Ghent University

Faculty of Sciences

Department of Biomedical Molecular Biology

Developing algorithms for the *in silico* Identification of Transcription Factor Binding Sites

Dissertation submitted in fulfillment of the requirement for the degree of

Doctor in Sciences, Bioinformatics

Academic year 2013-2014

Stefan BROOS

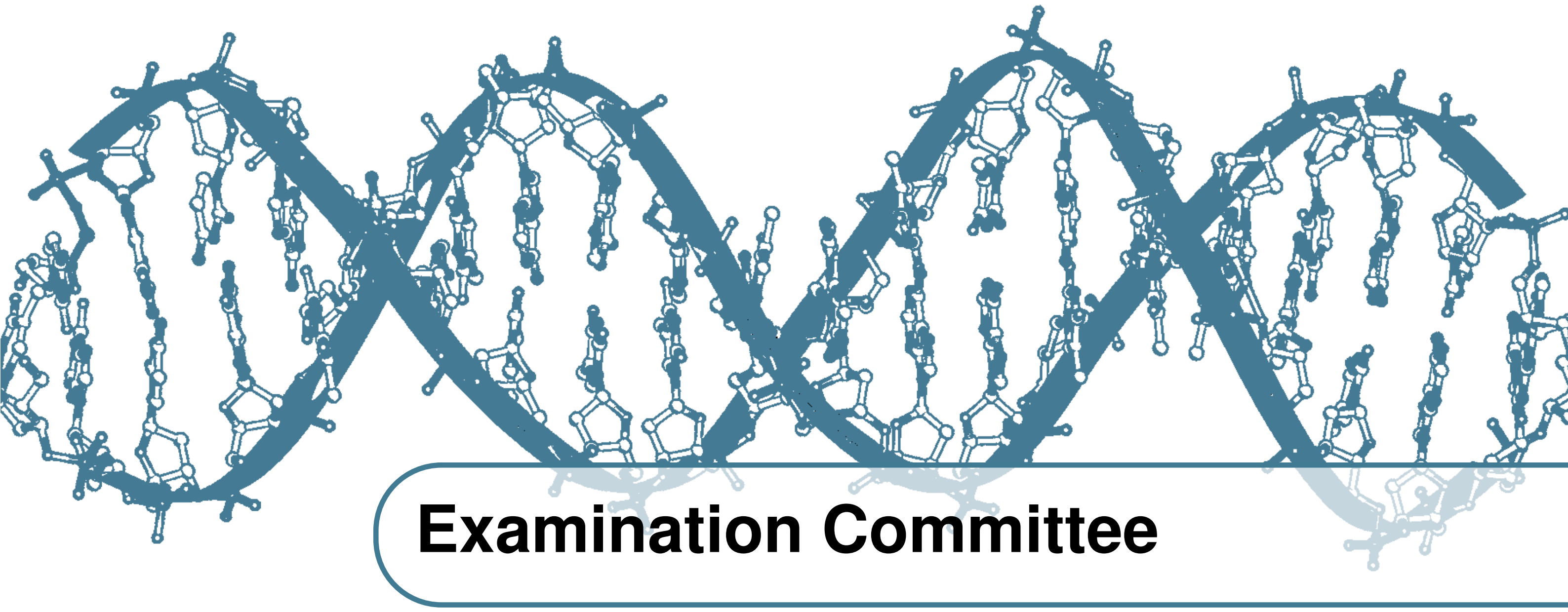
Promoter: Prof. Dr. Frans Van Roy

Co-Promoter: Prof. Dr. Pieter De Bleser

VIB / Inflammation Research Center

Bioinformatics Core Facility

Technologiepark 927 B-9052 Ghent (Zwijnaarde), Belgium



Examination Committee

- **Prof. Dr. Rudi Beyaert (chair)**
  Department of Biomedical Molecular Biology
  Ghent University
  VIB Inflammation Research Center

- **Prof. Dr. Frans Van Roy**
  Department of Biomedical Molecular Biology
  Ghent University
  VIB Inflammation Research Center

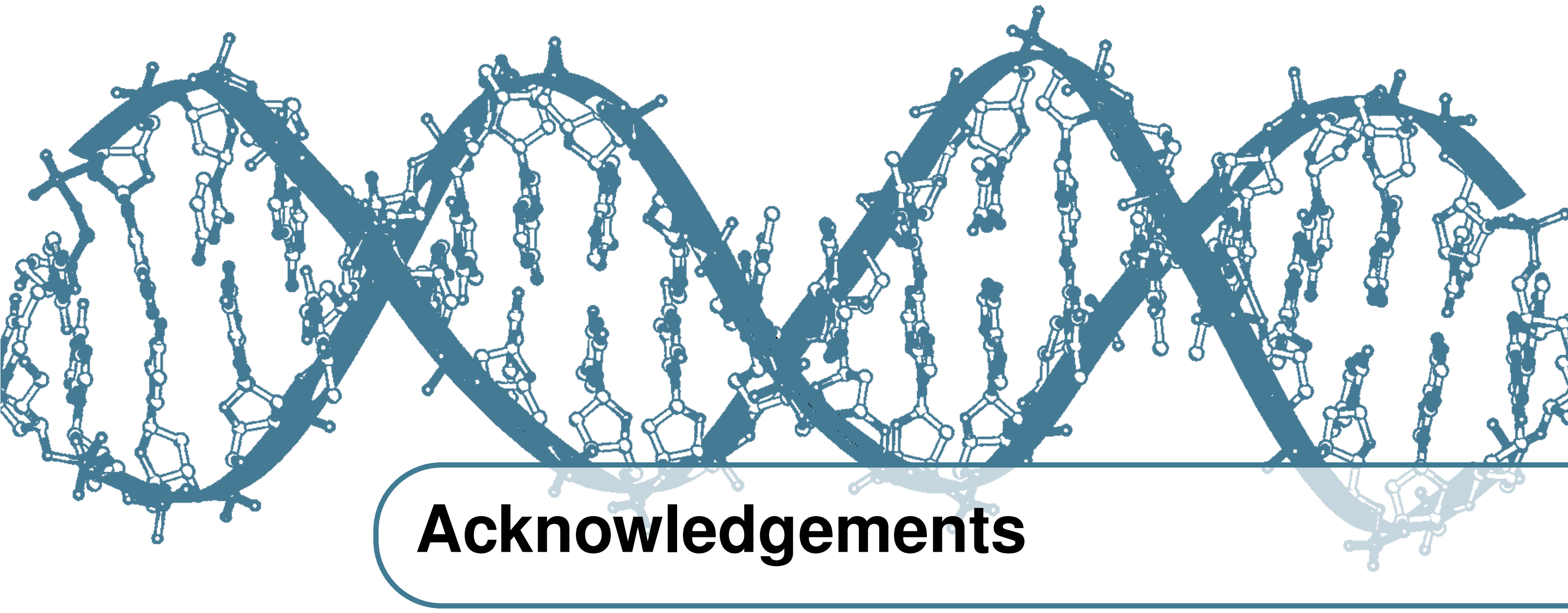
- **Prof. Dr. Pieter De Bleser**
  Department of Biomedical Molecular Biology
  Ghent University
  VIB Inflammation Research Center

- **Dr. Bram De Craene**
  Department of Biomedical Molecular Biology
  Ghent University
  VIB Inflammation Research Center

- **Prof. Dr. Klaas Vandepoele**
  Department of Plant Biotechnology and Bioinformatics
  Ghent University
  VIB Department of Plant Systems Biology

- **Prof. Dr. Ir. Jo Vandesompele**
  Department of Pediatrics and Medical Genetics
  Ghent University

- **Prof. Dr. Ir. Stein Aerts**
  Center for Human Genetics
  Catholic University of Leuven



Acknowledgements

First of all I would like to thank my promoters, Prof. Dr. Frans Van Roy and Prof. Dr. Pieter De Bleser for making this research possible. Thank you for the insightful discussions but also for the many pleasant conversations we had during my PhD period.

I am also very grateful to Bart Hooghe for the nice collaboration we had over the years, and the joint work that we have produced.

A big thank you to my colleagues, Arne, Paco and Liesbet for providing me with distraction during the hard times and the enjoyable conversations during our lunch breaks.

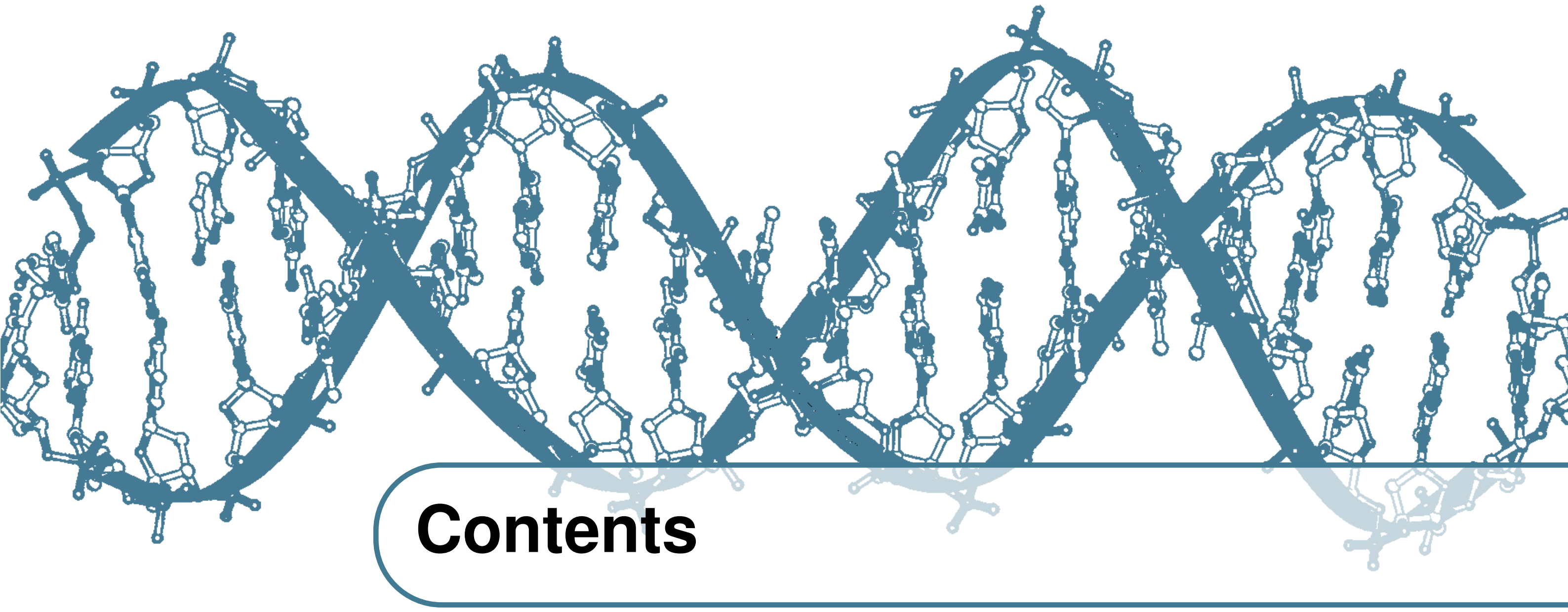
To my wife Karo, thank you for being there for and with me all the time. I love you so much! Thank you also for helping me so much by proofreading and helping with the formatting of some parts of this thesis (in addition to all your support in general!).

Special thanks to my parents for supporting me during my studies, both financially and emotionally and also to my brother Jelle and his girlfriend Kelly for the fun times we had together.

Hannes, Jasper, Pieter and Michiel, you were fantastic during our everyday coffee break (and outside). Without you I would not have survived my time at the VIB.

Tom, Alex, Dieter, Marlies, Lorin, Greet, Jordi, Michiel, Kevin, thanks for the fantastic moments we had (and will have).

I also want to acknowledge the IWT and UGent for supporting this work financially.



Contents

Examination Committee iii

Acknowledgements v

Glossary xi

List of Acronyms xvii

Summary xxiii

Samenvatting xxv

1 Introduction 1

1.1 The central dogma 1

1.2 A gene centric view 2

1.3 Junk DNA 3

1.4 Growing interest in the regulation of the genome 4

1.4.1 Transcription factors 4

1.4.2 Epigenetic modifications 6

1.4.3 Regulatory RNA molecules 8

1.4.4 Combinatorial effects of regulatory RNA molecules and transcription factors 11

1.4.5 Protein stability and localization 11

1.5 Experimental methods for the detection of TFBSs 12

1.5.1 Low-throughput methods 13

1.5.2 High-throughput methods 15

1.6 Bioinformatics approaches 17

1.6.1 Classical bioinformatics methods 17

1.7 Improving the accuracy of the *in silico* methods 24

1.7.1 Phylogenetic methods 24

1.7.2 Other approaches 27

1.8 The ENCODE project: a massive amount of data 30

1.9 Beyond the ENCODE project 31

1.10 Scope of this thesis 32

2 ConTra version 2011 35

2.1 Abstract 35

2.2 Introduction to ConTra v2 36

2.3 New features 37

2.4 Similar tools 42

2.5 Technical details and four-step analysis process 43

2.6 Author contributions 44

3 PhysBinder algorithm 45

3.1 Abstract 45

3.2 Introduction 46

3.3 Material and methods 48

3.3.1 Data 48
3.3.2 Structural characteristics 49
3.3.3 Random Forest algorithm 54
3.3.4 Building classification models 55
3.3.5 Evaluation of classification models 57

3.4 Results 58

3.4.1 Classification accuracy 59
3.4.2 Comparison with alternative sequence-based methods 64
3.4.3 Evaluation of a model on external data 70
3.4.4 Features contained in the models 70
3.4.5 Biological relevance of the selected features 72

3.5 Discussion 78

3.6 Author contributions 81

4 PhysBinder web-tool 83

4.1 Abstract 83

4.2 Introduction 84

4.3 Input and output 85

4.3.1 Input 85
4.3.2 Output 85
4.3.3 An example... 86

4.4 Technical details 88

4.5 Performance of currently available PhysBinder models 89

4.6 Author contributions 89

5 General discussion 93

5.1 What contributes to binding site specificity? 94

5.1.1 Direct readout 94

5.1.2 Indirect readout 96

5.1.3 *In vivo* problems 97

5.2 Why use prediction methods in a ChIP-seq world? 97

5.2.1 Ease of use 98

5.2.2 Validate the effect of SNPs 98

5.2.3 Choking on 10k genomes 99

5.3 Which approach is best? 100

5.3.1 PhysBinder or ConTra? 100

5.3.2 Other approaches that use structural characteristics 100

6 Future perspectives 103

6.1 Improvements of the Random Forest method 103

6.1.1 Aligning on a structural level 103

6.1.2 Integrating different data sources 104

6.2 What is next? 105

6.2.1 Micro-RNA target prediction 105

6.2.2 Integration of micro-RNA target prediction with TFBS predictions 106

6.2.3 PhysBinder explorer 106

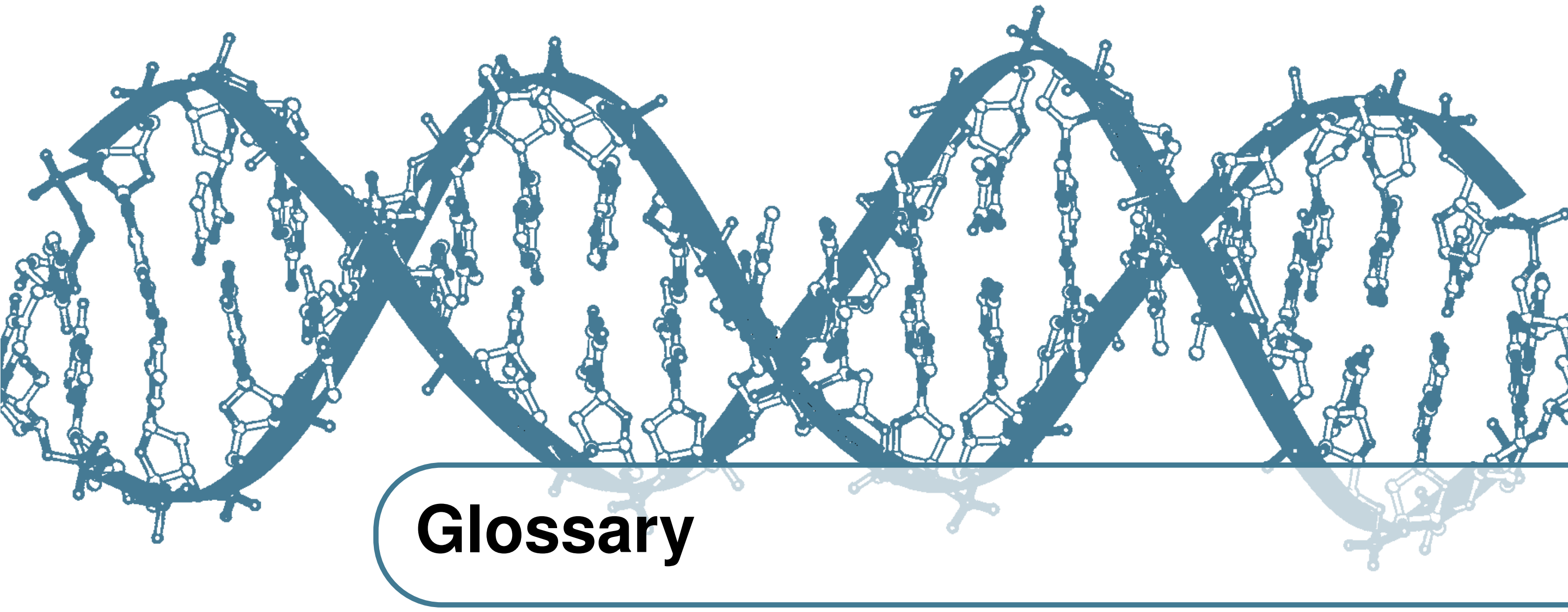
6.2.4 Cluster analysis with PhysBinder models 108

6.2.5 Correlating model score with peak score 109

6.2.6 Conclusion 109

Curriculum Vitae 135

A Appendix 139



Glossary

acetylation: a chemical reaction that attaches an acetyl group to another chemical compound. This reaction is important in molecular biology in a process called epigenetic regulation. In this process, an acetyl group is coupled to the tails of histones.

alignment: a representation of multiple sequences, where the sequences are arranged in such a way that similar parts of the different sequences are below each other. These alignments are often used in bioinformatics to get an idea about how conserved a region in the genome is through evolution.

biophysics: research field that studies the physics behind many biological processes. For example, the study of the cellular organization, physiology, energetics or dynamics of biological systems, and the mechanics of living organisms. In this thesis biophysics means the energetics and dynamics of the DNA molecule.

bromodomain: a recognition domain that is often found in regulatory proteins. This domain recognizes acetylated lysines on other proteins. This protein domain is very important for epigenetic regulation and the maintenance of epigenetic regulation.

classifier: a machine learning algorithm that can be taught to separate data into distinct classes. It is a supervised machine learning approach, which means that the algorithm relies on a training procedure. In this training procedure the algorithm learns which features are most informative for the separation in the different classes.

degenerate motifs: a sequence motif in which not every base pair is conserved in the alignment. In some parts of the motif, for example, multiple nucleotides are possible. To look for sequences that match a degenerate motif one has to use a technique called fuzzy matching.

Dnase I hypersensitive site: a site that is sensitive to cleavage by Deoxyribonuclease I. Sites that are sensitive to Dnase I are mostly active promoter regions. As a result of transcription factor binding, the histones are displaced at this location, making the site accessible for cleavage by Dnase I. These cleavage patterns can be statistically analyzed and used to delineate open chromatin.

Epigenetic modifications: modifications to the DNA that do not change the actual DNA sequence. Examples of these modifications are methylation of the DNA and modifications to histone proteins. Modifications to histone proteins are often described by a specific nomenclature. This nomenclature consists of the name of the histone (for example H3), followed by the letter code of the modified amino acid (e.g. lysine - K), followed by the type of modification (Me for methylation, Ac for acetylation, ...), followed by the number of modifications.

feature selection: this is the practice in machine learning of selecting a representative subset of features. This subset is then used to build a model. By reducing the effective number of features, the models become more sparse, as redundant features are removed.

futility theorem: the futility theorem was originally described by Wasserman and Sandelin in 2004 [1]. It states that most of the *in silico* predicted transcription factor binding sites are false positive predictions when tested in *in vivo* systems.

F-measure: or F1 score is a measure of the accuracy of a classifier. It is calculated by taking the harmonic mean of precision and recall (see formula 0.1).

Formula 0.1

$$\frac{2 \times precision \times recall}{precision + recall} \tag{1}$$

gel electrophoresis: gel electrophoresis is a technique that is used to separate large macromolecules. In molecular biology this technique is used to separate DNA, proteins and RNA molecules.

GFP: Green Fluorescent Protein or GFP is a protein that reflects green light when lit with an ultraviolet light source. It is commonly used as a labeling protein, in which it is fused to other proteins. The green fluorescence is used to report the location of the fusion protein.

histone: a protein that is responsible for the organization of the DNA. By wrapping around the histone protein, DNA becomes packaged into a nucleosome complex. These complexes have an important function in epigentic gene regulation. Due to methylation and acetylation of the histone tails, the DNA can be altered from an accessible state (open) to an inaccessible state (closed).

IUPAC: the International Union of Pure and Applied Chemistry is an international body that is responsible for rules nomenclature in chemistry. In molecular biology IUPAC is best known for the extended alphabet for nucleotides and amino acids.

***in vivo*:** means “in living”. It is used to indicate that experiments were done in living organisms or cells, compared to other test conditions.

in vitro: means “in glass”. This is a term often used to indicate that experiments were conducted in isolated fractions of a living organism. These experiments are also called “test tube experiments”.

in silico: experiments that are carried out *in silico* are performed on a computer or using computer simulations. More generally, in molecular biology this term is used to refer to bioinformatics or computational biology approaches.

information content: or IC is a metric from the field of information theory that is used to measure how different the distribution of a variable is compared to that of a random variable. It is often measured in bits or nats. In bioinformatics the metric is commonly used to calculate the quality of a PWM. The calculation of IC for a PWM is explained in formula 0.2.

Formula 0.2 D_i is the information content expressed in bits in position i of the PWM; $p_{b,i}$ is the probability of base b on position i.

$$D_i = 2 + \sum_b p_{b,i} log_2 p_{b,i} \tag{2}$$

junk DNA: junk DNA is DNA without an apparent function. Most of the junk DNA consists of pseudogenes and inactive retroviruses or transposons. Initially this term was used too liberal and almost all parts of the genome that are not coding for genes were called junk DNA. However, with the publication of the recent ENCODE experiments much of these early conclusions are disproved.

luciferase: is a protein that is used for bio-luminescence. It was first found in the firefly in which it is responsible for lighting up the body of the fly. In molecular biology it is used as a reporter protein. It is often fused to a promoter or a regulatory region to check the activity of the promoter, or regulatory region. Luciferase is also often used in microscopy and other types of imaging.

microarray: is a chip that is covered with a large number of experimental conditions. These conditions can be anything from antibodies to a string of DNA. It is often used in high-throughput experiments because it allows the rapid screening of thousands of experimental conditions.

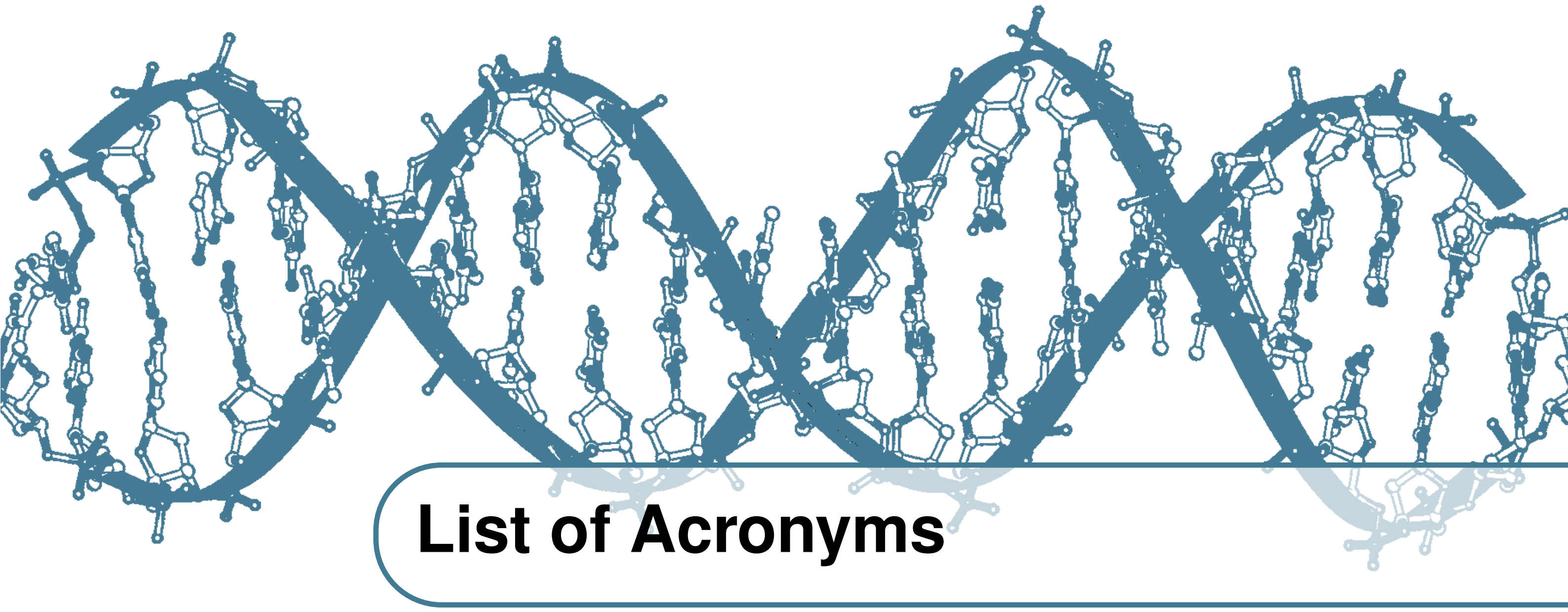
PCA: principal component analysis is a statistical procedure to reduce the dimensionality of a feature set. It does so by identifying the most important gradients in the data. PCA can be used to select the most important features of a data set. In PCA features are categorised in principal components. The first principal component is descriptive for the majority of the variance of the data set. The second component is second most important and so on.

phylogenetics: is an active field of research that is interested in the evolutionary links between species. With the help of multiple sequence alignments, phylogenetic researchers aim to unravel the evolutionary distance between the species. Species that have a high level of sequence conservation are assumed to be close descendants and *visa versa*.

pseudogene: a location in the genome with a strong resemblance to a functional gene that is not functional (anymore) due to the large amount of mutation the gene has accumulated (both in the regulatory regions and coding sequence). However, parts of pseudogenes can become active again at a later stage, for example by contributing exons to active genes. Pseudogenes are an important source of genetic variability.

Random Forest: is a machine learning method that uses an ensemble of classification and regression trees. In each node of the tree, a random set of features is selected that will make all decisions in that node. Each individual tree of the Random Forest is build with a bootstrap sample of the data set. The instances that were not used to build that particular tree are used to get an estimate on the error of the tree. Decisions are made by a majority vote of all the trees in the forest.

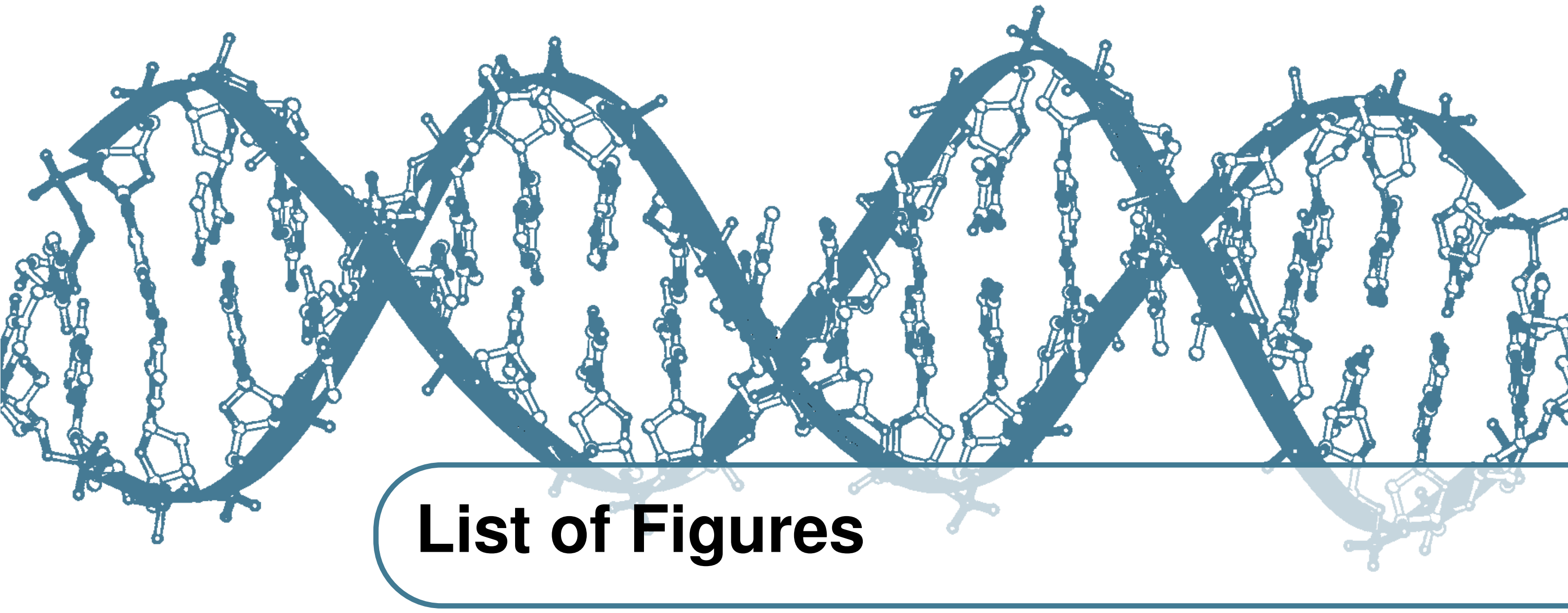
transposon: a mobile element in the DNA that can relocate. Most often these are non-coding elements, although some exceptions are known.



List of Acronyms

AUC: area under the curve
BIRS: best incremental ranked subset scheme
BLAT: BLAST-like alignment tool
bp: base pair(s)
CART: classification and regression tree
ChIP: chromatin immunoprecipitation
CRM: cis regulatory module
DBD: DNA binding domain
DE: direct evidence
DNA: deoxyribonucleic acid
EM: expectation maximization
EMSA: electrophoretic mobility shift assay
ENCODE: Encyclopedia of DNA Elements
FAIRE: Formaldehyde-Assisted Isolation of Regulatory Elements
FFL: feed-forward loop
FBL: feedback loop
FPR: false positive rate
GFP: Green Fluorescent Protein
GTF: general transcription factor

HGP: Human Genome Project
HMM: Hidden Markov model
HRE: hypoxia-response element
IUPAC: International Union of Pure and Applied Chemistry
JVM: Java Virtual Machine
MAF: multiple alignment format
MEME: Multiple Expectation Maximization for Motif Elicitation
minFN: minimal false negatives
miRNA: microRNA
modENCODE: Model Organism Encyclopedia Of DNA Elements
mRNA: Messenger RNA
MSA: multiple sequence alignment
NPD: nucleotides positional dependency
PAF: putative associated factor
PBM: protein-binding microarray
PCA: principal component analysis
PCR: polymerase chain reaction
PFM: positional frequency matrix
PWM: positional weight matrix
RE: response element
RF: Random Forest
RISC complex: RNA-induced silencing complex
RNA: ribonucleic acid
ROC: Receiver Operating Characteristic
SBE: sequential backwards elimination
SELEX: Systematic Evolution of Ligands by Exponential Enrichment
SNP: single-nucleotide polymorphism
SVM: support vector machine
TBA: threaded blockset aligner
TF: transcription factor
TFBS: transcription factor binding site
TPR: true positive rate
UTR: untranslated regions



List of Figures

1.1 The central dogma 2
1.2 Overview of transcriptional regulation mechanisms 5
1.3 Readers, writers and erasers of histone code 8
1.4 Biogenesis of small interfering RNAs and microRNAs 10
1.5 An overview of the ChIP-seq experimental procedure 18
1.6 Gibbs motif finder 23
1.7 An example of ConSite output 26
1.8 Overview of the ENCODE project 32

2.1 Pie chart of reference species 38
2.2 Evolutionarily conserved mechanism for microRNA-223 regulation . 41

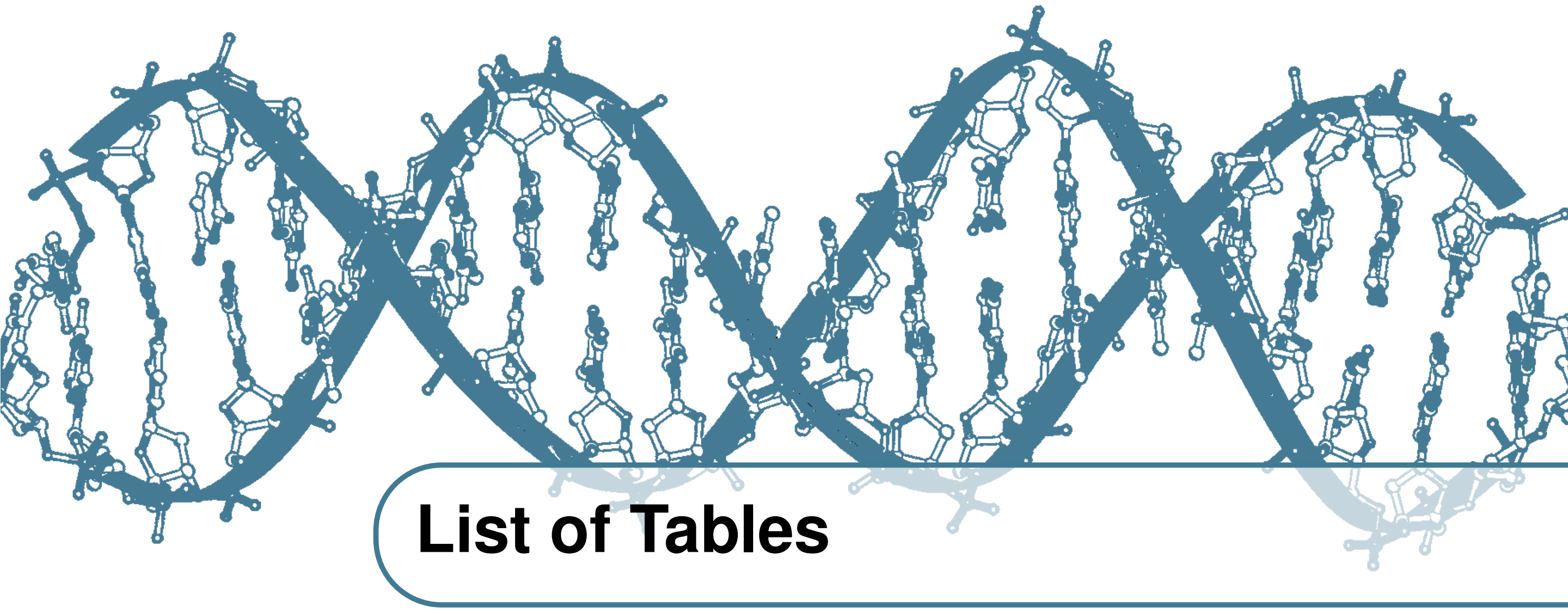
3.1 Overview of some base pair step parameters 51
3.2 Overview of our Random Forest approach 56
3.3 Accuracy of the eukaryotic classification models 60
3.4 Accuracy of the prokaryotic classification models 65
3.5 Pearson correlation analysis of the different features 71
3.6 Visualization of our integrative model for SP1 72

4.1 Example output of the PhysBinder tool 87

4.2 Visualization of PhysBinder results in the UCSC Genome Browser . 88

5.1 Graphical illustration of the different readout mechanisms 95

6.1 Output of the mirTra Prediction tool 107



List of Tables

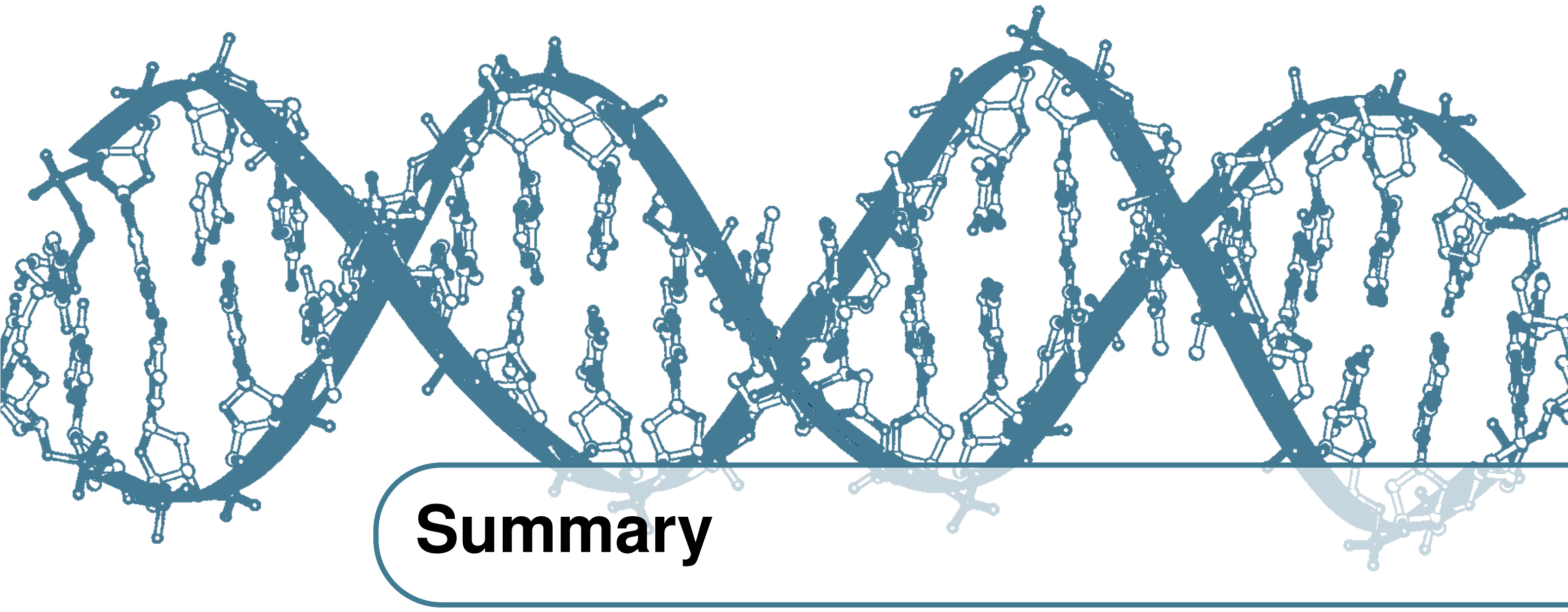
1.1 Nucleic acids IUPAC codes 19
1.2 Examples of biophysical and structural features 29

2.1 Distribution of TFBSs across the genome 37
2.2 Reference species in ConTra v2 39
2.3 Relative importance of the reference species in terms of PubMed hits 40

3.1 Accuracy of eukaryotic classification models; Part A 61
3.2 Accuracy of eukaryotic classification models; Part B 62
3.3 Accuracy of prokaryotic classification models; Part A 66
3.4 Accuracy of prokaryotic classification models; Part B 67
3.5 Accuracy of prokaryotic classification models; Part C 68
3.6 Performance of the TBP model on external ChIP-seq TBP data set . 70
3.7 Results of the PCA analysis for the eukaryotic transcription factors . 73
3.8 Results of the PCA analysis for the prokaryotic transcription factors . 74

4.1 ROC AUCs of currently available transcription factor models in PhysBinder 90
4.2 Continuation of table 4.1 91

6.1 Example of PhysBinder explorer output 108

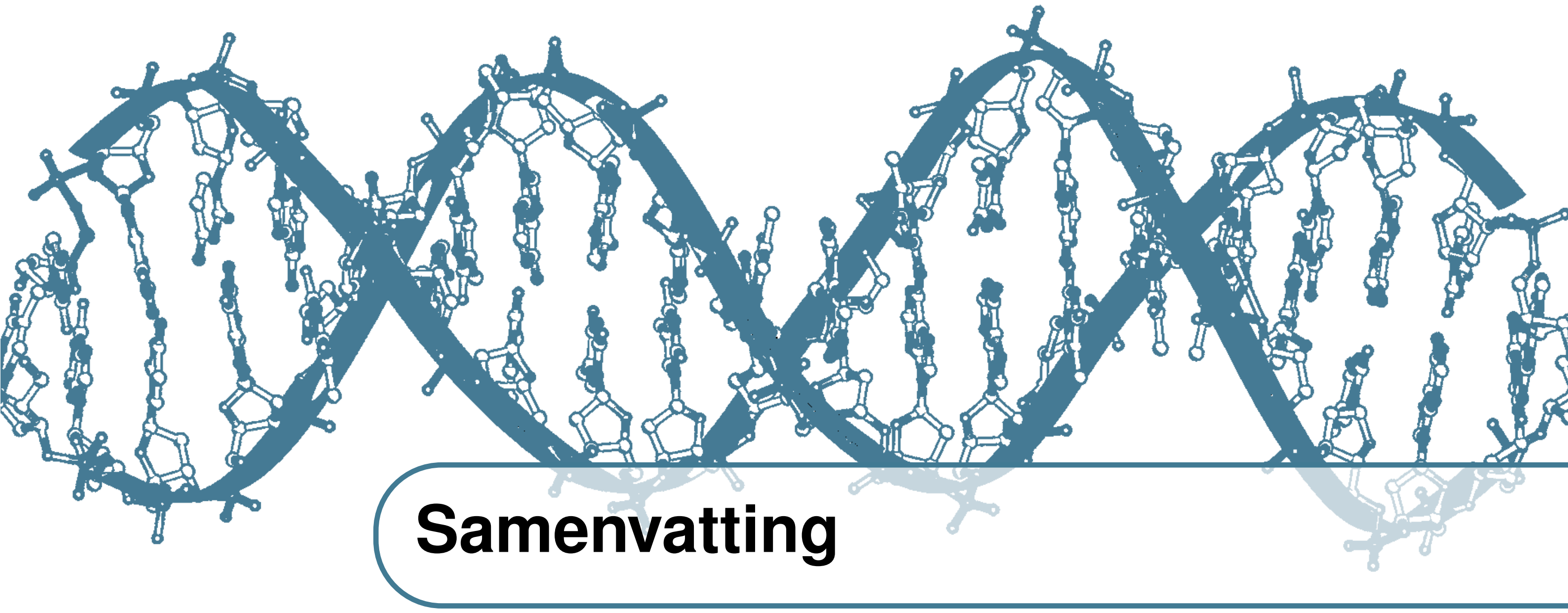


Summary

Modeling the specificity of transcription factors to the DNA is one of the challenges that has kept many bioinformatics researchers busy since the early beginnings. Initially it was expected that a universal recognition code describing the amino acid to base pair contacts would be able to describe protein-DNA complex formation. However, until this very day a universal recognition code has not yet been found and alternative methods became more important. Nowadays, methods that describe the specificity of only one transcription factor (or a small family of transcription factors) are used most often. These methods make use of a set of experimentally validated binding sites to construct a profile for each transcription factor. One of the oldest profile-based methods is the consensus sequence method. Consensus sequences consist of a simple text string in which each character of the string represents the most prevalent nucleotide in the corresponding position of DNA binding sites. As an extension to these consensus sequences, in 1982, Gary Stormo introduced the well-known and very popular positional weight matrix (PWM). These PWMs consist of a 4xL matrix, with L being the length of the binding sites. In each row of these matrices, the frequency of occurrence of one of the four nucleotides is given for a certain position in the binding sites. Even though these PWMs are a big improvement to the consensus sequences method, they also lead to many false positive predictions. Many alternative methods try to improve the accuracy of these PWMs, most of them with very limited success. In this thesis I will discuss the shortcomings of the previous generation of prediction methods and I will suggest new methods that overcome some of these shortcomings.

The first method that will be discussed in this thesis makes use of a multiple sequence alignment (MSA) to visualize evolutionary conserved transcription factor binding sites that are predicted with the PWM method. Binding sites that are conserved across all species in these alignments have a higher likelihood to be functional. Mutation of these binding sites would result in a less fit species, therefore mutations in these binding sites would have a negative effect. By inspecting these multiple sequence alignments for putative PWM hits we can reduce a large number of false positive predictions as false positive hits are less likely to be conserved.

A second contribution of this thesis to the improvement of prediction methods is the research on and development of a number of new methods that make use of the structure and the biophysical characteristics of protein-DNA complexes. These characteristics are often overlooked in the previous generation of prediction methods even though they are very important for binding specificity in many protein-DNA complexes. With the help of the Random Forest classification method and sequence-based structural and biophysical characteristics we managed to develop models that can predict transcription factor binding sites with a higher level of accuracy. Based on this method, we also developed a user-friendly web-tool that can make use of a large number of pre-calculated transcription factor models.

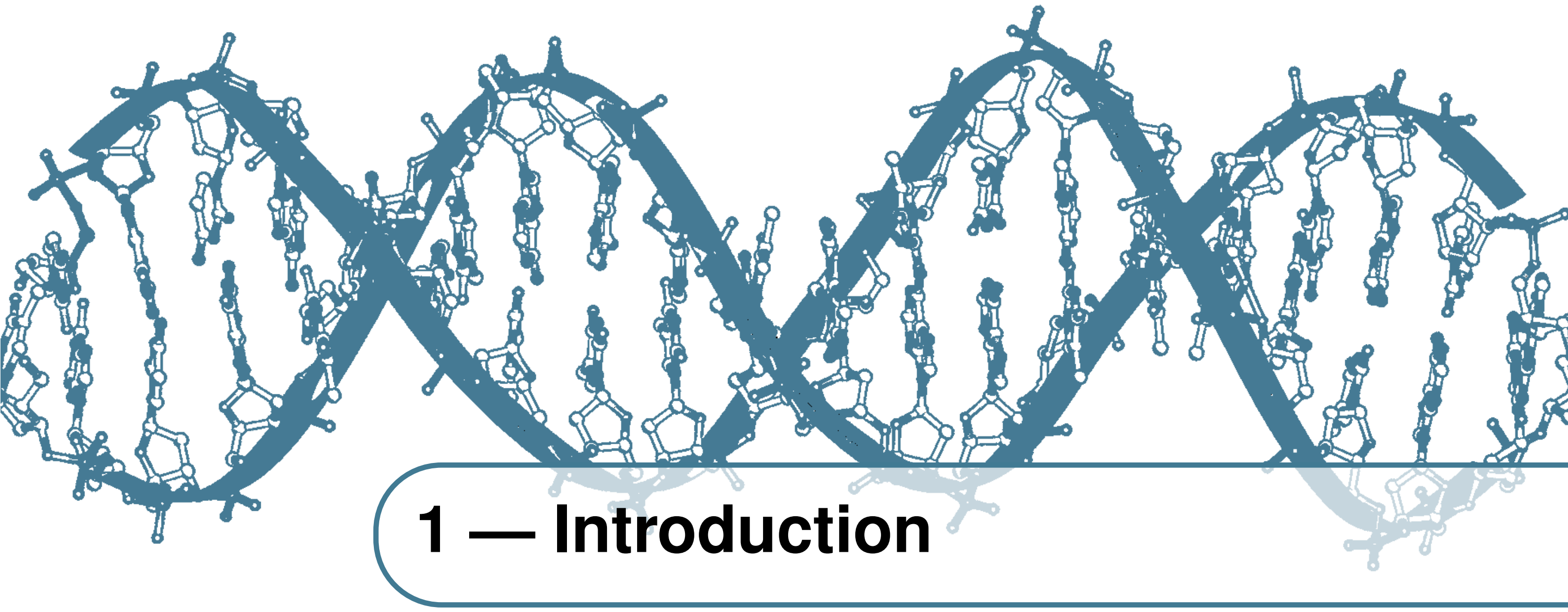


Samenvatting

Het modelleren van transcriptiefactor-DNA bindingsspecificiteit is een uitdaging die reeds vele generaties van wetenschappers bezig gehouden heeft. Initieel werd verwacht dat de specificiteit van DNA-bindende eiwitten simpelweg te beschrijven valt met behulp van een universele herkenningscode. Al snel bleek dat het vinden van deze universele code onmogelijk was en alternatieve ideeën wonnen aan terrein. Zo is het eenvoudiger gebleken om per transcriptiefactor (of per familie van transcriptiefactoren) een apart profiel op te stellen, gebruikmakende van reeds gekende bindingsplaatsen. Een eerste methode om op eenvoudige wijze het bindingsprofiel van een transcriptiefactor voor te stellen maakte gebruik van consensus sequenties en tekst-zoekalgoritmen. Consensus sequenties bestaan uit een tekst *string* waarbij voor elke positie in de bindingsplaats het meest voorkomende nucleotide gekozen wordt. Een bekend voorbeeld van zulke consensus sequenties is de sequentie van de TATAbox bindingsplaats (TATAAA). In navolging van deze consensus sequenties introduceerde Gary Stormo in 1982 de alom bekende *positional weight matrices* (PWMs) die tot op de dag van vandaag behoren tot één van de populairste methoden om transcriptiefactor bindingsplaatsen te modelleren. Deze matrices bevatten voor elke positie in de bindingsplaats een probabiliteitsscore voor elk van de vier nucleotiden. Alhoewel door de komst van deze PWMs vele van de tekortkomingen van de eenvoudige consensus sequenties verdwenen, geeft ook deze methode aanleiding tot een enorm grote hoeveelheid aan fout-positieve voorspellingen. Vele alternatieve methoden proberen deze fout-positieve voorspellingen te reduceren, maar met wisselend succes. In deze thesis vertrek ik van de problemen van de vorige generatie aan methodes, en suggeer ik een aantal nieuwe manieren om deze problemen aan te pakken.

Een eerste methode die ik in deze thesis zal bespreken maakt gebruik van *multiple sequence alignments* (MSAs) voor de visualisatie van evolutionair geconserveerde bindingsplaatsen, voorspeld met behulp van de eerder besproken PWM methode. Bindingsplaatsen die geconserveerd zijn doorheen meerdere species zijn met een hogere waarschijnlijkheid echte functionele bindingsplaatsen. Mogelijks zijn deze bindingsplaatsen van belang voor het species waardoor ze behouden worden. Door PWM voorspellingen in *multiple sequence alignments* uit te voeren kunnen reeds een heel aantal fout-positieve voorspelde bindingsplaatsen visueel gedetecteerd worden.

Een tweede bijdrage die deze thesis levert aan het verbeteren van voorspellingsmethodes is het onderzoek naar en het ontwikkelen van een set aan nieuwe methodes die gebruik kunnen maken van de structuur en de biofysische kenmerken van eiwit-DNA complexen. Deze kenmerken worden meestal over het hoofd gezien in de vorige generatie predictiemethodes maar zijn wel van groot belang in het verkrijgen van specificiteit bij eiwit-DNA binding. Met behulp van de *Random Forest* classificatiemethode en op sequentie gebaseerde structurele en biofysische kenmerken slaagden we er in om modellen op te stellen die met een hogere nauwkeurigheid bindingsplaatsen voorspellen. Met behulp van deze nieuwe methode ontwikkelden we tenslotte een gebruiksvriendelijke webtool die reeds een groot aantal transcriptiefactor modellen bevat.



1 — Introduction

1.1 The central dogma

The central dogma of molecular biology was first proposed by Francis Crick in 1958 [2]. It states that DNA (a gene) is converted into RNA (this process is called transcription) and subsequently RNA is converted into proteins (this process is called translation). Importantly, the central dogma maintains that the reverse pathway is not possible: a protein cannot be translated in RNA and RNA cannot be transcribed back into DNA. In the traditional view on molecular biology, DNA serves as the blueprint, RNA as the information transporting molecule and proteins as the functioning unit of our cells (see figure 1.1). However, since this dogma was first introduced by Francis Crick, a lot of exceptions have been found to it [3], and it is becoming clear that the central dogma is not a very accurate representation of reality. A first example of an exception to the dogma is the discovery of enzymes that can reverse the transcription of RNA back into DNA [4, 5]. In addition, many non-translated RNA molecules have been identified [6] such as tRNAs, micro RNAs (miRNAs) and long non-coding RNAs (lncRNAs). These RNA molecules are not encoding for proteins but have many specific functions such as transcriptional regulation (enhancer lncRNAs), post-transcriptional regulation (miRNAs), epigenetic regulation (lncRNAs) and even

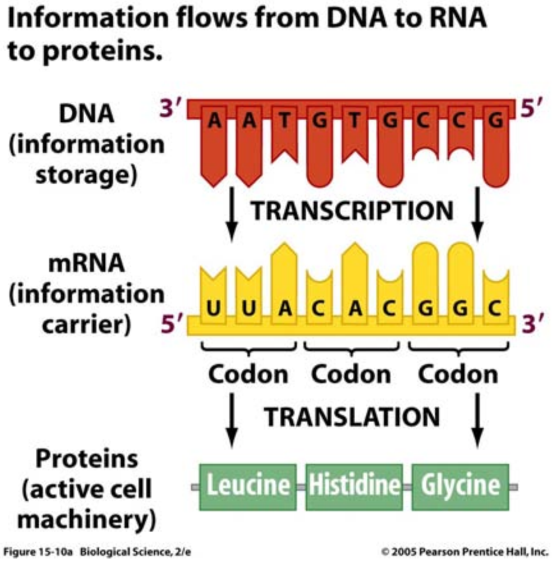

Information flows from DNA to RNA to proteins.
DNA (information storage)
3′
A A T G T G C C G
5′
TRANSCRIPTION
mRNA (information carrier)
5′
U U A C A C G G C
3′
Codon Codon Codon
TRANSLATION
Proteins (active cell machinery)
Leucine
Histidine
Glycine
Figure 15-10a Biological Science, 2/e
© 2005 Pearson Prentice Hall, Inc.

Figure 1.1: The central dogma of molecular biology. In this picture an overview is given of the main information flow that is central to molecular biology. DNA is transcribed into RNA, which in turn is translated into proteins. Since the central dogma was introduced in 1958, a lot of exceptions have been found to it. Picture taken from [7].

regulation of translation (tRNAs).

1.2 A gene centric view

For a long time, geneticists and molecular biologists have looked at the genome from a gene-centric point of view. In the light of the central dogma this misconception is easy to understand: genes are the precursors that are transcribed into RNA and this RNA will eventually be translated into proteins, the functioning units of our cells. From this perspective, genes were considered the functional blueprints in the genome. However, this view has led to an unduly focus on the functional analysis of genes while other parts of the genome have remained unexplored for a long period of time. At first, the number of functional genes was estimated around 100,000 [8, 9]. Interestingly, upon the completion of the Human Genome Project in 2003 the estimated number of genes has dropped steadily. Currently, researchers estimate that the number of genes is approximately 39,000 (20,687 protein-coding genes, 8800 small RNA molecules and

9600 long non-coding RNA molecules) [10], which is much less than the 100,000 genes biologists initially expected to be necessary to create a complex organism like the human being. Strikingly, the coding portions only account for ± 2% [10] of the DNA in the genome. What is the function of the remaining 98% of the DNA in the genome?

1.3 Junk DNA

For decades, most of the DNA in the genome was considered “junk DNA” [11], which means: DNA without a real function. Only genes and a small promoter region just upstream of the transcription start site were deemed functional by most scientists. The remainder of the DNA outside these regions was considered replication errors, pseudogenes and transposons. This view has changed dramatically with the recent publication of the ENCODE experiments [12]. Not only did the ENCODE consortium discover that up to 8% of the genome consists of transcription factor binding sites (a number that is expected to become even larger with the discovery of additional transcription factors), the project also revealed that up to ± 80% of the genome can be expected to be functional. This percentage is much larger than the initial estimates. However, since the publication of the ENCODE summary paper [10], the percentage of functional genome was received with some skepticism by a number of researchers [13, 14].

In this thesis I define a functional element as an element that enhances the fitness of an organism, or that increases the reproduction rate of the organism. I do believe that the percentage of functional elements as stated by the ENCODE project is an exaggeration, but nevertheless applaud the reopened debate about the nature of “junk” DNA. More recently a distinction was made between “junk” DNA and “garbage” DNA. “Junk” DNA is DNA that is neutral enough to keep. Garbage DNA should be removed from the genome because it is impairing an organism’s fitness. To quote Nobel Laureate Sydney Brenner:

“Some years ago I noticed that there were two kinds of rubbish in the world and that most languages have different words to distinguish them. There is

the rubbish we keep, which is junk, and the rubbish we throw away, which is garbage. The excess DNA in our genomes is junk, and is there because it is harmless, as well as being useless, and because the molecular processes generating extra DNA outpace those getting rid of it. Were the extra DNA to become disadvantageous, it would become subject to selection, just as junk that takes up too much space, or is beginning to smell, is instantly converted garbage."

However, calling large parts of the DNA "junk" because of the apparent lack of conservation or observable function is often too simplistic and highly unwanted in my opinion. For example, many parts of the DNA carry functions that are not yet known, such as regulatory functions.

1.4 Growing interest in the regulation of the genome

1.4.1 Transcription factors

A precise regulation of transcription is indispensable for the correct development and functioning of all living organisms. One of the most well-known mechanisms of transcriptional regulation is the regulation performed by transcription factors (see figure 1.2).
Transcription factors are a distinct class of proteins that influence the rate of transcription by binding to the DNA. The DNA binding sites to which these proteins bind are transcription factor binding sites (TFBSs) and they mostly consist of short degenerative motifs (motifs with a low degree of conservation). These TFBSs or response elements (REs) are recognized using electrostatic interactions and van der Waals forces. Note that also the structure of the DNA plays a major role in the recognition of TFBSs, as will be discussed in this thesis.

The binding of a transcription factor to the TFBS happens through a domain of the protein called the DNA binding domain (DBD). Sometimes these protein-DNA binding events regulate transcription positively (by recruiting RNA polymerase) whilst other times, they regulate transcription in a negative way (by blocking the recruitment of RNA polymerase). Most of the TFBSs are located in the promoter region, and in general, in enhancer regions. Although both types of regions affect

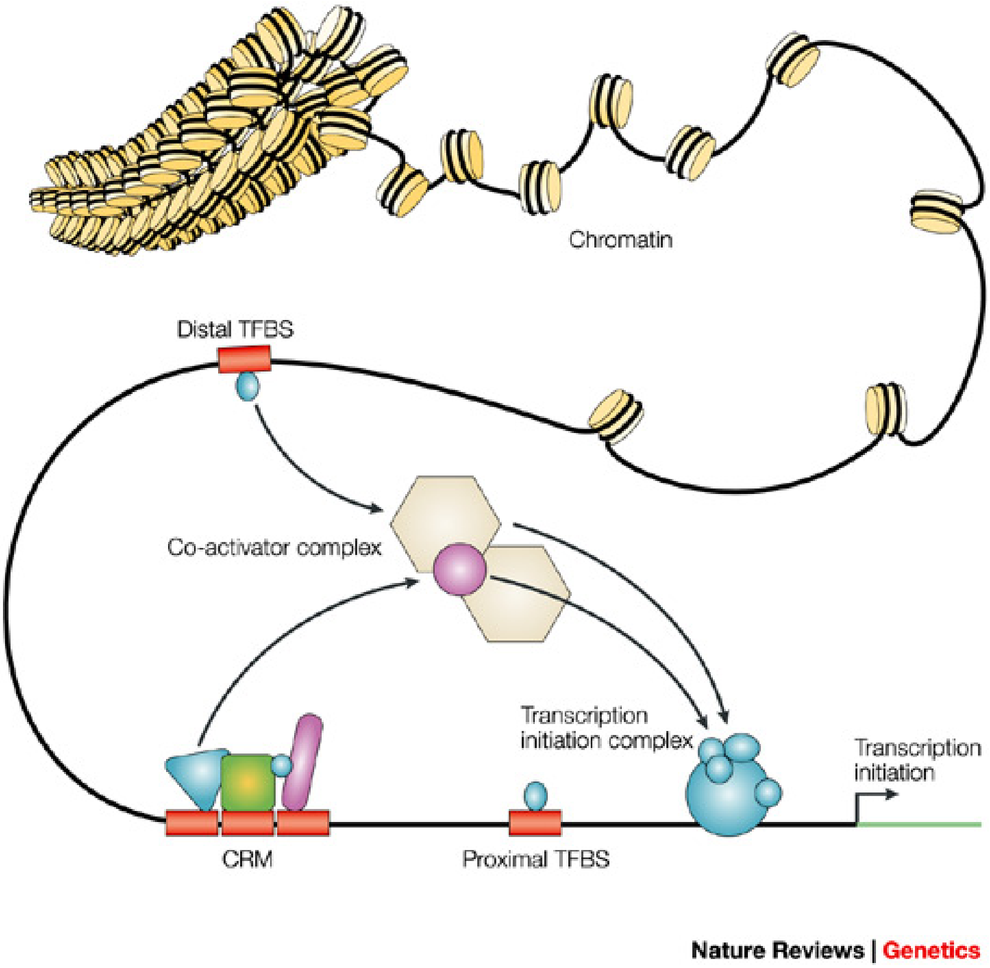


Figure 1.2: A general overview of the mechanisms that are regulating transcription. Taken from a review of Wasserman et al. in Nature Genetics [1]. In this overview, different types of binding sites are listed. Binding sites that are close to the transcription start site are known as proximal TFBSs. Binding sites that are further away are labeled distal TFBSs. The combination of different TFBSs organized in a regulatory module is called a cis regulatory module (CRM).

nearby genes, enhancer regions have been found that influence genes millions of base pairs away [15]. These distant enhancers are thought to interact with RNA polymerase through DNA looping. In some cases, transcription factors can also alter transcription by acetylating or deacetylating the histone proteins.

Transcription factors which are absolutely necessary to initiate transcription in eukaryotes are named the general transcription factors (GTFs) [16]. These GTFs interact directly with RNA polymerase. The other transcription factors outside the set of general transcription factors are mostly responsible for a correct spatio-temporal expression of genes. This implies that most of these non-GTF transcription factors play a major role in developmental processes, the response to signals and cell cycle control.

1.4.2 Epigenetic modifications

Epigenetic modifications concern all modifications to the genome that have no influence on the nucleotide sequence of the DNA. The two most well-known regulatory mechanisms on the epigenetic level are histone modifications and DNA methylation. Epigenetic changes caused by these mechanisms can change the level of transcriptional activity without altering the sequence of the DNA. Some of these epigenetic changes are maintained through cell division (for example cytosine methylation patterns are preserved through cell division with the help of Dnmt1) [17].

As a result of the epigenetic modification, the way in which the DNA is wrapped around the histone proteins is changed and the chromatin becomes remodeled. This remodeling can occur through the modification of one or more of the amino acids of the histone proteins or by the addition of methylgroups to the DNA. An overview of the different types of histone modifications is listed below (more information on the nomenclature of histone modifications can be found in the Glossary under “epigenetic modifications”).

- **H2A.Z** has a close resemblance to histone H2A (it is family member Z of histone H2A). Incorporation of histone H2A.Z is associated with ambient temperature changes (incorporation decreases as temperature rises). These histones are often found near transcription start sites, where they alter transcriptional activity.

- **H3K27ac** is involved in transcription initiation and is generally found in regions of open chromatin.
- **H3K27me3** histones pack the DNA of genes silenced by the polycomb protein. These genes are often silenced developmental genes that have no function anymore in the adult species.
- **H3K36me3** is thought to influence the elongation of RNA polymerase II when transcribing both coding and non-coding genes.
- **H3K4me1** is generally associated with enhancer or UTR regions of the genome.
- **H3K4me2** often co-occurs with CpG islands in promoter and enhancer regions.
- **H3K4me3** marks the promoter of actively transcribed genes or genes that will become active at a later stage.
- **H3K79me2** is associated with the region between the transcription initiation region and the elongation region.
- **H3K9ac** marks an open chromatin structure and actively transcribed regions.
- **H3K9me1** is often found in genomic regions in the euchromatin state.
- **H3K9me3** in contrast to H3K9me1 is associated with a heterochromatin state. It marks silenced and repressed genes.
- **H4K20me1** influences the accessibility of the genomic region. H4K20me1 is found at open and active genes.

There exist a number of different theories which describe how this remodeling has an influence on transcription. The classical theory claims that the methylation of histone tails changes the charge from positive to negative, resulting in different electrostatic interactions between the histone proteins and the negatively charged DNA. This change in electrostatic interaction loosens the binding of the DNA to the histones thus making the DNA more accessible to transcription factors and other parts of the transcriptional machinery. However, the classical explanation

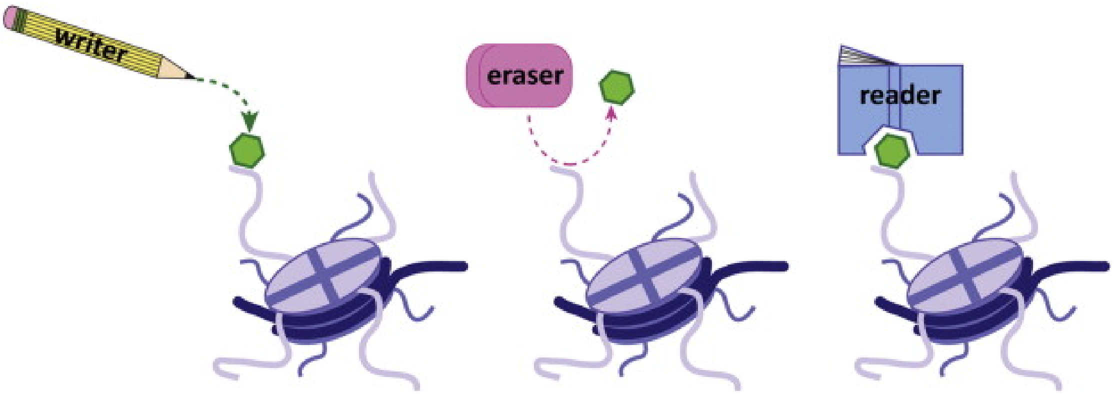


Figure 1.3: Three enzyme functions in epigenetic modifications. "writers" introduce modifications on the histone proteins, "erasers" can remove modifications and "readers" interpret the histone modification code. Figure taken from [18].

can be disproved by the observation that a trimethylation of lysine 9 of histone 3 is often associated with transcriptionally silent DNA. This contradicting observation has encouraged researchers to look for alternative explanations regarding the mechanisms responsible for the epigenetic regulation of transcriptional activity. One of these alternative theories is the "trans" model of epigenetic regulation. This model states that epigenetic modifications may introduce binding sites for certain proteins that can influence the chromatin state. Acetylated lysines, for example, can be bound by proteins that carry a bromodomain (this is a specific protein domain that recognizes acetylated lysines). These domains have been found in many transcription factors and they might be responsible for the changes in transcription and the remodeling of the chromatin. Enzymes that are able to recognize and interpret histone modifications are recently called "readers" of the histone code, enzymes that remove modification marks are termed "erasers" and enzymes that place these modifications on the histones are called "writers" (see figure 1.3).

1.4.3 Regulatory RNA molecules

On a post-transcriptional level, both expression and translation can be regulated through small regulatory RNA molecules. These molecules were first discovered in C.elegans in 1998 by Andrew Fire and Craig Mello [6]. The small RNA molecules are named microRNAs (miRNAs) and they protect our cell against viruses and transposons and they are involved in many biological processes. They are also responsible for the post-transcriptional regulation of protein levels by degradation of the mRNA and they are of great importance in a process called translational repression. Endogenous miRNAs are most often found in between coding genes and are also encountered in the introns of genes.

Regulatory RNA molecules originate in the nucleus when non-coding RNA molecules called primary miRNAs (pri-miRNAs) are processed by a protein called Drosha. Drosha cleaves the pri-miRNAs at the base of hairpin structures into fragments that are known as precursor miRNAs (pre-miRNAs). These pre-miRNAs are exported out of the nucleus and further processed by the RNase III enzyme Dicer. Dicer cleaves the hairpin structure of pre-miRNAs and by doing so it produces an imperfect RNA duplex. This duplex is converted into two single stranded RNA molecules and one of these strands (the guide strand) is incorporated into the RISC complex (RNA-induced silencing complex). The RISC complex, in turn, will cleave complementary mRNAs and repress translation (for more information, see figure 1.4).

Recently, some alternative modes of regulation by miRNAs were discovered in which the RNA molecule induces the expression of certain genes instead of blocking the expression. In these modes of regulation, the miRNA acts as an RNA “transcription factor” that binds to the promoter of the gene that is regulated. For example, it has been shown that miR-373 binds to a complementary site in the promoter of E-cadherin where it is responsible for inducing the expression of E-cadherin [19].

Another important class of RNA molecules that have a regulatory function are the long non-coding RNAs (lncRNAs). This class of RNA molecules consists of long polyadenylated RNAs that do not code for proteins. They are transcribed from intergenic or enhancer regions [21] and they often act on neighboring protein-coding genes. It is thought that enhancer lncRNAs act on their target promoters by establishing loops in the chromatin [22]. LncRNA molecules are also heavily involved in epigenetic regulation. They can act on the epigenetic level by actively recruiting chromatin modifiers as was shown by [23]. However,

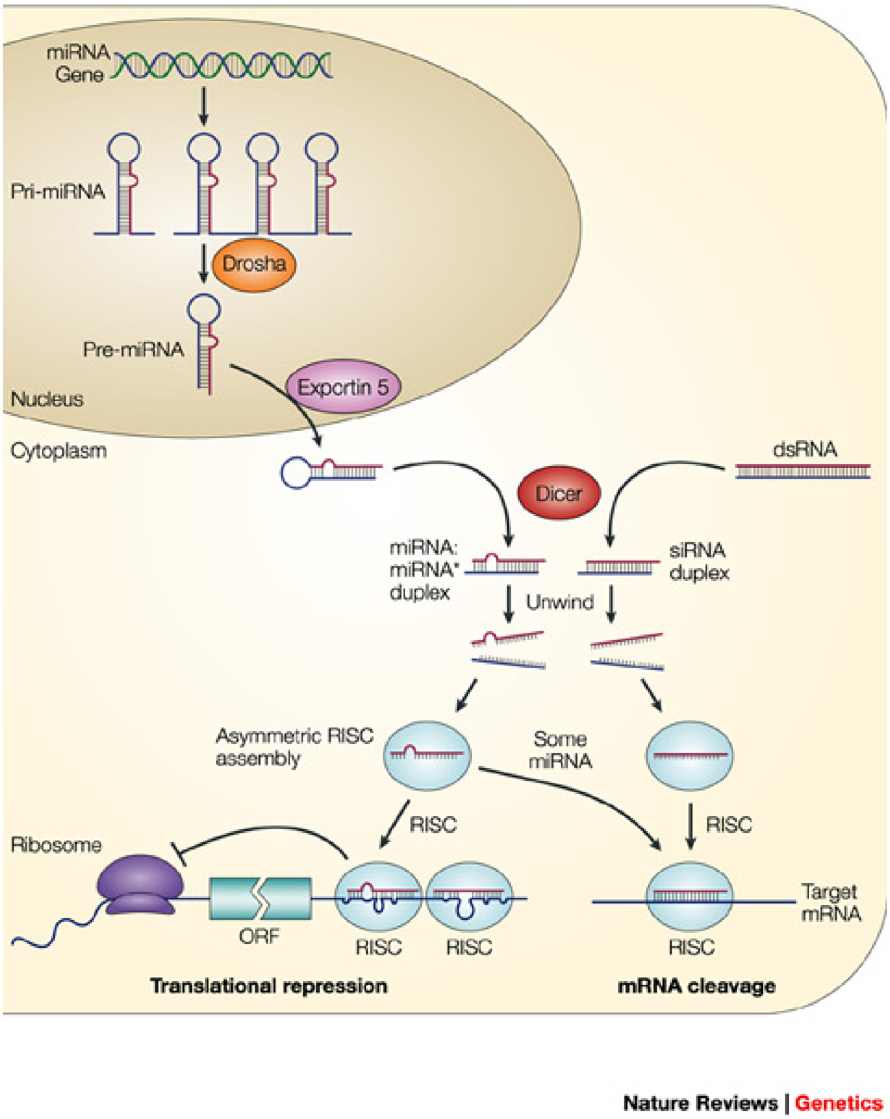


Figure 1.4: An overview of the pathways that are responsible for the biogenesis of small interfering RNAs and microRNAs and the interaction with their target sites. Transcripts are first processed by Drosha into pre-miRNAs. These pre-miRNAs are exported from the nucleus and processed by Dicer into a double stranded miRNA. One strand is incorporated in the RISC complex which cleaves the target site and represses translation. (image taken from [20]).

at this moment we are still discovering novel functions of lncRNAs each day and it is a very exciting area of research nowadays.

1.4.4 Combinatorial effects of regulatory RNA molecules and transcription factors

Transcription factors and miRNAs regulate the expression of genes in a closely related fashion and the regulatory effect of both molecules can be combined. This combination of transcription factors and miRNAs is able to regulate gene expression in a more advanced and precise way [24]. Both molecules can co-regulate the expression of genes but they can also regulate the expression of each other in what is known as regulatory loops. These loops can be reciprocal or unilateral and they can even consist of a double feedback loop [25].

- **unilateral loops**: The expression of the transcription factor is downregulated by the miRNA while the expression of the miRNA is upregulated by the transcription factor.
- **reciprocal loops**: Both the transcription factor and the miRNA regulate each other negatively.

The above loops can drastically reduce the amount of leaky expression [26]. For example, a miRNA can simultaneously repress a target gene and the transcription factor that is responsible for inducing the target gene. In this situation, the target gene can only be expressed when the miRNA is downregulated. A better insight into these types of complex interactions between transcription factors and miRNAs will be of great importance for unraveling many developmental processes and disease conditions.

1.4.5 Protein stability and localization

In addition to the transcriptional and post-transcriptional mechanisms, another important actor that influences cellular expression levels is protein stability. All mechanisms discussed previously control the rate of protein formation. However, if a protein is infinitely stable, the levels of that protein will theoretically reach an infinite concentration over time. In reality, protein concentrations in a cell are dependent both on the formation rate and degradation rate: the concentration of a protein at any moment is the difference between these two rates. This implies that the stability of a protein (which controls the degradation rate) is of equal importance in the regulation of expression levels. Changes in protein stability are often associated with diseases. Important determinants of protein stability are protein size, the number of amino acids that participate in hydrogen bonds and the number of hydrophobic amino acids of a protein [27].

Protein localization is important in the regulation of many cellular processes as it limits the scope of certain actors in the cell. Proteins can be sequestered in the cytoplasm, secreted outside the cell, localized in the mitochondria or in the nucleus. Targeting proteins to specific compartments is achieved with the help of a small polypetide chain of the protein that is known as the signal peptide. These signal peptides mostly reside at the N-terminal of the protein. Compartmentalization and protein localization are of great importance in transcription factor proteins. Transcription factors need to be localized in the nucleus where they can alter transcription. However, if another protein blocks the signal peptide of the transcription factor it cannot move into the nucleus. These blocking proteins often act as regulators of transcription factors.

1.5 Experimental methods for the detection of TFBSs

In this section I will discuss different types of experimental approaches that can be used to detect TFBSs. In particular, two distinct classes of experimental methods are described. First, I will present the low-throughput experimental methods. This class consists of methods that generally prioritize quality over quantity. In most low-throughput methods only one TFBS, or a small number of TFBSs, can be tested simultaneously. These methods are very labour-intensive and relatively costly but they yield results that are of a very high quality. Secondly, the class of high-throughput experimental methods are introduced. These methods allow researchers to validate a large number of TFBSs in one experiment. In general, high-throughput methods return a genome-wide list of binding sites or they return a full profile that describes all possible binding affinities of the transcription factor.

1.5.1 Low-throughput methods

Reporter gene constructs

One of the oldest ways to look for regulatory elements is by using reporter gene constructs. In this technique, a promoter of interest is placed in front of a fluorescent or luminescent reporter gene such as GFP or luciferase. This construct is then inserted into a cell culture or an animal model in which it can become expressed. Step by step, different types of modifications are made to the reporter gene construct such as nucleotide point mutations, deleting certain regions or inserting novel sequences in the construct. After each mutation, the expression level of the reporter gene is evaluated. Mutations that decrease the level of expression (measured by the intensity of the luminescence) are known as disrupting mutations whereas mutations that increase the expression level are termed enabling mutations. The reporter gene technique is very well-suited to characterize a promoter region in fine detail.

With a simple modification, this technique can also be used to study the activity of individual enhancers and binding sites. In this case, a minimal promoter is cloned in front of a reporter gene. However, because this promoter in itself is not sufficient to drive the expression of the reporter gene without additional enhancer sequences, putative regulatory sequences are inserted in front of this minimal promoter. If the regulatory sequence has enhancer capabilities, the reporter gene is expressed, which can be measured by the intensity of the luminescence.

Electrophoretic mobility shift assay

If a quick and low cost assay to study *in vitro* binding of proteins to DNA is preferred, Electrophoretic mobility shift assay (EMSA) can provide a good solution. EMSA is a variant of the electrophoresis technique that can be used to search for interactions between proteins and DNA. In this technique a protein-DNA mixture is separated on a polyacrylamide or agarose gel [28, 29]. Since the DNA fraction and the protein fraction have a different size and shape, the speed in which they migrate through the gel differs. One lane of the gel, the control lane, only contains the DNA fraction. The second lane, on the other hand, holds a protein-DNA mixture. If the protein fraction does not bind to the DNA, two bands will appear in this second lane: one for the DNA fraction and another one for the protein fraction. However, if the protein does bind to the DNA, three bands will appear: one band for the unbound DNA fraction, one band for the unbound protein fraction and a third band for the protein-DNA complex. The additional band is caused by an effect that is known as the band shift. This band shift is the result of the protein-DNA complex being slightly larger and slower moving through the gel. The ratio between the bound fraction and the unbound fraction of DNA can be used to determine the affinity of the protein for the DNA fragment.

To enlarge the shift between the bound and unbound fraction, one can add an antibody that recognizes the protein. By binding to the protein, this antibody makes the protein component bigger and the resulting shift larger. This technique is known as a supershift analysis. EMSA is usually followed by an *in silico* identification of the transcription factor. The DNA fragment is sequenced and the nucleotide sequence is compared to a database of known binding profiles. If the DNA fragment matches a particular binding profile, the search for the identity of the transcription factor becomes a lot easier.

Low-throughput Chromatin immunoprecipitation

Chromatin immunoprecipitation or ChIP is one of the most sensitive *in vivo* methods to detect DNA binding proteins. In Chromatin immunoprecipitation, DNA binding proteins are covalently bound to their *in vivo* TFBSs using formaldehyde or UV radiation. Next, the cells are lysated and the DNA is chopped into pieces using sonification. Special antibodies against the DNA binding protein are designed and the protein together with bound DNA are immunoprecipitated using this antibody. The cross-links between the protein and the DNA are reversed and the protein gets digested, after which the bound DNA fragments will be analyzed with (q)PCR. The low-throughput variant of this technique has mostly been replaced by high-throughput techniques such as ChIP-chip (ChIP on chip) and ChIP-seq (ChIP followed by next-generation sequencing).

1.5.2 High-throughput methods

Recently, technological advances in sequencing technology have led to a whole new branch of high-throughput methods such as the ChIP-seq method. These methods take advantage of new platforms for next-generation sequencing such as Illumina, 454 and Ion torrent. In addition to the next-generation sequencing methods, microarray technology can be used to identify sequences.

Systematic Evolution of Ligands by Exponential Enrichment

Systematic Evolution of Ligands by Exponential Enrichment (SELEX) has a long history in the discovery of drug molecules that bind to the DNA. The SELEX procedure is also easily adaptable for use in the detection of protein-bound DNA. The approach starts from a library of 10^{15} to 10^{16} sequences to which the protein of interest is added. Bound DNA samples are separated from unbound samples and the former are subsequently amplified by PCR. This process is repeated multiple times, causing the affinity of the selected DNA fragments to increase drastically. DNA fragments that survive all rounds of enrichment are sequenced and they are used to build a binding profile (for example a PWM). It is important to note that there is a risk of over-selection, since in reality many transcription factors will also bind to low affinity sites *in vivo*. As a result of this over-selection these low affinity sites remain undiscovered.

Protein-binding microarrays

Martha Bulyk and partners came up with the idea of protein-binding microarrays (PBMs) to study *in vitro* protein-DNA complexes [30]. These PBMs are specially designed microarrays that are coated with an exhaustive list of ungapped DNA k-mers. These k-mers are printed as double stranded DNA probes on the surface of the microarray. A transcription factor of interest is hybridized with the DNA on the array, unbound proteins are washed away and bound transcription factors are detected using labeled antibodies. The labeled microarray is scanned by a specialized device and the image is statistically analyzed. *In vitro* affinities are then calculated from the normalized intensities of the label for each probe on the microarray. These affinities are often converted into positional weight matrices for ease of use. Many examples of PBM PWMs can be found in the JASPAR library of transcription factor profiles.

ChIP-chip

Recent advances in genome-wide tiling arrays (chips) have enabled us to look for *in vivo* TFBSs on a genome-wide scale [31]. These chips contain tiled DNA fragments of the entire genome or of relevant portions of the genome (such as promoters) and they allow the mapping of binding sites across the genome. One of the first high-throughput extensions to classical Chromatin immunoprecipitation was implemented by adding such a genome-wide tiling array chip. Although this innovative approach allowed researchers to search genome-wide for *in vivo* binding sites, the resolution of the array was a limiting factor for species with a larger genome.

ChIP-seq

Due to recent breakthroughs in sequencing technology and the rise of next-generation sequencing platforms, new methods such as ChIP-seq were developed in which the shortcomings of the ChIP-chip technology are addressed. In the ChIP-seq technology, immunoprecipitated DNA is sequenced using one of the next-generation sequencing platforms (see figure 1.5). The identified fragments (reads) can be mapped back to a reference genome (this is a publicly available genome of the model organism on which the experiments were conducted). Regions of the reference genome where a significant number of reads are mapped are assumed to be bound by the transcription factor. However, as a result of an effect that is known as sequencing bias, separating bound from unbound regions is a non-trivial task and it requires a lot of statistical post-processing. Nevertheless, if a control sample is available (an input sample that followed the same experimental procedure, without the immunoprecipitation step), the statistical analysis is relatively easy since we can get a notion of the sequencing bias. In contrast, if no control sample is available, we have to use a statistical model such as a Poisson distribution to get an estimate of potential sequencing biases.

A big advantage of ChIP-seq is the much improved resolution compared to

ChIP-chip. Due to the improved resolution, target sites can be resolved on less than 200bp with this technique. This higher resolution is a very important feature when the bound regions are subsequently analyzed for motif enrichment because current motif enrichment algorithms are highly sensitive to noise. Another important advantage is that ChIP-seq has a higher specificity and sensitivity compared to ChIP-chip [32].

1.6 Bioinformatics approaches

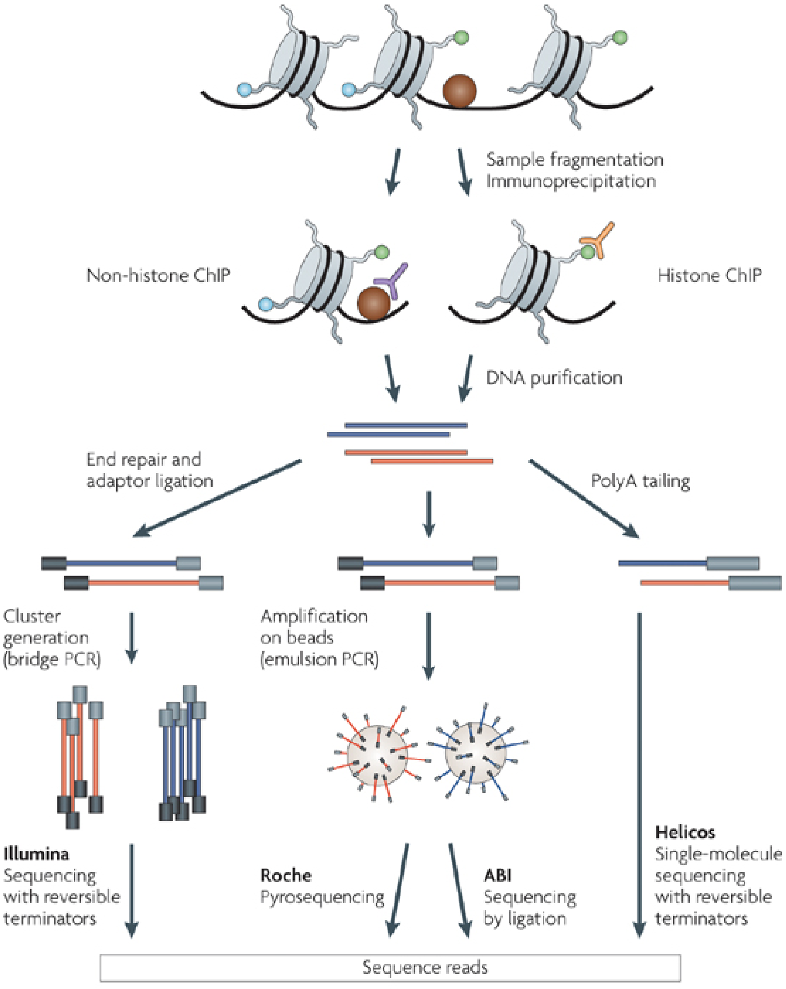
The bioinformatics algorithms that are used to predict binding sites are commonly separated into two classes. A first set of approaches enables researchers to look for novel motifs in sequences. These algorithms are called *de novo* approaches, because they detect novel binding sites and motifs from scratch. One of the most well-known *de novo* motif discovery tools is the MEME motif finder. A second set of approaches use existing information of already identified binding sites to build a model. This model can then be utilized to detect novel binding sites. In this thesis I will refer to this type of methods as model-based methods.

This section presents a short overview of some of the most frequently used algorithms. Note that I will not separate this overview into *de novo* and model-based methods, but I will rather discuss the methods in a chronological order. However, for clarity reasons I will mention for each method whether it is a *de novo* method or a model-based method.

1.6.1 Classical bioinformatics methods

Consensus sequences

A consensus sequence is a string of characters that depicts the nucleotide composition of an alignment of sequences. It is a model-based method as it relies on previously characterized binding sites. These consensus sequences are often used to represent the “average” TFBS sequence of a TF or a family of TFs. In most cases, not all nucleotides within a TFBS are conserved, so a modified sequence alphabet is necessary in order to account for degenerate nucleotides in the TFBS alignment. The IUPAC notation is an alphabet that is most frequently



Nature Reviews | Genetics

Figure 1.5: An overview of the ChIP-seq experimental procedure. Immunoprecipitated DNA is sequenced using one of the available next-generation sequencing technologies. *In silico*, the short sequence reads are mapped to a reference genome and a peak caller is used to identify bound regions. (image taken from a publication by Peter Park in Nature Genetics [33])

used in molecular biology. In table 1.1 the most used IUPAC codes are listed. One of the problems of consensus sequences is that no information is stored about the frequency of occurrence of the different nucleotides on each position. If, for example, in position one of the TFBS alignment, 99% of the time there is an A-nucleotide and the other 1% a T-nucleotide, this is indicated with a W code according to the IUPAC convention. It is immediately clear that information on the frequency of occurrence of each nucleotide is lost by the oversimplification introduced by this approach. However, one should appreciate the simplicity of the consensus approach. It is really straightforward to look for consensus sequences using the Perl regular expression engine.

IUPAC code	Nucleotides	Link
W	A and T	2 H-bonds
S	C and G	3 H-bonds
R	A and G	Purines
Y	C and T	Pyrimidines
K	G and T	
M	A and C	
B	C, G, and T	Not A
D	A, G, and T	Not C
H	A, C, and T	Not G
V	A, C, and G	Not T
N	All nucleotides	

Table 1.1: Nucleic acids IUPAC codes according to http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html

Positional weight matrices

The most common way to represent aligned binding sites and motifs nowadays is using positional weight matrices or PWMs. The positional weight matrix method is a model-based method that is constructed from a set of aligned experimentally validated TFBSs. PWMs are constructed by taking the following steps: first, on each position of the TFBS alignment, the occurrence of each nucleotide is counted. This results in a 4xL count matrix, with L being the length of the TFBS. Then, the count matrix is converted into a positional frequency matrix

(PFM) using formula 1.1 with p(b,i) being the probability of encountering base b on position i of the alignment; $c_{(b,i)}$ being the count of base b on position i.

Formula 1.1

$$p(b,i) = \frac{c_{(\mathrm{b,i})}}{N} \tag{1.1}$$

Finally, the PFM is converted into a PWM using formula 1.2. $W_{b,i}$ is the PWM value of base b in position i; p(b,i) and p(b) are respectively the probability to encounter base b on position i and the background probability of base b. Occasionally, a pseudocount function is used first to correct the PFM for small sample size.

Formula 1.2

$$W_{\mathrm{b,i}} = log_2 \frac{p(b,i)}{p(b)} \tag{1.2}$$

Currently, many online resources for PWMs are available, both publicly and commercially. The most well-known publicly available library is the JASPAR library [34], which contains over 500 PWMs of vertebrates, nematodes, insects, plants and yeast. A prominent commercially available collection of PWMs, the TRANSFAC library, is available from Biobase [35]. The TRANSFAC library has a larger collection of PWMs than the JASPAR library, but unfortunately many of the PWMs the library offers are redundant.

The PWM method, in contrast to the consensus method, does take into account positional frequencies. As Berg and von Hippel pointed out in their 1987 publication [36], the absolute PWM score is roughly correlated to the binding energy of the protein-DNA complex, but expressed in arbitrary units. Longer binding sites (longer matrix) will generally lead to higher absolute scores and higher binding energies, due to the additivity assumption (each extra base pair adds extra binding energy). However, the PWM approach does not take into account interactions between the base pairs, which is a serious oversimplification.

By far the most important problem the PWM method faces is the high level of false positive predictions. For some transcription factors with highly degenerate binding sites, such as sp1 or myoD, one binding site every few hundred base pairs is predicted. Genome-wide predictions with these PWMs would result in the prediction of millions of spurious binding sites. This number is much higher than the number of estimated *in vivo* binding sites, which is only a few thousands of binding sites. From these predicted binding sites, only a small fraction of 0.1-1% would have functional importance. This discrepancy between the predictions and *in vivo* functionality was termed the *futility theorem* by Wasserman and Sandelin in [1]. Most of the predicted binding sites are bound *in vitro*, but lack functionality in an *in vivo* context, possibly due to the inaccessibility of the binding site.

In order to reduce the number of false positive predictions, different sources of information should be incorporated into the prediction methods. One source of information that can be taken into account is the level of conservation of the binding regions in multiple species alignments. Methods that incorporate this type of data are called phylogenetic methods (for more information on these methods, see section 1.7.1).
Another way to reduce the number of false positive predictions is by adding a different type of data. For example, one may include information about the 3D structure of the DNA, or add information about the neighbouring regions.

Hidden Markov models

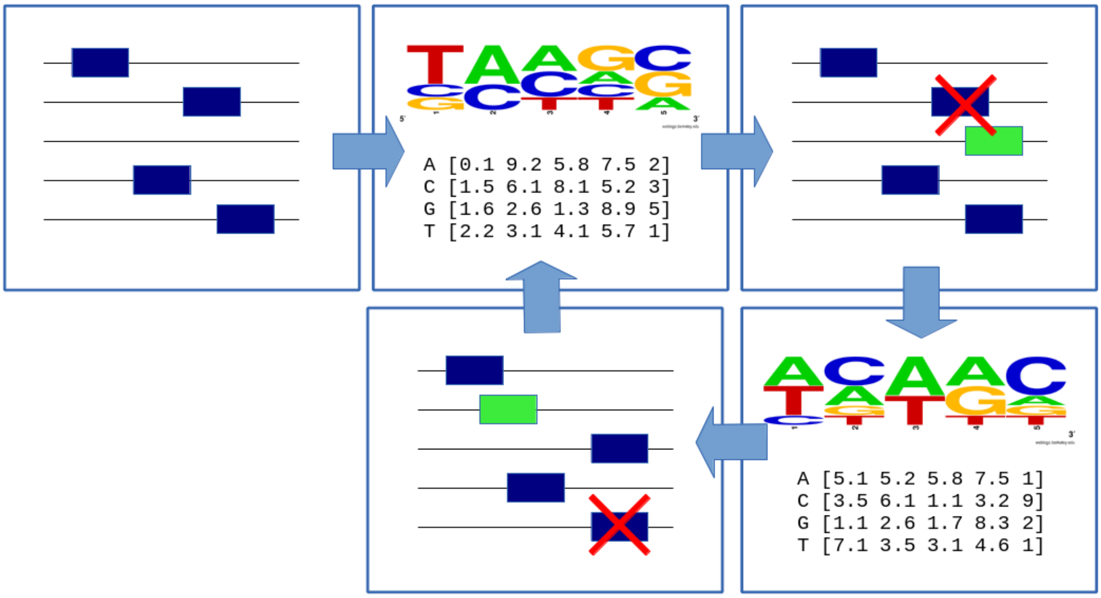
The Hidden Markov model (HMM) approach is a mathematical method that is very well-suited for biological sequence analysis. Similar to PWMs, HMMs can be used to search for certain sequence patterns. In the HMM method, binding sites are modeled with a technique termed Markov Chains. On each position of the TFBS, these Markov Chains give the probability of occurrence of each of the four nucleotides based on the nucleotide composition of preceding positions.

An important benefit of HMMs compared to PWMs is that HMMs implicitly take into account nucleotide dependencies. However, not all dependencies within the TFBS can be modeled with this approach. As Markov Chains only look at the preceding position (or multiple preceding positions in case of higher-order models), only the direct neighbours of each nucleotide are taken into account. Still, dependencies between non-neighbouring positions can be of great importance in TFBSs since protein-DNA complexes are essentially 3D complexes in which interactions in multiple dimensions are possible.

Gibbs sampler

Gibbs samplers were introduced in bioinformatics to look for sequence patterns and common motifs in a set of related sequences [37]. These shared motifs often indicate a shared biological function between the sequences that contain them. Gibbs samplers can be classified as *de novo* prediction methods that need no prior information about the input sequences.

In a Gibbs sampler, motifs are represented as a PWM. This PWM is used during the process of Gibbs sampling to search for certain sequence patterns. Most Gibbs samplers starts from a randomly chosen alignment of the sequences. In each sequence, a random starting point is chosen and the corresponding subsequences are used to build an initial (completely random) PWM. A random sequence ‘i’ is chosen and this sequence is scanned with the PWM constructed from the initial random alignment (without sequence i). In this way, each subsequence of sequence i can be scored. The position of the subsequence with the highest score is then used as the new starting position for this sequence in a new alignment. Then another random sequence ‘j’ is chosen and this sequence is scanned with the PWM that was build from the alignments of the other sequences (now also including sequence i). This process is repeated for a certain number of rounds, or it is repeated until the starting positions do not change anymore (see figure 1.6 for a graphical representation of the process). One of the problems with the aforementioned algorithm is that multiple runs on the same data set are necessary if TFBSs with varying length are considered. In order to address this issue, slightly different variations of the algorithm were proposed to alleviate this shortcoming.



```
A [0.1 9.2 5.8 7.5 2]
C [1.5 6.1 8.1 5.2 3]
G [1.6 2.6 1.3 8.9 5]
T [2.2 3.1 4.1 5.7 1]
```

```
A [5.1 5.2 5.8 7.5 1]
C [3.5 6.1 1.1 3.2 9]
G [1.1 2.6 1.7 8.3 2]
T [7.1 3.5 3.1 4.6 1]
```

Figure 1.6: Illustration of the Gibbs motif finder. Initially the starting positions of the binding sites on each sequence are randomly chosen. A PWM is build with all of these randomly chosen binding sites, except for one sequence. This PWM is then used to look for the highest scoring site in the hold-out sequence. Another sequence is left out and a novel PWM is build using the new alignment. Now this PWM is used to scan the hold-out sequence. This process continues until the starting positions of the binding sites do not change anymore (or very little).

Gibbs sampling is a heuristic approach, which means that the solution found is not always the most optimal one. Because of the random initial alignment, results of a sample run tend to differ each time. This problem can be solved by rerunning the sampling method a number of times, and then choosing the alignment or motif with the highest information content. However, as newer and better algorithms started to emerge, the use of Gibbs sampling algorithms has seen a decrease in popularity, mainly because of the unreliable results produced by the algorithm.

Multiple Expectation Maximization for Motif Elicitation

Multiple Expectation Maximization for Motif Elicitation or MEME is by far the most widely used *de novo* motif finder, originally developed by Timothy Bailey [38].

MEME is more robust against noise than Gibbs samplers, and it is able to cope with sequences that do not contain a binding site. At the heart of the MEME method lies an expectation maximization (EM) algorithm that is used to discover overrepresented motifs in the data set. The EM algorithm consists of an expectation step, in which the start positions of the most enriched motif are estimated, followed by a maximization step, in which the expected likelihood of the data set is maximized given the start positions of the motifs. Both steps are repeated for a certain number of iterations or until the parameters do not change anymore (or the changes are below a certain threshold). Another clever improvement of the MEME algorithm compared to a Gibbs sampler is that MEME can keep looking for other interesting motifs, even if a putative motif was already found. This is possible because MEME *erases* previously discovered motifs from the probabilistic space thus preventing that the same motif would be found again. In this way MEME can continue looking for other interesting lower ranked motifs.

1.7 Improving the accuracy of the *in silico* methods

All of the previously discussed methods are very prone to false positive predictions, because most binding sites and motifs are rather degenerated and of low information content. In an attempt to reduce these high levels of false positive predictions, multiple information sources were added to aid the accuracy of the predictions.

In this section I will discuss a few of these information sources and methods that can be utilized to reduce the number of false positive predictions.

1.7.1 Phylogenetic methods

A very popular way to reduce the false positive rate is by adding information about multiple species alignments. Methods that incorporate this type of information are called phylogenetic methods because they make use of phylogenetic data to filter out possible spurious results. The idea behind the phylogenetic approach is that functional genomic sequences have a higher degree of conservation in multiple species alignments because of a functional constraint: if a functional sequence is mutated this can result in a functional defect and a less fit organism. Although slight variations between the different phylogenetic methods exist, most methods follow roughly the same strategy. In most of these phylogenetic methods, a PWM collection is used to scan ultra-conserved regions (regions that are conserved in more than 90% of the organisms). The results from such an analysis most often are represented in conservation plots and in percent identity plots. Only those PWM hits that are conserved above a certain threshold are considered as bona fide hits. A few examples of commonly used phylogenetic methods are rVista, ConSite and ConTrav2.

rVista

rVista is a computational tool developed by Gabriela Loots to help the annotation of vertebrate genomes [39]. The method combines PWM predictions from the TRANSFAC library with alignments from the AVID [40] and the VISTA algorithm [41]. This approach claims to reduce the number of PWM false positive predictions by a fraction of 95% while still maintaining a large fraction of the experimentally validated sites.

ConSite

Shortly after the initial publication of JASPAR in 2004 [34], Sandelin, Wasserman and Lenhard alluded to the shortcomings of the PWM-based methods. In an attempt to address the issues raised, they developed the ConSite tool [42], which makes use of cross-species comparisons to reduce the number of false positive predictions. The tool uses ORCA alignments [43] and their publicly available and manually curated JASPAR library. In ConSite conservation is calculated by taking a sliding window approach in which the percentage conserved nucleotides are counted (see figure 1.7). The conservation score resulting from this calculation is visually represented as a percentage identity plot, and only windows with a score exceeding a certain threshold are further analyzed for binding sites with the JASPAR PWM library.

ConTrav2

See section 1.10 in "Scope of this thesis."

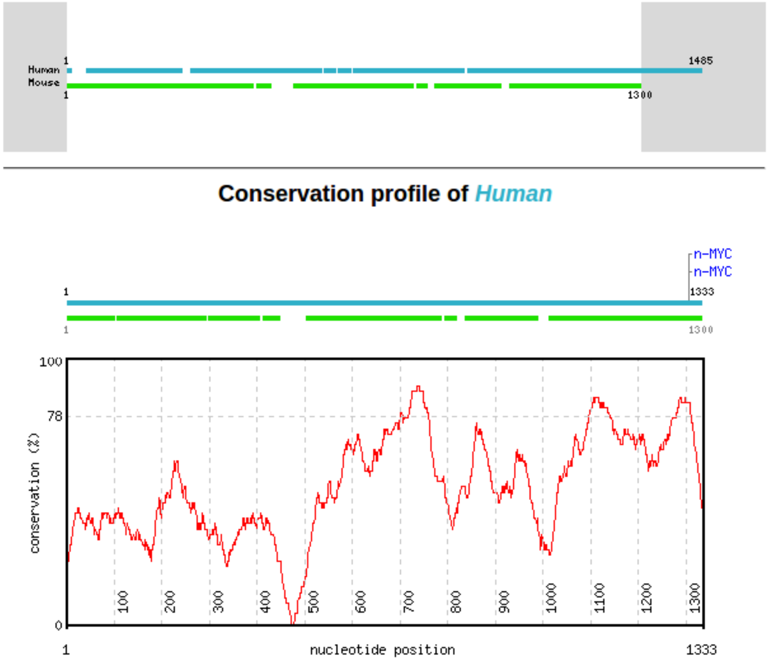

Conservation profile of *Human*

Figure 1.7: An example of ConSite output. In this example, a pairwise comparison between a human and mouse promoter is made. Both are scanned for N-MYC binding sites. In the top part a graphical representation is displayed that indicates the aligning regions of both sequences. In the lower part a conservation plot is shown with putative N-MYC hits indicated in blue.

Limitations of phylogenetic approaches

Although phylogenetic methods effectively reduce the number of false positive predictions, there are also some drawbacks to these approaches. For example, these methods do not provide us with any insight into the mechanics of transcription factor DNA specificity. Since cells do not have access to any evolutionary knowledge of regulatory regions, other mechanisms must be responsible for the specificity of protein-DNA complexes. One problem that might be responsible for the bad performance of *in silico* methods is that they often focus on a single binding site at a time. The *in vivo* situation is much more complex though, as multiple transcription factors can go into competition for the same binding site, or overlapping binding sites. However, this level of competition is also not captured by the phylogenetic methods. Furthermore, phylogenetic methods give no insight into the biophysical interactions between the DNA and TF and the mechanics that fine-tune the binding specificity. These biophysical interactions, and their use in the prediction of binding sites are the subject of next subsection.

1.7.2 Other approaches

The previously discussed methods rely on approaches that only take into account the primary sequence information about the binding site itself. However, other types of information can be incorporated to improve the accuracy of the *in silico* methods. This section presents an overview of two additional methods that can be used to enhance the predictions of binding sites. Both methods add an additional source of data to refine predictions. The first method involves the introduction of data on the surrounding environment of the binding site. The second method concerns the use of information about biophysical properties related to the flexibility, curvature and excluded surface of the DNA.

Incorporating information about the neighbourhood

Most algorithms only take into account the nucleotides from the binding site itself. This is a very conservative approach since numerous research papers have proven that there is a lot of information content in the neighbourhood of a binding site. This neighbourhood is a possible source of information about common cofactors, features that help the transcription factor to bind to the DNA, or other nucleotide signatures. Some prediction methods incorporate these broader regions in the prediction algorithm. One excellent example is the Dispare algorithm [44] which aims to extend PWMs beyond the boundaries of the binding site. In our PhysBinder approach (see 1.10, "Scope of this thesis" and chapter 3 for a thorough description of the algorithm) we also decided to include information about the flanks of the binding site. We take into account 50 nucleotides before and after the start position of the binding site. However, in some models the flanks are slightly shorter due to our feature selection algorithms.

Incorporating information about biophysical properties

Parker and his research team [45] discovered that DNA of TFBSs is more conserved at a 3D structural and biophysical level than at the primary sequence. This means that there is a stronger evolutionary constraint on the structure and physics of the chromatin than on the individual nucleotides. As a consequence, prediction methods that incorporate this structural or biophysical component are expected to make better predictions than methods that only look at the sequence level.

The structure and biophysical characteristics of the chromatin are mainly a product of the primary DNA sequence. This implies that given a certain DNA sequence one can infer most of the structure and biophysical characteristics of the corresponding chromatin bundle. However, the opposite is not necessarily possible because in many cases multiple distinct DNA sequences can lead to the same structural profile. Both the structure of the DNA and the primary sequence are thought to contain a different level of information. In addition to the primary DNA sequence, the structural level of the DNA can have a large impact on many biological mechanisms such as protein-DNA binding and transcription initiation due to the influence of the chromatin structure on protein-DNA complex formation.

Since most of the information about DNA structure is contained in the primary DNA sequence, we can use the DNA sequence to get an idea about DNA structure and use this information to get a better understanding about protein-DNA complex formation. To this end, we can use molecular mechanical simulations to unravel the biophysics and structure of the chromatin from the primary sequence with a very high level of precision. Unfortunately, these simulations are very resource intensive and they can take a long time to compute. Another approach that can be used to model the structure of the DNA is using “k-nucleotide” property lists. These lists contain pre-calculated structural values for each stretch of k-nucleotides. For example, dinucleotide lists will contain values for each pair of nucleotides. As most of the structural information can be described by these dinucleotides, dinucleotide scales are used most prevalently. Nevertheless, some structural characteristics (for example curvature and torsion) cannot be described by these dinucleotide tabels. In this case, higher-order scales are needed. Many of these scales are available from online resources such as DiproDB [46] and for this doctoral thesis we also calculated some additional feature scales. Examples of these scales are to be found in table 1.2.

Feature class	Feature name	Order of scale
DNA conformation	homogeneity A/B	dinucleotide
	uniformity A/B	dinucleotide
groove properties	bend towards groove	dinucleotide
	groove clash distance	dinucleotide
	groove clash size	dinucleotide
	groove width	dinucleotide
	hydration groove	dinucleotide
helix structure	curvature	trinucleotide
	torsion	tetranucleotide
	propeller twist	dinucleotide
AA pairing preferences	AA propensity	dinucleotide
solvent accessibility	solvent excluded surface	tetranucleotide
	solvent accessible surface	tetranucleotide
thermodynamic properties	enthalpy change	dinucleotide
	entropy change	dinucleotide
	free energy change	dinucleotide
	minimum energy	dinucleotide

Table 1.2: Some examples of biophysical and structural features. The features *curvature*, *torsion*, *solvent excluded surface* and *solvent accessible surface* are based on own calculations. The other features are available for download from DiProDB [46].

A large part of this thesis describes the use of structural and biophysical data in the prediction of transcription factor binding sites, hence I will discuss the use of the structural and biophysical state of chromatin in detail. Structural properties of a binding site are termed “indirect” components of the binding site and they are often very important properties of the binding site [47]. These indirect components, together with the direct contacts that are formed between the protein and the DNA, are largely responsible for the specific recognition of binding sites by transcription factors.

How can we incorporate these structural and biophysical features to enhance the binding site predictions? There are multiple approaches that can be used to achieve this goal. Similar to consensus sequences (discussed in section 1.6.1), one could use consensus vectors of structural properties to aid the prediction of binding sites. However, from our initial experiments we concluded that this approach does not necessarily lead to any improvements. A more suitable approach is the use of classification algorithms such as SVMs, naive Bayes, neural networks or Random Forests to incorporate these structural vectors. In this thesis I show that the use of these structural and biophysical features together with a Random Forest classification algorithm can lead to improvements to the classification accuracy of transcription factor binding sites. We made a comparison with thoroughly optimized PWM models (that use the Dispare algorithm for length optimalization) and compared our method with the latest alternative structure based approach (CRoSSeD [48]). From this comparison we conclude that our Random Forest based approach performs better overall (see chapter 3).

For the construction of these structural classification models, a lot of experimental data is needed. Lucky for us, recent developments in the ENCODE project have provided us with many of experimentally validated *in vivo* binding sites. These recent developments are discussed in the next section.

1.8 The ENCODE project: a massive amount of data

After the completion of the Human Genome Project in 2003, the Encyclopedia of DNA Elements (ENCODE) Consortium took over where the Human Genome Project (HGP) has ended [12] (see figure 1.8). The goal of the ENCODE project was to identify all major functional elements in the human genome. In a pilot phase that was used to optimize the experimental techniques, 1% of the total human genome was already studied. In this pilot phase some controversial discoveries were made that changed the way researchers regard functional elements in the genome. For example, a lot of regulatory sequences were found in the so called “junk DNA” in between genes and these elements were also found in the intronic regions.

In September 2012 the ENCODE project was finished with the aid of the techniques that were optimized in the pilot phase. During this project an enormous amount of previously unknown functional elements were discovered. Amongst these functional elements are some with a very narrow functionality while others play a role in approximately every pathway. These functional elements can be tissue specific (dependent on certain stimuli) or they can be ubiquitously active in all cells. The project concluded that around 80% of the human genome is functional in some way but this conclusion has become somewhat controversial since its initial publication [14, 49, 13]. Despite the controversy, the real beauty of the ENCODE project is that it provides researchers with genome-wide experimental data sets for a large number of transcription factors and epigenetic modifications. Some examples are: DNA methylation patterns of 82 cell lines; ChIP-seq experiments for 119 TFs in different cell lines; 12 histone ChIP-seq experiments for different cell lines; DNase-seq and DNase footprinting on a multitude of cell lines. In total 1640 different data sets were generated, which resulted in the publication of more than 30 papers in top journals such as Nature, Genome Biology and Genome Research. This collection of data sets will vastly enhance our knowledge of the human genome and it will certainly speed up the *in silico* research. In addition, these data sets are a very powerful tool to study the mechanisms that determine the specificity of protein-DNA complexes.

1.9 Beyond the ENCODE project

The ENCODE project has mainly focused on the characterization of functional genomic regions such as genes, regulatory elements and epigenetic modifications. Unfortunately, for many of these regions the precise molecular biological function remains unknown. For example, the regulatory effects and functions of many miRNAs and lncRNAs are still unclear. There is a lot of potential left in the research of the function and regulation of these RNA molecules. When knowledge about the functions of these non-coding RNAs is combined with data on transcription factor binding sites and epigenetic regulation this most definitely will give us a much broader insight into how pathways are regulated inside our cells.

A new initiative called modENCODE has been launched to ensure the continua-

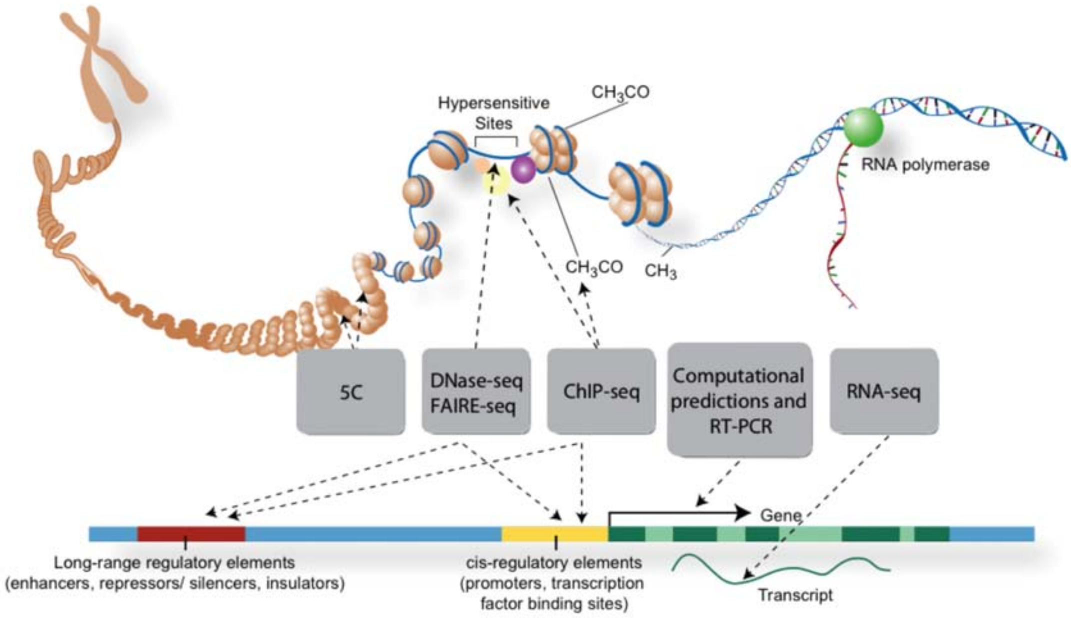

Hypersensitive Sites
CH_3CO
RNA polymerase
CH_3CO
CH_3
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

Figure 1.8: An overview of the experimental approaches used during the ENCODE project. The structural organization of the genome was assayed with the 5C approach. Hypersensitive sites were discovered using DNase-seq and FAIRE-seq. ChIP-seq experiments were performed to search for transcription factor binding sites. RNA-seq, finally, gives a broad idea about the transcriptional landscape of the genome. [Credits for the image: Darryl Leja (NHGRI), Ian Dunham (EBI)].

tion of the ENCODE project [50]. During the new project, functional elements of many of the model organisms in biotechnology will be studied. In this way, the biological validation of certain hypothesis that are impossible to test in humans (for ethical reasons) can be validated and the evolution of functional elements can be studied.

1.10 Scope of this thesis

In this PhD thesis I will describe a number of tools and algorithms that were developed both to enable and to improve the *in silico* identification of transcription factor bindings sites. This section presents a short overview of these different tools.

ConTrav2

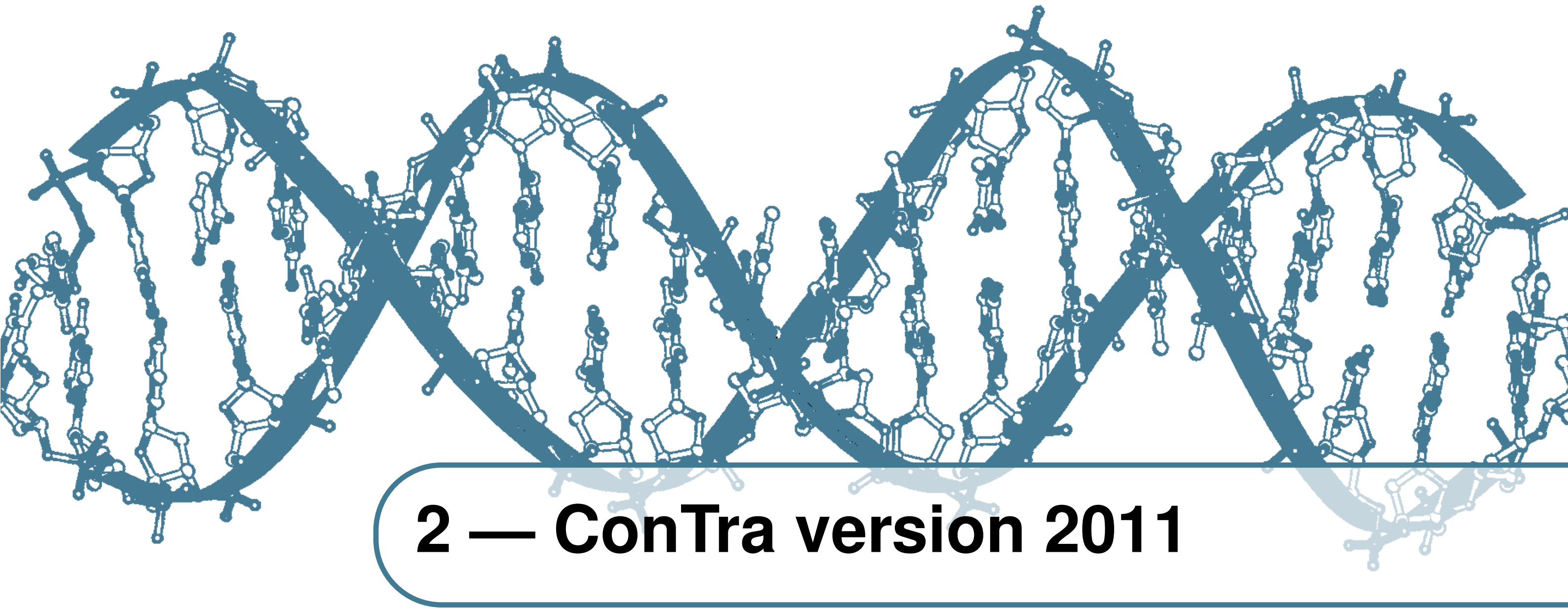
During this PhD we developed an update to the previously published ConTra web-tool. This web-tool provides the wet-lab biologist with a user-friendly tool to interactively explore and visualize TFBSs, predicted by position weight matrices (PWMs), in promoter alignments. The update was a result from the many requests from the wet-lab and our users and it offers several major improvements. The user interface is completely redesigned and is now far more intuitive to use. Furthermore, users are no longer restricted to the human genome when looking for transcripts. Multiz alignments of mouse, cow, chicken, frog, fish, insect, worm and yeast are now available and this has led to a further increase of our user community. In addition to promoter sequences, one can look for TFBSs in intron, 5'UTR and 3'UTR sequences in the updated version of ConTra. An additional library of protein-binding microarray (PBM) matrices increases the number of available PWM libraries. In this version users can now also upload their own PWM library of choice. In order to do this, they just have to upload a multi-FASTA file with the different binding sites; the web-tool will do the rest. Furthermore, the updated version no longer excludes non-coding genes from the analysis. For more information on the redesigned version of ConTra, and all the details, see chapter 2.

Random Forest biophysical algorithm

The majority of the work during this PhD was invested in the development of an *in silico* biophysical prediction method and in the construction of biophysical models for the prediction of transcription factor binding sites. Initially, we started by calculating structural and biophysical features from scratch. Starting from the primary sequence, we calculated a 3D model of the chromatin with the help of DNA bending models. We then used this 3D model to calculate certain features, such as the curvature, the torsion, solvent excluded and solvent accessible area of the DNA. As we started exploring different options to speed up the calculations, we decided to pre-calculate all possible structural values into a lookup table. Later, other features from literature such as the flexibility of the DNA and minor/major groove size were added and we also included the primary sequence into our models to capture direct readout. We experimented with a large number of classifiers and measured the performance of each of these classifiers. Based on the results we decided to use the Random Forest classifier as the main classifier in our algorithm. Since Random Forest classifiers do not cope well with a large number of features, several feature selection algorithms were implemented to reduce the number of features. This approach was implemented in a proof-of-principle and was published in [51]. In the publication describing the algorithm we show that we outperform current state-of-the-art approaches for *in silico* identification of transcription factor binding sites. We also discuss some problems with prokaryotic data sets and show that many of the shortcomings with these data sets are due to quality issues. For more information on the details of the algorithm, the implementation, the analysis and the proof of principle see chapter 3.

PhysBinder

Based on our Random Forest implementation we designed an intuitive and easy to use web-tool. The Random Forest algorithm was slightly adapted and a large number of transcription factor models were build using the publicly available data sets from ENCODE. For each of the models, we also offer a quality profile with ROC curves and an overview of all the features that are contained in the models. Users can submit FASTA sequences for analysis in the PhysBinder integrative algorithm or indicate a genomic region of interest. We also offer the option to overlay results with experimentally validated ENCODE binding sites and the option to visualize all results in the UCSC Genome Browser. In this way, we try to be a true companion of the UCSC Genome Browser and help wet-lab biologists steer their experiments. All details about this web-tool can be found in chapter 4.



2 — ConTra version 2011

This chapter is a redraft from the publication:

Stefan Broos, Paco Hulpiau, Jeroen Galle, Bart Hooghe, Frans Van Roy, and Pieter De Bleser **ConTra v2: a tool to identify transcription factor binding sites across species, update 2011**; Nucleic Acids Res. 2011 39: W74-W78

2.1 Abstract

Transcription factors are important gene regulators with distinctive roles in development, cell signaling and cell cycle and have been associated with many diseases. The ConTra v2 web server allows easy visualization and exploration of transcription factor binding sites (TFBS) in any genomic region surrounding coding and non-coding genes. In this new version users can choose from nine reference organisms ranging from human to yeast. ConTra v2 can analyze promoters, 5'UTRs, 3'UTRs and introns or any other genomic region of interest. Several hundreds of position weight matrices (PWM) are available to choose from or alternatively an own PWM for detecting specific binding sites can be uploaded. A typical analysis is run in four simple steps of choosing gene, transcript, region

of interest and selecting one or more TFBS. The ConTra v2 web server is free and open for everyone and available at http://bioit.dmbr.ugent.be/contrav2/.

2.2 Introduction to ConTra v2

Transcription factors (TFs) and microRNAs (miRNAs) are the key players of gene regulation in multicellular organisms [52]. Based on pairing between miRNAs and mRNAs, miRNA targets are predicted by searching for matches with the miRNA seed regions [53]. For detection of transcription factor binding sites (TFBS) the use of a position weight matrix (PWM) is the leading model. Such a PWM represents the sequence motif where a transcription factor can bind and is constructed using a set of known binding sequences. Traditionally regulation of genes by TFs is predicted by promoter analyses and experimentally determined by DNAse footprinting assays or Electrophoretic mobility shift assays (EMSA). Nowadays functional protein-DNA binding sites are more and more studied on a genome-scale basis using ChIP-seq. These studies indicate that only part of the functional TFBS are located in promoter regions next to intragenic regions and untranslated regions (UTR) which also contain a significant number of functional sites [54–56]. Regulatory sites in the first intron e.g. may interact with sites in the promoter region through DNA looping [57, 58]. Of the 2000 estimated human TFs about 300 are thought to bind to the core promoter with a role in the general transcription machinery while other TFs have a higher target specificity and regulate a smaller number of genes [59]. The latter can be expressed in almost all or only in a few tissues either having a broad or a specific function respectively. Over half of the human genes are believed to have alternative promoters [60] and consequently one should investigate promoters, UTRs and intronic regions of each transcript. In this update we describe the new features and expansions of the ConTra webserver. Transcription factor bindings sites can be detected and visualized in any genomic region of the known transcripts of a gene of interest. Starting from one of nine reference organisms, a scientist can easily investigate regulation on transcript level using the latest UCSC Multiz alignments which are automatically available through the ConTra interface. Alternatively sequence files and position weight matrices can be uploaded for analysis on own data.

% of TFBS in	TCF4	ZNF263	CTCF	NRSF	STAT1	cMyc	RANGE
introns	35-40%	30.40%	29%	24%	25%	NA	20-40%
5’-UTR	4%	NA	7%	12%	11%	NA	1-15%
3’-UTR	NA	NA	2%	1%	1%	NA	1-2%
promoter	10-20%	20%	13%	15%	15%	0-27%	10-20%

Table 2.1: Table with genome-wide distribution of binding sites. Only the regions relevant for ConTra v2 are included in the table. Results of the TCF4 binding sites are taken from Mokry et al.; ZNF263 are from Frietze et al.; CTCF, NRSF and STAT1 from Jothi et al.; cMyc from Cawley et al.

2.3 New features

The first edition of ConTra provided users with a flexible way to analyze promoter alignments [61]. Users were able to visualize or explore transcription factor binding sites (TFBSs) in the promoter region of a gene of interest. PWM libraries from the JASPAR CORE database and TRANSFAC database were used to identify TFBSs in a multispecies alignment with human as reference species. Even though the human genome is one of the most widely used reference genomes, the lack of other reference species and alignments was regarded as one of the most eminent shortcomings in the first edition of ConTra. Furthermore, only the promoter region could be analyzed for TFBSs. The 2011 update of ConTra adds several extra features. In addition to the promoter region, users can now look for TFBSs in 5’UTR, 3’UTR and introns. Many researchers suggest these regions are at least as important in transcriptional regulation as the promoter region itself [62, 55, 56, 54]. Mokry et al. [55] demonstrate that a large fraction (35-40%) of all TCF4 binding sites are intronic. Furthermore, considerable fractions of ZNF-263, CTCF, NRSF and STAT1 binding sites are located in 5’-UTR, 3’-UTR and intronic regions. A detailed overview on the relative importance of the aforementioned genomic regions is to be found in table 2.1.

In the first edition of ConTra, searching for TFBSs was only possible in multiple alignments in relation to the human genome, thus leaving many users empty handed. In ConTra v2 multiple alignments for mouse, chicken, cow, frog, zebrafish, fruitfly, roundworm and yeast were added. A detailed overview of the different genome assemblies, genes and Multiz alignments available in ConTra v2 can be found in table 2.2. Although the human genome is the most widely

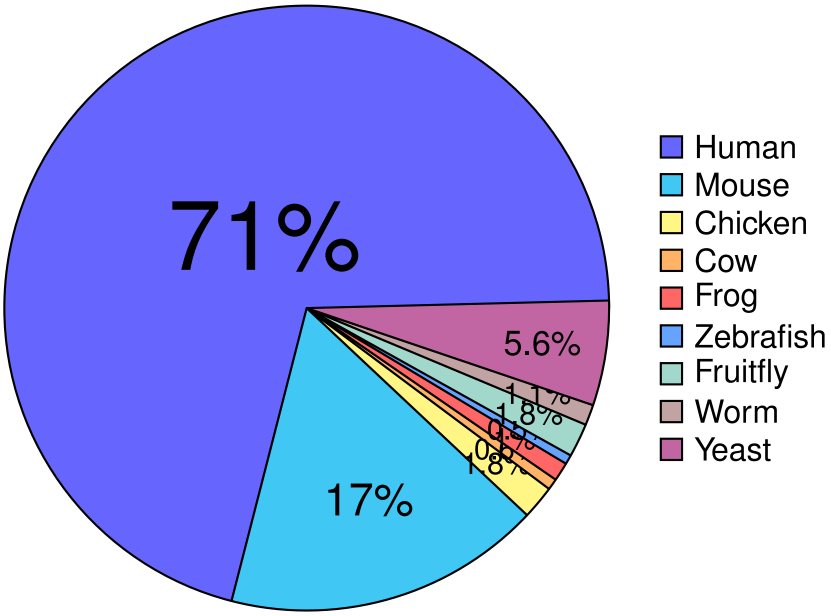


Figure 2.1: Pie chart representing the fractions of Pubmed hits of the different reference species on total. Percentages and search terms are taken from table 2.3.

studied genome, other model organisms should not be ignored. The importance of the different model organisms is illustrated by table 2.3 and figure 2.1, in which the popularity of the different organisms is compared in terms of PubMed hits.

In ConTra v2, searching for transcripts is possible using the HGNC gene name, symbol, alias, Ensembl gene ID (ENSG), the Entrez Gene ID, the RefSeq mRNA ID or the Ensembl transcript ID (ENST). For every species the most recent alignments are then automatically fetched from UCSC and further processed. Users can select binding motifs from different sources including the latest versions of the TRANSFAC database (update 2010.4) [63], the JASPAR core database update 2010 [64], the phyloFACTS database [65] and a protein-binding microarray (PBM) derived collection of homeodomain TF PWMs [66]. Furthermore, own PWMs can be constructed using the web interface. Creating a custom PWM is as easy as uploading a FASTA file containing aligned sequences. The ConTra v2 web interface automatically converts the data into the right format. In ConTra v2, non-coding genes are no longer excluded from the analysis. Often TFs and miRNAs work together in what is termed feedback loops (FBL) or feed-forward

Ref. species	Common name	Assembly	Genes	RefSeq	% coding	% non-coding	Ensembl transcripts	Alignment
H. sapiens	human	hg19	22167	37474	86.3	13.7	151222	multiz46way
M. musculus	mouse	mm9	21786	27621	93.3	6.7	88186	multiz30way
B. taurus	cow	bosTau4	11559	12427	97.7	2.3	31598	multiz5way
G. gallus	chicken	galGal3	4905	5176	90.1	9.9	23392	multiz7way
X. tropicalis	frog	xenTro2	8358	9695	99.8	0.2	28937	multiz7way
D. rerio	zebrafish	danRer6	13812	15776	95.6	4.4	32992	multiz6way
D. melanogaster	fruit fly	dm3	14230	23550	94.1	5.9	23017	multiz15way
C. elegans	worm	ce6	19903	24892	97.1	2.9	35019	multiz6way
S. cerevisiae	yeast	sacCer2	7130	na	na	na	7130	multiz7way

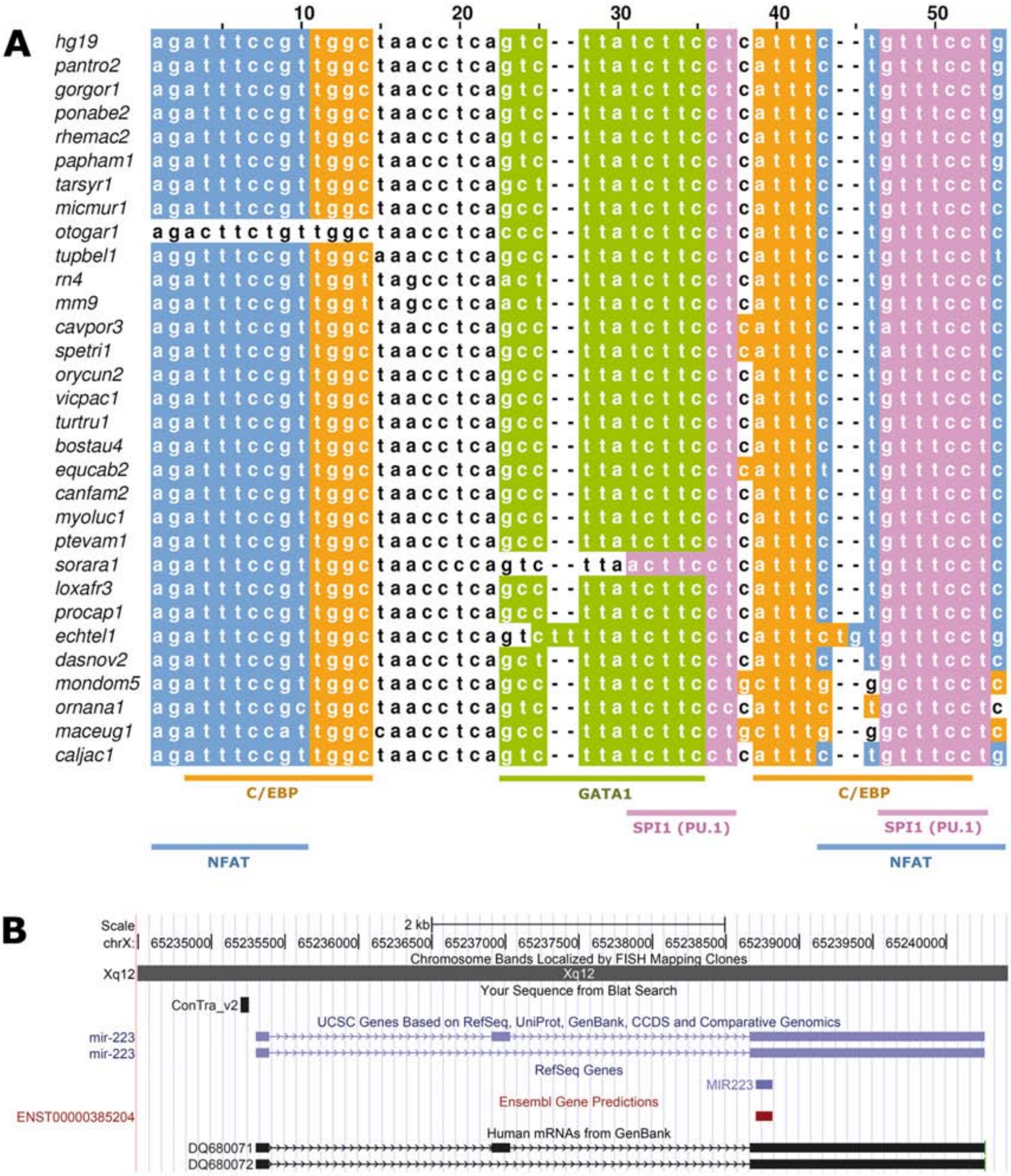
Table 2.2: Summary of the number of genes, non-coding genes and transcripts for each reference organism that can be analyzed in ConTra v2.

Popular name	Official name	Pubmed hits	In % total hits
Human (1)	Homo sapiens	1984438	71.06
Mouse (2)	Mus musculus	465858	16.68
Chicken (3)	Gallus gallus	51076	1.83
Cow (4)	Bos taurus	16479	0.59
Frog (5)	Xenopus tropicalis	26033	0.93
Zebrafish (6)	Danio rerio	14310	0.51
(Fruit)fly (7)	Drosophila melanogaster	49233	1.76
Nematoda (8)	Caenorhabditis elegans	29550	1.06
Yeast (9)	Saccharomyces cerevisiae	155458	5.57
	Total	2792435	100

Table 2.3: Table indicating the number of PubMed hits and relative percentages (search executed on 26/01/2011). It should be noted that the percentages only indicate the fraction on total PubMed hits on the nine reference species in ConTra v2. Search terms that we used: 1. “Homo sapiens” OR “humane”; 2. “‘Mus musculus” OR “mouse”; 3. “Gallus gallus” OR “chicken”; 4. “Bos taurus” OR “cow”; 5. “Xenopus tropicales” OR “frog”; 6. “Danio rerio” OR “zebrafish”; 7. “Drosophila melanogaster” OR “fly”; 8. “Caenorhabditis elegans” OR “worm”; 9. “Saccharomyces cerevisiae” OR “yeast”.

loops (FFL). In this type of regulatory network, TFs regulate the transcription of miRNAs, while the miRNAs control transcription of this TF. These loops regulate many important biological processes, like development and tumor formation [67]. Non-coding transcripts are treated as regular transcripts in ConTra, and can be analyzed in the same way. To verify whether the results on non-coding genes are meaningful, we looked for binding sites in the promoter region of microRNA223 with RefSeq accession number NR029637. Fukao et al. have shown that MIR233 is regulated by a wide range of transcription factors such as NFAT, C/EBP, GATA1 and PU.1 [68]. Analysis in ConTra v2 not only supports the presence of the binding sites for these TFs but also that these are evolutionary highly conserved (figure 2.2).

A wide variety of examples on the use of ConTra v2 can be found in the appendix of this thesis. Figures A.1-A.5 show results of ConTra v2 analyses on different genomic regions, using the UCSC multiz46way alignment based on the human hg19 reference sequence and illustrating curated binding sites from literature. Figure A.6 depicts an evolutionary conserved binding site in the second intron of the M. musculus Nestin gene, as described by Jin et al. [69]. In figure A.7



```
hg19      agatttccgttggctaacctcagtc--ttatcttcctcatttc--tgtttcctg
pantro2   agatttccgttggctaacctcagtc--ttatcttcctcatttc--tgtttcctg
gorgor1   agatttccgttggctaacctcagtc--ttatcttcctcatttc--tgtttcctg
ponabe2   agatttccgttggctaacctcagtc--ttatcttcctcatttc--tgtttcctg
rhemac2   agatttccgttggctaacctcagtc--ttatcttcctcatttc--tgtttcctg
papham1   agatttccgttggctaacctcagtc--ttatcttcctcatttc--tgtttcctg
tarsyr1   agatttccgttggctaacctcagct--ttatcttcctcatttc--tgtttcctc
micmur1   agatttccgttggctaacctcagcc--ttatcttcctcatttc--tgtttcctc
otogar1   agacttctgttggctaacctcaccc--ttatcttcctcatttc--tgtttcctc
tupbel1   aggtttccgttggcaaacctcagcc--ttatcttcctcatttc--tgtttcctt
rn4       agatttccgttggttagcctcaact--ttatcttcctcatttc--tgtttcccc
mm9       agatttccgttggttagcctcaact--ttatcttcctcatttc--tgtttcctc
cavpor3   agatttccgttggctaacctcagcc--ttatcttcctcatttc--tatttcctc
spetri1   agatttccgttggctaacctcagcc--ttatcttcctcatttc--tatttcctc
orycun2   agatttccgttggctaacctcagcc--ttatcttcctcatttc--tgtttcctc
vicpac1   agatttccgttggctaacctcagcc--ttatcttcctcatttc--tgtttcctc
turtru1   agatttccgttggctaacctcagcc--ttatcttcctcatttc--tgtttcctc
bostau4   agatttccgttggctaacctcagcc--ttatcttcctcatttc--tgtttcctc
equcab2   agatttccgttggctaacctcagcc--ttatcttcctcatttt--tgtttcctc
canfam2   agatttccgttggctaacctcagcc--ttatcttcctcatttc--tgtttcctc
myoluc1   agatttccgttggctaacctcagcc--ttatcttcctcatttc--tgtttcctc
ptevam1   agatttccgttggctaacctcagcc--ttatcttcctcatttc--tgtttcctc
sorara1   agatttccgttggctaaccccagtc--ttacttcctcatttc--tgtttcctc
loxafr3   agatttccgttggctaacctcagcc--ttatcttcctcatttc--tgtttcctc
procap1   agatttccgttggctaacctcagcc--ttatcttcctcatttc--tgtttcctc
echtel1   agatttccgttggctaacctcagtctttatcttcctcatttctgtgtttcctg
dasnov2   agatttccgttggctaacctcagct--ttatcttcctcatttc--tgtttcctc
mondom5   agatttccgttggctaacctcagcc--ttatcttcctgctttg--ggctttcctc
ornana1   agatttccgctggctaacctcagtc--ttatcttccccatttc--tgctttcctc
maceug1   agatttccattggccaacctcagcc--ttatcttcctgctttg--ggctttcctc
caljac1   agatttccgttggctaacctcagtc--ttatcttcctcatttc--tgtttcctg
```

Figure 2.2: Visualization of the evolutionarily conserved mechanism for microRNA-223 regulation in the promoter region, as described by Fukao et al. [68]. (A) Multiz alignment showing the conserved binding sites. In orange, the C/EBP transcription factor, predicted using the JASPAR positional weight matrix MA0102.2; in blue, the NFAT transcription factor (TRANSFAC M00935); in green; the GATA1 transcription factor (JASPAR MA0035.2); and in pink, the PU.1 transcription factor (JASPAR MA0080.2). The figure was created with the free multiple alignment editor Jalview using the ConTra FASTA and feature color (.fc) file on the results page. (B) Region of (A) was mapped using BLAT on the promoter region in the UCSC Genome Browser (black box). Blue box represents the microRNA location.

two Sine oculis (SO) binding sites are conserved in the second intron of the Drosophila Lz gene confirming the study of Yan et al [70]. Finally, the promoter of the S. cerevisiae PHD1 (FLO11) gene in figure A.8 shows two conserved TEA transcription factor binding sites supporting the regulatory mechanism reported by Heise et al [71]. Alignment files in the UCSC multiple alignment format (MAF), in multi-FASTA or clustal format can be uploaded if the genomic region e.g. from another reference organism or for a new transcript is not available in ConTra. On the help page of the website demos are available how to get such a MAF file in the UCSC Genome Browser, how to upload and analyze this file and how to use the feature color (fc) file and FASTA file on the result page to produce publication-quality figures similar to the ones in the appendix of this PhD thesis. If a PWM model for a particular transcription factor is missing in the available collections, uploading an own PWM is also possible. This can be either in the PWM format or alternatively less experienced users can simply upload an alignment file in multi-FASTA format which ConTra can automatically detect and build the position weight matrix.

2.4 Similar tools

- **MONKEY web server**: Identification of conserved TFBSs in multiple alignments using binding site-specific evolutionary model. This is not a web-tool and users have to use their own alignments.
- **NCBI Dcode.org**: Collection of dynamically interconnected tools. However, easy to get lost in.
- **FOOTER 2.0**: Web-tool for finding mammalian DNA regulatory regions using phylogenetic footprinting. This tool has a limited collection of PWMs and users have to upload their own sequences.
- **MAPPER**: HMM-based identification of TFBSs in multiple genomes.
- **PromoterPlot**: A graphical display of promoter similarities by pattern recognition. Depends on deprecated technology and special browser plugins. Does not work anymore (2013).
- **CONREAL web server**: Identification and visualization of conserved TFBSs. Only pairwise alignments. No graphical representation.

- **PAP 2.0**: Comprehensive workbench for mammalian transcriptional regulatory sequence analysis. Relies on JVM.
- **BLISS 2.0**: A web-based tool for predicting conserved regulatory molecules in distantly-related orthologous sequences. Only pairwise alignments. No graphical representation.
- **COTRASIF**: Conservation-aided transcription factor binding site finder. Uses old PWM databases. Only textual output and makes only use of promoter regions.
- **DoOPSearch**: A web-based tool for finding and analysing common conserved motifs in the promoter regions of different chordata and plants. Uses only the promoter region and only textual output.

2.5 Technical details and four-step analysis process

ConTra v2 runs on a CentOS 5 server configured with an Apache web server (version 2.2.3), MySQL server (5.0.77), PHP 5.1.6 and Perl 5.8.8. The interface is programmed in PHP, alignments are fetched from UCSC using Perl scripts. TFBS hits for a user-defined motif are calculated using the Match algorithm. An overview picture of these hits, created with Jalview, is embedded in the overview page with help of the Highslide thumbnail viewer (http://www.highslide.com). Dynamic visualization of different TF on the result page is accomplished using Javascript. Scores in the ConTra v2 exploration part are calculated in the same way as in the previous version of ConTra, but due to the inclusion of other genomic regions we no longer take into account the distance to the transcription start site. For each PWM in an alignment block a phylogenetic score is calculated using formula 2.1. In this calculation, n is the number of sequences in the alignment block; abs_score is the absolute phylogenetic score and rel_score is the relative phylogenetic score.

The ConTra v2 analysis consists of four steps. In step one, users have to choose whether they want to visualize or explore a gene of interest. In this step, it is also necessary to indicate the reference species and gene of interest. The second step lists a group of available transcripts for genes matching the search terms, from which one can be selected. For every gene, all possible RefSeq and Ensembl transcript variants are listed with a link to the genomic location in the respective genome browser. In this way, genes with alternative promoters, UTRs or alternative intronic regions can be analyzed for regulatory differences. In step three, different genomic regions for the selected transcript can be chosen (upstream, introns, 5'-UTR and 3'-UTR). The fourth and final step offers users an extensive choice of PWM motifs. Up to 25 PWM motifs can be taken into account for the analysis. For the visualization part, results are split into alignment blocks. These blocks consists of local alignments produced by the TBA program (threaded blockset aligner) [72]. In the exploration part, a list of PWMs is given, ranked according to the prediction score.

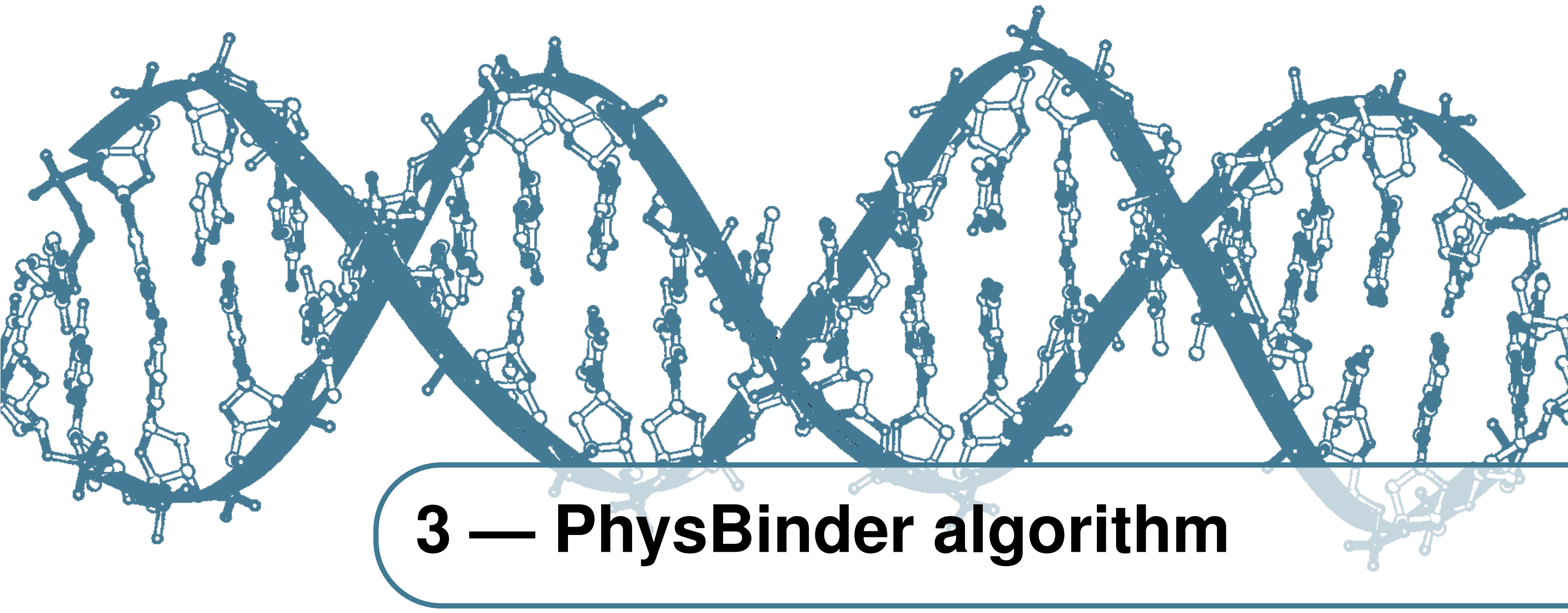
Formula 2.1

```
score = 0
for (i in 1:n){
   if (hit in sequence i){
      abs_score = abs_score + phylogenetic_distance_of_i_to_ref
   }

}
rel_score = abs_score/sum(all_i_distances)
```

2.6 Author contributions

SB and PH wrote the article. SB designed the web-tool and implemented the novel features. PH did the database design and wrote the database update scripts. JG was part of ConTra during his bachelor thesis. BH helped with the first version of ConTra. FVR supported the research. PDB initiated and supported the research and helped with testing.



3 — PhysBinder algorithm

This chapter is a redraft from the publication:

Bart Hooghe*, Stefan Broos*, Frans Van Roy, and Pieter De Bleser **A flexible integrative approach based on random forest improves prediction of transcription factor binding sites**; Nucleic Acids Res. 2012 40: e106 (*Joint first authors)

3.1 Abstract

Transcription factor binding sites (TFBSs) are DNA sequences of 6 to 15 base pairs. Interaction of these TFBSs with transcription factors (TFs) is largely responsible for most spatiotemporal gene expression patterns. Here, we evaluate to which extent sequence-based prediction of TFBSs can be improved by taking into account the positional dependencies of nucleotides (NPDs) and the nucleotide-sequence-dependent structure of DNA. We make use of the Random Forest algorithm to flexibly exploit both types of information. Results in this study show that both the structural method and the NPD method can be valuable for the prediction of TFBSs. Moreover, their predictive values seem

to be complementary, even to the widely used PWM method. This led us to combine all three methods. Results obtained for five eukaryotic TFs with different DNA-binding domains show that our method improves classification accuracy for all five eukaryotic TFs compared to other approaches. Additionally, we contrast the results of seven smaller prokaryotic sets with data of high-quality and show that with the use of data of high-quality we can significantly improve prediction performance. Models developed in this study can be of great use for gaining insight into the mechanisms of TF binding.

3.2 Introduction

DNA-binding specificity of TFs is traditionally viewed as consisting of a direct and an indirect readout component, and the proportion between them differs from one TF to another [73]. The direct readout mechanism is well-defined and involves recognition of specific DNA bases by amino acids. However, there is no deterministic recognition code for the interaction between DNA and protein sequences, essentially because of the influence of the three-dimensional structures of both macromolecules. The influence of the structure of the DNA-binding domain of the TF on the direct recognition code has been clearly shown for some TFs [74]. If DNA-binding specificity were determined only by direct readout, then a probabilistic approach to TF-DNA recognition would suffice. The direct readout does not, however, fully explain the observed variety of sequence composition and binding affinity of binding sites for a specific TF [75]. This is where the indirect readout mechanism comes in. Indirect readout is much less well-defined but takes into consideration protein-DNA interactions that depend on base pairs that are not directly contacted by the protein. These protein-DNA interactions essentially reflect the influence of the structure and thermodynamic properties of the DNA before or upon binding by the TF. DNA is flexible and exhibits sequence-dependent deviations from the idealized B-DNA structure: the deviations arise from the stacking interactions of successive dinucleotides [76, 77]. These structural details have usually been neglected in the analysis of TF-DNA interactions: a probabilistic approach to direct readout is most commonly used as the sole component for prediction of TFBSs, with varying degrees of success. Rohs et al. [78] recently emphasized the importance of the three-dimensional structures of both macromolecules. Direct readout and indirect readout were renamed as base readout and shape readout, respectively. Base readout was subdivided according to either the major or the minor groove of the DNA, whereas shape readout was subdivided into global and local shape recognition. It was argued that individual TFs combine multiple readout mechanisms to achieve DNA-binding specificity. Methods for identifying TFBSs can be classified into two main groups on the basis of the type of data used to model the TF-DNA binding specificity. Sequence-based methods model the binding specificity from a collection of aligned sequences known to bind the TF *in vitro* or *in vivo*. Structure-based methods use information from available crystal structures of TF-DNA complexes (reviewed in Ref [79]). Most sequence-based methods treat DNA as a uniform static structure that is independent of the nucleotide sequence. For example, the widely used position weight matrix (PWM) method [80] takes into account only the nucleotide frequency at each position of the TFBS and assumes independence between those positions. The assumption that the nucleotides add to the binding affinity of TFs independently from each other is called the ‘additivity’ assumption. Based on theoretical concerns and a few experiments for some TFs [81–84], the correctness of this assumption and the quality of the approximation it yields have been discussed in the previous years [85–87]. Most recently, thanks to larger amounts of experimental data, it was shown that for most TFs, dependencies exist between nucleotide positions in their binding sites [88]. This could be expected because it has been suggested that nucleotide positional dependencies observed within TFBSs arise from the structure and biophysical interactions of unbound and TF-bound DNA [85]. Nucleotide positional dependencies are symptoms of shape readout rather than base readout. Nowadays, many sequence-based methods try to model nucleotide dependencies between positions, and thus they implicitly recognize the structural aspects of TF-DNA binding. They yield accuracy improvement over the classic PWM method for most TFs (e.g. Refs [89–92]). A few publications present sequence-based methods that use sequence-dependent structural characteristics explicitly [48, 93–99]. Some of these methods, e.g. [48, 94], report higher accuracies than those obtained by methods that model only nucleotide dependencies. Structure-based methods, by definition, take into account at least some structural characteristics of TF-DNA binding. Some of these methods are valuable for comparative modeling and they seem promising for TFBS prediction as well (e.g. [100, 79]). However, none of the structure-based methods have offered substantial improvement on the PWM method yet.

In this manuscript we present a sequence-based method that uses the Random Forest (RF) algorithm with features that cover either nucleotide positional dependencies or nucleotide-sequence–dependent structural characteristics of the TFBS and its flanking sequences. We call the corresponding models the NPD model and the structural model. We also let our method combine both models and tried to integrate the PWM score in the combined model. The set of one-type models and combined models presented in this paper should be seen as the products of our flexible integrative method, which can easily determine the most appropriate model to use. We measure the accuracy with which our models separate TFBSs from randomly selected genomic sequences, and we compare this measured value to the accuracy of the classic PWM method and the most recent alternative method, namely CRoSSeD [48]. Results are given for five eukaryotic TFs that bind differently to DNA: HIF1 (zipper-type group/Helix-Loop-Helix family), P53 (zinc-coordinating group/Loop-Sheet-Helix family), SP1 (zinc-coordinating group/BetaBetaAlpha-zinc finger family), STAT1 (Stat protein family) and TBP (Beta-sheet group/TATA box-binding family) [101]. Our method was also used on seven prokaryotic data sets that were presented along with CRoSSeD [48] and a more recent data set for the Fis transcription factor [102].

3.3 Material and methods

3.3.1 Data

Positive sequences are those that are bound *in vivo* at least under some cellular conditions. They were extracted from various sources. Binding sites for HIF1, STAT1 and TBP were fetched from Pazar [103], for SP1 from TRANSFAC (licensed version 2008.4) [35], and for P53 from a paper [104]. TBP binding sites were from human, mouse and rat. The binding sites for the other TFs were all human. When necessary, TFBSs were mapped back to genomic coordinates.

PWMs available from TRANSFAC (licensed version 2008.4) [35] were used with the search algorithm MATCH [105] to align the fetched binding sites. These matrices were V$STAT1_01, V$SP1_Q2_01, V$TBP_01, and V$HIF1_Q3. The known TFBSs were positioned to the nearest TFBS predicted by the appropriate PWM using the TRANSFAC-given threshold values to minimize false negatives (minFN threshold values). These threshold values enable recognition of at least 90% of positive sequences, but come along with a high rate of false positives. We excluded the sequence if no predicted TFBS was found within 20 bp on either side of the position given by the database. The P53 binding sites from the paper were not re-aligned because they were already annotated in sufficient detail. We considered only P53 binding sites that were tagged as qualitative and gapless [104]. In this way, our data sets of positives consisted of 55 binding sites for HIF1, 87 for P53, 243 for SP1, 209 binding sites for STAT1, and 88 for TBP. In order to assess the performance on prokaryotic data sets, we used binding sites for AraC (13 sites), ArcA (44 sites), Fis (135 sites), FlhDC (12 sites), IHF (70 sites), LexA (13 sites) and PurR (17 sites) from the CRoSSeD article [48]. As an additional control for the prokaryotic data, we also used the large and qualitative ChIP-chip data set for Fis published by Cho et al. [102]. Negative or background sequences are randomly selected from the human or E. coli genome. We take ten times as many negative sequences as the corresponding number of positives. We must provide enough negatives to ensure consistency of results, but not so many that the RF algorithm could suffer from an imbalance of the training data set, which would cause the focus to be too much on the classification accuracy of the majority class.

3.3.2 Structural characteristics

Structural characteristics used for this manuscript comprise characteristics calculated from scratch (see below for curvature and torsion calculations) and characteristics extracted from the literature. Most of these are correlated to some extent, but we let a feature selection procedure decide which characteristics and combinations thereof are most useful for identifying binding sites for each TF. Each DNA-sequence–dependent structural characteristic is described by a list of all possible polynucleotides of a certain length, to which a numerical value describing the structural characteristic is assigned. For every characteristic, positions in a DNA sequence are scored by the value of the appropriate polynucleotide. The calculation of sequence-dependent structural values requires an assumption of a certain three-dimensional (3D) structure of the DNA. As we did not want to assume one specific DNA structural model, we implemented three different models: a model derived from protein-bound DNA [106], one from unbound DNA [97], and another from nucleosome-bound DNA [107, 108]. Each of these DNA structural models consists of values for all base pair step parameters (roll, twist, tilt, rise, shift and slide) for each dinucleotide or trinucleotide (For a visual overview of these base pair step parameters see figure 3.1).

This enabled us to convert DNA sequences into 3D coordinates by using the rebuilding part of 3DNA [109], a program for analysis, rebuilding and visualization of 3D nucleic acid structures. For each of the DNA structural models, we did this conversion on 10,000 randomly generated sequences of 100 bp. From the resulting 3D coordinates, we then calculated the values of our structural characteristics. Values calculated for a specific structural characteristic but with coordinates coming from different DNA structural models were eventually treated as values for different structural characteristics. Curvature and torsion of the helix's axis were calculated from the coordinates of this axis only, each for the highest possible resolution. The formulas we used are as follows:

Formula 3.1 Curvature: If a, b and c are three consecutive points on the helix's axis, then $\overrightarrow{U_A} = \overrightarrow{ab} \times \overrightarrow{ac}$ is orthogonal to the plan A formed by a, b and c. The curvature in b of the line containing a, b and c is given by the following equation:

$$C_b = \frac{2 \cdot |\vec{U_A}|}{|\overrightarrow{ab}| \cdot |\overrightarrow{bc}| \cdot |\overrightarrow{ac}|} \tag{3.1}$$

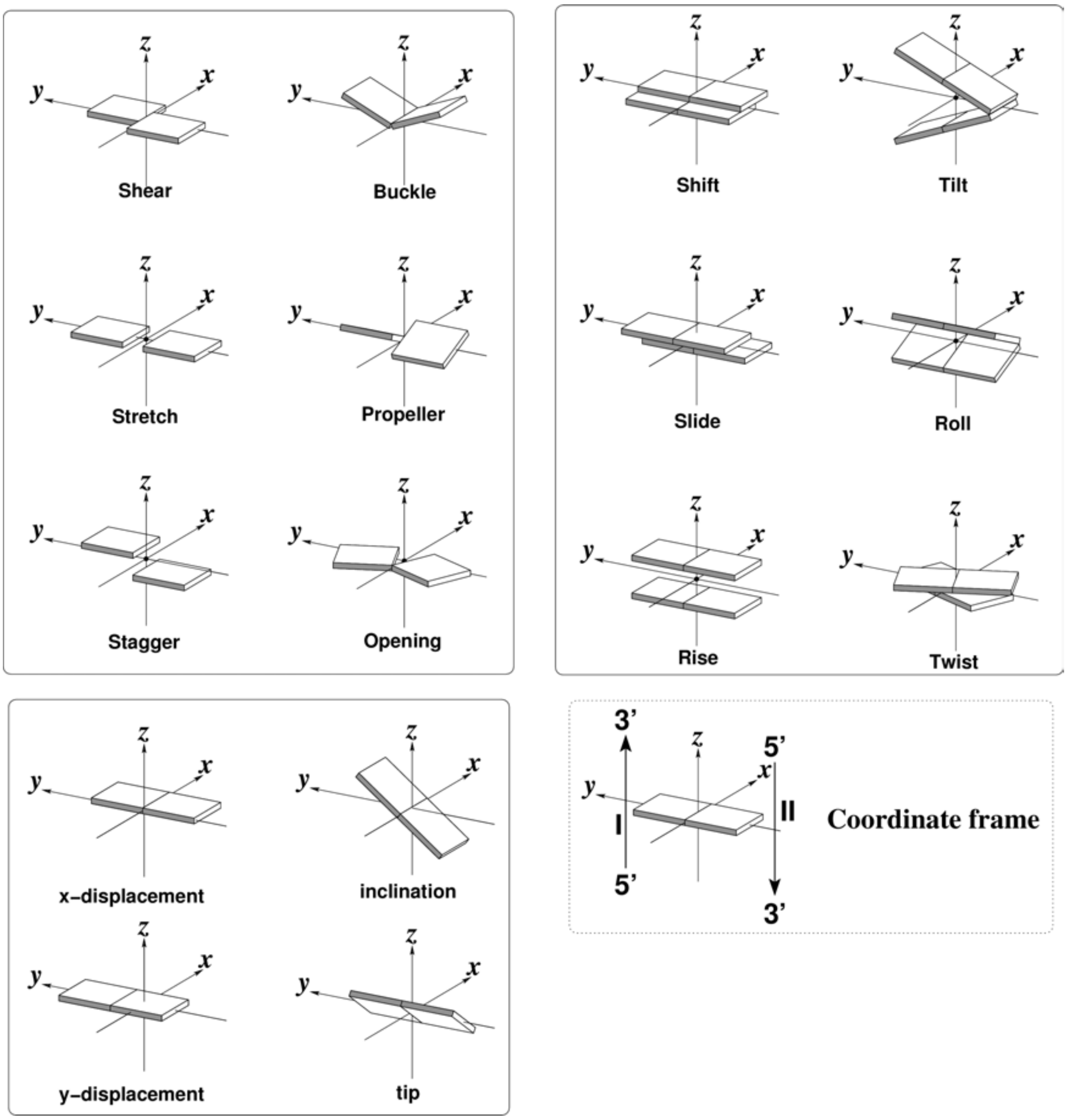


Figure 3.1: Visual overview of some of the base pair step parameters. These base pair step parameters enabled us to convert DNA sequences into 3D coordinates. Image taken from the 3DNA website (http://3dna.rutgers.edu/x3dna/examples) [109].

Formula 3.2 Torsion (dihedral angle): If a, b, c and d are four consecutive points on the helix's axis, then $\overrightarrow{U_A} = \overrightarrow{ab} \times \overrightarrow{ac}$ is orthogonal to the plane A formed by a, b and c, and $\overrightarrow{U_B} = \overrightarrow{bc} \times \overrightarrow{bd}$ is orthogonal to the plane B formed by b, c and d. Then the dihedral angle is given by the following equation:

$$T_{AB} = cos^{-1} \frac{|\overrightarrow{U_A} \cdot \overrightarrow{U_B}|}{|\overrightarrow{U_A}| \cdot |\overrightarrow{U_B}|} \tag{3.2}$$

These calculations provide a value for every base position. However, this value is calculated with coordinates of more than just this one base (see equations above) and these coordinates are dependent on the identity of neighboring bases. We sought to determine an accurate relation between sequence and calculated structural values, and so we took the shortest length of polynucleotides for which the relative standard deviation on the corresponding mean structural value was lower than 1%. This polynucleotide length is three, four or five, depending on the characteristic and the DNA structural model. The calculated values of sequence-dependent structural characteristics (curvature and torsion of helix's axis) are available from the authors upon request. Other structural characteristics used in this manuscript were extracted from the literature and comprise properties derived from either unbound or TF-bound DNA. They are all given as a value per dinucleotide, mostly with a considerably large standard deviation. The standard deviations, when expanding to polynucleotides longer than two bases, indicate that the structural characteristics of base pair steps depend on the identity of neighboring nucleotides. Although we used higher nucleotide lengths having nearly no standard deviation on their mean value for the structural characteristics we calculated ourselves, the calculation was still based on the assumption of DNA structural models described by only dinucleotides or trinucleotides. The structure of a dinucleotide is known to be influenced by the identity of the neighboring nucleotides [110, 93, 111]; and taking into account these next-nearest-neighbor effects might further improve the accuracy of the structural model. A description of the structural characteristics we used is given below.

Curvature and torsion describe the DNA backbone in its highest resolution and thus provide at least a measure of bending. The characteristic we implemented, **directed bending**, does the same [112]. Directed bending means the extent to which a dinucleotide tends to bend towards either the major or the minor groove when it is bound by a TF, and it is used as a measure of deformability of DNA. Values are determined on sequences bound by the TF CAP at sites where sequence dependence of bending is maximal [112]. Pre-bending of free DNA [113] and TF-induced bending [114] have been recognized for more than a decade as structural motifs common to many TF-DNA complexes.

Groove clash distance and size are both components of the clash function that was constructed to give a quantitative interpretation of the observed sequence dependence of TF-DNA interactions on DNA twist [115]. A steric clash between exocyclic groups results from out-of-plane base pair distortions. Its size is defined as the sum of the radii for the exocyclic groups interacting in the grooves. Clash distance is the distance between the centres of the interacting groups when they are in an “idealized” conformation.

Different geometries of the major and minor groove are taken into account and result in separate values per groove type [115]. **Groove shape** is an interesting characteristic to explore because it was recently acknowledged that most TFs recognize the minor groove width upon specific binding [116]. The value of **groove width** for prediction of TFBSs was suggested by Liu et al. in 2001 [97]. Minor groove opening is a measure of the degree to which a base step is open in the minor groove, and hence it is related to the above-mentioned measure of groove clash size. The values are derived from high-resolution crystal structures of unbound DNA in BI conformation [97].

Conformational tendency is measured by the standardized Pearson residuals for the test of uniformity or homogeneity of the individual dinucleotide steps over different conformations, i.e. structural types of DNA [117]. These values are derived from unbound DNA and represent the tendency of a dinucleotide to favor a specific DNA conformation. Uniformity of dinucleotides is tested between A-type, B-type and combined conformational families (A, B and A+B conformations) and within B-types of DNA (BI, BII, A/B, B/A, RESTB). RESTB is not assigned to any of the existing conformational families. We did not use the conformational tendencies of dinucleotides within A-forms of DNA because the dinucleotide AA/TT does not occur there [117]. Almost a third of dinucleotides from protein-DNA complexes adopt AI or AII conformations. This plasticity of DNA, which allows the conformation to change locally from the common B-form into an A-form, is one of the ways in which DNA achieves specificity in protein-DNA binding [118, 114, 119].

3.3.3 Random Forest algorithm

The Random Forest (RF) algorithm [120] (http://www.stat.berkeley.edu/ breiman/ RandomForests/cc_home.htm) is a tree-based machine learning algorithm and is the engine of both our structural method and our NPD method. It is an ensemble classifier that consists of many individual decision trees (CARTs: classification and regression trees) and outputs the class that is predicted by the majority of those trees. Tree-based methods consist of non-parametric statistical approaches for regression and classification analyses. Classification trees are grown by recursively partitioning the observations into subgroups with a more homogeneous categorical response. At each node, the explanatory variable giving the most homogeneous subgroups is selected. For the CART tree learning algorithm, this selection is based on Gini impurity, which is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset.

Tree-based methods can be very effective for selecting from large numbers of predictor variables those that best explain the observations. They make no implicit assumptions about the form of underlying relationships between the predictor variables and the response, and so they might detect non-linear associations. The RF methodology forms an ensemble of unpruned classification or regression trees (CARTs) by bootstrapping samples of the training data and using random feature selection in the tree induction process. It generally exhibits a substantial performance improvement over the single tree classifier such as CART and C4.5. The biggest disadvantage of RF is that its embedded feature selection procedure cannot handle large numbers of irrelevant features. For this reason, we performed a comprehensive filter feature selection and wrapper-based feature selection before the final model is trained (see next section: “Building classification models”). We used FastRandomForest (http://fast-random-forest.googlecode.com/), a parallellized implementation in Java. For further information, we refer to two publications that provide excellent explanations and examples on the use of RF for modeling dependencies among variables [121, 122].

3.3.4 Building classification models

In the first stage of building a classification model, one model per characteristic is built. The structural method uses the above-mentioned structural characteristics, whereas the characteristics of the NPD method are represented by mononucleotides and dinucleotides. Hence, each sequence from the positive and negative set is converted into a series of structural vectors or is split up into mononucleotides or dinucleotides (figure 3.2 A/B). We perform a comprehensive feature selection in order to obtain the final model.

A first round of feature selection is performed in a purely statistical way to make a basic selection of positions where a difference exists between the values for the characteristic of the positives and those of the negatives (so-called filter feature selection) (figure 3.2 C). The statistical tests are applied with mild threshold values in order not to exclude too many features and to permit detection of their interactions by the RF algorithm later on. For the structural model, we consider values for all positions in the TFBS and for the 50 bases flanking the start position of the binding site, as well as the mean value over all these positions, as features to be used in building the model. The Kolmogorov-Smirnov test at a false discovery rate threshold of 0.1 is used to determine the significance of differences between values at each position. The Wilcoxon rank test at a threshold of 0.05 is used to determine the significance of differences between values averaged over all 100 positions. For the NPD model, 50 mononucleotides

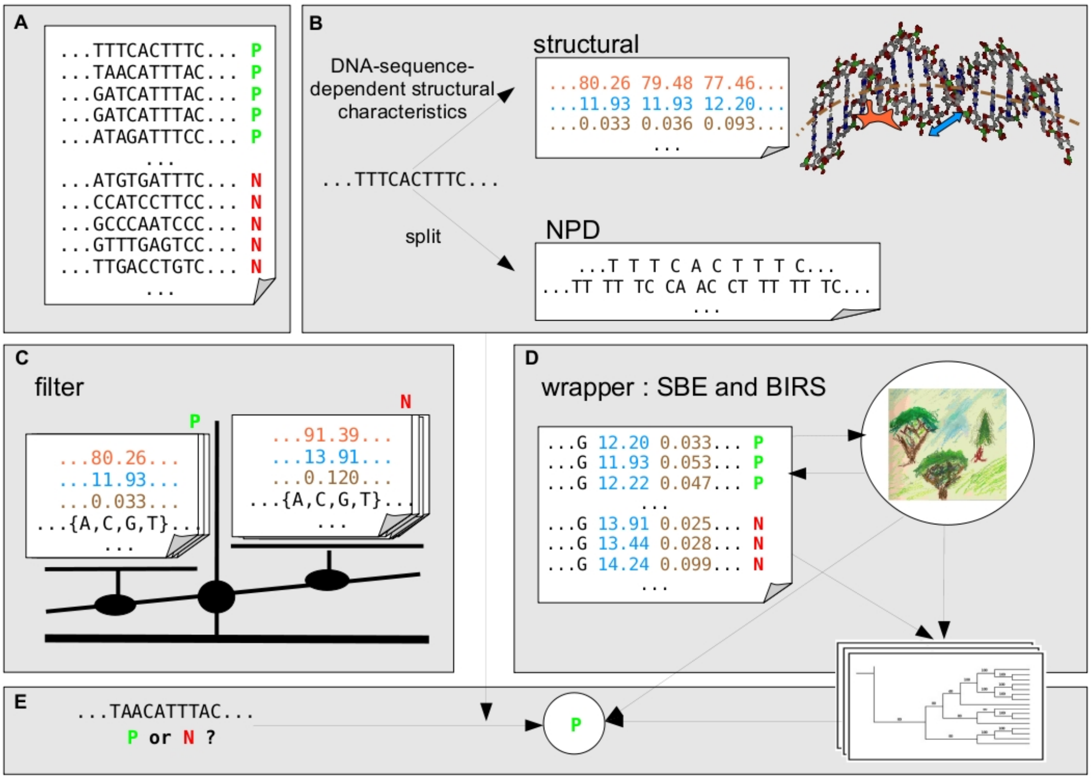


Figure 3.2: Overview of our approach: A, The input from which models are built consists of the two classes of nucleotide sequences that the method should learn to separate. One class contains positive sequences (P, green) known to be bound *in vivo*; the other contains negative sequences (N, red) highly unlikely to be bound *in vivo*. B, Each nucleotide sequence, from either class, is converted into multiple series of values; each series provides values for a specific DNA structural characteristic at all positions of the TFBS and its context (structural model), or simply consists of one-base or two-base parts of the sequence (NPD). C, Basic selection of relevant features (i.e. positions) is made by statistical comparison of distributions of values for positive and negative sequences with mild thresholds. D, Further selection is performed through wrapper-based feature selection, i.e. cross-validation performance evaluation with the Random Forest algorithm. Per characteristic, redundant features are removed by sequential backwards elimination (SBE). Several models with one characteristic might be merged through a best incremental ranked subset scheme (BIRS). The final NPD model and final structural model can be merged into one integrative model. E, The resulting model can be used by Random Forest to predict the likelihood that a nucleotide sequence is a TFBS, after converting the sequence into series of the features contained in the model.

flanking the TFBS start on both sides are considered. The basic selection of positions at which the mononucleotide distribution is different between positives and negatives is determined by the test for equality of proportions. More specifically, a position is selected when the sum of the logs of the p-values of proportion tests is significantly different from the background using a threshold of 0.1.

In the second round of feature selection, the preliminary model based on one characteristic is subjected to wrapper-based feature selection (figure 3.2 D). We repeatedly evaluate the accuracy of the model by cross-validation with the RF algorithm and remove features of the basic selection when this does not cause a significant decrease in accuracy (measured as either F-measure or AUC). AUC (area under the curve) represents the area under the ROC curve, whereas F-measure is the weighted harmonic mean of precision and recall. This procedure of removing insignificant features is also called sequential backwards elimination (SBE). It makes the model sparser, which permits better interpretation of the features it contains and which improves speed upon application. At this stage we end up with one model per characteristic. We rank all models according to their classification accuracy as determined by cross-validation (measured as AUC). Starting with the best performing one-characteristic model, we cumulatively merge it with lower-ranked models according to the BIRS (best incremental ranked subset) scheme; this implies the use of wrapper-based feature selection. Combined models, i.e. models that contain characteristics from two or three different categories (NPD, structural or PWM score) are simply built by merging two or more models that are restricted to one category. The process of finding the combination that gives the best model can be easily automated by an extra round of wrapper-based feature selection. When building PWMs for the eukaryotic sets, we automatically assigned their lengths by requiring that the start is on the assumed start position of the TFBSs and the end is characterized by three consecutive positions with an information content of at least 1.1. For the prokaryotic sets it was necessary to use the entire sequence length for the PWM.

3.3.5 Evaluation of classification models

The evaluation of classification models is based on their prediction scores and provides an estimation of the accuracy of their classification. The prediction score of both the structural method and the NPD method is the RF confidence score, which is assigned to each sequence and indicates the certainty with which this sequence is predicted to belong to either the positive or the negative class. In the case of PWMs, we used the matrix similarity score [105]. The evaluation of performance is visualized by ROC curves (Receiver Operating Characteristic) and precision-recall curves. Each ROC and precision-recall curve shown is derived from a threshold-based average of 20 curves. Data for each of these 20 curves were obtained by training the model with a randomly taken subset of 80% of the data and testing that trained model on the remaining 20%. Principle component analysis was performed on the full models using the Weka 3 suite [123] and used to select a top five feature set for each TF (default parameters).

3.4 Results

Based on the Random Forest (RF) algorithm [120], we initially built two types of models. The so-called structural model uses one or more structural characteristics by employing their values at specific positions or their average value over all positions in the TFBS and its flanking sequences. The so-called NPD model accounts for positional dependencies at the nucleotide level, utilizing only nucleotide identities (mononucleotides and dinucleotides). The procedure of building and using these models is depicted in figure 3.2 and explained in detail in Methods. We start by discussing the classification accuracy of the classic PWM method, the structural method, the NPD method and combinations thereof, and compare our integrative method with a recent alternative method. This evaluation is performed on five eukaryotic data sets of high quality and eight prokaryotic data sets. Seven of these prokaryotic data sets are rather small and less well-annotated. This led us to introduce a second, Fis data set of high quality in order to assess the influence of data quality on the performance of the different methods. As an additional confirmation of the validity of the RF method, we evaluate the integrative TBP model on external data. Finally, we look at the selected features in each model and try to relate these features to what has been reported in the literature.

3.4.1 Classification accuracy

The ROC curve (receiver operating characteristic) is a standard representation of the trade-off between false positive rate (FPR) and sensitivity. We use details of ROC curves to visualize the classification accuracy of the models. Regular ROC curves and their corresponding measure AUC (area under the curve) cover the full range of FPRs from 0 to 1 and are thus of not much use for estimating the discriminatory power of a predictor of TFBSs [124]. Genome-wide predictions performed with an FPR even as small as 0.01 are not really useful because they would return an overload of false positives, for example about six million for the human genome. Therefore, we focus on the part of the ROC curves that corresponds to the lower, more relevant range of FPR. We also take our most integrative model as a reference model and for each model we list the FPR that corresponds to the TPR that has an FPR of 0.01-0.1 for this reference model, corresponding to the bending point of the curves. Statistics of pairwise comparisons of these FPRs are provided as well. We compare our models with each other and also compare their accuracy with the accuracy of our home-made high-quality PWMs and with the most recently proposed alternative method, CRoSSeD [48]. The latter comparison will be discussed extensively in the next section.

For the eukaryotic transcription factors (figure 3.3 and tables 3.1-3.2), both structural and NPD models perform better than the PWM for four out of five TFs (HIF1, SP1, STAT1, TBP).

Overall, the NPD model performs better than the structural model (four out of five cases). This is logical because the structural method almost exclusively captures the shape readout mechanisms of DNA binding specificity. All base readout information gets lost upon conversion from a nucleotide sequence to vectors of structural characteristics. The NPD model, in contrast, is expected to capture base readout, as well as some portion of the shape readout that can be derived from nucleotide positional dependencies. Nevertheless, the structural models

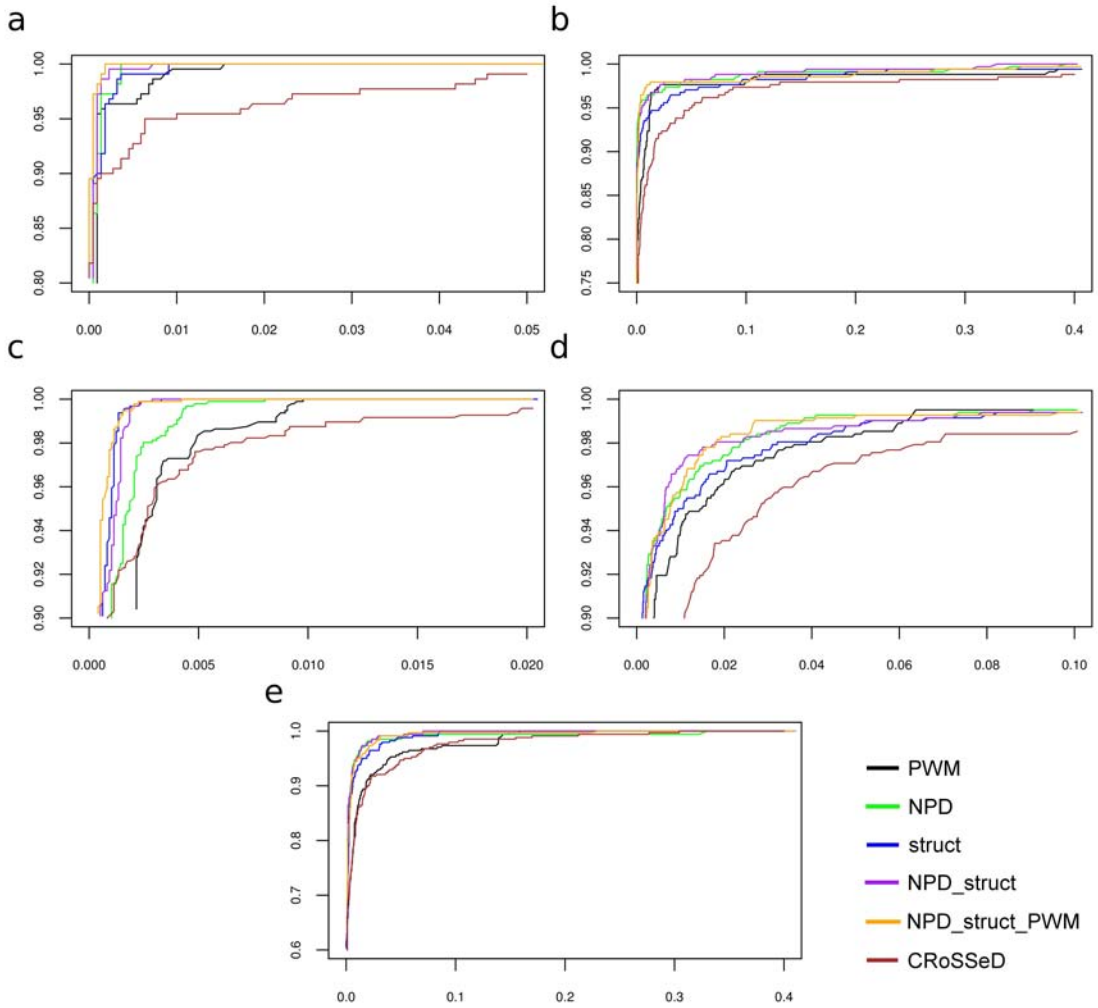


Figure 3.3: Accuracy of classification models in identifying TFBSs, as assessed for five eukaryotic TFs. Details of threshold-averaged ROC curves showing the trade-off between TPR (Y axis) and FPR (X axis); Classification models applied: PWM (black), NPD (green), struct (blue), NPD_struct (purple), NPD_struct_PWM (orange), CRoSSeD (brown). (A-E). ROC curves for various transcription factors: A. HIF1 B. P53; C. SP1; D. STAT1; E. TBP.

		A	B				
HIF1	**REF_TPR(FPR0.01):**	1 +- 0	**PWM**	**NPD**	**struct**	**NPD_struct**	**NPD_struct_PWM**
	PWM	0.015 +- 0.01423	1				
	NPD	0.00364 +- 0.00619	4.37E-03	1			
	struct	0.00909 +- 0.01216	1.44E-01	1.74E-01	1		
	NPD_struct	0.00682 +- 0.00715	8.10E-02	1.23E-01	9.07E-01	1	
	NPD_struct_PWM	**0.00182 +- 0.00373**	5.40E-04	4.00E-01	3.83E-02	0.0144	1
	CRoSSeD	0.12091 +- 0.1893	2.66E-04	4.07E-07	1.01E-05	1.68E-06	1.21E-07
P53	**REF_TPR(FPR0.01):**	0.976 +- 0.04	**PWM**	**NPD**	**struct**	**NPD_struct**	**NPD_struct_PWM**
	PWM	0.021 +- 0.0097	1				
	NPD	0.0391 +- 0.0113	1.90E-05	1			
	struct	0.075 +- 0.02072	8.97E-08	2.27E-06	1		
	NPD_struct	0.021 +- 0.0117	8.79E-01	7.48E-05	1.04E-07	1	
	NPD_struct_PWM	**0.00920 +- 0.00755**	3.60E-04	7.71E-08	6.05E-08	1.11E-03	1
	CRoSSeD	0.1213 +- 0.09719	2.25E-04	2.35E-02	2.79E-01	1.72E-04	1.12E-06
SP1	**REF_TPR(FPR0.01):**	1 +- 0	**PWM**	**NPD**	**struct**	**NPD_struct**	**NPD_struct_PWM**
	PWM	0.00977 +- 0.00371	1				
	NPD	0.00802 +- 0.00411	2.05E-01	1			
	struct	0.00329 +- 0.00270	3.36E-06	4.34E-04	1		
	NPD_struct	**0.00288 +- 0.00279**	2.75E-06	1.52E-04	5.77E-01	1	
	NPD_struct_PWM	0.00422 +- 0.00295	1.31E-05	3.83E-03	3.06E-01	1.58E-01	1
	CRoSSeD	0.06276 +- 0.04094	1.49E-07	9.92E-08	6.17E-08	5.92E-08	6.03E-08

Table 3.1: Accuracy of classification models in identifying TFBSs, as assessed for five eukaryotic TFs. A, FPR at the TPR that corresponds to a FPR of 0.01 for the reference model NPD_struct_PWM; this TPR REF_TPR(FPR0,01) is mentioned in the column header. The best model is indicated in bold B, Assessment of statistical significance in FPR difference by Wilcoxon two-tailed p-value; P<0.01 is indicated with a red background; P<0.05 is indicated with an orange background.

		A	B				
STAT1	**REF_TPR(FPR0.01):**	0.957 +- 0.026	**PWM**	**NPD**	**struct**	**NPD_struct**	**NPD_struct_PWM**
	PWM	0.0176 +- 0.00773	1				
	NPD	0.00993 +- 0.0047	1.30E-03	1			
	struct	0.01447 +- 0.00506	1.80E-01	5.85E-03	1		
	NPD_struct	**0.00658 +- 0.0041**	1.24E-05	2.94E-02	2.74E-05	1	
	NPD_struct_PWM	0.00885 +- 0.00460	2.98E-04	4.77E-01	9.60E-04	1.35E-01	1
	CRoSSeD	0.0311 +- 0.02175	2.26E-02	7.33E-05	7.10E-04	1.60E-05	4.49E-05

TBP	**REF_TPR(FPR0.01):**	0.95 +- 0.064	**PWM**	**NPD**	**struct**	**NPD_struct**	**NPD_struct_PWM**
	PWM	0.03835 +- 0.01678	1				
	NPD	0.00938 +- 0.00591	6.24E-07	1			
	struct	0.01449 +- 0.00677	9.36E-06	2.09E-02	1		
	NPD_struct	**0.00910 +- 0.00675**	4.65E-07	7.09E-01	1.07E-02	1	
	NPD_struct_PWM	0.00994 +- 0.00579	7.95E-07	7.02E-01	3.99E-02	4.07E-01	1
	CRoSSeD	0.05284 +- 0.0904	2.49E-02	1.05E-01	5.11E-01	8.40E-02	1.55E-01

Table 3.2: Continuation of table 3.1. Accuracy of classification models in identifying TFBSs, as assessed for five eukaryotic TFs. A, FPR at the TPR that corresponds to a FPR of 0.01 for the reference model NPD_struct_PWM; this TPR REF_TPR(FPR0,01) is mentioned in the column header. The best model is indicated in bold B, Assessment of statistical significance in FPR difference by Wilcoxon two-tailed p-value; non-significance due to extreme variance across repeats is indicated in red font. P<0.01 is indicated with a red background; P<0.05 is indicated with an orange background.

alone perform surprisingly well: they perform better than PWM in four out of five cases. For most eukaryotic TFs, merging the structural model with the NPD model leads to clear synergistic effects and achieves a classification accuracy that is superior to the accuracy of the separate models and PWM ('NPD_struct'). For three out of five eukaryotic transcription factors, inclusion of the PWM score even led to an additional improvement ('NPD_struct_PWM'). The Random Forest strategy significantly improved upon the PWM method for all eukaryotic TFs (tables 3.1 and 3.2).

For most prokaryotic models (figure 3.4 and tables 3.3-3.5), the NPD model and the structural model do not outperform the PWM. When considering the low-resolution prokaryotic data sets alone (figure 3.4 A-G), the structural or NPD model, or combinations thereof, perform better than the PWM model for only three out of seven TFs (ArcA,FlhDC and IHF).

Combining the NPD model and the structural model leads to an improvement in five out of seven cases when compared to the individual models. Adding the PWM score did not result in an additional improvement, except for AraC. Compared to the other prokaryotic models, the high-quality Fis model performs exceptionally well (figure 3.4 H). This result clearly demonstrates the importance of using qualitative data when building classification models.

As an additional test, we also looked into precision-recall curves of the classification models for a growing number of background sequences (available in the supplemental data of the published paper: http://nar.oxfordjournals.org/content/suppl/2012/03/17/gks283.DC1/nar-01633-met-n-2011-File010.pdf). With this type of analysis, we tested the models for their ability to cope with a growing number of background sequences. For each TF we compared the combined RF model with the PWM for ten different background sizes. We started with a 1:1 ratio and augmented the number of background sequences until we had a 1:10 ratio. Models that are less-suited to cope with many background sequences show a sharper decline in the precision-recall curves when facing more negative sequences. The prokaryotic models gave mixed results. Again, the high-quality Fis model performs exceptionally better than the other prokaryotic models. The RF models of ArcA and IHF perform equally well as the PWM, whereas the rest of the TFs did not benefit from the more complex RF model. However, unlike the prokaryotic models, the eukaryotic models gave consistent results. For all five eukaryotic TFs, the RF model turned out to be more robust against a growing number of background sequences compared to the simpler PWM model.

The difference in classification performance between the two Fis sets is striking (figure 3.4 H and table 3.5). The results indicate that with the high-quality Fis set the RF model can improve upon the PWM method. In this case, NPD_struct_PWM is the best model and it is significantly better than all other models. It is clear that the overall classification accuracy of all the methods we compared is much better with the more reliable Fis data set. We speculate that lack of improvement for the RF models in the majority of prokaryotic sets is due to their relatively small sizes and poor quality of annotation, as is illustrated with this example.

3.4.2 Comparison with alternative sequence-based methods

Differences between our method and others include accounting for the context of the TFBS, the use of several structural characteristics instead of just one, the use of structural values for specific positions rather than just the average value along the TFBS, the use of both structural characteristics and nucleotide positional dependencies, and the use of the Random Forest algorithm. Random Forest does not require any assumptions about the form of underlying relationships between the predictor variables and the response. Hence, there is no need to assume independence or uniform contribution of multiple structural characteristics. Some other sequence-based methods use additional types of data to reduce the FPR of TFBS prediction, such as phylogenetic conservation [125], genome annotation (e.g. Refs [126, 127]) or specific experimental results (e.g. Ref [128]). We only consider sequence-based methods not needing such additional information as methods comparable to ours. Some of these methods are SiteSleuth [93], promapper [99] and CRoSSeD [48]. Each of them is based on a different classification algorithm, namely, support vector machine, Bayesian network, and conditional random field, respectively. Furthermore, base readout and shape readout are captured in slightly different ways (e.g. other structural characteristics) and do not get equal chances due to arbitrary decisions. We conclude that with the exception of CRoSSeD [48], none of all previously presented methods

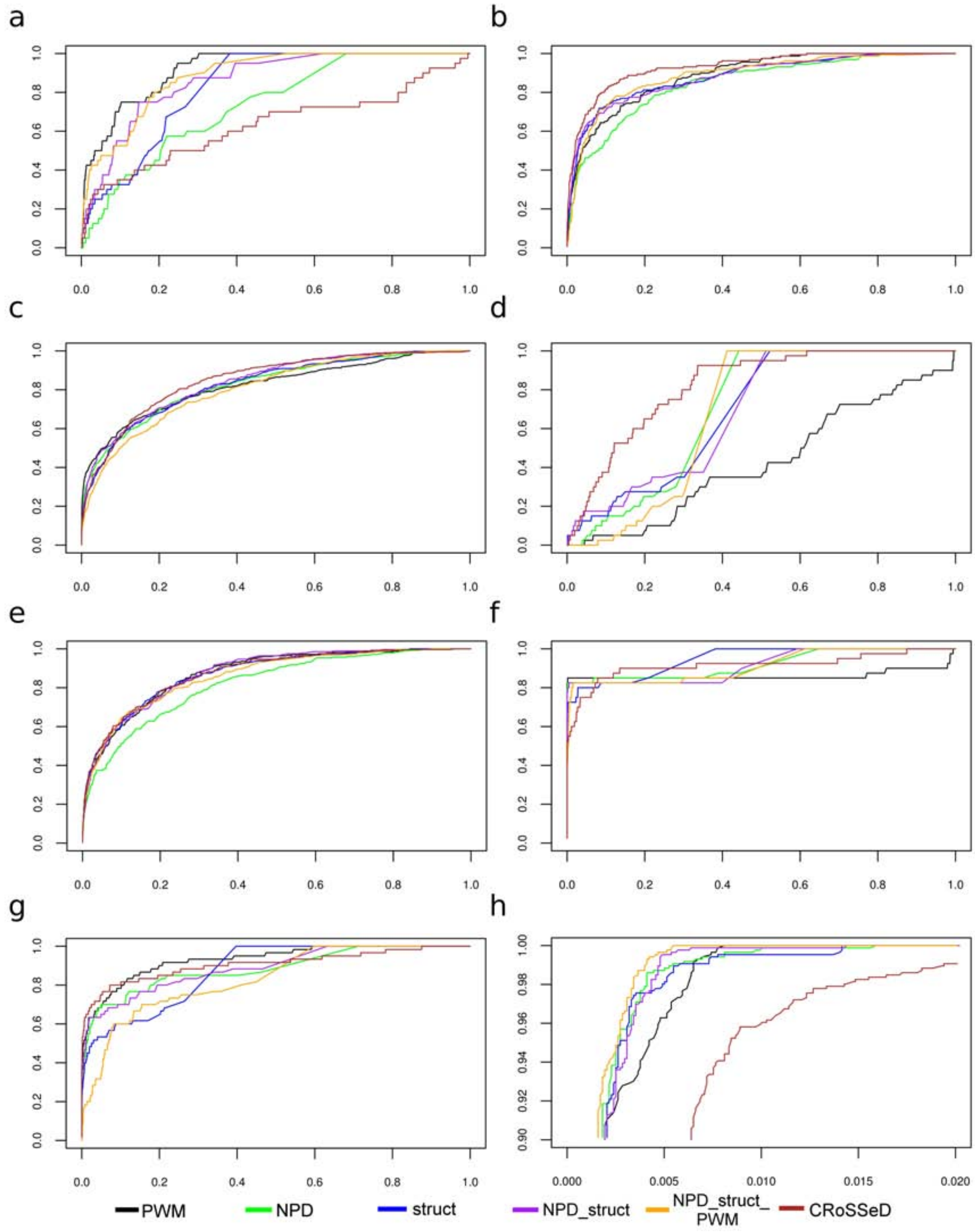


Figure 3.4: Accuracy of classification models in identifying TFBSs, as assessed for eight prokaryotic TFs. Threshold-averaged ROC curves showing the trade-off between TPR (Y axis) and FPR (X axis); Classification models applied: PWM (black), NPD (green), struct (blue), NPD_struct (purple), NPD_struct_PWM (orange), CRoSSeD (brown). (A-H). ROC curves for various transcription factors: A. AraC; B. ArcA; C. Fis; D. FlhDC; E. IHF. F. LexA. G. PurR. H. Fis (ChIP-chip set).

		A	B				
AraC	**REF_TPR(FPR0.05):**	0.45 +-0.2763	**PWM**	**NPD**	**struct**	**NPD_struct**	**NPD_struct_PWM**
	PWM	**0.02425 +- 0.0133**	1				
	NPD	0.1855 +-0.05867	6.54E-08	1			
	struct	0.1452 +- 0.03246	6.50E-08	2.45E-02	1		
	NPD_struct	0.0775 +- 0.02409	3.93E-07	2.30E-07	3.52E-07	1	
	NPD_struct_PWM	0.04225 +- 0.0283	0.0194	1.21E-07	1.39E-07	0.000473	1
	CRoSSeD	0.2185 +- 0.2306	0.249	8.07E-01	0.871	5.15E-01	2.97E-01

ArcA	**REF_TPR(FPR0.05):**	0.5625 +- 0.2088	**PWM**	**NPD**	**struct**	**NPD_struct**	**NPD_struct_PWM**
	PWM	0.058 +- 0.0178	1				
	NPD	0.1038 +- 0.02757	2.58E-06	1			
	struct	0.0355 +- 0.01169	1.05E-04	7.42E-08	1		
	NPD_struct	0.03025 +- 0.0118	7.42E-06	6.33E-08	1.05E-01	1	
	NPD_struct_PWM	0.04825 +- 0.0150	8.71E-02	1.72E-07	6.14E-03	4.00E-04	1
	CRoSSeD	**0.025 +- 0.01662**	9.82E-06	6.54E-08	0.0136	1.45E-01	1.61E-04

Fis	**REF_TPR(FPR0.05):**	0.3741 +- 0.1067	**PWM**	**NPD**	**struct**	**NPD_struct**	**NPD_struct_PWM**
	PWM	**0.01225 +- 0.0072**	1				
	NPD	0.02825 +- 0.0132	5.82E-05	1			
	struct	0.04075 +- 0.0165	4.16E-07	1.55E-02	1		
	NPD_struct	0.02375 +- 0.0186	3.84E-02	2.19E-01	4.15E-03	1	
	NPD_struct_PWM	0.0495 +- 0.02865	2.85E-05	1.68E-02	3.21E-01	4.93E-03	1
	CRoSSeD	0.03675 +- 0.9997	8.89E-05	3.74E-01	1.54E-01	1.03E-01	6.31E-02

Table 3.3: Accuracy of classification models in identifying TFBSs, as assessed for eight prokaryotic TFs. A, FPR at the TPR that corresponds to a FPR of 0.01-0.1 (due to poor classification performances of FlhDC) for the reference model NPD_struct_PWM; this TPR is mentioned in the column header. The best model is indicated in bold B, Assessment of statistical significance in FPR difference by Wilcoxon two-tailed p-value; non-significance due to extreme variance across repeats is indicated in red font. P<0.01 is indicated with a red background; P<0.05 is indicated with an orange background.

		A	B				
FlhDC	**REF_TPR(FPR0.1):**	0.025 +- 0.1118	**PWM**	**NPD**	**struct**	**NPD_struct**	**NPD_struct_PWM**
	PWM	0.0445 +-0.02077	1				
	NPD	0.038 +- 0.03454	3.42E-01	1			
	struct	**0.00125 +- 0.0028**	2.36E-08	4.77E-05	1		
	NPD_struct	0.0018 +- 0.0037	3.06E-08	8.33E-05	7.24E-01	1	
	NPD_struct_PWM	0.07875 +- 0.0276	2.80E-04	2.70E-04	2.59E-08	3.36E-08	1
	CRoSSeD	0.00425 +- 0.0150	3.15E-07	6.81E-05	5.00E-01	3.16E-01	4.29E-08

IHF	**REF_TPR(FPR0.05):**	0.475 +- 0.1043	**PWM**	**NPD**	**struct**	**NPD_struct**	**NPD_struct_PWM**
	PWM	0.05325 +- 0.0159	1				
	NPD	0.08625 +- 0.0244	3.66E-05	1			
	struct	0.04025 +- 0.0143	1.92E-02	2.77E-07	1		
	NPD_struct	0.04975 +- 0.0179	3.92E-01	2.48E-05	1.35E-01	1	
	NPD_struct_PWM	0.04725 +- 0.0131	2.25E-01	1.92E-06	1.69E-01	9.03E-01	1
	CRoSSeD	**0.03975 +- 0.0214**	5.38E-02	2.53E-06	1.00E+00	1.88E-01	2.89E-01

LexA	**REF_TPR(FPR0.05):**	0.825 +- 0.2447	**PWM**	**NPD**	**struct**	**NPD_struct**	**NPD_struct_PWM**
	PWM	**0 +- 0**	1				
	NPD	0.00125 +- 0.0028	4.00E-02	1			
	struct	0.08 +- 0.03742	7.80E-09	2.57E-08	1		
	NPD_struct	0.005 +- 0.00459	2.33E-05	3.69E-03	9.18E-08	1	
	NPD_struct_PWM	0.015 +- 0.00811	2.33E-05	3.69E-03	9.18E-08	1.00E+00	1
	CRoSSeD	0.0725 +- 0.1794	9.58E-03	2.78E-01	1.42E-04	2.63E-01	2.63E-01

Table 3.4: Continuation of table 3.3. Accuracy of classification models in identifying TFBSs, as assessed for eight prokaryotic TFs. A, FPR at the TPR that corresponds to a FPR of 0.01-0.1 (due to poor classification performances of FlhDC) for the reference model NPD_struct_PWM; this TPR is mentioned in the column header. The best model is indicated in bold B, Assessment of statistical significance in FPR difference by Wilcoxon two-tailed p-value; P<0.01 is indicated with a red background; P<0.05 is indicated with an orange background.

		A	B				
PurR	**REF_TPR(FPR0.05):**	0.3333 +- 0.2861	**PWM**	**NPD**	**struct**	**NPD_struct**	**NPD_struct_PWM**
	PWM	**0.00025 +- 0.0011**	1				
	NPD	0.00125 +- 0.0022	8.42E-02	1			
	struct	0.005 +- 0.00607	2.58E-04	1.46E-02	1		
	NPD_struct	**0.00025 +- 0.0011**	1.00E+00	8.42E-02	2.58E-04	1	
	NPD_struct_PWM	0.049 +- 0.02474	1.16E-08	3.98E-08	1.62E-07	1.16E-08	1
	CRoSSeD	0.00075 +- 0.0018	3.10E-01	4.47E-01	2.59E-03	3.10E-01	2.35E-08

Fis (*)	**REF_TPR(FPR0.01):**	1 +- 0	**PWM**	**NPD**	**struct**	**NPD_struct**	**NPD_struct_PWM**
	PWM	0.00776 +- 0.00358	1				
	NPD	0.01564 +- 0.00617	7.45E-05	1			
	struct	0.01416 +- 0.00416	3.45E-05	4.43E-01	1		
	NPD_struct	0.01427 +- 0.00583	4.40E-04	4.94E-01	9.13E-01	1	
	NPD_struct_PWM	**0.00537 +- 0.00299**	3.44E-02	1.29E-06	4.40E-07	6.21E-06	1
	CRoSSeD	0.03368 +- 0.0179	2.16E-07	4.72E-03	2.01E-05	5.02E-05	9.02E-08

Table 3.5: Continuation of table 3.4. Accuracy of classification models in identifying TFBSs, as assessed for eight prokaryotic TFs. A, FPR at the TPR that corresponds to a FPR of 0.01-0.1 (due to poor classification performances of FlhDC) for the reference model NPD_struct_PWM; this TPR is mentioned in the column header. The best model is indicated in bold B, Assessment of statistical significance in FPR difference by Wilcoxon two-tailed p-value; P<0.01 is indicated with a red background; P<0.05 is indicated with an orange background.

have made clear comparisons to show how accurately their method identifies TFBSs compared to methods modeling dependencies between nucleotide positions, and that CRoSSeD is the current best performing alternative method. Here, we clearly show the value of each of the ‘pure approaches’ (PWM, nucleotide positional dependencies, structural), and we show that integration of different approaches is beneficial to classification accuracy. We performed a quantitative comparison with the most recent alternative method, namely CRoSSeD. We compared our method with CRoSSeD both on the prokaryotic data set from the CRoSSeD article and on our eukaryotic data sets. The results on the eukaryotic data sets are depicted in figure 3.3. For all eukaryotic TFs, CRoSSeD separates TFBSs from non-TFBSs less accurately than the PWM. Our integrative model (‘NPD_struct’ and ‘NPD_struct_PWM’) performs significantly better than CRoSSeD for all eukaryotic TFs.

The prokaryotic data sets that were used originally come from RegulonDB [129] and are remarkably different from the eukaryotic data sets we used. Most of the prokaryotic data sets show very little sequence conservation and only expose weak signals over a long distance (see supplementary data of Meysman et al. [48]). The lack of strong nucleotide conservation in most prokaryotic data sets might have caused CRoSSeD to be developed with a different focus from our RF models. The different natures of the prokaryotic data sets are reflected by a much lower level of classification accuracy of the predictors and we were forced to list the FPR that corresponds to the TPR with an FPR of 0.05 or even 0.1 for the reference model ‘NPD_struct_PWM’, instead of the 0.01 used for the eukaryotic data sets (tables 3.1-3.2). Our ROC curves and some conclusions differ from those shown in the paper presenting CRoSSeD. The different results must have been caused by differences in the evaluation setup. Many papers, including Meysman et al., measure accuracy by the area under the ROC curve (AUC), but differences of its value might be irrelevant or even misleading, depending on the shapes of the ROC curves. Both CRoSSeD and our integrative method are among the best models in three out of seven cases (figure 3.4 A-G), but what is truly remarkable is that the PWM proves to be the best model in three out of seven cases when considering low FPRs only. We also compared our methods to the CRoSSeD method on the prokaryotic Fis set of high quality (figure 3.4 H).

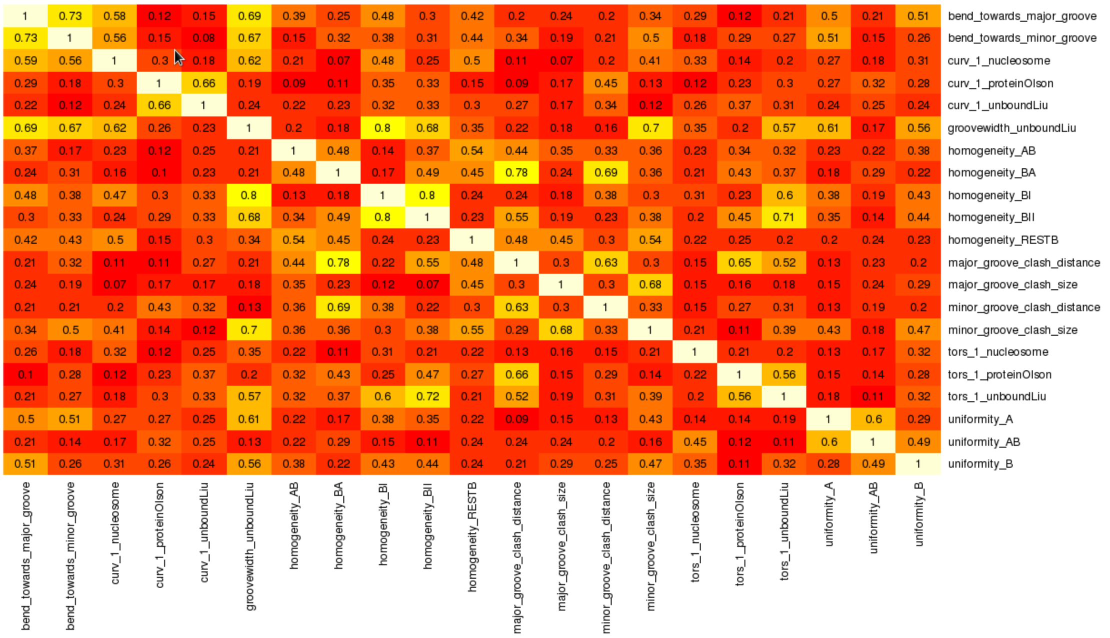
With this data set, the performance of all methods improves drastically. The RF method performs best, while the CRoSSeD method lags behind. These results make clear that data quality is an important determinant of model performance. From both comparisons with CRoSSeD, we conclude that our approach performs better overall. The small prokaryotic data sets did not fully meet the requirements of our qualitative approach to evaluation of models, and hence conclusions should be made carefully.

3.4.3 Evaluation of a model on external data

The seemingly small improvements in accuracy presented here may nevertheless make a huge difference when identifying TFBSs on large DNA sequences and genome-wide. Furthermore, it is interesting to evaluate models on data that do not originate from the same data set with which the models were built. In order to evaluate our method on external data, we tested the TBP model on an independent ChIP-seq experiment for TBP [55]. This is a very demanding test, since the models need to identify the TBP binding site in a wider peak region of the ChIP-seq experiment. The same is then repeated for a background with the same length distribution. In table 3.6 we compare the PWM method, our integrated model (containing structural and NPD characteristics) and the CRoSSeD tool in terms of ROC AUC for classification of sequences containing *in vivo* TBP binding sites and background sequences. Results clearly show that the PWM (AUC 0.535) and CRoSSeD (AUC 0.574) can barely discriminate between the TBP peaks and the background model, whereas our integrated model fulfills this task much better (AUC 0.774).

	PWM	RF model	CRoSSeD
ROC AUC	0.535	0.774	0.573

Table 3.6: Performance of the TBP model on external ChIP-seq TBP data set (Mokry et al. [55]) measured in ROC AUC.



bend_towards_major_groove	bend_towards_minor_groove	curv_1_nucleosome	curv_1_proteinOlson	curv_1_unboundLiu	groovewidth_unboundLiu	homogeneity_AB	homogeneity_BA	homogeneity_BI	homogeneity_BII	homogeneity_RESTB	major_groove_clash_distance	major_groove_clash_size	minor_groove_clash_distance	minor_groove_clash_size	tors_1_nucleosome	tors_1_proteinOlson	tors_1_unboundLiu	uniformity_A	uniformity_AB	uniformity_B	
1	0.73	0.58	0.12	0.15	0.69	0.39	0.25	0.48	0.3	0.42	0.2	0.24	0.2	0.34	0.29	0.12	0.21	0.5	0.21	0.51	bend_towards_major_groove
0.73	1	0.56	0.15	0.08	0.67	0.15	0.32	0.38	0.31	0.44	0.34	0.19	0.21	0.5	0.18	0.29	0.27	0.51	0.15	0.26	bend_towards_minor_groove
0.59	0.56	1	0.3	0.18	0.62	0.21	0.07	0.48	0.25	0.5	0.11	0.07	0.2	0.41	0.33	0.14	0.2	0.27	0.18	0.31	curv_1_nucleosome
0.29	0.18	0.3	1	0.66	0.19	0.09	0.11	0.35	0.33	0.15	0.09	0.17	0.45	0.13	0.12	0.23	0.3	0.27	0.32	0.28	curv_1_proteinOlson
0.22	0.12	0.24	0.66	1	0.24	0.22	0.23	0.32	0.33	0.3	0.27	0.17	0.34	0.12	0.26	0.37	0.31	0.24	0.25	0.24	curv_1_unboundLiu
0.69	0.67	0.62	0.26	0.23	1	0.2	0.18	0.8	0.68	0.35	0.22	0.18	0.16	0.7	0.35	0.2	0.57	0.61	0.17	0.56	groovewidth_unboundLiu
0.37	0.17	0.23	0.12	0.25	0.21	1	0.48	0.14	0.37	0.54	0.44	0.35	0.33	0.36	0.23	0.34	0.32	0.23	0.22	0.38	homogeneity_AB
0.24	0.31	0.16	0.1	0.23	0.21	0.48	1	0.17	0.49	0.45	0.78	0.24	0.69	0.36	0.21	0.43	0.37	0.18	0.29	0.22	homogeneity_BA
0.48	0.38	0.47	0.3	0.33	0.8	0.13	0.18	1	0.8	0.24	0.24	0.18	0.38	0.3	0.31	0.23	0.6	0.38	0.19	0.43	homogeneity_BI
0.3	0.33	0.24	0.29	0.33	0.68	0.34	0.49	0.8	1	0.23	0.55	0.19	0.23	0.38	0.2	0.45	0.71	0.35	0.14	0.44	homogeneity_BII
0.42	0.43	0.5	0.15	0.3	0.34	0.54	0.45	0.24	0.23	1	0.48	0.45	0.3	0.54	0.22	0.25	0.2	0.2	0.24	0.23	homogeneity_RESTB
0.21	0.32	0.11	0.11	0.27	0.21	0.44	0.78	0.22	0.55	0.48	1	0.3	0.63	0.3	0.15	0.65	0.52	0.13	0.23	0.2	major_groove_clash_distance
0.24	0.19	0.07	0.17	0.17	0.18	0.35	0.23	0.12	0.07	0.45	0.3	1	0.3	0.68	0.15	0.16	0.18	0.15	0.24	0.29	major_groove_clash_size
0.21	0.21	0.2	0.43	0.32	0.13	0.36	0.69	0.38	0.22	0.3	0.63	0.3	1	0.33	0.15	0.27	0.31	0.13	0.19	0.2	minor_groove_clash_distance
0.34	0.5	0.41	0.14	0.12	0.7	0.36	0.36	0.3	0.38	0.55	0.29	0.68	0.33	1	0.21	0.11	0.39	0.43	0.18	0.47	minor_groove_clash_size
0.26	0.18	0.32	0.12	0.25	0.35	0.22	0.11	0.31	0.21	0.22	0.13	0.16	0.15	0.21	1	0.21	0.2	0.13	0.17	0.32	tors_1_nucleosome
0.1	0.28	0.12	0.23	0.37	0.2	0.32	0.43	0.25	0.47	0.27	0.66	0.15	0.29	0.14	0.22	1	0.56	0.15	0.14	0.28	tors_1_proteinOlson
0.21	0.27	0.18	0.3	0.33	0.57	0.32	0.37	0.6	0.72	0.21	0.52	0.19	0.31	0.39	0.2	0.56	1	0.18	0.11	0.32	tors_1_unboundLiu
0.5	0.51	0.27	0.27	0.25	0.61	0.22	0.17	0.38	0.35	0.22	0.09	0.15	0.13	0.43	0.14	0.14	0.19	1	0.6	0.29	uniformity_A
0.21	0.14	0.17	0.32	0.25	0.13	0.22	0.29	0.15	0.11	0.24	0.24	0.24	0.2	0.16	0.45	0.12	0.11	0.6	1	0.49	uniformity_AB
0.51	0.26	0.31	0.26	0.24	0.56	0.38	0.22	0.43	0.44	0.24	0.21	0.29	0.25	0.47	0.35	0.11	0.32	0.28	0.49	1	uniformity_B

Figure 3.5: Pearson correlation analysis of the different features. Color indication of correlation goes from dark red (no correlation) over yellow (slight correlation) to white (high correlation). bend_towards_major_groove and bend_towards_minor_groove are characteristics included in the category referred to as “directed bending”; Homogeneity and uniformity are characteristics included in the category referred to as “conformational tendency”.

3.4.4 Features contained in the models

Appendix section A shows all features of the RF models. The features contained in the models can reveal aspects of the DNA-TF binding mechanism. Even though the prokaryotic models do not perform that well in terms of classification, the selected features can tell us something about the binding mode of these TFs. All TFs have different models with different characteristics, representing their DNA binding specificities. The structural characteristics are correlated to some extent (figure 3.5), but we let the feature selection procedures and the RF algorithm decide which features are most relevant for each TF.

It should be noted that for each TF both the structural model and the NPD model include features at positions that precede the actual TFBS. Moreover, each model contains one or more mean values as feature which implies that the global structural *in vivo* context of the TFBS is an important feature next to more

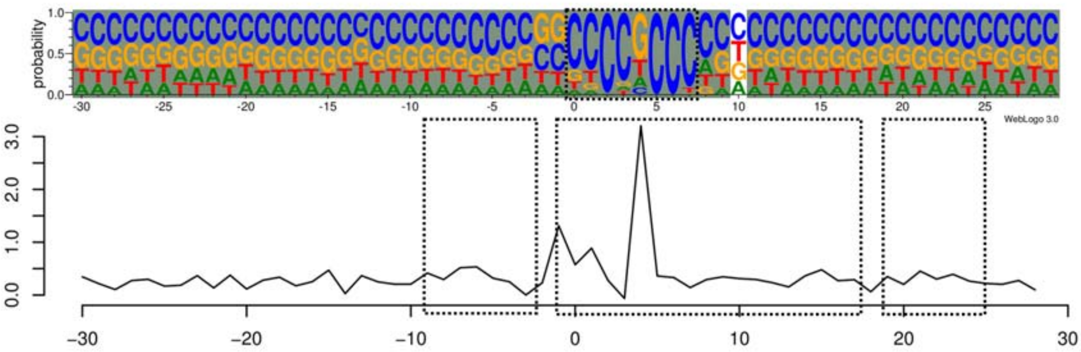


Figure 3.6: Visualization of our integrative model for SP1. Top, mononucleotide frequencies with the positions of the NPD model shown as shaded boxes. Bottom, average value of one of the structural characteristics contained in the structural model, namely conformational tendency RESTB; positions of the structural model are indicated by dotted-line boxes. PWM of SP1 is indicated on the sequence logo with dotted-line box. (X axes indicate position relative to the aligned start of the SP1 binding sites; Y axis value of the RESTB feature).

local shape readout mechanisms at or close to the binding site location. This global shape readout might reflect the general part of higher-order protein-DNA interactions that determine binding specificity and functionality: the tendency of a nucleosome to bind the region in which the TFBS is embedded [78]. It might thus be considered part of a so-called ‘general binding preference’ that was demonstrated to be important for improved prediction of TFBSs [127]. A visualization of the SP1 model (figure 3.6) clearly shows how the background genomic sequence in which SP1 binding sites are embedded is very similar to the consensus sequence of such sites. A PWM would thus predict many TFBSs, whereas the NPD model and structural model can look beyond position-independent nucleotide frequencies, each in its own way. In the next section, we will describe the most important features of each model, together with their biological relevance.

3.4.5 Biological relevance of the selected features

To assess the biological relevance of the selected features, we decided to do a principal component analysis (PCA) on the different TF models. For each model we selected the top five principal components, meaning the five most relevant

features according to the PCA (table 3.7 with PCA analysis of the eukaryote TFs; table 3.8 with PCA analysis of the prokaryote TFs).
We relate all of the selected features to what is known in the literature about structural protein-DNA complex formation. Unfortunately, for torsion related features we were unable to find explanations in the literature because this feature is not discussed in most protein-DNA reports. The PWM score is an important feature in most models and is considered the primary feature for direct readout. It should be noted that a strong deviation in the bending towards the major groove also means a deviation in the bending towards the minor groove. That is why we discuss these features as "bending towards the major/minor groove." The same goes for the conformational tendency of the DNA. We were able to explain most of the top features of each model, but unable to provide an explanation for the selected features for FlhDC. Although many prokaryotic classification models, in contrast to the eukaryotic models, did not result in any significant improvements over the simpler methods, the selected features and models can provide us with some valuable information about the binding mode of the protein. This information can be used to gain some insight even before any crystal structures are solved.

In most prokaryotic models, the role of direct readout is very important. This is represented by the PWM score feature. This feature will not be discussed separately for every transcription factor. It is striking that for prokaryotic transcription factors the PWM score is the best feature in six out of eight models while for eukaryotic transcription factors it is the best feature in only one out of five models. This can be explained by a recent systematic study on the differences between prokaryotic and eukaryotic TFBSs published by Wunderlich et al. [130], in which the authors calculated the average information content (IC) of both prokaryotic and eukaryotic TFBSs. They conclude that the average IC of a prokaryotic TFBS is 23 bits compared to 12.1 bits for eukaryotic TFBSs. This remarkable difference is mainly due to the shorter average length of the eukaryotic binding sites.

AraC, a regulator of the araBAD operon in E. coli, binds as a dimer to the DNA [131]. AraC proteins make all sequence specific contacts in the major

Table 3.7: Results of the PCA analysis for the eukaryotic transcription factors. For each TF model, we selected the five best features according to Weka PCA analysis.

Transcription Factor	Feature

HIF1	uniformity_A_fullseqmean
	dint_p5=CG
	PWMmatrixscore_general
	dint_p6=GT
	dint_p7=TG

P53	uniformity_A_fullseqmean homogeneity_BI_fullseqmean homogeneity_RESTB_fullseqmean PWMmatrixscore_general homogeneity_RESTB_p2
SP1	homogeneity_RESTB_fullseqmean PWMmatrixscore_general uniformity_AB_fullseqmean dint_p5=CC dint_p6=CC

STAT1	PWMmatrixscore_general
	dint_p13=AA
	dint_p5=TT
	dint_p12=GA
	dint_p7=TC

TBP	bend_towards_major_groove_fullseqmean bend_towards_minor_groove_fullseqmean homogeneity_BII_fullseqmean bend_towards_minor_groove_p8 bend_towards_major_groove_p8

Table 3.8: Results of the PCA analysis for the prokaryotic transcription factors. For each TF model, we selected the five best features according to Weka PCA analysis.

Transcription Factor	Feature
AraC	PWMmatrixscore_general
	minor_groove_clash_size_fullseqmean
	minor_groove_clash_size_p18
	monont_p19=G
	monont_p0=A
ArcA	PWMmatrixscore_general
	groovewidth_unboundLiu_fullseqmean
	groovewidth_unboundLiu_p0
	groovewidth_unboundLiu_p1
	groovewidth_unboundLiu_p-1
Fis	PWMmatrixscore_general
	PWMcorescore_general
	uniformity_A_p-2
	uniformity_A_fullseqmean
	uniformity_A_p-3
IHF	bend_towards_major_groove_fullseqmean
	bend_towards_minor_groove_fullseqmean
	PWMmatrixscore_general
	bend_towards_major_groove_p-6
	bend_towards_major_groove_p-7
FlhDC	PWMmatrixscore_general
	monont_p-3=C
	monont_p-20=G
	monont_p-3=T
	tors_1_nucleosome_p-7
LexA	minor_groove_clash_distance_p-8
	dint_p-8=GC
	PWMmatrixscore_general
	minor_groove_clash_distance_p-7
	minor_groove_clash_distance_p-9
PurR	PWMmatrixscore_general
	PWMcorescore_general
	monont_p-5=A
	monont_p-4=A
	monont_p1=T

groove. Structural reports indicate that both monomers of the dimeric AraC proteins are separated by an AT-rich linker, resulting in an overall bend and a smaller overall minor groove clash size [132]. This last feature is clearly reflected in the top five feature list of the AraC model.

In the ArcA model, groove width is a very important feature both as a positional feature and as a global mean feature. This is in agreement with the data on the OmpR/PhoB family of transcription factors, of which ArcA is a member [133, 134]. Just like clash size, width of both the major and the minor groove is an important feature in the winged helix-turn-helix (HTH) family of transcription factors. In this family of transcription factors, a helix is inserted in the major groove of the DNA, whereas the wings of the protein dimer are inserted in the minor groove [134].

Fis is known as one of the nucleoid-associated proteins (NAPS). Such proteins are responsible for the packing of the prokaryotic chromosome by bending and supercoiling of the DNA [135]. For Fis, two models are available: one with a limited number of binding sites and one more trustworthy ChIP-chip model, which we used as a quality control case. The smaller of the two models contains, among the direct readout features many features concerning the A/B-DNA tendency signifying the reported deviations from standard B-DNA [136]. The top features of the ChIP-chip model are a bit more diverse. Since Fis is one of the NAPS proteins, the appearance of the bending property in the list of PCA top features should come as no surprise. Other important features are both G/C mononucleotides on position 0 and +14. The presence of these features is very important because methylation of these positions on either strand is known to completely inhibit Fis binding [135]. The location of these nucleotides is in agreement with the major groove contacts by Fis. The TT dinucleotide feature is also an important *in vivo* feature: it corresponds to the center of the AT-track that is responsible for the bending properties of the DNA in the binding site [102].

The top five components in the IHF model consist mainly of features concerning DNA bending towards the major/minor groove. Since IHF is one of the most extreme DNA benders known, also called “the master bender” [137, 138], the inclusion and importance of the selected features should not be a surprise. This is also reflected in the RF model. The most important feature of this protein is the overall mean of the bend towards major/minor groove, making it one of the few prokaryotic models with a biophysical feature as a top feature, which is in agreement with the IHF's title as master bender.

For LexA, the most noticeable features are the minor groove clash size features between –7 and –9 (the linker region between two LexA half sites). This is also reported in the literature, where an unusually narrow minor groove and important clash interactions are observed in the linker region between two LexA half sites in order to fit into the network of interactions between the two half sites [139]. The selected GC dinucleotide feature is also of importance to the minor groove clash size: the occurrence of GC is disfavored because this dinucleotide has the largest minor groove clash size of all nucleotides. This is in agreement with earlier reports, which state that LexA has a preference for A/T-rich spacer regions [140, 139].

In the model of the purine repressor (PurR), the top five features consist only of monomeric sequence features and PWM scores. This suggests that this model focuses on the direct readout of PurR binding.

For the HIF1 transcription factor, three out of five top features are dinucleotide features. The dinucleotides together, one after the other, build the pattern 5'-CGTG-3', known as the hypoxia-response element (HRE). This pattern is the most important determining factor of HIF1 binding and is fully conserved in every HIF1 binding site. These HREs are cis-regulatory DNA sequences for the specific binding to HIF1 and are necessary for transcription upon hypoxic conditions [141–143]. The model was able to capture this sequence element very well.

For P53, the majority of important features concern the DNA conformation and the tendency to the A/B-DNA conformation. The DNA conformation is shown to be a very important determinant in the sequence-specific binding by P53.

Although P53 binding sites are very degenerate, P53 can bind strongly to a wide range of binding sites. It has been suggested that a shift to a non-standard B-DNA conformation can drastically alter the binding capacity of P53 and that this conformational shift is responsible for the specific binding to the wide variety of P53 motifs [144].

SP1 is known to unwind the DNA from 10.5 residues per turn to 11.2 residues per turn, thereby greatly distorting the standard B-structure of the DNA towards a more A-DNA oriented structure and other deviant structures [145, 146]. Two out of five top features of the SP1 model confirm the importance of DNA conformational features in aiding the binding specificity of SP1 to the DNA, both of which are global features. The other top features are more sequence-oriented. The two CC-dinucleotide features in the model are an indication of the cytosine enrichment in the canonical SP1 recognition element (CCCGCC). Furthermore, the importance of CC dinucleotides has been discussed by Zhu et al. [147] who found that methylation of the central CG dinucleotide did not impair SP1 binding, but methylation of the first CC dinucleotides significantly decreased SP1 binding specificity. This important feature of the specific binding of SP1 was correctly included as one of the top features in the RF model.

STAT1, like all other STATs, shows a very strong preference for sequences containing two palindromic half-sites (TTC...GAA), leading to a dyad symmetry where the STAT1 dimer can bind [148]. The inclusion of the dinucleotide features for AA, TT, GA and TC, together TTTC...GAAA, is the most specific variant of all STAT1 binding motifs according to an analysis made by Ehret et al. [149].

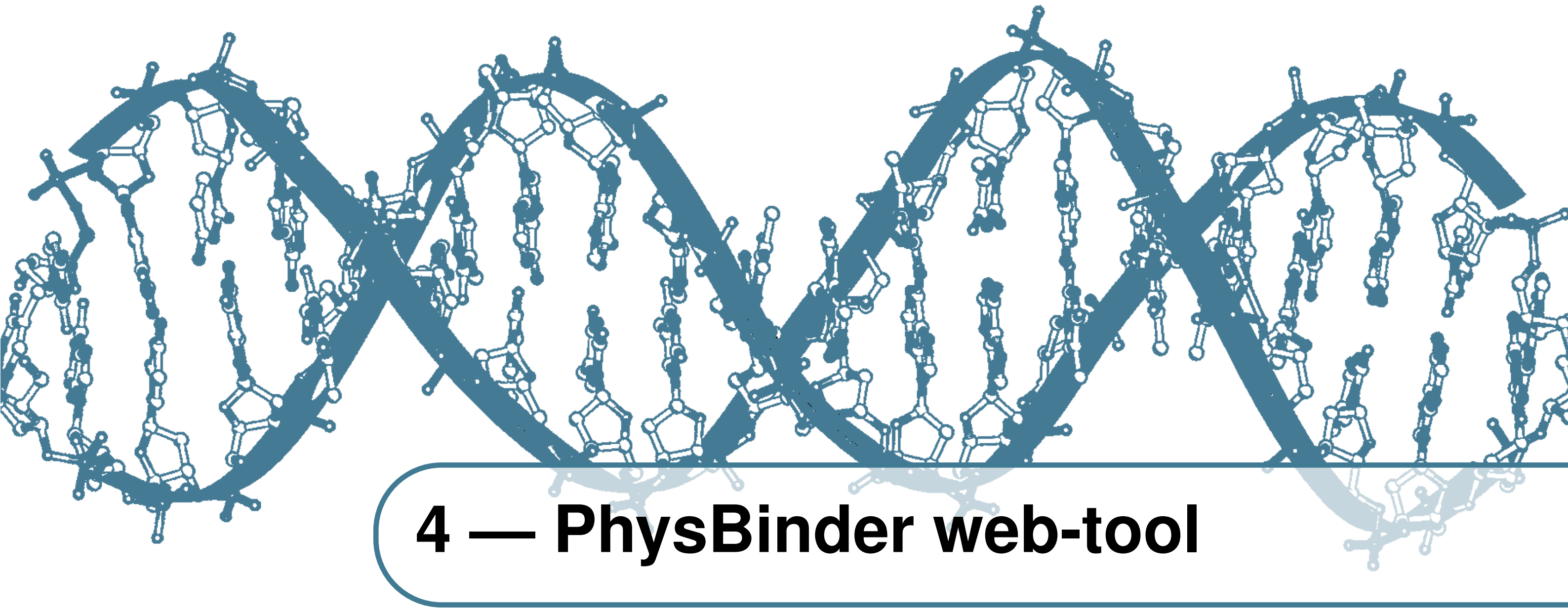
TBP is one of the most well-known DNA benders [150, 151] and it was shown that the unbound TATA box is already pre-bent [152]. The properties of introducing a kink in the DNA are also well reflected in the model. When looking at the top five features, four out of five top features contain properties about DNA bending, confirming the tendency of TBP to bend the DNA.

3.5 Discussion

It has been known for a few decades that the structure of DNA varies in a sequence-dependent manner [152, 77, 76]. Some recent papers stressed the importance of sequence-dependent structural properties of DNA by showing that they are much less diverse than the nucleotide sequences but at the same time they contain additional information [153, 45]. That makes the structure space better suited than the nucleotide sequence space for seeking patterns [154, 155, 45]. Several papers pointed specifically to the role of DNA shape in protein-DNA recognition [156, 45, 116, 157]. Rohs et al. published a comprehensive review on this topic [116]. In the past decade, only few proposed methods for TFBS identification explicitly took into account the nucleotide-sequence-dependent structural properties of DNA. However, many other methods implicitly capture some part of shape readout mechanisms of DNA-binding specificity when they model positional dependencies of nucleotides, and they tend to predict TFBSs more accurately than the widely used PWM. For prokaryotes, the apparent lack of improvement for the more complex RF models can have several causes. The size of these data sets is relatively small while complex models like the structural or NPD model might require bigger and better annotated data sets. The additional tests on the more qualitative Fis control set seem to confirm this hypothesis. A simpler method, like a PWM based strategy, was developed for use with small data sets and apparently performs quite well on most prokaryotic data sets. An alternative, more biological explanation for the poor performance of our models on prokaryotes lies in the differences between prokaryotic and eukaryotic transcription factors. A systematic analysis of the differences in binding strategy between prokaryotic and eukaryotic binding sites revealed that prokaryotic binding sites tend to be longer and that they have more information content [130]. In eukaryotes, the presence of the binding site alone is not enough and binding is often aided by signals in the flanking regions. Prokaryotes have few spurious binding sites, making the presence of one binding site alone a distinctive feature. This, in combination with the smaller and lower quality set of binding sites, might lead to an overall decrease in performance of the more complex models and give the more simple PWM an advantage, as revealed by comparing the two Fis sets. For eukaryotes, our results indicate that the inherent structural properties of DNA are involved in specific recognition by the TFs to an extent that depends on each TF, and that these properties can be used to refine predictions. Our results show that a purely structural model performs worse than a model capturing the positional dependencies of nucleotides most of the time. The latter type of model is represented in our comparison by our NPD model, which we believe models both base readout and a big portion of shape readout. The relative importance of the more simple NPD characteristic consequently cannot be ignored when analyzing TFBS binding patterns in the eukaryotic models. We demonstrate, however, that structural properties contain information other than the nucleotide sequence, and that the use of this information can be used to further improve classification accuracy. We demonstrate that the PWM score, which merely represents base readout in its most simple form, is sometimes complementary to the model combining the structural model and NPD model. Most importantly, we present an integrative approach that can easily combine two or three different approaches to establish the best possible prediction of TFBSs. Further improvements of our purely structural model might be achieved by using higher-resolution descriptions of structural characteristics and incorporation of additional ones, such as those available in the database for dinucleotide properties [46]. Additionally, input for sequence-based methods is currently gathered in a way that favors the performance of detection methods using nucleotide identities. Sequences containing TFBSs are aligned by methods focusing on nucleotide conservation only, such as existing PWMs or MEME [38]. It could be worthwhile to improve the alignment correction in a way that it takes into account structural vectors. This might even lead to a further improvement for the structural models. Shape readout is thought to fine-tune binding affinity rather than determine the binding event [78]. In this respect, the structural part of the combinatorial model might prove itself more important for discerning binding sites of TFs from the same TF family, as they have very similar or identical base readout mechanisms. Our method could also be useful for detecting binding sites of miRNAs because structure plays a dominant role in the RNA–RNA interaction [158]. Despite high-throughput experimental approaches to identification of TFBSs, improved *in silico* prediction of TFBSs is of great value. It allows more accurate identification of potential *in vivo* TFBSs on rapidly sequenced genomes and enhances our understanding of the TF binding processes. Our integrative method seems to be a good candidate for this purpose.

3.6 Author contributions

SB and BH both wrote the manuscript. BH mainly wrote the introduction and the material and methods section, SB wrote most of the results and the discussion. SB and BH designed the algorithm and did all the testing. SB designed a draft version of the algorithm during his master thesis and was the main developer of the feature selection steps. During the revision of the article, which was carried out by SB, SB build the final models. SB also did an additional precision-recall analysis for the final paper and did literature research about the relevance of the selected structural features in the models. Both authors are considered joint first authors. FVR supported the research project. PDB initiated and supported the research project.



4 — PhysBinder web-tool

This chapter is a redraft from the publication:

Stefan Broos, Arne Soete, Bart Hooghe, Raymond Moran, Frans Van Roy and Pieter De Bleser **PhysBinder: improving the prediction of transcription factor binding sites by flexible inclusion of biophysical properties**; Nucleic Acids Res. 2013 41: W531–W534

4.1 Abstract

The most important mechanism in the regulation of transcription is the binding of a transcription factor (TF) to a DNA sequence called the transcription factor binding site (TFBS). Most binding sites are short and degenerate, which makes predictions based on their primary sequence alone somewhat unreliable. We present a new web-tool that implements a flexible and extensible algorithm for predicting TFBSs. The algorithm makes use of both direct (the sequence) and several indirect readout features of protein–DNA complexes (biophysical properties such as bendability or the solvent-excluded surface of the DNA). This algorithm significantly outperforms state-of-the-art approaches for *in silico*

identification of transcription factor binding sites. Users can submit FASTA sequences for analysis in the PhysBinder integrative algorithm and choose from more than 60 different TF binding models. The results of this analysis can be used to plan and steer wet-lab experiments. The PhysBinder web-tool is freely available at http://bioit.dmbr.ugent.be/physbinder/index.php

4.2 Introduction

Proteins called transcription factors (TFs) are crucial for proper regulation of gene expression. They function by binding to regions of DNA called transcription factor binding sites (TFBSs). Two different mechanisms contribute to the TF–DNA binding specificity needed for correct regulation of gene expression: a direct readout component caused by direct contact between the amino acids of the protein and the bases of the DNA, and an indirect readout component caused by the global shape of the DNA and by conformational changes in both interaction partners [159, 160]. Traditional methods for predicting TFBSs tend to look at the direct readout component alone, and almost exclusively at the primary sequence. However, many of these widely used methods, such as positional weight matrices (PWMs), are afflicted by many false positive predictions, indicating the need for incorporating other discriminative features [1]. Recent evidence shows that sequence-dependent structural variations in the DNA account for a significant portion of the protein–DNA specificity [161, 45, 116]. Thus, it is expected to be beneficial to include structural features and nucleotide dependencies in the prediction models. In a recent publication we examined the effect of incorporating nucleotide position dependencies, which are related to the 3D structure of the DNA [85], on the prediction of TFBSs [51]. We also calculated structural features of the DNA and verified to which extent these features improve the prediction of TFBSs. We found that incorporation of both types of data can substantially enhance the prediction of TFBSs. Here, we present PhysBinder, a web-tool based on the flexible Random Forest algorithm published in [51]. We compiled more than 60 vertebrate TF models from various sources, but many more models will be offered in the future as new data become available. Binding sites for these models can be visualized together with the ENCODE TFBS data track of UCSC

Genome [162] in order to get a useful insight in the genomic context of the inspected region.

4.3 Input and output

4.3.1 Input

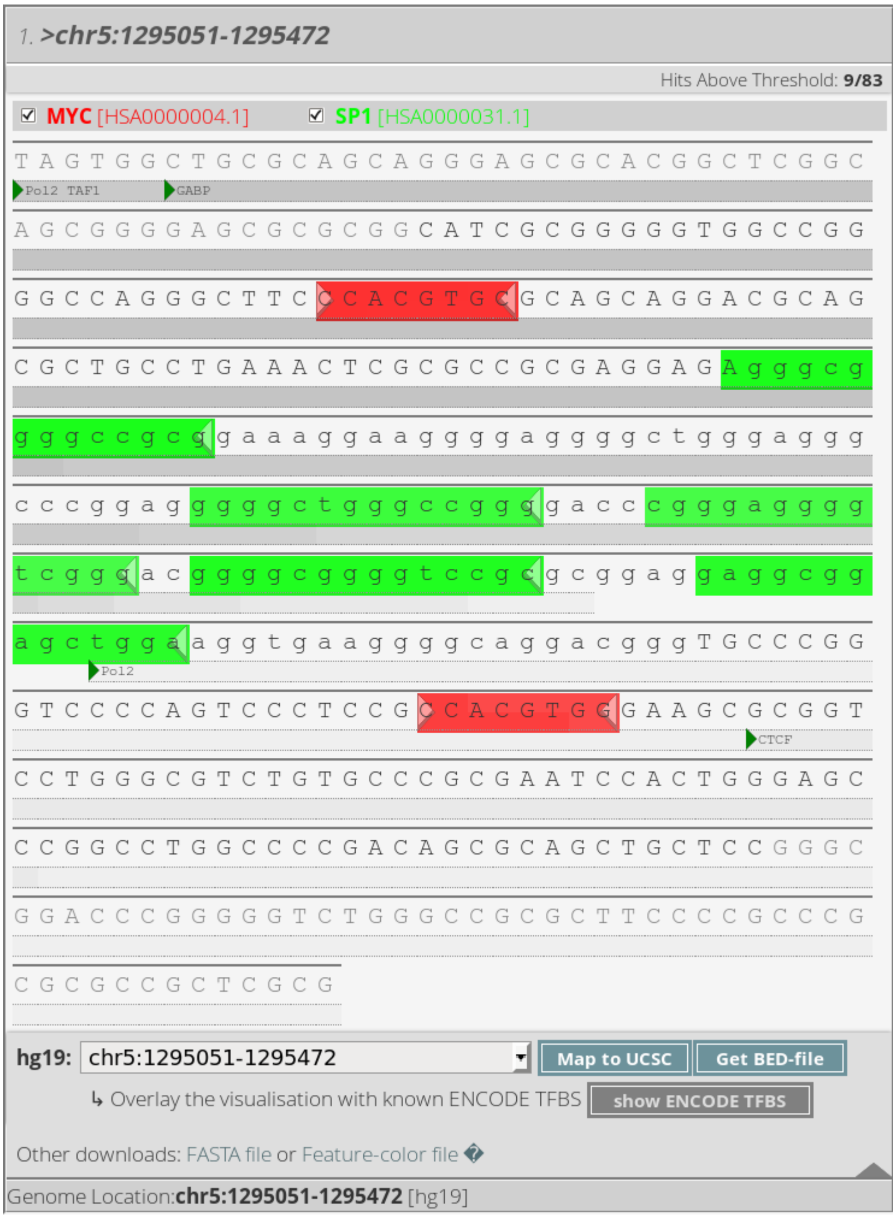
The PhysBinder web-tool is easy to use: for most parameters we offer default configurations to ensure a quick and easy workflow. Users just provide their sequences of interest and select the appropriate TF model information. Sequences can be uploaded by one of the following means: (i) pasting a set of FASTA-formatted sequences in the input field; (ii) uploading a file with FASTA-formatted sequences; (iii) indicating genomic regions in the "Fetch genomic regions" text field. Subsequently a model and a threshold are to be selected. We provide three pre-calculated thresholds: "Max. Precision," "Max. F-Measure," and an average of these two measures. A custom threshold can also be selected. More than 60 different TF models are now available on the PhysBinder website, but we expect to provide more models as additional data become available. Most of the PhysBinder models are compiled from recent ENCODE data [163] but other sources were also used (see 4.4 "Technical Details" for more information). TF models constructed from sequences that, according to the literature, clearly contain a sequence element associated with the TF are called "direct evidence" models (DEs). When an alternative consensus sequence is found or when no consensus sequence is known for a particular TF, we call the models "putative associated factors" (PAFs). Such a PAF might be a TF binding to multiple sequence elements or it might be a common cofactor (hence "putative associated factor"). By default, PhysBinder is configured to run in filter mode to speed up the calculations. In this mode, sequences are pre-filtered with a short PWM with very low thresholds, minimizing the number of false negative hits and effectively guaranteeing maximum recall.

4.3.2 Output

A summary table is given at the top of the results web page. This table can be sorted by model type or by input sequence, and for each model or sequence the number of hits is indicated. On this page, users can still alter the thresholds in order to increase or decrease the stringency of the binding site predictions. In the results section, binding sites are shown as sequences with a colored background (exemplified in figure 4.1). Clicking on the first nucleotide of such a colored sequence provides more details on the binding site. When clicked, a details window with the sequence logo of the binding site is shown (this logo was calculated on the model data) and the Random Forest score with a p-value is given as well. The relative position of the TFBS is shown, and if the genomic location of the sequence is known (because the user indicated this on the input page or performed a BLAT analysis of the sequence against a human or mouse reference genome), then the absolute coordinates of the binding sites are shown in the details window. Two additional options become available when the absolute position is known. For human sequences (hg18 and hg19), it is possible to integrate the most recent ENCODE data to get an overview of the transcription factors and RNA polymerase components that might bind within this genomic region. Predicted binding sites can also be visualized in the UCSC Genome Browser [164] (exemplified in figure 4.2). Using the checkboxes above the sequences or those on the right side of the screen, models can be dynamically shown or hidden to aid the interpretation of the results.

4.3.3 An example...

As an example (see figures 4.1 and 4.2), we examined the analysis performed by Kyo et al. [165] of the promoter of the human TERT gene, encoding the catalytic subunit of telomerase. These researchers identified a core promoter of 181 bp responsible for the transcriptional activity of the TERT gene. This 181-bp region, consisting of the 5’-UTR and the upstream promoter region, contains two E-boxes bound by MYC *in vivo*. Between these E-boxes Kyo et al. discovered and validated five GC-boxes that are bound by SP1. For illustrative purposes, we used the PhysBinder tool to look for SP1, MYC and TBP binding sites with default threshold settings in the same sequence they used [165] and we were readily able to confirm their findings. We unmistakably found the five SP1 binding sites flanked by two MYC binding sites, as reported in [165]. No TATA-box was found, and indeed this promoter was reported to lack such box [166].



1. >chr5:1295051-1295472

Hits Above Threshold: **9/83**

☑ **MYC** [HSA0000004.1] ☑ **SP1** [HSA0000031.1]

```
TAGTGGCTGCGCAGCAGGGAGCGCACGGCTCGGC
AGCGGGGAGCGCGCGGCATCGCGGGGTGGCCGG
GGCCAGGGCTTCCCACGTGCGCAGCAGGACGCAG
CGCTGCCTGAAACTCGCGCCGCGAGGAGAgggcg
gggccgcggaaaggaaggggaggggctgggaggg
cccggaggggctgggccggggacccgggaggggg
tcgggacgggcggggtccgccgcggaggaggcgg
agctggaaggtgaaggggcaggacgggTGCCCGG
GTCCCCAGTCCCTCCGCCACGTGGGAAGCGCGGT
CCTGGGCGTCTGTGCCCGCGAATCCACTGGGAGC
CCGGCCTGGCCCCGACAGCGCAGCTGCTCCGGGC
GGACCCGGGGGTCTGGGCCGCGCTTCCCCGCCCG
CGCGCCGCTCGCG
```

hg19: chr5:1295051-1295472

↳ Overlay the visualisation with known ENCODE TFBS

Other downloads: FASTA file or Feature-color file �

Genome Location:**chr5:1295051-1295472** [hg19]

Figure 4.1: Example output of the PhysBinder tool. All predicted TFBSs match the experimentally determined locations reported by Kyo et al. [165]. Detail of the results window: MYC binding sites (E-box) [HSA0000004.1] are shown in red. SP1 binding sites (GC-box) [HSA0000031.1] are shown in green. The default threshold (“Average”) was used for both models. Grey shaded bars indicate overlapping ENCODE tracks.

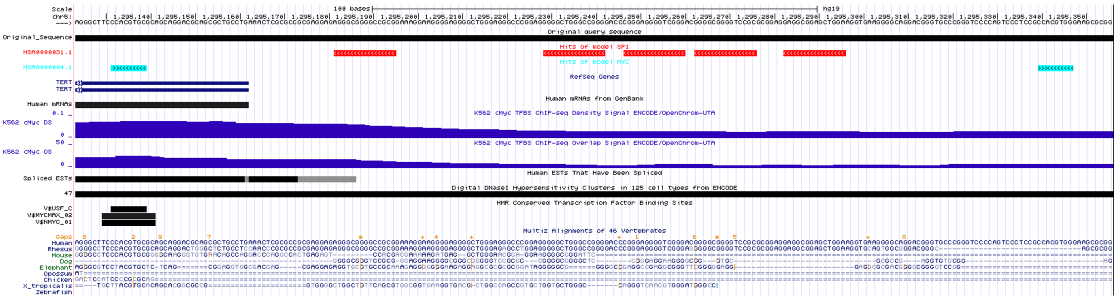


Figure 4.2: In this figure, the models of figure 4.1 are visualized in the UCSC Genome Browser. MYC binding sites are indicated in blue whereas SP1 binding sites are in red.

4.4 Technical details

The web-tool is hosted on a Linux CentOS 5 server with 32 GB of RAM, an Apache 2.2.3 web server, and PHP version 5.1.6. Web pages are written in the PHP and Javascript scripting languages. To map input sequences to mouse (mm10) or human (hg19) reference genomes, we use gfServer and Client binaries from UCSC, which makes it possible to BLAT sequences. ENCODE tracks are obtained from UCSC Genome [162]. Sequences can be fetched from 16 different species, obtained from UCSC Genome. Extensive help documentation is available on the PhysBinder website, including guidelines and tutorials to facilitate the interpretation of the PhysBinder results.

The backend of PhysBinder is programmed in a combination of Perl and R-script. The Random Forest classifier used in the backend is the “FastRandomForest” implementation. This is a multithreaded implementation of the Random Forest classifier in the Weka statistical package [167]. In our models we use a Random Forest with 100 trees. Most models are built from available ENCODE data of tier 1 cell lines, except for Esrrb [168], ETS1 [169], KLF4 [168], NANOG [168], Nmyc [168], STAT3 [168], TBP [103], Tfcp2l1 [168], TP53 [104] and Zfx [168]. All sequences were first aligned using the MEME motif aligner [38] on the STEVIN supercomputing infrastructure of Ghent University. To ensure the quality of input data, the resulting aligned sequence motifs were then manually searched for in the literature. If a motif is not yet reported in literature, the resulting model is called a putative associated factor (PAF). Otherwise, the model is termed a direct evidence model (DE). When available, 100 sequences were used to build the model. The other sequences were used for validation. More information on the different steps of the algorithm and on its validation has been reported by us previously [51]. Details about all models are available on the “models” page, where an overview can be found of all the features contained in the models, together with performance measures that were calculated on external test sets.

4.5 Performance of currently available PhysBinder models

An overview of the performance of the models that are currently implemented in PhysBinder can be found in tables 4.1 and 4.2. The ROC curves and the corresponding AUCs were generated on sequences that were not used to train the models or to calculate thresholds on. Background sequences were randomly drawn from the same reference genome as the positive sequences (ten times the number of positive sequences).

4.6 Author contributions

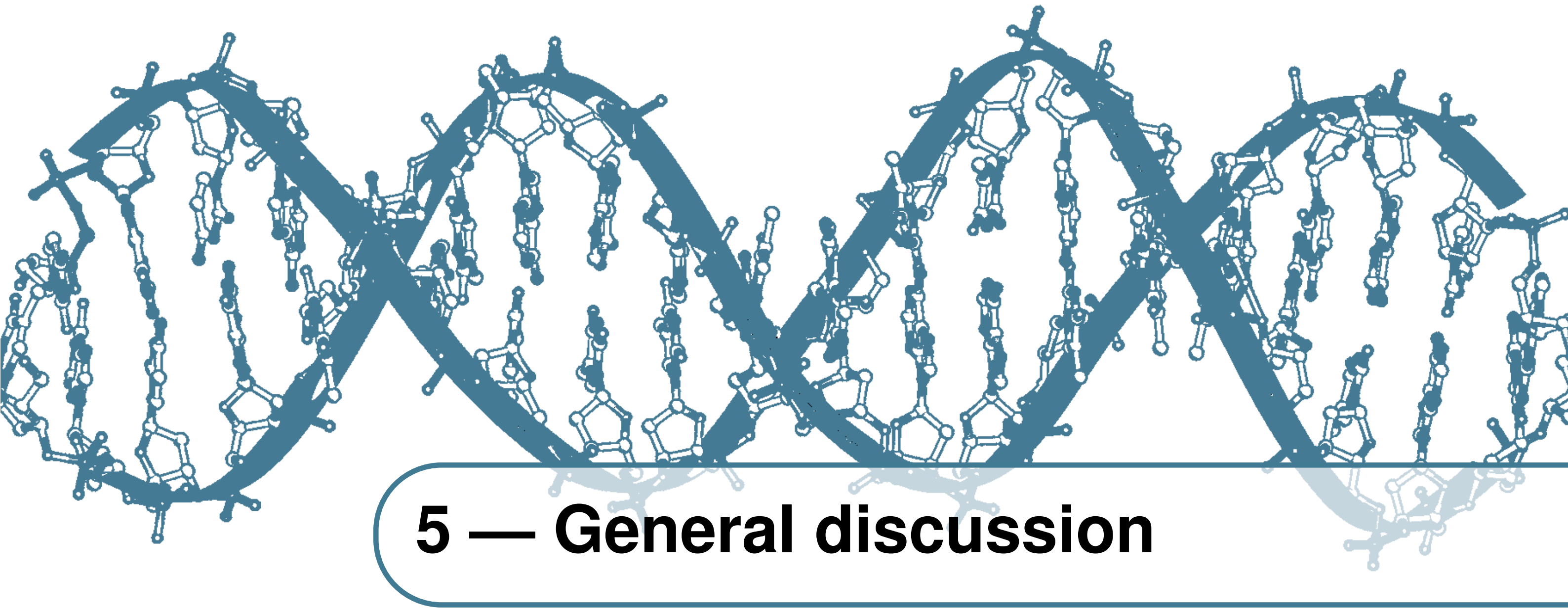
SB wrote the article. SB and AS developed the web-tool (SB focused on algorithmic implementations, AS on visualization). RM helped on PhysBinder during his internship in the bioIT Core. BH helped developing the original algorithm. FVR supported the research. PDB initiated and supported the research and helped testing.

Transcription Factor	ROC AUC pure	ROC AUC filter
ATF1	0.99848	0.98870
BACH1	0.99848	0.96643
BDP1	0.99818	0.99672
BRCA1	0.99798	0.96433
CREB1	0.99832	0.98467
CTCF	0.98408	0.96766
CTCFL	0.97942	0.97207
E2F1	0.95856	0.95856
EGR1	0.99699	0.96870
ELF1	0.99769	0.99473
ELK1	0.99865	1.00000
ELK4	0.96392	0.89158
ESRRA	0.98950	0.98303
ETS1	0.98858	0.96084
Esrrb	0.88453	0.89531
FOS	0.98767	0.96873
GABPA	0.99886	0.99061
GATA1	0.99398	0.99104
GATA2	0.99604	0.87464
GTF3C2	0.90912	0.81600
IRF3	0.94492	0.93765
JUN	0.99882	0.99131
JUNB	0.99869	0.98976
KLF4	0.96488	0.89591
MAFF	0.97770	0.96469
MAFK	0.97661	0.95590
MEF2C	0.95110	0.94209
MYC	0.99988	0.99000
Mycn	0.99869	0.97203
NANOG	0.97755	0.96919
NFYA	0.96805	0.95682
NFYB	0.99422	0.97054
GR	0.98114	0.98114
NRF1	0.99699	0.98030
POLR3A	0.99699	0.69214
POU5F1	0.95894	0.92797
PRDM1	0.98741	0.97699
RAD21	0.99363	0.99397

Table 4.1: ROC AUCs of currently available transcription factor models in PhysBinder. AUCs were calculated on a holdout set that was not used to train the models or to calculate thresholds on. ROC AUC pure: AUC without filtering with a PWM; ROC AUC filter: AUC when filtering with a PWM.

Transcription Factor	ROC AUC pure	ROC AUC filter
RELA	0.88408	0.89852
REST	0.99976	0.97143
REST (2)	0.99992	0.97449
RFX5	0.99991	0.99479
RUNX3	0.97696	0.92339
SIX5	0.99926	0.98838
SP1	0.97144	0.91736
SP2	0.97113	0.89983
SP4	0.99755	0.96804
SPI1	0.95121	0.96108
SREBF1	0.99157	0.76809
SRF	0.99852	0.96890
STAT1	0.98354	0.94069
STAT3	0.98725	0.96886
STAT3 (2)	0.93549	0.89987
TAF1	0.99942	0.94108
TAF7	0.99975	0.99975
TCF12	0.98174	0.95711
TEAD4	0.99730	0.98201
Tfcp2l1	0.76203	0.80366
USF1	0.99475	0.97971
USF2	0.99003	0.97133
YY1	0.95191	0.94495
ZEB1	0.98985	0.95170
ZNF274	0.99859	0.87247
Zfx	0.87386	0.88052

Table 4.2: Continuation of table 4.1. ROC AUCs of currently available transcription factor models in PhysBinder. AUCs were calculated on a holdout set that was not used to train the models or to calculate thresholds on. ROC AUC pure: AUC without filtering with a PWM; ROC AUC filter: AUC when filtering with a PWM.



5 — General discussion

The prediction of transcription factor binding sites is not straightforward. In fact, it is one of the oldest problems that bioinformatics is trying to solve; and it is a problem that has been proven difficult to solve. In the past, many algorithms were designed to address this issue, but few have had such an impact as the positional weight matrix (PWM), originally created by Gary D. Stormo [80]. Although the PWM method is the *de facto* standard nowadays, it is far from perfect. The predictions made by PWMs are plagued by an enormous amount of false positive predictions (the futility theorem). This implies that a simple approach such as the PWM does not capture the entire story of protein-DNA binding specificity. As a result, many researchers were encouraged to implement small improvements to the PWM method to overcome some of these problems [90, 170, 79, 171]. However, the majority of these suggested improvements were not very successful and most of them have found no adoption by the bioinformatics community today due to their complexity.

The focus of this PhD was to develop *in silico* methods that can accurately identify binding sites. During the course of the development two points were considered. First, the methods should provide some enhancements to already existing methods. Secondly, it does not matter how complicated the algorithm behind the method is, as long as the resulting web-tool is intuitive and easy to use. In my opinion, the second point is equally important as the first one. If the use of a method is overly complicated then it will fail to convince users. This means that it is very important to design an attractive interface and to choose good default parameters.

In this discussion I will clarify how binding site specificity is achieved and how I have included this in the methods. Another important point that I will tackle during this discussion is the relevance of *in silico* prediction methods in an era of ChIP-seq and other high-throughput techniques. In the last section of this discussion I compare the use of PhysBinder with that of our ConTra v2 webserver.

5.1 What contributes to binding site specificity?

In order to improve the prediction of binding sites, it is important to get a general idea about the mechanisms that define protein-DNA specificity. In literature typically two different mechanisms are defined that thrive protein-DNA specificity [78]. The first mechanism is called direct base readout and the second mechanism is termed indirect readout. Figure 5.1 illustrates these interactions. In this section I will discuss both mechanisms and their role in our biophysical prediction software. I will also present some potential issues of *in silico* prediction algorithms in *in vivo* systems.

5.1.1 Direct readout

A great deal of specificity between the DNA and protein is achieved from a direct contact between the amino acids and the base pairs of the DNA. This type of contact is also known as the direct readout of protein-DNA recognition. In direct readout the transcription factor senses the base pair composition of a stretch of the DNA by making hydrogen bonds between the amino acids of the protein and through van der Waals forces. A lot of these hydrogen bounds can be formed in the major groove of the DNA where a lot of functional groups reside that can form hydrogen bonds [173]. The nature of the hydrogen bond is very important for direct readout: it is highly relevant for protein-DNA specificity whether there

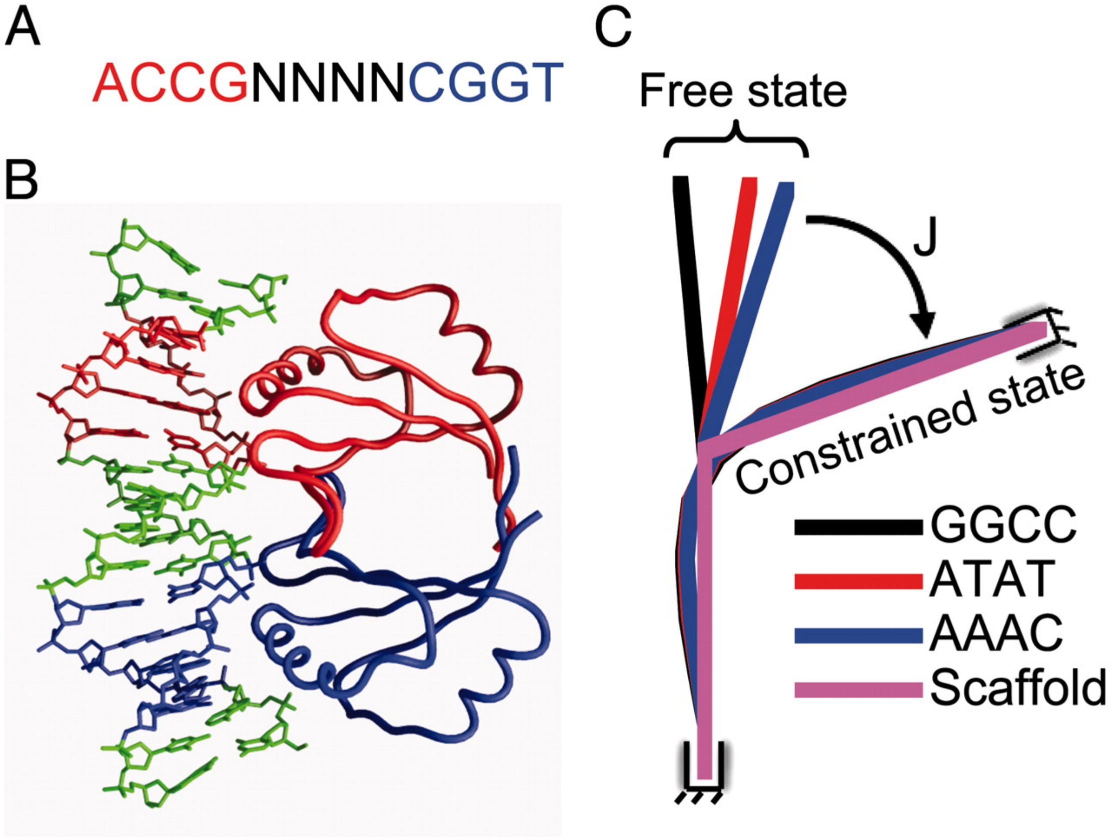


Figure 5.1: Example of the different readout mechanisms, illustrated by a figure of Zhang et al. [172]. (A) Consensus sequence of E2 proteins. (B) Both homodimers insert an alpha helix in the major groove of the E2 protein, contacting the ACCG sequences. By this insertion the protein can directly contact the base pairs of the binding site (direct readout). (C) However, as a result of the insertion of the alpha helices the conformation of the DNA is altered and a bend is introduced in order to get a good fit for both alpha helices (indirect readout). Binding sites with an AT rich spacer tend to have higher affinities for the E2 protein, since these AT rich sequences facilitate bending towards the minor groove.

is one or more hydrogen bonds between the amino acid and the corresponding base [174]. It is also possible that the amino acids form hydrogen bonds with more than one base or that solvent molecules participate in the hydrogen bonds.

van der Waals interactions between the protein and the DNA are possibly even more important than hydrogen bonds. An early analysis by Luscombe pointed out that on average 64.9% of all protein and DNA interactions are made by van der Waals contacts [174].

In our biophysical models we incorporate direct readout parameters in a number of ways: 1) In the models we include a vector of mononucleotides and/or dinucleotides. 2) A vector with amino acids propensities is included in the models. This vector indicates how often a certain amino acid is expected to bind to that part of the sequence (the affinity of the amino acid for a certain subsequence). 3) We include features such as groove width that indicate how accessible the functional groups of the DNA are. By including these different features we aim to cover direct readout effects in our models as much as possible. As a result of the feature selection mechanism a different subset of these features is combined in each model, depending on the nature of the protein-DNA complex. In this way, our models tend to reflect the real situation as much as possible.

5.1.2 Indirect readout

Binding specificities that are not resulting from the direct base readout between the protein and DNA binding site are referred to as indirect readout interactions. This type of specificity is mainly achieved by DNA flexibility and conformational changes in the protein-DNA complex, for example due to the bending of the DNA around the protein. These types of interactions do not necessarily occur at the actual binding site. It is also possible that sequence elements upstream or downstream of the binding site are essential to help with the change in conformation. For example, if the DNA is wrapped around the protein when binding, flexible sites just before and after the binding sites might be of importance.

In our biophysical models we include many features that represent the indirect readout component. For example, we incorporate structural features that give an indication about the curvature, torsion and bendability of the DNA. Furthermore, we also take into account features that concern the free energy change that happens when a protein binds to the DNA as well as features that tell us something about the stability of protein-DNA complexes. As previously mentioned in section 5.1.1, the exact content of each model is different. In some models more emphasis is given to the direct readout features whilst in other models the indirect readout component is more important. Given the dynamic composition of each model, we aim to get an accurate representation of reality.

5.1.3 *In vivo* problems

Although our models are very flexible with respect to feature composition, there are still some aspects of the *in vivo* situation that are difficult to account for. For example, it is possible that the transcription factor itself is not expressed (or in low concentrations) in the system of interest. Also, there is always the possibility of overlapping binding sites. In this way, the formation of a protein-DNA complex can be hindered by the binding of a neighbouring transcription factor. However, our models capture part of these overlapping binding sites since we also take into account up to 50bp of the flanks of each binding sites' start position. Another aspect that is difficult to capture in predictive models is that many transcription factors can compete for the same binding site, depending on changes in external stimuli. This competition between different transcription factors for the same binding site is very important for a precise regulation of many molecular pathways. At the same time, this competition makes it more complicated for *in silico* prediction methods to make realistic predictions. Finally, some epigenetic changes may influence the *in vivo* binding potential of a region as well.

5.2 Why use prediction methods in a ChIP-seq world?

Recent developments in high-throughput methods for the discovery of DNA binding sites have had an enormous influence on the computational field. One could ask if it still necessary to develop and improve *in silico* methods with all these high-throughput methods available? Genome-wide experiments become increasingly cheaper and require much less time to carry out than before. Despite these elements, there are still many reasons why computational prediction methods will remain important in life sciences.

5.2.1 Ease of use

One of the most obvious reasons why computational prediction methods are still important is their ease of use. Oftentimes, an *in silico* analysis is the first step to explore the systems that are regulating a gene of interest, or to validate whether multiple genes are regulated by the same transcription factors. For these research questions, a complete ChIP-seq experiment is often too costly and it will take too much time. With the *in silico* methods, the researcher can very quickly get a rough idea about the regulation of a gene. If the results of this step are satisfying, additional *in vitro* or even *in vivo* experiments can be conducted to confirm the predictions. Since our methods are developed with ease of use in mind, they can play an important role in the first level of screening.

5.2.2 Validate the effect of SNPs

Another important application of *in silico* prediction methods is the validation of SNP effects in binding sites and promoters. Much of the SNP analysis is focused on the coding regions of the genome. Researchers are most interested in whether a SNP may lead to truncations of the amino acid sequence, or whether it leads to changes in amino acids that are important for the structure of the protein. However, recently the focus is shifting towards disease related SNPs that are within regulatory regions. TFBS prediction methods can be used to get an estimate of the protein-DNA affinity change that the SNP introduces. Generally, a PWM-based method is used to predict changes in affinity of a binding site, but as was discussed earlier, the PWM method does not capture the entire story of binding specificity. The PhysBinder algorithm, developed during this PhD thesis, is better suited than the PWM-based method for a number of reasons. Our Random Forest models perform better than PWM models and they are easy to fine-tune, which suggests that the PhysBinder algorithm most likely gives a more accurate representation of binding specificity. Another important advantage of the Random Forest models is that they are very modular in nature. This modular nature enables us to add or remove features according to novel insights.

For example, a number of polymorphisms in the TATAbox are known that can weaken or enhance promoter activity [175]. A subset of these polymorphisms even lead to the creation or the deletion of TATAboxes and these events are often associated with human pathologies (such as hypertension and cancer). These disruptive polymorphisms can also reside outside the actual binding site (in the flanks of the binding sites). Even though these SNPs do not change the direct readout affinity of the TBP protein for the TATAbox, they do alter the binding specificity trough alteration of the indirect readout. One example of such a SNP that changes the indirect readout is a polymorphism in the tracts outside the TATAbox as described in [176]. It is important to note that these indirect readout mutations outside the TATAbox can modify the binding affinity as drastically as mutations in the TATAbox itself. These mutations most likely change the binding stability of the TBP-TATAbox complex as a result of changes in the biophysical properties of the flanking chromatin (e.g. due to changes in the bendability of these flanking regions [176]). Since our Random Forest models can incorporate a diverse set of indirect readout features (such as bendability), inside or outside the binding site, these models are particularly well-suited to analyze these types of polymorphisms. In fact, in our TBP model, the bendability of flanking regions is already one of the most important features for determining TBP binding sites (see chapter 3).

5.2.3 Choking on 10k genomes

Since the cost of a single ChIP-seq experiment has decreased vastly, the number of genomes that becomes available is astonishing. As a result of this recent increase in sequenced genomes it has become impossible, both financially and practically, to look for transcription factor binding sites on a genome-wide scale. In order to unravel the transcriptional regulatory mechanisms in these organisms, one can only rely on *in silico* methods. With more and more genomes being sequenced, the *in silico* methods will become even more important.

5.3 Which approach is best?

5.3.1 PhysBinder or ConTra?

As we have created two tools to predict transcription factor binding sites, one might ask which one we recommend the most. The answer is both. Most of the time we use both tools in complement. Interesting binding sites that are identified with PhysBinder and ConTra can be further analyzed in the wet-lab. We recommend ConTra to look for binding sites in multiple species alignments as PhysBinder currently has no option to look for binding sites in alignments. ConTra is also the tool of choice to generate attractive visualizations. These visualizations can be used in publications to indicate a conserved regulatory mechanism. PhysBinder, on the other hand, makes use of more sophisticated models compared to the PWM models that are used in ConTra. Hence, in a future version of ConTra we will possibly use the PhysBinder models instead of PWMs to scan the multiple sequence alignments. Furthermore, PhysBinder models will be refined with additional relevant features in the future (for example we will also include DNAse I hypersensitivity data). An additional benefit of using PhysBinder is the possibility to overlay results with the ENCODE tracks or to visualize results in the UCSC Genome Browser. The tracks from ENCODE and the UCSC Genome Browser can be very useful to generate additional hypotheses.

5.3.2 Other approaches that use structural characteristics

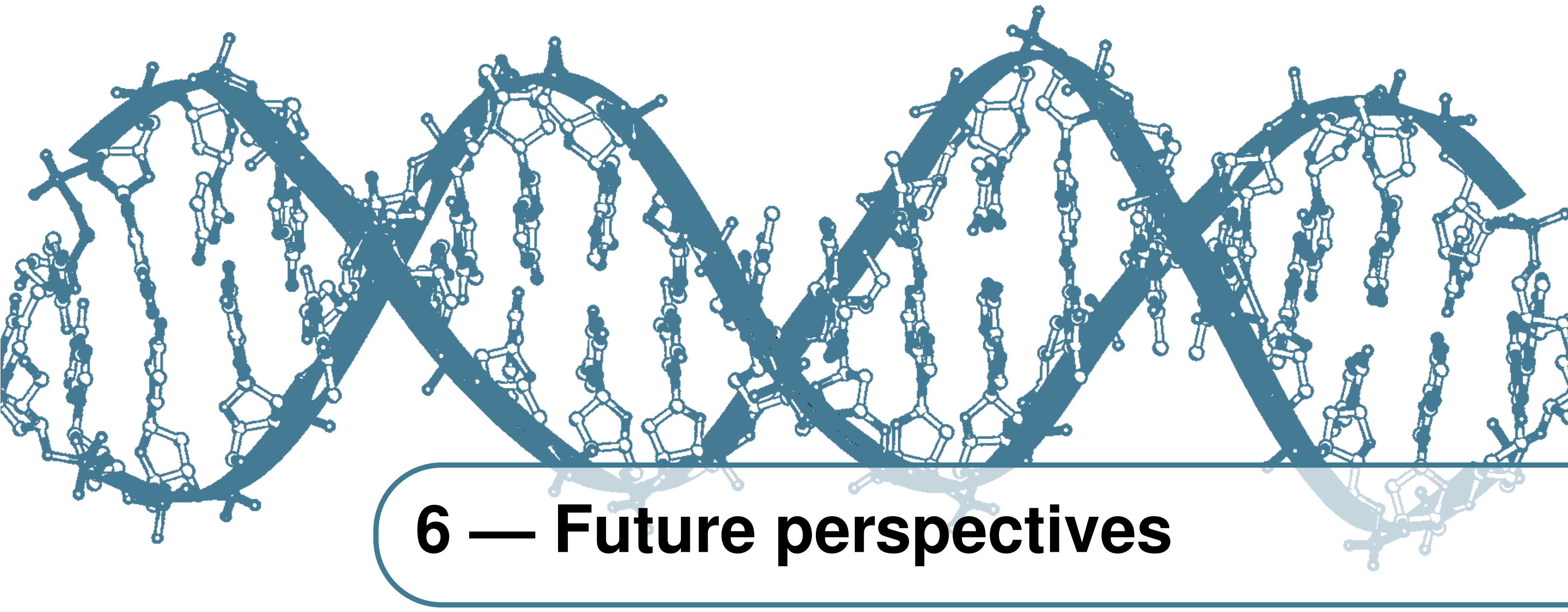
A few other attempts were made to incorporate structural features in order to aid the prediction of transcription factor binding sites.

- Karas et al., 1996 [96] developed a template-based method that was used to predict TATA boxes. This method uses only one structural characteristic at a time (we use multiple characteristics).
- Ponomarenko et al., 1999 [98] expanded the work by Karas et al. They added multiple biophysical parameters (parameters such as roll and twist). They compared themselves with available PWMs and concluded that they achieve similar accuracies as the PWMs.
- Steffen et al., 2002 [159] used a perceptron classifier that uses DNA

deformation energy to make predictions on the IHF transcription factor. However, no comparisons with other methods were made.

- Pudimat et al., 2005 [99] designed a tool called BioBayesNet that also uses the average values of structural properties in a restricted Bayesian network. They compared their method with the PWM for three TFs.
- Burden et al., 2005 [94] converted all JASPAR PWMs to structural profiles (based on base-pair step parameters). However, using the structural profiles together with the PWM did not improve the classification performance.
- Gunewardena et al., 2006 [95] evaluated a template-based approach to model the DNA structure. They made no comparisons and only show that the method can separate sequences from randomly generated sequences.
- Bauer et al., 2010 [93] used molecular dynamic simulations to obtain values for direct and indirect readout. These values are used in a support vector machine algorithm and their method was tested on RegulonDB.
- Meysman et al. 2011 [48] developed a tool called CRoSSeD. CRoSSeD makes use of the conditional random fields algorithm and also of several structural characteristics and the DNA sequence itself. This method is most comparable to ours (and publicly available) and it was also tested in the evaluation of our algorithm. It is mainly specialized in the prediction of prokaryotic transcription factor binding sites.
- Xu et al. 2013 [177] use PDB protein-DNA complex structures to generate an energy function that can be used to predict binding sites. However, they position themselves more as a *de novo* motif detector.

It should be mentioned, however, that most of these tools are not publicly available or are only implemented “in house” to make predictions about one particular transcription factor.



6 — Future perspectives

6.1 Improvements of the Random Forest method

In this section I will suggest possible future improvements to the PhysBinder algorithm. Since all of our Random Forest models are very modular in nature, the addition of new features and the removal of less predictive features is straightforward. Furthermore, there is potential to improve the alignment of the binding sites, as I will discuss later. I will also discuss some extensions to the PhysBinder web-tool that are being developed right now.

6.1.1 Aligning on a structural level

The current design of the Random Forest models requires a set of aligned binding sites with $\pm$ 50bp flanks. These binding sites are aligned on the basis of sequence conservation. In particular, the MEME motif finder is used to align binding sites on a common sequence motif (for more information about this motif finder see 1.6.1). Aligning binding sites based on a common sequence motif makes sense when training a purely sequence-based method. However, in our case, this type of alignment does not give optimal results since our models also act on a structural level. Aligning on the basis of a primary sequence will optimize sequence motifs but might disfavor certain structural or biophysical patterns in the binding sites. Some transcription factor binding sites are most likely not conserved on a sequence level but more on a structure level. Initially it might seem that the motifs of these binding sites have very low information content and that they are of bad quality. However, features other than primary sequence, such as the bendability of the DNA or the torsion of the DNA might be conserved. This is not visible at first, when looking at the sequence, but translating the sequences into biophysical features can reveal certain patterns. These patterns may get lost when we try to align on a sequence level. Based on these observations, it might be a better idea to build a structural motif aligner, roughly based on the MEME algorithm. This motif aligner will aim to optimize the maximum likelihood of structural motifs instead of sequence motifs. Another possible solution to the issues discussed above is the development of a post-processor for alignments. In this case, first a sequence-based motif aligner is used to discover motifs, then refinements are made to this initial motif alignment using a biophysical post-processor. This post-processor tries to optimize the biophysical patterns in the binding sites, aided by an initial sequence alignment. The advantage of this solution is that it will most likely be less resource intensive to optimize alignments when a sequence alignment has already taken place.

6.1.2 Integrating different data sources

As I discussed in section 5.1.3, the *in vivo* situation is much more complex than the *in vitro* situation. The affinity of a transcription factor for a certain DNA binding site can be very high in an isolated system. However, when multiple proteins compete for the same location on the DNA the absolute affinity falls short from telling the entire story. In this situation, it is also important to take into account the protein concentrations of the competing DNA-binders and the number of accessible binding sites. For example, it is possible that a highly interacting transcription factor in low concentration gets outnumbered by a transcription factor with lower affinity with very high concentrations. This means that the low affinity protein will bind to the binding site even though our methods predict that the high affinity transcription factor was much more likely to bind. To account for this effect, we should also include information on protein concentration in our predictions if we want to make accurate *in vivo* predictions. Unfortunately, including this information is not an easy task, as the protein concentration of a transcription factor is often dependent on many conditions. *In vivo* protein concentrations are influenced by the cell type, external stimuli, cell cycle, disease and many other factors. Nevertheless, by including a rough estimate of the concentration we can probably already get a large improvement on the *in vivo* predictions. For example, we can integrate RNA-seq data to estimate the abundance of the different transcription factors.

Other types of data that may be worth considering for the improvement of *in vivo* predictions are data on histone modifications and Dnase I hypersensitivity sites. These types of data hold information on the chromatin state of the DNA and the accessibility of the binding site.

6.2 What is next?

During this PhD thesis we focused on predicting DNA binding sites using novel and innovative methods. However, these methods can also be used to predict other types of interactions. Currently, the Random Forest classifier is very popular, and given the amount of time that was invested in optimizing the algorithms for use on genomic data, we can adapt these algorithms for use in other types of predictions (such as predictions of micro-RNA binding sites or splice sites). In this section I will discuss the use of the Random Forest method in the prediction of micro-RNA binding sites. Then I will also go into the integration of transcription factor binding site predictions with micro-RNA target predictions.

6.2.1 Micro-RNA target prediction

The Random Forest algorithm can easily be adapted to predict different types of interactions. More specifically, the algorithm can be modified to detect micro-RNA binding sites on messenger RNAs. In this case, the biophysical features (that are typical for DNA elements) which are normally included in the models can be substituted by features specific to RNA and RNA complexes. Some examples of RNA features that were found in the DiProDB are RNA hydrophilicity, RNA twist, rise, roll and slide, RNA stacking energy, RNA entropy, enthalpy and free energy [46]. A first proof-of-concept of our algorithm with these features has proven that there is a lot of potential for this approach. Currently, the available micro-RNA target prediction software performs rather poor, and most of the prediction results are unusable. Hence, there is room for well-performing micro-RNA target binding site prediction software to generate high-quality integrated networks.

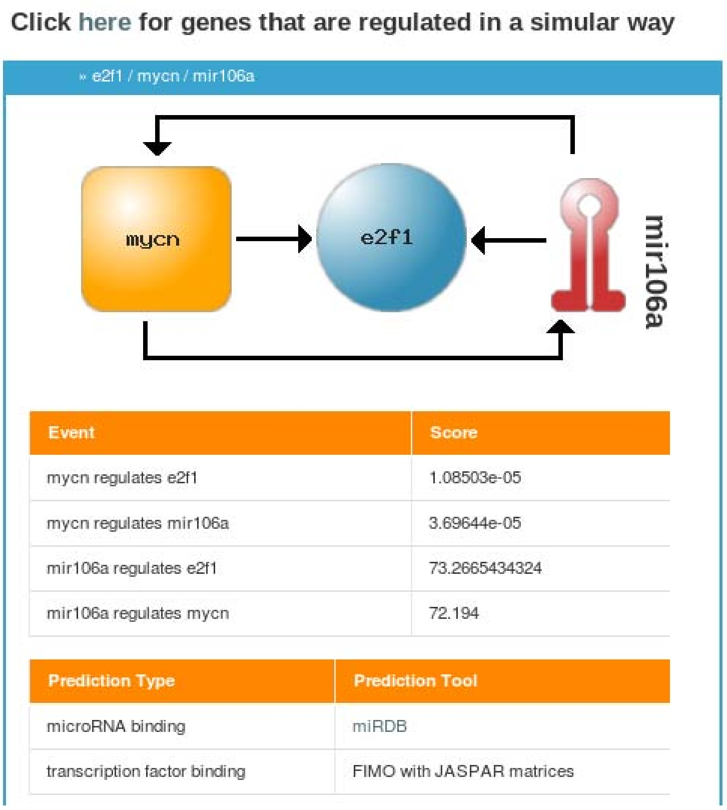
6.2.2 Integration of micro-RNA target prediction with TFBS predictions

The modified RNA version of the biophysical algorithm combined with the present TFBS prediction algorithm could further enhance our knowledge of regulatory networks, both on a translational and post-translational level. This knowledge is of great importance as the interplay between micro-RNAs and transcription factors largely remains unclear. Since both molecules are the primary gene regulators, their combinatorial logic is of great interest and it might give insights in many disease mechanisms. Subsequently, the combination of both types of regulatory information can be used to search for interesting feed-forward and feedback loops. These types of loops are often involved in important biological processes.

During my PhD research (together with Arne Soete) a tool was created that integrates experimentally validated micro-RNA target sites with predicted transcription factor binding sites (see figure 6.1). However, since the set of experimentally validated micro-RNAs target sites is rather limited, it would be interesting to couple the transcription factor binding sites predictions to high confidence micro-RNA target predictions.

6.2.3 PhysBinder explorer

To extend the functionality of the PhysBinder web-tool, we are developing an algorithm that uses PhysBinder predictions to look for overrepresented binding sites in a set of co-regulated genes. These PhysBinder predictions can be based on a set of promoter sequences or a list of RefSeq gene identifiers. The PhysBinder explorer algorithm will then uncover statistically enriched binding sites in



Click here for genes that are regulated in a simular way

Event	Score
mycn regulates e2f1	1.08503e-05
mycn regulates mir106a	3.69644e-05
mir106a regulates e2f1	73.2665434324
mir106a regulates mycn	72.194

Prediction Type	Prediction Tool
microRNA binding	miRDB
transcription factor binding	FIMO with JASPAR matrices

Figure 6.1: Example output of the miRTra prediction tool. This tool currently uses mirDB and PWM predictions from TRANSFAC and JASPAR (together with the fimo motif finder).

the regulatory regions of these genes using the PhysBinder models.

As an example, I have included an output table of the commandline version of PhysBinder explorer (table 6.1). In this example, a set of genes that are known to be regulated by the E2F family of transcription factors is used as input and a set of randomly selected promoter regions is used as background set. PhysBinder explorer correctly identifies the E2F family als primary regulators and even correctly identifies common co-factors and associated transcription factors.

In order not to limit the set of binding sites that can be found with this method

Transcription Factor	Up/Down	Significance
E2F4	+	2.4536e-06
E2F1	+	2.2026e-05
GTF2F1	+	0.0003789
ETS1	+	0.0006374
POU2F2	+	0.0010363
OCT2	+	0.0010363
NFKB	+	0.0017579
NFYA	+	0.0024182
NFYB	+	0.0024182
E2F6	+	0.002917
CHD2	+	0.0031662
HMGN3	+	0.0033353

Table 6.1: Example of PhysBinder explorer output when used on a set of genes that are all regulated by E2F transcription factors. PhysBinder explorer correctly identifies the importance of E2F transcription factors in the promoter regions of these genes. Parameters used in this prediction: Chi-squared test with a p-value cutoff of 0.01; a RefSeq set of co-regulated genes and a background set of randomly selected promoter regions.

(limited by the total number of models available in PhysBinder), we can also choose to first search for motifs in the sequences using a MEME-like motif finder. The motifs found can then be used to build novel PhysBinder models. After this step, the set of promoter sequences is compared to a set of randomly selected promoter sequences in order to validate the significance of the enrichment.

6.2.4 Cluster analysis with PhysBinder models

It might be interesting to investigate how the different transcription factors cluster together based on the features contained in the structural models. Such a cluster analysis can reveal novel transcription factor clusters based on their indirect readout preferences. For example, DNA bending transcription factors will possibly cluster closer together because of the shared extreme values in biophysical bending properties. For this type of analysis, models that have not been subjected to feature selection are preferred because otherwise models can contain a different set of features, which can hinder the cluster method. Both a cluster analysis of the total structural model and of individual characteristic models can provide unexpected insights about novel families based on indirect readout.

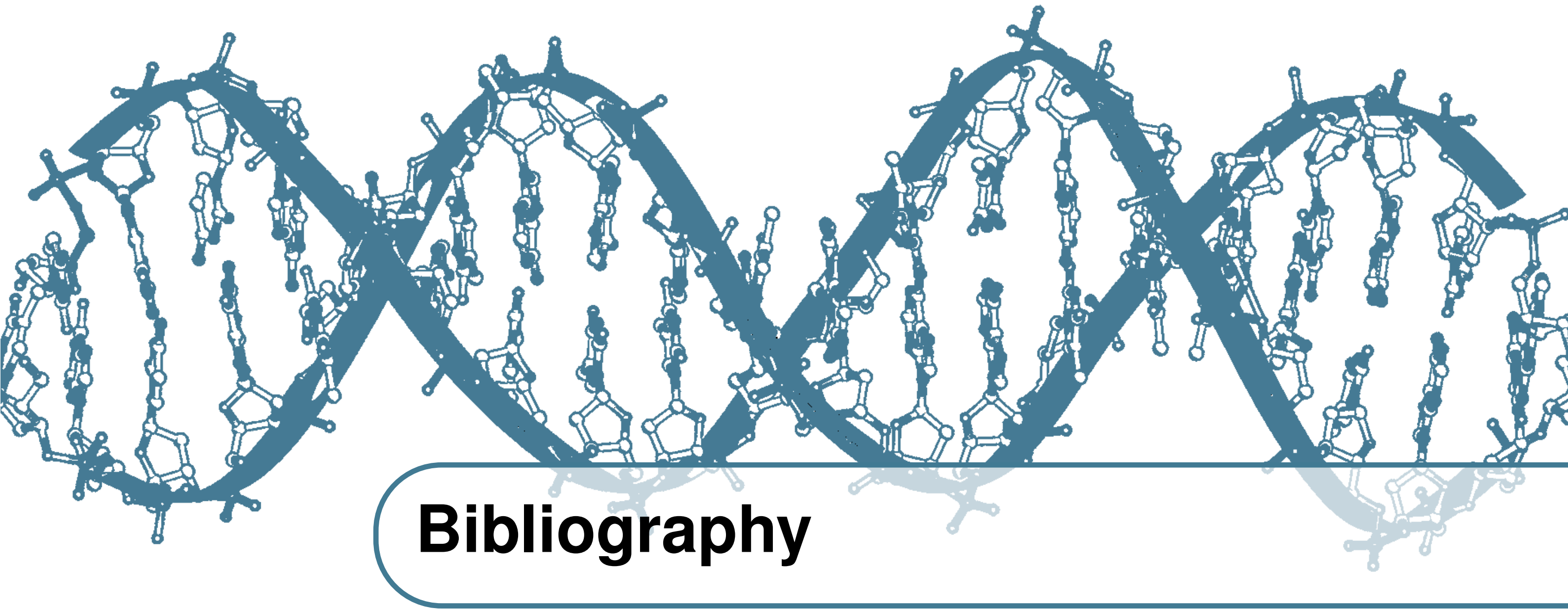
6.2.5 Correlating model score with peak score

As an additional validation of the algorithm we will correlate model scores with ChIP-seq peak scores. A high correlation between both scores is a good indication that the Random Forest score is a good measure of the binding affinity for the transcription factor to the DNA.

6.2.6 Conclusion

Improving functional *in vivo* binding site predictions with *in silico* approaches is a challenging task. During this PhD, I explored the possibilities of the structural and biophysical methods for this type of predictions. I am confident that the methods proposed are of great use in the study of regulatory genomics and that future approaches, based on these findings, will greatly enhance our knowledge about disease and health of human beings. Classifiers that use structural and biophysical characteristics are very suitable candidates for studying these fields. I also learned the importance of high-quality feature selection methods. With the help of these methods, it becomes possible to select the best features for each classification model. This makes the building of models very equivalent to playing with LEGO; you just use the bricks that get the job done.

For the future, I expect that data integration becomes more important than ever before. New data sets are becoming available on a daily basis, thanks to large scale initiatives such as the ENCODE and modENCODE projects. The integration of these data sets into current methods will help us get a better understanding of many *in vivo* pathways. Many researchers fear that this recent explosion of available data sets is too much to handle. In contrast, I regard them as an opportunity if we manage to adapt existing methods and feature selection algorithms. If we play our cards right, these data sets will enable many interesting discoveries.



Bibliography

[1] Wyeth W. Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4):276–287, Apr 2004.

[2] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, Aug 1970.

[3] Steven Henikoff. Beyond the central dogma. *Bioinformatics*, 18(2):223–225, Feb 2002.

[4] H. M. Temin and S. Mizutani. Rna-dependent dna polymerase in virions of rous sarcoma virus. *Nature*, 226(5252):1211–1213, Jun 1970.

[5] D. Baltimore. Rna-dependent dna polymerase in virions of rna tumour viruses. *Nature*, 226(5252):1209–1211, Jun 1970.

[6] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *Nature*, 391(6669):806–811, Feb 1998.

[7] S. Freeman. *Biological science*. Number v. 2. Pearson Prentice Hall, 2005.

[8] G. D. Schuler, M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tomé, A. Aggarwal, E. Bajorek, S. Bentolila, B. B. Birren, A. Butler, A. B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P. J. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, C. Edwards, J. B. Fan, N. Fang, C. Fizames, C. Garrett, L. Green, D. Hadley, M. Harris, P. Harrison, S. Brady, A. Hicks, E. Holloway, L. Hui, S. Hussain, C. Louis-Dit-Sully, J. Ma, A. MacGilvery, C. Mader, A. Maratukulam, T. C. Matise, K. B. McKusick, J. Morissette, A. Mungall, D. Muselet, H. C. Nusbaum, D. C. Page, A. Peck, S. Perkins, M. Piercy, F. Qin, J. Quackenbush, S. Ranby, T. Reif, S. Rozen, C. Sanders, X. She, J. Silva, D. K. Slonim, C. Soderlund, W. L. Sun, P. Tabar, T. Thangarajah, N. Vega-Czarny, D. Vollrath, S. Voyticky, T. Wilmer, X. Wu, M. D. Adams, C. Auffray, N. A. Walter, R. Brandon, A. Dehejia, P. N. Goodfellow, R. Houlgatte, JR Hudson, Jr, S. E. Ide, K. R. Iorio, W. Y. Lee, N. Seki, T. Nagase, K. Ishikawa, N. Nomura, C. Phillips, M. H. Polymeropoulos, M. Sandusky, K. Schmitt, R. Berry, K. Swanson, R. Torres, J. C. Venter, J. M. Sikela, J. S. Beckmann, J. Weissenbach, R. M. Myers, D. R. Cox, M. R. James, D. Bentley, P. Deloukas, E. S. Lander, and T. J. Hudson. A gene map of the human genome. *Science*, 274(5287):540–546, Oct 1996.

[9] F. Antequera and A. Bird. Number of cpg islands and genes in human and mouse. *Proc Natl Acad Sci U S A*, 90(24):11995–11999, Dec 1993.

[10] Bradley E. Bernstein, Ewan Birney, Ian Dunham, Eric D. Green, Chris Gunter, and Michael Snyder. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[11] S. Ohno. So much "junk" dna in our genome. *Brookhaven Symp Biol*, 23:366–370, 1972.

[12] Elizabeth Pennisi. Genomics. encode project writes eulogy for junk dna. *Science*, 337(6099):1159, 1161, Sep 2012.

[13] Sean R Eddy. The c-value paradox, junk dna and encode. *Curr Biol*, 22(21):R898–R899, Nov 2012.

[14] Dan Graur, Yichen Zheng, Nicholas Price, Ricardo B R Azevedo, Rebecca A Zufall, and Eran Elhaik. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of encode. *Genome Biol Evol*, 5(3):578–590, 2013.

[15] Axel Visel, Edward M. Rubin, and Len A. Pennacchio. Genomic views of distant-acting enhancers. *Nature*, 461(7261):199–205, Sep 2009.

[16] G. Orphanides, T. Lagrange, and D. Reinberg. The general transcription factors of rna polymerase ii. *Genes Dev*, 10(21):2657–2683, Nov 1996.

[17] Bradley E. Bernstein, Alexander Meissner, and Eric S. Lander. The mammalian epigenome. *Cell*, 128(4):669–681, Feb 2007.

[18] Kathryn E. Gardner, C David Allis, and Brian D. Strahl. Operating on chromatin, a colorful language where context matters. *J Mol Biol*, 409(1):36–46, May 2011.

[19] Robert F Place, Long-Cheng Li, Deepa Pookot, Emily J Noonan, and Rajvir Dahiya. Microrna-373 induces expression of genes with complementary promoter sequences. *Proc Natl Acad Sci U S A*, 105(5):1608–1613, Feb 2008.

[20] Lin He and Gregory J. Hannon. Micrornas: small rnas with a big role in gene regulation. *Nat Rev Genet*, 5(7):522–531, Jul 2004.

[21] Francesca De Santa, Iros Barozzi, Flore Mietton, Serena Ghisletti, Sara Polletti, Betsabeh Khoramian Tusi, Heiko Muller, Jiannis Ragoussis, Chia-Lin Wei, and Gioacchino Natoli. A large fraction of extragenic rna pol ii transcription sites overlap enhancers. *PLoS Biol*, 8(5):e1000384, May 2010.

[22] Michael Bulger and Mark Groudine. Functional and mechanistic diversity of distal transcription enhancers. *Cell*, 144(3):327–339, Feb 2011.

[23] Ahmad M. Khalil, Mitchell Guttman, Maite Huarte, Manuel Garber, Arjun Raj, Dianali Rivea Morales, Kelly Thomas, Aviva Presser, Bradley E.

Bernstein, Alexander van Oudenaarden, Aviv Regev, Eric S. Lander, and John L. Rinn. Many human large intergenic noncoding rnas associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A*, 106(28):11667–11672, Jul 2009.

[24] Qinghua Cui, Zhenbao Yu, Youlian Pan, Enrico O Purisima, and Edwin Wang. Micrornas preferentially target the genes with high transcriptional regulation complexity. *Biochem Biophys Res Commun*, 352(3):733–738, Jan 2007.

[25] Jacek Krol, Inga Loedige, and Witold Filipowicz. The widespread regulation of microrna biogenesis, function and decay. *Nat Rev Genet*, 11(9):597–610, Sep 2010.

[26] Reut Shalgi, Ran Brosh, Moshe Oren, Yitzhak Pilpel, and Varda Rotter. Coupling transcriptional and post-transcriptional mirna regulation in the control of cell fate. *Aging (Albany NY)*, 1(9):762–770, Sep 2009.

[27] S. Chakravarty and R. Varadarajan. Elucidation of determinants of protein stability through genome sequence analysis. *FEBS Lett*, 470(1):65–69, Mar 2000.

[28] M. Fried and D. M. Crothers. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res*, 9(23):6505–6525, Dec 1981.

[29] M. M. Garner and A. Revzin. A gel electrophoresis method for quantifying the binding of proteins to specific dna regions: application to components of the escherichia coli lactose operon regulatory system. *Nucleic Acids Res*, 9(13):3047–3060, Jul 1981.

[30] Cong Zhu, Kelsey J R P Byers, Rachel Patton McCord, Zhenwei Shi, Michael F Berger, Daniel E Newburger, Katrina Saulrieta, Zachary Smith, Mita V Shah, Mathangi Radhakrishnan, Anthony A Philippakis, Yanhui Hu, Federico De Masi, Marcin Pacek, Andreas Rolfs, Tal Murthy, Joshua

Labaer, and Martha L Bulyk. High-resolution dna-binding specificity analysis of yeast transcription factors. *Genome Res*, 19(4):556–566, Apr 2009.

[31] Christopher J Viggiani, Jennifer G Aparicio, and Oscar M Aparicio. Chip-chip to analyze the binding of replication proteins to chromatin using oligonucleotide dna microarrays. *Methods Mol Biol*, 521:255–278, 2009.

[32] Joshua W K. Ho, Eric Bishop, Peter V. Karchenko, Nicolas Nègre, Kevin P. White, and Peter J. Park. Chip-chip versus chip-seq: lessons for experimental design and data analysis. *BMC Genomics*, 12:134, 2011.

[33] Peter J Park. Chip-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–680, Oct 2009.

[34] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–D94, Jan 2004.

[35] V. Matys, E. Fricke, R. Geffers, E. Gössling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31(1):374–378, Jan 2003.

[36] O. G. Berg and P. H. von Hippel. Selection of dna binding sites by regulatory proteins. statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, 193(4):723–750, Feb 1987.

[37] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, Oct 1993.

[38] Timothy L. Bailey, Nadya Williams, Chris Misleh, and Wilfred W. Li. Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic Acids Res*, 34(Web Server issue):W369–W373, Jul 2006.

[39] Gabriela G Loots, Ivan Ovcharenko, Lior Pachter, Inna Dubchak, and Edward M Rubin. rvista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res*, 12(5):832–839, May 2002.

[40] Nick Bray, Inna Dubchak, and Lior Pachter. Avid: A global alignment program. *Genome Res*, 13(1):97–102, Jan 2003.

[41] Nameeta Shah, Olivier Couronne, Len A Pennacchio, Michael Brudno, Serafim Batzoglou, E. Wes Bethel, Edward M Rubin, Bernd Hamann, and Inna Dubchak. Phylo-vista: interactive visualization of multiple dna sequence alignments. *Bioinformatics*, 20(5):636–643, Mar 2004.

[42] Albin Sandelin, Wyeth W Wasserman, and Boris Lenhard. Consite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*, 32(Web Server issue):W249–W252, Jul 2004.

[43] Elodie Portales-Casamar, David Arenillas, Jonathan Lim, Magdalena I Swanson, Steven Jiang, Anthony McCallum, Stefan Kirov, and Wyeth W Wasserman. The pazar database of gene regulatory information coupled to the orca toolkit for the study of regulatory sequences. *Nucleic Acids Res*, 37(Database issue):D54–D60, Jan 2009.

[44] Isabelle da Piedade, Man-Hung Eric Tang, and Olivier Elemento. Dispare: Discriminative pattern refinement for position weight matrices. *BMC Bioinformatics*, 10:388, 2009.

[45] Stephen C J. Parker, Loren Hansen, Hatice Ozel Abaan, Thomas D. Tullius, and Elliott H. Margulies. Local dna topography correlates with functional noncoding regions of the human genome. *Science*, 324(5925):389–392, Apr 2009.

[46] Maik Friedel, Swetlana Nikolajewa, Jürgen Sühnel, and Thomas Wilhelm. Diprodb: a database for dinucleotide properties. *Nucleic Acids Res*, 37(Database issue):D37–D40, Jan 2009.

[47] M. Michael Gromiha, Jörg G. Siebers, Samuel Selvaraj, Hidetoshi Kono, and Akinori Sarai. Intermolecular and intramolecular readout mechanisms in protein-dna recognition. *J Mol Biol*, 337(2):285–294, Mar 2004.

[48] Pieter Meysman, Thanh Hai Dang, Kris Laukens, Riet De Smet, Yan Wu, Kathleen Marchal, and Kristof Engelen. Use of structural dna properties for the prediction of transcription-factor binding sites in escherichia coli. *Nucleic Acids Res*, 39(2):e6, Jan 2011.

[49] W. Ford Doolittle. Is junk dna bunk? a critique of encode. *Proc Natl Acad Sci U S A*, 110(14):5294–5300, Apr 2013.

[50] modENCODE Consortium, Sushmita Roy, Jason Ernst, Peter V Kharchenko, Pouya Kheradpour, Nicolas Negre, Matthew L Eaton, Jane M Landolin, Christopher A Bristow, Lijia Ma, Michael F Lin, Stefan Washietl, Bradley I Arshinoff, Ferhat Ay, Patrick E Meyer, Nicolas Robine, Nicole L Washington, Luisa Di Stefano, Eugene Berezikov, Christopher D Brown, Rogerio Candeias, Joseph W Carlson, Adrian Carr, Irwin Jungreis, Daniel Marbach, Rachel Sealfon, Michael Y Tolstorukov, Sebastian Will, Artyom A Alekseyenko, Carlo Artieri, Benjamin W Booth, Angela N Brooks, Qi Dai, Carrie A Davis, Michael O Duff, Xin Feng, Andrey A Gorchakov, Tingting Gu, Jorja G Henikoff, Philipp Kapranov, Renhua Li, Heather K MacAlpine, John Malone, Aki Minoda, Jared Nordman, Katsutomo Okamura, Marc Perry, Sara K Powell, Nicole C Riddle, Akiko Sakai, Anastasia Samsonova, Jeremy E Sandler, Yuri B Schwartz, Noa Sher, Rebecca Spokony, David Sturgill, Marijke van Baren, Kenneth H Wan, Li Yang, Charles Yu, Elise Feingold, Peter Good, Mark Guyer, Rebecca Lowdon, Kami Ahmad, Justen Andrews, Bonnie Berger, Steven E Brenner, Michael R Brent, Lucy Cherbas, Sarah C R Elgin, Thomas R Gingeras, Robert Grossman, Roger A Hoskins, Thomas C Kaufman, William Kent, Mitzi I Kuroda, Terry Orr-Weaver, Norbert Perrimon, Vincenzo Pirrotta, James W Posakony, Bing Ren, Steven

Russell, Peter Cherbas, Brenton R Graveley, Suzanna Lewis, Gos Micklem, Brian Oliver, Peter J Park, Susan E Celniker, Steven Henikoff, Gary H Karpen, Eric C Lai, David M MacAlpine, Lincoln D Stein, Kevin P White, and Manolis Kellis. Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330(6012):1787–1797, Dec 2010.

[51] Bart Hooghe, Stefan Broos, Frans van Roy, and Pieter De Bleser. A flexible integrative approach based on random forest improves prediction of transcription factor binding sites. *Nucleic Acids Res*, 40(14):e106, Aug 2012.

[52] Oliver Hobert. Gene regulation by transcription factors and micrornas. *Science*, 319(5871):1785–1786, Mar 2008.

[53] Robin C. Friedman, Kyle Kai-How Farh, Christopher B. Burge, and David P. Bartel. Most mammalian mrnas are conserved targets of micrornas. *Genome Res*, 19(1):92–105, Jan 2009.

[54] Raja Jothi, Suresh Cuddapah, Artem Barski, Kairong Cui, and Keji Zhao. Genome-wide identification of in vivo protein-dna binding sites from chip-seq data. *Nucleic Acids Res*, 36(16):5221–5231, Sep 2008.

[55] Michal Mokry, Pantelis Hatzis, Ewart de Bruijn, Jan Koster, Rogier Versteeg, Jurian Schuijers, Marc van de Wetering, Victor Guryev, Hans Clevers, and Edwin Cuppen. Efficient double fragmentation chip-seq provides nucleotide resolution protein-dna binding profiles. *PLoS One*, 5(11):e15092, 2010.

[56] Seth Frietze, Xun Lan, Victor X. Jin, and Peggy J. Farnham. Genomic targets of the krab and scan domain-containing zinc finger protein 263. *J Biol Chem*, 285(2):1393–1403, Jan 2010.

[57] Angeliki Magklara and Catharine L. Smith. A composite intronic element directs dynamic binding of the progesterone receptor and gata-2. *Mol Endocrinol*, 23(1):61–73, Jan 2009.

[58] Huilin Jin, Rob J. van't Hof, Omar M E. Albagha, and Stuart H. Ralston. Promoter and intron 1 polymorphisms of col1a1 interact to regulate transcription and susceptibility to osteoporosis. *Hum Mol Genet*, 18(15):2729–2738, Aug 2009.

[59] Peggy J. Farnham. Insights from genomic profiling of transcription factors. *Nat Rev Genet*, 10(9):605–616, Sep 2009.

[60] Kouichi Kimura, Ai Wakamatsu, Yutaka Suzuki, Toshio Ota, Tetsuo Nishikawa, Riu Yamashita, Jun-ichi Yamamoto, Mitsuo Sekine, Katsuki Tsuritani, Hiroyuki Wakaguri, Shizuko Ishii, Tomoyasu Sugiyama, Kaoru Saito, Yuko Isono, Ryotaro Irie, Norihiro Kushida, Takahiro Yoneyama, Rie Otsuka, Katsuhiro Kanda, Takahide Yokoi, Hiroshi Kondo, Masako Wagatsuma, Katsuji Murakawa, Shinichi Ishida, Tadashi Ishibashi, Asako Takahashi-Fujii, Tomoo Tanase, Keiichi Nagai, Hisashi Kikuchi, Kenta Nakai, Takao Isogai, and Sumio Sugano. Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res*, 16(1):55–65, Jan 2006.

[61] Bart Hooghe, Paco Hulpiau, Frans van Roy, and Pieter De Bleser. Contra: a promoter alignment analysis tool for identification of transcription factor binding sites across species. *Nucleic Acids Res*, 36(Web Server issue):W128–W132, Jul 2008.

[62] Simon Cawley, Stefan Bekiranov, Huck H. Ng, Philipp Kapranov, Edward A. Sekinger, Dione Kampa, Antonio Piccolboni, Victor Sementchenko, Jill Cheng, Alan J. Williams, Raymond Wheeler, Brant Wong, Jorg Drenkow, Mark Yamanaka, Sandeep Patel, Shane Brubaker, Hari Tammana, Gregg Helt, Kevin Struhl, and Thomas R. Gingeras. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. *Cell*, 116(4):499–509, Feb 2004.

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

[64] Elodie Portales-Casamar, Supat Thongjuea, Andrew T. Kwon, David Arenillas, Xiaobei Zhao, Eivind Valen, Dimas Yusuf, Boris Lenhard, Wyeth W. Wasserman, and Albin Sandelin. Jaspar 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 38(Database issue):D105–D110, Jan 2010.

[65] Xiaohui Xie, Jun Lu, E. J. Kulbokas, Todd R. Golub, Vamsi Mootha, Kerstin Lindblad-Toh, Eric S. Lander, and Manolis Kellis. Systematic discovery of regulatory motifs in human promoters and 3' utrs by comparison of several mammals. *Nature*, 434(7031):338–345, Mar 2005.

[66] Michael F. Berger, Gwenael Badis, Andrew R. Gehrke, Shaheynoor Talukder, Anthony A. Philippakis, Lourdes Peña-Castillo, Trevis M. Alleyne, Sanie Mnaimneh, Olga B. Botvinnik, Esther T. Chan, Faiqua Khalid, Wen Zhang, Daniel Newburger, Savina A. Jaeger, Quaid D. Morris, Martha L. Bulyk, and Timothy R. Hughes. Variation in homeodomain dna binding revealed by high-resolution analysis of sequence preferences. *Cell*, 133(7):1266–1276, Jun 2008.

[67] Naifang Su, Yufu Wang, Minping Qian, and Minghua Deng. Combinatorial regulation of transcription factors and micrornas. *BMC Syst Biol*, 4:150, 2010.

[68] Taro Fukao, Yoko Fukuda, Kotaro Kiga, Jafar Sharif, Kimihiro Hino, Yutaka Enomoto, Aya Kawamura, Kaito Nakamura, Tsutomu Takeuchi, and Masanobu Tanabe. An evolutionarily conserved mechanism for microrna-223 expression revealed by microrna gene profiling. *Cell*, 129(3):617–631, May 2007.

[69] Zhi-Gang Jin, Li Liu, Hua Zhong, Ke-Jing Zhang, Yong-Feng Chen, Wei Bian, Le-Ping Cheng, and Nai-He Jing. Second intron of mouse nestin gene directs its expression in pluripotent embryonic carcinoma cells through pou factor binding site. *Acta Biochim Biophys Sin (Shanghai)*, 38(3):207–212, Mar 2006.

[70] Huajun Yan, Jude Canon, and Utpal Banerjee. A transcriptional chain linking eye specification to terminal determination of cone cells in the drosophila eye. *Dev Biol*, 263(2):323–329, Nov 2003.

[71] Barbara Heise, Julia van der Felden, Sandra Kern, Mario Malcher, Stefan Brückner, and Hans-Ulrich Mösch. The tea transcription factor tec1 confers promoter-specific gene regulation by ste12-dependent and -independent mechanisms. *Eukaryot Cell*, 9(4):514–531, Apr 2010.

[72] Mathieu Blanchette, W James Kent, Cathy Riemer, Laura Elnitski, Arian F A. Smit, Krishna M. Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D. Green, David Haussler, and Webb Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, 14(4):708–715, Apr 2004.

[73] Guillaume Paillard and Richard Lavery. Analyzing protein-dna recognition mechanisms. *Structure*, 12(1):113–122, Jan 2004.

[74] Tommy Kaplan, Nir Friedman, and Hanah Margalit. Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput Biol*, 1(1):e1, Jun 2005.

[75] Kelly M. Thayer and D. L. Beveridge. Hidden markov models from molecular dynamics simulations on dna. *Proc Natl Acad Sci U S A*, 99(13):8642–8647, Jun 2002.

[76] Z. Shakked and D. Rabinovich. The effect of the base sequence on the fine structure of the dna double helix. *Prog Biophys Mol Biol*, 47(3):159–195, 1986.

[77] C. R. Calladine and H. R. Drew. Principles of sequence-dependent flexure of dna. *J Mol Biol*, 192(4):907–918, Dec 1986.

[78] Remo Rohs, Xiangshu Jin, Sean M. West, Rohit Joshi, Barry Honig, and Richard S. Mann. Origins of specificity in protein-dna recognition. *Annu Rev Biochem*, 79:233–269, 2010.

[79] Vladimir Espinosa Angarica, Abel González Pérez, Ana T. Vasconcelos, Julio Collado-Vides, and Bruno Contreras-Moreira. Prediction of tf target sites based on atomistic models of protein-dna complexes. *BMC Bioinformatics*, 9:436, 2008.

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

[81] Jiajian Liu and Gary D. Stormo. Context-dependent dna recognition code for c2h2 zinc-finger transcription factors. *Bioinformatics*, 24(17):1850–1857, Sep 2008.

[82] Martha L. Bulyk, Philip L F. Johnson, and George M. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5):1255–1261, Mar 2002.

[83] Jiajian Liu and Gary D. Stormo. Quantitative analysis of egr proteins binding to dna: assessing additivity in both the binding site and the protein. *BMC Bioinformatics*, 6:176, 2005.

[84] T. K. Man and G. D. Stormo. Non-independence of mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (qumfra) assay. *Nucleic Acids Res*, 29(12):2471–2478, Jun 2001.

[85] Andrija Tomovic and Edward J. Oakeley. Position dependencies in transcription factor binding sites. *Bioinformatics*, 23(8):933–941, Apr 2007.

[86] Panayiotis V. Benos, Martha L. Bulyk, and Gary D. Stormo. Additivity in protein-dna interactions: how good an approximation is it? *Nucleic Acids Res*, 30(20):4442–4451, Oct 2002.

[87] R. A. O'Flanagan, G. Paillard, R. Lavery, and A. M. Sengupta. Non-additivity in protein-dna binding. *Bioinformatics*, 21(10):2254–2263, May 2005.

[88] Ming Hu, Jindan Yu, Jeremy M G. Taylor, Arul M. Chinnaiyan, and Zhaohui S. Qin. On the detection and refinement of transcription factor binding sites using chip-seq data. *Nucleic Acids Res*, 38(7):2154–2167, Apr 2010.

[89] Eilon Sharon, Shai Lubliner, and Eran Segal. A feature-based approach to modeling protein-dna interactions. *PLoS Comput Biol*, 4(8):e1000154, 2008.

[90] Naum I. Gershenzon, Gary D. Stormo, and Ilya P. Ioshikhes. Computational technique for improvement of the position-weight matrices for the dna/protein binding sites. *Nucleic Acids Res*, 33(7):2290–2301, 2005.

[91] Voichita D. Marinescu, Isaac S. Kohane, and Alberto Riva. Mapper: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics*, 6:79, 2005.

[92] Brian T. Naughton, Eugene Fratkin, Serafim Batzoglou, and Douglas L. Brutlag. A graph-based motif detection algorithm models complex nucleotide dependencies in transcription factor binding sites. *Nucleic Acids Res*, 34(20):5730–5739, 2006.

[93] Amy L. Bauer, William S. Hlavacek, Pat J. Unkefer, and Fangping Mu. Using sequence-specific chemical and structural properties of dna to predict transcription factor binding sites. *PLoS Comput Biol*, 6(11):e1001007, 2010.

[94] Heather E. Burden and Zhiping Weng. Identification of conserved structural features at sequentially degenerate locations in transcription factor binding sites. *Genome Inform*, 16(1):49–58, 2005.

[95] Sumedha Gunewardena, Peter Jeavons, and Zhaolei Zhang. Enhancing the prediction of transcription factor binding sites by incorporating structural

properties and nucleotide covariations. *J Comput Biol*, 13(4):929–945, May 2006.

[96] H. Karas, R. Knüppel, W. Schulz, H. Sklenar, and E. Wingender. Combining structural analysis of dna with search routines for the detection of transcription regulatory elements. *Comput Appl Biosci*, 12(5):441–446, Oct 1996.

[97] R. Liu, T. W. Blackwell, and D. J. States. Conformational model for binding site recognition by the e.coli metj transcription factor. *Bioinformatics*, 17(7):622–633, Jul 2001.

[98] J. V. Ponomarenko, M. P. Ponomarenko, A. S. Frolov, D. G. Vorobyev, G. C. Overton, and N. A. Kolchanov. Conformational and physicochemical dna features specific for transcription factor binding sites. *Bioinformatics*, 15(7-8):654–668, 1999.

[99] Rainer Pudimat, Ernst-Günter Schukat-Talamazzini, and Rolf Backofen. A multiple-feature framework for modelling and predicting transcription factor binding sites. *Bioinformatics*, 21(14):3082–3088, Jul 2005.

[100] Alexandre V. Morozov and Eric D. Siggia. Connecting protein structure with predictions of regulatory sites. *Proc Natl Acad Sci U S A*, 104(17):7068–7073, Apr 2007.

[101] Debra L. Fulton, Saravanan Sundararajan, Gwenael Badis, Timothy R. Hughes, Wyeth W. Wasserman, Jared C. Roach, and Rob Sladek. Tfcat: the curated catalog of mouse and human transcription factors. *Genome Biol*, 10(3):R29, 2009.

[102] Byung-Kwan Cho, Eric M. Knight, Christian L. Barrett, and Bernhard Ø. Palsson. Genome-wide analysis of fis binding in escherichia coli indicates a causative role for a-/at-tracts. *Genome Res*, 18(6):900–910, Jun 2008.

[103] Elodie Portales-Casamar, Stefan Kirov, Jonathan Lim, Stuart Lithwick, Magdalena I. Swanson, Amy Ticoll, Jay Snoddy, and Wyeth W. Wasserman.

Pazar: a framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biol*, 8(10):R207, 2007.

[104] Sivakumar Gowrisankar and Anil G. Jegga. Regression based predictor for p53 transactivation. *BMC Bioinformatics*, 10:215, 2009.

[105] A. E. Kel, E. Gössling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender. Match: A tool for searching transcription factor binding sites in dna sequences. *Nucleic Acids Res*, 31(13):3576–3579, Jul 2003.

[106] W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock, and V. B. Zhurkin. Dna sequence-dependent deformability deduced from protein-dna crystal complexes. *Proc Natl Acad Sci U S A*, 95(19):11163–11168, Sep 1998.

[107] D. S. Goodsell and R. E. Dickerson. Bending and curvature calculations in b-dna. *Nucleic Acids Res*, 22(24):5497–5503, Dec 1994.

[108] S. C. Satchwell, H. R. Drew, and A. A. Travers. Sequence periodicities in chicken nucleosome core dna. *J Mol Biol*, 191(4):659–675, Oct 1986.

[109] Xiang-Jun Lu and Wilma K. Olson. 3dna: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat Protoc*, 3(7):1213–1227, 2008.

[110] Richard Lavery, Krystyna Zakrzewska, David Beveridge, Thomas C. Bishop, David A. Case, Thomas Cheatham, 3rd, Surjit Dixit, B. Jayaram, Filip Lankas, Charles Laughton, John H. Maddocks, Alexis Michon, Roman Osman, Modesto Orozco, Alberto Perez, Tanya Singh, Nada Spackova, and Jiri Sponer. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in b-dna. *Nucleic Acids Res*, 38(1):299–313, Jan 2010.

[111] Satoshi Fujii, Hidetoshi Kono, Shigeori Takenaka, Nobuhiro Go, and Akinori Sarai. Sequence-dependent dna deformability studied using molecular dynamics simulations. *Nucleic Acids Res*, 35(18):6063–6074, 2007.

[112] M. R. Gartenberg and D. M. Crothers. Dna sequence determinants of cap-induced bending and protein binding affinity. *Nature*, 333(6176):824–829, Jun 1988.

[113] J. D. Parvin, R. J. McCormick, P. A. Sharp, and D. E. Fisher. Pre-bending of a promoter sequence enhances affinity for the tata-binding factor. *Nature*, 373(6516):724–727, Feb 1995.

[114] R. E. Dickerson. Dna bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res*, 26(8):1906–1926, Apr 1998.

[115] A. A. Gorin, V. B. Zhurkin, and W. K. Olson. B-dna twisting correlates with base-pair morphology. *J Mol Biol*, 247(1):34–48, Mar 1995.

[116] Remo Rohs, Sean M. West, Alona Sosinsky, Peng Liu, Richard S. Mann, and Barry Honig. The role of dna shape in protein-dna recognition. *Nature*, 461(7268):1248–1253, Oct 2009.

[117] Daniel Svozil, Jan Kalina, Marek Omelka, and Bohdan Schneider. Dna conformations and their sequence preferences. *Nucleic Acids Res*, 36(11):3690–3706, Jun 2008.

[118] X. J. Lu, Z. Shakked, and W. K. Olson. A-form conformational motifs in ligand-bound dna structures. *J Mol Biol*, 300(4):819–840, Jul 2000.

[119] R. S. Spolar and MT Record, Jr. Coupling of local folding to site-specific binding of proteins to dna. *Science*, 263(5148):777–784, Feb 1994.

[120] E. Schapire Leo Breiman. Random forests. *Machine Learning*, 45:28, 2001.

[121] Heather J. Cordell. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*, 10(6):392–404, Jun 2009.

[122] Kathryn L. Lunetta, L Brooke Hayward, Jonathan Segal, and Paul Van Eerdewegh. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet*, 5:32, 2004.

[123] G. Holmes and M. A. Hall. A development environment for predictive modelling in foods. *Int J Food Microbiol*, 73(2-3):351–362, Mar 2002.

[124] Alejandra Medina-Rivera, Cei Abreu-Goodger, Morgane Thomas-Chollier, Heladia Salgado, Julio Collado-Vides, and Jacques van Helden. Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Res*, 39(3):808–824, Feb 2011.

[125] Zhaolei Zhang and Mark Gerstein. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J Biol*, 2(2):11, 2003.

[126] Vipin Narang, Ankush Mittal, and Wing-Kin Sung. Localized motif discovery in gene regulatory sequences. *Bioinformatics*, 26(9):1152–1159, May 2010.

[127] Jason Ernst, Heather L. Plasterer, Itamar Simon, and Ziv Bar-Joseph. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res*, 20(4):526–536, Apr 2010.

[128] Stephen A. Ramsey, Theo A. Knijnenburg, Kathleen A. Kennedy, Daniel E. Zak, Mark Gilchrist, Elizabeth S. Gold, Carrie D. Johnson, Aaron E. Lampano, Vladimir Litvak, Garnet Navarro, Tetyana Stolyar, Alan Aderem, and Ilya Shmulevich. Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics*, 26(17):2071–2075, Sep 2010.

[129] Socorro Gama-Castro, Verónica Jiménez-Jacinto, Martín Peralta-Gil, Alberto Santos-Zavaleta, Mónica I. Peñaloza-Spinola, Bruno Contreras-Moreira, Juan Segura-Salazar, Luis Muñiz-Rascado, Irma Martínez-Flores, Heladia Salgado, César Bonavides-Martínez, Cei Abreu-Goodger, Carlos Rodríguez-Penagos, Juan Miranda-Ríos, Enrique Morett, Enrique Merino, Araceli M. Huerta, Luis Treviño-Quintanilla, and Julio Collado-Vides. Regulondb (version 6.0): gene regulation model of escherichia coli k-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation. *Nucleic Acids Res*, 36(Database issue):D120–D124, Jan 2008.

[130] Zeba Wunderlich and Leonid A. Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet*, 25(10):434–440, Oct 2009.

[131] W. Hendrickson and R. Schleif. A dimer of arac protein contacts three adjacent major groove regions of the arai dna site. *Proc Natl Acad Sci U S A*, 82(10):3129–3133, May 1985.

[132] Y. Lu, C. Flaherty, and W. Hendrickson. Arac protein contacts asymmetric sites in the escherichia coli arafgh promoter. *J Biol Chem*, 267(34):24848–24857, Dec 1992.

[133] Alejandro Toro-Roman, Timothy R. Mack, and Ann M. Stock. Structural analysis and solution studies of the activated regulatory domain of the response regulator arca: a symmetric dimer mediated by the alpha4-beta5-alpha5 face. *J Mol Biol*, 349(1):11–26, May 2005.

[134] E. Martínez-Hackert and A. M. Stock. Structural relationships in the ompr family of winged-helix transcription factors. *J Mol Biol*, 269(3):301–312, Jun 1997.

[135] C. Q. Pan, S. E. Finkel, S. E. Cramton, J. A. Feng, D. S. Sigman, and R. C. Johnson. Variable structures of fis-dna complexes determined by flanking dna-protein contacts. *J Mol Biol*, 264(4):675–695, Dec 1996.

[136] H. Afflerbach, O. Schröder, and R. Wagner. Conformational changes of the upstream dna mediated by h-ns and fis regulate e. coli rrnb p1 promoter activity. *J Mol Biol*, 286(2):339–353, Feb 1999.

[137] T. D. Schneider. Strong minor groove base conservation in sequence logos implies dna distortion or base flipping during replication and transcription initiation. *Nucleic Acids Res*, 29(23):4881–4891, Dec 2001.

[138] A. Travers. Dna-protein interactions: Ihf–the master bender. *Curr Biol*, 7(4):R252–R254, Apr 1997.

[139] Adrianna P P. Zhang, Ying Z. Pigli, and Phoebe A. Rice. Structure of the lexa-dna complex and implications for sos box measurement. *Nature*, 466(7308):883–886, Aug 2010.

[140] L. K. Lewis, G. R. Harlow, L. A. Gregg-Jolly, and D. W. Mount. Identification of high affinity binding sites for lexa which define new dna damage-inducible genes in escherichia coli. *J Mol Biol*, 241(4):507–523, Aug 1994.

[141] G. Camenisch, D. M. Stroka, M. Gassmann, and R. H. Wenger. Attenuation of hif-1 dna-binding activity limits hypoxia-inducible endothelin-1 expression. *Pflugers Arch*, 443(2):240–249, Nov 2001.

[142] Shingo Kajimura, Katsumi Aida, and Cunming Duan. Understanding hypoxia-induced gene expression in early development: in vitro and in vivo analysis of hypoxia-inducible factor 1-regulated zebra fish insulin-like growth factor binding protein 1 gene expression. *Mol Cell Biol*, 26(3):1142–1155, Feb 2006.

[143] G. Michel, E. Minet, I. Ernest, I. Roland, F. Durant, J. Remacle, and C. Michiels. A model for the complex between the hypoxia-inducible factor-1 (hif-1) and its consensus dna sequence. *J Biomol Struct Dyn*, 18(2):169–179, Oct 2000.

[144] E. Kim, N. Albrechtsen, and W. Deppert. Dna-conformation is an important determinant of sequence-specific dna binding by tumor suppressor p53. *Oncogene*, 15(7):857–869, Aug 1997.

[145] Esther Marco, Raquel García-Nieto, and Federico Gago. Assessment by molecular dynamics simulations of the structural determinants of dna-binding specificity for transcription factor sp1. *J Mol Biol*, 328(1):9–32, Apr 2003.

[146] Y. Shi and J. M. Berg. Dna unwinding induced by zinc finger protein binding. *Biochemistry*, 35(12):3845–3848, Mar 1996.

[147] Wei-Guo Zhu, Kanur Srinivasan, Zunyan Dai, Wenrui Duan, Lawrence J. Druhan, Haiming Ding, Lisa Yee, Miguel A. Villalona-Calero, Christoph Plass, and Gregory A. Otterson. Methylation of adjacent cpg sites affects sp1/sp3 binding and activity in the p21(cip1) promoter. *Mol Cell Biol*, 23(12):4056–4065, Jun 2003.

[148] X. Chen, U. Vinkemeier, Y. Zhao, D. Jeruzalmi, JE Darnell, Jr, and J. Kuriyan. Crystal structure of a tyrosine phosphorylated stat-1 dimer bound to dna. *Cell*, 93(5):827–839, May 1998.

[149] G. B. Ehret, P. Reichenbach, U. Schindler, C. M. Horvath, S. Fritz, M. Nabholz, and P. Bucher. Dna binding specificity of different stat proteins. comparison of in vitro specificity with natural target sites. *J Biol Chem*, 276(9):6675–6688, Mar 2001.

[150] Z. S. Juo, T. K. Chiu, P. M. Leiberman, I. Baikalov, A. J. Berk, and R. E. Dickerson. How proteins recognize the tata box. *J Mol Biol*, 261(2):239–254, Aug 1996.

[151] Robyn M. Powell, Kay M. Parkhurst, and Lawrence J. Parkhurst. Comparison of tata-binding protein recognition of a variant and consensus dna promoters. *J Biol Chem*, 277(10):7776–7784, Mar 2002.

[152] N. A. Davis, S. S. Majee, and J. D. Kahn. Tata box dna deformation with and without the tata box-binding protein. *J Mol Biol*, 291(2):249–265, Aug 1999.

[153] Eleanor J. Gardiner, Christopher A. Hunter, Xiang-Jun Lu, and Peter Willett. A structural similarity analysis of double-helical dna. *J Mol Biol*, 343(4):879–889, Oct 2004.

[154] Thomas Abeel, Yvan Saeys, Eric Bonnet, Pierre Rouzé, and Yves Van de Peer. Generic eukaryotic core promoter prediction using structural features of dna. *Genome Res*, 18(2):310–323, Feb 2008.

[155] Jason A. Greenbaum, Bo Pang, and Thomas D. Tullius. Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res*, 17(6):947–953, Jun 2007.

[156] Remo Rohs, Sean M. West, Peng Liu, and Barry Honig. Nuance in the double-helix and its role in protein-dna recognition. *Curr Opin Struct Biol*, 19(2):171–177, Apr 2009.

[157] Tom Tullius. Structural biology: Dna binding shapes up. *Nature*, 461(7268):1225–1226, Oct 2009.

[158] Dang Long, Rosalind Lee, Peter Williams, Chi Yu Chan, Victor Ambros, and Ye Ding. Potent effect of target structure on microrna function. *Nat Struct Mol Biol*, 14(4):287–294, Apr 2007.

[159] N. R. Steffen, S. D. Murphy, L. Tolleri, G. W. Hatfield, and R. H. Lathrop. Dna sequence and structure: direct and indirect recognition in protein-dna binding. *Bioinformatics*, 18 Suppl 1:S22–S30, 2002.

[160] M Michael Gromiha, Joerg G. Siebers, Samuel Selvaraj, Hidetoshi Kono, and Akinori Sarai. Role of inter and intramolecular interactions in protein-dna recognition. *Gene*, 364:108–113, Dec 2005.

[161] Barry Honig and Remo Rohs. Biophysics: Flipping watson and crick. *Nature*, 470(7335):472–473, Feb 2011.

[162] Kate R. Rosenbloom, Cricket A. Sloan, Venkat S. Malladi, Timothy R. Dreszer, Katrina Learned, Vanessa M. Kirkup, Matthew C. Wong, Morgan Maddren, Ruihua Fang, Steven G. Heitner, Brian T. Lee, Galt P. Barber, Rachel A. Harte, Mark Diekhans, Jeffrey C. Long, Steven P. Wilder, Ann S. Zweig, Donna Karolchik, Robert M. Kuhn, David Haussler, and W James Kent. Encode data in the ucsc genome browser: year 5 update. *Nucleic Acids Res*, 41(Database issue):D56–D63, Jan 2013.

[163] I. Dunham, A. Kundaje, S.F. Aldred, P.J. Collins, C.A. Davis, F. Doyle, C.B. Epstein, S. Frietze, J. Harrow, and Kaul. R et al. An integrated encyclopedia

of dna elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[164] Robert M. Kuhn, David Haussler, and W James Kent. The ucsc genome browser and associated tools. *Brief Bioinform*, 14(2):144–161, Mar 2013.

[165] S. Kyo, M. Takakura, T. Taira, T. Kanaya, H. Itoh, M. Yutsudo, H. Ariga, and M. Inoue. Sp1 cooperates with c-myc to activate transcription of the human telomerase reverse transcriptase gene (htert). *Nucleic Acids Res*, 28(3):669–677, Feb 2000.

[166] I. Horikawa, P. L. Cable, C. Afshari, and J. C. Barrett. Cloning and characterization of the promoter region of human telomerase reverse transcriptase gene. *Cancer Res*, 59(4):826–830, Feb 1999.

[167] Eibe Frank, Mark Hall, Len Trigg, Geoffrey Holmes, and Ian H. Witten. Data mining in bioinformatics using weka. *Bioinformatics*, 20(15):2479–2481, Oct 2004.

[168] Xi Chen, Han Xu, Ping Yuan, Fang Fang, Mikael Huss, Vinsensius B. Vega, Eleanor Wong, Yuriy L. Orlov, Weiwei Zhang, Jianming Jiang, Yuin-Han Loh, Hock Chuan Yeo, Zhen Xuan Yeo, Vipin Narang, Kunde Ramamoorthy Govindarajan, Bernard Leong, Atif Shahab, Yijun Ruan, Guillaume Bourque, Wing-Kin Sung, Neil D. Clarke, Chia-Lin Wei, and Huck-Hui Ng. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117, Jun 2008.

[169] Peter C. Hollenhorst, Katherine J. Chandler, Rachel L. Poulsen, W Evan Johnson, Nancy A. Speck, and Barbara J. Graves. Dna specificity determinants associate with distinct transcription factor functions. *PLoS Genet*, 5(12):e1000778, Dec 2009.

[170] Yue Zhao, Shuxiang Ruan, Manishi Pandey, and Gary D Stormo. Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, 191(3):781–790, Jul 2012.

[171] Rahul Siddharthan. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PLoS One*, 5(3):e9722, 2010.

[172] Yongli Zhang, Zhiqun Xi, Rashmi S Hegde, Zippora Shakked, and Donald M Crothers. Predicting indirect readout effects in protein-dna interactions. *Proc Natl Acad Sci U S A*, 101(22):8337–8341, Jun 2004.

[173] N. C. Seeman, J. M. Rosenberg, and A. Rich. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci U S A*, 73(3):804–808, Mar 1976.

[174] N. M. Luscombe, R. A. Laskowski, and J. M. Thornton. Amino acid-base interactions: a three-dimensional analysis of protein-dna interactions at an atomic level. *Nucleic Acids Res*, 29(13):2860–2874, Jul 2001.

[175] L. K. Savinkova, M. P. Ponomarenko, P. M. Ponomarenko, I. A. Drachkova, M. V. Lysova, T. V. Arshinova, and N. A. Kolchanov. Tata box polymorphisms in human gene promoters and associated hereditary pathologies. *Biochemistry (Mosc)*, 74(2):117–129, Feb 2009.

[176] Hana Faiger, Marina Ivanchenko, Ilana Cohen, and Tali E. Haran. Tbp flanking sequences: asymmetry of binding, long-range effects and consensus sequences. *Nucleic Acids Res*, 34(1):104–119, 2006.

[177] Beisi Xu, Dustin E. Schones, Yongmei Wang, Haojun Liang, and Guohui Li. A structural-based strategy for recognition of transcription factor binding sites. *PLoS One*, 8(1):e52460, 2013.

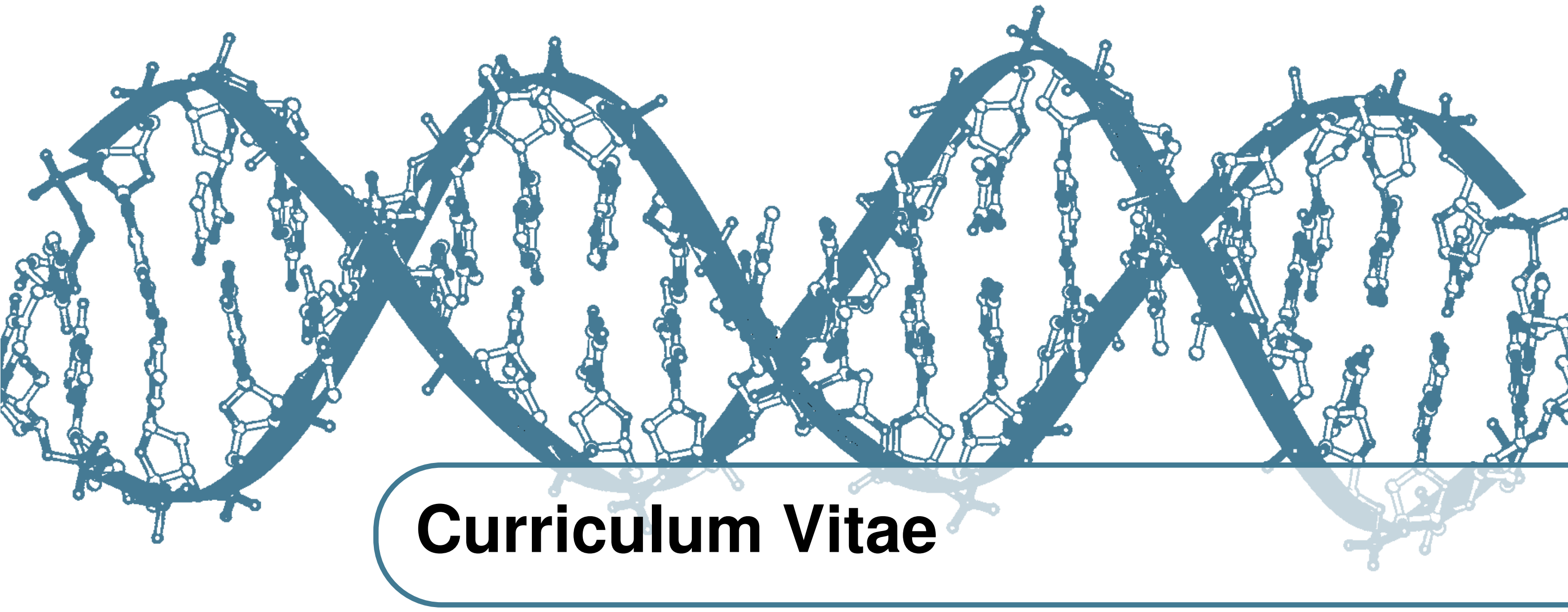
[178] James R. Dutton, Anthony Antonellis, Thomas J. Carney, Frederico S L M. Rodrigues, William J. Pavan, Andrew Ward, and Robert N. Kelsh. An evolutionarily conserved intronic region controls the spatiotemporal expression of the transcription factor sox10. *BMC Dev Biol*, 8:105, 2008.

[179] Marzia Bianchi, Rita Crinelli, Elisa Giacomini, Elisa Carloni, and Mauro Magnani. A potent enhancer element in the 5'-utr intron is crucial for

transcriptional regulation of the human ubiquitin c gene. *Gene*, 448(1):88–101, Dec 2009.

[180] Anne Saumet, Guillaume Vetter, Manuella Bouttier, Elodie Portales-Casamar, Wyeth W. Wasserman, Thomas Maurin, Bernard Mari, Pascal Barbry, Laurent Vallar, Evelyne Friederich, Khalil Arar, Bruno Cassinat, Christine Chomienne, and Charles-Henri Lecellier. Transcriptional repression of microrna genes by pml-rara increases expression of key cancer proteins in acute promyelocytic leukemia. *Blood*, 113(2):412–421, Jan 2009.

[181] Ali C. Ravanpay, Stacey J. Hansen, and James M. Olson. Transcriptional inhibition of rest by neurod2 during neuronal differentiation. *Mol Cell Neurosci*, 44(2):178–189, Jun 2010.



Curriculum Vitae

Personal information

Name: Stefan Broos
Address: Koningin Elisabethlaan 16/101
9000 Ghent (Belgium)
Email: stefan.broos@telenet.be
Phone: +32 (0)495/907521

Education

2009 – now, PhD in Bioinformatics
Topic: Developing algorithms for the *in silico* Identification of Transcription Factor Binding Sites

2009, Ghent University, Ghent
Master, Biochemistry and biotechnology, major bioinformatics (Graduated Magna Cum Laude)
Thesis: *In Silico* Identification of Transcription Factor Binding Sites

2004, Kardinaal Van Roey-Instituut, Vorselaar
Science-Mathematics (8 hours of mathematics)

Experience

2009 - now, VIB, Ghent
Researcher (PhD candidate)
– Developer of *in silico* algorithms for the prediction of transcription factor binding sites (based on the 3D structure of the DNA)
– Integrative data analyst
– Experience with recent high-throughput data analysis techniques

Training and Skills

Hands-on experience in a variety of bioinformatics methods
Experienced in the analysis of next-generation sequencing data and in the analysis of large biological data sets
Extensive knowledge of Perl, C, C++, Java, Python, MySQL, PHP, Git version control, LaTeX and the R software environment for statistical computing and graphics
Member of thesis jury Howest (2011 & 2012)
Guidance of master thesis students and master projects
Guidance of International University college students
Reviewer for Nucleic Acids Research; Oxford Journals
Much experience with high performance computing (super computing platforms)
Attended lectures on openMP and MPI (Specialist Workshops in Parallel Computing – 2013)
Developer of pattern recognition and machine learning algorithms
Thorough knowledge of the Linux operating system

Activities

2013, Presenter, Intelligent Systems for Molecular Biology 2013, Berlin
PhysBinder: improving the prediction of transcription factor binding sites by flexible inclusion of biophysical properties

2013, Lecturer for BITS training facility, Ghent
Topic: Introduction to gene regulation

2013, Lecturer for BITS training facility, Ghent
Topic: Comparative genomics in Eukaryotes

2012, Speaker at "Developed @ VIB", Blankenberge
MirTra a novel tool for the discovery of feed-forward and feedback loops of microRNAs and transcription factors

2012, Presenter, Intelligent Systems for Molecular Biology 2012, Long Beach, CA
A flexible integrative approach based on random forest improves prediction of transcription factor binding sites

2011, Lecturer for BITS training facility, Ghent
Topic: Comparative genomics in Eukaryotes

2011, Presenter, Intelligent Systems for Molecular Biology 2011, Vienna
ConTra v2: a tool to identify transcription factor binding sites across species, update 2011

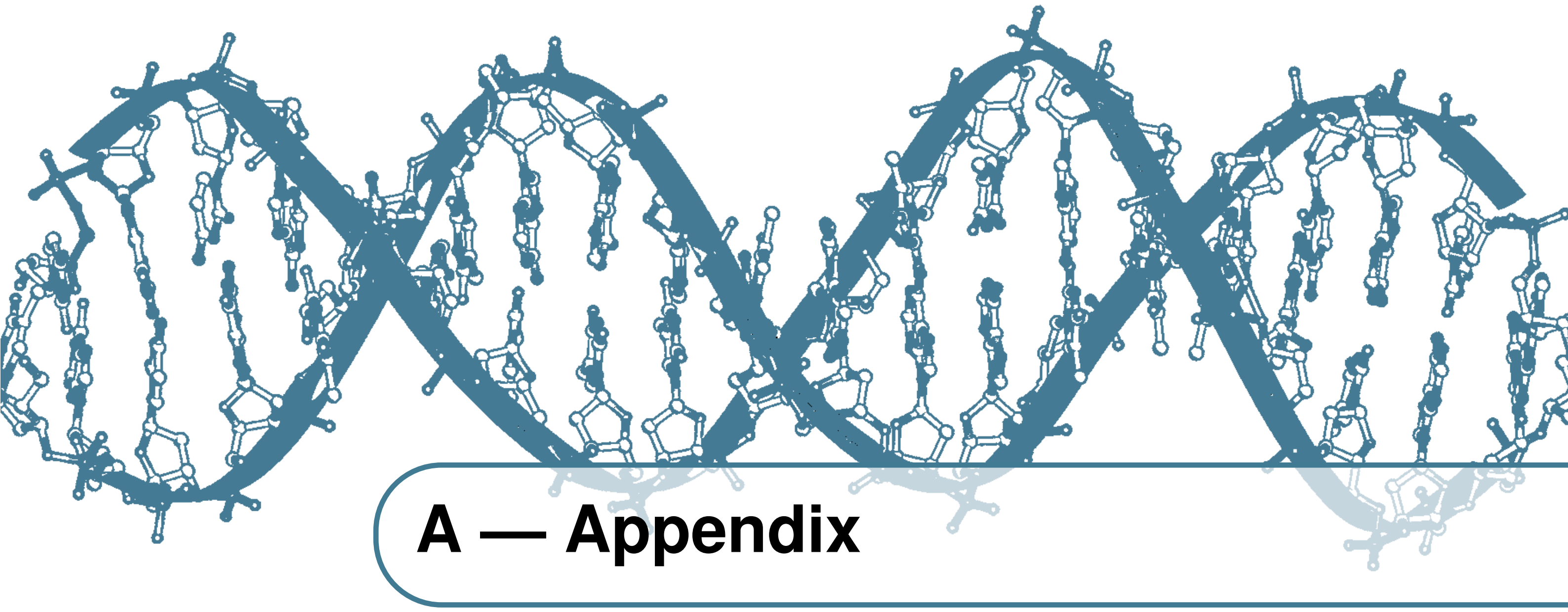
Publications

Stefan Broos, Arne Soete, Bart Hooghe, Raymond Moran, Frans Van Roy and Pieter De Bleser **PhysBinder: improving the prediction of transcription factor binding sites by flexible inclusion of biophysical properties**; Nucleic

Acids Res. 2013 41: W531–W534

Bart Hooghe*, Stefan Broos*, Frans Van Roy, and Pieter De Bleser **A flexible integrative approach based on random forest improves prediction of transcription factor binding sites**; Nucleic Acids Res. 2012 40: e106 (*Joint first authors)

Stefan Broos, Paco Hulpiau, Jeroen Galle, Bart Hooghe, Frans Van Roy, and Pieter De Bleser **ConTra v2: a tool to identify transcription factor binding sites across species, update 2011**; Nucleic Acids Res. 2011 39: W74-W78



A — Appendix

Features in the Random Forest models

HIF1:
dint: [-30,-3],[-1,14],[16,20],[22,26]
PWMmatrixscore: general
uniformity_A: [-9,-7],[-5,-3],[-1,10],[12,14],16,19,[22,23],fullseqmean
monont: -30,[-27,-26],-24,[-21,-17],-15,[-13,-12],[-9,-6],[-4,-2],[0,11],[13,14],[17,20],[23,24],26
homogeneity_BA: -7,-1,[1,9],fullseqmean

P53:
homogeneity_RESTB: [2,3],[5,8],[12,16],fullseqmean
tors_1_proteinOlson: -2,[1,5],8,[10,15],fullseqmean
dint: [-30,-25],-23,[-19,-18],[-15,-6],[-3,24],[26,27]
PWMmatrixscore: general
uniformity_A: [-8,-6],-3,[-1,0],[2,17],19,[21,23],fullseqmean
homogeneity_BI: -1,[1,17],[19,20],23,fullseqmean
PWMcorescore: general

SP1:
homogeneity_RESTB: [-9,-3],[-1,17],[19,24],fullseqmean
tors_1_proteinOlson: -30,-18,[-2,7],15,22
dint: -29,-24,[-17,-15],-12,[-7,-5],[-2,9],11,14,16,23,26
uniformity_B: -9,[-7,-5],[-1,8],[12,13],18,21,fullseqmean
uniformity_AB: [-6,-5],[-2,8],11,[14,17],[20,23],fullseqmean
PWMmatrixscore: general
monont: [-30,9],[11,29]
homogeneity_AB: [-8,-3],[-1,9],[11,16],fullseqmean

STAT1:
homogeneity_RESTB: -3,1,[4,15],17
tors_1_proteinOlson: [3,13]
dint: -21,-14,[-11,-9],-4,[0,18]
PWMmatrixscore: general
minor_groove_clash_distance: 2,[4,15],fullseqmean
homogeneity_AB: [5,9],[11,14],fullseqmean
homogeneity_BA: [4,15],17,fullseqmean

TBP:
tors_1_unboundLiu: [-11,-10],[-2,4]
tors_1_proteinOlson: [-2,2]
bend_towards_major_groove: -9,-3,[0,5],[7,8],[11,14],[16,18],21,fullseqmean
dint: [-29,-28],[-26,-25],[-22,-21],[-19,-17],-14,-12,[-8,-7],[-4,-3],[-1,10],[13,22],[25,26],28
homogeneity_BII: -9,-7,[-1,5],[7,10],13,17,20,24,fullseqmean
bend_towards_minor_groove: -9,-3,[0,5],[7,9],[12,16],19,[21,22],fullseqmean

ARAC:
PWMmatrixscore: general
monont: -11,-9,-6,0,5,13,19
minor_groove_clash_size: -7,[-2,0],18,fullseqmean

ARCA:
PWMmatrixscore: general
groovewidth_unboundLiu: [-8,-7],[-3,5],[9,10],fullseqmean
monont: -18,-14,[-7,-6],[-4,-2],[0,2],[4,5],7,10

FIS: PWMmatrixscore: general
uniformity_A: [-9,-8],[-5,1],[5,9],[16,18],fullseqmean
minor_groove_clash_distance: [-9,-8],[-5,-1],2,[5,6],15,17
major_groove_clash_distance: [-9,-8],[-5,-4],2,[5,6]
monont: -20,[-12,-11],-8,[-4,3],[6,7],9,[17,19]
PWMcorescore: general

FLHDC: tors_1_nucleosome: -7,-5,[1,2],fullseqmean
curv_1_unboundLiu: [-1,0],2,12,fullseqmean
PWMmatrixscore: general
monont: -20,-3
uniformity_B: [-9,-8],[-4,-3],2

IHF:
tors_1_proteinOlson: -9,[-7,-6],-3,5,9,13,fullseqmean
bend_towards_major_groove: [-9,16],18,fullseqmean
PWMmatrixscore: general
monont: [-20,-17],[-15,-11],[-9,-5],[-3,-2],[0,1],[3,7],[9,10],12,[14,16],19
bend_towards_minor_groove: [-9,18],fullseqmean

LEXA:
dint: [-9,-7],5

PWMmatrixscore: general
minor_groove_clash_distance: [-9,-3],[3,6],fullseqmean

PURR:
homogeneity_RESTB: [-7,-3],[0,2],12,fullseqmean
PWMmatrixscore: general
monont: -14,-9,-7,[-5,-2],[0,2],4,6

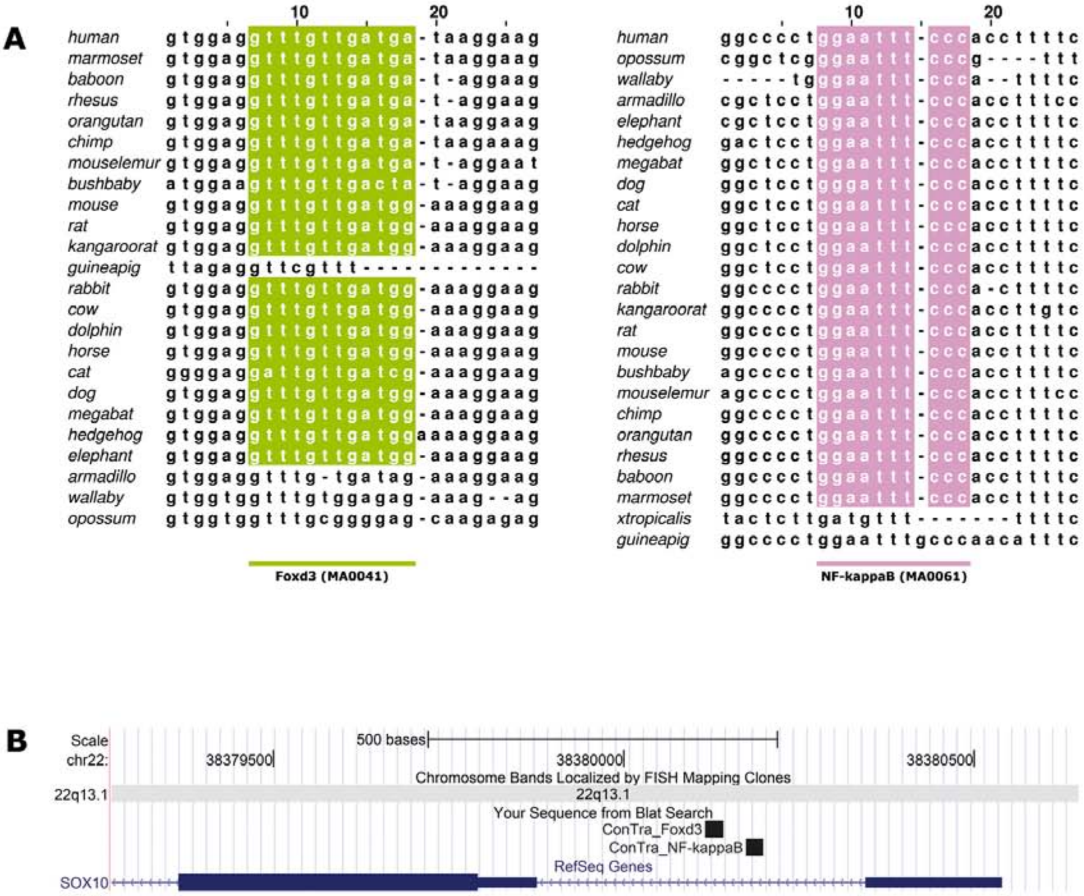
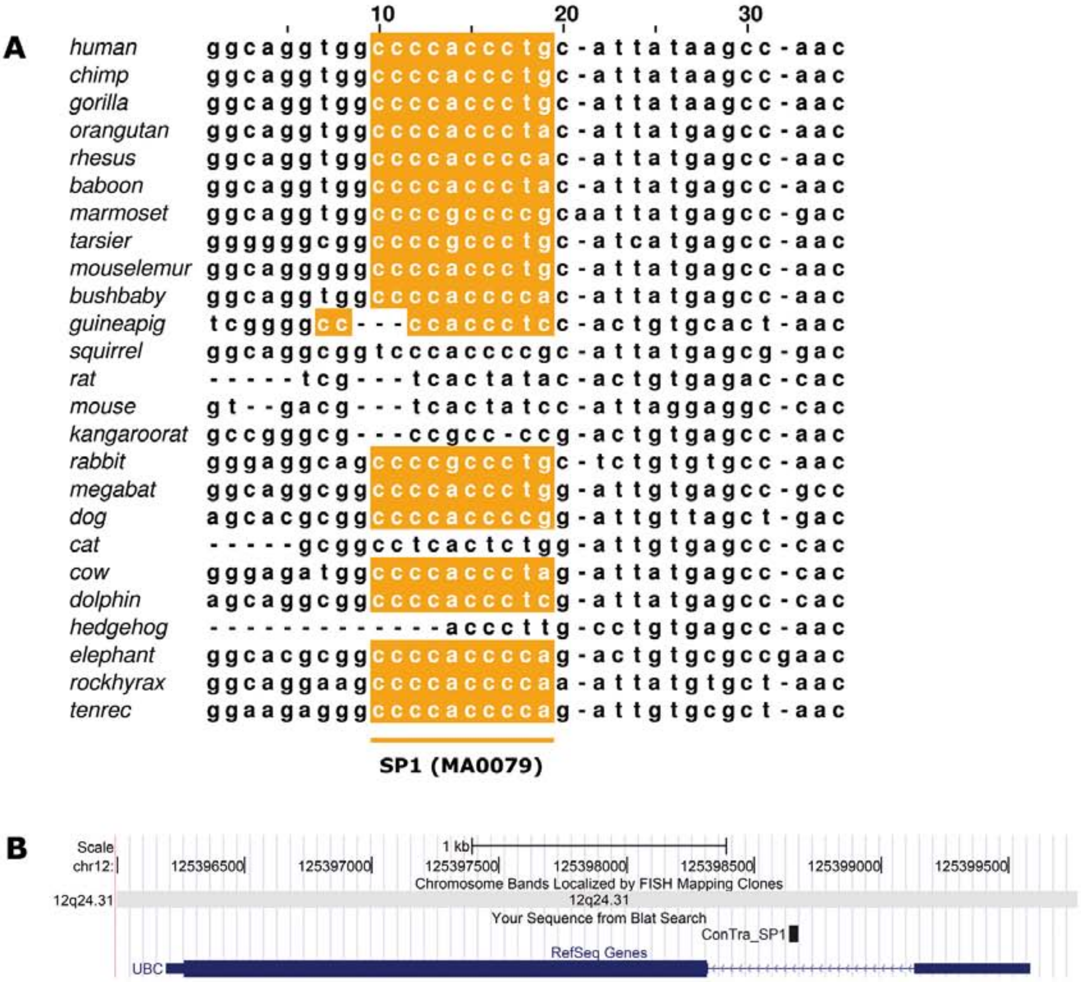

A
10
20
human gtggaggtttgttgatga-taaggaag
marmoset gtggaggtttgttgatga-taaggaag
baboon gtggaggtttgttgatga-t-aggaag
rhesus gtggaggtttgttgatga-t-aggaag
orangutan gtggaggtttgttgatga-taaggaag
chimp gtggaggtttgttgatga-taagaaag
mouselemur gtggaggtttgttgatga-t-aggaat
bushbaby atggaagtttgttgacta-t-aggaag
mouse gtggaggtttgttgatgg-aaaggaag
rat gtggaggtttgttgatgg-aaaggaag
kangaroorat gtggaggtttgttgatgg-aaaggaag
guineapig ttagaggttcgttt--------------
rabbit gtggaggtttgttgatgg-aaaggaag
cow gtggaggtttgttgatgg-aaaggaag
dolphin gtggaggtttgttgatgg-aaaggaag
horse gtggaggtttgttgatgg-aaaggaag
cat ggggaggattgttgatcg-aaaggaag
dog gtggaggtttgttgatgg-aaaggaag
megabat gtggaggtttgttgatgg-aaaggaag
hedgehog gtggaggtttgttgatggaaaaggaag
elephant gtggaggtttgttgatgg-aaaggaag
armadillo gtggaggtttg-tgatag-aaaggaag
wallaby gtggtggtttgtggagag-aaag--ag
opossum gtggtggtttgcggggag-caagagag
Foxd3 (MA0041)
10
20
human ggcccctggaattt-cccaccttttc
opossum cggctcgggaattt-cccg----ttt
wallaby ------tgggaattt-ccca--ttttc
armadillo cgctcctggaattt-cccacctttcc
elephant cgctcctggaattt-cccaccttttc
hedgehog gactcctggaattt-cccaccttttc
megabat ggctcctggaattt-cccaccttttc
dog ggctcctgggattt-cccaccttttc
cat ggctcctggaattt-cccaccttttc
horse ggctcctggaattt-cccaccttttc
dolphin ggctcctggaattt-cccaccttttc
cow ggctcctggaattt-cccaccttttc
rabbit ggcccctggaattt-ccca-cttttc
kangaroorat ggcccctggaattt-cccaccttgtc
rat ggcccctggaattt-cccaccttttc
mouse ggcccctggaattt-cccaccttttc
bushbaby agcccctggaattt-cccaccttttc
mouselemur agcccctggaattt-cccacctttcc
chimp ggcccctggaattt-cccaccttttc
orangutan ggcccctggaattt-cccaccttttc
rhesus ggcccctggaattt-cccaccttttc
baboon ggcccctggaattt-cccaccttttc
marmoset ggcccctggaattt-cccaccttttc
xtropicalis tactcttgatgttt-------ttttc
guineapig ggcccctggaatttgcccaacatttc
NF-kappaB (MA0061)
B
Scale
chr22:
500 bases
38379500
38380000
38380500
Chromosome Bands Localized by FISH Mapping Clones
22q13.1
22q13.1
Your Sequence from Blat Search
ConTra_Foxd3
ConTra_NF-kappaB
RefSeq Genes
SOX10

Figure A.1: Visualization of the evolutionary conserved Foxd3 and NF-kappaB binding sites in the first intronic region of the transcription factor SOX10, known for regulating spatiotemporal expression as described by Dutton et al [178]. (A) Multiz alignment showing the conserved Foxd3 site (in green), predicted using the JASPAR positional weight matrix MA0041 and the NF-kappaB site (in pink), predicted by the MA0061 matrix. The figure was created with the free multiple alignment editor Jalview using the ConTra FASTA and feature color (.fc) file on the results page. (B) Both regions from (A) were mapped using BLAT on intron 1 in the UCSC Genome Browser and are shown as black boxes. Blue boxes represent exon 1 (right) and exon 2 (left) with the blue arrows illustrating the intronic regions.



```
                    10        20        30
human      ggcaggtggccccaccctgc-attataagcc-aac
chimp      ggcaggtggccccaccctgc-attataagcc-aac
gorilla    ggcaggtggccccaccctgc-attataagcc-aac
orangutan  ggcaggtggccccacccctac-attatgagcc-aac
rhesus     ggcaggtggccccaccccac-attatgagcc-aac
baboon     ggcaggtggccccaccctac-attatgagcc-aac
marmoset   ggcaggtggccccgccccgcaattatgagcc-gac
tarsier    ggggggcggccccgccctgc-atcatgagcc-aac
mouselemur ggcagggggccccaccctgc-attatgagcc-aac
bushbaby   ggcaggtggccccaccccac-attatgagcc-aac
guineapig  tcggggcc---ccaccctcc-actgtgcact-aac
squirrel   ggcaggcggtcccaccccgc-attatgagcg-gac
rat        -----tcg---tcactatac-actgtgagac-cac
mouse      gt--gacg---tcactatcc-attaggaggc-cac
kangaroorat gccgggcg---ccgcc-ccg-actgtgagcc-aac
rabbit     gggaggcagccccgccctgc-tctgtgtgcc-aac
megabat    ggcaggcggccccaccctgg-attgtgagcc-gcc
dog        agcacgcggccccaccccgg-attgttagct-gac
cat        -----gcggcctcactctgg-attgtgagcc-cac
cow        gggagatggccccaccctag-attatgagcc-cac
dolphin    agcaggcggccccaccctcg-attatgagcc-cac
hedgehog   -------------acccttg-cctgtgagcc-aac
elephant   ggcacgcggccccaccccag-actgtgcgccgaac
rockhyrax  ggcaggaagccccaccccaa-attatgtgct-aac
tenrec     ggaagagggccccaccccag-attgtgcgct-aac
```

Figure A.2: Visualization of the evolutionary conserved Sp1 binding site in the second intron of the human UBC gene as described by Bianchi et al [179]. (A) Multiz alignment showing the conserved Sp1 site (in orange), predicted using the JASPAR positional weight matrix MA0079.2. (B) Region of (A) was mapped using BLAT on the first intronic region in the UCSC Genome Browser (black box). Blue boxes represents exons with the blue arrows illustrating the intronic regions.

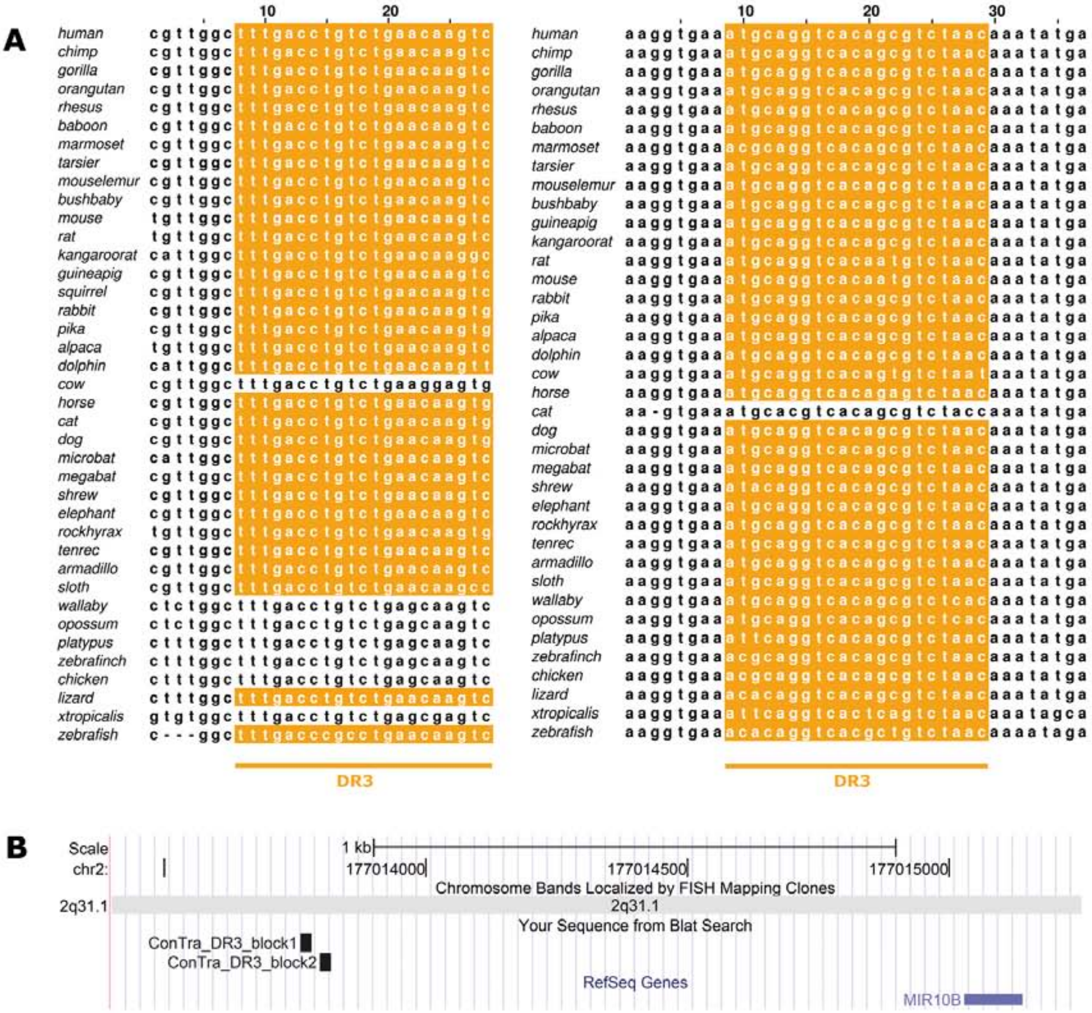


Figure A.3: Visualization of the evolutionary conserved DR3 binding site in the promoter region of microRNA-10b as described by Saumet et al [180]. (A) Multiz alignment showing the conserved DR3 site (in orange), predicted using the TRANSFAC M00966 matrix. (B) Both regions from (A) were mapped using BLAT on the promoter in the UCSC Genome Browser and are shown as black boxes. Blue box represents the microRNA location.

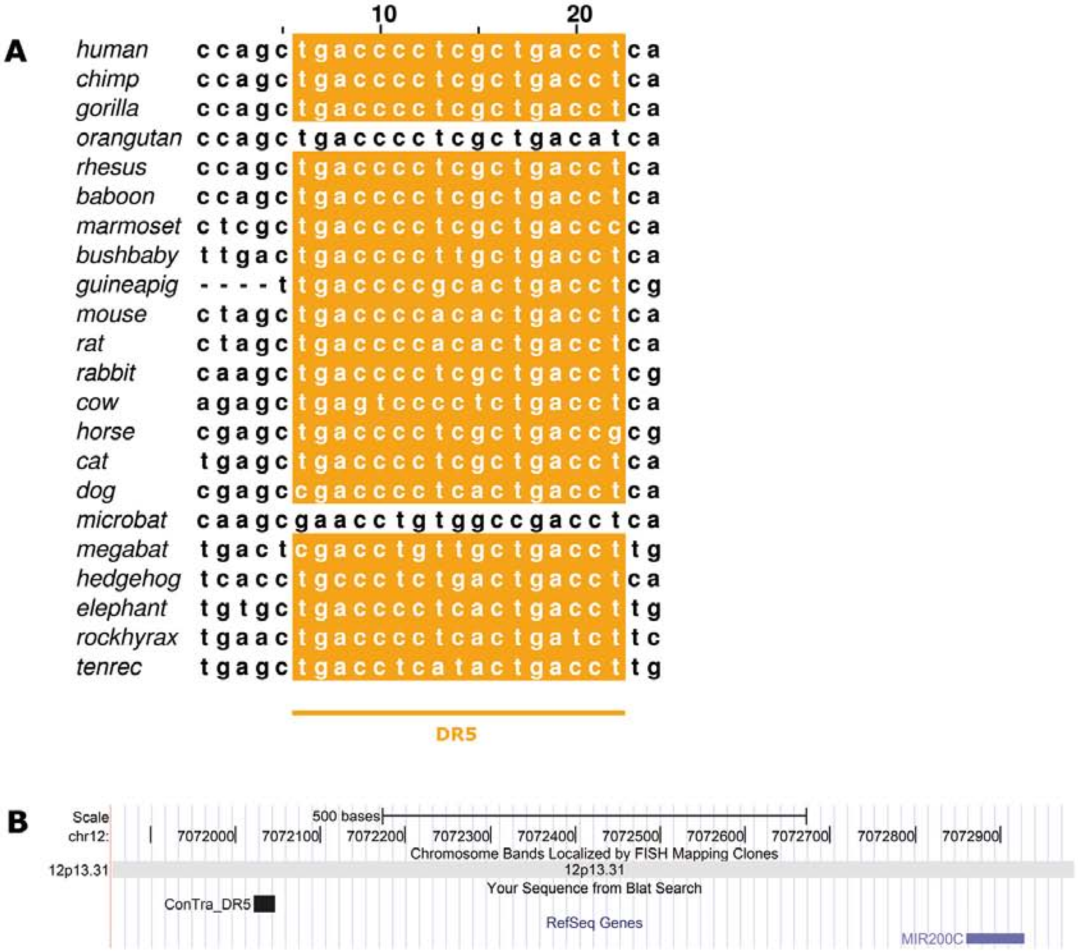

10
20
A
human ccagctgacccctcgctgacctca
chimp ccagctgacccctcgctgacctca
gorilla ccagctgacccctcgctgacctca
orangutan ccagctgacccctcgctgacatca
rhesus ccagctgacccctcgctgacctca
baboon ccagctgacccctcgctgacctca
marmoset ctcgctgacccctcgctgacccca
bushbaby ttgactgaccccttgctgacctca
guineapig ----ttgaccccgcactgacctcg
mouse ctagctgaccccacactgacctca
rat ctagctgaccccacactgacctca
rabbit caagctgacccctcgctgacctcg
cow agagctgagtcccctctgacctca
horse cgagctgacccctcgctgaccgcg
cat tgagctgacccctcgctgacctca
dog cgagccgacccctcactgacctca
microbat caagcgaacctgtggccgacctca
megabat tgactcgacctgttgctgaccttg
hedgehog tcacctgccctctgactgacctca
elephant tgtgctgacccctcactgaccttg
rockhyrax tgaactgacccctcactgatcttc
tenrec tgagctgacctcatactgaccttg
DR5
B
Scale
500 bases
chr12:
7072000
7072100
7072200
7072300
7072400
7072500
7072600
7072700
7072800
7072900
Chromosome Bands Localized by FISH Mapping Clones
12p13.31
12p13.31
Your Sequence from Blat Search
ConTra_DR5
RefSeq Genes
MIR200C

Figure A.4: Visualization of the evolutionary conserved DR5 binding site in the promoter region of microRNA-200c as described by Saumet et al [180]. (A) Multiz alignment showing the conserved DR5 site (in orange), predicted using the JASPAR MA0159.1 matrix. (B) Region of (A) was mapped using BLAT on the promoter in the UCSC Genome Browser (black box). Blue box represents the microRNA location.

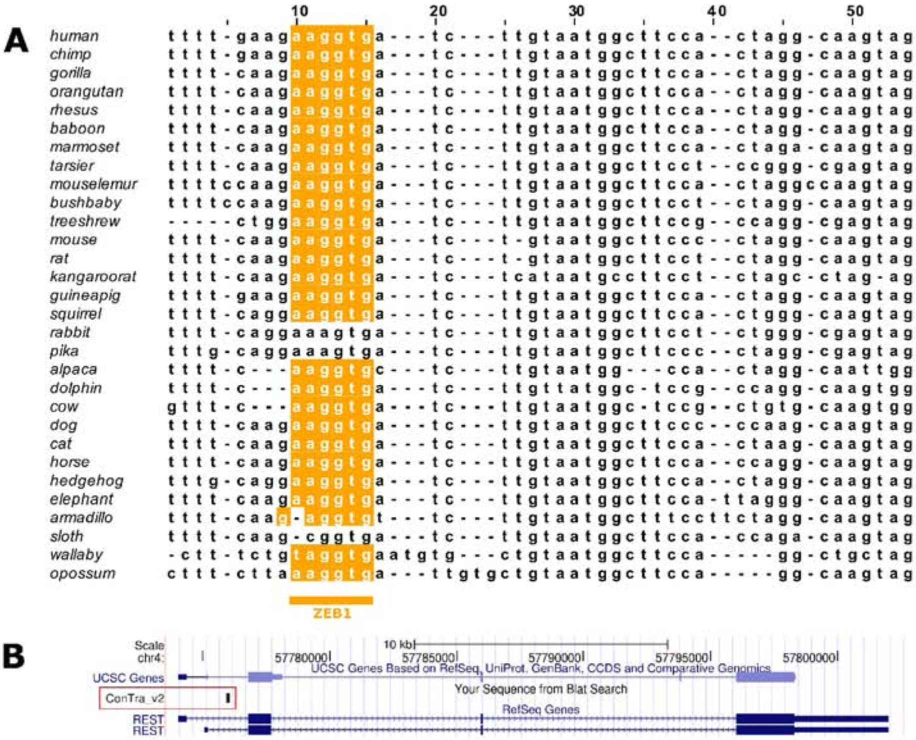

A
10
20
30
40
50
human tttt-gaagaaggtga---tc---ttgtaatggcttcca--ctagg-caagtag
chimp tttt-gaagaaggtga---tc---ttgtaatggcttcca--ctagg-caagtag
gorilla tttt-caagaaggtga---tc---ttgtaatggcttcca--ctagg-caagtag
orangutan tttt-caagaaggtga---tc---ttgtaatggcttcca--ctagg-caagtag
rhesus tttt-caagaaggtga---tc---ttgtaatggcttcca--ctagg-caagtag
baboon tttt-caagaaggtga---tc---ttgtaatggcttcca--ctagg-caagtag
marmoset tttt-caagaaggtga---tc---ttgtaatggcttcca--ctaga-caagtag
tarsier tttt-caagaaggtga---tc---ttgtaatggcttcct--ccggg-cgagtag
mouselemur ttttccaagaaggtga---tc---ttgtaatggcttcca--ctaggccaagtag
bushbaby ttttccaagaaggtga---tc---ttgtaatggcttcct--ctagg-caagtag
treeshrew -----ctggaaggtga---tc---ttgtaatggcttccg--ccagg-cgagtag
mouse tttt-caagaaggtga---tc---t-gtaatggcttccc--ctagg-caagtag
rat tttt-caagaaggtga---tc---t-gtaatggcttcct--ctagg-caagtag
kangaroorat tttt-caagaaggtga---tc---tcataatgccttcct--ctagc-ctag-ag
guineapig tttt-gaagaaggtga---tc---ttgtaatggcttcca--ctagg-caagtag
squirrel tttt-caggaaggtga---tc---ttgtaatggcttcca--ctggg-caagtag
rabbit tttt-caggaaagtga---tc---ttgtaatggcttcct--ctggg-cgagtag
pika tttg-caggaaagtga---tc---ttgtaatggcttccc--ctagg-cgagtag
alpaca tttt-c---aaggtgc---tc---ttgtaatgg---cca--ctagg-caattgg
dolphin tttt-c---aaggtga---tc---ttgttatggc-tccg--ccagg-caagtgg
cow gttt-c---aaggtga---tc---ttgtaatggc-tccg--ctgtg-caagtgg
dog tttt-caagaaggtga---tc---ttgtaatggcttccc--ccaag-caagtag
cat tttt-caagaaggtga---tc---ttgtaatggcttcca--ctaag-caagtag
horse tttt-caagaaggtga---tc---ttgtaatggcttcca--ccagg-caagtag
hedgehog tttg-caggaaggtga---tc---ttgtaatggcttcca--ctagg-caagtag
elephant tttt-caagaaggtga---tc---ttgtaatggcttcca-ttaggg-caagtag
armadillo tttt-caag-aggtgt---tc---ttgtaatggctttccttctagg-caagtag
sloth tttt-caag-cggtga---tc---ttgtaatggcttcca--ccaga-caagtag
wallaby -ctt-tctgtaggtgaatgtg---ctgtaatggcttcca-----gg-ctgctag
opossum cttt-cttaaaggtga---ttgtgctgtaatggcttcca-----gg-caagtag
ZEB1
B
Scale
chr4:
10 kb
57780000
57785000
57790000
57795000
57800000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
UCSC Genes
Your Sequence from Blat Search
ConTra_v2
RefSeq Genes
REST
REST

Figure A.5: Visualization of the evolutionary conserved ZEB1 binding site in the 5'-UTR region of transcription factor REST known for silencing activity of neuronal genes in nonneuronal cells as described by Ravanpay et al [181]. (A) Multiz alignment showing the conserved ZEB1 site (in orange), predicted using the JASPAR positional weight matrix MA0103.1. (B) Region of (A) was mapped using BLAT on 5'-UTR in the UCSC Genome Browser (black box). Blue boxes represent exons with the blue arrows illustrating the intronic regions.

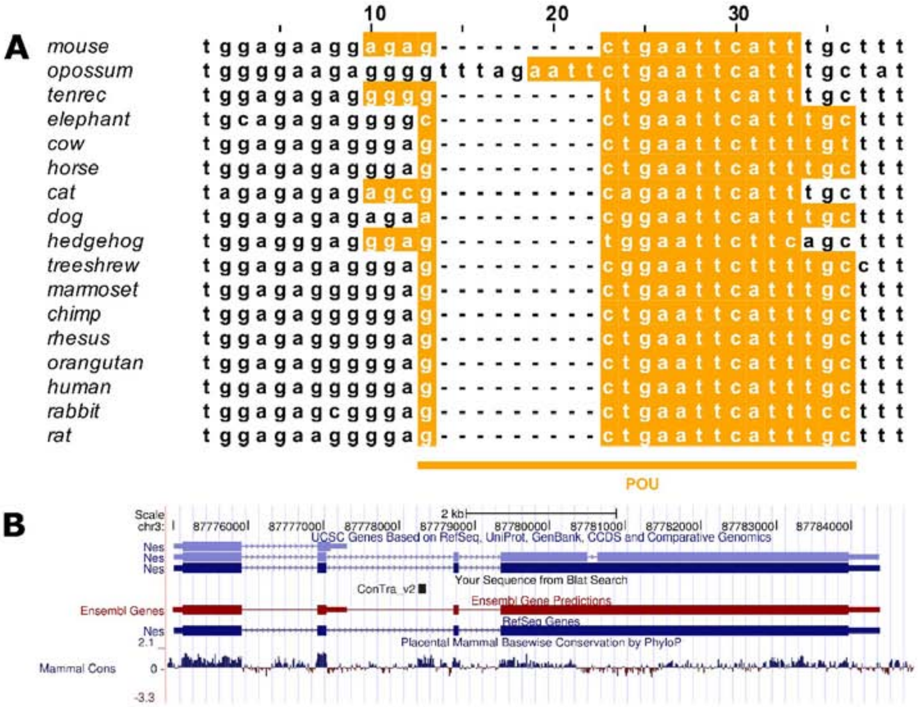

10
20
30
A
mouse t g g a g a a g g a g a g - - - - - - - - - c t g a a t t c a t t t g c t t t
opossum t g g g g a a g a g g g g t t t a g a a t t c t g a a t t c a t t t g c t a t
tenrec t g g a g a g a g g g g g - - - - - - - - - t t g a a t t c a t t t g c t t t
elephant t g c a g a g a g g g g c - - - - - - - - - c t g a a t t c a t t t g c t t t
cow t g g a g a g a g g g a g - - - - - - - - - c t g a a t t c t t t t g t t t t
horse t g g a g a g a g g g a g - - - - - - - - - c t g a a t t c a t t t g c t t t
cat t a g a g a g a g a g c g - - - - - - - - - c a g a a t t c a t t t g c t t t
dog t g g a g a g a g a g a a - - - - - - - - - c g g a a t t c a t t t g c t t t
hedgehog t g g a g g g a g g g a g - - - - - - - - - t g g a a t t c t t c a g c t t t
treeshrew t g g a g a g a g g g a g - - - - - - - - - c g g a a t t c t t t t g c c t t
marmoset t g g a g a g g g g g a g - - - - - - - - - c t g a a t t c a t t t g c t t t
chimp t g g a g a g g g g g a g - - - - - - - - - c t g a a t t c a t t t g c t t t
rhesus t g g a g a g g g g g a g - - - - - - - - - c t g a a t t c a t t t g c t t t
orangutan t g g a g a g g g g g a g - - - - - - - - - c t g a a t t c a t t t g c t t t
human t g g a g a g g g g g a g - - - - - - - - - c t g a a t t c a t t t g c t t t
rabbit t g g a g a g c g g g a g - - - - - - - - - c t g a a t t c a t t t c c t t t
rat t g g a g a a g g g g a g - - - - - - - - - c t g a a t t c a t t t g c t t t
POU
B
Scale
chr3:
2 kb
87776000
87777000
87778000
87779000
87780000
87781000
87782000
87783000
87784000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Nes
Nes
Nes
Your Sequence from Blat Search
ConTra_v2
Ensembl Gene Predictions
Ensembl Genes
RefSeq Genes
Nes
2.1
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
0
-3.3

Figure A.6: Visualization of the evolutionary conserved POU binding site in the second intron of the mouse NES gene as described by Jin et al [69]. (A) Multiz alignment showing the conserved POU site (in orange), predicted using the TRANSFAC M00133 positional weight matrix. (B) Region of (A) was mapped using BLAT on the second intronic region in the UCSC Genome Browser (black box). Blue boxes represent exons with the blue arrows illustrating the intronic regions.

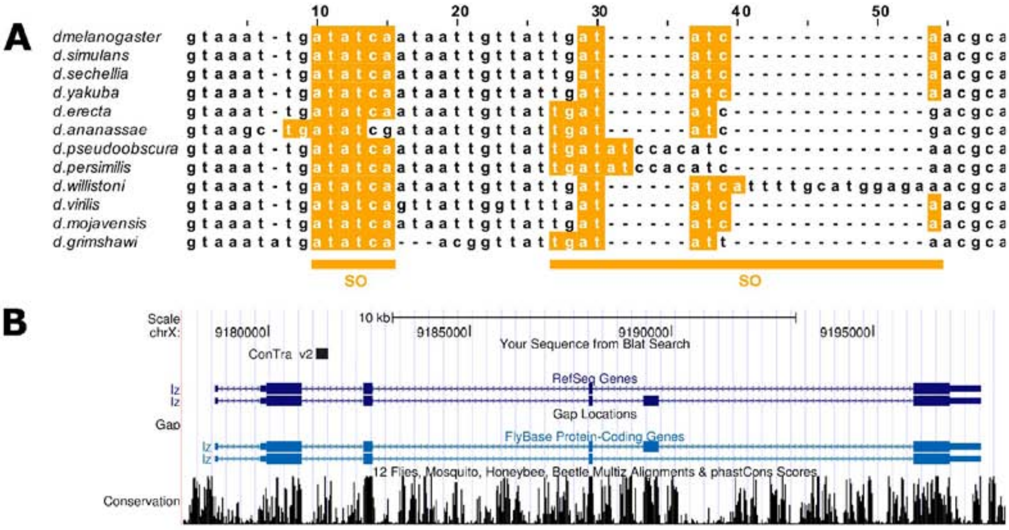

A
10
20
30
40
50
dmelanogaster gtaaat-tgatatcaataattgttattgat------atc--------------aacgca
d.simulans gtaaat-tgatatcaataattgttattgat------atc--------------aacgca
d.sechellia gtaaat-tgatatcaataattgttattgat------atc--------------aacgca
d.yakuba gtaaat-tgatatcaataattgttattgat------atc--------------aacgca
d.erecta gtaaat-tgatatcaataattgttattgat------atc--------------gacgca
d.ananassae gtaagc-tgatatcgataattgttattgat------atc--------------gacgca
d.pseudoobscura gtaaat-tgatatcaataattgttattgatatccacatc--------------aacgca
d.persimilis gtaaat-tgatatcaataattgttattgatatccacatc--------------aacgca
d.willistoni gtaaat-tgatatcaataattgttattgat------atcattttgcatggagaaacgca
d.virilis gtaaat-tgatatcagttattggttttaat------atc--------------aacgca
d.mojavensis gtaaat-tgatatcaataattgttattgat------atc--------------aacgca
d.grimshawi gtaaatatgatatca---acggttattgat------att--------------aacgca
SO
SO
B
Scale
chrX:
10 kb
9180000
9185000
9190000
9195000
Your Sequence from Blat Search
ConTra v2
RefSeq Genes
lz
lz
Gap Locations
Gap
FlyBase Protein-Coding Genes
lz
lz
12 Flies, Mosquito, Honeybee, Beetle Multiz Alignments & phastCons Scores
Conservation

Figure A.7: Visualization of the evolutionary conserved Sine Oculis (SO) binding site in the second intronic region of the fruitfly transcription factor Lozenge (LZ), involved in the prepatterning of photoreceptor precursors in the developing Drosophila eye as described by Yan et al [70]. (A) Multiz alignment showing the conserved SO site (in orange), predicted using the JASPAR positional weight matrix MA0246.1. (B) Region of (A) was mapped using BLAT on intron 2 in the UCSC Genome Browser (black box). Blue boxes represent exons with the blue arrows illustrating the intronic regions.

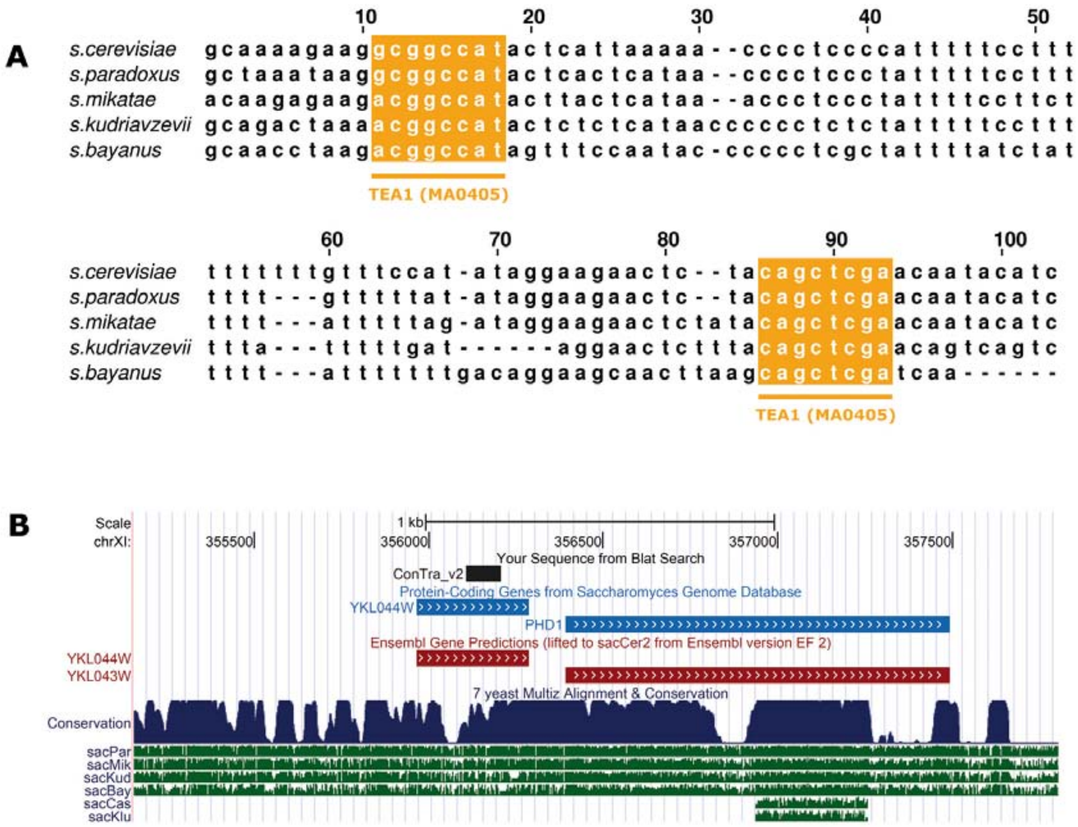

A
10
20
30
40
50
s.cerevisiae gcaaaagaaggcggccatactcattaaaaa--cccctccccattttttccttt
s.paradoxus gctaaataaggcggccatactcactcataa--cccctccctattttttccttt
s.mikatae acaagagaagacggccatacttactcataa--accctccctattttccttct
s.kudriavzevii gcagactaaaacggccatactctctcataacccccctctctattttttccttt
s.bayanus gcaacctaagacggccatagtttccaatac-cccccctcgctattttatctat
TEA1 (MA0405)
60
70
80
90
100
s.cerevisiae tttttttgtttccat-ataggaagaactc--tacagctcgaacaatacatc
s.paradoxus tttt---gtttttat-ataggaagaactc--tacagctcgaacaatacatc
s.mikatae tttt---atttttag-ataggaagaactctatacagctcgaacaatacatc
s.kudriavzevii ttta---tttttgat------aggaactctttacagctcgaacagtcagtc
s.bayanus tttt---atttttttgacaggaagcaacttaagcagctcgatcaa------
TEA1 (MA0405)
B
Scale
1 kb
chrXI:
355500
356000
356500
357000
357500
Your Sequence from Blat Search
ConTra_v2
Protein-Coding Genes from Saccharomyces Genome Database
YKL044W
PHD1
Ensembl Gene Predictions (lifted to sacCer2 from Ensembl version EF 2)
YKL044W
YKL043W
7 yeast Multiz Alignment & Conservation
Conservation
sacPar
sacMik
sacKud
sacBay
sacCas
sacKlu

Figure A.8: Visualization of the evolutionary conserved TEA1 binding site in the promoter region of the S. cerevisiae Phd1 (Flo11) gene as described by Heise et al [71]. (A) Multiz alignment showing the conserved TEA1 sites (in orange), predicted using the JASPAR positional weight matrix MA0405.1. (B) Region of (A) was mapped using BLAT on the promoter in the UCSC Genome Browser (black box). Blue and red boxes represent genes.