifs that are not specific of this dataset but with validated binding sites, such as Sp1, Oct1/2 or AP2. AP1, C/EBP, NFAT, STAT, IRF and PU.1 are well characterized transcription factors that interacts with NFkB in immune response processes. Out of these 7 important factors MOR found 6, oPOSSUM 3, TF-Explorer 4 and PEAKS 3. In addition of the particular algorithm used by each method, differences in the results, are related to the number of available matrices in the library, the number and the length of promoter regions analyzed or the background model.

Phylogenetic footprinting methods, such as oPOSSUM or TFM-Explorer are based on the assumption that functional TFBS are located on conserved regions. Those methods give similar results than MOR or PEAKS, that do not use ortholog promoters. There are some problems using phylogenetic footprinting on promoter regions: availability of ortholog promoters, sensibility of pairwise alignments in non-coding sequences and the assumption that all conserved sites are functional. We provide the conservation score of individual predictions together with the observed/expected ratio of conservation in promoters for a posteriori evaluation.

Using MOR we identify highly specific motifs in co-expressed datasets analyzed and for many of them there is experimental evidence of their implication in the regulation of such clusters. Although most transcription factor classes have a well defined binding motif and are found differentially on particular datasets, zinc-finger and ETS transcription factors present more diverse binding profiles depending on the factor. Five diverse zinc-finger binding motifs were found on 7 of 9 datasets: GGAG (several PWMs, Adhesion), GATC (GATA, CDAntigens), GNNTTCC (IK3, Metalion), GTCCG (MIZF, Mitichondrial A/B) and ACTTCC (STAF, Nuclear/Ribosomal). This is not surprising given that zinc-finger TFs (C_2H_2ZF) is the most extended family of human transcription factors covering around 53% of the repertory (Vaquerizas et al., 2009). Homeodomain TF family is the second most abundant in humans and accounts for over 20% of the repertory. We did not found homeodomain TFs because they are related to development processes that are not represented on the co-expression clusters. In a reference dataset of liver specific genes (Sui et al., 2007) MOR found over-represented specific hepatocite factors (HNF4 and HNF1) in addition to PWMs that correspond to homeodomain factors (data not shown). Other ubiquitous transcription factors such as Sp1, a GC-rich binding zinc-finger, were not over-represented in the datasets. This shows that the method can succesfully discriminate between specific and common transcriptional regulators.

In general, vertebrate conservation scores of predictions relative to mean conservation along individual promoters were higher than expected (see Figure 3). This fact indicate that those predictions greatly depart from noise and are likely to be regulatory regions.

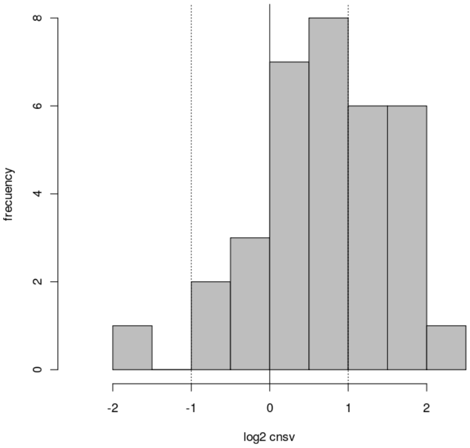
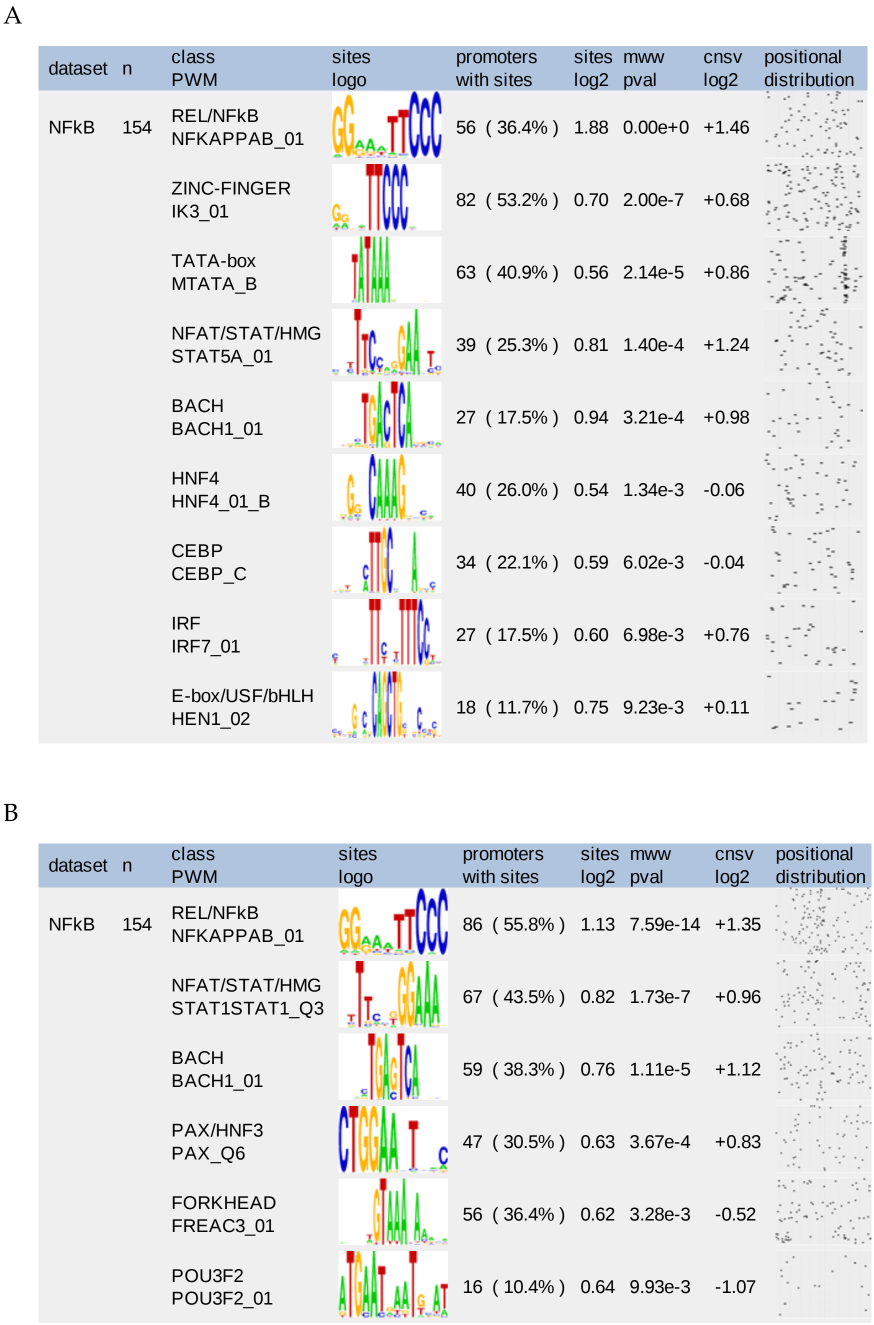


Figure 3: **Observed vs expected sequence conservation along significant predictions.** Distribution of Log2 ratios between observed and expected conservation scores (log2 cnsv) along predictions of non-redundant motifs in 9 dataset of co-expressed gene clusters. For most of those motifs (28 of 34) it was higher than expected.

Finally, we use MOR to identify TFs candidates in a particular microarray experiment. Nerve injures often leads to neuropathic pain syndromes. Experiments in microglia cells revaeled that transcriptional changes induced by interferon-γ (IFN-γ) are modulated by CB(2) receptor signaling. In this analysis, we evaluated 72 promoter regions of the genes differentially regulated after nerve injury in bot h wild type and knockout genotypes. Binding sites of IRF (interferon response element) factors were the only over-represented regulatory motifs (Racz et al., 2008). This result was consistent with experimental observations.

We have used a restrictive similarity cut-off on PWM matching, this is illustrated by the fact that in the NFkB dataset we don't detect all real binding sites. However, predictions are less noisy allowing to detect significant differences between a given dataset and background occurrences. Additionally, we may miss some important cis-regulatory motifs not represented by the those matrices. Unknown regulatory motifs can be obtained measuring k-mers instead of PWMs, but their interpretation is difficult.



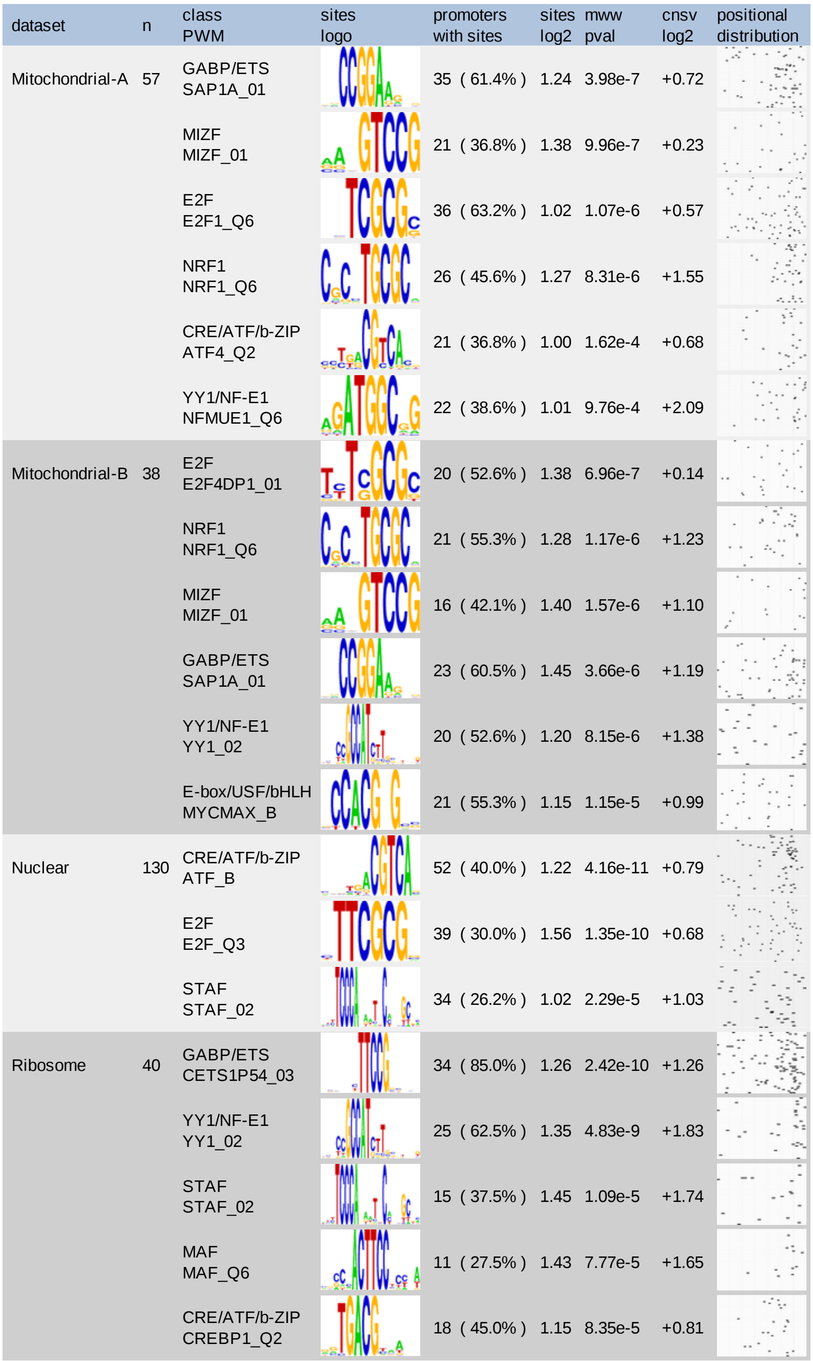
A

dataset	n	class PWM	sites logo	promoters with sites	sites log2	mww pval	cnsv log2	positional distribution
NFkB	154	REL/NFkB NFKAPPAB_01		56 (36.4%)	1.88	0.00e+0	+1.46	
		ZINC-FINGER IK3_01		82 (53.2%)	0.70	2.00e-7	+0.68	
		TATA-box MTATA_B		63 (40.9%)	0.56	2.14e-5	+0.86	
		NFAT/STAT/HMG STAT5A_01		39 (25.3%)	0.81	1.40e-4	+1.24	
		BACH BACH1_01		27 (17.5%)	0.94	3.21e-4	+0.98	
		HNF4 HNF4_01_B		40 (26.0%)	0.54	1.34e-3	-0.06	
		CEBP CEBP_C		34 (22.1%)	0.59	6.02e-3	-0.04	
		IRF IRF7_01		27 (17.5%)	0.60	6.98e-3	+0.76	
		E-box/USF/bHLH HEN1_02		18 (11.7%)	0.75	9.23e-3	+0.11	

B

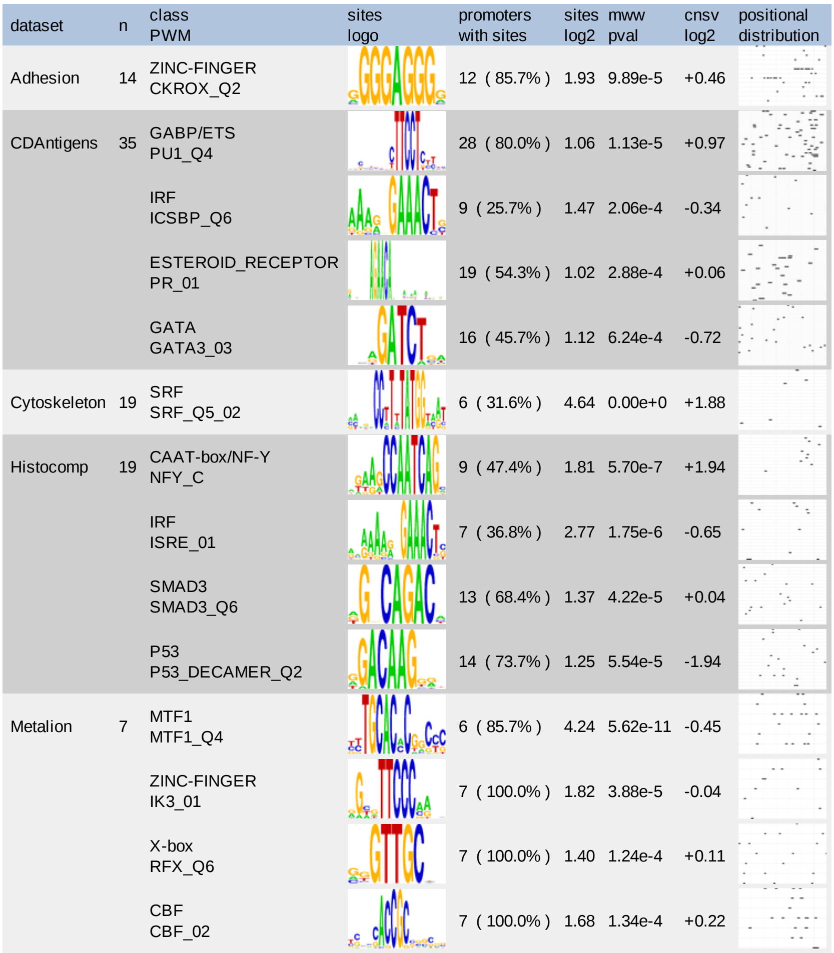
dataset	n	class PWM	sites logo	promoters with sites	sites log2	mww pval	cnsv log2	positional distribution
NFkB	154	REL/NFkB NFKAPPAB_01		86 (55.8%)	1.13	7.59e-14	+1.35	
		NFAT/STAT/HMG STAT1STAT1_Q3		67 (43.5%)	0.82	1.73e-7	+0.96	
		BACH BACH1_01		59 (38.3%)	0.76	1.11e-5	+1.12	
		PAX/HNF3 PAX_Q6		47 (30.5%)	0.63	3.67e-4	+0.83	
		FORKHEAD FREAC3_01		56 (36.4%)	0.62	3.28e-3	-0.52	
		POU3F2 POU3F2_01		16 (10.4%)	0.64	9.93e-3	-1.07	

Figure 4: **Significant regulatory motifs found in NFkB regulated genes promoters. (A) Analized from -1 Kbp. to +1 Kbp. relative to TSS. (B) Analized from -600**



dataset	n	class PWM	sites logo	promoters with sites	sites log2	mww pval	cnsv log2	positional distribution
Mitochondrial-A	57	GABP/ETS SAP1A_01		35 (61.4%)	1.24	3.98e-7	+0.72	
		MIZF MIZF_01		21 (36.8%)	1.38	9.96e-7	+0.23	
		E2F E2F1_Q6		36 (63.2%)	1.02	1.07e-6	+0.57	
		NRF1 NRF1_Q6		26 (45.6%)	1.27	8.31e-6	+1.55	
		CRE/ATF/b-ZIP ATF4_Q2		21 (36.8%)	1.00	1.62e-4	+0.68	
		YY1/NF-E1 NFMUE1_Q6		22 (38.6%)	1.01	9.76e-4	+2.09	
Mitochondrial-B	38	E2F E2F4DP1_01		20 (52.6%)	1.38	6.96e-7	+0.14	
		NRF1 NRF1_Q6		21 (55.3%)	1.28	1.17e-6	+1.23	
		MIZF MIZF_01		16 (42.1%)	1.40	1.57e-6	+1.10	
		GABP/ETS SAP1A_01		23 (60.5%)	1.45	3.66e-6	+1.19	
		YY1/NF-E1 YY1_02		20 (52.6%)	1.20	8.15e-6	+1.38	
		E-box/USF/bHLH MYCMAX_B		21 (55.3%)	1.15	1.15e-5	+0.99	
Nuclear	130	CRE/ATF/b-ZIP ATF_B		52 (40.0%)	1.22	4.16e-11	+0.79	
		E2F E2F_Q3		39 (30.0%)	1.56	1.35e-10	+0.68	
		STAF STAF_02		34 (26.2%)	1.02	2.29e-5	+1.03	
Ribosome	40	GABP/ETS CETS1P54_03		34 (85.0%)	1.26	2.42e-10	+1.26	
		YY1/NF-E1 YY1_02		25 (62.5%)	1.35	4.83e-9	+1.83	
		STAF STAF_02		15 (37.5%)	1.45	1.09e-5	+1.74	
		MAF MAF_Q6		11 (27.5%)	1.43	7.77e-5	+1.65	
		CRE/ATF/b-ZIP CREBP1_Q2		18 (45.0%)	1.15	8.35e-5	+0.81	

Figure 5: **Non-redundant motifs in Mitchondrial, Nuclear and Ribosome co-expressed gene clusters.**



dataset	n	class PWM	sites logo	promoters with sites	sites log2	mww pval	cnsv log2	positional distribution
Adhesion	14	ZINC-FINGER CKROX_Q2		12 (85.7%)	1.93	9.89e-5	+0.46	
CDAntigens	35	GABP/ETS PU1_Q4		28 (80.0%)	1.06	1.13e-5	+0.97	
		IRF ICSBP_Q6		9 (25.7%)	1.47	2.06e-4	-0.34	
		ESTEROID_RECEPTOR PR_01		19 (54.3%)	1.02	2.88e-4	+0.06	
		GATA GATA3_03		16 (45.7%)	1.12	6.24e-4	-0.72	
Cytoskeleton	19	SRF SRF_Q5_02		6 (31.6%)	4.64	0.00e+0	+1.88	
Histocomp	19	CAAT-box/NF-Y NFY_C		9 (47.4%)	1.81	5.70e-7	+1.94	
		IRF ISRE_01		7 (36.8%)	2.77	1.75e-6	-0.65	
		SMAD3 SMAD3_Q6		13 (68.4%)	1.37	4.22e-5	+0.04	
		P53 P53_DECAMER_Q2		14 (73.7%)	1.25	5.54e-5	-1.94	
Metalion	7	MTF1 MTF1_Q4		6 (85.7%)	4.24	5.62e-11	-0.45	
		ZINC-FINGER IK3_01		7 (100.0%)	1.82	3.88e-5	-0.04	
		X-box RFX_Q6		7 (100.0%)	1.40	1.24e-4	+0.11	
		CBF CBF_02		7 (100.0%)	1.68	1.34e-4	+0.22	

Figure 6: **Non-redundant motifs in Adhesion, CD-Antigens, Cytoskeleton, Histocompatibility and Metalion co-expressed gene clusters.**

REFERENCES

Atfi, A. and Baron, R. (2008). p53 brings a new twist to the smad signaling network. *Sci Signal*, 1(26):pe33.

Bellora, N., Farré, D., and Albà, M. M. (2007a). Peaks: identification of regulatory motifs by their position in dna sequences. *Bioinformatics*, 23(2):243–244.

Bellora, N., Farré, D., and Albà, M. M. (2007b). Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. *BMC Genomics*, 8:459.

Defrance, M. and Touzet, H. (2006). Predicting transcription factor binding sites using local over-representation and comparative genomics. *BMC Bioinformatics*, 7:396.

Hai, T. and Curran, T. (1991). Cross-family dimerization of transcription factors fos/jun and atf/creb alters dna binding specificity. *Proc Natl Acad Sci U S A*, 88(9):3720–3724.

John, L. B., Yoong, S., and Ward, A. C. (2009). Evolution of the ikaros gene family: implications for the origins of adaptive immunity. *J Immunol*, 182(8):4792–4799.

Karolchik, D., Kuhn, R. M., Baertsch, R., Barber, G. P., Clawson, H., Diekhans, M., Giardine, B., Harte, R. A., Hinrichs, A. S., Hsu, F., Kober, K. M., Miller, W., Pedersen, J. S., Pohl, A., Raney, B. J., Rhead, B., Rosenbloom, K. R., Smith, K. E., Stanke, M., Thakkapallayil, A., Trumbower, H., Wang, T., Zweig, A. S., Haussler, D., and Kent, W. J. (2008). The ucsc genome browser database: 2008 update. *Nucleic Acids Res*, 36(Database issue):D773–D779.

Koizumi, S., Suzuki, K., Ogra, Y., Yamada, H., and Otsuka, F. (1999). Transcriptional activity and regulatory protein binding of metal-responsive elements of the human metallothionein-iia gene. *Eur J Biochem*, 259(3):635–642.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006). Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–D110.

Miano, J. M., Long, X., and Fujiwara, K. (2007). Serum response factor: master regulator of the actin cytoskeleton and contractile apparatus. *Am J Physiol Cell Physiol*, 292(1):C70–C81.

Miura, F., Kawaguchi, N., Sese, J., Toyoda, A., Hattori, M., Morishita, S., and Ito, T. (2006). A large-scale full-length cdna analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci U S A*, 103(47):17846–17851.

Pahl, H. L. (1999). Activators and target genes of rel/nf-kappab transcription factors. *Oncogene*, 18(49):6853–6866.

Parker, S. F., Perkins, N. D., Gitlin, S. D., and Nabel, G. J. (1996). A cooperative interaction of human t-cell leukemia virus type 1 tax with the p21 cyclin-dependent kinase inhibitor activates the human immunodeficiency virus type 1 enhancer. *J Virol*, 70(8):5731–5734.

Prieto, C., Risueno, A., Fontanillo, C., and las Rivas, J. D. (2008). Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PLoS One*, 3(12):e3911.

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

Racz, I., Nadal, X., Alferink, J., Banos, J. E., Rehnelt, J., Martín, M., Pintado, B., Gutierrez-Adan, A., Sanguino, E., Bellora, N., Manzanares, J., Zimmer, A., and Maldonado, R. (2008). Interferon-gamma is a critical modulator of cb(2) cannabinoid receptor signaling during neuropathic pain. *J Neurosci*, 28(46):12136–12145.

Reichard, J. F., Sartor, M. A., and Puga, A. (2008). Bach1 is a specific repressor of hmox1 that is inactivated by arsenite. *J Biol Chem*, 283(33):22363–22370.

Richmond, A. (2002). Nf-kappa b, chemokine gene transcription and tumour growth. *Nat Rev Immunol*, 2(9):664–674.

Schaub, M., Myslinski, E., Schuster, C., Krol, A., and Carbon, P. (1997). Staf, a promiscuous activator for enhanced transcription by rna polymerases ii and iii. *EMBO J*, 16(1):173–181.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–1050.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., and Kasprzyk, A. (2009). Biomart–biological queries made easy. *BMC Genomics*, 10:22.

Stein, B., Cogswell, P. C., and Baldwin, A. S. (1993). Functional and physical associations between nf-kappa b and c/ebp family members: a rel domain-bzip interaction. *Mol Cell Biol*, 13(7):3964–3974.

Sui, S. J. H., Fulton, D. L., Arenillas, D. J., Kwon, A. T., and Wasserman, W. W. (2007). opossum: integrated tools for analysis of regulatory motif overrepresentation. *Nucleic Acids Res*, 35(Web Server issue):W245–W252.

Tait, A. S., Butts, C. L., and Sternberg, E. M. (2008). The role of glucocorticoids and progestins in inflammatory, autoimmune, and infectious disease. *J Leukoc Biol*, 84(4):924–931.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, 10(4):252–263.

Vlieghe, D., Sandelin, A., Bleser, P. J. D., Vleminckx, K., Wasserman, W. W., van Roy, F., and Lenhard, B. (2006). A new generation of jaspar, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res*, 34(Database issue):D95–D97.

Zhu, Y., Richardson, J. A., Parada, L. F., and Graff, J. M. (1998). Smad3 mutant mice develop metastatic colorectal cancer. *Cell*, 94(6):703–714.

4

Discussion

4.1 Thesis overview

Each research article included in this thesis includes a discussion of the corresponding results. This section does not attempt to go over the same points again but to provide some general remarks and future lines of research. The temporal thread behind the execution of the PhD project can be described as follows: (a) development and web server implementation of a method to assess the positional bias of regulatory motifs in biological sequences. (b) analysis of the relationship between promoter structure and gene expression pattern in terms of the content and arrangement of regulatory motifs. (c) development and implementation of a method to identify relevant regulatory motifs in co-regulated promoters.

In the first part of this thesis a novel method to detect motifs based on their position relative to a known functional element is presented (Bellora et al., 2007a). We evaluate PEAKS on yeast promoter sequences classified in different functional categories. Many well-studied transcription factor binding motifs have been shown to have positional biases in specific datasets (Ohler et al., 2002; Marino-Ramírez et al., 2004; Xie et al., 2005; Berendzen et al., 2006). This presumably reflects specific requirements of motif-binding proteins that need to interact with each other in order to regulate transcription. The identification of significant motif peaks can be used to increase the specificity of motif prediction, provide information on the promoter structure, and help discover regulatory motifs that are specifically involved in the regulation of genes with similar expression or function. PEAKS can also be used to analyze other types of motifs that show positional biases, such as splicing sites or transposon insertion sequences. In this respect, we are aware of several groups using our server for these other purposes. There were already a number of examples of the application CRMs or positional bias in large datasets, but this was the first systematic method to obtain significant biased motifs. PEAKS was the first web application published to identify motifs solely on the basis of positional bias. Some features could be further improved to enhance the performance. For example the calculation of the empirical p-value becomes too computationally intensive for large datasets. A possible solution is to use randomization of motif localization, which can significant increase speed. Other useful features would be to upload user-defined motifs, and to integrate motifs that are very similar into a single motif cluster. PEAKS, like many similar programs, does not provide accurate results for low abundance motifs, as the small samples does not allow for statistically robust comparisons.

The second part of this thesis is focused on the analysis of mouse promoters in relation to their expression in diverse anatomical systems (Bellora et al., 2007b). Our results strongly indicate that housekeeping gene promoters have a more simple motif arrangement than the class composed of promoters driving restricted tissue expression. This is not surprising, as distinct regulators are expected to control expression in different tissue types. Previous studies on the identification of tissue-specific motifs have been based on cross-species conservation and subsequent detection of tissue enrichment (Xie et al., 2005), or on the identification of cis-regulatory modules with high tissue specific expression predictive value (Smith et al., 2006). Many of the motifs that show significant positional bias in our analysis are located within the first 100 bp upstream of the TSS. This is not surprising considering that the sequences are anchored at the TSS in this analysis, and position dependencies between interacting motif-binding proteins are expected to be more relevant at short distances. In several tissue-restricted datasets we found motifs with positional bias much further upstream. This is consistent with the high upstream sequence conservation of tissue-specific promoters when compared to housekeeping promoters (Farré et al., 2007).

In the third part of this thesis, we investigated the presence of specific motifs potentially involved in regulation of co-expressed genes. Many specific binding sites are not biased relative to the TSS, so a global motif over-representation method is required. Although previous methods had been developed for this purpose, they did no allow for use of new collections of PWMs or k-mers. We develop one such method, Motif Over-Representation (MOR). We compared the results of four alternative methods, oPOSSUM, PEAKS, TF-Explorer and MOR, in a dataset of promoters containing a large number of experimentally validated TFBSs. We used MOR in diverse datasets of human and mouse promoters of co-expressed genes to identify a set of TF candidates, for which experimental testing could be conducted.

The lack of experimental data was an important limitation for the analysis and interpretation of putative regulatory motifs in gene promoters. On the one hand, we had to rely on the available experimental TFBSs for which there is partial promoter coverage and which is biased towards binding sites of TFs involved in disease (Vaquerizas et al., 2009). This was reflected in low availability of PWMs or problems to mapping k-mer results to validated TFBSs. On the other hand, alternative TSSs can be activated under diverse conditions. A quantitative analysis of promoter usage in different human and mouse tissues showed that differentially regulated alternative TSSs are a common feature in protein-coding genes (Carninci et al., 2006). DBTSS is a growing database that contains information of TSS positions with usage frequency and cell conditions (Wakaguri et al., 2008) but contained a low number of genes compared to experimental refSeq annotations when we performed our analysis. In order to work with the largest number of gene promoters we used one TSS per gene to define the promoter. Nowadays, many experimental groups are working on genome wide projects to characterize TFBSs and TSSs under specific cell conditions. Chip-seq is an experimental method that combines chromatin immunoprecipitation (ChIP) and DNA sequencing to map in vivo protein-DNA interactions. This method provides potentially unbiased genome-wide coverage for large-scale identification of TFBS or histone positions. Diverse genome-wide experiments have been performed for specific cell conditions and TFs, including STAT1 (Robertson et al., 2007) and NRSF (Johnson et al., 2007).

Our approaches were based on the relative density and position of cis-regulatory motifs in promoters. Diverse transcription factors compete for binding to cis-regulatory elements. In many cases we identify cis-regulatory elements and a cluster of putative TFs that are likelly to bound using the available experimental evidence. However, the affinities for a specific binding site and the nuclear concentrations of the active TFs may define the actual binding outcome.

For a given locus there is a genomic regulatory context defined by several regulatory elements and the resulting transcription is the output of many regulatory signals. In addition to the cis-regulatory motifs that are found in the proximal promoter sequence other regulatory elements define those input signals: chromatin regulation, distal enhancers, insulators, promoter physical structure, alternative promoters, methylation, etc. Integrating all the elements of this regulatory context is extremely difficult and was not within the scope of the work presented in this thesis. Moreover, post-transcriptional regulation by small-RNAs adds a new layer of complexity to the transcriptional landscape.

4.2 Future directions of research

Completion of a project is only the starting point for a new project. In this section, we outline three applications of the ideas discussed above.

Characterization of dependencies between regulatory elements

Using the large numbers of experimental TFBS that will become available with new technologies, it will be possible to apply local-over-representation methods to explore the putative dependencies between regulatory motifs. Those dependencies can indicate common interactions between TFs and other regulatory elements.

Finding TFs involved in gene co-regulation using validated binding sites

In our work we have used over-represented TFBSs prediction to identify TFs in a set of co-expressed genes. However, the same method can be applied to data on real binding sites. Alternative TSS usage together with experimentally validated binding sites in specific cell conditions could be used to analyze the enrichment of particular TFs in groups of co-expressed genes.

New predictive TFBS models based on TF binding affinity

Positional interdependencies, variable space lengths, or alternative binding domains are features that can not be modeled in PWMs. A probabilistic recognition model based on the DNA binding affinity of EGR has been proposed to identify putative binding sites (Benos et al., 2002b). Protein binding microarray (PBM) analysis is an *in vitro* technique that allows quantification of the affinity of a given TF to all possible 8-mers in an unbiased manner (Berger et al., 2008). Using this technique it is possible to discover subtle preferences in transcription factor binding affinities, dependencies between positions, and alternative usage of binding motifs (Badis et al., 2009). Novel DNA-binding models based on TF binding affinities (Newburger and Bulyk, 2009) and protein amino-acid sequence (Berger et al., 2008) will improve TFBS predictions.

5

Conclusions

The research carried out during my PhD has resulted in or contributed to:

1. The development of a novel systematic approach to identify motifs that show a preferential location in DNA sequences, a strong indication of the existence of functional constraints. The method was implemented on a web server to be used by the scientific community.

2. The identification of regulatory motifs over-represented on diverse tissues and housekeeping genes. YY1 and other general TFs have strong positional biases on housekeeping genes whereas tissue-specific TFs and putative novel regulatory motifs were found on diverse tissue restricted datasets. The differences are striking and serve to illustrate that many key specific regulatory signals may be present in the proximal promoter region in mammalian genes.

3. The development of a program to identify specific motifs in promoters of co-expressed genes. Discover TFs that are specific of genes with similar expression patterns by using this program. Data from experimentally-verified TFBS support the biological relevance of our findings.

Appendices

A

Articles

In this section are gathered other articles I have collaborated in:

- Housekeeping genes tend to show reduced upstream sequence conservation (Farré et al., 2007).
- Interferon-gamma is a critical modulator of CB(2) cannabinoid receptor signaling during neuropathic pain (Racz et al., 2008).
- Origin of primate orphan genes: a comparative genomics approach (Toll-Riera et al., 2009b).
- Evolution of primate orphan proteins (Toll-Riera et al., 2009a).
- Jagged1 is the pathological link between Wnt and Notch pathways in colorectal cancer (Rodilla et al., 2009).

My contribution to these articles was always related to the identification of regulatory motifs in gene promoter sequences. Since the original articles are too long to be included here, only the first page is included.

Farré D, Bellora N, Mularoni L, Messeguer X, Albà MM.
Housekeeping genes tend to show reduced upstream sequence conservation.
Genome Biol. 2007;8(7):R140.

Racz I, Nadal X, Alferink J, Baños JE, Rehnelt J, Martín M, Pintado B, Gutierrez-Adan A, Sanguino E, Bellora N, Manzanares J, Zimmer A, Maldonado R.

Interferon-gamma is a critical modulator of CB(2) cannabinoid receptor signaling during neuropathic pain.

J Neurosci. 2008 Nov 12;28(46):12136-45.

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Albà MM.
Origin of primate orphan genes: a comparative genomics approach.
Mol Biol Evol. 2009 Mar;26(3):603-12

Toll-Riera M, Castelo R, Bellora N, Albà MM.
Evolution of primate orphan proteins.
Biochem Soc Trans. 2009 Aug;37(Pt 4):778-82.

Rodilla V, Villanueva A, Obrador-Hevia A, Robert-Moreno A, Fernández-Majada V, Grilli A, López-Bigas N, Bellora N, Albà MM, Torres F, Duñach M, Sanjuan X, Gonzalez S, Gridley T, Capella G, Bigas A, Espinosa L.
Jagged1 is the pathological link between Wnt and Notch pathways in colorectal cancer.
Proc Natl Acad Sci U S A. 2009 Apr 14;106(15):6315-20.

B

List of publications

Articles

Bellora, N., Farré, D., and Albà, M. M. **Peaks: identification of regulatory motifs by their position in DNA sequences.** *Bioinformatics*, 23(2):243-244 (2007)

Farré, D., Bellora, N., Mularoni, L., Messeguer, X., and Albà, M. M. **Housekeeping genes tend to show reduced upstream sequence conservation.** *Genome Biology*, 8(7):R140 (2007)

Bellora, N., Farré, D., and Albà, M. M. **Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters.** *BMC Genomics*, 8:459 (2007)

Racz, I., Nadal, X., Alferink, J., Baños, J. E., Rehnelt, J., Martín, M., Pintado, B., Gutierrez-Adan, A., Sanguino, E., Bellora, N., Manzanares, J., Zimmer, A., and Maldonado, R. **Interferon-gamma is a critical modulator of CB(2) cannabinoid receptor signaling during neuropathic pain.** *The Journal of Neuroscience*, 28(46):12136-12145 (2008)

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., and Albà, M. M. **Origin of primate orphan genes: a comparative genomics approach.** *Molecular Biology and Evolution*, 26(3):603-612 (2009)

Rodilla, V., Villanueva, A., Obrador-Hevia, A., Robert-Moreno, A., Fernández-Majada, V., Grilli, A., López-Bigas, N., Bellora, N., Albà, M. M., Torres, F., Duach, M., Sanjuan, X., Gonzalez, S., Gridley, T., Capella, G., Bigas, A., and Espinosa, L. **Jagged1 is the pathological link between wnt and notch pathways in colorectal cancer.** *Proceedings of the National Academy of Sciences*, 106(15):6315-6320 (2009)

Toll-Riera, M., Castelo, R., Bellora, N., and Albà, M. M. **Evolution of primate orphan proteins.** *Biochemical Society Transactions*, 37(Pt 4):778-782 (2009)

Posters

2008: Society for Molecular Biology and Evolution Annual Meeting (SMBE), Barcelona (Spain)

Authors: Nicolás Bellora, Macarena Toll-Riera, M.Mar Albà

Title: Regulatory motif conservation on eukariotic promoters

2008: Society for Molecular Biology and Evolution Annual Meeting (SMBE), Barcelona (Spain)

Authors: Macarena Toll-Riera, Nina Bosch, Nicolás Bellora, Robert Castelo, Lluís Armengol, Xavier Estivill, M.Mar Albà

Title: How do novel genes arise? Insights from mammalian genome sequence comparisons

2008: VIII Jornadas de Bioinformatica, Centro de Investigación Príncipe Felipe, Valencia (Spain)

Authors: Nicolás Bellora, Domènec Farré, M. Mar Albà

Title: Identification of regulatory motifs by positional bias in mammalian promoters.

2007: The Biology of Genomes (CSHL), New York (USA)

Authors: Domènec Farré, Nicolás Bellora, M. Mar Albà

Title: Distinct architecture of housekeeping gene promoters versus tissue-specific gene promoters

2005: 4th European Conference on Computational Biology (ECCB) , Madrid (Spain)
Authors: Nicolás Bellora, Mar Albà
Title: Characterization of the proximal promoter region

2004: International Society for Computational Biology (ISMB) & ECCB, Glasgow (Scotland)
Authors: Nicolás Bellora, Mar Albà
Title: Computational prediction of signals regulating the Herpesvirus infection gene network

2003: VIII Congreso Nacional de Virología, III Jornada de Virología de Catalunya, Barcelona (Spain)
Authors: Ria Holzerdlandt, Nicolás Bellora, Paul Kellam, Mar Albà
Title: Identificación de genes virales de interacción con el huesped mediante técnicas de genómica comparativa.

C
IUPAC Ambiguity Code

Symbol	Meaning	Origin of designation
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
M	A or C	aMino
R	A or G	puRine
W	A or T	Weak interaction (3 H bonds)
S	C or G	Strong interaction (2 H bonds)
Y	C or T	pYrimidine
K	G or T	Keto
V	A or C or G	not-T (not-U), V follows U in the alphabet
H	A or C or T	not-G, H follows G
D	A or G or T	not-C, D follows C
B	C or G or T	not-A, B follows A
N/X	G or A or T or C	aNy

References

Arnone, M. I. and Davidson, E. H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–1864.

Atchison, M. L. (1988). Enhancers: mechanisms of action and cell specificity. *Annu Rev Cell Biol*, 4:127–153.

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.-F., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R., and Bulyk, M. L. (2009). Diversity and complexity in dna recognition by transcription factors. *Science*, 324(5935):1720–1723.

Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36.

Beer, M. A. and Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, 117(2):185–198.

Bellora, N., Farré, D., and Albà, M. M. (2007a). Peaks: identification of regulatory motifs by their position in dna sequences. *Bioinformatics*, 23(2):243–244.

Bellora, N., Farré, D., and Albà, M. M. (2007b). Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. *BMC Genomics*, 8:459.

Benos, P. V., Bulyk, M. L., and Stormo, G. D. (2002a). Additivity in protein-dna interactions: how good an approximation is it? *Nucleic Acids Res*, 30(20):4442–4451.

Benos, P. V., Lapedes, A. S., and Stormo, G. D. (2002b). Probabilistic code for dna recognition by proteins of the egr family. *J Mol Biol*, 323(4):701–727.

Berendzen, K. W., Stber, K., Harter, K., and Wanke, D. (2006). Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves. *BMC Bioinformatics*, 7:522.

Berger, M. F., Badis, G., Gehrke, A. R., Talukder, S., Philippakis, A. A., Pena-Castillo, L., Alleyne, T. M., Mnaimneh, S., Botvinnik, O. B., Chan, E. T., Khalid, F., Zhang, W., Newburger, D., Jaeger, S. A., Morris, Q. D., Bulyk, M. L., and Hughes, T. R. (2008). Variation in homeodomain dna binding revealed by high-resolution analysis of sequence preferences. *Cell*, 133(7):1266–1276.

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. (2006). Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11):1429–1435.

Blanco, E., Guigó, R., and Messeguer, X. (2007). Multiple non-collinear tf-map alignments of promoter regions. *BMC Bioinformatics*, 8:138.

Bulyk, M. L., McGuire, A. M., Masuda, N., and Church, G. M. (2004). A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in escherichia coli. *Genome Res*, 14(2):201–208.

Butler, J. E. F. and Kadonaga, J. T. (2002). The rna polymerase ii core promoter: a key component in the regulation of gene expression. *Genes Dev*, 16(20):2583–2592.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engström, P. G., Frith, M. C., Forrest, A. R. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A., and Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6):626–635.

Chalkley, G. E. and Verrijzer, C. P. (1999). Dna binding site selection by rna polymerase ii tafs: a taf(ii)250-taf(ii)150 complex recognizes the initiator. *EMBO J*, 18(17):4835–4845.

Chen, K. and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and micrornas. *Nat Rev Genet*, 8(2):93–103.

Davidson, E. H. (2001). *Genomic regulatory systems: development and evolution*. Academic Press, San Diego.

ENCODE Project Consortium (2004). The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640.

Farré, D., Bellora, N., Mularoni, L., Messeguer, X., and Albà, M. M. (2007). Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol*, 8(7):R140.

Farré, D., Roset, R., Huerta, M., Adsuara, J. E., Albà, L. R. M. M., and Messeguer, X. (2003). Identification of patterns in biological sequences at the alggen server: Promo and malgen. *Nucleic Acids Res*, 31(13):3651–3653.

Fischer, J. J., Toedling, J., Krueger, T., Schueler, M., Huber, W., and Sperling, S. (2008). Combinatorial effects of four histone modifications in transcription and differentiation. *Genomics*, 91(1):41–51.

FitzGerald, P. C., Shlyakhtenko, A., Mir, A. A., and Vinson, C. (2004). Clustering of dna sequences in human promoters. *Genome Res*, 14(8):1562–1574.

Fondon, J. W. and Garner, H. R. (2004). Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A*, 101(52):18058–18063.

Gershenzon, N. I. and Ioshikhes, I. P. (2005). Synergy of human pol ii core promoter elements revealed by statistical sequence analysis. *Bioinformatics*, 21(8):1295–1300.

Hahn, S. (2008). Transcriptional regulation. meeting on regulatory mechanisms in eukaryotic transcription. *EMBO Rep*, 9(7):612–616.

Hertz, G. Z., Hartzell, G. W., and Stormo, G. D. (1990). Identification of consensus patterns in unaligned dna sequences known to be functionally related. *Comput Appl Biosci*, 6(2):81–92.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502.

Juven-Gershon, T., Hsu, J.-Y., and Kadonaga, J. T. (2006). Perspectives on the rna polymerase ii core promoter. *Biochem Soc Trans*, 34(Pt 6):1047–1050.

Kadener, S., Fededa, J. P., Rosbash, M., and Kornblihtt, A. R. (2002). Regulation of alternative splicing by a transcriptional enhancer through rna pol ii elongation. *Proc Natl Acad Sci U S A*, 99(12):8185–8190.

Kawaji, H., Frith, M. C., Katayama, S., Sandelin, A., Kai, C., Kawai, J., Carninci, P., and Hayashizaki, Y. (2006). Dynamic usage of transcription start sites within core promoters. *Genome Biol*, 7(12):R118.

King, M. C. and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116.

Kolbe, D., Taylor, J., Elnitski, L., Eswara, P., Li, J., Miller, W., Hardison, R., and Chiaromonte, F. (2004). Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res*, 14(4):700–707.

Kurokawa, H., Motohashi, H., Sueno, S., Kimura, M., Takagawa, H., Kanno, Y., Yamamoto, M., and Tanaka, T. (2009). Structural basis of alternative dna recognition by maf transcription factors. *Mol Cell Biol*, 29(23):6232–6244.

Kutach, A. K. and Kadonaga, J. T. (2000). The downstream promoter element dpe appears to be as widely used as the tata box in drosophila core promoters. *Mol Cell Biol*, 20(13):4754–4764.

Lee, S., Kohane, I., and Kasif, S. (2005). Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes. *BMC Genomics*, 6:168.

Lejeune, F. and Maquat, L. E. (2005). Mechanistic links between nonsense-mediated mrna decay and pre-mrna splicing in mammalian cells. *Curr Opin Cell Biol*, 17(3):309–315.

Luscombe, N. M., Austin, S. E., Berman, H. M., and Thornton, J. M. (2000). An overview of the structures of protein-dna complexes. *Genome Biol*, 1(1):REVIEWS001.

Marino-Ramírez, L., Spouge, J. L., Kanga, G. C., and Landsman, D. (2004). Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res*, 32(3):949–958.

Masquilier, D. and Sassone-Corsi, P. (1992). Transcriptional cross-talk: nuclear factors crem and creb bind to ap-1 sites and inhibit activation by jun. *J Biol Chem*, 267(31):22460–22466.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006). Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–D110.

Miller, O. L. and Beatty, B. R. (1969). Visualization of nucleolar genes. *Science*, 164(882):955–957.

Morris, D. R. and Geballe, A. P. (2000). Upstream open reading frames as regulators of mrna translation. *Mol Cell Biol*, 20(23):8635–8642.

Morris, K. V. (2008). Rna-mediated transcriptional gene silencing in human cells. *Curr Top Microbiol Immunol*, 320:211–224.

Newburger, D. E. and Bulyk, M. L. (2009). Uniprobe: an online database of protein binding microarray data on protein-dna interactions. *Nucleic Acids Res*, 37(Database issue):D77–D82.

Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., MacIsaac, K. D., Rolfe, P. A., Conboy, C. M., Gifford, D. K., and Fraenkel, E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet*, 39(6):730–732.

Ohler, U., chun Liao, G., Niemann, H., and Rubin, G. M. (2002). Computational analysis of core promoters in the drosophila genome. *Genome Biol*, 3(12):RESEARCH0087.

Park, Y. G., Nesterova, M., Agrawal, S., and Cho-Chung, Y. S. (1999). Dual blockade of cyclic amp response element- (cre) and ap-1-directed transcription by cre-transcription factor decoy oligonucleotide. gene-specific inhibition of tumor growth. *J Biol Chem*, 274(3):1573–1580.

Pedersen, A. G., Baldi, P., Chauvin, Y., and Brunak, S. (1998). Dna structure in human rna polymerase ii promoters. *J Mol Biol*, 281(4):663–673.

Pikaard, C. S. (2006). Cell biology of the arabidopsis nuclear sirna pathway for rna-directed chromatin modification. *Cold Spring Harb Symp Quant Biol*, 71:473–480.

Qin, Z. S., McCue, L. A., Thompson, W., Mayerhofer, L., Lawrence, C. E., and Liu, J. S. (2003). Identification of co-regulated genes through bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol*, 21(4):435–439.

Racz, I., Nadal, X., Alferink, J., Banos, J. E., Rehnelt, J., Martín, M., Pintado, B., Gutierrez-Adan, A., Sanguino, E., Bellora, N., Manzanares, J., Zimmer, A., and Maldonado, R. (2008). Interferon-gamma is a critical modulator of cb(2) cannabinoid receptor signaling during neuropathic pain. *J Neurosci*, 28(46):12136–12145.

Riechmann, J. L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O. J., Samaha, R. R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J. Z., Ghandehari, D., Sherman, B. K., and Yu, G. (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, 290(5499):2105–2110.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. (2007). Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 4(8):651–657.

Rodilla, V., Villanueva, A., Obrador-Hevia, A., Robert-Moreno, A., Fernández-Majada, V., Grilli, A., López-Bigas, N., Bellora, N., Albà, M. M., Torres, F., Dunach, M., Sanjuan, X., Gonzalez, S., Gridley, T., Capella, G., Bigas, A., and Espinosa, L. (2009). Jagged1 is the pathological link between wnt and notch pathways in colorectal cancer. *Proc Natl Acad Sci U S A*, 106(15):6315–6320.

Sandelin, A., Wasserman, W. W., and Lenhard, B. (2004). Consite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*, 32(Web Server issue):W249–W252.

Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898.

Segal, E. and Widom, J. (2009). From dna sequence to transcriptional behaviour: a quantitative approach. *Nat Rev Genet*, 10(7):443–456.

Sharov, A. A., Dudekula, D. B., and Ko, M. S. H. (2006). Cisview: a browser and database of cis-regulatory modules predicted in the mouse genome. *DNA Res*, 13(3):123–134.

Shen, W. F., Detmer, K., Simonitch-Eason, T. A., Lawrence, H. J., and Largman, C. (1991). Alternative splicing of the hox 2.2 homeobox gene in human hematopoietic cells and murine embryonic and adult tissues. *Nucleic Acids Res*, 19(3):539–545.

Singer, V. L., Wobbe, C. R., and Struhl, K. (1990). A wide variety of dna sequences can functionally replace a yeast tata element for transcriptional activation. *Genes Dev*, 4(4):636–645.

Smale, S. T. (1997). Transcription initiation from tata-less promoters within eukaryotic protein-coding genes. *Biochim Biophys Acta*, 1351(1-2):73–88.

Smith, A. D., Sumazin, P., Xuan, Z., and Zhang, M. Q. (2006). Dna motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci U S A*, 103(16):6275–6280.

Song, C. Z., Loewenstein, P. M., Toth, K., and Green, M. (1995). Transcription factor tfiid is a direct functional target of the adenovirus e1a transcription-repression domain. *Proc Natl Acad Sci U S A*, 92(22):10330–10333.

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

Sui, S. J. H., Mortimer, J. R., Arenillas, D. J., Brumm, J., Walsh, C. J., Kennedy, B. P., and Wasserman, W. W. (2005). opossum: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res*, 33(10):3154–3164.

Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Nakamura, Y., Suyama, A., Sakaki, Y., Morishita, S., Okubo, K., and Sugano, S. (2001). Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res*, 11(5):677–684.

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., and Albà, M. M. (2009a). Origin of primate orphan genes: a comparative genomics approach. *Mol Biol Evol*, 26(3):603–612.

Toll-Riera, M., Castelo, R., Bellora, N., and Albà, M. M. (2009b). Evolution of primate orphan proteins. *Biochem Soc Trans*, 37(Pt 4):778–782.

Tompa, M. (1999). An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. *Proc Int Conf Intell Syst Mol Biol*, pages 262–271.

Trinklein, N. D., Aldred, S. J. F., Saldanha, A. J., and Myers, R. M. (2003). Identification and functional analysis of human transcriptional promoters. *Genome Res*, 13(2):308–312.

van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. *Trends Genet*, 19(9):479–484.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, 10(4):252–263.

Vlieghe, D., Sandelin, A., Bleser, P. J. D., Vleminckx, K., Wasserman, W. W., van Roy, F., and Lenhard, B. (2006). A new generation of jaspar, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res*, 34(Database issue):D95–D97.

Wade, P. A. (2001). Methyl cpg-binding proteins and transcriptional repression. *Bioessays*, 23(12):1131–1137.

Wakaguri, H., Yamashita, R., Suzuki, Y., Sugano, S., and Nakai, K. (2008). Dbtss: database of transcription start sites, progress report 2008. *Nucleic Acids Res*, 36(Database issue):D97–101.

Wang, D., Hu, Y., Zheng, F., Zhou, K., and Kohlhaw, G. B. (1997). Evidence that intramolecular interactions are involved in masking the activation domain of transcriptional activator leu3p. *J Biol Chem*, 272(31):19383–19392.

Wasserman, W. W. and Fickett, J. W. (1998). Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol*, 278(1):167–181.

Workman, C. T. and Stormo, G. D. (2000). Ann-spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput*, pages 467–478.

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V., and Romano, L. A. (2003). The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, 20(9):1377–1419.

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3′ utrs by comparison of several mammals. *Nature*, 434(7031):338–345.

Yazaki, J., Gregory, B. D., and Ecker, J. R. (2007). Mapping the genome landscape using tiling array technology. *Curr Opin Plant Biol*, 10(5):534–542.

Yuh, C. H., Bolouri, H., and Davidson, E. H. (1998). Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, 279(5358):1896–1902.

Notes