

Doctoraatsproefschrift nr. 954 aan de faculteit Bio-ingenieurswetenschappen van de K.U.Leuven

DETECTION OF REGULATORY MOTIFS BASED ON COEXPRESSION AND PHYLOGENETIC FOOTPRINTING

Valerie STORMS

Promotors:

Prof. dr. ir. K. Marchal (promotor)

Prof. dr. ir. B. De Moor (co-promotor)

Leden van de examencommissie:

Prof. dr. ir. R. Schoonheydt (voorzitter)

Prof. dr. ir. Y. Moreau

Prof. dr. ir. J. Michiels

Dr. ir. P. Monsieus

Dr. ir. L. Verlinden

Proefschrift voorgedragen tot
het behalen van de graad
van Doctor in de
Bio-ingenieurswetenschappen

Maart 2011

© 2009 Katholieke Universiteit Leuven, Groep Wetenschap & Technologie, Arenberg Doctoraatsschool, W. de Croylaan 6, 3001 Heverlee, België

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaandelijke schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

ISBN 978-90-8826-187-9
D/ 2011/11.109/10

Voorwoord

Het is vrij onwezenlijk je doctoraatswerk af te ronden, hetgeen gepaard gaat met je dank te betuigen aan iedereen die ertoe heeft bijgedragen. Het is een flashback naar vier jaar geleden en alles wat er in de vier volgende jaren gebeurd is. Alles begon toen Pieter Monsieurs, mijn toenmalig thesisbegeleider, polste of ik graag een doctoraat zou beginnen binnen de Bio-informatica groep van Prof. Kathleen Marchal. Dit kwam voor mij eerder overwacht, maar de fijne band met Pieter en Kathleen en hun passie voor wetenschappelijk onderzoek werkte aanstekelijk en dus heb ik niet lang getwijfeld.

Pieter, het onderwerp motiefdetectie heb ik van jou doorgekregen, evenals een groot deel van de expertise. Ik heb het heel fijn gevonden om jou als mentor en vriend te hebben en nu ook als jurylid.

De mogelijkheid om mijn doctoraat te starten en ook te voltooien heb ik te danken aan mijn promotor Prof. Kathleen Marchal. Kathleen, bedankt voor het helpen uitstippelen van mijn onderzoek, het kritisch beoordelen en aanvullen ervan. Ook bedankt voor jou vertrouwen in mij, hetgeen me meer vertrouwen heeft gegeven in mezelf. Daarnaast vond ik het ook super dat ik de mogelijkheid kreeg om les te geven aan de K.U.Leuven en om deel te nemen aan internationale conferenties. Geen moeite was jou teveel, ik zal nooit vergeten hoe je in Boston mijn ontbijt op de kamer kwam brengen zodat ik mijn kine oefeningen niet moest onderbreken, dit is slechts één voorbeeldje van hoe je altijd rekening hield met mijn gezondheid. Dat heeft het voor mij mogelijk gemaakt mijn mucoviscidose te combineren met mijn doctoraat.

Ook mijn collega's ('vrienden') bio-informatica stonden één voor één altijd klaar om te helpen, zowel met praktische zaken maar ook voor emotionele steun. Dankzij jullie, heb ik me heel erg thuis gevoeld in onze onderzoeksgroep en ben ik zelfs voor de eerste maal mee op groeps-weekend geweest. Dankje Riet, Aminael, Kristof, Carolina, Inge, Peyman, Sunny, Fu, Lore, Yan, Pieter, Ivan, Tim, Hui, Lyn, Abeer en Wouter. Marleen, ook jij was een heel fijne collega, ik heb veel van jou geleerd en bewonder je aanpak en moed om vanuit Qatar je onderzoek te doen. Ik hoop dat je volhoudt en je welverdiende doctoraatstitel behaalt.

Verder wens ik ook mijn co-promotor Prof. Bart De Moor en mijn assessoren Prof. Jan Michiels en Yves Moreau te bedanken voor hun opvolging van mijn doctoraatswerk en het kritisch evalueren van mijn doctoraatstekst. Prof. Robert Schoonheydt wil ik bedanken om mijn jury voor te zitten.

Een belangrijk deel van mijn doctoraat kwam tot stand dankzij de samenwerking met Legendo. Dankje Prof. Annemieke Verstuyf, Lieve Verlinden, Guy Eelen en Els

Vanoirbeek voor de leerrijke discussies. Het was voor mij heel fijn om op een biologische dataset onderzoek te doen en de resultaten met jullie te bespreken. Lieve, fijn dat je tijd kon maken om mijn manuscript na te lezen en in mijn jury te zetelen.

Mama en papa, Vicky, Daan en Alexis, jullie steun en liefde is alles voor mij! Vicky, jij was mijn vurigste supporter, mijn bondgenootje in dit leven. Hoe ver je ook bent, je bent voor altijd bij mij. Liesje, ik hou van je, dankje om er altijd te zijn. Aurelie, Myriam et Benoit, votre chaleur et amour me donnent l'espoir pour l'avenir, je vous aime. Ann en Paul, Lien en Nele, bedankt om er voor mij en Daan te zijn. Ingrid, jou wil ik bedanken voor de goede zorgen jaar in, jaar uit. Ook de rest van mijn familie, en vrienden, zowel uit Limburg als uit Leuven, dankje voor alle steun, om er steeds te zijn en voor het samen genieten.

Valerie

Abstract

Unraveling the mechanisms that regulate gene expression is a major challenge in biology. An important task in this challenge is to identify regulatory motifs or short sequences in the DNA that serve as binding sites for transcription factors (TFs). The first computational methods developed for the discovery of regulatory motifs searched for an overrepresented motif in a set of genes that were believed to contain several binding sites for the same TF (e.g. a set of coregulated genes from a single genome). But with the growing number of sequenced genomes, detecting motifs through ‘phylogenetic footprinting’ became feasible and the next generation of motif discovery algorithms has therefore integrated the use of orthology evidence in addition to coregulation information. Moreover, the more advanced motif discovery algorithms explicitly model the phylogenetic relatedness between the orthologous input sequences and thus should be well adapted towards using orthologous information.

In a first part of the study we evaluated the conditions under which complementing coregulation with orthologous information improves motif discovery for the class of probabilistic motif discovery algorithms with an explicit evolutionary model. We designed specific datasets, both synthetic and real, essential for the benchmarking of motif discovery algorithms that integrate orthologous information. Our results show that the nature of the used algorithm is crucial in determining how to exploit multiple species data in the best way to improve motif discovery performance. The use of an integrated evolutionary model that depends on reliable alignments of hard to align intergenic sequences seems to be the major bottleneck.

In a second part of the study we developed a complete workflow for motif discovery in eukaryotes: PHYLO-MOTIF-WEB. This workflow is unique as it allows for integrating epigenetic information (e.g. nucleosome occupancy and histone modifications) to guide the motif search to putative regulatory regions in the DNA, a necessary step considering the long non-coding sequences in eukaryotes. An asymmetric clustering algorithm, FuzzyClustering, was developed to summarize the results of multiple advanced motif discovery algorithms into an ensemble solution. PHYLO-MOTIF-WEB is easy accessible for non-expert users through a web server.

Finally, we applied PHYLO-MOTIF-WEB on a biological case to investigate the molecular mechanisms underlying the antiproliferative effects of vitamin D₃ on both human and mouse cell lines. We predicted *de novo* the regulatory motifs of some known TFs that possibly can be involved in the vitamin D₃ induced pathway. Further research is necessary to validate those predictions. Our results also show the potential of combining the results of multiple motif discovery algorithms, as a consequence of the diversity in their predictions.

Korte inhoud

Het ontrafelen van de mechanismen die genexpressie regelen is een grote uitdaging in de biologie. Een belangrijke taak binnen deze uitdaging is het identificeren van regulatorische motieven of korte DNA-sequenties die dienen als herkeningsplaats voor transcriptionele regulatoreiwitten. De eerste computationele methoden ontwikkeld voor de detectie van regulatorische motieven zochten naar een overgerepresenteerd motief in een reeks genen die verondersteld werden meerdere herkeningsplaatsen voor éénzelfde regulator te bevatten (bv. cogereguleerde genen afkomstig uit één organisme). Maar door de toename van het aantal gesequente genomen werd het mogelijk om *fylogenetische footprinting* toe te passen voor het detecteren van motieven, met als gevolg dat de volgende generatie motiefdetectie-algoritmen orthologe informatie integreert naast het gebruik van coregulatie informatie. De meest geavanceerde motiefdetectie-algoritmen modeleren ook de fylogenetische verwantschap tussen de orthologe inputsequenties waardoor deze algoritmen geschikt zouden moeten zijn voor het integreren van orthologe informatie.

In een eerste deel van de studie werden de voorwaarden geëvalueerd waaronder het combineren van coregulatie met orthologe informatie motiefdetectie verbetert voor de groep van probabilistische motiefdetectie-algoritmen met een expliciet evolutiemodel. Hiervoor werden nieuwe, geschikte datasets ontwikkeld, zowel synthetische als biologische, essentieel voor de benchmarking van motiefdetectie-algoritmen die orthologe informatie integreren. Onze resultaten illustreren dat de aard van het gebruikte motiefdetectie-algoritme essentieel is om mee te bepalen hoe orthologe informatie van meerdere organismen kan gebruikt worden om motiefdetectie te optimaliseren. Het gebruik van een geïntegreerd evolutiemodel dat afhangt van een betrouwbare alignering van moeilijk te aligneren intergenische sequenties schijnt het belangrijkste knelpunt te zijn.

In een tweede deel van de studie werd een volledig werkschema ontwikkeld voor motiefdetectie in eukaryoten: PHYLO-MOTIF-WEB. Dit werkschema is uniek aangezien het de integratie van epigenetische informatie mogelijk maakt (b.v. nucleosoom bezetting en histon modificaties) om op die manier de zoektocht naar motieven te sturen naar regio's in het DNA die mogelijk een regulatorische functie hebben, een noodzakelijke stap in eukaryoten omwille van de lange intergenische regio's. Een asymmetrisch clustering-algoritme, FuzzyClustering, werd ontwikkeld om de resultaten van meerdere geavanceerde motiefdetectie-algoritmen samen te vatten in een *ensemble* oplossing. PHYLO-MOTIF-WEB is via een webserver gemakkelijk toegankelijk voor gebruikers onervaren in motiefdetectie.

Tot slot hebben we PHYLO-MOTIF-WEB toegepast op een biologische dataset om het moleculair mechanisme onderliggend het antiproliferatieve effect van vitamine D₃ op zowel humane als muis cellijnen te onderzoeken. We voorspelden *de novo* de regulatorische motieven van enkele gekende regulators die mogelijk een rol spelen in de door vitamine D₃ geïnduceerde *pathways*. Verder onderzoek is nodig om deze predicties te valideren. Onze resultaten tonen het potentieel van het combineren van de resultaten van meerdere motiefdetectie-algoritmen, als gevolg van hun predictieve diversiteit.

Abbreviations

API	application programming interface
bp	base pairs
cDNA	complementary DNA
ChIP	chromatin immunoprecipitation
CRM	<i>cis</i> -regulatory module
Cs	consensus score
DHS	DNase hypersensitive sites
DNA	deoxyribonucleic acid
EM	Expectation-Maximization
ENCODE	Encyclopedia of DNA Elements
EPD	Eukaryotic Promoter Database
FDR	false discovery rate
FN	false negative
FP	false positive
GO	Gene Ontology
GTFs	general transcription factors
HMM	Hidden Markov Model
IC	information content
IUPAC	International Union of Pure and Applied Chemistry
MAP	maximum <i>a posteriori</i>
MASS	multiply aligned sequence set
MEME	Multiple EM for Motif Elicitation
MCMC	Markov Chain Monte Carlo
mRNA	messenger RNA
NHR	nuclear hormone receptor
PG	Phylogibbs
PIC	pre-initiation complex
PPV	positive predictive value
PS	Phylogenetic sampler
PSSM	position specific scoring matrix
PWM	position weight matrix
RNA	ribonucleic acid
RNAP	RNA polymerase
RR	recovery rate
rRNA	ribosomal RNA
RXR	retinoid X receptor

Abbreviations

SCPD	<i>Saccharomyces cerevisiae</i> Promoter Database
Sens	sensitivity
SPEC	specificity
spPPV	species-dependent PPV
spSens	species-dependent Sens
TF	transcription factor
TFBS	transcription factor binding site
TG	target gene
TP	true positive
TSS	transcription start site
UCSC	University of California Santa Cruz
UTR	untranslated region
VDR	vitamin D ₃ receptor
VDRE	vitamin D ₃ response element
1 α ,25(OH) ₂ D ₃	1 α ,25-dihydroxyvitamin D ₃
WM	weight matrix

Table of contents

Voorwoord	i
Abstract.....	iii
Korte inhoud.....	v
Abbreviations	vii
Table of contents	ix
Chapter 1 Introduction.....	1
1.1 Context of this thesis.....	1
1.2 Main players in transcriptional regulation	1
1.2.1 Prokaryotes versus eukaryotes.....	1
1.2.2 Modeling TF-DNA interactions.....	3
1.2.3 High-throughput experimental methods to uncover TF-DNA interactions	5
1.3 Computational approaches towards identifying TF-DNA interactions	8
1.3.1 Algorithms based on transcriptional coregulation	8
1.3.2 Algorithms based on comparative genomics	11
1.3.3 Algorithms for combinatorial motif discovery	13
1.4 The role of chromatin in transcriptional regulation	14
1.4.1 Histone modifications	14
1.4.2 DNA methylation and CpG islands	16
1.4.3 Epigenetic information in computational motif discovery	19
1.5 Objectives of the thesis	21

Chapter 2 Probabilistic motif discovery algorithms that incorporate

phylogeny23

- 2.1 Introduction..... 23
- 2.2 Models used by PG and PS..... 28
 - 2.2.1 Input sequences..... 28
 - 2.2.2 Motif model 28
 - 2.2.3 Background model..... 30
 - 2.2.4 Evolutionary model..... 31
- 2.3 The algorithms underlying PG and PS..... 32
 - 2.3.1 Algorithms to sample the search space 32
 - 2.3.2 Scoring methods..... 34
 - 2.3.3 Solutions and posterior probabilities 37
- 2.4 Discussion..... 38

Chapter 3 The effect of orthology and coregulation on detecting regulatory

motifs41

- 3.1 Introduction..... 41
- 3.2 Materials and Methods..... 41
 - 3.2.1 Motif discovery algorithms and parameter settings..... 41
 - 3.2.2 Synthetic datasets..... 42
 - 3.2.3 Real datasets..... 43
 - 3.2.4 Performance and quality measures 44
- 3.3 Results..... 45
 - 3.3.1 Design of the test datasets 45

Table of contents

3.3.2	Motif discovery in the coregulation space	47
3.3.3	Motif discovery in the combined coregulation-orthology space	48
3.3.4	Motif discovery in the orthologous space	54
3.4	Discussion	56
Chapter 4 PHYLO-MOTIF-WEB: an ensemble workflow on the web for <i>de novo</i> discovery of DNA binding sites using phylogeny		
61		
4.1	Introduction.....	61
4.2	Materials and Methods.....	63
4.2.1	User input.....	63
4.2.2	Additional information sources.....	64
4.2.3	Motif discovery following an ensemble strategy.....	65
4.2.4	Post-processing	67
4.3	The PHYLO-MOTIF-WEB web server.....	68
4.3.1	The ‘Run it’ webpage.....	68
4.3.2	The ‘Results’ webpage.....	70
4.4	Discussion	74
Chapter 5 <i>De novo</i> motif discovery in vitamin D₃ regulated genes		
77		
5.1	Introduction.....	77
5.2	Results.....	79
5.2.1	Microarray analysis.....	79
5.2.2	Identification of <i>cis</i> -regulatory elements	82
5.3	Material and Methods	95
5.3.1	Microarray analysis.....	95

Table of contents

5.3.2	Identification of <i>cis</i> -regulatory elements	96
5.4	Discussion	100
5.4.1	<i>De novo</i> motif discovery	100
5.4.2	<i>Cis</i> -regulatory modules	101
5.4.3	Future perspectives	102
Chapter 6 General discussion and perspectives		103
6.1	General discussion	103
6.2	Perspectives	109
6.2.1	The mode of action of vitamin D ₃	109
6.2.2	Integrating multiple information sources to predict TF binding	110
Appendix - Supplementary materials		111
Reference List		137
Publication list		151
Curriculum vitae		151

Chapter 1 Introduction

1.1 Context of this thesis

All cells of a living multi-cellular organism share the same DNA. Yet, they manifest tremendous variability in their structure, activities and interactions. The same applies for single-cell organisms, such as prokaryotes, since they can manifest many different phenotypes in response to environmental cues. Those variations arise through the differential deployment of the cell's genetic toolkit, namely differences in the expression of the genes. For most protein-coding genes the level of gene expression is mainly controlled at the level of transcription (Roeder, 2003). Specialized proteins, called transcription factors (TFs), bind regulatory DNA elements in a sequence-specific manner and, once bound, modulate the expression of neighboring genes. As straightforward as this may sound, years after sequencing the first genome, we still know very little about how this regulatory information is actually encoded in the genome. Deciphering the basic principles of transcriptional regulation underlying a living cell is a major challenge in biology. Such knowledge would allow us to better understand how cells work, how they respond to external stimuli and what goes wrong in diseases like cancer (which often involves disruption of gene regulation), and how they can be fought.

1.2 Main players in transcriptional regulation

Transcription is the process during which genetic information is transcribed from DNA to RNA. In all species, transcription begins with the binding of the RNA polymerase complex to a special DNA sequence at the beginning of the gene, known as the promoter. In this section we discuss the activation and repression of the RNA polymerase complex in both prokaryotes and eukaryotes (§ 1.2.1). TFs appear to be the main players in transcriptional regulation for both groups of organisms. As TFs act by binding to specific regions in the DNA we introduce the 'regulatory motif' as the TF binding specificity model (§ 1.2.2) and refer to high-throughput experimental methods to identify TF binding sites across the genome (§ 1.2.3).

1.2.1 Prokaryotes versus eukaryotes

In prokaryotes, all transcription is performed by a single type of RNA polymerase. This RNA polymerase contains four catalytic subunits and a single regulatory subunit, known as the sigma factor. Interestingly, several distinct sigma factors have been identified, and each of these oversees transcription of a unique set of genes. Sigma factors are thus discriminatory, as each binds a particular set of promoter sequences by recognizing a

specific DNA binding location. For example the major vegetative sigma 70 factor of *Escherichia coli* recognizes two conserved hexamers located at nucleotide positions -10 and -35 relative to the gene transcription start site (TSS), while the promoter regions recognized by sigma 54 (a sigma factor involved in *i.a.* nitrogen fixation) contain two conserved hexamers at positions -26 and -11 (also referred to as -24/-12 promoters) (Fischer, 1994). Therefore, while prokaryotes accomplish transcription of all genes using a single kind of RNA polymerase, the use of different sigma factor subunits provides an extra level of control that permits the cell to induce and repress different gene expression programs. However, this global regulation mechanism only permits to respond to general conditions, while often a more specific reaction is required. Therefore, in addition to the RNA polymerase, TFs that respond to specific conditions in the environment, can bind specific regions in the promoter region and facilitate or inhibit the binding and opening of the RNA polymerase and thus influence the transcription rate of the corresponding gene.

In prokaryotes genes are organized into operons, or clusters of coregulated genes. In addition to being physically close on the genome, these genes are regulated by the same promoter such that they are all turned on or off together. Grouping related genes under a common control mechanism allows prokaryotes to rapidly adapt to changes in the environment.

Eukaryotic cells are more complex than prokaryotes in many ways, including transcriptional regulation. In eukaryotic cells the DNA is wrapped around nucleosomes, globular complexes of histone proteins, to form the tightly packed chromatin (Luger et al., 1997). Chromatin structure plays a functional role in transcriptional regulation, by modulating the affinity of DNA to the transcriptional machinery (see § 1.4). Because of this tight packaging of DNA, RNA polymerase II, which is responsible for the transcription of protein-coding genes in eukaryotes, does not directly recognize the transcription start site (TSS) of the gene it will transcribe (Lee and Young, 2000). To guide the DNA binding of RNA polymerase II, other factors called general TFs (GTFs), will first assemble on the core promoter region, which includes the TSS of the gene as well as other binding sites recognized by different subunits of the GTFs (e.g. the TATA box) (Thomas and Chiang, 2006). After the GTFs form a complex with the core promoter, RNA polymerase II binds to it, forming a transcription initiation complex (TIC).

The main players regulating the formation and activity of the TIC can be classified into two groups based on their mode of activity (Narlikar and Ovcharenko, 2009):

- **Trans-acting factors** (not part of the DNA) are TFs, like activators and repressors that bind the DNA directly, usually in a sequence-specific manner, and influence the rate of transcription. Also non-DNA-binding proteins, like co-activators and co-repressors, recruited to the DNA by protein-protein-interactions, can act in a *trans*-manner to influence transcription e.g. chromatin remodeling proteins.
- **Cis-acting elements** are regions along the DNA that facilitate the binding of activators or repressors, or are responsible for changing the chromatin structure to either activate or repress transcription. Promoters, enhancers, silencers and insulators constitute the *cis*-acting elements in eukaryotic DNA. Those regulatory elements can be located thousands of bases away from the TSS, making it much more complex to identify them compared to the regulatory elements in prokaryotes.

Transcriptional regulation in eukaryotes is thus a collaborative effort between different TFs, chromatin remodeling complexes and other non-DNA-binding co-factors. These proteins can be either ubiquitous or cell type specific, but together activate or repress genes by targeting specific regulatory elements.

1.2.2 Modeling TF-DNA interactions

There are many ways of modeling the sequence specificity by which a TF binds to the DNA. Such a TF binding site model is also called a regulatory motif. To build a regulatory motif one usually starts from an alignment of experimentally defined TF binding sites that can be extracted from databases like TRANSFAC (Matys et al., 2006) and JASPAR (Bryne et al., 2008) (see Figure 1.1 A). The simplest representation of a regulatory motif is the consensus sequence, a string representation that contains at each position the most frequent nucleotide (see Figure 1.1 B) (Stormo, 2000). To allow for degeneracy at a specific position, the consensus sequence representation uses the IUPAC codes (Cornish-Bowden, 1985) for polymorphic nucleotides (see Table S1 in the Supplementary Materials). To capture variability of TF binding sites in a quantitative manner, the regulatory motif can also be represented as a matrix model. The simplest form of the matrix model is a count matrix that contains the nucleotide counts at each position of the binding site alignment (see Figure 1.1 C). From this count matrix other types of matrices can be deduced like the position probability matrix described by Thijs *et al.* (Thijs et al., 2002a) or the position weight matrix (PWM) also known as the position specific scoring matrix (PSSM) described by Stormo (Stormo, 2000).

Consensus sequences and simple matrix models like the PWM, ignore some of the complexities of protein–DNA interactions as they assume that positions within the binding site are independent (Stormo, 2000). While it is possible to use more complex matrix models to capture such inner dependencies, like in the study of King and Roth (King and Roth, 2003), this requires more TF binding site data to estimate the model's parameters. In case the data are limited, the risk exists that those complex models over fit the data and yield a poor representation of TF binding specificity. An important study by Benos *et al.* (Benos et al., 2002) suggested that while the consensus sequence and PWM may not fully capture all the subtleties of a protein's binding specificity, these simple and easily interpretable models usually provide a very good approximation to reality.

A more visual representation of a regulatory motif is a sequence logo (see Figure 1.1 D). A sequence logo consists of stacks of letters, one stack for each position in the alignment of binding sites. The overall height of each stack indicates the sequence conservation at that position (measured in bits), whereas the height of symbols within the stack reflects the relative frequency of the corresponding nucleotide at that position (Crooks et al., 2004).

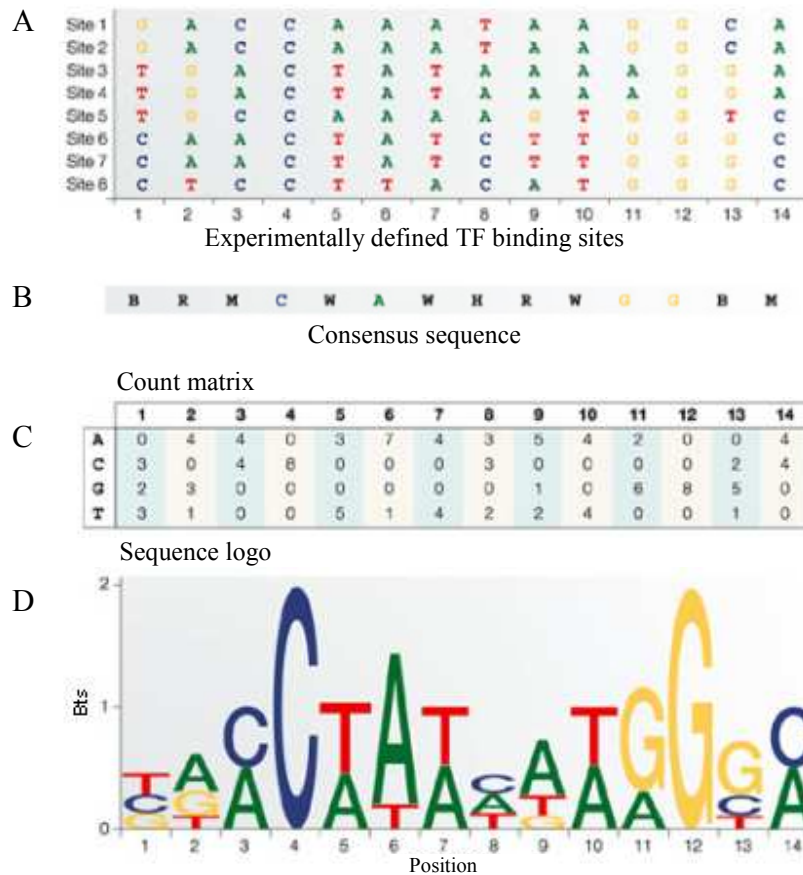


Figure 1.1 Modeling TF binding sites. (A) The aligned set of experimentally defined binding sites used to build a regulatory motif. The regulatory motif can be represented as (B) a consensus sequence, (C) a matrix model which represents the number of times each nucleotide is counted at each position of the alignment or (D) by a sequence logo, visually showing the information content and conservation at each of the alignment positions. Adapted from (Wasserman and Sandelin, 2004).

1.2.3 High-throughput experimental methods to uncover TF-DNA interactions

The classical experimental approach to characterize transcriptional regulatory elements in the DNA uses reporter gene assays and has been successful in various animal models (Allende et al., 2006; Muller et al., 1997). But assay-based methods are usually time-consuming and expensive. The development of less expensive, high-throughput experimental methods allows mapping of regulatory elements on a genome-wide scale and allows a global view of their biological roles. In this paragraph we highlight ‘chromatin immunoprecipitation’ (ChIP) as a high-throughput experimental method that can be used to map regulatory elements.

ChIP is a common method for detecting interactions between a protein and a DNA sequence *in vivo* (Kim and Ren, 2006). In recent years, this method has been combined with DNA microarrays (Derisi et al., 1997) and other high-throughput technologies to enable genome-wide identification of DNA-binding sites for various nuclear proteins.

The ChIP method treats living cells with a cross-linking agent, usually formaldehyde, which fixes proteins to their DNA substrates inside cells (Figure 1.2 a). Chromosomes are then extracted and fragmented by physical shearing or enzymatic digestion. Specific DNA sequences associated with a particular protein are isolated by immuno-affinity purification using a specific antibody against the protein (Figure 1.2 b). The purified DNA fragments are then assayed by microarrays or direct sequencing strategies (Figure 1.2 c). Combining ChIP and DNA microarrays (also referred to as ChIP-chip) has some limitations as the microarrays (mostly tiled genomic microarrays) do not contain repetitive DNA and they are affected by problems with cross-hybridization and varying oligomer affinities that cause background noise.

To overcome these problems, coupling ChIP with massively parallel sequencing of the recovered DNA fragments has been developed as a preferred strategy. Compared to ChIP-chip, ChIP followed by sequencing has an increased resolution in the detected binding sites and is much cheaper, especially for large genomes. Several forms of ChIP followed by sequencing have rapidly been developed and implemented e.g. Serial Analysis of Gene Expression (ChIP-SAGE) (Roh et al., 2005), Paired-End Tag (ChIP-PET) (Wei et al., 2006) and most recently ChIP-seq (Robertson et al., 2007), a very cost-effective strategy that makes use of the new sequencing technologies of Illumina (formerly Solexa) and Roche (454 Life Sciences). The main limitation of ChIP in general, is that a specific antibody needs to be created for the protein of interest; in many cases, such antibodies do not exist. Also the specificity of the antibody is critical for generating high-quality data. A major advantage is that the whole genome is tested for *in vivo* binding of the protein of interest, as *in vitro* experiments can never replicate *in vivo* conditions faithfully.

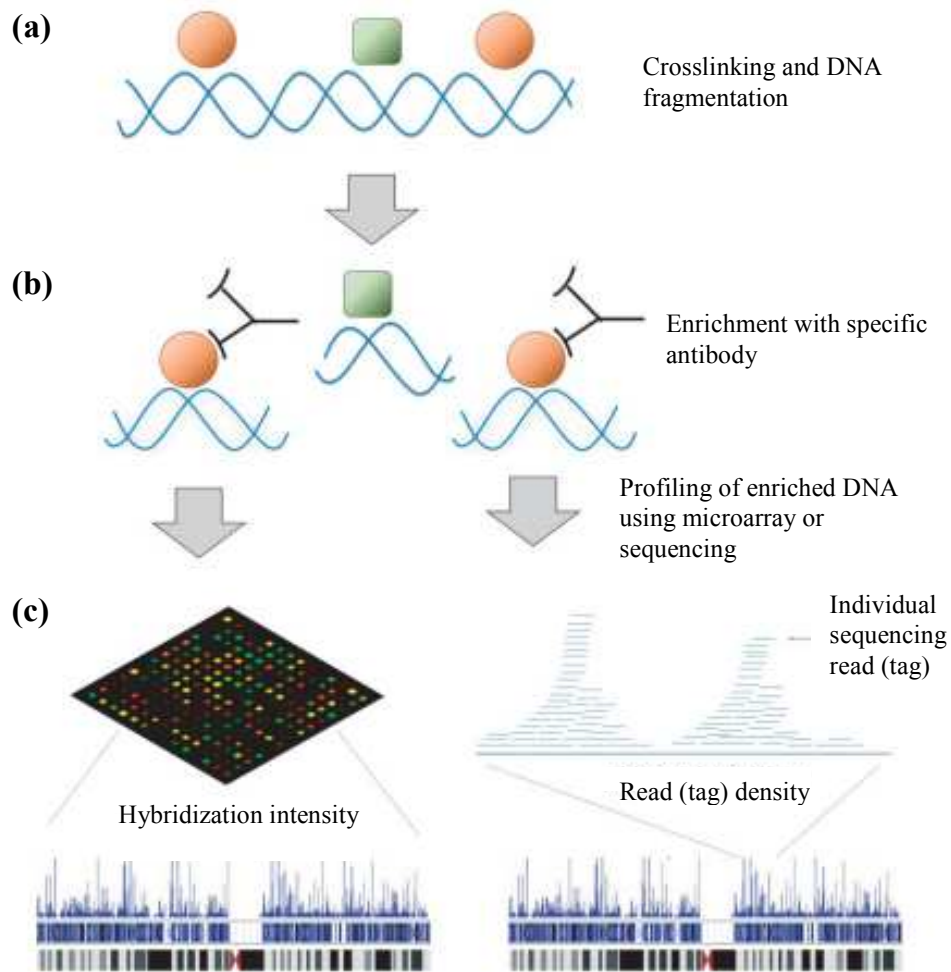


Figure 1.2 Overview of ChIP-chip and ChIP-seq. (a) Reversible cross-linking of DNA and protein is performed by treating the DNA–protein complex with formaldehyde. The cross-linked DNA–protein complex is fragmented by sonication. (b) An antibody specific to the protein of interest is used to enrich the DNA segments bound to the protein. (c) The purified DNA is profiled using a microarray (ChIP-chip) or direct sequencing (ChIP-seq). Adapted from (Kim and Park, 2011).

Depending on the function of the profiled protein, ChIP can detect different kinds of regulatory elements. One of the applications is the identification of direct downstream targets of TFs like e.g. STAT1 in human HeLa S3 cells by ChIP-seq (Robertson et al., 2007) or the binding sites of 203 transcriptional regulators in yeast by ChIP-chip (Harbison et al., 2004). These studies are revealing novel insights about general distribution of TF binding sites and complexity of regulatory mechanisms. ChIP-chip was also used to map active promoter regions in the human genome by profiling a component protein of the pre-initiation complex (PIC) (Kim et al., 2005a; Kim et al., 2005b). Insulator elements, which affect transcription by restricting enhancers from activating unrelated promoters, can be retrieved by ChIP-chip by profiling CTCF, a protein known to mediate insulator activity in vertebrates (Heintzman et al., 2009; Kim et al., 2007).

Human enhancers were located by targeting the transcriptional activator protein p300 (Heintzman et al., 2009). ChIP also plays an important role in unraveling chromatin structure by targeting covalent chromatin modifications (Schones and Zhao, 2008) (see § 1.4).

1.3 Computational approaches towards identifying TF-DNA interactions

As mentioned in the previous paragraph (§ 1.2) TFs bind DNA in a sequence-specific manner, and hence, detecting the binding specificities of individual TFs constitutes a first computational challenge: ‘*de novo* motif discovery’. These binding specificities or the regulatory motifs can then be used to determine genome-wide potential binding sites of the TF, which leads to a second computational challenge: ‘motif scanning’ (review article by (Wasserman and Sandelin, 2004)). In this paragraph we focus on the first challenge: identifying the locations of TF binding sites in a set of regulatory regions to define the TF target genes and its regulatory motif. Over the past few years, numerous computational strategies have become available for motif discovery and we classify them based on the information sources they use. Initially motif discovery started from a set of genes coregulated at the transcriptional level inferred from coexpression information or high-throughput experimental approaches (§ 1.3.1). The use of evolutionary conservation, information that can be extracted by comparative genomics, proved to be a successful extension for motif discovery (§ 1.3.2). More recently, methods have been developed to analyze composite regulatory elements, i.e. modules consisting of multiple binding sites bound by different TFs (§ 1.3.3). This paragraph reflects the trends in *de novo* motif discovery with focus on the used information sources rather than a complete overview of all methods developed in the field.

1.3.1 Algorithms based on transcriptional coregulation

High-throughput gene expression measurements by micro-arrays and ChIP experiments allow the identification of coexpressed and coregulated genes, respectively. In case of coexpressed genes, it is assumed that coexpression arises mainly from transcriptional coregulation. As coregulated genes are known to share some similarities in their regulatory mechanism, possibly at the transcriptional level, their promoter regions might contain binding sites for a common TF. Usually, a user provides a collection of non-coding regions of genes that are believed to be coregulated, and the computational tool identifies short DNA patterns (~ TF binding sites) that are statistically overrepresented in those regions (see Figure 1.3, on top). A statistically overrepresented pattern means a pattern that occurs more often than one would expect by chance, e.g. in the non-coding regions of a set of random genes (Figure 1.3, at the bottom).

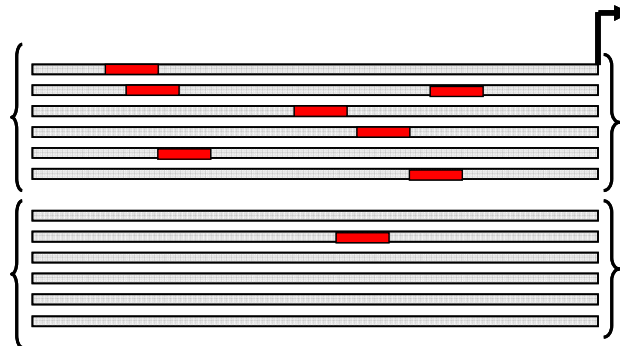


Figure 1.3 Overrepresented patterns in the regulatory sequences of coregulated genes. A set of transcriptionally coregulated genes can be inferred from high-throughput coexpression or ChIP studies. Computational algorithms were developed to identify short patterns (red) enriched among the promoters of those groups (top), in comparison to a set of promoters from random genes (bottom). The arrow indicates the transcription start sites of the corresponding genes.

We can categorize motif discovery algorithms into two major groups based on how they represent the regulatory motif: *enumerative methods*, representing the regulatory motif as a degenerated consensus sequence and *probabilistic methods* that represent the regulatory motif as a matrix model. There is also a smaller third group that represents regulatory motifs by using hidden Markov models. The latter models allow binding sites of varying length and correlations between bases at neighboring positions (Sandelin and Wasserman, 2005) (not further discussed here).

Enumerative algorithms examine the number of exact occurrences of all n-length patterns in the input sequences, and calculate which ones are most overrepresented (van Helden et al., 1998). But as the occurrence of exact patterns is too rigid for most real-world TF binding sites, one can also search for degenerate patterns (Sinha and Tompa, 2003; Tompa, 1999) and in addition, tools like Weeder (Pavesi et al., 2001) apply efficient data structures like suffix trees to decrease runtime for long DNA patterns. Enumerative approaches exhaustively explore the whole search space and therefore retrieve the global optimum. Because of the exact enumeration, these methods are limited to relatively simple patterns like short motifs with very few variations in the binding sites. In a recent assessment by Tompa *et al.* (Tompa et al., 2005), it was shown that an enumerative method like Weeder (Pavesi et al., 2001) can achieve very good results in predicting known motifs.

Probabilistic approaches represent the TF binding specificity by a matrix model (~motif model) and the remainder of the sequence is modeled by a background model. To find the parameters of the motif model, these methods use maximum likelihood estimation. The two most frequently used methods for maximizing the likelihood are expectation maximization (EM) and Gibbs sampling.

- EM (Bailey and Elkan, 1994; Lawrence and Reilly, 1990) is a local optimization procedure that is guaranteed to monotonically improve the expected likelihood, but it is sensitive to its initialization point and is therefore not guaranteed to converge to the global maximum. For this reason, motif discovery programs that use EM will typically restart the optimization from many distinct initialization points to improve the chances of converging to the global maximum. Multiple restarts also improve the chances of finding biologically relevant motifs that may not necessarily correspond to the global maximum. Interesting heuristics for selecting reasonable initialization points have been developed (Blekas et al., 2003). As example, we mention the popular program MEME (Bailey and Elkan, 1994) that uses EM in combination with multiple initializations to retrieve the global optimum.
- Gibbs sampling (Liu et al., 1995; Lawrence et al., 1993), the stochastic variant of EM, is now widely used in motif discovery. Gibbs sampling tends to provide a more robust optimization of the model parameters in order to avoid local optima. For stochastic algorithms like Gibbs sampling, multiple searches have to be performed within the input dataset, in order to confirm that the same matrix models are discovered starting from different initializations. Several improved versions of the initial Gibbs sampler (Lawrence et al., 1993) are now available like MotifSampler (Thijs et al., 2002a), Gibbs Recursive Sampler (Thompson et al., 2003) and BioProspector (Liu et al., 2001).

The assessment of different *de novo* motif discovery algorithms that only use coregulation information by Tompa *et al.* (Tompa et al., 2005) learned us that coregulation information can be sufficient to successfully discover regulatory motifs in yeast and prokaryotes, but not for motif discovery in higher organisms. Unlike prokaryotes, in which TF binding sites typically locate in promoter regions close to the TSS, TF binding sites in higher eukaryotes often locate in distal promoters or enhancers that can be located far from the TSS (§ 1.2.1). The longer distance between the TF binding sites and the TSS in higher eukaryotes, impose a greater computational challenge for motif discovery. Longer input sequences imply a larger search space and thus an increased risk of being trapped in local maxima by probabilistic methods. Therefore, the incorporation of auxiliary information, like evolutionary conservation, into such methods can be of significant benefit to discover motifs in higher organisms (see § 1.3.2 and § 1.3.3).

More recently, with the development of high-throughput experimental methods like ChIP (see § 1.2.3), *de novo* motif discovery evolves from a *gene-centered approach*, where the input consists of the non-coding sequences for a set of genes, to a *genome-wide approach*. Pipelines like MISCA (Boeva et al., 2010), CisGenome (Ji et al., 2008) and

W-ChIPMotifs (Jin et al., 2009), use standard probabilistic motif discovery tools like MEME (Bailey and Elkan, 1995) and the original Gibbs sampler (Lawrence et al., 1993) and even enumerative methods like Weeder (Pavesi et al., 2001) and MaMF (Hon and Jain, 2006) to perform *de novo* motif discovery on a subset of high-scoring DNA regions identified by ChIP assays. As most of those standard existing approaches can not computationally handle sets with thousands of candidate regulatory regions, they need to use an explicit significance threshold to retrieve only those regions with high TF binding probability. In contrast, algorithms specifically developed to perform motif discovery on ChIP-chip/seq datasets, like MatrixREDUCE (Foat et al., 2006) and cERMIT (Georgiev et al., 2010), use all the experimental data and their corresponding quantitative evidence (e.g. p-values of ChIP-chip experiments). Those approaches that make intelligent use of additional information like TF binding affinity consistently outperform the standard motif discovery tools. After mentioning this new tendency in motif discovery, we continue this introduction for the gene-centered motif discovery approaches, which are still in great demand as ChIP assays require substantial experimental efforts and are not yet commonly available for all TFs.

1.3.2 Algorithms based on comparative genomics

As a result of advances in DNA sequencing technologies, the number of closely related genomes being sequenced has increased tremendously. This has consequently led to the emergence of comparative studies focused on identifying functional elements in non-coding DNA sequences. Functional elements, including TF binding sites, are known to evolve at a slower rate than non-functional elements, and therefore well-conserved non-coding DNA sites should be good candidates for TF binding sites. This technique of delineating TF binding sites as conserved non-coding regions in the DNA is also called ‘phylogenetic footprinting’ (Duret and Bucher, 1997).

Many algorithms use evolutionary conservation information for *de novo* motif discovery, either as a pre- or post-processing step or by incorporating the conservation information into the motif finder itself. The former approach, where putative regions are filtered according to their conservation levels before applying conventional motif discovery (Harbison et al., 2004) or where predicted TF binding sites are post filtered by using conservation scores (Wasserman and Fickett, 1998), is quite straightforward. But this approach has the main drawback that any region with a conservation level below the chosen threshold is completely ignored, and thus TF binding sites that are not well conserved are not found by such methods. Thus, most conservation-based motif discovery algorithms use the latter approach, and incorporate the conservation information into the scoring function of the algorithm itself.

The first such algorithms that incorporated orthologous sequences like for example (Monsieurs et al., 2006; Marchal et al., 2004; Liu et al., 2004; Kellis et al., 2003; Wang and Stormo, 2003; Cliften et al., 2001; Gelfand et al., 2000), treated those orthologous sequences independently, thereby ignoring the underlying phylogeny that describes their relatedness. As a consequence those algorithms cannot distinguish between conserved DNA regions due to a short divergent time, from conserved DNA regions due to functionality.

The more advanced algorithms explicitly incorporate the relations between orthologous sequences by means of an evolutionary model, among many others PhyME (Sinha et al., 2004), OrthoMEME (Prakash et al., 2004), EMnEM (Moses et al., 2004a), Phylogibbs (Siddharthan et al., 2005) and Phylogenetic sampler (Newberg et al., 2007). Those algorithms require as input a predefined alignment of the orthologous regulatory regions and a phylogenetic tree defining the phylogenetic distances between the orthologous sequences. The main drawback of those algorithms that integrate conservation information by means of a predetermined ortholog alignment is that their performance strongly correlates with the quality of the alignment (Storms et al., 2010; Gordan et al., 2010; Ward and Bussemaker, 2008). How sensitive the algorithm is towards a bad-quality alignment also depends on how the algorithm intrinsically handles the alignment (Storms et al., 2010).

Since TF binding sites are usually short, sometimes degenerated, and often in reverse orientation or even relocated (Ludwig, 2002), alignment algorithms may not correctly align the binding sites within orthologous regulatory sequences. Especially when the sequences are very divergent, the background ‘noise’ of non-functional regions may be stronger than the ‘signal’ of conserved TF binding sites, preventing a correct alignment and often deteriorating motif discovery performance. Those observations inspired developers to create alignment-free approaches for using conservation information like for example the extension of the original Gibbs sampler by Li and Wong (Li and Wong, 2005) to find TF binding sites in multiple species independent of ortholog alignments by simultaneously sampling all orthologous and co-regulated sequences. In case of large input sets, this approach will become computational challenging. Another approach is the use of informative priors over DNA sequence positions based on a relaxed definition of evolutionary conservation: ‘a TF site within a regulatory region is considered to be conserved in an orthologous sequence if it occurs anywhere in that sequence, irrespective of orientation’(Gordan et al., 2010). Those priors can then be incorporated into an expectation maximization based approach like MEME (Bailey et al., 2010) or into a Gibbs sampling based algorithm like PRIORITY (Gordan et al., 2010).

All previous methods belong to the ‘multiple genes – multiple species’ category, which means that they were designed to search for motifs that are both overrepresented in a set

of coregulated genes from a reference species and conserved across related organisms. The combination of two information sources in motif discovery has shown large improvements compared with methods that only use one: transcriptional coregulation (see § 1.3.1) or evolutionary conservation (Blanchette and Tompa, 2003; Blanchette et al., 2002; Blanchette and Tompa, 2002).

1.3.3 Algorithms for combinatorial motif discovery

In eukaryotes, transcriptional regulation is often mediated by the concerted interaction of several TFs and cofactors (§ 1.2.1). The set of TF binding sites that attract interacting TFs often co-localize in the genome as modular structures of typically 50 bp to 1500 bp in size, forming a *cis*-regulatory module (CRM) (Jeziorska et al., 2009). As single TF binding sites are less likely to act as regulatory elements than TF binding sites occurring in clusters, co-localization can be used as an extra information source to improve motif discovery and forms the basis for the development of CRM discovery algorithms.

Such algorithms can be seen as extensions of ‘the standard *de novo* algorithms for single motif discovery’ to ‘algorithms for combinations of motifs’, by incorporating co-localization (Van Loo and Marynen, 2009). These methods are based on multiple-component motif models, where the singular motif models and their combinations are optimized simultaneously or iteratively. Joint modeling of TF binding sites in CRMs for a single species based on Gibbs sampling (Zhou and Wong, 2004) or expectation maximization (Segal and Sharan, 2005) demonstrated substantial improvement in *de novo* motif discovery.

Also successful was the combination of ‘co-localization’ and ‘comparative genomics’. This was done by the PRF-sampler (Grad et al., 2004) that first restricts the search space to regions conserved across different *Drosophila* species before searching for CRMs. Further improvement of CRM discovery performance was made when using evolutionary conservation in an alignment independent way, for example MultiModule (Zhou and Wong, 2007), EDGI (Sosinsky et al., 2007) and GibbsModule (Xie et al., 2008).

Previously mentioned methods search for CRMs without any prior information on the binding pattern of any relevant TF, which is often the case when the input consists of a set of genes identified in large-scale expression studies. But as the amount of TFs for which the regulatory motif is experimentally defined increases, discovery methods for CRMs that follow a slightly different approach were developed (Sun et al., 2009; Van Loo et al., 2008; Sharan et al., 2003; Aerts et al., 2003b). Those methods predict the set of regulatory motifs, responsible for the coregulation of the input genes, by using known motif matrix models from libraries and thus are expected to benefit greatly from novel technologies that construct these libraries (Van Loo and Marynen, 2009). Besides the

discovery of CRMs, many algorithms were developed to scan a set of sequences with a specific combination of known motif models, like Cluster-Buster (Frith et al., 2003) and Module-Scanner (Aerts et al., 2003a). These scanning approaches are currently the most advanced methods, although their applicability is limited to well-studied processes, for which the acting TFs and their motifs are known.

1.4 The role of chromatin in transcriptional regulation

DNA sequence information provides a basis for the prediction of TF binding sites, due to the sequence specificity of TF-binding events. However, DNA sequence alone is an impoverished source of information for the task of TF binding site prediction in eukaryotes as it always generates too many false positives. In the previous paragraph (§ 1.3) we already described some extra information sources that could increase motif discovery accuracy, namely the fact that TF binding sites tend to be more conserved than non-functional sites and binding sites of several TFs are often clustered together. Although those information sources showed promising improvements for the field of regulatory motif discovery, they do not allow distinguishing TF binding sites functional in one physiological condition or tissue from another.

In this paragraph we first introduce the structure of chromatin and how modification of chromatin structure influences eukaryotic transcriptional regulation. We distinguish two types of modifications that change chromatin structure: histone modifications (§ 1.4.1) and covalent DNA modifications like methylation (§ 1.4.2). As chromatin structure and its modifications can be inherited by the next generation, independent of the DNA sequence itself (Felsenfeld and Groudine, 2003), they are referred to as epigenetic traits. Due to the recent development of new experimental methodologies like ChIP (see § 1.2.3), an increased amount of experimental epigenetic data for several eukaryotic tissues and conditions becomes available. This inevitably creates a major computational challenge to incorporate those new data to improve the success rate of motif discovery in order to get novel insights into the mechanisms of gene regulation (§ 1.4.3).

1.4.1 Histone modifications

Chromatin is the complex of DNA and proteins in which the genetic material is packaged inside the cells of organisms with nuclei (Felsenfeld and Groudine, 2003). The nucleosome is the fundamental unit of chromatin and it is composed of an octamer of the four core histone proteins (H3, H4, H2A, H2B) around which 147 bp of DNA are wrapped. Histone modifications are post-translational modifications of the core histone proteins that constitute the nucleosome. The long and unstructured N-terminal tails by which histone proteins interact with neighboring nucleosomes are subject to various types

of covalent modifications, including lysine and arginine methylation, lysine acetylation and serine phosphorylation (Kouzarides, 2007). The use of modification-specific antibodies in ChIP-seq has revolutionized our ability to monitor the global incidence of histone modifications like acetylation and methylation in different cell lines for human (Wang et al., 2008; Barski et al., 2007) and mouse (Mikkelsen et al., 2007). Also ChIP-chip was used to map tri-methylation of lysine 4 of histone 3 (H3K4me3) (Guenther et al., 2007) or multiple histone acetylations and methylations (Koch et al., 2007) in different human cell lines.

There are two characterized mechanisms for the function of histone modifications in relation to transcriptional regulation:

- First they may affect higher-order chromatin structure by affecting the contact between different histones in adjacent nucleosomes or the interaction of histones with DNA (Hansen et al., 1998). Of all the known histone modifications, acetylation has the most potential to unfold chromatin since it neutralizes the basic charge of the lysine rich histone tails (Marks et al., 2001). In this way histone modifications control chromatin accessibility: either loosely packaged *euchromatin*, that allows access of the transcriptional machinery to the DNA and can be associated with transcriptional activation or highly compact *heterochromatin* associated with transcriptional repression (Sakabe and Nobrega, 2010; Kouzarides, 2007). Regions where local histone modifications displace the nucleosomes (nucleosome depleted regions) allow for easier digestion by DNase I. Two high-throughput methods: DNase-chip (Boyle et al., 2008; Xi et al., 2007; Crawford et al., 2006) and DNase-seq (Boyle et al., 2008), can be used to rapidly identify DNase I hypersensitive sites for any genomic region by either using tiled microarrays or sequencing. Mapping DNase I hypersensitive sites or open chromatin is an accurate method for identifying the location of active regulatory elements (promoters, enhancers, insulators, etc.).
- Secondly, histone modifications recruit non-histone proteins to the DNA, like enzymes that can further manipulate the chromatin structure or transcription regulatory protein complexes (Strahl and Allis, 2000). The pattern of histone modifications constitutes a ‘code’ that is read by the non-histone proteins and multi-protein complexes that form the transcription-activating and transcription-repressing molecular machinery (Strahl and Allis, 2000).

In mammalian systems the most established histone modifications that correlate with transcriptional *activation* are methylation of lysine 4 of histone 3 (H3K4me) and various histone acetylations in promoters and enhancers and those that correlate with *repression*

are trimethylation of lysine residues 27 (H3K27), 79 (H3K79) and 9 (H3K9) of histone 3 (Barski et al., 2007; Heintzman et al., 2007) see Figure 1.4.

As the histone modification patterns can differ for different classes of regulatory elements, they can be used to computationally predict new regulatory elements, like promoters and enhancers (Won et al., 2008). To illustrate: active human promoters and enhancers are both marked by nucleosome depletion, enrichment of histone acetylation and dimethylated H3K4 (H3K4me2), while monomethylated H3K4 (H3K4me1) is specific for enhancers regions allowing to distinguish between promoters and enhancers in the human genome (Heintzman et al., 2007). A more recent study of Heintzman *et al.*, also describes that enhancers, in contrast to promoters, are marked with highly cell-type-specific modification patterns (e.g. H3K4me1 is distributed in a cell-type specific manner) and thus enhancers strongly correlate to cell-type-specific gene expression programs (Heintzman et al., 2009).

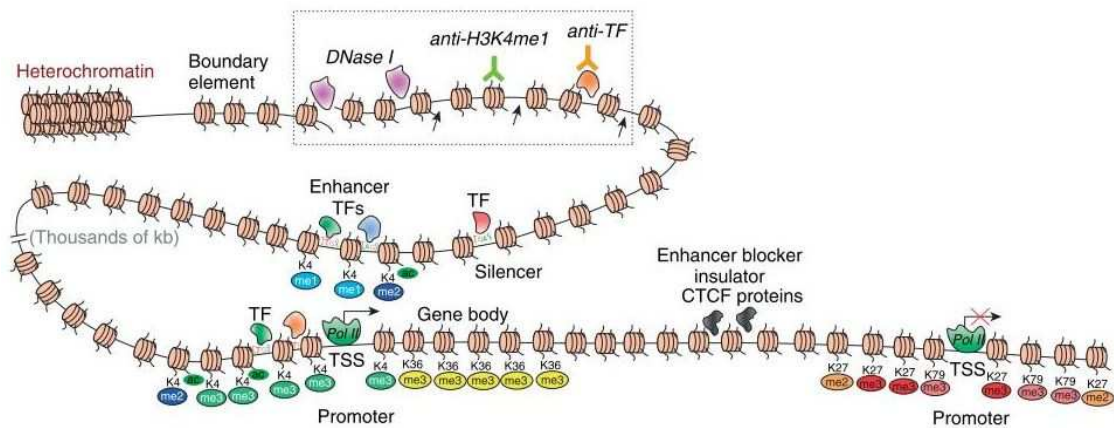


Figure 1.4 A schematic representation of transcription regulatory elements in the genome and two experimental identification methods. The promoter on the left (bottom) is activated by a distal enhancer that contains sequence-specific motifs to which TFs bind. The nearby silencer can control gene expression by competing with the enhancer or by invoking repressive chromatin through recruitment of histone deacetylases and methyltransferases. The enhancer blocker/insulator between the two promoters insures that only the gene on the left is transcribed. The boundary element (top left) prevents progression of heterochromatin on the euchromatic region. A few representative histone modifications are shown under histone tails: in green and blue hues, activating marks and in red and orange hues, repressive marks. The fraction of the genome that is covered by histone modifications is still unclear. Box: ChIP is represented by antibodies binding to a TF and to a nucleosome. The arrows represent DNA shearing that allows isolation of the bound sequences. DNase I digests accessible chromatin, represented by a discontinuity in the DNA. Taken from (Sakabe and Nobrega, 2010).

1.4.2 DNA methylation and CpG islands

Besides the histone proteins, also the DNA itself is subject to covalent chemical modification. DNA methylation (Weber and Schubeler, 2007) is the only epigenetic modification that directly affects the DNA. Biochemically, a hydrogen atom of the cytosine base is replaced by a methyl group. The gold standard for DNA methylation

mapping is bisulfite sequencing (as it achieves a single-bp resolution) that exploits the ability of bisulfite to convert the DNA methylation state into sequence-based information by conversion of unmethylated cytosines into uracils (Hajkova et al., 2002). Another method is methyl-DNA immunoprecipitation (MeDIP), a variant of ChIP-chip where purified DNA is immunoprecipitated with an antibody against methylated cytosines (Mohn et al., 2009). In mammals, DNA methylation is largely confined to cytosines in a CpG context ('CpG' stands for cytidine and guanosine, separated by a phosphate atom), which has two important implications. First, any genomic position that can be methylated is symmetric, i.e. there is a methylated or unmethylated cytosine on the forward strand as well as on the reverse strand. Therefore, after DNA replication a specific enzyme can read the DNA methylation pattern of the parent strand and faithfully copy it to the newly synthesized strand, thereby maintaining heritable DNA methylation patterns. Second, in mammalian genomes CpG dinucleotides occur in clusters, and the genomic regions with highest CpG density, termed CpG islands, usually (70-85%) exhibit very low levels of DNA methylation (Straussman et al., 2009). This because a methylated cytosine residue spontaneously deaminate to form a thymine residue; hence methylated CpG dinucleotides steadily mutate to TpG dinucleotides, which is evidenced by the underrepresentation of CpG dinucleotides in the human genome, except for the unmethylated CpG islands near promoter regions (Bock and Lengauer, 2008).

DNA methylation may affect transcriptional regulation in two ways:

- First, the methylation of DNA physically impedes the binding of transcriptional proteins to the DNA.
- Second DNA methylation fosters a locally more compact chromatin structure and hence represses transcription. Methylated CpG dinucleotide sites near a gene recruit specific DNA-binding proteins, which in turn recruit histone deacetylases and other chromatin remodeling proteins, resulting in inactive heterochromatin and silencing of gene expression (Felsenfeld and Groudine, 2003) (see Figure 1.5).

Unmethylated CpG islands are in contrast mediators of open chromatin structure and they frequently overlap with mammalian promoters (Antequera, 2003), enhancers and other regulatory elements (Bock and Lengauer, 2008). It is thus not surprising that unmethylated CpG islands are highly enriched for histone modification H3K4me (Straussman et al., 2009) and Ooi *et al.*, (Ooi et al., 2007) even suggest that the presence of H3K4me3 actually directed undermethylation by preventing the binding of the methylation complex.

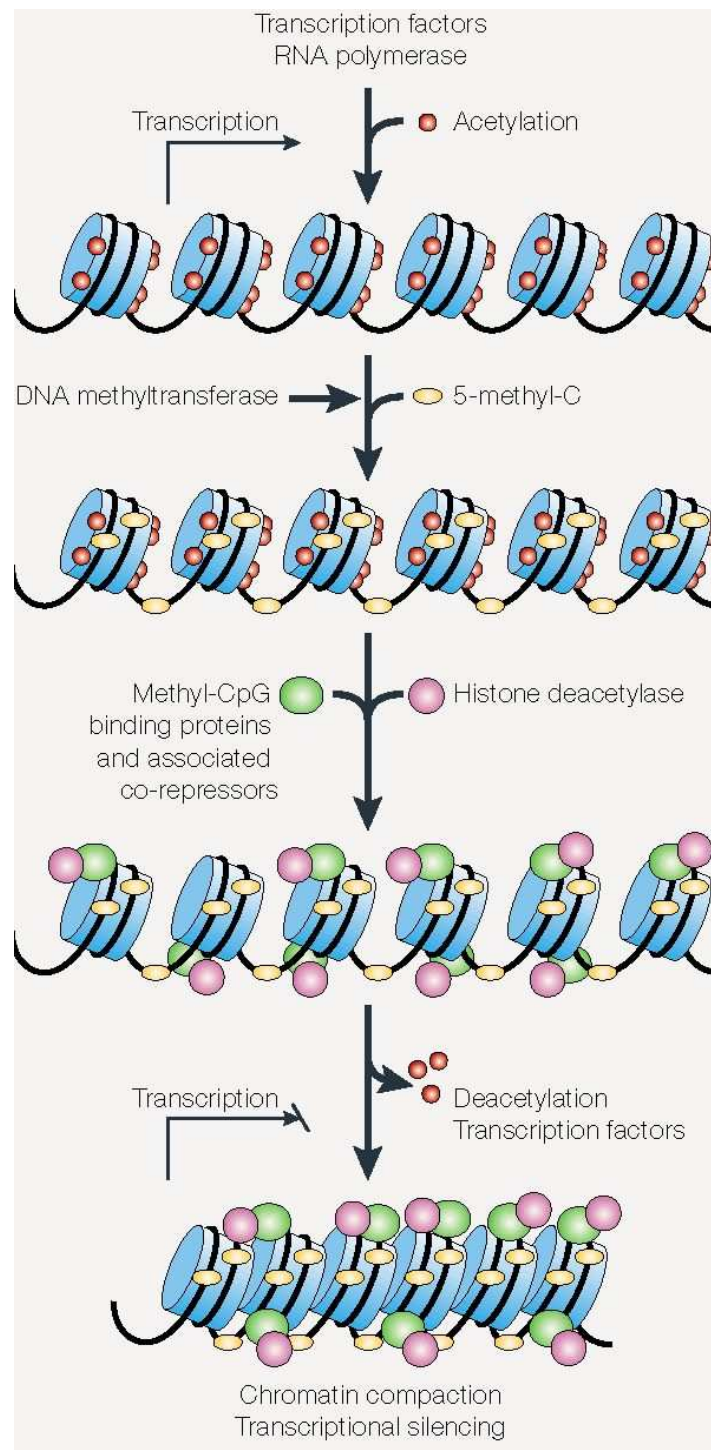


Figure 1.5 A transcriptionally active region targeted for silencing is proposed to acquire DNA methylation first, which then recruits the methyl-CpG binding proteins and their associated co-repressors and histone deacetylases (HDACs). As DNA methyltransferase 1 (DNMT1) can interact directly with histone deacetylase, it is also possible that transcription is first silenced by deacetylation by other tethering factors, after which the methylation machinery and the methyl-CpG binding proteins are recruited to 'cement' the promoter in the silent state. In either case, the deacetylated nucleosomes adopt a more tightly packed structure that inhibits the access of TFs to their binding sites. Adapted from (Robertson and Wolffe, 2000).

Computational prediction of DNA methylation is conceptually easier than the prediction of more volatile epigenetic mechanisms because DNA methylation patterns exhibit relatively low tissue specificity compared to other epigenetic information. For the prediction of methylated versus unmethylated CpG islands, the most predictive attributes included CpG-richness, specific DNA structure properties and repetitive DNA elements as well as certain TF binding sites (Straussman et al., 2009; Bock et al., 2006).

1.4.3 Epigenetic information in computational motif discovery

Rapid progress of experimental technologies has given rise to several initiatives like the ENCODE (Encyclopedia of DNA Elements) project (Birney et al., 2007) and the AHEAD (Alliance for Human Epigenomics and Disease) task force (Jones and Martienssen, 2005) to map functional elements and epigenetic traits. These projects are extremely important, not only in terms of applying and improving large-scale experimental methods, but also to make those data available for computational analysis and integration. This is particularly true for the ENCODE project (Birney et al., 2007), which has been designed from the onset as a close cooperation between experimental and computational biologists. The ENCODE project includes genome wide maps of DNase hypersensitive sites (DHS), DNA methylation, histone modifications and TF binding regions in various human cell lines. First only 1% of the human genome was targeted (Birney et al., 2007), and then was expanded to the entire human genome and genomes of model organisms (modENCODE) (Celniker et al., 2009). All the results of the ENCODE experiments are displayed on the UCSC Genome Browser (Thomas et al., 2007), which provides integrated visualization and standardized retrieval of various genome and epigenome datasets.

A few *de novo* motif discovery tools already integrate epigenetic information to gain performance accuracy. The use of epigenetic information can be integrated into the model or can be used in a discriminative way to reduce the search space in advance or to filter out retrieved binding sites that were not supported by the information. For now, most of the established *de novo* motif discovery approaches can only use this information in a discriminative way, except for BayesMD that uses a positional prior based on conservation and local sequence complexity (Tang et al., 2008), the new MEME (Bailey et al., 2010) and the PRIORIY algorithm (Gordan et al., 2010; Narlikar et al., 2006) that both make use of position-specific priors, based on for example epigenetic features. Position specific priors can easily be created based on different information sources using the PriorsEditor tool (Klepper and Drablos, 2010).

Compared to the limited use of epigenetic information in the *de novo* approach, motif scanning already describes many applications using this information source.

- Whittington *et al.*, (Whittington et al., 2009) showed that incorporating high-throughput histone modification data, such as H3K4me3 density, can greatly improve TF binding site prediction for a wide range of human and mouse TFs.
- Won *et al.* (Won et al., 2009) predicted CRMs in 1% of the human genome (ENCODE regions) for the HeLa cell line. Their strategy filters predictions based on their location relative to promoter and enhancer regions that were computationally predicted based on HeLa-specific histone modification data (Won et al., 2008).
- Won *et al.* (Won et al., 2010) developed ‘Chromia’ (CHROMatin based Integrated Approach) for genome-wide prediction of individual TF binding sites in mouse embryonic stem cells. This study differs from Won *et al.* (Won et al., 2009) as they used a genome-wide approach and fully integrated tissue-specific histone modification data instead of using it in a discriminative way.

All three studies (Won et al., 2010; Won et al., 2009; Whittington et al., 2009) used epigenetic data derived from the same tissue as for which they predicted TF binding sites. This corresponds with the observation that comparing five human tissues identified differences in the histone modification profiles, associated with transcriptional differences between the tissues (Koch et al., 2007).

The following two studies emphasize less the importance of using tissue specific epigenetic data for binding site prediction.

- Lahdesmaki *et al.* (Lahdesmaki et al., 2008) provided a probabilistic framework for integrating multiple data sources to predict TF binding per promoter region and in this way define genome-wide the target genes for a specific TF. Their framework can easily incorporate any information source that is indicative of TF binding. In this study they used computationally predicted nucleosome occupancy based on DNA sequence, which seemed too ‘static’ and thus not sufficiently informative to predict binding events.
- Ernst *et al.* (Ernst et al., 2010) used a large set of features to quantify binding preferences. The most informative features were based on histone modification levels (Barski et al., 2007) and DNase I hypersensitive locations (Boyle et al., 2008). The analysis in this paper showed that experimentally derived data in one tissue can be used to predict TF binding in another tissue.

1.5 Objectives of the thesis

In this section we will present, chapter-by-chapter, the objectives of this thesis. An overview of the relationships between the different chapters can be found in Figure 1.6.

Chapter 1 introduces transcriptional regulation and its main players, and summarizes the evolution of motif discovery with the focus on adding multiple information sources to improve its accuracy.

In **Chapter 2** the objective was the detailed study of two established motif discovery algorithms that integrate phylogeny. This was done by comparing their underlying models and algorithms. The choice for Phylogibbs (PG) (Siddharthan et al., 2005) and Phylogenetic sampler (PS) (Newberg et al., 2007) was based on their algorithmic similarity to the newly in-house developed algorithm: PhyloMotifSampler. PhyloMotifSampler is an extension of the MotifSampler algorithm which was developed by Gert Thijs (Thijs et al., 2002a). Similar to PG and PS, PhyloMotifSampler is a probabilistic algorithm that uses an evolutionary model to take into account the phylogenetic relatedness between orthologous sequences. Knowledge on the algorithmic background and the models used by two successful algorithms in the field was useful during the development of PhyloMotifSampler by our colleague Marleen Claeys.

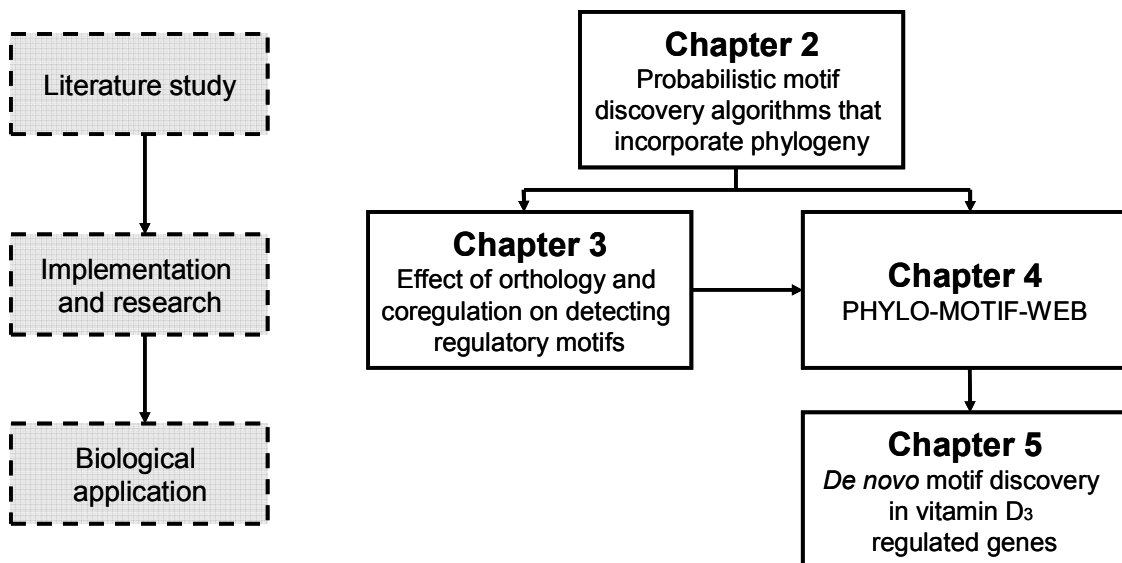


Figure 1.6 Overview of the relationships between the different chapters in this thesis.

In **Chapter 3** the first objective was to investigate the added value of using orthology information in combination with coregulation information in motif discovery. More specific we evaluated the conditions under which complementing coregulation with orthologous information improves motif discovery for the class of probabilistic motif discovery algorithms that incorporate phylogeny. We also investigated the effect of the type of data (e.g. the number of species, the evolutionary distances between the species

and the topology of the phylogenetic tree) on the performance of the motif discovery tools. Another objective of chapter 3 was to design datasets, both synthetic and real, covering different degrees of coregulation and orthologous information. In literature many benchmark datasets were described that only consist of coregulation information, in contrast, datasets suited to benchmark motif discovery tools that can integrate orthologous sequence data were very sparse.

In **Chapter 4** the goal was to develop an ‘ensemble strategy’ to comprise the results of multiple advanced motif discovery algorithms. Combining multiple algorithms may enhance motif discovery accuracy as was speculated by a number of studies (Tompa et al., 2005; Harbison et al., 2004). EMD (Hu et al., 2006) and SCOPE (Chakravarty et al., 2007; Carlson et al., 2007) are two *de novo* ensemble algorithms that combine multiple component algorithms to search for motifs in a set of coregulated genes. Both ensemble algorithms proved to be more successful than each of its component algorithms. We developed FuzzyClustering, a graph based clustering approach based on the work of Joshi *et al.* (Joshi et al., 2008) that instead of reporting the single best solution among all predictions, returns multiple local optima. As the posterior probability distribution is often multimodal, e.g. in case of weaker (~more degenerated) motifs, this ensemble solution will be more representative than the single best scoring solution.

Further on in Chapter 4, we wanted to make hard-to-use phylogenetic algorithms like PG and PS more accessible for non-expert users. To this end we developed a new workflow for motif discovery, PHYLO-MOTIF-WEB that applies an ensemble strategy on the results of multiple advanced motif discovery tools that can integrate phylogeny. This workflow provides all the necessary pre- and post-processing steps needed to identify *de novo* motifs in a biological dataset and is accessible through an easy to use web server. PHYLO-MOTIF-WEB also provides the option to use epigenetic information which is recommended when searching for motifs in eukaryotic species.

In **Chapter 5** the objective was to get more insight in the molecular mechanisms underlying the antiproliferative effects of vitamin D₃ on both human and mouse cell lines. In order to gain further insight in the molecular actions of vitamin D₃, we performed a comparative transcriptome analysis across human and mouse. Further, we would like to combine our *de novo* workflow, PHYLO-MOTIF-WEB, to identify ‘novel’ *cis*-regulatory elements, with the prediction of CRMs to unravel possible players in transcriptional regulation. This study was a combined effort of Fierro C., Storms V., Marchal K. and the Legendo (Laboratory for Experimental Medicine and Endocrinology) research group.

We end this thesis with a discussion of the main research and some perspectives for future research in **Chapter 6**.

Chapter 2 Probabilistic motif discovery algorithms that incorporate phylogeny

2.1 Introduction

With the growing number of sequenced genomes (Edwards et al., 2006; Venter et al., 2004; Breitbart et al., 2002), detecting motifs through ‘phylogenetic footprinting’ has become feasible. Several motif discovery algorithms have therefore integrated the use of orthology in addition to the frequently used coregulation information (Das and Dai, 2007). Most of the original motif discovery algorithms (Monsieurs et al., 2006; Marchal et al., 2004; Liu et al., 2004; Kellis et al., 2003; Wang and Stormo, 2003; Cliften et al., 2001; Gelfand et al., 2000; McGuire et al., 2000) could potentially incorporate orthologous sequences, but only by treating them independently and thus ignoring the underlying phylogeny that describes their relatedness. Because of this simplification, each orthologous sequence would contribute equally to the detected motif. This is counterintuitive as one would expect that a distantly related ortholog with a particular TF binding site contributes more information to the discovery of the motif than a more closely related ortholog with the same site conserved. On the other hand, the loss of a TF binding site in a distantly related ortholog should be penalized less than when this loss event occurs in a more closely related ortholog (Blanchette and Tompa, 2002). A number of more recent probabilistic motif discovery algorithms explicitly incorporate the relations between orthologous sequences by means of an evolutionary model, for example EMnEM (Moses et al., 2004a), OrthoMEME (Prakash et al., 2004), PhyME (Sinha et al., 2004), the method by Li and Wong (Li and Wong, 2005), Phylogibbs (Siddharthan et al., 2005), Tree Gibbs Sampler (Cai et al., 2007) and Phylogenetic sampler (Newberg et al., 2007).

In this chapter, we compare two well established, probabilistic motif discovery algorithms that explicitly incorporate phylogeny: Phylogibbs (PG) (Siddharthan et al., 2005) and Phylogenetic sampler (PS) (Newberg et al., 2007). We chose PG and PS as both algorithms can work on datasets including orthologs derived from more than two different species and datasets which contain only orthologous genes (i.e. no coregulation information). Another reason was the similarity between the algorithms underlying PG, PS and PhyloMotifSampler (a newly developed motif discovery algorithm by Marleen Claeys). A throughout study of PG and PS was thus very useful during the development and benchmarking of PhyloMotifSampler, an extension of MotifSampler (Thijs et al., 2002a) that can integrate phylogeny.

The theoretical comparison of their models (as described in § 2.2) and their algorithms (as described in § 2.3) also provides the necessary knowledge to correctly apply them on synthetic and real datasets, as is done further on in chapter 3 and chapter 5.

For developmental reasons, the weaknesses of both motif discovery algorithms are probably as interesting as their strengths, therefore we discuss the current limitations of this group of motif discovery algorithms in the discussion section (§ 2.4). We start this chapter with Table 2.1 that summarizes the most important characteristics of both algorithms.

Table 2.1 Summary of the most important characteristics of PG (Siddharthan and van Nimwegen, 2007; Siddharthan, 2007; van Nimwegen, 2007) and PS (Newberg et al., 2007; Thompson et al., 2007; Thompson et al., 2003).

Phylogibbs	Phylogenetic sampler
MODEL	
Input Sequences	
<p>-When used in the coregulation space, input sequences consist of intergenic regions of coregulated genes from one species. When used in the orthologous space, input sequences consist of orthologous intergenic regions that can optionally be prealigned. When used in the combined space, the input sequences consist of intergenic regions of coregulated genes complemented with the orthologs of these genes (that again can be optionally prealigned).</p> <p>-Both algorithms model the input sequences as generated by a <i>background model</i>. Some positions, deviating from the background model are assumed to be binding sites for TFs. These TF binding sites are modeled by a <i>motif model</i>, more specifically a position specific weight matrix (WM).</p>	
Motif model	
<p>-A WM of dimension (4xw) describes the probability of finding the respective nucleotides A, C, G, T at each position (from 1 to w) in the TF binding site. TF binding sites belonging to a specific TF are described by the same WM.</p> <p>-Both algorithms define TF binding sites differently for:</p> <p>1. Prealigned orthologous sequences</p>	
<p>-Assignment of a set of evolutionary related TF binding sites conserved across multiple species from the alignment. Terminology: a <i>window</i> (more specific a multi-species window).</p> <p>- Windows containing gaps can be split up into smaller windows without gaps and less species. PG can work on subparts of the alignment, allowing windows to be placed in conserved, well aligned regions as well as in unaligned regions. Therefore prealignments are made by a local alignment tool (Dialign was recommended) that annotates aligned and unaligned regions.</p>	<p>-Assignment of a set of evolutionary related TF binding sites conserved across all species from the alignment. Terminology: a <i>block</i>. Also used by PS is the term <i>MASS</i> defined as a set of aligned orthologs.</p> <p>-PS works on the alignment as a whole. Only sets of TF binding sites conserved over all species are taken into account, excluding all sets containing gaps. Prealignments for PS are based on a global alignment strategy and PS recommends ClustalW.</p>
2. Sequences from one species and unaligned orthologous sequences	
<p>Assignment of individual independent TF binding sites (a <i>window</i> now consists of one TF binding site, more specific a single-species window).</p>	<p>Assignment of individual independent TF binding sites (a <i>block</i> now consists of one TF binding site and a <i>MASS</i> is one sequence).</p>

PhyGibbs	Phylogenetic sampler
Background model	
Markov model with order n defined by the user. This model gives the probability of each nucleotide given the nucleotides on the n previous positions. Model parameters are estimated based on input sequences or based on an external file with intergenic sequences.	Position specific background model. This model gives the probability of each nucleotide on each position of the sequence and can be regarded as a zero-order Markov model. The model parameters are estimated based on the input sequences by a Bayesian segmentation algorithm (Liu and Lawrence, 1999).
Evolutionary model	
<p>-The model for evolution used by both algorithms is an <i>adapted F81 model</i> (Sinha et al., 2003):</p> <p>The adapted F81 model describes the probability $P_{ab}(t)$ that nucleotide a is mutated to nucleotide b over a time period t. The model assumes that all sequence positions evolve independently and at equal rates (γ) and the probability for fixation of a mutation at position i is proportional to the WM entry of that nucleotide at position i.</p> $P_{ab}(t) = \exp(-\gamma t) \delta_{ab} + (1 - \exp(-\gamma t)) WM_{b,i}$ <p>With γ = substitution rate, t =time, δ_{ab} is the Kronecker delta function that equals one for a=b and zero for a≠b and $WM_{b,i}$ is the WM entry of nucleotide b for position i (respectively the motif and background WM). Note that this model is an extension of the F81 model (Felsenstein, 1981) where fixation of a mutation is proportional to the frequency of that nucleotide in the data π_b (instead of $WM_{b,i}$).</p> <p>-The evolutionary relations between the species in the dataset are modeled by a phylogenetic tree with two properties:</p> <p>1. A set of phylogenetic distances: The phylogenetic distance between two species is modeled by:</p>	
The <i>proximity q</i> between two species = probability that no substitution took place per site. $q = \exp(-\gamma t)$ is used in the above evolutionary model.	The <i>branch length b</i> between two species = the expected number of mutations per site. $b = 3/4\gamma t$ is used in the above evolutionary model.
2. A topology (pattern of branching): Each branch connects a species/internal node with an internal node/ancestor.	
PG is only directly applicable to star topology trees where all species are directly descending from one common (unknown) ancestor.	PS is directly applicable to all tree topologies and thus allows for unknown internal nodes in the tree.
ALGORITHM	
-Goal: identify the positions of the TF binding sites hidden in the input sequences.	
-Method: explore the space of all possible solutions by MCMC (Liu, 2001) sampling.	
Sampling	
<p>-Collapsed Gibbs sampling: sample from a sequence of <i>posterior distributions</i> along a set of extensive moves.</p> <ol style="list-style-type: none"> 1. Start with a random positioning of windows, assigned to different TFs, also called a <i>configuration C</i>, based on prior information on the expected number of windows per TF in the data. 2. Construct the set of all possible configurations C' that differ in one single move from C. A move is e.g. changing the position of one single window or adding a new window. (see next page) 	<p>- Grouped Gibbs sampling: sample from a sequence of <i>conditional distributions</i> along a set of systematic moves.</p> <ol style="list-style-type: none"> 1. Start with a random positioning of blocks, assigned to different TFs, based on prior information on the expected number of blocks per TF in the data and maximum number of blocks per MASS. 2. Update the motif model based on all the current blocks (Model-update step). (see next page)

Phylogibbs	Phylogenetic sampler
Sampling (continued)	
<p>3. <i>Scoring</i>: calculate for each C' the posterior probability score.</p> <p>4. Sample a new configuration from this score distribution.</p> <p><u>This procedure (one cycle) is repeated for two phases:</u></p> <p>- 1) <i>simulated annealing</i> (Kirkpatrick et al., 1983) where one iterates to configuration C* with the highest posterior probability (=MAP). Instead of sampling from the normal score distribution a parameter β was introduced and sampling is done from a distribution which is proportional to (score)$^\beta$. By slowly increasing β the sampler will freeze into the global optimum C*.</p> <p>-2) <i>tracking</i> where posterior probabilities are assigned to the windows in C*.</p> <p>-> one initialization is sufficient -> short running time (minutes/hours)</p>	<p>3. Scoring: leave out the blocks for one MASS and calculate for each possible block in this MASS the conditional probability score.</p> <p>4. First sample the number of blocks for the MASS (recursive algorithm), then sample this number of blocks from the score distribution calculated in step 3 (Site-sampling step).</p> <p>5. Repeat steps 2 till 4 for each MASS in the dataset</p> <p><u>This procedure (one iteration) is repeated for two phases:</u></p> <p>-1) <i>burn-in iterations</i> to converge to an optimum. -2) <i>sampling iterations</i> to keep track of all sampled blocks to construct the solution afterwards.</p> <p>-> multiple initializations (seeds) recommended to avoid getting trapped in local optimum -> long running time (hours/days)</p>
Scoring	
<p>Score in above step 3</p> <p>-The <i>posterior probability</i> score of a configuration C is proportional with the probability that all windows in C are drawn from (unknown) motif WMs and that the background sequence is drawn from a known background model.</p> <p><u>-The motif WM is assumed to be unknown:</u> -to compute the probability that a window is drawn from an unknown motif WM, PG will use the conditional probability (this is the probability with a known motif WM) and scan this function over the entire WM space. Mathematically this resumes to solving an integral over all possible WMs, where the prior P(WM) is modeled by a Dirichlet prior distribution.</p> <p><u>For windows containing evolutionary related sites:</u> the scoring will include an <i>evolutionary model</i> and a phylogenetic tree to describe the probability that orthologous sites are related to a common ancestor site.</p>	<p>Score in above step 3:</p> <p>-The <i>conditional probability</i> of one block is proportional with the probability that the block is drawn from a known motif WM divided by the probability that the block is drawn from the known background model.</p> <p><u>-The motif WM is assumed to be known :</u> - update WM before score-computation (above step 2) :</p> <ul style="list-style-type: none"> • Sample a new motif WM from a Dirichlet distribution $Dir(\beta+c)$ where β = vector with pseudocounts for each nucleotide and c = vector with sequence weighted counts (*) for each nucleotide across all the blocks. • Accept the new motif WM with a probability proportional to the Metropolis Hastings ratio. This ratio is proportional with how good the new model explains the blocks versus the old model. <p>(*) Each orthologous sequence gets a weight based on the phylogenetic tree relating them by using the program Seq.weights.pl. For details on using sequence weights to build a motif WM see (Newberg et al., 2005).</p> <p><u>For blocks containing evolutionary related sites:</u> the scoring will include an <i>evolutionary model</i> and a phylogenetic tree to describe the probability that orthologous sites are related to a common ancestor site.</p>

Phylogibbs	Phylogenetic sampler
Scoring (continued)	
-For computational reasons, an approximation is needed to solve the integral. This approximation requires a star topology which makes it possible to directly obtain the joint probability of the evolutionary related nucleotides at the leaves of the tree. All other tree topologies are reduced to collections of star topologies.	-The <i>Felsenstein tree-likelihood algorithm</i> (Felsenstein, 1981) is used to handle all tree topologies. It is a recursive algorithm that marginalizes over all the interior nodes of the tree to obtain the joint probability of the nucleotides at the leaves of the tree.
Solution	
Maximum a posteriori (MAP) solution The output contains the configuration C* that has the highest posterior probability. It is an optimization based solution.	Ensemble centroid solution The centroid solution (also used in the ‘Gibbs Centroid sampler’ (Thompson et al., 2007)) is a collection of centroid TF binding sites composed by all the block positions that appear in at least half the sampling iterations over different initializations.
Posterior probabilities	
-During <i>the tracking phase</i> PG samples the distribution P(C S) of all configurations and compares each sampled configuration with the MAP configuration C* to assign posterior probabilities to all windows it reports. -The posterior probability of a window reports how strong this window is member of, or associated with C*. -Windows with a posterior probability higher than the chosen <i>tracking threshold</i> (T) are reported in the ‘track output’ file.	- A posterior probability is assigned to each centroid site based on the number of times the site or overlapping sites were sampled on the total number of iterations. -The <i>align-centroid</i> option aligns the centroid TF binding sites for a specific TF to construct a motif WM by using the ‘Gibbs recursive sampler’ (Thompson et al., 2003).

2.2 Models used by PG and PS

2.2.1 Input sequences

Both algorithms use as input a set of non-coding sequences, in which one expects a statistically overrepresented motif. The non-coding sequences are often the promoter sequences derived from a set of coregulated or orthologous genes or any combination thereof. Each input sequence (S_i) can be described as a background sequence modeled by a background model (β), interspersed with TF binding sites modeled by a motif model (θ) (see Figure 2.1).

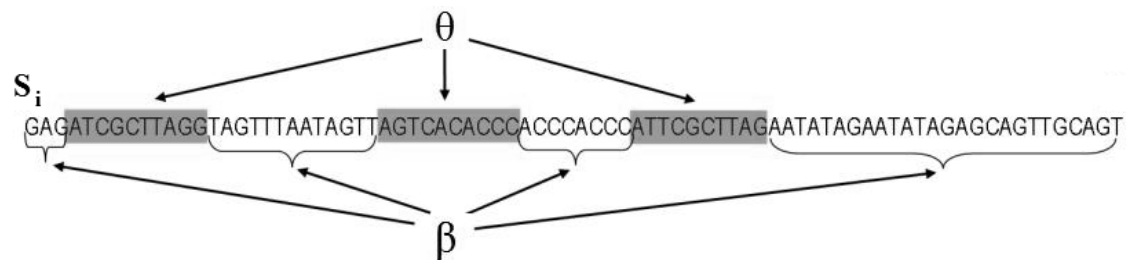


Figure 2.1 Input sequence (S_i) that consists of a set of TF binding sites (highlighted in gray) which are modeled by a motif model (θ), embedded in a background sequence which is modeled by the background model (β). Adapted from (van Nimwegen, 2007).

2.2.2 Motif model

The motif model (θ) is a position specific weight matrix (WM) of dimension ($4 \times w$) that describes the probability of finding the respective nucleotides A, C, G and T at each position (from 1 to w) in the TF binding site. TF binding sites that are bound by a common TF are described by the same motif model.

Both motif discovery algorithms start with a random assignment of TF binding sites to the input sequences, based on prior information on the number of different motifs and the number of TF binding sites per motif. Both algorithms can simultaneously search for multiple regulatory motifs. To keep the comparison more focused, the rest of this chapter will discuss both algorithms when searching for one regulatory motif. The random assignment of TF binding sites for one regulatory motif is straightforward in case of unaligned single sequences, but when the sequences are phylogenetically related, as is the case for orthologous non-coding sequences, they are first prealigned to delineate the regions that are orthologous. Then, TF binding sites are assigned to those sets of prealigned sequences. As both the alignment strategy and the positioning of TF binding sites differ for both tools, we explain them in more detail.

PS uses a global alignment strategy (ClustalW (Chenna et al., 2003)) while PG relies on a local one (Dialign (Morgenstern, 1999)). Being a global alignment strategy, ClustalW enforces the alignment to span the entire length of the sequences and by definition aligns all sequences. A ClustalW alignment thus contains the conserved regions in well aligned blocks, while the unconserved parts are usually located in ill-aligned, gapped regions. Because of these intrinsic properties, such global alignment strategies underperform when further related sequences and/or sequences with unequal length are included (Van Hellefont et al., 2005). The local alignment strategy Dialign identifies and aligns local regions of similarity within the sequences (aligned regions) and leaves the less conserved regions or sequences unaligned. Dialign thus explicitly annotates aligned and unaligned regions differently. Aligned regions can also cover a subset of the sequences only. It usually outperforms global strategies when unconserved sequences of unequal length are included.

The reason why both algorithms rely on different strategies to generate these prealignments stems from the difference in the way they use these prealignments. PS can only cope with global prealignments: from these global prealignments only regions that are gaplessly aligned over all species in the prealignment are considered as potential TF binding sites (also called blocks). Unaligned regions (corresponding to the gapped regions in a global prealignment) are ignored in the further analysis. For PG on the other hand, both the gaplessly aligned regions and unaligned regions are considered as potential TF binding sites, also called windows. For PG a window thus can contain a set of orthologous sites (multi-species window) as well as a single unaligned, independent site (single-species window). PG treats each of these subparts of the prealignment differently: for multi-species windows the phylogenetic relatedness between the sequences is taken into account by using an evolutionary model while the single-species windows are treated independently. PG thus benefits from using annotated, local prealignments as input. In Figure 2.2 the terms ‘window’ (used by PG) and ‘block’ (used by PS) are clarified.

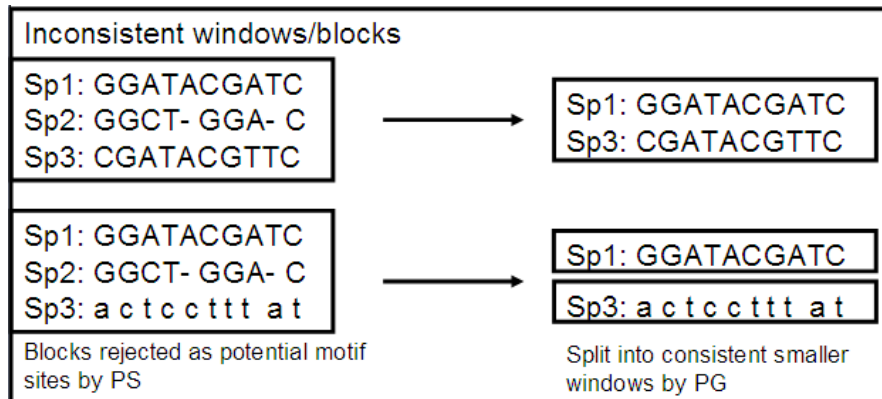


Figure 2.2 Explains the terms windows (used by PG) and blocks (used by PS). Both PG and PS have to assign sets of putative orthologous TF binding sites to a prealignment of orthologous sequences and these sets of putative orthologous TF binding sites are called windows (PG) or blocks (PS). When such a set of putative orthologous sites is not perfectly aligned over all sequences (contains gaps or unaligned parts) it is called an inconsistent window (PG) or an inconsistent block (PS). On the left of the figure, two examples of such an inconsistent window/block are shown: putative orthologous TF binding sites aligned over three sequences from respectively three species (Sp1, 2 and 3) that contain gaps and unaligned parts. PS does not consider inconsistent blocks for further analysis as they are rejected as potential TF binding sites. PG in contrast can split up inconsistent multi-species windows into smaller consistent windows that contain no gaps/unaligned parts by leaving out the ill aligned/unaligned sequences. These smaller consistent windows can still be multi-species windows (first example) that will be scored by an evolutionary model or they can become single-species windows (second example) that will be treated as independent TF binding sites.

2.2.3 Background model

The background sequence is modeled differently by both algorithms. PG uses an n^{th} -order Markov model that gives the probability of each nucleotide given the nucleotides on the n previous positions. The model parameters are estimated based on the input sequences or based on an external file provided by the user that contains non-coding sequences. PS does not assume homogeneity in the background composition of each input sequence. Therefore PS uses a Bayesian segmentation algorithm (Liu and Lawrence, 1999) to produce a position specific background model. This model gives the probability of each nucleotide on each position of the sequence and is estimated based on the input sequences.

2.2.4 Evolutionary model

Both algorithms account for the phylogenetic relatedness between orthologous sequences by scoring orthologous TF binding sites contained in blocks (for PS), or multi-species windows (for PG) by a tree-based evolutionary model. The model for the evolution of TF binding sites used by both algorithms is based on an *adapted* Felsenstein 1981 (F81) evolution model (Sinha et al., 2003) that describes the probability $P_{ab}(t)$ that nucleotide (a) is mutated to nucleotide (b) over a time period (t). The model assumes that all sequence positions evolve independently and at equal rates (γ) and the probability for fixation of a mutation at position i is proportional to the weight matrix (WM) entry of that nucleotide at position i .

$$P_{ab}(t) = \exp(-\gamma t)\delta_{ab} + (1 - \exp(-\gamma t))WM_{b,i} \quad (2.1)$$

With γ = the rate of nucleotide substitution (replacement by one of the four nucleotides) per unit time, t = time, δ_{ab} = the Kronecker delta function that equals 1 for $a = b$ and 0 for $a \neq b$ and $WM_{b,i}$ is the weight matrix entry of nucleotide b for position i (respectively the motif model θ or the background model β). Note that this model is an extension of the F81 model (Felsenstein, 1981) where fixation of a mutation is proportional to the frequency of that nucleotide in the data π_b (instead of $WM_{b,i}$).

The evolutionary relatedness between the orthologous sequences in the input set is modeled by a phylogenetic tree. A phylogenetic tree consists of nodes connected by branches. The input sequences are located at the external nodes of the tree. The internal nodes are the inferred ancestral sequences. The pattern of branching is called the ‘topology’ of the tree and the length of the branch connecting two nodes is a measure for the phylogenetic distance between them. PG and PS define the phylogenetic distance in a different way. PG introduced the ‘proximity’ (q) as distance measure, which corresponds to the probability that no substitution took place, per site.

$$q = \exp(-\gamma t) \quad (2.2)$$

Inserted in the evolutionary model we get:

$$P_{ab}(q) = q\delta_{ab} + (1 - q)WM_{b,i} \quad (2.3)$$

PS uses as distance measure the ‘branch length’ (b), which is defined as the expected number of mutations per site:

$$b = ut = \frac{3}{4}\gamma t \quad (2.4)$$

with u = the rate of nucleotide mutation (replacement by one of the three other nucleotides) per unit time and t = time. Mutation rate (u) thus differs from substitution rate (γ), as a ‘substitution’ of a nucleotide to itself is not counted and this is the source of the factor $\frac{3}{4}$. Inserted in the evolutionary model we get:

$$P_{ab}(b) = \exp(-\frac{4}{3}b)\delta_{ab} + (1 - \exp(-\frac{4}{3}b))WM_{b,i} \quad (2.5)$$

From those definitions we can deduce the following relation between ‘proximity’ used by PG and ‘branch length’ used by PS:

$$b = -\frac{3}{4}\ln(q) \quad (2.6)$$

Both algorithms also differ in the types of tree *topology* they can handle. PG is only directly applicable to star topology trees where all sequences are directly descending from one common (unknown) ancestor sequence. While PS is directly applicable to all tree topologies and thus allows for unknown internal nodes in the tree. This will be further explained in § 2.3.2.

2.3 The algorithms underlying PG and PS

2.3.1 Algorithms to sample the search space

The goal of both motif discovery algorithms is to detect the positions of the TF binding sites hidden in the set of input sequences. Because the space of all possible solutions is too large to search exhaustively, both tools use Markov Chain Monte Carlo (MCMC) sampling to efficiently explore the solution space (Liu, 2001). There are different MCMC based sampling methods, depending on how the transitions from one possible solution to another are defined; this is also called the ‘moveset’. PS uses one systematic move, typical for Gibbs samplers (explained beneath), while PG uses a specifically designed moveset with bigger moves. We illustrate some of the moves used by PG in Figure 2.4. These moves allow the PG algorithm to change between *configurations* (i.e. the positioning of the windows in the input sequences). For example, starting from one configuration, a new configuration can be obtained by shifting all the windows assigned to one TF, a few positions to the left (see panel 3 in Figure 2.4). Starting from a random configuration (C), PG calculates for all possible new configurations (defined by the

moveset) a posterior probability score and then samples a new configuration from the posterior probability score distribution. This iterative process of sampling new configurations based on their scores achieves convergence to the configuration with the highest posterior probability by simulated annealing. Simulated annealing implies that as the number of algorithm iterations increases, more and more weight is given to the configurations with highest posterior probability scores. Provided the annealing is done slowly enough, the final configuration will correspond to the globally optimal state.

The specifically designed moveset in combination with simulated annealing allows PG to search the whole solution space more efficiently, at the same time avoiding that the algorithm gets trapped in local optima and speeding up convergence. PG thus is an optimization based algorithm, designed to converge to the global optimum of the solution space during one run of the algorithm, which explains its relative short running time.

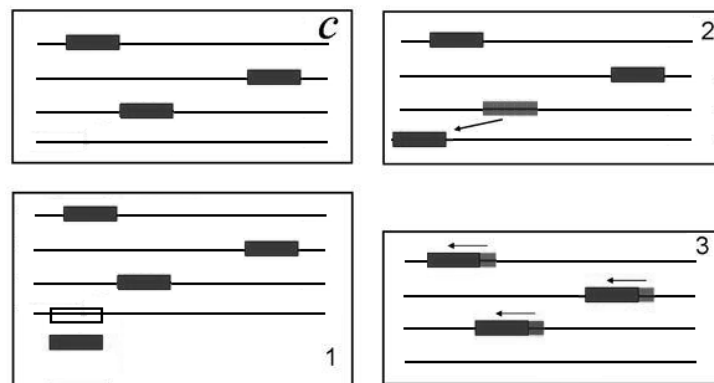


Figure 2.4 Moveset for PG. Left top corner: PG starts with a random positioning of windows (grey squares), also called a configuration C, based on prior information on the expected number of windows in the data. Then the algorithm constructs the set of all possible configurations C' that differ in one single move from C. Panel 1, 2 and 3 illustrate possible moves and the resulting configurations. Panel 1, the move equals adding an extra window. In panel 2, the move equals repositioning one window. In panel 3, the move equals shifting the positions of all the windows to the left. Adapted from (van Nimwegen, 2007).

The algorithm of PS also starts with the random assignment of blocks (a single TF binding site in case of unrelated sequences or a set of orthologous sites in case of orthologous sequences) to the input sequences based on prior information given by the user e.g. on the number of blocks per sequence (see Figure 2.5). Then, PS uses Gibbs sampling, a two step iterative procedure that consists of a 'site-sampling' and a 'model-update' step. During the 'site-sampling' step, the blocks in one MASS (a single sequence or one set of aligned orthologous sequences) are replaced by new blocks, by sampling from a conditional score distribution for all possible blocks in the MASS. As mentioned on top, the moves of PS, to shift between possible solutions are thus more restricted as those of PG, as all the other blocks remain fixed, except for the blocks in one MASS (see Figure 2.5). In the 'model-update' step, the motif model (θ) is updated based on all the current blocks. The two step iterative procedure is repeated during two phases, first a

number of *burn-in iterations* to converge to an optimum, followed by *sampling iterations* to keep track of all sampled blocks to construct an ensemble or centroid solution afterwards. As PS is more prone to converge in each run to a different local optimum it needs to be re-initialized multiple times in order to estimate the global solution. Because of the re-initializations of the algorithm, running times of PS are considerably larger than those of PG.

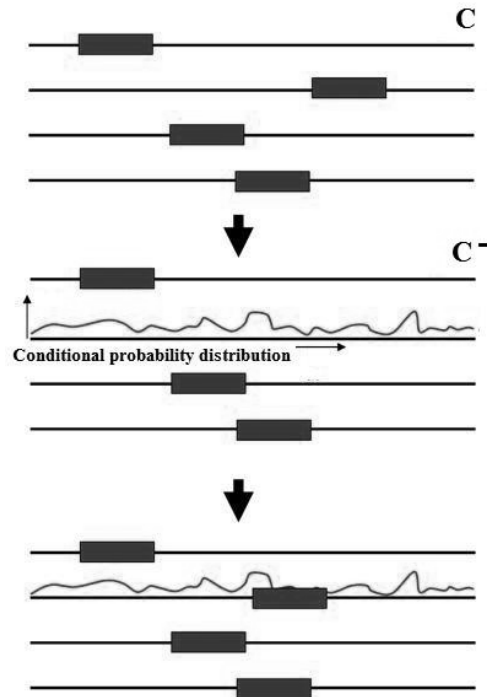


Figure 2.5 Moveset for PS. On top: PS starts with a random positioning of blocks (gray) for a specific TF, (i.e. configuration C), in the input set. This based on prior information on the expected number of blocks for this TF in the input set and the maximum number of blocks per MASS (single sequence or set of aligned orthologous sequences). In a following step of the algorithm, the blocks are left out for one MASS and the conditional probability score is calculated for each possible block in this MASS, conditional on the current values of the other blocks (C^-). Based on the conditional score distribution a number of new blocks are sampled and the motif model can be updated. Adapted from (van Nimwegen, 2007).

2.3.2 Scoring methods

PG calculates a *Bayesian posterior probability* $P(C | S)$ for every possible configuration (C) given the input set (S) by using the Bayes' theorem:

$$P(C | S) = \frac{P(S | C)P(C)}{\sum_{C'} P(S | C')P(C')} \quad (2.7)$$

where $P(C)$ is the prior probability of configuration (C) and $P(S | C)$ the likelihood of the input set (S) given the configuration (C). The denominator sums over all configurations and thus evaluates the prior probability of S , which can usually be treated as a constant

and ignored. The great advantage of a Bayesian approach is that prior information can easily be incorporated into the scoring scheme. In practice a uniform prior $P(C)$ that assigns equal probability to all configurations is too ignorant, that's why the user has to give some prior information regarding the number of windows and the number of different motifs and all other configurations get a probability of zero.

In case all configurations would be equally likely *a priori*, the posterior probability $P(C|S)$ is proportional to the likelihood $P(S|C)$, which equals the probability that all windows in C are drawn from the (unknown) motif model (θ) and that the background sequence is drawn from a known background model (β):

$$P(S|C, \theta, \beta) = P(S = C | \theta)P(S \neq C | \beta) \quad (2.8)$$

where $S = C$ refers to the windows (TF binding sites) and $S \neq C$ refers to the background sequence. To compute the probability that a window is drawn from an unknown motif WM, PG will use the conditional probability (this is the probability with a known motif WM) and scan this function over the entire motif WM space. Mathematically this resumes to solving an integral over all possible motif WMs, where the prior $P(WM)$ is modeled by a Dirichlet prior distribution $\text{Dir}(\gamma)$.

$$\int P(\text{Window} | WM)P(WM)\partial WM \quad (2.9)$$

$$P(WM) = \prod_{i=1}^m \prod_{\alpha} (WM_{i\alpha})^{\gamma-1} \quad (2.10)$$

Where m is the width of the motif WM and $WM_{i\alpha}$ is the probability of finding the nucleotide α at position i according to the motif WM. Parameter γ is generally referred to as a pseudocount that can be set by the user. Default γ is set to 1, so that a uniform prior is obtained, making all motif WMs *a priori* equally likely, which can be argued to reflect a state of complete ignorance about the motif WM. In reality, however, we know that for most positions in the TF binding site, regulatory factors tend to have distinct preferences for certain nucleotides. By setting $\gamma < 1$, more weight will be put on motif WM columns that are 'skewed', i.e. giving low probability to some nucleotides and high probabilities to others.

For single-species windows this integration (equation 2.9) can be done exactly. For multi-species windows, containing evolutionary related binding sites, the scoring will include an evolutionary model and a phylogenetic tree (T) to describe the probability that orthologous sites in a multi-species window evolved from a common ancestor site under the selective pressure that they remain binding sites for the same TF.

$$\int P(\text{multi-speciesWindow} | T, WM)P(WM)\partial WM \quad (2.11)$$

In principle we can analytically determine the value of this integral, but for computational reasons an approximation is used, as the complexity of the integral increases exponentially both with the number of orthologs and the number of windows. This approximation requires a star-like tree topology which makes it possible to directly obtain the joint probability of the evolutionary related nucleotides at the leaves of the tree. All other tree topologies are reduced to collections of star topologies.

PS calculates the *conditional probability* for each possible block in the MASS during the ‘site-sampling’ step of the algorithm. The conditional probability of a block is proportional with the probability that the block is drawn from a known motif model divided by the probability that the block is drawn from the known background model (β).

$$P(\text{block} | C) \approx \frac{P(\text{block} | \theta)}{P(\text{block} | \beta)} \quad (2.12)$$

With C = the configuration of blocks in the input set and θ = the motif model based on the blocks in configuration C . For blocks containing evolutionary related sites, the scoring will include an evolutionary model and a phylogenetic tree (T) to describe the probability that orthologous sites in a block evolved from a common ancestor site under the selective pressure that they remain binding sites for the same TF.

$$P(\text{block} | C, T) \approx \frac{P(\text{block} | \theta, T)}{P(\text{block} | \beta, T)} \quad (2.13)$$

To calculate $P(\text{block} | \theta, T)$ no integration is needed as the motif model is inferred during the ‘model-update’ step (see next section) and the Felsenstein tree-likelihood algorithm (Felsenstein, 1981) is used to handle all tree topologies. It is a recursive algorithm that marginalizes over all the interior nodes of the tree to obtain the joint probability of the nucleotides at the leaves of the tree. This means that PS works with a full phylogenetic model and does not need to make approximations regarding the topology.

In contrast to PG, PS accounts for the phylogeny at an additional level. During the ‘model-update’ step, a new motif model is sampled from a Dirichlet distribution $\text{Dir}(b+c)$ where b equals a vector with pseudocounts for each nucleotide and c equals a vector with sequence weighted counts for each nucleotide across all the current blocks. The pseudocount vector b compensates for zero occurrences in the nucleotide counts. To construct the vector c , the nucleotide counts for a fixed position over all blocks are weighted according to the phylogenetic weight of the sequences in which they were counted. All orthologous sequences have assigned a weight based upon the phylogenetic tree relating them (Newberg et al., 2005). This weighting scheme enables PS to give TF binding sites, conserved in distant orthologs, a higher weight/contribution to the motif

model, than sites conserved in close orthologs. After sampling, PS will accept this new motif model with a probability proportional to the Metropolis Hastings ratio. This ratio is proportional with how good the new model explains the blocks versus the old model.

2.3.3 Solutions and posterior probabilities

PG reports a *MAP (maximum a posteriori) solution*, meaning the configuration of windows with the maximum posterior probability. This MAP configuration is obtained during the simulated annealing phase of the algorithm (see § 2.3.1). This phase is followed by a ‘tracking’ phase, defined as the prolonged sampling after convergence, and measuring, for each possible window, how often it is co-clustered with one of the optimal MAP windows. Tracking provides a significance assessment for each window, expressing how strong this window is member of, or associated with the MAP configuration. Only windows with a posterior probability higher than a chosen tracking threshold are reported in the final solution. Tracking makes PG more robust against prior over- or under-estimations of the number of windows that occur in the data. Superfluous windows found during the simulated annealing phase, will not be stably associated with the MAP configuration during tracking and will be lost, while sites not found during simulated annealing can be picked up. The drawback is that the algorithm may miss a group of windows that are grouped a significant fraction of the samples, but that do not occur in the MAP configuration.

PS does not report the single best scoring solution, but garners information from the full ensemble of solutions. PS’s *ensemble centroid solution* (Thompson et al., 2007) represents all frequently visited solutions, during the iterations after convergence, across all re-initializations, also called the sampling iterations. To calculate the ensemble centroid solution, PS counts the occurrences of individual blocks during the sampling iterations. The blocks which occur in at least half the sampling iterations are defined as centroid blocks. The use of a fragmentation algorithm (Liu et al., 1995) to infer the widths of the blocks, causes variation in the lengths of the centroid blocks, making it difficult to see which positions are more highly conserved. As a solution, PS makes an alignment of them to construct a motif WM by using the ‘Gibbs recursive sampler’ (Thompson et al., 2003). In general, ensemble centroid solutions provide more accurate estimates compared to MAP solutions, which focus exclusively on the single most probable solution.

2.4 Discussion

Gibbs sampling was first used in the context of biological motif discovery by Lawrence *et al.* (Lawrence *et al.*, 1993). Since then, a significant number of Gibbs sampling based algorithms have been developed like (Thompson *et al.*, 2003; Thijs *et al.*, 2002a; Liu *et al.*, 2001) amongst many more. In this chapter we compared two of the more advanced Gibbs sampling based algorithms, Phylogibbs (PG) and Phylogenetic sampler (PS). They both integrate phylogeny by using a tree-based evolutionary model, they can sample simultaneously for multiple motifs and they assess the significance of their predictions by means of ‘tracking’ or constructing an ensemble centroid solution. Those improvements on motif discovery were explained in § 2.2 and § 2.3. In this section we briefly mention some of their limitations.

Common to all MCMC sampling based algorithms, like Gibbs samplers, they require long convergence times because they need to sample the search space for a long time. Especially for PS this results in a very long runtime, as it needs to be re-initialized multiple times to obtain an accurate ensemble solution. The convergence time increases with the size of the input set, the total number of TF binding sites and the number of motifs. Speed improvements were already proposed for the PG algorithm, by using a form of ‘importance sampling’ on top of the moveset used to sample the search space. When sampling a replacement window, PG now considers every available window and needs to calculate a posterior probability score for each possible new configuration. Using importance sampling would guide the sampler to spend most of its time in important parts of search space and only to select ‘important’ windows lowering the number of configurations to be scored (Siddharthan, 2008).

A limitation, more specific for algorithms incorporating phylogeny, is the quality of the ortholog alignment. It was shown (Storms *et al.*, 2010; Gordan *et al.*, 2010; Ward and Bussemaker, 2008) that errors in the ortholog alignment can have very deleterious effects on the performance of algorithms such as PG and PS. In chapter 3, we show that PG is more robust against bad quality alignments compared to PS. The local alignment strategy used by PG, in combination with the window-principle, allows PG to retrieve TF binding sites which are only partially conserved, meaning across a subset of aligned orthologous sequences. As a result, PG affords some flexibility in terms of the evolutionary distances spanned by the input sequences. For instance, the use of a distantly related ortholog will help pinpoint TF binding sites located in conserved regions but will not hamper the discovery of TF binding sites absent from that ortholog.

Other motif discovery algorithms try to overcome this limitation, by skipping the pre-alignment of orthologous sequences and searching both the space of multiple alignments and the space of possible TF binding sites at the same time (Cai *et al.*, 2007; Li and

Wong, 2005). This will create a very large search space, difficult to search effectively. Moreover, for closely related species, large segments of the orthologous non-coding sequences can be unambiguously aligned, and by pre-aligning these we significantly reduce the search space for the algorithm. A very recent Gibbs sampler, PRIORITY (Gordan et al., 2010), avoids the use of pre-alignments and does not take into account phylogenetic relationships between the orthologous sequences. To accomplish that, PRIORITY considers a TF binding site conserved in an orthologous sequence, if it occurs anywhere in that sequence, irrespective of its orientation. Based on this definition of conservation, informative priors over the input sequences are derived, which are incorporated into their Gibbs sampling algorithm.

Next, we discuss another common phenomenon that influences motif discovery; the absence of TF binding sites in the orthologous sequences. This can be the result of functional turnover of the orthologous sequence, which can be caused by TF binding site turnover, which is a common event during genome evolution, and plays a major role in shaping the regulatory circuitry of contemporary species (Wray, 2007). As TF binding sites are short and degenerate sequence patterns, they exhibit frequent turnover, even across phylogenetically closely related species, such as various *Drosophila* species (Moses et al., 2006). But in both motif discovery algorithms, no explicit functional turnover model has been used to infer TF binding turnover events. They only model sequence evolution at the nucleotide level, and lack the ability to capture the evolutionary dynamics of TF binding site turnover. The CSMET algorithm (Ray et al., 2008), explicitly models TF binding site turnover across species, through a ‘low resolution’ phylogeny, defined by a functional (motif or background) substitution process. Nucleotide substitution is modeled function-specific through ‘high resolution’ phylogenies. This makes that CSMET captures function-specific sequence evolution in every orthologous sequence rather than assuming that all aligned sequences evolved according to the same model (motif or background) (see Figure 2.6).

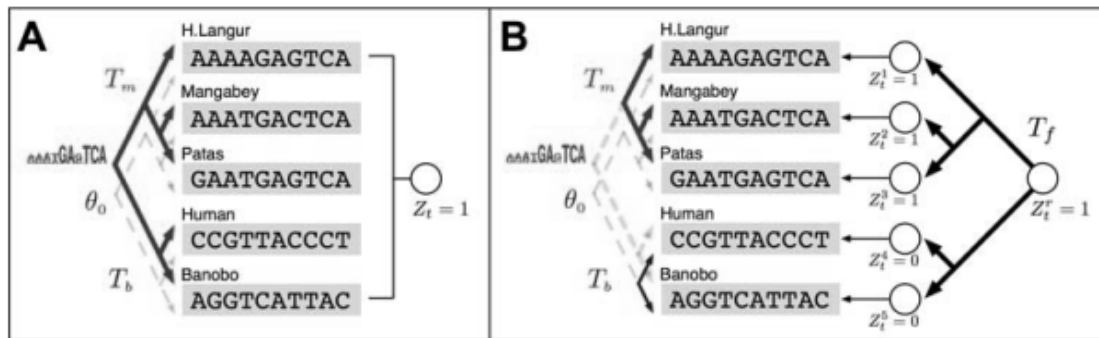


Figure 2.6 Modeling of TF binding site turnover by CSMET. (A) Algorithms like PG and PS do not incorporate an explicit model for function turnover. This means that the functionality indicator Z_t must apply to all the sequence in the alignment, in this case, all sequences evolve under a full motif phylogeny. Z_t specifies the functional state that can either be motif ($Z_t=1$) or background ($Z_t=0$). (B) CSMET, incorporates an explicit evolutionary model T_f for species-specific functional turnover, and partial motif or background phylogenies over subsets of sequences according to the turnover status. Adapted from (Ray et al., 2008).

PG and PS only model evolution at the nucleotide level, by means of a nucleotide substitution model that is based on the F81 evolution model (Felsenstein, 1981). This model is very limited as it assumes that all sequence positions evolve independently and at equal rates. Moses *at al.* (Moses et al., 2004b) shows that motif finding benefits from using more realistic evolutionary models such as HKY85 (Hasegawa et al., 1985), which captures the transition/transversion bias, or the model from Halpern and Bruno (HB) (Halpern and Bruno, 1998), which allows for position-specific variation in evolutionary rates.

To even further improve motif discovery, algorithms like PG and PS could integrate epigenetic information like open chromatin or histone modification data, both relating to functional active DNA. PG, that uses a Bayesian approach, which allows for easy integration of prior information, can for example favor configurations for which the TF binding sites lie in functional active DNA. PS can use a position-specific prior across the input sequences to guide the search to those regions that are functionally active.

Chapter 3 The effect of orthology and coregulation on detecting regulatory motifs

3.1 Introduction

The growing number of sequenced genomes allows integrating orthology evidence with coregulation information when searching for regulatory motifs. Moreover, the more advanced motif discovery algorithms explicitly model the phylogenetic relatedness between the orthologous input sequences and thus should be well adapted towards using orthologous information. So far no independent study has evaluated the extent of information contained within either the coregulation or the orthologous space and the conditions under which complementing both spaces improves motif discovery. In this chapter we performed such analysis by applying two of the more advanced motif discovery methods on both synthetic and real datasets with different properties. We chose for ‘Phylogibbs’ (PG) (Siddharthan et al., 2005) and ‘Phylogenetic sampler’ (PS) (Newberg et al., 2007) as both algorithms are specifically designed to integrate coregulation with orthology (therefore referred to as phylogenetic motif discovery algorithms in this study), neither of them is limited in the number of species that can be included and previous studies (Siddharthan et al., 2005) already described the superiority of PG in detecting motifs. As a comparison we included MEME (Bailey and Elkan, 1994) as a representative of algorithms that cannot explicitly incorporate phylogenetic relations (therefore referred to as a non-phylogenetic motif discovery algorithm).

3.2 Materials and Methods

3.2.1 Motif discovery algorithms and parameter settings

Three motif discovery algorithms were used: MEME (Bailey and Elkan, 1994), Phylogibbs (PG) (Siddharthan et al., 2005), and Phylogenetic sampler (PS) (Newberg et al., 2007). MEME is a probabilistic motif discovery tool that follows an optimization strategy based on Expectation Maximization (EM). As it was originally developed for detecting motifs in the coregulation space, it treats all input sequences independently, and does not explicitly model the phylogenetic relatedness between the input sequences. MEME searches for the most statistically overrepresented motif in the dataset (the one with the lowest E-value). Each TF binding site is reported with a p-value. We used MEME-4.00 with default parameters, we set the distribution of motifs to “anr (any number of repetitions)” and the maximum number of EM iterations to 500. We searched for a palindromic motif (-pal) in case of TyrR and LexA for the real data (see Text S3 in the Supplementary Materials).

In contrast, PG and PS, both developed for detecting motifs in the combined coregulation-orthology space do include a model to take into account the phylogenetic relatedness between orthologous input sequences. In chapter 2, we outlined the most important differences between both algorithms making it possible to view the results of this study in the light of these algorithmic characteristics. For PG we used Phylogibbs-1.0 and for PS we used Gibbs.x86_64. Before performing the tests on the synthetic and real datasets, we thoroughly tested the sensitivity of both algorithms towards parameter settings, not of primary importance for our main discussion, but that influence the results if not optimized. These tests and the optimized settings as applied in our analysis are summarized in Text S3. Most settings were not varied throughout the test runs except for the tracking threshold of PG that was set more stringent than its default value, unless indicated otherwise.

For PG, prealignments were made with Dialign (Morgenstern, 1999) (with the parameter $T=2$ to avoid long unaligned regions obtained with higher values of T). For PS, prealignments were obtained with ClustalW (version 1.83 (Chenna et al., 2003)) as suggested by the developers. For the difficult to align datasets we also performed tests with PS on prealignments obtained with Dialign (results in Text S2). For those tests the results were similar or worse than those obtained with prealignments from ClustalW, indicating that the observed differences between PS and PG are caused by the intrinsically different way they cope with the prealignments rather than to small differences in the used prealignments. In general, difficult to align sequences will be left unaligned with Dialign. This improves the alignment, but implies that those regions can no longer be used by PS (see also below). Therefore, for PS it is often more advantageous to use ClustalW instead of Dialign (which we therefore did in the remainder of the analysis).

3.2.2 Synthetic datasets

We created two synthetic motif weight matrices (WMs) as described previously (Siddharthan et al., 2005), both of width 13 bp, one with a high information content (IC) and one with a lower IC. TF binding sites sampled from these WMs were embedded at a randomly chosen position in a random background sequence of length 500 bp. Each ancestral sequence (~a background sequence containing an embedded TF binding site) was then evolved along a phylogenetic tree under a defined evolutionary model to create phylogenetically related sequences. For the background sequence we used the Jukes and Cantor (JC) model (Jukes and Cantor, 1969), for the embedded TF binding sites an adapted Felsenstein (F81) model (Sinha et al., 2003). Details on the construction of the WMs and the evolutionary related sequences are in Text S1 (Supplementary Materials).

For the experimental setup we simulated datasets for the coregulation space, the orthologous space and the combined coregulation-orthology space. For the coregulation space, we simulated the intergenic sequences of ten genes in a reference species (the species exhibiting a proximity of 0.80 to the ancestral species was considered the reference species). In each of these 10 sequences a TF binding site, drawn from a common motif WM, was embedded. For the combined space, we extended the coregulation space by simulating the orthologous intergenic sequences for each of the ten coregulated reference genes, according to a phylogenetic tree that describes the relatedness of the orthologous sequences to the ancestral sequence. The topology of the phylogenetic tree was varied between a star topology (equal or unequal distances) and a tree topology with internal nodes. The orthologous space consisted of the intergenics of a single reference gene together with its simulated orthologs. For all trees used in our tests, the Newick format is given in Table S4.

3.2.3 Real datasets

The real datasets are derived from Gamma-proteobacterial and *Saccharomyces* intergenic sequences. Also here, datasets were obtained with either a high IC or a low IC motif. For the coregulation space we selected target genes in *Escherichia coli* for the regulators LexA and TyrR and in *Saccharomyces cerevisiae* for the regulators URS1H and RAP1. To extend these datasets in the combined space, we searched for all target genes their corresponding orthologs in respectively other Gamma-proteobacterial or *Saccharomyces* species. The real datasets for the orthologous space only consist each time of one single target of the regulator in the reference species and its corresponding orthologs. In this case we selected as reference target, a gene that contained exactly one copy of the TF binding site in its upstream region, in order not to confound coregulation with orthologous information (as the presence of multiple copies confers coregulation information). For the real data, we defined the upstream region as the intergenic region between the start codon of the gene and, depending on its orientation the start or stop of the previous coding gene. Details on the construction of the real datasets are in Text S1 and Table S2, the phylogenetic trees that relate the intergenic sequences of respectively the bacterial and yeast species are depicted in Figure S1. The Newick formats of the trees are given in Table S4.

3.2.4 Performance and quality measures

- *Predicted TF binding sites*: TF binding sites predicted by MEME correspond to all sites obtained from the Expectation Maximization based solution. For PG, the ‘predicted TF binding sites’ are the TF binding sites from the tracked maximum *a posteriori* solution. For PS we defined the ‘predicted TF binding sites’ as the sites returned after running the ‘align-centroid’ option on the collection of centroid TF binding sites. More information on the output of PG and PS can be found in chapter 2 (§ 2.3.3 ‘Solutions and posterior probabilities’). A ‘predicted motif model’ is the WM constructed from the predicted TF binding sites for a specific transcription factor.
- *Number of datasets/runs with an output (D1/R1)*: For the synthetic data we had 100 input datasets per test. D1 gives the number of datasets for which the algorithm returned an output, irrespective of whether this output is correct or not. For the real data we only had one input dataset per test, so here we re-ran the algorithm ten times to get ten outputs for one input dataset. R1 represents the number of runs for which the algorithm returns an output. PG and PS internally evaluate their results and only report for each run or dataset the solutions that exceed a certain threshold. As a result for PG and PS, D1 and R1 sometimes are smaller than the number of runs. In contrast, MEME by default reports all retrieved results irrespective of their scores and therefore the number of datasets or runs with an output by definition equals the number of runs.
- *Recovery rate (RR)*: The RR determines the percentage of the output (D1 for synthetic datasets) (R1 for real datasets), for which the predicted motif model corresponds to the ‘correct’ motif model. If a match is found between the predicted and the correct motif model, the recovery is one, otherwise zero. Motif models were compared with MotifComparison (Thijs et al., 2002b). For the synthetic data the correct motif model was based on the embedded TF binding sites, for the real data on the annotated TF binding sites in the reference species (*E. coli* or *S. cerevisiae*). Predicted models in the real data contain besides contributions from sites in the reference species also contributions from yet unannotated sites present in the orthologs. This sometimes causes discrepancy between the predicted and the correct motif model. For this reason predicted motif models that did not pass the MotifComparison threshold were retained if both the species-dependent positive predictive value and species-dependent sensitivity (for definitions see below) were above 50% or if one of the two measures was higher than 80%.

- *Positive predictive value (PPV) and sensitivity (Sens)*: The PPV [$PPV = TP / (TP + FP)$] is a measure for the percentage of true positive (TP) sites amongst the predicted sites (TP+FP). A TP site corresponds in the synthetic datasets to the embedded sites and in the real datasets to the annotated sites. The false positives (FP) correspond to predicted sites, other than those embedded or annotated. The Sens [$Sens = TP / (TP + FN)$] is a measure for the percentage of true sites (TP+FN) that are found by the algorithm, with FN = false negatives corresponding to embedded or annotated sites not recovered by the algorithm. When a predicted site covers at least half the length of the embedded or annotated site, it is considered as a true positive site. For the less studied species other than *E. coli* or *S. cerevisiae* no judgement can be made on whether sites are true or false. Therefore, we defined the species-dependent PPV (spPPV) and species-dependent sensitivity (spSens) by only taking into account the sites predicted/annotated for the genes of the reference species. In the output tables of the Results section the PPV, spPPV, Sens and spSens are described per test and represent the mean of these values over all datasets/runs with a correct output (recovery equal to one) within a single test.

3.3 Results

3.3.1 Design of the test datasets

In this study we assessed the specific contribution of the coregulation, the orthologous and the combined space on motif discovery (Figure 3.1 Panel A). Success rates observed in the coregulation space were treated as baseline levels. In the combined space we tested under which conditions adding orthologs improved the baseline success rate observed in the coregulation space. We tested the effect of changing the topology by which the orthologs are related, the phylogenetic distances and the number of the added orthologs (Figure 3.1 Panel B). Lastly, we evaluated the success rate of the algorithms when only orthologous information is available, also by using different conditions. In each space we performed tests on datasets with different signal to noise ratios (Figure 3.1 Panel C). We refer to ‘changing the signal to noise ratio’ as any manipulation that lowers/increases the degree to which the motif is statistically overrepresented in the dataset compared to the background e.g. by changing the degree of degeneracy of the motifs or by leaving out TF binding sites.

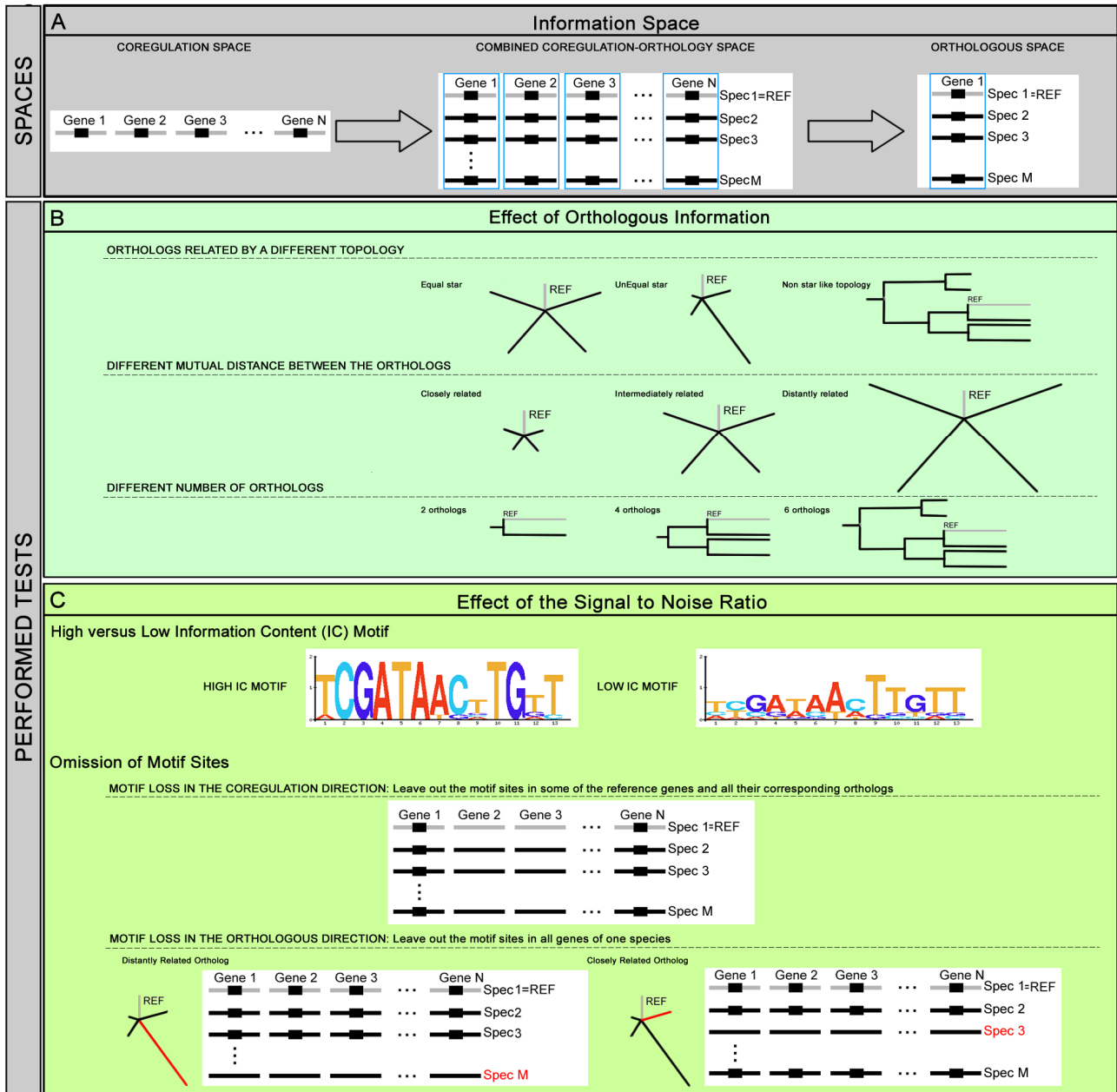


Figure 3.1. Caption on next page.

Figure 3.1. (On previous page) Overview of the test setup. Panel A presents the three different information spaces in which motif discovery was assessed: the coregulation, the combined coregulation-orthology and the orthologous space. The coregulation space consists of a set of non-coding sequences from a reference species (Spec1 = REF) that each contain at least one TF binding site for a common TF (indicated by Gene 1 to Gene N). For the combined space, we extend the coregulation space with orthologous sequences selected from different species (indicated by Spec 2 to Spec M). One reference gene together with its orthologs is referred to as an *orthologous set* (indicated by a blue frame). The combined space thus consists of multiple orthologous sets while the orthologous space consists of a single orthologous set. We assessed the specific contribution of each space to the success rate of motif discovery by performing the tests summarized in panels B and C. At first we tested the effect of adding different types of orthologous information as shown in Panel B. These tests involve changing the topology by which the orthologs are related (equal, unequal star and non star like topology), changing the mutual distance between the orthologs (represented by elongating the branches of the tree) and using datasets with a different number of orthologs. Secondly, the effect of altering the signal to noise ratio of the datasets on the accuracy of the results was tested 1) by changing the degree of degeneracy of the motifs and 2) by omitting motifs sites. We differentiate between leaving out TF binding sites in the coregulation direction versus their omission in the orthologous direction as is illustrated for a dataset in the combined space.

3.3.2 Motif discovery in the coregulation space

Datasets consist of sets of coregulated genes from the reference species. We tested the ability of the algorithms to recover motifs in datasets with different signal to noise ratios. The most trivial task consists of detecting a high IC motif in a dataset where each sequence contains a motif instance (Figure 3.2 (A)). We also assessed whether the motif discovery tools could recover motifs in datasets with lower signal to noise levels e.g. by searching for a low IC motif (Figure 3.2 (B)) or by searching for a high IC motif in a dataset where not all sequences contain a motif instance (Figure 3.2 (C)). We applied those tests on both synthetic and real datasets. Figure 3.2 summarizes the results for the synthetic datasets (from Table S5 (A) and Table S6) as these reflect the most important tendencies. Details on the results for the real datasets can be found in Table S5 (D).

Results were evaluated by ‘performance measures’ and ‘quality measures’. The ‘performance measures’ describe whether the motif discovery tool is able to retrieve the motif model of the embedded motif in a particular test. They correspond to *the number of datasets with an output* (D1) and the *recovery rate* (RR) that indicates the percentage of outputs in which a correct motif was predicted. The ‘quality measures’ defined as the *positive predictive value* (PPV) and the *sensitivity* (Sens) describe whether and how many of the true embedded TF binding sites contribute to the predicted motif model. In the figures the number of datasets with an output (D1) is indicated by a clear box. The number of those datasets that has a correct outcome (D1*RR) is indicated by the black area in the clear box. A larger fraction of the black area in the box (RR) indicates that a larger fraction of the output is correct. The best results are thus obtained if most of the outputs contain a true motif model (largely filled boxes) of a high quality (the latter is indicated by the PPV and Sens approaching 100).

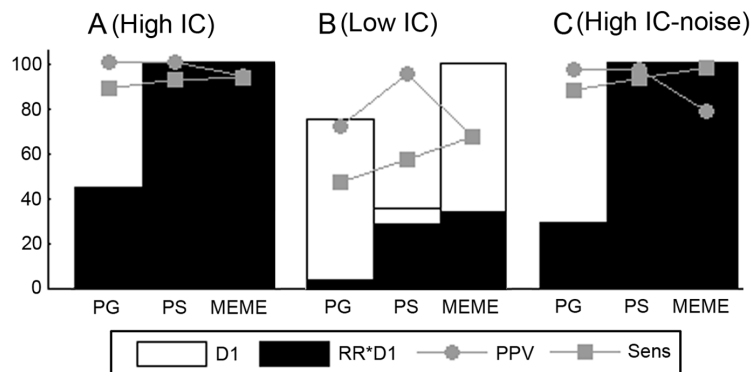


Figure 3.2. Results for motif discovery in the coregulation space. Each dataset consists of ten coregulated genes from the reference species (proximity 0.80). Panel A displays the results for a synthetic dataset in which all sequences contain a site sampled from a high IC motif (A). Panel B shows the results for a dataset in which all sequences contain a site sampled from a low IC motif (B) and panel C shows the results of a dataset where the TF binding site is missing in two out of ten sequences. The remainder of the sequences contains a TF binding site sampled from the high IC motif. Results were assessed by the performance measures D1: the number of datasets with an output out of 100 datasets, D1*RR: the number of datasets with a correct output and the quality measures PPV (the percentage of true sites among the predicted TF binding sites, averaged over all correct outputs) and Sens (the percentage of the true sites recovered by the algorithm, averaged over all correct outputs).

In Figure 3.2 we see that the results were consistent for the three tested algorithms. It shows that coregulation information is sufficient to detect the correct motif provided that the motif has a high IC. For a low IC motif, both the RR and the motif quality (assessed by PPV and Sens) drop. More specifically we had to lower the tracking threshold T of PG to 0.05 in order to still retrieve this low IC motif. Lowering the tracking threshold results for PG in general, in a higher number of datasets with an outcome (D1), but at the cost of a decreased RR and PPV. Of the three algorithms tested, PS performed best for these low IC motifs with a RR equal to of 80.6%, compared to a RR of 34% for MEME and 5.3% for PG. As shown in Figure 3.2 (C), all algorithms are quite robust against the presence of sequences without TF binding site provided the motif itself is sufficiently pronounced. Based on these results, we expect that including orthologous information will be beneficial if it increases the signal to noise ratio in the dataset e.g. when searching for a low IC motif.

3.3.3 Motif discovery in the combined coregulation-orthology space

In this section we assessed to what extent adding orthologous information to the coregulation space improves motif discovery. For the algorithms that rely on a phylogenetic model we expect that their results will depend on the accuracy with which the used phylogenetic tree approximates the true phylogenetic distances between the used intergenic sequences. For real data approximating an optimal tree is not obvious as the intergenic sequences can not accurately be aligned. The best results were obtained with a tree that is based on a 'neutral' evolution rate. Using a protein tree seriously deteriorated

the results obtained by the phylogenetic motif discovery algorithms as the true evolution rate of the intergenic sequences is underestimated (for more details see Table S3). In all tests, we used for the phylogenetic algorithms the tree based on a neutral evolution rate. If the input sequences were left unaligned, PG and PS will just like MEME treat the sequences independently.

3.3.3.1 Effect of the phylogenetic distances between the orthologs

Datasets consist of coregulated genes in the reference species (coregulation space) complemented with their respective orthologs (orthologous space). A reference gene together with its orthologs constitutes an orthologous set. For the first set of tests, the relatedness between the sequences in an orthologous set was modeled by a '*star topology with equal distances*'. Each orthologous set consists of the reference sequence (proximity of 0.80) and four *equally* distant orthologs. The tests consist of changing the distance (~"proximity") for these four orthologs that were added to each coregulated reference gene.

All results for a high and low IC motif resumed in Table S5 (A) reflect the same tendency, summarized for one representative example in Figure 3.3. Figure 3.3 shows how the discovery of a low IC motif was affected by adding to a set of coregulated reference genes (Figure 3.3 (A)), either closely related orthologs (proximity of 0.90, Figure 3.3 (B)), intermediately (proximity of 0.50, Figure 3.3 (C)) or distantly related orthologs (proximity of 0.20, Figure 3.3 (D)). Adding orthologous information improved the RR for all algorithms (fraction of the black area). The best results were obtained for a proximity of 0.50 (Figure 3.3 (C)) and under these optimal conditions, algorithms that use an evolutionary model clearly outperform the non-phylogenetic motif discovery algorithm in finding high quality motifs (both Sens and PPV). All algorithms are sensitive towards deviations from the optimal phylogenetic distance between the added orthologs. Too closely related orthologs (Figure 3.3 (B)) imply many local optima and this resulted for all algorithms, compared to the more optimal situation, mainly in a decrease of the RR. This drop in RR was most obvious for MEME as it does not use an evolutionary model. PS performed best (highest RR) for these datasets where the motif is less pronounced. Adding orthologs that were all very distantly related (Figure 3.3 (D)) was mainly deleterious for the phylogenetic algorithms as they depend on the quality of the prealignments: misalignment of TF binding sites or gaps introduced within the sequence of the TF binding sites make it harder or even impossible to retrieve these TF binding sites, which resulted in a lower motif quality (especially a lower Sens) for PG and PS compared to MEME. In some cases leaving the distant orthologs unaligned can compensate for the loss in sensitivity (Table S5 (A)).

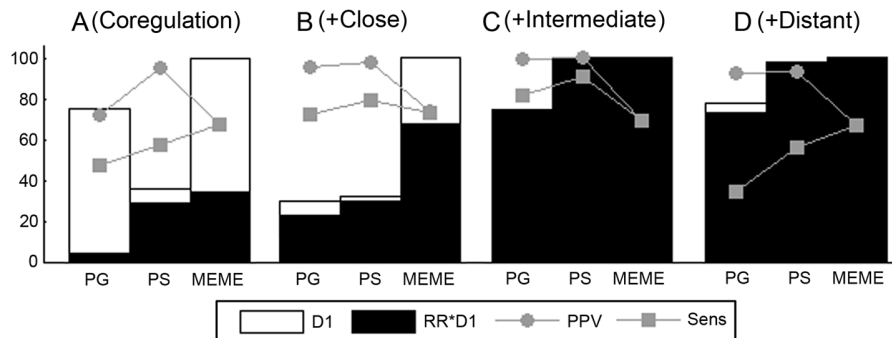


Figure 3.3. Effect of adding orthologs with distinct phylogenetic distances on motif discovery in the combined space. Results are displayed on the retrieval of a low IC motif in a synthetic dataset. Panel (A) shows the results for the coregulation space that consists of ten coregulated reference genes. The remaining panels represent the results for the combined space that consists of the ten coregulated reference genes together with their orthologs, also referred to as ten orthologous sets. Each orthologous set consists of five prealigned sequences related through an equal star topology: the reference sequence with proximity 0.80 and four equally distant sequences with proximities of respectively 0.90 (B), 0.50 (C) and 0.20 (D). For the measures D1, D1*RR, PPV and Sens see Figure 3.2.

In a second set of tests, we examined if adding one distantly related ortholog to a set of closely related orthologs reduces the number of local optima and hence improves the motif discovery results. To this end we used for each coregulated reference gene an orthologous set for which the relatedness was modeled by a ‘*star topology with unequal distances*’. Each orthologous set consists of four closely related orthologs with proximities of respectively 0.80 (the ortholog of the reference species), 0.90, 0.85 and 0.75 and one distantly related ortholog with a proximity of 0.20. All results represented in Table S5 (B) confirmed our expectation: compared to using orthologous sets containing only the four closely related orthologs, adding one distantly related ortholog to the orthologous set of each coregulated reference gene improved the RR of all algorithms. For the phylogenetic algorithms the number of datasets with an output (D1) increased, especially for the low IC motif. The increase in RR was sometimes at the expense of a small sensitivity (Sens) loss for the predicted motif, which was mainly caused by the algorithms not being able to detect the TF binding sites in the distant orthologs. This was confirmed by specifically calculating the sensitivity in the distant ortholog (species-dependent sensitivity, spSens) which was indeed lower than the overall sensitivity (results in Table S5 (C)). In this example where the synthetic datasets were particularly easy to prealign (equal sequence lengths), including the distant ortholog in the prealignment of the orthologous sets improved the results of both phylogenetic algorithms.

3.3.3.2 Effect of the number of added orthologs

For each dataset, we started off with a real set of coregulated genes in the reference species (the target genes of respectively LexA, TyrR in *E. coli* and URS1H, RAP1 in *S. cerevisiae*) and tested the effect of gradually adding more distant orthologs to these

reference genes. All results are shown in Table S5 (D). As for most tests the performance parameters (R1 and RR) reached their maximum level, the most striking results for both the bacterial and yeast datasets relate to changes in motif quality. To visualize this tendency in motif quality observed for both the bacterial and yeast datasets we used a combined ‘quality’ metric, the F-value, defined as the harmonic mean of spPPV (species-dependent PPV) and spSens (species-dependent sensitivity). Figure 3.4 displays the difference between the F-value obtained from searching in the combined coregulation-orthology space and the F-value obtained from searching in the coregulation space only. The results are shown for datasets in which for each coregulated gene the orthologous sets contain respectively two (Figure 3.4 (A)), four (Figure 3.4 (B)), five (yeast)/six (bacteria) prealigned orthologs (Figure 3.4 (C)), and five/six unaligned orthologs (Figure 3.4 (D)). A positive value of the F-value difference thus indicates a positive effect on the motif quality of adding orthologs to the coregulation space, while a negative value indicates the negative effect.

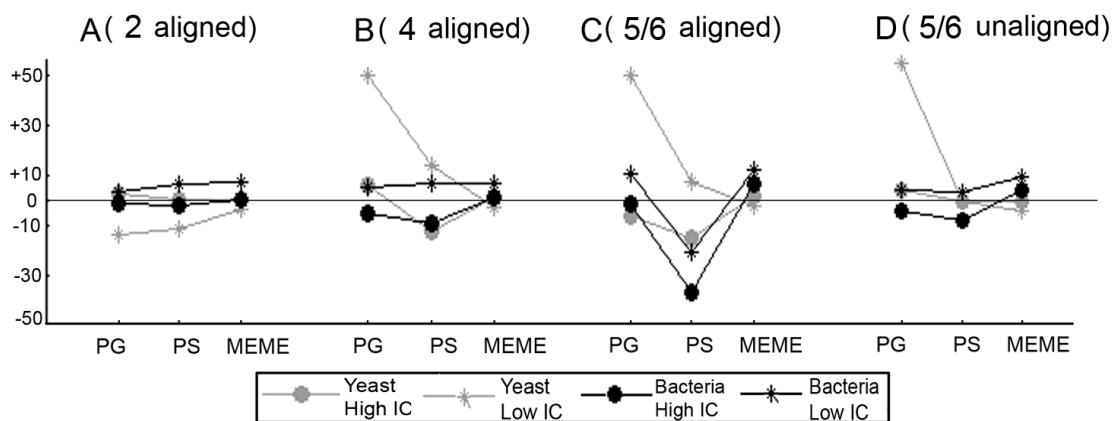


Figure 3.4. Effect of the number of added orthologs on motif discovery in the combined space. Results on the retrieval of both a high and a low IC motif are displayed for the real datasets: 1) results from the Gamma-proteobacterial datasets are indicated as black curves and 2) those of the *Saccharomyces* dataset are indicated as gray curves. Results for the high IC motif are indicated by circles and correspond to those obtained for LexA (bacterial dataset) or URS1H (yeast dataset), results for the low IC motif are indicated by stars and correspond to those obtained for TyrR (bacterial dataset) or RAP1 (yeast dataset). The panels represent the results of a dataset containing for each coregulated reference gene two (A), four (B) and six (for the bacterial datasets) or five (for the yeast datasets) prealigned orthologs (the reference gene included) (C). Panel (D) represents the results of a dataset containing for each coregulated reference gene six or five unaligned orthologs (the reference gene included). Results were assessed by the F-value defined as the harmonic mean of the spPPV (the percentage of true sites amongst the predicted TF binding sites for the reference species, averaged over all correct outputs) and the spSens (the percentage of the true sites found by the algorithm for the reference species, averaged over all correct outputs). The reference species are respectively *E. coli* (bacterial data) or *S. cerevisiae* (yeast data). The Y-axis represents the difference between the F-value obtained from searching motifs in the combined coregulation-orthology space and the F-value obtained from searching in the coregulation space only.

In general the results confirm what we already observed for the synthetic data (see previous section: ‘Effect of the phylogenetic distances between the orthologs’): at first, adding orthologous information has more impact on the results when searching for a low

IC motif than when searching for a high IC motif. Adding orthologs barely improved the motif quality when searching for a high IC motif (LexA and URS1H) (Figure 3.4).

Secondly, the quality of the motifs retrieved by the phylogenetic tools is more sensitive towards the type of orthologs that was added than MEME because their results depend on the correctness of the prealignments. Figure 3.4 (C) shows that for PS, the F-value difference dropped drastically when adding the more distantly related orthologs that can no longer be accurately aligned with the closely related ones. The effect was more pronounced for the bacterial datasets that were the most difficult to prealign. As a result leaving all orthologs unaligned in those cases of misalignment (Figure 3.4 (D)) improved the quality of the motifs retrieved by PS. All four panels in Figure 3.4 show that for MEME, the effect of adding orthologs on the quality of the retrieved motif is rather small.

Additional tests on synthetic data (see Table S7) ensured us that the differences in performance between the motif discovery algorithms we observed when adding orthologs could indeed be attributed to the gradually increased phylogenetic relatedness between the added orthologs, rather than to the intrinsically different way PG and PS handle non star like topologies in their phylogenetic model.

3.3.3.3 Simulation of motif loss in the orthologous and coregulation direction

Previous tests showed that adding orthologs was beneficial, provided that they contain the TF binding site. However, adding orthologous sequences from species in which the mode of regulation is not conserved will increase the noise in the input datasets (Perez and Groisman, 2009; Tanay et al., 2005). Here we simulated this situation by adding orthologs to a set of coregulated reference genes, but assuming that all added sequences derived from one species did not contain the TF binding site. The relatedness between the sequences in the orthologous sets was modeled by a star topology with unequal distances. Figure 3.5 summarizes these results for a high IC motif (as given in Table S6). Figure 3.5 (A) shows the reference level of performance when a TF binding site is present in all sequences of the orthologous sets. In the remainder of the panels the results are shown of replacing in each orthologous set the TF binding site by a random site in the sequence derived from either a closely related species (proximity 0.75, Figure 3.5 (B)) or a distantly related species (proximity 0.20, Figure 3.5 (C)).

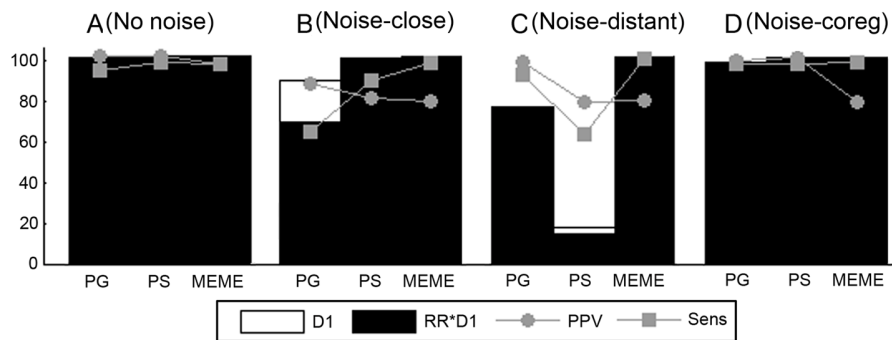


Figure 3.5. Effect of motif loss on motif discovery in the combined space. The results are displayed for a synthetic dataset containing sites sampled from a high IC motif. Each dataset consists of ten coregulated reference genes complemented with their orthologs, also referred to as ten orthologous sets. Each orthologous set consists of five prealigned sequences related through an unequal star topology: four closely related orthologs with proximities of respectively 0.80 (reference ortholog), 0.90, 0.85 and 0.75 and one distantly related ortholog with a proximity of 0.20. Panel (A) represents the results when a TF binding site is present in all sequences of the orthologous sets. Panels (B) and (C) display the results when motif loss occurs in all sequences derived from respectively a closely ($q=0.75$) or a distantly ($q=0.20$) related species. Panel (D) shows the results when motif loss occurs in two out of ten coregulated reference genes and in all their corresponding orthologs. For the measures D1, RR*D1, PPV and Sens see Figure 3.2.

As shown in Figure 3.5 (B and C), all three algorithms were affected by adding orthologs without TF binding site. For PG the absence of the TF binding sites in closely related orthologs (Figure 3.5 (B)) had a more pronounced negative influence (drop in RR, PPV and mainly Sens) than when the TF binding site was absent in the distantly related orthologs (Figure 3.5 (C)). For PS the situation is reversed: the presence of distant orthologs without TF binding site resulted in a drastic drop in D1 and in the Sens compared to the reference situation (where the TF binding site was present in all orthologs) (Figure 3.5 (A)) or to the situation where the TF binding site was absent in the closely related orthologs (Figure 3.5 (B)). The difference in response between PG and PS towards the absence of TF binding sites is related to the intrinsically different way they treat the prealignments (see also Table S6 for more information). When the TF binding site is missing in the distant orthologs, regions that normally would contain the TF binding site will be left unaligned by Dialign or will result in a gapped alignment by ClustalW. In neither case PS will search for motifs in these regions of the prealignment while PG will correctly treat these regions independently and search for motifs in the remaining part of the prealignment. Missing TF binding sites in the close orthologs on the other hand are better handled by PS as it relies on a global alignment strategy. As closely related orthologs align any way well over the total length of their sequence, a global alignment is not too much disturbed by a missing TF binding site in one of these close orthologs. For a local alignment this often interferes with the correct identification of the orthologous regions.

In addition, for PS also the weighting scheme used during the update step of the motif WM affects its specific behavior towards missing TF binding sites in distantly related sequences. Distantly related orthologs get a higher weight than closely related ones, so a false positive site in a distant ortholog has a more negative impact on the WM update than a false positive site in a close ortholog.

For MEME mainly the motif quality (more in particular the PPV) was decreased by omitting TF binding sites, but in contrast to what was observed for the phylogenetic algorithms this effect was largely independent of the type of ortholog from which the sites were omitted (Figure 3.5 (B, C)). By setting the number of asked TF binding sites equal to the number of input sequences, the number of sites we searched is overestimated when leaving out TF binding sites. This effect of overestimating the number of TF binding sites affects the quality of the motif retrieved by MEME that does not internally filter out low quality TF binding sites.

As for the coregulation space, we also tested for the combined space the effect of missing TF binding sites in *the coregulation direction*. This situation was mimicked by assuming that two of the reference genes were not truly coregulated with the other genes. The TF binding site is thus absent in these two genes and in their respective orthologs. Figure 3.5 (D) shows that this had almost no effect on the results, except for a PPV drop in case of MEME with the same reason as above.

Figure 3.5 (D) also shows that omitting TF binding sites in the coregulation direction has less drastic effects on the results (most obvious for PG) when also the orthologs are provided than in the absence of the orthologs (Figure 3.2 (C)), even though some of the orthologs might not contain the TF binding site.

3.3.4 Motif discovery in the orthologous space

Lastly we assessed the performance of the algorithms in the presence of only orthologous information. We used a test setup similar as in the combined coregulation-orthology space, but instead of using a set of coregulated reference genes complemented with their orthologs, we used only one reference gene together with its orthologs (~ one orthologous set). The tests consist of changing for this orthologous set the number of orthologous sequences and their phylogenetic relatedness (equal or unequal star topology). We also assessed in real datasets the effect of gradually adding more orthologs with increasing phylogenetic distance to the orthologous set.

All results for the synthetic data are shown in Table S8 (A). Figure 3.6 shows representative results for the tested algorithms in detecting respectively a single embedded high (at the top) and low (at the bottom) IC motif. Figure 3.6 (A) and (B) show the results for the orthologous set containing respectively five and ten orthologs related

through an equal star topology with proximity 0.50. Figure 3.6 (C) shows the results for the orthologous set containing five orthologs related through an equal star topology with proximity 0.90 and Figure 3.6 (D) shows the results for the orthologous set containing five orthologs related through the earlier described unequal star topology (see previous section § 3.3.3: ‘Motif discovery in the combined space’). All algorithms performed best on datasets with 10 prealigned orthologs related to each other with a proximity of 0.50 (Figure 3.6 (B)). For this setting, PG and PS outperformed MEME (higher RR and motif quality), especially for the low IC motif. However, for PS the number of datasets with an output was extremely low ($D1 < 10$).

By keeping track of the motif positions sampled during the early iteration stage of PS, we noticed that the sampler explored the solution space less for the prealigned input than when leaving the sequences unaligned. By getting stuck in non-overlapping local optima for each re-initialization, no centroid output could be obtained (low D1).

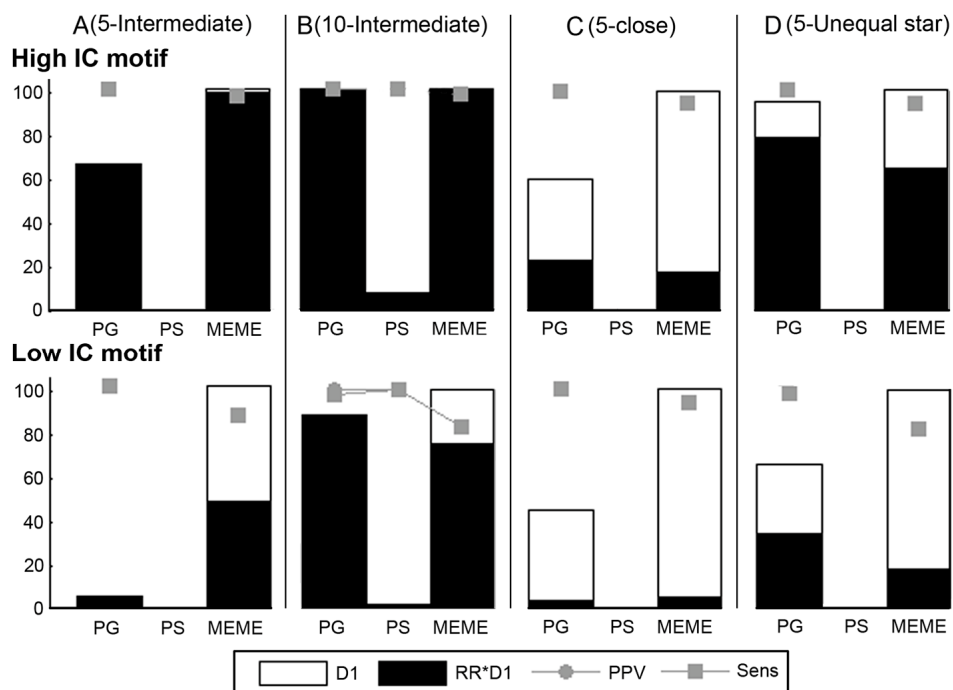


Figure 3.6. Results for motif discovery in the orthologous space. Results are displayed for a synthetic dataset with TF binding sites sampled from a high IC (on top) and a low IC motif (below). Each dataset consists of only one reference gene and its orthologs, referred to as one orthologous set. Panel (A) and (B) represent the results when the orthologous set contains respectively five and ten prealigned orthologs related through an equal star topology with a proximity of 0.50. Panel (C) represents the results when the orthologous set contains five prealigned orthologs related through an equal star topology with a proximity of 0.90 and panel (D) represents the results when the orthologous set contains five prealigned orthologs related through an unequal star topology. Note that for most tests the PPV equaled the Sens resulting in overlapping dots. For the measures D1, RR*D1, PPV and Sens see Figure 3.2.

The performance of all algorithms dropped when the number of prealigned orthologs was lowered to 5 (Figure 3.6 (A, C and D)) in which case PS even did not longer retrieve an output. Using too closely related orthologs (Figure 3.6 (C)) resulted in a severe further decrease of the RR for both MEME and PG (despite lowering the tracking threshold). As was also the case in the combined space, we can increase the information level of the datasets by adding one distant ortholog through the use of a ‘star topology with unequal distances’ (Figure 3.6 (D)): this improved the performance (D1 and RR) of both PG and MEME considerably compared to the situation with closely related orthologs of equal phylogenetic distance.

For the real datasets, we used two reference targets genes of LexA, two of TyrR, two of URS1H and two of RAP1 each containing exactly one TF binding site for their respective regulators and we added to each of these individual genes their orthologs resulting in 8 datasets in total. As was done in the combined space, these orthologs were added gradually with increasing phylogenetic distances. The results on the real datasets in the orthologous space (Table S8 (B)) were rather poor and similar to what was observed for the synthetic data: when the dataset contains too few closely related orthologs (less than six for bacterial genes and four for the yeast genes) the algorithms failed in detecting the motif (data not shown). Increasing the information level of the datasets by adding extra orthologs resulted in PG and MEME becoming able to retrieve the motif for at least some of the datasets. For PG the best results were obtained by including the phylogenetic relatedness by means of a prealignment. PS totally failed on the real data in the orthologous space, even for the maximum number of orthologs (irrespective of whether they were aligned or left unaligned).

3.4 Discussion

In this work, we tested the impact of using coregulation and/or orthologous information on the efficiency of regulatory motif discovery by two representative motif discovery algorithms with an evolutionary model. We designed appropriate benchmark datasets and made an exhaustive evaluation of both algorithms together with MEME, a well-known reference algorithm (All benchmark datasets are available at http://homes.esat.kuleuven.be/~kmarchal/Supplementary_Storms_Valerie_PlosONE). Parameter tuning required a detailed analysis of how parameters influence test results. This analysis (see also Text S3) together with guidelines describing how the selection of the best tool depends on the composition of the dataset is summarized in Table 3.1.

Chapter 3 - The effect of orthology and coregulation on detecting regulatory motifs

Table 3.1 Summary of user-guidelines.

PROBLEM	CONSTRUCTING DATASET	PREFERRED TOOLS	REMARKS
1. COREGULATION SPACE			
Maximizing the signal to noise ratio in the dataset (i.e. the enrichment of TF binding sites in the dataset) improves the success rate.	Only select sequences that are likely to contain the motif. Keep the input sequences as short as possible. Adding orthologs (see 2: combined space) improves the success rate at a low signal to noise ratio.	PS: the ensemble centroid solution guarantees a high success rate for datasets with low signal to noise ratios. MEME: easy to use with performances comparable to those of PG and PS.	Both PG and PS provide a statistical procedure to filter out non-significant TF binding sites => Overestimating the 'expected number of TF binding sites' affects the performance less than underestimating them. For MEME misestimating the expected number of TF binding sites affects the motif quality.
2. COMBINED SPACE			
It is crucial to use a phylogenetic tree that reflects the true evolutionary distances between the intergenic sequences.	Use a tree based on a neutral evolution rate or a protein tree with corrected distances to prevent underestimating the evolution rate.	Both PG and PS are sensitive to overestimating the evolutionary proximity of the orthologous intergenic regions.	The type of topology (star, tree like structure) does not affect the performance of the phylogenetic tools.
<p>The characteristics of the added orthologs: mainly the evolutionary distance between them influences the results by affecting the trade-off between align-ability of the orthologs and the information level of the dataset.</p> <p>Close orthologs: the dataset contains little information</p> <p>Intermediate orthologs: this is the optimal situation.</p> <p>Distant orthologs: the dataset contains more information, but the alignment might get deteriorated.</p>	<p>Close ~ $q=0.90$, the orthologs align for almost 100 %. In this case add at least one distant ortholog to increase the information level of the dataset.</p> <p>Intermediate ~ $q=0.50$, the sequences can be aligned and contain sufficient information (clear phylogenetic shadowing of the motif).</p> <p>Distant ~ $q=0.20$, the prealignment looks bad. For the phylogenetic tools it is better to leave the difficult to align sequences unaligned.</p>	<p>Close: the phylogenetic tools outperform MEME because of the multiple local optima in the data.</p> <p>Intermediate: for this optimal evolutionary distance the phylogenetic tools outperform MEME.</p> <p>Distant: an unreliable prealignment deteriorates the results of the phylogenetic tools. MEME performs better under those conditions. In general PG better handles these difficult to align datasets than PS.</p>	<p>The number of orthologs to be added is of less importance for the success rate. Good results can already be obtained with 4 orthologs, provided that they have a good evolutionary distance.</p> <p>PG is easier to use than PS: 1) when the dataset contains a different number of orthologs per gene, PG adapts the input phylogenetic tree automatically while for PS it needs manual interference. 2) PS has a long running time compared to PG and MEME.</p>

PROBLEM	CONSTRUCTING DATASET	PREFERRED TOOLS	REMARKS
2. COMBINED SPACE (continued)			
Motif Loss in a closely related ortholog or in a distantly related ortholog increases the noise in the dataset.	Avoid sequences for which one expects that the mode of regulation has changed (mostly the distantly related sequences).	PG/PS performs better if the motif is omitted in the distant/close ortholog. MEME: not dependent on the type of ortholog.	
3. ORTHOLOGOUS SPACE			
The same issues as in 2 are valid regarding the phylogenetic tree and the characteristics of the orthologs.	The more orthologs are added, the better the results.	PG performs best when the orthologs are prealigned and slightly outperforms MEME. PS underperforms in the orthologous space.	Observing a PG output that only contains unaligned TF binding sites indicates that the input tree underestimates the true evolution rate. In that case, lower the proximities.

From our results it appeared that coregulation data allow all three motif discovery algorithms to retrieve the motif if the signal to noise ratio in the data is high. In real life situations it is more common to encounter datasets with a low signal to noise ratio, as biologists often define coregulated gene sets based on results derived from noisy high throughput experiments. Moreover the length of the intergenic sequences can be long compared to the length of the TF binding sites (Van Hellemonst et al., 2005) and often the motifs themselves are heavily degenerated. Under such conditions, adding orthologous information to the coregulation space can improve the results. There seems to exist an optimal phylogenetic distance between the added orthologs, for which all algorithms retrieved the best results. This optimal distance corresponds to orthologs that are still alignable, but show a sufficient level of divergence so that non functional background sequences are no longer conserved and the signal of the conserved TF binding site stands out in the background sequence. For applications of phylogenetic footprinting, where motifs are searched for in the orthologous space, there is still room for improvement. All three algorithms performed poor, partially because they were originally developed and tuned towards searching for motifs in the coregulation or the combined coregulation-orthology space.

In all tests we observed some reoccurring effects that can be explained by the algorithmic specificities of the applied motif discovery algorithms.

At first we consistently observed that PS outperforms PG and MEME when the signal to noise ratio drops in the dataset. This is because PS uses an ensemble of solutions to define the statistically most overrepresented motif in the dataset whereas both PG and MEME report a single optimal solution. Especially in the presence of multiple local

optima, such ensemble strategies have proven to be more successful in estimating the true optimum than searching for a single optimal solution (Reddy et al., 2007). However, this advantage of using an ensemble solution comes at the expense of much longer running times (e.g. a dataset with 10 orthologous sets each containing five orthologs had a running time around 8 hours, for PS, compared to several minutes for PG and MEME).

Secondly, we would expect that modeling the relation between orthologous sequences when searching for motifs in the combined or the orthologous space would improve results over those obtained with MEME, or with PG and PS when leaving the sequences unaligned. Using an evolutionary model in combination with a tree that correctly represents the phylogenetic distances between the used sequences is indeed advantageous when adding closely related sequences. Closely related sequences that are treated independently harm motif discovery by inducing multiple local optima as observed for MEME. PG and PS can better handle this problem of local optima as they constrain the search space by prealigning conserved regions and by treating those regions simultaneously. In addition their evolutionary model helps to distinguish conservation due to evolutionary proximity from conservation due to functionality as the prealignment itself is often uninformative (Blanchette and Tompa, 2002; Tompa, 2001). Adding distantly related orthologs usually relieves the problem of the local optima, but often occurs at the cost of the motif quality as TF binding sites in the distant orthologs are harder to find (less similar to the other TF binding sites) or the distant orthologs disturbs the prealignment needed for the phylogenetic algorithms. The accuracy of the prealignment seemed in general the major bottleneck for the phylogenetic motif finders. PG in general handles better these difficult to align datasets by combining a local alignment strategy with a more flexible way of assigning TF binding sites. The different way of treating the prealignment by PG and PS also explains the different behavior of PG and PS towards omitting TF binding sites in the orthologous direction. For PS we also observed that the use of a weighting scheme in a non-ideal situation (incorrect prealignment and missing TF binding sites in the distant ortholog) negatively influences the results. This implies that when using PS, the user can better omit distant sequences for which he is not sure that the mode of regulation is still conserved.

Lastly, all used algorithms underperform when searching for motifs in only a set of orthologous sequences. This effect was most pronounced for PS that only retrieved an output when leaving the sequences unaligned and suppressing the use of the phylogenetic model. This failure of PS relates to the fact that the sampling ‘model-update step’ (see chapter 2, § 2.3.2 ‘Scoring methods’) does not sufficiently explore the search space in the absence of coregulated information. PG which uses a different search strategy better explores the search space in the orthologous space.

Having an insight in this relation between the obtained results and the working principles of the algorithms provides developers hints for further improvements. For instance the ease with which a basic algorithm as MEME can be used largely compensates for the slightly higher accuracy that is obtained with the more complex phylogenetic algorithms. Based on our experience we would therefore suggest of using MEME to get a first insight into the data. This will help tuning the parameters of the more complex phylogenetic algorithms that on their turn can further improve the results e.g. by retrieving more ‘true’ and less ‘false positive’ sites. User-friendliness is one of the major issues in determining which algorithm to use. Most of the current phylogenetic algorithms are still in their developmental phase and do not yet provide the same user-friendliness as more settled algorithms such as MEME. Moreover, as the quality of the results of the phylogenetic algorithms heavily depends on the correctness of the prealignments, developing ways to account for phylogenetic relatedness, independent of a prealignment is a future challenge. Care should also be taken when introducing specific ways to model the relation between the orthologous sequences. For instance, for PS the use of the weighting scheme has a very counterintuitive effect when TF binding sites are missing in the orthologous direction. The development of algorithms that can better cope with phenomena of ‘TF binding site turnover’ during evolution (Ray et al., 2008) will hopefully result in more realistic and informative models. Lastly the ensemble strategy of PS definitely is useful, but can be computationally limiting.

Chapter 4 PHYLO-MOTIF-WEB: an ensemble workflow on the web for *de novo* discovery of DNA binding sites using phylogeny

4.1 Introduction

The *de novo* identification of regulatory motifs is one of the oldest problems in bioinformatics with nearly a hundred algorithms published in the last 25 years. Despite the abundance of motif discovery algorithms, most programs are difficult for non-expert users to readily apply to their uncharacterized datasets. This is unintentionally a consequence of motif discovery algorithms getting more and more complex and thus requiring a lot of pre- and post-processing steps. This increase in algorithmic complexity stems from attempts to improve motif discovery performance by incorporating additional information to guide the motif search.

Initially motif discovery was performed on the non-coding sequences of several coregulated genes from a single genome. Web servers like MEME SUITE (Bailey et al., 2009), Motif Tool Manager (Phan and Furlotte, 2008), SCOPE (Carlson et al., 2007), iMotifs (Piipari et al., 2010) and many others provide easy access and use of such ‘coregulation-based’ motif discovery algorithms. Although those motif finders have been shown to work successfully in yeast and other lower organisms, they suffer from low specificity and sensitivity in higher organisms due to the low signal to noise ratio of short, degenerated binding sites in long background sequences (Das and Dai, 2007; Tompa et al., 2005).

Large-scale genome sequencing provided a lot of new information and led to a successful extension for motif discovery: the use of ‘evolutionary conservation’ by means of ortholog alignments, also known as phylogenetic footprinting (Duret and Bucher, 1997). Initially, the information extracted from ortholog alignments was used in a discriminative way as a pre-processing step to restrict the search space (e.g. MEME_c (Harbison et al., 2004)) or to confirm/filter predicted binding sites in a post-processing step (Gelfand et al., 2000; Wasserman and Fickett, 1998). This was followed by an integrated approach, where we can distinguish between algorithms that use the full alignment information completed with a phylogenetic tree and an evolutionary model (Siddharthan et al., 2005; Sinha et al., 2004; Prakash et al., 2004; Blanchette and Tompa, 2003) and algorithms that include the alignment information into positional priors (Gordan et al., 2010; Bailey et al., 2010).

For long, ‘evolutionary conservation’ was the most powerful remedy against many false-positive annotations (Vingron et al., 2009) but new techniques like Chromatin immunoprecipitation on chip (ChIP-chip) or followed by DNA sequencing (ChIP-Seq) introduced new informative DNA features that help to differentiate true binding sites (which are bound in vivo by their TF) from random ones (which are not bound in vivo). Features like the binding regions of other TFs and epigenetic factors such as nucleosome occupancy, DNase hypersensitivity and histone modifications have already been applied and shown to improve performance for motif scanning purposes (Won et al., 2010; Ernst et al., 2010; Lahdesmaki et al., 2008). Except for MEME version 4.2+ (Bailey et al., 2010) and PRIORITY (Gordan et al., 2010; Narlikar et al., 2006) that can incorporate positional priors based on multiple DNA features, most of the established *de novo* motif discovery algorithms do not yet support the direct use of such information (Klepper and Drablos, 2010).

In this chapter, we present a new web server for *de novo* motif discovery: PHYLO-MOTIF-WEB. PHYLO-MOTIF-WEB is unique compared to other existing web servers, as it covers all the different pre- and post-processing steps needed to identify potential motifs in a set of non-coding sequences. The PHYLO-MOTIF-WEB workflow allows for integrating DNA features like nucleosome occupancy and histone modifications to define putative regulatory regions in the DNA and it applies an ensemble strategy on the results of multiple advanced motif discovery tools that can integrate phylogeny.

4.2 Materials and Methods

The PHYLO-MOTIF-WEB workflow consists of four main steps as show in Figure 4.1. In this section we explain each of the steps in more detail.

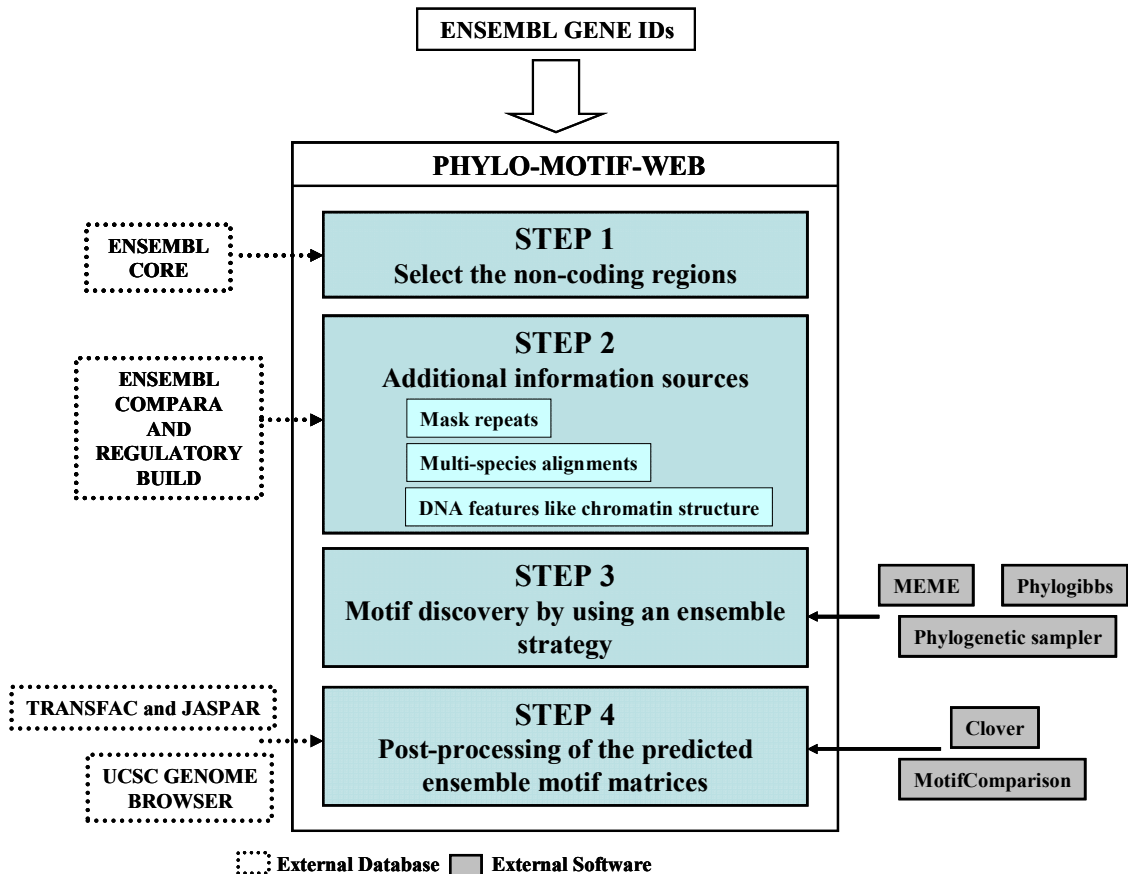


Figure 4.1 Overview of the different sub-steps of PHYLO-MOTIF-WEB. The user input consists of one or more Ensembl gene IDs for a reference species. Step1: Selects the non-coding sequence regions for each reference gene by using the Ensembl Core API modules. Step 2: Allows the user to use additional information sources to guide the motif search e.g. mask repeating patterns or add multi-species alignments by using the Ensembl Compara API modules or use DNA features to restrict the search space by using the Ensembl Regulatory Build. Step 3: Performs motif discovery following an ensemble strategy that clusters the results of multiple runs (covering different parameter settings) of one or more component algorithms like e.g. MEME, Phylogibbs and Phylogenetic Sampler. Step 4: Evaluation of the predicted ensemble motif matrices by calculating their overrepresentation in the input sequences versus random sequence sets using Clover, comparing each predicted motif matrix to TRANSFAC/JASPAR motif matrices using MotifComparison and visualization of the corresponding binding sites in the reference genome using the UCSC Genome Browser.

4.2.1 User input (see Figure 4.1: step1)

PHYLO-MOTIF-WEB is developed to perform a gene-centered motif discovery approach in eukaryotes. The input consists of the Ensembl gene IDs for a set of genes from a reference species that are believed to be regulated by a common TF. For each gene, we select the non-coding sequence region up- and downstream of the transcription

start site (TSS) using Ensemble core API modules (Hubbard et al., 2009). As those non-coding regions are often long in order to ensure covering of all the TF binding sites, restriction to regions with a high ‘regulatory potential’ is necessary to guarantee an accurate performance. For now, the Ensembl database contains the needed additional information sources to define ‘regulatory potential’ (see beneath) only for human (*Homo sapiens*) and mouse (*Mus musculus*).

4.2.2 Additional information sources (see Figure 4.1: step 2)

PHYLO-MOTIF-WEB allows for masking repeating patterns like tandem repeats or transposable elements helping the motif discovery to focus on true binding sites.

Evolutionary conservation is integrated in a non-discriminative way by providing for each reference sequence region, the full alignments blocks, retrieved from Ensembl using the compara API modules. The alignment blocks stem from the multi-species genome alignments of 11 eutherian mammals created by EPO (the Ensembl ‘Enredo, Pecan, Ortheus’ pipeline for whole-genome multiple alignments). As evolutionary closely related species show high sequence similarity and low discriminative power in multiple alignments, we leave the sequences for following species out of the alignment blocks: chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*) and rat (*Rattus norvegicus*), marmoset (*Callithrix jacchus*). The species for which the sequences are included: human, mouse, macaque (*Macaca mulatta*), cow (*Bos taurus*), horse (*Equus caballus*) and dog (*Canis familiaris*). The user can choose beforehand for which species he likes to include the sequences in the alignment blocks (default all six species), but as not every genomic region of the reference species can be aligned to every other genome, this will anyway result in alignment blocks with varying sizes (i.e. different number of orthologous sequences). Besides the alignment blocks, PHYLO-MOTIF-WEB also provides a phylogenetic tree to describe the evolutionary relationships between the orthologous sequences in the alignment block. The phylogenetic tree is based on a neutral evolution rate (3rd position of four-fold degenerated codons) and was proven to be most suited to capture the relationships between the non-coding sequences of the six species (Storms et al., 2010). Providing the alignment blocks in combination with a phylogenetic tree allows for optimal use of orthology information where non-conserved TF binding sites can still be retrieved.

More recently, many other features are described as indicators for DNA regions with ‘regulatory potential’ and thus can be used to improve the motif discovery performance. Chromatin structure is an indirect indicator of regulatory elements. DNase hypersensitive sites, i.e. nucleosome depleted regions that are easily digested by the DNase I enzyme, are associated with regulatory elements (Boyle et al., 2008). Similarly some regulatory elements are known to be enriched for certain histone modifications (Kouzarides, 2007).

Genome-wide profiles of various histone modifications using ChIP experiments have revealed the locations of several putative regulatory regions in different cell types (Barski et al., 2007; Heintzman et al., 2007). Also the location of other functional TF binding sites identified by ChIP-chip or ChIP-Seq is a useful feature as the transcriptional regulation in eukaryotes is often mediated by the concerted interaction of several TFs that bind in close proximity (Van Loo and Marynen, 2009; Gotea and Ovcharenko, 2008). Most of these features are cell type and condition specific, and so ideally used to discover cell type and condition specific TF binding sites (Whittington et al., 2009). However the study of Ernst *et al.* (Ernst et al., 2010), shows that experimentally derived data in one cell type (e.g. DNase hypersensitivity or histone modifications) can be used to predict TF binding in another cell type. PHYLO-MOTIF-WEB uses the ‘Regulatory Build’ pipeline of Ensembl that combines specific DNA features to provide the annotation of potential regulatory regions within the genome (Hubbard et al., 2009). For now available are: Human Regulatory Build version 8 and Mouse Regulatory Build version 3. The user can choose to use regions identified across multiple cell types (based on open chromatin defined by DNase I hypersensitivity mapping and the locations of other TF binding sites) and can extend those in a cell type specific manner (based on histone modifications assayed by ChIP). The regulatory region annotations can be used to narrow down the long non-coding input sequences to hotspots of regulatory elements in a pre-processing step or to evaluate predicted binding sites in a post-processing step.

4.2.3 Motif discovery following an ensemble strategy (see Figure 4.1: step 3)

PHYLO-MOTIF-WEB integrates multiple *de novo* motif discovery algorithms by using an ensemble strategy. For now, the user can add following algorithms to the ensemble strategy: MEME (Bailey and Elkan, 1994) an established algorithm to find TF binding sites in sets of coreregulated genes, Phylogibbs (Siddharthan et al., 2005) and Phylogenetic sampler (Newberg et al., 2007), both more advanced algorithms specialized in integrating orthology information. It is often hard for non-expert users to use these advanced motif discovery algorithms as they have a lot of parameters to tune and their optimal settings depend on the type of input and on the algorithmic properties (Storms et al., 2010). PHYLO-MOTIF-WEB overcomes this pitfall as it runs each component algorithm over the range of most likely parameter settings, leaving the user with very few undefined settings. For each run of each component algorithm, the output consists of a list of predicted motifs in the form of position specific probability matrices, each with their corresponding individual TF binding sites. Our asymmetric clustering approach, also referred to as ‘FuzzyClustering’, resumes all the predicted motif matrices into one final list of putative regulatory motif matrices and works on the level of the ‘individual TF binding sites’. FuzzyClustering is based on the sequential cluster extraction algorithm,

first described in Inoue *et al.* (Inoue and Urahama, 1999), who demonstrated its use in image segmentation and later applied by Joshi *et al.* (Joshi *et al.*, 2008) to extract clusters from gene expression data. The FuzzyClustering algorithm starts with the construction of an overview matrix, as shown in Figure 4.2, to keep track which TF binding sites (columns in the overview matrix) were retrieved for each predicted motif (rows in the overview matrix). To construct the overview matrix, overlapping TF binding sites are merged into one representative binding site in case their overlapping region succeeds the overlap threshold (default = minimum 50% overlap to merge two TF binding sites). Weak TF binding sites can be filtered out the overview matrix by putting a threshold (default = 0.05) on the TF binding site's individual probability. The individual probability for a TF binding site is defined as the ratio of 'the number of times the binding site appeared over all the predicted motifs' and 'the total number of predicted motifs' as shown in Figure 4.2. As each motif discovery algorithm assigns a weight (i.e. the posterior probability score in case of PS and PG and a p-value in case of MEME) to their predicted individual binding sites, those weights can also be integrated in the calculation of the individual binding site probability. This by multiplying the number of times the binding site appeared over all the predicted motifs with the square root of the binding site's weight. In case of overlapping binding sites, the average of their weights is taken and assigned to the representing binding site.

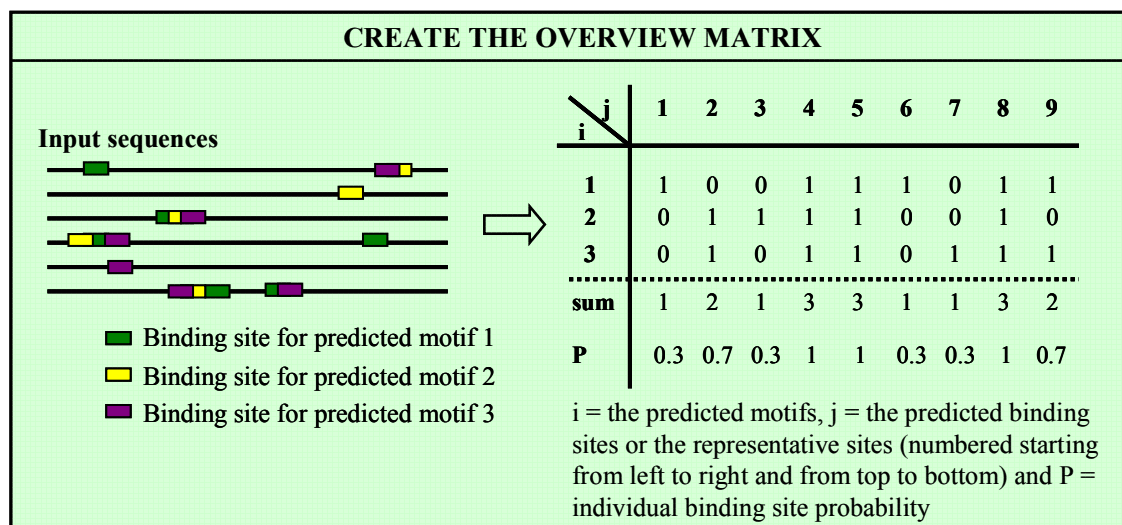


Figure 4.2: Construction of the overview matrix that is based on the predicted individual binding sites. For each predicted motif (3 in total) the corresponding TF binding sites are shown in the input sequences. In case the individual binding sites sufficiently overlap they are merged into one representative binding site (e.g. in the first sequence the binding sites for predicted motif 2 and 3 are merged into one representative binding site). The overview matrix contains '1' in case the (representative) binding site (column j) was retrieved for the predicted motif (row i), if not it contains '0'. In case weights were assigned to the individual TF binding sites by the motif discovery algorithm, the '1' is replaced by the 'binding site weight' in the overview matrix. For each (representative) binding site the individual binding site probability (p) is calculated. In this figure we did not integrate weights.

Based on this overview matrix an asymmetric clustering approach will sequentially group sets of TF binding sites that significantly appear together in sets of predicted motifs. So we get clusters of the form (A, B), where A represents the set of predicted motifs that contribute to B, the set of significantly co-occurring TF binding sites. This allows for each set of TF binding sites to evaluate the fraction of predicted motifs that is covered (i.e. the prediction rate) and to trace back the contribution level of each component algorithm (PG, PS and MEME). TF binding sites and predicted motifs may be assigned to different clusters (referring to the 'fuzzy' aspect of this approach) each with a membership probability that represents their relative importance compared to other TF binding sites/predicted motifs in the cluster. Those membership probabilities are also affected by the original weight assigned to the TF binding site by the motif discovery algorithm.

For each asymmetric cluster reported by our approach, the TF binding sites are aligned to construct a position specific probability matrix (i.e. the ensemble motif), where the contribution of each binding site is weighted by its membership probability. To enhance biological meaningful results we select those ensemble motif matrices that show a reasonable consensus score (default threshold value = 0.5) and have binding sites in a minimal fraction of the input sequences (default = at least one binding site in 50% of the input sequences). The overview matrix is only based on binding sites retrieved for the input sequences of the reference species and not those retrieved for the orthologous sequences, this in order to create species-specific ensemble motif matrices.

4.2.4 Post-processing (see Figure 4.1: step 4)

The post-processing of PHYLO-MOTIF-WEB allows for an optimal evaluation of the final ensemble motif matrices. Each motif matrix is visually presented by its motif logo; the location of the corresponding binding sites on the input sequences is shown graphically and can also be viewed in the UCSC Genome Browser (<http://genome.ucsc.edu/>) to place them in their broader genomic and epigenetic context. The contribution level of each component algorithm is indicated by a pie chart. To evaluate if the motif matrix is specific for the input set, Clover (Frith et al., 2004) is used to calculate the p-value for the motif matrix's overrepresentation in the input set versus random sequence sets sampled from the reference genome. Each final ensemble motif matrix can be compared to known motif matrices in databases as TRANSFAC (Matys et al., 2006) and JASPAR (Bryne et al., 2008) using the MotifComparison software (Thijs et al., 2002b). MotifComparison relies on the Kullback-Leibler distance to measure the similarity between two motif matrices. Matches with a Kullback-Leibler distance < 0.40 (default) and a minimum overlap of 5 bp or a maximum shift of 4 bp between both matrices were regarded as significant.

4.3 The PHYLO-MOTIF-WEB web server

4.3.1 The ‘Run it’ webpage

On the homepage (http://homes.esat.kuleuven.be/~bioi_marchal/PMW/index.html) of the PHYLO-MOTIF-WEB web server, the user can choose to run the PHYLO-MOTIF-WEB workflow by clicking on ‘Run it’. The ‘Run it’ webpage (as shown in Figure 4.3) allows the user to provide personal information like an email address to which the results of the PHYLO-MOTIF-WEB workflow will be sent. Further on, the user can input the Ensembl gene IDs for the chosen reference species and select the regions for motif discovery under ‘Step 1: Select the non-coding regions’. Under ‘Step 2: Additional information sources’ the user can choose 1) to use evolutionary conservation information by selecting the species in which he expects the TF binding sites to be conserved, 2) to use cell type specific DNA features or DNA features applicable to many, different cell types and 3) to use repeat masking. Under ‘Step 3: Motif discovery using an ensemble strategy’ the user can control the parameters for *de novo* motif discovery and for the asymmetric clustering approach (FuzzyClustering).

Figure 4.3 (on next page) A part of the ‘Run it’ webpage of PHYLO-MOTIF-WEB, where the user can input the Ensembl gene IDs for the reference species and select the regions for motif discovery relative to the TSS (Step 1), add additional information sources (Step 2) and control the parameters for *de novo* motif discovery and the asymmetric clustering approach (Step 3).

Step 1: Select the non-coding regions

Please select the [reference species](#) for which you like to predict regulatory motifs: Human

Enter one or multiple [gene Ensembl IDs](#) from the reference species which you believe share a common regulatory motif:

ENSG00000145386
 ENSG00000115966
 ENSG00000070501
 ENSG00000146469
 ENSG00000170345
 ENSG00000110092


Set the length of the region [upstream](#) of the TSS:


Set the length of the region [downstream](#) of the TSS:


Step 2: Additional information sources


Integrate [evolutionary conservation information](#) by means of an alignment: Yes


To create the multiple alignment: Please select the species in which you believe the motif will be conserved:


 Human

 Rhesus monkey

 Mouse

 Cow

 Dog

 Horse

Integrate [DNA features](#) by using the Ensembl Regulatory Build pipeline: Yes

Across all cell types
 Cell type specific

Perform [RepeatMasking](#): No

Step 3: Motif discovery by using an ensemble strategy

Please select one or more [motif discovery algorithms](#) to run on your input sequences:

MEME [\[1\]](#)
 Phylogibbs [\[2\]](#)
 Phylogenetic sampler [\[3\]](#)

Please set [the range of values](#) for the following parameters by using a comma-separated line:

The number of different motifs:

The motif width:

The number of motif sites per sequence:

Also search for palindromes? Yes

Set the parameter values for the [Asymmetric Clustering Approach](#):

Individual binding site probability:

Consensus score threshold:

Input coverage threshold:

As an example, we ran PHYLO-MOTIF-WEB on a synthetic dataset consisting of 7 human genes, each containing a binding site of the CREB/ATF TF family in their promoter region (Hon and Jain, 2006). A region of 1000 bp up- and 200 bp downstream of the TSS was selected per gene, evolutionary conservation was integrated by means of a human-mouse pairwise alignment and only regulatory regions across all cell types were used (see Figure 4.3) (all information was gathered from Ensembl database version 59, August 2010). *De novo* motif discovery was performed by PG, PS and MEME, each searching for 1 or 2 regulatory motifs, of width 6 or 8 bp, with 0.5 (4 binding sites in 7 sequences) or 1 TF binding site per input sequence. As the consensus sequence recognized by the CREB/ATF family of TFs equals 5'-TGACGTCA-3', we also searched for palindromic motifs. In case the user has no clue on the number of different motifs, their widths and the corresponding number of TF binding sites present in the input sequences, he can run each motif discovery algorithm over multiple values for each of those parameters. However, the amount of different values for each parameter negatively influences the running time of PHYLO-MOTIF-WEB. The set of motif matrices, predicted across all algorithms and parameter values, is reduced to a set of ensemble motif matrices by FuzzyClustering, using the default settings except for the individual binding site probability that was lowered to 0.005, in order to improve sensitivity.

4.3.2 The 'Results' webpage

When clicking on 'submit' the user gets an overview of the chosen parameter values and can proceed to the 'Results' webpage which is presented in Figure 4.4 and Figure 4.5 for the CREB/ATF dataset. Figure 4.4 presents the top of the 'Results' webpage, containing a first table with the genomic coordinates of the selected non-coding regions for each gene. In case the user chose to restrict the non-coding regions to the regulatory regions annotated by Ensembl and/or to use evolutionary information by means of alignment blocks, the genomic coordinates corresponding to those regions are also included in the first table. The second table in Figure 4.4 contains the input files for each participating motif discovery algorithm, and a picture of the phylogenetic tree in case the tree was required by the algorithm. After running each motif discovery algorithm, the predicted motif matrices are resumed in a matrix file and the corresponding TF binding sites in a binding site file, which can be downloaded from the third table.

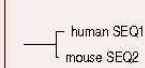
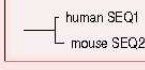
Figure 4.4 (on next page) The first part of the 'Results' webpage. The first table contains the genomic coordinates (chromosome number (Chrom), Start, End and Strand) of the non-coding regions selected for each input gene (identified by a NR and Ensembl ID). In case of regulatory regions and/or alignment information, this table also contains the genomic coordinates that correspond to those regulatory regions (RR) or alignment blocks (AB). For each alignment block, the number of orthologous sequences is given (Orth). Notice that the length of the selected region (Length) also contains the gaps in case of alignment blocks. The second table contains the input files for the motif discovery algorithms and a picture of the phylogenetic tree in case evolutionary conservation was integrated. The third table contains the motif discovery output files: the 'motif matrix files' and the corresponding 'binding site files'.

Results for CREBATF3

The output files of STEP 1 and STEP2: "Select the non-coding regions and add additional information sources"

NR	Ensembl ID	RR	AB	Orth	Chrom	Start	End	Strand	Length
1	ENSG00000169252	--	--	--	5	148205156	148206356	1	1200
2	ENSG00000145386	--	--	--	4	122744887	122746087	-1	1200
3	ENSG00000115966	--	--	--	2	176032910	176034110	-1	1200
4	ENSG00000070501	--	--	--	8	42195009	42196209	1	1200
5	ENSG00000146469	--	--	--	6	153070933	153072133	1	1200
6	ENSG00000170345	--	--	--	14	75744531	75745731	1	1200
7	ENSG00000110092	--	--	--	11	69454873	69456073	1	1200
7	ENSG00000110092	1	1	2	11	69454873	69456073	1	5087
6	ENSG00000170345	1	1	2	14	75744531	75745731	1	1776
5	ENSG00000146469	1	1	2	6	153071791	153072133	1	372
4	ENSG00000070501	1	1	2	8	42195009	42196209	1	4040
3	ENSG00000115966	1	2	1	2	176032910	176033655	-1	1662
3	ENSG00000115966	1	1	1	2	176033656	176033672	-1	17
1	ENSG00000169252	1	1	2	5	148205294	148206356	1	1858
2	ENSG00000145386	1	1	2	4	122744887	122745662	-1	1326

All the files with the genomic coordinates, can easily be downloaded through this link [Non-coding-sequence-coordinates.tar](#)

Algorithm	Input file	Phylogenetic tree
MEME	INPUT 1.fasta	
PHYLOGIBBS	INPUT 1.fasta	
PHYLOGENETIC SAMPLER	INPUT 1.fasta	

The output files of STEP 3: "Motif discovery by using an ensemble strategy"

Algorithm	Motif matrix file	Binding sites file
MEME	MATRICES	Binding sites
PHYLOGIBBS	MATRICES	Binding sites
PHYLOGENETIC SAMPLER	MATRICES	Binding sites

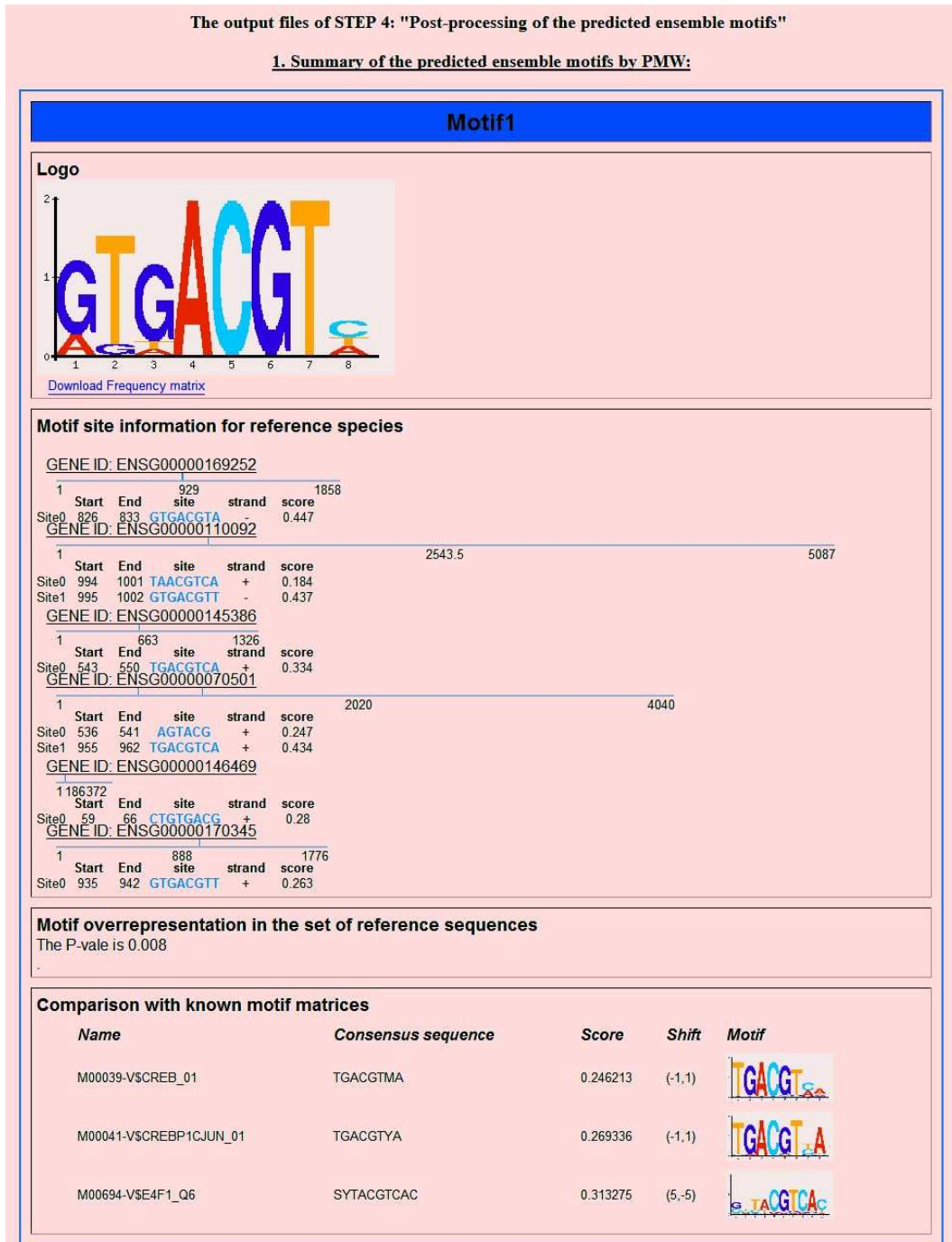


Figure 4.5 The second part of the 'Results' webpage. All ensemble motifs predicted by FuzzyClustering are ranked according to their sequential retrieval. Here we only present the first ensemble motif. Per ensemble motif, the following information is provided: the motif logo and position specific frequency matrix, the location and scores of the TF binding sites in the reference sequences, the p-value for the motif's overrepresentation in the reference sequence set and a comparison with known motif matrices from TRANSFAC and JASPAR.

Figure 4.5 shows the remainder of the ‘Results’ webpage, namely the evaluation of each ensemble motif predicted by FuzzyClustering. For the CREB/ATF dataset, multiple ensemble motifs were predicted, ranked according to their sequential retrieval. For clarity, Figure 4.5 only shows the first and thus most significant ensemble motif. Per ensemble motif the following information can be retrieved on the ‘Results’ page as shown in Figure 4.5:

- The *motif logo* and the possibility to download the corresponding position specific frequency matrix. By looking at the motif logo, we could already see the resemblance with the consensus binding sequence of CREB/ATF.
- The *TF binding sites* that make up the ensemble motif, with their relative positions in the reference input sequences and their membership scores (i.e. a score that represents the relative importance of the binding site compared to the other binding sites that were assigned to the motif) as assigned by the FuzzyClustering algorithm. We checked the overlap between the TF binding sites predicted by PHYLO-MOTIF-WEB and the ‘true’ binding sites of CREB/ATF in the benchmark dataset. In total there were 8 predicted binding sites for this ensemble motif; 6 true positives (i.e. they overlap with the true binding sites) and 2 false positives (no overlap with true sites). This resulted in a very high sensitivity (i.e. 6 of the 7 true sites were retrieved) and positive predictive value (i.e. only 2 of the 8 predicted sites were false positive). Moreover, the 2 false positive predictions had the lowest membership scores (0.184 and 0.247).
- The *p-value for the overrepresentation* of the ensemble motif in the reference sequence set versus random sequence sets. For ensemble motif 1, a p-value of 0.008 indicated that this motif is very specific for the reference sequence set.
- The *comparison* of the ensemble motif matrix with known motif matrices from the TRANSFAC and JASPAR databases. As shown in Figure 4.5, ensemble motif 1 resembled the CREB motif matrix from TRANSFAC with a Kullback-Leibler distance < 0.40 (default).
- A pie chart that shows the *contribution of each motif discovery algorithm* to the ensemble motif (still to be implemented). Here, only motif predictions made by PG contributed to ensemble motif 1, with a prediction rate of 12% (data not shown).
- Per gene, the ‘Results’ webpage also contains a link to view the TF binding sites in their genomic context through the *UCSC Genome Browser* (version hg19/Feb. 2009). An example can be found in Figure 4.6, for the CREB/ATF binding site retrieved in the promoter region of ENSG00000169252 (*ADRB2*).

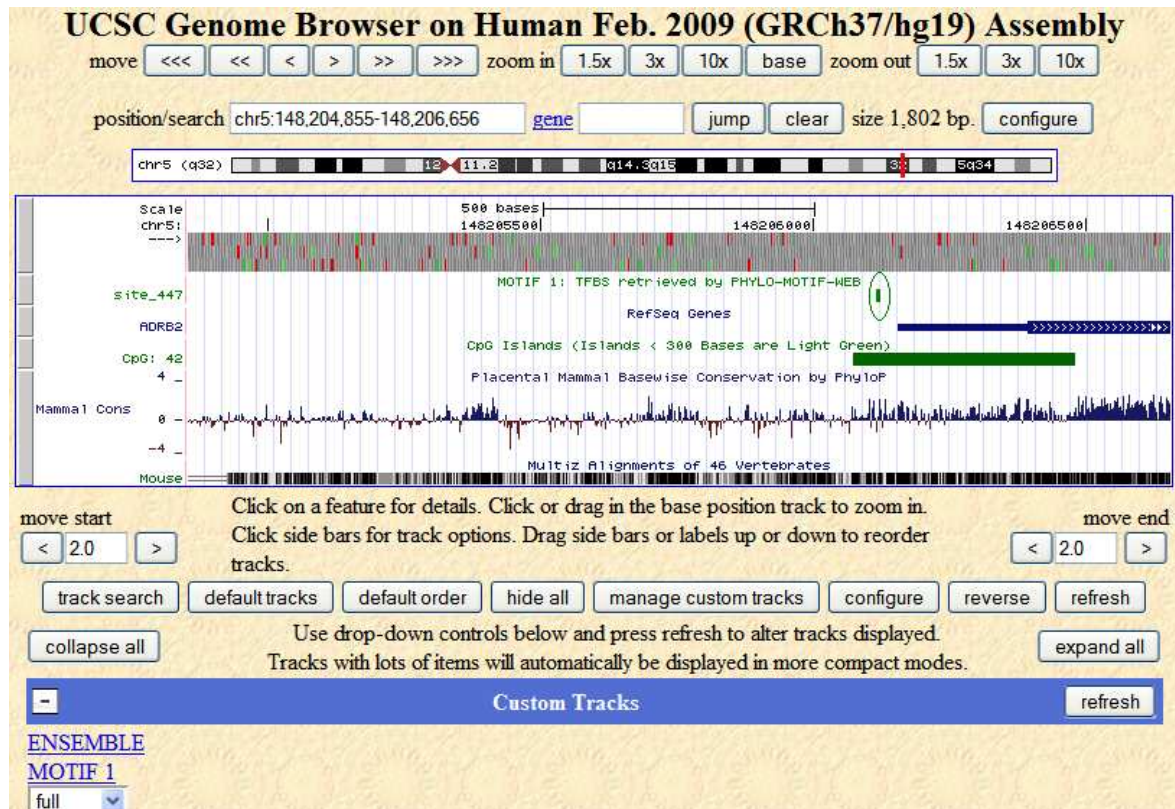


Figure 4.6 Visualization of the CREB/ATF binding site (green circle) in the promoter region of ENSG00000169252 (i.e. the ADRB2 RefSeq gene indicated in blue), through a Custom Track in the UCSC Genome Browser (hg19). Other DNA features can be visualized by activating their tracks, e.g. CpG islands indicated in green.

4.4 Discussion

In this chapter, we present a ‘complete’ *de novo* motif discovery workflow that covers all the different pre- and post-processing steps needed to identify potential motifs in a set of non-coding sequences and this in the form of an easy to use web server: ‘PHYLO-MOTIF-WEB’.

PHYLO-MOTIF-WEB allows for the easy integration of multiple different information sources that were shown to improve motif discovery performance, like evolutionary conservation and DNA features like chromatin structure. For example, PHYLO-MOTIF-WEB provides the multi-species alignments and the corresponding phylogenetic trees as required by motif discovery tools like PG and PS, in order to maximally exploit evolutionary conservation information. This makes PHYLO-MOTIF-WEB more user-friendly compared to other web services like e.g. PhyloScan (Palumbo and Newberg, 2010) that do not provide the ortholog alignment nor the most optimal phylogenetic tree. Also DNA features like nucleosome occupancy and histone modifications can be integrated in the workflow by using the Regulatory Build pipeline of the Ensembl

database that annotates ‘putative regulatory regions’ in the DNA in a cell type specific manner. For algorithms that can not yet integrate this type of information, PHYLO-MOTIF-WEB provides the option to use this information in a discriminative way, to narrow down the sequence search space as much as possible beforehand, not superfluous as the regulation of genes in higher eukaryotes likely depends on a complex interplay between proximal and distal enhancers (Luster and Rizzino, 2003).

Finally PHYLO-MOTIF-WEB applies an ensemble strategy on the results of multiple advanced de novo motif discovery tools. The use of ensemble strategies already proved its benefits (improving both the sensitivity and specificity) over the use of a single component algorithm by exploiting the synergetic prediction capability of multiple algorithms (Chakravarty et al., 2007; Hu et al., 2006; Hu et al., 2005). We presented a clustering-based ensemble strategy, ‘FuzzyClustering’ that groups sets of TF binding sites, jointly predicted during multiple runs of one or more component algorithms, into one (or more) higher quality motifs. As FuzzyClustering works on the level of the TF binding sites, it is able to filter out false positive TF binding sites and pick up weak TF binding sites that would get lost in a motif-level clustering approach.

Besides the contribution level of each component algorithm, also the overrepresentation of the predicted motifs in the input set compared to random sequence sets and the comparison with known database motifs helps the user to assess the statistical and biological significance of the predicted ensemble motifs.

Chapter 5 *De novo* motif discovery in vitamin D₃ regulated genes

5.1 Introduction

Vitamin D₃ (cholecalciferol) can be derived from nutrition, but the main supply of vitamin D₃ derives from production in the skin after exposure to ultraviolet light from the sun, which converts 7-dehydrocholesterol to form previtamin D₃, which is rapidly converted to vitamin D₃ (Holick, 2004) (see Figure 5.1). Therefore, vitamin D₃ is not considered as a true vitamin but rather as a precursor of a hormone. Whether it is made in the skin or ingested, vitamin D₃ is first hydroxylated in the liver to 25-hydroxyvitamin D₃, and then in the kidneys to its biologically active form, 1 α ,25-dihydroxyvitamin D₃ or 1 α ,25(OH)₂D₃ (Holick, 2004). As shown in Figure 5.1, the active metabolite of vitamin D₃ plays an important role in the regulation of numerous physiological and cellular processes including calcium/phosphate homeostasis, bone maintenance, cell growth, cell differentiation, and the immune system (Bouillon et al., 2008; Lin and White, 2004; Gurlek et al., 2002; Jones et al., 1998). The most widely accepted role of 1 α ,25(OH)₂D₃ is the regulation of calcium and phosphate metabolism as it is important for the absorption of these essential minerals in the intestine, and for their mobilization in bone tissues (Rachez and Freedman, 2000). Besides this ‘classical’ effect of vitamin D₃, the hormone is also a modulator of the immune response (Bouillon et al., 2008) and has a potent growth-inhibitory or antiproliferative and prodifferentiating action on different cell types including malignant cancer cells (Deeb et al., 2007).

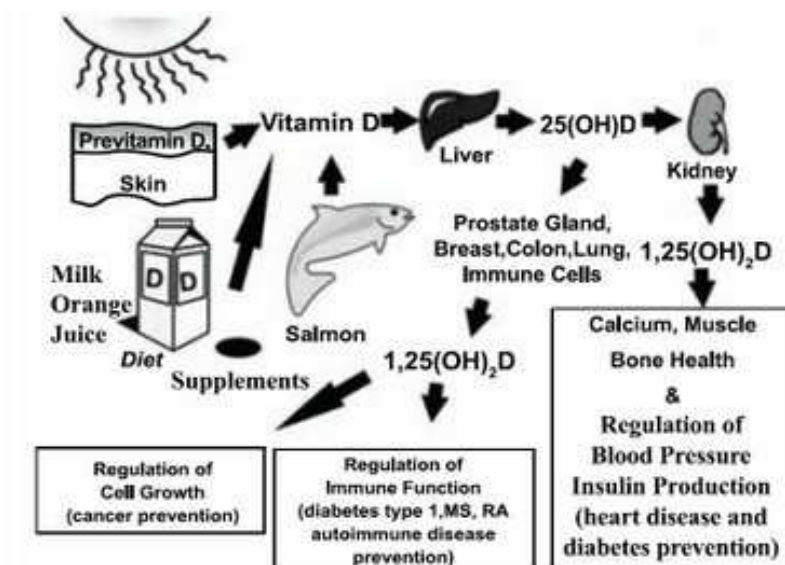


Figure 5.1 Metabolism of vitamin D₃ to 1 α ,25(OH)₂D₃ in the kidney and other organs and the biological consequences. Taken from (Holick, 2004).

The wide range of therapeutic and health-related benefits of vitamin D₃ relates to the fact that the vitamin D receptor (VDR), through which 1 α ,25(OH)₂D₃ exerts its genomic effects, is present in most cells and tissues in the body (Zehnder et al., 2001). This vitamin D receptor belongs to the class of nuclear receptors which are highly conserved throughout evolution and regulate gene expression in a ligand-dependent manner (Haussler et al., 1998). Upon 1 α ,25(OH)₂D₃ binding, the VDR acts as a ligand-activated TF and translocates to the nucleus. Before the VDR is able to influence transcription it heterodimerizes with the retinoid X receptor (RXR). This VDR/RXR complex binds chromatin rapidly at regulatory regions called vitamin D responsive elements (VDREs). These elements are generally composed of two hexameric binding sites interspaced by a varying number of nucleotides (Haussler et al., 1998) (more details follow in § 5.2.2.1). Many genes regulated by vitamin D₃ have multiple VDREs in their promoter region, sometimes far away from the coding region, e.g. in distal enhancer regions.

Modulation of gene expression by ligand-activated VDR/RXR bound to the DNA is mediated through the recruitment of co-activator protein complexes. These co-activating proteins, like the members of the SRC/p160 and the CBP/p300 protein families, may induce histone acetylation, which results in an open chromatin structure, creating a chromatin surrounding permissive for gene transcription (Rachez and Freedman, 2000). Subsequently, general TFs (GTFs) as well as RNA polymerase II are recruited to the TSS by VDR/RXR-interacting multi-protein complexes like DRIP (vitamin D Receptor Interacting Proteins) (Rachez and Freedman, 2000). This will induce transcription of the neighbouring gene. Gene transcription can also be affected by ATP-dependent chromatin remodeling complexes, like SWI/SNF, that interact with VDR and mediate a locally low nucleosome occupancy by histone displacement (Li et al., 2007). Unliganded VDR is kept transcriptionally silent, even when present in the nucleus and bound to chromatin, by one or more co-repressors like SMRT (silent mediator for retinoid and thyroid hormone receptors) and NCoR (nuclear receptor co-repressors). Those co-repressors are able to deacetylate (directly or indirectly) histones and thus keep the chromatin in a densely packed configuration which is inaccessible for the transcriptional machinery (Tagami et al., 1998).

The antiproliferative effects of 1 α ,25(OH)₂D₃, in combination with the presence of VDR in a wide variety of cell types opens perspectives for the use of this molecule in the treatment of cancer and other hyperproliferative disorders, however its calcemic side effects hamper its therapeutic applications. Therefore several research groups designed synthetic analogs of 1 α ,25(OH)₂D₃ with an improved antiproliferative/prodifferentiating action and lower calcemic effects (Eelen et al., 2007; Guyton et al., 2003; Verlinden et al., 2000; Bouillon et al., 1995). However, the exact nature of the 1 α ,25(OH)₂D₃ mediated signaling cascade that relates to the antiproliferative effects remains not

completely understood. A transcriptome study by Vanoirbeek *et al.* (Vanoirbeek *et al.*, 2009) revealed that a superagonistic vitamin D₃ analog (WY1112) induces the same set of genes as 1 α ,25(OH)₂D₃, but the level of induction of the individual genes is higher. A better understanding of the 1 α ,25(OH)₂D₃ signaling cascade will allow identifying regulated key target genes that could be used as important markers for the activity profile of newly developed analogs.

The setup of this study is based on a very interesting observation; the antiproliferative phenotype induced by vitamin D₃ treatment was observed for both human (*Homo sapiens*) breast cancer and mouse (*Mus musculus*) osteoblastic cell lines. Therefore, we performed a comparative transcriptome analysis to gain better insights in the molecular mechanism underlying the antiproliferative effects of vitamin D₃. In experiments limited to a single species, it is often hard to filter false positives from a set of predicted differentially expressed genes that have potential to be physiologically important. However, evolutionary conservation is a powerful criterion to identify genes that are functionally important from a set of coexpressed genes. The availability of transcriptome data for both human and mouse, allowed us to focus on genes that showed a conserved coexpression behaviour, across both species, which could be of particular interest for the common antiproliferative phenotype. The next challenge was to elucidate the transcriptional regulation underlying the conserved coexpression behaviour. Although not always the case (Ludwig *et al.*, 2000), we assumed that also the transcriptional regulation had been conserved and searched for conserved *cis*-regulatory elements that govern the coexpression of the genes. The results of the microarray analysis are described in section § 5.2.1 and the identification of *cis*-regulatory elements for one specific, well conserved coexpression cluster, are described in section § 5.2.2.

5.2 Results

5.2.1 Microarray analysis

In order to gain better insights in the molecular mechanism underlying the antiproliferative effects of vitamin D₃, a comparative transcriptome analysis for human and mouse was performed. The Legendo research group performed microarray time-series experiments, for both human and mouse, on vitamin D₃ treated cells versus control cells. For mouse, a cDNA platform was used to examine the expression profile of 21492 genes in osteoblastic MC3T3-E1 cells (Verlinden *et al.*, 2005), while for human, the Affymetrix platform was used, to measure the expression profile of all human coding genes in MCF-7 breast cancer cells (Vanoirbeek *et al.*, 2009). Mouse bone cells are a very classic target of VDR action in contrast to the human breast cancer cells. However, the Legendo research group chose to work with human breast cancer cells as they express

VDR and they show clear growth inhibition after vitamin D₃ treatment. Also their experience on manipulating this type of cells supported their decision. The different microarray platforms and tissues, used for both species, make this comparative transcriptome analysis less straightforward; differences in gene expression observed between both species can be due to species-specific effects, tissue effects or platform effects. Therefore we will focus on the similarities in gene expression between both species. Those can maybe relate to the common phenotype observed after vitamin D₃ treatment. A statistical analysis, to derive, for each species, the differentially expressed genes, followed by a clustering and comparative gene expression analysis, revealed one interesting cluster of coexpressed genes which is conserved across human and mouse (for all details see Materials and Methods). This conserved cluster consists of 10 genes, upregulated in time after vitamin D₃ treatment compared to no treatment (control state), as shown in Figure 5.2.

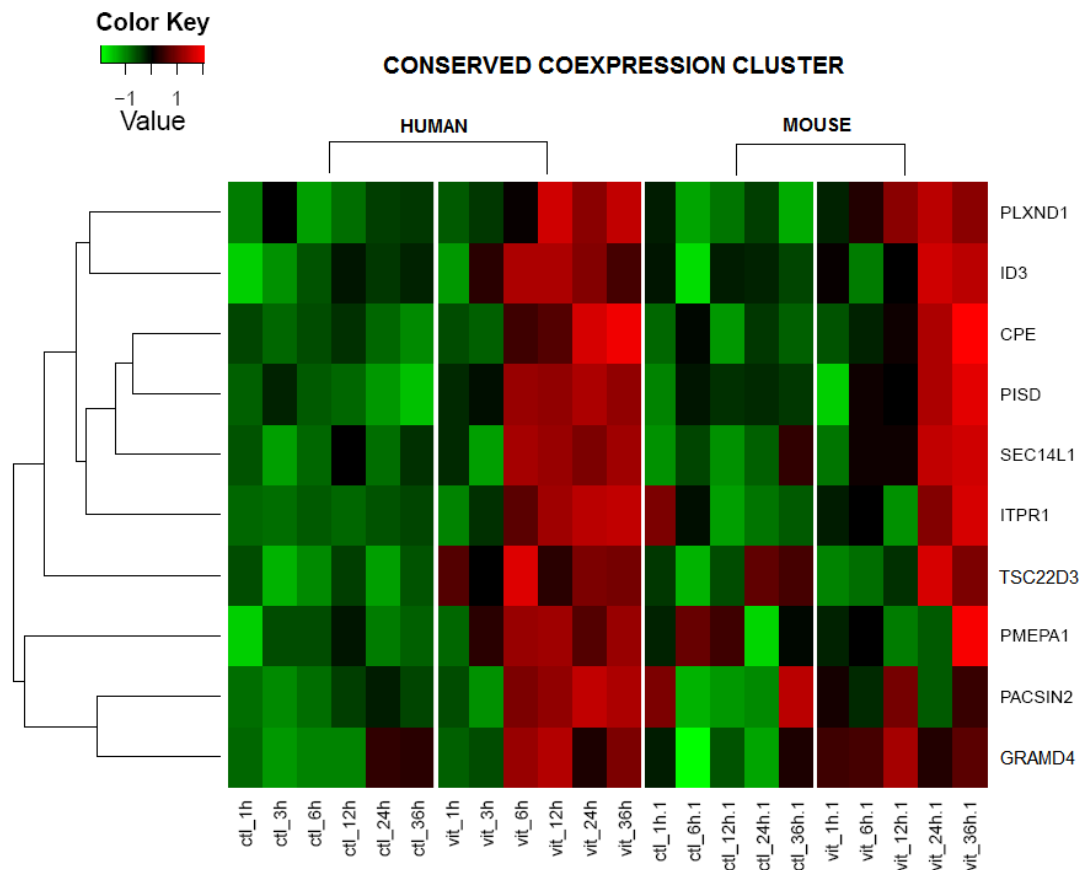


Figure 5.2 Expression profiles for the conserved coexpression cluster which consists of 10 genes (see Table 5.1). The expression profiles for 10 genes (rows) in non-treated cells (ctl) and in vitamin D₃ treated cells (vit) for different time points (1h, 3h, 6h, 12h, 24h, 36h) (columns), in both human (left) and mouse (right). Notice, for mouse the gene expression for the 3h time point is not measured. The first (ctr) and second (vit) block of columns represent the expression values for human; the third (ctr) and fourth (vit) block of columns represent the expression values for mouse.

To get a first indication, on how those 10 genes fit in the antiproliferative effect of vitamin D₃, we considered their individual functions. We looked up the GO terms concerning the biological processes in which the 10 genes are involved, for both human and mouse, in the amiGO database version 1.7 (Ashburner et al., 2000). Table 5.1 enlists for each gene, the gene name and a short description as well as the GO terms for the biological processes that are linked to this gene.

Table 5.1 The genes identified as being upregulated in both human and mouse after cell treatment with vitamin D₃. For each gene belonging to the conserved coexpression cluster (see Figure 5.2), a number (Nr) and the gene name (Name) are given, followed by a short description (Description) and a list of all the associated GO terms concerning biological processes (GO biological process).

Nr	Name	Description	GO biological process
1	<i>PLXND1</i>	Plexin-D1	Multicellular organismal development Signal transduction Patterning of blood vessels
2	<i>ID3</i>	DNA-binding protein inhibitor ID-3	Multicellular organismal development Negative regulation of transcription Regulation of transcription Regulation of DNA replication Positive regulation of apoptosis Epithelial cell differentiation
3	<i>CPE</i>	Carboxypeptidase E	Insulin processing Proteolysis Protein modification process
4	<i>PISD</i>	Phosphatidylserine decarboxylase proenzyme	Phospholipid biosynthetic process
5	<i>SEC14L1</i>	SEC14-like protein 1	Transport
6	<i>ITPR1</i>	Inositol 1,4,5-trisphosphate receptor type 1	Calcium ion (transmembrane) transport and homeostasis Response to hypoxia Cell death Signal transduction
7	<i>TSC22D3</i>	TSC22 domain family protein 3	Regulation of transcription, DNA-dependent Anti-apoptosis Response to osmotic stress
8	<i>PMEPA1</i>	Prostate transmembrane protein, androgen induced 1	Androgen receptor signaling pathway
9	<i>PACSIN2</i>	Protein kinase C and casein kinase substrate in neurons protein 2	Endocytosis Signal transduction Actin cytoskeleton organization
10	<i>GRAMD4</i>	GRAM domain-containing protein 4 (Death-inducing protein)	Apoptosis

Although some genes are active in the same processes, no strong tendency to a specific biological pathway is revealed. Based on the GO annotations, it is not possible to identify one or several biological processes that are strongly represented in this conserved coexpression cluster. Despite this, two biological processes could be directly related to an antiproliferative phenotype i.e. cell death and apoptosis, in which *ID3*, *ITPR1* and *GRAMD4* are involved.

Further on, we describe other biological processes exerted by more than one gene of the conserved coexpression cluster. The very general process of multicellular organismal development, which covers any biological process of which the specific outcome is the progression of a multicellular organism over time, was assigned to *PLXND1* and *ID3*. Three genes (*PLXND1*, *ITPR1* and *PACSIN2*) were associated with signal transduction, the process whereby a signal is converted into a form where it can ultimately trigger a change in the state or activity of a cell. *SECI4L1* is involved in transport, a very general process by which substances such as macromolecules, small molecules or ions are moved into, out of or within a cell, or between cells. More specific is the process of calcium ion transport assigned to *ITPR1* and endocytosis, a vesicle-mediated transport process in which cells take up external materials, assigned to *PACSIN2*. *ID3* and *TSC22D3* are both involved in regulation of transcription (~ the process that modulates the frequency, rate or extent of the synthesis of RNA on a template of DNA).

Then, we have some biological processes, specific for one gene of the coexpression cluster, like patterning of blood vessels (i.e. the process that regulates the coordinated growth and sprouting of blood vessels) (*PLXND1*); epithelial cell differentiation (i.e. the process whereby a relatively unspecialized cell acquires specialized features of an epithelial cell) (*ID3*); insulin processing (i.e. the formation of mature insulin by proteolysis of the precursor preproinsulin), protein modification and proteolysis (CPE); phospholipid biosynthetic process (i.e. the chemical reactions and pathways resulting in the formation of phospholipids) (*PISD*); response to hypoxia (i.e. lowered oxygen tension) (*ITPR1*); response to osmotic stress (i.e. an increase or decrease in the concentration of solutes outside the organism or cell) (*TSC22D3*); Androgen receptor signaling pathway, any series of molecular signals generated as a consequence of an androgen binding to its receptor (*PMEPA1*); actin cytoskeleton organization (*PACSIN2*). As explained further in the text, we will add an extra gene to the analysis, *PDLIM2*, which is an adaptor protein located at the actin cytoskeleton that promotes cell attachment, and is necessary for the migration capacity of cells.

5.2.2 Identification of *cis*-regulatory elements

To unravel *cis*-regulatory elements that possibly play a role in the transcriptional regulation of the coexpressed gene set in both human and mouse, we apply a *de novo* motif discovery approach followed by a motif scanning approach. The purpose of *de novo* motif discovery is to detect novel motifs, i.e. TF binding sites for a common regulator shared between multiple genes. For many eukaryotic TFs the regulatory motif is already known and stored in databases like TRANSFAC (Matys et al., 2006) and JASPAR (Bryne et al., 2008). This allows comparing *de novo* predicted motifs with database motifs which can provide insight in the TFs potentially involved in the

transcriptional regulation of the coexpressed gene set. In a second step, we will analyse if the binding sites of the *de novo* predicted motifs co-locate in the DNA, to form *cis*-regulatory modules (CRMs). CRMs are common in higher eukaryotes like human and mouse, due to the more complex combinatorial interactions between TFs which allow for biochemical specificity of transcription. The detection of CRMs can produce predictions of substantially better specificity than the analysis of isolated binding sites.

For *de novo* motif discovery we applied PHYLO-MOTIF-WEB which was described in chapter 4. PHYLO-MOTIF-WEB combines the results of multiple *de novo* motif discovery algorithms and allows the use of additional information sources like ‘evolutionary conservation’ and ‘regulatory potential’ as annotated by the Regulatory Build pipeline of the Ensembl database. The collection of *de novo* predicted motifs were then used by CPMModule (Guns et al., 2010) to detect CRMs statistically overrepresented in the gene set compared to genomic background sequences. CPMModule stands for ‘*cis*-regulatory module detection by constraint programming’ which combines the principles of itemset mining and constraint programming. The results of PHYLO-MOTIF-WEB for *de novo* motif discovery are presented in §5.2.2.1 and the results of CPMModule in §5.2.2.2.

5.2.2.1 *De novo* motif discovery

Per input setting, PHYLO-MOTIF-WEB runs three different algorithms for motif discovery: Phylogibbs (PG) (Siddharthan et al., 2005), Phylogenetic sampler (PS) (Newberg et al., 2007) and MEME (Bailey and Elkan, 1994). This is followed by an asymmetric clustering approach (FuzzyClustering) to summarize all predictions made by PG, PS and MEME into a final set of ensemble motif matrices. The parameter settings for each of the motif discovery algorithms and the ensemble strategy are described in Material and Methods. As we search for conserved *cis*-regulatory elements, we will integrate evolutionary conservation information. Both PG and PS were specifically developed to maximally exploit evolutionary conservation information. Provided with a multiple alignment of orthologous sequences, they both search for TF binding sites that evolved according to a tree-based evolutionary model (see chapter 2). The user has to provide PHYLO-MOTIF-WEB with a set of Ensembl gene IDs from a reference species (i.e. the species for which the user likes to infer the regulatory motifs) and PHYLO-MOTIF-WEB will align the corresponding non-coding sequences with their orthologous sequences selected from a set of user chosen species.

For this study, we provided PHYLO-MOTIF-WEB with the Ensembl gene IDs corresponding to the genes from the conserved coexpression cluster (see Figure 5.2). As shown in Table 5.2, we ran PHYLO-MOTIF-WEB on four different input settings (A, B, C and D), related to the selected reference species, the number of species included in the alignment (~alignment type) and the number of input genes.

Table 5.2 The four different input settings for the PHYLO-MOTIF-WEB workflow.

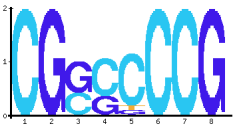
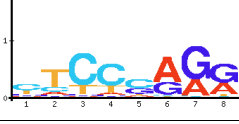
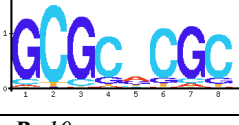
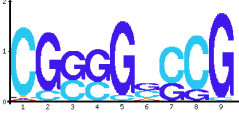
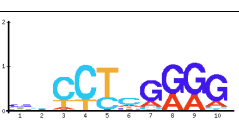
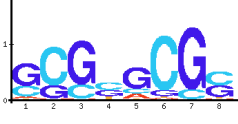
Setting	Reference species	Alignment type	Input genes
A	human	Pairwise human-mouse	10
B	mouse	Pairwise human-mouse	10
C	human	Six-species alignment	10
D	human	Six-species alignment	11


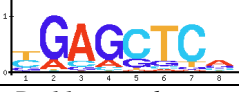
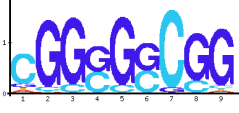
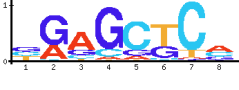
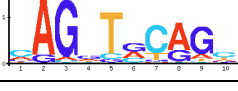
The reference species was chosen to be either human (setting A, C and D) or mouse (setting B) as shown in Table 5.2. If the reference species equals human, the human Ensembl gene IDs are used as input and the motif matrices predicted by PHYLO-MOTIF-WEB are human-specific (i.e. only based on TF binding sites predicted for the human input sequences and not on binding sites predicted for the orthologous sequences). When mouse was chosen as the reference, the same holds, but now mouse Ensembl IDs are used as input and the obtained motifs are mouse-specific. For each input gene, we selected 4000 bp centered on the TSS (transcription start site). This region was further restricted to its ‘regulatory regions’ as annotated by the Regulatory Build Pipeline of Ensembl (see Material and Methods). Reducing the input will limit the motif search space to those regions with high potential of containing ‘functional’ TF binding sites. Effectively reducing the search space reduces the risk of finding false positive predictions but comes at the expense of a lower sensitivity i.e. TF binding sites lying outside those regulatory regions will be missed. As mentioned on top, we will search for conserved *cis*-regulatory elements by incorporating ‘evolutionary conservation’ by means of an alignment. For input settings A and B we used a pairwise human-mouse alignment, for settings C and D a six-species alignment (see Material and Methods). For setting D we added an extra gene, *PDLIM2* (PDZ and LIM domain 2), also referred to as *mystique*. This gene was not retrieved by the microanalysis as explained in the Material and Methods section, but it is of particular interest for the Legendo research group as it is involved in cancer metastasis. On top, the Legendo research group could experimentally verify a binding site for VDR upstream of *PDLIM2*’s TSS.

Table 5.3 summarizes, per input setting (A, B, C and D), all the predicted ensemble motifs together with following characteristics; their motif logo, the consensus score of their motif matrix, the number of TF binding sites predicted to be present in the sequences of the reference species, the number of target genes of the reference species that contain a TF binding site, the names of the target genes and a p-value for the overrepresentation of the ensemble motif in the reference sequence set compared to

random sequence sets. The two last characteristics mentioned in Table 5.3 are derived from the Ensemble strategy (FuzzyClustering) that clusters TF binding sites which were jointly retrieved in a significant number of motifs predicted by PG, PS and MEME, into ensemble motifs. Per ensemble motif, Table 5.3 mentions the percentage of predicted motif matrices that contributed to the ensemble motif (i.e. the prediction rate) and the motif discovery algorithms by which those contributing motifs were predicted (see also Material and Methods).

Table 5.3 Ensemble motifs predicted by PHYLO-MOTIF-WEB, for the four different input settings (A, B, C and D). Each ensemble motif is indicated by its ID (the setting number followed by a serial number). Then, the motif logo (Logo) is shown, the consensus score (Cs), the number of TF binding sites predicted for the sequences of the reference species (TFBSs), the number of target genes (TGs), the target gene names (Target gene names), a p-value for the overrepresentation of the ensemble motif (P-value), the motif discovery algorithms with a contribution level different from zero (Tools) and the prediction rate (PR).

ID	Logo	Cs	TFBSs	TGs	Target gene names	P-value	Tools	PR
Setting A: 10 genes, human=reference, pair-wise alignment with mouse								
A1		1.65	12	5	<i>PLXND1, ID3, SEC14L1, PMEPA1, GRAMD4</i>	0.686	MEME	9%
A2		0.59	28	9	<i>PLXND1, ID3, PISD, SEC14L1, ITPR1, TSC22D3, PMEPA1, PACSIN2, GRAMD4</i>	0.223	PS	18%
A3		1.08	44	9	<i>PLXND1, ID3, PISD, SEC14L1, ITPR1, TSC22D3, PMEPA1, PACSIN2, GRAMD4</i>	0.61	PG MEME	14%
Setting B: 10 genes, mouse=reference, pair-wise alignment with human								
B1		1.23	25	8	<i>PLXND1, PISD, SEC14L1, ITPR1, TSC22D3, PMEPA1, PACSIN2, GRAMD4</i>	0.061	MEME	7%
B2		0.71	31	6	<i>PLXND1, PISD, SEC14L1, TSC22D3, PMEPA1, PACSIN2</i>	0.551	PS	11%
B3		0.84	77	10	<i>PLXND1, ID3, CPE, PISD, SEC14L1, ITPR1, TSC22D3, PMEPA1, PACSIN2, GRAMD4</i>	0.058	PG	23%

ID	Logo	Cs	TFBSs	TGs	Target gene names	P-value	Tools	PR
Setting C: 10 genes, human=reference, six-species alignment								
C1		1.30	82	9	<i>PLXND1, ID3, PISD, SEC14L1, ITPR1, TSC22D3, PMEPA1, PACSIN2, GRAMD4</i>	0.72	MEME	8%
C2		0.74	50	8	<i>PLXND1, ID3, PISD, SEC14L1, ITPR1, PMEPA1, PACSIN2, GRAMD4</i>	0.02	PG PS	20%
Setting D: 11 genes, human=reference, six-species alignment								
D1		1.27	85	9	<i>PLXND1, ID3, PISD, SEC14L1, TSC22D3, PMEPA1, PACSIN2, GRAMD4, PDLIM2</i>	0.82	MEME	13%
D2		0.67	39	10	<i>PLXND1, ID3, PISD, SEC14L1, ITPR1, TSC22D3, PMEPA1, PACSIN2, GRAMD4, PDLIM2</i>	0.24	PG	10%
D3		0.77	17	7	<i>PLXND1, ID3, PISD, SEC14L1, ITPR1, TSC22D3, PDLIM2</i>	0.09	PS	10%

Motifs A1, B1, C1 and D1

Across all four input settings, PHYLO-MOTIF-WEB predicted a similar GC-rich motif matrix represented respectively by the motifs A1, B1, C1 and D1, referring to the first predicted ensemble motif for each setting. Motifs A1, C1 and D1 are human-specific while motif B1 models TF binding sites for mouse sequences. In general, over all input settings, only motif matrices predicted by MEME contributed to this GC-rich ensemble motif and the prediction rate ranged between 7% for motif B1 and 13% for motif D1. We calculated for each ensemble motif the p-value for its overrepresentation in the input sequence set compared to a set of random sequence sets of the same size. Except for the mouse-specific motif B1 (p-value = 0.061), the p-values for the human-specific motifs were quite high; 0.686 (motif A1), 0.72 (motif C1) and 0.82 (motif D1). So, especially in case of human, those GC-rich ensemble motifs are not very specific for the upregulated gene set and probably correspond to a very general/common regulatory motif that occurs frequently in the promoter regions of human genes. This is also reflected by the number of target genes in our input set; all genes except for *CPE* contain at least one binding site for this GC-rich motif across the four input settings.

When we compared the four ensemble motif matrices A1, B1, C1 and D1 with motif matrices in the TRANSFAC and JASPAR databases (see Table 5.4), they matched with the motif matrices for the following TFs: SP1 (all four ensemble motifs), EGR-1 (A1, C1 and D1), LRF (motif B1 and C1) and TEAD2 (motif D1). SP1 appears the most significant match with p-values ranging between 0.00004 for motif A1 and 0.00059 for motif B1.

Table 5.4 Comparison between predicted and known motif matrices. The columns are: the ID of the predicted motif matrix according to Table 5.3 (ID), the names of the TFs that correspond to regulatory motif matrices that match the predicted motif matrix (TF), the consensus sequence for the matching regulatory motifs (Consensus sequence), the p-value for the comparison between the predicted motif matrix and the known motif matrices (p-value) and the number of nucleotides that overlap between the predicted and the known motif matrix (Overlap). P-values and overlap values are given for the optimal trade-off and orientation between both matrices. For degenerated nucleotide symbols, see IUPAC code in Table S1 in the Supplementary Materials.

ID	TF	Consensus sequence	p-value	Overlap
A1	SP1	GCCCCGCCCC	0.00004	6
	EGR-1	CCCCCCCCrCCCC	0.00014	8
B1	SP1	nnGGGGCGGGGnn	0.00044	7
	LRF	GGGGkynnb	0.00406	7
C1	SP1	GGGGCGGGGC	0.00059	6
	EGR-1	GGGGyGGGGCGGG	0.00094	8
	LRF	GGGGkynnb	0.00148	7
D1	EGR-1	GGGGyGGGGCGGG	1.8e-05	9
	TEAD2	GvGGmGG	0.00010	7
	SP1	GGGGCGGGGC	0.00019	7

SP1 (specificity protein 1) is known to act as a master regulator of TFs among which several cell cycle regulators. The regulatory motif for SP1 was also retrieved by Prakash and Tompa (Prakash and Tompa, 2005), who performed a genome-wide discovery of regulatory elements in vertebrates through comparative genomics. Also here the SP1 motif was predicted for a large number of genes and well conserved across vertebrate species. Interesting in the light of this vitamin D₃ study, is the study of Huang *et al.*, (Huang *et al.*, 2004) which confirms a physical interaction between the SP1 and the VDR proteins, required for the induction of $p27^{Kip1}$ expression after vitamin D₃ supply. Their results suggest that VDR is involved in the induction of $p27^{Kip1}$, by interacting with SP1 to modulate the expression of $p27^{Kip1}$ that lacks a VDR response element (VDRE) in its promoter region. EGR-1 (Early growth response 1) belongs to the EGR family of zinc-finger proteins. It is a nuclear protein and functions as a transcriptional regulator. The products of target genes it activates are required for differentiation and mitogenesis. Studies on human breast cancer cells suggest that EGR-1 is a tumor suppressor gene (Liu *et al.*, 2007). LRF (also known as ZBTB7A or Pokemon) is involved in cell cycle progression and was described as a transcriptional repressor involved in oncogenesis (Maeda *et al.*, 2007; Maeda *et al.*, 2005).

TEAD2 (TEA domain family member 2) or ETF, is a TF which plays a key role in the Hippo signaling pathway, a pathway involved in organ size control and tumor suppression by restricting proliferation and promoting apoptosis (Zhao et al., 2008).

Motifs A3 and B3

Motifs A3 (human-specific) and B3 (mouse-specific) are two very similar GC-rich ensemble motifs predicted by PHYLO-MOTIF-WEB. In contrast to the SP1-like ensemble motifs, these two motifs were retrieved only for input settings A and B. When integrating more species in the alignment (as was done for input settings C and D), this motif was not longer retrieved by PHYLO-MOTIF-WEB. A reason can be that the TF binding sites are not well conserved across all six species in the alignment and thus harder to discover. As shown in the results Table 5.3, motif A3 has a prediction rate of 14% (i.e. the fraction of predicted motif matrices that contributed to the ensemble solution). Those predictions were made by both MEME and PG, which adds extra confidence to this motif. Motif B3, the mouse-specific motif, has a high prediction rate of 23% and was based on motif matrices predicted by PG. Similar as for the SP1-like ensemble motifs, the p-value for the overrepresentation of motif A1 (p-value = 0.61) in the human sequence set is much higher than for motif B1 (p-value = 0.058) in the mouse sequence set. A reason for this returning phenomenon is maybe the smaller GC-content of the mouse background sequence set (46.9%) compared to the human one (50.5%). Many binding sites were retrieved for both motifs, with on average 4.4 (motif A3) and 7.7 (motif B3) binding sites per target gene. As shown in Table 5.5, both ensemble motif matrices match with the regulatory motif matrix of the ZFP161 protein with p-values equal to 0.00056 (motif A3) and 0.000079 (motif B3). Motif A3 also matches the regulatory motif matrix of the NRF1 protein (p-value = 0.00020).

Table 5.5 Comparison between predicted and known motif matrices. For column information see Table 5.4.

ID	TF	Consensus sequence	p-value	Overlap
A3	NRF1	yGCGCATGCG	0.00020	8
	ZFP161	yCGCGCsC	0.00056	6
B3	ZFP161	GsGCGCGr	7.9e-05	6

ZFP161 (zinc finger protein homologous to ZFP161 in mouse) or ZF5 is the transcriptional repressor of c-myc (~oncogene) and appears to be expressed ubiquitously, but its message seems particularly abundant in certain differentiated tissues with little mitotic activity. Furthermore, it has growth-suppressive activity when overexpressed in mouse. Therefore, ZFP161 was proposed to have a function in arresting cell division and maintaining a differentiated state (Numoto et al., 1995). NRF1 (nuclear respiratory factor 1) activates the expression of some key metabolic genes regulating cellular growth and nuclear genes required for respiration and mitochondrial DNA transcription and replication.

NRF1 is also a strong biological and positional candidate to contribute to type 2 diabetes susceptibility (Liu et al., 2008). Elkon *et al.* (Elkon et al., 2003) found that NRF1 co-occurred significantly with both SP1 and TEAD2 (see motifs A1, B1) in promoters of genes expressed in a cell cycle dependent manner.

Motifs A2 and B2

The following two ensemble motifs are motif A2 (human-specific) and motif B2 (mouse-specific). Similar to motifs A3 and B3, they were only retrieved for the human-mouse alignment and lost in case of the six-species alignment (input settings C and D). Motif A2 has the highest prediction rate (18%) of all ensemble motifs in setting A. The contributing matrices were all predicted by PS. Motif B2 has a prediction rate of 11% and also here, PS was the only contributing algorithm. The p-values for the overrepresentation equalled 0.223 in case of motif A2 (human sequence set) and 0.551 in case of motif B2 (mouse sequence set). Motif A2 has binding sites in almost every human input gene (except for the *CPE* gene), while binding sites for motif B2 were only retrieved in six out of ten mouse input genes as shown in Table 5.3. When we compare both motif matrices (A2 and B2) to known matrices from TRANSFAC and JASPAR we get diverse matching profiles as shown in Table 5.6. The two common matches are the regulatory motif matrix of OLF1 (p-value = 0.00331 for A2 and p-value = 0.00131 for B2) and the regulatory motif matrix of EBF (p-value = 0.0011 for A2 and p-value = 0.00676 for B2).

Table 5.6 Comparison between predicted and known motif matrices. For column information see Table 5.4.

ID	TF	Consensus sequence	p-value	Overlap
A2	LYF1 (Ikaros)	yCTCCCAAA	0.00091	8
	EBF	TCyCwrGGGAm	0.00110	7
	OLF1	nCmnyvTCyCTrGGGAvThGnn	0.00331	8
	STAT	TCCmAGAAnnnnn	0.00405	7
B2	DEAF1	nCGnnyTCGGGnrTTCCGdArnnn	4.5e-05	10
	AP2ALPHA	sCynnnGGC	0.00123	8
	OLF1	nCmnyvTCyCTrGGGAvThGnn	0.00131	10
	AP2GAMMA	sCCnrGGC	0.00162	8
	EBF	TCyCwrGGGAm	0.00676	9

EBF (early B-cell factor) is a protein family that consists of four members: EBF1 till EBF4, with EBF1 (synonymous to OLF1) a regulator of B-cell differentiation. Further, motif A2 matches with the motif matrices of LYF1 (i.e. a transcriptional regulator in the development of lymphocytes, B-cells and T-cells) and STAT. Motif B2 matches with the motif matrices of DEAF1 and the AP2 family of TFs. Notice that the motif matrices of OLF1 and DEAF1 are only partially covered by the predicted motif matrices (~ the length of the predicted motifs A2 and B2 is much shorter).

Motif D3

Motif D3 (human-specific) is a special motif as it was only retrieved for input setting D that includes *PDLIM2*. The motif has a prediction rate of 10% and the predicted matrices that contributed were all retrieved by PS. The p-value for its overrepresentation in the human sequence set, compared to random sequence sets equals 0.09. The total number of TF binding sites, divided over its 7 target genes, equals 17 of which 5 binding sites are located in the promoter region of *PDLIM2*. As shown in Table 5.7, motif matrix D3 matches significantly with the motif matrices of ZEB1 and LMO2.

Table 5.7 Comparison between predicted and known motif matrices. For column information see Table 5.4.

NR	TF	Consensus sequence	p-value	Overlap
D3	ZEB1	GnmCAGGTGynb	0.00030	9
	LMO2	CnnCAGGTGbnn	0.00077	9

ZEB1 (zinc finger E-box binding homeobox 1, also called deltaEF1 or AREB6) may modulate the levels of *VDR* expression during differentiation in embryonal development, as well as in cancer cells (Lazarova et al., 2001). ZEB1 is also a critical inducer of epithelial to mesenchymal transitions (EMT) in some cell types during development and cancer metastasis, which suggests that ZEB1 is a key player in late stage carcinogenesis (Eger et al., 2005). The LMO2 (LIM domain only 2) protein has a central and crucial role in hematopoietic development and is highly conserved across vertebrates (Yamada et al., 1998).

Motifs C2 and D2

Motifs C2 and D2, both human-specific, were only retrieved when using the six-species alignment. The use of the six-species alignment further reduces the search space and allows retrieving less pronounced (~degenerated) motifs. As input settings C and D only differ by one input gene, *PDLIM2*, it is not surprising that motif matrices C2 and D2 are similar and align very well (Tomtom p-value = 8.34e-06). The TF binding sites assigned to both motifs by the FuzzyClustering algorithm largely overlap (data not shown) and almost all 11 human genes contain at least one binding site in their non-coding region (except for *CPE* and *TSC22D3*). Motif C2 has a high prediction rate (20%) and its contributing motif matrices were predicted by both PG and PS. Motif D2 has a prediction rate of 10% and only predictions made by PG contributed to this ensemble motif. The p-values for the overrepresentation of both motifs in the human sequence set equal 0.022 for motif C2 and 0.24 for motif D2. When we compare both motif matrices with the TRANSFAC and JASPAR motif matrices we find a match between motif matrix C2 and the motif matrix for T3R (p-value = 0.01214) and a match between motif matrix D2 and the VDR motif matrix (p-value = 0.0124) as shown in Table 5.8. Notice that motifs C2 and D2 only cover a small part of the T3R and VDR motifs e.g. the optimal offset for motif D2 and the VDR motif equals seven, meaning that motif D2 matches only the second half of the VDR motif.

Table 5.8 Comparison between predicted and known motif matrices. For column information see Table 5.4.

NR	TF	Consensus sequence	p-value	Overlap
C2	T3R	SnnTrAGGTACGsn	0.01214	8
D2	VDR	nGGnnAnnnrGnnCA	0.0124	8

T3R (3,5,3'-triiodothyronine receptor) shares functional binding sites with VDR (vitamin D₃ receptor) as it is also a member of the nuclear hormone receptor (NHR) family which binds the DNA predominantly as a heterodimer with the Retinoid X Receptor (RXR) (Schrader et al., 1995). It is not surprising to predict a motif similar to the regulatory motif of VDR in a set of human genes upregulated after vitamin D₃ treatment, as many of the known biological effects of active vitamin D₃ are direct. This means that upon binding of 1 α ,25(OH)₂D₃, the VDR heterodimerizes with RXR and binds directly to vitamin D₃ response elements (VDREs) in the promoter region of its target genes, to induce or repress their expression (Bouillon et al., 2008). Those VDREs are scattered throughout the genome as they are positioned within approximately 100.000 bp either 5' or 3' of the TSS of the target gene (Haussler et al., 2010; Rachez and Freedman, 2000). Promoter analysis of genes transcriptionally regulated by 1 α ,25(OH)₂D₃ has identified many VDREs with distinct structures. The most common VDRE type, designated DR3, contains two direct repeats of 6 bp separated by a 3 bp spacer as shown in Figure 5.3. The sequence of the hexameric repeats, or half-sites, varies considerably, but a general consensus sequence of AGGTCA has been established (Haussler et al., 1998). Besides the variation in the consensus sequence of the two half-sites, also the length of the spacer separating them can vary, e.g. DR4 and DR6 (Carlberg, 1995). Additionally, other natural occurring VDREs have been reported like the inverted repeat (IR) (AGGTCA-spacer-TGACCT) (Echchgadda et al., 2004), and the everted repeat (ER) (TGACCT-spacer-AGGTCA) (Thompson et al., 2002), where the length of the spacer can vary from zero to eight nucleotides (Sandelin and Wasserman, 2005). For transcriptional activation, VDR occupies the 3' half-site whereas RXR binds the 5' half-site of VDRE site (Schrader et al., 1995) as shown in Figure 5.3.

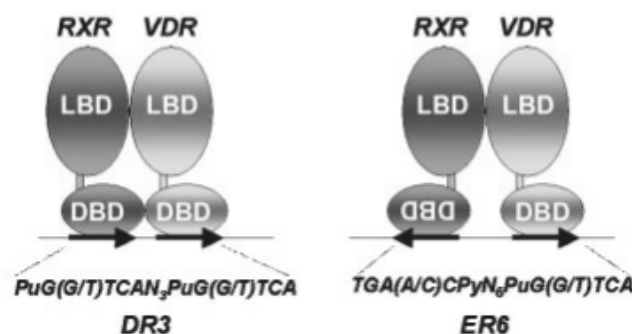


Figure 5.3 Schematic representation of binding of RXR-VDR heterodimers to vitamin D₃ response elements (VDREs) in the form of DR3 (left) and ER6 (right) motifs. Both the RXR and the VDR protein belong to the nuclear hormone receptor family and consist of a ligand-binding domain (LBD) and a DNA-binding domain (DBD). For the consensus motifs, Pu stands for purine (A or G) and Py stands for pyrimidine (T or C). Taken from (Lin and White, 2004).

However, *de novo* predicted motif D2 only corresponds to one half-site of the VDR motif. This can relate to our parameter settings as we searched for motifs with a motif width equal to 8 bp (see Material and Methods). To retrieve the full VDR regulatory motif, we also performed *de novo* motif discovery for a motif width equal to 15 bp. But those motifs with a longer motif width were too degenerate and didn't correspond to any known regulatory motif form the TRANSFAC/JASPAR database.

It is hard to discover *de novo* the full length VDR motif, not only because the VDREs can be positioned far up- or downstream of the respective target genes, but also because of the diverse VDRE configurations (DR, IR, ER) and the variable spacer lengths. As the *de novo* motif discovery tools, integrated in the PHYLO-MOTIF-WEB workflow, all model a regulatory motif by using a position weight matrix, they are not able to represent variable spacing and inversions of the half-sites within the TF binding sites. This inflexibility is usually not a concern because the structural constraints for monomeric protein-DNA interaction generally preclude such variation, but it is a substantial problem for the analysis of VDR target sites.

Therefore, we used a motif scanning approach specifically designed to predict binding sites for nuclear hormone receptors (NHRs) like VDR; the NHR-scan algorithm (Sandelin and Wasserman, 2005) (see Material and Methods). The results of the NHR-scan on the entire 4000 bp region for the 11 human genes are summarized in Table S9 (Supplementary Material). As shown in Table S9, there is overlap between the binding sites predicted by the NHR-scan and the binding sites of the *de novo* predicted motif D2 that resembles the VDR half-site motif. Table S9 also shows that NHR-scan retrieved a ER6-type binding site upstream of *PDLIM2*, which was experimentally confirmed being a functional RXR-VDR binding site by the Legendo research group. The location of this binding site (chr8:22434438-22434455) was previously predicted *in silico* by Wang *et al.* (Wang *et al.*, 2005). This binding site was not retrieved by PHYLO-MOTIF-WEB as it is not located in a regulatory region and not conserved for mouse as shown in Figure 5.4. This non-conservation can indicate that this VDR binding site is not responsible for the conserved gene expression seen across human and mouse.

Homo sapiens	TG-A--CCCAGCAGG---GGTCA
Mus musculus	TG-GACCCTAGCCAGGATGGTTAA

Figure 5.4 Conservation of the experimentally defined VDRE upstream of *PDLIM2*. The genome coordinates of this binding site are chromosome 8, position 22434438-22434455.

Other human genes that contain an ER6-type binding site with a score equal or better than the score of the true ER6-type binding site are: *CPE* that contains 3 ER6-type binding sites, *PMEPA1* that contains 2 ER6-type binding sites and *PISD*, *SEC14L1* and *GRAMD4* that contain each 1 ER6-type binding site. Also the most common DR3-type

binding site for VDR was predicted for a few genes as shown in Supplementary Table S9. However, none of those VDR binding sites were retrieved by the *de novo* approach.

5.2.2.2 *Cis*-regulatory module (CRM) detection

In this section we describe the results of CPModule (see Material and Methods). As input we used the entire 4000 bp region centered on the TSS for each of the 11 human input genes (includes *PDLIM2*). Using the entire 4000 bp allows to pickup TF binding sites missed in the *de novo* approach due to restriction of the search space to the regulatory regions. Further, we provided the set of *de novo* discovered motif matrices predicted by PHYLO-MOTIF-WEB in § 5.2.2.1. For this collection of motif matrices, CPModule will predict the subset of motifs for which the corresponding binding sites significantly collocate in CRMs in the human sequence set. As default, you can use the entire motif matrix collection from TRANSFAC and JASPAR. However, as we already had a clue on which motifs may play a role in the upregulation of our gene set, we could exclude possible noise by only providing the *de novo* discovered motif matrices. As we searched for CRMs in the human sequence set, only the human-specific motif matrices were qualified (input settings A, C and D). After leaving out the redundant (~similar) motif matrices we got; motifs A2, A3, D1, D2 and D3.

We ran the CPModule algorithm for different module sizes (i.e. maximum length of the region spanned by the CRM), ranging from 150 bp up to 400 bp. The CRMs, specific to our human sequence set and not to a background model (p-value < 0.001), are summarized in Table 5.9. For each module size, we enumerate the predicted CRMs, characterized by their rank (based on the p-value), their corresponding motif set, the number of target genes (i.e. the genes that contain at least one CRM in their selected 4000 bp region) and the names of the non-target genes.

Table 5.9 Results of CPModule for the human sequence set (~ 11 upregulated genes after vitamin D₃ treatment). For module sizes between 200 and 300 bp and for module sizes between 350 and 400 bp, CPModule predicted the same set of CRMs. For each CRM we report the rank, the corresponding motif set, the number of target genes and the names of the non-target genes.

Module size (bp)	Rank	Motif set	Number of target genes	Non-target gene names
150	1	D3-D2	9	<i>ID3, PDLIM2</i>
	2	D3-A2	11	/
	3	D2-A2	10	<i>PACSN2</i>
200-300	1	A2-A3	7	<i>ID3, TSC22D3, PMEPA1, PACSN2</i>
	2	D3-D2-A2	10	<i>PACSN2</i>
350-400	1	D1-A2	7	<i>PLXND1, ID3, SEC14L1, PDLIM2</i>
	2	A2-A3	7	<i>ID3, TSC22D3, PMEPA1, PACSN2</i>
	3	D3-D2-A2	10	<i>PACSN2</i>

For a module size of 150 bp, CPModule predicted three different CRMs, each containing TF binding sites for two different motifs: D3-D2, D3-A2 and D2-A2. Each of those CRMs covers at least 81.82% of the input sequences. When we increased the module size up to 200 or 300bp, CPModule predicted two CRMs: the first ranked CRM contains TF binding sites for two motifs: A2 and A3, the second ranked CRM combines the previous two-motif-CRMs to a three-motif-CRM, containing TF binding sites for motifs D3, D2 and A2. When we further increased the module size up to 350 or 400 bp, CPModule predicted the same CRMs as for module sizes equal to 200 or 300 bp, except for one new CRM containing TF binding sites for two motifs: D1 and A2. As shown in Table 5.9, all the *de novo* predicted motifs are contained in at least one CRM, specific for the 11 human sequences. Striking, is the presence of TF binding sites for motif A2 in almost every predicted CRM, which is maybe related to the degenerated nature of the motif matrix (consensus score of motif A2 equals 0.59).

The predicted CRMs are the DNA footprints of a set of TFs that co-operate to regulate gene expression. Motif D2, which resembles one half-site of the VDR regulatory motif, frequently co-occurs with two other regulatory motifs, one which resembles the ZEB1 motif (motif D3) and one which resembles the EBF motif (motif A2). When we added two extra motif matrices, representing the full DR3-type and the full ER6-type binding sites to the analysis, CPModule also predicted a significant CRM, which contains TF binding sites for the DR3, ZEB1 and EBF like motifs (data not shown).

5.3 Material and Methods

5.3.1 Microarray analysis

5.3.1.1 Differentially expressed genes

The microarray analysis was performed by a colleague of the CMPG Bioinformatics group Carolina Fierro. Data pre-processing and statistical analysis were adapted to the technical platforms and experimental design of both human and mouse microarray experiments. Mouse experiments were performed on a cDNA two-color microarray platform, measuring 5 time points (1, 6, 12, 24 and 36 hours after control and vitamin D₃ treatment) each with 2 technical replicates (dye-swap). The human dataset consists of 6 time points (1, 3, 6, 12, 24 and 36 hours after control and vitamin D₃ treatment) each with 2 biological replicates. The human experiments were performed on the Affymetrix platform. For both human and mouse datasets, an F-test was used to detect differentially expressed genes. For the mouse data we used a stringent false discovery rate (FDR) cutoff of 3% (Storey and Tibshirani, 2003) that resulted in 740 differentially expressed genes. For human, the biological replicates greatly differed, so we used a much less stringent FDR cutoff of 20%, which resulted in 143 differentially expressed genes.

5.3.1.2 Selection of human and mouse genes

We focused our analysis on genes with orthologs in both human and mouse (as defined in Ensembl version 54), which were present in both microarray platforms. Non-specific probes (i.e. probes that match to more than one gene) were left out the analysis. Since many probes could match the same gene, we averaged the profiles of probes representing the same gene only if the Euclidean distance between two profiles was lower than 3. In case there were no probes selected for this Euclidian distance cutoff, the gene was discarded from the analysis. The intersection of differentially expressed genes, i.e., genes differentially expressed in both species, consisted of only 9 genes. As this number is very limited, we used the union (i.e. differentially expressed in one of the two species) of differentially expressed genes. The final union consists of 505 mouse genes, orthologous to 515 human genes.

5.3.1.3 Detection of conserved expression behaviour

For those 505 mouse genes, orthologous to 515 human genes, we applied a clustering approach to group the genes with a similar expression profile over time. In this study, we used 'Euclidian distance' to assess the relationship between two gene expression profiles. We used the differential cluster analysis (DCA) (Ihmels et al., 2005) that first clusters the genes in one species into multiple coexpressed groups. In a second step, DCA will for each coexpressed group in the first species re-cluster the genes, based on the profiles of their orthologs in the other species, into a predefined number of clusters (two). For the

DCA we started by clustering the human data (hierarchical clustering to obtain 7 clusters) and reorder each human cluster based on hierarchical clustering over the mouse profiles. We finally obtained one partially conserved cluster, three clusters with split conservation and three clusters with no conservation between human and mouse. For the first two cases we selected three sub-clusters; a sub-set of genes, coexpressed in both human and mouse. From these three sub-clusters, only one clearly represents genes up-regulated in both human and mouse and consists of 10 genes (see Figure 5.2). This conserved coexpression cluster will be used to infer *cis*-regulatory elements.

5.3.1.4 The *PDLIM2* gene

On the human microarray *PDLIM2* was represented by two probes, each with a different expression profile and therefore not picked-up by our microarray analysis (see Euclidian distance cutoff = 3). This could be the result of different splicing variants of the *PDLIM2* gene. As one of the two probes shows a clear upregulated expression profile in human, consistent with the expression profile in mouse, it is justified to include *PDLIM2* for input setting D.

5.3.2 **Identification of *cis*-regulatory elements**

5.3.2.1 *De novo* motif discovery by PHYLO-MOTIF-WEB

For input settings A and C, we used the human Ensembl gene IDs while for input setting B the mouse Ensembl gene IDs which correspond to the 10 genes in the conserved coexpression cluster and which were retrieved in the Ensembl database version 59 (August 2010). For input setting D, the human Ensembl gene ID for *PDLIM2* was added.

For each input gene, we selected the 4000 bp genomic region, centered on the TSS of the gene. As many eukaryotic TF binding sites locate in the 5' untranslated region (5'UTR), the first intron or the proximal promoter region, we may also capture those regions within our selection. The 4000 bp regions were restricted to their 'regulatory regions' as defined by the Ensembl Regulatory Build pipeline (version 59), thereby also excluding the protein coding regions (exons). As there are no cell type specific regulatory regions available for the human breast cancer cell-type and the mouse osteoblastic cell-type, we used 'MultiCell-type' regulatory regions based on features like DNase I, which is known to mark accessible chromatin, binding site locations of other TFs defined by ChIP-chip or ChIP-seq experiments and binding sites for the enhancer binding factor CTCF.

The genomic coordinates of the 'regulatory regions', were used to retrieve the alignment blocks from the Ensembl database. For input settings A and B, we selected the pairwise alignment blocks between human and mouse, for input settings C and D we selected the six-species alignment blocks to further restrict the search space. The genome sequences of those six species (human, mouse, macaque (*Macaca mulatta*), cow (*Bos taurus*), horse

(*Equus caballus*) and dog (*Canis familiaris*)) show a good trade-off between still alignable and informative and were recommended by Eric van Nimwegen (author of the PG software). In case of PG and PS, PHYLO-MOTIF-WEB provided the alignment blocks for each input gene, for MEME on the other hand, the orthologous sequences were left unaligned.

PHYLO-MOTIF-WEB predicts a set of ensemble motif matrices based on clustering the results of three motif discovery algorithms; PG, PS and MEME. A major advantage of PHYLO-MOTIF-WEB is that the user can chose a range of parameter values and PHYLO-MOTIF-WEB will run each motif discovery algorithm for all possible combinations of those parameter values. Table 5.10 summarizes the parameter values for *de novo* motif discovery by PG, PS and MEME.

Table 5.10 Parameter values used for the three motif discovery algorithms (PS, PG and MEME) implemented in the PHYLO-MOTIF-WEB workflow. We have the ‘general’ parameters, applicable for each motif discovery algorithm and the algorithm-specific parameters ‘PS’, ‘PG’ and ‘MEME’. For more details on the parameters of PG and PS and their meanings see Supplementary Text S3.

	Parameter	Value
General	Motif type	Search for both normal and palindromic motifs
	Number of different motifs	1, 2 or 3
	Motif width	8, 15
	Number of TF binding sites per motif	(0.50, 1 or 2 TF binding sites per input sequence) * the number of input sequences
PS	Maximum number of TF binding sites per sequence	2
	Threshold on the posterior probability	0.25 (1/(1+gamma) with gamma = 3)
	Number of seeds	1
	Burn-in iterations	1000
	Sampling iterations	2000
PG	Iterations during simulated annealing	100
	Iterations during tracking	100
	Pseudocount	1
	Tracking threshold	0.05
MEME	P-value threshold	0.01

PG and MEME use a third order Markov background model based on the promoter regions selected from the Eukaryotic Promoter Database (EPD) (Schmid et al., 2006). PS builds its own position specific background model based on the set of input sequences.

The ensemble strategy of PHYLO-MOTIF-WEB uses the FuzzyClustering algorithm to reduce the large and redundant collection of motif matrices predicted by PG, PS and MEME. FuzzyClustering acts on the level of the TF binding sites and in order to retrieve species-specific motif matrices, only the TF binding sites which were predicted for the reference species (human or mouse), were used as input. FuzzyClustering clusters pairs of TF binding sites with a statistically relevant jointly occurrence, across all predicted motifs, into meaningful ensemble motifs, that have a higher likelihood to represent a true motif than the original set of predicted motifs. FuzzyClustering was run in the ‘traceback’

mode (-t), so that it reports the percentage of predicted motifs that contributed to the ensemble motif (~prediction rate) and the motif discovery algorithms by which those contributing motif matrices were predicted (contribution level of each motif discovery tool). The threshold for the individual binding site probability (-p) was set to 0.005. The consensus cutoff parameter (-c), defines the cutoff on the consensus score of the ensemble motif and was set to 0.50 (default). The consensus score is a measure for the degeneracy of a motif matrix: a non-degenerated motif has a score equal to 2 while a motif with a uniform nucleotide distribution (fully degenerated) has a score 0. The sequence volume (-i), which stands for the minimum number of reference sequences that should contain at least one TF binding site of the ensemble motif was set to 0.20. For more details see chapter 4.

To evaluate if an ensemble motif matrix is specific for the input set we used Clover (Frith et al., 2004) to calculate the p-value for the motif matrix's over-representation in the set of sequences for the reference species versus random sequence sets. The reference sequence set consisted of the full 4000 bp regions for each gene (human or mouse). The random sequence sets were sampled from a collection of 4000 bp regions centered on the TSS for all genes annotated in the Eukaryotic Promoter Database (EPD), for either human or mouse. Clover calculated for each ensemble motif a 'raw score', indicating how strongly the motif is present in the reference sequence set and a corresponding p-value that assigns the statistical significance of this raw score. The number of randomizations (-r) was set to 1000. A p-value of zero means that the 'raw score' for the motif in the random sequence set could never equal to the one obtained for the reference sequence set in any of the 1000 randomizations.

A common question within the context of *de novo* motif discovery is whether a newly discovered motif resembles any previously discovered motif in an existing database. To answer this question, we used the software Tomtom (Gupta et al., 2007) that uses the 'Pearson correlation coefficient' as statistical measure of motif-motif similarity and searches a database of 'target' motifs with a given 'query' motif. As query motif we used the ensemble motifs predicted by PHYLO-MOTIF-WEB and searched for significant matches in the TRANSFAC and JASPAR databases. A match was found significant if the E-value < 10 (i.e. the expected number of times that the given query would be expected to match a target as well or better than the observed match in a randomized target database of the given size). The E-value is the result of applying a form of Bonferroni correction for multiple tests (matching a query motif against multiple target motifs) which assumes that the targets are independent of one another. The correction consists of multiplying the motif p-value by the number of targets in the database. In Table 5.3 we report the motif p-value for the optimal offset and orientation of the query and target motif.

5.3.2.2 NHR-scan

The NHR-scan (Sandelin and Wasserman, 2005), is a flexible Hidden Markov Model framework capable of predicting NHR binding sites by using a model that allows for variable spacing and orientation of half-sites. Essentially the model consists of three 'match state chains' - corresponding to each type of binding site configuration (direct, inverted and everted repeats), and one 'background state' - corresponding to no prediction. Each 'match state chain' is composed of two half-site models and a spacer model separating them. From the background state, it is possible to move to each type of match state. To identify candidate binding sites, the NHR-scan implemented the Viterbi algorithm. Given the model, the Viterbi algorithm is applied to identify 'the most probable chain of states' that is consistent with the observed sequence. We ran the NHR-scan algorithm with the 'probability for entering match states' set to the default value of 0.01. This value corresponds to the transition probability from the background state to the different match state chains. The probability of entering match states can be viewed as a sensitivity and selectivity trade-off. Higher values will result in more predictions: more true sites will be detected but more false sites will be reported.

5.3.2.3 CRM detection

The CPModule algorithm (Guns et al., 2010) consists of 3 phases: screening, mining and ranking. The 'itemset mining phase' of CPModule is preceded by a 'scanning phase' to identify putative TF binding sites for each *de novo* predicted motif matrix. Scanning is performed by Clover (Frith et al., 2004) under low stringency conditions which requires further filtering of the TF binding sites. This filtering was done based on whether or not the binding sites were located in regions of transcriptionally active chromatin (~low nucleosome occupancy). Also binding sites in regions with a GC-content > 80% were filtered out. For the itemset mining, CPModule applies the framework of constraint programming for itemset mining. The constraints involve the putative binding sites of TFs, the number of sequences in which they co-occur and the proximity of the binding sites (~ module size). We searched for CRMs which occur in at least 60% of the input sequences and with module sizes ranging between 150 and 400 bp. For each potential CRM, CPModule assesses its specificity for the input sequences compared to a background model by calculating a p-value, and ranks the potential CRMs accordingly.

5.4 Discussion

In this chapter, we aimed to elucidate the molecular mechanisms underlying the antiproliferative effects of vitamin D₃ on human breast cancer cells and mouse bone cells. We performed a comparative transcriptome analysis across human and mouse cell lines that both showed decreased cell growth after vitamin D₃ treatment. We could extract a cluster of genes, which were all upregulated after vitamin D₃ treatment, in both species. To elucidate the transcriptional regulation underlying the conserved coexpression behaviour of the genes, we assumed that also the transcriptional regulation had been conserved and searched for conserved *cis*-regulatory elements in the gene's non-coding sequences.

5.4.1 *De novo* motif discovery

First, we applied a *de novo* strategy to retrieve completely novel motifs, without any prior information on the motif type. This approach revealed a collection of putative motifs including two GC-rich motifs, which resemble the regulatory motifs of the SP1 and ZFP161/NRF1 TFs. The SP1-like motif was predicted for each of the four input settings, while the ZFP161/NRF1-like motif was only predicted for settings A and B, using the human-mouse pairwise alignment. Both GC-rich motifs seemed not very specific for our human sequence set. Indeed, SP1 and NRF1 motifs occur in the promoter regions of many human genes and are well conserved among vertebrates (Xie et al., 2005; FitzGerald et al., 2004). Both motifs are also enriched in the promoters of genes that function in cell cycle and are thus involved in controlling the cell cycle (Elkon et al., 2003). Notice, that only MEME and PG, not PS, contributed to those GC-rich motifs. In contrast to MEME and PG, PS uses a position specific background model that takes into account the locally varying GC-content of the input sequences (see chapter 2). Variations in GC-content are common in vertebrate promoter regions as they may contain CpG islands (i.e. stretches of frequently unmethylated CpG dinucleotides that are transcriptionally active). The GC-rich binding sites of SP1 often locate in CpG islands and even contribute to the maintenance of their hypomethylated state (Brandeis et al., 1994), making it very hard for PS to discover those binding sites.

In case we searched for motifs conserved across the human-mouse pairwise alignment, a motif that resembles the regulatory motif of EBF was predicted, only by PS. The degenerated nature of this motif matrix can explain its low overrepresentation for the upregulated sequence set and its presence in every CRM predicted by CPModule. The TFs that could possibly bind this motif are mainly involved in B-cell differentiation.

When searching for motifs conserved across multiple mammalian species, both PG and PS predicted a motif that resembles one half-site of the VDR motif. This may indicate

direct regulation of the upregulated genes by VDR in response to vitamin D₃ treatment. We used NHR-scan, a scanning approach specifically developed to predict NHR binding sites, to exactly locate possible VDR binding sites in the promoter region of each human gene. NHR-scan could predict the true VDRE in the promoter region of *PDLIM2*.

When we added the *PDLIM2* gene also to our *de novo* approach (input setting D), PS predicted an extra motif, which resembles the regulatory motif of ZEB1. Most of the corresponding TF binding sites were indeed retrieved in the non-coding region of *PDLIM2*. Both proteins (PDLIM2 and ZEB1) are involved in metastasis, i.e. the process by which cancer spreads to surrounding tissues. The process of metastasis requires that epithelial cells undergo a de-differentiation process known as epithelial-mesenchymal transition (EMT). ZEB1 is involved in EMT, as it downregulates E-cadherin, an intercellular adhesion protein, in order to increase cell motility (Sanchez-Tillo et al., 2010). However, other studies also indicate that the levels of different ZEB1 co-factors are critical for the action of ZEB1 (repressive or activating) on E-cadherin (Pena et al., 2005). Further on ZEB1 was also reported as a transcriptional activator of VDR (Lazarova et al., 2001). PDLIM2, suppresses anchorage-independent growth of cancer cells, which suggests a possible tumor suppressor function (Loughran et al., 2005). We may suggest for ZEB1 a possible role as transcriptional repressor or activator of the *PDLIM2* gene.

Conclusively, the *de novo* approach benefitted from using different motif discovery algorithms that each contributed different motifs, related to their algorithmic backgrounds. Both, very common motifs such as the SP1 and NRF1-like motifs as well as motifs more specific for the set of upregulated genes such as the VDR and ZEB1-like motifs were predicted.

5.4.2 *Cis*-regulatory modules

In higher eukaryotes, TFs rarely operate by themselves, but rather bind to DNA in cooperation with other DNA-binding proteins. To infer possible combinatorial regulation for the set of human sequences, we searched for combinations of TF binding sites corresponding to the *de novo* predicted motif matrices. In a study of Blanchette *et al.* (Blanchette et al., 2006), where they predicted CRMs for the whole human genome, 58% of the predicted CRMs had a module size of less than 500 bp. In this study, we considered module sizes of 150 bp up to 400 bp. For a module size of 350 bp, CPModule predicted three CRMs that were specific for the human sequence set compared to random sequence sets. Each of those CRMs contained the EBF-like motif, in combination with the SP1-like motif (first CRM) or the NRF1-like motif (second CRM) or in combination with both the VDR and the ZEB1-like motifs (third CRM). None of those TFs are known to co-act in transcriptional regulation following the literature.

The scanning phase of CPModule is still open for improvement. TF binding sites located in promoter regions with a high GC-content (>50%) were filtered out. This threshold is very stringent because many CRMs in the human genome overlap with CpG islands (Blanchette et al., 2006) and we therefore relaxed this threshold to 80%. Next, the scanning method does not correct for the high number of individual binding sites predicted for degenerated motif matrices (e.g. motif A2 ~ EBF-like motif). Only the statistical significance of the resulting CRMs was calculated. Further, we noticed overlap between TF binding sites predicted for different motif matrices. As the regulatory motifs of different TFs may resemble each other, especially if they come from the same family, it can be useful to integrate an extra constraint that avoids overlap between predicted TF binding sites. A possible improvement, more specific for the case of VDR binding sites, is the use of the NHR-scan predictions instead of using the default PWM screening with the motif matrix of VDR.

5.4.3 Future perspectives

The research performed in this chapter focussed specifically on the cluster of coexpressed genes which was conserved across the human and mouse cell lines. The motifs predicted for this small set of genes and the corresponding regulators could be promising elements to further elucidate the molecular pathways induced by vitamin D₃. For future research it would be interesting to also investigate the presence of these motifs in the rest of the human/mouse genome. For example, the cluster of genes upregulated after vitamin D₃ treatment in human breast cancer cells was only partially conserved in mouse bone cells. We could investigate the complete set of human genes upregulated after vitamin D₃ for the presence of the predicted motifs.

Although motif discovery can indicate *cis*-regulatory elements, overrepresented in the input sequence set, it remains hard to determine ‘functional’ *cis*-regulatory elements. We used PHYLO-MOTIF-WEB, which allows integrating many information sources to guide the motif search as much as possible to functional DNA regions, e.g. nucleosome depleted DNA regions (~DNase I hypersensitivity information), the locations of other TF binding sites, evolutionary conserved regions *etc.* Ideally, PHYLO-MOTIF-WEB would also use information on chromatin modifications (e.g. histone modifications), but for now, this cell type and condition dependent information is lacking for the human and mouse cell lines used in this study.

Chapter 6 General discussion and perspectives

6.1 General discussion

Although there is an enormous collection of tools available, the discovery of TF binding sites (motifs) stays a challenging problem. The evolution in motif discovery was characterized by optimizing existing algorithms to maximally exploit all kinds of information sources. In this thesis, we focused on the class of *probabilistic* motif discovery tools that try to optimize the use of *phylogenetic information*.

In **chapter 2**, we made a theoretical comparison between two well established, probabilistic motif discovery algorithms that use a tree-based evolutionary model to integrate phylogeny; Phylogibbs (PG) and Phylogenetic sampler (PS), also referred to as phylogenetic motif finders. In **chapter 3**, we evaluated the conditions under which complementing coregulation with orthologous information improves motif discovery for this class of phylogenetic motif finders. We designed appropriate benchmark datasets and made an exhaustive evaluation of both algorithms together with MEME, as a representative of algorithms that cannot explicitly incorporate phylogenetic relations.

First, we learned very useful hints for further improvement of motif discovery from the relation between ‘the working principles of the algorithms’ as explained in chapter 2 and ‘their performance results’ as obtained in chapter 3.

- In this way, we showed that the use of an ensemble strategy, to estimate the true optimum in the dataset, was more successful than searching for a single optimal solution. This was especially true for datasets showing a low signal to noise ratio, as is often the case for biological datasets, e.g. a set of coexpressed genes derived from a microarray experiment likely contains targets of more than one regulatory protein, lowering the relative overrepresentation of a particular motif. However, this advantage of using an ensemble solution comes at the expense of much longer running times, as it is computationally very demanding.
- Another important observation was that motif discovery performance of phylogenetic motif finders depends on the quality of the ortholog alignments, as was also stated by (Gordan et al., 2010; Ward and Bussemaker, 2008). The deleterious effects of errors in the ortholog alignments mainly arose when using evolutionary distant orthologs, difficult to prealign correctly. However, adding distant orthologs to the alignment usually relieved the problem of multiple local optima induced by a set of closely related orthologs. Our results showed that combining a local alignment strategy with a more flexible way of assigning TF binding sites (e.g. TF binding sites may be absent in some of the orthologs) could

increase robustness against difficult to align datasets, compared to using a global alignment strategy with a very rigid assignment of sites. Moreover, this local alignment strategy indirectly accounts for TF binding site turnover, which is a common event during genome evolution, and plays a major role in shaping the regulatory circuitry of contemporary species (Wray, 2007).

- In general, motif discovery by the phylogenetic algorithms was poor for the orthologous space, where evolutionary conservation forms the only information source. In case the dataset only contains one set of aligned orthologous sequences, we expect that the adequacy of both the evolutionary model and the phylogenetic tree will be important to correctly predict the motif. But, if the motif discovery algorithm is unable to sufficiently explore the search space (i.e. the predictions depend on the initialization of the algorithm) for this type of data, no results will be retrieved. In contrast to the moves used by a typical Gibbs sampler, which failed to deviate from the initialization point, the moveset defined by PG seemed more successful to explore the orthologous information space.

Secondly, chapter 3 showed that the success rate of combining coregulation and orthology information depends on the complex relation between the algorithm and the dataset.

- The performed tests illustrated that the nature of the used algorithm is crucial in determining how to exploit multiple species data in the best way to improve motif discovery performance. Most influencing the results were ‘the phylogenetic distances between the orthologous sequences’ as they affect the trade-off between align-ability of the orthologs and the information level of the dataset.
- The topology of the tree (star- or tree-like structure) that describes the relatedness of the orthologous sequences does not affect the performance of the phylogenetic tools. However, the tree should reflect the true evolutionary distances between the non-coding orthologous sequences, as both phylogenetic motif finders are sensitive to underestimating those distances. This is the case when using a species tree based on the conservation of rRNA or protein-coding DNA sequences.
- The number of added orthologs mainly influenced the results for the orthologous space, with as rule: the more orthologs added the better the results.

The results of the extended benchmarking in chapter 3 also showed that the performance of a motif discovery algorithm strongly depends on the dataset specificities. Each algorithm performed better than the others on some type of data, which can be explained by the differences in their algorithmic backgrounds. The same observation was made by Tompa *et al.* (Tompa et al., 2005) for a set of algorithms performing in the coregulation space. This diversity in performance has led to the idea that ensemble methods,

comprising the results of multiple motif finders may lead to improvements in prediction accuracy. Diversified predictions by different motif finders would contribute in increasing the sensitivity of the final prediction (Hu et al., 2006). As dozens of motif discovery algorithms are available today, the ensemble approach is especially promising to use them to our advantage.

In **chapter 4** we developed a workflow for *de novo* motif discovery in eukaryotes, PHYLO-MOTIF-WEB, which can easily be accessed through its web service. PHYLO-MOTIF-WEB applies an *ensemble strategy* to comprise the results of multiple motif discovery algorithms, for now PG, PS and MEME. To the best of our knowledge, this is the first time that an ensemble approach is used in the motif discovery problem that incorporates phylogenetic motif finders. Moreover, the workflow covers all the different pre- and post-processing steps needed to identify potential motifs in a set of non-coding sequences, making hard-to-use phylogenetic tools like PG and PS more accessible to the non-expert user. For the ensemble strategy we developed a clustering algorithm, FuzzyClustering that sequentially groups a set of TF binding sites that significantly appear together in a set of predicted motifs. For each asymmetric cluster, all the individual binding sites and motif predictions get a membership probability assigned that reflects the relative importance of a binding site/predicted motif in the extracted cluster. Then, the TF binding sites are aligned to construct a position specific weight matrix (i.e. the ensemble motif). We quote some advantages of using FuzzyClustering.

- Compared to algorithms that cluster on the level of the motif matrix (Habib et al., 2008; Romer et al., 2007; Thijs et al., 2002a), FuzzyClustering can distinguish between more and less likely TF binding site predictions, based on their membership scores. This allows better fine-tuning of sensitivity and specificity of the results.
- For each ensemble motif, FuzzyClustering can evaluate the fraction of predicted motifs that significantly contributed and trace back the contribution level of each component algorithm (PG, PS and MEME). This makes it possible to put more confidence in an ensemble motif supported by a high fraction of the predicted motifs retrieved by multiple motif finders.
- PS also uses an ensemble strategy to cluster the results across multiple initializations and iterations of the algorithm itself (see chapter 2). The difference with FuzzyClustering is that PS does not account for the co-occurrence of different TF binding sites per iteration of the algorithm. Instead, PS just counts individual binding site frequencies and in case the algorithm was asked to search for multiple motifs simultaneously, an extra post-processing step is required to group the significant binding sites into different motifs.

- FuzzyClustering integrates the score, assigned to each TF binding site by the motif discovery algorithm, into the cluster extraction process. In Hu *et al.* (Hu *et al.*, 2006) they not only propose to use the scores assigned by the motif discovery algorithm, but also weigh those scores by the overall accuracy of the algorithm itself. Although, determining algorithm accuracy seems not straightforward, especially as it depends on many factors (see chapter 3).
- To enhance biological meaningful results, the ensemble motif is only reported if the consensus score is sufficiently high, TF binding sites are sufficiently distributed over the input sequences and a sufficient number of predicted motifs contributed to the ensemble motif.

Using FuzzyClustering to resume the results of multiple motif discovery algorithms requires a more extended benchmarking (research in progress). This to further elucidate the ‘fuzzy’ aspect (i.e. TF binding sites and predicted motifs may be assigned to different clusters with different membership probabilities) and how this relates with possible redundancy in the final set of ensemble motifs. To enhance the diversity in the predictions we ran each motif discovery algorithm for different parameter settings. Varying the parameters is reasonable since the optimal parameter settings for a given dataset cannot be estimated in advance. However, a possible downside is that bad choices for the parameters can deteriorate the quality of the predictions, resulting in poor consensus building.

PHYLO-MOTIF-WEB performs *de novo* discovery of motifs in eukaryotes, which is extremely difficult, due to the low signal to noise ratio. Integration of extra information like evolutionary conservation can partially increase this signal to noise ratio, but more efficient would be to restrict or prioritize the search space to those DNA regions that are likely involved in transcriptional regulation. Epigenetic information like chromatin structure can provide information on ‘transcriptionally active’ DNA regions. However, chromatin structure is tissue and condition dependent. Due to the recent development of new experimental methodologies like ChIP, an increased amount of experimental epigenetic data for several eukaryotic tissues and conditions becomes available. This inevitably creates a major computational challenge to incorporate those new data to improve motif discovery (e.g. allows distinguishing TF binding sites functional in one physiological condition or tissue from another).

Although extremely promising, our knowledge is still limited, as the pattern of chromatin modifications present in the cell, like histone modifications and the methylation of DNA, constitute a ‘code’ that is not fully understood yet. More knowledge on which chromatin modifications co-locate with transcriptionally active regions like promoters or enhancers or even with the binding locations of specific TFs will enhance usability. PHYLO-

MOTIF-WEB provides an option to use epigenetic information to restrict the motif search space, by integrating the information provided by the Regulatory Build Pipeline of Ensembl (Hubbard et al., 2009).

- This pipeline combines a variety of genome-wide epigenomic and genomic data sets to annotate *potential regulatory regions* within the genome. The vast majority of features are derived from ChIP-seq data.
- Annotations were made across multiple cell types, based on open chromatin defined by DNase I hypersensitivity mapping and FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) and the genome-wide binding locations of CTCF (i.e. marker for enhancer/insulator regions) and other TFs. The multiple cell type annotations can be extended in a cell type specific manner, based on histone modifications assayed by ChIP.

The use of epigenomic data, preferably cell type and condition specific, can help to predict the actual functionality of *cis*-regulatory elements as this depends on chromatin accessibility by the regulatory proteins. However, actual *in vivo* TF binding can only be validated by experiments like ChIP preferably combined with expression data to check for transcriptional activity.

In **chapter 5** we focussed on a biological topic, namely gaining more insight in the mode of action underlying the antiproliferative effects of vitamin D₃ on human breast cancer cells and mouse bone cells. We performed a comparative transcriptome analysis to study the transcriptional response towards vitamin D₃ treatment, in both human and mouse cell lines. We could extract a cluster of genes, which were all upregulated after vitamin D₃ treatment, in both species. Despite a few genes involved in cell death or apoptosis, no biological process was overrepresented in this cluster of genes. This heterogeneous nature of the extracted cluster can reflect the multiple sites of action of vitamin D₃, maybe suggesting the involvement of different regulatory proteins.

To elucidate the transcriptional regulation behind the conserved coexpression behaviour of the genes, we assumed that also the transcriptional regulation had been conserved and searched for conserved *cis*-regulatory elements in the gene's non-coding sequences. We first performed a *de novo* approach by using PHYLO-MOTIF-WEB and then analysed if the binding sites of the *de novo* predicted motifs co-locate in the DNA, to form CRMs. The *de novo* approach predicted on one hand very common motifs, known to be involved in cell cycle regulation, such as the SP1- and NRF1-like motifs, but also more specific motifs such as the VDR- and ZEB1-like motifs. The VDR-like motif suggests direct regulation of the genes by the active metabolite of vitamin D₃. The ZEB1-like motif can be interesting because ZEB1 is a transcriptional activator of *VDR* and is involved in the process of metastasis. Although none of those individual motifs was convincingly

overrepresented in the coexpressed gene set, particular combinations were. For example, the co-location of the binding sites for VDR and ZEB1 was significant for the human coexpressed gene set. This shows that the detection of CRMs can produce predictions of substantially better specificity than the analysis of isolated binding sites.

As standard probabilistic motif discovery tools usually fail to recover the full-length VDR motif, we also used NHR-scan that integrates a Hidden Markov Model framework to account for the diverse configurations (DR, IR and ER) and the variable spacer lengths of the VDR motif. NHR-scan could predict the experimentally verified ER6-type binding site upstream of *PDLIM2*.

In silico prediction of VDR target genes, following a *gene-centered approach* is almost impossible as the VDR binding locations are scattered throughout the genome. The study of Ramagopalan *et al.* (Ramagopalan *et al.*, 2010), presented a comprehensive high-resolution map of VDR binding throughout the human genome defined by CHIP-seq in lymphoblastoid cells. After calcitriol stimulation of the cells, they identified 2776 VDR binding sites which were mainly located in introns (36%) and in regions at least 5000 bp away from the first or last exon of the gene (28%). The remaining VDR binding sites (36%) were located within 5000 bp of the gene's first or last exon. This *genome-wide approach* can be combined with *in silico* motif discovery to retrieve the exact VDR binding motif. In the study of Ramagopalan *et al.*, the DR3-type of VDR motif was most significantly enriched within the VDR binding intervals. They also described that those VDR binding intervals were significantly enriched in regions associated with active chromatin such as DNase I-hypersensitive sites, CTCF binding locations and specific histone modifications (e.g. H3K4me3 and H3K27ac).

6.2 Perspectives

6.2.1 The mode of action of vitamin D₃

In this thesis, we only performed a transcriptome analysis where we studied the gene expression profiles after vitamin D₃ treatment and predicted *in silico* TF-DNA interactions. In order to obtain a more comprehensive view on the molecular actions of 1 α ,25(OH)₂D₃ (i.e. the active metabolite of vitamin D₃), we plan to reconstruct its regulatory network by the integrated analysis of different 'omics' data sets. A multi-level approach using high-throughput ChIP-seq, and expression profiling (microarray and 2D-DIGE) will be used to detect 1 α ,25(OH)₂D₃-induced changes at the chromatin, the mRNA and the protein level in one cell type (see Figure 6.1).

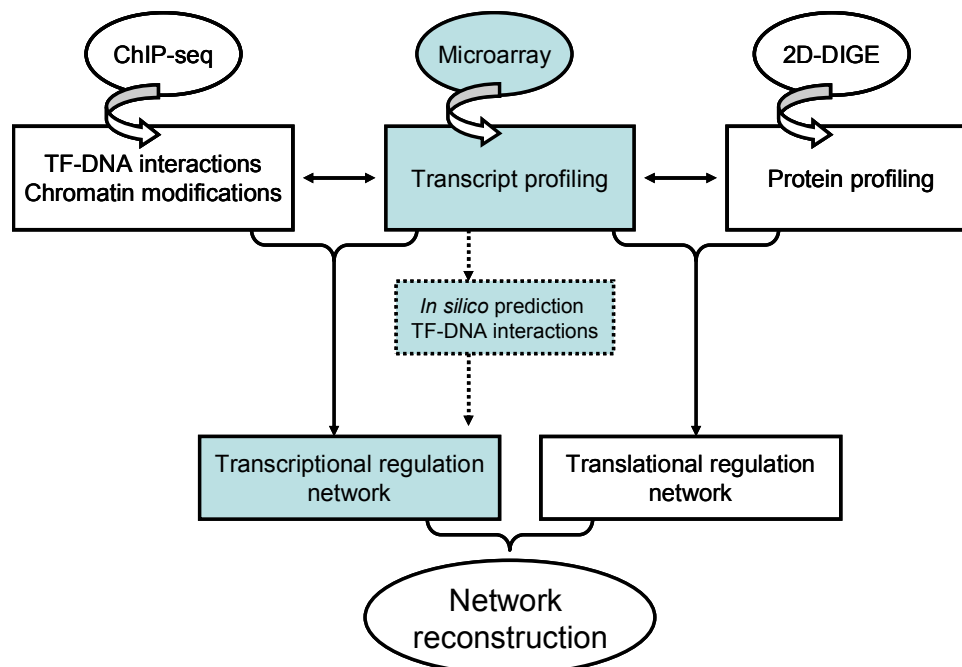


Figure 6.1 A multi-level approach by integrating different 'omics' data sets to get a comprehensive view on the 1 α ,25(OH)₂D₃ regulatory network. The boxes colored in blue, show the one-level transcriptome analysis performed in this thesis.

New and existing bio-informatics tools will be used to analyze and integrate the data at different levels in order to reconstruct the 1 α ,25(OH)₂D₃ regulatory network. The resulting comprehensive view will allow new and crucial signaling cascades in the actions of 1 α ,25(OH)₂D₃ to be uncovered and characterized. Integration of all different sources is important to overcome the limitations of each of the individual techniques. As an example: despite being bound in ChIP-seq data, some target genes might be found transcriptionally inactive in the expression data, indicating the necessity of temporally and/or tissue-restricted transcriptional co-regulators. On the other hand, expression data

does not allow distinctions to be made between primary events directly controlled by TF binding and secondary transcriptional changes, while ChIP-seq can.

6.2.2 Integrating multiple information sources to predict TF binding

Recent progress of experimental technologies like high-throughput sequencing and ChIP-based approaches will have a tremendous impact on our understanding of the transcriptional regulatory mechanisms. New datasets like genome-wide TF binding locations, histone modifications, DNA methylation and nucleosome occupancy will become available for an increasing amount of cell types and physiological conditions, all providing new evidence in the search for regulatory regions in the genome. With the growing number of information sources, computational methods for integrating these diverse data sources can further improve the prediction of TF binding. To enhance the integration of these new data in computational approaches and to enable a more faithful construction of the transcriptional regulatory network, initiatives like the ENCODE project (Birney et al., 2007), together with the UCSC Genome Browser (Thomas et al., 2007) provide an integrated visualization and standardized retrieval of various genome and epigenome datasets to the research community.

Appendix - Supplementary materials

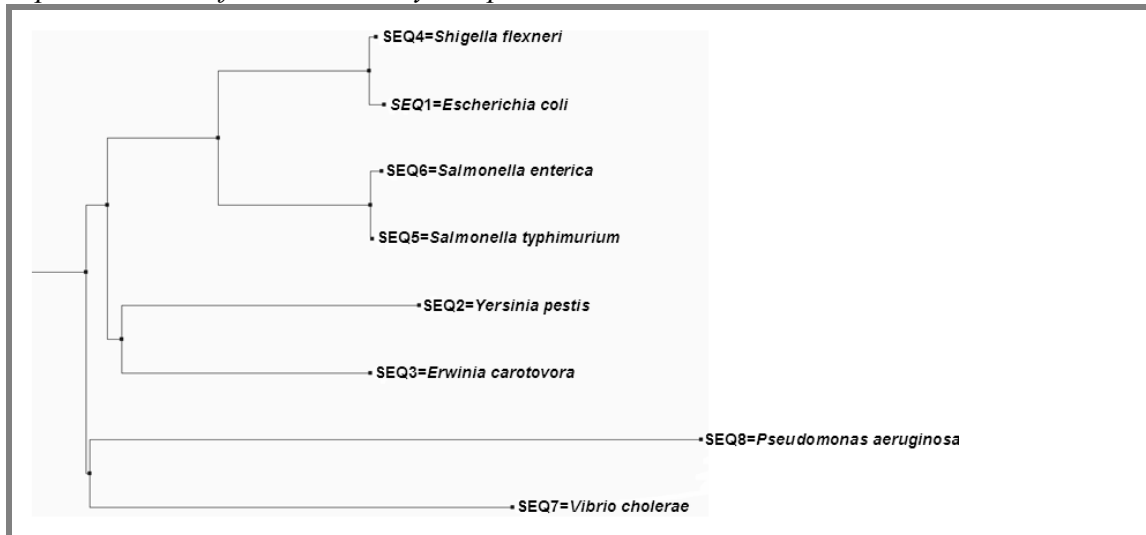
Chapter 1: Introduction

Table S1 IUPAC (International Union of Pure and Applied Chemistry) codes used to denote ambiguous positions in nucleotide sequences. Adapted from (Pavesi et al., 2004).

IUPAC	Nucleotides	Mnemonics
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
R	A or G	puRine
Y	C or T	pYrimidine
M	C or A	aMino group at common position
K	T or G	Ketogroup at common position
W	T or A	Weak hydrogen bonding
S	C or G	Strong hydrogen bonding
B	C, T or G	not A
D	A, T or G	not C
H	A, T or C	not G
V	A, C, or G	not T
N	A, C, G or T	aNy

Chapter 3: The effect of orthology and coregulation on detecting regulatory motifs

Figure S1 depicts the phylogenetic trees used to relate the eight Gamma-proteobacterial species and the five *Saccharomyces* species.



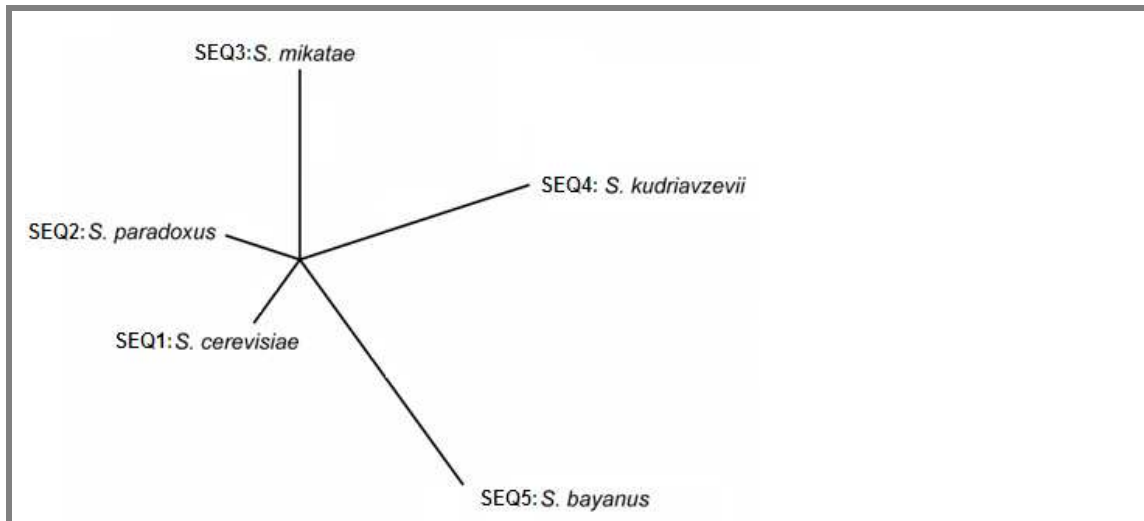


Figure S1 Phylogenetic trees relating the eight Gamma-proteobacterial species (on top) and the five *Saccharomyces* species (below). Both trees (Newick formats in Table S4) are based on a neutral evolution rate and reflect the phylogenetic relatedness between the intergenic sequences of eight Gamma-proteobacterial species (on top) or five *Saccharomyces* species (below). For the tests on real data for the combined and the orthologous space, we added orthologs with increasing phylogenetic distances to the reference species *E.coli* (in case of the bacterial datasets) or *S. cerevisiae* (in case of the yeast datasets). Subsets of these orthologs (reference species included) used throughout the tests were selected as follows:

For the Gamma-proteobacteria:

- 2 orthologs = SEQ1+SEQ4; 4 orthologs = SEQ1+SEQ4+SEQ5+SEQ6;
- 6 orthologs = SEQ1+SEQ4+SEQ5+SEQ6+SEQ2+SEQ3;
- 7 orthologs = SEQ1+SEQ4+SEQ5+SEQ6+SEQ2+SEQ3+SEQ8
- and 8 orthologs = SEQ1+SEQ4+SEQ5+SEQ6+SEQ2+SEQ3+SEQ7+SEQ8.

For the Saccharomyces species:

- 2 orthologs = SEQ1+SEQ2; 4 orthologs = SEQ1+SEQ2+SEQ3+SEQ4; 5 orthologs = SEQ1+SEQ2+SEQ3+SEQ4+SEQ5.

Table S2 explains the composition of the real datasets for the Gamma-proteobacterial and the *Saccharomyces* species.

Table S2 Composition of the real datasets for the Gamma-proteobacterial and the *Saccharomyces* species.

GAMMA-PROTEOBACTERIA								
R	HIGH IC – LexA							
T	LexA	PolB	RecN	RpsU*	SulA	UvrA*	UvrB	UvrD
M	3	1	2	1	1	1	1	1
O	8	8	8	8	6	8	8	8
S ^[1]	/	/	/	/	7,8	/	/	/
R	LOW IC – TyrR							
T	AroF	AroG	AroL	Mtr*	TyrB*	TyrP	TyrR	
M	3	1	5	1	1	2	2	
O	8	5	6	7	7	6	8	
S ^[1]	/	2,7,8	7,8	7	7	7,8	/	

SACCHAROMYCES SPECIES										
R	HIGH IC - URS1H									
T	AGP1	SPO16	REC104	IME2*	REC114*	MEK1	HOP1	MSH5	MRPL27	POP4
M	1	1	1	1	1	2	1	1	1	1
O	5	5	5	5	5	5	5	5	5	5
S ^[2]	/	/	/	/	/	/	/	/	/	/
R	LOW IC - RAP1									
T	HIS4*	RPL11B	ENO1*	OPI3	AVT4	BUD22	RPS2	YEF3	SNF4	RPL5
M	1	1	1	1	1	1	1	1	1	1
O	5	5	5	5	5	5	5	5	5	5
S ^[2]	/	/	/	/	/	/	/	/	/	/

Rows: **R**: the regulator, **T**: the target genes for the regulator used to compose the dataset in the coregulation space in respectively the reference species *E. coli* and *S. cerevisiae*, **M**: number of TF binding sites for the specific regulator in the selected upstream region of those reference genes, **O**: the total number of orthologs that could be used for each target gene in the orthologous or combined space (the number includes the gene in the reference species itself) and **S**: the species for which no ortholog was retrieved. Species names are represented by the following numbers:^[1] The gamma proteobacteria: 1=*Escherichia coli*, 2=*Yersinia pestis*, 3=*Erwinia carotovora*, 4=*Shigella flexneri*, 5=*Salmonella typhimurium*, 6=*Salmonella enterica*, 7=*Vibrio cholerae*, 8=*Pseudomonas aeruginosa*.

^[2] For both, URS1H and RAP1, all 10 target genes in the reference species had orthologs in all 4 other *Saccharomyces* species.

Note that the used version of PS can not handle a dataset for which the MASSES (~a prealigned set of orthologs for one gene) contain a different number of orthologs. Therefore, we left out the gene AroG in the TyrR datasets for the tests with PS. We applied this correction also in the coregulation space to better compare results obtained from the combined and orthologous space with those from the coregulation space. *Target genes in the reference species for respectively LexA, TyrR, URS1H and RAP1: each of those was used together with its respective orthologs for the tests in the orthologous space.

Table S3 shows the effect of different types of phylogenetic trees on the results of the phylogenetic algorithms (PG and PS) for the Gamma-proteobacterial datasets in the combined coregulation-orthology space.

Table S3 The effect of different types of phylogenetic trees on the results of the phylogenetic algorithms (PG and PS) for the Gamma-proteobacterial datasets in the combined coregulation-orthology space.

GAMMA-PROTEOBACTERIA								
SETUP	HIGH IC - LexA				LOW IC - TyrR			
Results of PG								
Tree type	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
Neutral	10	100	98.6	75.5	8	100	96.9	67.5
Protein	2	50	85.7	54.5	2	0	/	/
Corrected	10	90	95.6	72.7	8	75	92.8	61.1
Results of PS								
Tree type	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
Neutral	10	90	69	42.2	10	100	85	36.4
Protein	0	/	/	/	0	/	/	/
Corrected	10	80	71.4	45.5	10	100	100	40

Performance and quality measures: **R1**: the number of runs with an output out of the 10 runs on one real dataset, **RR (%)**: Recovery Rate: the percentage of the output (R1) for which the correct motif was retrieved (correct outputs), **spPPV (%)**: species-dependent PPV: the percentage of true sites among the predicted sites for the reference species, averaged over all correct outputs, **spSens (%)**: species-dependent

Appendix – Supplementary materials

Sens: the percentage of the true sites in the reference species found by the algorithm, averaged over all correct outputs. *E. coli* is the reference species. The dataset of each regulator consists of 8 (LexA) or 7 (TyrR) target genes from the reference species (see Table S2), together with their orthologs in 5 additional species (Figure S1 lists from which species these orthologs were derived). Each reference sequence together with its orthologs was prealigned (6 sequences in total). The phylogenetic relatedness between the orthologous sequences is modeled by one of the three different ‘tree types’: a Neutral, a Protein or a Corrected tree (Newick formats in Table S4).

Additional information on the phylogenetic trees used for the synthetic and real datasets

The construction of the tree used to model the phylogenetic relatedness between the intergenic sequences of the Gamma-proteobacteria was based on coding sequence alignments. Three different trees were obtained : 1) a tree made by PhyML (Guindon and Gascuel, 2003) with as input the alignment of 30 concatenated protein sequences over all eight species (~Protein tree), 2) the branch lengths of the previous tree multiplied by factor 13.5 as described by Newberg *et al.* (Newberg et al., 2007) (~Corrected tree) and 3) a tree based on neutral evolution rates by only taking into account the evolution of the third positions of four fold degenerate codons (~Neutral tree) (kindly provided by Erik Van Nimwegen). As can be seen in Table S3, for both PG and PS, the overall highest recovery rate (RR) was obtained with the tree based on a neutral evolution rate. The tree based on protein alignments did most of the time not result in any output as it overestimates the relatedness between the intergenic sequences. As intergenic sequences are expected to evolve faster than coding sequences, correcting the branch lengths of the protein tree allowed to better approximate the true relatedness between the intergenic sequences and the obtained results became more comparable to those obtained with the neutral tree. In all subsequent analyses, the tree based on the neutral evolution rate was used (Siddharthan et al., 2005). Figure S1 shows the neutral evolution tree for both the Gamma-proteobacterial and the *Saccharomyces* species. The Newick formats of all trees, used for tests on synthetic and real datasets, are given in Table S4.

Table S4 contains the Newick formats for all the phylogenetic trees that were used in the tests on the synthetic and real datasets.

Table S4 Newick format for all the phylogenetic trees used in the tests on the synthetic and real data. The distances are given in proximities (q).

Synthetic data		
Topology		Proximities
Star	Equal distances ^[1]	(SEQ1:0.80,SEQ2:q,SEQ3:q,SEQ4:q,SEQ5:q); (5orthologs) (SEQ1:0.80,SEQ2:q,SEQ3:q,SEQ4:q,SEQ5:q, SEQ6:q,SEQ7:q, SEQ8:q,SEQ9:q,SEQ10:q); (10 orthologs)
	Unequal distances	(SEQ1:0.80,SEQ2:0.90,SEQ3:0.85,SEQ4:0.75,SEQ5:0.20);
Tree (Newberg et al., 2007)		((((SEQ1:0.83,(SEQ2:0.89,SEQ3:0.91):0.84):0.95, (SEQ4:0.97,SEQ5:0.99):0.82):0.93,SEQ6:0.70);

Real data ^[2]		
Gamma-proteobacteria		
Topology		Proximities
Tree	Neutral	(((SEQ1:0.95,SEQ4:0.98):0.61,(SEQ5:0.99,SEQ6:0.96):0.61):0.70,(SEQ2:0.38,SEQ3:0.45):0.95):0.94,(SEQ7:0.25,SEQ8:0.14):0.99);
	Protein	(((SEQ1:0.99,SEQ4:0.99):0.97,(SEQ5:0.99,SEQ6:0.99):0.97):0.94,(SEQ2:0.92,SEQ3:0.91):0.96):0.79,SEQ7:0.76,SEQ8:0.38);
	Corrected ^[3]	(((SEQ1:0.99,SEQ4:0.92):0.68,(SEQ5:0.99,SEQ6:0.97):0.66):0.43,(SEQ2:0.32,SEQ3:0.29):0.58):0.04,SEQ7:0.02,SEQ8:0.0);
Saccharomyces species		
Topology		Proximities
Star	Neutral	(SEQ1:0.80,SEQ2:0.80,SEQ3:0.58,SEQ4:0.50,SEQ5:0.45);

^[1] The value for q varies between 0.90, 0.50 and 0.20 for different tests. For example: q=0.90 describes a phylogenetic tree for very closely related orthologs.

^[2] For the real data we replaced the species names by SEQnumbers. For the **Gamma-proteobacteria**: 1=Escherichia coli, 2=Yersinia pestis, 3=Erwinia carotovora, 4=Shigella flexneri, 5=Salmonella typhimurium, 6=Salmonella enterica, 7=Vibrio cholerae, 8=Pseudomonas aeruginosa. For the **Saccharomyces species**: 1=S. cerevisiae, 2=S. paradoxus, 3=S. mikatae, 4=S. kudriavzevii, 5=S. bayanus,

^[3] The relation between the proximities for the Protein and the Corrected tree is given by: (proximity of the Corrected tree) = (proximity of the Protein tree)^{13.5} as described in (Newberg et al., 2007). The proximities are rounded up to two decimals after the comma.

Table S5 consists of Tables S5 (A, B, C and D) containing the results of PG, PS and MEME in the coregulation space and in the combined coregulation-orthology space for both the synthetic and real datasets.

All data presented in Table S5 (A, B and C), showing the results of the three algorithms on the *synthetic datasets* in the coregulation space and in the combined space are described in the main text. Table S5 (D) shows the results of the three algorithms on the *real datasets* in the coregulation and combined space. For the results on the real data in the coregulation space, a trend similar as for the synthetic datasets in Table S5 (A) was observed, though less pronounced. For both, the Gamma-proteobacterial and yeast datasets, all algorithms retrieved the high IC motif (LexA or URS1H) with a high RR and motif quality. For a low IC motif (TyrR or RAP1) the quality of the motifs retrieved by all three algorithms, especially by PG and PS, was characterized by a very pronounced drop in sensitivity (Sens). As for the synthetic data in Table S5 (A), PG showed for a low IC motif a weaker performance (R1 and RR) compared to PS and MEME.

Table S5 A Results of PG, PS and MEME on synthetic datasets for the ‘star topology with equal distances’ given in proximities (q). The results for the ‘coregulation space’ are given as reference values (REF).

SYNTHETIC DATA								
SETUP	HIGH IC				LOW IC			
Results of PG								
Proximity q	D1	RR	PPV	Sens	D1	RR	PPV	Sens
REF: Coregulation space	44	100	99.3	88.4	75*	5.3*	72.2*	47.5*
0.90	88	100	99.4	98	30	76.7	94.8	72.2
0.50	99	100	100	98.9	74	100	98.5	80.9
0.20	100	100	98.5	76.9	77	94.8	92	33.9
0.20 (unaligned)	95	100	96.7	91	42	97.6	89.9	50

SYNTHETIC DATA								
SETUP	HIGH IC				LOW IC			
Results of PS								
Proximity q	D1	RR	PPV	Sens	D1	RR	PPV	Sens
<i>REF: Coregulation space</i>	100	100	99.4	91.9	36	80.6	94.9	57.2
0.90	100	100	99.8	98.1	32	93.8	97.6	79
0.50	100	100	100	97.9	100	100	99.5	90.6
0.20	100	100	97.9	78.7	97	100	92.7	55.5
0.20 (unaligned)	100	100	99.3	95.8	100	100	92.5	46.4
Results of MEME								
Proximity q	D1	RR	PPV	Sens	D1	RR	PPV	Sens
<i>REF: Coregulation space</i>	100	100	93.1	92.7	100	34	67.4	67.4
0.90	100	100	95.6	94.9	100	67	73.2	73.0
0.50	100	100	94.8	94.5	100	100	68.8	68.8
0.20	100	100	94.2	94.1	100	100	66.9	66.8

Performance and quality measures: **D1**: the number of datasets with an output out of the 100 synthetic datasets, **RR (%)**: Recovery Rate: the percentage of the output (D1) for which the correct motif was retrieved (correct outputs), **PPV (%)**: Positive Predictive Value: the percentage of true sites among the predicted TF binding sites, averaged over all correct outputs, **Sens (%)**: Sensitivity: the percentage of the true sites found by the algorithm, averaged over all correct outputs. * Tracking threshold PG equal to 0.05 (instead of 0.50).

The synthetic datasets contain in ‘*the coregulation space*’ the 10 coregulated sequences from the reference species each of which has one TF binding site embedded. In the ‘combined space’ the datasets contain 10 orthologous sets, (an orthologous set is defined as one reference gene and its orthologs). Each orthologous set contains 5 orthologous sequences that are related through a star topology with equal distances (Newick format in Table S4) and contain one embedded TF binding site per sequence (high IC or low IC). They can be aligned or left unaligned.

Note that in the *coregulation space* true motifs recovered by PG and PS in general exhibit a higher PPV than motifs recovered by MEME, while for the Sens the opposite is true. This is a consequence of the different working regime of the Sens/PPV trade off which in each of the algorithms is being used. For MEME and PS this trade off is fixed and can not be user specified.

Table S5 B Results of PG, PS and MEME algorithms on synthetic datasets for a ‘**star topology with unequal distances**’ (~ four closely related orthologs and one distantly related ortholog).

SYNTHETIC DATA								
SETUP	HIGH IC				LOW IC			
Results of PG								
Unequal star topology	D1	RR	PPV	Sens	D1	RR	PPV	Sens
Only 4 orthologs aligned	91	98.9	99.7	99	36	72.2	94.7	71.2
All 5 orthologs aligned	99	100	99.7	92.7	69	94.2	96.7	64.3
Distant ortholog unaligned	91	100	98.7	96.3	44	90.9	92.6	68.6
Results of PS								
Unequal star topology	D1	RR	PPV	Sens	D1	RR	PPV	Sens
Only 4 orthologs aligned	100	100	99.8	98	59	89.8	96.2	72.6
All 5 orthologs aligned	100	100	99.8	96.9	100	100	98.8	81.9
Distant ortholog unaligned	100	100	99.7	98.3	90	96.7	96.8	67.9

SYNTHETIC DATA								
SETUP	HIGH IC				LOW IC			
Results of MEME								
Unequal star topology	D1	RR	PPV	Sens	D1	RR	PPV	Sens
Only 4 orthologs	100	100	96.4	95.9	100	84	71.5	71.3
All 5 orthologs	100	100	96.1	95.7	100	93	68.4	68.1

Performance and quality measures: idem as in Table S5 A. Each synthetic dataset in the combined space consists of 10 coregulated sequences from the reference species together with their orthologs. Each reference sequence together with its orthologs is referred to as an orthologous set. A synthetic dataset thus consists of 10 orthologous sets. Each orthologous set contains in total 5 orthologs that are related through a star topology with unequal distances (Newick format in Table S4) and contain one embedded TF binding site per sequence (high IC or low IC). **Only 4 orthologs aligned:** we search for motifs in a dataset for which the 10 orthologous sets only contain the 4 closest related orthologs that were aligned. **All 5 orthologs aligned:** the 10 orthologous sets contain 5 prealigned orthologous sequences. **Distant ortholog unaligned:** the 10 orthologous sets contain all 5 orthologs, but only the 4 closely related ones are aligned and the most distant ortholog is left unaligned.

Table S5 C The species-dependent quality parameters for PG, PS and MEME on results obtained in the combined space with synthetic datasets containing sequences related through a **star topology with unequal distances**. This Table complements Table S5 B, with values for the species-dependent PPV and species-dependent sensitivity.

SYNTHETIC DATA								
SETUP	HIGH IC				LOW IC			
Results of PG								
Unequal star topology	PPV	Sens	spPPV	spSens	PPV	Sens	spPPV	spSens
All 5 orthologs aligned	99.7	92.7	98.5	84.9	96.7	64.3	96.5	47.8
Distant ortholog unaligned	98.7	96.3	96.9	92.4	92.6	68.6	92.9	47.8
Results of PS								
Unequal star topology	PPV	Sens	spPPV	spSens	PPV	Sens	spPPV	spSens
All 5 orthologs aligned	99.8	96.9	99.7	96.7	98.8	81.9	98.8	81.9
Distant ortholog unaligned	99.7	98.3	99	94.9	96.8	67.9	95.3	49.4
Results of MEME								
Unequal star topology	PPV	Sens	spPPV	spSens	PPV	Sens	spPPV	spSens
All 5 orthologs	96.1	95.7	98	92.2	68.4	68.1	73	54.5

Performance and quality measures: idem as in Table S5 A, except for spPPV (%): species-dependent PPV: the percentage of true sites among the predicted sites for the reference species, averaged over all correct outputs, spSens (%): species-dependent Sens: the percentage of the true sites in the reference species found by the algorithm, averaged over all correct outputs. For this specific case the reference species equals the distantly related species (proximity 0.20). Each synthetic dataset consists of 10 coregulated genes in the reference species together with their orthologs, thus containing 10 orthologous sets. Each orthologous set contains 5 orthologous sequences that are related through a star topology with unequal distances (Newick format in Table S4) and contain one embedded TF binding site per sequence (high IC or low IC). **All 5 orthologs aligned:** the 10 orthologous sets contain 5 prealigned orthologous sequences. **Distant ortholog unaligned:** the 10 orthologous sets contain all 5 orthologs, but only the 4 closely related ones are aligned and the most distant ortholog is left unaligned.

Appendix – Supplementary materials

Table S5 D Results of PG, PS and MEME on real datasets (**Gamma-proteobacterial** and **Saccharomyces species**) for motif discovery in the ‘combined coregulation-orthology space’. Results for the ‘coregulation space’ are given as reference values (*REF*).

GAMMA-PROTEOBACTERIA								
SETUP	HIGH IC - LexA				LOW IC – TyrR			
Results of PG								
# orthologs	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
<i>REF: coregulation space</i>	10	100	98	81.8	8	100	92	58.3
2	10	100	94	81.8	7	100	91.6	61.9
4	10	100	89	74.5	10	100	91.9	67.3
6	10	100	98.6	75.5	8	100	96.9	67.5
6 (unaligned)	10	100	87.9	81.8	10	50	95.3	62.7
Results of PS								
# orthologs	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
<i>REF: coregulation space</i>	10	100	100	81.8	10	100	100	57.1
2	10	100	96.7	79.1	10	100	100	64.3
4	10	100	90	71.8	10	100	98.9	63.6
6	10	90	69	42.4	10	100	85	36.4
6 (unaligned)	10	100	79.8	84.5	10	100	100	61.4
Results of MEME								
# orthologs	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
<i>REF: coregulation space</i>	10	100	90.9	90.9	10	100	73.3	73.3
2	10	100	90.9	90.9	10	100	80	80
4	10	100	90.9	90.9	10	100	84.6	73.3
6	10	100	100	90.9	10	100	85.7	80
SACCHAROMYCES SPECIES								
SETUP	HIGH IC – URS1H				LOW IC – RAP1			
Results of PG								
# orthologs	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
<i>REF: coregulation space</i>	10	100	93.3	98.2	9*	22.2*	87.5*	20*
2	9	100	96.3	99	7*	28.6*	100*	10*
4	5	100	100	100	5*	40*	70.3*	90*
5	5	100	96	80	4*	75*	67.4*	93.3*
5 (unaligned)	2	50	100	100	8*	37.5*	79.1*	100*
Results of PS								
# orthologs	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
<i>REF: coregulation space</i>	10	100	92.5	100	10	100	87.6	71
2	10	100	91.7	100	10	100	75	60
4	10	100	83.9	81.8	10	100	90	90
5	10	100	84	75.5	10	100	88.8	79
5 (unaligned)	10	100	91.7	100	10	100	87.6	71

SACCHAROMYCES SPECIES								
SETUP	HIGH IC – URS1H				LOW IC – RAP1			
Results of MEME								
# orthologs	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
<i>REF: coregulation space</i>	10	100	100	100	10	100	90	90
2	10	100	100	100	10	100	81.8	90
4	10	100	100	100	10	100	81.8	90
5	10	100	100	100	10	100	81.8	90

Performance and quality measures: **R1:** the number of runs with an output out of the 10 runs on one real dataset, **RR (%)**: Recovery Rate: the percentage of the output (R1) for which the correct motif was retrieved (correct outputs), **spPPV (%)**: species-dependent PPV: the percentage of true sites among the predicted sites for the reference species, averaged over all correct outputs, **spSens (%)**: species-dependent Sens: the percentage of the true sites in the reference species found by the algorithm, averaged over all correct outputs. The reference species equals *E. coli* (bacterial data) or *S. cerevisiae* (yeast data).

Gamma-proteobacteria: The dataset of each regulator consists of 8 (LexA) or 7 (TyrR) target genes from the reference species (Table S2), together with their orthologs selected from additional species. In total we have “# orthologs” orthologs per gene (the ortholog of the reference species included), all prealigned or all left unaligned, related through the neutral tree for the Gamma-proteobacteria. **Saccharomyces species:** The dataset of each regulator (URS1H, RAP1) consists of 10 target genes from the reference species (Table S2), together with their orthologs selected from additional species. In total we have “# orthologs” orthologs per gene (the ortholog of the reference species included), all prealigned or all unaligned, related through the neutral tree for the *Saccharomyces* species. Table S4 shows the Newick formats of both trees and Figure S1 lists which species are used for each ‘# orthologs’ for both the bacterial and yeast species. * Tracking threshold PG equal to 0.05 (instead of 0.50).

Table S6 shows the effect of leaving out motif sites in both the coregulation space and the combined coregulation-orthology space for a synthetic dataset containing sites sampled from a high IC motif.

Table S6 The effect of leaving out TF binding sites in both the ‘coregulation space’ (on top) and in the ‘combined coregulation-orthology space’ (below). Results are displayed for a synthetic dataset containing sites sampled from a high IC motif.

SYNTHETIC DATA					
COREGULATION SPACE	SETUP	HIGH IC			
	Results of PG				
	Number of TF binding sites/gene	D1	RR	PPV	Sens
	<i>Ref: 10(1,1,1,1,1,1,1,1,1,1)</i> ^[a]	44	100	99.3	88.4
	10(1,1,1,1,1,1,1,1,0,0) ^[b]	29	100	96.9	87.1
	Results of PS				
	Number of TF binding sites/gene	D1	RR	PPV	Sens
	<i>Ref: 10(1,1,1,1,1,1,1,1,1,1)</i> ^[a]	100	100	99.4	91.9
	10(1,1,1,1,1,1,1,1,0,0) ^[b]	100	100	96.9	92.4
	Results of MEME				
	Number of TF binding sites/gene	D1	RR	PPV	Sens
	<i>Ref: 10(1,1,1,1,1,1,1,1,1,1)</i> ^[a]	100	100	93.1	92.7
	10(1,1,1,1,1,1,1,1,0,0) ^[b]	100	99	77.7	97

COMBINED COREGULATION-ORTHOLOGY SPACE	SETUP		HIGH IC			
	Results of PG					
	Number of TF binding sites/gene, ortholog		D1	RR	PPV	Sens
	<i>Ref: all genes and all orthologs contain the motif site</i> ^[a]		99	100	99.7	92.7
	TF binding site absent in distant ortholog (q=0.20) for all		76	100	97.6	91.4
	TF binding site absent in close ortholog (q=0.75) for all		77	88.3	86.2	63.3
	TF binding site absent in all orthologs for two out of ten		97	100	98.2	96.6
	Results of PS					
	Number of TF binding sites/gene, ortholog		D1	RR	PPV	Sens
	<i>Ref: all genes and all orthologs contain the motif site</i> ^[a]		100	100	99.8	96.9
	TF binding site absent in distant ortholog (q=0.20) for all		18	83.3	78.1	62.7
	TF binding site absent in close ortholog (q=0.75) for all		99	99	79.3	88.3
	TF binding site absent in all orthologs for two out of ten		100	100	99.5	96.5
	Results of MEME					
	Number of TF binding sites/gene, ortholog		D1	RR	PPV	Sens
	<i>Ref: all genes and all orthologs contain the motif site</i> ^[a]		100	100	96.1	95.7
TF binding site absent in distant ortholog (q=0.20) for all		100	100	79.0	98.6	
TF binding site absent in close ortholog (q=0.75) for all		100	100	77.7	96.8	
TF binding site absent in all orthologs for two out of ten		100	100	77.9	97.1	

Performance and quality measures: **D1:** the number of datasets with an output out of the 100 synthetic datasets, **RR (%)**: Recovery Rate: the percentage of the output (D1) for which the correct motif was retrieved (correct outputs), **PPV (%)**: Positive Predictive Value: the percentage of true sites among the predicted TF binding sites, averaged over all correct outputs, **Sens (%)**: Sensitivity: the percentage of the true sites found by the algorithm, averaged over all correct outputs. **Number of motif sites/gene** [e.g. 10(1,1,1,1,1,1,1,1,0,0)]: the number before the brackets represents the total number of coregulated sequences for the reference species present in the dataset. The numbers between brackets indicate gene per gene the number of embedded TF binding sites in the gene's intergenic sequences. For 'the coregulation space' each synthetic dataset consists of 10 sequences from the reference species with one/zero TF binding site embedded. For the 'combined space' each synthetic dataset contains 10 orthologous sets (an orthologous set is defined as one reference sequence and its orthologs). Each orthologous set consists in total of 5 prealigned orthologs, related through a **star topology with unequal distances** (Newick format in Table S4 on the Supplementary) and one/zero embedded TF binding site per sequence. Note that for MEME the orthologs are always unaligned.

^[a] Ask the motif discovery algorithm to search for the exact number of TF binding sites present in the dataset.

^[b] Ask the motif discovery algorithm to search for 1 TF binding site per sequence, hereby slightly overestimating the number of TF binding sites present in the dataset.

As mentioned in chapter 2, the prealignment of the orthologous sequences plays a role in how missing TF binding sites affect the motif discovery results in the combined space, more specifically when TF binding sites were omitted in all the sequences of one of the species that were added as additional orthologous information to the coregulated gene set of the reference species. Here we provide a more in depth explanation why this is the case.

The performance of PG was most deteriorated if the TF binding sites were omitted in the sequences from a closely related species while for PS the performance was most affected if the TF binding sites were absent in the sequences of a distantly related species.

The local alignment strategy used in combination with PG will leave the distant orthologs that shows low similarity with the closely related orthologs unaligned. PG thus will only

take into account the alignment of the closely related orthologs and will treat the distant orthologs as independent, unaligned sequences. However, when TF binding sites are missing in the sequences of a closely related species, this might interfere with finding the correct alignment of the orthologous TF binding sites present in the other closely related sequences. If so, these TF binding sites will not longer be captured in the same window and this will result in a decrease in sensitivity of the retrieved motifs. Misalignment will also increase the chance to capture a false positive site in the window, resulting in a decrease of the PPV (see Table S6). The effect of a false positive TF binding site in a window is dependent on the phylogenetic distance of the ortholog for which this site was retrieved: a window will be more penalized during scoring when containing a false, non conserved site present in a closely related ortholog than when present in a distantly related one. This explains why PG is more sensitive to the presence of noisy sequences in closely related species.

For PS only the regions that are gaplessly aligned over all species in the alignment are considered as potential TF binding sites (blocks). A missing TF binding site will result in rejection of a block or it will be replaced by a false positive TF binding site. This last effect can be observed in the Table S6_species_specific, by comparing the overall PPV (over all species) to the species-dependent PPV (= the PPV in the reference species that has a proximity of 0.80): the overall PPV is 80% of the species-dependent PPV, indicating that one site in a block of five TF binding sites is indeed a false positive one. Moreover, the absence of TF binding sites in a distant ortholog interferes more in obtaining a good global alignment than when absent in a close ortholog that aligns any way well over the remainder of its sequence. This explains why PS can cope better with the absence of TF binding sites in closely related orthologs than in distant orthologs

Table S6_species_specific Species-dependent motif quality parameters for the results obtained by PS in the ‘combined coregulation-orthology space’ when leaving out TF binding sites. Results are displayed for a synthetic dataset containing sites sampled from a high IC motif.

Unequal star topology with	PPV	Sens	spPPV	spSens
TF binding site absent in distant ortholog (q=0.20) for all genes ^[b]	78.1	62.7	97.7	62.7
TF binding site absent in close ortholog (q=0.75) for all genes ^[b]	79.3	88.3	99.2	88.4

Performance and quality measures: idem as in Table S6 except for the spPPV (=species-dependent PPV) and spSens (=species-dependent Sensitivity), both measured for the reference species (q=0.80).

Each synthetic dataset contains 10 orthologous sets (an orthologous set is defined as one reference sequence and its orthologs). Each orthologous set consists in total of 5 prealigned orthologs, related through a **star topology with unequal distances** (Newick format in Table S4 on the Supplementary) and one/zero embedded TF binding site per sequence. ^[b] Ask the motif discovery algorithm to search for 1 TF binding site per sequence, hereby slightly overestimating the number of TF binding sites present in the dataset.

Table S7 gives the results of the phylogenetic algorithms when using orthologs related through a non star like tree topology for the synthetic data in the combined coregulation-orthology space.

Table S7 Results of the phylogenetic algorithms when using orthologs related through a **non star like topology** for the synthetic datasets in the combined coregulation-orthology space.

SYNTHETIC DATA								
SETUP	HIGH IC				LOW IC			
Results of PG								
Number of TF binding	D1	RR	PPV	Sens	D1	RR	PPV	Sens
10(1,1,1,1,1,1,1,1,1,1)	97	100	99.8	99.8	63	90.5	95.4	78.7
Results of PS								
Number of TF binding	D1	RR	PPV	Sens	D1	RR	PPV	Sens
10(1,1,1,1,1,1,1,1,1,1)	100	100	99.8	99.5	83	95.2	98.6	80.1

Performance and quality measures: **D1:** the number of datasets with an output out of the 100 synthetic datasets, **RR (%)**: Recovery Rate: the percentage of the output (D1) for which the correct motif was retrieved (correct outputs), **PPV (%)**: Positive Predictive Value: the percentage of true sites among the predicted TF binding sites, averaged over all correct outputs, **Sens (%)**: Sensitivity: the percentage of the true sites found by the algorithm, averaged over all correct outputs. Each synthetic dataset consists of 10 coregulated genes in the reference species together with their orthologs, thus containing 10 orthologous sets. Each orthologous set contains 6 prealigned orthologous sequences that are related through a non star like topology (Newick format in Table S4) and contain one embedded TF binding site per sequence (high IC or low IC).

The effect of using orthologs related through a non star like topology on the accuracy of the phylogenetic algorithms

An intrinsic property of PG is that it can only handle star topologies directly during the score calculation of phylogenetically related TF binding sites as a result of the way it calculates these posterior probability scores (see chapter 2, Table 2.1: ‘Scoring’). To calculate the posterior probabilities the algorithm solves the integral over all possible motif WMs. When an evolutionary model is included, this integral is solved by making an approximation that requires a star topology tree. Any other topology, deviating from a star topology needs to be converted to a collection of star topologies first. PS on the contrary can cope directly with different topologies. It uses a conditional probability for which the motif WM is known and no integration is needed. Moreover, it uses the Felsenstein tree-likelihood algorithm where internal nodes are allowed.

To test whether this intrinsic difference between both algorithms in treating topologies deviating from a star topology has an effect on the performance, we ran both algorithms on synthetic datasets in the combined coregulation-orthology space that exhibited a non star like topology (created by using the phylogenetic tree described in Newberg *et al.* (Newberg *et al.*, 2007)). Each dataset contains 10 coregulated reference genes, each supplemented with 5 additional orthologs. The evolutionary distances are sufficiently close to guarantee that the intergenic sequences can reliably be aligned. Results are shown in Table S7.

These results suggest that for this non star like topology PS seemingly outperforms PG. Motif discovery resulted in more datasets with an output, a slightly higher recovery rate (RR) and a slightly higher quality for the datasets with a correct output, given by the values of the PPV and sensitivity. However, this better performance of PS over PG was

also seen for star topologies in the combined coregulation-orthology space (see Table S5 A and B). So it might be a general tendency observed for all topologies and does not really prove that PS is less sensitive than PG in handling topologies different from a star. This observation was also confirmed by the results on the real datasets which have a non star like topology and for which both algorithms showed comparable results. Conclusively, using a topology different from a star does not result in striking differences between PG and PS in retrieving the true motifs.

Table S8 consists of Tables S8 (A and B) containing the results of PG, PS and MEME in the orthologous space.

Table S8 (A) shows the results of the three algorithms on the *synthetic datasets* in the orthologous space and Table S8 (B) shows the results of the three algorithms on the *real datasets* in the orthologous space. These results were all described in the main text.

Table S8 A Results of PG, PS and MEME on synthetic datasets in the orthologous space.

SYNTHETIC DATA								
SETUP	HIGH IC				LOW IC			
Results of PG								
Topology - # orthologs	D1	RR	PPV	Sens	D1	RR	PPV	Sens
EST (0.50) - 5	66	100	100	100	6	100	100	100
EST (0.50) - 10	100	100	100	100	88	98.9	100	97.1
EST (0.50) - 10 (unaligned)	62	100	99.4	93.5	20	85	98.9	73.5
EST (0.90) - 5	60*	38.3*	100*	100*	45*	8.89*	100*	100*
EST (0.90) - 10	97*	75.3*	100*	100*	87*	27.6*	100*	100*
EST (0.90) - 10	96*	64.6*	97.5*	100*	89*	34.8*	96*	99.7*
UEST- 4	77*	48.1*	100*	100*	67*	25.4*	100*	100*
UEST - 5 (+distant)	94*	83*	99.7*	99.7*	66*	51.5*	98.2*	98.2*
Results of PS								
Topology - # orthologs	D1	RR	PPV	Sens	D1	RR	PPV	Sens
EST (0.50) - 5	0	/	/	/	0	/	/	/
EST (0.50) - 10	7	100	100	100	2	100	100	100
EST (0.50) - 10 (unaligned)	100	100	100	94.4	79	94.9	99.4	66.3
EST (0.90) - all settings	0	/	/	/	0	/	/	/
UEST - all settings	0	/	/	/	0	/	/	/
Results of MEME								
Topology - # orthologs	D1	RR	PPV	Sens	D1	RR	PPV	Sens
EST (0.50) - 5	100	98	96.7	96.7	100	48	86.6	86.6
EST (0.50) - 10	100	100	97.4	97.4	100	75	82.9	82.9
EST (0.90) - 5	100	17	94.1	94.1	100	6	93.3	93.3
EST (0.90) - 10	100	59	90.2	90.2	100	20	83.5	83.5
UEST- 4	100	38	98.7	98.7	99	18.2	90.3	90.3
UEST - 5 (+distant)	100	64	93.4	93.4	100	18	82.2	82.2

Performance and quality measures: **D1:** the number of datasets with an output out of the 100 synthetic datasets, **RR (%):** Recovery Rate: the percentage of the output (D1) for which the correct motif was retrieved (correct outputs), **PPV (%):** Positive Predictive Value: the percentage of true sites among the predicted TF binding sites, averaged over all correct outputs, **Sens (%):** Sensitivity: the percentage of the

Appendix – Supplementary materials

true sites found by the algorithm, averaged over all correct outputs. Each synthetic dataset contains 1 single orthologous set consisting of ‘# orthologs’ orthologs (i.e. one sequence of the reference species together with all its orthologs), either all being prealigned or all left unaligned, related through a ‘Topology’ topology. EST = Equal Star Topology with proximity q equal to 0.50 or 0.90 and UEST = Unequal Star Topology with four closely related orthologs and one distantly related ortholog. For the Newick formats of the equal and unequal star topology see Table S4. * Tracking threshold PG equal to 0.05 (instead of 0.50). Note that for MEME all orthologs are unaligned.

Table S8 B Results of PG, PS and MEME on real datasets (**Gamma-proteobacterial** and **Saccharomyces species**) in the orthologous space.

GAMMA-PROTEOBACTERIA									
SETUP		HIGH IC - LexA				LOW IC - TyrR			
Results of PG									
	# orthologs	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
G1	6	10	0	/	/				
	8(LexA)/7(TyrR)	10	0	/	/	10*	0*	/*	/*
	8/7 (unaligned)	7	0	/	/	7*	14.3*	50*	100*
G2	6	10	100	100	100				
	8(LexA)/7(TyrR)	10	100	100	100	10*	100*	100*	100*
	8/7 (unaligned)	5	40	100	100	9*	88.9*	49.4*	100*
Results of PS									
	# orthologs	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
G1	6	3	0	/	/				
	8(LexA)/7(TyrR)	0	/	/	/	0	/	/	/
	8/7 (unaligned)	2	0	/	/	8	/	/	/
G2	6	0	/	/	/				
	8(LexA)/7(TyrR)	0	/	/	/	0	/	/	/
	8/7 (unaligned)	1	100	100	100	8	/	/	/
Results of MEME									
	# orthologs	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
G1	6	10	0	/	/				
	8(LexA)/7(TyrR)	10	0	/	/	10	0	/	/
G2	6	10	100	100	100				
	8(LexA)/7(TyrR)	10	100	100	100	10	100	50	100
SACCHAROMYCES SPECIES									
SETUP		HIGH IC – URS1H				LOW IC – RAPI			
Results of PG									
	# orthologs	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
G1	4	0*	/*	/*	/*	10*	0*	/*	/*
	5	0*	/*	/*	/*	10*	0*	/*	/*
	5 (unaligned)	4*	0*	/*	/*	9*	88.9*	100*	100*
G2	4	10*	100*	100*	100*	9*	0*	/*	/*
	5	10*	100*	100*	100*	10*	0*	/*	/*
	5 (unaligned)	10*	100*	100*	100*	10*	0*	/*	/*

SACCHAROMYCES SPECIES									
SETUP		HIGH IC – URS1H				LOW IC – RAP1			
Results of PS									
	# orthologs	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
G1	4	0	/	/	/	0	/	/	/
	5	0	/	/	/	4	0	/	/
	5 (unaligned)	0	/	/	/	4	50	100	100
G2	4	0	/	/	/	0	/	/	/
	5	0	/	/	/	0	/	/	/
	5 (unaligned)	4	25	100	100	0	/	/	/
Results of MEME									
	# orthologs	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
G1	4	10	0	/	/	10	100	100	100
	5	10	0	/	/	10	100	100	100
G2	4	10	100	100	100	10	0	/	/
	5	10	100	100	100	10	0	/	/

Performance and quality measures: **R1**: the number of runs with an output out of the 10 runs on one real dataset, **RR (%)**: Recovery Rate: the percentage of the output (R1) for which the correct motif was retrieved (correct outputs), **spPPV (%)**: species-dependent PPV: the percentage of true sites among the predicted sites for the reference species, averaged over all correct outputs, **spSens (%)**: species-dependent Sens: the percentage of the true sites in the reference species found by the algorithm, averaged over all correct outputs. The reference species equals *E. coli* (bacterial data) or *S. cerevisiae* (yeast data). Each real dataset contains one single orthologous set consisting of ‘# orthologs’ orthologs (i.e. one sequence of the reference species together with all its orthologs), either all being prealigned or all left unaligned, related through a neutral species tree (Newick formats for the Gamma-proteobacterial or Saccharomyces species trees in Table S4). Figure S1 lists which species were used for each ‘# orthologs’ for both the bacterial and yeast species. We generated for each regulator two test sets. **Gamma-proteobacteria**: for LexA we selected respectively the targets G1=RpsU and G2=UvrA and for TyrR the targets G1=Mtr or G2=TyrB, each time complemented with their respective orthologs. For TyrR targets, orthologs were only retrieved for 7 out of the 8 Gamma-proteobacterial species. **Saccharomyces species**: for URS1H we selected the targets G1= IME2 or G2= REC114 and for the regulator RAP1 the targets G1= HIS4 or G2= ENO1, each time complemented with their respective orthologs. Targets in the reference species contained exactly one TF binding site for the regulator in their intergenic region.

* Tracking threshold PG equal to 0.05 (instead of 0.50). For MEME all orthologs are unaligned.

Note that for the real data in the orthologous space it is hard to judge on results with a RR=0%, as this might both refer to the discovery of a false positive motif or the discovery of a true positive different from the annotated motif (so the presence of a second more strong local optimum).

Text S1 provides additional information on the construction of the synthetic and real datasets.

Synthetic datasets

Synthetic motif weight matrices (WMs) were constructed as in Siddharthan *et al.* (Siddharthan *et al.*, 2005); for each position in the motif WM we picked a random “consensus” nucleotide, set the probability of that nucleotide to p and set the probabilities of the other nucleotides to $(1-p)/3$, where p is called the polarization of the motif WM. We created two different motif WMs: 1) a high IC motif WM of width 13 bp with $p=0.90$ for each position and 2) a more degenerated, low IC motif WM of width 13 bp with for each position $p=0.75$. TF binding sites were sampled from both motif WMs to create input sequences containing respectively a high or low IC motif. We embedded each

sampled TF binding site at a randomly chosen position in a background sequence of length 500 bp that was randomly generated. Each ancestral sequence (~a background sequence containing an embedded TF binding site) was then evolved along a phylogenetic tree under a defined evolutionary model to create phylogenetically related sequences. For the background sequence (whole sequence except the TF binding sites) we used the Jukes and Cantor (JC) model (Jukes and Cantor, 1969), for the embedded TF binding sites an adapted Felsenstein (F81) model (Sinha et al., 2003). To simulate evolution under the JC model we used the software Rose (Stoye et al., 1998). Rose creates, guided by a phylogenetic tree and an evolutionary model, a family of evolutionary related sequences starting from an ancestral sequence by insertion, deletion and substitution of characters. The branch length of the Rose input tree equals the expected number of substitutions per 100 sites (default branch length multiplied by 100). By setting the ‘mutation probability’ parameter to 1.0 for all sites in the ancestral sequence, the number of substitutions introduced for each 100 sites of the sequence on average equals the branch lengths of the tree. Of course, the number of observed mutations might be smaller due to back-mutations, especially if the branch length is large. The insertion and deletion threshold was set to zero. For the embedded TF binding sites we simulated the evolution according to the adapted F81 evolutionary model; we used the approach described in Siddharthan *et al.* (Siddharthan et al., 2005). Evolutionary related TF binding sites were created starting from the embedded ancestral TF binding site following a phylogenetic tree with proximities (q). For each position in the orthologous TF binding site, the probability of finding a nucleotide equal to the ancestral nucleotide is q and the probability of a mutation is $(1-q)$. When the ancestral nucleotide was mutated, it was replaced by a new nucleotide sampled from the motif WM.

Real datasets

For the real data we constructed datasets containing a high IC or a more degenerated, low IC motif for both the Gamma-proteobacterial and for the *Saccharomyces* species. To construct the datasets for the Gamma-proteobacteria in the coregulation space, we selected according to RegulonDB (Huerta et al., 1998), genes in *Escherichia coli* that contain at least one annotated TF binding site for respectively the regulators LexA and TyrR. To extend the LexA and TyrR datasets in the combined space, we searched for each of their target genes the corresponding orthologs in other Gamma-proteobacteria (*Shigella flexneri*, *Salmonella typhimurium*, *Salmonella enterica*, *Yersinia pestis*, *Erwinia carotovora*, *Vibrio Cholerae* and *Pseudomonas aeruginosa*). For the ortholog discovery we used the reciprocal smallest distance approach (RSD) (Wall et al., 2003). For the orthologous space we selected *E. coli* genes that have exactly one annotated LexA or TyrR TF binding site and their corresponding orthologs. For the yeast datasets in the coregulation space, we selected target genes in *Saccharomyces cerevisiae* for respectively the regulators URS1H and RAP1 based on annotated TF binding sites in SCPD (Zhu and Zhang, 1999) and SwissRegulon (Pachkov et al., 2007). Corresponding orthologs for all target genes were retrieved from the *Saccharomyces* Genome Database (SGD project. "*Saccharomyces* Genome Database" <http://www.yeastgenome.org/>, accessed 14 March 2009) in case of *Saccharomyces paradoxus*, while for *Saccharomyces mikatae*, *Saccharomyces kudriavzevii* and *Saccharomyces bayanus* we used data from the Washington University group (Cliften et al., 2003). For the orthologous space we selected *S. cerevisiae* genes that have exactly one annotated URS1H or RAP1 TF binding

site and their corresponding orthologs. Table S2 gives the exact composition for each real dataset.

Text S2 provides additional information on the pre-processing of the alignments.

Both phylogenetic algorithms need as input a prealignment of the orthologous sequences in order to account for their phylogenetic relatedness by the use of an evolutionary model (see chapter 2). In this section of the supplementary we describe the results for using PS in combination with the local alignment strategy Dialign (Morgenstern, 1999) instead of the default global alignment strategy ClustalW (Chenna et al., 2003).

For datasets that contain easy to align sequences (e.g. the synthetic datasets with sequences of equal length) the global and local alignment strategies perform equally well and result in similar high quality alignments. So in those cases there is no difference between using ClustalW or Dialign. For the more difficult to align sequences (which in our analysis corresponded to the bacterial and yeast datasets that consisted of intergenic sequences of different length), the local alignment strategy can perform better than a global one.

To rule out that for those difficult to align sequences, the difference in performance we observed between PG and PS was due to the difference in the used alignment procedure rather than in the intrinsically different way both algorithms handle the alignments, we also run tests of PS with alignments obtained by Dialign (results in Table A below). When doing so, the regions that are well aligned over all sequences are extracted from the local prealignment prior to providing them as input to PS. Using a local alignment strategy did not improve the results obtained by PS compared to using ClustalW. When using a local alignment strategy on those difficult to align datasets most regions will be left unaligned which usually improves the quality of the prealignment. However, as PS can only search for motifs in the regions that are well aligned over all orthologs, the prealignment does not longer contain input information for PS. Therefore, it is often more advantageous to use PS in combination with ClustalW than with Dialign (which we therefore did in the remainder of the analysis).

Appendix – Supplementary materials

Table A Results of PS when using a local alignment strategy on the difficult to align real datasets in the combined coregulation-orthology space.

GAMMA-PROTEOBACTERIA								
SETUP	HIGH IC - LexA				LOW IC – TyrR			
Results of PS								
# orthologs = 6	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
<i>Global alignment strategy</i>	10	90	69	42.4	10	100	85	36.4
Local alignment strategy	10	100	100	33.6	0	/	/	/
SACCHAROMYCES SPECIES								
SETUP	HIGH IC – URS1H				LOW IC – RAP1			
Results of PS								
# orthologs = 5	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
<i>Global alignment strategy</i>	10	100	84	75.5	10	100	88.8	79
Local alignment strategy	10	100	16	11.8	0	/	/	/

Performance and quality measures: **R1**: the number of runs with an output out of the 10 runs on one real dataset, **RR (%)**: Recovery Rate: the percentage of the output (R1) for which the correct motif was retrieved (correct outputs), **spPPV (%)**: species-dependent PPV: the percentage of true sites among the predicted sites for the reference species, averaged over all correct outputs, **spSens (%)**: species-dependent Sens: the percentage of the true sites in the reference species found by the algorithm, averaged over all correct outputs. The reference species equals *E. coli* or *S. cerevisiae*. **Gamma-proteobacteria:** The dataset of each regulator consists of 8 (LexA) or 7 (TyrR) target genes from the reference species (Table S2), together with their orthologs in 5 additional species (Figure S1 lists from which species these orthologs were derived). Each reference sequence together with its orthologs was prealigned so in total we have 6 prealigned orthologs, related through the neutral tree for the Gamma-proteobacteria. **Saccharomyces species:** The dataset of each regulator (URS1H and RAP1) consists of 10 target genes from the reference species (Table S2), together with their orthologs in 4 additional species (Figure S1 lists from which species these orthologs were derived). Each reference sequence together with its orthologs was prealigned so in total we have 5 prealigned orthologs, related through the neutral tree for the *Saccharomyces* species. Table S4 shows the Newick formats of both trees. We provide **as input to PS** the regions of a local prealignment (Dialign, threshold = 2) that were conserved over all 6 or 5 orthologs. As *reference* also the results of PS are given that were obtained with the full ClustalW prealignment.

Text S3 provides additional information on the parameters of PG and PS.

As both phylogenetic algorithms have some more specific parameters to integrate e.g. the phylogeny, we discuss their parameters in more detail in this section. Table B shows the parameter settings used in this study for PG and PS for all performed tests on synthetic and real datasets and is followed by a description of these parameters.

Table B The parameter settings for PG and PS on the synthetic and real datasets.

PG			
Parameter	Symbol	Synthetic datasets	Real datasets
Input tree	-L	Phylogenetic tree in Newick format, distances given by proximities	Gamma-proteobacterial tree or <i>Saccharomyces</i> species tree both with proximities
Order of Markov model for background probabilities	-N	-1	1 (default)
Alignment level	-D	0: unaligned sequences 1: aligned sequences	0: unaligned sequences 1: aligned sequences

Appendix – Supplementary materials

PG			
Parameter	Symbol	Synthetic datasets	Real datasets
Motif width	-m	13	20 (LexA) 18 (TyrR) 13 (URS1H) 10 (RAP1)
Expected number of windows (unless mentioned differently in tables)	-I	- unaligned: number of embedded TF binding sites in the input data - aligned/partially aligned: number of embedded TF binding sites in the input data divided by the number of orthologs per gene	-unaligned: number of annotated TF binding sites present in <i>E. coli/S. cerevisiae</i> multiplied by the number of orthologs per gene -aligned: number of annotated TF binding sites present in <i>E.coli/S. cerevisiae</i>
Palindromic motif	-C	No palindromic TF binding sites	-No C for LexA, URS1H and RAP1 -C for TyrR (see Table C)
Tracking threshold (unless mentioned differently in tables)	-E	0.50	0.50
Number of cycles during annealing/tracking	-S	100	100
Reverse complement	-r	Use r (only search on the forward strand)	Use r (only search on the forward strand)
PS			
Parameter	Symbol	Synthetic datasets	Real datasets
Input tree	Tree*	For each MASS a separate phylogenetic tree relating only the species used in that MASS. Newick format with distances given by branch lengths.	For each MASS a Gamma-proteobacterial tree or a <i>Saccharomyces</i> species tree, only relating the species for that MASS (branch lengths).
Sequences weights	Weights*	Made by Seq.weights.pl	Made by Seq.weights.pl
Background composition model	-B	Made by unifiedcpp.opteron	Made by unifiedcpp.opteron
Alignment of the centroid TF binding sites	-Align_centroid	Use Align_centroid	Use Align_centroid
Reverse complement	-r	Use r (only search on the forward strand)	Use r (only search on the forward strand)
Palindromic motif	-R	No palindromic TF binding sites	No -R for LexA, URS1H and RAP1 -R 1,1,9 for TyrR (see Table C)
Motif width	/	13	20 (LexA) 18 (TyrR) 13 (URS1H) 10 (RAP1)

PS			
Parameter	Symbol	Synthetic datasets	Real datasets
Expected number of TF binding sites (unless mentioned differently in tables)	/	Number of embedded TF binding sites in the input data	Number of annotated TF binding sites in <i>E. coli/S. cerevisiae</i> multiplied by the number of orthologs per gene
Maximum number of TF binding sites per sequence	-E	1	Maximum number of annotated TF binding sites present in one of the genes in the input file
Prior distribution on the number of TF binding sites per sequence	Blocks*	0.50 0.95	No prior info
Bayesian sampling	-bayes	2000,8000	2000,8000
Number of seeds (re-initializations)	-S	20	20
Nucleotide alphabet	-n	Use n	Use n

*For PS the prior information is gathered in a ‘prior file’ (-P). In this prior file, terms as ‘Tree’, ‘Weights’ and ‘Blocks’ are used to specify the parameter for which additional information was provided (see also in the description of the parameters).

The description of the parameters of both phylogenetic motif discovery algorithms

Parameters describing the motifs:

both algorithms need an initial guess on *the expected number of motifs* present in the input sequences and for each motif *the expected number of motif sites*. This information is captured in the parameter –I for PG and for PS this information is mentioned just after the program name in the command line. The format to describe this information is for both algorithms the same, e.g. ‘10,10’ when we expect two different motifs, each with 10 expected TF binding sites. In this work in principle all described datasets contain exactly one motif. Both algorithms were therefore asked to search exactly for one motif model per dataset. For PG, the expected number of TF binding sites is described by the expected number of ‘windows’ (see chapter 2, Table 2.1: ‘Motif model’). For a dataset that contains only unaligned sequences, a window always equals a single TF binding site (~single-species windows) while for a dataset with prealigned sequences, a window equals or a single unaligned TF binding site or a set of multiple aligned TF binding sites (~multi-species windows). So to define the expected number of windows the user has to take into account if the sequences in the dataset are prealigned or not. For unaligned sequences the expected number of windows equals the total number of embedded TF binding sites for the synthetic data and the number of TF binding sites in *E.coli/S. cerevisiae* multiplied by the number of orthologs for the real data (assuming that all orthologous genes contain the same number of TF binding sites). For prealigned sequences the expected number of windows was set to the number of embedded TF binding sites divided by the number of orthologs for the synthetic data and for the real data to the number of TF binding sites in *E. coli/S. cerevisiae*.

For PS, the expected number of TF binding sites present in the input sequences was chosen equal to the total number of embedded TF binding sites for the synthetic data and for the real data equal to the number of TF binding sites in *E. coli/S. cerevisiae* multiplied by the number of orthologs. PS has two extra parameters compared to PG concerning prior information on the number of TF binding sites per sequence. The first parameter is *-E*, the maximum number of motif sites per sequence. This parameter was set to one for the synthetic data and equal to the highest number of TF binding sites present in one of the input sequences for the real data. The second parameter describes the prior probabilities for finding zero, one until *-E* sites per sequence. We set the prior probabilities for finding zero and one TF binding site per sequence to respectively 0.50 and 0.95 for the synthetic data, while for the real data we used uniform probabilities to find zero, one until *-E* TF binding sites per sequence. For the synthetic data a prior probability for finding zero TF binding sites per sequence equal to 0.50 had a positive influence in the presence of noisy sequences (sequences without TF binding sites), but did not deteriorate the performance of detecting the true motifs in the absence of noise (data not shown). This prior information is provided to the algorithm through a prior file (*-P*) that can contain different types of prior information each labeled by a short term, in this case '*>Blocks*'.

Other prior information on the motif is the *motif width* (number of conserved motif positions): for both algorithms this parameter equaled the total motif length (13 bp for the synthetic data, 20 bp respectively 18 bp for LexA and TyrR and 13 bp respectively 10 bp for URS1H and RAP1). This motif width is not restrictive for PS, because for the default settings of the algorithm the *fragmentation option -F* is turned on which means that PS allows conserved motif positions to be interrupted by degenerated motif positions and as such alters and optimizes the length of the motif.

All algorithms allow the option to search for special motif types such as palindromic motifs. In Table C (beneath) we tested the effect of using a specific *model for palindromic motifs* or not, on the real LexA and TyrR datasets. We found that for both phylogenetic algorithms, the TyrR datasets gave the best results when choosing for a palindromic model, while for the LexA datasets the opposite was true (probably because the degenerate spacer has less palindromic properties). So for the TyrR datasets we used parameter *-C* for PG and parameter *-R1,1,9* for PS (*-R1,1,9* indicates that the first (1) motif model we search for is palindromic in positions 1 through 9, implying automatically a corresponding position, the same distance away from the opposite end of the motif that is also palindromic). For the synthetic and yeast data (URS1H and RAP1), the TF binding sites were all non-palindromic.

To search for motifs on one strand only and not on the complementary strand we used parameter *-r* for both algorithms.

Table C The effect of using a palindromic motif model for PG, PS and MEME when searching for the LexA and TyrR motif in the **Gamma-proteobacterial** datasets in the coregulation space.

GAMMA-PROTEOBACTERIA								
SETUP	HIGH IC - LexA				LOW IC - TyrR			
Results of PG								
Model	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
Palindromic	9	100	96.4	79.8	8	100	92	58.3
Non Palindromic	10	100	98	81.8	5	80	90.7	58.3
Results of PS								
Model	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
Palindromic	10	100	88.9	72.7	10	100	100	57.1
Non Palindromic	10	100	100	81.8	10	100	100	44
Results of MEME								
Model	R1	RR	spPPV	spSens	R1	RR	spPPV	spSens
Palindromic	10	100	90.9	90.9	10	100	73.3	73.3
Non Palindromic	10	100	81.8	81.8	10	100	73.3	73.3

Performance and quality measures: **R1**: the number of runs with an output out of the 10 runs on one real dataset, **RR (%)**: Recovery Rate: the percentage of the output (R1) for which the correct motif was retrieved (correct outputs), **spPPV (%)**: species-dependent PPV: the percentage of true sites among the predicted sites for the reference species, averaged over all correct outputs, **spSens (%)**: species-dependent Sens: the percentage of the true sites in the reference species found by the algorithm, averaged over all correct outputs. *E. coli* is the reference species. **Model**: ‘Palindromic’ stands for using the palindromic parameters (PG: -C, PS: -R 1,1,10 for LexA and -R 1,1,9 for TyrR and MEME: -pal), ‘Non Palindromic’ stands for no use of the palindromic parameters. **Gamma-proteobacteria**: The dataset of each regulator consists of 8 (LexA) or 7 (TyrR) target genes in *E. coli* (Table S2).

Parameters influencing the search algorithm:

The PG algorithm can be split up in simulated annealing and tracking (see chapter 2, Table 2.1: ‘Sampling’), *the number of cycles* for each phase (-S) was set to the default value of 100. One of the parameters often mentioned in the results section is the *tracking threshold* -T, which determines the trade-off between sensitivity and PPV (it corresponds to the frequency with which TF binding sites are co-sampled with the reference configuration during the tracking cycles). By increasing the default threshold of 0.05 to 0.50 we observed a drastic increase in PPV at the expense of only a slight drop in sensitivity, resulting in a better overall performance (data not shown). Unless indicated explicitly we always used T= 0.50. PS consists of burn-in and *sampling iterations* (see chapter 2, Table 2.1: ‘Sampling’) respectively set to 2000 and 8000 iterations (-bayes). The number of re-initializations (-S) was set to 20, to avoid that the algorithm reports a local optimum as is the case for low values of -S. When using *the fragmentation option* -F (see paragraph motif width) we also turned on the option *-align_centroid* to obtain a motif WM (this to align the centroid motif sites of different length resulting from the -F option). For PS the stringency of the centroid sites can not be altered by the user.

Parameters relating to the use of the phylogeny when working with prealigned orthologous sequences.

Both algorithms need as input a *phylogenetic tree* in Newick format with distances described by proximities for PG and branch lengths for PS. PS requires a separate tree for each MASS containing only the species present in this particular set of aligned orthologs (provided through the prior file -P indicated by ‘>TREE’). The provided perl script

(Seq.weights.pl) was used to calculate the sequence weights (see chapter 2, Table 2.1: ‘Scoring’) for each MASS based on the corresponding tree. The obtained sequence weights were provided to the algorithm through the prior file, indicated by ‘>WEIGHTS’. PG works with one tree relating all species present in the input dataset. An additional parameter in the PG algorithm is *the alignment level (-D)*. This parameter indicates if the sequences in the dataset are prealigned (-D=1 or -D=2) or not (-D=0). For prealigned sequences the user can specify by using -D the placement of the windows: -D=1 (splits up multi-species windows containing gaps into smaller windows without gaps) or -D=2 (gapped windows will be left out) (see also Text S2). For both the synthetic and real data -D was set to zero for the unaligned data and to one for the prealigned data.

Parameters describing the background model:

PG uses an N^{th} order Markov model (-N). For the synthetic data we used (-N=-1) indicating a single nucleotide frequency of 0.25 for A, C, G and T. For the real data we tested different background models: Markov models trained on the input sequences with order 0, 1 and 3 and a single nucleotide background model derived from the full *E. coli/S. cerevisiae* genome. Except for the Markov model order 3 (for which the input sequences were not sufficiently long to calculate reliably the correlated counts), all other background models derived from the input sequences gave comparable results (data not shown). For further tests on the real data, we used the default order 1 background model derived from the input sequences. PS uses a special *position specific background model* that gives the probabilities of observing each of the four nucleotides at each position in the sequence. This background model is derived by running a Bayesian segmentation algorithm (unifiedcpp) provided by PS, on the set of input sequences (for both the real and the synthetic datasets). The generated background model was provided to the algorithm by parameter -B.

Chapter 5 *De novo* motif discovery in vitamin D₃ regulated genes

Table S9 Results of the NHR-scan method on the entire 4000 bp region centered on the TSS of each human gene (in total 11 human genes) The table displays the gene number (Nr), gene name (Name), the binding site-type (Type), the binding site sequence (Sequence), the genomic start position (Genomic start), the position relative to the TSS (TSS) which is negative if the binding site is located upstream of the TSS, the location of the binding site inside (=1) or outside (=0) the regulatory regions as annotated by Ensembl (R) and the logarithm of the Viterbi score (Score). The rows marked in light gray are TF binding sites that overlap with *de novo* predicted binding sites for motif D2. The row marked in dark gray corresponds to the experimentally confirmed VDRE site upstream of the *PDLIM2* gene.

Nr	Name	Type	Sequence	Genomic start	TSS	R	Score
1	<i>PLXND1</i>	DR1	TGACCTTGGAACC	129326683	-1034	1	-19.9331
		DR2	GGGTCACAGGTGCA	129323838	1810	1	-21.4774
		IR1	GGGCCACTGGCTG	129326470	-821	1	-21.1364
		IR1	GGTGCACAGTCCT	129324597	1052	1	-20.4125
2	<i>ID3</i>	DR4	TGACCTCGGAGGAGCT	23886343	-73	1	-25.1957
		DR2	TGAACCTGTGGCCT	23884739	1533	1	-21.0615
3	<i>CPE</i>	DR4	TGGCCTCAAGTGATCC	166280377	-1969	0	-24.25
		DR1	TGGTGATAGGTCA	166282251	-95	0	-19.3872
		DR4	TGTACTGCTGTGAACA	166282954	608	0	-25.0523
		ER6	TGAACCTGCACAGGTTAT	166281947	-399	0	-27.6249

Appendix – Supplementary materials

Nr	Name	Type	Sequence	Genomic start	TSS	R	Score
		ER6	TGAAATGTTTAAAGTTTA	166282121	-225	0	-26.4722
		ER4	TGAGCCCGGGAGGTCA	166283592	1246	0	-25.1342
		ER6	TGAAATAGGTGAAGTGCA	166284129	1783	0	-26.5557
4	<i>PISD</i>	DR4	TGACCTCAGGTGATCC	32060126	-1723	0	-22.3667
		DR2	TGACCTCGTGATCC	32059755	-1350	0	-20.3773
		DR3	TGCCCTAGTTGAACT	32059472	-1068	0	-21.3738
		DR2	GGATCACGAGGTCA	32059219	-814	0	-20.4523
		DR2	TGGCCACTGAACT	32058268	137	0	-21.6818
		DR3	AGTTCACGAGGGACA	32057884	520	1	-23.3265
		DR4	GGGTTAGGTTAGGTAA	32057692	711	0	-24.1713
		DR3	TTACCTTGTGACCC	32057642	762	1	-23.341
		DR4	GGGTCTCTGAGTCCA	32057589	814	0	-25.1072
		DR4	GGATCACTGAGGTCA	32056948	1455	0	-22.2023
		ER6	TGTCCTGTGGGAAGGACA	32057762	639	1	-27.4598
5	<i>SEC14L1</i>	DR2	AGTTCAGCAGGTCA	75135223	-1782	0	-18.3712
		DR4	TGACCTCAGGTGATCC	75135481	-1524	0	-22.2917
		DR2	TGACCTCGTGATCC	75135809	-1196	0	-20.4523
		DR2	GGGTGATGAGGCCA	75135983	-1022	0	-22.0171
		DR4	TGACCTCCCGAAACT	75136641	-364	1	-24.9787
		ER6	AGACCTGGCTGTGGTTCA	75137759	754	1	-26.3309
6	<i>ITPR1</i>	ER8	TGAAGTGTGGACGAAGCTCA	4535756	722	1	-29.6821
7	<i>TSC22D3</i>	DR1	TCACCTCTGACCT	107021087	-527	1	-18.493
		ER8	TCAACTAAGGTGGCAGGTAA	107020246	307	1	-30.129
8	<i>PMEPA1</i>	DR1	TGGCCTTTGGCCT	56288314	-1734	0	-20.5451
		DR2	TGACCTCTGACCT	56288281	-1702	0	-17.4549
		DR4	ATGTCACACGAGGTCA	56288209	-1632	0	-22.6401
		DR3	TGAACTGGGTGTCCT	56287761	-1183	0	-22.5829
		DR5	GGATCAATGGGAGGTCA	56287202	-626	0	-25.2079
		DR3	TGACCTCCTTCAACC	56286916	-338	0	-21.3264
		DR2	GGTTCAGTGGGCCA	56285575	1004	1	-22.2064
		IR1	GGGCCTGGGTCTCT	56286658	-78	0	-21.1962
		IR1	GGGTCAAGGACCC	56285402	1178	1	-19.2777
		ER6	AGCACTCAGCCGAGGTCA	56287396	-821	0	-27.4286
		ER6	TGAATGGGACCCAGGTCA	56286891	-316	0	-27.5239
9	<i>PACSLN2</i>	DR2	TGACCTCGTGATCC	43412931	-1793	0	-20.4523
		DR2	TGACCTTGTGATCT	43412430	-1292	0	-19.8384
		DR4	TTACCTTCCCTGAACT	43411513	-377	1	-22.8915
		DR1	AGGTCAAGGGTCTG	43411000	139	1	-19.0009
		DR2	TGCCCTTGTGACCC	43410241	897	1	-19.6395
		IR1	AGGTGAGTGGCCG	43411016	123	1	-21.1649
		IR1	GGGTCCGGGGCCT	43410714	425	1	-20.503
10	<i>GRAMD4</i>	DR1	TGGCCTCTGCCCC	46970765	-1144	0	-19.8597
		DR2	AGGTCTGGCGGTCA	46971542	-367	1	-22.0877
		DR3	AGGTCACTGAGTTGA	46971974	65	1	-22.4748
		DR2	AGGGCACTGGGGCA	46971993	84	1	-22.4536
		DR3	CGACCTGGTTGCCCC	46973219	1310	0	-23.4304
		IR1	TGGTCACAGTGCT	46971229	-680	1	-21.0472

Appendix – Supplementary materials

Nr	Name	Type	Sequence	Genomic start	TSS	R	Score
		IR0	AGTCCCTGACCC	46972617	708	0	-19.4268
		ER6	TGCACCCCTGCCAGGTCC	46973844	1935	1	-27.9547
11	<i>PDLIM2</i>	DR4	TGACCTCAGGTGATCC	22433793	-1999	0	-22.2917
		DR2	AGGTGAGCAGGGCA	22433984	-1808	0	-21.4339
		DR4	TTACCTAACATGAGCC	22434016	-1776	0	-25.017
		DR4	AGGTTGCTGAAGGTCA	22435368	-424	0	-24.0336
		DR2	AGGCCAGCGGTCA	22435516	-276	0	-22.4536
		DR3	TGGCCTCCTTGCCCT	22437438	1646	1	-22.7564
		IR1	AGGTCAGGGACTT	22434222	-1570	0	-19.3915
		IR1	AGGGCAGTGGCCT	22434391	-1401	0	-18.7749
		IR1	GGGACAGAGACCA	22434661	-1131	0	-20.7429
		IR0	GGGTGAAGACCT	22435957	165	0	-19.6841
		IR1	GGGCCATCGGGCT	22436977	1185	1	-20.6213
		IR1	GGCTCAAAGTCT	22437032	1240	1	-20.8036
		ER1	TGAAGTCAGGTCC	22434122	-1670	0	-20.9458
		ER6	TGACCCAGCAGGGGTCA	22434438	-1354	0	-26.3359

Reference List

- Aerts,S., Thijs,G., Coessens,B., Staes,M., Moreau,Y. and De Moor,B. (2003a) Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res*, 31, 1753-1764.
- Aerts,S., Van Loo,P., Thijs,G., Moreau,Y. and De Moor,B. (2003b) Computational detection of cis -regulatory modules. *Bioinformatics*. 2003. Oct. ;19. Suppl 2:II5. -III4., 19 Suppl 2, II5-III4.
- Allende,M.L., Manzanares,M., Tena,J.J., Feijoo,C.G. and Gomez-Skarmeta,J.L. (2006) Cracking the genome's second code: enhancer detection by combined phylogenetic footprinting and transgenic fish and frog embryos. *Methods*, 39, 212-219.
- Antequera,F. (2003) Structure, function and evolution of CpG island promoters. *Cell Mol. Life Sci.*, 60, 1647-1658.
- Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., Harris,M.A., Hill,D.P. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25, 25-29.
- Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, 37, W202-W208.
- Bailey,T.L., Boden,M., Whittington,T. and Machanick,P. (2010) The value of position-specific priors in motif discovery using MEME. *BMC. Bioinformatics*, 11, 179.
- 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.
- Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 3, 21-29.
- Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, 129, 823-837.
- Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, 30, 4442-4451.
- Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E., Kuehn,M.S., Taylor,C.M. et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447, 799-816.
- Blanchette,M., Bataille,A.R., Chen,X., Poitras,C., Laganriere,J., Lefebvre,C., Deblois,G., Giguere,V., Ferretti,V., Bergeron,D., Coulombe,B. and Robert,F. (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.*, 16, 656-668.
- Blanchette,M., Schwikowski,B. and Tompa,M. (2002) Algorithms for phylogenetic footprinting. *J Comput. Biol*, 9, 211-223.

Reference List

- Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res*, 12, 739-748.
- Blanchette,M. and Tompa,M. (2003) FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res.*, 31, 3840-3842.
- Blekas,K., Fotiadis,D.I. and Likas,A. (2003) Greedy mixture learning for multiple motif discovery in biological sequences. *Bioinformatics*, 19, 607-617.
- Bock,C. and Lengauer,T. (2008) Computational epigenetics. *Bioinformatics*, 24, 1-10.
- Bock,C., Paulsen,M., Tierling,S., Mikeska,T., Lengauer,T. and Walter,J. (2006) CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet.*, 2, e26.
- Boeva,V., Surdez,D., Guillon,N., Tirode,F., Fejes,A.P., Delattre,O. and Barillot,E. (2010) De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic Acids Res.*, 38, e126.
- Bouillon,R., Carmeliet,G., Verlinden,L., van Etten,E., Verstuyf,A., Luderer,H.F., Lieben,L., Mathieu,C. and Demay,M. (2008) Vitamin D and human health: lessons from vitamin D receptor null mice. *Endocr. Rev.*, 29, 726-776.
- Bouillon,R., Okamura,W.H. and Norman,A.W. (1995) Structure-function relationships in the vitamin D endocrine system. *Endocr. Rev.*, 16, 200-257.
- Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132, 311-322.
- Brandeis,M., Frank,D., Keshet,I., Siegfried,Z., Mendelsohn,M., Nemes,A., Temper,V., Razin,A. and Cedar,H. (1994) Sp1 elements protect a CpG island from de novo methylation. *Nature*, 371, 435-438.
- Breitbart,M., Salamon,P., Andresen,B., Mahaffy,J.M., Segall,A.M., Mead,D., Azam,F. and Rohwer,F. (2002) Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U. S. A.*, 99, 14250-14255.
- Bryne,J.C., Valen,E., Tang,M.H., Marstrand,T., Winther,O., da,P., I, Krogh,A., Lenhard,B. and Sandelin,A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, 36, D102-D106.
- Cai,X., Hu,H. and Li,X.S. (2007) Tree Gibbs Sampler: identifying conserved motifs without aligning orthologous sequences. *Bioinformatics*, 23, 2013-2014.
- Carlberg,C. (1995) Mechanisms of nuclear signalling by vitamin D3. Interplay with retinoid and thyroid hormone signalling. *Eur. J. Biochem.*, 231, 517-527.
- Carlson,J.M., Chakravarty,A., DeZiel,C.E. and Gross,R.H. (2007) SCOPE: a web server for practical de novo motif discovery. *Nucleic Acids Res.*, 35, W259-W264.
- Celniker,S.E., Dillon,L.A., Gerstein,M.B., Gunsalus,K.C., Henikoff,S., Karpen,G.H., Kellis,M., Lai,E.C., Lieb,J.D., MacAlpine,D.M., Micklem,G., Piano,F. et al. (2009) Unlocking the secrets of the genome. *Nature*, 459, 927-930.
- Chakravarty,A., Carlson,J.M., Khetani,R.S. and Gross,R.H. (2007) A novel ensemble learning method for de novo computational identification of DNA binding sites. *BMC. Bioinformatics*, 8, 249.

Reference List

- Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, 31, 3497-3500.
- Cliften,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, 301, 71-76.
- Cliften,P.F., Hillier,L.W., Fulton,L., Graves,T., Miner,T., Gish,W.R., Waterston,R.H. and Johnston,M. (2001) Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.*, 11, 1175-1186.
- Cornish-Bowden,A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.*, 13, 3021-3030.
- Crawford,G.E., Davis,S., Scacheri,P.C., Renaud,G., Halawi,M.J., Erdos,M.R., Green,R., Meltzer,P.S., Wolfsberg,T.G. and Collins,F.S. (2006) DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat. Methods*, 3, 503-509.
- Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, 14, 1188-1190.
- Das,M.K. and Dai,H.K. (2007) A survey of DNA motif finding algorithms. *BMC Bioinformatics.*, 8 Suppl 7, S21.
- Deeb,K.K., Trump,D.L. and Johnson,C.S. (2007) Vitamin D signalling pathways in cancer: potential for anticancer therapeutics. *Nat. Rev. Cancer*, 7, 684-700.
- Derisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278, 680-686.
- Duret,L. and Bucher,P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct. Biol.*, 7, 399-406.
- Echchgadda,I., Song,C.S., Roy,A.K. and Chatterjee,B. (2004) Dehydroepiandrosterone sulfotransferase is a target for transcriptional induction by the vitamin D receptor. *Mol. Pharmacol.*, 65, 720-729.
- Edwards,R.A., Rodriguez-Brito,B., Wegley,L., Haynes,M., Breitbart,M., Peterson,D.M., Saar,M.O., Alexander,S., Alexander,E.C., Jr. and Rohwer,F. (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC. Genomics*, 7, 57.
- Eelen,G., Gysemans,C., Verlinden,L., Vanoirbeek,E., De Clercq,P., Van Haver,D., Mathieu,C., Bouillon,R. and Verstuyf,A. (2007) Mechanism and potential of the growth-inhibitory actions of vitamin D and analogs. *Curr. Med. Chem.*, 14, 1893-1910.
- Eger,A., Aigner,K., Sonderegger,S., Dampier,B., Oehler,S., Schreiber,M., Berx,G., Cano,A., Beug,H. and Foisner,R. (2005) DeltaEF1 is a transcriptional repressor of E-cadherin and regulates epithelial plasticity in breast cancer cells. *Oncogene*, 24, 2375-2385.
- Elkon,R., Linhart,C., Sharan,R., Shamir,R. and Shiloh,Y. (2003) Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.*, 13, 773-780.
- Ernst,J., Plasterer,H.L., Simon,I. and Bar-Joseph,Z. (2010) Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.*, 20, 526-536.
- Felsenfeld,G. and Groudine,M. (2003) Controlling the double helix. *Nature*, 421, 448-453.

Reference List

- Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17, 368-376.
- Fischer, H.M. (1994) Genetic regulation of nitrogen fixation in rhizobia. *Microbiol. Rev.*, 58, 352-386.
- FitzGerald, P.C., Shlyakhtenko, A., Mir, A.A. and Vinson, C. (2004) Clustering of DNA sequences in human promoters. *Genome Res.*, 14, 1562-1574.
- Foat, B.C., Morozov, A.V. and Bussemaker, H.J. (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, 22, e141-e149.
- Frith, M.C., Fu, Y., Yu, L., Chen, J.F., Hansen, U. and Weng, Z. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.*, 32, 1372-1381.
- Frith, M.C., Li, M.C. and Weng, Z. (2003) Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.*, 31, 3666-3668.
- Gelfand, M.S., Koonin, E.V. and Mironov, A.A. (2000) Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res.*, 28, 695-705.
- Georgiev, S., Boyle, A.P., Jayasurya, K., Ding, X., Mukherjee, S. and Ohler, U. (2010) Evidence-ranked motif identification. *Genome Biol.*, 11, R19.
- Gordan, R., Narlikar, L. and Hartemink, A.J. (2010) Finding regulatory DNA motifs using alignment-free evolutionary conservation information. *Nucleic Acids Res.*, 38, e90.
- Gotea, V. and Ovcharenko, I. (2008) DiRE: identifying distant regulatory elements of co-expressed genes. *Nucleic Acids Res.*, 36, W133-W139.
- Grad, Y.H., Roth, F.P., Halfon, M.S. and Church, G.M. (2004) Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in *Drosophila melanogaster* and *D.pseudoobscura*. *Bioinformatics*, 20, 2738-2750.
- Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R. and Young, R.A. (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130, 77-88.
- Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, 52, 696-704.
- Guns, T., Sun, H., Marchal, K. and Nijssen, S. (2010) Cis-regulatory module detection using constraint programming.
- Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. and Noble, W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, 8, R24.
- Gurlek, A., Pittelkow, M.R. and Kumar, R. (2002) Modulation of growth factor/cytokine synthesis and signaling by 1 α ,25-dihydroxyvitamin D(3): implications in cell growth and differentiation. *Endocr. Rev.*, 23, 763-786.
- Guyton, K.Z., Kensler, T.W. and Posner, G.H. (2003) Vitamin D and vitamin D analogs as cancer chemopreventive agents. *Nutr. Rev.*, 61, 227-238.
- Habib, N., Kaplan, T., Margalit, H. and Friedman, N. (2008) A novel Bayesian DNA motif comparison method for clustering and retrieval. *PLoS Comput. Biol.*, 4, e1000010.
- Hajkova, P., El-Maarri, O., Engemann, S., Oswald, J., Olek, A. and Walter, J. (2002) DNA-methylation analysis by the bisulfite-assisted genomic sequencing method. *Methods Mol. Biol.*, 200, 143-154.

Reference List

- Halpern,A.L. and Bruno,W.J. (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.*, 15, 910-917.
- Hansen,J.C., Tse,C. and Wolffe,A.P. (1998) Structure and function of the core histone N-termini: more than meets the eye. *Biochemistry*, 37, 17637-17641.
- Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J., Jennings,E.G., Zeitlinger,J. et al. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431, 99-104.
- Hasegawa,M., Kishino,H. and Yano,T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22, 160-174.
- Haussler,M.R., Haussler,C.A., Whitfield,G.K., Hsieh,J.C., Thompson,P.D., Barthel,T.K., Bartik,L., Egan,J.B., Wu,Y., Kubicek,J.L., Lowmiller,C.L., Moffet,E.W. et al. (2010) The nuclear vitamin D receptor controls the expression of genes encoding factors which feed the "Fountain of Youth" to mediate healthful aging. *J. Steroid Biochem. Mol. Biol.*, 121, 88-97.
- Haussler,M.R., Whitfield,G.K., Haussler,C.A., Hsieh,J.C., Thompson,P.D., Selznick,S.H., Dominguez,C.E. and Jurutka,P.W. (1998) The nuclear vitamin D receptor: biological and molecular regulatory properties revealed. *J. Bone Miner. Res.*, 13, 325-349.
- Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W., Ching,K.A., ntosiewicz-Bourget,J.E. et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459, 108-112.
- Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van,C.S., Qu,C., Ching,K.A., Wang,W., Weng,Z. et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, 39, 311-318.
- Holick,M.F. (2004) Sunlight and vitamin D for bone health and prevention of autoimmune diseases, cancers, and cardiovascular disease. *Am. J. Clin. Nutr.*, 80, 1678S-1688S.
- Hon,L.S. and Jain,A.N. (2006) A deterministic motif finding algorithm with application to the human genome. *Bioinformatics*, 22, 1047-1054.
- Hu,J., Li,B. and Kihara,D. (2005) Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res.*, 33, 4899-4913.
- Hu,J., Yang,Y.D. and Kihara,D. (2006) EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences. *BMC. Bioinformatics*, 7, 342.
- Huang,Y.C., Chen,J.Y. and Hung,W.C. (2004) Vitamin D3 receptor/Sp1 complex is required for the induction of p27Kip1 expression by vitamin D3. *Oncogene*, 23, 4856-4861.
- Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L., Coates,G., Fairley,S. et al. (2009) Ensembl 2009. *Nucleic Acids Res.*, 37, D690-D697.
- Huerta,A.M., Salgado,H., Thieffry,D. and Collado-Vides,J. (1998) RegulonDB: a database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res.*, 26, 55-59.
- Ihmels,J., Bergmann,S., Berman,J. and Barkai,N. (2005) Comparative gene expression analysis by differential clustering approach: application to the *Candida albicans* transcription program. *PLoS Genet.*, 1, e39.

Reference List

- Inoue,K. and Urahama,K. (1999) Sequential fuzzy cluster extraction by a graph spectral method. *Pattern Recognition Letters*, 20, 699-705.
- Jeziorska,D.M., Jordan,K.W. and Vance,K.W. (2009) A systems biology approach to understanding cis-regulatory module function. *Semin. Cell Dev. Biol.*, 20, 856-862.
- Ji,H., Jiang,H., Ma,W., Johnson,D.S., Myers,R.M. and Wong,W.H. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, 26, 1293-1300.
- Jin,V.X., Apostolos,J., Nagisetty,N.S. and Farnham,P.J. (2009) W-ChIPMotifs: a web application tool for de novo motif discovery from ChIP-based high-throughput data. *Bioinformatics*, 25, 3191-3193.
- Jones,G., Strugnell,S.A. and DeLuca,H.F. (1998) Current understanding of the molecular actions of vitamin D. *Physiol Rev.*, 78, 1193-1231.
- Jones,P.A. and Martienssen,R. (2005) A blueprint for a Human Epigenome Project: the AACR Human Epigenome Workshop. *Cancer Res.*, 65, 11241-11246.
- Joshi,A., Van de Peer,Y. and Michoel,T. (2008) Analysis of a Gibbs sampler method for model-based clustering of gene expression data. *Bioinformatics*, 24, 176-183.
- Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. *Mammalian protein metabolism*. Academic Press, New York, pp. 21-132.
- Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423, 241-254.
- Kim,T.H., Abdullaev,Z.K., Smith,A.D., Ching,K.A., Loukinov,D.I., Green,R.D., Zhang,M.Q., Lobanenko,V.V. and Ren,B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, 128, 1231-1245.
- Kim,T.H., Barrera,L.O., Qu,C., Van,C.S., Trinklein,N.D., Cooper,S.J., Luna,R.M., Glass,C.K., Rosenfeld,M.G., Myers,R.M. and Ren,B. (2005a) Direct isolation and identification of promoters in the human genome. *Genome Res.*, 15, 830-839.
- Kim,T.H., Barrera,L.O., Zheng,M., Qu,C., Singer,M.A., Richmond,T.A., Wu,Y., Green,R.D. and Ren,B. (2005b) A high-resolution map of active promoters in the human genome. *Nature*, 436, 876-880.
- Kim,T.H. and Ren,B. (2006) Genome-wide analysis of protein-DNA interactions. *Annu. Rev. Genomics Hum. Genet.*, 7, 81-102.
- Kim,T.M. and Park,P.J. (2011) Advances in analysis of transcriptional regulatory networks. *Wiley. Interdiscip. Rev. Syst. Biol. Med.*, 3, 21-35.
- King,O.D. and Roth,F.P. (2003) A non-parametric model for transcription factor binding sites. *Nucleic Acids Res.*, 31, e116.
- Kirkpatrick,S., Gelatt,C.D., Jr. and Vecchi,M.P. (1983) Optimization by Simulated Annealing. *Science*, 220, 671-680.
- Klepper,K. and Drablos,F. (2010) PriorsEditor: a tool for the creation and use of positional priors in motif discovery. *Bioinformatics*, 26, 2195-2197.

Reference List

- Koch,C.M., Andrews,R.M., Flicek,P., Dillon,S.C., Karaoz,U., Clelland,G.K., Wilcox,S., Beare,D.M., Fowler,J.C., Couttet,P., James,K.D., Lefebvre,G.C. et al. (2007) The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res.*, 17, 691-707.
- Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, 128, 693-705.
- Lahdesmaki,H., Rust,A.G. and Shmulevich,I. (2008) Probabilistic inference of transcription factor binding from multiple data sources. *PLoS One.*, 3, e1820.
- Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262, 208-214.
- Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7, 41-51.
- Lazarova,D.L., Bordonaro,M. and Sartorelli,A.C. (2001) Transcriptional regulation of the vitamin D(3) receptor gene by ZEB. *Cell Growth Differ.*, 12, 319-326.
- Lee,T.I. and Young,R.A. (2000) Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.*, 34, 77-137.
- Li,B., Carey,M. and Workman,J.L. (2007) The role of chromatin during transcription. *Cell*, 128, 707-719.
- Li,X. and Wong,W.H. (2005) Sampling motifs on phylogenetic trees. *Proc Natl Acad Sci U. S. A.*, 102, 9481-9486.
- Lin,R. and White,J.H. (2004) The pleiotropic actions of vitamin D. *Bioessays*, 26, 21-28.
- Liu,J., Liu,Y.G., Huang,R., Yao,C., Li,S., Yang,W., Yang,D. and Huang,R.P. (2007) Concurrent down-regulation of Egr-1 and gelsolin in the majority of human breast cancer cells. *Cancer Genomics Proteomics.*, 4, 377-385.
- Liu,J.S. (2001) *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics.
- Liu,J.S. and Lawrence,C.E. (1999) Bayesian inference on biopolymer models. *Bioinformatics*, 15, 38-52.
- Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *Journal of the American Statistical Association*, 90, 1156-1170.
- Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, :127-38., 127-138.
- Liu,Y., Liu,X.S., Wei,L., Altman,R.B. and Batzoglou,S. (2004) Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res.*, 14, 451-458.
- Liu,Y., Niu,N., Zhu,X., Du,T., Wang,X., Chen,D., Wu,X., Gu,H.F. and Liu,Y. (2008) Genetic variation and association analyses of the nuclear respiratory factor 1 (nRF1) gene in Chinese patients with type 2 diabetes. *Diabetes*, 57, 777-782.

Reference List

- Loughran,G., Healy,N.C., Kiely,P.A., Huigsloot,M., Kedersha,N.L. and O'Connor,R. (2005) Mystique is a new insulin-like growth factor-I-regulated PDZ-LIM domain protein that promotes cell attachment and migration and suppresses Anchorage-independent growth. *Mol. Biol. Cell*, 16, 1811-1822.
- Ludwig,M.Z. (2002) Functional evolution of noncoding DNA. *Curr. Opin. Genet. Dev.*, 12, 634-639.
- Ludwig,M.Z., Bergman,C., Patel,N.H. and Kreitman,M. (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, 403, 564-567.
- Luster,T.A. and Rizzino,A. (2003) Regulation of the FGF-4 gene by a complex distal enhancer that functions in part as an enhanceosome. *Gene*, 323, 163-172.
- Maeda,T., Hobbs,R.M. and Pandolfi,P.P. (2005) The transcription factor Pokemon: a new key player in cancer pathogenesis. *Cancer Res.*, 65, 8575-8578.
- Maeda,T., Merghoub,T., Hobbs,R.M., Dong,L., Maeda,M., Zakrzewski,J., van den Brink,M.R., Zelent,A., Shigematsu,H., Akashi,K., Teruya-Feldstein,J., Cattoretti,G. et al. (2007) Regulation of B versus T lymphoid lineage fate decision by the proto-oncogene LRF. *Science*, 316, 860-866.
- Marchal,K., De Keersmaecker,S., Monsieurs,P., van Boxel,N., Lemmens,K., Thijs,G., Vanderleyden,J. and De Moor,B. (2004) In silico identification and experimental validation of PmrAB targets in *Salmonella typhimurium* by regulatory motif detection. *Genome Biol.*, 5, R9.
- Marks,P., Rifkind,R.A., Richon,V.M., Breslow,R., Miller,T. and Kelly,W.K. (2001) Histone deacetylases and cancer: causes and therapies. *Nat. Rev. Cancer*, 1, 194-202.
- 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. et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34, D108-D110.
- McGuire,A.M., Hughes,J.D. and Church,G.M. (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.*, 10, 744-757.
- Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P., Lee,W., Mendenhall,E. et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448, 553-560.
- Mohn,F., Weber,M., Schubeler,D. and Roloff,T.C. (2009) Methylated DNA immunoprecipitation (MeDIP). *Methods Mol. Biol.*, 507, 55-64.
- Monsieurs,P., Thijs,G., Fadda,A.A., De Keersmaecker,S.C., Vanderleyden,J., De Moor,B. and Marchal,K. (2006) More robust detection of motifs in coexpressed genes by using phylogenetic information. *BMC. Bioinformatics.*, 7, 160.
- Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics.*, 15, 211-218.
- Moses,A.M., Chiang,D.Y. and Eisen,M.B. (2004a) Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac. Symp Biocomput.*, 324-335.
- Moses,A.M., Chiang,D.Y., Pollard,D.A., Iyer,V.N. and Eisen,M.B. (2004b) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol*, 5, R98.

Reference List

- Moses, A.M., Pollard, D.A., Nix, D.A., Iyer, V.N., Li, X.Y., Biggin, M.D. and Eisen, M.B. (2006) Large-scale turnover of functional transcription factor binding sites in *Drosophila*. *PLoS Comput. Biol.*, 2, e130.
- Muller, F., Williams, D.W., Kobolak, J., Gauvry, L., Goldspink, G., Orban, L. and Maclean, N. (1997) Activator effect of coinjected enhancers on the muscle-specific expression of promoters in zebrafish embryos. *Mol. Reprod. Dev.*, 47, 404-412.
- Narlikar, L., Gordan, R., Ohler, U. and Hartemink, A.J. (2006) Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics*, 22, e384-e392.
- Narlikar, L. and Ovcharenko, I. (2009) Identifying regulatory elements in eukaryotic genomes. *Brief. Funct. Genomic. Proteomic.*, 8, 215-230.
- Newberg, L.A., McCue, L.A. and Lawrence, C.E. (2005) The relative inefficiency of sequence weights approaches in determining a nucleotide position weight matrix. *Stat. Appl. Genet. Mol. Biol.*, 4, Article13.
- Newberg, L.A., Thompson, W.A., Conlan, S., Smith, T.M., McCue, L.A. and Lawrence, C.E. (2007) A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction. *Bioinformatics.*, 23, 1718-1727.
- Numoto, M., Yokoro, K., Yanagihara, K., Kamiya, K. and Niwa, O. (1995) Over-expressed ZF5 gene product, a c-myc-binding protein related to GL1-Kruppel protein, has a growth-suppressive activity in mouse cell lines. *Jpn. J. Cancer Res.*, 86, 277-283.
- Ooi, S.K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.P., Allis, C.D., Cheng, X. and Bestor, T.H. (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, 448, 714-717.
- Pachkov, M., Erb, I., Molina, N. and van Nimwegen, E. (2007) SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res.*, 35, D127-D131.
- Palumbo, M.J. and Newberg, L.A. (2010) PhyloScan: locating transcription-regulating binding sites in mixed aligned and unaligned sequence data. *Nucleic Acids Res.*, 38 Suppl, W268-W274.
- Pavesi, G., Mauri, G. and Pesole, G. (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, 17 Suppl 1, S207-S214.
- Pavesi, G., Mauri, G. and Pesole, G. (2004) In silico representation and discovery of transcription factor binding sites. *Brief. Bioinform.*, 5, 217-236.
- Pena, C., Garcia, J.M., Silva, J., Garcia, V., Rodriguez, R., Alonso, I., Millan, I., Salas, C., de Herreros, A.G., Munoz, A. and Bonilla, F. (2005) E-cadherin and vitamin D receptor regulation by SNAIL and ZEB1 in colon cancer: clinicopathological correlations. *Hum. Mol. Genet.*, 14, 3361-3370.
- Perez, J.C. and Groisman, E.A. (2009) Evolution of transcriptional regulatory circuits in bacteria. *Cell*, 138, 233-244.
- Phan, V. and Furlotte, N.A. (2008) Motif Tool Manager: a web-based framework for motif discovery. *Bioinformatics*, 24, 2930-2931.
- Piipari, M., Down, T.A., Saini, H., Enright, A. and Hubbard, T.J. (2010) iMotifs: an integrated sequence motif visualization and analysis environment. *Bioinformatics*, 26, 843-844.

Reference List

- Prakash,A., Blanchette,M., Sinha,S. and Tompa,M. (2004) Motif discovery in heterogeneous sequence data. *Pac. Symp Biocomput.*, 348-359.
- Prakash,A. and Tompa,M. (2005) Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol.*, 23, 1249-1256.
- Rachez,C. and Freedman,L.P. (2000) Mechanisms of gene regulation by vitamin D(3) receptor: a network of coactivator interactions. *Gene*, 246, 9-21.
- Ramagopalan,S.V., Heger,A., Berlanga,A.J., Maugeri,N.J., Lincoln,M.R., Burrell,A., Handunnetthi,L., Handel,A.E., Disanto,G., Orton,S.M., Watson,C.T., Morahan,J.M. et al. (2010) A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. *Genome Res.*, 20, 1352-1360.
- Ray,P., Shringarpure,S., Kolar,M. and Xing,E.P. (2008) CSMET: comparative genomic motif detection via multi-resolution phylogenetic shadowing. *PLoS. Comput. Biol.*, 4, e1000090.
- Reddy,T.E., DeLisi,C. and Shakhnovich,B.E. (2007) Binding site graphs: a new graph theoretical framework for prediction of transcription factor binding sites. *PLoS. Comput. Biol.*, 3, e90.
- 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. et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, 4, 651-657.
- Robertson,K.D. and Wolffe,A.P. (2000) DNA methylation in health and disease. *Nat. Rev. Genet.*, 1, 11-19.
- Roeder,R.G. (2003) Lasker Basic Medical Research Award. The eukaryotic transcriptional machinery: complexities and mechanisms unforeseen. *Nat. Med.*, 9, 1239-1244.
- Roh,T.Y., Cuddapah,S. and Zhao,K. (2005) Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.*, 19, 542-552.
- Romer,K.A., Kayombya,G.R. and Fraenkel,E. (2007) WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches. *Nucleic Acids Res.*, 35, W217-W220.
- Sakabe,N.J. and Nobrega,M.A. (2010) Genome-wide maps of transcription regulatory elements. *Wiley. Interdiscip. Rev. Syst. Biol. Med.*, 2, 422-437.
- Sanchez-Tillo,E., Lazaro,A., Torrent,R., Cuatrecasas,M., Vaquero,E.C., Castells,A., Engel,P. and Postigo,A. (2010) ZEB1 represses E-cadherin and induces an EMT by recruiting the SWI/SNF chromatin-remodeling protein BRG1. *Oncogene*, 29, 3490-3500.
- Sandelin,A. and Wasserman,W.W. (2005) Prediction of nuclear hormone receptor response elements. *Mol. Endocrinol.*, 19, 595-606.
- Schmid,C.D., Perier,R., Praz,V. and Bucher,P. (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res.*, 34, D82-D85.
- Schones,D.E. and Zhao,K. (2008) Genome-wide approaches to studying chromatin modifications. *Nat. Rev. Genet.*, 9, 179-191.

Reference List

- Schrader,M., Nayeri,S., Kahlen,J.P., Muller,K.M. and Carlberg,C. (1995) Natural vitamin D3 response elements formed by inverted palindromes: polarity-directed ligand sensitivity of vitamin D3 receptor-retinoid X receptor heterodimer-mediated transactivation. *Mol. Cell Biol.*, 15, 1154-1161.
- Segal,E. and Sharan,R. (2005) A discriminative model for identifying spatial cis-regulatory modules. *J. Comput. Biol.*, 12, 822-834.
- Sharan,R., Ovcharenko,I., Ben Hur,A. and Karp,R.M. (2003) CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, 19 Suppl 1:I283-I291., I283-I291.
- Siddharthan,R. (2007) Parsing regulatory DNA: general tasks, techniques, and the PhyloGibbs approach. *J. Biosci.*, 32, 863-870.
- Siddharthan,R. (2008) PhyloGibbs-MP: module prediction and discriminative motif-finding by Gibbs sampling. *PLoS Comput. Biol.*, 4, e1000156.
- Siddharthan,R., Siggia,E.D. and van Nimwegen,E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS. Comput Biol.*, 1, e67.
- Siddharthan,R. and van Nimwegen,E. (2007) Detecting regulatory sites using PhyloGibbs. *Methods Mol. Biol.*, 395, 381-402.
- Sinha,S., Blanchette,M. and Tompa,M. (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, 5, 170.
- Sinha,S. and Tompa,M. (2003) YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, 31, 3586-3588.
- Sinha,S., van Nimwegen,E. and Siggia,E.D. (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, 19 Suppl 1:I292-I301., I292-I301.
- Sosinsky,A., Honig,B., Mann,R.S. and Califano,A. (2007) Discovering transcriptional regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting. *Proc. Natl. Acad. Sci. U. S. A.*, 104, 6305-6310.
- Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U. S. A.*, 100, 9440-9445.
- Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, 16, 16-23.
- Storms,V., Claeys,M., Sanchez,A., De Moor,B., Verstuyf,A. and Marchal,K. (2010) The effect of orthology and coregulation on detecting regulatory motifs. *PLoS. ONE.*, 5, e8938.
- Stoye,J., Evers,D. and Meyer,F. (1998) Rose: generating sequence families. *Bioinformatics.*, 14, 157-163.
- Strahl,B.D. and Allis,C.D. (2000) The language of covalent histone modifications. *Nature*, 403, 41-45.
- Straussman,R., Nejman,D., Roberts,D., Steinfeld,I., Blum,B., Benvenisty,N., Simon,I., Yakhini,Z. and Cedar,H. (2009) Developmental programming of CpG island methylation profiles in the human genome. *Nat. Struct. Mol. Biol.*, 16, 564-571.
- Sun,H., De Bie,T., Storms,V., Fu,Q., Dhollander,T., Lemmens,K., Verstuyf,A., De Moor,B. and Marchal,K. (2009) ModuleDigger: an itemset mining framework for the detection of cis-regulatory modules. *BMC. Bioinformatics*, 10 Suppl 1, S30.

Reference List

- Tagami,T., Lutz,W.H., Kumar,R. and Jameson,J.L. (1998) The interaction of the vitamin D receptor with nuclear receptor corepressors and coactivators. *Biochem. Biophys. Res. Commun.*, 253, 358-363.
- Tanay,A., Regev,A. and Shamir,R. (2005) Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc. Natl. Acad. Sci. U. S. A.*, 102, 7203-7208.
- Tang,M.H., Krogh,A. and Winther,O. (2008) BayesMD: flexible biological modeling for motif discovery. *J. Comput. Biol.*, 15, 1347-1363.
- Tavera-Mendoza,L., Wang,T.T., Lallemand,B., Zhang,R., Nagai,Y., Bourdeau,V., Ramirez-Calderon,M., Desbarats,J., Mader,S. and White,J.H. (2006) Convergence of vitamin D and retinoic acid signalling at a common hormone response element. *EMBO Rep.*, 7, 180-185.
- Thijs,G., Marchal,K., Lescot,M., Rombauts,S., De Moor,B., Rouze,P. and Moreau,Y. (2002a) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.*, 9, 447-464.
- Thijs,G., Moreau,Y., De Smet,F., Mathys,J., Lescot,M., Rombauts,S., Rouze,P., De Moor,B. and Marchal,K. (2002b) INCLUSIVE: INtegrated Clustering, Upstream sequence retrieval and motif Sampling. *Bioinformatics*, 18, 331-332.
- Thomas,D.J., Rosenbloom,K.R., Clawson,H., Hinrichs,A.S., Trumbower,H., Raney,B.J., Karolchik,D., Barber,G.P., Harte,R.A., Hillman-Jackson,J., Kuhn,R.M., Rhead,B.L. et al. (2007) The ENCODE Project at UC Santa Cruz. *Nucleic Acids Res.*, 35, D663-D667.
- Thomas,M.C. and Chiang,C.M. (2006) The general transcription machinery and general cofactors. *Crit Rev. Biochem. Mol. Biol.*, 41, 105-178.
- Thompson,P.D., Jurutka,P.W., Whitfield,G.K., Myskowski,S.M., Eichhorst,K.R., Dominguez,C.E., Haussler,C.A. and Haussler,M.R. (2002) Liganded VDR induces CYP3A4 in small intestinal and colon cancer cells via DR3 and ER6 vitamin D responsive elements. *Biochem. Biophys. Res. Commun.*, 299, 730-738.
- Thompson,W., Rouchka,E.C. and Lawrence,C.E. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, 31, 3580-3585.
- Thompson,W.A., Newberg,L.A., Conlan,S., McCue,L.A. and Lawrence,C.E. (2007) The Gibbs Centroid Sampler. *Nucleic Acids Res.*, 35, W232-W237.
- Tompa,M. (2001) Identifying functional elements by comparative DNA sequence analysis. *Genome Res*, 11, 1143-1144.
- 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.*, 262-271.
- Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J., Makeev,V.J., Mironov,A.A. et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, 23, 137-144.
- van Helden,J., Andre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281, 827-842.
- Van Hellemont,R., Monsieurs,P., Thijs,G., De Moor,B., Van de Peer,Y. and Marchal,K. (2005) A novel approach to identifying regulatory motifs in distantly related genomes. *Genome Biol*, 6, R113.

Reference List

- Van Loo,P., Aerts,S., Thienpont,B., De Moor,B., Moreau,Y. and Marynen,P. (2008) ModuleMiner - improved computational detection of cis-regulatory modules: are there different modes of gene regulation in embryonic development and adult tissues? *Genome Biol.*, 9, R66.
- Van Loo,P. and Marynen,P. (2009) Computational methods for the detection of cis-regulatory modules. *Brief. Bioinform.*, 10, 509-524.
- van Nimwegen,E. (2007) Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC. Bioinformatics*, 8 Suppl 6, S4.
- Vanoirbeek,E., Eelen,G., Verlinden,L., Marchal,K., Engelen,K., De Moor,B., Beullens,I., Marcelis,S., De Clercq,P., Bouillon,R. and Verstuyf,A. (2009) Microarray analysis of MCF-7 breast cancer cells treated with 1,25-dihydroxyvitamin D3 or a 17-methyl-D-ring analog. *Anticancer Res.*, 29, 3585-3590.
- Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W., Fouts,D.E., Levy,S. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304, 66-74.
- Verlinden,L., Eelen,G., Beullens,I., Van Camp,M., Van hummelen,P., Engelen,K., Van Hellefont,R., Marchal,K., De Moor,B., Fojijer,F., Te Riele,H., Beullens,M. et al. (2005) Characterization of the condensin component Cnap1 and the protein kinase Melk as novel E2F-target genes down-regulated by 1,25-dihydroxyvitamin D3. *J. Biol Chem.*
- Verlinden,L., Verstuyf,A., Van Camp,M., Marcelis,S., Sabbe,K., Zhao,X.Y., De Clercq,P., Vandewalle,M. and Bouillon,R. (2000) Two novel 14-Epi-analogues of 1,25-dihydroxyvitamin D3 inhibit the growth of human breast cancer cells in vitro and in vivo. *Cancer Res.*, 60, 2673-2679.
- Vingron,M., Brazma,A., Coulson,R., van,H.J., Manke,T., Palin,K., Sand,O. and Ukkonen,E. (2009) Integrating sequence, evolution and functional genomics in regulatory genomics. *Genome Biol.*, 10, 202.
- Wall,D.P., Fraser,H.B. and Hirsh,A.E. (2003) Detecting putative orthologs. *Bioinformatics.*, 19, 1710-1711.
- Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 19, 2369-2380.
- Wang,T.T., Tavera-Mendoza,L.E., Laperriere,D., Libby,E., MacLeod,N.B., Nagai,Y., Bourdeau,V., Konstorum,A., Lallemand,B., Zhang,R., Mader,S. and White,J.H. (2005) Large-scale in silico and microarray-based identification of direct 1,25-dihydroxyvitamin D3 target genes. *Mol. Endocrinol.*, 19, 2685-2695.
- Wang,Z., Zang,C., Rosenfeld,J.A., Schones,D.E., Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Peng,W., Zhang,M.Q. and Zhao,K. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, 40, 897-903.
- Ward,L.D. and Bussemaker,H.J. (2008) Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics*, 24, i165-i171.
- Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, 278, 167-181.
- Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, 5, 276-287.

Reference List

- Weber, M. and Schubeler, D. (2007) Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Curr. Opin. Cell Biol.*, 19, 273-280.
- Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y., Weng, Z., Liu, J., Zhao, X.D. et al. (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell*, 124, 207-219.
- Whittington, T., Perkins, A.C. and Bailey, T.L. (2009) High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Res.*, 37, 14-25.
- Won, K.J., Agarwal, S., Shen, L., Shoemaker, R., Ren, B. and Wang, W. (2009) An integrated approach to identifying cis-regulatory modules in the human genome. *PLoS One.*, 4, e5501.
- Won, K.J., Chepelev, I., Ren, B. and Wang, W. (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics*, 9, 547.
- Won, K.J., Ren, B. and Wang, W. (2010) Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol.*, 11, R7.
- Wray, G.A. (2007) The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.*, 8, 206-216.
- Xi, H., Shulha, H.P., Lin, J.M., Vales, T.R., Fu, Y., Bodine, D.M., McKay, R.D., Chenoweth, J.G., Tesar, P.J., Furey, T.S., Ren, B., Weng, Z. et al. (2007) Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet.*, 3, e136.
- Xie, D., Cai, J., Chia, N.Y., Ng, H.H. and Zhong, S. (2008) Cross-species de novo identification of cis-regulatory modules with GibbsModule: application to gene regulation in embryonic stem cells. *Genome Res.*, 18, 1325-1335.
- 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, 338-345.
- Yamada, Y., Warren, A.J., Dobson, C., Forster, A., Pannell, R. and Rabbitts, T.H. (1998) The T cell leukemia LIM protein Lmo2 is necessary for adult mouse hematopoiesis. *Proc. Natl. Acad. Sci. U. S. A.*, 95, 3890-3895.
- Zehnder, D., Bland, R., Williams, M.C., McNinch, R.W., Howie, A.J., Stewart, P.M. and Hewison, M. (2001) Extrarenal expression of 25-hydroxyvitamin d(3)-1 alpha-hydroxylase. *J. Clin. Endocrinol. Metab.*, 86, 888-894.
- Zhao, B., Ye, X., Yu, J., Li, L., Li, W., Li, S., Yu, J., Lin, J.D., Wang, C.Y., Chinnaiyan, A.M., Lai, Z.C. and Guan, K.L. (2008) TEAD mediates YAP-dependent gene induction and growth control. *Genes Dev.*, 22, 1962-1971.
- Zhou, Q. and Wong, W.H. (2007) Coupling hidden Markov models for the discovery of Cis-regulatory modules in multiple species. *The Annals of Applied Statistics*, 1, 36-65.
- Zhou, Q. and Wong, W.H. (2004) CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci. U. S. A.*, 101, 12114-12119.
- Zhu, J. and Zhang, M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, 15, 607-611.

Publication list

Storms, V., Claeys, M., Sanchez, A., De Moor, B., Verstuyf, A., Marchal, K. (2010). The effect of orthology and coregulation on detecting regulatory motifs. *PLoS. ONE.*, 5(2), e8938.

Sun, H., De Bie, T., **Storms, V.**, Fu, Q., Dhollander, T., Lemmens, K., Verstuyf, A., De Moor, B., Marchal, K. (2009). ModuleDigger: an itemset mining framework for the detection of *cis*-regulatory modules. *BMC. Bioinformatics*, 10 Suppl 1, S30.

Zhao, H., Cloots, L., Van den Bulcke, T., Wu, Y., De Smet, R., **Storms, V.**, Meysman, P., Engelen, K., Marchal, K. (2010). Query-based biclustering of gene expression data using Probabilistic Relational Models. Accepted for publication in *BMC Bioinformatics*.

Curriculum vitae

Valerie Storms was born on September 8th, 1983 in Hasselt, Belgium. In 2001 she started her education in Bioscience Engineering at the Katholieke Universiteit Leuven, where she received a Masters degree in Bioscience Engineering, option ‘Cell and Gene Technology’ in 2006. In October 2006, she started her PhD at the Katholieke Universiteit Leuven under the supervision of Prof. Kathleen Marchal and Prof. Bart De Moor.