In silico Analysis of Polycomb/Trithorax Response Elements in Drosophila and Mammals Based on DynScan: An Alignment Independent Sensitivity Increasing Framework for Cross-species Prediction of Regulatory Elements

März 2008

Dissertation zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.) der Technischen
Fakultät der Universität Bielefeld
vorgelegt von

Arne Hauenschild

Contents

List of Figures

List of Tables

Acknowledgements

First I want to thank my supervisors. Marc Rehmsmeier for constant input, ideas and suggestions in the last few years that made this work possible in the first place and for always offering guidance if requested. Leonie Ringrose for all biological input in form of explanations and showing the big pictures, and especially for all the experiments and hard work on the manuscript, which I am sure will get the attention it deserves. Robert Giegerich for consulting in strategic questions. My colleague Thomas Fiedler for providing us with the jPREdictor and for all the helpful discussions. Ding Jia for providing additional ideas and nice working climate. My friend Julia for sharing good as well as bad experiences during the long work on our theses. My girlfriend Jennifer for showing me life outside university. My parents, without their support neither the diploma nor this thesis would have been possible.

1 Introduction

With ongoing progresses in genome sequencing, more and more sequence data is becoming available. The task after the assembly is to identify the functional elements inside genomes. Different techniques are established to find or predict genes, in bacteria and viruses tools like Glimmer [1] make use of prior knowledge of CG content or codon usage of the analyzed species. In eukaryotes, where introns and exons have to be identified, conservation plays a bigger role, implemented in tools like Genescan/Twinscan [2] or Genewise [3]. The results of predictions as well as experimentally gained and verified gene positions and structures are stored in annotation databases, like Ensembl [4], RefSeq [5] or species specific like FlyBase [6] and WormBase.

However, even if all genes can be determined and stored in databases, knowledge about gene positions and protein functions is still not sufficient to explain all observed phenotypic variations or to explain how embryonic development works. To do this, it is necessary to understand which mechanisms are involved in the regulation of gene expression. What determines which genes are activated or repressed in different cells at different time points? How can phenotypic divergence be explained between two closely related species that use the same gene base? Functional genomics is meant as a general term for the field of molecular biology that tries to shed light on function and regulation of genes.

Regulation can take place at almost any level of the pathway from the DNA to the transcription by polymerase, the splicing step in eukaryotes, the ribosomal translation into the completed protein and finally to the function of the protein. Regulatory effects start with direct DNA modification, either on nucleotide level by chemical modifications like DNA methylation, or on structure level like chromatin remodeling and histone modifications. All these effects can force the DNA into another form, which prevents the transcription start site to be accessed by the polymerase.

The next regulatory effects take place on the transcription level. Common regulatory elements on this level are promoters, which contain elements recognized by the polymerase. The promoters within one species share common basic elements, which work as a general method to attract the polymerase and are the standard elements that enable transcription. In case of eukaryotes, four different elements are characterized. The first one is the Inr element (a pyrimidin rich region around the start nucleotide), present in most but not all promoters. Secondly, the TATA box is located -30 bp upstream of the transcription start site and can be found in most genes, except for housekeeping genes. As a third element, many promoters additionally contain down stream promoter elements (DPE), consisting of a gene type specific sequence. Finally, regions of GC rich boxes are located in many promoters and act as regulatory elements as well. Besides those basic promoter elements, special regulatory elements exist in cell specific and development specific genes, located in close distance to the transcription start site as well as several kilobases away.

Elements that are outside the promoter but increase the transcriptional activity of a gene are called enhancers and belong to the class of *cis*-regulatory elements. The latin term "cis" can be approximately translated as "on this side" and is used to mark regulatory elements that are located in the vicinity of their regulated gene. Remember that a gene's vicinity can span several thousand bases. Further types of those elements are insulators, regions that separate regulatory elements. This is important if two adjacent genes differ in their transcriptional status and are regulated by different classes of elements. Another *cis*-regulatory class are Polycomb/Trithorax Response elements (PRE/TRE), initially discovered in *Drosophila melanogaster* [7]. Once gene expression patterns of developmental genes are set in the early developmental stages by activators and repressors, the transcriptional decision is maintained through cell division cycles. PRE/TREs are key players in the epigenetic system. They allow the inheritance of gene transcription patterns from one cell generation to the next without involvement of DNA mutations.

A common attribute of *cis*-regulatory elements is that they are targets of DNA binding proteins, called transcription factors. Within a *cis*-regulatory element multiple transcription factor binding sites can be contained. In order to understand how the regulation of a specific gene works on transcriptional level, the position and function of the corresponding *cis*-regulatory elements are of huge interest. Based on known transcription factor binding sites, different methods exist to identify the element's position. In vitro, chromatin immunoprecipitation (ChIP) is one way to detect positions of protein binding. Methods working in silico search for representations (motifs) of binding sites. The prediction of protein binding positions is usually combined with a statistic to choose a certain level of specificity or sensitivity. If single motif occurrence is not sufficient to predict the location of a *cis*-regulatory element, but different motif combinations have to be taken into account, the statistical problem grows bigger.

We present an enhancement for existing prediction tools of *cis*-regulatory elements that uses a comparative approach to input results gained in one species as prior knowledge into another, named *DynScan*. The idea is first mentioned in [8], where it was tested for PRE/TRE predictions in two Drosophila species. However is that study, instead of being implemented in a generic way, the analyis was mainly done manually. In this work, the method is generalized in a way that it can be applied to any bioinformatics tool that is based on scoring continuous sequences in at least two related species and applying cut-offs to ensure some statistical relevance. Algorithms focusing on motif matching or prediction of *cis*-regulatory elements usually fall into this category. Increasing a prediction result by introducing comparative genomics is not a novel approach, the first methods are about 20 years old, originally used in 1988 on primates [9]. However, in contrast to regular phylogenetic footprinting, which relies completely on sequence conservation of the element that is to be found, the method described here works independent of conservation, but gives higher rewards the closer an element in two species is located to the respective orthologous site. Furthermore, our method is implemented in form of a framework, allowing the user to arbitrarily choose the underlying scoring algorithm.

As an application we choose the jPREdictor software [10], which we use to predict PRE/TREs in multiple Drosophila species. The choice of that algorithm is made because the software PREdictor [11] already demonstrated in 2003 the general possibility of computational PRE/TRE prediction. Furthermore the analysis placed emphasis on specificity, i.e. the goal was to be confident in the predicted results, not to predict as many elements as possible. This gives an excellent study case for our *DynScan* method, which is aimed to increase sensitivity while keeping specificity. The data presented in this thesis shows that *DynScan* predicts novel elements in all of the Drosophila species we used. Furthermore, the specificity is not only statistically determined, but biological experiments done in collaboration in Leonie Ringrose's lab at the "Institue of Molecular Biotechnology GmbH Vienna, Austria" show high accuracy of the additionally predicted results.

Another part of this work is the prediction of potential novel motifs that are part of *cis*-regulatory elements. The number of available motif prediction tools keeps growing, although many implementations already exist based on enumerative approaches of candidates as well as probabilistic models that try to sample most promising signals from a noisy background. As shown in a recent comparison [12], none of those shows sensitivity above 10% in different test cases. Instead of trying to develop yet another general purpose motif prediction tool, we introduce different ways to combine existing complementary methods and furthermore to automatically validate prediction results based on existing real life data.

With ongoing progresses in experimental methods such as large scale chromatin immunoprecipitation (ChIP on chip), which requires less financial effort than it did a few years ago, motif prediction tools can benefit from larger sets of confirmed sequences that contain the unknown binding sites, and negative sets that can serve as background. We demonstrate an enhancement of motif enumeration tools that directly makes use of such data sets. In addition, we developed a prediction pipeline that combines existing methods on large sets of experimental data to increase prediction accuracy. Furthermore, a motif evaluation algorithm is introduced to rate each motif in relation to the biological data.

In recent studies genome-wide experimental data for potential PRE/TRE related sequences in human and mouse have been published [13, 14], but no functional motifs are known so far. We chose the data as an application for our motif prediction pipeline and evaluating algorithm.

Structure of this work

In Chapter 2 background information are provided, that explain the basis of our work, seperated into the bioinformatic and biological aspects. First the general approach of phylogenetic footprinting and its limits are explained. Afterwards, I summarize the different techniques for prediction of motifs as representations of transcription factor binding sites and give an overview of the most prominent tools. Subsequently, the biological background is explained, which is necessary for the understanding of the application of our bioinformatics method. I explain how a special regulatory element, the Polycomb/Trithorax Response elements work and which methods can be used to find and to validate those elements in vivo and in silico. In Chapter 3, I describe our novel method called *DynScan*, which uses location rather than sequence conservation to increase statistical sensitivity without losing specificity. The purpose as well as the implementation details are provided. Furthermore a method is introduced to rate a sequence's potential to gain specific regulatory functionality by minor mutations of transcription factor binding sites.

The second method chapter (Chapter 4) covers various novel approaches to use motif predictions in order to separate sequences that contain regulatory elements from a set of background sequences. An algorithm is described that allows an evaluation of the contribution of predicted motifs in the prediction of specific elements of interest. The described methods are applied on two related but differing tasks:

First, the *DynScan* method is used to increase the sensitivity of the prediction of Polycomb/Trithorax Response elements in Drosophila, as shown in Chapter 5. Based on the jPREdictor, multiple Drosophila species are scored and *DynScan* tries to predict new hits in each species based on results in other species. The results are presented in combination with experimental data supplied by our collaboration partner, and the evolutionary study is done on some of the experimental data. The PRE/TRE prediction is compared to three biological studies and the *DynScan* benefit is shown in that context.

Second, we use the motif prediction methods in order to identify potential novel motifs involved in a elements similar to fly PREs/TREs in mammalian data (Chapter 6). Based on recently published new ChIP data, the motif search is performed in different ways, and the resulting motifs are put into the validating process. Furthermore, additional statistical studies and their results on possible overrepresentations of dinucleotide distributions are included. The thesis concludes with a discussion of the described results.

2 Background

2.1 Bioinformatics

2.1.1 Phylogenetic footprinting

A common task in bioinformatics is to extend knowledge gained about some element in a single species to other species, in order to learn more about the analyzed element of interest. For example, known coding regions inside one genome are used to detect coding regions in other species because those regions are expected to be conserved, if not on DNA level then at least on the protein level. Gene prediction tools like Genescan/Twinscan or Genewise [2, 3] use conservation as a main criterion. When it comes to RNA, the prediction of miRNAs as well as their targets is usually based on strong conservation of the targeted UTRs and the miRNA's sequence. The classical algorithms Miranda [15] or mirScan [16] make use of this approach.

Following this idea, conservation can be used to detect regulatory elements without knowing their location in any of the observed species, but by scanning for positions showing higher conservation than adjacent regions. In this case, the conservation can be considered a signal for a selective pressure, and thus for a biological function of those regions. Because these spots of high conservations can be thought of as a footprint left by the phylogeny during evolution, the method was called "Phylogenetic Footprinting", first introduced in 1988 for the predicton of *cis*-regulatory elements involved in the expression of embryonic A and B globulin in primates [9]. Since then, the technique has been implemented into different algorithms like Footprinter [17] and is mainly used for the prediction of transcription factor binding sites in non-coding DNA regions.

Usually, a gene of interest is chosen in a set of different species and the non-coding regions upstream of the orthologous promoters are searched for spots of significantly higher sequence conservation. While these steps are shared by most approaches, the differences lie only in the calculation of the significance value. All approaches are affected by the same fundamental difficulties. A comparison of different applications of phylogenetic footprinting [18] came to the conclusion that the identification of orthologous genes and, in the next step, promoter regions can be difficult in practice due to often incomplete annotations. Especially if distant species are involved, the alignment of the promoter regions can be another difficult and error prone task. As a consequence, the method leads to good results in only some cases, highly depending on the quality of the data. In most cases, phylogenetic footprinting is weakened by the available annotations. Furthermore, some *cis*-regulatory elements like enhancers can consist of multiple clusters of transcription factor binding sites. If the sites' order is not conserved, the whole element could be missed by phylogenetic footprinting. Such motifs turnovers are widely described in literature [19, 20, 21]. Furthermore, a bioinformatics study by Emberly et al. [22] revealed that clusters of functional motifs inside enhancers are not neccessarily located inside conserved blocks, even if the clusters are preserved between species [23].

In general, methods for the prediction of any kinds of elements that rely completely on conservation will miss elements that are not conserved above background. Nevertheless, elements that are present in different species might at least occur within the same locus, therefore it sounds reasonable to reward if position conservation is present, instead of relying on it completely. A method that concentrates on finding elements occurring at orthologous positions leaves out all situations in which functionally analogous elements exist within the same locus, but without showing significant conservation or are even not sequence conserved at all. This observation is independent of the focus of the method, whether single motifs are searched or if complete functional elements like enhancers are to be detected.

2.1.2 Motif prediction

The task to find potential transcription factor binding sites in a list of sequences that share common properties has been addressed by multiple bioinformatic tools based on various approaches in the last decade. Such properties could be that the sequences are upstream sequences of co-regulated genes. Especially growing biological capabilities like genome-wide ChIP experiments that collect large amounts of data require the use of such tools. In general, methods to find the possibly degenerated functional subsequences in a set of related sequences can roughly be separated into two categories.

1. Enumerative approaches

2. Probabilistic approaches

Both methods will be described and examples of implementations will be shown.

Enumeration

The straight-forward way is to exhaustively enumerate all possible k-words (for a given k) and search for most overrepresented ones. In order to define overrepresentation, a background model is required, such as the number of occurrences in a second set of negative sequences or a probabilistic model based on Markov Chains.

Some transcription factor binding sites can be represented in form of k-word motifs, if a strongly conserved core binding site is present. In other cases, the motif representation has to be less strict, accepting different nucleotides to match at some positions. In this case, a pure k-word approach will reach its limits and the possibility to handle motifs in a degenerated form is required.

One possibility is to use clustering methods to combine related k-words into a single model that describes a transcription factor binding site. Such clustering techniques are described in the next section. Another possibility to introduce degeneration into enumerative motif prediction is to extend the alphabet from nucleotides [A,C,G,T] to the IUPAC code[1]. To keep the search space manageable (4^k vs. 14^k; e.g. $k = 10$: ˜1 m vs. ˜289 bn), usually only a small number of positions are taken from the IUPAC code. An implementation is the tool YMF [24], depending on the value k, two or three positions can be degenerated. Alternatively, a motif can be seen as a consensus sequence of a motif matrix, containing positions of allowed mismatches. In the implementation of Weeder [25], the number of mismatches depends on the word length. By default, for words of length 6 one mismatch is allowed, respectively two in 8 and three in 10-words. The input sequences are transformed into a suffix tree, in which each of the possible words can be searched in $O(k)$. The matching words are combined into a common matrix. Furthermore, in a second step the matrix is taken to score the found words. Those receiving the highest scores are used to build up a second matrix that is reported as the final prediction result. Enumeration approaches are exact in the sense that exhaustively the entire search space is covered. On the other hand, the search space may cover only parts of the real world possibilities to achieve a usable running time and space consumption behavior.

Probabilistic approaches

The representation of motifs commonly used is a position specific scoring (PSSM), weighting (PWM), or probability matrix (PSPM). Instead of determining which nucleotides are allowed (and therefore which ones are forbidden) at each position, relations are provided, so as an exmaple it can be expressed that in all known sequences

[1] The nucleotide IUPAC code contains symbols for each possible subset of [A,C,G,T]. For example B means "not A" [C,G,T], or Y means pyrimidin [C,T].

of a motif an 'A' occurs four times as often as 'C' at a specific position. The subject of a matrix based motif prediction is to optimize a PWM and the corresponding binding probabilities of a binding site. Usually, the optimization is either done by deterministic optimization or by probabilistic optimization. A common method for deterministic optimization is Expectation Maximization, shown briefly in the following section. In case of probabilistic optimization for motif prediction, Gibbs sampling is the most prominent technique. Because Gibbs sampling is also used by different motif clustering algorithms, the mathmatical basics will be described in the following.

Weighting

A matrix is a model (M) to describe the observed data (D), i.e. the potential binding sites given the sequences. The probability of the model cannot be calculated directly. Instead, following Bayes theorem, the likelihood of the model is used: $L(M|D) = P(D|M)$. The probability that a sequence is generated by the matrix in relation to the probability that it has been generated by the null-model is the weight of that sequence for a specific matrix.

$$W(D|M) = \log \frac{P(D|M)}{P(D|M_0)}$$

$W(S|M)$ is the log-likelihood of a sequence S given a matrix. Accordingly, the log-likelihood of a matrix is therefore:

$$W(M) = \sum_{i \in D} \log \frac{P(D_i|M)}{P(D_i|M_0)}$$

Expectation Maximization (EM)

The idea is to build a preliminary PWM and use it to search the sequences in order to find new matching elements. As soon as additional sites are detected, the matrix is updated (Figure 2.1).

In the first step the matrix is initialized with a single (randomly chosen) k-word, combined with background sequences or pseudoocounts to allow matching of words similar to the chosen one as well. Without introducing additional occurrences into the matrix, the probability of every other word would be zero.

Each k-word in the sequence is weighted according to the matrix and a provided null-model, as described above. By EM a weighted average is chosen from these weights, and the matrix is updated. By repetition of the steps the maximum log likelihood of the model can be reached.

One typical implementation of this approach is MEME [27]. In order to avoid

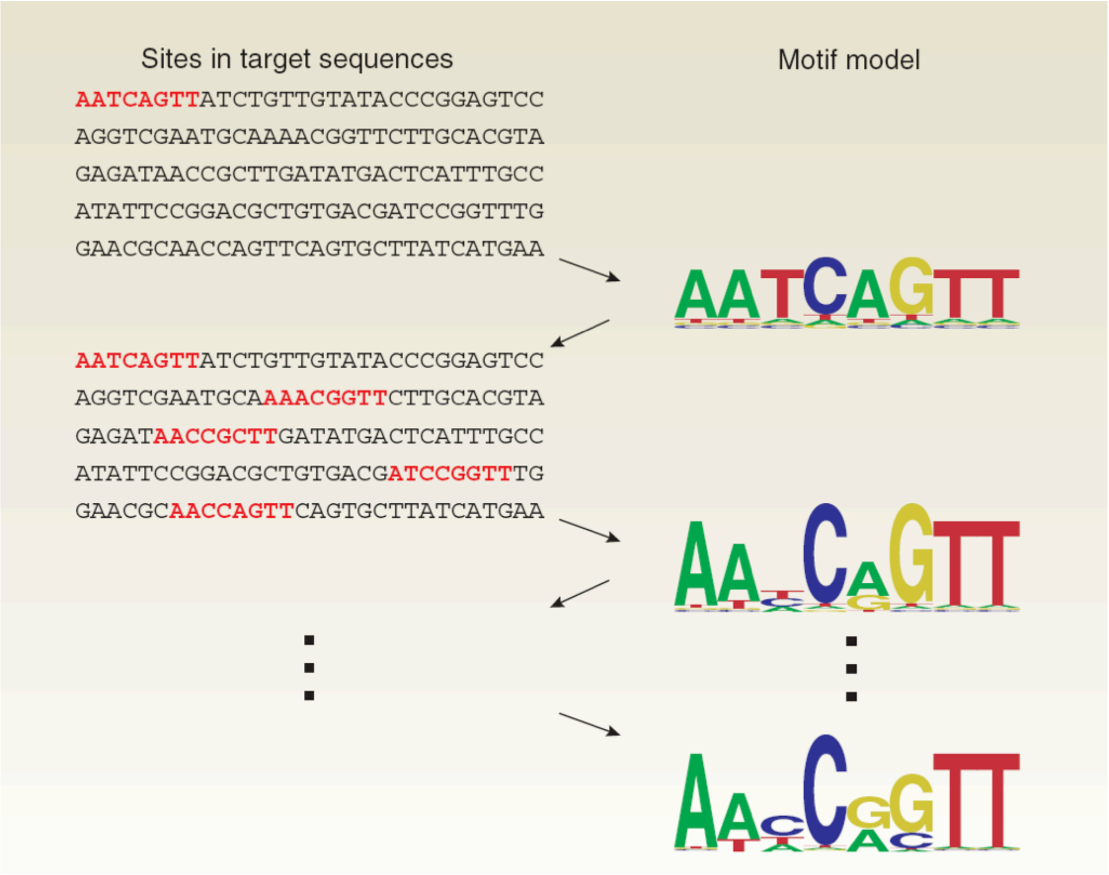
Sites in target sequences
Motif model
AATCAGTTATCTGTTGTATACCCGGAGTCC
AGGTCGAATGCAAAACGGTTCTTGCACGTA
GAGATAACCGCTTGATATGACTCATTTGCC
ATATTCCGGACGCTGTGACGATCCGGTTTG
GAACGCAACCAGTTCAGTGCTTATCATGAA
AATCAGTTATCTGTTGTATACCCGGAGTCC
AGGTCGAATGCAAAACGGTTCTTGCACGTA
GAGATAACCGCTTGATATGACTCATTTGCC
ATATTCCGGACGCTGTGACGATCCGGTTTG
GAACGCAACCAGTTCAGTGCTTATCATGAA

Figure 2.1: Expectation maximization: Select a single site (shown in red), then iterate between assigning new sites to the matrix (right) and updating the matrix (left). Figure taken from [26].

running into a local maxima, each k-word in the sequence is only used in a single iteration. Then the highest weighting one is selected and iterated until convergence. Because the initial choice of a k-word directly affects the outcome, different runs on the same sequence set could lead to different results.

Gibbs sampling

The general approach is similar to EM, but instead of taking a weighted average across all sites, a weighted sample is chosen. Gibbs sampling itself has its roots in statistical mechanics, the first adaptation to bioinformatics was published in 1993 [28] and was used to detect local multiple alignments. Interestingly, the term 'motif' was not used up to that point.

First, we need a formal definition of the motif finding problem. Given a scoring function $f(y_1, y_2, ..., y_n)$, the problem is to find a vector $\vec{y}$ that maximizes f. Let p be a probability distribution with $p \sim f$. If f is large at the optimum, than sampling from p will most likely provide an optimal result. In some cases, sampling from the joint probability distribution is not feasable, instead we sample from the conditional distribution where all parameters are fixed except for one. The Gibbs sampling is based on a Monte Carlo Markov Chain simulation. The Markov Chain is chosen to has p as its steady state. By running the simulation long enough and sampling from it, an approximation of the steady state can be found. What is Gibbs sampling used for? The input is a probability distribution $p(y_1, y_2, ..., y_n)$ where $y \in S$. In case of motif prediction, S denotes all different motif positions in the input sequences. The complexity of $|S|^n$ might be hard to manage, but $|S|$ can be processed. The output of the Gibbs sampling is a (random) $\vec{y}$ chosen from p.

The first step is to build up the Markov Chain that simulates p. As mentioned, p is a conditional probability distribution. The vectors $\vec{y}$ and $\vec{y\prime}$ differ in one position only.

$$\vec{y} = (y_1, ..., y_m, ..., y_n)$$

$$\vec{y\prime} = (y_1, ..., y\prime_m, ..., y_n)$$

The transition probability T is defined as:

$$T(\vec{y} \rightarrow \vec{y\prime}) = \frac{1}{n} \frac{p(y_1, ..., y\prime_m, ..., y_n)}{\sum_{y_m} p(y_1, ..., y_m, ..., y_n)}$$

Can we be sure that p is the steady state distribution of the Markov Chain? A steady state is present if two requirements are fulfilled, the *global balance* and the *detailed* *balance*. The *global balance* definition says:

$$\pi T = \pi$$

The *detailed balance* constraint is:

$$\pi(\vec{y})T(\vec{y} \rightarrow \vec{y\prime}) = \pi(\vec{y\prime})T(\vec{y\prime} \rightarrow \vec{y})$$

It can be shown that the *global balance* is fulfilled if the *detailed balance* is fulfilled. Furthermore, by setting T to the defintion above, the term evaluates to true. This proves that the Markov Chain simulates p.

A motif prediction based on Gibbs sampling works similar to EM. First, for each of n sequences in the input set, one start positions $(y_1, ..., y_n)$ is chosen randomly. Given a fixed motif length k, we achieve a set of k-words that are combined into a common PWM.

Second, one sequence is taken out of the set and the weight of each k-word in that sequence towards the matrix is calculated. As a result each position in the sequence is assigned to a weight.

Third, the former start position in the removed sequence (y_m) is replaced by a new one $(y\prime_m)$, picked randomly according to the weights. The higher the weight the higher the probability of the k-word under the matrix and the more likely it is that the start position of that k-word is chosen by random. The matrix is updated in relation to the new sequences.

The steps two and three are iterated until convergence.

Gibbs sampling has been implemented into multiple different motif prediction tools. Enhancements such as MotifSampler [29] use higher order Markov Chains as background to avoid a bias towards repetitive elements in the weighting step. Other applications like PhyloGibbs [30] work on aligned sequences and introduce conserved positions as an additional constraint.

Using Motif Predictions

In [12] 13 different implementations were tested on multiple sets of eukaryotic sequences (yeast, Drosophila, mouse), into which known binding sites were inserted. None of the tools was able to predict more than 10% of the motifs. Furthermore, the simpler and faster approach of Weeder perfoms surprisingly well. In general, the overlap between the tools depends on the method they are based on. Different implementations of similar algorithms come to similar results. Different methods on the other hand can be combined to increase the outcome. The enumarative Weeder was shown to be complementary to MEME, while similar algorithms showed higher overlap. In order to maximize the chance of good prediction results, some aspects have to be considered [26].

- Use a combination of multiple tools. For instance, MotifSampler, based on Gibbs sampling is complementary to enumerative tools such as Weeder.
- The selection of the input data has a huge impact on the result, independent from the prediction method. Therefore the use of different data sets could improve the results.
- Masking found motifs and run the prediction again can lead to new results. This way strong signals that may cover other significant motifs can be removed.
- Especially with tools such as MEME, which select one motif each time and try to optimize its likelihood, multiple runs may provide different results.

2.1.3 Clustering

Multiple non-degenerated words found by enumeration approaches can be representations of the same transcription factor binding site. If probabilistic prediction methods are used, different matrices found by different tools, or by one tool in different sequences can also refer to the same motif. As long as it is known which words or sequences belong together, combining them into a common matrix is simple. If the related elements are mixed with a set of unrelated elements however, a clustering method has to be found. In general, several different statisticial techniques for clustering of observations are known. For example, K-means clustering is based on a predefined number of resulting clusters, which is obviously difficult to provide in our case of motif clustering. Several applications for motif clustering exist, such as CompareAce, PROCSE, Mat-Compare, TREG, and YRSA [31, 32, 33, 34, 35], but all these tools lack some of the functions needed in this work. The different motif prediction tools work on score matrices as well as on consensus sequences. Therefore it is mandatory that the clustering algorithm is capable of dealing with both. Furthermore, the clustering tool should be able to create non-redundant motif sets by combining motifs on its own until optimal clusters are created. Two different tools following completely different approaches are chosen and described in the following.

Hierarchical clustering

One can think of a motif clustering as a two step approach. First, a similarity function has to be defined to decide which motifs are to be combined into a larger cluster, resulting in a growing tree in which the motifs are represented as leaves and clusters as nodes. The root eventually represents the all containing cluster. Second, a criterium has to be found to decide where to cut the tree in order to get a set of distinct clusters (in this case score matrices) that in an optimal case only contains motifs related to the same transcription factor. The tool MATLIGN [36] is based on hierarchical clustering.

In the first step the algorithm does a pairwise comparison of each motif in the set and combines the two most similar ones recursively, until eventually one single cluster remains. The similarity between two matrices is calculated by a dynamic programming algorithm that is based on Gotoh's algorithm for gapped alignments [37]. Different scoring functions are implemented in MATLIGN, namely Kendaull's tau rank correlation coefficient, Spearnan's rank correlation coefficient, Pearson's correlation coefficient, and the normalized Euclidean distance. By default, the product of the scores is used to calculate the pairwise distance. If a new cluster is built, the distances to all other motifs are recalculated as the average of the motifs in the cluster and the remaining motifs. The clustering procedure ends in a hierarchical tree whose root represents the all containing cluster. In a second step the number of clusters is optimized based on silhouette values, which give an evaluation of the classification of the clusters:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i),\, b(i)\}}$$

where $a(i)$ gives the average distance of an element to all other elements in the same cluster and $b(i)$ gives the average distance of the element to all elements inside the closest different cluster. The result $s(i)$ is an estimation of the classification, reaching from -1, which shows poor clasification to $+1$. The maximum silhouette value gives a cut-off for the hierarchical tree, resulting in a list of optimal clusters under this model.

Bayesian clustering

This method is implemented in the clustering part of the Phyloclus [38] framework. In contrast to the hierarchical clustering, no explicit separation into two independent steps of clustering and choosing the stop criterion is done. Alternatively, a Bayseian approach is implemented. The probability of a cluster given the motifs would allow straight forward maximization. However, that probability cannot be calculated directly. Therefore the likelihood is taken instead:

$$L(\Theta_i|Y_i) = P(Y_i|\Theta_i)$$

where Y_i is a count matrix indexed by i, and Θ_i is the set of clusters indexed by i.

Therefore the probability of a matrix given the set of clusters is:

$$P(Y_i|\Theta_i) = \prod_{j=1}^{w} P(Y_{ij}|\Theta_{ij})$$

where the column is indexed by j. $Y_{ij} = (Y_{ijA}, Y_{ijC}, Y_{ijG}, Y_{ijT})$ and w is the number of columns. This can be approximated by a multinomial distribution. The basic implementation idea of Phyloclus is now to use Gibbs sampling to determine whether to add a motif i to an existing set of clusters without i (z_{-i}) as the c-th cluster ($P(z_i = c|z_{-i}, Y)$ or to leave the motif unclustered ($P(z_i = 0|z_{-i}, Y)$. Because in each step all parameters stay fixed except for one (i), a conditional probability distribution is modeled and Gibbs sampling can be used to sample z. The result is the cluster set that gives the highest posterior probability $P(z|Y)$.

2.2 Biology

2.2.1 What are Polycomb/Trithorax Repsonse Elements?

Homeotic genes are genes involved in the regulation of morphogenesis during the development of an organism and have been found originally in flies [39]. Hox genes are homeobox-containing genes, usually occurring in so-called Hox clusters. In the mean time it has been discovered that hox complexes exist in all bilaterally symmetrical species [40]. Their main function is to regulate the development of the body axis [41], for example in the Drosophila fly Hox genes determine which parts of the embryo develop into each of the segments of the adult fly. In Drosophila two hox complexes exist, the Bithorax complex and the Antennapedia complex.

During the analysis of the Bithorax complex (BX-C) in Drosophila different studies came to the conclusion that the regulation of the Hox genes is done in two consecutive steps [7, 42]. In the initiation step during the first few hours of embryonic development, gene expression patterns of homeotic genes are set up by activators and repressors. These factors occur at different concentrations in different tissues of the embryo, so depending on the future cell identity, different types of “blueprints” are used by the cells [43]. Once the genes’ expression states are set up, a second mechanism takes over to maintain the transcriptional status through various cycles of cell division. This is done by the so-called Polycomb/Trithorax Response Elements (PRE/TREs). The identity of each cell is passed on its daughter cells even if the adult state is reached. Proteins of the Polycomb group (PcG) maintain repression states of regulated genes while Trithorax group Proteins (trxG) act antagonistically, keeping the

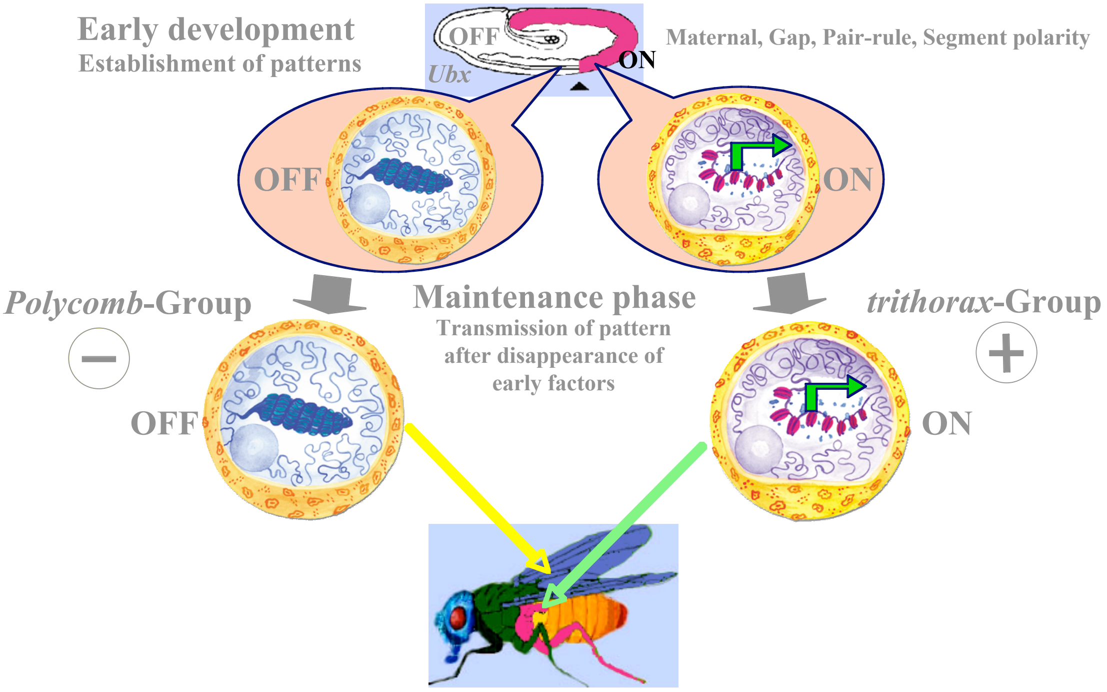
Early development
Establishment of patterns
OFF
Ubx
ON
Maternal, Gap, Pair-rule, Segment polarity
OFF
ON
Polycomb-Group
Maintenance phase
Transmission of pattern after disappearance of early factors
trithorax-Group
−
+
OFF
ON

Figure 2.2: Transcriptional status of gene *ultrabithorax* as an example for PRE/TRE regulation. Picture taken from Polycomb Teaching Page (`http://www.igh.cnrs.fr/equip/cavalli/link.PolycombTeaching.html`). Transcriptional patterns set up in larval state are tissue specifically maintained into the adult state, resulting in the development of different body parts. Depending on the torso segment, legs, wings or halteres are developed.

gene in its active state. It has been shown that PRE/TREs are able to keep the status of a gene even in the absence of the initial activation or repression factors [44].

For example, in the early larval development state the gene *Ultrabithorax* (*ubx)* inside the BX-C is activated in some cells and repressed in others. In the following maintenance phase the repression state is “frozen” by the PcG proteins, while the gene is kept active in other cells by the trxG proteins, even after the initial repressors and activators have disappeared. In the adult fly the *Ultrabithrox* gene product determines the development of wings and legs in the second and third thoracic segment. Cells of the second segment in which the gene is switched off are part of the wings, in the third segment the activated gene leads to haltere development (Figure 2.2). A mutation in the *ubx* gene causes a turning of cells of the thoracic segment three into their segment two counterpart; the halteres are transformed into a second wing pair, the third leg pair is transformed into the second one. *Ultrabithorax* is one of the first genes that were experimentally verified to be targets of PRE/TRE regulation [45, 46].

2.2.2 Clusters of PRE/TREs

The segment identity in Drosophila is set up and maintained by two different clusters of homeotic genes, the Bithorax complex (BX-C) [47] and the Antennapedia complex (Ant-C) [48]. The BX-C consists of only three protein coding genes *Abdominal-A* (*abd-A*), *Abdimonal-B* (*Abd-B*) and *Ultrabithorax* (*Ubx*) and is about 300 kb long.

The DNA binding proteins ABD-A and ABD-B specify the parasegments PS7-PS14 [49, 50]. The genes in the BX-C occur in the order the segments are effected, the order therefore is *abd-A*, *Abd-B* and *UBX*. This order as well as the genes' sequences are conserved from flies to mouse and human. However, in mouse and human the BX-C and Ant-C are clustered together into a common Hox cluster that is duplicated four times.

In Drosophila, the BX-C genes are regulated by *cis*-regulatory elements in 9 infra-abdominal regions (iab1-iab9), one domain for each segment. Each of the domains contains enhancers and PRE/TREs.

2.2.3 How are PRE/TREs working?

In general PcG and trxG proteins are able to modify the chromatin structure to maintain either a gene's activation or repression state. The PcG proteins in Drosophila can be separated into two categories, the Polycomb repressive complex 1 proteins (PRC1), and the Polycomb repressive group 2 proteins (PRC2). The PRC2 consists of four main proteins, Enhancer of Zeste (E(Z)) [51], Extra sex combs (ESC) [52], Suppressor of zeste-12 (SU(Z)12) [53], and the nucleosome-remodeling factor 55 (NURF-55) [54] (reviewed in [55]). Through E(Z) the PRC2 leads to trimethylation of lysine 27 histone H3 [56] and to a lesser extent H3K9me3, both provide binding sites for the Polycomb (PC) protein. Polycomb together with Polyhomeotic (PH), Posterior Sex Combs (PSC) and dRING are members of PRC1. Recent studies suggest that additional PcG members might play a role, such as Polycomb-like [57]. Beside the two Polycomb repressive complexes two additional complexes Pcl-PRC2 and PhoRC have been described [58, 59].

Although a lot of the details about the way PRE/TREs work have not been resolved so far, some key steps are known. The PRE/TREs in Drosophila are DNA regions containing clusters of binding sites for a number of different proteins which directly or indirectly recruit PcG protein binding. One of the most important proteins is Pleihomeotic (Pho) [60] which, once bound, recruits the PRC2 protein E(Z) [61]. This leads to H3K27me3 methylation, which is detected by PC and recruits PRC1 proteins. Additionally, Pho can recruit PC direclty, as shown in [62]. Additional to Pho other DNA binding motifs that assist the Pho function occur in PRE/TREs, such as

Pleihomeotic-like (Phol) [63], GAGA factor (GAF) [64] and Zeste (Z) [65]. More recent studies implied that additional binding motifs play a role like DSP1 [66], Grainyhead and members of the SP1/KLF family [67]. Although it has not determined completely so far which motifs are sufficient for PRE/TRE functionality, there is strong evidence that GAF, Pho and Z are necessary and act together to recruit PRC2 proteins.

Available experimental methods such as antibody based ChIP experiments or transgenic flies are based on detecting binding sites of PcG proteins or demonstrate the inhibition of gene transcription and its genetic dependence on PcG genes. In this case, the functionality to maintain the repression state is targeted by the experiments, a validation of PREs. Corresponding methods for the antagonistically acting trithorax group proteins are used in a similar way. Thus the same genomic region can act in both ways, depending on the transcriptional state that is maintained. In this work I will use the term PRE to refer to regions targeted by PcG proteins, without making any assumption about possible trxG related functionality unless stated otherwise.

In [66] a minimal PRE containing only these motifs embedded in a random bacterial DNA showed no PRE functionality, while the presence of an additional Dsp1 led to PcG protein recruiting and PcG-dependent silencing. Dsp1 binding can be observed at many locations to different binding motifs, the representation used by Dejardin et. al was GAAAA. Although a colocalisation of Dsp1 and Pho has been found on polytene chromosomes, it also has been shown that Dsp1 functions as a trxG protein. Grainyhead and Sp1/KLF [67, 68] have been observed to be important for PcG recruiting in specific PREs. However, a general colocalisation with PcG or trxG proteins on polytene chromosomes is missing.

How is silencing working? The exact mechanisms of PcG mediated silencing are not known so far. What is known is that PREs are bound by all known PcG protein complexes and that histone methylation plays an important role. A comparison of the *ubx* gene in its on and off state by quantiative ChIP analysis [69] suggests that trimethylation in the promoter region and coding regions is important. While extensive trimethylation of promoter, coding regions, and an upstream contol position can be observed in the off state, only the upstream control shows trimethylation in the on state.

2.2.4 Methods of PRE detection

Originally only a handful of PREs have been known, all located in *Drosophila melanogaster*. Recently, in two different studies Chromatin Immunoprecipitation (ChIP) has been used [70, 71], covering different regions of the Drosohpila genome while in a third study the DamID technique was applied [72]. The approach is similar in all three cases, regions at which a binding of PcG proteins can be shown are very likely to be PREs. In contrast to the direct DNA binding proteins like GAF, which is involved in heat shock reactions as well, no additional function than gene repressing by PRE binding is known for Polycomb in Drosophila.

Polytene pictures

A common method for the search of binding sites of a specific protein in a genome-wide scale in Drosophila is the use of immunofluorescence on polytene chromosomes. In general polytene chromosomes are the result of multiple rounds of DNA replications without cell division (endoreplication). Various copies of homologous chromatids remain inside the same cell and are banded together to large chromosomes. Those structures have been found in larvae of some two-winged flies like *Drosophila melanogaster* and in few plants in ovary and immature seed tissues. In Drosophila, chromosomes of salivary glands in larvae form giant polytene chromosomes to allow the production of huge amounts of glue, required for pupation. Immunofluoresence is a technique to label protein specific anti-bodies that can be detected for example by a fluorescence microscope. A usual method is to use an anti-body against the protein (or antigen in general) of interest in the first step, which is brought on the polytene chromosomes. In a second step labeled anti-bodies against the first anti-body make binding sites of the first protein visible. Because the binding sites are evaluated by visual techniques, polytene chromosomes are used to enhance visual evaluation. The positions are usually derived from cytological maps. Therefore multiple binding sites in one region cannot be distinguished, so neither the exact amount of binding sites nor their exact genomic location can be determined. Nevertheless, the technique can show binding of a protein in general and provides valuable data like clusters of active cytological location or rough estimates about the number of binding sites in general.

Chromatin Immunoprecipitation

ChIP, or the large scale approach ChIP on chip, is used to detect the binding of the protein of interest (poi) at a specific time point in a chosen cell line. Compared to immunofluorescence on polytene chromosomes, this method allows to achieve a higher resolution i.e. more exact positions can be determined.

In the initial step the poi is cross-linked to the DNA, e.g. by formaldehyde, in the studied cells, therefore only proteins that are bound at that moment can be detected. In the next step the DNA is sheared into smaller pieces by sonication. The fragments are immunoprecipitated with an antibody against the poi. The DNA from the samples is purified. To test single positions for poi binding, location specific primers can be used for PCR. By quantitative PCR the amount of enrichment of the immunoprecipitated samples in relation to a background can be measured. A large-scale alternative of the last steps is to use DNA microarrays; which is called ChIP on chip. In this case the poi containing DNA samples are purified, amplified and labeled, e.g. by a fluorescent marker. The single stranded DNA fragments are brought onto DNA microarrays, which cover a larger range of the genome. The labeled fragments hybridize on the microarrays, allowing an identification of the corresponding genome locations.

Although ChIP experiments are widely used for all different kind of protein binding search, some aspects raise minor concerns about specificity and sensitivity when it comes to PRE studies. First, the proteins have to be bound to the DNA at the moment of the cross-linking in vivo in the cell line used. As already shown in the *ubx* example, Polycomb binding is highly tissue specific. Depending on which cells are used, even known PREs cannot be detected by ChIP in every case. For example, the well characterised fab7 PRE is missed in some of the ChIP experiments [11, 71]. Second, antibody affinities have a huge impact on the ChIP outcome. Furthermore, to ensure specificity a statistical threshold is used, therefore observed enrichments of true protein bindings may fall below this threshold. For genome-wide mammalian PcG ChIP studies a false-negative rate of 30% has been reported [73].

On the other hand, a positive ChIP result does not necessarily prove PRE activity at the observed position. It has been shown that PcG proteins can loop from a PRE/TRE site to other sites like the regulated promoters [74]. In this case, ChIP experiments will show protein binding at the promoter although no protein recruiting takes place there.

Another experimental technique for detection of protein binding is DamID. The protein of interest, in our case Polycomb again, is fused to a DNA methyltransferase. The resulting fusion protein is expressed by the transfection of cultured cells. The methyltransferase is linked to the positions at which the poi binds, leading to a higher methylation of that region. In contrast to ChIP, where due to the cross-linking only a snapshot of protein binding can be observed, DamID allows in theory the detection of all spots at which the poi bound to during the observation time. Although this is the main advantage of the technique in theory, it raises new problems. PRE/TRE are usually only less than 1 kb long, in Tolhuis et al. [72] however, DamID experiments showed regions with an average size of 30 kb. Traces of temporary binding of Polycomb to the DNA cannot answer the question of whether it actually functions in the region.

Schwartz et al. [71]performed genome-wide ChIP on chip experiments on Sg4 cells, probing for binding of the PcG proteins PC, E(Z) and PSC. Additionally, trimethylation of histone H3 Lys27 (me3K27) was searched. Regions that show a binding of all four factors are called “strong” sites, while regions bound by at least two factors are “weak” sites. Exact positions are not provided in the publication, instead names of genes that are assumed to be PRE regulated in each cytological position are given. All in all, 187 “strong” and 73 “weak” genes are detected.

Negre et al. [70] used Drosophila embryos to evalute PC, PH and GAF binding in 7 Mb of the X chromosome and 3 Mb of chromosome 2L. Regions that show a binding of all three factors are assumed to be PREs. In the evaluated regions 41 PREs are found, the average length of each hit is 5 kb.

Tolhuis et al. [72] used DamID on Kc cells and evaluated binding profiles on chromosomes 2L and 4, 11 Mb of chromosome 2R, and 2Mb of the X chromosome. They found 131 hit bands with an average length of 28 kb.

Transgenic flies

Although ChIP experiments are able to provide locations of PRC1 and PRC2 protein bindings, the presence of PRE functionality cannot be completely proven this way. By using transgenic reporter flies some aspects of Polycomb mediated silencing can be shown. When a PRE is brought into a P-element vector upstream of the *miniwhite* reporter gene, it silences the expression of the *miniwhite* gene in Drosophila. The gene product is required for the red eye color in white mutants. Darker eyes can be observed in miniwhite homozygotes in contrast to whiter eyes in miniwhite heterozygotes. If the gene is PRE silenced it can be observed that now the silencing is enhanced in flies that are homozygous for the insertion, giving a lighter eye color. This typical PRE behavior is called paring sensitive silencing [75]. Another typical PRE behavior is variegation of the eye color, which also can be observed in transgenic miniwhite reporter flies [76]. Furthermore, a lack of PRE repression of the miniwhite gene in PcG mutants can be seen by the eye color as well as a lack of miniwhite activation in trxG mutants.

In-silico prediction

In 2003 Ringrose et al. [11] presented a software tool called PREdictor that is designed to predict PREs in *Drosophila melanogaster*. It works by searching for consensus sequences of a few known DNA binding motifs that are involved in Polycomb recruiting. The motifs used are Gaga factor/Pipsqueak, Zeste, Engrailed and three versions of Pho/Pho-like. The algorithm uses a sliding window to calculate scores for an input sequence. The default parameters are a window of size 500 bp that is shifted by 100 bp each step. Within each window the number of motif pairs is counted. A pair is defined as two motifs occurring within a distance of 0-220 bp. Each possible motif pair is assigned a specific weight, the sum of the weights of all found motif pairs inside the window is the window's score. The weights are determined based on motif pair occurrences in a positive training set (the model) in relation to a negative training set (the background). The model consists of twelve already known PREs, mainly located inside the Bithorax complex. As background 16 promoters of genes are chosen that contain some of the motifs. These are for example heat shock genes that contain GAF motifs but are not involved in Polycomb regulation. The weight of a motif pair m is the log-odds score calculated as

$$\log \frac{f(m|model)}{f(m|background)}$$

The function f simply counts the motif pair occurrences combined with sequence length normalization. PREdictor assigns a score value to each window of an input sequence, a higher score indicates a more probable PRE location. To assess the specificity, a non-parametric empirical statistic is used. The software scores random data of 100 times the length of the *Drosophila melanogaster* genome. The score that gives 100 hits in the random sequence is the cut-off at which one false positive can be expected in the real genome, therefore an E-value of 1 is calculated. The PREdictor predicts 167 PREs genome-wide, out of which 43 were tested by ChIP in S2 cells. An enrichment over 2-fold was observed in 29 cases, 14 were enriched less than 2-fold. Out of those another 10 were strongly enriched for PcG proteins in other cell types, or were confirmed by transgenic asssays [11, 71, 72].

The algorithm gives best results for prediction PREs that regulate homeotic genes, in that case all predicted PREs could be tested positively. This observation might be due to the fact that training was done on classical canonical PREs that contain clusters of known motifs. A class of PRE that can be found near homeotic genes. It is not clear if there are different kinds of PREs, regions that recruit PcG proteins by other mechanims than the mentioned motifs. In that case, the PREdictor would miss those elements.

A newer and more general rewrite is the jPREdictor [10]. It introduces a more generalized form of motifs and motif pairs. Instead of relying on consensus sequences only, it is now possible to provide motifs as position specific weight matrices (PSWM), as position specific score matrices (PSSM), as regular expressions or simply as gapless multiple alignments. Motifs can be combined to so-called multi-motifs, the minimum and maximum distance between each single motif can be set independently. Most functions are accessible through a graphical user interface.

Some of the new functions were demonstrated in the publication [10] in form of another PRE prediction run. The different Pho descriptions used in the original version are combined into a single matrix, the motif DSP1 was added in form of a Pho:Dsp1 pair, which serves as a single element and is allowed to form motif pairs with the other single motifs. As a result, the number of predicted elements raises to 306. On the other hand, the new set is partially overlapping with the old one, instead of just extending it. A validation of the prediction results by ChIP or transgenic experiments is lacking. For our analysis in this work the old set will be used because it is based on a stronger validated background.

2.2.5 Experiments in mammals

Up to this point no PREs have been found in vertebrates. However, there are strong indications that elements similar to PREs exist. Two studies came to the conclusion that PcG proteins are able to keep embryonic stem cells pluripotent by maintaining differentiation genes silenced [13, 14]. In contrast to fly hox genes, where PcG proteins maintain transcriptional decisions over many cell generations into the adult fly, the effect lasts less long in mammals. During the embryonic development differentiation of ES takes place at some point, which requires the PcG protein to end the silencing function. One theory is that trxG proteins are able to switch PRE/TREs from repressing to activation state, but validation is missing. If PcG proteins show similar functions in mammals too, why have no PREs been determined so far? Although most members of the PcG show strong conservation, no functional analogs for the DNA binding proteins GAF/PSQ and Z exist in vertebrates. Analogs for Pho and DSP1 are named YY1 and HMGB2.

In recent studies, ChIP experiments are presented searching the mouse and human genome for binding of PRC1 and PRC2 proteins. In mouse, Boyer et al. [13] searched a region 8 kb downstream until 2 kb upstream of 15742 genes for Phc1 and Rnf2 (both PRC1) as well as Suz12 and Eed (both RPC1) bindings and H3k27me3 methylation. Suz12 has been shown to be involved in trimethylation of H3K27 as well as H3K9 in mammals [77].

The five factors overlap at 561 positions, out of which 512 occur within a radius of 1 kb around a transcription start site. Lee at al. [14] performed genome-wide ChIP experiments in human to detect Suz12 enriched sites, discovering 3465 positions. Again, an enrichment of binding sites within promoters can be observed, around 80% of all found positions are located within short distance to a transcription start site.

2.2.6 CpG islands

The dinucleotide CG is usually rare in vertebrates because the cytosine tends to be methylated. The methylation works as a signal during DNA replication and helps to distinguish the parent strand from the newly synthesized one. Methylated cytosines are deaminated and turned into uracil, which then is replaced by thymin. If not prevented in some way, CG dinucleotides will be turned into CT dinucleotides eventually, given a long enough period of time. Nevertheless, in promoters regions of unusual high CG content exist in mammals, called CpG islands. The ‘p’ refers to the phosphodiester bond between the two nucleotides and is used to distinguish CpG islands from simple CG dinucleotides. The definition of a CpG island has changed over time. Originally, a region of at least 200 bp and a CG percentage of $\geq 50\%$ and an expected/observed ratio ≥ 0.6 was seen as a CpG island [78]. The observed/expected ratio for a sequence of length N is calculated as:

$$Obs/Exp = \frac{Number\,of\,CpG}{Number\,of\,C + Number\,of\,G} \times N$$

The parameters were changed during the analysis of CpG islands in the human chromosomes 21 and 22 [79]. In order to rule out ALU repeats, the minimum length is increased to 500 bp, Obs/Exp to ≥ 0.65 and %CG to $\geq 55\%$. The numbers of genes having CpG islands within the promoter sequences differ between various studies. In general, about 40% of mammalian promoters contain CpG islands [80], while 72% of human promoters have a high CpG content [81]. Older publications stated 56% in human and 40% in mouse [82].

Usually DNA methylation of CG dinucleotides leads to their avoidance, an effect that is prevented in CpG islands. The DNA methylation state of CpG islands in promoters is thought to have a regulatory effect on gene transcription by modifying chromatin structure. Methylated CpG promoters restrict transcription, whereas unmethylated CpG promoters allow gene expression [81]. On the other hand, strong CpG island promoters are normally unmethylated, even if the gene is inactive [83]. Promoters that show only weak CpG island enrichment can be both methylated and unmethylated, presumably depending on the gene function, in somatic cells germline-specific genes are methylated. In case of unmethylated but still inactive CpG island containing promoters, enrichment of histone methylation can be observed. In particular, elevated levels of dimethylation of Lys4 of histone H3 have been found, which maybe act as a chromatin marker to prevent DNA methylation. Vire et al. [84] showed that a direct connection between DNA methylation of CpG islands and chromatin methylation mediated by Polycomb group protein EZH2 exists (Figure 2.3). Polycomb proteins might be able to prevent transcription by acting over CpG islands. .

Another relation between CpG islands in mammals and targets of Polycomb has been suggested by Eden et al. [85]. As an example for their motif prediction approach, Eden et al. ran their software on sequences taken from human cancer cell lines that show

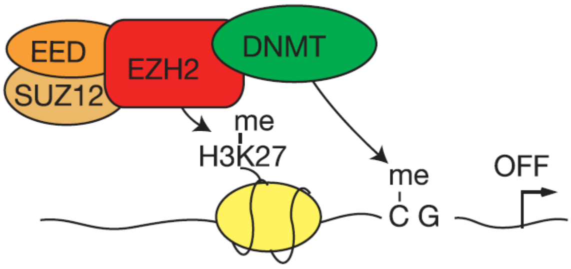

Figure 2.3: EZH2 controls CpG methylation, in the context of the PRC2/3 complexes, through direct physical contact with DNA methyltransferases (DNMT). Figure taken from [84].

CpG methylation and compared the predicted motifs to those found to be bound by PcG proteins. According to their statements, most of their novel motifs are similar to DNA sequence elements that are bound by PcG proteins. Remarkably, they even considered the elements predicted by PREdictor in Drosophila as Polycomb bound regions, although most Drosophila PcG related DNA binding proteins do not even exist in human. Their motifs are basically CA repeats. They ruled out also found CG rich motifs due to possible bias by the CpG islands. Based on this data they suggest a mechanical linking between CpG methylation and histone methylation.

3 Method: *DynScan* - Beyond conservation constraints

3.1 The idea

In order to allow an alignment independent prediction of various elements, in this chapter we will present a method and its implementation named *DynScan.* Our method can make use of the knowledge of orthologous regions, but does not depend on an above average conservation. The method can be applied to any kind of scoring algorithm and is meant to increase sensitivity without losing specificity by using prediction results in one species to adjust the statistics for a search in other species.

The main idea is to perform a genome-wide prediction with an arbitrary scoring algorithm in one species in the first step, followed by a search restricted to the same loci in other species (see Figure 3.1). In each step the search radius is increased around the orthologous position while dynamically adjusting the cut-off to always keep the same specificity level. The cut-off directly depends on the length of the searched region and the number of elements searched in total to guarantee an overall E-value of 1. The smaller the region, the lower the cut-off. The more searches are performed the greater is the totaly searched sequence and the higher the cut-off becomes. Because the cut-off is dynamically adjusted to these two parameters, the package is called *DynScan*.

Although a direct conservation of the searched element is not required, the closer the analogous element is located to the orthologous region, the lower is the required cut-off. This way conserved positions are rewarded but not necessary. It is plausible to expect analogous functional elements to occur in the same locus, not necessarily fixed at the same position, because functional regions like promoters or coding regions are known to be conserved in most cases and elements targeting them can be expected to be located within range. On the other hand, *cis*-regulatory elements like enhancers can be located in a wide radius around the regulated gene [86], single enhancers have been located directly at the transcription start site as well as several 10s of kilobases upstream or inside introns [87, 88]. The method described in the next section is designed to be more sensitive the closer an analogous element occurs to orthologous site, but also increases sensitivity for elements which are up to several kilobases away.

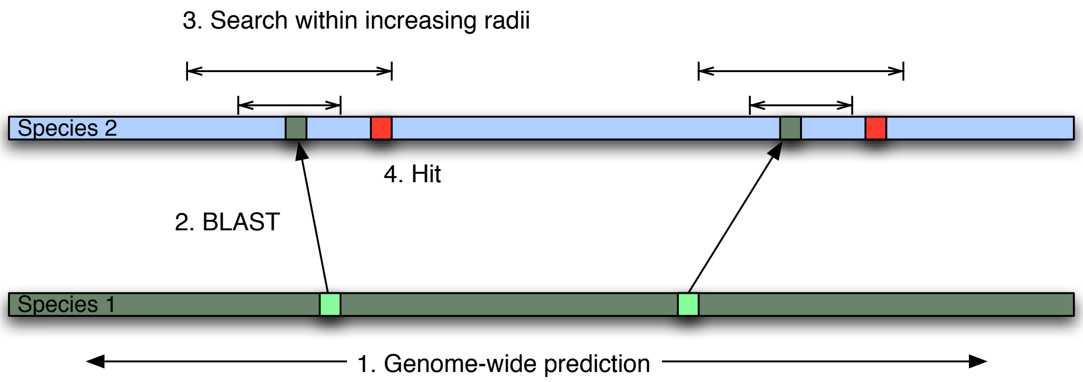

Figure 3.1: *DynScan*: For each element predicted genome-wide in one species (1) the orthologous region in another species is searched (2), increasing the search radius and the cut-off stepwise (3). Search is stopped if either a new element is found or if the cut-off reaches the genome-wide one (4).

3.2 *DynScan* initilization

The requirements are a scoring algorithm that gives a score for each position of a given sequence, such as the jPREdictor, which has been described in Chapter 2.2.4, and a null-model used for the background score distribution. The algorithm can now be applied to a set of genomes, which are known to contain elements that the scoring algorithm can predict in general. Each genome in the set is to be scored, without applying a cut-off at this point. The result should be a score value for each genomic position.

In the initialization phase, background sequences based on the provided null-model are generated and scored. Because the cut-off calculation is based on those data, larger data sets allow more accurate results. As a rule of thumb, random data of 100 times the length of the real data lead to significant results.

The choice of the null-model is independent of the *DynScan* package, which merely works on the provided data. Therefore any model that can be combined with the scoring algorithm is imaginable. Common null-models are random data following the real genome's nucleotide base composition (0-order Markov chain), random data conserving longer nucleotide runs (higher order Markov chains), or shuffled versions of real genomes.

The number of hits at each possible integer score in the null-model gives the background distribution required for a DynScan run. In the actual implementation, the background data is stored in a PostgreSQL database. Additionally, the prediction algorithm of choice is used to score each genome in order to store each position's score in the database as well. Once all genomes are scored and the background scores are calculated, the initialization phase is finished.

3.3 *DynScan* running phase

The emphasis of the whole method is placed on keeping specificity, i.e. we try to minimize the false positive predictions, while still being able to raise sensitivity. This is done for several reasons. First, as can be seen in the analysis of a real world example in Chapter 5, the choice of the null-model can have a largebig influence on the calculated E-values. Applying strong constraints at the start helps to increase the result's plausibility. Even if a change of the null-model later indicates that the E-value used might be in fact higher than expected at the beginning, we will see that the overall outcome of the prediction is only slightly affected and the confidence in the results is not weakened. Second, because predictions made in one species are used as prior knowledge in other species, reliable prediction results are required.

In the first step of the running phase, the cut-off corresponding to an E-value of 1 is calculated for any of the genomes in the set. That means we expect one hit to reach this score by chance when scoring the genome. This significance level of hits is used throughout the whole prediction process. The exact cut-off is selected from the database as the lowest score that can be observed in the background at least n-times, where n is the length of the background data divided by the genome's length. Using this cut-off for a genome-wide prediction gives a list of hits out of which we expect one to be false-positive in each genome.

All non-overlapping positions in the genomes reaching the cut-off are called “static” elements and are kept in a database table. Because the score of each position within each genome is already stored in the database, the static hits are created out of this data in order to build non-overlapping hits. Each static element is assigned to a unique identification number. Once the genome-wide significant hits are written into the database, the orthologous positions for each hit inside the other species are determined. The *DynScan* implementation uses BLAST [89] to search against indeces built from each genome in the set. The BLAST search leads to multiple high scoring pairs (hsp) for each query element, which are also stored in the database. The orthologous position of each element in each other species is then determined as the longest continuous sequence covering multiple hsps to which the following criteria can be applied:

- The hsp's E-value is ≤ 0.005
- All hsps are located in the same chromosome

- All hsps are located on the same strand
- The distance between two adjacent hsps must not exceed 1 kb

The intention of the chosen criteria is to ensure that the derived orthologous region actually hits the correct locus. In case no hit matching the criteria is found in a species, no further actions are taken for that element. The E-value cut-off is meant to filter out ambiguous results. The other two criteria are used to assemble a region that is not completely covered by one single BLAST hit into a common region. For example, if the query sequence has a length of 1 kb, and BLAST finds three hits of 150 bp each with an E-vlaue <=0.005 and located on the same strand with a distance of 500 bp between each adjacent hsp, the whole orthologous region will be assumed to be reaching from the beginning of the first until the end of the third hsp. In this example, the region will be of length 1450, showing 450 bp of gaps. The heuristic used by BLAST assumes that the orthologous positions contain short seeds (default for DNA search is 11) with high percentage of identity, which are later extended in both directions. Although it is possible to miss the orthologous region due to this heuristic, the emphasis on specificity justifies the use of BLAST. Because the DyScan algorithm only needs the homologous loci and not an alignment of the complete query sequence, single hsps covering only small subsets of the query with high probability are enough to proceed.

Once the orthologous positions for each element in the static table within each other genomes are known, the main scoring can take place. The hypothesis is that functional elements might not be completely conserved, but are still located around the same locus in different species, to allow a regulation of the same gene for example. For some *cis*-regulatory elements the presence within a specific distance to a gene can be enough to act as a regulator, while the exact position does not affect the biological function. The algorithm now takes all static elements from the first species and counts the number of orthologous regions (n), determined as described above. In the first step, the search region is set to contain the orthologous region only. Depending on the assumed average length of the element of interest, the search radius around the center of the orthologous position has to be chosen. A common value could be 1 kb. The score cut-off for the prediction around each of the n regions is set to reflect an E-value of $1/n$ for a sequence length of for example 2 kb (assume a radius of 1 kb), which is equivalent to an E-value of 1 for a sequence of length $2000n$.

The scored region is now extended stepwise, each time a new cut-off is calculated to leave the overall E-value set to 1. Increasing radii lead to increasing cut-offs, so the greater the distance of an analogous element to the orthologous region, the higher the score that has to be reached. This way conservation is not required, but elements that have constrained position during evolution get a bonus by the scoring scheme. The algorithm looks in the prediction score database table for the nearest position within the search radius that reaches the cut-off. The procedure stops once an element is found, or if the increasing cut-off comes too close to the genome-wide one, so that the dynamic search is stopped and the nearest genome-wide predicted element is taken. The result of the search is an element in each other species, which in each case is considered to be the functional analog of the query element.

Because the overall E-value is set to 1, only one of the predicted functional analogs in all orthologous loci is expected to be false-positive. Furthermore, because the initial genome-wide search in each species was based on the same stringent cut-off, except for one all other loci indeed serve as prior knowledge of likely elements whereabouts in the target species, which allows a high confidence in the data. The resulting elements are stored in the database, so that for each static element it can be looked up where the assumed functional analogs in the other species are located. In the following section, the implementation details including the underlying database schema are shown. *DynScan* has been used for the prediction of PREs, the steps taken and a detailed evaluation of the achieved results will be presented in Chapter 5.

3.4 Technical details

3.4.1 Database layout

The database layout (Figure 3.2) allows the addition of other genomes without affecting previously stored data. Species are stored in the *species* table, identified by a unique primary key (*id*) and provided with a version identifier. The scores for each position within each genome, identified by a species id, the chromosome, and the base pair position are stored in the *prediction_score* table. In order to allow different prediction runs on the same genome, each run is identified by the "run" parameter, which refers to an entry in the *run* table. The *run* table again keeps track of the parameter set used for each prediction, containing an unique identifier, as well as references to the *motif_set*, *null_model* and *training_set* tables, together with additional parameters "length" and "step". A description field allows comments for each run. The parameters are chosen to reflect the possible parameters used by the jPREdictor, but other prediction software can be used easily either by adding other parameters to the table or by using the description fields in the *ecalc_motif_set*, *ecalc_null_model* and *ecalc_training_set* tables. The cut-off for any given E-value and sequence length is calculated based on the *ecalc_scores* table, that contains the number of hits found for each integer score within the background identified by the "run" parameter. Based on a given E-value, the *score* table entries can be evaluated and all non-overlapping hits

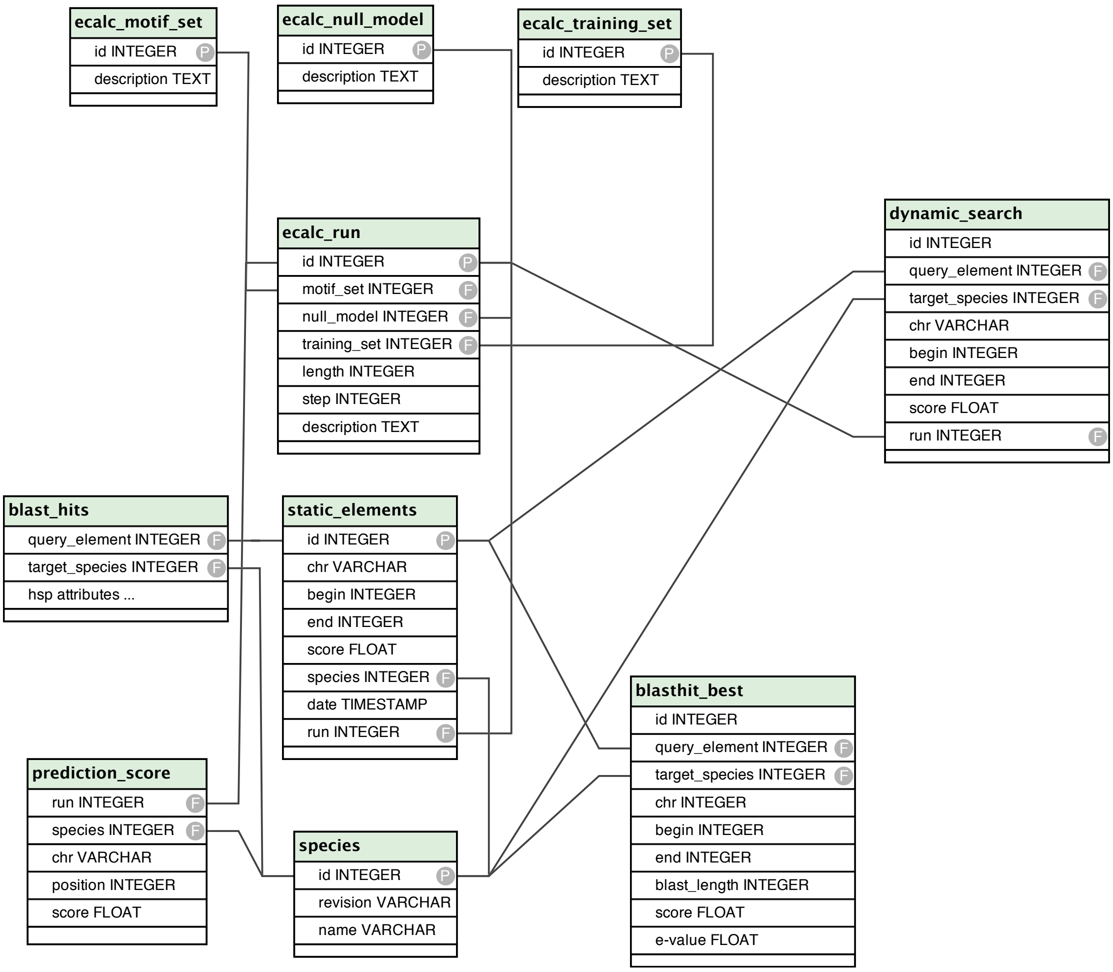

Figure 3.2: SQL schema of the database. Lines denote foreign key constraints. Attributes labeled as "P" are primary keys, the ones labeled as "F" are foreign keys. Only most elementary tables are shown.

that reach the score are stored in the *static_hits* table, identified by a unique primary key id and stored with location (chromosome, begin and end) as well as the maximum score of each hit, a reference to the *species* table, a timestamp of the database entry and a reference to the “run” parameter sets.

This way it is possible to combine all genome-wide predicted elements in one table, regardless of the species or the parameters used for the prediction run. The results of a BLAST search of each genome-wide predicted element within the other genomes of the species set used for *DynScan* are stored in the *blast_hits* table. It contains a reference to the genome-wide element used as query, a reference to the targeted species, a timestamp of the BLAST search and of course the BLAST results, namely the chromosome (named "hit_accession"), and the details of each found high scoring pair (hsp). The stored hsp details are score ("hsp_score"), start and end position in the query sequence ("hsp_query_from", "hsp_query_to"), start and end positions in the target sequence ("hsp_hit_from", "hsp_hit_to"), number of gaps, length of the alignment, percentage of identity, number of similarities ("hsp_positive") and most importantly the E-value of the hits. Out of the single hsp the positions of the orthologous regions are determined and stored in the *blasthit_best* table, that contains an unique primary id key, references to the query element in the static table, and a reference to the targeted species as well as the position's location, stored in the chromosome, begin and end fields. The sum of the length of all hsps in one region is written to the "blast_length" field. The results of the dynamic search are stored in the *dynamic_search* table, identified by a unique id. Each stored element contains a reference to the query element in the *static_hits* table, the id of the species the element is located in ("target_species"), the position ("chr", "begin", "end"), the element's score and a reference to the parameter set used in the prediction (“run”). Additional information about each element is stored in *static_elements_gene* and *dynamic_elements_gene* tables. These tables contain cross-references for each element's id to gene ids of the nearest located gene. The genes are stored in the *genes_positions* table, which holds locations of known genes inside the species. The table's attributes are a unique id for each gene, a species reference, a gene's global id referring to the entry in the original genome database, a name, the position (chromosome, begin and end), and globally unique identifiers. Which identifiers are chosen depends on the species; in case of Drosophila, FlyBase [6] identifiers are used.

3.4.2 Implementation

The *DynScan* algorithm is designed as a software package written in Perl. Each step in the dynamic search is implemented as a standalone script, all combined by a main

```
>Chr2
 250    6.554
 260    7.445
 270    8.556
 280    9.0
 290    10.12
 300    8.4
```

Figure 3.3: Example of jPREdictor raw score output. Sequence identifier is provided in FASTA format, position and score are seperated by any whitespace character.

wrapper script that serves as the main component once the initial phase is finished. Several helper tools are provided to set up the requirements.

Species identifiers and names must first be stored into the database, as well as identifiers for the different prediction parameter sets. The only prediction parameters required by the dynamic search are the width in case a sliding window is used and the value by which the window is shifted in each step. This is important for later calculations of begin and end positions of hits. Once this is done, a helper script can be used to store positions and their scores for each genome used in the database. The format expected is the one used by jPREdictor (see example Figure 3.3). Chromosome names are given in form of FASTA format, in the other lines position and score values are expected, separated by a whitespace character. As the position of a scored window the center should be provided, so that start and end positions are calculated based on the window width. The required data for the E-value calculation are scoring results of significant amounts of data following some null-model.

In addition to the introduction of the different species to the database and the import of prediction scores, the scripts need to know some parameters, which are to be provided in form of a configuration file (syntax in Figure 3.4). Mandatory settings are the BLAST index for each species and the path where the fasta files of each element will be stored together with the location of the genomes' fasta files. A list of distinct non-overlapping predicted PREs with an E-value of 1 or less can be built out of the stored postion scores in the database.

For each entry in the *static_element* table, an additional script creates a fasta file inside the directory set in the configuration file by cutting the region out of the provided genomes' fasta files. The files' filenames contain the element's database id as well as the chromosomal position. The *DynScan* software can now be started by the main script (1_add_species.pl), which needs to be provided with a species name and a

```
#Generic path to chromosome. “CHR” is replaced by specific
value according to names used in database
$chr_path{pseudoobscura_2.0}=/path/to/chr/dpse/CHR.fa;
#Path to genome-wide predicted elements
$pre_path{pseudoobscura_2.0}=/path/to/elements/;
#Path to BLAST indeces
$blast_path=/path/to/blast/indeces/;
#Name of species specific BLAST index
$blast_index{pseudoobscura_2.0}=dpse-all-chr-2.0;
#Prefix for database tables
${test_string}=debug;
#Number of parallel processes
$jobs=4
#Database specific “run” parameter (refers to motif set,
training set,...)
$run=1
```

Figure 3.4: *DynScan* configuration file. Parameters are described in Perl syntax. Lines beginning with "#" are comments.

configuration file (options described in Appendix A.1). A dynamic search will be performed against all other species. The software iterates over all species in the database and selects a new target species in each run while the query species stays the one provided. The subsequent steps will be called automatically but can be run manually if needed.

In the first step, each static element in the query species stored in the database serves as input for a BLAST search against the index of the target species of the run. The input file is taken from the file system, identified by the element's database id; the location of the target index is taken from the configuration file. In order to enhance running time, multiple BLAST jobs can be run simultaneously. The number of threads can be set in the configuration file. The result of each BLAST run is a single XML file, which is stored in a temporary directory and parsed in the same step. The BLAST results are written to the *blast_hits* database table.

In the second step, the orthologous positions are determined by another script, which takes query and target species as options. By default, the script is called automatically with the actual query and target species. For each static element of the query species in the database that has a BLAST result in the target species, the homologous position is determined by applying the criteria mentioned in Section 3.3. If existent, the determined region is stored in the *blasthit_best* table. Because the script iterates over all static elements, it is possible to gain benefit from parallelization again, especially if the script is run on a multi-processor or multi-core machine.

Once the homologous positions are known, they can be used as prior knowledge in the dynamic search, done in the third step by the next script, which again takes the query and target species as options. The main script calls this script with the current values for query and target species. For the actual dynamic search, each entry in the *blasthit_best* table is taken that is referenced to a static element in the query species and belongs to the target species. Around the center of the region, increasing search radii are used. The steps by which the radius is increased are set in a hash inside the script. The default values, used for the prediction od *cis*-regulatory elements with the jPREdictor, are 1 kb, 10 kb and 20 kb. The required cut-offs at each step are depending on the radius and the overall searched number of orthologous regions and are calculated at run time. The radii can be set to arbitrary values. The radius hash is processed step by step with increased values. Each time the analyzed region is set to the center of the orthologous region, extended in both directions by the radius. If the highest score in this region, looked up in the score table of the database, reaches the current cut-off a hit has been found. The position of the highest score is then extended to the maximum region that scores above the cut-off and returned as the assumed functional analog. If all radii in the hash table have been processed without finding a dynamic hit, the nearest located static hit on the same chromosome is returned.

The returned hit region is written to the *dynamic_element* table. Because the steps can be done independently for each orthologous region, they can be run in parallel threads again. After all *blasthit_best* entries have been processed, the steps one until three are repeated with switched query and target species. The complete run is done for every other species in the database. This way each species can be added separately by a run of the main script "1_add_species.pl". Adding species number n requires $2(n-1)$ dynamic searches. The most essential scripts are listed in Appendix A.1.

3.5 Evolutionary studies

Regulatory elements, which consist of several clustered transcription factor binding sites, can be predicted computationally by software such as the jPREdictor. With the *DynScan* software we presented a method to increase sensitivity of such predictions. But if we are interested in the question, what is necessary to turn a given not-functional sequence into a predictable element, a new approach is required. The question might arise if *DynScan* reveals that a functionally analogous element is shifted outside the orthologous region to different spots in various species. Which of the regions that are functional in one but not functional in other species is the evolutionary ancestor? We base our method on the idea of motif presites, regions that have a higher predisposition to become a functional motif than other surrounding regions. For enhancers it is known that presites of motifs are overrepresented in some regions [90], allowing the gain of a regulatory element within fewer generations than in regions without or with fewer presites. Presites are defined as sequences that need only minor mutations to turn into a transcription factor binding site. We call regions that contain clusters of presites, or even mixtures of functional but not sufficient motifs and presites, pre-elements in general. In case of the Polycomb/Trithorax Response Elements we use the term pre-PRE.

As a second aspect we have to keep in mind the frequency of presites expected to be found by chance within regions of given window length. For each motif, we calculate the Hamming distance at each position in a sufficient amount of background data, e.g. a complete chromosome or even a whole genome. The calculation is done for each window of motif width, which is slid in steps of one. The Hamming distances for each motif are therefore between 0, which is a direct match, and the motif length, which means no overlap at all.

Within each window the numbers of occurrences of the same distance are summed up for each motif. Depending on each motif's length and degeneration, different distances are chosen to define a presite Higher numbers of presites within a window mean higher probabilities to gain a motif by mutation. We explicitly do not consider motif pairs but look for windows in which different motifs have a high chance of being gained within few mutation steps. Based on background number of presites within a window, we can calculate the p-value for a specific number of sites to be found in a single window simply by counting how many percent of all windows in background have at least the same number of presites. The overall p-value for a window is the product all motif's p-values in this window. The smaller the p-value, the higher the chance that the region spanned by the window becomes functional by random mutations.

However, the model only expresses at which positions more presites with small Hamming distances to motifs are present and therefore have a higher chance of gaining a motif within less mutation steps than other regions. Nevertheless, even windows having less presites than the background could gain motifs by only one single mutation, but the chances are lower. As long as a presite occurs within a window, it cannot be said that this sequence cannot be turned into an active element by random mutation, but we expect such gains to occur at other positions with a higher chance.

4 Method: Motif Prediction and Evaluation

The *DynScan* package relies on a working scoring algorithm that calculates scores of genomic positions, as for example jPREdictor. In any case, some kind of description of the elements of interest is required. In the example of jPREdictor, a set of motifs has to be provided which are representative for a searched regulatory element. For some elements, such as Drosophila PREs or enhancers, some of the transcription factors involved as well as their DNA binding sites are known. If motifs are not known, however, we have to start there first before we can even consider using jPREdictor in combination with *DynScan*. A method of finding motifs must first be found. Motif prediction is a difficult and error prune task. Although several prediction tools already exist, relying on only a single one of them limits the chances of receiving good results. Here we describe several new methods we have developed to combine motif prediction tools and to evaluate their outcome.

4.1 k-word approach

Enumerative approaches for motif prediction work by looking for all words of a specific length and try to filter out those that are statistically overrepresented. Different tools already exist that aim to extend the approach by introducing degeneration. Our method takes a different approach. Examination the binding sites of some of the known DNA binding transcription factors in PREs shows that concentrating on non-degenerated motifs alone does not inevitably rule out any sensitivity. The core binding site of Pho for example is GCCAT, the GAGA factor is described as GAGAG. Furthermore, a clustering of overrepresented k-words into a degenerated matrix is possible in a subsequent step, as well as using such words as prior knowledge in other motif predictions. Instead of extending the k-words to degeneration as performed by Weeder or YMF, our focus lies on finding motifs for a more specific case. A prediction of regulatory elements performed by jPREdictor based on motifs requires an additional positive and negative training set to calculate each motif's weight. If the training set contains a set of sequences that share a common functionality that is absent in the negative set, this fact could be reflected in the motif prediction as well. The k-word search therefore favors words that are overrepresented in the set of positive sequences in relation to the negative sequences. Furthermore, because the searched motifs are meant to be functional elements of all positive sequences, we want to reward if they occur equally spread over all of those.

The basic idea is now to search for words with a fixed length over the alphabet $\{A, C, G, T\}$ that occur more often in a positive training set than in a negative training set and appear equally distributed in all positive sequences.

Occurrences For each possible k-word m out of all 4^k possible ones for a given k, an occurrence score O can be calculated as

$$O(m) = \log \frac{f(m|M)}{f(m|B)}$$

where f counts the occurrences of motif m in a set of given sequences, normalized by the length of the sequences; M and B denote the model and background.

Distribution As a value representing the distribution of a motif throughout the positive training set, we use the joint entropy.

$$H(m) = -\sum_{s \in M} p \cdot \log(p)$$

where $p = \frac{f(m|s)}{f(m|M)}$. If no motif is present in s, $f(m|s)$ returns a small pseudocount.

The value p gives the probability for each motif m to be located in sequence s. If a motif occurs with the same frequency in all sequences of the positive training set M, the joint entropy becomes maximal. In general the value gets higher the more sequences in the set contain the motif to an equal amount. If the motif occurs $n-$times in all sequences but for example $5n-$times in a single sequence in the set, the entropy value will be lower as if the motif occured only $n-$times in that sequence as well. We made this decision based on the assumption that all sequences in the set are representatives of the same functional element, sharing a common structure of functional motifs. The ideal motif we want to find occurs therefore equally often in all members of the positive set. In real application data, the ideal motif might not be existent, so that different numbers of occurrences might be observed in different sequences. In that case the optimal entropy value will not be reached, but in relation to motifs that are absent in a large fraction of the sequences, the value will be higher. The "quality" of a motif is now determined as a combination of its overrepresentation in the positive set and its relative entropy.

The score S of a motif m is defined as

$$S(m) = O(m) \cdot H(m).$$

Motifs that are most overrepresented in the positive training set and are equally spread through all of its sequences receive the highest overall score. One could think to weight the influence of the two factors $O(m)$ and $H(m)$ by introducing a parametric sum instead,

$$S(m) = w_1 O(m) + w_2 S(m)$$

allowing to arbitrarily place the emphasis on either the motif's occurrences or entropy, by using different values for w_1 and w_2. If prior knowledge is present about the data in an application that requires a specific weighting, this option can be used. Both definitions of $S(m)$ are implemented in the same Perl script.

4.2 Phylogenetic Footprinting Pipeline

In [23] a comparative analysis of 12 Drosophila species is presented that demonstrates the potential of phylogenetic methods. The authors were able to provide novel gene predictions as well as new miRNA genes. Phylogenetic footprinting can also be used to predict novel transcription factor binding sites. In case the element of interest is usually located at homologous positions in multiple species, phylogenetic footprinting approaches may be used that consider local alignments showing higher conservation as more likely motif positions. Two examples of those tools are Footprinter [17] and Phylogibbs [30], both are regular motif prediction tools that have a rewarding function for putative motifs that occur at conserved positions in the input sequences. But even if no explicit phylogenetic footprinting is implemented into a prediction algorithm, we still can make use of the technique by running a prediction on homologous sequences of an element known in at least one species. For example, if biological data show the presence of a regulatory element in the same conserved promoter region within human and mouse, both sequences could serve as input in a single prediction run, expecting the same motif to occur in both species. Furthermore, homologous sequences taken from other species could be added into the prediction input.

General suggestions for dealing with motif prediction tools include [12, 26]:

- Try substantial amounts of input data to minimize effect of biasing sequences.
- Choose multiple tools that are based on different approaches.
- Remove found hits and run prediction again.

- Combine related results into common motif representations.

Additionally, as described before, a motif should be contained in a large fraction of the input sequences. Motifs appearing only rarely in the input sequences are less likely to play an important role in the input sequences' common functionality.

Pipeline

We combine those criteria into a prediction pipeline, that takes large sets of aligned sequences as input, runs different prediction tools, masks the hits in each run, clusters predicted motifs, and finally checks whether the motifs appear in a substential number of the input sequences (Figure 4.2).

The pipeline itself is implemented in Perl. The input sequences are received by querying the UCSC genome database for Multiz17way [91] alignments via HTTP. For a set of given positions in one sequence, the alignments are requested and a configured set of sequences is kept. For example, if the provided positions are from the human genome, only sequences from mammals can be kept. Although it can vary how many alignments are available, depending on the query positions, constraints can be defined such as that at least human and mouse have to be present.

The alignments are prepared for three different prediction tools. MEME is chosen exemplarily for an approach based on Expectation Maximization. The alignments are converted into gapless sets of five sequences each to meet MEMEs constraints on input data. Larger sets of sequences between 2 kb and 4 kb each lead to a running time of more than 24 hours on a 2.4 GHz dual-opteron system, so restricting the number of alignments to five each time reduces time consumption. The hits reported in the first run are masked and the search is repeated in order to allow the detection of lower scoring motifs.

Additionally to MEME the tool Footprinter is run on the input data. It works directly on alignments and rewards motifs showing a high conservation in all provided species. However, in each step only one alignment can be handled at a time. The parameters are set for a search for motifs with up to one mutation within each branch of the phylogenetic tree. Filtering of low complexity regions is enabled to avoid a bias towards repetitive regions. Even in a single alignment up to 15 predicted motifs can be observed giving sequences of 4 kb each. The total amont of distinct motifs is smaller, several reported hits belong to the same matrix. Therefore related motifs have to be combined in common matrices before a further processing of the predicted motifs can take place. A direct clustering of all Footprinter runs, especially if the input set contains 100 sequences or more, is beyond computational limits for the clustering tools MATLIGN and Phyloclus. The problem can be avoided by introducing a two-step

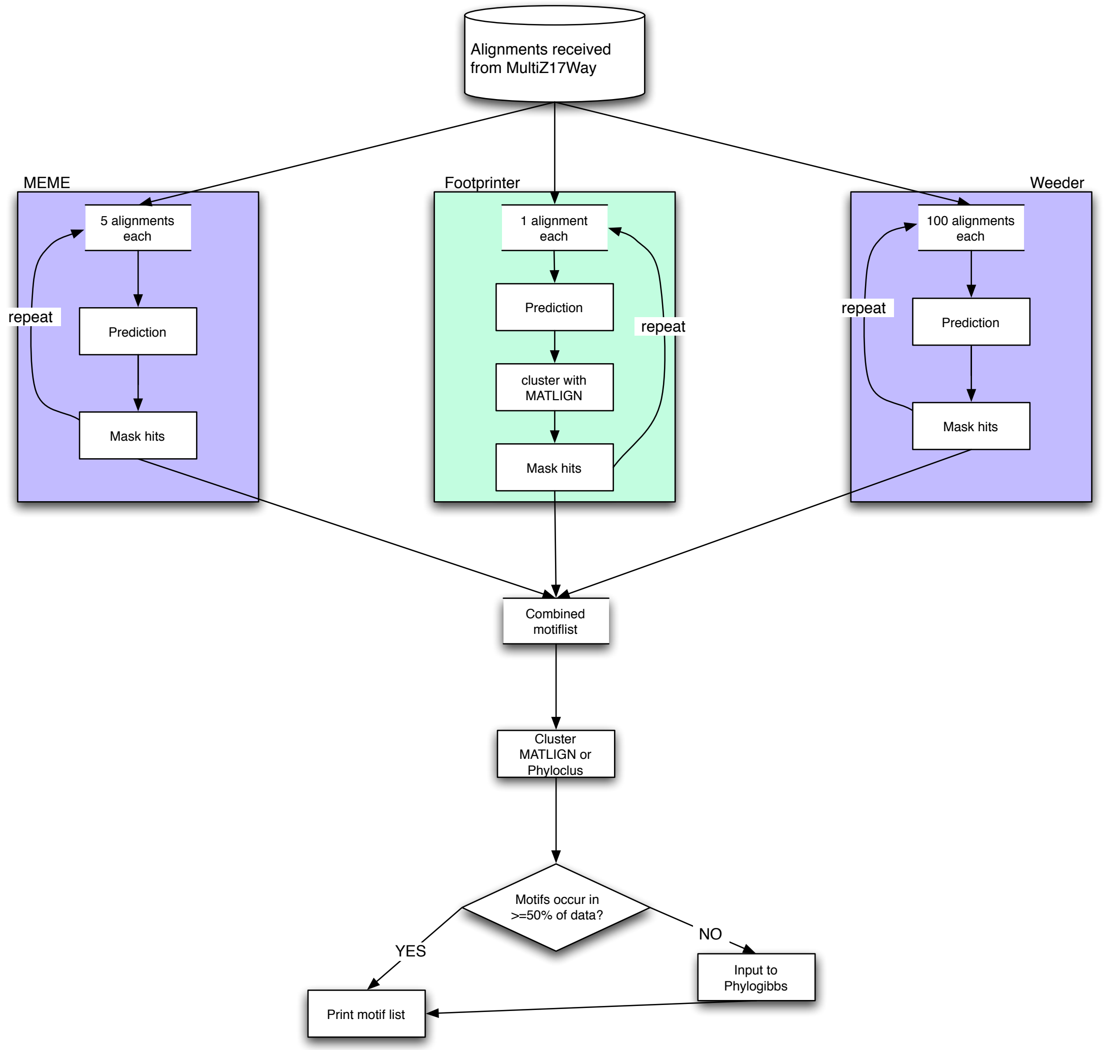

Figure 4.1: Phylogenetic pipeline. Input alignments are called from UCSC database and transformed into input for MEME, Footprinter and Weeder. Motifs are predicted and hits masked. In case of Footprinter, hits of each alignment are clustered. Prediction and masking are repeated. All results are combined in a single list and clustered. If clustered motifs can be found in $\geq 50\%$ of the input, motif is reported as hit. Otherwise Phylogibbs is called to align additionally predicted motifs to the existing clusters.

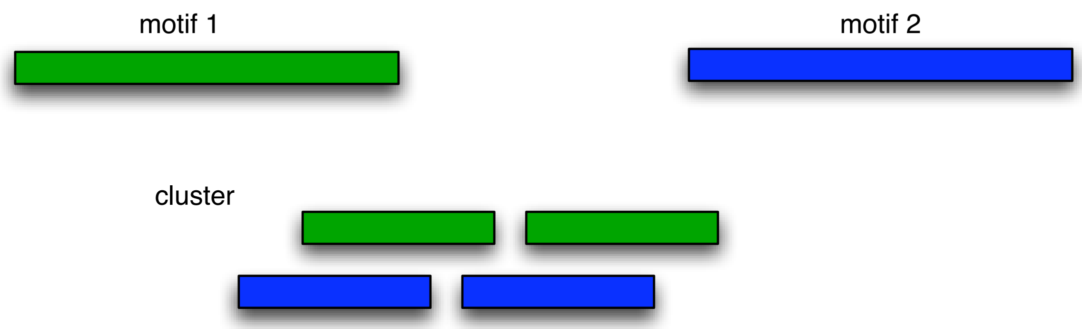

Figure 4.2: Allowing gaps in MATLIGN clustering could lead to matrices that are not matched by original motifs. Neither motif 1 nor motif 2 are matching the cluster directly in this example. If the matrix is treated like any position probability matrix and used in additional software, the elements building the matrix would not be found.

clustering. First, each alignment's prediction is clustered separately to get an overlap free list of predicted motifs for each prediction run. The resulting clusters are then clustered a second time. As done in the MEME part, the hits are masked in each step and the prediction is iterated.

Moreover, Weeder is used as an example for an enumerative approach that has been shown to be complementary to probabilisitc predictions [12, 25]. The input alignments are transformed into gapless FASTA files of up to 100 alignments at once. The parameters are set to search for motifs of length six, eight, and ten with none, two, or three mutations allowed. Again, hits are masked and the search is repeated.

All three methods report a list of putative motifs. Because in each prediction step only subsets of the input have been processed, multiple reported motifs could be part of the same transcription factor binding site representation. The different motifs are converted into the input format of the clustering tools MATLIGN and Phyloclus. MATLIGN is set to not allow spacers in the input motifs to prevent the creation of matrices matching the input motifs only if gaps are included (Figure 4.2). Additionally the input motifs are extended by “N” at the beginning and end. Otherwise it can happen that the input motifs are not matching the generated cluster (Figure 4.3). .

The result is two motif lists which are overlap free in the sense of the clustering algorithm. Each new motif is checked whether it meets the requirement that it should be contained in at least 50% of the input sequences. If so, it is reported as a putative result. If not, the motif can either be a false positive and is not related to the input sequences' biological functionality, or the motif lacks of sensitivity. In the latter case, trying to align other subsequences of the input data could lead to a more sensitive

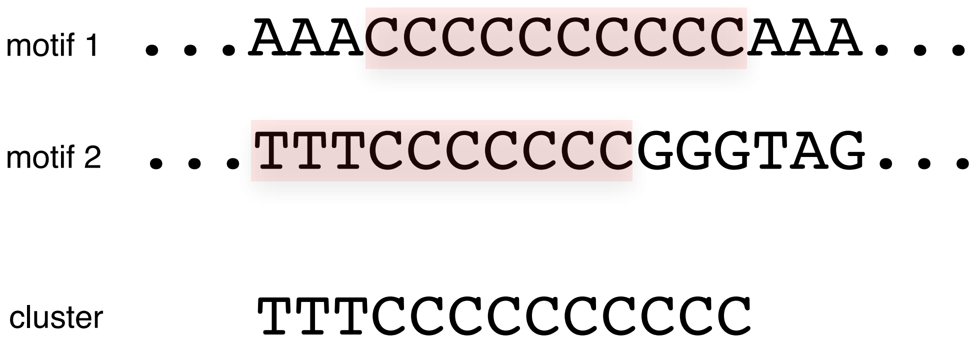
motif 1 ...AAACCCCCCCCCCCAAA...
motif 2 ...TTTCCCCCCCGGGTAG...
cluster TTTCCCCCCCCCC

Figure 4.3: MATLIGN clustering problems: Red regions show motifs predicted in two sequences that are clustered by MATLIGN. Because the flanking regions of the motifs are lost, the clustering could lead to a matrix that don't match the original motifs. This can be prevented by adding 'N' to the motifs before the clustering is done.

matrix.

Prior knowledge

Phylogibbs offers the possibility to take a set of matrices as input sequences to which predicted motifs are aligned if possible. All motifs that do not occur in at least 50% of the input data serve as prior knowledge for a Phylogibbs search. The motifs provided to Phylogibbs have to be of the same length as the motifs that are searched. The motif length parameter is set to 10 in our pipeline. Longer motifs are just cut-off while shorter motifs are padded by ‘N’ per default by Phylogibbs. The clusters created by Phyloclus have the same length as the input sequences, no further processing is required. Because the MATLIGN output varies in length and a restriction to the first 10 positions does not reflect the positions of the individual cluster members, another criterion has to be chosen. The continuous part taken from each input cluster has to be a subsequence of as many cluster members as possible.

For example, the best MATLIGN result of the Footprinter output is shown in Figure 4.4. The bars are the positions of the individual clustered motifs, the common region is chosen as the longest overlap.

Phylogibbs is run on each alignment seperately, provided with the clustered motifs as input. For each of those clusters it is counted in how many of the input sequences it has been found in the Phylogibbs run. Again, the threshold is set to 50%.

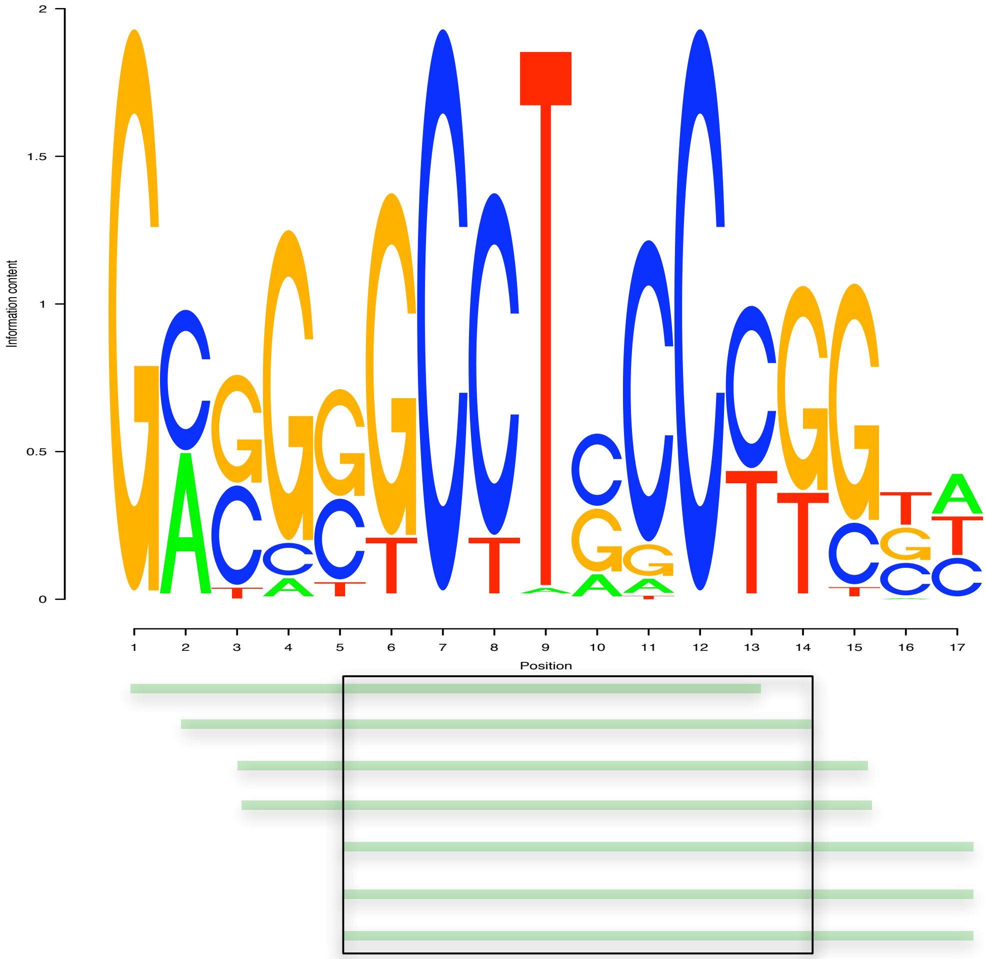

Figure 4.4: Selecting input as prior knowledge for Phylogibbs from MATLIGN clusters. The motif example shows largest Footprinter cluster of length 17. The sequences of length 10 each, that are combined into the cluster are indicated as green bars. To get again a cluster of length 10 that can be used as Phylogibbs input, the 10 bp subsequence (rectangle) is chosen to cover as many cluster elements as possible.

4.3 Evaluation algorithm

The problem of assessing a set of motifs in order to distinguish specific and significant motifs from statistically random hits or motifs that do not have a positive impact in further use, occurs in most cases that involve a motif prediction. Therefore in this chapter a generalized method for choosing potential motifs that can be used in a motif based prediction of elements of interest is presented. The motif based prediction software for regulatory elements we choose is jPREdictor.

In 2003, when the first PREs in Drosophila were predicted genome-wide, some of the involved transcription factors and their binding motifs were known, namely Zeste, GAF, and Pho. In addition the En1 site has been used, which has been found to be conserved in the Engrailed PRE. But as long as not all binding motifs for an element class of interest are identified, it is necessary to decide on a set of motifs to use in a prediction. This problem occurs in the motif based prediction of any kind of elements, let it be PREs in flies or mammals, or enhancers.

Because the calculation of the cut-off to be used in predictions is based on the number of occurrences found for each score in a prediction run in a null-model, the choice of the motif-set directly influences the cut-off and therefore has an effect on the sensitivity and specificity of the whole prediction. Every motif or multi-motif within the motif-set with a positive or negative weight raises or lowers the cut-off because each motif has a chance greater zero to occur even in a random sequence. Every time such a motif is found in a scored window, the score is influenced. Trivially, the smaller the motif set is, the smaller the chance of finding a member of the set in a random sequence. Every motif inserted into the motif set influences the cut-off (if the weight is not 0). Thus a motif set restricted to important motifs only is preferable. The basic considerations are not affected if the cut-off calculation is based on a probabilistic model instead of the empirical statistic. It does not matter whether the chance to find a motif pair is determined by random data or is calculated directly. The reason is that a positive weight only gives the relation between the numbers of occurrences in a model and in a background, independent of the actual expectation of an occurrence. For example, if a motif pair that is built of two single nucleotides ‘A’ and ‘C’ occurs only slightly more often in the model with a weight of only ‘0.2’, the cut-off will still be very high and completely bias the prediction output. The A:C motif pair occurrences will contribute almost exclusively to any score in either the background or in model sequences. Real element specific motifs that receive high weights of maybe ≥ 8 but occur only a few times in each of the real sequences, will be masked and have little to none impact on the scoring.

This observation leads to the conclusion that a good motif-set should only contain those motifs or multi-motifs whose contribution to the separation of positive and negative training sets is stronger than their effect on the cut-off calculation. The question to be answered now, is how can we evaluate the effect of each multi-motif on the prediction and keep only those that give a benefit. Furthermore, multi-motifs that are kept in the motif list are meant to be characteristic for the elements in the positive training set and therefore their weights should not be altered too strongly if single sequences are removed out of the training set.

Furthermore, motif pairs are weighted according to their occurrences in a positive set in relation to a negative set. In case of jPREdictor, the distribution of motif pairs within a set is not considered. Basically, the weight is the same if a motif pair occurs one time in each of n sequences in the model as if it occurs n times in one sequence of the model. In practice, there might be small differences due to pseudo-counts, but the general description demonstrates the potential problem. In our evaluation step, a weight will be put into relation to the total number of hits and the number of sequences it occurs in.

Given a set of potential motifs and a positive (M) and negative (B) training set, the optimal motif set to separate the two training sets is derived as the result of a multi step pipeline. In the first step all motifs are combined to all possible motif pairs P, then each pair is weighted according to the training sets. The "net" effect of each pair on the output of a set of sequences S is calculated as

$$\sigma_{S_m} = \sum_{s \epsilon S} O_{(s,m)} \cdot W_m$$

where W_m is the weight of motif pair $m \in P$ and $O_{(s,m)}$ is the number of occurrences of that pair in sequence s of training set S. The motif pair with the biggest proportion of the positive training set's score is therefore $argmax_m(\sigma_{M_m})$. Each motif pair in σ_M is now either kept in the motif set or removed, depending on its relation to the maximum score and to its value in σ_B. For positive σ_{M_m}, a motif pair is chosen to be part of the motif set if

$$(\sigma_{M_m} \geq \frac{maxscore}{\mu_p})\, and\, (\sigma_{M_m} \geq \mu_s \cdot \sigma_{B_m}).$$

The μ parameters are scaling parameters that determine the minimum fraction of the maximal or minimal score (μ_p, μ_n) and the factor between scores in model and in background (μ_s).

A negative value means a negative weight, indicating that a motif pair is underrepresented in the positive training set in relation to the negative set. Negative weights lower the score at positions that are less likely to be an element of interest and therefore increase the prediction's specificity. Motif pairs for which σ_{M_m} is negative are kept if

$$(\sigma_{M_m} \leq \frac{minscore}{\mu_n})\, and\, ((\sigma_{M_m} \geq \mu_{s\prime} \cdot \sigma_{B_m})\, or\, (\sigma_{B_m} \geq \mu_{s\prime} \cdot \sigma_{M_m}))$$

This step restricts the motif set to the motif pairs providing the highest information content. The values μ_s and $\mu_{s\prime}$ give the factor between the model and background σ-values for a motif pair m if the values are positive (μ_s) or negative ($\mu_{s\prime}$). The parameters μ_p, μ_n, μ_s and $\mu_{s\prime}$ are chosen depending on the actual oberserved distribution of the σ-values.

Robustness test

Although positive weights reflect an overrepresentation of motifs in the positive training set, a positive weight alone is not sufficient to prove that the motifs are symptomatic for the element of interest. Because the distribution within the training sets is not considered in the weighting step, strong motif repeats in one single sequence can bias the weight. In the last pipeline step decribed above a motif set has been built and can be used to score the training sets. To avoid the time consuming cut-off calculation at this point in the pipeline, a preliminary cut-off is set to $2\cdot((highest\, score\, in\, background)+1)$. The 1 is added in case no motif pair is found in the background which would lead to a cut-off of 0.

The common way in statistics to determine potential bias in a given set of data points is resampling, either in form of bootstrapping or as a jack-knife test. The latter is applied to our set to rule out single sequences that bias σ_M. In [92] a similar method is used in relation to motif predictions. They removed random single sequences from the prediction process to observe their effect on the outcome. In our case, all sequences scoring above the preliminary cut-off within the positive training set are removed in the next pipeline step. These sequences are detected by the prediction, using the motif set derived in the first steps. Biased weights due to motif repeats in single sequences will result in a high score above the cut-off in only few positive sequences. If we remove the sequences that contain repeats, the motif weights will be drastically decreased, while the weights of motifs that are equally overrepresented in most sequences will show robustness. Repeating the pipeline multiple times eventually removes the biasing motifs and sequences. As the result, either a motif set and positive training set are found that are representative for the element of interest, or no single motif sets can be found that can predict a large amount of elements.

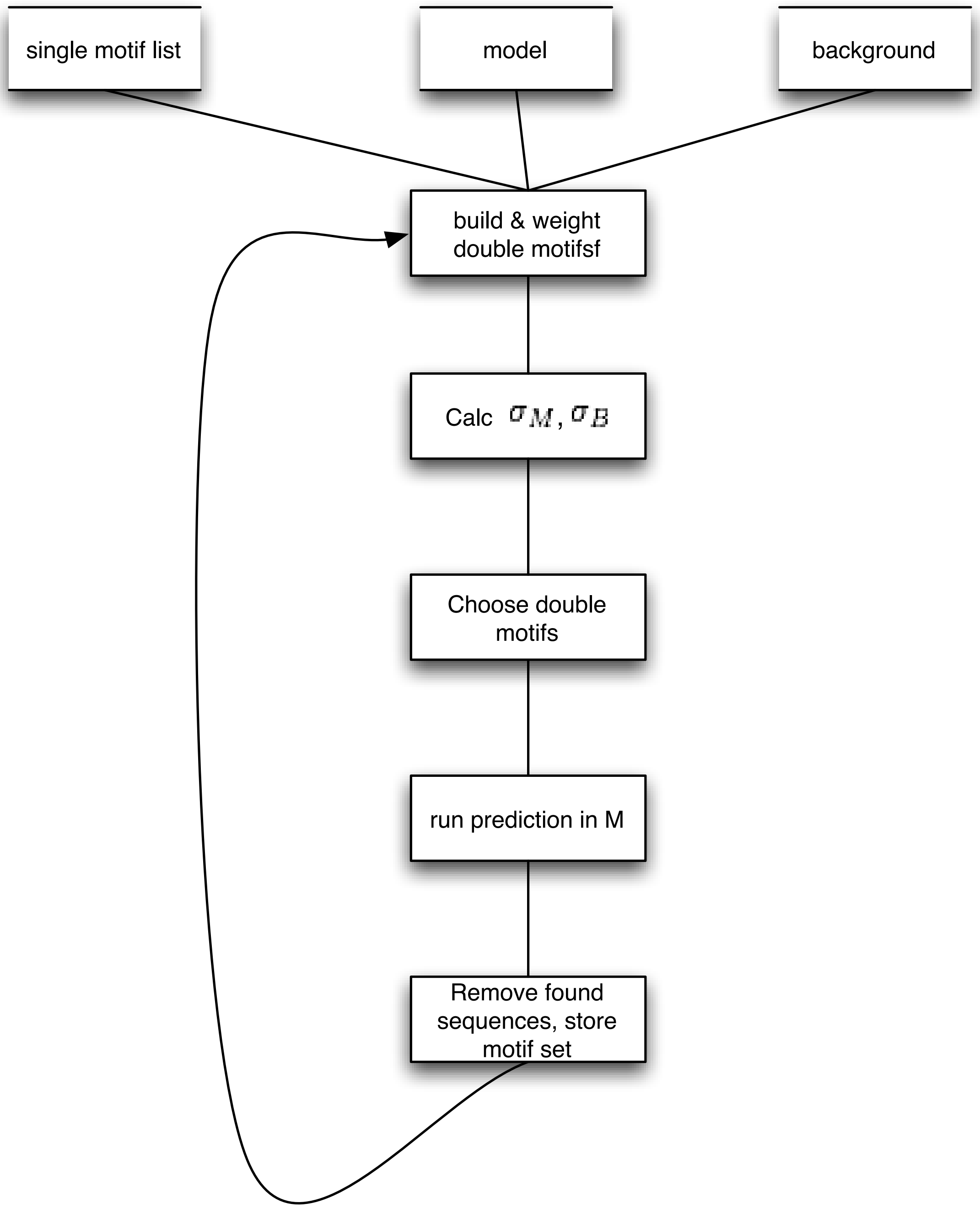
single motif list
model
background
build & weight double motifsf
Calc σ_M, σ_B
Choose double motifs
run prediction in M
Remove found sequences, store motif set

Figure 4.5: Motif evaluating pipeline: A list of single motifs, a set of positive sequences (model) and a set of negative sequences (background) are provided. Motifs are combined to pairs and weighted, σ-scores are calcutated for model and background, highest scoring motif pairs are selected and used in a prediction in the model. Positive sequences scoring above preliminary cut-off are removed from model and pipline is iterated.

5 Application and Results: Fly PREs

5.1 The search and results

The prediction of Polycomb Response Elements in Drosophila seems to be well suited as an application for the *DynScan* package. The original PREdictor already showed in 2003 [11] that a prediction of PREs in *Drosophila melanogaster (D.mel)* is possible. As described in Section 2.2.4, the specificity had the main emphasis in the prediction, which let to reduced sensitivity. Because a large number of the 167 predicted PREs have been experimentally verified, the confidence in the chosen significance level is justified. Furthermore, the later developed jPREdictor allows to base the extended prediction on the same motif and training sets while further making use of additional features. Thus the *DynScan* requirements are fulfilled, jPREdictor serves as an algorithm that can score whole genomes if a set of motifs and training sets for motif weighting are provided. These parameters are taken from the original PREdictor publication, because they already have been partially validated by ChIP and transgenic experiments.

For the comparative dynamic search, a set of species needs to be supplied. The existence of PREs in *D.mel* is well confirmed, whereas there is as yet no information about PRE locations and functions inside other Drosophila species. The results that the *DynScan* method can provide may give further insights into PRE functionality and give a better understanding about the essential functional parts inside the PRE regions.

The species set we used for *DynScan* consists of five Drosophila species (see tree in Figure 5.1). Additional to *D.mel* (version 4 [94]), we added *Drosophila pseudoobscura* (*D.pse*) (version 2.0 [94]), *Drosophila yakuba* (*D.yak*) (version 1.0), *Drosophila simulans* (*D.sim*) (version 1.0) and parts of *Drosophila erecta* (*D.ere*) (Comparative Assembly Freeze 1 [95]). Four of the species are part of the melanogaster subgroup, while *D.pse*, as part of the *obscura* subgroup, is ˜25 million years apart. *D.mel*, *D.yak*, *D.sim*, and *D.pse* are the only species within the 12 Drosophila species that have been assembled to chromosomes. The results presented in this chapter are gained from these species unless stated otherwise. The recently published Comparative Assembly Freeze 1 was chosen for specific analysis, as will be shown later in this chapter, but

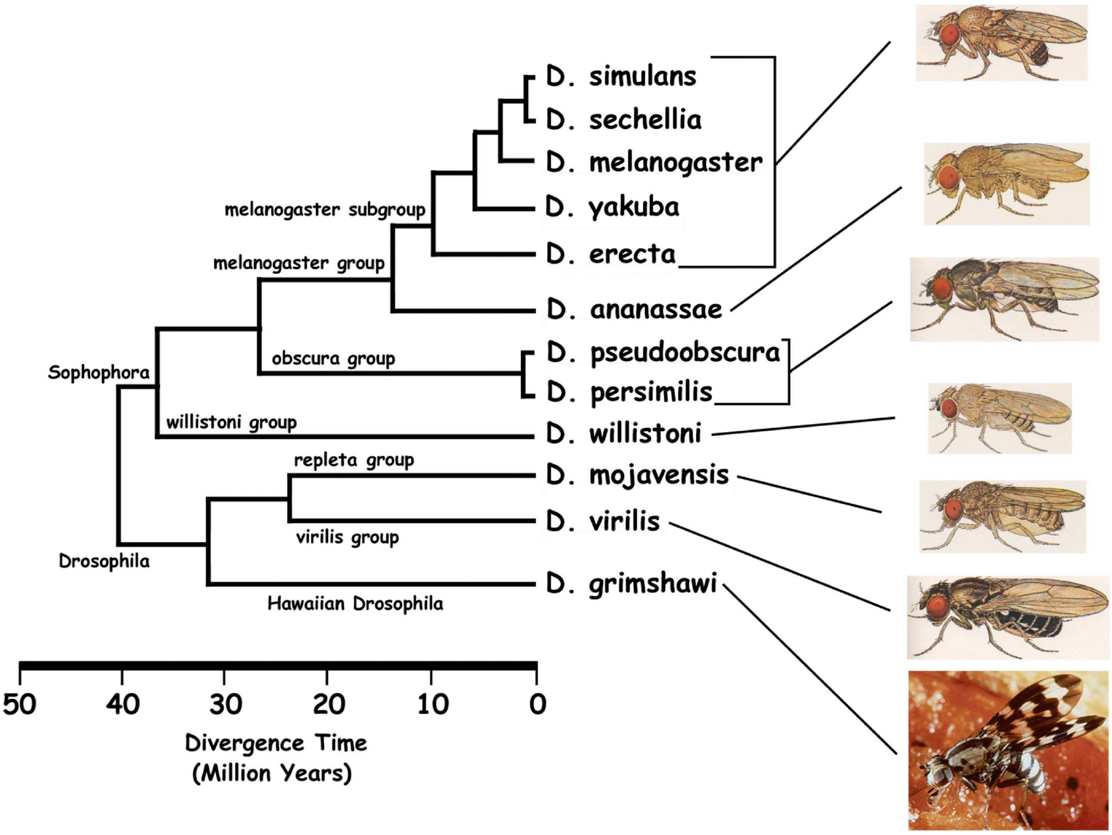

Figure 5.1: Phylogenetic tree of Drosophila [93].

did not make its way into the complete *DynScan* run, because *DynScan* requires long sequences around the orthologous positions in order to dynamically search the loci. Unfinished assemblies in form of contigs are avoided to minimize assembly related bias on the results.

In the first step, all species are scored by jPREdictor, using the same parameters as the original PREdictor. Because analyses of the effect of different parameters were not done in 2003, the scoring parameters are reconsidered. The original parameters (window width 500, shifted by 100 in each step) predict 167 PREs above the genome-wide cut-off of 157 that corresponds to an E-value of 1. The maximum distance between two single motifs to be taken as a motif pair is 220. The distance directly influences the weights, because even if the same two motifs occur in the model and the background, the distance will most likely be different. Because of the low number of sequences in the training sets, using individual weights for each motif pair to maximize each weight could lead to biased results. A thorough statistical analysis of distance distribution is not within the scope of this work. The chosen step width of 100 on the other hand, could lead to a miss of motif pairs if one motif falls into another window than the other one. The smaller the step width, the less likely this happens. Using a step width of 1 however, results in lower running time, and more important, the size of the score values increases by factor 100. As can be seen in the database layout described in Section 3.4.1, for each scored position three integers (run, species, position), one float (score) and one set of characters with variable length (chromosome) have to be stored. This sums up to 32 bytes, plus the overhead for the database index on the chromosome field. At a step width of 1, the database data would be at least 32 times the genome length. The tradeoff between the most accurate step width and the space overhead is at a step width of 10.

The species *D.mel*, *D.pse*, *D.sim*, *D.yak* and *D.ere* are added to the *species* database table and are referenced by an identification number. The chosen species contain all fully assembled genomes, as well as *D.ere*, which has been chosen because it is evolutionary located between *D.pse* and the melanogaster species and about 80% of the genome is available in scaffolds with a length of at least 50 kb. All species are scored by jPREdictor with a window width of 500 and a step size of 10. The motif set and the training set are the ones used by PREdictor. All scores and positions are copied into the database. For the E-value calculation, 20 GB of random data following the *D.mel* nucleotide distribution are created and scored, the number of hits for each integer score is also stored in the database. This parameter setting is identified by run number 1. Out of this data the entries for the *static_element* table are created, at an E-value of 1 we now predict 201 PREs genome-wide in *D.mel*, in contrast to the originally found 167 in 2003.

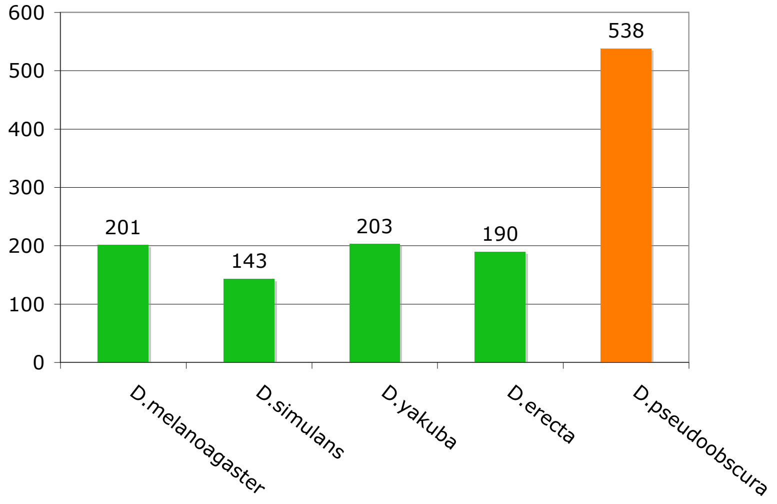

Figure 5.2: Number of predicted PREs in different species. The cut-off is set to 157, which reflects an E-value of 1 in *Drosophila melanogaster.* Motif and trainging sets are taken from original publication [11], window size is 500, shifted by 10.

Genome-wide predictions

For each scored species, the same cut-off (in this case 157) is used to build the genome-wide prediction table. The numbers of predicted elements inside the different species differ drastically, as shown in Figure 5.2. Inside the melanogaster subgroup, the number varies around 200; the smaller amount of 143 in *D.sim* could in part be explained by the fact that 18% of the genome are either annotated as 'N' or are provided in forms of random reads for each chromosome. In *D.pse*, 538 hits can be predicted, which cannot be explained by the difference in genome-length. Although the estimated *D.pse* genome is about 18% longer than the *D.mel* genome [96] (while [95] estimate a 9% longer genome), the effect on the cut-off is small. The chosen cut-off reflects an E-value of 1 in *D.mel* and thus an E-value of 1.18 in *D.pse*, so according to our statistic, at least 536 elements are true positive. The question to be answered now is whether the increase of hits in *D.pse* has a biological background, or shows a weakness in the prediction's parameters or the statistic. Due to the lack of different decriptions of the functional transcription factor binding sites in other species than

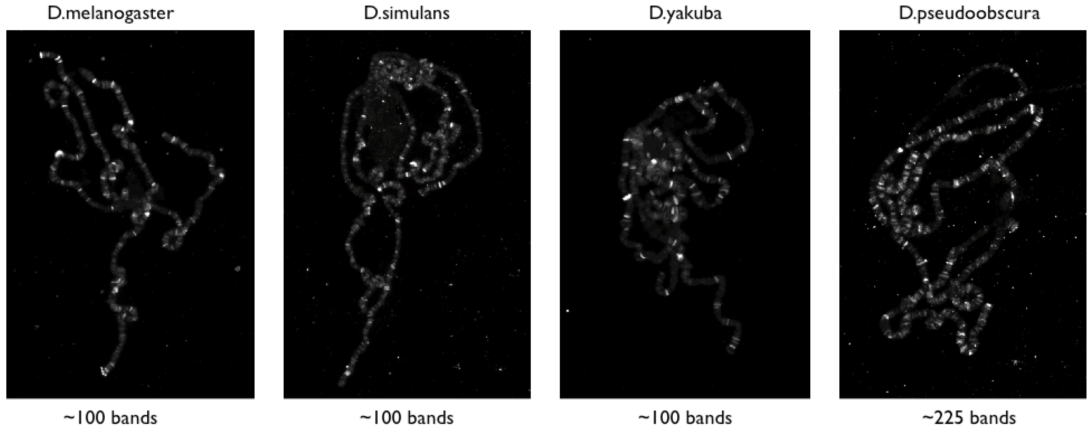

Figure 5.3: Immunofluorescence on polytene chromosomes with anti PC antibodies in four species. Pictures taken from [97].

D.mel, we took the same motif set under the assumption that the proteins and even more the binding sites are conserved within all Drosophila species.

To validate this hypothesis, our collaboration partner has performed different biological tests [97]. Pictures of Polycomb protein distribution made by immunofluorescence using anti Polycomb antibody on polytene chromosomes prepared from third instar larvae of four species show different numbers of bands in different species (Figure 5.3). The number within the melanogaster subgroup varies arounf 100. In *D.pse* on the other hand, about 220 bands have been identified. Because the resolution of this kind of experiments does not allow an exact determination of single PRE locations, the total number of PREs within each species remains unknown. One band of Polycomb binding can contain multiple PREs, as for example the Bithorax complex, in which we predict 7 PREs in *D.mel*. The number of bands is not only consistent with our prediction, but also matches the number of detected bands of H3K27 methylation (Figure 5.4). As can be seen in Figure, , a strong increase of Polycomb binding as well as histone methylation can be observed in *D.pse* in relation to the melanogaster subgroup species. The experimental data indicates an increase of regions that contain PREs of 72% between *D.mel* and *D.pse*. Our prediction detects 178% more PREs in *D.mel* than in *D.pse*.

The difference in PRE numbers gives a first impression of potential dynamics in the evolution of such elements. We refer to this observation as the first type of evolutionary plasticity.

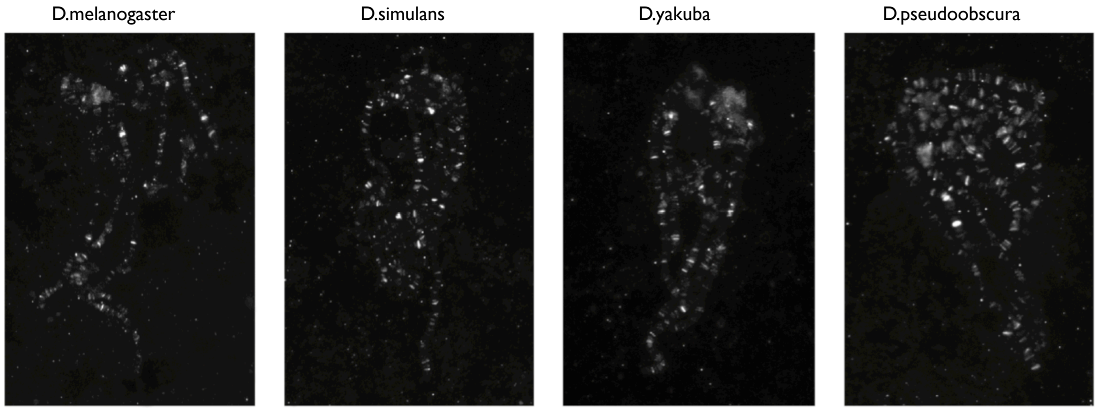

Figure 5.4: Immunostaining with anti histone H3 K27 me3 antibody on four species.

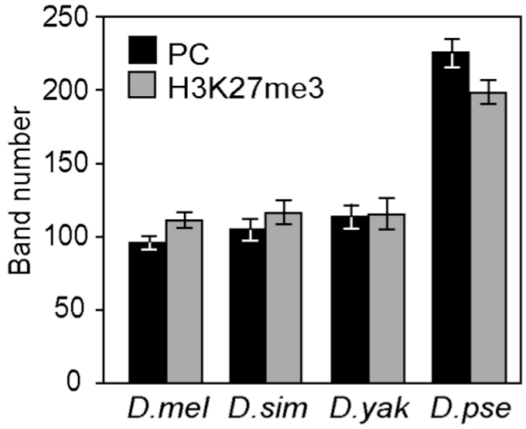

Figure 5.5: Average band numbers from polytene chromosomes as shown in Figure 5.3 and Figure 5.4. Error bars give standard deviation.

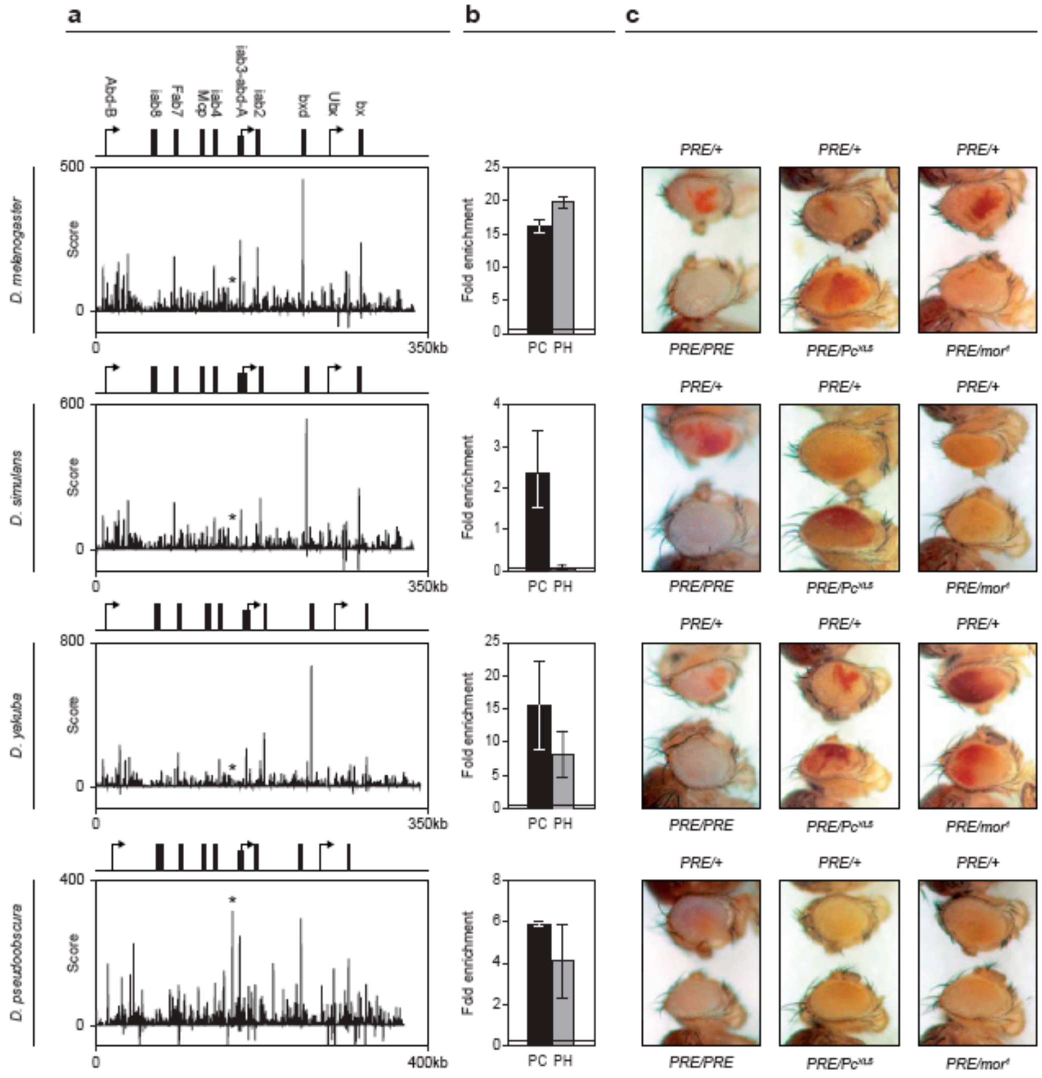

Figure 5.6: Experimental validation of PRE prediction: **a**) Prediction scores of Bithorax complexes. Positions of homeotic genes promoters and experimentally verified PREs are given above each plot. Orthologous positions of additionally predicted *D.pse* PRE at iab3 marked by asterisks. **b)** ChIP analysis of Polycomb (PC) and Polyhomeotic (PH) enrichments on the bxd PRE in embryos of four species. Error bars indicate standard deviation. Horizontal lines represent mean enrichments of negative control fragments that were present at detectable levels in all samples. **c)** Transgenic reporter assays for 1.6kb centered around PRE bxd. Top row: the *D.mel* bxd PRE26 was cloned upstream of the miniwhite reporter gene. The eyes of transgenic flies show variegation, pairing sensitive silencing (left panel), loss of silencing in a PcG mutant background (middle panel) and loss of activation in a trxG mutant background (right panel). Bottom three rows: miniwhite reporter constructs containing 1.6kb of *D.sim*, *D.yak* and *D.pse* sequences orthologous to the *D.mel* bxd PRE were injected into *D.mel* embryos. All show behaviour similar to the *D.mel* bxd PRE. All figures are from [97].

Bithorax complex

The role of the three genes inside the Bithorax complex as well as the location of some PREs within regulatory regions of BX-C are well studied (Section 2.2.2). As a positive control experiment, we score the Bithorax complexes of four species and see a significant peak at the position of the bxd [98] PRE (Figure 5.6). By ChIP experiments, it can be shown that Polcycomb group proteins are enriched at the bxd position in all four species. This shows that the prediction parameters based on *D.melanogaster* PREs lead to high score peaks at positions in other Drosophila genomes, and that these positions are actually bound by Polycomb group proteins. Additional transgenic reporter array experiments in *D.mel* show typical PRE behavior of the predicted bxd PRE from all species (pairing sensitive silencing, variegation, loss of silencing in PcG mutants). All those experiments show that the transcription factor binding sites used in the motif set are in fact conserved in other Drosophila species as well. Thus the requirements of the dynamic search are fulfilled - the jPREdictor in combination with the motif and training set can be used to score and to predict PREs in different Drosophila species.

Dynamic PRE search

The dynamic search by *DynScan* is performed in each direction between the four species *D.mel*, *D.pse*, *D.sim*, and *D.yak*. Because *D.ere* is only available in form of assembled scaffolds instead of complete chromosomes, it is not considered for the genome-wide dynamic search. As search radii we took 1 kb, 10 kb, and 20 kb. The chosen 1 kb reflects the case that the analog is directly located at the orthologous position. Predicted PREs are between 500 and 1000 base pairs long, the orthologous sites however can be shorter, depending on the locations of the hsps.

The first category of dynamically predicted PREs contains those that are located within 1 kb around the orthologous site. In this case, the found PREs occur at conserved positions. As can be seen in Table 5.1, the cut-off required to get an E-value of 0.005 is calculated as 70. If the number of orthologous positions is 200, (remember we have 201 genome-wide hits in *D.mel*), the overall E-value is 1, instead of 157 in the genome-wide search.

The next radius is set to 10 kb to detect analogous PREs that are not overlapping with the orthologous site, but are still close enough to allow the categorization of assumed functional analogs. The cut-off used in case of 200 searches is 102.

In the last step the search radius is set to 20 kb. The analogous PREs have now moved more than 10 kb, but are is still required to be inside the same locus. Because the hypothesis is that some *cis*-regulatory elements can be at different positions in different genomes, but the functional analog is at least located inside the same locus,

E-value	1 kb	10 kb	20 kb
1	14	34	42
0.1	34	61	70
0.01	61	95	104
0.005	70	104	114
0.001	94	127	139

Table 5.1: Cut-off scores for different E-values and different sequence length. Ranges are given as radius. Scores calculated empirically in 20 GB random data, created as 0-order Markov Chain and following *D.mel* nucleotide composition.

the dynamic search stops after a radius of 20 kb is searched. The confidence that the nearest predicted element is the query element's functional analog decreases with growing distances. That explains why no more dynamic steps are considered after 20 kb and the nearest genome-wide predicted PRE is taken.

In addition to the position of the statically and dynamically predicted PRE, the nearest located genes are stored in the database. The table *genes_positions* contains the positions of all known genes in *D.mel* and *D.pse*, received from FlyBase. The database layout has been described in Section 3.4.1. For each predicted PRE in any of the two species, the nearest gene is determined. If the PRE is overlapping directly with the transcription start site of a gene, this gene is taken as the nearest one and the distance is set to zero, otherwise the distance to the two nearest transcription start sites is calculated. Because a PRE that is located in intergenic regions could regulate a gene located upstream as well as downstream, the two nearest genes are stored.

The distribution of distances between the homologous site and the nearest predicted functional analog is supposed to give first insights into the dynamic search's outcome. Because in the first step the cut-off is the lowest, even weak signals within 1 kb around the homologous position are detected. Despite the lowered cut-off, it can be observed that in multiple cases no score of at least 70 can be found within a 2 kb window. Considering that although the homologous positions contain by definition BLAST hsps and therefore highly conserved subsequences, in some cases not enough motif pairs are conserved to gain a score peak.

The distance definition takes into account the fact that the center of the region spanned by the hsps can vary from the query sequence's center. The distance is calculated as the absolute value of the difference (D1-D2) between the distances between the center of sequences defined by the hsps and the center of the PRE in the query sequence (D1) and the center of the PRE in the target sequence (D2) (Figure 5.7).

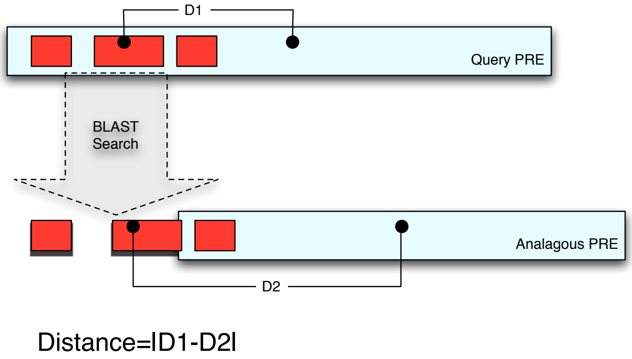

Figure 5.7: Distance definition: The blue bars indicate the PRE in the target and the query species. The contained red bars show the single HSPs of a BLAST search, starting with the query PRE. D1 is defined as the difference between the center of the query PRE and the center of the BLAST query hits. D2 is defined as the difference between the center of the BLAST hits and the center of the nearest predicted PRE. The overall distance is the difference between D1 and D2.

The numbers of dynamically predicted PREs in *D.mel* in relation to the distances are presented in Figure 5.8. The percentage of analagous PREs in *D.mel* in close position (<1 kb) to the homologous regions is above 80%, if the dynamic search is performed within the melanogaster subgroup. Still the predicted PREs do not cover the exact BLAST positions in most of the cases, although a slight overlap can be observed. In general, we can say that sequences homologous to PREs in one species lack PRE features in the other, suggesting that they have lost (or never acquired) PRE functionality. Instead, we can detect PREs in non-homologous regions nearby that are assumed to be functionally analogous. The hypothesis of the dynamic search is that *cis*-regulatory elements may not be sequence conserved but occur in orthologous loci in related species. The observed behavior of PRE evolution supports this hypothesis and gives a first impression of high evolutionary dynamics in the development of these *cis*-regulatory elements. The distance distributions reflect the phylogenetic distances between the species, inside the melanogaster subgroup the vast majority of assumed functionally analogous elements are in close proximity to the homologous positions, whereas the situation is different if the divergence between the species increases.

Although more static PREs are predicted in *D.pse* and therefore more loci are dynamically searched in the melanogaster subgroup species, even the absolute number of hits within a 1 kb radius is lower. Out of 531 *D.pseudoobcura* PREs, only 120 lead to a functional analog within 1 kb in *D.mel*. Furthermore, in 41% of the cases no analog

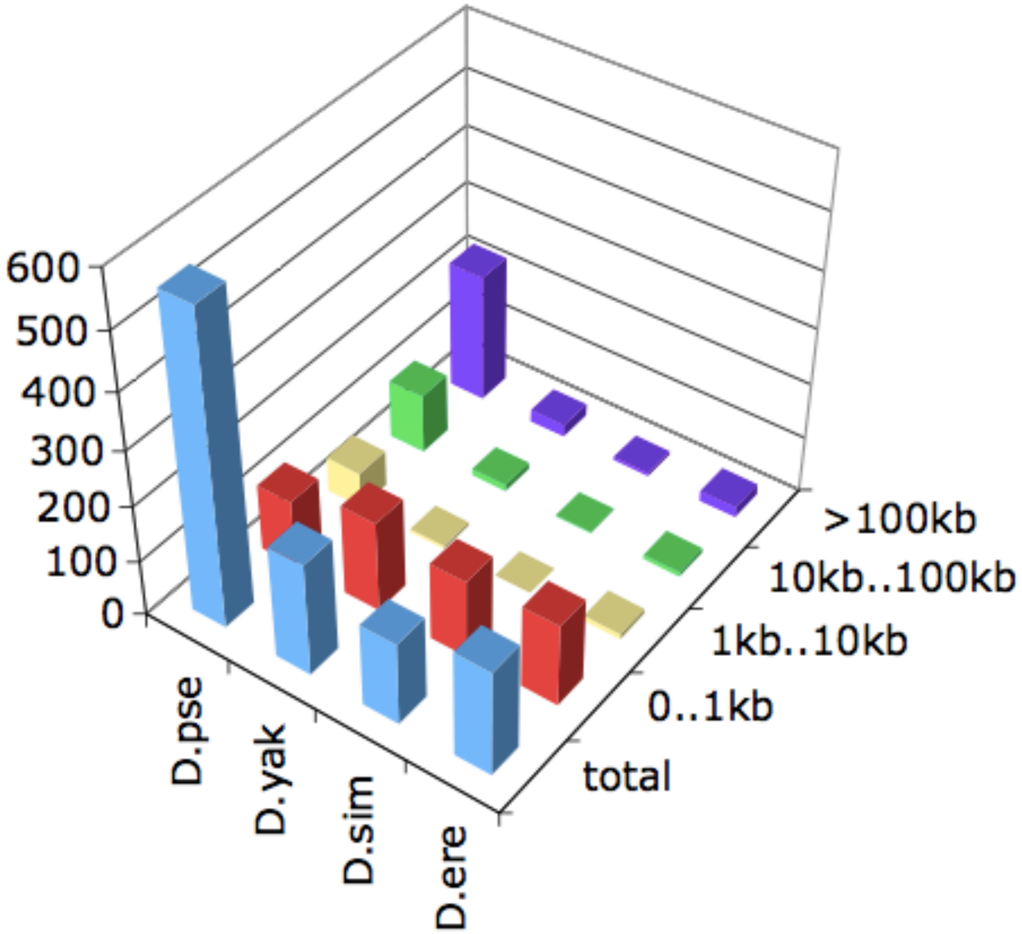

Figure 5.8: Results of dynamic search in *D.melanogaster*. The plot shows the result of the dynamic searches starting from *D.pse*, *D.yak*, *D.sim* and *D.ere*. The numbers of PREs predicted within *D.mel* at different radii are given in different colors (1 kb=red, 10 kb=yellow, 20 kb=green, >20 kb=dark purple).

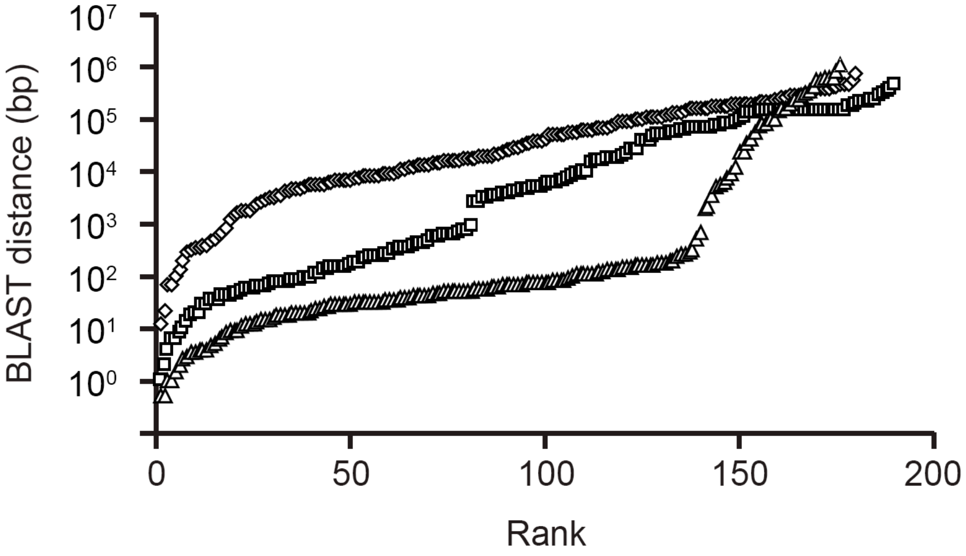

Figure 5.9: Distances between orthologous regions and predicted analogs. Triangles: genome-wide predicted static *D.melanogaster* PREs versus *D.yakuba* analogs. Boxes: genome-wide predicted *D.melanogaster* PREs versus *D.pseudoobscura* analogs. Diamonds: 1 kb sequences randomly chosen from the *D.melanogaster* genome versus *D.pseudoobscura*. The numbers of random sequences on each chromosome equal the numbers of predicted static PREs on that chromosome.

can be found within a 20 kb radius, indicating that no analog may exist. This is consistent with the results of the static prediction and the polytene pictures – although the overall number of PREs in all Drosophila species remains unknown, the data indicate at least twice as many PREs in *D.pse* than in *D.mel.*

As an additional control for the *DynScan* method, in Figure 5.9 the results of dynamic searches of the 201 static *D.mel* PREs against *D.yak* and *D.pse* are shown, together with 201 randomly chosen and BLASTed *D.mel* positions in *D.pse*. The results show that closer phylogenetic divergence leads to smaller distances (80% of the *D.mel* PREs have an assumed functional analog within 1 kb in *D.yak*) and that random data lead to only very few hits within 1 kb (less than 10%).

Validation of Evolutionary Plasticity

Especially between *D.pse* and the melanogaster subgroup species we can find footprints of strong evolutionary dynamics in the PRE sequences. It can be seen that the positions of analogous PREs appear to be independent of direct sequence conservation, an observation we refer to as the second type of evolutionary plasticity. In order to validate the prediction, our collaboration partner performed ChIP on chosen examples. The experiments are designed to test the presence of protein binding at the positions of predicted analogous PREs as well as the absence of binding at the corresponding homologous positions. The examples are chosen to cover the different observed aspects of PRE divergence. Experimental targets are chosen from the following categories of dynamically predicted PREs:

- PREs that show no movement in any species

  There are nine examples in which a PRE can be predicted directly at the homologous positions in all four species that shows a genome-wide significant prediction score.

- PREs that are located next to but not directly at the homologous position in other species

  In this category lies the majority of PREs predicted by *DynScan*. A PRE can be found in one species by a genome-wide search but the homologous positions in other species show no score peak. Instead, within 1 kb-20 kb distance a PRE can be predicted with an either locally or even genome-wide significant score.

- PREs that show no functional analog in other species

  Especially if the search is based on static *D.pse* PREs, multiple cases do not show an assumed functional analog within 20 kb in melanogaster subgroup species. One example is the additional peak in the *D.pse* Bithorax complex that does not have a counterpart in the other species (Figure 5.6 a).

Example bxd

The bxd PRE inside the Bithorax complex regulates the *Ultrabithorax* gene. It is located at the same position in all four tested species and shows a score above 157 in every case. The ChIP experiments detect an enrichment of Polycomb group proteins at all positions, and an enrichment of Polyhomeotic proteins in all species except for *D.sim* (Figure 5.6 a,b). The bxd PRE serves as a positve control, because it is characterised in *D.mel*. The detected enrichment in *D.mel* shows that the ChIP experiments lead to correct results.

Example eyes absent

The next example of a PRE that is conserved in its position is the PRE located inside the first intron of the gene *eyes absent* in *D.mel* and *D.pse* (Figure 5.10). The PRE is

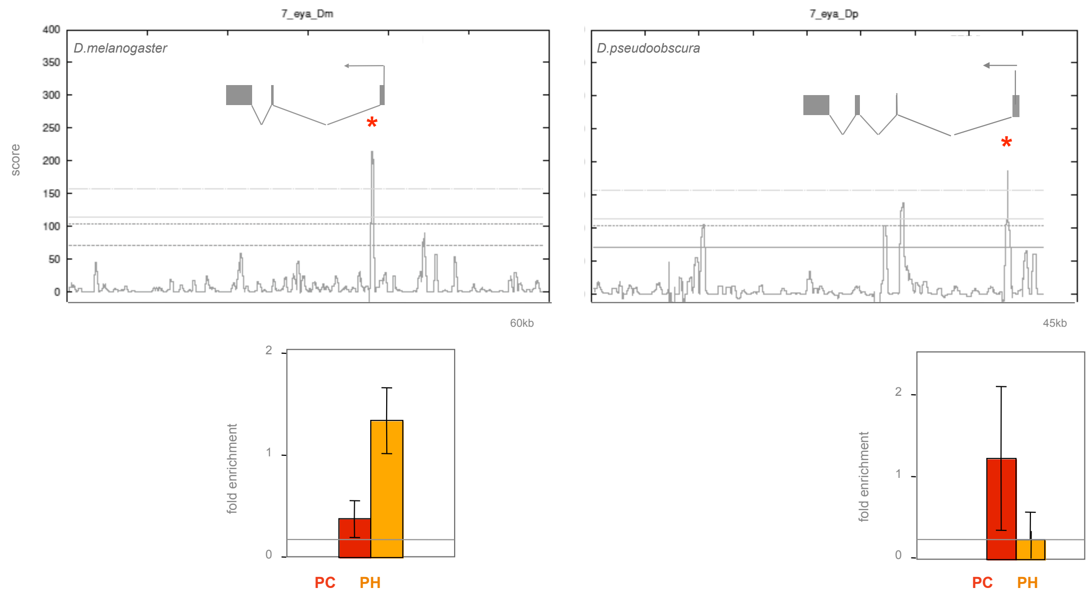

Figure 5.10: PRE inside the *eyes absent* gene. The red asterisks indicate homologous positions. Polycomb protein (PC) and Polyhomeotic protein (PH) are tested by ChIP.

also predicted at the homologous positions in the other analyzed species, but the ChIP experiments have been perfomed only on the former two. By ChIP an enrichment of Polyhomeotic group proteins can be detected in *D.mel* while in *D.pse* Polycomb group proteins are enriched. Both experiments indicate a functional PRE at the same position in the same gene, supporting the prediction results.

Example spalt major

The PRE inside the first intron of the gene *spalt major* in *D.mel* scores above the genome-wide cut-off and shows strong ChIP enrichment of Polycomb group proteins (Figure 5.11). The same situation can be observed in the other two melanogaster subgroup species. In *D.pse* however, no genome-wide significant score is present. Instead, in a distance of 782 bp a score of 107 is found, which is significant in the dynamic search step only. As can be shown by ChIP, a strong Polycomb group protein enrichment is detected, but additionally Polyhomeotic is enriched in *D.pse*, which is not the case in the other tested species. This could indicate a different structure of the *D.pse* PRE, which recruites different proteins than the other species' PREs. The assumption is consistent with the fact that only minor parts of the *D.pse* PRE sequence are conserved in the other PREs (5.11 d).

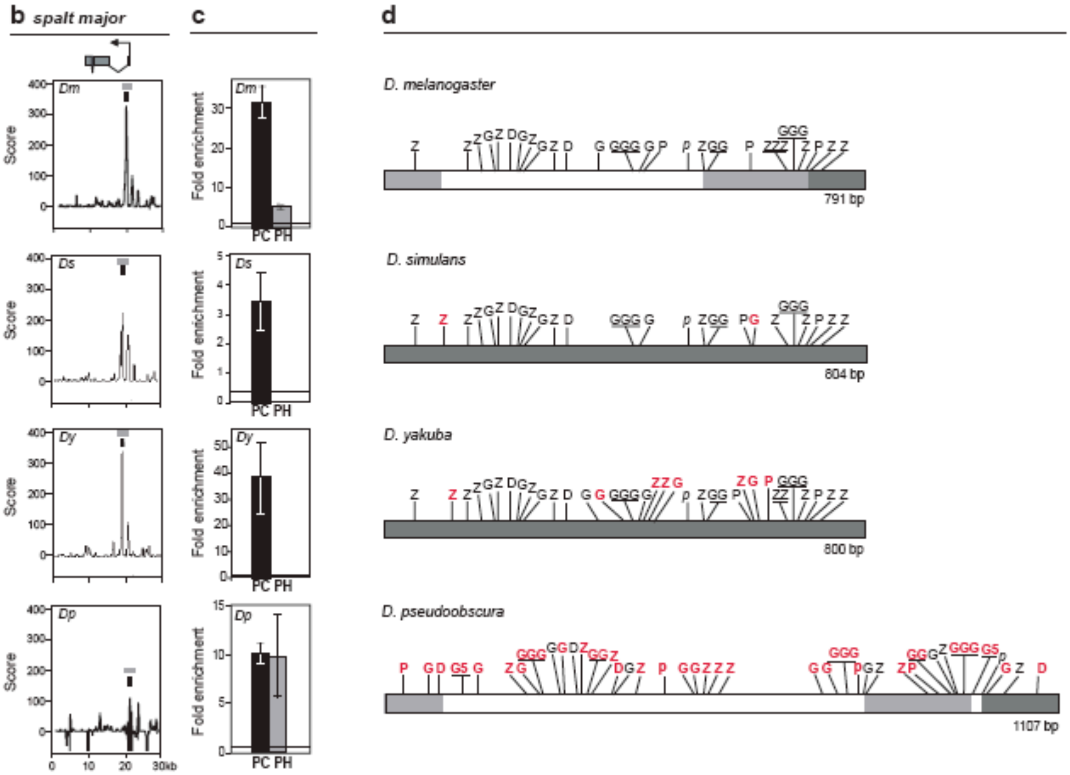

Figure 5.11: PRE close to *Spalt major* (*salm*) promoter. **b)** Prediction score plots. First row shows *D.mel*, second *D.sim*, third *D.yak*, and fourth *D.pse*. Black boxes indicate PCR fragments used for real time PCR detection in ChIP analysis shown in **c**, grey boxes indicates region shown in **d**. **c)** PC and PH ChIP for predicted position, error bars give standard deviation. **d)** The core *D.mel* PRE and the orthologous regions from the other three species are shown. Conservation between *D.mel* and *D.pse* is marked on the diagrams for these two species: Dark grey: regions of over 70% identity. Light grey: 50%-70% identity. Motif positions are indicated above the figure. Motifs shown in red on *D.sim*, *D.yak* and *D.pse* are not present in the *D.mel* PRE. D = Dsp1; Z = Zeste; G = GAF; P = Pho extended site (PF or PM; p = Pho core site (GCCAT). Underlined motifs indicate overlapping runs of motif separated by 2 bases. G5, G7, G9 = 5, 7 or 9 GA repeats. Figure taken from [97].

Example trachealess

In the last example the PRE could only be found by the dynamic search in *D.pse*. Although the position was not exactly conserved, the sequences at least partially overlapped. The PRE of the gene *trachealess* is an example for stronger PRE movement (Figure 5.12). In *D.mel*, *D.sim*, and *D.yak* the PRE is located in the promoter region of the gene. In all three cases the score reaches the cut-off of 157. ChIP experiments show strong Polycomb enrichment in all melanogaster subgroup species and additionally Polyhomeotic enrichment in *D.yak*. The ChIP results in *D.pse* at the homologous position show only a weak signal of Polycomb binding and no significant Polyhomeotic folding is present. A dynamic search started from either of the melanogaster subgroup PRE positions leads to the prediction of a functionally analogous PRE in *D.pse*, located around 4.4 kb upstream of the transcription start side inside an intron. At this position strong enrichments of Polyhomeotic proteins can be detected, indicating a functional PRE site. The score of the PRE falls below the genome-wide cut-off, but shows significance in the scored 10 kb radius around the homologous site.

To evaluate whether this site in fact is the functional analog to the promoter PREs in the other species instead of just a second PRE, the homologous positions of the *D.pse* PRE within *D.mel*, *D.yak* and *D.sim* are also tested by ChIP. Neither significant Polycomb nor Polyhomeotic group protein fold enrichments can be detected in any of the species. The data support the theory that only one PRE regulates the *trachealess* gene in the four species. The location of the PRE seems to be not conserved, it can be located inside the promoter as well as inside an intron, depending on the species.

Example decapentaplegic (dpp)

The *trachealess* example illustrated the case of two different positions in close distance to each other in different species, regulating the same gene. The PREs of the gene *decapentaplegic* serve as an example for three different PRE positions, which can be identified by the dynamic search within the four different species (Figure 5.13). In *D.pse* a PRE is predicted genome-wide directly within the promoter region of the gene *dpp* (position 3). The dynamic search finds a locally significant score peak 5 kb upstream (position 2) in *D.yak* and 10 kb upstream (position 1) of the promoter in *D.mel* and *D.sim*. ChIP detects enrichment of Polycomb as well as Polyhomeotic proteins at position 3 only in *D.pse*. In the other three species neither a score peak nor protein binding can be observed, indicating that position 3 functions only in *D.pse* as a PRE. At position 2, Polycomb enrichment can be found again in *D.pse* and additionally in *D.yak*, strikingly consistent with the prediction results. Furthermore, Polyhomeotic proteins also bind position 2 in *D.pse*, which is not the case in *D.yak*, but in *D.mel* and *D.sim*

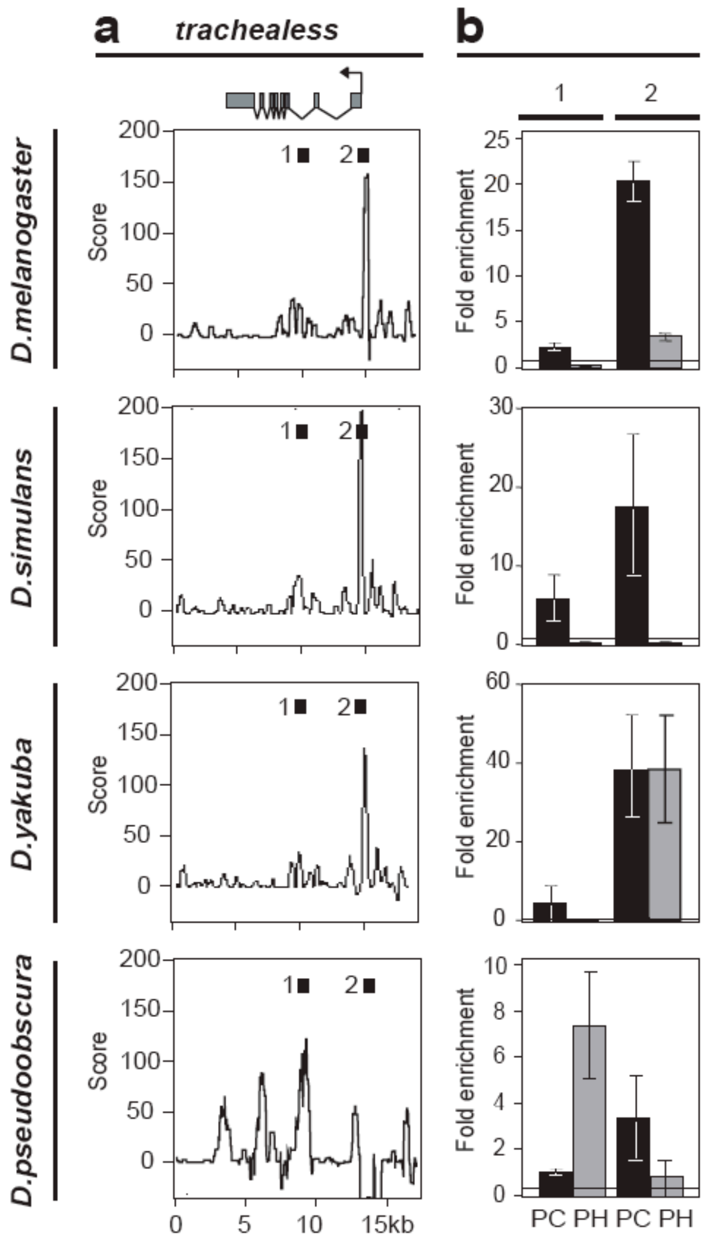

Figure 5.12: Predicted PREs within promoter region and intron of *trachealess* gene. **a)** Prediction scores in four species, gene location shown above. Two positions marked by black boxes are orthologous positions that are predicted in at least one species. **b)** ChIP for PC and PH on the positions 1 (first column) and 2 (second column) marked in **a)**.

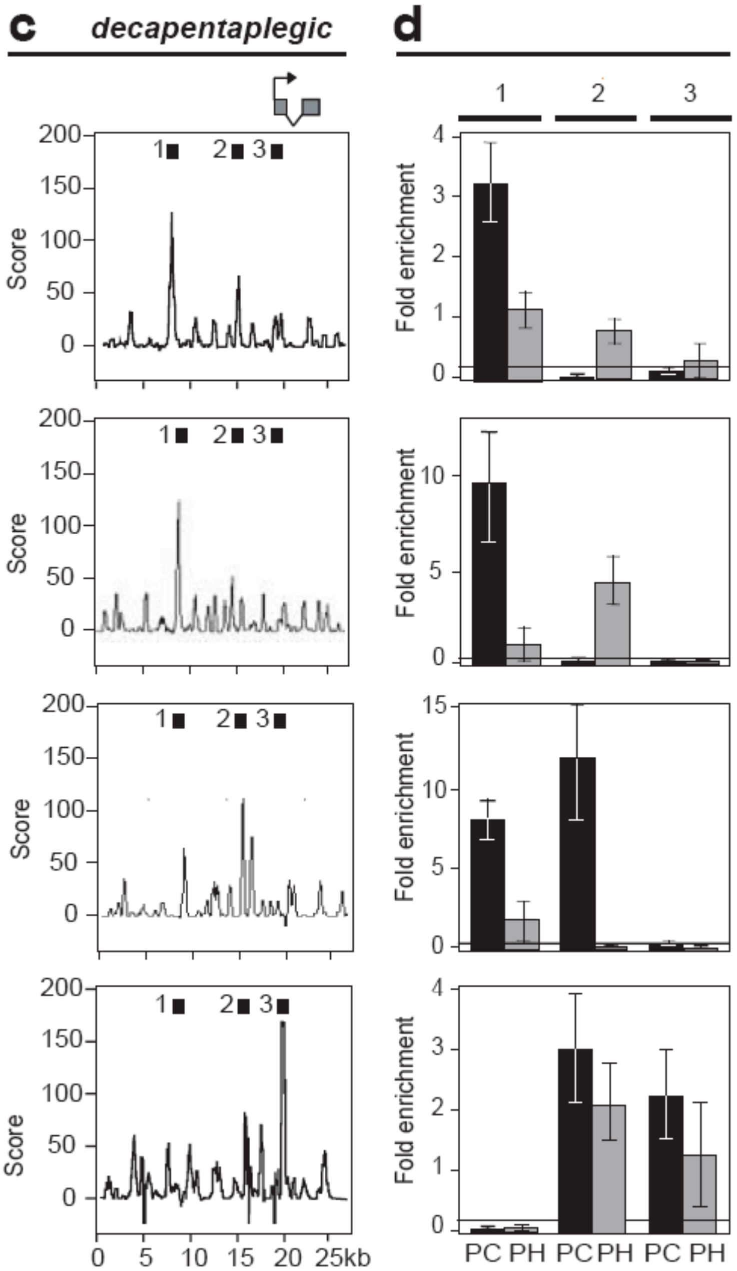

Figure 5.13: Three predicted PRE positions upstream of the gene *decapentaplegic* in four species. Score plots and ChIP data for marked positions shown as described in Figure 5.12.

at least weak binding can be seen. While the position 2 seems to be a strong functional PRE in *D.pse* and *D.yak*, the weak binding in the *D.mel* and *D.sim* species gives no clear answer. In both species the prediction score is below the dynamic cut-offs, but at position 1 a significant peak is present exclusively in these species. According to Polycomb enrichment, position 1 shows PRE functionality in all melanogaster subgroup species but not in *D.pse*. Although it cannot be answered clearly which PREs are the functional analog of which other species' PREs, the complete region demonstrates the possible dynamics in PRE evolution. It might be the case that two PREs are necessary in the upstream region of the *dpp* gene in order to achieve a regulatory effect. A deeper analysis of the possible evolutionary processes is presented in Section 5.2.

Motif turnover

According to the previous observations and the experimental results, functionally analogous PREs can occur in close distance to the homologous positions or even up to several kilobases away. Nevertheless, in some cases a PRE is predicted or even validated at conserved positions in several genomes. Following the idea of phylogenetic footprinting, the motifs, as the functional elements, should show stronger sequence conservation than non-functional regions in the PREs. Two aspects are considered in the next step. First, does the sequence conservation provide any information about functional elements inside the PREs, and second, do the same motifs occur in the same order in all species? Even the PREs that are predicted genome-wide at the exact same position in multiple species show different score levels, which indicates that at least minor changes in the motif composition are to be expected. As an example serves the PRE of the gene *eyes absent*. As already shown, the PRE is validated in *D.mel* and *D.pse*, located at the same position in both species. Additionally, at the homologous positions in *D.yak* and *D.sim* a PRE can be found as well by a genome-wide prediction. Thus the same region in four species scores above the genome-wide threshold. In Figure 5.14 a multiple alignment of the extended PRE regions in the four species is shown, created by MLAGAN [99] and visualized by the Vista browser [100]. The sequences are taken from our prediction and extended by 1 kb in each direction to cover non-PRE regions as well. The alignments are shown in relation to the *D.mel* sequence. Except for a few gapped regions, the complete sequences of *D.yak* and *D.sim* show conservation of at least 70%. In *D.pse*, due to the bigger evolutionary distance, only four positions reach a 70% conservation level, out of which two are located outside the PRE sequence. According to these data, a significant increase of sequence conservation at functionally active sites cannot be observed. Nevertheless, the two conserved positions in the *D.pse* PRE sequence require a deeper analysis.

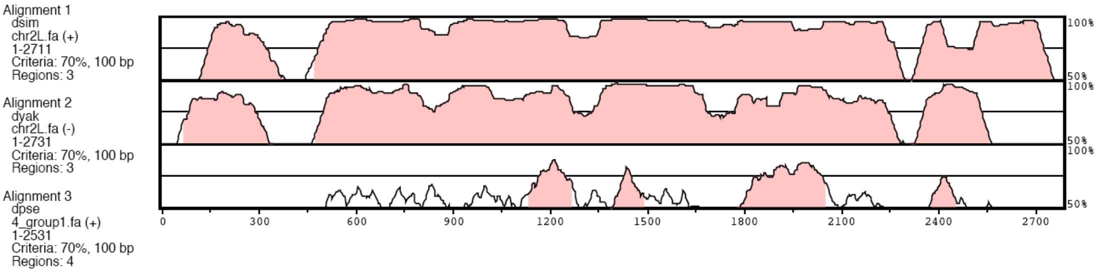

Figure 5.14: Conservation plot of the extended (±1 kb) *eyes absent* PRE in relation to *D.melanogaster* in the species *D.simulans*, *D.yakuba* and *D.pseudoobscura*. Colored regions show a conservation ≥70%. Y-axis gives conservation in %, X-axis refers to *D.melanogaster* sequence position. Alignment created by MLAGAN.

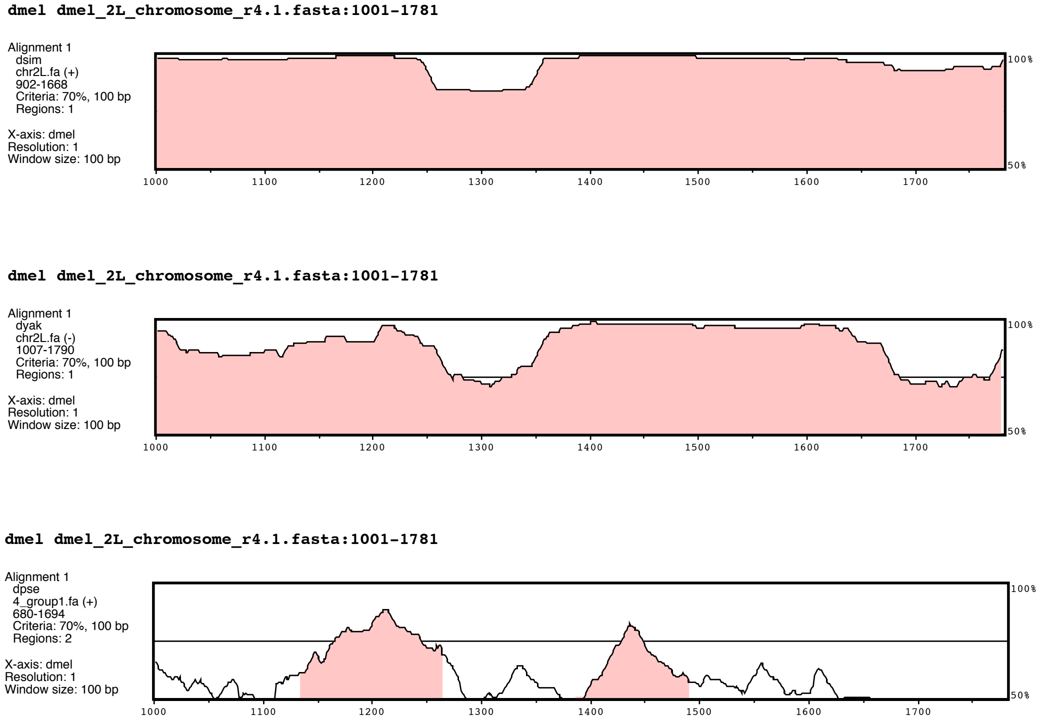

Figure 5.15: Conservation plot of the *eyes absent* PRE regions in relation to the predicted *D.melanogaster* PRE sequence in the species *D.simulans*, *D.yakuba* and *D.pseudoobscura*. Colored regions show a conservation ≥70%. Y-axis gives conservation in %, X-axis refers to *D.mel* sequence position. Alignment created by MLAGAN.

A concentration on the *D.mel* PRE region (Figure 5.15) reveals again high conservation throughout the complete regions in *D.yak* and *D.sim* whereas in *D.pse* only two single peaks in the center show high conservation. This indicates that major parts of the *D.mel* PRE are not highly conserved in *D.pse*, although functional elements are expected to be located in these parts as well. To evaluate this theory, a comparison of the motif occurrences inside the four PRE sequences is required (Figure 5.16).

The different motifs in the motif set and additionally Dsp1 and SP1/KLF are drawn as colored boxes at the corresponding positions in the four species. The PRE sequences are displayed as bars, arranged in relation to the *D.mel* PRE. A comparison of the motifs between *D.mel* and *D.sim* shows high overlap of motif order and position. Motifs that are not exactly aligned in the figure can still occur at the same position because gaps are not represented. The only noticeable difference is an additional Dsp1 binding site occurrence in *D.mel*. The sequence in *D.yak* is reverse complementary to the other species. Compared to *D.mel* it can be observed that again most motifs are conserved in their order, but a few differences are present. The first Sp1/KLF motif in *D.mel* is exchanged by a Dsp1 motif, furthermore three Dsp1 motifs occur between the Gaga-Zeste motif pairs and the nearest Pho cluster. In *D.yak* one Dsp1 motif is located on the other side of the Gaga-Zeste cluster. Instead of the Zeste in *D.mel* an additional Dsp1 can be found in *D.yak*. Furthermore the distances between several motifs are different in *D.yak*.

In general, it can be observed that inside the melanogaster subgroup, as reflected by a conservation of at least 70% in the whole sequence, the motifs are conserved to a major extent. The situation changes if the *D.mel* motifs are compared to *D.pse*. Except for two small Pho clusters and an adjacent Dsp1 motif, hardly any similarity can be detected. This introduces the third type of evolutionary plasticity, even in the rare cases where the PREs occur at homologous and therefore at least partially sequence conserved sites, the motif composition can be independent of the conservation. The rearrangement of motifs is called motif turnover. Although it has been known so far that motif turnover exists [19, 20, 21], the extent to which it can be found between melanogaster subgroup species and *D.pse* has not been observed earlier.

5.2 Explaining evolutionary plasticity : pre-PREs?

The observation of described types of evolutionary plasticity gives rise to the question at which point in evolution a region gained or lost its PRE functionality. To answer this question we want to analyze multiple aspects of this topic. First, does our prediction in multiple species follow the phylogenetic tree of the Drosophila species? Second,

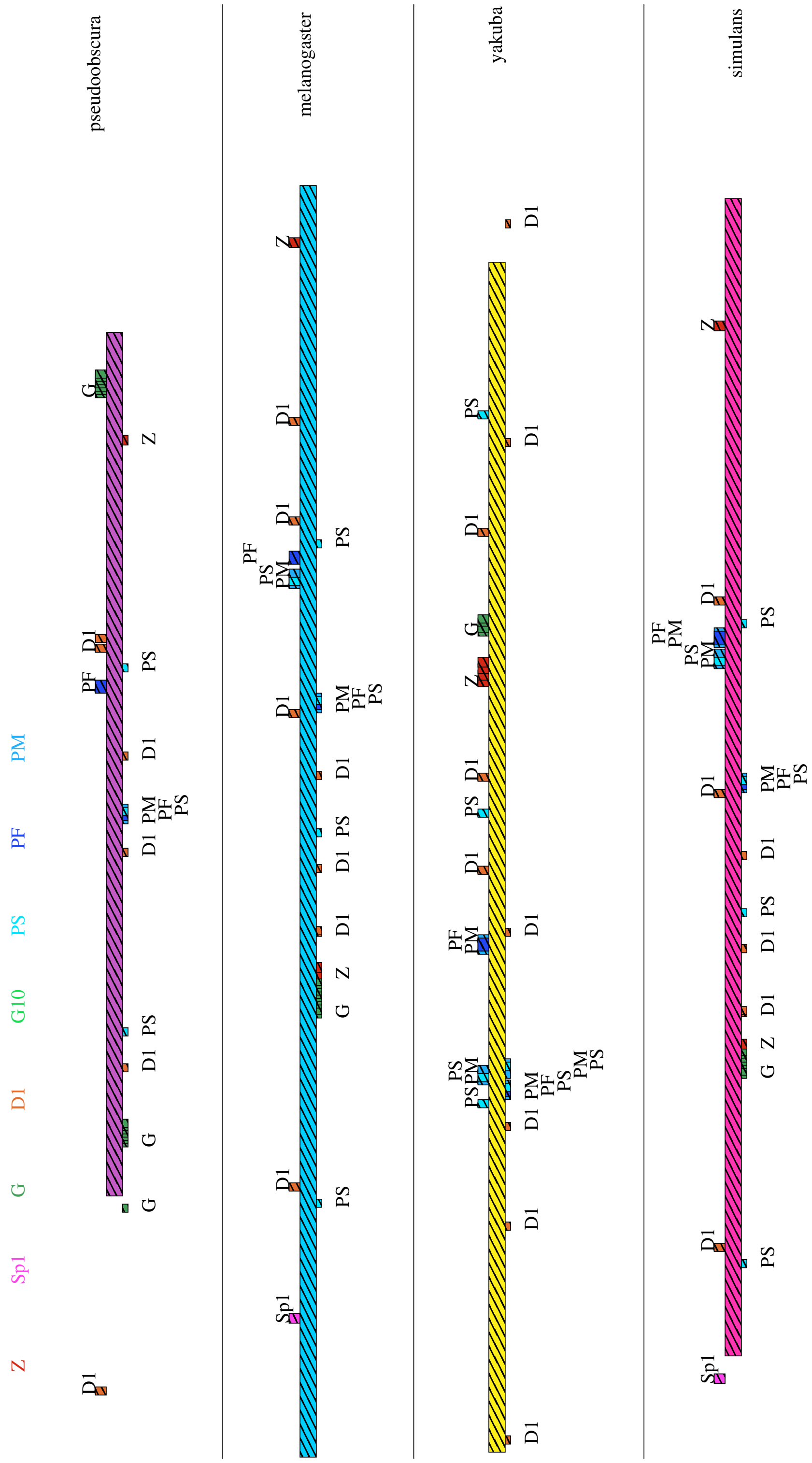

Figure 5.16: Motif composition in the PREs of the *eyes absent* gene in the species *D.pse*, *D.mel*, *D.yak* and *D.sim*. The motifs are Zeste (red), Sp1/KLF (pink), GAF (green), Dsp1 (orange), GAF 10-repeats (light green), and the three different Pho definitions PF, PM, PS (light blue, blue, dark blue). Boxes above the sequence bars are found on the + strand, boxes below the bars on the - strand. Length of each bar shows orthologous region in relation to *D.mel*.

even if we can predict a PRE in one species, but don't see a predicted PRE in the homologous region in another species, does this conserved region contain "presites" of the binding motifs, favouring a PRE gain in this region over adjacent regions?

Phylogenetic trees and PRE gains

The recent release of the sequenced genomes of 12 Drosophila species [95] provides us with the data needed to trace single PREs through their evolutionary history. A complete dynamic search between all 12 species is not performed because most genomes are not completely assembled yet. On the other hand, for questions that cannot be answered based on the completed genomes alone, additional data can provide further information. During the dynamic search on the four species, an additional PRE has been predicted genome-wide in *D.pse* inside the Bithorax complex. The PRE is located upstream of the gene *abd-A* within the infraabdominal region 3, named as iab3-abd-A. As shown in Figure 5.18, an enrichment of Polycomb and Polyhomeotic protein binding can only be detected by ChIP in *D.pse*, but not in *D.mel*, *D.sim*, or *D.yak*. Transgenic reporter assays of 1.6 kb of the predicted *D.pse* PRE show typical PRE behavior like variegation, paring sensitive silencing and response to Polycomb group proteins. None of these are present at the orthologous region in *D.mel*.

The additional *D.pse* PRE in the usually strongly conserved Bithorax locus could either be a relic of an ancestral Bithorax complex that has lost PRE functionality at this position during the evolution, or the development of an additional PRE during the *D.pse* development. To answer this question, a 5 kb region at the same region in all 12 genomes has been scored (Figure 5.17). It can be seen that none of the melanogaster subgroup species reaches a score above 50 and no ChIP enrichment can be detected at those positions. The dynamic search predicts the iab3 PRE in only four out of 12 species, in *D.pse*, *D.persimilis*, *D.willistoni*, and *D.ananassae*. In the most parsimonious tree two mutations are sufficient to simulate the evolution. In addition to a loss during the separation into the melanogaster subgroup, either another loss during the development of the *D.mojavensis*, *D.grimshawi*, and *D.virilis* branch, or an early gain are required. In both cases, according to this tree, the iab3 PRE we found in *D.pse* is not a novel gain in that species, but a loss in the melanogaster species.

Motif presites and pre-PREs

Presites definition

We already saw PREs that are located at different positions in different species. How is it possible that clusters of functional motifs are moved to different locations, inde-

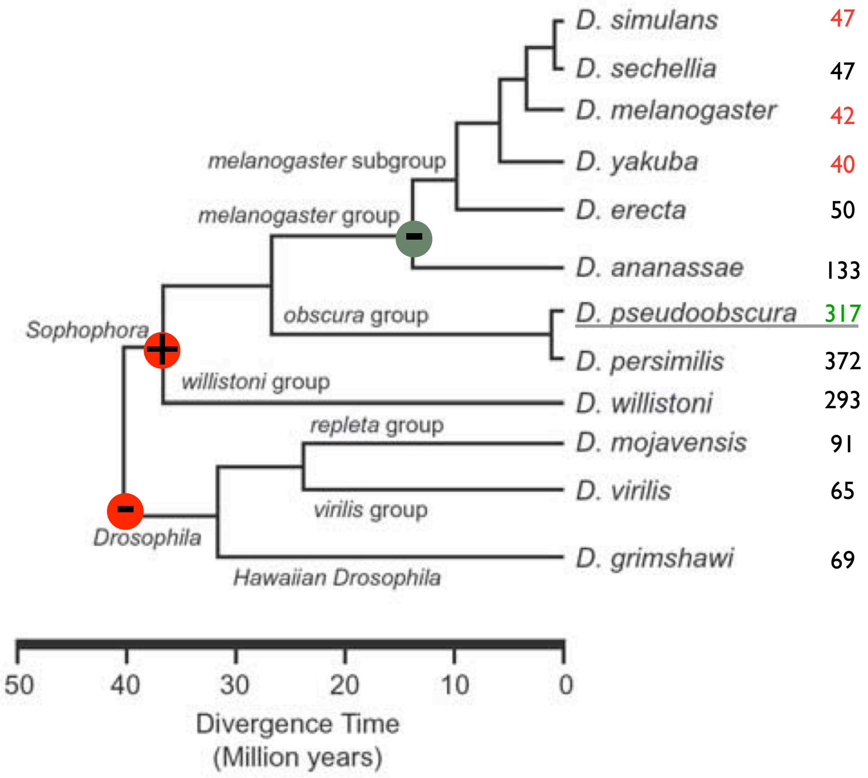

Figure 5.17: Phylogenetic tree of the iab3-abd-A PRE. Numbers give highest prediction score within an orthologous 5 kb window. Green number shows validated presence of PRE functionality by ChIP. Red numbers show negative ChIP results. Underlined species is query sequence for the BLAST search in the other species. Green '-' indicate a PRE loss in the most parsimonious tree. Red circles show either gain or loss at two different positions.

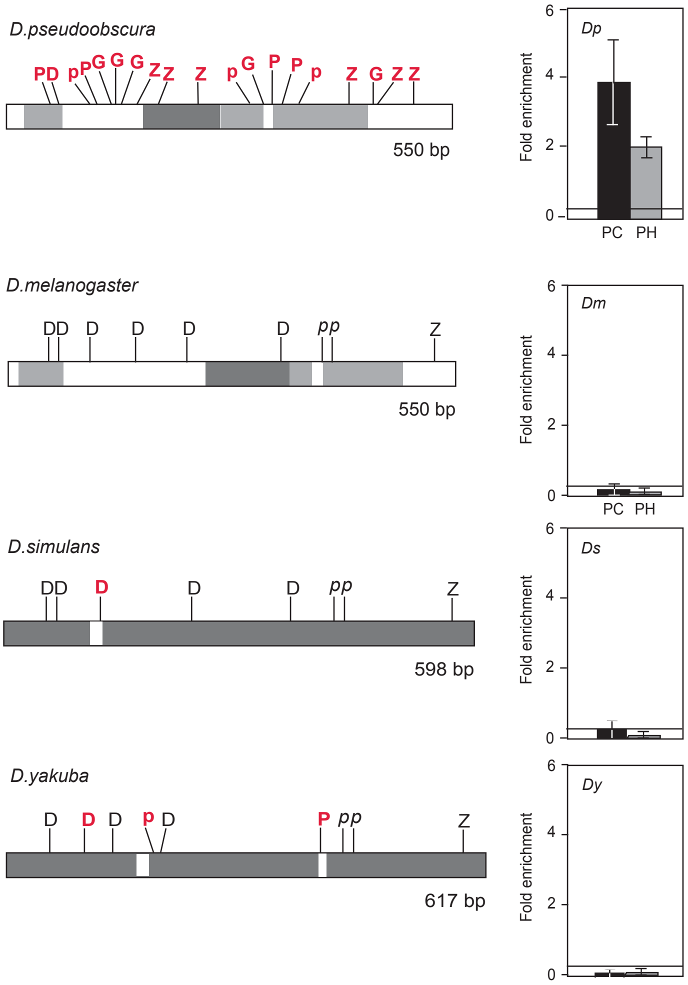

Figure 5.18: Additional prediction hit iab3-abd-A in *D.pse* shows no ChIP enrichment of PC nor PH in *D.mel*, *D.sim*, and *D.yak*. D = Dsp1; Z = Zeste; G = GAF; P = Pho extended site (PF or PS); p = Pho core site (GCCAT). Motifs shown in red are absent in *D.mel*. Figure taken from [97]

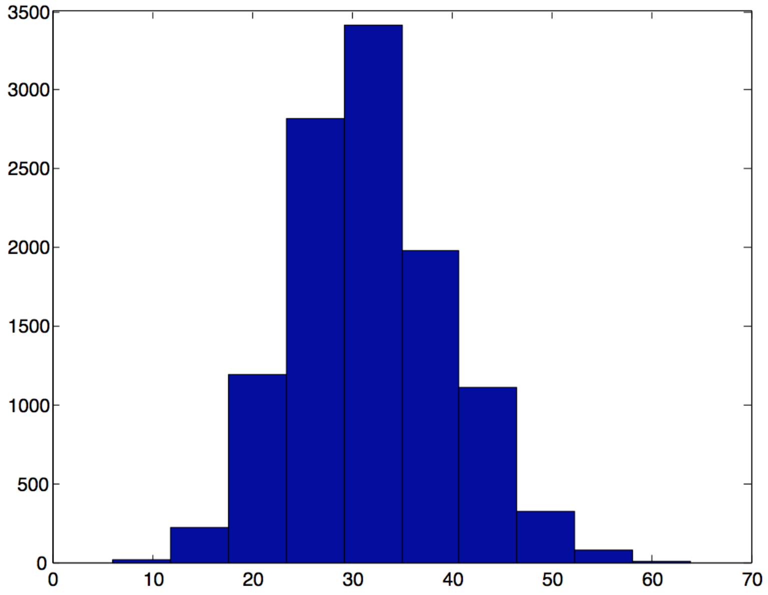

Figure 5.19: Distribution of number of sites with a Hamming distance of 1 to PM or PF binding sites within a 2.5 kb window in chromosome 3R of *Drosophila melanogaster.*

pendent of sequence conservation?

We apply the presite model shown in Section 3.5 to regions in which we can predict and furthermore validate strong evolutionary shift of PRE locations. The window size is set to 2.5 kb in order to make sure that the regions completely cover potential PREs, while on the other hand still keep a resolution at least on the level of ChIP experiments. As background the chromosome 3R of *Drosophila melanogaster* is chosen, because the observed examples as well as the motif rich Bithorax complex are located on that chromosome.

The motifs used are Zeste, GAGA, Engrailed, G10 and Pho, which is a combination of the PF and PM motifs. For example, the distribution of the number of sites with a Hamming distance of one to either PF or PM within a window of 2.5 kb is shown in Figure 5.19 for chromosome 3R and serves as a background for the analysis of specific subsequences.

For Zeste, Gaga and Pho only one mutation is allowed, while in case of the longer Engrailed motif three mutations are accepted. The average number of presites within a 2.5 kb window in the background are shown in Table 5.2.

Motif	#mutations	Ø sites
Zeste	1	56.4
GAF	1	56.5
Pho	1	32.1
En1	2	3.6
En1	3	26.8
G10	2	2.5
G10	3	15.1

Table 5.2: Average numbers of presites for different motifs with fixed number of mutations allowed in a 2.5 kb window in chromsome 3R of *D.mel*.

Example Decapentaplegic

The PREs of the gene *decapentaplegic* are the first region analyzed by the presite model because they serve as an example for the plasticity of PRE locations. In the four species analyzed, the PRE has been detected in three adjacent but distinct places. In *D.mel* (Figure 5.13), the PRE is located around 15 kb upstream of the gene while in *D.pse* a PRE is located directly next to the transcription start site. A third PRE site is in between the first two, showing strong ChIP enrichment in *D.yak* and *D.pse* and also weaker enrichment in *D.mel* and *D.sim*. While the first and second sites are ChIP enriched in different species, the third one near the transcription start site only shows enrichment in *D.pse*. The prediction scores of the three sites show only a genome-wide significant peak in *D.pseudoobscura* at the third position. A dynamic search based on the *D.pse* PRE leads to the prediction of a PRE at positions 1 in *D.sim*, 3 in *D.yak* and at both positions in *D.mel*. All three species show no score peak at the third position.

Presites of Pho, Gaga, Zeste and engrailed are searched in the complete region in all species combined in a p-value for each window as described in Section 3.5. The chosen maximal Hamming distance is 1 for Zeste, GAF, and Pho as well as 2 for En1 and G10. In *D.mel* (Figure 5.20) the highest Pho presite density is shown at position 1, overlapping with the highest score peak within a 50 kb radius. Because only Hamming distances of one are counted as Pho presites, the score peak does not reflect the presite presence but in fact the accumulation of matching Pho motifs which cannot be seen in the presite plot. Other motif presites are not that significantly enriched in position 1, dropping the combined p-value to 0.001. The lowest p-value (< 0.0001) can be found at position 3, showing higher pre-PRE potential there than at any other site in the region. Nevertheless, Pho presites are the least enriched with a motif specific p-value of 0.28. Furthermore, the low and unsignificant prediction score at that position might

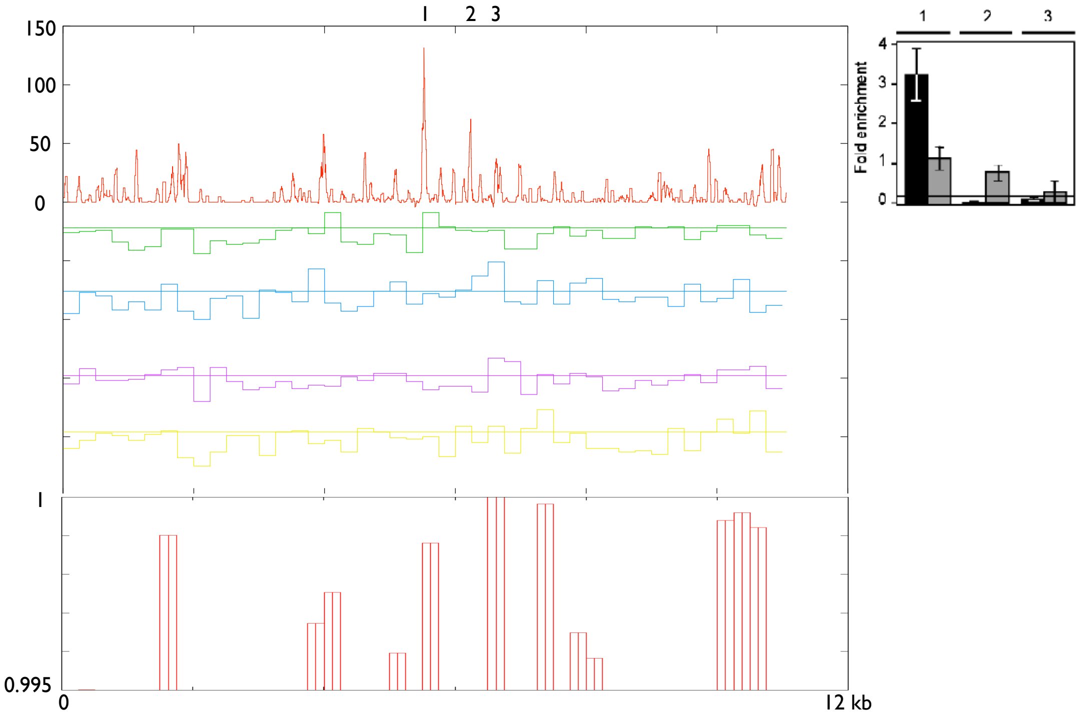

Figure 5.20: *D.mel*: Prediction score and presites around the *dpp* gene. Green: Pho (PF+PS) presites with Hamming distance one. 75% of background are below green line. Blue: Zeste presites with distance one. Purple: En1 presites with distance three. Yellow: GAGA presites with distance one. Lables 1,2,3 indicates ChIP tested positions. ChIP results taken from [97] given on the right. Red bars below shows the overall p-value, defined as $1 - \Pi_{presite}(pvalue_{presite})$.

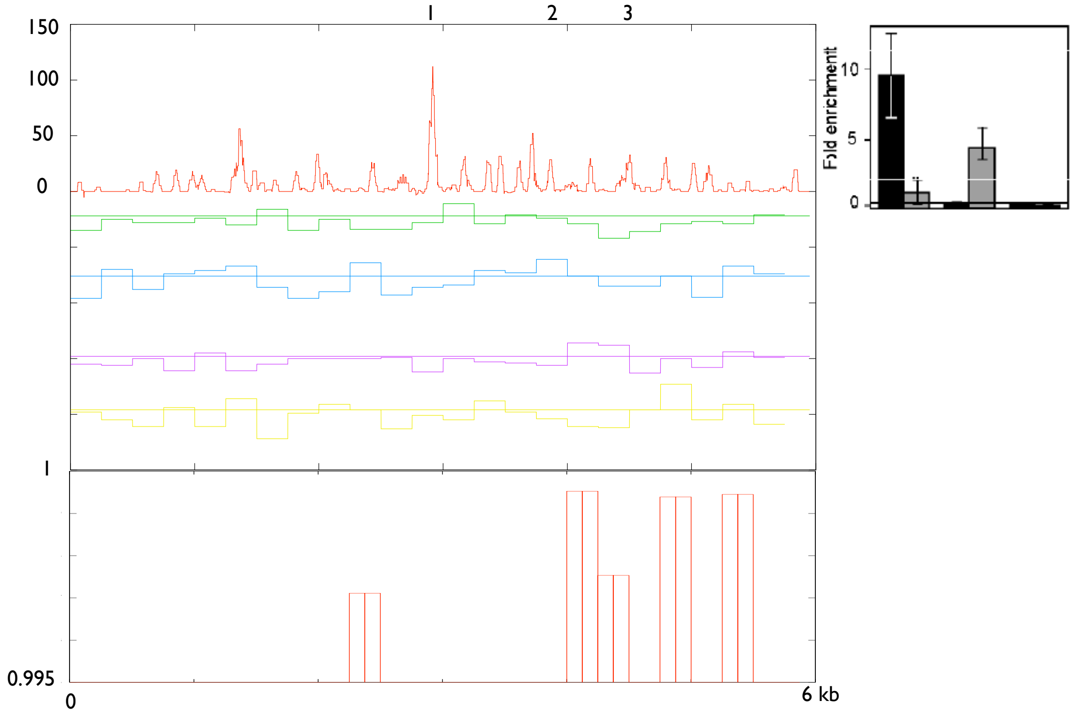

Figure 5.21: *D.simulans*: Prediction score and presites around the *dpp* gene as described for *D.mel* in Figure 5.20. ChIP results taken from [97].

be explained by the small number of found Pho motifs which is consistent with the ChIP data that shows the lowest enrichment of all species in *D.mel*. The data suggests that there might be a pre-PRE at position three mainly defined by GAGA, Zeste and Engrailes motifs, but the site is not active due to only few Pho motifs and presites.

The situation in the other three species differs from the one observed in *D.mel* in some cases, allowing different interpretations. In *D.pse* (Figure 5.23) positions 2 and 3 are shown to be ChIP enriched, in contrast to position 1. Looking at the presites found at the positions we can see an enrichment of GAGA and Engrailed at position 1, a score for Zeste which is still above 75% of the background, and only low Pho potential, similar to the observations of position 3 in *D.mel*. Interestingly, Pho presites are not enriched in any position within the analyzed region, the prediction of position 3 is due to exact Pho matches. At the ChIP enriched position 2, except for high numbers of GAGA presites, no other presites can be found significantly above background. The positive ChIP result cannot be explained by the model since neither Pho motifs nor Pho presites appear to be enriched. One possible explanation is that the motifs used are not

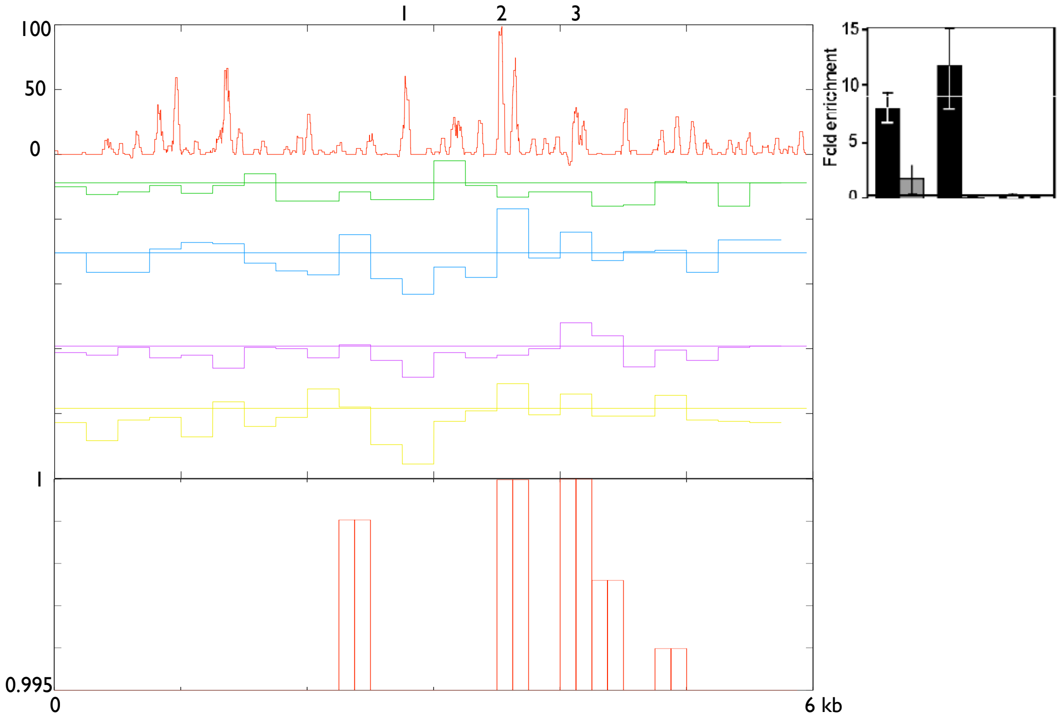

Figure 5.22: *D.yak*: Prediction score and presites around the *dpp* gene as described for *D.mel* in Figure 5.20. ChIP results taken from [97].

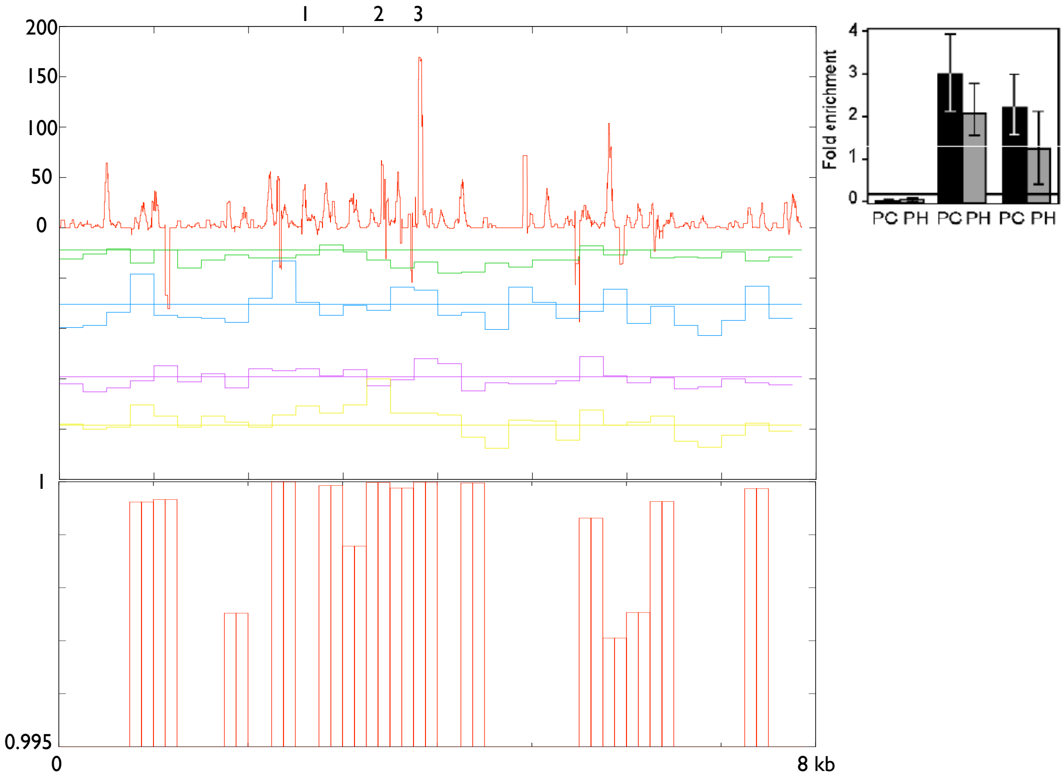

Figure 5.23: *D.pse*: Prediction score and presites around the *dpp* gene as described for *D.mel* in Figure 5.20. ChIP results taken from [97].

able to detect all various kinds of Pho motifs. The prediction's emphasis on specificity inevitably leads to a lack of sensitivity as already discussed in previous chapters. In *D.sim* (Figure 5.26) and *D.yak* (Figure 5.27), ChIP results indicate PRE activity at the positions 1 and 2 but not in 3. Interestingly, in *D.yak* the lowest combined p-value within the analyzed area can be observed at position 3, although Pho presites are again not overrepresented in contrast to the other motifs. In *D.sim* only En1 presites are found above background at position 3 while especially Pho occurrences are rare. This observation does not allow any concrete statement because a function of En1 sites in Polycomb recruitment has not been verified.

Example trachealess

As the second example serves the PRE of the gene *trachealess* which has been analyzed by ChIP in all for species. As can be seen by the ChIP data, two different positions might act as a PRE regulating the gene, depending on the species. In *D.mel* (Figure 5.24), the prediction reveals a high and genome-wide siginificant score peak near the transcription start site at which position ChIP shows strong protein bindings (position 2). The same position in *D.pse* (Figure 5.25) shows no significant prediction score, the nearest potential PRE predicted by dynamic search is located 5 kb downstream inside an intron (position 1). The ChIP experiments indicate only weak protein binding at position 2 but high fold enrichment at position 1 in *D.pse*. The other two species *D.sim* (Figure 5.26) and *D.yak* (Figure 5.27) have high score peaks at position 2, covered by ChIP and low scores as well as low to none enrichment at position 1. Using the presite method we want to explore whether the non-enriched regions 1 or 2 in the different species show pre-PRE potential. Interestingly, the highest pre-PRE scoring site in *D.mel* is actually at position 2, where high numbers of Zeste, GAGA and Engrailed presites can be observed, although the region already scores above the genome-wide cut-off and therefore contains direct motif matches. One possible explanation is that the accumulation of presites favored the development of the PRE at this position and the remaining presites are not necessary to be mutated into motifs to keep the PRE functional. This theory is supported by the lower number of Pho presites in combination with the found overrepresentation of Pho motifs which might be sufficient for a PRE at this position. In this case the found GAF, Zesete and En1 presites could be considered as an artifact of PRE development. At position 1 only Pho presites are found strongly overrepresented with a p-value of 0.04. This could be a hint to the importance of Pho for PRE functionality. At two more positions, p-values $\leq$0.999 can be seen. Because these regions have not been experimentally tested in any species, it cannot be said for sure whether there are really pre-PRE. Nevertheless, the lack of Pho presites at both positions could be the reason that PRE functionality is present at position 1 in other species. In *D.pse*, the most significant p-value can be observed at position 2. The negative prediction score is caused by repeats of the GAGA motif because this motif paired with itself is underrepresented in the positive training set. At the same position, strong enrichments of Zeste and Engrailed presites can be found, which allows the assumption that minor mutations in this region might add additional motifs to the present GAGA motifs, enabling PRE functionality.

The situation found in *D.yak* and *D.sim* can neither be used to confirm nor to disprove the model. In both cases there actually is a significant amount of presites at position 1 but in comparison to adjacent windows it cannot be seen that there is higher pre-PRE potential at that position. The plots of both species show similar values at some positions, again indicating a third site approximately 1 kb in front of position 1 that shows high presite accumulations, although Pho is the only presite not overrepresented. This observation matches with the one made in *D.pse*. Further experimental studies of this position in various species could detect even more PRE positions in the *trh* region. A fourth PRE that is present in only some species would demonstrate even stronger evolutionary changes in that region.

Result

The hypothesis of the pre-PRE model follows the one used in former presites based evolutionary studies. MacArthur et al. [90] calculated the “output” of a sequence based on the score a motif PSSM provides in combination with a clustering factor. This factor favors occurrences of motifs in close proximity to each other, preferably located on the same side of the DNA strand. The output is calculated for each possible mutation at each position and steps increasing the output are defined as a selective advantage and hence introduced into the population. While their model is used to give an estimate about the required numbers of generation to develop enhancer functionality at different position using three different motif PSSMs, we want to find sites of higher probability to gain PRE functionality during evolution to give possible explanations for the observed plasticity. One has to keep in mind that it is unknown what is the minimum requirement to make a PRE functional. Although it can be said that some of the involved motifs are known, it remains unclear which motif combination is sufficient. We could use the genome-wide prediction cut-off as the stop criterion, but this way we restrict the potential of the presite analysis to the types of PRE that are preferably detected by the prediction.

Nevertheless, a second model was used mutating a set of randomly chosen bases in each step, keeping only those sets that increased the prediction score. Once the

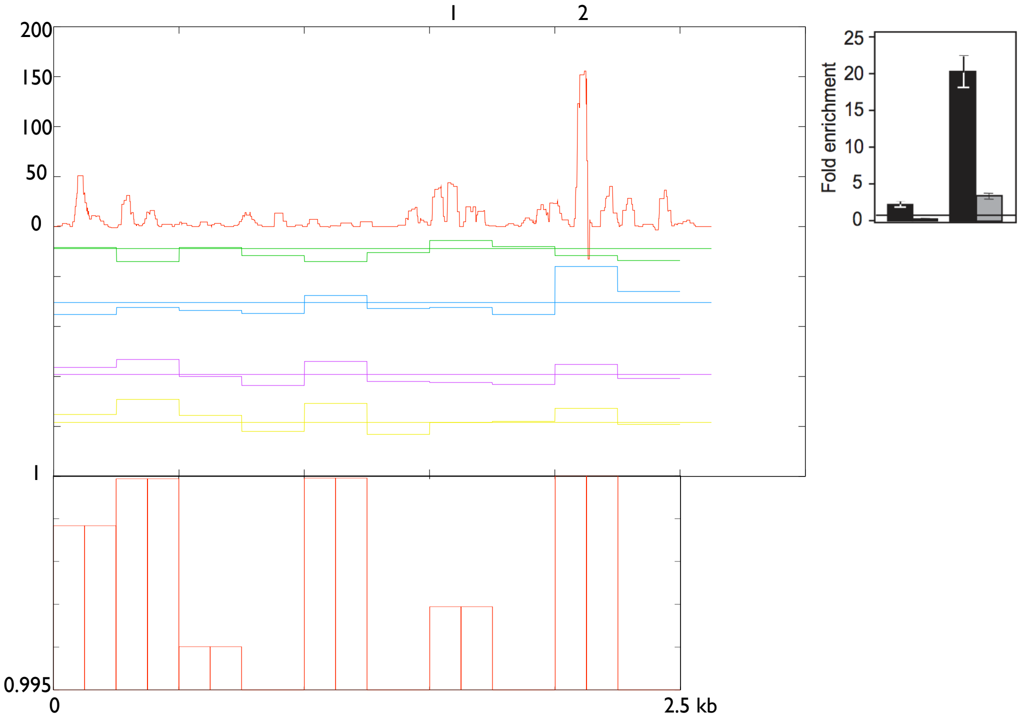

Figure 5.24: *D.mel*: Prediction score and presites around the *trh* gene, description as in Figure 5.20. ChIP results taken from [97].

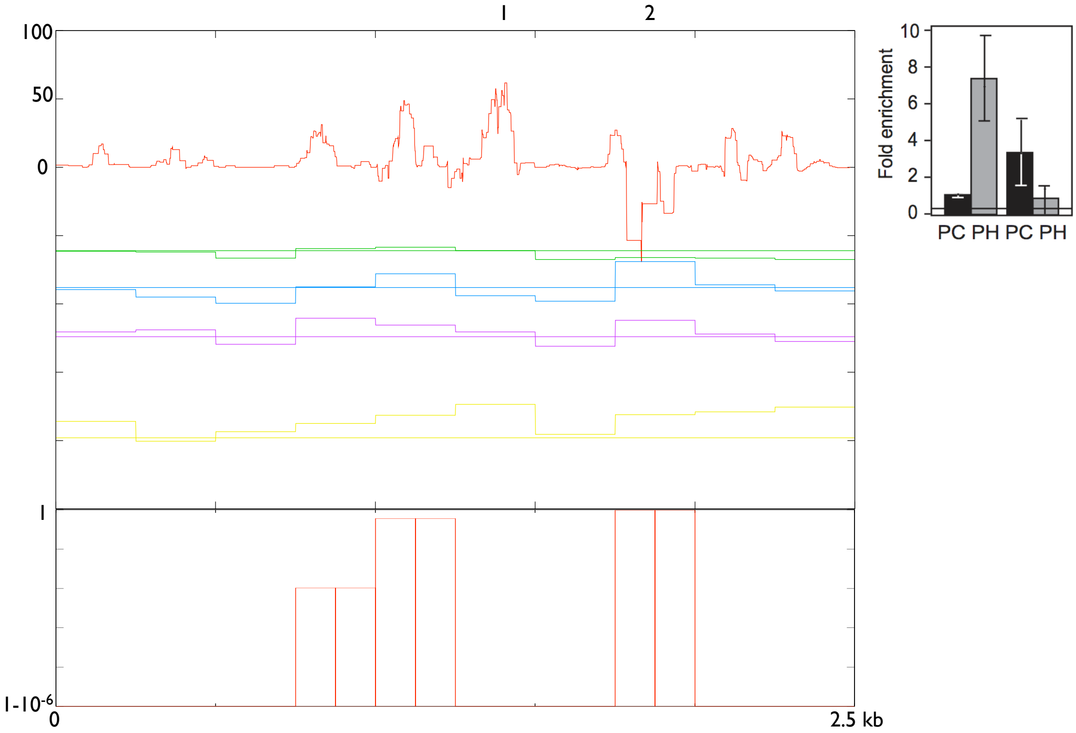

Figure 5.25: *D.pse*: Prediction score and presites around the *trh* gene, description as in Figure 5.20. ChIP results taken from [97].

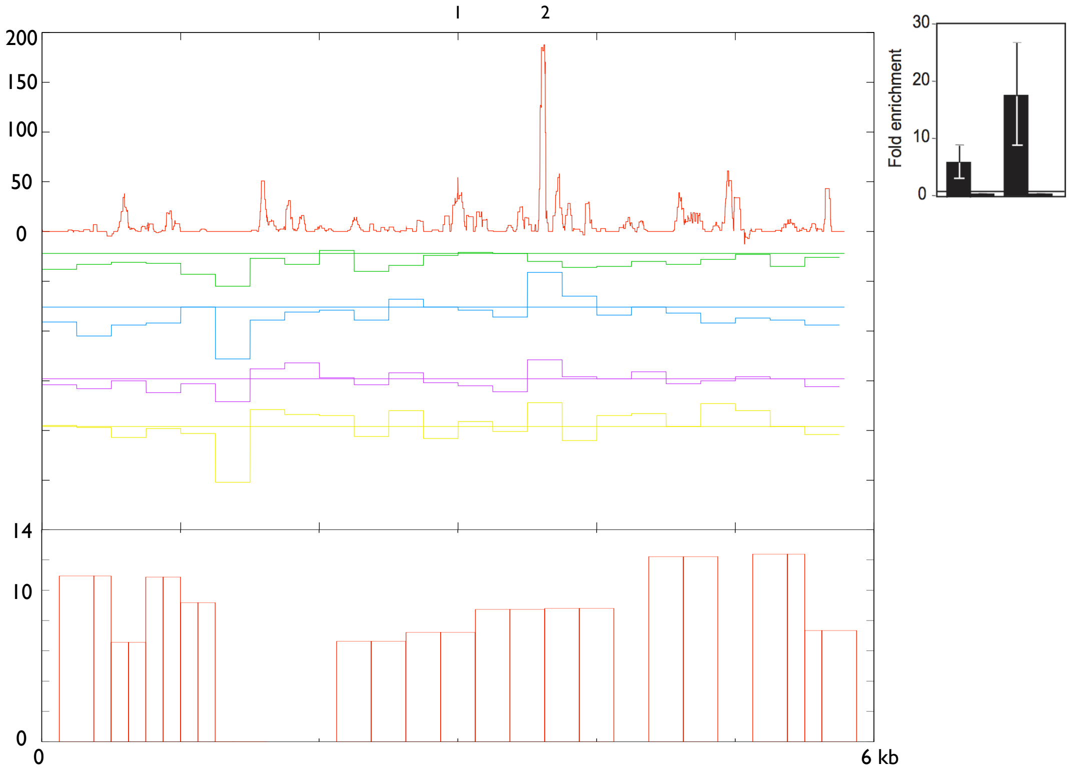

Figure 5.26: *D.sim*: Prediction score and presites around the *trh* gene, description as in Figure 5.20. ChIP results taken from [97]. Scale of Y-axis on second plot in log scale $1 - \log(\Pi_{presite}(pvalue_{presite}))$.

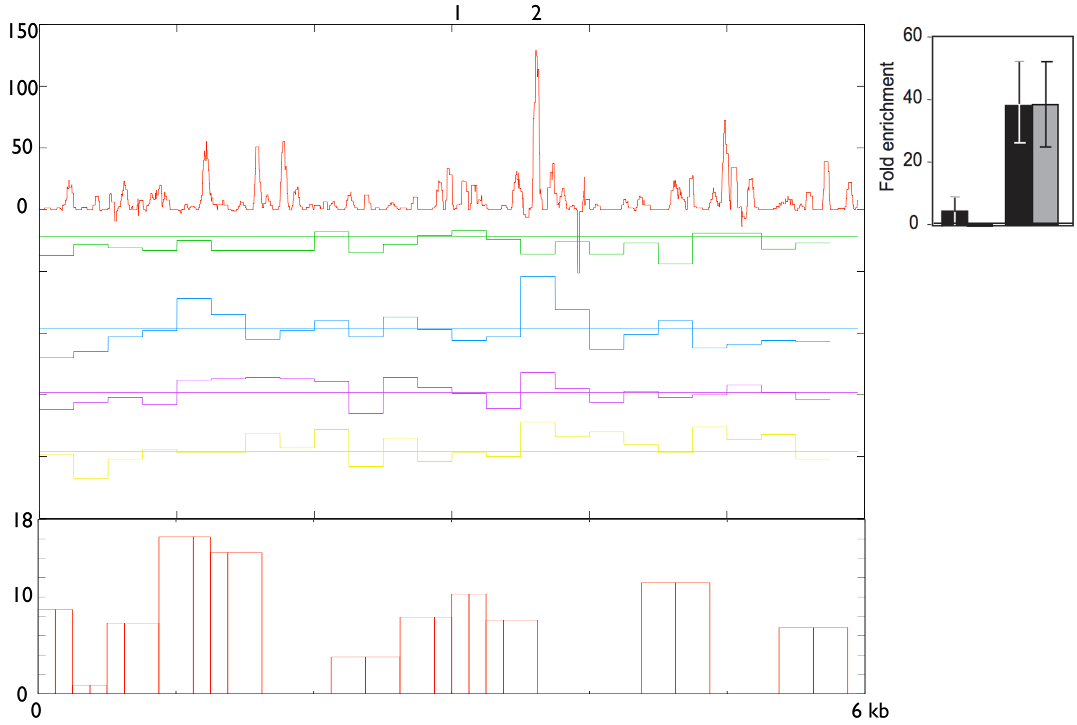

Figure 5.27: *D.yak*: Prediction score and presites around the *trh* gene, description as in Figure 5.20. ChIP results taken from [97]. Scale of Y-axis on second plot in log scale $1 - \log(\Pi_{presite}(pvalue_{presite}))$.

prediction score cut-off was achieved, the Hamming distance was calculated in those parts of the sequence that are part of a motif in either the original sequence or the newly mutated one. Sequences of length 5-7 kb were used at different positions around the *dpp*and *trh* genes, in each step mutating 20-50 bases. The results showed a large variation in the Hamming distance, which can be understood as the required steps to turn the sequence into a high scoring one, going from 9 to 100 required steps. A comparison of the average or mean number of mutations showed a direct correlation between the difference in initial prediction score of two sequences and the steps required to reach the cut-off. The higher the prediction score, the less steps are neccessary to reach the cut-off score. This model could not give additional information to the prediction score in terms of pre-PRE sites and therefore the motif based distance model was chosen.

In general, the data suggest that the presite model can provide further insights into PRE plasticity. The pre-PRE theory might be capable of explaining the evolutionary mechanism involved in PRE gaining. As been shown in the *dpp* example, orthologous sites of active PREs in one species that neither show ChIP binding nor high prediction scores can be matched with a significant p-value to predicted pre-PRE sites. Similar observations in the *trh* example show high pre-PRE potentail at non-functional sites in one species that are validated PREs in homologous position in other species.

5.3 Comparing *DynScan* to genome-wide ChIP

In the original PREdictor run, with emphasis on high specificity, 167 PREs were predicted in *D.mel*, containing only one false-positive to be expected by chance (E-value of 1). The calculation of the E-value is based on counting false-positive hits in background sequences generated from a null-model, which in this case is a zero-order Markov Chain.

Inevitably, a focus on high specificity leads to a lack of sensitivity. As mentioned by Ringrose et al. [11] more than 50% of the known PREs in their positive training set scored below the stringent score cut-off and since were not part of the 167 predicted PREs. In the mean time multiple publications [70, 71, 72] gave positions of regions expected to contain PREs generated by different methods (see Section 2.2.4).

All these publications compared their results to the 167 previously predicted PREs and found at most a 20% overlap (Figure 5.28). Because new motifs involved in PRE functionality have been published, the authors suggested to use these additional motifs in the PRE prediction to improve sensitivity.

Furthermore, the absence of experimentally supported regions in the area of most

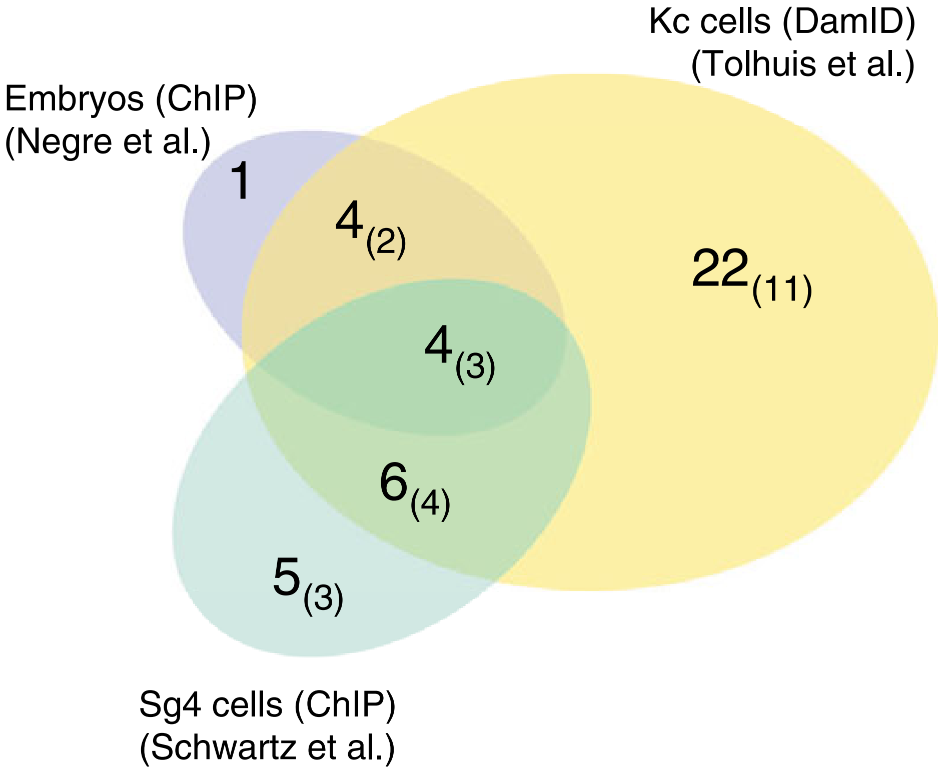

Figure 5.28: Overlap between Polycomb targets in three different studies for common tested regions. Numbers in brackets show predicted overlaps at cut-off of 70 (Figure from [101]).

of the predicted PREs let to a questioning of the specificity of [11], indicating these predicted PREs to maybe be false-positive. Nevertheless, the transgenic experiments performed on the prediction results by Ringrose et at. [11] as well as the ChIP experiments justified the use of PRE predictions as a *DynScan* application, leading to even stronger indications of the method's specificity (see Section 5). Additional motifs had not been introduced in the *DynScan* search to allow the evaluation of the method on an already published basis.

In this chapter the effect of *DynScan* on the sensitivity of PRE prediction will be analyzed in relation to large scale ChIP experiments, as well as the effect of additional motifs.

Original set

Because of the strong emphasis on specificity, several real PREs score below the chosen threshold of 157 [11]. It is no surprise that in a lot of the regions given in the three large scale experimental studies [70, 71, 72], none of the original 167 PREs are located. For example, the PcG protein enriched regions in [70] have an overall length of 57kb, covering only 0.05% of the whole genome and the 131 enriched regions from [72] cover only 3.2%. Therefore overlaps with predicted 167 PREs are very unlikely, which means that every single match raises the confidence of the biological methods as well as our prediction. Furthermore, most of the predicted PREs are not covered by the given regions, which does not necessarily weakens the specificity of the prediction but questions the sensitivity of the biological experiments for two reasons. First, out of 43 tested PREs that were predicted in [11], 41 are true positives, while the remaining two are not proven to be false positives. Secondly, the overlap between the regions in all three papers [70, 71, 72] is as little as it is with the predicted PREs, indicating that a negative result in a genomic region in each of the papers is not sufficient to show absence of PRE functionality (Figure 5.28). Furthermore, PRE activity is cell type specific, depending on the cell lines used in the experiments different results are to be expected. In Table 5.3 the comparison of the biological regions with the PREs predicted based on the original motif set is shown. The Schwartz domains are provided in form of cytological positions and corresponding genes. For a first analysis of the overlap, the large cytological regions are used. The significance of the overlaps can be assessed by hypergeometric distribution, considering the genome as a set of distinct subsequences of the lengths given in the table. Counting the overlaps between experimental data and predictions is then modeled as a drawing experiment.

As already stated only very few predicted PREs are located in those regions.

Source	#Sequences	Length	# Predictions/# Overlaps	p-value
Schwartz et al. [71]	96	110 kb	41/29	$6.4e^{-7}$
Tolhuis et al. [72]	131	28 kb	7/7	0.16
Negre et al. [70]	141	5 kb	0	-

Table 5.3: Overlap between predictions with classic motif set (201 elements) and three biological studies. Numbers of total overlapping PREs and number of distinct biological regions are shown. P-value calculation based on hypergeometric distribution.

Source	# Sequences	Length	# Predictions/# Overlaps	p-value
Schwartz et al. [71]	96	110 kb	112/55	$2.5e^{-13}$
Tolhuis et al. [72]	131	28 kb	22/16	0.13
Negre et al. [70]	141	5 kb	2/2	0.55

Table 5.4: Overlap between predictions with extended motif set (603 elements) and three biological studies. Numbers of total overlapping PREs and number of distinct biological regions are shown. P-value based on hypergeometric distribution.

Motif-sets

The original prediction was based on a limited motif set consisting of Engrailed, Gaga, G10, Zeste and three different Pho motifs. This motif set will be referred to as “original motif-set”. In the mean time new studies stated that other DNA binding motifs may play a significant role in PRE functionality, like DSP1 (GAAAA) and Sp1/KLF (RRGGYG).

Furthermore. the different Pho definitions have been combined to a single matrix in the jPREdictor work [10]. Combining the jPREdictor motif set with Sp1/KLF gives us a new “extended" motif set. Additionally the training sets are also extended (sequences from personal communications with Thomas Fiedler). A prediction based on this extended motif set has a cut-off of 114 at an E-value of 1. The number of genome-wide predicted elements rises to 603. The increased number of predicted elements also increases the number of overlaps with the three studies (Table 5.4). An increase of predicted PREs by factor three leads to an increased overlap of almost the same factor. A difference in specificity of both different motif sets cannot be observed in this case.

Source	# Sequences	Score cut-off	# Overlaps with prediction
Schwartz et al. [71]	96	85	72
Tolhuis et al. [72]	131	72	43
Negre et al. [70]	41	43	12

Table 5.5: Number of PC domains containing at least one PRE predicted with the extended parameter set. The cut-off for PRE prediction is set to reflect an E-value of 1 depending on the amount and length of Polycomb regions.

Searching regions only

In three cases multiple genomic regions are given that are shown to be bound by Polycomb and hence are expected to contain PREs. We used these regions as prior knowledge to our search by setting the score cut-off equivalent to an E-Value of 1 for searching these domains only. This is similar to the *DynScan* approach, but instead of using predictions in another species as prior knowledge, experimental data is used.

For example according to Negre et al. [70] we can expect to find PREs in 41 regions with average length of ca. 5 kb. To ensure the overall E-Value to be 1 the individual E-Value is set to $\frac{1}{41}$ for each region of 5 kb length. The cut-off drops from 114 for a genome-wide search with the extended motif-set to 43, allowing a more sensitive prediction by still expecting only one false-positive by chance thus the specificity is kept the same (Table 5.5).

DynScan

The *DynScan* approach as well as the PRE prediction based on the classic motif set has been presented in sections 3.2-5. The same approach is repeated based on the extended motif set in a similar way.

In *D.mel* we predict 603 PREs with a genome-wide treshold of at least 114 using the new motif-set. In *D.pse*, *D.yak* and *D.sim* we predict 2457, 681 and 516 respectively with a score of at least 114. The comparative search in *D.mel* starting with *D.pse*, *D.yak* and *D.sim* combined with the 603 genome-wide predicted Dmel PREs reveals 1683 distinctively predicted PREs in *D.mel*, i.e. no overlaps between PREs are allowed. The results of both *DynScan* runs are compared to the three biological studies. It can be seen that the number of overlaps grows with increased number of hits, but strikingly so does the significance of the overlaps. According to the data, the specificity of PREs predicted by *DynScan* is not lower than of genome-wide predicted ones, regardless of the motif set (Table 5.6).

Source	# Sequences	Motif set	# Predictions/# Overlaps	p-value
Schwartz et al. [71]	96	classic	83/45	$3.2e^{-9}$
		extended	278/78	$2.5e^{-17}$
Tolhuis et al. [72]	131	classic	18/14	0.08
		extended	66/40	0.07
Negre et al. [70]	41	classic	1/1	0.76
		extended	8/6	0.15

Table 5.6: Overlap of *DynScan* results with experimental data, based on search of *D.pse, D.yak*, and *D.sim* predictions in *D.mel*. Classic motifset contains PM, PS, PF, Zeste, GAF, G10, and En1 motifs, extended set contains additional DSP1/KLF and DSP1 motifs and additional sequences in training set. Numbers of predicted elements in all regions as well as distinct regions are given.

Motif set influence

Up to this point, we used two different parameter sets, the "classic" one and the "extended" one. The "classic" version has been used in our *DynScan* application. Additionally, a third motif set has been published as an example application of jPREdictor [10]. The latter one leads to a higher number of predicted PREs but lacks of experimental validation yet. As it has been shown, the number of hits predicted with the "extended" set is three times higher than with the "classic" set. But how accurate are the predictions with the newer sets?

To get an idea of the effect of the motif sets, we compare different sets to our previously shown ChIP data. In *D.mel,* we predicted a PRE by *DynScan* at position 1 of Figure 5.29 using the classic parameters (already discussed for Figure 5.13). The motif set described in the jPREdictor publication combines PM, PF, and PS to a single matrix and adds a Pho:DSP1 motif pair. Even with *DynScan* no PRE can be predicted at position 1 based on that set. Instead, the score at position 2 is significant genome-wide. The ChIP data for position 2 indicate possible PH binding, but to a lesser extend than at position 1. Adding the SP1/KLF motif drops the score at position 2 below the genome-wide cut-off but increases the score at position 1 above the 1 kb radius cut-off. Finally, extending the training sets with additional sequences provided by Thomas Fiedler only effects the scores to a minor extend.

In *D.sim*, a similar situation can be observed (Figure 5.30). The PC enriched position 1 has a signficant score within a 20 kb radius only with the classic set. In the other sets, position 2 receives higher scores while position 1 falls below any significance value. Furthermore, position 3 gets high scores if SP1/KLF is added, although neither PC nor

PH enrichment can be detected.

In *D.yak* position 1 shows PC binding but scores below cut-off with all sets (Figure 5.31), while the highest score is based on the classic set. The enriched position 2 is found by *DynScan* with all four sets. However, none leads to a genome-wide prediction hit. The non enriched region 3 gets the least significant score with the classic set. In the other cases the score is significant in a 2 kb window.

Finally, in *D.pse* positions 2 and 3 show PC and PH enrichment. Position 3 is found in a genome-wide search based on the classic set only (Figure 5.32) while at position 2 a PRE is predicted genome-wide by SP1/KLF containing sets.

In all four species the classic set gives the highest overlap with the ChIP data in the *dpp* example. Furthermore it leads to a prediction of a reasonable number of PREs out of which a large fraction has been validated experimentally.

Cut-off influence

It can be seen that the overlap between the different predictions and the experimental results strongly depends on the chosen experiment. While the overlap with Schwartz et al. [71] can be risen to almost 50% with the classical motif set and even higher with the extended training set, the overlap with the sequences provided by Negre et al. [70] remains poor.

Furthermore, one has to keep in mind that it remains to be seen whether the large scale ChIP on chip experiments find only functional PREs or are influenced by indirect binding as described in Section 2.2.4, or detect binding to elements that use different factors for recruitment. Because Schwartz et al. [71] provided the most comprehensive study, we concentrate on those results to examine the influence of the cut-off on the overlap. The scoring is done by the classic motif set only, due to the fact that the classic prediction results has shown to be more consistent with our experimental results.

Additionally to the cytological positions, the identifiers of presumably regulated genes are provided, although without exact coordinates. Still, taking ±10 kb around each gene restricts the length of hits to 4 Mb. All in all 187 genes are provided, that are categorized as "strong" hits (i.e. binding of all four factors). For each gene the highest prediction score is taken and transformed into an E-value based on the total 4 Mb instead of the whole genome. The number of hits at each E-value is set into relation to the number of hits with at least the same score in a set of randomly picked genes. The random background search is repeated several times to determine the error rate as standard deviation (Figure 5.33).

Rising the E-value from 1 to 10 means that according to the null-model only nine

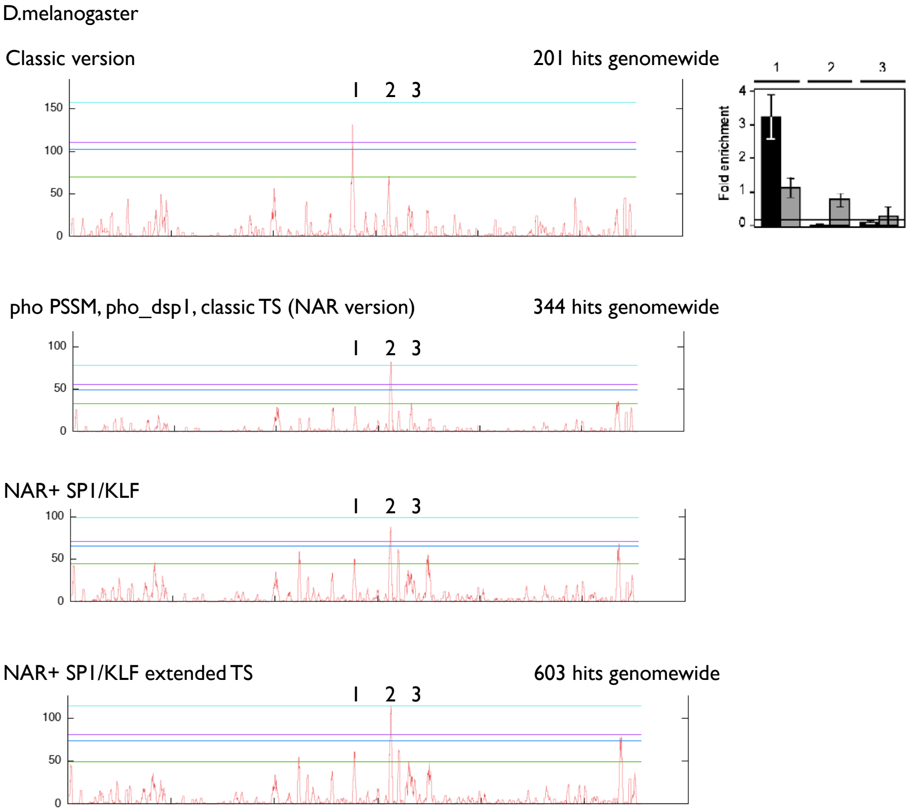

Figure 5.29: Influence of four different parameter sets on prediction of *dpp* PRE in *D.melanogaster.* Classic set (GAF, Z, En1, PF, PM, PS, G10) prediction scores shown in first plot, set from [10] (NAR version) in second plot, NAR + Sp1/KLF motif in third plot, NAR+Sp1/KLF weighted on extended training set in fourth plot. *DynScan* cut-offs for 1 kb, 10 kb, 20 kb, and genome-wide search marked green, blue, purple, and light blue. Positions tested by ChIP provided as 1,2,3. ChIP results given in plot on the right (taken from [97]).

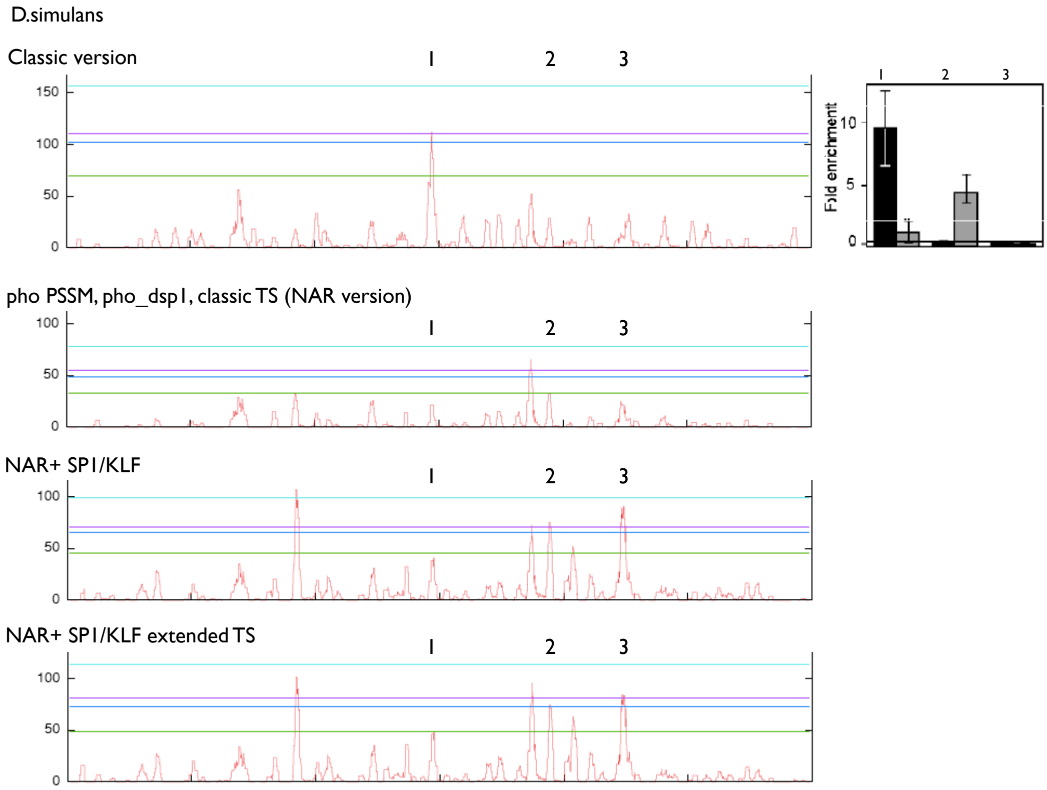

Figure 5.30: Influence of four different parameter sets on prediction of *dpp* PRE in *D.simulans*. Figure explained in Figure 5.29.

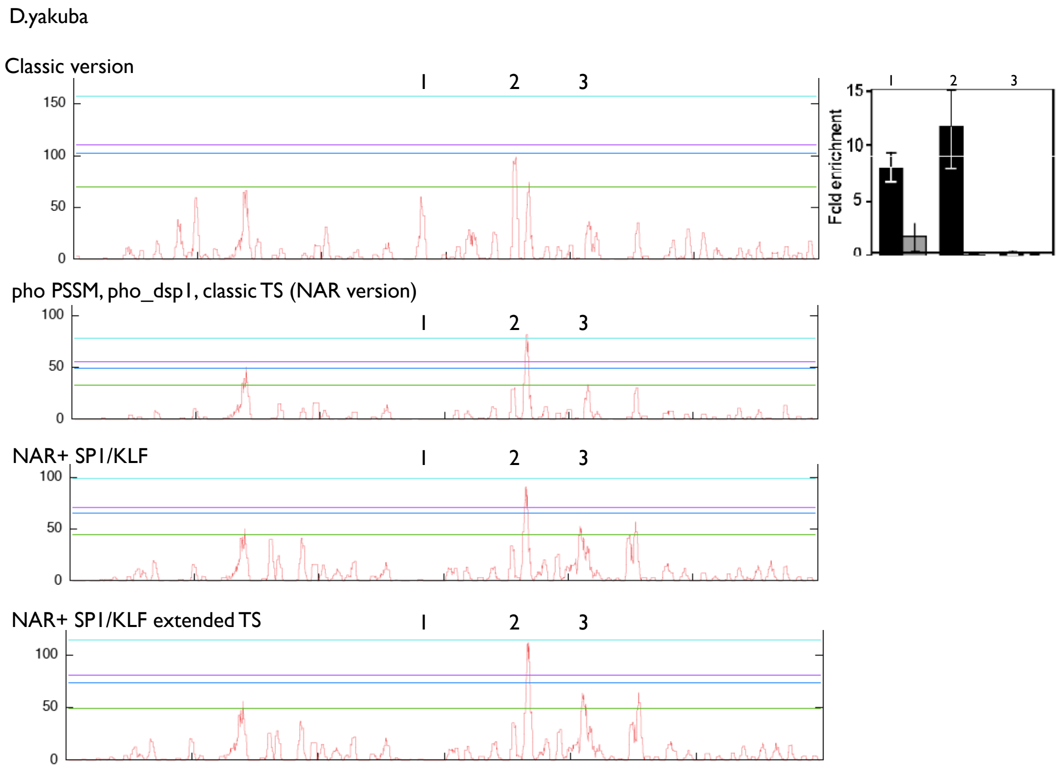

Figure 5.31: Influence of four different parameter sets on prediction of *dpp* PRE in *D.yakuba*. Figure explained in Figure 5.29.

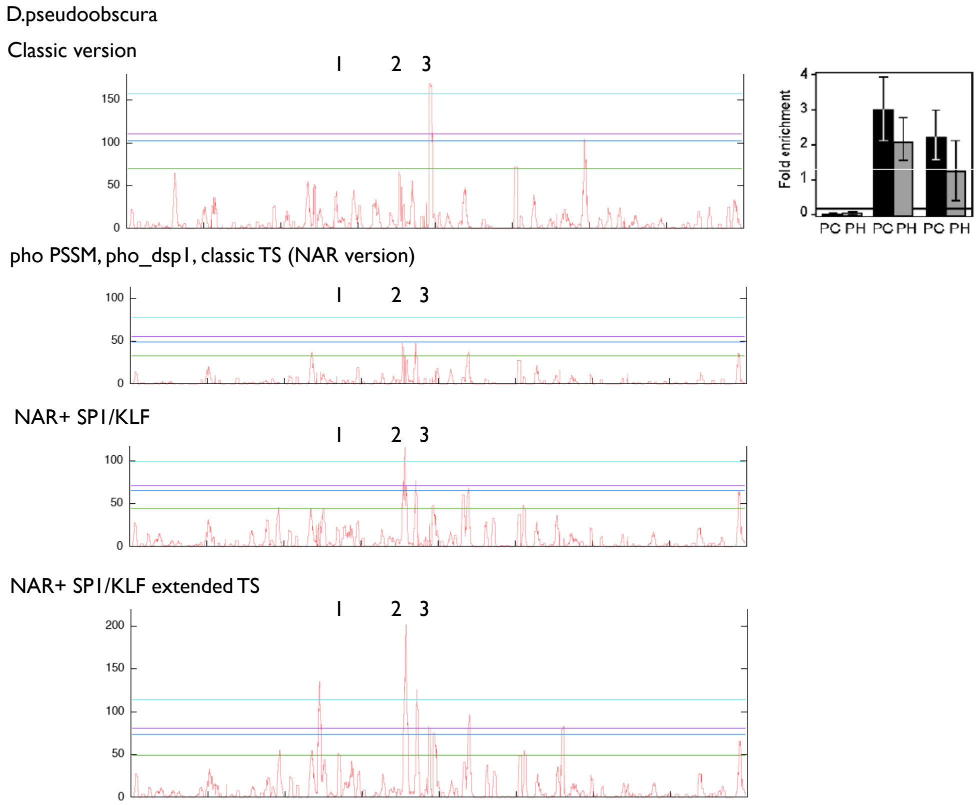

Figure 5.32: Influence of four different parameter sets on prediction of *dpp* PRE in *D.pseudoobscura*. Figure explained in Figure 5.29.

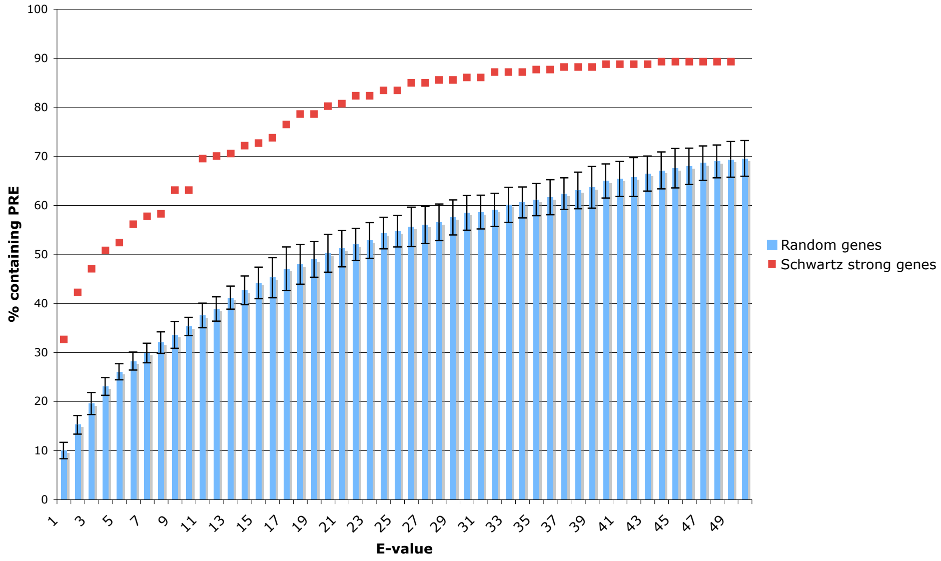

Figure 5.33: Number of hits in 187 “strong” Schwartz genes at different E-values in relation to random background. E-value calculation based on a overall gene sequence length of 4 Mb. Error bars show standard deviation.

additional false-positives are to be expected. Nevertheless, the number of hits is doubled, at an E-value of 10 around 60% of the 187 genes can be found as well as a third of the random genes. While there still is a strong enrichment of overlaps with Schwartz' genes in relation to random genes, a few concerns about the reliability of the null-model may be necessary. The extreme rise of predicted elements by minor increases of the E-value can be observed at any sequence length, in this 4 Mb example as well as in a genome-wide prediction. One possible explanation might be potential "pre-PREs" as described in Section 5.2 that occur in multiple regions and score only a little bit under the cut-off. Depending on the null-model (higher order Markov Chains or shuffled genome data) the corresponding cut-off for an E-value of 1 varies[1]. In each case, the discrepancy between the number of expected additional hits and observed additional hits stays. For this reason, the Drosophila prediction data presented in this work is based on the classical motif set at an stringent E-value. Results based on these parameters have shown reliability in various experiments.

[1] Impact of different null-models on PRE predictions is evaluated by Thomas Fiedler and is not described in detail in this work.

6 Application and Results: Mammalian "PREs"

6.1 Building training sets

The prediction of PREs in Drosophila served as a first application for the dynamic search. The results show the general potential of *DynScan* to enhance sensitivity. Furthermore, the data give very interesting insights into biological aspects of PREs in flies. In mammals, no PREs have been characterized so far, but genome-wide ChIP data are available in mouse [13] and human [14], giving positions of binding sites of PcG proteins such as Suz12. None of these PcG proteins binds directly to the DNA, but possibly to some transcription factors that are recruited by DNA motifs. The lack of validated DNA binding motifs prevents a direct jPREdictor prediction and therefore a dynamic PRE search in mammals. For that reason, in this chapter an analysis of the ChIP data [13, 14] is presented combined with an application of the methods described in Chapter 4, which aimsto identify mammalian PREs, if existent. No binding factors for proteins involved in PcG recruiting have been identified outside Drosophila so far, but homologs of Pleihomeotic (YY1) and Dsp1 (HMGB2) are known. Because it is known that YY1 is involved in PcG protein recruiting [102], the corresponding binding sites should be overrepresented in the Suz12 bound region in contrast to Suz12 unbound background sequences. The Pho motif was provided in form of three different consensus sequences (PS, PM, PF) for the original prediction, while in the jPREdictor publication multiple Pho binding sites were combined into a single matrix. The weight of this matrix based on the original Drosophila model and background sets is 1.7, a Pho-Pho motif pair reaches a weight of 2.7. In Drosophila PREs, the Pho motif is obviously overrepresented. Because YY1 is also involved in PcG protein recruiting, its binding sites should be overrepresented in mammalian PREs. To calculate weights for YY1 single and double motifs, mammalian model and background need to be provided. We therefore build training sets based on the ChIP data recently made available.

Murine Training Set Design

To cover a wide range of possible types of potential mammalian PREs, different training sets are used, built out of mouse as well as human sequences. In mouse, ChIP results for the four proteins Eed, Phc1, Rnf2, and Suz12 as well as H3k27me methylation are available in form of single positions. The main criteria for sequences to be part of the positive training set is that all five factors have to be located inside a 2 kb window. We require the overlap to maximize the specificity of the experimental result. Furthermore, if a set of motifs is present in the sequences, shorter sequences give higher weights due to the sequence length normalization. The criterion that all sequence have to be completely within 2 kb limits the overall number to 71 sequences, covering 17 different chromosomes. We combine subsets of 10-20 sequences to positive training sets. The background is built out of the promoters of chosen genes that are very unlikely to be PcG protein targets. First, the ChIP data for the genes must not indicate binding for any PcG group protein nor H3k27me3 methylation. Furthermore, the function of the gene should raise the confidence in a PcG independent regulation. For example, genes that are involved in reaction to external stimuli such as heat shock genes can be assumed not to be involved determination of cell identity, the same refers to house keeping genes which are meant to be essential for the cell and therefore cannot be switched off by PcG proteins. The negative training set consists of 20 sequences of 2 kb each which are cut out 2 kb upstream of the transcription start site of such genes.

The ChIP experiments were performed on a list of 15742 chosen genes, provided in form of Entrez gene identifiers and assembled from different databases, based on NCBI build 34. To rebuild the dataset, we store all gene annotations from Ensembl, RefSeq and UCSC databases in a local PostgreSQL database. The local schema contains an unique integer identification number, a species id, a source database specific geneid, the location of the gene, a reference to the source database, and corresponding Entrez identifiers. A gene's location is stored as a 4-tupel (begin, end, chromosome, and strand). The begin position refers to the transcription start site, the end gives the end of the 3' UTR. The Ensembl database contains 38000 genes out of which 21155 can be referenced to 19930 distinct Entrez identifiers. From the RefSeq database 20329 genes are stored in our database, cross referenced to 18441 different Entrez identifiers. Finally, gene annotations from the UCSC "known genes" database are added, 23723 genes are cross-referenced to 15037 Entrez genes. In total, we store references to 18424 different Entrez identifiers. Unfortunately, there is an n:m relation between Entrez identifiers and annotated genes. Different genes, even in the same source database, can be assigned the same Entrez identifier. In contrast, also the opposite occurs, the same gene can be cross-referenced to multiple Entrez identifiers. Furthermore, annotations for the same gene in the different databases provide different transcription start site positions. For example, the Entrez identifier 66640, a gene that has been positively tested for all five factors, is cross-referenced to a gene starting at position 52,310,128 according to the UCSC "known genes" database and a gene starting at position 52,309,174 according to Ensembl database. Out of the 15742 tested genes, 15729 can be referenced to at least one entry in any of the three databases.

Instead of running future predictions on the whole mouse genome we take the promoter regions of those genes as the dataset. Boyer et al. [13] performed ChIP tests on sequences reaching from -8 kb until +2 kb around the transcription start site of each of the 15742 genes. Our dataset is set to ±10 kb in order to make sure that the tested regions are included, although the exact position depends on the annotation database used. Overlapping 20 kb regions are combined to a single one, so the 15729 Entrez identifiers are found within 12370 blocks of 20 kb each.

Human Training Set Design

Additionally, training sets are built for the human ChIP experiments. Lee et al. [14] provide genome-wide ChIP data detecting Suz12 sites at 3465 different positions with a length between 21 bp and 35665 bp. In the supplements, the authors give gene identifiers from various databases for genes with a transcription start site within 1 kb around the Suz12 enriched sites. Furthermore, genome-wide polymerase II activity is given in the supplement, together with ChIP results for Suz12, Eed, and H3K27me3 methylation on promoter arrays on selected genes. We combine all data in a local database. First, gene annotations taken from RefSeq, Ensembl and UCSC are stored in the database using the same table in which the mouse genes are stored (*mammal_gene_positions*). The genes are cross-referenced to Entrez identifiers if corresponding information are provided in the source databases. For human, the local database contains around 158000 entries, referenced to 19911 different Entrez genes. Again, annotations in Ensembl and RefSeq for the same Entrez identifier can differ in the gene coordinates. The data given by Lee et al. [14] are stored in two local tables, in (*human_chip_data*) the locations of Suz12 enriched sites are stored together with gene identifiers if a transcription start site is within a 1 kb radius. In addition, we cross-referenced the gene names given by Lee et al. as database specific identifiers to the Entrez identifiers that are stored in the *mammal_gene_positions* table. The promoter array results are combined with the polymerase II activities and are stored in the second table (*human_tested_genes*). Entrez identifiers serve as unique primary key, bindings of Suz12, Eed, H3K27me3, and polymerase II are stored as Booleans. For further analysis, out of the overall 3465 ChIP regions only those are chosen that do not exceed 2 kb in length (2484). The sequences are set to ±2 kb around the ChIP regions' center positions to cover a possible error range of ChIP positions. Lee et al. observed an enrichment of CpG islands around the ChIP positions which we have to keep in mind for the design of the training sets.

The PRE prediction in Drosophila has been found to show highest accuracy for "classic" PREs, namely the PREs that regulate homeobox transcription factors. Assuming that different classes showing different types of motif occurrences exist in mammals as well, one human training set is built out of ChIP regions that are located inside the promoters of homeobox transcription factors. The databases RefSeq and Ensembl provide information about homeobox containing genes. Our local ChIP data containing database is queried for all Suz12 enriched sites that are located around 2 kb of the transcription start site of the homeotic genes and do not exceed 3 kb. These constraints limit the number to 100 ChIP sequences which are split equally on two different human positive training sets. To build a negative training set, the combined results of human ChIP experiments are considered. The supplements of the genome-wide experiments give a list of Suz12 bound genes and polymerase II activity. Furthermore experiments in selected promoters show ChIP results for the PcG proteins Suz12 and Eed, as well as H3K27me3 methylation.

Now all genes that show no binding for Suz12, Eed, or methylation but show polymerase II activity are potential members of a negative training set. Negative results for PcG proteins indicate no PcG silencing, which is even stronger supported by active transcription indicating polymerase II activity. These criteria are served by only 237 Entrez genes, out of which 194 have an unambiguous position entry in the local gene database. Sequences of ± 2 kb around the transcription start site of the genes are cut out. The negative training sets are chosen out of these sequences.

As general negative background sequences for statistical tests all genes are chosen that have an unambiguous RefSeq to Entrez relation and have no positive Suz12 near the promoter regions. The sequences are chosen ±2 kb around the transcription start site. All in all, 14907 sequences are selected for background tests. The different data sets are summerized in Table 6.1.

6.2 CpG islands

Lee et al. [14] observed an enrichment of CpG islands near Suz12 enriched sites: "It is interesting that 40% of all SUZ12 bound regions are within 1 kb of CpG islands, given

Species	Type	No. of seq.	Length	Source
Mouse	Positive	71	2 kb	Overlap of Suz12, Eed, Phc, Rnf2, H3K27me3 within 2 kb
Mouse	Negative	20	2 kb	Collection of promoters tested negatively. Heat shock and house keeping genes preferred
Mouse	Positive	20	2 kb	Subset of set mouse set #1
Human	Positive	2484	4 kb	Set of all Suz12 enriched sites with length $\leq$4 kb. All normalized to 4 kb length.
Human	Negative	14907	4 kb	Sequences taken $\pm$2 kb around TSS of all Entrez genes that can unambiguously be cross-referenced to RefSeq genes without Suz12 enrichment
Human	Positive	50	4 kb	$\pm$2 kb of TSS of homeotic genes that contain Suz12 hits
Human	Negative	50	4 kb	$\pm$2 kb of TSS of genes that are negative for SuZ12 and show strong enrichment of low density CpG islands
Human	Positive	50	4 kb	Second positive TS, same type of sequences
Human	Negative	50	4 kb	Second negative TS

Table 6.1: Overview of the different mammalian training sets. Length refers to each sequence in set.

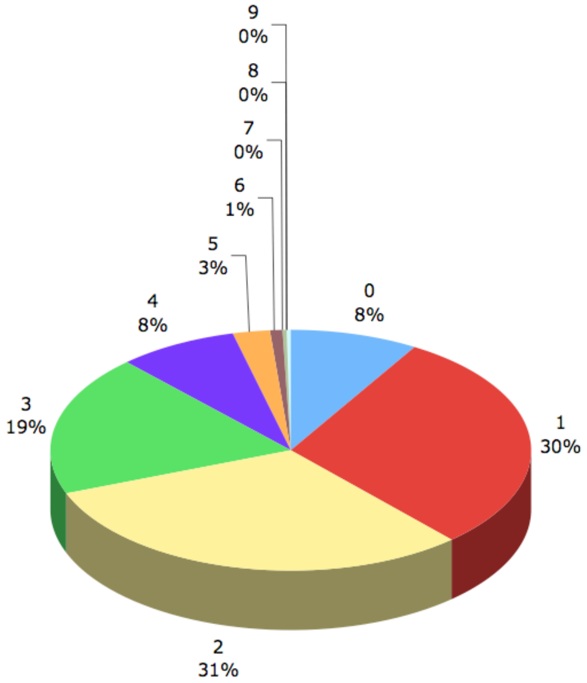

Figure 6.1: Number of low density CpG islands in 2484 Suz12 enriched sites of 4 kb each. Definition of CpG island in this figure: C+G content $\geq$ 50%, Observed/Expected CG ratio $\geq$ 0.6, length $\geq$ 200bp.

the recent discovery of a mechanistic link between PcG proteins and DNA methyltransferoses (Vire et al. [84]).” Their observation requires a more detailed analysis to estimate the effect of CG rich regions on motif occurrences and motif predictions. First, a clear definition of a CpG island is necessary. We used both parameter sets described in the background on CpG islands (Section 2.2.6). First, CpG island are searched that show an observed CG to expected CG ratio $\geq$ 0.6, a C+G content $\geq$ 50% and a length $\geq$ 200 bp. CpG islands are searched by newcpgreport from the EMBOSS [103] package.

Within the 2484 Suz12 sequences of 4 kb each, only around 8% contain no CpG island at all. Almost two thirds even contain two or more islands (Figure 6.1). Do we observe an enrichment of CpG islands in the Suz12 bound sequences? According to Antequera and Bird [82], 56% of all human genes are associated with a CpG island. As described in Section 2.2.6, at the time of the publication in 1993 only limited gene annotations could be used, it is not sure how accurate the data really is. Furthermore a direct comparison with the CpG island occurrences inside the Suz12 sequences mentioned by Lee et al. [14] is not possible due to a lack of a clear definition of CpG

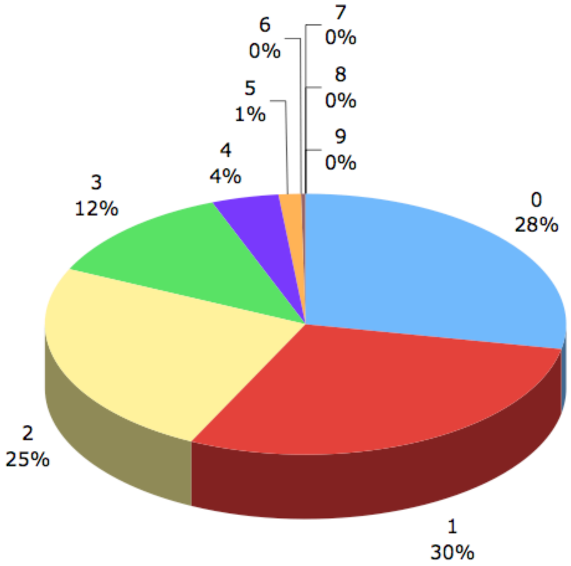

Figure 6.2: Number of low density CpG islands in 14907 Suz12 unbound promoter regions of 4 kb each. CpG island definition used in this figure: C+G content $\geq$ 50%, Observed/Expected CG ratio $\geq$ 0.6, length $\geq$ 200bp.

islands, sequence length and distances between promoters and CpG islands. Therefore as a background the list of 14907 genes is used. The sequences have the same length as the chosen Suz12 enriched sites (4 kb) and cover ±2 kb around the TSS. Because it is known that CpG islands often overlap the promoter and extend up to 1 kb downstream, the background set should cover as many CpG islands as possible. Out of the 14907 sequences, 28% show no CpG island at all and only 42% contain more than one (Figure 6.2). Interestingly, if the 14907 sequences are set to cover the region 4 kb upstream until TSS instead of ±2 kb around the TSS, the number of sequences without a CpG island raises to 52%. This supports the observation that CpG islands can extend into the downstream region of promoters. Our data suggest that CpG islands are strongly overrepresented in Suz12 bound regions. On the other hand, the definition of a CpG island is no longer the commonly used one. Instead, Takai and Jones [79] suggested to set the parameters to demand a C+G ratio $\geq$ 55%, a observed CpG to expected CpG ratio $\geq$ 0.65 and a length $\geq$ 500.

Again, the set of 2484 Suz12 regions of 4 kb each is searched. In 45% no CpG can be detected, 46% contain one CpG island and 9% two or three islands (Figure 6.3). In the background set, 62% of the promoters have a CpG island. Still CpG islands seem to be overrepresented in Suz12 bound regions, but to a lower extend compared with the previous CpG island definition. However, CpG island occur favored within promoter

		No. of strong CpG islands			
Dist to TSS	% of all Suz12	0	1	2	3
≤2 kb	60%	35%	55%	8%	1%
2−20 kb	30%	70%	27%	3%	0%
>20 kb	10%	67%	27%	5%	0%

Table 6.2: Number of high density CpG islands within Suz12 region in relation to distance to nearest transcription start site. Distance is calculated between center of Suz12 region to nearest TSS in local database (first column). Proportion of regions for different distances given in % (second column), number of high density CpG islands reaching from none to three are given in percent (right columns).

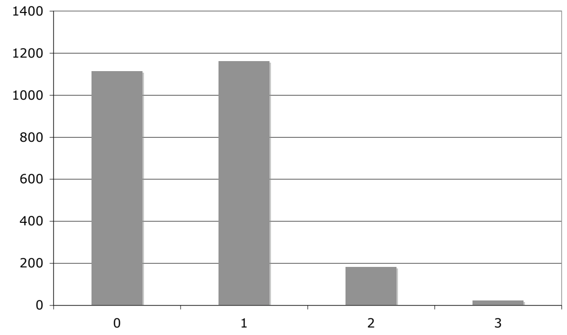

Figure 6.3: Numbers of high density CpG islands in 2484 Suz12 enriched promoters.

regions but the Suz12 sites are gained from genome-wide experiments while the background data consists of promoters only. The net enrichment in Suz12 regions could be higher. Therefore the distance of Suz12 enriched sites to the nearest transcription start site has to be taken into account.

Within a radius of 2 kb around the center of 60% of the Suz12 sites a transcription start site is found in the local database (Table 6.2). In this case the ChIP regions are most likely located inside a promoter and only 35% show no CpG island. If the distance to the nearest TSS is between 2 kb and 20 kb, which can be observed in 30% of the cases, 70% of the 4 kb regions do not contain a CpG island. If the distance exceeds 20 kb (10% of the ChIP regions), the proportion of regions without CpG island is very similar.

Thus it can be seen that promoters that have a Suz12 enriched site show high density CpG island occurrences in 65%, while promoters without a Suz12 region only contain a CpG island in 38% of the cases. Low density CpG islands occur in more than 92% of all Suz12 regions and in almost 100% of Suz12 regions inside promoters. Can this observation be used as a criterion for a genome-wide mammalian PRE or at least Suz12 recruiting site prediction? Instead of the whole human genome, promoter regions of all RefSeq annotated genes that can be referenced to Entrez identifiers are used. Each sequence is taken ± 5 kb around each TSS, all in all 15110 sequences are chosen, out of which 2022 overlap with a Suz12 enriched site. A high density CpG island can be found in 5853 sequences while out of these 1396 overlap with a Suz12 site. Although only 39% of the 10 kb promoters contain a strong CpG island, 69% of the Suz12 regions are found within these sequences. The p-value calculated by the hypergeometric distribution is $1.3e^{-213}$. This leads to the conclusion that CpG islands are overrepresented in Suz12 bound regions for both definitions of CpG islands used.

CpG islands in mouse - applying a filter

In human, Suz12 enriched sites are significantly enriched with CpG islands, regardless of the exact CpG island definition. This observation will now be tested on the mouse data. According to Antequera and Bird [82], in mouse fewer genes are associated with CpG islands (40% vs. 56% in human), but the average size of a CpG island is bigger. We want to see if this has an effect on the relation between Suz12 enriched sites and CpG islands. Our complete mouse dataset as described above covers 15729 genes in 12370 non-overlapping regions of 20 kb each. This time none of the classic CpG island definitions are used but the parameters are chosen to be sensitive enough to detect a CpG island in almost all of the 512 regions that have an overlap of all factors tested in mouse and are located within 1 kb around a transcription start site. CpG islands are predicted by newcpgreport from the EMBOSS package. The cut-off is chosen to reflect about 20 CpG repeats per 200 bp window. Out of the 512 sequences 473 are found this way. Genome-wide 9120 promoters are found. Therefore 92% of the ChIP regions are located within 60% of the tested regions. Again a significant enrichment of CpG island can be observed. Furthermore, the presence of CpG island might be used as a filter criterion to reduce the search space by 40% while only losing 8% sensitivity.

As described in Section 2.2.6, CpG islands occur around unmethylated promoters of house keeping or other essential genes that are usually expressed. In a regulatory context, CpG islands also occur methylated around promoters of regulated and therefore sometimes repressed genes. The overrepresentation of CpG islands in Suz12 bound regions is a strong indication of methylated CpG islands, which regulate homeotic genes or genes involved in stem cell proliferation in general.

Preventing CpG bias in human training set design

As described earlier in Section 6.1, we built two positive homeotic training sets for human sequences. The corresponding negative training set, again 50 sequences each, are chosen to circumvent potential biasing due to CpG island inside the positive sets. The definition of a CpG island in this case is the most sensitive one to get around the same number of CG repeats in model and background. Only 5% of the sequences in the positive training set contain no low density CpG island, around two-third contain more than two islands. The negative sequences are chosen preferably from the list of promoters that are negative for PcG proteins but show active polymerase II. Additional promoters without Suz12 enrichment in the genome-wide experiment have to be added to meet the CpG island requirements. This way we make sure that the level of CpG enrichment is equal in model and background sequences.

6.3 Motif predictions in mammals

Common motif prediction tools search for statistically overrepresented words in a provided sequence in relation to a null-model (see Section 2.1.2). Usually one single sequence or a set of related sequences can be provided. A subset of such algorithms relies on conservation of functional regions inside an alignment. In order to find motifs that are involved in PcG protein recruiting in Suz12 enriched sites, two tasks have to be performed. First, motifs that are generally involved in Suz12 recruiting should be contained in at least most of Suz12 enriched sequences (and therefore the positive training sets). To consider this in the motif prediction, the tools have to be able to work on multiple sequences at once. The second task is to eliminate the influence of other functional elements on the prediction. The Suz12 enriched sites occur preferably inside promoters, which could lead to the prediction of promoter specific motifs. Furthermore CpG islands are overrepresented in the sequences, which could lead to false positive detection of CG rich motifs. This can be prevented by using negative training sets as null-model, either directly in the prediction or in a following filtering and evaluation step.

6.3.1 Running k-words approach

A first search for motifs in Suz12 regions is perfomed by applying our k-word method described in Section 4.1. Model and background are chosen from the first of the two

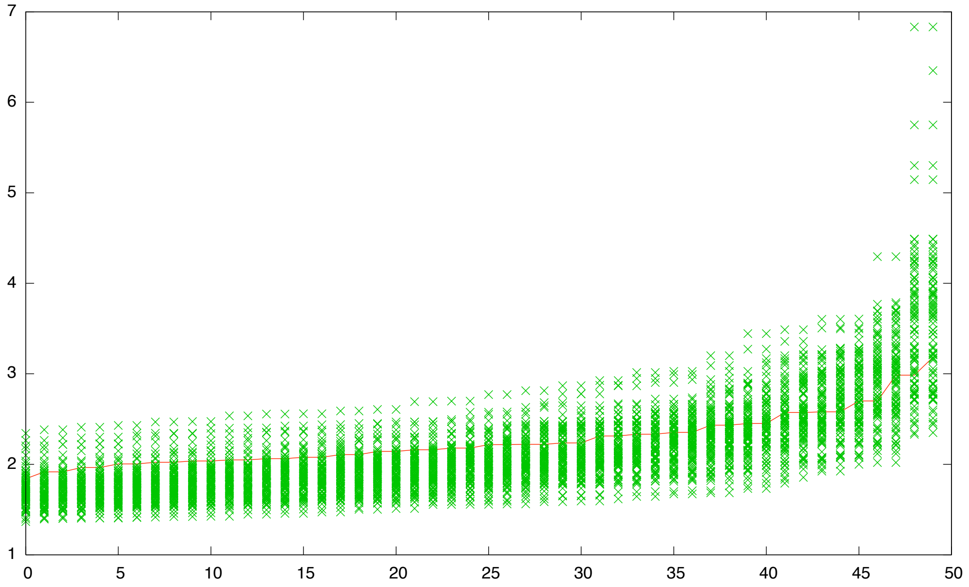

Figure 6.4: Highest scoring k-words ($k = 6$) in homeotic training set (red) and 100 randomly created sets of model and background taken from the list of Suz12 unbound promoters. X-axis=rank of hit, Y-axis=score value.

homeotic training set pairs, the positive as well as the negative training set contain 50 sequences of 4 kb each. The parameter k is set to 6, 8, and 10. As a background test the same search was performed on randomly chosen sequences. Out of the 14907 negative promoters, 100 pairs (model, background) of 50 sequences each are chosen randomly. The scores of the 50 highest scoring motifs for $k = 6$ show that values for words observed in the real sets occur in the random data as well (Figure 6.4). The promoters bound by Suz12 in the real training sets show a similar behavior to any set of chosen Suz12 unbound promoters of the same length. Even the best hit scores below most of the highest scoring hits in the random data. Are the scores of the real data below the maximized value at each position only, or does a single set exists that scores higher than the real data? We test this by calculating the sum of scores for each random set independently. The sum for the real data is 113.25 and the maximal sum of scores in the background is 142.6. For k=8 it can be seen that the scores for the k-words found in the real data are higher than almost all values in the background except for the best five hits (Figure 6.5). The sum of scores is this time 286.97, which is higher than any sum found in the 100 random sets (maximum is 281.49) With increasing k, (i.e. $k = 10$) the real data separate strongly from the background (Figure 6.6), again except only for the highest scoring hits in the background. Nevertheless, the higher scores of the real

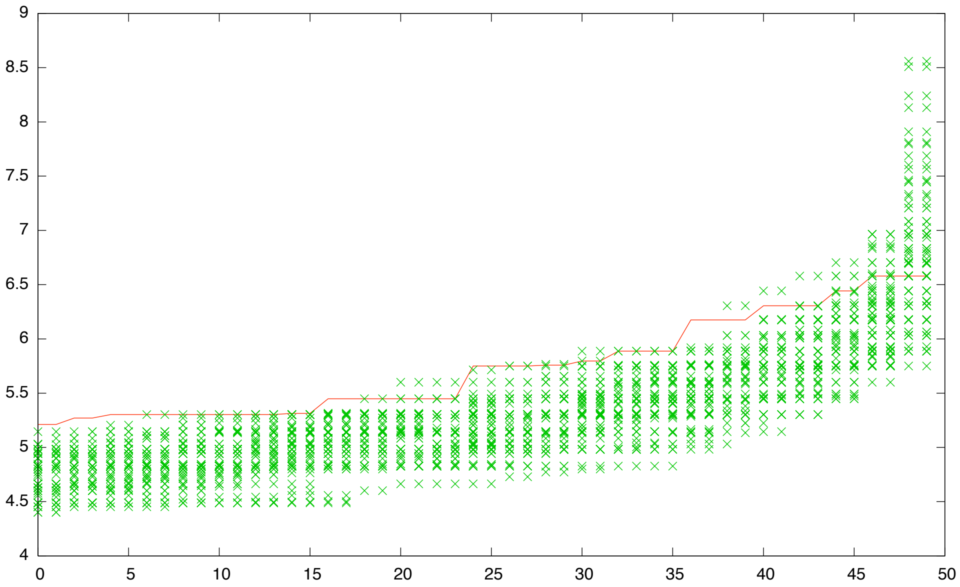

Figure 6.5: Highest scoring k-words ($k = 8$) in homeotic training set (red) and 100 background sequences chosen randomly from list of Suz12 unbound promoters. X-axis=rank of hit, Y-axis=score value

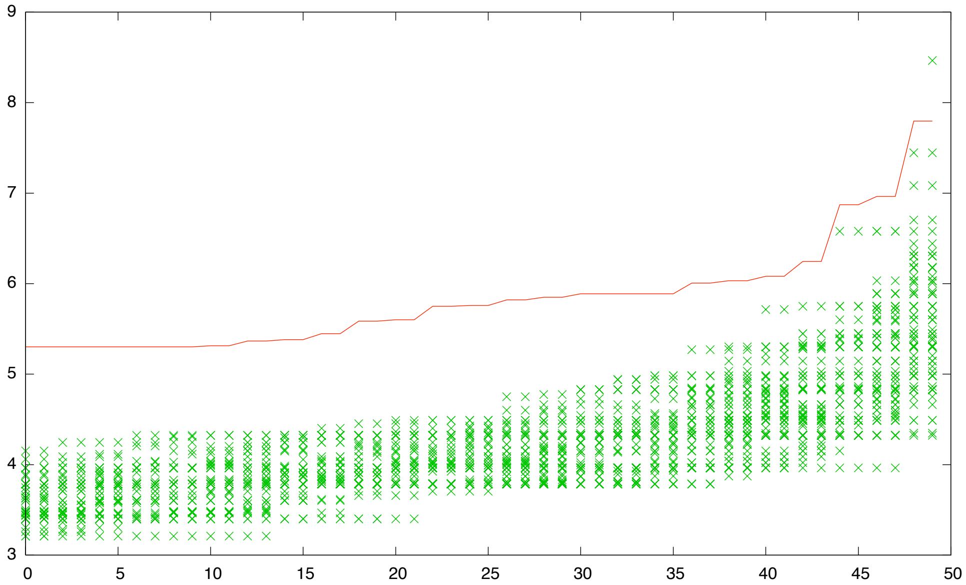

Figure 6.6: Highest scoring k-words ($k = 10$) in homeotic training set (red) and 100 background sequences chosen randomly from list of Suz12 unbound promoters. X-axis=rank of hit, Y-axis=score value.

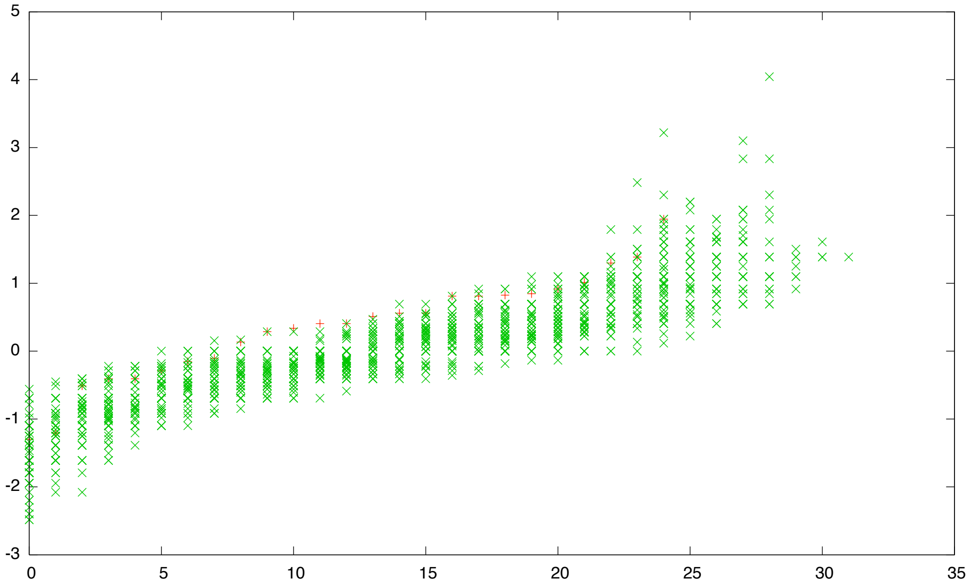

Figure 6.7: Weights of highest scoring 10-mers in second homeotic training (red) set vs. background (green), derived by taking highest scoring 10-mers from 100 random sets and weighting in 100 different sets.

data in contrast to the background can be considered significant, indicating that Suz12 specific motifs might exist (real sum=292.08, max random sum=238.73). On the other hand, none of the highest scoring motifs occurs in more than 22 (GAGGAGGAGG) sequences. A possible explanation is that motifs of this length are degenerated and the limits of this approach are simply reached. Nevertheless, the best hits are considered potential motifs for a prediction of human PRE-like or at least Suz12 bound sequences.

So far the k-word analysis has been performed on one of the two homeotic training sets. Words receiving a high score in the first set are assumed to show equal values in the second if it is about Suz12 related motifs. We assessed this assumption by using the jPREdictor to weight the highest scoring words found in the first homeotic set in the second. For $k = 6$ and $k = 8$ even best scoring words showed none to low significance, compared to the random experiments. In contrast to $k = 10$, so only the latter one is considered in the weighting step. Because the 50 best hits contain reverse complements, the dataset can be restricted to 25 sequences, or a few more in case of palindromes. The best scoring words for each of the 100 background pairs have been weighted in 100 different randomly chosen sets of Suz12 negative promoter sequences (Figure 6.7), model and background are both selected from the same data source. It can be seen that even in the real data, negative weights occur, stating that words overrepresented in one positive training set are underrepresented in another positive training set, although both sets should give positive weights to motifs related to observed Suz12 enrichment. Furthermore, none of the best 25 positions is outside the range of the random background, as it has been with the overall score values in Figure 6.6. A list of the highest scoring 8-mers and 10-mers is provided in annex A.2 and A.1.

6.3.2 Running prediction tools

A simple search for k-words over the alphabet [A,C,G,T] gives only few signals that score above the empirically determined background. Furthermore, a following weighting step in another set of training sequences even removes most of the highest scoring candidates. Compared to the background, the highest weights are still not highly significant. On the other hand, at least for *k=10* the analysis shows that some words are overrepresented and can be considered significant according to the statistic used. The observation that none of the highest scoring words occur in at least half of the positive sequences suggests that multiple words belong to the same motif. It is therefore necessary to introduce degenerated motifs by extending the alphabet to the IUPAC set or by using matrices. This can be achieved by either considering degenerated motifs directly in the prediction step, as will be seen in this chapter, or by adding a clustering step to combine similar motifs into a degenerated motif description. The latter one will be used in combination with a pipeline approach in the next chapter.

The algorithmic problem of a motif prediction has been addressed in many different implementations; a collection will be used in this chapter. To add the possibility to use a negative training set instead of a generic null-model, we will apply the motif evaluation pipeline based on the jPREdictor (Section 4.5) on the results. In [12] 13 tools were assessed by using known eukaryotic motifs taken from TransFac [104]. The tested tools cover the enumerative approaches as well as deterministic optimization and probabilistic optimization (Section 2.1.2). The results show that no single tool is able to predict all motifs in the sets. The selection of tools chosen in our analysis represents all three approaches.

Weeder [25] showed some of the best results in the different evaluating steps taken by Tompa et al. [12]. It is based on an enumerative approach and contains a clustering method to combine similar overrepresented hits. Version 1.3 has been installed locally to run on our X86 Solaris computers. The parameters are set to search for motifs of length 6, 8, or 10 that occur in at least half of the sequences and can appear on both strands. The statistical background depends on the species, we use either predefined mouse or human nucleotide probabilities. In detail, Weeder has been run

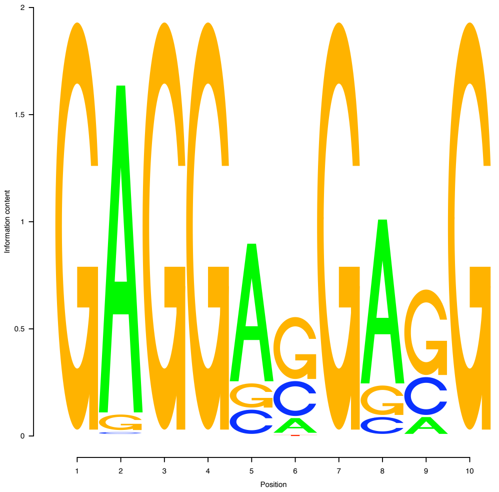

Figure 6.8: Weeder: highest scoring predicted motif in human homeotic training set. Logo created by SeqLogo [105].

on one of the two positive human homeotic sets and a 20 sequences murine set. The highest scoring hit (Figure 6.8) in the human set is of length 10 (Weeder score 1.13). Strikingly, the highest scoring 10-mer (GAGGAGGAGG) in our previous k-word analysis (Section 6.3.1) matches the matrix found by Weeder. The clustering procedure done by Weeder led to degenerations of the motif especially at positions 6 and 9. In total the matrix consists of 704 sequences found in all 50 sequences. In the murine sequences, the highest scoring motif has a length of eight. The matrix logo shows that the motif mainly consists of C and G nucleotides only. If the prediction is biased by CpG islands in the training set, a weighting against CpG rich negative sets will lead to low weights. Remember we already built such a negative training set and are going to use it in a motif evaluation step, described later in Section 6.3.4.

The classic probabilistic motif prediction approach is based on the Gibbs sampling algorithm, implemented in various tools like for example AlignACE [31], GLAM [106], SeSiMCMC [107], or MotifSampler [29]. The latter one extends Gibbs sampling to higher order Markov background models and allows one motif to occur multiple times within one sequence. We used MotifSampler on our human and murine training sets. The software has been run locally on a X86 Linux system. As background serve precompiled human (or mouse) upstream regions as a 3rd order Markov chain. The motif lengths are set to 8 and 10, the prior probability to find the motif in each of the sequences is 0.5, i.e. we expect the motif to occur in at least half of the sequences in the positive set. The maximum number of one motif in each sequence is not restricted

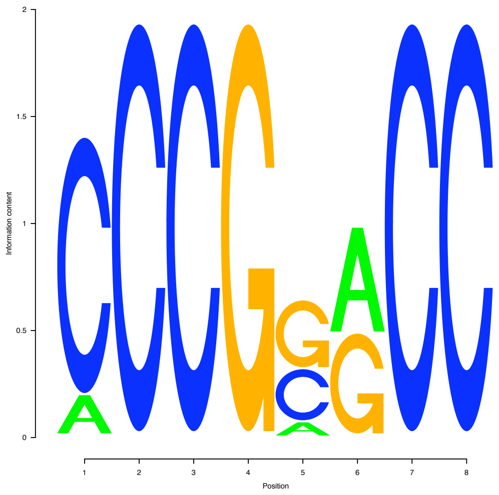

Figure 6.9: MotifSampler: Best hit in murine training set. Logo created by SeqLogo [105].

and the reverse strand is considered as well. Adjacent motifs are allowed to overlap. In the human set, a consensus sequence of GGCGGCGG is returned, the murine motif is slightly more degenerated (Figure 6.9), but still a bias towards CpG islands is very likely. Nevertheless, the motif is kept for future evaluation.

As a third tool Improbizer [108] , which is based on Expectation Maximization (see Chapter 2.1.2), is used on the training sets. Run as an online tool, the available process time is limited to 5 minutes. To avoid these constraints, Improbizer is run locally on a X86 Linux system. The background can be provided in form of a negative sequence, out of which lower order Markov chains are created. For human, the 50 negative regions from the homeotic set are used. These sequences contain a similar CpG island enrichment as the positive sequences. Calculating the background on all negative promoter regions instead, leads to a CG enrichment in the positive data against the background again, as can be seen in the murine data. The human prediction reveals two "TA" rich motifs of length 10-40, which are stored in the hit list of potential motifs, too.

Furthermore, the tool MEME [27] is applied on the same human and murine training sets. Motifs of length between six and eight are searched in up to 20 Expectation Maximization iterations (see Figure 2.1 in Section 2.1.2). In the homeotic human training set, a poly-A motif is reported. In mouse, the consensus sequence of the best motifs is TTTTTTTT, but an A is allowed at different positions.

The last two prediction runs revealed possible motifs poly-T, poly-A, or AT-repeats, which could either be a different motif or an artefact of a clustering of reverse poly-A and poly-T motifs. Poly-A as well as poly-T of length 6 to 10 are added to the list of potential motifs, together with the same number of AT repeats.

Instead of relying on the created training sets completly, we performed additional predictions with the mentioned tools on different subsets of murine and human Suz12 regions, which are chosen randomly from the set of all bound regions of length ≤ 4 kb. This time, instead of concentrating on promoters only, all positive ChIP regions are considered. Furthermore, to remove the CpG bias, the CpG islands within the sequences are masked out in some steps. In addition we ran the tool RepeatMasker [109] to remove low complexity regions in some of the sequences. As a result, different potential motifs are predicted, like 'CTAATG' found by Weeder in human sequences. Reported motifs are contained within the list of potential motifs.

Furthermore, the motifs found by the software Drim [85] (see Section 6.2) are also added to the list. The TransFac database version 11.1 contains 822 different matrices, which we search within our training sets. In the package the tools "match" and "patch" are provided to search given sequences for matches against the database. Match searches in given sequences for hits against the stored PSSMs, while patch searches consensus strings in the input. We ran both tools on our different human and murine sequences and kept those reported motifs that appear in at least 50% of each input set. In case of the human homeotic set, the search for TransFac motifs has been performed on the negative set as well. All motifs that occur in most of the negative sequences as well are not taken into further consideration. A list of the best hits of each prediction method in the human homeotic set is shown in Table 6.3.

6.3.3 Phylogenetic footprinting

In flies, PREs occur in different positions, in promoters as well as in introns or intergenic regions. As showed in the Drosophila application (Section 5), the mean distance to the nearest assumed analogous PRE increases with growing phylogenetic distance. In mammals, the situation might be different. Human and mouse show a higher sequence similarity than any of the Drosophila species. Furthermore, Suz12 sites are mainly detected in promoter regions, most of them even in very close distance to a transcription start site. To check for conservation of Suz12 regions, the distances of BLAST sites of each murine Suz12 enriched sites to the nearest human Suz12 sites are calculated.

In mouse, 1800 Suz12 enriched regions are considered which are located in promoter sequences. Each of those is BLASTed against the human genome with an E-value cut-off of 10^{-4}.

Prediction	Motif	
Weeder [25]		
	CTCSBCSA	
	GAGGVVGVVG	
	GAVGRVGA	
	CTCSBCSA	
MotifSampler [29]		
	GGCGGCGG	
Improbizer [108]		
	ATATTATTATATAAATAAATATATTTATGTAAATATTATAAAATTCA	
	AATATTATTAATATATATAAATAAATAAAA	
MEME [27]		
	AAAAAAAA	
TransFac		
	Oct1	RTAATNA
	Pax6	TTYACGCWTSA
	PPARG	TAGGTCA
	FAC1	CACAACA
	VDR	GGGKNARNRRGGWSA
	RFX	SHGWTGCSD
	POU32F2	TTATGYTAAT
	RFX1	HRGYAAC
	NKX	HSYCACTTS
	GATA-4	AGATADMAGGGA
	CdxA	AWTWMTR

Table 6.3: Results of motif predictions on suz12 bound regions that are located in promoter regions of human homeotic transcription factors.

Due to the stringent cut-off, only 931 homologous positions can be identified in human. About two-third (around 600) of those positions are overlapping with a human Suz12 site, 100 are within a distance of 10 kb to the nearest Suz12 region, and in 200 cases no Suz12 site is found in human anywhere in the area around the gene. The data show that the nearest Suz12 site is either located very closely to the homologous site, or is more than 10 kb away. Especially if one considers a false-negative rate of the ChIP experiments of up to 30%, even the 22% examples without a detected Suz12 region nearby the homologous region do not suggest any "evolutionary plasticity", as observed in the Drosophila study. A phylogenetic footprinting approach is reasonable and presented in the following.

Data

This observation is now used in a phylogenetic footprinting approach based on alignments of Suz12 enriched sites in different mammalian species. The alignments are taken from the Multiz17way [91] entries in the UCSC genome database [110]. The positions of selected Suz12 enriched sites, like the ones in the homeotic training set, are sent by a Perl script via HTTP requests. Only sequences from the species human, mouse, rat, dog, and chimpanzee are kept in the alignments. All in all 100 alignments are received, based on human Suz12 positions. The sequences are chosen by length (≤ 2 kb for the human Suz12 site) and availability as alignment. At least the murine sequence has to be contained in the alignment.

Phylogibbs

The tool Phylogibbs [30] extends traditional Gibbs sampling to a phylogenetic footprinting approach. It works on one alignment and looks for motifs that show conservation above background. Because Phylogibbs handles each alignment independently, further actions have to be taken to find those motifs that are potentially related to Suz12 recruiting and occur therefore in most sequences. Phylogibbs returns matrices of degenerated motifs in each run, which have to be clustered to a single motif if they refer to the same transcription factor binding site. Previously used tools have implicit built in clustering methods. As standalone clustering applications, MATLIGN [36] and Phyloclus [38] are chosen because both are designed to work on matrices. Phylogibbs version 1.1 is run on all 100 alignments provided in FASTA format, including gaps. Parameters set are "-D 1" (consider phylogeny), "-m 10" (motif length of 10), "-S 100" (100 steps in the tracking phase), "-N 0" (order of background Markov chain), and the option "-L (((hg17:0.85,panTro1:0.9):0.6,(mm7:0.8,rn3:0.9):0.7):0.9,canFam2:0.7)" (pro-

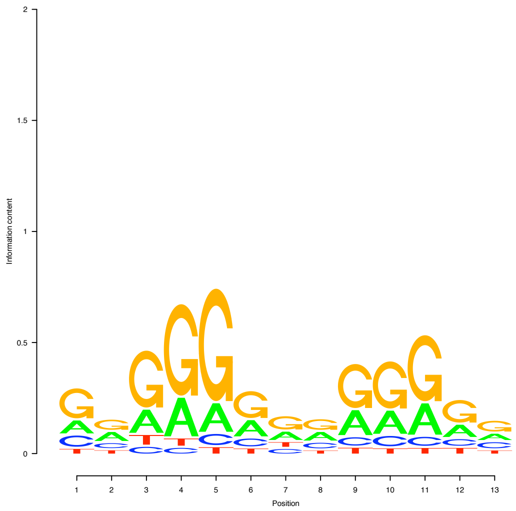

Figure 6.10: MATLIGN cluster of Phylogibbs results. Biggest cluster (26 members) shown.

vides phylogenetic tree). The output of the tracking phase is converted into input files for MATLIGN as well as Phyloclus. The number of combined motifs differs drastically between the two tools. MATLIGN combines 31 motifs into one, reaching an information content of 0.75 at the highest position (Figure 6.10). Phyloclus combines only eight predicted motifs in four clusters, which consist of two motifs each (an example is given in Figure 6.11). Nevertheless, both clustered motifs show big similarity if the reverse complement is taken. The motifs match G or A repeats, respectively the complementary C or T.

Phyloclus is run with 100 iterations in the pre-processing step, the motif length is set to 10. The same parameters are used in the post-processing step.

Additionally, the 50 sequences in the homeotic training set are searched by Phylogibbs. Again, results are clustered by MATLIGN to combine predicted motifs in different sequences. The largest cluster consists of five elements and matches the GA rich 10-mer found by both our k-word approach and Weeder. The reverse complementary CT rich motifs are combined in a second cluster.

Applying phylogenetic footprinting pipeline

A combination of phylogenetic footprinting, masking of found hits, combining multiple prediction tools, and clustering is implemented in the phylogenetic footprinting pipeline described in Section 4.2. We used the pipeline on Suz12 data. The input is the same as in the single Phylogibbs run, 100 positions of human Suz12 enriched regions

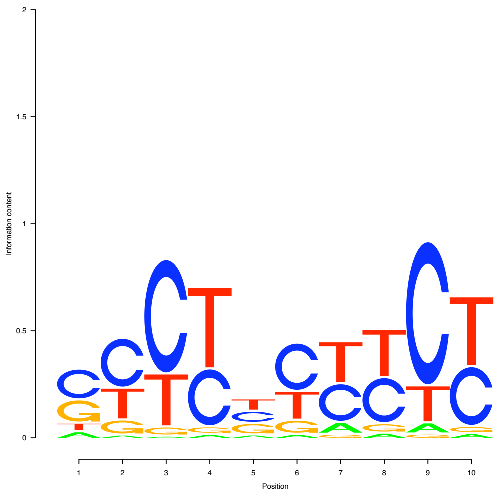

Figure 6.11: Phyloclus cluster of Phylogibbs results. Biggest cluster (2 members) shown.

are provided to the pipeline, which queries the UCSC database for Multiz17way alignments. Again, the constraints are set to limit the species in the alignments to mammals only (human, mouse, dog, rat, and chimp). In addition to the human sequence, murine sequences have to be present in the alignment. All in all, 100 alignments serve as pipeline input.

In the fist step, the alignments are transformed into the different input formats and processed by MEME, Footprinter and Weeder. After each run, hits are masked and the pipeline is repeated. The results in each step are added to a global motif list for further processing. After the first MEME run, 11%, and after the second run 19% of the input sequences are masked. In total, 126 motifs are predicted by MEME, out of which 4 are clustered together by MATLIGN. The motifs found in the second iteration are similar to the ones found in the first step. In both cases mainly repetitive motifs are reported. Even after the second iteration, the resulting cluster shows mainly GA repeats.

In the first iteration of Footprinter, run on unmasked input alignments, the best overall cluster consists of seven clustered motifs. That means that the motif can only be found in seven out of 100 sequences. In the second iteration, only four single alignment clusters are combined to one cluster. None of the predicted motifs meets the constraints to occur in at least 50% of the input sequences. Nevertheless, all motifs show a low rate of degeneration at some positions, the maximum information content reaches up to one. Additionally to MATLIGN, the two-step clustering of Footprinter results is also done by Phyloclus, which puts 714 out of 1092 overall motifs in 315 different clusters. In average, a cluster consists of less than three elements. The biggest cluster is built from 10 elements, but like MATLIGN, the degeneration rate is low (MASMAGCCGS). While the MEME results are very similar to the previous run on not aligned sequences, Footprinter reports non-repetitive motifs that show low degeneration rates, but occur in only very few sequences in the input data.

The third prediction method implemented into the pipeline is Weeder. All 100 alignments, which sum up to 460 sequences in total, are put into one single Weeder run. The motif occurring most often is of length six (AGCGCG), found in 200 out of 460 single sequences. Longer motifs are only found in eight or nine different sequences.

Since none of the motifs predicted in the pipeline so far occurs in at least 50% of the input data, they are used as prior knowledge in the Phylogibbs step. Except for the short Weeder motif, no other motif occurs more often than in 10% of the input. The biggest Phyloclus cluster after two Footprinter iterations combines ten elements, the second one nine, and another four clusters have four or five members. The MATLIGN clustering results of the same Footprinter predictions contain only up to seven elements. These clusters combined with the MEME and Weeder outputs are given to Phylogibbs, which is run on each single input alignment independently.

The Phyloclus cluster of the Footprinter predictions that contains five elements is matched to nine motif predictions by Phylogibbs. The other clusters occur in even less alignments in the Phylogibbs run. Phylogibbs applied to MATLIGN clusters returns motifs that occur in almost all input sequences, but consist of single nucleotide repeats with low information content. The Weeder and MEME results lead to predictions of widely found motifs, the AGCGCG motif is turned into GC-repeats by Phylogibbs and is found in most of the input sequences.

The resulting motifs of the phylogenetic footprinting pipeline are either found in only a small subset of the input sequences, or show strong rates of degeneration, or are part of repetitive regions. Nevertheless, the results are kept as potential motifs for an evaluation step.

6.3.4 Motif evaluation

The runs of different motif prediction tools on the different sets of Suz12 enriched sequences led to a list of potential motifs, despite the observed tendency to repetative sequences. A motif list containing all potential binding site descriptions includes about 60 entries. The motif evaluation algorithm described in Section 4.3 is now applied to evaluate the motifs' ability to seperate a positive set of Suz12 enriched promoters from sets of not enriched promoters. As a control experiment, the same algorithm is applied

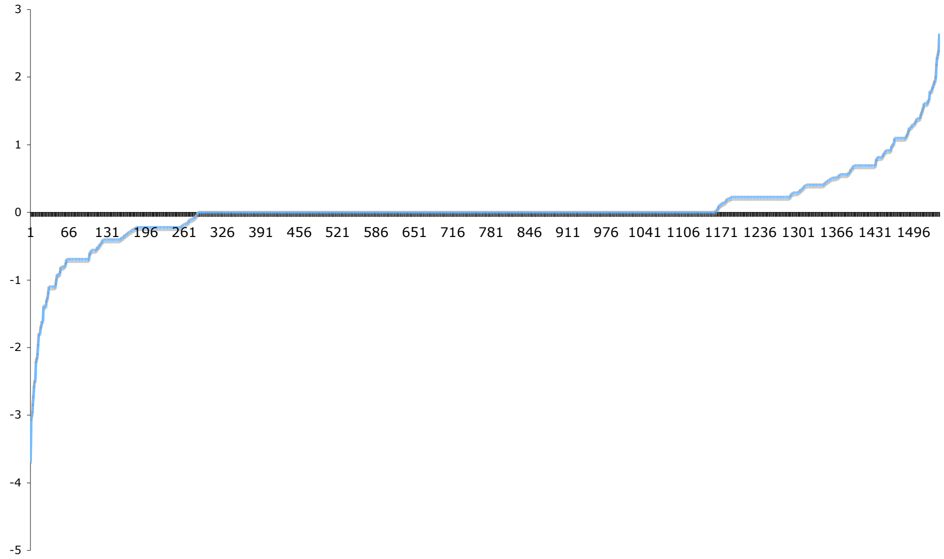

Figure 6.12: Weights of all motif pairs build from 60 single motifs. Most pairs have weight close to 0, around 100 have a weight smaller -1 or greater +1.

to Drosophila data, in which case the motifs are known and have shown to be able to at least partially seperate model from background.

Results

Homeotic 1

As a first application, the pipeline is used to evaluate our potential mammalian motif set, which contains 60 single motifs. The first training set used as model and background is the human homeotic set 1, as described in Table 6.1. The distribution of motif pair weights (Figure 6.12) shows that most motif pairs get a weight close to 0, while few pairs show strong positive or negative weights and therefore should have the biggest impact on the prediction score. If all motif pairs exceeding an arbitrary weight threshold of 1 are selected, the motif set would contain about 60 motif pairs.

However, comparing the highest score of σ_M with σ_B (Figure 6.13) filters out most of these motif pairs ($\mu_m = 10$, $\mu_b = 5$, $\mu_s = 3$). Even high weighting motifs can get high scores in single sequences in the negative set. This reduces the difference between σ_M

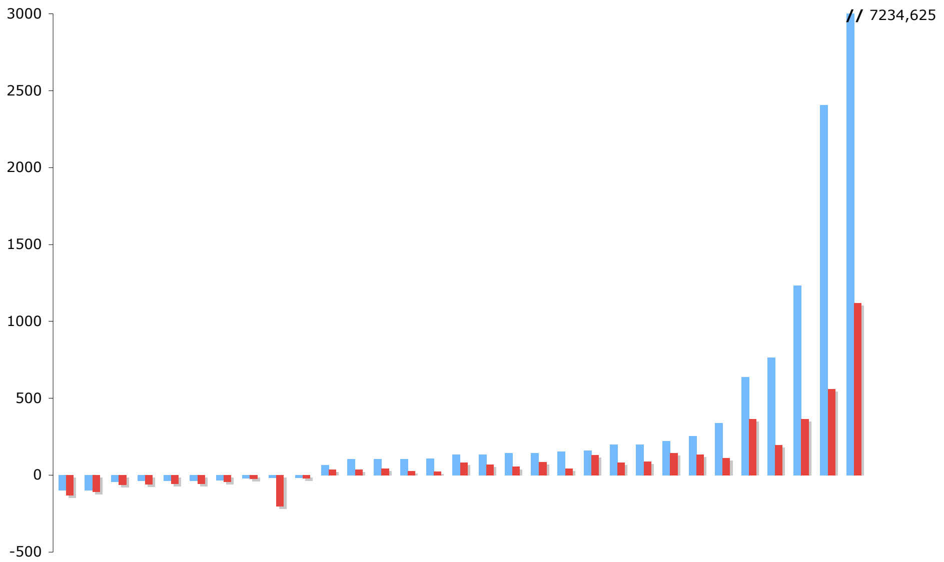

Figure 6.13: σ_M (blue) and corresponding σ_B (red) values for the 10 lowest and 20 highest scores in σ_M. X-axis gives the motif pairs, Y-axis gives corresponding values cropped to 3000. Highest σ_M value at 7234.

and σ_B, and adds only low information content to the motif set, leading to an increased cut-off, which only weakens sensitivity without strengthen specificity. Furthermore, one motif pair alone contributes extremely high to the scores, with a more than three times higher value of σ_M than the other motif pairs. In the first pipeline run, only one sequence in the positive training set scores above the preliminary cut-off, and only four motif pairs are kept in the motif set from which the highest scoring pair alone is sufficient to reach the required cut-off in the found sequence. These results strongly indicate, that the found sequence biases the motif weights due to repeats of the motif sequence. The motif pair providing alone about 75% of the overall score is the drim4 (DGAGAGV) motif paired with itself.

In the second iteration the same motif pairs are used, but the one high scoring sequence is removed from the positive training set. The background remains the same. While in the first iteration the values of σ_M reach up to 7234, the achieved maximum is the second iteration is only 154. Furthermore, the preliminary cut-off drops from 1024 to 45 and is not reached by any sequence in the positive set. Although more

Motif 1	Motif 2
CACACACA	GGGGTNCC
CACACACA	GCTGCNBB
CACACACA	GGGRTGGG

Table 6.4: Motifs pairs left after second iteration in homeotic set 2

than 60 potential motifs are introduced and combined to 1230 motif pairs ($\binom{60}{2} + 60$), none of the pairs can be used to distinguish the positive from the negative training set.

As mentioned during the motif prediction runs on human sequences, the generated motif list contains poly-A and poly-T respectively as well as AT repeats of various length. None of those received a positive weight in the homeotic training sets, neither as single nor as double motifs. The single motif AT$_{10}$ alone received a weight of -2.1 and poly-T$_{10}$ of -1.1. The double motif AT$_6$: AT$_{10}$ is strongly underrepresented in the positive set with a weight of -5.9. The T$_{10}$: T$_{10}$ double motif is weighted -2.4, again strongly underrepresented in homeotic promoters.

Homeotic 2

A control run on the second homeotic set does not lead to a hit in the positive training set in the first iteration. However, if the required distance between the score for a motif pair in the model and the background is increased (i.e. $\mu_s = 5$ instead of 3), the drim4:drim4 double motif is removed ($\sigma_M = 2700$ and $\sigma_B = 800$ in homeotic set 2). The resulting motif set consists of three pairs with positive (Table 6.4) and six pairs with negative scores. As can be seen in the table, each of the three motif pairs contain the same repetitive single motif. The preliminary cut-off is 61, which is reached by only 3 out of 50 sequences in the positive set. Again, a repetitive element is part of every motif pair.

Murine sets

Because the murine training set has not been designed to be enriched with CpG islands, the result reflects the previously observed CpG island enrichment in Suz12 bound regions. Motifs that show high CG content receive very high values for σ_M. Removing the motifs or the CpG rich sequences in the positive set leads to an absence of significant motif scores. After the first iteration, no separation between model and background can be achieved.

Drosophila

The Drosophila PRE prediction serves as a positive control. The training sets as well as the motif sets are taken from the original 2003 analysis [11] (En1, GAF, G10, PF, PM, PS, Z). The evaluation algorithm selects 13 out of all 28 motif pairs. Each of the seven single motifs occurs in at least one motif pair, which means that every motif used in the study actually contributes to a separation between model and background. The highest σ_M value is achieved by the motif pair GAF:PF. Seven out of the twelve sequences in the positive training set score above the preliminary cut-off. The genome-wide cut-off drops from 157 to 152.

As a control for the robustness test, the positive training set is extended by a 3 kb random sequence, into which 24 copies of the word (GTGTGTGT) are inserted. The same word is entered into the motif list as an additional motif "Test". The result of the evaluation algorithm is as expected, in the first iteration only the motif pair "Test:Test" is selected and only the placed GT rich sequence is found. The sequence is automatically removed and the next iteration is run on the original set.

6.4 Dinucleotide repeats (CpG islands, GA repeats)

All approaches used for the prediction of potential motifs that are related to Suz12 recruiting in mammals, returned some motifs of single or dinucleotide repeats (see Section 6.3). The analysis of CpG islands in the Section 6.2 explains CG rich k-words, but additionally GA and CA rich motifs are found. In flies, GA repeats up to the length of ten are known as the G10 motif, a double repeat of the GAGA factor binding site. The GA repeats predicted in mammals on the other hand, show strong rates of degeneration, some reported matrices have a basic consensus sequence of poly-R[1]. Repeats of C and A are found to a lesser extent. Interestingly, although AT repeats are reported as well, the double motif TATATA:TATATA shows a high difference between σ_M and σ_B (-36 to -3460) and is more likely to occur in Suz12 unbound promoters.

A motif evaluating pipeline run on a motif set containing additionally AC, AG, and AG repeats of length 6, 8, and 10 calculates high σ_M values in the homeotic set 1 and 2 in the first iteration (Table 6.5), while in the murine set no motif pair meets the requirements. Still, only 3 out of 50 sequences score above the preliminary cut-off. Removing the found sequences in the second and third iteration and running the pipeline again shows no enrichment of AG repeats anymore, only 6% of the positive sequences contain AG repeats. Instead, to a lesser extent AC repeats are reported in another four positive sequences.

[1] R in IUPAC = purine (A or G)

Motif	σ_M	σ_B
AT6:AT6	-232	-2159
AT6:AT8	-200	-5632
...		
AG10:AG8	1392	111
AG10:AG6	1689	220
AG6:AG6	3177	459
AG6:AG8	3241	434

Table 6.5: Sigma-values (σ) of dinucleotide repeats in human homeotic set. Pairs with similar values for positive and negative set not shown.

Depending on the training set, repeats of AG or AC seem to be overrepresented. Analyzing those repeats on a pure motif level just leads to the shown results. Alternatively, one could define repeats not in form of degenerated motifs, but similar to CpG island as regions of overrepresented AG (or AC) content.

AG repeats

We define an AG "island" as a sequence with a length ≥ 40 that has an AG content $\geq 75\%$ and an expected vs. observed ratio of AG dinucletotides of ≥ 0.6. In the homeotic sets, 60% of the positive sequences contain at least one "island" (30 out of 50 in each set). Of all 2484 Suz12 regions of length 4 kb, 47% contain at least one "island" (in total 1182). In promoters without Suz12 enrichment, an "island" is found in 39% (5857 out of 14907) of the cases, which indicates an enrichment in Suz12 regions.

Considering the background set of 15110 human promoters (10 kb each), 42% contain an "island" (in total 6461). All in all 1335 promoters are Suz12 enriched, out of which 56% additionally contain an AG "island" (749). The p-value calculated by hypergeometric distribution is *9e-25*. Therefore AG "islands" are significantly enriched in Suz12 bound region. The overlaps of promoters with high density CpG islands, AG "islands", and Suz12 enrichments (Figure 6.14) show that 8% of all promoters are bound by Suz12, 11.5% of the AG "island" containing promoters are bound by Suz12, 14.7% of the CpG island containing promoters are bound by Suz12, and finally 18.9% of the CpG island and AG "islands" containing promoters are bound by Suz12.

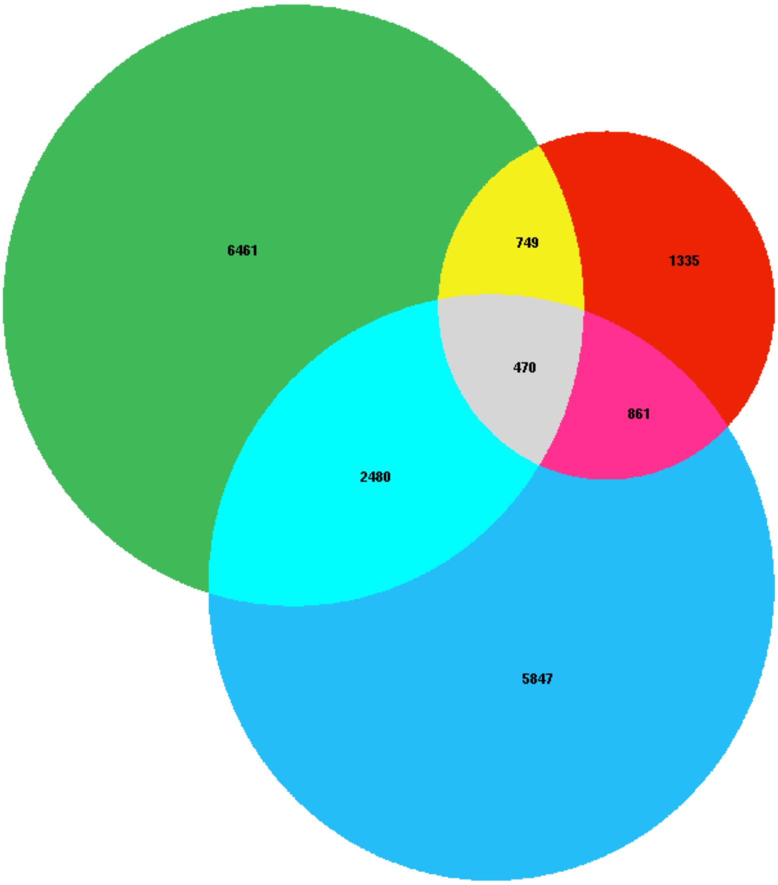

Figure 6.14: Overlaps of 15110 promoters that are Suz12 enriched (red), high density CpG islands (blue), or AG "islands" (green). Overlaps of CpG and AG is shown in cyan, of CpG and Suz12 in pink, of AG and Suz12 in yellow and overlaps of all three in grey.

AC repeats

In a similar way regions of enriched AC repeats are determined, defined as regions of length ≥ 50, an observed/expected AC ratio of ≥ 1 and an AC percentage $\geq 65\%$. The parameters are chosen to find 60% of the homeotic training set again. A search in all 2484 Suz12 regions of 4 kb each, shows a similar result of 1461 hits (59%). The set of 15000 negative 4 kb promoters contains 8579 AC regions (57%), therefore low to none enrichment is detected.

Additionally, the set of 15110 human promoters of 10 kb each is tested. In the longer sequences more AC regions can be found, 12,802/15,110 (85%) contain at least one AC region. In 1335 promoters, a Suz12 enriched site is reported, out of which 1172 also have an AC region (88%). The p-value of 0.0006 shows only low significance. A slight enrichment of AC repeats might be present in Suz12 bound regions, but is too low to be used in predictions.

7 Discussion

In Chapter 3 we described a framework for the enhancement of predictions that are based on genome-wide scoring, called *DynScan*. The method takes benefit from prediction runs in related species and uses those as prior knowledge to increase sensitivity without losing specificity by scoring stepwise increased windows around orthologous sites. The threshold directly depends on the search space, rewarding small radii around the orthologous sites while still staying alignment independent.

The prediction of Polycomb/Trithorax Response Elements (PREs) in Drosophila serves as an application for the method (Section 5). PREs are *cis*-regulatory elements that take over from enhancers once expression patterns of developmental genes are set and maintain the status over many cell division cycles. The jPREdictor [10] and the corresponding parameters gained in the preceeding PREdictor [11] work provide a tool capable of a highly specific but less sensitive PRE prediction. Using jPREdictor in combination with *DynScan* on a set of four different Drosophila species not only increases sensitivity, but also reveales extraordinary insights into dynamic processes during PRE evolution, which we refer to as evolutionary plasticity. All observations are strikingly consistent with the results of biological experiments, proving the general capabilities of our method.

First type of evolutionary plasticity

First, the number of predicted PREs differs drastically between different Drosophila species, an observation that is also reflected in the number of different bands in giant polytene chromosomes. Although the exact number of PREs cannot be determined by any available method, the prediction as well as the experimental data show around twice as many hits in *D.pse* than in any melanogaster subgroup species. We can say that species that are more closely related to each other such as the members of the melanogaster subgroup show similar amounts of PREs while, the number can differ drastically if the evolutionary divergence grows larger.

A possible biological reason for this observation is that an orthologous gene is regulated by a different number of PREs in different species. Following that idea, additional PREs in *D.pse* could be a relic of ancestral loci that contained multiple PREs for the regulation of a single gene. For example, we can detect an additional PRE in the *D.pse* Bithorax complex that can only be predicted in species outside the melanogaster subgroup. We assume that this PRE regulates the gene *Abd-A,* which is regulated by three PREs in *D.mel* (iab-2, iab-3, and iab-4). The additionally predicted PRE in D.pse is located between iab-3 and iab-4. In species where the additional PRE is absent, its functionality could be taken over by the remaining PREs near the gene. On the other hand, *D.pse* shows phenotypical diversity from *D.mel*. The species differ dramatically in the number of sex-combs, a trait that is known to depend on PcG regulation [97]. Furthermore the species differ in size, body shape, color and even the choice of habitat. Differences that at least partially could be related to the observed variety of PREs.

Second type of evolutionary plasticity

As the second type of evolutionary plasticity we observe that genomic positions of PREs change rapidly during evolution. A search between more closely related species like *D.mel* and *D.sim* finds an assumed functionally analogous PRE in close distance to the orthologous PRE positions in most of the cases. However, the orthologous position itself shows no significant prediction score in the vast majority of elements. The distribution of distances between orthologous site and nearest predicted PRE strongly correlates with the phylogenetic distance between the species. While between *D.mel* and *D.sim* most analogs are located within 1 kb around the orthologous position, the majority of the *D.pse* PREs show no functional analog within 10 kb around their orthologous site in *D.mel*. This can be explained by the first type of evolutionary plasticity. For many additional PREs in *D.pse* no functional analog exist in *D.mel.* This even supports *DynScan*'s claimed specificity. Although a lower cut-off is used, no hit is found within 1 kb in those cases. This effect is expected considering the difference in overall PRE numbers and only supports our method.

ChIP experiments on the orthologous sites as well as on the predicted analogs were performed on chosen examples. In cases where the orthologous sites show neither PC nor PH protein binding, strong enrichments can be detected at the predicted functional analagous sites. This demonstrates actual PcG protein recruiting on the sites that are predicted by *DynScan* and show no genome-wide significant score. Additionally, transgenic fly assays confirm PRE functionality in terms of pairing sensitive silencing, eye color variegation and response to PcG and trxG mutations in several selected cases.

Third type of evolutionary plasticity

Finally, we demonstrate that even for those PREs that indeed are positionally conserved, clusters of functional motifs are usually located outside of highly conserved spots. Even PREs at orthologous positions differ in composition and positions of functional motifs. Our data show that motif turnover is not an exception but the rule in PRE development, even between more closely related species.

DynScan specificity and sensitivity

A comparison of the prediction results with available *D.mel* ChIP data shows again the boost in sensitivity gained from *DynScan* (Section 5.3). It can be shown by p-value calculation that the *DynScan* predicted PREs are of same specificity based on the ChIP experiments as the regular predicted ones. Furthermore, a comparison of different prediction parameters shows that *DynScan* leads to a higher increase of sensitivity by keeping specificity level than can be achieved by adding DSP1 or SP1/KLF as additional motifs. The *dpp* examples demonstrates that the "classic" motif set shows more coverage with experimental results. Adding new motifs could increase sensitivity in some cases, e.g. in *D.pse* a PcG protein enriched site near the *dpp* gene can be predicted, but other validated PREs fall below the threshold and regions without protein binding that scored below any cut-off earlier are increased in scores. Changing the prediction's parameter set affects already validated prediction results, while applying *DynScan* only allows additional hits, without changing previous results.

Still not all ChIP positions provided by Schwartz et al. [71] can be confirmed by our prediction. This can be explained by the genome-wide ChIP inherent problems. Enrichment of PcG proteins alone does not prove the presence of canonical PREs that contain the few known motifs as functional elements. The algorithm is trained on such PREs regulating homeotic genes. Thus other types of regions that recruit totally different types of DNA binding proteins may escape detection, and to this point even biological characterization. Furthermore, the presence of PcG proteins in promoter regions does not necessarily prove a recruitment of DNA binding proteins at those positions. Instead, indirect binding by looping from nearby motif containing PRE regions has been reported [74]. In this case no positive prediction result can be expected at those promoter positions.

Finally low sensitivity of genome-wide ChIP experiments may lead to further discrepancies between prediction and experiments. Absence of genome-wide ChIP enrichment does not contradict prediction results at a high scoring position. PRE activity is strongly tissue specific. Second, antibody affinities affect the ChIPs sensitivity. Furthermore, thresholds are applied to ensure statistical significance, which allow real positives to fall below the cut-off. The estimated false-negative rate of 30% for genome wide ChIP on chip data [13, 14] explains why some of the predicted PREs are not confirmed by genome-wide ChIP experiments but are confirmed in other experiments including the Fab-7 PRE [11], which demonstrates that the predicted PREs are not false positives.

In this work, we showed that the *DynScan* method indeed increases sensitivity in case of PRE predictions. Other applications could also take benefit from *DynScan*. For obvious reasons, any prediction that can be based on the jPREdictor will work. An example would be enhancer prediction. Other scoring algorithms beside jPREdictor could also work, as long as a cut-off is calculated as a trade-off between specificity and sensitivity. With ongoing improvements in large-scale biological experiments such as genome-wide ChIP on chip, it also could become possible to skip the prediction step and to run *DynScan* on those data directly. The Suz12 experiments in human and mouse showed which amounts of data are to be expected in the future, although similar studies in different Drosophila species are still missing. Nevertheless, since the evaluation of ChIP data is based on thresholds to ensure statistical significance, each genomic positions is assigned some kind of significance value, such as fold enrichment. Given the raw scores for each genomic positions in different species and the desired significance level, we could use *DynScan* to increase sensitivity of large scale ChIP experiments.

Evolutionary studies

To give answers about potential mechanisms involved in the observed dynamics of PRE development, we followed the idea of pre-PREs, regions that contain motifs or presites that can be turned into functional motifs by minor mutations but have not developed PcG protein recruiting potential (Section 5.2). Presites adjacent to an existing PRE may acquire new functional motifs and replace former functional motifs in other positions, allowing the PRE to “creep” from one site to the other. Sequence insertions could accelerate this process. We observe such a local shift in the *spalt major* PRE, in which a single motif cluster spanning approximately 600 bp in the *D.mel* PRE has split into two clusters in *D.pse*, which are separated by an insertion of a few hundred base pairs. We followed the theory by developing a scoring scheme for presites of different transcript factor binding sites (Section 5.2) and applying the scoring on orthologous positions of experimentally tested PREs that show a lack of PRE functionality but have a validated PcG binding within a few kilobases. The respective inactive sites show higher PRE potential scores than other regions nearby, indicating that presites indeed exist and serve as a highly probable explanation for shifting of PRE positions as well as for possible de-novo evolution of PREs. The data further show that the Pho motifs seem to play a very important role and might be neccessary for PRE functionality, while other motifs might assist in PcG protein recruiting but are not neccessarily essential.

Tracking PREs through their evolution by scoring regions around orthologous positions in all twelve Drosophila species again shows strong dynamics. For example, in case of an additional Bithorax complex PRE that is predicted and validated in *D.pse* but not in *D.mel*, *D.sim*, or *D.yak,* it can be assumed that the functionality changed at two points during the evolution. This shows that the PRE was lost in the Bithorax complex of the melanogaster subgroup species but survived in *D.pse*.

In other cases, a PRE can be predicted in some but not all species, independent of their phylogenetic relation. In general, the absence of a significant prediction score does not prove absence of PRE functionality. On the other hand, the sensitivity strongly depends on the "type" of PREs, classical homeotic PREs are the prediction's main target. If a PRE is predicted in some species, but no score peak is found within a reasonable radius around the same position in other species, at least it can be suggested that the type of PRE found in some species is absent in others. A loss of functionality as well as a possible substitute by a different kind of PRE are both signs of dynamics in PRE evolution.

Mammalian studies

The success in the PRE application for *DynScan* and the availability of large-scale ChIP data for PcG proteins in mammals made us take a deeper look into potential mammalian PREs, although not a single one has been found experimentally yet. We ran different motif prediction algorithms that cover all known theoretical approaches (Section 6.3), and implemented a pipeline based on phylogenetic footprinting that considers all common "rules" of motif prediction: Different tools are used, results of each run are masked, different clustering strategies are performed on results and different source sequences are chosen (Section 4.2). Furthermore an evaluating algorithm has been developed that tests potential motifs for their capability to separate model from background sequences by considering motif weights and number of occurrences in positive and negative training sets, combined with a robustness test (Section 4.3).

The method has been tested on Drosophila sequences and shows that it is able to detect biasing sequences that have strong influence on the scoring output by containing repetitive elements. Furthermore the method puts additional constraints on jPREdictor's weighting scheme, forcing even high weighted motif pairs to occur in a sufficient amount of positive sequences. This way the net effect of motifs on separating positive from negative sequences can be estimated (Section 6.3.4).

All motif prediction data in human and mouse test sequences indicate the absence of classic transcription factor binding sites for recruiting DNA binding members of the PcG complex. Instead, all used methods come to the conclusion that repetitive elements, namely CpG islands and GA rich regions are the only predicted motifs. We thereofore suggest that CpG islands are part of Polycomb induced gene silencing. We proved that CpG islands are statistically strongly enriched around Suz12 sites in mouse and human (Section 6.2). This fact could be used as filter for potential future computational predictions. Concentrating on CpG islands only reduces the search space by 40% while only losing 8% sensitivity.

To a lesser extent this is also true for GA rich regions which are still overrepresented in Suz12 regions but with lower significance. One could speculate that GA repeats we found on any strand are signs of any GA, AG, or reverse complimentary TC, CT enrichment, which could be involved into looping of DNA on itself in form of a triplex structure which builds a DNA-H form [111]. Such DNA structures, involved in transcription silencing, have been observed for different kinds of monopurine-monopyrimidine repeats like C-G or CT-GA.

In summary, with *DynScan* we presented a novel approach that reveals extraordinary dynamics in PRE development in Drosophila, adding new knowledge about evolution of *cis*-regulatory elements in general. Furthermore, by combining current knowledge about motif predictions with a new evaluation algorithm, we indicate different and unrevealed processes to be involved in PcG protein recruitment in mammals.

Bibliography

[1] Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–9 (2007).

[2] Flicek, P., Keibler, E., Hu, P., Korf, I. & Brent, M. R. Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Res* **13**, 46–54 (2003).

[3] Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988–95 (2004).

[4] Hubbard, T. J. *et al.* Ensembl 2007. *Nucleic Acids Res* **35**, D610–7 (2007).

[5] Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61–5 (2007).

[6] Wilson, R. J., Goodman, J. L. & Strelets, V. B. FlyBase: integration and improvements to query tools. *Nucleic Acids Res* **36**, D588–93 (2008).

[7] Busturia, A., Casanova, J., Sanchez-Herrero, E. & Morata, G. Structure and function of the bithorax complex genes of Drosophila. *Ciba Found Symp* **144**, 227–38; discussion 239–42, 290–5 (1989).

[8] Hauenschild, A. *Vergleichende Analyse von Polycomb/Trithorax Response Elements in Drosophila* . Ph.D. thesis, Bielefeld University (2005).

[9] Tagle, D. A. *et al.* Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* **203**, 439–55 (1988).

[10] Fiedler, T. & Rehmsmeier, M. jPREdictor: a versatile tool for the prediction of cis-regulatory elements. *Nucleic Acids Res* **34**, W546–50 (2006).

[11] Ringrose, L., Rehmsmeier, M., Dura, J.-M. & Paro, R. Genome-Wide Prediction of Polycomb/Trithorax Response Elements in Drosophila melanogaster. *Dev Cell* **5**, 759–771 (2003).

[12] Tompa, M. *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23**, 137–44 (2005).

[13] Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349–53 (2006).

[14] Lee, T. I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301–13 (2006).

[15] John, B. *et al.* Human MicroRNA targets. *PLoS Biol* **2**, e363 (2004).

[16] Lim, L. P. *et al.* The microRNAs of Caenorhabditis elegans. *Genes Dev* **17**, 991–1008 (2003).

[17] Blanchette, M. & Tompa, M. FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res* **31**, 3840–2 (2003).

[18] Prakash, A. & Tompa, M. Discovery of regulatory elements in vertebrates through comparative genomics. *Nat Biotechnol* **23**, 1249–56 (2005).

[19] Costas, J., Casares, F. & Vieira, J. Turnover of binding sites for transcription factors involved in early Drosophila development. *Gene* **310**, 215–20 (2003).

[20] Dermitzakis, E. T. & Clark, A. G. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* **19**, 1114–21 (2002).

[21] Ludwig, M. Z. *et al.* Functional evolution of a cis-regulatory module. *PLoS Biol* **3**, e93 (2005).

[22] Emberly, E., Rajewsky, N. & Siggia, E. D. Conservation of regulatory elements between two species of Drosophila. *BMC Bioinformatics* **4**, 57 (2003).

[23] Stark, A. *et al.* Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450**, 219–32 (2007).

[24] Sinha, S. & Tompa, M. YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* **31**, 3586–8 (2003).

[25] Pavesi, G., Mereghetti, P., Mauri, G. & Pesole, G. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res* **32**, W199–203 (2004).

[26] D'Haeseleer, P. How does DNA sequence motif discovery work? *Nat Biotechnol* **24**, 959–61 (2006).

[27] Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28–36 (1994).

[28] Lawrence, C. E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–14 (1993).

[29] Thijs, G. *et al.* A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17**, 1113–22 (2001).

[30] Siddharthan, R., Siggia, E. D. & van Nimwegen, E. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* **1**, e67 (2005).

[31] Pietrokovski, S. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res* **24**, 3836–45 (1996).

[32] van Nimwegen, E., Zavolan, M., Rajewsky, N. & Siggia, E. D. Probabilistic clustering of sequences: inferring new bacterial regulons by comparative genomics. *Proc Natl Acad Sci U S A* **99**, 7323–8 (2002).

[33] Schones, D. E., Sumazin, P. & Zhang, M. Q. Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics* **21**, 307–13 (2005).

[34] Roepcke, S., Grossmann, S., Rahmann, S. & Vingron, M. T-Reg Comparator: an analysis tool for the comparison of position weight matrices. *Nucleic Acids Res* **33**, W438–41 (2005).

[35] Sandelin, A., Hoglund, A., Lenhard, B. & Wasserman, W. W. Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes. *Funct Integr Genomics* **3**, 125–34 (2003).

[36] Kankainen, M. & Loytynoja, A. MATLIGN: a motif clustering, comparison and matching tool. *BMC Bioinformatics* **8**, 189 (2007).

[37] Gotoh, O. An improved algorithm for matching biological sequences. *J Mol Biol* **162**, 705–8 (1982).

[38] Jensen, S. T., Shen, L. & Liu, J. S. Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. *Bioinformatics* **21**, 3832–9 (2005).

[39] McGinnis, W., Hart, C. P., Gehring, W. J. & Ruddle, F. H. Molecular cloning and chromosome mapping of a mouse DNA sequence homologous to homeotic genes of Drosophila. *Cell* **38**, 675–80 (1984).

[40] Duboule, D. The rise and fall of Hox gene clusters. *Development* **134**, 2549–60 (2007).

[41] McGinnis, W. & Krumlauf, R. Homeobox genes and axial patterning. *Cell* **68**, 283–302 (1992).

[42] Maeda, R. K. & Karch, F. The ABC of the BX-C: the bithorax complex explained. *Development* **133**, 1413–22 (2006).

[43] Moehrle, A. & Paro, R. Spreading the silence: epigenetic transcriptional regulation during Drosophila development. *Dev. Genet.* **15**, 478–484 (1994).

[44] Cavalli, G. & Paro, R. Epigenetic inheritance of active chromatin after removal of the main transactivator. *Science* **286**, 955–8 (1999).

[45] Chan, C. S., Rastelli, L. & Pirrotta, V. A Polycomb response element in the Ubx gene that determines an epigenetically inherited state of repression. *EMBO J* **13**, 2553–64 (1994).

[46] Simon, J., Chiang, A., Bender, W., Shimell, M. J. & O'Connor, M. Elements of the Drosophila bithorax complex that mediate repression by Polycomb group products. *Dev Biol* **158**, 131–44 (1993).

[47] Lewis, E. B. A gene complex controlling segmentation in Drosophila. *Nature* **276**, 565–70 (1978).

[48] Kaufman, T. C., Seeger, M. A. & Olsen, G. Molecular and genetic organization of the antennapedia gene complex of Drosophila melanogaster. *Adv Genet* **27**, 309–62 (1990).

[49] Sanchez-Herrero, E., Vernos, I., Marco, R. & Morata, G. Genetic organization of Drosophila bithorax complex. *Nature* **313**, 108–13 (1985).

[50] Tiong, S., Bone, L. M. & Whittle, J. R. Recessive lethal mutations within the bithorax-complex in Drosophila. *Mol Gen Genet* **200**, 335–42 (1985).

[51] Jones, R. S. & Gelbart, W. M. Genetic analysis of the enhancer of zeste locus and its role in gene regulation in Drosophila melanogaster. *Genetics* **126**, 185–99 (1990).

[52] Glicksman, M. A. & Brower, D. L. Misregulation of homeotic gene expression in Drosophila larvae resulting from mutations at the extra sex combs locus. *Dev Biol* **126**, 219–27 (1988).

[53] Birve, A. *et al.* Su(z)12, a novel Drosophila Polycomb group gene that is conserved in vertebrates and plants. *Development* **128**, 3371–9 (2001).

[54] Tie, F., Furuyama, T., Prasad-Sinha, J., Jane, E. & Harte, P. J. The Drosophila Polycomb Group proteins ESC and E(Z) are present in a complex containing the histone-binding protein p55 and the histone deacetylase RPD3. *Development* **128**, 275–86 (2001).

[55] Bantignies, F. & Cavalli, G. Cellular memory and dynamic regulation of polycomb group proteins. *Curr Opin Cell Biol* **18**, 275–83 (2006).

[56] Muller, J. *et al.* Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. *Cell* **111**, 197–208 (2002).

[57] Savla, U., Benes, J., Zhang, J. & Jones, R. S. Recruitment of Drosophila Polycomb-group proteins by Polycomblike, a component of a novel protein complex in larvae. *Development* **135**, 813–7 (2008).

[58] Nekrasov, M. *et al.* Pcl-PRC2 is needed to generate high levels of H3-K27 trimethylation at Polycomb target genes. *EMBO J* **26**, 4078–88 (2007).

[59] Muller, J. & Kassis, J. A. Polycomb response elements and targeting of Polycomb group proteins in Drosophila. *Curr Opin Genet Dev* **16**, 476–84 (2006).

[60] Brown, J. L., Mucci, D., Whiteley, M., Dirksen, M.-L. & Kassis, J. A. The Drosophila Polycomb Group Gene pleiohomeotic Encodes a DNA Binding Protein with Homology to the Transcription Factor YY1. *Mol Cell* **1(7)**, 1057–1064 (1998).

[61] Czermin, B. *et al.* Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. *Cell* **111**, 185–96 (2002).

[62] Mohd-Sarip, A., Venturini, F., Chalkley, G. E. & Verrijzer, C. P. Pleiohomeotic can link polycomb to DNA and mediate transcriptional repression. *Mol Cell Biol* **22**, 7473–83 (2002).

[63] Brown, J. L., Fritsch, C., Mueller, J. & Kassis, J. A. The Drosophila pho-like gene encodes a YY1-related DNA binding protein that is redundant with pleiohomeotic in homeotic gene silencing. *Development* **130**, 285–94 (2003).

[64] Strutt, H., Cavalli, G. & Paro, R. Co-localization of Polycomb protein and GAGA factor on regulatory elements responsible for the maintenance of homeotic gene expression. *EMBO J* **16(12)**, 3621–3632 (1997).

[65] Hur, M.-W., Laney, J. D., Jeon, S.-H., Ali, J. & Biggin, M. D. Zeste maintains repression of Ubx transgenes: support for a new model of Polycomb repression. *Development* **129**, 1339–1343 (2002).

[66] Dejardin, J. *et al.* Recruitment of Drosophila Polycomb group proteins to chromatin by DSP1. *Nature* **434**, 533–8 (2005).

[67] Brown, J. L. *et al.* An Sp1/KLF binding site is important for the activity of a Polycomb group response element from the Drosophila engrailed gene. *Nucleic Acids Res* **33(16)**, 5181–5189 (2005).

[68] Blastyak, M. R. K. K. F., A. & Gyurkovics, H. Efficient and specific targeting of Polycomb group proteins requires cooperative interaction between Grainyhead and Pleihomeotic . *Mol. Cell. Biol.* **26**, 1434–1444 (2006).

[69] Papp, B. & Muller, J. Histone trimethylation and the maintenance of transcriptional ON and OFF states by trxG and PcG proteins. *Genes Dev* **20**, 2041–54 (2006).

[70] Negre, N. *et al.* Chromosomal distribution of PcG proteins during Drosophila development. *PLoS Biol* **4**, e170 (2006).

[71] Schwartz, Y. B. *et al.* Genome-wide analysis of Polycomb targets in Drosophila melanogaster. *Nat Genet* **38**, 700–5 (2006).

[72] Tolhuis, B. *et al.* Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster. *Nat Genet* **38**, 694–9 (2006).

[73] Ringrose, L. Polycomb comes of age: genome-wide profiling of target sites. *Curr Opin Cell Biol* **19**, 290–7 (2007).

[74] Cleard, F., Moshkin, Y., Karch, F. & Maeda, R. K. Probing long-distance regulatory interactions in the Drosophila melanogaster bithorax complex using Dam identification. *Nat Genet* **38**, 931–5 (2006).

[75] Kassis, J. A. Unusual properties of regulatory DNA from the Drosophila engrailed gene: three "pairing-sensitive" sites within a 1.6-kb region. *Genetics* **136**, 1025–38 (1994).

[76] Fauvarque, M. O. & Dura, J. M. polyhomeotic regulatory sequences induce developmental regulator-dependent variegation and targeted P-element insertions in Drosophila. *Genes Dev* **7**, 1508–20 (1993).

[77] de la Cruz, C. C. *et al.* The polycomb group protein SUZ12 regulates histone H3 lysine 9 methylation and HP1 alpha distribution. *Chromosome Res* **15**, 299–314 (2007).

[78] Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J Mol Biol* **196**, 261–82 (1987).

[79] Takai, D. & Jones, P. A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* **99**, 3740–5 (2002).

[80] Fatemi, M. *et al.* Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. *Nucleic Acids Res* **33**, e176 (2005).

[81] Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103**, 1412–7 (2006).

[82] Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A* **90**, 11995–9 (1993).

[83] Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* **39**, 457–66 (2007).

[84] Vire, E. *et al.* The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* **439**, 871–4 (2006).

[85] Eden, E., Lipson, D., Yogev, S. & Yakhini, Z. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol* **3**, e39 (2007).

[86] Berman, B. P. *et al.* Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biol* **5(9)**, R61 (2004).

[87] Stark, K., Kirk, D. L. & Schmitt, R. Two enhancers and one silencer located in the introns of regA control somatic cell differentiation in Volvox carteri. *Genes Dev* **15**, 1449–60 (2001).

[88] Knippers, R. *Molekulare Genetik*, vol. 8 (Georg Thieme Verlag, 2001).

[89] Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–402 (1997).

[90] MacArthur, S. & Brookfield, J. F. Expected rates and modes of evolution of enhancer sequences. *Mol Biol Evol* **21**, 1064–73 (2004).

[91] Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708–15 (2004).

[92] Elemento, O., Slonim, N. & Tavazoie, S. A universal framework for regulatory element discovery across all genomes and data types. *Mol Cell* **28**, 337–50 (2007).

[93] Consortium, Drosophila 12 Genomes. Assembly/Alignment/Annotation of 12 related Drosophila species (2008). URL `http://rana.lbl.gov/drosophila/`.

[94] Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287**, 2185–95 (2000).

[95] Clark, A. G. *et al.* Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**, 203–18 (2007).

[96] Richards, S. *et al.* Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res* **15**, 1–18 (2005).

[97] Hauenschild, A., Ringrose, L., Altmutter, C., Paro, R. & Rehmsmeier, M. Evolutionary plasticity of Polycomb/Trithorax Response Elements in Drosophila species (2008). In Revision.

[98] Horard, B., Tatout, C., Poux, S. & Pirrotta, V. Structure of a polycomb response element and in vitro binding of polycomb group complexes containing GAGA factor. *Mol Cell Biol* **20**, 3187–97 (2000).

[99] Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13**, 721–31 (2003).

[100] Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32**, W273–9 (2004).

[101] Ringrose, L. & Paro, R. Polycomb/Trithorax response elements and epigenetic memory of cell identity. *Development* **134**, 223–32 (2007).

[102] Atchison, L., Ghias, A., Wilkinson, F., Bonini, N. & Atchison, M. L. Transcription factor YY1 functions as a PcG protein in vivo. *EMBO J* **22**, 1347–58 (2003).

[103] Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276–7 (2000).

[104] Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 374–8 (2003).

[105] Bembom, O. seqLogo: An R package for plotting DNA sequence logos. *http://works.bepress.com/bembom/11/* (2007).

[106] Frith, M. C., Hansen, U., Spouge, J. L. & Weng, Z. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res* **32**, 189–200 (2004).

[107] Favorov, A. V. *et al.* A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics* **21**, 2240–5 (2005).

[108] Ao, W., Gaudet, J., Kent, W. J., Muttumu, S. & Mango, S. E. Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science* **305**, 1743–6 (2004).

[109] Smit, H. R. . G. P., AFA. RepeatMasker Open-3.0. *http://www.repeatmasker.org* (1996-2004).

[110] Karolchik, D. *et al.* The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36**, D773–9 (2008).

[111] Mirkin, S. M. *et al.* DNA H form requires a homopurine-homopyrimidine mirror repeat. *Nature* **330**, 495–7 (1987).

A Appendix

A.1 *DynScan* package elements

copy_score_in_db.pl

Does a copy of a given score file in the database table instead of using insert to speed up the process.

make_hit_bands.pl

Builds a list of non-overlapping regions in a given genome which score above the cut-off and stores them in table *static_pres*

extract_pre_sequences.pl

Creates a single FASTA file for each distinct PRE

1_add_species.pl –q SPECIES –c FILE

Main part of the *DynScan* package. Runs a dynamic search between query species provided by "–q" option and all other species in database. Query species has to be entered in database first. Additonally required parameters are taken from configuration file provided by option "-c"

2_find_orthologous_hits.pl –q SPECIES –t SPECIES –c FILE

Evaluates the best BLAST hits according to the described criteria and stores the result in the database. Task is performed for each element in the query species ("-q") against the target species ("-t"). Required parameters are taken from configuration file provided by parameter "–c".

3_run_prediction_complete.pl –q SPECIES –t SPECIES –c FILE

Runs the dynamic search around the homologous sites in the database for the provided query ("-q") and target species ("-t"). Required parameters are taken from configuration file provided by parameter "–c".

A.2 Motif evaluation

evaluate_motif_list.pl -m MODEL -b BACKGROUND -o MOTIFLIST -r RUN

Builds all possible double motifs out of the given list of single motifs ("-o"). All motif pairs are weighted acccording to the provided positive ("-m") and negative ("-b") training sets. Those motif pairs that help to distinguish the sets best are kept in a jPREdictor option file, the sequences in the model are either copied in a FASTA file of found or missed sequences. The run identifier "-r" determines the ID used in the names of the created files.

A.3 Phylogenetic footprinting pipeline

footprinter_pipeline.pl -r RUN

Runs Footprinter on a remote Linux system. Option "-r" determines iteration. Sequences are taken from default pipeline Footprinter directory. Hits are masked, resulting sequences are stored in iteration+1 directory. Hits are transfered into clustering input format and clustered by MATLIGN. Clustering results of all single sequences are clustered again.

meme_pipeline.pl -r RUN

Runs MEME on all alignemnts. Option "-r" determines iteration. Input sequences are built as blocks out of Multiz17Way alignments first, MEME is run, results are transfered into input for clustering tools. Hits are masked and new sequences in directory iteration+1 are created.

weeder_pipeline -r RUN

Runs Weeder on a set of alignments. All sequences are searched at once, hits are masked. Hits are transformed into input for clustering tools.

count_tree.pl -t treefile -c clusterfile [-p] [-l] [-j] [-T]

Evaluates MATLIGN output and outputs all clusters sorted by the number of elements. Files containing the tree and cluster have to be provided. Option -p (print) gives a complete output of all clusters, -l (logo) creates motif logo in PDF format of largest cluster. Output can be got in TransFac style "[-T]" or in jPREdictor format "-j". The variable "$max" determines how many of the best clusters are reported by default if is "-p" is omitted. If "$texoutput" is set to 1, logos will be combined in a LaTex file.

A.4 Supplementary data

10-mer	# in homeotic pos set	# number in homeotic neg set	# number of diff seq in pos set
GGCTGCAGCG	9	1	8
AGAGAGCGAG	10	1	8
CCTCTTCCTC	14	2	11
CTCCCTCTTC	10	1	8
CTTTTTAAAA	10	1	8
AGGAGGAGGA	42	6	15
CCCGGCCGCC	13	2	13
CCTGGGCTGC	9	1	9
CGGGCCCGGC	9	1	9
CTTCTCTCCC	9	1	9
GCCCGCCGGC	9	1	9
GCCGGGGCGC	9	1	9
TCTGGAACCA	9	1	9
AGGAGGAGAG	10	1	9
AGGGGGAGAA	10	1	9
GCGCCGCTCC	10	1	9
AGGCCGGGGC	10	1	10
CCTTCTCTCC	10	1	10
CTCCCCTCCA	10	1	10
CTCCTCTCCC	10	1	10
GGTTCCAGAA	10	1	10
GAGAAAGGGA	12	1	9
GCCCGCGCGC	14	1	10
TTTTTAAAAA	26	2	12
GAGGGGGAGA	17	1	12
GAGAGAAAGA	14	1	14

Table A.1: Highest scoring (entropy∗log odds score) 10-mers in human homeotic training set. Column two gives absolute numbers in all 50 positive sequences, column three give numbers in negative set. Column four shows number of distinct positive sequences that contain at least one hit.

8-mer	# in homeotic pos set	# number in homeotic neg set	# number of diff seq in pos set
CGTCCGTC	13	1	9
CAGCTCAA	15	2	14
AAACCTCT	10	1	10
CAAGTGGA	10	1	10
CACGGGAC	10	1	10
CTGGAATC	10	1	10
GTTTAATA	10	1	10
AATTAAGA	12	1	9
AAGAAGTC	11	1	10
AGACATCC	11	1	10
AGTCTAGA	11	1	10
ATCTTAAT	11	1	10
AGACTGAA	11	1	11
CAGTGCAC	11	1	11
AATGAATT	13	1	10
CAGTGGAG	21	3	20
AACTTTCA	12	1	11
ACGGGAAG	12	1	11
ATGAACAA	12	1	12
CTAGAAGC	12	1	12
CCCCTAGA	13	1	12
CCGCGGAA	13	1	12
AGTGAGGC	14	1	12
AGCCAAAG	13	1	13
GGACAGCC	13	1	13

Table A.2: Highest scoring (entropy$*$log odds score) 8-mers in human homeotic training set. Column two gives absolute numbers in all 50 positive sequences, column three give numbers in negative set. Column four shows number of distinct positive sequences that contain at least one hit.