Department of Computer Science
Series of Publications A
Report A-2007-7

Computational Methods for Locating and Analyzing Conserved Gene Regulatory DNA Elements

Kimmo Palin

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium XII, University Main Building, on November 23rd, 2007, at noon.

University of Helsinki
Finland

Contact information

Postal address:
Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Email address: postmaster@cs.Helsinki.FI (Internet)

URL: http://www.cs.Helsinki.FI/

Telephone: +358 9 1911

Telefax: +358 9 191 51120

ISSN 1238-8645
ISBN 978-952-10-4352-9 (paperback)
ISBN 978-952-10-4353-6 (PDF)
URL: http://urn.fi/URN:ISBN:978-952-10-4353-6
Computing Reviews (1998) Classification: G.2.1, G.3, I.6.5, J.3
Helsinki 2007
Helsinki University Printing House

Computational Methods for Locating and Analyzing Conserved Gene Regulatory DNA Elements

Kimmo Palin

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
kimmo.palin@helsinki.fi
http://www.cs.helsinki.fi/u/kpalin/

PhD Thesis, Series of Publications A, Report A-2007-7
Helsinki, November 2007, 130 pages
ISSN 1238-8645
ISBN 978-952-10-4352-9 (paperback)
ISBN 978-952-10-4353-6 (PDF)
URL: http://urn.fi/URN:ISBN:978-952-10-4353-6

Abstract

This thesis presents methods for locating and analyzing cis–regulatory DNA elements involved with the regulation of gene expression in multicellular organisms. The regulation of gene expression is carried out by the combined effort of several transcription factor proteins collectively binding the DNA on the cis–regulatory elements. Only sparse knowledge of the 'genetic code' of these elements exists today. An automatic tool for discovery of putative cis–regulatory elements could help their experimental analysis, which would result in a more detailed view of the cis–regulatory element structure and function.

We have developed a computational model for the evolutionary conservation of cis–regulatory elements. The elements are modeled as evolutionarily conserved clusters of sequence–specific transcription factor binding sites. We give an efficient dynamic programming algorithm that locates the putative cis–regulatory elements and scores them according to the conservation model. A notable proportion of the high–scoring DNA sequences show transcriptional enhancer activity in transgenic mouse embryos.

The conservation model includes four parameters whose optimal values are estimated with simulated annealing. With good parameter values the model discriminates well between the DNA sequences with evolutionarily conserved

cis–regulatory elements and the DNA sequences that have evolved neutrally. In further inquiry, the set of highest scoring putative cis–regulatory elements were found to be sensitive to small variations in the parameter values.

The statistical significance of the putative cis–regulatory elements is estimated with the Two Component Extreme Value Distribution. The p–values grade the conservation of the cis–regulatory elements above the neutral expectation. The parameter values for the distribution are estimated by simulating the neutral DNA evolution.

The conservation of the transcription factor binding sites can be used in the upstream analysis of regulatory interactions. This approach may provide mechanistic insight to the transcription level data from, e.g., microarray experiments. Here we give a method to predict shared transcriptional regulators for a set of co–expressed genes.

The EEL (Enhancer Element Locator) software implements the method for locating putative cis–regulatory elements. The software facilitates both interactive use and distributed batch processing. We have used it to analyze the non–coding regions around all human genes with respect to the orthologous regions in various other species including mouse. The data from these genome–wide analyzes is stored in a relational database which is used in the publicly available web services for upstream analysis and visualization of the putative cis–regulatory elements in the human genome.

Computing Reviews (1998) Categories and Subject Descriptors:
G.2.1 Combinatorics: Combinatorial algorithms
G.3 Probability and statistics: Contingency table analysis, Probabilistic algorithms (including Monte Carlo)
I.6.5 Model development
J.3 Life and medical sciences

General Terms:
Algorithms, Bioinformatics

Additional Key Words and Phrases:
gene regulation, cis–regulatory module, transcription factor binding site, comparative genomics

Acknowledgements

My first and foremost thanks goes to my adviser Esko Ukkonen who has taken care of my academic well–being and guided me through all of these years in the University. His continuing support has led me past the ups and downs encountered on the road to dissertation.

Maybe the biggest factor for me to finally reach this goal is the outstanding guidance and support by Jussi Taipale. His astonishing intellect, great intuition and unfaltering confidence have saved my day (not to mention my results) many times. His can–do attitude combined with his skill of doing make him truly a great scientist that can be regarded as an example for anyone.

I am indebted to Outi Hallikas and Teemu Kivioja who have greatly expanded my knowledge of biology, bioinformatics, biomedical research and, in the general stuff happening in Meilahti. I am especially grateful for Teemu for being my walking dictionary of information about graduating as a PhD. from the Department of Computer Science.

The friends at work deserve my sincere thank you for dragging me to lunch or coffee every now or then. Esa, Pasi, Veli, Matti, Janne, Juho, Taneli and others have been great company and offered their invaluable intellectual support. Ari Rantanen deserves a special thank you firstly for the rope and secondly for the example of writing a dissertation without using it.

I would like to thank Alvis Brazma who is a great contact in the heart of European bioinformatics and Jaak Vilo who was my first role model when entering the world of research. I also thank Martin Vingron for giving me an idea that almost worked and thus gave rise to some of the better ideas in this thesis. Margus Lukk I thank for teaching me much of the biology I know. A big thanks also to Matthias Berg and Jaakko Nyman who did a terrific job implementing some of my ideas. I also thank Thomas Schlitt for great collaboration during my yeast years.

The invaluable comments by Sami Kaski and Harri Lähdesmäki were vital for getting this manuscript to a publishable shape. Marina Kurtén

did a great work for checking the text for language errors. Petri Kutvonen, Pekka Niklander and the whole computing facilities staff have made the work on the department computers a breeze.

I thank Heikki Mannila for taking part in my work as the second adviser and also as the chair of the graduate school ComBi, that I gratefully acknowledge for the funding and for the contacts created at the seminars and the BREW workshops. I am also grateful for the funding provided by the From Data to Knowledge research unit and the European Union.

Finally, I thank my parents, Asko and Helvi, and my family, Merja, Emma and Hugo just for being there.

Contents

Chapter 1

Introduction

In this thesis I have developed computational methods and tools for finding and analyzing genomic DNA sequences that regulate the gene expression in multicellular organisms. The gene expression, the process where genomic DNA is *transcribed* to RNA, is the main determinant of gene activity within a living cell. The transcribed RNA is usually *translated* to proteins that make up the cell taking care of such things as cell structure, metabolism, cell–to–cell signaling, transcription, translation and also the regulation of gene expression. Aberrant regulation of the gene expression is the cause of many developmental defects and malignancies such as cancer [TB01, BC00].

The gene expression is regulated by *transcription factor* proteins binding to their *binding sites* on the DNA near the controlled gene. Several of these binding sites are clustered close together to form a *cis–regulatory element* which enables temporally and spatially specific expression response dependent on the bound transcription factors [SBL92, AD97]. Finding the cis–regulatory elements in the genome would greatly advance our understanding of the *gene regulatory networks* underlying development and cancer [Tho17, Tur52, DRO$^+$02, TB01].

The approach taken in this work for finding the *cis–regulatory elements* is both comparative and model–based. On one hand the system looks for sequences preserved in the genomes of two species over a long history of independent evolution and on the other hand it looks for sequences whose content seems feasible under current biological understanding. This kind of approach has been highly successful in gene finding and we can exploit it also in finding the cis–regulatory elements that are less conserved than the protein–coding parts of the genome [OMRM$^+$99, PAA$^+$03].

The main computational problem in finding the putative cis–regulatory elements is the question how to efficiently find the segments of DNA that follow the cis–regulatory element model the best. In this work this problem

is solved with a dynamic programming algorithm similar to the well–known local sequence alignment [SW81]. The implementation is fast enough for pairwise analysis of practically whole genomes.

The genome–wide analysis is made possible by the availability of several mammalian genome sequences together with the growing databases of transcription factor binding site motifs [VAM+01, LLB+01, MHE+05, WLTB+02, GWM+04, LTWM+05, VSDB+06, WDKK96, BPQ+06]. These resources, together with the forthcoming computational and experimental methods will dramatically increase our understanding of the biological problems in transcription regulation and gene regulatory networks [PR01, PPS+06, PHB05, SHH+04, PAM+06, HPS+06, RRW+00].

1.1 Contributions and Organization

This thesis is structured as follows. Chapter 2 provides the necessary molecular biology background about genes, transcription and gene regulation. It also provides a brief introduction to the laboratory methods that were used while testing the predicted regulatory elements in living organisms.

Chapter 3 presents a novel mathematical model for the cis–regulatory elements and their conservation in evolution. The best conserved genomic region according to this model can be found by comparing two or more genomic DNA sequences with the algorithm given in this chapter. The chapter also presents the experimental evidence that shows the success of the model in locating cis–regulatory elements that work as expression enhancers in mammalian development. The chapter concludes by surveying the previous work on the field and comparing it to the newly introduced method.

Chapter 4 takes a closer look at the parameters of the cis–regulatory element model and develops methods for estimating good parameter values. It also experiments with small variations in the parameter values and analyzes their effect on the results.

Chapter 5 considers the statistical significance of the predicted regulatory elements and surveys the local sequence features that may confound the interpretation of the associated quality scores. The results provide a way of comparing the significance of the cis–regulatory element predictions from sequences with different length.

Chapter 6 shows how the cis–regulatory element predictions can be used to find the causative transcription factors for changes in the expression of a set of genes. This extends the now commonplace expression cluster analysis by providing putative regulators for the clusters of genes [ESBB98].

Chapter 7 describes the various interfaces to the software developed as part of this work. The chapter briefly introduces a simple point–and–click interface for interactive use and a scripting interface for the high performance computing system that allows distributing the regulatory element prediction jobs to a clustered set of computers. The various web interfaces for the precomputed cis–regulatory element predictions are also described here. The final Chapter 8 concludes the thesis and presents a few avenues for future work.

The thesis is centered around the system for locating putative cis–regulatory elements from mammalian genomes. This problem is approached from several viewpoints and the main contributions of this thesis are

- The mathematical model for the evolutionary conservation of cis–regulatory elements active during development. The model for the cis–regulatory element conservation in two species was developed together by the author and Professor Jussi Taipale. The success of the model was demonstrated by a series of laboratory experiments [HPS+06].

- An efficient dynamic programming algorithm for locating the well–conserved putative cis–regulatory elements from the genomic sequences of two or more species. The algorithm locates both the best conserved element and, in a particular sense, suboptimal elements.

- A procedure for estimating the parameter values for the cis–regulatory element conservation model. The procedure uses a custom scoring function for the set of parameters and optimizes the function by Simulated Annealing.

- The analysis of the stability of the cis–regulatory element predictions under varying parameter values and set of input motifs.

- A statistical model of the distribution of the cis–regulatory element conservation scores under the neutral evolution [Kim68]. This statistical model is conditional on the GC—content of the analyzed sequences and the parameter values used in the cis–regulatory element conservation model.

- A method for predicting DNA–binding transcription factors regulating a given set of genes. The regulatory interaction is inferred by analyzing the transcription factor binding sites conserved in the well–conserved cis–regulatory elements. The method utilizes the well–known independence tests on contingency tables but the construction of the tables is somewhat intricate.

- A software implementation called Enhancer Element Locator, or EEL, of the cis–regulatory element location method. The software is efficient and versatile enough to be used both in an interactive and batch–processing environment.

- A distributed scripting environment for the comparison of non–coding genomic sequences near human genes to the orthologous sequences in any species annotated in the Ensembl database [HAB+07].

- A database of putative cis–regulatory elements for regions around all human and mouse ortholog genes. This data is distributed over the Internet via a web–based query interface and via the Distributed Annotation System (DAS) [DJD+01].

- A web service for predicting the regulators for a given set of genes. The service protocol was colloquially defined in a meeting at the European Bioinformatics Institute on May 2007, chaired by Misha Kapushesky. The service was designed and the implementation was supervised by the author. The implementation was made by Jaakko Nyman.

The main results of this thesis, the conserved cis–regulatory module model and the associated software and databases, were published previously [HPS+06]. The description of the user–friendly point-and-click interface for the implemented software was also published in a separate article [PTU06]. Since this thesis includes the essence of these articles in full, they are not cited later in the text.

Chapter 2

Preliminaries

This chapter introduces some of the basic concepts of molecular biology that are used later in this work [CR05]. The description will ignore many known or suspected biological phenomena and concentrates on the features and assumptions that this work is based on.

The rest of this chapter is organized as follows. Section 2.1 briefly introduces genes and genomes in the detail that is used later in the work. Section 2.2 concentrates on the gene expression and its regulation. The final Section 2.3 introduces some of the laboratory methods that were used in verification of the predictions made with the conserved cis–regulatory module model given in Chapter 3.

2.1 Genes and the genome

The genetic information that is passed from parents to their offspring is encoded in the genome. Physically the genome is a small number, 46 in human, of long deoxyribonucleic acid (DNA) molecules called chromosomes. The DNA consists of four kinds of nucleotides distinguished by their bases, adenine, cytosine, guanine and thymine. The nucleotides are joined with a sugar backbone to a linear sequence forming a long helical polymer [WC53]. The DNA polymer naturally binds another DNA molecule with a complementary sequence of bases, thymine binding adenine and guanine binding cytosine. The ends of the polymer have unbound atoms named 5' and 3', hence the structure of a DNA molecule can be fully described by listing the nucleotide bases in sequence. The DNA sequence is conventionally described as a string of characters A, C, G, and T representing the bases from the 5' to the 3' end of the molecule.

In this work a gene is a section of the genome that can be transcribed to an RNA and codes for a protein. In molecular biology the gene is sometimes considered to include also the transcribed but non–protein–coding DNA sequences and the DNA sequences regulating the genes activity. In genetics, genes are defined differently as the basic hereditary unit. In concrete physical terms this hereditary unit can coincide with the genes of molecular biology or it can be something else such as a single nucleotide polymorphism.

The RNA is a polymer similar to DNA with the deoxyribose sugars replaced with ribose and thymine base replaced with uracil. The transcribed RNA contains the same sequence of bases as the DNA that was used as a template for the transcription, although in complementary bases [WC53].

For most mammalian genes the protein–coding parts are not continuous in the DNA sequence but there are large non–coding sections in the middle of the gene. The protein–coding parts of the gene are called exons and the non–coding parts are called introns. The newly transcribed primary RNA containing the whole gene is processed to messenger RNA by splicing off the introns leaving only the coding sequence which is later translated to a protein according to the genetic code [OJWM92].

The proteins are the workhorse of the cell. They take care of most of the activities of a living organism like maintaining the structure of the cells, catalyzing the metabolic reactions, transporting signals between the cells and fighting foreign objects like viruses and bacteria. The proteins are also responsible for the molecular machinery upholding life itself: copying the DNA, transcribing the RNA and translating the proteins. The time and place when and where a gene is turned into a functional protein is of utmost importance to the organism. That process is also regulated by the proteins with guidance of the non–coding parts of the genome.

2.2 Gene regulation

The transcription of a gene is called expression and the number of transcripts, i.e., the messenger RNA molecules, available at a given time is called its expression level. The expression level depends on the rates of transcription and RNA degradation. The genes, or more specifically their expression, is regulated by certain proteins, called transcription factors (TFs), according to guides provided by the DNA sequence.

The two parts of the regulatory machinery, the protein factors and the DNA elements, are called trans– and cis–acting elements respectively. The names rise from the viewpoint of the gene: the DNA elements are

'here' (lat. cis) on the same molecule and the transcription factors are 'there' (lat. trans) on the separate protein molecules.

The transcription factors fall into two different categories. Some factors are general in the sense that they are always needed for transcription of any gene while others are specific in the sense that they only enhance or repress the expression of their target genes.

The general transcription factors bind to the basal promoter sequence proximal to the transcription start site of the gene (See Figure 2.1). These factors form a basal transcription initiation complex and recruit the RNA polymerase to transcribe the gene.

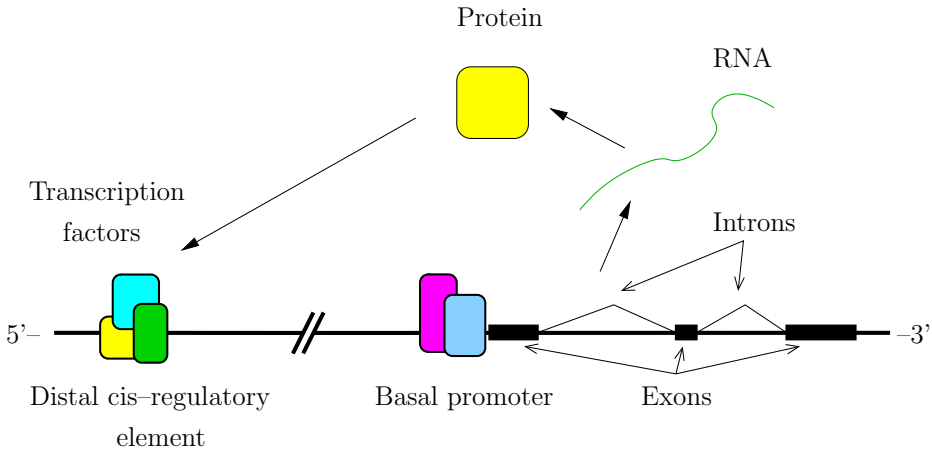

Figure 2.1: Functional elements on the genome.

The rate of transcription, how often the transcription is initiated, is regulated by the specific transcription factors that either promote or inhibit the transcription start. The specific transcription factors work by binding to the DNA in certain sites recognized by their sequence and further binding to each other and to the transcription initiation complex.

The transcription factor binding sites are short, typically 6–20 base pairs long, sequences of DNA that have above average affinity of binding the particular transcription factor. The binding sites tolerate a few nucleotide 'errors' so the binding site motifs are not exact sequences but more like distributions of DNA sequences.

The transcription factor binding sites, just like the factors themselves, come in two flavors. The sites binding the general factors are located on the basal promoter proximal to the gene. The sites binding the specific

factors are located on the cis–regulatory elements, such as enhancers, which may be located more distal to the gene [BRS81].

The specific factors can form complex logic gates by having their binding sites close to each other in the cis–regulatory element. The clustering of the binding sites is needed also to obtain enough affinity for interaction with the basal transcriptional machinery.

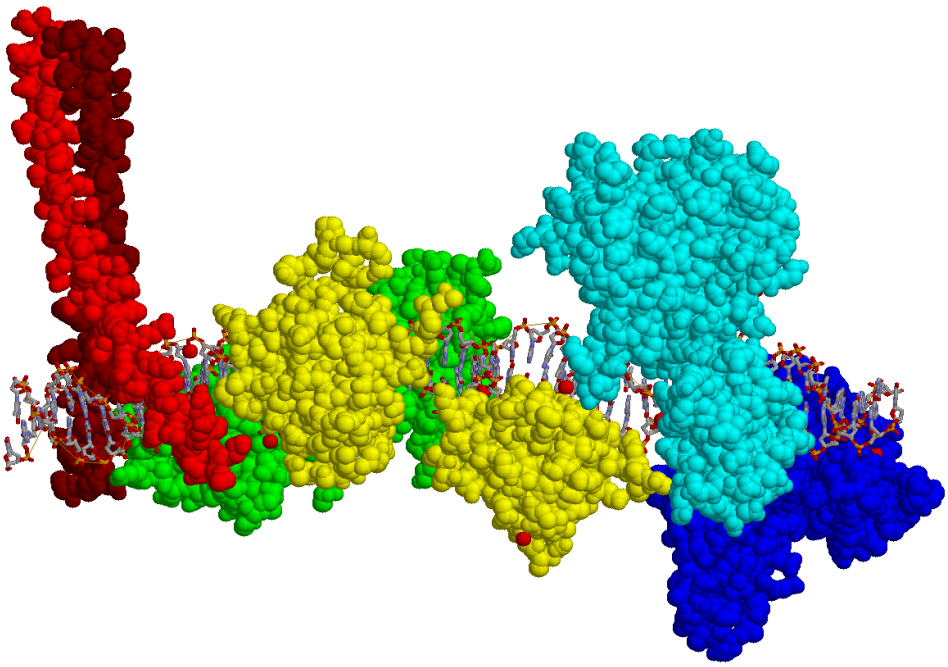

Figure 2.2: Structure of the *IFN–β* enhancer with bound transcription factors c–Jun, ATF–2, IRF–3A, IRF–7B, IRF–3C, IRF–7D, p50 and RelA [PMH07].

A special case of transcription factors binding a cis–regulatory element can be seen in Figure 2.2. The figure displays the molecular structure of a well–known cis–regulatory element, the enhancer of the gene Interferon β [PMH07]. Here the enhancer is short and the binding sites are close to each other and sometimes even overlapping. In qualitative agreement with the conserved cis–regulatory module model of Chapter 3, the eight transcription factors binding to the DNA interact mostly with the adjacent factors.

The cis–regulatory elements, together with the transcription factors binding them, are involved in several types of regulatory activities. They

can work as basal enhancers or repressors affecting gene expression in all conditions, they can be condition–specific working only in a certain developmental stage or a tissue or they can work as insulators, insulating a gene from a distal enhancer or a repressor. The repressors and insulators are difficult to test experimentally so in this work the main interest is on the cis–regulatory elements acting as enhancers.

2.3 Experimental laboratory methods

Chapter 3 includes experimental results obtained with methods that are not well known in the Bioinformatics community. Because of this it is beneficial to review these methods that provide results looking superficially similar but still test different phenomena. The two experimental protocols are the *in-situ hybridization* for visualizing gene expression and the *transgenic reporter assay* for analyzing enhancer activity.

The in–situ hybridization visualizes the tissue–specific expression of a particular gene in an organism, in this case in a mouse embryo [RB93, HAM$^+$95]. The process uses labeled RNA probes that recognize the messenger RNA of the target gene by sequence complementarity. The probes hybridize with the expressed messenger RNA and the label on the probe can be bound with an antibody. Finally the antibody is stained with a dye and the location of the targeted RNA becomes visible. The several steps of washing, hybridizing and staining in this protocol have many pitfalls and require three days of lab work.

The in–situ hybridized mouse embryos can be processed intact and viewed "whole mount" under the microscope provided they are harvested from early to mid–gestation; before 13 days after the fertilization. Figures 3.9A–F on Page 31 and Figures 3.10 on Page 32 display whole mount images of in–situ hybridized mouse embryos at 9.5–11.5 days after the fertilization.

The transgenic reporter assay tests the activity of a DNA sequence as an expression enhancer for a marker gene. First, the DNA sequence to be tested is cloned with the Polymerase Chain Reaction and inserted into a vector carrying a *LacZ*–gene and a minimal promoter, which is not alone able to express the *LacZ*–gene. The enhancer–LacZ fusion construct is microinjected into the pronucleus of a single–celled mouse zygote which is transferred to the mouse uterus and grown up to the developmental stage of interest before harvesting [NGVB06a, NGVB06b]. The product of *LacZ* expression, the β–galactosidase protein, is then stained and the result is viewed under the microscope [Nag03, Kob07].

The result of the β–galactosidase staining can be seen in Figures 3.9G–I, 3.11 and 3.12 on Pages 31, 34 and 35. Figures 3.11C–E and 3.12D–G are images of sectioned embryos providing a more detailed view of the expression [NGVB07].

As Figure 3.9 shows, the two methods provide similar images even though their interpretation is crucially different. The in–situ protocol visualizes the native expression of the target gene, which may, or may not, result from the predicted enhancer. The transgenic assay visualizes the expression driven by the predicted enhancer, which may, or may not, regulate the expected target gene.

The images should be interpreted by comparing them with each other and the expected location of the expression. The genes in Figure 3.9 for example are assumed to be regulated by the SHH signaling protein produced in locations seen in Figure 3.9A. The genes in Figures 3.9B–E show expression in a subset of the tissues expressing SHH. The gene *GPC3* in Figure 3.9F shows expression in the sclerotome, a tissue induced by Sonic Hedgehog signaling [FPC$^+$95]. The transgenic embryo in Figure 3.9I also shows marker expression in the sclerotome. Together these two figures suggest that the computationally located cis–regulatory element acts as a sclerotome–specific transcriptional enhancer of the *GPC3* gene at embryonic day 12.5.

Chapter 3

Computational model of conserved cis–regulatory modules

The main contribution of this work is the novel *enhancer element locator* (EEL) method for locating putative cis–regulatory elements from long non–coding DNA sequences of vertebrates and other higher eukaryotes [CIG+70]. The method is based on the *conserved cis–regulatory module model* which evaluates the evolutionary conservation of cis–regulatory element–like sequences of transcription factor binding sites.

The rest of this chapter is organized as follows. The conserved cis–regulatory module model is described in Section 3.1. Section 3.2 discusses the subproblem of finding the individual transcription factor binding sites and estimating their affinity to bind the transcription factors. Section 3.3 provides an algorithm for finding the parts of the binding site sequences that best fit the conserved cis–regulatory module model. Section 3.4 discusses the computational complexity of the conserved cis–regulatory module–finding algorithm. Section 3.5 presents the laboratory results verifying the method. Discussion about previous work on computational cis–regulatory element finding is postponed to Section 3.6 so the other methods can be compared with the newly introduced model. The final Section 3.7 concludes the chapter.

3.1 Conserved cis–regulatory module model

The enhancer element locator tool is provided with evolutionarily related DNA sequences G and G' and the *binding site motifs* for a set of *transcription factors* $\mathcal{F}$. The DNA sequences are first scanned with the motifs and the produced sequences of binding sites are compared for agreement

according to the conserved cis–regulatory module model. The output from the conserved cis–regulatory module–finding algorithm is a list of subsequences of the input DNA, sorted according to the fit to the conserved cis–regulatory module model and annotated with a quality score and with the conserved transcription factor binding sites.

For our purposes a *binding site* is a quadruplet $s = (f, p, q, W)$ where $f \in \mathcal{F}$ names the binding transcription factor (orthologous factors in the different species are considered equal), p is the start and q is the end position of the binding site on the DNA sequence, and W is the binding affinity (the binding force) of the transcription factor f to the indicated DNA sequence. The binding sites in the DNA sequence G are given in the sequence $\mathrm{S} = (s_i)_{i=1}^{\mathrm{L}}$ in ascending order of the start positions (and similarly in $\mathrm{S}' = (s'_j)_{j=1}^{\mathrm{L}'}$ for DNA sequence G'). The locations and the affinities of the binding sites are found with the methods of Section 3.2 after which the underlying DNA sequences are no longer used and all information about the cis–regulatory modules are in S and S′.

A *cis–regulatory module* (CRM) E is a sequence of binding sites in ascending order of their start positions that do not overlap and lie within 1000bp of their adjacent binding sites. Two CRMs $E = (e_i)_{i=1}^{l}$ and $E' = (e'_j)_{j=1}^{l}$ from different sequences are *conserved* if they bind the same sequence of transcription factors. A binding site is conserved if it is part of a conserved CRM.

The conserved CRM model has two components: some *bonus* is given for each binding site in the conserved CRM and some *penalty* is given for the distance between the adjacent conserved binding sites and for the difference of these distances in the two sequences. The bonus for binding sites e_i and e'_i with affinities W_i and W'_i is just

$$\lambda(W_i + W'_i) \tag{3.1}$$

where λ is a weighting parameter.

The penalty is defined with the notations in Figure 3.1. Let (s, s') and (t, t') be pairs of conserved binding sites adjacent in the conserved CRMs. The distance between s and t is $x = p_t - q_s - 1$ base pairs and the distance between s' and t' is $x' = p_{t'} - q_{s'} - 1$ base pairs. The angular distance of the binding sites on the DNA helices is different in the two species by $\Delta\phi$ radians. In this setting, the penalty score is

$$\mathrm{F}(x, x') = \mu\frac{x + x'}{2} + \nu\frac{(x - x')^2}{x + x'} + \xi\frac{(\Delta\phi)^2}{x + x'} \tag{3.2}$$

where μ, ν and ξ are the weighting parameters. In the special case when the binding sites touch each other, $x = x' = 0$, the value $\mathrm{F}(0, 0) = 0$.

The twist angle $\Delta\phi = (x - x')\frac{2\pi}{10.4} - 2k\pi$ where k is an integer such that $\Delta\phi$ will be the minimum distance to full rotation, i.e., $-\pi \leq \Delta\phi < \pi$. This models the regular B–DNA having one rotation for every 10.4 base pairs on average [WDT+80]. Motivation for the penalty score in Equation (3.2) is provided later.

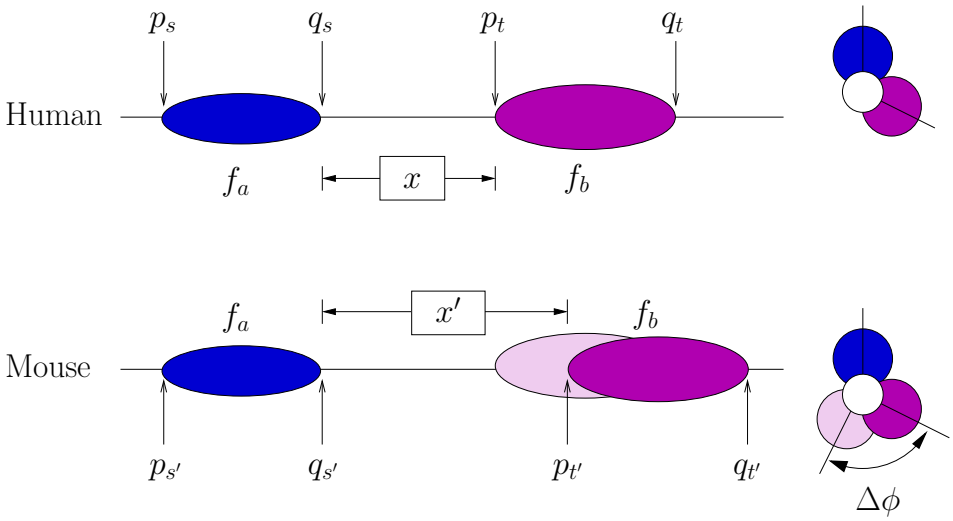

Figure 3.1: Conserved cis–regulatory module model. The distance between the binding sites of transcription factors f_a and f_b is $x = p_t - q_s - 1$ nucleotides on the human sequence and $x' = p_{t'} - q_{s'} - 1$ nucleotides on the mouse sequence. On the DNA helices, the angle between the binding sites for the two factors is different in human and mouse by $\Delta\phi$ radians.

The *enhancer element locator score* $\mathrm{S}(E, E')$ of the conserved cis–regulatory modules E and E' is the sum of the bonuses from Equation (3.1) subtracted with the penalties from Equation (3.2). The enhancer element locator score for the conserved cis–regulatory modules E and E' is computed as

$$\mathrm{S}(E, E') = \sum_{i=1}^{l} \lambda(W_i + W'_i) - \sum_{i=2}^{l} \mathrm{F}(p_i - q_{i-1} - 1, p'_i - q'_{i-1} - 1). \quad (3.3)$$

For a pair of DNA sequences y the enhancer element locator score $EEL(y)$ is the maximum enhancer element locator score for cis–regulatory modules conserved in the sequences y.

The main idea behind the conserved cis–regulatory module model is that transcription factors binding to adjacent sites are likely to cooper-

ate with each other. This cooperation takes the form of protein–protein interactions [JT96]. The DNA binding transcription factors and other regulatory proteins interact via the surface formed by the DNA bound factors. The physical constraints for this interaction sets the maximum distance of 1000bp between the adjacent binding sites in a CRM.

The first term $\mu\frac{x+x'}{2}$ of Equation (3.2) penalizes for loose clustering of the binding sites proportionally to the distance between the adjacent sites. This term keeps the binding sites clustered so the bound transcription factors can properly interact. It is likely that the optimal distance between the adjacent transcription factors is greater than zero bases but the optimal distance is unknown and is probably different for different pairs of factors so we see this linear scoring as reasonable approximation of the true clustering affinity. The maximum distance of 1kbp between adjacent binding sites is more than what is biologically expected but it is chosen large to account for the uncertainties in the binding site predictions and in the mammalian enhancer structure [LT03].

The last two terms of Equation (3.2) model the energy needed to compress and twist the DNA helices to a common conformation in the two species. The second term $\nu\frac{(x-x')^2}{x+x'}$ of Equation (3.2) models the energy needed to compress the two DNA molecules to a common length [BSG$^+$03]. The formula arises from the energy needed to displace a spring with spring constant $\nu/(x+x')$ the amount of $|x-x'|$. The final term $\xi\frac{(\Delta\phi)^2}{x+x'}$ models the energy needed to twist the DNA helices so the adjacent transcription factors can reach a common 3D structure on both DNAs.

Any change in the length of DNA between two transcription factors that bind near each other is expected to change the protein–protein interaction surface, and thus result in changes in the avidity of other proteins of the transcription regulatory machinery [KA04, AK05]. When the length of the DNA between adjacent binding sites has changed, the protein–protein interactions are better accommodated when the transcription factors are farther apart on the DNA. The longer distance allows the DNA to bend and compensate for the difference. Therefore the penalty for distance change becomes smaller with increase in average distance between adjacent binding sites. The model of penalizing for the DNA length changes is supported by the observation that long insertions and deletions are strongly disfavored in the evolution of cis–regulatory elements [CCB$^+$05].

The conserved CRM model does not include the distance of the module from the transcription start site(s) of the target gene. This is required by the traditional definition of an enhancer which states that the enhancer regulates gene expression regardless of its position with respect to the regulated

gene [BRS81] and the same assumption is used when testing the enhancer activity with the transgenic assay.

Notable restriction for the applicability of the conserved cis–regulatory module model is the requirement of conservation. The model does not help in finding novel or non–conserved cis–regulatory modules. Similarly, the conserved cis–regulatory module model does not handle binding site turnover when a binding site is replaced with a novel site in different position. The proposed model is most fit for finding conserved *enhanceosomes* with highly specific structure [AK05].

3.1.1 Conservation in multiple species

A sum–of–pairs approach extends the conserved cis–regulatory module model to consider multiple evolutionarily related sequences. Let us call the set $\mathcal{E} = \{E_i\}$ of k cis–regulatory modules conserved, if E_i and E_j are conserved for all $i \neq j$. Now the sum–of–pairs enhancer element locator score for the set $\mathcal{E}$ of conserved cis–regulatory modules is

$$\mathrm{S}(\mathcal{E}) = \sum_{i=2}^{k} \sum_{j<i} \mathrm{S}(E_i, E_j). \tag{3.4}$$

The requirement of pairwise conservation of the modules forces all binding sites to be conserved in all sequences. This might be an undesirable feature in some cases, when the module is fully conserved in only a subset of the species. Anyway, the conserved cis–regulatory module model forbids insertions or deletions of binding sites in a conserved CRM. Insertions and deletions would change the dependencies between the adjacent binding sites in an unpredictable manner because of the dramatic change in the 3D structure of the assembled transcription factor complex.

3.2 Affinity of Protein–DNA binding

The enhancer element locator model uses two sequences of transcription factor binding sites annotated with DNA location, bound factor and binding affinity. Probabilistic binding motif models provide reasonable estimates for the binding affinity when the binding probability is compared to assumed background sequence probability. In the following we use the *positional frequency model* which is simple, commonly used and the only probabilistic binding motif model with abundant data available in public databases [Sto00, VSDB$^+$06, BEFK03, SS07].

Let G denote a DNA sequence of length n, and let $\{M_f | f \in \mathcal{F}\}$ be the set of *binding motifs* for the transcription factors $\mathcal{F}$. If the transcription factor f has binding sites of length m_f then the binding motif M_f is a *positional frequency matrix* defining a probability distribution for DNA sequences of length m_f. When each nucleotide exerts independent binding affinity towards the factor, the binding motif can be seen as m_f–dimensional multinomial distribution, where the multinomial has four possible outcomes, one for each kind of nucleotide [Sto00].

The binding sites of a transcription factor f are located from the DNA by matching the positional frequency matrix M_f to the DNA sequence G. The affinity of the transcription factor f to each substring $g = g_1 g_2 \dots g_{m_f}$ of length m_f is estimated with the positional weight matrix by

$$W(g) = \log_2 \frac{\mathrm{P}_{M_f}(g)}{\mathrm{P}_b(g)} \tag{3.5}$$

where $\mathrm{P}_{M_f}(g)$ is the likelihood of sequence g being generated by the distribution M_f, and $\mathrm{P}_b(g)$ is the likelihood of g being generated by the *background nucleotide distribution* [HB86, SF98, Sto00]. The log likelihood ratio of Equation (3.5) is considered as a proxy for the actual binding affinity and we do not associate it with any statistical interpretation e.g. for hypothesis testing. Because of the affinity interpretation it is natural that the values $W(g)$ can be larger for transcription factors with longer binding sites.

The background nucleotide distribution $\mathrm{P}_b(\cdot)$ is commonly assumed as Markov of some fairly low degree. Markov background models are known to improve the ab initio motif finding and they are used to weight the binding site detection away from low complexity regions, such as AT–repeats towards more rare CpG–containing sequences [TLM$^+$01]. The likelihood of sequence g in the k:th order Markov background is

$$\mathrm{P}_k(g) = \prod_{i=1}^{m_f} \mathrm{P}(g_i | g_{i-1} \dots g_{i-k}). \tag{3.6}$$

Since g is a part of the long sequence G, the non–positive indexes of g_i refer to the genomic nucleotides preceding the subsequence g in G. In the beginning of the genomic sequence G, the order of the Markov model is decreased to fit the available sequence information.

An important special case for the background distribution is the zeroth order Markov which stands for the independent and identical distribution of the nucleotides. Let the likelihood of nucleotide 'A' in the zeroth order Markov background model be p_A and the number of As in the sequence g

be A_g (similarly for C, G and T). Now the likelihood of sequence g in zeroth order Markov model is

$$\mathrm{P}_0(g) = \prod_{i=1}^{m_f} p_{g_i} = p_A^{A_g} \cdot p_C^{C_g} \cdot p_G^{G_g} \cdot p_T^{T_g}. \tag{3.7}$$

This distribution has the good property that the null probabilities p_* can be included straight into the log likelihood ratios in Equation (3.5) so the weighting scheme can be described with a single weight matrix and the total score, or affinity, can be computed with a simple sum over the log likelihood ratio matrix. Sophisticated weight matrix–matching techniques can obtain small speed improvements over the trivial $\mathrm{O}(nm_f)$ matching algorithm [BSH+04, PRU07]. The zeroth order Markov background assumption is usually made to obtain analytical results e.g. about optimality of the positional weighting or exact p–value thresholds [Sto00, HLS94, Sta89].

The parameter values for the background distribution can be estimated from non–regulatory sample sequences. The values can be estimated from long genomic sequence, for example from the whole human genome since only about 4% of it is under purifying selection and the rest is likely non–coding and non–regulatory [LPH06]. The statistical significance of individual binding sites could theoretically be improved by using a sequence–specific background distribution, which takes into account such things as the local GC–content variation (isochores), but that would provide a different estimate of the binding affinities for same binding sequence depending on the genomic environment [Ber89]. This is not desirable in our application since we interpret the score in Equation (3.5) as the actual binding affinity which should not depend on any other feature but the binding motif and the DNA sequence [HB86, SF98].

High sequence–specific affinity allows us to provide a discrete list of binding sites on the DNA for a given transcription factor. The low affinity protein–DNA binding is governed by non–sequence specific forces which counters the assumption about the cumulative effect of weak binding sites [MQ07, RKMV07]. We have found that using binding sites with affinity greater than 9 provides good results with the enhancer element locator.

3.3 Finding the conserved cis–regulatory modules

The conserved cis–regulatory module with the maximal enhancer element locator score can be found with dynamic programming similar to the Smith–Waterman local alignment algorithm [SW81]. The *binding site alignment*

algorithm aligns the two sequences of binding sites and traces the subsequences that best fit the conserved CRM model.

The algorithm takes as inputs two binding site sequences S and S′ of length L and L′. The sequences are *aligned* by filling a matrix D of size L×L′. The cell $D_{i,j}$ of the matrix holds the enhancer element locator score for the highest scoring conserved cis–regulatory module whose last conserved pair of binding sites are s_i from the sequence S and s'_j from the sequence S′.

The dynamic programming matrix is computed with the recursion

$$D_{i,j} = \begin{cases} \lambda w_{ij} + \max\limits_{\substack{0<p_i-q_k<1000 \\ 0<p'_j-q'_l<1000}} \left\{0, D_{k,l} - F\left(p_i - q_k - 1, p'_j - q'_l - 1\right)\right\} & , \text{ if } f_i = f'_j \\ -\infty & , \text{otherwise.} \end{cases} \tag{3.8}$$

where $w_{ij} = (W_i + W'_j)$. The matrix does not need initialization because of inclusion of the finite term λw_{ij} to the maximization.

The score of the best conserved cis–regulatory module is ultimately found by finding $\max D_{i,j}$. The structure of the conserved cis–regulatory module is found with a *backtrace* $bt(i_1, i'_1) = (i_j, i'_j)_{j=1}^{l}$ from a finite valued cell D_{i_1,i'_1}. The backtrace from (i_1, i'_1) is a list of (i_1, i'_1) and the indexes for the cells visited by following the *backtrace pointers* from the cell (i_j, i'_j), for $j = 1, \ldots$. The backtrace pointer in the cell (i_j, i'_j) points to the cell (i_{j+1}, i'_{j+1}) that maximized Equation (3.8). The backtrace ends if the backtrace pointer does not exist i.e., $D_{i,j} = \lambda w_{ij}$.

We now proceed to show that the described procedure uncovers the CRMs with the maximal enhancer element locator score.

Theorem 3.1. *The backtrace* $bt(i_1, i'_1) = (i_j, i'_j)_{j=1}^{l}$ *defines a pair of conserved cis–regulatory modules* $(s_{i_j})_{j=l}^{1}$ *and* $(s'_{i'_j})_{j=l}^{1}$.

Proof.

The backtrace from a finite valued cell (i_1, i'_1) can take one of two cases. First, if $D_{i_1,i'_1} = \lambda w_{i_1 i'_1}$, then the backtrace from (i_1, i'_1) defines the CRMs (s_{i_1}) and $(s'_{i'_1})$ of length one. Second, the backtrace of length $l > 1$ from (i_1, i'_1) extends the CRMs of length $l - 1$ defined by the backtrace from (i_2, i'_2). These shorter CRMs are extended with binding sites within non–negative distance of at most 1000bp from s_{i_2} and $s'_{i'_2}$ resulting in CRMs. The CRMs are conserved since all pairs $(s_{i_j}, s'_{i'_j})$ of binding sites contribute a finite score in Equation (3.8) and hence bind the same transcription factors.□

Because of Theorem 3.1, we can set $\mathrm{S}(bt(i,i'))$ to be the enhancer element locator score for the CRMs defined by the backtrace $bt(i,i')$. We can state a theorem

Theorem 3.2. *For backtrace* $bt(i_1,i'_1)=(i_j,i'_j)_{j=1}^{l}$*, the enhancer element locator score* $S(bt(i_1,i'_1))=\mathrm{D}_{i_1,i'_1}$.

Proof. If the CRMs are of length one, $l=1$, then $\mathrm{S}(bt(i_1,i'_1))=\lambda w_{i_1,i'_1}=\mathrm{D}_{i_1,i'_1}$. Assume now that the Theorem holds for $l-1$. If the CRMs are of length $l>1$, then

$$\begin{aligned}\mathrm{S}(bt(i_1,i'_1)) &= \sum_{j=1}^{l}\lambda w_{i_j i'_j}-\sum_{j=2}^{l}\mathrm{F}(p_{i_j}-q_{i_j-1}-1,p'_{i'_j}-q'_{i'_j-1}-1)\\ &= \mathrm{S}(bt(i_2,i'_2))+\lambda w_{i_1 i'_1}-\mathrm{F}(p_{i_1}-q_{i_1-1}-1,p'_{i'_1}-q'_{i'_1-1}-1)\\ &= \mathrm{D}_{i_1,i'_1},\end{aligned}$$

because $bt(i_2,i'_2)$ is of length $l-1$. □

Now we can show the correctness of the binding site alignment algorithm.

Theorem 3.3. *The cis–regulatory module conserved in* S *and* S′ *that has the maximal enhancer element locator score, is defined by* $bt(\arg\max_{i,i'} D_{i,i'})$.

Proof. The values $\mathrm{D}_{i,j}$ stand for the maximal enhancer element locator score for CRMs ending in sites s_i and s'_j, because of Theorem 3.2 and the fact that the recursion (3.8) always takes the maximum value. The maximal enhancer element locator score is thus given by $\max \mathrm{D}_{i,j}$. The associated CRMs are defined by $bt(\arg\max_{i,i'} D_{i,i'})$ because of Theorem 3.2. □

3.3.1 Sub-optimal solutions

The biologist using a cis–regulatory element finding tool is not interested only in the absolutely highest scoring element within his pair of sequences but he most likely also wants to see a few lower scoring predictions. What exactly is meant by "lower scoring prediction" is not obvious in this context. The main question is about removing the effect of the higher scoring elements.

If one simply takes the backtrace from the second–highest scoring cell in the matrix D, one almost certainly ends up with the same CRMs as previously, just with the last, or few last binding sites changed. This is probably not what one would want to present to the end user.

Another option is to reject all lower–scoring CRMs that backtrace to the same first pair of binding sites that are included in higher–scoring CRM.

This approach would give separate CRMs for each rank. The downside of this is the rejection of a high number of feasible CRMs, some of which might share only a single binding site with any of the previously seen CRMs.

Our concept of suboptimal alignments follows the work of Waterman and Eggert [WE87]. In their method, the suboptimal alignments are the highest–scoring local alignments, given they do not contain previously aligned pairs of characters (their work was related to the Smith–Waterman type of local biopolymer sequence alignment).

The suboptimal alignments are computed by similar recursion formula as in Equation (3.8). The only difference is the addition of a branch

$$\mathrm{D}_{i,j} = -\infty \quad \text{, if i and j are aligned in some higher–scoring alignment,} \tag{3.9}$$

which forces the algorithm to use alternate paths instead of the previously aligned pairs of sites.

The Waterman–Eggert algorithm does not recompute the whole matrix D for each suboptimal alignment. The effect of the higher–scoring conserved CRMs is removed from matrix D by recomputing only the part that is affected by the high–scoring conserved CRMs. The low index limit for the affected part of matrix D is equal to the indexes of the first aligned pair of binding sites on the alignment and the effect can last up to 1000bp, the distance of the maximization in Equation (3.8), from the last sites whose alignment score $\mathrm{D}_{i,j}$ is changed. In the worst case, the Waterman–Eggert type of suboptimal alignment algorithm has to recompute the whole matrix for each suboptimal alignment recovered but that worst case is unlikely in practice.

3.3.2 Sparse implementation

Equation (3.8) has a distinct property that immediately suggests implementation of the dynamic programming in a sparse matrix [EGGI92a, EGGI92b, HS77]. It is worthy to consider here the actual implementation of the dynamic programming since the computational complexity and the overall efficiency of the cis–regulatory module finding method depends on it heavily.

A compressed row implementation of the sparse matrix D allows a time– and space–efficient algorithm that discards the negative infinite values of the matrix without any extra effort. The main idea is to implement the dynamic programming over *inverted index* of the input sequence S' [Knu98, Moo21, ZM06]. An inverted index is a set of lists, each list containing the occurrence locations of a particular word or a character in a string. In our case the lists contain the occurrences of the binding sites for a particular transcription factor in the input sequences.

Definition 3.1 (Inverted index)**.** An inverted index for a binding site sequence $\mathrm{S} = (s_i = (f_i, p_i, q_i, W_i))_{i=1}^{\mathrm{L}}$ is a mapping $\pi : \mathcal{F} \times \mathbb{N} \mapsto \mathbb{N}$ where

$$\pi(f, j) = k,$$

such that s_k is the j:th binding site of transcription factor f in sequence S.

A schematic representation of the inverted index is displayed in Figure 3.2. With the inverted index, we can find all binding sites of a particular transcription factor in time linear in the number of those sites. Given the inverted indexes of two transcription factor binding site sequences, we can take a Cartesian product between the lists defined by a transcription factor f which provides the locations of the pairs of the sites in the two sequences that bind f. These are exactly the locations of D that have a finite value and propose aligning two binding sites of the transcription factor f. All finite value cells of D are covered by taking the Cartesian products for the lists of all factors in $\mathcal{F}$.

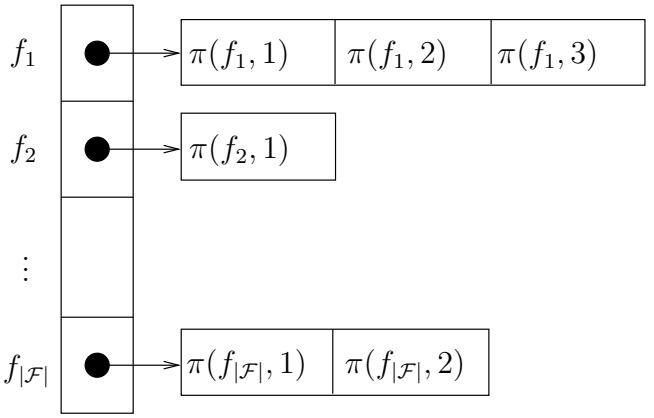

Figure 3.2: Inverted index for transcription factor binding site sequence

The Cartesian product of the inverted indexes can be used as the basis of matrix D. All finite valued cells of matrix D used to align sequences S and S′ can be represented with $\mathrm{D}_{i,\pi'(f_i,j)}$ for some i and j. The structure of the described sparse matrix is shown in Figure 3.3.

A caveat in using the sparse matrix of Figure 3.3 in the dynamic programming is in quickly finding the cells used in the maximization. This can be achieved by maintaining a pointer $prevI$ to sequence S and a set of pointers $prev[f]$ to the inverted index of S′. These pointers point no further than the first site at most 1000bp before the site currently being aligned. This construction allows finding the sites needed in the maximization branch

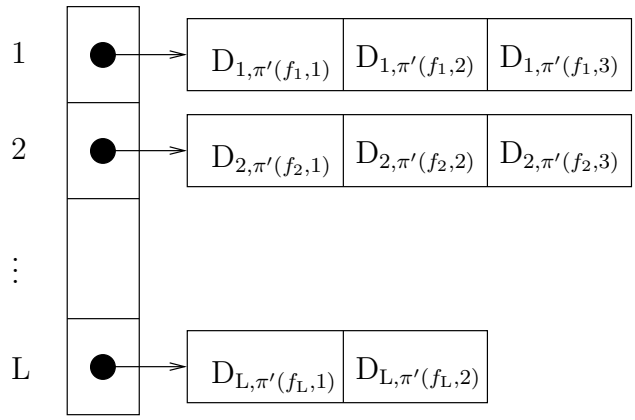

Figure 3.3: Sparse matrix for two dimensional dynamic programming

of Equation (3.8) in amortized constant time. The pseudocode for the whole binding site alignment algorithm is provided as Algorithm 1.

The main idea of the algorithm is to choose one sequence as the base, which defines the transcription factor whose binding sites will be aligned next, and to use the inverted index of the other sequence to locate the corresponding sites quickly. This algorithm can be easily extended to multiple alignment of Equation (3.4) just by using the inverted indexes of the additional sequences. A sample of a sparse dynamic programming matrix for three sequences is given in Figure 3.4. The actual implementation of the arbitrary–dimensional dynamic programming is somewhat tedious because of the long chains of pointers to the matrix values and the maintenance of several layers of *prev* pointers.

3.4 Computational complexity

In the worst case for the binding site alignment algorithm, all transcription factor binding motifs match both DNA sequences of length n in all locations. This would result in binding site sequences of length

$$\mathrm{L} = \sum_{f \in \mathcal{F}} (n - m_f + 1) \approx n|\mathcal{F}|. \tag{3.10}$$

In this case a trivial implementation of Equation (3.8) evaluates the maximization branch almost n^2 times. The maximization itself looks for $999^2|\mathcal{F}|^2$ cells in the matrix in worst case and in average case for large L. Together, this results in worst case time complexity of $\mathrm{O}\left(|\mathcal{F}|^4 n^2\right)$ and high constant

Algorithm 1 Binding site alignment algorithm

Require: Transcription factor binding site sequences S and S′ and the inverted index π' of the latter.

Ensure: Matrix M, such that $M[i,j] = \mathrm{D}_{i,\pi'(f_i,j)}$.

$prevI \leftarrow 1$
for $i = 1, \ldots, \mathrm{L}$ **do** {Over S}
  while $p_i - q_{prevI} \geq 1000$ **do** {Skip distant sites}
    $prevI \leftarrow prevI + 1$
  end while
  $prev[f] \leftarrow 1$ **for all** $f \in \mathcal{F}$
  $j \leftarrow 1$; $rj \leftarrow \pi'(f_i, j)$
  while $rj < \mathrm{L}'$ **do** {Over π'}
    $pi \leftarrow prevI$; $head \leftarrow 0$
    while $p_i - q_{pi} > 0$ **do** {Maximization over S}
      while $p_{rj} - q_{\pi'(f_{pi}, prev[f_{pi}])} \geq 1000$ **do** {Skip distant sites}
        $prev[f_{pi}] \leftarrow prev[f_{pi}] + 1$
      end while
      $pj \leftarrow prev[f_{pi}]$
      while $p_{rj} - q_{\pi'(f_{pi}, pj)} > 0$ **do** {Maximization over π'}
        $extend \leftarrow M[pi, pj] - \mathrm{F}\left(p_i - q_{pi} - 1, p_{rj} - q_{\pi'(f_{pi}, pj)} - 1\right)$
        $head \leftarrow \max(head, extend)$
        $pj \leftarrow pj + 1$
      end while
      $pi \leftarrow pi + 1$
    end while
    $M[i,j] \leftarrow \lambda(w_i + w_{rj}) + head$
    $j \leftarrow j + 1$; $rj \leftarrow \pi'(f_i, j)$
  end while
end for

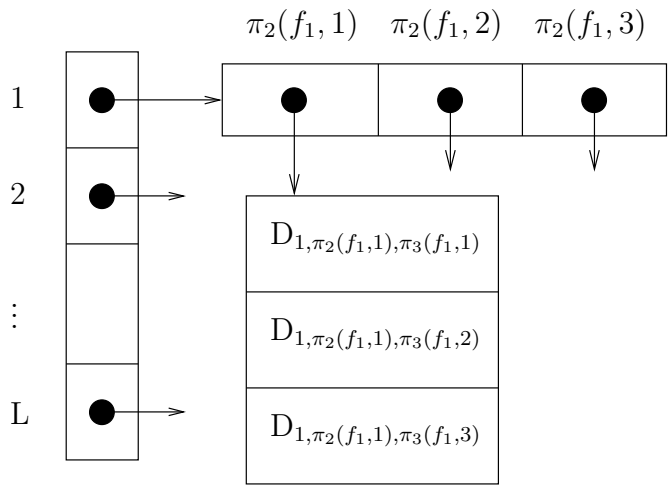

Figure 3.4: Sparse matrix for three–dimensional dynamic programming of sequences S_1, S_2 and S_3 with inverted indexes π_1, π_2 and π_3.

factors for the need of referencing more than one million cells of matrix D for computing a single additional value. Fortunately the sparse implementation presented as Algorithm 1 brings the computational complexity to a usable realm for practical use cases.

Since Equation (3.8) takes the negative infinity branch when factors f_i and f'_j differ, one can use the sparse representation of matrix D without loss of information. The missing cells of D can also be omitted from the maximization, because the default value of the matrix cells is negative infinite and the maximization on Equation (3.8) always includes the finite value λw_{ij}. By maintaining the *prev* pointers, the dynamic programming does not need random access to matrix D hence Algorithm 1 does not incur significant time penalty for sparseness. Updating the *prev* pointers takes an amortized constant time since they are increased at most as many times as there are cells to be filled. Algorithm 1 reduces the worst–case time complexity to $O(n^2)$ since it automatically disregards the cells standing for non–matching transcription factors.

In practical usage, the worst–case bound for the complexity is not extremely important. Because of the thresholding, all transcription factor binding motifs do not match all positions of the input DNA sequences. This alone sets the time complexity for the trivial algorithm to $O(L^2)$ for sequences of L sites, where $L << |\mathcal{F}|n$ for all reasonable thresholds. The number of binding sites L can be controlled with the affinity cutoff threshold which can be chosen, e.g., for each factor separately so that

only expected percentage p of the nucleotides in the background sequence match the binding motif [Sta89, BSH+04, BEKF05]. This kind of thresholds provide the expected number of sites $\mathbf{EL} = pn|\mathcal{F}|$ and the expected time complexity of $\mathrm{O}\left(p^4|\mathcal{F}|^4n^2\right)$ for the trivial implementation. This is an improvement since $p < 1$.

The expected time and space complexity of the sparse matrix method depends on the number of cells that need to be filled. Since we need to fill only the cells corresponding to the sites of the same transcription factor on both sequences, the number of filled cells equal

$$\sum_{i=1}^{\mathrm{L}} \left|\left\{s'_j | f'_j = f_i, s'_j \in S'\right\}\right| \tag{3.11}$$

which is the same as the size of the Cartesian product of the inverted indexes computed as

$$\sum_{f \in \mathcal{F}} |\pi(f, \cdot)| \cdot |\pi'(f, \cdot)|. \tag{3.12}$$

If all transcription factors have an equally distributed number of binding sites in both sequences, each term in Formula (3.11) is expected $\frac{\mathrm{L}}{|\mathcal{F}|}$. This results in $\mathrm{O}(\frac{\mathrm{L}^2}{|\mathcal{F}|})$ cells filled in the sparse matrix. If the cutoff thresholds have been selected such that each transcription factor matches proportion p of the nucleotides, one has to fill $\sum_{\mathcal{F}}(pn)^2 = p^2n^2|\mathcal{F}|$ cells of the matrix and filling each takes time $\sum_{\mathcal{F}}(999p)^2 = \mathrm{O}(p^2|\mathcal{F}|)$. In total the expected time complexity of the sparse implementation of recursion (3.8) equals $\mathrm{O}(p^4n^2|\mathcal{F}|^2)$. The matrix sparseness also speeds up the constant factors in the maximization step since the negative infinite cells are automatically skipped.

3.5 Experimental Results

The conserved cis–regulatory module model is implemented in the enhancer element locator tool. Our group and others have used the software successfully in real biological settings [VJM+07]. EEL recovers known cis–regulatory elements from fly sequences and the application of EEL to non–coding regions around all human and mouse ortholog genes reveals several genes whose expression coincide with the known expression patterns of their predicted regulators. EEL is also able to find *in vivo* active enhancer elements near genes known to be regulated in a particular distal enhancer element dependent fashion.

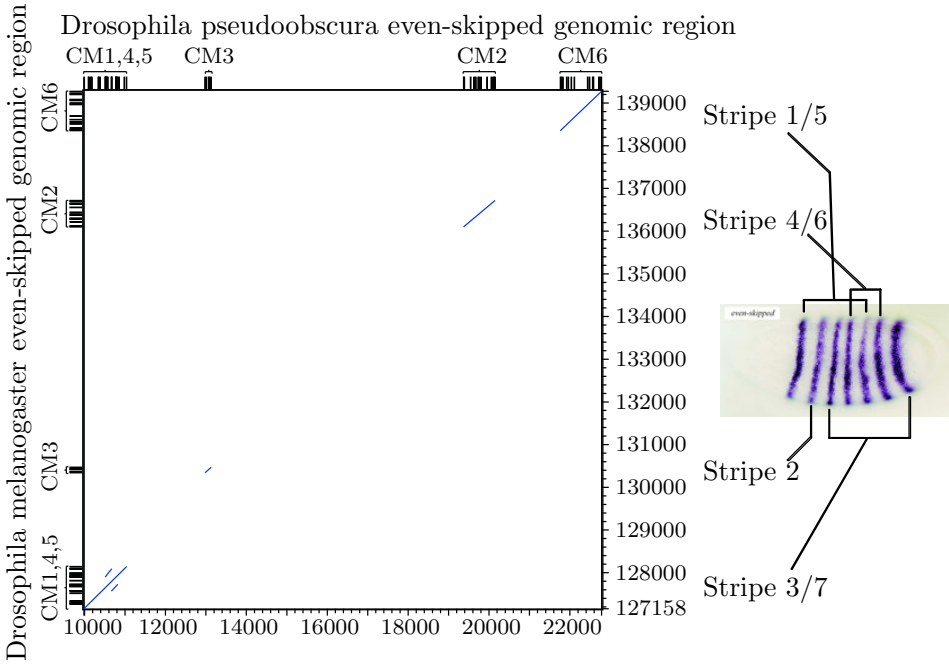

Figure 3.5: Enhancers regulating the drosophila *eve*–gene.

3.5.1 Known invertebrate elements

The first sanity check for the method was the recovery of the known enhancer elements from the well–studied model of early body patterning of the Drosophila embryo and specifically, the stripe patterning of the even–skipped gene *eve* [BNP+02].

Eve is one of the genes that define the anterior–posterior (head–tail) body patterning of the Drosophila fruit flies. In early embryonic development it is expressed in seven well–defined stripes that are separated by regions lacking *eve* expression. Decades of research have revealed that the *eve* expression is regulated by concentration gradients of five transcription factors that work through four distinct cis–regulatory elements [MS86, SBL96, SBL92].

We tested EEL by applying it to the genomic DNA around the *eve* gene in Drosophila Melanogaster and Drosophila Pseudoobscura, two highly divergent flies that have gained an estimated 1.79 synonymous substitutions per site since their divergence some $25 - 55$ million years ago [ACH+00, RLB+05] (as compared to 0.45 substitutions per site between human and mouse [WLTB+02]). The analysis used parameters $\lambda = 2$, $\mu = 0.5$, $\nu = 1$

and $\xi = 1$ and the five binding motifs of the known regulators *Hunchback*, *Caudal*, *Knirps*, *Bicoid* and *Kruppel* [BNP+02].

The result of this initial check is shown in Figure 3.5. The left–hand panel is a dot plot (generated with gff2aplot [AGW03]) showing the six highest–scoring conserved cis–regulatory modules, the short lines near the diagonal, between the binding site sequences on the two species. The right and bottom sides of the square give the coordinates with respect to the DNA. The tick and text annotations on the left and top of the plot provide the locations of the conserved transcription factor binding sites and the positions and ranks of the conserved cis–regulatory modules with respect to the enhancer element locator scores. On the right, there is a picture of a drosophila embryo with in–situ staining of the *eve* expression, along with the known positions of the enhancers regulating each stripe [BNP+02].

Quite incredibly, the enhancer element locator was able to find all four enhancers within the six highest–ranking CRMs, while the CRMs ranking fourth and fifth were alternatives to the highest–ranking CRM. Even though a lot of prior information went into the Drosophila test and similar results can be obtained with much less sophisticated means this gives confidence to proceed to analyzing mammalian sequences [BNP+02].

3.5.2 Known mammalian elements

The mammalian genomes have significantly more non–coding sequence than the invertebrate fly genomes. The human genome has about 14 times as much DNA per gene as does Drosophila Melanogaster [HAB+07]. Most of this excess is non–functional, but some of it must also code for the differences in the human and fly phenotypes [NZPF+04]. It has been shown that even within species, genes with complex expression patterns tend to have more adjacent non–coding DNA allowing the placement of the complex cis–regulatory elements [NHC04].

We tested the enhancer element locator with human and mouse genomic sequences around the *MyoD1* gene. The parameter values ($\lambda = 2$, $\mu = 0,12$, $\nu = 200$ and $\xi = 200$) were re–optimized for a random sample of human and mouse sequences with methods similar to the ones described in Chapter 4. The analysis used all 107 good quality binding motifs that were at our disposal at the time [VSDB+06]. The resulting dot plot is shown in Figure 3.6.

The enhancer element locator was able to locate the known distal enhancer element of the *MyoD1*–gene as the second–highest scoring cis–regulatory module in the 50kbp region [GBF+95]. The highest–scoring element overlapped the coding region of a nearby gene and the third–highest

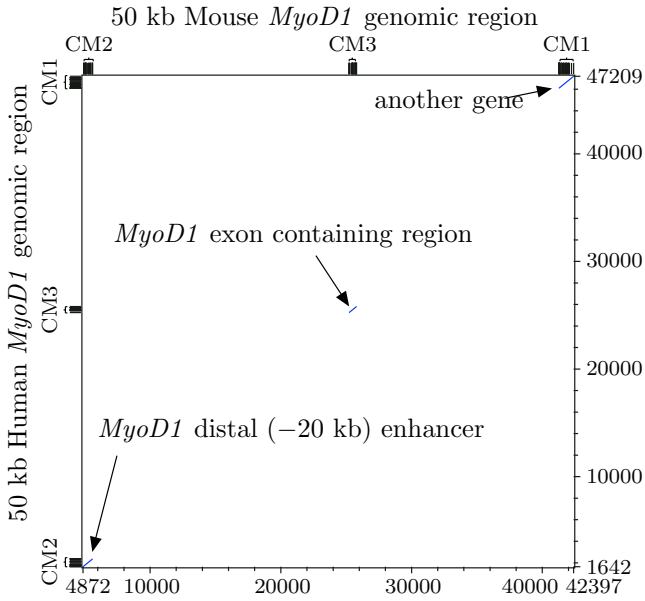

Figure 3.6: Enhancers regulating the mammalian *MyoD1*–gene.

scoring element overlapped with the *MyoD1* coding region. This experiment shows that the enhancer element locator can find functional cis–regulatory modules even in mammalian sequences.

With some faith in the implemented method, we applied it to all human genes that have orthologs in mouse, rat, chicken or fugu in the Ensembl database (version 23) [HAB+07]. The input sequences consisted of the genomic DNA spanning from 100kbp 5'–upstream to 100kbp 3'–downstream of the gene, and the known coding regions were masked so the binding sites do not overlap with exons. One of the sequences were reverse complemented if needed so the orthologous genes are transcribed in a common direction. The computations were distributed to 25 computers in the Biomedicum Bioinformatics Unit computing cluster with the methods of Section 7.2.

With the whole genome analysis we concentrated on the targets of two major cellular signaling pathways Hedgehog and Wnt [IM01, LN04]. These pathways direct growth and patterning during embryonic development and regulate the stem-cell number in adult epithelium [TB01]. Their dysfunction during embryonic development causes various dramatic defects such as cyclopism (holoprosencephaly) and limb malformation and they are commonly mutated in various types of cancers such as basal cell carcinoma and colorectal cancer [TB01, LN04].

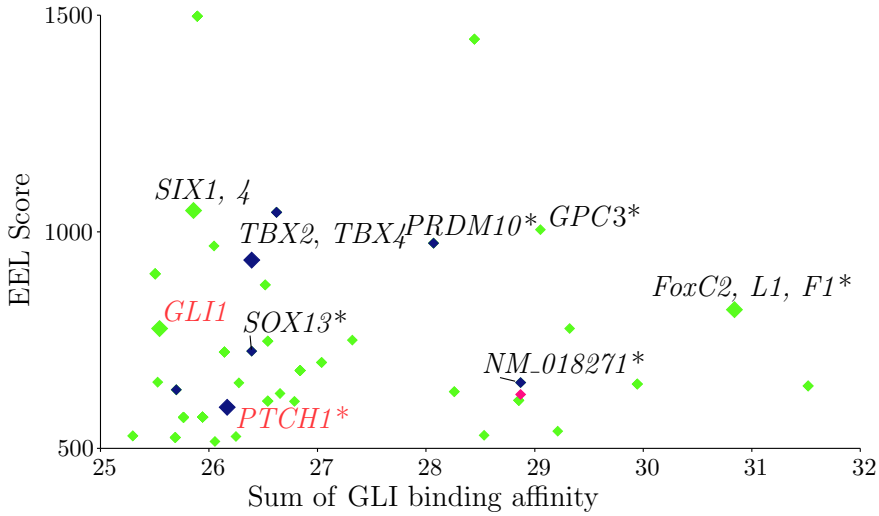

Figure 3.7: Mammalian genes on the Hedgehog signaling pathway transcriptionally regulated by GLI. Human elements with at least two GLI binding sites conserved in mouse (green), rat (blue) and chicken (red). Red–labeled genes are known targets and genes marked with an asterisk are analyzed further below [HPS+06].

The transcriptional response to Hedgehog and Wnt activity is regulated by transcription factors GLI and TCF7L2 (also known as Tcf-4) respectively. We narrowed down on the most likely targets of these factors by concentrating on the CRMs that contain at least two high affinity binding sites for the transcription factor of interest, obtain a high enhancer element locator score and have a biologically motivated maximum length of 2000bp. Figures 3.7 and 3.8 show the genes with the most extreme values. These criteria provide us with 42 putative GLI targets and 132 TCF7L2 targets that were conserved in mouse. Of these 7 and 59 were conserved in rat and 1 and 14 were conserved in mouse, rat and chicken.

Previously only three genes, *GLI1*, *PTCH1* and *HNF-3β*, were known to be directly regulated by GLI and Figure 3.7 contains two of these (*GLI1* and *PTCH1*) [LPCiA97, AKK+04, SHNK97]. Similarly Figure 3.8 contains one, *AXIN2*, of the three direct targets of TCF7L2. In total, six predicted TCF7L2–target elements were located close to known Wnt inducible genes, *AXIN2*, *LEF-1*, *LMX1A*, *c-MET*, *CDX2* and *c-MYC*.

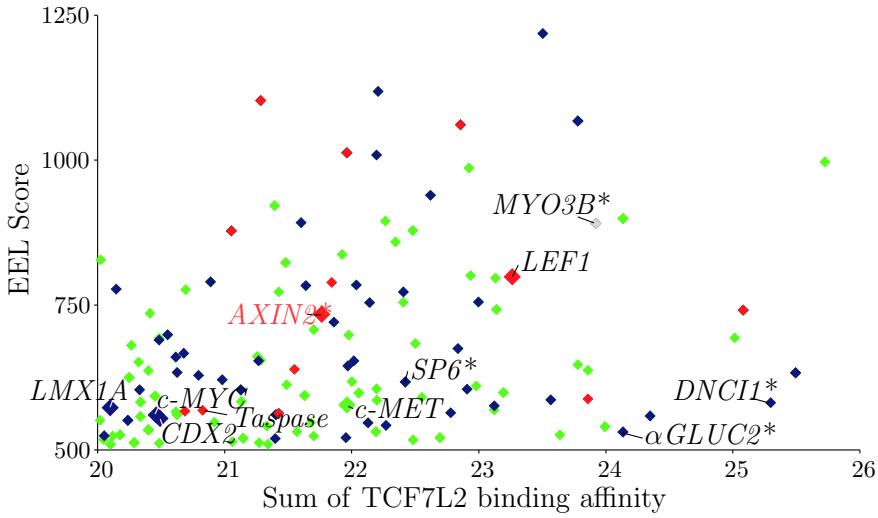

Figure 3.8: Mammalian genes on the Wnt signaling pathway transcriptionally regulated by factor TCF7L2. Human elements with at least two TCF7L2 binding sites conserved in mouse (green), rat (blue) and chicken (red). Red–labeled genes are known targets and genes marked with an asterisk are analyzed further below [HPS+06].

3.5.3 Novel mammalian elements

These positive results were further confirmed with in–situ expression analysis, which displays the expression of a single target gene confined to a particular location. The expression of *Shh*, the gene transporting the hedgehog signal from cell to cell and activating the transcription factor GLI, is shown in Figure 3.9A on whole–mount mouse embryo on the embryonic day E9.5. Of the 16 tested genes close to the high–GLI–affinity CRMs, ten were expressed in a restricted pattern in the analyzed stage (E9.5-E11.5). Five of these, shown as 3.9B–F were expressed in a manner consistent with regulation by Shh.

Some genes, such as the well known *PTCH1* show expression in exactly the same tissues as *Shh* (Figure 3.9B) but more commonly the expression is restricted only to a subset of the targeted cells, as in Figures 3.9C–F. Two of the putatively Shh–regulated genes, *SOX13* and *GPC3* (Figs. 3.9E–F) have not previously been identified as Hedgehog targets.

To verify that the found sequence elements function as independent expression enhancers, we tested their ability to direct *LacZ* expression in transgenic mouse embryos to tissues that are defined by Shh. Indeed, three

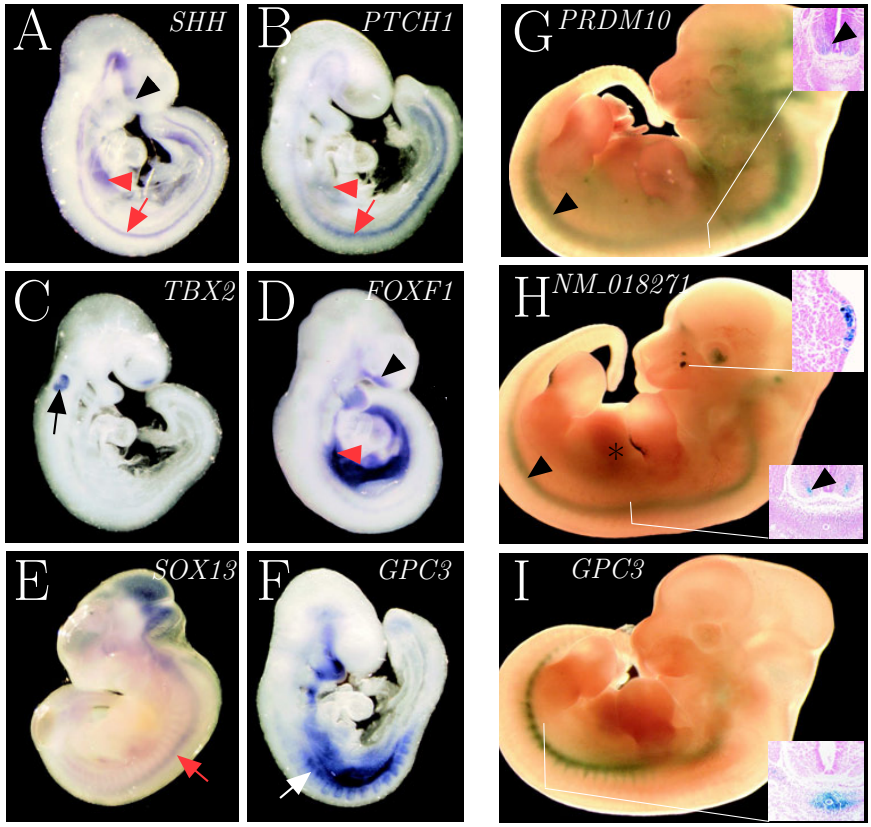

Figure 3.9: In-Situ hybridization of predicted GLI target genes (A–F) and marker expression driven by predicted GLI dependent enhancers (G–I). Highlighted nasal process (black arrowhead), gut (red arrowhead), ventral neural tube (red arrow), ventral auditory vesicle (black arrow) and sclerotome (white arrow). Embryos A–D and F at E9.5, E at 11.5 and G–I at E12.5 [HPS$^+$06].

out of the four (Missing *ATXN2L*) high scoring predicted GLI–regulated enhancer elements directed the expression in the expected manner in almost all of the transgenic embryos (5–6 out of 6 embryos). Figures 3.9G–I display the LacZ expression driven by these elements at E12.5.

It seems that the gene *GPC3* is regulated by more than one enhancer and the other enhancers drive the expression on the anterior parts of the embryo. This can be seen by comparing the mRNA expression of *GPC3* in Figure 3.9F (which is similar to its expression at E12.5 [PPP+98]) to the LacZ expression of its enhancer in Figure 3.9I.

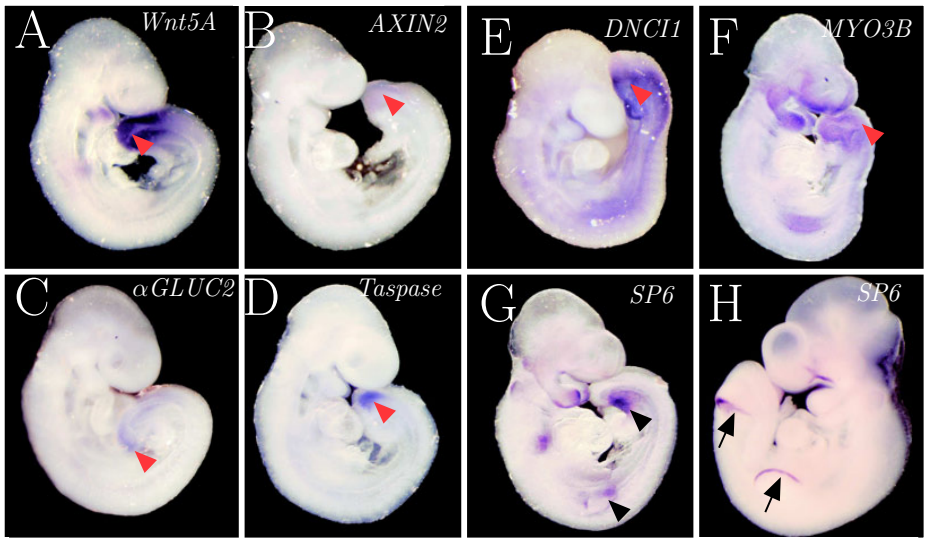

Figure 3.10: Mammalian genes regulated by the transcription factor TCF7L2 (also known as TCF-4) on the Wnt signaling pathway. Red arrowhead: tail bud, black arrowhead: limb buds, black arrows: apical ectodermal ridge. Embryos A-G at E9.5, H at E10.5 [HPS+06].

The results for Wnt target genes regulated by TCF7L2 were as positive. As seen in Figure 3.10A and B, for *Wnt4A* and *AXIN2* at E9.5, the Wnt target genes are typically expressed in the tail, whose formation is known to depend on Wnt signaling [HVB+00]. Another structure whose formation depends on Wnt signals is the apical ectodermal ridge, the location where limbs grow distal from the body and later turn into fingers.

Of the 25 predicted and tested Wnt target genes, 12 were found to be expressed in a restricted pattern at E9.5. Five of these genes that had not previously been described as Wnt targets were expressed in a pattern

consistent with Wnt regulation as seen in Figure 3.10E-H. In addition, four genes, including a known Wnt target *LEF1*, had an expression pattern somewhat consistent with Wnt regulation.

These experimental results show that EEL is able to locate novel enhancers of developmentally regulated genes. The regulation of developmental genes is particularly complex since their expression is spatially and temporally highly specific. The complexity of the regulation is partly the reason why all tested predictions did not work in the laboratory. The failure to find expression on single developmental time point does not mean that the enhancer is not functional in other time points. There are also technical difficulties with transgenesis and in–situ hybridization methods that may materialize in some conditions resulting in aberrant results such as completely white or blue embryos after staining.

3.5.4 Oncogenic regulatory elements

The final test case for the practical utility of EEL is in detection of enhancers for the individual genes that have particular biomedical interest. This practicality was shown by locating functional enhancers near *c–Myc* and *MYCN* proto–oncogenes.

The *c-Myc* is a known target of the Wnt pathway and its expression seems to depend on unidentified distal enhancer elements [LPBM94, HSR$^+$98]. Understanding its regulation has great clinical and scientific importance as it is commonly deregulated in several cancers and its tumorigenesis is one of the reasons preventing the clinical applications of induced Pluripotent Stem cells, which have great potential in treatment of various other diseases such as diabetes or Alzheimer's [CM99, OIY07].

Upon enhancer element locator analysis, the *c-Myc* locus was found to contain TCF7L2–binding sites in CRMs having the third and fifth highest scores in the region. The elements were found to direct LacZ expression in distinct neural tissues, as seen in Figure 3.11.

The *MYCN* oncogene is a similarly important target of the Hedgehog signaling pathway whose deregulation causes various cancers, most commonly neuroblastoma [WAM$^+$97]. Also *MYCN* is known to be regulated by previously unknown distal enhancers that we could find as the fifth and the seventh highest scoring CRMs in the region. Figure 3.12 shows the marker expression driven by the two located elements that include GLI binding sites. Two other elements with GLI binding sites were located in the region but they failed to show consistent tissue–specific expression at E12.5.

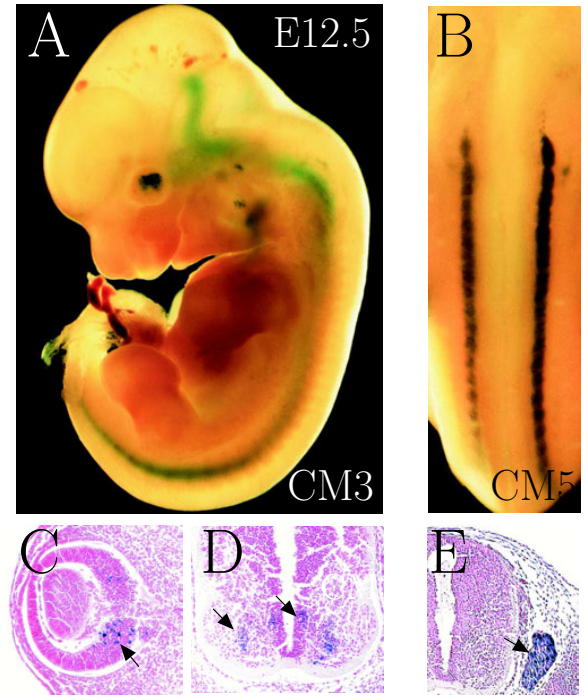

Figure 3.11: Elements near *c–Myc* gene driving tissue–specific marker expression at E12.5. CM3 drives expression to the ventral neural tube (A and D), the eye (C) and CM5 to the dorsal root ganglia (B and E) [HPS+06].

Notable features of the seventh ranking element CM7 on Figure 3.12A and D is the expression in the developing tooth buds where Shh is known to act as a mitogen driving the formation of teeth [DLB+00]. The marker expression driven by CM5 in Figures 3.12B, E and F are consistent with the mRNA expression of *MYCN* at E12.5 [KCR03]. The activity of CM5 in Figure 3.12C and G at postnatal day 3 is also consistent with the response to Shh and its association with childhood neural tumors like neuroblastoma and medulloblastoma [DiA99, KCR03].

3.6 Previous work

Over the last years, there has been considerable interest in finding the cis–regulatory modules computationally. Many different methods have been published that can be classified into several categories. The methods can be described as combinatorial or probabilistic according to the way they model

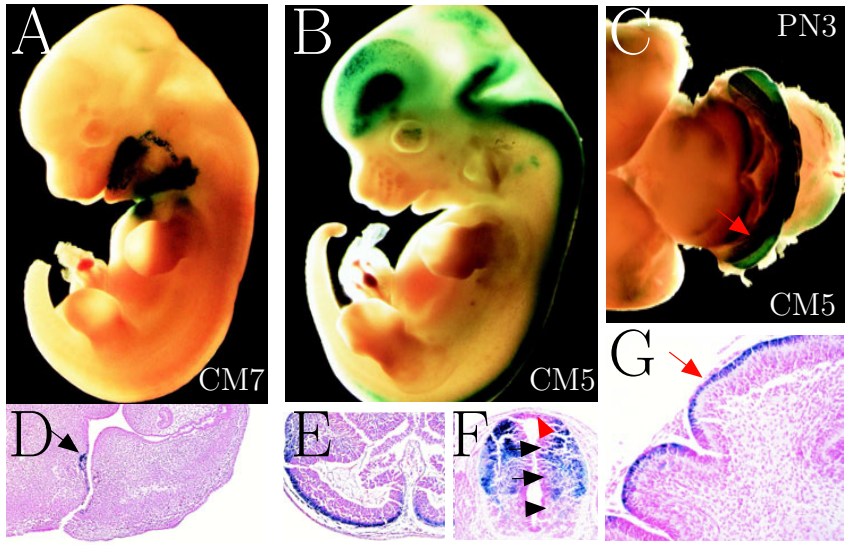

Figure 3.12: Elements near *MYCN* gene driving tissue–specific marker expression. A, B, D–F are at E12.5, C and G at Postnatal (PN) day 3. CM7 drives expression to the neck and mouth (A) and to the developing toothbud (D). CM5 directs expression to the forebrain (B, E) and to the dorsal aspect of the neural tube (F). At PN3 CM5 drives expression to the external layer of the cerebellum (C, G) [HPS$^+$06].

the cis–regulatory modules. These categories are not completely exclusive and features of both can be found in most of the methods. The probabilistic and the combinatorial categories can be further subdivided to single–sequence and multiple–sequence methods.

The early work by Berman and others showed that the distal enhancer elements can be found in the genome of Drosophila Melanogaster simply by locating a sequence window with a high number of transcription factor binding sites [BNP$^+$02]. This simple combinatorial method relies on the relatively compact genome of the invertebrate fly and the well specified set of transcription factors of interest.

The windowed counting methods have been used many times with good success on the analysis of Drosophila early development [MZM$^+$04, MMML02, HGCM02]. Several other methods have improved the binding site counting method by searching dense clusters that are significant according to some statistical model [Wag99, RRP02, JAWL03, BN03, SOBHK03, Pap07, Kre04].

A complementary approach to the ones mentioned above are the probabilistic models that consider the likelihoods of binding sites and their distances. These methods interpret the score of the binding site motif match either as the affinity or the likelihood of the site [FLW03, FHW01, FSHW02, SS05, RVGS02, SvNS03].

The problem of cis–regulatory module finding from a single sequence is naturally extended to usage of multiple sequences. These methods commonly work on the binding sites that lie on aligned DNA, hence they depend on the correctness of the underlying whole genome alignment which can be suspicious around the patchily conserved cis–regulatory modules [LOP+02, MCG06]. The multiple sequence methods can also be divided into combinatorial and probabilistic flavors [AVLT+03, DCG05, PHB05, BBC+06, SvNS03].

A new method by Bais and others combines the binding site annotation and DNA alignment into a single procedure [BGV07]. The binding site containing parts of the alignment are scored according to a binding–motif–dependent scoring function and the intervening non–binding parts are scored according to the traditional substitutions–insertions–deletions type of score [HB98]. The output from their dynamic programming algorithm is the highest–scoring DNA alignment which is annotated with likely the transcription factor binding sites. The method is good for detailed analysis of relatively short sequences but the computational complexity is too high for analysis of long sequences with many transcription factor binding motifs [RKMV07].

The method by Blanco and others aligns sequences of binding sites in much the same way as EEL but their alignment is global instead of local [BMG03, BMSG06]. The global binding site alignment is not able to locate novel CRMs from long sequences but demands exact knowledge about the location of the regulatory element. Their scoring penalizes also for the unaligned binding sites which makes the system intolerant of false positive binding site predictions. The unaligned sites penalty gets high when the input contains conserved low affinity binding sites embedded in abundant non–conserved binding sites with similar affinity.

Many procedures try to solve a somewhat different problem of finding novel cis–regulatory elements similar to a sample of CRMs with highly specific structure and function [CRB97, KW01, ATC+03, ON05]. This approach has the advantage of providing an exact and easy–to–test hypothesis about the element activity and it has provided successful results in flies and to some extent in mammals [WF98, KW01, HGCM02, MZM+04, DCK+05]. The downside of these methods is the need for detailed information about

the structure of the modules of interest. This structure might be discovered with some of the tools that find common cis–regulatory modules from a set of co–regulated genes [ZW04, TPW+04, Kre04, GL05, ZSC05, DG07] (review [JW06]).

Some methods find cis–regulatory elements by discriminating the regulatory DNA regions from the neutrally evolved sequences and also from the sequences conserved for a reason other than transcriptional regulation [EHL+03, TTK+06, PPS+06, PBW06]. These methods are universal in the sense that they do not use or need prior transcription factor binding information. The output provides the putative cis–regulatory DNA sequences but give no clue of the transcription factors binding them. The tissue specificity of the putative regulatory elements can be tested in the wetlab or predicted with other computational methods [PAM+06, PLNO07].

Some methods mentioned, and many more, concentrate only on the proximal promoter regions. These methods are too time– or space–consuming for finding distal cis–regulatory modules from long sequences. The time is spent in either stochastic algorithms like Expectation Maximization and Markov Chain Monte Carlo or in enumerating a large number of binding site or nucleotide combinations [DLR77, GSR96, Kre04, BGV07]. Furthermore, the numerous methods for ab initio binding motif finding are neither applicable for finding distal cis–regulatory modules nor successful with mammalian data [TLB+05].

The question 'which of the cis–regulatory module finding methods is the best' is not well posed at the moment. There is no known 'gold standard' set of distal cis–regulatory elements from mammalian genomes against which one could compare the computational predictions. The largest available set of in–vivo verified enhancer elements is still narrow and it is heavily biased towards extremely conserved DNA elements [PAM+06]. Using synthetic sequences in the evaluation would give an unreasonable edge to the method whose underlying CRM model is closest to the model generating the test data.

3.7 Conclusions

We have developed a biochemically motivated conserved cis–regulatory module model for locating distal cis–regulatory elements. The model and the associated algorithms are both computationally efficient and biologically successful. The conserved CRM model is partly based on the long history of position–specific scoring matrices and partly it is a novel way of clustering

the evolutionarily conserved transcription factor binding sites to close vicinity of each other [Sto00].

The main novelty of the cis–regulatory module model is the way the module conservation is handled. The DNA sequence underlying the module is allowed to evolve freely under the constraint that the protein–DNA and protein–protein interactions must be preserved. This model allows an efficient dynamic programming type of algorithm to find the highly conserved cis–regulatory modules from two evolutionarily related sequences. A model for multi–species CRM conservation is also presented, but with an exponential slowdown with respect to the number of species considered.

The dynamic programming algorithm employed in finding the well–conserved CRMs has few particularities that make its implementation somewhat more complicated than the traditional Smith–Waterman type of DNA alignment [SW81]. The most notable features are the sparseness of the dynamic programming matrix and the large number of potentially extendable alignments that have to be considered when trying to append a new pair of conserved binding sites to an existing alignment. The former feature can be used to ease the difficulty presented by the latter by executing the dynamic programming over inverted indexes.

The conserved CRM model and the actual software implementation have been successfully applied to find known and novel developmentally regulated genes and enhancers [VJM+07]. While the theoretical bases of the conserved cis–regulatory module model are still somewhat hypothetical, the practical applicability of the system is well demonstrated and it lays the foundation for the future improved models of the second genetic code, the code of the regulatory effects of the DNA binding transcription factors.

Chapter 4

Parameter estimation and stability

The conserved cis–regulatory module model of Chapter 3 was developed from a biochemical background with highly hypothetical knowledge of the evolution of functional cis–regulatory elements in complex organisms. This results in a flexible description of nature and it requires the user to specify the many small details of the model. These details include

- the three free parameters of the scoring function,
- the set of transcription factor binding motifs,
- the distribution assumption about the genomic background and
- the thresholds for calling a transcription factor motif match a binding site.

The choice for the set of transcription factor binding motifs is in principle limited by the amount of information available but one might still be interested in some 'robustness' features of the motif set [VSDB$^+$06]. This way the similar transcription factor binding motifs would not block each other too much, hence hampering the post-processing and analysis of the resulting cis–regulatory module predictions.

A reasonable approach for selecting the correct parameter values is to use heuristics to estimate their goodness. Having some goodness measure, we can apply a numerical optimization procedure to find a close–to–optimal set of parameter values. For goodness measures without reasonably computable gradients one has to resort to computation intensive optimization procedures, such as Simulated Annealing [KGV83]. This approach has at least two problems. It is not clear how to measure the goodness of a set of parameters for the conserved cis–regulatory module model and it is not certain that

Simulated Annealing, or any other local search method, finds a set of parameter values anywhere near the global optimum.

The rest of this chapter is organized as follows. Section 4.1 describes the Simulated Annealing optimization algorithm and Section 4.2 provides the evaluation function for a set of parameter values. Section 4.3 discusses the ways of obtaining DNA sequence that do not contain conserved CRMs. These are needed in the parameter value evaluation as examples of wrong predictions. Sections 4.4 and 4.5 apply the optimization procedure to two types of training data. Section 4.6 considers omitting some of the parameters in the conserved CRM model in order to obtain a simpler and more stable model. Section 4.7 analyzes the sensitivity of the prediction results to small variations in the parameter values. Finally, Section 4.8 concludes the chapter.

4.1 Simulated Annealing for model parameter optimization

Simulated Annealing is a general–purpose optimization method for non–linear or combinatorial problems [KGV83, MF00]. It draws from the parallelism with annealing of crystals in statistical mechanics, where it is important to carefully anneal a liquid so it can reach the lowest energy state and not get stuck in local optima which would result in faulty crystals and brittle materials. We use Simulated Annealing to optimize the parameter values because it does not utilize the gradient of the target function which would be difficult to compute in our setting.

Fundamentally, Simulated Annealing is a stochastic hill–climbing procedure augmented with a chance of escaping the local optima by having a certain probability of moving from a good to a worse state. This probability is controlled with a temperature value T which is slowly decreased over the run of the Simulated Annealing procedure. Because of the stochastic nature of the search the procedure is restarted multiple times from independent locations of the search space. The multiple local optima from the restarts allow the evaluation of the overall convergence of the procedure.

High values of T allow the optimization procedure to take downhill steps almost at random and allow the system to cover large areas of the search space without getting stuck to local maxima. When the system is annealed, i.e., the temperature T is lowered, the procedure becomes more and more keen on taking moves only to better states.

The detailed algorithm of Simulated Annealing for estimating the parameter values for the enhancer element locator tool is illustrated as Algorithm 2.

Algorithm 2 Simulated Annealing for EEL parameter optimization

Require: Evaluation function $eval()$. Cooling rate r. Threshold temperatures T_{max} and T_{min}.
Ensure: Locally optimal set of parameters v_o.

```
Initialize set of current parameters v_c at random.
T ← T_max
v_o ← v_c
repeat
  for j = 1, ..., 10 do
    Select a new set of parameters v_n in the neighborhood of v_c.
    if eval(v_c) < eval(v_n) then
      v_c ← v_n
      if eval(v_o) < eval(v_n) then
        v_o ← v_n
      end if
    else if random[0,1) < e^((eval(v_n)−eval(v_c))/T) then
      v_c ← v_n
    end if
  end for
  T ← r · T
until T < T_min
```

The algorithm tries random steps in the parameter space and takes the step either if the resulting parameter values are better than the current ones or with a probability dependent on the decrease of the quality of the parameters and the current temperature.

The parameters can be initialized by a random draw from their domain. Each of the parameters μ, ν and ξ are constrained to be greater than zero. While there is no hard top limit for these values we choose to initialize the parameters by random values from uniform distribution

$$\begin{aligned} \xi, \nu &\sim \mathrm{U}[0, 500] \\ \mu &\sim \mathrm{U}[0, 5] \end{aligned}$$

which covers a significant proportion of the reasonable parameter values based on preliminary tests. The greater range for ξ and ν is due to the expected importance of the conservation of the binding site distances.

As for the neighborhood of the parameter values, Michalewicz and Fogel suggest using normal distribution with the mean at the current value and standard deviation equaling one sixth of the range of the variable [MF00]. Even though the ranges of our variables are unbounded, we apply this

advice and define the neighborhood as normally distributed with the mean of current values and the standard deviation being one third of the absolute value of the parameter, hence the new parameter value x_n is sampled according to

$$x_n \sim \max\left\{0.01, \mathrm{N}\left(x_c, \frac{x_c}{3}\right)\right\} \tag{4.1}$$

for the current parameter value x_c. The maximization is needed to keep the parameter value strictly positive. This kind of neighborhood allows the system to explore the parameter space quite extensively while still maintaining close control of the parameters with small absolute values that presumably are more sensitive to changes.

The choice of the starting and ending temperatures T_{max} and T_{min} are heavily dependent on the scores provided by the evaluation function $eval()$. The choice of the evaluation function along with the temperature thresholds are discussed in Section 4.2.

Final detail of the Simulated Annealing procedure is the cooling rate. While a certain version of Simulated Annealing is able to recover the global optimum with probability one, the annealing speed has to be impractically slow to reach this. To make the procedure practically feasible, the cooling speed is usually chosen to be exponential with some predetermined constant rate $r < 1$ as is done in Algorithm 2. In our practical implementation the cooling rate was set to value $r = 0.9$. The choice of the temperature limits and the cooling rate are interwoven with the speed of the algorithm since the outer loop of Algorithm 2 is executed k times until

$$T_{min} > T_{max} r^k, \tag{4.2}$$

which results in

$$k = \left\lceil \frac{\log T_{max} - \log T_{min}}{-\log r} \right\rceil. \tag{4.3}$$

The inner loop is executed $10k$ times so the evaluation function is executed $10k + 1$ times in total. In our case evaluating the parameter values is the most time–consuming part of the Simulated Annealing procedure since we are using quite a computation intensive–evaluation function. These time complexities are given for a single run of the Simulated Annealing procedure so each restart increases the time requirements respectively.

4.2 Evaluation of a set of parameter values

The most crucial part of any parameter–learning system is the evaluation function, which determines the quality of a given set of parameter values.

In a cis–regulatory module–prediction context this function should provide high scores for correct cis–regulatory element predictions and low scores for incorrect ones.

At least two approaches can be used for estimating the quality of the parameter values. One can consider all conserved sequences as correct predictions and try to maximize the score difference between the real orthologous sequences and the simulated non–conserved sequences. On the other hand, if one has data on working cis–regulatory elements, one could try to optimize their score against the sequences without functional elements.

We now assume that we have two sets of pairs of sequences: The positive pairs $\mathcal{P}$ likely to contain cis–regulatory elements and the negative pairs $\mathcal{N}$ that do not contain functional elements. With the methods of Chapter 3, we can compute the enhancer element locator score $EEL_p(x)$ as the maximal enhancer element locator score for the pair of DNA sequences x, with the parameter values p. Now we can evaluate the set of parameter values p by setting

$$eval(p) = \frac{\sum\limits_{x \in \mathcal{P}} EEL_p(x)}{\max\limits_{y \in \mathcal{N}} EEL_p(y)}. \tag{4.4}$$

This function will give better scores for parameter values that give a high score for the highest–scoring prediction in the positive sequences and scales the scoring range to be relative to the scores from the negative sequences. Evaluation scores larger than $|\mathcal{P}|$ give an indication for 'good' parameter values although the actual score is quite sensitive to outliers due to the small data size and the maximization in the denominator of Equation (4.4).

The starting temperature of the Simulated Annealing procedure should be selected such that the search explores the parameter space relatively freely. This can be achieved by having the downhill stepping probability close to half. If we see the evaluation function of equation (4.4) as a sum of independent and identically distributed random variables we can select T_{max} by probabilistic considerations. By the central limit theorem, the value $eval(v_c)$ is normally distributed in the limit of large $\mathcal{P}$ and by the variance sum rules its variance is equal to $|\mathcal{P}|s^2$, where

$$s^2 = \mathrm{var}\left(\frac{EEL_p(X)}{\max\limits_{y \in \mathcal{N}} EEL_p(y)}\right) \tag{4.5}$$

is the variance of the terms of Equation (4.4). The variance s^2 can be easily approximated when computing the $eval(v_c)$ for the first time as

$$s^2 = \frac{1}{|\mathcal{P}| - 1} \sum_{x \in \mathcal{P}} \left(\frac{EEL_p(x)}{\max_{y \in \mathcal{N}} EEL_p(y)} - \frac{eval(p)}{|\mathcal{P}|} \right)^2 . \tag{4.6}$$

Assuming that $eval(v_n)$ has the same mean and variance as $eval(v_c)$, and $eval(v_c) > eval(v_n)$, it follows from the normal approximation that

$$eval(v_c) - eval(v_n) < \Phi^{-1}(0.75)\sqrt{|\mathcal{P}|s^2} \approx 0.6745\sqrt{|\mathcal{P}|s^2} \tag{4.7}$$

approximately with probability 0.5. This would suggest using the starting temperature

$$T_{max} = \frac{0.6745\sqrt{|\mathcal{P}|s^2}}{\ln 2}, \tag{4.8}$$

which we expect to provide downhill stepping probability of 0.5 about half of the time.

The above analysis provides a handle on the initial temperature choice but it does make some approximating assumptions. Firstly it assumes constant variance of the enhancer element locator scores over the full parameter range. Also the normal approximation for the evaluation score distribution is likely not very exact due to the relatively small number $|\mathcal{P}|$ of summed variables.

4.3 Simulating neutrally evolved DNA sequences

To identify the conserved cis–regulatory modules from the orthologous DNA sequences, we must be able to distinguish the conservation of the cis–regulatory modules from the neutrally evolved sequence around it [Kim68]. The neutral theory of evolution assumes an ancestral sequence which has duplicated and changed over time into present–day sequences without any evolutionary pressure in any direction and most of the mutations have been selectively neutral. This section surveys methods for modeling neutral evolution and Section 4.3.1 gives the required parameter values for simulating the neutral evolution between human and mouse.

There has been extensive theoretical work on the issue of how the nucleotide base substitutions occur and get fixed in the DNA over generations [JC69, Tav86, Fel81, HKY85, HB98]. The standard way of modeling the substitutions is to utilize a continuous time Markov model which has

the state space of bases in a given site. The various substitution models vary by the amount of biological details modeled and by the number of parameters required. The simplest substitution model is the Jukes–Cantor model which assumes that all four bases and all substitutions between them are equiprobable [JC69]. The other extreme of the substitution models is the generalized time–reversible model which allows the genomic base distribution and the substitution probabilities between any two bases to be set arbitrarily and independently from each other [Tav86].

Several other substitution models have been proposed to take into account such biologically relevant features as a biased base distribution, a difference in the rate of transition (A $\leftrightarrow$ G or C $\leftrightarrow$ T) and transversion (A $\leftrightarrow$ C, G $\leftrightarrow$ C, A $\leftrightarrow$ T or G $\leftrightarrow$ T) substitutions or both [Kim80, HKY85, Fel81]. The base distribution and the transitions to transversions ratio can differ significantly from their trivial values in the Jukes–Cantor substitution model.

The neutral parameter values for a substitution model can be estimated with a maximum likelihood fit to the synonymous codons of conserved genes [Yan97]. The synonymous codons can be reliably aligned since they are embedded in the highly conserved protein coding regions. The synonymous codons do not affect the amino–acid sequence of the encoded protein so they are assumed to be selectively neutral. Even though the exact evolutionary parameter values vary between the genomic locations, the overall averages of these parameters can be found in the literature [RSK03, KS02].

In addition to point substitutions, the genomes have undergone extensive insertion, deletion (*indel*) and rearrangement mutations. As compared to substitutions, these types of mutations have received much less interest from researchers; no standard framework exists for modeling or simulating their evolution. The DNA alignment procedures handle the insertions and deletions with linear or affine gap penalties that are useful in practice but are not founded on any evolution model. The early work on indel models provided only single residue indel events or events concerning indivisible fragments where deletion cannot occur inside an old insertion [TKF91, TKF92]. Recently a new time-reversible indel model was proposed which resolves some of these problems but does not take a stand about the length distribution of the insertions and deletions [MLH04].

Because the indel events can occur on top of each other, i.e., a deletion can occur within an insertion, the length distribution of the indel events cannot be directly deduced from the alignment of distantly related sequences [OSK04]. By assuming that the indel length distribution is independent of the number of the indel events, that is the evolutionary distance, one can estimate the indel length distribution with the gap lengths in

the alignment of closely related genomes. Because of the small evolutionary distance between the closely related species, it is unlikely for the indel events to overlap in the genome. The gap length distribution in the alignment of human and chimpanzee genomes follows the Zipf–distribution (i.e., power–law) to some extent, but there are noticeable peaks around the lengths of the common transposable DNA elements [VAM$^+$01, LLB$^+$01, MHE$^+$05, OSK04]. Each linage specific insertion of a transposable element results in an alignment gap of the length of the element.

4.3.1 Evolutionary parameter values between human and mouse

The mouse, Mus Musculus, is the most important model animal in the research of mammalian disease and development. It has long been used as a model animal for many kinds of research and its genome has been thoroughly sequenced [WLTB$^+$02]. The mouse genome is useful in comparative analysis with human because of the 'optimal' evolutionary distance between the two species. Neutrally evolved human and mouse sequences have gathered about $0.45 - 0.58$ substitutions per site after their latest common ancestor some 65 million years ago [WLTB$^+$02, OSK04, CBS$^+$04, KS02]. This evolutionary distance has allowed the neutrally evolving sequence to become almost random while the coding regions of most of the genes that are under strong stabilizing pressure have accumulated only few amino acid substitutions making their alignment and homology detection relatively easy [WLTB$^+$02].

Different nucleotide bases have not been substituted uniformly to each other. The ratio of transition (A $\leftrightarrow$ G or C $\leftrightarrow$ T) and transversion (A $\leftrightarrow$ C, G $\leftrightarrow$ C, A $\leftrightarrow$ T or G $\leftrightarrow$ T) substitutions during the evolution of human and mouse has not been the trivial $1:1$ as assumed in the Jukes–Cantor model but it has been estimated to an average of $3.66:1$ [RSK03].

In contrary to substitutions, the insertions and deletions have largely been disregarded in most phylogenetic analyzes even though they are important features in the cis–regulatory module conservation [CCB$^+$05]. Recent work suggests that the number of indel events in the genome is about 10 fold less than the number of substitution events. This results on about 0.045 indels per site between human and mouse [OSK04, LPH06].

In the simulation studies of Section 4.4 and Section 5.3, we used the HKY–substitution model with 0.45 substitutions per site and the indel lengths from the Zipf-distribution with the same mean as observed for human–chimpanzee alignment gaps [HKY85, Zip35, OSK04]. This results in insertions and

deletions of length L such that

$$\mathrm{P}(\mathrm{L} = l) = \frac{l^{-1.8}}{\zeta(1.8)} \tag{4.9}$$

where $\zeta(x) = \sum_{n=1}^{\infty} \frac{1}{n^x}$ is the Riemann zeta-function. Since the indel event lengths generated by the transposable elements are quite large, of the length a few hundred base pairs, we ignore them in our simulations since already much shorter indels destroy the conserved cis–regulatory module for the reasonable CRM model parameter values.

4.4 Parameter learning from simulated data

The choice of the sequences for sets $\mathcal{P}$ and $\mathcal{N}$ is crucial for the evaluation function of Equation (4.4). It would be beneficial for set $\mathcal{P}$ to contain sequences with real functional and conserved cis–regulatory elements while set $\mathcal{N}$ should contain none.

Since we do not have an unbiased sample of functional cis–regulatory elements to use as the positive training data, we chose $\mathcal{P}$ to contain sequences around genes that are developmentally regulated. To be specific, we used the mouse Gene Expression Database from the Jackson Laboratories to search for the genes that show expression in *in situ* hybridization at embryonic time point E12.5 and show no expression at time point E15 or later [HBF$^+$04]. This developmental stage at mid gestation is interesting for the research into vertebrate development and it is commonly used in experimental enhancer studies [PAM$^+$06].

Of the 23 genes obtained from the expression database, we rejected three (*Hoxa2*, *Hoxa6*, *Hoxd13*) for notoriously high conservation and somewhat unique regulation [LM06]. One gene (*Adra2b*) did not have a unique homolog in human and was discarded for that. Finally for set $\mathcal{P}$ we selected the sequences of the ten shortest mouse genes along with their human orthologs and included 100000 base pairs on both sides of the gene. All exons in the sequences, including the exons of other genes, were masked out. The sequences around the nine other orthologous genes were retrieved similarly and used as a test set for the found optimum. The genes for the optimization are listed in top and and the genes for the test are listed in bottom of Table 4.1.

The set $\mathcal{N}$ of negative sequences is generated with the methods of Section 4.3 by using the human sequences from $\mathcal{P}$ as ancestors. The GC–content of the generated sequences is set to match the real human sequence. Ten pairs of sequences were generated for $\mathcal{N}$ with this method such that they are related similarly to neutrally evolved human and mouse sequences.

Mouse Gene	Mouse gene Id	Human gene	Human gene Id
Lsr	ENSMUSG00000001247	LSR	ENSG00000105699
Dll3	ENSMUSG00000003436	DLL3	ENSG00000090932
Mfge8	ENSMUSG00000030605	MFGE8	ENSG00000140545
Plxna3	ENSMUSG00000031398	PLXNA3	ENSG00000130827
Lmo2	ENSMUSG00000032698	INHA	ENSG00000123999
Armcx2	ENSMUSG00000033436	ARMCX2	ENSG00000184867
Foxc1	ENSMUSG00000050295	FOXC1	ENSG00000054598
Nkx2-9	ENSMUSG00000058669	NKX2-8	ENSG00000136327
Acvr2b	ENSMUSG00000061393	ACVR2B	ENSG00000114739
Sox13	ENSMUSG00000070643	SOX13	ENSG00000143842
Meox1	ENSMUSG00000001493	MEOX1	ENSG00000005102
Dlx6	ENSMUSG00000029754	DLX6	ENSG00000006377
Myf5	ENSMUSG00000000435	MYF5	ENSG00000111049
Casz1	ENSMUSG00000028977	CASZ1	ENSG00000130940
Lmo2	ENSMUSG00000032698	LMO2	ENSG00000135363
Acvr1b	ENSMUSG00000000532	ACVR1B	ENSG00000135503
Rarg	ENSMUSG00000001288	RARG	ENSG00000172819
Gsx2	ENSMUSG00000035946	GSX2	ENSG00000180613
Mitf	ENSMUSG00000035158	MITF	ENSG00000187098

Table 4.1: Genes selected for $\mathcal{P}$ [BEB$^+$06, HAB$^+$07, EDS$^+$06]

4.4.1 Results

The Simulated Annealing procedure with 12 independent restarts ultimately found the best parameter values $\xi = 704.844$, $\mu = 0.211751$ and $\nu = 0.236472$. These parameter values have evaluation score $eval(p) = 33.0$. Because the evaluation score is more than three times larger than $|\mathcal{P}|$ we conclude that the enhancer element locator is able to distinguish between conserved and neutrally evolved sequences.

There are some notable issues about the given optimal parameters. Figure 4.1 displays the evaluation scores for the parameters tested during Simulated Annealing. The color coding is with respect to the evaluation score. It is somewhat surprising to see such a clear effect of most of the parameters to the evaluation score.

The interesting features of Figure 4.1 are the evaluation score peaks for μ and ν and the asymptotic behavior of those parameters. The evaluation score $eval()$ converges to the limiting value of ≈ 10 for large values of μ which is understandable since large values of μ result in short enhancers that occur randomly on both positive and negative sequences. On the other hand, the limit ≈ 22 for high values of ν is quite unexpected. It seems

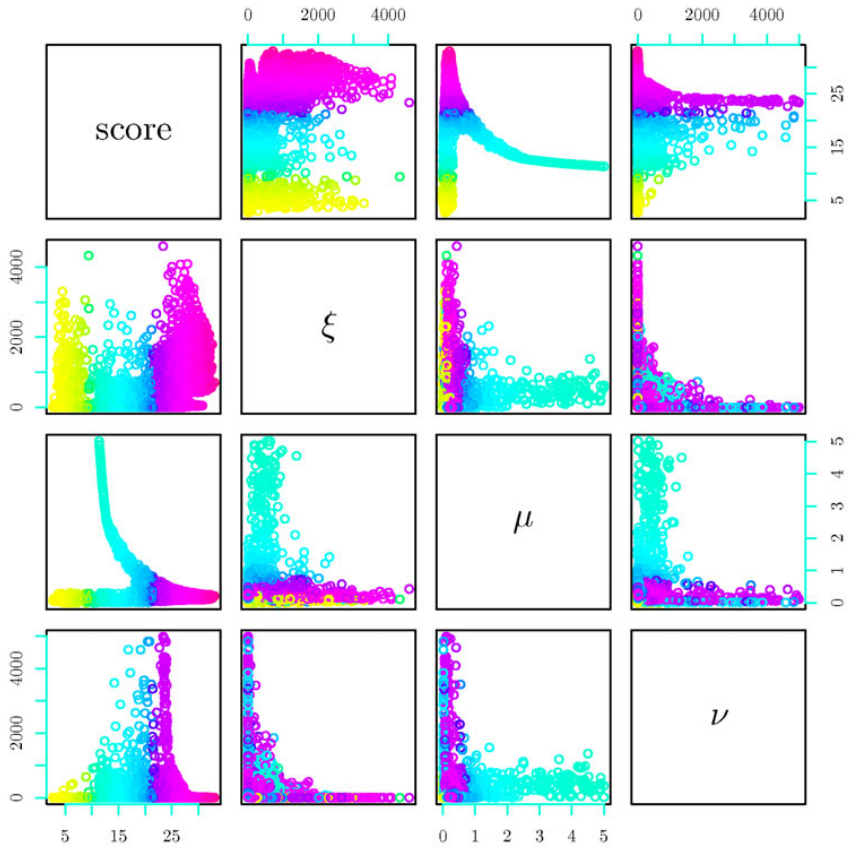

Figure 4.1: Evaluation scores with simulated data for sample of parameters.

that variation in the distance between adjacent binding sites is completely forbidden when the penalty grows.

Found optimal parameter values are robust with respect to the training data. The optimization with the training data did not improve the evaluation score significantly when starting from the above optimum. The optimum had the seventh highest evaluation score (68.5) out of one hundred evaluated parameter value settings (the best evaluation score 75.47) in the neighborhood. This 're–recovery' of the optimal parameter values provides additional confidence for the parameter values found through the Simulated Annealing procedure.

4.5 Parameter learning from a sample of positive enhancers

Pennacchio and others have devised a relatively high–throughput transgenic marker assay for characterizing developmental enhancers *in vivo* and used it to test several hundreds of sequences for enhancer activity [PAM+06]. Their data provides the first reasonably sized set of positive examples of cis–regulatory element sequences. Unfortunately, the sequences selected for testing were not selected in an unbiased manner but they were selected according to strong and deep conservation in various vertebrate species [PAM+06]. This extreme conservation hinders the usage of the Pennacchios dataset for the parameter evaluation.

Another problem with the Pennacchios dataset is the inevitable ambiguity of the negative results. A missing expression in the tested E11.5 stage does not mean a missing expression in all other developmental or adult stages and conditions. Nevertheless we applied the parameter estimation method to this data to see whether we can distinguish the functional enhancers from the non–functional ones on that single embryonic time point.

The data used for optimizing the parameter values for finding enhancers active at the developmental stage E11.5 was downloaded from the enhancer browser [PAM+06] `http://enhancer.lbl.gov/` (Jan. 23. 2007). Out of the 114 mouse sequences showing positive enhancer activity at E11.5, the ten shortest ones were selected, while taking care that the selected sequences are physically distant from each other in the genome. Out of the 155 sequences tested, but not showing enough activity to be classified as positive, we selected ten sequences slightly longer than the positive enhancers selected above. Confining to a small set of ten sequences is needed to enable parameter optimization in reasonable time with the Simulated Annealing.

The location of the sequences in the mouse genome were mapped to the orthologous regions in the human genome with the liftOver–software tool with the chain files provided by the University of California at Santa Cruz [KKZ+07]. The sequences along with 10000bp on each side of the tested enhancer was downloaded from the Ensembl web site [HAB+07]. Tandem repeats and exons within the sequence regions were masked out (data from Ensembl 41). The flanking regions were selected shorter than usually to speed up the parameter evaluation. This should not have a significant effect on the results since the highest–scoring predictions tend to be in the highly conserved middle part of the sequence.

To avoid the possibility of having functional enhancers in the negative set, we randomly permute the local alignments between human and mouse sequences that obtain a BLAST e–value less than 10^{-25} [AMS+97]. Since we apply the permutation to the alignment, this procedure conserves the percentage identity on the highly conserved sequence but destroys the conserved binding sites and gaps between them.

4.5.1 Results

The best parameter values for the Pennacchio dataset were $\xi = 29.29$, $\mu = 0.416689$ and $\nu = 4725.07$. The evaluation score for these with the Pennacchios dataset was $eval(p) = 9.834684$. These parameter values differ significantly from the ones derived from the simulated data in Section 4.4.1. The two most important differences are the significantly higher value for μ and the lower evaluation score. The values obtained for ξ and ν are not significant since the sequences in the Pennacchios dataset are so conserved that there is not enough insertions and deletions to be counteracted with those penalty parameters. This can also be seen in Figure 4.2 where parameters ξ and ν faithfully reproduce each other's behavior with respect to the evaluation score and also with respect to each other.

As seen in Figure 4.2, while the best parameter values provide as good scores for positive enhancers on average as for the randomized enhancers on maximum, the found maximum evaluation score is somewhat above the 'bad' value ≈ 8 which is obtained in the limit of large μ.

When the above optimal parameter values were tested with the independent set of similarly chosen set of sequences, the optimization score dropped to 5.4 and the new optimum obtained evaluation score 9.6 with completely different set of parameter values. These results suggest that the conserved CRM model is not able to improve the enhancer predictions based on extreme sequence conservation. Therefore I do not consider the parameter values derived in this section any further.

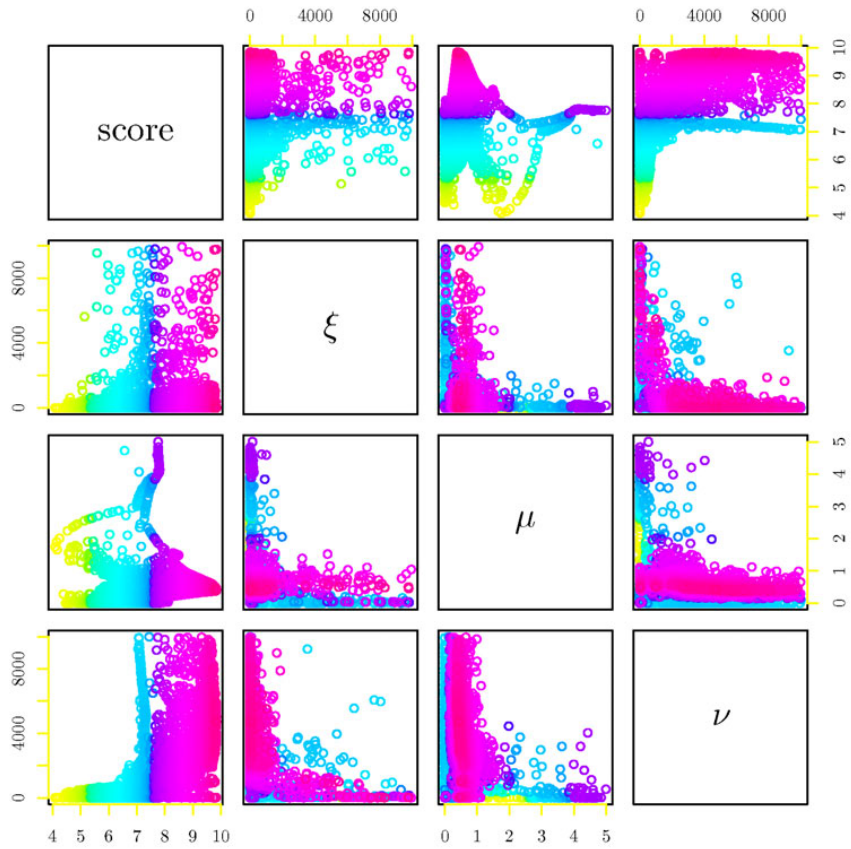

Figure 4.2: Evaluation scores with the Pennacchios data for sample of parameters.

4.6 Effects of individual parameters

The reasonably small changes in the parameter values can have significant effects on the quality of the results obtained with the conserved cis–regulatory module model. As seen in Figure 4.1 some of the parameters, like μ and to a lesser extent ν have distinct peaks in the quality landscape while ξ has only a weak and wide peak with a plateau not much worse.

Since the parameters ξ and ν weight a similar feature in the conserved cis–regulatory module model, the indel length either monotonically or cyclically, one might question whether one could give up from one of the parameters so the parameter learning would become easier by having one less degrees of freedom. This idea can be supported by the strict L–shape of the ξ vs. ν plots in Figures 4.1 and 4.2. To answer this question we used the Simulated Annealing procedure with the simulated data keeping either ξ or ν parameter fixed to zero.

After three independent restarts, the evaluation score decreased from 33.0 to 27.35 and 27.22 for omitted ν and ξ respectively. The optimal parameter values, $\mu = 0.189 - 0.208$ and $\nu = 320.0 - 358.3$ with omitted ξ and $\xi = 2261.49 - 3735.27$ and $\mu = 0.227 - 0.248$ with omitted ν were recovered repeatedly. For omitted ν case one of the restarts resulted in a local optima of score 26.6 with higher μ and much lower ξ.

The drop of 6 in the optimal evaluation score is quite large when seen in the context of Figure 4.1. The large drop of the optimal score when one of the parameters is omitted, shows that all of the model parameters are important in distinguishing the orthologous sequences from the simulated ones. While the optimal parameter values found in Section 4.4 and 4.5 are located in quite an unrestricted landscape it seems that both ξ and ν are important and neither of them can be categorically discarded.

4.7 Stability of the CRM predictions

Now that we know the set of parameter values that provide good separation between real and randomized sequences, we would also like to know about the sensitivity of the actual results to small variations in these values. Since the user is only interested in at most a few dozen best–ranking cis–regulatory modules in a pair of sequences, we are interested in how much this set changes with the small parameter value perturbations.

Quantifying the change in the results is not completely obvious since there are at least three properties changing concurrently. First, the ranking of the conserved cis–regulatory modules can change, second, the contents,

the factors binding the CRM can change and finally, the region of the CRM on the DNA can change. These properties are interrelated in the sense that in order to compare the rankings of the top–k cis–regulatory modules we need to resolve the discrepancies between the modules induced by the changes in content and location [FKS03]. These discrepancies could be solved by matching the most similar enhancers in the two lists but this method would disregard the scale of the changes in the enhancer content and position.

Since the absolute ranking of the predictions is not as important as scoring good predictions 'high' and bad predictions 'low', the stability of the enhancer predictions can be quantified by the overlap, on the DNA or on the individual binding sites, of the two sets of conserved CRMs. The DNA overlap measure gives the percentage of DNA that is covered by enhancers in both sets, out of the DNA that is covered by enhancers in either set. An alternative is the site overlap measure which gives the percentage of binding sites shared by the two sets of CRMs.

The two measures give quite similar results but emphasize different features of the prediction. The DNA overlap measure is relevant for stability of the results in transgenic assays that use the cloned DNA in a reporter vector and does not care about the actual factors binding the DNA. The site overlap, on the other hand, is relevant when interested in the biochemical mechanisms of the enhancer and one is looking for targets of a specific transcription factor.

The effect of perturbed parameter values on the stability of the prediction results was tested according to both DNA and site overlap measures for the 20 highest–scoring cis–regulatory modules. The overlap measures were averaged over the 10 pairs of real orthologous DNA sequences given in Table 4.1. The prediction made with the optimal parameters from Section 4.4.1 ($\xi = 704.844$, $\mu = 0.211751$ and $\nu = 0.236472$) is compared to predictions with the altered parameters with the DNA overlap measure in Figure 4.3 and the site overlap measure in Figure 4.4. In each panel, the color indicates the overlap with the missing parameter fixed to the optimal value.

Even though the overlap stays well above overlaps of independent IID sequences (where the DNA overlap over 1000 trials is always less than 0.08) over a wide range of parameter values, the results still show quite strong variation with slight perturbations of the parameter values. As expected, the sensitivity to μ (the distance penalty) is much stronger than to ν or ξ. The dependence of ξ and ν is also expected but their dependence on μ is somewhat surprising.

When comparing the two overlap measures one notices that the site overlap measure drops to about 50% faster than the DNA overlap but stays

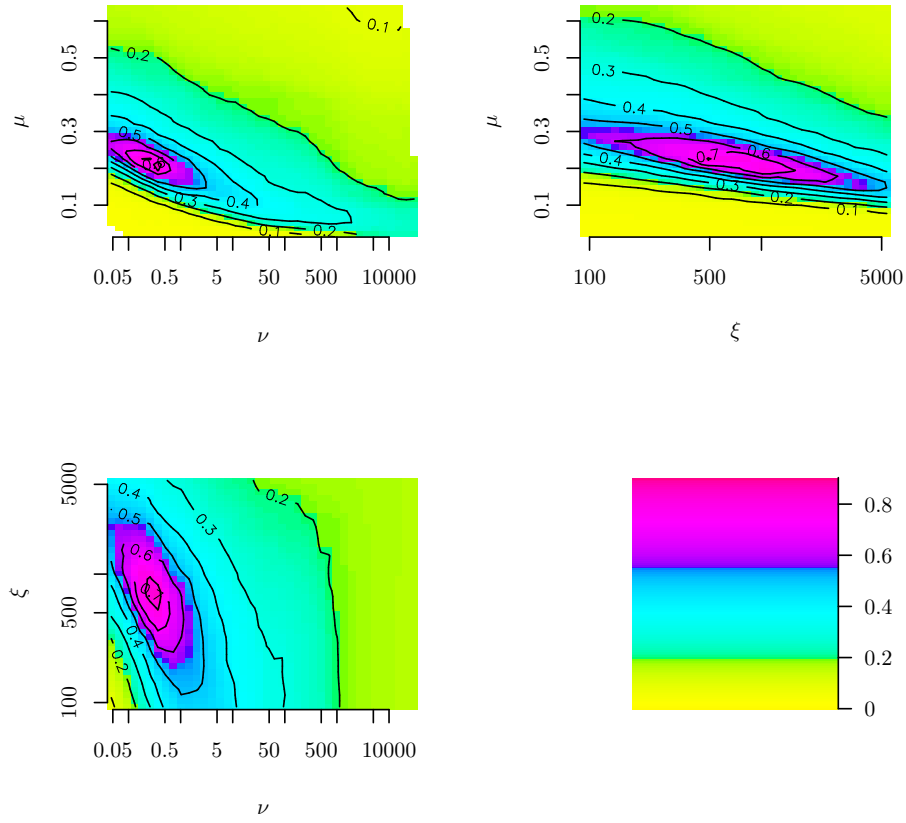

Figure 4.3: DNA overlap between alignments with optimal and modified parameters.

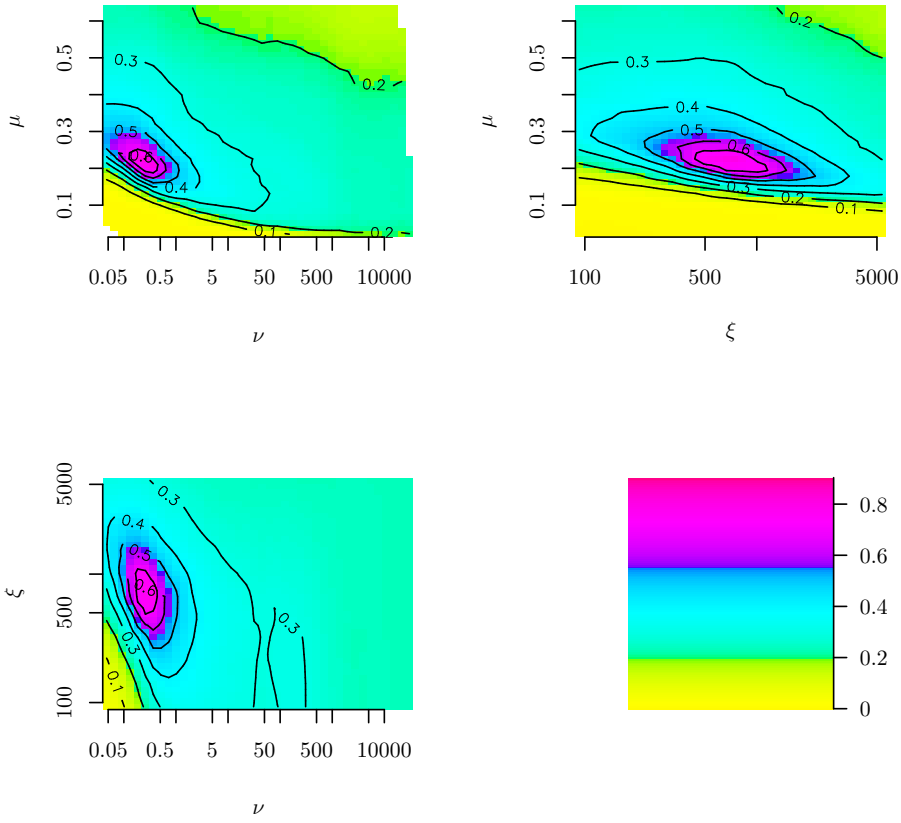

Figure 4.4: Site overlap between alignments with optimal and modified parameters.

around 30% longer when the parameter values are increased. This results from the best conserved parts of the DNA that were also included in the original CRMs and contain well spaced, conserved binding sites. The continued decrease in the DNA overlap results from shortening the overall length of the predicted cis–regulatory modules by requiring more 'tight' clusters.

4.7.1 Sensitivity to the binding sites

One of the input parameters for the conserved cis–regulatory module model has not been addressed above. The alignment results are also sensitive to the cutoff for the transcription factor binding site matches. The enhancer element locator scores change monotonically with the cutoff threshold but more importantly, as seen in Figure 4.5 the prediction overlap drops strongly with quite small changes from the value 9 used. Already the change of 0.5 in the cutoff makes the overlap drop to half. An increase or decrease of one makes the results almost unrelated.

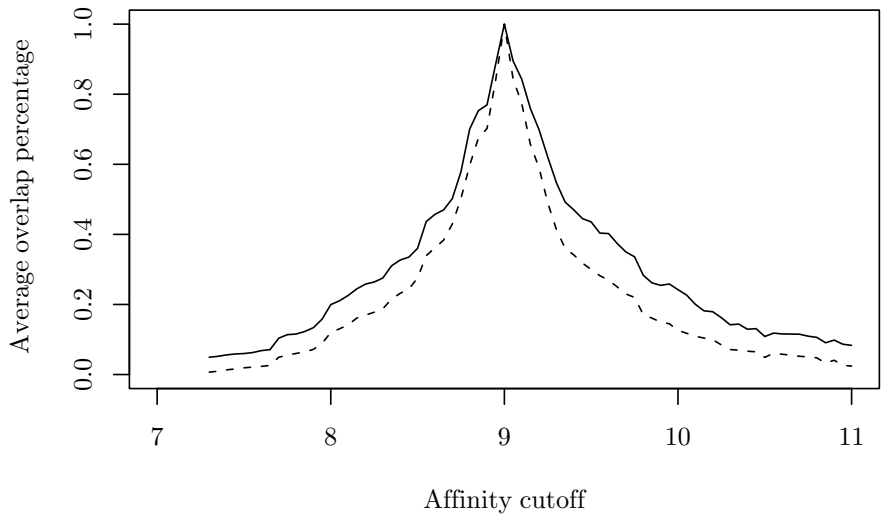

Figure 4.5: Effect of altered binding site cutoffs. Solid line for DNA overlap percentage and dashed line for site overlap percentage.

The sensitivity about the set of transcription factor motifs used depends strongly on the promiscuity of the motifs. As expected if the site occurs often in the input sequences its removal has a strong effect on the overlaps of the CRM predictions. Removal of the most common individual motifs, each counting for more than 3% of the binding sites, can drop the DNA overlap down to 70% and the site overlap down to 60%. On the other hand, removing individual motifs accounting for less than 3% of the sites, rarely (once out of 116 motifs) drops the DNA overlap below 90%.

Even though the number of binding sites correlates strongly with the importance of the motif, the relation is not perfect. Many of the transcription factor binding motifs in the dataset are redundant due to similar DNA binding domains in the transcription factor proteins and removing the binding motifs individually from the analysis results in only slight variation in the results.

The effects of the removal of individual motifs is not completely cumulative since 66% (83 out of 125) of the rarely occuring motifs can be removed with the DNA overlap staying above 50%. The DNA overlap stays above 90% after removal of 38% of the rarest motifs.

The choices for the set of transcription factor binding site motifs and the affinity cutoffs have strong effects on the resulting cis–regulatory modules. Making the correct choices requires intimate biological knowledge of the system under study and the choices have implications also on the stability of the results and the speed of the computation. These detailed parameter settings are highly context specific and they are left to the end user of EEL.

4.8 Conclusions

Here we have shown how to find parameter values for the conserved cis–regulatory module model, which provide good separation between the orthologous and the non–orthologous DNA sequences. The parameter values for separating functional from non–functional enhancers are more difficult to find, possibly due to lack of training data which is unbiased with respect to sequence conservation.

The parameter μ, which controls the density of the conserved cis–regulatory modules, is quite sensitive to small numerical changes while the other two parameters ξ and ν tolerate an order of magnitude changes without completely changing the prediction result.

The sensitivity of the results to the set of DNA binding motifs depends on the promiscuity of the motif. A motif with many sites has much higher

importance than one with only few sites but binding motif redundancy allows stable prediction of the DNA location of the cis–regulatory modules even with missing motifs.

Chapter 5

Statistical significance of the conserved cis–regulatory modules

The methods of Chapter 3 locate and score the conserved cis–regulatory modules within a sequence region. Even though the enhancer element locator score is related to the free energy of the DNA binding complex, the unit of this score is unknown. It is also difficult to evaluate the absolute scale of the enhancer element locator scores so that one could say whether a conserved cis–regulatory module with a given score is somehow significant and deserves further analysis.

In frequentist statistics, the significance is related to the surprise of observing the observed value, a statistic. This surprise is measured as *p–value*, which is the probability of observing a statistic as extreme, or more, given that the null–hypothesis is true. The null–hypothesis is the uninteresting *status quo*, the case which we would like to disprove. In the case of conserved cis–regulatory modules, the null–hypothesis is that the sequences do not contain conserved cis–regulatory modules.

This chapter provides a statistical model for estimating the significance of the enhancer element locator scores from DNA sequences of different lengths. Section 5.1 first introduces Monte Carlo methods for estimating the statistical significance. Since the Monte Carlo methods suffer from high run–time requirement in practice, Section 5.2 shows how the null distribution of the enhancer element locator scores observed in independent, identically distributed sequences fits the extreme value distribution of the Karlin–Dembo form [Gum58, KA90, KD92]. Finally, Section 5.3 shows how the scores from evolutionarily related sequences can be modeled

with a Two Component Extreme Value distribution with a novel set of parameters [RFV84].

5.1 Significance estimation by direct Monte Carlo simulation

A simple Monte Carlo method can estimate the p–values $\rho = \mathrm{P}(S \geq t)$, where $S = EEL(x)$, for the conserved cis–regulatory module model. Sampling the null sequence distribution provides DNA sequence pairs x_i, $i = 1, \ldots, N$, that do not contain conserved cis–regulatory modules and EEL provides the enhancer element locator scores $s_i = EEL(x_i)$ for these sequences. These scores estimate the p–value of score t as

$$p_{\mathrm{MC}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{I}_{s_i \geq t} \tag{5.1}$$

where $\mathrm{I}_{s_i \geq t}$ is the indicator function that obtains value one if $s_i \geq t$ and zero otherwise. Equation (5.1) gives an unbiased estimate p_{MC} of the real p-value ρ since

$$\mathbf{E}p_{\mathrm{MC}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{E}\mathrm{I}_{s_i \geq t} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{P}(S \geq t) = \rho. \tag{5.2}$$

Unfortunately p_{MC} converges to its expectation only with linear pace with respect to the increase in the sample size N. The random variable p_{MC} has a variance

$$\mathrm{Var}(p_{\mathrm{MC}}) = \frac{1}{N^2} \sum_{i=1}^{N} \mathrm{Var}(\mathrm{I}_{s_i \geq t}) = \frac{1}{N^2} \sum_{i=1}^{N} \rho(1-\rho) = \frac{\rho(1-\rho)}{N} \tag{5.3}$$

which is quite large.

We can give a hint of the estimation accuracy by looking at the *confidence intervals* (CI) for the scaled, binomially distributed random variable p_{MC}. Figure 5.1 shows the expectation of p_{MC} with 95% confidence intervals for sample size $N = 10^5$. Clearly this number of samples is not enough to estimate even the relatively high p-values in the range of $10^{-4} - 10^{-5}$. As a rule of thumb one needs at least $10^{2-\log_{10}\rho}$ samples to estimate ρ accurately [BEKF05]. According to the normal approximation this results in 40% accuracy with 95% confidence.

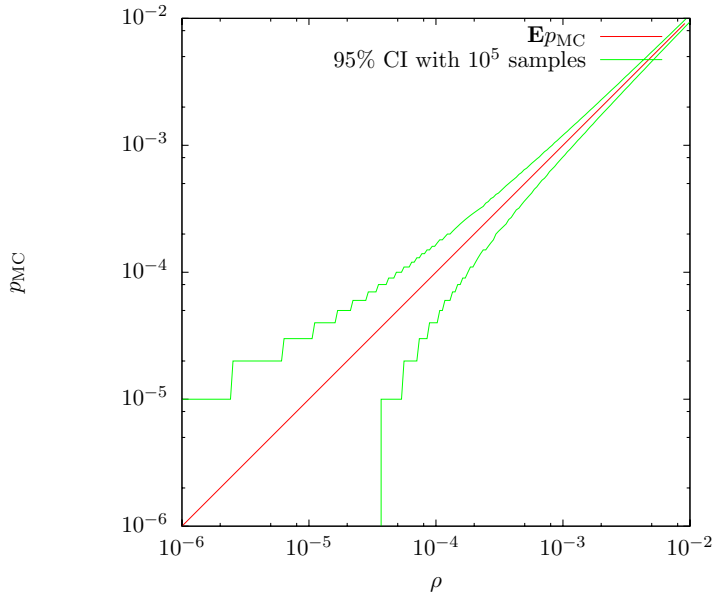

Figure 5.1: The theoretical accuracy of the Monte Carlo estimate of p–values.

5.1.1 Importance Sampling

Importance sampling could improve the convergence of the simple Monte Carlo method [HH64]. The importance sampling obtains more accurate estimate of the distribution tail by drawing more samples for the rare events. The method draws samples from an alternative distribution with probability density function $q(\cdot)$, that must have $q(x) > 0$ for all x for which $p(x) > 0$ [Har02]. Importance sampling for the p–value estimation weights the samples according to the *importance weighting factor* $w(x) = \frac{p(x)}{q(x)}$. The importance sampling procedure gives an estimator

$$p_{\mathrm{IS}} = \frac{1}{N} \sum_{i=1}^{N} w(y_i) \mathrm{I}_{S \geq t}, \tag{5.4}$$

with samples y_i from distribution $q(\cdot)$. Equation (5.4) provides an unbiased estimator of ρ since

$$\mathbf{E}_q w(x) \mathrm{I}_{S \geq t} = \sum_x \mathrm{I}_{S \geq t} \frac{p(x)}{q(x)} q(x) = \sum_x \mathrm{I}_{S \geq t} p(x) = \rho, \tag{5.5}$$

where the sums cover the whole support of $q(\cdot)$.

The choice of the weighting distribution $q(\cdot)$ is crucial for the efficiency of the importance sampling procedure. The variance of the estimator p_{IS} depends on how the distribution $q(\cdot)$ weights the 'important' tail of the original distribution $p(\cdot)$. The distribution $q(\cdot)$ must also enable efficient sampling which is not guaranteed for a complex distribution in a large space [Har02].

In general, the importance sampling method can give a similar variance estimator with an order of magnitude smaller sample size [HH64, Bun01]. The importance sampling could bring a notable speed improvement for the Monte Carlo method if pursued further. The improvement might be worthwhile if the null distribution for the enhancer element locator score would not allow analytical approximation.

5.2 Extreme Value Distribution of the alignment scores

To ease the computational burden of the Monte Carlo simulations of Section 5.1 it is beneficial to analytically approximate the null distribution of the enhancer element locator scores. An analytical null distribution provides a quick method for evaluating the statistical significance of the enhancer element locator scores and the formulas can take care of some of the individual features of the alignments, such as the lengths of the sequences.

The best known rigorous work on the statistical significance of local alignment is the result by Karlin and Dembo which show that the scores of the gapless local DNA alignment are distributed according to the Gumbel flavor of the Extreme Value Distribution [KD92, KA90, Gum58]. More exactly, consider a scoring scheme, which has an expected negative score and some probability of a positive score for an alignment of two characters. For gapless local alignment of independent, identically distributed (*IID*) DNA sequences of lengths n and m the probability of the highest score X to exceed a threshold t is approximately

$$\mathrm{P}(X > t) \approx 1 - e^{-\gamma nmp^t} \tag{5.6}$$

where γ and p depend on the scoring system used and they are found by solving two simple equations. This formula is used in the well known sequence alignment program BLAST to annotate the local alignments with e-values, which is the expected number of alignments scoring as good or better by random chance in the database of given size [AGM$^+$90]. Some authors argue that Equation (5.6) holds approximately also for Smith–Waterman local alignment allowing gaps [SW81, Mot92, WV94]. Equation (5.6) is

used heuristically also in the newer gapped versions of BLAST even though there is no theoretical justification for this [AMS+97].

The intuition behind the Karlin–Dembo formula in Equation (5.6) is in the extreme value theory of normal random variables. For a series of N independent, identically distributed normal random variables $(X_1, \ldots, X_N)$, the maximum $M_N = \max X_i$ is known to be distributed according to the Gumbel distribution with

$$\mathrm{P}(M_N \leq t) = \mathrm{P}(X_1 < t)^N \simeq e^{-KNe^{\lambda(t-\mu)}}, \tag{5.7}$$

for large N with some parameters K, λ and μ [Gum58]. The alignment algorithms can be thought as computing the maximum over the random variables that give the score of the best local alignment beginning from a certain pair of characters in the sequences to be aligned. Each of these nm random variables can be considered (approximately) normal as sums of independent, identically distributed random variables. We apply these heuristic considerations also to the enhancer element locator scores to obtain Equation (5.6) for their distribution even though the binding sites do induce some Markov–dependency to the binding site sequences.

Here we show that Equation (5.6) fits the enhancer element locator scores from independent IID DNA sequences. The approach will be empirical in the sense that we show that the data from simulations fit the model well but we are not able to provide formal proofs of our claim. We show the applicability of the Extreme Value Distribution by fitting it to the enhancer element locator scores obtained from IID nucleotide sequences.

The Monte Carlo estimated empirical cumulative distribution function provides us with a simple way of estimating the parameter values for the Extreme Value Distribution from the simulated data. By

$$\ln(-\ln(\mathrm{P}(S < t))) - \ln(nm) \approx \ln\gamma + t\ln p \tag{5.8}$$

the parameters γ and p can be estimated easily with the linear least squares methods.

First we show that the empirical cumulative distribution function of the enhancer element locator scores from independent DNA sequences with IID nucleotide distribution follows the form of Equation (5.6). Figure 5.2 displays the Monte Carlo estimated empirical cumulative distribution function $\mathrm{P}_e(\cdot)$ of the enhancer element locator scores on a set of 2000 pairs of IID sequences of length 2^{13} bp. The transform $\ln(-\ln(\mathrm{P}_e(S < t)))$ follows

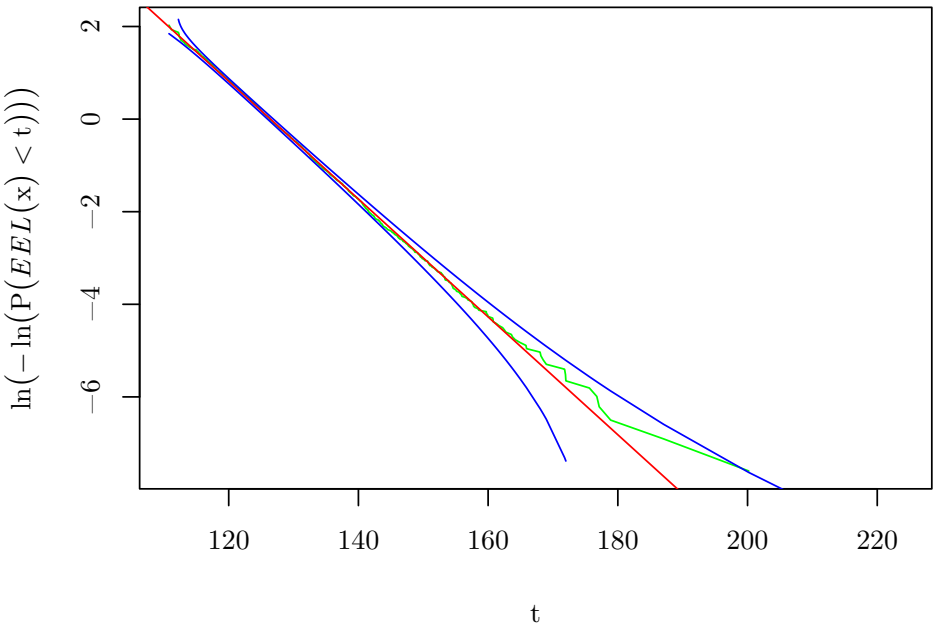

Figure 5.2: Empirical cumulative distribution function (green line) of the enhancer element locator scores in independent pairs of IID DNA sequences of length 128kbp. Least squares fitted analytical distribution (red) and 95% confidence intervals (blue) from normal approximation are provided.

a remarkably straight line with respect to score t and the least squares fitted line explains (coefficient of determination, r^2) 99.8% of the variation. The 95% confidence intervals cover practically all of the simulated data.

The appropriateness of the *length normalization*, the product nm in Equation (5.6), for the enhancer element locator scores can be seen in Figure 5.3 where we have plotted $-\log_{10} \mathrm{P}_e(S > t)$ against the theoretical probabilities from Equation (5.6) with fixed parameter values γ and p. The logarithm is there to show the fit on the important low probability region of the distributions. The least squares fit of the line in Equation (5.8) explains more than 99.9% of the variation. None of the sequence lengths show an appreciable amount of deviation from the theoretical line with confidence intervals. The expected edge effects are not seen in this length range and the shorter sequences are of no interest since we are targeting the analysis of long sequences of several hundred thousand bases.

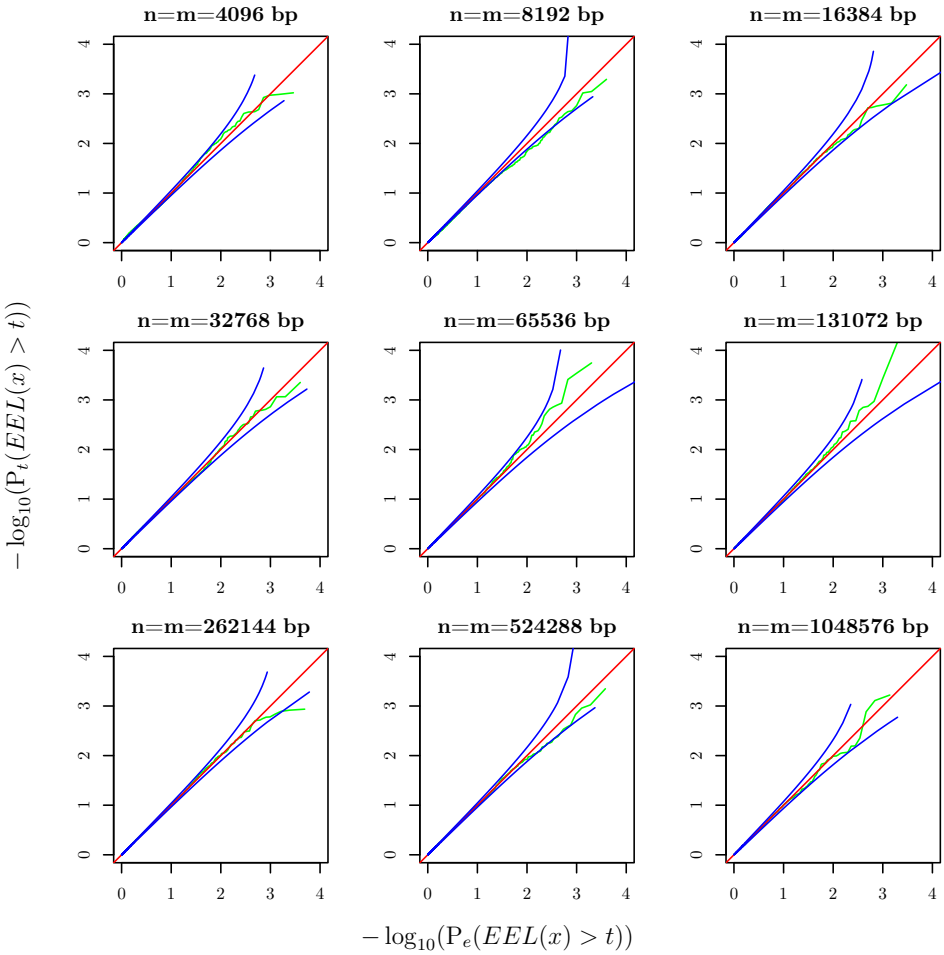

Figure 5.3: Quantile–to–Quantile plots of the least squares fitted analytical distribution versus the Monte Carlo estimated distribution (green line) for independent IID DNA sequences of indicated length. Optimal fit (red) and 95% confidence intervals (blue) from normal approximation.

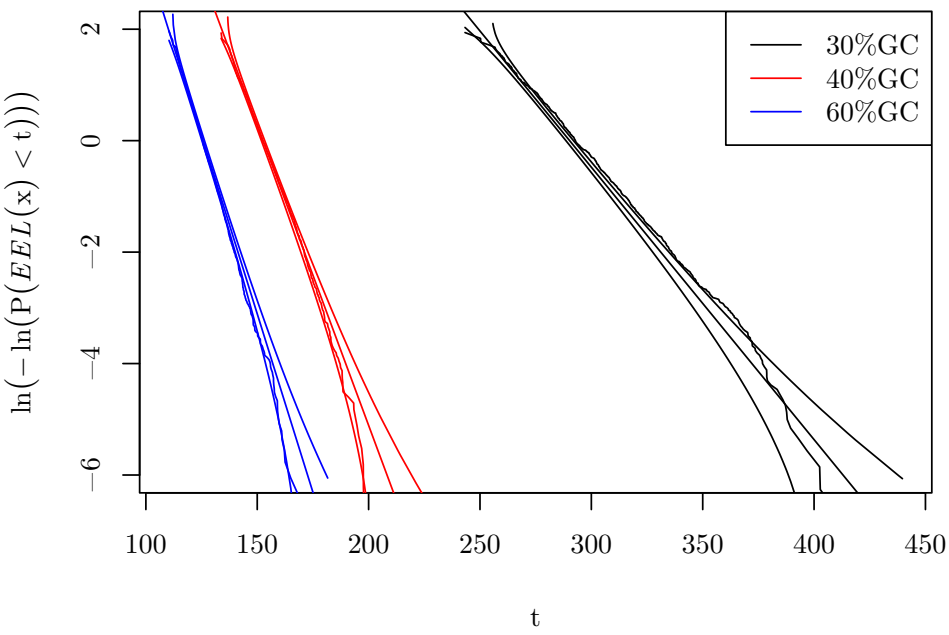

Figure 5.4: Monte Carlo estimated cumulative distribution functions, least squares fitted analytical distributions and 95% confidence intervals for the scores from independent IID sequences of 128kbp with varying GC contents.

Somewhat unexpectedly the proportion of nucleotides G and C in the analyzed sequences has a strong effect on the score distribution. Figure 5.4 shows $\ln(-\ln(\cdot))$ plots for IID sequences with GC contents of 30, 40 and 60 percent (The graph with 50% GC is in Figure 5.2). It is apparent that the scores tend to be significantly larger when the GC content of the sequences drop.

The effect of the GC content is likely due to the specific set of transcription factor binding motifs that were used in the analysis. The set of motifs that were used in the simulations and in the genome-wide predictions have a total GC content of 55% which makes the set slighly GC enriched. When looking at individual motifs, only 44% were GC enriched so it seems that the effect of the GC content on the scores comes from an added number of sites in the sequence.

5.3 Two Component Extreme Value distribution for evolutionarily related sequences

The probability model in Equation (5.6) does not fit the enhancer element locator scores from orthologous DNA sequences. The reason is that on the orthologous DNA sequences, the conserved cis–regulatory modules tend to lie on the about n pairs of orthologous nucleotides instead of on the nm independent pairs of nucleotides (without loss of generality we assume $n \leq m$). This results in two types of independent random variables competing to obtain the maximum enhancer element locator score: the nm random variables $S_a^{(i,j)}$ from the independent, non–orthologous, positions and the n variables $S_b^{(i)}$ from the dependent, orthologous, positions.

The extreme values $S_a = \max S_a^{(i,j)}$ and $S_b = \max S_b^{(i)}$ of these variables are independent and both distributed according to the Gumbel distribution but with different parameter values. While the distribution of S_a is given in Equation (5.6), the variable S_b is distributed as

$$\mathrm{P}(S_b > t) \approx 1 - e^{-\beta n q^t} \tag{5.9}$$

for some parameter values β and q.

Our interest is concentrated on the highest enhancer element locator score $S = \max(S_a, S_b)$ from a pair of orthologous DNA sequences. Because of independence this distribution is given by

$$\mathrm{P}(S < t) = \mathrm{P}(S_a < t)\mathrm{P}(S_b < t) = e^{-(\gamma n m p^t + \beta n q^t)} \tag{5.10}$$

which is the *Two Component Extreme Value* distribution (TCEV) [RFV84]. The four parameters of the TCEV distribution, γ, β, p and q, can be estimated with non–linear least squares regression. Table 5.1 gives the parameter value estimates for simulated orthologous sequences on evolutionary distances 0.45 and 0.6 substitutions per site (the ancestral sequence was IID with 50% GC). These distances were selected to illustrate the TCEV distribution on evolutionary distances relevant in mammalian comparative genomics. As expected the relative weight $\frac{\gamma}{\beta}$ of the independent component S_a increases with the evolutionary distance.

The fit of TCEV distribution for the enhancer element locator scores is seen in Figures 5.5-5.7. Figure 5.5 shows $\ln(-\ln(\mathrm{P}(S < t)))$ with respect to the threshold t in sequences neutrally evolved to distance 0.45 substitutions per site. The lines in the figure stand for the Monte Carlo estimated distribution (green), the fitted Karlin–Dembo (orange) and the TCEV distribution (red) with 95% confidence intervals (blue) for sampling from the TCEV

Parameter	0.45 subs/site	0.6 subs/site
γ	$1.848 \cdot 10^{-9}$	0.0010562
β	0.02018	0.012453
p	0.96481	0.874131
q	0.94681	0.931166

Table 5.1: Two Component Extreme Value distribution parameters.

distribution. Some features are noted when comparing Figure 5.5 to similar Figure 5.2 from the IID sequences. These are the shifted horizontal axis and the slight upward bend of the fitted TCEV distribution. The horizontal shift is expected as the sequences are better conserved. The upward bend is slight and is most visible on the low scoring and low probability region that includes the cis–regulatory modules on the independent non–orthologous positions.

The practical importance of the TCEV distribution is illustrated with the Quantile–to–Quantile plots in Figure 5.6 and 5.7. These figures compare the Monte Carlo estimated distribution to the fitted analytical distributions on sequences derived from ancestral sequences of lengths $n = m = 32000$, 128000, 256000 and 512000. Figure 5.6 shows the distributions for the sequences with 0.45 substitutions per site (0.225 substitutions each from the ancestral sequence) and Figure 5.7 shows the distributions for the sequences with 0.6 substitutions per site. It can be noted that neither the quadratic length normalization of Equation (5.6) (dotted magenta line) nor the linear length normalization of Equation (5.9) (dashed green line) fit the data on every sequence length. The TCEV distribution (solid orange line) lies closest to the diagonal almost uniformly, i.e., fits the data well.

Comparison of Figure 5.6 and 5.7 reveals the gradual shift from the linear length normalization of the orthologous nucleotides to the quadratic length normalization of the non–orthologous nucleotides as the evolutionary distance increases. When the evolutionary distance increases even more, e.g., 0.9 substitutions per site, the orthologous sequences become essentially independent and the quadratic length normalization takes over, i.e., $\beta = 0$. In small evolutionary distances, on the other hand, the sequence length itself becomes the limiting factor for S and all Extreme Value considerations fail because $S_b^{(i)}$ are no longer (approximately) identically distributed. This situation can possibly be avoided by choosing another set of parameter values for the conserved cis–regulatory module model which would make the expected EEL score negative also on the orthologous nucleotides.

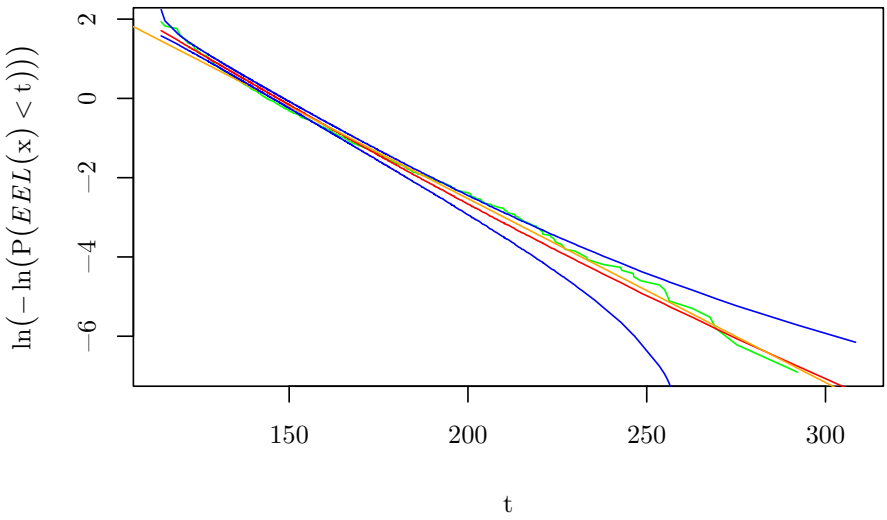

Figure 5.5: Cumulative distribution function of the scores from sequences of that have evolved neutrally to similar distance as human and mouse. The half–way ancestral sequence of length 128kbp. Monte Carlo estimated (green) and fitted TCEV (red) and Karlin–Dembo (orange) distribution are given. Blue lines show the 95% confidence intervals for Monte Carlo estimation of the TCEV distribution.

5.4 Conclusions

The significance of the conserved cis–regulatory modules found with the methods of Chapter 3 is a crucial questions for the practical applicability of the enhancer element locator tool. While the biological significance has previously been shown with the transgenic constructs and in-situ hybridizations, it has remained difficult to asses the overall quality of the predictions by automatic or non–expert means [VJM$^+$07].

The main difficulty with assessing the statistical significance of the conserved cis–regulatory modules is the choice of an appropriate null hypothesis. Depending on the users' biological insight and exact application field, the null hypothesis can take into account such things as isochores (i.e., GC content

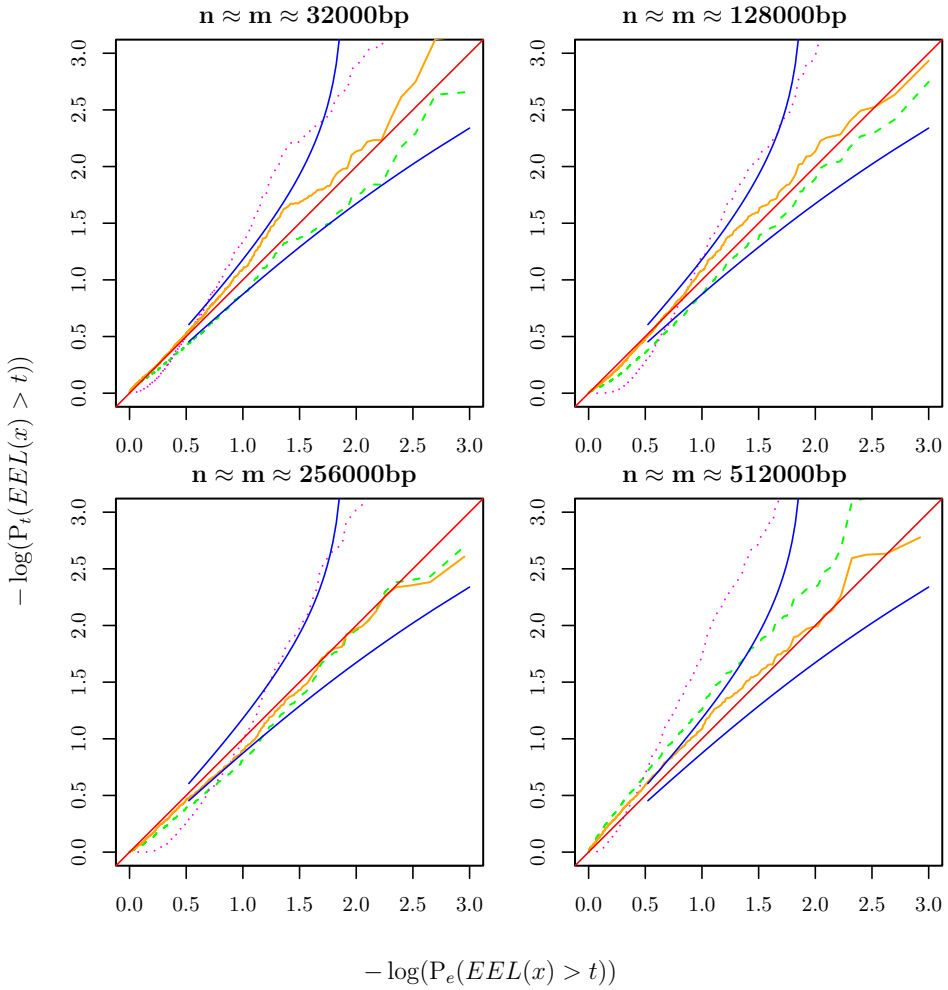

Figure 5.6: Quantile–to–Quantile plot of the least squares fitted analytical distributions versus the Monte Carlo estimated distribution for sequences that have evolved neutrally with 0.45 substitutions per site, have GC content of 50% and share a half–way ancestral sequence of indicated length. Diagonal (red) is the optimal result with 95% confidence intervals (blue) from normal approximation. The dotted magenta line is the best fit for Karlin–Dembo formula Eq. (5.6), the dashed green line is the best fit for Eq. (5.9) and the solid orange line is the best fit for the TCEV distribution in Eq.(5.10).

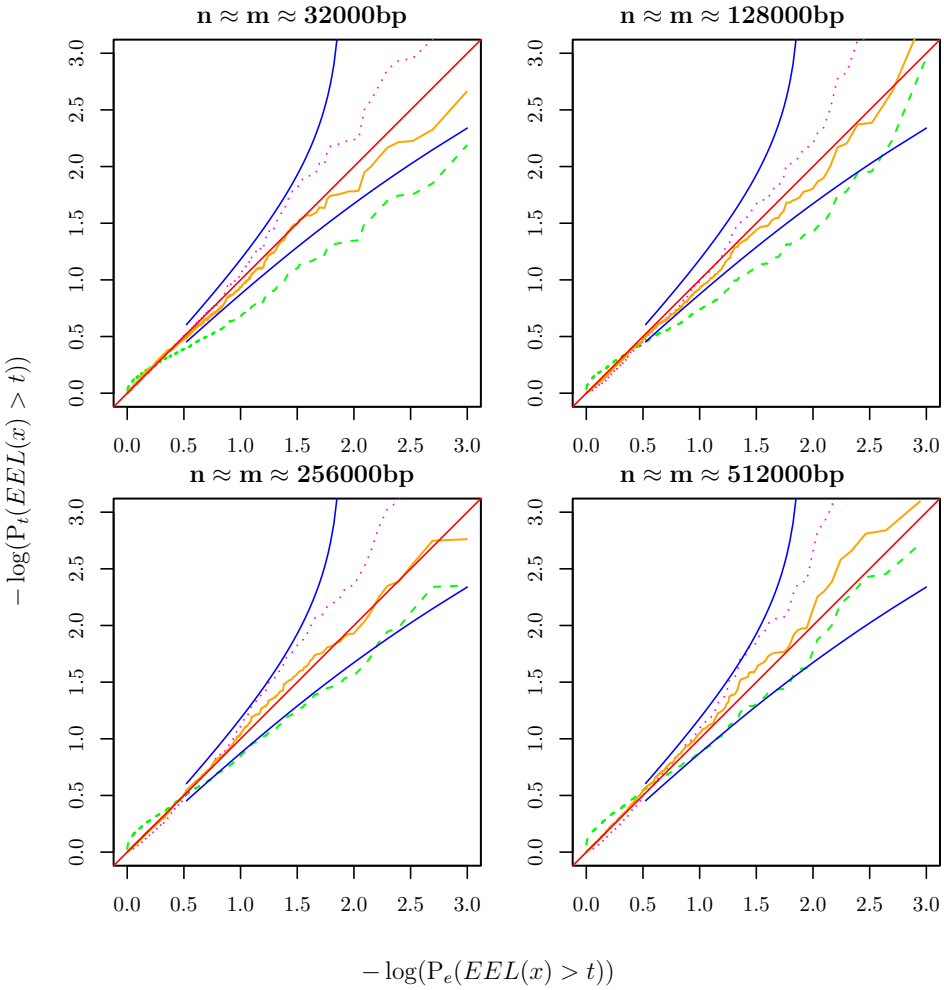

Figure 5.7: Quantile–to–Quantile plot of the least squares fitted analytical distributions versus the Monte Carlo estimated distribution for sequences that have evolved neutrally with 0.6 substitutions per site, have GC content of 50% and share a half–way ancestral sequence of indicated length. Diagonal (red) is the optimal result with 95% confidence intervals (blue). The dotted magenta line is the best fit for Karlin–Dembo formula Eq. (5.6), the dashed green line is the best fit for Eq. (5.9) and the solid orange line is the best fit for the TCEV distribution in Eq.(5.10).

variation), neutral conservation and varying evolutionary rates in different parts of the genome [Kim68, WSL89].

The neutral evolutionary conservation of the DNA sequences results in higher EEL scores and slightly altered score distribution as compared to unrelated DNA sequences. These effects are modeled with the Two Component Extreme Value distribution which separately considers the similarities between orthologous and non–orthologous nucleotides and provides a method to give accurate p–values for the enhancer element locator scores from sequences of different lengths. Since the evolutionary distance in substitutions per site vary across species and even across genomic locations, it is necessary to use different TCEV parameter values with predictions made from different sequences.

Maybe the most inherent problem with the probabilistic modeling in this setting is the variation of the input parameter values. Since only about 10% of the transcription factor binding motifs are known the set of binding motifs is bound to grow and each additional binding motif will invalidate the parameter values of the null distribution [VSDB$^+$06].

The methods provided in this chapter and especially the ones in Section 5.3 give us good tools for assigning the p–values for the conserved cis–regulatory modules. The computing power required for estimating the parameter values of the TCEV distribution is still fairly reasonable in the context of the genome–wide predictions, where heavy computation is needed for analyzing many sequences with identical parameter values and transcription factor motifs.

Chapter 6

Regulatory mechanism discovery from whole genome predictions

Microarrays and other high throughput technologies find the expression levels for all genes in a given sample [SSDB95]. The obtained data is often used to cluster the genes according to the similarity of their expression levels [ESBB98]. This cluster analysis provides sets of co–expressed genes and it is natural to hypothesize that the co–expressed genes are also co–regulated. The interesting question now is to identify the transcription factors regulating the set of co–regulated genes.

The spatial and temporal regulation of gene expression in multicellular organisms pose a problem for the high throughput expression measurement technologies. It is laborious and costly to analyze the expression of all genes in many tissues. Especially many of the interesting developmental stages are practically inaccessible to microarray analysis. These difficulties could be alleviated by finding the genes that putatively have expression patterns similar to a single gene of interest. These hypotheses could then be tested on the more accurate but lower throughput in-situ assays.

In this chapter we provide methods for analyzing the conserved cis–regulatory modules within a species. First, in Section 6.1 we develop a statistical test for identifying the overrepresented transcription factor binding sites within the conserved cis–regulatory modules of a given set of genes. These sites are likely to bind the transcription factors regulating the given genes.

In Section 6.2 we analyze the distance metrics between individual conserved cis–regulatory modules. The conserved cis–regulatory modules are clustered according to a distance metric and some of the clusters are found to be enriched of modules associated with genes that are expressed in specific tissues.

6.1 Finding a common regulatory mechanism

Probably the most common question a biologist has, after conducting a laborious and expensive high–throughput microarray experiment, is 'What transcription factor is causing these genes to change their expression?' While it might be possible to find the answer by looking for a transcription factor changing its expression, the answer is quite likely hidden in the cis–regulatory modules around the genes.

Most previous systems considering regulators of sets of mammalian genes have concentrated on the proximal promoters [WS04, ATC+03, CNM+06, KM04, HB06]. The methods analyzing the distal cis–regulatory elements usually concentrate on the DNA conservation so the approach does not provide tools for predicting the regulatory factors [PHB05, PPS+06, PLNO07]. When the cis–regulatory module analysis uses conserved transcription factor binding sites, the question of transcription factors explaining the gene expression arises naturally [BBC+06]. Our main contribution on the regulatory mechanism finding is the integration of several pairwise analyzes to a single test in Section 6.1.2.

6.1.1 Using one pairwise comparison

In statistical terms, we would like to answer the question, 'which transcription factors bind the CRMs of the genes of interest more often than expected by chance?' For a pairwise comparison of two genomes, this question can be answered with Fisher's exact test for proportions [PUBV02, Fis22]. Simply put, if we have C co–regulated genes in a genome of G genes and the transcription factor of interest is present in the cis–regulatory modules of T genes, then the number of co-regulated genes with the transcription factor is X which, if the factors in the cis–regulatory modules are chosen randomly, is distributed according to hypergeometric distribution. The situation is depicted in the contingency Table 6.1.

A technical caveat with the presented gene counts, apart from deciding what are the co–regulated genes and the real CRMs, arises with the size of the 'genome' G. Since the exact number of genes in any genome is not known and we have data only from orthologous genes, the number G should be the number of genes that have orthologs in both species considered. Disregarding this detail, and using e.g., the size of the union of the two genomes, results in a bias towards genes without the factor that are not co–regulated.

The significance of observing k or more co–regulated genes with a binding site for the given transcription factor is assessed with the p–value which is

	With factor	Without factor	
Co-regulated	X	$C - X$	C
Not co-regulated	$T - X$	$G - C - T + X$	$G - C$
	T	$G - T$	G

Table 6.1: Contingency table of conserved binding sites. Number of co-regulated genes C. Total number of orthologous genes G. Genes with conserved site for the factor T. Co-regulated genes with the given factor X.

the probability

$$\mathrm{P}(X \geq k) = \sum_{x=k}^{\min\{C,T\}} \frac{\binom{C}{x}\binom{G-C}{T-x}}{\binom{G}{T}}. \tag{6.1}$$

That is the probability of observing k or more co-regulated genes with binding sites for the given factor given that the sites are allocated independently for the genes.

The significance evaluation of 2×2–contingency tables is classical statistics and it is well known that Equation (6.1) can be approximated with χ^2–statistic with one degree of freedom if all of the cells in Table 6.1 are large. Since this is not commonly the case in our scenario, Equation (6.1) should be evaluated exactly whenever feasible. Tables with fairly large values can be evaluated quickly with modern computers using efficient algorithms for Fisher's exact test [CFJ93, MP86, RD06].

6.1.2 Using multiple pairwise comparisons

The contingency table assignment of Section 6.1.1 applies when searching common mechanisms from a pair of species but today there are many full genome sequences available for genome–wide comparisons that can provide added power to the mechanism discovery.

Let us now consider a situation with m pairwise comparisons each sharing one species, say human. Similarly to the two species case, let G be the number of genes that have orthologs in *all* considered species and C the number of co–regulated genes (that have orthologs in all species). Now let $G_{\overline{x}}$ for $\overline{x} = (x_1, \ldots, x_m) \in \{0,1\}^m$ be the number of genes that have a conserved binding site for the given transcription factor in exactly those pairwise comparisons i for which $x_i = 1$. Similarly $C_{\overline{x}}$ is the number of genes that are co–regulated and have a conserved binding site for the given transcription factor in the comparisons i for which $x_i = 1$. It follows that $G_{\overline{x}}$ partitions the G orthologous genes into distinct sets (and similarly $\sum C_{\overline{x}} = C$).

The contingency table to be tested now is 2×2^m and the values in the co–regulated row are $C_{\overline{x}}$ and in the not co–regulated row $G_{\overline{x}} - C_{\overline{x}}$ for all $\overline{x} \in \{0,1\}^m$. The contingency table for the case of two genome–wide alignments between three species is given in Table 6.2.

	11	01	10	00	
Co-regulated	C_{11}	C_{01}	C_{10}	C_{00}	C
Not co-regulated	$G_{11} - C_{11}$	$G_{01} - C_{01}$	$G_{10} - C_{10}$	$G_{00} - C_{00}$	$G - C$
	G_{11}	G_{01}	G_{10}	G_{00}	G

Table 6.2: Contingency table of conserved sites. Number of co-regulated genes C. Total number of genes G. Genes with conserved site in given species $C_{\overline{x}}$.

Fisher's exact test and the χ^2–test apply for this table as well as Table 6.1. Extra care should be taken if using the χ^2–test since the table comparing data from many species grows exponentially with the number of considered pairwise alignments and the set of co–regulated genes is quickly dispersed around the table thus providing small expected values for each cell and so violating the assumptions of the χ^2–test. If Fisher's exact test is too complex to compute, a simulation solution usually provides reasonable approximation of the p–values [Pat81].

6.1.3 Results

The mechanism discovery method of Section 6.1.2 was tested in recovery of the regulators of tissue–specific expression. We used a set of microarray measurements of gene transcription in 60 healthy human tissues to discover the genes that are differentially expressed in a specific part of the body, thus showing a somewhat complex regulatory pattern [SWB+04]. The expression data was downloaded from the ArrayExpress [PKS+07] (Accession code E-AFMX-5). The raw data was summarized with gcrma by using the non–standard probesets made for measuring expression of Ensembl genes (Probeset version 8) [WIG+03, DWB+05]. The differentially expressed genes were located with the Limma library from the Bioconductor/R package [vHHG04, Smy04, GCB+04, RD06]. The genes were deemed tissue specific when they had p–value less than 10^{-6} for tissue–specific expression in some tissue after Holm's correction [Hol79].

We analyzed the over-representation of the 112 transcription factors in the pairwise comparisons between human and dog, cow and mouse. The mechanism discovery was executed for all 60 tissues independently. All putative cis–regulatory modules scoring 350.0 or more, and having a length

of less than 2kbp were considered functional. The p–values obtained from Fisher's test were corrected with Holm's method and tests with a corrected p–value less than 0.01 were deemed significant. Table 6.3 on pages 80–81 lists the significant tissue–motif pairs.

Some of the mechanism predictions are corroborated by literature. For example *Snail* homologs *Scratch1* and *Scratch2* are active in neuronal differentiation and are actively transcribed in developing and adult brain [NWS$^+$01, MN06]. Also *MEF2C* (*SQUA* ortholog) isoform and *c-FOS* are active in brain [MCY$^+$93, SSC88]. The results also recover the brain associated regulatory activity of the Acute Myeloid Leukemia gene *AML-1* which has recently been implicated in quite unexpected role in human brain tumors [PSB$^+$02]. Some of the used and significant binding motifs, like the MYB-ph3 from the plant Petunia Hybrida and the Broad Complexes from Drosophila, do not have clear human orthologs and the significance of these results is very difficult to assess.

Table 6.3 is heavily biased towards brain–related factors and tissues due to a similar bias in the expression data set used and also due to the complex regulation of the mammalian neural system. The mechanism discovery method seems to find reasonable correlations also for other than brain–related tissues. For example the smooth muscle regulator MEF2 is correlated with uterine tissue and for the lung tissue, the most correlated factor ($p < 10^{-4}$) is c-ETS-1 which is known to be expressed in lung carcinomas [BGD$^+$95, MBMO95]. The method presented provides a quick and sound method for making hypothesis for mechanisms of altered gene expression.

6.2 Measures for cis–regulatory module similarity

Even though the enhancer element locator tool is based on the conservation of the existence and the order of the transcription factor binding sites within two species, one is often interested in finding similarities between cis–regulatory modules for different genes in a single species. The genes expressed in similar fashion in a single species are likely to be regulated by similar cis–regulatory modules but since their cis–regulatory elements have evolved independently we should not assume that the exact order of the transcription factor binding sites would be the same. For all but very recent gene duplications, it is not realistic to assume complete structural conservation of the cis–regulatory modules within species.

This problem setting is somewhat similar to matching CRMs with known structure to the DNA or de novo inference of CRMs for sets of co–regulated

Tissue	Factor	p–value	Corrected p–value
fetal brain	Broadc3	4.2e-10	2.8e-06
occipital lobe	SQUA	5.4e-10	3.6e-06
occipital lobe	Broadc3	1.3e-09	8.5e-06
occipital lobe	Athb-1	1.4e-09	9.2e-06
occipital lobe	TCF11-MafG	1.4e-09	9.5e-06
occipital lobe	HNF-3beta	1.8e-09	1.2e-05
occipital lobe	Sox-5	2.1e-09	1.4e-05
occipital lobe	Tcf4	6.0e-09	4.0e-05
fetal brain	HLF	8.3e-09	5.6e-05
fetal brain	SOX-9	1.9e-08	0.00013
occipital lobe	MNB-1A	2.9e-08	0.00020
fetal brain	Hunchback	3.1e-08	0.00021
fetal brain	AML-1	3.1e-08	0.00021
fetal brain	cEBP	4.2e-08	0.00028
occipital lobe	SOX-9	5.9e-08	0.00039
prefrontal cortex	FREAC-2	8.2e-08	0.00055
fetal brain	c-FOS	8.4e-08	0.00056
occipital lobe	FREAC-2	8.4e-08	0.00056
prefrontal cortex	Athb-1	8.5e-08	0.00057
fetal brain	COUP-TF	8.5e-08	0.00057
fetal brain	Sox-5	9.0e-08	0.00061
occipital lobe	Gfi	1.2e-07	0.00079
occipital lobe	Irf-1	1.3e-07	0.00088
fetal brain	Snail	1.4e-07	0.00095
occipital lobe	HNF-1	1.4e-07	0.00096
fetal brain	Tcf4	1.5e-07	0.00099
prefrontal cortex	Chop-cEBP	1.5e-07	0.00099
uterus	cEBP	1.5e-07	0.00103
fetal brain	Chop-cEBP	1.7e-07	0.00114
uterus	MEF2	2.0e-07	0.00135
prefrontal cortex	SOX-9	2.0e-07	0.00137
amygdala	FREAC-2	2.2e-07	0.00148
uterus	Irf-1	2.3e-07	0.00152
occipital lobe	Broadc4	2.4e-07	0.00164
fetal brain	Broadc2	2.6e-07	0.00171
fetal brain	FREAC-2	2.9e-07	0.00195
prefrontal cortex	TCF11-MafG	3.0e-07	0.00202
fetal brain	MYB-ph3	3.9e-07	0.00261
occipital lobe	Nkx	4.1e-07	0.00271

Continued...

Tissue	Factor	p–value	Corrected p–value
occipital lobe	SOX17	4.3e-07	0.00285
fetal brain	Nkx	4.4e-07	0.00294
prefrontal cortex	Broadc3	5.0e-07	0.00331
occipital lobe	Broadc2	5.6e-07	0.00372
occipital lobe	cEBP	6.7e-07	0.00446
prefrontal cortex	MEF2	7.1e-07	0.00475
cingulate cortex	NF-kB	7.4e-07	0.00491
fetal brain	Pax6	8.0e-07	0.00537
caudate nucleus	FREAC-2	8.2e-07	0.00549
occipital lobe	Hunchback	1.0e-06	0.00687
uterus	HNF-3beta	1.2e-06	0.00769
fetal brain	Tal1beta-E47S	1.2e-06	0.00831
prefrontal cortex	E4BP4	1.4e-06	0.00911
occipital lobe	jsmad5	1.5e-06	0.00985

Table 6.3: List of active transcription factors in certain tissues

genes [KW01, ZW04, HGCM02]. The following inquiry is mostly executed to provide additional tools for analysis of the enhancer elements located with the methods presented earlier in this work.

The cis–regulatory elements active in similar conditions are assumed to have similar transcription factor binding sites, hence the distance between two cis–regulatory modules may be measured as the distance between two sets of transcription factor binding sites. For our analysis, a transcription factor binding site is an object with class identity (the binding transcription factor), affinity and position. Of these, we choose to disregard the position because of the aforementioned lack of common ancestry.

The count of the binding sites and their total affinity are the two features that make two cis–regulatory modules similar to each other. Any measure of distance between the histograms of the transcription factor binding site counts could serve as a distance measure between two cis–regulatory modules.

The distance between histograms can be defined as almost any kind of common distance between vectors or distributions such as L_n, χ^2 or Kullback–Leibler–distance [PBRT99]. In image retrieval a popular distance measure is the quadratic form $\sqrt{(\bar{e}-\bar{h})^T A(\bar{e}-\bar{h})}$, where $\bar{e}, \bar{h} \in \mathbb{R}^{|\mathcal{F}|}$ are the binding site histograms mapped to a real space and $A \in \mathbb{R}^{|\mathcal{F}|\times|\mathcal{F}|}$ is a weight matrix [NBE+93]. An important special case of the quadratic form distance is the Mahalanobis distance

$$\mathrm{d}_M(\bar{e}, \bar{h}) = \sqrt{(\bar{e}-\bar{h})^T \Sigma^{-1} (\bar{e}-\bar{h})} \tag{6.2}$$

when $A = \Sigma^{-1}$ is the inverse of the co–variance matrix of the binding site histograms [Mah36]. A further special case is when the bins of the histogram are independent and the co–variance matrix is diagonal. Finally, when the variances of all of the bins are equal to one, the quadratic form distance reduces to the Euclidean norm.

Mahalanobis distance is intimately connected with linear and quadratic discriminant analysis [DMJRM00]. If the data x comes from one of two multivariate Gaussian distributions with means μ_1 and μ_2 and covariance Σ, then the optimal prediction for source distribution of x is given by $\arg\min_i \mathrm{d}_M(x, \mu_i)$.

The histogram methods have fundamental issues with the sparsity of the cis–regulatory modules. The good quality cis–regulatory modules (score at least 400 and length at most 2000bp) contain binding sites for 9 to 54 distinct transcription factors out of the total 112. The counts on the individual bins, i.e., the number of binding sites for an individual transcription factor on

one module, is at most nine. Sparsity of this data makes the straightforward histogram methods inappropriate for the task at hand.

Apart from counting the occurrences of the binding sites, it is also useful and necessary to consider the binding affinity of the sites [RKMV07]. These considerations suggest *a proportional affinity weighted mapping* of the conserved cis–regulatory modules to an Euclidean space.

With the proportional affinity weighted mapping we aim to give strong weight to the transcription factors with close to maximum affinity towards the cis–regulatory module. Specifically we map a conserved cis–regulatory module e to a vector $\bar{e}$ in the Euclidean space $\mathbb{R}^{|\mathcal{F}|}$. Each dimension of the space stands for one transcription factor and the components of the vectors range from zero to one representing the binding strength between the cis–regulatory module and the respective transcription factor. In detail, if the conserved cis–regulatory module e has k binding sites for transcription factor f, each having an affinity of proportion r_i of the maximum, the cis–regulatory module is represented by a vector $\bar{e}$ for which

$$e_f = 1 - \prod_{i=1}^{k}(1 - r_i^2). \tag{6.3}$$

The formula in Equation (6.3) has the desired properties that factors with close to maximum affinity sites get values close to one and having more sites strictly increases the value while it is difficult to obtain high values by having only sites with low relative affinity. After mapping the cis–regulatory modules to the Euclidean space, their distance is defined as the Euclidean distance between their representative vectors.

6.2.1 Cis–regulatory module clustering

To test the proportional affinity weighted mapping and the accompanied Euclidean distance, we clustered the high quality cis–regulatory modules around genes that are differentially expressed in the human tissues according to the data in Section 6.1.3. The genes were called differentially expressed with a false discovery rate of 0.01 to obtain large number of likely elements [BH95, BY01]. The method used to estimate the false discovery rate does not assume independence between the tests, since many of the samples in the expression data are from similar tissues or even contained in another [BY01] (e.g., 17 parts of the brain and whole brain and fetal brain.)

The CRMs were clustered with the Affinity Propagation algorithm which strongly resembles traditional k–centers clustering [FD07]. Given a similarity function $s(i, j)$ between data points i and j, the algorithm assigns each

data point i to an *exemplar* data point c_i maximizing the total similarity $\sum_{i=1}^{N} s(i, c_i)$ between the data points and the exemplars. The special case of self similarity $s(i, i)$ provides the *preference* of the data point i to act as an exemplar. These preferences affect the number of resulting clusters, i.e., the number of used exemplars. While the method is heuristic, the algorithm is very fast and the results are stable. Affinity Propagation seems to produce better clusterings than the traditional k–centers clustering method when measured with the within cluster similarity $\sum_{i \neq c_i} s(i, c_i)$.

We used Affinity Propagation with the mapping of Equation (6.3) and the Euclidean distance to obtain partition of cis–regulatory modules to 50 clusters. The number of clusters is chosen to be in the scale of number of tissues present in the expression data while accounting for some of the tissue type redundancy. The exemplar preferences were equal for all data points and they were selected with a binary search to provide the desired number of clusters. Of the 50 obtained clusters, 26 clusters were enriched (Fisher's exact test. False Discovery Rate 0.05 [BY01]) with cis–regulatory modules of genes with differential expression in a certain tissue. In total 70 tissue–cluster pairs were found significant and differentially expressed genes for 22 tissues were overrepresented in at least one cluster. Genes for some of the tissues were overrepresented in multiple, up to ten, clusters. Since the tissues and the gene expression patters are not independent, the clusters tend to be enriched with genes expressed in related tissues. These dependencies are taken into account when estimating the false discovery rate so the statistical conclusions from these experiments remain sound.

This clustering result shows that the signal recovered by the enhancer element locator is biologically relevant. Most of the overrepresented tissues are parts of the brain although pituitary, uterus, placenta and thyroid are present in some clusters.

The actual mapping of the cis–regulatory modules to the Euclidean space does not have an extreme effect on the clustering results. Trying out 20 exponents uniformly in the range $1 - 10$ in Equation (6.3) resulted in 15–26 significant clusters with no clear trend. Exponential mapping

$$e_f = 1 - \prod_{i=1}^{k} (1 - e^{t(r_i - 1)}) \tag{6.4}$$

with t values between one and 30 provided an even smaller number of significant enrichments with peak value 20 with $t = 5$ and decline to about 6 enriched clusters with the larger values of t. Several other methods for weighting the binding site occurrences did not reach the quality of mapping in Equation (6.3). These results suggest limited success in predicting

the tissue specificity of cis–regulatory modules, especially in tissues that are not characterized with well–known transcription factors with well–known binding motifs.

6.3 Conclusion

In summary, the multiple hypothesis corrected Fisher's test is sufficient for providing important mechanistic insight to the regulatory system of a set of co–regulated genes. The results, like the ones in Table 6.3, require watchful interpretation because the binding site motifs often do not uniquely define the binding transcription factor but merely a structural family of factors whose members have similar binding preferences and hence bind to similar sites [MAB07].

The mechanism recovery relies strongly on the availability of the correct binding site motifs. The methods for measuring protein–DNA binding affinities accurately in high throughput have recently been developed but current knowledge of the binding affinities is still sparse [BPQ+06]. This situation is expected to change in the near future with unbiased, high accuracy measurements of binding affinities for almost all human DNA binding transcription factors.

The between–CRM similarity measures are possibly even more hampered by the lack of correct binding affinities. Since many of the binding sites in the cis–regulatory modules are falsely predicted, either because of missing or incorrect affinity measurements, the binding–site–based similarity measures cannot be very accurate in predicting the correct regulatory behavior of complex cis–regulatory modules.

Chapter 7

Software for cis–regulatory module finding

The enhancer element locator method as described in Chapter 3 has been implemented in the EEL software which is freely available on the Internet under the GNU General Public License [Sta02, Sta91]. The user interfaces of the software is written in programming language Python which interfaces the time–consuming core algorithms written in C/C++ [vR96].

Using two different programming languages for the implementation is justified by the rapid application development allowed by Python and the time– and memory–efficient algorithms and data structures allowed by C++ and its Standard Template Library [SLL95]. The co–operation of the two environments is seamless as the Python interpreter is natively implemented in C and the Python/C API is in the absolute core of the Python reference implementation [vRDJ06].

The EEL software has been developed with two separate use cases in mind and both have been provided with their own user interfaces. One use case is when the user is interested in a few of his favorite sequences and prefers a simple point-and-click kind of interface. The second use case is for a more genomic approach where the user wants to compute the analysis for all genes of a given pair of species. This use case requires a capability for batch processing of the genes with no user interaction.

The rest of this chapter is organized as follows. At first, Section 7.1 introduces the user–friendly point-and-click kind of interface for the integrated and portable EEL software. Section 7.2 introduces the system that allows computing a large number of EEL analyzes in a batch–processing environment. Section 7.3 describes the methods to manage and publish the data obtained from large computations. Finally, Section 7.4 gives some concluding remarks.

7.1 User–friendly integrated software

To accommodate the needs of a casual user interested in analyzing few of his favorite sequences with EEL, the software provides a simple point-and-click user interface as seen in Figure 7.1. This interface is implemented with the Tkinter module which is part of the Python standard library and interfaces the Tk toolkit. Because of these standard modules, the whole software suite, including the look-and-feel, is portable to any computer system with a C++-compiler, Python and the Tk library. The software is running on Linux, Microsoft Windows and Apple MacOS X systems but it should be easy to port to any system from high–end mobile phones to mainframes and supercomputers [Nok05].

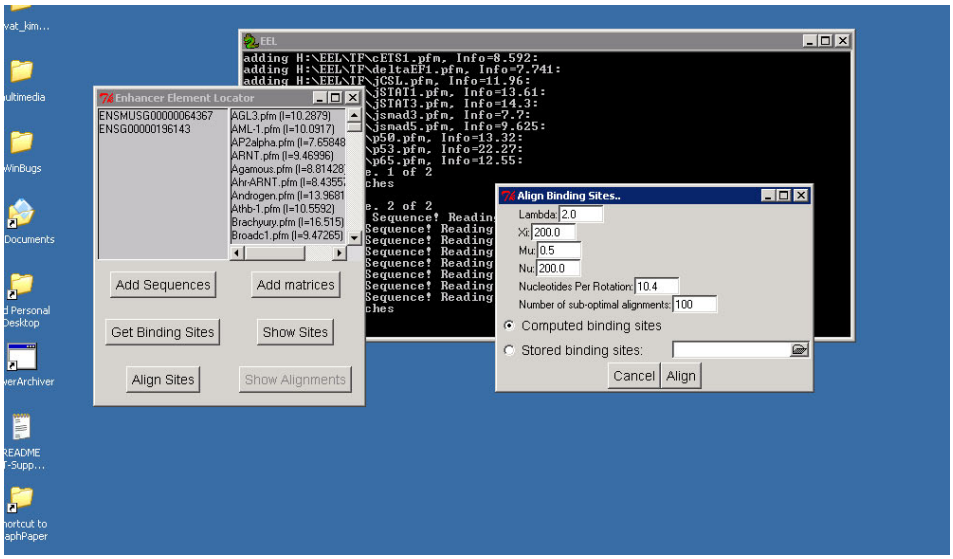

Figure 7.1: EEL point-and-click interface on Microsoft Windows

The graphical interface allows the user to execute the basic analysis operations with ease but more specialized commands have to be executed on the command line or via the console interface introduced in Section 7.2. Briefly explained, the user first adds the sequences and the binding motif matrices for the analysis by clicking the buttons on the left–most window in Figure 7.1. After choosing the input data, the user clicks on 'Get Binding Sites' and 'Align Sites' to do the actual analysis. The locations of the binding site matches and the final result, the conserved cis–regulatory modules, can be either viewed on the display or saved to a file for later inspection.

The graphical user interface has been designed to be straightforward so the user would need only minimal computer experience. The only choices the user has to make are the parameter values for the analysis that might be difficult to decide as seen in Chapter 4.

7.2 High performance clustering

The amount of data gathered in the biomedical research has grown exponentially for at least three decades and the leading edge research is moving more and more to the genomic scale where the experiments and the analyzes are made for full genomes or for all genes at once [KAA+07, PKS+07]. For bioinformatics software this shift in paradigm requires a support for distributed computing that can apply the methods developed to large datasets in an efficient manner.

The EEL software supports distributed computing in the sense that many individual EEL processes can be controlled to work each on their own set of input sequences. This is efficient and the work for analyzing e.g., all human and mouse ortholog genes can be distributed to an almost arbitrary number of processes. The granularity according to which the work is split into individual jobs is limited only by the number of orthologous sequences to be analyzed e.g., 23723 in the case of human and mouse from the Ensembl version 44 [HAB+07].

The clustered computing system is implemented in various scripting languages including Bash, Perl and Python [Ram02, WS00]. The actual computation work is done on regular Linux computers networked together as a cluster or a grid [Fos02]. The environment used at the Biomedicum Bioinformatics Unit uses Sun Grid Engine as the software controlling the batch queues but the scripts can easily be adapted to work with other batch scheduler systems as well [Gen02].

The input data management in the system is highly distributed so that up-to-date DNA sequences and annotations from the current Ensembl release is used. The diagram in Figure 7.2 shows the logical division of different parts of the system. Physically the parts can be located on one computer or be distributed around the world and be connected via the Internet. The price paid for the up-to-dateness of the data is the amount of work needed to maintain the distributed data management system in working order. This workload is mainly due to the third party databases and services that are typically under active development and can change from time to time.

The work flow of the system in Figure 7.2 is straightforward. The settings for the EEL program are given in configuration files that detail what

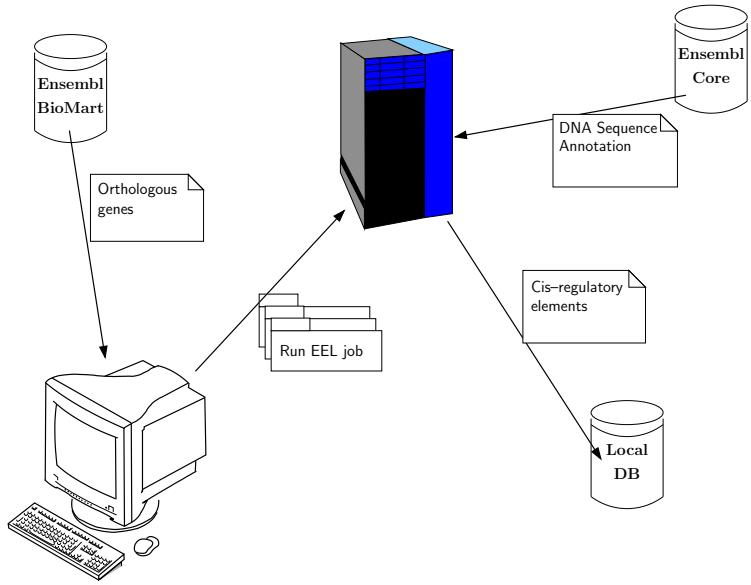

Figure 7.2: Coupled databases for running EEL on a computing cluster.

parameters and transcription factor binding motifs to use, what EEL commands to run, what preprocessing to do with the input sequences and where to get the input and store the output. These files are relatively static after the initial parameters have been decided.

When the settings files have been created, the process of computing the cis–regulatory module predictions for all genes of human and other species is very simple. A single Bash script downloads and presents to the user a list of species from the BioMart service and, after the user choice, downloads the list of orthologous genes [KKS+04]. The same script also provides an option to generate the (Bash) job scripts that run the actual analyzes and an option to submit the jobs to the batch queue.

To avoid the relatively high time cost of a job startup, each batch job analyzes several pairs of sequences (default 100 pairs). While running on a node of the computing cluster, the job script spawns two child processes that communicate with each other via named pipes. The first process is a Perl script that downloads the orthologous DNA sequences from the Ensembl database and does the requested preprocessing, e.g. masks the Ensembl annotated coding regions [HAB+07]. The second process is the EEL program that reads the preprocessed sequences via a named pipe and computes the cis–regulatory module prediction writing the result to a file in the given directory.

The asynchronous sequence fetcher and EEL analysis processes make the job script somewhat complex but it significantly speeds up the total throughput as the computation–intensive analysis process does not need to wait for the IO–intensive fetcher process. The implementation of the sequence fetcher in Perl is justified by the availability of the Ensembl Perl API which allows easy interfacing with Ensembl data from within Perl programs [HAB+07].

Finally, when all orthologous gene pairs have been analyzed, the output data is loaded from the files on the file system to a MySQL database which facilitates storing and warehousing the data [MyS07].

7.3 Data warehousing

After committing a significant computing power to obtain a large number of predictions from the genome–wide conserved cis–regulatory module analysis, it is important to make this information useful for the end user. While the data is safely stored in the relational database management system, it is not a particularly easy interface for most people with biologist backgrounds. There is a serious threat of creating a write–only database, a data tomb [Fay98]. To make the data generally available, we store the data in a data warehouse for easy access and we also provide an On-line Analytical Processing (OLAP) service utilizing the stored data [CD97].

7.3.1 Raw data queries

The detailed information about the cis–regulatory module predictions from various analyzes is publicly available via a web form query interface[1] seen in Figure 7.3. The cis–regulatory module query is always targeted on a particular species in a particular pairwise comparison and the results are reported with respect to the genes of the chosen species.

The results can be limited to the CRMs of particular genes or the CRMs containing binding sites for particular transcription factors. The minimum enhancer element locator score and the maximum length of the module can also be set. The system lists the conserved cis–regulatory modules fulfilling the requirements in the decreasing order of the enhancer element locator score. The report also includes the genes associated with the modules and the genomic regions containing the module. All included information is annotated with appropriate hyperlinks to the Ensembl database. The list

[1] at `http://sysdb.cs.helsinki.fi/u/tkt_bsap`

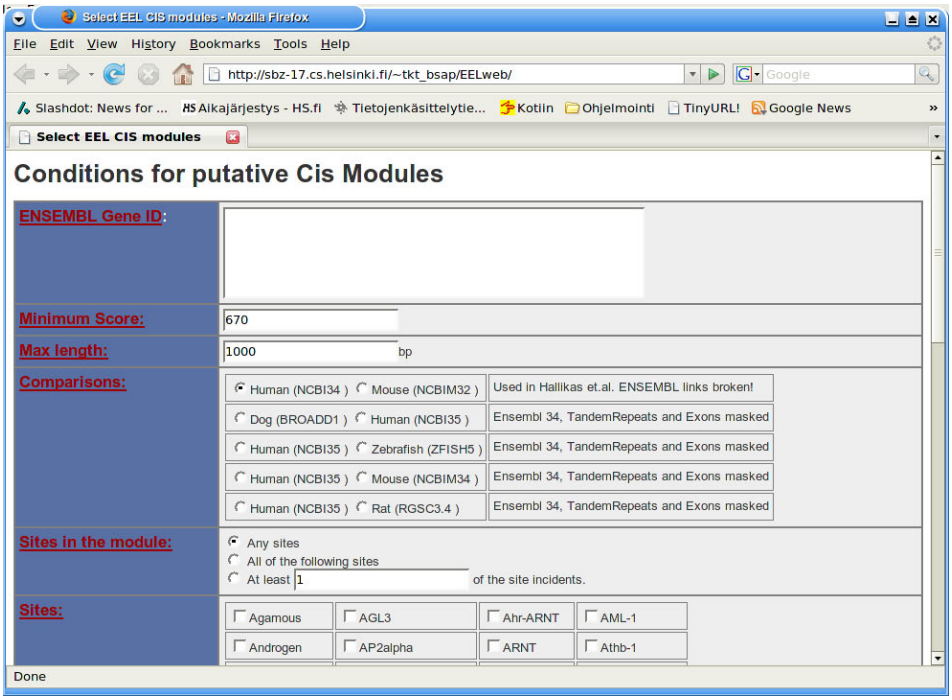

Figure 7.3: Query input view

also includes a link to a more detailed view of the conserved cis–regulatory module.

The detailed view of an individual cis–regulatory module, as seen in Figure 7.4 reports the exact location of the conserved cis–regulatory module in chromosomal DNA coordinates, the associated orthologous genes, and the detailed positions, transcription factors and affinities of the binding sites forming the module. The DNA locations and the gene names are hyperlinked to the Ensembl database which provides a wealth of additional genomic annotations.

7.3.2 Distributed Annotation System server

The Distributed Annotation System (DAS) is a protocol for sharing genomic sequence annotation across various independent data providers [DJD+01]. The annotation data is transported as XML over http with ReST type architecture, which makes programmatic usage of DAS simple both for the client and the annotation server [Fie00]. The annotation data can be visualized

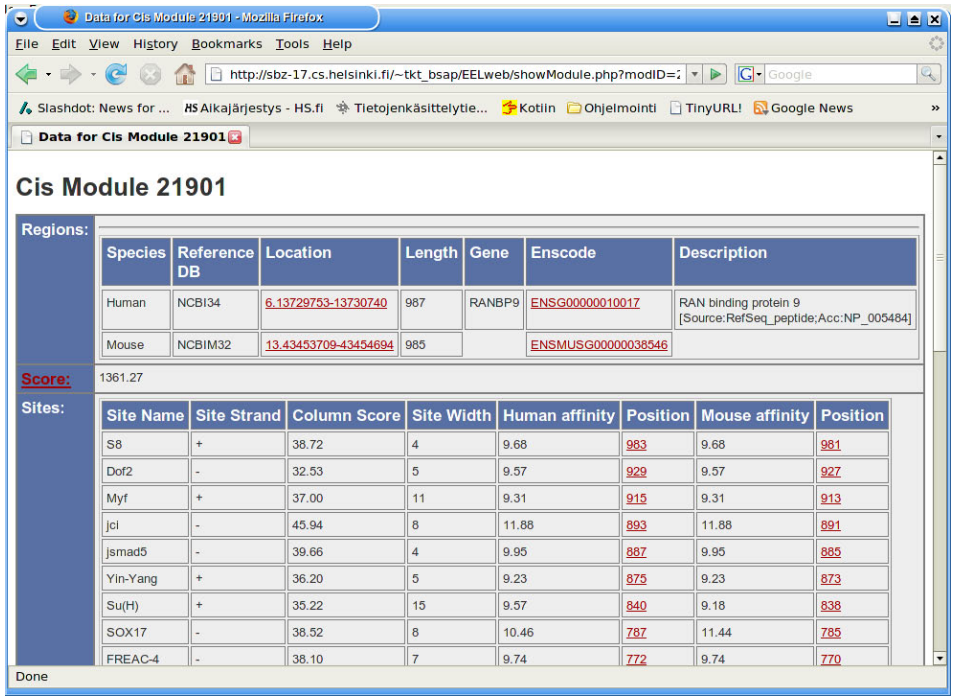

Figure 7.4: Detailed view of a conserved cis–regulatory module

on the genome browser websites or on dedicated DAS clients [HAB+07, KKZ+07, PDH05, JVD+05].

The conserved cis–regulatory modules from the human and mouse genome–wide comparison are also published as a DAS server[2]. The server receives a query for annotations for a given genomic region, defined in chromosomal coordinates, and returns the conserved cis–regulatory modules (score > 400 and length $<$ 1kbp) on that region.

An example of how the DAS data is visualized in the Ensembl genome browser is seen in Figure 7.5 [HAB+07]. The figure displays a 200kbp genomic region from the human chromosome 10 with four Ensembl annotated genes. In addition, there is the NCBI35cis track, which shows the locations of the three conserved cis–regulatory modules in this region as small black rectangles. By clicking a black rectangle on the web page, the user can view more detailed information about the module including a hyperlink to the detailed web page of Figure 7.4.

[2]at `http://sysdb.cs.helsinki.fi/das`

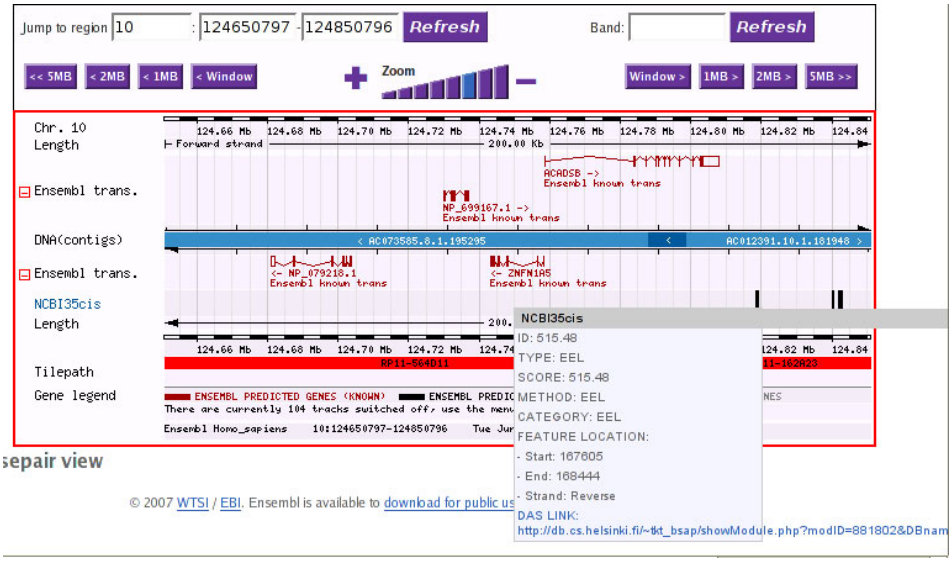

Figure 7.5: Enhancer annotation track on the Ensembl genome browser

7.3.3 OLAP web service for regulatory mechanism discovery

The previous user interfaces for the genome–wide data provide detailed information for a user interested in specific genes or a specific region of the genome but offers very little for the user interested in a set of genes. To serve the users interested in the regulators of a set of known genes, we have implemented the mechanism discovery methods of Chapter 6 as an On-Line Analytical Processing web service.

The service implements a simple ReST style stateless protocol, receiving the list of Ensembl stable gene identifiers and the associated Ensembl release number as the query and returning the names and p–values for the transcription factors as the result [Fie00]. The p–values test the hypothesis whether the transcription factor binds the CRMs of these genes at background frequency or not.

The return value of the web service is a well–structured XML message together with a XSLT and CSS style sheets [BPSM$^+$06, Cla99, LB96]. This relatively complicated architecture provides an easy programmatic access

to the service and a standard and controlled way of converting the response to a human readable format[3].

The service provides programmatic access over the hypertext transport protocol (http) and the resulting XML can be parsed with equally ubiquitous parser libraries. The XML format is vital for including the service as a building block to larger bioinformatics work flows [OGA+06].

The attached style sheets allow modern web browsers to render the XML to a human readable format in a standard and controlled way. This feature is important for including the web service as a route for continued analysis to systems such as the ArrayExpress data warehouse that can generate clusters of putatively co-regulated genes [PKS+07]. The request to our server is equal to submitting a web form and the response can be given to the web browser just like any other web page.

7.4 Conclusions

A common problem in academic bioinformatics software development is ignoring the end users who most likely are not as computer-literate as the person developing the software. The other similarly problematic extreme is the software that cannot be used in non-interactive environments such as grids or clusters.

The enhancer element locator tries to avoid these problems by providing multiple optional interfaces for various user profiles. A biologist more inclined towards traditional single gene study will probably find the point–and–click kind of interface acceptable while a bioinformatician might favor the command line interface which allows scripting and clustering.

The high performance clustering allows computing the EEL predictions for all human genes which results in a massive dataset that allows asking and answering completely different kind of questions as compared to single gene analysis. To make these queries generally available we have implemented a set of web interfaces that provide the data, annotations and analyzes to the general public over the Internet.

[3]The protocol was agreed upon at a meeting of the Biosapiens Network of Excellence Work Package 102 at the European Bioinformatics Institute in May 2007

Chapter 8

Conclusions

In this thesis I have presented the conserved cis–regulatory module model for evolutionary conservation of cis–regulatory elements in multicellular organisms. The model has been utilized to locate elements that enhance tissue–specific gene expression during mammalian development.

The model allows a sparse dynamic programming type of algorithm to find a list of conserved DNA sequences that are most like cis–regulatory elements. The algorithm is implemented in an efficient manner which allows both interactive use and highly parallel use for practically whole genome studies. The software efficiency enables the estimation of reasonable parameter values for the cis–regulatory module model by computationally intensive search methods.

This thesis also analyzed the statistical distribution of the enhancer element locator scores expected under neutrally evolved genomic sequences. The analytical distribution developed allows estimating the statistical significance of the located cis–regulatory modules under biologically relevant null hypothesis.

The available high–power computing clusters, the public genome information databases and the efficiency of the implemented software allow cis–regulatory module prediction for all genes in various genomes [HAB+07]. This prediction data is used by the web interfaces and services to provide predictions of new targets for known transcription factors or predictions of the transcription factors actuating the transcriptional response observed in a laboratory experiment.

The goal of this work has been to create a practical tool for analyzing distal enhancers of mammalian genes expressed in various stages of development. I see that the implemented Enhancer Element Locator tool meets this objective. The biological relevance of the predictions made was

shown in the original publication and later other groups have also verified the biological relevance of the results [VJM+07].

The possibilities for future work in this field are almost endless. The increase in the number and quality of the transcription factor binding site motifs will improve the practical prediction results in the short term and enable the development of more detailed models in the long term. The cis–regulatory module model will also be improved when the knowledge about the biochemistry and the evolution of cis–regulatory elements improve [PMH07, WHA+03].

A practical short–term improvement to the presented method is the development of an efficient multi–species version of the Enhancer Element Locator. The sum–of–pairs model presented in this thesis does not allow efficient and exact inference hence there is clear room for developing heuristic multiple alignment methods for cis–regulatory module discovery.

References

[ACH+00] M. D. Adams, S. E. Celniker, R. A. Holt et al. The genome sequence of Drosophila melanogaster. *Science*, 287(5461):2185–2195, 2000.

[AD97] M. Arnone and E. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124:1851–1864, 1997.

[AGM+90] S. Altschul, W. Gish, W. Miller, E. Myers and D. Lipman. Basic local alignment search tool. *Molecular Biology and Evolution*, 215(3):403–410, 1990.

[AGW03] J. Abril, R. Guigo and T. Wiehe. gff2aplot: Plotting sequence comparisons. *Bioinformatics*, 19(18):2477–2479, 2003.

[AK05] D. Arnosti and M. Kulkarni. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry*, 94(5):890–898, 2005.

[AKK+04] M. Agren, P. Kogerman, M. Kleman, M. Wessling and R. Toftgard. Expression of the PTCH1 tumor suppressor gene is regulated by alternative promoters and a single functional Gli-binding site. *Gene*, 330:101–14, 2004.

[AMS+97] S. Altschul, T. Madden, A. Schaffer et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[ATC+03] S. Aerts, G. Thijs, B. Coessens et al. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Research*, 31(6):1753–1764, 2003.

[AVLT+03] S. Aerts, P. Van Loo, G. Thijs, Y. Moreau and B. De Moor. Computational detection of cis-regulatory modules. *Bioinformatics*, 19(Suppl. 2):ii5–ii14, 2003.

[BBC+06] M. Blanchette, A. Bataille, X. Chen et al. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Research*, 16(5):656–668, 2006.

[BC00] M. Bienz and H. Clevers. Linking Colorectal Cancer to Wnt Signaling. *Cell*, 103(2):311–320, 2000.

[BEB+06] J. Blake, J. Eppig, C. Bult et al. The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Research*, 34(Suppl. 1):D562–567, 2006.

[BEFK03] Y. Barash, G. Elidan, N. Friedman and T. Kaplan. Modeling dependencies in protein-DNA binding sites. *Proceedings of the seventh annual international conference on Computational molecular biology*, pages 28–37, 2003.

[BEKF05] Y. Barash, G. Elidan, T. Kaplan and N. Friedman. CIS: compound importance sampling method for protein–DNA binding site p-value estimation. *Bioinformatics*, 21(5):596–600, 2005.

[Ber89] G. Bernardi. The Isochore Organization of the Human Genome. *Annual Review of Genetics*, 23(1):637–659, 1989.

[BGD+95] I. Bolon, V. Gouyer, M. Devouassoux et al. Expression of c-ets-1, collagenase 1, and urokinase-type plasminogen activator genes in lung carcinomas. *American Journal of Pathology*, 147(5):1298–1310, 1995.

[BGV07] A. Bais, S. Grossmann and M. Vingron. Simultaneous alignment and annotation of cis-regulatory regions. *Bioinformatics*, 23(2):e44–e49, 2007.

[BH95] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B(Statistical Methodology)*, 57(1):289–300, 1995.

[BMG03] E. Blanco, X. Messeguer and R. Guigo. Alignment of Promoter Regions by Mapping Nucleotide Sequences into Arrays of Transcription Factors Binding Motifs. Tech. rep., Universitat Pompeu Fabra, Universitat Politecnica de Catalunya, 2003. Also a poster in RECOMB'03. http://genome.imim.es/ eblanco/work/.

[BMSG06] E. Blanco, X. Messeguer, T. Smith and R. Guigó. Transcription Factor Map Alignment of Promoter Regions. *PLoS Computational Biology*, 2:e49, 2006.

[BN03] T. Bailey and W. Noble. Searching for statistically significant regulatory modules. *Bioinformatics*, 19(Suppl. 2):ii16–ii25, 2003.

[BNP+02] B. P. Berman, Y. Nibu, B. D. Pfeiffer et al. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):757–762, 2002.

[BPQ+06] M. Berger, A. Philippakis, A. Qureshi et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology*, 24:1429–1435, 2006.

[BPSM+06] T. Bray, J. Paoli, C. Sperberg-McQueen, E. Maler and F. Yergeau. Extensible Markup Language (XML) 1.0 (Fourth Edition) . Tech. rep., W3C Recommendation, 2006.

[BRS81] J. Banerji, S. Rusconi and W. Schaffner. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 Pt 1):299–308, 1981.

[BSG+03] Z. Bryant, M. D. Stone, J. Gore et al. Structural transitions and elasticity from torque measurements on DNA. *Nature*, 424(6946):338–341, 2003.

[BSH+04] M. Beckstette, D. Strothmann, R. Homann, R. Giegerich and S. Kurtz. PoSSuMsearch: Fast and Sensitive Matching of Position Specific Scoring Matrices using Enhanced Suffix Arrays. *Proceedings of the German Conference on Bioinformatics*, pages 53–64, 2004.

[Bun01] R. Bundschuh. Rapid significance estimation in local sequence alignment with gaps. *Annual Conference on Research in Computational Molecular Biology*, pages 77–85, 2001.

[BY01] Y. Benjamini and D. Yekutieli. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[CBS+04] G. Cooper, M. Brudno, E. Stone et al. Characterization of evolutionary rates and constraints in three Mammalian genomes. *Genome Research*, 14(4):539–48, 2004.

[CCB+05] R. A. Cameron, S. H. Chow, K. Berney et al. An evolutionary constraint: strongly disfavored class of change in DNA sequence during divergence of cis-regulatory modules. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11769–74, 2005.

[CD97] S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. *ACM SIGMOD Record*, 26(1):65–74, 1997.

[CFJ93] D. Clarkson, Y. Fan and H. Joe. A remark on algorithm 643: FEXACT: an algorithm for performing Fisher's exact test in rxc contingency tables. *ACM Transactions on Mathematical Software*, 19(4):484–488, 1993.

[CIG+70] G. Chapman, E. Idle, T. Gilliam et al. Dirty Hungarian Phrasebook. in *Monty Python's Flying Circus*, 1970.

[Cla99] J. Clark. XSL Transformations (XSLT) Version 1.0. Tech. Rep. 11, W3C Recommendation, 1999.

[CM99] M. Cole and S. McMahon. The Myc oncoprotein: a critical evaluation of transactivation and target gene regulation. *Oncogene*, 18(19):2916–24, 1999.

[CNM+06] L. Chang, R. Nagarajan, J. Magee, J. Milbrandt and G. Stormo. A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles. *Genome Research*, 16(3):405–413, 2006.

[CR05] N. Campbell and J. Reece. *Biology*. Pearson, Benjamin Cumings, 2005.

[CRB97] E. Crowley, K. Roeder and M. Bina. A Statistical Model for Locating Regulatory Regions in Genomic DNA. *Journal Of Molecular Biology*, 268(1):8–14, 1997.

[DCG05] I. Donaldson, M. Chapman and B. Gottgens. TFBScluster: a resource for the characterization of transcriptional regulatory networks. *Bioinformatics*, 21(13):3058–3059, 2005.

[DCK+05] I. Donaldson, M. Chapman, S. Kinston et al. Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development. *Human Molecular Genetics*, 14(5):595–601, 2005.

[DG07] I. Donaldson and B. Gottgens. CoMoDis: composite motif discovery in mammalian genomes. *Nucleic Acids Research*, 35(1):e1, 2007.

[DiA99] N. Dahmane and A. i Altaba. Sonic hedgehog regulates the growth and patterning of the cerebellum. *Development*, 126:3089–3100, 1999.

[DJD+01] R. Dowell, R. Jokerst, A. Day, S. Eddy and L. Stein. The Distributed Annotation System. *BMC Bioinformatics*, 2(1):7, 2001.

[DLB+00] H. Dassule, P. Lewis, M. Bei, R. Maas and A. McMahon. Sonic hedgehog regulates growth and morphogenesis of the tooth. *Development*, 127:4775–4785, 2000.

[DLR77] A. Dempster, N. Laird and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Series B(Statistical Methodology)*, 39(1):1–38, 1977.

[DMJRM00] R. De Maesschalck, D. Jouan-Rimbaud and D. Massart. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000.

[DRO+02] E. Davidson, J. Rast, P. Oliveri et al. A Genomic Regulatory Network for Development. *Science*, 295(5560):1669–1678, 2002.

[DWB+05] M. Dai, P. Wang, A. Boyd et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research*, 33(20):e175, 2005.

[EDS+06] T. Eyre, F. Ducluzeau, T. Sneddon et al. The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Research*, 34(Suppl. 1):D319–321, 2006.

[EGGI92a] D. Eppstein, Z. Galil, R. Giancarlo and G. Italiano. Sparse dynamic programming I: linear cost functions. *Journal of the ACM*, 39(3):519–545, 1992.

[EGGI92b] D. Eppstein, Z. Galil, R. Giancarlo and G. Italiano. Sparse dynamic programming II: convex and concave cost functions. *Journal of the ACM*, 39(3):546–567, 1992.

[EHL+03] L. Elnitski, R. Hardison, J. Li et al. Distinguishing Regulatory DNA From Neutral Sites. *Genome Research*, 13(1):64, 2003.

[ESBB98] M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–14868, 1998.

[Fay98] U. Fayyad. Mining Databases: Towards Algorithms for Knowledge Discovery. *Data Engineering Bulletin*, 21(1):39–48, 1998.

[FD07] B. Frey and D. Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972, 2007.

[Fel81] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[FHW01] M. Frith, U. Hansen and Z. Weng. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, 17(10):878–889, 2001.

[Fie00] R. Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. Ph.D. thesis, University Of California, Irvine, 2000.

[Fis22] R. A. Fisher. On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[FKS03] R. Fagin, R. Kumar and D. Sivakumar. Comparing Top k Lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003.

[FLW03] M. Frith, M. Li and Z. Weng. Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Research*, 31(13):3666–3668, 2003.

[Fos02] I. Foster. What is the Grid? A Three Point Checklist., 2002.

[FPC+95] C. Fan, J. Porter, C. Chiang et al. Long-range sclerotome induction by sonic hedgehog: direct role of the amino-terminal cleavage product and modulation by the cyclic AMP signaling pathway. *Cell*, 81(3):457–465, 1995.

[FSHW02] M. Frith, J. Spouge, U. Hansen and Z. Weng. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Research*, 30(14):3214–3224, 2002.

[GBF+95] D. Goldhamer, B. Brunk, A. Faerman et al. Embryonic activation of the myoD gene is regulated by a highly conserved distal control element. *Development*, 121:637–649, 1995.

[GCB+04] R. C. Gentleman, V. J. Carey, D. M. Bates et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[Gen02] W. Gentzsch. Sun Grid Engine: Towards Creating a Compute Power Grid. *Proceedings of the First IEEE/ACM International Symposium on Cluster Computing and the Grid*, 2002.

[GL05] M. Gupta and J. Liu. De novo cis-regulatory module elicitation for eukaryotic genomes. *Proceedings of the National Academy of Sciences*, 102(20):7079–7084, 2005.

[GSR96] W. Gilks, D. Spiegelhalter and S. Richardson. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, 1996.

[Gum58] E. Gumbel. *Statistics of Extremes*. Columbia University Press, 1958.

[GWM+04] R. Gibbs, G. Weinstock, M. Metzker et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–521, 2004.

[HAB+07] T. J. P. Hubbard, B. L. Aken, K. Beal et al. Ensembl 2007. *Nucleic Acids Research*, 35(Suppl. 1):D610–617, 2007.

[HAM+95] D. Henrique, J. Adam, A. Myat et al. Expression of a Delta homologue in prospective neurons in the chick. *Nature*, 375(6534):787–790, 1995.

[Har02] A. Hartmann. Sampling rare events: Statistics of local sequence alignments. *Physical Review E*, 65(5):56102, 2002.

[HB86] P. Hippel and O. Berg. On the Specificity of DNA–Protein Interactions. *Proceedings of the National Academy of Sciences*, 83(6):1608–1612, 1986.

[HB98] A. Halpern and W. Bruno. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution*, 15:910–917, 1998.

[HB06] B. Huber and M. Bulyk. Meta-analysis discovery of tissue-specific DNA sequence motifs from mammalian gene expression data. *BMC Bioinformatics*, 7(1):229, 2006.

[HBF+04] D. Hill, D. Begley, J. Finger et al. The mouse Gene Expression Database (GXD): updates and enhancements. *Nucleic Acids Research*, 32(Suppl. 1):D568–D571, 2004.

[HGCM02] M. Halfon, Y. Grad, G. Church and A. Michelson. Computation-Based Discovery of Related Transcriptional Regulatory Modules and Motifs Using an Experimentally Validated Combinatorial Model. *Genome Research*, 12(7):1019–1028, 2002.

[HH64] J. Hammersley and D. Handscomb. *Monte Carto Methods*. London: Methuen & Co Ltd. New York: John Wiley & Sons Inc., 1964.

[HKY85] M. Hasegawa, H. Kishino and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

[HLS94] J. M. Heumann, A. S. Lapedes and G. D. Stormo. Neural Networks for Determining Protein Specificity and Multiple Alignment of Binding Sites. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, pages 188–194. AAAI Press, Menlo Park, California, 1994.

[Hol79] S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[HPS+06] O. Hallikas, K. Palin, N. Sinjushina et al. Genome-wide Prediction of Mammalian Enhancers Based on Analysis of Transcription-Factor Binding Affinity. *Cell*, 124(1):47–59, 2006.

[HS77] J. Hunt and T. Szymanski. A fast algorithm for computing longest common subsequences. *Communications of the ACM*, 20(5):350–353, 1977.

[HSR+98] T. He, A. Sparks, C. Rago et al. Identification of c-MYC as a Target of the APC Pathway. *Science*, 281(5382):1509–1512, 1998.

[HVB+00] J. Huelsken, R. Vogel, V. Brinkmann et al. Requirement for β-Catenin in Anterior-Posterior Axis Formation in Mice. *The Journal of Cell Biology*, 148(3):567–578, 2000.

[IM01] P. Ingham and A. McMahon. Hedgehog signaling in animal development: paradigms and principles. *Genes & Development*, 15(23):3059–3087, 2001.

[JAWL03] O. Johansson, W. Alkema, W. Wasserman and J. Lagergren. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, 19(1):i169–76, 2003.

[JC69] T. Jukes and C. Cantor. Evolution of protein molecules. *Mammalian Protein Metabolism*, 3:21–132, 1969.

[JT96] S. Jones and J. Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 93(1):13–20, 1996.

[JVD+05] P. Jones, N. Vinod, T. Down et al. Dasty and UniProt DAS: a perfect pair for protein feature visualization. *Bioinformatics*, 21(14):3198–3199, 2005.

[JW06] H. Ji and W. Wong. Computational Biology: Toward Deciphering Gene Regulatory Information in Mammalian Genomes. *Biometrics*, 62(3):645–663, 2006.

[KA90] S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the United States of America*, 87:2264–2268, 1990.

[KA04] M. Kulkarni and D. Arnosti. cis-Regulatory Logic of Short-Range Transcriptional Repression in Drosophila melanogaster. *Molecular and Cellular Biology*, 25(9):3411–3420, 2004.

[KAA+07] T. Kulikova, R. Akhtar, P. Aldebert et al. EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Research*, 35(Database issue):D16, 2007.

[KCR03] A. Kenney, M. Cole and D. Rowitch. Nmyc upregulation by sonic hedgehog signaling promotes proliferation in developing cerebellar granule neuron precursors. *Development*, 130(1):15–28, 2003.

[KD92] S. Karlin and A. Dembo. Limit Distributions of Maximal Segmental Score among Markov-Dependent Partial Sums. *Advances in Applied Probability*, 24(1):113–140, 1992.

[KGV83] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

[Kim68] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217(129):624–6, 1968.

[Kim80] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120, 1980.

[KKS+04] A. Kasprzyk, D. Keefe, D. Smedley et al. EnsMart: A Generic System for Fast and Flexible Access to Biological Data. *Genome Research*, 14(1):160–169, 2004.

[KKZ+07] R. M. Kuhn, D. Karolchik, A. S. Zweig et al. The UCSC genome browser database: update 2007. *Nucleic Acids Research*, 35(Suppl. 1):D668–673, 2007.

[KM04] S. Karanam and C. Moreno. CONFAC: automated application of comparative genomic promoter analysis to DNA microarray datasets. *Nucleic Acids Research*, 32(1):W475–W484, 2004.

[Knu98] D. Knuth. *The art of computer programming, volume 3: sorting and searching*. Addison Wesley Longman Publishing Co., Inc. Redwood City, CA, USA, 1998.

[Kob07] A. Kobayashi. Imaging X-gal-Stained Mouse Embryos. *Cold Spring Harbor Protocols*, 2007(8):pdb.prot4724–, 2007.

[Kre04] G. Kreiman. Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucleic Acids Research*, 32(9):2889–2900, 2004.

[KS02] S. Kumar and S. Subramanian. Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):803–808, 2002.

[KW01] W. Krivan and W. Wasserman. A Predictive Model for Regulatory Sequences Directing Liver-Specific Transcription. *Genome Research*, 11(9):1559, 2001.

[LB96] H. k. W. Lie and B. Bos. Cascading Style Sheets, level 1. Tech. rep., W3C Recommendation, 1996.

[LLB+01] E. Lander, L. Linton, B. Birren et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[LM06] D. Lemons and W. McGinnis. Genomic Evolution of Hox Gene Clusters. *Science*, 313(5795):1918–1922, 2006.

[LN04] C. Logan and R. Nusse. The Wnt Signaling Pathway In Development And Disease. *Annual Review of Cell and Developmental Biology*, 20(1):781–810, 2004.

[LOP+02] G. Loots, I. Ovcharenko, L. Pachter, I. Dubchak and E. Rubin. rVista for Comparative Sequence-Based Discovery of Functional Transcription Factor Binding Sites. *Genome Research*, 12(5):832–839, 2002.

[LPBM94] A. Lavenu, S. Pournin, C. Babinet and D. Morello. The cis-acting elements known to regulate c-myc expression ex vivo are not sufficient for correct transcription in vivo. *Oncogene*, 9(2):527–36, 1994.

[LPCiA97] J. Lee, K. Platt, P. Censullo and A. i Altaba. *Gli1* is a target of Sonic hedgehog that induces ventral neural tube development. *Development*, 124:2537–2552, 1997.

[LPH06] G. Lunter, C. Ponting and J. Hein. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Computational Biology*, 2:e5, 2006.

[LT03] M. Levine and R. Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945):147–51, 2003.

[LTWM+05] K. Lindblad-Toh, C. Wade, T. Mikkelsen et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–819, 2005.

[MAB07] S. Mahony, P. Auron and P. Benos. DNA Familial Binding Profiles Made Easy: Comparison of Various Motif Alignment and Clustering Strategies. *PLoS Computational Biology*, 3(3):e61, 2007.

[Mah36] P. Mahalanobis. On the generalized distance in statistics. *Proc Natl Inst Sci India*, 2(1):49–55, 1936.

[MBMO95] J. Molkentin, B. Black, J. Martin and E. Olson. Cooperative activation of muscle gene expression by MEF2 and myogenic bHLH proteins. *Cell*, 83(7):1125–1136, 1995.

[MCG06] E. Margulies, C. Chen and E. Green. Differences between pairwise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends in Genetics*, 22(4):187–193, 2006.

[MCY+93] J. McDermott, M. Cardoso, Y. Yu et al. hMEF2C gene encodes skeletal muscle-and brain-specific transcription factors. *Molecular and Cellular Biology*, 13(4):2564–2577, 1993.

[MF00] Z. Michalewicz and D. B. Fogel. *How to Solve It: Modern Heuristics*. Springer Verlag, 2000.

[MHE+05] T. Mikkelsen, L. Hillier, E. Eichler et al. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437:69–87, 2005.

[MLH04] I. Miklós, G. Lunter and I. Holmes. A" Long Indel" Model For Evolutionary Sequence Alignment. *Molecular Biology and Evolution*, 21(3):529–540, 2004.

[MMML02] M. Markstein, P. Markstein, V. Markstein and M. Levine. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proceedings of the National Academy of Sciences*, 99(2):763–768, 2002.

[MN06] F. Marın and M. Nieto. The Expression of Scratch Genes in the Developing and Adult Brain. *Developmental Dynamics*, 235:2586–2591, 2006.

[Moo21] O. L. Moore. Index For Numbers. U.S. Patent # 1,463,656, 1921. Granted Jul. 31, 1923.

[Mot92] R. Mott. Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bulletin of Mathematical Biology*, 54(1):59–75, 1992.

[MP86] C. Mehta and N. Patel. ALGORITHM 643: FEXACT: a FORTRAN subroutine for Fisher's exact test on unordered r× c contingency tables. *ACM Transactions on Mathematical Software*, 12(2):154–161, 1986.

[MQ07] S. Maerkl and S. Quake. A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. *Science*, 315(5809):233, 2007.

[MS86] P. Macdonald and G. Struhl. A molecular gradient in early Drosophila embryos and its role in specifying the body pattern. *Nature*, 324(6097):537–545, 1986.

[MyS07] MySQL AB. *MySQL Reference Manual*. MySQL AB, 2007.

[MZM+04] M. Markstein, R. Zinzen, P. Markstein et al. A regulatory code for neurogenic gene expression in the Drosophila embryo. *Development*, 131(10):2387–2394, 2004.

[Nag03] A. Nagy. *Manipulating the Mouse Embryo: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, 3 edn., 2003.

[NBE+93] W. Niblack, R. Barber, W. Equitz et al. QBIC project: querying images by content, using color, texture, and shape. *Proc. SPIE*, 1908:173–187, 1993.

[NGVB06a] A. Nagy, M. Gertsenstein, K. Vintersten and R. Behringer. Microinjection of Mouse Zygotes. *Cold Spring Harbor Protocols*, 2006(16):pdb.prot4397–, 2006.

[NGVB06b] A. Nagy, M. Gertsenstein, K. Vintersten and R. Behringer. Oviduct Transfer. *Cold Spring Harbor Protocols*, 2006(18):pdb.prot4379–, 2006.

[NGVB07] A. Nagy, M. Gertsenstein, K. Vintersten and R. Behringer. Sectioning Mouse Embryos. *Cold Spring Harbor Protocols*, 2007(1):pdb.prot4703–, 2007.

[NHC04] C. Nelson, B. Hersh and S. Carroll. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biology*, 5(4), 2004.

[Nok05] Nokia. *Python for Series 60 Platform. API Reference*, 2005.

[NWS+01] E. Nakakura, D. Watkins, K. Schuebel et al. Mammalian Scratch: A neural-specific Snail family transcriptional repressor. *Proceedings of the National Academy of Sciences of the United States of America*, 98(7):4010, 2001.

[NZPF+04] M. Nobrega, Y. Zhu, I. Plajzer-Frick, V. Afzal and E. Rubin. Megabase deletions of gene deserts result in viable mice. *Nature*, 431(7011):988–993, 2004.

[OGA+06] T. Oinn, M. Greenwood, M. Addis et al. Taverna: Lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, 18(10):1067–1100, 2006.

[OIY07] K. Okita, T. Ichisaka and S. Yamanaka. Generation of germline-competent induced pluripotent stem cells. *Nature*, 448(7151):313–317, 2007.

[OJWM92] S. Osawa, T. Jukes, K. Watanabe and A. Muto. Recent evidence for evolution of the genetic code. *Microbiology and Molecular Biology Reviews*, 56(1):229–264, 1992.

[OMRM+99] S. O'Brien, M. Menotti-Raymond, W. Murphy et al. The Promise of Comparative Genomics in Mammals. *Science*, 286(5439):458, 1999.

[ON05] I. Ovcharenko and M. Nobrega. Identifying synonymous regulatory elements in vertebrate genomes. *Nucleic Acids Research*, 33(1):W403–W407, 2005.

[OSK04] A. Ogurtsov, S. Sunyaev and A. Kondrashov. Indel-Based Evolutionary Distance and Mouse-Human Divergence. *Genome Research*, 14(8):1610, 2004.

[PAA+03] G. Parra, P. Agarwal, J. Abril et al. Comparative Gene Prediction in Human and Mouse. *Genome Research*, 13(1):108, 2003.

[PAM+06] L. A. Pennacchio, N. Ahituv, A. M. Moses et al. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499 – 502, 2006.

[Pap07] D. Papatsenko. ClusterDraw web server: a tool to identify and visualize clusters of binding motifs for transcription factors. *Bioinformatics*, 2007.

[Pat81] W. Patefield. Algorithm AS 159: An Efficient Method of Generating Random $R \times C$ Tables with Given Row and Column Totals. *Applied Statistics*, 30(1):91–97, 1981.

[PBRT99] J. Puzicha, J. M. Buhmann, Y. Rubner and C. Tomasi. Empirical Evaluation of Dissimilarity Measures for Color and Texture. *Seventh International Conference on Computer Vision*, 02:1165, 1999.

[PBW06] N. Pierstorff, C. Bergman and T. Wiehe. Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA. *Bioinformatics*, 22(23):2858, 2006.

[PDH05] A. Prlic, T. Down and T. Hubbard. Adding some SPICE to DAS. *Bioinformatics*, 21(Suppl. 2), 2005.

[PHB05] A. Philippakis, F. He and M. Bulyk. ModuleFinder: A Tool for Computational Discovery of Cis Regulatory Modules. In *Proceedings of the Pacific Symposium of Biocomputing*, 2005. `http://helix-web.stanford.edu/psb05/`.

[PKS+07] H. Parkinson, M. Kapushesky, M. Shojatalab et al. ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Research*, 35(Suppl. 1):D747–750, 2007.

[PLNO07] L. Pennacchio, G. Loots, M. Nobrega and I. Ovcharenko. Predicting tissue-specific enhancers in the human genome. *Genome Research*, 17(2):201, 2007.

[PMH07] D. Panne, T. Maniatis and S. Harrison. An Atomic Model of the Interferon-β Enhanceosome. *Cell*, 129(6):1111–1123, 2007.

[PPP+98] M. Pellegrini, G. Pilia, S. Pantano et al. Gpc 3 expression correlates with the phenotype of the Simpson-Golabi-Behmel syndrome. *Developmental Dynamics*, 213(4):431–439, 1998.

[PPS+06] S. Prabhakar, F. Poulin, M. Shoukry et al. Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Research*, 16(7):855–863, 2006.

[PR01] L. A. Pennacchio and E. M. Rubin. Genomic strategies to identify mammalian regulatory sequences. *Nature Reviews Genetics*, 2(2):100–109, 2001.

[PRU07] C. Pizzi, P. Rastas and E. Ukkonen. Fast search algorithms for position specific scoring matrices. In *First International Conference on Bioinformatics Research and Development — BIRD 2007*, 2007.

[PSB+02] C. Perry, E. Sklan, K. Birikh et al. Complex regulation of acetylcholinesterase gene expression in human brain tumors. *Oncogene*, 21:8428–8441, 2002.

[PTU06] K. Palin, J. Taipale and E. Ukkonen. Locating potential enhancer elements by comparative genomics using the EEL software. *Nature Protocols*, 1(1):368 – 374, 2006.

[PUBV02] K. Palin, E. Ukkonen, A. Brazma and J. Vilo. Correlating gene promoters and expression in gene disruption experiments. *Bioinformatics*, 18(Suppl. 2):172–180, 2002.

[Ram02] Ramey, C. and Fox, B. The GNU Bash Reference Manual, 2002.

[RB93] B. Rosen and R. Beddington. Whole-mount in situ hybridization in the mouse embryo: gene expression in three dimensions. *Trends in Genetics*, 9(5):162–7, 1993.

[RD06] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006.

[RFV84] F. Rossi, M. Fiorentino and P. Versace. Two-Component Extreme Value Distribution for Flood Frequency Analysis. *Water Resources Research*, 20(7), 1984.

[RKMV07] H. Roider, A. Kanhere, T. Manke and M. Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134, 2007.

[RLB+05] S. Richards, Y. Liu, B. Bettencourt et al. Comparative genome sequencing of Drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution. *Genome Research*, 15(1):1–18, 2005.

[RRP02] M. Rebeiz, N. Reeves and J. Posakony. SCORE: A computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, 99(15):9888, 2002.

[RRW+00] B. Ren, F. Robert, J. J. Wyrick et al. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–2309, 2000.

[RSK03] M. Rosenberg, S. Subramanian and S. Kumar. Patterns of Transitional Mutation Biases Within and Among Mammalian Genomes. *Molecular Biology and Evolution*, 20(6):988–993, 2003.

[RVGS02] N. Rajewsky, M. Vergassola, U. Gaul and E. D. Siggia. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics*, 3(1):30, 2002.

[SBL92] S. Small, A. Blair and M. Levine. Regulation of even-skipped stripe 2 in the Drosophila embryo. *EMBO Journal*, 11(11):4047–4057, 1992.

[SBL96] S. Small, A. Blair and M. Levine. Regulation of Two Pair-Rule Stripes by a Single Enhancer in the Drosophila Embryo. *Developmental Biology*, 175(2):314–324, 1996.

[SF98] G. D. Stormo and D. S. Fields. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci*, 23(3):109–113, 1998.

[SHH+04] P. Sabo, R. Humbert, M. Hawrylycz et al. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proceedings of the National Academy of Sciences of the United States of America*, 101(13):4537–4542, 2004.

[SHNK97] H. Sasaki, C. Hui, M. Nakafuku and H. Kondoh. A binding site for Gli proteins is essential for HNF-3β floor plate enhancer activity in transgenics and can respond to Shh in vitro. *Development*, 124:1313–1322, 1997.

[SLL95] A. Stepanov, M. Lee and H.-P. Laboratories. *The Standard Template Library*. Hewlett-Packard Laboratories, Technical Publications Dept, 1995.

[Smy04] G. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):3, 2004.

[SOBHK03] R. Sharan, I. Ovcharenko, A. Ben-Hur and R. Karp. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, 19(1):i283–91, 2003.

[SS05] E. Segal and R. Sharan. A Discriminative Model for Identifying Spatial cis-Regulatory Modules. *Journal of Computational Biology*, 12:822–834, 2005.

[SS07] E. Sharon and E. Segal. A Feature-Based Approach to Modeling Protein-DNA Interactions. *Lecture Notes in Computer Science*, 4453:77, 2007.

[SSC88] S. Sagar, F. Sharp and T. Curran. Expression of c-fos protein in brain: metabolic mapping at the cellular level. *Science*, 240(4857):1328, 1988.

[SSDB95] M. Schena, D. Shalon, R. W. Davis and P. O. Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235):467–470, 1995.

[Sta89] R. Staden. Methods for calculating the probabilities of finding patterns in sequences. *Computer Applications in the Biosciences*, 5(2):89–96, 1989.

[Sta91] R. Stallman. GNU General Public License. *GNU Project–Free Software Foundation*, 1991.

[Sta02] R. Stallman. *Free Software, Free Society: Selected Essays of Richard M. Stallman*, chap. The Free Software Definition. GNU Press, 2002.

[Sto00] G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[SvNS03] S. Sinha, E. van Nimwegen and E. Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19(Suppl. 1):i292, 2003.

[SW81] T. F. Smith and M. S. Waterman. Identication of common molecular subsequences. *Journal Of Molecular Biology*, 147:195–197, 1981.

[SWB+04] A. Su, T. Wiltshire, S. Batalov et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067, 2004.

[Tav86] S. Tavaré. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.

[TB01] J. Taipale and P. Beachy. The Hedgehog and Wnt signalling pathways in cancer. *Nature*, 411:349–354, 2001.

[Tho17] D. Thompson. *On Growth and Form*. Cambridge University Press, 1917.

[TKF91] J. Thorne, H. Kishino and J. Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, 33(2):114–124, 1991.

[TKF92] J. Thorne, H. Kishino and J. Felsenstein. Inching toward reality: An improved likelihood model of sequence evolution. *Journal of Molecular Evolution*, 34(1):3–16, 1992.

[TLB+05] M. Tompa, N. Li, T. Bailey et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23:137–144, 2005.

[TLM+01] G. Thijs, M. Lescot, K. Marchal et al. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 17(12):1113–1122, 2001.

[TPW+04] W. Thompson, M. Palumbo, W. Wasserman, J. Liu and C. Lawrence. Decoding Human Regulatory Circuits. *Genome Research*, 14(10a):1967, 2004.

[TTK+06] J. Taylor, S. Tyekucheva, D. King et al. ESPERR: Learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Research*, 16(12):1596, 2006.

[Tur52] A. M. Turing. The Chemical Basis of Morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37–72, 1952.

[VAM+01] J. C. Venter, M. D. Adams, E. W. Myers et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[vHHG04] A. von Heydebreck, W. Huber and R. Gentleman. Differential expression with the bioconductor project. *Bioconductor Project Working Papers. Working Paper 7*, 2004.

[VJM+07] S. Vokes, H. Ji, S. McCuine et al. Genomic characterization of Gli-activator targets in sonic hedgehog-mediated neural patterning. *Development*, 134(10):1977, 2007.

[vR96] G. van Rossum. Foreword. In M. Lutz, ed., *Programming Python*. O'Reilly, 1996.

[vRDJ06] G. van Rossum and F. Drake Jr. Python/C API Reference Manual, 2006.

[VSDB+06] D. Vlieghe, A. Sandelin, P. J. De Bleser et al. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Research*, 34(Suppl. 1):D95–97, 2006.

[Wag99] A. Wagner. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, 15(10):776–784, 1999.

[WAM+97] W. Weiss, K. Aldape, G. Mohapatra, B. Feuerstein and J. Bishop. Targeted expression of MYCN causes neuroblastoma in transgenic mice. *EMBO Journal*, 16:2985–2995, 1997.

[WC53] J. Watson and F. Crick. A Structure for DNA. *Nature*, 171:737–738, 1953.

[WDKK96] E. Wingender, P. Dietze, H. Karas and R. Knuppel. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Research*, 24(1):238–241, 1996.

[WDT+80] R. Wing, H. Drew, T. Takano et al. Crystal structure analysis of a complete turn of B-DNA. *Nature*, 287(5784):755–758, 1980.

[WE87] M. Waterman and M. Eggert. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *Journal Of Molecular Biology*, 197:723–728, 1987.

[WF98] W. Wasserman and J. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *Journal Of Molecular Biology*, 278(1):167–181, 1998.

[WHA+03] G. A. Wray, M. W. Hahn, E. Abouheif et al. The Evolution of Transcriptional Regulation in Eukaryotes. *Molecular Biology and Evolution*, 20(9):1377–1419, 2003.

[WIG+03] Z. Wu, R. Irizarry, R. Gentleman, F. Murillo and F. A. Spencer. A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Tech. rep., John Hopkins University, Department of Biostatistics Working Papers, Baltimore, MD, 2003.

[WLTB+02] R. Waterston, K. Lindblad-Toh, E. Birney et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–62, 2002.

[WS00] L. Wall and R. Schwartz. *Programming Perl*. O'Reilly Cambridge, Mass, 2000.

[WS04] W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, 2004.

[WSL89] K. Wolfe, P. Sharp and W. Li. Mutation rates differ among regions of the mammalian genome. *Nature*, 337(6204):283–285, 1989.

[WV94] M. S. Waterman and M. Vingron. Sequence Comparison Significance and Poisson Approximation. *Statistical Science*, 9(3):367–381, 1994.

[Yan97] Z. Yang. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences*, 13(5):555–556, 1997.

[Zip35] G. K. Zipf. *The Psycho-Biology of Language*. Houghton Mifltm Co, 1935.

[ZM06] J. Zobel and A. Moffat. Inverted files for text search engines. *ACM Computing Surveys*, 38(2), 2006.

[ZSC05] Z. Zhu, J. Shendure and G. Church. Discovering functional transcription-factor combinations in the human cell cycle. *Genome Research*, 15(6):848, 2005.

[ZW04] Q. Zhou and W. Wong. CisModule: De novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proceedings of the National Academy of Sciences*, 101(33):12114–12119, 2004.

TIETOJENKÄSITTELYTIETEEN LAITOS
PL 68 (Gustaf Hällströmin katu 2 b)
00014 Helsingin yliopisto

DEPARTMENT OF COMPUTER SCIENCE
P.O. Box 68 (Gustaf Hällströmin katu 2 b)
FIN-00014 University of Helsinki, FINLAND

JULKAISUSARJA **A**

SERIES OF PUBLICATIONS **A**

Reports may be ordered from: Kumpula Science Library, P.O. Box 64, FIN-00014 University of Helsinki, FINLAND.

A-1998-3 E. Sutinen: Approximate pattern matching with the q-gram family. 116 pp. (Ph.D. thesis).

A-1999-1 M. Klemettinen: A knowledge discovery methodology for telecommunication network alarm databases. 137 pp. (Ph.D. thesis).

A-1999-2 J. Puustjärvi: Transactional workflows. 104 pp. (Ph.D. thesis).

A-1999-3 G. Lindén & E. Ukkonen (eds.): Department of Computer Science: annual report 1998. 55 pp.

A-1999-4 J. Kärkkäinen: Repetition-based text indexes. 106 pp. (Ph.D. thesis).

A-2000-1 P. Moen: Attribute, event sequence, and event type similarity notions for data mining. 190+9 pp. (Ph.D. thesis).

A-2000-2 B. Heikkinen: Generalization of document structures and document assembly. 179 pp. (Ph.D. thesis).

A-2000-3 P. Kähkipuro: Performance modeling framework for CORBA based distributed systems. 151+15 pp. (Ph.D. thesis).

A-2000-4 K. Lemström: String matching techniques for music retrieval. 56+56 pp. (Ph.D.Thesis).

A-2000-5 T. Karvi: Partially defined Lotos specifications and their refinement relations. 157 pp. (Ph.D.Thesis).

A-2001-1 J. Rousu: Efficient range partitioning in classification learning. 68+74 pp. (Ph.D. thesis)

A-2001-2 M. Salmenkivi: Computational methods for intensity models. 145 pp. (Ph.D. thesis)

A-2001-3 K. Fredriksson: Rotation invariant template matching. 138 pp. (Ph.D. thesis)

A-2002-1 A.-P. Tuovinen: Object-oriented engineering of visual languages. 185 pp. (Ph.D. thesis)

A-2002-2 V. Ollikainen: Simulation techniques for disease gene localization in isolated populations. 149+5 pp. (Ph.D. thesis)

A-2002-3 J. Vilo: Discovery from biosequences. 149 pp. (Ph.D. thesis)

A-2003-1 J. Lindström: Optimistic concurrency control methods for real-time database systems. 111 pp. (Ph.D. thesis)

A-2003-2 H. Helin: Supporting nomadic agent-based applications in the FIPA agent architecture. 200+17 pp. (Ph.D. thesis)

A-2003-3 S. Campadello: Middleware infrastructure for distributed mobile applications. 164 pp. (Ph.D. thesis)

A-2003-4 J. Taina: Design and analysis of a distributed database architecture for IN/GSM data. 130 pp. (Ph.D. thesis)

A-2003-5 J. Kurhila: Considering individual differences in computer-supported special and elementary education. 135 pp. (Ph.D. thesis)

A-2003-6 V. Mäkinen: Parameterized approximate string matching and local-similarity-based point-pattern matching. 144 pp. (Ph.D. thesis)

A-2003-7 M. Luukkainen: A process algebraic reduction strategy for automata theoretic verification of untimed and timed concurrent systems. 141 pp. (Ph.D. thesis)

A-2003-8 J. Manner: Provision of quality of service in IP-based mobile access networks. 191 pp. (Ph.D. thesis)

A-2004-1 M. Koivisto: Sum-product algorithms for the analysis of genetic risks. 155 pp. (Ph.D. thesis)

A-2004-2 A. Gurtov: Efficient data transport in wireless overlay networks. [B 141 pp. (Ph.D. thesis)

A-2004-3 K. Vasko: Computational methods and models for paleoecology. 176 pp. (Ph.D. thesis)

A-2004-4 P. Sevon: Algorithms for Association-Based Gene Mapping. 101 pp. (Ph.D. thesis)

A-2004-5 J. Viljamaa: Applying Formal Concept Analysis to Extract Framework Reuse Interface Specifications from Source Code. 206 pp. (Ph.D. thesis)

A-2004-6 J. Ravantti: Computational Methods for Reconstructing Macromolecular Complexes from Cryo-Electron Microscopy Images. 100 pp. (Ph.D. thesis)

A-2004-7 M. Kääriäinen: Learning Small Trees and Graphs that Generalize. 45+49 pp. (Ph.D. thesis)

A-2004-8 T. Kivioja: Computational Tools for a Novel Transcriptional Profiling Method. 98 pp. (Ph.D. thesis)

A-2004-9 H. Tamm: On Minimality and Size Reduction of One-Tape and Multitape Finite Automata. 80 pp. (Ph.D. thesis)

A-2005-1 T. Mielikäinen: Summarization Techniques for Pattern Collections in Data Mining. 201 pp. (Ph.D. thesis)

A-2005-2 A. Doucet: Advanced Document Description, a Sequential Approach. 161 pp. (Ph.D. thesis)

A-2006-1 A. Viljamaa: Specifying Reuse Interfaces for Task-Oriented Framework Specialization. 285 pp. (Ph.D. thesis)

A-2006-2 S. Tarkoma: Efficient Content-based Routing, Mobility-aware Topologies, and Temporal Subspace Matching. 198 pp. (Ph.D. thesis)

A-2006-3 M. Lehtonen: Indexing Heterogeneous XML for Full-Text Search. 185+3 pp.(Ph.D. thesis).

A-2006-4 A. Rantanen: Algorithms for ^{13}C Metabolic Flux Analysis. 92+73 pp.(Ph.D. thesis).

A-2006-5 E. Terzi: Problems and Algorithms for Sequence Segmentations. 141 pp. (Ph.D. Thesis).

A-2007-1 P. Sarolahti: TCP Performance in Heterogeneous Wireless Networks.(Ph.D. Thesis).

A-2007-2 M. Raento: TCP Exploring privacy for ubiquitous computing: Tools, methods and experiments. (Ph.D. thesis).

A-2007-3 L. Aunimo: Methods for Answer Extraction in Textual Question Answering 127+18 pp. (Ph.D. Thesis).

A-2007-4 T. Roos: Statistical and Information-Theoretic Methods for Data Analysis. 82+75pp. (Ph.D. Thesis).

A-2007-5 S. Leggio: A Decentralized Session Management Framework for Heterogeneous Ad-Hoc and Fixed Networks. 230 pp. (Ph.D. Thesis).

A-2007-6 O. Riva: Middleware for Mobile Sensing Applications in Urban Environments. 195 pp. (Ph.D. thesis).

A-2007-7 K. Palin: Computational Methods for Locating and Analyzing Conserved Gene Regulatory DNA Elements. 130 pp. (Ph.D. thesis).