Computational Prediction of Functional Similarity of CRMs

Hashem Koohy

Systems Biology Centre

Warwick

A thesis submitted for the degree of

PhilosophiæDoctor (PhD)

2010 October

1. Examination Advisor: Alison Rodger

2. External Examiner: Thomas Down

3. Internal Examiner: Matthew Turner

Day of the defense:

Signature from head of PhD committee:

Abstract

Transcriptional regulation of genes is fundamental to all living organisms. The spatial, temporal and condition-specific expression levels of genes are in part determined by inherited regulatory codes in non-coding regions of the DNA. A large set of methods have been proposed to detect conserved regions of regulatory DNA by means of sequence alignments. However, it has become clear that some regulatory regions do not show statistically significant alignments even in the presence of functional conservation. Therefore, detecting and characterising elusive regulatory codes remains a challenging problem.

In this thesis we develop and validate a novel computational alignment-free model for detection of functional similarity of regulatory sequences. We show that our model can detect functional links between pairs of sequences that do not align with a significant score. We apply the model to a) detect enhancers within the same genome that are likely to have similar functions and b) to detect functionally conserved enhancer regions in orthologous genomes. Our method finds regulatory codes that are common to groups of similar enhancers and consistent with previous biological knowledge.

The inputs for our model are two sequences that we wish to compare in terms of their functional similarity as well as a set of transcription factor motifs.

The mathematical framework of our model is built on two main components: In the first model component, each sequence is mapped to a vector of estimated occupancy levels for all motifs. These vectors are representing which motifs at what multiplicity and specificity are present in each sequence.

In the second model component, a statistical approach is established where we first estimate a probability distribution of motif occupancy levels for sequences that function similar to the template sequence. We then compute a statistical similarity score to evaluate if the sequences are more similar to each other than to random background sequences.

Two applications of this model are presented: First it is applied to a set of experimentally validated non-alignable enhancers from *D. melanogaster*. We show that:

- Our model can detect statistical links between these enhancers,
- Weak binding sites can make a strong contribution to sequence similarity,
- Our model treats statistically significant presence and absence of motifs symmetrically. Similarity of sequences, therefore, can be based on a combination of the two. We show examples of motifs making contributions to sequence similarity through their absence.
- Using our model, we can create a network of similarities among the fly enhancers. Groups of enhancers in this network show common regulatory codes. One of these regulatory codes is strongly supported by existing experimental data.

In the second application of our model we predict functional subregions of a known *D. melanogaster* enhancer. To achieve this, we first show that the model can detect the orthology of this enhancer between 10 Drosophila species. We then demonstrate how this statistical link can be used to predict functional subregions within this enhancer.

To those who valued the freedom and democracy so much that they risked their lives in Iranian streets and prisons.

Acknowledgements

First and foremost, I would like to acknowledge MOAC director Prof. Alison Rodger, director of Warwick Systems Biology Centre, Prof. David Rand and their management team for such a nice scientific environment and for their incredible support over the past four years.

Surely, this thesis would not have been possible without encouragement and efforts from my supervisors. I therefore like to express my gratitude to: Dr. Sascha Ott for his very valuable assistance and advice and Prof. Georgy Koentges for his very kind financial support.

I also wish to thank my advisory committee: Dr. Magnus Richardson, Dr. Till Bretschneider and Dr. Katherine Denby for their suggestions and comments along the path of the project.

I am indebted to many of my colleagues and friends including Dr. Miguel A. Juárez and people in both SO and GK's group for the discussions with them that led to further improvement of the thesis.

My PhD project was partially supported by HFSPO.

Last, but certainly not least, I would like to thank my wife Behnoosh and my son Behrad who have supported me through the difficult moments and helped to celebrate the joyous ones.

Contents

List of Figures

List of Tables

1

Introduction

The fascinating process of animal development starts from a single fertilized egg which develops into an embryo as embryonic cells divide and differentiate into diverse cell types leading to adult body formation and completion of the organism. This accurate process is regulated under an instruction written in the genomic DNA sequence and under a mechanism which is known as gene regulation.

The gene regulation mechanism in eukaryotic organisms takes place at a variety of different levels including gene localization inside the nucleus, transcription, RNA processing, mRNA stability and translation. In a multicellular animal, although different cell types possess the same genomic DNA sequence, they exhibit different gene expression profiles that are regulated at the transcription level. In other words, at this level, it is controlled when transcription starts and how much RNA is created.

The transcriptional regulation is one of the most fundamental mechanisms employed by the cell to ensure coordinated expression of its numerous genes. A key component of this process are the interactions between some proteins and corresponding DNA sequences. However, there are other components and events involved transcriptional regulation including chromatin structure and modification states. The interplay of these events in the complex control of transcription is sometimes called transcriptional regulatory code. Understanding which proteins are required for expression of different genes, where exactly they bind, under what conditions they are activated and which genes they are regulating is all part of deciphering the transcriptional regulatory code.

Despite of many advances in recent years (38; 62; 63; 80 and 20), the deciphering of the genome's regulatory code remains far from complete. This is mainly because of the complex control of transcription in eukaryotic cells. For example, transcriptional initiation of a gene demands combinatorial interactions of some proteins with the corresponding DNA subsequences, remodeling of local chromatin structure as well as the different types of histone modifications. In addition, in some genomes, the transcriptional regulatory sequences for a gene may be scattered over large regions and sometimes hundreds of kilobases away from the transcription starting sites. Therefore, unlike the protein coding sequences, integrating information over these various layers of control makes deciphering the regulatory code far from straightforward.

Our general goal is to contribute to on-going effort of deciphering the regulatory code. However, we should clarify that within the gene regulation machinery we only focus on the transcription level. Furthermore, by a regulatory code in this context we mean a distribution of different motifs in a genomic regulatory sequence (this will be defined in the following subsection) that are recognized by proteins in different levels and therefore directing different spatio-temporal expression patterns. Our emphasis will be to have a predictive and quantitative model of the transcriptional behaviours encoded by DNA sequence. We are ignoring the fact that a motif can be recognized by different proteins. We are assuming that the regulatory sequence is a linear sequence and do not take into account nucleosomes.

1.1 Basics and terminologies

In the following subsections we will provide the reader with some background and basic terminologies that will be used frequently throughout the rest of this thesis.

1.1.1 Regulatory sequences

Transcription factors (TFs) are proteins that regulate transcription, the process by which messenger RNA is synthesised from a DNA template. TFs facilitate

or inhibit recruitment of the RNA polymerase by binding to DNA, usually near the gene that they regulate. We should note that any transcription factor may recognize more than one site (mismatches and variations often occur). The collection of these short patterns are called motifs. Motifs are usually represented by position weight matrices (see Section 1.1.2). Detection of such short motifs in the DNA sequence is therefore of great importance in the study of gene regulation.

The genomic regions that are bound by TFs and control spatio-temporal gene expression patterns are called cis-regulatory modules (CRMs). These are called 'cis' because usually they are located at the same locus of the DNA molecule as their target genes. But 'trans' are usually referred to some proteins that bind to 'cis' elements (binding sites). These proteins are some times produced by some genes where as they dictate expression of different genes.

It is well-known that regulatory sequences makes only a small fraction of the 95% of the mammalian genome that does not encode proteins. But these regions are crucial in determination of the level, location and chronology of gene expression (54).

CRMs are built of clusters of binding sites (which are called regulatory elements) for specific sets of TFs and are thought to integrate the bound factors' cues. These regions broadly fall into two categories: *promoters* and *enhancers*. Promoters are proximal to the gene transcription starting site (TSS) and act as a binding site for RNA polymerase and from which transcription is initiated. Enhancers are, on the other hand, independent of the gene positions and can be found upstream, downstream or within a target or neighbouring gene (25). Enhancers (as their names imply) contribute to enhance the transcription.

An initial step in the analysis of any gene is the identification of CRMs.

1.1.2 Position weight matrices

The most common representation of binding sites is the position weight matrix (PWM) which is also called position specific scoring matrix (PSSM). In this representation, a motif with length L is represented by a $4 \times L$ matrix M where each possible base i, at each position j, is assigned a probability P_{ij} where $i \in \mathcal{A} = \{A, C, G, T\}$ and $j \in \{1, \cdots, L\}$. The probability of a specific sequence

given the model M is the product of probabilities of each particular nucleotide occurring at that position. For example, given the matrix M, a sequence like $S = S_1 S_2 \cdots S_L$ is associated with the probability $P(S|M) = \prod_{i=1}^{L} P_{S_i i}$.

Although an underlying assumption in a PWM is the independency of the positions in the binding site, this type of presentation is widely used and believed to be a reasonable approximation to the factor binding specificity.

The sequence logo that was first introduced by Schneider in (61) is a visual depiction of a PWM. In this graphical representation, each stack is associated with the information content of the base frequencies at that position which is $I_i = \log_2 |\mathcal{A}| + \sum_{\mathcal{A}} P_{S_i i} \times \log_2(P_{S_i i})$. According to this equation, positions can contain information in a range of 0 at positions where all four bases occur equally, to 2 bits at positions that are perfectly conserved, (for more information the reader is referred to 12).

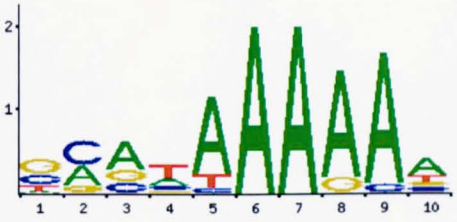

Figure 1.1: Logo representation of a Position Weight Matrix (Hunchback).

We must also note that the probability of a given sequence is usually calculated with respect to a background distribution (or model; denoted it by B) that the sequence might belong to. Markov models are the most commonly used models for the background distribution of nucleotides in different genomes. In this context we use a uniform zeroth order Markov model for the background model i.e., $P_B(A) = P_B(C) = P_B(G) = P_B(T) = 0.25$. Therefore the probability of sequence S, given this background model is $P(S|B) = (\frac{1}{4})^L$. This implies that the binding specificity of this sequence can be considered as $P_M(S)/P_B(S)$. The (base 2) log of this quantity is usually called the log odds ratio and denoted by $\mathcal{L}$ i.e., $\mathcal{L}(S) = log(P_M(S)/P_B(S)) = L \log 4 + \sum_{i=1}^{L} log P_{S_i i}$. A prior belief of binding likelihood can be added to this equation: $\mathcal{L}(S) = L \log 4 + \sum_{i=1}^{L} log P_{S_i i} + V$.

One common task in the analysis of regulatory DNA sequences is to search for potential transcription factor binding sites (TFBSs) within DNA regions of interest. For example, one may have a gene or set of genes whose expression is of interest and wants to find potential sites governing their regulation.

To accomplish this task one needs a database of regulatory motifs and an implementation of the PWM models in which the significance of the potential sites is determined. Among others, two databases that include comprehensive information about TFs are commonly used. The TRANSFAC database (47) provides extensive data on experimentally characterised TFs in several organisms, known binding sites, the PWM models and genes that are regulated by specific TFs.

Another recently developed and widely used database is JASPAR (60) which is an open access database for eukaryotic TF binding profiles. JASPAR has a smaller set but is believed to be less redundant than TRANSFAC. Two examples of widely used implementations of the PWM models are PATSER (26) and MATCH (33). However, in our analysis we used an implementation of the PWM model called BiFa tool (unpublished tool developed by N. Dyer and J. Reid). The reason why the BiFa tool is used in our model to score the binding strengths is explained in Subsection 3.1.

1.2 Motivations of the project

As we earlier mentioned, CRMs carry regulatory elements that are necessary to the specification of the spatio-temporal gene expression patterns. Understanding the rule by which modules process these regulatory elements is key to understanding the transcriptional processes.

The growing scientific interest in gene regulation means that it will a significant advantage to be able to detect the cis-regulatory modules in newly sequenced genomes that are homologous to known enhancers and/or promotors.

Despite the importance of the regulatory sequences in gene regulation, our ability to detect these sequences and also to predict their functions is very limited. This contrasts with non-coding sequences, where the wide-spread availability and

study of complementary DNAs (which are used for gene cloning) and proteins has made identification and prediction of their functions possible (54).

In the sequence comparison context, the most well-studied framework is measuring the sequence similarity between proteins or coding sequences in order to detect the homology. The basic local alignment search tool (BLAST) (3) is the most widely used alignment tool for this purpose. But, it is not very suitable in comparison of DNA regulatory sequences where, in contrast to the coding sequences, they demonstrate less significant alignments. This case may arise:

- where two sequences being compared are not orthologous (we note that the orthologous sequences are referred to those that share a common ancestor), yet functionally related. In Chapter 4, we will demonstrate a set of non-alignable enhancers in which a subset of enhancers is likely to be functionally related.

- where the sequences are evolutionarily highly diverged yet maintaining similar functions. Recently Hare et al. in (24) detailed evidence of some eve modules that produce near identical regulatory outputs where in more distantly related *D. wilstoni* and *D. virilis* groups only 29% of modules were conserved in these species.

Thus for comparison of DNA regulatory sequences alignment-free models are required.

The first alignment-free sequence comparison model proposed in 1986 by Blaisdell (8), and from that time it has received a great deal of attention by researchers. The overwhelming majority of reports about alignment-free models have been published over last 10 years (1; 20; 31; 62; 63; 77). These published models can be categorized into two groups.

Models in the first group are based on the principle that CRMs with similar functions should share some binding sites for the same transcription factors. These common binding sites are likely to be the key factor in driving similar expression patterns. In Chapter 2 we will provide the reader with an overview of some of key models in this group. We will see that these models are widely applicable to any type of data even protein sequences, but the results are, not

informative enough. For a review of these type of models the reader is referred to (44).

Models in the second group, on the other hand, are aimed at predicting spatio-temporal gene expression patterns from the regulatory sequences. In Chapter 2 we will review some of these models. Although these models advance our understanding of how genomic sequences are translated into transcriptional outputs, the complexity and extreme data dependency of the models in this group do not allow for a wide application of these models as a sequence comparison tool.

Having seen some advances in both of these groups, leading to more annotation of regulatory sequences and further understanding of regulatory systems, there has been very few successful attempts at using them for the comparison of regulatory modules. Indeed, our ability to quantify functional (dis)similarity of two regulatory modules, will help us to detect other enhancers in the same genome that are likely to have similar functions to the given enhancer. It also can be used to detect functionally conserved enhancer regions in orthologous genomes even if the enhancers do not align.

Here, we present a regulatory region scoring (RRS) model that overcomes this problem in some of its recent applications presented in this thesis. Our model takes as input a template sequence, a test sequence and a set of transcription factors motifs for which we need binding affinity and also the concentration of factors. As output, RRS provides the user with some statistical similarity scores and a list of factors that contribute to this (dis)similarity.

The mathematical and computational framework of the RRS has two main components. In the first model component, we establish a mathematical concept that represents what proteins, in what level of specificity and multiplicity are bound to the module. In the second model component, we estimate a probability distribution of motif occupancy levels for sequences that are functionally similar to the template sequence. We then compute a Bayes factor to evaluate if the test sequence is more similar to the template sequence or more similar to random background sequences.

Relative to the above mentioned families of models, the reader may wonder where the RRS stands in relation to existing models. Throughout Chapter 2 we shall try to convince the reader that there is a gap between these families

of models. The former family of models is defined very generally and is widely applicable, but some natural principles underlying transcriptional control, such as TF competition, motif degeneracy, and effects of weak binding sites, are completely ignored. Consequently, the results are less conclusive. Whereas the latter is based on a mechanistic understanding of the regulation of gene expression by predicting expression patterns using TF occupancy and interaction and is too dependent on a specific combination of data sets to be generally applicable. The key idea of the development of the RRS that we shall try to bring to the reader's attention throughout this thesis, was to enhance the conclusiveness of the results and lessen the data dependency of the model by borrowing the key ideas of each family of models so as to get more accurate results on a wider range of data.

This thesis consists of five chapters. In the first chapter, we provide the reader with a brief background and also the clarification and/or motivations of the problem. In the second chapter, we will briefly review some of the existing models, emphasising their strengths and pointing out their weaknesses. There has been an enormous amount of published work on alignment-free methods applied for detection and/or comparison of the regulatory modules as well as predicting expression profiles from the regulatory modules (recently, it has been also used as a motif finding tool see 21). Reviewing all of these reported models is out of the scope for this chapter. We consider those models that, to some extent, have had an influence on the establishment of our model. The third chapter is devoted to our regulatory region scoring model including its mathematical foundations and its computational framework. This is followed by two applications of the RRS. The first application is presented in Chapter 4 where the RRS is used to detect functional and/or evolutionary links between some non-alignable enhancers with a strong statistical significance. We will also identify groups of enhancers that are likely to be similarly regulated. Chapters 3 and 4 are based on our published paper (38).

Chapter 5 is devoted to the second application of our model. In this chapter, we first demonstrate how the RRS detects orthology between some fly species. Some of the orthologous sequences with (relatively) high statistical significant RRS scores are then used for our *in silico* predictions of functional subregions

of a *D. melanogaster* enhancer that are likely to drive expression patterns in a subset of projection neurons in the *D. melanogaster* olfactory system.

It is widely thought that the targeting specificity of the projection neurons in the fly olfactory system is controlled by a transcriptional code but very little of the underlying mechanism is understood. Therefore we are aiming to open some new insights into this poorly understood notion by predicting functional subregions and their key regulators using our RRS model.

The underlying project of this chapter is a close collaboration with our collaborators at Stanford University . Here the emphasis is on the bioinformatical side of the project (For the biological side of this project the reader is referred to Chapter 4 of 71). This project is still ongoing and a manuscript of both bioinformatical and biological results of this project is under preparation.

Finally, we would like to further clarify that each chapter in this thesis ends with a conclusion subsection in which we provide the reader with brief findings as well as some future directions specific to that chapter. We believe that this will help readers who are interested in only some parts the thesis to follow their interests easily.

2

Existing Models

It is widely accepted that cis-regulatory modules are key for establishment of precise spatio-temporal gene expression patterns. Some recent studies show that CRMs may function similarly in different species despite substantial sequence divergence (45 and 24). This implies that, firstly, alignment-based sequence comparison tools are not applicable for further decoding the conserved function of such CRMs and secondly, that some CRMs must share common patterns that drive almost identical regulatory outputs but possibly with different arrangements of binding sites. When different, but functionally related enhancer loci in the same species are considered, then alignment-based tools are not normally suitable for regulatory sequence comparisons as these sequences are not orthologous.

Recently, there has been a great deal of attention on alignment-free methods to further reveal the mechanism of transcription control (see 78). Among these methods, two families are of particular interest for us within this project. We call them *data intensive* and *general* models. The former is based on a mechanistic understanding of the regulation of gene expression by predicting expression patterns using TF occupancy and interaction and is too dependent on a specific combination of data sets to be generally applicable. The latter family of models is defined very generally and is widely applicable, but some natural principles underlying transcriptional control, such as TF competition, motif degeneracy, and effects of weak binding sites are completely ignored. Consequently, the results are less conclusive.

In the following two sections we shall review some of the models in any of these families.

2.1 Data intensive models

Recent studies show that some CRMs with the same function may have strikingly different architectures (10). A big challenge in the field is now to predict the activity of a CRM based on its organisation. This has been recently attempted by many researchers, but among others three closely related computational modelling approaches (in order: 62; 80 and 32) have been at the center of debate by making new insights of our understanding from the regulatory code. These models are aimed at predicting spatio-temporal gene expression patterns from the regulatory sequences. They all follow the same idea but differ mainly in input and slightly in structure. As representative of data intensive models, we will review these three approaches in this section.

2.1.1 A thermodynamic model for prediction of gene expression patterns

In theoretical gene regulation frameworks, thermodynamically motivated models (for the sake of simplicity, from now on we will call them thermodynamic models) are based on the assumption that the level of gene expression is proportional to the equilibrium probability that RNA polymerase is bound to the promoter of interest. This is perhaps the most attractive feature of these models for theoretical scientists interested in gene regulation, because it avoids the difficult task of computing gene expression from the concentration of proteins produced by the gene of interest.

These models are established, however, based on some different assumptions that can be problematic. The equilibrium assumption itself can be considered the most critical one that according to our best knowledge has not been systematically evaluated yet (see 7 and 63). The second problematic assumption in these models is that the gene expression level is considered proportional to the probability of

promoter occupancy by the RNA polymerase. This assumption can mean ignorance of several different mechanisms that do occur between polymerase binding and the existence of a functional gene product. For a more detailed review of thermodynamic models in gene regulation frameworks including their modeling and applications, the reader is referred to (7 and 6).

Despite of these critical assumptions, there are some reports showing that these models are very instructive and predictive (see 20; 22; 62 and 66).

In this subsection, we well review only one of these thermodynamic models that has been established by Segal et. al. (62), in which the reader can see that the developers are strongly motivated by some previous work for example (14; 17; 74; 78 and 59).

Similar to the others, this model is based on the above mentioned thermodynamic equilibrium assumption. In other words the probability of polymerase occupancy is computed from the intrinsic equilibrium affinities and concentrations of the transcription factors. The gene expression level is considered to be proportional to the polymerase occupancy.

This thermodynamic model for prediction of gene expression patterns made use of TF expression levels as well as the arrangement and quality of their binding affinity to predict the expression profile of an arbitrary DNA sequence. The authors achieved this by generating a model based on the biochemical properties and binding site preferences of eight key TFs (Bicoid, Hunchback, Caudal, Kruppel, Giant, TorRE, Knirps and Tailless) of the early Drosophila segmentation network. For previous related work see

This model (in this context we call it thermodynamic model) is based on a thermodynamic equilibrium (between DNA-binding proteins) assumption. The probability of polymerase occupancy is computed from the intrinsic equilibrium affinities and concentrations of the transcription factors (TFs). The gene expression level is considered to be proportional to the polymerase occupancy.

This model takes into account some important aspects of TF-DNA interaction including competition of TFs for TF binding sites, self-cooperativity of TFs, and the effects of weak binding sites.

2.1.1.1 Mathematical framework of the model

The thermodynamic model takes three input parameters: Module sequence, concentration of any of the factors under analysis at any anterior-posterior (AP) position and also binding affinity of the factors. As output, it provides the reader with a prediction of expression pattern that the given sequence might have as a profile over the AP axis. Figure 2.1 on page 14 is an schematic depiction of this model.

The mathematical structure of this model is built by two main components.

Throughout the first model component, each factor views the sequence in a unique way - called binding landscape - depending on its recognition specificity at any set of concentrations of the DNA binding proteins. The range of this binding landscape is key to cooperative and competitive binding interactions between the factors and the DNA sequence. According to this binding landscapes, one may see a particular arrangement of molecules along the DNA sequence which includes specification of the precise position and orientation at which each molecule is bound. Any of these distributions of a set of molecules bound to the sequence is called a binding configuration or more precisely a valid binding configuration by not allowing overlapped molecules (from now on by a configuration we will mean a valid binding configuration).

It is worth pointing out that different interpretations of this idea have been applied for other organisms including bacteria (7), yeast (20) and mammals (22 and 66).

It is then argued that any of these distinct configurations convey a distinct transcriptional behaviour.

Therefore, according to this framework the key question turns to further understand these binding configurations. For this, all possible configurations are taken into account and each configurations is associated with a statistical weight.

We should note that in this context, the binding affinity that can be considered as the strength of binding that is measured by using a position weight matrix model. In other words, lets assume that $S = S_1 \cdots S_l$ and position weight matrix M are given. Then the binding affinity of S is defined as $\frac{P(S|M)}{P(S|B)}$, where the numerator means probability of the sequence using the weight matrix model M

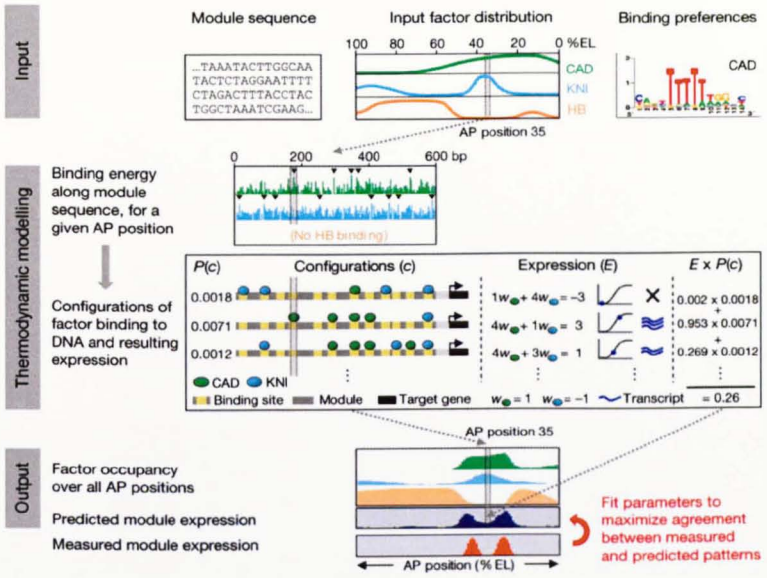

Figure 2.1: An schematic illustration of the thermodynamic model. This figure has been taken from (62).

and the denominator means the probability of the sequence given the background model B.

Assuming that molecules bind independently, a statistical weight of a configuration is defined as the product of contributions of any of the molecules bound to the sequence within the given configuration. The contribution of each of the binding events is in turn computed from the concentration of the corresponding factor and affinity of the binding site that the molecule is occupying. Thus, for a set of n transcription factors i.e. $\{TF_1, \cdots, TF_n\}$, if we assume that N molecules m_i of these factors are bound to the sequence within the configuration c, then we can write:

$$W(c) = \prod_{i=1}^{N} \tau(m_i) \times F(m_i, P_i) \tag{2.1}$$

where P_i is the interval of the DNA sequence that has been occupied by the molecule m_i, $\tau(m_i)$ is the concentration of the m_i and $F(m_i, P_i)$ is the binding affinity of the interval P_i for molecule m_i. It worth pointing out that firstly the linear dependency does not model saturation effects, and we are not dealing with

situations where concentrations are known in this work, but simply assume a constant and identical concentration for all TFs. Secondly, for a given state of a thermal system - in statistical mechanics and thermodynamic contexts - $F(m_i, P_i)$ is called Boltzmann factor and is defined as the exponential of minus its energy which is measured in $k_B T$ units. More precisely, the energetic contribution of the binding of molecule m_i to the sequence from position P_i to position $P_{i+L_{i-1}}$ with L_i being the binding interval length is defined as:

$$F(m_i, P_i) = e^{-\frac{E_i}{k_B \cdot T}} \tag{2.2}$$

For more details the reader is referred to (7 and 63).

The normalised statistical weight of each configuration is then defined as the probability of that configuration, that is :

$$P(c) = \frac{W(c)}{\sum_{c' \in C} W(c')} \tag{2.3}$$

All in all, at the end of the first model component the user is provided with the occupancy distribution of the molecules on the target DNA sequence.

The second model component on the other hand translates this occupancy distribution into a level of gene expression in other words $P(E|c)$ which is discussed below.

We should recall that the probability of the gene expression is assumed to be proportional to the probability of the RNA polymerase binding and is denoted by $P(E)$. The overall probability that polymerase is binding is obtained from the weighted sum of the polymerase binding at every configuration, with the weight of each configuration is being its probability:

$$P(E) = \sum_{c \in C} P(c) P(E|c) \tag{2.4}$$

in which $P(E|c)$ is interpreted as a translation of expression level driven by the configuration c. The underlaying assumption at this level is that each factor bound in the configuration contributes independently to the expression outcome, with activators contributing positively and repressors contributing negatively. The authors employ a logistic function to translate these contributions into expression. In other words, if we assume that a configuration c has built up by

binding N molecules $m_1, \cdots, m_N$ at positions $P_1, \cdots, P_N$ to the DNA sequence, then the probability of expression can be expressed as:

$$P(E|c) = logit(w_0 + \sum_{i=1}^{N} w_{m_i}) = \frac{1}{1 + e^{-(w_0 + \sum_{i=1}^{N} w_{m_i})}} \tag{2.5}$$

where w_0 is the basal expression level and w_i is the expression contribution of the molecule i. From this equation one may see that the parameters are the same for all sequences and also in longer sequences all of the factors would be able to simultaneously have their effects. To overcome this problem the authors normalised the input of the logistic function by dividing it by the length of the sequence.

2.1.1.2 Parameter fitting and validation of the model

As parameter fitting of this model, 44 gap and pair-rule gene modules with known expression patterns were used. By comparing the predicted expression patterns of these models with measured expression patterns, and devising a learning algorithm they trained the parameters of the model. For any factor these parameters included a) the absolute concentration of the factor *in vivo*, b) the transcription rate resulting from its interactions with the basal machinery, c) the strength of binding cooperativity and d) the strength of the PSSM which was representing the factors' binding preferences. The model then was used to predict expression patterns for 11 *D. melanogaster* and 15 *D. pseudoobscura* modules. The result of this analysis is presented in Table 2.1 on page 16.

Species	number of modules	good	fair	poor
D. melanogaster	11	4	4	3
D. pseudoobscura	15	2	9	4

Table 2.1: Results of predictions of expression patterns for 11 *D. melanogaster* and 15 *D. pseudoobscura* modules. Predictions were subjectively classified into three categories: good, fair and poor.

2.1.1.3 Conclusion

This thermodynamic model advances our understanding of how genomic sequences are translated into transcriptional outputs. It shows that knowing the TF concentration at different AP positions as well as the arrangement and quality of the binding sites can be sufficient to explain the segmentation pattern in fly species.

The knowledge about these two key parameters of the model, however, is accounted as the main drawback of the model. On one hand detailed knowledge of biochemical TF properties is often not available. On the other hand, detailed knowledge of some spatial expression patterns of a number of related enhancers and their key regulators is required which is again not always available. Furthermore, the number of configurations is an exponential function of the length of the sequence and the number of TFs which makes computation of occupancy level of factors very expensive and almost impossible for genome wide applications.

Finally, according to (63), although the underlying thermodynamic assumption of this model has been successfully used in some other models, it remains unclear how and even whether regulatory systems equilibrate.

2.1.2 Global predictions of regulatory module activity

In Section 2.1.1 we argued that a key factor for the thermodynamic model was knowledge about the concentration of proteins which are rarely available. To overcome this problem, Zinzen et al. (80) decided to predict enhancers' activity solely from their TF binding site patterns. They established a novel approach based on comprehensive catalogue of CRMs involved in Drosophila mesoderm development that are bound by five key factors.

In this section we briefly review this model. For the sake of simplicity we call it *Zinzen* model.

2.1.2.1 Computational framework

Using chromatin immunoprecipitation combined with microarray (ChIP-chip) assays Zinzen et al. determine the genome wide distribution of binding sites of five

key factors of mesoderm and muscle (Twist, Mef 2, Tinman, Bagpipe, Biniou) at 5 different time points (spanning the majority of stages when each TF is expressed), resulting in high resolution binding data for 15 developmental conditions.

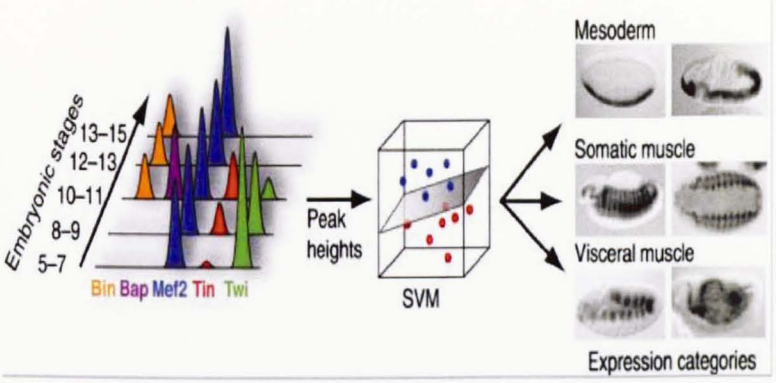

Figure 2.2: An schematic illustration of the Zinzen model. This figure has been taken from (72).

With this protocol, they found in total 19522 binding sites that were clustered into 8008 distinct CRMs. In order to investigate whether combinatorial transcription factor binding is predictive of CRM activity, they collected a reference data set of enhancers (CAD : the CRM activity database) with characterized tissue-specefic expression pattern. They then identified 310 among 8008 CHIP-CRMs that were overlapping with CAD. From these 310, 87 fall into one five broad and partially overlapping categories: early mesoderm, visceral (gut) muscle, somatic muscle, meso and somatic muscle and visceral.

They trained a machine learning algorithm called support vector machine (SVM) with the respective CRM activity information. This was first used for 310 known CRMs, by excluding each CRM in turn for testing, and training the SVM with the remaining ones.

2.1.2.2 Conclusion

A novelty in the Zinzen model was that the developer used a ChIP approach not only to predict the location of CRMs but also to predict their spatio-temporal activity. Also, the user does not need detailed knowledge of the system, including

estimates of transcription factor concentrations, their affinity for various sequence motifs and cooperativity and competition between transcription factors.

Impressively, the model was able to predict the expression pattern of the modules with a high accuracy, in other words, 71% of the predictions turn out to be correct: the enhancers drive expression of transgenic reporters specifically in the predicated regions and not in other mesodermal tissues.

Despite of the high accuracy of predictions, the Zinzen model is still intensively based on *in vivo* activity data which is not often available.

In comparison of the Zinzen model with the thermodynamic model, one can argue that both are novel strategies for predicting CRM expression pattern, but are strongly dependent on availability of experimental data. The thermodynamic model looks powerful when a detailed knowledge of concentration of key factors at different developmental stages is available, but it does not need a whole map of CRMs. On the other hand, the Zinzen model, does not require detailed biochemical information about regulators but rather requires *in vivo* TF binding and CRM activity data.

Another drawback of the presented Zinzen model is that it is based on a machine learning algorithm where its robustness and reliability is not addressed therefore further applications of this model in a wider range of data is required and will provide further insights into its usability.

Finally, the authors in (80) argue that their previous data for binding profiles of transcription factors were not of enough quality to model the CRM activity. However, there is no clear definition of quality level of the data that will be enough for the CRM activity prediction. On the other hand, for generating high resolution data, they performed Chip-on-chip on each TF at consecutive time points in 5 different developmental stages. This procedure provided them with binding data for 15 developmental conditions. But, as far as we can see, there is no relationship between this binding data with the level of accuracy of the model. In other words, how much of this binding data is required for some statistically significant predictions.

We should leave reviewing of this model at this level, the interested reader is referred to (57 and 72) for more details.

2.1.3 Quantitative analysis of CRMs using pattern generating potentials

Recently, a new computational approach for annotation of genomic sequences was established by Kazemian et al. (32). This model that we will call it the *regression-based* model is based on a pattern generating potential and similar to the thermodynamic model, it uses both the DNA binding specificity and concentration of transcription factors. However, as will be described through the next subsection, the binding specificities as well as the input for the logistic function are computed quite differently.

The regression-based model is the first model in this family that can be used in a genome-wide manner to identify modules by scanning genomic sequences for the potential to generate all or part of the expression pattern of a flanking gene.

As output, it provides the user with a location of a module as well as an estimation of its potential expression pattern. Furthermore, based on an *in silico* genetic analysis, a transcriptional regulatory network is constructed in which each edge depicts the direct contribution of individual factor with an associated estimate for its statistical significance.

In the following subsection we will provide the reader with more details of mathematical and computational framework of the regression-based model.

2.1.3.1 Computational framework

We would like to recall that the thermodynamic model is constructed based on two components, one that is estimating the occupancy level of factors in a given sequence based on Equation 2.3 and the other that is translating this occupancy level into an expression pattern using Equation 2.5. But a key issue with computations of these quantities is the enormous number of configurations that increases exponentially as a function of length of the sequence and the number of factors. Although the authors used a dynamic programming approach to address this computational cost, it still prevents the model from having a wider range of applications.

The regression-based model, on the other hand, is a new strategy to tackle this problem. The mathematical structure of the model is similar to the ther-

modynamic model built of two parts. First, a cross-species comparison strategy is used and transcription factor binding specificity profiles are computed. Next, a logistic regression function is employed to combine factor motif scores with transcription factor expression information to predict the module activity. The details of these procedures are as follows:

- **Computation of binding specificities:** The basic idea of this approach was that CRMs with conserved activity across Drosophila species will maintain some binding activity for each TF while binding sites in non-functional regions will be less conserved. They used the Hidden Markov Model-based Stubb (67) program to generate genome profiles of binding motif scores for a set of 10 TFs including BCD, CAD, HB, KNI, KR, GT, HKB, TLL, FKH and CIC. For the sake of generality we will denote the set of TFs as: $\mathcal{F} = \{F_1, \cdots, F_N\}$.

  They then created a multi-species motif profile by averaging the motif profiles from the *D. melanogaster* and 10 other Drosophila genomes (averaging scores from orthologous 500bp regions). However, the averaging was not just the additive mean of the scores. In order to reflect the evolutionary distances among the species, the motif score of a region was defined as a random variable evolving according to the Brownian motion process along the branches of a phylogenetic tree. The average was thus defined as the expected tree-wide average of this variable given its observed value in the extant species. Using this approach, each module l was associated with a motif score C_i^l for any $i \in \mathcal{F}$. For more details of this averaging scheme the reader is referred to (73).

- **Employment of a logistic regression model:** Within this model, the AP axis is divided into 100 bins. Lets assume that the concentration of any factor $i \in \mathcal{F}$ at bin b is equal to γ_{ib}. Then the predicted expression level for the CRM l at bin b is defined as:

$$E_{l,b} = logit(w_0^l + \sum_{i \in \mathcal{F}} w_i \gamma_{ib} C_i^l) = \frac{1}{w_0^l + e^{-(\sum_{i \in \mathcal{F}} w_i \gamma_{ib} C_i^l)}} \tag{2.6}$$

where the w_0^l is the basal expression level of CRM l and w_i is called the regression coefficient (positive for activators and negative for repressors) for each factor i. The basal expression and regression coefficient are free parameters of the regression model and are learned by applying the model to 46 modules with known expression profiles.

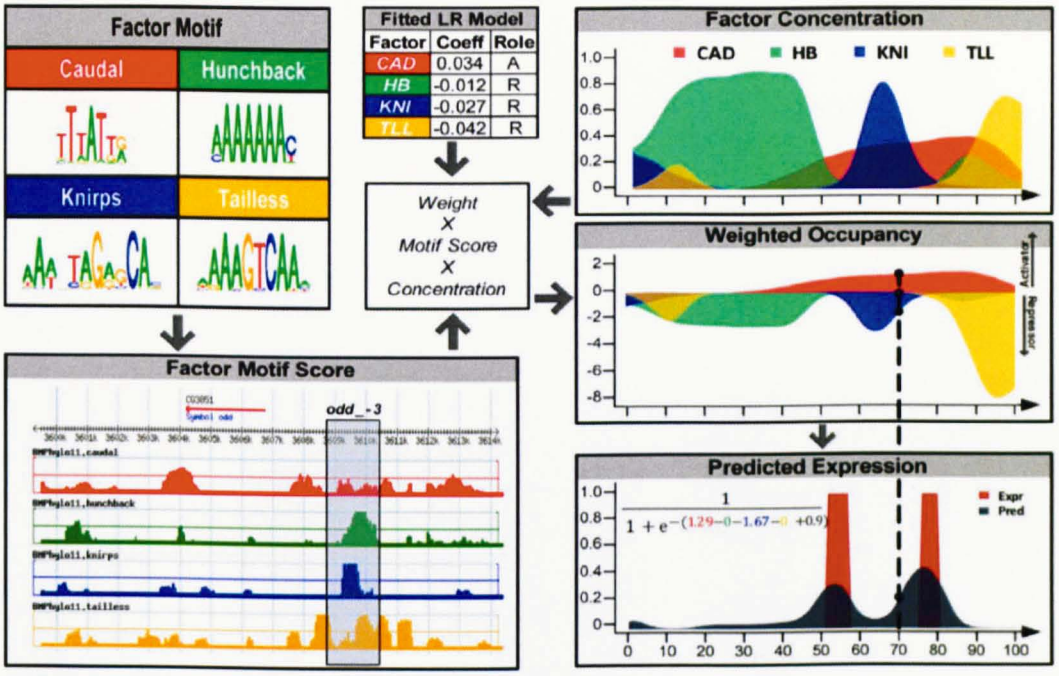

Figure 2.3: An schematic illustration of the regression-based model. This figure has been taken from (32).

From the Equations 2.5 and 2.6 we can see that both the thermodynamic and the regression-based model are using the same logistic models to translate the differently computed occupancy level of motifs into the expression level, but the input of logistic functions is different. The authors are claiming that this logistic model is simpler than the one used in thermodynamic model in a sense that they have fewer number of free parameters to be learnt from data (2 vs 3) and that the regression-based model has the advantages of incorporating multiple species comparisons and of computation that is order of magnitude faster. But from our point of view, the ability of incorporating multi-species comparisons makes regression-based model more dependent to data than its counterpart. It

is worth mentioning that although a direct comparison of these models has not been presented, the authors are claiming that the regression-based model is as effective as the thermodynamic model.

2.1.3.2 Detection of regulatory modules with the regression-based model

In the paper under review, the authors presented a measure of similarity between a genomic sequence activity (predicted expression by regression-based model) with a gene's endogenous expression pattern. This scoring scheme was called pattern generating potential (PGP). Given a predicted expression profile (real numbers between 0 and 1 for each bin along AP axis) and endogenous expression profile (again numbers ranging from 0 to 1) the PGP was defined as:

$$PGP = 0.5 \times (1 + \frac{\sum_b E_{g,b} \times \hat{E}_{g,b}}{\sum_b E_{g,b}} - 3 \times \frac{\sum_b (1 - E_{g,b}) \times \hat{E}_{g,b}}{\sum_b (1 - E_{g,b})}) \qquad (2.7)$$

where $E_{g,b}$ is endogenous expression value of the gene g in bin b and $\hat{E}_{g,b}$ is the predicted expression value. We should note that the $\frac{\sum_b E_{g,b} \times \hat{E}_{g,b}}{\sum_b E_{g,b}}$ is in fact the average of the predicted expression in expressed bins and is called the reward term whereas the $\frac{\sum_b (1-E_{g,b}) \times \hat{E}_{g,b}}{\sum_b (1-E_{g,b})}$ is the average of the predicted expression in non-expressed bins and called the penalty term. The difference of reward and penalty is indeed the PGP score, the coefficient 3 in the penalty term of Equation 2.7 is just a weight. The PGP scores are linearly scored as $y = 0.5 = 0.5x$.

This scoring scheme inferred a genome wide application of the regression-based model for detection of CRMs in the following way: A genomic region consisting of gene transcript and 10kb of its upstream and downstream region is scanned with windows of fixed length (for instance 1kb, colour-filled rectangles in Figure 2.4). The predicted expression profile of each window (open blue and green rectangles in the same figure) is then compared with the endogenous expression (open red rectangle) of the gene leading to PGP scores that are plotted as a function of the genomic coordinate of the window (as is depicted in Figure 2.4).

The PGP was first tested on 22 genes regulated by 46 CRMs and then applied to a collection of 144 genes where the authors identified 123 putative CRMs from 68 genes.

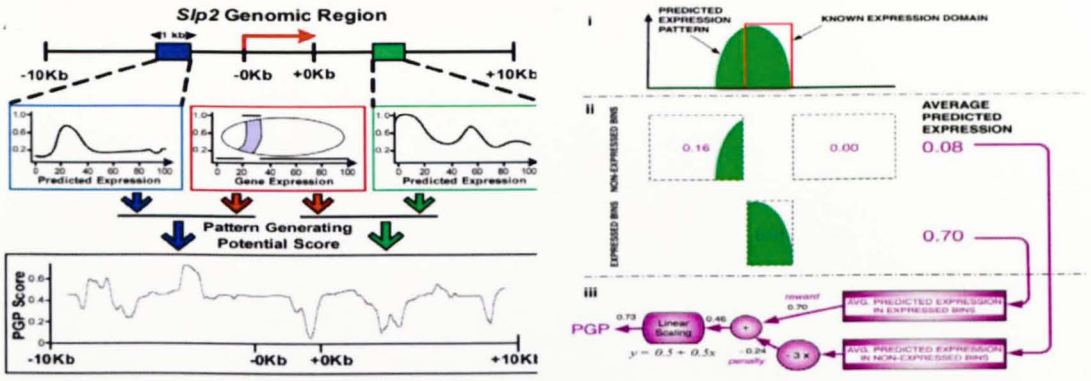

Figure 2.4: An schematic illustration of the PGP scoring scheme. This figure has been adapted from (32).

2.1.3.3 Inferring transcription factor and regulatory module interaction networks

The authors then used this formalism to infer factor-module interactions as a network. The principal idea was simple: the PGP method was working as a function of binding specificities of TFs as well as the concentration of factors. Therefore it was possible to computationally assess the contribution of each TF by setting its concentration to 0 and compare this *in silico* mutant to the concentration of the wild type. For any TF, in order to test the statical significance of its mutation, they measured the root mean square error (RMSE) between predicted expression profile of 1000 random permutations of that TF's concentration (blue histogram in Figure 2.5) and the true expression. They set up an empirical $p-$value for the RMSE which reflects how important this factor is to the CRM expression.

Top right panel in part A of Figure 2.5 on page 25, depicts the true (red) and predicted (blue) expression profiles. The reader also can see the effect of *in silico* mutant of three factors (CAD, HB and TLL) in red border rectangles and the corresponding RMSE score as a red dot in any of the histograms.

2.1.3.4 Conclusion

The regression-based model can be used for genome-wide predictions of CRMs and their potential activity as well as to examine the effect of each motif on each putative CRM and empirical assessment of its statistical significance. In this

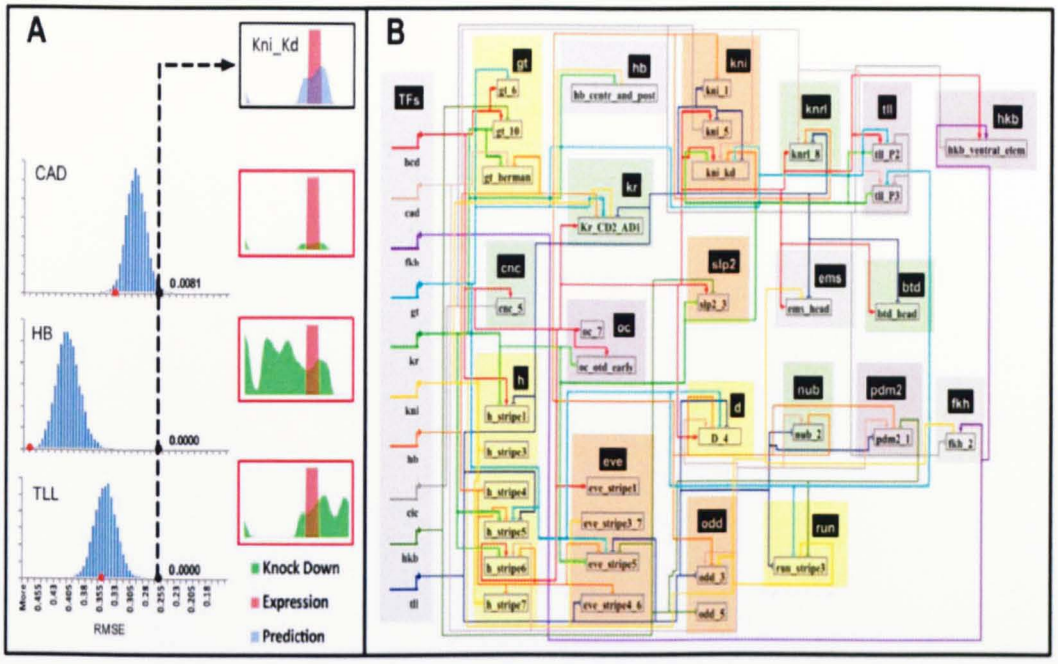

Figure 2.5: A: Illustration of formalism for construction of factor-module interactions network. B: Predicted regulatory network for 10 TF and 35 experimentally characterised CRMs. This figure has been taken from (32).

sense, the approach presented by Kazemian et al. provides the user with a multifunctional method for the analysis of CRMs that promises further annotation of the regulatory sequences.

However, unlike the thermodynamic model, the regression-based model does not capture some of the known mechanistic features of a regulatory module function such as the synergy between pair of motifs. And unlike to the Zinzen model, it lacks the *in vivo* context of ChIP data.

We should mention that in an attempt to compare the performance of the regression-based model to the Zinzen model, the developers of the regression-based model replaced the motif scores of 8 TFs with ChIP scores and retained the regression-based model using these data, but it did not lead to superior predictions.

The regression-based model is only applicable to systems where the adequate expression data are available for relevant TFs, CRMs and target genes. Thus it

seems more dependent to data than two other counterparts.

Finally, all three above reviewed studies are based on some machine learning algorithms. Therefore, the abundance as well as the quality of a training set for these machine learning algorithms is a fundamental requirement that might affect the quantity and quality of their results. A direct comparison of these three models, will reveal the robustness of these algorithms in particular with respect to the over-fitting problem. Obviously further investigation of disagreements of such a study will enhance our understanding of the regulatory code.

2.2 General models

This family of alignment-free methods is mostly based on the rationale that functionally similar sequences must share some common words. Within these methods each sequence is mainly associated with a vector of k-mer counts. A distance function for these vectors is then defined (1; 8; 31; 77 and 40).

In this section we will be reviewing only three of these methods (in a chronological order) as representatives of this family. Throughout, we are hoping to convince the reader that this family of models is defined very generally and therefore is widely applicable, but some natural principles underlying transcriptional control such as TF competition, motif degeneracy, cooperativity of binding sites, effects of weak binding sites and concentration of factors are completely ignored.

2.2.1 Metrics for comparing regulatory sequences on basis of pattern counts

The key idea behind this model (we call it the *Poisson-based* model (77)) was that the presence of common motifs in the regulatory regions of two sequences (genes) might be considered as a measure of similarity, and presence of different motifs as a sign of dissimilarity. Therefore common putative regulatory properties of genes can be captured by defining a pattern count-based similarity and/or dissimilarity function.

2.2.1.1 Computational framework

The functional similarity of two sequences a and b in the Poisson-based framework is defined as:

$$M^{ab} = S^{ab} - \alpha D^{ab} + \beta \tag{2.8}$$

where S^{ab} and D^{ab} are respectively similarity and dissimilarity metrics (as defined below), α is a positive weighting parameter, which can be tuned arbitrarily to give more emphasis on the common (low values) or distinct (high values) occurrences between two sequences and β is offset to ensure the that metric is always positive.

In this model the data set is considered as a matrix N, containing n rows (one per sequence) and p columns (one per pattern). N_i^a corresponds to the number of occurrences of pattern i in sequence a. In order to define a (dis)similarity between two sequences (a and b) a Poisson distribution is employed.

Each pattern i is characterised by a prior probability f_i, indicating the probability to find an occurrence at any position of a sequence. Prior probabilities can be calculated either on the basis of the data set itself, or on the basis of an external background model. The expected number of occurrences m_i is obtained by multiplying the prior probability f_i by the number of possible positions T for the pattern:

$$m_i = f_i T = f_i(L - w + 1) \tag{2.9}$$

where L is the length of the sequence and w the length of the pattern. (For simplicity, assume all the sequences have the same lengths). Let us denote the cumulative function of the Poisson distribution by $F(x, m_i)$, that is the probability to observe at most x occurrences, when the expected value is m_i. Thus for a single gene a and single pattern i, the probability to observe at least N_i^a occurrences is obtained by:

$$P(x \geq N_i^a) = \begin{cases} 1 - F(N_i^a, m_i) & \text{if } N_i^a > 0 \\ 1 & \text{if } N_i^a = 0 \end{cases} \tag{2.10}$$

It is clear that when N_i^a increases (i.e. for over-represented patterns) $F(N_i^a, m_i) \rightarrow 1$ and consequently $P(x \geq N_i^a) \rightarrow 0$ i.e., the low values of $P(x \geq N_i^a)$ correspond to overrepresented patterns.

The contribution of each pattern i to the similarity of a pair of sequences is then calculated on the basis of the probability of common counts. For this, lets assume that $C^{ab} =< C_1^{ab}, \cdots, C_p^{ab} >$, where $C_i^{ab} = min(N_i^a, N_i^b)$ is the number of common counts for pattern i.

Now the probability to observe at least C_i^{ab} occurrences of pattern i in each sequence, is the product of the probabilities (under the assumption of independency):

$$P(x \geq C_i^{ab}) = \begin{cases} [1 - F(C_i^{ab}, m_i)]^2 & \text{if } C_i^{ab} > 0 \\ 1 & \text{if } C_i^{ab} = 0 \end{cases} \tag{2.11}$$

This probability is then converted into a similarity metric as:

$$s_i^{ab} = 1 - P(x \geq C_i^{ab}) \tag{2.12}$$

reflecting how exceptional is to find at least C_i^{ab} common occurrences of pattern i in a pair of sequences. For a multi-variate similarity, the score then can be defined either as additive mean which is defined as:

$$S_{add}^{ab} = \frac{1}{p} \sum_{1=1}^{p} s_i^{ab} \tag{2.13}$$

or to consider a joint probability simultaneously, and applying geometric mean:

$$s_{prod}^{ab} = 1 - \sqrt[p]{\prod_{i=p}^{p} P(x \geq C_i^{ab})}. \tag{2.14}$$

From this similarity metric one can see that :

- A pair of sequences that do not share a common motif are obtaining 0 as their similarity score.
- High number of occurrences of a single motif or multiple occurrences of different motifs increase the similarity score.
- Patterns with low prior probabilities contribute more than those with higher prior probabilities.

For establishment of a dissimilarity metric, the author calculates the probability of the distinct occurrences, i.e., those found in one sequence but not in the other one. For this, it is assumed that pattern i has occurred N_i^a and N_i^b times respectively in sequences a and b, and that $N_i^a \leq N_i^b$, then the contribution of this motif to dissimilarity can be defined as:

$$d^{ab}_{distinc_i} = |F(N_i^b, m_i) - F(N_i^a, m_i)|, D^{ab}_{distinct} = \frac{1}{p}\sum_{i=1}^{p} d_i^{ab} \tag{2.15}$$

In order to capture the degree of over-represention of a motif which is indicated by low values of the probability to observe at least x occurrences: $P(x \geq N_i^a) = 1 - F(N_i^a - 1, m_i)$, the author defined another catalogue of dissimilarity metric as:

$$\begin{aligned} d^{ab}_{over_i} &= |P(x \geq N_i^a) - P(x \geq N_i^b)| \\ &= |F(N_i^a - 1, m_i) - F(N_i^b - 1, m_i)| \\ D^{ab}_{over} &= \frac{1}{p}\sum_{i=1}^{p} d_i^{ab} \end{aligned} \tag{2.16}$$

From Equations 2.15 and 2.16 on page 29, one can see that: a) a motif with the same number of occurrences in both sequences has a 0 contribution to the dissimilarity definition, b) high number of distinct counts of a motif and also high number of different motifs occurring with different counts in both sequences increases the dissimilarity.

Finally, the author defines the mixed metric as Equation 2.8 on page 27, in which some key points are worth highlighting: a) motifs found in both sequences are contributing positively whereas motifs found in one sequence but not in the other are contributing negatively, b) score 0 means that either none of the sequences contains any occurrences of any motif or common and distinct occurrences of motifs are compensating each other's effect.

2.2.1.2 Conclusion

The Poisson-based model is easy to implement and computationally efficient algorithm. However, there are some points that we would like to bring them to the

reader's attention. Firstly, there is no significance defined to the final similarity (dissimilarity) metric i.e., Equation 2.8. In other words, for instance, 5 motifs with the same number of occurrences in both sequences has the same effect as 100 motifs with same number of occurrences in both sequences. Secondly, there is no evidence to show why the underlying Poisson distribution is an appropriate distribution for the occurrences of motif in a sequence, in particular, this means that we are assuming that the occurrences of a motif in a regulatory sequence is only by chance, which seems unrealistic. Thirdly, a big concern that the user might have about this model is that he/she requires a prior knowledge about motifs. Finally, as a minor technical point, it might worth mentioning that from a mathematical point of view the term 'metric' is inappropriate in particular for Equation 2.8. For instance,we know that as a (mathematical) metric (function) the score 0 corresponds only to the same sequences which is not true in this definition.

2.2.2 Fixed-length word distribution model

The model we will be reviewing in this section is called *D2z* and established by Kantorovitz et al. (31). The D2z model is based on comparing the frequencies of all fixed-length words in the two sequences. In this way sequences are mapped to to vectors by the counts of (for instance) k-mers. The vectors obtained in this way, represent the original sequences with a fixed resolution k. Then the basic logic is that similar sequences will share more words. This is being quantified by defining different techniques.

2.2.2.1 Computational framework

Lets assume that $\mathcal{A} = \{A, C, G, T\}$ is the alphabet set, and the background model is a Markov model of order ω (we note that different sequences may fit different background models). We suppose that $A = A_1 A_2 \cdots A_{n_1}$ and $B = B_1 B_2 \ldots B_{n_2}$ are two sequences that we wish to measure their similarities in terms of distributions of $k-$mers. The D_2 statistics (42) is defined to be the number of $k-$mer matches between two sequences A and B, including overlaps. It is

originally computed as:

$$D_2(A,B) = \sum_{(i,j)\in I} Y(i,j) \tag{2.17}$$

where $Y(i,j)$ is the indicator variable between the $k-$words starting at position i in A and B, and the index set $I = \{(i,j) \ : \ 1 \leq i \leq n_1 - k + 1, \ and \ 1 \leq j \leq n_2 - k + 1\}$. One may note that:

$$D_2(A,B) =< N^A, N^B >= \sum_{w\in W} N_w^A N_w^B \tag{2.18}$$

where similar to what we defined in Section 2.2.1.1, N_w^A is the number of occurrences of the word w in sequence A and $w \in W = \{w_1, w_2, \cdots w_{4^k}\}$.

In order to measure the number of standard deviations by which the observed value of D_2 deviates from the mean, the authors presented a normalised version of the $D2$ score:

$$D2z(A,B) = \frac{D_2(A,B) - E(D_2)}{\sigma(D_2)} \tag{2.19}$$

where $E(D_2)$ and $\sigma(D_2)$ are the expectation and the standard deviation of the D_2 respectively. For computations of these parameters, two different computational algorithms based on independent and identically distributed random variables IID, and also Markov model (MM) is presented.

2.2.2.2 Conclusion

In applications where several different distributions are to be compared the normalization of the D2z becomes very useful as different background distributions of the sequences are taken into account. This makes it possible to compare sequences from different species.

Besides, we can see that this model is relatively easy to implement and also can be adapted to a more limited set of $k-$mers, in order to reduce the computational expenses. It can be used for any sort of sequences (even protein sequences). However, it is too theoretical. In other words, some particular limitations of this method can be listed as:

1. Not all functional motifs in a pair of sequences are in the form of 6-mers. So by considering only k-mers as patterns underlying functional similarity of a pair of sequences, some motifs which contribute to the gene expression pattern may be overlooked.

2. Not all k-mers are biologically meaningful words, hence using all 6-mers may mean introducing some noise to the model and furthermore, we may want to compare two sequences just based on a subset of meaningful words.

3. Within the D2z framework, degeneracy of TF binding motifs is not accounted for. So different 6-mers are treated separately even if they only differ in one base.

4. The framework does not allow for a sequence and its reverse complement to be combined for the purposes of assessing possible TF binding.

2.2.3 Identifying regulatory modules by word profile similarity

Most recently, Garmay Leung et al. (40) came up with a different idea for comparison of vectors of counts of $k-$mers associated to two sequences. They presented their solutions as a model called word profile hits or *WPH* in short. In this framework, given a sequence (for example a CRM), the WPH algorithm uses its word composition to search other putative CRMs with similar word composition. In the following subsections we shall provide the reader with more details of the WPH framework. We should mention that in this study the authors were only interested in compositions of $8-$mers. Therefore, by a word profile of a sequence they mean its $8-$mer composition.

2.2.3.1 Computational framework

In this framework, the similarity of two sequences is determined by comparing the degree of word overlap between two profiles with the expected overlap given the number of words in each sequence. To see this in more details, we need to

establish some notations. We will use '8−mer' and 'word' to refer to the same object in this section.

Lets assume that two sequences A and B are given and we wish to measure their functional similarity based on WPH framework. Lets also assume that $W(A)$ and $W(B)$ are the sets of all 8−mers occurred in sequences A and B respectively. A 1−neighbour of a word $w \in W(A)$ is a word w' which has maximum 1 mismatch with w. The set of words in 1−neighbourhood of $W(A)$ is denoted as $W'(A)$ (the number of allowed mismatches is considered as a free parameter). We should note that $W(A) \subseteq W'(A)$. A word $w \in W(A)$ contributes to the observed word overlap $ov_{A\to B}$ if a 1−neighbour of w occurs in B. With this definition, it is clear that each pair of sequences defines two overlaps ($ov_{A\to B}$ and $ov_{B\to A}$) that lead to two similarity scores $z_{A\to B}$ and $z_{B\to A}$ which are defined in the rest of this subsection.

The probability of the overlaps is calculated by employing a Poisson distribution with mean $\lambda = |W(A)/n|$ where $n = 32896$ is the number of unique 8−mers (a word is mapped to itself and its reverse complement). Therefore the probability that a given word w occurs at least once in A is equal to:

$$p_w(A) = 1 - e^{-|W(A)/n|} \tag{2.20}$$

and the probability of a 1−neighbour of a given word w in A is:

$$p_{w'}(A) = 1 - e^{-|W'(A)/n|} \tag{2.21}$$

This implies that a given word w occurs in A and its 1−neighbour occurs in B with the probability:

$$p_{ov}(A \to B) = p_w(A)p_{w'}(B) \tag{2.22}$$

Let $X^w_{A\to B}$ be the indicator variable representing whether the word w occurs in A and one of its 1-neighbours say w' occurs in B.

The authors then assume that each word occurs independently and therefore one can use a binomial distribution with the following properties:

$$
\begin{aligned}
Pr[X^w_{A\to B} = 1] &= p_{ov}(A \to B) \\
X_{A\to B} &= \sum_{w\in\mathcal{A}} X^w_{A\to B} \\
E[X_{A\to B}] &= Pr[X^w_{A\to B} = 1] \cdot n \\
\sigma_{A\to B} &= \sqrt{Pr[X^w_{A\to B} = 1] \cdot n \cdot (1 - Pr[X^w_{A\to B} = 1] \cdot n)}
\end{aligned} \tag{2.23}
$$

note that $\mathcal{A} = \{A, C, G, T\}$.

Similar to D2z model, the overlap score is defined as:

$$z_{A\to B} = \frac{V_{A\to B} - E[X_{A\to B}]}{\sigma_{A\to B}} \tag{2.24}$$

where $V_{A\to B}$ is the actual overlap, $E[X_{A\to B}]$ is the expected overlap and $\sigma_{A\to B}$ is the standard deviation. However, to make the scores symmetric, they defined the final similarity of sequences as $Z(A, b) = min(z_{A\to B}, z_{B\to A})$. Taking minimum is to ensure that similarity requires many words in A to have $1-$neighbours in B and vice versa.

In a series of analyses, the authors noticed that upon applying this scoring scheme sequences with similar GC-content are clustered together. Therefore they decided to bin together words with equal GC-ratio and calculating the probability of word overlap for each bin. That is they argued that for a fixed word length k, there are n_r words for each GC-ratio $r = 0, 1/k, 2/k, \cdots, 1$. Let $W_r(A)$ be the set of words in A with GC-ratio of r, and similarly $W'_r(A)$ be the set of words in the $1-$neighbourhood of $W_r(A)$. Then the word occurrence probability for a given GC-ratio r is as: $p_{w_r}(A) = 1 - e^{-|W_r(A)/n|}$ and $p_{w'_r}(A) = 1 - e^{-|W'_r(A)/n|}$. Similar to Equation 2.22 the corresponding pairwise word overlap probability between sequences A and B for words with a given GC-ratio is: $p_{ov_r}(A \to B) = p_{w_r}(A)p_{w'_r}(B)$ and overall probability of word overlap is defined as sum over all possible GC-ratios:

$$p_{ov}(A \to B) = \sum_r \frac{n_r}{n} p_{ov_r}(A \to B) \tag{2.25}$$

Figure 2.6 on page 35 shows how this scheme can be used to identify sub-sequences in the target sequence with similar sequence composition to a given

CRM's word composition: First the CRM is split to subsequences of length 500bp. Each of these subsequences then is associated with their word profiles. Finally, using the above mentioned scoring scheme, the target sequence is searched for subsequences with similar word profiles.

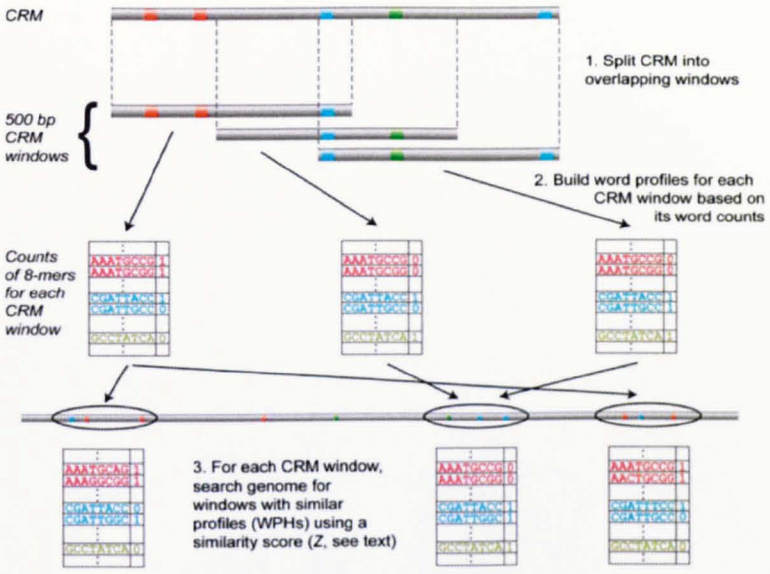

Figure 2.6: An schematic overview of WPH model. The figure has been taken from (40).

2.2.3.2 Conclusion

The reader might have noticed that the WPH is a combination of the Poisson-based and D2z model. In comparison to the Poisson-based model, it provides a better estimation of the mean for the Poisson distribution. In comparison to the D2z model, they do not consider distribution of all $k-$mers in both sequences, but those $k-$mers that up to $1-$neighbourhood have occurred in both sequences. Furthermore, considering $1-$neighbourhood of a word equal to its own occurrence is one step development, while comparing to D2z.

But similar problems still remain:

- By only considering $8-$mers, some functional words are overlooked.
- By considering all $8-$mers, it is very likely to introduce noise to the system.

- Occurrence of each word in this scheme is equiprobable.
- There is no guarantee that one might not need to do some other corrections (for example for AT rich sequences, similar to GC-biases correction)

3

Regulatory Region Scoring (RRS) Model

Some recent comparative studies have revealed that regulatory regions can retain function over large evolutionary distance, even though the DNA sequences are divergent and difficult to align. It is also known that such enhancers can drive very similar expression patterns. This poses a challenge for in the *in silico* detection of biologically related sequences, as they can only be discovered using alignment-free methods. Our main objective in this chapter is to present a new computational framework called Regulatory Region Scoring (RRS) model for detection of functional conservation of regulatory sequences using predicted occupancy levels of transcription factors of interest. Our goals are:

1. To be able to detect functionally similar enhancer regions even if the enhancer regions do not align.

2. To find groups of similar enhancers and determine relevant sequence features shared among enhancers within a group.

The RRS model takes as input a pair of sequences and a set of TF motifs. We call one of the sequences the *template sequence* and the other the *test sequence*. The task is to judge whether the test sequence has the potential to drive similar expression patterns as the template sequence, assuming expression is driven by the given set of motifs. We do not use any cutoff for probabilities of binding of

these motifs to the sequences and so allow weak binding events and even absence of motifs to contribute to sequence similarity. The output from the RRS model is a statistical similarity score and a list of motifs that contribute to that similarity score.

The model is built of two main components: one component associates each sequence with a mathematical vector reflecting which proteins with what multiplicity and what specificity have the potential to be bound to the sequence. We call the elements of these vectors *motif occupancy values* or, in short, *o-values*. These vectors give an indication of the potential enhancer function of the given sequences. As the reader might notice, some parts of this component are a modification of the thermodynamic model that was reviewed in Subsection 2.1.1, meaning that to some extent we are accepting both equilibrium assumption and that the gene expression level is considered proportional to the probability of promoter occupancy by the RNA polymerase. The second component estimates a probability distribution of motif *o-value* vectors for sequences that function similar to the template sequence. We then compute a Bayes factor to evaluate if the test sequence is more similar to the template sequence or more similar to random background sequences (Figure 3.1 shows a simplified schematic illustration of the RRS concept).

We like to draw the reader's attention to the point that the RRS has been developed to be able to learn parameters from both randomly picked and randomly generated sequences. However, as the reader will notice, within this project we preferred to learn the model from the randomly picked sequences. This is because we believed that it is not possible to capture all the genome features (such as repeat elements, low complexity DNA and ect) with randomly generated sequences.

In the rest of this chapter we first provide the reader with mathematical foundations of the RRS model in Section 3.1. The main focus of this section therefore is establishing the feasibility of computation of the o-values. This section is very mathematically oriented. For those readers with less mathematical background, we will try to keep the coherence of the story in the next sections by repeating some of the essential equations in a less mathematically oriented language. Then, in Section 3.2, we show how the *o-values* are defined and computed. Section 3.3 is

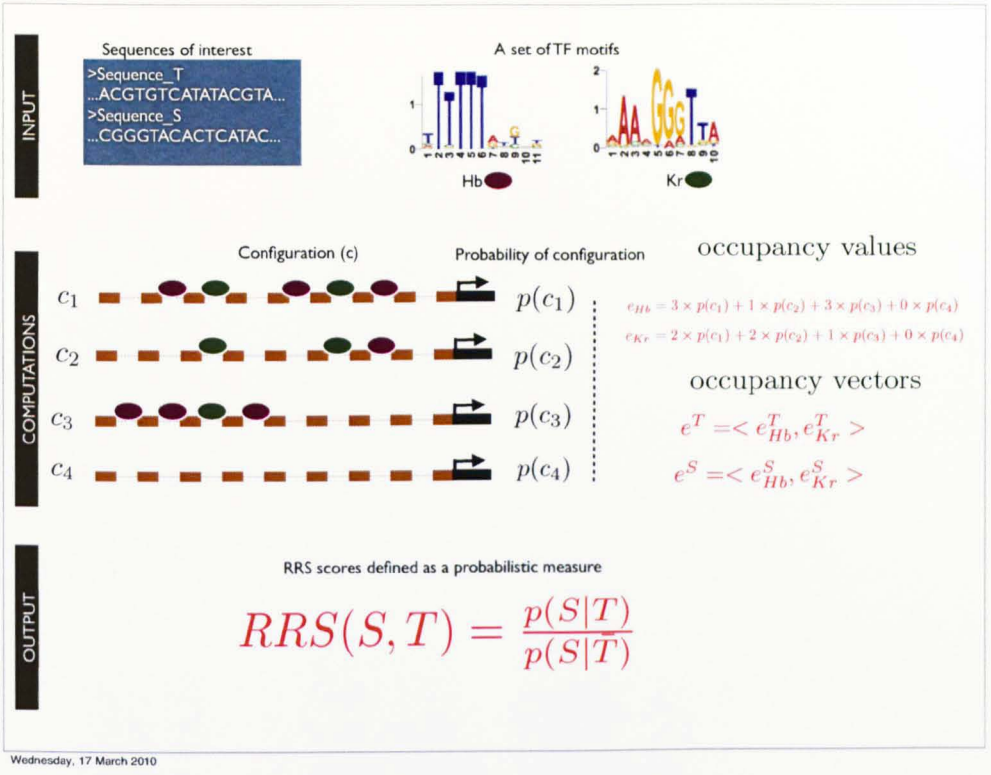

Figure 3.1: A simplified schematic illustration of the RRS model. As input, the RRS model takes two sequences and a set of transcription factor motifs (here just Hb and Kr for the sake of simplicity). Probabilities of configurations of TFs on the sequences (four possible configurations illustrated) can be computed. Using dynamic programming, expected numbers of proteins binding each motif (motif *o-values*) are computed integrating over the space of all possible configurations. The vector of motif *o-values* for each sequence is taken to represent its potential regulatory function to some extent. We then define a probabilistic score for the similarity of a pair of sequences. The score is defined as the ratio of the probability that the motif e-values for test sequence S were drawn from the same distribution as for template sequence T over the probability of e-values for S being drawn from the distribution for random background sequences.

devoted to establishment of our similarity score function. The parameter fitting procedure of our model is discussed in Section 3.4.

3.1 Mathematical framework of the RRS model

Throughout this section we are going to provide the reader with the mathematical foundations of the RRS model. For this, a given sequence is first associated with a set of binding configurations. Then any of these configurations in turn is associated with its probability. In the following, the expected number of occurrences of a motif is defined. Apart from presenting these terminologies and definitions in detail, we will put a particular emphasis on mathematical feasibility of computations of the probability of each configuration and also the expected number of occurrences of each motif in a given sequence.

In what follows, we will assume a template sequence T, a test sequence S and a set of transcription factor motifs $\mathcal{M} = \{M_1, \cdots, M_n\}$. We shall denote the length of a sequence T by L_T or simply by L, if there is no risk of confusion and the length of a motif M by $|M|$.

Definition 1 *A site s in a sequence T with length L is defined as an element of $\mathcal{M} \times \{1, \ldots, L\}$, i.e., $s = (M, P_i)$ for some $M \in \mathcal{M}$, and $|M| \leq P_i \leq L$ where P_i is the position of the last nucleotide of the motif in the sequence T.*

We use the term configuration to denote a particular arrangement of protein molecules along the DNA sequence, which is defined by the sites at which each molecule is bound to the sequence. In other words:

Definition 2 *A configuration c with N molecules bound to a sequence is defined as $c = \{(M_i, P_i) | 1 \leq i \leq N,\ M_i \in \mathcal{M}\}$, where M_i is the i-th molecule bound at a position P_i.*

Valid configurations are those in which sites do not overlap:

Definition 3 *A valid configuration is a configuration $c = \{(M_i, P_i) | 1 \leq i \leq N,\ M_i \in \mathcal{M}\}$ in which for any given (M_{i_1}, P_{i_1}) and (M_{i_2}, P_{i_2}), either $P_{i_1} \leq P_{i_2} - |M_{i_2}|$ or $P_{i_2} \leq P_{i_1} - |M_{i_1}|$ holds.*

From now on, we will be only interested in valid configurations and we will denote the set of valid configurations by C. However, for the sake of our argument we like to introduce a particular subset of C. That is the set of those configurations

that have exactly j occurrences of the motif M up to position P_i of the sequence and there is no site after P_i, i.e.,

$$C_{i,j}^{M} = \{c \in C \mid (\forall (M, P_k) \in c, (P_k \leq P_i) \wedge (|\{(M', P_k) \in c | \ M' = M\}| = j))\}$$

The following lemma shows that the set of $C_{i,j}^{M}$s is indeed a partition of the set C. We note that for a given sequence T with length L and a motif M with length $|M|$, the maximum number of occurrences of M over T is $J_M = L/|M|$.

Lemma 4 *Assume that T is a sequence with length L, M is an arbitrary motif and J_M is the maximum number of occurrences of M over all valid configurations, then*

1. $C_{L,i}^{M} \cap C_{L,j}^{M} = \emptyset$ *for any* $0 \leq i \neq j \leq J_M$;
2. $\bigcup_{k=0}^{k=J_M} C_{L,k}^{M} = C$

Proof. The first part is a direct application of the definition. For the second part, let's assume that $c \in C$ is an arbitrary configuration. If there is no occurrences of M over c, then $c \in C_{L,0}^{M}$. If there are more than zero occurrences of M over c, then we may assume that the j is the position of the last occurrence of M over c that will imply that $c \in C_{L,j}^{M}$. This means that $\bigcup_{k=0}^{k=J_M} C_{L,k}^{M} \supseteq C$. The other side of this inclusion is obvious. ■

Now let's assume that a configuration c with N molecules bound to the sequence is given i.e., $c = \{(M_i, P_i) | 1 \leq i \leq N, \ M_i \in \mathcal{M} \}$. If we further assume that molecules bind independently then the statistical weight of this configuration is defined as the product of the contribution of each of the binding events. But the contribution of each molecule is in turn a function of binding affinity and concentration parameter. In other words:

Definition 5 *If we denote the sequence at binding interval of molecule M_i at position P_i by B_i, then the statistical weight is defined as:*

$$W(c) = \prod_{i=1}^{N} \frac{p(B_i|M_i)}{p(B_i|\bar{M}_i)} = \prod_{i=1}^{N} \frac{p(M_i|B_i)}{p(\bar{M}_i|B_i)} \times \frac{p(\bar{M}_i)}{p(M_i)} \tag{3.1}$$

In which, $p(B_i|M_i)$ means the probability of subsequence B_i using the the corresponding PSSM model and $p(B_i|\bar{M}_i)$ means the probability of subsequence B_i given the background model (uniform $0-$order Markov model in our case). $\frac{p(B_i|M_i)}{p(B_i|\bar{M}_i)}$ is the contribution of each binding molecule, $\frac{p(M_i|B_i)}{p(\bar{M}_i|B_i)}$ is considered as the binding affinity and $\frac{p(\bar{M}_i)}{p(M_i)}$ is considered as the concentration parameter.

In our model, the BiFa tool (see Subsection 5.2.2) is used to score the strength of bindings i.e., $\frac{p(M_i|B_i)}{p(\bar{M}_i|B_i)}$. This is because in the BiFa tool a Bayesian approach is implemented to compute these scores which is equivalent to what has been used in (62). As our model, to some extent, is a modification of (62) therefore one may agree that it was reasonable to use an equivalent scoring scheme. Besides, according to the developers (see Figure 3.2 on page 42) of the BiFa tool, it is more sensitive than the currently used model in the TRANSFAC database.

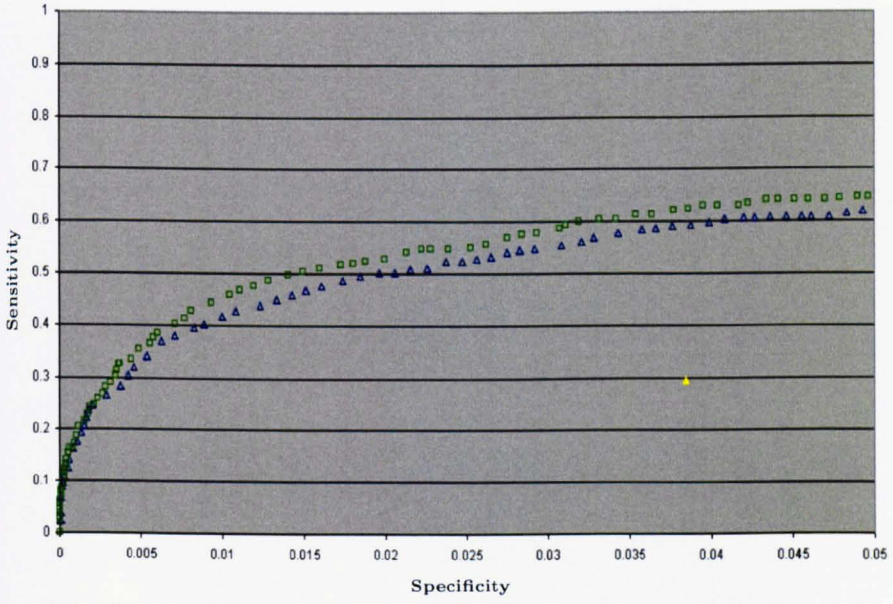

Figure 3.2: Comparison of sensitivity of BiFa scores vs two other models. Green: Bayesian model used in BiFa for scoring binding strengths. This scoring model has been used in our algorithm. The model underlying the Blue curve is a frequentist statistic provided as an alternative within BiFa. That is, given a position weight matrix score x, what is the likelihood of observing a score $\geq x$ by chance. The yellow triangle shows the performance of the score implemented in the TRANSFAC database. This figure has been provided by the developers of BiFa tool.

We should note that according to this definition, statistical weight associated to the empty configuration i.e., the configuration without any molecule bound to it is 1 (product over an empty set). We also note that this definition enables weak binding events to be included in the model. Assume that in a configuration c we have a molecule that has been weakly bound to the sequence many times. If, for the sake of simplicity we assume an equal binding affinity a ($a > 1$) in K positions, then the contribution of this factor to the $W(c)$ is equal to a^K. Depending on K, this might be a strong contribution.

The probability of each configuration c is then defined as $p(c) = \frac{W(c)}{\sum_{c \in C} W(c)}$. We use the same dynamic programming technique as in (62) to compute this probability. The core of our model, however, is where we define the expected number of occurrences of each motifs in a sequence. For a given sequence T and a given motif M, the expected number of occurrences of M over T is defined as:

$$e_M^T = \sum_{c \in C} p(c) I_M(c) \tag{3.2}$$

where $I_M(c)$ is the number of occurrences of motif M in the sequence over the configuration c. This equation is of particular interest as it contains both the multiplicity and specificity of a binding event of a protein to the sequence respectively in $I_M(c)$ and $p(c)$. However, as already mentioned, our main emphasis in this section is the mathematical proof of feasibility of computation of this term. To achieve this we need to establish some more notations.

Notation 6 *In the rest of this chapter we define:* $P_{L,j}^M := \sum_{C_{L,j}^M} p(c)$, $W_{L,j}^M := \sum_{C_{L,j}^M} W(c)$ *and* $Z := \sum_{c \in C} W(c)$

where L is the length of the sequence T, M is the motif, j is the number of occurrences of M in sequence T.

Lemma 7 *For a sequence T with length L and a motif $M \in \mathcal{M}$,* $\sum_{j=0}^{j=J_M} P_{L,j}^M = 1$.

Proof.

$$\begin{aligned}
&\sum_{c\in C} p(c) = 1 \\
\Rightarrow &\sum_{\cup_{j=0}^{j=J_M} C_{L,j}^M} p(c) = 1 && \text{(According to Lemma 4)} \\
\Rightarrow &\sum_{C_{L,0}^M} p(c) + \cdots + \sum_{C_{L,J_M}^M} p(c) = 1
\end{aligned}$$

■

Lemma 8 *For a sequence T with length L and a motif $M \in \mathcal{M}$, $Z = \sum_{j=0}^{J_M} W_{L,j}^M$.*

Proof. Similar to proof of the Lemma 7. ■

Corollary 9 *For a sequence T with length L, a motif $M \in \mathcal{M}$, $0 \leq i \leq L$ and $0 \leq j \leq J_M$, $ZP_{i,j}^M = W_{i,j}^M$*

Proof. Proof is straightforward from Notation 6 and Lemmas 7 and 8. ■

Lemma 10 *Suppose T is a sequence of length L and M is a motif from $\mathcal{M}$ then*

$$e_M^T = \sum_{j=1}^{J_M} P_{L,j}^M \cdot j$$

Proof.

$$\begin{aligned}
e_M^T &= \sum_{c\in C} P(c) I_M(c) \\
&= \sum_{\cup_{j=0}^{j=J_M} C_{L,j}^M} p(c) \times j && \text{(According to Lemma 4)} \\
&= \sum_{j=0}^{J_M} (\sum_{C_{L,j}^M} p(c) \times j) \\
&= \sum_{j=0}^{J_M} (j \times \sum_{C_{L,j}^M} p(c)) \\
&= \sum_{j=0}^{J_M} (j \times P_{L,j}^M)
\end{aligned}$$

■

Assume that $M \in \mathcal{M}$ and c is a configuration in $C_{i,j}^M$. We remember that c is a configuration in which up to position i of the sequence there are exactly j occurrences of M. One may consider three possibilities for this configuration.

- (M, i) is an element of c. Let us denote the set of these type of configurations with C_1^M, meaning that position i has been occupied by M.
- (M', i) is an element of c, where $M \neq M' \in \mathcal{M}$. Let us denote the set of these type of configurations with C_2^M, meaning that position i has been occupied by another motif.
- There is no element X in $\mathcal{M}$ such that $(X, i) \in c$. In other word, position i of the sequence is left unoccupied. Let us denote the set of these type of configurations with C_3^M.

It is not difficult to observe that $C_{i,j}^M = C_1^M \cup C_2^M \cup C_3^M$ and consequently:

$$P_{i,j}^M = \sum_{c \in C_{i,j}^M} p(c) = \sum_{C_1^M} p(c) + \sum_{C_2^M} p(c) + \sum_{C_3^M} p(c) \tag{3.3}$$

The following three lemmas are in fact main tools for the proof of the main theorem of this section. We should recall that in the following B is the sequence at the binding interval of molecule M, i.e., B is the $S[i - |M|, i])$ subsequence.

Lemma 11 *For any motif $M \in \mathcal{M}$ and with the notations shown above, the following equation holds:*

$$\sum_{C_1^M} p(c) = P_{i-|M|,j-1}^M \frac{p(B_i|M)}{p(B_i|\bar{M})}.$$

Proof. Suppose $c \in C_1^M$, then (M, i) is an element of c. This also implies that c has exactly $j - 1$ occurrences of M up to position $i - |M|$. If we assume $|C_1^M| = t$, then we can write:

$$\begin{aligned}
\sum_{C_1^M} p(c) &= p(c_1) + \cdots + p(c_t) \\
&= \frac{W(c_1)}{Z} + \cdots + \frac{W(c_t)}{Z} \\
&= \frac{1}{Z}\left(\prod_{c_1} \frac{p(B_1|M_1)}{p(B_1|\bar{M}_1)} + \cdots + \prod_{c_t} \frac{p(B_t|M_t)}{p(B_t|\bar{M}_t)}\right)
\end{aligned}$$

we know that the last site in any of configurations $c_1, \cdots c_t$ is (M, i). By separating the contribution of (M, i), we can re-write the last equation as:

$$\frac{1}{Z}\Big(\prod_{(M_1,P_1)\neq(M,i)} \frac{p(B_1|M_1)}{p(B_1|\bar{M}_1)} \times \frac{p(B|M)}{p(B|\bar{M})} + \cdots + \prod_{(M_t,k_t)\neq(M,i)} \frac{p(B_t|M_t)}{p(B_t|\bar{M}_t)} \times \frac{p(B|M)}{p(B|\bar{M})}\Big) =$$

$$\frac{1}{Z}\Big(\prod_{(M_1,k_1)\neq(M,i)} \frac{p(B_1|M_1)}{p(B_1|\bar{M}_1)} + \cdots + \prod_{(M_t,k_t)\neq(M,i)} \frac{p(B_t|M_t)}{p(B_t|\bar{M}_t)}\Big) \times \frac{p(B|M)}{p(B|\bar{M})} \leq$$

$$P^M_{i-|M|,j-1} \times \frac{p(B|M)}{p(B|\bar{M})}$$

To prove the other side of this inequality, let's suppose that c is an element of $C^M_{i-|M|,j-1}$. In other words c is a configuration with $j-1$ occurrences of M up to position $i-|M|$. We can write:

$$\begin{aligned} p(c)\frac{p(B|M)}{p(B|\bar{M})} &= \frac{W(c)}{Z}\frac{p(B_i|M)}{p(B_i|\bar{M})} \\ &= \frac{1}{Z}\Big(\prod_c \frac{p(B'|M')}{p(B'|\bar{M}')} \times \frac{p(B|M)}{p(B|\bar{M})}\Big) \\ &= p(c_1) \qquad \text{(where } c_1 \text{ is an element of } C^M_1\text{)} \end{aligned} \tag{3.4}$$

This implies that

$$\sum_{C^M_{i-|M|,j-1}} p(c)\frac{p(B|M)}{p(B|\bar{M})} \leq \sum_{C^M_1} p(c_1)$$

which completes the proof. ■

Lemma 12 *For any motif $M \in \mathcal{M}$, the following equation holds:*

$$\sum_{C^M_2} p(c) = \sum_{M'\in\mathcal{M}, M'\neq M} P^M_{i-|M'|,j}\frac{p(B'|M')}{p(B'|\bar{M}')}$$

Proof. Let us assume that c is an element of C^M_2. Then according to the definition of C^M_2, there exists an $M' \neq M$ in $\mathcal{M}$ such that $c = (M', i)$. With a similar argument to the proof of Lemma 11, we may write:

$$\sum_{C^M_2} p(c) \leq \sum_{M'\in\mathcal{M}\setminus\{M\}} P^M_{i-|M'|+1,j} \times \frac{p(B'|M')}{p(B'|\bar{M}')}$$

For the other side of this inequality again we suppose that c is an element of $C^M_{i-|M'|,j}$ where $M' \neq M$ is a motif in $\mathcal{M}$. Therefore c has j occurrences of M up to position $i - |M'|$. Similar to the proof of Lemma 11, we can write:

$$\begin{aligned} p(c)\frac{p(B'|M')}{p(B'|\bar{M}')} &= \frac{W(c)}{Z}\frac{p(B'|M')}{p(B'|\bar{M}')} \\ &= \frac{1}{Z}\Big(\prod_{C^M_{i-|M'|,j}} \frac{p(B|M)}{p(B|\bar{M})} \times \frac{p(B'|M')}{p(B'|\bar{M}')}\Big) \\ &= p(c_2) \qquad \text{(where } c_2 \text{ is an element of } C^M_2\text{)} \end{aligned}$$

and so the proof is completed. ■

Lemma 13 *For any motif $M \in \mathcal{M}$, the following equation holds:*

$$\sum_{C^M_3} p(c) = P^M_{i-1,j}$$

Proof. Any configuration $c \in C^M_3$ has j occurrences of M up to position i, but the position i itself is left unoccupied. This means that c is a configuration in $C^M_{i-1,j}$. And obviously any configuration $c \in C^M_{i-1,j}$ has j occurrences of M up to position i but the position i itself remains unoccupied. Meaning that c is an element of C^M_3. Therefore we have:

$$\sum_{C^M_3} p(c) = P^M_{i-1,j}$$

■

We are now in a position to present the main theorem of this section that guaranties a dynamic programming method for computation of the expected number of occurrences of motif a $M \in \mathcal{M}$ in sequence a T. i.e., e^T_M.

Theorem 14 *Suppose T is a sequence with length L, M is a motif from $\mathcal{M} = \{M_1, \cdots, M_n\}$, J_M is the maximum number of occurrences of M over T, $0 \leq i \leq L$ and $0 \leq j \leq J_M$, then*

$$P^M_{i,j} = P^M_{i-1,j} + P^M_{i-|M|,j-1}\frac{p(B|M)}{p(B|\bar{M})} + \sum_{M' \in \mathcal{M}, M' \neq M} P^M_{i-|M'|,j}\frac{p(B'|M')}{p(B'|\bar{M}')} \tag{3.5}$$

Proof. Before proving the theorem in general, we like to pay attention to some boundary conditions. We should recall that $W(\emptyset) = 1$ and consequently $\sum_{\{\emptyset\}} p(c) = \sum_{\{\emptyset\}} \frac{W(c)}{Z} = \frac{1}{Z}$. If for a motif $X \in \mathcal{M}$, $i \leq |X|$ then $P^X_{i,j} = \sum_{\emptyset} p(c) = 0$ and therefore the corresponding term would be cancelled out for the 3.5 and therefore we will not have any negative values for position indices. Similarly if $j = 0$ then the second term of the Equation 3.5 will be zero and hence Equation 3.5 is modified as:

$$P^M_{i,j} = P^M_{i-1,j} + \sum_{M' \in \mathcal{M}, M' \neq M} P^M_{i-|M'|,j} \frac{p(B'|M')}{p(B'|\bar{M}')}$$

Therefore without loss of generality we may assume that $i \geq \max\{|M||M \in \mathcal{M}\}$ and $j \geq 1$. Now according to Lemmas 11, 12, and 13 we can write:

$$\begin{aligned} P^M_{i,j} = \sum_{C^M_{i,j}} p(c) &= \sum_{C^M_1} p(c) + \sum_{C^M_2} p(c) + \sum_{C^M_3} p(c) \\ &= P^M_{i-|M|,j-1} \frac{p(B_i|M)}{p(B_i|\bar{M})} + \sum_{M' \in \mathcal{M}, M' \neq M} P^M_{i-|M'|,j} \frac{p(B'|M')}{p(B'|\bar{M}')} + P^M_{i-1,j} \end{aligned}$$

This finishes the proof. ■

Theorem 15 *With the above mentioned notations we have:*

$$W^M_{i,j} = W^M_{i-1,j} + W^M_{i-|M|,j-1} \frac{p(B|M)}{p(B|\bar{M})} + \sum_{M' \in \mathcal{M}, M' \neq M} W^M_{i-|M'|,j} \frac{p(B'|M')}{p(B'|\bar{M}')} \tag{3.6}$$

Proof. See Lemma 9 and Theorem 14 ■

3.2 Occupancy values of proteins binding a sequence (motif *o-values*)

In this section we shall explain how in our model the expected number of occurrences of a given motif in a given sequence is computed. However, as we promised in Section 3.1, we will repeat the key ideas of the RRS model in a less mathematical language with the hope of keeping the coherence of the story for those readers with less mathematical background who might have skipped the Section 3.1.

We assume a template sequence T, a test sequence S, and a set of transcription factor motifs $\mathcal{M} = \{M_1, \cdots, M_n\}$. We use the term configuration to denote a particular arrangement of protein molecules along the DNA sequence, which is defined by the intervals at which each molecule is bound to the sequence. Valid configurations are those in which binding intervals do not overlap. By assuming molecules are bound to sequence independently, we then associate a statistical weight $W(c)$ to any valid configuration c (see Equation 3.1) which is the product of contribution of each binding event. The contribution of any of these binding events are in turn a function of function of binding affinity and concentration parameter.

The probability of each configuration c is then defined as $p(c) = \frac{W(c)}{\sum_{c \in C} W(c)}$ where C is the set of all valid configurations. We use the same dynamic programming technique as in (62) to compute this probability.

There can be more than one expressed protein species that can bind to a given motif. In the absence of information on either the number of protein species capable of binding a motif or the nuclear concentrations of these proteins we assume the total nuclear concentration of such proteins to be equal for each motif and set $\frac{p(\bar{M}_i)}{p(M_i)}$ to a constant value. Where such information is available it can be integrated into the RRS model by setting the concentration parameters accordingly. When the concentration parameter is set to a constant value, it determines the average density of proteins bound to DNA within our model. We chose 15 as the setting for the concentration parameter and confirmed that results presented in this work are robust as long as the concentration parameter is set such that the protein density is realistic. Note that the scaling of this parameter depends on the scaling of the binding affinity and therefore the absolute value does not have a direct interpretation.

Intuitively, this probability distribution over all possible configurations should reflect a number of aspects of enhancer function in a natural way. Overlapping binding sites will compete with each other, high affinity binding sites will attract a binding molecule more often, and weak binding sites can exert an effect if they are present in numbers. Proteins are more likely to interact with the polymerase if they occupy the enhancer more often. Therefore, a key quantity relevant to the

function of an enhancer is the expected number of copies of a given protein that bind to motifs in the enhancer (T):

$$e^T_{M_i} = \sum_{c \in C} p(c) I_{M_i}(c) \tag{3.7}$$

in which $I_{M_i}(c)$ is the number of occurrences of motif M_i in configuration c. This definition is of particular interest because it captures both the specificity and multiplicity of a binding event of a protein to the sequence in the $p(c)$ and $I_{M_i}(c)$ terms respectively. A dynamic programming approach is used to compute each occupancy value. Finally the sequence T is associated with the vector of occupancy values, that is, $E^T =< e^T_{M_1}, \cdots, e^T_{M_n} >$ and similarly sequence S is associated with $E^S =< e^S_{M_1}, \cdots, e^S_{M_n} >$. Our results show that these occupancy values are length dependent. We divide them by the length of the sequences to normalise them. Therefore, each of these vectors summarises the combined specificity and multiplicity that each protein is likely to bind to each of the sequences.

3.3 Similarity scores

Our aim in this section is to define a similarity function over the space of vectors of occupancy values to extract the similarity of a given pair of *o-values*. Having observed *o-values* from the template sequence, E^T, we want to test if the vector of *o-values* from the test sequence, E^S, has been drawn from the same distribution or from a random background distribution. The logarithm of motif *o-values* in randomly picked sequences from the genome of the species of interest approximates a normal distribution (see 3.3).

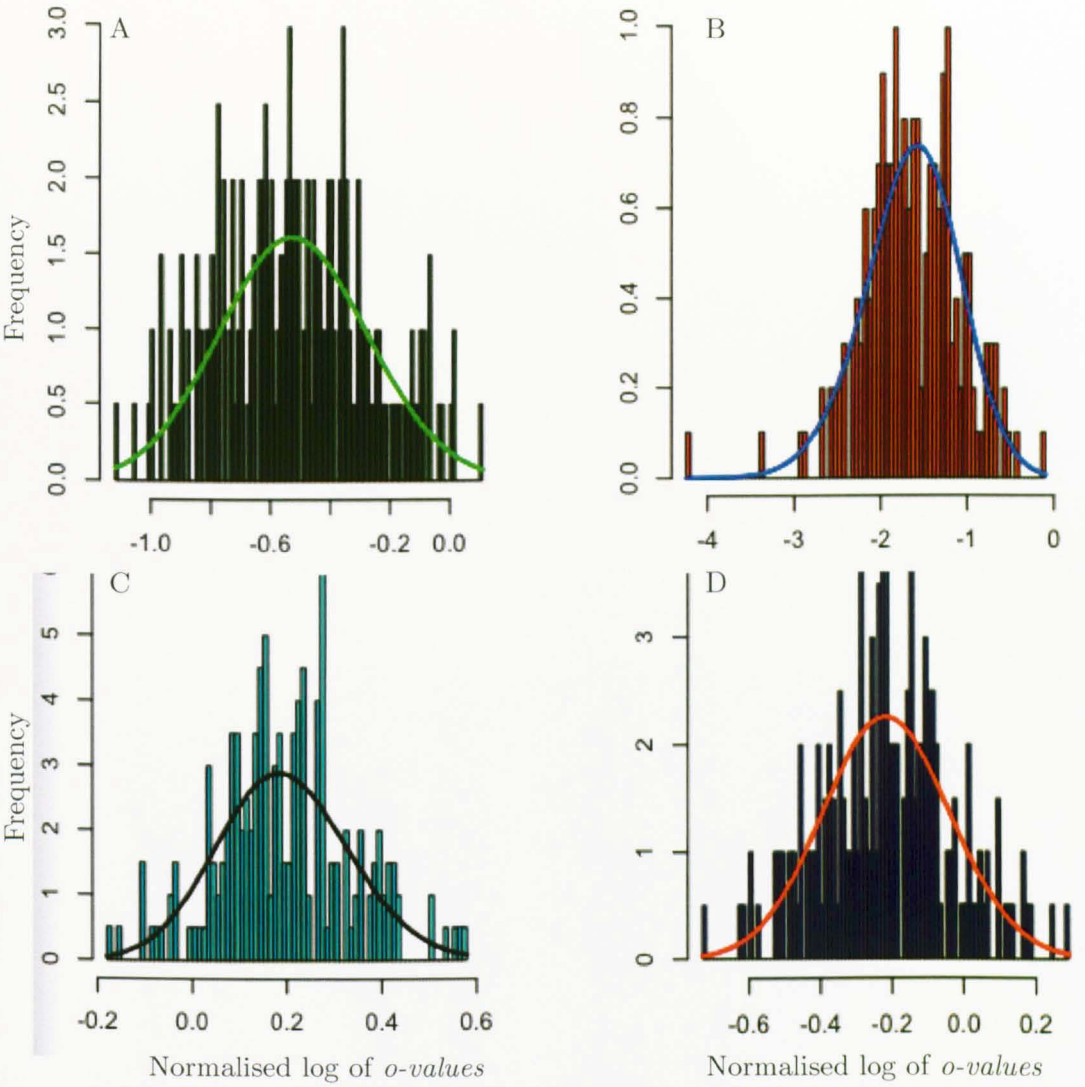

Figure 3.3: The normal distribution is a fairly good approximation for normalised and logged *o-values*. A: *o-values* for motif M00092 in 1000 randomly picked sequences of length 1000. B: motif M00488 with random sequences of length 300. C: motif M00093 with random sequences of length 3000. D: motif M00696 with random sequences of length 1700.

Therefore, the probability of a motif o-vector such as $E^S =< e^S_{M_1}, \cdots, e^S_{M_n} >$ can be obtained from a multivariate normal distribution. For the sake of simplicity, we shall consider an independent multivariate normal distribution. This

means that the probability of the *o-vector* E^S under the random model is $p(E^S|R) = \prod_{i=1}^{n} p(e^S_{M_i}|\mu = \mu_{R_i}, \sigma = \sigma_{R_i})$, where, μ_{R_i} and σ_{R_i} are the mean and standard deviation of *o-values* for motif i in randomly picked sequences. The probability that E^S has been drawn from the same distribution as the template is $p(E^S|T) = \prod_{i=1}^{n} p(e^S_{M_i}|\mu = e^T_{M_i}, \sigma = \sigma_{R_i})$. We define the RRS score as:

$$RRS(S|T) = \frac{p(E^S|T)}{p(E^S|R)} \tag{3.8}$$

The first point to note about this definition is that it is asymmetric but one may define it as an average to make it symmetric, i.e. $RRS(S,T) = (RRS(S|T) + RRS(T|S))/2$. However, it is sensible to work with the asymmetric version, in particular when comparing two sequences from different species.

The second point is that, in the current version we are using a single sequence as template. This limits our prior information about the distribution of the *o-values* in the template sequence. In other words, for each motif M_i we use only $\mu = e^T_{M_i}$ as the mean and $\sigma = \sigma_{R_i}$ as the standard deviation of the distribution. However, if we know that some enhancers are driving almost similar expression pattern, then it is better to consider these set of sequences as template and consequently feed more accurate mean and standard deviation of the distribution into the model.

The third point that makes this definition more realistic and useful is the contribution of the individual motifs:

$$f(e^S_{M_i}) := \frac{p(e^S_{M_i}|\mu = e^T_{M_i}, \sigma = \sigma_{R_i})}{p(e^S_{M_i}|\mu = \mu_{R_i}, \sigma = \sigma_{R_i})} \tag{3.9}$$

for any motif M_i, where $1 \leq i \leq n$. For any test sequence S, one can consider Equation 3.9 as a function of variable $e^S_{M_i}$ with three extra parameters: $e^T_{M_i}$, μ_{R_i}, and σ_{R_i}. The following cases illustrate this definition and its usage in the rest of this paper:

1. if $e^T_{M_i} \approx \mu_{R_i}$ (see Figure 3.4A), then $f(e^S_{M_i})$ can be considered as a constant function with value ≈ 1 (Figure 3.4D). This means that if the expected number of occurrences of this motif in the template sequence is very close to the average of its expected number of occurrences in the random sequences, then the overall RRS score for the test sequence will be largely

independent of number of occurrences of this motif in the test sequence. In biological terms, if the test sequence shares a regulatory code with the template sequence, but also contains additional binding sites, then these additional sites do not reduce the sequence similarity.

2. if $e^T_{M_i} > \mu_{R_i}$, then $f(e^S_{M_i})$ is an increasing function. More accurately, if we assume that $e^T_{M_i} > A > \mu_{R_i}$ where A is the intersection point of the two distribution curves (Figure 3.4), then $f(e^S_{M_i}) \leq 1$ if $e^S_{M_i} \leq A$ else it is greater than one. This case occurs when the motif is strongly present in the template sequence. Accordingly, the greater the motif *o-value* in the test sequence, the greater the contribution of the motif (Figure 3.4 parts B and E). Note that a strongly negative RRS score in this case implies poor presence of the motif in the test sequence.

3. Similarly, if $e^T_{M_i} < \mu_{R_i}$, then $f(e^S_{M_i})$ is a decreasing function. In other words, $f(e^S_{M_i}) > 1$, if $e^S_{M_i} < A$ (where $e^T_{M_i} < A < \mu_{R_i}$ is the intersection point of two curves) then the motif will be assigned a contribution greater than one, otherwise $f(e^S_{M_i})$ has a value less than one, contributing negatively to sequence similarity (Figure 3.4 parts C and F).

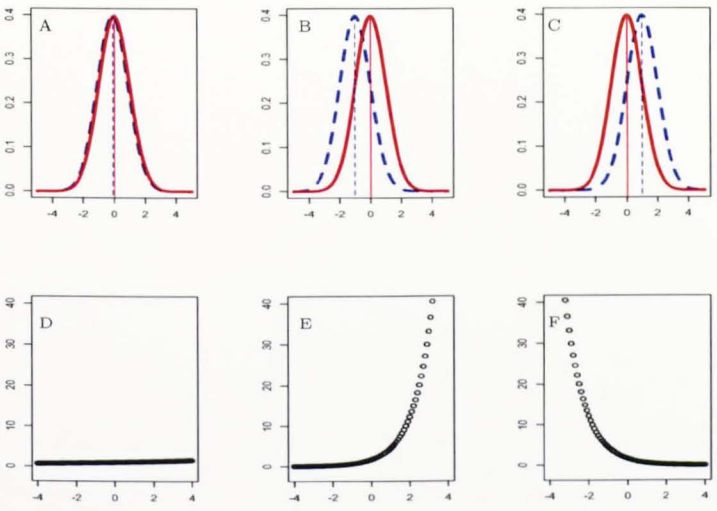

Figure 3.4: Illustration of the RRS similarity score for an individual motif. There are three possibilities. A,D: The motif is neither significantly present nor absent in the template sequence. The distribution of motif *o-values* in sequences with the same function as the template sequence (solid line) is estimated to be equal to the random background (dashed line). In this case irrespective of the motif *o-value* in the test sequence, the function $f(e^S_{M_i})$ (see Equation 3.9) is constant (D). B, E: The motif *o-value* in the template is higher than in random sequences (B), in this case $f(e^S_{M_i})$ is an increasing function (E). C, F: The motif *o-value* is lower than in random sequences, indicating significant absence. In this case $f(e^S_{M_i})$ is decreasing (F).

3.4 Parameter fitting

Given the sequence T and the motif $M \in \mathcal{M}$, the model requires tree parameters: binding probabilities of the motif at each position of the sequence, maximum number of occurrences of the motif over the sequence T and the concentration of the corresponding factors.

For calculations of binding binding probabilities in this model we used an implementation of the PWM (see Section 1.1.2) model called BiFa tool (unpublished tool developed by N. Dyer and J. Reid). We should recall that we do not use predetermined thresholds for binding probabilities, allowing both weak and

strong factor binding to contribute.

In the following two subsections we will try to clarify how the other two parameters can be fitted into the model.

3.4.1 Maximum number of occurrences of a motif in a sequence

The maximum number of occurrences of the motif M over the sequence T is theoretically defined as $J_{\max} = \frac{L}{|M|}$ where L is the length of the sequence T and $|M|$ is the length of the motif M. However, using these theoretically defined number of occurrences of each of the motifs might be computationally expensive and one may like to see how robust the results are with respect to fewer values for $J_{\max}$s. To clarify this, first we would like to recall that the number of configurations exponentially increase as a function of number of motifs. To illustrate this further, consider a simple example where we have a sequence with length 1000bp, a set of motifs each of which have a length equal to 10bp and also that factors can only bind in positions $1, 11, 21, \cdots, 991$, then even in this very simplified example the number of configurations is equal to 10^{100}.

In order to see how we can reduce this computational cost, we should remember that for a given motif M we have $Z = \sum_C W(c) = \sum_{j=0}^{J_{\max}} W_{L,j}^M$ (see Notation 6 and Lemma 8), where $W_{L,j}^M$ is the sum of statistical weights over all configurations with exactly j occurrences of M. However, the number of configuration with exactly j occurrences of M exponentially decreases when j increases. In our simplified example $C_{L,0}^M = 9^{100}$ where as $C_{L,100}^M = 9^0 = 1$. Consequently $W_{L,j}^M$ is an exponentially decreasing function of j, that means that for a big enough j, we may assume that $W_{L,j}^M \simeq W_{L,k}^M$ for any $k \gg j$. This is illustrated in Figure 3.5 on page 56 where the logarithm of statistical weight i.e., $\log W_{L,j}^M$ is plotted as a function of j i.e., different number of occurrences of the motif M for 10 different motifs. The sequence in this figure was of length 450pb and it was randomly picked from the *D. melanogaster* genome. The motifs illustrated in this figure are top 10 motifs in Table 4.2 on page 64. Therefore, it can be concluded that for computations of the statistical weight over all configurations i.e., Z, one may not require to take maximum number of occurrences of each motif M as $\frac{L}{|M|}$, instead

any number around 15 will provide him/her with an accurate approximation that will lead saving computational costs.

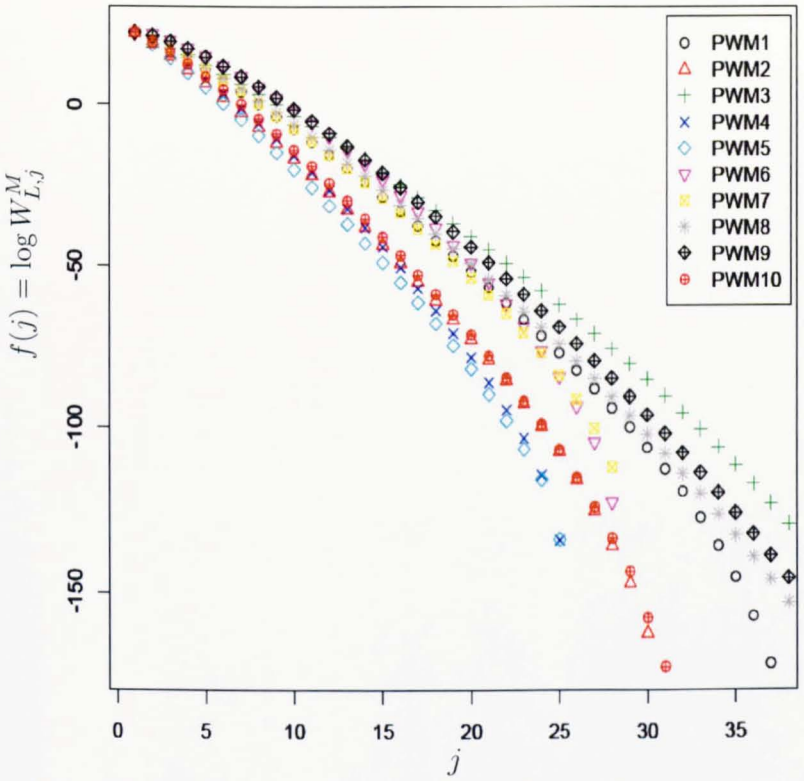

Figure 3.5: Illustrated here is the logarithm of sum of statistical weights, in other words, $W^M_{L,j}$ as a function of j which is the number of occurrences of the motif M. This is depicted for 10 different motifs. The sequence used in this analysis was a randomly picked sequence from the*D. melanogaster* genome with length 450pb.

3.4.2 Robustness of the concentration parameter

We note that there can be more than one expressed protein species that can bind to a given motif. In the presence of information on either the number of proteins species capable of binding to a motif or the concentration of the corresponding proteins, then Equation 3.1 on page 41 is re-written as:

$$W(c) = \prod_{i=1}^{N} \frac{p(M_i|B_i)}{p(\bar{M}_i|B_i)} \times \gamma(s_i, c_i) \tag{3.10}$$

where $\gamma(s_i, c_i)$ is the concentration parameter as a function of s_i which is the protein species that recognizes the motif M_i and c_i which is the corresponding protein concentration (in one point of AP axis). However, in the absence of such information we assume that the total nuclear concentration of such proteins to be equal for each motif and set $\frac{p(M_i)}{p(M_i)}$ to a constant value, that can be considered as the average density of proteins bound to DNA within our model. We also note that the scaling of the this parameter depends on the scaling of the binding affinity and, therefore, the absolute value does not have a direct interpretation. We chose 15 as the setting for concentration parameter and confirmed that (see Figures 3.6 and 3.7 on pages 58 and 59 respectively) the result presented in this project is robust as long as the concentration parameter is set such that the protein density realistic. Our observations show that this can range from 10 to 100. Intuitively, protein density close to zero is meaningless and extremely high protein density can mean the system reaches a saturated point, and also, we should note that proteins make only a fraction of the cell volume.

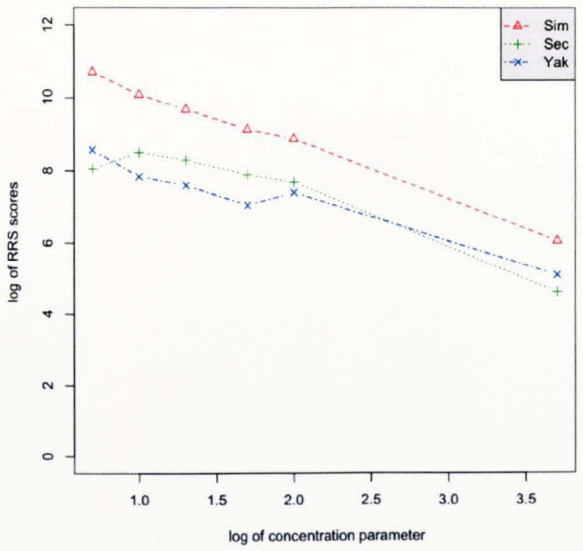

Figure 3.6: The RRS scores of a functional subregion of *D. melanogaster* vs its orthologous in *D. simulans*, *D. sechellia* and *D. yakuba* using 6 different concentration parameters which are $5, 10, 20, 50, 100$ and 5000 are illustrated. Note that the numbers in both x and y axes are log transformed. More information about these sequence can be found in Subsection 5.3.1. As we can see, the RRS is not considerably varying for any concentration from 10 to 50. It worth pointing out that, theoretically, the RRS scores for a concentration close to zero is not defined. The RRS scores for big concentration are statistically less significant as there are some random sequences obtaining higher scores, when the subsequence of *D. melanogaster* compared to 1000 randomly picked sequences from *D. simulans*.

In Figure 3.7 on page 59 we are illustrating the RRS scores of a subregion of *D. melanogaster* (BiFa-Only see Subsection 5.3.1 for more details about this sequence) vs its orthologs from *D. simulans* in 6 different concentrations (green vertical lines) and at each case the subsequence from *D. melanogaster* is compared with 1000 randomly picked sequences from *D. simulans* to show the statistical significance of the RRS scores at that concentration.

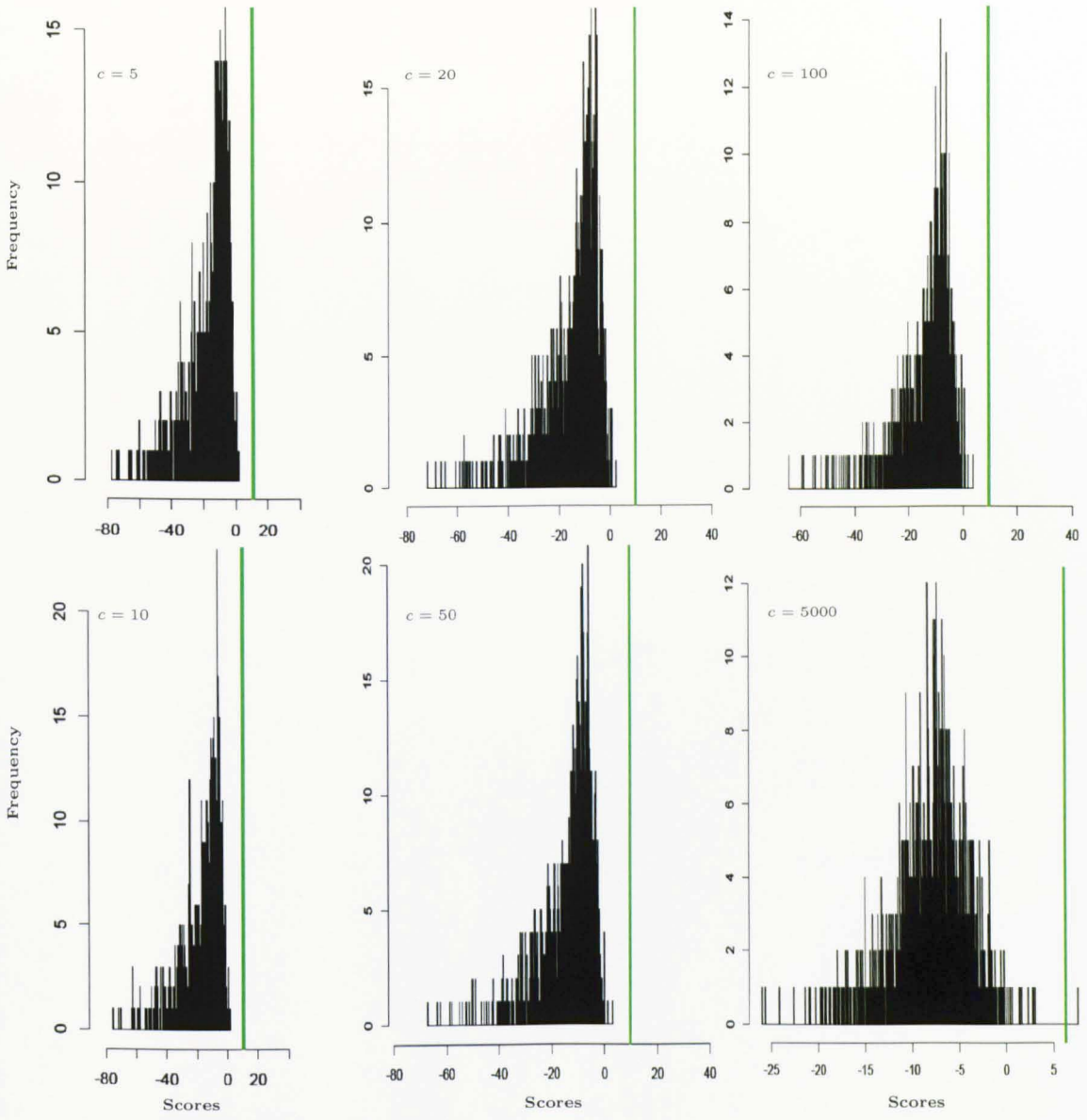

Figure 3.7: RRS scores of a functional subregion of *D. melanogaster* vs its orthologous sequence from *D. simulans* at concentrations $5, 10, 20, 50, 100$ and 5000. Green vertical lines show to these scores. The statistical significance of each score can be seen when compared to the scores of *D. melanogaster* vs 1000 randomly picked sequences form the *D. simulans* genome, computed with the corresponding concentration parameter.

3.5 Conclusion

We have presented an alignment-free method for detection of functional conservation of the regulatory sequences based only on occupancy level of some transcription factors of interest. It has been designed such that it is less data-dependent with a wider range of applications and more conclusive results. This model can be used for comparison of regulatory sequences where sequences are functionally related but are not orthologous (see Chapter 4). The RRS can also be used for comparison of regulatory sequences from different species where they have undergone a substantial evolutionary divergence (Chapter 5). Finally, we would like to close this chapter by listing some finer points and shortcomings of our model where further development may lead to a more accurate model.

- In the current version of the RRS we use a set of known TF motifs, focusing the sequence analysis on validated motifs. However, there may be yet unknown binding motifs relevant to the function of the sequences analysed. We could introduce some complementary sequence patterns into the analysis to test for a possible contribution to sequence similarity.

- There are further sources of prior knowledge that could be fed into the analysis in principle. For example, we are assuming equal concentrations of all regulators even though these will vary in different cell types. Some motifs belong to particular pathways which may be of particular interest in some cases. It would be possible to define a weight for such subsets of motifs.

- Within the current version, the synergy between pair of motifs is ignored, but there are some reports that regulation of some fly enhancers requires synergy between pairs of motifs (65).

- Rather than using a single template sequence, it would be possible to use multiple template sequences with similar expression pattern. This should help to define a more accurate distribution of motif occupancy levels.

- Given the key regulators of an enhancer and concentration of factors at different position of the AP axis, the RRS can be modified in a similar

way to the regression-based model (see 32 or Subsecection 2.1.3.1 for more information) to predict the expression profile of the enhancer. For this, one may employ the same regression function and use the expected number of occurrences of each motif (i.e., *o-value*) instead of the motifs score defined in regression-based model. This might help lessening the data dependency of the regression-based model, where for calculations of the motif scores one need a cross-species comparison. Furthermore, a direct comparison of this modified RRS with existing models that are predicting expression profiles, may help further improvements of any of these models and may provide more insights into the regulatory mechanism.

4

Functional Links Between Non-Alignable Enhancers

In this chapter we demonstrate how the RRS can be used to detect functional links between a set of enhancers that do not show any alignment conservations (non-overlapping enhancers from *D. melanogaster*). These type of applications might be of great importance in situations where a set of co-regulated genes in a single species is given and it is aimed for searching for some subregions that are likely to mediate similar expression profiles.

In what follows, after a brief introduction, we first give more details of the data sets that were used for this analysis. We then present the results at each corresponding subsection.

It is worth pointing out that a slightly modified version of this chapter has been published in (38).

4.1 Introduction

Our goal in this chapter is to evaluate if the RRS can distinguish functionally/evolutionarily related sequence pairs (positive sets) from the sequence pairs randomly picked from the genome (negative sets). For this, we apply it to the same fly data sets as used in (31) as is explained in Section 4.2.1. We first demonstrate that the distribution of alignment significance levels, or e-values in short, of positive sets is not significantly different from the distribution of alignment

e-values of negative sets. Using RRS however, there are 40 pairs of sequences (edges in Graph 4.3) whose scores are significantly greater than the scores obtained using random pairs. The statistical significance of some of these scores are highlighted. We show that according to the RRS results, a subset of these 40 enhancers are regulated by the regulator BCD (subgraph highlighted by rectangles in Figure 4.3). This finding is of particular significance as it has been experimentally confirmed by (52). Finally, we do some analysis firstly to show the contribution of strongly absent motifs to the similarity of a pair of sequences and secondly to highlight the substantial contribution of weak binding sites in our model scheme.

4.2 Discussion and results

4.2.1 Data sets

This study uses four data sets of experimentally confirmed fly enhancer sequences (same data sets as are used in (31)): 82 FLY_BLASTODERM, 23 FLY_PNS, 9 FLY_TRACHEAL and 17 FLY_EYE enhancers. For each of these positive sets we associate a corresponding negative set of sequences randomly picked from non-coding regions of the same genome. Thus each real enhancer had a randomly picked counterpart of the same length (Table 4.1). To establish the discriminatory capabilities of the RRS, scores were calculated for each possible pair of sequences in the positive sets and in the negative sets. A comparison of these two sets of results was done by sorting all scores and then looking at top $K = \frac{k(k-1)}{2}$ pairs, where k is the number of enhancers in that set. For the set of TF motifs, we used 67 insect-specific PSSMs available in the TRANSFAC database, (47). The full list of the motif-IDs is presented in Table 4.2.

	POSITIVE SET$_{(Real)}$	NEGATIVE SET$_{(Random)}$
BLASTODERM	82	82
EYES	17	17
PNS	23	23
TRACHEAL	9	9
GLOBAL	131	131

Table 4.1: Sequences used in this analysis

Table 4.2: List of motifs used throughout work.

Motif ID	Length	Gene	Consensus
M00009	8	ttk	GGTCCTGC
M00012	9	cf2	RTATATRTA
M00013	9	CF2	GTATATATA
M00016	17	E74	NNAYCCGGAAGTNNKN
M00018	19	Ubx	NNNNNNTTAATKGNNNNNN
M00019	16	Dfd	NNNNNNTTAMYNNNN
M00020	12	Ftz	ANWGCAATTAAG
M00021	10	Kr	AMYGGGTTAW
M00022	10	Hb	SMANAAAAAA
M00028	5	Hsf	AGAAN
M00043	11	Dl	GGGTTTTTCCN
M00044	14	Sn	ASCACCTGTTNNCA
M00060	13	Sn	NNRACAGGTGYAN
M00067	14	H(d)	NNGGCACGCGMCNN
M00090	14	Abd	NSNTTATGGCNNN
M00091	18	BR-C	WNRTAATARACAARWNWN
M00092	16	BR-C	NNBTNTNCTATTTNTT
M00093	15	BR-C	NANTAAACTARANNN
M00094	13	BR-C	WWWRTAAASAWAA
M00110	16	Elf	NNKWNYGGTTTTGWAN
M00111	9	Cf1	GGGGTCAYS
M00112	9	Cf1	GGGGTCACG
M00120	11	Dl	HGRGAAAANCV
M00140	8	Bcd	SGGARRAA
M00163	15	HSTF	AGAANAGAANAGAAN
M00164	15	HSTF	AGAANAGAANNTTCT

Continued on next page

Table 4.2 – continued from previous page

Motif ID	Length	Gene	Consensus
M00165	15	HSTF	AGANNTTCTAGAAN
M00166	15	HSTF	NTTCTAGAANAGAAN
M00171	16	Adf	CCGCYGCYGYNGCCGV
M00234	13	Su	ANYGTGGGAAMCN
M00259	21	STAT	NNNNNTTTCCSGGAAANNNNN
M00266	16	Croc	WANAATAAATATNNNN
M00270	13	GCM	NNACCCGCATNNN
M00283	16	Zeste	NNWNTTGAGTGNNNNN
M00362	11	TCF-A	CTTTGATCTT
M00455	10	dri	NNRATTAATN
M00461	15	Ovo	NNNWGTAACNGNNNN
M00487	11	mtTFA	KNCTTATCNNN
M00488	14	DREF	ASCTATCGATADNY
M00629	10	Eve	TNWSSYCTGC
M00662	7	SGF	TTRTKCA
M00666	9	Sry	CGCATCWCT
M00679	8	Tll	AAGTYWAR
M00696	7	En	YCAATTA
M00710	8	Zen	WCATTWAM
M00723	11	GAGA	ASWGAGMGNRA
M00923	21	Adf	VCGCYGCMGYCGCGTGMCNGCG
M00934	11	Zeste	NWNTTGAGTGN
M00951	8	Grainyhead	ACYGGTTT
M01083	10	Abd	NNAAATNNNN
M01084	12	Antp	AAWAAMMATWAN
M01086	15	BYN	ARAAWTCRCACTTWN
M01087	23	CEBPA	WNWWNTKTGBVATCAKYYNTNNN
M01088	12	Deaf	GYBMTTCGGNTG
M01089	12	Kr	NNAACCCTTNN
M01090	8	Mad	GMGACGVN
M01091	7	Prd	AAATTRY
M01092	16	TCF	RNNNATCAAARNNNNN
M01094	7	Abd	CATAAAA
M01095	8	Ap	NNNATTDT
M01096	7	brk	GCGCCAG
M01097	10	cad	NNNTTTNYGN
M01098	16	Cf1-a	BWKAATNAATTNAWAN
M01099	18	Kni	NNNNNAAANTGGRNNNNN

Continued on next page

Table 4.2 – continued from previous page

Motif ID	Length	Gene	Consensus
M01101	8	Ovo	TAACRGTW
M01102	7	Sd	CATTYCN
M01103	14	Twi	CATRTGTKNHGCNN

4.2.2 Statistical links between sequences

We first used a local sequence alignment tool from the NCBI (`http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi`; 'Blast 2 Sequences') as well as an implementation of the Smith-Waterman algorithm (the water tool from the EBI; `http://www.ebi.ac.uk/Tools/emboss/align/index.html`) to show that these sequences are not alignable. The best hit found over all of these sets for BLAST had an e-value of 1e-08 corresponding to a stretch of $23bp$ from a pair in the negative BLASTODERM set (4.3). Figure 4.1 shows the results for both algorithms in BLASTODERM positive and negative sets. Therefore by looking at only the alignment scores, one cannot say if a particular pair is likely to be from the positive set or negative set.

	Positive Set	Negative Set
Set Name	e-value (length of aligned subsequence)	e-value (length of aligned subsequence)
BLASTODERM	$7e-06$ (19)	$1e-08$ (23)
EYE	0.003 (13)	$1e-04$ (15)
PNS	$3e-04$ (20)	$5e-05$ (17)
TRACHEAL	0.022 (13)	0.003 (18)

Table 4.3: This table shows the alignment significance levels (e-values) of the best hit for each pair of sequences within the positive and negative sets.

The functional conservation of these sequences presents a very different picture. To examine this, we looked at the RRS scores for all pairs of sequences in any of both positive and negative sets. For instance, in BLASTODERM enhancers, 43 out of 50 top scores belong to pairs from the positive set. The best (log

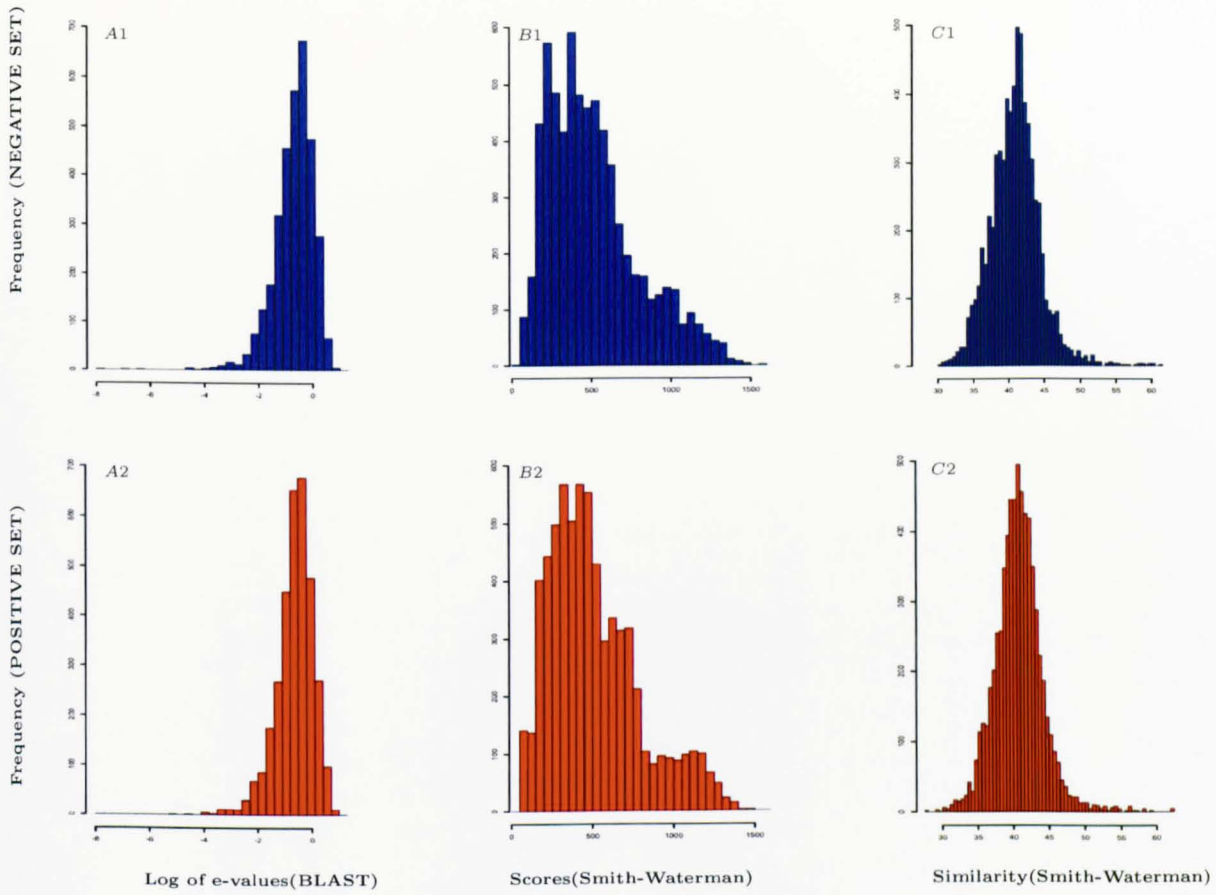

Figure 4.1: Distributions of alignment scores for BLAST and Smith-Waterman are not significantly different between positive and negative sequences. (A1 and A2) The log of e-values of BLAST applied to BLASTODERM positive set (red) and negative set (blue). (B1 and B2) Scores of Smith-Waterman algorithm applied to BLASTODERM positive (red) and negative (blue) sets. (C1 and C2) Same as B1/B2 but for sequence similarity instead of scores (as defined by water-tool).

of) RRS score was 9.64 corresponding to the comparison of *eve_stripe*1 (length 801*bp*) with *oc_otd-186* (length 187*bp*). To check the statistical significance of the RRS score, we compared *eve_stripe*1 with 1000 sequences randomly picked from the longest chromosome of the *D.melanogaster* genome, with length ranging from 100*bp* up to 3000*bp*. Interestingly, when comparing *eve_stripe*1 with these random sequences, no pairs gave an RRS score with log greater than 0. The result of this analysis is illustrated in Figure 4.2*A* in which the vertical dashed line is a reference line to show the position of the RRS score from *eve_stripe*1 vs *oc_otd-186* and the black histogram is the distribution of the RRS scores of *eve_stripe*1 vs 1000 randomly picked sequences.

We went on to consider what motifs contribute to the functional conservation that is seen. If the log of the score for a specific motif is greater than 1 (see Section 3.3), this indicates a significant similarity between the presence of the motif in the template and test sequence either by multiplicity or by specificity. An RRS score around zero is expected for a random DNA sequence and scores of less than -1 indicates a significant dissimilarity between the presence of the motif in the two sequences. RRS scores of all 67 insect motifs individually computed. Figure 4.2*B* depicts the distribution of these scores. As we can see, there are 3 factors that are assigned scores greater than 1. These factors are (in descending order): Bicoid (BCD), Krúppel (KR) and fushi tarazu (FTZ). This means that according to our model these three factors are main functional similarity-makers of this pair of enhancers. In comparison to the background sequences, all of these three factors are strongly presented in both of these sequences (see Section 4.2.4). This finding is of particular significance as it is supported by (52) where they show both computationally and experimentally that the regulation of the eve1 plus 10 other CRMs are strongly dependent to the regulator BCD. This suggests that the BCD is a regulator for *oc_otd-186*, too. We will come back to this point in more detail in Section 4.2.3.

4.2.3 Identification of enhancers with similar function

In order to make a more global analysis of these enhancers rather than analysing each individual set of enhancers we put all 131 enhancers into one set (referred to

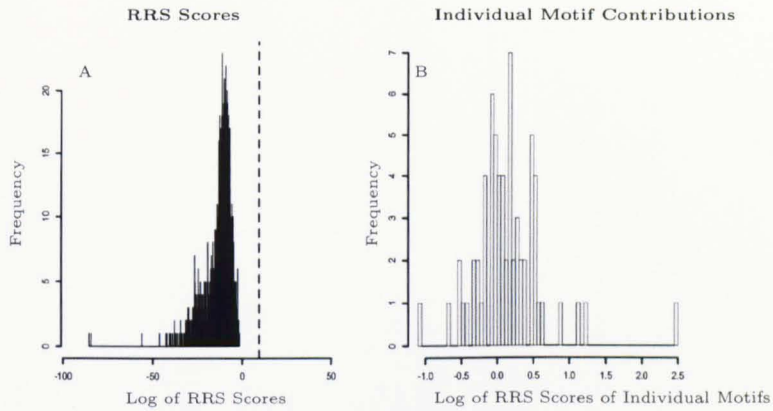

Figure 4.2: A: Illustrating the statistical significance of the RRS score of *eve_stripe*1 vs *oc_otd-186*. The dashed vertical line shows the log of RRS score from this pair which is 9.64. The black histogram shows the distribution of log of RRS scores of *eve_stripe*1 vs 1000 randomly picked sequences from *D.melanogaster* longest chromosome. B: Depiction of the contribution of individual motifs in the RRS scheme. Shown here, is the distribution of the individual motif scores in comparison of *eve_stripe*1 vs *oc_otd-186*. Three strongly positively contributed factors that are obtaining scores above 1, in descending order, are: BCD, KR and FTZ. The factor that is negatively contributing to this scheme i.e., obtaining a score less than -1 is SRY_β

as G_Positive set). Similarly all 131 randomly picked counterpart sequences were placed into another set called G_Negative set. The RRS scores were computed and a directed graph was generated in which each node is an enhancer from the G_Positive set and each edge represents a high RRS score for two corresponding nodes. The threshold for inclusion of edges was set above the maximum score within the G_Negative set (equal to 3). Therefore, only enhancer pairs that are scored above any pair from the G_Negative set are shown. The resulting graph (see Figure 4.3) shows the RRS prediction of the functional and/or evolutionary relationship of the enhancers associated to the top 43 scores from the G_Positive set. From this graph, we can see that only 34 enhancers (nodes) are associated to these 43 scores (edges). Thus some of the enhancers are paired together more often than would be expected by random chance alone. For instance HLHg* is paired with 6 other enhancers ($p < 1e-04$, p-value of binomial test for one node out of 131 to be part of 6 or more edges). The presence of a large number of high-scoring edges and the dense connectivity of the graph confirm that the RRS uncovers statistically significant structure in this data set.

We might want to think of the subgraph highlighted by rectangular nodes as a core subgraph because: firstly, all four of the nodes are from BLASTODERM enhancers, secondly it contains a pair that gets the highest score in BLASTODERM enhancers and thirdly it satisfies a transitivity property. Focusing more deeply on this subgraph reveals that, according to our analysis, the factor Bicoid (BCD) is the most strongly contributing factor in the functional similarity of any pair in this subgraph. This significant finding is experimentally supported by (52) where the regulation of the *eve_stripe*1, *eve_stripe*2 and *hstripe*0 and 8 more CRMs are reported to be strongly dependent on the activator BCD. They also showed that many of the BCD-dependent CRM contain a cluster of the gap protein Krüppel which is again in a high agreement with ours (see Table 4.4) in that in all of these five comparisons KR is either the second or third strongly contributed factor. We must recall that according to our model, a motif can obtain a high score either by its strong presence (because of multiplicity or specificity) or by strong absence in both sequences. It is also important to note that the five enhancers in this subgraph are regulated by a set of common factors (as colour-coded in Table 4.4), and this might be the reason that RRS can almost distinguish it as a subgraph.

Table 4.5 provides similar results for the subgraph with octagon shaped nodes distinguished by the RRS and a set of common motifs that we predict to regulate that subgraph.

Pair of Enhancers	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
eve_stripe1 vs oc_otd-186	I$BCD_01	I$KR_01	I$FTZ_01	I$HSF_03	I$HSF_04
eve_stripe2 vs oc_otd-186	I$BCD_01	I$KR_01	I$FTZ_01	I$HSF_03	I$MAD_Q6
hstripe0 vs oc_otd-186	I$BCD_01	I$KR_01	I$MAD_Q6	I$HAIRY_01	I$HSF_04
hstripe0 vs eve_stripe1	I$BCD_01	I$STAT_01	I$KR_01	I$GCM_01	I$EVE_Q6
eve_stripe2 vs eve_stripe1	I$BCD_01	I$FTZ_01	I$KR_01	I$GCM_01	I$TTK69_01

Table 4.4: The top five factors that are strongly contributing to the functional similarities of each pair in the subgraph highlighted by rectangles in Figure 4.3.

Pair of Enhancers	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
hstripe5 vs tllk10	I$ABDB_01	I$DL_01	I$KR_Q6	I$ADF1_Q6	I$SN_01
AbdBIAB vs tllk10	I$ABDB_01	I$KR_Q6	I$DL_01	I$SN_01	I$FTZ_01
dppdlmel vs tllk10	I$DL_01	I$SN_01	I$KR_Q6	I$ADF1_Q6	I$FTZ_01
clusterat55 vs tllk10	I$KR_01	I$ABDB_01	I$SN_02	I$FTZ_01	I$BRCZ3_01

Table 4.5: The top five factors that are strongly contributing to the similarity of the enhancers in the subgraph highlighted by octagon shape (see 4.3). Factors are ordered by their contribution. Colour-coding represents factor identity.

Overall, these findings reveal that our model indeed captures some of the core principles governing functional conservation of modules and hence performs much better than random expectation.

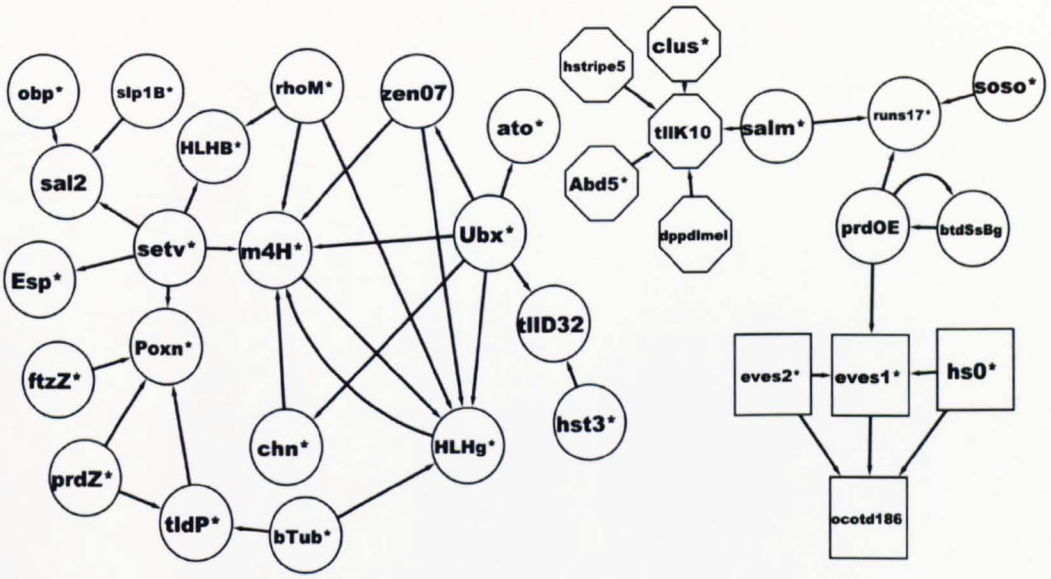

Figure 4.3: This graph represents the functional relationship of some of the top scored enhancers from the G_Positive set. Each node represents an enhancer and $Enhancer_1 \rightarrow Enhancer_2$ means that the $\log(RRS(Enhancer_1, Enhancer_2)) \geq 3$. The threshold 3 is to filter out other scores that are less than a score from the G_Negative set. Asterisks indicate abbreviated names. Full names of these enhancers are provided in Table 4.6 on page 73.

Enhancer Abbreviated Name	Enhancer Full Name
Slam	salm_salm _TSE_TRACHEAL
clus	cluster_at_55C_CE8016
rhom	rho_MLE-long_TRACHEAL
HLHB	HLHmbeta_enhancer
obp	Obp56a_prom
Sal2	salm_sal242S_PNS
m4H	m4_HZm4
Esp	EsPIPNC
HLHg	HLHmgamma_HZmgammaKX
bTub	betaTub60D_beta3-14/vm1
tldP	toldPromoterfusionright
prdZ	prdzebraenhancer
ftzZ	ftz_zebra_element
Ubx	Ubx_abx17
hs3	h_stripe3_ET38
chn	chn_SOP
soso	so_so10_EYE
serv	Ser_IV-3.0_EYE
ato	ato_RE
Poxn	Poxn_9
runs17	run_stripe17
Abd5	Abd-B_IAB5
Slp1	slp1_slp_B
eves2	eve_stripe12
eves1	eve_stripe1
hs0	h_stripe0

Table 4.6: Abbreviated names and full names for enhancers highlighted by star sign in 4.3 on page 72.

4.2.4 Contributions of motif absence and weak binding sites

We are interested in whether the strong absence of a motif in a pair of sequences can underly the statistically significant similarities we observed. We looked for motifs that are associated with a relatively high RRS score but whose associated *o-values* are lower than the *o-values* of the motif in random sequences. In Section 3.2 and Figure 3.4 we considered two situations where a motif is assigned a high RRS score because the motif is strongly present or it is strongly missing in both sequences. The strong presence may be more intuitive and it is illustrated in Figure 4.4 (parts $A1$ and $A2$) where we can see both RRS scores for any of the 67 used motifs in the comparison of the *eve_stripe*1 and *oc_otd-186* ($A1$) and also the normalised vectors of *o-values* for *eve_stripe*1 in red and *oc_otd-186* in blue ($A2$). The yellow base line is to show the *o-values* from the background (random sequences). Motifs 24, 8 and 7 associated with the top three RRS scores (in order) in *eve_stripe*1 vs *oc_otd-186* comparison. The reader can see from $A2$ that for all of these three motifs, the motif *o-values* are considerably higher than the background. This is called strong presence of motifs in both sequences. However, the interesting part is shown in parts $B1$ and $B2$ of Figure 4.4 where first we can see again in $B1$ the contribution of the individual motifs to the RRS scores of Ubxabx17EYE vs tllD32 and in $B2$ the *o-values* from Ubxabx17EYE in red, tllD32 in blue and motifs that are obtaining the top three RRS scores. We see that all three motifs are associated with *o-values* lower than the background (strong absence of motifs) but these contribute to the RRS score and, therefore, to the recognition of functional conservation.

The contribution of weak binding sites to the RRS scores can be seen in Figure 4.2B,C. The log of RRS score for *eve_stripe*1 vs *oc_otd-186* is 9.64. This is the sum of scores of each motif. The four motifs making the strongest contribution only contribute about half of this score (Figure 4.2C) while any RRS score above 0 is still significantly different from noise as none of the random sequences evaluated in Figure 4.2B had a score above 0. Therefore, the similarity of these two enhancers cannot be solely attributed to strong binding sites, but is influenced significantly by contribution of other motif even weak binding sites. This is consistent with

previous findings in (62), the authors hypothesis the effect of weak binding sites in functional similarity of two sequences.

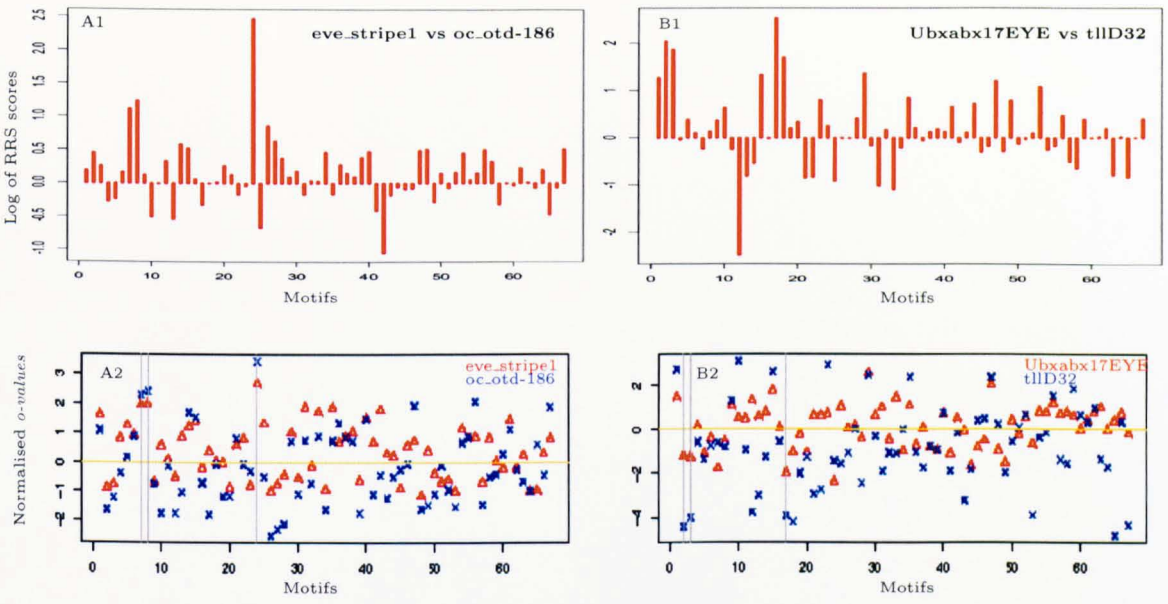

Figure 4.4: A1 shows the log of RRS scores for each of the 67 insect motifs that were used for the comparison of *eve_stripe*1 vs *oc_otd-186*. Motifs 24, 8 and 7 (in descending order) are the three top contributors to this comparison. A2 illustrates the *o-values* of these motifs from *eve_stripe*1 (red) and from *oc_otd-186* in blue. The y-axis is the number of standard deviations that an o-value deviates from the mean. The yellow base line shows the background *o-values*. The vertical lines highlight the positions of the top three motifs by RRS score. The main feature of A1 and A2 is that motifs with high RRS scores (A1) have *o-values* considerably higher than background level (A2), indicating strong presence of the motifs. B1 and B2 show an example where strong absence of motifs contributes to the statistical link between the sequences. B1 shows the individual contributions of each of the motifs in the comparison of $Ubxabx17EYE$ with $tllD32$. In B2 the *o-values* of the motifs from $Ubxabx17EYE$ are shown in red and those from $tllD32$ in blue. The three motifs that contribute strongly to the RRS scores (motifs 17, 2 and 3 in descending order) all have *o-values* less than background. This is referred to as a strongly absent motif.

4.2.5 Comparison of performance of RRS against some of the existing models

In this subsection we will present a comparison of RRS performance versus top three best performing models that were benchmarked by Kantorovitz et al. (31). The best performing model in that benchmark was the D2z model that we reviewed it in Chapter 2, however we have not reviewed the other two models i.e., p.a.5.3 and ed.6. For more details about these models the reader is referred to (31). We would also like to draw the reader's attention to the point that a direct comparison of our model with data intensive models is not possible as they are not defined as sequence comparison tools, but they try to predict qualitative gene expression patterns from the regulatory modules.

In order to assess the performance of our model versus these three models, we took the same approach as to Kantorovitz's in (31). In other words each pair of sequences in BLASTODERM positive set was compared by any of these four models, and so was each pair in BLASTODERM negative set. That is 3321 comparisons in each of the sets from each of the models. It was then assessed if the sequences in the positive set score higher than sequences in the negative set. This was done by sorting scores from all pairs, whether they were from the positive set or the negative set. Then we look at the top 300 scores and counted the number of scores from the positive set as correct predictions for any of the models. This analysis was repeated for the three other sets described in Subsection 4.2.1, i.e., EYES, PNS and TRACHEAL and we obtained almost the same results as to BLASTODERM that has been described here. Figure 4.5 on 77 shows the results of this comparison.

From this analysis, one may draw the conclusion that our model is not outperforming the D2z model. It is counted as the second best performing, although competing with best performing model. Regarding to this conclusion, it can be argued that in D2z model, the similarity of a pair of sequences is based on the distribution of all possible 6−mers 4096 words, whereas in our model the similarity of a pair of sequences is based on the distribution of only 69 meaningful motifs. Therefore, this comparison is not a fair comparison. we should acknowledge the idea from the examiners of this thesis for a more meaningful comparison

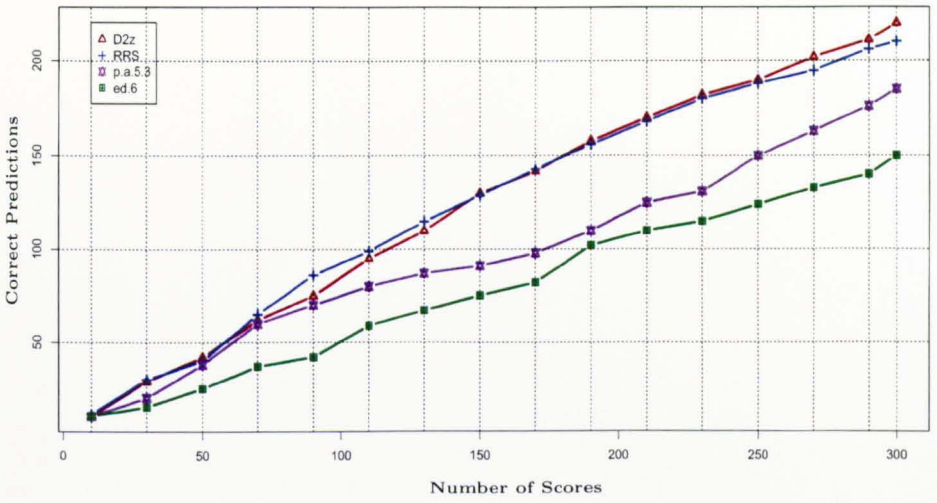

Figure 4.5: Performance of similarity scores form RRS, D2z, p.a.5.3 and ed.6 applied on Fly BLASTODERM.

as a future direction. That is, by computing the number of correct predictions of RRS using different random subsets of our motif set, one may compute how the number of correct predictions varies as a function of number of motifs. Performing this analysis with a big enough number of random subsets will assure that the variation in number of correct predictions is not a consequence of some strongly contributing motifs in a given subset.

4.3 Conclusion

We have demonstrated that our model can be used for comparison of regulatory sequences where sequences are functionally related but are not orthologous. For statistical validation of the RRS scores, the sequences that obtained top scores were compared with 1000 randomly picked sequences and showed that it is highly unlikely to get such high RRS scores just by chance. We have shown that the RRS can significantly detect the functional and/or evolutionary similarities of the regulatory sequences. In particular, the RRS can categorise some enhancers that are regulated by a set of common factors, a result that was in strong agreement

with experimentally validated reports. Based on the predictions of our model, we have proposed the hypothesis that the strong absence of a motif in a pair of sequences might be a feature for functional conservation.

In this analysis we used a set of high quality fly motifs that were available at the time of the analysis. However, as a future direction one may conduct similar analysis with a bigger set of motifs, for instance, vertebrate motifs. In addition, there might be unknown binding motifs relevant to the function of the sequences. Therefore introducing some complementary sequence patterns into the analysis to test for possible contribution to sequence similarity can be another option for further development of the model.

5

Prediction of Functional Regions of a Fly Enhancer

It is widely believed that the targeting specificity of the projection neurons (PNs) in the fly olfactory system is controlled by a transcriptional code. However, the underlying mechanism is not well understood and according to our current knowledge only a few of the key regulators of this mechanism have been identified (37; 70 and 35). On the other hand, it is well-known that the structure of the antennal lobe (AL) is highly conserved across Drosophila species (53 and 16). Therefore, one may hypothesize that an enhancer region that drives an expression pattern in a subset of PNs in *D. melanogaster* is likely to have a similar function in other Drosophila species.

In this chapter, we will present our *in silico* predictions of functional subregions of a fly enhancer region. According to these predictions, our collaborators at Stanford University (Maria Spletter and Liqun Luo) identified putative boundaries of the subregions. Then to test these predictions and dissect enhancer function, they generated some deletion constructs within the enhancer region.

Throughout our analysis, three approaches were tried: an alignment-based method, a motif-based method and our recently developed alignment-free method. The alignment-based method identified a region that was well conserved between some of the Drosophila species. The motif-based method revealed four regions with a high density of motifs. Some initial experiments based on these identified regions from the alignment-based and motif-based approaches raised the

requirement of predictions of shorter subregions. The final detailed predictions of these subregions were made by the RRS, our innovative alignment-free model. We show that the RRS can detect the orthology between 10 Drosophila species. Three of these orthologous sequences were assigned statistically very significant RRS scores. The top eight predicted key regulators of these three orthologous sequences are presented. We also demonstrate how one of these orthologous sequences is used to predict functional subregions within the *D. melanogaster* enhancer.

It is also shown that our model can construct the phylogenetic tree of 10 Drosophila species with a high level of accuracy from only the orthologous regulatory sequences of these species and distributions of 67 input PWMs.

5.1 Introduction

It is widely accepted that the precise connectivity of neural circuits (in the olfactory system) is mainly regulated by transcription factors that determine the particular set of guidance factors a neuron expresses (51 and 56). However, very little is known about the underlying transcriptional regulation and the identity of the main regulators (transcription factors).

We are aiming to provide new insights into this poorly understood area by predicting functional subregions of a *D. melanogaster* enhancer region that are likely to drive expression in subsets of PNs.

For this, we will make use of our understanding of the mechanism of the very well-studied fly olfactory system. In the fly olfactory system (see Figure 5.1), about 1300 olfactory receptor neurons (ORNs) are converged into about 50 glomeruli ($\sim$ 30 : 1). Those ORNs that are expressing the same olfactory receptor (OR) are converged into a single glomerulus. These 50 glomeruli are diverged into about 150 PNs ($\sim$ 1 : 3). In other words, each PN belongs to one of 50 unique groups based on which glomerulus they are connecting to. These cell types (groups) are determined by genes that they express. Expression of the corresponding genes, in turn, is regulated by many factors that bind a regulatory sequence (usually) upstream of the transcription start site. The regulatory sequence upstream of one of these genes (i.e. the *oaz* gene) is called GH146-Gal4

enhancer region. GH146-Gal4 is a P-element insertion 290bp off the transcription start site of the *oaz* gene located on chromosome 2R of the *D. melanogaster* genome (see Figure 5.4 on page 89). GH146-Gal4 labels around 90 of the 150 PNs in the AL and displays a relatively stable expression pattern in three lineages of PNs called anterodorsal, lateral and ventral (further details about GH146-Gal4 can be found in (29) and (71)). The enhancer region of our interest that is known to drive expression in these 90 PNs is just upstream of the GH146-Gal4 insertion point and therefore we call it GH146 enhancer region or GH146 enhancer in short. Therefore the fly AL is a well-studied system in which the identification of enhancer regions that are driving expression patterns in PNs can be assayed from the expression of subsets of PNs.

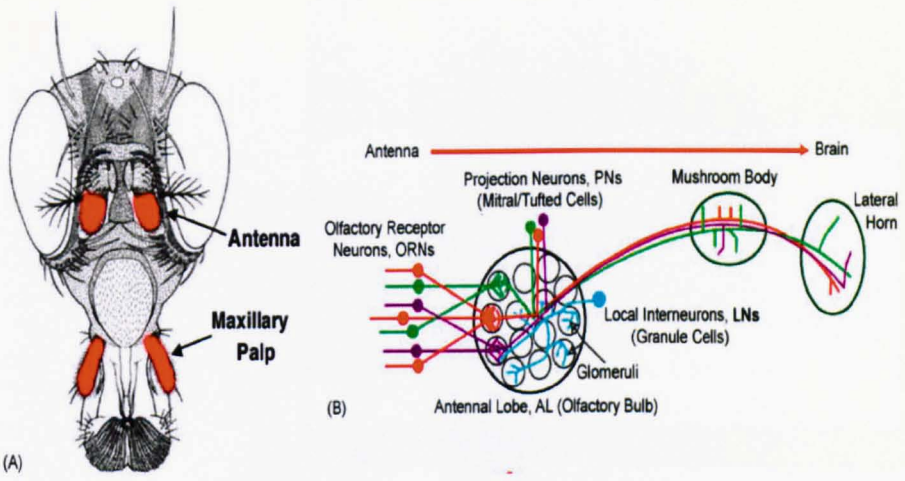

Figure 5.1: Illustrated here is a simplified schematic of the fly olfactory organs and its mechanism. (A) Olfactory organs are depicted in red. The upper structure contains about 1200 receptor neurons, while the maxillary palp (the bottom structure) contains about 120. (B) A simplified schematic cartoon of the olfactory mechanism showing how ORN are converged into a glomerulus and how glomeruli are diverged to higher brain centres such as mushroom bodies and lateral horns. Illustration has been taken from (27).

One way of gaining genetic access to different classes of PNs is assembling a collection of Gal4 enhancer trap lines that label a subset of PNs. But the small soma size (cell body) of PNs and limited amount of tissue precludes biochemical

methods, such as chromatin immunoprecipitation and its many variations, making it biologically difficult to use PNs as a model to investigate enhancer elements (see 71). Therefore, Bioinformatics becomes an alternative potential approach to the problem.

This project was a close collaboration with our experimental collaborators and is still ongoing. We hope that these predictions in combination with some additional experiments will provide new insights into rules of enhancer function in PNs and identification of some of the key regulators. One may consider this chapter as a bioinformatical counterpart of Chapter 4 in (71).

This chapter is mainly devoted to our *in silico* predictions of functional subregions of GH146-Gal4 enhancer region that were used by biologists to identify putative boundaries of enhancer regions. Some deletion constructs in the GH146 enhancer region were made to test our predictions. The experimentally evaluation of these predictions is still ongoing and not fully completed.

In the following, we will refer to conserved regulatory sequences (detected by the alignment-based ReMo algorithm, see Subsection 5.2.1) as ReMos and regions detected by the motif-based tool (binding factor analysis tool, see Subsection 5.2.2) as BiFa regions. A pair of sequences that obtains a statistically significant RRS score, will be called functionally conserved. The reader may note that in this context, 'regions', 'subregions', 'subsequences' and 'intervals' are considered synonyms.

5.2 Methods

This study was mainly based on regulatory elements of the GH146 enhancer. This sequence is about $4kb$ long, located upstream of *oaz* and is believed to contain most of the regulatory elements (see 71).

Identification of the homologous regulatory sequences between Drosophila species was done based on the fact that the *oaz* gene is present in all 12 Drosophilidae species (BLAST on http://flybase.org/blast/ was used). However, due to high level of repetitive elements upstream of the *oaz* gene in *D. persimilis* and *D. wilistoni*, these two species were excluded from the analysis. Figure 5.2 on page 83 shows the evolutionary relationship of the fly species.

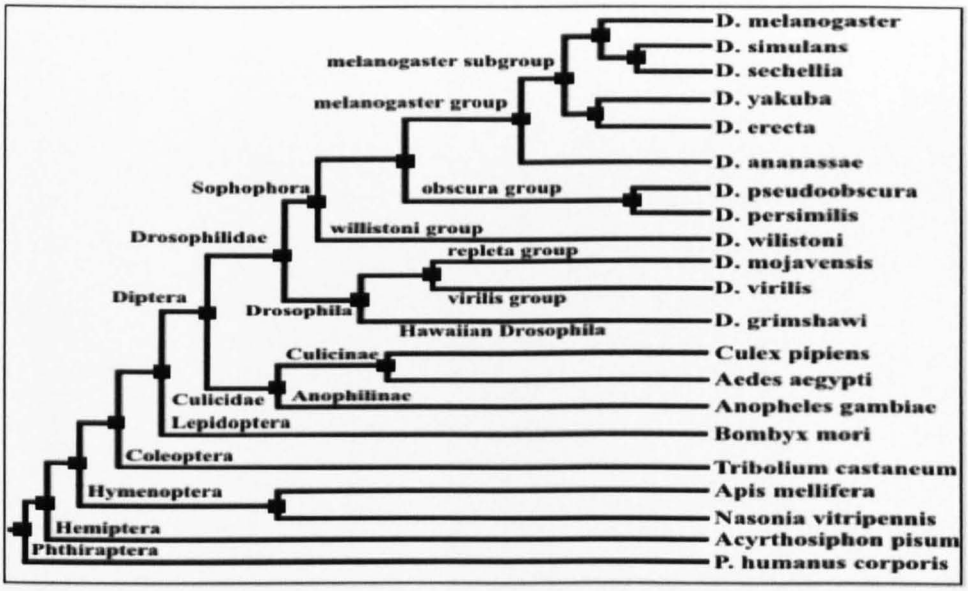

Figure 5.2: Phylogenetic tree of the fly species. The divergence period for the Drosophilidae species is estimated about 50 million years.

As the first series of experiments, three constructs were made by our collaborators and the expression pattern driven from the corresponding sequences were evaluated (more details about these analysis can be found in 71). These constructs were called GH146-Full, ReMo-Only and BiFa-Only. The reasons of making (and also naming of) these constructs are provided in the following.

The first construct i.e., the GH146-Full construct included the whole $4kb$ GH146 enhancer (see the bottom blue rectangle in Figure 5.4 on page 89). The

experiment based on this construct confirmed that this sequence contains most of the key binding sites (the experiment showed the full expression pattern in PNs see Figure 5.5:D1 on page 89).

Thus our task was to see if we can predict some functional subregions of the *D. melanogaster* GH146 enhancer that are likely to drive the same expression pattern as the whole GH146 enhancer. For this, three methods were applied: an alignment-based algorithm called the ReMo algorithm, a motif-scanning based model called the BiFa tool and the RRS, our newly developed alignment-free method. The two former models are unpublished methods developed by Sascha Ott and John Reid.

In the following subsection we provide the reader with a brief outline of the ReMo algorithm. For a more comprehensive description of this algorithm, the interested readers are referred to Appendix A where he/she can see that this algorithm is more sensitive than its (publicly available) counterparts, in particular, for detection of short conserved stretches of sequences.

5.2.1 Outline of the ReMo algorithm

We should recall that within this project, a pair of genomic sequences is called alignment-conserved or sequence-conserved if their optimal alignment has a statistically significant score and the sequences are not repeats. The algorithm employed to comprehensively detect alignment conserved non-coding regions at the *oaz* locus as potential conserved regulatory modules essentially computes an optimal alignment for every pair of $100bp$-fragments, comparing *D. melanogaster* to each of the other Drosophila species. For instance, when comparing two sequences of $100kb$ the algorithm compares in the order of 10^{10} pairs of 100-mers. The statistical evaluation of sequence alignment scores is greatly simplified by this approach as all aligned sequences have the same length.

The analysis based on the ReMo algorithm identified four well-conserved regulatory modules that were called ReMos A, B, C and D positioned $3713 - 4637bp$ upstream of the *oaz* gene (ReMo-C was significantly conserved between all 10 species). Therefore, the second construct called ReMo-Only was made by our

collaborators that included ReMos A, B, C and D (ReMos are shown as golden rectangles in Figure 5.4, the ReMo-Only construct is seen as a blue rectangle).

5.2.2 Outline of the BiFa tool

The BiFa tool is an implementation of a PWM model (see Subsection 1.1.2), in which a 0-order Markov model was used to evaluate matches found in GH146 *D. melanogaster* against the background. The likelihood of binding in each species was evaluated individually and the geometric mean to aggregate likelihoods across species was used. Briefly, sequences were scanned using fly and vertebrate PWMs extracted from the TRANSFAC database. A PWM of length L induces a distribution over L-mers that models binding sites for the transcription factor(s) it represents. Figure 5.3 on page 86 shows and example output from the BiFa tool.

The BiFa tool analysis revealed four regions with high density of factors in the *D. melanogaster* GH146 enhancer region. These regions were called BiFaA, BiFaB, BiFaC and BiFaD (yellow rectangles in Figure 5.4 on page 89). BiFaA overlapped with the ReMo-Only region. A subsequence positioned $1300 - 3300bp$ upstream of the *oaz* gene in *D. melanogaster*, consisting mainly of the other three regions (i.e., BiFaB, BiFaC and BiFaD) was called BiFaOnlyDmel. For the sake of shortness we will call it BiFaDmel in this text (see purple rectangle in Figure 5.11 on page 101). The homologs of this sequence in other species named similarly, for example, in *D. simulans* the homologous sequence is called BiFaDsim. Therefore, the third construct aimed to test the significance of BiFaDmel and was called BiFa-Only construct (see blue rectangle in Figure 5.4 on page 89)

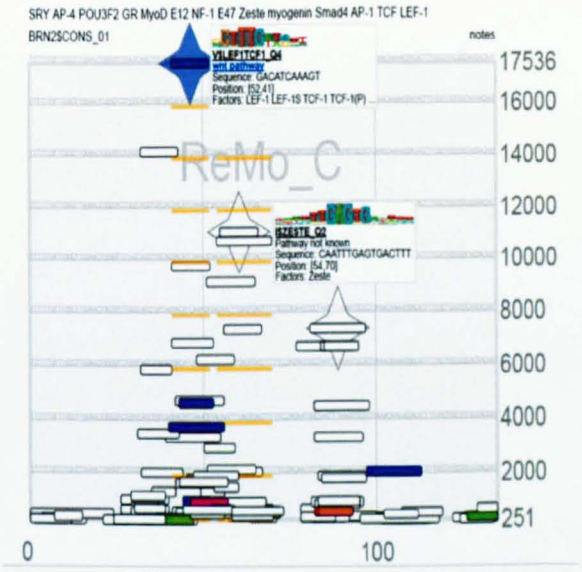

Figure 5.3: An example of the BiFa tool output that depicts the distribution of the fly and vertebrate motifs exported from the TRANSFAC that are obtaining a score above a threshold in ReMo C. In this figure the x-axis is the nucleotides from $5'$ to $3'$. The rectangles in the figure are depicting motifs that are scoring above the threshold. Y-axis shows the significance of the occurrence of the motifs. For instance, the LEF1TCF1 factor with y-component 17536 means that in a sequence with length 17536 one may expect to see one occurrence of this motif. If a set of motifs are known (to the BiFa tool) to belong to the same signaling pathway, they are colour-coded. In this figure factors corresponding to the top two scored motifs are illustrated. Diamonds are to highlight motifs that are conserved in the same order in the other species.

5.3 Discussion and results

A set of initial experiments based on the three above mentioned constructs revealed that GH146-Full drives a full expression pattern in PNs (Figure 5.5:D1 on page 89). They also found that expression in PNs is completely lost in the ReMo-Only construct (see Figure 5.5:E1 on page 89), whereas the BiFa-Only construct drives almost the same expression pattern in PNs as the GH146-Full (see Figure 5.5:F1 on page 89).

The main conclusion of these experiments was that the main regulatory elements are distributed in the BiFa-Only region. It is also possible that the deletion of such a large region of sequence may disrupt higher order interactions that may result in not having expression from the ReMo-Only construct.

For the next step, our task was to further narrow down the functional part of the GH146 enhancer by predicting shorter subsequences of the BiFa-Only region that were likely to drive an expression pattern similar to the expression pattern of the BiFa-Only construct. However, it was too hard to make predictions of these functional regions with either BiFa tool or ReMo algorithm any longer. The main problem with BiFa tool were:

- Although the BiFa tool is very useful for some analyses, (for instance, one may find some motifs strongly distributed over some subregions,) judgement about significance of motif-rich subregions is dependent on the user. In other words, there is no mathematical or computational way to provide the user with boundaries of the motif-rich subregions with their associated significance. This is becoming a more serious obstacle when the user needs to judge about multiplicity vs specificity of the motifs or vice versa. Besides, in the BiFa tool the motif scores are computed individually and independent of the other motifs, whereas we needed a tool to provide us with a score associated to any input sequence that reflects its potential activation level based on the set of motifs.

- The contribution of weak binding sites is ignored whereas we have some recent evidence showing the strong contribution of weak binding sites in expression of an enhancer (38 and 62).

- Many of the motifs occurring in a given genomic sequence are overlapping and the BiFa tool has not been implemented such as to be able to consider competitions of different factors for these overlapping motifs. Figure 5.3 on page 86 illustrates an example output of the BiFa tool.

And with respect to the ReMo algorithm:

- The ReMo algorithm did not detect any significant conservation between BiFa-Only regions of these 10 species.

- Recently, we have had some reports that the regulatory regions can retain function over large evolutionary distances, even though the DNA sequences are divergent and difficult to align. Therefore, if an alignment-based method such as ReMo algorithm does not detect any conservation, it does not necessarily mean lack of functional conservation.

To overcome these limitations of alignment-based and motif-scaning-base algorithms we developed and applied the RRS our alignment-free model.

The experimental results based on predictions of the RRS will be of great significance. In essence, agreements of the RRS predictions with the experimental results will mean that our model understands the regulatory code governing the fly olfactory wiring specificity, whereas the failures of our model will be as instructive as it successes. They will suggest that some input factors and some higher interaction rules are not captured, but also that the model does not artificially compensate for these missing features.

For the sake of completeness, in what follows, we first discuss the data sets that used for the RRS analysis for detection of functional subregions of the BiFaDmel. This is followed by presenting analysis of orthology detection of *D. melanogaster* in some other drosophilas. The orthologous sequence was then used to obtain statistical links between species in order to predict the optimal position of the functional subregions of the BiFaDmel.

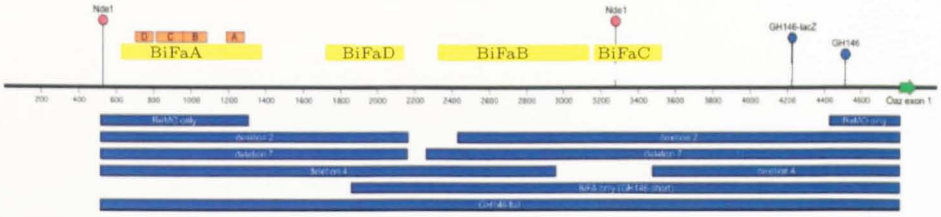

Figure 5.4: GH146 enhancer regions and some of the deletion constructs. Golden rectangles known as ReMo A, B, C and D are sequence-conserved regions detected by the ReMo algorithm. The ReMo-Only construct (blue rectangle underneath the map) was made to test significant of this prediction. Four yellow rectangles are motif-rich subregions and identified by the BiFa tool. BiFaA is overlapped with ReMo region, but the other three BiFa regions made up the BiFa-Only construct (the second blue rectangle from the bottom). The GH146-Full construct included the entire $4kb$ upstream of the *oaz* (the last blue rectangle). Deletions 2, 4 and 7 in this figure were made based on the RRS predictions. Blue circles are GH146 P-element and lacZ insertion points. The amber circles called NdeI1 are restriction enzyme sites. The green arrow is the start of the *oaz* gene.

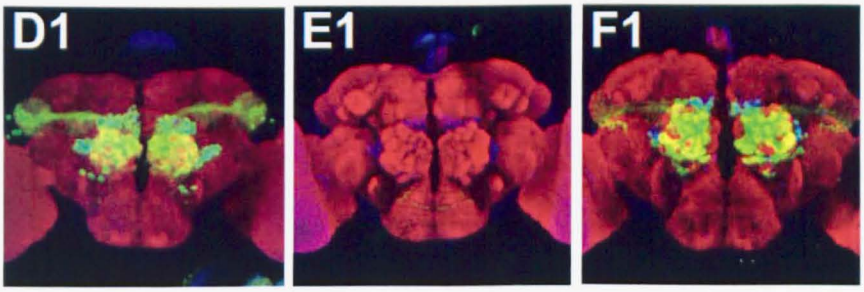

Figure 5.5: Expression patterns driven from the GH146-Full construct indicated as D1, from the ReMo-Only as E1 and from the BiFa-Only as F1. For more details of this figure the reader is referred to (71).

5.3.1 Identifying sequence regions for analysis

As observed, the ReMo C was conserved in all of the Drosophila species and also that BiFa-Only construct in *D. melongaster* (denoted as BiFaDmel) drove the same expression pattern as the GH146-Full. By using BLAST for any other species, we identified the position of a subsequence that was conserved with ReMo C and also the *oaz* gene in that species genome. Within this interval, a subse-

quence with the same length as BiFaDmel was chosen such that its distance from the ReMo C counterpart was the same as the distance of the BiFaDmel from the ReMo C in the *D. melanogaster* genome. In this way, corresponding to any of the 10 species, we obtained the BiFa-Only sequences (see Table 5.1 on page 90). These 10 sequences plus 67 fly PWMs as described in Chapter 4 (see Table 4.2 on page 64) were used for the RRS analysis.

Species	Sequence Name
D. melanogaster	BiFaDmel
D. simulans	BiFaDsim
D. sechellia	BiFaDsec
D. yakuba	BiFaDyak
D. erecta	BiFaDere
D. ananassae	BiFaDana
D. pseudoobscura	BiFaDpse
D. mojavensis	BiFaDmoj
D. virilis	BiFaDvir
D. grimshawi	BiFaDgri

Table 5.1: BiFa-Only Regions and corresponding species used for the RRS analysis

5.3.2 Detection of orthology between Drosophila species

Our next step was to detect the functional conservation of BiFaDmel in other species. For this, the RRS scores of BiFaDmel as the template vs any of the other 9 sequences as the test sequence was computed. We found that BiFaDsim, BiFaDsec and BiFaDyak in order were the top three functionally conserved sequences to BiFaDmel.

The RRS results of the comparisons of the BiFaDmel vs the BiFaDsim, BiFaDpse and BiFaDgri are illustrated in the Figure 5.6 on page 93. In the left-hand side of the figure (A1, A2 and A3), we have illustrated the log of the RRS scores of comparisons of BiFaDmel vs BiFaDsim (which is about 12), BiFaDpse (about -3) and BiFaDgri (about -4) as vertical green lines. The significance of these

scores can be seen when compared to the scores of BiFaDmel vs 1000 randomly picked sequences from *D. simulans* (black histogram in A1), *D. pseudoobscura* (A2) and *D. grimshawi* (A3).

From A1, we can see that the log of the RRS score for BiFaDmel vs BiFaDsim is around 12 whereas the maximum score of BiFaDmel vs 1000 random sequences is about 3. This fact supports the statistical significance of the RRS score of BiFaDmel vs BiFaDsim.

From A2, it is clear that the log of the RRS score for BiFaDmel vs BiFaDpse is around -3 and that only 5 out of 1000 of the random sequences are obtaining a score greater than or equal to the score of BiFaDmel vs BiFaDpse. Although this orthology signal detected by the RRS might not look very strong, it becomes interesting when we note that the alignment-based ReMo algorithm does not detect any conserved subregions in these region of the sequence (see part B of Figure 5.7 on page 106).

And finally from A3 we see that the log of the RRS BiFaDmel vs BiFaDgri is about -4 and about 30 of the random sequences are obtaining a score greater than or equal to the score of BiFaDmel vs BiFaDgri. This score may not seem statistically very significant in the first instance, we may argue that firstly it is still greater than 97% of the scores from the random sequences secondly *D. grimshawi* was the most distant species in our analysis and thirdly the alignment-based comparison of BiFaDmel vs BiFaDgri does not show any statistically significant sequence conservation in this region (see part C of Figure 5.7 on page 106). Therefore, the BiFaDgri still can be suggested as orthologous.

In the right-hand side of the figure (B1, B2 and B3), the contribution of the individual motifs in any of these three comparisons are depicted. As we explained in Section 3.3, in the RRS framework, an individual motif score around zero is expected from a random DNA sequence, but the greater the scores means the stronger presence of the motif (either by multiplicity or by specificity).

A very interesting point to note in this figure is that in comparison of the BiFaDmel vs the BiFaDsim where we had the strongest RRS score among the other species, we can see a strong right-hand side tail in the histogram of individual motif contributions (B1). This is means that motifs distributed in this tail (for a list of top 8 contributors see Figure 5.2 on page 95) are significantly present in

both of the sequences and that the greater the score, the greater the contribution of that motif in functional conservation of that pair of sequences. But in comparison of BiFaDmel vs BiFaDpse and BiFaDmel vs BiFaDgri where the scores were not as significant as BiFaDmel vs BiFaDsim, the histograms of individual motif contributions are almost a normal distribution with mean zero. This can be interpreted that occurrences of most of the motifs over these two comparisons (B2 and B3) are the same as their occurrences in the random sequences.

Another point of interest is the common regulators of these top three RRS-scored sequences. In Table 5.2 on page 95 we have presented 8 key regulators from comparisons of BiFaDmel vs any of the BiFaDsim, BiFaDsec and BiFaDyak. The common regulators of these comparisons have been colour-coded in Table 5.2 (where a factor being common at least between two sequences has been coloured and non-common regulators have been left with a white background). One can easily see that the number of common regulators in this table is a direct proportion to the RRS score of BiFaDmel vs that species. In other word, the BiFaDsim gets the most significant RRS score and the corresponding row of the table (row1) is fully coloured whereas in the row corresponding to the BiFaDyak we see only four coloured cells. This fact is supporting the contribution of the key regulators in these orthologous sequences.

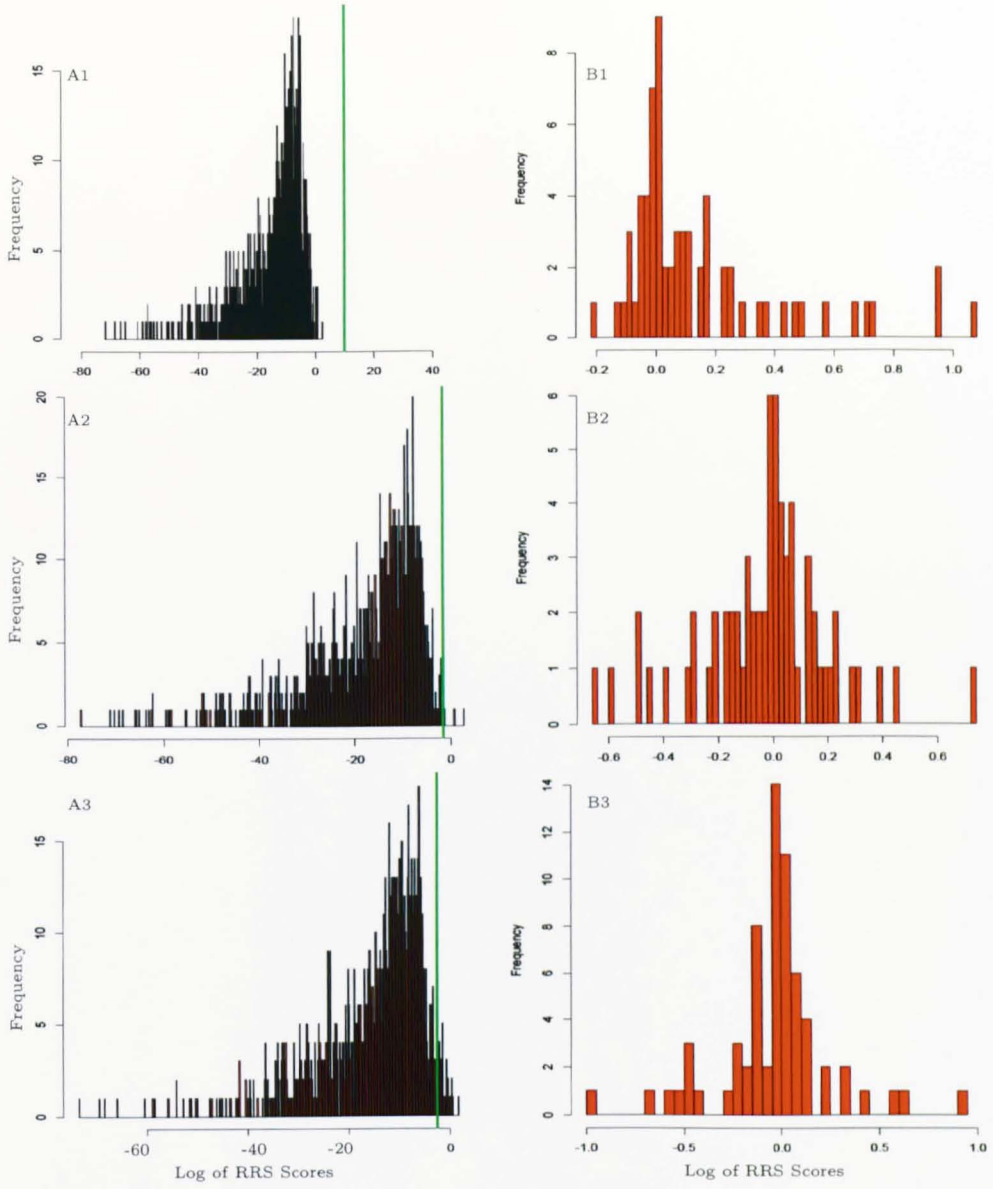

Figure 5.6: A1: Illustrating the statistical significance of the RRS score of BiFaDmel vs BiFaDsim. The green vertical line shows the log of the RRS score from this pair which is around 12. The black histogram shows the distribution of log of the RRS scores of BiFaDmel vs 1000 randomly picked sequences from the *D. simulans* genome. A2 and A3: same as A1, but comparison of BiFaDmel vs BiFaDpse and BiFaDgri respectively. From A2 one can see that around five random sequences are scoring greater the score of BiFaDmel vs BiFaDpse. This number in A3 increases to 30 sequences (out of 1000), reducing the significance of the RRS score, but still can be considered as a signal of orthology detection by the RRS. Figures B1 ,B2 and B3 are depictions of the contribution of any of the 67 motifs (used in this analysis) in the RRS scheme, BiFaDmel vs BiFaDsim in B1, BiFaDmel vs BiFaDpse in B2 and BiFaDmel vs BiFaDgri in B3. For the list of top eight contributors see Table 5.2 on page 95

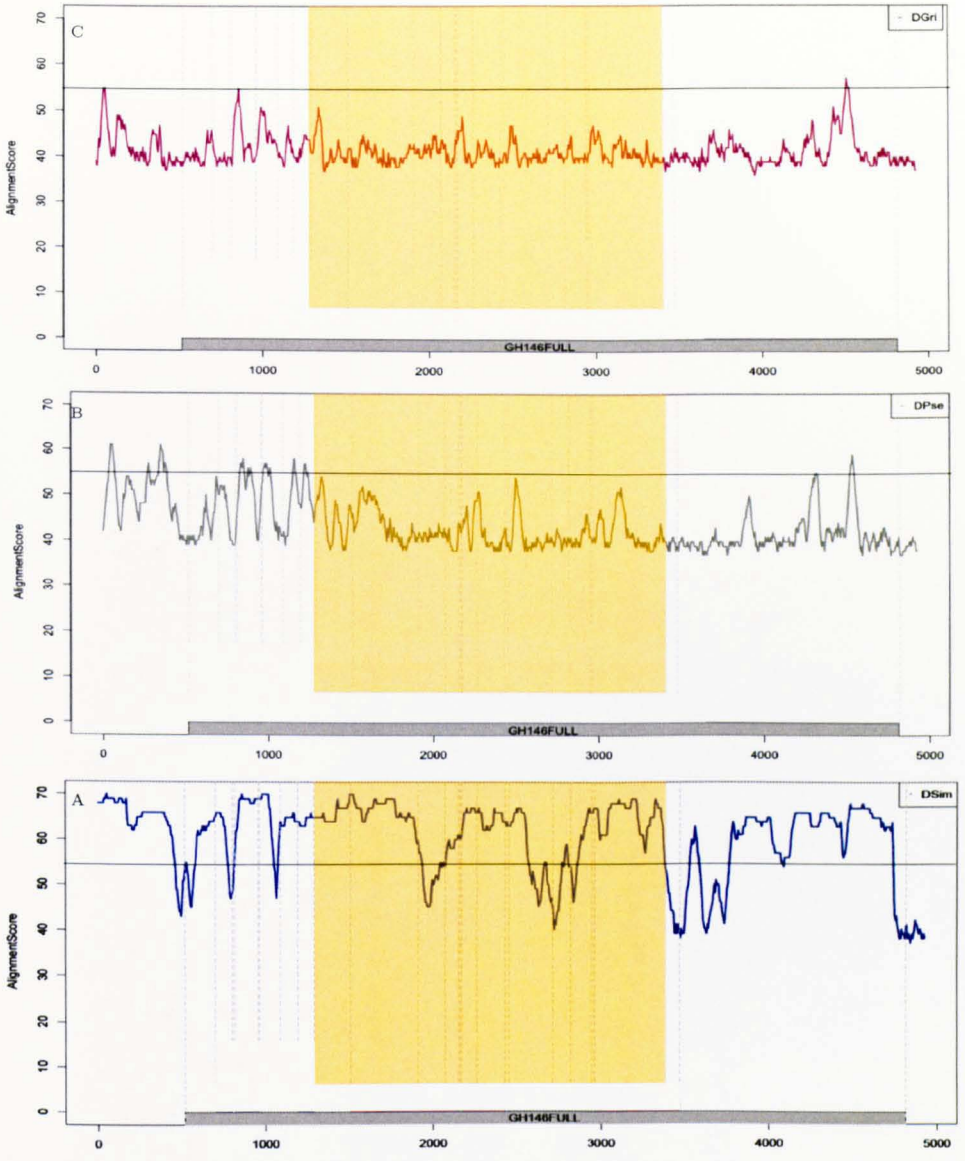

Figure 5.7: Alignment profiles (ReMo algorithm) of the $5kb$ upstream of the gene *oaz* in *D. simulans*, *D. pseudoobscura* and *D. grimshawi* vs $5kb$ upstream the *oaz* in *D. melanogaster*. The shadowed area is to highlight the BiFa-Only region- the region of interest. The horizontal lines are reference lines (equal to 55 in this figure) to show the significance of the alignment scores.

On the other hand, one may also note that, the RRS suggests BiFaDsim, BiFaDsec and BiFaDyak as (in order) the strongest functionally linked sequences to the BiFaDmel and in these comparisons we have listed 8 top contributors. However, for instance in comparison of BiFaDmel vs BiFaDsim, the sum of contributions of these 8 regulators is about 6, and by subtracting this number from the real score (in order to ignore the contributions of key regulators) which was about 12, we will still have a score about 6 which is higher than scores of BiFaDmel vs randomly picked sequences. Overall, we may conclude that in order for a pair of sequences be considered functionally linked by the RRS with a high statistical significance, the contribution of the key regulators are necessary but not sufficient.

The last point that we would like to make in this section is that according to same table, for example, Engrailed (the factor identity is I$En) and Kruppel (the factor identity is I$KR) are listed between top 8 regulators of BiFaDmel vs BiFaDsim and BiFaDsec but not in BiFaDmel vs BiFaDyak. On the other hand, *D. melanogaster, D. simulans, D. sechellia* and *D. yakuba* are belonging to the melanogaster subgroup in the fly phylogenic tree (see Figure 5.2 on page 83). One may expect that a significant occurrence of a motif in three species of the this subgroup (i.e., BiFaDmel, BiFaDsim and BiFaDsec) might imply its occurrence in BiFaDyak. But in our example, the occurrences of the I$En and I$KR in BiFaDyak are not statistically significant. This might mean that according to the RRS results, we have had a loss of these motifs over the evolution.

Pair of Enhancers	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8
Mel vs Sim	I$TCF_1	I$FTZ_1	I$SN_1	I$En_Q6	I$Kr_Q6	I$DL_1	I$UBX_1	I$DL_2
Mel vs Sec	I$En_Q6	I$Kr_Q6	I$SN_1	I$UBX_1	I$DL_2	I$HSF_Q4	I$TCF_1	I$ADF_Q6
Mel vs Yak	I$FTZ_1	I$SN_1	I$DRI_1	I$TLL_Q5	I$STAT_Q1	I$DL_1	I$ZEN_Q6	I$TCF_1

Table 5.2: The top eight factors (in a descendent order) that are strongly contributing to the functional similarities of BiFaDmel vs any of BiFaDsim, BiFaDsec and BiFaDyak. Colour-coding is to highlight the common regulators. Factors that are common in at least two species have been coloured the same, the rest have been left with a white background.

5.3.3 Results of *in silico* deletions in *D. melanogaster*

As mentioned, the main idea was to detect subregions of BiFaDmel that are likely to drive an expression pattern similar to the expression pattern driven by BiFaDmel itself, and thereby further defining the function enhancer boundaries. Having observed a statistically significant link between functional conservation of BiFaDmel and BiFaDsim in Section 5.3.2, it was natural to take BiFaDsim as the template sequence and then scan subregions of BiFaDmel as the test sequence for functionally similar subsequences. One way of doing that was to delete some subsequences of BiFaDmel and find deletions which induce a drop in RRS score. For such deletions the drop in the RRS scores means that the functional similarity of the BiFaDsim and the BiFaDmel can no longer be detected by the RRS method.

Therefore, in a sliding manner, with step size $25bp$ we deleted subsequences of a fixed window length from the BiFaDmel and each time the remaining subsequence was considered as a test sequence. This scenario was repeated with different window lengths including $50, 100, 150, 200, 250, 300, 350$ and 500bp and the results were plotted.

The results of this deletion analysis with window lengths $100, 150, 250$ and 500 are shown in Figure 5.8 on page 98. With respect to these predictions we would like to make the following points:

- The troughs in these profiles mean that by deleting the corresponding window, we have had an extreme loss of the RRS score which in turn means that deleted window must be the most functionally similar subsequence to the template sequence.

- x-axis depicts the length of step size. Because the step size for this analysis was $25bp$, in order to get the starting position of the deletion window, one may need to multiply the numbers corresponding to any of the troughs by 25.

- As we can see, the starting positions for the suggested deletion windows are dependent to the deletion window length. In other words, if one needs a deletion with length 100bp, then the RRS is suggesting a subsequence

starting from $50 \times 25 = 1250$ in BiFaDmel, whereas if one needs a deletion with length $250pb$, then the suggested starting position is $55 \times 25 = 1375$.

- The x-component of the last point in any of these profiles is less than or equal the the length of BiFaDmel sequence minus the deletion window length.
- One may observe that the shorter the deletion window length, the sharper the corresponding profile.

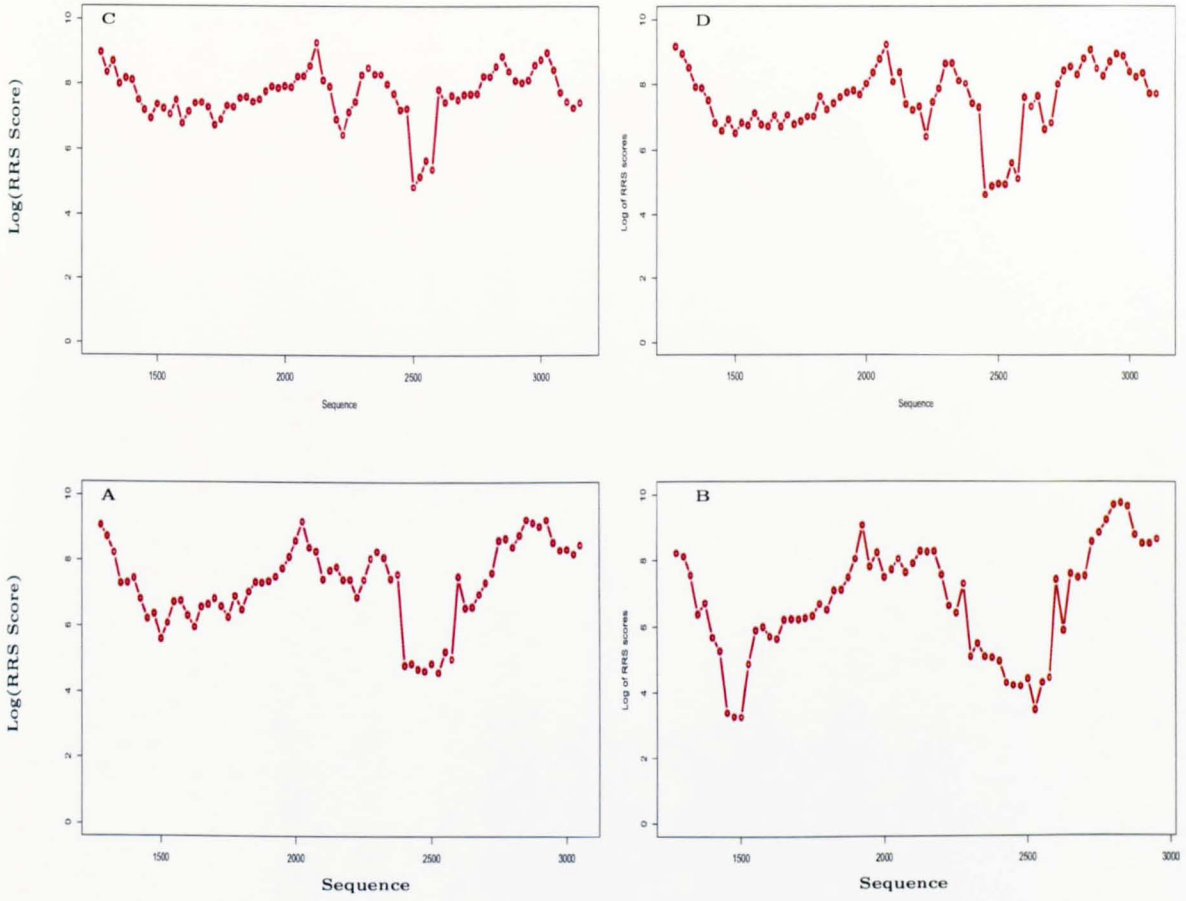

Figure 5.8: Plots from deletion subsequences: The first plot shows the *RRS* scores of BiFaDsim vs some subsequences of BiFaDmel each of which was obtained by deleting a region with length 300*bp* as depicted in part A. Parts B, C and D are the same but with window lengths 200, 100 and 150*bp* accordingly. Note that numbers in x-axis are based on coordinates of GH146Full sequence.

We repeated the same analysis but with BiFaDsec, BiFaDyak and even with BiFaDmel itself as template sequences. Interestingly, the peaks and troughs in corresponding output profiles were in a high agreement with those suggested from the analysis of BiFaDsim. Whereas when we chose the BiFa region from a more distant species for example BiFaDgri as template sequence, we had a flatter profile. This makes the results of these predictions more significant because we have already observed that these were the best functionally conserved sequences

to the BiFaDmel. Figure 5.9 on page 99 shows the profiles of deletion analysis where the template sequence was BiFaDyak.

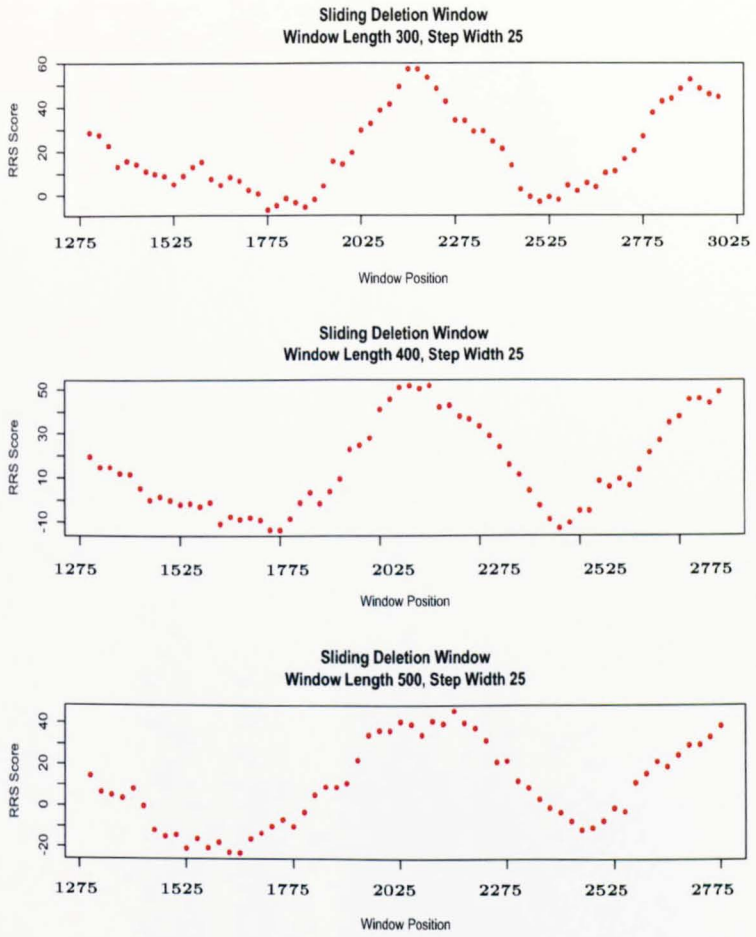

Figure 5.9: Plots from deletion subsequences: The first plot shows the *RRS* scores of BiFaDyak vs some subsequences of BiFaDmel each of which was obtained by deleting a region with length $300bp$ starting at positions $0, 25, 50, \cdots$. The second and third plots show the same for window lengths of 400 and $500bp$. Note that numbers on the x-axis must be multiplied by the stepsize (i.e. 25) to obtain the deletion position in the sequence.

5.3.3.1 Experimental results of our deletion predictions

According to these RRS deletion predictions, our collaborator made 7 deletion constructs. Figures 5.11 on page 101 and 5.4 on page 89 are showing more details

about how and where these deletion constructs were made. Not all of these constructs have been experimentally tested. Continued experiments are likely to reveal more details about the significance of these predictions and consequently about the logic of this enhancer region. However, experiments based on some of these deletions (deletions 2,4 and 7) revealed that the GH146 expression pattern is differently affected by different deletions (see Figure 5.10 on page 101) suggesting that deletions 2 and 7 are likely to contain repressor elements whereas deletion 4 is likely to contain some promoter elements.

From the experiments completed so far one may argue that the effect of these deletions seem to be more phenotypic. For instance, deletions 2 and 7 are both overlapped and both drived expression of some cells outside of the PNs. For more details of results of the completed experiments the reader is referred to Chapter 4 of (71).

Although making a final conclusion for this project requires all the experiments from the deletion constructs to be completed, based on current state of the project the following discussion can be made:

On one hand, we have observed (both theoretically and and experimentally) that BiFaDmel is the main functional region of the GH146 enhancer. On the other hand, bioinformatical analyses suggest that some subregions of the BiFaDmel are likely to have the same expression pattern as the BiFaDmel itself. But the result of experiments are not as significant as the bioinforamtical evidence. The simplest conclusion that one can draw is that the bioinformatical analysis was not accurate enough. This might be due to the inappropriateness of the PWMs used in these analysis. However, we can argue that the expression of the BiFaDmel is likely to be a result of a combinatorial effect of some shorter functional subregions. At this moment, I do not know how one can experimentally test this hypothesis.

A further step that will lead towards a more confident conclusion is to perform some experiments in which each deletion construct is accompanied with an eqi-length control deletions corresponding to the peak of that deletion profile. It will be also very informative to see the affect of deleting two regions corresponding to two non-overlapped troughs of an RRS deletion profile made in one construct.

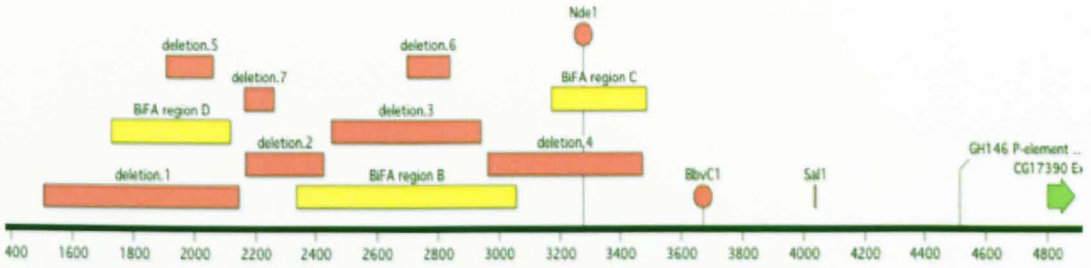

Figure 5.11: According to the RRS predictions for deletions regions, our collaborators made some constructs (red rectangles).

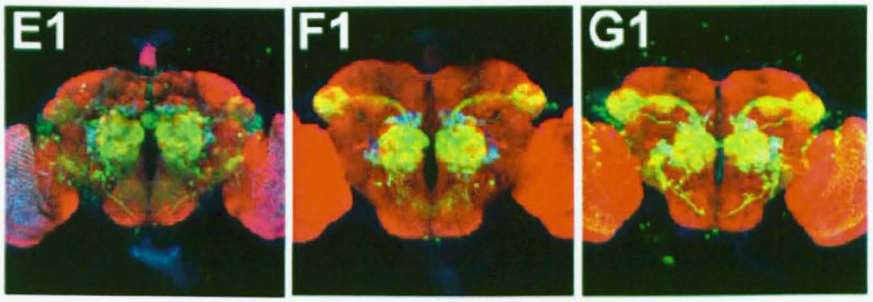

Figure 5.10: Expression patterns: E1 drived from the construct of deletion 2, F1 drived from deletion 4, G1 from deletion 7.

5.3.4 Reconstruction of a phylogenetic tree from regulatory sequences

We found that the RRS was able to detect the evolutionary links between the species with a high level of accuracy. The resulting phylogenetic tree was of significant interest because it was made only from the BiFaDmel and its homologous in 9 other species (i.e., BiFaDsim, BiFaDsec, BiFaDyak, BiFaDere, BiFaDana, BiFaDpse, BiFaDmoj, BiFaDvir and BiFaDgri) and 67 PWMs that used in this analysis. Figure 5.12 on page 103 shows the heatmaps and phylogenic tree made by the RRS. (A) is the heatmap that made only from these sequences and (B) is the heatmap that made from these sequences plus two randomly picked se-

quences as controls. Each row in heatmp was considered as the template and each column was considered as the test sequence. Therefore similarity of a pair of sequences can be judged by comparing colours of related rows. The trees in the left hand-side of the heatmaps (made by the similarity of rows) reflect the functional similarity of the species. We see that the RRS can distinguish the random sequences as outliers.

One may argue that there must be a pattern of occurrences of some of the (possibly key) regulators governing this evolutionary link between these species that are picked up by the RRS. To address this question lets once again have a look to the (log of) the RRS score for BiFaDmel vs BiFaDsim which was nearly 12. In Table 5.2 on page 95 we have represented the top eight key contributors of this similarity score. The overall contribution of these 8 regulators is about 6.5. We should note that practically it is almost impossible to force these 8 regulators to score zero (by deleting or filtering their sites), because according to the RRS framework these scores are made up by looking through all the possible configurations and accounting even very weak binding site effects. But, for a moment lets assume we have managed to force these regulators to obtain an overall zero contribution, then the rest of motifs will assign a score around 6 to this pair which is still significantly more than scores of BiFaDmel vs randomly picked sequences from the *D. simulans*. Meaning that the functional similarity of a pair of sequences in the RRS framework is influenced by contribution of weak binding sites too. Therefore it is really too hard to propose a simple pattern behind the phylogenetic tree made by the RRS, as it seems to be made by more than a simple pattern.

As a future direction point, it worth mentioning that according to this analysis the BiFaDsim, BiFaDsec and BiFaDayk that were detected as orthologous sequences to the BiFaDmel are very likely to drive a similar expression pattern when planted to the *D. melanogaster* genome. This has not been experimentally tested yet. An experiment targeting this hypothesis will provide new insights into evolutionary significance of the enhancer regions.

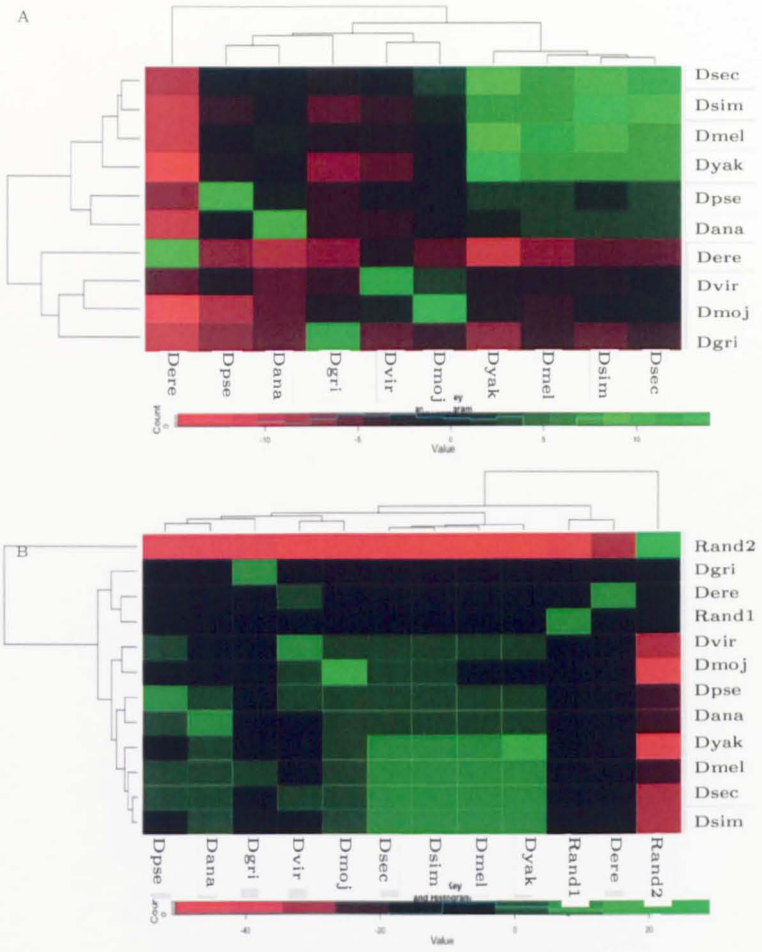

Figure 5.12: A: Heatmap made by RRS similarity (log) scores from BiFaDmel, BiFaDsim, BiFaDsec, BiFaDyak, BiFaDere, BiFaDana, BiFaDpse, BiFaDmoj, BiFaDvir, and BiFaDgri. Please note that in this figure for the sake of simplicity a sequence name such as BiFaDmel has been denoted by Dmel and so on for other sequences. B: The same as A but with two extra randomly picked sequences from the *D. simulans*.

5.3.5 Consistency of the RRS predictions with alignment based methods

At this stage one may wonder if these deletion predictions made by the RRS are producible by any of the alignment based methods and also that how and why these agreements and/or disagreements are for. Our objective in this section is to address this question. For this, we made the alignment-based ReMo algorithm profiles of $5kb$ upstream of gene *oaz* of any other 9 species vs $5kb$ upstream of the *oaz* in *D. melanogaster*. Although we looked at the alignment profiles of $5kb$ upstream of the gene to get an overall image, one may need to concentrate only on BiFaDmel (shown as orange shadowed area in Figure 5.13 on page 106) region. This is because we had some both theoretical and experimental evidence, as explained, that this region was driving the same expression pattern as the GH146-Full and the original idea was to dissect this region and detect its functional subregions with the RRS. The results of this alignment-based analysis have been presented in Figure 5.13 on page 106 and Figure 5.7 on page 94.

In both the RRS and the ReMo algorithms, similarity scores are linear to the evolutionary distance of the species under comparison to *D. melanogaster*. This can be seen from Figures 5.6 on page 93 and 5.7 on page 94. In both algorithms, BiFaDsim is the most similar sequence to BiFaDmel and the similarity score falls in more distant species such as BiFaDpse and BiFaDgri sequences. However, it seems that the significance of similarity level in the RRS model is higher for BiFaDpse and BiFaDgri. For instance, in Figure 5.7 on page 94 where *D. grimshawi* is compared to *D. melanogaster*, it is too hard to point out any conserved subregions in the BiFa-Only region.

From Figure 5.13 on page 106, we can see that deletions 1,2,7, and 4 (red rectangles in the figure) which were made based on our RRS predictions are in a high agreement with some peaks of the alignment profiles, whereas deletions 5, 3 and 6 (golden rectangles) are matching with some troughs of the profiles. From these comparisons, we can not draw any conclusion about the level of the significance of any the RRS or the alignment-based model. To make this point more clear we should note that:

- In Chapter 4 and also (45) and (24) we have seen examples of functionally conserved non-alignable sequences. Detection of these types of functionally conserved sequences has been the initial reason of developing alignment-free DNA sequence comparison algorithms including the RRS. Therefore, we will not be surprised if any of these predictions made by the RRS were not identified by any of the alignment-based methods.

- On the other hand, we do not expect the RRS to detect all the regions that have been identified as conserved sequences by the alignment-based method as functionally conserved regions. The reason is that the RRS judgement about the similarity of a pair of sequences is based on the distribution of a set of PWMs that was passed to it as an input (in this analysis only 67 fly PWMs). According to the appropriateness of these motifs, we may or may not have significant RRS score for a pair of sequences with high level of sequence conservation.

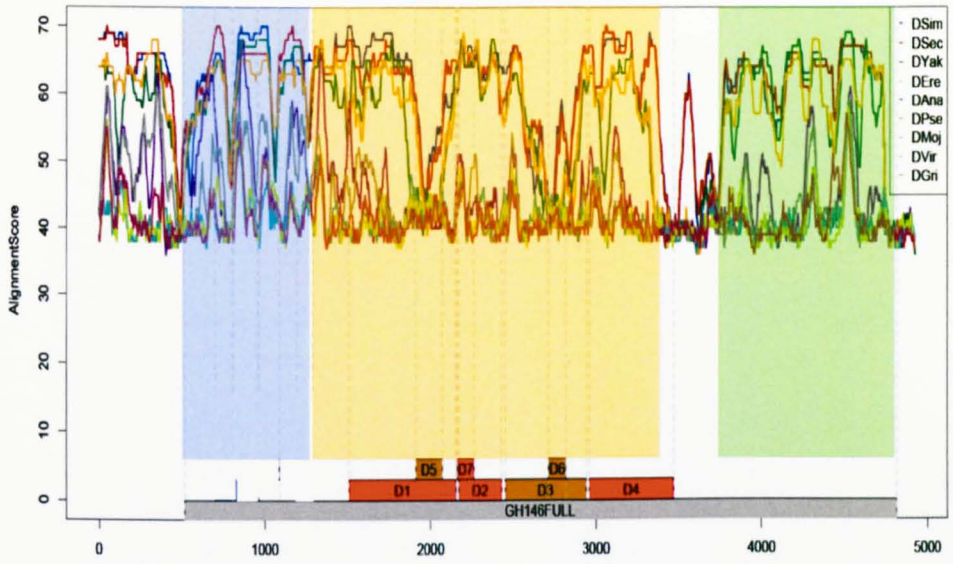

Figure 5.13: Comparison of predictions of the RRS with the alignment based method. Shown here are the ReMo algorithm profiles of the $5kb$ upstream of the gene *oaz* of 9 Drosophila species used in this analysis compared with their homologs in *D. melanogaster*. The x-axis is the sequence position (5' to 3' upstream of the *oaz*) and y-axis is the alignment scores. The grey rectangle in the x-axis is to show the area that drove a full expression pattern in PNs. The red rectangles D1, D2, D7 and D4 are the RRS deletion predictions that look to be in a good agreement with some peaks from the alignment profiles. The golden rectangles D5, D3 and D6 are the RRS deletion prediction that look in a disagreement with the sequence conserved subregions detected by the ReMo algorithm. The area shadowed as orange is to highlight the BiFa-Only region that drove the same expression pattern as the GH146-Full and therefore is the region of the interest. The area shadowed as blue is to highlight the ReMo-only region, and the area shadowed as green is the promoter area and of no interest in this analysis.

5.4 Conclusion

We have presented our predictions of the functional subregions of the GH146. These deletion predictions were made by our RRS model. However, prior to the development of the RRS, an alignment based-model (the ReMo algorithm), and a motif-scanning based model (BiFa tool) were used and identified ReMo-Only BiFa-Only regions. These enhancer subregions were corroborated by some experiments and revealed that although the ReMo-Only is conserved in almost all of the species, the ReMo-Only construct is not capable of driving PN expression pattern. On the other hand, the BiFa-Only region recapitulate the expression pattern, suggesting that the functional enhancer region lies in the BiFa-Only region. Our RRS model was then used to predict these functional regions. According to these deletion predictions, 7 deletion constructs were made, but the function of all of these 7 constructs have not been yet completely experimented. The results from three of these constructs revealed that the effect of these deletions is likely to be phenotypic and also that deletions contain both promoter and repressor elements. We also found that the RRS looks to be capable of picking up the evolutionary links between species surprisingly from only (short) regulatory sequences and a (small) set of PWMs.

This project is still ongoing and we believe that cross-referencing results from the underlying experiments to our predictions will make new insights into the regulatory code in fly olfactory system and also will signify our model development. But based on currently existing results, we can set up the following discussions and future directions to this project.

- According to the RRS results (see 5.12), BiFaDsim, BiFaDsed and BiFaDyak are functionally conserved to the BiFaDmel with a high statistical significance (this is supported by alignment-base tools as well, see 5.13A). This suggests that rather than taking a single template sequence, a multi-template version of the RRS where the set of BiFaDmel, BiFaDsim, BiFaDsec and BiFaDyak will be considered as the template set, will strength the significance the RRS results.

- Another pertinent point to make is that the result of the RRS is strongly dependent to the set of input PWMs. Thus a more appropriate set of PWMs will lead to more accurate, conclusive and meaningful results. On the one hand, the computational expenses is not allowing to take a very big set of PWMs (for example all the available PWMs), on the other hand a set of PWMs with high level of redundancy may introduce some noise to the model. Thus, a set of non-redundant PWMs that includes all the possible key regulators of the systems is suggested.

- Some control experiments are required to evaluate the significance of the RRS predictions for instance for a construct corresponding to a trough of an RRS profile and another construct made for either a peak or a plateau area of the profile would reveal the significance of the RRS predictions. The control experiments can be based on some constructs corresponding to the peaks of the RRS profiles. In addition, some experiments assessing the combinatorial effects of shorter functional subregions will enhance our understanding from the transcriptional machinery.

Appendix A

Loss-free Identification of Alignment-Conserved CRMs

In this appendix we provide the reader with a comprehensive description of the ReMo algorithm that we applied in Chapter 5 to detect alignment-conserved non-coding subregions in *D. melanogaster* GH146 enhancer. This description includes the proof of correctness, and evaluates the algorithm's running time. Please note that this data has been provided by developers of the algorithm and therefore analysis and results mentioned in this appendix are to show the advantages of the ReMo-algorithm. There is no direct relationship between this analysis and my PhD project, and I have had no contribution to this analysis.

A.1 Introduction

We define a pair of genomic sequences as alignment-conserved if their optimal alignment has a statistically significant score and the sequences are not repeats. Using alignments rather than TF binding motifs to identify potential CRMs provides a relatively unbiased approach as CRMs containing yet undescribed binding motifs can be identified as well.

The most frequently used algorithms for CRM-detection are members of the BLAST-family. These are heuristic algorithms that can not guarantee to find weakly conserved regions, but are relatively fast and, therefore, currently being employed by browsers for non-coding conserved regions (5; 34).

We have used the algorithmic techniques described below to perform a loss-free genome-wide scan for conserved non-coding regions in the vicinity of mouse and fugu genes. We found that about half of the alignment-conserved regions between mouse and fugu show a sequence similarity below 70 percent (see Figure A.1). Given that BLAST was found to fail in more than 60 percent of cases in a study based on randomly generated sequences (41), this heuristic is bound to miss a substantial number of biologically relevant regions, in particular for distantly related species.

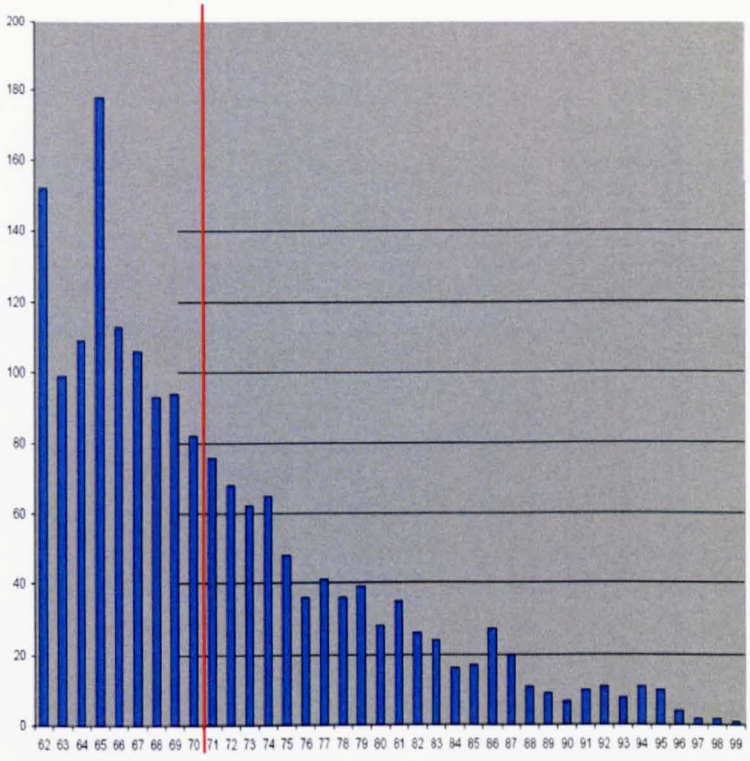

Figure A.1: Number of conserved regions in fugu detected in the vicinity of 10272 mouse genes including most transcription factors, segmented by maximal degree of conservation in windows of 100 bases. For a conservation of 62 to 64, numbers are reduced to those regions for which significant conservation was also found in at least one other species than mouse and fugu.

As an alternative approach optimal local alignments of upstream regions have been employed (13). However, these can fail to detect biologically significant

conservation on short stretches as long but meaningless alignments can cross the alignment path of the shorter alignment in the Smith-Waterman matrix (called *shadow effect* (4; 68)). To avoid this problem a method for maximising the ratio of alignment score to sequence lengths has been proposed (4). As the authors indicate themselves their method suffers from the dependence on a parameter for which no general selection rule has been given. Therefore, even this method is not guaranteed to find all alignment-conserved CRMs.

A.2 Naive Algorithm

The following algorithm provides a straightforward approach to ensure detection of all short alignment-conserved regions within two stretches s and t of genomic DNA (such as the upstream regions of two orthologous genes). The basic idea is to compute an optimal alignment for every pair of short substrings of s and t.

Algorithm 16

Step 1:	Read input: strings s, t, step width w, and a window length l
Step 2:	Compute the minimal alignment score S that is still statistically significant for two sequences of length l.
Step 3:	Compute number of window positions: $n_1 = \lfloor((\|s\| - \|l\|)/\|w\|) + 1\rfloor$ $n_2 = \lfloor((\|t\| - \|l\|)/\|w\|) + 1\rfloor$
Step 4:	Initialise variables: set $R = \emptyset$
Step 5:	For all pairs (i, j) with $1 \leq i \leq n_1$ and $1 \leq j \leq n_2$:
Step 5A:	Apply the Needleman-Wunsch algorithm to compute the optimal alignment score of substrings $s[(i-1)w, (i-1)w + l - 1]$ and $t[(j-1)w, (j-1)w + l - 1]$ N = optimal alignment score of window-pair (i, j)
Step 5B:	If $N \geq S$, then add (i, j, N) to set R.
Step 6:	Output: R

As Needleman-Wunsch alignments require $O(l^2)$ dynamic programming (DP) steps the feasibility of Algorithm 16 is limited. For example, if two 100kb sequences are considered, and the step width is set to $w = 5$, a total of about 400

million optimal alignments, each requiring 10,000 DP steps are needed. While this is feasible for the comparison of a limited number of genes across a limited number of species, it is not realistically applicable to genome-wide scans.

Finding a reasonable setting for the window length l is not a problem in practice as it is sufficient for the window to cover only a part of a CRM. In this case a number of significant window pairs will be found which can be grouped and displayed as a single block of conserved sequence.

The statistical evaluation of sequence alignment scores is greatly simplified by our approach as all aligned sequences have the same length.

A.3 Our Algorithm

The key idea to improve the sliding-window approach of Algorithm 16 is to make use of previously computed alignment scores for other pairs of windows in order to reduce the CPU-time needed to do the computation for following window pairs. This is done by deriving upper and lower bounds for the alignment score of a given window pair, before the application of Needleman-Wunsch is considered. If the upper bound is lower than the cut-off S, the alignment would not be part of the final output and can be omitted. If the lower bound is high, an alignment has to be computed, but the Needleman-Wunsch matrix can be restricted to a tight corridor around the main diagonal as alignment paths that deviate from this corridor would not be optimal. A full applicaton of Needleman-Wunsch is only required if neither bound provides a computational saving.

We also add the computation of conservation profiles for each input sequence to the algorithm. These are informative in practice, but are not part of the speed improvement over Algorithm 16.

Algorithm 17

Step 1: Read input: strings s, t, step width w, and a window length l

Step 2: Compute the minimal alignment score S that is still statistically significant for two sequences of length l.

Step 3: Compute number of window positions:
$n_1 = \lfloor((|s| - |l|)/|w|) + 1\rfloor$
$n_2 = \lfloor((|t| - |l|)/|w|) + 1\rfloor$

Step 4: Initialise variables:
set $R = \emptyset$
vector P_1 of length n_1, $\forall i : P_1[i] = 0$
/* conservation profile first sequence */
vector P_2 of length n_2, $\forall j : P_2[j] = 0$
/* conservation profile second sequence */
$n_1 \times n_2$ matrix $M_{\min}$, $\forall i, j : M_{\min}[i, j] = -\infty$
/* to store lower bounds for alignment scores */
$n_1 \times n_2$ matrix $M_{\max}$, $\forall i, j : M_{\max}[i, j] = \infty$
/* to store upper bounds for alignment scores */

Step 5: For all pairs (i, j) with $1 \leq i \leq n_1$ and $1 \leq j \leq n_2$ (in any order):

Step 5A: Compute lower bound:

$m_1 = \max\{M_{\min}[i-1, j], M_{\min}[i, j-1], M_{\min}[i, j+1], M_{\min}[i+1, j]\}$

/* best score moving sideways */

$m_2 = \max\{M_{\min}[i-1, j-1], M_{\min}[i+1, j+1]\}$

/* best score moving on one diagonal */

$m_3 = \max\{M_{\min}[i-1, j+1], M_{\min}[i+1, j-1]\}$

/* best score moving on other diagonal */

$b_L = \max\{m_1 - w + 2w\delta, m_2 - w, m_3 - 2w + 4w\delta\}$

Step 5B: Compute upper bound:

$m_1 = \min\{M_{\max}[i-1, j], M_{\max}[i, j-1], M_{\max}[i, j+1], M_{\max}[i+1, j]\}$

/* best score moving sideways */

$m_2 = \min\{M_{\max}[i-1, j-1], M_{\max}[i+1, j+1]\}$

/* best score moving on one diagonal */

$m_3 = \min\{M_{\max}[i-1, j+1], M_{\max}[i+1, j-1]\}$

/* best score moving on other diagonal */

$b_U = \min\{m_1 + w - 2w\delta, m_2 + w, m_3 + 2w - 4w\delta\}$

Step 5C: Compute minimum score to influence final results:

$A = \min\{P_1[i], P_2[j], S\}$

Step 5D: If $(b_U < A)$ then jump to Step 5J

Step 5E: Compute corridor of interest:

$C = \lceil \frac{l - M_{\min}[i,j]}{1-2\delta} \rceil$

Step 5F: Apply the Needleman-Wunsch algorithm to compute the optimal alignment score of substrings $s[(i-1)w, (i-1)w + l - 1]$ and $t[(j-1)w, (j-1)w + l - 1]$

N = optimal alignment score of window-pair (i, j)

Only compute the corridor of the Needleman-Wunsch matrix that is within C positions off the main diagonal.

Step 5G: If $N \geq S$, then add (i, j, N) to set R.

Step 5H: If $N \geq P_1[i]$ then set $P_1[i] = N$

Step 5I: If $N \geq P_2[j]$ then set $P_2[j] = N$

Step 5J: Store computed bounds:

$M_{\min}[i, j] = b_L$ (or N if computed)

$M_{\max}[i, j] = b_U$ (or N if computed)

Step 6: Output: R, P_1, and P_2

For readibility we ignore undefined indexing of matrices $M_{\max}$ and $M_{\min}$ such as $M_{\max}[0, 0]$ - these would be replaced by ∞ or $-\infty$ in real programme code.

We decided to employ the original Needleman-Wunsch algorithm as a subroutine for Step 5F (49), since the existing subquadratic algorithms do not make an

improvement for this application (2; 11; 46).

A.4 Correctness

We only need to prove that the upper and lower bounds computed in Algorithm 17 are correct. We formulate our Lemma for the special case of a match-score of 1 and a mismatch-score of 0, but similar results can be derived for general alignment scores. We employed this scoring matrix for our work as it reflects our limited knowledge of nucleotide frequencies in CRMs.

Lemma 18 *Let Σ be an alphabet. Let $s,t,u,v \in \Sigma^+$ such that $s = \alpha x, u = x\beta, t = \gamma y, v = y\delta$ for some $\alpha,\beta,\gamma,\delta \in \Sigma^*$. Let $N(\cdot,\cdot)$ denote the optimal alignment score of two strings when using a match-score of 1, a mismatch score of 0, and a gap-penalty $-\frac{1}{Z}$, for $Z \in]0,\infty[$.*

1. $N(u,v) \geq N(s,t) - (\max\{|\alpha|,|\gamma|\} + \frac{||\alpha|-|\gamma||+||\beta|-|\delta||}{Z})$
2. $N(u,v) \leq N(s,t) + (\max\{|\beta|,|\delta|\} + \frac{||\alpha|-|\gamma||+||\beta|-|\delta||}{Z})$
3. $N(s,v) \geq N(u,t) - (|\beta| + |\gamma| + \frac{|\alpha|+|\beta|+|\gamma|+|\delta|}{Z})$
4. $N(s,v) \leq N(u,t) + (|\alpha| + |\delta| + \frac{|\alpha|+|\beta|+|\gamma|+|\delta|}{Z})$

Proof. 1.) Let $p,q \in (\Sigma \cup \{-\})^+$ be an optimal alignment for s and t. For any given string $\theta \in (\Sigma \cup \{-\})^+$ let $T(\theta)$ denote the string that is derived from θ by removing all gap-characters. Let $p',q',\epsilon,\omega \in (\Sigma \cup \{-\})^+$ such that $p = \epsilon p', q = \omega q', x = T(p'), y = T(q')$, and $|\epsilon|$ as well as $|\omega|$ maximal.
Case 1: $|\beta| \geq |\delta|, |p'| \geq |q'|$
Let $r_1 = p'\beta$ and $r_2 = (-)^{|p'|-|q'|}q'(-)^{|\beta|-|\delta|}$. Obviously $|r_1| = |r_2|$ holds. Let $S(r_1,r_2)$ denote the alignment score of r_1 and r_2. Then $N(u,v) \geq S(r_1,r_2)$, since $u = T(r_1)$ and $v = T(r_2)$. As in the alignment (r_1,r_2) q' is aligned to the same suffix of p' as in alignment (p,q), we have:

$$S(r_1,r_2) \geq N(s,t) - \max\{|\alpha|,|\gamma|\} - \frac{||\alpha|-|\gamma||}{Z} - \frac{|\beta|-|\delta|}{Z}$$

Here $\max\{|\alpha|,|\gamma|\}$ is an upper bound for the number of matches that are lost by removing ϵ and ω. $\frac{||\alpha|-|\gamma||}{Z}$ is an upper bound for the number of additional gaps

at the beginning of r_2.

Case 2: $|\delta| \geq |\beta|, |p'| \geq |q'|$

For $r_1 = p'\beta(-)^{|\delta|-|\beta|}$ and $r_2 = (-)^{|p'|-|q'|}q'\delta$ the following inequality holds:

$$S(r_1, r_2) \geq N(s,t) - \max\{|\alpha|, |\gamma|\} - \frac{||\alpha| - |\gamma||}{Z} - \frac{|\delta| - |\beta|}{Z}$$

Case 3: $|\beta| \geq |\delta|, |p'| < |q'|$

For $r_1 = (-)^{|q'|-|p'|}p'\beta$ and $r_2 = q'\delta(-)^{|\beta|-|\delta|}$ the same inequality as in Case 1 holds (see Equation A.1).

Case 4: $|\beta| < |\delta|, |p' < |q'|$

For $r_1 = (-)^{|q'|-|p'|}p'\beta(-)^{|\delta|-|\beta|}$ and $r_2 = q'\delta$ the same inequality as in Case 3 holds (see Equation A.1).

Hence the claim holds in all cases.

2.) We use $R : \Sigma^* \rightarrow \Sigma^*$ to denote the reversion function for strings. For any $x_1, x_2 \in \Sigma^*$ $N(x_1, x_2) = N(R(x_1), R(x_2)$ holds, since alignment scores are invariant under string reversions. Therefore, we have

$$\begin{aligned}
N(u,v) &= N(R(u), R(v)) \\
&= N(R(\beta)R(x), R(\delta)R(y)) \\
&\leq N(R(x)R(\alpha), R(y)R(\gamma)) + \max\{|R(\beta)|, |R(\delta)|\} + \\
&\quad \frac{||R(\beta)| - |R(\delta)|| + ||R(\alpha)| - |R(\gamma)||}{Z} \\
&= N(s,t) + \max\{|\beta|, |\delta|\} + \frac{||\beta| - |\delta|| + ||\alpha| - |\gamma||}{Z}
\end{aligned}$$

3.) This statement can be seen by first applying statement 2 and then statement 1 both of which are already proven:

$$\begin{aligned}
N(s,v) &= N(\alpha x, v) \\
&\geq N(x\beta, v) - |\beta| - \frac{|\alpha| + |\beta|}{Z} \\
&= N(u, y\delta) - |\beta| - \frac{|\alpha| + |\beta|}{Z} \\
&\geq N(u, \gamma y) - |\gamma| - \frac{|\gamma| + |\delta|}{Z} - \beta - \frac{|\alpha| + |\beta|}{Z} \\
&= N(u,t) - (|\beta| + |\gamma| + \frac{|\alpha| + |\beta| + |\gamma| + |\delta|}{Z})
\end{aligned}$$

Sequence Lengths	$w = 5$	$w = 4$	$w = 3$	$w = 2$	$w = 1$
$15,000 \times 17,500$	308	365	431	543	881
$15,000 \times 23,500$	414	483	571	722	1155
$48,000 \times 48,000$	2508	2891	3316	4069	6368
$80,500 \times 93,000$	7100	8075	9212	11334	18338
$120,000 \times 120,000$	14802	16915	19292	23555	37305

Table A.1: Effect of Step Width on CPU time

4.) Using statement 3 we conclude:

$$\begin{aligned} N(s,v) &= N(R(x)R(\alpha), R(\delta)R(y)) \\ &\leq N(R(\beta)R(x), R(y)R(\gamma)) + |R(\alpha)| + |R(\delta)| \\ &\quad + \frac{|R(\alpha)| + |R(\beta)|}{Z} + \frac{|R(\gamma)| + |R(\delta)|}{Z} \\ &= N(u,t) + |\alpha| + |\delta| + \frac{|\alpha| + |\beta| + |\gamma| + |\delta|}{Z} \end{aligned}$$

■

A.5 Performance

The asymptotic order of Algorithm 17 is still in $O(|s||t|l^2)$ for a constant step width w, but it makes a substantial improvement over Algorithm 16 in practice. Table A.1 shows the effect of step width on CPU time (in seconds) and provides examples of running time on real biological sequences using a 3GHz Linux machine.

Computing conservation profiles increases the CPU time as potential updates of the profiles have to be considered in Step 5C, resulting in fewer omissions of alignments. However, these increases are modest as the profiles will quickly reach values near the cut-off S during the execution of the algorithm.

Glossary

AP Anterior Posterior, page 11

BiFa tool Binding Factor analysis tool, page 78

BLAST Basic Local Alignment Search Tool, page 5

CAD CRM Activity Database, page 16

ChIP-chip Chromatin Immunoprecipitation combined with microarray, page 15

CRM cis-Regulatory Module, page 3

IID Independent and Identically Distributed, page 28

MM Markov Model, page 28

PGP Pattern Generating Potential, page 20

PSSM Position Specific Scoring Matrix, page 3

PWM Position Weight Matrix, page 3

ReMo GUI Regulatory Module Graphical User Interface, page 78

RRS Regulatory Region Scoring Model, page 6

TF Transcription Factor, page 2

TFBS Transcription Factor Binding Site, page 4

TSS Transcription Starting Site, page 3

Bibliography

[1] S. Aerts, P. Van Loo, G. Thijs, Y. Moreau, and B. De Moor. Computational detection of cis-regulatory modules. *Bioinformatics*, 19 Suppl 2:ii5–14, 2003. 6, 26

[2] A. Aggarwal, M.M. Klawe, M. Shlomo, P. Shor, and R. Wilber. Geometric applications of a matrix-searching algorithm. *Algorithmica*, 2:195–208, 1987. 115

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990. 6

[4] A. N. Arslan, O. Egecioglu, and P. A. Pevzner. A new approach to sequence comparison: normalized sequence alignment. *Bioinformatics*, 17(4):327–37, 2001. 111

[5] E. Berezikov, V. Guryev, and E. Cuppen. Conreal web server: identification and visualization of conserved transcription factor binding sites. *Nucleic Acids Res*, 33(Web Server issue):W447–50, 2005. 109

[6] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: applications. *Curr Opin Genet Dev*, 15(2):125–35, 2005. 12

[7] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev*, 15(2):116–24, 2005. 11, 12, 13, 15

[8] B. E. Blaisdell. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc Natl Acad Sci U S A*, 83(14):5155–9, 1986. 6, 26

[9] C. D. Brown, D. S. Johnson, and A. Sidow. Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science*, 317(5844):1557–60, 2007.

[10] R. A. Cameron and E. H. Davidson. Flexibility of transcription factor target site position in conserved cis-regulatory modules. *Dev Biol*, 336(1):122–35, 2009. 11

[11] M. Crochemore, G.M. Landau, and M. Ziv-Ukelson. A subquadratic sequence alignment algorithm for unrestricted scoring matrices. *SIAM J. Comput.*, 32:1654–1673, 2003. 115

[12] P. D'Haeseleer. What are dna sequence motifs? *Nat Biotechnol*, 24(4):423–5, —2006—. 4

[13] C. Dieterich, H. Wang, K. Rateitschak, H. Luz, and M. Vingron. Corg: a database for comparative regulatory genomics. *Nucleic Acids Res*, 31(1):55–7, 2003. 110

[14] M. Djordjevic, A. M. Sengupta, and B. I. Shraiman. A biophysical approach to transcription factor binding site discovery. *Genome Res*, 13(11):2381–90, 2003. 12

[15] Thomas Down. *Computational localization of promoters and transcirption start sites in mammalian genomes.* PhD thesis, Cambridge University, 2003.

[16] *Drosophila* 12 Genomes Consortium et. al. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450(7167):203–18, 2007. 79

[17] B. C. Foat, A. V. Morozov, and H. J. Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by matrixreduce. *Bioinformatics*, 22(14):e141–9, 2006. 12

[18] B. C. Foat, A. V. Morozov, and H. J. Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by matrixreduce. *Bioinformatics*, 22(14):e141–9, 2006.

[19] R. Fuchs. From sequence to biology: the impact on bioinformatics. *Bioinformatics*, 18(4):505–6, 2002.

[20] J. Gertz, E. D. Siggia, and B. A. Cohen. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, 457(7226):215–8, 2009. 2, 6, 12, 13

[21] R. Gordan, L. Narlikar, and A. J. Hartemink. Finding regulatory DNA motifs using alignment-free evolutionary conservation information. *Nucleic Acids Res*, 38(6):e90, 2010. 8

[22] S. Gupta, J. Dennis, R. E. Thurman, R. Kingston, J. A. Stamatoyannopoulos, and W. S. Noble. Predicting human nucleosome occupancy from primary sequence. *PLoS Comput Biol*, 4(8):e1000134, 2008. 12, 13

[23] D Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer science and Computational Biology.* Cambridge University Press., 1997.

[24] E. E. Hare, B. K. Peterson, V. N. Iyer, R. Meier, and M. B. Eisen. Sepsid even-skipped enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. *PLoS Genet*, 4(6):e1000106, 2008. 6, 10, 105

[25] N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford, and B. Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, 39(3):311–8, 2007. 3

[26] G. Z. Hertz and G. D. Stormo. Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–77, 1999. 5

[27] G. S. Jefferis and T. Hummel. Wiring specificity in the olfactory system. *Semin Cell Dev Biol*, 17(1):50–65, 2006. 81

[28] G. S. Jefferis, E. C. Marin, T. Komiyama, H. Zhu, T. Chihara, D. Berdnik, and L. Luo. Development of wiring specificity of the *Drosophila* olfactory system. *Chem Senses*, 30 Suppl 1:i94, 2005.

[29] G. S. Jefferis, E. C. Marin, R. F. Stocker, and L. Luo. Target neuron prespecification in the olfactory map of *Drosophila*. *Nature*, 414(6860):204–8, 2001. 81

[30] D. M. Jeziorska, K. W. Jordan, and K. W. Vance. A systems biology approach to understanding cis-regulatory module function. *Semin Cell Dev Biol*, 2009.

[31] M. R. Kantorovitz, G. E. Robinson, and S. Sinha. A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, 23(13):i249–55, 2007. 6, 26, 30, 62, 63, 76

[32] M. Kazemian, C. Blatti, A. Richards, M. McCutchan, N. Wakabayashi-Ito, A. S. Hammonds, S. E. Celniker, S. Kumar, S. A. Wolfe, M. H. Brodsky, and S. Sinha. Quantitative analysis of the *Drosophila* segmentation regulatory network using pattern generating potentials. *PLoS Biol*, 8(8), 2010. 11, 20, 22, 24, 25, 61

[33] A. E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender. Match: A tool for searching transcription factor binding sites in dna sequences. *Nucleic Acids Res*, 31(13):3576–9, 2003. 5

[34] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at ucsc. *Genome Res*, 12(6):996–1006, 2002. 109

[35] T. Komiyama, W. A. Johnson, L. Luo, and G. S. Jefferis. From lineage to wiring specificity. pou domain transcription factors control precise connections of *Drosophila* olfactory projection neurons. *Cell*, 112(2):157–67, 2003. 79

[36] T. Komiyama and L. Luo. Development of wiring specificity in the olfactory system. *Curr Opin Neurobiol*, 16(1):67–73, 2006.

[37] T. Komiyama and L. Luo. Intrinsic control of precise dendritic targeting by an ensemble of transcription factors. *Curr Biol*, 17(3):278–85, 2007. 79

[38] H. Koohy, N. P. Dyer, J. E. Reid, G. Koentges, and S. Ott. An alignment-free model for comparison of regulatory sequences. *Bioinformatics*, 26(19):2391–7, 2010. 2, 8, 62, 87

[39] S. C. Kou, Q. Zhou, and W. H. Wong. Equi-energy sampler with applications in statistical inference and statistical mechanics. *Ann Statist*, 34(4):1581–619, 2006.

[40] G. Leung and M. B. Eisen. Identifying cis-regulatory sequences by word profile similarity. *PLoS One*, 4(9):e6901, 2009. 26, 32, 35

[41] M. Li, B. Ma, D. Kisman, and J. Tromp. Patternhunter ii: highly sensitive and fast homology search. *J Bioinform Comput Biol*, 2(3):417–39, 2004. 110

[42] R. A. Lippert, H. Huang, and M. S. Waterman. Distributional regimes for the number of k-word matches between two random sequences. *Proc Natl Acad Sci U S A*, 99(22):13980–9, 2002. 30

[43] Y. H. Loh, Q. Wu, J. L. Chew, V. B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, K. Y. Wong, K. W. Sung, C. W. Lee, X. D. Zhao, K. P. Chiu, L. Lipovich, V. A. Kuznetsov, P. Robson, L. W. Stanton, C. L. Wei, Y. Ruan, B. Lim, and H. H. Ng. The oct4 and nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet*, 38(4):431–40, 2006.

[44] P. V. Loo and P. Marynen. Computational methods for the detection of cis-regulatory modules. *Brief Bioinform*, 2009. 7

[45] M. Z. Ludwig, A. Palsson, E. Alekseeva, C. M. Bergman, J. Nathan, and M. Kreitman. Functional evolution of a cis-regulatory module. *PLoS Biol*, 3(4):e93, 2005. 10, 105

[46] W.J. Masek and M.S. Paterson. A faster algorithm computing string edit distances. *Journal of Computer and System Sciences*, 20:18–31, 1980. 115

[47] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31(1):374–8, 2003. 5, 63

[48] A. V. Morozov and E. D. Siggia. Connecting protein structure with predictions of regulatory sites. *Proc Natl Acad Sci U S A*, 104(17):7068–73, 2007.

[49] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–53, 1970. 114

[50] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–53, 1970.

[51] S. Nobrega-Pereira and O. Marin. Transcriptional control of neuronal migration in the developing mouse brain. *Cereb Cortex*, 19 Suppl 1:i107–13, 2009. 80

[52] A. Ochoa-Espinosa, G. Yucel, L. Kaplan, A. Pare, N. Pura, A. Oberstein, D. Papatsenko, and S. Small. The role of binding site cluster strength in bicoid-dependent patterning in *Drosophila*. *Proc Natl Acad Sci U S A*, 102(14):4960–5, 2005. 63, 68, 70

[53] M. Pelandakis and M. Solignac. Molecular phylogeny of *Drosophila* based on ribosomal rna sequences. *J Mol Evol*, 37(5):525–43, 1993. 79

[54] L. A. Pennacchio and E. M. Rubin. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet*, 2(2):100–9, 2001. 3, 6

[55] N. Pierstorff, C. M. Bergman, and T. Wiehe. Identifying cis-regulatory modules by combining comparative and compositional analysis of dna. *Bioinformatics*, 22(23):2858–64, 2006.

[56] F. Polleux, G. Ince-Dunn, and A. Ghosh. Transcriptional regulation of vertebrate axon guidance and synapse formation. *Nat Rev Neurosci*, 8(5):331–40, 2007. 80

[57] J. Rister and C. Desplan. Deciphering the genome's regulatory code: the many languages of dna. *Bioessays*, 32(5):381–4, 2010. 19

[58] J. Rister and C. Desplan. Deciphering the genome's regulatory code: the many languages of dna. *Bioessays*, 32(5):381–4, 2010.

[59] H. G. Roider, A. Kanhere, T. Manke, and M. Vingron. Predicting transcription factor affinities to dna from a biophysical model. *Bioinformatics*, 23(2), 2007. 12

[60] A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–4, 2004. 5

[61] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–100, 1990. 4

[62] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature*, 451(7178):535–40, 2008. 2, 6, 11, 12, 14, 42, 43, 49, 75, 87

[63] E. Segal and J. Widom. From dna sequence to transcriptional behaviour: a quantitative approach. *Nat Rev Genet*, 10(7):443–56, 2009. 2, 6, 11, 15, 17

[64] Eilon Sharon and Eran Segal. A feature-based approach to modeling protein-dna interactions. In *RECOMB*, pages 77–91, 2007.

[65] M. Simpson-Brose, J. Treisman, and C. Desplan. Synergy between the hunchback and bicoid morphogens is required for anterior patterning in *Drosophila*. *Cell*, 78(5):855–65, 1994. 60

[66] S. Sinha, A. S. Adler, Y. Field, H. Y. Chang, and E. Segal. Systematic functional characterization of cis-regulatory motifs in human core promoters. *Genome Res*, 18(3):477–88, 2008. 12, 13

[67] S. Sinha, E. van Nimwegen, and E. D. Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19 Suppl 1:i292–301, 2003. 21

[68] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–7, 1981. 111

[69] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–7, 1981. Journal Article England.

[70] M. L. Spletter, J. Liu, H. Su, E. Giniger, T. Komiyama, S. Quake, and L. Luo. Lola regulates *Drosophila* olfactory projection neuron identity and targeting specificity. *Neural Dev*, 2:14, 2007. 79

[71] Maria Lynn Spletter. *Cell identity and wiring specificity in the Drosophila olfactory system.* PhD thesis, Stanford University, 2009. 9, 81, 82, 83, 89, 100

[72] A. Stark. Learning the transcriptional regulatory code. *Mol Syst Biol*, 5:329, 2009. 18, 19

[73] E. A. Stone and A. Sidow. Constructing a meaningful evolutionary average at the phylogenetic center of mass. *BMC Bioinformatics*, 8:222, 2007. 21

[74] A. Tanay. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res*, 16(8):962–72, 2006. 12

[75] A. Tanay. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res*, 16(8):962–72, 2006.

[76] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–44, 2005.

[77] J. van Helden. Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics*, 20(3):399–406, 2004. 6, 26

[78] S. Vinga and J. Almeida. Alignment-free sequence comparison - a review. *Bioinformatics*, 19(4):513–23, 2003. 10, 12

[79] Y. X. Zhang, K. Perry, V. A. Vinci, K. Powell, W. P. Stemmer, and S. B. del Cardayre. Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature*, 415(6872):644–6, 2002.

[80] R. P. Zinzen, C. Girardot, J. Gagneur, M. Braun, and E. E. Furlong. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269):65–70, 2009. 2, 11, 17, 19

Declaration

I herewith declare that the research submitted in this thesis was conducted by myself under the supervision of Dr. Sascha Ott at the Warwick Systems Biology Centre and Prof. Georgy Koentges at the Department of Biological Sciences of the University of Warwick.

No parts of this work have previously been submitted to be considered for a degree or other qualifications. All sources of information have been specifically acknowledged in form of references.