International
Max Planck Research School
for Computational Biology
and Scientific Computing

Freie Universität Berlin

Analysis of Long-Distance Gene Regulatory Elements

Jonathan Göke

Dissertation zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Betreuer: Prof. Dr. Martin Vingron

Berlin, Mai 2012

Erstgutachter: Prof. Dr. Martin Vingron

Zweitgutachter: Prof. Dr. James Adjaye

Tag der Disputation: 6. Juli 2012

To the memory of Hannes Luz.

Acknowledgements

This dissertation would not have been possible without the guidance and the help of many people who in one way or another contributed to the preparation and completion of this study.

First and foremost, I would like to thank my supervisor, Martin Vingron, for his support, suggestions and ideas throughout my time in his group. I am especially grateful to him for encouraging me in the difficult phases of my research. I would like to thank James Adjaye for the collaboration that aroused my interest in stem cell research. This work would not have been possible without his invaluable feedback and enthusiasm. Deepest gratitude are also due to the members of the supervisory committee, Knut Reinert and Stefan Haas from whose knowledge and assistance this study greatly profited. It is an honour for me to be part of the International Max Planck Research School for Computational Biology and Scientific Computing and I would like to thank Kirsten Kelleher and all my fellow Ph.D. students.

This thesis would not have been possible without the help of many people, in particular I would like to thank Marcel Schulz and Julia Lasserre for their invaluable contribution to this work and to the publication that resulted from this research (Göke *et al.*, 2012). I want to thank Marcel Schulz for guiding me during my first year and for all the time and effort that he invested throughout the development of the alignment-free sequence comparison method $N2$. I am very grateful to Julia Lasserre for her help, assistance and detailed feedback in particular for Chapter 3 and Section 4.3. Furthermore, I owe my deepest gratitude to all the people that have been involved in this work. I would like to thank Marc Jung for performing the OCT4 ChIP-Seq experiment and for his feedback on the manuscript, and Sarah Behrens for her help with data processing, her feedback on the manuscripts and on her assistance with word statistics. I would like to thank David Weese and Manuel Holtgrewe for their assistance with SeqAn, I feel honoured to be the first external contributor. I would like to thank Morgane Thomas-Chollier for helpful discussion and providing tools for analysis of ChIP-Seq data, Sean O'Keeffe for his help with visualisation of genomic data and Lukas Chavez and Bernd Timmermann for their contribution to the manuscript. I would like to thank Sebastiaan Meijsing for comments on the initial manuscript, Matthew Huska for proofreading, and Huck Hui

Ng for ideas and suggestions on the analysis of genome-wide binding data in embryonic stem cells. My special thanks go to my office mates Akdes Serin, Alena Myšičková and Christian Rödelsperger, whose company and friendship I will always remember. In my daily work I have been blessed with a wonderful group of fellow students and colleagues and I want to thank Mike Love, Annalisa Marsico, Yves Clement, Hugues Richard, Juliane Perner, Brian Cusack, Matthias Heinig, Ho-Ryun Chung, Ruping Sun, Anne-Katrin Emde, Christian Diener, Federico Squartini, Paz Polak, Rosa Karlic, Ewa Szczurek, Marta Luksza, Roman Brinzanik, Holger Klein, Roland Krause and all the current and past members of the Vingron Department which I forgot to mention.

Finally, my warmest gratitude goes to my parents, and to Lukas, Sebastian and Tina.

Gewidmet Hannes Luz, in Erinnerung.

Contents

1 Introduction 1

1.1 Outline of the Thesis . . . 3

1.2 Background: Introduction to Molecular Genetics . . . 5

1.2.1 Replication, Transcription and Translation of Genetic Information 6

1.2.2 Structure of the DNA in the Cell . . . 7

1.3 Regulation of Gene Expression . . . 8

1.3.1 Epigenetic Regulation of Gene Expression . . . 9

1.3.2 DNA Looping and Chromatin Architecture . . . 10

1.4 Identification of Transcription Factor Binding Sites . . . 11

1.4.1 Computational Approaches . . . 12

1.4.2 Experimental Approaches . . . 12

2 Combinatorial Binding at Enhancers in ESCs 15

2.1 Introduction . . . 15

2.1.1 Methods I: Overview of the Data Used in this Study . . . 17

2.1.2 Methods II: Data Processing . . . 20

2.2 Results . . . 21

2.2.1 Combinatorial Binding and Transcription . . . 22

2.2.2 Combinatorial Binding and Evolution . . . 26

2.2.3 Conserved Combinatoriallly Bound Loci are Active in Development 28

2.2.4 Gene Regulatory Hotspots . . . 31

2.3 Discussion . . . 33

2.4 Conclusion . . . 36

3 Pairwise Comparison of Enhancer Sequences 37

3.1 Introduction . . . 37

3.2 Background: Word Statistics . . . 39

3.2.1 Background Models for Biological Sequences . . . 40

3.2.2 Alignment-Free Sequence Comparison Methods . . . 47

3.3 The N2 Similarity Score . . . 51

3.3.1 Definition: N2 . . . 51

3.3.2 Word Statistics for Word Neighbourhood Counts 52
3.3.3 Implementation, Instances, and Availability of N2 55
3.4 Results 58
3.4.1 Choice of Parameters for N2 58
3.4.2 N2 can be Computed Quickly 59
3.4.3 N2 is Robust Against Single Sequence Noise 60
3.4.4 Pairwise Comparison: Simulations 62
3.4.5 Pairwise Comparison of Developmental Enhancers 64
3.5 Discussion 66
3.5.1 Conclusion 68

4 Large-Scale Analysis of Enhancer Sequences 71
4.1 Introduction 71
4.2 Motif Finding: N2 Word Ranking 73
4.2.1 Algorithm: ALF-M 73
4.2.2 Motif Finding: Simulations 76
4.2.3 Motif Finding: Enhancers 77
4.2.4 ALF-M Identifies Conserved Motifs in Developmental Enhancers 78
4.2.5 Discussion 80
4.3 Classification: The N2 Kernel Function 81
4.3.1 The N2 Kernel Function 82
4.3.2 Classification: Simulations 83
4.3.3 Prediction of Developmental Enhancers with N2 84
4.3.4 Prediction of the Epigenetic State of Enhancers 86
4.3.5 Discussion 87
4.4 Clustering: N2 Similarity Measure 89
4.4.1 Simulations 90
4.4.2 Clustering of Developmental Enhancers 92
4.4.3 Discussion 96

5 Summary 97

VI Appendix 101
VI.1 Supplementary Information for Chapter 2 101
VI.2 Supplementary Information for Chapter 3 105
VI.3 Supplementary Information for Chapter 4 110

Bibliography 113

List of Abbreviations 127

Notations 129

Glossary 131

Availability 133

Zusammenfassung 135

Ehrenwörtliche Erklärung 137

List of Tables

1.1 Histone Modifications 10

2.1 Mapping Statistics 18

3.1 Running Time in O-Notation 60
3.2 Simulations with Implanted k-mers, Random Strand, $k = 6$ 63
3.3 Simulations with Implanted k-mers, Mismatch, $k = 6$ 64
3.4 Comparison of Different Methods on Tissue-Specific Enhancers 65

4.1 Classification of Tissue-Specific Enhancers 84

VI.1 Simulations with Implanted k-mers, Random Strand, $k = 5$ 105
VI.2 Simulations with Implanted k-mers, Mismatch, $k = 5$ 106
VI.3 Influence of Repeats 106

List of Figures

1.1 Mammalian Cell Types 1
1.2 DNA 5
1.3 Replication, Transcription, and Translation 6
1.4 Chromatin Structure 7
1.5 Regulation of Gene Expression 9
1.6 Modelling Transcription Factor Binding Sites 11
1.7 ChIP-Sequencing 14

2.1 Combinatorial Binding at the Sox2 Locus 19
2.2 Co-localisation of Proteins at the DNA 21
2.3 Mediator Co-localises at Combinatorially Bound Loci 22
2.4 Combinatorial Binding at Enhancers and Promoters 24
2.5 Functional Analysis of Combinatorially Bound Enhancers 25
2.6 Combinatorial Binding and Evolution 27
2.7 Combinatorial Binding Identifies Developmental Enhancers 29
2.8 Conservation of Gene Regulatory Hotspots 31
2.9 The Gene Regulatory Hotspot Downstream of *SOX21* 32
2.10 Gene Regulatory Hotspots 34

3.1 Statistics for Word Counts 43
3.2 Markov Models for DNA Sequences 44
3.3 Dotplot 48
3.4 Running Time Comparison 59
3.5 Influence of Single Sequences on Pairwise Scores. 61
3.6 Precision-Recall Curve for Pairwise Comparison of Enhancers, $k = 6$. 64
3.7 Tissue-Specificity of Enhancers, $k = 6$ 66

4.1 Motif Finding with N2: ALF-M 75
4.2 ALF-M Motif Finding, Simulated Sequences 76
4.3 ALF-M Motif Finding, Enhancer 77
4.4 ALF-M Identifies Conserved Motifs 79
4.5 Classification with $N2$, Simulations 83

4.6 Classification of Embryonic Enhancers with $N2$ 85
4.7 Prediction of the Epigenetic State of Enhancers 86
4.8 Clustering of Simulated Sequences with $N2$ and k-means 90
4.9 Hierarchical Clustering of Simulated Sequences with $N2$ 91
4.10 Clustering of Embryonic Heart Enhancers 92
4.11 Hierarchical Clustering identifies Homogeneous Groups of Enhancers . 93
4.12 Dotplot of Sequences From Two Clusters 94
4.13 Clustering of Mouse Developmental Enhancers 95

VI.1 Comparison of Different Cutoffs for Peak Calling 101
VI.2 Combinatorial Binding and Mediator Co-localisation ('stringent cut-off') 102
VI.3 Combinatorial Binding and Conservation ('stringent cut-off') 103
VI.4 Combinatorial Binding at Developmental Enhancers ('stringent cut-off') 104
VI.5 Integrating Data from Multiple Cell Lines Identifies Functional Binding 105
VI.6 Influence of Single Sequences on Pairwise Scores, $N2$ Variants 107
VI.7 Influence of Mismatch Weights a_w on Pairwise Scores, Simulations . . 107
VI.8 Influence of Mismatch Weights a_w on Pairwise Scores, Enhancers . . . 108
VI.9 Precision-Recall Curve, Enhancers, $k = 4$ 108
VI.10 Precision-Recall Curve, Enhancers, $k = 5$ 108
VI.11 Tissue-Specificity of Enhancers, $k = 5$ 109
VI.12 Classification with $N2$, Simulations 110
VI.13 Classification of Embryonic Enhancers with $N2$ (M,L,H) 111
VI.14 Classification of Embryonic Enhancers with $N2$, $k = 6$ 112

1 Introduction

The diversity of life is amazing: It is estimated that there are more than 10 million living species on Earth today (Alberts *et al.*, 2002). Even though all these species appear very different, they all share one common property: heredity. The parent organism passes on all the information that accounts for the characteristics of the offspring. Heredity is the basis for evolution, the process of change that is shaping the diversity of species we observe today.

All living organisms consist of cells, sometimes one, sometimes many millions. Mammalian organisms, like human or mouse, consist of hundreds of different cell types which build the tissues. These cells are very different in morphology and function (Figure 1.1). For example, neuronal cells can span more than a meter, whereas cells of the immune system are usually less than 100 μm in size (Alberts *et al.*, 2002). Nevertheless, their genetic basis is largely identical: all cells that form the adult organism originate from one single cell, the fertilised oocyte (zygote). The zygote contains the information handed down from the parents that defines the development of the offspring. Every time the cell divides, this genetic information is passed on to the daughter cells. Therefore, all cells of a living organism, whether they are neurons or immune cells, share the same genetic information.

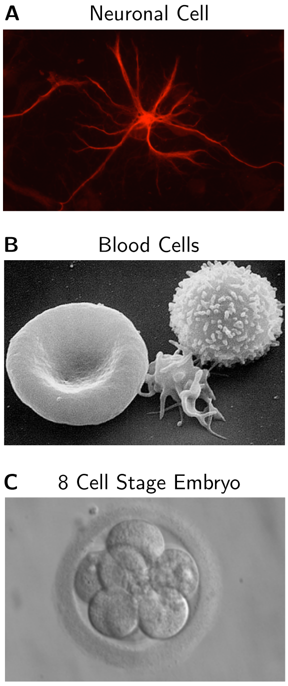


Figure 1.1: (A) Neuronal cell (Endo *et al.*, 2009) **(B)** Blood cells (Wikimedia Commons, 2012c) **(C)** 8 cell stage embryo (nature.com, 2012)

One key question is how the same genotype can give rise to this large variety of cell types. It is known that the cell tightly controls which part of the genetic information is active for every specific cell through regulation of gene expression. Many different factors govern gene expression, amongst others external factors, the cellular milieu or DNA accessibility (Coller and Kruglyak, 2008). The importance of the different influences on gene expression was studied using a mouse model of what is known as Down syndrome in humans, a disease where the affected carries an additional copy of chromosome 21 (Wilson *et al.*, 2008). However, instead of having three copies of the same chromosome, the researchers used a mouse strain carrying a human chromosome 21. This model enabled them to determine on a large scale, whether inter-species differences in transcriptional regulation are primarily directed by human genetic sequence or mouse nuclear environment. Strikingly, they found that in homologous tissues, genetic sequence is largely responsible for directing transcriptional programs, whereas inter-species differences seemed to play a secondary role. This experiment demonstrates that the information about when and where genes are active is largely written into the DNA.

The genetic 'switches' that regulate gene expression are recognised and interpreted by sequence-specific DNA binding proteins, the transcription factors. A large fraction of transcription factor binding sites is located at long distances to the gene that is regulated (*enhancers*, Tjian and Maniatis (1994); Heintzman and Ren (2009)). Long-distance gene regulation is crucial for correct gene expression and facilitates the formation of different cell types during development which all originated from the same, single cell (Maston *et al.*, 2006). This thesis combines experimental data with computational and statistical methods to study the properties and characteristics of long-distance gene regulatory elements in mammalian cells.

1.1 Outline of the Thesis

Chapter 1: Introduction and Background to Molecular Genetics

This chapter reviews the basics of molecular genetics and gives an overview of computational and experimental methods for identification of gene regulatory elements.

Chapter 2: Combinatorial Binding at Enhancers in Embryonic Stem Cells

In this chapter, genome-wide binding data of transcription factors and co-factors is integrated to study the influence of combinatorial binding at long-distance enhancers on transcription and evolution of gene regulation. This work was published in 2011 (Göke *et al.*, 2011).

Chapter 3: Alignment-Free Pairwise Comparison of Enhancer Sequences

Here, a novel alignment-free method, $N2$, is presented, which measures the pairwise sequence similarity of regulatory sequences, analogous to alignments for protein-coding sequences. $N2$ is applied to tissue-specific mammalian developmental enhancers. The method was published in 2012 (Göke *et al.*, 2012).

Chapter 4: Large-Scale Analysis of Developmental Enhancer Sequences

In contrast to Chapter 3 which is restricted to the case of pairwise sequence comparison, this chapter aims at analysing large-scale enhancer data sets. The $N2$-based word statistics are utilised to study sequence-specific properties of developmental enhancers. First, a motif finding algorithm is presented (ALF-M). Second, $N2$ is used as a kernel function to classify and predict regulatory potential of DNA sequences. Finally, $N2$ is used to study the heterogeneity of tissue-specific enhancer data sets.

Chapter 5: Summary

This chapter provides a brief summary of the thesis.

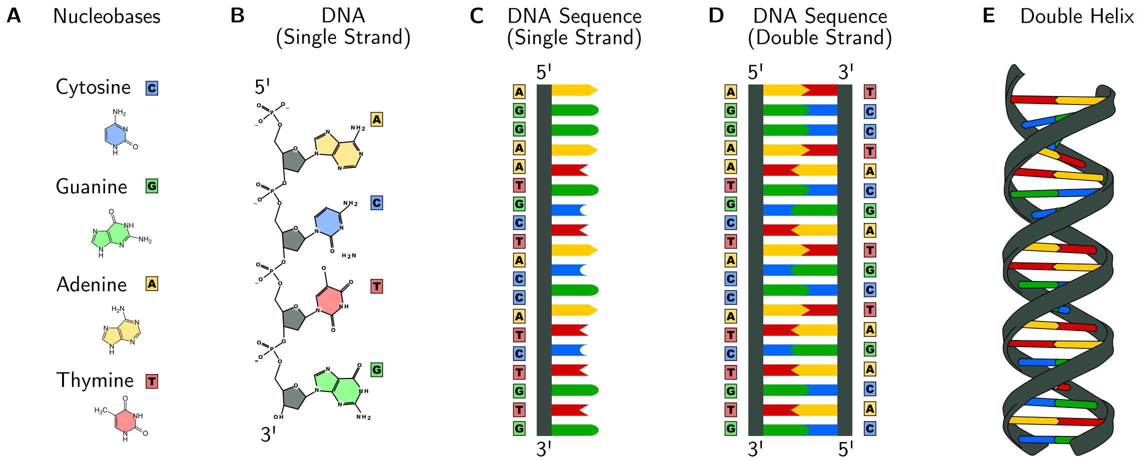

A Nucleobases
Cytosine C
Guanine G
Adenine A
Thymine T
B DNA (Single Strand)
5'
A
C
T
G
3'
C DNA Sequence (Single Strand)
5'
A G G A A T G C T A C C A T C T G T G
3'
D DNA Sequence (Double Strand)
5' 3'
A G G A A T G C T A C C A T C T G T G
T C C T A C G A T G C T A G A C A C
3' 5'
E Double Helix

Figure 1.2: The structure of DNA sequences. **(A)** The basic unit of DNA sequences are the nucleobases, Adenine (`A`), Cytosine (`C`), Guanine (`G`) and Thymine (`T`). **(B)** DNA is a polymer of nucleotides, nucleobases connected by a sugar phosphate backbone. **(C)** Schematic view of a single strand DNA sequence. **(D)** Schematic view of a double stranded DNA sequence. The two strands are connected by base pairing and form reverse complementary sequences. **(E)**. DNA forms a double helical structure in the cell. Illustrations A and E are based on Wikimedia Commons (2012h), B is based on Wikimedia Commons (2012f).

1.2 Background: Introduction to Molecular Genetics

The hereditary information that is passed on from parents to offsprings and from a parent cell to daughter cells is stored in the deoxyribonucleic acid (DNA). The DNA is a pair of long polymer chains of smaller subunits, the nucleotides. A nucleotide is composed of a sugar and phosphate backbone and one of the four nucleobases: Adenine (`A`), Cytosine (`C`), Guanine (`G`) and Thymine (`T`) (Figure 1.2A). These nucleotides are attached to each other in a strictly linear fashion forming the sequence that encodes the genetic information (Figure 1.2B-C). This code is universally readable, so that bacterial DNA for example, which is inserted into a human cell will be correctly interpreted, and vice versa (Alberts *et al.*, 2002). The genome, the entirety of all the hereditary information which is necessary to form a living organism, is thus encoded as a long sequence of `A`s, `C`s, `G`s and `T`s.

The polymer of nucleotides forms a single DNA strand. In mammalian cells, every DNA molecule consists of two such strands, the forward and the backward strand. Both strands are connected by base pairing, `A` in one strand always binds to `T` on the other strand and a `C` binds to a `G` (Figure 1.2D). The two DNA strands therefore form complementary sequences. These complementary DNA strands are tightly twisted around each other, forming a double helix (Figure 1.2E).

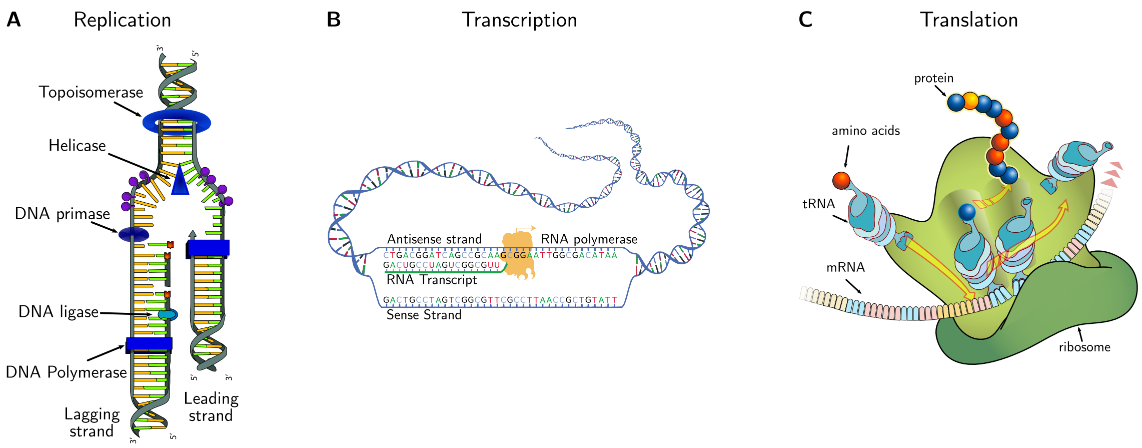


Figure 1.3: (A) During the process of replication, the DNA is unwound and copied. Replication of DNA is required to pass on the genetic information to the daughter cells. **(B)** During transcription, the information encoded in the DNA is read and RNAs are synthesised according to the genetic sequence. **(C)** Translation is the process that synthesises proteins from mRNAs. Proteins are the major functional molecules in the cell and required for almost all biological processes. Illustration A is based on Wikimedia Commons (2012g), B is based on Wikimedia Commons (2012a), C is based on Wikimedia Commons (2012e).

1.2.1 Replication, Transcription and Translation of Genetic Information

The power of this double helical structure becomes apparent during the process that copies the genetic information: *replication* (Figure 1.3A). Every time a cell divides, the DNA is exactly replicated and transmitted to the daughter cells. First, the double helix is unwound so that both strands become accessible. Then the cellular machinery uses these single strands to synthesise their complementary copies. This way, the original double stranded DNA is replicated. Replication and DNA synthesis is common to all living organisms and forms the basis for inheritance of genetic information.

The genetic information which is stored in the DNA is read through a process called *transcription* (Figure 1.3B). Specific sequences in the genome, the genes, are recognised by the cellular transcriptional machinery. The DNA of these genes is then used as a template to synthesise ribonucleic acid (RNA). Similar to DNA, RNAs are polymers of nucleotides with a slightly different backbone (ribose instead of deoxyribose) and a different alphabet (`A`, `C`, `G`, and `U` (Uracil) instead of `T`). RNAs are exact copies from DNA sequences using a slightly different language: `A`, `C`, `G` and `T` from the DNA are transcribed into the complementary `U`, `G`, `C` and `A` of the RNA. These RNAs are much smaller and flexible and provide the first step in interpreting the genetic information encoded in the DNA. The enzyme that catalyses transcription of protein-coding genes

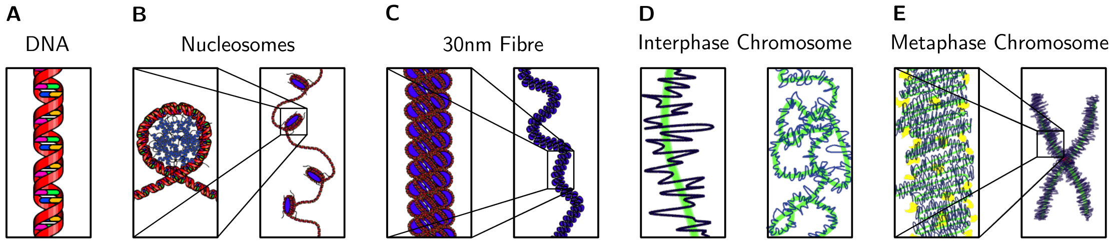


Figure 1.4: Chromatin Structure.**(A)** DNA double helix. **(B)** DNA is wrapped around nucleosomes. **(C)** Cellular DNA is packaged into a 30 nm fibre. **(D)** Chromatin during interphase. **(E)** During metaphase, the densely packaged chromosomal structures known from karyotypes can be observed. Illustration is based on Wikimedia Commons (2012b)

and several small RNAs is RNA polymerase II (POLII). POLII interacts with many different proteins at the transcription start site (TSS) in order to initiate and elongate transcription (Figure 1.5). POLII forms the core of the basal transcriptional machinery that is needed in every cell type to maintain active transcription.

RNAs can have a variety of functions and there are many different classes of RNAs. The best studied class of RNAs are the protein-coding messenger RNAs (mRNAs). During a process called *translation*, proteins are synthesised according to the mRNA sequence (Figure 1.3C). Similar to DNA and RNA, proteins are polymers of smaller subunits, the amino acids. A sequence of three nucleotides corresponds to exactly one amino acid. This way, the mRNA sequence is translated into a sequence of amino acids. These amino acid chains build three dimensional structures, the proteins, which are able to fulfil a large variety of functions. Proteins catalyse the large majority of chemical processes in the cell, they participate in all major pathways and ultimately decide the phenotype of all cells. Proteins are therefore the molecules that carry out the function encoded in the DNA.

1.2.2 Structure of the DNA in the Cell

The DNA can be displayed as a sequence of nucleotides, however, to understand molecular genetics, the structure and cellular organisation of the DNA is important. In mammalian cells, the DNA is located in the nucleus. The genetic information is distributed to different DNA molecules, the chromosomes. The human genome consists of 24 chromosomes. 22 of these chromosomes can be found in every human cell (autosomes) and two chromosomes (X and Y) are sex-specific (sex chromosomes). Every somatic human cell contains 46 chromosomes, two copies of every autosome and two sex chromosomes (diploid cells). The cells of the germ line are haploid, they contain a single set of 23 chromosomes.

Stretched out from end to end, the DNA of the smallest human chromosome (chromosome 22) would extend about 1.5 *cm* (Alberts *et al.*, 2002). The average diameter of a mammalian nucleus is approximately 6 micrometers (Alberts *et al.*, 2002). Clearly, cellular DNA needs to be compressed in an organised manner to facilitate controlled replication and transcription and to avoid damage. Indeed, chromosome 22 measures only about 2 μm in its most compact form in the cell. The protein-DNA complex that is responsible for DNA packaging is called chromatin (Figure 1.4).

The first level of DNA packaging is achieved by the histone proteins. Approximately 147 base pairs (bp) of DNA are wrapped around a set of 8 histones forming the nucleosome, the basic unit of DNA packaging in the cell (Figure 1.4A,B). Nucleosomes are connected by a 60 to 80 bp long linker DNA, such that the DNA is organised in a chain of Nucleosomes ("beads on a string"). In the cell, the DNA is further condensed into a 30 nm fibre (Figure 1.4C). The accessibility of the DNA is regulated by the level of condensation. In an 'open chromatin' structure, the DNA is accessibly by DNA binding proteins and genes can be transcribed. The active chromatin formation is referred to as 'euchromatin'. In contrast, tightly condensed, inaccessible, and transcriptional inactive chromatin is referred to as 'heterochromatin'. Both euchromatin and heterochromatin are local structures, this way the same chromosomes can have both accessible and inaccessible DNA. The highest level of DNA packaging into the most compact form can only be observed at specific stages during cell division when the chromosomes form the typical structure observed in karyotypes (Figure 1.4E).

1.3 Regulation of Gene Expression

Gene expression is the process that begins with reading the genetic information and leads to the synthesis of a functional gene product. The process from reading the genetic information on the DNA to the synthesis of a functional gene product is called gene expression. Changes in gene expression can lead to cell division, differentiation or proliferation. Tight regulation of gene expression is therefore crucial to ensure correct embryonic development, but it is also involved in almost all physiological processes in adults. Even minor errors in transcriptional regulation can lead to severe misbuildings and diseases, such as cancer, heart failure or developmental disorders (Kleinjan and van Heyningen, 2005).

The large diversity of transcripts and gene expression patterns that leads to the formation of different cellular phenotypes is obtained through cell type-specific transcription factors (Maston *et al.*, 2006; Coller and Kruglyak, 2008; Wilson *et al.*, 2008). These transcription factors recognise specific nucleotide sequences in order to regulate gene expression in *cis*. Such *cis*-regulatory sequences can be proximal (*promoter*) or many

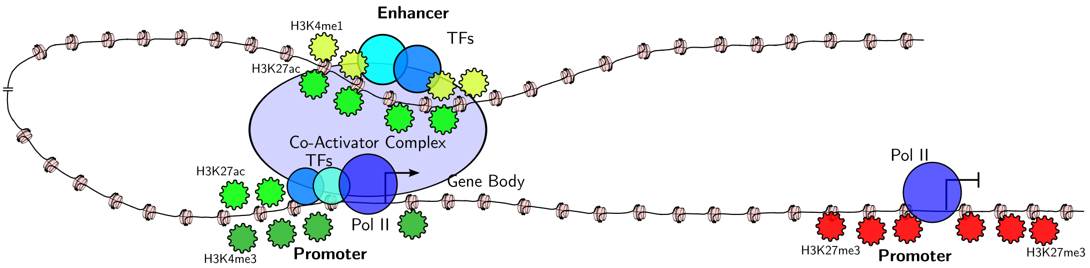


Figure 1.5: Regulation of Gene Expression. Genes are controlled though proximal (*promoter*) and distal (*enhancer*) regulatory elements. The promoter on the left is activated through binding of sequence specific transcription factors and interaction with a distant enhancer, leading to active transcription of the gene. Co-activator complexes such as p300 and Mediator and chromatin remodelling complexes such as Cohesin connect distal with proximal regulatory elements. Histones near actively transcribed genes frequently show H3K4me3 at the promoter and H3K36me3 at the gene body. Enhancers are marked by H3K4me1 or H3K27ac. Transcription can be silenced by proteins which mediate repressive histone modifications at enhancers or promoters through recruitment of histone deacetylases and methyltransferases. The promoter on the right shows such repressive histone marks (H3K27me3) which ensure that the gene is silenced. See Sakabe and Nobrega (2010) for a review.

kilo bases distant (*enhancer*) to the TSS of the gene which is regulated (Tjian and Maniatis (1994); Heintzman and Ren (2009), Figure 1.5) .

Transcription factors form a very divergent protein family, many thousand genes encode for such proteins in the human genome (Vaquerizas *et al.*, 2009). Every transcription factor has a DNA-binding domain which recognises a specific DNA sequence (DNA *motif*). Transcription factors contain additional domains to integrate external signalling, interact with transcriptional co-activators or chromatin modulator complexes to initiate, enhance or repress transcription.

1.3.1 Epigenetic Regulation of Gene Expression

The word epigenetics summarises modifications which influence the cellular phenotype and which can be inherited through cell division without changing the DNA sequence (Berger *et al.*, 2009). Epigenetic modifications that influence gene expression are DNA methylation and modifications of histone proteins. DNA methylation involves methylation of a CpG dinucleotide which ensures long-term, almost irreversible gene silencing (Bird, 1986; Chavez *et al.*, 2010). In contrast, modification of histone proteins are reversible and highly dynamic.

Epigenetics: 'An epigenetic trait is a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence.'(Berger *et al.*, 2009)

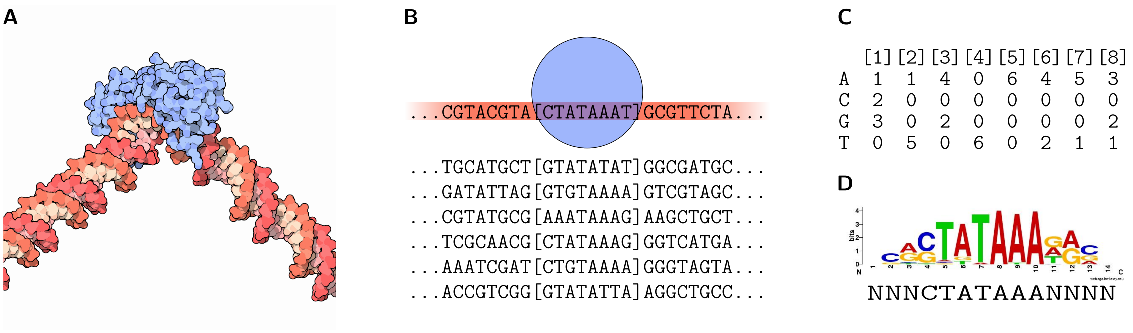
Modification	Location	Associated transcriptional activity
H3K4me3	Promoter	Active
H3K4me1	Enhancer	Active
H3K27ac	Enhancer, Promoter	Active
H3K27me3	Promoter	Repressed
H3K36me3	Gene body	Active

Table 1.1: Histone modifications at regulatory elements in mammalian cells. Active enhancers are marked by H3K27ac and H3K4me1, active promoters are marked by H3K4me3. Promoters and enhancers near repressed genes show increased levels of H3K27me3. H3K36me3 is typically enriched at the gene body of transcribed genes.

Histones are the building blocks of the nucleosomes, the structures that are responsible for chromatin assembly and condensation (see Section 1.2.2). Nucleosomes consist of two copies each of the four core histone proteins H2A, H2B, H3 and H4. These histone proteins can be post-transcriptionally modified at their N-terminus (histone tail). The best studied modifications are acetylation and methylation of specific lysine residues (Table 1.1). Enhancers and promoters can be marked by different histone modifications, dependent on their transcriptional activity (Figure 1.5). Active enhancers are usually marked by monomethylation of H3 Lysine 4 (H3K4me1) and acetylation of H3 Lysine 27 (H3K27ac). Active promoters are marked by trimethylation of H3 lysine 4 (H3K4me3). Promoters near genes which are repressed are frequently marked by trimethylation of H3 lysine 27 (H3K27me3). Promoters which are marked by H3K4me3 and H3K27me3 at the same time are called bivalent domains (Bernstein *et al.*, 2006). Bivalent domains occur in early stages of embryogenesis near genes involved in developmental processes and cell fate determination (Bernstein *et al.*, 2006). Recent literature also suggested that enhancers can be marked by H3K27me3 (Rada-Iglesias *et al.*, 2011). These enhancers were termed poised enhancers, as they seem to be poised for activation after loss of this histone modification. Transcriptional regulation is tightly linked to epigenetic modifications of histone tails. The proteins that catalyse these reactions interact with sequence-specific transcription factors and the basal transcriptional machinery (Kouzarides, 2007). Similarly, chromatin accessibility and binding of transcription factors are influenced by the epigenetic state of regulatory elements (Kouzarides, 2007).

1.3.2 DNA Looping and Chromatin Architecture

Currently, it is assumed that binding of specific transcription factors at distant enhancers and at the promoter, and their interaction with co-activators and chromatin modulators



```
...CGTACGTA[CTATAAAT]GCGTTCTA...

...TGCATGCT[GTATATAT]GGCGATGC...
...GATATTAG[GTGTAAAA]GTCGTAGC...
...CGTATGCG[AAATAAAG]AAGCTGCT...
...TCGCAACG[CTATAAAG]GGTCATGA...
...AAATCGAT[CTGTAAAA]GGGTAGTA...
...ACCGTCGG[GTATATTA]AGGCTGCC...
```

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

Figure 1.6: Modelling Transcription Factor Binding Sites. **(A)** The transcription factor TBP (blue) binds to a specific sequence (TATA) in the DNA (red) (Figure from Goodsell (2005)). **(B)** Transcription factors bind the DNA in a stochastic manner, which enables the recognition of different DNA sequences that resemble a core structure. **(C)** Binding motifs of transcription factors can be described using position frequency matrices. The columns contain the nucleotide frequencies for every position, as obtained from experimental data. **(D)** Sequence logo visualisation of the DNA binding motif of the TATA-box binding transcription factor with the consensus sequence displayed below.

is required to bring the DNA of enhancers into close proximity to the TSS in order to facilitate cell type-specific gene expression. The link between distant enhancers and the promoter can be provided by co-activator proteins, such as p300 or the Mediator complex (Kagey *et al.*, 2010). Transcription factors interact with these co-activators, which in turn interact with the basal transcriptional machinery. Chromatin loops are established through interaction of Mediator with the Cohesin complex which can form rings to connect DNA segments (Kagey *et al.*, 2010). The interaction of general co-activator complexes with transcription factors ensures cell type-specific DNA looping and accordingly cell type-specific regulation of gene expression.

1.4 Identification of Transcription Factor Binding Sites

One of the primary step towards analysing and understanding *cis*-regulatory sequences is the identification of transcription factor binding sites in the genome (Vingron *et al.*, 2009). Binding of transcription factors at the DNA occurs in a stochastic manner, dependent on the biophysical properties of the interaction of the binding domain and the DNA sequence (von Hippel and Berg, 1986; Roider *et al.*, 2007). This stochastic binding leads to the effect that every transcription factor is able to bind to a variety of DNA sequences that resemble a core structure (Figure 1.6A-B). The DNA binding motif of transcription factors can be described using a position frequency matrix, which summarises the nucleotide counts at every position of the binding motif (Figure 1.6C). This matrix can be transformed into a position weight matrix (PWM), for example using log-odd ratios.

The DNA binding motif of the transcription factor can be visualised as a sequence logo (Schneider and Stephens (1990), Figure 1.6D). Sequence logos show the frequency of every nucleotide at every position. The height displays the information content, indicating the flexibility of the motifs (Schneider *et al.*, 1986).

1.4.1 Computational Approaches

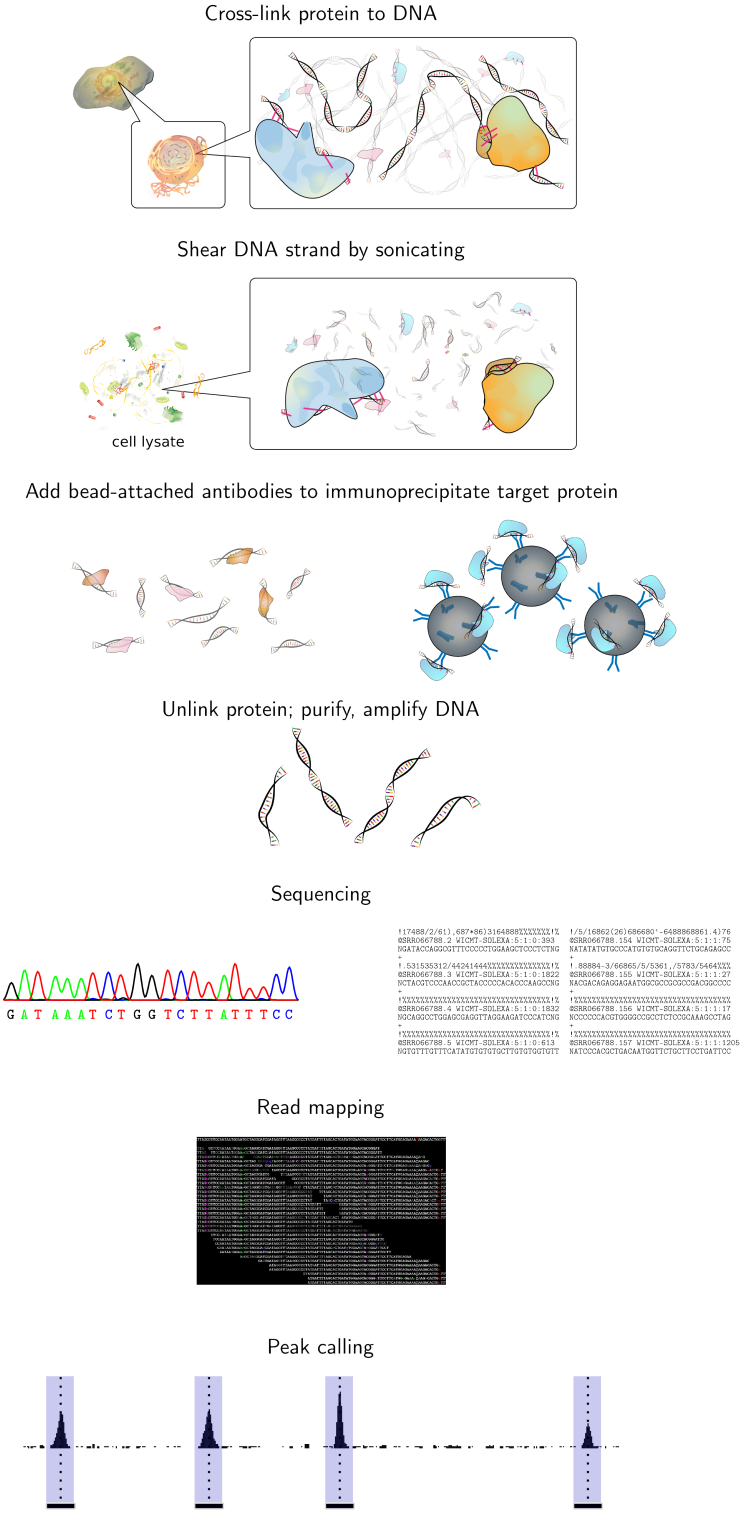
The binding affinity of a transcription factor to the DNA can be estimated by the similarity of the nucleotide sequence to the known binding motif. This similarity can be used to estimate the regulatory potential of genomic DNA and to predict *cis*-regulatory sequences. Depending on the motif, a transcription factor is expected to have a high affinity binding site approximately every thousand base pairs. However, *in vivo* only a minority of potential binding sites are bound by the transcription factor. Chromatin accessibility, co-factor binding, or protein-protein interactions ensure the correct, cell type-specific binding of transcription factors in the cell. Since sequence-based approaches are independent from the cell type, they are limited in their ability to predict genome-wide locations.

1.4.2 Experimental Approaches

Mapping of genome-wide DNA-protein interactions became possible with the availability of next generation sequencing technologies, in particular in combination with large scale chromatin immunoprecipitation (ChIP-Seq; Johnson *et al.* (2007); Mardis (2007); see Figure 1.7 for the detailed work-flow). In the first step, the DNA is experimentally fixed to the proteins that are bound. Secondly, the DNA is sheared into small pieces which are still cross-linked to the bound proteins. The protein of interest is then immunoprecipitated using a specific antibody. The DNA that is bound by the transcription factor is purified, amplified and sequenced. This procedure results in many million short DNA sequences (reads).

In order to obtain the genome-wide binding locations of the transcription factor,the reads are mapped against a reference genome. The loci in the genome which are bound by the transcription factor are enriched in reads, the mapping procedure results in so-called *peaks* (Figure 1.7). The boundaries of these peaks can be identified using peak-calling software (Zhang *et al.*, 2008), resulting in a set of genome-wide locations of *cis*-regulatory elements.

ChIP-Seq can be used to identify cell type-specific transcription factor binding sites. The technology can also be employed to identify loci that are bound by co-factors such as the p300, resulting in data sets of tissue-specific enhancers (Visel *et al.*, 2009).


Cross-link protein to DNA
Shear DNA strand by sonicating
cell lysate
Add bead-attached antibodies to immunoprecipitate target protein
Unlink protein; purify, amplify DNA
Sequencing
G A T A A A T C T G G T C T T A T T T C C
Read mapping
Peak calling

```
!17488/2/61),687*86)3164888%%%%%%%!%  !/5/16862(26)686680'-6488868861.4)76
@SRR066788.2 WICMT-SOLEXA:5:1:0:393   @SRR066788.154 WICMT-SOLEXA:5:1:1:75
NGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCTNG  NATATATGTGCCCATGTGTGCAGGTTCTGCAGAGCC
+                                     +
!.531535312/44241444%%%%%%%%%%%%%%!%  !.88884-3/66865/5/5361,/5783/5464%%%
@SRR066788.3 WICMT-SOLEXA:5:1:0:1822  @SRR066788.155 WICMT-SOLEXA:5:1:1:27
NCTACGTCCCAACCGCTACCCCCACACCCAAGCCNG  NACGACAGAGGAGAATGGCGCCGCGCCGACGGCCCC
+                                     +
!%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%!%  !%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@SRR066788.4 WICMT-SOLEXA:5:1:0:1832  @SRR066788.156 WICMT-SOLEXA:5:1:1:17
NGCAGGCCTGGAGCGAGGTTAGGAAGATCCCATCNG  NCCCCCCACGTGGGGCCGCCTCTCCGCAAAGCCTAG
+                                     +
!%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%!%  !%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@SRR066788.5 WICMT-SOLEXA:5:1:0:613   @SRR066788.157 WICMT-SOLEXA:5:1:1:1205
NGTGTTTGTTTCATATGTGTGTGCTTGTGTGGTGTT  NATCCCACGCTGACAATGGTTCTGCTTCCTGATTCC
```

Figure 1.7: ChIP-Sequencing. Genome-wide DNA-protein interactions can be identified using chromatin immunoprecipitation followed by high throughput sequencing. See Section 1.4.2 for a description. Illustration based on Wikimedia Commons (2012d).

2 Combinatorial Binding at Enhancers in Embryonic Stem Cells

Transcription factors frequently interact in order to regulate gene expression. Many transcription factor binding sites have been identified, but the influence of combinatorial binding on co-factor interaction and evolution of gene regulation has not been investigated. In the following chapter, I integrate genome-wide binding data from mouse and human embryonic stem (ES) cells to study the role of combinatorial binding at long-distance gene regulatory elements.

2.1 Introduction

ES cells are derived from the inner cell mass of the blastocyst (Thomson *et al.*, 1998; Evans and Kaufman, 1981). During the course of normal development, implantation of the blastocyst results in further differentiation into distinct cell types of the three primary germ layers that will later form the tissues and organs of the developing embryo. ES cells form the in-vitro model of the inner cells mass, as they can differentiate into all somatic cell types. This pluripotent capacity of ES cells is maintained through a network of transcription factors, co-activators and chromatin modulators (Babaie *et al.*, 2007; Chen *et al.*, 2008; Jung *et al.*, 2010). The importance of transcriptional regulation in ES cells was demonstrated in a ground-breaking experiment, that showed that the transcription factors, OCT4, SOX2, NANOG, and KLF4 can induce an artificial pluripotent state in somatic cells (Takahashi *et al.*, 2007; Yu *et al.*, 2007; Nakagawa *et al.*, 2008). Expression of these four factors was sufficient to obtain ES cell-like induced pluripotent stem (iPS) cells, showing that the pivotal step in inducing and maintaining the pluripotency occurs at the level of genomic DNA by the binding of transcription factors to regulate gene expression. Identification and analysis of these binding sites is therefore highly important to understand pluripotency.

Many large-scale data sets have been produced for ES cells using the ChIP-Seq technology (Table 2.1). ChIP-Seq data pinpoints many thousands of transcription factor binding site candidates genome-wide. However, the high sensitivity comes along with a low specificity. For example, binding events detected with the ChIP-Seq technology can be indirect, non-functional for the cell type which is analysed, or technical artifacts, making identification of functional sites challenging. Nevertheless, in order to understand pluripotency at the level of transcriptional regulation, it is crucial to identify a reliable set of regulatory elements that actively contribute to the regulation of gene expression.

Mouse is a popular model system for human diseases and development. Coding sequences show a remarkably high level of conservation between mouse and human, orthologous genes have 82% identical amino acid sequences in average (Church *et al.*, 2009). Yet, the largest fraction of the genome is non-coding and shows much stronger divergence. Genome-wide binding events of OCT4 and NANOG show less than 5% conservation in mouse and human ES cells (Kunarso *et al.*, 2010), despite their conserved function for embryonic development. A study of genome-wide binding in liver tissue reported the same with only about 7% conserved binding events for the liver transcription factors CEBP and HNF4 between mouse and human (Schmidt *et al.*, 2010). These data show how fast *cis*-regulatory elements can evolve compared to coding sequence, yet it is unknown what discriminates conserved from non-conserved binding events. Sequence conservation has been used to search for enhancers, but sequence conservation alone is insufficient to estimate conservation of binding events (Blow *et al.*, 2010). Furthermore, genome-wide comparisons give average values over all observed binding events independently from their biological relevance. Since the ChIP-Seq technology identifies not only functional binding events, the level of binding conservation is currently unknown for a highly confident set of enhancers.

Sequence Conservation: Conserved similarity of the DNA sequence at orthologous loci

Binding Conservation: Binding of orthologous transcription factors at orthologous loci

This chapter addresses the role of combinatorial binding in embryonic stem cells and early mammalian development. Genome-wide binding data of the key transcription factors OCT4, SOX2 and NANOG is integrated with co-activator binding data, histone modification profiles and gene expression data. The integrated data is used to identify enhancers in ES cells and it is shown that these are frequently active during embryonic development. Additionally, combinatorial binding in mouse and human ES cells is compared to more precisely understand the evolution of gene regulation at long-distance regulatory elements.

2.1.1 Methods I: Overview of the Data Used in this Study

For this study, genome-wide binding data of the transcription factors Oct4, Sox2 and Nanog in mouse ES (mES) cells (Chen *et al.*, 2008; Marson *et al.*, 2008) was integrated with binding data of the transcriptional co-activators p300 (Chen *et al.*, 2008) and Mediator (subunits Med1 and Med12) (Kagey *et al.*, 2010) and with binding data for the Cohesin complex (subunits Smc1 and Smc3) and CTCF (Kagey *et al.*, 2010). These co-factors are important to activate gene expression by linking regulatory elements with the basal transcriptional machinery (see Section 1.3). Mouse developmental enhancers were obtained from Blow *et al.* (2010). Potential binding events were identified using MACS (Zhang *et al.*, 2008) ('peaks', see Section 1.4.2). All peaks with a p-value $> 1e-05$ and peaks that were detected in the control data were discarded. As a control for the influence of the p-value cutoff, the data was analysed using only the top 10% of peaks (sorted by p-value) from every experiment ('stringent data set'). I intentionally did not choose a false discovery rate (FDR) cutoff, since the FDR (as estimated by MACS) is heavily dependent on the control data (Zhang *et al.*, 2008) which is lacking for some experiments (see Appendix, Figures VI.1, VI.2, VI.3, VI.4 for a comparison of different cutoffs). To compare genome-wide binding in mouse and human ES cells, data from human cells was processed in the same way (see Table 2.1 for a complete listing of accession numbers, mapped reads and number of peaks). To investigate cell type differences, genome-wide binding data of OCT4 from human embryonal carcinoma cells (NCCIT) (Jung *et al.*, 2010) was produced in collaboration with the laboratory of James Adjaye. Important insights have been obtained from studies using ChIP-on-chip data (Boyer *et al.*, 2005), however due to its limitation to promoter regions, this data was not integrated into this analysis. The complete data is available at the European Nucleotide Archive (see Table 2.1) and can be accessed at `http://enhancer.molgen.mpg.de`, where I provide a human and mouse genome browser displaying genome-wide binding profiles, major histone modifications and RNA-seq data (Lister *et al.*, 2011). Figure 2.1 shows the aligned *SOX2* locus in the mouse and human genomes along with the data used for this study.

Study	Genome	Cell Type	Protein	GEO/ENA ID	Aligned Reads	Peaks
Göke *et al.* (2011)	hg19	NCCIT	OCT4	ERS071642	10064877 (26%)	4359
Göke *et al.* (2011)	hg19	NCCIT	Control	ERS071643	6926105 (27%)	-
Kunarso *et al.* (2010)	hg19	H1	OCT4	SRR037059-SRR037060	7353539 (57%)	19214
Kunarso *et al.* (2010)	hg19	H1	NANOG	SRR037061-SRR037063	7702058 (35%)	81891
Kunarso *et al.* (2010)	hg19	H1	CTCF	SRR037064-SRR037066	9731072 (46%)	77750
Kunarso *et al.* (2010)	hg19	H1	Control 1	SRR139068-SRR139071	9312693 (58%)	-
Kunarso *et al.* (2010)	hg19	H1	Control 2	SRR139072-SRR139074	8832219 (51%)	-
Lister *et al.* (2009)	hg19	H1	OCT4	SRR027915-SRR027916	661981 (6%)	3404
Lister *et al.* (2009)	hg19	H1	NANOG	SRR027965-SRR027966	7604042 (34%)	60209
Lister *et al.* (2009)	hg19	H1	SOX2	SRR027964	4242324 (47%)	33353
Lister *et al.* (2009)	hg19	H1	p300	SRR027920	3189661 (35%)	16206
Lister *et al.* (2009)	hg19	H1	Control	SRR018484	7471411 (43%)	-
Lister *et al.* (2011)	hg19	H1	RNA-Seq	SRR094768-SRR094775	-	-
NIH (2011)	hg19	IMR90	H3K27ac	SRR029631-SRR029632	19926175 (68%)	-
Chen *et al.* (2008)	mm9	E14	Oct4	SRR002012-SRR002015	7924650 (33%)	9029
Chen *et al.* (2008)	mm9	E14	Nanog	SRR002004-SRR002011	9979157 (35%)	19702
Chen *et al.* (2008)	mm9	E14	Sox2	SRR002023-SRR002026	8136347 (35%)	9012
Chen *et al.* (2008)	mm9	E14	p300	SRR023866-SRR023869	8633464 (26%)	481
Chen *et al.* (2008)	mm9	E14	CTCF	SRR001985-SRR001987	6211286 (26%)	52474
Chen *et al.* (2008)	mm9	E14	Control	SRR001996-SRR001999	6939857 (29%)	-
Marson *et al.* (2008)	mm9	V6.5	Oct4	SRR015151	4024970 (46%)	38774
Marson *et al.* (2008)	mm9	V6.5	Nanog	SRR015149-SRR015150	7466443 (42%)	22962
Marson *et al.* (2008)	mm9	V6.5	Sox2	SRR050356-SRR050357	7218075 (36%)	25306
Marson *et al.* (2008)	mm9	V6.5	Control	SRR015157-SRR015158	5878858 (51%)	-
Kagey *et al.* (2010)	mm9	V6.5	Med1	SRR058987-SRR058988	28391093 (61%)	27698
Kagey *et al.* (2010)	mm9	V6.5	Med12	SRR058985-SRR058986	22776494 (60%)	34318
Kagey *et al.* (2010)	mm9	V6.5	Nipbl	SRR058989-SRR058990	31250219 (57%)	21464
Kagey *et al.* (2010)	mm9	V6.5	Smc1	SRR058981-SRR058982	24176890 (62%)	48257
Kagey *et al.* (2010)	mm9	V6.5	Smc3	SRR058983-SRR058984	22917455 (64%)	35539
Kagey *et al.* (2010)	mm9	V6.5	Control	SRR058997	3639594 (50%)	-
Creyghton *et al.* (2010)	mm9	V6.5	H3K27ac	SRR066766-SRR066767	21872571 (67%)	-
Creyghton *et al.* (2010)	mm9	NPC	H3K27ac	SRR066773	8838081(71%)	-

Table 2.1: Mapping statistics. NPC: neuronal progenitor cells. Peaks: Peaks after cleaning

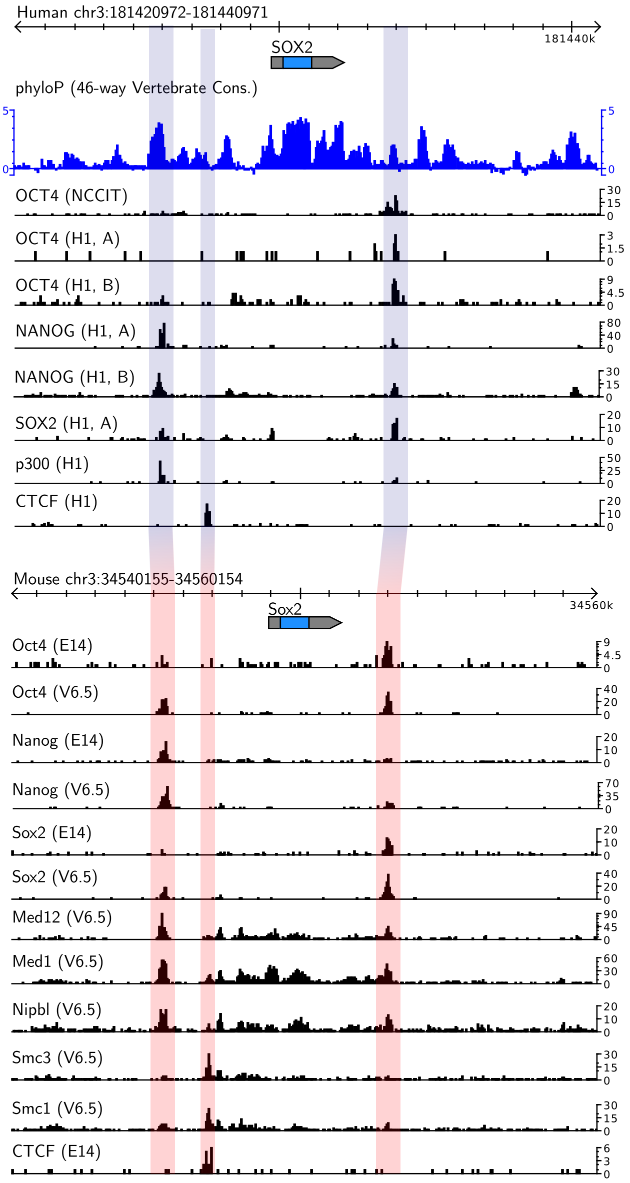


Figure 2.1: Overview of genome-wide binding data in human and mouse embryonic stem cells and embryonal carcinoma cells. Shown is the locus of the *SOX2* gene in the human genome (top), along with mapped reads for OCT4, SOX2, NANOG and p300. Individual experiments are shown as separate tracks. The orthologous locus in the mouse genome is aligned at the bottom along with mapped reads from the individual experiments. The dark blue track indicates sequence conservation (Pollard *et al.*, 2010). The highlighted areas correspond to regulatory elements bound by different combinations of transcription factors and co-activator complexes.

2.1.2 Methods II: Data Processing

Bowtie (0.12.5) was used to map the sequencing reads (Langmead *et al.*, 2009) with options -m 1 and -v 2 which guarantees that only those reads are kept that map uniquely and that contain at most two mismatches when being aligned to the reference. All coordinates refer to the reference genome versions hg19 and mm9.

Peak calling was done using MACS (1.4.0) (Zhang *et al.*, 2008) on the resulting BED files with control data as summarised in Table 2.1. The MACS default parameters were used, i.e. a p-value cutoff of 10^{-5}, except for the tag and effective genome size which had to be adjusted for every experiment, and -mfold 5,30. MACS was run on every negative control data set to obtain unspecific peaks. All peaks from the original experiment that overlapped with peaks from the control data were removed using BEDTools (Quinlan and Hall, 2010). The resulting numbers of final peaks after this 'cleaning' procedure are shown in Table 2.1.

Pre-computed whole genome alignments (Fujita *et al.*, 2011) were used to compare binding events from mouse and human ES cells. The peaks from the mouse-ChIP-Seq experiments were aligned to the human genome using the UCSC LiftOver tool (-minmatch 0.1) (Fujita *et al.*, 2011).

To analyse the binding combinations, the different sets of peaks were integrated into a binary matrix with rows for every genomic locus which is bound at least once, and columns for every factor (i.e. ChIP-Seq experiment). The entry [locus X, factor Y] in this matrix is set to 'true', if factor Y binds at locus X in the genome. All data sets were iteratively integrated by extending the length of the combined regulatory sites to span the overlapping peaks. The significance of the number of overlapping peaks for two genome-wide binding profiles was estimated using a hypergeometric test. For this test, it is assumed that only 25% of the genome can be bound by transcription factors. This way it is accounted for mapping limitations in repetitive sequences and genome-wide binding preferences. Furthermore, it is assumed that peaks overlap by 1 bp in average to obtain conservative estimates of p-values. The p-values of the hypergeometric test estimate the probability to observe the same number (or more) of shared binding events for position-randomised data sets. Clustering was done using the z-scores obtained from the hypergeometric tests. Enrichment of histone modifications was calculated on the highest 10% of peaks. All analysis was carried out with R (R Development Core Team, 2010), Bioconductor and peakAnalyzer.

2.2 Results

As a first step toward analysing combinatorial binding, I calculated the amount of pairwise co-localisation at the DNA in mouse ES cells (Figure 2.2). Co-localisation was estimated with a hypergeometric test (Section 2.1.2) and these estimates were used to cluster the experiments (Figure 2.2A). The clustering identified three distinct groups: enhancer binding (Oct4, Nanog, Sox2), insulator binding/ chromatin architecture (CTCF, Cohesin subunits Smc1 and Smc3a), and transcriptional co-activation (Mediator subunits Med1 and Med12). Interestingly, pairwise co-localisation as estimated from genome-wide data on DNA-protein interactions reproduces known protein-protein interactions (Manke *et al.*, 2003): CTCF interacts with Cohesin at insulator elements, Oct4, Sox2 and Nanog interact at enhancers, and Mediator plays a central role by integrating signals from distant regulatory elements with Cohesin (Figure 2.2B). To test whether this

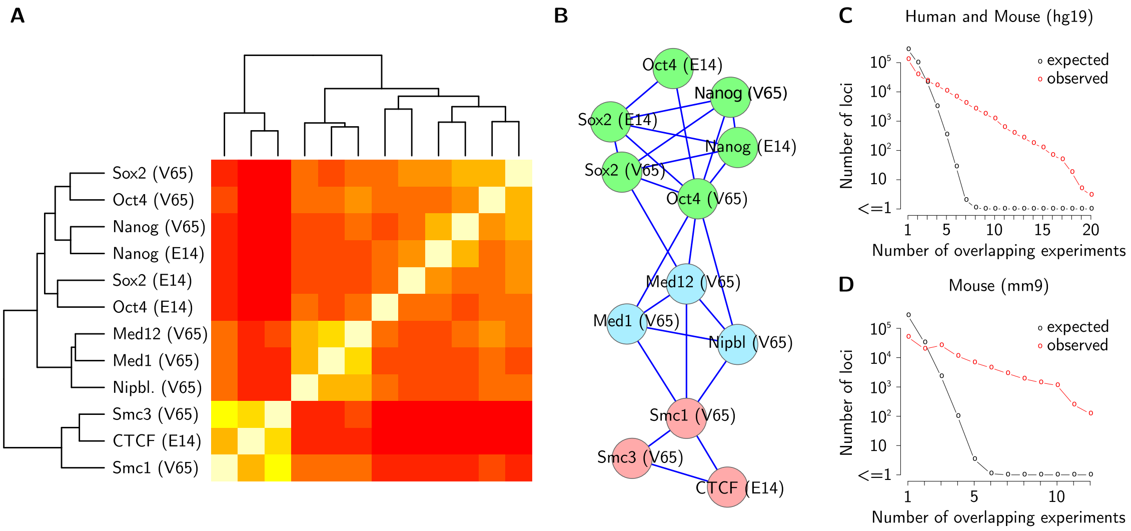


Figure 2.2: Co-localisation of proteins at the DNA reflect known protein-protein interactions. **(A)**Clustering of genome-wide binding profiles from mES cells based on the number of shared binding events identifies three main classes: Enhancer binding (Oct4, Sox2, Nanog), Insulator binding/ Chromatin looping (CTCF, Smc1, Smc3) and Mediator associated binding (Med1, Med12, Nipbl). **(B)** Protein-protein interaction network inferred from genome-wide binding data. Edges represent the pairwise similarities with the highest z-scores. **(C-D)** The number of co-localising proteins is much higher than expected by chance, both for mouse binding data (mm9, D) and human binding data together with the aligned mouse data (hg19, C). Randomised data sets show only very few cases where more than five experiments overlap (black line). The data used in this study show much stronger co-localisation with many loci where binding was detected in more than five experiments (red line).

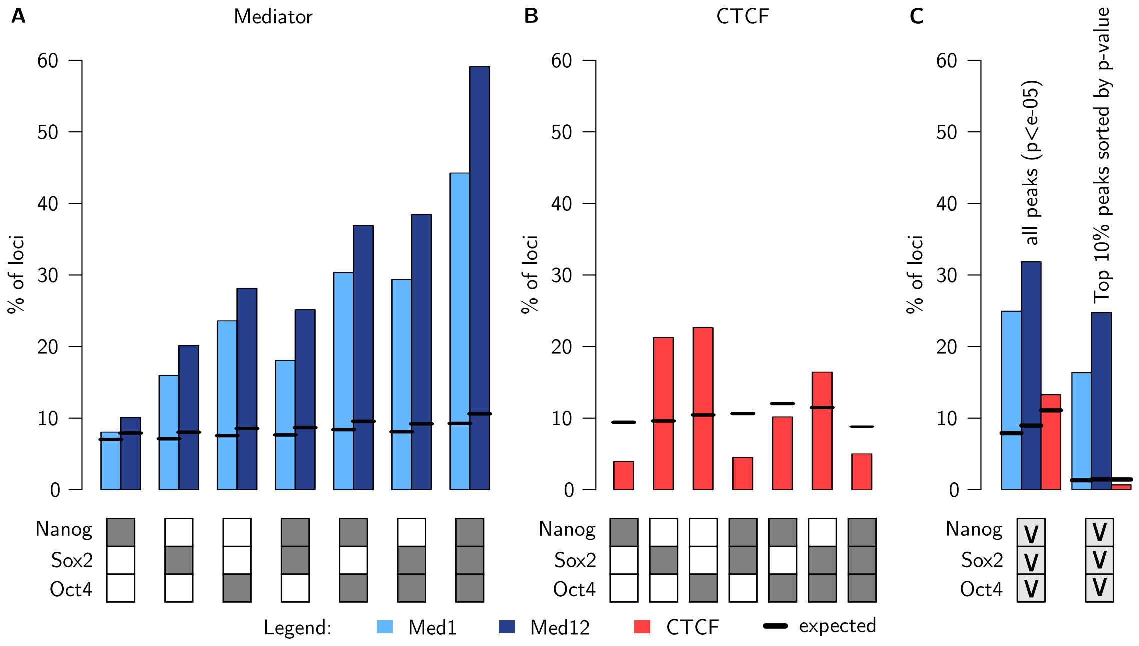


Figure 2.3: Mediator co-localises with Oct4, Sox2 and Nanog at combinatorially bound enhancers. **(A)** Bars indicate the fraction of loci where Med1 and Med12 binding can be observed, separated by the combination of Oct4, Sox2 and Nanog as indicated by the boxes below. Dark boxes indicate binding, white boxes indicate no binding. Med1 and Med12 preferentially co-localise when Oct4, Sox2 and Nanog bind simultaneously (combinatorially bound loci). **(B)** CTCF co-localisation with Oct4, Sox2 and Nanog. CTCF serves as a control to estimate unspecific binding. The binding combination has no influence on CTCF co-localisation, confirming that Mediator co-localisation is not caused by unspecific enrichment. **(C)**. Comparison of different peak-calling cutoffs. Light grey boxes ('v') indicate binding of at least one factor ('OR' relation). Combinatorial binding is more sensitive than a stringent control of false positives: the 10% most significant peaks are significantly associated with Med1 and Med12, however, the overall fraction is much lower compared to combinatorially bound loci.

amount of overlap of transcription factor binding events can be expected by chance, I calculated the overlap of position-randomised data sets (Figure 2.2C-D). Overall, the overlap observed in the data is much higher than expected by chance. These results support that the combination of binding events reflects functional interactions between the proteins themselves.

2.2.1 Combinatorial Binding and Transcription

The data indicates that interactions are not restricted to pairs of proteins, but rather extend to larger complexes. For example, Oct4, Nanog and Sox2 show co-localisation of all pairwise combinations, indicating that the combination of all three proteins might be required to regulate gene expression. Active enhancers are frequently bound by transcriptional co-activators such as Mediator, any influence of the binding combination should therefore be reflected in different levels of recruitment of the Mediator complex. To investigate the influence of higher order combinations, I calculated the fraction of loci that co-localise with the Mediator subunits Med1 or Med12 for all possible combinations of Oct4, Sox2 and Nanog (Figure 2.3A).

Between 5% and 30% of loci bound by Oct4, Nanog or Sox2 individually co-localise with Med1 or Med12 (Figure 2.3A). In contrast, loci bound by Oct4, Nanog and Sox2 simultaneously (further referred to as combinatorially bound loci) co-localise much more frequently with Med1 (44%) and Med12 (59%). Since these loci vary by number and size, the expected overlap from randomised data sets was calculated (hypergeometric test, see Section 2.1.2). These tests confirmed that the overlap of Med1 and Med12 with combinatorially bound loci is significantly higher than expected by chance (Med1: z-score 155.9, Med12: z-score 215.5).

Co-localisation of DNA binding proteins could be unspecific, for example due to binding at open chromatin regions (see Park (2009) for a review). In such a scenario, the increased co-localisation of Mediator at loci bound by multiple transcription factors could be an artifact. Unspecific co-localisation is not accounted for with the theoretical expected overlap. However, co-localisation levels of a factor which is known to be unrelated, such as CTCF, would be affected (Handoko *et al.*, 2011; Kagey *et al.*, 2010; Kunarso *et al.*, 2010). Figure 2.2 shows that CTCF largely binds to different regions than the enhancer binding proteins Oct4, Sox2 and Nanog, therefore CTCF co-localisation should be depleted at combinatorially bound loci. Indeed, this is confirmed by the data (Figure 2.3B). CTCF co-localisation is significantly depleted at loci bound by Oct4, Sox2 and Nanog simultaneously (z-score=-10.5). This suggests that combinatorial binding reduces unspecific co-localisation and confirms the association of the Mediator complex with combinatorially bound loci.

Next, I investigated whether the ChIP-Seq signal (‘binding intensity’) can be used to obtain enhancers which co-localise with Mediator independently of the binding combination. The fraction of loci where Mediator co-localises with at least one of the three transcription factors Oct4, Sox2 or Nanog, for the full data set and the stringent data set that only contains the top 10% peaks with the highest binding signal was calculated (Figure 2.3C). In the full data set, 25% of all loci show co-localisation with Med1. In the stringent data set with the high intensity binding peaks, 16% of all loci show co-localisation with Med1. This shows that choosing a cutoff on the binding intensity alone has a lower sensitivity in identifying enhancers that co-localise with Mediator compared to combinatorial binding (44% of loci show co-localisation). Integration of binding combinations has a similar effect in both the stringent and the full data set (Figure VI.2),

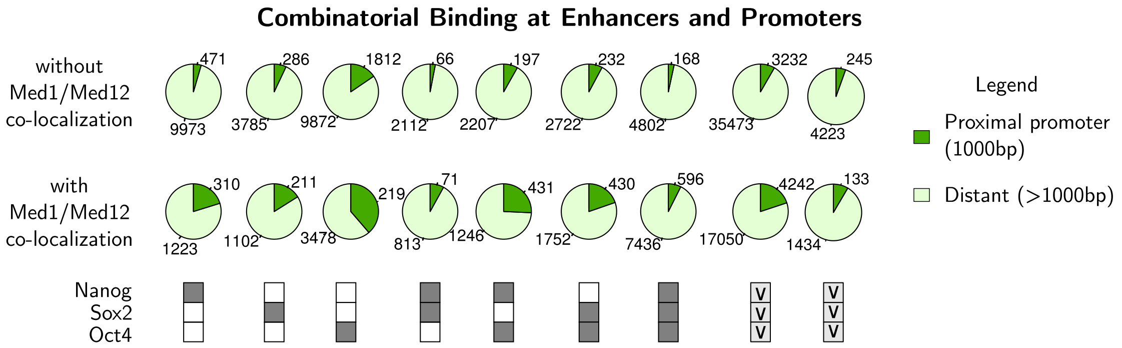


Figure 2.4: The majority of loci bound by Oct4, Sox2 and Nanog is more than 1000 bp distal from the nearest transcription start sites for all possible combinations (indicated by boxes below). Mediator co-localisation mainly occurs at distal regulatory sites, showing that the increased co-localisation of Med1 and Med12 at combinatorially bound loci is specific to enhancers.

confirming that the particular choice of p-value cutoff is of little importance in this analysis.

Since Mediator occupies many promoters in the genome, transcription factors that bind preferentially to promoter regions would be expected to show co-localisation with Med1 or Med12. To test whether the interaction between Oct4, Sox2 and Nanog occurs mainly at the promoter thereby causing the observed Mediator co-localisation, the fraction of promoters and enhancers was calculated for all binding combinations (Figure 2.4). The majority of loci bound by Oct4, Sox2 and Nanog are at distant regulatory elements (61%-97%), even when Mediator co-localisation can be observed. This shows that the increased overlap at combinatorially bound loci reflects specific binding at distant regulatory elements and is not caused by simultaneous occupation of the proximal promoter of actively transcribed genes.

The strong association of combinatorial binding with Mediator suggests that Mediator bound loci are functionally different from loci without Mediator binding. Histone modifications and gene expression indicate the activity of the regulatory elements (Section 1.3.1). Here, I analysed the enrichment of histone marks and gene expression of nearby genes to test for functional differences. Combinatorially bound loci occupied by Mediator are strongly enriched in H3K27ac, a mark for active enhancers (Rada-Iglesias *et al.*, 2011; Creyghton *et al.*, 2010), compared to loci without Mediator co-localisation (Figure 2.5A). To test the effect of Mediator co-localisation on gene expression, I performed a gene set enrichment analysis (GSEA, Subramanian *et al.* (2005)) using expression data from mES cells and differentiated cells after 14 days (Sene *et al.*, 2007). All genes were sorted according to their expression change between ES cells and differentiated cells and then the enrichment scores were calculated for genes near loci bound by

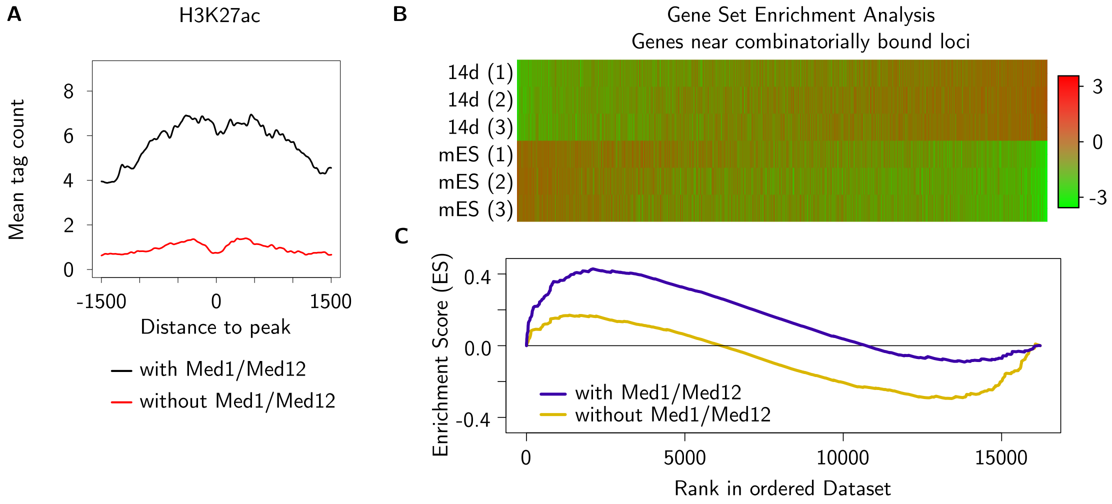


Figure 2.5: Functional analysis of combinatorially bound enhancers. **(A)** Average H3K27ac profile in mES cells around combinatorially bound loci. Loci bound by Oct4, Sox2 and Nanog together with Mediator are enriched in H3K27ac, a mark associated with active enhancers (black line). In contrast, loci without Mediator co-localisation show a much weaker enrichment (red line) suggesting that Mediator associates with active enhancers. **(B-C)** Gene Set Enrichment Analysis (GSEA) of genes near combinatorial binding events. **(B)** Expression of genes in mES cells (V6.5) and differentiated cells after 14 days (14d), sorted by the signal-to-noise ratio obtained from the GSEA software (Subramanian *et al.*, 2005). **(C)** The random walk that describes the gene set enrichment over genes sorted by their rank according to signal-to-noise ratio (similar sorting as in B). Group 1 (Oct4, Sox2, Nanog and Med1/Med12 in blue) is enriched in genes active in mES cells (enrichment score 0.43, p-value$< 10^{-3}$), group 2 (Oct4, Sox2, Nanog without Med1/Med12 in yellow) is enriched in genes active in differentiated cells (enrichment score -0.3, p-value=0.05).

Oct4, Sox2 and Nanog with Med1/Med12 (group 1) and without Med1/Med12 (group 2) (Figure 2.5B-C). Group 1 is significantly enriched in genes expressed in ES cells (enrichment score 0.43, p-value$< 10^{-3}$). Interestingly, group 2 shows a stronger enrichment in genes which are expressed in differentiated cells (enrichment score -0.3, p=0.05), suggesting that Oct4, Nanog and Sox2 might co-occupy poised enhancers. Both histone profiles and gene expression data support the notion that combinatorial binding identifies enhancers in embryonic stem cells while Mediator co-localisation determines their activity.

Active enhancers are frequently bound by transcriptional co-regulator complexes, show specific histone modifications and are associated with increased expression of nearby genes. The above results demonstrate that combinatorial binding has an influence on all three aspects, suggesting that combinatorially bound loci represent an important set of enhancer in ES cells.

2.2.2 Combinatorial Binding and Evolution

Using data from human ES (hES) cells, I investigated whether combinatorial binding can help in discriminating conserved and non-conserved binding events. To test this, whole genome binding data for OCT4, SOX2 and NANOG from human ES cells (Kunarso *et al.*, 2010; Lister *et al.*, 2009) and OCT4 from human embryonal carcinoma (EC) cells was analysed. EC cells are the malignant counterpart of ES cells (Przyborski *et al.*, 2004), however, they possess a distinct set of binding sites, extending the repertoire of potential OCT4 bound loci (Jung *et al.*, 2010). I used whole-genome alignments to assign binding events in mES cells to their orthologous loci in the human genome, retaining only those that could be aligned uniquely (Fujita *et al.*, 2011) (Section 2.1.2). A binding event is termed 'conserved' if binding of the same factor can be observed at the aligned loci in the human and mouse genome.

For every combination of OCT4, SOX2 and NANOG binding, the fraction of conserved binding was calculated (Figure 2.6A). Indeed, combinatorial binding is a good predictor for conservation: Less than 5% of individual binding events are conserved, which is less than expected. In contrast, about 15% of binding events at loci which are simultaneously co-occupied by OCT4, SOX2 and NANOG in hES cells show conserved binding of the respective transcription factor in mES cells (z-scores = 33.6, 41.3, 31.1). To test if combinatorial binding itself is conserved, the binding combination at conserved binding events in mouse was calculated for all combinations of OCT4, SOX2 and NANOG in human cells (Figure 2.6D, top). 53% of combinatorial binding events in human are simultaneously occupied by Oct4, Sox2 and Nanog in mouse, showing that combinatorial binding is likely to be a conserved property of regulatory elements in ES cells.

To investigate whether increased binding conservation at combinatorially bound loci is caused through unspecific effects, I calculate the fraction of loci bound by OCT4, SOX2 or NANOG in human ES cells that show CTCF binding in mES cells. Since these transcription factors do no co-localise with CTCF (Figure 2.3B), there should be no association between combinatorial binding in human and CTCF binding in mouse. Indeed, CTCF binding is significantly depleted at combinatorially bound loci (z-score = −6.8). CTCF binding can be observed at higher levels for all other binding combinations, most prominently OCT4 with SOX2. The combination of OCT4 and SOX2 without NANOG scarcely occurs genome-wide (159 times, the combination OCT4, SOX2 and NANOG occurs 6698 times). The high levels of CTCF at OCT4/SOX2 loci is therefore likely to be unspecific and of low relevance, which is confirmed by the stringent data set (Figure VI.3). This shows that the increased binding conservation at loci occupied by OCT4, SOX2 and NANOG is specific to the combination of transcription factors.

I further tested if a cutoff that selects the peaks with the highest binding intensities similarly identifies conserved binding events. For all loci which are bound by OCT4,

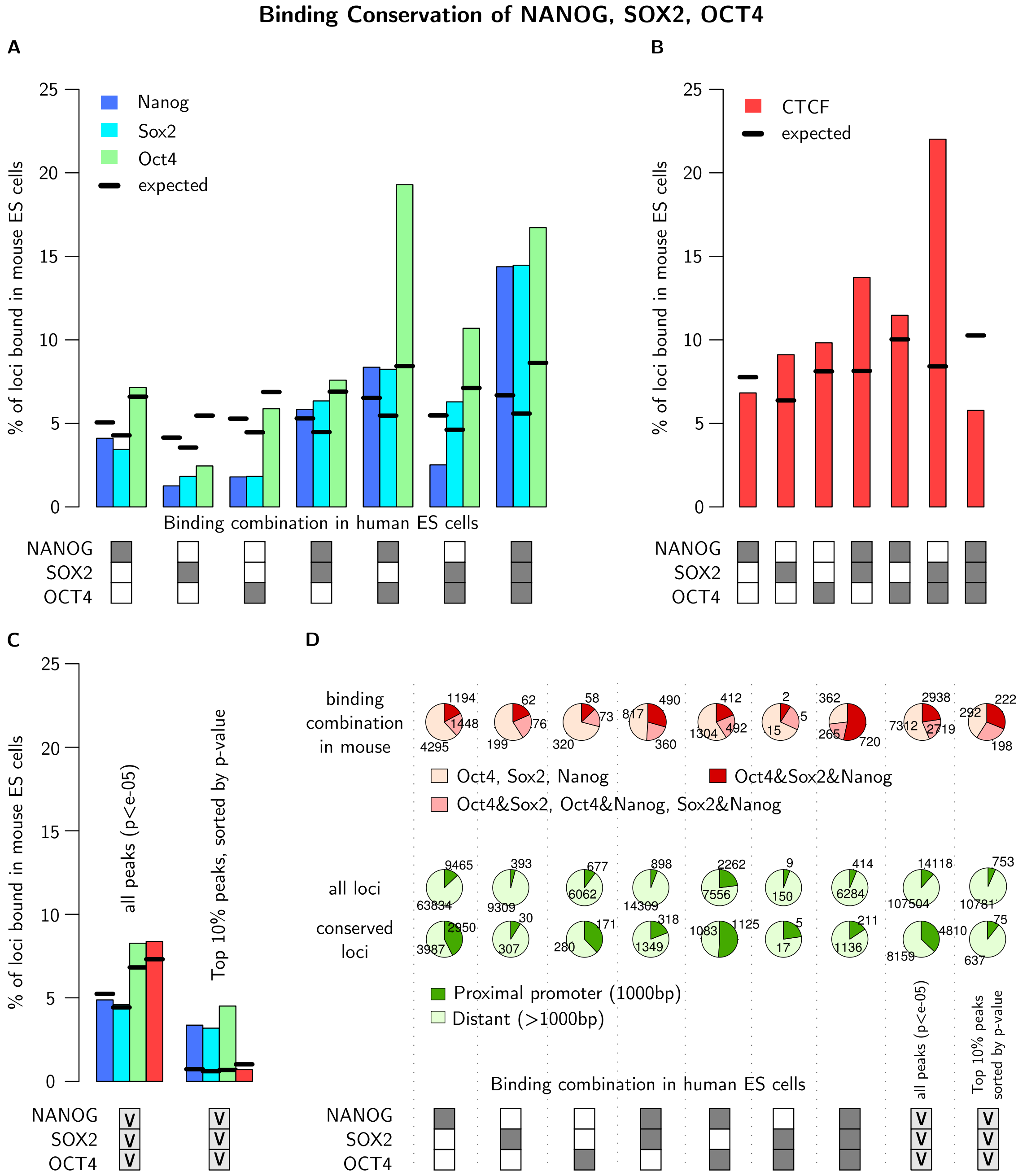


Figure 2.6: The combination of OCT4, SOX2 and NANOG influences conservation of binding events. **(A)** Bars indicate the fraction of loci where binding of Nanog, Sox2 and Oct4 can be observed at the orthologous loci in mouse ES cells for all combinations of OCT4, SOX2 and NANOG in human ES cells (dark boxes: binding; white boxes no binding). Loci simultaneously occupied by OCT4, SOX2 and NANOG in human show the largest fraction of conserved binding for Oct4, Sox2 and Nanog in mouse. **(B)** Estimation of unspecific binding (CTCF) in mouse at loci bound by OCT4, SOX2, or NANOG in human. CTCF is not enriched at combinatorially bound loci, confirming that this effect is specific. **(C)** Comparison of different peak-calling cutoffs. Light grey boxes with 'v' indicate binding of at least one factor. **(D)** Top: The fractions of binding combinations in mES cells at conserved loci (for all combinations of binding in human cells as indicated by the boxes below). Combinatorial binding of Oct4, Sox2 and Nanog in mES cells is much higher at combinatorially bound loci in human, suggesting that combinatorial binding is conserved in evolution. Bottom: The fraction of proximal and distant binding sites for conserved and non-conserved binding events, split up according to the binding combinations indicated below. The majority of conserved binding events are distant regulatory elements.

SOX2 or NANOG in human ES cells, the percentage of Oct4, Sox2, Nanog and CTCF binding in mouse ES cells was calculated for the full and the stringent data set that contains only the top 10% of peaks (Figure 2.6C). Less than 5% of binding events are conserved between mouse and human in the stringent data set. This is higher than expected, however, combinatorial binding is a more sensitive indicator for conservation (3-5% conserved binding for p-value cutoff vs. 14-17% for combinatorial binding). Many true binding sites will be lost in the stringent data set, increasing the number of false negatives. This is avoided using combinatorial binding. On the other hand, CTCF levels are strongly reduced in the stringent data set (Figure VI.3), showing that the number of false positives is higher when the full data set is used. It is likely that combinatorial binding decreases the number of false positives, as it is unlikely that these will be identified in multiple experiments.

Interestingly, the fraction of loci within the proximal promoter (±1000 bp) is higher for conserved binding events compared to non-conserved binding (Figure 2.6D), thus suggesting that the promoter is under stronger evolutionary constraint. However, the majority of binding events are distant from the predicted transcription start sites. The increased level of conservation at combinatorially bound loci is therefore not caused by a bias towards promoter binding, but specific to enhancers.

2.2.3 Conserved Combinatoriallly Bound Loci are Active in Development

The outcome of transcription factor binding events is ultimately determined by the function of the genes that they regulate. There are 720 conserved loci bound by OCT4, SOX2 and NANOG in human and mouse ES cells, associated with 608 genes nearby. Amongst the putative target genes are *OCT4*, *SOX2*, *LEFTY1*, *JARID2* and many other well known factors associated with pluripotency.

To obtain a more general picture of the downstream target genes of conserved combinatorial binding events I performed a Gene Ontology (GO) enrichment analysis using the Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean *et al.*, 2010). GREAT calculates enrichment of biological processes against a background set to correct for the bias introduced through large non-coding regions near developmental genes. All combinatorially bound regions in human were selected as background and the GO term enrichment was calculated for the subset of combinatorial binding events which are conserved between mouse and human (Figure 2.7A). Conserved combinatorially bound loci are significantly enriched in the terms pattern specification process (p-value = 4.7e-13), rationalisation (p-value = 2.5e-12) and developmental induction (p-value = 8.4e-8). Even though the background set is already enriched in developmental GO terms (amongst others developmental induction, p-value = 2e-8), the association of conserved

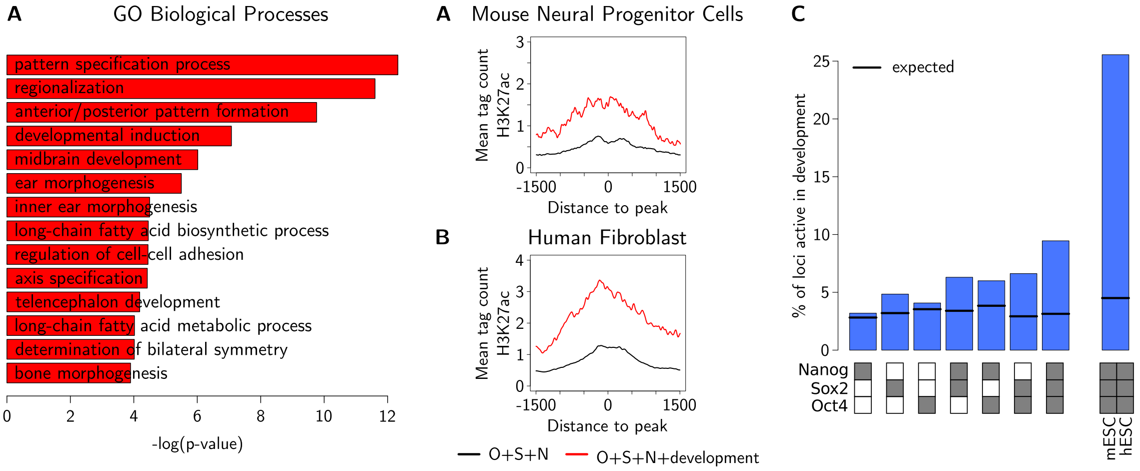


Figure 2.7: Conserved combinatorially bound loci are active in development. **(A)** GO enrichment analysis of conserved combinatorially bound loci, using all combinatorially bound loci as background. Genes near conserved loci are significantly enriched in processes important for development and differentiation. **(B)** Average mouse neural progenitor cell H3K27ac profile around loci bound by Oct4, Sox2 or Nanog in mES cells (O+S+N). Enhancers which are active in mouse development are enriched in H3K27ac in neural progenitor cell (red line) supporting that these elements play a role after differentiation of embryonic stem cells. **(C)** Average H3K27ac profile in human embryonic fibroblasts around loci bound by OCT4, SOX2 or NANOG in hES cells. Enhancers bound by OCT4, SOX2 or NANOG which are active in mouse development (red line) are enriched in H3K27ac in human fibroblast cells supporting that many of these enhancers are developmentally active in human as well. **(D)** Fraction of loci which show developmental activity in mouse; boxes below indicate the combination of Oct4, Sox2 and Nanog.

combinatorial binding events with developmental processes is even stronger. In support of this, genes such as *SOX21*, *FGF4*, *NEUROG3* and *CDX2* which are located near conserved combinatorial binding events have been shown to be important for directing differentiation of ES cells (Mallanna *et al.*, 2010; Spence *et al.*, 2011; Chawengsaksophak *et al.*, 2004).

The GO enrichment analysis showed that binding of Oct4, Sox2 and Nanog frequently occurs near developmental genes and gene expression data suggests that genes near combinatorial binding events are indeed up-regulated after differentiation (Figure 2.5B). The majority of loci bound by Oct4, Sox2 and Nanog together with Med1 or Med12 are likely to act as enhancers in embryonic stem cells. However, the function of loci near inactive genes is unclear. Since many of these genes are active during development, I tested whether Oct4, Sox2 and Nanog co-occupied loci act as early developmental enhancers. I used a set of tissue-specific enhancers obtained from mouse embryos at day e11.5 (Visel *et al.*, 2009; Blow *et al.*, 2010), a stage where neither Oct4 nor Nanog is expressed. Surprisingly, combinatorially bound loci in ES cells overlap significantly with these developmental enhancers (9%, z-score = 27.5).

The histone modification H3K27ac is associated with active enhancers. I calculated the enrichment of H3K27ac in mouse neural progenitor (NP) cells derived from mouse ES cells (Creyghton *et al.*, 2010) at all loci bound by Oct4, Sox2 and Nanog (Figure 2.7B). The subset of loci which are additionally active during development shows an enrichment of H3K27ac in NP cells, supporting that these indeed become active after differentiation. Next I tested whether these mouse developmental enhancers can be used to estimate human developmental enhancers. I calculated enrichment of H3K27ac in human embryonic lung fibroblast cells (IMR90, NIH (2011)) for all loci bound by OCT4, SOX2 and NANOG in human ES cells (Figure 2.7C). The subset of combinatorially bound loci which align with mouse developmental enhancers shows an enrichment in H3K27ac in IMR90 cells, suggesting that these enhancers can be used to estimate activity during human embryonic development.

Identification of developmental enhancers in human is very difficult, the best known examples were found using sequence conservation (Visel *et al.*, 2007; Pennacchio *et al.*, 2006). The above evidence suggests that Oct4, Sox2 and Nanog frequently bind to developmental enhancers. Interestingly, less than 5% of individually bound loci are active during development, whereas 10% of combinatorially bound loci show developmental activity in the mouse embryo (Figure 2.7C). Most strikingly, 26% of enhancers which show conserved combinatorial binding in mouse and human ES cells are active during development. This suggests that conserved combinatorial binding of Oct4, Sox2 and Nanog can be used to identify enhancers active during human embryonic development.

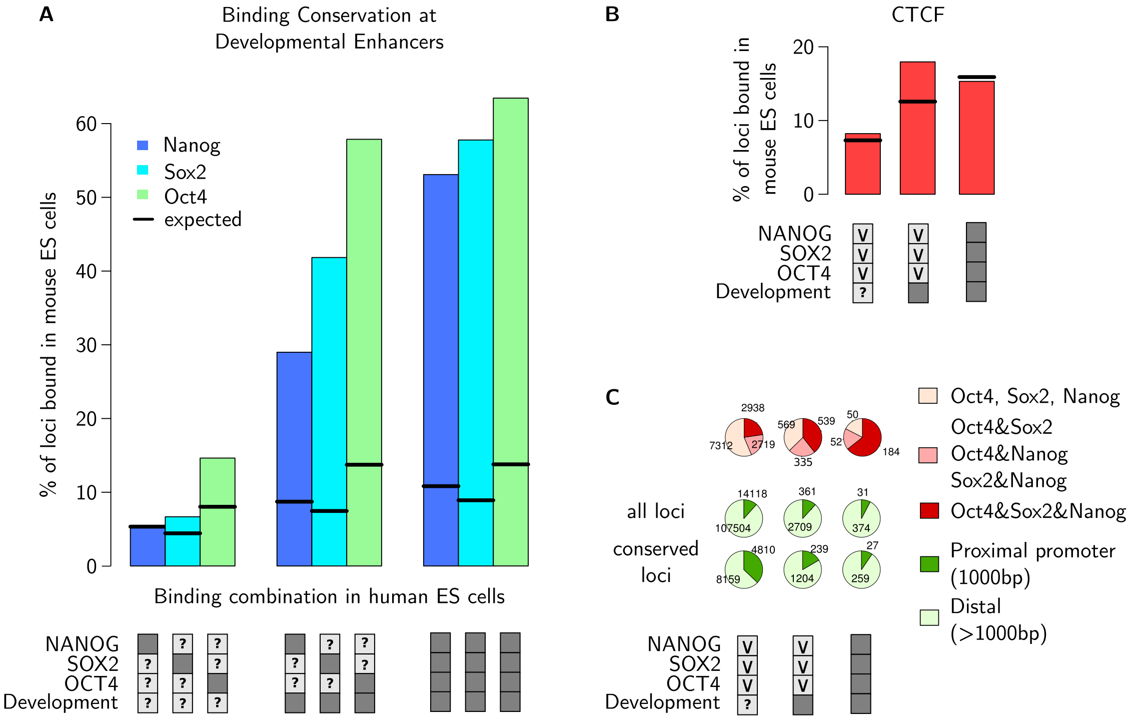


Figure 2.8: Conservation of gene regulatory hotspots. **(A)** Bars indicate the fraction of loci where binding of Nanog, Sox2, and Oct4 can be observed at the orthologous locus in mouse ES cells for all combinations of OCT4, SOX2 and NANOG in human ES cells discriminated by developmental activity as indicated by the boxes below. Dark boxes indicate binding, '?' indicates no restriction. Combinatorial binding events at developmentally active enhancers show the highest levels of binding conservation between mouse and human ES cells (>50%). **(B)** The level of CTCF binding is not affected by developmental activity, confirming that this effect is specific for conserved combinatorially bound loci. **(C)** Top: The fractions of binding combinations in mES cells at conserved loci (for all combinations indicated by the boxes below). The majority of conserved binding events at developmentally active enhancers where OCT4, SOX2 and NANOG bind simultaneously show combinatorial binding of Oct4, Sox2 and Nanog in mouse ES cells. Bottom: The fraction of proximal and distal binding sites for conserved and non-conserved binding events.

2.2.4 Gene Regulatory Hotspots: A Model for Highly Conserved Regulatory Elements

Conserved combinatorial binding hints at developmental activity. Vice versa, enhancers which are active during development show high conservation of binding events in ES cells (Figure 2.8A). Strikingly, 63%, 58% and 53% of OCT4, SOX2 and NANOG binding events are conserved in mouse at enhancers that are active in early development. This number is drastically higher than previous estimations (Kunarso *et al.*, 2010) and shows

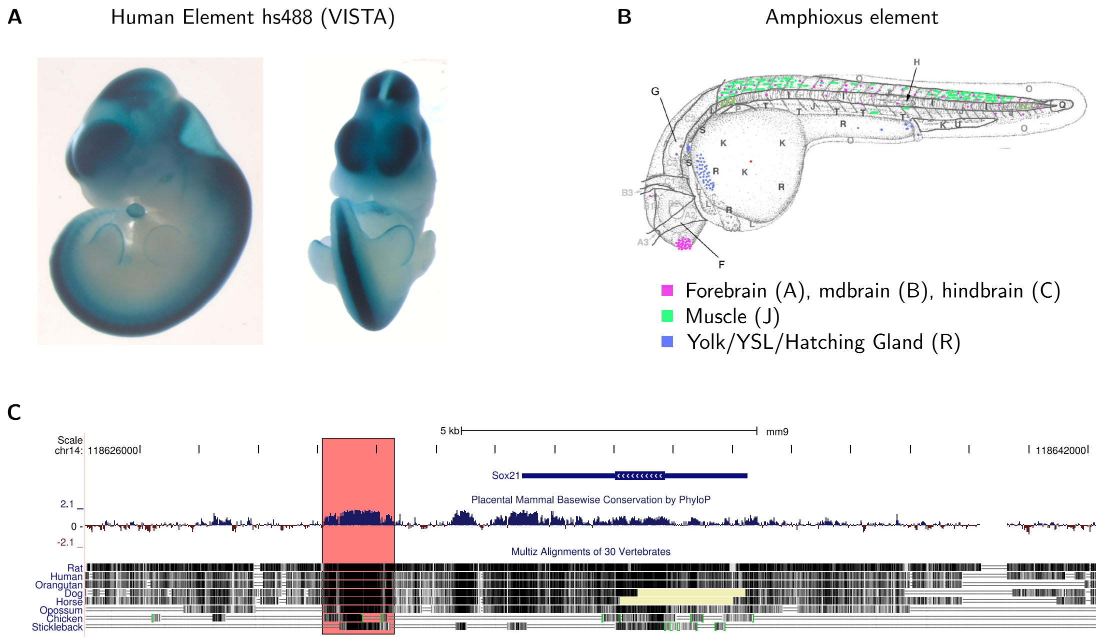

A
Human Element hs488 (VISTA)
B
Amphioxus element
Forebrain (A), mdbrain (B), hindbrain (C)
Muscle (J)
Yolk/YSL/Hatching Gland (R)
C
Scale
chr14: 118626000
5 kb
mm9
118642000
Sox21
Placental Mammal Basewise Conservation by PhyloP
2.1
0
-2.1
Multiz Alignments of 30 Vertebrates
Rat
Human
Orangutan
Dog
Horse
Opossum
Chicken
Stickleback

Figure 2.9: The gene regulatory hotspot downstream of *SOX21*. **(A)** The human sequence of the conserved combinatorially bound regulatory element downstream of *SOX21* shows reproducible activity during mouse development. Figure reproduced from the VISTA enhancer browser (Visel *et al.*, 2007). **(B)** The orthologous sequence of this regulatory element from amphioxus was tested in zebrafish where it showed reproducible activity in forebrain. Figure reproduced from Hufton *et al.* (2009). **(C)** Screenshot from the UCSC genome browser of the regulatory element downstream of *SOX21*. This enhancer is bound by Oct4, Sox2 and Nanog in mouse and human ES cells and acts as a developmental enhancer during mouse embryogenesis. Both, sequence and expression pattern is conserved between human, mouse and amphioxus.

that combinatorial binding together with developmental activity of the bound loci are strong indicators for binding conservation in embryonic stem cells.

The prominent difference in conservation between individual, isolated binding events and combinatorial binding events at enhancers which are active in multiple cell types suggests the existence of *gene regulatory hotspots* which are highly conserved in evolution (Figure 2.10). These hotspots are enhancers which recruit multiple, interacting transcription factors in pluripotent cells where they can be in an active or poised state. The very same element recruits different sets of transcription factors after differentiation and during development.

The element downstream of *SOX21* is such an example and illustrates the intimate connection between embryonic stem cells, pluripotency and development (Figure 2.9). SOX21 plays a pivotal role during brain development by promoting neuronal differentiation (Sandberg *et al.*, 2005). The downstream regulatory element is ultra-conserved with high sequence similarity in human, mouse and zebrafish, where it is always in close proximity to the *SOX21* gene (Figure 2.9C). The *cis*-regulatory element is bound by OCT4, SOX2, NANOG and p300 in human ES cells and Oct4, Sox2, Nanog and p300 in mouse ES cells. During mouse midbrain and forebrain development, this element is bound by the enhancer binding protein p300 and expression data shows that Sox21 is indeed over-expressed in forebrain compared to the whole embryo at day 11.5 (Visel *et al.*, 2009). The human element was tested *in vivo* in mouse and showed reproducible activity in forebrain, midbrain, hindbrain and neural tube (Figure 2.9A Visel *et al.* (2007)). The same element is conserved in amphioxus where it is associated with the *SOX21* ortholog *soxB2*. The amphioxus sequence was tested in zebrafish where it showed reproducible activity in forebrain (Figure 2.9B, (Hufton *et al.*, 2009)). The conserved enhancer downstream of *SOX21* is therefore a unique example of a functionally and genetically ultra-conserved *cis*-regulatory element that is bound in ES cells and active during development. This finding is indeed remarkable as it has been estimated that amphioxus split from vertebrates about 550 million years ago (Putnam *et al.*, 2008).

2.3 Discussion

Protein interactions are required for all major cellular processes. Transcription factors such as OCT4 and NANOG interact to regulate gene expression (van den Berg *et al.*, 2010). Since transcription factor recognise specific DNA motifs, it has been proposed that their interaction is encoded in the DNA and that the binding combination effectively determines the regulatory outcome. Computational approaches to study binding combinations have been limited to promoters as enhancers are very difficult to identify using sequence alone (Section 1.4.1). In this chapter, combinatorial binding at enhancers was investigated by analysing *in vitro* co-localisation of transcription factors. This analysis revealed that combinatorial binding impacts recruitment of transcriptional co-activators, histone modifications and has a significant impact on gene expression. Furthermore, combinatorial binding events are more frequently conserved between mouse and human, suggesting that combinatorial binding at enhancers increases the evolutionary constraint.

The enormous amount of genome-wide binding data produced in recent years has improved our understanding of the self-renewing and pluripotent state of embryonic stem cells (Boyer *et al.*, 2005; Chen *et al.*, 2008; Lee *et al.*, 2006). By integrating data from ES cells with developmental enhancers I demonstrated that the very same regulatory elements bound by key pluripotency factors in ES cells frequently act as enhancers during early development. This finding provides an unknown link between the gene regulatory networks of ES cells and early development at the level of transcriptional regulation.

The finding that binding at developmental enhancers is highly conserved in mouse and human ES cells suggests that these gene regulatory hotspots are crucial for the maintenance of the pluripotent state (Figure 2.10). It is likely that these elements are poised for activation (Rada-Iglesias *et al.*, 2011; Creyghton *et al.*, 2010), and an open chromatin state might be maintained throughout development to enable recruitment of transcription factors, co-activators, or histone modification proteins throughout cellular specification. Enhancers bound in multiple developmental stages by multiple factors influence gene expression in numerous cell types from pluripotent cells to at least cells of the mouse embryo at day 11.5. The existence of such gene regulatory hotspots could explain the extraordinarily high level of binding conservation observed in ES cells, since mutations of these elements would influence a major part of early embryogenesis. In contrast to these hotspots, loss of individual binding events can more easily be substituted by nearby binding events, and is likely to influence only a limited number of cell types. This analysis therefore suggests that the fast evolutionary rewiring of regulatory networks indeed mainly affects individual binding events, while combinatorial binding at gene regulatory hotspots is under stronger evolutionary constraint.

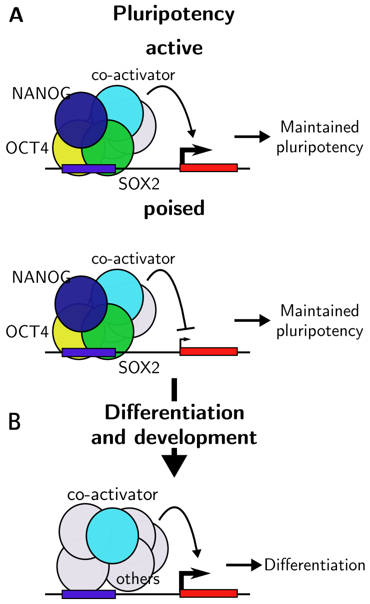


Figure 2.10: Schematic view of gene regulatory hotspots.

The definition of combinatorial binding in this study relies on the ChIP-Seq technology. Here, combinatorially bound loci are defined as genomic regions were binding of different transcription factors in similar cell types can be observed. These experiments are independent of each other and reflect measures from a mixed population of cells. Co-localisation could therefore be observed without direct physical interactions (for example through indirect interaction or competitive binding). However, results of this and other studies (van den Berg *et al.*, 2010; Lemischka, 2010) support that co-localisation as observed by the ChIP-Seq technology indeed reflects combinatorial binding.

One of the difficulties in analysing genome-wide data sets is how to discriminate true binding sites from false positive binding sites. It is impossible to identify a set of exclusively true binding sites, due to technical limitations, but also due to biological variation since many binding events will only be important under specified developmental cues.

A more stringent p-value cutoff decreases the fraction of false binding sites in the data while at the same time true positive binding events will be lost. Combinatorial binding is likely to select for true binding sites as well, since non-functional binding events are unlikely to be detected in multiple experiments. However, combinatorial binding is different from a stringent control of false positives as can be seen by Mediator co-localisation and binding conservation (Figures 2.3 and 2.6). It has been shown that groups of transcription factor binding sites are more likely to be conserved than isolated sites (Hemberg and Kreiman, 2011) which supports the value of combinatorial binding for transcriptional regulation. This is an important insight for future studies, which should consider the combination of transcription factors for defining regulatory networks.

One limitation of the ChIP-Seq technology is that combinatorial binding cannot be excluded. Weak or sporadic combinatorial binding events might be missed and therefore wrongly assigned as individual binding events (false negatives). However, the results obtained using two different cutoffs, a loose cutoff (full data set) with few false negatives and a very stringent cutoff (stringent data set) with many false negatives largely agree. Combinatorial binding events (OCT4, SOX2, NANOG) consistently show the strongest association with Mediator and highest levels of binding conservation. This suggests that the influence of the p-value cutoff and false negative binding events is limited on this analysis.

Most of the binding data in this study is obtained from embryonic stem cells. In mouse, data from two mouse embryonic stem cell lines (V6.5 and E14) was integrated. Interestingly, loci bound in both cell lines are much more likely to show co-localisation with Mediator (Figure VI.5A). In human, the available data was extended by OCT4 ChIP-Seq from embryonal carcinoma cells to obtain data from different cell lines. Loci bound in both EC and ES cells are much more likely to show combinatorial binding (Figure VI.5B). Therefore employing closely related cell lines is a biologically relevant approach for identifying important binding sites when data on combinatorial binding is not available.

A number of genes also show strong species-specific binding patterns, most prominently *Esrrb*, an interaction partner of Oct4 (van den Berg *et al.*, 2010) which is almost exclusively bound in mouse ES cells. It would be of interest to more deeply investigate genes that show strong species-specific binding patterns. Such an analysis could help to better understand the differences of mouse and human ES cells.

2.4 Conclusion

Developmental cues that lead to differentiation of cells during early embryogenesis involves binding of transcription factors at regulatory sequences in the genome. I have demonstrated that in ES cells, the combination of transcription factors that bind to regulatory elements is important for transcriptional activation. Combinatorial binding of OCT4, SOX2 and NANOG identifies enhancers characterised by H3K27ac and Mediator co-localisation. Many of these combinatorially bound enhancers are active during early development. The comparison of mouse and human ES cells showed that both combinatorial binding and multiple activity of enhancers in ES cells and development increase the evolutionary constraint. This analysis suggests that the fast evolutionary rewiring of regulatory networks mainly affects individual binding events. In contrast to these events, there is a group of conserved enhancers in the genome which recruit multiple interacting factors and are active in multiple tissues of the developing embryo (Figure 2.10). Many of these 'gene regulatory hotspots' are under strong evolutionary constraints and seem to play a major role by linking the regulatory networks of cellular differentiation during early mammalian development.

3 Alignment-Free Pairwise Comparison of Enhancer Sequences

Sequence similarity has been used to estimate the functional similarity of protein-coding genes. The similarity of coding sequences is usually estimated with alignments. In contrast to protein-coding genes, these alignment methods fail in the identification of functionally similar regulatory sequences like enhancers. Enhancers which drive expression in the same tissues frequently share the same transcription factor binding sites, therefore the number of shared DNA words can be used to compare such sequences. In this chapter, I present a novel, alignment-free sequence comparison method, $N2$, which can be used to calculate the pairwise sequence similarity of regulatory sequences.

3.1 Introduction

Sequence dependent gene regulation is mainly achieved through the binding of transcription factors at enhancers and promoters. The promoter is crucial to activate gene expression and it integrates different regulatory signals. Most frequently, the same promoter is active in all tissues where the gene is expressed. Enhancers occur much more frequently in the genome than promoters and multiple enhancers are associated with every gene (Heintzman *et al.*, 2009). This enables enhancers to influence gene expression in a much more cell type-specific manner (Chen *et al.*, 2008; Goto *et al.*, 1989; Small *et al.*, 1991; Zinzen *et al.*, 2009).

In Chapter 2 I showed that the cooperativity of transcription factors is important for transcriptional activation. Studies in *Drosophila* showed that the combination of binding sites together with the set of transcription factors actively recruited to an enhancer determines its cell type-specificity (Goto *et al.*, 1989; Small *et al.*, 1991; Zinzen *et al.*, 2009). More generally speaking, regulatory sequences with a similar binding site content can be expected to drive similar expression patterns. This is analogous to coding sequences, where sequence similarity has been used for many years to estimate functional similarity. The pairwise similarity of coding sequences is usually computed using global (Needleman and Wunsch, 1970) or local (Smith and Waterman, 1981) alignments. This approach works well for sequences which are at least partially alignable, however this is not the case for non-homologous enhancers. The location and orientation of binding sites in enhancers that show similar cell type-specific activity may differ widely (Davidson, 2006), making it impossible to produce alignments.

Alignment-free methods compare sequences according to their word content, see Vinga and Almeida (2003) and Bolshoy (2003) for an overview. The initial purpose was to design a fast and accurate measure of pairwise (dis-)similarity that could be used in databases where traditional alignments were too slow (Blaisdell, 1986; Hide *et al.*, 1994; Carpenter *et al.*, 2002). In the meantime, alignment-free methods have been applied in other contexts such as phylogeny (Wu *et al.*, 2009) and motif finding (Gordân *et al.*, 2010). The idea to describe a sequence by its word content directly fits the model of combinatorial binding in enhancers, where it is assumed that a similar function is reflected in a similar binding site content.

Word count-based methods have been used to compare regulatory sequences (Kantorovitz *et al.*, 2007; van Helden, 2004). However, these methods calculate the similarity of sequences based on exact word counts, whereas transcription factor binding sites are generally more flexible patterns (Section 1.4). Furthermore, the genomic orientation of enhancers is most often unknown, therefore it is important to compare sequences according to the word counts on both strands simultaneously. As an example, the word w = `CATAAT` might be bound by the same transcription factor as the words `CTTAAT` and `ATTATG`, the former having one substitution, the latter being on the reverse strand. Exact word comparison methods consider these words dissimilar, highlighting the need of a much more flexible approach for comparison of regulatory sequences.

More generally, let $n(w)$ be the set of words which are similar to w (the 'neighbourhood' of w). To overcome the limitation of exact word comparison methods, a similarity measure that compares sequences based on word neighbourhoods needs to be developed. Theoretical approaches that consider approximate word matches have been studied before (Forêt *et al.*, 2006; Burden *et al.*, 2008; Forêt *et al.*, 2009; Burden *et al.*, 2012), however no applicable method has been published for the purpose of pairwise comparison.

In this chapter I will give a background to word statistics and an overview of current alignment-free sequence comparison methods. I will then introduce $N2$, an alignment-free sequence comparison method that is based on the concept of word neighbourhoods to overcome the limitations of exact word comparison methods. I compare $N2$ to other alignment-free methods on simulated sequences and tissue-specific enhancer sequences identified *in vivo* in mouse embryos. I have implemented $N2$ and other alignment-free similarity measures as part of the open source C++ library SeqAn (Doering *et al.*, 2008). The fully documented code and an executable version (ALF) is available online (`http://www.seqan.de/projects/alf/`).

3.2 Background: Word Statistics

Nomenclature: Word Statistics

S	sequence
$S[i \dots i{+}j]$	sub-sequence from position i to $i{+}j$
l	length of sequence
A	alphabet, `A,C,G,T` in the case of DNA
w	word/ k-mer
k	length of word/k-mer
N_w^S, N_w	number of occurrences of w in sequence S
$Y_i(w)$	binary indicator for an occurrence of w starting at position i
D	set of all words w of length k ('dictionary')
$\|D\|$	size of D (4^k for DNA)
N^S	vector of word occurrences N_w^S for all $w \in D$

Regulatory sequences are defined by the set of transcription factor binding sites. Since these binding sites are most often unknown, regulatory sequences can be compared using the set of all possible words. However, this introduces noise, as the majority of words is unlikely to act as a binding site. Therefore it is important to identify words which are likely to be biologically relevant. Words with frequencies close to the expected frequency are of less biological interest, whereas words which occur more often than expected will be more relevant, presuming that this behaviour is due to a specific biological function. In order to decide if something occurs more often than expected, one first needs to know what to expect.

The occurrence of a transcription factor binding site in a DNA sequence can be described in general terms as the occurrence of a word in a text of letters from a specific alphabet. For a given alphabet A, let S be a sequence ('text') of length l, with every letter $S[i] \ \in A \ \forall \ i = 1 \dots l$ and let w be a word of length $k < l$. The number of occurrences of the word w in the sequence S (*word count*) is then described as

$$N_w^S = \sum_{i=1}^{l-k+1} \mathbf{1}(S[i \dots i+k-1] = w) \ .$$

N_w is a random variable, the distribution of which is dependent on the sequence length, sequence composition, word length, and word composition. For example, the words $w_1 =$ `AAAA`, $w_2 =$ `CGCG` and $w_3 =$ `CAGT` have entirely different word count distributions (Figure 3.1). Alignment-free comparison methods rely on an accurate description of the distribution of N_w^S in order to correctly weight word counts[1]. In the following, I will introduce models that describe the expected behaviour and variance of the word count N_w.

[1] The superscript indicator for sequence S will be omitted in the unambiguous case of a single sequence

3.2.1 Background Models for Biological Sequences

The aim of the model is to provide an accurate description of the phenomenon under study (Robin *et al.*, 2005). In order to study word statistics, it is assumed that the sequence $S = S[1]S[2]\dots S[l]$ was generated by a sequence of random variables $X_1X_2\dots X_l$. The possible values of X_i are the letters $x \in A$, $A = \mathtt{A,C,G,T}$ in the case of DNA, with $\sum_{x \in A} P(X_i = x) = 1$.

Bernoulli Model

The Bernoulli model assumes that the X_i are independently identically distributed (i.i.d.) random variables. For every position i in S, the distribution of letters $x \in A$ is described by $\mu(x)$, that is $P(X_i = x) = \mu(x)$ with $\sum_{x \in A} \mu(x) = 1$. The probability that a word w occurs at a specific position i in sequence S (further referred to as the *word probability* $\mu(w)$) is then the product of the probabilities of the letters of w:

$$\mu(w) = P(S[i \dots i+k-1] = w) = \prod_{j=1}^{k} \mu(w[j]) \ . \tag{3.1}$$

Formula 3.1 gives the probability that w occurs at any position i in sequence S, where the position is defined by the first letter of w. Let $Y_i(w)$ be the binary variable that indicates if an occurrence of w starts at position i in S:

$$Y_i(w) = \begin{cases} 1 & \text{if } S[i \dots i+k-1] = w \\ 0 & \text{otherwise .} \end{cases}$$

The indicator $Y_i(w)$ is a Bernoulli distributed random variable with parameter $p = P(Y_i(w) = 1) = \mu(w)$. The expected value $\mathbb{E}$ and variance $\mathbb{V}$ of $Y_i(w)$ are:

$$\mathbb{E}[Y_i(w)] = \mu(w) \tag{3.2}$$

$$\mathbb{V}[Y_i(w)] = \mu(w)(1 - \mu(w)) \ . \tag{3.3}$$

The properties of $Y_i(w)$ will be helpful to calculate the expected value and variance of the word counts N_w, as these can be modelled using $Y_i(w)$:

$$N_w = \sum_{i=1}^{l-k+1} Y_i(w) \ . \tag{3.4}$$

Dependence between word occurrences. Importantly, $Y_i(w)$ and $Y_j(w)$ are not independent for close positions i and j ($|i - j| \leq k - 1$). For example, the probability that the word $w = \texttt{AAAA}$ occurs at position $i+d$ is much higher when w already occurred at position i for $d = 1, 2, 3$, as it may overlap with itself, that is

$$P(Y_i(w) = 1 \ , Y_{i+d} = 1) \neq P(Y_j(w) = 1)P(Y_{j-1} = 1) \ .$$

This dependency can be captured by calculating the *word overlap indicator* ϵ, which indicates for every position $u = 1 \dots k$ of w if w can overlap with itself:

$$\epsilon_u(w) = \begin{cases} 1 \text{ if } w[k-u+1 \dots k] = w[1 \dots u] \\ 0 \text{ otherwise} \ . \end{cases} \tag{3.5}$$

The word overlap indicator can now be used to calculate the probability of observing dependent word occurrences for all possible word overlaps $|i-j| \leq k-1$:

$$P(Y_i(w) = 1 \ , Y_j(w) = 1) = \mu(w)\epsilon_{k-|i-j|}(w) \prod_{d=k-|i-j|+1}^{k} \mu(w[d]) \ . \tag{3.6}$$

Expected Counts. The word count N_w is a sum of random variables (Formula 3.4). The expected value of a sum of (dependent) random variables equals the sum of their expected values (Formula VI.1). Therefore, Formula 3.2 can be used to calculate the expected number of occurrences of the word w in the sequence S, $\mathbb{E}[N_w]$ (*expected counts*):

$$\begin{aligned} \mathbb{E}[N_w] =& \mathbb{E}[\sum_{i=1}^{l-k+1} Y_i(w)] = \sum_{i=1}^{l-k+1} \mathbb{E}[Y_i(w)] = \sum_{i=1}^{l-k+1} \mu(w) \\ =& (l-k+1)\mu(w) \ . \end{aligned}$$

Variance of the word count. Since N_w is a sum of dependent random variables, the variance of the word count $\mathbb{V}[N_w]$ can be computed according to

$$\mathbb{V}[N_w] = \mathbb{V}[\sum_{i=1}^{l-k+1} Y_i(w)] = \sum_{i=1}^{l-k+1} \sum_{j=1}^{l-k+1} \mathbb{C}\text{ov}[Y_i(w), Y_j(w)] \ . \tag{3.7}$$

The definition of the covariance (Formula VI.2) gives

$$\mathbb{C}\text{ov}[Y_i(w), Y_j(w)] = \mathbb{E}[Y_i(w) \times Y_j(w)] - \mathbb{E}[Y_i(w)]\mathbb{E}[Y_j(w)] \ . \tag{3.8}$$

The $Y_i(w)$ are Bernoulli distributed random variables where the expected value equals the probability of success:

$$\begin{aligned} \mathbb{C}\text{ov}[Y_i(w), Y_j(w)] =& P(Y_i(w) \times Y_j(w) = 1) - P(Y_i(w) = 1)P(Y_j(w) = 1) && (3.9) \\ =& P(Y_i(w) = 1 \ , Y_j(w) = 1) - P(Y_i(w) = 1)P(Y_j(w) = 1) \ . && (3.10) \end{aligned}$$

Since the dependency in the Bernoulli model only extends to positions where overlaps are possible, all $Y_i(w)$ and $Y_j(w)$ are independent for $|i-j| > k-1$, that is

$$P(Y_i(w) = 1\ ,\ Y_j(w) = 1) = P(Y_i(w) = 1)P(Y_j(w) = 1)\ ,\ \text{therefore} \tag{3.11}$$

$$\mathbb{C}\text{ov}[Y_i(w), Y_j(w)] = 0 \text{ for } |i-j| > k-1\ . \tag{3.12}$$

In the case of $|i-j| \leq k-1$ overlaps are possible and $P(Y_i(w) = 1\ ,\ Y_j(w) = 1)$ is defined according to Formula 3.6. Together with Formula 3.3, the covariance can be calculated:

$$\mathbb{C}\text{ov}[Y_i(w), Y_j(w)] = \begin{cases} \mu(w)(1-\mu(w)) \text{ for } \ i = j \\ \mu(w)\epsilon_{k-|i-j|}(w) \displaystyle\prod_{d=k-|i-j|+1}^{k} \mu(w[d]) - \mu(w)^2 \ \text{ for } |i-j| \leq k-1 \\ 0 \quad \text{otherwise}\ . \end{cases}$$

Applied to Formula 3.7 we obtain the exact word count variance for the Bernoulli model.

Application of the Bernoulli Model. Figure 3.1A shows the observed word count distribution for the words $w_1 = \texttt{AAAA}$, $w_2 = \texttt{CAGT}$, and $w_3 = \texttt{CGCG}$ for 15000 sequences of length 15000 bp sampled according to the Bernoulli model, with $\mu(\texttt{A}) = 0.3$, $\mu(\texttt{C}) = 0.2$, $\mu(\texttt{G}) = 0.2$, $\mu(\texttt{T}) = 0.3$. The expected value and variance exactly reproduce the observed mean and variance. The figure also shows the importance to calculate the exact variance which accounts for word overlaps. While ignoring dependencies of Y_i still accurately measures the variance for non-overlapping (CAGT) and rare words (CGCG), the variance for words with frequent self-overlaps (AAAA) is under-estimated.

Applying this to sequences sampled according to the dinucleotide frequency observed in the mouse genome shows that the Bernoulli model is still very limited (Figure 3.1B). Both the expected value and the variance do not resemble the observed mean or variance. For DNA sequences, the assumption that every nucleotide is independent from its neighbouring nucleotides is simply not correct, the CG dinucleotide (abbreviated as CpG) is very rare in mammalian genomes (Gardiner-Garden and Frommer, 1987). This directly leads to the formulation of Markov chains as a model for DNA sequences.

Markov Model

The main goal of Markov models is to include neighbouring dependencies of nucleotides in DNA sequences. As stated above, it is assumed that the sequence $S = S[1]S[2]\ldots S[l]$ was generated by a collection of random variables $X_1X_2\ldots X_l$, in other words, S is a realisation of the discrete stochastic process $(X_t)_{t\in\mathbb{N}^+}$. In contrast to the Bernoulli model, the assumption that all X_i are independent is dropped, instead it is assumed

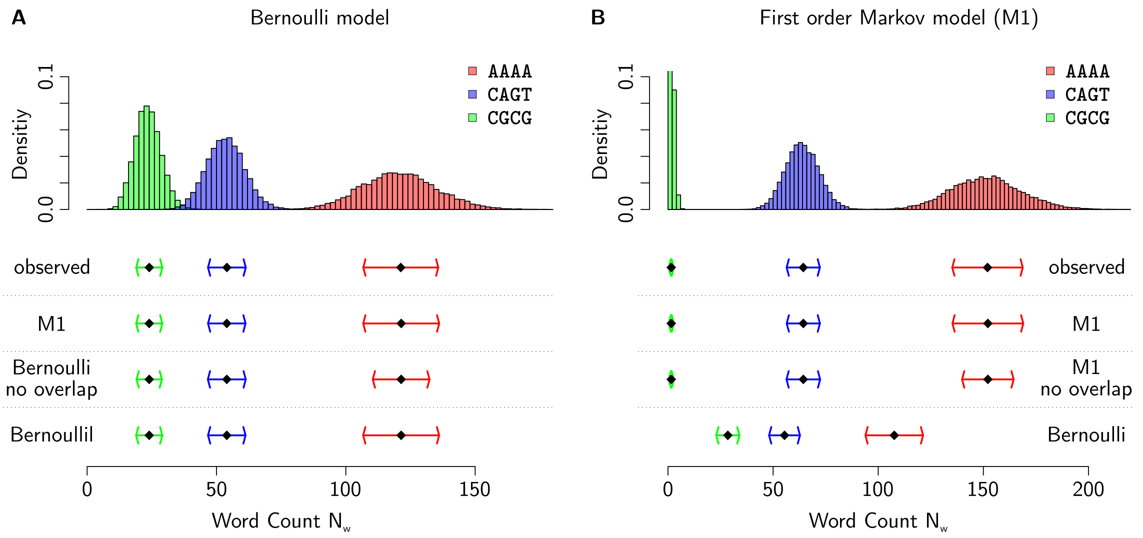


Figure 3.1: Word Counts for w_1 = AAAA (red), w_2 = CAGT (blue), and w_3 = CGCG (green) in 15000 sequences of length 15000 bp each. Black dots indicate the expected value and the arrows indicate the standard deviation for the simulated distributions for each word. **(A)** Sequences were generated according to i.i.d. nucleotides. The expected value and variance for the Bernoulli model and first order Markov model correctly resemble the observed values. Ignoring word dependencies leads to an underestimation of the word count variance. **(B)** Sequences were generated according to the dinucleotide distribution in the mouse genome. The expected value and variance for the first order Markov model correctly resembles the observed values, whereas the Bernoulli model fails. Similar to i.i.d. sequences, ignoring word dependencies leads to an underestimation of the word count variance.

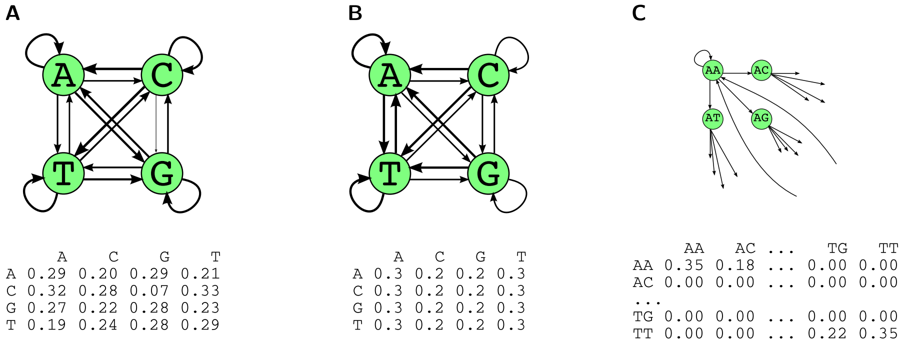
that the conditional probability distribution of future states of the stochastic process depends only on the present state, not on the sequence of events that preceded (Markov property):

$$P(X_{i+1} = x_{i+1} \mid X_i = x_i, \ldots, X_0 = x_0) = P(X_{i+1} = x_{i+1} \mid X_i = x_i)$$

The conditional probability $P(X_{i+1} = x_{i+1} \mid X_i = x_i)$ is called *transition probability*. The transition matrix Π is defined as the matrix containing all transition probabilities from state a to b ($\pi(a,b) = P(X_{i+1} = b \mid X_i = a)\ \forall\ a, b \in A$). Since X_{i+1} is necessarily drawn from the state space defined by the alphabet A, it holds that

$$\sum_{b \in A} \pi(a,b) = 1 \ .$$

For every row of the transition matrix, the sum of all elements is equal to 1. For DNA sequences, homogeneity of the stochastic process is generally assumed, that is



	A	C	G	T
A	0.29	0.20	0.29	0.21
C	0.32	0.28	0.07	0.33
G	0.27	0.22	0.28	0.23
T	0.19	0.24	0.28	0.29

	A	C	G	T
A	0.3	0.2	0.2	0.3
C	0.3	0.2	0.2	0.3
G	0.3	0.2	0.2	0.3
T	0.3	0.2	0.2	0.3

	AA	AC	...	TG	TT
AA	0.35	0.18	...	0.00	0.00
AC	0.00	0.00	...	0.00	0.00
...					
TG	0.00	0.00	...	0.00	0.00
TT	0.00	0.00	...	0.22	0.35

Figure 3.2: Markov Models for DNA Sequences. **(A)** First order Markov model for DNA sequences. The graph shows the states (nodes) and transition probabilities (edge width). The transition matrix is shown below. **(B)** Bernoulli Model ('zero order Markov model'). There are no dependencies of neighbouring nucleotides, therefore all rows of the transition matrix are equal. **(C)** Higher order Markov models can be reduced to first order Markov models by extending the state space. This example shows a second order Markov model.

the transition probability $\pi(a, b)$ does not depend on the position in the sequence. A Markov process, or Markov chain, can be visualised as a graph where the states are the nodes and the edges are associated with the transition probabilities (Figure 3.2A). The Bernoulli model is a specific Markov model with equal transition probabilities for all states, that is $\pi(a, b) = \mu(b) \; \forall \; b \in A$ ('zero' order Markov model, Figure 3.2B). Markov models of order m, where X_i depends on $X_i, \ldots, X_{i-m}$ can effectively be reduced to a first order Markov model by extending the state space to oligonucleotides of length m (Figure 3.2C). This observation will be used again later in this chapter.

Stationarity. Since X_1 has no predecessor, it has to be sampled from a particular distribution $\mu = \{\mu(\mathtt{A})\mu(\mathtt{C})\mu(\mathtt{G})\mu(\mathtt{T})\}$. In the following, it is assumed that the Markov chain is *stationary*, that is X_{i+1} has the same distribution as X_i. Accordingly, since X_1 is sampled from μ, all X_i are distributed according to μ. For every letter $b \in A$, it holds

$$\mu(b) = \sum_{a \in A} \mu(a)\pi(a, b) \; .$$

μ is called the *stationary distribution*. In sufficiently long DNA sequences all 16 dinucleotides occur, which ensures irreducibility[1] and aperiodicity[2] of the Markov chain. Given a finite state space, irreducibility, and aperiodicity of the Markov chain, the

[1] A Markov Chain is irreducible if it is possible to get to any state from any state.

[2] A Markov chain is aperiodic if for all states i there exists a t such that for all $t' > t$: $P(X_{t'} = i \mid X_t = i) > 0$.

stationary distribution $\mu = \mu\Pi$ always exists and is unique (Robin *et al.*, 2005). Stationarity of the Markov chain ensures that for every position i in the sequence S, the probability that a letter a occurs is $\mu(a)$. The probability $\mu(w)$ that a word w occurs at a specific position i therefore depends on the probability that the first letter occurs, $\mu(w[1])$ and can be calculated as follows:

$$\mu(w) = \mu(w[1]) \times \prod_{j=2}^{k} \pi(w[j-1], w[j]) \ . \tag{3.13}$$

Dependence between occurrences. Similarly to the Bernoulli model, the probability to observe occurrences of a word w at positions that would allow self-overlaps can be computed using the word overlap indicator ϵ (Formula 3.5):

$$P(Y_i(w) = 1 \ , Y_j(w) = 1) \tag{3.14}$$

$$= \mu(w)\epsilon_{k-|i-j|}(w) \prod_{d=k-|i-j|+1}^{k} \pi(w[d-1], w[d]) \text{ for } |i-j| <= k-1 \ . \tag{3.15}$$

Additionally, non-overlapping words are dependent in the Markov model, as the probability to observe the first letter of the word w ($w[1]$) depends on its last letter $w[k]$. To observe occurrences of w at positions i and j, w has to occur at position i ($\mu(w)$), the first letter of w has to occur exactly $|j-i|-k+1$ positions after the last letter of the occurrence of w at position j ($\pi^{|j-i|-k+1}(w[k], w[1])$), and $w[1]$ has to be followed by $w[2 \dots k]$:

$$P(Y_i(w) = 1 \ , Y_j(w) = 1) \tag{3.16}$$

$$= \mu(w)\pi^{|j-i|-k+1}(w[k], w[1]) \prod_{j=2}^{k} \pi(w[j-1], w[j]) \tag{3.17}$$

$$= \frac{\mu(w)^2}{\mu(w[1])}\pi^{|j-i|-k+1}(w[k], w[1]) \text{ for } |i-j| > k-1 \ . \tag{3.18}$$

Expected counts. Similarly to the Bernoulli model, the number of occurrences is a sum of (dependent) Bernoulli distributed random variables (Formula 3.4). Together with Formula 3.13 the expected value of $\mathbb{E}[N_w]$ can be computed as

$$\mathbb{E}[N_w] = \mathbb{E}[\sum_{i=1}^{l-k+1} Y_i(w)] = \sum_{i=1}^{l-k+1} \mathbb{E}[Y_i(w)] = \sum_{i=1}^{l-k+1} \mu(w) \tag{3.19}$$

$$= (l-k+1)\mu(w) \ . \tag{3.20}$$

Variance of the word count. The variance of the word count $\mathbb{V}[N_w]$ is the variance of the sum of dependent random variables:

$$\mathbb{V}[N_w] = \mathbb{V}[\sum_{i=1}^{l-k+1} Y_i(w)] = \sum_{i=1}^{l-k+1} \sum_{j=1}^{l-k+1} \mathbb{C}\mathrm{ov}[Y_i(w), Y_j(w)] \; . \tag{3.21}$$

The covariances can be calculated using the expected values (Formula 3.8). Since the Y_i are Bernoulli distributed random variables, the expected value equals the probability of success (Formula 3.10). The calculation of the covariance can be separated into three cases, identical positions $i = j$ (Formula 3.3), overlapping positions $|i - j| \leq k - 1$ (Formula 3.15), and non-overlapping positions $|i - j| > k - 1$ (Formula 3.18):

$$\mathbb{C}\mathrm{ov}[Y_i(w), Y_j(w)] = \begin{cases} \mu(w)(1 - \mu(w)) \text{ for } \; i = j \\ \mu(w)\epsilon_{k-|i-j|}(w) \displaystyle\prod_{d=k-|i-j|+1}^{k} \pi(w[d-1], w[d]) - \mu(w)^2 \; \text{ for } |i-j| \leq k-1 \\ \frac{\mu(w)^2}{\mu(w[1])} \pi^{|j-i|-k+1}(w[k], w[1]) - \mu(w)^2 \; \text{ for } |i-j| > k-1. \end{cases} \tag{3.22}$$

Together with Formula 3.21, the variance of the word count can be calculated.

Application of the Markov Model. For sequences that were generated according to the Bernoulli model (i.i.d.) the Markov model of first order correctly estimates the expected value and variance of the word counts for w_1 = `AAAA`, w_2 = `CAGT`, and w_3 = `CGCG` (Figure 3.1A). Similarly, when applied to sequences that have the same dinucleotide composition as the mouse genome, the Markov model accurately captures both mean and variance (Figure 3.1B). In this scenario, the Bernoulli model fails. Ignoring dependencies of overlapping words leads to an under-estimation of the word count variances (Figure 3.1B).

Estimation of the Markov Model Parameters

The true transition probabilities are unknown, therefore they have to be estimated from the observed sequence. Estimation of the Markov model parameters can be done using the maximum likelihood method. The likelihood $\mathcal{L}$ of the parameters θ for the statistical model given an observation S equals the probability to occur under this model:

$$\mathcal{L}(\theta \mid S) = P(S \mid \theta) \; .$$

In the maximum likelihood method, the true model parameters θ_0 are estimated by the value $\hat{\theta}$ which maximises $\mathcal{L}(\theta \mid S)$:

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta \mid S) \ .$$

In the case of DNA sequences, the maximum likelihood method is used to calculate the transition matrix $\hat{\Pi}$ with transition probabilities $\hat{\pi}(a,b)$ which most likely explain the sequence S:

$$\hat{\Pi} = \arg\max_{\Pi} \mathcal{L}(\Pi \mid S) \ .$$

The maximum likelihood estimator of transition probabilities of a first order Markov model for DNA sequences is given by (Durbin *et al.*, 1998)

$$\hat{\pi}(a,b) = \frac{N(ab)}{\sum_{x \in A} N(ax)} \ .$$

The same holds true for the Bernoulli model and Markov models of order $m > 1$, where the state space can be extended from nucleotides to oligo-nucleotides of length m.

3.2.2 Alignment-Free Sequence Comparison Methods

The development of alignment-free sequence comparison methods was initially driven by two main reasons. Technically, alignments can be slow to compute, alignment-free methods provided the means for significant speed-up (Blaisdell, 1986; Pevzner, 1992). Biologically, alignments assume sequences to be of linear order, which is not presupposed by alignment-free methods. The second reason makes alignment-free methods particularly suitable for comparing regulatory sequences, where linearity can not be assumed.

Traditionally, the idea of alignment-free methods is to compare two sequences S_1 and S_2, of length l_1 and l_2, based on the common word content using word frequency statistics to estimate similarity. Let D denote the *dictionary*, the set of all words w of length k (*k-mers*) and let $|D|$ be the size of the dictionary (4^k in the case of DNA sequences). Every sequence S of length l is then associated with the word count vector consisting of all word counts as defined in Formula 3.4:

$$\begin{aligned} N^S =& (N^S_{w_1}, N^S_{w_2}, \ldots, N^S_{w_{|D|}}) \ , \text{ with} \\ N^S_w =& \sum_{i=1}^{l-k+1} Y_i(w) \ . \end{aligned}$$

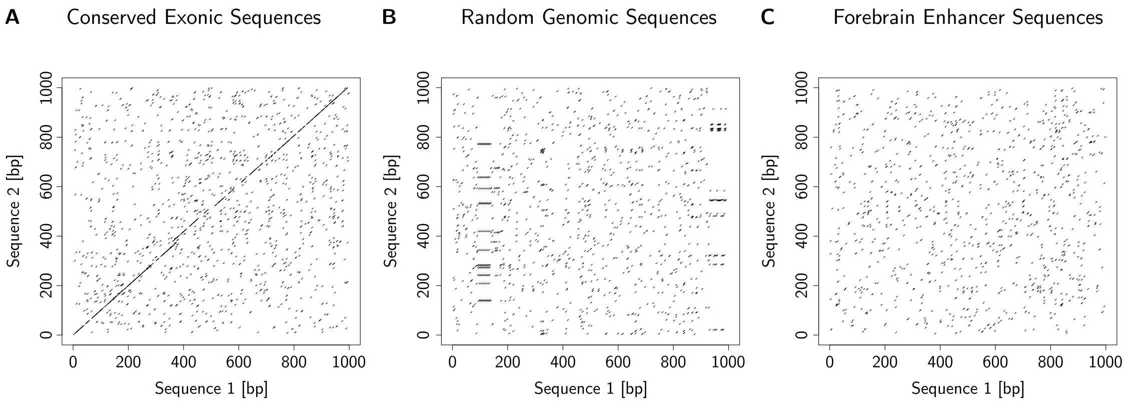


Figure 3.3: Dotplot, every line indicates a word of length 5 that is similar at the respective positions in both sequences. **(A)** Dotplot of conserved exonic sequence from (*JARID2*) in mouse and human. **(B)** Dotplot of two randomly selected sequences from the mouse genome. **(C)** Dotplot of two enhancers that drive expression during mouse forebrain development (Visel *et al.*, 2009).

Alignment-Free Similarities

D2. The most simple way to calculate a similarity of sequences using their word counts is to calculate their inner product, $D2$ (Lippert *et al.*, 2002):

$$\begin{aligned} D2(S_1, S_2) &=< N^{S_1}, N^{S_2} > \\ &= \sum_{w \in A} N_w^{S_1} \times N_w^{S_2} \, . \end{aligned}$$

The $D2$ similarity equals the sum of all pairs of word occurrences:

$$\begin{aligned} D2(S_1, S_2) &= \sum_{w \in D} \sum_{i=1}^{l_1-k+1} \sum_{j=1}^{l_2-k+1} Y_i^{S_1}(w) Y_j^{S_2}(w) \\ &= \sum_{i=1}^{l_1-k+1} \sum_{j=1}^{l_2-k+1} \mathbf{1}(S^1[i \dots i+k-1] = S^2[j \dots j+k-1]) \, . \end{aligned}$$

This can be visualised as a *dotplot* (Figure 3.3), where a dot at position (i, j) indicates a word of length k that is similar at position i in sequence S_1 and at position j in S_2. In this representation, $D2$ is exactly the number of dots in a dotplot. Similar sequences have a larger number of shared words (Figure 3.3A), whereas two random genomic sequences show only randomly matching words (Figure 3.3B). In a dotplot, regulatory sequences that drive similar expression patterns are almost indistinguishable from random sequence pairs (Figure 3.3C), as they only share a limited number of similar words.

The dotplot representation highlights the major limitations of the $D2$ score. Firstly, the score is directly dependent on the length of the sequences, larger sequences will results in more matching words and higher pairwise scores. Secondly, words which occur very frequently in only one sequence will still introduce a high pairwise score (Figure 3.3B), therefore $D2$ may measure single sequence noise rather than sequence similarity. Thirdly, in mammalian sequences the expected number of word occurrences varies strongly dependent on the word probability and overlap structure (Figure 3.2). High similarities calculated with $D2$ therefore may reflect an inappropriate background model (uniform i.i.d) instead of high similarity.

D2z. The $D2$ z-score ($D2z$) was proposed to obtain a standardised $D2$ score (Kantorovitz *et al.*, 2007):

$$D2z(S_1, S_2) = \frac{D2(S_1, S_2) - \mathbb{E}[D2(S_1, S_2)]}{\sqrt{\mathbb{V}[D2(S_1, S_2)]}} \ .$$

The $D2$ z-score corrects for length differences and incorporates a background model for word probabilities, thereby addressing some of the limitations of $D2$. One limitation of $D2z$ is that it is biased towards single sequence noise similar to $D2$ (Reinert *et al.*, 2009).

D2*. The $D2^*$ score (Reinert *et al.*, 2009) standardises the word counts instead of their inner product. $D2^*$ is defined as the inner product of the standardised word counts:

$$D2^*(S^1, S^2) = \sum_{w \in D} \frac{(N_w^{S_1} - \mathbb{E}[N_w^{S_1}])(N_w^{S_2} - \mathbb{E}[N_w^{S_2}])}{\sqrt{(l_1 - k + 1)\mu(w)}\sqrt{(l_2 - k + 1)\mu(w)}} \ . \tag{3.23}$$

Let $\mu(w)$ be the probability of w, the expected value of N_w^S is then estimated by $\mathbb{E}[N_w^S] = (l - k + 1)\mu(w)$. The authors assume a Poisson distribution, which implies that the variance is equal to the expected value. This assumption gives reasonable estimates for rare, non-overlapping words (Robin *et al.*, 2005). $D2^*$ was originally proposed with a Bernoulli background model for the computation of $\mu(w)$, where the background model is estimated on the concatenation of both sequences. For this study, I extended $D2^*$ to use Markov background models of higher order.

Other extensions of $D2$ for Bernoulli sequences. Several other methods have been proposed that approach some of the limitations of the $D2$ score. The distribution for the $D2$ score has been derived which allows the estimation of the significance for pairwise comparisons (Forêt *et al.*, 2006). The $D2$ score has been extended to approximate word matches (Burden *et al.*, 2008; Forêt *et al.*, 2006, 2009; Burden *et al.*, 2012), and a weighted word match statistic has been proposed (Jing *et al.*, 2011).

Alignment-Free Distances

$\mathbf{d^E}$. Among the first alignment-free distance measures proposed to compare sequences was the squared euclidean distance of the word count vectors, d^E (Blaisdell, 1986):

$$d^E(S_1, S_2) = \|N^{S_1} - N^{S_2}\|^2 = \sum_{w \in D} (N_w^{S_1} - N_w^{S_2})^2 \ . \tag{3.24}$$

Several limitations apply to the d^E distance measure as discussed for $D2$, for example it does not account for length differences, expected word frequencies or word overlaps.

$\mathbf{d^M}$. The squared Mahalanobis distance was proposed as a generalised statistical distance measure that corrects for all word correlations. The Mahalanobis distance is a multinomial generalisation of the distance from a distribution to its mean vector (Wu *et al.*, 1997):

$$d^M(S_1, S_2) = (N^{S_1} - N^{S_2}) \times \Sigma^{-1} \times (N^{S_1} - N^{S_2}) \ . \tag{3.25}$$

$$= \sum_{w \in D} \sum_{w' \in D} (N_w^{S_1} - N_w^{S_2}) \sigma_{ww'}^{inv} (N_{w'}^{S_1} - N_{w'}^{S_2}) \ . \tag{3.26}$$

The Mahalanobis distance requires the computation of the inverse of the covariance matrix (Σ^{-1}), which has a determinant near zero and therefore is almost singular (Vinga and Almeida, 2003). Up to word length $k = 4$, the pseudo inverse has been used as an estimate (Wu *et al.*, 1997).

$\mathbf{d^S}$. The squared standardised euclidean distance is the special case of the squared Mahalanobis distance where all covariances are assumed to be zero (Wu *et al.*, 1997):

$$d^S(S_1, S_2) = \sum_{w \in D} \frac{(N_w^{S_1} - N_w^{S_2})^2}{\sigma_{ww}} \ . \tag{3.27}$$

In the case of regulatory sequences, euclidean distance based measures are less sensitive compared to the inner product, as the quadratic term enforces that large word count differences have stronger impact on the score than any small numbers of similar words. Therefore, this analysis is restricted to alignment-free methods based on the inner product ($D2$, $D2z$, $D2^*$, $N2$).

Other Methods

The above methods summarise alignment-free distance measures based on k-mer content. Many other methods have been proposed, using angle metrics (Stuart *et al.*, 2002a,b), information theory (Wu *et al.*, 2001; Fernandes *et al.*, 2009), Chaos Theory (Almeida *et al.*, 2001; Joseph and Sasikumar, 2006; Almeida and Vinga, 2009), Kolmogorov complexity (Li *et al.*, 2001), Poisson approximations (van Helden, 2004), or compositional spectras (Kirzhner *et al.*, 2002; Bolshoy *et al.*, 2010), see Vinga and Almeida (2003) for an overview. A variety of alignment-free sequence comparison methods have been applied for phylogenetic analyses and whole genome comparisons (Karlin and Ladunga, 1994; Kirzhner *et al.*, 2003; Sims *et al.*, 2009a,b; Bolshoy *et al.*, 2010), see Bolshoy (2003) for an overview.

3.3 The N2 Similarity Score

Some of the limiting factors of alignment-free sequence comparison methods is their restriction to exact word counts or their narrow assumptions on word probabilities and variances. In the following $N2$ is presented, a novel alignment-free measure that allows for approximate word matches and incorporates the exact expected value and covariance of word counts with an additional improvement in running time compared to other inner product-based alignment-free methods.

3.3.1 Definition: N2

$N2$ overcomes the restriction to exact word counts by introducing the concept of *word neighbourhood counts.*

Weighted Word Neighbourhood Count. Let $n(w)$ be the set of words in the neighbourhood of the word w. The neighbourhood may be defined appropriately for every application, for example, to fit transcription factor binding motifs, to allow for reverse complement word counts or to include mismatches. Integrating neighbourhood counts for every word w reduces the influence of w itself. This leads to word count 'smoothing', i.e., inexact words are considered similar, but also to 'blurring', since inexact words might not be related. To control for these effects, every word w' in $n(w)$ is associated with a weight $a_{w'}$ which may differ for the considered application. The weighted word neighbourhood count $N^S_{n(w)}$ for every word $w \in D$ in sequence S can be defined as follows:

$$\begin{aligned} N^S_{n(w)} &= \sum_{w' \in n(w)} a_{w'} N^S_{w'} && (3.28) \\ &= \sum_{w' \in n(w)} a_{w'} \sum_{i=1}^{l-k+1} Y_i(w')\ . && (3.29) \end{aligned}$$

Standardised Weighted Word Neighbourhood Count. Depending on the choice of the neighbourhood $n(w)$, the word neighbourhood count $N_{n(w)}$ will be a sum of highly dependent random variables. Additionally, the variance of individual word counts should be considered, since, for example, a high number of `CAGCTG` occurrences is more informative than a high count of self overlapping words such as `AAAAAA` where a Poly-A stretch of length 15 already gives 10 occurrences. Also, some words are more likely to occur than others, `GC`-rich words for example are less frequent in mammalian genomes than `AT`-rich words. In order to use word neighbourhood counts for alignment-free sequence comparison, the counts have to be corrected for inter-variable dependency, word count variances and word probabilities. For $N2$, this is achieved by standardising the word neighbourhood counts:

$$\tilde{N}^S_w = \frac{N^S_{n(w)} - \mathbb{E}[N^S_{n(w)}]}{\sqrt{\mathbb{V}[N^S_{n(w)}]}} \; . \tag{3.30}$$

Since the word counts are dependent, the covariance of all words in the word neighbourhood has to be computed to obtain $\mathbb{V}[N^S_{n(w)}]$. The formulae for the expected value and variance of the weighted word neighbourhood counts are derived in Section 3.3.2.

Normalised Standardised Weighted Word Neighbourhood Count. The standardised neighbourhood count is normalised using the Euclidean norm ($\|\cdot\|$). This results in the normalised and standardised weighted word neighbourhood count vector $\hat{N}^S = (\hat{N}^S_{w_1}, \hat{N}^S_{w_2}, \ldots, \hat{N}^S_{w_{|D|}})$, with

$$\hat{N}^S_w = \frac{\tilde{N}^S_w}{\|\tilde{N}^S\|} \; . \tag{3.31}$$

N2. Using the above declarations, the $N2$ similarity of two sequences is defined as the inner product of their normalised standardised word neighbourhood count vectors:

$$\begin{aligned} N2(S_1, S_2) &=< \hat{N}^{S_1}, \hat{N}^{S_2} > && (3.32)\\ &= \sum_{w \in D} \hat{N}^{S_1}_w \times \hat{N}^{S_2}_w \; . && (3.33) \end{aligned}$$

As a consequence of the normalisation, $N2$ fulfils the properties $-1 \leq N2(S_1, S_2) \leq 1$ and $S_1 = S_2 \Rightarrow N2(S_1, S_2) = 1$, i.e., equal sequences will always have the maximum pairwise similarity of 1.

3.3.2 Word Statistics for Word Neighbourhood Counts

$N2$ can be computed with Markov models of any order. In the following the word statistics for $N2$ are derived assuming a first order Markov model with similar notation as introduced in Section 3.2[1].

Dependence Between Occurrences of Multiple Words. Similar to exact word counts, $N^S_{n(w)}$ is a sum of dependent Bernoulli distributed variables Y_i. However, in the case of exact word counts, only dependencies of self-overlapping words had to be considered ($Y_i(w) = 1$ and $Y_j(w) = 1$). In contrast, depending on the choice of $n(w)$, $N^S_{n(w)}$ incorporates dependencies of word pairs ($Y_i(w) = 1$ and $Y_j(w') = 1$). For example, occurrences of overlapping words such as $w = \texttt{C}\underline{\texttt{AAAA}}$ and $w' = \underline{\texttt{AAAA}}\texttt{A}$ are highly correlated, that is $P(Y_i(w) = 1\ ,\ Y_j(w') = 1)$ depends on the overlap structure of w and w'. The overlap indicator ϵ can be extended to word pairs (Robin *et al.*, 2005):

$$\epsilon_u(w, w') = \begin{cases} 1 \text{ if } w[k-u+1 \dots k] = w'[1 \dots u] \\ 0 \text{ otherwise .} \end{cases}$$

This can be used to calculate the probability to observe an occurrence of the word w at position i while considering any potential overlap with w' at position j:

$$P(Y_i(w) = 1\ ,\ Y_j(w') = 1) \tag{3.34}$$

$$= \mu(w)\epsilon_{k-|i-j|}(w, w') \prod_{d=k-|i-j|+1}^{k} \pi(w'[d-1], w'[d]) \text{ for } |i-j| < k\ . \tag{3.35}$$

Due to dependency of neighbouring nucleotides (Markov assumption), the probability to observe the first letter of the word w' ($w'[1]$) at position j depends on the last letter of the occurrence of the word w ($w[k]$) at position $i < j - k + 1$. The probability that the first letter of w' occurs exactly $|j - i| - k + 1$ positions after the last letter of w starting at position i equals $\pi^{|j-i|-k+1}(w[k], w'[1])$, therefore:

$$P(Y_i(w) = 1\ ,\ Y_j(w') = 1) \tag{3.36}$$

$$= \mu(w)\pi^{|j-i|-k+1}(w[k], w'[1]) \prod_{j=2}^{k} \pi(w'[j-1], w'[j]) \tag{3.37}$$

$$= \frac{\mu(w)\mu(w')}{\mu(w'[1])}\pi^{|j-i|-k+1}(w[k], w'[1]) \text{ for } |j-i| > k-1\ . \tag{3.38}$$

Expected Word Neighbourhood Count[2]. In Section 3.2, the expected value for

[1] Let the sequences be modelled by a first-order homogeneous stationary Markov chain with transition probabilities $\pi(i, j)$. The probability $\mu(w)$ that a word w occurs at a specific position i depends on the probability that the first letter occurs, denoted $\mu(w[1])$ (stationarity of the Markov chain) and can be calculated according to Formula 3.13.

[2] For clarity, the superscript indicator for sequence S is omitted in the following.

exact word counts was introduced. In the case of weighted word neighbourhood counts, the expected value $\mathbb{E}[N_{n(w)}]$ has to be extended to cover all words in the neighbourhood $n(w)$. $N_{n(w)}$ can be modelled as a sum of dependent Bernoulli distributed random variables (Formula 3.29). The expected value of the word neighbourhood counts, $\mathbb{E}[N_{n(w)}]$, can then be calculated according to:

$$
\begin{aligned}
\mathbb{E}[N_{n(w)}] &= \mathbb{E}\left[\sum_{w' \in n(w)} a_{w'} N_{w'}\right] \\
&= \sum_{w' \in n(w)} a_{w'} \mathbb{E}[N_{w'}] = \sum_{w' \in n(w)} a_{w'} \mathbb{E}[\sum_{i=1}^{l-k+1} Y_i(w')] \\
&= \sum_{w' \in n(w)} a_{w'} \sum_{i=1}^{l-k+1} \mathbb{E}[Y_i(w')] \\
&= \sum_{w' \in n(w)} a_{w'} \sum_{i=1}^{l-k+1} \mu(w') \\
&= \sum_{w' \in n(w)} a_{w'} (l-k+1) \mu(w') \; .
\end{aligned}
$$

Variance of the Weighted Word Neighbourhood Count. Since the word neighbourhood count is the number of occurrences of potentially overlapping words, the variance of the word neighbourhood count has to include all possible overlaps from words in $n(w)$:

$$
\begin{aligned}
\mathbb{V}[N_{n(w)}] &= \mathbb{V}\left[\sum_{w' \in n(w)} a_{w'} N_{w'}\right] && (3.39) \\
&= \sum_{w' \in n(w)} \sum_{w'' \in n(w)} a_{w'} a_{w''} \mathbb{C}\mathrm{ov}[N_{w'}, N_{w''}] \; . && (3.40)
\end{aligned}
$$

Therefore, the variance of the weighted word neighbourhood count, $\mathbb{V}[N_{n(w)}]$, equals the sum of the weighted covariances of the word counts for all pairs of words (w', w'') in the neighbourhood $n(w)$. The covariance of word counts, $\mathbb{C}\mathrm{ov}[N_w, N_{w'}]$, can be calculated using the property that they form a sum of Bernoulli distributed random variables

(Formula 3.4):

$$\mathbb{C}\mathrm{ov}[N_w, N_{w'}] = \mathbb{C}\mathrm{ov}[\sum_{i=1}^{l-k+1} Y_i(w), \sum_{j=1}^{l-k+1} Y_j(w')] \tag{3.41}$$

$$= \sum_{i=1}^{l-k+1} \sum_{j=1}^{l-k+1} \mathbb{C}\mathrm{ov}[Y_i(w), Y_j(w')] \; . \tag{3.42}$$

The covariance can be calculated using the expected values:

$$\mathbb{C}\mathrm{ov}[Y_i(w), Y_j(w')] = \mathbb{E}[Y_i(w) \times Y_j(w')] - \mathbb{E}[Y_i(w)]\mathbb{E}[Y_j(w')] \; . \tag{3.43}$$

The expected value of $Y_i(w)$ equals the probability of success:

$$\mathbb{C}\mathrm{ov}[Y_i(w), Y_j(w')] = P(Y_i(w) \times Y_j(w') = 1) - P(Y_i(w) = 1)P(Y_j(w') = 1) \tag{3.44}$$

$$= P(Y_i(w) = 1 \; , Y_j(w') = 1) - P(Y_i(w) = 1)P(Y_j(w') = 1) \; . \tag{3.45}$$

The calculation of the covariance for word pairs can be separated into the individual cases (similar to the covariance of word counts for equal words introduced in Formula 3.22): identical positions for identical words ($i = j, w = w'$) and non-identical words ($i = j, w \neq w'$, Formula 3.45), overlapping positions $|i-j| < k$ (Formula 3.35), and non-overlapping positions $|i - j| > k - 1$ (Formula 3.38):

$$\mathbb{C}\mathrm{ov}[Y_i(w), Y_j(w')] = \begin{cases} \mu(w)(1-\mu(w)) \text{ for } \; i = j, w = w' \\ -\mu(w)\mu(w') \; \text{ for } \; i = j, w \neq w' \\ \mu(w)\epsilon_{k-|i-j|}(w,w') \prod_{d=k-|i-j|+1}^{k} \pi(w'[d-1], w'[d]) - \mu(w)\mu(w') \text{ for } |i-j| < k \\ \frac{\mu(w)\mu(w')}{\mu(w'[1])} \pi^{|j-i|-k+1}(w[k], w'[1]) - \mu(w)\mu(w') \text{ for } |j-i| > k-1 \; . \end{cases}$$

Together with Formula 3.40 and 3.42, the variance of the weighted word neighbourhood counts can be calculated.

3.3.3 Implementation, Instances, and Availability of N2

The calculation of the scores is divided into two steps, a pre-processing step and a comparison step.

Pre-processing. The pre-processing step, as outlined in the following, is run for every sequence individually. The running time of this step depends on the length of the input sequences l, the Markov model's order m, the word length k and the average size of the word neighbourhoods. First, the background Markov model is estimated using a maximum likelihood approach (see Section 3.2) on every sequence ($O(4^m)$), then the words are counted ($O(l)$), and the word probabilities and covariances are calculated ($O(4^k \text{NeighbourhoodSize}^2)$). The neighbourhood word count variance is computed using Formula 3.40. The covariance of word counts (Formula 3.42) can be computed by (Robin *et al.*, 2005):

$$\mathbb{C}\text{ov}[N_w, N_{w'}] = \tag{3.46}$$

$$\mu(w)\sum_{d=1}^{k-1}(l-k-d+1)\left[\epsilon_{k-d}(w,w')\prod_{j=k-d+1}^{k}\pi(w'[j-1],w'[j])-\mu(w')\right] \tag{3.47}$$

$$+\mu(w')\sum_{d=1}^{k-1}(l-k-d+1)\left[\epsilon_{k-d}(w',w)\prod_{j=k-d+1}^{k}\pi(w[j-1],w[j])-\mu(w)\right] \tag{3.48}$$

$$+\mu(w)\mu(w')\sum_{t=1}^{l-2k+1}(l-2k-t+2)\left[\frac{\pi^t(w[k],w'[1])}{\mu(w'[1])}+\frac{\pi^t(w'[k],w[1])}{\mu(w[1])}-2\right] \tag{3.49}$$

$$-(l-k+1)\mu(w)\mu(w')\;. \tag{3.50}$$

In the case where $w = w'$, we have $\mathbb{C}\text{ov}[N_w, N_{w'}] = \mathbb{V}[N_w]$. The word count variance can be calculated as follows (Robin *et al.*, 2005):

$$\mathbb{V}[N_w] = \tag{3.51}$$

$$(l-k+1)\mu(w)[1-\mu(w)] \tag{3.52}$$

$$+2\mu(w)\sum_{d=1}^{k-1}(l-k-d+1)\left[\epsilon_{k-d}(w)\prod_{j=k-d+1}^{k}\pi(w[j-1],w[j])-\mu(w)\right] \tag{3.53}$$

$$+2[\mu(w)]^2\sum_{t=1}^{l-2k+1}(l-2k-t+2)\left[\frac{1}{\mu(w[1])}\pi^t(w[k],w[1])-1\right]\;. \tag{3.54}$$

Terms (3.54) and (3.49) are costly to compute and have minor effects on the variance and covariance. Let the sequence $S = S[1]\dots S[l]$ be a realisation of the irreducible, aperiodic Markov chain $X_1 \dots X_l$ on the finite alphabet A (Section 3.2), then the distribution of X_i converges to the stationary distribution μ:

$$\lim_{i\to\inf} P(X_i = a) = \mu(a)\;. \tag{3.55}$$

Furthermore, the convergence rate is exponential (Robin *et al.*, 2005), therefore the limiting distribution is reached quickly. Here, it is assumed that the limiting distribution is reached for $t >= k$. This way, dependencies from non-overlapping word occurrences are neglected by assuming that $\mu(w'[1]) \approx \pi^t(w[k], w'[1])$ for $t >= k$.

Since the neighbourhoods of different words w_1 and w_2 can contain similar word pairs ($\{w', w''\} \in n(w_1) \cap n(w_2)$), the same covariance will be used multiple times. At the same time, some word pairs never occur in the same neighbourhood ($\{w_1, w_2\} \notin n(w) \forall\ w \in D$), these covariance values won't be needed. Therefore, every covariance term is computed at its first occurrence to dynamically fill the covariance matrix. This procedure allows the pre-computation of all required covariance values without pre-computing unnecessary covariance terms.

As the last pre-processing step, the standardised normalised word neighbourhood counts (Formula 3.31) are calculated for every sequence. The total complexity of the pre-processing is linear in the number of input sequences n:

$$O(n(l + 4^m + 4^k \text{NeighbourhoodSize}^2)) \, .$$

Comparison. In the comparison step, the inner product of the standardised normalised word neighbourhood counts is computed for all pairs of sequences. The running time of this step depends on the word length k and is quadratic in the number of input sequences n. Due to the pre-processing, the comparison step has the same complexity as calculating the inner product ($D2$), $O(n^2 4^k)$.

Masked Sequences

Repeats such as SINE elements have a substantial influence on pairwise scores (horizontal lines in Figure 3.3B). Repeat-masking can be used to hide those repetitive elements by replacing nucleotides with the letter `N` (RepeatMasker (`www.repeatmasker.org`), TandemRepeatsFinder, (Benson, 1999)). For $N2$, any repeat-masked sequence is split into a set of repeat-free sub-sequences by cutting out all masked regions. Words are counted in this set such that no artificial words are created by concatenation. The length of the repeat-free sequence is estimated by (number of counted words) $+ k - 1$.

Instances of N2

The most basic instance of $N2$, with $n(w) = w$ will be referred to as $N2^*$. In the implementation of $N2$, $n(w)$ may be extended to include its reverse complement (rc),

$$n_{rc}(w) = \{w, rc(w)\}$$

all words equal to w with one mismatch (mm, hamming distance $dist_{hamming} \leq 1$),

$$n_{mm}(w) = \{w' | dist_{hamming}(w, w') \leq 1\}$$

or the combination of both (mm, rc), where

$$n_{mm,rc}(w) = \{w', rc(w') | dist_{hamming}(w, w') \leq 1\} \, .$$

In the following, these instances are referred to as $N2^{rc}$, $N2^{mm}$, $N2^{mm,rc}$. The word count of w (and its reverse complement when selected) is always weighted with $a_w = 1$, for all other words w' in $n(w)$ an alternative weight $a_{w'}$ may be chosen. The weights for mismatch neighbourhood counts are indicated in superscript, with $a_{w'} = 1$ ($N2^{mm(1.0)}$) if not stated otherwise. Note that the neighbourhood definition for $n_{mm}(w)$ and $n_{mm,rc}(w)$ only covers direct neighbours, not neighbours of neighbours.

Availability

The implementation that I provide for $N2$ is part of the SeqAn library (Doering *et al.*, 2008) where the fully documented source code and a pre-compiled executable version (ALF) is available for download (`http://www.seqan.de/projects/alf/`). ALF requires a set of n sequences in fasta-format as input and returns a matrix with all pairwise similarity scores. The word length k (default $k = 5$) and the background model order (default 1) may be chosen manually and the normalised standardised word neighbourhood counts may be returned to obtain additional information on important words (see Section 4.2).

Additionally, I implemented the $D2$, $D2^*$[1] and $D2z$[2] scores in the SeqAn library (Doering *et al.*, 2008) and ALF can be used to calculate these scores.

3.4 Results

3.4.1 Choice of Parameters for N2

The choice of parameters will influence the results obtained from alignment-free comparisons. For $N2$, the main parameters are the order of the Markov model m, the length of the k-mers k and the weights of the words in the neighbourhood (a_w).

Markov Model Order. Calculation of the expected value and variance of the word counts assumes that the background model that describes the sequence is known. For $N2$, the background model is estimated separately for every sequence using a maximum likelihood approach. Since `CpG` dinucleotides in mammalian genomic sequences are very rare (Gardiner-Garden and Frommer, 1987), a Bernoulli background model is insufficient to estimate word probabilities. This can be seen on simulations, where the first

[1] $D2^*$ was originally proposed with a Bernoulli background model for the computation of $\mu(w)$. The implementation provides the extended score that allows usage of higher order Markov background models.

[2] Repeat-masked sequences are treated equally for all methods. Note that this is slightly different to the original method proposed for $D2z$, which introduced artificial words by concatenating sequences.

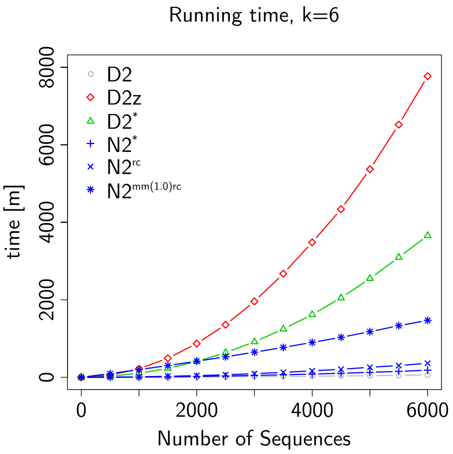


Figure 3.4: Running time comparison. All pairwise scores were calculated for random sequences of length 1000 bp, $k = 6$, Markov model of order 1.

order Markov model consistently outperforms the Bernoulli model (Figure 3.1). The optimal order for the Markov background model for enhancer sequences is an unknown function of organism complexity and sequence length. Due to the limited size of enhancer sequences, estimating higher order Markov models likely results in overfitting and poor estimates. This analysis will therefore rely on a first order Markov chain as background model for all methods throughout this analysis.

Word Length k and Word Neighbourhood Weights a_w. Since enhancer sequences have no apparent preferential orientation of transcription factor binding sites, reverse complement words are always weighted similar to the original words ($w' \in n_{rc}(w) \rightarrow a_{w'} = a_w$). For the mismatch neighbourhood variant $n_{mm}(w)$, the choice of a_w is connected to the choice of the word length k. Therefore, all combinations $k = 4, 5, 6$ and mismatch weights $a_w = \{1, 0.75, 0.5, 0.25, 0.1, 0.05, 0.01, 0.001\}$ were tested (Appendix, Figures VI.8, VI.9, VI.10). This analysis indicates that a_w should be larger for higher values of k where the expected number of k-mer occurrences is below 1. Different parameters might improve results for different data sets (Kantorovitz *et al.*, 2007). However, to have a consistent and comparable setup, I selected $k = 6$ and mismatch weights of 1 as reasonable parameters throughout the analysis. For completeness, results for $k = 4$ and $k = 5$ and different choices of a_w can be found in the Appendix (Figures VI.9,VI.10).

3.4.2 N2 can be Computed Quickly

Genome-wide data sets consist of many thousand regulatory sequences. The computation of pairwise similarities needs to be efficient for large-scale usage. The running

	Running time in O notation
$D2$	$O(nl + n^2 4^k)$
$D2z$	Kantorovitz *et al.* (2007)
$D2^*$	$O(n^2(l + 4^k + 4^m))$
$N2$	$O(n(l + 4^m + 4^k \text{NeighbourhoodSize}^2) + n^2 4^k)$

Table 3.1: Running time of the different methods in O-notation. n: number of sequences; l: average sequence length; k:k-mer size; m: Markov model order. The running time for $D2^*$ is dominated by the quadratic term. The running time for $N2$ is dominated by the linear term (pre-processing).

time of each method was estimated on sets with various numbers of sequences where the matrix of all pairwise similarities was computed (quadratic number of scores computed). The methods show strong differences in practise (Figure 3.4), but $N2$ and its variants are always faster than the other methods with a statistical model for realistically chosen numbers. Computing pairwise scores for 5000 enhancers with $k = 6$ takes 2 hours (h) for $N2^*$ (4h for $N2^{rc}$, 20h for $N2^{rc,mm}$), it takes about 42 h for $D2^*$ and 91 h for $D2z$.

The computation of $N2$ is dominated by the pre-processing step which scales linearly in the number of sequences since the neighbourhood counts are calculated once for every sequence in advance (Figure 3.4, Table 3.1, see Section 3.3.3). In contrast, $D2z$ and $D2^*$ cannot pre-compute normalised counts like $N2$, and scale quadratically in the number of sequences. $D2z$ calculates z-scores on pairs of sequences which are not pre-processed (Kantorovitz *et al.*, 2007), and $D2^*$ calculates the background model on the concatenation of sequences which cannot be pre-computed (Reinert *et al.*, 2009). While this is likely to increase the accuracy of the model, running times are drastically higher. Computing pairwise scores for realistically large data sets is therefore nearly impossible for both $D2z$ and $D2^*$. This makes the $N2$ score very attractive for large-scale applications such as classification of regulatory sequences, or applications that support pre-computed data structures such as database searches.

3.4.3 N2 is Robust Against Single Sequence Noise

Ideally, the pairwise score between two sequences should reflect the sequences' similarity. However, in practise, word count-based methods can be heavily influenced by noise specific to individual sequences, meaning that some sequences will intrinsically have high (or low) scores (Lippert *et al.*, 2002; Reinert *et al.*, 2009). Without proper correction, the pairwise score is an attribute of the individual sequence rather than of the pair of sequences. This is especially prominent for $D2$, where a high number of occurrences of

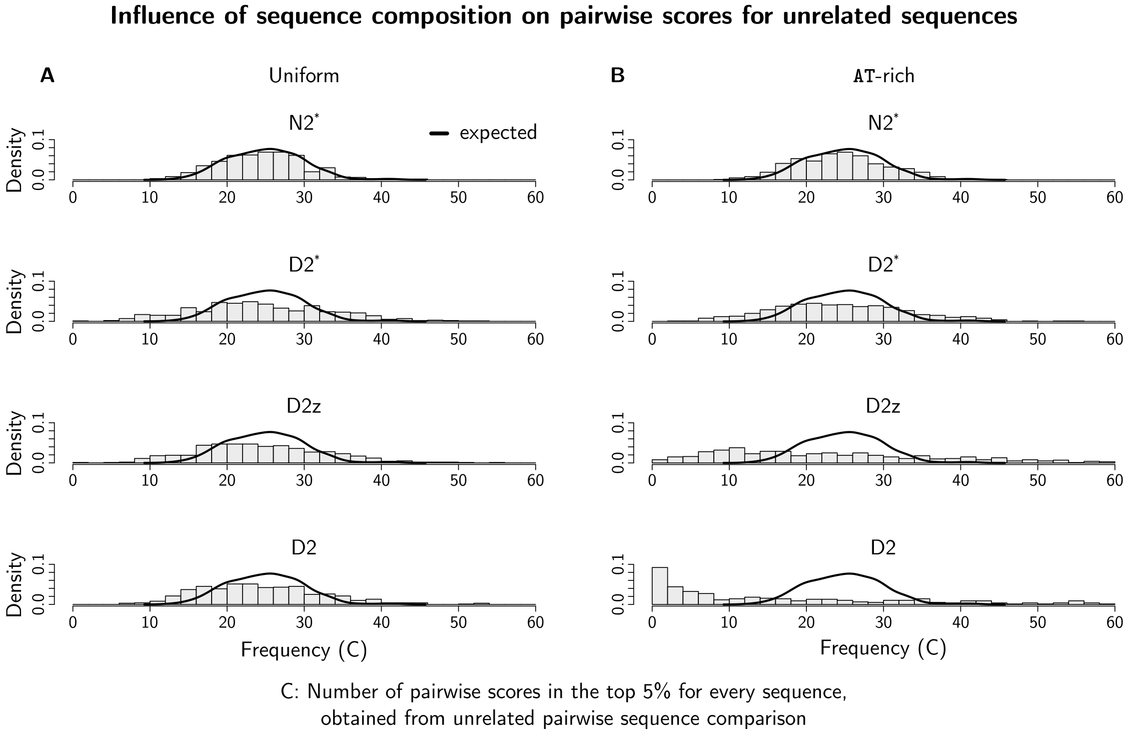


Figure 3.5: Influence of single sequences on pairwise scores. All pairwise scores for 500 sequences generated by the same model were calculated. C_i measures the number of sequence pairs for sequence S_i among the highest 5% of all scores ('high scoring pairs'). Since all sequences were created using the same model, the distribution of $C = \{C_1, \ldots, C_i\}$ from alignment-free methods should be similar to the distribution of C obtained from a random scoring method ('expected', black line). A different distribution would indicate that the number of high scoring pairs is strongly dependent on the individual sequence, indicating that pairwise scores are dependent on the single sequence noise rather than on the similarity of the sequence pair. (A) Uniform nucleotide distribution, all methods show the expected behaviour. (B) AT rich nucleotide distribution, $D2$ and $D2z$ differ from the expected behaviour, showing that these pairwise scores are strongly influenced by the sequence composition.

a repetitive self-overlapping word (such as AAAAA) in one sequence will always induce high pairwise scores.

To quantify the influence of single-sequence-specific noise on pairwise scores, I studied the behaviour of $D2$, $D2z$, $D2^*$ and $N2^*$ for scoring pairs of unrelated sequences simulated by the same background model. Scores for all sequence pairs (S_i, S_j) were calculated for 500 such unrelated sequences. A threshold t was chosen to select the top 5% highest scoring sequence pairs ('high scoring pairs'). For every sequence S_i, the number of high scoring pairs is defined as C_i: $C_i = \sum_j \mathbf{1}(\text{score}(S_i, S_j) \geq t)$. Since all sequences were generated by the same model, the expected value of C_i, $\mathbb{E}[C_i]$, is equal for all sequences S_i. Here, 5% of the 499 sequence pairs of S_i are expected to have a score greater than t, thus $\mathbb{E}[C_i] = 24.95$. As a reference, $C = \{C_1, \ldots, C_i\}$ was calculated for sequence pairs with randomly assigned scores. This method is not influenced by the sequence at all and therefore recapitulates the expected behaviour for unrelated sequence pairs (Figure 3.5, black line). Then C was calculated for the four alignment-free sequence comparison methods.

The distribution of C when $N2^*$ is used is close to the expected distribution for unrelated sequences (Figure 3.5). This shows that $N2$ is robust against single-sequence-specific noise as the numbers of high scoring sequence pairs are not influenced by the individual sequences (see Supplementary Figures VI.6 and VI.6 for $N2^{rc}$ and $N2^{mm,rc}$).

In contrast, $D2$ and $D2z$ show a very different distribution of C from the expected behaviour in the non-uniform case. Figure 3.5 B shows that the number of high scoring pairs strongly varies, suggesting that the expected number for C_i is different for every sequence S_i, even though all sequences were generated by the same model. This shows that the number of high scoring pairs detected with these methods is strongly influenced by the individual sequence, indicating that pairwise scores measure the individual sequence composition and not the similarity of the sequence pair. Prior work comparing regulatory sequences using alignment-free methods did not consider this effect (Kantorovitz *et al.*, 2007; Dai *et al.*, 2008). The above results confirm that neither the $D2$ nor the $D2z$-score should be applied to real biological sequences (Lippert *et al.*, 2002; Reinert *et al.*, 2009).

Other sequence noise such as repeats and regions of low complexity occurs frequently in genomic data. $N2$ is more robust to this type of noise than $D2^*$ and $D2z$ due to its correction for word overlaps and normalisation of counts (Appendix, Table VI.3). This analysis suggests that $N2$ should be used when repeat-masking is not an option.

3.4.4 Pairwise Comparison: Simulations

First, I tested the performance of $N2$ on simulated data. Random sequences were generated with a similar dinucleotide content as the mouse genome (Thomas-Chollier *et al.*, Performance with implanted k-mers, random strand

	5%-Precision		AUC ROC		AUC PR	
Motif setting:	m1r8	m4r2	m1r8	m4r2	m1r8	m4r2
$D2$	0.88	0.59	0.72	0.54	0.72	0.54
$D2z$	0.91	0.64	0.74	0.56	0.73	0.56
$D2^*$	0.87	0.66	0.71	0.58	0.70	0.57
$N2^*$	0.86	0.65	0.71	0.58	0.70	0.57
$N2^{rc}$	**0.93**	**0.71**	**0.77**	**0.60**	**0.77**	**0.59**

Table 3.2: Simulations with implanted k-mers. Comparison of the different methods ($k = 6$, $mo = 1$) when the genomic orientation of the motif is unknown. Bold numbers indicate best performance.

2011) (mm9) as background sequences ('negative set'). Then, m randomly chosen motifs of length 5 were implanted r times into the same background sequences to simulate enhancers ('positive set'; m1r8: $m = 1$, $r = 8$; m4r2: $m = 4$, $r = 2$). Following Kantorovitz *et al.* (2007), all pairwise scores were computed for the corresponding negative and the positive sets. The pairwise scores from the negative and the positive sets were then combined and ranked. Based on this ranked list, the performance of the $D2$-based methods for pairwise sequence comparison was compared using the area under ROC curve (AUC ROC) and area under Precision-Recall curve (AUC PR). Furthermore, the interpolated precision at 5% recall was estimated, '5%-Precision' for short. Results show average values over 25 simulations, each time drawing 100 random sequences of length 1000 bp and inserting random motifs, thus covering different motif compositions in an unbiased way. The performance was tested with word size $k = 6$ using a first order Markov model for word probabilities (see Appendix, Tables VI.1 and VI.2 for $k = 5$).

Two different settings were simulated to evaluate the performance of the neighbourhood concept of $N2$. First, randomly sampled 5-mers were implanted into the forward and backward strand of the sequences to simulate the orientation independence of binding sites in enhancers. The $N2^{rc}$ variant was specifically designed for this scenario and, indeed, $N2^{rc}$ performs best (Table 3.2). Second, words were randomly sampled and implanted with one mismatch at a random position to simulate more flexible motifs. The $N2^{mm}$ variant was designed for this scenario as it considers the word neighbourhood for the similarity. In these simulations, the $N2^{mm}$ variant with mismatch weights $a_w = 1.0$ shows the best performance, demonstrating the value of neighbourhood counts to score sequences with approximate word matches (Table 3.3, see Appendix Figure VI.7 for different choices of a_w). These simulations confirm the value of extending exact word count methods to word neighbourhoods.

Performance with implanted k-mers, mismatch

	5%-Precision		AUC ROC		AUC PR	
Motif setting:	m1r8	m4r2	m1r8	m4r2	m1r8	m4r2
$D2$	0.59	0.51	0.53	0.48	0.53	0.49
$D2z$	0.59	0.54	0.54	0.51	0.53	0.51
$D2^*$	0.60	0.54	0.54	0.51	0.54	0.51
$N2^*$	0.59	0.54	0.54	0.51	0.54	0.51
$N2^{mm(0.01)}$	0.60	0.54	0.55	0.51	0.54	0.51
$N2^{mm(1.0)}$	**0.65**	**0.55**	**0.57**	**0.52**	**0.57**	**0.53**

Table 3.3: Simulations with implanted k-mers. Comparison of the different methods ($k = 6$, $mo = 1$) when motifs are sampled from all k-mers with one mismatch to the word. Bold numbers indicate best performance.

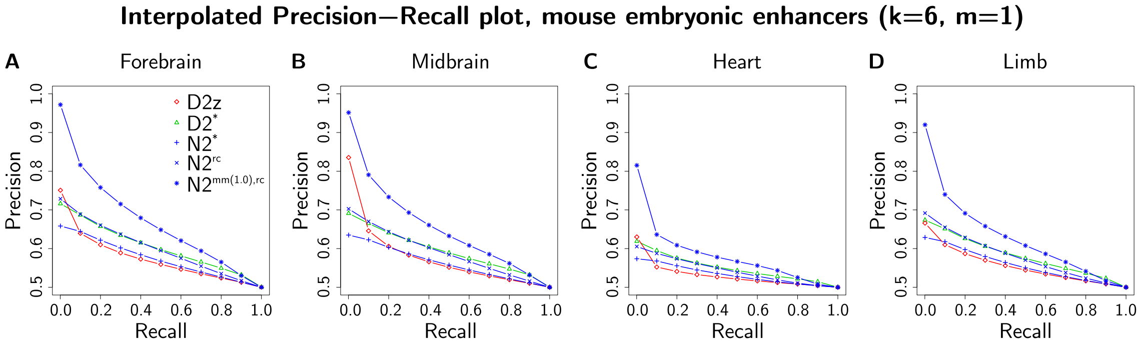


Figure 3.6: Precision-Recall curve for enhancers active during mouse development. The plots show the precision average over 25 samples each time drawing 500 enhancer sequences ('positive') and 500 unrelated genomic sequences of equal length as the enhancers ('negative'). **(A)** Forebrain. **(B)** Midbrain. **(C)** Heart. **(D)** Limb.

3.4.5 Pairwise Comparison of Developmental Enhancers

The above simulations demonstrated the ability of $N2$ to distinguish artificial enhancers from unrelated sequences. Currently, our knowledge on regulatory sequences is limited and simulations can only approximate the real nature of enhancers. Tissue-specific enhancers in mouse embryos have been identified in a genome-wide manner using the co-activator protein p300 (Visel *et al.* (2009); Blow *et al.* (2010), Chapter 2). Here, these data sets are used to test whether alignment-free methods are able to discriminate *in vivo* identified enhancers that show similar activity from genomic background. Enhancers active in forebrain, midbrain, limb and heart tissue of the developing mouse embryo were used as as positive sets (Visel *et al.*, 2009; Blow *et al.*, 2010). All se-

Performance on tissue-specific enhancer sequences

	5%-Precision				AUC ROC				AUC PR			
Tissue:	F	M	L	H	F	M	L	H	F	M	L	H
$D2$	0.61	0.64	0.55	0.50	0.55	0.55	0.50	0.45	0.54	0.55	0.51	0.47
$D2z$	0.66	0.69	0.63	0.56	0.57	0.57	0.56	0.53	0.57	0.57	0.55	0.52
$D2^{*}$	0.71	0.70	0.67	0.60	0.62	0.62	0.59	0.55	0.60	0.60	0.58	0.54
$N2^{*}$	0.65	0.64	0.62	0.58	0.58	0.57	0.56	0.53	0.57	0.56	0.55	0.53
$N2^{rc}$	0.71	0.67	0.68	0.60	0.61	0.59	0.58	0.55	0.60	0.58	0.58	0.55
$N2^{mm(1.0),rc}$	**0.84**	**0.82**	**0.79**	**0.66**	**0.66**	**0.64**	**0.63**	**0.57**	**0.66**	**0.64**	**0.63**	**0.57**

Table 3.4: Comparison of the different methods on tissue-specific enhancers. Bold numbers indicate the best performance. Positive sequences were obtained by ChIP-Seq of p300 in forebrain (F), midbrain (M), limb (L), and heart (H) tissue of the mouse embryo. Negative sequences were randomly sampled from the mouse genome. All pairwise scores were computed with repeats masked, $k = 6$, background Markov model of order 1. Results show average values over 25 samples each time drawing 500 sequences.

quences were obtained from the UCSC pre-masked genome sequence (mm9, RepeatMasker (`www.repeatmasker.org`) and TandemRepeatsFinder (Benson, 1999)). Pairwise scores from these tissue-specific enhancers were compared with pairwise scores from genomic sequences of the same length randomly sampled from the mouse genome, ensuring a maximum of 30% of repetitive sequence for every negative sample. To obtain accurate estimations, the average over 25 samples was calculated, each time drawing 500 sequences from the positive set. Using the same evaluation measures as in the previous section, the ability of alignment-free sequence comparison methods to detect functional similarity of regulatory sequences was measured.

Figure 3.6 and Table 3.4 show the results for pairwise comparison of tissue-specific enhancers with alignment-free methods. Across all tissues, $N2^{mm(1.0),rc}$ gives the best results, demonstrating that $N2$ is most suitable to detect tissue-specific activity of regulatory sequences. The results also confirm the value of the word neighbourhood concept: comparing $N2^{rc}$ with $N2^{*}$ shows that the neighbourhood extension to the reverse complement is always preferable (Table 3.4). Extending the word neighbourhood to all words with one mismatch ($N2^{mm(1.0),rc}$) further improves the results by 6-15% (Table 3.4). These results support the usage of $N2$ with word neighbourhood counts to score the similarity of regulatory sequences.

Tissue-Specificity of Enhancers

The above results indicate that tissue-specific enhancer sequences indeed have a similar word content. However, a comparison of ChIP-Seq data with randomly sampled genomic

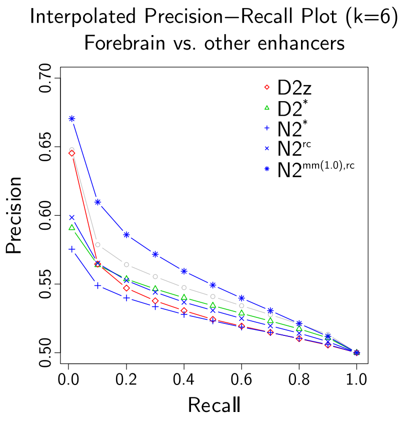


Figure 3.7: Tissue-specificity of enhancers. Precision-Recall curve for forebrain enhancers in the mouse. Enhancers active in different tissues were used as the background set.

sequences might be biased towards measuring similarities introduced by the technology, such as similar GC content. To test this, I verified whether it is possible to discriminate enhancers according to the tissue where they drive expression. For that purpose, all pairwise scores of enhancers active in the same tissue (‘positive set’) and all pairwise scores between enhancers active in other tissues (‘negative set’) were calculated, discarding all enhancers active in multiple tissues. To correct for length differences between data sets from different tissues, 750 bp in the middle of the reported enhancer sequences were selected. Figure 3.7 shows that tissue-specific enhancers can be discriminated by alignment-free methods (see Appendix, Figure VI.11 for the other data sets and $k = 5$). While the performance decreases compared to using random sequences as the negative set, these results show that activity in a similar tissue is indeed reflected in a higher sequence similarity. Again, the neighbourhood extensions of $N2$ improves the results, further highlighting the value of this concept for regulatory sequences.

3.5 Discussion

Section 3.4 showed that $N2$ improves alignment-free sequence comparison through its flexible extension to word neighbourhood counts, thereby covering approximate and orientation independent word matches. Previously, the $D2z$ score has been extended to allow for approximate matching words using estimates for the expectations and the variances based on a Bernoulli background model, however, no implementation is available (Forêt *et al.*, 2006; Burden *et al.*, 2008; Forêt *et al.*, 2009; Burden *et al.*, 2012). The framework that is presented here is much more general and powerful. $N2$ allows for any desired word neighbourhood and associates words with weights such that the signal of words matching exactly is not lost. Furthermore, $N2$ can be computed on any background model order, which is essential to properly describe genomic sequences. Finally, $N2$ is much faster than $D2z$ even without approximate matching, suggesting that a z-score calculation for an approximate $D2$ score would be infeasible for any data set of realistic size.

The differences between $N2^*$ as used in this study and $D2^*$ are mainly due to the estimation of the background model. The better performance of $D2^*$ on real biological data suggests that the concatenation of the sequences improves the accuracy of the background model. However, it drastically increases the running time. The improvement due to the extension to the word neighbourhood ($N2^*$ vs. $N2^{mm,rc}$) is better than the improvement due to different background model estimates ($N2^*$ vs. $D2^*$, see Table 3.4).

The simulation studies demonstrated that $N2$ performs well on the task it was designed for, namely finding similarities between sequences based on shared words. Extending the word neighbourhood to the reverse complement ($N2^{rc}$) improves the performance, highlighting that binding sites can occur on both strands of the enhancer. Extending the neighbourhood to words with one mismatch ($N2^{rc,mm}$) further improves the performance on experimentally identified enhancers. This suggests that there are subtle signals like a common content of similar but not equal words which are characteristic of genomic enhancers.

For many transcription factors, DNA binding motifs have been identified (Matys *et al.*, 2003; Bryne *et al.*, 2008). Known transcription factor binding motifs have been used for detection of regulatory sequences (Klein and Vingron, 2007), identification of interacting transcription factors (Mysickova and Vingron, 2012) or comparison of regulatory sequences (Koohy *et al.*, 2010). If the binding motif for regulatory sequences is known, incorporation of these motifs will improve the power of regulatory sequence comparison. However, the majority of binding sites in enhancers is unknown, and integration of annotated binding sites might easily introduce biases. Therefore, $N2$ is based on sequence information alone to provide a similarity measure that is independent of prior knowledge.

Here it is assumed that a high number of shared words represents a similar binding site content of enhancers. This assumption is violated by repeats, having a high number of shared words only due to high sequence similarity. For this reason, repeats are masked before calculating pairwise scores. Although some transcription factor binding sites have been found in repetitive sequences (Kunarso *et al.*, 2010; Zemojtel *et al.*, 2009), the sequence similarity of repeats is largely unrelated to regulatory activity and will eclipse any shared word count from common DNA binding motifs. The usage of repeat masked sequences is therefore generally recommended when comparing regulatory elements.

$N2$ is defined as the inner product of the standardised word count vectors. Many other methods have been proposed and applied to compare word count vectors (Section 3.2.2). Similarly, the neighbourhood word counts used for $N2$ could be combined in other ways, for example using the euclidean distance to obtain a distance measure. A ranked list-based comparison of standardised word counts could also be applied to minimise the influence of words with small weights. Here, the inner product was used for $N2$ as it guarantees specific properties which are required for large-scale applications such as classification (Chapter 4).

The idea of the $N2$ score is that enhancers which share the same binding sites have a similar word content. $N2$ is based on standardised word counts, therefore every word is associated with a k-mer weight that corresponds to a z-score. These scores can be used to identify over-represented words and reconstruct the binding sites within enhancers. This way, the $N2$ k-mer weights can directly be used to identify the motifs that lead to the increased similarity of tissue-specific enhancers (Section 4.2).

Alignment-free methods have been used to predict *cis*-regulatory modules in flies and mouse (Kantorovitz *et al.*, 2009; Lee *et al.*, 2011). The results on pairwise comparison of enhancers suggests that the $N2$ similarity could as well be used to predict the regulatory outcome of enhancers. $N2$ fulfils all properties of a Kernel function and can therefore directly be used for support vector machine based classification and prediction (Section 4.3).

The tissue specific enhancers used in this study were identified by binding of the enhancer-associated protein p300. However, the transcription factors that bind to these enhancers are unknown. Even for enhancers from the same tissue it is likely that different sets of transcription factors are recruited and that the set of transcription factor binding sites varies. The performance of any task that relies on tissue-specific enhancers (such as pairwise comparison and classification) is therefore limited by the heterogeneity of the data sets. $N2$ provides a similarity measure for enhancers and can be used to identify more homogeneous clusters in the data (Section 4.4). Usage of such homogeneous data sets might further improve detection and prediction of tissue-specific regulatory elements.

3.5.1 Conclusion

In this Chapter, $N2$ was presented, a novel alignment-free measure of sequence similarity that overcomes the limitations imposed by traditional exact word count-based methods. $N2$ includes the general concept of weighted word neighbourhood counts and it improves the ability to detect similarity between regulatory sequences. The task of pairwise comparison of regulatory sequences is much harder than traditional pairwise alignment since only very few shared words might lead to a similar activity. Application of $N2$ to large-scale data sets of mammalian enhancers demonstrated that pairwise sequence similarity of non-homologous regulatory elements is able to estimate similar *in vivo* activity. The observation that word count-based similarity measures are able to detect tissue-specific activity of enhancers suggests that enhancers contain scattered binding sites that contribute to their tissue-specificity. This result supports the importance of combinatorial binding in transcriptional regulation (Chapter 2). Availability of methods such as $N2$ which are able to compare enhancers based on sequence information alone will greatly help to improve our understanding of the sequence-dependent regulatory code that enables the establishment of a large diversity of cell types coded in one genomic sequence.

4 Large-Scale Analysis of Developmental Enhancer Sequences

4.1 Introduction

Binding of transcription factors is one of the key events that governs transcriptional regulation and gene expression. Transcription factors recognise specific DNA sequences to regulate gene expression. These DNA binding motifs are required for many cellular processes, and consequently they show higher levels of sequence conservation (He *et al.*, 2011). The combination of transcription factor binding events (Chapter 2) and the combination of binding patterns (Chapter 3) define the cell type-specific activity of regulatory elements. The developmental enhancers analysed in Chapters 2 and 3 were identified by the enhancer binding protein p300, which does not bind to the DNA directly. Instead, intermediate transcription factors bind to those enhancers and define their activity. Enhancers which are active in the same tissue show higher sequence similarity, but the transcription factor binding sites are still unknown. Furthermore, it is likely that these binding sites vary and that tissue-specific enhancers represent a heterogeneous data set of enhancers with only partially similar binding sites. In order to better understand gene regulation during development, the sequence-specific properties of tissue-specific developmental enhancers have to be studied.

Availability of genome-wide data provides the possibility to use tools such as motif finding, or classification. Chapter 3 aimed at estimation of the similarity of two enhancer sequences (pairwise comparison) without any additional knowledge. While both motif finding and classification require much larger data sets and would fail for this task, $N2$ can easily be used and extended to the analysis of large-scale data sets.

First, a $N2$-based motif finding algorithm is presented. This algorithm is applied to large-scale enhancer data sets to identify transcription factor binding sites. Secondly, $N2$ is used as a kernel function for support vector machine (SVM) classification and prediction of enhancers. Finally, heterogeneity of tissue-specific enhancers is investigated by combining clustering with the $N2$ measure of similarity. All of these tasks are different in their goal, yet connected through the usage of the same $N2$-based word statistics.

4.2 Motif Finding: N2 Word Ranking

The task of *motif finding* is to identify statistically over-represented signals in a set of sequences. Transcription factors regulate gene expression by binding to specific DNA sequences, finding transcription factor binding sites in enhancers is therefore an instance of the motif finding problem. The DNA binding sequence of transcription factors can be identified experimentally with the ChIP-Seq technology which results in a large set of sequences bound by the protein of interest (Figure 1.7). The DNA binding motif of the transcription factor can then be identified by extracting over-represented words.

The set of enhancers analysed in Chapters 2 and 3 act by recruiting specific transcription factors, which activate or repress gene expression. The developmental enhancers were identified by binding of p300, but the transcription factors which are recruited to the enhancers remain unknown. Knowledge of the DNA motifs in these enhancers is a first step to identifying these transcription factors and to understand the functionality of tissue-specific regulatory sequences.

Motif finding algorithms usually have two phases, derivation of an initial motif, and subsequent refinement. MEME (Bailey *et al.*, 2009), one of the most widely used tools for motif finding, uses an expectation-maximisation algorithm. MEME makes prior assumptions on how and where motif occurrences appear, such as one occurrence per sequence (OOPS). While MEME is sensitive for finding motifs, it cannot be used for large-scale data sets with several thousand sequences. Many other motif finding algorithms have been developed, some of which are specifically aimed at analysing large-scale data sets (Sinha and Tompa, 2002, 2003; Tompa *et al.*, 2005; Pavesi *et al.*, 2006; Chakravarty *et al.*, 2007; Bailey, 2011; Machanick and Bailey, 2011; Huggins *et al.*, 2011; Thomas-Chollier *et al.*, 2012; Ma *et al.*, 2012), see Das and Dai (2007) for an overview.

Identification of over-represented words is the basis for alignment-free sequence comparison. The $N2$ method (Chapter 3) calculates z-scores for the number of occurrences for every possible word of a specific length in a sequence. Instead of calculating pairwise similarity, these z-scores can be directly used as a measure for over-representation. Consequently, the standardised word counts from $N2$ provide the starting point for a motif finding algorithm.

4.2.1 Algorithm: ALF-M

The $N2$-based motif finding algorithm (ALF-M) starts with calculating z-scores for all words of length k and returns the motifs associated with highest over-represented words. Similar to other algorithms, ALF-M can be divided into two steps, derivation of a starting pattern (steps 1-4, Figure 4.1) and refinement to obtain a more precise motif description (steps 5-6, Figure 4.1).

1. **Calculate k-mer weights**
   Firstly, all words of a given length k (k-mers) are counted in the set of input sequences. As described in Chapter 3, the standardised word neighbourhood counts are calculated and normalised (further referred to as k-mer weights). The k-mer weights provide the estimation on the over-representation for every word (z-scores).

2. **Rank k-mers**
   All k-mers are ranked according to their weight. The n k-mers with the highest weight will provide the initial seed k-mers for steps 3-6.

3. **Extend k-mers**
   Recursively, the k-mers with the highest weights are aligned to the seed k-mer to obtain an extended pattern.

4. **Calculate Weighted PFM**
   A position frequency matrix (PFM) is calculated from the extended k-mer alignment. In this step, the k-mer weights are used to obtain a weighted PFM. Based on the PFM, a consensus motif is computed where all positions below a threshold (0.8) are masked with the wild-card letter `N`.

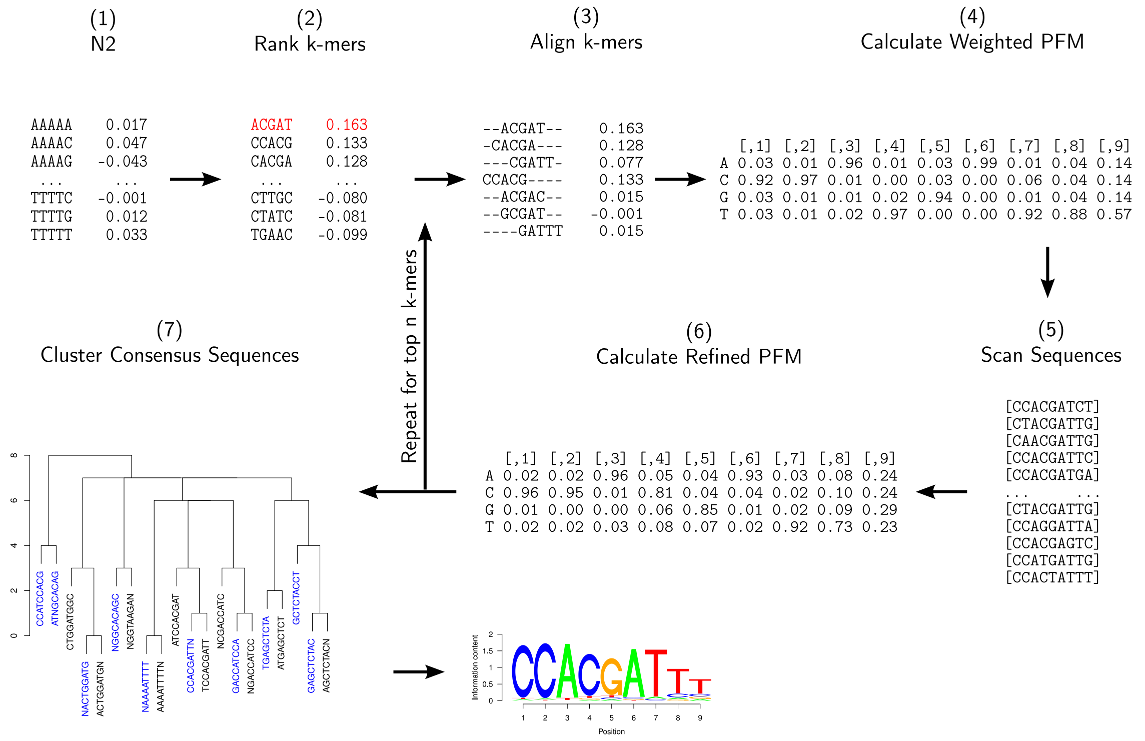
5. **Scan Sequences**
   The consensus motif is now used to scan all sequences for occurrences. By default, two mismatches are allowed in the refinement step, however, this can be adjusted freely.

6. **Calculate Refined PFM**
   Based on all matches of the consensus motif in the set of sequences, the refined position frequency matrix is calculated.

   Steps 3-6 are repeated using the n highest ranking k-mers as seeds.

7. **Cluster motifs and return non-redundant hits**
   Since the highest ranking k-mers might represent sub-words from the same motif, redundant words have to be identified. This is achieved by clustering of all consensus motifs. The tree is cut into m sub-clusters, from which a single motif is returned.

Significance of the Motifs. The k-mer weights are standardised word counts. Every word count is a sum of (dependent) random variables. If word overlaps and nucleotide dependencies are ignored, the word count forms a sum of independent Bernoulli variables. In that case, the central limit theorem states that the standardised word counts



```
AAAAA    0.017
AAAAC    0.047
AAAAG   -0.043
 ...      ...
TTTTC   -0.001
TTTTG    0.012
TTTTT    0.033
```

```
ACGAT    0.163
CCACG    0.133
CACGA    0.128
 ...      ...
CTTGC   -0.080
CTATC   -0.081
TGAAC   -0.099
```

```
--ACGAT--     0.163
-CACGA---     0.128
---CGATT-     0.077
CCACG----     0.133
--ACGAC--     0.015
--GCGAT--    -0.001
----GATTT     0.015
```

```
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
A 0.03 0.01 0.96 0.01 0.03 0.99 0.01 0.04 0.14
C 0.92 0.97 0.01 0.00 0.03 0.00 0.06 0.04 0.14
G 0.03 0.01 0.01 0.02 0.94 0.00 0.01 0.04 0.14
T 0.03 0.01 0.02 0.97 0.00 0.00 0.92 0.88 0.57
```

```
[CCACGATCT]
[CTACGATTG]
[CAACGATTG]
[CCACGATTC]
[CCACGATGA]
...      ...
[CTACGATTG]
[CCAGGATTA]
[CCACGAGTC]
[CCATGATTG]
[CCACTATTT]
```

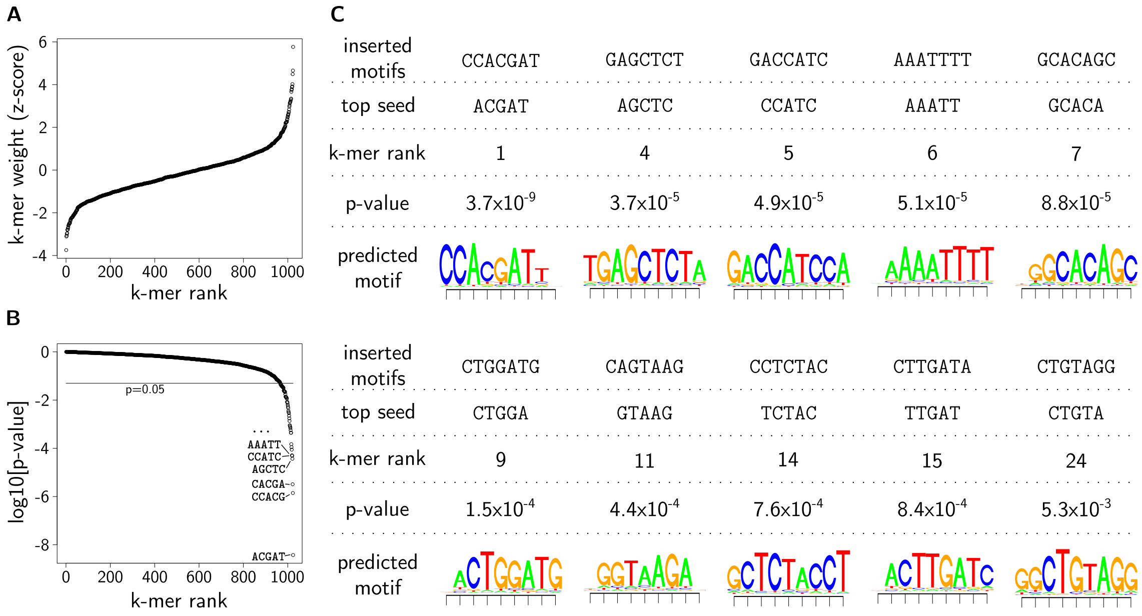
```
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
A 0.02 0.02 0.96 0.05 0.04 0.93 0.03 0.08 0.24
C 0.96 0.95 0.01 0.81 0.04 0.04 0.02 0.10 0.24
G 0.01 0.00 0.00 0.06 0.85 0.01 0.02 0.09 0.29
T 0.02 0.02 0.03 0.08 0.07 0.02 0.92 0.73 0.23
```

Figure 4.1: Motif finding with N2 word statistics: ALF-M. For every k-mer a z-score is calculated that estimates the level of over-representation (k-mer weights). The k-mers with the highest weights are used as seeds, which are extended and refined to obtain a PFM description of the motif. Finally, a clustering approach is used to return a set of non-redundant motifs.

converge to the standard Gaussian distribution (Reinert *et al.*, 2009). The asymptotic normality of the word counts can be established for higher order Markov models as well (Robin *et al.*, 2005). Therefore, for large sequences, the standardised word counts follows an approximately standard normal distributed z-score, for which a significance estimate (p-value) can be calculated. For ALF-M, the p-value of the motif is estimated by the probability to observe the same or a higher z-score for the seed k-mer by chance.

Parameters. The parameters that can be decided by the user are the word length k for the initial ranking, the number of k-mers n which will be used as seeds, the number of mismatches d for the refinement step, and the number of motifs m which will be returned after the clustering. The length e for the extension and alignment step can also be chosen freely (default 2).

Running Time. The running time for ALF-M is dependent on the total length of all



inserted motifs	CCACGAT	GAGCTCT	GACCATC	AAATTTT	GCACAGC
top seed	ACGAT	AGCTC	CCATC	AAATT	GCACA
k-mer rank	1	4	5	6	7
p-value	$3.7x10^{-9}$	$3.7x10^{-5}$	$4.9x10^{-5}$	$5.1x10^{-5}$	$8.8x10^{-5}$
predicted motif					

inserted motifs	CTGGATG	CAGTAAG	CCTCTAC	CTTGATA	CTGTAGG
top seed	CTGGA	GTAAG	TCTAC	TTGAT	CTGTA
k-mer rank	9	11	14	15	24
p-value	$1.5x10^{-4}$	$4.4x10^{-4}$	$7.6x10^{-4}$	$8.4x10^{-4}$	$5.3x10^{-3}$
predicted motif					

Figure 4.2: ALF-M correctly identifies motifs in simulated sequences (1000 sequences of length 1000 bp, 10 motifs of length 7 bp inserted each into 100 sequences. **(A)** K-mer ranking by k-mer weights (z-scores). **(B)** P-value estimates for every k-mer. **(C)** Implanted motifs, top k-mer that is part of this motif with k-mer rank and p-value, and the motif predicted by ALF-M.

sequences l, the order of the background Markov model mo, the choice of k, the size of the word neighbourhood, and the number of k-mers used as seed n:

$$O(l + 4^{mo} + 4^k \text{NeighbourhoodSize}^2 + nl + n)$$

In practise, ALF-M is very fast as the running time is linear in the length of sequences analysed. For example, finding 20 motifs in 2000 sequences of length 500 using $k = 5$ and $mo = 3$ takes only a few seconds.

4.2.2 Motif Finding: Simulations

To test the ability of ALF-M to identify motifs I sampled sequences of length 1000 bp with a similar word composition as the mouse genome (Thomas-Chollier *et al.*, 2011). 10 random words of length 7 were implanted each into 100 sequences ($m1r1$) and ALF-M was used to identify these 10 motifs in the combined data set of 1000 sequences (Figure 4.2).

Within the top 24 k-mers, ALF-M returned significant motifs for all implanted words. Among these 24 k-mers, only 5 k-mers were not fully part of an inserted motif. In total, ALF-M identified 59 out of all 1024 k-mers as significantly over-represented (p-value

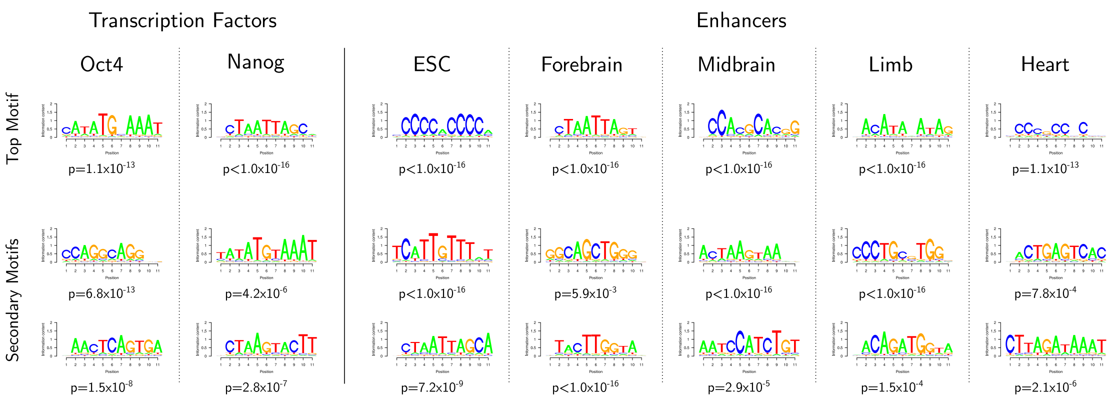


Figure 4.3: Motif finding with ALF-M identifies known and novel transcription factor binding motifs. ALF-M can be applied to data which is specific for a single transcription factor (left) and to enhancers where the transcription factors are unknown (right).

$<$0.05). 26 from all possible 30 5-mers that resemble the inserted motifs were among those significant k-mers, and 37 overlapped by at least 4 nucleotides. This shows that the p-value estimates from ALF-M are good indicators of statistically over-represented words.

The k-mer based approach of ALF-M leads to identification of motifs which are longer than the inserted words. This can be changed by choosing smaller values for the word length k or extension length e, without affecting the k-mer ranking. These result demonstrate the sensitivity of ALF-M and the $N2$ based word ranking, as it accurately identified words of length 7 which occur only once in 10% of all sequences.

4.2.3 Motif Finding: Enhancers

In chapters 2 and 3, two different kinds of data sets were analysed: binding data from transcription factors such as Oct4 and Nanog, and binding data from the general enhancer binding protein p300. For Oct4, and Nanog, the binding motif is known (Schöler *et al.*, 1989, 1990; Loh *et al.*, 2006). In contrast, the transcription factors that bind at enhancers identified with p300 are unknown. I applied ALF-M to both kinds of data sets (Marson *et al.*, 2008; Visel *et al.*, 2009; Blow *et al.*, 2010; Creyghton *et al.*, 2010), selecting 2000 sequences of length 500 bp each time, to measure its ability to recover known and novel transcription factor binding motifs.

Oct4 binds to the octamer motif (`ATGCAAAT`). ALF-M identifies this motif as the highest scoring motif in the Oct4 data set (see Figure 4.3 for a summary of the motifs identified with ALF-M in all data sets). Nanog is a homeobox transcription factor, and the primary motif for Nanog identified by ALF-M is the homeobox binding motif.

Nanog frequently co-localises with Oct4 in embryonic stem cells (Chapter 2) and ALF-M identifies the octamer motif among the significant secondary motifs in the Nanog binding data. Several other secondary motifs were identified by ALF-M confirming the power of the $N2$ word count based approach for motif finding in transcription factor binding data.

The enhancer binding protein p300 is not associated with a single transcription factor, therefore motif finding in these enhancers might help identifying different factors bound in the distinct tissues. In embryonic stem cells, p300 binds to enhancers were Oct4, Sox2 and Nanog binds (Chen *et al.*, 2008; Creyghton *et al.*, 2010). The Sox2 binding motif (`CATTGTT`, Chen *et al.* (2008)) is identified in the p300 binding data, as is the primary Nanog binding motif (Figure 4.3, ESC). Interestingly, the primary motif for the p300 data in embryonic stem cells resembles the Klf4 motif (`CCCACC`). This might reflect that p300 frequently binds to promoters, and Klf4 is an important, promoter-specific transcription factor involved in the maintenance of pluripotency (Chen *et al.*, 2008). This analysis of p300 binding data from embryonic stem cells shows that ALF-M is able to identify motifs which correctly reflect the expected binding combinations (Chapter 2).

Much less is known about enhancers from mouse forebrain, midbrain, limb and heart tissue (Thomas-Chollier *et al.*, 2012). Enhancers from similar tissues can be compared using $N2$ (Chapter 3), which indicates that these are bound by specific transcription factors. In contrast to pairwise comparison, which is based on only two sequences, motif finding can use a much larger data set. Applied to these developmental enhancers, ALF-M returns homeobox motifs (`TAAT`), E-Box motifs (`CAGCTG`), GATA motifs (`GATA`) and many other well known transcription factor binding sequences (Figure 4.3). Interestingly, motifs identified in the Nanog data sets can be identified in forebrain and midbrain enhancers. This is in line with results presented in Chapter 2, where it was shown that combinatorial binding of Nanog with Oct4 and Sox2 identifies highly conserved developmental enhancers. Identification of shared motifs in the Nanog and p300 data sets further supports the notion of gene regulatory hotspots, highly conserved regulatory elements that are bound in multiple tissues.

4.2.4 ALF-M Identifies Conserved Motifs in Developmental Enhancers

The key step for motif finding with ALF-M is computation of the standardised word counts ($N2$). All motifs that are reported are essentially extensions of the k-mers with the highest weights (z-scores). In order to estimate the value of the k-mer ranking, I investigated the sequence conservation (PhyloP) of all words in the sequences (Pollard *et al.*, 2010).

The top motif identified in forebrain enhancers (`CTAATTA`) is the refined, extended

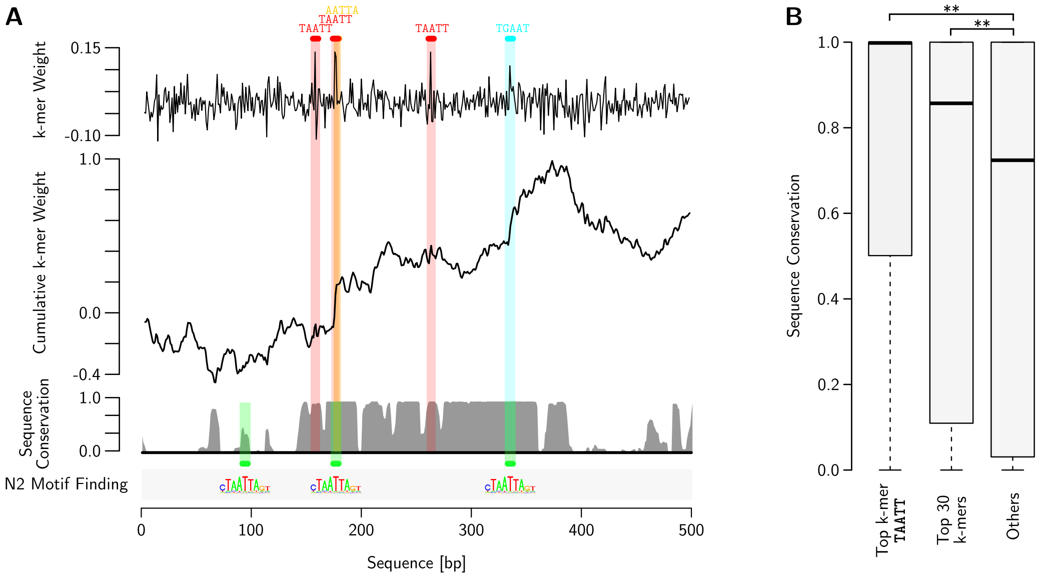


Figure 4.4: ALF-M Identifies Conserved Motifs in Developmental Enhancers. **(A)** Shown is an enhancer sequence that drives expression in mouse embryonic forebrain development. (Top) K-mer weights (z-scores) and cumulative k-mer weights for every 5-mer at every position. Below is shown the sequence conservation (PhyloP, Pollard *et al.* (2010)). Words with high k-mer weights fall into the conserved parts of this enhancer sequence. Occurrences of the four k-mers with the highest weights are highlighted. Below are depicted the occurrences of the motif which is predicted by ALF-M. **(B)** Sequence conservation for all occurrences in the full data set of forebrain enhancers. (Left) Sequence conservation of the k-mer with the highest rank (TAATT). (Middle) Sequence conservation of the top 30 k-mers with the highest ranks. (Right) Sequence conservation of all other k-mers. The k-mers with the highest k-mer weights show significantly higher sequence conservation (Wilcoxon test: $p < 10^{-16}$).

motif of the k-mer TAATT, which had the highest z-score ($p < 10^{-16}$). For many forebrain enhancers, both the refined motif and the k-mer itself fall into highly conserved regions (Figure 4.4A). Similarly, the cumulative sum of k-mer weights increases in conserved regions, while it decreases in non-conserved parts (Figure 4.4A). To estimate the significance of this observation, the average sequence conservation was calculated for all occurrences of the k-mer with the highest weight (TAATT), the conservation for top 30 k-mers with the highest weights, and the conservation for all other k-mers in all forebrain enhancers (Figure 4.4B). Sequence conservation of the single top k-mer and of the 30 top k-mers is significantly higher compared to the conservation of all other k-mers (Wilcoxon test: $p < 10^{-16}$). This strongly demonstrates that the standardised word counts used by $N2$ and ALF-M indeed identify functional, highly conserved sequence motifs which provide a proper seed for motif finding in regulatory sequences.

4.2.5 Discussion

Word statistics and k-mer counts have been used previously for motif finding algorithms (Thomas-Chollier *et al.*, 2012; Fratkin *et al.*, 2006; Ma *et al.*, 2012). The algorithm used by ALF-M is closest to RSAT's peak-motif software (Thomas-Chollier *et al.*, 2012) which is based on word over-representation statistics (van Helden *et al.*, 1998). However, peak-motif does not account for word-overlaps as ALF-M does (Section 3.3.2). These differences will likely reflect the ranking of significant motifs. Among the significant motifs in the Oct4, Sox2, Nanog and p300 data sets, peak-motif and ALF-M identify similar patterns (Figure 4.3, Thomas-Chollier *et al.* (2012)). Furthermore, both methods have quick running times, reflecting similar approaches for motif identification. In contrast to peaks-motif, ALF-M is able to integrate approximate word count statistics. It would be of interest to investigate in detail the differences when mismatches are included compared to exact word counts.

Currently, ALF-M refines the PWM only once by searching the input sequences for occurrences of the consensus pattern. In this step, the flexibility of DNA binding motifs is captured by allowing mismatches for pattern matching. Utilising a PWM-based search algorithm might further improve the accuracy of the final motif returned by ALF-M. Furthermore, multiple rounds of refinement and searching could be employed to return more sensitive results.

In summary, these results highlight the versatility of the $N2$ word statistics for the identification of transcription factor binding motifs. The analysis of the k-mer weights showed that the most significantly over-represented k-mers identified by ALF-M are highly conserved, confirming the value of the $N2$-based ranking. The results on simulations and on enhancer data showed that ALF-M recognises both known and novel motifs, making it a powerful and quick tool for motif identification in large-scale data sets. ALF-M can be used within R, making it particularly interesting as part of R ChIP-Seq analysis pipelines.

4.3 Classification: The N2 Kernel Function

The genome-wide identification of regulatory sequences revealed that hundred thousands of enhancers are most likely active in every cell type (Heintzman *et al.*, 2009). For some cell types, like embryonic stem cells, the repertoire of regulatory sequences is already mapped to a large degree (see Chapter 2). However, the data sets of developmental enhancers only cover a few thousands identified at a specific time point and tissue during development (Blow *et al.*, 2010; Visel *et al.*, 2009). Clearly, many more enhancers that drive partially similar expression patterns exist, but their locations are unknown. The identification of genomic enhancers and the estimation of their activity are therefore major goals in molecular genetics.

Sequence conservation has been used to identify novel enhancers (Pennacchio *et al.*, 2006). Highly conserved non-coding sequences frequently act as enhancers during development. However, sequence conservation alone is unable to predict the precise expression pattern driven by the regulatory elements, this has to be tested using experimental essays (Visel *et al.*, 2007). Additionally, many enhancers show only very weak sequence conservation. Even though they are crucial for embryogenesis (Blow *et al.*, 2010), these enhancers cannot be identified using comparative genomics.

Both the identification of enhancers and the prediction of their activity are essentially classification problems. Classification describes the procedure that assigns a new observation into a known class. The identification of novel enhancers can be achieved by classification of non-coding sequences into enhancer or random genomic background, using the experimental data and randomly sampled background as known classes. Similarly, an enhancer candidate sequence can be classified into one of the known groups of tissue-specific enhancers to predict its regulatory activity.

Binary classification problems can be solved using support vector machines (SVMs) (Cortes and Vapnik, 1995). Given a set of observations from two different classes (training data), the SVM predicts for new data points (test data) to which of the two classes it belongs. SVMs have been applied successfully to DNA sequences where they are able to predict locations of functional elements such as promoters, enhancers or splice sites (Sonnenburg *et al.*, 2006, 2007; Rätsch *et al.*, 2006; Ben-Hur *et al.*, 2008; Lee *et al.*, 2011).

Classification with SVMs is achieved by mapping the input data into a high-dimensional feature space, where a hyperplane is constructed that optimally separates the two classes. DNA sequences can be transformed into word count representations which can then be classified with SVMs (Leslie *et al.*, 2002, 2004). The feature space representation needs not to be calculated explicitly if a kernel function is known that defines an inner product on the feature space-transformed input data. The $N2$ method introduced in Chapter 3 fulfils the properties of a kernel function, as it is defined as the inner product of feature-space transformed sequences, where the word neighbourhood counts represent the feature space. Therefore, $N2$ can directly be plugged into an SVM for classification of DNA regulatory sequences.

4.3.1 The N2 Kernel Function

Let X be the (input) space, a function $K : X \times X \to \mathbb{R}$ is called Kernel, if there exists a feature space F and a function $\phi : X \to F$ such that for all $x, y \in X$, $K(x, y)$ equals the inner product $< \cdot, \cdot >$ of $\phi(x)$ and $\phi(y)$:

$$K(x, y) =< \phi(x), \phi(y) > \ \forall\, x, y \in X \ .$$

Since $N2$ is defined as the inner product of the normalised standardised word neighbourhood counts $\hat{N}$ (feature space), $N2$ provides a kernel function on the set of sequences S:

$$N2(S_1, S_2) =< \hat{N}^{S_1}, \hat{N}^{S_2} > \ \forall\, S_1, S_2 \in S \ .$$

The matrix of all pairwise $N2$ scores forms the kernel matrix which will be used for classification. The same instances of $N2$ will be used as defined in Section 3.3.3. Among the other alignment-free sequence comparison methods (Chapter 3), only $D2$ is a kernel function with precisely defined ϕ (Spectrum kernel, Leslie *et al.* (2002)). Nevertheless, even though both $D2^*$ and $D2z$ scores do not form an inner product space, they provide similarity matrices which I will use for SVM-based classification to show the comparison.

Classification with the $N2$ Kernel

To classify regulatory sequences with the $N2$ kernel function, an SVM model is built using training data. The optimal hyper-plane can be constructed by solving a constrained quadratic programming problem with cost parameter C that controls the penalty for misclassified training data points (Karatzoglou and Meyer, 2006). Here, I used the R package kernlab (Karatzoglou *et al.*, 2004) to train the SVM ($C - svc$, $C = 1$). The raw decision values of the support vector model are then calculated on the test data to predict their class.

The accuracy of $N2$ for classification of regulatory sequences is estimated using 10-fold cross-validation. The data is randomly partitioned into 10 equally sized groups. Every single group is used exactly once as test data, while the remaining 9 groups are used as training data. The performance is evaluated using the average ROC curve calculated on the decision values from the test set. The parameters were fixed to k-mer size $k = 5$ and Markov model order 1 (see Appendix Figure VI.14 for $k = 6$).

The developmental enhancers were obtained using p300 ChIP-Seq data from Visel *et al.* (2009) and Blow *et al.* (2010). Reads were mapped using Bowtie (Langmead

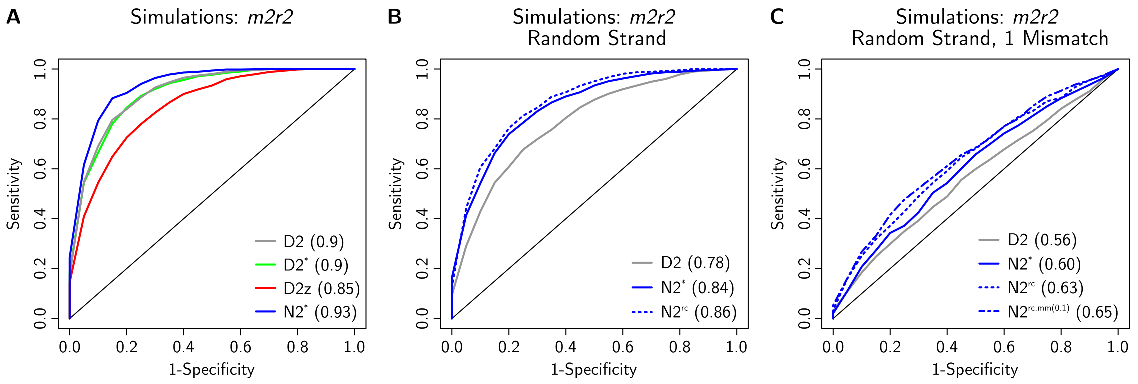


Figure 4.5: Classification with the $N2$ kernel function, simulated sequences. Shown are the average ROC curves after 10-fold cross-validation, numbers indicate the AUC. **(A)** Comparison of the alignment-free sequence comparison methods for classification ($m2r2l5$). **(B)** Words are implanted on a random strand of the sequences to simulate the orientation independence of transcription factor binding sites. $N2^{rc}$ shows the best performance. **(C)** Words were implanted on a random strand with one mismatch to model the flexibility of transcription factor binding sites. Here, $N2^{rc,mm}$ performs best.

et al., 2009), peaks were called using MACS (Zhang *et al.*, 2008). Enhancer sequences were obtained by selecting 1000 bp around the centre of the peak. All data sets were analysed using the top 1000 peaks (sorted by p-value).

4.3.2 Classification: Simulations

To test the accuracy of $N2$ for classification, 1000 regulatory sequences and 1000 random genomic sequences of length 1000 bp were simulated. Random genomic sequences were simulated such that they have the same dinucleotide content as the mouse genome (negative set). Regulatory sequences were simulated similarly and $m = 2$ motifs of length $l = 5$ were each implanted $r = 2$ times into every sequence (positive set, $m2r2l5$, see Appendix Figure VI.12 for $m1r1l7$, $m1r4l6$, $m2r4l5$).

$N2^*$ accurately classifies positive and negative sequences (AUC: 0.93, Figure 4.5A). Furthermore, $N2^*$ outperforms the simple string kernel $D2$ (AUC: 0.9). For these simulations, the inner product based alignment-free sequence comparison methods $D2^*$ (AUC: 0.9) and $D2z$ (AUC: 0.85) provide reasonable results as well, even though they do not fulfil the properties of a kernel function.

The novelty of $N2$ for classification is its ability to include standardised word neighbourhood counts. To test the usability of the neighbourhood concept, two different scenarios were simulated, one that accounts for orientation independent transcription

Classification of tissue-specific enhancer sequences

Tissue:	ESC-F	ESC-M	ESC-L	ESC-H	F-M	F-L	F-H	M-L	M-H	L-H
$D2$	0.84	**0.86**	0.79	0.76	0.63	0.74	0.82	0.76	0.81	0.78
$D2z^{\dagger}$	0.47	0.35	0.55	0.62	0.53	0.58	0.42	0.50	0.32	0.51
$D2^{*\dagger}$	0.80	0.82	0.71	0.72	0.62	0.73	0.81	0.70	0.79	0.73
$N2^{*}$	0.85	0.81	0.83	0.77	0.67	0.79	0.84	0.75	0.78	0.80
$N2^{rc}$	0.87	0.84	**0.85**	**0.81**	0.69	0.81	0.87	**0.79**	0.82	0.82
$N2^{mm(0.1),rc}$	**0.89**	**0.86**	**0.85**	**0.81**	**0.72**	**0.83**	**0.88**	**0.79**	**0.84**	**0.83**

† Not a kernel function.

Table 4.1: Classification of tissue-specific developmental enhancers. Values indicate the AUC after 10-fold cross validation. ESC: embryonic stem cells; F: forebrain; M: midbrain; L: limb; H: heart.

factor binding sites, and one that additionally includes mismatches.

The orientation independence of transcription factor binding sites was simulated by implanting words on the forward or backward strand with equal probability ($m2r2l5$ random strand, see Appendix Figure VI.12 for $m1r1l7$, $m1r4l6$, $m2r4l5$). In this simulation, the $N2^{rc}$ (AUC: 0.86) variant performs better than the other inner product based alignment-free comparison methods (Figure 4.5A). This shows that the extension of the neighbourhood to score words on both strands (rc) improves the classification.

In order to simulate the variability of transcription factor binding sites, words were implanted on the forward or backward strand with one mismatch to the original word ($m2r2$, random strand, 1 mismatch, see Appendix Figure VI.12 for $m1r1l7$, $m1r4l6$, $m2r4l5$). On these simulations, the mismatch variant of $N2$, $N2^{rc,mm(0.1)}$ outperforms the other methods (Figure 4.5B). Together, these simulations demonstrate the value of the extended word neighbourhood counts used by the $N2$ kernel function for classification of sequences.

4.3.3 Classification of Developmental Enhancers with N2

Classification of regulatory sequences can be used to identify novel enhancers. For example, the sequences upstream of coding genes could be classified according to a known group of enhancers or as random genomic background. To estimate the power of $N2$ for this task, the developmental enhancers used in Section 4.4 were used as the positive class (Blow *et al.*, 2010; Visel *et al.*, 2009; Creyghton *et al.*, 2010). To obtain data for the negative class, random positions in the mouse genome were sampled that do not overlap any known enhancer element. Only sequences containing less then 20% repetitive elements were retained for the analysis.

Classification of sequences as enhancer or background can be achieved with an AUC between 0.81 (limb) and 0.88 (forebrain) when the $N2^{rc,mm(0.1)}$ kernel with word size

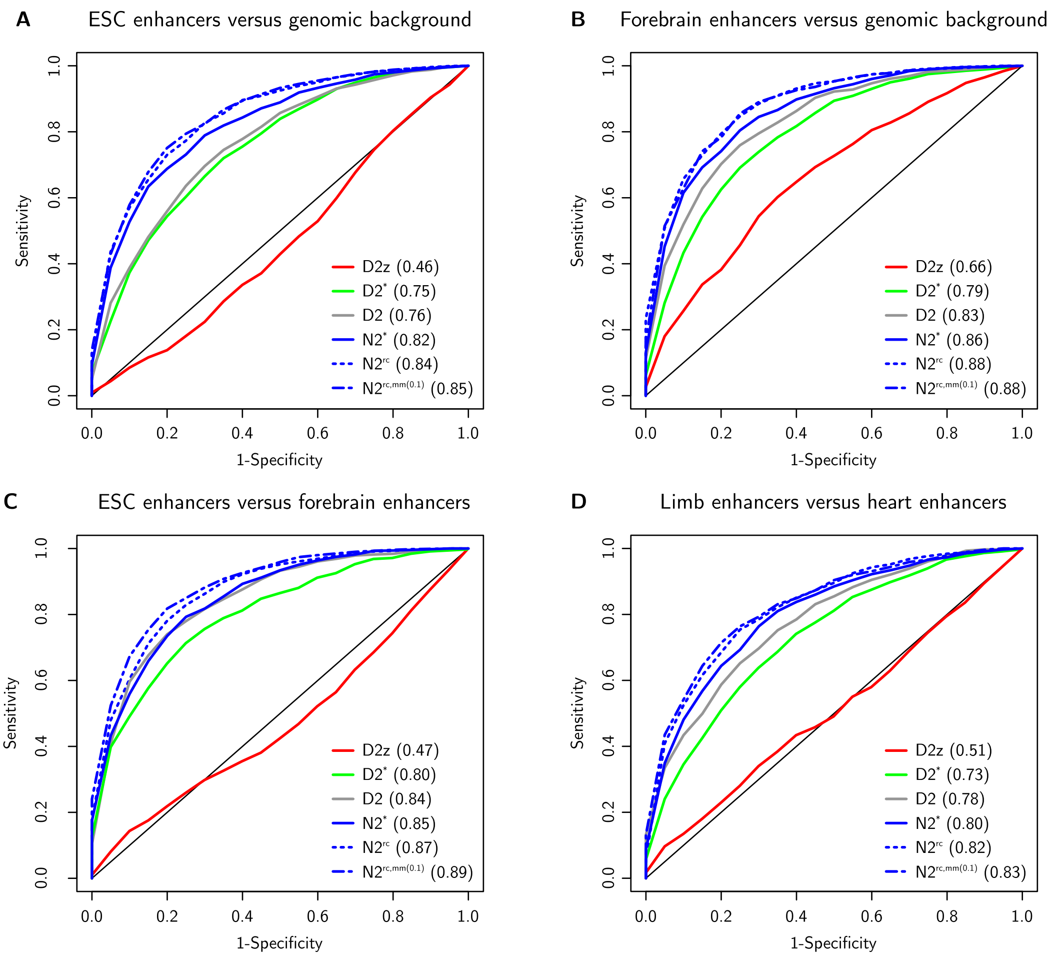


Figure 4.6: Classification of embryonic enhancers using the $N2$ kernel function, shown are the average ROC curves after 10-fold cross-validation, numbers indicate AUC. **(A)** Classification of embryonic stem cell (ESC) enhancers (positive) and random genomic background (negative set). **(B)** Classification of embryonic forebrain enhancers (positive) and random genomic background (negative set). **(C)** Classification of embryonic stem cell enhancers (positive set) and embryonic forebrain enhancers (negative set). **(D)** Classification of embryonic limb enhancers (positive set) and embryonic heart enhancers (negative set).

$k = 5$ is used (Figure 4.6A-B, Appendix Figure VI.13, Appendix Figure VI.14 for $k = 6$). For all tissues, $N2^*$ outperformed the $D2$ string kernel and both $D2^*$ and $D2z$. The extension of $N2$ to count words on both strands $N2^{rc}$ and to include mismatches $N2^{rc,mm}$ improved the performance across all data sets.

The second application for classification of regulatory sequences is the prediction of tissue-specific enhancer activity. For example, a novel identified enhancer sequence can be classified according to the known groups of developmental enhancers. I verified the

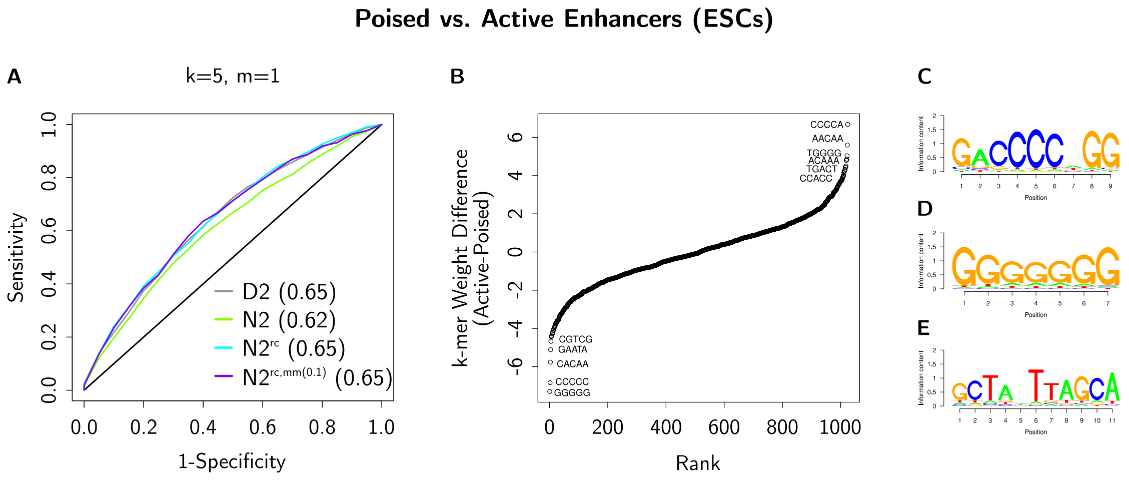


Figure 4.7: Prediction of the epigenetic state of enhancers. **(A)** Active and poised enhancers can be classified using the $N2$ kernel, suggesting that the cell type-specific epigenetic state can be partly predicted from sequence alone.**(B)** Ranking of the k-mer weight differences for active and poised enhancers. **(C)** The primary motif associated with the k-mers enriched in active enhancers over poised enhancers. **(D)** The primary motif associated with the k-mers enriched in the poised class of enhancers compared to active enhancers. **(E)** The primary motif in poised enhancers in human embryonic stem cells is similar to the motif identified in developmental enhancers.

power of $N2$ to correctly predict tissue-specific activity using enhancers from two different tissues, one as the positive and the other as the negative class (Figure 4.6C-D, Table 4.1). Across all combinations of tissues, the $N2^{rc,mm(0.1)}$ kernel most accurately predicted the tissue-specific activity of enhancers. Depending on the pair of tissues used as positive and negative sequences, the AUC varies between 0.72 (forebrain versus midbrain) and 0.89 (embryonic stem cells versus forebrain). The comparison of the different alignment-free methods shows that $N2^*$ generally outperforms $D2$. $D2^*$ performs worse than $D2$ and $D2z$ frequently yields random predictions (Table 4.1).

4.3.4 Prediction of the Epigenetic State of Enhancers

Not all enhancers identified in a specific tissue drive similar expression patterns. While most enhancers are active, some enhancers are poised for activation (Rada-Iglesias *et al.*, 2011). Active and poised enhancers have first been described in human embryonic stem cells (Rada-Iglesias *et al.*, 2011). Active enhancer are marked by the histone modifications H3K4me1 and H3K27ac which are associated with active transcription. In contrast, poised enhancers are marked by the histone modification H3K27me3 which mainly occurs near repressed genes. Both classes are bound by the enhancer binding protein p300, therefore both active and poised enhancers fall into the same class of tissue-specific enhancers, even though they have highly different functionality. One of the key questions is whether the active and poised state of enhancers is reflected in differences of the DNA sequence. In other words, can tissue-specific enhancers be sub-classified into the active and poised epigenetic state based on sequence information alone?

Data sets of active and poised enhancers have been published for human embryonic stem cells (Rada-Iglesias *et al.*, 2011). Here, I used these poised enhancers as the positive set and the active enhancers as the negative set. The classification of enhancers by epigenetic state was then tested using 10-fold cross-validation (Figure 4.7A). Indeed, active and poised enhancers can be classified using the $N2$ string kernel (AUC:0.65). This strongly supports that the epigenetic state of enhancers can be partially predicted from the DNA sequence.

These results suggest that differences in DNA sequence cause epigenetic differences of poised and active enhancers. To analyse sequence-specific differences, I calculated the difference in k-mer weights from active and poised enhancers (Figure 4.7B). These k-mer weight differences were ranked and used as input for ALF-M to identify motifs specifically associated with poised and active enhancers (Figure 4.7C-D). The identified motifs might hint at transcription factors that interact with the histone modifying enzymes to establish the epigenetic state. Among the motifs which are specifically enriched in poised enhancers is the same motif which was identified in developmental enhancers in mouse (Figure 4.7C-D). The close connection of embryonic stem cells and developmental enhancers was highlighted in Chapter 2, where it was shown that enhancers are frequently bound in multiple cell types (*gene regulatory hotspots*). This analysis suggests, that poised enhancers in embryonic stem cells similarly act as such gene regulatory hotspots during early mammalian development.

4.3.5 Discussion

The k-spectrum kernel function (Leslie *et al.*, 2002), has been extended to integrate mismatches (Leslie *et al.*, 2004). These kernels have been applied to classification of the developmental enhancers from midbrain, forebrain and limb used here (Visel *et al.*, 2009; Lee *et al.*, 2011). The authors reported that an extended k-spectrum kernel based on normalised, non-redundant, reverse complement word counts yielded the best results (Lee *et al.*, 2011).

These results are not directly comparable with results in Section 4.3.3 due to differences in data sets and classification methods. For example, Lee *et al.* (2011) use sequences of different length whereas the sequences used here are all of equal length. Furthermore, the random genomic sequences are not the same, which will influence the ROC curve. Additionally, Lee *et al.* (2011) use a different platform for classification

(Sonnenburg *et al.*, 2010) which will yield different results. The AUC values reported in Section 4.3.3 are lower compared to Lee *et al.* (2011) for the k-spectrum kernel $D2$ when enhancers are classified against random genomic sequences ($D2$, $k = 5$, AUC: 0.74 (limb), 0.81 (midbrain), 0.83 (forebrain); Lee *et al.* (2011): Supplementary Figure S1A, $k = 5$, AUC: 0.90 (limb), 0.90 (midbrain) and 0.92 (forebrain)) whereas forebrain enhancers classified against midbrain enhancers give higher AUC values ($D2$, $k = 5$: 0.68; Lee *et al.* (2011), only $k = 6$: 0.56). Another apparent difference is that including mismatches improved the results with $N2$, whereas the usage of the mismatch-kernel gave slightly poorer performance in Lee *et al.* (2011). However, for a balanced comparison, the usage of the same classification pipeline and data sets is required.

$N2$ was designed for pairwise comparison of enhancers, therefore the statistical background model is calculated on the individual sequences. For the task of classification, a much larger data set could be used to estimate the background model. This would drastically reduce the pre-processing time as all neighbourhood word count covariances could be precomputed at once for fixed length sequences (Section 3.3.3). Therefore, estimation of the background model on the full set of sequences could yield more robust estimations of the Markov model, and would enable the usage of much larger choices of k through the decrease in running time.

Results from simulated sequences and enhancers show that the $N2$ string kernel can be used to accurately classify regulatory sequences. Similarly to pairwise comparisons, the reverse complement and the neighbourhood instances yielded additional improvement. Interestingly, $N2$ is able to predict a poised state of enhancers in embryonic stem cells. Therefore, this is the first report of sequence-based classification of enhancers into epigenetically distinct groups, suggesting that epigenetic differences are partially caused by sequence-specific properties, such as nucleotide composition, or binding site content. These results show that $N2$ provides a novel string kernel that accurately predicts the regulatory potential in DNA sequences.

4.4 Clustering: N2 Similarity Measure

Mammalian enhancer sequences show only limited sequence similarity (Section 4.3.3, Chapter 3.4.5). Even regulatory sequences which drive expression in the same tissue are highly divergent. This might be expected as enhancers frequently drive expression in multiple cell types (Section 2.2.4) and have sometimes antagonistic functionality (Rada-Iglesias *et al.* (2011); Section 4.3.4). Therefore, data sets of tissue-specific enhancers are highly heterogeneous collections of sequences.

Heterogeneity of data sets strongly influences their analysis. Both pairwise sequence comparison and classification of tissue-specific enhancers is limited to the subset of truly similar enhancers, yet this subset is unknown. For example, enhancers active in mouse embryonic heart tissue can only partly be discriminated from random genomic sequences (Chapter 3). Even tough these enhancers were identified in the same tissue, it is very likely that they differ substantially in sequence composition and transcription factor binding site content. In general, some enhancers will be bound by the same set of transcription factors and share the same binding sites, while others will be bound by different transcription factors. Therefore, due to the data heterogeneity, it won't be possible to perfectly discriminate them from enhancers active in other tissues.

A partitioning of enhancers into more homogeneous groups would provide a more precise basis to study tissue-specific regulation of gene expression. The idea of cluster analysis is to partition objects into groups (*clusters*) such that all objects in one cluster are more similar to each other than to the objects in other clusters (Jain *et al.*, 1999). Clustering is a very general method and can be applied to every data set on which a similarity can be defined. Whole-genome alignment-free sequence comparison methods have been used in combination with clustering for phylogenetic analyses (Bolshoy *et al.*, 2010). The alignment-free $N2$ method provides an estimate on the pairwise similarity of regulatory sequences (see Chapter 3). The distribution of all pairwise $N2$ scores therefore gives an estimate on the homogeneity of a cluster of enhancers. In order to maximise the homogeneity of enhancers, they can be clustered such that their pairwise $N2$ scores are maximised within the clusters.

Centroid-based clustering algorithms, such as k-means, minimise the distance from the data points to the cluster centres (MacQueen, 1967; Lloyd, 1982). For k-means, the number of clusters k is fixed, and the algorithms searches for the cluster centres and assigns all data points to the nearest centroid such that the distance is minimised. k-means can be applied using the $N2$ scores directly, however, the number of clusters in the enhancer data is unknown. Hierarchical clustering groups objects based on their distance, such that similar objects will be closer, resulting in a dendrogram-like ordering (see Figure 2.2). In contrast to k-means, the number of clusters is not a priori fixed, but determined by the maximum distance allowed between objects within the clusters.

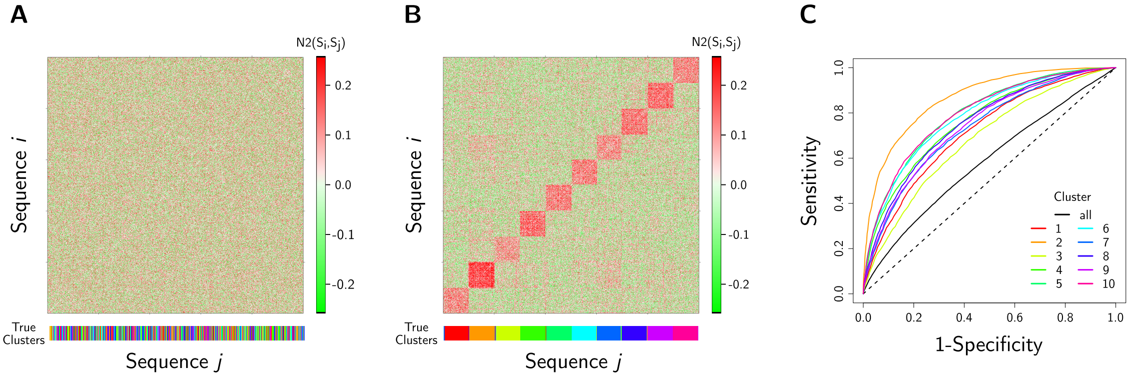


Figure 4.8: Clustering of simulated sequences with $N2$ and k-means. **(A)** Shown are all pairwise scores for 1000 simulated sequences in random order. The true groups are displayed at the bottom. **(B)** The same set of sequences after k-means clustering. The 10 groups of sequences are largely identified. The true clusters are displayed at the bottom. **(C)** ROC curve, for pairwise scores from positive sequences versus negative sequences, before and after clustering. Clustering identifies more homogeneous groups, improving the detection of pairwise similarity.

4.4.1 Simulations

To simulate a heterogeneous cluster of regulatory sequences, two words were inserted each four times into 100 random sequences of length 1000 bp ($m2r4$, see Section 4.2.2). This was repeated 10 times to obtain a data set of 1000 sequences, consisting of 10 true clusters. Only the sequences within the true clusters share the same motifs and are therefore similar, while all other sequence pairs correspond to wrongly labelled true positives. For this data set all pairwise similarities were calculated using $N2$ (Figure 4.8A).

k-means

In the simulation setup, the number of true clusters (10) is known, therefore k-means was used to predict exactly 10 clusters (Figure 4.8B). On average, 97% of the predicted clusters belonged to the same group of sequences ('true cluster'). This shows that the $N2$ similarity in combination with k-means results in highly homogeneous clusters.

To verify the benefit of prior clustering, the power of $N2$ to detect sequence similarity for every predicted cluster was tested. All sequence pairs from sequences with inserted motifs were used as true positives, and pairwise scores from sequences without inserted motifs were used as false positives (see Section 3.4.4). The ROC curve displays the true positive rate (sensitivity) versus the false positive rate (1-specificity), and the area under the ROC curve (AUC) corresponds to the probability that a randomly selected

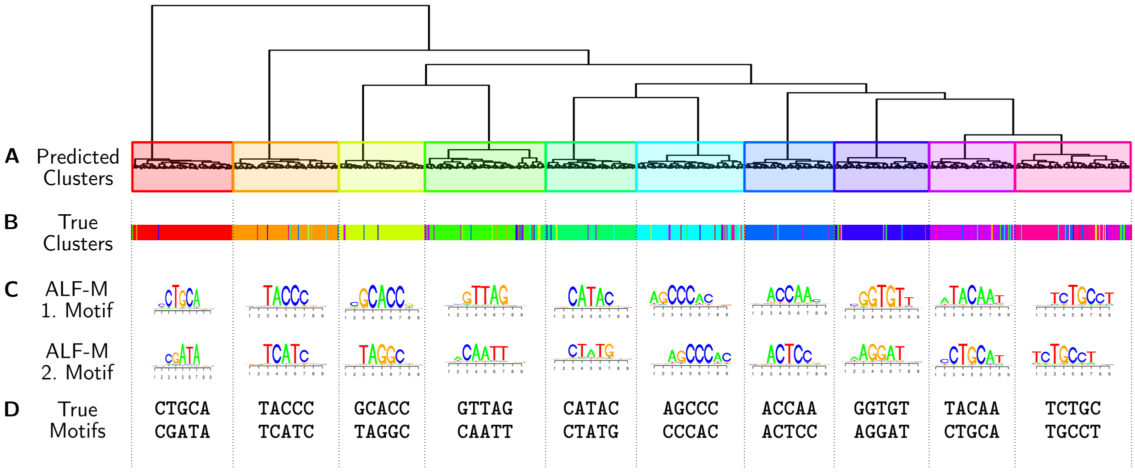


Figure 4.9: Hierarchical clustering of simulated sequences with $N2$. **(A)** Dendrogram showing the hierarchical clustering of 1000 simulated sequences consisting of then 10 groups. The tree was cut at the height that resulted in 10 clusters, the resulting predicted clusters are highlighted. **(B)** The true clusters of sequences according to the implanted motifs. **(C)** The two primary motifs identified using ALF-M. **(D)** Inserted words, every word-pair was inserted four times into 100 sequences. The clustering exactly corresponds to these implanted motifs.

true positive will be scored higher than a randomly selected false positive.

The ROC curve prior to clustering (Figure 4.8C, black line) includes 90% of random sequence pairs as wrongly labelled true positives (inter-cluster sequence comparisons). The AUC is higher for all clusters compared to the full, heterogeneous data set (Figure 4.8C). Since the intra-cluster $N2$ similarity was maximised during cluster generation, this behaviour is expected. However, this simulation demonstrates the value of having homogeneous data sets, as the ROC curve reflects both sequence similarity and homogeneity of the data.

Hierarchical Clustering

For real enhancer sequences, the optimal number of clusters is unknown. Hierarchical clustering provides a means to determine the number of clusters in the data, therefore the same simulated data was used as for k-means. The distances for all sequences were estimated by calculating the euclidean distance on the pairwise $N2$ scores, clustering was performed using Ward's methods (Ward, 1963) (Figure 4.9A). To estimate homogeneity of the predicted clusters, the dendrogram was cut at the height that resulted in 10 clusters. For these predicted clusters, in average 86% of the sequences belonged to

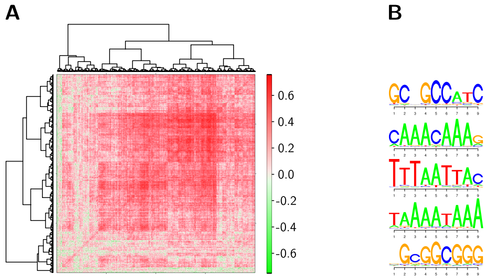


Figure 4.10: Clustering of embryonic heart enhancers.**(A)** Heatmap showing the pairwise $N2$ similarities of embryonic heart enhancers ordered according to the hierarchical clustering. **(B)** The top motifs identified by ALF-M for five clusters.

the same true cluster (Figure 4.9B). While this is lower compared to k-means, it still drastically increases the homogeneity compared to the full data set, suggesting that hierarchical clustering could be used on real enhancer sequences when the number of clusters is unknown.

To further demonstrate the power of homogeneous data sets, the $N2$ based motif finding algorithm ALF-M was applied to the predicted clusters (Figure 4.9C-D). For every cluster, the two highest ranking motifs identified by ALF-M corresponded to the implanted words. This demonstrates the power of word based similarity measures to partition data sets of regulatory sequences based on the transcription factor binding site content.

4.4.2 Clustering of Developmental Enhancers

Enhancer sequences identified in the same tissue have higher pairwise $N2$ scores (Chapter 3). However, the clustering simulations (Figure 4.8) showed that the ability of $N2$ to compare regulatory sequences might be influenced by the heterogeneity of the data. Enhancers active in mouse embryonic heart showed the smallest difference to random sequences (Figure 3.6, Table 3.4), suggesting that this data set consists of a more heterogeneous collection of regulatory sequences than enhancers of other tissues.

In order to study the heterogeneity of heart enhancers, all pairwise $N2^{rc,mm}$ scores were calculated on 2000 heart enhancers (Blow *et al.*, 2010) of length 1000 bp, each having maximal 30% of its sequence covered by repeats. Similarly to the simulated sequences, a hierarchical clustering was computed (Figure 4.10A). First, ALF-M was used to identify motifs in five sub-clusters. While some clusters have more similar

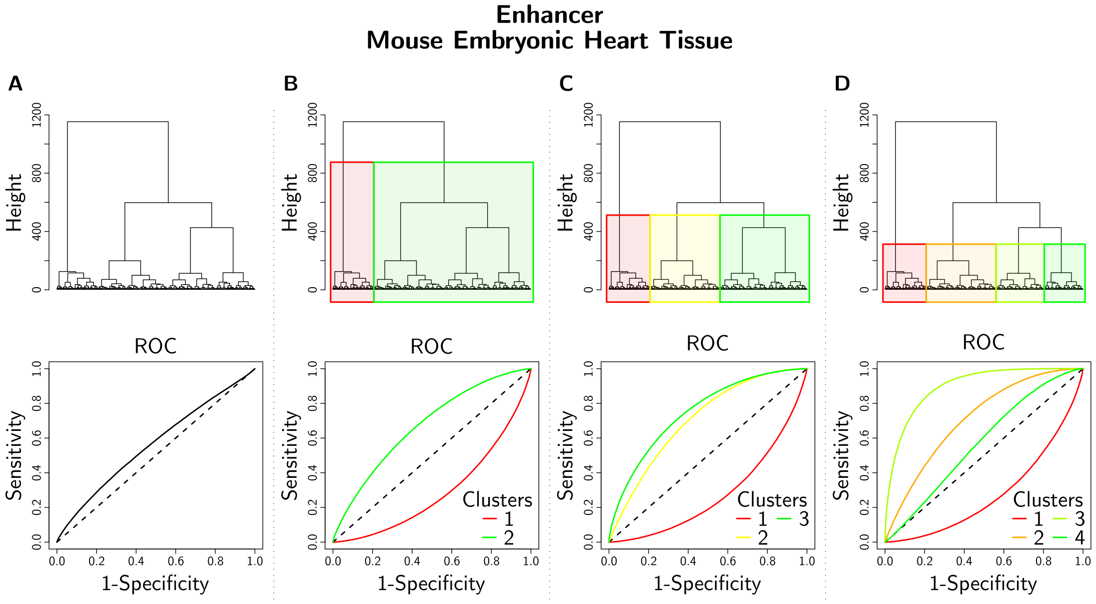


Figure 4.11: Hierarchical clustering identifies homogeneous groups of enhancers. The hierarchical clustering of mouse embryonic heart enhancers is shown on top. The homogeneity of the clusters was estimated by a ROC curve (bottom). **(A)** Full data set, single cluster. **(B)** Two clusters. **(C)** Three clusters. **(D)** Four clusters.

motifs, others are very different (Figure 4.10B). This shows that clustering using the $N2$ similarity results in clusters having different nucleotide and word compositions.

Since $N2$ was used to obtain the clustering, $N2$ gives an estimate on the cluster homogeneity. Similar to the simulations, the ROC curves before and after clustering were calculated using pairs of sequences randomly selected from the mouse genome to estimate the false positive rate. The number of clusters was stepwise increased as the true number is unknown (Figure 4.11).

The ROC curve calculated on the full data set confirms that pairwise scores from heart enhancers are close to pairwise scores from random sequences (Figure 4.11A). Splitting the data into two clusters already demonstrates the heterogeneity of the heart enhancer data set (Figure 4.11B). A small cluster has pairwise scores lower than random, these sequences are more dissimilar than is expected by chance. In contrast, the large majority of sequences are more similar than expected by chance, as is reflected by the ROC curve. Splitting the data into three (Figure 4.11C) and four (Figure 4.11D) clusters further separates heart enhancers into more homogeneous clusters. Before clustering, the AUC is 0.56. After partitioning the data set into 4 clusters, the AUC for the largest cluster (705 sequences) is 0.70, for the most homogeneous cluster (487 sequences, cluster 3) 0.91, and for the cluster of dissimilar sequences (416 sequences, cluster 1) the AUC

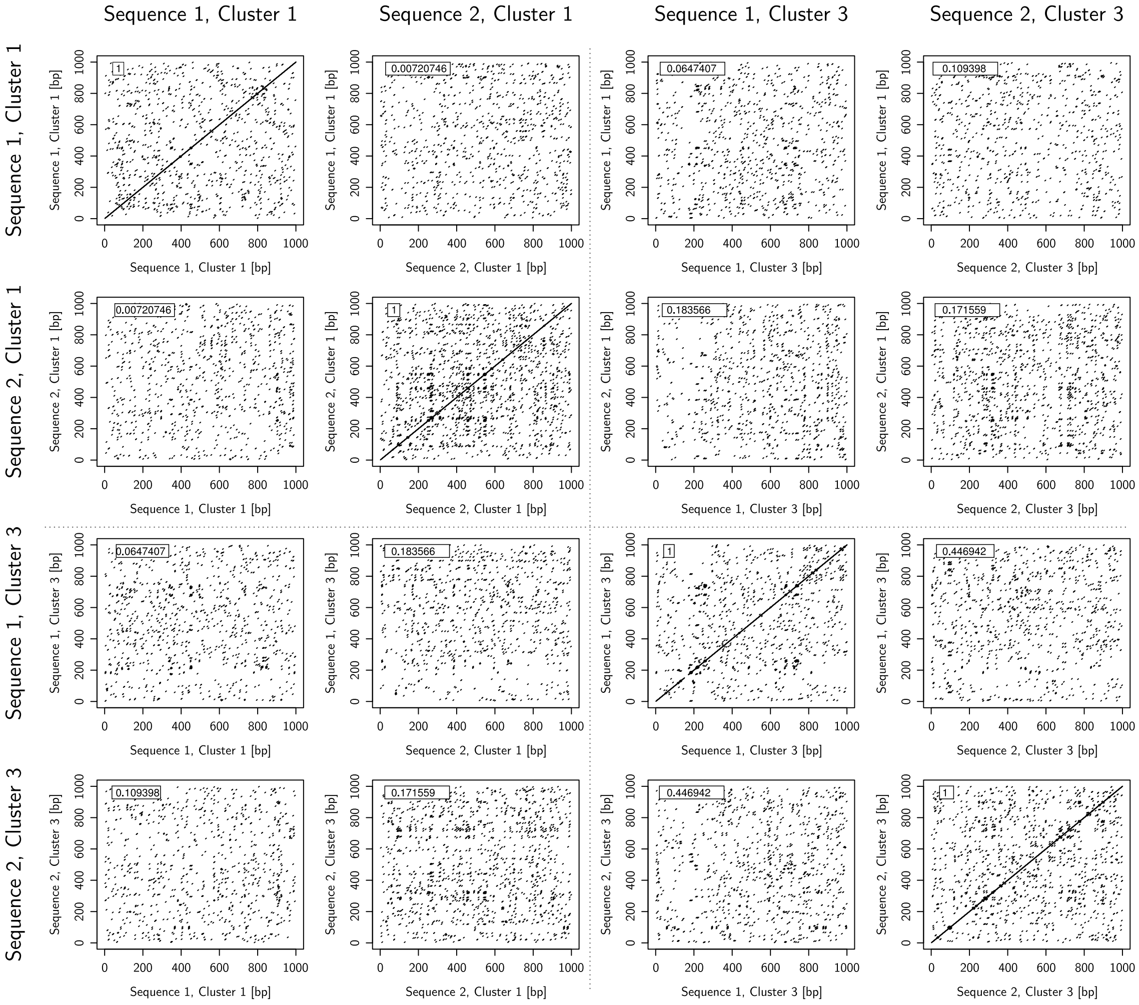


Figure 4.12: Dotplot (word size 5) of sequences from clusters 1 (dissimilar) and cluster 3 (highly similar), $N2^{rc,mm}$ scores are indicated in the top left corner for every pair. Pairwise scores are higher for sequences from cluster 3 compared to sequences from cluster 1. However, the sequence similarity can not be estimated from the dotplots or alignments.

is 0.30. This confirms that the poor performance of the heart enhancer data set in Chapter 3 can be partly explained by the heterogeneity of the data and in particular by a sub-population of sequences (cluster 1) that do not seem to share many binding sites as quantified by the N2 similarity.

Several sequence properties could have an impact on the clustering with $N2$. For example, repetitive sequences (high word count, sequence noise, see Section 3.4.3) or large regions of sequence similarity (alignable sequences) could induce high $N2$ scores. One way to visualise pairwise sequence comparisons is a *dotplot* (see Section 3.2.2).

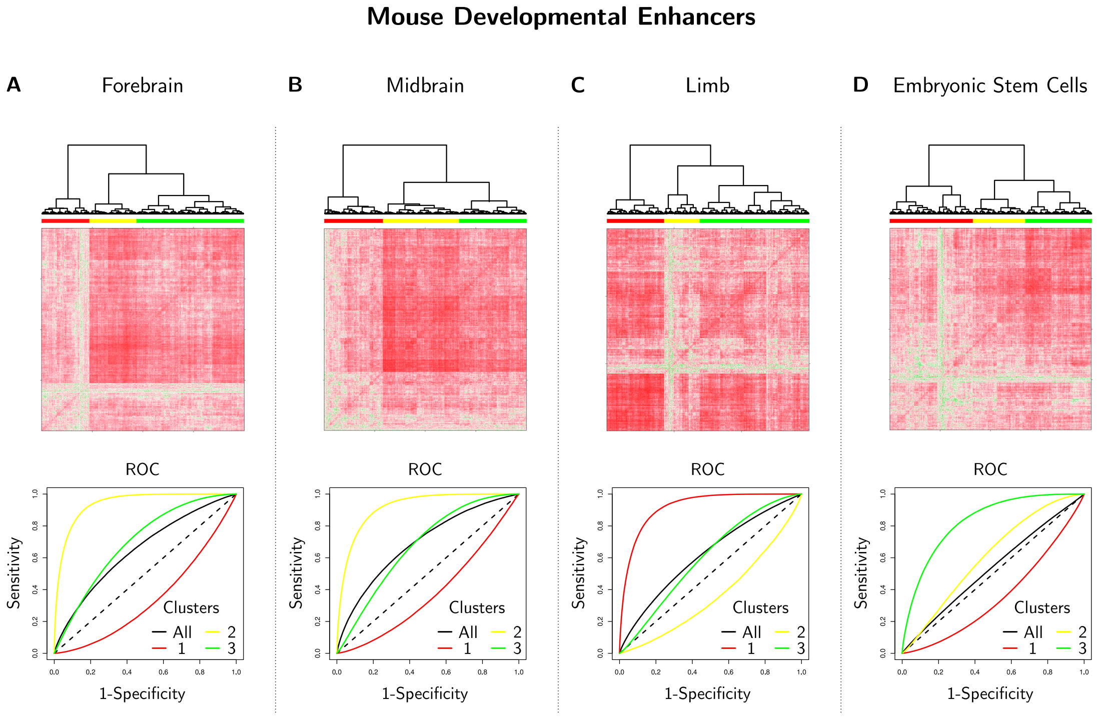


Figure 4.13: Clustering of Mouse Developmental Enhancers. Enhancer data sets can be separated into more homogeneous sub-clusters. **(A)** Forebrain. **(B)** Midbrain. **(C)** Limb. **(D)** Embryonic stem cells.

Figure 4.12 shows the dotplots of all pairwise comparisons for two sequences from cluster 1 (low $N2$ scores) and two sequences from cluster 3 (high $N2$ scores). All sequence pairs, between and across the two clusters, show similar amounts of matching words, therefore neither repetitive sequences nor high sequence similarity (alignability) seem to influence the clustering. Furthermore, the dotplots demonstrate the power of alignment-free sequence comparison over traditional alignments, as none of the heart enhancers show stretches of similar words that could have been captured with local or global alignments.

To test whether the other enhancer data sets show similar characteristics, the same analysis was performed on forebrain (Visel *et al.*, 2009), midbrain (Visel *et al.*, 2009; Blow *et al.*, 2010), limb (Visel *et al.*, 2009), and embryonic stem cell enhancers (Creyghton *et al.*, 2010) (Figure 4.13). Across all data sets, clustering results in a partitioning of the data into one cluster of dissimilar sequences, one cluster of highly similar sequences and one cluster of sequences with intermediate similarity, likely reflecting intra-cluster heterogeneity.

4.4.3 Discussion

Clustering of regulatory sequences requires a sensitive measure of pairwise similarity. In contrast to large-scale applications such as motif finding and classification, pairwise sequence similarity is restricted to two sequences and therefore a very difficult task. $N2$ provides a sensitive measure of similarity for regulatory sequences and thereby facilitates a large-scale cluster analysis.

The analysis of enhancer sequence data emphasises the influence of data homogeneity on the results. Global measures such as ROC curves and AUC values are strongly limited by heterogeneous data and provide only partial information on the performance of the method. For example, pairwise comparison of regulatory sequences with $N2$ can be achieved with an AUC of 0.91 for mouse embryonic heart enhancers after clustering. It is very likely that other tasks such as classification and prediction (Section 4.3) or motif finding (Section 4.2) will similarly profit from more homogeneous data and yield more sensitive results.

Since the numbers of true clusters is unknown for enhancer data, hierarchical clustering was used here. Yet, the simulations showed that k-means might result in more homogeneous clusters. Hierarchical clustering could be used to estimate the number of clusters, and k-means could then be used for the final clustering. The two clustering methods used here only provide a small subset and many other methods exist that have specific advantages (Jain *et al.*, 1999). Nevertheless, usage of hierarchical clustering combined with the $N2$ similarity measure already highlighted the heterogeneity of the data.

All enhancer data sets could be partitioned into smaller clusters that were more homogeneous. The structure of the data (Figure 4.13) shows that all data sets consist of 2-3 larger clusters. Yet, even these clusters consist of many hundred sub-clusters. While further partitioning of the data would result in smaller data sets that are of limited use for large-scale applications, the data structure is still of interest. It is very likely that all the small sub-clusters of sometimes only a few sequences represent enhancers with a large number of shared motifs, while the larger clusters represent enhancers with fewer shared words. Such a structure of regulatory sequences would enable the cell to direct gene expression programs on a larger scale (tissue, larger cluster) and on a more refined scale (cell types, developmental stages, smaller clusters). Therefore, the structure of the clustering of long-distance regulatory elements might reflect the structure of tightly connected transcriptional regulatory networks that facilitate the concerted embryonic development in mammals.

5 Summary

Mammalian organisms consist of several hundred different cell types. Although every cell has the same repertoire of genes only a subset will be expressed to enable cell type-specific functions. Regulation of gene expression is organised in a highly connected manner through the binding of transcription factors at specific DNA sequences (Chan *et al.*, 2011; Lee *et al.*, 2002; Davidson, 2006). These *cis*-regulatory elements can be found in close proximity to the transcription start site (promoters) or can be many kilobases distant (enhancers). Most of our knowledge of transcriptional regulation was obtained from studies of promoters, since enhancers are much harder to identify and study (Heintzman and Ren, 2009). However, enhancers are crucial for cellular differentiation and embryonic development (Rada-Iglesias *et al.*, 2011). This thesis deals with the analysis of such long-distance regulatory elements.

Transcription factors typically do not act in isolation but instead act in a combinatorial manner to regulate cell type-specific gene expression. For example, in mouse embryonic stem cells (ESCs), the key factors Oct4, Sox2 and Nanog are frequently found at the same distal regulatory elements (Figure 2.2). Chapter 2 investigates the influence of combinatorial binding of Oct4, Sox2 and Nanog on transcription and evolution of gene regulation. It is shown that in contrast to loci where only one of the transcription factors binds, loci with multiple factors binding are enriched in several properties associated with active enhancers. These properties include binding by the transcriptional co-activator Mediator, enrichment in the histone modification H3K27ac, and increased expression of nearby genes. The target genes of these combinatorially bound enhancers are frequently active during embryonic development. A comparison with enhancers from different developmental stages of the mouse embryo (Blow *et al.*, 2010; Visel *et al.*, 2009) shows that the same elements which are bound by Oct4, Sox2 and Nanog in ESCs frequently drive expression in developmental tissues.

It has been described that only a minority of enhancers (2-7%) shows conserved binding between mouse and human (Schmidt *et al.*, 2010; Kunarso *et al.*, 2010). A comparison of genome-wide binding data from OCT4, SOX2 and NANOG in human ESCs with mouse ESCs shows that combinatorially bound enhancers are more frequently conserved (15%). More than 50% of enhancers which are bound by all three factors and which are active in development show binding conservation, suggesting that these are specifically important for embryonic development.

ChIP experiments typically identify many thousands of binding sites, which raises the question: which of these elements are actually relevant to regulate the expression of associated genes? In Chapter 2 it is shown that integration of combinatorial binding identifies highly conserved developmental enhancers important for pluripotency and embryogenesis, highlighting the importance of combinatorial regulation of gene expression at distal regulatory elements.

The large-scale identification of regulatory elements in recent years is comparable to the large-scale identification of protein coding genes after the initial sequencing of the human genome. For many years, sequence similarity has been used to estimate the functional similarity of protein coding genes. Sequence similarity is usually computed using global (Needleman and Wunsch, 1970) or local alignment tools (Smith and Waterman, 1981) such as BLAST (Altschul *et al.*, 1990). In contrast to coding sequences, these alignment methods fail in the identification of functionally similar regulatory sequences. In Chapter 3 a novel alignment-free sequence comparison method ($N2$) is presented, which can be used to compare regulatory elements.

The basic idea of $N2$ is to estimate the similarity of regulatory sequences by the number and combination of shared words. This approach is motivated by the observation, that regulatory sequences which drive expression in the same cell type have similar combinations of transcription factor binding sites (Chapter 2). Therefore, sequences that share similar combinations of words have a high $N2$ similarity score.

The novelty of $N2$ is the flexible framework of standardised word neighbourhood counts. This framework extends existing alignment-free sequence comparison methods to approximate word matching. Application of $N2$ to developmental enhancers demonstrates that $N2$ is able to identify enhancers which are active in the same tissue using sequence information alone.

Even though $N2$ was designed for pairwise sequence comparison, it provides many possibilities for large-scale data analysis of regulatory sequences. Chapter 4 utilises $N2$ and the underlying word statistics to analyse the sequence-specific properties of mammalian enhancers in embryonic stem cells and early embryonic development.

Since the regulatory function of enhancers is determined by their transcription factor binding site content, the knowledge of these binding sites is of great interest. In Section 4.2, ALF-M is presented, a de novo motif finding algorithm. The initial step is provided using the $N2$ word statistics: a k-mer ranking is created based on the standardised word neighbourhood counts. Application of ALF-M to enhancers in embryonic stem cells identifies the motifs that correspond to the combinatorial regulation of gene expression reported in Chapter 2. Applied to developmental enhancers, ALF-M identifies several candidate transcription factor binding motifs which have significantly increased levels of sequence conservation.

In Section 4.3, $N2$ is used as a kernel function for SVM-based classification of regulatory sequences. The $N2$ instance that includes mismatches and reverse complement word counts ($N2^{rc,mm}$) achieves a maximum AUC of 0.89 for classification of developmental enhancers, confirming that the word content is largely sufficient to determine the tissue-specific activity of regulatory sequences.

Section 4.4 investigates the heterogeneity of large-scale enhancer data sets. The $N2$ similarity measure is used to cluster developmental enhancers into homogeneous groups. This analysis highlights that tissue-specific enhancer data sets most likely consist of enhancers which are only partially similar. Such a clustering might provide the basis for a detailed analysis of transcriptional regulatory networks based on the binding site content of long-distance enhancers.

In summary, this thesis presents new insights into the combinatorial regulation of gene expression in embryonic stem cells and provides a novel method for sensitive pairwise comparison of enhancers and in-depth analysis of large-scale data sets of regulatory elements.

VI Appendix

VI.1 Supplementary Information for Chapter 2

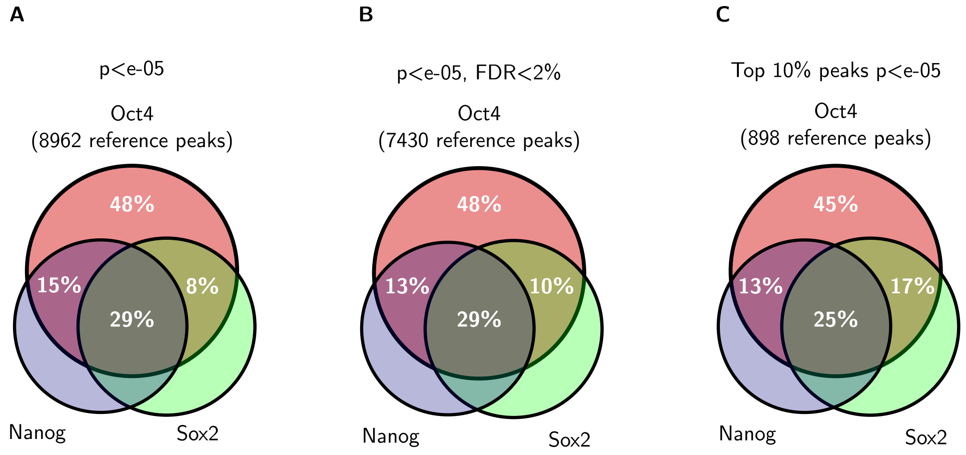


Figure VI.1: Comparison of different cutoffs for peak calling. The diagram shows the percentage of Oct4 bound loci that are bound by Nanog and Sox2. The observed level of co-localisation is comparable across data sets with different cutoffs. **(A)** Full data set, all peaks with p<e-05. **(B)** FDR controlled data set, peaks with p<e-05 and FDR<2%. **(C)** Stringent cutoff, the 10% most significant peaks from all peaks with p<e-05.

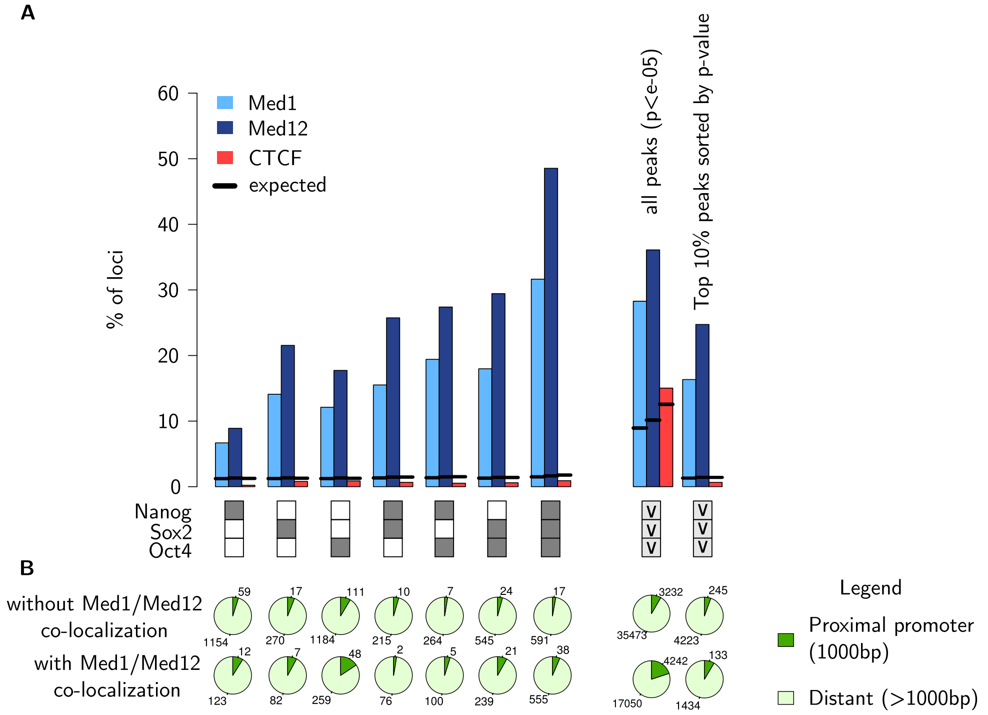


Figure VI.2: Mediator co-localises with Oct4, Sox2 and Nanog at combinatorially bound enhancers. For every data set, only the 10% most significant peaks of all peaks with p<e-05 are considered ('stringent data set'). **(A)** Bars indicate the fraction of loci where Med1, Med12 and CTCF binding can be observed, depending on the combination of Oct4, Sox2 and Nanog, indicated by boxes below. Dark boxes indicate binding, light grey boxes with 'v' indicate binding of at least one factor ('OR' relation). Both Med1 and Med12 preferentially co-localise at loci bound by Oct4, Sox2 and Nanog simultaneously. CTCF serves as a control to estimate unspecific binding. **(B)** The majority of loci bound by Oct4, Sox2 and Nanog are more than 1000 bp away from the nearest transcription start sites for all possible combinations (indicated by boxes above). Mediator co-localisation mainly occurs at distant regulatory sites, showing that the increased overlap of Med1/Med12 at combinatorially bound loci is not caused by promoter specific co-localisation

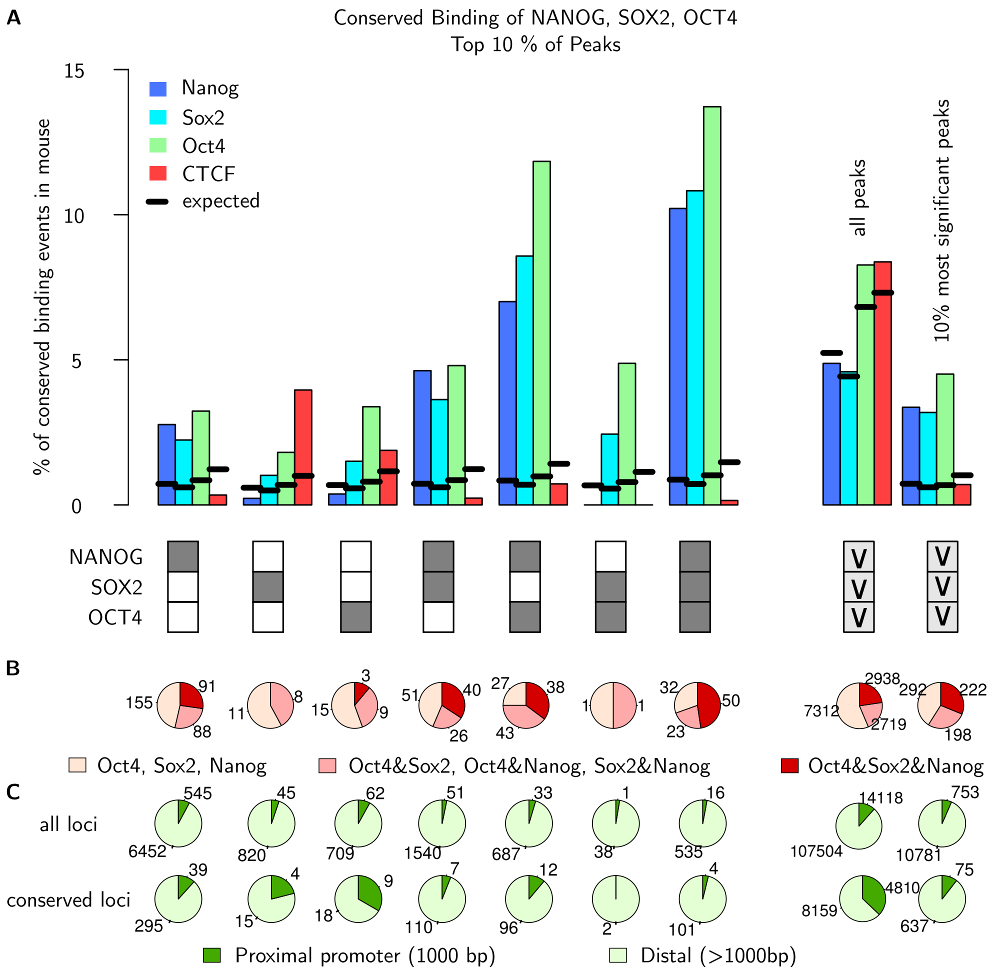


Figure VI.3: The combination of OCT4, SOX2 and NANOG influences conservation of binding events. For every data set, only the 10% most significant peaks of all peaks with p<e-05 are considered. **(A)** Bars indicate the fraction of loci where binding of Nanog, Sox2, Oct4 or CTCF can be observed at the orthologous locus in mouse ES cells for all combinations of OCT4, SOX2 and NANOG in human ES cells as indicated by the boxes below. Dark boxes indicate binding, light grey boxes with 'v' indicate binding of at least one factor ('OR' relation). Combinatorial binding of OCT4, SOX2 and NANOG shows the largest fraction of conserved binding for Oct4, Sox2 and Nanog in mouse. **(B)** The fractions of binding combinations in mES cells at conserved loci (for all combinations of binding in human cells as indicated by the boxes above). Combinatorial binding of Oct4, Sox2 and Nanog in mES cells is much higher at combinatorially bound loci in human, suggesting that combinatorial binding is conserved in evolution. **(C)** The fraction of proximal and distant binding sites for conserved and non-conserved binding events, split up according to the combinations of binding as indicated by the boxes above. The majority of conserved binding events are distant regulatory elements.

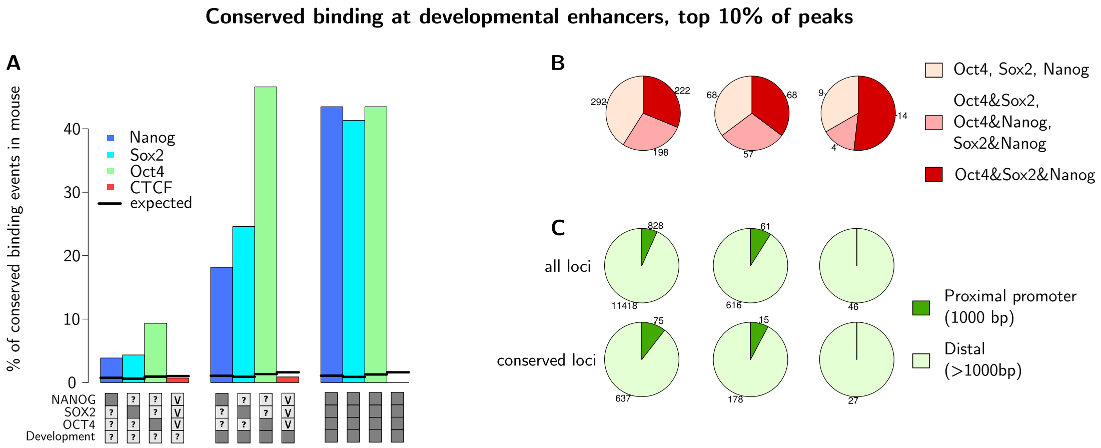


Figure VI.4: Combinatorial binding in ES cells is highly conserved at developmental enhancers. For every data set, only the 10% most significant peaks of all peaks with p<e-05 are considered. **(A)** Bars indicate the fraction of loci where binding of Nanog, Sox2, Oct4 and CTCF can be observed at the orthologous locus in mouse ES cells for all combinations of OCT4, SOX2 and NANOG in human ES cells discriminated by developmental activity as indicated by the boxes below. Dark boxes indicate binding, light grey boxes with 'v' indicate 'OR' relation, '?' indicates no restriction. Combinatorial binding events at developmentally active enhancers show the highest levels of binding conservation between mouse and human ES cells (>40%). **(B)** The fractions of binding combinations in mES cells at conserved loci (for all combinations indicated by the boxes above). The majority of conserved binding events at developmentally active enhancers where OCT4, SOX2 and NANOG bind simultaneously show combinatorial binding of Oct4, Sox2 and Nanog in mouse ES cells. **(C)** The fraction of proximal and distant binding sites for conserved and non-conserved binding events (split up according to the combinations of binding as indicated by the boxes above). The majority of conserved binding events are distant regulatory elements.

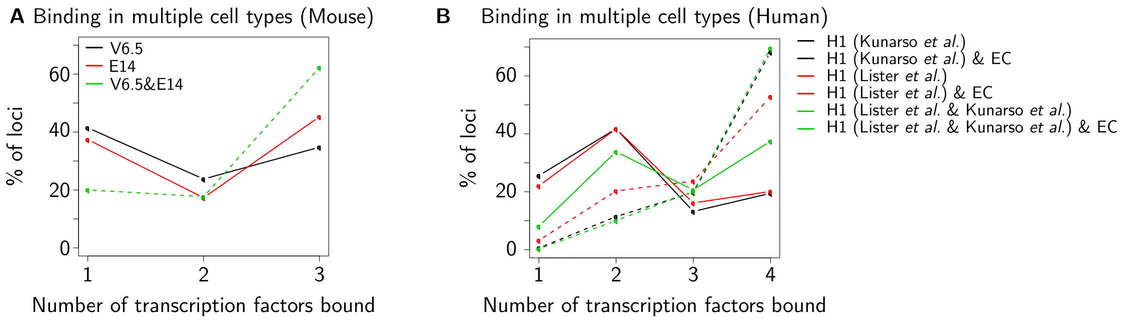


Figure VI.5: Integrating data from different cell lines identifies functional binding events. **(A)** Shown is the fraction of loci where one, two or three (Oct4, Sox2, Nanog) transcription factor binding events can be observed. Binding events detected in both cell lines are more frequently bound by multiple transcription factors (dotted lines). **(B)** Shown is the fraction of loci where one, two, three or four different factors are binding (OCT4, SOX2, NANOG, p300). The fraction of loci bound by all four factors is much higher when data from embryonic stem cells and embryonal carcinoma cells are combined (dotted lines).

VI.2 Supplementary Information for Chapter 3

$$\mathbb{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{E}[X_i]\ . \tag{VI.1}$$

$$\mathbb{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\ . \tag{VI.2}$$

Performance with implanted k-mers, random strand, $k = 5$

	5%-precision		AUC (ROC)		AUC (PR)	
Motif setting:	m1r8	m4r2	m1r8	m4r2	m1r8	m4r2
$D2$	0.88	0.59	0.74	0.54	0.73	0.53
$D2z$	**0.96**	0.67	0.79	0.58	0.79	0.57
$D2^{*}$	0.92	0.69	0.75	0.59	0.74	0.58
$N2^{*}$	0.91	0.68	0.75	0.59	0.74	0.58
$N2^{rc}$	0.95	**0.73**	**0.81**	**0.62**	**0.81**	**0.61**

Table VI.1: Simulations with implanted k-mers. Comparison of the different methods ($k = 5$, $mo = 1$) when the genomic orientation of the motif is unknown. Bold numbers indicate best performance.

Performance with implanted k-mers, mismatch, $k = 5$

	5%-precision		AUC (ROC)		AUC (PR)	
Motif setting:	m1r8	m4r2	m1r8	m4r2	m1r8	m4r2
$D2$	0.60	0.51	0.53	0.47	0.53	0.47
$D2z$	0.62	0.54	0.56	0.51	0.55	0.51
$D2^*$	0.63	**0.55**	0.56	**0.52**	0.55	0.51
$N2^*$	0.62	0.54	0.56	**0.52**	0.55	0.51
$N2^{mm(0.1)}$	0.63	0.54	0.56	**0.52**	0.55	0.51
$N2^{mm(1.0)}$	**0.65**	0.54	**0.57**	**0.52**	**0.57**	**0.52**

Table VI.2: Simulations with implanted k-mers. Comparison of the different methods ($k = 5$, $mo = 1$) when motifs are sampled from all k-mers with one mismatch to the word. Bold numbers indicate best performance.

Influence of repeats, k-mer model, m4r2

	5%-precision		AUC ROC		AUC PR	
Repeat masked:	N	Y	N	Y	N	Y
$D2$	0.52	0.52	0.51	0.50	0.50	0.49
$D2z$	0.67	0.82	0.67	0.71	0.62	0.69
$D2^*$	0.91	**0.97**	0.82	0.81	0.81	0.83
$N2^*$	**0.94**	**0.97**	**0.90**	**0.90**	**0.89**	**0.89**

Table VI.3: Influence of repeats on pairwise sequence comparison. Values are averages over 25 simulations, numbers in bold indicate the best performing method for the given setting. Mammalian genome sequences contain to a large degree repetitive sequences. I studied the influence of repeats on pairwise scores by randomly selecting sequences from the mouse genome and implanting 4 words of length 6 each twice into the sequences. Repeat-masking improves the results. $N2$ is most robust against repeats, whereas $D2z$ is strongly influenced by repeats.

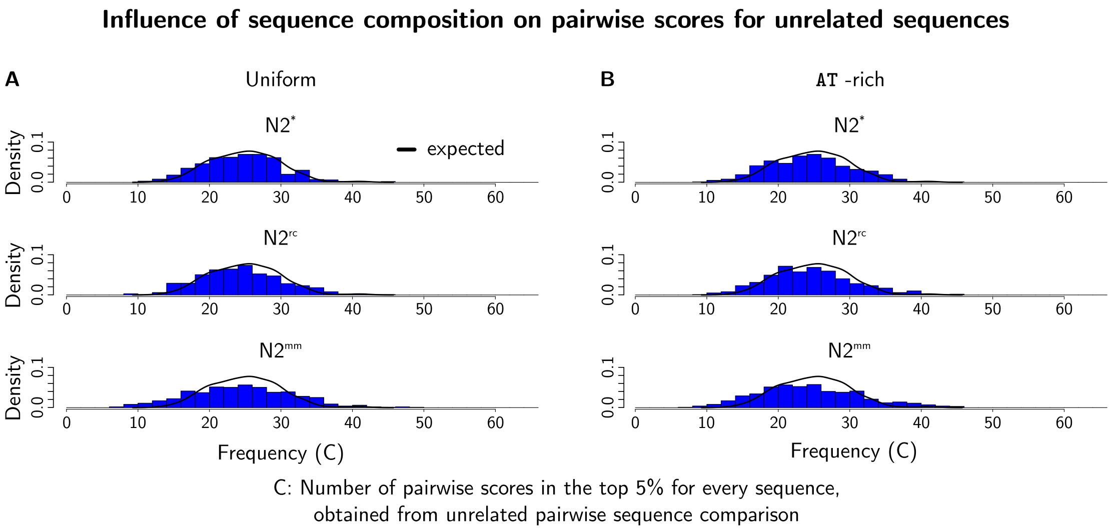


Figure VI.6: Influence of single sequences on pairwise scores on the $N2$ variants. **(A)** Uniform nucleotide distribution. **(B)** `AT`-rich nucleotide distribution. Alignment-free sequence comparison methods can easily be influenced by nucleotide composition and low complexity or repetitive sequences (see Section 3.4.3). The $N2$ variants that use mismatches and the reverse complement as extended word neighbourhood are robust against a change of nucleotide distribution, they perform as expected (black line).

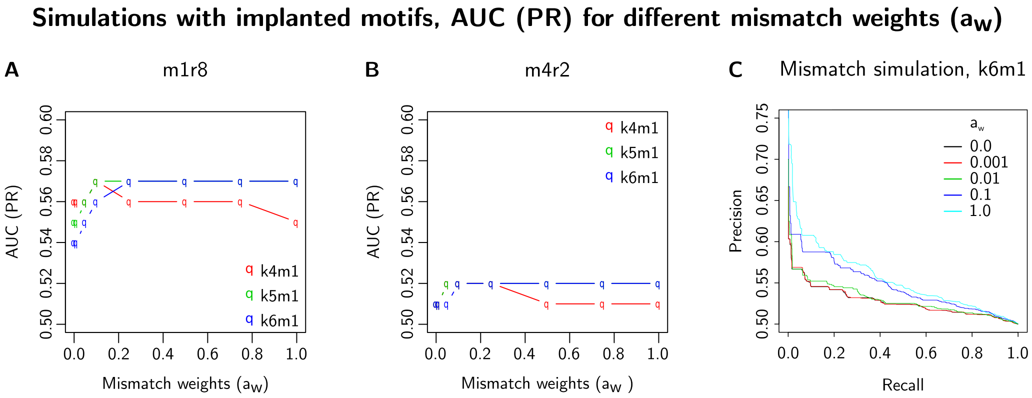


Figure VI.7: Influence of the mismatch weights a_w on the pairwise scores of simulated sequences.**(A)** AUC values for simulated sequences ($m1r8$). **(B)** AUC values for simulated sequences ($m4r2$). **(C)** Precision-Recall plot for simulations with implanted k-mers (mismatch model) for different choices of a_w. The combination of $k = 6$ and $k = 5$ with $a_w = 1.0$ produced the best results. For $k = 4$, smaller mismatch weights gave better results.

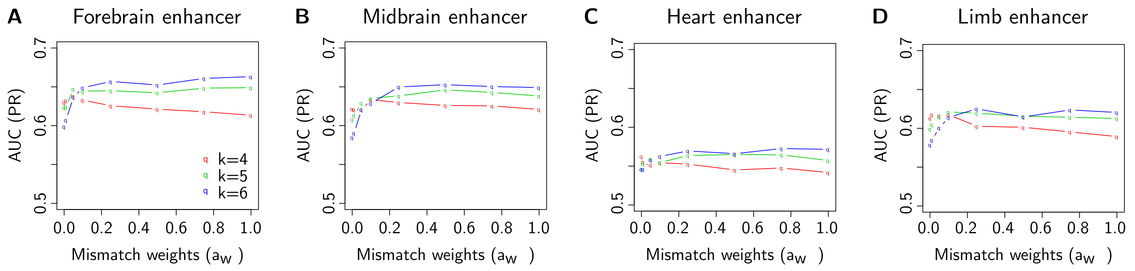


Figure VI.8: Influence of mismatch weights a_w and word length k on the pairwise scores of developmental enhancers. In all data sets, the combination of $k = 6$ and $a_w = 1.0$ produced the best results. For $k = 4$, smaller mismatch weights gave better results, probably because words in the neighbourhood are more likely to occur by chance. **(A)** Forebrain. **(B)** Midbrain. **(C)** Heart. **(D)** Limb.

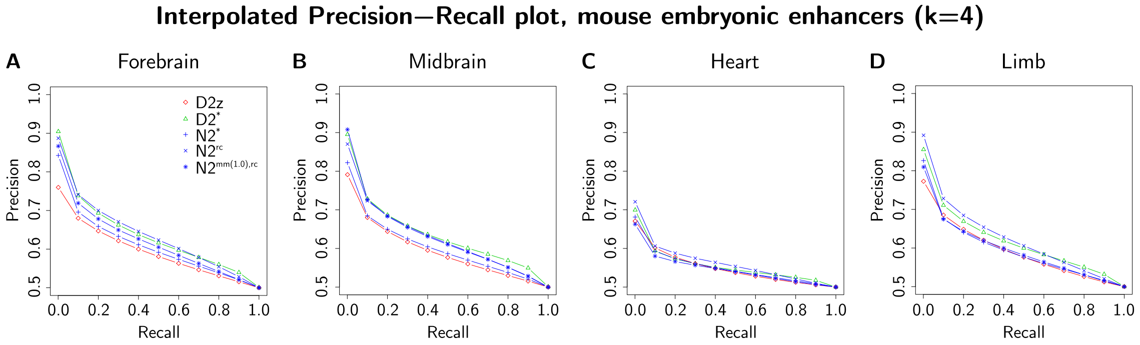


Figure VI.9: Precision-Recall curve for enhancers active during mouse development, $k = 4$. The plots show the precision average over 25 samples each time drawing 500 enhancer sequences ('positive') and 500 unrelated genomic sequences of equal length as the enhancers ('negative'). Results are shown for enhancers active in forebrain **(A)**, midbrain **(B)**, heart **(C)**,limb **(D)**.

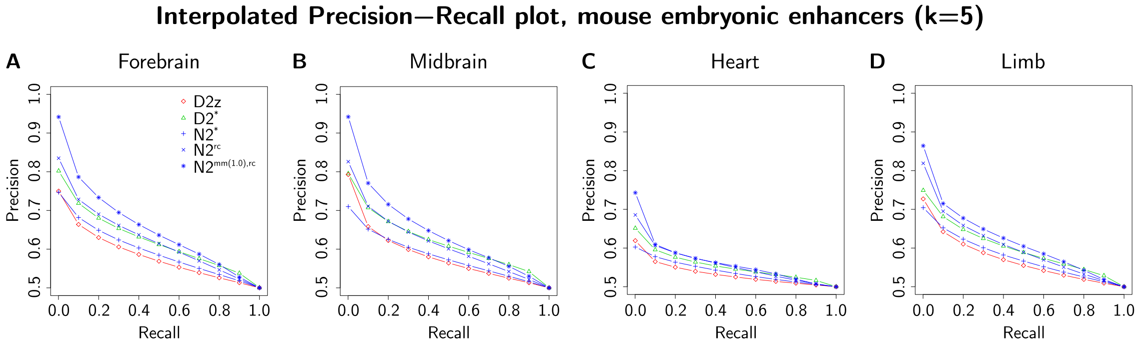


Figure VI.10: Precision-Recall curve for enhancers active during mouse development, $k = 5$. The plots show the precision average over 25 samples each time drawing 500 enhancer sequences ('positive') and 500 unrelated genomic sequences of equal length as the enhancers ('negative'). Results are shown for enhancers active in forebrain **(A)**, midbrain **(B)**, heart **(C)**,limb **(D)**.

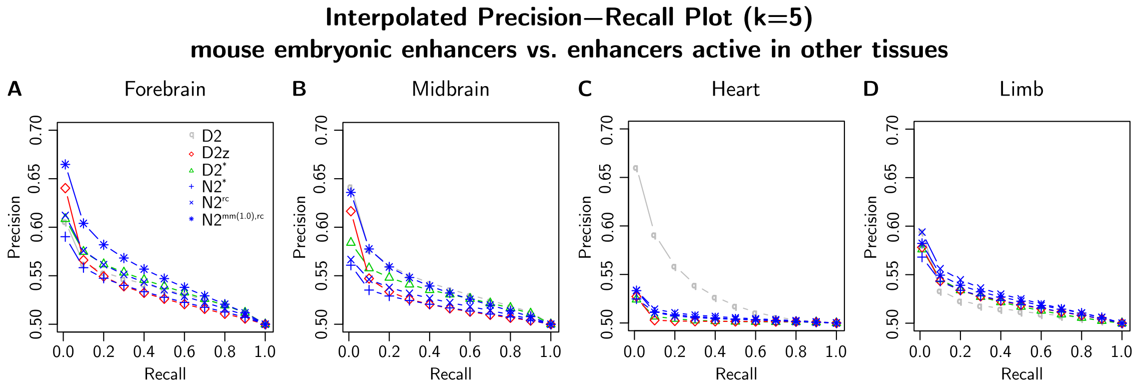


Figure VI.11: Tissue-specificity of enhancers, $k = 5$. Precision-Recall curve for mouse enhancers active in the same tissue versus enhancers active in different tissues. The performance is reduced compared to randomly selected genomic sequences. Nevertheless, enhancers active in the same tissue have higher pairwise scores. Enhancers obtained from embryonic heart tissue (D) can not be distinguished from other embryonic tissues by $N2$, $D2z$ or $D2^*$. Midbrain and forebrain enhancers are both active in neuronal tissues and show greater similarities then heart enhancers, which might lead to the poor performance. Heart enhancers show much weaker sequence conservation Blow *et al.* (2010), it is therefore expected that this set assembles a highly divergent set of enhancers with a limited number of shared transcription factor binding sites. Interestingly, the heart data set is the only data set were $D2$ shows better performance than the other methods, this might be caused by repetitive sequences which escaped repeat-masking.

VI.3 Supplementary Information for Chapter 4

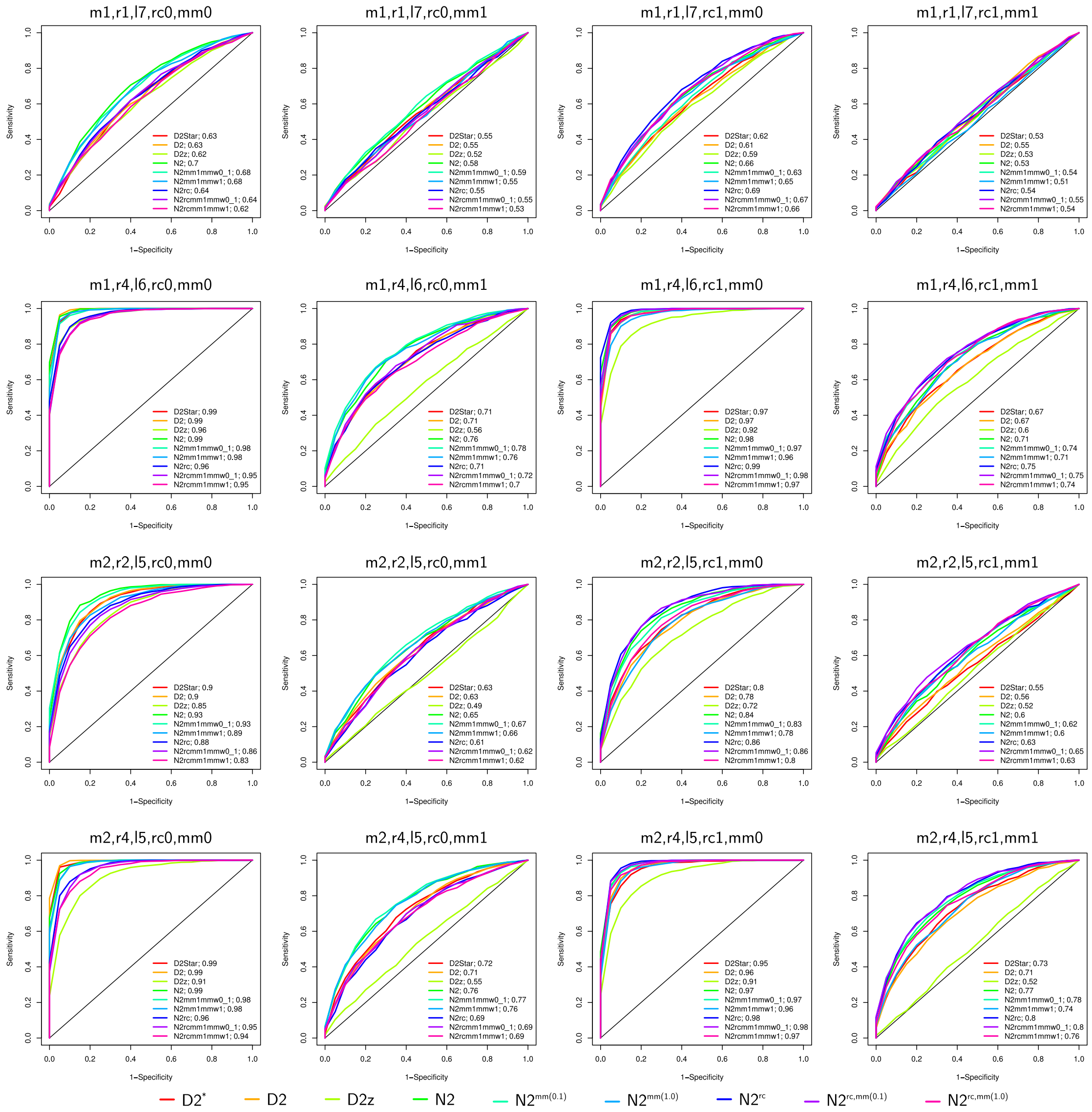


Figure VI.12: Classification with $N2$, simulated sequences. Shown are the ROC curves for different methods, parameters, and simulations. The numbers indicate the AUC.

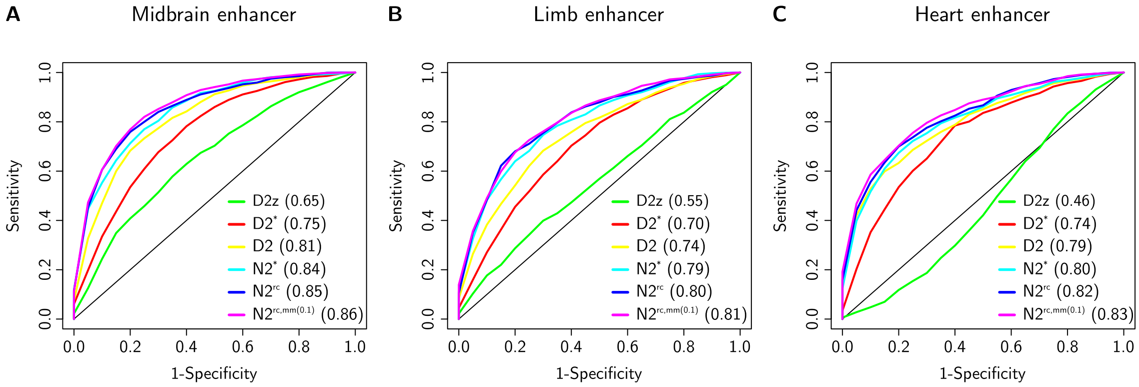


Figure VI.13: Classification of embryonic enhancers (positive set) versus random genomic sequences (negative set), numbers show the AUC (10-fold cross-validation). **(A)** Midbrain enhancers. **(B)** Limb enhancers. **(C)** Heart enhancers.

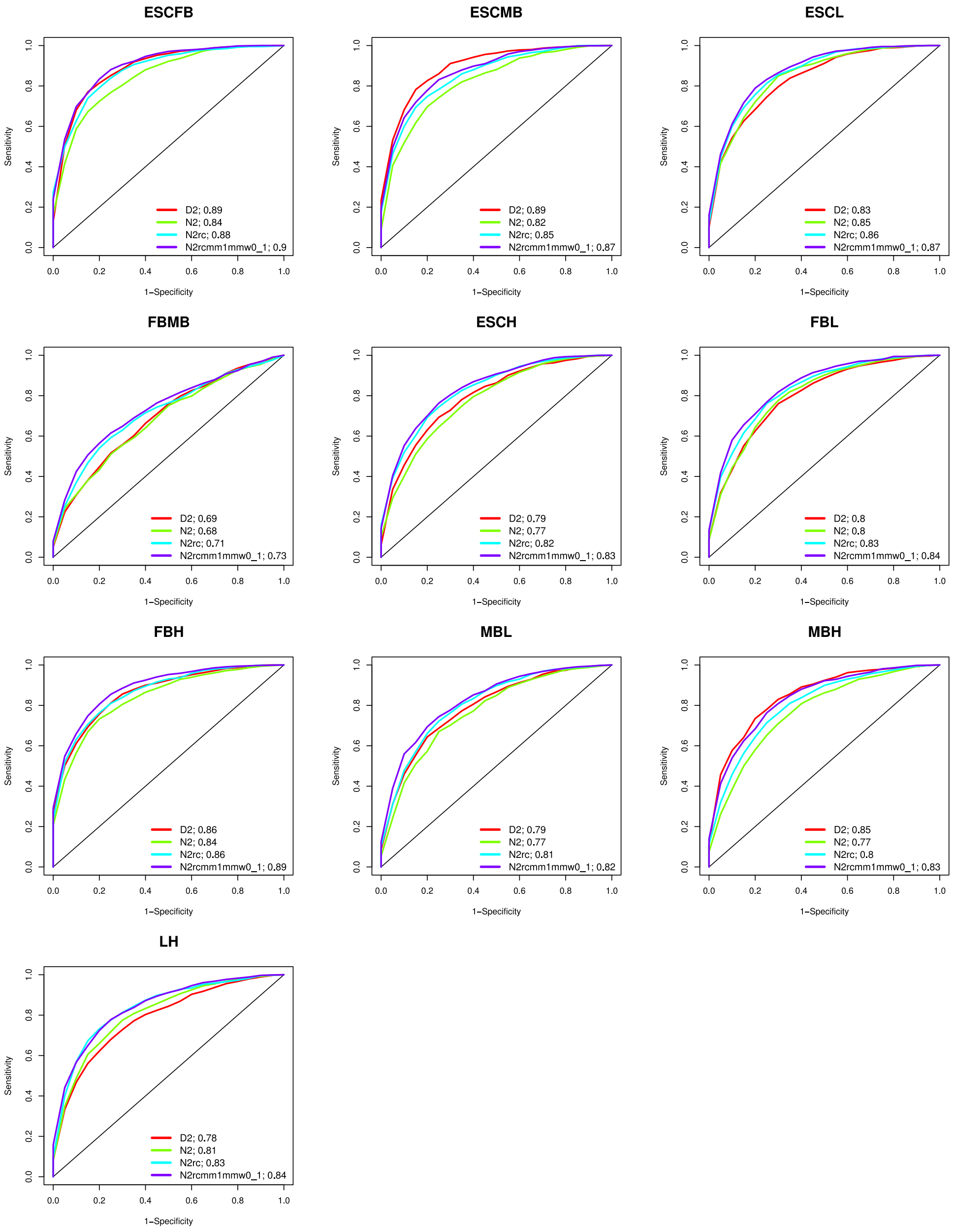


Figure VI.14: Classification of tissue-specific embryonic enhancers (positive set) versus enhancers active in other tissues (negative set), numbers show the AUC (10-fold cross-validation), $k = 6$.

Bibliography

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular biology of the cell*. Garland Science Taylor & Francis Group, 4 edition.

Almeida, J. S. and Vinga, S. (2009). Biological sequences as pictures: a generic two dimensional solution for iterated maps. *BMC Bioinformatics*, **10**, 100.

Almeida, J. S., Carriço, J. A., Maretzek, A., Noble, P. A., and Fletcher, M. (2001). Analysis of genomic sequences by chaos game representation. *Bioinformatics*, **17**(5), 429–437.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, **215**(3), 403–410.

Babaie, Y., Herwig, R., Greber, B., Brink, T. C., Wruck, W., Groth, D., Lehrach, H., Burdon, T., and Adjaye, J. (2007). Analysis of oct4-dependent transcriptional networks regulating self-renewal and pluripotency in human embryonic stem cells. *Stem Cells*, **25**(2), 500–510.

Bailey, T. L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**(12), 1653–1659.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*, **37**(Web Server issue), W202–W208.

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS Comput Biol*, **4**(10), e1000173.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, **27**(2), 573–580.

Berger, S. L., Kouzarides, T., Shiekhattar, R., and Shilatifard, A. (2009). An operational definition of epigenetics. *Genes Dev*, **23**(7), 781–783.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., and Lander, E. S. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**(2), 315–326.

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature*, **321**(6067), 209–213.

Blaisdell, B. E. (1986). A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc Natl Acad Sci U S A*, **83**(14), 5155–5159.

Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Bristow, J., Ren, B., Black, B. L., Rubin, E. M., Visel, A., and Pennacchio, L. A. (2010). ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet*, **42**(9), 806–810.

Bolshoy, A. (2003). DNA sequence analysis linguistic tools: contrast vocabularies, compositional spectra and linguistic complexity. *Appl Bioinformatics*, **2**(2), 103–112.

Bolshoy, A., Volkovich, Z., Kirzhner, V., and Barzily, Z. (2010). *Genome Clustering: From Linguistic Models to Classification of Genetic Texts (Studies in Computational Intelligence)*. Springer.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R., and Young, R. A. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**(6), 947–956.

Bryne, J. C., Valen, E., Tang, M.-H. E., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*, **36**(Database issue), D102–D106.

Burden, C., Jing, J., and Wilson, S. (2012). Alignment-free sequence comparison for biologically realistic sequences of moderate length. *Statistical Applications in Genetics and Molecular Biology*, **11**(1).

Burden, C. J., Kantorovitz, M. R., and Wilson, S. R. (2008). Approximate word matches between two random sequences. *Ann. Appl. Probab*, **18**, 1–21.

Carpenter, J. E., Christoffels, A., Weinbach, Y., and Hide, W. A. (2002). Assessment of the parallelization approach of d2-cluster for high-performance sequence clustering. *J Comput Chem*, **23**(7), 755–757.

Chakravarty, A., Carlson, J. M., Khetani, R. S., and Gross, R. H. (2007). A novel ensemble learning method for de novo computational identification of DNA binding sites. *BMC Bioinformatics*, **8**, 249.

Chan, Y. S., Yang, L., and Ng, H.-H. (2011). Transcriptional regulatory networks in embryonic stem cells. *Prog Drug Res*, **67**, 239–252.

Chavez, L., Jozefczuk, J., Grimm, C., Dietrich, J., Timmermann, B., Lehrach, H., Herwig, R., and Adjaye, J. (2010). Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res*, **20**(10), 1441–1450.

Chawengsaksophak, K., de Graaff, W., Rossant, J., Deschamps, J., and Beck, F. (2004). Cdx2 is essential for axial elongation in mouse development. *Proc Natl Acad Sci U S A*, **101**(20), 7641–7645.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y.-H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.-K., Clarke, N. D., Wei, C.-L., and Ng, H.-H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**(6), 1106–1117.

Church, D. M., Goodstadt, L., Hillier, L. W., Zody, M. C., Goldstein, S., She, X., Bult, C. J., Agarwala, R., Cherry, J. L., DiCuccio, M., Hlavina, W., Kapustin, Y., Meric, P., Maglott, D., Birtle, Z., Marques, A. C., Graves, T., Zhou, S., Teague, B., Potamousis, K., Churas, C., Place, M., Herschleb, J., Runnheim, R., Forrest, D., Amos-Landgraf, J., Schwartz, D. C., Cheng, Z., Lindblad-Toh, K., Eichler, E. E., Ponting, C. P., and Consortium, M. G. S. (2009). Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol*, **7**(5), e1000112.

Coller, H. A. and Kruglyak, L. (2008). Genetics. it's the sequence, stupid! *Science*, **322**(5900), 380–381.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**, 273–297.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., and Jaenisch, R. (2010). Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*, **107**(50), 21931–21936.

Dai, Q., Yang, Y., and Wang, T. (2008). Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. *Bioinformatics*, **24**(20), 2296–2302.

Das, M. K. and Dai, H.-K. (2007). A survey of DNA motif finding algorithms. *BMC Bioinformatics*, **8 Suppl 7**, S21.

Davidson, E. H. (2006). *Gene Regulatory Networks In Development and Evolution*. Academic Press.

Doering, A., Weese, D., Rausch, T., and Reinert, K. (2008). SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, **9**, 11.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Endo, M., Antonyak, M. A., and Cerione, R. A. (2009). Cdc42-mtor signaling pathway controls hes5 and pax6 expression in retinoic acid-dependent neural differentiation. *J Biol Chem*, **284**(8), 5107–5118.

Evans, M. J. and Kaufman, M. H. (1981). Establishment in culture of pluripotential cells from mouse embryos. *Nature*, **292**(5819), 154–156.

Fernandes, F., Freitas, A. T., Almeida, J. S., and Vinga, S. (2009). Entropic profiler - detection of conservation in genomes using information theory. *BMC Res Notes*, **2**, 72.

Forêt, S., Kantorovitz, M. R., and Burden, C. J. (2006). Asymptotic behaviour and optimal word size for exact and approximate word matches between random sequences. *BMC Bioinformatics*, **7 Suppl 5**, S21.

Forêt, S., Wilson, S. R., and Burden, C. J. (2009). Characterizing the d2 statistic: word matches in biological sequences. *Stat Appl Genet Mol Biol*, **8**(1), Article 43.

Fratkin, E., Naughton, B. T., Brutlag, D. L., and Batzoglou, S. (2006). MotifCut: regulatory motifs finding with maximum density subgraphs. *Bioinformatics*, **22**(14), e150–e157.

Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., Li, C. H., Meyer, L. R., Pohl, A., Raney, B. J., Rosenbloom, K. R., Smith, K. E., Haussler, D., and Kent, W. J. (2011). The ucsc genome browser database: update 2011. *Nucleic Acids Res*, **39**(Database issue), D876–D882.

Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *J Mol Biol*, **196**(2), 261–282.

Goodsell, D. S. (2005). TATA-binding protein. *RCSB Protein Data Bank*.

Gordân, R., Narlikar, L., and Hartemink, A. J. (2010). Finding regulatory DNA motifs using alignment-free evolutionary conservation information. *Nucleic Acids Res*, **38**(6), e90.

Goto, T., Macdonald, P., and Maniatis, T. (1989). Early and late periodic patterns of even skipped expression are controlled by distinct regulatory elements that respond to different spatial cues. *Cell*, **57**(3), 413–422.

Göke, J., Jung, M., Behrens, S., Chavez, L., O'Keeffe, S., Timmermann, B., Lehrach, H., Adjaye, J., and Vingron, M. (2011). Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development. *PLoS Comput Biol*, **7**(12), e1002304.

Göke, J., Schulz, M. H., Lasserre, J., and Vingron, M. (2012). Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts. *Bioinformatics*, **28**(5), 656–663.

Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., Lee, C. W. H., Ye, C., Ping, J. L. H., Mulawadi, F., Wong, E., Sheng, J., Zhang, Y., Poh, T., Chan, C. S., Kunarso, G., Shahab, A., Bourque, G., Cacheux-Rataboul, V., Sung, W.-K., Ruan, Y., and Wei, C.-L. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet*, **43**(7), 630–638.

He, Q., Bardet, A. F., Patton, B., Purvis, J., Johnston, J., Paulson, A., Gogol, M., Stark, A., and Zeitlinger, J. (2011). High conservation of transcription factor binding and evidence for combinatorial regulation across six drosophila species. *Nat Genet*, **43**(5), 414–420.

Heintzman, N. D. and Ren, B. (2009). Finding distal regulatory elements in the human genome. *Curr Opin Genet Dev*, **19**(6), 541–549.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M., and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*.

Hemberg, M. and Kreiman, G. (2011). Conservation of transcription factor binding events predicts gene expression across species. *Nucleic Acids Res*, **39**(16), 7092–7102.

Hide, W., Burke, J., and Davison, D. B. (1994). Biological evaluation of d2, an algorithm for high-performance sequence comparison. *J Comput Biol*, **1**(3), 199–215.

Hufton, A. L., Mathia, S., Braun, H., Georgi, U., Lehrach, H., Vingron, M., Poustka, A. J., and Panopoulou, G. (2009). Deeply conserved chordate noncoding sequences preserve genome synteny but do not drive gene duplicate retention. *Genome Res*, **19**(11), 2036–2051.

Huggins, P., Zhong, S., Shiff, I., Beckerman, R., Laptenko, O., Prives, C., Schulz, M. H., Simon, I., and Bar-Joseph, Z. (2011). Decod: fast and accurate discriminative dna motif finding. *Bioinformatics*, **27**(17), 2361–2367.

Jain, A., Murty, M., and Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys*, **31**(3), 264–323.

Jing, J., Wilson, S., and Burden, C. (2011). Weighted k-word matches: a sequence comparison tool for proteins. *ANZIAM J*, **52**, C172–C189.

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

Joseph, J. and Sasikumar, R. (2006). Chaos game representation for comparison of whole genomes. *BMC Bioinformatics*, **7**, 243.

Jung, M., Peterson, H., Chavez, L., Kahlem, P., Lehrach, H., Vilo, J., and Adjaye, J. (2010). A data integration approach to mapping OCT4 gene regulatory networks operative in embryonic stem cells and embryonal carcinoma cells. *PLOS One*, **5**(5).

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J., and Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, **467**(7314), 430–435.

Kantorovitz, M. R., Robinson, G. E., and Sinha, S. (2007). A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, **23**(13), i249–i255.

Kantorovitz, M. R., Kazemian, M., Kinston, S., Miranda-Saavedra, D., Zhu, Q., Robinson, G. E., Göttgens, B., Halfon, M. S., and Sinha, S. (2009). Motif-blind, genome-wide discovery of cis-regulatory modules in drosophila and mouse. *Dev Cell*, **17**(4), 568–579.

Karatzoglou, A. and Meyer, D. (2006). Support vector machines in r. *journal of Statistical Software*, **15**(9).

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, **11**(9), 1–20.

Karlin, S. and Ladunga, I. (1994). Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci U S A*, **91**(26), 12832–12836.

Kirzhner, V., Paz, A., Volkovich, Z., Bolshoy, A., and Nevo, E. (2002). Compositional spectrum - revealing patterns for genomic sequence characterization and comparison. *Physica A*, **312**, 447–457.

Kirzhner, V., Nevo, E., Korol, A., and Bolshoy, A. (2003). A large-scale comparison of genomic sequences: one promising approach. *Acta Biotheor*, **51**(2), 73–89.

Klein, H. and Vingron, M. (2007). Using transcription factor binding site co-occurrence to predict regulatory regions. *Genome Inform*, **18**, 109–118.

Kleinjan, D. A. and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet*, **76**(1), 8–32.

Koohy, H., Dyer, N. P., Reid, J. E., Koentges, G., and Ott, S. (2010). An alignment-free model for comparison of regulatory sequences. *Bioinformatics*, **26**(19), 2391–2397.

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, **128**(4), 693–705.

Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet*, **42**(7), 631–634.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**(3), R25.

Lee, D., Karchin, R., and Beer, M. A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res*.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, **298**(5594), 799–804.

Lee, T. I., Jenner, R. G., Boyer, L. A., Guenther, M. G., Levine, S. S., Kumar, R. M., Chevalier, B., Johnstone, S. E., Cole, M. F., ichi Isono, K., Koseki, H., Fuchikami, T., Abe, K., Murray, H. L., Zucker, J. P., Yuan, B., Bell, G. W., Herbolsheimer, E., Hannett, N. M., Sun, K., Odom, D. T., Otte, A. P., Volkert, T. L., Bartel, D. P., Melton, D. A., Gifford, D. K., Jaenisch, R., and Young, R. A. (2006). Control of developmental regulators by polycomb in human embryonic stem cells. *Cell*, **125**(2), 301–313.

Lemischka, I. R. (2010). Hooking up with oct4. *Cell Stem Cell*, **6**(4), 291–292.

Leslie, C., Eskin, E., and Noble, W. S. (2002). The spectrum kernel: a string kernel for svm protein classification. *Pac Symp Biocomput*, pages 564–575.

Leslie, C. S., Eskin, E., Cohen, A., Weston, J., and Noble, W. S. (2004). Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**(4), 467–476.

Li, M., Badger, J. H., Chen, X., Kwong, S., Kearney, P., and Zhang, H. (2001). An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, **17**(2), 149–154.

Lippert, R. A., Huang, H., and Waterman, M. S. (2002). Distributional regimes for the number of k-word matches between two random sequences. *Proc Natl Acad Sci U S A*, **99**(22), 13980–13989.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*.

Lister, R., Pelizzola, M., Kida, Y. S., Hawkins, R. D., Nery, J. R., Hon, G., Antosiewicz-Bourget, J., O'Malley, R., Castanon, R., Klugman, S., Downes, M., Yu, R., Stewart, R., Ren, B., Thomson, J. A., Evans, R. M., and Ecker, J. R. (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, **471**(7336), 68–73.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, **28**, 129–137.

Loh, Y.-H., Wu, Q., Chew, J.-L., Vega, V. B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., Wong, K.-Y., Sung, K. W., Lee, C. W. H., Zhao, X.-D., Chiu, K.-P., Lipovich, L., Kuznetsov, V. A., Robson, P., Stanton, L. W., Wei, C.-L., Ruan, Y., Lim, B., and Ng, H.-H. (2006). The oct4 and nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet*, **38**(4), 431–440.

Ma, X., Kulkarni, A., Zhang, Z., Xuan, Z., Serfling, R., and Zhang, M. Q. (2012). A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. *Nucleic Acids Res*.

Machanick, P. and Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**(12), 1696–1697.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press*, pages 281–297.

Mallanna, S. K., Ormsbee, B. D., Iacovino, M., Gilmore, J. M., Cox, J. L., Kyba, M., Washburn, M. P., and Rizzino, A. (2010). Proteomic analysis of sox2-associated proteins during early stages of mouse embryonic stem cell differentiation identifies sox21 as a novel regulator of stem cell fate. *Stem Cells*, **28**(10), 1715–1727.

Manke, T., Bringas, R., and Vingron, M. (2003). Correlating protein-DNA and protein-protein interaction networks. *J Mol Biol*, **333**(1), 75–85.

Mardis, E. R. (2007). ChIP-seq: welcome to the new frontier. *Nat Methods*, **4**(8), 613–614.

Marson, A., Levine, S. S., Cole, M. F., Frampton, G. M., Brambrink, T., Johnstone, S., Guenther, M. G., Johnston, W. K., Wernig, M., Newman, J., Calabrese, J. M., Dennis, L. M., Volkert, T. L., Gupta, S., Love, J., Hannett, N., Sharp, P. A., Bartel, D. P., Jaenisch, R., and Young, R. A. (2008). Connecting microrna genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**(3), 521–533.

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, **7**, 29–59.

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

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol*, **28**(5), 495–501.

Mysickova, A. and Vingron, M. (2012). Detection of interacting transcription factors in human tissues using predicted DNA binding affinity. *BMC Genomics*, **13 Suppl 1**, S2.

Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., Okita, K., Mochiduki, Y., Takizawa, N., and Yamanaka, S. (2008). Generation of induced pluripotent stem cells without myc from mouse and human fibroblasts. *Nat Biotechnol*, **26**(1), 101–106.

nature.com (2012). Online, 26/03/2012. `http://blogs.nature.com/a_mad_hemorrhage/2011/05/15/formal-definitions-and-complex-systems`.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48**(3), 443–453.

NIH (2011). Roadmap epigenomics. http://www.roadmapepigenomics.org.

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, **10**(10), 669–680.

Pavesi, G., Mereghetti, P., Zambelli, F., Stefani, M., Mauri, G., and Pesole, G. (2006). MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Res*, **34**(Web Server issue), W566–W570.

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., Plajzer-Frick, I., Akiyama, J., Val, S. D., Afzal, V., Black, B. L., Couronne, O., Eisen, M. B., Visel, A., and Rubin, E. M. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**(7118), 499–502.

Pevzner, P. A. (1992). Statistical distance between texts and filtration methods in sequence comparison. *Comput Appl Biosci*, **8**(2), 121–127.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, **20**(1), 110–121.

Przyborski, S. A., Christie, V. B., Hayman, M. W., Stewart, R., and Horrocks, G. M. (2004). Human embryonal carcinoma stem cells: models of embryonic development in humans. *Stem Cells Dev*, **13**(4), 400–408.

Putnam, N. H., Butts, T., Ferrier, D. E. K., Furlong, R. F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.-K., Benito-Gutiérrez, E. L., Dubchak, I., Garcia-Fernàndez, J., Gibson-Brown, J. J., Grigoriev, I. V., Horton, A. C., de Jong, P. J., Jurka, J., Kapitonov, V. V., Kohara, Y., Kuroki, Y., Lindquist, E., Lucas, S., Osoegawa, K., Pennacchio,

L. A., Salamov, A. A., Satou, Y., Sauka-Spengler, T., Schmutz, J., Shin-I, T., Toyoda, A., Bronner-Fraser, M., Fujiyama, A., Holland, L. Z., Holland, P. W. H., Satoh, N., and Rokhsar, D. S. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature*, **453**(7198), 1064–1071.

Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6), 841–842.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, **470**(7333), 279–283.

Reinert, G., Chew, D., Sun, F., and Waterman, M. S. (2009). Alignment-free sequence comparison (i): Statistics and power. *J Comput Biol*.

Robin, S., Rodolphe, F., and Schbath, S. (2005). *DNA, Words and Models*. Cambridge University Press.

Roider, H. G., Kanhere, A., Manke, T., and Vingron, M. (2007). Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, **23**(2), 134–141.

Rätsch, G., Sonnenburg, S., and Schäfer, C. (2006). Learning interpretable svms for biological sequence classification. *BMC Bioinformatics*, **7 Suppl 1**, S9.

Sakabe, N. J. and Nobrega, M. A. (2010). Genome-wide maps of transcription regulatory elements. *Wiley Interdiscip Rev Syst Biol Med*, **2**(4), 422–437.

Sandberg, M., Källström, M., and Muhr, J. (2005). Sox21 promotes the progression of vertebrate neurogenesis. *Nat Neurosci*, **8**(8), 995–1001.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Flicek, P., and Odom, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, **328**(5981), 1036–1040.

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

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J Mol Biol*, **188**(3), 415–431.

Schöler, H. R., Balling, R., Hatzopoulos, A. K., Suzuki, N., and Gruss, P. (1989). Octamer binding proteins confer transcriptional activity in early mouse embryogenesis. *EMBO J*, **8**(9), 2551–2557.

Schöler, H. R., Dressler, G. R., Balling, R., Rohdewohld, H., and Gruss, P. (1990). Oct-4: a germline-specific transcription factor mapping to the mouse t-complex. *EMBO J*, **9**(7), 2185–2195.

Sene, K. H., Porter, C. J., Palidwor, G., Perez-Iratxeta, C., Muro, E. M., Campbell, P. A., Rudnicki, M. A., and Andrade-Navarro, M. A. (2007). Gene function in early mouse embryonic stem cell differentiation. *BMC Genomics*, **8**, 85.

Sims, G. E., Jun, S.-R., Wu, G. A., and Kim, S.-H. (2009a). Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions. *Proc Natl Acad Sci U S A*, **106**(8), 2677–2682.

Sims, G. E., Jun, S.-R., Wu, G. A., and Kim, S.-H. (2009b). Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proc Natl Acad Sci U S A*, **106**(40), 17077–17082.

Sinha, S. and Tompa, M. (2002). Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*, **30**(24), 5549–5560.

Sinha, S. and Tompa, M. (2003). YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*, **31**(13), 3586–3588.

Small, S., Kraut, R., Hoey, T., Warrior, R., and Levine, M. (1991). Transcriptional regulation of a pair-rule stripe in drosophila. *Genes Dev*, **5**(5), 827–839.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, **147**(1), 195–197.

Sonnenburg, S., Zien, A., and Rätsch, G. (2006). ARTS: accurate recognition of transcription starts in human. *Bioinformatics*, **22**(14), e472–e480.

Sonnenburg, S., Schweikert, G., Philips, P., Behr, J., and Rätsch, G. (2007). Accurate splice site prediction using support vector machines. *BMC Bioinformatics*, **8 Suppl 10**, S7.

Sonnenburg, S., Raetsch, G., Henschel, S., Widmer, C., Behr, J., Zien, A., de Bona, F., Binder, A., Gehl, C., and Franc, V. (2010). The shogun machine learning toolbox. *Journal of Machine Learning Research*, **11**, 1799–1802.

Spence, J. R., Mayhew, C. N., Rankin, S. A., Kuhar, M. F., Vallance, J. E., Tolle, K., Hoskins, E. E., Kalinichenko, V. V., Wells, S. I., Zorn, A. M., Shroyer, N. F., and Wells, J. M. (2011). Directed differentiation of human pluripotent stem cells into intestinal tissue in vitro. *Nature*, **470**(7332), 105–109.

Stuart, G. W., Moffett, K., and Leader, J. J. (2002a). A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol Biol Evol*, **19**(4), 554–562.

Stuart, G. W., Moffett, K., and Baker, S. (2002b). Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics*, **18**(1), 100–108.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**(43), 15545–15550.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, **131**(5), 861–872.

Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C., Thieffry, D., and van Helden, J. (2011). RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res*, **39**(Web Server issue), W86–W91.

Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., and van Helden, J. (2012). Rsat peak-motifs: motif analysis in full-size chip-seq datasets. *Nucleic Acids Res*, **40**(4), e31.

Thomson, J. A., Itskovitz-Eldor, J., Shapiro, S. S., Waknitz, M. A., Swiergiel, J. J., Marshall, V. S., and Jones, J. M. (1998). Embryonic stem cell lines derived from human blastocysts. *Science*, **282**(5391), 1145–1147.

Tjian, R. and Maniatis, T. (1994). Transcriptional activation: a complex puzzle with few easy pieces. *Cell*, **77**(1), 5–8.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., Moor, B. D., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, **23**(1), 137–144.

van den Berg, D. L. C., Snoek, T., Mullin, N. P., Yates, A., Bezstarosti, K., Demmers, J., Chambers, I., and Poot, R. A. (2010). An oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell*, **6**(4), 369–381.

van Helden, J. (2004). Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics*, **20**(3), 399–406.

van Helden, J., André, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol*, **281**(5), 827–842.

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

Vinga, S. and Almeida, J. (2003). Alignment-free sequence comparison-a review. *Bioinformatics*, **19**(4), 513–523.

Vingron, M., Brazma, A., Coulson, R., van Helden, J., Manke, T., Palin, K., Sand, O., and Ukkonen, E. (2009). Integrating sequence, evolution and functional genomics in regulatory genomics. *Genome Biol*, **10**(1), 202.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. A. (2007). VISTA enhancer browser–a database of tissue-specific human enhancers. *Nucleic Acids Res*, **35**(Database issue), D88–D92.

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. (2009). Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**(7231), 854–858.

von Hippel, P. H. and Berg, O. G. (1986). On the specificity of DNA-protein interactions. *Proc Natl Acad Sci U S A*, **83**(6), 1608–1612.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**(301), 236–244.

Wikimedia Commons (2012a). Online, 26/03/2012. `https://upload.wikimedia.org/wikipedia/commons/3/36/DNA_transcription.svg`.

Wikimedia Commons (2012b). Online, 26/03/2012. `http://upload.wikimedia.org/wikipedia/commons/4/4b/Chromatin_Structures.png`.

Wikimedia Commons (2012c). Online, Electron Microscopy Facility at The National Cancer Institute at Frederick, 26/03/2012. `http://upload.wikimedia.org/wikipedia/commons/2/24/Red_White_Blood_cells.jpg`.

Wikimedia Commons (2012d). Online, Jkwchui, 27/03/2012. `http://upload.wikimedia.org/wikipedia/commons/1/15/ChIP-sequencing.svg`.

Wikimedia Commons (2012e). Online, LadyofHats, 26/03/2012. `http://upload.wikimedia.org/wikipedia/commons/b/b1/Ribosome_mRNA_translation_en.svg`.

Wikimedia Commons (2012f). Online, Madeleine Price Ball, 26/03/2012. `http://commons.wikimedia.org/wiki/File:DNA_chemical_structure.svg`.

Wikimedia Commons (2012g). Online Mariana Ruiz, 26/03/2012. `http://upload.wikimedia.org/wikipedia/commons/8/8f/DNA_replication_en.svg`.

Wikimedia Commons (2012h). Online, Sponk , 26/03/2012. `http://upload.wikimedia.org/wikipedia/commons/f/fd/Difference_DNA_RNA-DE.svg`.

Wilson, M. D., Barbosa-Morais, N. L., Schmidt, D., Conboy, C. M., Vanes, L., Tybulewicz, V. L. J., Fisher, E. M. C., Tavaré, S., and Odom, D. T. (2008). Species-specific transcription in mice carrying human chromosome 21. *Science*, **322**(5900), 434–438.

Wu, G. A., Jun, S.-R., Sims, G. E., and Kim, S.-H. (2009). Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. *Proc Natl Acad Sci U S A*, **106**(31), 12826–12831.

Wu, T. J., Burke, J. P., and Davison, D. B. (1997). A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. *Biometrics*, **53**(4), 1431–1439.

Wu, T. J., Hsieh, Y. C., and Li, L. A. (2001). Statistical measures of DNA sequence dissimilarity under markov chain models of base composition. *Biometrics*, **57**(2), 441–448.

Yu, J., Vodyanik, M. A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J. L., Tian, S., Nie, J., Jonsdottir, G. A., Ruotti, V., Stewart, R., Slukvin, I. I., and Thomson, J. A. (2007). Induced pluripotent stem cell lines derived from human somatic cells. *Science*, **318**(5858), 1917–1920.

Zemojtel, T., Kielbasa, S. M., Arndt, P. F., Chung, H.-R., and Vingron, M. (2009). Methylation and deamination of CpGs generate p53-binding sites on a genomic scale. *Trends Genet*, **25**(2), 63–66.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of chip-seq (macs). *Genome Biol*, **9**(9), R137.

Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E. E. M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, **462**(7269), 65–70.

List of Abbreviations

A	adenine
C	cytosine
G	guanine
T	thymine
AUC	area under curve
bp	base pair
ChIP	chromatin immunoprecipitation
DNA	deoxyribonucleic acid
EC cell	embryonal carcinoma cells
ES cell	embryonic stem cell
ESC	embryonic stem cell
FDR	false discovery rate
GO	gene ontology
GSEA	gene set enrichment analysis
hES cell	human embryonic stem cell
i.i.d.	independent and identically distributed
iPS cell	induced pluripotent stem cell
mES cell	mouse embryonic stem cell
mRNA	messenger RNA

NP cell	neural progenitor cell
PR	precision recall
PWM	position weight matrix
RNA	ribonucleic acid
ROC	receiver operating characteristic
SVM	support vector machine
TSS	transcription start site

Notations

S	sequence
$S[i...i+j]$	sub-sequence from position i to i+j
l	length of sequence
A	alphabet, `A,C,G,T` in the case of DNA
w	word/ k-mer
k	length of word/ k-mer
N_w^S, N_w	number of occurrences of w in sequence S
$Y_i(w)$	binary indicator for an occurrence of w starting at position i
D	set of all words w of length k ('dictionary')
$\|D\|$	size of D (4^k for DNA)
N^S	vector of word occurrences N_w^S for all $w \in D$
ϵ	word overlap indicator
$\mu(w)$	word probability
$\mathbb{E}[X]$	expected value
$\mathbb{V}[X]$	variance
$\mathbb{C}\mathbf{ov}[X_1, X_2]$	covariance
$\pi(a,b)$	transition probability
μ	stationary distribution
$\mathcal{L}$	likelihood

$n(w)$	word neighbourhood
a_w	word weight
$N_{n(w)}$	weighted word neighbourhood count
$\tilde{N}_w$	standardised, weighted word neighbourhood count
$\hat{N}_w$	normalised, standardised, weighted word neighbourhood count

Glossary

Promoter Regulatory DNA sequence upstream of transcription start sites.

Enhancer Regulatory DNA sequence that enhances transcription, can be several kilo bases distal from the nearest transcription start site.

Motif Sequence motif, here: a nucleotide pattern which is recognised by a transcription factor.

Peak Accumulation of reads from ChIP-Seq data, resulting in a peak-like shape in genome browser visualisations. A peak corresponds to a DNA-protein interaction.

Word Count The number of occurrences of a word w in a sequence.

Word Probability The probability that a word w occurs at a specific position in a sequence.

Expected Count The expected number of occurrences of a word w in a sequence.

Dictionary Set of all words w of length k (D)

k-mer Word of length k

Dotplot Graphical representation of the similarity of two sequences S^1 and S^2, a dot at position (i, j) indicates a similar word at position i in S^1 and position j in S^2.

Word Neighbourhood Count The number of occurrences of all words w' which are in the neighbourhood of the word w.

Motif Finding Identification of functional patterns (*motifs*) in sequences.

Availability

OCT4 Chip-Seq in NCCIT Cells

OCT4 ChIP-Seq data in NCCIT cells (Göke *et al.*, 2011) is publicly available at the European Nucleotide Archive (`http://www.ebi.ac.uk/ena/`, accession number ERP001004).

Public Data Sets

All data sets used in this work have been downloaded in fastq format from the European Nucleotide Archive, the accession numbers are summarised in Table 2.1.

Data Access

The mapped sequencing data can be accessed at `http://enhancer.molgen.mpg.de`, where I provide a human and mouse genome browser.

N2

The code for $N2$, $D2$, $D2^*$, $D2z$ and the underlying word statistics is available at `http://www.seqan.de/` along with examples and complete documentation. The executable (ALF) can be downloaded at `http://www.seqan.de/projects/alf/`.

Zusammenfassung

Der Menschliche Organismus besteht aus vielen hundert verschiedenen Zelltypen. Jede Zelle besitzt das gleiche Repertoire an Genen, von denen jedoch nur ein Teil exprimiert wird. Die große Vielfalt an verschiedenen Zellen wird durch zelltypspezifische Regulation der Genexpression ermöglicht. Die Information, wann und wo ein Gen aktiv ist, ist in der DNA kodiert und kann durch DNA-bindende Proteine, den Transkriptionsfaktoren, gelesen werden. Die DNA-Bindestellen können direkt neben einem Gen liegen (Promoter), aber auch viele tausend Basenpaare entfernt sein (Enhancer). Enhancer spielen eine wichtige Rolle in der Zelldifferenzierung und der Embryonalentwicklung und sind entscheidend daran beteiligt, dass sich die große Vielfalt von Zelltypen im ausgewachsenen Organismus bilden kann. Diese Dissertation beschäftigt sich mit der Analyse von solchen Enhancern, regulatorischen Sequenzen die weit entfernt von Genen deren Expression steuern.

Zunächst wird eine Einführung in die Grundlagen der molekularen Genetik und Genregulation gegeben (Kapitel 1). Im zweiten Kapitel werden genomweite Datensätze von DNA-Bindestellen von Transkriptionsfaktoren in embryonalen Stammzellen integriert um den Einfluss der Kombination von DNA-bindenden Proteinen auf die Transkription und auf die Evolution von regulatorischen Sequenzen zu analysieren. Anschließend (Kapitel 3) wird eine neue, nicht-alignment-basierende Methode ($N2$) vorgestellt, welche die paarweise Ähnlichkeit von regulatorischen Sequenzen messen kann, analog zu Alignments von Protein-kodierenden Genen. $N2$ wird auf gewebespezifische regulatorische Sequenzen angewendet und es wird gezeigt, dass Enhancer-Sequenzen die in demselben Gewebe aktiv sind eine höhere $N2$-Ähnlichkeit aufweisen. Kapitel 4 verwendet die Wort-Statistiken auf denen $N2$ basiert um große Datensätze regulatorischer Sequenzen zu analysieren. Die vielfältigen Möglichkeiten die, $N2$ bietet, werden anhand von aktuellen Forschungsfragen (Sequenzmotif-Identifizierung, Klassifizierung, Clusteranalyse) aufgezeigt. Abschließend (Kapitel 5) werden die Ergebnisse in einem gemeinsamen Kontext zusammengefasst.

Die Ergebnisse aus Kapitel 2 wurden im Dezember 2011 veröffentlicht (Göke *et al.*,

2011), die Ergebnisse aus Kapitel 3 wurden im Januar 2012 veröffentlicht (Göke *et al.*, 2012).

Zusammengefasst verschafft diese Arbeit neue Erkenntnisse in die kombinatorische Regulation der Genexpression und präsentiert eine neue Methode für den paarweisen Vergleich von Enhancern, die abschließend auf die Analyse großer Datensätze angewendet wird.

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Jonathan Göke

Berlin, Mai 2011