Genome-wide prediction of developmental enhancers and analysis of evolutionary plasticity in *Drosophila*

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

vorgelegt von
Jia Ding

September 2009

Acknowledgements

First I want to express a clear and heartfelt thank you to my supervisor, Dr. Marc Rehmsmeier, for his constant input, ideas and suggestions during my time as a graduate student in Bielefeld. His initiative made this work possible in the first place and he was always helpful in offering guidance when requested. Prof. Robert Giegerich helped with proofreading and much valued suggestions and corrections. My office-mate Dr. Thomas Fiedler provided me with jPREdictor and was always present for helpful discussion. Dr. Arne Hauenschild was always ready to serve as a critical sounding board during our discussions. Dr. Leonie Ringrose provided useful biological input by offering experimental validation and a thorough biological discussion of my *in-silico* results. I reserve a big thank you for my parents: without their support the long road from bachelor degree in Shanghai to this doctoral thesis in Germany would not have been possible. Finally I would like to thank Francis Dierick for his patience with me over the years and his moral support.

Contents

List of Figures

List of Tables

Abstract

The early *Drosophila* embryo is a model for the study of transcriptional control of development. A pre-requisite for the development of the *Drosophila* embryo is a precise and coordinated control of gene expression, both spatially and temporally. This process of complex transcription regulation is thought to be achieved by the combinatorial action of multiple transcription factors binding to modular units of *cis*-regulatory DNA sequences. The transcription factors Bicoid (BCD), Caudal (CAD), Hunchback (HB), Kruppel (KR), and Knirps (KNI) are crucial in patterning the anterior-posterior axis of the embryo by acting at very early stages of *Drosophila* development.

In prior studies, functional tests of 37 predicted targets of the five above-mentioned motifs have been completed. A positive training set of 15 sequences and a negative training set of 18 sequences have been constructed for embryonic enhancer prediction.

Clustering of transcription factor binding sites is the traditional approach for *cis*-regulatory element prediction, but several drawbacks exist.

In contrast to popular clustering approaches, my proposed method utilizes paired motifs to identify enhancers from non-functional elements. Application of this paired motifs approach achieves a genome-wide prediction with high specificity (94%) and sensitivity (60%). Paired motif prediction performs better than single motif prediction when considering motif weight, separation of positive/negative training sets and the total number of predicted enhancers.

The availability of Gene Ontology information improves genome-wide prediction by enabling the use of a subset prediction method, restricting the search to regions flanking embryonic genes; more candidates are included while still maintaining a good specificity.

The genomes of multiple *Drosophila* species provide an excellent model for comparative analysis. I present a dynamic search approach, which is unbiased in terms of sequence conservation, and has the potential to find non-conserved enhancers.

In total I predicted 135 enhancers in *D. melanogaster* including 37 novel and 27 known enhancers. 71 enhancers of my prediction overlap with experimentally verified binding regions but not with characterized known enhancers; they are likely to be functional elements and good candidates for experimental validation.

Additionally, I confirmed that enhancer elements are indeed subject to fast evolutionary changes. First, the number of enhancers varies widely across *Drosophila* species. Second, the positions of embryonic enhancers are independent of sequence conservation. Third, motif re-arrangement in homologous enhancers is rather frequent and rapid. Fourth, enhancer gain and loss analysis shows that 9 enhancers have been gained in *D. melanogaster* during evolution. From this observation I speculate that embryonic enhancers can originate from non-functional sequence.

This prediction method has been proven to work on embryonic enhancers as well as PRE/TREs, hence exploring its effectiveness outside the embryonic domain or even in different species is worthy of further research. The prediction results are available online http://bibiserv.techfak.uni-bielefeld.de/jpred_en.

1. Introduction

1.1. Hierarchical cascade of regulatory gene expression in the *Drosophila* embryo

A major challenge in interpreting genome sequences is understanding the mechanisms that govern gene expression. It is well known that complex temporal and spatial patterns are involved in the transcription of protein-coding genes; these patterns can significantly influence the differentiation and development of eukaryotic embryos [1]. Thus, any attempt at fundamentally understanding embryonic development is deeply intertwined with knowledge of decoding the transcriptional control of patterned gene expression [2].

Molecular and genetic studies of *D. melanogaster* have led to profound advances in understanding the regulation of development [3]. Each cell in the embryo receives its ultimate function after an orderly cascade of gene expression patterns during embryonic development [4].

These cascades can be observed in one of the most important phenomena in development: the formation of embryonic axes such as the Anterior-Posterior (AP) axis. This axis extends from head to tail and is one of the foundations of body segmentation [5]. A hierarchical cascade of regulatory gene expression driven by the establishment of a diffusion-mediated morphogen gradient which has been observed to cause segmentation in *Drosophila* embryogenesis. This particular cascade results in the formation of head, thorax and abdomen in insect species [6] (see Figure 1.1).

Gene cascades include "maternal effect genes", "gap genes", "pair-rule genes", "segment polarity genes" and "Hox genes". The functions of these genes will be explained in this chapter.

1.1.1. Maternal effect genes

A broad determination of anterior and posterior regions, defined by long-range gradients of maternally expressed genes is observed in *Drosophila* segmentation [8].

These maternal effect genes are expressed in the mother's ovaries and produce messenger RNAs which are then positioned in various regions of the egg. The production of anterior structures is regulated by the maternal proteins Bicoid (BCD) and Hunchback (HB); the posterior parts of the embryo are regulated by the maternally specified Nanos and Caudal (CAD) [5]. These maternal factors consistently act as activators [9].

1.1.2. Gap genes

After initiation, further maternal gradients define the position of the gap gene domains, which specify blocks of adjacent segments [9]. Gap genes, particularly *Kruppel (kr)* , *Knirps (kni)*

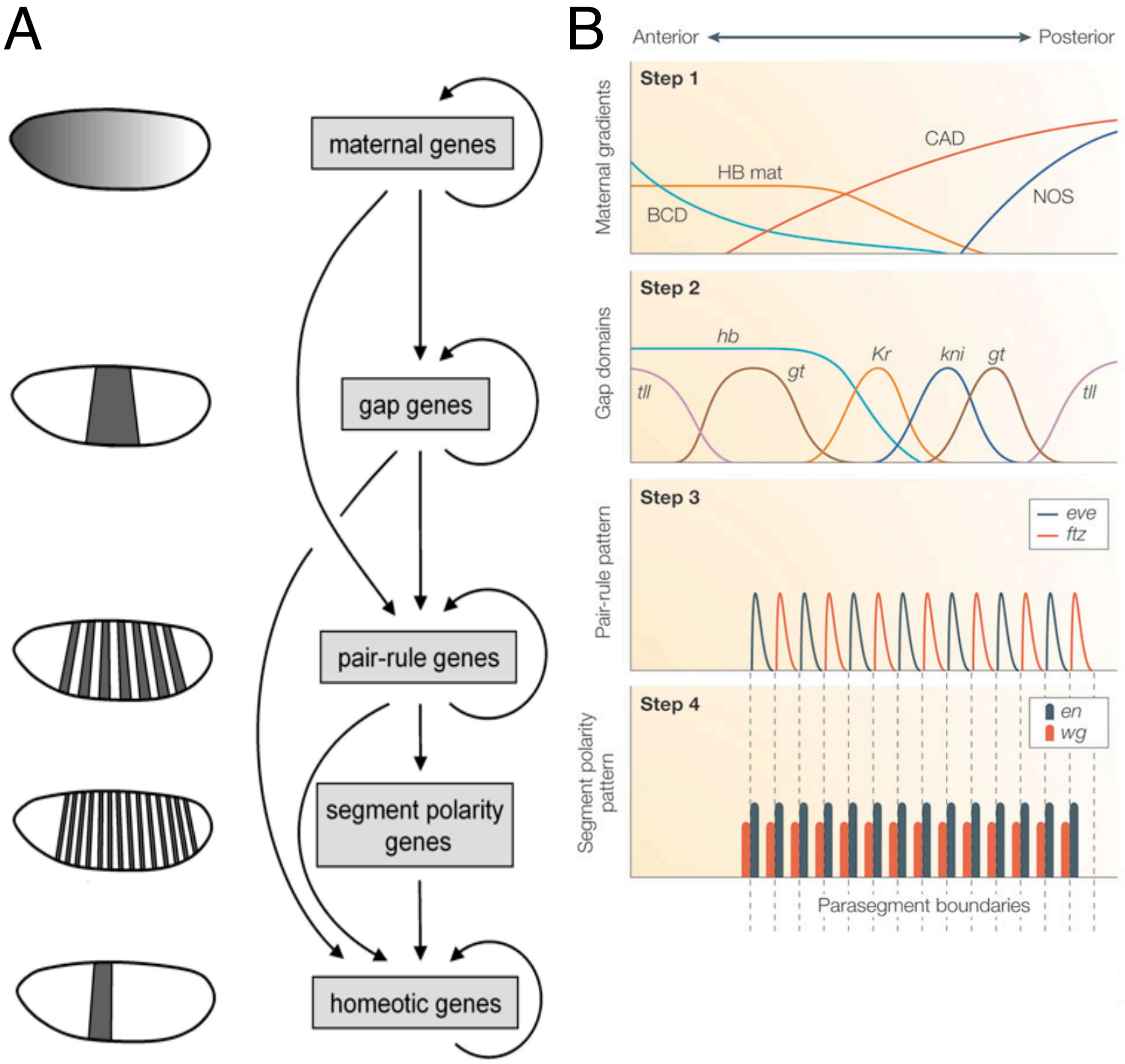


Figure 1.1.: Hierarchical cascade of regulatory gene expression in *Drosophila* embryo. Schematic depiction of the regulatory relationships within the segmentation gene network (figure A from [2]). Segmentation genes encode a cascade of interacting transcription factors that generate progressively finer patterns of gene expression in the blastoderm-stage embryo (figure B from [7]). In step 1, translation of localized maternal transcripts of gene *cad*, *hb* and *bcd*, and the diffusion of their protein products, generates protein gradients along the egg. The resulting egg contains the HB protein in the anterior half (maternal HB, HB mat), and long-range gradients of both BCD (high at the anterior) and CAD (high at the posterior). In step 2, the signals from these maternal proteins activate a small set of zygotic gap genes at specific positions along the AP axis of the egg. In step 3, transcriptional interactions between primary pair-rule genes (for example, *even-skipped*) and the other genes that they regulate (for example, *fushi tarazu*) refine the domains of expression into periodic stripes. In step 4, the activities of the pair-rule proteins result in the activation of the segment polarity genes. The boundary between engrailed expressing cells and their anterior neighbors wingless becomes the parasegment boundary.

and *Giant (gt)*, are expressed in specific broad domains [5, 10].

The gap nomenclature to designate these genes was introduced after observing the effect of mutation: gap genes literally cause gaps to appear in the phenotype of the developing structure by causing the loss of central elements in the embryo [11].

All gap genes encode transcription factors and cross-regulate each other to refine borders of expression. All gap factors exercise regulation through repression, except for the bimodal factor Hunchback which is also an activator [2].

In classical genetic and molecular studies it has been observed that the maternal *bcd*, *hb*, and *cad* gradients are essential for the establishment of gap gene expression patterns [12, 13]. However, a large degree of uncertainty still exists when trying to pinpoint the precise combinations of maternal morphogens involved in the control of the gap system [4].

1.1.3. Pair-rule genes

The pair-rule genes further split the embryo into periodic units which can be observed in the blastoderm as seven stripes of gene expression. Auto- and cross-regulation of all pair rule genes refine the borders of the pair-rule stripes of expression. Deletion of portions of alternate segments is usually caused by mutations of pair-rule genes, such as *fushi tarazu* (*ftz*)[5].

1.1.4. Segment polarity genes

Segment polarity genes divide the early embryo into a repeating series of segmental primordia along the AP axis and are expressed as 14 stripes. Mutations in segmentation genes cause the embryo to lack certain segments or parts of segments [5].

Unlike gap and pair-rule genes, not all segment polarity genes encode transcription factors and some are involved in the encoding of cell signaling proteins [14]. It has been long observed that anterior regions typically have faster generation of pair-rule and segment polarity stripes than posterior regions [15].

1.1.5. Hox genes

During the later stages of the cascade in nearly all insects and vertebrates, Hox genes are involved in identifying individual segments and determining which body parts they will become. Physical deformities were observed after mutations in specific Hox genes, causing e.g. legs to be grown out of the fly's head or wings be to grown in places where they shouldn't be [6]. The complement of Hox genes is made up of two clusters, the *Antennapedia* complex, which controls the anterior third, and the *Bithorax* complex which takes care of the posterior two thirds of the fly[16, 17]. The *Drosophila* genes that belong to the Hox genes family are *Antennapedia (Antp), Deformed (DFD), labial (lab), Sex combs reduced (Scr), Abdominal-A/B (Abd-A/B)* and *Ultrabithorax (Ubx)* [18, 19] (see Figure 1.2).

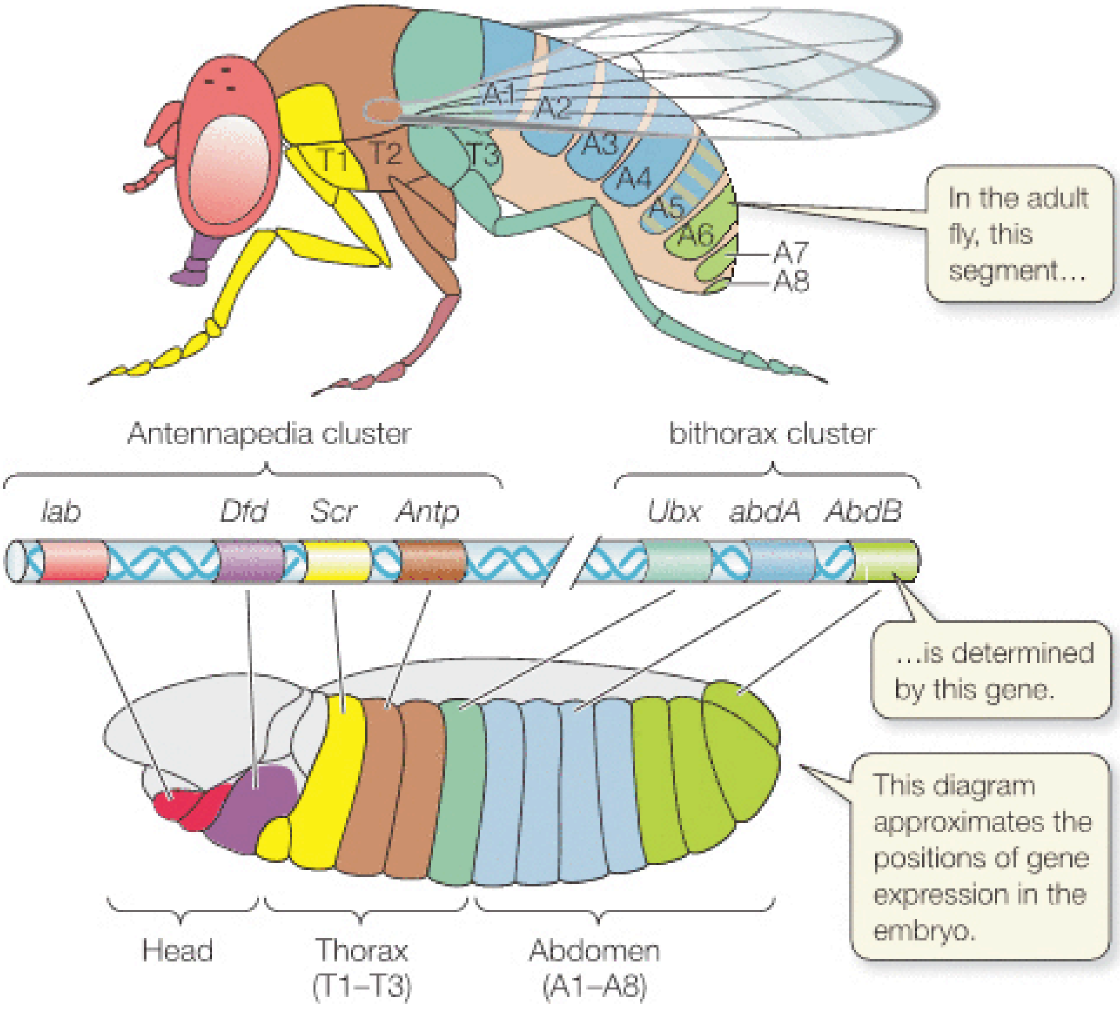

A1
A2
A3
A4
A5
A6
A7
A8
T1
T2
T3
In the adult fly, this segment…
Antennapedia cluster
bithorax cluster
lab
Dfd
Scr
Antp
Ubx
abdA
AbdB
…is determined by this gene.
This diagram approximates the positions of gene expression in the embryo.
Head
Thorax (T1–T3)
Abdomen (A1–A8)

Figure 1.2.: Hox Genes in *Drosophila*. Two clusters of genes on chromosome 3 (center) determine segment function in the adult fly (top). These genes are expressed in the embryo (bottom) long before the structures of the segments actually appear [6].

1.2. Combinatorial interaction of bound transcription factors

The nature of regulation within the segmentation gene network is almost completely transcriptional. A large number of the segmentation genes are utilized several times in different phases of development. Several segmentation genes are transcription factors themselves; their principal targets are other segmentation genes acting at the same level or downstream [2, 20]. Multiple bound transcription factors act combinatorially to confer specific transcriptional activity along the AP axis in the early *Drosophila* embryo. It is helpful to understand how they are able to perform co-regulation by looking at their respective functions.

Maternal and gap-gene transcripts generate the transcription factors BCD, CAD, HB, KR, and KNI. Acting at the very early stages of *Drosophila* development, they originate the AP axis of the embryo [21, 22] and ensure normal development (see Table 1.1).

1.2.1. Bicoid

The *Bicoid (bcd)* gene, a well-known maternal patterning gene, was discovered in a large-scale screen for female-sterile mutants in *Drosophila* [25]. Development of head and thorax in the larva is heavily influenced by the function of this gene. However, its activity declines markedly with increased distance from the anterior part of the embryo in cleavage stage [26].

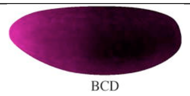
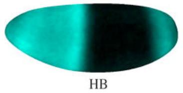
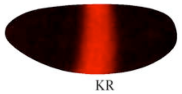
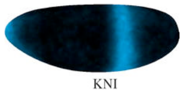
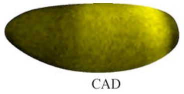
Two crucial characteristics of *bcd* can be observed: (1) *bcd* encodes a transcription factor including a homeodomain, and (2) its maternal mRNA forms a gradient along the AP axis of the embryo at stages preceding cellular blastoderm [26]. However, BCD concentration required for target gene activation is believed to be in excess at every position along the AP axis [27]. Positional information of this gradient is transferred to specific downstream gap genes occupying a well defined spatial domain [10].

Development is a precise process. The transmission of errors and their amplification to downstream genes must be avoid. The establishment of a morphogen gradient is a good example of an error-prone process [10]. Embryos lacking maternally expressed *bcd* fail to develop anterior segments, including the head and thorax (see Figure 1.3).

1.2.2. Hunchback

The function of *Hunchback (hb)* is crucial in the establishment of an AP gradient of gene activity during the transition from unfertilized egg to developing zygote. The HB protein establishes a distinct boundary at the middle of the embryo with great positional precision [29]. BCD shares responsibility for HB activation with self-regulation by HB itself. HB is built up from maternal and zygotic contributions, and provides positional information for other gap genes, such as *Kruppel (Kr), Knirps (kni), and Giant (gt)*, and for the homeotic gene *Ultrabithorax (Ubx)*. Removing both maternal and zygotic *hb* expression results in severe deletions and polarity reversals of the most anterior segments [30]. Hunchback acts both as an activator for the anterior gap gene function and as a co-activator with BCD. It extends the effective range of *bcd* by shifting the effective morphogenetic activity of *bcd* towards the posterior segment. Both activation and repression of transcription can be regulated by *hb*; the

Table 1.1.: List of five transcription factors BCD, HB, CAD, KR and KNI. TG* transcription factor target genes from [1, 23]. mRNA Expression Pattern* images from [24].

Name	**FBgn**	**Keywords**	**Binding domain**	**TF**	**TG***	**mRNA Expression Pattern***
Bicoid	FBgn0000166	Anterior group	Homeodomain	BCD	*tll, eve, ems, Kr, kni, salm, h, hb, spalt*	
Hunchback	FBgn0001180	gap gene	Zinc finger protein	HB	*abd-A, Ubx, eve, Kr, kni, salm, h, en, hb*	
Kruppel	FBgn0001325	gap gene	Zinc finger	KR	*abd-A, Ubx, ko, eve, kni, salm, h, en, hb*	
Knirps	FBgn0001320	gap gene	Steroid receptor	KNI	*Ubx, eve, Kr, h*	
Caudal	FBgn0000251	gap gene	Homeodomain	CAD	*ftz, kni, salm, en, h*	

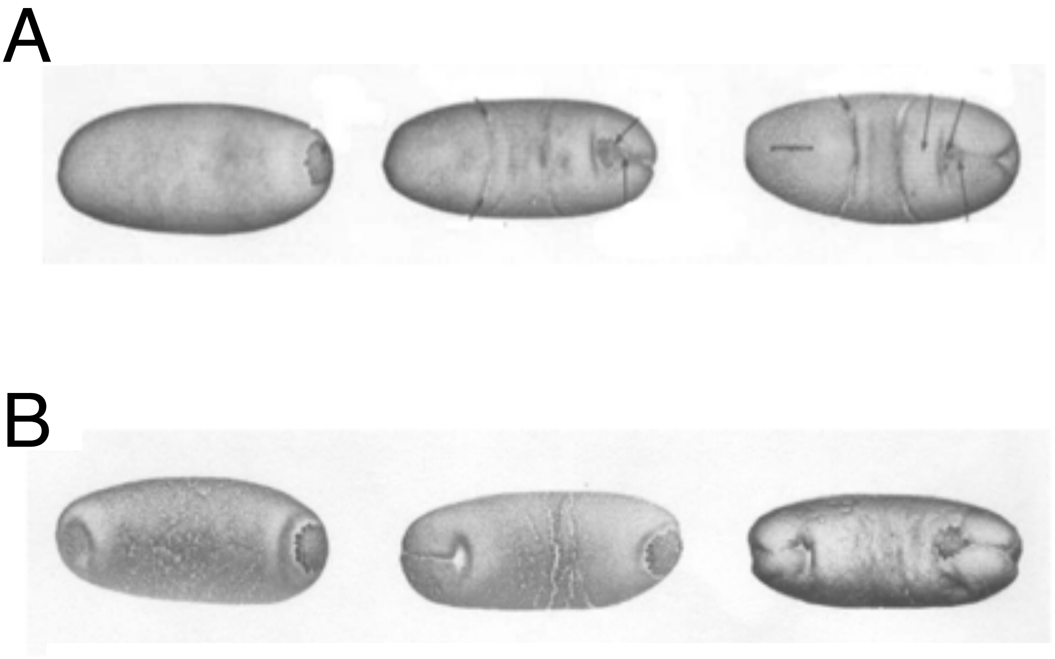


Figure 1.3.: Comparing normal developing embryo with mutant embryo. Normal development of a *D. melanogaster* embryo (A) compared with development in an embryo which has a mutation in just one gene-*Bicoid* (B). Expression of the Bicoid gene in normal development tells the developing embryo where its head should be. In the normal embryo, the left hand side will eventually be the head and the right hand side will be the tail of the larva. In the Bicoid mutant embryo, the head end (left hand side) of the embryo has formed a groove similar to that on the tail of the embryo. The embryo does not form a head, rather it forms two tail ends. Images, courtesy of SIBE [28].

placement of both anterior and posterior gap genes is determined by it. At low concentrations, HB activates gene expression, whereas at high concentrations it mediates repression [4].

1.2.3. Kruppel

Kruppel (kr) one of the gap genes is homologous to *hb*; they share four zinc finger domains. The domain of early *kr* expression is in the center of the embryo. Expression of *kr* forms a single broad band, like a belt, around the middle of the embryo. The *kr* gene is expressed primarily in parasegments 4–6, central to the *Drosophila* embryo. An absence of the KR protein causes the embryo to lack these regions. The KR transcript appears over the region where HB is on the decline [5].

1.2.4. Knirps

The gap gene *Knirps (kni)* is expressed in the two polar domains along the AP axis in the blastoderm; three modules driving the AP expression have been identified [31, 2]. If KNI activity is lacking, *kr* gene expression spills over into the posterior domain [5]. KNI is known as a repressor working at short range. These types of repressors are thought to provoke enhancer silencing by causing changes in chromatin structure at the local level rather than competitively blocking the binding of inducers [32].

1.2.5. Caudal

During early embryogenesis in *Drosophila*, *Caudal (cad)* mRNA is distributed as a gradient achieving its peak level at the posterior end of the embryo. The abdominal gap gene *kni* is activated by this gradients. The *cad* gene is expressed both maternally and zygotically. Its zygotic expression in the blastoderm consists of a single posterior stripe. This suggests that the CAD homeodomain transcription factor might play a role in establishing the posterior domains of the embryo, which undergo gastrulation and give rise to the posterior gut.

1.2.6. Combinatorial interaction of transcription factors in the regulation of *even-skipped*

The combinatorial interaction of the above transcription factors determines time- and tissue-specific gene activation or repression [33]. These early maternal and gap genes (*bicoid, hunchback, kruppel, knirps, caudal*) encode transcription factors that co-regulate in order to segment the embryonic trunk along the AP axis.

For example, the pair-rule gene *even-skipped* (*eve*) causes seven transversal stripes along the AP axis of the blastoderm embryo to be produced. The second stripe in *even-skipped* (*eve*) is for the most part regulated by multiple binding sites for five TFs. BCD and HB coordinately act as activators, whereas KR and GT repress expression in the embryo, restricting *eve* expression output to a narrow stripe lying between the expression of KR and GT [34, 35, 36, 24]. Spatiotemporal control of *eve* stripe 2 expression is controlled through the interaction of these regulation signals. In *kr* mutant embryo, *eve* expression pattern of stripes 2, 3 and 4/6 are fused into two broad bands; some of the *eve* stripes will disappear,

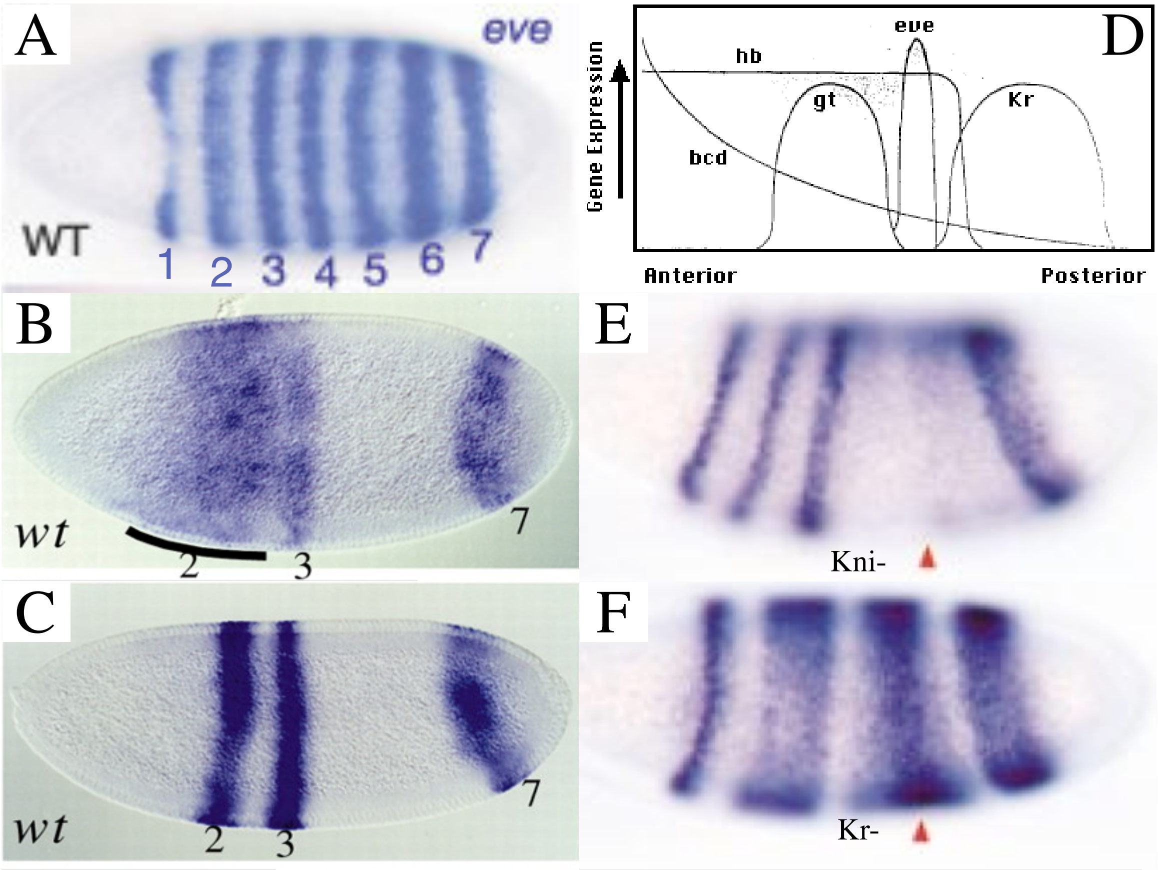


Figure 1.4.: Expression pattern of *even-skipped* gene. (A) wild-type *eve* expression pattern with 7 stripes.

The wild-type *eve* 2 enhancer is first activated in a broad anterior domain (marked by the broad line, B), which is then refined to a stripe (C) [36].

(D) Binding of bicoid and hunchback proteins stimulates transcription of *eve*. Repressors KR and GT delimit expression borders of the *even-skipped* stripe 2 [36].

(E) *eve* expression pattern in Knirps– embryo. There is a loss of stripes 4, 5 and 6 [37].

(F) *eve* expression pattern in Kruppel– embryo. Stripes 2, 3 and 4/6 are fused into two broad bands [37].

if *kni* is not expressed (see Figure 1.4). Thus, phenotypic consistency for *eve* expression or even general embryo development is strictly controlled by genes switching on and off at the correct time and in the correct place. The typically complex expression patterns found in segmentation genes have made them one of the preferred areas of study for understanding transcription control *in-vivo*.

1.3. *cis*-regulatory DNA elements

The core functioning of gene regulatory networks consists of genes encoding transcription factors and the *cis*-regulatory elements that control the expression of said genes [38]. The expression patterns of these segmentation genes typically show high complexity, and in many cases different aspects of the pattern are controlled by separate *cis*-regulatory elements. One such example can be seen in the expression of the seven *even-skipped* early transverse stripes, which is regulated by five distinct *cis*-regulatory elements [2]. Individual *cis*-regulatory sequences influence target genes by the combinatorial action of multiple transcription factors [39, 40, 2, 41, 1].

Cis-regulatory DNA elements include enhancers, silencers, isolators, insulators and Polycomb/Trithorax response elements (PRE/TREs). They are typically 500–1000 bp long and often contain binding sites for transcription factors (TFs). However, the number of binding sites and the spacing between them may vary, and binding-site sequences are often so degenerated that they can only be identified by probabilistic methods [42].

Enhancers, one kind of CREs, are traditionally defined by their ability to recapitulate an aspect of the endogenous gene activity when linked to a reporter gene in a position and orientation independent manner [43]. The position of these enhancers can vary: they may be located upstream, downstream, or within the gene they control [44, 45].

Some CREs may function as silencers regulating both active and passive repression mechanisms [46]. Both enhancers and silencers may operate over large distances to regulate the genes in these domains [47].

Insulators constitute another class of *cis*-regulatory elements. They create boundaries in chromatin, delineating the ranges over which other regulatory influences take effect. Two types of insulators have been observed: enhancer-blocking insulators and barrier insulators. The former are DNA elements that disrupt communication between discrete regulatory sequence elements (typically enhancers, silencers and promoters) when positioned between them. The latter prevent the spread of heterochromatin [48, 49, 50].

Polycomb and trithorax group proteins (PcG and TrxG) are conserved chromatin components that maintain the transcriptional state of the developmental control genes such as Hox genes in *Drosophila*. After initial setup in the early phases of embryogenesis, Hox genes are initialized by a cascade of maternal and zygotic transcription factors which are the products of the interplay between pair-rule, gap and segmentation genes [51, 52]. At mid-embryogenesis, most of these initial transcription factors have disappeared, but the transcriptional state of homeotic genes is preserved throughout life due to the action of PcG and TrxG proteins. PcG proteins are responsible for maintaining the repressed state of homeotic genes while TrxG factors aid in the maintenance of their active state [52]. The regulatory DNA elements to which these PcG and TrxG factors bind are called PcG and TrxG response elements (PREs and TREs). Several DNA binding proteins such as PHO, Dsp1, GAF and

Zeste have been suggested to recruit PcG proteins to PREs [53].

1.4. Identification of *cis*-regulatory elements by computational methods

The number of regulatory elements in the genome is considered to be very high; Davidson [54] suggests that the number of CREs might be 5 to 10 times the number of genes in the genome. Much less is known about CREs in general than about coding sequences, although they are quite important and prevalent. This is largely because of the difficulties involved in detecting CREs by bioinformatics tools [55].

The compilation of multiple genome sequences and the rising amount of large-scale gene expression data have contributed to the maturation of bioinformatics methods for the analysis of sequences that regulate gene expression [56].

In the last 25 years, many computational methods for both modeling and identification of DNA regulatory elements have been developed (see[57, 56, 58] for an overview). The common workflow for the prediction of *cis*-regulatory elements is shown in Figure 1.5. Broadly speaking, most of these methods fall into either or both of two classes: transcription factor binding sites (TFBSs) clustering or inter-species conservation. In the former class, CREs are defined as those regions that contain a defined number and/or combination of specific TFBSs. In the latter class, CREs are predicted based on sequence conservation between multiple related species [55].

1.4.1. Prediction of CREs by clustering

In a typical scenario, transcription factors with previously defined binding specificities are known, and one wants to perform genome-wide discovery of modules (and genes) targeted by these factors [39].

A functional CREs typically contains binding sites for multiple transcription factors with most of the sites represented multiple times [55], because a region with multiple putative TFBSs is more likely to be functional element than a region with only a single site [59].

Clustering of the relevant binding sites within a small interval can be made with the assumption that TFs co-operate as a functional complex in regulating gene expression [2]. Discovering these statistically significant clusters of predicted occurrences of input transcription factor motifs is one of the computational solutions for CREs detection [39].

1.4.1.1. "Homotypic" versus "Heterotypic" clustering methods and their advantages and disadvantages

Several approaches have been developed to identify TFBS clusters in genomic DNA [57, 56, 58]. Such methods can be either assigned to the homotypic clustering group, containing multiple sites for one particular TF, or assigned to the heterotypic clustering category, containing one or more binding sites for multiple transcription factors (TFs) [60, 57].

Using heterotypic clustering, Berman *et al.* [22] identified CREs that are adjacent to genes expressed in the early embryo with any combination of TFBSs for BCD, HB, KR, KNI and

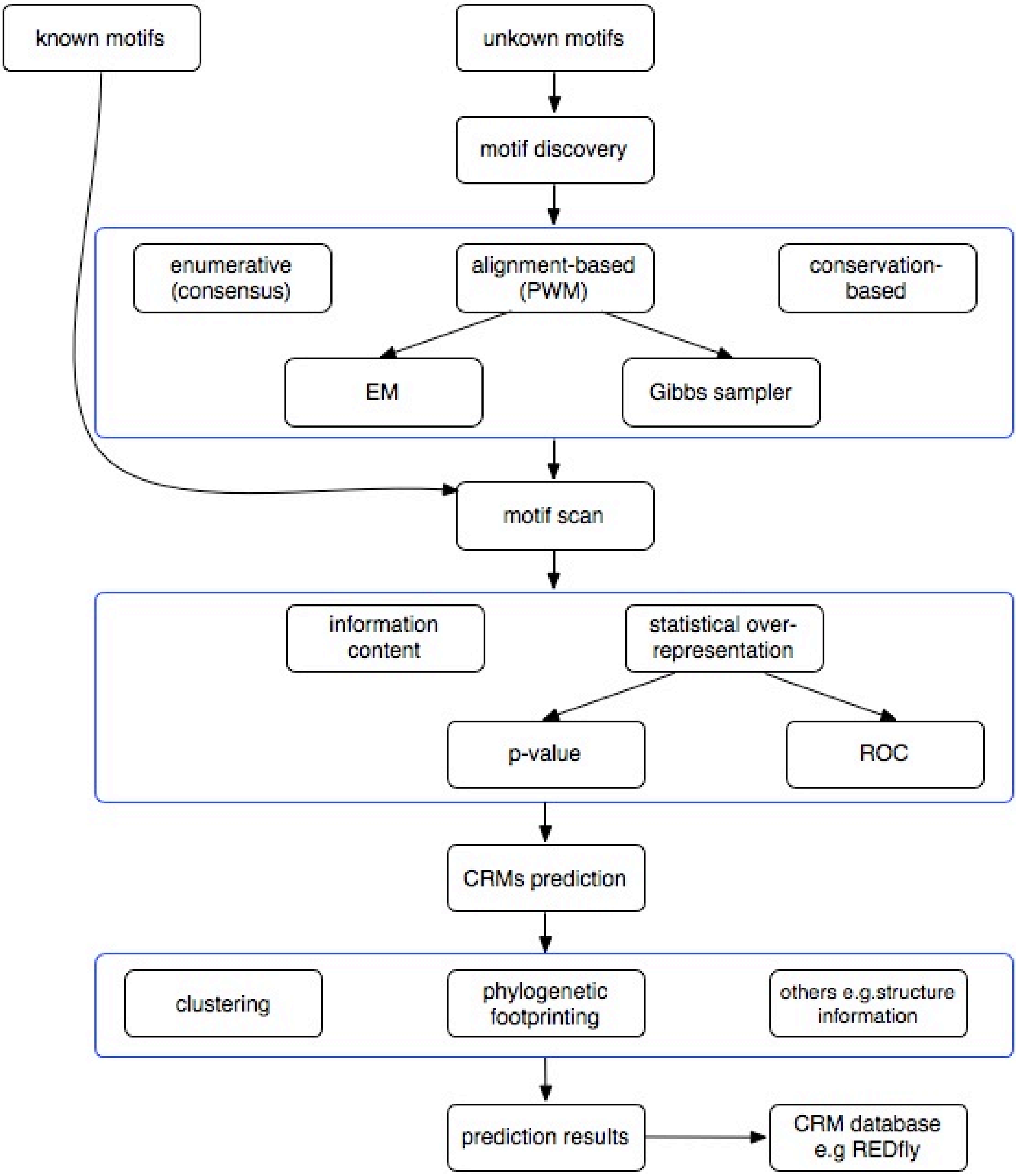


Figure 1.5.: CREs prediction workflow. The prediction procedure includes motif discovery, motif scan and CREs prediction. When motifs are known in advance, the motif discovery step can be omitted.

CAD factors [61]. Other examples of heterotypic TFBS clusters for several well characterized TFs have been discussed in [2, 40, 62, 63].

Using homotypic clustering, Ochoa-Espinosa *et al.* [64] identified 11 BCD-dependent CREs. Such clustering methods have also been applied in the studies of single TFs such as: Dorsal [65] and Suppressor of Hairless [66]. In over 60 known CREs from 20 *Drosophila* developmental genes, Lifanov *et al.* [67] found evidence that each type of recognition motif can form significant clusters within the regulatory regions of the corresponding TF.

Of course, both clustering methods have their advantages and disadvantages. For example, the homotypic method was able to form significant clusters within the regions regulated by the corresponding BCD TF, but it failed to detect all *eve* stripes enhancers [67]. In another example, 7 of 11 BCD-dependent CREs were not predicted by the heterotypic clustering method but were by homotypic clustering [64], showing that the homotypic method applied where appropriate can be useful.

A search for homotypic clusters, which requires clustering of each binding motif, tends to have a higher degree of selectiveness than searching with heterotypic cluster models [67]. The homotypic clustering method provides an initial separate consideration of distinct binding motifs which is relatively more flexible, as one can describe the final heterotypic clustering method as a combined specific homotypic clustering method, built for each motif separately [67]. However, the detection of individual binding sites by homotypic clustering runs the risk of making many false negative predictions [67, 68]. Usually, more than a single type of recognized motif is contained within the regulatory modules of transcription sequences in eukaryotes, thus heterotypic clustering models are more potent for the identification of CREs [67]. Moreover, position and distance specificity should be taken into consideration to determine CREs [69]. The heterotypic clustering method is more advanced in the construction of biologically relevant, complex regulatory models with multiple TFs.

The comparison of advantages and disadvantages between these two methods suggests that the two types of methods (homotypic and heterotypic) complement each other [61].

1.4.1.2. Prediction procedure of clustering methods

The expression patterns of the embryonic developmental genes are quite complex, their control regions often contain multiple separate CREs controlling different aspects of the pattern [39].

In order to identify embryonic related CREs, the prediction procedure often starts with scanning motifs. For motifs represented as a consensus sequence, scanning is accomplished by searching for subsequences that match the consensus word, with a pre-specified threshold on the number of allowable errors. For motifs represented with PWMs, a threshold score has to be specified to define a motif match. Methods of assessing cutoff thresholds for motif matches usually include measure of information content [70] or statistical overrepresentation (e.g. hypergeometric p-value or ROC score) [71, 72, 73, 74, 75]. The region with a high density of TFBSs suggests one CREs [59].

1.4.1.3. Difficulties of *cis*-regulatory elements prediction by clustering methods

CREs have distinct features that as a group distinguish them from other types of DNA sequences, thus TFBS clustering is able to provide classification of sequences as CREs. However, these differences are typically not sufficient enough to reliably classify a given unknown sequence as regulatory or non-regulatory. Clustering alone is not a sufficient marker of regulation.

Berman *et al.* [22] demonstrated the searching for clusters of predicted TFBSs and identified 37 regions in the *D. melanogaster* genome with high densities of predicted binding sites for five TFs involved in anterior-posterior embryonic patterning. The separation between the positive CREs and negative CREs in these 37 regions is too poor to allow for successful discrimination.

In another example, a region with high density of TFs has been reported by experimental test in a region containing the gene *CG13334*. But until now, no CREs has been discovered in this locus [24] (see Figure A.1).

In the context of the blastoderm set of modules, it has been widely observed that a control region may have multiple CREs. If multiple known CREs lie in the same control region, the prediction task is more demanding than when each control region has exactly one CREs. The predictor has to have the additional ability to decide if there are one or more CREs in any particular input sequence. For example, the control region of *even-skipped* consists of five CREs. Berman et al. [22] and Lifanov et al. [67] were not able to identify all of them.

1.4.2. Prediction of CREs by comparative methods

As an alternative, conservation of non-coding sequences among divergent species has been used widely as a predictor of CREs [76, 77]. Comparative analysis of multiple genomes in a phylogenetic framework positively affecting both precision and sensitivity, thus producing more robust results than single-genome analysis [78].

Notable success has been achieved with comparative approaches [79, 80, 81, 77]. For example, Berman *et al.* demonstrated that the predictive value of TFBSs clustering approaches could be enhanced significantly by incorporating certain sequence conservation criterion [82]. The underlying assumption is that orthologous DNA sequences that serve a function common to the species have changed significantly less than neutral DNA over a sufficient phylogenetic distance [77]. Phylogenetic footprinting [83] is a comparative genomics approach to identify *cis*-regulatory elements that are conserved in homologous sequences across multiple species [58, 84, 85].

1.4.2.1. Prediction procedures by phylogenetic footprinting

There are three components to the existing phylogenetic footprinting algorithms [56]:

1. Defining a suitable set of orthologous sequences in relatively closely related species

2. Aligning the sequences of these orthologous species

3. Identifying segments of significant conservation

First, the diversity in evolutionary distance of the 12 available *Drosophila* genomes provides an excellent model for phylogenetic footprinting tests on non-coding conserved sequences [78, 86, 85, 87, 88].

If the species are very closely related, divergence of the non-functional sequences may not be high enough to allow functional sequence motifs to be identified; conversely, in remotely related species, the short conserved sequences have no difference from the background sequences [78]. For example, almost every CREs of *even-skipped* was predicted correctly for *D. melanogaster* / *D. pseudoobscura* and *D. melanogaster* / *D. ananassae.* The results were neither satisfactory for the closer relative *D. melanogaster* / *D. erecta* nor for the furthest relative *D. melanogaster* / *D. mojavensis*. This suggests that reference species selection should be carried out in relatively closely related species or in species whose evolutionary distance is similar to that in the training dataset [89]. Thus, defining orthologous sequences under common evolutionary pressure is the first important step for phylogenetic footprinting.

Second, once orthologous sequences are obtained, they must be aligned to identify segments of similarity. There are two broadly used strategies for such alignments: one for the detection of short similar segments and the other to provide an ideal description of similarity across a whole pair of sequences [56, 82].

For the former, a local alignment tool such as BLASTZ [90] identifies short segments of exact identity and defines alignments by extending the analysis from the edges of each seed, whereas global alignment is produced by using the Needleman-Wunsch algorithm [91]. LAGAN [92] searches for best alignment over the entire length of the sequence using local similarities as anchors [56, 82].

Once an alignment is defined, several tools are helpful in the interpretation of the data. The patterns of nucleotide identity in the alignment are analyzed and conserved regions such as regulatory regions are classified [56].

1.4.2.2. Disadvantage of *cis*-regulatory elements prediction by phylognentic footprinting

Most classical phylogenetic footprinting methods are based on alignment and thus use sequence alignment to perform their predictions. As a result, a high sensitivity to alignment errors or alignment uncertainty is apparent in these methods [88].

Stark *et al.* [86] selected 12 *Drosophila* genomes for analysis and used them for the location of new functional elements. The fact that only a 59% similarity between different alignment strategies for regulatory motif instances was found shows the prime importance of alignment accuracy in phylogenetic footprinting [88]. Different designs of global and local alignment algorithms have different capabilities of detection of conservation [93, 85, 82].

In addition to the disadvantage of its dependency on quality of sequence alignment, phylogenetic footprinting has a secondary weakness due to the fact that *cis*-regulatory elements are not always conserved across species. Known regulatory regions showed only slightly more conservation compared to the remaining non-coding sequences in a comparison of *D. melanogaster* and *D. pseudoobscura* [94]. For example, it has been shown that the enhancers can maintain their function although several functional recognition sequences from the *eve* stripe 2 enhancer in *D. melanogaster* are no longer present in other *Drosophila* species [95, 33, 42]. Presumably, the loss of binding sites is covered by a gain elsewhere in the enhancer [24]. Additionally, some other apparently constrained non-coding DNA sequences have little or no obvious function [77].

Comparative genomics is a good indicator of function; phylogenetic footprinting can readily identify strongly conserved motifs. But, sequence conservation is a poor guide for CREs prediction. There is no reason to expect that all CREs will be under the same level of evolutionary constraint, and certainly many genes show differences in expression between species; in these cases the sequences of CREs should have changed [77]. The need for the development of comparative methods that go beyond measures of sequence identity and the need for experimental assays of regulatory activity is undeniable.

1.4.3. Combination of clustering and phylogenetic footprinting

Besides simple sequence conservation and TFBS clustering, I am particularly interested in developing better methods to improve the identification of regulatory sequences.

In this study, programs have been developed which incorporate the advantages of both clustering and sequence conservation methods. The prediction results provide answers to many questions related to the genes regulatory networks involved in embryonic patterning of *Drosophila* during evolution.

2. Results

2.1. Genome-wide computational identification of embryonic enhancers

2.1.1. Number of predicted enhancers

Previously, Hauenschild and Ringrose [96, 97] predicted specific classes of *cis*-regulatory elements, mainly Polycomb/Trithorax Response Elements (PRE/TREs) in *Drosophila* species. Other types of *cis*-regulatory elements such as enhancers may benefit from a similar approach. Here, I extend the work of [96, 97] to the prediction of developmental enhancers and to other fly species.

A computational approach from Berman *et al.* [1] tried to predict *cis*-regulatory elements in body patterning in the fly, and predictions were experimentally verified. The underlying principle in Berman's algorithm was to detect conserved clusters of transcription factor binding sites. My method is different: the approach is neither based on sequence conservation nor on classical clustering methods; it is a genome-wide computational approach for searching *Drosophila* embryonic enhancers. This method uses weighted paired motifs for enhancer prediction and is based on the scoring algorithm which has been implemented in jPREdictor [98]. 18 positive and 15 negative training sets were extracted from [1]. Additionally, five motifs are supplied in previous literature: Bicoid (BCD), Caudal (CAD) and Hunchback (HB) for the maternal factors; Kruppel (KR) and Knirps (KNI) as gap factors [12, 99, 100, 101, 30]. They act at very early stages of *Drosophila* development and are crucial in patterning the Anterior-Posterior (AP) axis of the embryo [22]. The requirements for my approach are fulfilled once I have prediction tools and training sets; prediction can begin. The jPREdictor algorithm slides a window across the whole genome; it counts motif pairs in each window and then assigns a score to each window. Paired motif weights are calculated as log-odds scores based on positive and negative training sets. The five transcription factors' position weight matrices (PWMs) are incorporated in the jPREdictor option file with a threshold, see Section 3.1.4.

The Motif PWMs' thresholds were selected to maximize the occurrence of motifs in the positive training set and minimize occurrences in the negative training set. For each motif, the motif weight for thresholds from 3 to 8 was calculated by paired motif settings (see Figure 2.1). First, the threshold was selected by the correspondence of the highest possible weight in the threshold region, such as motif CAD and HB. Second, if the weight remained the same value in a range of thresholds, the starting point of the threshold region was selected. E.g. the threshold region from 5.4 to 7 for KNI showed a constant weight in paired motif setting; giving a final threshold of 5.4. Third, the first peaks were considered to be the threshold for motif BCD and KR. Because the second peak value is too high, it will dramatically reduce the number of motifs that can be discovered. Furthermore, the additional

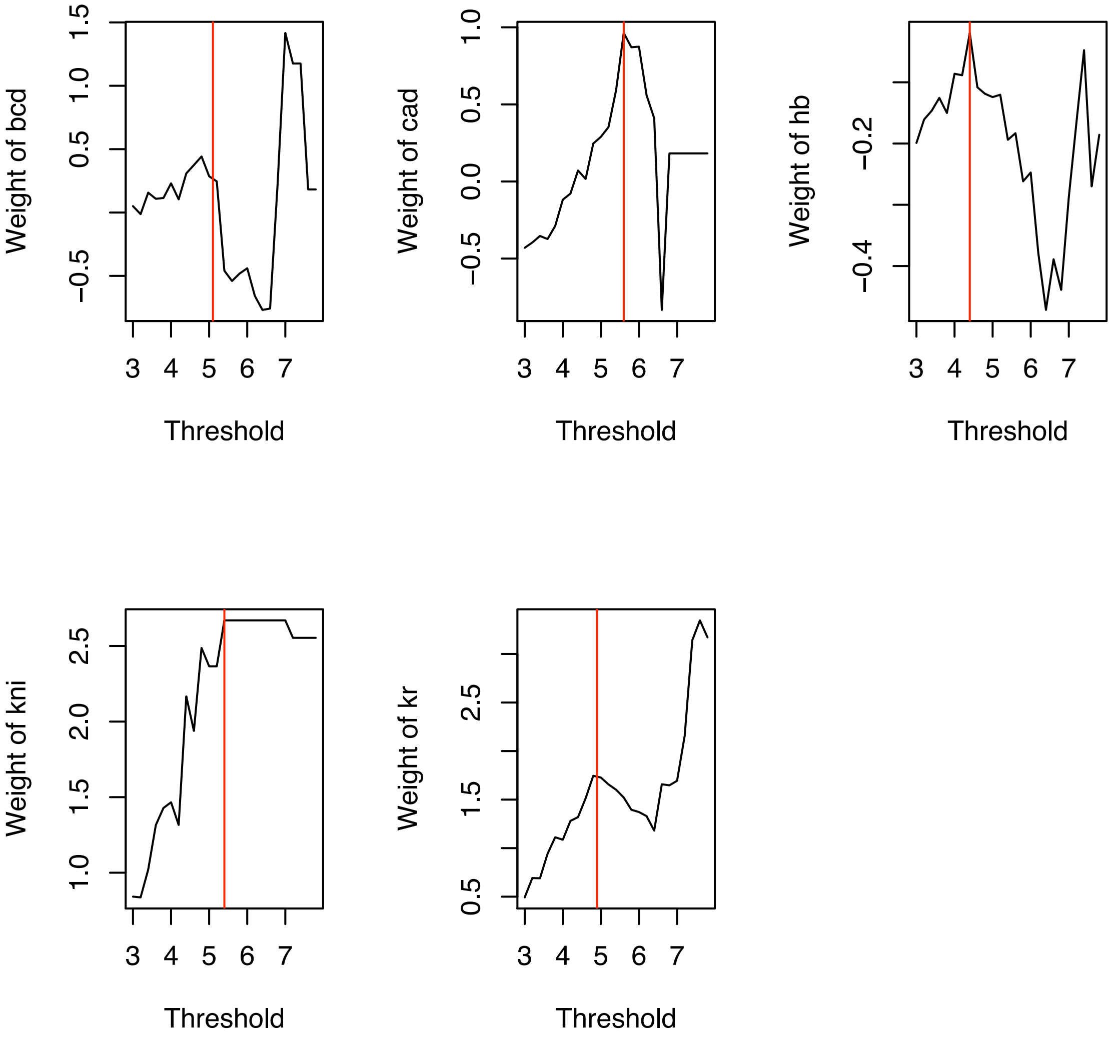


Figure 2.1.: Threshold setting by paired motifs. Motif weight is calculated with thresholds from 3 to 8 in steps of 0.2. The X-axis represents the region of PWM scores. The Y-axis represents the range of paired motif weights which distinguish paired motifs from positive/negative training sets. The selected thresholds for each motif are: BCD 5.1, CAD 5.6, HB 4.4, KNI 5.4, KR 4.9, which are the red vertical lines in this figure.

information from the single motif threshold plot defined the threshold of motif BCD to be 5.1. The final selected thresholds for each motif are: BCD 5.1, CAD 5.6, HB 4.4, KNI 5.4, KR 4.9 (see Figure 2.2).

The distribution of all motif combinations can be seen in Figure 2.3. On the one hand, the threshold of motif KNI shows the strongest impact on the positive training set. The self-paired motif KNI-KNI still has the highest weight. On the other hand, motif HB has an exceptionally low weight and remains a negative value. Which means, self-paired HB-HB occurs slightly more often in the negative training set. However, although self-paired HB-HB has the lowest weight, once HB pairs with the other motif, it starts to make a contribution to the positive training set.

Three parameters were selected in my approach: a window size of 700bp, window shift of 10bp and a paired motif distance of 150bp. Finally, the best parameters were retained after trying out every possible combination. The Kolmogorov-Smirnov (KS) test was used to tell the difference between the positive and the negative training sets. A $p-value$ of $3.4e-4$ suggests the difference is rather significant.

With these parameters the highest score of the positive training set (103.3) is about twice as high as the highest score of the negative training set (53.5). Additionally, 9 out of 15 elements in the positive training set have a higher score than the highest scoring in the negative training set (see Figure 2.4).

jPREdictor was used with these selected parameters to search for developmental enhancers in eight *Drosophila* genomes. All the scores and positions of each genome are stored in a database. In order to specify the significance of the scores, the jPREdictor run was then repeated with the same parameter settings on a random control data-set, which is 100 times larger (about 13G) than the *D. melanogaster* genome.

The choice of E-value or score cutoff is a critical issue, as the requirement for a more stringent match (a higher cutoff) is likely to result in fewer false-positive predictions but can potentially result in more sites being missed (false negatives) because they might just occur outside cutoff regions. The same kind of problem occurs when a larger E-value is used: the assumption is that more enhancers will be predicted, but a greater number of these ‘hits’ may be false positives because of the lower cutoff.

For an E-value of 1.0, only 1 false positive is expected for all of my predicted enhancers in *D. melanogaster*. My main focus is not on discovering all embryonic enhancers, but rather on finding real enhancers, thus a stringent cutoff retains high specificity. Taking into account a cutoff value of 50 which corresponds to an E-value of 1, I found 92 enhancers in *D. melanogaster*.

The same procedure was repeated on the other seven *Drosophila* species. The numbers of predicted enhancers in each species are listed in Figure 2.5.

The obtained genome-wide prediction numbers for *D. simulans*, *D. yakuba*, *D. pseudoobscura* and *D.persimilis* are less than 50. In *D. melanogaster* and *D. erecta*, there are less than 100 enhancers which have been found. However, in *D. sechellia* and *D. ananassae*, this number dramatically increases to several hundreds. The reason for this high prediction level is that the genome sequences for *D. sechellia* and *D. ananassae* are only available as scaffold versions, resulting in a high level of repeat sequences. For example, in *D. sechellia*, 204 statically predicted enhancers always have BLAST hits in *D. melanogaster*, but 137 of them have the same BLAST locus in chromosome X. The low quality of these genomes which contain too many repeats, causes the high number of genome-wide static predictions. An

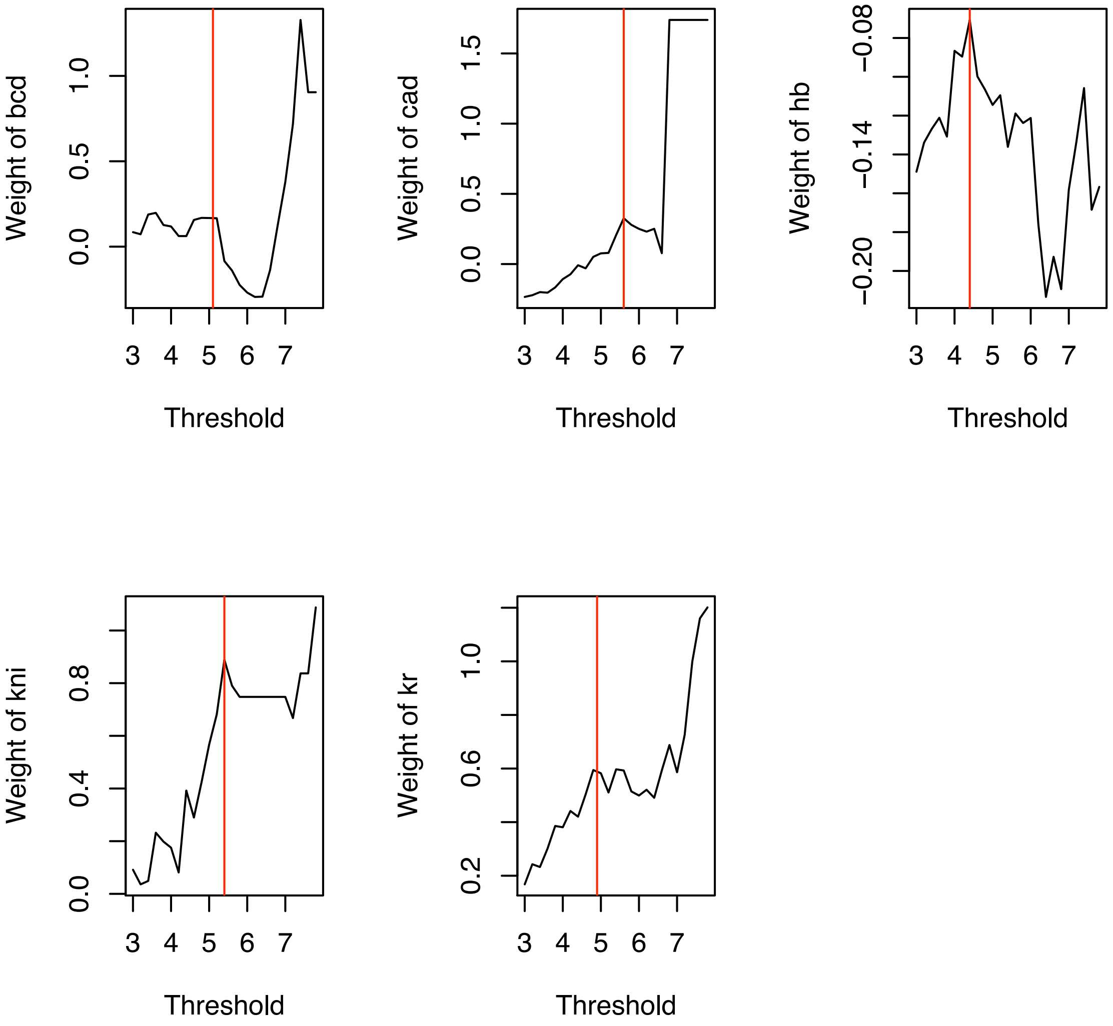


Figure 2.2.: Threshold setting of single motifs. Motif weight is calculated with thresholds from 3 to 8 in steps of 0.2. The X-axis represents the region of PWM scores. The Y-axis represents the range of single motif weights which distinguish single motifs from positive/negative training sets. The selected thresholds for each motif are: BCD 5.1, CAD 5.6, HB 4.4, KNI 5.4, KR 4.9, which are the red vertical lines in this figure.

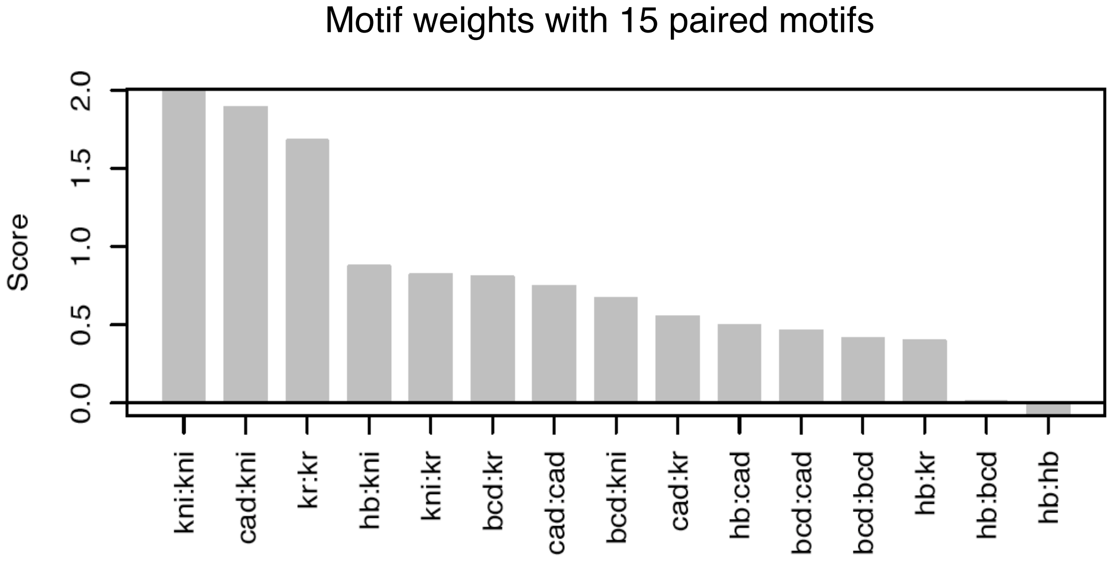


Figure 2.3.: The distribution of 15 paired motifs weights. The X-axis is the name of 15 pairwise motif combinations. The Y-axis is the motif weight in a range of (-0.0835, 2.0106).

equivalent observation can be made in *D. ananassae*.

2.1.2. Statistical evaluation - specificity & sensitivity

Sensitivity and specificity were assessed for the genome-wide prediction. First, sensitivity and specificity were calculated based on proven experimental results [1]. The jPREdictor algorithm was run on both Berman's positive and negative training sets. For each of Berman's enhancers, a score was assigned. I sorted these scores by positive and negative training sets. Specificity was calculated as the proportion of true positives over combined true positives and false positives, the latter being the number in the negative training set. There was only one element's score above my genome-wide prediction with E-value 1, cutoff 50. I found 17 true negative and the sum of true negative and false positive is 18. The sensitivity($= \frac{TP}{TP+FN}$) of 60% and specificity($= \frac{TN}{TN+FP}$) of 94.4% were calculated based on the training set of Berman (see Figure 2.6).

Second, specificity was also calculated in a different way, which was based on the E-value definition. Each of the specificities are calculated by 'predicted enhancers without False Positive (FP) estimation' divided by 'predicted enhancers'; e.g. when $E - value = 1$, $specificity = \frac{91}{92} = 98.9\% \approx 1$. This genome-wide expected specificity is in excellent agreement with the first specificity (94.4%) which was calculated base on 32 verified elements.

As a conclusion, the statistical analysis indicates that the prediction method is appropriate for *Drosophila* developmental enhancers prediction, with a prediction specificity of 94.4%; a sensitivity of 60% compared to Berman's prediction result when E-value is 1.

The same procedures are repeated to predict developmental enhancers from E-value 0.1 to

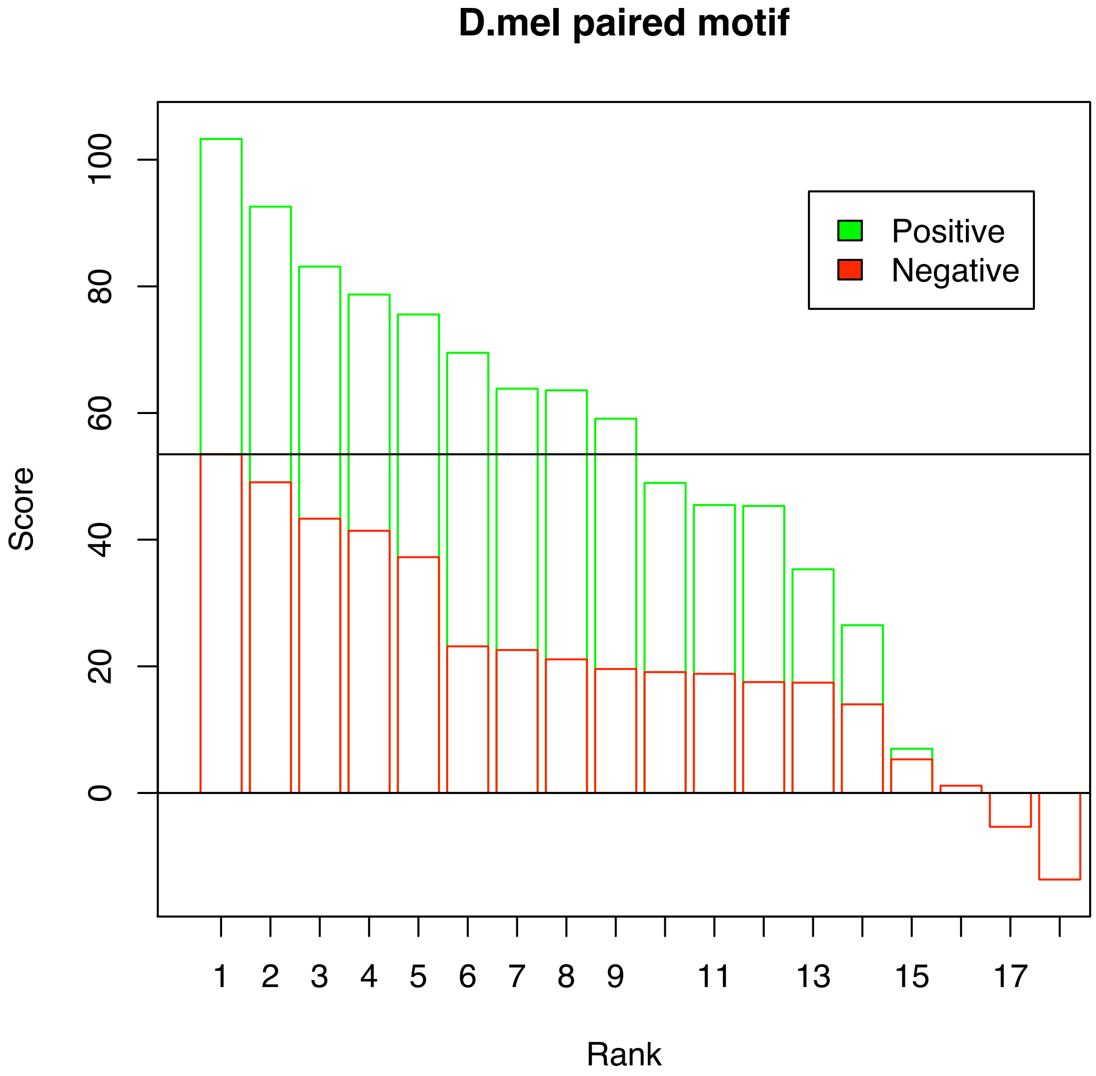


Figure 2.4.: The separation of positive and negative training sets with paired motif setting. Green bars represent 15 positive training set. Red bars show 18 negative training set. The black line separate the number of enhancers which have score higher than the highest score in the negative training set.

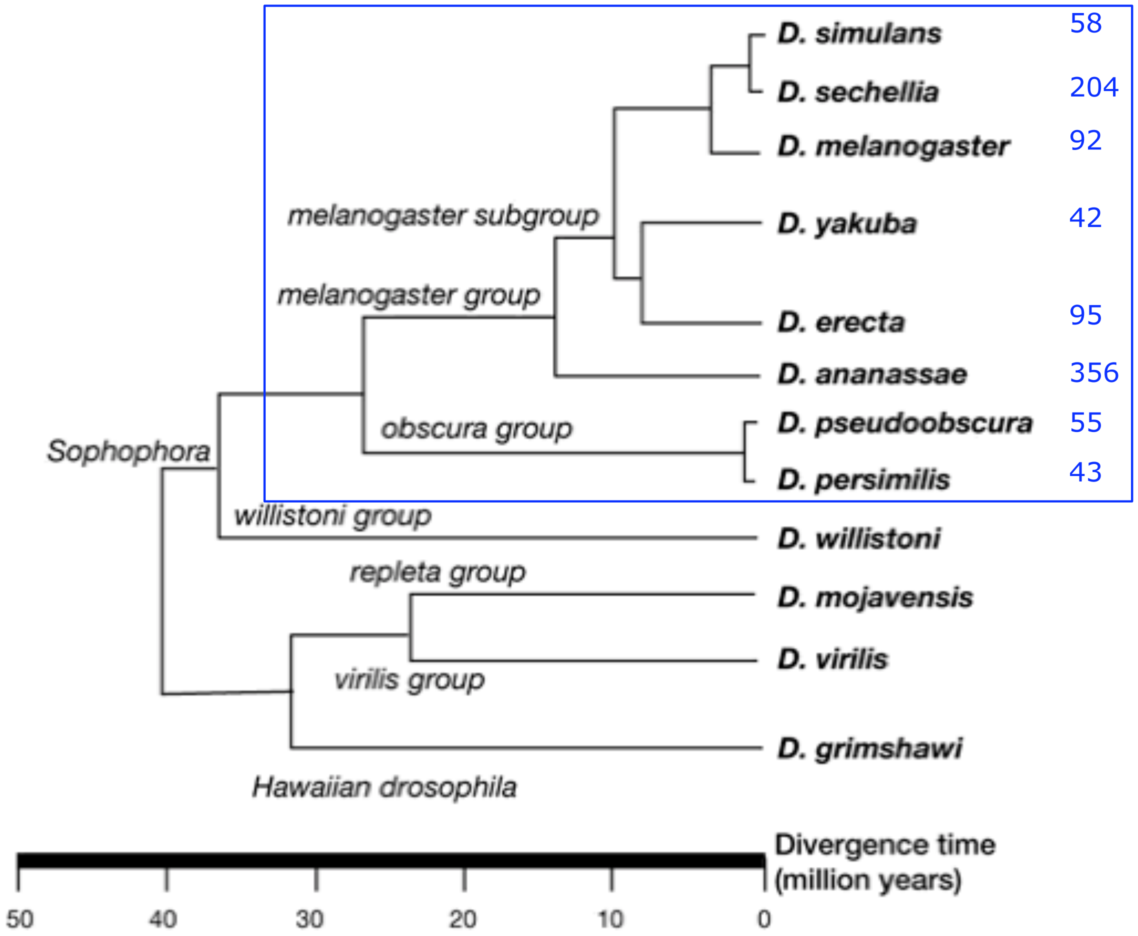
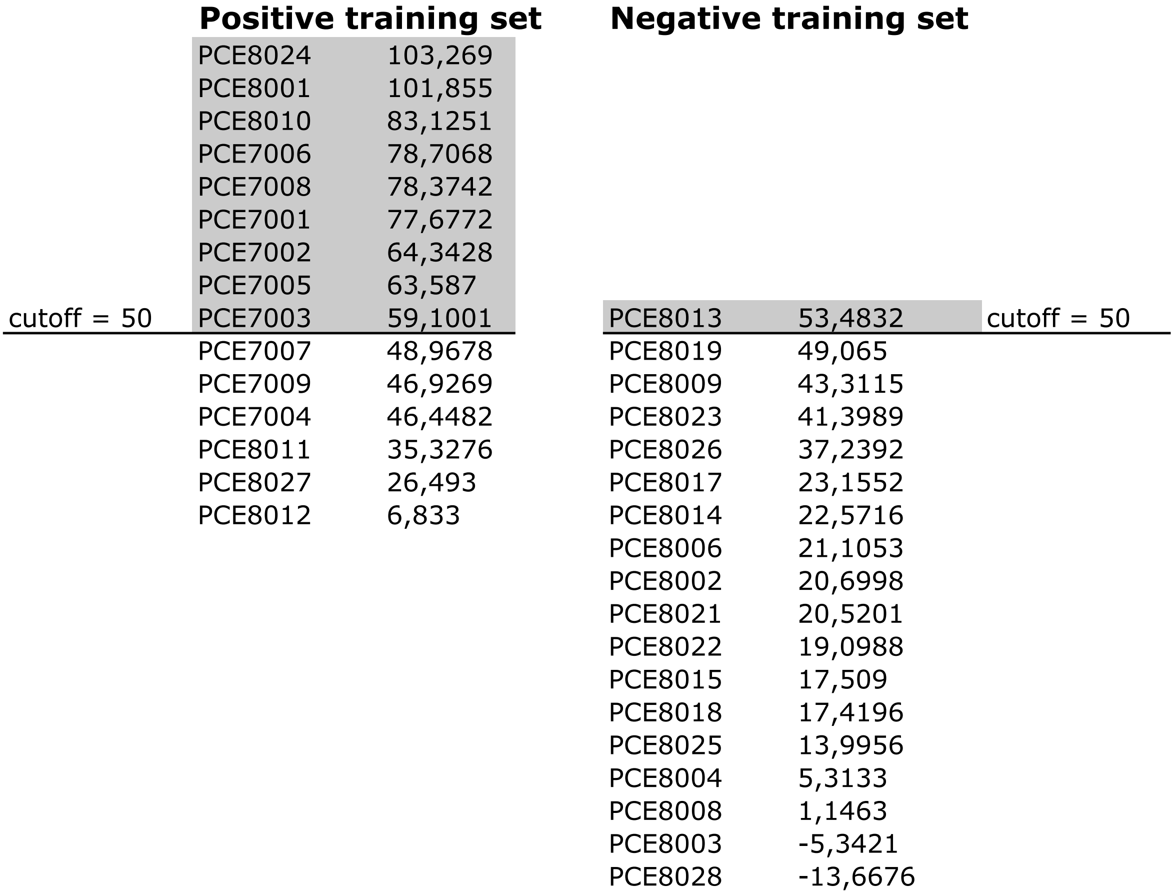


Figure 2.5.: Phylogenetic tree of *Drosophila*. The eight *Drosophila* species the analysis mainly focuses on, are highlighted with a blue frame. The numbers of predicted embryonic enhancers are listed next to the species names [102].



	Positive training set		Negative training set		
	PCE8024	103,269			
	PCE8001	101,855			
	PCE8010	83,1251			
	PCE7006	78,7068			
	PCE7008	78,3742			
	PCE7001	77,6772			
	PCE7002	64,3428			
	PCE7005	63,587			
cutoff = 50	PCE7003	59,1001	PCE8013	53,4832	cutoff = 50
	PCE7007	48,9678	PCE8019	49,065	
	PCE7009	46,9269	PCE8009	43,3115	
	PCE7004	46,4482	PCE8023	41,3989	
	PCE8011	35,3276	PCE8026	37,2392	
	PCE8027	26,493	PCE8017	23,1552	
	PCE8012	6,833	PCE8014	22,5716	
			PCE8006	21,1053	
			PCE8002	20,6998	
			PCE8021	20,5201	
			PCE8022	19,0988	
			PCE8015	17,509	
			PCE8018	17,4196	
			PCE8025	13,9956	
			PCE8004	5,3133	
			PCE8008	1,1463	
			PCE8003	-5,3421	
			PCE8028	-13,6676	

Figure 2.6.: Calculation of specificity and sensitivity with paired motifs setting on Berman's 15 positive and 18 negative training sets. The scores of each element are listed with IDs defined by Berman. The scores above the cutoff 50 are highlighted in grey. The final $Sensitivity = \frac{TP}{TP+FN} = \frac{9}{15} = 60\%, Specificity = \frac{TN}{TN+FP} = \frac{17}{18} \approx 1$
.

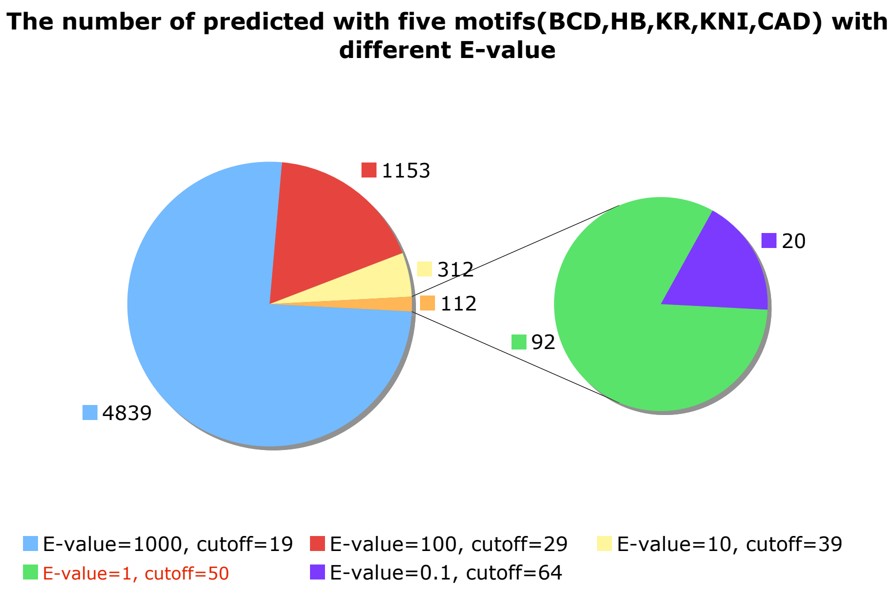


Figure 2.7.: The number of enhancer candidates with five motifs (BCD, HB, KR, KNI, CAD) in whole genome-wide prediction. The plot shows each cutoff corresponding E-value and the number of predictions. A small E-value setting produces less enhancers. The green slice is the main focus of this research.

E-value 1000. The higher the E-value, the more candidates can be predicted. But the number of FP is also increased. More stringent E-values reduce the number of predicted enhancers but increase the specificity (see Figure 2.7). In my work, the prediction with E-value of 1 balances both specificity and sensitivity, thus it is the main focus.

2.1.3. Enhancer location

Blanchette [103] documented the tendency for certain TFs to bind modules located in specific regions with respect to their target genes in human. To test this hypothesis in *Drosophila*, I checked the preference of embryonic enhancers' locations in my prediction.

I classified enhancers and their neighboring genes purely by their locations to have a brief overview of predicted enhancer location in the whole genome. I defined 7 types describing the positions of genes and enhancers (see Figure 2.8). The distribution of predicted enhancers suggests Type 4 and 5 are most frequent; Type 2 and 3 are possible; Type 1 is likely to occur; Type 0 and 6 are quite unlikely for enhancer location.

Among 92 genome-wide enhancers in *D. melanogaster*, 12 enhancers are close to the regions of TSS and the sites of termination of transcription with the distance less than 1kb; 57 of them are 1-10kb away from TSS or the sites of termination of transcription; 21 enhancers are possibly inside the intron; 68 enhancers are more than 10kb away from their neighboring genes. This preference of predicted enhancer locations is consistent with common biological agreement [104].

Sixty enhancers have their neighboring genes with a distance less than 10kb. Such distance

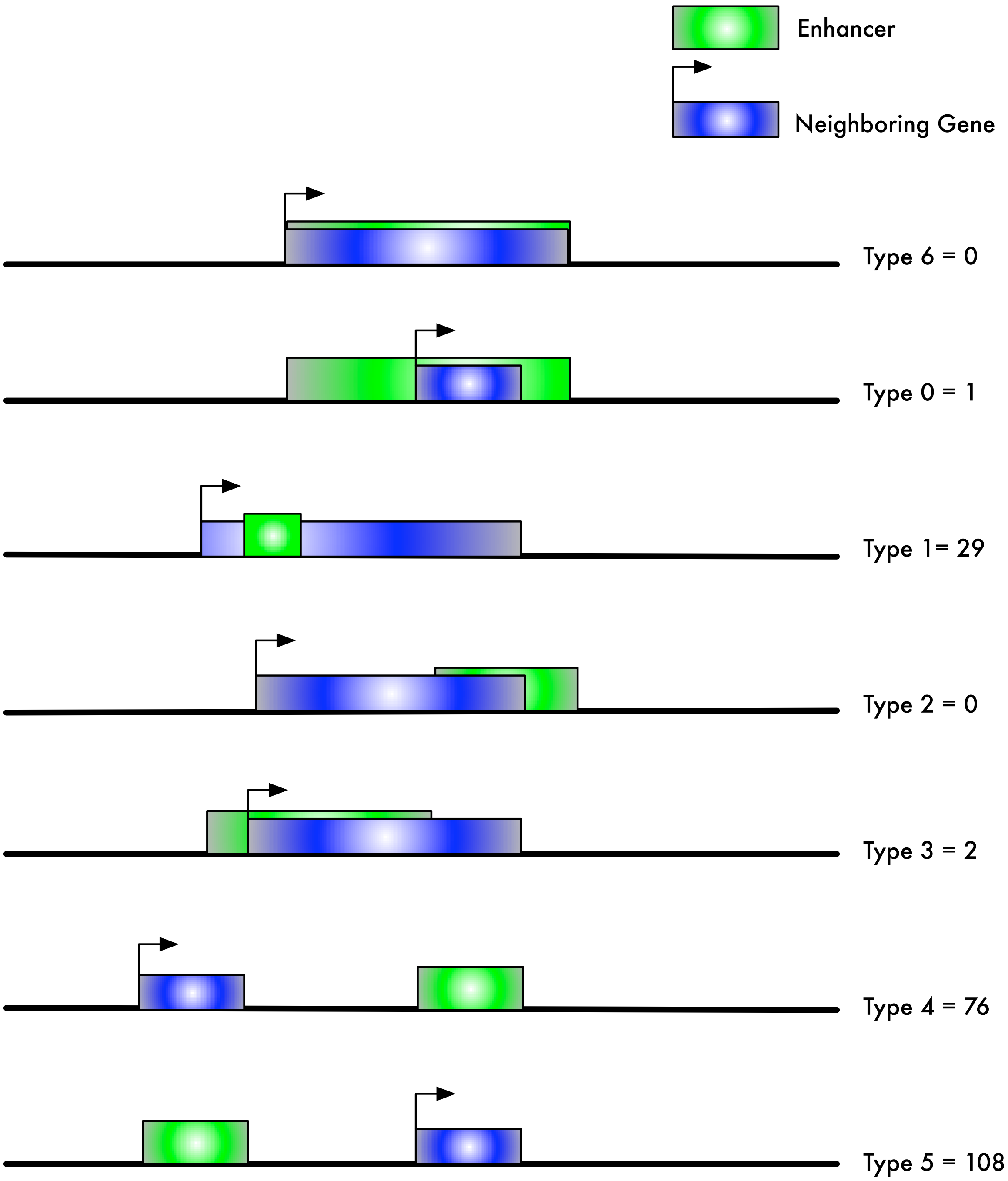


Figure 2.8.: Six types of enhancer locations relative to their neighboring genes. The neighboring gene is colored in blue and the predicted enhancer in green. Type 0 means the gene is located inside the predicted enhancer. Type 6 means the gene and the predicted enhancer are exactly overlapping. Type 1 means the predicted enhancer is located inside the gene. Type 2 means the predicted enhancer overlaps with its 5' gene. Type 3 means the predicted enhancer overlaps with its 3' gene. Type 4 means the predicted enhancer has no overlap with its nearest 5' gene, and is downstream of the gene. Type 5 means the predicted enhancer has no overlap with its nearest 3' gene and is upstream of the gene. The number of enhancers for each type of location is listed on the right side of graph.

Table 2.1.: Comparison of the number of predicted enhancers in the whole genome versus the embryonic subsets

Species	Genome size	Number of prediction	Cutoff
D. melanogaster genome-wide	132M	92	50
D. melanogaster subset	21M	57	41
Percentage(%)	15.9%	62.0%	

is relatively close, thus, these enhancers are able to regulate these close genes. The remaining 32 enhancers are more than 10kb away from their neighboring genes.

Information including types, gene orientation of enhancers with their neighboring gene can be accessed at my project webpage http://bibiserv.techfak.uni-bielefeld.de/jpred_en/.

2.2. Subset prediction increases the sensitivity of embryonic enhancer discovery by using Gene Ontology information

In the genome-wide enhancer prediction, the cutoff has to be very stringent in order to keep high specificity. E-value of 1 is equal to the cutoff value of 50. However, there are some elements which are just below the genome-wide cutoff and have been strictly excluded from the genome-wide prediction. Such elements could be in close proximity to embryonic development related genes. I speculated these elements to be the potential regulatory elements for those annotated genes. I therefore investigated a solution and carried out a variant search, where not the entire genome is scanned, but the regions around these small sets of genes.

Gene Ontology can be used as a filter for the query sequences, so that it is not necessary to scan the whole genome. The genes with the terms related to ‘embryo development’ from Gene Ontology’s biological process refer to the group of genes that are expressed in relation to early embryogenesis. The genes annotated with embryonic development related labels in Flybase gave 2813 *Drosophila* genes. The regions 10kb up/downstream around these selected genes were then searched. I therefore only needed to investigate my prediction in this 21M region in *D. melanogaster*. The same search procedure for genome-wide prediction was repeated on this subset region.

Illumined by the GO approach, the searching region was limited to specific GO-supported regions. This gives an embryonic subsets size of 21M, which is more than 6 times smaller than the entire genome (132M). The cutoff for subset prediction is 41 (E-value=1), which is equivalent to E-value 7 in the genome-wide prediction. Meanwhile, the genome-wide cutoff value of 50 is with E-value of 0.14 in the subset prediction. For E-value 1 in subset prediction, according to Figure 2.6, the sensitivity calculation based on positive training set is increased to be 80%. Therefore, if I choose E-value of 1 for the subset prediction, the specificity is kept to be stringent and sensitivity has been improved.

Moreover, 57 enhancers are predicted in the embryonic subset region. With only 16% of the genome size, 62% of the genome-wide enhancers could be found, when E-value is 1 (see Table 2.1).

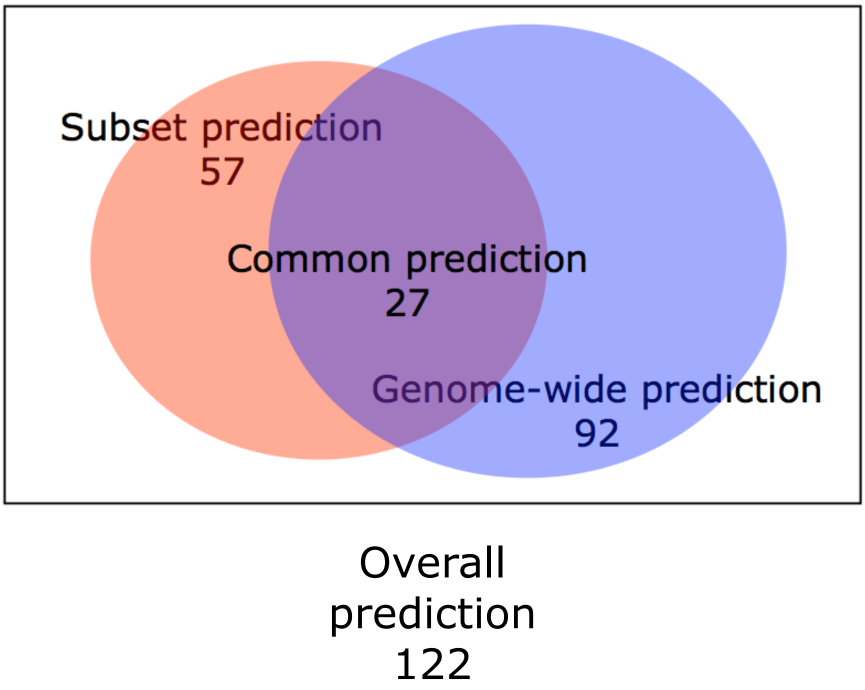


Figure 2.9.: Venn plot of genome-wide and subset prediction results. The common prediction of both methods is 27, the total prediction is 122.

What needs to be emphasized is that the results from genome-wide and subset prediction have a certain overlap, but they are not necessarily to be exactly the same element. The 27 common predictions by both methods have very high scores and are known to regulate embryonic genes. The light blue part shows 65 enhancers only found by genome-wide prediction. They are above score cutoff 50, but might not have neighboring genes from the embryonic category. The light red part indicates 30 enhancers which have only been found by the GO subset method (see Figure 2.9). These enhancers have neighboring genes from the embryo class, and the score is above 41 but lower than 50. So, the GO subset method makes up for the missing candidates from the genome-wide prediction. The total number of predicted enhancers becomes 122.

even-skipped locus is a detailed example that indicates a lower score cutoff for enhancer prediction by using the subset method. From the score plot of *even-skipped* region (see Figure 2.10), stripe 2, stripe 5 and stripe 3/7 are above the stringent genome-wide cutoff of 50 (E-value 1). Now, after applying the GO subset approach, the score cutoff gets significantly lower, resulting in inclusion of stripe 4/6 in the prediction.

As a short summary, GO subset prediction serves as a filter to improve the detection *in-silico*. The region's size has been minimized, the score cutoff is decreased and the specificity still remains as good as genome-wide. GO subset method improves genome-wide prediction by including more candidate enhancers as illustrated in Figure 2.9.

Strictly speaking, currently it is only possible to apply the subset method to *D. melanogaster*. Although the GO terms are species independent, GO annotation is only available in *D. melanogaster*. Nevertheless, by using BLAST on the embryonic genes in *D. melanogaster*,

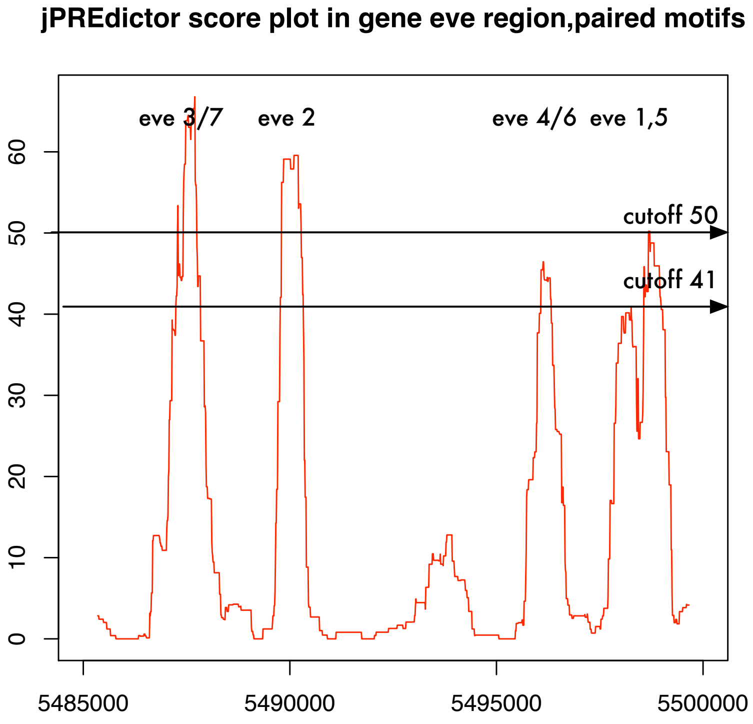


Figure 2.10.: Score plot of *even-skipped* enhancers with both genome-wide and subset cutoff. The x-axis is the position of each enhancer. The y-axis is the value of the prediction score. The GO subset method selects more *even-skipped* enhancers.

the homologous regions in other species can be suggested to have embryonic function as well. Thus, the GO subset method could possibly be applied to these species.

2.3. Dynamic search aids in discovery of conserved and non-conserved enhancers by using comparative genomics information

2.3.1. Evolutionary divergence of *Drosophila* species

The availability of complete sequences from 12 *Drosophila* genomes [86] in 2007 brings us the opportunity to carry out comparative analysis of *cis*-regulatory elements. As can be seen from the phylogenetic tree in Figure 2.5, some of the species, such as *D. melanogaster* and *D. simulans* are very closely related. Both of them separated from *D. yakuba* approximately 5 million years ago. The species in the melanogaster subgroup share the same chromosome arms. *D. pseudoobscura* is from the obscura group with an evolutionary distance of about 50 million years, and is one of the well-studied fruit fly species besides *D. melanogaster*. For the moment, the whole-genome sequences of *D. sechellia*, *D. ananassae*, *D.persimilis*, *D. erecta* are publicly available as scaffold versions. Since the evolutionary distance separating *D. melanogaster* and *D.grimshawi* is greater than the evolutionary distance separating any two mammals when generation time is taken into account [86], I mainly focus on the research of 8 *Drosophila* species from the obscura and melanogaster group. The great diversity among these *Drosophila* species in evolutionary distance (from 5 million to 50 million years) makes this 8 species set ideal for investigation of evolutionary forces on regulatory elements. Dynamic search was carried out for comparative analysis of *cis*-regulatory elements. The results of this dynamic prediction provide insights into the evolutionary forces working on embryonic enhancers.

2.3.2. Number of enhancers predicted by dynamic search

In the previous chapters, the genome-wide method successfully predicted enhancers by using a single genome. In this chapter, I will apply a method that uses multiple species comparison to assess the evolutionary dynamics based on single genome prediction. Since this comparative method goes beyond ordinary sequence conservation, the method is named as “Dynamic search”. Obviously, the genome-wide or embryonic subset predictions are statical methods compared to a multiple species dynamic search.

For every genome-wide predicted enhancer in each of the species, the putative functional analog was searched in each of the other species. This search procedure is performed in all eight pairwise species with a sequence radius setting from 1kb, 10kb to 20kb.

The 1kb radius surrounding the BLAST hit from genome-wide prediction of source species *D. melanogaster*, limited the search region to a sequence length of 92kb. The dynamically predicted enhancers in target species are mainly conserved within the orthologous region of source enhancer sites. Once the radii are set to be 10kb or 20kb, enhancers start to move away from the orthologous sites, but are believed to stay in the same locus and are expected to preserve the function. The list of search radii and corresponding score cutoffs is presented in

Table 2.2.: Score cutoff with different radius in dynamic search and in static search (E-value 1).

	1kb	10kb	20kb	subset	genome-wide
Cutoff	17	27	30	41	50
Sequence size	92kb	920kb	1.84mb	21mb	132mb

Table 2.2. According to the table, the score cutoff is lower once the search region is smaller. After finishing scanning the functionally analogous region, the number of dynamically predicted enhancers per pairwise query/target species in each direction of the eight *Drosophila* species is shown in Figure 2.11. The further the radius, the less possibility for the dynamic prediction to have preserved the function from the query element. Moreover, a large radius could cover larger search regions, this will cause less difference in cutoff for genome-wide and dynamic search. This is the reason why the radius setting stops at 20kb.

In summary, dynamic search increased the number of predictions and found additional enhancers previously undiscovered by the static methods.

2.3.3. Genomic position of enhancers is not conserved during evolution

Dynamic search provides the opportunity to observe the evolution of predicted enhancers. The hypothesis of dynamic search is that predicted analogous *cis*-regulatory elements might not be sequence conserved, but at least located around the same orthologous locus. In order to verify this hypothesis, I investigated the distances between the original best BLAST regions and the nearest comparatively predicted locus in four *Drosophila* species including *D. melanogaster*, *D. simulans*, *D. yakuba*, *D. pseudoobscura*. I noticed that analogous enhancers in different *Drosophila* species do not necessarily include the conservation of position. Figure 2.12 shows the dynamic search result of 92 source static *D. melanogaster* against *D. simulans*, *D. yakuba*, *D. pseudoobscura*. I observed a growth tendency of the distance between the predicted enhancer and its BLAST locus.

Eighty-five out of 92 statically predicted enhancers from source species *D. melanogaster* have BLAST hits in target species *D. simulans*. About 80.4% (74 out of 92) dynamically predicted enhancers are close to their best BLAST hit within a distance of 1kb (orange line in Figure 2.12). I observed a similar distance tendency in *D. melanogaster* and *D. yakuba* comparison; 91 out of 92 enhancers have BLAST hits in target species D. yakuba, and 78.2% (72/92) of the dynamic predictions are close to the orthologous locus within 1kb. However, in the evolutionary less conserved species *D. melanogaster* and *D. pseudoobscura*, only 71 dynamically predicted enhancers have BLAST hits, and 40.2% (37/92) of dynamically predicted enhancers in target species *D. pseudoobscura* are found at a distance below 1kb (red line in Figure 2.12). Because *D. melanogaster* and *D. simulans* are evolutionary closely related species; their functionally analogous enhancers stay close to the homologous position. But, *D. pseudoobscura* has diverged from the melanogaster subgroup about 50 million years ago. This divergence reflects the increase in distance. This observation is consistent with the species evolutionary distance from the phylogenetic tree (see Figure 2.5).

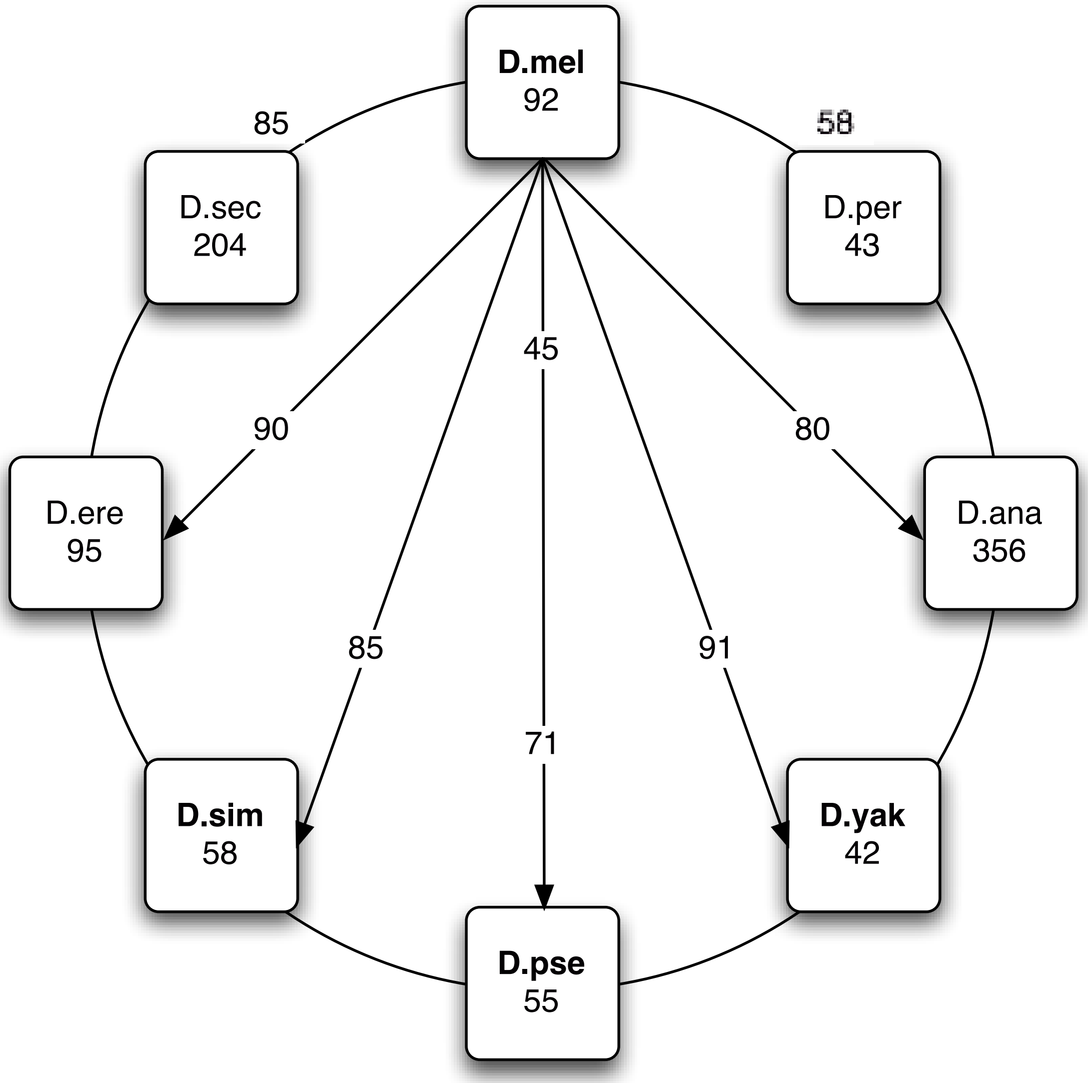


Figure 2.11.: The number of dynamically predicted enhancers per pairwise query/target species in each direction of the eight *Drosophila* species.

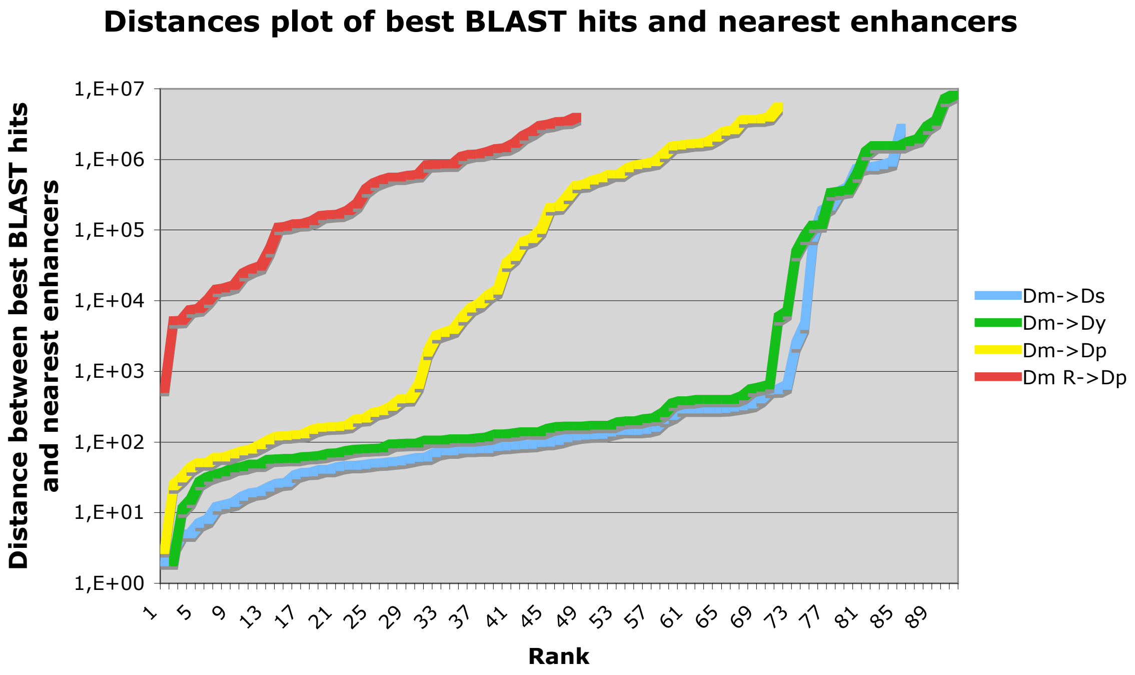


Figure 2.12.: Distance plot of predicted enhancer locus and its nearest BLAST hit (Y-axis) against number of predicted enhancers found (X-axis) within this distance. For target/source taken from *D. melanogaster* (Dm), *D. pseudoobscura* (Dp), *D. simulans* (Ds), D. yakuba (Dy). Dm R means random dataset generated by *D. melanogaster* predicted enhancers.

To make sure that the distribution of this distance plot is not arbitrary, I made the same distance comparison with the random data. Although Schaeffer *et al.*[105] mentioned that *cis*-regulatory sequences are slightly more conserved than random sequences, I expect to find larger distances in random / *D. pseudoobscura* than in *D. melanogaster* / *D. pseudoobscura*. The 92 statically predicted enhancers are distributed in 5 chromosomes in *D. melanogaster*; 36 enhancers in chromosome X, 22 enhancers in chromosome 3L, 15 enhancers in chromosome 2L, 9 enhancers in chromosome 2R and 10 enhancers in chromosome 3R. For each chromosome, I generated 10 sets of randomly positioned 1kb sequence, totally 920 random elements. For each random element, normal dynamic search was done to search for target hits in *D. pseudoobscura*, the distances between the BLAST hits and nearest predicted locus were again measured. To validate the randomization of the random set, I checked the distribution of the random BLAST distances by QQ-plot [106]. Most of the points (70%) lie on a straight line, up to 1.2e+06. The majority of random distances follows an exponential distribution (see Figure 2.13). Although the positions of random elements are randomly selected in each chromosome, these random elements have to strictly follow the real chromosome distribution. That is the reason for about 30% of 'larger' distances. The 'very large' distances (y-value > 4e+06) might be due to the limited length of the chromosomes, which means the 'very large' distance reaches the length of the chromosome.

One of these random distance sets was introduced into the distance plot (see upper left corner of Figure 2.12). Even the distances plot between *D. melanogaster* and the furtherest divergent species *D. pseudoobscura* is showing a smaller distance difference (y-value) than the random set. The random set is the shortest line overall. Fewer random elements have BLAST hits in *D. pseudoobscura* and the distances are much larger than the real enhancer sets.

In conclusion, the distribution shown in my distance plot follows the phylogenetic tree in *Drosophila*. Thus, it reflects the truth that some predicted enhancers are evolutionarily constrained, and for others that their genomic positions change rapidly during evolution.

2.4. Gain and Loss analysis of enhancers during evolution

2.4.1. Three groups of excluded enhancers

In order to study the evolutionary gain and loss of enhancers among *Drosophila* species I focus on eight species for which I made my enhancer genome-wide static predictions and they are with divergence time 0-30 million years from the phylogenetic tree (see Figure 2.5). To do so, I first had to check the presence of the predicted enhancers in all eight species by both static and dynamic prediction methods. In dynamic search, if the distance between the enhancer and the BLAST homologous locus is less than 10 kb, I positively indicate the predicted enhancers in the target species to be functionally analogous enhancers of the source species. There are three cases I did not take into count. I pointedly excluded these to build the gain and loss occurrence pattern.

First, the source species must be *D. melanogaster*. One reason is that *D. melanogaster* is the model species and my research is mainly carried out on enhancer presence in *D. melanogaster*. Another purpose is to exclude duplicates in the gain and loss analysis. *D.*

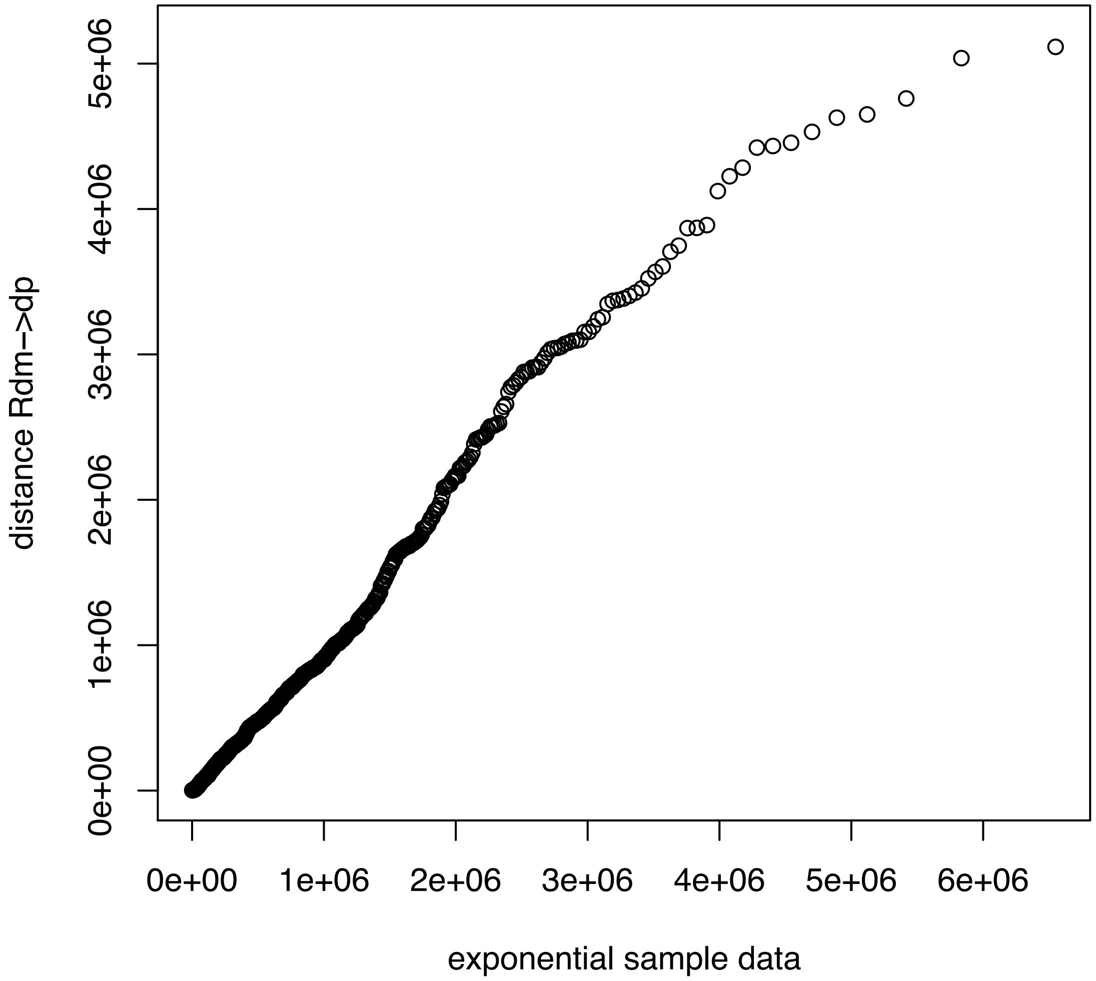


Figure 2.13.: The distances distribution of randomly generated elements from *D. melanogaster* to target species *D. pseudoobscura*. The random set is generated by randomly positioned 1kb sequences in the same chromosome distribution from real *D. melanogaster* enhancers. The distances from the random set's BLAST loci to the dynamically prediction in *D. pseudoobscura*, the majority (about 70%) of them follows an exponential distribution.

melanogaster and *D. simulans* are very closely related species; a static enhancer in *D. melanogaster* might be the same hit as a dynamic search hit from *D. simulans*. For example, the statically predicted enhancer in *D. simulans* (2L;16225400-16226299, ID:19219) has a dynamic prediction in *D. melanogaster* (2L;16526150-16527730), which is the same locus for the statically predicted enhancer of *D. melanogaster* (2L;16526490-16527529, ID:19120). If the analysis were made both from the source species of *D. melanogaster* and *D. simulans*, the same enhancer would have been counted twice. However, they both would represent the same type of gain and loss tree.

Second, if the static enhancer in source species has no BLAST hit in the target species, I designate this as target species absence group. For example, in the previous statically predicted enhancer (2R;14767220-14768089, ID:19131) , the presence of the analogous enhancer was indicated in *D. ananassae*, *D. simulans* and *D. erecta* because of the 10kb distance restriction. But, the lack of orthologous loci in *D. pseudoobscura* and *D.persimilis* suggests that the embryonic enhancer might never appear in these target species. There are 42 enhancers in *D. melanogaster* belonging to this target species absence group. They are not suitable for gain and loss analysis.

Third, enhancers which have full presence in all eight species are excluded, because the common ancestor automatically becomes present, such cases are found in this example of statically predicted enhancers (2L;3608980-3610169, ID:19111). There is no need to include these enhancer for gain and loss analysis, as they are totally conserved during the 30 million years of evolution. There are 21 enhancers which belong to this full presence group. However, if more information for the outlying species were to become available, the full presence in these eight species would influence the common ancestor and thus change the overall gain and loss analysis.

Finally, the remaining 29 enhancers are valid candidates to carry out the gain and loss analysis. If the distance between the enhancer and the BLAST homologous locus is further than 10kb, I assume that this enhancer is not present(0) in this target species. Otherwise, I assign presence(1) for these enhancers. The presence/non-presence(0,1) occurrence pattern is built with this data. Each column of the matrix is unique (see Table 2.3). This concludes the preparation for further gain and loss analysis.

2.4.2. Maximum parsimony method

Maximum Parsimony (MP) evaluates trees on the basis of the minimum number of character state changes required to generate the data on a given tree [107]. This method was first applied to estimate gain and loss of embryonic enhancers in *Drosophila* species.

Sixteen possible combinations of enhancer gain and loss types (see Table 2.4) were generated based on the occurrence pattern in Table 2.3. The numbers of gained enhancers and their neighboring genes are listed in Table 2.4. Every type of phylogenetic trees is available in Figure 2.14 and Figure 2.15. Enhancer state of *D. melanogaster* is always shown as presence (full dot), non-presence (hollow dot) and split-presence (half black and white), because the three excluded cases have been filtered (section 2.4.1).

Trees 1-5 colored in green are the types where *D. melanogaster* gains enhancers when I compare the enhancer presence status of *D. melanogaster* and its common ancestor. In the example of Tree1 (enhancer ID:19148,19180,19184), the branch of *D. melanogaster* is a full black dot which means the enhancer is present in the species, and the ancestor on

Table 2.3.: Occurrence patterns of enhancer presence/non-presence in each species. *Each row of occurrence patterns is in the species order of *D. simulans*, *D. sechellia*, *D. melanogaster*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. persimilis*. The right column of the table shows the *D. melanogaster* enhancer IDs belonging to the occurrence pattern.

Occurrence pattern*	**Enhancer IDs**
11111000	19148 19180 19184
11100000	19125
11110000	19159 19171
00100100	19122 19198
00111000	19173
11100100	19150
11111100	19112 19117 19124 19158
01111100	19172
00111100	19194
11110010	19183
11101011	19140
11111101	19115 19139 19143
11111110	19114
11111011	19118 19120 19121 19166 19177
00100111	19196
11101111	19153

Table 2.4.: Tree types in Figure 2.14 and Figure 2.15 by MP method. This table summarizes information of tree types, including the status of every tree type, the number of enhancers which belong to the tree type and the detailed enhancer IDs, neighboring genes' names. Column N* indicates the number of of *D. melanogaster* enhancers. The types in green are the trees which have gained enhancers during evolution. The types in red are the ones which are unclear. The types in black remain at their ancestor status.

Tree ID	gain and loss	N*	Enhancer ID	Neighboring Gene Name
1	gain	3	19148 19180 19184	*CG12525,CG32048,CG14177,CG12661,rdgA,CG10962,CG12662,fz4,CG9650*
2	gain	1	19125	*CG15477,CG31677*
3	gain	2	19159 19171	*CG18516,CG5302,CG4116,kirre,N*
4	gain	2	19122 19198	*CG15147,CadN,CG15454,CG15453*
5	gain	1	19173	*CG2875,AlstR,Ilp7*
6	not sure	1	19150	*CG6128,Mob1,CG7394*
7	not sure	4	19112 19117 19124 19158	*E5,ems,fs(2)ltoPP43,CG13962,nht,esg,TpnC25D,tkv,CG14033*
8	not sure	1	19172	*CG12535,CG14269*
9	not sure	1	19194	*dpr8,CG18313,CG9411*
10	not sure	1	19183	*CG12075,Moe,CG1885*
11	presence	1	19140	*CG32387,CG14826*
12	presence	3	19115 19139 19143	*CG32369,Pdp1,CG32365,vvl,Prat2,nub,pdm2,CG15485*
13	presence	1	19114	*bun,CG15489*
14	presence	5	19118 19120 19121 19166 19177	*beat-IIIc,CG6380,Tpr2,CG5953,CG31816,nht,esg,lace,CG15256,rk,CG33681,bgm*
15	presence	1	19196	*CG12433,CG8949*
16	presence	1	19153	*rpr,skl*

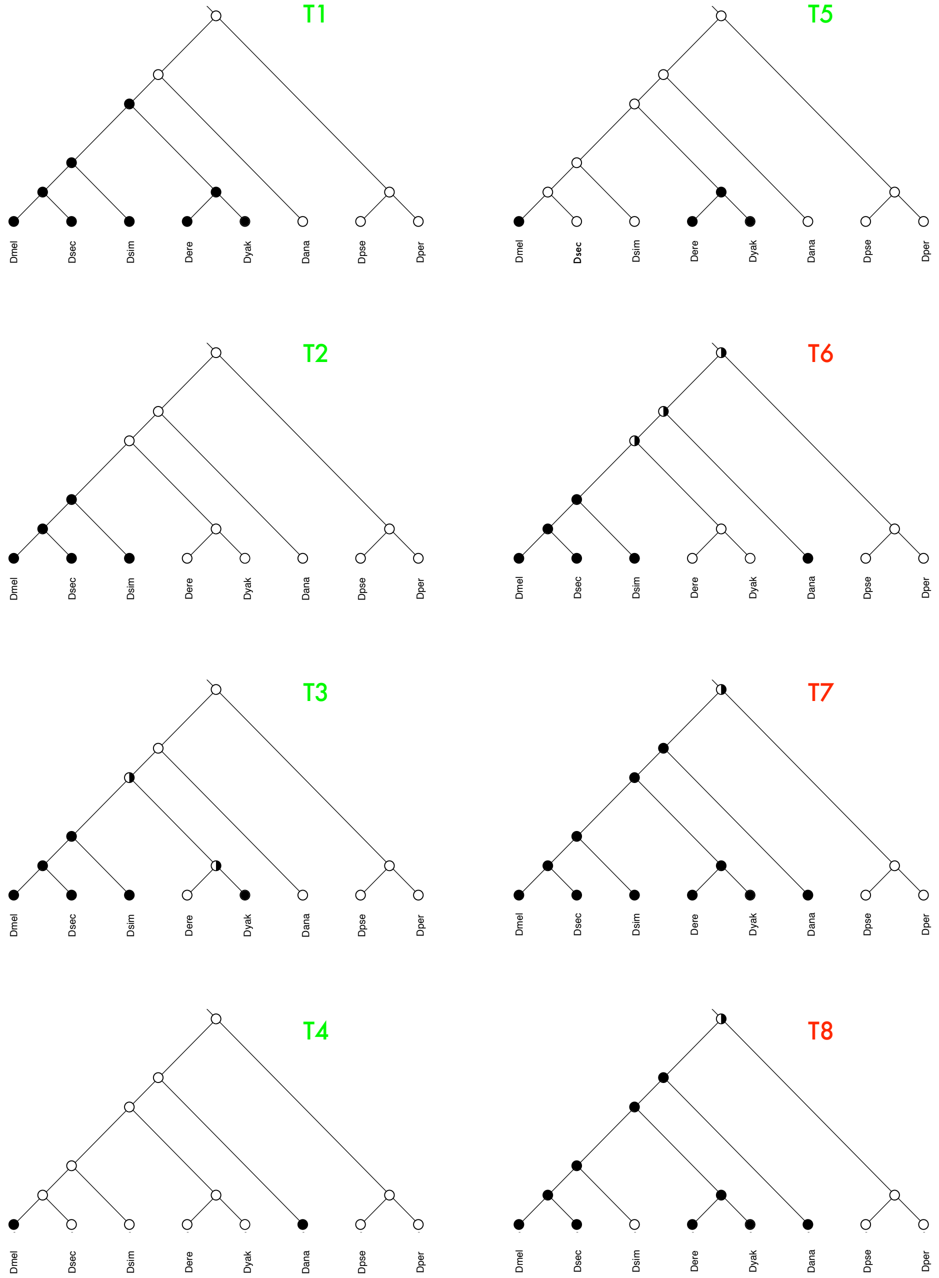


Figure 2.14.: Sixteen types of MP Trees. The tree is drawn by following the rule of phylogenetic distance. At the edge of the tree, the presence/non-presence of enhancers are presented with filled/hollow circles. The common ancestor has three possible statuses - presence(full), non-presence(hollow), split-presence (half black and white). The names of the species are listed below each tree. Each tree is unique, and is identified by numerical types. The presence/non-presence/split-presence status of the tree root is calculated by MP method. The types in green are the trees which have gained enhancers during evolution. The types in red are the ones which are unclear. The types in black remain at their ancestor status. (see Figure 2.15)

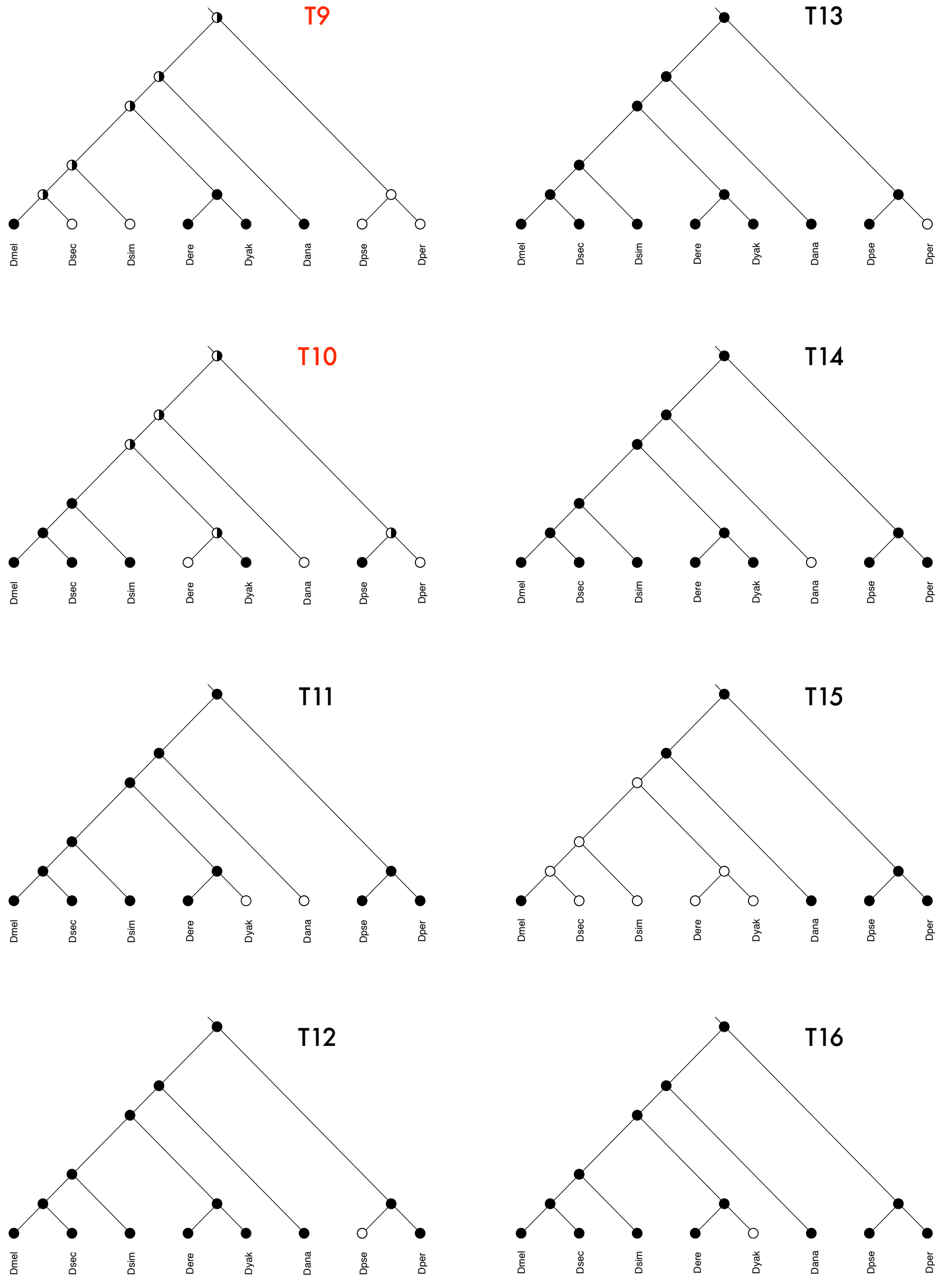


Figure 2.15.: Sixteen types of MP Trees. The tree is drawn by following the rule of phylogenetic distance. At the edge of the tree, the presence/non-presence of enhancers are presented with filled/hollow circles. The common ancestor has three possible statuses - presence(full), non-presence(hollow), split-presence (half black and white). The names of the species are listed below each tree. Each tree is unique, and is identified by numerical types. The presence/non-presence/split-presence status of the tree root is calculated by MP method. The types in green are the trees which have gained enhancers during evolution. The types in red are the ones which are unclear. The types in black remain at the ancestor status. (see Figure 2.14)

the top of the tree is a hollow dot that shows absence of enhancer. Following the rule of the least numbers of changes of the MP method, I assume that the enhancers present in *D. melanogaster* are those enhancers which have been gained during evolution.

Trees 6-10 colored in red represent situations where enhancer gain and loss is not clear. Additional out-group species are required to decide on their gain and loss status. For example, in Tree 9, the status of seven other enhancers influences the common ancestor. Although the enhancer is fully present in *D. melanogaster*, the uncertainty of the ancestor (half full, half hollow dot) leaves this type of tree in an indeterminate state.

The enhancers in Trees 11-16 remain present in *D. melanogaster*; meaning no enhancer gain and loss occurs during evolution. Unless additional out-group species are taken into account and influence the state of common ancestors, I believe they are evolutionary stable.

With the MP method, nine enhancers in *D. melanogaster* are gained, eight enhancers are not clear, and twelve enhancers are stable during evolution.

2.4.3. Maximum likelihood method

The Maximum Likelihood (ML) method introduces the concept of likelihood and has the advantage of representing the status of common ancestors by p-value. In previous analysis, MP method by default considers an equal parameter for gain and loss. However, in the real world sequence sites evolve non-identically over time [108]. The ML method allows a bias in gains versus losses and reconstructs ancestral character states based on the estimated parameters. Although both methods perform well most of the time [109], it is necessary to implement the ML method in the gain and loss analysis, in order to compare the difference of the results obtained by using both methods.

Again, 16 possible combinations of enhancer gain and loss types (see Table 2.5) were generated based on the occurrence pattern in Table 2.3 and 16 types of trees are generated by ML method. This time, the common ancestor in the tree has more than three statuses: presence (full), non-presence (hollow), split-presence (half black and white). Instead, a p-value is listed next to the tree types (see Figure 2.16 and Figure 2.17) .

The smaller the p-value is, the higher likelihood of enhancer absence in the ancestor. A *D. melanogaster* enhancer is defined to be gained if the p-value of its common ancestor is smaller than 0.5.

The common ancestors of Trees 1-8 in green have a p-value smaller than 0.5. These trees have gained enhancers in *D. melanogaster* during evolution. For Tree 9 in red, the gain and loss status is uncertain, because the p-value is 0.5. The situation corresponds to Tree 6-10 in MP method. Trees 10-16 in black have not gained or lost enhancers in *D. melanogaster*, since their p-value is larger than 0.5.

2.4.4. Embryonic enhancers can originate from non-functional sequence

The asymmetrical Markov k-state two-parameter model (AsymmMk model) is supported in the likelihood reconstruction method. The model is used in estimating transition rates among all possible pairs of states. The two parameters can be forward rate (one for the rate of change from state from 0 to 1) and backward rate (one for the rate of change from 1 to 0). Table 2.6

Table 2.5.: Tree types in Figure 2.16 and Figure 2.17 by ML method. This table summarizes information of tree types, including the status of every tree type, the number of enhancers which belong to the tree type and the detailed enhancer IDs, neighboring genes' names. Column N* indicates the number of of *D. melanogaster* enhancers. The types in green are the trees which have gained enhancers during evolution (p-value<0.5). The types in red are the ones which are unclear (p-value=0.5). The types in black remain at their ancestor status (p-value>0.5).

Tree ID	**gain and loss**	**N***	**Proportional likelihoods**	**Enhancer ID**	**Neighboring Gene Name**
1	gain	3	0.064	19148 19180 19184	*CG12525,CG32048,CG14177,CG12661,rdgA,CG10962,CG12662,fz4,CG9650*
2	gain	1	0.126	19125	*CG15477,CG31677*
3	gain	2	0.156	19159 19171	*CG18516,CG5302,CG4116,kirre,N*
4	gain	2	0.250	19122 19198	*CG15147,CadN,CG15454,CG15453*
5	gain	1	0.305	19173	*CG2875,AlstR,Ilp7*
6	not sure	1	0.413	19150	*CG6128,Mob1,CG7394*
7	not sure	4	0.427	19112 19117 19124 19158	*E5,ems,fs(2)ltoPP43,CG13962,nht,esg,TpnC25D,tkv,CG14033*
8	not sure	1	0.427	19172	*CG12535,CG14269*
9	not sure	1	0.500	19194	*dpr8,CG18313,CG9411*
10	not sure	1	0.624	19183	*CG12075,Moe,CG1885*
11	presence	1	0.750	19140	*CG32387,CG14826*
12	presence	3	0.819	19115 19139 19143	*CG32369,Pdp1,CG32365,vvl,Prat2,nub,pdm2,CG15485*
13	presence	1	0.832	19114	*bun,CG15489*
14	presence	5	0.832	19118 19120 19121 19166 19177	*beat-IIIc,CG6380,Tpr2,CG5953,CG31816,nht,esg,lace,CG15256,rk,CG33681,bgm*
15	presence	1	0.839	19196	*CG12433,CG8949*
16	presence	1	0.874	19153	*rpr,skl*

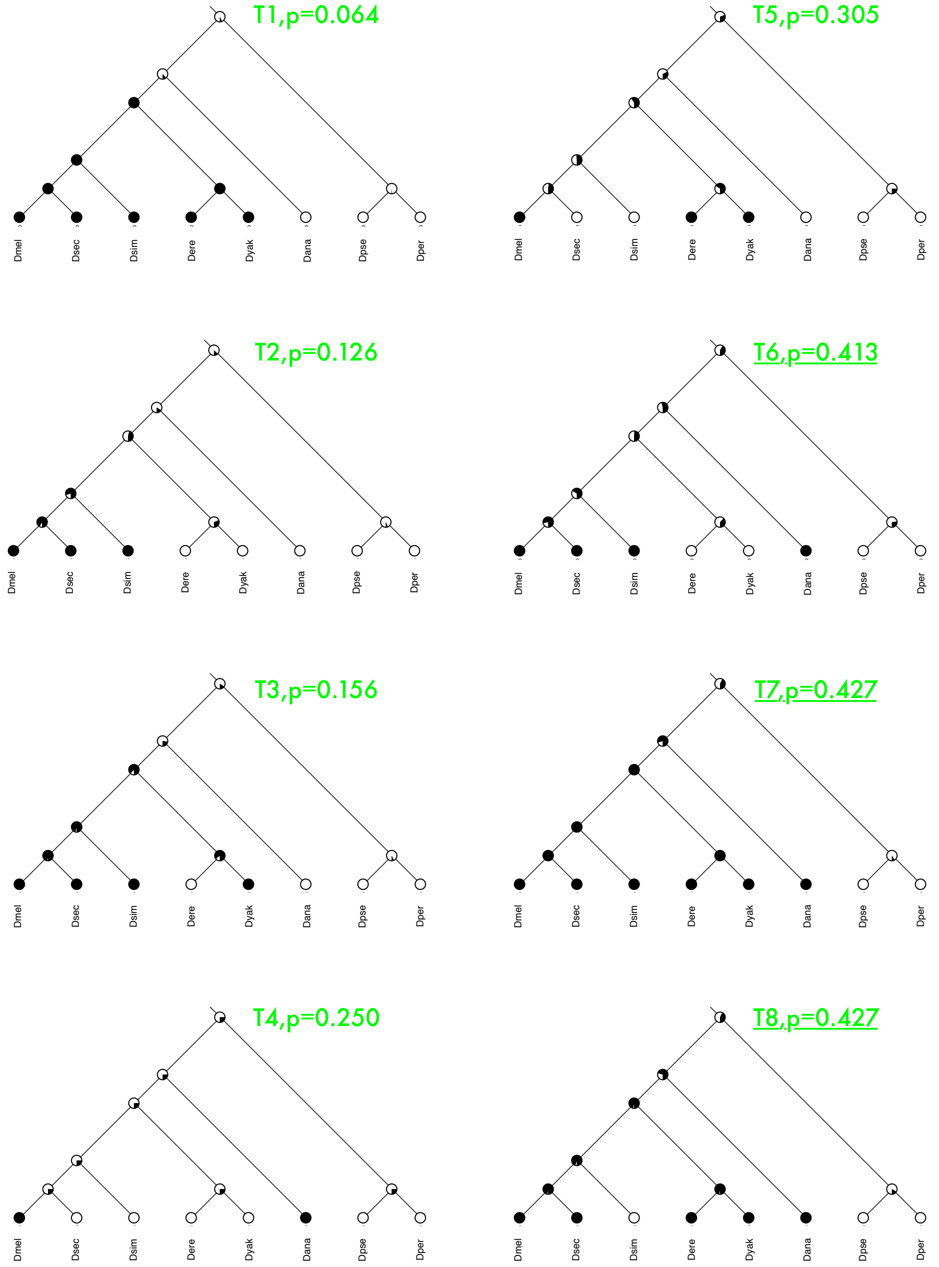


Figure 2.16.: Sixteen types of ML Trees. The tree is drawn by following the rule of phylogenetic distance. At the edge of the tree, the presence/non-presence of enhancers are presented with filled/hollow circles. The names of the species are listed below each tree. Each tree is unique, and identical by numerical types. The status of the common ancestor is calculated by ML method and can be identified by p-value. Thus a high amount of black in the common ancestor indicates a high p-value. The types in green are the trees which have gained enhancers during evolution. The types in red are the ones which are unclear. The types in black remain at their ancestor status. The tree types by ML method which show disagreement with MP method are underlined. (See Figure 2.17)

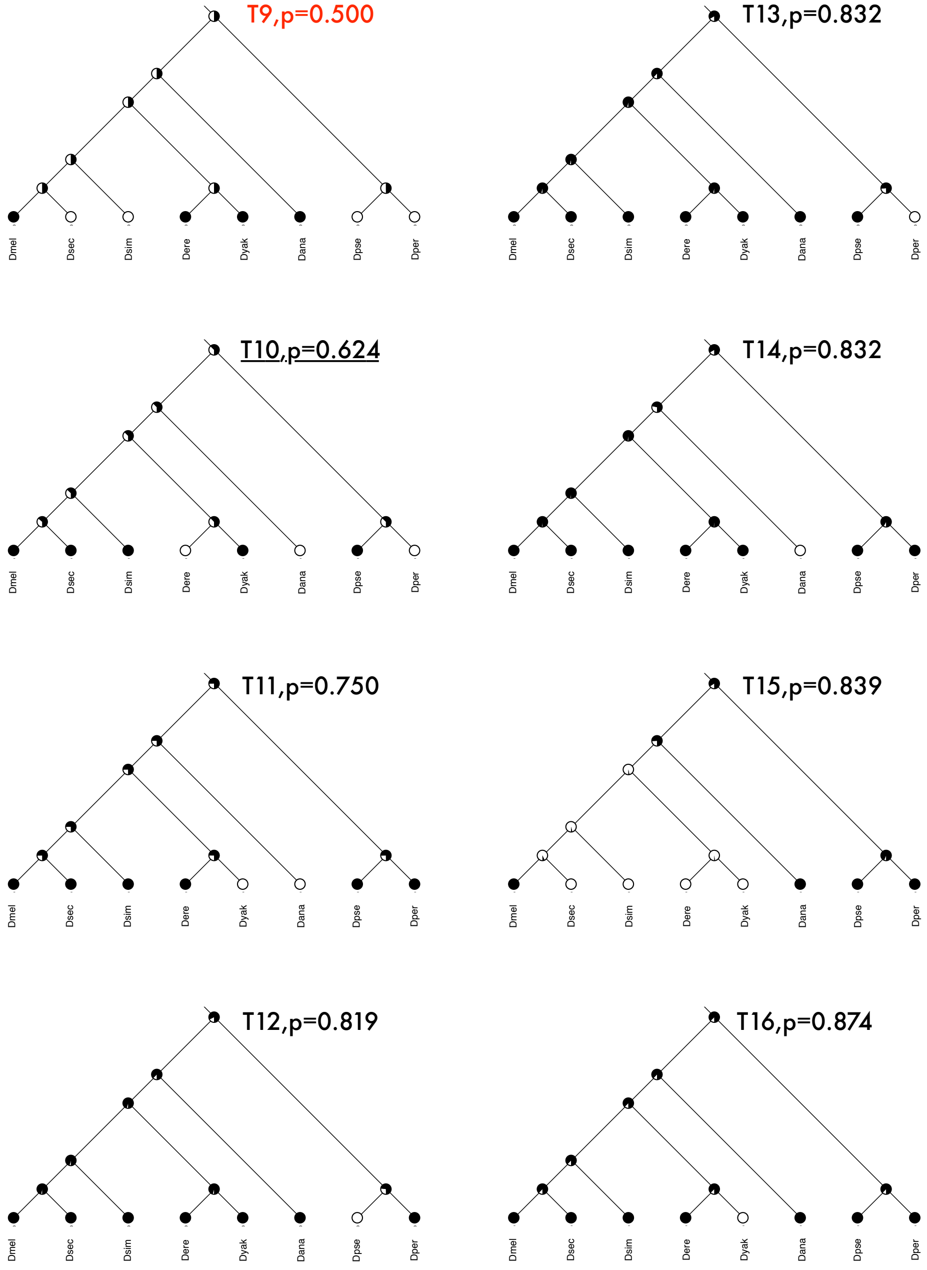


Figure 2.17.: Sixteen types of ML Trees. The tree is drawn by following the rule of phylogenetic distance. At the edge of the tree, the presence/non-presence of enhancers is presented with filled/hollow circles. The names of the species are listed below each tree. Each tree is unique, and identical by numerical types. The status of the common ancestor is calculated by ML method and can be identified by p-value. The blacker the common ancestor, the higher the p-value. The types in green are the trees which have gained enhancers during evolution. The types in red are the ones which are unclear. The types in black remain at their ancestor status. The tree types by ML method which show disagreement with MP method are underlined. (See Figure 2.16)

Table 2.6.: Forward and backward rate in 16 tree types by maximum likelihood method. The "forward" rate means the rate of change from state from 0 to 1 and the "backward" rate means the rate of change from 1 to 0 in asymmetrical Markov k-state 2 parameter model. Tree types with significant bias in gain versus loss are highlighted in blue.

Tree ID	forward rate	backward rate
1	0.020	0.021
2	0.038	0.084
3	0.040	0.063
4	0.957	2.872
5	0.107	0.192
6	0.112	0.115
7	0.040	0.019
8	0.064	0.054
9	4.536	4.536
10	0.644	0.387
11	1.119	0.373
12	0.239	0.037
13	0.226	0.035
14	0.143	0.025
15	0.064	0.038
16	2.238	0.320

lists the transition rates of all ancestors per tree type. Except for tree type 4,11 and 16, thirteen tree types have consistent forward and backward rate. Thus, there is no significant bias in gains versus losses. Therefore, parsimony and likelihood methods are both suitable for ancestral state reconstruction.

The same conclusion can be proven by comparing the generated trees from MP and ML methods in Figure 2.14 and Figure 2.16. Trees 1-5 have gained enhancers in *D. melanogaster*, no matter which methods I have used, but some conflicts seem to happen in Trees 6-10, presented as underlined in Figure 2.16. The p-value of 0.5 is the cutoff defined as separator for gain enhancers, Trees 6-10 from likelihood method have a p-value nearly close to 0.5. Therefore, both methods draw nearly the same conclusion on the enhancer gain and loss analysis. Of course, additional information from more *Drosophila* species can always improve the analysis precision.

Overall, the exact number of gained enhancer during evolution could be concluded from MP and ML analysis. Nine enhancers out of 92 (10%) from the genome-wide static prediction in *D. melanogaster* have been gained during evolution. The table lists the 9 embryonic enhancers which have been gained in D. melanogaster, together with enhancer neighboring genes and their distance to the enhancer (see Table 2.7).

Gene *CG4116* and *kirre* have two neighboring enhancers 19170 and 19171. Thus, only one of these 9 enhancers (19171) was gained from the genes which already have one enhancer (19170). In total, 8 enhancers are associated with genes that previously had no enhancers. The gain and loss analysis gives evidence that embryonic enhancers can originate from non-functional sequences, and moreover suggests that genes can gain new regulation in embryo development.

Below, I discuss a selection of detailed examples with gene annotation.

In Tree 4, examples of enhancer 19122 and 19198. Enhancers present in *D. melanogaster*, *D. ananassae* but do not occur in closer related species such as *D. sechellia*, *D. simulans*, *D. erecta* and *D. yakuba*. Both neighboring genes (*CG15454* and *CG15453*) of enhancer 19198 have an annotation of regulation of transcription. Neighboring gene *CG7100* of enhancer 19122 also has an annotation of embryonic nervous system.

In Tree 3, neighboring genes of enhancer 19159 have no embryonic related annotation. Neighboring gene *CG3936* of enhancer 19171 has annotation of transcription activator activity, protein binding, specific transcriptional repressor activity, regulation of developmental process and organ morphogenesis. It is not clear why this enhancer is only absent in *D. erecta*.

In Tree 2, the function of neighboring genes (*CG15477, CG31677*) of enhancer 19125 are unknown yet. Since enhancers are gained individually in *D. simulans*, *D. sechellia* and *D. melanogaster*, and not in any of further species, I suspect this gained enhancer only started to appear around 3 million years ago.

2.5. Enhancer plasticity in *Drosophila* species

In vertebrates, comparative genomics methods are commonly used to identify conserved non-coding sequences [1]. Genome-wide enhancer prediction has been demonstrated by using static and dynamic search; these prediction results could be used for cross-species comparison and evaluation of potential dynamics in evolution. Recently, binding site reorganization

Table 2.7.: Gained enhancers in *D. melanogaster* and distance to their neighboring genes. Distances of zero indicate that enhancer and gene overlap.

ID	chr	gene begin	gene end	distance	gene ID	gene Name
19122	2L	17621676	17622274	776	*CG15147*	*CG15147*
19122	2L	17645795	17735450	21626	*CG7100*	*CadN*
19125	2L	20532234	20532752	51018	*CG15477*	*CG15477*
19125	2L	20591182	20591789	6223	*CG31677*	*CG31677*
19148	3L	9571801	9572142	24772	*CG14177*	*CG14177*
19148	3L	9516187	9517946	28354	*CG12525*	*CG12525*
19148	3L	9521167	9568461	0	*CG32048*	*CG32048*
19159	3R	11374350	11378494	45856	*CG18516*	*CG18516*
19159	3R	11458605	11464977	33516	*CG5302*	*CG5302*
19171	X	2956234	2988959	0	*CG3653*	*kirre*
19171	X	2991028	3028418	10999	*CG3936*	*N*
19171	X	2922924	2923718	55422	*CG4116*	*CG4116*
19173	X	3513768	3514428	64209	*CG13317*	*Ilp7*
19173	X	3427877	3429776	18664	*CG2875*	*CG2875*
19173	X	3430694	3510945	0	*CG2872*	*AlstR*
19180	X	7038312	7085656	70743	*CG9650*	*CG9650*
19180	X	6951508	6955528	11342	*CG4626*	*fz4*
19184	X	8768459	8768854	53436	*CG12661*	*CG12661*
19184	X	8828890	8829405	5491	*CG12662*	*CG12662*
19184	X	8770420	8870175	0	*CG10966*	*rdgA*
19184	X	8818129	8894208	0	*CG10962*	*CG10962*
19198	X	20402838	20404590	24009	*CG15453*	*CG15453*
19198	X	20375447	20376689	1441	*CG15454*	*CG15454*

has been reported in the *even-skipped* enhancers of *Drosophila* and *Sepsid* [110]; plasticity of other *cis*-regulatory such as PRE/TREs has been unveiled [96]. All of this evidence could be useful to answer several questions in my prediction work:

D. simulans and *D. yakuba* are species closely related to *D. melanogaster* and they belong to the same melanogaster subgroup. *D. pseudoobscura* diverged about 50 million years ago from the melanogaster subgroup. Is it possible that the melanogaster subgroup has the same amount of enhancers regulating embryo development? Since *D. pseudoobscura* is a relatively further species, might it contain a rather different number of enhancers? How common is position flexibility in analogous enhancers and is it important for gene regulation? Do conserved enhancers have high sequence conservation and motif conservation? All these questions will be discussed in this chapter.

2.5.1. First type of plasticity - the number of enhancers varies widely across species

In genome-wide prediction (see Figure 2.5), *D. melanogaster* has about 40% more enhancers than *D. simulans* although *D. simulans* is the most closely related species to *D. melanogaster*. Another melanogaster subgroup species, *D. yakuba,* is in a similar situation, and the number of predicted enhancers in *D. yakuba* is even less than half of the prediction for *D. melanogaster*. For the evolutionary furthest species, the prediction number of *D. pseudoobscura* only matches the number of *D. simulans*. However, the genome sizes of the *D. melanogaster* subgroup are more or less the same as *D. pseudoobscura*. The numbers of predicted enhancers in these four *Drosophila* species are various when I compare the results from the genome-wide static method (see Figure 2.18). About 10% of the genome in *D. simulans* contains 'N' nt, which means any nucleotide including A, T, C or G. This might partly explain a shortage of predicted enhancers in *D. simulans* compared to *D. melanogaster*. Nevertheless, it's not likely that 40% of enhancers disappear in these 10% regions.

Previously, Hauenschild and Ringrose [97, 96] performed biological tests to compare the cytological positions of the predicted PRE/TREs with immunocytologically mapped PcG and TrxG binding sites. The *in-silico* result of PRE/TREs shows excellent agreement with experimental results. The architecture and evolution of PRE/TREs might be different from embryonic enhancers. Enhancer position can also be verified by immunocytologically mapping.

Additionally, I observed that embryonic enhancers can originate from a non-functional sequence (section 2.4.4). This could serve as another explanation for the various prediction numbers.

To summarize, there is no correlation between the species genome sizes and the number of embryonic enhancers predicted during evolution among species. On average, the number of predicted enhancers in *D. melanogaster* are 40% to 50% more than in the other three species. I consider this to be a first type of enhancer plasticity.

2.5.2. Second type of plasticity - enhancer position plasticity

Although conservation does imply function, it does not follow that all functional elements must be conserved, nor that non-conserved DNA has no function [96]. Investigating the

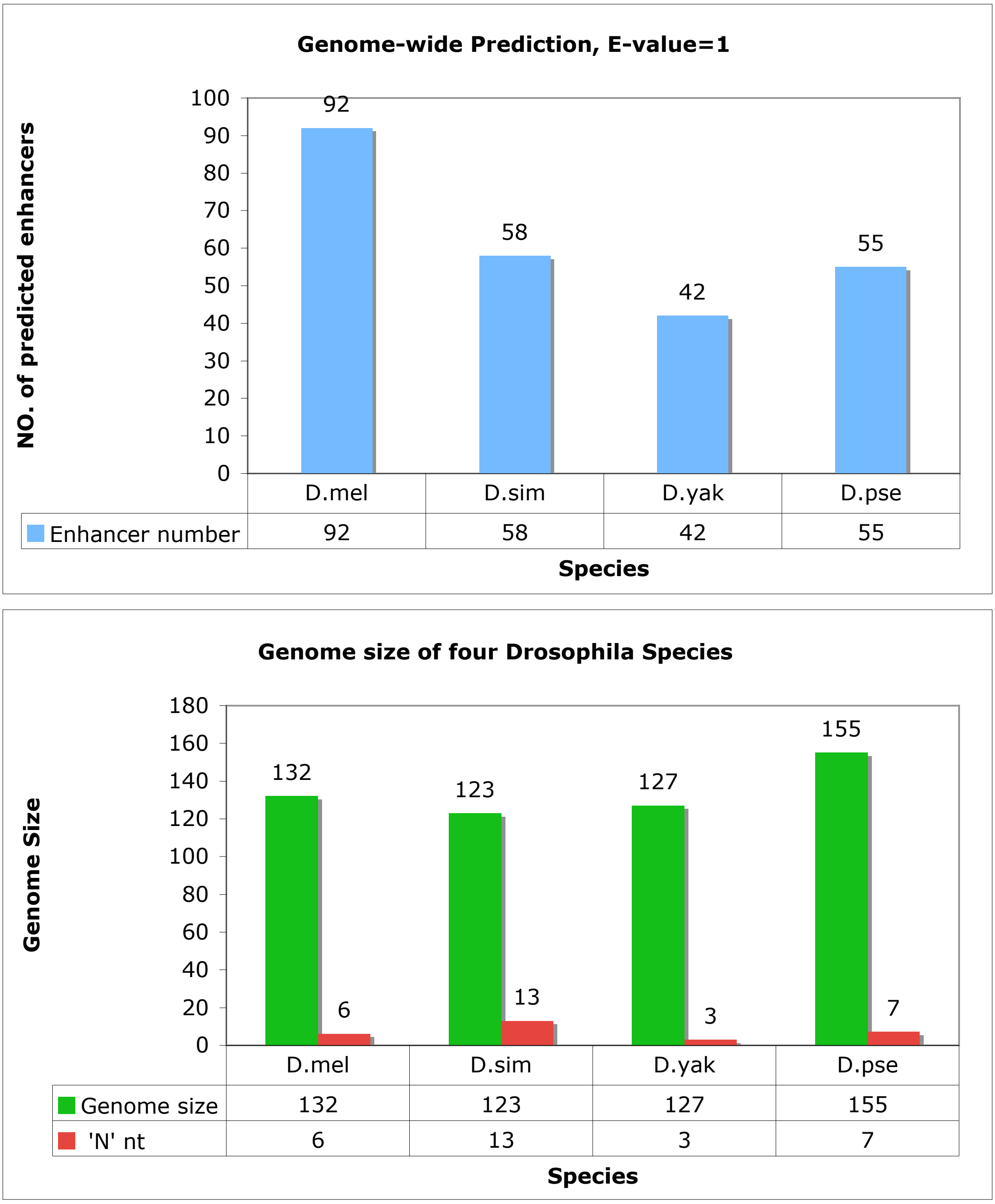


Figure 2.18.: The number of predicted enhancers and genome size are various in different species. The cutoff is 50, which corresponds to a genome-wide E-value of 1 in *D. melanogaster*. Parameters were set as: window size 700bp, paired motif distance 150bp, and window shift in step of 10bp.

distances between the original best BLAST regions and the nearest comparatively predicted loci (Figure 2.12), it is clear that analogous enhancers in different *Drosophila* species do not necessarily include the conservation of position. Sometimes the distances from the BLAST hit to the functionally analogous enhancer are closer than 1kb, sometimes they could be further than 10kb. Therefore, enhancers are categorized by chosen distances, so that the degree and popularity of this enhancer evolutionary plasticity can be verified.

First category

The first category of enhancer position plasticity is the case where the BLAST distance is smaller than 1kb (see Figure 2.12). In total, 50 enhancers remain well conserved in *D. melanogaster* and *D. pseudoobscura*. 20 of them are the predictions from *D. pseudoobscura* to *D. melanogaster*, 30 enhancers are the predictions from *D. melanogaster* to *D. pseudoobscura*. Some of these enhancers are well-known ones, such as the enhancers *odd* (ID:19111), *eve* (IDs:19126,19127), *kni* (ID:19154), *run* (ID:19200), *ftz* (ID:19157), *gt* (ID:19169), *hb* (ID:19321) and *hairy* (ID:19146). Some of the enhancers' neighboring genes have an annotation related to embryonic development, such as *SoxN* (ID:19113) which has an annotation of embryonic nervous system development, or *Dl* (ID:19162) with regulation of developmental process. The list of predicted enhancers is shown in Table A.1. Twenty-three of these well-conserved enhancers have been experimentally verified [67, 2, 111]; nineteen of enhancers' TFBSs have been verified by [24]; five predictions previously have not been experimentally verified both at the enhancer level and at TFBS level, but they all have neighboring genes with embryonic-development related annotation; the last three predictions are new, they have never been reported in any literature and their TFBSs have not been verified.

In the first category of enhancer position plasticity, enhancers have not moved far away from the orthologous locus. There is a high level of sequence conservation among the predictions. In fact, most well-known enhancers are likely from this category, because previous prediction methods have mainly been based on clustering of TFBS and conservation of non-coding sequences [22, 1, 67, 2].

A first detailed example of this category is the striped expression pattern of the pair-rule gene *even-skipped* which is established by five stripe-specific enhancers. Each enhancer responds in a unique way to gradients of positional information in the early *Drosophila* embryo [36]. One of the prominent example enhancers is *even-skipped* stripe 2. The experimental results from [95] show that *even-skipped* stripe 2 expression is functionally conserved to a remarkable degree in *D. melanogaster*, *D. erecta*, *D. yakuba*, *D. pseudoobscura*. This conclusion is fully consistent with my score plot among four species (see Figure 2.19). In *D. melanogaster* and *D. simulans*, the score values of enhancers are all above the genome-wide cutoff. However, in *D. yakuba* and *D. pseudoobscura*, the enhancer can only be found once dynamic search is applied. The evolutionary divergence of the *even-skipped* stripe 2 enhancers has no discernible effect on either the timing or spatial localization of stripe 2 expression [95]. Enhancers belonging to this first category are considered functionally conserved because of their position/sequence conservation.

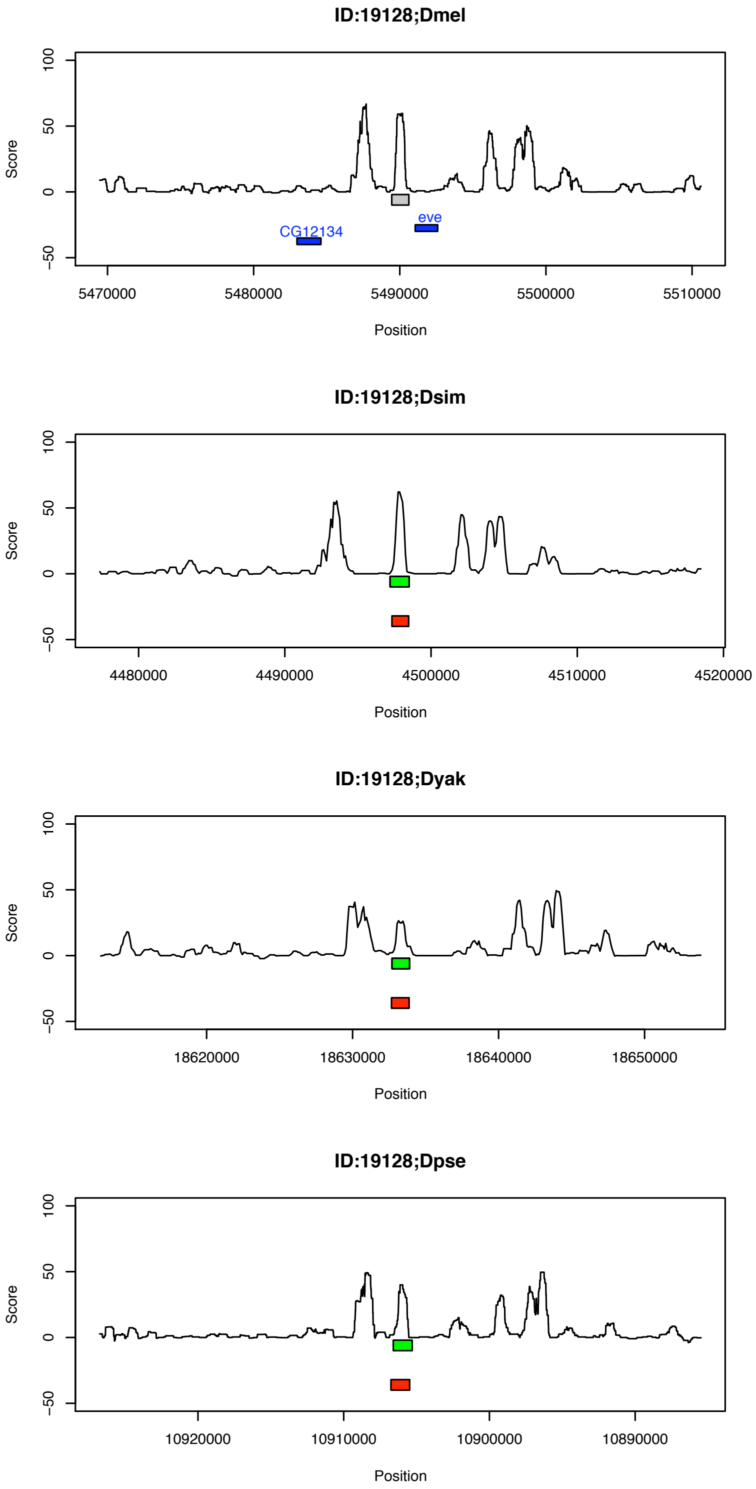


Figure 2.19.: Score plots of *even-skipped* stripe 2 enhancer in four species. Enhancer *eve* stripe 2 belonging to the first category are considered functionally conserved because of their position conservation. The enhancer *eve* stripe 2 in *D. melanogaster* is shown in grey. The BLAST hits from source species *D. melanogaster* to the other three species are shown in red. The dynamically predicted enhancers are drawn in green. The blue bars denote positions of neighboring genes. X-axis refers to sequence positions. Y-axis shows score values.

Second category

Some statically predicted enhancers have orthologous sites in the target species, however, these orthologous sites have no score peaks or the scores are not high enough to be identified as dynamic predictions. The actual functionally analogous sites move away from the orthologous sites in the target species. If the distance is less than 10kb, I speculate dynamically predicted enhancers are not conserved with the source static enhancer, but are still functionally analogous to the original enhancer. These enhancers which are not position conserved but locus conserved are grouped in the second category.

In total, I found 8 enhancers in the dynamic prediction between *D. melanogaster* and *D. pseudoobscura* in both prediction directions. Six enhancers are highly bound by multiple factors - BCD, CAD, HB, KR, KNI [24]; three of these six enhancers have embryonic genes close by; the last two enhancers have not been predicted or characterized before.

Enhancer 19190 (X;12898030-12898899) illustrates the second category of enhancer plasticity analysis. Figure 2.20 shows score plots of predicted enhancers in four species. In *D. melanogaster*, the scores reach 56 and are above the genome-wide cutoff. In closely related species, such as *D. simulans* and *D. yakuba*, the BLAST hits overlap well with the dynamically predicted enhancers, and the scores are significant enough for dynamic prediction. But, once I look at the evolutionary more divergent species *D. pseudoobscura*, the predicted enhancer at least moved 3500bp away from the original BLAST hit. Li *et al.* [24] has experimentally verified that this enhancer is highly bound by embryonic factors in *D. melanogaster*, but more experiments are needed to evaluate wether the dynamically predicted enhancers in *D. pseudoobscura* is functionally analogous to the enhancers in the melanogaster subgroup.

Enhancer 19115 (2L;12660910-12661999) is another example showing that enhancers in close but different positions in different species regulate the same gene *pdm2*. In *D. melanogaster*, an enhancer is predicted by the genome-wide method, and located at the beginning of the gene *pdm2* intron. In all *D. melanogaster* subgroup species, the scores are above the genome-wide cutoff of 50. By dynamic search, a significant score peak with value of 43 is identified to be an enhancer in *D. pseudoobscura*. This enhancer has moved away from the BLAST homologous locus for about 10kb, but is still inside *pdm2*. Gene *pdm2* is very long; about 28kb. I speculate that in *D. melanogaster*, *D. simulans* and *D. yakuba*, the enhancer which regulates *pdm2* lies inside the promoter region of *pdm2*. The functionally analogous enhancer in *D. pseudoobscura* locates in the downstream region of the gene. It is quite likely that the enhancers from this category are functionally conserved without position conservation. Lab experiments will be carried out to verify this type of plasticity.

Third category

Once the dynamic prediction is too far away from the original site, it might be an entirely different enhancer. The third category are for enhancers which have moved away from the BLAST locus more than 10kb; I assume they are no longer functionally analogous in target species. There are 48 enhancers who belong to this category.

Enhancer position plasticity

As a conclusion, some enhancers such as the ones from the first category (*eve, ftz*) are positionally consistent during evolution. Others such as the enhancers from the second category

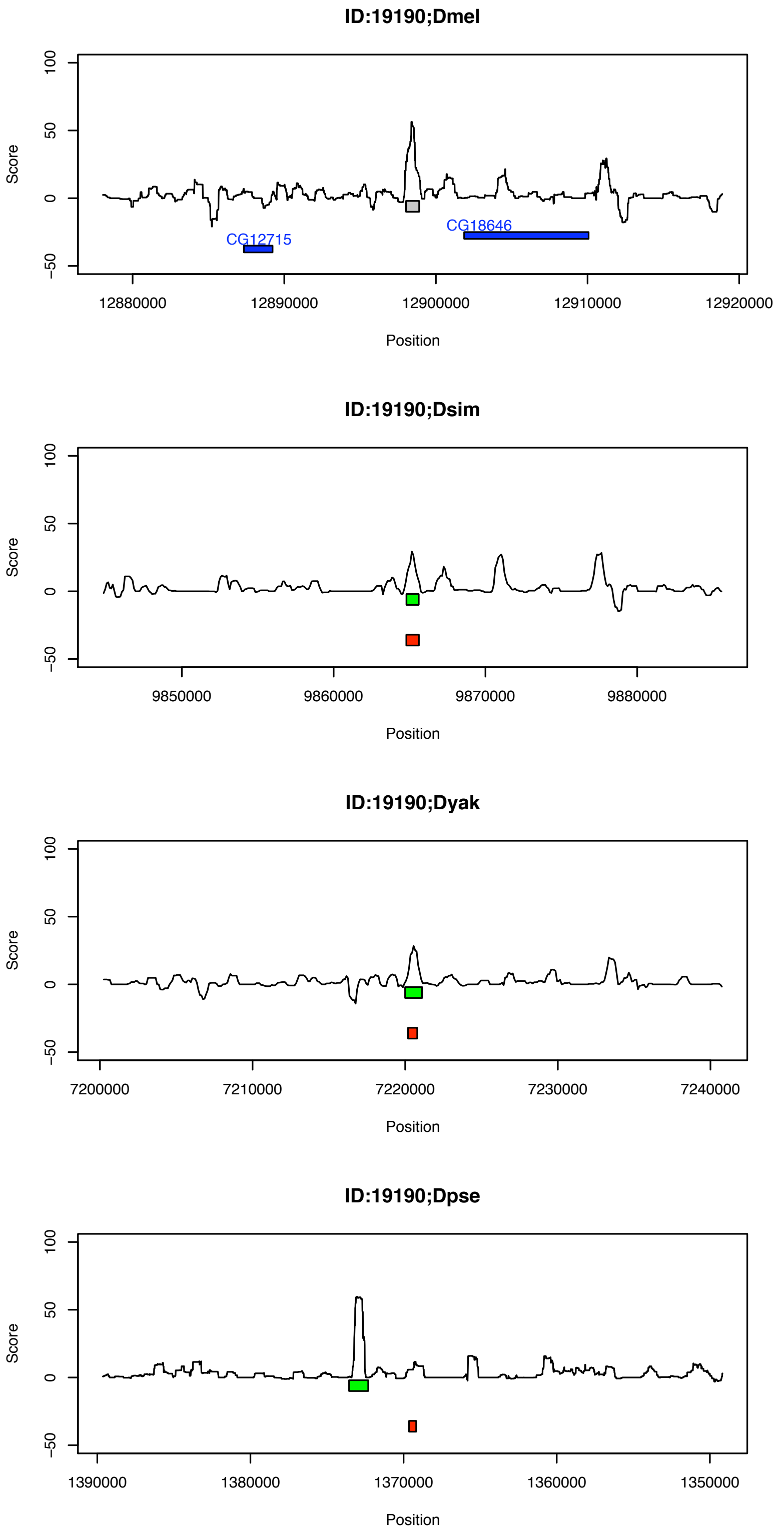


Figure 2.20.: Score plots of enhancer 19190. The enhancer region in *D. melanogaster* is shown in grey. Enhancer 19190 belongs to the second category, that is considered functionally conserved without position conservation. The BLAST hits from source species *D. melanogaster* to the other three species are shown in red. The dynamically predicted enhancers are drawn in green. The blue bars denote positions of neighboring genes. X-axis refers to sequence positions. Y-axis shows score values.

are functionally conserved and independent of sequence conservation. The observation that the positions of embryonic enhancers are independent of sequence conservation is defined as the second type of enhancer plasticity. This is the first time, that the functionally analogous enhancers are able to be reported in multiple *Drosophila* species in the whole genome range. I suggest this feature of position plasticity is not unique for embryonic enhancers [96]. Additionally, the dynamic search approach is unbiased in terms of evolutionary conservation, and has the power of identifying both conserved and non-conserved regulatory elements.

2.5.3. Third type of plasticity - motif turnover

From the analysis of the second type of plasticity, some of the enhancers are observed to be conserved among all species, as the distances between the BLAST locus and nearest prediction are less than 1kb. Initially, I speculated that the full conservation of individual motifs and complete occurrences of every motif might be the reason for the enhancer full presence in all *Drosophila* species. The enhancers which have been predicted in *D. melanogaster* as well as their dynamic prediction in other species are selected as the candidates to verify this speculation.

In the full presence enhancer list (see Table A.1 to A.3), the highest score of the enhancer from source species *D. melanogaster* is 105.714 (3R:2693200-2694579, *ftz*). The one with the lowest score is 50.211 (2R:5498330-5499059, *eve*). The significant difference of enhancer scores suggests motif weights or motif density are various among enhancers. Moreover, I compared the score plot of the *ftz* enhancers in four *Drosophila* species; their distributions are quite various (see Figure 2.21). Enhancer *ftz* has the highest score value of 105.714 only in *D. melanogaster*, the score value drops to be 60.3435 in *D. pseudoobscura*. Although the sequences of *ftz* enhancers are highly conserved (see Figure 2.22) , it does not necessarily imply that motifs inside enhancers have been evolutionarily static.

The explanation could be that motifs rearranged in the homologous enhancers during evolution, such as some motifs which exist in one *Drosophila* species, but not in the other species, or motifs located in different position/order in homologous enhancers across species. I define these rearrangements as motif turnover, also known as the third type of plasticity [96].

Several examples of *Drosophila* regulatory sequence conservation over long evolutionary distances have previoulsy been reported on *eve* stripe 2 [95, 33, 42, 110, 113]. Thus, *even-skipped* stripe 2 regulation once again becomes the preeminent model for the study of binding site turnover. In [42] experimental tests were performed on the *eve* stripe 2 region in *D. melanogaster*, *D. erecta*, *D. yakuba*, *D. pseudoobscura*. After comparison of TFBS occurrences in Ludwig *et al.* and my prediction, I found out Ludwig *et al.* selected the motif occurrences with prior knowledge, because by the threshold he declared in his paper, more motifs should be present in his research. A further analysis of threshold is presented on page 64.

To gain a general view, I made analysis without the influence of prior knowledge, and was able to get a general view of motif composition and organization of the *eve* stripe 2 enhancer in four *Drosophila* species.

First, in my analysis I have made a comparison of the score plots of four *eve* stripe 2 enhancers regions. Certainly, the tendency of score value is a direct expression of density of motifs. The higher the score, the more motifs (see Figure 2.19).

Next, a multiple sequence alignment of these regulatory regions was drawn by MLAGAN [92] (see Figure 2.23) and the sequence conservation was included into Figure 2.24. The

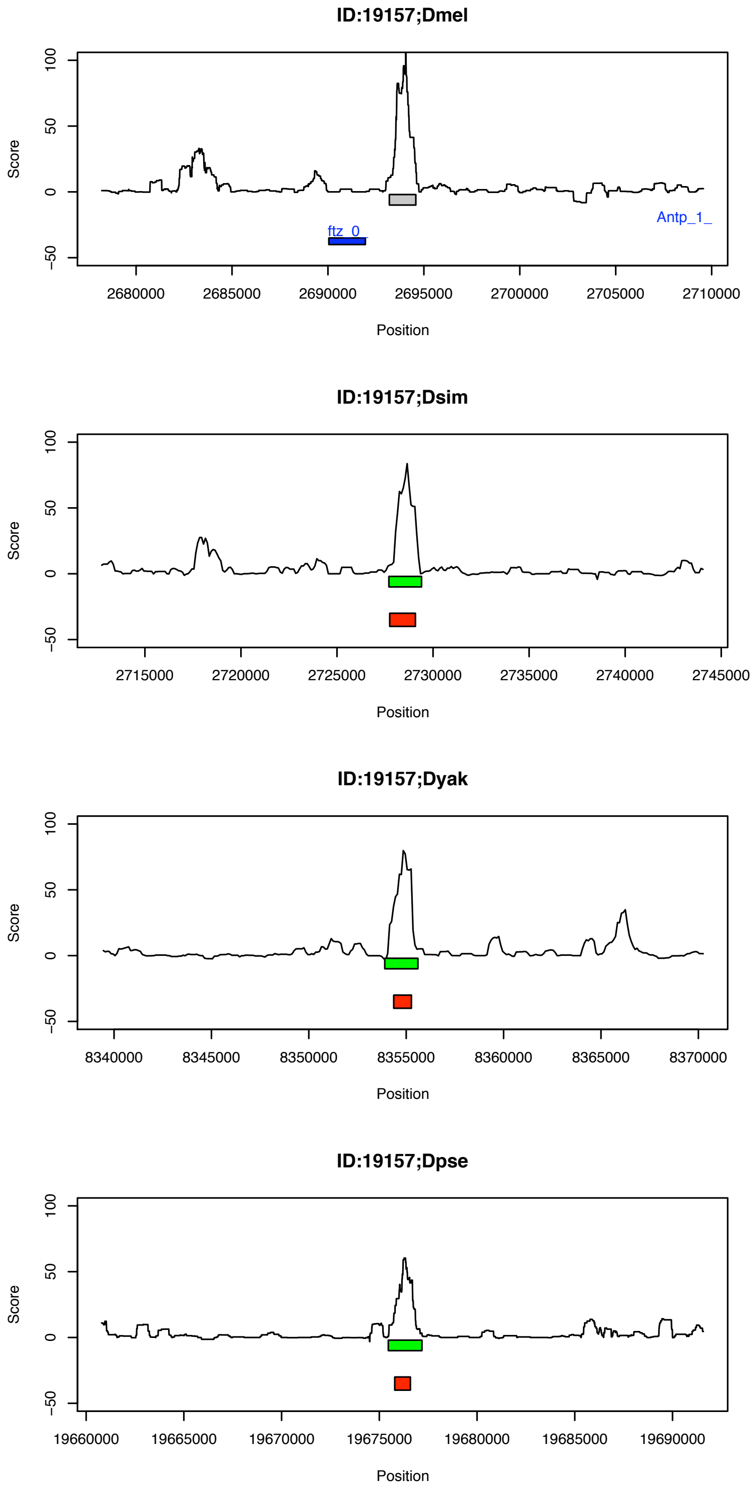


Figure 2.21.: Score plots of enhancer *ftz*. The enhancer region in *D. melanogaster* is shown in grey. Score distributions of enhancer *ftz* in different species are various, which imply motifs inside enhancer have been evolutionarily dynamic. The BLAST hits from source species *D. melanogaster* to the other three species are shown in red. The dynamically predicted enhancers are drawn in green. The blue bars denote positions of neighboring genes. X-axis refers to sequence positions. Y-axis shows score values.

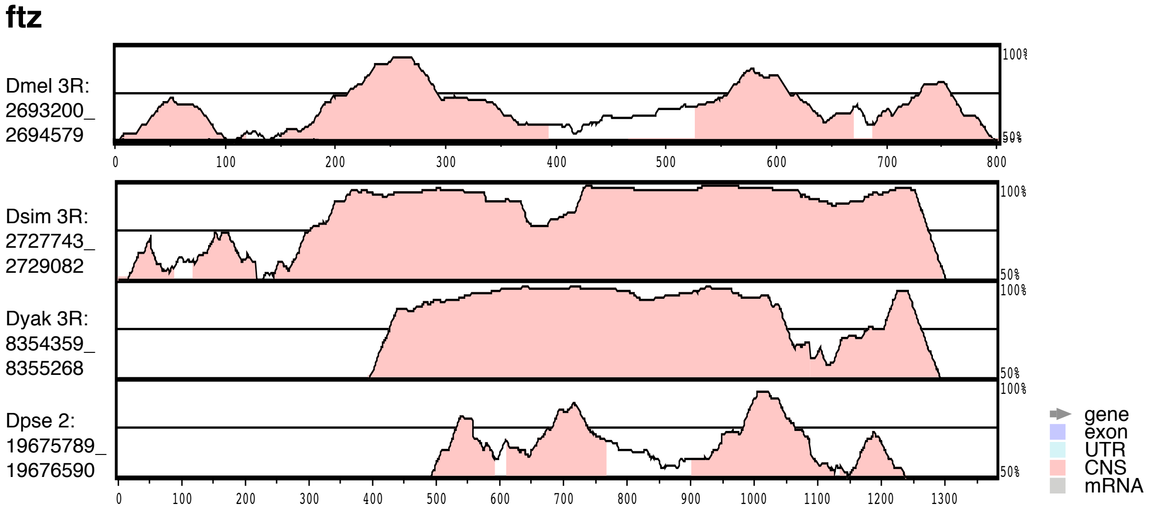


Figure 2.22.: Conservation plot of enhancer *ftz* with multiple alignments among *D. melanogaster* and the other three species *D. simulans*, *D. yakuba*, *D. pseudoobscura*. Regions colored in pink have a conservation percentage >70%; regions in white have a conservation percentage in the range 50%-70%. X-axis represents *D. melanogaster* sequence positions. Y-axis represents conservation percentage. The alignment program is mVISTA-LAGAN [112].

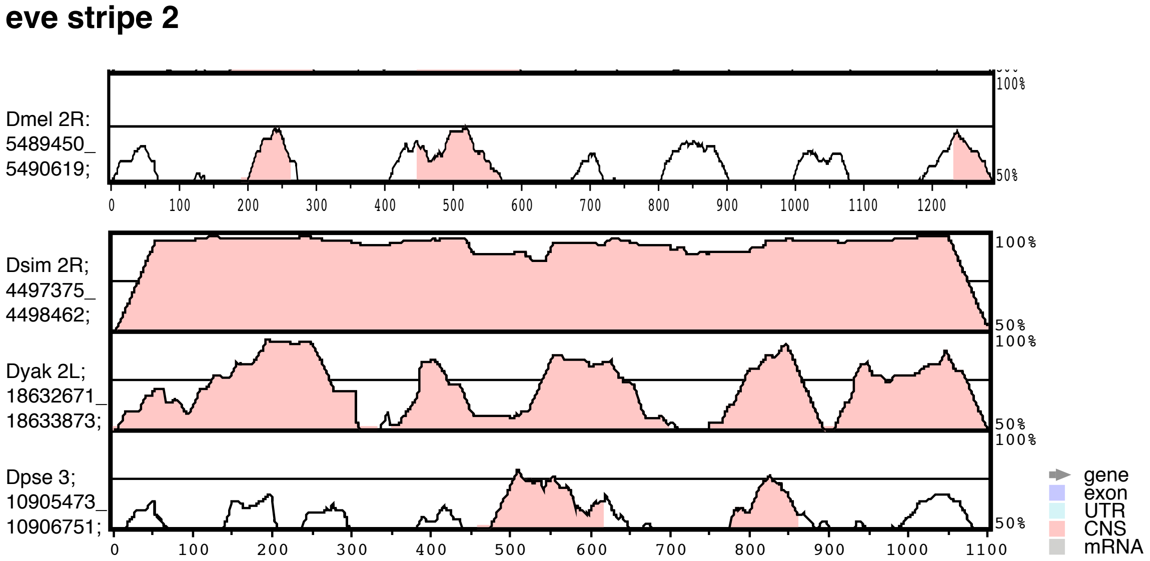


Figure 2.23.: Conservation plot of enhancer *eve* stripe 2 with multiple alignments among *D. melanogaster* and the other three species *D. simulans*, *D. yakuba*, *D. pseudoobscura*. Regions colored in pink have a conservation percentage >70%; regions in white have a conservation percentage in the range 50%-70%. X-axis represents *D. melanogaster* sequence positions. Y-axis represents conservation percentage. The alignment program is mVISTA-LAGAN [112].

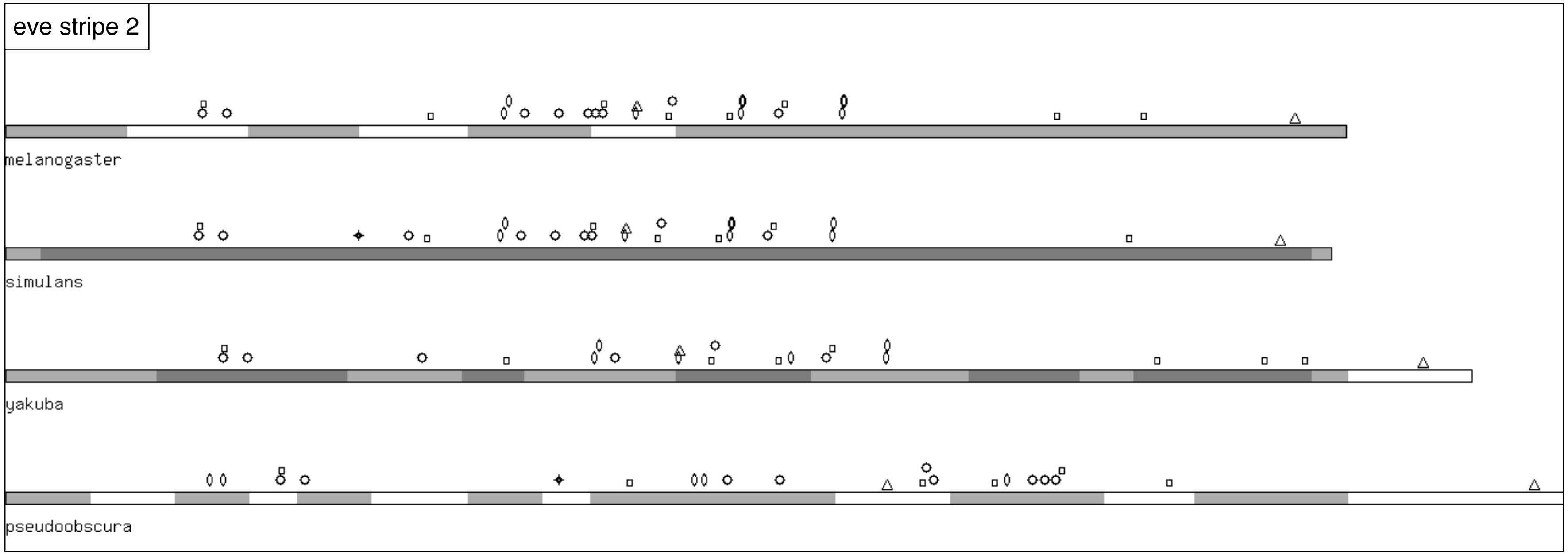


Figure 2.24.: Motif turnover in enhancer *eve* stripe 2 region. The length of the bars show sequence region in *D. melanogaster* and orthologous regions in *D. simulans*, *D. yakuba*, *D. pseudoobscura*. Sequence regions shaded grey represent percentage of pair-wise conservation. Dark grey: high conservation (>70%); Light grey: medium conservation(50%-70%); White: low conservation(<50%). Motif positions are indicated above each bar. The five motifs are Bicoid (circle), Hunchback (oval), Kruppel (square), Knirps (star) and Caudal (triangle).

predicted sequences are displayed as bars, the length of the bar indicates the relative length in four species in Figure 2.24. The information of sequences conservation is presented by the deep/light grey bars. The enhancer regions were classified by level of conservation: high conservation (>70%), medium conservation (50%-70%) and low conservation (<50%). Because of the close evolutionary distance among *D. melanogaster*, *D. simulans* and *D. yakuba*, the majority of enhancer sequences are conserved above 50% in these species, indicated by grey bars. However, for further species such as *D. pseudoobscura*, the sequence conservation gets poorer; very few pieces of sequence have a conservation level above 50%. The grey bars in *D. melanogaster* and *D. pseudoobscura* are drawn by pair-wise alignment between these two species.

The occurrences of the TFBS of each motif are represented above the enhancer bars. Except for the first KNI (star), the remaining motifs in *D. simulans* are conserved with *D. melanogaster*. There is still a large number of binding sites conserved between *D. melanogaster* and *D. yakuba*, despite the fact that the third BCD (circle) might get a bit far away from the second KR (square), plus the absence of some BCD (circle) and KR (square) before the position of CAD (triangle). For the further species, motif turnover is rather rapid between *D. melanogaster* and *D. pseudoobscura*. The enhancer starts with two HB (oval), which never occur in *D. melanogaster*, *D. simulans* and *D. yakuba*. Some motifs originally found in *D. melanogaster* can not be found in *D. pseudoobscura*. The distance between adjacent motifs is also different. The complete alignment and motif occurrences of enhancer *eve* strip 2 is shown in Figure A.2.

Another example of motif turnover is the enhancer which regulates the gene *ftz* (see Figure 2.25). If I subdivide the enhancers in *D. melanogaster* into three parts, the second and third parts of motifs are very well conserved, the motifs in the first part are halfway conserved in *D. melanogaster* and *D. simulans*. However, in *D. yakuba*, the complete first part of motifs seem to be removed. More motif turnover exists in the first part of *D. pseudoobscura* enhancer. The complete alignment and motif occurrences of enhancer *ftz* is shown in appendix A.3.

In the example of *eve* stripe 3/7, score plots were drawn in four species (see Figure 2.26). Enhancers are highly conserved in all species, except two peaks which appear in the enhancer region of *D. yakuba*. The motif distance in *D. yakuba* might be larger than in *D. melanogaster*, *D. simulans* and *D. pseudoobscura*. Thus, one enhancer has been split into two parts, each part being about 700bp. One explanation could be stripe 3 and stripe 7 have evolved to be regulated by two short enhancers in *D. yakuba*, however, in *D. melanogaster*, *D. simulans* and *D. pseudoobscura*, one long enhancer is sufficient for two stripe regulation. This speculation needs further proof.

A similar analysis has been repeated on the remaining fully conserved enhancers. Strikingly, motif turnover is quite common for embryonic enhancers in *Drosophila* species. Sometimes motif turnover is much more significant in some of the enhancers than the others, such as enhancers that regulate *ftz* compared with enhancers of *eve*. The activity of motif turnover is within a certain degree, if only compared within *D. melanogaster* sister texa, there might be not much turnover. However, once further species are introduced, motif turnover could be much more rapid [113]. The degree of motif turnover of the same enhancer is various depending on the evolutionary distance of species.

Previous studies have demonstrated that appropriate regulation of the *even-skipped* stripe enhancers relies on the close proximity of multiple binding sites for both activators and re-

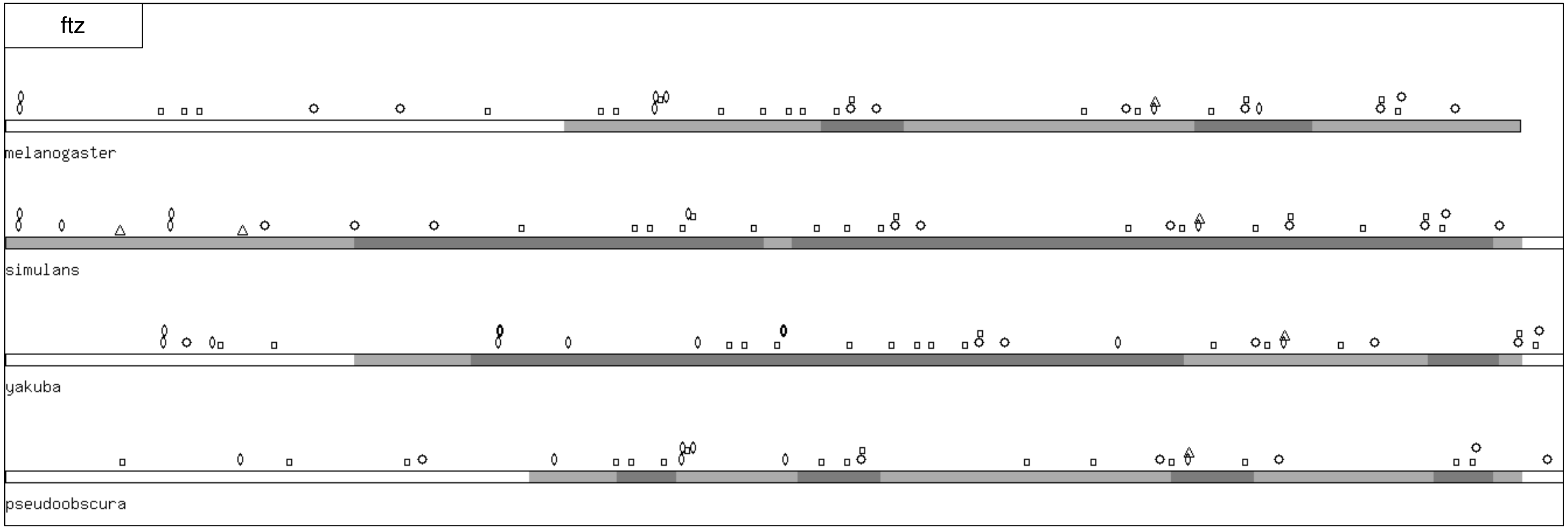


Figure 2.25.: Motif turnover in enhancer *ftz* region. The length of the bars show sequence region in *D. melanogaster* and orthologous regions in *D. simulans*, *D. yakuba*, *D. pseudoobscura*. Sequence regions shaded grey represent percentage of pair-wise conservation. Dark grey: high conservation (>70%); Light grey: medium conservation(50%-70%); White: low conservation(<50%). Motif positions are indicated above each bar. The five motifs are Bicoid (circle), Hunchback (oval), Kruppel (square), Knirps (star) and Caudal (triangle).

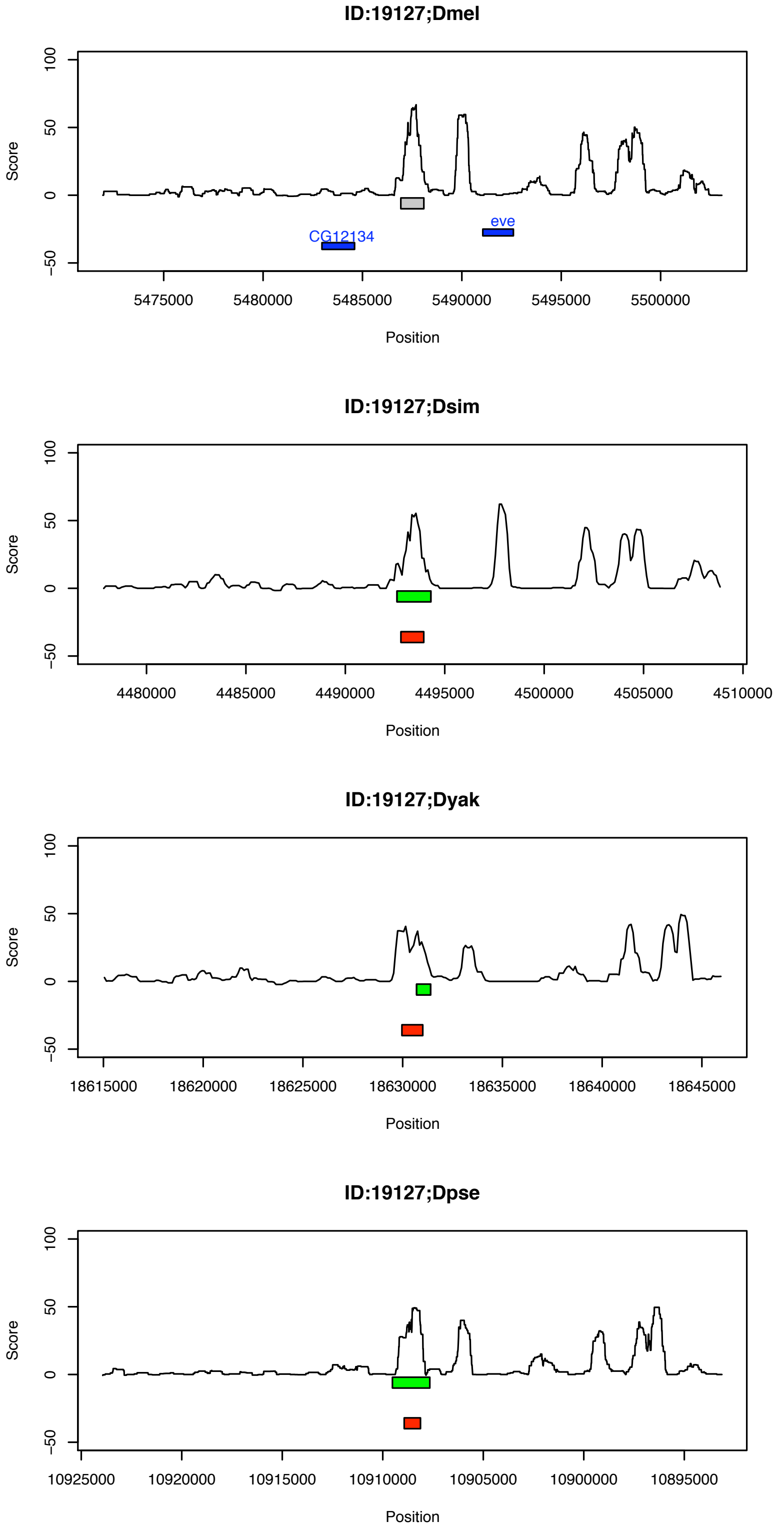


Figure 2.26.: Score plots of enhancer stripe 3/7. The enhancer region in *D. melanogaster* is shown in grey. Enhancer 19190 belongs to the second category, which is considered functionally conserved without position conservation. The BLAST hits from source species *D. melanogaster* to the other three species are shown in red. The dynamically predicted enhancers are drawn in green. The blue bars denote positions of neighboring genes. X-axis refers to sequence positions. Y-axis shows score values.

pressors [113, 34, 35]. I try to further explore the motif turnover activity by checking adjacency of these motifs. To do so, the binding sites are classified into three types based on their proximity to the neighboring binding sites: first, overlapping sites that share one or more nucleotides with another binding site; second, close sites that are within 10 nucleotides of another site but do not overlap, and the remaining binding sites which are isolated sites [113]. The conservation of each type of sites is checked in all four species. From the examples of enhancer *eve* stripe 2, stripe 3/7 and *ftz* in Figure 2.27, overlapping paired motifs are more extremely conserved in four species, close paired motifs are still conserved in *D. melanogaster* subgroup species and isolated motifs are often minimally (present in *D. melanogaster* and *D. simulans*) or non-conserved (only present in *D. melanogaster*).

Motif turnover from overlapping and close binding sites might be more difficult than that from isolated sites. The explanation could be found in from the annotation of motifs. HB, CAD and BCD are activators [114], KR and KNI are repressors. The development of the embryo requires very precise regulation from these motifs. Once they are adjacently paired together, it is much harder to influence close sites by mutation than isolated sites, because they have to perform regulation cooperatively by binding both TFs.

Similar analysis of motif proximity has been discussed by [113] with extreme motif conservation of both *Drosophila* and *Sepsid*. Although my definition of extremely conserved paired motifs is slightly different from [113] - I only compare four *Drosophila* species, the conclusion of motif proximity is the same. Overlap and close motifs are more conserved than isolated motifs during evolution.

Furthermore, I applied the same method on other enhancers such as *ftz* and *eve* stripe 3/7. On the one hand, in the example of *ftz*, overlapping sites are still greatly conserved in all species, as shown before in Figure 2.25. But the number of isolated sites is much larger than in the example of *eve* stripe 2. On the other hand, in the example of *even-skipped* stripe 3/7 region, the number of overlapping and close sites are significantly larger in evolutionary close species.

Previously, [115] did motif turnover analysis only on single zest binding site. Ludwig *et al.* [95, 33, 42] showed in a series of papers that the *eve* stripe 2 enhancer in *Drosophila* species drives a stripe 2 pattern in transgenic *D. melanogaster* embryos despite the imperfect conservation of multiple functional binding sites [95, 33, 42, 110, 113]. Hare *et al.* [113] examined the pioneering work on binding site turnover of *eve* stripe 2 between *Drosophila* and the even further evolved *Sepsid*. My work of motif turnover discovery firstly extends the number of motifs by comparing with [115]; secondly it extends the range of enhancers genome-wide. The large number of sites is helpful to detect relationships between the spacing of paired motifs and conservation for embryonic factors. Moreover, my analysis crosses an evolutionary distance of up to 50MB years, which could possibly be extended to a comparison of even further evolutionary species, and up to the complete set of 12 *Drosophila* species.

2.6. Overall prediction number and distribution among prediction methods

Overall, I have tried three methods for enhancer prediction (see Table 2.8). The final pre-

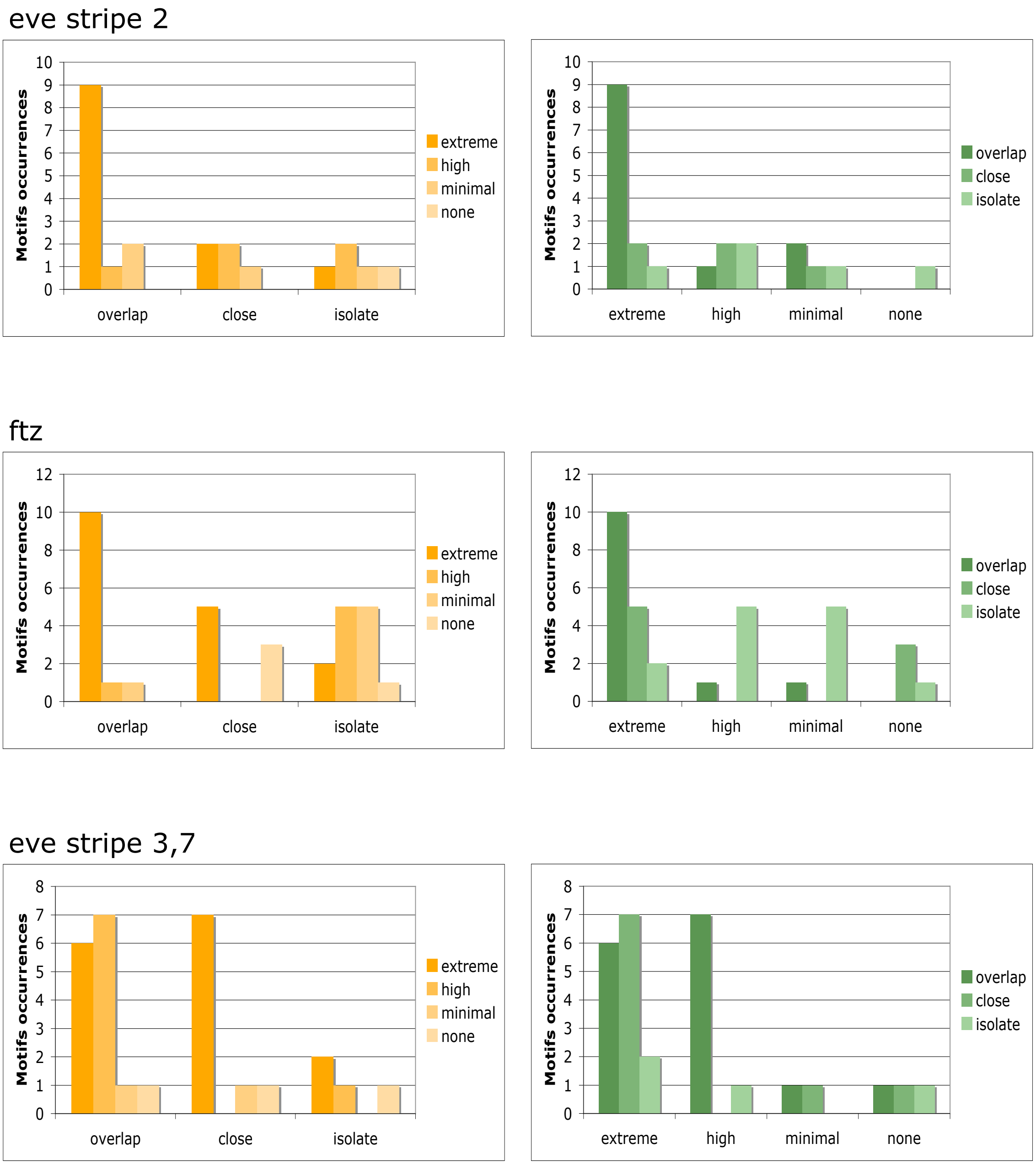


Figure 2.27.: Motif proximity influences motif evolution. The conservation of binding sites was checked in four species with the examples of enhancer *eve* stripe 2, stripe 3/7 and *ftz*. Overlapping(dark green) paired motifs are more extremely (dark orange) conserved in four species, close paired (medium dark green) motifs are conserved in *D. melanogaster* subgroup species (medium dark orange) and isolated (light green) motifs are often minimally or non-conserved (light orange).

Table 2.8.: Overall prediction methods

Method	Scale	Genome	Cutoff
Genome-wide static search	whole genome	single	high
GO subset	subset	single	low
Dynamic	whole genome	multiple	low

diction result is based on the combined results of all three methods. Since *D. melanogaster* is the model species, the final results are mainly generated for *D. melanogaster*. The overall prediction is the result of a genome-wide static method, the result of GO subset and the result of dynamic search from source species *D. pseudoobscura* to *D. melanogaster*. *D. pseudoobscura* was chosen for the dynamic analysis because it appeared to have the appropriate degree of sequence divergence from *D. melanogaster* in order to locate more embryonic enhancers. The final number is generated by every unique prediction which means the predicted enhancer has no single base pair overlap with the others (see Figure 2.28). The unique prediction from the genome-wide method is 56 (light purple), from the GO embryonic subset method is 28 (light yellow) and from dynamic search is 13 (light blue). By using static subset and dynamic search, I made a new prediction and found additional enhancers previously undiscovered by the static method. The common prediction between genome-wide and subset is 16, between subset and dynamic search is 2 and between genome-wide and dynamic is 9. The common prediction for all three methods is 11 (white). Summarizing all the prediction results gives us the final embryonic enhancer amount of 135 in *D. melanogaster*.

2.7. Prediction evaluation

2.7.1. Motif threshold evaluation - comparison with Ludwig and Berman

Ludwig *et al.* and Berman *et al.* [33, 95, 42, 22, 1] have used the same motifs for the study of embryo regulatory networks and both of them have experimentally verified the function of motifs in *even-skipped* stripe 2 locus. To compare the motif occurrence in my work with their published results of *eve* stripe 2 enhancer, it is helpful to evaluate my method of the motif threshold selection.

Berman *et al.*, Ludwig *et al.* and I share the same alignment matrix of motif discovery. The common motifs used in all three researches are BCD, KR, HB (see Figure 2.29). The formula used for transforming the alignment matrix to a weight matrix is shown below in equation 2.1 on section 3.1.1.

Berman *et al.*, Ludwig *et al.* perform the binding-site prediction by using Patser [117], which is able to select a threshold automatically on the basis of the information content of the matrix and the sequence size. Ludwig *et al.* set thresholds to be –ln(P) = –6.1 for Bicoid, ln(P) = –8.06 for Hunchback, and ln(P) = –6.65 for Kruppel, to recover all the known binding sites [42]. Berman *et al.* set thresholds to be –ln(P) = –6.33 for Bicoid, ln(P) = –8.19 for Hunchback, and ln(P) = –8.7 for Kruppel [1]. Berman's thresholds are more stringent than Ludwig *et al.*, which means less binding sites will be discovered.

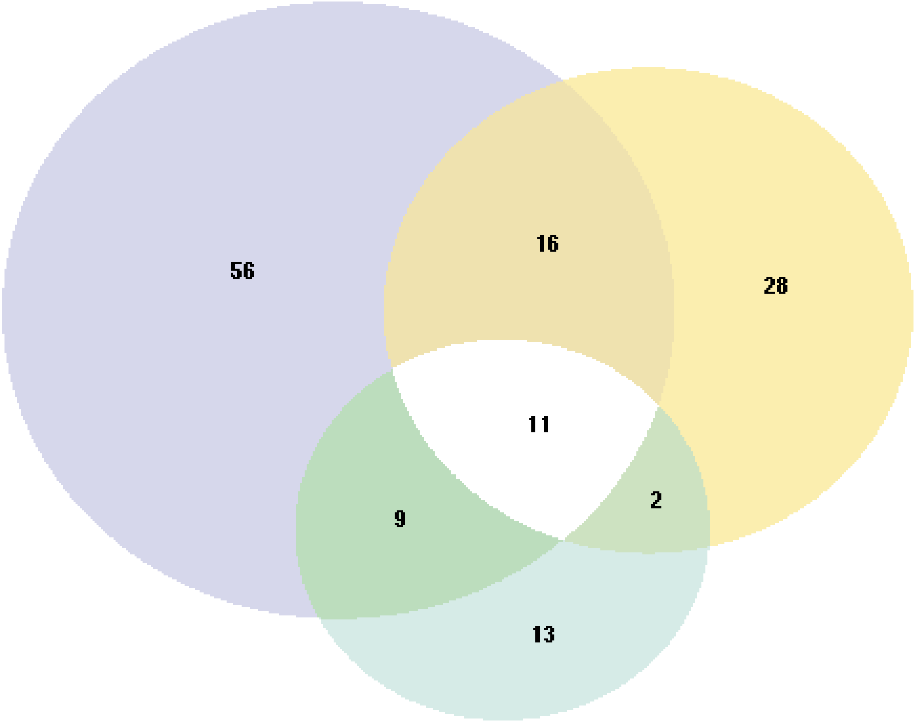


Figure 2.28.: Venn plot of prediction result by genome-wide, subset and dynamic search. The unique prediction from genome-wide method is 56 (light purple), from subset method is 28 (light yellow) and from dynamic search is 13 (light blue). The common prediction between genome-wide and subset is 16, between subset and dynamic search is 2 and between genome-wide and dynamic is 9. The common prediction for all three methods is 11 (white). Summarizing all the prediction results gives us the final embryonic enhancer amount of 135 in *D. melanogaster*.

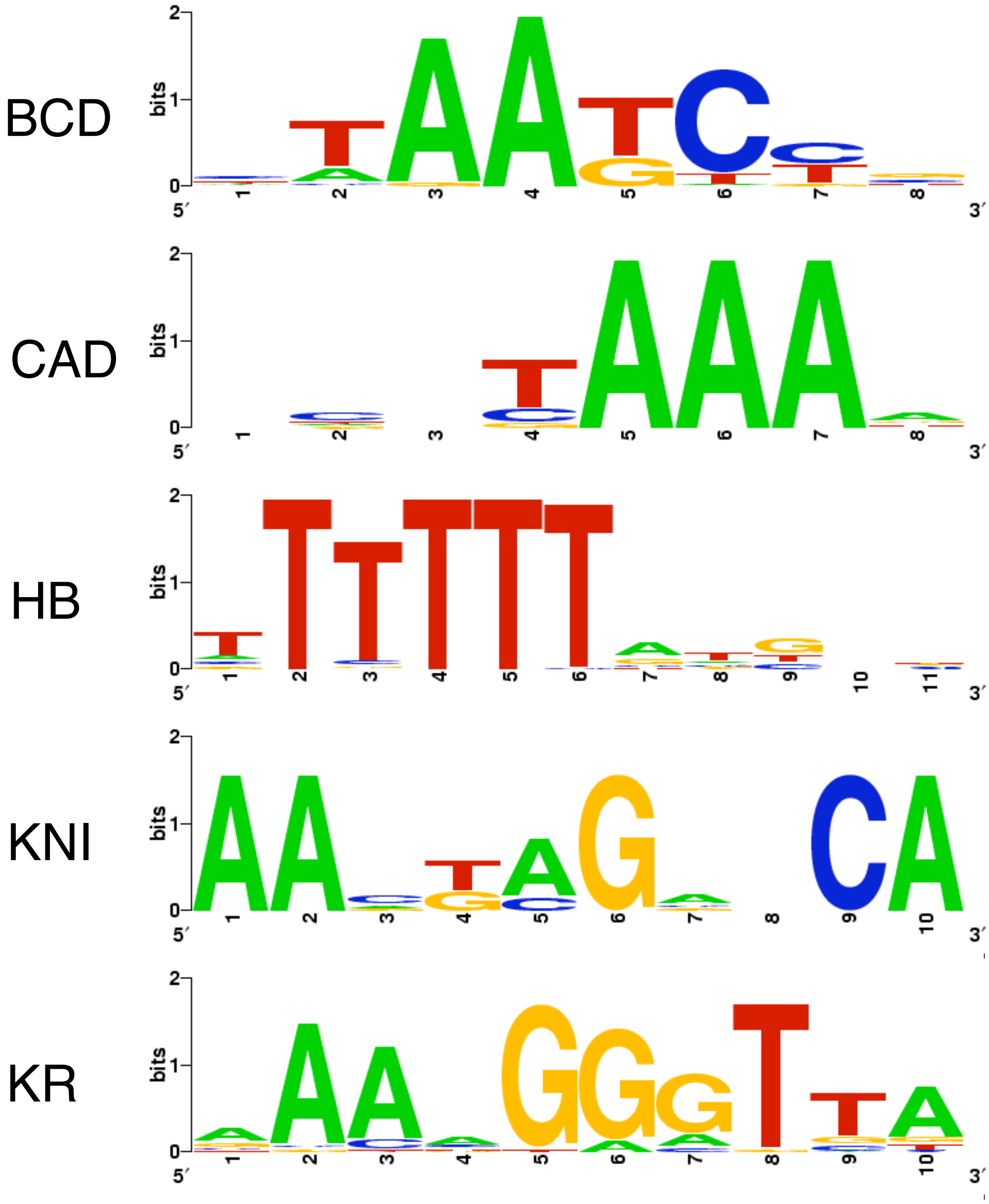


Figure 2.29.: Motif logos. Motif alignment matrices were downloaded from Berman [1][1]. Motif logos were constructed with Weblogo [116].

I use a different program to search for motifs. To transform Position Frequency Matrix (PFM) to Position Weight Matrix (PWM), motif weight calculation by jPREdictor is performed according to formulas shown in Section 3.1.1. As a comparison, the calculation of Patser is shown:

$$P_{Patser}(b,i) = \frac{n_{b,i} + s(b)}{N + \sum\limits_{b' \varepsilon \{A,C,G,T\}} s(b')} \tag{2.1}$$

$P_{Patser}(b,i)$: the corrected probability of base b in position i; $n_{b,i}$: letter b is observed at position i of this alignment; N : the total number of sequences; $s(b)$: the a-priori probability of the letter b ; $\sum\limits_{b' \varepsilon \{A,C,G,T\}} s(b') = 1$.

$$W_{Patser}(b,i) = ln \frac{P_{Patser}(b,i)}{p(b)} \tag{2.2}$$

$W_{Patser}(b,i)$ is the PWM value of base b in position i; $p(b)$ is the background probability of b.

During the calculation of PWM, on average the position weight matrix score value by Patser is slightly different than jPREdictor's calculation. Since the difference is so minor and the principle of PWM calculation is the same in both programs, the influence on PWM calculation can be ignored.

Although the program I used is not the same as Ludwig *et al.* and Berman *et al.*, the same binding sites should possibly be recovered in all three research approaches. I have made a comparison of my TFBS with Berman and Ludwig's to judge the correctness of my PWM thresholds.

I set PWM score thresholds to be 5.1 for BCD, 4.9 for KR, 4.4 for HB. Almost every motif from Ludwig *et al.* can be recovered by jPREdictor, except KR-6 (ATAACCCAAT), BCD-3 (TATAATCGC) . In order to find out the reason for these two missing motifs, I repeated binding site discovery with the same parameters setting as Ludwig *et al.* and Berman *et al.* described in the publication by Patser, as can be seen from Table 2.9.

The ancestral *even-skipped* stripe 2 enhancer lacking the binding site BCD-3 (TATAATCGC) would not properly activate stripe 2 expression in *D. melanogaster* [34]. Since the authors made experimental tests, and they know this binding site is a very important functional element, they permit such a high threshold -6.10 for BCD-3 (TATAATCGC).

Once I lower my motif weight threshold of BCD from 4.4 to 3.8, I am also able to recover BCD-3 (TATAATCGC) in *even-skipped* stripe 2 region. But, because I carry out a mass search in the whole genome, and I do not have prior knowledge of binding sites threshold selection, I have to keep the threshold at a relatively high and strict value in order to keep the prediction quality. The number of false positive enhancer predictions will be reduced in the prediction procedure of the genome-wide scale search.

Another example is KR-6 (ATAACCCAAT). Because of the similarity of the alignment matrix between KR and BCD. There are several sites both KR and BCD are able to bind transcription factors. Such as, KR-5 (TTAATCCGTT) which shares the binding site with BCD-5 (GTTAATCCG) and KR-3 (GAAGGGATTAG) which shares with BCD-1 (GAAGGGATTAG). But Ludwig *et al.* didn't report binding site GAGCTTAA of BCD with KR-1 (TTAACCCGTTT), both Berman *et al.* and I believe that both TFBS exist. The reason of this KR/BCD

Table 2.9.: Motif threshold evaluation by comparison with Ludwig *et al.* and Berman *et al.* [42, 1]. Patser* means PWM score by Patser; P* means -ln(P) value; T* means -ln(P) threshold; BS* means TFBS discovered in my prediction; Score* means PWM score threshold.

	PatserBS	Patser*	P*	TF	LudwigBS	T*	BermanBS	T*	BS*	Score*
BCD						-6.10		-6.33		5.10
	AAAAGCTG	4.50	-6.13							
	TTAATCCG	7.21	-9.81	BCD-5	GTTAATCCG		TTAATCCG		TTAATCC	
	GAGATTAT	5.71	-7.28	BCD-4	GAGATTATT		GAGATTAT		GAGATTA	
	ATAATCGC	4.47	-6.10	**BCD-3**	TATAATCGC					
	GGGATTAG	7.37	-10.55	BCD-2	GGGATTAGC		GGGATTAG		GGGATTA	
	TCAAGCCC	4.58	-6.22							
	CCAATCCC	5.32	-6.84				CCAATCCC		CCAATCC	
	CCAATCCC	5.32	-6.84				CCAATCCC		CCAATCC	
	CCAATCCC	5.32	-6.84				CCAATCCC		CCAATCC	
	CCAATCCC	5.32	-6.84				CCAATCCC		CCAATCC	
	GGGATTAG	7.37	-10.55	BCD-1	GAAGGGATTAG		GGGATTAG		GGGATTA	
	CGACTTAG	4.80	-6.38				CGACTTAG			
	CTGATCCG	4.91	-6.50				CTGATCCG			
	GAGCTTAA	6.08	-7.75				GAGCTTAA		GAGCTTA	
KR						-6.65		-8.70		4.90
	TATTGGGTTA	4.34	-6.65	**KR-6**	ATAACCCAAT					
	TAATCCGTTT	8.35	-10.36	KR-5	TTAATCCGTT		TAATCCGTTT		TAATCCGTT	
	GACCGGGTTG	6.48	-8.39							
	CAAGGGCTTG	4.51	-6.77							
	CAATCCCTTG	5.01	-7.14						CAATCCCTT	
				KR-4	ACCGGGTTGC				GACCGGGTT	
	GAAGGGATTA	7.1	-8.97	KR-3	GAAGGGATTAG		GAAGGGATTA		GAAGGGATT	
	AACTGGGTTA	7.58	-9.47	KR-2	ACTGGGTTAT		AACTGGGTTA		AACTGGGTT	
	TAACCCGTTT	10.16	-12.75	KR-1	TTAACCCGTTT		TAACCCGTTT		TAACCCGTT	
	AACTGGGTTC	5.04	-7.17						AACTGGGTT	
	CAAATGGTT	6.10	-8.06						CAAATGGTT	
	CAACGGGGTG	4.85	-7.02							
	AACGGGGTGG	4.87	-7.04							
HB						-8.06		-8.19		4.40
	GTCATAAAAAC	7.57	-9.31	HB-3	CATAAAAACA		GTCATAAAAAC		GTCATAAAAA	
	ttatTTTTTTGCGCCgact	8.37	-10.92	HB-2	TTATTTTTTT		TTTTTTGCGCC		ATTTTTTTGCGC	
	acgaTTTTTTTGGCCaaac	6.61	-8.06	HB-1	CGATTTTTTT				ATTTTTTTGGCC	

co-occurrence might be that BCD is a functional activator while KR is a repressor. Thus the motifs are competing to be bound by TFs so that they are able to carry out the regulation correctly.

From these motif threshold comparisons, it becomes apparent that there is a high agreement between Berman *et al.* and my motif thresholds. Additionally, there are some inconsistencies in Ludwig's [42] research, which might be worthy of further discussion. First, Ludwig *et al.* declared that binding sites HB-1 (CGATTTTTTT), HB-2 (TTATTTTTTT) are both discovered by Patser with a threshold of $-ln(P) = -8.06$. But the real HB-1 and HB-2 are with binding sequences of TTTTTTGCGCC and TTTTTTTGGCC according to the alignment matrix. My suspicion is that Ludwig *et al.* retained some nucleotides of the HB motif out of the alignment matrix (see Figure 3.1), as he knew the real binding sites by experimental test. However, Berman *et al.* and I, we both did not include the two beginning nucleotides in TFBS of HB-1 and HB-2. Second, by the thresholds Ludwig *et al.* declared in his publication, Patser could report totally 14 binding sites of BCD, 12 binding sites of KR, which does not match the validated numbers of totally 5 binding sites of BCD, 6 binding sites of KR. It is not clear why Ludwig *et al.* only took these 11 motifs for experimental tests. Comparatively, the motif threshold selection from Berman *et al.* is more precise as a reference for evaluation.

Moreover, five motifs' binding densities are possibly recovered by scoring individual motifs. There is a good correspondence between my method and verified ChIP-chip experiments [24] (see Figure 2.30, black bars).

Overall, almost all of the TFBS were able to be recovered by my threshold selection method. With the references from Ludwig *et al.*, Berman *et al.* and Li *et al.* [24, 42, 1], my selection of threshold was approved both by biological experiment and statistical calculation (see Section 2.1.2). The value of threshold was strict enough to guarantee a genome-wide search. Additionally, I have proved that the motif threshold can be well defined by jPREdictor besides Patser.

2.7.2. Chromosomal rearrangement and Muller elements

During the evaluation procedure, I observed some gene and enhancer shuffling. Such as a statically predicted enhancer in *D. melanogaster* (2L; 20159600-20160800; 19125, neighboring genes *CG15477, CG31677*) whose analogous enhancer in *D. yakuba* is located at chromosome (2R; 7028600-7029599).

Although *Drosophila* species vary in their number of chromosomes, there are six fundamental chromosome arms common to all species. For easier denotation of chromosomal homology, these six arms are referred to as 'Muller elements' after Hermann J. Muller, and are denoted AF. These elements have rearranged following chromosome fusions. The specifics of these rearrangements are shown in Figure 2.31. From the Muller Element Arm Synteny, Flybase has provided an explanation of the inversion in *D. erecta* and *D. yakuba*. There is a shared pericentric inversion at the base of the B.C element (2L.2R in D. melanogaster). Thus relative to *D. melanogaster* the B and C elements are now mixed from telomere to centromere. The new order is B/C* and C/B*.

Although most orthologous genes are found on the same Muller element, there is extensive gene shuffling within Muller elements between even moderately diverged genomes [78]. In contrast to the strong syntenic conservation in *Drosophila*, the order of genes along chro-

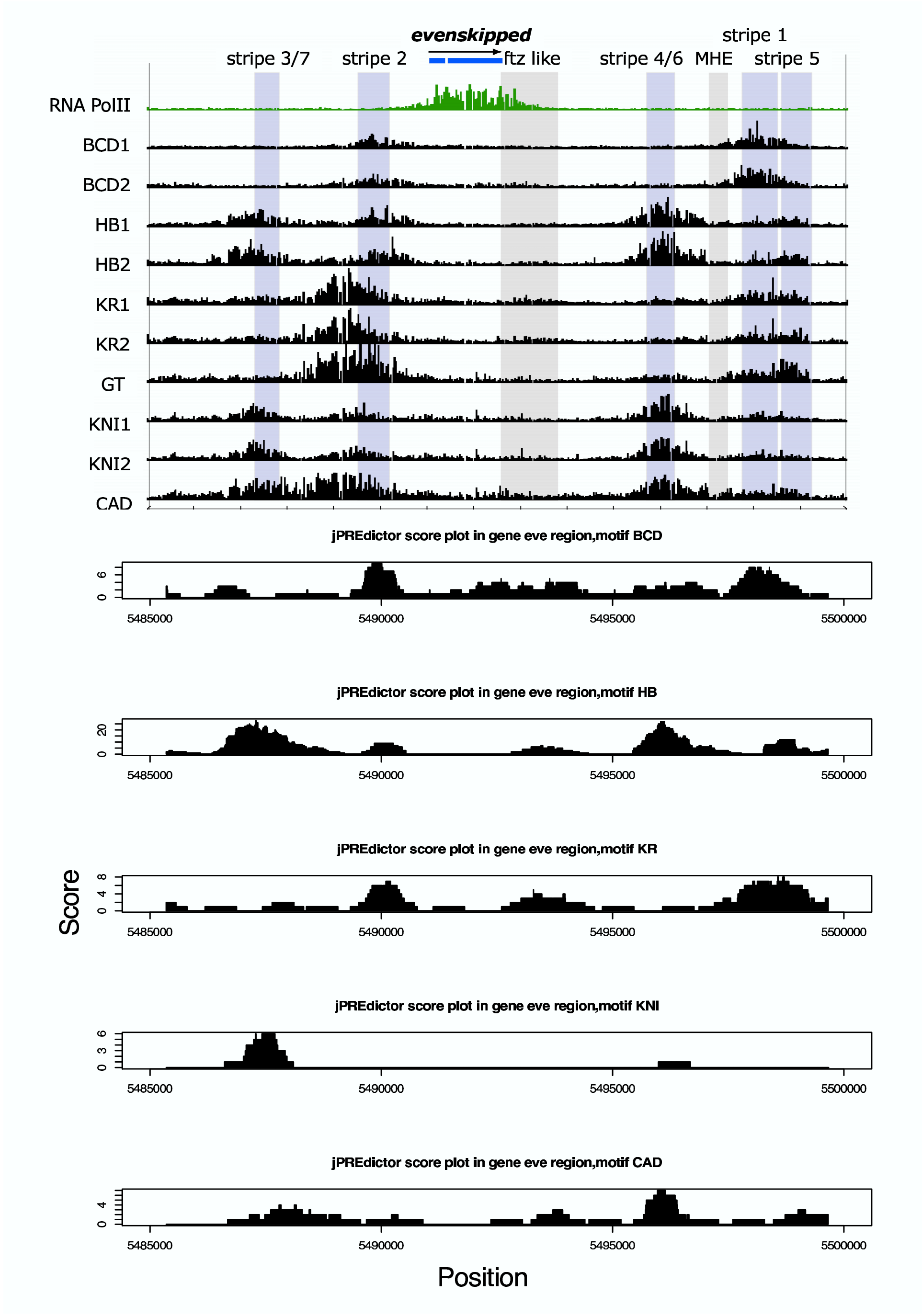


Figure 2.30.: Comparison with the score plots of binding sites by paired motif settings and corresponding *even-skipped* locus oligonucleotide ratio scores from ChIP-chip experiments. (First figure) *In-vivo* binding to the *even-skipped* locus oligonucleotide ratio scores for all ChIP-chip experiments across the well-characterized *even-skipped* locus. Data are shown for RNA PolII and six factors. The light-blue boxes mark the positions of experimentally characterized AP enhancers regulating stripe 1, 2, 3/7, 4/6, and 5. For comparison, the grey boxes mark the positions of the *ftz-like* enhancer and the muscle and heart enhancer (MHE), which are not regulatory elements in the blastoderm. This figure is adapted from [24]. (Second figures) Score plots of every motif by paired motif settings in the *even-skipped* locus.

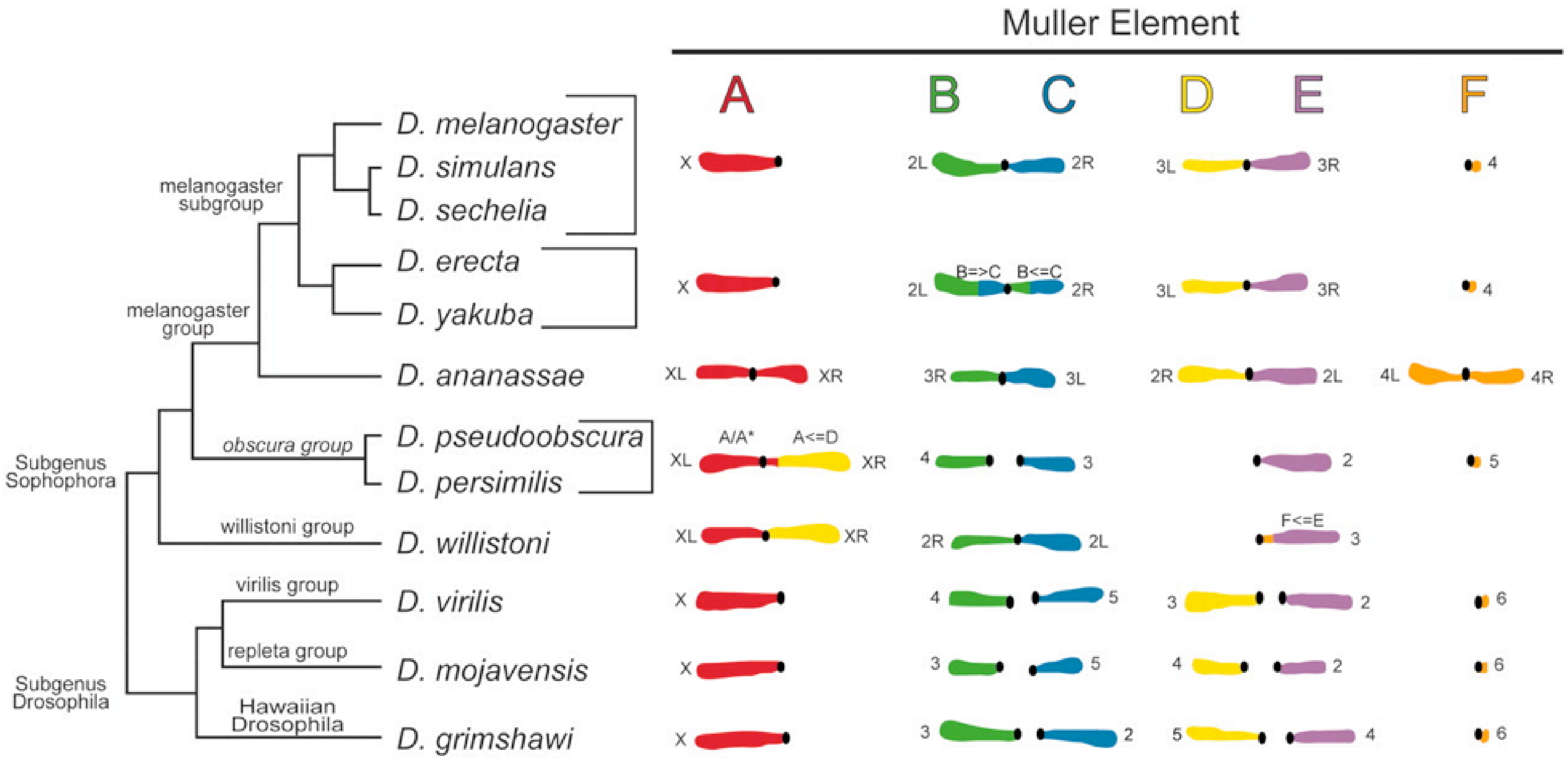


Figure 2.31.: Karyotypic and syntenic relationships of the 12 sequenced species of the genus Drosophila [105]. (Left) The phylogenetic relationships of the 12 species. The members of the two main subgenera, *Drosophila* and Sophophora, are distinguished by the configuration of their autosomes. The ancestral pattern is shown by the subgenus *Drosophila*, which has all acrocentric chromosomes. The members of the subgenus Sophophora differ from this pattern in having varying numbers of fusions of these elements. (Right) The chromosomal arms are separated and aligned as Muller syntenic elements. Each element is colored differently and the arms are also designated by their conventional numbering. There is not a simple one-to-one correspondence for all of the species chromosome arms and a single Muller element. This lack of correspondence is associated with identified fusion and/or inversion events that re-associate all or portions of arms/elements in six of the species relative to D. melanogaster. The black dots designate the positions of the centromeres.

mosome arms is poorly conserved due to the accumulation of inversions that shuffle gene order [105, 118, 119]. I suggest when a gene is rearranged into a different chromosome, the enhancer is together with this gene rearranged into a different region, so that this enhancer can still play its role in gene regulation.

Thus, this observation of analogous enhancers rearranged into different chromosomes is not a false prediction, but a feature of Muller elements.

2.7.3. Prediction validation with published data at TFBS and enhancer level

The biological relevance of the 135 final predictions can be evaluated by using the literature as a guideline [22, 1, 67, 111, 120]. I evaluated the predicted enhancers by measuring the extent to which they overlap well-characterized *cis*-regulatory elements found in resources such as the REDfly database and other publications [22, 1, 67, 111, 120]. I also measured the common region between predicted enhancers and *in-vivo* binding regions. Overall, 27 enhancers show overlap with characterized embryonic enhancers. The detailed examples of bound regions found near well-characterized target genes *h* and *hb* by ChIP-chip experiment match well with my prediction score plots (see Figure 2.32). Li *et al.* [24] gave a large number of *in-vivo* binding regions identified by ChIP-chip experiment; he believes the gap and maternal factors may regulate a much broader array of genes and CREs than the small collection of known target elements. 71 enhancers of my prediction overlap with his experimentally tested bound regions but not with known CREs. My prediction allows his initial suspicion to be brought into the realm of reality. Moreover, 37 enhancers are newly discovered (see Figure 2.33). Further detailed information is available online at http://bibiserv.techfak.uni-bielefeld.de/jpred_en/.

2.7.4. Prediction validation by expression patterns of adjacent embryonic genes

The Berkeley Drosophila Genome Project (BDGP) [121] has catalogued the expression patterns of a large number of genes in *D. melanogaster*, at various stages of development. As a first step towards experimental validation of my prediction, I used this database. In total, 54 predicted enhancers are adjacent to 45 genes which are identified with anterior-posterior patterns in the blastoderm during stages 4-6 (see Figure 2.34, Figure 2.35). ID19200 (X; 20494690-20497109) and ID19351 (X; 20489980-20491200) have the same adjacent gene *run*, but they are identified as two enhancers. The same applies to ID19115 (2L; 12660750-12662219) and ID19333 (2L; 12616350-12617690), both of them regulate gene *nub*. 6 out of 54 enhancers are newly predicted elements that have no overlap with previously characterized enhancers [22, 1, 67, 2, 111, 24, 120]. Their enhancer IDs have been highlighted in red in Figure 2.34 and Figure 2.35. Enhancer ID 19163 (3R; 17435680-17436549) is an exceptional example. This enhancer has two neighboring genes: *InR* and *E2f*. Gene *E2f,* which is 9477bp away from the enhancer, has embryonic expression. However, embryonic motifs are not enriched in this enhancer region [24]. Some of the other predicted enhancers such as 19116 are verified by [24], but no gene expression images can be found in this database. In this case, more experiments on gene expressions have to be done.

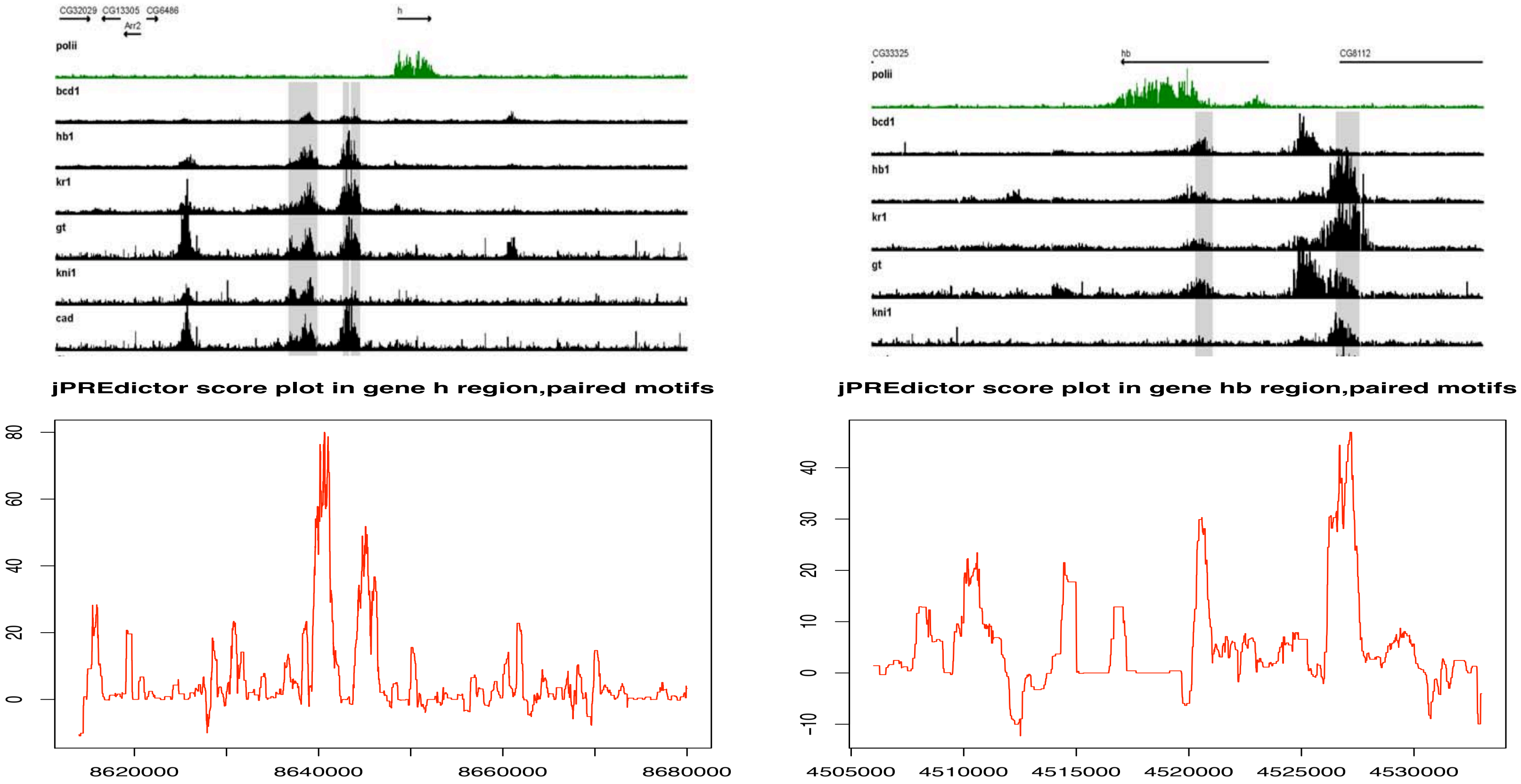


Figure 2.32.: Prediction score plots match ChIP-chip oligonucleotide ratio scores for selected targets. (First image) Bound regions found near well-characterized target genes *h* and *hb*, which overlap known CREs - shown in grey. (Second image) Score plots of gene *h* and *hb,* score peaks match known CREs.

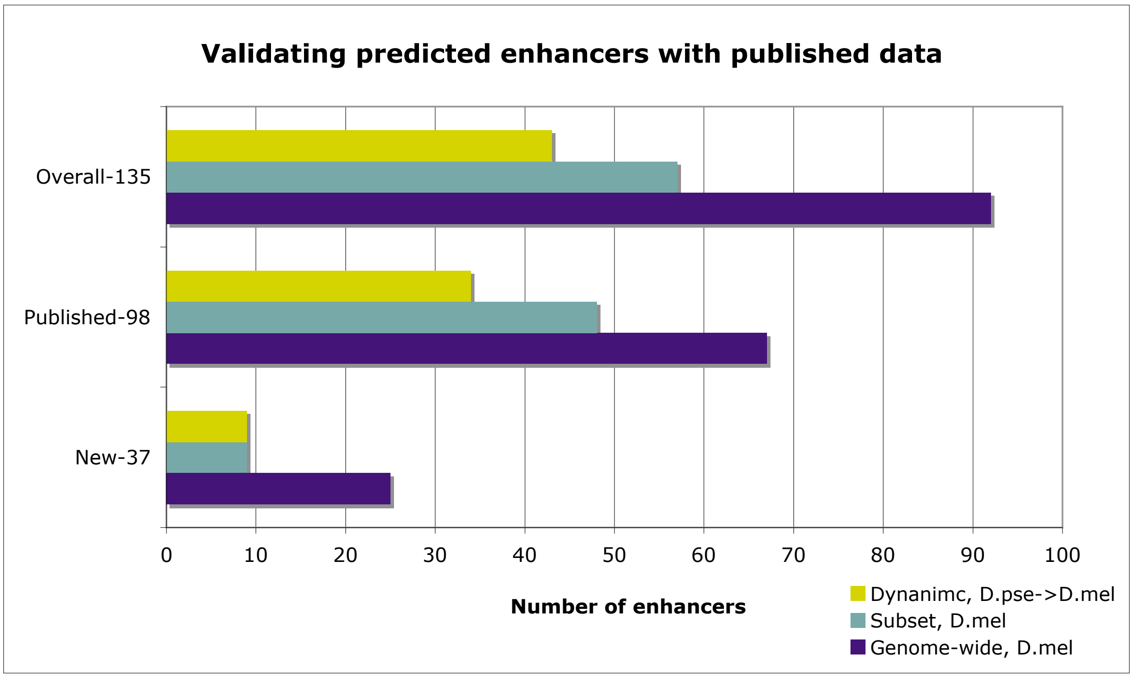


Figure 2.33.: The numbers of overall, published and newly predicted enhancers by genome-wide, embryonic subset and dynamic search. X-axis is the number of enhancers. Y-axis is the summary of unique prediction numbers by three methods. The category of "published-98" includes verified enhancers [22, 1, 67, 2, 111, 24, 120] and TFBSs [24].

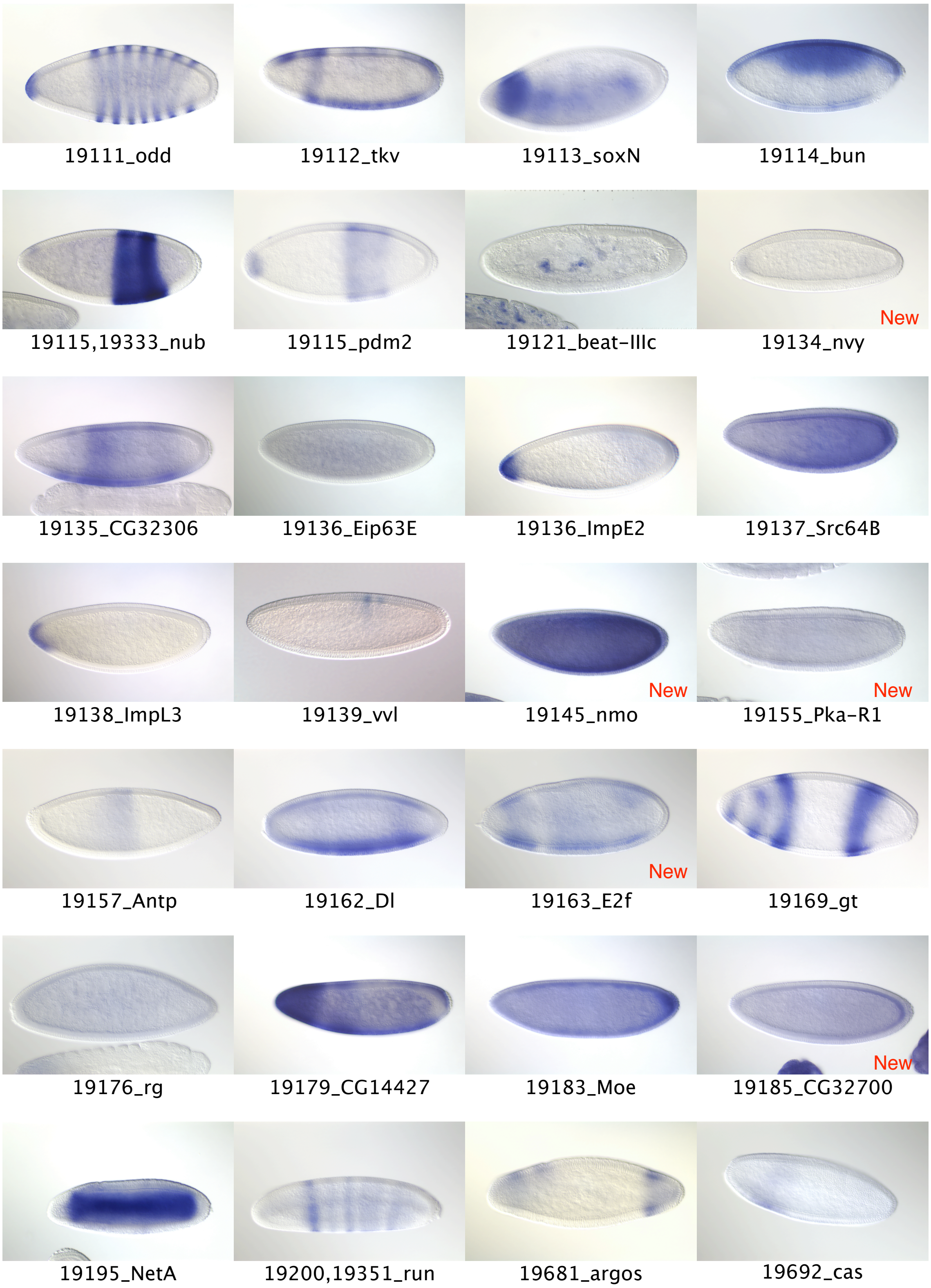


Figure 2.34.: Wild-type embryonic expression patterns of genes adjacent to predicted embryonic enhancers. The images were obtained from the BDGP Embryonic Expression Pattern Database [122]. Patterns of genes that neighbor newly predicted enhancers are marked with "New" in red. Enhancer IDs and gene names are presented below the expression images. See Figure 2.35.

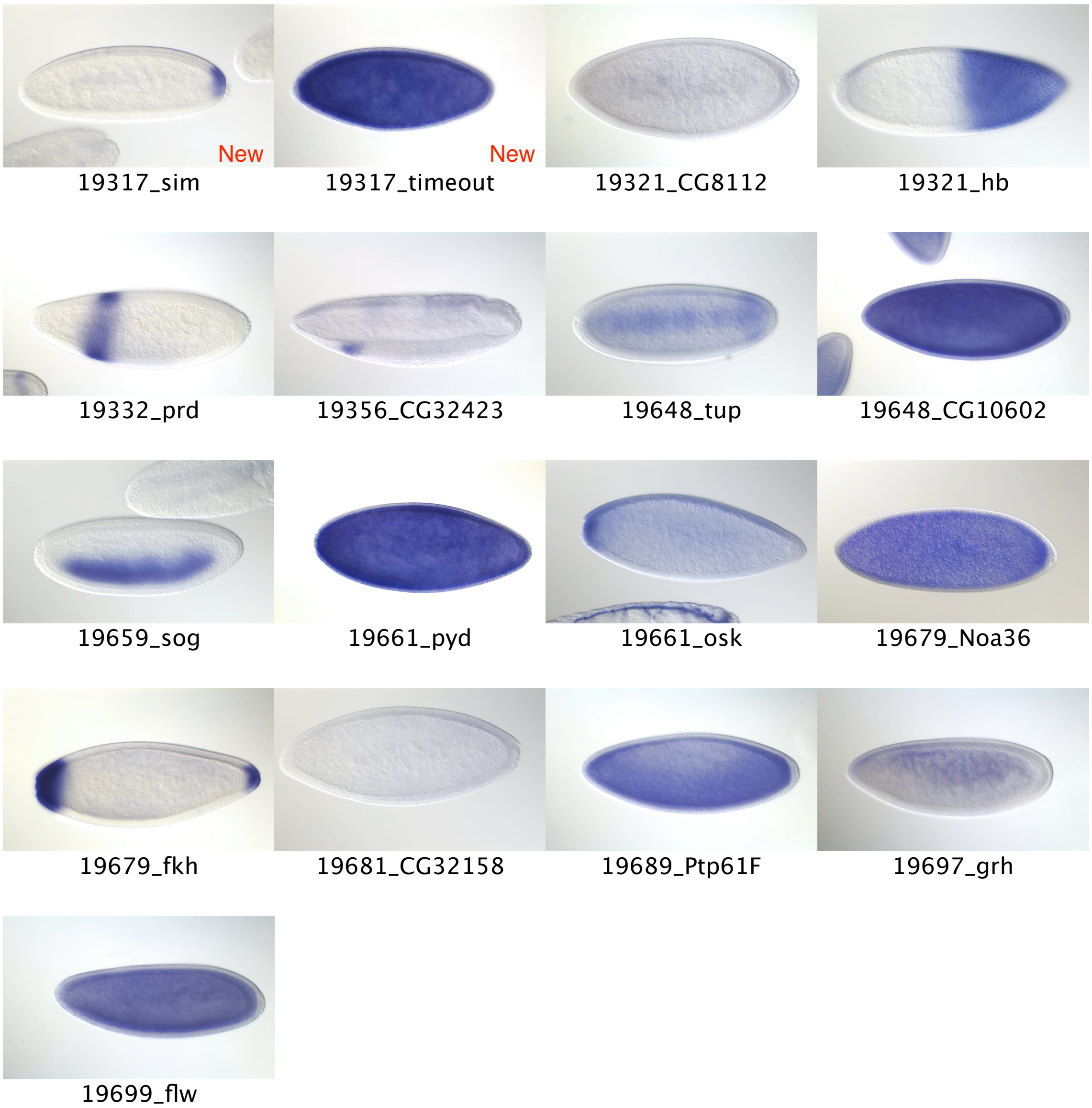


Figure 2.35.: Wild-type embryonic expression patterns of genes adjacent to predicted embryonic enhancers. The images were obtained from the BDGP Embryonic Expression Pattern Database [122]. Patterns of genes that neighbor newly predicted enhancers are marked with “New” in red. Enhancer IDs and gene names are presented below the expression images. See Figure 2.34.

In summary, 40% of 135 predicted enhancers have genes with embryonic pattern nearby, including 6 new predictions. Moreover, gene expression of the remaining predictions could also potentially exhibit embryonic patterns.

2.7.5. Prediction validation by characterizing adjacent gene "Gene Ontology" enrichment

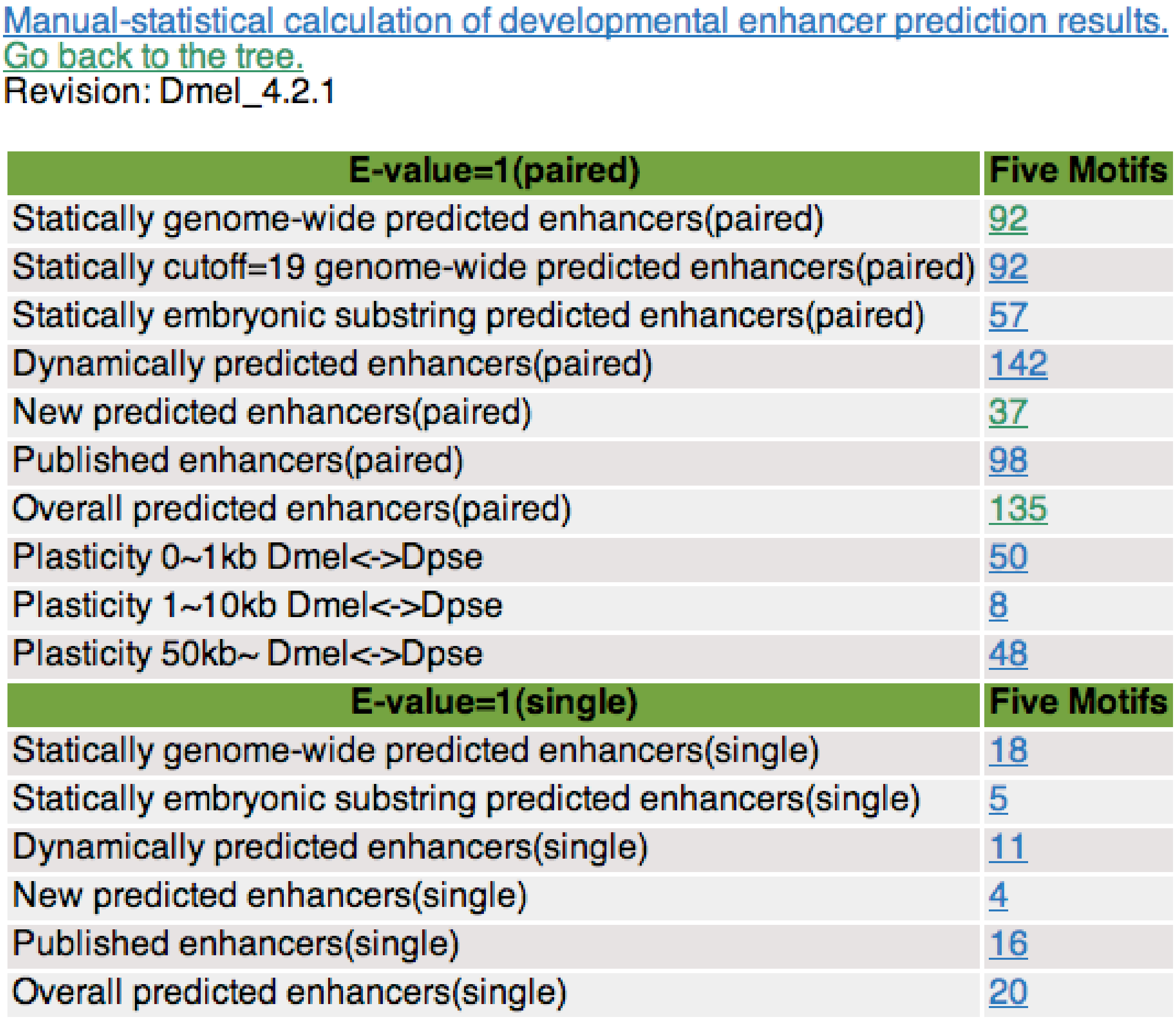
To further probe the associations between predicted embryonic enhancers and their proximate genes, I evaluated the enrichment of gene ontology (GO) terms of the predicted enhancers in three gene association p-value lists. These lists contain known enhancers (see Table A.7 to A.13), *in-vivo* motif binding enhancers (see Table A.4 to A.6) and newly predicted enhancers which I have defined respectively as first, second and third group.

The Gene Ontology database provides a comprehensive and standardized set of annotations for biological processes, molecular functions or cellular components. I chose biological processes to characterize these genes by using GOtermfinder [123]. There is a consistent enrichment of GO terms associated with embryonic development, such as "blastoderm segmentation", "embryonic pattern specification" and "cellular developmental process" etc for the first and second groups. Strikingly, although 49 out of 182 genes could not be identified by GO, 33 GO terms are abundantly annotated with "embryonic development".

However, for the third group consisting of new predictions, no GO terms could be determined as they have not yet been incorporated in the GO database. It could be speculated that these genes have not been annotated and would make excellent candidates for further research.

2.7.6. A website of predicted embryonic enhancers

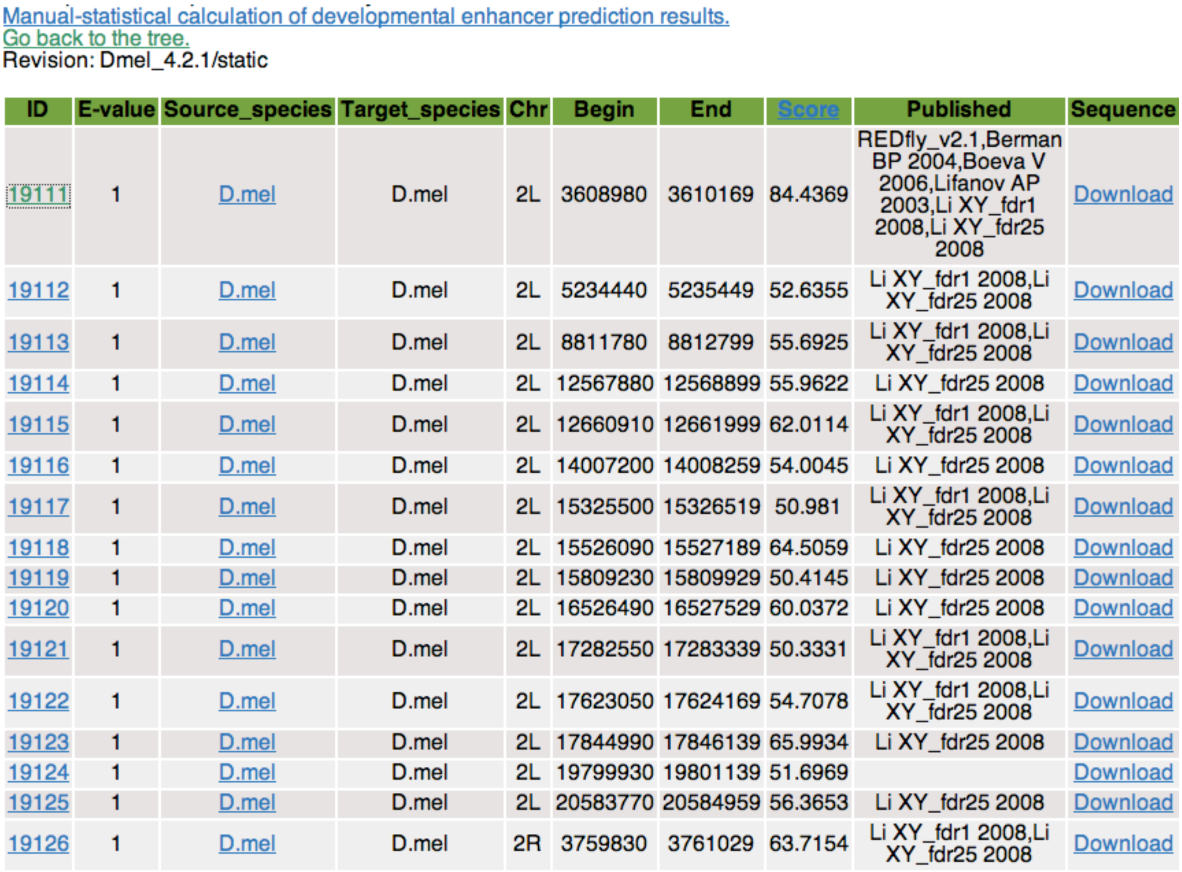
The predicted embryonic enhancers by genome-wide, GO subset and dynamic search are stored in a database which can be accessed online http://bibiserv.techfak.uni-bielefeld.de/jpred_en/. The screenshots of the website are shown in Figure 2.36 and Figure 2.37. The final predictions in eight *Drosophila* species are presented with a hyperlink from their phylogenetic tree. The majority of research has been done on *D. melanogaster*, thus the webpage of *D. melanogaster* contains prediction results by every individual method (genome-wide, subset or dynamic), plus the overall prediction. Tables for overall new predictions and known predictions are listed separately. A graphical view of the score distribution for enhancer plasticity is also provided. The prediction of enhancers in *D. melanogaster* has been performed both by paired-motif setting and by single-motif setting. For the remaining seven species, prediction results are available for genome-wide and dynamic search. Information of each predicted enhancer includes its genomic position and score. Genomic sequences of individual prediction or the whole set of enhancers can also be downloaded in fasta format from the web site. Associated literature proves the prediction to be either a known enhancer or a new element. Each enhancer locates either upstream or downstream of its neighboring genes within certain distances or inside genes. The website contains links to Flybase and BDGP to give an overview of embryonic gene expression.



Manual-statistical calculation of developmental enhancer prediction results.
Go back to the tree.
Revision: Dmel_4.2.1

E-value=1(paired)	Five Motifs
Statically genome-wide predicted enhancers(paired)	92
Statically cutoff=19 genome-wide predicted enhancers(paired)	92
Statically embryonic substring predicted enhancers(paired)	57
Dynamically predicted enhancers(paired)	142
New predicted enhancers(paired)	37
Published enhancers(paired)	98
Overall predicted enhancers(paired)	135
Plasticity 0~1kb Dmel<->Dpse	50
Plasticity 1~10kb Dmel<->Dpse	8
Plasticity 50kb~ Dmel<->Dpse	48
E-value=1(single)	**Five Motifs**
Statically genome-wide predicted enhancers(single)	18
Statically embryonic substring predicted enhancers(single)	5
Dynamically predicted enhancers(single)	11
New predicted enhancers(single)	4
Published enhancers(single)	16
Overall predicted enhancers(single)	20

Figure 2.36.: Screenshot webpage of *D. melanogaster* contains prediction results by every method (genome-wide, subset or dynamic), plus the overall prediction. The prediction of enhancers in *D. melanogaster* has been performed both by paired-motif setting and by single-motif setting.



Manual-statistical calculation of developmental enhancer prediction results.
Go back to the tree.
Revision: Dmel_4.2.1/static

ID	E-value	Source_species	Target_species	Chr	Begin	End	Score	Published	Sequence
19111	1	D.mel	D.mel	2L	3608980	3610169	84.4369	REDfly_v2.1,Berman BP 2004,Boeva V 2006,Lifanov AP 2003,Li XY_fdr1 2008,Li XY_fdr25 2008	Download
19112	1	D.mel	D.mel	2L	5234440	5235449	52.6355	Li XY_fdr1 2008,Li XY_fdr25 2008	Download
19113	1	D.mel	D.mel	2L	8811780	8812799	55.6925	Li XY_fdr1 2008,Li XY_fdr25 2008	Download
19114	1	D.mel	D.mel	2L	12567880	12568899	55.9622	Li XY_fdr25 2008	Download
19115	1	D.mel	D.mel	2L	12660910	12661999	62.0114	Li XY_fdr1 2008,Li XY_fdr25 2008	Download
19116	1	D.mel	D.mel	2L	14007200	14008259	54.0045	Li XY_fdr25 2008	Download
19117	1	D.mel	D.mel	2L	15325500	15326519	50.981	Li XY_fdr1 2008,Li XY_fdr25 2008	Download
19118	1	D.mel	D.mel	2L	15526090	15527189	64.5059	Li XY_fdr25 2008	Download
19119	1	D.mel	D.mel	2L	15809230	15809929	50.4145	Li XY_fdr25 2008	Download
19120	1	D.mel	D.mel	2L	16526490	16527529	60.0372	Li XY_fdr25 2008	Download
19121	1	D.mel	D.mel	2L	17282550	17283339	50.3331	Li XY_fdr1 2008,Li XY_fdr25 2008	Download
19122	1	D.mel	D.mel	2L	17623050	17624169	54.7078	Li XY_fdr1 2008,Li XY_fdr25 2008	Download
19123	1	D.mel	D.mel	2L	17844990	17846139	65.9934	Li XY_fdr25 2008	Download
19124	1	D.mel	D.mel	2L	19799930	19801139	51.6969		Download
19125	1	D.mel	D.mel	2L	20583770	20584959	56.3653	Li XY_fdr25 2008	Download
19126	1	D.mel	D.mel	2R	3759830	3761029	63.7154	Li XY_fdr1 2008,Li XY_fdr25 2008	Download

Figure 2.37.: Screenshot webpage of genome-wide prediction in *D. melanogaster*. Information of each predicted enhancer includes its genomic position and score. Genomic sequences of individual predictions or the whole set of enhancers can also be downloaded in fasta format. Associated literature proves the prediction to be either a known enhancer or a new element.

2.8. Paired motifs distinguish embryonic enhancers from non-embryonic elements

In section 2.5.3, it seemed that overlap and adjacent motifs are preferred by orthologous enhancers in *Drosophila*. This observation supports the initial idea of using paired motifs for embryonic enhancer prediction. Here, I would like to make a full comparison of paired or single motifs settings, and their influence on the prediction result.

2.8.1. Motif weight distribution of single and paired motifs

Motif weight reflects the motif's relative abundance between a positive training set of sequences and a negative training set of sequences. There are no constraints on motif order in both single and paired motifs cases. However, for paired motifs, the distances between two motifs are predefined; motifs are allowed to pair with themselves. The positive motif weight indicates that the motif occurs more often in the positive training set than in the negative training set; the negative motif weight indicates that the motif occurs more often in the negative training set than in the positive training set. A weight close to zero represents equal abundance in both training sets. See chapter 3 for weight calculation in detail. From figure 2.38, the weights of abundant motifs are more significantly amplified in paired motifs than in single motifs. Especially, motif HB, which shows a negative weight in single motif setting, is mainly making a positive contribution after pairing with other motifs. Moreover, the influence of HB-HB negative weights has been minimized due to the increase in the range of values for paired motifs. For example, single HB has a weight of -0.0707 relative to the highest weight from KNI of 0.889. However, the weight of paired HB is -0.0835, and is comparatively less significant than the highest weight from self paired KNI of 2.010.

2.8.2. Distinguishing enhancers from non-embryonic elements

In order to distinguish enhancers from non-enhancers by using the five motifs, I tried jPREdictor both for single and paired motif settings in positive and negative training sets. In the paired motif setting, there are 9 enhancers that can be distinguished from non-enhancers, because their score is larger than the highest value of the negative training set. However, with single motifs setting, only 4 enhancers are distinguishable from the negative training set and only barely so. Moreover, the difference between the highest and lowest score value is much more significant in paired motif setting than in the single one. For example, the range of values is 103.2687 to -13.6676 in paired motif setting and 9.955 to -1.495 in single motif setting. There are even two elements in the positive training set that have a negative score value with single motif setting, while none of the elements have a negative value with paired motif setting (see Figure 2.39). Using paired motif offers a better separation than using single motif in training sets and genome-wide data.

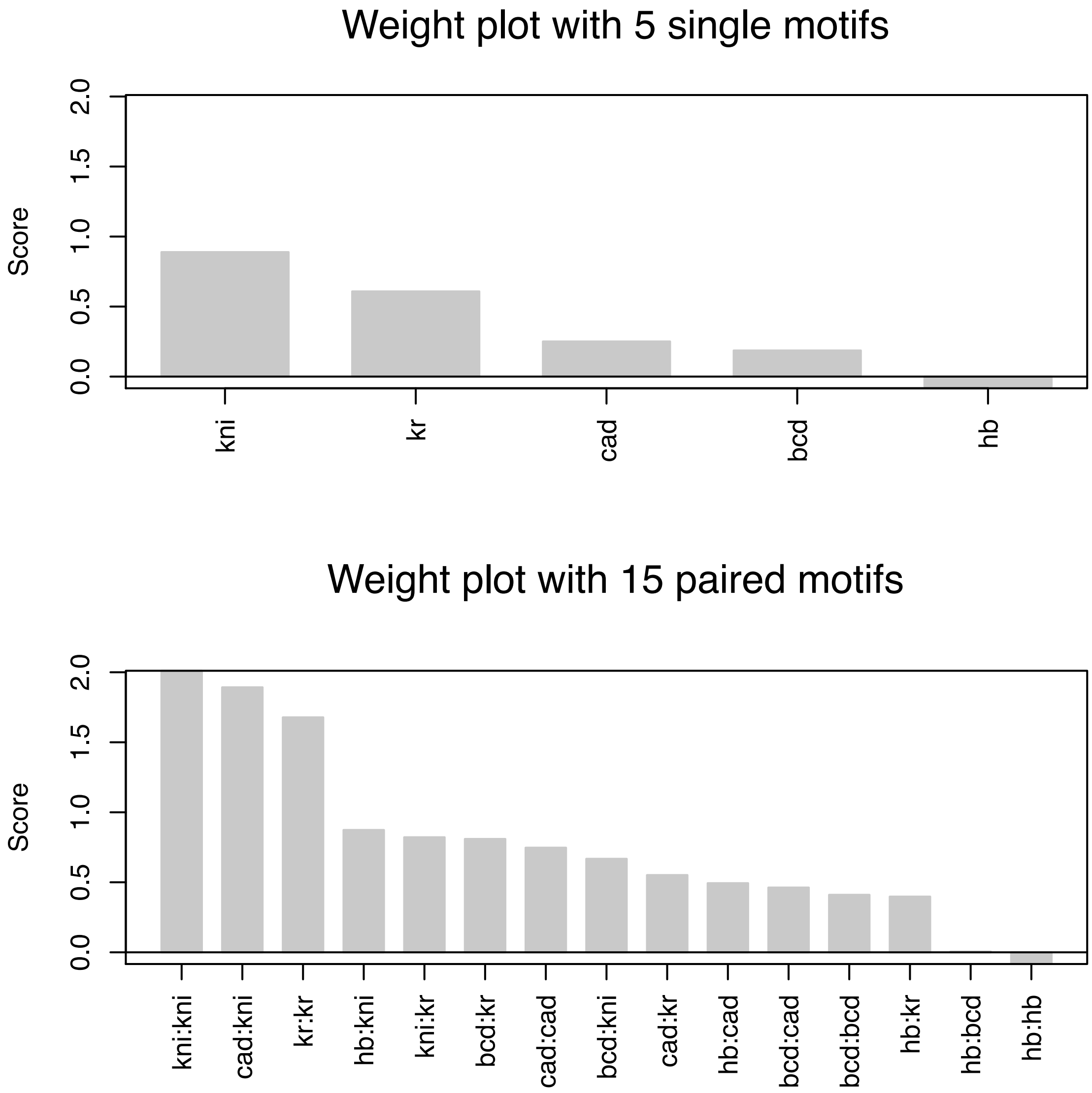


Figure 2.38.: Comparison of motif weights with single or paired motif settings. The first figure shows the distribution of 5 single motifs' weight. The second figure shows the distribution of 15 paired motifs' weights. The X-axis is the name of 15 pairwise motif combinations. The Y-axis is the motif weight.

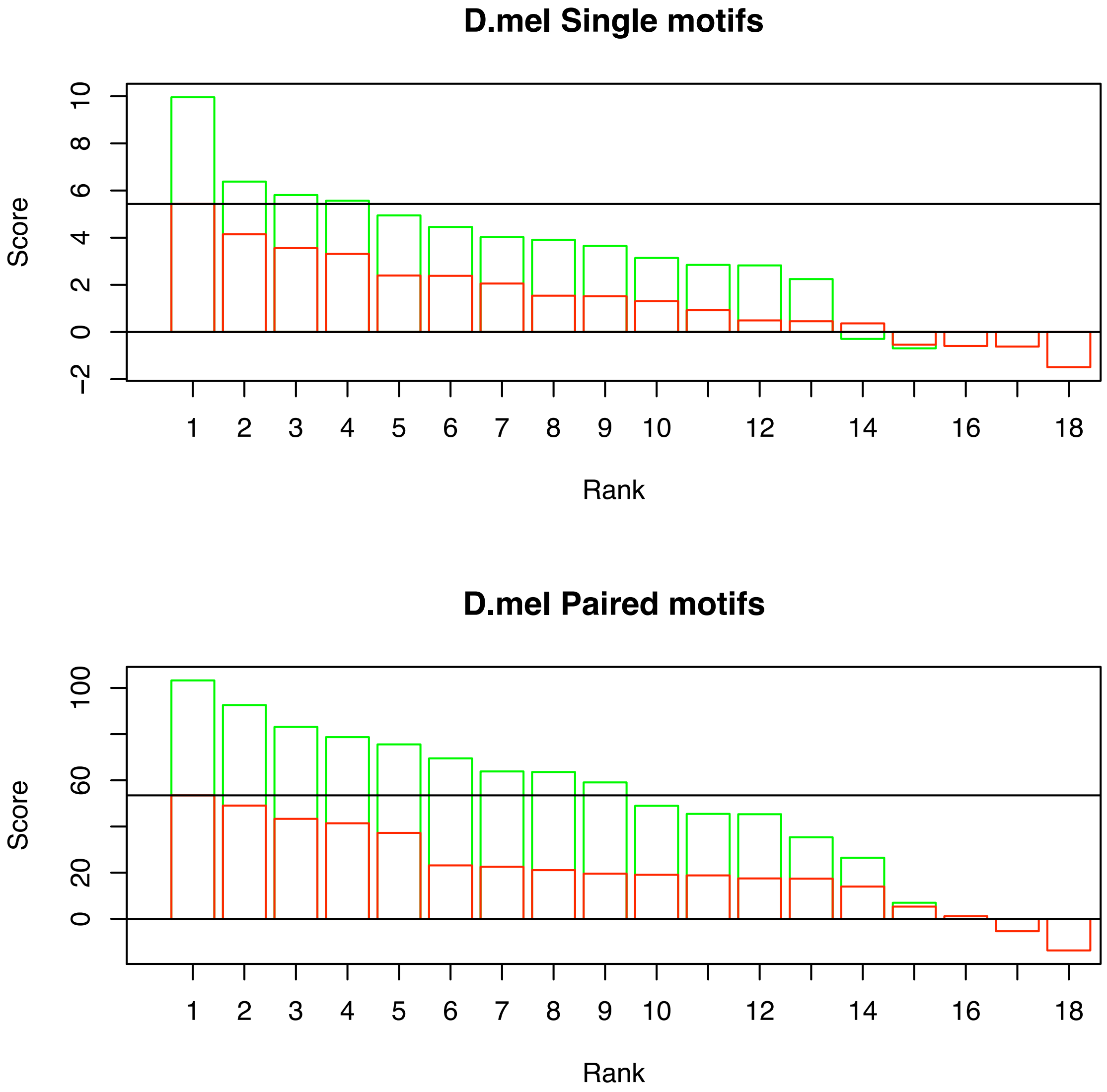


Figure 2.39.: Separation of positive and negative training sets with single/paired motif settings. Green bars represent 15 positive training set. Red bars show 18 negative training set. The black line separates the number of enhancers which have a score higher than the highest score in the negative training set.

2.8.3. The number of predicted enhancers and sensitivity & specificity analysis in *D. melanogaster*

The comparison on the prediction numbers is based on overall enhancer prediction data in species of *D. melanogaster*. By using the paired motif setting, 92 enhancers are discovered to have embryonic function for E-value of 1 in the genome-wide prediction. With an additional 42 enhancers by subset and dynamic methods, the total number of enhancers is 135 (see Figure 2.40). However, with the same method, the same parameters and the same E-value of 1, only 18 enhancers are found with the single motif setting genome-wide. The total number is only 20, after summing up all the single predictions. Meanwhile, if the prediction is only applied on 37 training sets, 9 elements have a score value above the genome-wide cutoff of 50 (E-value of 1) in the positive training set with paired motif setting. But only 1 element is selected with a score higher than cutoff of 7 (still E-value of 1). The calculation of sensitivity and specificity is introduced to allow a statistical comparison. The sensitivity of paired motifs setting is 60% with a high specificity of 94%. But, the sensitivity dramatically decreases to 6% in single motifs setting.

2.8.4. Prediction of five *even-skipped* enhancers

The striped expression pattern of the pair-rule gene *even-skipped* (*eve*) is established by five stripe-specific enhancers, each of which responds in a unique way to gradients of positional information in the early *Drosophila* embryo [36].

In 2008, Li *et al.* [24] performed an experimental validation of five embryonic motifs in the *even-skipped* gene regions. In Figure 2.41, the density of TF bindings is represented by black bars. The verified positions of enhancers regulating stripe 1, stripe 2, stripe 3/7, stripe 4/6, and stripe 5 are marked with light blue blocks. The grey blocks mark the positions of two enhancers that do not respond to maternal or gap factors in the blastoderm, the *ftz-like* [124] enhancer and the muscle and heart enhancer (MHE) [125]. The exact sequence region has been chosen to draw the jPREdictor score plot in a red line by both single and paired motif setting.

Paired motifs performed better than single motifs from several aspects in the example of *even-skipped* region. First, there is a good correspondence of the experimental data with paired motif score plot in the *even-skipped* locus. The light blue regions match well with the prediction peaks. The paired motifs method successfully predicted five enhancers, whereas the single motifs method barely predicted three enhancers, missed the stripe 4/6 enhancer and predicted the *ftz-like* enhancer as a false positive. In the region from 5493kb to 5495kb, the distance between two close single motifs is too far away to be allowed to pair together, since the parameter for motif distance is less than 150bp. The motif weight from each motif could not be summed up into a significant score value. Thus, no enhancer is found in this region with paired motif setting. However, in single motifs, every motif occurrence could be counted, so that the motif weight is added into score value and cause one FP enhancer. Third, with paired motif setting, closely paired motifs amplified the score, the peak at 5496k is formed for presence of the stripe 4/6 enhancer. But in single motif setting, there is no significant score peak to be detected as an enhancer. Fourth, in paired motifs plot, there are clearly two predicted enhancers - stripe 1 and stripe 5. However, the single motifs method failed to separate them. Fifth, the score range for paired motifs is from 0 to 60. The strong

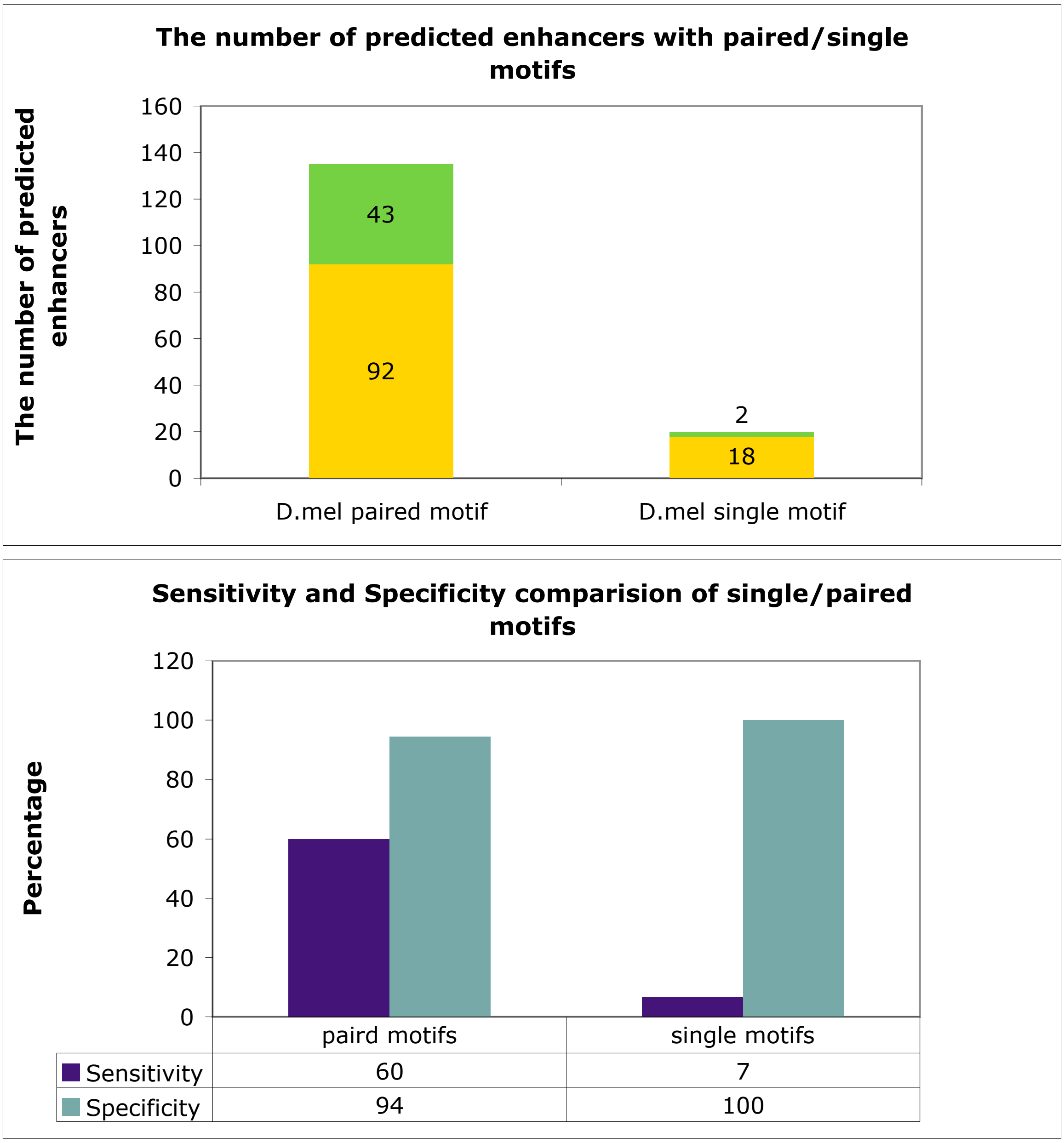


Figure 2.40.: Comparison of prediction numbers and sensitivity/specificity (for both single/paired motif settings) with positive and negative training sets.

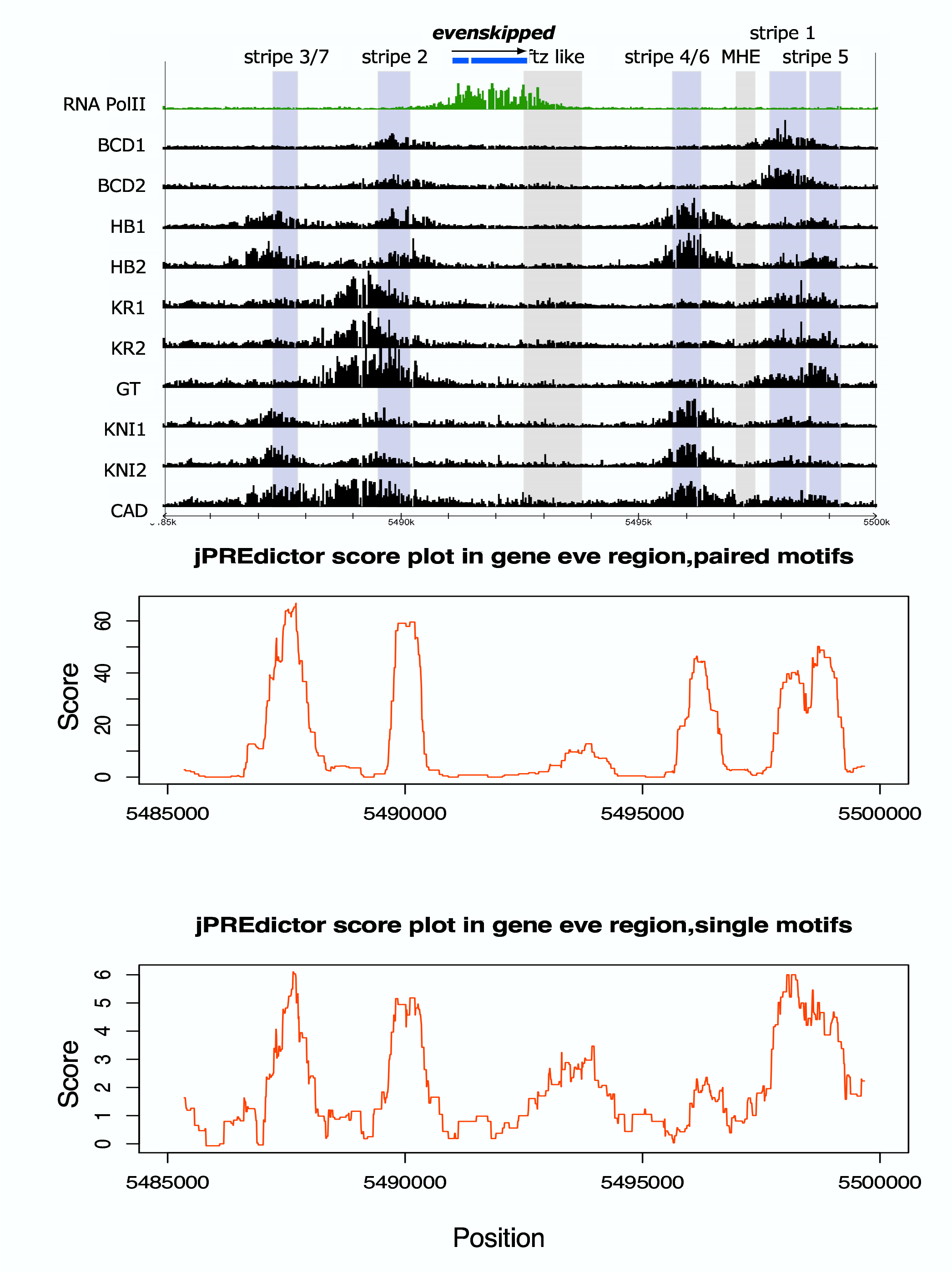


Figure 2.41.: Comparison with the score plots by both single/paired motif settings in *eve* locus. (First figure) *In-vivo* binding to the *even-skipped* locus oligonucleotide ratio scores for all ChIP-chip experiments across the well-characterized *even-skipped* locus. Data are shown for RNA PolII and six factors. The light-blue boxes mark the positions of experimentally characterized AP enhancers regulating stripe 1, 2, 3/7, 4/6, and 5. For comparison, the grey boxes mark the positions of two enhancers that do not respond to these factors in the blastoderm, the *ftz-like* enhancer and the muscle and heart enhancer (MHE). This figure is adapted from [24]. The second and third figure shows score plots with paired/single motif settings in the *even-skipped* locus.

score peaks in Figure 2.41 are clearly distinguishable from the background. As a negative control, a random sequence was generated by shuffling the *even-skipped* locus with markov-order 1. Scores for these random sequence were calculated for both single and paired motifs setting. Figure 2.42 shows that there are no peaks that score higher than the real data and the peaks are not discernible from the background. Using single motifs, the scoring region differs only in value from 0 to 6, making distinguishing the real data from background more difficult than when using paired motifs where the scoring region differs from 0 to 70.

2.8.5. Paired motifs perform better prediction than single motifs

Using paired motif setting produces higher motif weights, performs better separation of enhancer / non-enhancer and predicts more enhancers with high specificity and sensitivity. It is quite convincing that usage of paired motifs is much more suitable for embryonic enhancer prediction than using a single motif setting.

2.9. Comparison of enhancers' and PRE/TREs evolutionary plasticities - differences and similarities

Cis-regulatory DNA elements can be generally classified into two classes. One such class of *cis*-regulatory DNA elements is enhancers, which initialize the regulation of genes expression. Another important class is Polycomb/Trithorax response elements (PRE/TREs), which regulate several hundred developmental genes and are vital for maintaining cell identities [96].

Although PREs are similar to enhancers in many ways, the most important functional difference between these two classes of elements is that enhancers respond to local differences in concentration of the transcription factors that bind them, whereas the Polycomb group (PcG) and Trithorax group (TrxG) proteins are ubiquitously expressed [96].

I have documented three kinds of evolutionary plasticity of embryonic enhancers, including: the numbers of enhancers, non-conservation of enhancers position and motif turnover in positionally conserved enhancers. Similar plasticity of PRE/TREs has been discussed recently in [96]. A comparison of evolutionary plasticity study on enhancer and PRE/TRE should help to gain a better understanding of embryonic developmental *cis*-regulatory elements.

2.9.1. The number of enhancers is lower than the number of PRE/TREs in *Drosophila* embryo regulation network

For static genome-wide prediction, 92 embryonic enhancers have been predicted in *D. melanogaster* with motifset BCD, CAD, KR, KNI, HB. With the same prediction method, 201 PRE/TREs have been discovered, although the parameters for genome-wide prediction are slightly different (i.e 500bp window size moves in 10bp step and motif distance is 220bp). PRE/TREs prediction was accomplished with a different motifset - PM, PS, PF, Zeste, GAF,

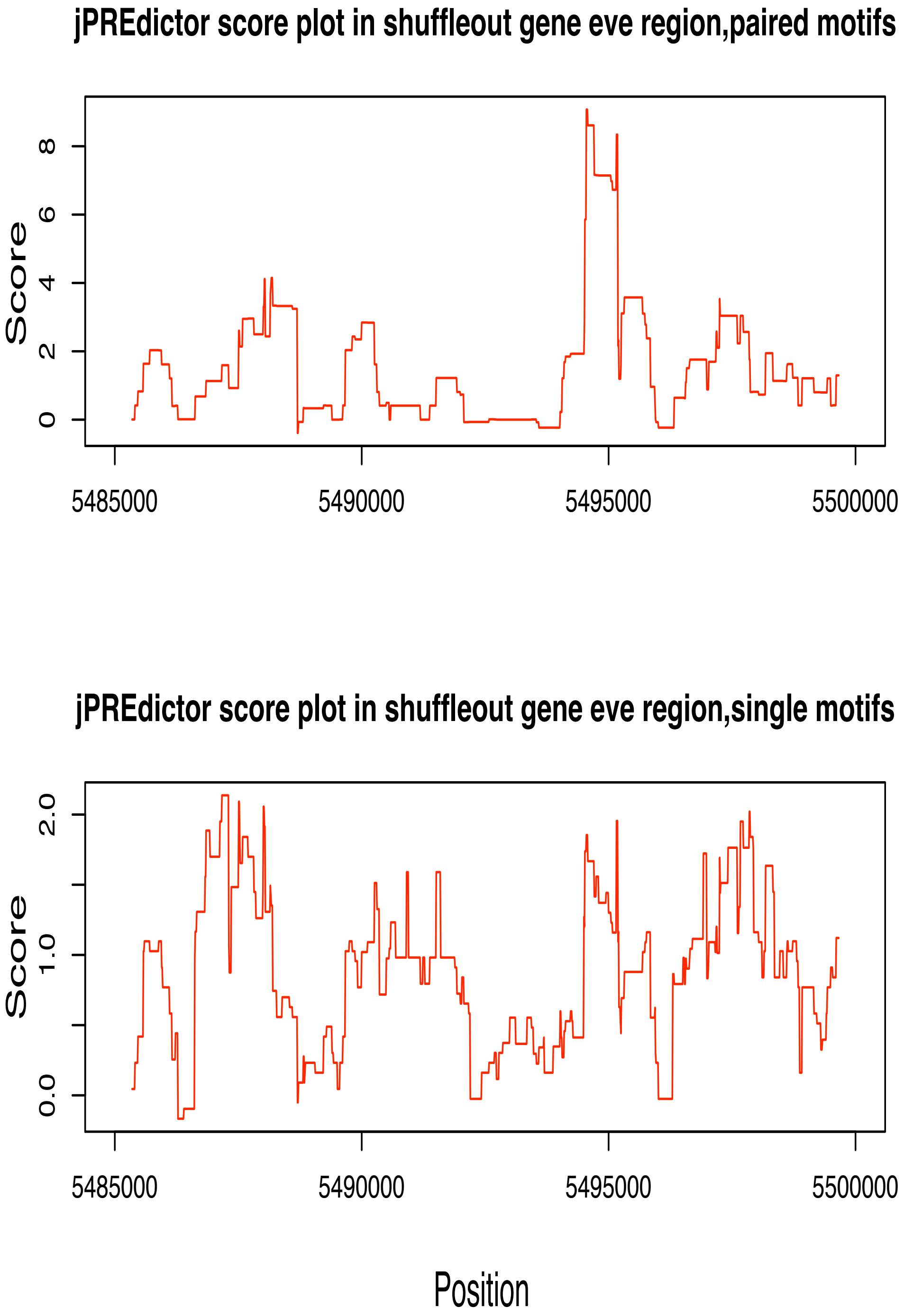


Figure 2.42.: Comparison of the score plots by both single/paired motif settings with shuffled *eve* sequence in *eve* locus. Score plots by paired/single motifs settings in the *even-skipped* locus, which sequence has been shuffled in markov order 1. See Figure 2.41.

G10, En1, DSP1/KLF and DSP1. The number of PRE/TREs is more than double the number of embryonic enhancers in *D. melanogaster*.

The first type of plasticity varies between these two *cis*-regulatory elements. For enhancer prediction, *D. melanogaster* is the species that contains the largest number of enhancers. There are only 55 enhancers in *D. pseudoobscura*, which is 60% (55 out of 92) of the *D. melanogaster* number. On the other hand the number of PRE/TREs in *D. melanogaster* is only 37% of the number in *D. pseudoobscura* (201 in comparison to 538). Data source of PRE/TREs prediction is available in reference [96].

Therefore, embryonic enhancers have the first plasticity - different numbers of enhancers in different species. But, enhancer plasticity shows different feature from PRE/TREs.

2.9.2. Enhancers and PRE/TREs do not regulate the same genes

Embryonic enhancers act at very early stages of *Drosophila* development and are crucial in patterning the anterior-posterior axis of the embryo. PRE/TREs can regulate several hundred developmental genes including regulation of morphogenetic pathways [96]. In order to ascertain the genes that commonly associate with predicted enhancers and PREs in *D. melanogaster*, genes that are in the neighborhood of the enhancers and PRE/TREs are compared. Although both kinds of *cis*-regulatory element are for developmental genes regulation, only two neighboring genes were found in common.

Enhancer *eve* stripe 2 (2R:5489450-5490619) and PRE/TREs (2R:5490190-5491169) are overlapping, thus this PRE/TRE is potentially co-regulating with *eve* stipe 2 enhancer. The same situation occurs for enhancer (3L:20630320-20631769) and PRE/TREs (3L:20629150-20629879), which seem to co-regulate the closest gene *kni*.

Furthermore, in order to gain an overall comparison, the distance between the nearest enhancers and PRE/TREs were calculated as shown in Figure 2.43. The majority of enhancers are far away from their nearest PRE/TREs, which suggests they probably do not regulate the same genes. The distance distribution in the real dataset has no difference from the random dataset.

Although both *cis*-regulatory elements play an important role in embryo development; it may be speculated that PRE/TREs ubiquitous regulation is different with enhancers regulation of target genes, which are expressed at specific positions within the gradient.

2.9.3. Positional plasticity is various for enhancers and PRE/TREs

I have observed enhancer position plasticity in section 2.3.3. The same type of plasticity has been reported in PRE/TREs in [96]. Thus, in order to examine the common behavior of this plasticity, a distance between analogous enhancer to its BLAST locus and analogous PRE/TREs to its BLAST locus was calculated in Figure 2.44. Although the number of enhancers is much smaller than the number of PRE/TREs, position plasticity is commonly observed in both *cis*-regulatory elements. For embryonic enhancers, 40%(37 out of 92) analogous enhancers are within 10kb from BLAST locus. On the other hand, 55%(110 out of 201) analogous PRE/TREs are less than 10kb away from BLAST locus. Therefore, PRE/TREs have a lower rate of position plasticity than embryonic enhancers.

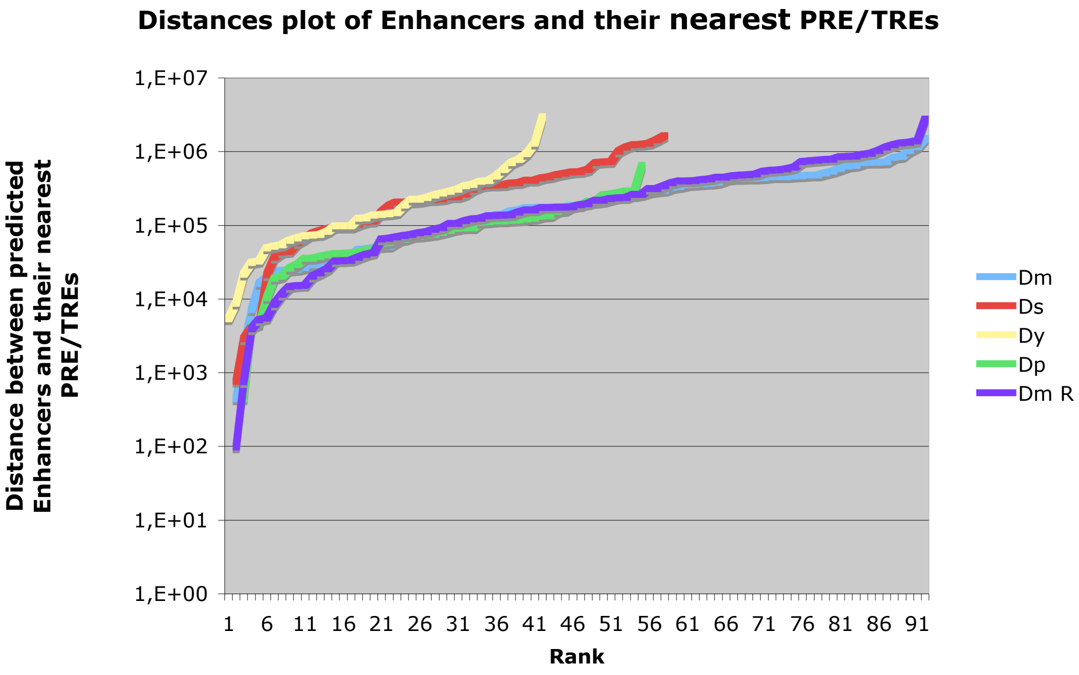


Figure 2.43.: Distances plot of predicted enhancers and their nearest PRE/TREs in species of *D. melanogaster* (Dm), *D. pseudoobscura* (Dp), *D. simulans* (Ds), *D. yakuba* (Dy). (Dm R) means random dataset generated by *D. melanogaster* predicted enhancers. X-axis is the distance rank from *D. melanogaster* prediction; Y-axis is the distance between predicted enhancers and their nearest PRE/TREs.

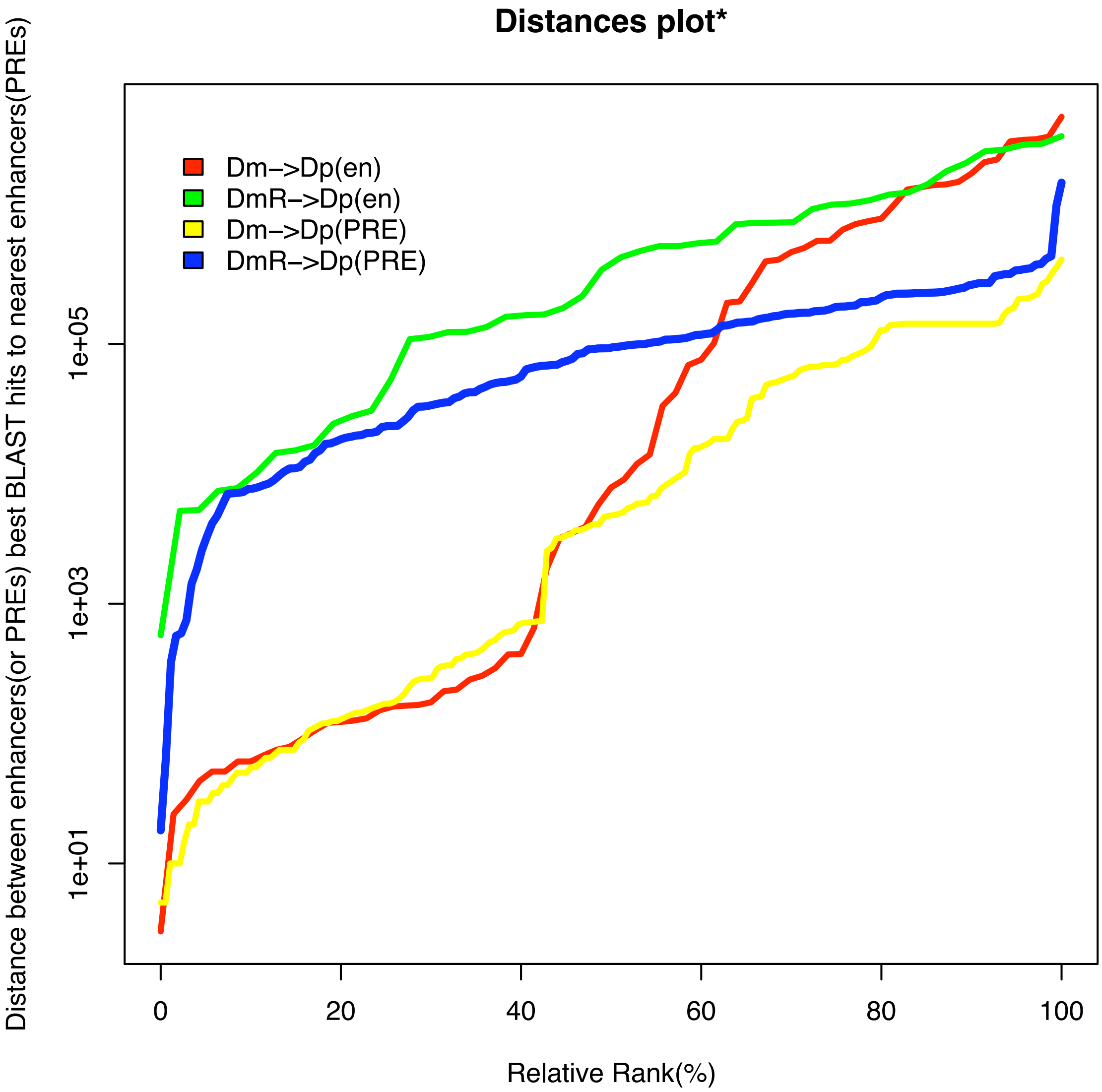


Figure 2.44.: Distances plot* of orthologous regions and functional analogs between enhancers and PRE/TREs. X-axis shows relative enhancers rank by elements percentage, Y-axis is log distance of predicted enhancer locus and its nearest BLAST hit. For target/source taken from *D. melanogaster* (Dm), *D. pseudoobscura* (Dp), *D. simulans* (Ds), *D. yakuba* (Dy). (Dm R) means random dataset generated by *D. melanogaster* predicted enhancers.

2.9.4. New elements can arise from nonfunctional sequence both for enhancers and PRE/TREs

In section 2.4.4 and Table 2.7, 10% (9 out of 92) embryonic enhancers have been gained in D. melanogaster and 8 out of 9 enhancers are associated with genes that previously had no enhancers. The data from PRE/TREs prediction indicates the same observation, although the percentage of gained PRE/TREs is slightly larger than enhancers. 16%(33 out of 201) PRE/TREs are inferred to have been gained in D. melanogaster using a BLAST distance of 10 kb. As a preliminary conclusion, new elements may arise from nonfunctional sequence both for enhancers and PRE/TREs, but further experimental validation is required.

3. Methods

3.1. Initial preparation

3.1.1. Selection of motifs and construction of Position Weight Matrices

Multiple bound transcription factors act combinatorially to confer specific transcriptional activity in the early *Drosophila* embryo. Bicoid, Hunchback, Knirps, Kruppel and Caudal binding sequences were compiled from several papers. The experimentally verified binding sequences from different resources are accessible online [22].

Based on an alignment of all known sites, the frequencies of different nucleotides are recorded for each position, producing position frequency matrix (PFM) (see Figure 3.1). The PFM is recalculated to weights by converting the occurrence probabilities to a log-scale in order to be able to consider the background distribution. This conversion of PFM into PWM allows a different kind of quantitative descriptions to be used for the known binding sites for a TF [126, 56].

Before transforming the downloaded alignment matrix into a position weight matrix (PWM), a gap check was performed to assure consistency of motifs. Gaps in the nucleotide alignments were left out as shown in Figure 3.1. Final alignment matrices were verified so that the sum of unique nucleotides equals the number of alignments. For example, when the first and second columns of CAD alignment matrix are removed, gaps are present, hence the overall number of occurrences is less than 34. The final matrix contains the last 8 columns in which no gaps are present.

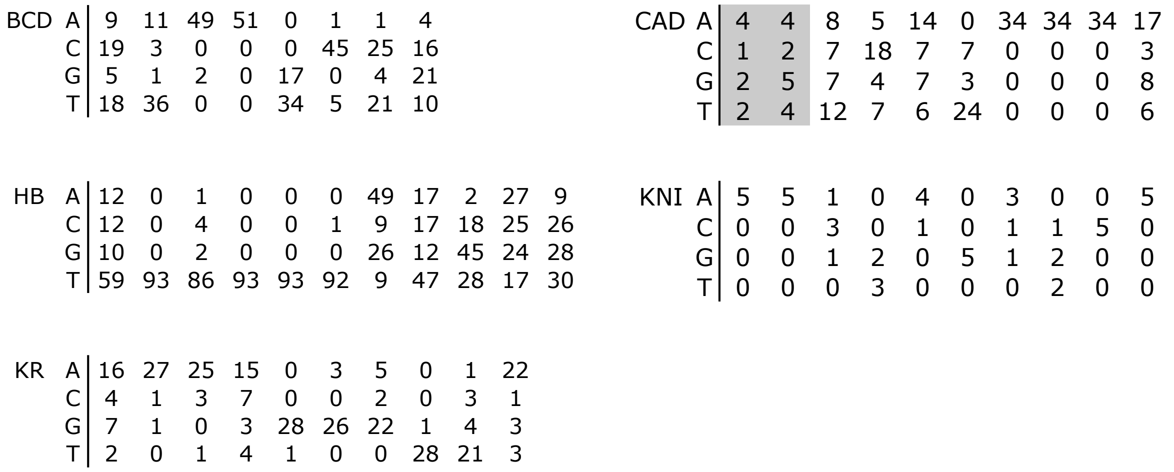
The log-scale conversion from PFM to PWM starts with the calculation of probabilities of observing a given nucleotide in PFM [117, 56]:

$$f(b,i) = \frac{n_{b,i}}{N} \tag{3.1}$$

$n_{b,i}$: letter b is observed at position i of this alignment; N : the total number of sequences; $f(b,i)$: the frequency of letter b at position i.

Then, the genome nucleotide distribution is taken into account in the conversion. The genome of *D. melanogaster* has a background distribution p(A) = 0.2877, p(C) = 0.2124, p(G) = 0.2124, p(T) = 0.2877, which shows a non-uniform nucleotide distribution. The nucleotides A and T occur more often than nucleotides C and G. The PWM is constructed by dividing nucleotide probabilities from equation 3.1 (see also: Figure 3.1 and Table 3.1) with background probabilities and transforming the divided values into a log scale:

$$W(b,i) = ln(\frac{f(b,i)}{p(b)} + c) \tag{3.2}$$



BCD	A	9	11	49	51	0	1	1	4
	C	19	3	0	0	0	45	25	16
	G	5	1	2	0	17	0	4	21
	T	18	36	0	0	34	5	21	10

CAD	A	4	4	8	5	14	0	34	34	34	17
	C	1	2	7	18	7	7	0	0	0	3
	G	2	5	7	4	7	3	0	0	0	8
	T	2	4	12	7	6	24	0	0	0	6

HB	A	12	0	1	0	0	0	49	17	2	27	9
	C	12	0	4	0	0	1	9	17	18	25	26
	G	10	0	2	0	0	0	26	12	45	24	28
	T	59	93	86	93	93	92	9	47	28	17	30

KNI	A	5	5	1	0	4	0	3	0	0	5
	C	0	0	3	0	1	0	1	1	5	0
	G	0	0	1	2	0	5	1	2	0	0
	T	0	0	0	3	0	0	0	2	0	0

KR	A	16	27	25	15	0	3	5	0	1	22
	C	4	1	3	7	0	0	2	0	3	1
	G	7	1	0	3	28	26	22	1	4	3
	T	2	0	1	4	1	0	0	28	21	3

Figure 3.1.: Motif alignment matrices were download from Berman *et al*.. [1]. PWMs were constructed from these matrices. Gaps (the first and second columns in grey) of CAD alignment matrix were left out.

Table 3.1.: The Knirps (KNI) motif as PWM. The most significant scores are in blue. Background distribution is p(A) = 0.2877, p(C) = 0.2124, p(G) = 0.2124, p(T) = 0.2877.

A	1.25	1.25	-0.35	-4.61	1.03	-4.61	0.74	-4.61	-4.61	1.25
C	-4.61	-4.61	1.04	-4.61	-0.05	-4.61	-0.05	-0.05	1.55	-4.61
T	-4.61	-4.61	-0.05	0.64	-4.61	1.55	-0.05	0.64	-4.61	-4.61
G	-4.61	-4.61	-4.61	0.74	-4.61	-4.61	-4.61	0.34	-4.61	-4.61

$W(b,i)$ is the PWM value of base b in position i; $p(b)$ is the background probability of b; $c = 0.01$ denotes a small pseudo-count.

The final log-scale matrix is referred to as a PWM. As an example, see the PFM for KNI in Table 3.1. A quantitative score for a potential binding sequence can be generated by summing up the relevant nucleotide PWM values at each position (using equation 3.3).

$$S = \sum_{i=1}^{w} W_{l_i,i} \tag{3.3}$$

S is the PWM score of a sequence; w is the width of the PWM; l_i is the nucleotide in position i in an input sequence.

Such a PWM profile allows to assay any sequence for binding site potential through assignment of a quantitative score. Fast and intuitive visual verification of pattern characteristics can be performed through the generation of sequence logos (see Figure 2.29). This is done by making the height of each nucleotide letter proportional to its probability of occurrence and adjusting the height of the entire stack to the information content.

The transforming procedure from PFM to PWM is included in jPREdictor [98].

3.1.2. Selection of training sets

The training sets neither have to contain the same number of sequences, nor have the same sequence length; however, they should be in FASTA format. Both positive and negative training sets could be very similar to each other in terms of containing a specific composition of motifs involved in a known functional pattern or in terms of relatively high conservation of primary sequences. In this study, the five embryonic motifs are allowed to occur in both positive and negative training sets. Fifteen sequences were selected as a positive training set, since they have been experimentally proved to be functional enhancers regulating *runt, eve*, *hairy* and some other genes. Eighteen sequences were chosen as the negative training set since they do not appear to have enhancer activity [1].

3.1.3. Motif weight calculation

Each motif weight was assigned by counting the motif occurrences in a positive training set (model) versus a negative training set (background). A higher motif weight means these motifs are more abundant in the positive training set, also called over-represented. For example, this is the case for Knirps (see Figure 2.3). A lower weight means lower presence, a weight of zero means an equal distribution between positive and negative training sets and a negative weight means higher occurrence in negative training set than positive training set, also called under-represented. The formula of motif weight:

$$w(M) = \ln \frac{f(M|P)}{f(M|N)}. \tag{3.4}$$

$w(M)$: natural logarithm of the frequency of a motif in the P (positive training set) and in the N (negative training set)

The motif weight calculations for BCD, CAD, HB, KR and KNI were completed with both single and paired motifs settings. N single motifs can comprise $\frac{N^2+N}{2}$ paired motifs. Thus 5 motifs will form 15 motif pairs (with self-coupling). The actual calculation of motif weight was done by jPREdictor.

The motif weights later used to scan for enhancers may be different when different motif thresholds are assigned, even though the 5 motif PWMs are fixed. Thus a good enhancer prediction starts with a good motif threshold selection.

3.1.4. PWM threshold selection

If a sum-score exceeds a defined threshold, a match is found [98] (see the equation 3.3). A higher threshold, which requires a more stringent match, is likely to result in fewer accepted binding sites. Also, the threshold can not be set too low, since that will basically match every sequence piece. The rationale behind threshold selection is to try to maximize the occurrence of motifs in the positive training set and to minimize occurrences in the negative training set. Every motif PWM threshold was defined by trying out every motif weight with thresholds from 3 to 8 in steps of 0.2. The positional probabilities are multiplied in the PWM. The threshold was finally selected by the correspondence of the highest possible weight in the threshold region (see Section 2.1.1). Certainly, the final threshold of the PWM would be smaller than the maximum possible score.

3.1.5. Sections of the option file

jPREdictor is a program written in Java to support the genome-wide prediction of *cis*-regulatory regions on the basis of predefined motifs [98]. In order to use it, an option file has to be prepared with required information, which I have described in previous sections. After motif collection, PWM construction, training set selection and motif threshold calculation, the data needed for the option file consists of these further basic elements:

1. Motif name, Motif PWM, Motif Background, Threshold.
2. Training data: positive and negative training sets,
3. Sequences file.
4. Distances in case of MultiMotifs.

(see Table A.14 for a detailed example of an option file).

3.2. Scoring method in general

3.2.1. Scoring procedure

Every single motif or paired motif was assigned a specific motif weight (see Figure 2.3). Then, the weights of all motifs were used to derive score profiles for the sequences file, which is either the whole *Drosophila* genome or random sequences. The window score was assigned by taking the sum of the motif weights inside a window of a specific width [97, 98].

$$S(m) = \sum_{m} w(m)o(m)\,. \tag{3.5}$$

Table 3.2.: *Drosophila* genome versions [127, 128]

Species	Genome size	Version
D. melanogaster	132M	Flybase r4.2.1, Assembly Apr.2004
D. simulans	123M	Assembly Apr.2005
D. yakuba	127M	Assembly Apr.2004
D. pseudoobscura	155M	Flybase r2.0, Assembly Nov.2004
D. sechellia	169M	Assembly Oct.2005
D. ananassae	235M	Assembly Aug.2005
D. erecta	155M	Assembly Aug.2005
D. persimilis	191M	Assembly Oct.2005

$S(m)$ is the motif score, $w(m)$ is the weight of motif m and $o(m)$ is the number of occurrences of motif matches in the given window.

Scores for each sliding window (700bp) were calculated in 10bp steps across the entire sequence, until the score plot covered the whole sequence.

3.2.2. Genome version

Drosophila genome sequences were compiled from Flybase [127]. The eight selected species are from the melanogaster and obscura group. Each species' genome version is specified in Table (see Table 3.2). Genome-wide and subset searches were carried out by using a single genome. Dynamic search was performed by using paired genomes.

3.2.3. Null model

By default, the null model for the prediction is a random control sequence that was generated with the same nucleotide composition as the actual *D. melanogaster* genome (0-order Markov chain), but is about 100 times larger than the entire genome.

Except this default setting for random sequence generation, there are two other ways for creating random data, either by following higher order Markov chains, or by shuffling real genomes. The latter version is created by concatenating fragments of length 10 that were randomly chosen from the *D. melanogaster* genome, and is called "shuffled-out" [129] and is the default setting for jPREdictor.

The choice of the null-model (0/higher-order Markov Chains or shuffled genome data) will influence the later cut-off calculation [129].

3.2.4. Determing cutoff for CRE identification

In order to determine the significance of the real sequence scores, jPREdictor was run both on the real genome sequence and on a random control sequence with the same settings. Sequence regions with high scores are assumed to be functional elements, while regions with low scores are assumed to be background (non-functional elements). In order to identify real CREs from non-CREs (noise), a score cut-off has to be defined. Such a score cutoff is expressed in terms of an E-value. For a given score, the E-value is the number of times one expects to find that score (or higher) in the real genome sequence [97]. When the E-value is 1, only 1 false positive is expected to occur by chance in the genome. A sequence element whose score exceeds the cutoff will be counted as a predicted enhancer.

Analysis of the random sequence showed that an E-value of 1 corresponds to a score of 50. I chose this E-value as the cutoff for the prediction of genome-wide enhancers in *Drosophila*, and thus sequences that score below 50 will be excluded in this analysis.

Cut-off calculation can be applied to other E-values as well for E-values from 0.1 to 1000. More stringent E-values reduce the number of predicted enhancers but increase specificity.

For this research project I mainly focus on the enhancer prediction with E-value setting of 1. Although there may be many TPs in the genome that have a score below the cutoff for E-value 1, the aim of this embryonic enhancer study is not to find all embryonic enhancers, but to find real enhancers. Later methods such as subset and dynamic search will make up for this deficiency, and optimize both sensitivity and specificity (see 2.2 and 2.3).

In the actual implementation, both real genomic data such as every score value corresponding to a position as well as background data such as cutoff, number of FP occurrences and E-values are stored in a PostgreSQL database.

3.2.5. Parameter selection

Four parameters need to be considered in the approach: PWM threshold (which has been discussed previously in section 3.1.4), window size, window shift and paired motif distance. Parameter selection mainly focuses on the combinations between window size and paired motif distance. Optimizing these parameters helps us to distinguish between positive and negative training sets.

First, window size tells the width of the sequences in which searching for motifs and scoring is done. This window width was selected to be 700bp in this study. Second, window shift is defined in terms of a base pair value which represents how the window slides across the sequence. jPREdictor's default window shift of 10bp was used. Third, paired motifs distance has to be taken into account. A paired motif was defined as two motifs occurring in any orientation on either strand within a distance of 0 to 150 bp.

The first criterion for judging the quality of the parameter settings is to count how many scores in the positive training set are higher than the highest score in the negative training set. I took the 15 highest scores from each sequence in the positive training set's sequence file, and the 18 highest scores from each sequence in the negative training set's sequence file. Simply counting these scores can give an initial impression of how well a given parameter could separate positive and negative training sets. As a second criterion, a Kolmogorov-Smirnov (KS) test was applied to the whole training set to give an idea whether the difference is significant by p-value. KS-test calculation is done by using the R command ks.test [130]. I retained the parameter that gives a p-value smaller than 0.001 in the KS-test.

A window shift of 10bp is the default setting of jPREdictor. The scoring result will be much more precise, when the window shift is set to be 1, but a compromise between computational efficiency and quantity needs to be struck. Thus, a default setting of 10bp is a reasonable choice.

3.3. Genome-wide extent of enhancer borders

Due to the selection of the window size parameter used in the genome scoring procedure, the minimum length of predicted enhancers is 700bp. A typical CRE has a length of approximately 500-1000 bp [42]. The actual length of enhancers should not be limited to an artificial window size. In order to achieve a greater precision in the predicted enhancer's position, especially in the case of short enhancers of less than 700bp length, the prediction widths are adjusted.

The score cutoff is recalculated on the sequences of predicted enhancers while keeping an E-value of 1. The sum of the total number of predictions is taken in one genome, e.g. in *D. melanogaster*, keeping the total sequence length in mind. The same prediction procedure was repeated on the random control region as described in section 3.2.3. Once the search region is limited to the predicted enhancer region, the cutoff for this specific small region is 19 and is lower than the genome-wide of 50.

The corresponding cutoff of the total motif occurrences can be decided by using the following formula:

$$E = \frac{Ls}{Lr} * O. \tag{3.6}$$

where E is the E-value, O is occurrence, Lr is the length of random sequence, Ls is the length of searching sequence (prediction sequence).

In genome-wide search, an E-value of 1 corresponds to 1 FP in real data and 100 occurrences in random data (because $Lr = 100Ls$). Given the same E-value, the longer the searching sequence, the higher the cutoff. As a comparison, a cutoff of 19 (E-value of 1) for the locally extended prediction regions corresponds to an E-value of 1000 in the genome-wide prediction. This leads to a highly increased number of candidate enhancers compared to a genome-wide prediction run with the same E-value setting of 1. Since the attempt is to extend enhancers instead of finding new ones, this lowered cutoff 19 was used to scan over the genome where regions of predicted enhancers exist. In the end, every initial genome-wide prediction region was extended (see Figure 3.2).

3.4. Gene Ontology subset approach

As mentioned previously in section 3.2, there may be many enhancers that have a score just below the genome-wide cutoff for E-value 1. I applied a subset prediction to deal with this situation.

The Gene Ontology (GO) database offers a stringent database with terms linked to biological processes, including specific groups related to embryonic development in *Drosophila* species. Generally, there are four ways of querying the GO database: via AmiGO, SQL, Perl or XML/RDF. I opted for using a local copy of the MySQL database for performance reasons (database schema shown in Figure 3.3). The database 2009-06 is the latest version and the syntax has slightly modified since my queries in May 2006. For example, "SELECT * FROM instancedata" replaces "SELECT distinct xrefkey FROM gene_product". For performance reasons, the results of my initial queries were imported into my local database. For

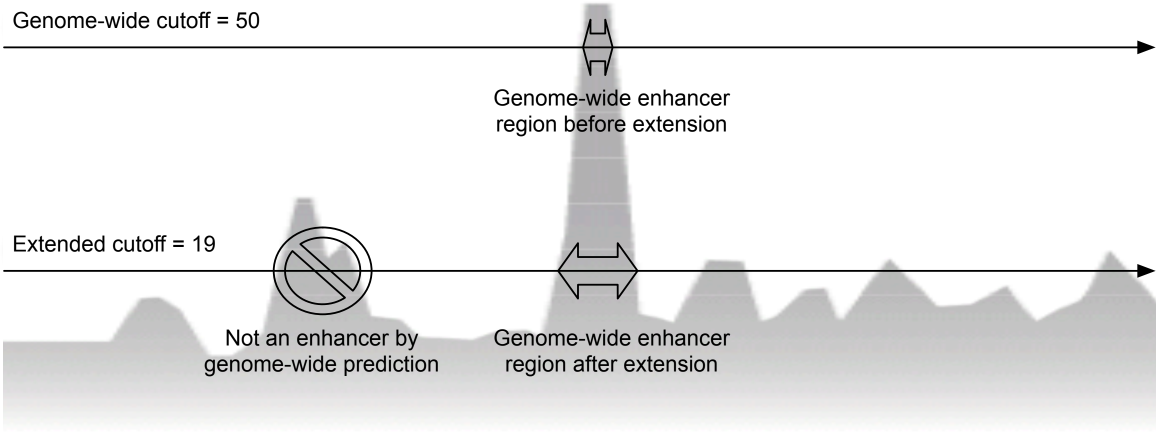

Genome-wide cutoff = 50
Genome-wide enhancer region before extension
Extended cutoff = 19
Not an enhancer by genome-wide prediction
Genome-wide enhancer region after extension

Figure 3.2.: Genome-wide prediction with extension of enhancer region. With the genome-wide cutoff, one enhancer could be predicted. The enhancer region could be extended by lowering the cutoff from 50 to 19.

example to fetch every descendent of "embryo", I performed the following query (valid for 2006-05 GO database schema):

```
SELECT rchild.* FROM term AS rchild, term AS ancestor, graph_path WHERE
graph_path.term2_id = rchild.id and graph_path.term1_id = ancestor.id and ancestor.name
regexp 'embryo';
```

In the subset prediction, all the children of terms "embryonic development" and every descendent were fetched from the Gene Ontology database. GO labels are species independent. The number of genes annotated with embryonic development related labels in Flybase amounts to 2813 in *Drosophila* and these genes correspond to 26486 GO IDs. By incorporating the GO approach, it is possible to limit the searching region to the flanking region of specific GO-supported embryonic genes (10kb up- and downstream region). The embryonic subset prediction ended up with a 21Mb sequence file, or 17.5% of the whole genome (132Mb). The cutoff for this subset prediction was decreased to 41 from the genome-wide cutoff 50 after calculation with equation 3.6. As a result, additional enhancers were found by this method.

3.5. Neighboring gene types

Enhancers are independent of their position and orientation with respect to the transcriptional initiation site [132]. Enhancers may be located in the intergenic region, either upstream or downstream of the regulated gene and some enhancers might even map to intronic regions [133]. Therefore, classifying enhancers by their position relative to neighboring genes helps to fix the location of predicted enhancers in the whole genome. I defined 7 types describing

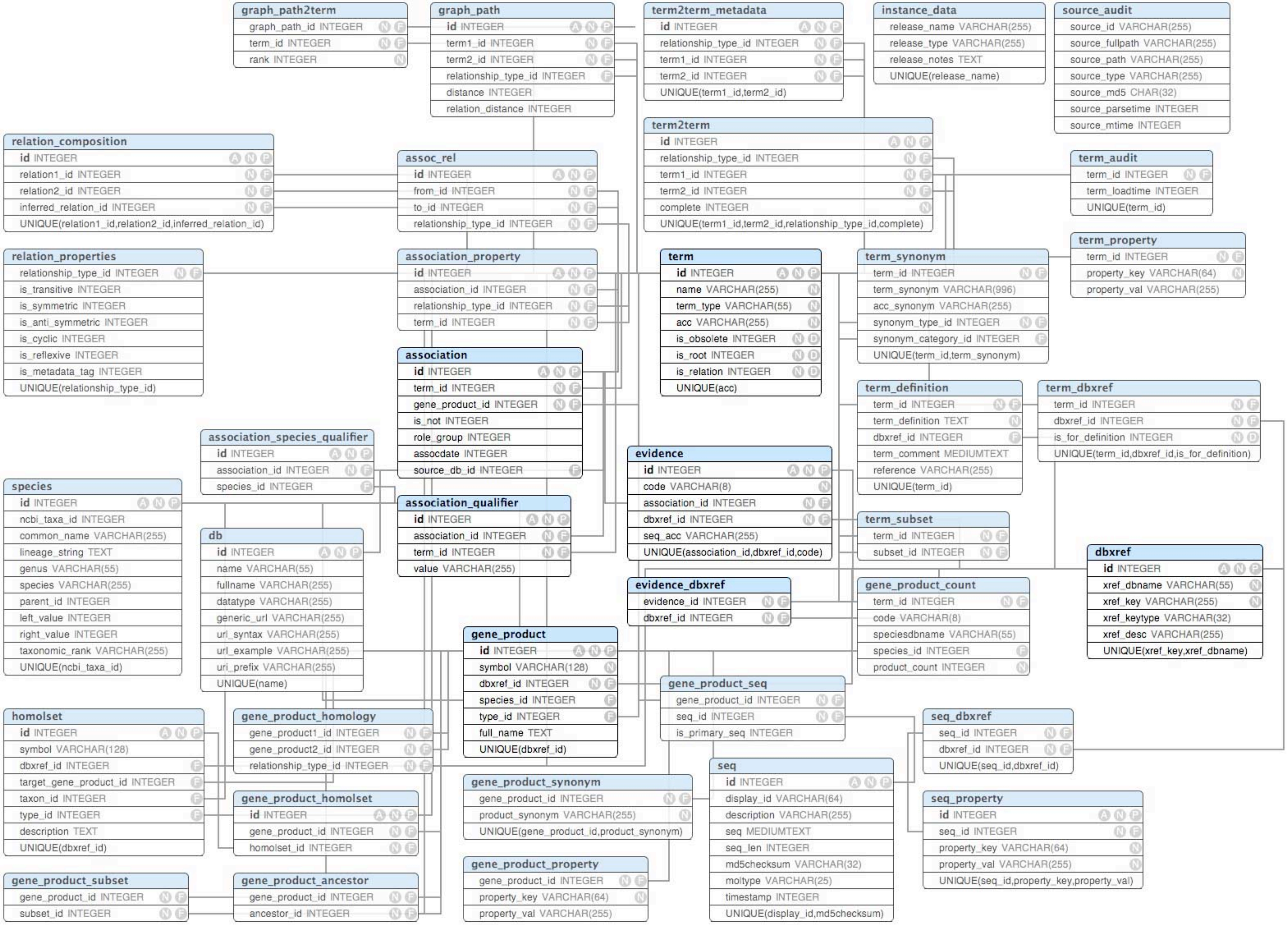


Figure 3.3.: Gene Ontology schema diagram (2009-06) from [131]. The schema for the GO database consists of tables for storing the terms and structure of the GO ontologies, along with gene product and annotation data. The central tables are term, term2term, association and gene_product.

the positions of genes and enhancers Figure 2.8.

- Type 0 means the gene is located inside the predicted enhancer.
- Type 1 means the predicted enhancer is located inside the gene.
- Type 2 means the predicted enhancer overlaps with its 5' gene.
- Type 3 means the predicted enhancer overlaps with its 3' gene.
- Type 4 means the predicted enhancer has no overlap with its nearest 5' gene, and is downstream of the gene.
- Type 5 means the predicted enhancer has no overlap with its nearest 3' gene and is upstream of the gene.
- Type 6 means the gene and the predicted enhancer are exactly overlapped.

For type 4 and 5, the distance between enhancer and nearest gene was calculated from edge to edge. For the other types, the distances are assigned as 0.

This enhancer type classification has been applied not only in *D. melanogaster* but also in some other *Drosophila* species.

Because of the limitations of the currently available gene annotation version, well annotated gene sequences are only available for *D. melanogaster* and *D. pseudoobscura*. Thus, the gene information of *D. simulans* and *D. yakuba* are generated by blasting the corresponding genes from *D. melanogaster*. The BLAST approach is appropriate for 2 reasons: first, *D. simulans* and *D. yakuba* share the same chromosome arms with *D. melanogaster* [105]; second, they are evolutionarily close to *D. melanogaster* (see Figure 2.31).

3.6. Dynamic search

Enhancer prediction based on binding data from a single species is referred as 'static search' in this project. This static enhancer prediction was applied in 8 *Drosophila* species: *D. melanogaster*, *D. pseudoobscura*, *D. simulans*, *D. yakuba*, *D. sechelia, D. erecta, D. persimilis* and *D. ananassae.*

However, computational searches for conserved sequences might only identify a small fraction of the enhancers in the genome [134]. Additionally, parallel studies in multiple species allow an explicit comparison of the evolutionary changes in regulatory sequences. To characterize these events, a 'dynamic search' method was developed which is described below.

Dynamic search [96] consists of four steps: 1) searching, 2) BLAST, 3) selection, 4) reproduction. The workflow is presented in Figure 3.4. I applied these four steps in order to perform cross-species prediction of embryonic enhancers.

First, a static search for the embryonic enhancers is carried out in the whole genomes of all eight *Drosophila* species. The predictions were retained for an E-value of 1, which corresponds to a genome-wide cutoff score of 50.

Second, (i) The homologous regions of predicted enhancers in each of the other *Drosophila* species are determined by doing a BLAST search with the first step's predicted enhancers as

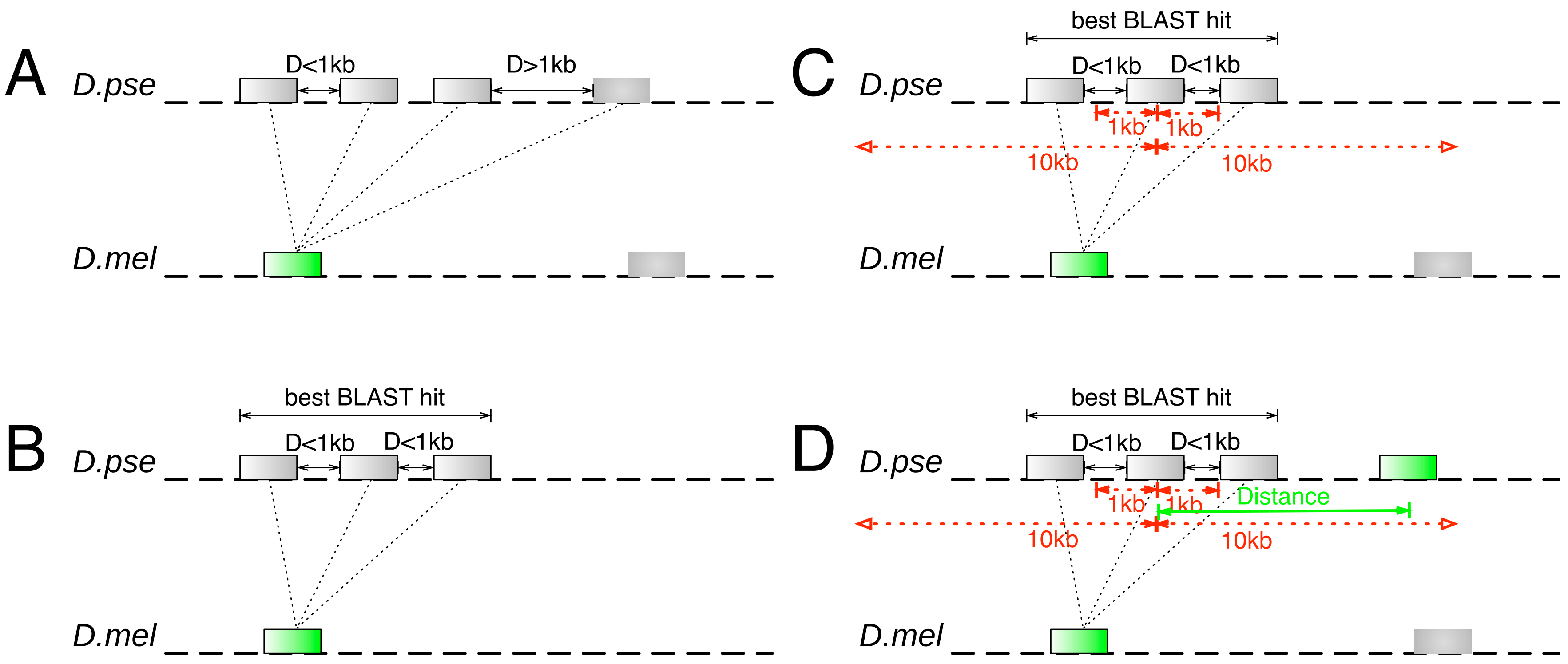


Figure 3.4.: Dynamic search Diagram. Figure3.4(A) An enhancer in query species *D.mel* could have several BLAST hits in a target species such as *D.pse*. Distances between neighbouring BLAST hits can be smaller or bigger than 1kb. Figure3.4(B) Several BLAST hits in target species *D.pse* are combined into one best BLAST hit. Figure3.4(C) Search analogous enhancer in the best BLAST hit of the target species with 1kb, 10kb radius. Figure3.4(D) Predicted analogous enhancer in target species with a distance smaller than 10kb.

query sequences. Finding multiple BLAST hits for each query enhancer is possible (see Figure 3.4(A)). (ii) If the distances between the neighboring BLAST hits are smaller than 1kb, these hits are grouped together and defined to be one best BLAST hit (see Figure 3.4(B)). (iii) After finding every query enhancer's best BLAST hit in the target species, the middle of this best BLAST hit is extended with different radius settings (1kb and 10kb) in both directions (see Figure 3.4(C)), defining the dynamic search region.

Third, the score cutoff was recalculated in the dynamic search regions. Overall searching sequence (*Ls*) defined in equation 3.7 is:

$$Ls = \frac{2 * R * N}{G} \tag{3.7}$$

R is the radius length; *N* is the number of genome-wide predicted enhancers in source species; *G* is the length of the whole genome.

According to equation 3.6 and 3.7, the smaller search region causes a lower cutoff while specificity remains the same (E-value of 1). The cutoff value of the 1kb radius region is considerably lower than the genome-wide static cutoff. Any element above the dynamic score cutoff was assigned to be a dynamically predicted enhancer.

Fourth, if no enhancer is found in a 1kb dynamic search region, the third selection step is repeated with radii of 10kb and 20kb (see Figure 3.4(D)). If still no elements can be identified, the nearest statically predicted enhancer is assigned to be a prediction hit.

In the end, for every genome-wide predicted enhancer in each of the species, the putative functional analogs were searched in each of the other species. This search was performed in both directions. With dynamic search, the number of predicted enhancers can be increased, while keeping the specificity.

The data for dynamic search is stored in a PostgreSQL database.

3.7. Evolutionary analysis

3.7.1. Distance definition

Observing the distance from the dynamically predicted enhancer to the orthologous locus helps to gain knowledge on the functional evolution of enhancers among *Drosophila* species. Three radii were set up as 1kb, 10kb and 20kb in the dynamic search.

A radius of 20kb is considered to be the upper limit between source and target elements because for larger radii the difference between dynamic search cutoff and genome-wide prediction cutoff will become almost indistinguishable. The dynamic distance definition is given by measuring the absolute value between the center of the target element and the center of the best BLAST hit (see Figure 3.4(d)). The distance distributions between each pair of species are generated by querying the PostgreSQL tables (dynamic_search) and (blasthit_best). A distance plot that represents the best BLAST hit and its nearest enhancer was drawn based on this data.

As a control, a random data distance distribution was prepared for comparison. The sequence parts themselves were not randomized by shuffling the nucleotides, but by randomizing the position of each sequence in the chromosome. This randomization approach is different from usual random sequence generating methods such as 0- or higher order Markov chains, which are not suitable in this application. The idea behind this random sequence generation method is to randomly pick sequences according to the distribution of real positional elements in the source species. The generation proceeds in this way:

First, the number of enhancers distributed in each chromosome of source species *D. melanogaster* was queried from the database table (static_pres) . Second, each random element was assumed to be 1000bp long, since the enhancer length is 500-1000bp [42]. From each chromosome, a number of random positions were generated equal to the number of real enhancers. Since the end of random position should not exceed the maximum number of base pairs available in the chromosome, the random numbers were generated from 0 to the maximum chromosome length minus 1000. The start point of the generated random sequence is this random position, and the end point is situated 1000bp further. Third, these random sequence pieces were treated as the source element in the dynamic search to look for hits in the target species *D. pseudoobscura.*

In the end, the distances distribution of the random elements to target species was added to the real species distribution and displayed in the distance plot (see Figure 2.12).

3.7.2. Enhancer gain and loss analysis

The eight species on which genome-wide enhancer predictions were performed were used for a study of further evolutionary gain and loss. *D. melanogaster* is commonly used as a model species; most of the existing experiments and available data are related to *D. melanogaster*. For this reason, it is practical to count the situations for which enhancers are present in *D. melanogaster*. Thus, the main dataset for gain and loss analysis consists of statically predicted enhancers in *D. melanogaster* and their functionally analogous elements in the target species.

The first criterion of enhancer gain and loss analysis is to check for presence of the predicted enhancers in eight species. Non-/presence of enhancers was classified by the distance definition described in the previous section. Enhancers that stay more or less in the same place or were less than 10kb away from the orthologous region, they were defined to be present (assigned 1) in the target species. If the distance is larger than 10kb, the dynamic predicted enhancer might not be the functional analog from the source species, hence enhancers were considered as non-present in the target species.

After collecting all the presence and non-presence data, a matrix for phylogeny estimation was built, in which 8 rows refer to 8 species. The number of columns depends on the number of analogous enhancers from the source species *D. melanogaster* to the target species. Maximum Parsimony (MP) [135, 136] and Maximum Likelihood (ML) [137, 138] are two common methods for phylogeny estimation and both methods perform well on average according to previous publication [108]. MP is a nonparametric, binary encoding method which finds the ancestral states that require the minimum number of steps of character changes in a given tree. Recent phylogenetic analysis have turned away from MP towards the probabilistic techniques of ML [108]. ML tests the hypotheses about trait evolution by summation of the probabilities over all possible states at each node of the tree [139]. The likelihood of the observed data is shown below:

$$L(m|d) \quad \propto \quad P(d|m) \tag{3.8}$$

$$= \sum_{a=0}^{1}\sum_{b=0}^{1} w(a)\, P(d\,|\,m, n) \tag{3.9}$$

$P(d\,|\,m)$ represents the probability of the observed data given the model of evolution. If the tree contains two nodes $n = \{a, b\}$ with the root a , data observed as $d = \{0, 1\}$ and prior weight $w(a)$; then the likelihood of the root $a = 0$ is $L(n = 0, 0\,|\,d)$ or $L(n = 0, 1\,|\,d)$; the likelihood of the root $a = 1$ is $L(n = 1, 0\,|\,d)$ or $L(n = 1, 1\,|\,d)$.

On a tree with eight *Drosophila* species, there are multiple ways to reconstruct the ancestral character states. The evolutionary analysis tool Mesquite [140] can construct ancestor states by implementing the ML method as shown in equation 3.9, and is an ideal tool for enhancer gain and loss analysis.

The phylogenetic tree representing the branching history of descent linking *Drosophila* species is required by Mesquite and the phylogeny is available on [141], values for each branch length are shown below:

```
(((((dmel:4.71283,dsec:3.65378):1.07592,dsim:7.28995):1.03928,
(dere:4.31795,dyak:4.8472):0.67972):8.69153,dana:7.31418):3.70898,
(dpse:5.45575,dper:6.20616):3.22088);
```

This *Drosophila* phylogeny together with the initially generated presence matrix were imported into the program.

MP method follows the rule of the least numbers of changes and the command for MP analysis by Mesquite is:

```
stored Trees > Trace > Reconstruction Method > Parsimony Ancestral States
```

The analysis procedure ends by generating the possible combinations of enhancer gain and loss trees. There are three ancestor states which are present (1), non-present (0) and uncertain (0/1).

The ML method finds the ancestral states that maximize the probability of the observed states under the evolutionary model [139] and the command for ML analysis by Mesquite is:

```
stored Trees > Trace > Reconstruction Method > Likelihood Ancestral States > Stored
Probability Model > Asymm.2param
```

The "Asymmetrical Markov k-state 2 parameter" model allows a bias in gains versus losses, hence it was selected for evolutionary analysis. This Asymm.2param model has two parameters: forward rate and backward rate; the instantaneous rate matrix is shown in Table 3.3. Each of the source enhancers in *D. melanogaster* (or in other words: each column of the matrix) corresponds to one tree. Every node of the tree has a proportional likelihood. If the ancestor's p-value is smaller than 0.50, most likely this enhancer was not present (0) in the ancestor species. Since the corresponding enhancer was defined to be present in *D. melanogaster*er, its tree was assigned to have gained enhancers during evolution. If the common ancestor has a p-value larger than 0.5, the statically predicted enhancers neither gained nor lost during evolution. For the trees with an ancestor p-value equal to 0.5, it is still to be decided if the enhancers from source *D. melanogaster* are either gained or lost. In this case, additional information outside of these 8 species is needed for detailed analysis.

3.7.3. Motif turnover

To analyze the evolutionary turnover of binding sites, the following case was used: when the distances between the best BLAST hits and the prediction loci are less than 1kb, the predicted

Table 3.3.: Four possible transitions between beginning and end of a branch of length t. α : the forward transition rate from $0 \rightarrow 1$. β : the backward transition rate from $1 \rightarrow 0$.

state at the beginning of branch	state at the end of branch 0	 1
0	$P_{00}(t) = 1 - P_{01}(t)$	$P_{10}(t) = \frac{\beta}{\alpha+\beta}(1 - exp[-(\alpha+\beta)t])$
1	$P_{01}(t) = \frac{\alpha}{\alpha+\beta}(1 - exp[-(\alpha+\beta)t])$	$P_{11}(t) = 1 - P_{10}(t)$

enhancers are considered to be conserved among all species. For every enhancer, its sequence from the source species *D. melanogaster* and its analogous sequences in the target species *D. yakuba*, *D. simulans* and *D. pseudoobscura* were selected from every species' genome fasta files. Analogous sequences of *D. pseudoobscura* were converted to their reverse-complement counterpart.

Next, these four pieces of sequence were input into mVISTA [112] which is a tool to align and compare sequences from multiple species. The default option LAGAN was selected for global multiple sequence alignment. The enhancer regions were classified by level of conservation: high conservation (>70%), medium conservation(50%-70%) and low conservation(<50%). The exact positions of conservation regions were stored and were later included in motif turnover plots.

The position of every single motif inside each enhancer has to be decided. Five motifs are symbolized as BCD (circle), HB (oval), KR (square), KNI (star) and CAD (triangle). For two positionally overlapping but different motifs, these were shown as one stacked on top of the other (see Figure 2.24 and Figure 2.25). For two positionally overlapping but identical motifs, only the first motif was drawn in the figure. After locating all motifs in four species, some of the motifs were observed to be highly conserved in all species; others show evolutionary turnover. The binding sites that have motif turnover are highlighted in red.

In order to analyze the feature of motif proximity, the online multiple alignment server MAVID [142] was used to generate plain aligned sequences. The nucleotides were highlighted in different colors to represent the location of each motif; BCD (yellow), CAD (blue), KR (red), KNI (green) and HB (grey) (see Figure A.2 and Figure A.3). Following the ideas presented by [143], the motifs in *D. melanogaster* were classified as "overlap", "close" or "isolate" manually. Overlapping sites share one or more nucleotides between two motifs. Close sites are within 10bp of each other. The remaining sites are isolated motifs. After the classification, every motif in *D. melanogaster* is checked for how well it aligns with sites in the three other species. If a motif is conserved in all four species, it is named an "extremely conserved" motif. If a motif is conserved in melanogaster subgroup species, it is named a "highly conserved" motif. If a motif is only conserved in *D. melanogaster* and *D. simulans*, it is named a “minimally conserved” motif, because these two species are least divergent during evolution among all the species considered here. If a motif only appears in *D. melanogaster*, it is named a "non conserved" motif. Finally, the figure explaining how proximity of binding sites influences motifs during evolution was drawn (see Figure 2.27).

3.8. Overall prediction

For the final prediction, a combined prediction was performed based on three methods: genome-wide (after region extension) in *D. melanogaster*, GO subset in *D. melanogaster* and dynamic search from source species *D. pseudoobscura* to target species *D. melanogaster*. The reason I chose dynamic search from *D. pseudoobscura* to *D. melanogaster* is that *D. pseudoobscura* and *D. melanogaster* are the most-studied *Drosophila* species and they are both frequently used to carry out comparative research. Moreover, the divergence distance from *D. melanogaster* to *D. pseudoobscura* is the furthest among the eight species studied here.

The common overlap from all three prediction methods was assigned as type A. After exclusion of all enhancers of type A, I checked the common predictions for every other method. Next, I left out one prediction and checked the overlap (type B) and difference (type C) for the remaining two predictions. This procedure was repeated until all three prediction comparisons were finished. At this point, results in three types of pairwise methods show overlaps. Bgs stands for the type of enhancers that overlap by genome-wide and subset methods. Bds is the type for enhancers in common in dynamic search and subset search. Bgd is the type of overlap in genome-wide and dynamic method.

For type A and B, if there is any base pair overlap from enhancers found by the three methods, the final location for these three grouped enhancers starts at the min beginning position and max end position for these enhancers. Any of these overlapped enhancers picked from three methods do not need to be checked again, as they are commonly predicted elements.

After excluding all type A, Bgs, Bgd, and Bds, only enhancers predicted solely by one of the three methods are left. For these three unique types of prediction, I name Cg for the genome-wide prediction method, Cd for the dynamic prediction and Cs for the subset method.

In order to clearly show the overlap in prediction results between each individual method, a venn plot was drawn showing the number of enhancers for each prediction method. I modified the function called vennX in matlab [144], redefined the preferred color for each section, saved it as file vennX_colored.m and finished drawing by applying the following command

```
vennX_colored( [ Cg Bgs Cs Bds Cd Bgd A ], Resolution )
```

Finally, the overall prediction number of enhancers should be the sum of enhancers of types A, Bgs, Bgd, Bds, Cg, Cd, Cs.

3.9. Comparison with published enhancers

Several publications have experimentally verified embryonic enhancers [1, 42, 33, 95]. I count these enhancers as the first source of published data. Since the different publications used different *Drosophila* sequence versions, an initial sequence check was done by blasting the enhancer sequences reported in the publications with sequence version 4.2.1 which I have used in the prediction, so that the correct sequence positions could be defined. Although slight differences in nucleotide position can be observed between the published data, some of the predicted enhancers can still be clearly validated with the publications despite these minor sequence differences. The second source of evaluation is from the research finished by [24], he verified embryonic bound regions of 5 transcription factors experimentally.

The following logic was used to assign an enhancer as ’published’ or ’not published’: first, if the enhancers in my prediction showed overlap with experimentally verified enhancers, these enhancers were classified as published ones. These overlaps also proved my prediction results were valid. Second, if the predicted enhancers showed overlap with experimentally verified bound regions, I grouped these enhancers into published ones as well. Although identified TFs bound regions (multiple TFBSs) are not equal to discovery of enhancers, it strongly suggests that these predicted enhancers are the most interesting candidate for functional enhancers. These bound regions were considered as strong support for evaluating new enhancers. Third, if a predicted enhancer did not have any single base pair overlap with published enhancers and TFs bound regions, these enhancers are defined as newly discovered elements.

3.10. Website

The prediction results of genome-wide, subset and dynamic search for *Drosophila* species are available online.

```
http://bibiserv.techfak.uni-bielefeld.de/jpred_en/
```

The main page presents a phylogenetic tree of *Drosophila* species as a navigation tool to explore the result sets. The web-based interface was built in Perl/CGI and calls the PostgreSQL enhancer tables. Current available database tables are:

static_pres genome-wide enhancers in eight *Drosophila* species;

nextgene genome-wide enhancers’ neighboring genes in *D. melanogaster*;

dyn_search dynamic search prediction;

blast_hits all of the BLAST hit results with different E-values;

blasthit_best the best BLAST hit in the target species;

ecalc_run description of null models;

prediction_score scores for sliding window scanning whole genome;

gene_positions gene annotation information including gene ID, name, position and chromosome;

species *Drosophila* species information include ID and name;

published_crm publication verification.

3.11. Prediction pipeline

Scripts made for embryonic enhancer analysis are shown as pipeline (see Figure 3.5).). These scripts are mainly written in Perl, R is mostly applied to draw plots. All the scripts can be accessed at /vol/fpsearch/jiading/scripts.

A brief explanation of every script is shown below:

pn_separation.R separation of positive and negative training sets based on the Berman *et al.* data

motif_weight.sh motif weight calculation

motif_weight.R motif weight plot of 5 single and 15 paired motifs

ks.pl statistical significance analysis for positive and negative training set separation

calc_cutoff.sh cutoff calculation

comprise_bands.sh get occurrences of hits in the background model

jpred_score.pl score real sequence

above_cutoff.sh get enhancer elements above score cutoff

copy_scores.pl copy scores from genome-wide prediction into jPREdictor_score table

copy_enhancers.pl copy statically predicted elements into static_pres table

copy_hits.pl copy occurrences of hits in the background model and corresponding score cutoff value into database

go_DBI.pl access embryonic terms from gene ontology database

substring.pl fetch substring sequences

dyn_search.pl dynamic search *(original scripts from Arne Hauenschild)

gene_position.pl prepare gene tables for species Dmel, Dpse, Dsim, Dyak

neighbor_gene.pl fetch enhancers' neighboring genes

shuffle_sequence.pl shuffle sequence to generate random sequence score plot

position_plasticity.pl draw score plot for enhancer positional plasticity analysis

analogous_seq.pl get statically and dynamically predicted analogous sequences

motif_position.pl find every motif position inside enhancer

motif_turnover.pl generate html file to show motif turnover and sequence alignment

motif_occur.sh get motif occurrences inside enhancer

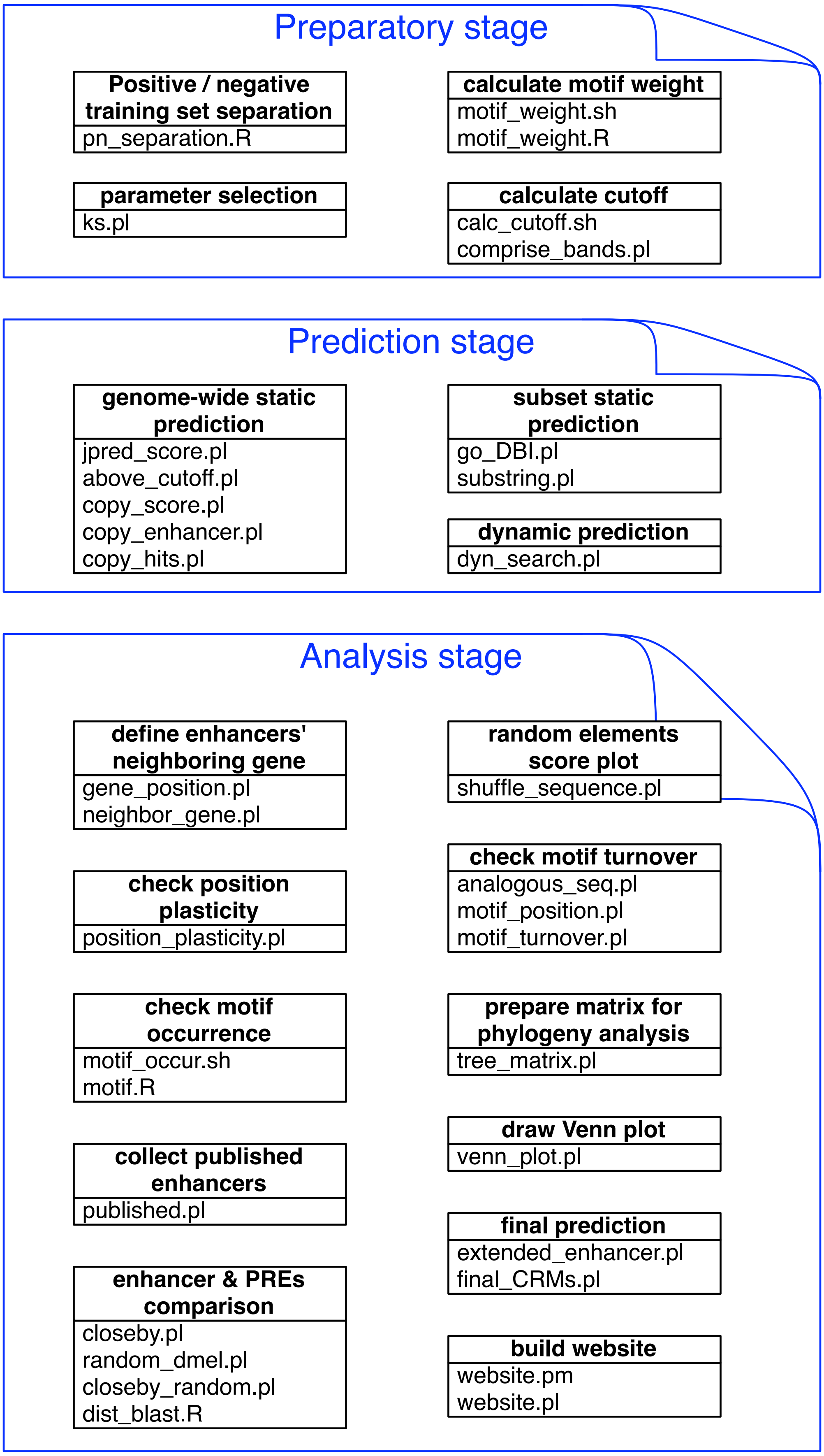


Figure 3.5.: Embryonic enhancer prediction pipeline.

motif.R draw motif occurrences plot

tree_matrix.pl generate matrix for tree construction in D.mel

published.pl prepare tables for published enhancers, sequence position matching latest flybase version

venn_plot.pl check overlaps of three prediction methods

extended_enhancer.pl extend sequence regions of genome-wide predicted enhancers

final_CREs.pl prepare final prediction, classify enhancers as new or published elements

closeby.pl check if predicted enhancers have close-by PREs

random_dmel.pl generate random positional elements by using predicted melanogaster enhancers

closeby_random.pl check if random elements have close-by PREs

dist_blast.R compare the tendency of enhancers & PREs positional plasticity

website.pl build website

4. Discussion

4.1. Comparison to other *cis*-regulatory element predictions

During the past 25 years multiple computational methods for modeling and identifying of DNA regulatory elements have been developed (see the references therein[57, 56, 58]). Although over-representation of transcription factor binding sites (TFBSs) in regulatory sequences has been intensively exploited by many algorithms, it is still a difficult problem to distinguish regulatory from other genomic DNA. In this study I have described my computational method for the systematic prediction of enhancers in eight *Drosophila* species.

The existing methods for CREs *(cis*-regulatory elements) prediction are based either on TFBS clustering or phylogenetic footprinting. The latter is based on sequence conservation. Both methods assume that CREs have common properties that can provide a signal for their identification [62, 57, 56, 145, 55, 146], as shown in the introduction chapter. I have mentioned some of the most representative prediction programs. Many different CREs prediction programs based on clustering or phylogenetic footprinting are available, every one having specialties and limitations (see Table 4.1). In this section my improvements on both classical types of method will be discussed.

4.1.1. Using paired motifs improves clustering

In this project, I have developed approaches that are appropriate from both a biological and a computational point of view have been developed to identify valid pairs of motifs.

From a biological point of view, different TFs bind to an enhancer, and each factor can bind to multiple sites within it [43]. The five motifs selected in my project are all from maternal and gap regulatory stages; some of them are functional activators (such as BCD and HB); others act as repressors (KNI, KR and CAD). An appropriate regulation of the embryonic enhancers relies on the close proximity of multiple binding sites for both activators and repressors [147, 34, 148, 35, 113].

From a computational biology point of view, first, clustering paired motifs is more advanced than using homotypic single motifs. The latter method barely obtained adequate performance in the prediction of CREs. For example, Lifanov *et al.* [67], by using homotypic regulatory clusters, only identified *eve* stripe 2 and stripe 1. Second, clustering paired motifs is more accurate than using multiple single motifs. For example, Berman *et al.* [22] identified 37 presumed regulatory elements by simply scanning for multiple individual motifs in a sequence window, however, only 15 of these are true positives. Thus, even when the motifs are known, using single motifs clustering for enhancer prediction is not accurate, and has a high potential for false positive occurrences. As it was shown in section 2.8.3, 92 enhancers were prediected with E-value of 1 (one falso positive expected) by using paired

Table 4.1.: List of CREs prediction programs

Author	**Year**	**Motif**	**Program name**	**Feature**	**Main contribution**
Berman	2002	BCD, KR, KNI, CAD, HB	CIS-analyst	web-based visualization tool	genome-wide prediction in melanogaster, find 37 embryonic enhancer
Berman	2004	BCD, KR, KNI, CAD, HB	eCIS-analyst	visualization and prediction tool	use *D. pseudoobscura* to improve prediction
Sosinsky	2007	BCD, KR, KNI, CAD, HB	EDGI	motif-discovery and local permutation-clustering algorithm	identify enhancers as evolutionarily conserved order-independent clusters of short conserved motifs
Lifanov	2003	BCD	homotypic method	single motif clustering	identification of similarly regulated genes
Li	2007		FTT-Z, YMF	motif-finding program	find biases in CREs type
Rajewsky	2002	BCD, KR, KNI, CAD, HB, GT, TLL, DLL, torRE	Ahab	thermodynamic model for CREs detection	detect clusters of weak sites
Sinha	2006		Stubb	extension of Ahab, handle two-species data within its probabilistic framework	cross-species comparison improve genome-wide prediction
Ivan	2008	not pre-required	CSam, D2Z-set	simulated annealing	prediction by scanning gene battery without knowledge of motifs

motifs. Third, a cluster of paired motifs is able to discover functional enhancers. The five motifs used in my research have to bind cooperatively to their TFs in order to become functional enhancers and thus initialize gene regulation. For example, paired motifs show good performance in the detection of *eve* stripe enhancers (see Figure 2.41). Even in this difficult case where multiple known enhancers lie in the same control region; the exact number, position and coverage of enhancers is determined. The clustered motifs correspond well with the prediction score plot (see Figure 2.30).

In short, usage of paired motifs is the first improvement to other clustering methods.

4.1.2. Dynamic search improves phylogenetic footprinting

Using phylogenetic footprinting to identify strongly conserved motifs in distant yet related species is possible, but it comprises a great risk of missing non-conserved functional regulatory elements [85]. However, a growing body of evidence suggests the biological importance of these non-conserved sequences [149, 94, 150, 59, 151, 152]. A genome-wide comparison between *Drosophila* species shows only a slight difference in conservation between known regulatory regions and other non-coding regions, illustrating the difficulty of discovering real regulatory elements [94, 59]. Thus, the importance of my work lies in showing that functionality of CREs does not necessarily follow sequence conservation.

In my work, the comparative method named as “Dynamic search” goes beyond ordinary sequence conservation, although it uses the basic alignment tool BLAST. After applying the search radius around the best BLAST locus, the cutoff value can be decreased safely; hence, this dynamic search method has the power of finding non-conserved enhancers which are up to more than 10 kilo base pairs away from the initial orthologous region.

4.2. Novel knowledge gained from embryonic enhancer prediction

In comparison to previous research, the study presented here innovates not only in the discovery of new enhancers related to embryonic development, but also in its novel understanding of evolution in regulatory elements, its exploration of enhancer plasticity, and its comparison of enhancers and PRE/TREs. It suggests that computational approaches can be an effective addition to experimental methods in the analysis of transcriptional networks. It offers new insights into of the regulatory mechanisms involved in gene expression.

4.2.1. Identification of new enhancers

The study expands our knowledge of the segmentation gene network by increasing the number of computationally discovered regulatory elements to a grand total of 135 in *D. melanogaster*.

48 of the associated genes have been shown experimentally to be required for the segmentation of the embryo [2]. Assuming that every predicted element regulates a single gene, I speculate that at least 135 genes are involved in the gene control network.

The most striking result of these new predictions is that more than half of the identified enhancers overlap with *in-vivo* binding regions from a ChIP-chip experiment [24], but not with any known regulatory elements. This strongly suggests the prediction method is good at finding real new functional elements.

4.2.2. Enhancer plasticity

It has been suggested that highly conserved elements are preferentially located in the vicinity of genes coding for transcription factors involved in early development [153, 154, 155]. For the first time, a large number of embryonic development enhancers is available for plasticity analysis. From this new data emerges the fact that enhancers are not as evolutionarily constrained as has been expected before. The three types of plasticity discovered in this study show how evolutionarily flexible enhancers can really be.

4.2.2.1. First type of plasticity - prediction number plasticity

Feature The numbers of predicted enhancers differ dramatically between multiple *Drosophila* species in single genome-wide analysis. According to my prediction result, *D. melanogaster* has more embryonic developmental enhancers than the three other species. I speculate the prediction shows around twice as many hits in *D. melanogaster* as in *D. simulans*, *D. yakuba*, and *D. pseudoobscura* although the exact numbers of embryonic enhancers are unknown right now. Moreover, it seems this difference in number of predictions is not correlated to evolutionary distances among species. Although *D. simulans* and *D. yakuba* are both closely related to *D. melanogaster* and all belong to the melanogaster subgroup, they do not have the same number of regulatory elements. This indicates that not only the numbers of enhancers differ among species, but also the numbers of the genes they regulate, especially the same gene might have different numbers of enhancers sometimes.

Explanation A possible biological reason for this observation could be that *D. melanogaster* gained several enhancers to regulate the neighboring genes expression pattern; or it could be that an orthologous gene is regulated by a different number of embryonic enhancers in different species. There are examples for this latter suspicion: the genes *nub* and *run* both have two enhancers in *D. melanogaster*. This may be explained by varying gene response dependent on different concentration levels of TFs, resulting in a different number of required binding sites, thus in a different number of enhancers [156].

Value of research A similar genome-wide analysis on developmental enhancers was done by Berman *et al.* [22, 1], 37 enhancers were tested in *D. melanogaster*. The authors assumed enhancers in *D. pseudoobscura* are conserved with their corresponding element in *D. melanogaster*. In their analysis, the enhancers predicted in *D. pseudoobscura* are simply the aligned sequences from the prediction in *D. melanogaster*. Thus, the authors logically expected the number of enhancers in *D. melanogaster* and *D. pseudoobscura* to be the same. Moreover, the authors did not make predictions for the other *Drosophila* species.

Contrary to Berman, my analysis was performed at the genome-wide level, was not based on alignment, and included *D. yakuba* and *D. simulans*, doubling the number of species considered. Due to this higher number of species and non-reliance on alignment, it was possible to establish a more reliable comparison of prediction numbers. I observed that the number of enhancers varies among species. This novel discovery was named 'first type of plasticity'.

4.2.2.2. Second type of plasticity - prediction position plasticity

Feature From the analysis of enhancer position plasticity, embryonic enhancers may be organized in two groups of regulatory elements: one group consists of enhancers that have conserved their position in evolution, the other group contains those enhancers that have moved.

Explanation Even though conservation of sequence is a potent indicator of function, *cis*-regulatory sequences with conserved regulatory function are sometimes no longer alignable due to strong divergence [157, 158, 159, 152]. This may explain at a biological level why the phenomenon of positional plasticity is possible to occur in the first place.

Enhancers may appear to move because of a loss in one site and a gain in another. For example, enhancers 19115 and 19190 in *D. pseudoobscura* seem to move several kb away from their orthologous BLAST hits in *D. melanogaster* (see section 2.5.2). The enhancer at the original orthologous site appears lost in *D. pseudoobscura*, but a newly positioned enhancer may create a complement for the same gene expression pattern. Because of this functional complementation, such evolved changes in the enhancers in *D. pseudoobscura* may have little or undetectable impact on spatiotemporal control of gene expression, making observation of this interplay challenging [95, 42, 143].

If enhancer position appears to be flexible, it begs the question: how can enhancers maintain their role in regulation? Recent reports suggest that enhancer DNA may be neighboring the promoter while the in-active intervening sequence is "looped" out [160, 161]. Such a looping model brings activator proteins bound to distant enhancer elements into proximity with TFs complexes interacting with associated sequences. Although these embryonic enhancers may move away from the orthologous sites, distant enhancers could regulate genes without altering the level of expression. Nevertheless, experimental verification is necessary to get a better understanding of such enhancer position plasticity.

Value of research Up until now, most of the enhancers that have been identified were found based on the assumption that they are highly conserved. In reality, the regulatory network in the early developmental stage is much more flexible in evolutionary terms than expected. Thus a traditional phylogenetic footprinting method may not be suitable for the detection of developmental enhancers.

Attempts to overcome this issue have been made before, e.g. Sosinsky *et al.* [162] suggested an alignment-free method for enhancer identification. Instead of applying a non-alignment method on the CREs containing multiple heterotypic TFBSs, the author focusses on detecting individual TFBSs. Strictly speaking, since TFBSs are only a component of enhancers, the author's method is for alignment-free motif discovery, not for complete enhancer identification.

In short, the value of discovering positional plasticity lies in the fact that it act as a reminder to take this weakness of phylogneetic footprinting into account when attempting genome-wide regulatory element discovery. Further research of positional plasticity may lead to new insights into the mechanisms of gene regulation.

4.2.2.3. Third type of plasticity - motif turnover

Feature After comparing composition and organization of multiple motifs in enhancers, I observed that motif turnover is prevalent even for positionally constrained enhancers. Motif turnover also happens in closely related species and such turnover is more rapid for further divergent species of *Drosophila*.

Additionally, after motif proximity analysis, I observed that overlapping or closely paired motifs are evolutionarily more constrained and that hence it is harder for motif turnover to occur than for isolated motifs. A similar analysis has been done before[143], my results remain consistent with these observations even after using different enhancers in different species.

Explanation First, the effects of insertion and deletion constitute a major cause of sequence variation in *Drosophila* and thus insertion and deletion can be considered a major contributing factor to binding site turnover [163]. The loss of motifs in one region of an enhancer may be dampened by the occurrence of new mutations, creating a complementary site elsewhere in the same enhancer [143]. Second, specific factors may influence the conservation and turnover of binding sites. Sometimes TFs bind cooperatively, other times they bind competitively to motifs [164, 34, 156]; thus motifs might have to relocate in order to conserve the site's binding affinity for the transcription factors. Third, only 25% of non-conserved binding sites are estimated to be functional [115]. If functional binding sites happen to be inside functional enhancers [24], I might assume that the non-functional motifs just happen to be present inside the regulatory region by chance, thus they are free for rearrangement or gain and loss across species without affecting regulatory function.

One explanation of low turnover in closely paired motifs can be found in motif proximity analysis: if a site is close to or overlapping with another site, such paired sites may be less likely to be affected by turnover during evolution, as the cooperative activity of proximal motifs leads to a stronger constraint during evolution. A different way of explaining low turnover in closely paired motifs is by considering the influence of random mutations: random mutations are far less likely to produce pairs of adjacent sites than single sites[143].

Value of research In my study, I extended the number of predicted enhancers and performed a systematic exploration of motif turnover in CREs sequences. Earlier attempts to characterize the evolutionary patterns of motifs used a few well-studied *even-skipped* enhancers. These studies were thus limited to a small range of CREs[42, 143].

I also demonstrated that the degree of motif turnover does not influence the previously established hypothesis that proximate motifs are more likely to be conserved during evolution. For example, a large fraction of the binding sites in *eve* enhancers are conserved across *Drosophila* species and their rate of motif turnover is much less than for the enhancer *ftz*. A repetition of the paired motif proximity analysis specifically for *ftz* shows a high consistency with the result for *eve* enhancers, showing that the rate of motif turnover is not influenced by motif proximity.

4.2.3. Discussion on enhancer gain and loss

Feature and explanation In my study, 9 enhancers out of 92 (10%) in *D. melanogaster* have most likely evolved from non-functional sequences. It has been shown theoretically that the processes of mutation and selection can quickly produce TFBSs in enhancers even from random DNA. Usually, these motifs differ from functional binding sites only by a single substitution, in which case they are called presites [165]. These presites may be understood as a precursor of enhancer gain from random DNA. In this context, the enhancer gain itself can be seen as compensation against evolutionary change or as establishment of new function.

First, the process of compensation against evolutionary change may explain the fact that only one out of these 9 enhancers was gained from genes that already have a single regulatory element. Random mutation on presites in the vicinity of existing enhancers can reasonably be expected to create a new enhancer. This new enhancer may not compromise existing regulatory function directly, but effects the existing enhancer to perform its function less well while at the same time compensating for this loss with its own increased regulation; the combined roles of these enhancers canceling out the effect of random mutation [166].

Such a compensatory theory might also explain the evolution of stripe enhancers in multiple species. In the score plot of enhancer stripe 3/7 (see Figure 2.26), there is always a single score peak in *D. melanogaster*, *D. simulans* and *D. pseudoobscura*. But, surprisingly, I observed a tendency of two peaks in one sequence region in *D. yakuba*. One could speculate that stripe 3 and 7 are both regulated by one enhancer in *D. melanogaster*, *D. simulans* and *D. pseudoobscura*. However, in *D. yakuba*, my observation suggests that the ancestral enhancer has split into now-separate elements to govern one gene expression.

Second, the establishment of new function may explain the observation that 8 enhancers are associated with genes that previously had no enhancers. From this observation, it seems more probable that a novel gene expression pattern will arise from these evolutionary changes. Random mutations may acquire all of the necessary binding sites and eventually form enhancers, even if the process takes a long time and is relatively difficult.

Value of research Until very recently, there have been very few direct empirical examples linking CREs evolution to morphological evolution [167, 168]. This might be because it is relatively complicated to experimentally verify changes in the very early gene expression patterns. Thus, most of the research has been undertaken on later stages, for example body pigmentation by the Abdominal-B Hox protein and its gain and loss in *Drosophila* evolution [169, 170, 168]

My analysis of embryonic enhancer gain and loss helps to grow knowledge in the early morphological stage and draws the focus away from pure TFBSs gain and loss analysis. The latter method is appropriate for individual binding site studies, but enhancer gain and loss is better understood when the regulatory function of whole components instead of only motifs is considered.

4.2.4. Are embryonic enhancers more constrained than PREs?

Recently, a similar plasticity analysis has been carried out on other *cis*-regulatory elements, the Polycomb/Trithorax response elements (PRE/TREs), which maintain transcription states at later stages after the enhancers have established the developmental gene expression [96].

Embryonic enhancers and PRE/TREs act at different developmental stages and have an entirely different composition of motifs; a systematic comparison helps to gain knowledge on evolutionary plasticity overall.

Feature and explanation The numbers of predicted embryonic enhancers is less than 100 elements. However, the number of the predicted PRE/TREs exceeds 500 elements. The largest number of embryonic enhancers is found in *D. melanogaster*. The situation is just the opposite in PRE/TREs which have the smallest number of prediction in *D. melanogaster*. The number of PRE/TREs in *D. pseudoobscura* is twice as many as embryonic enhancers in *D. melanogaster*.

One might argue that these differences reflect the functional difference between embryonic enhancers and PRE/TREs: activation/repression vs. maintenance. The early regulation network is hierarchical. The evolutionary modifications in regulatory connections occur not just between two consecutive levels. Regulatory evolution takes advantage of transcription factors throughout multiple genetic hierarchies to generate new regulatory connections [168]. Thus, more PRE/TREs might be needed to maintain every regulation status (such as maternal or gap stage) from embryonic enhancers. Nevertheless, it is not clear why *D. pseudoobscura* has the least number of developmental enhancers but the largest amount of PRE/TREs.

Additionally, the number of analogous enhancers within close distances from their BLAST locus is less than the number of analogous PRE/TREs (see Figure 2.44). Thus, I probably can speculate that enhancers display a similar evolutionary plasticity as PRE/TREs, but enhancers have an even higher rate of plasticity.

Value of research Previous studies often suggest that during evolution enhancers of developmental genes may be less sensitive to mutation than PREs [96, 82]. My observation is just the opposite. A different study from [7] might support my results. The degree of conservation of genes that function in successive steps of the segmentation cascade is various and is represented by the width of the hourglass in Figure A.4. The earliest stage of the cascade, determined by maternal gradients, has diverged significantly between arthropod groups and has the lowest rate of conservation. Gap gene homologues can be found in all arthropods, but their function in segmentation is variable. The genes in the later cascades are often more conserved, both molecularly and functionally. My predictions are mainly made at early cascades (maternal and gap), thus, evolution is much more rapid and flexible than expected.

4.3. Outlooks and conclusions

Although a good prediction has been finished in this project, there remains much knowledge to be gained before a complete understanding of these developmental networks can be reached. Some aspects which are worth further exploration are listed below.

4.3.1. Impact of motif selection on CREs identification

It is widely known that BCD, CAD, HB, KR and KNI act together as maternal or gap factors. Different studies include slightly different motifs for embryonic enhancers. For example, Li *et al.* [24] used one additional gap factor GT; Schroeder *et al.* [2] inputs two more maternal factors: the Torso-response element (TorRE), Stat92E (D-Stat); and one gap factor Tailless (TLL).

By scanning through the two training sets with the tool MatInspector, STAT's binding site consensus sequence-TTCCCGGAA- was spotted significantly in the positive training set. Adding such a motif as a co-acting factor to BCD, KR, KNI, HB and CAD gives a better separation between the two training sets (see Figure A.5). Especially, given that the literature [171] supports that the JAK-STAT pathway is connected to early *Drosophila* development. Adding additional collections of motifs might improve prediction but until further understanding of biological function and the impact of adding more motifs is gained, I elect to err on the conservative side and limit my choice of motifs to these five well-established ones even though adding STAT may be valid once more experimental validation is performed.

4.3.2. Properties of motif proximity

My motif turnover analysis confirms Hare *et al.*'s [143] observation that paired proximate motifs are more likely to be conserved than isolated motifs. In a future study, it would be interesting to know if such a proximity property is limited to embryonic enhancers, or if it is a general principle which may also apply to completely different *cis*-regulatory elements (e.g. PRE/TREs). Furthermore, the individual interplay between proximate motifs may need to be detailed. For example, both KR and TLL are repressors, but TLL is more conserved if it is adjacent to some other site, while KR is more conserved if it overlaps with another site [156]. If exact knowledge of proximity dependent motif conservation is gained, a putative answer may be derived to the question of how ultra-conserved regulatory elements can be maintained by evolution.

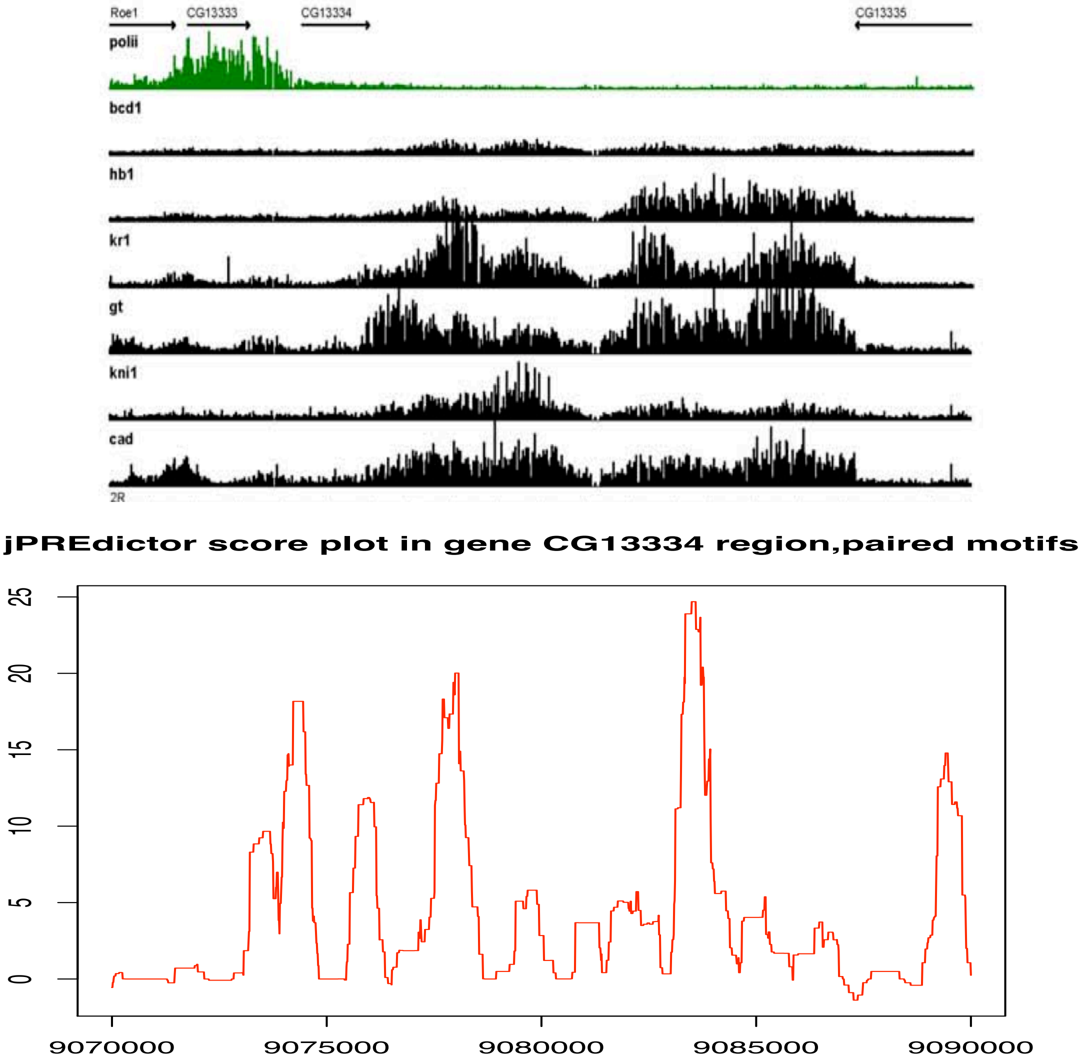
4.3.3. Applications outside of *Drosophila* embryogenesis

I believe that the computational methods presented in this thesis can be used outside of the research area of developmental enhancers in *Drosophila* species. For example, my approach can be applied to identify tailless enhancers from the house fly Musca domestica [158] and the single-minded enhancer from the mosquito Anopheles gambiae [172] because they drive similar patterns as their endogenous orthologs in *D. melanogaster* embryos[143].

4.3.4. Conclusions

In this study I have described computational methods for the systematic prediction of enhancers in multiple *Drosophila* species. In addition to the discovery of many new elements, I gained specific biological insights into embryonic development. The evolution of regulatory sequences is a dynamic process. Progress has been made in understanding enhancer gain and loss, which is important for understanding the full picture of enhancer evolution, the origins of diversity and the mechanisms of gene regulation in multiple species. Ultimately, follow-up experiments are a necessary step towards a comprehensive understanding of enhancer evolutionary plasticity in multiple animal genomes.

A. Supplementary tables and figures



jPREdictor score plot in gene CG13334 region,paired motifs

Figure A.1.: ChIP-chip oligonucleotide ratio score plot and prediction score plot near gene *CG13334*. Genes *CG13333* and *CG13334* that have unknown function in the early embryo (second image - score plot) but are bound at moderate to high levels by multiple gap factors (first image - ChIP-chip score plot).

Table A.1.: List of enhancers belong to first category of enhancer positional plasticity. SS: source species; TS: target species; Dm: *D. melanogaster*; Dp: *D. pseudoobscura*; Dist: distance from middle of best BLAST hit to mid of enhancer.

ID	SS	TS	Chr	position	Score	Dist	E/T/N	Gene	Flybase annotation
19111	Dm	Dp	2L	3608980_3610169	84.44	79	En	*odd*	embryonic segment
19113	Dm	Dp	2L	8811780_8812799	55.69	164	TFBS	*SoxN*	embryonic nervous system development
19114	Dm	Dp	2L	12567880_12568899	55.96	24	TFBS	*bun*	regulation of developmental process
19119	Dm	Dp	2L	15809230_15809929	50.41	151	TFBS	*CG18109, CG18518*	unkown
19120	Dm	Dp	2L	16526490_16527529	60.04	61	TFBS	*Tpr2, CG5953, CG31816*	unkown
19121	Dm	Dp	2L	17282550_17283339	50.33	75	TFBS	*beat-IIIc, CG6380*	unkown
19123	Dm	Dp	2L	17844990_17846139	65.99	91	TFBS	*CadN2, CG5674*	unkown
19127	Dm	Dp	2R	5486930_5488089	66.78	61	En	*eve*	embryonic development
19128	Dm	Dp	2R	5489450_5490619	59.56	166	En	*eve*	embryonic development
19129	Dm	Dp	2R	5498330_5499059	50.21	123	En	*eve*	embryonic development
19136	Dm	Dp	3L	3497800_3498519	51.38	407	TFBS	*Eip63E*	embryonic development via the syncytial BLASToderm
19140	Dm	Dp	3L	7173460_7174589	62.56	218	TFBS	*CG32387, CG14826*	unkown
19146	Dm	Dp	3L	8639310_8641549	80.02	648	En	*hairy*	embryonic pattern specification
19147	Dm	Dp	3L	8644750_8645459	51.83	43	En	*hairy*	embryonic pattern specification
19149	Dm	Dp	3L	10303270_10304349	61.89	68	TFBS	*CG6559, CG12362*	unkown
19153	Dm	Dp	3L	18372190_18373109	53.01	31	TFBS	*rpr*	regulation of developmental process
19154	Dm	Dp	3L	20630320_20631769	77.27	51	En	*kni*	anterior/posterior axis specification
19157	Dm	Dp	3R	2693200_2694579	105.71	131	En	*ftz*	embryonic development
19160	Dm	Dp	3R	14744960_14745979	52.98	174	New	*CG17836*	regulation of transcription

Table A.2.: List of enhancers belong to first category of enhancer positional plasticity. SS: source species; TS: target species; Dm: *D. melanogaster*; Dp: *D. pseudoobscura*; Dist: distance from middle of best BLAST hit to mid of enhancer.

ID	SS	TS	Chr	position	Score	Dist	E/T/N	Gene	Flybase annotation
19161	Dm	Dp	3R	14996730_14997939	76.85	279	TFBS	*sqz*	neuron development
19162	Dm	Dp	3R	15192230_15193379	76.98	121	TFBS	*Dl*	regulation of developmental process
19166	Dm	Dp	3R	26504060_26504899	53.79	260	TFBS	*CG15541, CG1342*	unkown
19169	Dm	Dp	X	2286590_2288039	101.86	126	En	*gt*	zygotic determination of anterior/posterior axis, embryo
19174	Dm	Dp	X	4576440_4577669	60.24	106	New	*CG6986, CG12683*	unkown
19177	Dm	Dp	X	5056070_5057019	50.13	410	TFBS	*rg*	adult segment
19182	Dm	Dp	X	7501780_7502599	54.76	51	New	*ct*	organ morphogenesis
19183	Dm	Dp	X	8739300_8740269	59.87	212	TFBS	*Moe*	anterior/posterior axis specification
19193	Dm	Dp	X	13379350_13380279	67.84	3	New	*CG15757, CG11816*	unkown
19199	Dm	Dp	X	20462340_20463509	54.00	161	En	*Cyp6v1*	unkown
19200	Dm	Dp	X	20495040_20496589	76.59	319	En	*run*	reproductive developmental process
19315	Dp	Dm	2	4863830_4864959	56.22	544	New	*Tl*	embryonic pattern specification
19317	Dp	Dm	2	11144450_11145239	50.84	181	New	*sim*	embryonic neuron;system development
19320	Dp	Dm	2	19675860_19676759	60.34	12	En	*ftz*	embryonic development
19321	Dp	Dm	2	27179860_27181039	77.65	477	En	*hb*	regulation of developmental process
19332	Dp	Dm	4_group2	876960_877969	61.05	217	En	*prd*	periodic partitioning by pair rule gene
19333	Dp	Dm	4_group3	3353800_3354859	54.28	192	En	*nub*	embryonic neuron
19334	Dp	Dm	4_group3	5132300_5133509	76.52	152	En	*odd*	embryonic segment

Table A.3.: List of enhancers belong to first category of enhancer positional plasticity. SS: source species; TS: target species; Dm: *D. melanogaster*; Dp: *D. pseudoobscura*; Dist: distance from middle of best BLAST hit to mid of enhancer.

ID	SS	TS	Chr	position	Score	Dist	E/T/N	Gene	Flybase annotation
19338	Dp	Dm	4_group3	10028910_10029809	50.85	371	New	*CG31925, CG31928*	unkown
19339	Dp	Dm	4_group3	10718810_10720049	80.17	98	TFBS	*CadN*	embryonic nervous system
19346	Dp	Dm	XL_group1a	3496050_3496779	52.07	24	En	*gt*	zygotic determination of anterior/posterior axis, embryo
19350	Dp	Dm	XL_group1e	9165940_9167159	76.61	440	En	*CG11692, Cyp6v1*	unkown
19351	Dp	Dm	XL_group1e	9185040_9186089	57.46	384	En	*run*	embryonic segment
19353	Dp	Dm	XL_group3a	1372350_1373509	59.53	381	New	*hep*	embryonic development via the syncytial BLASToderm
19355	Dp	Dm	XR_group6	6956520_6957799	82.19	622	New	*fz*	organ morphogenesis
19356	Dp	Dm	XR_group6	7470390_7471199	50.34	63	TFBS	*CG32423, CG10677, CG4669*	unkown
19358	Dp	Dm	XR_group6	8560780_8561539	50.75	159	TFBS	*dpr10, CG6628, ect*	unkown
19363	Dp	Dm	XR_group8	5144430_5145149	50.01	360	TFBS	*vvl*	dendrite morphogenesis
19365	Dp	Dm	XR_group8	8115190_8116109	53.11	41	En	*kni*	anterior/posterior axis specification
19366	Dp	Dm	XR_group8	8931800_8932509	50.04	554	En	*hairy*	embryonic pattern specification
19367	Dp	Dm	XR_group8	8936570_8938419	72.12	649	En	*hairy*	embryonic pattern specification

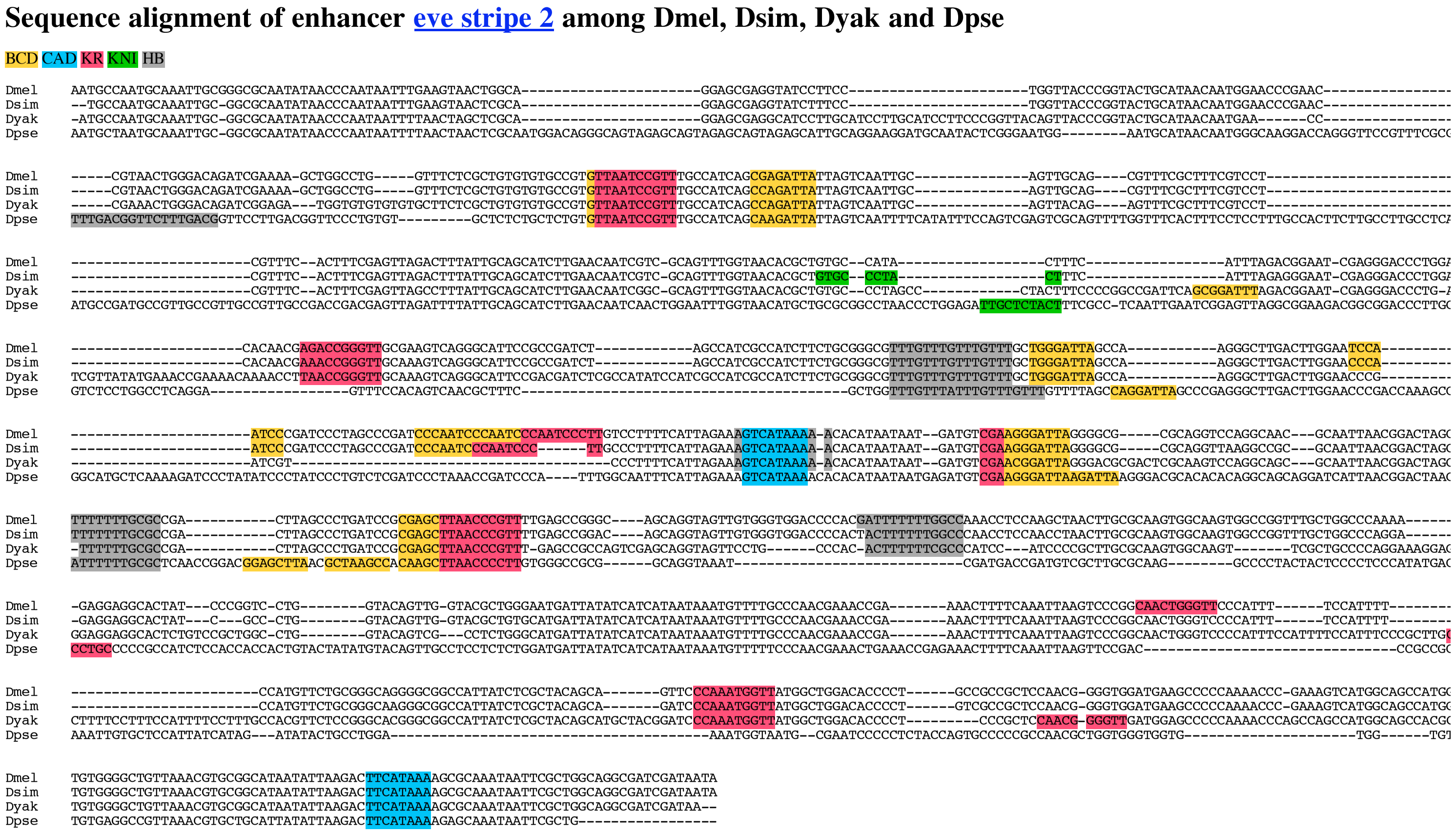

Sequence alignment of enhancer eve stripe 2 among Dmel, Dsim, Dyak and Dpse
BCD
CAD
KR
KNI
HB
Dmel
Dsim
Dyak
Dpse

Figure A.2.: Motif positions in aligned eve stripe 2 enhancer. The nucleotides were highlighted in different colors to represent the location of each motif; BCD (yellow), CAD (blue), KR (red), KNI (green) and HB (grey).

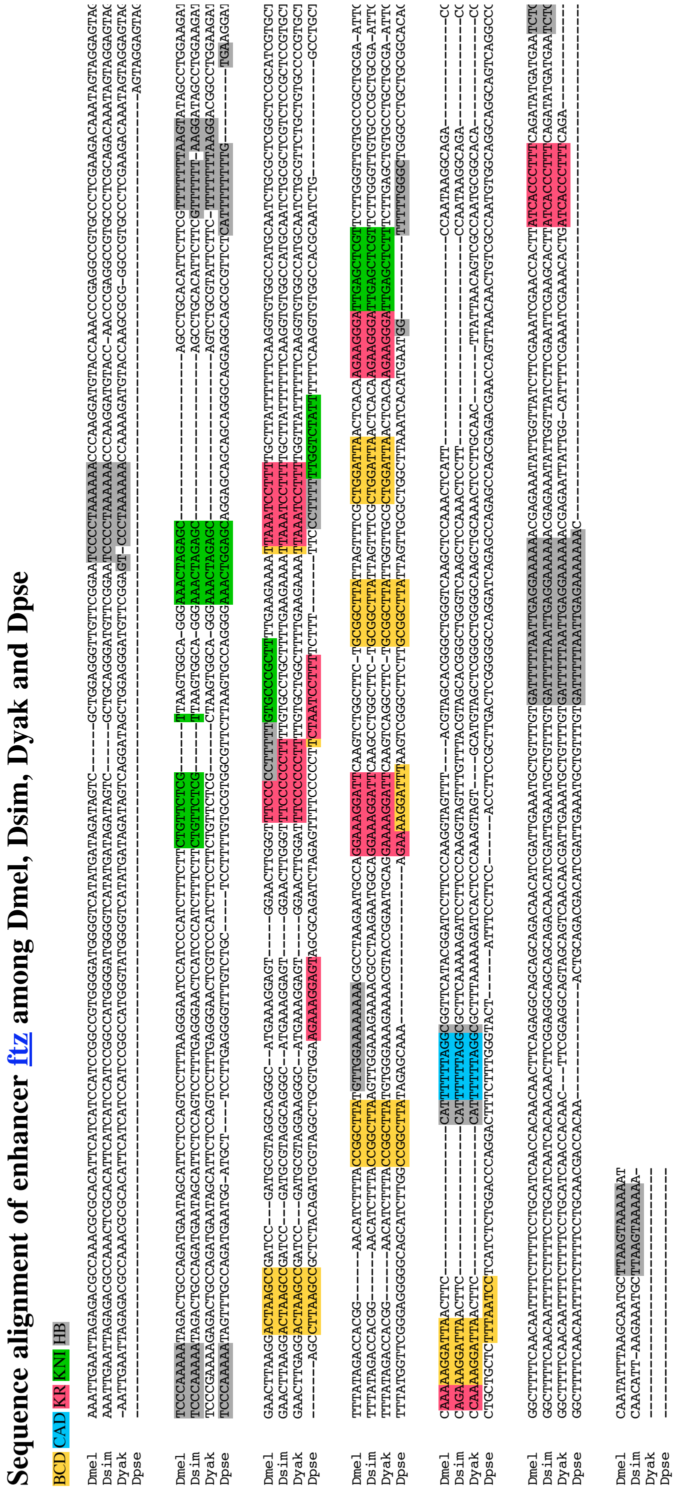

Sequence alignment of enhancer ftz among Dmel, Dsim, Dyak and Dpse
BCD CAD KR KNI HB
Dmel
Dsim
Dyak
Dpse

Figure A.3.: Motif positions in *ftz* enhancer after sequence alignment. The nucleotides were highlighted in different colors to represent the location of each motif; BCD (yellow), CAD (blue), KR (red), KNI (green) and HB (grey).

Table A.4.: Terms from the process Ontology with p-value<=0.01, category of *in-vivo* motif binding enhancers.

Term	p-value	Annotated Genes
morphogenesis of an epithelium	2.97E-07	*CG14026, CG7527, CG17697, CG31349, CG7892, CG3936, CG14728, CG10619, CG5462, CG7100, CG10701, CG7524, CG10579, CG14426, CG7734, CG3619*
post-embryonic development	1.05E-06	*CG1934, CG14026, CG2102, CG17697, CG5799, CG4319, CG18076, CG14728, CG9224, CG7734, CG3758, CG3653, CG7892, CG12154, CG3936, CG5461, CG5462, CG10701, CG2096, CG10579, CG11354, CG3619*
metamorphosis	1.00E-05	*CG1934, CG14026, CG17697, CG5799, CG4319, CG18076, CG9224, CG7734, CG3758, CG7892, CG12154, CG3936, CG5461, CG5462, CG2096, CG10701, CG10579, CG11354, CG3619*
instar larval or pupal development	1.61E-05	*CG1934, CG14026, CG17697, CG5799, CG4319, CG18076, CG14728, CG9224, CG7734, CG3758, CG7892, CG12154, CG3936, CG5461, CG5462, CG10701, CG2096, CG10579, CG11354, CG3619*
instar larval or pupal morphogenesis	2.86E-05	*CG1934, CG14026, CG17697, CG5799, CG4319, CG18076, CG9224, CG7734, CG3758, CG7892, CG12154, CG3936, CG5461, CG5462, CG2096, CG10701, CG11354, CG3619*
post-embryonic morphogenesis	3.38E-05	*CG1934, CG14026, CG17697, CG5799, CG4319, CG18076, CG9224, CG7734, CG3758, CG7892, CG12154, CG3936, CG5461, CG5462, CG2096, CG10701, CG11354, CG3619*
fusion cell fate specification	3.56E-05	*CG3758, CG31349, CG3936, CG3619*
imaginal disc morphogenesis	5.08E-05	*CG1934, CG14026, CG17697, CG5799, CG18076, CG9224, CG7734, CG3758, CG7892, CG12154, CG3936, CG5462, CG2096, CG10701, CG11354, CG3619*
establishment or maintenance of cell polarity	9.97E-05	*CG7527, CG17697, CG31349, CG7892, CG3936, CG4626, CG7100, CG5462, CG10701, CG3619*
anatomical structure morphogenesis	0.000185038	*CG14026, CG7527, CG18657, CG17697, CG5799, CG4319, CG10037, CG6775, CG14728, CG7734, CG10901, CG3758, CG7892, CG12154, CG3936, CG4626, CG10619, CG11711, CG5461, CG2096, CG8705, CG7524, CG14426, CG3619, CG1934, CG18076, CG8896, CG7100, CG9224, CG3653, CG31349, CG12653, CG5557, CG5462, CG10701, CG10579, CG18024, CG11354*
imaginal disc development	0.000319691	*CG1934, CG14026, CG17697, CG5799, CG18076, CG9224, CG7734, CG3758, CG7892, CG12154, CG3936, CG10619, CG4220, CG5462, CG2096, CG10701, CG11354, CG3619*

Table A.5.: Terms from the process Ontology with p-value<=0.01, category of *in-vivo* motif binding enhancers.

Term	p-value	Annotated Genes
cellular developmental process	0.00034942	*CG14026, CG2102, CG7527, CG18657, CG17697, CG4319, CG10037, CG6775, CG10901, CG3758, CG7892, CG12154, CG3936, CG4626, CG10619, CG11711, CG5461, CG8705, CG7524, CG14426, CG3619, CG33175, CG18076, CG7100, CG9224, CG15489, CG31349, CG3653, CG5557, CG5462, CG10701, CG18024*
nervous system development	0.000577705	*CG14026, CG2102, CG7527, CG18657, CG17697, CG4319, CG18076, CG10037, CG14728, CG7100, CG7734, CG3758, CG15489, CG12154, CG3936, CG5557, CG10619, CG12287, CG5461, CG10701, CG8705, CG7524, CG18024, CG3619*
embryonic morphogenesis	0.000803918	*CG14026, CG31349, CG12653, CG3936, CG14728, CG10619, CG8896, CG5462, CG9224, CG7524, CG14426, CG7734*
epithelial cell type specification, open tracheal system	0.000867576	*CG3758, CG31349, CG3936, CG3619*
organ morphogenesis	0.000909882	*CG1934, CG14026, CG7527, CG17697, CG5799, CG18076, CG6775, CG7100, CG9224, CG7734, CG3758, CG7892, CG12154, CG3936, CG10619, CG5461, CG5462, CG10701, CG2096, CG8705, CG11354, CG3619*
cell fate commitment	0.000931147	*CG14026, CG2102, CG17697, CG3758, CG31349, CG15489, CG18076, CG10037, CG3936, CG10619, CG5461, CG10901, CG3619*
anatomical structure development	0.000961004	*CG14026, CG2102, CG7527, CG18657, CG17697, CG5799, CG4319, CG10037, CG6775, CG14728, CG7734, CG10901, CG15284, CG3758, CG8930, CG7892, CG12154, CG3936, CG4626, CG11711, CG10619, CG12287, CG5461, CG2096, CG8705, CG7524, CG14426, CG3619, CG1934, CG18076, CG8896, CG7100, CG9224, CG3653, CG15489, CG31349, CG12653, CG5557, CG4220, CG5462, CG10701, CG10579, CG11354, CG18024*
cell fate determination	0.00271111	*CG14026, CG2102, CG17697, CG15489, CG18076, CG10037, CG3936, CG5461, CG10901*
cell-cell adhesion	0.003273342	*CG7527, CG17697, CG32387, CG7100, CG3653, CG31349, CG3936*
ommatidial rotation	0.003333344	*CG7527, CG17697, CG7100, CG7892, CG3619*
regulation of cellular process	0.00371927	*CG1343, CG14026, CG32678, CG2102, CG17697, CG5799, CG4319, CG10037, CG7734, CG10901, CG15284, CG3758, CG8930, CG7892, CG12154, CG3936, CG4626, CG10966, CG10619, CG12287, CG5461, CG2096, CG8705, CG7524, CG14426, CG17888, CG3619, CG33175, CG18076, CG8896, CG7100, CG9224, CG6380, CG15489, CG31349, CG12653, CG32683, CG5557, CG18646, CG4220, CG13701, CG5462, CG10701, CG11354, CG18024*

Table A.6.: Terms from the process Ontology with p-value<=0.01, category of *in-vivo* motif binding enhancers.

Term	p-value	Annotated Genes
morphogenesis of embryonic epithelium	0.004074842	*CG14026, CG31349, CG3936, CG14728, CG10619, CG5462, CG7524, CG14426, CG7734*
central nervous system development	0.00467819	*CG2102, CG3758, CG15489, CG4319, CG18076, CG12154, CG3936, CG14728, CG7524, CG18024, CG3619*
dorsal closure	0.004826394	*CG14026, CG31349, CG3936, CG14728, CG10619, CG5462, CG7524, CG7734*
cell communication	0.004929378	*CG33175, CG14026, CG32678, CG17697, CG4319, CG18076, CG8896, CG7100, CG9224, CG6380, CG7734, CG15284, CG31349, CG8930, CG7892, CG32683, CG3936, CG4626, CG10966, CG18646, CG10619, CG5462, CG5461, CG4220, CG10701, CG2096, CG7524, CG10617, CG4349, CG18024, CG3619*
organ development	0.005252955	*CG1934, CG14026, CG2102, CG7527, CG17697, CG5799, CG4319, CG18076, CG6775, CG14728, CG7100, CG9224, CG7734, CG3758, CG3653, CG15489, CG7892, CG12154, CG3936, CG5557, CG10619, CG5462, CG5461, CG4220, CG10701, CG2096, CG8705, CG7524, CG11354, CG3619*
regulation of biological process	0.005712321	*CG1343, CG14026, CG32678, CG2102, CG17697, CG5799, CG4319, CG10037, CG7734, CG10901, CG15284, CG3758, CG8930, CG7892, CG12154, CG3936, CG4626, CG10966, CG10619, CG12287, CG5461, CG2096, CG8705, CG7524, CG14426, CG17888, CG3619, CG33175, CG18076, CG8896, CG7100, CG9224, CG4346, CG6380, CG15489, CG31349, CG12653, CG32683, CG5557, CG18646, CG4220, CG13701, CG5462, CG10701, CG11354, CG18024*
establishment of ommatidial polarity	0.005758217	*CG7527, CG17697, CG7100, CG7892, CG3936, CG3619*
morphogenesis of a polarized epithelium	0.008536607	*CG7527, CG5462, CG17697, CG7100, CG7892, CG3936, CG3619*
homophilic cell adhesion	0.008564357	*CG7527, CG17697, CG32387, CG7100, CG3653*
biological regulation	0.008745426	*CG1343, CG14026, CG32678, CG2102, CG17697, CG5799, CG4319, CG10037, CG14728, CG7734, CG10901, CG15284, CG3758, CG8930, CG7892, CG12154, CG3936, CG4626, CG10966, CG10619, CG12287, CG5461, CG2096, CG8705, CG7524, CG4349, CG14426, CG17888, CG3619, CG33175, CG18076, CG8896, CG7100, CG9224, CG4346, CG6380, CG15489, CG31349, CG12653, CG32683, CG5557, CG18646, CG4220, CG13701, CG5462, CG10701, CG10617, CG11354, CG18024*
cell differentiation	0.009889293	*CG33175, CG14026, CG2102, CG7527, CG18657, CG17697, CG4319, CG18076, CG10037, CG6775, CG7100, CG9224, CG10901, CG3758, CG3653, CG31349, CG15489, CG12154, CG3936, CG5557, CG10619, CG5461, CG10701, CG8705, CG7524, CG18024, CG3619*

Table A.7.: Terms from the process Ontology with p-value<=0.01, category of known enhancers.

Term	p-value	Annotated Genes
periodic partitioning by pair rule gene	1.37E-13	*CG6494, CG2047, CG3851, CG6716, CG2328, CG1849*
head segmentation	2.83E-13	*CG7952, CG2988, CG12154, CG9786, CG6494, CG1028, CG2047, CG2328*
BLASToderm segmentation	3.03E-12	*CG7952, CG4717, CG3851, CG2988, CG1849, CG12154, CG9786, CG6494, CG1028, CG2047, CG6716, CG2328*
pattern specification process	5.12E-12	*CG4717, CG3851, CG10002, CG2047, CG1028, CG2331, CG7952, CG6246, CG4531, CG2988, CG9786, CG12154, CG1849, CG6494, CG1214, CG6716, CG2328*
regionalization	4.60E-11	*CG4717, CG3851, CG10002, CG2047, CG1028, CG2331, CG7952, CG4531, CG2988, CG9786, CG12154, CG1849, CG6494, CG1214, CG6716, CG2328*
embryonic pattern specification	6.70E-11	*CG7952, CG4717, CG3851, CG2988, CG1849, CG12154, CG9786, CG6494, CG1028, CG2047, CG6716, CG2328*
segmentation	8.60E-11	*CG7952, CG4717, CG3851, CG2988, CG1849, CG12154, CG9786, CG6494, CG1028, CG2047, CG6716, CG2328*
posterior head segmentation	6.71E-10	*CG6494, CG7952, CG2047, CG2988, CG2328, CG9786*
regulation of transcription from RNA polymerase II promoter	1.65E-08	*CG7952, CG4717, CG3851, CG6246, CG5058, CG1849, CG10002, CG6494, CG1028, CG2047*
embryonic development	1.12E-07	*CG9181, CG7952, CG4717, CG3851, CG2988, CG1849, CG12154, CG9786, CG10002, CG6494, CG2047, CG1028, CG6716, CG2328*
regulation of transcription	1.22E-07	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG7952, CG6246, CG9930, CG2988, CG9786, CG12154, CG1849, CG6494, CG6716, CG2328*

Table A.8.: Terms from the process Ontology with p-value<=0.01, category of known enhancers.

Term	p-value	Annotated Genes
regulation of transcription, DNA-dependent	1.29E-07	*CG7952*, *CG4717*, *CG3851*, *CG9930*, *CG6246*, *CG2988*, *CG5058*, *CG1849*, *CG12154*, *CG10002*, *CG6494*, *CG2047*, *CG1028*, *CG6716*, *CG2328*
regulation of gene expression	2.01E-07	*CG4717*, *CG3851*, *CG5058*, *CG10002*, *CG2047*, *CG1028*, *CG9181*, *CG7952*, *CG6246*, *CG9930*, *CG2988*, *CG9786*, *CG12154*, *CG1849*, *CG6494*, *CG6716*, *CG2328*
regulation of gene expression	2.01E-07	*CG4717*, *CG3851*, *CG5058*, *CG10002*, *CG2047*, *CG1028*, *CG9181*, *CG7952*, *CG6246*, *CG9930*, *CG2988*, *CG9786*, *CG12154*, *CG1849*, *CG6494*, *CG6716*, *CG2328*
trunk segmentation	2.18E-07	*CG6494*, *CG2047*, *CG4717*, *CG2328*, *CG9786*
regulation of macromolecule biosynthetic process	3.81E-07	*CG4717*, *CG3851*, *CG5058*, *CG10002*, *CG2047*, *CG1028*, *CG7952*, *CG6246*, *CG9930*, *CG2988*, *CG9786*, *CG12154*, *CG1849*, *CG6494*, *CG6716*, *CG2328*
regulation of macromolecule biosynthetic process	3.81E-07	*CG4717*, *CG3851*, *CG5058*, *CG10002*, *CG2047*, *CG1028*, *CG7952*, *CG6246*, *CG9930*, *CG2988*, *CG9786*, *CG12154*, *CG1849*, *CG6494*, *CG6716*, *CG2328*
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	4.08E-07	*CG4717*, *CG3851*, *CG5058*, *CG10002*, *CG2047*, *CG1028*, *CG7952*, *CG6246*, *CG9930*, *CG2988*, *CG9786*, *CG12154*, *CG1849*, *CG6494*, *CG6716*, *CG2328*
regulation of RNA metabolic process	4.15E-07	*CG7952*, *CG4717*, *CG3851*, *CG9930*, *CG6246*, *CG2988*, *CG5058*, *CG1849*, *CG12154*, *CG10002*, *CG6494*, *CG2047*, *CG1028*, *CG6716*, *CG2328*
cell fate commitment	4.46E-07	*CG4717*, *CG6246*, *CG4531*, *CG5058*, *CG1849*, *CG9786*, *CG6494*, *CG2047*, *CG1214*, *CG2328*
transcription from RNA polymerase II promoter	4.84E-07	*CG7952*, *CG4717*, *CG3851*, *CG6246*, *CG5058*, *CG1849*, *CG10002*, *CG6494*, *CG1028*, *CG2047*
regulation of cellular metabolic process	5.07E-07	*CG4717*, *CG3851*, *CG5058*, *CG10002*, *CG2047*, *CG1028*, *CG9181*, *CG7952*, *CG6246*, *CG9930*, *CG2988*, *CG9786*, *CG12154*, *CG1849*, *CG6494*, *CG6716*, *CG2328*

Table A.9.: Terms from the process Ontology with p-value<=0.01, category of known enhancers.

Term	p-value	Annotated Genes
nervous system development	5.10E-07	*CG9181, CG11430, CG4717, CG6246, CG2988, CG4531, CG5058, CG1849, CG12154, CG9786, CG6494, CG2047, CG1214, CG2328, CG2331*
transcription	6.00E-07	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG7952, CG6246, CG9930, CG2988, CG9786, CG12154, CG1849, CG6494, CG6716, CG2328*
regulation of biosynthetic process	6.10E-07	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG7952, CG6246, CG9930, CG2988, CG9786, CG12154, CG1849, CG6494, CG6716, CG2328*
regulation of cellular biosynthetic process	6.10E-07	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG7952, CG6246, CG9930, CG2988, CG9786, CG12154, CG1849, CG6494, CG6716, CG2328*
periodic partitioning	6.62E-07	*CG6494, CG2047, CG3851, CG6716, CG2328, CG1849*
transcription, DNA-dependent	6.96E-07	*CG7952, CG4717, CG3851, CG9930, CG6246, CG2988, CG5058, CG1849, CG12154, CG10002, CG6494, CG2047, CG1028, CG6716, CG2328*
RNA biosynthetic process	7.22E-07	*CG7952, CG4717, CG3851, CG9930, CG6246, CG2988, CG5058, CG1849, CG12154, CG10002, CG6494, CG2047, CG1028, CG6716, CG2328*
&regulation of macromolecule metabolic process	1.09E-06	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG9181, CG7952, CG6246, CG9930, CG2988, CG9786, CG12154, CG1849, CG6494, CG6716, CG2328*
regulation of macromolecule metabolic process	1.09E-06	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG9181, CG7952, CG6246, CG9930, CG2988, CG9786, CG12154, CG1849, CG6494, CG6716, CG2328*
regulation of metabolic process	1.83E-06	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG9181, CG7952, CG6246, CG9930, CG2988, CG9786, CG12154, CG1849, CG6494, CG6716, CG2328*
generation of neurons	6.39E-06	*CG9181, CG4717, CG6246, CG2988, CG4531, CG1849, CG12154, CG9786, CG6494, CG1214, CG2328, CG2331*

Table A.10.: Terms from the process Ontology with p-value<=0.01, category of known enhancers.

Term	p-value	Annotated Genes
organ development	7.49E-06	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG2331, CG7952, CG6246, CG4531, CG2988, CG9786, CG12154, CG1849, CG6494, CG1214, CG2328*
neurogenesis	8.78E-06	*CG9181, CG4717, CG6246, CG2988, CG4531, CG1849, CG12154, CG9786, CG6494, CG1214, CG2328, CG2331*
ventral cord development	9.70E-06	*CG6246, CG2988, CG5058, CG1849, CG9786*
system development	1.52E-05	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG2331, CG11430, CG9181, CG7952, CG6246, CG4531, CG2988, CG1849, CG12154, CG9786, CG6494, CG1214, CG2328*
central nervous system development	2.81E-05	*CG6246, CG2988, CG5058, CG1849, CG12154, CG9786, CG2047, CG2328*
cellular biopolymer biosynthetic process	6.28E-05	*CG33158, CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG7952, CG6246, CG9930, CG2988, CG9786, CG12154, CG1849, CG7925, CG6494, CG6716, CG2328*
biopolymer biosynthetic process	6.28E-05	*CG33158, CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG7952, CG6246, CG9930, CG2988, CG9786, CG12154, CG1849, CG7925, CG6494, CG6716, CG2328*
regulation of cellular process	8.25E-05	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG2331, CG11430, CG9181, CG7952, CG9930, CG6246, CG2988, CG4531, CG1849, CG12154, CG9786, CG6494, CG6716, CG1214, CG2328*
gene expression	9.18E-05	*CG33158, CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG9181, CG7952, CG6246, CG9930, CG2988, CG1849, CG9786, CG12154, CG7925, CG6494, CG6716, CG2328*
cellular macromolecule biosynthetic process	0.000127822	*CG33158, CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG7952, CG6246, CG9930, CG2988, CG9786, CG12154, CG1849, CG7925, CG6494, CG6716, CG2328*
macromolecule biosynthetic process	0.000128989	*CG33158, CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG7952, CG6246, CG9930, CG2988, CG9786, CG12154, CG1849, CG7925, CG6494, CG6716, CG2328*

Table A.11.: Terms from the process Ontology with p-value<=0.01, category of known enhancers.

Term	p-value	Annotated Genes
RNA metabolic process	0.000176543	*CG7952, CG4717, CG3851, CG9930, CG6246, CG2988, CG5058, CG1849, CG12154, CG10002, CG6494, CG2047, CG1028, CG6716, CG2328*
regulation of biological process	0.000187533	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG2331, CG11430, CG9181, CG7952, CG9930, CG6246, CG2988, CG4531, CG1849, CG12154, CG9786, CG6494, CG6716, CG1214, CG2328*
neuron differentiation	0.00026688	*CG9181, CG4717, CG2988, CG4531, CG1849, CG12154, CG6494, CG1214, CG2328, CG2331*
stem cell differentiation	0.000268705	*CG6246, CG2988, CG5058, CG1849, CG9786*
multicellular organismal development	0.000324393	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG2331, CG11430, CG9181, CG7952, CG6246, CG2988, CG4531, CG1849, CG12154, CG9786, CG6494, CG6716, CG1214, CG2328*
cell fate determination	0.000350794	*CG4717, CG6246, CG5058, CG2328, CG1849, CG9786*
neuroBLAST differentiation	0.000363342	*CG6246, CG2988, CG1849, CG9786*
stem cell fate determination	0.000363342	*CG6246, CG5058, CG1849, CG9786*
germ-band extension	0.000370405	*CG9181, CG2328, CG1849*
ganglion mother cell fate determination	0.000370405	*CG6246, CG5058, CG9786*
cell differentiation	0.000376236	*CG9181, CG4717, CG6246, CG2988, CG4531, CG5058, CG1849, CG12154, CG9786, CG6494, CG2047, CG1214, CG2328, CG2331*

Table A.12.: Terms from the process Ontology with p-value<=0.01, category of known enhancers.

Term	p-value	Annotated Genes
zygotic determination of anterior/posterior axis, embryo	0.000420556	*CG7952, CG4717, CG12154, CG9786*
stem cell fate commitment	0.000484195	*CG6246, CG5058, CG1849, CG9786*
anatomical structure development	0.000516852	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG2331, CG11430, CG9181, CG7952, CG6246, CG4531, CG2988, CG1849, CG12154, CG9786, CG6494, CG1214, CG2328*
anterior head segmentation	0.000554442	*CG1028, CG2988, CG12154*
tripartite regional subdivision	0.000660515	*CG7952, CG4717, CG2328, CG12154, CG9786*
determination of anterior/posterior axis, embryo	0.000660515	*CG7952, CG4717, CG2328, CG12154, CG9786*
multicellular organismal process	0.000905105	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG2331, CG11430, CG9181, CG7952, CG6246, CG2988, CG4531, CG1849, CG12154, CG9786, CG7925, CG6494, CG6716, CG1214, CG2328, CG32401*
biological regulation	0.001105485	*CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG2331, CG11430, CG9181, CG7952, CG9930, CG6246, CG2988, CG4531, CG1849, CG12154, CG9786, CG6494, CG6716, CG1214, CG2328*
embryonic axis specification	0.001176034	*CG7952, CG4717, CG2328, CG12154, CG9786*
cellular biosynthetic process	0.0012226	*CG33158, CG4717, CG3851, CG5058, CG10002, CG2047, CG1028, CG7952, CG6246, CG9930, CG2988, CG9786, CG12154, CG1849, CG7925, CG6494, CG6716, CG2328*
cellular developmental process	0.001290419	*CG9181, CG4717, CG6246, CG2988, CG4531, CG5058, CG1849, CG12154, CG9786, CG6494, CG2047, CG1214, CG2328, CG2331*

Table A.13.: Terms from the process Ontology with p-value<=0.01, category of known enhancers.

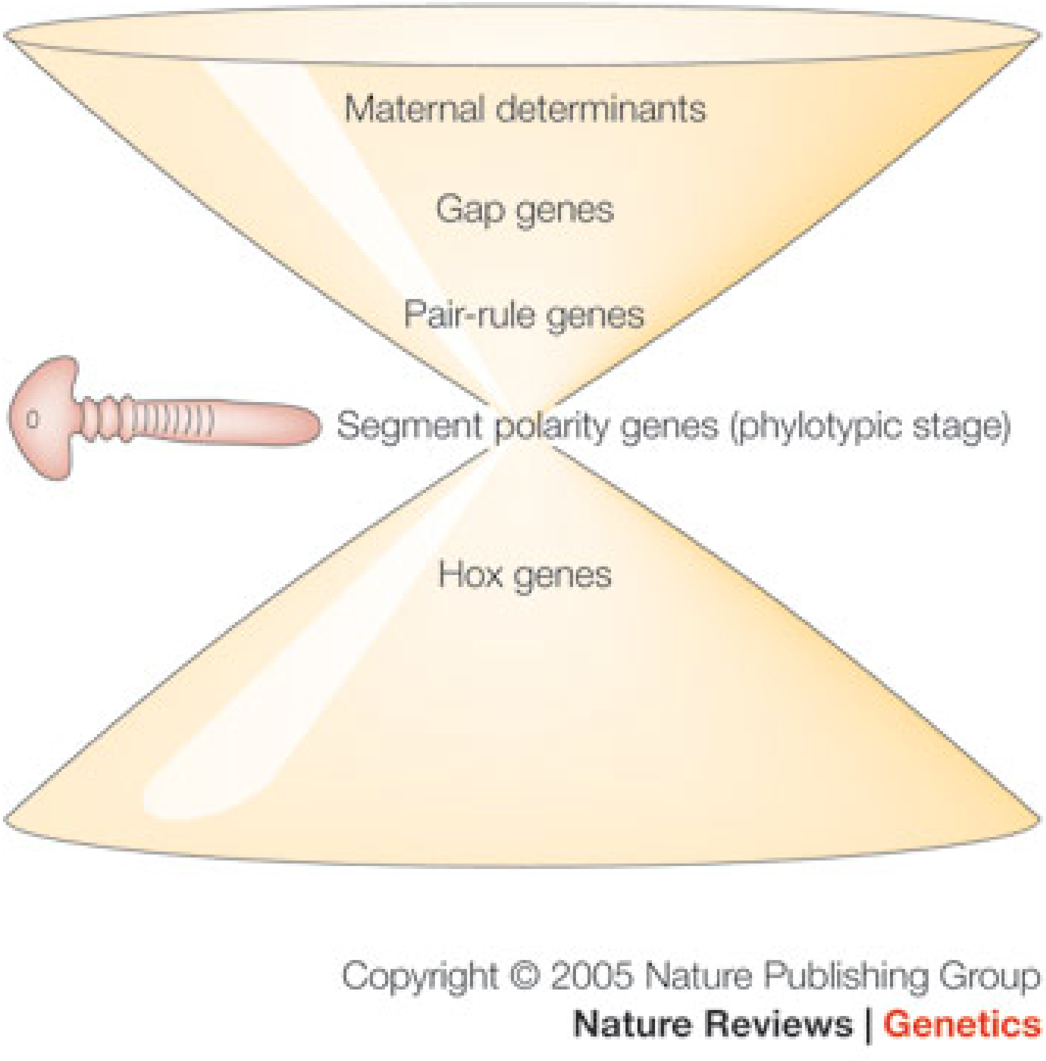
Term	p-value	Annotated Genes
biosynthetic process	0.001804278	*CG33158*, *CG4717*, *CG3851*, *CG5058*, *CG10002*, *CG2047*, *CG1028*, *CG7952*, *CG6246*, *CG9930*, *CG2988*, *CG9786*, *CG12154*, *CG1849*, *CG7925*, *CG6494*, *CG6716*, *CG2328*
anterior/posterior axis specification	0.002128033	*CG7952*, *CG4717*, *CG2328*, *CG12154*, *CG9786*, *CG2331*
central nervous system segmentation	0.002337	*CG2988*, *CG12154*
brain segmentation	0.002337	*CG2988*, *CG12154*
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	0.002340374	*CG4717*, *CG3851*, *CG5058*, *CG10002*, *CG2047*, *CG1028*, *CG7952*, *CG6246*, *CG9930*, *CG2988*, *CG9786*, *CG12154*, *CG1849*, *CG6494*, *CG6716*, *CG2328*
cell motion	0.002434403	*CG9181*, *CG4717*, *CG4531*, *CG1849*, *CG9786*, *CG2047*, *CG1214*, *CG2328*
localization of cell	0.00280617	*CG9181*, *CG4717*, *CG4531*, *CG1849*, *CG9786*, *CG2047*, *CG1214*, *CG2328*
developmental process	0.003277111	*CG4717*, *CG3851*, *CG5058*, *CG10002*, *CG2047*, *CG1028*, *CG2331*, *CG11430*, *CG9181*, *CG7952*, *CG6246*, *CG2988*, *CG4531*, *CG1849*, *CG12154*, *CG9786*, *CG6494*, *CG6716*, *CG1214*, *CG2328*
cell development	0.004159131	*CG9181*, *CG4717*, *CG6246*, *CG2988*, *CG4531*, *CG1849*, *CG12154*, *CG9786*, *CG6494*, *CG1214*, *CG2328*, *CG2331*
anterior/posterior pattern formation	0.00428314	*CG7952*, *CG4717*, *CG2328*, *CG12154*, *CG9786*, *CG2331*
neuron development	0.005272689	*CG9181*, *CG4717*, *CG2988*, *CG4531*, *CG1849*, *CG12154*, *CG2328*, *CG2331*
negative regulation of transcription from RNA polymerase II promoter	0.005301728	*CG6494*, *CG7952*, *CG4717*, *CG10002*
anterior region determination	0.006997632	*CG12154*, *CG9786*
specification of segmental identity, antennal segment	0.006997632	*CG1028*, *CG2988*
regulation of cell fate specification	0.007352843	*CG4531*, *CG2328*, *CG9786*

Table A.14.: The format of complete option file is shown in table, with five motifs occurrence probabilities.

```
[Motif]
name = BCD
motif = TABLE_PROB
#pos A C G T
-4 9 19 5 18
-3 11 3 1 36
-2 49 0 2 0
-1 51 0 0 0
1 0 0 17 34
2 1 45 0 5
3 1 25 4 21
4 4 16 21 10
background = 0.287 0.213 0.213 0.287
threshold = 5.1
[Motif]
name = CAD
motif = TABLE_PROB
#pos A C G T
-4 8 7 7 12
-3 5 18 4 7
-2 14 7 7 6
-1 0 7 3 24
1 34 0 0 0
2 34 0 0 0
3 34 0 0 0
4 17 3 8 6
background = 0.287 0.213 0.213 0.287
threshold = 5.6
[Motif]
name = HB
motif = TABLE_PROB
#pos A C G T
-6 12 12 10 59
-5 0 0 0 93
-4 1 4 2 86
-3 0 0 0 93
-2 0 0 0 93
-1 0 1 0 92
1 49 9 26 9
2 17 17 12 47
3 2 18 45 28
4 27 25 24 17
5 9 26 28 30
background = 0.287 0.213 0.213 0.287
threshold = 4.4
[Motif]
name = KR
motif = TABLE_PROB
#pos A C G T
-5 16 4 7 2
-4 27 1 1 0
-3 25 3 0 1
-2 15 7 3 4
-1 0 0 28 1
1 3 0 26 0
2 5 2 22 0
3 0 0 1 28
4 1 3 4 21
5 22 1 3 3
background = 0.287 0.213 0.213 0.287
threshold = 4.9

[Motif]
name = KNI
motif = TABLE_PROB
#pos A C G T
-5 5 0 0 0
-4 5 0 0 0
-3 1 3 1 0
-2 0 0 2 3
-1 4 1 0 0
1 0 0 5 0
2 3 1 1 0
3 0 1 2 2
4 0 5 0 0
5 5 0 0 0
background = 0.287 0.213 0.213 0.287
threshold = 5.4

[Sequence]
positive_training_set_filename=POS_FA
negative_training_set_filename=NEG_FA
sequence_filename=SEQUENCE_FA
[MultiMotifList]
distance=0,150
HB,BCD,CAD,KNI,KR
```



Maternal determinants
Gap genes
Pair-rule genes
Segment polarity genes (phylotypic stage)
Hox genes
Copyright © 2005 Nature Publishing Group
Nature Reviews | Genetics

Figure A.4.: Conservation of the segmentation cascade in arthropods. The degree of conservation of genes that function in successive steps of the segmentation cascade are represented by the width of the hourglass. Image courtesy of Peel [7].

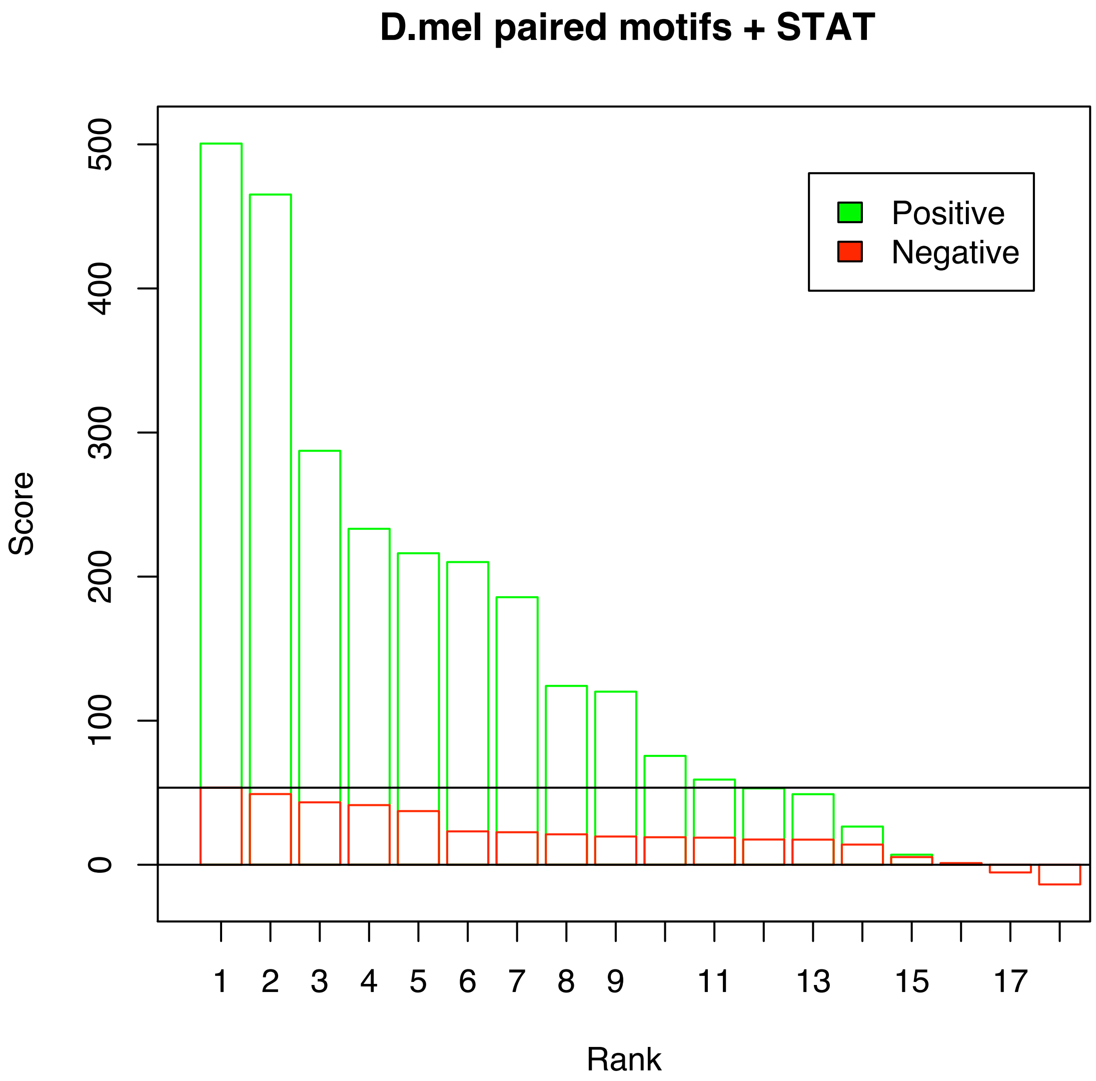


Figure A.5.: The separation of positive and negative training set with additional motif STAT. Green bars represent 15 positive training set. Red bars show 18 negative training set. The black line separate the number of enhancers which have score higher than the highest score in negative training set.

B. List of jPREdictor commands

In this project, jPREdictor was used to weight motifs, search for motifs in a sequence, calculate cutoff values, ect. All these tasks were performed using a command-line interface, which gives more options and parameters and makes configuration of the task more flexible. Especially, some special tasks can only be started via command line, such as the calculation of PSSM probabilities of motifs. The list of useful jPREdictor commands implemented in the project is shown below:

```
java -jar /vol/fpsearch/jPREdictor.jar -o OPTION_FILE -a -w WINDOW -v > OUTPUT
```

Paired motifs is the default setting by jPREdictor. The command line for scoring paired motifs is shown above.

parameter -A: switch off sequence length normalization, where the number of motif occurrences is divided by the length of the sequence.

parameter -v: Verbose mode. Prints some status information in addition to the results to standard out.

```
java -jar /vol/fpsearch/jPREdictor.jar -o OPTION_FILE -a -w WINDOW -v -p single > OUTPUT
```

To score sequence file with single motif setting, the default setting must be switch off by using parameter “–p” with option “single”.

```
java -jar /vol/fpsearch/jPREdictor.jar -o OPTION_FILE -w WINDOW -v -p single –forcesearchmotifs -t > OUTPUT
```

“—forceSearchMotifs": forces jPREdictor to find every motif occurrence in a sequence file. Figure 2.24 is an implementation of this command.

```
java -jar /vol/fpsearch/jPREdictor.jar -o OPTION_FILE -a -w WINDOW –forceWeightMotifs
```

“—forceWeightMotifs": forces to weight either paired or single motifs in the option file. Figure 2.38 is an implementation of this command.

```
/homes/tfiedler/pub/bin/mksequ -b 100 21205609 | java -jar /vol/fpsearch/jPREdictor.jar -o OPTION_FILE -a -w WINDOW –forceWeightMotifs –cutoffCalc -f - > OUTPUT
```

“—cutoffCalc": performs a cut-off calculation by scoring a sequence file.

```
java -jar /vol/fpsearch/jPREdictor.jar –PSSMprobs -o OPTION_FILE > OUTPUT
```

“—pssmProbs”: calculates a complete p-value distribution for any given PSSM. The PFMs of motifs must be provided in an option file first. Figure 2.2 is an implementation of this command.

C. Abbreviations

BX-C	Bithorax complex
PREs	PcG response elements
TREs	TrxG response elements
PcG	polycomb group
TF	transcription factors
AP	anterior–posterior
BCD	Bicoid
CAD	Caudal
ChIP/chip	chromatin immunoprecipitation coupled with DNA microarray hybridization
CREs	*cis*-regulatory module
DV	dorsal–ventral
FDR	false-discovery rate
GO	gene ontology
GT	Giant
HB	Hunchback
KNI	Knirps
KR	Kruppel
PWM	position weight matrix
eve	*even-skipped*
ftz	*fushi tarazu*

Bibliography

[1] B. P. Berman, B. D. Pfeiffer, T. R. Laverty, S. L. Salzberg, G. M. Rubin, M. B. Eisen, and S. E. Celniker, "Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in drosophila melanogaster and drosophila pseudoobscura.," *Genome Biol*, vol. 5, no. 9, p. R61, 2004.

[2] M. D. Schroeder, M. Pearce, J. Fak, H. Fan, U. Unnerstall, E. Emberly, N. Rajewsky, E. D. Siggia, and U. Gaul, "Transcriptional control in the segmentation gene network of drosophila.," *PLoS Biol*, vol. 2, p. E271, Sep 2004.

[3] M. N. Arbeitman, E. E. M. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. A. Krasnow, M. P. Scott, R. W. Davis, and K. P. White, "Gene expression during the life cycle of drosophila melanogaster.," *Science*, vol. 297, pp. 2270–2275, Sep 2002.

[4] D. Papatsenko and M. S. Levine, "Dual regulation by the hunchback gradient in the drosophila embryo.," *Proc Natl Acad Sci U S A*, Feb 2008.

[5] S. F. Gilbert, *Developmental Biology, 6th Edition*. Sinauer associates, INC., 2000. ISBN 0-87893-243-7.

[6] T. Phillips, "Genetic signaling: Transcription factor cascades and segmentation," *Nature Education*, 2008.

[7] A. D. Peel, A. D. Chipman, and M. Akam, "Arthropod segmentation: beyond the drosophila paradigm.," *Nat Rev Genet*, vol. 6, pp. 905–916, Dec 2005.

[8] D. S. Johnston and C. Nuesslein-Volhard, "The origin of pattern and polarity in the drosophila embryo.," *Cell*, vol. 68, pp. 201–219, Jan 1992.

[9] D. Tautz, "Segmentation.," *Dev Cell*, vol. 7, pp. 301–312, Sep 2004.

[10] B. Houchmandzadeh, E. Wieschaus, and S. Leibler, "Establishment of developmental precision and proportions in the early drosophila embryo.," *Nature*, vol. 415, pp. 798–802, Feb 2002.

[11] http://www.sdbonline.org/fly/segment/kruppel1.htm.

[12] R. Rivera-Pomar and H. Jackle, "From gradients to stripes in drosophila embryogenesis: filling in the gaps.," *Trends Genet*, vol. 12, pp. 478–483, Nov 1996.

[13] M. Huelskamp, C. Pfeifle, and D. Tautz, "A morphogenetic gradient of hunchback protein organizes the expression of the gap genes kruppel and knirps in the early drosophila embryo.," *Nature*, vol. 346, pp. 577–580, Aug 1990.

[14] http://studentreader.com/tag/even-skipped/.

[15] M. J. Pankratz, E. Seifert, N. Gerwin, B. Billi, U. Nauber, and H. Jaeckle, "Gradients of kruppel and knirps gene products direct pair-rule gene stripe patterning in the posterior region of the drosophila embryo.," *Cell*, vol. 61, pp. 309–317, Apr 1990.

[16] A. Lempradl and L. Ringrose, "How does noncoding transcription regulate hox genes?," *Bioessays*, vol. 30, pp. 110–121, Feb 2008.

[17] R. K. Maeda and F. Karch, "The abc of the bx-c: the bithorax complex explained.," *Development*, vol. 133, pp. 1413–1422, Apr 2006.

[18] E. B. Lewis, "A gene complex controlling segmentation in drosophila.," *Nature*, vol. 276, pp. 565–570, Dec 1978.

[19] E. B. Lewis, B. D. Pfeiffer, D. R. Mathog, and S. E. Celniker, "Evolution of the homeobox complex in the diptera.," *Curr Biol*, vol. 13, pp. R587–R588, Aug 2003.

[20] J. Smith and E. H. Davidson, "A new method, using cis-regulatory control, for blocking embryonic gene expression.," *Dev Biol*, vol. 318, pp. 360–365, Jun 2008.

[21] D. Niessing, R. Rivera-Pomar, A. L. Rosee, T. Haeder, F. Schoeck, B. A. Purnell, and H. Jaeckle, "A cascade of transcriptional control leading to axis determination in drosophila.," *J Cell Physiol*, vol. 173, pp. 162–167, Nov 1997.

[22] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen, "Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the drosophila genome.," *Proc Natl Acad Sci U S A*, vol. 99, pp. 757–762, Jan 2002.

[23] S. Aerts, M. Haeussler, S. van Vooren, O. L. Griffith, P. Hulpiau, S. J. M. Jones, S. B. Montgomery, C. M. Bergman, and O. R. A. Consortium, "Text-mining assisted regulatory annotation.," *Genome Biol*, vol. 9, no. 2, p. R31, 2008.

[24] X.-Y. Li, S. Macarthur, R. Bourgon, D. Nix, D. A. Pollard, V. N. Iyer, A. Hechmer, L. Simirenko, M. Stapleton, C. L. L. Hendriks, H. C. Chu, N. Ogawa, W. Inwood, V. Sementchenko, A. Beaton, R. Weiszmann, S. E. Celniker, D. W. Knowles, T. Gingeras, T. P. Speed, M. B. Eisen, and M. D. Biggin, "Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm.," *PLoS Biol*, vol. 6, p. e27, Feb 2008.

[25] C. N. Nuesslein-Volhard, "The bicoid morphogen papers (i): account from cnv.," *Cell*, vol. 116, pp. S1–5, 2 p following S9, Jan 2004.

[26] A. Spirov, K. Fahmy, M. Schneider, E. Frei, M. Noll, and S. Baumgartner, "Formation of the bicoid morphogen gradient: an mRNA gradient dictates the protein gradient.," *Development*, vol. 136, pp. 605–614, Feb 2009.

[27] A. Ochoa-Espinosa, D. Yu, A. Tsirigos, P. Struffi, and S. Small, "Sackler special feature: Anterior-posterior positional information in the absence of a strong bicoid gradient.," *Proc Natl Acad Sci U S A*, Feb 2009.

[28] http://www.ltscotland.org.uk/nq/resources/dynamicdevelopment/moviesflycompare.htm.

[29] H. Hardway, B. Mukhopadhyay, T. Burke, T. J. Hitchman, and R. Forman, "Modeling the precision and robustness of hunchback border during drosophila embryonic development.," *J Theor Biol*, vol. 254, pp. 390–399, Sep 2008.

[30] F. J. P. Lopes, F. M. C. Vieira, D. M. Holloway, P. M. Bisch, and A. V. Spirov, "Spatial bistability generates hunchback expression sharpness in the drosophila embryo.," *PLoS Comput Biol*, vol. 4, p. e1000184, Sep 2008.

[31] M. J. Pankratz, M. Busch, M. Hoch, E. Seifert, and H. Jaeckle, "Spatial control of the gap gene knirps in the drosophila embryo by posterior morphogen system.," *Science*, vol. 255, pp. 986–989, Feb 1992.

[32] K. Lunde, B. Biehs, U. Nauber, and E. Bier, "The knirps and knirps-related genes organize development of the second wing vein in drosophila.," *Development*, vol. 125, pp. 4145–4154, Nov 1998.

[33] M. Z. Ludwig, C. Bergman, N. H. Patel, and M. Kreitman, "Evidence for stabilizing selection in a eukaryotic enhancer element.," *Nature*, vol. 403, pp. 564–567, Feb 2000.

[34] S. Small, A. Blair, and M. Levine, "Regulation of even-skipped stripe 2 in the drosophila embryo.," *EMBO J*, vol. 11, pp. 4047–4057, Nov 1992.

[35] D. N. Arnosti, S. Barolo, M. Levine, and S. Small, "The eve stripe 2 enhancer employs multiple modes of transcriptional synergy.," *Development*, vol. 122, pp. 205–214, Jan 1996.

[36] L. P. M. Andrioli, V. Vasisht, E. Theodosopoulou, A. Oberstein, and S. Small, "Anterior repression of a drosophila stripe enhancer requires three position-specific mechanisms.," *Development*, vol. 129, pp. 4931–4940, Nov 2002.

[37] Y. Nibu, H. Zhang, E. Bajor, S. Barolo, S. Small, and M. Levine, "dctbp mediates transcriptional repression by knirps, kruppel and snail in the drosophila embryo.," *EMBO J*, vol. 17, pp. 7009–7020, Dec 1998.

[38] E. H. Davidson and M. S. Levine, "Properties of developmental gene regulatory networks.," *Proc Natl Acad Sci U S A*, vol. 105, pp. 20063–20066, Dec 2008.

[39] S. Sinha, Y. Liang, and E. Siggia, "Stubb: a program for discovery and analysis of cis-regulatory modules.," *Nucleic Acids Res*, vol. 34, pp. W555–W559, Jul 2006.

[40] N. Rajewsky, M. Vergassola, U. Gaul, and E. D. Siggia, "Computational detection of genomic cis-regulatory modules applied to body patterning in the early drosophila embryo.," *BMC Bioinformatics*, vol. 3, p. 30, Oct 2002.

[41] Y. H. Grad, F. P. Roth, M. S. Halfon, and G. M. Church, "Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in drosophila melanogaster and d.pseudoobscura.," *Bioinformatics*, vol. 20, pp. 2738–2750, Nov 2004.

[42] M. Z. Ludwig, A. Palsson, E. Alekseeva, C. M. Bergman, J. Nathan, and M. Kreitman, "Functional evolution of a cis-regulatory module.," *PLoS Biol*, vol. 3, p. e93, Apr 2005.

[43] M. I. Arnone and E. H. Davidson, "The hardwiring of development: organization and function of genomic regulatory systems.," *Development*, vol. 124, pp. 1851–1864, May 1997.

[44] H. Lodish, A. Berk, L. S. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology, 4th Edition*. W. H. FREEMAN AND COMPANY, 2000. ISBN 0-7167-3136-31986.

[45] L. A. Lettice, S. J. H. Heaney, L. A. Purdie, L. Li, P. de Beer, B. A. Oostra, D. Goode, G. Elgar, R. E. Hill, and E. de Graaff, "A long-range shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly.," *Hum Mol Genet*, vol. 12, pp. 1725–1735, Jul 2003.

[46] S. Ogbourne and T. M. Antalis, "Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes.," *Biochem J*, vol. 331 (Pt 1), pp. 1–14, Apr 1998.

[47] L. Valenzuela and R. T. Kamakaka, "Chromatin insulators.," *Annu Rev Genet*, vol. 40, pp. 107–138, 2006.

[48] M. Gaszner and G. Felsenfeld, "Insulators: exploiting transcriptional and epigenetic mechanisms.," *Nat Rev Genet*, vol. 7, pp. 703–713, Sep 2006.

[49] B. Burgess-Beusse, C. Farrell, M. Gaszner, M. Litt, V. Mutskov, F. Recillas-Targa, M. Simpson, A. West, and G. Felsenfeld, "The insulation of genes from external enhancers and silencing chromatin.," *Proc Natl Acad Sci U S A*, vol. 99 Suppl 4, pp. 16433–16437, Dec 2002.

[50] A. M. Bushey, E. R. Dorman, and V. G. Corces, "Chromatin insulators: regulatory mechanisms and epigenetic inheritance.," *Mol Cell*, vol. 32, pp. 1–9, Oct 2008.

[51] N. Negre, J. Hennetin, L. V. Sun, S. Lavrov, M. Bellis, K. P. White, and G. Cavalli, "Chromosomal distribution of pcg proteins during drosophila development.," *PLoS Biol*, vol. 4, p. e170, Jun 2006.

[52] L. Ringrose and R. Paro, "Epigenetic regulation of cellular memory by the polycomb and trithorax group proteins.," *Annu Rev Genet*, vol. 38, pp. 413–443, 2004.

[53] B. Schuettengruber, M. Ganapathi, B. Leblanc, M. Portoso, R. Jaschek, B. Tolhuis, M. van Lohuizen, A. Tanay, and G. Cavalli, "Functional anatomy of polycomb and trithorax chromatin landscapes in drosophila embryos.," *PLoS Biol*, vol. 7, p. e13, Jan 2009.

[54] D. EH., "In: The regulatory genome: Gene regulatory networks in development and evolution, 1st.," *Burlington, MA: Academic Press*, 2006.

[55] L. Li, Q. Zhu, X. He, S. Sinha, and M. Halfon, “Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses.,” *Genome Biol*, vol. 8, p. R101, Jun 2007.

[56] W. W. Wasserman and A. Sandelin, “Applied bioinformatics for the identification of regulatory elements.,” *Nat Rev Genet*, vol. 5, pp. 276–287, Apr 2004.

[57] M. L. Bulyk, “Computational prediction of transcription-factor binding site locations.,” *Genome Biol*, vol. 5, no. 1, p. 201, 2003.

[58] D. GuhaThakurta, “Computational identification of transcriptional regulatory elements in dna sequence.,” *Nucleic Acids Res*, vol. 34, no. 12, pp. 3585–3598, 2006.

[59] X. Yu, J. Lin, D. J. Zack, and J. Qian, “Identification of tissue-specific cis-regulatory modules based on interactions between transcription factors.,” *BMC Bioinformatics*, vol. 8, p. 437, 2007.

[60] A. Wagner, “Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes.,” *Bioinformatics*, vol. 15, pp. 776–784, Oct 1999.

[61] A. Ochoa-Espinosa and S. Small, “Developmental mechanisms and cis-regulatory codes.,” *Curr Opin Genet Dev*, vol. 16, pp. 165–170, Apr 2006.

[62] M. S. Halfon, Y. Grad, G. M. Church, and A. M. Michelson, “Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model.,” *Genome Res*, vol. 12, pp. 1019–1028, Jul 2002.

[63] S. Aerts, J. van Helden, O. Sand, and B. A. Hassan, “Fine-tuning enhancer models to predict transcriptional targets across multiple genomes.,” *PLoS ONE*, vol. 2, no. 11, p. e1115, 2007.

[64] A. Ochoa-Espinosa, G. Yucel, L. Kaplan, A. Pare, N. Pura, A. Oberstein, D. Papatsenko, and S. Small, “The role of binding site cluster strength in bicoid-dependent patterning in drosophila.,” *Proc Natl Acad Sci U S A*, vol. 102, pp. 4960–4965, Apr 2005.

[65] M. Markstein, P. Markstein, V. Markstein, and M. S. Levine, “Genome-wide analysis of clustered dorsal binding sites identifies putative target genes in the drosophila embryo.,” *Proc Natl Acad Sci U S A*, vol. 99, pp. 763–768, Jan 2002.

[66] M. Rebeiz, N. L. Reeves, and J. W. Posakony, “Score: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. site clustering over random expectation.,” *Proc Natl Acad Sci U S A*, vol. 99, pp. 9888–9893, Jul 2002.

[67] A. P. Lifanov, V. J. Makeev, A. G. Nazina, and D. A. Papatsenko, “Homotypic regulatory clusters in drosophila.,” *Genome Res*, vol. 13, pp. 579–588, Apr 2003.

[68] J.-V. Turatsinze, M. Thomas-Chollier, M. Defrance, and J. van Helden, “Using rsat to scan genome sequences for transcription factor binding sites and cis-regulatory modules.,” *Nat Protoc*, vol. 3, no. 10, pp. 1578–1588, 2008.

[69] S. Vardhanabhuti, J. Wang, and S. Hannenhalli, “Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation.,” *Nucleic Acids Res*, vol. 35, no. 10, pp. 3203–3213, 2007.

[70] X. Liu, D. L. Brutlag, and J. S. Liu, “Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes.,” *Pac Symp Biocomput*, pp. 127–138, 2001.

[71] Y. Barash, G. Bejerano, and N. Friedman, *A Simple Hyper-Geometric Approach for Discovering Putative Transcription Factor Binding Sites*. 2001.

[72] K. D. MacIsaac and E. Fraenkel, “Practical strategies for discovering regulatory dna sequence motifs.,” *PLoS Comput Biol*, vol. 2, p. e36, Apr 2006.

[73] K. T. Takusagawa and D. K. Gifford, “Negative information for motif discovery.,” *Pac Symp Biocomput*, pp. 360–371, 2004.

[74] J. M. Claverie and S. Audic, “The statistical significance of nucleotide position-weight matrix matches.,” *Comput Appl Biosci*, vol. 12, pp. 431–439, Oct 1996.

[75] J. Vilo, A. Brazma, I. Jonassen, A. Robinson, and E. Ukkonen, “Mining for putative regulatory elements in the yeast genome using gene expression data.,” *Proc Int Conf Intell Syst Mol Biol*, vol. 8, pp. 384–394, 2000.

[76] K. A. Frazer, L. Elnitski, D. M. Church, I. Dubchak, and R. C. Hardison, “Cross-species sequence comparisons: a review of methods and available resources.,” *Genome Res*, vol. 13, pp. 1–12, Jan 2003.

[77] D. C. King, J. Taylor, Y. Zhang, Y. Cheng, H. A. Lawson, J. Martin, E. N. C. O. D. E. groups for Transcriptional Regulation, M. S. Analysis, F. Chiaromonte, W. Miller, and R. C. Hardison, “Finding cis-regulatory elements using comparative genomics: some lessons from encode data.,” *Genome Res*, vol. 17, pp. 775–786, Jun 2007.

[78] A. G. Clark and D. G. Consortium, “Evolution of genes and genomes on the drosophila phylogeny.,” *Nature*, vol. 450, pp. 203–218, Nov 2007.

[79] M. A. Nobrega, I. Ovcharenko, V. Afzal, and E. M. Rubin, “Scanning human gene deserts for long-range enhancers.,” *Science*, vol. 302, p. 413, Oct 2003.

[80] L. Elnitski, W. Miller, and R. Hardison, “Conserved e boxes function as part of the enhancer in hypersensitive site 2 of the beta-globin locus control region. role of basic helix-loop-helix proteins.,” *J Biol Chem*, vol. 272, pp. 369–378, Jan 1997.

[81] G. G. Loots, R. M. Locksley, C. M. Blankespoor, Z. E. Wang, W. Miller, E. M. Rubin, and K. A. Frazer, “Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons.,” *Science*, vol. 288, pp. 136–140, Apr 2000.

[82] T. Dickmeis and F. Mueller, “The identification and functional characterisation of conserved regulatory elements in developmental genes.,” *Brief Funct Genomic Proteomic*, vol. 3, pp. 332–350, Feb 2005.

[83] D. A. Tagle, B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess, and R. T. Jones, “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*, vol. 203, pp. 439–455, Sep 1988.

[84] X. Wang, J. Gu, M. Q. Zhang, and Y. Li, “Identification of phylogenetically conserved microrna cis-regulatory elements across 12 drosophila species.,” *Bioinformatics*, vol. 24, pp. 165–171, Jan 2008.

[85] E. H. Margulies, “Confidence in comparative genomics.,” *Genome Res*, vol. 18, pp. 199–200, Feb 2008.

[86] A. Stark, M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, J. G. Ruby, J. Brennecke, H. F. curators, B. D. G. Project, E. Hodges, A. S. Hinrichs, A. Caspi, B. Paten, S.-W. Park, M. V. Han, M. L. Maeder, B. J. Polansky, B. E. Robson, S. Aerts, J. van Helden, B. Hassan, D. G. Gilbert, D. A. Eastman, M. Rice, M. Weir, M. W. Hahn, Y. Park, C. N. Dewey, L. Pachter, W. J. Kent, D. Haussler, E. C. Lai, D. P. Bartel, G. J. Hannon, T. C. Kaufman, M. B. Eisen, A. G. Clark, D. Smith, S. E. Celniker, W. M. Gelbart, and M. Kellis, “Discovery of functional elements in 12 drosophila genomes using evolutionary signatures.,” *Nature*, vol. 450, pp. 219–232, Nov 2007.

[87] E. H. Margulies and E. Birney, “Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes.,” *Nat Rev Genet*, vol. 9, pp. 303–313, Apr 2008.

[88] R. Satija, L. Pachter, and J. Hein, “Combining statistical alignment and phylogenetic footprinting to detect regulatory elements.,” *Bioinformatics*, vol. 24, pp. 1236–1242, May 2008.

[89] B. Wilczynski, N. Dojer, M. Patelak, and J. Tiuryn, “Finding evolutionarily conserved cis-regulatory modules with a universal set of motifs.,” *BMC Bioinformatics*, vol. 10, p. 82, Mar 2009.

[90] S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller, “Human-mouse alignments with blastz.,” *Genome Res*, vol. 13, pp. 103–107, Jan 2003.

[91] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins.,” *J Mol Biol*, vol. 48, pp. 443–453, Mar 1970.

[92] M. Brudno, C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, N. I. S. C. C. S. Program, E. D. Green, A. Sidow, and S. Batzoglou, “Lagan and multi-lagan: efficient tools for large-scale multiple alignment of genomic dna.,” *Genome Res*, vol. 13, pp. 721–731, Apr 2003.

[93] A. Prakash and M. Tompa, "Measuring the accuracy of genome-size multiple alignments.," *Genome Biol*, vol. 8, no. 6, p. R124, 2007.

[94] E. Emberly, N. Rajewsky, and E. D. Siggia, "Conservation of regulatory elements between two species of drosophila.," *BMC Bioinformatics*, vol. 4, p. 57, Nov 2003.

[95] M. Z. Ludwig, N. H. Patel, and M. Kreitman, "Functional analysis of eve stripe 2 enhancer evolution in drosophila: rules governing conservation and change.," *Development*, vol. 125, pp. 949–958, Mar 1998.

[96] A. Hauenschild, L. Ringrose, C. Altmutter, R. Paro, and M. Rehmsmeier, "Evolutionary plasticity of polycomb/trithorax response elements in drosophila species.," *PLoS Biol*, vol. 6, p. e261, Oct 2008.

[97] L. Ringrose, M. Rehmsmeier, J.-M. Dura, and R. Paro, "Genome-wide prediction of polycomb/trithorax response elements in drosophila melanogaster.," *Dev Cell*, vol. 5, pp. 759–771, Nov 2003.

[98] T. Fiedler and M. Rehmsmeier, "jpredictor: a versatile tool for the prediction of cis-regulatory elements.," *Nucleic Acids Res*, vol. 34, pp. W546–W550, Jul 2006.

[99] M. Klingler, "The organization of the antero-posterior axis.," *Semin Cell Biol*, vol. 1, pp. 151–160, Jun 1990.

[100] A. P. McGregor, "How to get ahead: the origin, evolution and function of bicoid.," *Bioessays*, vol. 27, pp. 904–913, Sep 2005.

[101] D. Lebrecht, M. Foehr, E. Smith, F. J. P. Lopes, C. E. Vanario-Alonso, J. Reinitz, D. S. Burz, and S. D. Hanes, "Bicoid cooperative dna binding is critical for embryonic patterning in drosophila.," *Proc Natl Acad Sci U S A*, vol. 102, pp. 13176–13181, Sep 2005.

[102] http://rana.lbl.gov/drosophila/.

[103] M. Blanchette, A. R. Bataille, X. Chen, C. Poitras, J. Laganiere, C. Lefebvre, G. Deblois, V. Giguere, V. Ferretti, D. Bergeron, B. Coulombe, and F. Robert, "Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression.," *Genome Res*, vol. 16, pp. 656–668, May 2006.

[104] D. Dorsett, "Distant liaisons: long-range enhancer-promoter interactions in drosophila.," *Curr Opin Genet Dev*, vol. 9, pp. 505–514, Oct 1999.

[105] S. W. Schaeffer, A. Bhutkar, B. F. McAllister, M. Matsuda, L. M. Matzkin, P. M. O'Grady, C. Rohde, V. L. S. Valente, M. Aguade, W. W. Anderson, K. Edwards, A. C. L. Garcia, J. Goodman, J. Hartigan, E. Kataoka, R. T. Lapoint, E. R. Lozovsky, C. A. Machado, M. A. F. Noor, M. Papaceit, L. K. Reed, S. Richards, T. T. Rieger, S. M. Russo, H. Sato, C. Segarra, D. R. Smith, T. F. Smith, V. Strelets, Y. N. Tobari, Y. Tomimura, M. Wasserman, T. Watts, R. Wilson, K. Yoshida, T. A. Markow, W. M. Gelbart, and T. C. Kaufman, "Polytene chromosomal maps of 11 drosophila species: the order of genomic scaffolds inferred from genetic and physical maps.," *Genetics*, vol. 179, pp. 1601–1655, Jul 2008.

[106] https://svn.r-project.org/R/trunk/src/library/stats/R/qqplot.R.

[107] M. J. Sanderson and J. Kim, "Parametric phylogenetics?," *Syst Biol*, vol. 49, pp. 817–829, Dec 2000.

[108] B. Kolaczkowski and J. W. Thornton, "Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous.," *Nature*, vol. 431, pp. 980–984, Oct 2004.

[109] D. M. Hillis, J. P. Huelsenbeck, and C. W. Cunningham, "Application and accuracy of molecular phylogenies.," *Science*, vol. 264, pp. 671–677, Apr 1994.

[110] E. E. Hare, B. K. Peterson, and M. B. Eisen, "A careful look at binding site reorganization in the even-skipped enhancers of drosophila and sepsids.," *PLoS Genet*, vol. 4, p. e1000268, Nov 2008.

[111] V. Boeva, M. Regnier, D. Papatsenko, and V. Makeev, "Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression.," *Bioinformatics*, vol. 22, pp. 676–684, Mar 2006.

[112] http://genome.lbl.gov/vista/mvista/submit.shtml.

[113] E. E. Hare, B. K. Peterson, V. N. Iyer, R. Meier, and M. B. Eisen, "Sepsid even-skipped enhancers are functionally conserved in drosophila despite lack of sequence conservation.," *PLoS Genet*, vol. 4, p. e1000106, Jun 2008.

[114] T. Hader, A. L. Rosee, U. Ziebold, M. Busch, H. Taubert, H. Jackle, and R. Rivera-Pomar, "Activation of posterior pair-rule stripe expression in response to maternal caudal and zygotic knirps activities.," *Mech Dev*, vol. 71, pp. 177–186, Feb 1998.

[115] A. M. Moses, D. A. Pollard, D. A. Nix, V. N. Iyer, X.-Y. Li, M. D. Biggin, and M. B. Eisen, "Large-scale turnover of functional transcription factor binding sites in drosophila.," *PLoS Comput Biol*, vol. 2, p. e130, Oct 2006.

[116] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, "Weblogo: a sequence logo generator.," *Genome Res*, vol. 14, pp. 1188–1190, Jun 2004.

[117] G. Z. Hertz and G. D. Stormo, "Identifying dna and protein patterns with statistically significant alignments of multiple sequences.," *Bioinformatics*, vol. 15, no. 7-8, pp. 563–577, 1999.

[118] C. Segarra and M. Aguade, "Molecular organization of the x chromosome in different species of the obscura group of drosophila.," *Genetics*, vol. 130, pp. 513–521, Mar 1992.

[119] C. Segarra, E. R. Lozovskaya, G. Ribo, M. Aguade, and D. L. Hartl, "P1 clones from drosophila melanogaster as markers to study the chromosomal evolution of muller's a element in two species of the obscura group of drosophila.," *Chromosoma*, vol. 104, pp. 129–136, Nov 1995.

[120] M. S. Halfon, S. M. Gallo, and C. M. Bergman, “Redfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in drosophila.,” *Nucleic Acids Res*, Nov 2007.

[121] http://www.fruitfly.org/cgi-bin/ex/basic.pl.

[122] http://www.fruitfly.org/cgi-bin/ex/insitu.pl.

[123] http://go.princeton.edu/cgi-bin/GOTermFinder.

[124] M. Fujioka, Y. Emi-Sarker, G. L. Yusibova, T. Goto, and J. B. Jaynes, “Analysis of an even-skipped rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients.,” *Development*, vol. 126, pp. 2527–2538, Jun 1999.

[125] M. S. Halfon, A. Carmena, S. Gisselbrecht, C. M. Sackerson, F. Jimenez, M. K. Baylies, and A. M. Michelson, “Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors.,” *Cell*, vol. 103, pp. 63–74, Sep 2000.

[126] G. D. Stormo, “Dna binding sites: representation and discovery.,” *Bioinformatics*, vol. 16, pp. 16–23, Jan 2000.

[127] F. Consortium, “The flybase database of the drosophila genome projects and community literature.,” *Nucleic Acids Res*, vol. 31, pp. 172–175, January 2003.

[128] http://genome.cse.ucsc.edu/cgi-bin/hgGateway.

[129] T. Fiedler, *The jPREdictor and its Application to Motif Discovery, Motif Clustering and the Prediction of Polycomb/Trithorax Response Elements*. PhD thesis, Technischen Fakultaet der Universitaet Bielefeld, April 2008.

[130] http://cran.r-project.org/doc/manuals/R-intro.html#Examining-the-distribution-of-a-set-of-data.

[131] http://www.geneontology.org/GO.database.shtml.

[132] J. Banerji, S. Rusconi, and W. Schaffner, “Expression of a beta-globin gene is enhanced by remote sv40 dna sequences.,” *Cell*, vol. 27, pp. 299–308, Dec 1981.

[133] J. R. Manak, S. Dike, V. Sementchenko, P. Kapranov, F. Biemar, J. Long, J. Cheng, I. Bell, S. Ghosh, A. Piccolboni, and T. R. Gingeras, “Biological function of unannotated transcription during the early development of drosophila melanogaster.,” *Nat Genet*, vol. 38, pp. 1151–1158, Oct 2006.

[134] P. J. Wittkopp, “Evolution of cis-regulatory sequence and function in diptera.,” *Heredity*, vol. 97, pp. 139–147, Sep 2006.

[135] G. Sawa, J. Dicks, and I. N. Roberts, “Current approaches to whole genome phylogenetic analysis.,” *Brief Bioinform*, vol. 4, pp. 63–74, Mar 2003.

[136] W. M. Fitch, “Toward defining the course of evolution: Minimum change for a specific tree topology,” *Systematic Zoology*, vol. 20, no. 4, pp. 406–416, 1971.

[137] M. Pagel, “Inferring the historical patterns of biological evolution.,” *Nature*, vol. 401, pp. 877–884, Oct 1999.

[138] C. R. Hardy, “Reconstructing ancestral ecologies: challenges and possible solutions.,” *Diversity & Distributions*, vol. 12, pp. 7–19, Oct 2006.

[139] M. Pagel, “The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies.,” *Systematic Biology*, vol. 48, pp. 612–622, 1999.

[140] W. P. Maddison and D. R. Maddison, “Mesquite: a modular system for evolutionary analysis.,” 2008.

[141] http://insects.eugenes.org/species/news/genome-summaries/gene-phylogeny.html.

[142] N. Bray and L. Pachter, “Mavid: constrained ancestral alignment of multiple sequences.,” *Genome Res*, vol. 14, pp. 693–699, Apr 2004.

[143] E. E. Hare, B. K. Peterson, V. N. Iyer, R. Meier, and M. B. Eisen, “Sepsid even-skipped enhancers are functionally conserved in drosophila despite lack of sequence conservation.,” *PLoS Genet*, vol. 4, p. e1000106, Jun 2008.

[144] http://www.mathworks.de/matlabcentral/fileexchange/6116.

[145] S. Sinha and E. D. Siggia, “Sequence turnover and tandem repeats in cis-regulatory modules in drosophila.,” *Mol Biol Evol*, vol. 22, pp. 874–885, Apr 2005.

[146] H. Li and W. Wang, “Dissecting the transcription networks of a cell using computational genomics.,” *Curr Opin Genet Dev*, vol. 13, pp. 611–616, Dec 2003.

[147] S. Gray and M. Levine, “Short-range transcriptional repressors mediate both quenching and direct repression within complex loci in drosophila.,” *Genes Dev*, vol. 10, pp. 700–710, Mar 1996.

[148] S. Small, A. Blair, and M. Levine, “Regulation of two pair-rule stripes by a single enhancer in the drosophila embryo.,” *Dev Biol*, vol. 175, pp. 314–324, May 1996.

[149] S. Fisher, E. A. Grice, R. M. Vinton, S. L. Bessling, and A. S. McCallion, “Conservation of ret regulatory function from human to zebrafish without sequence similarity.,” *Science*, vol. 312, pp. 276–279, Apr 2006.

[150] E. P. Consortium, “Identification and analysis of functional elements in 1the human genome by the encode pilot project.,” *Nature*, vol. 447, pp. 799–816, Jun 2007.

[151] A. R. Borneman, T. A. Gianoulis, Z. D. Zhang, H. Yu, J. Rozowsky, M. R. Seringhaus, L. Y. Wang, M. Gerstein, and M. Snyder, “Divergence of transcription factor binding sites across related yeast species.,” *Science*, vol. 317, pp. 815–819, Aug 2007.

[152] S. W. Doniger and J. C. Fay, “Frequent gain and loss of functional transcription factor binding sites.,” *PLoS Comput Biol*, vol. 3, p. e99, May 2007.

[153] C. Plessy, T. Dickmeis, F. Chalmel, and U. Strahle, “Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes.,” *Trends Genet*, vol. 21, pp. 207–210, Apr 2005.

[154] A. Woolfe, M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, S. F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y. J. K. Edwards, J. E. Cooke, and G. Elgar, “Highly conserved non-coding sequences are associated with vertebrate development.,” *PLoS Biol*, vol. 3, p. e7, Jan 2005.

[155] M. E. Alonso, B. Pernaute, M. Crespo, J. Gomez-Skarmeta, and M. Manzanares, “Understanding the regulatory genome.,” *Int J Dev Biol*, Nov 2008.

[156] J. Kim, X. He, and S. Sinha, “Evolution of regulatory sequences in 12 drosophila species.,” *PLoS Genet*, vol. 5, p. e1000330, Jan 2009.

[157] D. Tautz, “Evolution of transcriptional regulation.,” *Curr Opin Genet Dev*, vol. 10, pp. 575–579, Oct 2000.

[158] N. S. Wratten, A. P. McGregor, P. J. Shaw, and G. A. Dover, “Evolutionary and functional analysis of the tailless enhancer in musca domestica and drosophila melanogaster.,” *Evol Dev*, vol. 8, no. 1, pp. 6–15, 2006.

[159] A. P. McGregor, P. J. Shaw, and G. A. Dover, “Sequence and expression of the hunchback gene in lucilia sericata: a comparison with other dipterans.,” *Dev Genes Evol*, vol. 211, pp. 315–318, Jun 2001.

[160] L. Sipos and H. Gyurkovics, “Long-distance interactions between enhancers and promoters.,” *FEBS J*, vol. 272, pp. 3253–3259, Jul 2005.

[161] J. Ma, *Gene expression and regulation.* Springer, 2006. ISBN 7040176750, 9787040176759.

[162] A. Sosinsky, B. Honig, R. S. Mann, and A. Califano, “Discovering transcriptional regulatory regions in drosophila by a nonalignment method for phylogenetic footprinting.,” *Proc Natl Acad Sci U S A*, vol. 104, pp. 6305–6310, Apr 2007.

[163] W. Huang, J. R. Nevins, and U. Ohler, “Phylogenetic simulation of promoter evolution: estimation and modeling of binding site turnover events and assessment of their impact on alignment tools.,” *Genome Biol*, vol. 8, no. 10, p. R225, 2007.

[164] K. Struhl, “Gene regulation. a paradigm for precision.,” *Science*, vol. 293, pp. 1054–1055, Aug 2001.

[165] S. MacArthur and J. F. Y. Brookfield, “Expected rates and modes of evolution of enhancer sequences.,” *Mol Biol Evol*, vol. 21, pp. 1064–1073, Jun 2004.

[166] M. Ridley, *Evolution, 3th Edition.* Wiley-Blackwell, 2004. ISBN 1405103450, 9781405103459.

[167] X. Wang and H. M. Chamberlin, “Multiple regulatory changes contribute to the evolution of the caenorhabditis lin-48 ovo gene.,” *Genes Dev*, vol. 16, pp. 2345–2349, Sep 2002.

[168] B. Prud’homme, N. Gompel, and S. B. Carroll, “Emerging principles of regulatory evolution.,” *Proc Natl Acad Sci U S A*, vol. 104 Suppl 1, pp. 8605–8612, May 2007.

[169] S. Jeong, A. Rokas, and S. B. Carroll, “Regulation of body pigmentation by the abdominal-b hox protein and its gain and loss in drosophila evolution.,” *Cell*, vol. 125, pp. 1387–1399, Jun 2006.

[170] S. Jeong, M. Rebeiz, P. Andolfatto, T. Werner, J. True, and S. B. Carroll, “The evolution of gene regulation underlies a morphological difference between two drosophila sister species.,” *Cell*, vol. 132, pp. 783–793, Mar 2008.

[171] R. Yan, S. Small, C. Desplan, C. R. Dearolf, and J. E. Darnell, “Identification of a stat gene that functions in drosophila development.,” *Cell*, vol. 84, pp. 421–430, Feb 1996.

[172] M. Markstein, R. Zinzen, P. Markstein, K.-P. Yee, A. Erives, A. Stathopoulos, and M. Levine, “A regulatory code for neurogenic gene expression in the drosophila embryo.,” *Development*, vol. 131, pp. 2387–2394, May 2004.