Comparative and compositional features of cis-regulatory modules in *Drosophila melanogaster*

Dissertation
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Nora Pierstorff
aus Neuss

2006

1. Gutachter: Prof. Thomas Wiehe

2. Gutachter: Prof. Diethard Tautz

Tag der letzten mündlichen Prüfung: 14.2.2007

Acknowledgements

This thesis and myself have been guided by many complaisant supporters during the last three years.

First I would like to thank my supervisor Prof. Thomas Wiehe for his patience and the insistence to do things more properly than fast. He was always available for all questions and I felt spoilt when I listened to PhD students from other groups talking about their supervisors.

Prof. Diethard Tautz and Prof. Michael Laessig gave very useful hints and reminded me to stick to practical biological questions. Especially the decision to use comparative data was made with their help. The supervision of additional experienced scientists conveyed an additional amount of confirmation during the work for this thesis. I am grateful for this support.

I also would like to thank Prof. Siegfried Roth and Dr. Röbbe Wünschiers who complete the examination committee with Prof. Thomas Wiehe and Prof. Diethard Tautz.

Ph.D. Casey Bergman hosted me in Manchester (UK) and taught me the fashionable way to present results. He shared his large knowledge about Drosophila with me and introduced some hypotheses to me.

Special thanks to Rodrigo Nunes da Fonseca, who was always willing to discuss my questions, produced some of the results in Chapter 2 and corrected the biological background in the introduction. Dale Richardson crossed my way at the end of the thesis and planned to learn something about gene regulation in plants. His work led to some results, which are mentioned in the last Chapter of this thesis.

During the last 2 years of my thesis the group of Thomas Wiehe grew and additional PhD-students accompanied Thomas, Anton, Frank and me. This very friendly and lively environment cheered up my days.

I received additional financial but more importantly mental support from my parents and my siblings Bärbel, Klaus and Martha. They supported me with PhD experience, distraction from scientific problems when it was necessary and fulfilling my horse riding duties, when I was in Manchester. I want to thank especially for financing our horse "Francesca". This is special even in our family where education is not wanted to be limited by financial issues.

Abstract

Transcriptional regulation of genes is a crucial process in every living system. The difficulty of uncovering details of this process increases with the complexity of the analysed species. The goal of this work is to develop a tool that helps uncovering this process in different species by localizing those genomic regions which are involved in transcriptional regulation.

We developed two methods addressing this question. `Shureg` is applicable to all species but produces results of low quality. `CisPlusFinder` can only be applied to species, where enough closely related species are sequenced. The CRM predictions of `CisPlusFinder` are of higher accuracy than has generally been acchieved thus far.

We performed a whole genome scan of the species *D. melanogaster* and concluded the results with the aim to formulate new hypotheses concerning transcriptional regulation.

We found evidence that transcriptional regulation is of extremely high relevance for higher organisms and that the extended use of complex gene regulation played a major role in the split of multicellular organisms from single cell organisms. We could also support the idea that sequences performing regulatory function shape the genome architecture of *D. melanogaster*, which corresponds to a high DNA loss and genome density in this species as it was stated before.

Zusammenfassung

Transkriptionelle Genregulation ist ein äusserst wichtiger Prozess jedes biologischen Systems. Die Komplexität und Anzahl verschiedener regulatorischer Mechanismen steigt mit der Komplexität des Organismus. Das Ziel dieser Doktorarbeit ist die Entwicklung eines neuen Computerprogramms, dass bei der Aufklärung der transkriptionellen Regulation verschiedener Spezies hilft, indem es die genomischen Regionen lokalisiert, die in diesen Prozess involviert sind.

Wir haben zwei Methoden entwickelt, die dieses Problem bearbeiten. `Shureg` ist auf alle Sequenzen anwendbar und benutzt keine Informationen, die nicht in der untersuchten Sequenz enthalten sind. Die Ergebnisse dieser Methode sind sehr ungenau. `CisPlusFinder` kann nur auf solche Spezies angewendet werden, für die nahe verwandte Spezies sequenziert sind, die Informationen über die Sequenzkonservierung zwischen verschiedenen Spezies liefern. Die Vorhersagen des `CisPlusFinders` sind von besserer Qualität als bis jetzt mit theoretischen Methoden erreicht werden konnte.

Um neue Hypothesen über die Vorgänge der transkriptionellen Regulation formulieren zu können, haben wir `CisPlusFinder` auf das komplette Drosophilagenom angewendet. Die so errechneten Vorhersagen wurden ausgewertet. Wir konnten Theorien und Abschätzungen, die von Wissenschaftlern gemacht wurden, bestätigen und neue Erkenntnisse gewinnen.

Insbesondere weisen unsere Ergebnisse darauf hin, daß transkiptionelle Regulation eine immens wichtige Rolle in höheren Organismen inne hat. Die Anwendung komplexer Regulationsmechanismen auf viele Gene scheint mit dem evolutionären Split zwischen einzelligen und mehrzelligen Organismen einhergegangen zu sein. Außerdem können wir die Theorie unterstützen, daß Sequenzen, die regulatorische Funktionen ausführen, die Genomarchitektur von *D. melanogaster* stark beeinflussen. Diese Idee stimmt mit früheren Aussagen über einen hohen DNA-Verlust und eine sehr hohen Genomdichte in Drosophila überein.

Contents

Chapter 1

Introduction

One of the goals of computational biology is the assignment of biological function to parts of genomes of all kind of organisms and species. The function of DNA, which was discovered first, is the protein coding function. By now it is known that many other functions are performed by the DNA. RNAs which catalyse biochemical reactions themselves or accomplish other enzymes to function are also coded in DNA sequence. At the moment it is known that a high amount of DNA which was believed to be noncoding until now is transcribed into RNA. Manak *et al.* (2006) found that 30% of the DNA which is transcribed during the first hour of Drosophila development is unannotated. Wong *et al.* (2001) state that most of the human genome is transcribed. The role or function of this transcription is not known yet. Krasilnikov *et al.* (1999) found that the process of transcription influences the DNA structure and suggest that this is a further function of DNA transcription.

The function of DNA which is central to this thesis is the transcriptional regulation of genes. Not all proteins coded in the genome are needed in all tissues at all times during the life span of an organism. Gene products can even be lethal, if they are present in the wrong cell. The aim of this thesis is to help to uncover mechanisms of transcriptional regulation by identifying the genomic regions of the genome where the information about gene regulation is coded.

1.1 Biological background of transcriptional regulation

The mechanism of transcriptional regulation differs a lot between eukaryotes and prokaryotes. Prokaryotes have very compact genomes that contain mainly coding regions and short intergenic regions. The DNA regulating a gene is mainly localized upstream of its transcription start sites and downstream of the stop-codon of the previous gene. The polymerase binding site which is directly upstream of a gene is called *"proximal promoter"*. Figure 1.1 shows a very well studied example of prokaryotic transcriptional regulation, the *lac-operon.*

Jacob & Monod (1961) discovered the coregulation of three genes in the lac-operon and described the rules of its regulation. The lac-operon encodes three genes whose products are necessary to catabolise lactose. The genes are arranged on one strand following each other and transcribed into a single mRNA. Their regulation is performed by a single promoter. The bacterium needs these genes to be expressed only if lactose is present in the cell. If lactose is absent the enzymes are dispensable. In the presence of glucose and lactose the transcription of the lac-operon is disadvantageous for the bacterium, because the glucose metabolism is more profitable than the lactose-metabolism. Figure 1.1 shows two possible states of the lac-operon. The first part of Figure 1.1 shows the state of the cell where glucose is present and lactose is absent. In this case a repressor is bound to the DNA. The repressor is located downstream of the polymerase binding site and upstream of the first gene at a position called operator. When the repressor is bound to the operator, the polymerase cannot bind to its binding site and transcription is suppressed.

The second part of Figure 1.1 shows the case that lactose is present in the cell. In

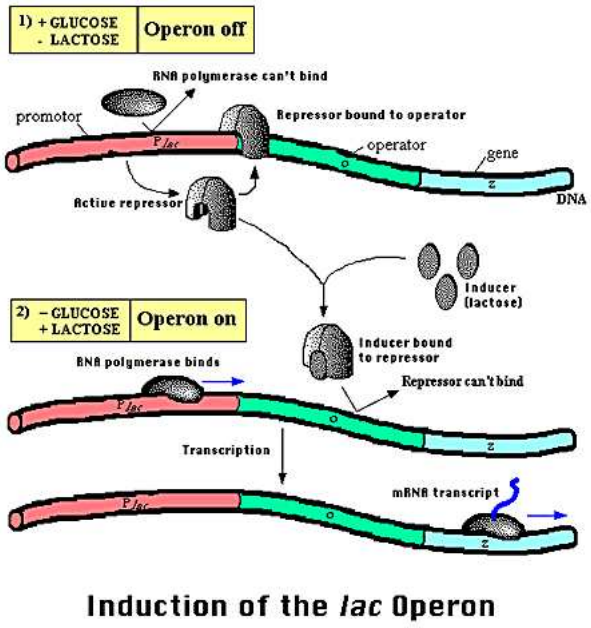


Figure 1.1: schematic overview over the regulation of the lac-operon
Source: Access Excellence Resource Center
http://www.accessexcellence.org/RC/VL/GG/induction.html

this case the lactose molecules bind to the repressor molecule and prevents the repressor from binding to the DNA. The polymerase binding site is then available to the enzyme and transcription proceeds.

To ensure that the expression of the lac-operon is not activated as long as glucose is present in the cell there is an additional step of regulation. A binding site for *CAP (catabolite activator protein)* is positioned upstream of the polymerase binding site. CAP is a protein which needs the cofactor *cAMP (cyclic Adenosinmonophosphat)* to perform DNA binding. cAMP is accumulated in cells when the glucose level is low. To enable a polymerase binding at the polymerase binding site CAP needs to be bound to its binding site upstream of the polymerase binding site. If CAP is not activated by cAMP, which signals a lack of glucose in the cell, the polymerase cannot bind and the operon cannot be expressed.

The regulation of the lac-operon is a very efficient gene regulation that encounters three cases. In multicellular organisms like eukaryotes gene regulation is much more complex. The main part of the genome is non coding in higher eukaryotes. In Drosophila roughly 20% of the genome is covered by protein coding genes (Halligan & Keightley, 2006). Some of the long intergenic regions are known to contain *cis-regulatory modules (CRMs)* to control the expression of the surrounding genes. A very well explored example of complexly regulated genes is *even-skipped (eve)*. *Eve* is expressed in the early development of *Drosophila melanogaster* and is involved in the segmentation of the embryo. The expression pattern and the distribution of the regulating modules are shown in Figure 1.2. A misregulation or a loss of function of *eve* is lethal for the fruit fly.

As Figure 1.2 shows *eve* is expressed in seven stripes along the anterior posterior body axis. Simon *et al.* (1990) and Celniker *et al.* (1990) were the first to show the expression of certain patterns by distinct CRMs in *D. melanogaster*. Figure 1.2 shows six different modules to regulate the 7-stripe expression pattern of *even-skipped* in blue color. CRMs indicated in a different color drive transcription of this gene at different time points or in different tissues.

Expression by distinct CRMs means that one CRM acts independently of the other CRMs regulating the same gene. A reporter gene construct with the stripe-3 enhancer, proximal

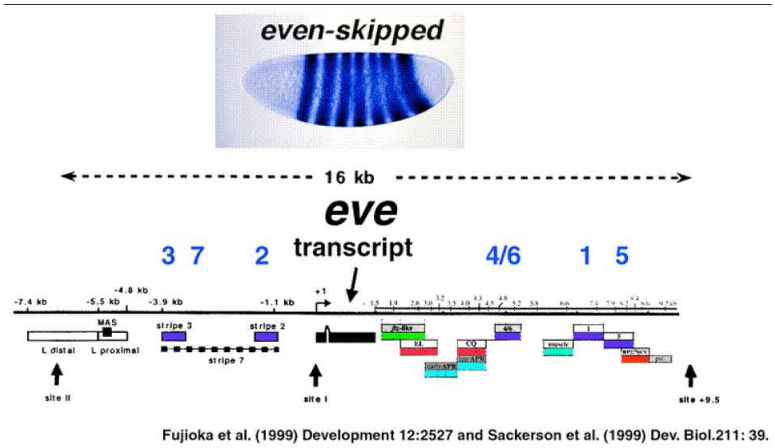


Figure 1.2: scheme of the 7-stripe expression pattern of *even-skipped* in *D. melanogaster* Source: Lecture course: Evolution of regulatory processes (D. Arnosti, Apr. 2004)

promoter and a reporter gene leads to the expression of the reporter gene at the position of *eve* stripe-3. Figure 1.2 shows that the regulation of gene expression in Drosophila does not exclusively take part in the upstream sequence of the gene. CRMs can also be located downstream of the regulated gene. We assume that CRMs are located downstream of the stop codon of the previous gene and upstream of the startcodon of the following gene. Hence, we assume that no coding region is positioned between a CRM and the gene it regulates.

Wide parts of the mechanism of gene regulation are still unknown. The function of a CRM is determined by the interaction with transcription factors. Every CRM contains different *transcription factor binding sites (TFBS)*. The concentration of a transcription factor in the nucleus of a cell and the affinity of the TFBS to its transcription factor result in a bound TFBS or a free TFBS. The combination of free and bound TFBS result in a well defined expression pattern. For some well studied cases in the Drosophila development it is known which transcription factors have to be bound to which TFBS to drive the given expression pattern. In many cases the details are not known yet. Another open question in gene regulation is the process which connects the binding of the transcription factors with the resulting expression of the gene. A conformational change of the DNA to cause a physical contact between the CRM and the proximal promoter is assumed. But there is no experimental evidence for this or any other hypothesis.

1.1.1 Segmentation and dorsal-/ventral patterning of *Drosophila melanogaster*

The embryonic development of Drosophila melanogaster is a very complex process including many transcription factors regulating modular expression. This process has been studied extensively and the regulation of the genes involved in these process are the genes whose regulation is best known. The studies about this process by C. Nüsslein Vollhard, E. Wieschhaus and coworkers (Nüsslein-Vollhard & Wieschhaus, 1980) have been honored with a noble prize in 1995. Large parts of this thesis focus on this well studied process and a short overview of the principles of this process is given here.

Two very important steps of embryonic development are the formations of two axis. The anterior posterior axis and the dorsal ventral axis have to be determined in the *D. melanogaster* embryo. These axes are determined in the one celled egg by maternal transcription factors.

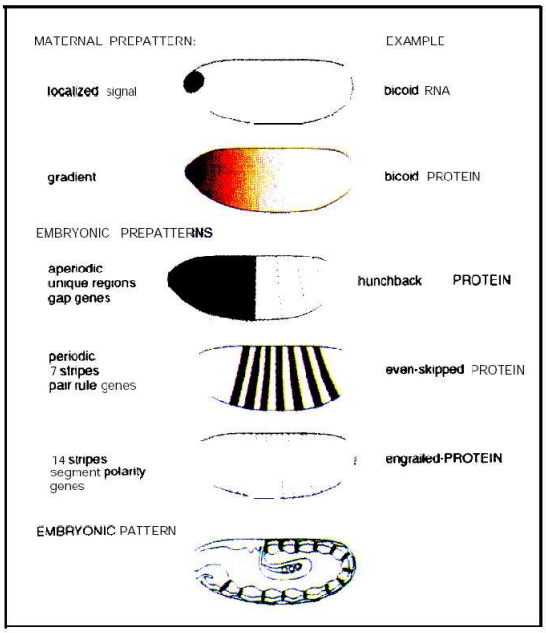


Figure 1.3: scheme of the anterior posterior patterning of *D. melanogaster*
Source: Nüsslein-Vollhard (1995)

The anterior posterior axis

A scheme of the determination of the anterior posterior axis is taken from Nüsslein-Vollhard (1995) and shown in Figure 1.3.

The maternal factors are present with a concentration gradient along the embryo. This concentration gradient and a certain affinity of TFBS result in a defined expression pattern. The affinity of the TFBS determines the necessary transcription factor concentration to be bound. The expression of the gene is then enhanced or silenced by the binding of the maternal transcription factor in the region where the transcription factor concentration exceeds the required threshold.

The genes directly regulated by maternal factors are called *gap genes* and are expressed in large unique regions as shown in Figure 1.3 for the gap gene *hunchback*. At the moment fifteen genes are known to perform the function of gap genes. They are called *buttonhead (btd), cap'n'collar (cnc), caudal (cad), knot (kn), crocodile (croc), empty spiracles (ems), giant (gt), huckebein (hkb), hunchback (hb), knirps (kn), krueppel (kr), orthodenticle (otd), sloppy paired 1 (slp1), sloppy paired 2 (slp2)* and *tailless*. Because every gap gene is expressed in a different unique region of the embryo, different parts of the embryo are marked by different combinations of present and absent gap genes. These different combinations of gap genes regulate the so called *pair rule* genes. The pair rule gene example shown in Figure 1.3 and in Figure 1.2 is the gene *even-skipped* which is one of the most extensively studied genes in all higher eukaryotes. Other known pair rule genes are called *fushi tarazu (ftz), hairy (h), odd paired (opa), odd skipped (odd), paired (prd), runt (run), sloppy paired (slp)* and *Tenascin major (Ten-M)*

All nine known pair rule genes are expressed in a pattern of seven stripes regularly along the embryo. The striped patterns of the different pair rule genes differ in stripe-width, inter-stripe distance and the distance from the first stripe to the anterior border of the embryo or the distance between the last stripe and the posterior end. The different combinations of pair rule genes combined with the presence or absence of gap proteins at certain positions of the embryo subdivide the embryo further and the *segment polarity* genes are expressed in 14 stripes along the embryo. Every segment of the Drosophila embryo has a determined

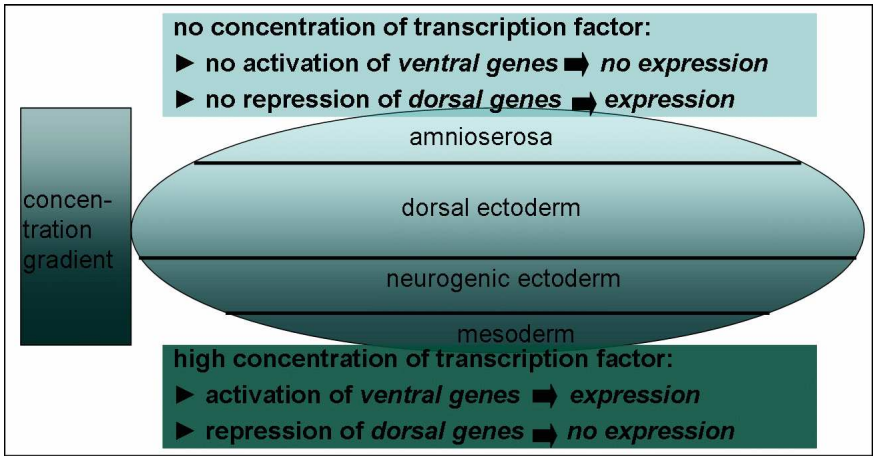


Figure 1.4: scheme of the dorsal ventral patterning of *D. melanogaster*

function in the later completely developed Drosophila organism, which has been defined by the maternal gradient already and is not changed any more.

The dorsal ventral axis

The determination of the dorsal ventral axis is less complex than the segmentation patterning along the anterior posterior axis. Figure 1.4 shows a scheme describing the regulation of this process. Only one group of maternal genes is necessary to establish a nuclear localization gradient of one transcription factor. The maximal concentration of the transcription factor is at the ventral side of the embryo and the minimal concentration is at the dorsal side of the embryo. The enhanced target genes of this transcription factor are expressed at the ventral end of this embryo and at the same time the genes expressed at the dorsal side of the embryo are silenced by this transcription factor.

1.2 Motivation of the prediction of CRMs

The correct regulation of transcription is of very high importance for a cell. One or multiple substitutions which change the affinity of one or more TFBS of a CRM in a way that the expression pattern changes can be lethal for an organism. Many diseases can be caused by the misregulation of genes. Many Modifications which turn a sane cell into a cancer cell should induce cell apoptosis of the single cell by upregulating apoptosis factors. A disruption of this mechanism can lead to the amplification of cancer cells and results in a tumor.

If the function of one transcription factor is modified by a substitution in the coding sequence the expression pattern of many proteins is effected. These processes cannot be understood without detailed knowledge about the regulatory processes of a cell.

Another reason to study gene regulation was published by Levine & Tijan (2003). Sequencing of many different organisms including nematode worms, mice and humans showed a much higher conservation of the coding sequences than expected. Vertebrate genomes have only about twice the number of genes that invertebrate genomes have, and the increase is primarily due to the duplication of existing genes rather than the invention of new ones.

These little differences can hardly account for the evolutionary differences between these species. Many differences are believed to be caused by differences in gene regulation. Also the complexity of the organisms is believed to be reflected in more elaborate regulation of gene expression instead of a higher complexity of the proteome.

Compared to the importance of gene regulation for different research branches our knowledge about eukaryotic gene regulation is tremendously small. This thesis is concerned only with transcriptional regulation. It is known that genes are also regulated on the level of

translation. But even less is known about this process and our target is exclusively the transcriptional regulation.

All we know about this mechanism is the fact that the expression level of a gene is determined by the binding of transcription factors to the TFBS located in CRMs. It is only assumed that this binding causes a conformational change of the DNA to affect the binding of the RNA-Polymerase to the core-promoter sequence. Neither is anything known about the conformational changes nor do we have any knowledge about the way bound CRMs interact with the core promoter.

The current datasets of known CRMs are not sufficient to find new regularities in gene regulation. Also the definition of a CRM is still fuzzy, since we do not know in which way it is limited from the surrounding sequence. Capelson & Corces (2004) review about some CRMs which are limited by boundary elements formed by chromatin structure. Since we are not able to correlate pure sequence with chromatin structure and do not know a sufficient number of boundary elements to make any assumption at the moment, we cannot define the boundaries of CRMs precisely.

Another problem is the lack of knowledge about CRM density along the sequence. Since we do not know how many CRMs are needed to regulate a gene or to control a certain expression pattern we cannot tell, if we know all CRMs of a certain region or if another CRM is located between a CRM and the regulated gene. If we can reproduce the expression pattern of one gene at a specific stage of the cell, we do not know the expression patterns at different time points and we do not know, if there are more CRMs active in different cell stages within the region. We also do not know exactly, if there are CRMs located within coding region or how frequently they occur.

All these questions have to be answered by further computational and experimental analysis of transcriptional regulation. The first task is to localize more CRMs to get a sufficient amount of data to perform significant computational analysis to formulate hypotheses which can then be tested experimentally. To decrease the cost of experimental CRM detections good predictions are crucial.

1.3 Known approaches to CRM prediction

Many methods have been developed to solve the problem of CRM identification according to different initial situations and for different organisms. Given the case that our goal is to uncover the regulation of one gene we need to choose a method which is adapted to our information about the gene. Experiments can provide the following different classes of information to us.

1. Coregulated genes: genes that share the same expression pattern and are likely to be regulated by the same mechanism

2. Regulating transcription factors: other verified TFBS of the transcription factors regulating the explored gene

3. Comparative data: homologous genes of the explored gene from other species which are assumed to be regulated by a conserved mechanism

4. No extrinsic information: The gene and its surrounding intergenic region containing the CRMs regulating the explored gene.

1.3.1 Analysing coregulated genes

The approach to use coregulated genes to find motifs which are common in their upstream regions has been successfully used in prokaryotes and lower eukaryotes such as yeast. Many methods following this approach have been developed. The first published program of this kind is `MEME` written by Bailey & Elkan (1994). `AlignACE` (Hughes *et al.*, 2000) and `Motifsampler` (Thisj *et al.*, 2001) were developed to be applicable to sequences of higher

eukaryotes as well as small genomes. `MOPAC` (Ganesh *et al.*, 2003) predicts the motifs directly from microarray data and performs also the clustering of the expression profiles. `BioProspector` (Liu *et al.*, 2004) was developed in combination with other programs to permit a combined analysis of multiple approaches which results in highly evaluated predictions in many cases.

The application of these methods to higher eukaryotes causes different problems. One problem is the size of the intergenic regions. Prokaryotes and lower eukaryotes have short well defined intergenic regions which can be completely analysed for CRM prediction. Those intergenic regions are rarely believed to perform any other functions than gene regulation. Intergenic regions in higher eukaryotes also code for functional RNA molecules, such as RNA genes, micro RNAs and others. They contain different origins of replication and some properties of the 3-dimensional structure of the DNA are also believed to be coded in the intergenic sequences of higher eukaryotes (Lankas *et al.*, 2000). In eukaryotes CRMs can be positioned upstream and downstream of the gene, within introns or overlapping coding regions.

Analysing whole intergenic regions surrounding an eukaryotic gene means analysing long sequences performing all kinds of known and unknown function. Not only motifs representing TFBS should be contained in all the sequences, but different motifs performing different functions are found with this approach. Tompa *et al.* (2005) tested 13 different programs using the information from coregulated genes to predict TFBS in eukaryotes and found at most 22% of all sites found. The highest accuracy measured by the correlation coefficient, defined by Burset & Guigo (1996) is 0.2. This result shows that the approach to search common motifs in upstream regions of coregulated genes is inapplicable to CRM prediction in higher eukaryotes. The approach to find common upstream motifs has been successfully applied to core-promoter analysis by Ohler (2006). Core-promoter are assumed to be within several hundred bp upstream of a gene. The sequences which have to be analysed for this purpose are of processable length.

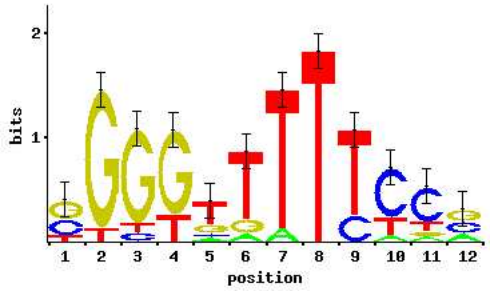
1.3.2 Analysing the regulation of a gene knowing the regulating transcription factors

The most promising approach to CRM prediction requires detailed biological knowledge about the regulation of the analysed gene. If the set of transcription factors which regulate the gene is known and if other TFBS bound by the used transcription factor are known, computer programs search for more occurrences of these TFBS.

The databases `TRANSFAC` (Matys *et al.*, 2003) and `JASPAR` (Stormo, 2000) are databases of transcription factors and known sequences of TFBS for different species. These big collections are crucial for the practicability of this approach. To conclude different TFBS of the same transcription factor the binding sites have to be aligned. Matrices as shown in Figure 1.5(b) are calculated from aligned TFBS and PWMs represented by motifs as shown in Figure 1.5(a) are computed.

Day & McMorris (1992) were the first to implement this approach. They represented binding sites of a certain transcription factor as a consensus of all known binding sites of this transcription factor and searched for all motifs matching this consensus in the analysed sequence. Due to the degeneracy of the TFBS of one transcription factor the consensus of all binding sites of one transcription factor has many degenerate positions and many hits which do not overlap a functional TFBS may be found. Osada *et al.* (2004) compare the consensus method by Day & McMorris (1992) with two methods using *position-weight matrices (PWMs)* (Staden, 1984, Berg & v.Hippel, 1987) and show very clearly that the use of PWMs improves this approach a lot.

All the methods mentioned so far have been developed and applied to prokaryotic organisms, e.g. *E. coli*. But a lot of methods following this approach have been implemented for *D. melanogaster* and mammals. Rajewsky *et al.* (2002) developed the method `Ahab`, applied it to *D. melanogaster* and verified some of their predictions experimentally. Also Johansson *et al.* (2003), Bailey & Noble (2003) and Frith *et al.* (2003) developed PWM matching pro-



(a) motif-representation

A	0	0	0	0	1	1	1	0	0	1	1	3
C	5	0	1	0	1	0	0	0	3	9	9	5
G	6	12	11	10	3	2	0	0	0	0	1	5
T	2	1	1	3	8	10	12	13	10	3	2	0

(b) matrix of aligned TFBS

Figure 1.5: representation of the alignment of TFBS bound by the transcription factor *dorsal* Source: TFBS database: Jaspar (Stormo, 2000)

grams for Drosophila. Chan & Kibler (2005) compared these different methods. The results are shown in Chapter 3 of this thesis.

The difficulty of CRM prediction and TFBS localization increases with genome complexity and the average length of intergenic regions. Therefore the analysis of mammalian gene regulation is even more challenging compared to insects. `Match` (Kel *et al.*, 2003), `MatInspector` (Cartharius *et al.*, 2005) and `AliBaba` (Grabe, 2004) are programs which were developed to locate TFBS in mammals using PWMs. `Match` and `MatInspector` are commercialized methods using commercialized databases. `AliBaba` is freely available. PWMs are accessed from the freely available `TRANSFAC`-version or have to be provided to the program. The usability of these programs is very much limited by the quality of the PWMs. If few TFBS of a transcription factor are known the resulting PWM is not sensitive and many TFBS are missed. If the bound TFBS vary in a high degree the resulting PWM is highly unspecific which results in a lot of false positive results. For this reason Kechris *et al.* (2004) assigned informations contents to PWMs and use these to calculate the significance of a hit.

Another method to distinguish true positive matches of PWMs from false positive matches uses the concept of comparative genome analysis. Comparative analysis of genomic sequences assumes that the function of DNA is conserved between species. Blanchette *et al.* (2006) developed a program that scans orthologous sequences for matches of PWMs and increases the score of a hit, if it is conserved between species. This program was applied to mammals. Moses *et al.* (2004), Sinha *et al.* (2004) and Berman *et al.* (2004) developed programs following the same approach and applied them to Yeast and Drosophila. This approach applied to lower organisms resulted in accurate results and the results in higher eukaryotes were very useful for the design of experiments to identify CRMs.

Another modification of this approach using PWMs was applied by Markstein *et al.* (2002) to *D. melanogaster*. They searched for very well defined modules of TFBS of the transcription factor *dorsal* and *twist*. They score cooccurrences of PWM-hits, which have a certain distance from each other. They could identify many genes and the according

CRMs regulated by the given set of transcription factors. Palin *et al.* (2006) combined this approach with inter-species conservation of TFBS and applied this approach successfully to mammals.

1.3.3 Analysing homologous genes from different species

One approach to predict CRMs is based on an idea which has already been successfully applied to gene prediction and other annotation problems. Loots *et al.* (2000), Bergman & Kreitman (2001), Boffelli *et al.* (2003), Frazer *et al.* (2004), Johnson *et al.* (2004) and Woolfe *et al.* (2004) showed that conservation of noncoding DNA between genomes is a good indicator of biological function. Emberly *et al.* (2003) investigated the conservation of TFBS between *D. melanogaster* and *D. pseudoobscura* using the alignment programs `LAGAN` (Brudno *et al.*, 2003) and `SMASH` (Zavolan *et al.*, 2003). They calculated the overlap between conserved blocks returned by their alignment programs and annotated TFBS and stated that the overlap between conserved blocks and TFBS is still statistically significant but not much greater than by chance. These authors themselves suspected that alignment methods which are adapted to a conservation of regulatory function could result in a more clear conservation of regulatory function.

The correlation between sequence conservation and gene regulation is determined by the following conditions.

- According to Bilu & Barkai (2005) the average length of a TFBS is 12.5bp.
- A TFBS does not have to be perfectly conserved. Some positions can undergo substitutions without causing a difference in the affinity of the binding site to its transcription factor.
- Wilson (1975), Tautz (2000) and Khaitovich *et al.* (2006) stated that phenotypic differences between species may be attributable to differences in gene regulation.

Hence, signals of conserved gene regulation are short conserved motifs containing mismatches. If a pair of closely related species is analysed a signal of a conserved TFBS is not distinguishable from background conservation, because its difference from background conservation is not statistically significant. If distantly related species are analysed TFBS can have undergone too many substitutions to be detectable any more.

In lack of information about the binding transcription factors or coregulated genes sequence conservation is still the best approach to predict CRMs and the described problems are handled in different ways. Grad *et al.* (2004) developed a method called `PFR` which successively searches for conserved non-coding sequences and conserved non-coding subsequences between *D. melanogaster* and *D. pseudoobscura*. A first oder Markov chain is then trained with the transitions between the 5mer prefix and 5mer suffix of all 6mers. This Markov chain is used to discriminate regulatory and background sequence in non-conserved non-coding sequences. Additional modules which are not detectable by conservation are found this way. The results of this method applied to *D. melanogaster* are shown in Chapter 3. This method misses a high amount of known CRMs. Another problem of this method is that it is only applicable to genes which are regulated by the same transcription factors. The 5mers which are found to contribute mainly to the score of a region are contained within known TFBS. Thus only CRMs of coregulated genes can be found using this method.

Loots & Ovcharenko (2004) developed a method called `rVista` which identifies conserved regions between mammals and tests them afterwards for occurrences of known TFBS to exclude false positive predictions. This method returns high quality results but can only be applied to cases where the regulating TFBS are known.

1.3.4 *ab-intio*-prediction of CRMs using only the analysed gene and the neighbouring intergenic regions

In some cases no additional information than the gene and the neighbouring intergenic sequence is given. Neither closely related species nor coregulated genes are sequenced and

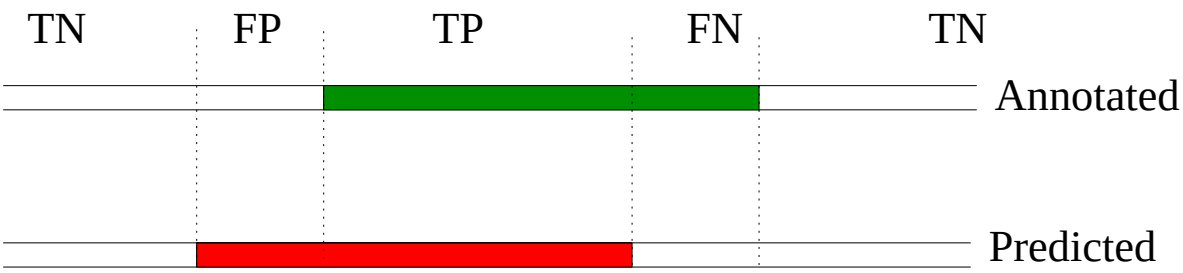


Figure 1.6:

no TFBS of the analysed organism are annotated yet. In this case a method has to use the statistical properties of regulatory DNA to localize CRMs. Since many details of the process of transcriptional regulation are not yet discovered the presence of TFBS is the only known reliable property of DNA which has to be detected by an ab-initio method. Kassis *et al.* (1989), Stanojevic *et al.* (1991) and Small *et al.* (1992) reported the presence of accessory weak binding sites in the DNA around a functional TFBS. The presence of weak binding sites combined with the fact, that TFBS are often repeated within a narrow region leads to redundancy of some motifs in a CRM, which can be detected by looking for locally overrepresented motifs.

There are two hypotheses why accessory weak binding sites are expected to occur close to a functional one. Kim *et al.* (1987), Khory *et al.* (1990) and Coleman & Pugh (1995) assume that degenerate binding sites close to a functional TFBS are necessary to guide the transcription factor to its binding site. Dermitzakis *et al.* (2003) assume that degenerate binding sites near or within CRMs build an evolutionary stock and can be substituted into functional TFBS quickly, if there is a need to adapt to environmental changes or if functional TFBS get lost by mutation.

Papatsenko *et al.* (2002) and Abnizova *et al.* (2005) developed methods following this approach and faced the following problems. One problem of this approach is the false positive rate. The DNA contains many motifs, which are overrepresented for various reasons. It is not possible to decide why a certain motif is overrepresented. Another problem is that the length and the allowed number of mismatches is not known for overrepresented functional and degenerate binding sites. These properties probably change for binding sites of different transcription factors and cannot be generally formulated. An additional problem is that we do generally not know anything about properties, such as oligo-nucleotide composition, of the region between TFBS. A CRM contains probably more information than the TFBSs and their affinities. But we do not exactly know which information they carry or how it is coded. Long intergenic regions may also carry more functions than we know so far, which may acquire overrepresented motifs. We can be sure that other proteins in addition to transcription factors bind specifically to the DNA (e.g. splicing enhancers, replication factors). We cannot exclude the possibility that a functional binding site for a splicing enhancer requires degenerate binding sites within the intron and we do not know any way to distinguish between a TFBS and a binding site for a splicing enhancer in this case.

1.4 Evaluation of CRM predicting methods

Assigning a function to a stretch of DNA or extracting all sequences with a certain function from a genome by computational methods has been a task since DNA has been sequenced. Methods have been developed to predict protein coding genes, alternatively spliced exons or introns, 5' and 3' untranslated regions and CRMs. The aim of the first prediction methods was to predict protein coding genes with exact start and stop codons and precise exon-intron boundaries.

Different groups have developed a variety of methods. To compare the accuracy of

different methods the overlap between known annotated genes and predicted genes was calculated. These calculations have been transferred to the evaluation of CRM prediction by Chan & Kibler (2005). As it was done for genes, predictions and annotations are partitioned into *TP (true positives)*, *FP (false positives)*, *TN (true negatives)* and *FN (false negatives)*. Predictions are evaluated on CRM and nucleotide level. Annotated CRMs are counted as TP, if their overlap with a predicted CRM extends 50bp. Otherwise they are counted as FN. Predicted CRMs are counted as TP, if they overlap an annotated CRM for at least 50bp. Otherwise they are counted as FP. TN cannot be counted on CRM level, because it is not possible to partition not annotated, not predicted sequence into countable truely not predicted CRMs.

Figure 1.6 shows the way nucleotides are partitioned into TP, FP, TN and FN. If a predicted CRM partially overlaps an annotated CRM, the nucleotides within the overlap are counted as TP. Nucleotides within the predicted CRM but not within the annotated CRM are counted as FP and predictions within the annotated exon but not within the prediction are counted as FN. Every nucleotide that is neither overlapped by a prediction nor by an annotation is counted as TN.

The value to calculate the amount of missed annotations is *sensitivity (SN)* and the formula is

$$SN = TP/(TP + FN) \tag{1.1}$$

The *positive predicted value (PPV)* describes the amount of predictions which is true and is calculated by

$$PPV = TP/(TP + FP) \tag{1.2}$$

The amount of false positive predictions within not annotated sequence is computed using *specificity (SP)*. The according formula is

$$SP = TN/(TN + FP). \tag{1.3}$$

PPV, which describes the amount of true predictions within all predictions strongly depends on the length of the analysed sequence. If the sequence is short the probability of a false positive prediction is limited by space. For this reason the comparison of the PPV values of different programs is only possible, if it was calculated on the same datasets. If programs have been applied to different datasets only SN and SP can be compared. Since SP can only be calculated on nucleotide level a comparison based on false positive predictions is only possible, when the performance of different methods is calculated on the same dataset.

The transfer of these measures to the evaluation of CRM predicting methods entails the following problems. It is possible that annotated CRMs exceed the boundaries of the functional CRM causing many false negative nucleotides. It is also possible that additional functional regions of a CRM exist outside an annotated CRM giving rise to nucleotides which are classified as false positive. More false positive nucleotides and predictions are caused by the fact that many CRMs are still unknown and not annotated yet. Thus false positive predictions are expected and high specificity may not necessarily reflect good accuracy of a program.

Another problem is that the equation *one prediction=one CRM* is not true for most of the prediction programs. Some CRMs are hit by two predictions and the mid-part of the annotation is missed. It is also possible that one prediction hits two or more annotations. PPV on CRM level is calculated assuming the equation above.

Because of these considerations we developed a standardized version of PPV to train our method, denoted PPV', which calculates the length-adjusted probability that a prediction hits an annotated CRM. This measure on the nucleotide level is defined as

$$PPV' = \frac{\frac{TP}{P}}{\frac{TP}{P} + \frac{FP}{N}},$$

where P means the number of nucleotides covered by annotated CRMs and N means the number of nucleotides not covered by annotated CRMs. In addition penalizing predictions that overlap annotated CRMs but do not hit the exact boundaries of the annotations should

be avoided. Thus, for a CRM prediction that hits an annotated CRM, the number of false positive nucleotides outside of the boundary of the hit annotation should not be used to calculate PPV′ or SP. Likewise, nucleotides in annotated CRMs outside the boundary of an overlapping prediction should not counted as false negatives in calculating SN.

The standardized PPV can be used to compare the amount of false positive predictions on different datasets. It is particularly useful to get an idea of the performance of a method during its development. The most meaningful measure to evaluate the amount of found annotations is SN on CRM level.

Chapter 2

Localizing CRMs using statistical analysis of the DNA and shortest unique substrings

2.1 Introduction

CRM predicting methods using information about the binding transcription factors or coregulated genes are limited by the requirement of extrinsic information. To overcome these limitations we developed a complete ab initio approach which uses exclusively the explored sequence as input. This method, `Shureg`, follows the approach 1.3.4, which is based on the assumption that locally overrepresented strings are caused by the presence of TFBS. There are two theories which support this approach. Both theories assume that the overrepresentation is caused by repeated degenerate binding sites.

1. The degenerate binding sites in the surrounding of the functional binding site should bind the transcription factor with low affinity and lead the protein to the functional binding site.

2. The degenerate binding sites are seen as a repository for new binding sites which can be made functional through few mutations when adaptation is necessary.

2.2 `Shureg`

The method `Shureg` can be partitioned into three large steps.

1. Calculation of the length of local shortest unique substrings *(shustrings)* and their *neighbours* at all positions relative to a defined window.

2. Calculation of P-values

3. Counting extreme shustrings

2.2.1 The concept of shortest unique substrings *shustrings*

The concept of shustrings was introduced by Haubold *et al.* (2005) to investigate correlations between sequence complexity and sequence function in a very efficient manner. Shustrings are substrings of a sequence which loose their property of uniqueness if they are shortened by one. A shustring analysis of a sequence can be done in two different ways. Global or local shustrings can be calculated. Global shustrings are unique relative to the explored sequence and shortest relative to all other unique substrings of the text. Local shustrings are the

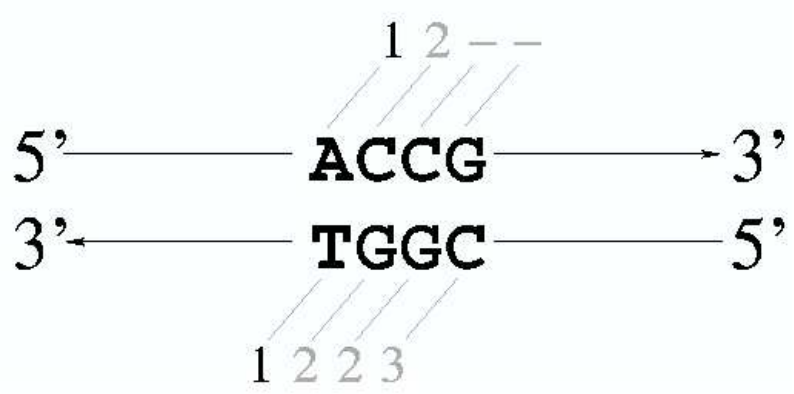


Figure 2.1: Local shustrings of the string ACCG and its complement.

5' CCCGGGAAATTTCTCCTTATAGAGCGCCAACTGTGGTACGATCGTTGCAGGCTAAGTCACATGACC 3'
3' GGGCCCTTTAAAGAGGAATATCTCGCGGTTGACACCATGCTAGCAACGTCCGATTCAGTGTACTGG 5'

Windowsize: 66		
Shustringlength:	forward:	4
	backward:	4
Number of neighbours:	forward:	1
	backward:	1

Figure 2.2: window of 66bp with shustrings and neighbours of both directions

shortest shustrings starting at a certain position. Shustrings starting at other positions can be shorter.

We use the text “ACCG” as example to explain this concept. The text “ACCG” has ten substrings {A, AC, ACC, ACCG, C, CC, CCG, C, CG, G}. Eight substrings are unique. The substrings “A” and “G” are the shortest unique substrings in the set of unique substrings and they are called *global shustrings*. Local shustring lengths are shown in Figure 2.1. A local shustring length is defined for every position of a sequence and its complement. The local shustring at position 1 is “A”. The local shustring length at this position is 1. “C” at position 2 is not unique. We elongate the substring one base and the resulting shustring “CC” is unique. The local shustring length starting at position 2 is 2. At position 3 the longest possible substring is “CG”. This word is repeated at position 1 of the complement string. We cannot elongate the string in a way that it fulfills the unique criteria. For this reason we put a sentinel at the end of the text. A sentinel is a letter which is not contained in the text, for example “$”. Then “GC” is elongated to “GC$” and the shustring length at position 3 is 3.

Another feature than the local shustring length describing the complexity of a sequence is the *number of neighbours of a shustring*. Neighbours can be regarded as maximal repeats contained in the window. When we calculate the length of a shustring, we also know, which strings are the exact copies of the strings, which are one basepair shorter than the shortest unique string. The exact copies are called neighbours and their number of occurrences is called number of neighbours. An example of a text of 66bp is given in Figure 2.2. The local shustrings starting at the middle position of the window are marked with red letters. The neighbours of these shustrings are marked with blue letters.

`Shureg` is a method to measure local representation of patterns. Therefore local shus-

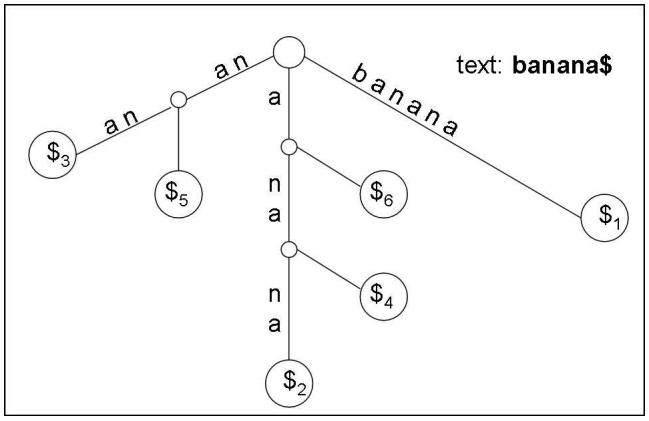


Figure 2.3: Suffix tree of the string `banana` - Every suffix of `banana` is represented as a path from the root to a leaf.

trings are calculated for every position of the explored sequence relative to a surrounding window of user defined size. Also the number of neighbours is counted in the local surrounding window.

For this purpose a window is shifted in 1bp steps along the sequence. The shustring length and the number of neighbours is calculated for the middle position of every window for the backward and the forward strand. Subsequences of $1/2*windowsize$ at the beginning and end of the whole sequence cannot be analysed, because the surrounding window does not exist completely.

2.2.2 Computation of shustringlength and number of neighbours using suffix trees

To calculate Shustrings and their numbers of neighbours a suffix tree is built for every excised window of the analysed sequence. The used algorithm to built suffix trees in linear time and memory requirements, named *Ukkonen algorithm* is explained in Gusfield (1997). We explain an efficient way to read the shustring length and number of neighbours for a certain position given a tree.

Figure 2.3 shows the suffix tree for the word `banana`. All 6 suffixes `banana`, `anana`, `nana`, `ana`, `na` and `a` are represented by paths from the root to leafs. Every leaf represents one suffix and holds the information about the starting position of the assigned suffix in the string. Inner nodes show multiple occurrences of the substring starting at the root ending at the inner node. The number of leafs below an inner node is equal to the number of occurrences of the represented substring.

Substrings which are unique in the explored string end below the last inner node of a path from the root to a leaf representing one suffix. To find the minimal length of a unique substring we have to follow the assigned path through the tree to the last inner node and elongate this substring by one. The paths representing the shortest unique substrings starting at positions 3 and 4 of the word `banana` are shown in Figures 2.4 and 2.5.

The number of neighbours of a shustring corresponds to the number of leafs following the last inner node in the substring path. These nodes are colored in pink in Figures 2.4 and 2.5. In both demonstrated cases the number of neighbours is equal to one.

The number of steps analysing one position is linear to the user-defined windowsize: The complexity of an analysis of a sequence of length n with a window size w is then $O(n * w)$. To implement the program we used a collection of C programs called `strmat`, which was initiated by D. Gusfield and developed at the UC Davis.

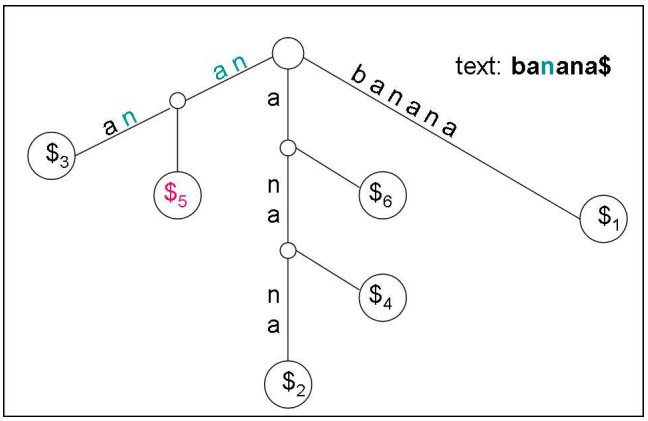


Figure 2.4: Suffix tree of the string `banana` - the shustring starting at position 3 is marked in green. Leafs of paths representing neighbours are marked in pink.

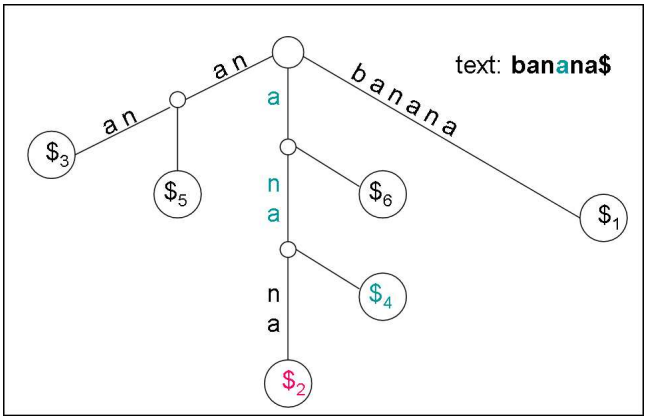


Figure 2.5: Suffix tree of the string `banana` - the shustring starting at position 4 is marked in green. Leafs of paths representing neighbours are marked in pink.

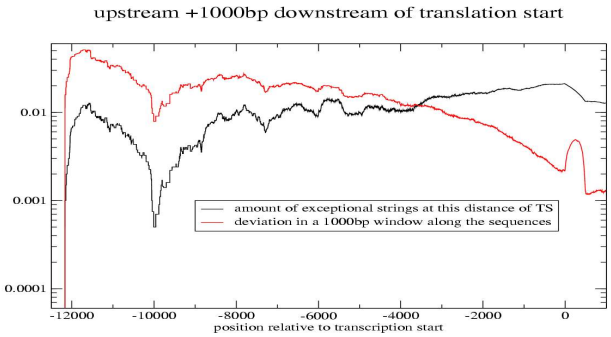


Figure 2.6: **Analysis of yeast upstream regions aligned at the TSS of all annotated yeast genes**
black line: `Shureg`-score summed over all upstream regions.
red line: variation of the `Shureg` scores along the sequences.
The positions are evaluated based on their distance to the downstream TSS.

2.2.3 Significance of the number of neighbours dependent on the length of local shustrings

In the second step we calculate one P-value for each direction at every position. One P-value describes the probability of the shustring to have at least the observed length and the observed number of neighbours, considering the gc-content of the sequence and the gc-content of the repeated part of the shustring. The number of neighbours depending on the length of the shustring is Poisson distributed. Because we calculate the cumulative probability according to a Poisson distributed probability we use the gamma-function and the formula is

$$\text{P-value} = e^{\lambda} * \frac{(1 - \Gamma(\#(neighbours), \lambda))}{(e^{\lambda} - 1)}$$

with

$$\lambda = 2 * 3 * w * (0.5 - gcb)^{len - \#gc} * (gcb)^{\#gc},$$

where w is window size, $\#gc$ means the number of Guanines and Cytosines in the shustring, gcb means the gc-content in the background sequence and len is the shustring length.

We consider positions where the P-value for the shustring and its number of neighbours is smaller than 0.05 to be a signal of a CRM. We find positions where the shustring is unexpectedly long which have at least one neighbour and we find short shustrings with an unexpected high number of neighbours. Both properties represent overrepresentation.

2.2.4 Counting extreme shustrings

The output of `Shureg` is a curve. The number of positions with P-values smaller than 0.05 in a smoothing window is counted. The window size is a variable parameter. The number of extreme positions divided by the doubled window size is the score of the position to be contained in a regulatory region. The window size has to be doubled, because we are observing forward and backward strand.

2.3 Results

2.3.1 Application to yeast

To calculate the graph shown in Figure 2.6 we applied our method to the whole yeast genome, aligned the yeast genes at their transcription start sites *(TSS)* and added the score of all positions relative to the TSS. The result is shown in Figure 2.6.

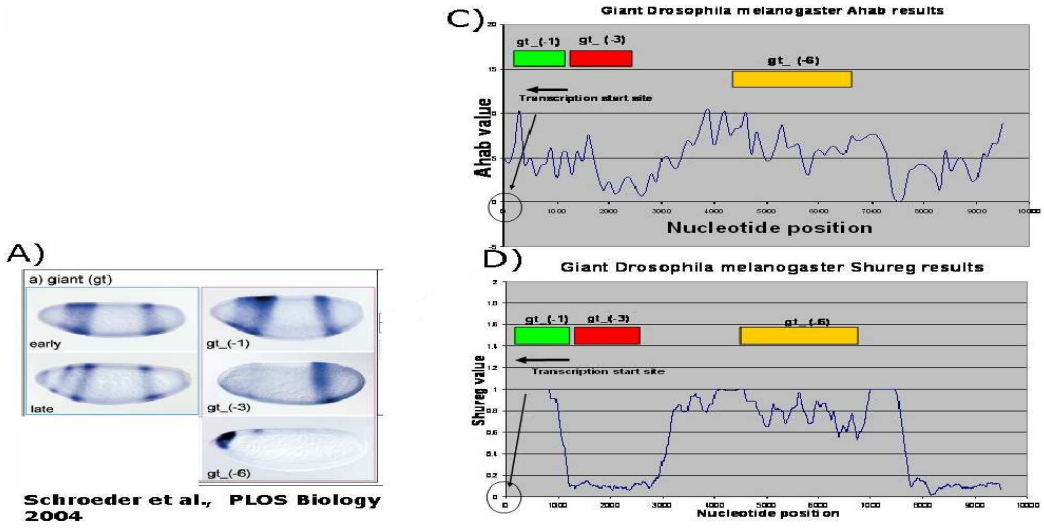


Figure 2.7: Analysis of CRMs of the gene *giant*
A) expression patterns of CRMs
C) prediction of `Ahab`
D) prediction of `Shureg`

A fact which is clearly shown in Figure 2.6 is that the `Shureg` score becomes higher the closer the position is to the TSS. The high scores between 10000bp and 12000bp upstream are caused by one repeat in the only intergenic region which is of that length. Yeast contains only one upstream region of this length. Thus, the sum of scores is divided by one at this position and one repeat can push the score to a huge degree.

2.3.2 Application to Drosophila developmental genes regulated by well defined CRMs

The program was applied to the intergenic regions surrounding the developmental genes of *Drosophila melanogaster*. The regulation of these genes is well known and a lot of experiments have been done to define the regions regulating these genes.

We compared the results of `Shureg` with the results of `Ahab` (Rajewsky *et al.*, 2002). `Ahab` is a program which matches PWM's of known transcription factors to the DNA and scores every region according to the ratio

$$\frac{P(\text{seq}|\text{seq matches a PWM})}{P(\text{seq}|\text{seq is background sequence})}.$$

If the score exceeds a certain threshold, the region is considered to be of regulatory function. The `Ahab` program cannot be applied to genes whose regulating transcription factors are not known.

Figures 2.7, 2.8 and 2.9 show the predictions of the regulatory regions of the genes *giant (gt)*, *hairy (h)* and *short gastrulation (sog)*. The upstream region of *gt* contains three known CRMs. The expression patterns driven by the CRMs are shown in Figure 2.7A. The enhancer labeled *gt_(-1)* enhances the early expression pattern. The enhancers labeled *gt_(-3)* and *gt_(-6)* drive the late expression pattern. gt_(-6) effects the expression of the anterior domain and *gt_(-3)* induces one stripe in the posterior part of the embryo.

gt_(-1) is clearly found by `Ahab` and `Shureg`. `Ahab` produces a peak in the region of *gt_(-3)* and predicts the enhancer with a low score. `Shureg` does not find *gt_(-3)*. *gt_(-6)* is also predicted by both programs. Both programs predict a large region containing *gt_(-6)*.

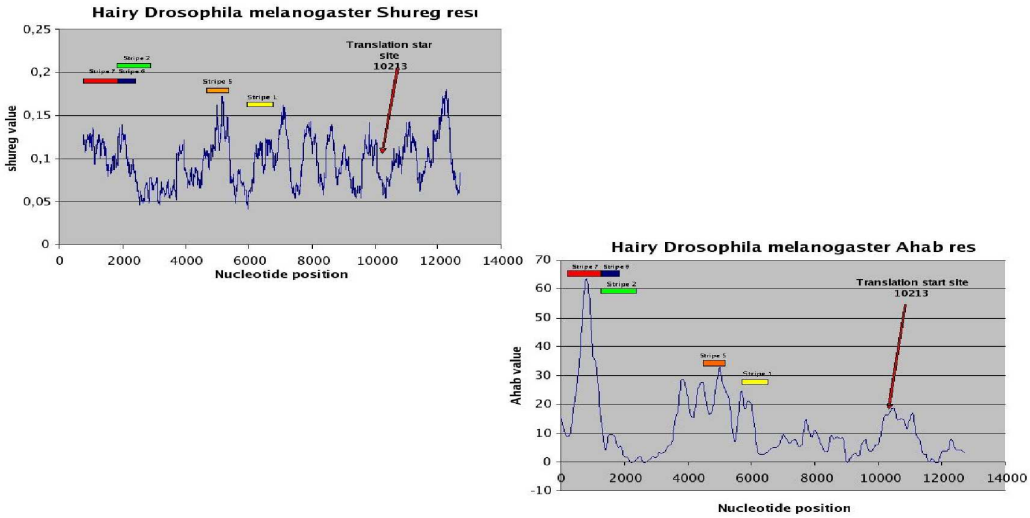


Figure 2.8: Analysis of CRMs of the gene *hairy*
Prediction of the programs `Shureg` and Ahab

The prediction of the CRMs of the *hairy* gene is shown in Figure 2.8. In the *hairy* upstream region five enhancers are experimentally verified *(stripe7, stripe6, stripe2, stripe5, stripe1)*. *Hairy* shows an expression pattern of seven stripes during the Drosophila development. Every enhancer drives the expression of one stripe. This is a strong indicator that there are still unknown enhancers in this region.

`Ahab` identifies the stripe7 enhancer with very high significance. Also the stripe5 and the stripe1 enhancers are predicted by `Ahab`. The overlapping stripe6 and stripe2 enhancers are missed in the `Ahab` prediction. Additional to the experimentally verified enhancers `Ahab` predicts two CRMs. It is not possible to evaluate the quality of this prediction, because nothing is known about the regulatory properties of these regions.

`Shureg` finds highly significant overrepresentation in the enhancers, stripe7, stripe6, stripe2 and stripe5. The stripe1 enhancer is also predicted, but not as strong as the others. Downstream of the stripe1 enhancer we find four additional peaks. Another CRM is predicted between the stripe2 and the stripe5 enhancer. This prediction overlaps one of the additional `Ahab` predictions. `Shureg` does not miss one of the known CRMs regulating the transcription of *hairy*. But it predicts five more CRMs. It is unlikely that all of these predictions are true. It is not possible to evaluate these predictions with a reasonable amount of confirmation.

Figure 2.9 shows the predictions for the region surrounding the gene *sog*. One CRM is experimentally verified to regulate the transcription of this gene. This region is marked with a white square. The two exons including the intron with the CRM are marked by a yellow and a red square.

In Figure 2.9C the `Ahab` prediction is shown. The PWM's, which are experimentally verified to be transcription factors of *sog* are given as input for `Ahab`. The one known enhancer of *sog* is clearly found. Additional predictions have a much lower score.

Figure 2.9D shows the `Shureg` prediction. This prediction is unspecific. We see a peak in the region of the experimentally verified CRM but almost the whole intron is scored higher to be a CRM.

In Figure 2.9B another `Ahab` output is shown. Here we simulate the situation that the transcription factors of the gene are not known. The input additional to the sequence consists of all known PWM's of Drosophila transcription factor binding sites. The `Ahab` prediction without the knowledge of the responsible transcription factors is also very unspecific.

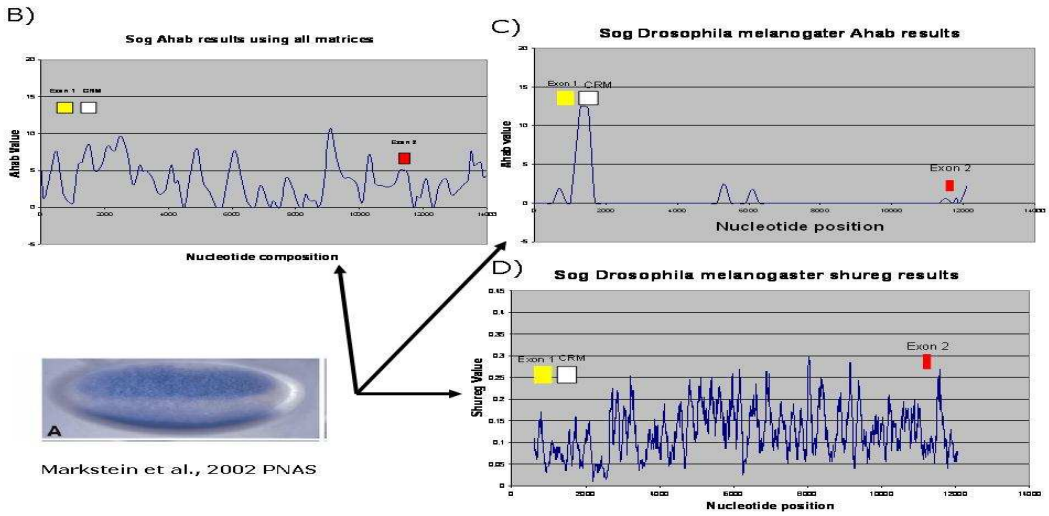


Figure 2.9: Analysis of CRMs of the gene *short gastrulation*
A) expression patterns of CRMs
B) prediction of the program `Ahab` using all known PWM's
C) prediction of `Ahab` using PWM's of transcription factors which are known to regulate *sog*
D) prediction of `Shureg`

2.4 First Conclusion

The predictions of `Ahab` match the known enhancers in some cases better than the predictions of `Shureg` do. `Ahab` misses only two of all known enhancers and is able to predict seven enhancers correctly. `Shureg` predicts also seven of nine known enhancers. One enhancer is missed and the prediction for the *sog* region is not reasonable. `Ahab` predicts two additional regulatory regions and `Shureg` predicts five additional CRMs. As we can see from this analysis the false positive predictions have to be reduced.

Until this point we ignore one very important fact in the evaluation of the programs. we do not consider the information about the regulation of the gene which has to be provided to the programs. If this is considered, we have to evaluate the results differently. `Ahab` helps to complete the knowledge about the regulation of a well investigated gene. The transcription factors, which bind to the CRMs have to be known. `Shureg` predicts the CRMs based on the raw sequence data and does not use any extrinsic information.

Regarding this fact `Ahab` still performs better than `Shureg` in some well known cases. When we analyse genes which are not included in a well understood process `Ahab` cannot be applied in a meaningful way.

Figure 2.9 shows a case, where neither `Ahab` nor `Shureg` are able to locate the CRM, if the regulating transcription factors are not given to the program. To improve the predictions of `Shureg` further analysis was done to find properties of CRMs, which enable us to distinguish between motifs which are overrepresented because of regulatory function and those which are overrepresented because of other reasons.

2.5 Further analysis

2.5.1 Method

To improve the method `Shureg` one of our goals is to distinguish between true positive *(TP)* and false positive *(FP)* predictions. Following this goal, we had to find a bigger dataset to partition the predictions into TP and FP predictions. The regions which are not predicted

to be regulatory are partitioned in the true negative *(TN)* and false negative *(FN)* set. The dataset we used was collected by Lifanov *et al.* (2003). It contains 74 CRMs regulating 20 Drosophila developmental genes.

The analysis is performed in a way that all CRMs are found and that FP are not avoided to strictly. Then we search for special features of TP peaks which are not present in the FP peaks.

For the evaluation and partitioning of the peaks every single position is considered . If the score at one position of a CRM is too low it is considered not to be regulatory even when all neighbouring positions exceed the threshold. Also at a boundary of a CRM a position which is evaluated below the score is predicted as a FN. This evaluation is very strict given the fact that exact boundaries of CRMs are rarely known.

To find the threshold for what score a position is considered to be regulatory or not, we calculate the mean of the whole region and define the threshold as $(1/2 * mean)$. Despite the strict evaluation we achieve a sensitivity of 0.8069 and a specificity of 0.1918.

In a next step we analyse the shustrings at the positions in the different sets TP, FP, TN and FN. First we calculate the base frequencies in all shustrings of the four sets. Here we find very similar values in all four groups.

In a next step we want to observe the structure of the shustrings. For this purpose we calculate two properties:

1. number of different bases in a shustring having a base frequency larger than 0.1 in the shustring.

2. transition probabilities within one sequence

number of different bases in a shustring with a base frequency larger than 0.1

Because we cannot distinguish between the four groups using the base frequencies, we want to investigate the base distribution of the regions. For this purpose we want to know the number of different bases in shustrings. In case of low sequence complexity we expect long shustrings consisting of one or two different bases. In case of high sequence complexity we expect shustrings containing three or four bases. We did this analysis two times. First we analysed all found shustrings divided in the four subsets TP, FP, FN and TN. The result is shown in Table 2.1. In the second analysis we count the number of different bases only in shustrings with a p-value < 0.05. Table 2.2 shows the result of this analysis.

When we evaluate the results we can cluster the subsets into two groups. We expect the same results for TP and FN, because these positions are known to be in CRMs and for the subsets FP and TN, because they are not known to be part of a CRM.

Table 2.1 shows a marginal trend of the known regulatory shustrings to be more complex. The amount of shustrings representing mainly long stretches of one base is higher in regions which are not known to be regulatory. Especially FN lack stretches of simple sequence order. They are enriched in shustrings containing three different bases. This trend is shown more clearly in Table 2.2.

But also in Table 2.2 we can see only a trend and not a statistically significant difference. Thus we do further analysis and calculate base transition probabilities in one sequence.

2.5.2 transition probabilities within one sequence

This analysis resembles a Markov model of order 1. We calculate the probability of a base to occur at a position observing the base at the previous position. The results are shown in Figure 2.10 and 2.11.

Most of the results do not show a clear difference. Remarkable in all pictures is the fact that the bar of the subset FP and the bar of the subset TN are of almost identical height

	4	3	2	1
TP	30.48%	52.99%	16.13%	0.40%
	18.55%	18.40%	18.46%	17.64%
FP	30.57%	52.78%	16.20%	0.45%
	34.83%	34.32%	34.71%	37.24%
FN	30.78%	54.02%	14.96%	0.24%
	4.48%	4.49%	4.09%	2.57%
TN	30.37%	53.11%	16.10%	0.42%
	42.87%	42.79%	42.74%	42.55%

Table 2.1: **Composition of shustrings from the different subsets TP, FP, FN, TN.** To distinguish between the shustrings from these four subsets, we calculated the base frequencies within every shustring and counted the number of bases with a frequency > 0.1. These numbers are the headers of the columns. The black row describes the amount of all shustrings of one subset which contains a certain number of different bases. The values in the blue row tell the amount of shustrings which contain the certain number of bases in one subset.

	4	3	2	1
TP	28.15%	52.28%	17.88%	1.69%
	20.89%	20.65%	19.65%	19.61%
FP	27.59%	51.55%	18.99%	1.87%
	37.54%	37.33%	38.28%	39.92%
FN	26.76%	58.41%	14.83%	-
	1.30%	1.51%	0.19%	-
TN	27.64%	51.90%	18.88%	1.76%
	40.26%	40.51%	41.00%	40.47%

Table 2.2: **Composition of the shustrings with a p-value < 0.05 from the different subsets TP, FP, FN, TN.**
To distinguish between the shustrings which cause peaks from these four subsets, we calculated the base frequencies within every shustring and counted the number of bases with a frequency > 0.1. black and blue rows have the same meaning as in Table 2.1

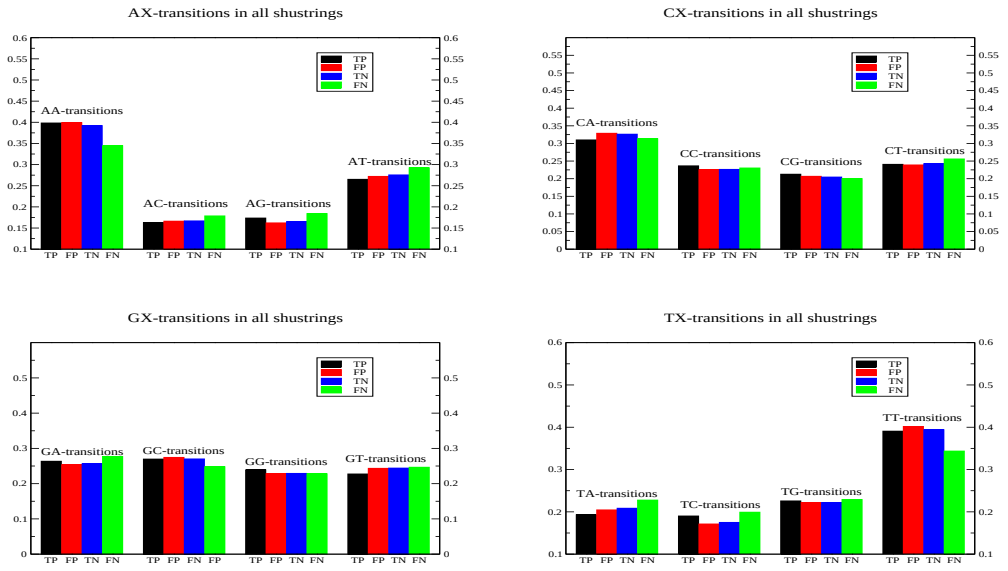


Figure 2.10: base transition probabilities in all shustrings divided in the four subsets TP, FP, FN. TN

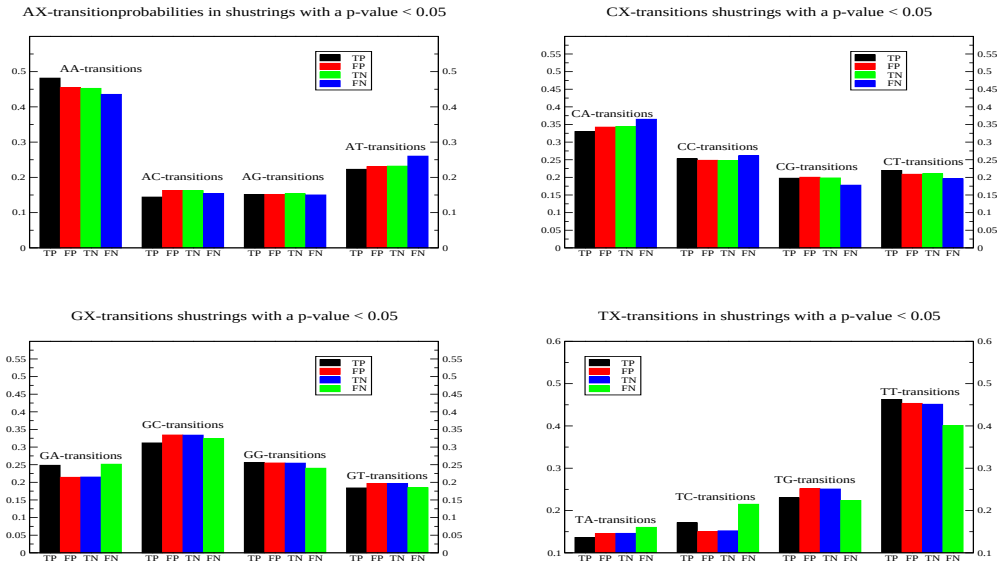


Figure 2.11: base transition probabilities of the shustrings with a p-value < 0.05 from the four subsets TP, FP, FN, TN

in every case. The height of the TP-bar and the height of the FN-bar differ from the FP and TN values in many cases. But the deviations from these bars do not only differ in amount but also in leading sign. Only in one case we see a trend in the same direction. The probability of a cytosine following a thymine is higher in the FN-subset and in the TP-subset than in the FP-subset and TN-subset. This effect becomes more clear, if we consider only the shustrings, whose position have a p-value < 0.05, as shown in Figure 2.11.

2.5.3 Analysis of upstream regions of homologous genes

Another possible way to improve the accuracy of our program is to include information about sequence conservation. Our first result is shown in Figure 2.12. For the right picture in Figure 2.12 we aligned the fushi tarazu *(ftz)* region of the D. melanogaster genome with the ftz region from the D.virilis genome. To compute shustrings we excised two 1000bp windows around the aligned positions. One window is taken out of the melanogaster sequence and one window is taken out of the virilis sequence. Now we calculate the shustring starting at the aligned position according to the forward and backward strand of both sequences (4000bp). In the next step we calculate a p-value to find a shustring of this length in a 4000bp stretch. Extremely short and extremely long shustrings are evaluated with a small p-value. In the next step we count the positions with extraordinary shustrings in a 200bp window and divide the number of positions with a p-value < 0.05 by 200bp.

Figure 2.12 shows clearly that information from a second sequence can improve the prediction a lot.

2.6 Concluding remarks

`Shureg` gives high scores to regions, where overrepresented motifs are present. To gain a low p-value a shustring has to have more neighbours than expected. That means the shustring without the last position which is a repeat of maximal length has to be repeated more times than we would expect. A long shustring with one neighbour can also have a low p-value, because it is unlikely to have an exact repeat of a long stretch in the sequence.

Our results show clearly that CRMs contain overrepresented motifs. In yeast gene regulation is simpler than in higher eukaryotes and CRMs are located in nearby upstream region

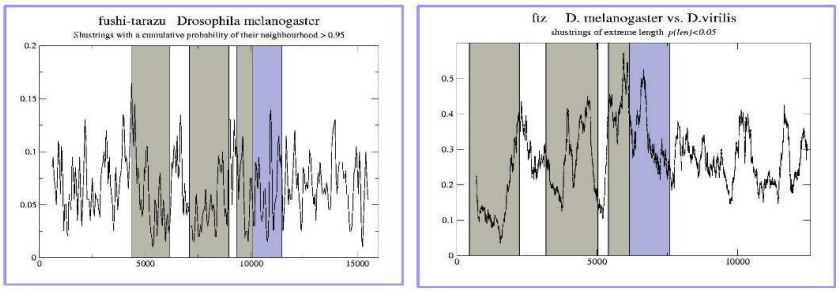


Figure 2.12: 2-dimensional analysis of th CRM of the gene fushi tarazu.
In the left picture we show the same analysis as in Figure 2.7, 2.8 and 2.9. In the right picture we calculate shustrings in a window, that contains 1000bp of D. melanogaster sequence and 1000bp of D. virilis sequence.
Regulatory regions are marked with a gray background. Blue background shows the location of the CDS.

of the gene. Figure 2.6 shows that the number of significantly overrepresented patterns increases when the distance to the gene is decreased.

In Drosophila most CRMs are enriched in shustrings with a high number of neighbours. But many regions which are not verified to be a CRM contain overrepresented patterns too. We do not know about the regulatory function and cannot evaluate the quality of these predictions.

To learn more about the regions which are highlighted by `Shureg` we did some further analysis of the data and used an additional nice property of shustrings. Shustrings show special properties of the DNA more clearly than the raw sequence. If we investigate a GC-rich sequence, the GC-content is even higher in the shustrings, because GC-rich shustring are in average longer and have a denser neighbourhood than AT-rich shustrings, because GC-rich motifs are more likely to occur. This way also other properties of the DNA are shown more clearly with shustrings and might be found in our analysis.

But also the enhancement of DNA-properties does not show significant differences between DNA accomplishing regulatory function and background sequence. These results can be caused by two different effects of noisy data. One possible reason is just the theoretical chance to get the result by analysing samples of different size. The other possible reason is that the known regulatory modules are not minimal. Thus they can contain non CRM positions at their borders. The other fact is that the sequences contain unknown CRMs. The positions which are clustered into the false group can destroy or lower the effect.

A first result of this chapter is that CRMs contain overrepresented patterns. But local overrepresentation can be caused by other reasons and appears also in regions of the DNA, which accomplishes other functions. Using locally overrepresented patterns as the only sign of regulatory function leads to a high rate of false positive predictions.

The concluding result is that the current data situation does not allow to develop a reasonable program to predict CRMs completely ab initio. Figure 2.12 shows that the information about sequence conservation helps a lot to identify CRMs. Based on this result we developed the method `CisPlusFinder` which is explained in Chapter 3.

Chapter 3

Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA

3.1 Introduction

Based on the results of Chapter 2 we decided to develop a method to identify CRMs based on information about sequence conservation between species. Grad *et al.* (2004) compare *D. melanogaster* with *D. pseudoobscura* to identify CRMs in highly conserved regions. Boffelli *et al.* (2003) use a method called phylogenetic shadowing which calculates a homology measure between more than two species to predict functional regions. Methods using conservation signals alone face the problem that TFBS conservation necessary to maintain regulatory function may not be significantly higher than in non-binding sequence (Emberly *et al.*, 2003) and TFBSs can be gained or lost even when CRM function is conserved (Ludwig *et al.*, 2000). In addition, sequence conservation depends on the local mutation rate and selective constraints which vary over the genome.

Here, we show that CRMs can be identified in *D. melanogaster* by combining conservation signals with the property of TFBS to contain overrepresented motifs. Our method, called `CisPlusFinder` (Pierstorff *et al.*, 2006), locates sequences that are both perfectly conserved in multiple genomes and contain an overrepresented core motif as signal of a TFBS, and subsequently clusters sequences with these properties to predict CRMs. We developed and applied our method in the genus Drosophila, a species group that is well suited for a multiway comparative analysis since draft genome sequences are available for *D. melanogaster* and 11 other species at http://rana.lbl.gov/drosophila/. Our results indicate that combining conservation and compositional signals from a set of multiple closely related species which sum to a sufficiently high substitution rate can lead to better CRM predictions than has generally been achieved thus far.

3.2 Methods

3.2.1 Overview over `CisPlusFinder`

The motivating idea behind `CisPlusFinder` is to find regions in a target genome which contain a high density of *Perfect Local Ungapped Sequences (PLUSs)* that are shared among a set of closely related species. Individual PLUSs are selected on the basis that: (1) the length of a PLUS is unlikely relative to random occurrence; (2) at least one core motif contained in a PLUS, which is essential for the binding of the transcription factor, is locally

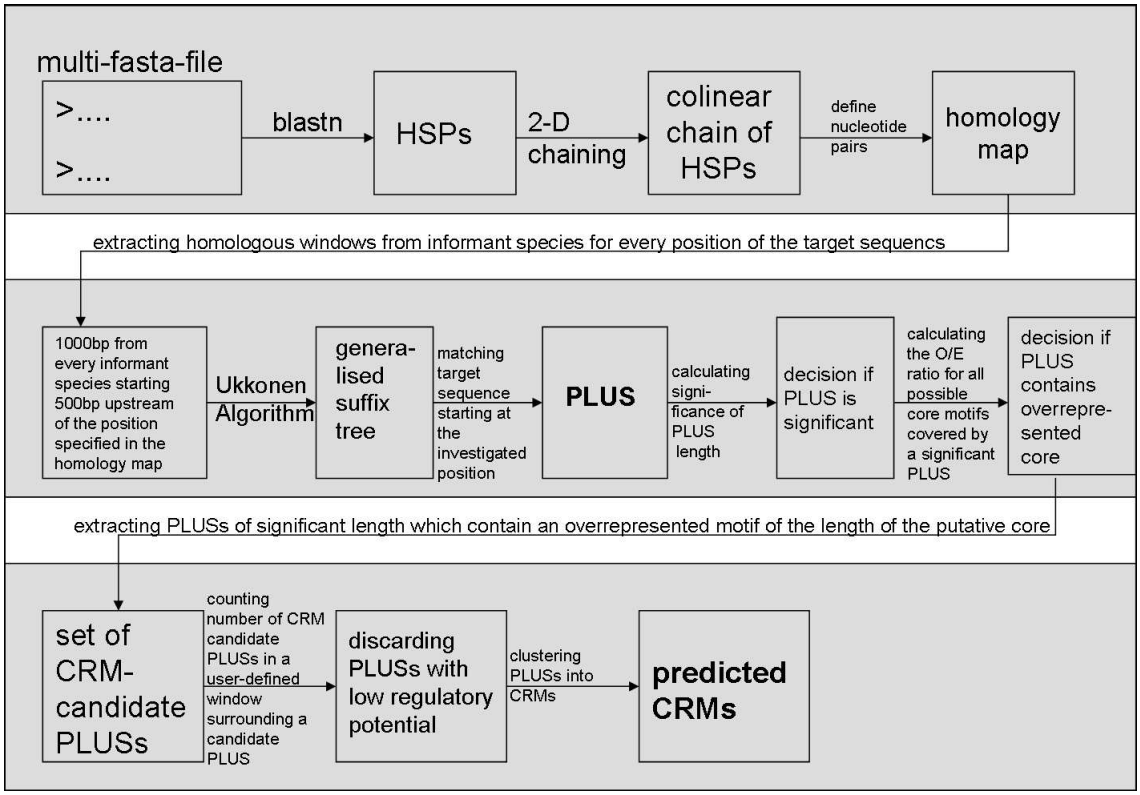


Figure 3.1: Flowchart describing the `CisPlusFinder` method.

overrepresented, and (3) some additional PLUSs fulfilling conditions (1) and (2) occur within the close neighbourhood. Underlying these requirements is the idea that PLUSs contain core TFBS motifs, and that local clustering of PLUSs represents the comparative genomic signal of clustered TFBS that typify CRMs.

Figure 3.1 shows a flowchart describing `CisPlusFinder`. The first section of Figure 3.1 shows the calculation of the nucleotide homology map, which is based on chained `blastn`-hits (Altschul *et al.*, 1990). A detailed description of the method is given in Section 3.3.1. The second part of Figure 3.1 depicts the stepwise calculation of putative PLUSs, which have a significant length and contain within them a core motif that is overrepresented in a local window centered on the PLUS. In order to find the set of putative PLUSs suffix trees are built using Ukkonen's algorithm (Gusfield, 1997) as explained in Section 3.3.2. The equations to calculate the significance of the length of a PLUS and the overrepresentation of possible core motifs are explained in the following Section 3.3.3. The last section of the flowchart depicts the selection of PLUSs which possess at least the minimal regulatory potential, and the clustering of PLUSs into CRMs. The clustering process is explained in Section 3.3.6. King *et al.* (2005) assign a regulatory potential to a motif according to supervised machine learning. Their definition of "regulatory potential" is not related to ours.

3.3 Detailed description of the `CisPlusFinder` method

3.3.1 Computing the homology map

To locate PLUSs in a set of homologous sequences we calculate a set of pairwise alignments in several steps. For each contiguous subsequence of arbitrary length in the target genome, we obtain orthologous regions for each of the informant species using precomputed

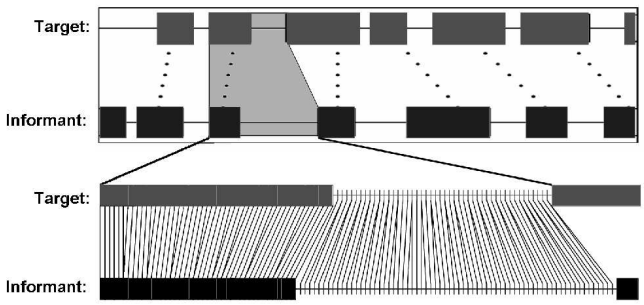


Figure 3.2: Scheme of homology map as basis for the window analysis. The upper part of the figure shows chained HSPs indicated as blocks. Regions between HSPs are shown as lines. The lower part shows an alignment of single positions. Regions outside of HSPs are paired and gaps are distributed in relative proportions across the segments. Gaps in HSPs are distributed in relative proportions across the HSP.

pairwise whole genome alignments, such as those produced with `Lagan` (Brudno *et al.*, 2003) accessible at the VISTA server (Couronne *et al.*, 2003). These homologous sequences are locally aligned using default parameters of `Blastn` (Altschul *et al.*, 1990) and collinear HSPs are chained with a two-dimensional chaining method (Wiehe *et al.*, 2001). The resulting pairwise alignments are used to assign every position in the target genome to a homologous position of the informant sequence to create a nucleotide level homology map as shown in Figure 3.2. Simple gaps in informant species relative to the target genome are treated by assigning two target positions to the same position in the informant sequence. Regions not covered by HSPs (such as those arising from small inversions) are forced to pair in the nucleotide homology map as shown in Figure 3.2.

3.3.2 Calculating PLUS length for every position using suffix trees

Based on the nucleotide homology maps of the pairwise alignments the following steps are performed for every position of the target sequence .

1. Extract a window of +/- 500bp around the homologous position in the informant species.
2. Build a generalized suffix tree of the 1000bp for each informant sequence and its reverse complement (Gusfield, 1997) as shown in Figure 3.3.
3. Match the target sequence starting at the investigated position against the tree and find the maximal PLUS which is shared by all species. An efficient way to do this is demonstrated in Figure 3.4. Only the maximal PLUS is recorded when one PLUS is totally contained in another one. Partially overlapping PLUSs from different positions are considered separately.
4. Calculate the significance of the PLUS x dependent on its length, `GC`-content and the `GC`-content of the informant sequence using Equation (3.1). The required significance level of a PLUS is 0.01.
5. Test all possible core motifs (c-mers covered by a PLUS) if they are locally overrepresented in a user defined window of size w in the target sequence and all informant sequences.
6. Count significantly long PLUSs containing an overrepresented core in a window of $r$bp surrounding the explored PLUS. The PLUS is discarded if less than $m-1$ other PLUSs are found in this window.

strings:

ACCGGTTACG$

GAATTCGGTA$

CGGTTAATCC$

TCCCGGTTAT$

CGGCGGTATT$

suffixes starting with C:

C$	3;10
CC$	3;9
CCGGTTAT$	4;3
CCGGTTACG$	1;2
CCCGGTTAT$	4;2
CG$	1;9
CGGTA$	2;6
CGGTTAT$	4;4
CGGTATT$	5;4
CGGTTACG$	1;3
CGGTTAATCC$	3;1
CGGCGGTATT$	5;1

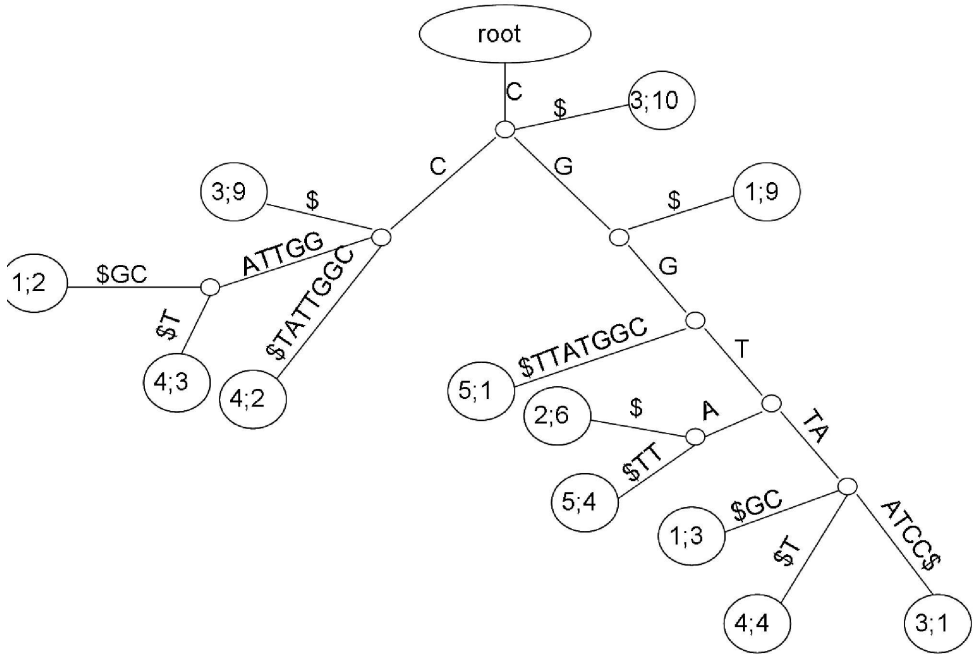


Figure 3.3: section of the generalized suffix tree of the five strings in the upper left corner; The section represents only the suffixes starting with "C". The leafs of a generalized suffix tree contain information about the string and the starting position of the suffix.

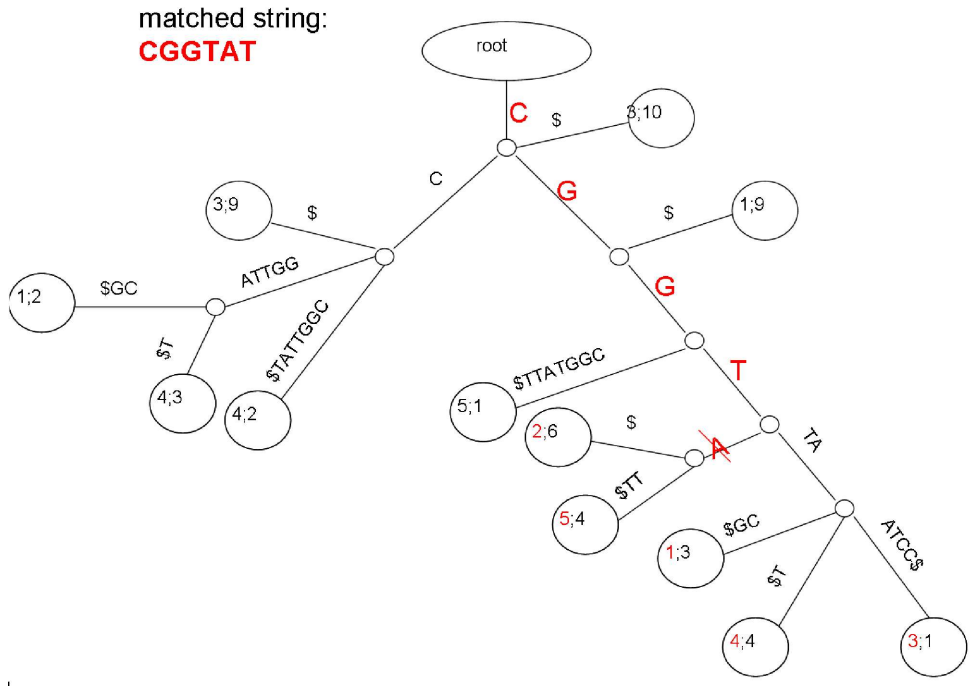


Figure 3.4: path along a string through the generalized suffix tree; The path of the string "CGGTAT" is marked with red letters. A PLUS has to be substring of all species. The maximal PLUS length is reached, if not one leaf from every string is in the subtree below the last node of the matching path. The PLUS ends after the fourth position of the target sequence and is "CGGT".

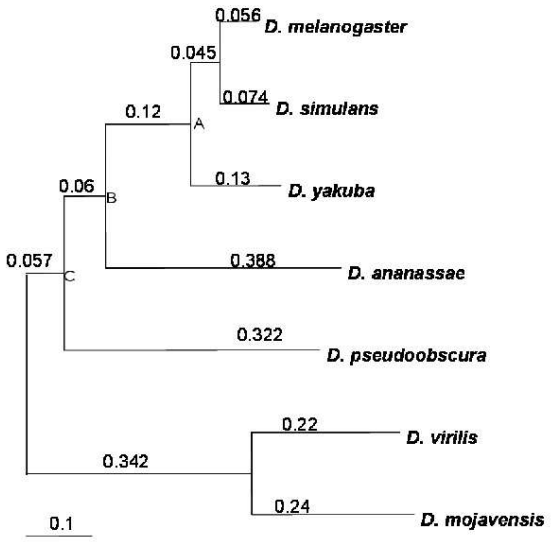


Figure 3.5: Phylogenetic tree of seven Drosophila species, based on non-conserved regions (data from A. Siepel, pers. comm.) and created with `PhastCons` (Siepel *et al.*, 2005) based on a multiple alignment using `Multiz` (Blanchette *et al.*, 2004)

3.3.3 Significance of perfectly conserved matches in multiple alignments

Crucial for the idea of `CisPlusFinder` is the choice of the informant species according to the target species. Species should be close enough such that homologous functional sites are conserved, but distant enough such that PLUSs of the size of a typical binding site, around 10bp in length, are unlikely to be due to common ancestry.

Given an x-mer in a target sequence, let π_x^{cons} be the probability to find in each of a set of evolutionarily related informant sequences a perfectly matching x-mer. Further let π_x^{rand} be the probability to find a perfectly matching x-mer in a random sequence with the same `GC`-content as the informant sequence. Below we show how these two probabilities can be put in relation to each other.

To test `CisPlusFinder` we used a subset of twelve currently available assembled genome sequences of Drosophila species which can be found at `http://rana.lbl.gov/drosophila/`. This set of genomic sequences provides an excellent resource to find PLUSs in *Drosophila melanogaster* with informant sequences from congeneric species. The phylogenetic relationship between seven Drosophila species is shown in Figure 3.5. The tree is based on distances in nonconserved regions, which have been compiled by A. Hinrichs and A. Siepel using `PhastCons` (Siepel *et al.*, 2005) and a `Multiz` (Blanchette *et al.*, 2004) multiple alignment. It agrees with the commonly accepted Drosophila phylogeny (Powell, 1997). Numbers along the branches denote per-site substitution rates. Since a given site may be affected by multiple substitution events, the number of mismatches in the underlying alignment usually underestimates the substitution rate. Models of evolution establish a relationship between substitution rate, k, and fraction of mismatches, d. Applying the simplest such model, the Jukes-Cantor model (Jukes & Cantor, 1969), we expect for a substitution rate of $k = 0.9$ between *D. melanogaster* and *D. virilis* a fraction $d = (3/4)(1-\exp(-4k/3)) \approx 0.524$ of mismatches in a pairwise (gap free) alignment. Consequently, one expects a fraction of matches of $1 - d = 0.476$. In the following we call the fraction of matches *identity*. When considering more than two species we calculate the expected number of mismatches in the multiple alignment using the pairwise mismatches. We start with the two most closely related species and proceed by considering the substitution rate between the internal node joining the for-

mer two species and the next closely related species (represented by an external node in the phylogenetic tree). For each such pair we calculate the pairwise identity. The identity of the complete multiple alignment (the fraction of columns where all nucleotides are identical) is then approximated by the product of the pairwise identities. For instance, from the data shown in Figure 3.5, we calculate for the five species *D. melanogaster*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura* and *D. virilis* the following values. $d_1 = d_{\mathrm{mel-yak}} = 0.199$, $d_2 = d_{\mathrm{A-ana}} = 0.369$, $d_3 = d_{\mathrm{B-pse}} = 0.299$ and $d_4 = d_{\mathrm{C-vir}} = 0.421$. Therefore, the expected identity in the multiple alignment is

$$\mathrm{id} = \prod_{i=1}^{4}(1 - d_i) \approx 0.205\,.$$

Assuming independence of sites, the probability that an x-mer of *D. melanogaster* is perfectly conserved in the other four species is

$$\pi_x^{\mathrm{cons}} = \mathrm{id}^x\,.$$

For instance, the probability of a completely conserved 10-mer is $\pi_{10}^{cons} \approx 1.3 \cdot 10^{-7}$. In a sequence of length l, and considering forward and backward strand of the DNA simultaneously, the probability to find at least one completely conserved x-mer is

$$p_l = 1 - (1 - \pi_x^{\mathrm{cons}})^{2l}\,. \tag{3.1}$$

For $l = 1000$bp and $x = 10$, we obtain

$$p_{1000} = 1 - (1 - \pi_{10}^{\mathrm{cons}})^{2000} \approx 0.00026\,.$$

In the above calculation we again assume independence of sites. Therefore, the calculation is only approximately valid. However, it can be shown that the error incurred is very small as far as first order moments are concerned (Haubold *et al.*, 2005).

To see how the number $p_{1000} \approx 0.00026$ above compares with perfect matches in random sequences we now calculate the background frequency of matches of length x in a second sequence with a given GC-content $P(\mathtt{GC})$. Considering a specific x-mer which contains at k $(0 \le k \le x)$ positions either of the two nucleotides G or C the probability, π_x^{rand}, to find a perfect match of this x-mer is

$$\pi_x^{\mathrm{rand}} = (\frac{1}{2} * P(\mathtt{GC}))^k * (\frac{1 - P(\mathtt{GC})}{2})^{x-k}\,.$$

Here it is assumed that G and C nucleotides and A and T nucleotides each occur with the same frequency. This is justified by the fact that both strands, forward and backward, are taken into account and that DNA is complementary. For instance, for a GC-content of $P(\mathtt{GC}) = 0.424$ (the average GC-content in the *D. melanogaster* genome) we find that π_x^{rand} ranges from $1.83 \cdot 10^{-7}$ (for $k = 10$) to $3.93 \cdot 10^{-6}$ (for $k = 0$). The lower value is close to $\pi_{10}^{\mathrm{cons}} \approx 1.3 \cdot 10^{-7}$ calculated above.

Thus, the probability of perfectly conserved 10-mers in the multiple alignment of the five species considered above is smaller than or similar to the probability of perfect matches of 10bp in two random (i.e. evolutionarily un-related) sequences. Therefore, it is conservative for this set of species to calculate the significance of a perfectly matching x-mer from a pairwise alignment of two random sequences and we use π_x^{rand} instead of π_x^{cons} in Equation (3.1) to calculate the significance level in step 4 of Section 3.3. This approximation becomes incorrect if the evolutionary distance between the compared species becomes smaller or if fewer species are used in the multiple alignment. Figure 3.6 shows the probability for a sequence fragment of length $l = 1000$bp to contain at least one defined 10mer representing a putative PLUS depending on the GC-content of the background sequence. The figure also shows the probability for seeing at least one perfectly conserved 10-mer in a multiple pairwise alignment of the five species mentioned above (black horizontal line).

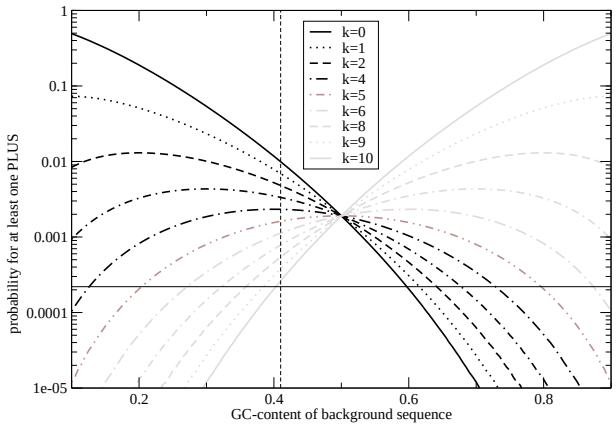


Figure 3.6: Probability to find at least one perfectly matching 10-mer in a random sequence of 1000bp (forward and backward strands) depending of the GC-content of the background sequence (x-axis). Shown are the probabilities for a 10-mer with k = 0,...,k = 10 of its nucleotides to be G or C. For a GC-content of 0.424 (the average in the Drosophila genome; vertical dashed line) the probabilities for a random match of length $x = 10$ are larger than the probability for a perfectly conserved 10-mer between the five Drosophila species analysed ($p = 0.00022$, horizontal solid line).

3.3.4 Excluding exceedingly long PLUSs

While a minimal length of an acceptable PLUS is given by its p-value shown in Eq. (3.1), an upper bound of its length is derived from an heuristic argument. Searching only those PLUSs which possibly contain a core motif of a transcription factor binding site (TFBS) and attempting to avoid perfectly conserved, but interspersed or simple ubiquitous repetitive elements we exclude those PLUSs from further consideration which exceed the length of an average TFBS (12.5bp) (Bilu & Barkai, 2005) by more than 10bp.

3.3.5 Overrepresentation of the TFBS core motif

Several authors (Papatsenko *et al.*, 2002, Abnizova *et al.*, 2005) have noted that TFBSs often contain substrings which are locally overrepresented compared to random occurrence. We measure the amount of overrepresentation of a substring of length c by comparing the observed number of occurrences within a certain region, O, to the expected number of occurrences, given the GC-content of the region, E. Again assuming independence of sites, the expected value $E = E_{c,k,w,\mathtt{GC}}$ is

$$E = (\frac{P(\mathtt{GC})}{2})^k * (\frac{1 - P(\mathtt{GC})}{2})^{c-k} * 2w\,,$$

where $P(\mathtt{GC})$ is the GC-content and w is the length of the investigated sequence. If the O/E-ratio of at least one c-mer contained in a PLUS exceeds a certain threshold t, the *PLUS* is considered to be overrepresented.

3.3.6 Clustering PLUSs into CRMs

To predict CRMs, we clustered all PLUSs with a distance smaller than 250bp into a single CRM, which may tend to merge neighbouring/overlapping CRMs. The predicted CRM starts 125bp upstream of a PLUS and ends 125bp downstream of the last PLUS in the cluster.

Table 3.1: Summary of CRM datasets used in this study

	`Papatsenko`	`HexDiff`
number of target genes	39	16
number of CRMs	95	52
References	Papatsenko *et al.* (2002) Papatsenko & Levine (2005)	Chan & Kibler (2005)

	`REDfly` (all)	`REDfly` (nonredundant)
number of target genes	203	165
number of CRMs	610	386
References	Gallo *et al.* (2005)	Gallo *et al.* (2005)

3.3.7 Measuring prediction accuracy

To evaluate our CRM predictions we use the measures used by Chan & Kibler (2005): SN and PPV on nucleotide and CRM level and SP on the nucleotide level. All measures are already explained in Chapter 1.4. Chan & Kibler (2005) calculated these measures for six programs and we took their results. To train CisPlusFinder we optimized the performance indicators SN on CRM level and PPV', a modification of PPV which is also explained in Chapter 1.4.

To assure equal conditions for the evaluation by Chan & Kibler (2005) and by us we used the script by Chan & Kibler (2005) to calculate the performance values for CisPlusFinder and other methods using sequence conservation as input. This script considers a CRM to be discovered if the overlap between it and a predicted CRM exceeds 50bp and therefore excludes CRMs shorter than 50bp from the analysis.

3.3.8 Datasets

To develop and test the algorithm we used three datasets containing CRMs from genes involved in embryonic patterning in *D. melanogaster* (Table 3.1).

To develop the `CisPlusFinder` method and train its parameters we used the combined datasets from Papatsenko *et al.* (2002) and Papatsenko & Levine (2005).

To compare results from `CisPlusFinder` with other CRM detection methods, we used the dataset and published results of Chan & Kibler (2005). The `HexDiff` dataset shares 12 genes with the Papatsenko dataset, and therefore this evaluation dataset is not independent from our training dataset, as it is the case for many of the PWM-based methods tested in Chan & Kibler (2005). We also note that the CRM annotation for the `HexDiff` and Papatsenko datasets differ for these 12 genes, with 34 of 44 annotated `HexDiff` CRMs overlapping 35 of 41 annotated Papatsenko CRMs.

To test our method on a dataset of CRMs that is fully independent from the training data and that covers a wider range of biological processes, we used the `REDfly` database curated by Gallo *et al.* (2005). To construct the `REDfly` nonredundant dataset, we chose every CRM that is verified by an *in vivo* reporter construct and regulates a target gene not found in one of the other CRM datasets used to train and evaluate `CisPlusFinder`. For every target gene, the region containing the gene and all annotated CRMs are downloaded and extended 10kb upstream of the first CRM and 10kb downstream of the last CRM.

As an additional experiment, we tested whether the PLUSs that underlie CRM predictions by `CisPlusFinder` correspond to annotated TFBSs using the `flyreg` database curated by Bergman *et al.* (2005), which are positioned in CRMs from the `HexDiff` dataset.

3.4 Results

3.4.1 Training

Choice of informant species

To train the parameters of `CisPlusFinder` we attempted not to miss CRMs in case they are absent from one or more informant sequences. Therefore, we selected manually the informant species from the set of *D. yakuba, D. ananassae, D. pseudoobscura* and *D. virilis*. Table 3.2 shows which informant species were used for which target gene.

Table 3.2: Usage of species for parameter training. '+' indicates presence and '−' indicates absence.

target gene	*D. yakuba*	*D. pseudoobscura*	*D. ananassae*	*D. virilis*
ady	-	+	+	-
brk	+	+	+	+
dpp	+	+	+	+
hbr	+	+	+	+
htl	+	+	+	+
ind	+	+	+	+
m8	+	+	+	+
mes3	+	-	-	+
mes5	+	+	+	-
phm	+	-	-	+
rho	+	+	+	+
sim	+	+	+	+
sna	+	+	+	+
sog	+	+	+	+
ths	+	+	+	+
toll	+	-	+	+
twi	+	+	+	+
vnd	+	+	+	+
vn	+	-	-	+
zen	+	+	+	+
abda	+	+	+	+
btd	+	+	+	-
dll	+	+	+	+
ems	+	+	+	+
en	+	+	+	+
eve	+	+	+	+
ftz	+	+	+	+
gsb	+	+	+	+
gt	+	+	+	+
hairy	+	+	+	+
hb	+	+	+	+
kni	+	+	+	+
kr	+	+	+	+
otd	+	+	+	+
prd	+	+	+	+
runt	+	+	+	+
sal	+	+	+	+
tll	+	+	+	+
ubx	+	+	+	+

Parameters for Drosophila

For *D. melanogaster* and the informant species, we systematically varied the following five parameters, and measured the accuracy (PPV′) on the training dataset. For other species the parameters can easily be retrained. We found the following combination to be optimal:

1. length of overrepresented core of a PLUS: $c = 5$
2. threshold for the minimal O/E-ratio: $t = 2$
3. window size for measuring overrepresentation: $w = 1000$
4. window size for measuring regulatory potential: $r = 500$
5. minimal regulatory potential: $m = 3$

Accuracy on the training set

With optimized parameters we found 94 of the known 95 CRMs in the Papatsenko dataset. Only the CRM regulating the gene *zen* is missing in our prediction, which lacks a third PLUS to reach the required regulatory potential. Sensitivity on CRM level of our method is 0.99 and PPV′ is 0.67. Specificity on nucleotide level is 0.60 and sensitivity on nucleotide level is 0.56.

3.4.2 Comparison with other methods on `HexDiff` dataset

To compare the performance of `CisPlusFinder` with other methods, we applied it to the dataset used by Chan & Kibler (2005). In this study the methods `Ahab` (Rajewsky *et al.*, 2002), `ClusterBuster` (Frith *et al.*, 2003), `MSCAN` (Johansson *et al.*, 2003), `MCAST` (Bailey & Noble, 2003), `LWF` (Papatsenko *et al.*, 2002) and `HexDiff` (Chan & Kibler, 2005) were applied to a set of CRMs active in early Drosophila development and performance measures of these methods were computed. The results are summarised in Table 3.3. Details on the parameters used for these method are described in the publication of Chan & Kibler (2005). A problem of this analysis is that the thresholds chosen by Chan & Kibler (2005) were optimized to increase specificity according to the `HexDiff` annotation. Lowering the thresholds would increase sensitivity and decrease specificity for these methods. We did not test to lower the thresholds and do not account this fact in our analysis. We added to these results the published predictions from an additional method developed by Grad *et al.* (2004), which uses only comparative data and a training set of known CRMs. Furthermore, we generated predictions for the `HexDiff` dataset using `Stubb` (Sinha *et al.*, 2004) and `ecis-analyst` (Berman *et al.*, 2004), both of which use information about conservation of known TFBSs. To evaluate these methods we used default parameter settings and input matrices from Chan & Kibler (2005).

Table 3.3 shows that `CisPlusFinder` has a higher sensitivity on the CRM and nucleotide level than any other method. The second highest sensitivity is achieved by `Stubb` which misses five CRMs found by `CisPlusFinder`. Both CRMs missed by `CisPlusFinder` have high local substitution rates, and one CRM missed by `CisPlusFinder` (regulating the expression of *oc*) was also not found by the method of Grad *et al.* (2004).

Figure 3.5 to 3.21 in the Appendix of this chapter show the predictions of all programs listed in Table 3.3, the `HexDiff` annotation and the current `REDfly`-based annotation for all `HexDiff` regions. Some CRMs are found by all methods, such as the *eve-stripe3* enhancer at position 5,487,313 in Figure 3.9, suggesting that a combined approach to CRM prediction, as is increasingly used for gene prediction (Allen & Salzberg, 2005), may yield accurate CRM annotations. Some CRMs are exclusively found by `CisPlusFinder`, such as the *prd PMFE* CRM at position 12,078,033 in Figure 3.18.

Table 3.3: Comparison of different CRM prediction methods on the `HexDiff` dataset - Accuracy of different CRM prediction methods on a set of 52 CRMs active in early Drosophila development. (Results in the top section of the table are taken from Chan & Kibler (2005)).

	SN(CRM)	PPV(CRM)	SP(nuc.)	SN(nuc.)	PPV(nuc.)	hit CRMs	missed CRMs
HexDiff	69	34	94	39	36	36	16
Ahab	44	58	98	22	55	23	29
ClusterBuster	60	26	95	34	37	31	21
MSCAN	65	19	91	27	21	34	18
MCAST	83	11	70	48	13	43	9
LWF	52	11	90	13	11	27	25
Stubb (pse-default)	67	24	92	29	24	35	17
Stubb (vir-PhastCons)	87	11	71	43	12	45	7
ecis-analyst	61	65	92	45	33	32	20
CisPlusFinder	**96**	**11**	**54**	**60**	**11**	**50**	**2**
PFR	23	36	83	10	5	12	40

3.4.3 Analysis of false positive predictions

Table 3.3 also shows a relatively low specificity achieved by `CisPlusFinder`. However it is obvious from Figures 3.5 to 3.21 that there are predictions made by `CisPlusFinder` that do not overlap with the `HexDiff`-based annotation (black boxes) but do correspond to annotated CRMs in the `REDfly` database (gray boxes). To see if the low specificity of `CisPlusFinder` results from an incompletely annotated evaluation dataset, we extracted all CRMs from the `REDfly` database that regulate genes in the the `HexDiff` dataset. The `REDfly` database contains 125 CRMs for the `HexDiff` regions in contrast to 52 CRMs currently annotated, and also has improved annotations for *hairy* and *oc*. 120 of the 125 `REDfly` CRMs overlap with our predictions, and 5 of the annotated CRMs cannot be found. The number of false positive `CisPlusFinder` CRM predictions is reduced from 408 to 338. The number of nucleotides in false positive `CisPlusFinder` CRM predictions is reduced by 15.41% from 296,005 to 250,388.

3.4.4 Correlation between PLUSs and known TFBS

To test whether PLUSs are caused by TFBSs, we downloaded all annotated TFBSs that are found in the `HexDiff` regions from the `flyreg` database (Bergman *et al.*, 2005). This resulted in a dataset of 376 TFBSs regulating 10 target genes bound by 27 different transcription factors. For 67 of the TFBS, the binding factors are not known. The average length of the TFBS in our subset is 18bp.

135 TFBS (36%) overlap with 115 PLUSs found by `CisPlusFinder`. The average length of a hit TFBS is 19bp and the average length of a hitting PLUS is 14bp. In the 602,508bp analysed sequence 4,243 PLUSs are found, covering 8.65% of the sequence. The 376 TFBSs cover 1.07% of the sequence. The hit TFBSs cover 0.39% of the sequence. The probability to have this overlap by chance is $3.12 * 10^{-4}$. These results indicate that while many PLUSs correspond to TFBSs, many TFBSs do not appear in PLUSs because of the degree of substitution and turnover in this species set (Emberly *et al.*, 2003, Ludwig *et al.*, 2000).

3.4.5 Application to simulated data

To investigate our predictions in unconstrained sequences we applied it to simulated data. We simulated the evolution of the 16 `HexDiff` sequences as ancestral sequences along the tree shown in Figure 3.5 using the method `CisEvolver` (Pollard *et al.*, 2006), which can evolve sequences according to the HKY85-model (Hasegawa *et al.*, 1985) without any selective constraint. We extracted the simulated sequences of *D. melanogaster* and the informant sequences *D. yakuba*, *D. ananassae*, *D. pseudoobscura* and *D. virilis* from the generated sequences and applied `CisPlusFinder` to it. `CisPlusFinder` did not predict any CRM in the simulated *D. melanogaster* sequences. Out of 642,508bp of the simulated *D. melanogaster* sequence roughly 60% are covered by chained alignments. No significant PLUS can be found in these aligned sequences. We conclude that PLUSs occur preferentially in sequences which are subject to functional constraint.

3.4.6 Application to the `REDfly` test set

Accuracy on the `REDfly` test set

The results of `CisPlusFinder` applied to the `REDfly` CRMs are shown in Table 3.4. Our method predicts CRMs that overlap 289 of 386 annotated CRMs in the `REDfly` dataset. We compared our results with the output of `Stubb` using the same set of PWMs as above, which may limit `Stubb` to predict CRMs regulated by transcription factors binding the CRMs in the Chan & Kibler (2005) regions. When *D. virilis* is used as a reference species `Stubb` predicts CRMs that overlap 236 annotated `REDfly` CRMs.

Table 3.4: Accuracy of `CisPlusFinder` and `Stubb` on the nonredundant `REDfly` dataset

	SN(CRM)	PPV(CRM)	SP(nuc.)	SN(nuc.)	PPV(nuc.)	hit CRMs	missed CRMs
CisPlusFinder	**75**	**13**	**64**	**47**	**19**	**289**	**97**
Stubb (pse-default)	24	19	94	7	18	92	294
Stubb (vir-PhastCons)	62	12	75	27	16	238	148

Further analysis of false negatives

97 CRMs are missed by `CisPlusFinder` in the `REDfly` regions. Assuming that the function of a CRM is conserved in all informant species, it can be missed by `CisPlusFinder` for two reasons. One reason is that the number of TFBSs in this region is lower than the minimal required regulatory potential. The other reason is that the local substitution rate of a region is very high, which means that the probability that a TFBS causes a PLUS is reduced.

To understand the basis of false negative predictions, we define missed CRMs as those which do not overlap at all with any `CisPlusFinder` prediction. Thus 9 CRMs that overlap with `CisPlusFinder` predictions for less than 50bp and were treated as false negatives above, are treated as true positives in this analysis. One missed CRM is not present in the chained alignment and excluded from further consideration. For 13 missed CRMs, the informant sequences contain ambiguous or unassembled sequences, i.e. stretches of `Ns`. The very strong requirement of perfect ungapped alignment makes `CisPlusFinder` very sensitive to unassembled sequences and `Ns` in the sequence. For this reason 6 missed CRMs with a `N`-content above 60% are excluded from further consideration. After excluding these 16 exceptions, 81 false negative CRMs are left for further analysis.

To find which CRMs are missed because they do not contain enough PLUSs to give rise to the required regulatory potential, we chose the subset of missed CRMs that contain putative PLUSs but are not predicted by `CisPlusFinder` (named *lowRP* in the following). *lowRP* contains 31 missed CRMs from 26 different regions. The remaining 50 missed CRMs contain *no PLUS* and map to 30 different regions. We speculate that these missed CRMs in *no PLUS* regions might be caused by a high local substitution rate.

To test this hypothesis, we downloaded the `PhastCons` scores for multiple alignments of the *D. melanogaster* genome with 8 other species *D. simulans, D. yakuba, D. ananassae, D. pseudoobscura, D. virilis, D. mojavensis, A. gambiae* and *A. mellifera*. These alignment scores are calculated for every position of the genome of *D. melanogaster*. The average conservation scores for the different subsets of CRMs are given in Table 3.5, which shows that many missed CRMs are positioned in regions of the genome that are poorly conserved. CRMs with *lowRP* show an intermediate level of conservation, and positively predicted CRMs show the highest average level of conservation. We suggest that TFBS in the *no PLUS* regions have undergone a higher rate of evolutionary turnover because of a high local substitution rate, and thus CRMs are not identified since the number of PLUSs is too low. In contrast to positively predicted CRMs both classes of missed CRMs have a lower conservation score than their surrounding regions.

3.5 Discussion

Table 3.3 shows that `CisPlusFinder` is able to localize Drosophila CRMs better than the other methods applied thus far, which were also trained and developed with CRMs involved in early Drosophila development. The results of `CisPlusFinder` applied to the `REDfly` dataset (Table 3.4) show that `CisPlusFinder` is not limited to a special group of CRMs,

Table 3.5: Relationship between local conservation score and `CisPlusFinder` CRM prediction accuracy - Average `PhastCons` score according to the multiple alignment of 9 insects for predicted CRMs, CRMs containing putative PLUSs with low regulatory potential (lowRP) and CRMs where no PLUSs are found. Standard errors are given in brackets.

	found	lowRP	no PLUS
number of CRMs	297	31	50
number of regions	106	26	30
average length of predicted CRM	2444	1194	533
average conservation score of the CRM	0.53 (0.14)	0.40 (0.16)	0.25 (0.19)
average conservation score of the region	0.48 (0.063)	0.46 (0.064)	0.38 (0.10)

and that our method can predict CRMs that are involved in biological processes different from those in the training data. This generality of `CisPlusFinder` results from the fact that it does not require any *a priori* knowledge about the binding transcription factors regulating a target gene and allows a whole genome scan using `CisPlusFinder`. The whole genome scan of *D. melanogaster* was performed for this PhD-thesis. The results are shown in Chapter 4.

However, we found that the specificity of our method is the lowest compared to those that condition predictions on TFBS models. To test if `CisPlusFinder` predicts a high number of false positives or, more interestingly, a high number of undiscovered CRMs, we compared our prediction with an updated annotation of the `HexDiff` regions (section 3.4.3). These results clearly indicate that the low specificity of `CisPlusFinder` is in part due to unannotated CRMs in the `HexDiff` regions, underlying both the importance of high quality CRM annotations as well as the difficulty in evaluating the specificity of CRM prediction methods.

The results reported in Table 3.5 also show that the performance of `CisPlusFinder` depends on the local substitution rate of the region investigated. In regions with high substitution rates, CRMs cannot be identified with the same accuracy as those in regions with low substitution rates. There may also be cases where the substitution rates are extremely low such that the probability to find a PLUS by chance is very high. One example for this case is the surrounding sequence of the gene *ems* which is shown in Figure 3.8. In these cases more species should be included in the analysis.

A final result of our study is that PLUSs are valuable signals for predicting CRMs. This finding is not incompatible with the occurrence of substitutions in TFBS and binding site turnover (Emberly *et al.*, 2003, Ludwig *et al.*, 2000). PLUSs likely correspond to and are caused by TFBSs, but not every TFBS causes a PLUS. Hence `CisPlusFinder` can find CRMs that may have undergone functional substitutions, when at least some of the TFBSs are conserved strongly enough to give rise to PLUSs.

Appendix to Chapter 3
Visualization of HexDiff regions

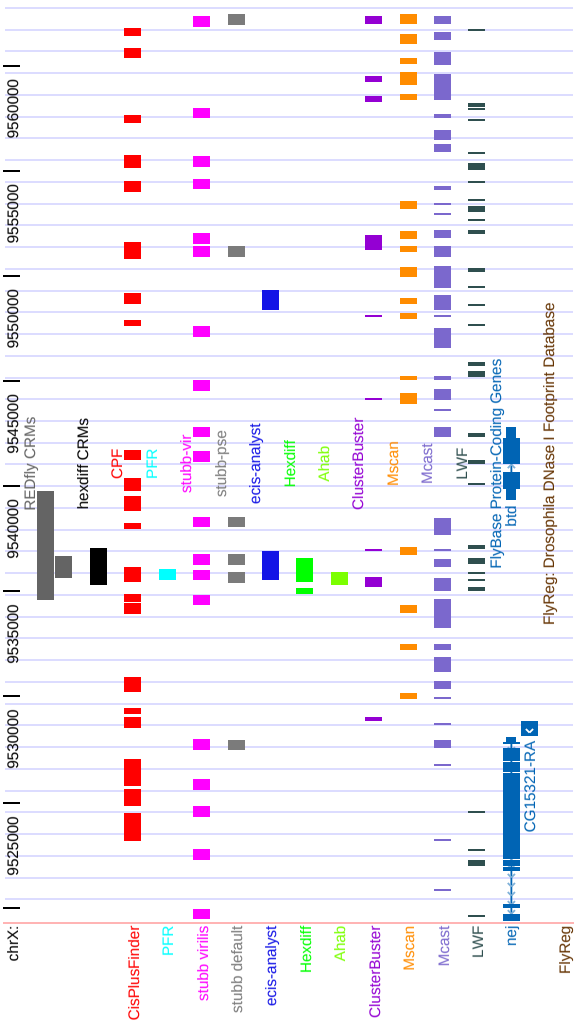
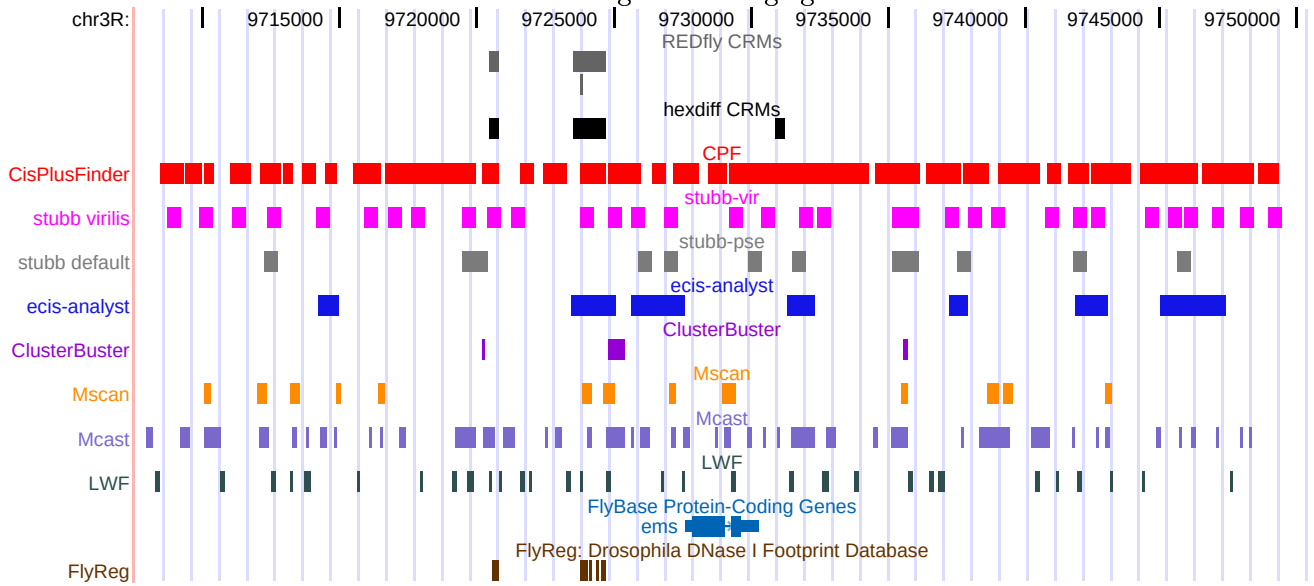


Figure 3.7: targetgene: **btd**

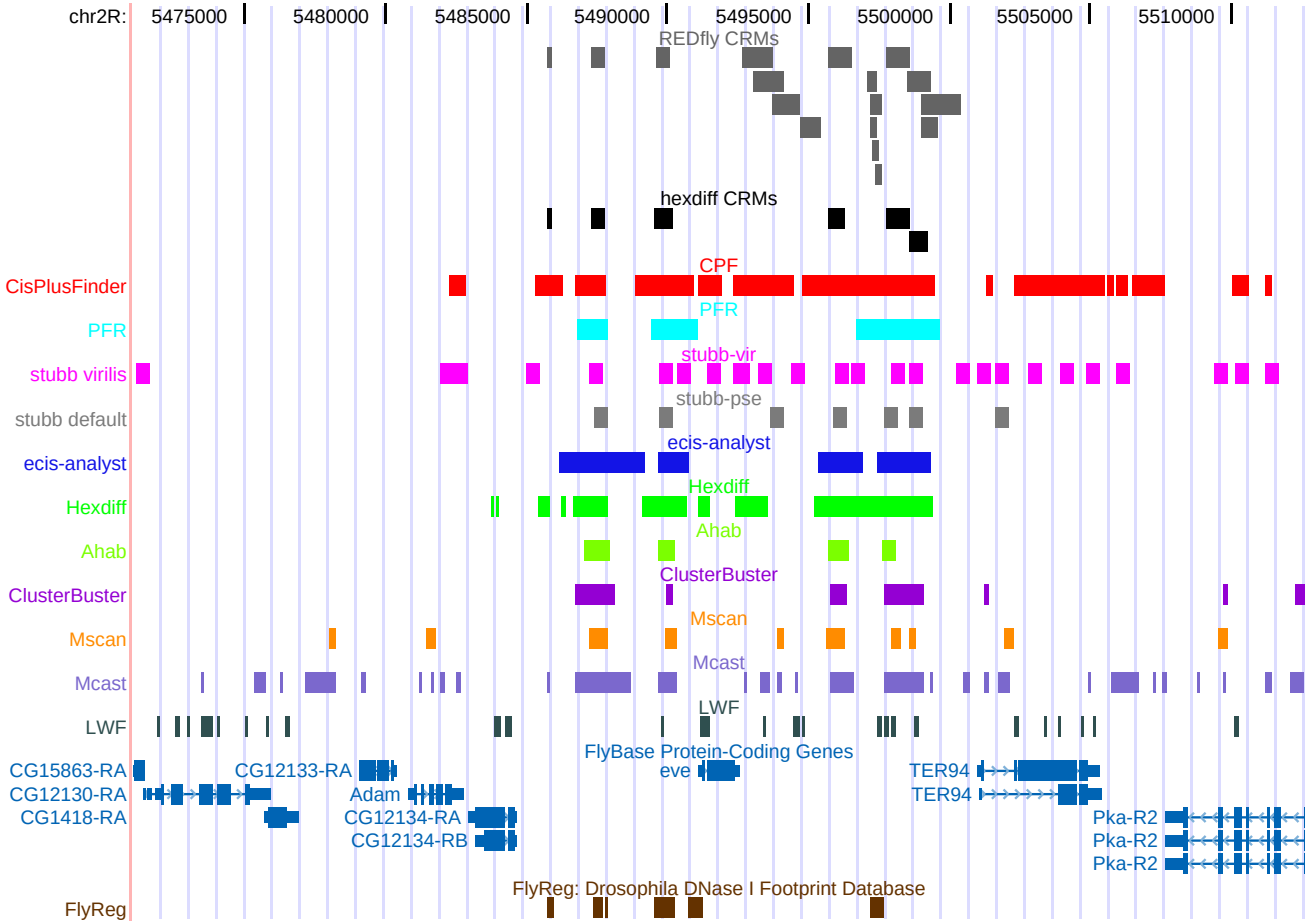
Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *btd* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Figure 3.8: targetgene: **ems**



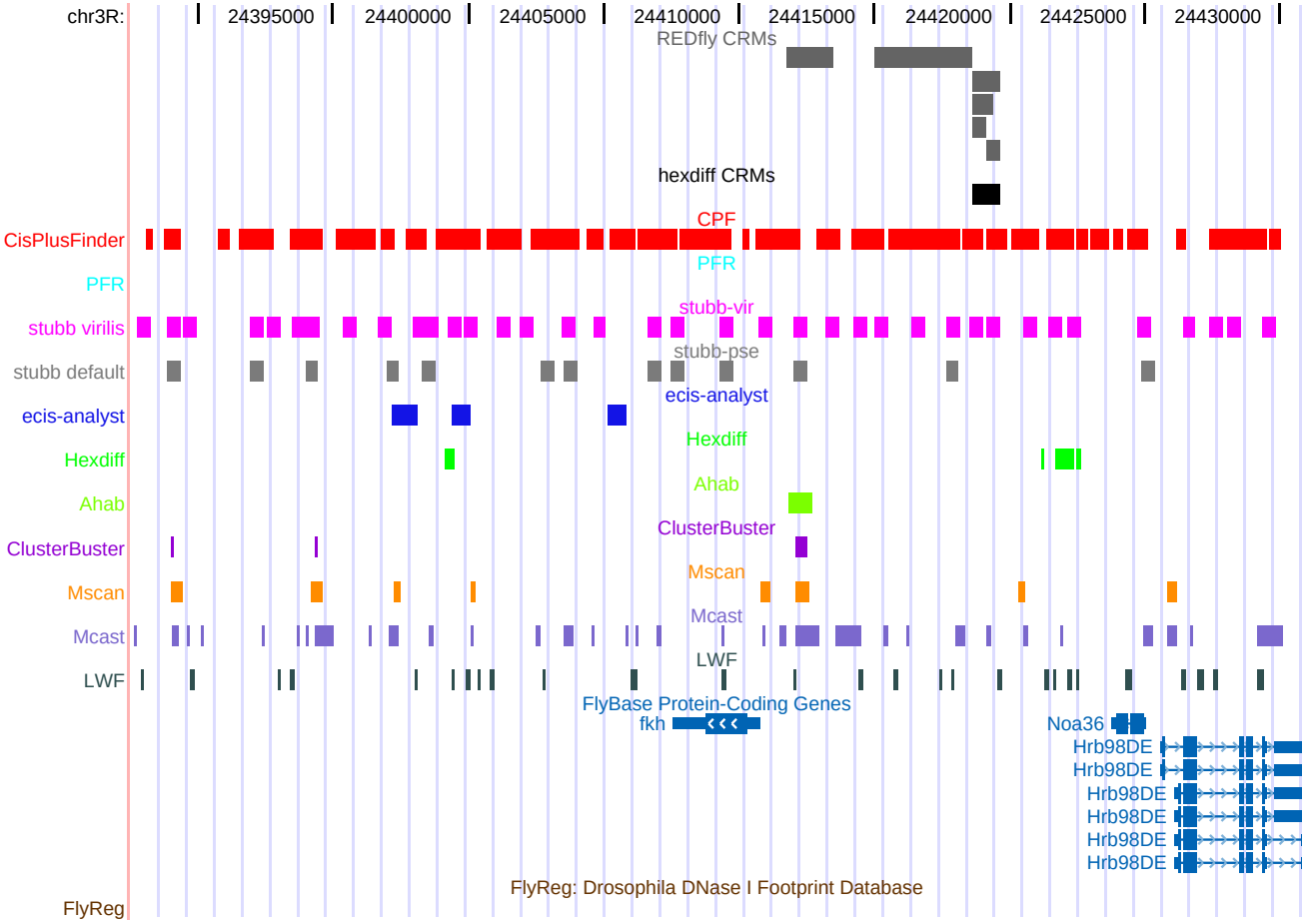
Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *ems* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Figure 3.9: targetgene: **eve**



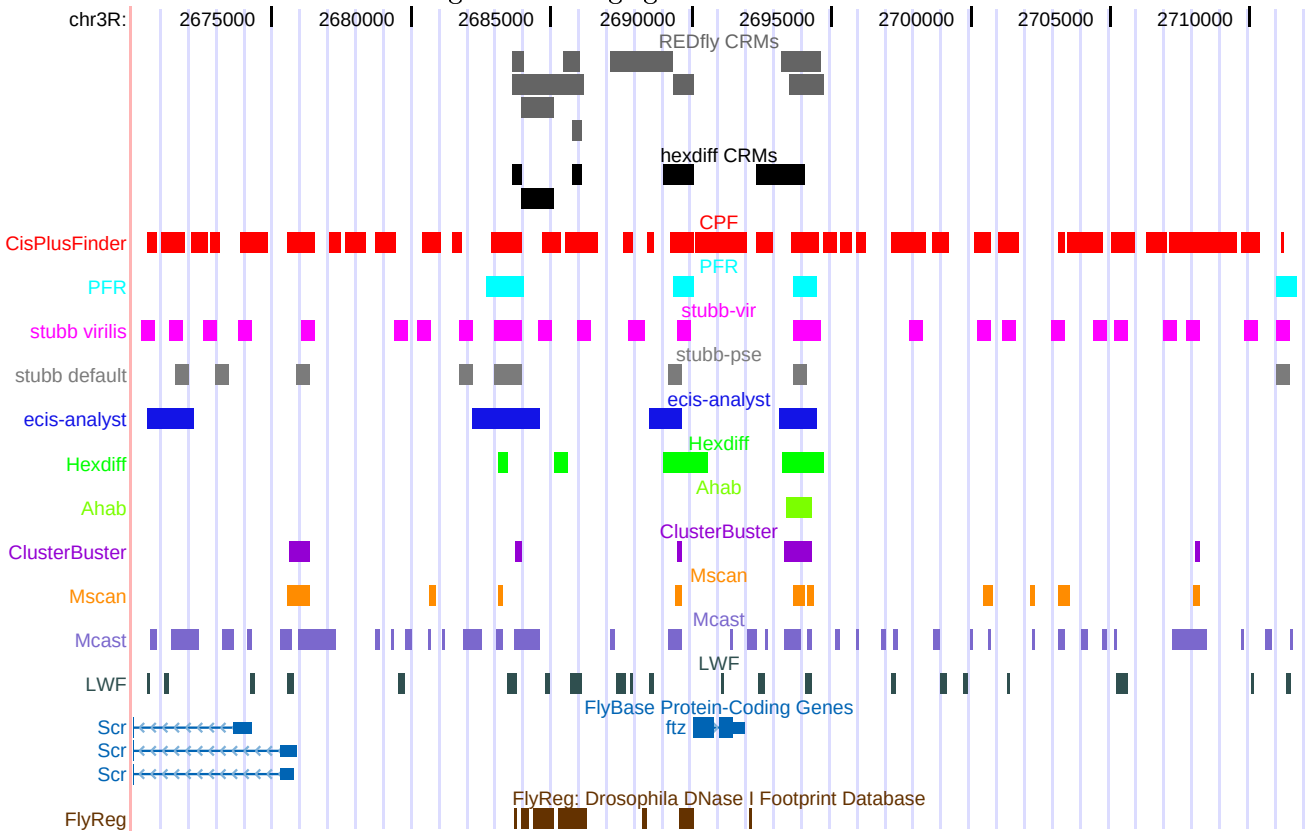
Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *eve* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Figure 3.10: targetgene: **fkh**



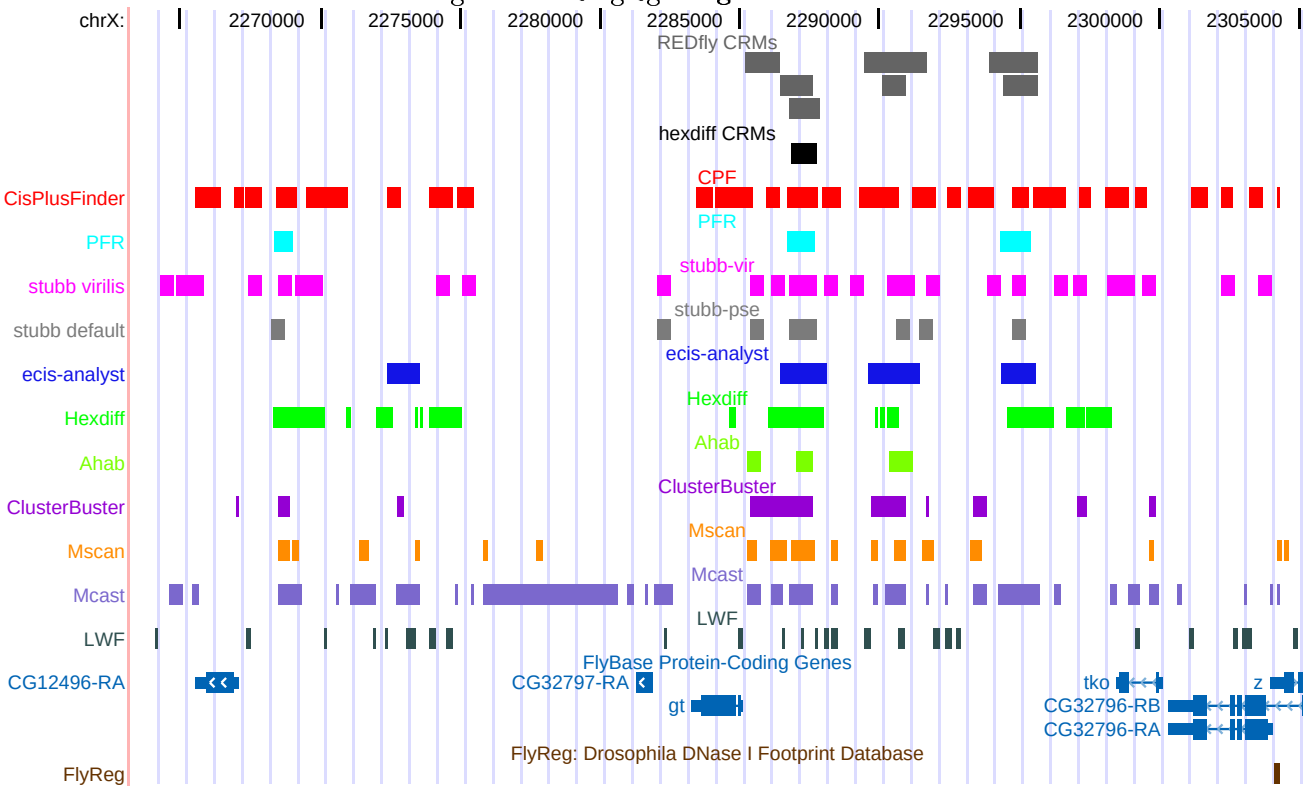
Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *fkh* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Figure 3.11: targetgene: **ftz**



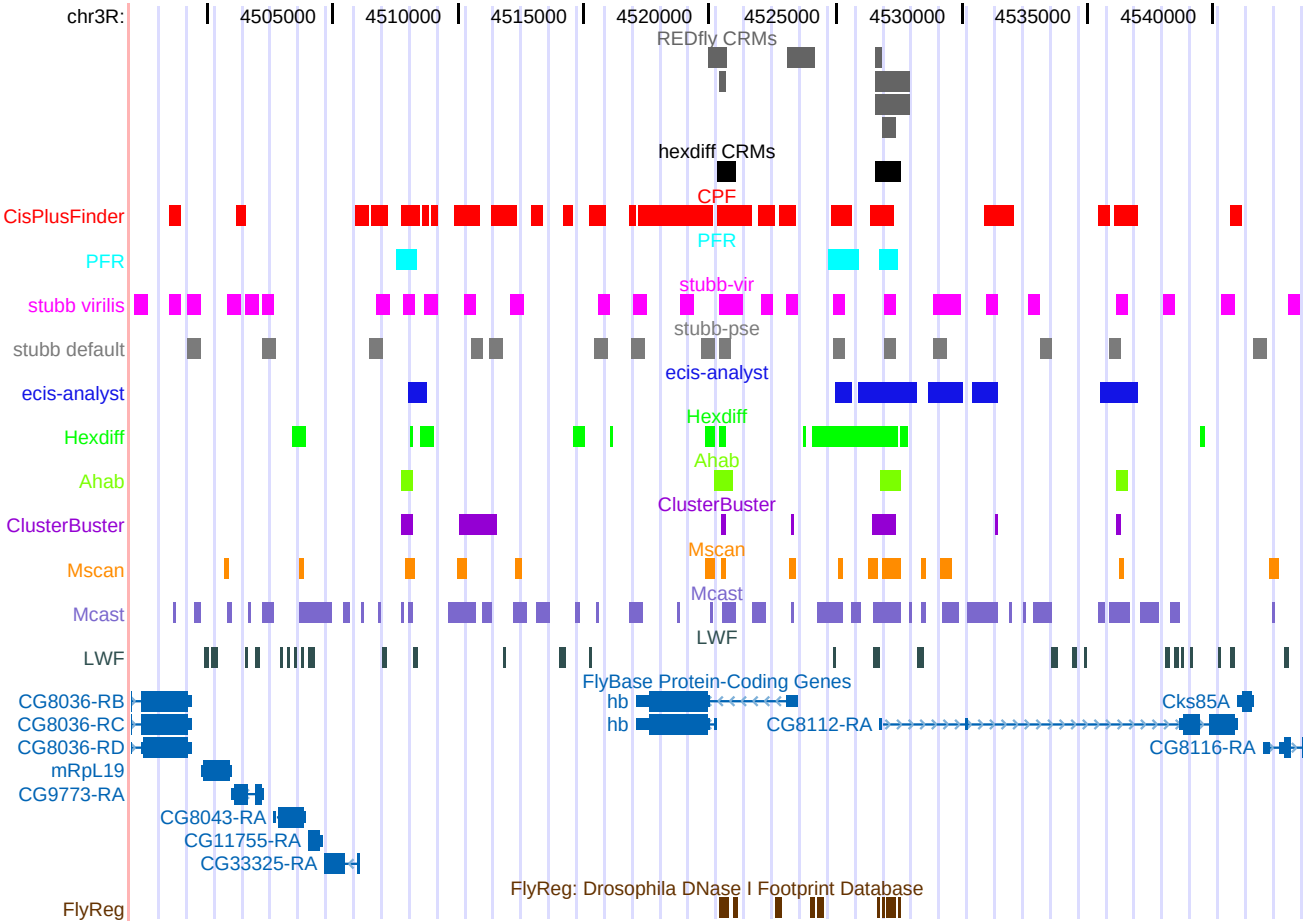
Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *ftz* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Figure 3.12: targetgene: **gt**



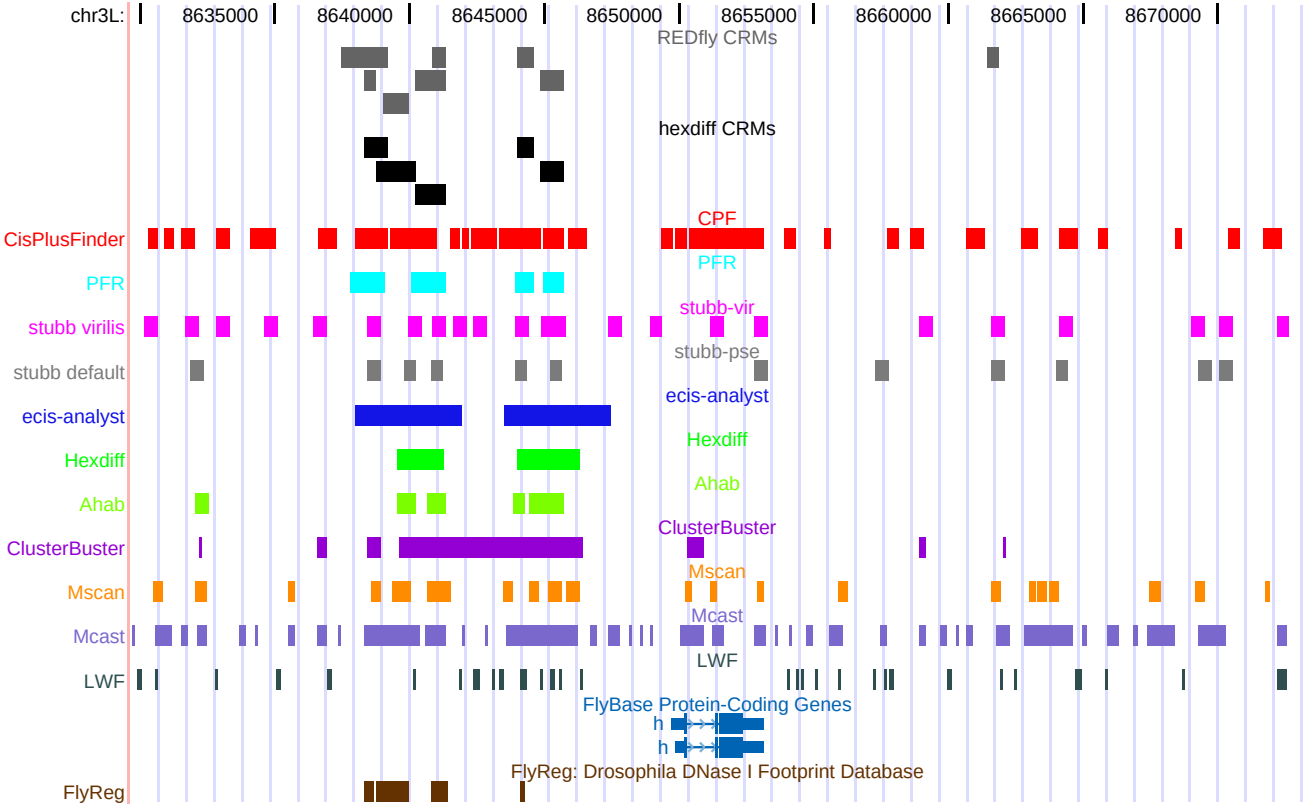
Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *gt* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Figure 3.13: targetgene: **hb**



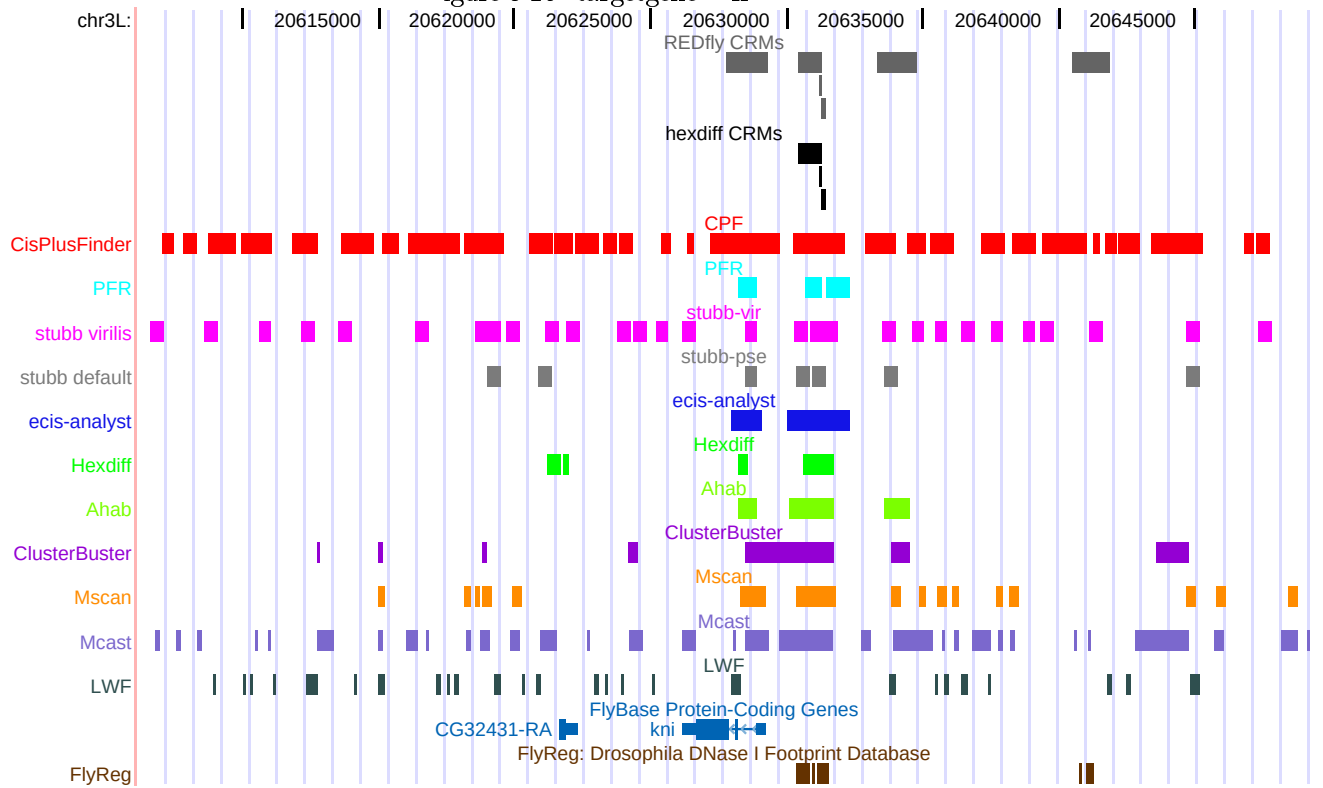
Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *hb* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Figure 3.14: targetgene: **hairy**



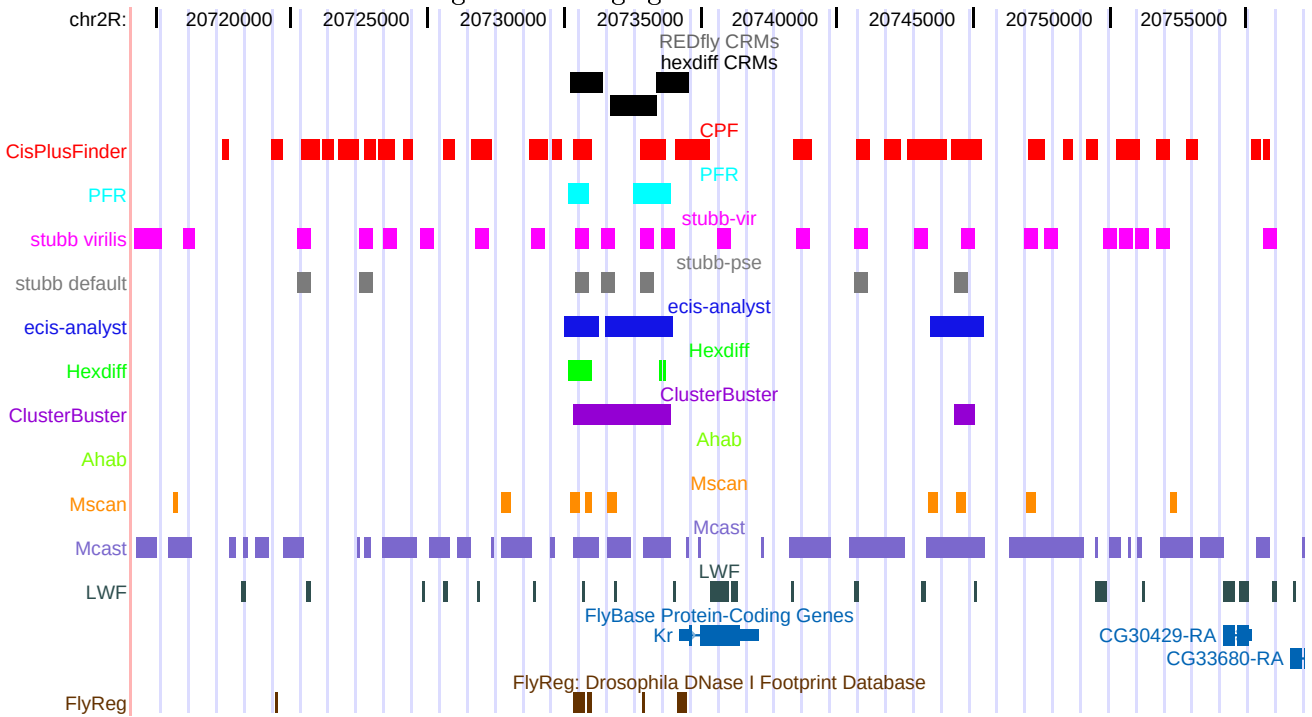
Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *hairy* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Figure 3.15: targetgene: **kni**



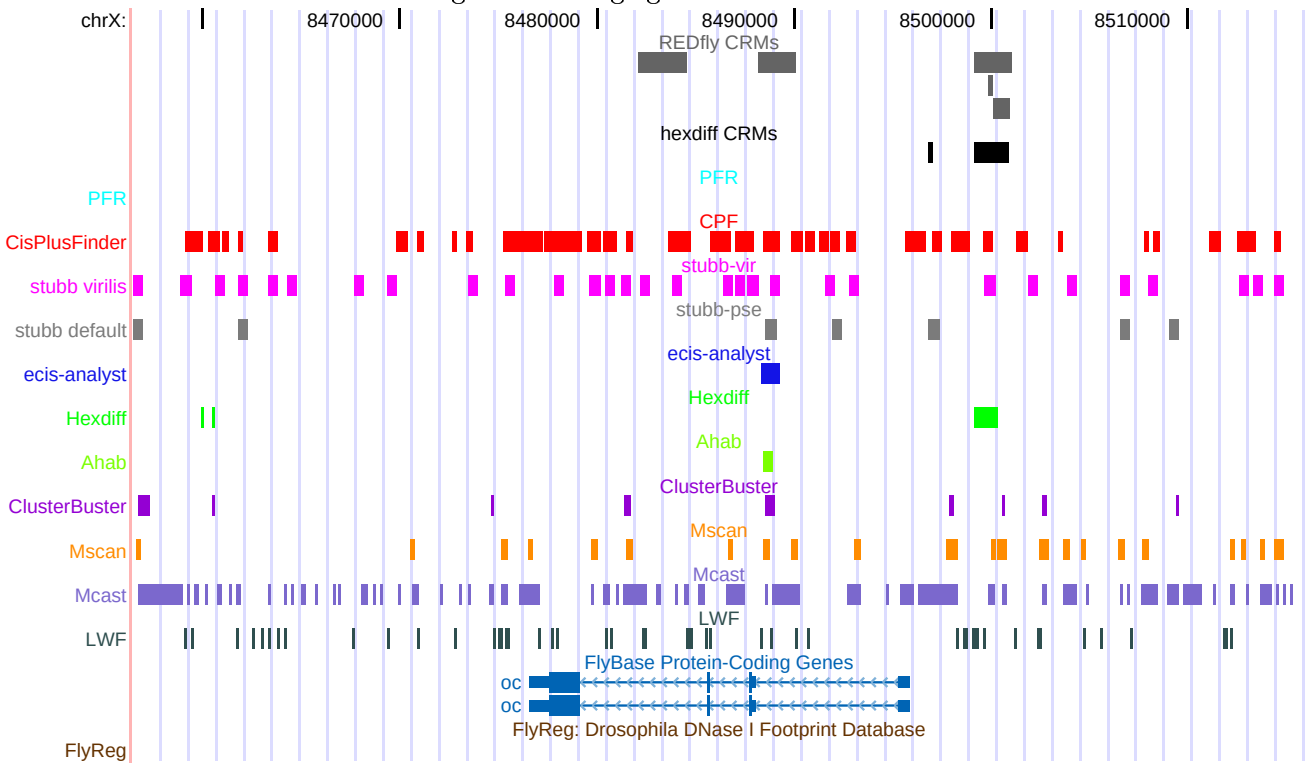
Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *kni* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Figure 3.16: targetgene: **kr**



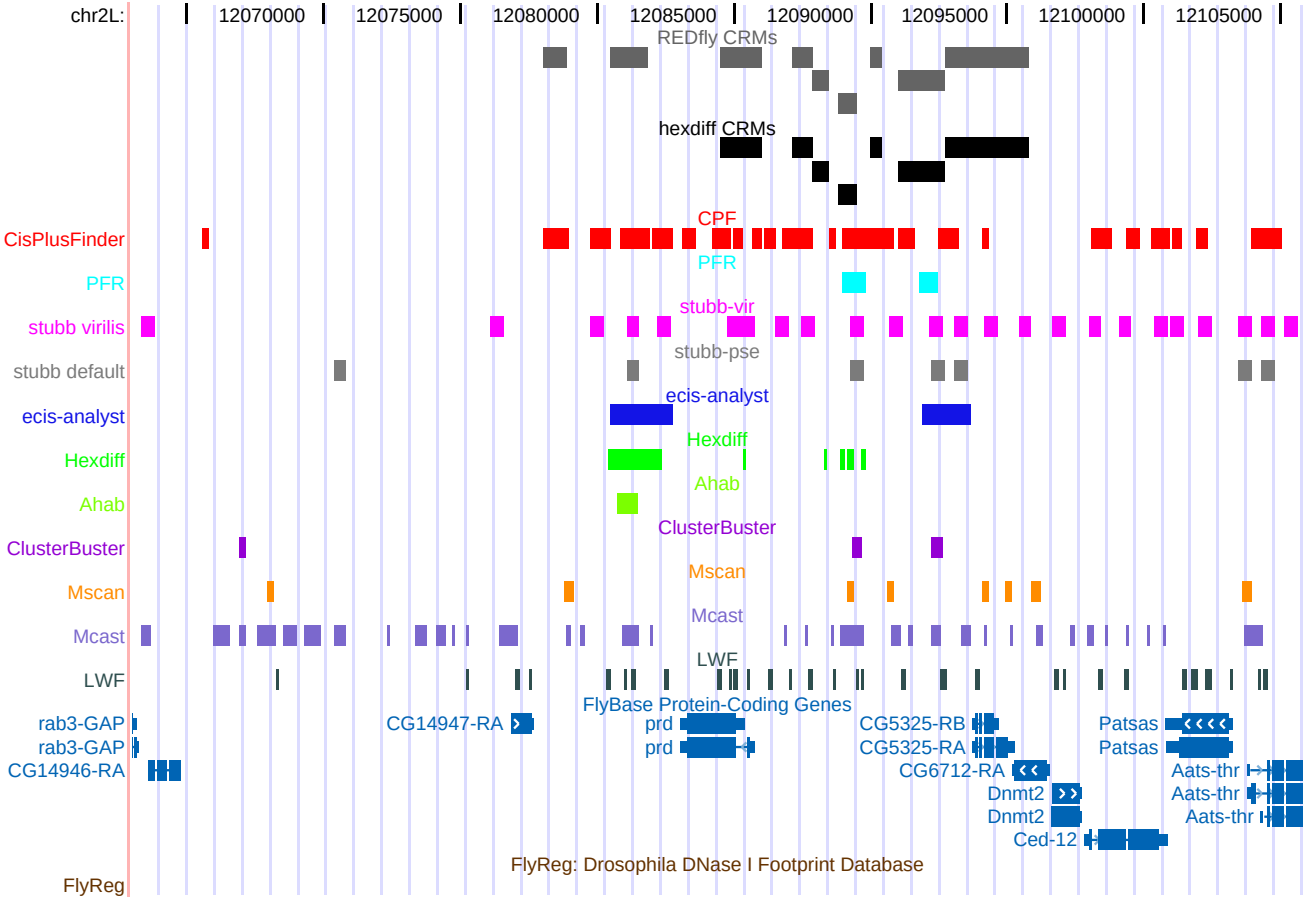
Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *kr* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Figure 3.17: targetgene: **oc**



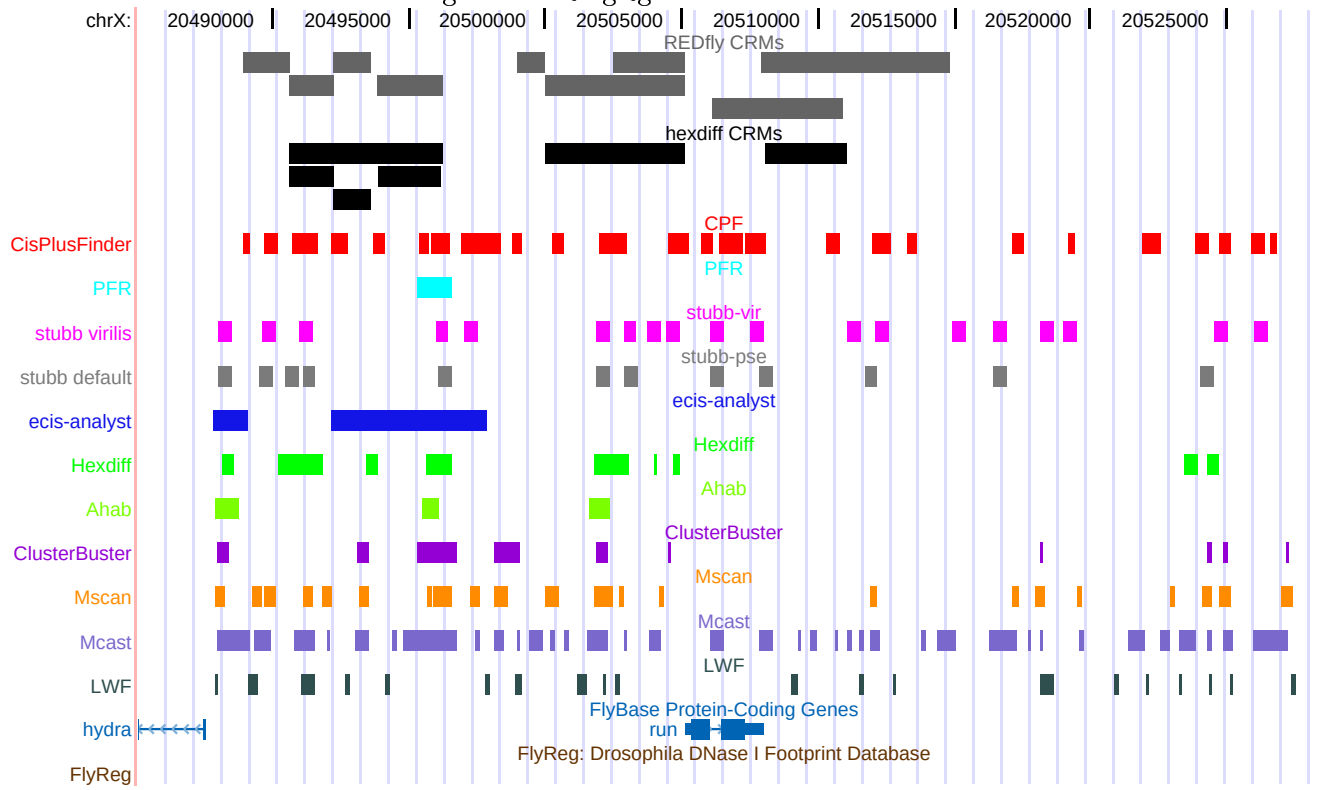
Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *oc* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Figure 3.18: targetgene: **prd**



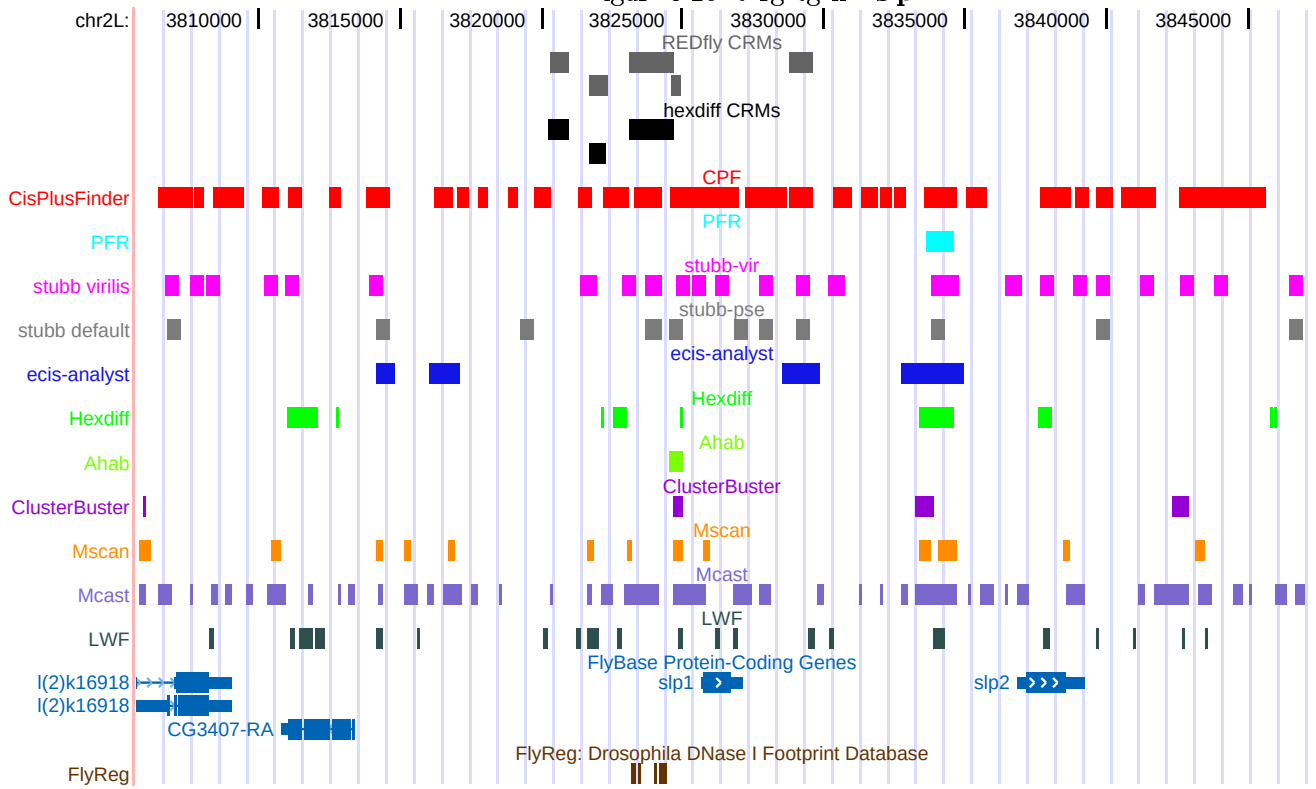
Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *prd* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Figure 3.19: targetgene: **runt**



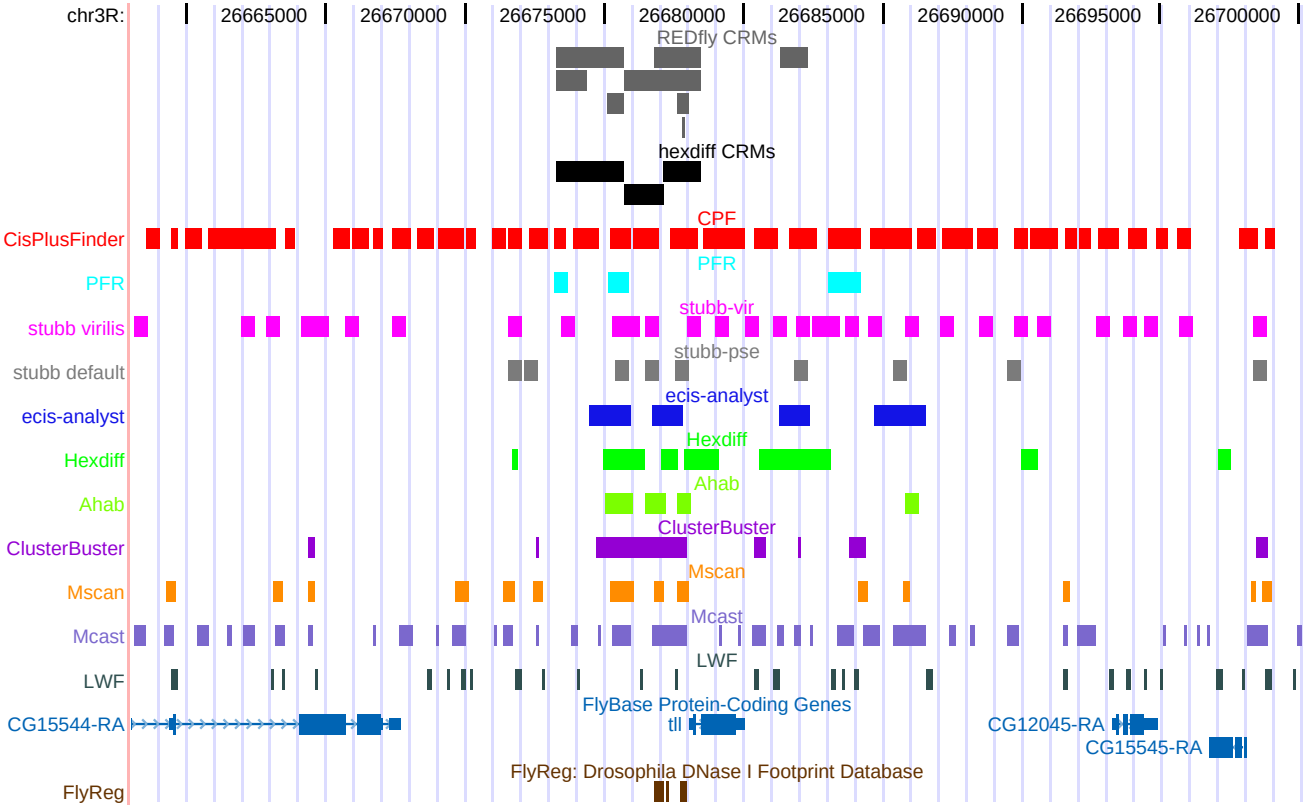
Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *runt* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Figure 3.20: targetgene: **slp1**



Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *slp1* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Figure 3.21: targetgene: **tll**



Visualization of the predictions of the methods listed in Table 3.3. The top two tracks show the `REDfly` (2006) and `HexDiff` (2005) annotations for the genomic regions surrounding the target gene *tll* using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

Chapter 4

Whole genome scan of *Drosophila melanogaster* using `CisPlusFinder`

4.1 Introduction

This chapter describes the results of a whole genome scan *(WGS)* in *D. melanogaster*. A CRM prediction along the complete genome of *D. melanogaster* can fulfill two functions.

1. The results can provide further support to the program `CisPlusFinder` and prove our concept.
2. If the predictions are shown to be reasonable further hypotheses concerning gene regulation can be formulated.

Another question which we would like to answer is if the need and the presence of CRMs in a genome influence the architecture and structure of the *D. melanogaster* genome, which is very compact.

4.2 Methods

To perform a WGS of the *Drosophila melanogaster* genome we used the pairwise whole genome alignments accessible at VISTA (Couronne *et al.*, 2003). These alignments were computed using the software `Lagan` (Brudno *et al.*, 2003). We downloaded the pairwise alignments of *D. melanogaster* with *D. yakuba, D. ananassae, D. pseudoobscura* and *D. virilis* respectively. The pairwise whole genome alignments match the *D. melanogaster* chromosomes with fragments of the other species. To get four sequences, which represent the homologues to a complete chromosome we concatenated the fragments into one sequence and applied `CisPlusFinder` to *D. melanogaster* chromosomal sequences downloaded from NCBI (Benson *et al.*, 2006) using the concatenated homologue sequences parsed out of the whole genome alignments as informant species. The procedure and the parameters were exactly the same as described in Chapter 3.

We applied `CisPlusFinder` to each chromosome separately and analysed the results separately to get an idea about the consistency of the results.

4.3 Results

4.3.1 Distribution of predicted CRMs along the genome

A biologist expects a meaningful whole genome CRM prediction to result in clusters of CRMs around genes and long stretches of DNA performing a different function or junk DNA. The degree of clustering of the predictions is observable when we look at the lengths of sequences

Table 4.1: Lengths of fragments between predicted CRMs

chromosome	Ø(length between CRMs)	max(length between CRMs)	min(length between CRMs
2L	1690.85bp	156,306bp	
2L (exons excluded)	1961.29bp	156,306bp	
2R	1811.3bp	265,427bp	
2R (exons excluded)	2116.11bp	265,427bp	
3L	1923.15bp	1,136,665bp	1bp
3L (exons excluded)	2178.87bp	1,136,665bp	
3R	1726.31bp	983,499bp	
3R (exons excluded)	1914.38bp	983,499bp	
X	3484.71bp	1,131,842bp	
X (exons excluded)	4166.02bp	1,131,842bp	

between predicted CRMs. Averages, maxima and minima of the length distribution of interCRM-distances are given in Table 4.1.

Lunter *et al.* (2006) compare the distribution of indels along the genome to the geometrical distribution and show that this distribution matches a random distribution of indels along the genome very well. We assume a random annotation of CRMs to result in the same distribution as a random annotation of indels. Therefore we expect a geometrical distribution of the interCRM length in random case. To test this hypothesis we simulated the random distribution by replacing every predicted region by a single X and shuffling the sequence of `As`, `Cs`, `Gs`, `Ts`, `Ns` and `Xs`. Then the distances between Xs in the shuffled sequence were calculated. To compare the random placement of CRMs with a geometrical distribution, we performed the Kolgomorov-Smirnov-Test. The results are shown in the last two columns of Table 4.2 and they indicate very clearly that the random distribution matches the geometrical distribution very well.

The first two columns of Table 4.2 show the Kolgomorov-Smirnov-statistic for the overlap between the distribution of the predicted CRMs and a geometrical distribution. We can see that these two distributions deviate significantly from each other as we expected in case of a meaningful clustered prediction.

To see, if the predicted CRMs are clustered around genes, we computed the correlation between prediction density and gene density. For this purpose we slided a window of 1Mb in 500kb steps along the genome and calculated gene and CRM coverage for these windows. The values are plotted in Figure 4.1. The Pearson-Product-Moment correlation coefficient between gene density and CRM prediction density is $\approx +0.3$. We calculated the significance of this correlation using a t-test and found that the non-directional probability to find this correlation coefficient in not correlated data is 0.000006

4.3.2 Genome coverage by CRMs

After we demonstrated that the results of the WGS make sense we estimated the part of the genome which is responsible for gene regulation. We calculated the coverage of the genome by predicted CRMs and by found PLUSs. The results are shown in Table 4.3. Information about the lengths of CRMs and PLUSs are listed in Table 4.4.

Table 4.3 gives an average coverage of the genome by predicted CRMs of 25%. In Chapter 3 values for the specificity and the PPV of `CisPlusFinder` are given. If they were trustworthy, we could calculate the number of false positive predictions and give a corrected estimation of our prediction. But since we do not know the true annotation of most of the target regions, we can give a lower boundary and an upper boundary of true CRMs. The

Table 4.2: Significance of clustering of CRMs

chromosome	KS-statistic (real data)	probability (real data)	KS-statistic (randomized)	probability (randomized)
2L	0.165672	0.000000	0.004699	0.989919
2L (exons excluded)	0.146582	0.000000	0.010380	0.358331
2R	0.162417	0.000000	0.008094	0.676163
2R (exons excluded)	0.143028	0.000000	0.009409	0.554876
3L	0.154508	0.000000	0.009584	0.362567
3L (exons excluded)	0.172658	0.000000	0.008253	0.611463
3R	0.183645	0.000000	0.004824	0.948616
3R (exons excluded)	0.147298	0.000000	0.005455	0.902842
X	0.188192	0.000000	0.010567	0.569474
X (exons excluded)	0.166219	0.000000	0.012999	0.399747

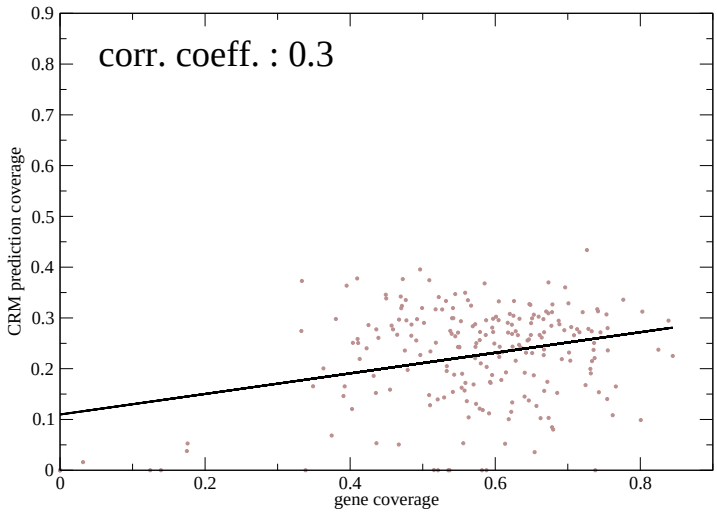


Figure 4.1: Correlation between gene density and CRM density

specificity on nucleotide level on the HexDiff regions using the new annotation is 25%. If we conclude that 25% of our predictions are true only 6% of the genome are predicted to perform regulatory function. Since we do not know the true annotations of the analysed regions, our lower boundary of the regulatory part of the genome of *D. melanogaster* is 6%.

To calculate an upper boundary we observe the sensitivity of `CisPlusFinder`. If we assume, that we find 96% of all true CRMs the upper boundary for the regulatory part of the genome of *D. melanogaster* is 26%.

The amount of coding region within the Drosophila genome is approximately 20%. It is possible, that the amount of regulatory sequence is of the same size as the amount of protein coding sequence. Figure 4.1 shows that the CRM prediction density rarely exceeds 0.4. The majority of gene coverage values is larger than 0.4. The gene density grows approximately three times faster than the more stable prediction density. Even if the gene coverage corresponds to the CRM coverage the conclusion that the amount of regulatory DNA for one gene corresponds to its length is probably not true.

4.3.3 Chromosome coverage by CRMs

To observe the consistency of our prediction we analysed every chromosome separately. Table 4.3 shows the coverage by predicted CRMs respectively for every chromosome. The results on autosomes are similar to each other. This fact provides further support to our result. But Table 4.3 also indicates very clearly, that the amount of predicted CRMs on the X-chromosome is smaller than the amount on autosomes. The amount of the X-chromosome covered by predicted CRMs is only 54% of the average coverage of the autosomes. In Table 4.5 we calculate different properties of the chromosomes to explain the difference.

One possible explanation for the lower coverage of the X-chromosomes by predictions could be a higher substitution rate of the X-chromosome compared to autosomes. Reinhold (1998) was the first to provide evidence that the X-chromosome has a higher evolutionary turnover than the autosomes. This idea has been intensely discussed since then. There are two possible explanations for the increase of the substitution rate in the X-chromosome.

One possible explanation is the lower population size of the X-chromosome compared to autosomes because of its haploidy in males. The population size of the X-chromosome is only 3/4 of the population size of autosomes. A smaller population size decreases the efficiency of purifying selection and therefore leads to an increased substitution rate.

The other theory explaining the higher substitution rate of the X-chromosome is called the fast-X evolution (Counterman *et al.*, 2003). The fast-X evolution assumes that the rate of positively selected mutations is higher than the rate of neutral substitutions in the genome. Because X-chromosomes are haploid in males, advantageous substitutions on the X-chromosome are carried in the phenotype directly and are fixed more rapidly than on autosomes.

To test, if the lower coverage of the X-chromosome by CRMs is an effect of a higher substitution rate, we used the conservation scores calculated using the program PhastCons (Siepel *et al.*, 2005) downloadable at UCSC (Kent *et al.*, 2002). We averaged the PhastCons-scores over every chromosome. The results are given in the second column of Table 4.5. The conservation score of the X-chromosome is in fact 20% smaller than the average autosomal conservation score. This difference is strong but not sufficient to explain the 46% difference in coverage of the different chromosomes.

Another explanation for a lower CRM density could be a lower gene density. We calculated the gene densities of the chromosomes and found a lower gene density in the X-chromosome. The gene density, the gene coverage and the average gene length are given in Table 4.5 for every chromosome separately. The gene density in the X-chromosome amounts to 80% of the average gene density in autosomes but the gene coverage adds up to 95% of the average coverage in autosomes. The average gene length is 15% longer in the X-chromosome than in the autosomes. To test if the lower amount of genes in the X-chromosome effects the prediction coverage we calculate the number of predicted CRMs per gene. The number of predicted CRMs regulating one gene is 77% of the average number of predictions per gene

Table 4.3: Genome coverage by predicted CRMs

chromosome	bp covered by CRMs	% of chromosom covered by CRMs	bp covered by PLUS	% of chromosom covered by PLUS
2L	5520339bp	24.64%	841765bp	3.76%
2L (exons excluded)	4845963bp	21.63%	738208bp	3.19
2R	5054727bp	26.06%	903249bp	4.66%
2R (exons excluded)	4383366bp	22.60%	668934bp	3.45%
3L	5910440bp	24.86%	938690bp	3.95%
3L (exons excluded)	5282969bp	22.22%	837093bp	3.52%
3R	7753512bp	27.79%	1227473bp	4.4%
3R (exons excluded)	7115240bp	25.5%	1119333bp	4.01%
X	3067891bp	13.98%	499301bp	2.25%
X (exons excluded)	2537541bp	11.42%	408396bp	1.84%

in autosomes. The average number of predictions in intergenic regions adds up to 74% of the average value in autosomes.

Lower gene density and the higher substitution rate are causing a decrease in prediction density on the X-chromosome. But we cannot exclude the possibility that additional factors are influencing the prediction rate.

4.3.4 Correlation between Gene ontologies and CRM density

Nelson *et al.* (2004) suggest in their publication that long intergenic regions are caused by the need of a complex regulation of one or both of the neighbouring genes. To support this hypothesis they showed results which we reproduce in Figure 4.2, where they correlated *GO (gene ontology)* terms with the length of the adjacent intergenic regions of *D. melanogaster* and *Ciona intestinalis*. If their hypothesis is true we would expect an enrichment of CRM predictions in long intergenic regions. To test this we calculated the average length of intergenic regions with predictions and compared this value to the average length of intergenic regions of the whole chromosome in Table 4.6. CRMs tend to be predicted in intergenic regions which are between 4 and 5 times longer than the average intergenic regions. This result supports the theory that long intergenic regions are essential for the Drosophila organism to regulate the genes expressed in complex expression patterns.

As a second approach to support or reject the hypothesis that intergenic regions are hosts of CRMs we downloaded all genes whose functional annotations contained the GO terms used by Nelson *et al.* (2004) from flybase (Grumbling *et al.*, 2006). The considered GO terms and their identification numbers are listed in Table 4.7. To see, if the trend shown in Figure 4.2 is consistent with the current GO annotation we calculated the mean length of the intergenic regions per gene. The intergenic regions of one gene include the complete upstream region starting at the end of the previous gene, the downstream sequence ending at the start of the following gene and all introns. The mean intergenic length for the GO terms used by Nelson *et al.* (2004) are shown by the white bars in the background of Figure 4.3. The black bars in Figure 4.3 show the mean number of bp covered by CRMs per gene for the different GO terms. For this value the length of all CRMs between the stop codon of

Table 4.4: Lengths of predicted CRMs

chromosome	number of CRMs	Ø(CRM length)	number of PLUSs	Ø(PLUSlength)
2L	8799	627.38bp	68479	12.29bp
2L (exons excluded)	7930	611.09bp	60091	12.28bp
2R	7912	638.87bp	74723	12.09bp
2R (exons excluded)	7090	618.25bp	55352	12.09bp
3L	9239	639.73bp	77029	12.19bp
3L (exons excluded)	8443	625.72bp	68252	12.26bp
3R	11653	665.37bp	100988	12.15bp
3R (exons excluded)	10842	656.27bp	91616	12.22bp
X	5494	558.41bp	42138	11.85bp
X (exons excluded)	4723	537.27bp	34266	11.92bp

the upstream gene and the transcription start site of the downstream gene are summed up and the average value of all genes of one GO term is calculated.

The general trend in our new results using current GO annotations is consistent with the results of Nelson *et al.* (2004). The intergenic length surrounding the genes performing functions which do not require complex regulation *("ribosomal constituent", "general transcription factors")* are neighbouring shorter intergenic regions than the genes performing functions requiring a very complex gene regulation *("receptors", "embryonic development", "cell differenciation", "specific transcription factors", "pattern specification")*. The results of genes involved in *"metabolism"* are not consistent. According to the current GO-annotations the length of intergenic sequences adjacent to genes performing metabolic functions is nearly as long as the mean intergenic length of genes involved in embryonic development (white bars in Figure 4.3). According to the annotation of Nelson *et al.* (2004) this length was nearly as short as the length corresponding to *"ribosomal constituent"* (Figure 4.2).

The black bars of Figure 4.3 show a higher consistency with Figure 4.2. The mean length of regulatory sequence regulating one gene differs for different GO terms. The longest regulatory sequences regulate genes performing *"pattern specification"*. For GO terms which are correlated to simple expression patterns *"ribosomal constituent", "general transcription factors"* and *"metabolism"* the mean length or regulatory sequence is smaller than 5.3kb. *"Receptor"* genes and genes involved in *"embryonic development"* are regulated by approximately 10kb. The mean length of regulatory sequences regulating genes involved in *"cell differenciation", "specific transcription factors"* and *"pattern specification"* is between 15kb and 30kb.

Another result is shown in Figure 4.4. It shows the amount of bp in the intergenic regions which are predicted to perform regulatory function. Except for the two outliers *"metabolism"* and *" pattern specification"* the amount of regulatory sequence is steady in the intergenic regions. The prediction density is not dependent on the length of the intergenic region or the GO term of the neighbouring genes.

But since we know that the absolute number of basepairs involved in regulation of genes with a more complex expression pattern is higher than the number of basepairs involved in regulation of a gene with a simple expression pattern, we have to ask ourselves, if this causes the difference in prediction density of the chromosomes. To test this we grouped the genes of different GO terms according to the chromosomes where they are positioned. The numbers are given in Table 4.8. According to the lower prediction density of chromosome X we would expect a lower amount of genes of complex expression pattern on this chromosome.

Table 4.5: General properties of the chromosomes;
This table lists some of the general properties of the chromosomes of Drosophila melanogaster. *inter* is used as an abbreviation of the word intergenic. "#(inter CRMs)/#(inter regions)" stands for the number of CRMs predicted in intergenic regions divided by the total number of intergenic regions of the chromosome.

chromosome	PP-score	gene density (gene/bp)	coverage by genes (%)	average genelength (bp)	#(coding bp) per gene	#(inter CRMs)/ #(inter regions	#(inter CRMs) / #inter with pred	#(all CRMs)/ #genes
2L	0.46	$1.23 * 10^{-4}$	58	4742	852	1.45	5.76	3.20
2R	0.46	$1.45 * 10^{-4}$	62	4302	756	0.99	4.42	2.63
3L	0.47	$1.19 * 10^{-4}$	60	4991	880	1.99	8.24	3.26
3R	0.48	$1.28 * 10^{-4}$	61	4773	814	1.61	6.83	3.27
X	0.4	$1.05 * 10^{-4}$	57	5413	916	1.11	4.99	2.37
average	0.45	$1.24 * 10^{-4}$	60	4844	843.6	1.43	6.05	2.95
$\frac{X}{avg(autosome)}$	0.86	0.82	0.95	1.15	1.11	0.74	0.79	0.77

Table 4.6: Length of intergenic regions hosting CRM predictions compared to the average intergenic length of one chromosome

chromosome	length (intergenic with prediction)	length (all intergenic)	intergenic with prediction / all intergenic
2L	11399	3414	3.34
2R	9836	2591	3.80
3L	17216	3386	5.08
3R	11625	3058	3.80
X	17865	4152	4.30
average	13594	3320	4.06
X/avg(autosome)	1.43	1.33	1.07

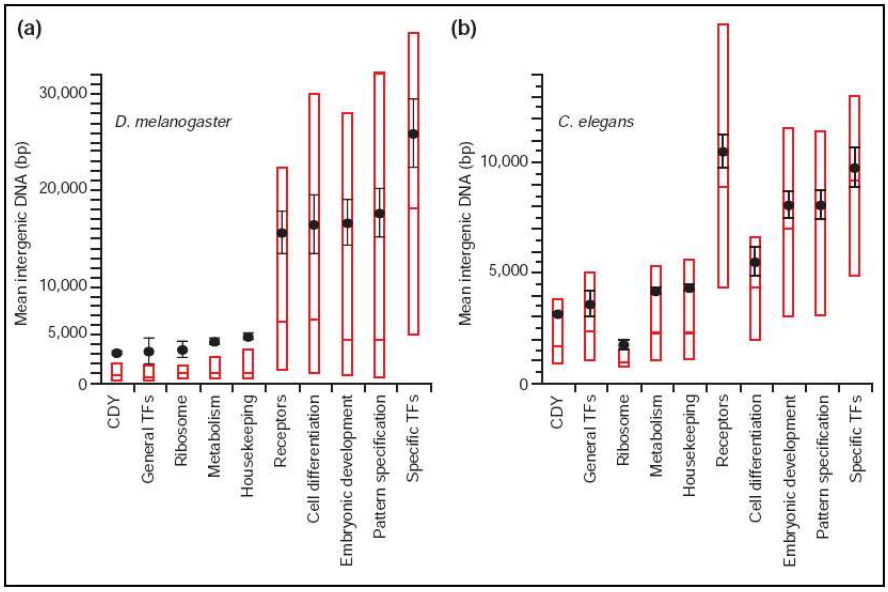


Figure 4.2: Correlation between gene ontologies and the length of the intergenic regions neighbouring these genes

Table 4.9 shows the ratio

$$\sum \text{genes with complex expression patterns} / \sum \text{genes with simple expression patterns}$$

for every chromosome. This ratio is not significantly different in the X-chromosome from the ratios in the other chromosomes. The chromosome where this rate is clearly lower is the autosome 2L. But we cannot find a lower prediction density in this chromosome.

4.3.5 Genomic context of predicted CRMs

Another question we address here concerns the genomic context of our predictions. In Section 4.3.1 we analyse the predicted CRMs without respect to their genomic context. In this section we want to observe the genomic context of our predictions. We classified all predicted CRMs according to their position into the classes *"intergenic"*, *"intron"*, *"start"*, *"stop"*, *"gene"*, *"exon"*, *"acceptor"*, *"donor"* and *"inexon"*. "intergenic" and "intron" contain CRMs which do not overlap any coding sequence. CRMs in the groups "start" and "stop" overlap boundaries of genes. CRMs which are classified into "gene" or "exon" cover a complete gene or complete exon, respectively. "Acceptor" and "donor" mean that the CRMs

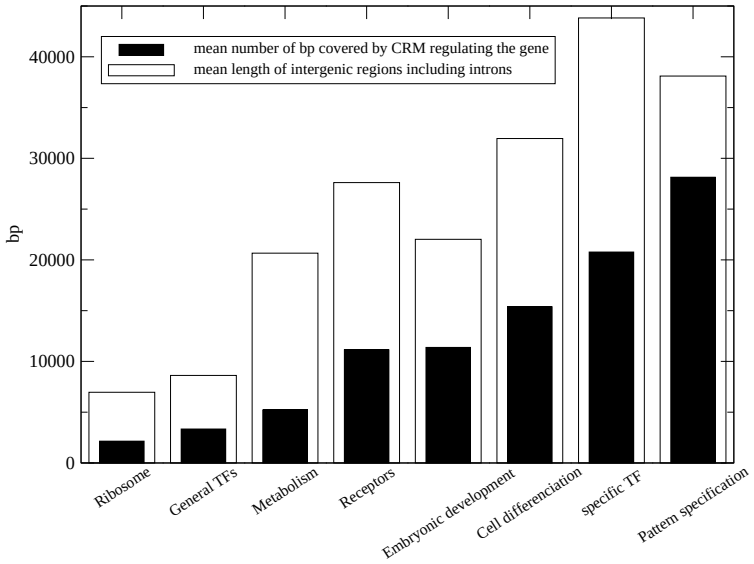


Figure 4.3: Correlation between gene ontologies, length of intergenic regions neighbouring these genes and size of regulatory sequence around and within the gene

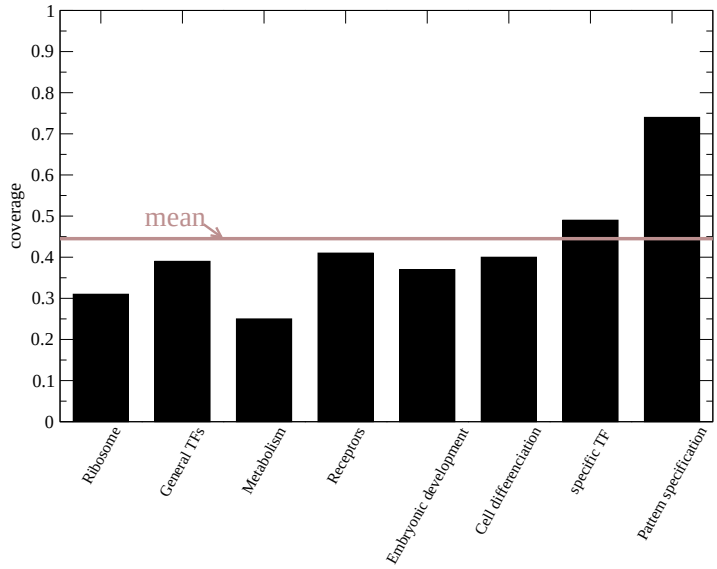


Figure 4.4: Correlation between gene ontologies and the amount of sequence covered by CRMs

Table 4.7: assignment of GO numbers to GO terms

GO-number	GO-term
0003704	specific RNA polymerase transcription factors
0004872	receptor activity
0008152	metabolism
0016251	general Polymerase II TF
0003735	structural constituent of ribosome
0007389	pattern specification
0009790	embryonic development
0030154	cell differenciation

Table 4.8: Number of genes of GO terms on the different chromosomes; the genes were downloaded from flybase (Grumbling *et al.*, 2006).

chromosome	GO:							
	0003704	0004872	0008152	0016251	0003735	0007389	0009790	0030154
2L	13	19	19	14	29	1	4	2
2R	13	17	16	10	25	1	7	2
3L	8	18	16	12	25	1	8	0
3R	18	19	23	16	39	4	6	6
X	10	22	23	10	22	2	7	2

Table 4.9: Number of genes of complex and simple expression profiles on different chromosomes

chromosome	$\sum$ complex expression pattern	$\sum$ simple expression pattern	ratio
2L	39	62	0.63
2R	40	51	0.78
3L	41	53	0.77
3R	63	78	0.81
X	43	55	0.78

in these groups overlap the acceptor or donor of an exon. Predicted CRMs are classified as "inexon", if they are completely positioned within one exon. The results of this grouping are given in Table 4.10 and 4.11.

The most peculiar result is the fact that there are still "inexon"-predictions, when all PLUSs overlapping annotated exons are deleted. This happens, because of inconsistent annotations which are caused by alternative splicing or annotation errors. To mask out PLUSs overlapping exons the first annotation of *CDS (coding sequence)* is taken. Other overlapping annotations of CDS are disrespected. To calculate the genomic context of a CRM we test if it overlaps with any exon. For this reason it might be the case that we do not mask out all PLUSs contained in exons but keep them, if the exons they overlap are alternatively spliced.

Another result is that these results are consistent for all five chromosomes. We do not see any extreme outlier values in Table 4.11. This fact backs up our results and we can assume that the distribution of CRMs across the different groups describing the genomic context is true. If we assume this and partition the dataset into CRMs overlapping coding regions and CRMs that are completely located in intergenic regions we find that 26.26% of the CRMs overlap with coding regions, if we do not discard PLUSs from coding regions. But even, if we discard PLUSs from coding regions still 18.09% of CRMs overlap exons. The remaining 81.09% are completely positioned in noncoding sequence meaning intronic or intergenic sequence.

In prokaryotes and yeast CRMs are mainly positioned upstream of a gene. The fact that many downstream CRMs are known and experimentally verified in *D. melanogaster* arises the question, if there is a tendency of CRMs to be upstream or downstream of the regulated target gene in *D. melanogaster*. We tried to address this question by grouping the CRMs and the intergenic regions which contain predicted CRMs according to the orientation of their adjacent genes. This grouping is explained in Figure 4.5 and the results are summarized in the Tables 4.12 and 4.13. The first number in each cell describes the number of CRMs in the addressed group of intergenic regions. The parenthesized numbers are the total number of intergenic regions containing predictions with adjacent genes of the described orientation.

To declare a tendency towards a relative position of CRMs to the regulated target genes, we expect a clear difference between the amount of CRMs in the divergent group and the amount of CRMs in the convergent group. An optimal example result would be $|divergent| > |tandem - reverse| = |tandem| > |convergent|$. As Table 4.13 clearly shows we cannot find any tendency of the CRMs to be more likely upstream than downstream of any gene and we assume that there is no preference against one or the other model.

We compared the number of CRMs and the number of intergenic regions to ensure that our result is not caused by the total number of intergenic regions of these groups in the *D. melanogaster* genome. If we did not calculate the number of intergenic regions of the groups a high amount of tandem-reverse-CRMs could be caused by the fact, that many more tandem-reverse-intergenic regions contain CRMs. But the consistency of ratios represented in Table 4.13 shows that this is not the case. To confirm this we calculated the average number of predicted CRMs per intergenic region with predictions in every group and for every chromosome and show these results in Table 4.14. Our result remains consistent. These data show a weak but insignificant trend of CRMs to be positioned upstream of a gene.

To prove that no tendency can be found in our data we need to exclude one more possibility to explain our result. Another way of explaining our result is a bias in the grouping of all intergenic regions of *D. melanogaster*. If *D. melanogaster* has a high amount of intergenic regions of the convergent-group in its genome, the fact that this is not reflected in the amount of convergent-regions containing CRM predictions would be very striking. For this reason we calculated the average number of CRM-predictions per intergenic regions in the genome for each group separately. As Table 4.15 shows no significant tendency can be found here either.

Table 4.10: Genomic context of CRM predictions (absolute number of CRMs in relative position to a gene); CRMs are grouped according to their genomic context. The terms describing the groups are explained in the text.

chromosome	inter-genic	intron	start	stop	gene	exon	donor	acceptor	inexon
2L	3990	2298	495	460	0	353	248	575	391
2L (exons excluded)	3989	2308	371	346	0	211	194	401	120
2R	2978	2304	478	495	0	435	225	628	358
2R (exons excluded)	2990	2310	402	405	0	28	182	436	118
3L	4654	2485	319	304	0	602	224	358	296
3L (exons excluded)	4659	2495	260	252	0	300	161	236	84
3R	5737	3361	405	367	0	706	287	485	297
3R (exons excluded)	5752	3373	337	313	0	398	214	345	112
X	2578	1395	171	147	0	368	192	347	298
X (exons excluded)	2552	1376	120	105	0	177	103	196	98

Table 4.11: Genomic context of CRM predictions (amount of CRMs in relative position to a gene). The terms describing the groups are explained in the text.

chromosome	inter-genic	intron	start	stop	gene	exon	donor	acceptor	inexon
2L	0.45	0.26	0.05	0.06	0	0.04	0.03	0.07	0.04
2L (exons excluded)	0.50	0.29	0.04	0.06	0	0.03	0.02	0.05	0.02
2R	0.38	0.29	0.05	0.08	0	0.05	0.03	0.08	0.05
2R (exons excluded)	0.42	0.33	0.04	0.07	0	0.03	0.03	0.06	0.02
3L	0.50	0.27	0.03	0.04	0	0.07	0.02	0.04	0.03
3L (exons excluded)	0.55	0.30	0.03	0.03	0	0.04	0.02	0.03	0.01
3R	0.49	0.29	0.03	0.04	0	0.06	0.02	0.04	0.03
3R (exons excluded)	0.53	0.31	0.03	0.03	0	0.04	0.02	0.03	0.01
X	0.47	0.25	0.03	0.03	0	0.07	0.03	0.06	0.05
X (exons excluded)	0.54	0.29	0.02	0.03	0	0.04	0.02	0.04	0.02

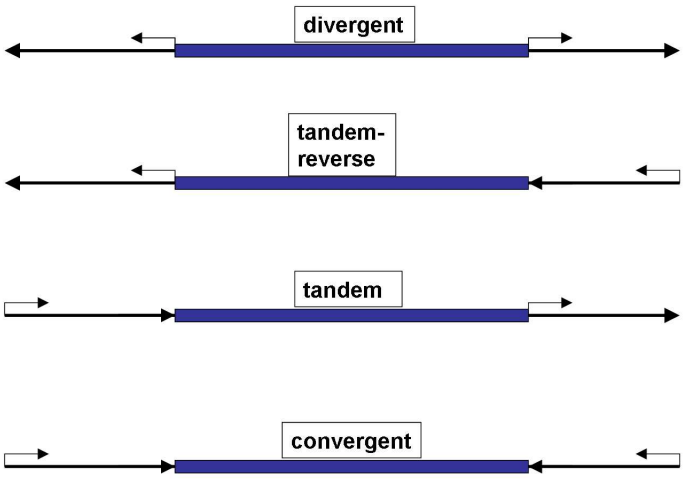


Figure 4.5: Illustration of the grouping of intergenic regions; *divergent* means that the two adjacent genes to the intergenic regions are oriented with the startcodon towards the intergenic region. *tandem-reverse* means that the gene upstream of the intergenic region is positioned with the startcodon towards the intergenic region and the downstream gene points the stopcodon in the direction of the intergenic region. *tandem* means that both adjacent genes are expressed on the forward strand and *convergent* means that the upstream and downstream gene point their stopcodon towards the described intergenic regions.

Table 4.12: absolute number of CRMs in four groups of intergenic regions. The regions are grouped by the orientation of the surrounding genes. An explanantion of the group labels is given in Figure 4.5

chromosome	divergent	tandem-reverse	tandem	convergent
2L	1338 (195)	690 (141)	818 (160)	1144 (197)
2L (exons excluded)	1340 (198)	682 (139)	817 (161)	1150 (198)
2R	855 (176)	689 (145)	621 (159)	813 (194)
2R (exons excluded)	861 (180)	688 (146)	625 (159)	816 (193)
3L	1124 (181)	1476 (186)	1027 (143)	1027 (177)
3L (exons excluded)	1126 (180)	1480 (187)	1027 (143)	1026 (177)
3R	1598 (230)	1395 (222)	1325 (181)	1419 (207)
3R (exons excluded)	1604 (232)	1402 (224)	1326 (182)	1420 (208)
X	634 (132)	760 (146)	480 (109)	704 (130)
X (exons excluded)	636 (134)	761 (147)	480 (108)	675 (128)

Table 4.13: Number of CRMs in four groups of intergenic regions relative to the total number of CRMs or intergenic regions on the chromosome. The regions are grouped by the orientation of the surrounding genes as in Table 4.12

chromosome	divergent	tandem-reverse	tandem	convergent
2L	0.34 (0.28)	0.17 (0.20)	0.21 (0.23)	0.29 (0.28)
2R	0.28 (0.26)	0.23 (0.22)	0.20 (0.24)	0.27 (0.29)
3L	0.24 (0.26)	0.32 (0.27)	0.20 (0.21)	0.22 (0.26)
3R	0.28 (0.27)	0.24 (0.26)	0.23 (0.22)	0.25 (0.25)
X	0.25 (0.26)	0.29 (0.28)	0.19 (0.21)	0.27 (0.25)
average	0.27 (0.27)	0.25 (0.25)	0.21 (0.22)	0.27 (0.27)

Table 4.14: Number of predicted CRMs per intergenic regions with prediction. Standard deviations are given in parentheses.

chromosome	divergent	tandem-reverse	tandem	convergent
2L	6.86 (8.68)	4.89 (5.56)	5.11 (5.81)	5.81 (8.70)
2R	4.86 (8.11)	4.75 (5.15)	3.91 (5.06)	4.19 (7.76)
3L	6.21 (8.32)	7.94 (12.4)	7.18 (8.86)	5.80 (8.42)
3R	6.95 (10.43)	6.28 (7.79)	7.32 (10.24)	6.86 (11.0)
X	4.80 (4.81)	5.21 (6.37)	4.40 (4.76)	5.42 (4.89)
average	6.07	5.96	5.68	5.63

Table 4.15: number of predicted CRMs per intergenic region

chromosome	divergent	tandem-reverse	tandem	convergent
2L	1.77 (5.33)	1.17 (3.46)	1.24 (3.55)	1.51 (4.70)
2R	1.08 (4.33)	0.94 (3.00)	0.88 (2.90)	1.03 (3.27)
3L	1.53 (4.89)	2.26 (7.37)	1.45 (5.28)	1.39 (4.70)
3R	1.66 (5.88)	1.74 (4.98)	1.37 (5.66)	1.47 (5.06)
X	1.01 (2.94)	1.37 (4.08)	0.94 (2.78)	1.12 (2.60)
average	1.43	1.51	1.25	1.59

4.3.6 The distance between a CRM and the regulated target gene

Another question concerning gene regulation regards the distance between a CRM and its regulated gene. For *D. melanogaster* Ochoa-Espinosa *et al.* (2005) assume that all CRMs regulating a target gene are positioned within 20kb upstream of the transcription start site and 20kb downstream of the stop codon. We are not aware of other estimates of this distance in *D. melanogaster*.

To measure the distance between the predictions and the closest genes the distances between the mid-points of the predicted CRMs and the closest genes are calculated. Only predictions which are positioned completely within an intergenic region are considered.

Figure 4.6 shows that most of the predicted CRMs are positioned closely to genes. To get a clearer picture the distances of a CRM to the closest genes were grouped according to their lengths. The result of this analysis is shown in Figure 4.7. The mid point of 51% of the predicted CRMs is positioned within a 5000bp distance from the closest gene. 90.7% of all predicted CRMs are found within 20kb around a gene. 3% of the predictions have a distance higher than 30kb from the closest gene.

Figure 4.1 shows that CRMs are more likely to be predicted near genes than in genomic regions of low gene density. To see, if CRMs are more likely to be positioned in a certain part of the intergenic region we calculated the relative distance of CRMs to the nearest gene. For this purpose the distance between the mid point of a predicted CRM and the start or stop codon of the nearest gene is calculated and this value is divided by the length of the intergenic region. The result of this analysis is shown in Figure 4.8. We cannot detect any preference to any relative position within the intergenic region. CRMs are as likely to be predicted in the middle of intergenic regions as they are near the edges of intergenic regions. The drop of the curve towards the relative distance "0" is caused by the fact, that we observed exclusively intergenic regions. Predictions which have their midpoint very close to a gene are likely to overlap the gene.

4.4 Discussion

We analysed the results of the WGS considering two questions. One question was, whether `CisPlusFinder` is applicable to a complete genome. The other question concerns conclusions

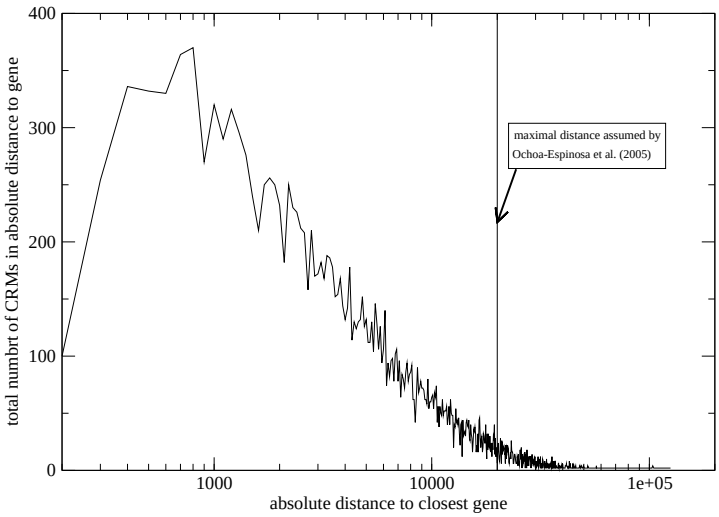


Figure 4.6: Number of occurrences of absolute distances between CRM and the regulated gene

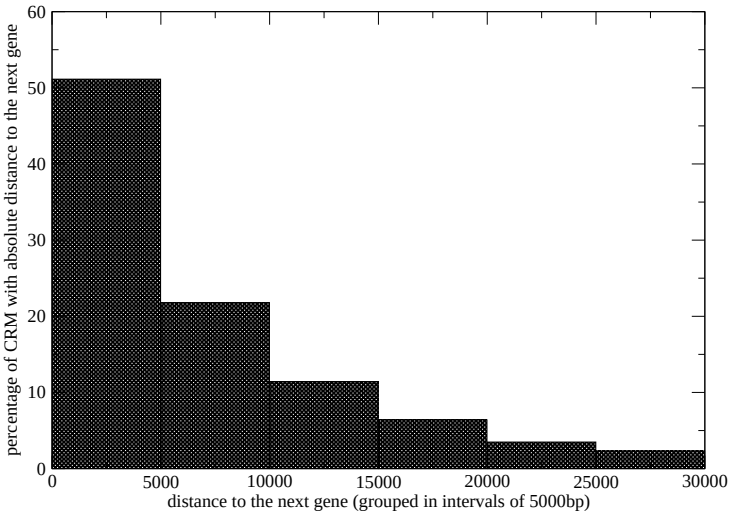


Figure 4.7: Distance between CRM and the regulated target gene; The distances are grouped into 6 groups (0-5000, 5000-10000, 10000-15000, 15000-20000, 20000-25000, 25000-30000)

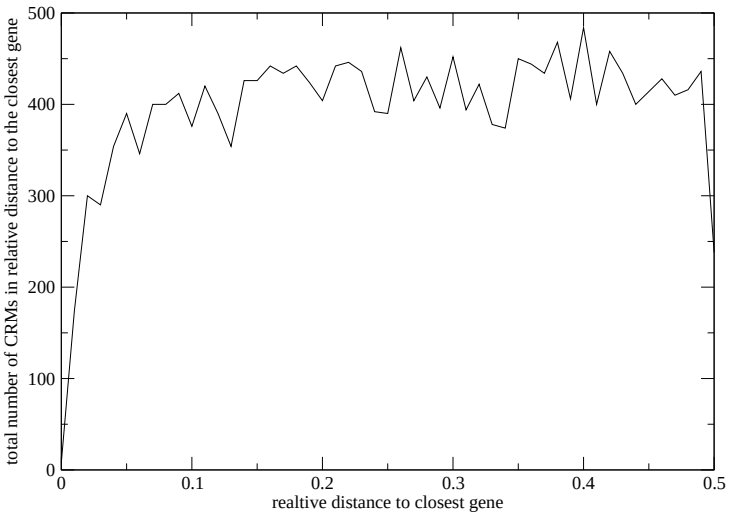


Figure 4.8: Number of occurrences of relative distances between CRM and the regulated gene

which can be drawn from a global survey of regulatory sequences.

Acceptance of prediction

In Chapter 3 the method `CisPlusFinder` was tested on different sets of experimentally verified CRMs and was shown to be a very valuable predictor of regulatory function within short genomic sequences. In Section 4.3.1 we showed that the prediction of `CisPlusFinder` of the whole genome of *D. melanogaster* can be regarded as meaningful. Predictions are nonrandomly distributed along the sequence and their density is significantly positively correlated with gene density. The proportion of predicted sequence is roughly constant on the autosomes and the average length of the prediction is within the expected range between 500bp and 1000bp. The average length of PLUSs is approximately 12.5, the average length of a TFBS calculated by Bilu & Barkai (2005) (see Table 4.4).

Prediction coverage on the X-chromosome

In Section 4.3.3 we noticed a difference in the prediction density of the X-chromosome compared to the autosomes. We find that the effect can be explained by a combination of a higher substitution rate and a lower gene density. But 14% weaker sequence conservation are unlikely to cause 23% less CRMs per gene. To elucidate this we suggest one of the following explanations.

- Assuming a higher neutral substitution rate of the X-chromosome because of the smaller population size of this allosome guides us to the following explanation. 57% of the X-chromosome are covered by genes. The higher evolutionary turnover is mainly measured in intergenic regions, because substitutions in genes are less likely to be neutral. When we assume that the conservation scores for the regions covered by genes are similar, the difference in substitution rates has to be higher in intergenic regions than in average along the whole genome and can probably cause the found difference in prediction density.

- If we assume the fast-X evolution, introduced by Counterman *et al.* (2003), we expect most substitutions to be functional. The difference in substitution rate is higher within coding regions than in noncoding regions and the assumption in the last item is wrong under this explanation. But, if we assume that most substitutions are functional, we expect a very high turnover of TFBS concentrated on sites, which change the affinity of the TFBS to the binding transcription factor. If this assumption is true it is very clear, that TFBS are less likely to cause the occurrence of PLUSs in the X-chromosome. This presumption would explain our result as well.

Since it is not known whether the higher substitution rate of the X-chromosome is caused by a higher rate of neutral or functional substitutions we tried to support one of the hypotheses with our result. As explained above the CRMs predicted by `CisPlusFinder` can be explained by and support both theories in the same way and we cannot draw further conclusions.

The correlation between the function and the complexity of the regulation of a gene

Section 4.3.4 follows up on an idea introduced by Nelson *et al.* (2004). They found a very clear relationship between the GO terms assigned to a gene and the lengths of the intergenic regions adjacent to the gene or introns within the gene (Figure 4.2). Based on this result they stated that the length of these noncoding regions is necessary within the genome to host the modules which perform the complex regulation of these genes.

We found evidence which supports their hypothesis. We confirm their trend using current annotations and find a stable amount of regulatory sequence in intergenic regions independent on the length of the intergenic region or the GO term of the gene. This gives evidence that CRMs need this amount of DNA to code their function and cannot be compressed. Figure 4.1 indicates as well that an upper limit for the density of CRMs exist. We believe that the need of complex gene regulation influences the genome architecture and cause long intergenic regions which are still believed to be nonfunctional by many scientists.

Predicted CRMs and their genomic context

In prokaryotes and yeast most sequences involved in gene regulation are located upstream of the regulated gene. Some rare cases are known where transcription factor binding sites are located downstream of the transcription start site. According to this we expect to find a tendency of CRMs in *D. melanogaster* to be positioned upstream of the gene they regulate. This trend would be found by the classification of CRMs in intergenic regions as it was done in Section 4.3.5. Surprisingly no trend was found in our data. This fact supports the assumption that an increase in the complexity of gene regulation accompanied the split between simple organisms such as prokaryotes and yeast and more complex organisms such as *D. melanogaster*.

Distance between CRMs and their target gene

To estimate the physical distances between CRMs and the gene they regulate, we calculated the distances between the mid points of the predictions and the start or stop of the closest gene. We found that most of the predictions were positioned within a 20kb distance around the gene.

The maximal distance we could find between the mid point of a CRM and the closest gene is 125100bp. This distance is much higher than expected and can be explained in different ways. One explanation is that predictions with such a large distance to the closest gene are false positive predictions and no functional CRMs. Another possible explanation is that the chromatin structure of the DNA between the CRM and the closest gene connects the gene and the CRM to enable their interaction. The structure of the DNA between a CRM and its gene could also be regulated by binding proteins and this could be a further mechanism of transcriptional regulation.

If we assume that this prediction is not caused by a functional CRM, the finding of sequence conservation within an intergenic region of this length remains. If a cluster of PLUSs is found in such a big distance from the gene the probability that this sequence performs a function is very high. This function could be distinct from transcriptional regulation. The finding of functional sequence in the middle of extremely long intergenic regions remains surprising.

Sequence conservation and regulatory function

The regions located by `CisPlusFinder` are characterized by sequence conservation and statistical properties of the DNA. Sequence conservation is generally seen to be a signal of functional DNA (Loots *et al.*, 2000, Bergman & Kreitman, 2001, Boffelli *et al.*, 2003, Frazer *et al.*, 2004, Johnson *et al.*, 2004, Woolfe *et al.*, 2004). Despite protein coding genes and transcriptional regulation other functions of genomic DNA are known, e.g. non coding RNAs, origins of replication, regulation of alternatively spliced exons and introns and the definition of recombination hot spots.

`CisPlusFinder` tries to exclude these regions by excluding PLUSs whose length exceeds the length of a TFBS significantly (Chapter 3) and requiring overrepresentation of a core motif. But we cannot exclude the possibility that we find regions, which are conserved because of functional constraints different from transcriptional regulation. Especially biological processes requiring the binding of proteins might misguide our method. A more detailed analysis of other possible functions of our predictions is a challenging task for the future.

Summarizing we note that a WGS with `CisPlusFinder` is possible. This WGS indicates that the method can be extended from the analysis of short intergenic regions to a large scale application. Many results of the WGS correspond to the expectations. Therefore we can assume with a high level of confirmation that the straight forward application of `CisPlusFinder` to a complete genome leads to a reasonable prediction which allows to draw conclusions about general rules of transcriptional regulation.

The most striking result of the WGS is the great relevance of transcriptional regulation for the complex multicellular organism *D. melanogaster*. We found further evidence for the idea that regulatory function influences genome architecture which was believed to be determined by random genome duplications, insertions and deletions. This result is another signal of the dense packing of the Drosophila genome caused by the high rate of DNA loss reported by Petrov & Hartl (1998).

Another aspect which is made clear by this WGS is that the difference between simple organisms like prokaryotes and yeast and a complex organism like *D. melanogaster* is accompanied by an increase in complexity of transcriptional regulation. The high complexity of gene regulation has been detected in embryonic development. It is crucial to organize the complex anatomy of *D. melanogaster*. This complex gene regulation requires CRMs which are located upstream or downstream of the gene. We can assume that the DNA between a CRM and the regulated gene has certain properties concerning structure and bendability but cannot make any further assumptions about it at this stage.

The occurrence of predictions in great distance from the closest gene are signals of sequence conservation within regions where it was completely unexpected. This fact combined with the high loss of DNA within the Drosophila group supports the idea that the nonfunctional part of the genome of *D. melanogaster* is very small.

Chapter 5

Conclusion and Perspectives

At the beginning of this thesis the aim was to develop a computational method to identify CRMs which can be applied to a complete genome. In Chapter 2 we present the method `Shureg`, which uses exclusively information from the analysed sequence itself and is therefore applicable to any sequence from any species. The accuracy of this method is quite low and cannot be improved at the current data status without using additional information.

Additional information which is not specific to certain genes or groups of genes and which is available for many sequences can be obtained from sequence comparison between different species. For this reason we developed the method `CisPlusFinder`, described in Chapter 3, which searches for conservation patterns, that are caused by regulatory function. We applied and evaluated `CisPlusFinder` on experimentally verified CRMs from *D. melanogaster*. It is hard to measure the performance of a method with the current data status, but the predictions of `CisPlusFinder` seem to be trustworthy and the number of missed CRMs is very low.

In Chapter 4 we applied the method `CisPlusFinder` to the whole genome of *D. melanogaster* and acchieved results that allows us to formulate new hypotheses concerning transcriptional regulation. We used *D. yakuba, D. ananassae, D. pseudoobscura* and *D. virilis* as informant species. The finding of a CRM by this application of the method means that the found CRM contains at least three TFBS which are perfectly conserved through all these species beyond a significant length.

Conservation of gene regulation on transcriptional level

Emberly *et al.* (2003) state that TFBS are significantly conserved between *D. melanogaster* and *D. pseudoobscura* but that this conservation is not much greater than by chance. Dermitzakis & Clark (2002) find a very high evolutionary turnover rate within TFBS between human and rodent species. The reasonable performance of `CisPlusFinder` shows that the process of transcriptional regulation of most genes is conserved within the Drosophila species. We show in Section 3.4.4 that one third of the known TFBS are shown to be highly conserved.

These statements do not contradict each other. The fact that CRMs cause a sufficient number of PLUSs does not require that all TFBS are perfectly conserved. The number of TFBS which are perfectly conserved by selective pressure or by chance need to be high enough to cause the demanded number of PLUSs. Those TFBS which do not cause PLUSs might undergo a high turnover.

Regarding the assumptions that only few substitutions within a TFBS change its affinity to the binding transcription factor and that even the change of the binding affinity is not lethal in most cases the conservation rate of TFBS is still higher than expected. One reason might be the fact that bound transciption factors protect the DNA from factors that advance substitutions. Another reason might be that most nonlethal substitutions within TFBS are negatively selected. The effect of negative selection is much bigger in species of a high population size like *D. melanogaster* than in species of small population size like

most mammals. The probability of a substitution that compensates the disadvantageous substitution is very small and the disadvantageous allele might be removed from the population before the balancing substitution has happened. This explains also the comparatively high turnover within mammals reported by Dermitzakis & Clark (2002). The population size of mammals is drastically smaller than the population size of fruit flies. Substitutions which are weakly negatively selected remain in the mammal populations much longer than in Drosophila population and the probability that they are compensated by a balancing substitution is bigger in mammals.

Relevance of the complexity of transcriptional regulation for *D.melanogaster*

We described already that the high conservation of gene regulation within the rapidly evolving Drosophila species is surprising. This is a first hint that the process of gene regulation is of very high relevance for fruit flies. Our analysis brought up more results indicating this idea.

CRMs regulating the genes of *D. melanogaster* do not show any preference to a relative position to the regulated gene. They occur as likely upstream as downstream. Also a tendency towards a certain relative distance to the gene could not be observed. These findings present a drastic difference between simple organisms and species of high complexity. Simple species locate the complete DNA necessary to perform regulatory function upstream of the DNA and the amount of DNA which is necessary to regulate one gene is rarely longer than several hundred basepairs. In *D. melanogaster* 49% of all CRMs are found to be located more than 5kb distant from the gene they regulate. In mammals the maximal distance between a CRM and the gene it regulates is believed to be even longer. The absolute amount of sequence to host the regulatory modules increases with the complexity of the organism. The extension of a very specific, highly complex mechanism to regulate the transcription to most of the genes accompanied the split of complex species from simple species like prokaryotes and yeast. This step seems to be combined to the introduction of multicellular organisms which assign every cell to a certain tissue.

Another hint for the extremely high relevance of transcriptional regulation is the support of the theory that transcriptional regulation shapes genome architecture. This theory was introduced by Nelson *et al.* (2004) and we could find further evidence supporting it. This theory reveals that long intergenic regions are protected against the high DNA loss in Drosophila melanogaster, because they perform regulatory function. We found that long intergenic regions adjacent to genes which are expressed in complex patterns are hosts of CRMs. The fact that the CRM coverage is stable in short and long intergenic regions indicates that there is a maximal coverage by CRMs which cannot be exceeded without disturbing the function. Given this evidence we assume that these long intergenic regions are necessary to host all required CRMs and a loss of them is lethal for the fruit fly.

This work gives strong evidence that the increase in genome complexity which is correlated to an increase of the complexity of organisms is mainly caused by an increase in complexity of gene regulation. The mechanism used by higher organisms to regulate gene expression occupies a large amount of non coding DNA in *D. melanogaster*. Regulatory processes have been well maintained and probably improved through evolution. By now we can have only an idea about the relevance of transcriptional regulation for one species or the evolution. Further projects in the future will probably provide further evidence to the immense weight of transcriptional regulation.

Sequence conservation in intergenic sequences

Another result of this thesis is the finding of sequence conservation within very long intergenic sequences. Figure 5.1 shows different conservation measures for an intergenic sequence of length 251418bp. `CisPlusFinder` predicts CRMs in this region whose distance to the closest gene is much higher than expected. The maximal distance of a CRM within this intergenic region to the closest gene is 125100bp. These CRMs are not experimentally tested yet and the occurrence of sequence rearrangements decreases the trust in this predictions.

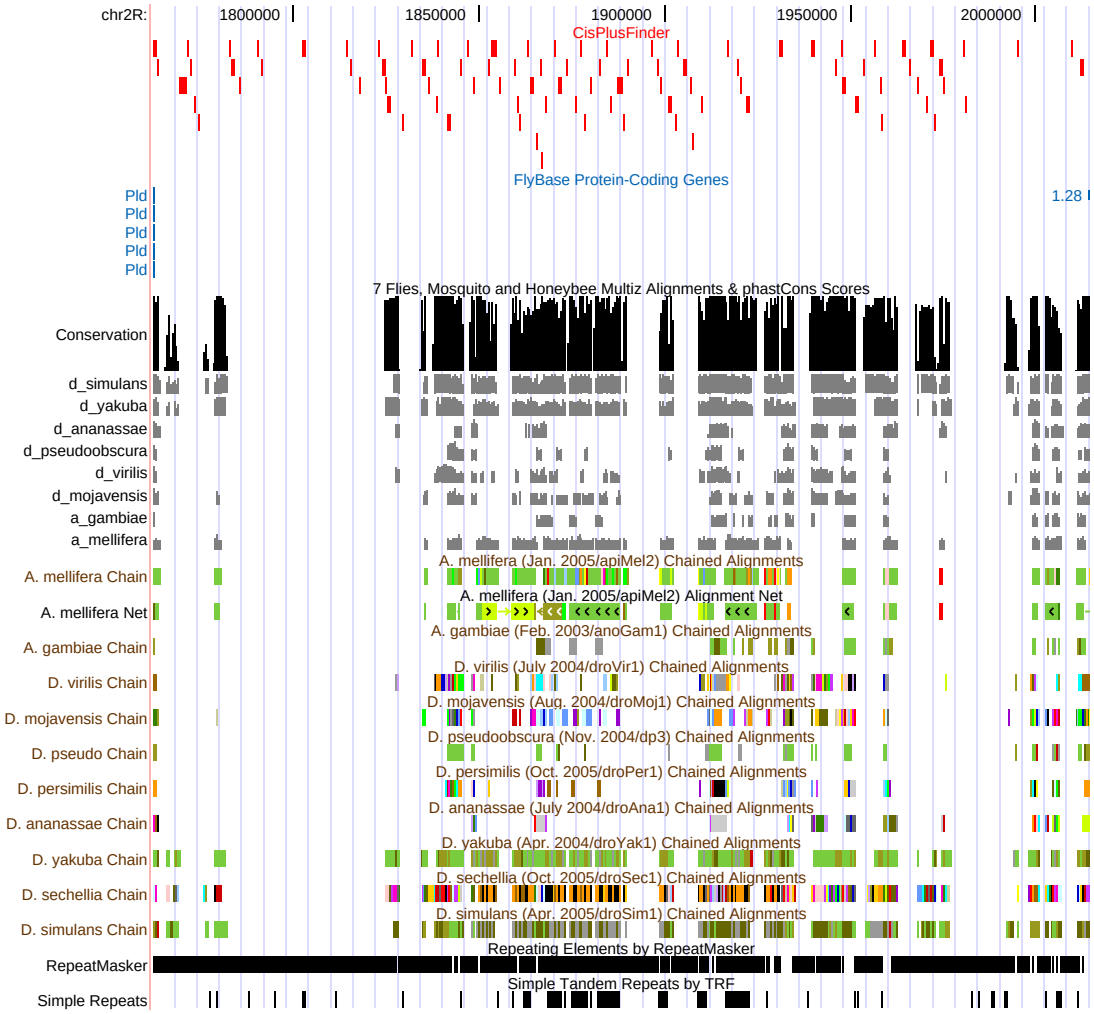


Figure 5.1: Predictions of `CisPlusFinder`, PhastCons conservation score and the chained alignment fragments within a conserved intergenic region, which is 251418bp long. This graphic was generated using the UCSC Browser (Kent *et al.*, 2002) http://genome.ucsc.edu/.

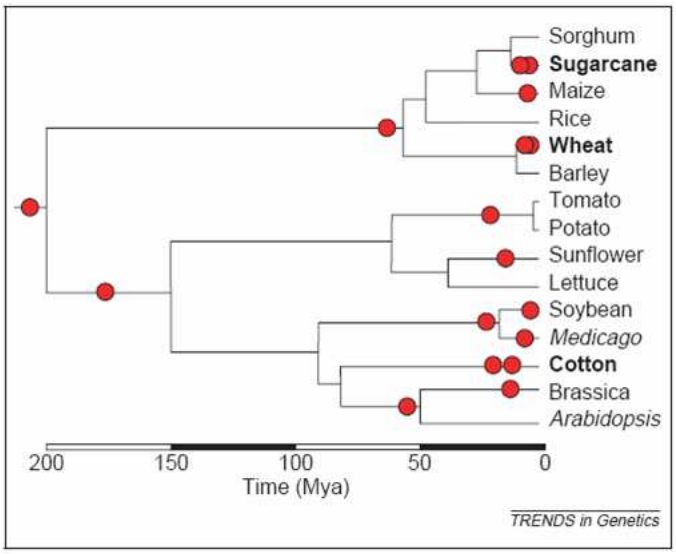


Figure 5.2: Phylogenetic tree of different plants; Red dots represent whole genome duplications.
Source: Lockton & Gaut (2005)

But the fact that the finding of sequence conservation between different species is consistent between different methods supports the idea that some function is associated to this intergenic region.

Predictions of `CisPlusFinder` are also found in regions, where PhastCons does not indicate sequence conservation between *D. melanogaster* and all informant species. This points to a conservation structure which corresponds to the conservation structure of CRMs. Rather short motifs have to be perfectly conserved within an environment of low sequence conservation. Assumptions about the function would be guesses at this moment. We have to wait until further experiments are done to develop an idea about the incidents in this or other long, well conserved intergenic regions.

Extending `CisPlusFinder` to other species

We state in Chapter 3, that the method `CisPlusFinder` is easily extendible to other species, if enough species in appropiate evolutionary distance are sequenced. The extension of `CisPlusFinder` for the application to different species would be very helpful for the clarification of their transcriptional regulation. To test, if our statement proves true, we planned to apply `CisPlusFinder` to *Arabidopsis thaliana*, a small flowering plant and a model organism in plant genetics.

Figure 5.2 shows a tree of different plant species. Some of them are completely sequenced. For the other species whole genome sequences will be available in the near future. Every red dot in Figure 5.2 represents a whole genome duplication. For the purpose of applying `CisPlusFinder` we need to chose informant species, where no genome duplication happened since the split between them and *A. thaliana*. After a whole genome duplication genes as well as the regulation of genes are free to undergo functional substitutions as long as one of both gene copies ensures its functionality. The probability of one or multiple TFBS causing a PLUS is arbitrarily reduced in this case.

Figure 5.2 shows that no appropriate informant species for *A. thaliana* is completely sequenced at the moment. To test the applicability of `CisPlusFinder` to *A. thaliana* in cases, where enough appropriate informant species are sequenced we analysed one very well explored gene, the *chalcone synthase*. Koch *et al.* (2000) identified a region which is involved in the transcriptional regulation and obtained sequences for this region from 15 species within the Arabidopsis and Arabis family and one outgroup from the family of Brassicaceae. The

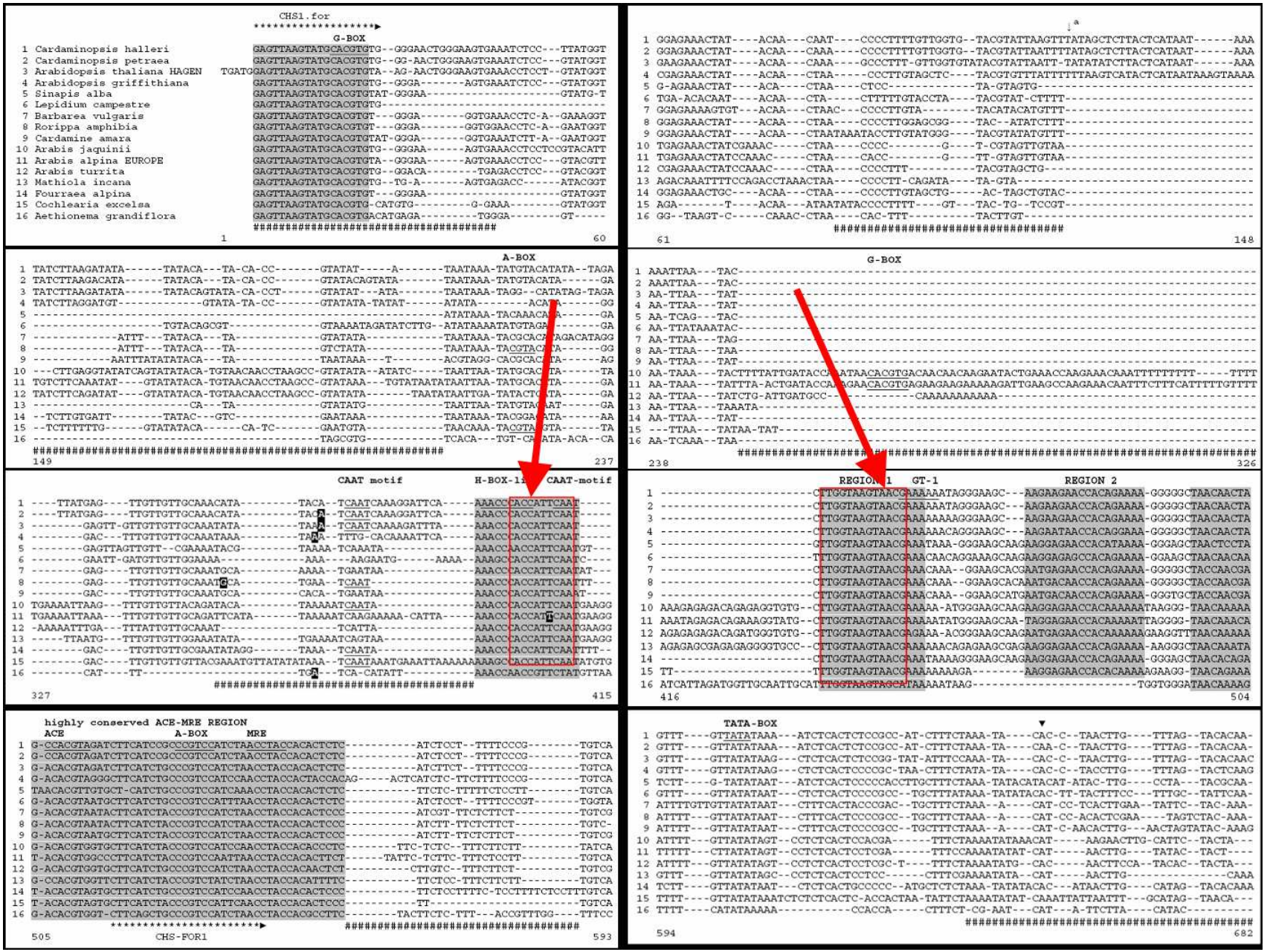


Figure 5.3: Alignment of a regulatory region of the *chalcon synthase* from *A. thaliana* and 14 informant species; The arrows and red boxes show TFBS causing PLUSs.
The alignment is taken from Koch *et al.* (2000).

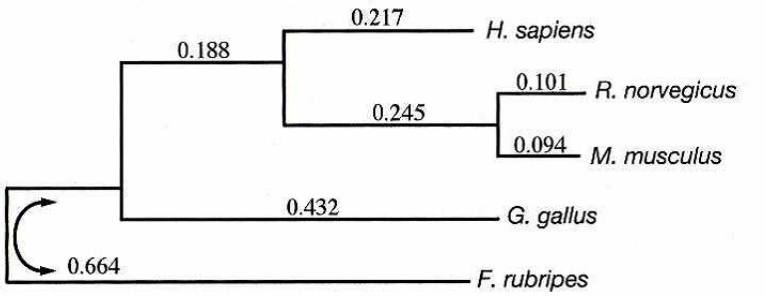


Figure 5.4: Phylogenetic tree of human, mouse, rat, chicken and pufferfish based on non-conserved regions and created with `PhastCons` (Siepel *et al.*, 2005) based on a multiple alignment using `Multiz` (Blanchette *et al.*, 2004)
Source: Siepel *et al.* (2005)

alignment of the homologue sequences of all species is shown in Figure 5.3. We applied `CisPlusFinder` to the set of 15 species from species belonging to the Arabidopsis and Arabis family. We could detect the motifs, which are marked with red boxes as PLUSs of significant length. This result gives some confidence that the application of `CisPlusFinder` combined with the sequencing of closely related species of *A. thaliana* could help in discovering transcriptional regulation in plants.

Another set of species where the unravelling of transcriptional regulation is a big goal is formed by mammals. Figure 5.4 shows the phylogenetic tree of humans, rodents, chicken and pufferfish. The comparison of this tree and the tree of the Drosophila species shown in Figure 3.5 brings up a striking difference. The substitution rate between the phenotypically very different mammals *Homo sapiens* and *Mus musculus* is approximately equal to the substitution rate between the phenotypically closely related Drosophila species *D. melanogaster* and *D. ananassae*.

This fact complicates the analysis of mammals. Many species need to be included to acchieve the required evolutionary distance between the target species and the informant species. The inclusion of every additional species can effect that a CRM cannot be detected any more, because it is not present in the added species.

As a first approach to test `CisPlusFinder` on mammalian sequences we applied it to a test set of human using the informant species mouse, dog, cow and opossum. 131 experimantally verified CRMs were downloaded from the database ORegAnno (Montgomery *et al.*, 2006). The results are shown in Table 5.1. The sensitivity of `CisPlusFinder` applied to mammals is very low in comparison to the results in Drosophila. The transcriptional regulation seems to be less conserved between mammals even when the substitution rate is decreased. To extend `CisPlusFinder` in a way that it can be applied to mammals requires some modifications of the method. The requirement of perfectly conserved TFBS is probably too strong within this species group. The allowance of mismatches will cause many false positive predictions, because of the low substitution rate between the species. Pennacchio *et al.* (2006) prove experimentally the regulatory function of highly conserved regions of *H. sapiens*. We can conclude from their analysis that regulatory function can cause a very strong conservation signal within mammals, if the function is conserved. The close examination of these cases could lead to new insights into the conservation of mammalian gene regulation.

Table 5.1: Accuracy of `CisPlusFinder` applied to sequences of *H. sapiens*

	SN(CRM)	PPV(CRM)	SP(nuc.)	SN(nuc.)	PPV(nuc.)	hit CRMs	missed CRMs
CisPlusFinder	**0.28**	**0.11**	**0.14**	**0.05**	**0.99**	**37**	**94**

Bibliography

Abnizova,I., te Boekhorst,R., Walter,K., Gilks,W. (2005) Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the *Drosophila* genome: the fluffy-tail test, *BMC Bioinformatics*, **6**, 109.

Allen,J., Salzberg,S. (2005) JIGSAW: integration of multiple sources of evidence for gene prediction *Bioinformatics*, **21**, 3596-3603.

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

Bailey,T., Noble,W. (2003) Searching for statistically significant regulatory modules. *Bioinformatics*, **19 Suppl 2**, II16-II25.

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

Benson,DA., Karsch-Mizrachi,I., Lipman,Dj., Wheeler,DL. (2006) GenBank *Nucleic Acids Res.*, **1;34**, D16-20.

Berg,O., von Hippel,P. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters *J. Mol. Biol.* **193**,723-750

Bergman,C., Carlson,J., Celniker,S. (2005) *Drosophila* DNase we footprint database: A systematic genome annotation of transcription factor binding sites in the fruitfly, *D. melanogaster Bioinformatics*, **21**, 1747-1749.

Bergman,C., Kreitman,M. (2001) Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. *Genome Res.* **11**:1335-1345

Berman,B., Nibu,Y., Pfeiffer,B., Tomancak,P., Celniker,S., Levine,M., Rubin,G., Eisen,M. (2004) Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci.*, **2**, 757-762.

Bilu,Y., Barkai,N. (2005) The design of transcription factor binding sites is affected by combinatorial regulation. *Genome Biology*, **6**, 103.

Blanchette,M., Kent,W., Riemer,C., Elnitski,L., Smit,A., Roskin,K., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E., Haussler,D., Miller D. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708-715.

Blanchette,M., Bataille,A., Chen,X., Poitras,C., Laganiere,J., Lefebvre,C., Debloise,G., Giguere,V., Ferretti,V., Bergeron,D., Coulombe,B., Robert,F. (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression *Genome Res.*, **16**, 656-668.

Bofelli,D., McAuliffe,J., Ovcharenko,D., Lewis,K., Ovcharenko,I., Patcher,L., Rubin,E. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391-1394.

Brudno M., Do C., Cooper G., Kim M., Davydov E., Green E., Sidow A., Batzoglou S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, **13**, 721-731.

Burset,M., Guigo,R. Evaluation of gene structure prediction programs. *Genomics* **34**,353-367

Cartharius,K., Frech,K., Grote,K., Klocke,B., Haltmeier,M., Klingenhoff,A., Frisch,M., Bayerlein,M., Werner,T. (2005) MattInspector and beyond: promoter analysis based on transcription factor binding sites *Bioinformatics* **21** 2933-2942

Celniker,S., Sharma,S., Keelan,D., Lewis,E. (1990) The molecular genetics of the bithorax complex of Drosophila: cis-regulation of the abdominal-B domain

Capelson,M., Corces,V. (2004) Boundary elements and nuclear organization *Biol. of the Cell* **96** 617-629

Chan,B., Kibler,D. (2005) Using hexamers to predict *cis*-regulatory motifs in *Drosophila*. *BMC Bioinformatics*, **6**, 262.

Coleman,R., Pugh,B. (1995) Evidence for functional binding and stable sliding of the TATA binding protein on nonspecific DNA. *J. Biol. Chem.*, **270**, 13850-13859.

Counterman,B., Ortiz-Barrientos,D., Noor,M. (2003) Using comparative genomic data to test for fast-X evolution *Evolution* **58:3** 656-660

Couronne,O., Poliakov,A., Bray,N., Ishkanov,T., Ryaboy,D., Rubin,E., Patcher,L., Dubchak,I. (2003) Strategies and tools for whole genome alignments. *Genome Res.* **13:1** 73-80.

Day,W.H., McMorris,F. (1992) Critical comparison of consensus methods for molecular sequences *Nucl. Acids Res.* **20**,1093-1099

Dermitzakis,E., Clark,A. (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover *Mol. Biol. and Evol.* **19**:1114-1121

Dermitzakis,E., Bergman,C., Clark,A. (2003) Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites. *Mol. Biol. Evol.* **20(5)**:703-714

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

Frazer,K., Tao,H., Osoegawa,K., de Jong,P., Chen,X., Doherty,M., Cox,D. (2004) Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional, *Genome Res.* **14**;367-372

Frith,M., Li,M., Weng,Z. (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.*, **31**, 3666-3668.

Gallo,S., Li,L., Halfon,M. (2005) REDfly: a regulatory element database for *Drosophila*. *Bioinformatics*, **22**, 381-183.

Ganesh,R., Ioerger,T., Siegele,D. (2003) MOPAC: Motif finding by preprocessing and agglomerative clustering from microarrays. *Pac. Sym. on Biocomputing* **8**:41-52

Grabe,N. (2004) AliBaba2: context specific identification of transcription factor binding sites *In Silico Biol.* **2(1)**:S1-1

Grad,Y., Roth,F., Halfon,M, Church,G. (2004) Prediction of similarly acting *cis*-regulatory modules by subsequence profiling and comparative genomics in *Drosophila melanogaster* and *D. pseudoobscura*. *Bioinformatics*, **20**, 2738-2750.

Grumbling,G., Strelets,V., The FlyBaseConsortium (2006) FlyBase: anatomical data, images and queries *Nucl. Acids Res.* **34**: D484-D488

Gusfield, D. (1997) Algorithms on strings, trees and sequences. *Cambridge University Press.*

Halligan D., Keightley P. (2006) Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Research* **16** 875-884

Hasegawa,M., Kishino,H., Yano,T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160-174.

Haubold, B., Pierstorff, N., Moeller, F., Wiehe, T. (2005) Genome comparison without alignment using shortest unique substrings. *BMC Bioinformatics*, **6**, 123.

Hughes,J.D., Estep,P.W., Tavazole,S., Church,G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae *J. Mol. Biol.* **296** 1205-1214

Jacob, F., Monod, J (1961) Genetic regulatory mechanisms in the synthesis of proteins *J. Mol. Biol.* **3** 318-356.

Johansson,O., Alkema,W., Wassermann,W., Lagergren,J. (2003) Identification of functional clusters of transcription factor binding motifs in genomic sequences: the MSCAN algorithm. *Nucleic Acid Res.*, **19 Suppl I**, 169-176.

Johnson,D., Davidson,B., Brown,C., Smith,W., Sidow,A. (2004) Noncoding regulatory sequences of Ciona exhibit strong correspondence between evolutionary constraint and functional importance. *Genome Res.* **14**:2448-2456

Jukes, T.H., Cantor, C.R. Evolution of protein molecules. In *Mammalian Protein Metabolism*, H.N. Munro, ed., *Academic Press*, New York, 21-232.

Kassis,J., Desplan,C., Wright,D., O'Farrel,P (1989) Evolutionary conservation of homeodomain binding sites and other sequences upstream and within the major transcription unit of the Drosophila segmentation gene engrailed. *Mol. Cell Biol.* **9**:4304-4311

Kel,A., Gossling,E.,Reuter,I., Cheremushkin,E., Kel-Margoulls,O., Wingender,E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucl. Acids Res.* **31**:3576-3579

Kent,W., Sugnet,C., Furey,T., Pringle,K., Zahler,A., Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996-1006.

Khaitovich,P., Tang,K., Franz,H., Kelso,J., Hellmann,I., Enard,W., Lachmann,M,. Paabo,S. (2006) Positive selction on gene expression in the human brain. *Curr. Biol.*, **16**, R356-R358.

Kim,J., Takeda,Y., Matthews,B. ,Anderson,W. (1987) Kinetic studies on Cro repressor-operator DNA interaction. *J. Mol. Biol.*, **196**, 149-158.

King,D., Taylor,J., Elnitski,L., Chiaromonte,F., Miller,W., Hardison,R. (2005) Evaluation of regulatory potential and conservation scores for detecting *cis*-regulatory modules in aligned mammalian genome sequences. *Genome Res.*, **15**, 1051-1060.

Kechris,K., Zwet,E., Bickel,P., Eisen,M. (2004) Detecting DNA regulatory motifs by incorporating positional trends in information content *Genome Biology*45:R50

Khory,A. ,Lee,H. ,Lillis,M. ,Lu,P. (1990) Lac repressor-operator interaction: DNA length dependence *Biochim. Biophys. Acta*, **1087**, 55-60.

Koch,M., Haubold,B., Mitchell-Olds,T. (2000) Comparative Evolutionary Analysis of Chalcone Synthase and Alcohol Dehydrogenase Loci in Arabidopsis, Arabis, and Related Genera (Brassicaceae) *Mol. Biol. and Evol.* **17**:1483-1498

Krasilnikov,A., Podtelezhnikov,A, Vologodskii,A., Mirkin,S. (1999) Large scale Effects of Transcriptional DNA Supercoiling in Vivo *Journal of Mol. Biol.* **292**, 1149-1160

Lankas,F., Sponer,J., Hobza,P., Langowski,J. (2000) Sequence-dependent Elastic properties of DNA *Journ. of Mol. Biol.* **299**, 695-709

Lifanov,A., Makeev,V., Nazina,A., Papatsenko,D. (2003) Homotypic regulatory clusters in *Drosophila*, *Genome Res.*, **13**, 579-588.

Liu,Y., Wei,L., Batzoglou,S., Brutlag,D., Liu,J., Liu,X. (2004) A suite of web-based programs to search for transcriptional regulatory motifs, *Nucl. Acids Res.* **32**

Levine,M., Tijan,R. (2003) Transcription regulation and animal diversity, *Nature*, **424**, 147-151.

Lockton,S., Gaut,B. (2005) Plant conserved non-coding sequences and paralogue evolution *Trends Genet.* **21**:60-65

Loots,G., Locksley,R., Blankespoor,C., Wang,Z., Miller,W., Rubin,E., Frazer,K. (2000) Identifaication of a coordinate regulator of interleukins 4, 13 and 5 by cross-species sequence comparisons. *Science* **288**:136-140

Loots,G., Ovcharenko,I. (2004) rVista 2.0: evolutionary analysis of transcritption factor binding sites.*Nucl. Acids Res.* **32**:W217-W221

Ludwig,M., Bergman,C., Patel,N., Kreitman,M., (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, **403**, 564-567.

Lunter,G., Ponting,C., Hein,J. (2006) Genome-Wide identification of human functional DNA using a neutral indel model *PLOS Comp. Biol.* **2:1** e5

Manak,J., Dike,S., Sementchenko,V., Kapranov,P., Biemar,F., Long,J., Cheng,L., Bell,I., Ghosh,S., Piccolboni,A., Gingeras,T. (2006) Biological function of unannotated transcription during the early development of Drosophila melanogaster *Nature Genetics* **38**,1151-1158

Markstein,M., Markstein,P., MArkstein,V., Levine,M. (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila genes *PNAS* **99** 763-768

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

Montgomery,S., Griffith,O., Sleumer,M. Bergman,C., Bilenky,M., Pleasance,E., Prychyna,Y., Zhang,X., Jones,S. (2006) ORegAnno: An open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation *Bioinformatics*

Moses,A., Chiang,D., Pollard,D., Iyer,V., Eisen,M. (2004) Monkey: identifying conserved transcription factor binding sites in multiple alignments using a binding site-specific evolutionary model *Genome Biology* **5**:R98

Nelson,CE., Hersh,BM., Carroll,SB. (2004) The regulatory content of intergenic DNA shapes genome architecture. *Genome Bio.* **5(4)**:R25

Nüsslein-Vollhard,C., Wieschhaus,E., (1980) Mutations affecting segment number and polarity in *Drosophila Nature* **287** 795-801

Nüsslein-Vollhard,C. (1995) The identification of genes controlling development in flies and fishes *Nobel Lecture, December 8, 1995*

Ochoa-Espinosa,A., Yucel,G.,Kaplan,L., Pare,A., Pura,N., Oberstein,A., Papatsenko,D., Small.,S. (2005) The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. *PNAS* **102**:4960-4965

Ohler,U. (2006) Identification of core promoter modules in *Drosophila* and their application in accurate transcription start site prediction *Nuc. Acids Res.*

Osada,R., Zaslavsky,E., Singh,M. (2004) Comparative analysis of methods for representing and searching for transcription factor binding sites *Bioinformatics* **20** 3516-3525

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

Papatsenko,D., Levine,M. (2005) Quantitative analysis of binding motifs meditating diverse spatial readouts of the Dorsal gradient in the *Drosophila* embryo, *Proc. Natl. Acad. Sci.*, **102**, 4966-4971.

Papatsenko,D., Makeev,V., Lifanov,A., Regnier,M., Nazina,A., Desplan,C. (2002) Extraction of functional binding sites from unique regulatory regions: the *Drosophila* early developmental enhancers, *Genome Res.*, **12**, 470-481.

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

Petrov,D., Hartl,D. (1998) High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species group. *Mol. Biol. and Evol.* **15** 293-302

Pierstorff,N., Bergman,CM., Wiehe,T. (2006) Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA *Bioinformatics* **22(23)**:2858-2864

Pollard,D., Moses,A., Iyer,V., Eisen,M. Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. *BMC Bioinformatics*, **7**:376.

Powell,J. (1997) Progress and prospects in evolutionary biology: the *Drosophila* model. *Oxford University Press.*

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

Reinhold,K.(1998) Sex linkage among genes controlling sexually selected traits. *Behav. Ecol. Sociobiol.*, **44**:1-7

Siepel,A., Bejerano,G., Pedersen,J., Hinrichs,A., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L., Richards,S., Weinstock,G., Wilson,R., Gibbs,R, Kent,W., Miller,W., Haussler,D. (2005) Evolutionary conserved elements in vertebrate, insect, worm and yeast genomes. *Genome Res.*, **15**, 1034-1050.

Simon,J., Peifer,M., Bender,W., O'Connor,M., (1990) Regulatory elements of the Bithorax Complex that control expression along the Anterior-Posterior axis. *EMBO J* **9**:3945-3956

Sinha,S., Schroeder,M., Unnerstall,U., Gaul,U., Siggia,E. (2004) Cross-species comparison significantly improves genome-wide prediction of *cis*-regulatory modules in *Drosophila*. *BMC Bioinformatics*, **5**, 129.

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

Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucl. Acids Res.* **12**,505-519

Stanojevic,D., Small,S., Levine,M. (1991) Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo *Science* **254**:1385-1387

Stormo,G.D. (2000) DNA binding sites: representation and discovery *Bioinfomratics* **16(1)** 16-23

Tautz,D. (2000) Evolution of transcriptional regulation. *Curr. Opin. Genet. Dev.*, **10**, 575-579.

Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouze,P., Moreau,Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling *Bioinformatics* **17** 1113-1122

Tompa,M., Li,N., Bailey,T., Church,G., De Moor,B., Eskin,E., Favorov,A., Frith,M., Fu,Y., Kent,W., Makeev,V., Mironov,A., Noble,W., Pavesi,G., Pesole,G., Regnier,M., Simonis,N., Sinha,S., Thijs,G., V. Helden,J., Vandenbogaert,M., Weng,Z., Workman,C., Ye,C., Zhu,Z. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotech.* **23**:137-144

Wiehe,T., Gebauer-Jung,S., Mitchell.Olds,T., Guigo,R. (2001) SGP-1: prediction and validation of homologous genes based on sequence alignments *Genome Res.*, **9**, 1574-1583.

Wilson,A. (1975) Evolutionary importance of gene regulation. *Stadler Genetics Symposium*, **7**, 117-134.

Wong,G., Passey,D., Yu,J. (2001) Most of the human genome is transcribed *Genome. Res.* **11**,1975-1977

Woolfe,A., Goodson,M., Goode,D., Snell,P., McEwen,G., Vavouri,T., Smith,S., North,P., Callaway,H., Kelly,K., Walter,K., Abnizova,I., Gilks,W., Edwards,Y., Cooke,J., Elgar,G. (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLOS Biol.* **3(1)**:e7

Zavolan,M., Rajewsky,N., Socci,N., Gasterland,T., SMASHing regulatory sites in DNA by human-mouse sequence comparisons. *Proceedings of the 2003 IEEE Bioinformatics Conferene (CSB2003)* 277-286

Ich versichere, daß ich die von mir vorgelegte Dissertation selbstständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; daß diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; daß sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, daß ich eine solche Veröffentlichung vor Abschluß des Promotionsverfahren nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Thomas Wiehe betreut worden.

Teilpublikation: Pierstorff *et al.* (2006)
Pierstorff,N., Bergman,CM., Wiehe,T. (2006) Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA *Bioinformatics* **22(23)**:2858-2864

Lebenslauf

20.03.1980	in Neuss geboren Staatsangehörigkeit: deutsch
07/1986-07/1990	Grundschule Wevelinghoven
08/1990-06/1998	Gymnasium Marienberg, Neuss Abschluss: Allgemeine Hochschulreife
10/1998-11/2003	Studium der Bioinformatik an der Eberhardt-Karls-Universität, Tübingen **Diplomarbeit: Experimental study of comparative gene prediction programs** Abschluss: Diplom-Bioinformatikerin (Univ.)
01/2004	Beginn der Doktorarbeit