Ultra-Conserved Regions
Analyzing the Tip of the Iceberg

Universität Leipzig

Fakultät für Mathematik und Informatik

Institut für Informatik

Diplomarbeit

Leipzig, Januar 2008	vorgelegt von:	Sven Findeiß
	geb. am:	16.09.1982
	Studiengang:	Informatik
	Studienrichtung:	Bioinformatik

Contents

Danksagung

Ich möchte mich bei all denen bedanken, die dem Gelingen dieser Arbeit beigetragen haben.

Besonderen Dank an Prof. Peter F. Stadler und Sonja J. Prohaska für die interessante Aufgabenstellung, die vielen Diskussionen, weiterführenden Ideen und der Möglichkeit frei und selbstständig zu arbeiten.

Für die gute Zusammenarbeit vielen Dank an: Alexander Donath, Gin T. Onic, Jörg Lehmann, Kristin Reiche, Manja Marz, Stephan Bernhart, Steve Hoffmann sowie allen anderen, die mich bei dieser Arbeit unterstützt haben[1]. Außerdem möchte ich an dieser Stelle auch Petra Pregel, für die vielen gemeinsamen Cappuccinos und Jens Steuck, für jegliche administrative Hilfe danken. Natürlich auch ein großes Dankeschön an Axel Krinner für die vielen gemeinsamen und arbeitsreichen Nächte sowie einen neuen Freizeitsport.

Herzlichen Dank auch an meine Familie, Mitbewohner und Freunde dafür, dass sie mir während des Studiums immer wieder Mut und Rückhalt gegeben haben.

[1]Namen sind alphabethisch sortiert

Zusammenfassung

In den letzten Jahren wurden hoch konservierte Regionen innerhalb von Gruppen verschiedenster Spezies analysiert. Diese Diplomarbeit beschäftigt sich mit der Untersuchung von Ultra Konservierten Regionen (UCRs) innerhalb der erst vor kurzem sequenzierten zwölf *Drosophila* Genome. Eine neu entwickelte Technik, welche auf einer Seed-Extension Methode (SEM) basiert, ermöglicht es UCRs schnell und effizient zu detektieren. Unabhängig von der Anzahl der zu analysierenden Spezies kann ein multiples Sequenz Alignment (MSA) auf UCRs durchsucht werden. Im Vergleich zu anderen Methoden, welche Schnittmengen von paarweisen Alignments oder komplexe phylogenetische Modelle benutzen, ist diese hier vorgestellte Methode einfach und sehr effizient.

Die gefundenen UCRs umfassen ca. 1% des euchromatischen *Drosophila melanogaster* Genoms. Außergewöhnliche Bedingungen sind nötig um genomische Regionen über Millionen von Jahre hinweg konstant zu halten. Innerhalb dieser Studie wird die Hypothese betrachtet, dass UCRs Ergebnisse von mehreren, überlappenden Annotationen sind.

Eine Unterteilung in einmalig und mehrfach auftretende UCR-Sequenzen hilft bei der Zuordnung der jeweiligen Funktion(en). UCR-Sequenzen, die nur einmal innerhalb eines Genoms auftreten, weisen Funktionen auf, die spezifisch für diese Region sind. Wohingegen mehrfach auftretende UCR-Sequenzen bestimmte Funktionen innerhalb eines immer wieder vorkommenden Kontextes haben. Ungefähr 87% der betrachteten UCR-Sequenzen treten nur einmal innerhalb des *D. melanogaster* Genoms auf. Dies unterstützt die Annahme, dass UCRs meist kontextabhängige Funktionen haben und erschwert die Analyse eines solchen Datensatzes.

Des weiteren werden UCRs innerhalb von proteinkodierenden Sequenzen gesondert betrachtet. Speziell werden die Verbindung von UCRs und der beobachtbaren Codon Usage untersucht.

Um die Funktion von UCRs im Zusammenhang mit ncRNAs zu betrachten, werden die Sekundärstruktur und die zugehörige thermodynamische Stabilität für alle UCR-Sequenzen bestimmt und analysiert.

Durch die genaue Analyse einzelner UCRs ist es uns gelungen neue und bisher unbekannte Funktionen dieser Elemente zu finden. Dennoch bleibt der Grund für die Existenz der UCRs weitestgehend ungeklärt und es bedarf weiterer biologischer und bioinformatischer Analysen unseres Datensatzes.

Abstract

In the last years highly conserved elements were analyzed in several phyla. We have conducted a genome-wide detection of ultra-conserved regions (UCRs) within a set of twelve *Drosophila* genomes. For the efficient detection of UCRs we developed a novel technique based on a seed-extension method (SEM). This technique enables us to scan multiple-sequence alignments (MSA) containing any number of species for UCRs. It is a simple and efficient method compared to others, which use intersections of pairwise alignments or complex phylogenetic models.

The detected UCRs cover roughly 1% of the *Drosophila melanogaster* euchromatic sequences. Exceptional constraints are needed to keep genomic regions constant at the sequence level over million of years. Within this study we address the hypothesis that UCRs result from cases of multiple, overlapping sequence features.
The distinction between (non-)unique UCR sequences helps to elucidate their function. Unique occuring UCRs seem to have a region specific role whereas multiple occuring UCRs show common functions within a certain context. Approximately 87% of the UCR sequences are unique occuring motifs within the *D. melanogaster* genome. This leads to the assumption that most of the UCRs have context specific functions. Therefore, the analysis of the whole set is difficult.
Furthermore, UCRs within protein coding sequences are analyzed. Especially, the relation of ultra-conserved regions and the observed codon usage bias is examined.
Additionally, the functionality of UCRs in the context of (putative) non-coding RNA genes is analyzed. Therefore, the secondary structure and the corresponding thermodynamic stability is regarded for the complete UCR pool.

Individual case studies showed region specific features of UCRs and enabled us to find new functions of these elements. Anyway, the reasons for the observed UCRs remain largely unknown and biological as well as bioinformatic analysis of the detected UCR pool are needed.

Chapter 1

Introduction

1.1 Motivation

In computational biology comparative sequence analysis has become an essential component for the analysis of genome function. Comparative genomics takes the information of shared signatures across a set of genomes. It involves the use of computer programs that can line up genomic sequences and look for regions of (dis)similarity among them. Which regions are important depends on the subject of interest. In the case of protein-coding sequences it might be relevant to identify regions which are shared across a (sub)set of sequences, so called 'domains'. On the other hand, it is also of interest to look at the regions in-between such domains and to analyse why the observed differences occur. All these approaches work because of the evolutionary constraints assumed to keep functional regions much more conserved than non-functional ones. Therefore, conserved sequences are tend to be functional. With respect to species specific elements the reverse assumption, that functional elements have to be conserved is not mandatory. During the last years the detection and analysis of highly conserved regions more and more into the focus of interest. Several approaches with a wide range of definitions for conserved regions are available.

Usually, the term 'ultra-conserved region' (UCR) is used to define blocks of sequences which are identical in at least two sequences. Such a conserved block gains more interest if the included sequences are only distantly related. UCRs are intriguing because they have been repeatedly observed in several animal groups, associated with different functions. They have received a lot of attention [1, 2, 3, 4, 5, 6], but beyond a few anecdotal examples it is still unknown what their function(s) might be.

It is difficult to assess specific functions to UCRs due to the (nearly) perfect conservation of these elements. Ussulay, the observed variation within an alignment helps to detect and analyse sequence features. Moreover, information in the patterns of sequence change within an alignment reveals evolutionary signatures that can be used for systematic genome wide annotations. Functional elements that tolerate 'silent' mutations, such as protein-coding exons, non-coding RNAs and regulatory motifs are detectable using feature specific signatures.

For instance, protein-coding sequences are subsumed of nucleotide triplets (codons) at the DNA level. The first and second nucleotide of such a codon is usually constant within orthologous sequences but the third position varies. These information of constant (first and second) and varible (third) codon-positions are used to detect and analyze protein-coding sequences.

The (nearly) perfect conservation and therefore the missing variation within UCRs makes it difficult to assess specific functions to UCRs not only in the context of protein-coding sequences.

This thesis is based on *Drosophila melanogaster* data and the eleven recently sequenced *Drosophilids*. The phylogenetic relation and the grouping of the twelve *Drosophilids* is shown in 1.1. At the moment, the analysis of these species is one of the hottest topics in (computational) biology. A paper about comparative genomic approaches to study the evolution of genes and genomes on the *Drosophila* phylogeny [8] was recently published in Nature. This work was done by several research groups around the world. *D. melanogaster* is a well studied model organism with a compact, completely sequenced genome, and well established physical and functional maps [9, 10]. One of the main reasons for this is the high analogy of regulatory-mechanisms and orthologous genes in *D. melanogaster* and human. Due to an eleven times faster rate of evolution the sequence divergence within the *melanogaster* subgroup (see Figure 1.1) is equivalent to the divergence among mammals. Therefore, the twelve species and their subgroups are an ideal model system for genomic scale sequence comparisons.

Given a definition for UCRs, it would be more interesting to observe such an UCR element e.g. between *Drosophila simulans* and *Drosophila grimshawi* than between

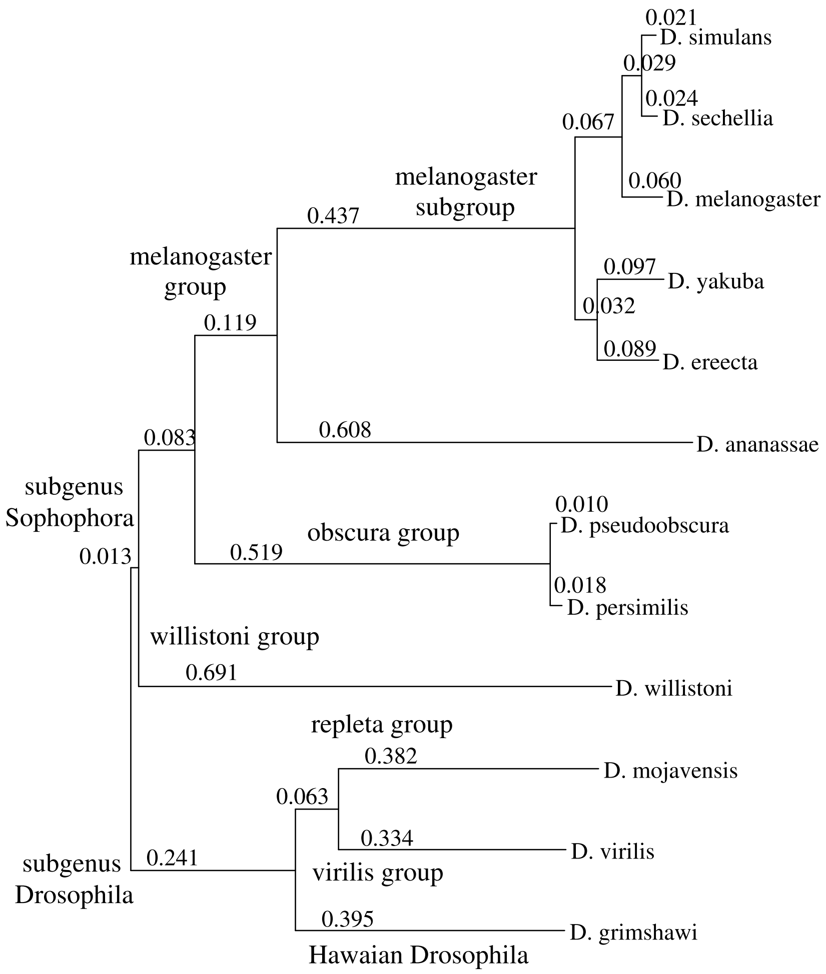


Figure 1.1: Tree of divergence for the twelve *Drosophilids*. The relative distances are calculated from mutations at four-fold degenerated sites (cp. [7]).

Drosophila simulans and *Drosophila sechellia*. Purifying selection is assumed to keep this UCR constant at the sequence level. Within this thesis we analyze the tip of the iceberg since the detection procedure is based on multiple-sequence alignments (MSAs) and restricts UCRs to be conserved between all twelve *Drosophilids*.

1.2 Subject of this thesis

Several publications which address the detection and analysis of conserved regions between two or more species are available (non exhaustive summary see Appendix A.1).

In this thesis a novel method to detect ultra-conserved regions (UCRs) in MSAs will be described (see Chapter 3). Due to the fact that there is no exact definition for UCRs in the literature, a first task is to find such a definition. This has to be done with respect to the varying quality of the current twelve *Drosophila* assemblies. Given a definition, we try to analyze reasons for the existence of the UCR set among all *Drosophilids*. Is there another feature than the conservation which connects the whole set or is it possible to describe subsets with common features?

The aim of this assessment is to find reasons **why** UCRs exist. To elucidate the function/features of UCRs, Chapter 5 addresses four hypotheses. In Section 5.2 UCRs within protein-coding sequences are analyzed. Especially, the codon-usage within UCRs is compared to the codon-usage over all protein-coding sequences.
The reported functionality of highly conserved elements in the context of putative RNA genes was mentioned in different studies before [1, 11, 12]. Within Section 5.3, the secondary structures and the corresponding thermodynamical stability for the complete UCR pool is regarded. Predictions of non-coding RNAs [13] gets more an more into the focus of UCR research, due to the fact that small RNA molecules (e.g miRNA, snRNA) are known as regulatory elements.
The distinction into (non-)unique UCR sequences might help to elucidate the function of these elements. Unique occuring UCR sequences might have a region specific function. If a UCR sequence occure several times within one genome it might act as a regulatory element (e.g. transcription factor binding site) or a common (sub-)element of a sequence features (e.g. (un-)structured sub-sequence of an non-coding RNA). The uniqueness of all UCR sequences within the genome of *D. melanogaster* is addressed in Section 5.4.
Finally the hypothesis that UCRs are conserved because of several overlapping sequence features is addressed in Section 5.5.

Chapter 2

Background

During the last years more and more genomes were completely sequenced. With the onset of new techniques the sequencing time for a complete genome decreases, while the accuracy increases. A large sample of whole genome alignments with different species compositions is available for almost each of the model organisms. Despite this tremendous progress, it is still not clear how many bases of a genome are of functional importance. Regulatory proteins as well as non-coding RNAs are known to influence transcription and translation, initiate replication, altering chromatin structure and regulate several other cellular processes.

One strategy to find functional sequences within a genome is to search for conserved regions in at least two species. Typically, models of molecular evolution consider homogeneous mutation rates along the genome. The observed differences in sequence divergence are attributed to different substitution and fixation rates in functionally conserved regions. Assuming, that mutations in functional conserved regions are deleterious and therefore they are rapidly eliminated. Within two related species the observed fraction of conserved regions might also be caused by reduced mutation rates. Andrew G. Clark [14] termed such regions 'mutational coldspots'. Based on a pairwise species comparison of conserved non-coding sequences Bergman et al. [15] showed that these regions are clustered within the *D. melanogaster* genome. Furthermore the length of spacer between such conserved non-coding elements is not random as expected for unconstrained regions and random evolution. These findings indicate region specific constraints on the conserved non-coding regions. Also, the works of Retelska et al. [3] and Drake et al. [16] addressed and disproved the hypothesis of 'mutational coldspots'. Both studies used single nucleotide polymorphism (SNP) data of humans to analyze the occurring mutations within conserved non-coding regions. The variation of a single nucleotide within DNA fragments between members of the same species is a so-termed SNP. If conserved non-coding regions are simple 'mutational coldspots' the percentage of SNPs should be significantly lower than in constraint regions. The studies showed similar SNP occurrences within both types of regions and therefore disprove the hypothesis of 'mutational coldspots'.

However, negative (purifying) selection is assumed as the primary reason for cross-species conservation [15, 16]. Specifically, evidence that UCRs are ultra-selected is reported [17]. If a genomic region is under negative selection it is assumed to be of functional importance in the composition as it is observed. Approximately 1.5% of the human genome is postulated to code for proteins and about 5% to be under purifying selection [18, 19]. If 5% of the human genome are under purifying selection and therefrom 1.5% protein-coding the non protein-coding proportion has to account for approximately 70% $\left(\frac{0.05-0.015}{0.05} \times 100 = 70\%\right)$. Novel techniques to detect functional regions within the 'dark matter of the genome' (the remaining non protein-coding 70%) [12, 20] are needed.

Usually, the analysis of genomic sequences of at least two species starts with an alignment. A sequence alignment is a way of arranging the primary sequences of either DNA, RNA, or protein to identify regions of similarity that maybe a consequence of functional, structural, or evolutionary relationship between the sequences.

Pairwise sequence alignment methods are used to find the best-matching local or global alignments of two sequences. For the calculation of pairwise alignments several algorithms are established. To produce local alignments via heuristic implementations of the Smith-Waterman algorithm [21] or global alignments via heuristic implementations of the Needleman-Wunsch algorithm [22] is a common task in the field of bioinformatics. Within a pairwise alignment a conservation pattern can be defined (according to [2]) through an ordered set S of gaps, mismatches and ungapped conserved blocks and their corresponding lengths. A region within S is 'conserved' if both sequences share a minimum pairwise identity over a certain length. Shin et al. [23] for example define evolutionary conserved regions by a minimum length of 70nt and 80% sequence identity between two species.

Multiple sequence alignments (MSAs) are an extension of pairwise alignments to incorporate the information of more than two sequences at a time. In the last years several tools to compute whole genome MSAs have been developed. They are mostly based either on orthologous maps, like `PECAN`[1] and `MAVID` [24, 25], or they use phylogenetic trees, as the `threaded blockset aligner (TBA)` [26] does. Due to missing information about the phylogeny between species and the fact that orthologous maps are often based on protein coding sequences these approaches might be biased. Therefor, `NcDNAlign` (submitted [27]) represents a third class of alignment tools. Only the genomic sequence of each species is needed. Without any additional information (e.g. precomputed maps or phylogeny) it is possible to calculate the MSA. MSAs are often used to identify conserved regions across a group of sequences [11, 12, 16].

In general pairwise alignments are more convenient to analyse statistics of conservation patterns within different functional sequences [2] than MSAs are. However, Margulies et al. [28] describe the differences between pairwise and multiple-sequence alignment methods. They used a set of 31 vertebrate genomes and aligned them stepwise with `TBA` [26]. Specifically, alignments were generated using the human and platypus sequences and then an increasing number of additional species were added, one at a time, based on the increasing phylogenetic distance to human. In each step they extracted the human-platypus sequences and compared their alignability to the pairwise alignment of human and platypus generated with `BLASTz` [29]. Interestingly, an extracted alignment covers 79-115% more sequences than the corresponding pairwise alignments do. This percentage increases for more distantly related species. Margulies et al. suspect that this is due to the presence of additional sequences broadly distributed across the phylogenetic tree. Therefore the evidence of an enhanced statistical power to align regions, which are too diverged to be pairwise aligned, is given. This advantage of MSA will help to find short regulatory elements which might be missed within pairwise alignments.

[1] `http://www.ebi.ac.uk/~bjp/pecan/`

Obviously, is it easier to define a conserved region as a block of minimum length n and percent identity of 95% in pairwise sequence alignments than in MSAs. A block with 100% conservation and a defined length is also easy to detect within a MSA. To get a definition for less conserved regions (e.g. mean pairwise identity of 95%) is difficult with respect to MSAs. For instance Siepel et al. [12] used a phylogenetic Hidden Markov Model (phylo-HMM) to calculate length dependent log-odds scores for each column in a MSA with respect to a reference sequence. Two independent models have to be estimated from the MSA to calculate such a log-odds score. Usually, a model for a conserved state and a second one for the unconstrained background is used. Such a score reflects how much more likely an alignment column is under the conserved state than under the non-conserved. Using an already scored MSA and scanning for blocks which have log-odds scores above a given threshold helps to identify conserved regions.

In this diploma thesis ultra-conserved regions (UCRs) among the recently sequenced *Drosophilids* are analyzed. To define UCRs in general is problematic due to many similar terms for conserved regions and the usage of different length and conservation constraints. Additional terms for conserved regions are highly conserved elements (HCE) [12] and ultra-conserved elements (UCE) [5]. In the context of non-coding regions additional terms are also common. For instance, conserved non-coding regions (CNC) [3] are labeled as conserved non-genic (CNG) regions [30] and conserved non-coding elements (CNE) [31]. Several approaches are described in the literature to detect these elements.

Nevertheless, to divide these approaches in two classes one can consider the alignment input as a criterion. On the one hand pairwise alignments are used (e.g. [2, 6, 30]) and on the other hand MSAs (e.g. [11, 12]). In MSAs a definition is not as straightforward as in pairwise sequence alignments where a minimum length and percent identity, are sufficient to define UCRs.

The approach presented here is based on a seed-extension method (SEM) to detect blocks with a minor conservation within MSAs. SEM means that the MSA is scanned for perfectly conserved blocks of specified minimum length l in a first step of the UCR detection procedure. This block is considered as a seed of length l. These seeds are extended in both direction separately up to a minimum conservation of 95% regarding to the consensus sequence of the alignment (see Chapter 3). This novel technique enables us to scan MSAs containing any number of species for UCRs. It is a simple and efficient approach compared to methods which use intersections of pairwise alignments [4, 32] or complex phylogenetic models [12].

There is also a wide range of specified minimum lengths for conserved regions. Elements from 10nt [15] up to more than 200nt [5, 31] are analysed. Which minimum length is considered depends on the input alignments and the way to analyse the detected conserved regions. A non exhaustive summary of different definitions with the considered measurements of conservation and the minimum length is given in Appendix A.1.

In the last years highly conserved elements were analyzed from yeasts [12], worms [12, 33] and insects [1, 2, 4, 12, 15, 32, 34] as well as vertebrates [3, 5, 6, 12, 11, 16]. It is commonly known that short non-coding DNA fragments which are highly conserved in vertebrates are absent from *Drosophila* genomes [3, 30, 31]. Interestingly, near orthologous genes in worms, insects and vertebrates a core set of conserved non-coding elements is reported [33]. Since many of the *D. melanogaster* protein-coding genes have orthologous loci within the human genome these findings indicate a context specific usage of highly conserved elements and their presents or absents within other species.

Nevertheless, in general highly conserved elements show similar functions in all phyla. For instance, Sandelin et al. [6] describe that developmental genes in vertebrates are surrounded by many non-coding UCRs. Furthermore, a strong positional correlation between UCRs and transcription factor coding genes is noted. This association extends up to hundreds of kilo bases (kb) from the transcription start site of the gene (e.g. UCR enrichment starts 300kb in front of the homeobox genes in human). This linkage between highly conserved elements and transcription factor coding genes as well as developmental regulators is reported in the literature [1, 5, 32, 35] many times. A known function of such elements is to work as (distant) regulators of (developmental) genes [31, 36]. Also, *in vivo* characterization of specific highly conserved regions revealed this role as regulatory element [15, 23, 37, 38]. Bergman et al. [15], for example showed that conserved non-coding sequence clusters are suitable to predict developmental regulated enhancers in the apterous region (*ap*). The LIM-homeobox transcription factor *ap* is expressed in many tissues of *D. melanogaster* [39]. Engström et al. [32] conclude that clusters of highly conserved elements play a major role in the context of developmental regulation in most, if not all, Metazoa.

In contrast, Ahituv et al. [40] investigated the necessity of UCRs *in vivo* by removing four non-coding ultra-conserved elements from the mouse genome. To maximize the likelihood of phenotypic effects they deleted elements that function as enhancers in a mouse transgenic assay and that are near genes which are known to show effects on the phenotype if they are up or down regulated. Despite the careful selection of the removed UCRs the mutated mice were viable and fertile. They failed to reveal any critical abnormalities for a variety of phenotypes including growth, longevity, pathology, and metabolism. Of course, the deletion of UCRs must not show phenotypic effects but this study indicates how difficult it is to assess the function of UCRs. In addition it should be noted that publications of failed experiments are as important as successful ones.

UCRs are also reported in the context of alternative splicing and nonsense-mediated decay (NMD) [5, 41]. Specifically Ni et al. [41] describe a class of highly conserved stop-codon containing exons. These exons contain stop codons in all three reading frames. Integration of such a special exon into the transcript is leading to NMD. These types of exons frequently occur within genes which encode for splicing activators and are enriched in UCRs.

A striking example of intron retention was discovered by Glazov et al. [1, 42] within the homothorax (*hth*) gene locus of *Drosophila*. They showed that an UCR element which covers an intron-exon junction is able to form a stable stem-loop structure. This stable structure seems to be the reason for intron retention within *hth* transcripts. A second UCR element, located closely to the first one, shows functional importance as a *cis*-regulator and thereby prevents transcripts with retained introns. Glazov et al. report the complementary sequences of both UCRs which could form an alternative secondary structure to the one described above. The switch between two stable secondary structures might regulate intron retention. If the transcript folds into the structure which covers the intron-exon junction the result is intron retention. Otherwise, a base-pairing between the two complementary UCR elements leaves the intron-exon junction open for splicing.

Another class of functional important UCRs is associated with several non-coding RNAs. Beside studies which search for highly conserved non-coding RNA like sequences in general [1, 11, 12, 20] more precised analysis [4, 43, 44] reveal several functional UCRs.

Within the human genome it is known that an ultra-conserved region is transcribed to generate *Evf-2*, an alternative splice form of the non-coding gene *Evf-1* [43]. Both transcripts are located in the *Dlx* gene loci. The six *Dlx* genes encode for transcription factors which are arranged in three convergently transcribed clusters [45]. *Evf-2* was shown to cooperate with the homeodomain protein *Dlx-2* to increase the activity of the *Dlx-5/6* enhancer as a target. This suggest the possibility that UCRs can act as functional *cis*- as well as *trans*-regulatory elements [43].

Furthermore, UCRs were also found to encode for non-coding RNAs in the context of human leukemia and carcinomas [44]. The blepharophimosis syndrome seems to be caused by conserved non-coding sequences [46]. This indicates that UCRs play also a functional role in different diseases. On the other hand a study about autism [47] and another one analyzing multiple sclerosis [48] revealed no evidence that variations of yet known UCRs influence this diseases.

All these findings indicate a broad functional range of UCRs. Nevertheless, beyond a few anecdotal examples the reason for their existence is still unknown.

Usually, feature specific variation helps to elucidate the possible function of elements. Moreover, information in the patterns of sequence change within an alignment reveals evolutionary signatures that can be used for systematic genome wide annotations. Protein-coding regions show highly constrained codon substitutions. These regions tolerate mutations that preserve the amino acid translation. Therefore, the information of silent nucleotide substitutions at the third codon position and the usually constant first and second codon position are used to detect protein-coding regions. Tolerated substitutions within non-coding RNA genes are known to preserve base pairings and structures (e.g. compensatory mutations) [49]. For instance a characteristic conservation profile (conserved stem and variable loop structure) for microRNAs is used to detect this type of non-coding RNAs. Finally, regulatory motifs tolerate a local movement and nucleotide substitutions consistent with degeneracy patterns [50, 51].

Especially, position weight matrices (PWM) are used to detect regulatory motifs. A PWM contains scores which give a weighted match to any sub-string of a fixed length. However, it is a difficult task to assess functions due to the (nearly) perfect conservation within UCRs.

Testing of novel comparative genomic approaches is usually done within a model system. The set of twelve closely and distantly related *Drosophilids* provides an ideal model for calculating and accelerating comparative genomic methods. The phylogeny and evolutionary genetics of the twelve *Drosophilids* form a well known range of divergence times [7, 8, 52]. As described by Richards et al. [53], significant changes in the genomic sequences (71% identity in alignable regions, covering approximately 40% of the genome) are observable between *D. melanogaster* and *D. pseudoobscura*. These changes are comparable to those between humans and rodents (69% identity in alignable regions, covering 40% of the genome) [18], which diverged earlier, but have longer generation times than insects. Another way to compare the divergence between *Drosophilids* and vertebrates is based on the analysis of four-fold degenerated sites. The third codon position of amino acids which are encoded by four codons is termed four-fold degenerated sites. The corresponding codons of such an amino acid are constant at the first and second position. The third codon position is variable and could show any nucleotide (Adenin, Guanin, Cytosin, Thymin). Assuming, that a mutation at the third codon position has no effect on the process of transcription and translation mutations without functional changes are observed. Using four-fold degenerated sites pairwise and multi-species distances are estimatable for a set of species (for review see [7, 52]). Stark et al. [52] used `PHYML v2.4.4` [54] to compare pairwise as well as multi-species distances (in substitutions per four-fold degenerate site) of *Drosophilids* (reference genome *D. melanogaster*) and vertebrates (reference genome human). Figure 2.1 shows the comparison of both distances measurements with respect to *D. melanogaster* and human. The depicted pairwise distances for flies are comparable to those shown in Figure 1.1. With respect to the multi-species distances the estimated sequence divergence of the twelve *Drosophilids* is comparable to mammals.

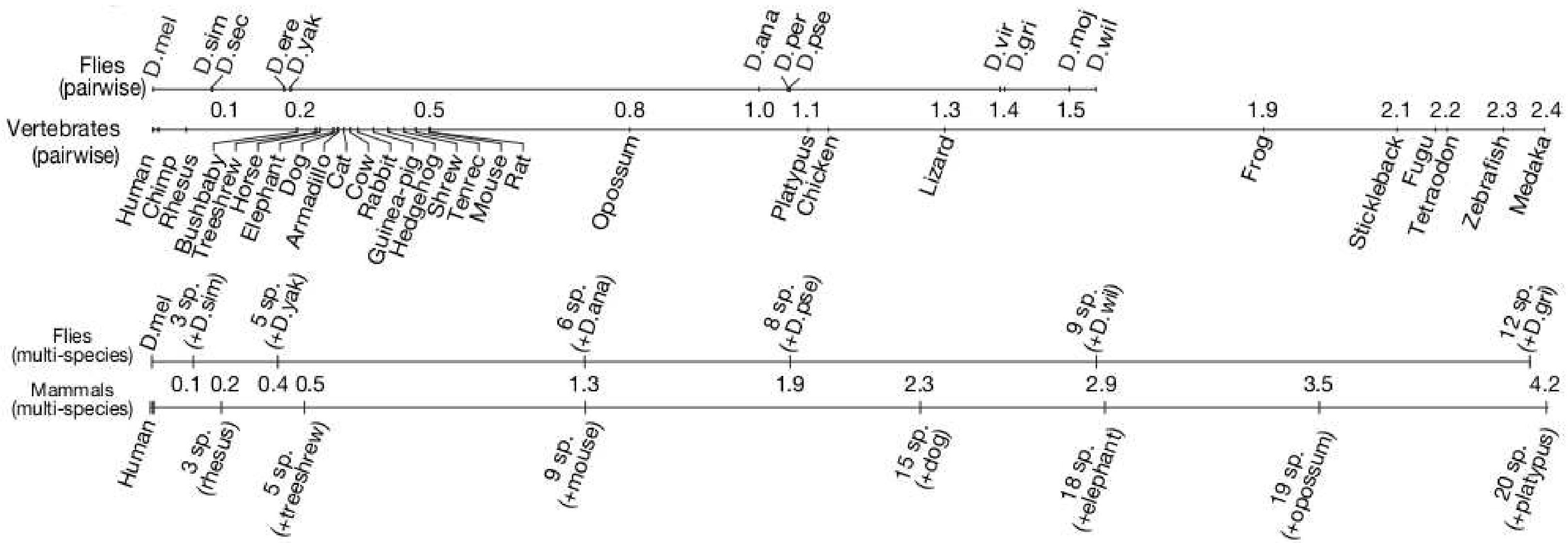


Figure 2.1: Pairwise and multi-species distances (in substitutions per four-fold degenerate site) are depicted with respect to *D. melanogaster* and human as reference genomes. As noted by Stark et al. species with longer branches (e.g. human-mouse) show higher pairwise distances which do not always reflect the order of divergence. In contrast the multi-species distances include all species within a phylogenetic clade (e.g. human-mouse distance is estimated using nine species). Using the multi-species distances the sequence divergence of the twelve *Drosophilids* is comparable to mammals. (Figure is copied from Stark et al. [52])

Large scale analysis, as for instance the mammalian ENCODE project [55], describe the necessity of novel comparative genomic approaches. The development of innovative tools and algorithms based on MSAs is needed. Due to the large mammalian genomes it will take a number of years before more than twelve genomes of the mammalian clade with an almost complete coverage are available. This causes the *Drosophilids* to be an ideal model system for genomic scale sequence comparisons with implications to comparative genomics among vertebrates. Smaller genomes with a high degree of similar features to the mammalian genome are the fundament for testing these approaches. Each of the compact and feature-rich *Drosophila* genomes is $\sim 1/30^{th}$ in size compared to the human and mouse genomes [2]. This means that they are less expensive in terms of computation-time for comparative genomic approaches. Additionally, their comparative importance is given by the fact that the the majority of human disease-associated genes have orthologous loci within the *Drosophila* genome [56]. Using a variety of *Drosophilids* also provides the ability to study and analyze the function and interaction of pathways shared by *Drosophila* and human. The complete sequencing of the euchromatic proportion of the *D. melanogaster* genome [9, 57] facilitates additional comparative genomic approaches as well.
The active fly community in both biology and bioinformatic research is spread all over the world and has established databases (e.g. Flybase [58, 59], Flyexpress [60]) to share information and annotations. In addition, the small genome size as well as the established physical and functional maps [9, 10] for *Drosophila melanogaster* are supporting aspects. *Drosophila* has long served as a model system for developing and testing evolutionary principles at morphological and cytological level. Detailed analysis of the evolutionary principles on DNA and protein level support the choice of the *Drosophila* genus. Furthermore, the biological tools for the investigation of *Drosophila* genetics are well tuned. This makes a direct biological validation of innovative bioinformatic approaches possible. The choice of using the *Drosophilids* as model system is also supported by the rapid and cost-effective transgenic system which can be adapted for mutation studies.

[2] `http://insects.eugenes.org/species/about/genome-proposals/GenomesWP2003.pdf`

The analysis of UCRs in the *Drosophila* genus will show which genomic regions are ultra-selected within all twelve *Drosophilids*. It is essential to identify the constraints which keep such an UCR element constant at the sequence level. If we are able to elucidate these constraints for insects the implication of similar mechanisms can be drawn for vertebrates.

Chapter 3

UCR detection

Detection of not perfectly conserved blocks within an multiple-sequence alignment (MSA) is a difficult task due to a missing conservation measurement. The novel technique presented here is based on a three step seed-extension method (SEM). SEM means that the MSA is scanned for perfectly conserved blocks of a specified minimum length l in a first step. This block will further be considered as a seed. In a second step these seeds are extended to ultra-conserved regions (UCRs) until the conservation falls below a given threshold regarding to the consensus sequence of the alignment. Fusion of overlapping UCR candidates is the third and last step within this procedure. The next three sections give detailed descriptions of this method.

3.1 Scanning for seeds

An UCR is defined as a block of totally conserved columns which are connected by a few less conserved columns in between. If a MSA is scanned for such a block the process of detection is direction dependent. On the one hand one never knows at which position it is needed to allow less conserved columns in order to reach a totally conserved block. On the other hand if a totally conserved block is observed it might be possible to include previously seen alignment columns. To prevent this direction dependency the first step of the UCR detection process is to scan for so-called 'seeds'. A seed is defined as a block of fully conserved ungapped sequences within a MSA (see Figure 3.1). The input alignment is scanned from left to right for columns with only one character (not equal to the gap character) out of alphabet $\mathbb{A}$ among all species. The alphabet $\mathbb{A}$ contains the four DNA nucleotides and the gap character ($\mathbb{A} = \{A, C, T, G\} \cup \{-\}$). If at least l of these totally conserved columns are detected within a block of contiguous sequences, this block is considered as a seed of length l.

3.2 Extending the seeds

With respect to sequencing errors, e.g. doubling or missing of a nucleotide (see [61]) it is useful to permit a few columns with a conservation of less than 100% in the alignment block. The aim of this second step is to extend a detected seed of length l by some less conserved columns and therefore to reach additional fully conserved columns. Two adjacent blocks of 100% conservation which are separated by these less conserved columns are combined within this step.

Since we are searching for UCRs, the extension process should not accept all types of less conserved alignment columns. The procedure is therefore limited a) to columns containing not more than two different types of characters and b) to prevent fuzzy ends by the restriction that the last column has to be fully conserved. At this point a measurement of conservation for extended MSA blocks is needed.

A first intuitive approach might be to calculate the conservation dependent on the number of sequences, the length of the current extended seed and the number of mismatches in all columns of the according alignment. Formula 3.1 represents a way to calculate such a conservation for a seed of length l, the current extension of length l_e, the total number of mismatches m, and the number of sequences n involved in the alignment. The total number of mismatches is the sum over all sequence positions which are unequal to the consensus.

$$\text{conservation} = \frac{n(l + l_e) - m}{n(l + l_e)} \times 100 \tag{3.1}$$

The consensus sequence of a MSA represents the MSA as a single sequence by displaying only the most common character for each alignment position. Figure 3.1 displays an example for a detected seed of length $l = 21$ and an extension to the right of length $l_e = 11$. In this alignment $m = 8$ and the calculated conservation for the extended seed is 98%.

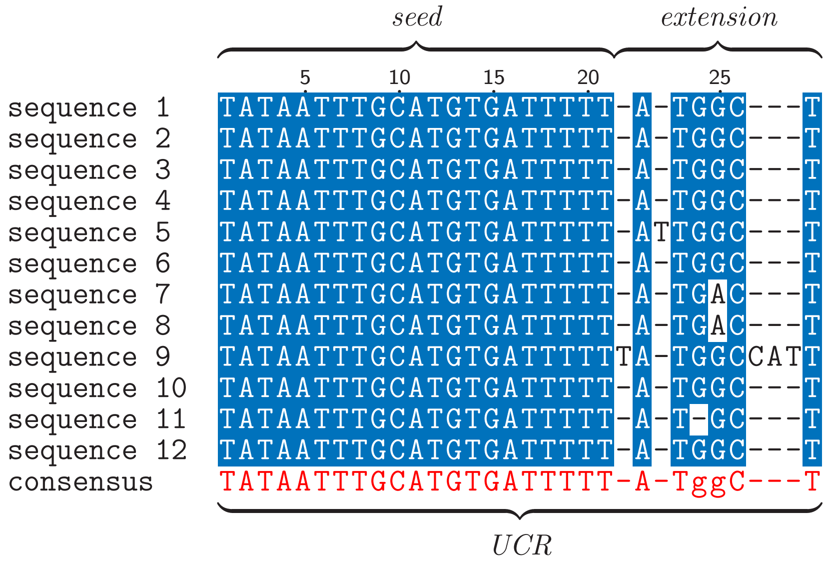


Figure 3.1: Example alignment with seed length 21, number of mismatches 8 and the length of the extension 11. The conservation value (calculated by Formula 3.1) 98%.

It is easy to see that mismatches accumulate across columns and rows. Due to the number of fully conserved columns and the current count of mismatches the extension might be fuzzy. It is possible that long stretches of less conserved columns are included and the resulting alignment region could no longer be treated as ultra-conserved.

Another way to get a conservation measurement for the extension is to separate the accumulation of columns and rows and use the minimal conserved sequence as a reference for the whole block. For this purpose the conservation between each sequence and the consensus is calculated using Formula 3.2. Where i is the i-th sequence in the alignment and the function $\mathrm{mismatches}(i, c)$ returns the number of mismatches between sequence i and the consensus sequence c.

$$\mathrm{conservation}(i) = \left(1 - \frac{\mathrm{mismatches}(i, c)}{l + l_e}\right) \times 100 \tag{3.2}$$

The extension stops if one of the involved sequences falls below a specified conservation threshold. This measurement is more strict than the first approach and the resolving alignment is more conserved over all columns. Figure 3.2 displays the resulting example alignment if 95% is used as the conservation threshold. 'sequence 9' falls below this threshold if one column more is added to the block (cp. Figure 3.1). This is a fair

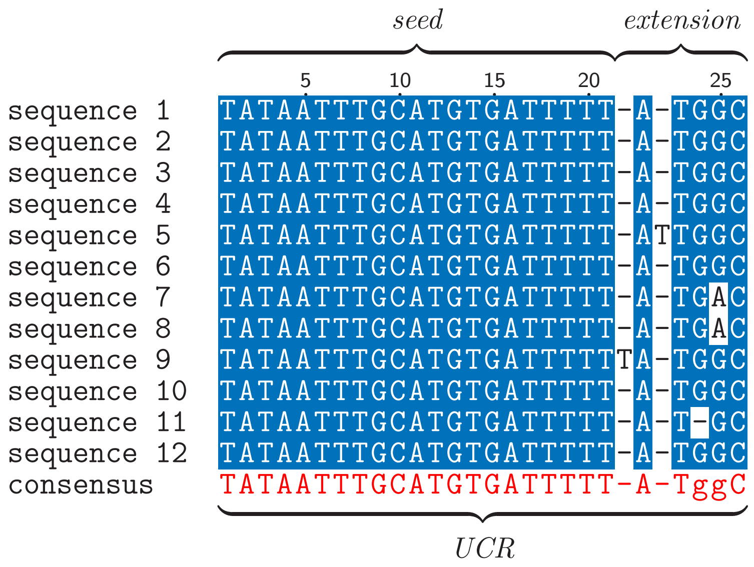


Figure 3.2: Example alignment with seed length of 21 and extension length of 7. The conservation calculated by Formula 3.2 is 96%.

measurement because of the fact that only a few less conserved columns are permitted and if one sequence diverge from the alignment the procedure stops.

Each seed is extended separately in both directions with the conservation measurement given by Formula 3.2. A set of UCRs, where the elements can overlap, is the outcome of this step. The next and last step of the UCR detection procedure addresses this redundancy.

3.3 Fusion of overlapping UCRs

The seeds which were detected by the first step are non overlapping, fully conserved blocks. After extending each seed in both directions it is possible that adjacent UCRs overlap.

Two kinds of overlaps are possible. The first one happens if it is possible to extend a seed along less conserved alignment columns and to reach another seed region. In fact, the seed reached is a block of fully conserved alignment columns and will be included into the extension. It is not necessary that both UCRs reach each other. If the detected seeds differ in length it is common that the longer one includes the seed region of the

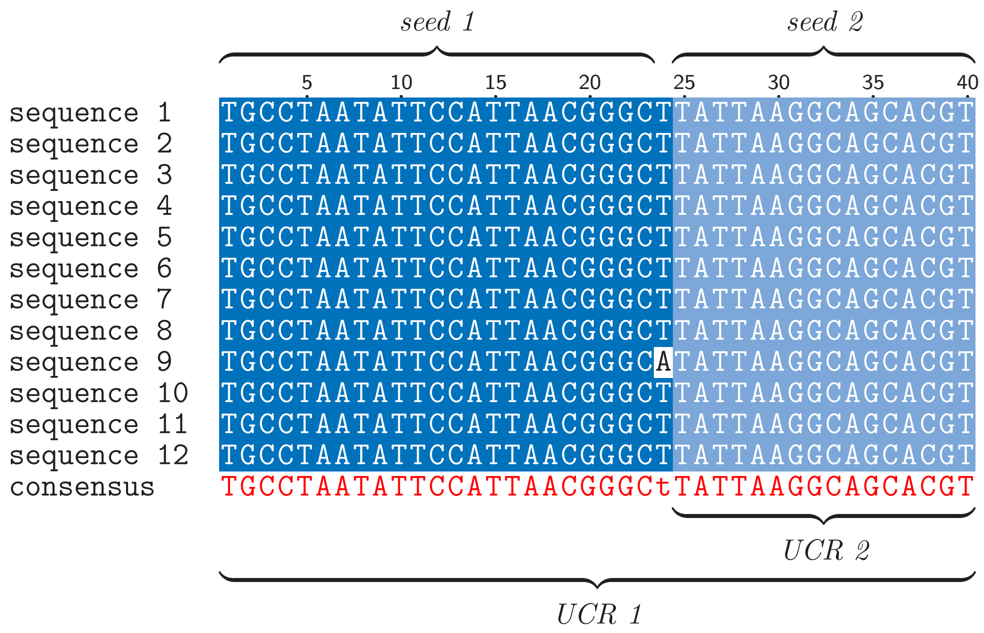


Figure 3.3: UCR 1 and UCR 2 overlap in the shaded region. If the conservation threshold is set to 95% *seed* 1 is able to extend through the less conserved column 24. On the other hand *seed* 2 stop the extension procedure in front of this column. If the threshold is lower than 95%, both UCRs would completely include each other.

shorter one but not the other way around. Figure 3.3 displays an example for such an overlap where the conservation threshold is set to 95%. Seed 1 represents the longer one and is able to extend through alignment column 24 (uncolored 'A') while seed 2 is not able to include this column and the extension process stops beforehand.

The second case appears if two UCRs overlap in their extensions only. This happens if both extensions were stopped at a less conserved alignment column before they reach the other seed. Figure 3.4 displays an example for this kind of overlap. The depicted UCR 1 and 2 overlap in the shaded region.

All types of overlap are treated by calculating the union of both UCR regions. Given two overlapping UCRs each defined by a start and an end point (start $<$ end, relative to the input alignment), the union goes from the lowest start to the largest end point. In the example of Figure 3.4 the resulting UCR would span from alignment position 1 to 50. For the sake of clarity, both UCRs are displayed with an extension in just one direction. Of course an extension in the other direction could also appear and in this case the resulting UCR would include this extended region as well.

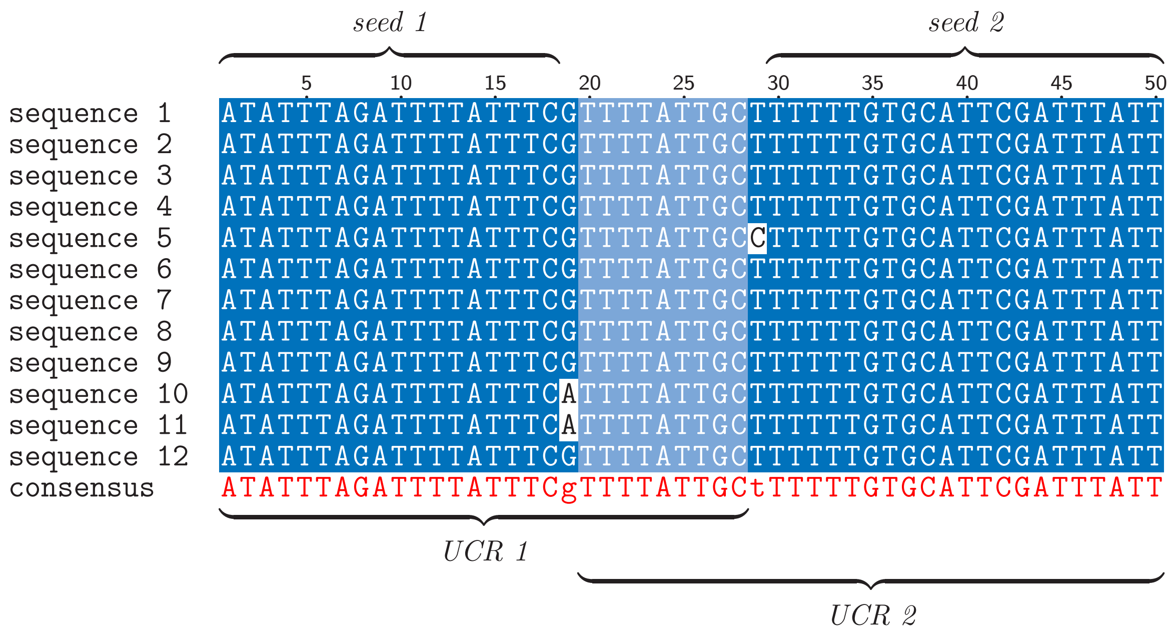


Figure 3.4: Example where both seeds are able to extend along one less conserved alignment column but stop at the next (assumptive conservation threshold 95%). The shaded region marks their overlap.

Chapter 4

Dataset

4.1 Materials

Genomes and Alignments

Two multiple whole-genome alignments were downloaded from `AAA-fly Wiki`[1]. The `PECAN`[2] alignments are available at `http://www.sanger.ac.uk/Users/td2/pecan-CAF1/` and the `MAVID` [24, 25] alignments used at `http://www.biostat.wisc.edu/~cdewey/fly_CAF1/`. For further information on how these alignments were generated see the descriptions at the corresponding web pages. Both alignments are generated on basis of the drosophilid genome sequences available at `CAF1 Freeze homepage`[3]. The `MAVID` alignments contain heterochromtic *Drosophila melanogaster* sequences which are not included in `CAF1 Freeze`. In Rose et al. [13] it is also noted that `MAVID` alignments seem to be biased towards protein coding regions. Due to this fact the `PECAN` alignments are preferred. Following the comments on their web page the reconciled sequences are excluded. The unassigned sequences (chrU) are not mentioned in the `PECAN` alignments and by that it is also excluded from further analysis.

[1] http://rana.lbl.gov/drosophila/wiki/index.php
[2] `http://www.ebi.ac.uk/~bjp/pecan/`
[3] http://rana.lbl.gov/drosophila/caf1.html

Annotations

Annotations for CDS, 5' and 3' UTR, gene, intergenic region, intron, pseudogene, transcript (mRNA), translation (protein), transposable element, transposable element insertion site, enhancer, poly-A site, protein binding site, regulatory region, miRNA, ncRNA, tRNA, and miscRNA were downloaded from the `FlyBase Release 4.3` [59, 62, 58]. Additional annotations for ESTs, repeats (`RepeatMasker` [63], `Tandem Repeat Finder` [64]), mRNA, and RefSeq mRNA were downloaded from the `UCSC Genome Browser` [65, 66] [4]. The recently published Transfrag dataset [67] was downloaded from the affymetrix transcriptome webpage[5]. For additional regulatory element information annotated cis-regulatory modules (CRMs) and transcription factor binding sites (TFBSs) were downloaded from `REDfly database version 2.0`[6]. After inspection of the available annotations the subset summarized in Table 4.1 is used for further analysis.

4.2 Parameter estimation

For the procedure described in Chapter 3 two main parameters have to be estimated. These are the minimum length l of a seed and the conservation threshold c. In addition the influence of a parameter u which defines the number of characters unequal to the consensus will be considered.

Minimum seed length (l):
Determining the minimum length of a seed is a non trivial task. No measurement to estimate the minimum length of a significant block within a MSA is available. There is no way to calculate the probability that a block of length l is observed by chance. The main problem is that a conserved block can not be seen as an uncoupled event within a MSA. Several factors, such as the algorithm used to create the alignment and the context in which a block of length l is observed influences such a measurement.

[4] `ftp://hgdownload.cse.ucsc.edu/goldenPath/dm2/database`
[5] `http://transcriptome.affymetrix.com/download/publication/dros_dev/transfrags/`
[6] http://www.flyreg.org/

Table 4.1: List of used annotations for all further analyses. The column 'Number' displays the total count of entries for each type of annotation. The entries can overlap with each other. Small RNAs is hereby the summarized set of all known rRNAs, tRNAs, snRNAs, snoRNAs, miRNAs and ncRNAs.

Annotation	**Source**	**Number**
CDS	`Flybase Release 4.3`	85778
intron	`Flybase Release 4.3`	50980
five prime UTR	`Flybase Release 4.3`	18900
three prime UTR	`Flybase Release 4.3`	13561
small RNAs	`Flybase Release 4.3`	679
transposons	`Flybase Release 4.3`	6003
Repeat Masker	`UCSC` dm2	74418
ESTs	`UCSC` dm2	608028
TransFrags	Manak 2006	81331
TFBS	`REDfly database version 2.0`	1345
CRM	`REDfly database version 2.0`	665

Nevertheless, two different approaches can be used to estimate the minimum seed length l.

The first is based on a single sequence measurement. Given a DNA sequence S of length N. An estimation for the mimimum length l of a sub-string S' to be observed at least once in S is calculatable using Formula 4.1 [68].

$$1 \geq N \times \left(\frac{1}{4}\right)^{l} \tag{4.1}$$

The term $\frac{1}{4}$ stands for the nucleotide frequency if it is assumed that all four nucleotides appear equally distributed in the genome. For N, the shortest sequence or genome in the dataset, is chosen because of the fact that sub-sequence S' has to be in each of the aligned sequences. In our dataset the genome of *Drosophila melanogaster* with approximately 118 million bases is the shortest one. Assuming a fix $N = 118$ million bases, the value of the exponent l is 14 to get the Formula true. Occurrence of certain sequence is not expected by chance. This single sequence measurement gives a lower bound for the parameter l.

The second approach is to use an already existing MSA. Given a MSA we are able

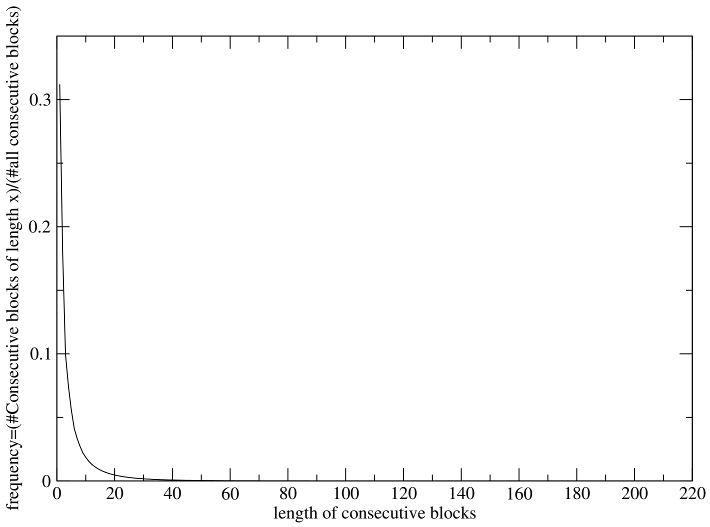


Figure 4.1: This Figure displays the distribution of possible consecutive blocks for each length x. The y values are normalized by the total number of consecutive blocks.

to extract the number of fully conserved blocks. A fully conserved block means a set of adjacent alignment columns which contain only one type of nucleotide. The minimum length of such a block is one and the largest block length depends on the longest observed block (`lob`). Given a list of blocks and their corresponding length it is possible to evaluate the number of consecutive blocks from one up to the `lob`. To get a distribution where the area under the curve sums up to one, we calculate the frequency f as given by Formula 4.2.

$$f = \frac{\#\text{consecutive blocks of length } x}{\#\text{all consecutive blocks}} \tag{4.2}$$

All alignment columns which contain twelve *Drosophilids* and only one type of nucleotide character were extracted from the `PECAN` alignments. Approximately 11.6 million fully conserved alignment columns were detected. This indicates that approximately 10% of the *D. melanogaster* genome are conserved across all *Drosophilids*. The `lob` spans 209 alignment columns. Figure 4.2 displays the distribution of all possible consecutive blocks of length x calculated by Formula 4.2. We assume that the start of the five percent quantile marks the minimum length l. For the `PECAN` alignments the five percent quantile starts at length 20. This leads to the conclusion that a block of minimum length 20 is significant in the set of consecutive blocks and by that in the set

Table 4.2: Possible combinations of seed lengths l (columns) and conservation thresholds c (rows). Each entry is calculated by Formula 4.3 and m_j denote the number of mismatches in both directions.

	$l = 10$	$l = 15$	l=20
0.95	$m_j = 0$ $1 - 0 = 1$	$m_j = 0$ $1 - 0 = 1$	$m_j = 1$ $1 - \frac{2}{22} \approx 0.91$
0.90	$m_j = 1$ $1 - \frac{2}{12} \approx 0.83$	$m_j = 1$ $1 - \frac{2}{17} \approx 0.88$	$m_j = 2$ $1 - \frac{4}{24} \approx 0.83$
0.85	$m_j = 1$ $1 - \frac{2}{12} \approx 0.83$	$m_j = 2$ $1 - \frac{4}{19} \approx 0.79$	$m_j = 3$ $1 - \frac{6}{26} \approx 0.77$

of `PECAN` alignments.

These two approaches suggest that the parameter l should be set between 14 and 20. In all further analyses the minimum length of a seed and by that the minimum length of an UCR is set to 15.

Conservation threshold (c):

As described in Section 3.2, this threshold is needed to define a minimum value of conservation between each sequence and the consensus. If the conservation of one sequence falls under the given threshold the extension process is stopped. Each seed is extended twice and therefore the conservation of a complete extended seed is measured by Formula 4.3, where n is the number of sequences in the alignment, m_1 the number of mismatches, and l_1 the length of an extension to the left and the pair m_2, l_2 the equivalent for the right extension.

$$\text{conservation(UCR)} = \min_{i=1..n} \underbrace{\left(1 - \frac{m_1 + m_2}{l_1 + l + l_2}\right)}_{T} \tag{4.3}$$

To obtain the effective conservation value for a given threshold c the underbraced term T has to be maximal. This is the case if the length l_j of each extension is equal to the number of mismatches m_j with $j \in \{1, 2\}$. Table 4.2 shows a set of seed lengths (columns), conservation thresholds (rows), and the result for each combination of them. With respect to the definition of UCRs the conservation threshold for each extension is set to 95%. An effective conservation of 90% for a maximal extended UCR with seed length 19 is the result. The prevention of fuzzy ends by trimming an extended seed back to the last fully conserved alignment column corrects this worst case estimation to a higher percentage of 91%.

Number of characters unequal to the consensus sequence (u):
As described above a less conserved column is only included if it contains two characters out of the alphabet $\mathbb{A}$ at maximum. Note that $\mathbb{A}$ contains the four DNA nucleotides (A, G, C, T) and the gap character $(-)$. However, it was not noted so far how these two characters are distributed. In the set of twelve *Drosophilids* the value u can be set from one up to six. If, for example, a column with 9 A's and 3 G's is observed the consensus character is A and the character unequal to the consensus is G. Choosing a value of 3 for the parameter u would lead to an inclusion of that column, since $u = 3$ states that there are at most 3 consensus unequal characters allowed. On the other hand, choosing $u = 2$ would lead to a rejection of that column, since there are more characters unequal to the consensus than allowed. It is obvious that the total number of UCR columns will grow if u is increased. Furthermore all UCR columns calculated with $u = 1$ are included in the result set which is calculated with value 2. This parameter should permit less conserved columns with respect to sequencing errors (e.g. doubling or missing of a nucleotide [61]) and columns with subgroup specific mutations from one nucleotide to another should be prevented. Since sequencing errors are not that systematic as subgroup specific mutations are, u should be set to one or two.
For the analysis presented here the value $u = 2$ is used.

Chapter 5

Results

The aim of this Chapter is to find reasons for the existence of UCRs. In the first section, features of the detected UCR set are summarized and compared to previous findings. In addition, this chapter addresses four hypotheses to elucidate reasons for the existence and possible functions of UCRs. In Section 5.2 especially UCRs within coding sequences are in the focus. Mainly the relation of UCRs and codon usage bias is examined. The prediction of non-coding RNAs [13] becomes more and more the focus of research. For example, small non-coding RNA molecules (e.g. miRNAs, snRNAs) are regulatory elements and the reported functionality of highly conserved elements in the context of putative non-coding RNA genes is mentioned in different studies [1, 11, 12]. In Section 5.3 the secondary structure and the corresponding thermodynamic stability is regarded for the complete UCR pool. The distinction between (non-)unique UCR sequences could help to elucidate their function. Uniqueness of UCR sequences within the genome of *D. melanogaster* is addressed in Section 5.4. Finally, the hypothesis that UCRs are conserved because of several overlapping sequence features is tested in Section 5.5.

5.1 Abundance of UCRs

As described in section 4.2 the minimum length l is set to 15, the conservation threshold c to 95% and the number u characters unequal to the consensus is set to 2. Only alignments containing all twelve *Drosophila* sequences are considered.

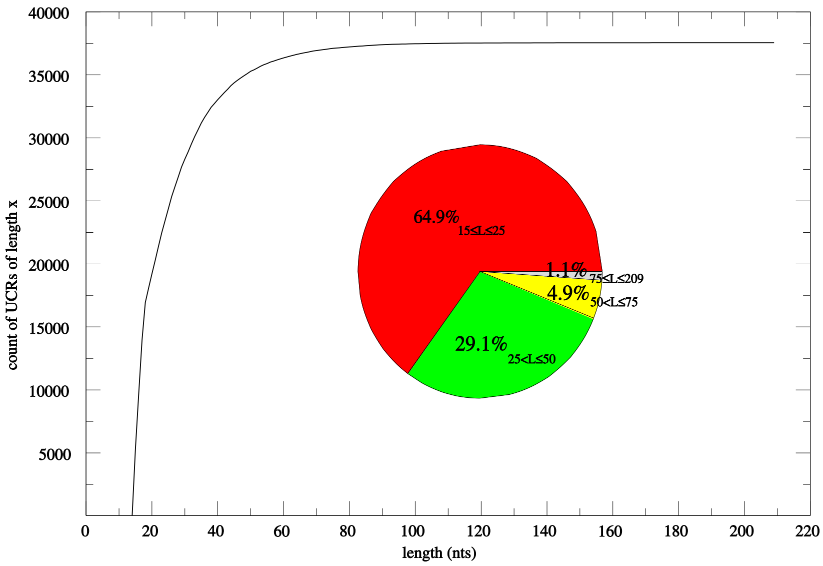


Figure 5.1: The cumulative curve displays the length distribution of the detected 37.554 UCRs (parameters: $l = 15, c = 95\%, u = 2$). The included pie-chart shows a partitioning into four length-classes and illustrates especially the fraction of UCRs longer than 100 which are not presentable by the curve. A block of 209 totally conserved alignment columns is the longest observed one within the result set.

The set of UCRs detected with our method contains 37,554 elements with 944,210 totally conserved and 22,620 less conserved alignment columns. The minimum length is 15 (seed length) and the obtained maximum length is 209. The UCR of length 209 is a totally conserved block across all twelve *Drosophila* sequences. Figure 5.1 displays the length distribution of the detected UCR set. As expected the number of UCRs with a length of less or equal to 25 alignment columns forms the largest fraction of 64.9% (depicted in the pie-chart). These findings are in agreement with previous ones [34]. The fraction of short UCRs includes the unextended seeds of length 15. Furthermore, it should be noted that all extended seeds with length less than 22 are totally conserved blocks. Due to the chosen conservation threshold of 95% a less conserved column could be included and will be kept if a length of 22 is reached. The large fraction of 64.9% could correspond to small regulatory elements (e.g. small non-coding RNAs or transcription factor binding sites). The `REDfly database version 2.0`[1] which is a curated collection of known *Drosophila* transcriptional cis-regulatory modules (CRMs) and transcription factor binding sites (TFBSs) contains 1345 TFBSs with a mean length of approximately 20nts. Indeed, these could not elucidate 64.9% of small UCRs. Nevertheless, the lack of annotations for such regulatory elements in *D. melanogaster* is known and therefore short UCRs might help to find novel elements within this class [69]. The uniqueness of all UCR sequences is addressed by a whole genome `BLAST` search for the complete UCR pool. This might give an explanation for the large set of small UCRs which could be associated either with specific or general regulatory elements (see Section 5.4).

D. melanogaster represents the best characterised species in the set of twelve *Drosophilids*. It is a well studied model organism with a compact, completely sequenced genome and well established physical and functional maps [9, 10]. All this information is provided by several databases at the World Wide Web (see Appendix A.2). The other sequenced *Drosophilids* are not that well characterised and most of their annotations (if available) are based on comparative genomic data to *D. melanogaster*. Due to these facts all further analyses are based on annotations of *D. melanogaster*.

The used *D. melanogaster* euchromatic genomic sequences of chromosome 2L, 2R, 3L, 3R, 4 and X sum up to approximately 118,4 million bases. We summarize the genomic regions which are annotatable as intron, coding sequence (CDS), 3' or 5' untranslated region (UTR) into the term 'genic'. The considered euchromatic genomic sequence is subdivided into 57% genic and 43% intergenic regions. According to the UCR detection procedure one percent (964,450nts) of the *D. melanogaster* euchromatic sequence is ultra-conserved. The partitioning into genic and intergenic applied to the UCR set results in 48% genic and 52% intergenic sequences (for details see Appendix A.3). Note, all previous counts are based on alignment columns and therefore, the given counts vary with respect to the *D. melanogaster* euchromatic sequence. Retelska et al. [3] obtained a similar partitioning for their so called conserved non-protein-coding (CNC) DNA regions in *D. melanogaster* and *D. virilis*. Overall *D. melanogaster* and *D. virilis* conserved sequences the CNCs account for 47.69% with conservation threshold

[1] `http://redfly.ccr.buffalo.edu/`

of 80% and 53.9% for a conservation threshold of more than 90%. The fraction of CNC elements (> 90% conservation) are comparable to the intergenic fraction above.
The conservation of the genic regions can be explained by general evolutionary constraints due to requirements of the protein bio-synthesis and structure. The percentage of UCRs in intergenic regions suggests similar constraints that cause the high amount of conservation within these regions. However, this indicates that the importance of non-coding regions within *Drosophila* genomes is normally underestimated.

5.1.1 Comparison to previous studies

The datasets of Glazov et al. [1] and Papatsenko et al. [2] are used to compare the detected UCR set with previous studies of UCRs within the *Drosophila* genus. The first set (referred to as gUCR) is based on a pairwise `BLASTz` alignment [29] of *D. melanogaster* and *D. pseudoobscura*. Both *Drosophila* species diverged approximately 25-30 million years ago [70]. In the study of Glazov et al. UCRs with a minimum length of 50nt and 100% percent identity are analyzed. The second set of Papatsenko et al. is based on a pairwise `LAGAN` alignment [71] of *D. melanogaster* and *D. virilis*. These two species diverged approximately 40 million years ago[2]. Within this study all identical ungapped matches of 60nt minimum length are analyzed. To get a dataset which is comparable to the gUCRs (*geq* 50nts, 100%) the pairwise-`LAGAN` alignments (used by Papatsenko et al. [2]) were downloaded from `http://pipeline.lbl.gov/data/mel_vir/` and all conserved regions with minimum length of 50nt and 100% percent identity were extracted (referred to as pUCR). Obviously the data set of Papatsenko et al. is completely covered within the extracted pUCR set.

Both sets (gUCR and pUCR) are based on *D. melanogaster* coordinates/sequences of the BDGP release 3.1. A sequence mapping to BDGP release 4.3 is needed to compare these two approaches with all detected UCRs of 50nt minimum length from our set (referred to as dUCR). A whole genome `BLAST` search for each of the gUCR and pUCR sequences is used to determine their BDGP 4.3 coordinates. Only elements with an unique entire `BLAST` hit in the *D. melanogaster* euchromatic genome (BDGP release 4.3) were used. In addition the mapped loci had to be covered within `PECAN` alignment regions containing all twelve *Drosophila* sequences. These criterias are fulfilled by 88% of the gUCRs and 82% of the pUCRs.

Since the dUCRs were limited to contain all twelve *Drosophilids* the set is much more restricted than the other two (gUCR as well as pUCR). To determine the influence of this restriction we extracted a pairwise dataset from the `PECAN` MSAs containing twelve *Drosophilids*. All *D. melanogaster*-*D. virilis* sequences are used to extract identical ungapped matches with a minimum length of 50nt. After filtering (only uniquely occurring UCRs) the extracted dataset contains 4,532 elements (referred to as eUCR).

[2] `http://insects.eugenes.org/species/about/genome-proposals/GenomesWP2003.pdf`

Table 5.1: Summary of the mapping and filtering procedure. To get four comparable sets of unique occurring UCRs (with respect to BDGP release 4.3) with minimum length of 50nt.

	gUCR	**pUCR**	**eUCR**	**dUCR**
alignment	pairwise `BLASTz`	pairwise `LAGAN`	extracted pairwise `PECAN` alignments	MSA `PECAN`
species	*melanogaster-pseudoobscura*	*melanogaster-virilis*	*melanogaster-virilis*	twelve *Drosophilids*
UCRs $\geq 50nt$	23,669	8,712	4,617	2,451
unique in BDGP 4.3	23,323	7,976	4,532	2,424
`PECAN` covered[3]	20,921 (1,346,402nt)	7,128 (450,298nt)	4,532 (285,065nt)	2,424 (157,706nt)
genic proportion	41% (546,250nt)	41% (185,545nt)	47% (133,180nt)	43% (68,079nt)

Finally, only uniquely occurring UCRs with a minimum length of 50nt are used to form four well defined and comparable datasets. For all these sets Table 5.1 summarizes the results of the mapping and filtering procedure. The four UCR sets show a similar proportion of genic nucleotides. The partitioning into genic and intergenic applied to the complete UCR set results in 48% genic and 52% intergenic nucleotides. Long UCR elements (50nt minimum length) seem to be slightly shifted towards intergenic regions in comparison to the complete set.

D. melanogaster and *D. pseudoobscura* diverged approximately 25-30 million years ago [70]. Comparing this to the divergence time of *D. melanogaster* and *D. virilis* the gUCR set is based on the more closely related species. Therefore, the huge amount of 20,921 UCRs in this dataset was expected.

Comparing the MSA extracted eUCRs with the pUCR set shows a 1.6-fold UCR enrichment between *D. melanogaster* and *D. virilis* within the pairwise `LAGAN` alignments. If the compared set is additionally restricted to contain twelve *Drosophila* sequences

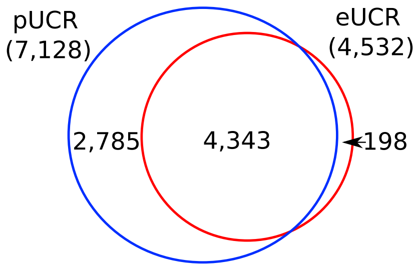


Figure 5.2: The venn diagram displays the overlap of the pUCR and eUCR set. The overlap counts 4343 elements. With respect ot the eUCRs 96% are covered.

(dUCR set) the UCR count is approximately 2-fold in the eUCR set and 3-fold in the pUCRs with respect to the dUCRs. This indicates the observable differences between pairwise and multi-sequence alignments as well as their underlying algorithm. A MSA algorithm is optimized to align as many sequences as possible within a region. Therefore, pairwise aligned regions might be dismissed.

However, calculating regional overlaps of all sets enables us to analyze the data sets in detail. Since the different UCR sets might vary in length a minimum coverage of 90% is applied. An overlap of the eUCR set and the pUCRs results in a coverage of 96% with respect to the eUCRs (see Figure 5.2). This indicates comparable features of these two pairwise datasets. Applying an overlap of the dUCR set to each of the three pairwise datasets results in a mean coverage of 81% (gUCRs 2036/2424 = 84%; pUCRs 2001/2424 = 82%; eUCRs 1892/2424 = 78%). With respect to the extracted pairwise alignment this reveals the information that the dUCR set is subsumed by ∼78% fully conserved elements and the remaining ∼22% are either short extended seeds (seed length less than 50nt) or UCRs of different lengths. Detailed analysis of all dUCRs shows that 16% (402/2451) are extended seeds of a length less than 50nt. Therefore, these elements are not expected to be covered within the pairwise datasets. The coverage of 82% (pUCRs) and 84% (gUCRs) is expected and explainable by UCRs with a seed length of less than 50nt.

Since the dUCR set is based on a MSA and a restrictive definition of ultra-conserved regions is used, short UCR elements (length $\leq$ 50nt) of the complete UCR pool might cover additional sequences of the pUCR as well as gUCR set. However, testing the comparability of the dUCR set to previous findings was the aim of this Section. The results described above show that we analyze 'the tip of the iceberg' with such a restrictive dUCR set which is based on a MSA and detected with our innovative technique (see Chapter 3). Nevertheless, the dUCR set shows a good accordance with previous approaches and their findings will be applicable to our data as well.

5.2 CDS and codon usage

Hypothesis:

The term protein bio-synthesis subsumes a multi-step process and describes the pathway from a DNA encoded gene to the corresponding protein. This central process is subdivided into three main steps: Transcription, post-transcriptional processing and translation. Figure 5.3 depicts a simplified work-flow of this process. During transcription the protein-coding locus is transcribed into the pre-messenger RNA (pre-mRNA). The pre-mRNA is subdivided into exonic and intronic sequences. Post-transcriptional processing is necessary to get the mature mRNA (mRNA) out of the pre-mRNA. This term summarizes several procedures: Splicing (separate exonic and intronic sequences), capping (add 5'-cap-structure), polyadenylating (append a poly-A-tail at the 3' prime end) and editing (modification of single or multiple nucleotides) of the mRNA sequence. Outcome of this second step is the mRNA combined of a continuous protein-coding sequence, so called open reading frame (ORF) and the flanking untranslated regions (5' and 3' UTR). Only the ORF sub-sequence of the mRNA encodes for the corresponding protein. In other words the ORF of a transcript defines the exact template for the finally translated sequence.

The process of translation is based on the genetic code. Therefore, the genetic code is a mapping between tri-nucleotide RNA sequence (codon) and the corresponding amino acid. These amino acids are protein basic modules with several specific attributes and functions. The genetic code is redundant but not ambiguous. It is possible that one amino acid is encoded by several codons (redundancy) but one codon encodes not more than one amino acid (ambiguity). Anyway, four RNA nucleotides Adenin (A), Guanin (G), Cytosin (C) and Uracil (U) are combinable to 64 triplets. Only 20 amino acids have to be encoded, therefore it is possible to encode one amino acid by a set of triplets. Further on, we term amino acids with more than one corresponding codon 'n-fold degenerated amino acid' where n is equal to the number of possible codons. An amino acid with two corresponding codons is named two-fold degenerated amino acid. Usually the triplets for one amino acid differ at the third codon position (see Table 5.2). For example Histidin (His) is encoded by CAU and CAC. Exceptions are

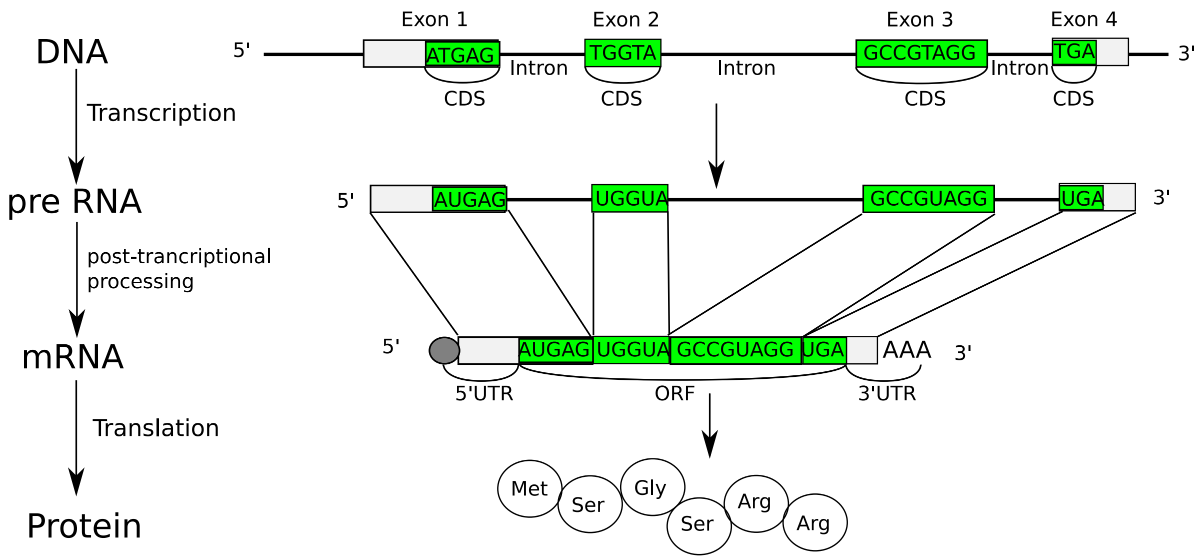


Figure 5.3: Simplified work-flow of the protein bio-synthesis. This process is subdivided into transcription, post-transcriptional processing and translation. On DNA and mRNA level the protein-coding portion of exons (depicted in green) is subdivided into protein-coding sequences (CDS). After the post-transcriptional processing these sequences are arranged in a continuous open reading frame (ORF). The ORF forms the exact template of the translated protein.

Arginin, Leucin and Serin with six corresponding codons. In these cases the first and/or second codon position could also differ between triplets. Conservation is a common feature within protein-coding sequences (CDS) due to the constant sides of codons and the specific function of amino acids. However, in sense of ultra-conserved regions within coding sequences (CDS-UCR) the high amount of conservation is of interest. Assuming that mutations occur randomly along the genome a mechanism is needed to keep genomic regions constant. In protein-coding sequences usually this mechanism is expected on functional and structural level. A structural domain is an element of the overall structure within a protein that is self-stabilizing and often folds independently from the rest of the polypeptide chain. Domains are named and singled out because of their important role in the biological function within the protein they belong to. The importance of several domains results in a high amount of conservation. If the composition of an amino acid chain changes, especially in domains, the encoded protein could change or lose the structure and therefore its function. Obviously, essential proteins (e.g. proteins involved in embryonic development) are maintained over long

Table 5.2: The universal genetic code is the fundament of translation. Twenty amino acids and three special stop codons are encoded by 64 triplets. The triplet AUG codes for Methionin and serves as translation initiation site. The three stop codons UAA, UAG and UGG are triplets with no corresponding amino acid and translation terminates.

		Second Base				
		U	C	A	G	
First Base	U	UUU (Phe)	UCU (Ser)	UAU (Tyr)	UGU (Cys)	U
		UUC (Phe)	UCC (Ser)	UAC (Tyr)	UGC (Cys)	C
		UUA (Leu)	UCA (Ser)	UAA (STOP)	UGA (STOP)	A
		UUG (Leu)	UCG (Ser)	UAG (STOP)	UGG (Trp)	G
	C	CUU (Leu)	CCU (Pro)	CAU (His)	CGU (Arg)	T
		CUC (Leu)	CCC (Pro)	CAC (His)	CGC (Arg)	C
		CUA (Leu)	CCA (Pro)	CAA (Gln)	CGA (Arg)	A
		CUG (Leu)	CCG (Pro)	CAG (Gln)	CGG (Arg)	G
	A	AUU (Ile)	ACU (Thr)	AAU (Asn)	AGU (Ser)	T
		AUC (Ile)	ACC (Thr)	AAC (Asn)	AGC (Ser)	C
		AUA (Ile)	ACA (Thr)	AAA (Lys)	AGA (Arg)	A
		AUG (Met)	ACG (Thr)	AAG (Lys)	AGG (Arg)	G
	G	GUU (Val)	GCU (Ala)	GAU (Asp)	GGU (Gly)	T
		GUC (Val)	GCC (Ala)	GAC (Asp)	GGC (Gly)	C
		GUA (Val)	GCA (Ala)	GAA (Glu)	GGA (Gly)	A
		GUG (Val)	GCG (Ala)	GAG (Glu)	GGG (Gly)	G

Third Base

evolutionary time.

An amino acid with less than four codons is strongly limited in the mutable codon positions and possible nucleotides. If for instance the third position of the Histidin codon is mutated from U or C into an A or G the triplet encodes Glutamin and not anymore Histidin. This amino acid alteration may change structure and function of the whole protein and therefore the substitution is prevented. Given an ultra-conserved subsequence of a protein the usage of amino acids which are encoded by one (Methionin and Tryptophan), two (Asparagin, Aspartic acid, Cystein, Glutamin, Glutamic acid, Histidin, Lysin, Phenylalanin, Tyrosin) or three (Isoleucin) triplets might keep the nucleotide composition more constant than others because of their strict limitation in possible third codon nucleotides.

The addressed hypothesis is that one- to three-fold degenerated amino acids occur more frequently within CDS-UCRs than others. Furthermore, it might be possible that some codon types are more likely within CDS-UCRs than others. A special nucleotide composition (e.g. preferred nucleotides at specific codon positions) might be observable for these codons.

Testing:

Out of the 964.450 UCR nucleotides 11% (101.318nts) are part of CDS. Indeed several regions which are annotated as CDS could for example also be annotated as UTR, intron or non-coding RNAs. However, there is no obvious explanation which functional role is more important if only one function should be considered. Within this approach we used all CDS annotated (`FlyBase Release 4.3`) sequences.

After the post-transcriptional processing the protein-coding portion of exons (named CDS and depicted in green in Figure 5.3) is arranged in a continuous sequence the so-called open reading frame (ORF). The ORF is translated from the 5' to the 3' end. It begins with the start codon AUG and is translated in three-way steps until a stop codon in frame terminates the process (simplified work-flow see Figure 5.3). However, given a CDS it is not compulsive that the frame starts at the first sequence position. Three reading frames (denoted with 0,1 and 2) are possible for each CDS with already known reading direction: If the ORF starts/continues at the first sequence position of the CDS the reading frame is 0, if the ORF continues at the second sequence position of the reading frame is 1 and if the ORF continues at the third sequence position of the CDS the reading frame is 2. As depicted in Figure 5.3 the reading frame of the second and third CDS starts not at the first position. The encoded Serin, for example is subdivided on DNA and pre-mRNA level. Two nucleotides are located at the first and the last one at the second CDS. The reading frame of the second CDS is 1 and only one complete amino acid is encoded (simplified example).

The `FlyBase Release 4.3` annotations for the CDS did not provide information about the reading frame and therefore it is necessary to reconstruct the corresponding ORF.

Each known transcript is identifiable by an unique seven digit `FlyBase`-transcript-number (FBtr). Anyway, given an ORF for one FBtr, which is a combination of CDS, the frame to each segment is assignable. For each FBtr we reconstruct the ORF out of the annotated CDS. Only ORFs without an in-frame stop codon, a length divisible by three and encoded either on the plus or minus strand (no trans-splicing) are used. To prevent the overestimation of alternative transcripts all CDS annotated regions with the same start and end positions, strand and frame are unified.

Assuming a set of frame-annotated CDS and the corresponding intersection of the CDS-UCR set it is possible to estimate the codon usage of UCRs on the one hand and overall CDS sequences on the other hand. This is done by counting the appearing codons in all annotated CDS-UCRs and CDSs. For comparison of the occurring amino acids the fraction of an amino acid count to the total number of considered codons is calculated (referred to as occurrence rate).

Results:

The bar-plot in Figure 5.4 depicts the occurrence rate (y-axis) for each amino acid (x-axis) within CDS-UCRs and CDSs. From left to the right the number of possible codons (number in brackets) for the amino acids decreases. As expected, the occurrence rate of amino acids with more than three codons is within the CDS set greater than in CDS-UCRs. This changes with the three-fold degenerated amino acid, Isoleucin (Ile). With respect to the CDS set the three stop codons occur approximately 2-fold enriched within CDS-UCRs. This indicates a correlation between ultra-conserved regions and 3'UTR of transcripts. The tendency that amino acids with less than four corresponding codons are more frequent in CDS-UCR than in CDS is observable. Nevertheless, it is only a tendency and questionable why the occurrence rate of Glutamic acid (two-fold) and Glutamin (two-fold) is higher within CDS compared to CDS-UCRs.

Amino acids within the CDS-UCR set might be restricted to a subset of preferred codons and therefore the ultra-conservation is explainable. Independently of the corresponding codon count preferred codons for the one or the other set might be observed overall amino acids.

For each codon the occurrence rate relative to the complete set (either CDS-UCR

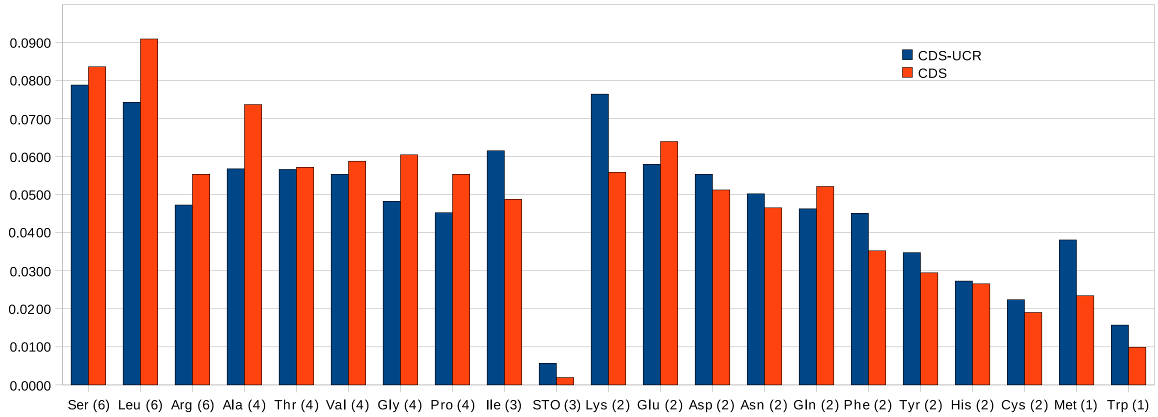


Figure 5.4: Amino acid usage within the euchromatic portion of the *D. melanogaster* genome. For comparison of the occurring amino acids within both sets (CDS-UCR and CDS) the fraction of each amino acid count to the total number of considered codons is calculated. This occurrence rate (y-axis) of each amino acid (x-axis) is depicted. From left to the right the number of corresponding codons (noted in brackets) decreases. The total number of considered codons for the CDS set is 6,808,743 and 30,250 for the CDS-UCR set.

or CDS) is known. To estimate if it is more likely to observe a codon t within the CDS-UCR set or in the CDS set the log-odds ratio $R(t)$ is calculated:

$$R(t) = log_2 \left(\frac{f_{\text{CDS-UCR}}(t)}{f_{\text{CDS}}(t)} \right) \tag{5.1}$$

Base two logarithm (log_2) is used, in which case the unit of $R(t)$ is a bit. $R(t)$ is positive if it is more likely to observe the corresponding codon within CDS-UCRs than in CDS set. A negative value of $R(t)$ indicates that the corresponding codon is more likely to be observed within the CDS set. Figure 5.5 depicts $R(t)$ (x-axis) for each amino acid and their corresponding codons (y-axis). From top down the number of corresponding codons increases.

Amino acids encoded by less than four codons are more likely to be observed within the CDS-UCRs (similar to Figure 5.4). No alternative codon is possible for Methionin as well as Tryptophan. Therefore, the corresponding log-odds score has to be treated differently from the remaining amino acids. Indeed the positive log-odds score indicates that it is more likely to observe these amino acids ultra-conserved.

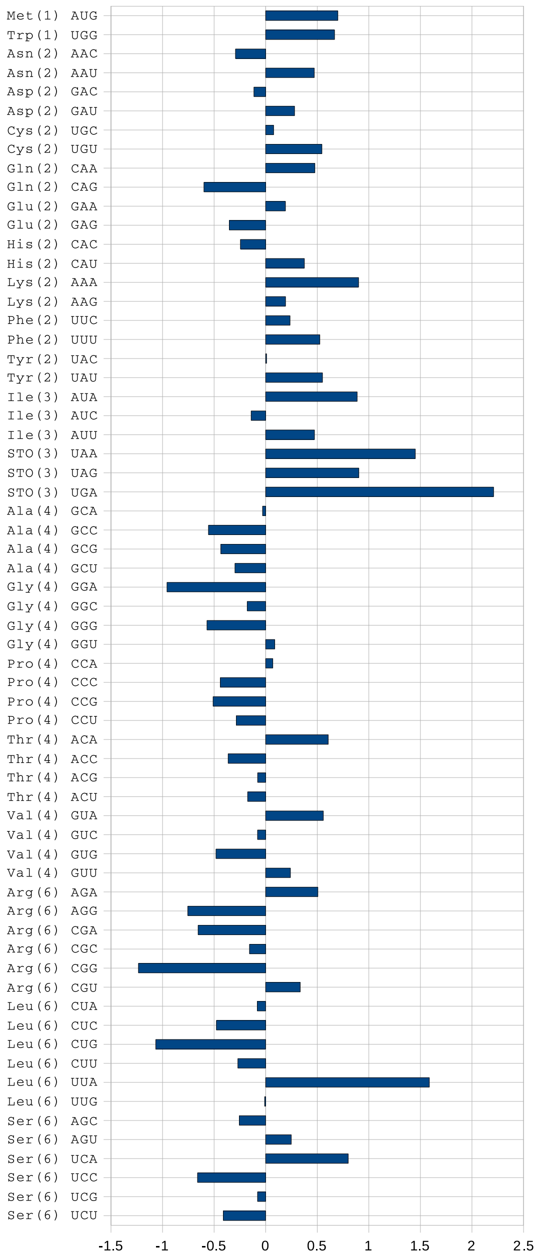


Figure 5.5: For each codon the log-odds ratio $R(t)$ is calculated using Formula 5.1. This ratio indicates how more likely it is to observe a codon within the CDS-UCR than in the CDS.

Nevertheless, this ratio is supported by the functional and/or structural importance of Methionin and Tryptophan but not by their codon composition.

Interestingly, codons with an Uracil and especially Adenin at the third codon position occur more frequent within the CDS-UCRs (left hand side of Figure 5.6). This is true especially for the subset of two- and three-fold degenerated amino acids (center of Figure 5.6). Asparagin, for instance shows a positive ratio for the triplet AAU and on the other hand a negative one for AAC. It is not always a pairing of a positive and negative log-odds ratio but the value for Uracil and especially Adenin ending codons is always greater than the value for Guanin and Cytosin.

Siepel et al. [12] reported for insects as well as vertebrates that an unexpected large fraction of highly conserved bases is located within the 3'UTR. Therefore, it is expected that the surrounding region of a stop codon tend to be ultra-conserved. The high log-odds scores of all three stop codons are supporting aspects for the association of UCRs and 3'UTRs. The ultra-conserved regions at the three prime end of the ORF might be constrained by transcript stability and/or translatability.

As mentioned before four- and six-fold degenerated amino acids tend to occur more

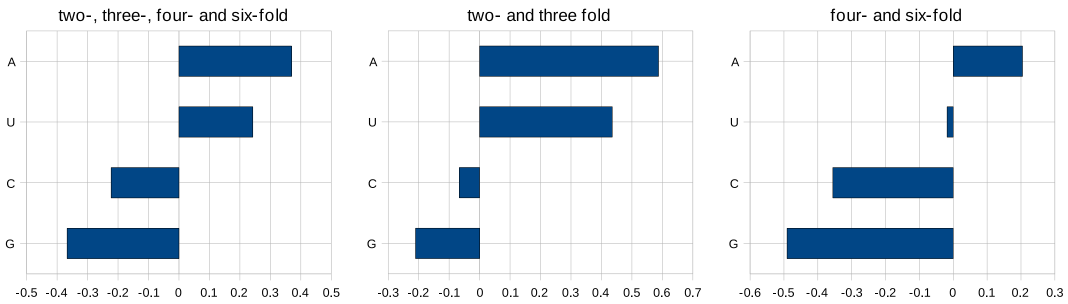


Figure 5.6: Subsumed log-odds ratio separated for each nucleotide at the third codon position. At the left hand side the log-odds ratio is subsumed over all two-, three-, four- and six-fold degenerated amino acids. Note that the stop codons have no corresponding amino acid and are therefore not regarded within this plot. The two other plots are subsumed either from two- and three-fold (center) or four- and six-fold degenerated amino acids (right hand side). Overall amino acids with Uracil and especially Adenin ending codons are more likely to occur within the CDS-UCRs.

frequent within the CDS set than in CDS-UCRs. This is also reflected in Figure 5.5. With respect to some exceptions most log-odds scores of the four- and six-fold degenerated amino acids are negative. Nevertheless codons with an Uracil and especially Adenin at the third codon position are more likely to occur within CDS-UCRs than others. The right hand side of Figure 5.6 displays the log-odds ratios overall four- and six-fold degenerated amino acids separated for each nucleotide at the third codon position. A good example is the four-fold degenerated amino acid, Valin. The codons ending with an Adenin or Uracil show positive log-odds ratios. In contrast the other codons have negative log-odds ratios.

To analyze the observed differences between CDS and CDS-UCR in more detail an expectation measurement based on nucleotide frequencies is calculated. For the complete set of CDS the mono-nucleotide frequencies are countable. Assuming that the mono-nucleotide frequencies ($f(A), f(T), f(G)$ and $f(C)$) are known and the three codon positions are independent, Formula 5.2 calculates an expectation frequency f_m of a codon composed out of the three nucleotides x, y and z ($x, y, z \in \{A, C, G, T\}$).

$$f_m(x, y, z) = f(x) \times f(y) \times f(z) \tag{5.2}$$

All CDS and their contained codons are known. This facilitates the opportunity to take position dependent nucleotide frequencies into account. For each nucleotide three frequencies are needed if a position dependent frequency of nucleotides is used . Anyway, this value ignores the known fact that the di-nucleotide content effects such a measurement and should also be taken into account. For each codon x_1, y_2, z_3 we have two di-nucleotides x_1, y_2 and y_2, z_3 and their corresponding frequencies $f(x_1, y_2)$ and $f(y_2, z_3)$ within the CDS set. Using position dependent mono- as well as di-nucleotide frequencies of the complete CDS set enables us to calculate an expectation value $f_p(x_1, y_2, z_3)$ for each codon:

$$
\begin{aligned}
f_p(x_1, y_2, z_3) &= f(x_1) \times \frac{f(x_1, y_2)}{f(x_1)} \times \frac{f(y_2, z_3)}{f(y_2)} \\
&= \frac{f(x_1, y_2) \times f(y_2, z_3)}{f(y_2)} \qquad (5.3)
\end{aligned}
$$

This measurement reflects how often the codon x_1, y_2, z_3 is expected given the mono- as well as di-nucleotide frequencies of the CDS set. The next step is the calution of the differnce Δf:

$$\Delta f(x_1, y_2, z_3) = f_o(x_1, y_2, z_3) - f_p(x_1, y_2, z_3) \qquad (5.4)$$

The value of $\Delta f(x_1, y_2, z_3)$ (using Formula 5.4) is zero if the expected frequency f_p is equal to the observed frequency f_o. Furthermore, the observed frequency f_o is completely explainable by mono- and di-nucleotide frequencies in this case. A positive value of $\Delta f(x_1, y_2, z_3)$ indicates underrepresentation of the codon compared to the expected value. Otherwise the difference is negative and therefore an overrepresented codon frequency is indicated. Over- as well as underrepresentation of a codon frequency could be explained by additional constraints acting on the codon. Figure 5.7 depicts the difference $\Delta f(x_1, y_2, z_3)$ for each codon within both sets (UCR-CDS as well as CDS). In both sets over- as well as underrepresented frequencies are observed. Interestingly, the sum over the absolute values of $\Delta f(x_1, y_2, z_3)$ for all codons is by a factor of 1.8 greater in CDS-UCRs than in CDS.

Discussion

As mentioned above amino acids which are encoded by less than four codons occur more frequently within CDS-UCRs than in CDS. Nevertheless, to decide for a given polypeptide chain if a sub-sequence is ultra-conserved is not possible. Only a slightly

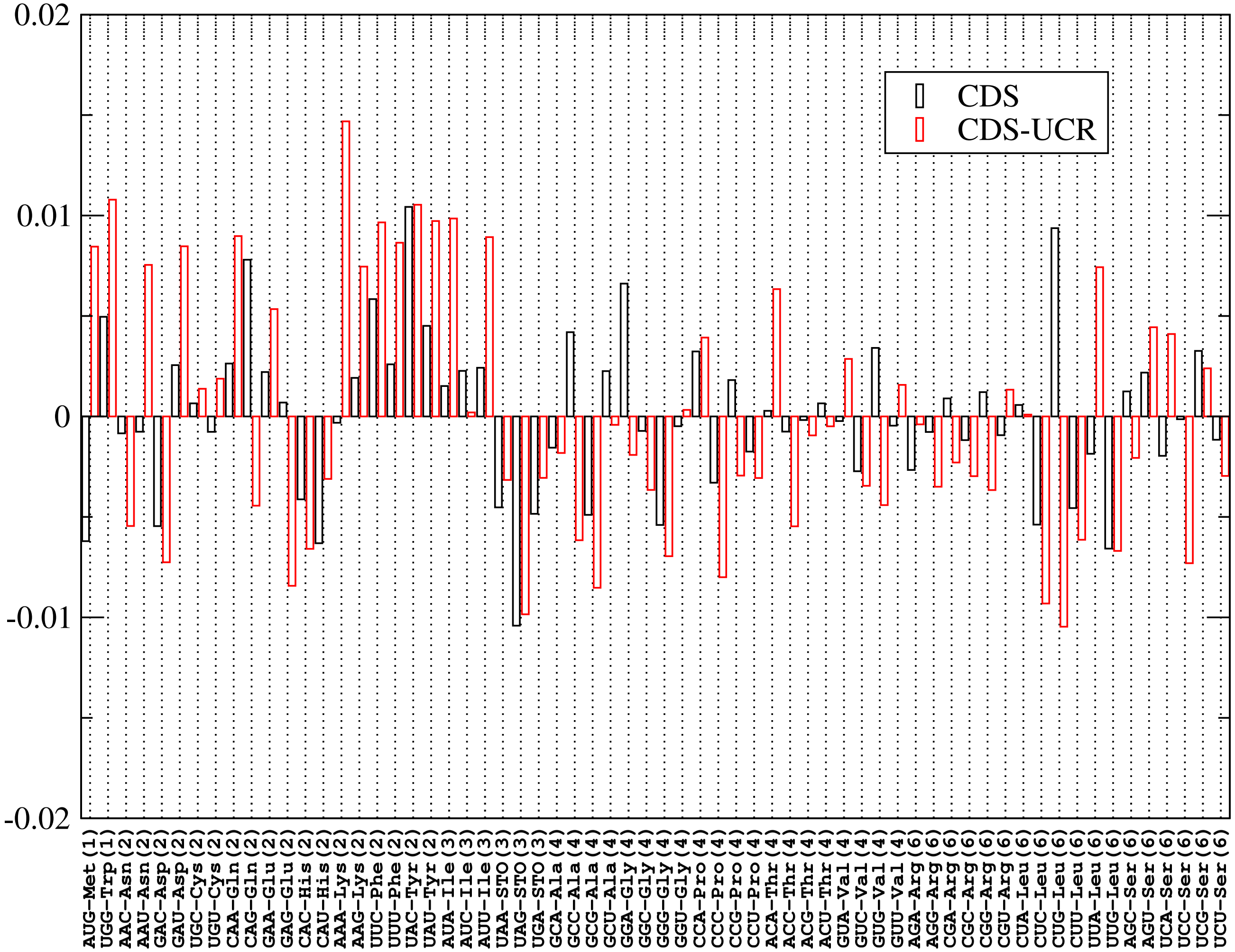


Figure 5.7: Difference of the expected codon frequency f_p and the observable codon frequencies f_o in CDS and CDS-UCR sets. If the bar is above or below the zero level the observed codon frequency is not completely explainable by the estimated frequency f_p. This indicates that additional constraints are reflected within the observed codon frequency f_o.

tendency is observable and therefore the constraints which keep UCRs constant within coding sequences is not detectable at the CDS level.

We observed a higher divergence from the expected value (by a factor of 1.8) between UCR in CDS and CDS in general. This hints at stronger or additional unknown constraints on UCR in general then the surrounding CDS. Furthermore, the data suggest that the constraints on UCRs are different from those on CDS as indicated by divergence from the expected value in opposit directions.

The finding that codons within CDS-UCRs tend to prefer codons with an Uracil and especially Adenin nucleotide at the third codon position is contrary to the reported codon usage biases for the *Drosophila* genus. Vicario et al. [72] reported that in most cases the preferred codons in all twelve *Drosophilids* are those with Guanin and especially Cytosin in the third codon position. One of the possible reasons for this contrary codon usage within CDS-UCRs might be caused by region dependent occurrences of these elements. It is known for *D. melanogaster* that preferred codons near exon boundaries tend to be enriched for Adenin and depleted in Cytosin [73]. Additionally, for vertebrates it is reported that UCRs frequently overlap with splice site junctions [5]. This seems also to be true for *Drosophilids* and our UCR set. Approximately 36% (1,517/4,221) of all CDS associated UCRs overlap with intron/exon junctions. Therefor the occurrence of UCRs near exon boundary is assumed.

5.3 Structured UCRs

Hypothesis:

Is there an evidence for ultra-conserved functional elements in non-coding RNAs or do UCRs encode for non-coding RNAs?
A common problem for researchers working with RNA is the prediction and calculation of RNA secondary structures. The RNA inventory is larger than the protein-coding mRNA inventory [55, 74]. Several other RNA molecules (e.g. tRNA, rRNA, snRNA) are needed to produce the protein of the DNA sequence. In the last years RNA molecules come more and more in the focus of interest. So called non-coding RNAs were found to regulate several processes like transcription and translation. For example, miRNAs are short molecules (21-24nt) that attenuate protein bio-synthesis by binding complementary sites in target mRNAs. Most non-coding RNAs are characterized by a specific secondary and/or tertiary structure.
A RNA is defined by the primary structure which is the order of RNA nucleotides (Adenine A, Guanine G, Cytosine C, Uracil U) in 5'→3' direction. The functional conformation of an RNA molecule is its secondary structure. The secondary structure contains Watson-Crick (A-U, C-G), wobble (G-U) and other non-canonical base-pairs between nucleotides of the RNA sequence. A single-stranded RNA is folded into a structure composed of hairpins, bulges and interior loops. Figure 5.8 illustrates all types of RNA secondary structure elements implemented in the `Vienna RNA Package`[4]. Given a possible RNA single stranded sequence several programs (e.g. `RNAfold` [76], `Mfold` [77]) are available to predict the most reliable RNA secondary structure. The structure of an RNA molecule can be predicted computationally by calculating the minimum free energy (MFE) for all different combinations of hydrogen bindings and domains. These methods are based on minimizing the free energy of the RNA molecule and maximizing the number of favorable base-pairs. Each base-pair decreases the stacking energy and stabilizes the secondary structure. On the other hand non-complementary bases increase the stacking energy and destabilize the secondary structure [78, 79].

[4] `http://www.tbi.univie.ac.at/~ivo/RNA/`

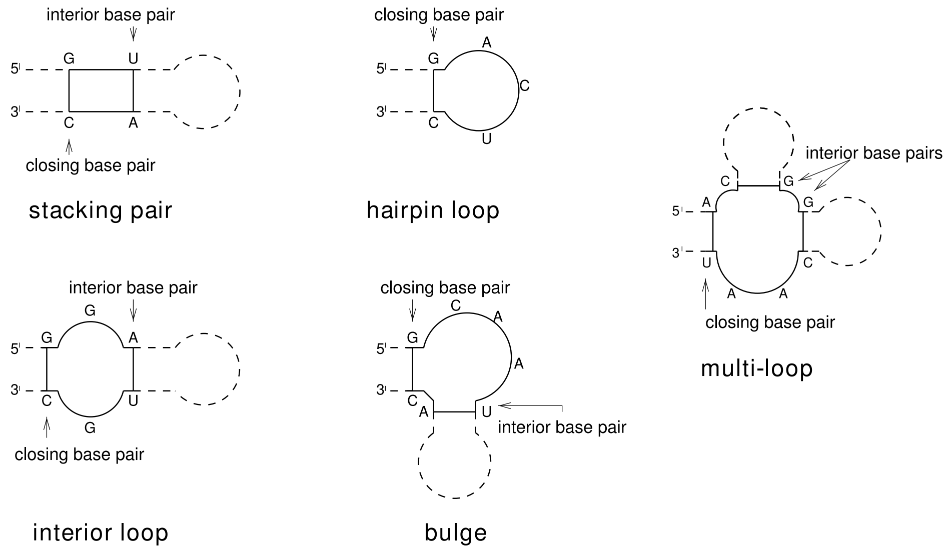


Figure 5.8: Depicted are all allowed base-pairs and their resulting substructure elements (with friendly permission of Christoph Flamm [75]).

Testing:

The stability of a secondary structure is quantified as the amount of free energy which is stored by forming base-pairs. Positive free energy denotes that the supply of external energy is required to form a secondary structure; negative free energy catalyzes the folding procedure. Free energies are additive, so one can determine the total free energy of a secondary structure by adding all the component free energies. The more negative the free energy of a structure, the more likely is the formation of that structure, since there is more energy accumulated. This finding is used to predict the secondary structure of a particular sequence by calculating its minimum free energy (MFE). However, the MFE depends on the length and the nucleotide composition of the sequence, and is therefore difficult to interpret in absolute terms.

The thermodynamical stability of an RNA secondary structure relative to a set of shuffled sequences with the same dinucleotide composition is estimated by a variant of the Z-score procedure. We used the program `RANDFOLD` from Bonnet et al. [80] which applies `RNAfold` for the MFE calculation and shuffles explicit the 'native' RNA sequence. `RANDFOLD` takes the RNA molecule, estimates the 'native' MFE, shuffles the

Table 5.3: Length distribution of the folded UCRs relative to the complete set. The bracket percentage shows the fraction of structured UCRs relative to the complete set with length l.

length l	37,554 complete set	22,800 folded UCRs (61%)
$l \leq 25$	24,365	10,278 (42%)
$25 > l \leq 50$	10,918	10,252 (94%)
$l > 50$	2,271	2,270 (∼100%)

sequence n-times and calculates the MFE for each shuffled sequence. The outcome is a distribution of possible MFE values for a sample of RNA sequences with the same nucleotide content. This randomization test results in the frequency p given by Formula 5.5. Where N is the number of iterations and R is the number of shuffled sequences that have a MFE less or equal to the 'native' MFE.

$$p = \frac{R}{N+1} \tag{5.5}$$

A high number of iteration is needed to reach a good estimation of p [81].
However, the RNA coding strand is not known for UCRs since conservation is direction independent. The explicit shuffling is expensive in terms of computation-time. Therefore, the MFE is calculated in advance for the UCR-sequence on the top strand as well as its reverse complement with `RNAfold`. Assuming, that the sequence with the smaller MFE is more stable, its p-value is calculated using 1000 iterations and dinucleotide shuffling. This procedure is applied to each of the 37,554 UCRs.

Results:

Sixty-one percent of the UCR set folds into a more or less stable secondary structure (structured UCR iff MFE < 0). Table 5.3 shows the length distribution of the folded UCRs relative to the complete set. Calculating the MFE depends on the length of the given sequence and the increasing probability to form a base-pair in longer sequences. These limitations explain that the largest fraction of unfolded UCRs (58%) is less or equal to 25nt long. To estimate the degree of similarity between previously known non-coding RNAs and the structured UCRs a regional overlap and sequence similarity is used. The overlap of UCRs associated with annotated non-coding RNAs in *D. melanogaster* is calculated. Either the UCR or the non-coding RNA has to cover more than 90% of the other. The same regional overlap is used for the `RNAz` predicted non-coding RNAs from Rose et al. [13]. A `BLAST` search is applied to identify sequence similarities between structured UCRs and known non-coding RNAs from Rfam version 0.7 [82], NONCODE version 1.0 [83], miRBase version 9.0 [84], snoRNABase version 3[5], ncRNA database[6] and the *D. melanogaster* specific predictions in the tRNA database [85]. To determine the function of UCRs in the context of non-coding RNAs either a minimum of 90% of the UCR or the non-coding RNA have to be identical.

Table 5.4 summarizes the similarity between known and predicted non-coding RNAs and the structured UCR set. The number of `RNAz` loci and tRNA database hits are explicitly listed with respect to the fact, that these annotations are based on computational predictions either of non-coding RNAs or tRNAs without any *in vivo* confirmation. For further information about these predictions see Rose et al. [13] for the `RNAz` loci and Lowe et al. [85] for the tRNA database.

Figure 5.9 displays a graphical overview of the structured UCR set. For each structured UCR the MFE (x-axis) and the frequency (y-axis) is plotted. The covered non-coding RNA classes are miRNA (63), tRNA (37), snRNA (7), snoRNA (7) and non-protein-coding genes with different or unknown function (12, 8). An additional set of 131 structured and 84 unstructured UCRs is located within the Bithorax Complex (BX-C). The *Drosophila* BX-C contains nine homeotic genes [86] and ranges from the start of Ultrabithorax (*Ubx*) to the end of Abdominal-B (*Abd-B*). In context of non-coding RNAs bithoraxoid (*bxd*) and *iab-4* are the most interesting loci of the complex. The molecular function of *bxd* is described as specific transcriptional repressor [87]. The longest of four transcripts (bxd-RB CR31273-RD) contains 17 structured UCR elements with a length between 15 and 91 nucleotides (mean 30). Figure 5.10 depicts the four known *bxd* transcripts and the recently discovered *tre* non-coding RNAs [88]. Depicted in detail are two structured UCR elements (UCR_22934 and UCR_22936) in the *tre-3* and *tre-2* transcripts. This observation of ultra-conserved substructures in known complex non-coding RNA transcripts suggest the possibility to detect/verify other non-coding RNAs. Structure based analysis seem to play an important role to detect the

[5] `http://www-snorna.biotoul.fr/index.php`

[6] `http://biobases.ibch.poznan.pl/ncRNA/`

[2] complete listing of different non-coding RNAs see Appendix A.4

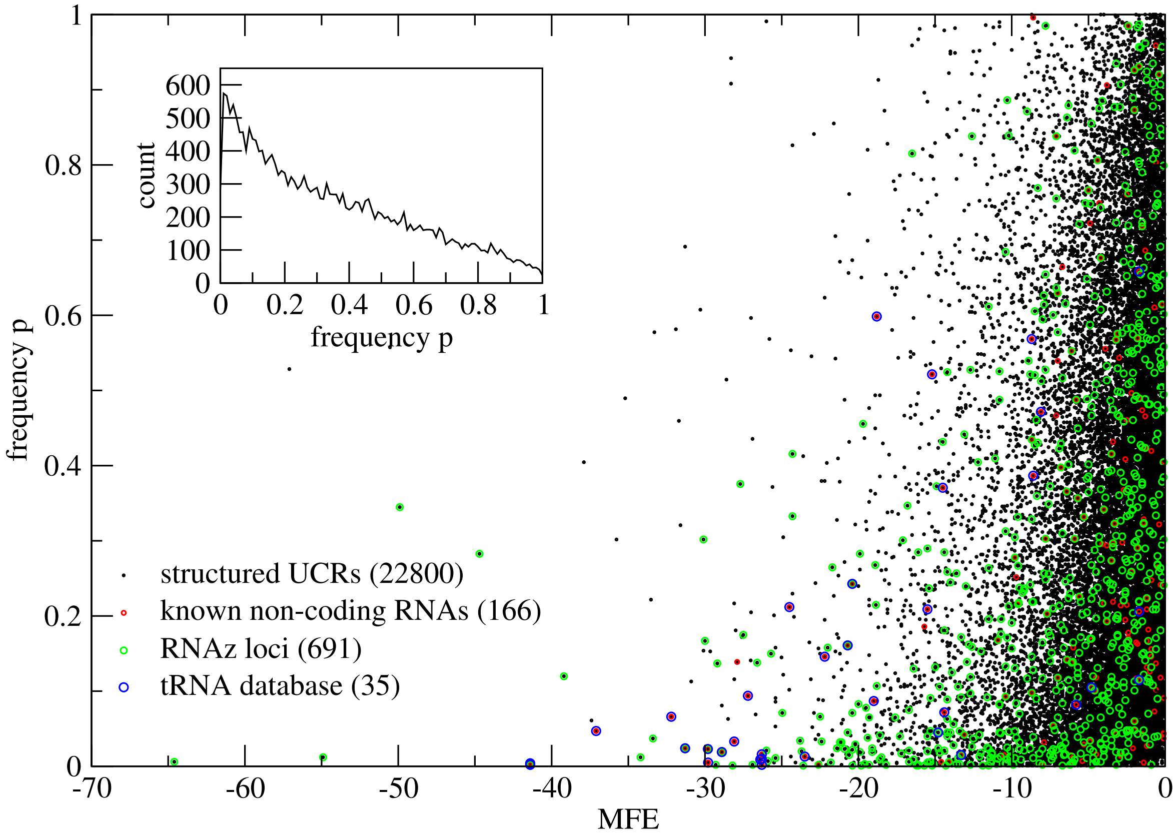


Figure 5.9: Graphical overview of structured UCRs. For each UCR the MFE (x-axis) and the frequency p (y-axis) is calculated. The p-value is calculated using $p = \frac{R}{N+1}$ where N is the number of iterations and R is the number of shuffled sequences with a MFE less or equal to the native sequence (for details see description above). The small included plot depicts the distribution of the p-values. On the x-axis the p-value and on the y-axis the corresponding count is noted. Points which are associated with known non-coding RNAs (red), `RNAz` loci (green) and/or tRNA database entries (blue) are highlight with colored circles.

Table 5.4: Annotation summary of the 22,800 structured UCRs (MFE < 0). The data set is subdivided into UCRs which overlap with CDS annotated regions or not (second and third column). The CDS overlapping subset contains 166 UCRs which overlap `RNAz` loci (p-value ≥ 0.9) from Rose et al. [13] and 3 previous annotated non-coding RNAs. Non-protein-coding regions contain 20,054 structured UCRs with 525 `RNAz` loci, 163 previous known non-coding RNAs and 35 tRNAs annotated in tRNA database. For comparison the set of unstructured UCRs is also shown. The counts of 'known non-coding RNAs' and '`RNAz` loci' is less than a third of the structured UCR counts. In the context of non-coding RNAs the proportion of unstructured UCRs is explainable by non-coding RNA loci which form (long) paired (stem) or unpaired (loop) substructures. If a (short) UCR element is located within such a substructure it is counted as unstructured.

	structured UCRs			unstructured UCRs		
	CDS	others	total	CDS	others	total
total	2,746	20,054	22,800	918	13,836	14,754
known non-coding RNAs	3[2]	163[2]	166	1[2]	58[2]	59
`RNAz` loci (32,758)	166	525	691	161	27	188
tRNA database (291)	0	35	35	0	1	1

constraints which keep UCRs constant at the sequence level.

The *iab-4* RNA gene contains 9 structured (8 with unknown function and 1 miRNA) UCR elements with length of 22 to 71 nucleotides (mean 42). The most stable UCR (UCR_22982, MFE=-22.60 $\frac{kcal}{mol}$, p=0.002, length=68nt) covers nearly the complete mir-iab-4 stem-loop (depicted in Figure 5.11). The mature miRNAs arise from the larger stem-loop structure (precursor) processed by ribonuclease III-like nuclease [89, 90]. Interestingly this miRNA locus resides only on the longest of two *iab-4* transcripts. This leads to a complex regulatory network of the miRNA locus. Ronshaugen et al. [91] presented the evidence that both *iab-4* miRNAs selectively attenuate *Ubx* activity *in vivo*. This attenuation results in a classical homeotic mutant phenotype: the star halters are transformed into a second pair of wings. Ronshaugen conclude that the whole *iab-4* transcription unit encodes a non-coding homeotic regulatory gene.

The findings of several structured and unstructured UCRs in the BX-C support the evidence of other non-coding regulatory loci like the *iab-4* and *bxd* complex or additional non-coding RNAs within these loci.

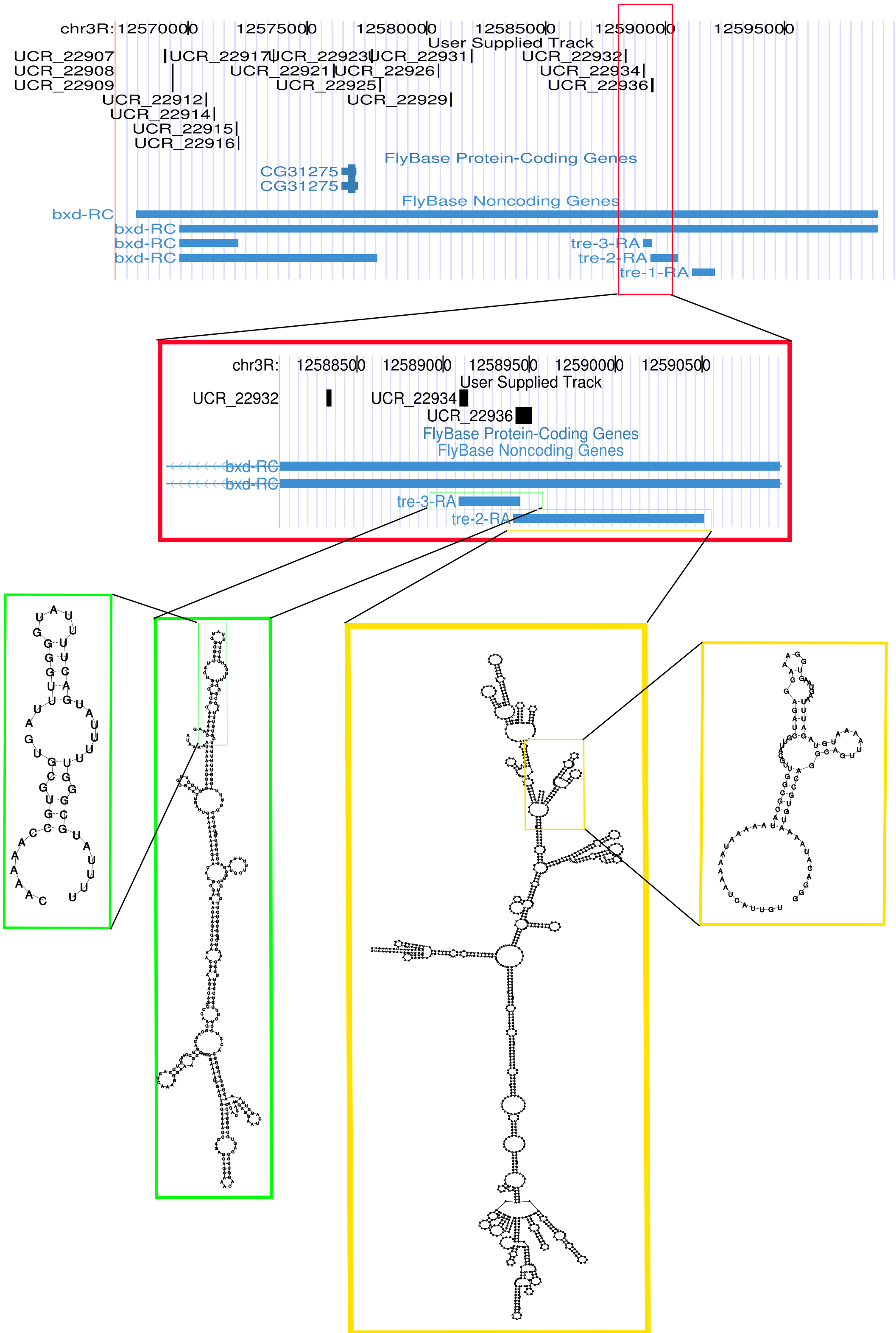
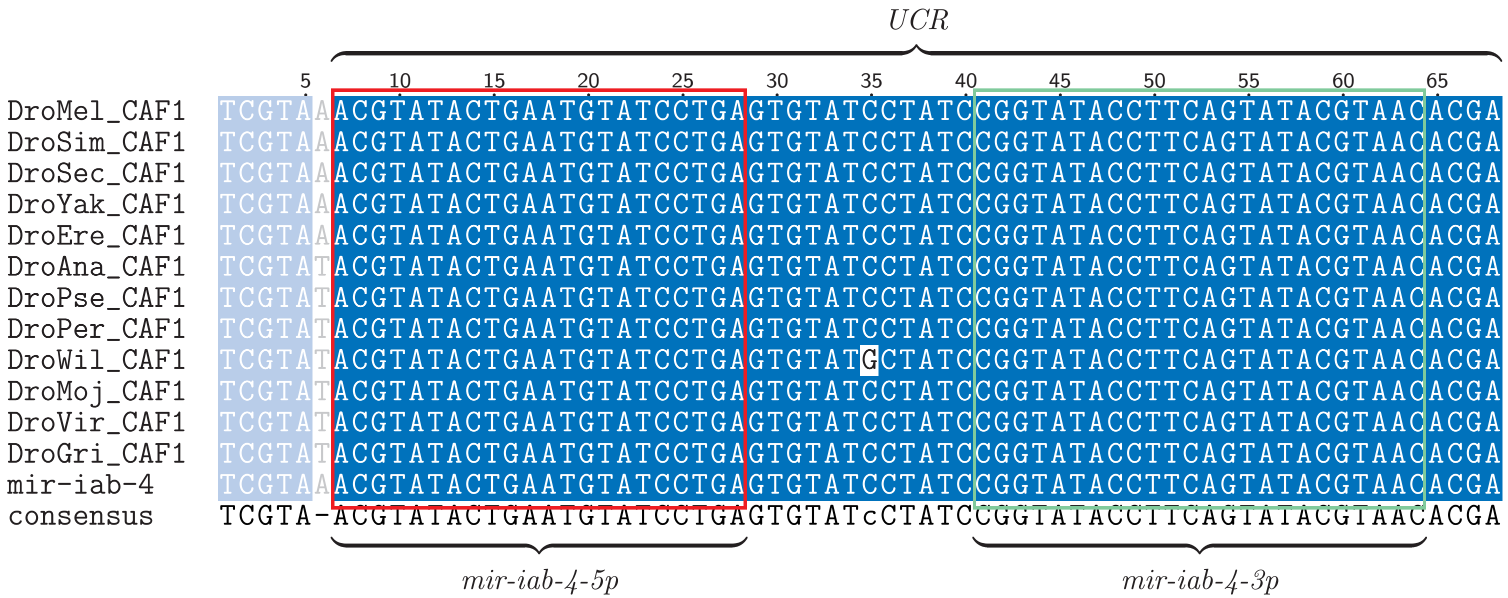


Figure 5.10: The genomic region of the non-coding RNA *bxd* with all structured UCRs as 'User Supplied Track', the 'FlyBase Protein Coding Genes' and 'FlyBase Noncoding Genes' displayed with the `UCSC Genome Browser` [65, 66]. Two sub-regions also known as non-coding RNAs are depicted in a higher resolution. Both *tre-3-RA* and *tre-2-RA* contain a structured UCR element close to their 3' ends.

a)



b)

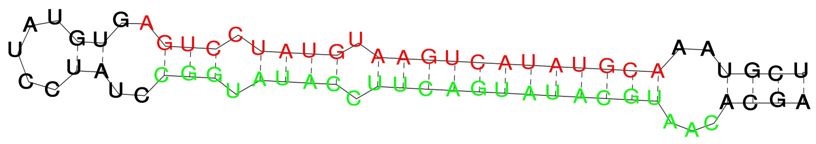


Figure 5.11: a) The alignment of the iab-4 stem-loop sequence with two mature miRNAs identified on both arms of the precursor. The mature RNAs mir-iab-4-5p and mir-iab-4-3p are marked either in red or green in the alignment and the secondary structure b).

As described above, 37 UCRs are associated with known and predicted tRNAs. Thirty-three of these UCRs are associated with `tRNAscan` predictions as well as already annotated as tRNAs. The remaining four UCRs are subdivided into 2 UCRs which are associated with `tRNAscan` predictions and 2 already annotated UCRs. Detailed analysis of the two UCRs which show only an overlap to `tRNAscan` predictions offer a strong signal to encode for 'real' tRNAs. The first one is UCR_13163 (chr2R, 6920606-6929677, 72nt, MFE $= -26.30\frac{kcal}{mol}$, $p = 0,009$) which covers the complete predicted tRNA no.15 on arm_2R in the tRNA database[7]. The MFE and the p-value correspond to already annotated tRNAs (see Figure 5.9). UCR_13164 (chr2R, 6921139-6921191, 52nt, MFE $= -14.80\frac{kcal}{mol}$, $p = 0.045$) on the other hand, shows a high sequence similarity to tRNAs encoding for the Threonin anti-codon. The complete tRNA coding sequence is covered if the only one nucleotide adjacent UCR_13165 (chr2R, 6921193-6921213, 21nt, MFE $= -2.3\frac{kcal}{mol}$, $p = 0.334665$) is also taken into account (see Figure 5.12). This region has also a `RNAz` predicted locus (chr2R, 6921117-6921227, 111nt) which supports the assumption of an unknown encoded tRNA.

Focusing on the top structured UCR candidates UCR_33374 (chr3R, 6411318-6411500, 183nt, MFE$= -64.60\frac{kcal}{mol}$, $p = 0.006$) points to the homothorax (*hth*) gene. This gene locus is known to encode for a homeodomain containing protein [92], which is involved in various developmental processes (e.g. eye development in *D. melanogaster* [93]). The *hth* gene coding region contains 274 structured UCRs. Figure 5.13 depicts the arrangement of the total 394 UCRs in the region of the known *hth* transcripts. The UCR_33374 element in the context of *hth* was mentioned by Glazov et al. [1]. They used pairwise sequence alignments from *D. melanogaster*, *D. pseudoobscura* and *A. gambia* for the process of UCR detection. The reported UCR set includes all identical non-gaped matches longer than 50nt between these three insect genomes. The UCR which corresponds to UCR_33374 with 183nt in length is a locus of 197nts between *D. melanogaster* and *D. pseudoobscura* and 92nt identical match between *D. melanogaster* and *A. gambia* (see Figure 5.14). UCR_33374 covers the donor of the last common intron-exon junction (indicated by the red line in Figure 5.13). Glazov et al. conclude that the secondary structure of this UCR locus blocks the splicing of the according intron and intron retained transcripts are the result. By Northern blot hybridization

[7] `http://lowelab.ucsc.edu/GtRNAdb/Dmela/Dmela-by-locus.html`

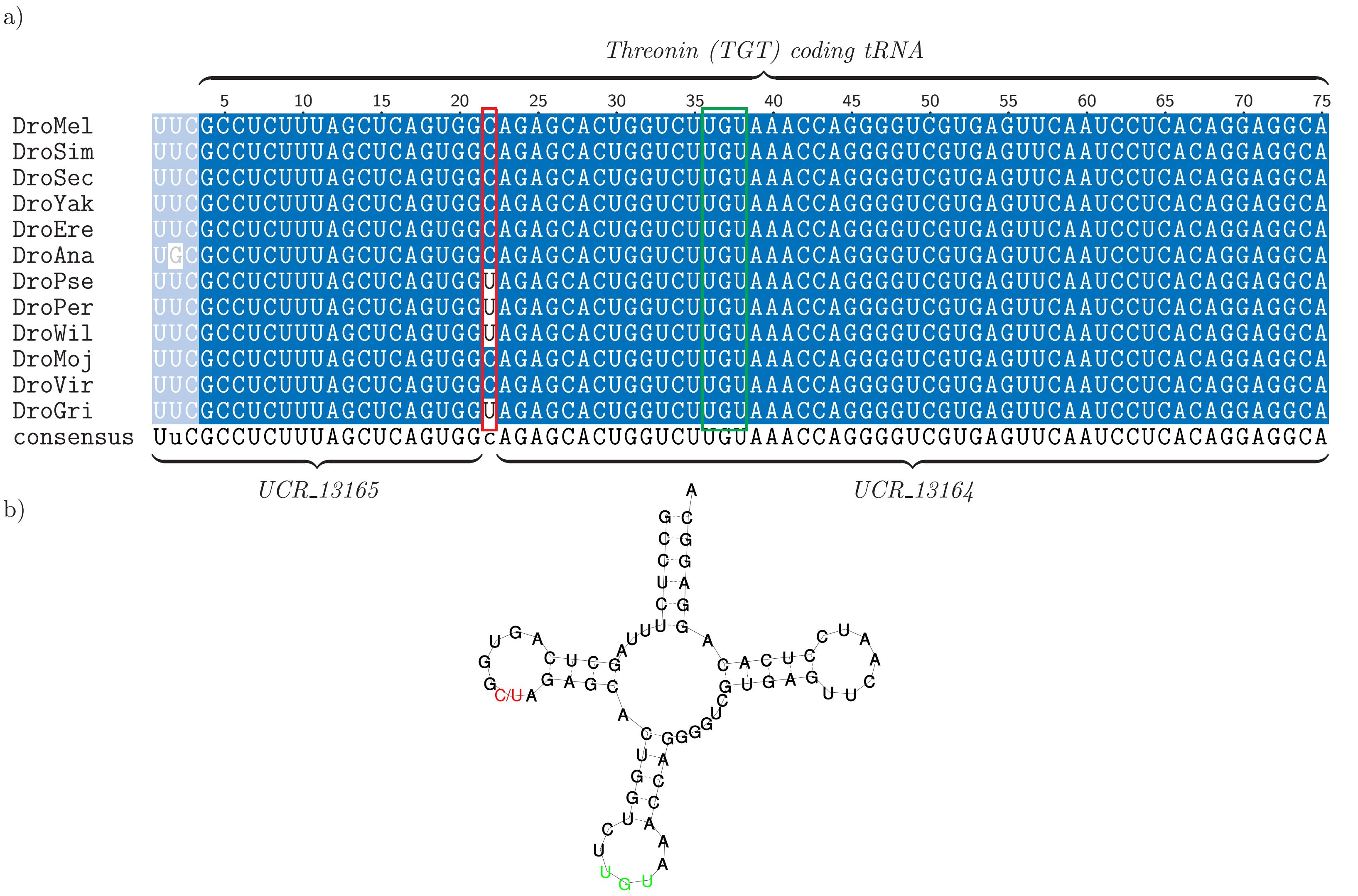


Figure 5.12: Threonin coding tRNA combined out of two UCR loci (UCR_13165 and UCR_13164). The red marked region in the alignment a) and the secondary structure of the tRNA b) indicates the less conserved alignment position in red. Highlighted in green is the anti-codon region of the tRNA molecule.

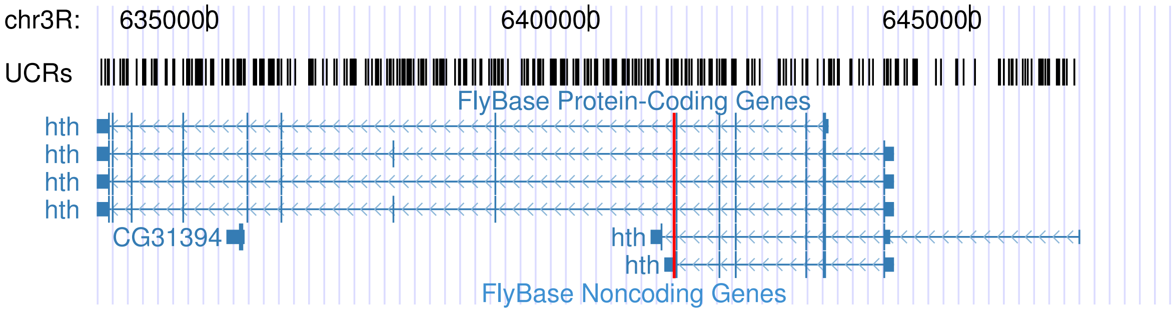


Figure 5.13: Distribution of overall UCRs (structured and unstructured) within the known *hth* gene locus. The red line marks the longest UCR which covers the donor of the last common intron-exon junction.

they showed that the majority of steady state RNAs exceeds the previous known length. Several publications have demonstrated that intron retention (IR) is a common feature in alternative spliced transcripts (e.g. [94, 95, 96, 97]). As mentioned by Gougle et al. [94] an inserted 15bp long hairpin structure is able to inhibit almost the complete splicing within yeast. In the context of UCRs the work of Phillipps et al. [95] shows that alternative splicing is often associated with sequence conservation.

The timing of the structural pre-mRNA formation seems to play an important role in regulation of splicing and gene expression. In 2006 Glazov et. al published a second paper about splicing control of the *hth* transcript. They describe a mechanism to form alternative structures were a closely located second UCR element shows functional importance as a cis-regulator. The 45nt long UCR_33369 (MFE $-4.60\frac{kcal}{mol}$, $p = 0.727$) is the corresponding element within our set. UCR_33369 is located in the retained intron and is 1048nt distant to UCR_333474. As depicted in Figure 5.15 both elements show complementary sequences and could also form a stable secondary structure. The switch between two stable secondary structures might regulate intron retention. If the transcript folds into the structure which covers the intron-exon junction intron retention results. Otherwise, a base-pairing between the two complementary UCR elements leaves the intron-exon junction open for splicing. Several other proteins or non-coding RNAs can be involved to regulate this mechanism (discussed [42]).

These findings indicate that non-coding RNAs contribute to the functionality of the UCR pool in many cases.

```
                         5    10   15   20   25   30   35   40   45   50   55   60
UCR_33374  ------------ACATCGCACACCGATGGAGCTAGTACACCAGACGTTGTGAGTACCCCCTTCGCGGGAGCACAT
dmel_dpse  AACGCCGACTCCACATCGCACACCGATGGAGCTAGTACACCAGACGTTGTGAGTACCCCCTTCGCGGGAGCACAT
dmel_anop  ------------------------------------ACACCAGACGTTGTGAGTACCCCCTTCGCGGGAGCACAT

            65   70   75   80   85   90   95   100  105  110  115  120  125  130  135
UCR_33374  GGTCCTATATTGGCGAGCTACAATGCGGTGGTGCACCCATGTTCGTAGTACGATAACGACCTTCTTGCAAAAACG
dmel_dpse  GGTCCTATATTGGCGAGCTACAATGCGGTGGTGCACCCATGTTCGTAGTACGATAACGACCTTCTTGCAAAAACG
dmel_anop  GGTCCTATATTGGCGAGCTACAATGCGGTGGTGCACCCATGTTCGTAGTACGA----------------------

            140  145  150  155  160  165  170  175  180
UCR_33374  AAGCGGGTGCTCACAATGCCCTGGGATTAAATGAATCAATAATAA--
dmel_dpse  AAGCGGGTGCTCACAATGCCCTGGGATTAAATGAATCAATAATAATA
dmel_anop  -----------------------------------------------
```

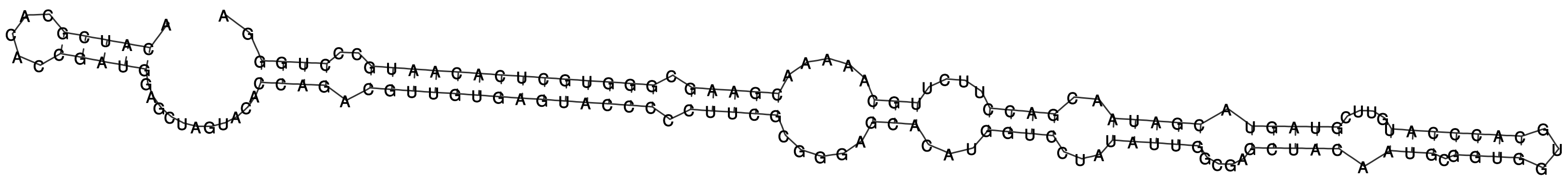


Figure 5.14: a) Alignment of UCR_33374 within the *hth* gene locus and the corresponding UCR-sequences used from Glazov et al. [1]. b) Illustrated is the stable hairpin structure of UCR_33374.

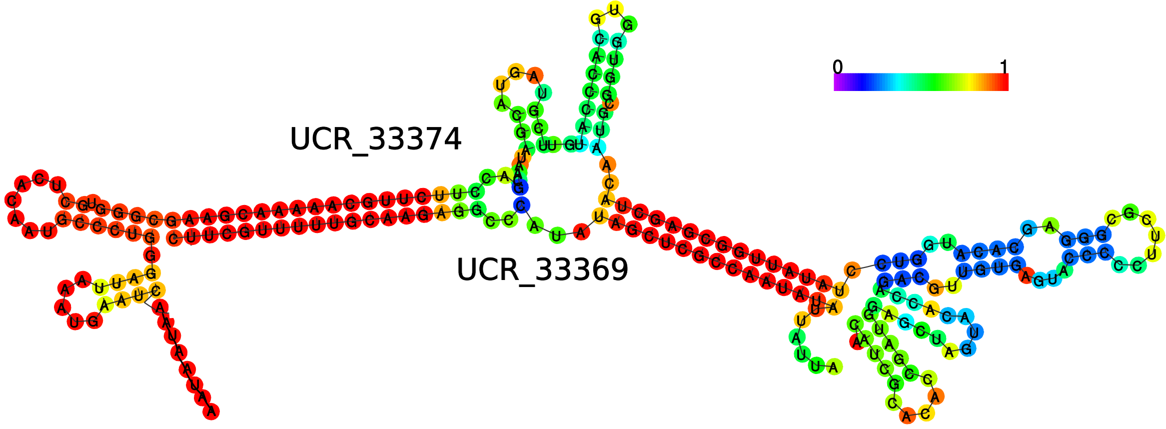


Figure 5.15: Secondary structure of UCR_33369 and UCR_33374. The complex is folded using *RNAco-fold*. Color code indicates the probability of being paired for each base within the complex. Important are the two regions were both UCRs base-pair with each other with probability one.

5.4 Uniqueness of UCR sequences

Hypothesis:

The `PECAN` alignments contain all syntenic aligned genomic sequences of the twelve *Drosophilids*. No filtering for repetitive elements has been performed so far. Therefore, the amount of 37,554 UCRs could be caused by aligned repetitive sequences. Nevertheless, the arising question is why they are ultra-conserved across all flies. If repetitive sequences are junk of the genome there is no obvious reason to keep them constant over 40 million years. On the other hand, cis-regulatory elements are known as common features within genomes. These regulatory elements work as enhancer and silencer of the corresponding genes. A set of different genes could be regulated by enhancer and/or silencer element which occur multiple times.
A conservative `BLAST` search of each UCR-sequence to the *D. melanogaster* euchromatic chromosome arms determines the occurrences of each element.

Testing:

The procedure described in Chapter 3 results in a set of alignment blocks with an effective conservation of at least 91%. The `BLAST` search within one genome (in this case *D. melanogaster*) of the considered alignments should also respect this conservation measurement. For each UCR element the sequence of *D. melanogaster* and the consensus sequence is used to calculate the minimum-matches of a `BLAST` hit. The minimum-match count results from a pairwise sequence comparison of the *D. melanogaster* and the consensus sequence (see Figure 5.16). For each element of the UCR set a whole genome `BLAST` search is performed and the hits are filtered by a minimum alignable sequence with length equal or greater than the minimum-match count. This measurement respects the underlying UCR alignment with an effective conservation of at least 91%.

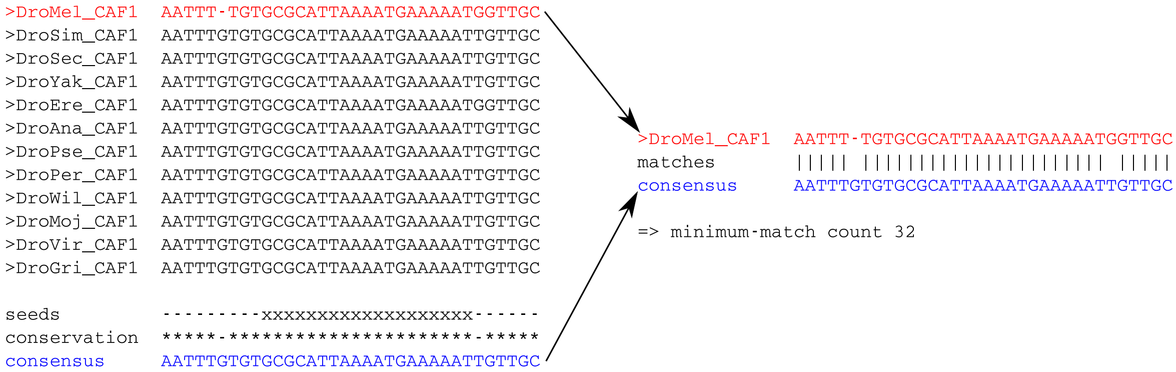


Figure 5.16: The minimum-match count results from a pairwise comparison between the species of interest and the consensus sequence of the alignment. The species of interest and the consensus sequence are extracted from the UCR alignment. A pairwise comparison of the extracted sequences and the number of exact matches results in the minimum-match count.

Results:

Approximately 87% (32,554) of the complete UCR sequences are unique within the *D. melanogaster* genome. This indicates a region specific function of these elements. Such a large fraction of unique conserved elements is reported within the human genome by Margulies et al. [11]. Derti et al. [98] use ultra-conserved elements of Bejerano et al. [5] and note their single copy within the haploid chromosome set. UCRs are absent only on chromosome Y (normally present in one copy in male) and chromosome 21 (causes trisomie 21 an autosomal aneuploidy). They conclude that the exact count of two copies are needed within a diploid cell. Of course their UCRs are much longer ($\geq$ 200nt) than those of our set forming 87%. Nevertheless, it is important to note that at least a single copy of UCRs of certain lengths (15-209nt; mean $\sim$27nt) is observed within the *D. melanogaster* genome.

However, within this section the remaining 13% are in the focus of interest. Figure 5.17 illustrates the length distribution of the multiple occurring UCRs. The length distribution is fractionated into three classes. First the UCR element longer than 110nt is of interest. A second class is formed by elements occurring less than 10 times (length 30nt up to 85nt) within the *D. melanogaster* genome. The remaining UCR elements with a length less than 30nt form the third group. All three classes are analyzed in

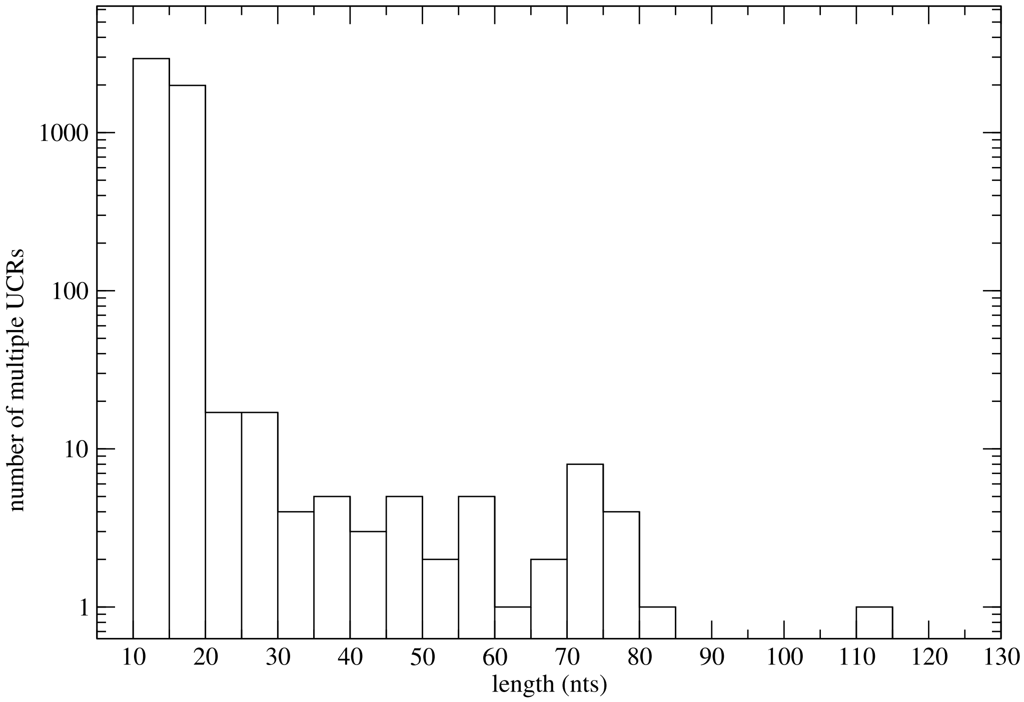


Figure 5.17: This histogram depicts the length distribution of the 5000 multiple UCR elements. As expected a large fraction of multiple UCRs is of length less than 20nt. Most of the UCRs with length greater than 30nt correspond to known non-coding RNAs (e.g. snRNA, tRNA).

detail within the next paragraphs.

The longest multiple occurring UCR element (UCR_27191; length 111nt) maps three times to an U6 spliceosomal small nuclear RNA (snRNA) cluster on chromosome arm 3R (FBgn0004188, FBgn0004189, FBgn0004190). The recently published paper [99] and an unpublished work of Marz et al. (personal communication) about the detection and verification of snRNA genes report these three U6 loci as functional elements. U6 is the only completely covered snRNA within the UCR set (FBgn0004190). Sub-regions of U4 (FBgn0003930) and U5 (FBgn0003938) of the major spliceosom are at least covered by two UCRs. Within the whole genome `BLAST` search UCR_18168 (length 47nt) and

UCR_18169 (length 18) map seven times to U5 (143nt) coding regions with a distance of two nucleotides. This distance is caused by an alignment column with more than two characters unequal to the consensus (see Appendix A.5). These seven U5 snRNA loci also correspond to the reported ones in [99] and Marz et al. For U4 (FBgn0003930) we report two corresponding UCRs (UCR_3725, length 24nt; UCR_3726, length 28nt) which map two out of three reported U4 snRNA loci within the *D. melanogaster* `BLAST` search. The missed U4 (FBgn0041719) snRNA is caused by the sequence difference of this locus compared to the other loci (see Appendix A.2). If both UCR elements occur their distance is always nine nucleotides within the *D. melanogaster* genome.

Nevertheless, only within the gene loci FBgn0003930 (U4), FBgn0003938 (U5) and FBgn0004190 (U6) UCRs are detected. Detailed analysis of the `PECAN` alignments corresponding to the snRNA U4 (FBgn0003931) coding region revealed that only the *melanogaster* subgroup is aligned. This suggests two explanations: a) the locus is specific for the *melanogaster* subgroup and per default not detectable with the procedure presented here or b) alignment artifacts causing 'miss-aligned' loci. To address the problem of 'miss aligned' sequences a `BLAST` search of the five UCR sequences within the *Drosophilids* is used. Table 5.5 summarizes the number of `BLAST` hits for each species in the context of snRNAs. All UCR sequences appear at least once in each genome. This is not surprising since we observed them as ultra-conserved alignment blocks.

Of interest is the fact, that the sequence of UCR_3725 appears only in some genomes while UCR_3726 occurs at least twice within known U4 loci. If both UCRs appear twice within one genome (e.g *D. sechellia*) the spacer between these hits is always nine nucleotides long. The observation that the sequence of UCR_3726 occurs at least twice in each genome and the fact that Mount et al. report three U4 loci in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. pseudoobscura* and *D. persimilis* indicates a specific function of UCR_3726 and maybe connected with UCR_3725. The count of UCR_3726 occurrences and the predicted snRNAs reported by Marz et al. fit well (except the *melanogaster* subgroup). The difference of UCR_3725 and the predictions of Marz et al. suggest that the function of this sub-sequence vary within the *Drosophilids*.

In case of the U5 snRNA UCR_18168 and UCR_18169 occur in tandem. Interestingly,

Table 5.5: Within the snRNAs U4, U5 and U6 ultra-conserved (sub)sequences are detected. To verify if these UCR-sequences are common motifs of those snRNAs we blasted all of them against each of the twelve genomes. To compare the occurrences of UCR-sequences in combination with detected snRNA counts of Marz et al. (unpublished, personal communication) and Mount et al. [99] their findings are also included. Positions with '-' indicate absent species in the set of Mount et al. Counts noted in brackets indicate the minimum and maximum of snRNAs predicted by Marz et al. without filtering these loci by promoter analysis.

		dsim	dsec	dmel	dyak	dere	dana	dpse	dper	dwil	dgri	dmoj	dvir
U4	UCR_3725 (24nt)	1	2	2	2	1	2	1	1	1	1	2	2
	UCR_3726 (28nt)	3[a]	2	2	2	3	2	3[a]	3[a]	3	3	3	3
	Marz et al.	3	3	3	3	3 (3-4)	2	3 (3-4)	3 (3-4)	3 (3-7)	3	3	3
	Mount et al. [99]	3	3	3	3	-	-	2	3	-	-	-	-
U5	UCR_18168 (47nt)	8	7[a]	7	8	7	4	7[a]	7[a]	8	7	6	6
	UCR_18169 (18nt)	8	6	7	8	7	2	6	6	7	4	4	5
	Marz et al.	8 (7-9)	7 (7-9)	7	8	7 (7-9)	4	7 (6-7)	7 (7-8)	8 (8-11)	7	6 (6-8)	6
	Mount et al. [99]	6	7	7	8	-	-	7	7	-	-	-	-
U6	UCR_27191 (111nt)	3	3	3	3[a]	5	2	4[a]	3	4[a]	2	3	2
	Marz et al.	3	3 (3-5)	3	3 (3-5)	4 (7-8)	2	3 (3-4)	3 (3-4)	1 (0-1)	3	3 (3-5)	2 (5-6)
	Mount et al. [99]	3	3	3	3	-	-	3	3	-	-	-	-

[a]Listed is the maximum match count for the UCR sequence. The maximum match count allows with respect to the UCR length a maximum of two mismatches within the entire `BLAST` hit.

the counts of the longer UCR-sequence (UCR_18168; 47nt) correspond to the findings of Marz et al. while the shorter UCR-sequence (UCR_18169; 18nt) diverge. This might have different meanings. If the tandem of both UCRs is required to form a functional snRNA locus the counts of the predictions overestimate the counts of U5 loci. It is also possible that the idea of snRNA verification using UCR-sequences is not working.
The U6 snRNA is completely covered by the longest UCR (111nt) within this set. Our `BLAST` results fit to the trusted snRNA loci of Marz et al. All counts in brackets denote the minimum and maximum count of possible snRNAs within one species. These snRNA loci were analyzed in more detail (see Marz et al.) and the counts in front (without a bracket) are assumed to be trusted snRNAs. The difference between the maximum and the trusted number of snRNAs could be caused by so called pseudogenes. Generally, pseudogenes are defunct relatives of known genes that have lost their coding ability or otherwise they lack essential gene features (e.g. promoter, CpG islands and splice sites). With the presented UCR-sequence strategy we were able to distinguish predicted and functional snRNA loci of U4, U5 and U6 from pseudogenes (e.g. U5 in *Drosophila willistoni*).

Is there an evidence for specific functions of ultra-conserved sub-regions in snRNAs? It is possible that UCRs fold into stable substructures and/or their function corresponds to specific binding. It is known that U4 and U6 form an RNA-RNA duplex during the first steps of the splicing process. Forming this U4-U6 complex also initiates the binding of U5. The secondary structure of the formed complex as well as the structure of the constituents plays an important role during the binding-process.
The secondary structure of the U4 (FBgn0003930) and U6 (FBgn0004190) snRNA alignments is calculated with `RNAalifold`[8] [100] of the `Vienna RNA Package`. Using `RNAalicofold` [101] the MFE secondary structure of the U4-U6 complex is estimated as well. Figure 5.18 depicts both the secondary structures of U4 (left), U6 (right) separately and the U4-U6 complex (center). The secondary structure of the U4-U6 complex corresponds to the complex-structure in human described by Shukla et al. [102]. The sequence of U6 is well conserved within the *Drosophilids*. However, the secondary structure, which is shown on the right-hand side, contains many A-U base-pairs. Such

[8] `http://www.tbi.univie.ac.at/~ivo/RNA/`

an A-U base-pair forms two hydrophobic bonds. An G-C base-pair on the other hand contains three hydrophobic bonds and maintains the stability of the structure. This enrichment of weak A-U base-pairs might explain the nearly unfolded U6 snRNA within the duplex.
On the other hand U4 consists of three G-C rich stem-loop structures (stem I, II, III). These three stems are maintained within the U4-U6 complex. The red encircled nucleotides mark the UCR elements within the U4 snRNA. UCR_3725 folds into a stable stem-loop structure with a MFE of $-7.9\frac{kcal}{mol}$ and is maintained within the complex (stem II). UCR_3726 , which is of comparable length, is A-U rich and folds in a less stable structure (MFE $= -1.8\frac{kcal}{mol}$). Two-third of UCR_3726 are unpaired within the complex which might be important for the binding process of U5. As noted in Table 5.5 the counts of trusted snRNA loci and the corresponding occurrence of UCR_3726 support this assumption.
At the moment there is no algorithm available to fold more than two alignments at a time. To get an idea if UCR_3726 within the U4 is involved in the binding-process of U5 we used `RNAnofold` (unpublished, personal communication with Stephan Bernhart). `RNAnofold` is an extension of `RNAcofold` which is able to handle three sequences and calculates the MFE structure of the complex. As speculated before the unstructured UCR_3726 works as a contact point for U5. The U4-U6 complex is opened and U5 is able to bind. The complex of these three snRNAs indicates two possible binding areas to form the major spliceosom. U6 is open within the unpaired 19 nucleotides at the top and UCR_18168 might play a similar role as UCR_3726 does.

The next set of multiple occurring UCRs covers 100 of 292 annotated tRNAs within *D. melanogaster*. The length of these UCR elements within tRNA loci range from 14 up to 84 nucleotides with a mean of 51nt. Annotated tRNAs have a mean length of 76 nucleotides (range from 71nt to 186nt). The sequence conservation of tRNAs is a well known and common feature. With respect to the variable sequence positions within a tRNA a combination of adjacent UCRs covers one gene locus. This observation is consistent to the previous findings in Section 5.3.

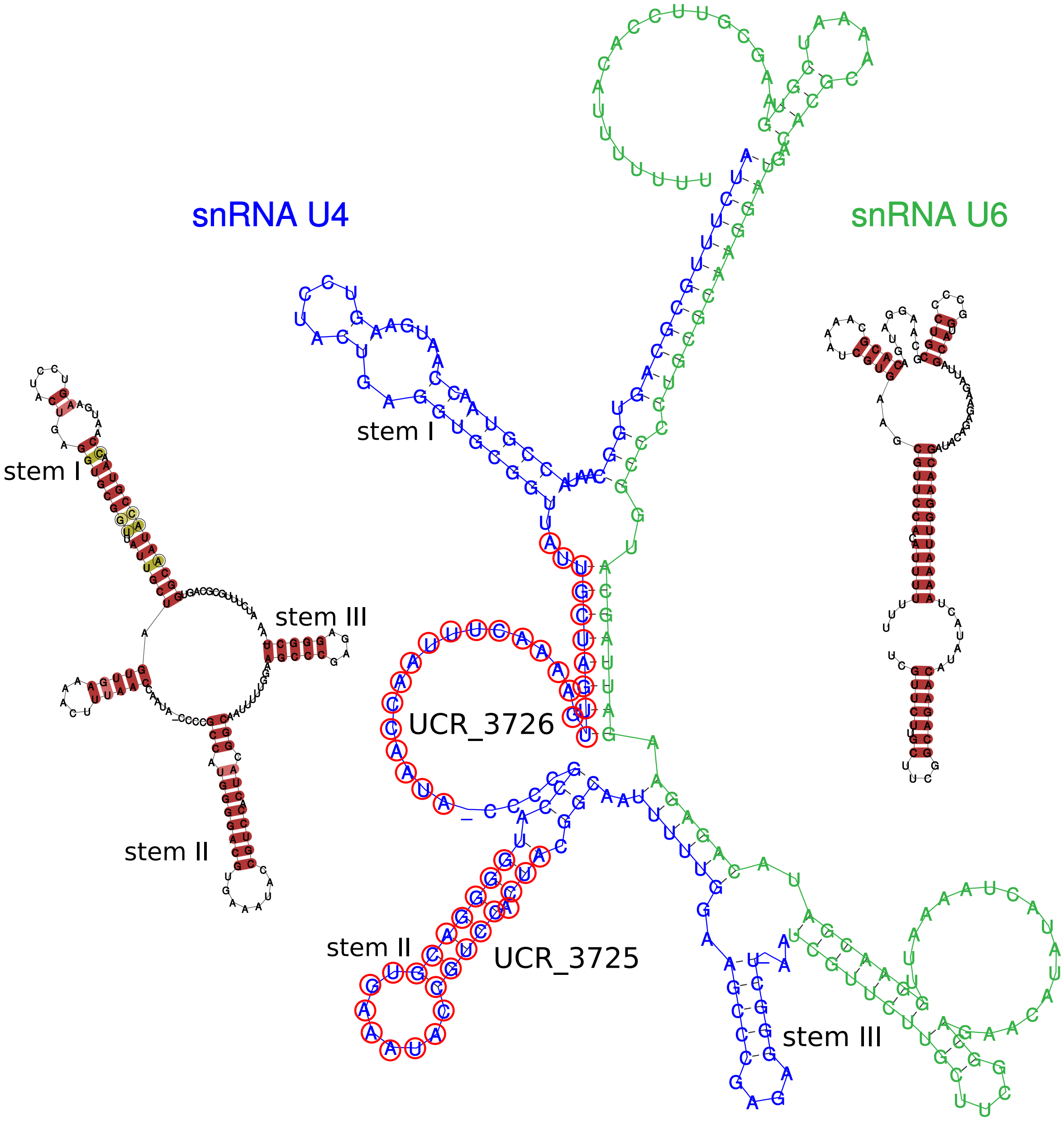


Figure 5.18: Secondary structures of U4 (left) and U6 (right) separately and folded as a complex (center). The red encircled nucleotides mark the UCR elements within the U4 snRNA. U6 is completely covered by UCR_27191.

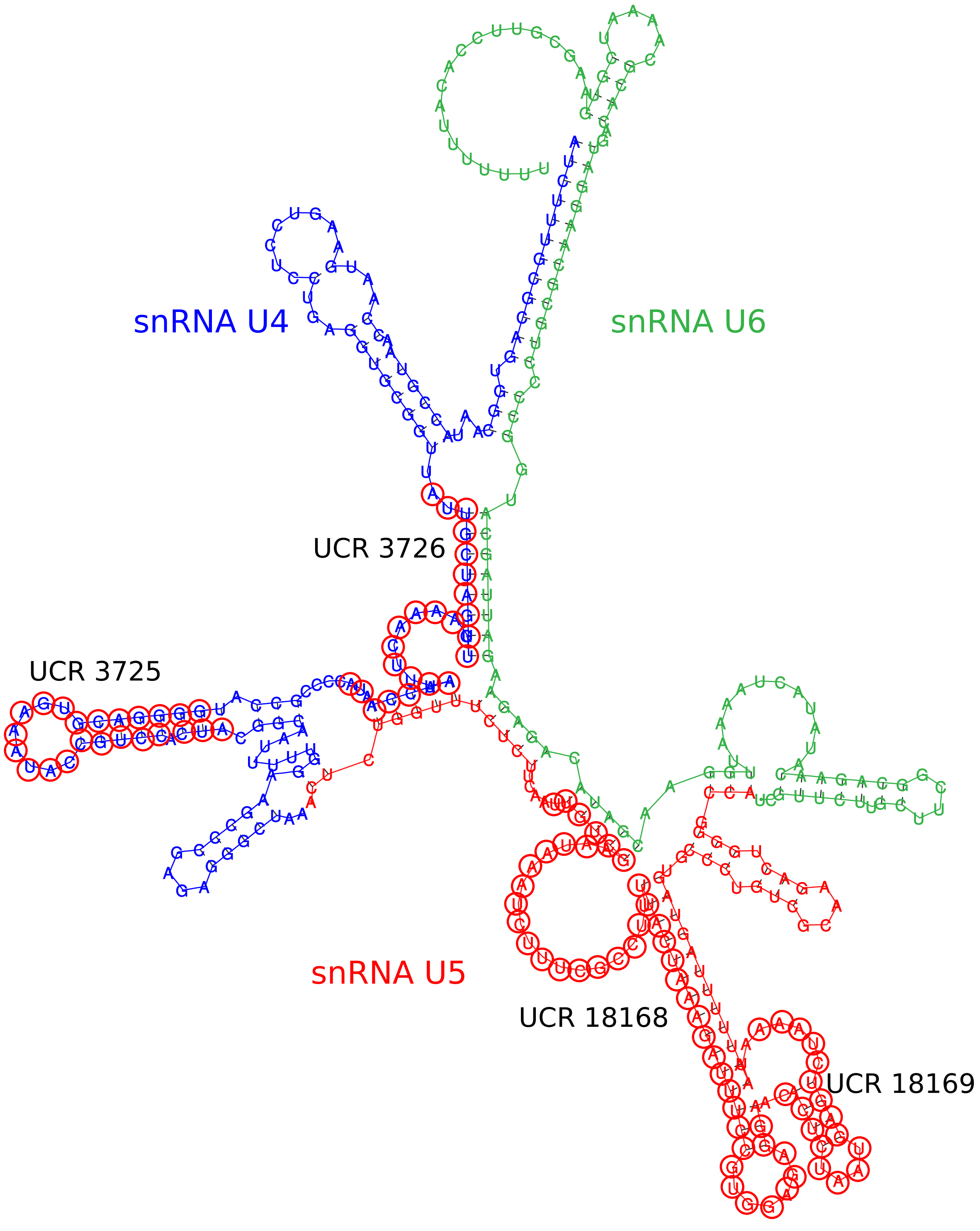


Figure 5.19: Secondary structure of the U4-U5-U6 complex. The red encircled nucleotides mark UCRs within U4 and U5. UCR_27191 covers the complete U6 snRNA locus.

For the analysis of the third class of short multiple occurring UCR sequences the `REDfly database version 2.0` is used. According to the `REDfly database version 2.0` web page:

> "`REDfly` is a curated collection of known *Drosophila* transcriptional cis-regulatory modules (CRMs) and transcription factor binding sites (TFBSs). Despite more than 20 years of experimental determination of these elements, the data have never been collected into a single searchable database. `REDfly` seeks to include all experimentally verified fly regulatory elements along with their DNA sequence, their associated genes, and the expression patterns they direct. Expression patterns are annotated using a defined anatomy ontology to enable high interoperability with `FlyBase`, `FlyExpress`, the `BDGP` in situ hybridization database, and other model organism resources."

An overlap of annotated CRMs and the set of multiple UCRs is used. Sixty-four UCRs overlap CRM annotated regions with a minimum of 15nt and maximum of 27nt (mean 16nt). This result is consistent with the postulated hypothesis that cis-regulatory elements are typically less than 50nt long [53]. Since cluster of cis-regulatory elements are aranged within CRMs many of the short UCR sequences might have cis-regulatory functions. One third of the known CRMs (664) are associated with short UCR-sequences. The detected 64 UCR sequences within this class map 239 times to the *D. melanogaster* genome and fall into 220 annotated CRMs. Indeed the number of 239 UCR related loci could not elucidate the large number of short multiple UCRs. However, it is known that *D. melanogaster* extensively lacks annotation for regulatory elements [69].

Glazov et al. [1] reported only 19 TFBSs associated elements within their UCR set (minimum length 50nt). To compare our results with these previous findings the overlap of the complete UCR set (multiple as well as unique occurring UCRs) with up to date TFBS annotations of the `REDfly database version 2.0` is calculated. Only 99 UCRs (9 multiple, 90 unique) are associated with known TFBSs. At least nine of these UCR sequences have a length greater than 49nt. With respect to the strict UCR definition within this thesis our results are in agreement with the reported findings of Glazov et al. [1].

Discussion:

This section showed the functional importance of UCRs in the context of non-coding RNAs. For the class of snRNAs, especially U4, U5 and U6, ultra-conserved sequences enable us to distinguish between plausible gene loci and pseudogenes. The reason why we detect UCRs only in three snRNAs could be foundet in the alignment algorithm. The `PECAN` alignments are based on homology maps produced by `MERCATOR`[9]. Since the homology maps are based on exonic annotations the algorithm favours to align protein-coding sequences. The consequence is that small non-coding regions are not aligned if surrounding genes prevent this. Using novel alignment approaches like `NcDNAlign` ([27], submitted) and/or aligning only non-protein-coding sequences could prevent these effects. Due to the fact that insects contain only few pseudogenes, the distinction between pseudo- and functional genes should be tested in other phyla as well.
The missing annotations of TFBS within the *Drosophila* genus could also be addressed by motif-finding techniques like `MEME` [103] or `AlignACE` [104]. Predictions of these tools in comparison to the large fraction of short multiple occurring UCR sequences might be a powerful and informative set. Xie et al. [51] described in a recently published study an approach to search for regulatory motifs within conserved regions of the human genome. Using this established approach in combination with our dataset will help to discover new regulatory elements within the *D. melanogaster* genome.

Anyway, the same which was previously concluded for vertebrates seems to be true in insects as well: 'Forming an UCR needs more than traditional DNA binding by transcription factors' [6].

[9] `http://www.biostat.wisc.edu/~cdewey/mercator/`

5.5 Multiple functional constraints on UCRs

Hypothesis:

D. melanogaster is a well characterised and annotated species. This leads to the assumption that all possible, in the meaning of known and predictable, annotations are available for this species. Addressing the question why UCRs are ultra-conserved the idea could come in mind, that more features like CDS, UTR and small RNA overlap each other within UCRs (also discussed in [5, 12]). In other words the conservation is caused by purifying selection due to constraints of several overlapping features. For sequence positions with an annotated CDS and a small RNA as an anti-sense transcript the purifying selection and therefore the conservation might be greater than on sequence positions with only one or no feature. One could also assert that alternative spliced genes and/or alternative transcript regions are more regulated than others.

Testing:

To address this hypothesis two comparable and independent sets of alignment positions are needed. The first set is well defined by the alignment columns of the UCR set. For comparison it is not possible to use the complementary set of all other alignment columns (containing 12 species). Conserved (also ultra-conserved) alignment positions are under negative selection because of their assumed functionality. This common assumption prohibits the random selection of any other alignment column. It is obvious, that not conserved (no function) alignment columns are more frequently selected for the non-UCR set and per default ultra-conserved ones (function) for the UCRs. This would lead to the result, that UCR alignment columns are more functional than random chosen (more frequently not conserved) alignment columns.
Programs are available to calculate scores which reflect the purifying selection for each alignment position. For instants the log-odds score of phastOdds (program of the `PhastCons-Package` [12]) was applied to all MSA containing twelve sequences. The `PhastCons-Package` is designed to identify conserved elements in MSA. This approach is founded on a phylo-HMM which estimates two models. One for the feature of interest

(e.g. UCRs) and the other for the background (e.g. four-fold degenerated sites). Given these two models a log-odd score is assignable for each alignment column within a MSA. This score reflects how much more likely an alignment column was under the conserved than under the non-conserved state of the phylo-HMM. The huge amount of conservation within UCRs leads to the result that each non-UCR set only consists of alignment positions with the same conservation (almost 100%). This is the case for all approaches. Obviously, it is easier and faster to extract all alignment positions with 100% conservation than calculating in a time intensive procedure a log-odd score and extract afterwards all perfectly conservered alignment columns.

Finally, we use all MSAs which contain twelve *Drosophilids* sequences and extract the perfectly conserved alignment columns. The set of extracted alignment columns is subdivided into 944,210 UCR and 10,655,983 other columns. To address the hypothesis above we take all available annotations for CDS, 3' UTR, 5'UTR, introns, repeats, small RNAs, transposons and TransFrags into account (for more information see Chapter 4). For each extracted alignment position an annotation vector is prepared. The vector contains 15 elements two for each feature (coding direction dependent) with exception of the TransFrag data-set which has no annotated direction. The feature vector is filled with the count of features i at sequence position j corresponding to the *D. melanogaster* genome annotations.

All in all, we have two independent sets of annotation vectors. One for perfectly conserved UCR columns and the other for all perfectly conserved columns outside the UCR set. To compare both sets we sampled 100 times a subset of 10,000 columns of each set.

Results:

Interestingly, Figure 5.20 depicts the opposite of the expected. On the x-axis the sum over all counts of the annotation-vector is drawn versus the count of sampled observations (y-axis) with annotation-sum x. To get the variation for each point the standard deviation is marked with error bars around the mean. The distinction of the two curves is significant if the difference between two mean values at position x is at least twice of the standard deviation. If Figure 5.20 would display the expecta-

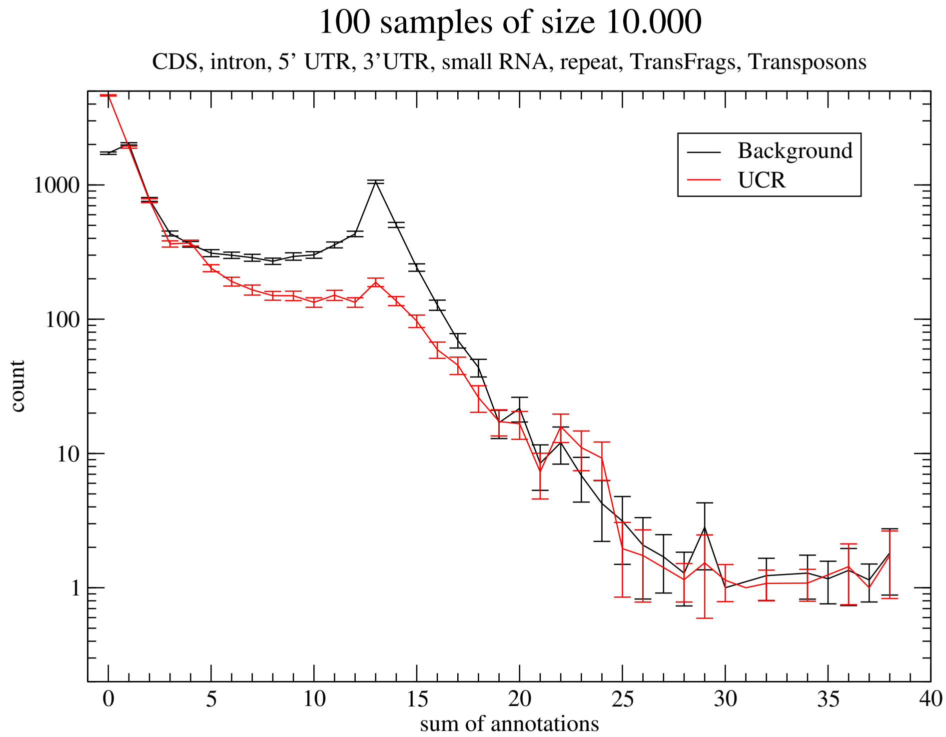


Figure 5.20: Depicted is the sum of annotations for sampled subsets of size 10,000. Standard deviation bars indicate the significance of observable difference for UCRs and the background.

tions described above the UCR curve (red) should be several times greater than the background. Only at x equal to zero the UCR curve is significant greater than the background. This leads to the conclusion that UCRs have more often no known annotation than the background. Furthermore, the background shows significantly more annotations for overlapping annotations from four up to sixteen overlapping features. To verify if these observations are also true for subsets of the annotation vector Figure 5.21 displays the same plot for genic elements (CDS, intron, 5' UTR and 3' UTR) and their constituents. Focusing on the plot of genic annotations two points unequal to zero are of interest. The first spike at x equal to 11 results from the intron annotation which reflects the same spike in the UCR curve. The second spike at x equal to 17 has no apparent corresponding source and therefore it is a subsumed effect of the three considered annotations.

Discussion:

All in all, this section reflects no evidence of ultra-conserved regions just by the overlap of several features within the same region. As mentioned before, we chose the model system of the twelve *Drosophilids* to use a well studied system (especially with respect to *D. melanogaster*) and the currently known genomic elements. One could argue that not all features are known within the *D. melanogaster* genome. Therefore the question would be the other way around: Which model-system should be used to analyze hypothesis like the one mentioned within this section?
Our results support the idea of other mechanisms to keep UCRs constant. Engström et al. [32] showed that highly-conserved elements are clustered and enriched within regions coding for developmental transcriptional regulators. They reported examples of genomic loci were at least two co-expressed developmental regulator genes are spanned by a cluster of highly-conserved elements. The two genes suggest a combined purifying selection on the surrounding cluster. Furthermore it is known that Polycomb binding sites are also enriched around developmental regulators [105]. Combining these information large regulatory regions might be conserved due to several associated and adjacent factors and not by the enrichment local features.

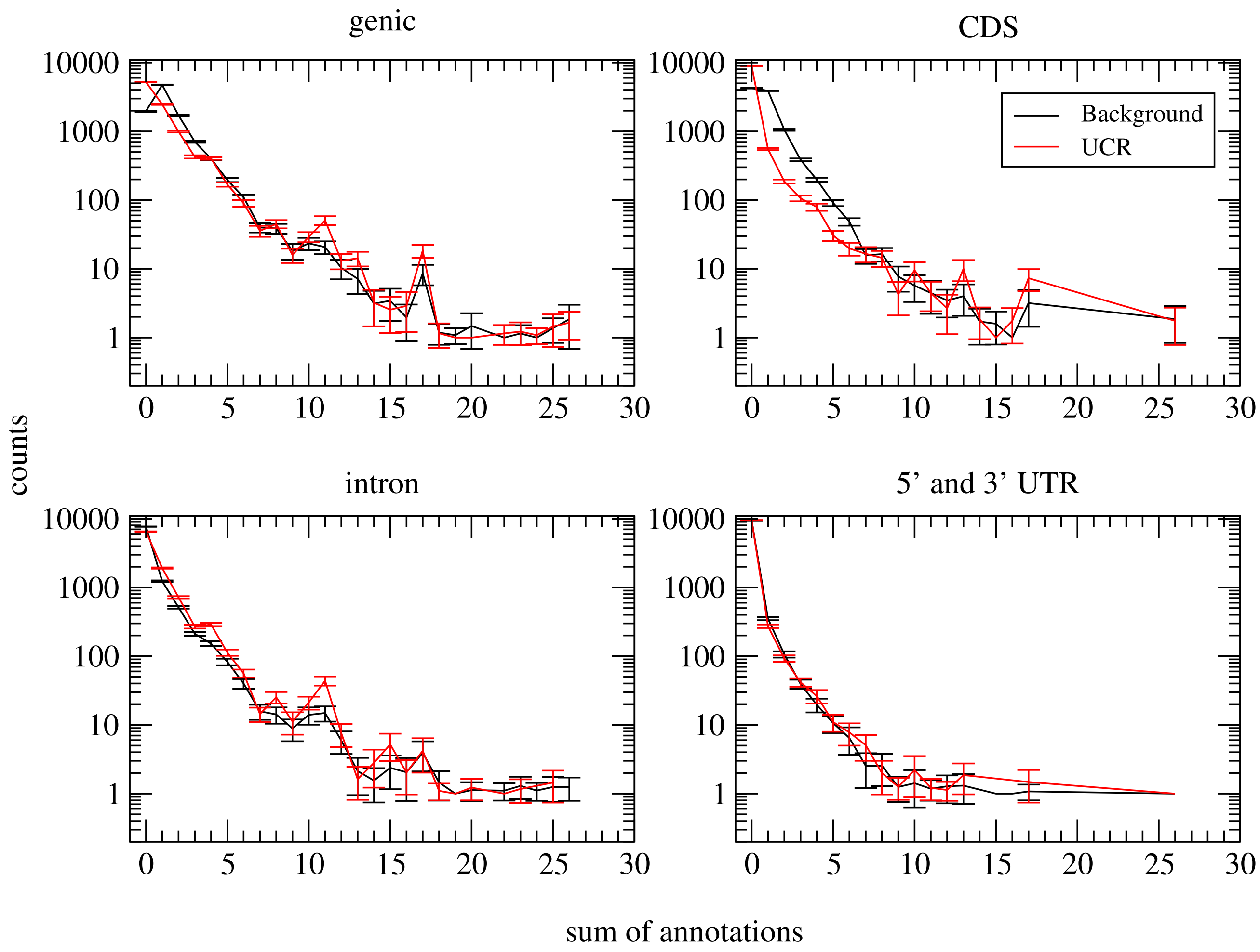


Figure 5.21: The four plots illustrate the sum of annotations for genic regions and their constituents. As observed before the difference of the UCR curves (red) does not support the hypothesis that UCRs could be conserved due to multiple functions (annotations).

Chapter 6

Discussion

In this work the occurrence of UCRs within twelve *Drosophila* genomes was investigated. For the detection of UCRs a novel technique which is based on a seed-extension method (SEM) was constructed. This innovative method enables us to scan multiple sequence alignments (MSA) containing any number of species. The presented SEM is subdivided into three steps. The first and essential step is to search for perfectly conserved blocks of a defined minimum length l. Specifying this minimum length l is a non trivial task. The probability to observe a (perfectly) conserved block of length l by chance within a MSA depends on several factors. Such a measurement is influenced by the algorithm which is used to create the alignment and the context in which a block of length l is observed. Therefore, a (perfectly) conserved block can not be seen as an uncoupled event. So far, there is no adequate measurement for (perfectly) conserved blocks within MSAs available.

Furthermore, the robust and accurate calculation of multiple sequence alignments still a difficult task. Imperfections of the alignment algorithm results in false negative (failure to detect an UCR) as well as false positive (inappropriate prediction of UCRs) detected elements. Therefore, the outcome of the presented SEM as well as any other method applied to detect highly conserved elements strongly depends on the used alignment algorithm. It is important to test the presented SEM on the basis of other input alignments as well as within other phyla.

However, we were able to show that the outcome of our SEM fits very well to previous approaches. Therefore, their findings were applicable to our UCR set as well.

The ability to detect UCRs within MSAs is only a first step. The real challenge is to find the proper meaning of UCRs and the constraints which keep them constant at the sequence level. Exceptional constraints are needed to keep genomic regions constant over million of years. Previous studies showed that this kind of conservation was not explainable by a single functional sequence class.

We addressed the hypothesis that multiple, overlapping sequence features cause the high amount of conservation. If, for instance, a non-coding RNA overlaps with an alternative spliced gene locus the underlying sequence is ultra-selected on structural as well as functional level. Finally, no evidence for UCRs was found by the overlap of several sequence features. Furthermore, UCRs show more often no annotation than the chosen background. This leads to the assumption that not the overlap of local features maintains the extreme amount conservation, but the interaction of closely as well as distantly located elements seems to maintain the conservation of UCRs. Therefore, UCRs could function as 'hubs' within regulatory networks. The interaction of UCRs with regulatory proteins as well as non-coding RNAs could elucidate the high amount of conservation. Many scenarios of interacting elements are imaginable and therefore the detection and analysis of regulatory networks is a difficult task for genic as well as intergenic regions.

Testing of stable secondary structures reveals several UCR sequences which show regulatory features within protein-coding as well as non-coding regions.

For the class of snRNAs, especially U4, U5 and U6, ultra-conserved (sub-)sequences enable us to distinguish between plausible gene loci and pseudogenes. In addition we found several known non-coding RNAs which show (structured) ultra-conserved subregions. Predictions of non-coding RNA loci often showed the problem of a large false positive rate. Filtering of the predictions by additional methods is needed.

A recently published paper about the identification of novel *D. melanogaster* microRNAs used the 'most conserved' track of the `UCSC Genome Browser` in combination with non-coding RNA predictions of `RNAz` [106]. Based on the computational predictions 22 novel miRNAs were found. Instead of the 'most conserved' track (short) UCRs might also help to detect miRNAs as well as other maybe unknown non-coding RNAs.

The combination of the detected UCR set and the prediction of evolutionary conserved non-coding RNA secondary structures will help to extend the annotation of structured elements within *Drosophila*. Using the detected UCR set in combination with the available `RNAz` screen [13] results in subset of candidates which have to be proven by experimental evidence.

The genic proportion of the detected UCR set (48%) indicates also the importance of ultra-selected regions within protein-coding genes. Of course it was reported that regulatory elements, for instance miRNA target sites, cause UCRs within protein-coding genes. Nevertheless, case studies are needed to elucidate region specific functions of UCRs. Case studies of UCRs showed the regulation of alternative splicing as well as nonsense-mediated decay within specific gene loci. An striking example of an ultra-conserved region in the context of intron retention was previously mentioned by Glazov et al. [1]. The corresponding element of comparable length (197nts *D. melanogaster*-*D. pseudoobscura* vs. 183nts twelve *Drosophilids*) was also detected within our dataset. Therefore, the fact that the intron retaining UCR is conserved within all *Drosophilids* is proven.

Additional case studies of other UCRs within protein-coding genes might discover novel regulatory features of these elements. For instance, the mechanism of trans-splicing could also be regulated by UCRs. In *D. melanogaster* the trans-spliced gene mod(mdg4) is known [107, 108, 109]. The term trans-splicing describes a previously 'ignored' mechanism to encode transcripts on both antiparallel DNA strands. The complex protein-coding locus of mod(mdg4) encodes more than 30 isoforms [109]. All transcripts contain four common 5'-exons and at least one alternative 3'-exon. Three of the four common exons are coding exons. Only two UCRs (UCR_25155, 17nt; UCR_25156, 31nt) are observed within this complex gene locus. Interestingly, UCR_25155 overlaps the boundary of the protein-coding sequence within the common exon-2 and UCR_25156 overlaps the boundary of the protein-coding sequences within the common exon-4. Both UCRs contain protein-coding as well as non-coding nucleotides. Therefore, the existence of both elements is not explainable by protein dependent constraints. Since mod(mdg4) is known to be a trans-spliced gene locus both UCRs could help to identify the mechanism of trans-splicing in detail. One or both UCRs could function as sequence motifs which are needed to combine the common coding exons with the alternative exon(s).

In conclusion, it has to be summarized that the detection of highly as well as ultra-conserved regions is only a first step. Within this thesis several hypothesis to elucidate reasons for the existence of UCRs were addressed. Our as well as previous findings indicate a broad functional range of UCRs. Nevertheless, beyond a view anecdotal examples the reason for the existence of UCRs is still unknown. This is manly caused by region specific functions of almost all UCR sequences (87% are unique within *D. melanogaster*). Therefore, further biological as well as bioinformatic analysis are needed to unravel the mystery of UCRs.

Appendix A

Appendix

A.1 Definitions for conserved regions

Publication	Name	Percent Identity (%)	minimum Length (nt)	Alignment	Species
Glazov 2005 [1]	uc-elements	100	≥50	pairwise	*D. melanogaster*, *D. pseudoobscura* and *A. gambia*
Papatsenko 2006 [2]	ultraconserved blocks	100	≥60	pairwise	*D. melanogaster* and *D. virilis*
Retelska 2007 [3]	conserved non-protein-coding (CNC) DNA	≥80	-	pairwise	human, chicken and *D. melanogaster*, *D. virilis*
Bergman 2001 [34]	-	≥70	≥8	pairwise	*D. melanogaster* and *D. virilis*

Publication	Name	Percent Identity (%)	minimum Length (nt)	Alignment	Species
Bergman 2002 [15]	conserved non-coding sequences (CNCSs)	≥90	≥10	pairwise	*D. melanogaster*, *D. erecta*, *D. pseudoobscura*, *D. willistoni*and *Drosophila littoralis*
Tran 2006 [4]	ultraconserved sequences	100	≥19	ternary intersection method, pairwise	insects (*D. melanogaster*, *A. gambia*and *Apis mellifera*) and vertebrates (human, frog, mouse and rat)
Engström 2007 [32]	highly conserved noncoding elements (HCNEs)	≥98	≥50	pairwise	*D. melanogaster*, *D. pseudoobscura*, *D. ananassae*, *D. virilis*, *D. mojavensis*and *A. gambia*
Vavouri 2007 [33]	worm conserved non-coding elements (wCNEs)	-	≥30 consecutive nucleotides identical	ternary intersection using `MegaBlast`	*Caenorhabditis elegans, Caenorhabditis briggsae* and *Caenorhabditis remanei*

Publication	Name	Percent Identity (%)	minimum Length (nt)	Alignment	Species
Margulies 2003 [11]	Multi-species Conserved Sequences (MCSs)	specific scores for MSA	≥25	MSA	human (chromosome 7q31), chimpanzee, baboon, cat, dog, cow, pig, mouse, rat, chicken, fugu and tetraodon
Siepel 2005 [12]	Highly Conserved Elements (HCE)	special log-odds score	-	MSA	four different phyla: vertebrate, arthropods, nematodes and fungi
Bejerano 2004a [5]	ultra-conserved elements (UCE)	100	≥200	pairwise	human, mouse and rat
Bejerano 2004 [20]	-	special p-value	≥50	three-way	human, mouse and rat
Dermitzakis 2002 [30]	conserved non-genic (CNG) regions	≥70	≥100	pairwise	chromosome 21 human and orthologous mouse regions
Woolfe 2005 [31]	conserved non-coding elements (CNE)	≥74	≥100	pairwise	human and pufferfish

Publication	Name	Percent Identity (%)	minimum Length (nt)	Alignment	Species
Sandelin 2004 [6]	ultra conserved region (UCR)	≥95	≥50	pairwise	human, mouse and pufferfish
Drake 2006 [16]	conserved non-coding sequences (CNCs)	70-100	100-200	pairwise and MSA	different sets of vertebrate genomes
Shin 2005 [23]	evolutionary conserved region (ECR)	≥70	≥80	pairwise	human, *Danio rerio* and *Fugu rubripes*

Table A.1: List of previous publications (this list is not exhaustive) about highly conserved regions. Noted are the definitions with the considered measurements of conservation and the minimum length.

A.2 Annotation and sequence databases

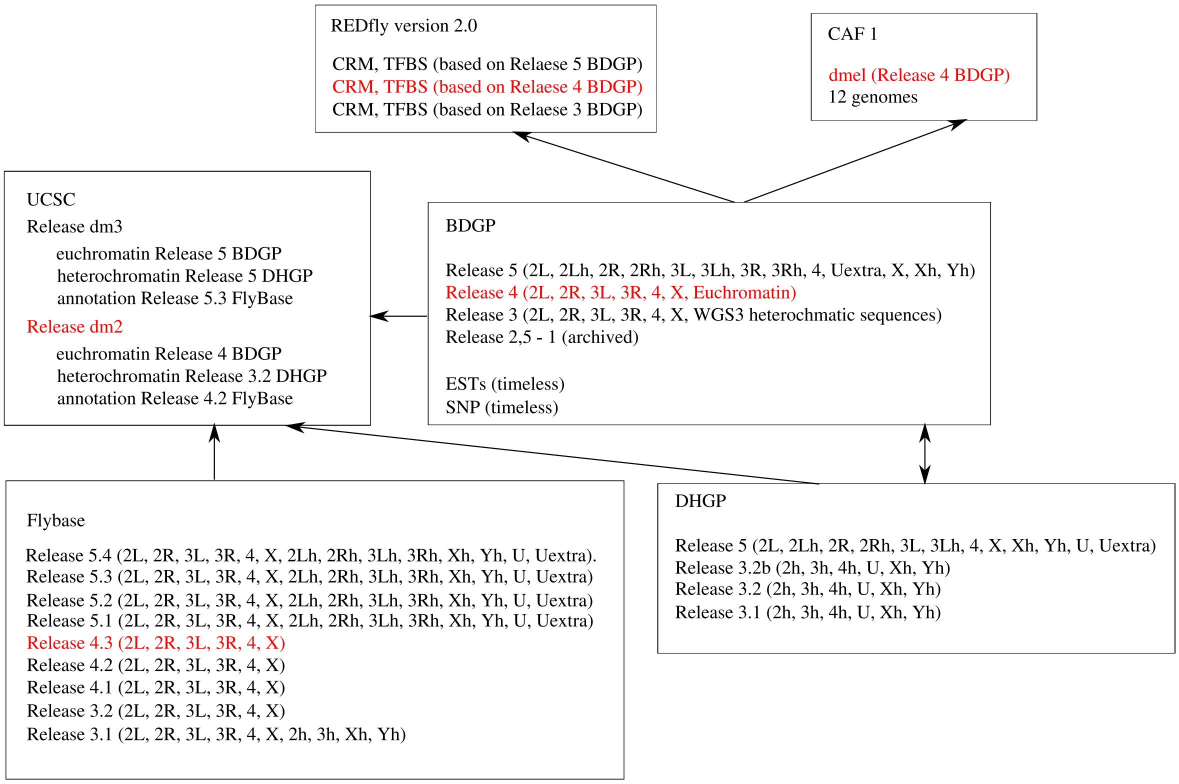


Figure A.1: *Drosophila melanogaster* genomic sequences and the accordingly annotation databases. In here only the red marked sequences and annotations were used. An arrow indicates that database x is based on or applies database y.

A.3 Sequence size of euchromatic chromosome arms

Table A.2: Sequence size of six euchromatic used chromosome arms. First row of each chromosome summarizes the distribution of the chromosome into genic and intergenic regions. The second row displays the same partioning for the corresponding UCR subset.

chromosome	size in bp	genic in bp (%)	intergenic in bp (%)
2L	22407834	12489956 (56%)	9917878 (44%)
	178272	90364 (51%)	87908 (49%)
2R	20766785	12352591 (59%)	8414194 (41%)
	186634	89292 (48%)	97342 (52%)
3L	23771897	13425453 (56%)	10346444 (44%)
	185932	86923 (47%)	99009 (53%)
3R	27905053	16243981 (58%)	11661072 (42%)
	354691	163346 (46%)	191345 (54%)
4	1281640	888725 (69%)	392915 (31%)
	439	439 (100%)	0 (0%)
X	22224390	11975032 (54%)	10249358 (46%)
	58482	30150 (52%)	28332 (48%)
total	118357599	67375738 (57%)	50981861 (43%)
	964450	460514 (48%)	503936 (52%)

A.4 non-coding RNAs

To estimate the degree of similarity between previously known non-coding RNAs (named 'small RNA') and the (un)structured UCRs a regional overlap and sequence similarity is used. The overlap of UCRs associated to annotated non-coding RNAs in *D. melanogaster* is calculated. Either the UCR or the non-coding RNA has to cover more than 90% of each other. A `BLAST` search is applied to identify sequence similarities between (un)structured UCRs and known non-coding RNAs from Rfam version 0.7 [82], NONCODE version 1.0 [83], miRBase version 9.0 [84], snoRNABase version 3[1], ncRNA database[2] and the *D. melanogaster* specific predictions in the tRNA database [85]. To determine the function of UCRs in the context of non-coding RNAs either a minimum of 90% of the UCR or the non-coding RNA have to be identical.

[1] `http://www-snorna.biotoul.fr/index.php`
[2] `http://biobases.ibch.poznan.pl/ncRNA/`

Table A.3: Listing of unstructured UCRs associated with known non-coding RNAs.

	RNA type	**source**	**count**
unstructured CDS	miRNA	miRBase	1
unstructured others	miRNA	miRBase	5
	tRNA	small RNA	1
		Rfam	1
	snoRNA	small RNA	7
		snoRNAdb	4
		Rfam	1
	within *bxd*	small RNA	13
	within *iab*	small RNA	5
	Group I catalytic introns	Rfam	7
	XIST	NONCODE	1
	UR1	NONCODE	1
	KS-1	NONCODE	1
	BORG	NONCODE	1
	unkown	small RNA	7
		ncRNAdb	3
sum	non-coding RNAs	all	59

Table A.4: Listing of structured UCRs associated with known non-coding RNAs.

	RNA type	**source**	**count**
structured CDS	miRNA	miRBase	3
structured others	miRNA	small RNA	42
		miRBase	18
	tRNA	small RNA	35
		tRNAdb	2
	snRNA	small RNA	7
	snoRNA	small RNA	6
		snoRNAdb	1
		Rfam	5
	within *bxd*	small RNA	17
	within *iab*	small RNA	8 (+1miRNA)
	RNAseP	small RNA	3
	Group I catalytic introns	Rfam	2
	signal recognition particle	Rfam	1
	T-box leader element	Rfam	1
	His-1	NONCODE	1
	7H4	NONCODE	1
	PCGEM1	NONCODE	1
	KS-1	NONCODE	1
	unkown	small RNA	8
		ncRNAdb	3
sum	non-coding RNAs	all	166

A.5 snRNAs

```
              5    10   15   20   25   30   35   40
U4.1      ATCTTTGCGCAGAGGCGATATCGTAACCAATGAAGTTCTA-CTGA
U4.2      ATCTTTGCGCAGTGGCAATACCGTAACCAATGAAGTCCTC-CTGA
U4.3      AACCTTGTGCAGTGGCAACATCGCAAGCAATGAAGTTCCAACTGA
consensus AtCtTTGcGCAGtGGCaAtAtCGtAAcCAATGAAGTtCta-CTGA

                      UCR_3726
          45   50   55   60   65   70   75   80   85
U4.1      GGTGCGATTATTGCTAGTTGAAAACTTTAACCAATACCCCGCC--
U4.2      GGTGCGGTTATTGCTAGTTGAAAACTTTAACCAATACCCCGCC--
U4.3      GCTGCGATTATTGCTAGTTGAAAACTAAAACCAATATCTCGCCCA
consensus GgTGCGaTTATTGCTAGTTGAAAACTttAACCAATAcCcCGCC--

                  UCR_3725
              90   95   100  105  110  115  120  125  130
U4.1      --ATGGGGACGTGAAATACCGTCCACTACGGCAATTTTTGGAAGC
U4.2      --ATGGGGACGTGAAATACCGTCCACTACGGCAATTTTTGGAAGC
U4.3      GCGTAAGGATCTACGAT-CTTTAAGCTAAGGCAATTTTTTTAGGC
consensus --aTggGGAcgTgaaATaCcgTccaCTAcGGCAATTTTTggAaGC

              135  140
U4.1      CC-GAGAGGGCCA-
U4.2      CC-GAGAGGGCTAA
U4.3      CCCAAGTGGGCTGA
consensus CC-gAGaGGGCtaa
```

Figure A.2: Alignment of the three U4 snRNA loci U4.1 (FBgn0003931; chr2L:19806663-19806804) U4.2 (FBgn0003930; chr2L:21202629-21202771) and U4.3 (FBgn0041719; chr2L:5565619-5565766) within the *Drosophila melanogaster*genome. At locus U4.2 two UCRsare detected and used to search for the other two U4 loci.

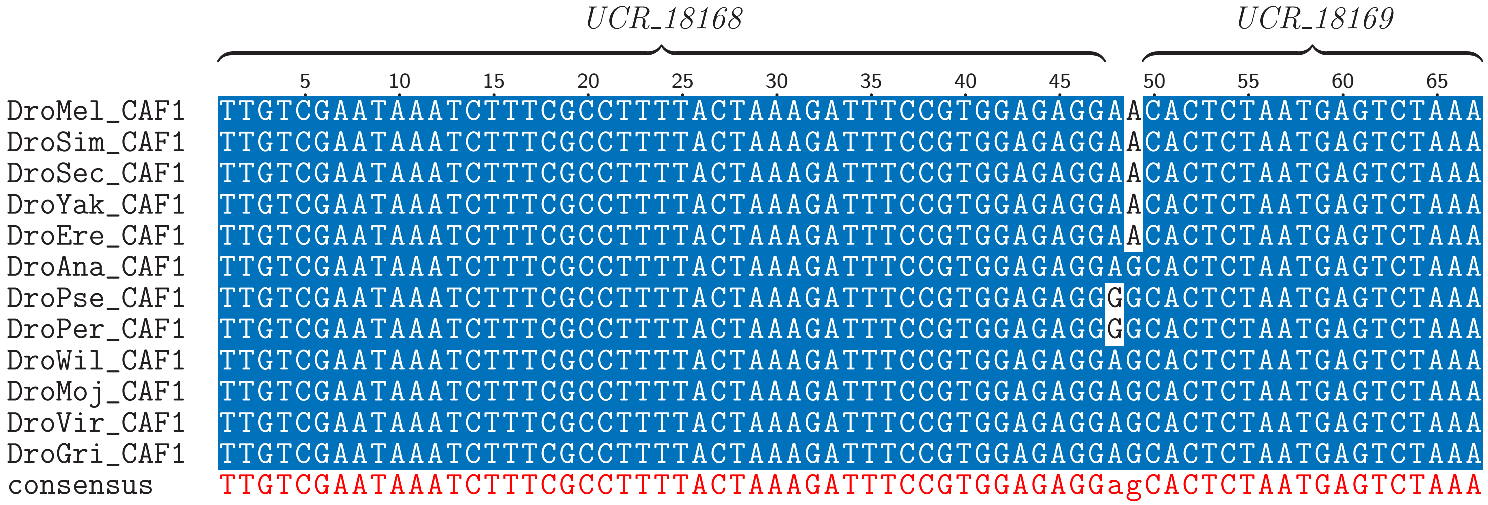


Figure A.3: As mentioned in Section 5.4 the snRNA U5 covers two UCR elements (UCR_18168 and UCR_18169). These are separated in all genomes by at least two nucleotides. The nucleotide composition of alignment column 49 causes the separation of both UCRs. Only alignment columns with not more than two consensus unequal characters are permitted (see Section 4.2). In case of column 49 the *melanogaster* subgroup has an A instead the consensus G of remaining species.

List of Figures

List of Tables

Bibliography

[1] E. A. Glazov, M. Pheasant, E. A. McGraw, G. Bejerano & J. S. Mattick. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res*, 15(6):800–808, 2005

[2] D. Papatsenko, A. Kislyuk, M. Levine & I. Dubchak. Conservation patterns in different functional sequence categories of divergent Drosophila species. *Genomics*, 88(4):431–442, 2006

[3] D. Retelska, E. Beaudoing, C. Notredame, C. V. Jongeneel & P. Bucher. Vertebrate conserved non coding DNA regions have a high persistence length and a short persistence time. *BMC Genomics*, 8(1):398, 2007

[4] T. Tran, P. Havlak & J. Miller. MicroRNA enrichment among short 'ultraconserved' sequences in insects. *Nucleic Acids Res*, 34(9):e65, 2006

[5] G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick & D. Haussler. Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325, 2004

[6] A. Sandelin, P. Bailey, S. Bruce, P. G. Engstrm, J. M. Klos, W. W. Wasserman, J. Ericson & B. Lenhard. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, 5(1):99, 2004

[7] K. Tamura, S. Subramanian & S. Kumar. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Mol Biol Evol*, 21(1):36–44, 2004

[8] D. . G. Consortium. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167):203–218, 2007

[9] M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y. H. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. Gabor, J. F. Abril, A. Agbayani, H. J. An, C. Andrews-Pfannkoch, D. Baldwin, R. M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, S. M. Dietz, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. Howland, M. H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G. H. Karpen, Z. Ke, J. A. Kennison, K. A. Ketchum, B. E. Kimmel, C. D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A. A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T. C. McIntosh, M. P. McLeod, D. McPherson, G. Merkulov, N. V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S. M. Mount, M. Moy, B. Murphy, L. Murphy, D. M. Muzny, D. L. Nelson, D. R. Nelson, K. A. Nelson, K. Nixon, D. R. Nusskern, J. M. Pacleb, M. Palazzolo, G. S. Pittman, S. Pan, J. Pollard, V. Puri, M. G. Reese, K. Reinert, K. Remington, R. D. Saunders, F. Scheeler, H. Shen, B. C. Shue, I. Sidn-Kiamos, M. Simpson, M. P. Skupski, T. Smith, E. Spier, A. C. Spradling, M. Stapleton,

R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A. H. Wang, X. Wang, Z. Y. Wang, D. A. Wassarman, G. M. Weinstock, J. Weissenbach, S. M. Williams, WoodageT, K. C. Worley, D. Wu, S. Yang, Q. A. Yao, J. Ye, R. F. Yeh, J. S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X. H. Zheng, F. N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H. O. Smith, R. A. Gibbs, E. W. Myers, G. M. Rubin & J. C. Venter. The genome sequence of Drosophila melanogaster. *Science*, 287(5461):2185–2195, 2000

[10] R. A. Hoskins, C. R. Nelson, B. P. Berman, T. R. Laverty, R. A. George, L. Ciesiolka, M. Naeemuddin, A. D. Arenson, J. Durbin, R. G. David, P. E. Tabor, M. R. Bailey, D. R. DeShazo, J. Catanese, A. Mammoser, K. Osoegawa, P. J. de Jong, S. E. Celniker, R. A. Gibbs, G. M. Rubin & S. E. Scherer. A BAC-based physical map of the major autosomes of Drosophila melanogaster. *Science*, 287(5461):2271–2274, 2000

[11] E. H. Margulies, M. Blanchette, N. I. S. C. C. S. Program, D. Haussler & E. D. Green. Identification and characterization of multi-species conserved sequences. *Genome Res*, 13(12):2507–2518, 2003

[12] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller & D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–1050, 2005

[13] D. Rose, J. Hackermueller, S. Washietl, K. Reiche, J. Hertel, S. Findeiss, P. Stadler & S. Prohaska. Computational RNomics of Drosophilids. *BMC Genomics*, 8(1):406, 2007

[14] A. G. Clark. The search for meaning in noncoding DNA. *Genome Res*, 11(8):1319–1320, 2001

[15] C. M. Bergman, B. D. Pfeiffer, D. E. Rincn-Limas, R. A. Hoskins, A. Gnirke, C. J. Mungall, A. M. Wang, B. Kronmiller, J. Pacleb, S. Park, M. Stapleton, K. Wan, R. A. George, P. J. de Jong, J. Botas, G. M. Rubin & S. E. Celniker.

Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. *Genome Biol*, 3(12):RESEARCH0086, 2002

[16] J. A. Drake, C. Bird, J. Nemesh, D. J. Thomas, C. Newton-Cheh, A. Reymond, L. Excoffier, H. Attar, S. E. Antonarakis, E. T. Dermitzakis & J. N. Hirschhorn. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat Genet*, 38(2):223–227, 2006

[17] S. Katzman, A. D. Kern, G. Bejerano, G. Fewell, L. Fulton, R. K. Wilson, S. R. Salama & D. Haussler. Human genome ultraconserved elements are ultraselected. *Science*, 317(5840):915, 2007

[18] M. G. S. Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002

[19] R. A. Gibbs, G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren, S. Scherer, G. Scott, D. Steffen, K. C. Worley, P. E. Burch, G. Okwuonu, S. Hines, L. Lewis, C. DeRamo, O. Delgado, S. Dugan-Rocha, G. Miner, M. Morgan, A. Hawes, R. Gill, Celera, R. A. Holt, M. D. Adams, P. G. Amanatides, H. Baden-Tillson, M. Barnstead, S. Chin, C. A. Evans, S. Ferriera, C. Fosler, A. Glodek, Z. Gu, D. Jennings, C. L. Kraft, T. Nguyen, C. M. Pfannkoch, C. Sitter, G. G. Sutton, J. C. Venter, T. Woodage, D. Smith, H.-M. Lee, E. Gustafson, P. Cahill, A. Kana, L. Doucette-Stamm, K. Weinstock, K. Fechtel, R. B. Weiss, D. M. Dunn, E. D. Green, R. W. Blakesley, G. G. Bouffard, P. J. D. Jong, K. Osoegawa, B. Zhu, M. Marra, J. Schein, I. Bosdet, C. Fjell, S. Jones, M. Krzywinski, C. Mathewson, A. Siddiqui, N. Wye, J. McPherson, S. Zhao, C. M. Fraser, J. Shetty, S. Shatsman, K. Geer, Y. Chen, S. Abramzon, W. C. Nierman, P. H. Havlak, R. Chen, K. J. Durbin, A. Egan, Y. Ren, X.-Z. Song, B. Li, Y. Liu, X. Qin, S. Cawley, K. C. Worley, A. J. Cooney, L. M. D'Souza, K. Martin, J. Q. Wu, M. L. Gonzalez-Garay, A. R. Jackson, K. J. Kalafus, M. P. McLeod, A. Milosavljevic, D. Virk, A. Volkov, D. A. Wheeler, Z. Zhang, J. A. Bailey, E. E. Eichler, E. Tuzun, E. Birney, E. Mongin, A. Ureta-Vidal, C. Woodwark, E. Zdobnov, P. Bork, M. Suyama, D. Torrents, M. Alexandersson, B. J. Trask,

J. M. Young, H. Huang, H. Wang, H. Xing, S. Daniels, D. Gietzen, J. Schmidt, K. Stevens, U. Vitt, J. Wingrove, F. Camara, M. M. Alb, J. F. Abril, R. Guigo, A. Smit, I. Dubchak, E. M. Rubin, O. Couronne, A. Poliakov, N. Hbner, D. Ganten, C. Goesele, O. Hummel, T. Kreitler, Y.-A. Lee, J. Monti, H. Schulz, H. Zimdahl, H. Himmelbauer, H. Lehrach, H. J. Jacob, S. Bromberg, J. Gullings-Handley, M. I. Jensen-Seaman, A. E. Kwitek, J. Lazar, D. Pasko, P. J. Tonellato, S. Twigger, C. P. Ponting, J. M. Duarte, S. Rice, L. Goodstadt, S. A. Beatson, R. D. Emes, E. E. Winter, C. Webber, P. Brandt, G. Nyakatura, M. Adetobi, F. Chiaromonte, L. Elnitski, P. Eswara, R. C. Hardison, M. Hou, D. Kolbe, K. Makova, W. Miller, A. Nekrutenko, C. Riemer, S. Schwartz, J. Taylor, S. Yang, Y. Zhang, K. Lindpaintner, T. D. Andrews, M. Caccamo, M. Clamp, L. Clarke, V. Curwen, R. Durbin, E. Eyras, S. M. Searle, G. M. Cooper, S. Batzoglou, M. Brudno, A. Sidow, E. A. Stone, J. C. Venter, B. A. Payseur, G. Bourque, C. Lpez-Otn, X. S. Puente, K. Chakrabarti, S. Chatterji, C. Dewey, L. Pachter, N. Bray, V. B. Yap, A. Caspi, G. Tesler, P. A. Pevzner, D. Haussler, K. M. Roskin, R. Baertsch, H. Clawson, T. S. Furey, A. S. Hinrichs, D. Karolchik, W. J. Kent, K. R. Rosenbloom, H. Trumbower, M. Weirauch, D. N. Cooper, P. D. Stenson, B. Ma, M. Brent, M. Arumugam, D. Shteynberg, R. R. Copley, M. S. Taylor, H. Riethman, U. Mudunuri, J. Peterson, M. Guyer, A. Felsenfeld, S. Old, S. Mockrin, F. Collins & R. G. S. P. Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–521, 2004

[20] G. Bejerano, D. Haussler & M. Blanchette. Into the heart of darkness: large-scale clustering of human non-coding DNA. *Bioinformatics*, 20 Suppl 1:I40–I48, 2004

[21] T. F. Smith & M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, 1981

[22] S. B. Needleman & C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, 1970

[23] J. T. Shin, J. R. Priest, I. Ovcharenko, A. Ronco, R. K. Moore, C. G. Burns & C. A. MacRae. Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. *Nucleic Acids Res*, 33(17):5437–5445, 2005

[24] N. Bray & L. Pachter. MAVID multiple alignment server. *Nucleic Acids Res*, 31(13):3525–3526, 2003

[25] N. Bray & L. Pachter. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res*, 14(4):693–699, 2004

[26] M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler & W. Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, 14(4):708–715, 2004

[27] D. Rose, J. Hertel, K. Reiche, P. F. Stadler & J. Hackermller. NcDNAlign - Plausible Multiple Alignments of Non-Protein-Coding Genomic sequences. 2007. Submitted

[28] E. H. Margulies, C. W. Chen & E. D. Green. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends Genet*, 22(4):187–193, 2006

[29] S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler & W. Miller. Human-mouse alignments with BLASTZ. *Genome Res*, 13(1):103–107, 2003

[30] E. T. Dermitzakis, A. Reymond, R. Lyle, N. Scamuffa, C. Ucla, S. Deutsch, B. J. Stevenson, V. Flegel, P. Bucher, C. V. Jongeneel & S. E. Antonarakis. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature*, 420(6915):578–582, 2002

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

[32] P. G. Engstrm, S. J. H. Sui, O. Drivenes, T. S. Becker & B. Lenhard. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res*, 2007

[33] T. Vavouri, K. Walter, W. R. Gilks, B. Lehner & G. Elgar. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biol*, 8(2):R15, 2007

[34] C. M. Bergman & M. Kreitman. Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. *Genome Res*, 11(8):1335–1345, 2001

[35] L. A. Pennacchio, N. Ahituv, A. M. Moses, S. Prabhakar, M. A. Nobrega, M. Shoukry, S. Minovitsky, I. Dubchak, A. Holt, K. D. Lewis, I. Plajzer-Frick, J. Akiyama, S. D. Val, V. Afzal, B. L. Black, O. Couronne, M. B. Eisen, A. Visel & E. M. Rubin. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502, 2006

[36] G. Bejerano, C. B. Lowe, N. Ahituv, B. King, A. Siepel, S. R. Salama, E. M. Rubin, W. J. Kent & D. Haussler. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, 441(7089):87–90, 2006

[37] F. Poulin, M. A. Nobrega, I. Plajzer-Frick, A. Holt, V. Afzal, E. M. Rubin & L. A. Pennacchio. In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics*, 85(6):774–781, 2005

[38] H. G. Belting, C. S. Shashikant & F. H. Ruddle. Modification of expression and cis-regulation of Hoxc8 in the evolution of diverged axial morphology. *Proc Natl Acad Sci U S A*, 95(5):2355–2360, 1998

[39] B. Cohen, M. E. McGuffin, C. Pfeifle, D. Segal & S. M. Cohen. apterous, a gene required for imaginal disc development in Drosophila encodes a member of the LIM family of developmental regulatory proteins. *Genes Dev*, 6(5):715–729, 1992

[40] N. Ahituv, Y. Zhu, A. Visel, A. Holt, V. Afzal, L. A. Pennacchio & E. M. Rubin. Deletion of ultraconserved elements yields viable mice. *PLoS Biol*, 5(9):e234, 2007

[41] J. Z. Ni, L. Grate, J. P. Donohue, C. Preston, N. Nobida, G. O'Brien, L. Shiue, T. A. Clark, J. E. Blume & M. Ares. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev*, 21(6):708–718, 2007

[42] E. A. Glazov, M. Pheasant, S. Nahkuri & J. S. Mattick. Evidence for control of splicing by alternative RNA secondary structures in Dipteran homothorax pre-mRNA. *RNA Biol*, 3(1):36–39, 2006

[43] J. Feng, C. Bi, B. S. Clark, R. Mady, P. Shah & J. D. Kohtz. The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev*, 20(11):1470–1484, 2006

[44] G. A. Calin, C.-G. Liu, M. Ferracin, T. Hyslop, R. Spizzo, C. Sevignani, M. Fabbri, A. Cimmino, E. J. Lee, S. E. Wojcik, M. Shimizu, E. Tili, S. Rossi, C. Taccioli, F. Pichiorri, X. Liu, S. Zupo, V. Herlea, L. Gramantieri, G. Lanza, H. Alder, L. Rassenti, S. Volinia, T. D. Schmittgen, T. J. Kipps, M. Negrini & C. M. Croce. Ultraconserved Regions Encoding ncRNAs Are Altered in Human Leukemias and Carcinomas. *Cancer Cell*, 12(3):215–229, 2007

[45] G. Panganiban & J. L. R. Rubenstein. Developmental functions of the Distal-less/Dlx homeobox genes. *Development*, 129(19):4371–4386, 2002

[46] D. Beysen, J. Raes, B. P. Leroy, A. Lucassen, J. R. W. Yates, J. Clayton-Smith, H. Ilyina, S. S. Brooks, S. Christin-Maitre, M. Fellous, J. P. Fryns, J. R. Kim, P. Lapunzina, E. Lemyre, F. Meire, L. M. Messiaen, C. Oley, M. Splitt, J. Thomson, Y. V. de Peer, R. A. Veitia, A. D. Paepe & E. D. Baere. Deletions involving long-range conserved nongenic sequences upstream and downstream of FOXL2 as a novel disease-causing mechanism in blepharophimosis syndrome. *Am J Hum Genet*, 77(2):205–218, 2005

[47] E. Richler, J. G. Reichert, J. D. Buxbaum & L. A. McInnes. Autism and ultra-conserved non-coding sequence on chromosome 7q. *Psychiatr Genet*, 16(1):19–23, 2006

[48] M. Ban, M. Maranian, T. W. Yeo, J. Gray, A. Compston & S. Sawcer. Ultraconserved regions in multiple sclerosis. *Eur J Hum Genet*, 13(9):998–999, 2005

[49] A. F. Bompfunewerer, C. Flamm, C. Fried, G. Fritzsch, I. L. Hofacker, J. Lehmann, K. Missal, A. Mosig, B. Muller & S. J. Prohaska. Evolutionary patterns of non-coding rnas. *Theory in Biosciences*, 123(4):301–369, 2005

[50] X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander & M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, 2005

[51] X. Xie, T. S. Mikkelsen, A. Gnirke, K. Lindblad-Toh, M. Kellis & E. S. Lander. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A*, 104(17):7145–7150, 2007

[52] A. Stark, M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, J. G. Ruby, J. Brennecke, H. F. curators, B. D. G. Project, E. Hodges, A. S. Hinrichs, A. Caspi, B. Paten, S.-W. Park, M. V. Han, M. L. Maeder, B. J. Polansky, B. E. Robson, S. Aerts, J. van Helden, B. Hassan, D. G. Gilbert, D. A. Eastman, M. Rice, M. Weir, M. W. Hahn, Y. Park, C. N. Dewey, L. Pachter, W. J. Kent, D. Haussler, E. C. Lai, D. P. Bartel, G. J. Hannon, T. C. Kaufman, M. B. Eisen, A. G. Clark, D. Smith, S. E. Celniker, W. M. Gelbart, M. Kellis, M. A. Crosby, B. B. Matthews, A. J. Schroeder, L. S. Gramates, S. E. S. Pierre, M. Roark, K. L. Wiley, R. J. Kulathinal, P. Zhang, K. V. Myrick, J. V. Antone, W. M. Gelbart, J. W. Carlson, C. Yu, S. Park, K. H. Wan & S. E. Celniker. Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature*, 450(7167):219–232, 2007

[53] S. Richards, Y. Liu, B. R. Bettencourt, P. Hradecky, S. Letovsky, R. Nielsen, K. Thornton, M. J. Hubisz, R. Chen, R. P. Meisel, O. Couronne, S. Hua, M. A. Smith, P. Zhang, J. Liu, H. J. Bussemaker, M. F. van Batenburg, S. L. Howells, S. E. Scherer, E. Sodergren, B. B. Matthews, M. A. Crosby, A. J. Schroeder,

D. Ortiz-Barrientos, C. M. Rives, M. L. Metzker, D. M. Muzny, G. Scott, D. Steffen, D. A. Wheeler, K. C. Worley, P. Havlak, K. J. Durbin, A. Egan, R. Gill, J. Hume, M. B. Morgan, G. Miner, C. Hamilton, Y. Huang, L. Waldron, D. Verduzco, K. P. Clerc-Blankenburg, I. Dubchak, M. A. F. Noor, W. Anderson, K. P. White, A. G. Clark, S. W. Schaeffer, W. Gelbart, G. M. Weinstock & R. A. Gibbs. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res*, 15(1):1–18, 2005

[54] S. Guindon & O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5):696–704, 2003

[55] E. N. C. O. D. E. P. Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, 2007

[56] G. M. Rubin, M. D. Yandell, J. R. Wortman, G. L. G. Miklos, C. R. Nelson, I. K. Hariharan, M. E. Fortini, P. W. Li, R. Apweiler, W. Fleischmann, J. M. Cherry, S. Henikoff, M. P. Skupski, S. Misra, M. Ashburner, E. Birney, M. S. Boguski, T. Brody, P. Brokstein, S. E. Celniker, S. A. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R. F. Galle, W. M. Gelbart, R. A. George, L. S. Goldstein, F. Gong, P. Guan, N. L. Harris, B. A. Hay, R. A. Hoskins, J. Li, Z. Li, R. O. Hynes, S. J. Jones, P. M. Kuehl, B. Lemaitre, J. T. Littleton, D. K. Morrison, C. Mungall, P. H. O'Farrell, O. K. Pickeral, C. Shue, L. B. Vosshall, J. Zhang, Q. Zhao, X. H. Zheng & S. Lewis. Comparative genomics of the eukaryotes. *Science*, 287(5461):2204–2215, 2000

[57] S. E. Celniker, D. A. Wheeler, B. Kronmiller, J. W. Carlson, A. Halpern, S. Patel, M. Adams, M. Champe, S. P. Dugan, E. Frise, A. Hodgson, R. A. George, R. A. Hoskins, T. Laverty, D. M. Muzny, C. R. Nelson, J. M. Pacleb, S. Park, B. D. Pfeiffer, S. Richards, E. J. Sodergren, R. Svirskas, P. E. Tabor, K. Wan, M. Stapleton, G. G. Sutton, C. Venter, G. Weinstock, S. E. Scherer, E. W. Myers, R. A. Gibbs & G. M. Rubin. Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. *Genome Biol*, 3(12):RESEARCH0079, 2002

[58] G. Grumbling & V. Strelets. FlyBase: anatomical data, images and queries. *Nucleic Acids Res*, 34(Database issue):D484–D488, 2006

[59] M. A. Crosby, J. L. Goodman, V. B. Strelets, P. Zhang, W. M. Gelbart & F. Consortium. FlyBase: genomes by the dozen. *Nucleic Acids Res*, 35(Database issue):D486–D491, 2007

[60] B. V. Emden, H. Ramos, S. Panchanathan, S. Newfeld & S. Kumar. Flyexpress: An image-matching web-tool for finding genes with overlapping patterns of expression in drosophila embryos. Technical report, Arizona State University, Tempe, Arizona 85287-5301, USA, 2006

[61] B. Ewing, L. Hillier, M. C. Wendl & P. Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, 8(3):175–185, 1998

[62] R. A. Drysdale, M. A. Crosby & F. Consortium. FlyBase: genes and gene models. *Nucleic Acids Res*, 33(Database issue):D390–D395, 2005

[63] A. Smit, R. Hubley & P. Green. Repeatmasker open-3.0, 1996-2004. `http://www.repeatmasker.org`

[64] G. Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, 27(2):573–580, 1999

[65] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, W. J. Kent & U. of California Santa Cruz. The UCSC Genome Browser Database. *Nucleic Acids Res*, 31(1):51–54, 2003

[66] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler & W. J. Kent. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, 32(Database issue):D493–D496, 2004

[67] J. R. Manak, S. Dike, V. Sementchenko, P. Kapranov, F. Biemar, J. Long, J. Cheng, I. Bell, S. Ghosh, A. Piccolboni & T. R. Gingeras. Biological func-

tion of unannotated transcription during the early development of Drosophila melanogaster. *Nat Genet*, 38(10):1151–1158, 2006

[68] Y. Fofanov, Y. Luo, C. Katili, J. Wang, Y. Belosludtsev, T. Powdrill, C. Belapurkar, V. Fofanov, T.-B. Li, S. Chumakov & B. M. Pettitt. How independent are the appearances of n-mers in different genomes? *Bioinformatics*, 20(15):2421–2428, 2004

[69] S. M. Gallo, L. Li, Z. Hu & M. S. Halfon. REDfly: a Regulatory Element Database for Drosophila. *Bioinformatics*, 22(3):381–383, 2006

[70] J. R. Powell. Progress and Prospects in Evolutionary Biology : The Drosophila Model (Oxford Series in Ecology & Evolution). Oxford University Press, USA, 1997

[71] M. Brudno, C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, N. I. S. C. C. S. Program, E. D. Green, A. Sidow & S. Batzoglou. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*, 13(4):721–731, 2003

[72] S. Vicario, E. Moriyama & J. Powell. Codon usage in twelve species of drosophila. *BMC Evol Biol*, 7(1):226, 2007

[73] T. Warnecke & L. D. Hurst. Evidence for a Trade-Off between Translational Efficiency and Splicing Regulation in Determining Synonymous Codon Usage in Drosophila melanogaster. *Mol Biol Evol*, 2007

[74] J. S. Mattick. A new paradigm for developmental biology. *J Exp Biol*, 210(Pt 9):1526–1547, 2007

[75] C. Flamm, I. Hofacker & P. Stadler. Computational chemistry with RNA secondary structures. 2004

[76] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker & P. Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatsh Chem*, 125:167–188, 1994

[77] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31(13):3406–3415, 2003

[78] M. Zuker & P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–148, 1981

[79] P. Schuster, W. Fontana, P. F. Stadler & I. L. Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. *Proc Biol Sci*, 255(1344):279–284, 1994

[80] E. Bonnet, J. Wuyts, P. Rouz & Y. V. de Peer. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20(17):2911–2917, 2004

[81] E. S. Edgington. Randomization tests, volume 77. Marcel Dekker, Inc., New York, NY, USA, 1986. 0-824-77656-9

[82] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy & A. Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, 33(Database issue):D121–D124, 2005

[83] C. Liu, B. Bai, G. Skogerb, L. Cai, W. Deng, Y. Zhang, D. Bu, Y. Zhao & R. Chen. NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res*, 33(Database issue):D112–D115, 2005

[84] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman & A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34(Database issue):D140–D144, 2006

[85] T. M. Lowe & S. R. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, 25(5):955–964, 1997

[86] E. B. Lewis. A gene complex controlling segmentation in Drosophila. *Nature*, 276(5688):565–570, 1978

[87] S. Petruk, Y. Sedkov, K. M. Riley, J. Hodgson, F. Schweisguth, S. Hirose, J. B. Jaynes, H. W. Brock & A. Mazo. Transcription of bxd noncoding rnas promoted by trithorax represses ubx in cis by transcriptional interference. *Cell*, 127(6):1209–1221, 2006

[88] T. Sanchez-Elsner, D. Gou, E. Kremmer & F. Sauer. Noncoding RNAs of trithorax response elements recruit Drosophila Ash1 to Ultrabithorax. *Science*, 311(5764):1118–1123, 2006

[89] Y. Lee, K. Jeon, J.-T. Lee, S. Kim & V. N. Kim. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J*, 21(17):4663–4670, 2002

[90] Y. Zeng. Principles of micro-RNA production and maturation. *Oncogene*, 25(46):6156–6162, 2006

[91] M. Ronshaugen, F. Biemar, J. Piel, M. Levine & E. C. Lai. The Drosophila microRNA iab-4 causes a dominant homeotic transformation of halteres to wings. *Genes Dev*, 19(24):2947–2952, 2005

[92] C. Y. Pai, T. S. Kuo, T. J. Jaw, E. Kurant, C. T. Chen, D. A. Bessarab, A. Salzberg & Y. H. Sun. The Homothorax homeoprotein activates the nuclear localization of another homeoprotein, extradenticle, and suppresses eye development in Drosophila. *Genes Dev*, 12(3):435–446, 1998

[93] J. Bessa, B. Gebelein, F. Pichaud, F. Casares & R. S. Mann. Combinatorial control of Drosophila eye development by eyeless, homothorax, and teashirt. *Genes Dev*, 16(18):2415–2427, 2002

[94] V. Goguel, Y. Wang & M. Rosbash. Short artificial hairpins sequester splicing signals and inhibit yeast pre-mRNA splicing. *Mol Cell Biol*, 13(11):6841–6848, 1993

[95] D. L. Philipps, J. W. Park & B. R. Graveley. A computational and experimental approach toward a priori identification of alternatively spliced exons. *RNA*, 10(12):1838–1844, 2004

[96] N. J. Sakabe & S. J. de Souza. Sequence features responsible for intron retention in human. *BMC Genomics*, 8:59, 2007

[97] S. Tan, J. Guo, Q. Huang, X. Chen, J. Li-Ling, Q. Li & F. Ma. Retained introns increase putative microRNA targets within 3' UTRs of human mRNA. *FEBS Lett*, 581(6):1081–1086, 2007

[98] A. Derti, F. P. Roth, G. M. Church & C. ting Wu. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet*, 38(10):1216–1220, 2006

[99] S. M. Mount, V. Gotea, C.-F. Lin, K. Hernandez & W. Makalowski. Spliceosomal small nuclear RNA genes in 11 insect genomes. *RNA*, 13(1):5–14, 2007

[100] I. L. Hofacker, M. Fekete & P. F. Stadler. Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, 319(5):1059–1066, 2002

[101] J. Jöris. Secondary structure prediction for aligned RNA sequences of heterodimers. Master's thesis, Universität Leipzig, 2007

[102] G. C. Shukla & R. A. Padgett. U4 small nuclear RNA can function in both the major and minor spliceosomes. *Proc Natl Acad Sci U S A*, 101(1):93–98, 2004

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

[104] F. P. Roth, J. D. Hughes, P. W. Estep & G. M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol*, 16(10):939–945, 1998

[105] B. Tolhuis, E. de Wit, I. Muijrers, H. Teunissen, W. Talhout, B. van Steensel & M. van Lohuizen. Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster. *Nat Genet*, 38(6):694–699, 2006

[106] T. Sandmann & S. M. Cohen. Identification of Novel Drosophila melanogaster MicroRNAs. *PLoS ONE*, 2(11):e1265, 2007

[107] R. Dorn, G. Reuter & A. Loewendorf. Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in Drosophila. *Proc Natl Acad Sci U S A*, 98(17):9724–9729, 2001

[108] M. Gabler, M. Volkmar, S. Weinlich, A. Herbst, P. Dobberthien, S. Sklarss, L. Fanti, S. Pimpinelli, H. Kress, G. Reuter & R. Dorn. Trans-splicing of the mod(mdg4) complex locus is conserved between the distantly related species drosophila melanogaster and d. virilis. *Genetics*, 169(2):723–736, 2005

[109] V. Krauss & R. Dorn. Evolution of the trans-splicing Drosophila locus mod(mdg4) in several species of Diptera and Lepidoptera. *Gene*, 331:165–176, 2004

Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Leipzig, 11.01.2008