# 4 CORG - a promoter annotation framework

This chapter presents, the CORG (Comparative Regulatory Genomics) pipeline, our infrastructure for large-scale analysis of putative promoter regions. All relevant design issues as well as corresponding software modifications and implementations will be discussed here. Many of the aspects have been published previously e.g. (Dieterich et al., 2002, methodical aspects) and (Dieterich et al., 2003b, database related issues)).

## 4.1 Definition of an upstream region

The notion of "promoter region" deserves some further explanation in the context of our approach. Typically, though not exclusively, we expect conserved regulatory regions to appear in the vicinity of the promotor of a gene. Since we seldom know the precise location of the start of transcription, we chose to compare the sequence regions upstream of the start of translation from orthologous genes. Of course, we are very generous in defining the extent of this upstream sequence such that we can be confident that it will encompass the start of transcription and the promotor. To get an idea of the size of "upstream" regions, we investigated the distance distribution between verified genomic transcription start sites and translation start sites (Figure 4.1).

The Eukaryotic Promoter Database (Schmid et al., 2004) was the first available dataset to address this issue. The vast majority of known promoters are closer than 1 Kb to the translation start site (median at 303 bp). Our observations indicate that most promoter regions should be captured in a sequence window of 15 Kb size (95 % quantile is at ≈ 16 Kb). Figure 4.1A summarizes these observations. The size of a promoter region may be bounded by the size of the corresponding intergenic region. If an annotated gene happens to lie within the primary sequence window, the promoter region will be shortened to exclude exonic sequence and for the sake of computation time.

The overall impression did not change with the advent of new large-scale efforts to map transcription starts. Figure 4.1 B+C show the distance distributions for two recent projects: the human section of DBTSS (Suzuki et al., 2004) and the H-InvDB project

(Imanishi et al., 2004). Section 4.3.4 discusses how these resources are integrated into the CORG project.

## 4.2 Sequence retrieval and preprocessing

In Section 2.3.1, we presented phylogenetic footprinting as an appropriate tool for the detection of regulatory elements in promoter regions. A prerequisite for this approach is the definition of groups of homologous, or better orthologous, gene loci. This step deserves much attention since a wrong grouping of sequences results in the comparison of unrelated sequences. Traditionally, phylogenetic relations[1] are inferred *in silico* from protein sequence similarity searches. Alternatively, whole genomic regions can be classified based on gene order and content (synteny).

### 4.2.1 Phylogenetic relationships of genes

In this work, we take a gene-centered view of phylogeny. Homology among proteins and genes is often concluded on the basis of sequence similarity, especially in bioinformatics. Many algorithms exist to cluster protein sequences into sequence families, which are sets of mutually homologous sequences. Traditionally, one would define homology based on protein sequence comparisons following the procedure outlined by Tatusov et al. (1997). In this approach, best reciprocal similarity matches are computed. A triplet of mutually best matching sequences from three different species founds a group of orthologous genes. However, sequence similarity may be misleading (i.e. in terms of functional homology) or inconclusive (i.e. failure to resolve ortholog-paralog relations). An example of the first case is presented by Fessele et al. (2002). They showed for the human RANTES/CCL5 gene that orthology inference based on protein sequence similarity does not always guarantee functional similarity. Their results indicate that the murine CCL5 exerts a similar function as the human GRO gene product.

Secondly, a failure in resolving ortholog-paralog relations is particularily disturbing in the context of CORG since it will ultimately lead to large groups of homologous genes. An exhaustive pairwise cross-species comparison of all promoter regions within a gene group becomes intractable if the member count increases dramatically.

To mitigate the latter problem, Birney et al. (2004) further improved the detection of phylogenetic relationships by taking information on conserved synteny into account. We retrived all components from the undirected graph of EnsEMBL homologous gene pairs. All members of a component form a homologous group of genes in CORG.

[1] A phylogeny (or phylogenesis) is the origin and evolution of a set of organisms, usually species.

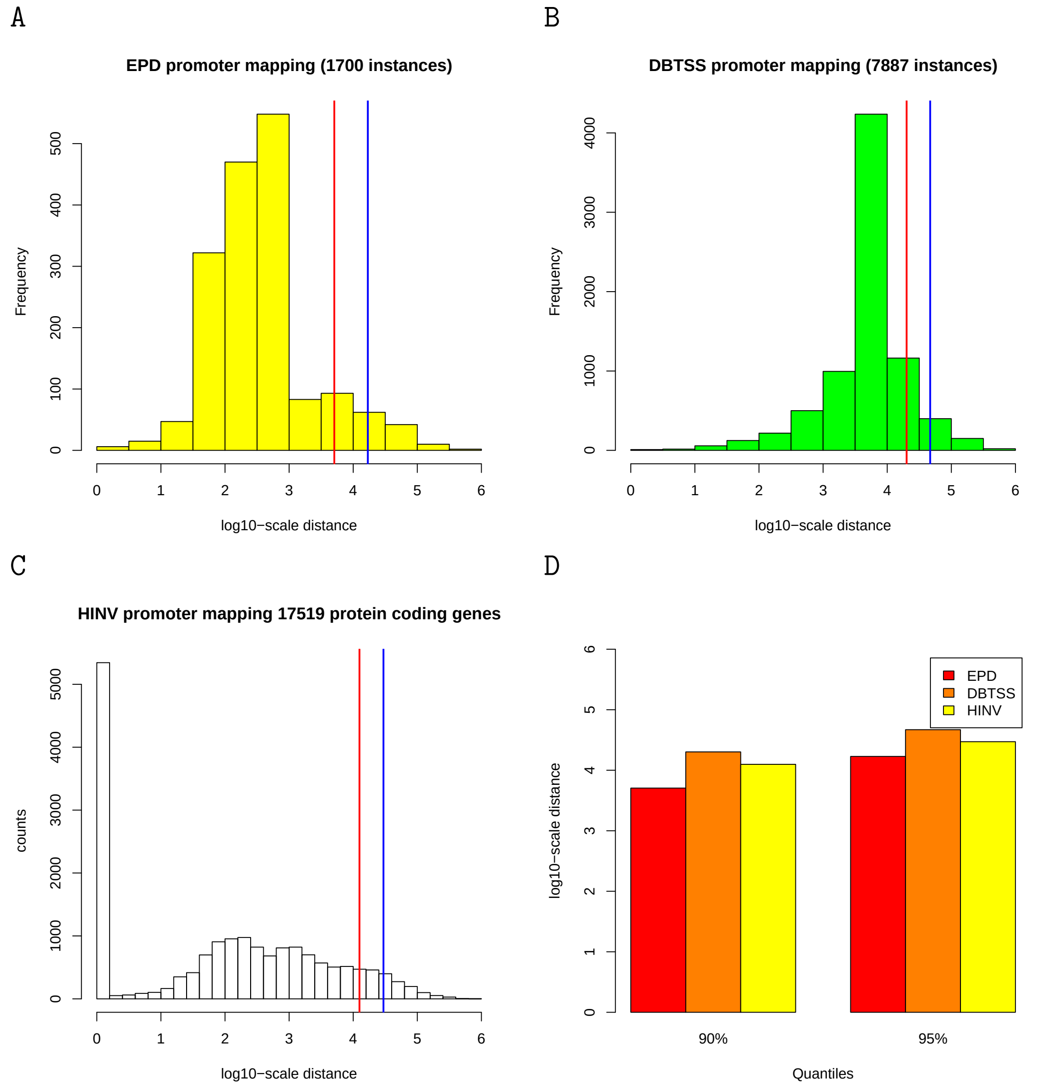


**Figure 4.1: Distributions of distance between start of transcription and translation.** Histograms of observed genomic distances between start sites of transcription and translation in man. Red and blue lines indicate the 90% and 95% quantiles, respectively. Distances greater than $10^6$ bp were exluded from the analysis as they mostly occur due to mismappings in the ENSEMBL database ($\leq$ 1% of instances). A: Distance distribution based on 1,700 promoters from the EPD database. B: 7,887 mappings of DBTSS promoters. C: Distribution over H-InvDB with 17,519 entries D: Comparative display of quantile values for A-C. Quantile values (from left to right): 3.705005; 4.302720; 4.097500; 4.227985; 4.670103; 4.471784.

| Algorithm | % aligned sequence | % “Sensitivity” |
|---|---|---|
| SIM | 23 | 98 |
| BBA | 24 | 90 |

**Table 4.1: Wasserman et al. dataset.** Summary of the results on the comparison of 27 man-rodent promoter pairs. The first column lists the employed algorithms where SIM is an implementation of the Waterman-Eggert algortihm as used in CORG and BBA stands for `Bayes Block Aligner`, which is the algorithm used in Wasserman et al. (2000). The second column shows the overall proportion of aligned human sequence. The third column shows what percentage of “known” binding sites were inside the aligned sequence regions.

**Pros and cons of footprinting** Wasserman et al. (2000) presented a fundamental study on the benefits of phylogenetic footprinting to detect regulatory elements. They compared a set of 28 man-rodent orthologous gene pairs that are specifically upregulated in skeletal muscle, and for which there is considerable genomic sequence available. We could reconstruct their test setting for 27 man-rodent promoter pairs.

In a direct comparison (Table 4.1), the SIM algorithm performed worse in terms of quality but was superior in terms of speed (2 orders of magnitude). Binding sites of all three major muscle-specific transcription factors (MYF, SRF and MEF2) can be computationally identified. A detailed example of one promoter pair is shown in Figure 4.2.

Contrary to Wasserman et al. (2000), Dermitzakis and Clark (2002) are less optimistic on the average number of conserved sites in man-rodent comparisons. They constructed a test set of 20 man-rodent promoter pairs for which extensive experimental data were available. A total of 64 alignable binding sites have been identified in these 20 regions, out of which 33 have shared function between human and rodents (mouse or rat), 14 are human specific and 17 are rodent specific. As a consequence, they estimate that 32%-40% of the human functional sites are not functional in rodents. The dataset of Dermitzakis and Clark (2002) is a better representative of promoter regions than the one of Wasserman et al. (2000) as it is unbiased towards specific gene groups (i.e. skeletal muscle-specific genes).

### 4.2.2 Initial sequence processing

Once all groups of orthologous genes are defined, we retrieve the corresponding promoter regions (see Section 4.1 for an explanation) from the EnsEMBL core databases. We employ `RepeatMasker` (Smit and Green) to mask promoter regions for repeats, which would interfere with our computation of $p$-values and are largely neglegible in terms of gene regulation.

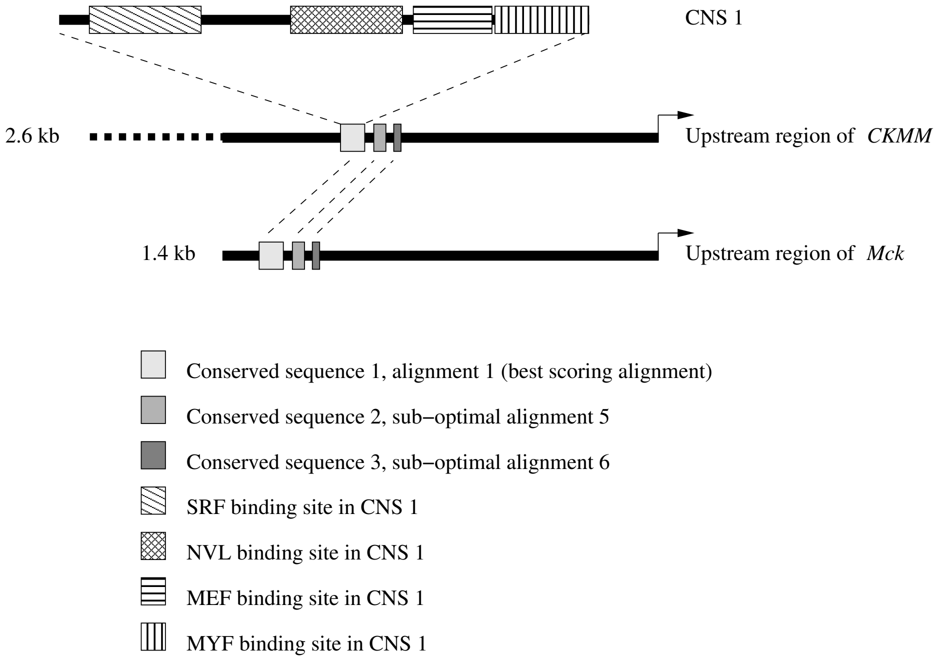

CNS 1
2.6 kb
Upstream region of *CKMM*
1.4 kb
Upstream region of *Mck*
Conserved sequence 1, alignment 1 (best scoring alignment)
Conserved sequence 2, sub–optimal alignment 5
Conserved sequence 3, sub–optimal alignment 6
SRF binding site in CNS 1
NVL binding site in CNS 1
MEF binding site in CNS 1
MYF binding site in CNS 1

**Figure 4.2: Phylogenetic footprinting in the promoter region of *CKM*.** Promoter regions for human and murine creatine kinase, M chain. A battery of colinear local alignments is depicted as gray-scale boxes. The best alignment is enlarged and shown to largely consist of four conserved binding sites. CNS stands for Conserved Noncoding Sequence.

## 4.3 The notion of conserved non-coding blocks

We now introduce a new term for local suboptimal alignments in non-coding DNA regions: **Conserved Non-coding Blocks (CNBs)**. There are several programs available to detect local similarities with gaps between two sequences and we have already given an overview on those (see Section 3.1.4). In this thesis, CNBs are computed using an implementation of the Waterman-Eggert algorithm and a DNA scoring scheme, which is based on the two-parameter Kimura model and normalized to a distance of $n$ PAM (States et al., 1991) where $n$ depends on the species pair under comparison. In earlier work (Dieterich et al., 2003b), we opted for a pragmatic choice of the scoring scheme for man-mouse comparison. This choice was mainly motivated from individual promoter pair comparisons and is now supported by in-depth analysis of whole-genome alignments. We will now have a look at the procedure of CNB detection.

**Table 4.2: Program workflow adaptation for batch alignments**

```
1:  Read in commandline options
2:  while ((A,B) $\leftarrow$ read_sequence_pair) do
3:          $B' \leftarrow reverse(B)$
4:          Compute random scores SIM($A, B'$)
5:          Estimate parameters $(\gamma, p) \leftarrow$ gsl_fit_linear
6:          Compute alignment scores SIM($A, B$)
7:          Report $n$ best alignments with $P(score|rank) > threshold$
8:  end
```

### 4.3.1 Adaptation of the SIM implementation of the Waterman-Eggert algorithm.

The original implementation as obtained from the Globin Gene Server (http://globin.cse.psu.edu/html/software.html) was modified with respect to statistical assessment, file handling and score matrix input. These modifications (listed below) became necessary due to the high-throughput context of the CORG analysis pipeline.

**Input, line 1+2.** Sequence files that contain paired FASTA entries substitute for the single sequence files. A score function can be globally set for the whole comparison or for each individual comparison (batch file). Additionally, a p-value cutoff needs to be given.

**Parameter estimation of random score distribution, line 5.** The program computes random alignment scores on the sequence pair ($A$,$B$). The number $N_t$ of local gapped alignments exceeding a threshold $t$ has a Poisson distribution with mean $\gamma mnp^t$ where $mn$ is the length of the sequence search space and parameters $\gamma$ and $p$ are unknown for the gapped case. Linear regression on the two-dimensional data space ($N_t$ vs. $t$) yields both parameters $\gamma$ and $p$ for the Poisson model of random scores (see Figure 4.3).

**Significance computation, line 7.** $P$-values are computed according to the order statistics as in Section 3.2.2. Subroutines have been introduced for this purpose.

### 4.3.2 Detection of CNBs

With the computational method at hand, we had to decide on an appropriate way to score our pairwise alignments. Since we were mainly interested in highly conserved regulatory elements, we demanded an average similarity level at least as high as the average exon conservation. Makalowski et al. (1996) report in an initial analysis of 1196 orthologous man-rodent mRNAs an average degree of nucleotide identity for coding exons of 85%. This number gained support from a recent genome-wide analysis of Waterston et al. (2002), who found an average degree of conservation of around 87.3%.

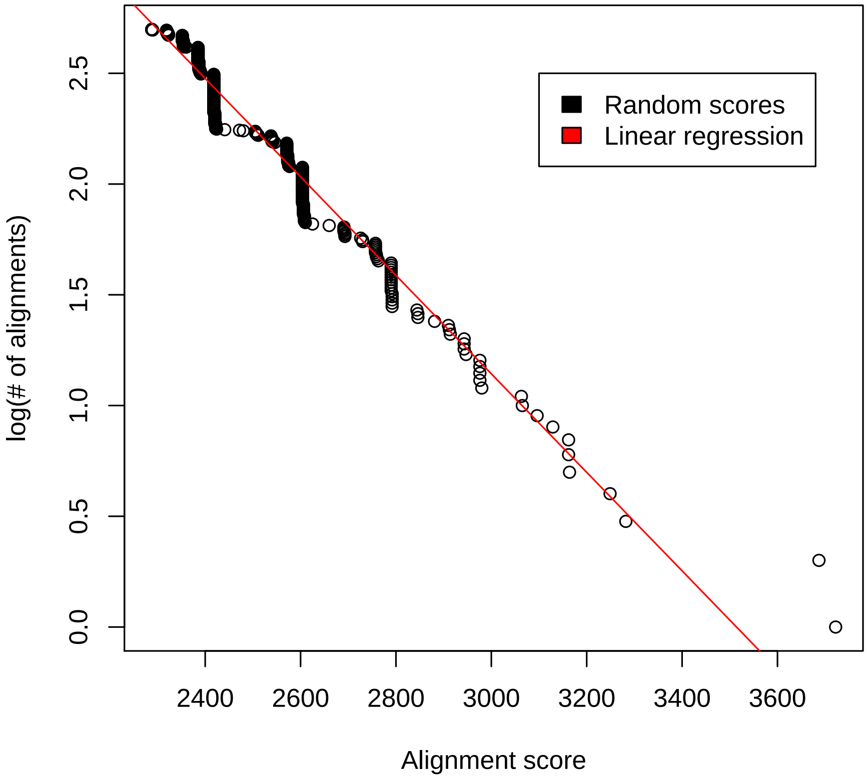


**Figure 4.3: Parameter estimation via linear regression of sampled random scores.** This plot shows the distribution of scores of 500 random alignments. Two orthologous sequences were masked for mammalian wide repeats. One was reversed and both then aligned with a Kimura two-parameter model scoring matrix of PAM distance 10.

Having these values in mind, we computed our man-mouse CNBs at two stringency levels: a) 1 PAM and b) 10 PAM. These choices correspond to an expected conservation level of 99 % and 90.7 %. By using Kimura's two-parameter model (K2M), we took the mutational bias between transition and transversion rate into account. In our model, a transition is three times more likely to occur than a transversion. Large-scale studies on man-rodent sequence evolution provide extensive evidence for this choice (Hardison et al., 2003; Cooper et al., 2004). For sequence comparisons of other vertebrate species pairs, we opted for an arbitary scoring scheme choice based on the

K2M model at a PAM distance of 50. This corresponds to an average conservation level of 64.2%. States et al. (1991) studied the *efficiency* of nucleotide scoring matrices. They define *efficiency* as the ratio of the observed average bit score at a PAM distance $D$ to the optimal average bit score at the actual PAM distance of the scoring matrix. The authors support our choice by demonstrating that the 90% efficiency level of the uniform mutational model matrix PAM-47, as used in `BLASTN`, spans a distance range from 20 to 68 PAM. We deem this range sufficient for detecting phylogenetic footprints, which stem from conserved transcription factor binding sites.

The choice of gap penalties is mainly constrained by the phenomenon of phase transition (see Section 3.2.2). Previous work (Cusack, 2001) suggests a gap opening penalty of 11× `match score` and a gap extension of penalty of 0.1× `match score`.

A maximum of 100 local alignments are computed for each pairwise sequence comparison an upstream region of 15 Kb. If the p-value of any alignment exceeds a threshold of 0.001, alignment sampling is stopped and the corresponding alignment is discarded. All signficant CNBs are then stored into an MySQL database. CNBs are the essence of CORG and subsequent annotation with various features is build on top of them.

### 4.3.3 Extension to multi-species comparison

Comparative approaches gain power from the inclusion of sequences from more than two species (McCue et al., 2002). Multi-species comparisons help to reduce the level of noise as supporting evidence in terms of conservation stems from many observations. The buildup of multiple alignments from pairwise alignments (CNBs) is subsequently summarized.

All CNBs from pairwise sequence alignments are split up into groups as defined by gene homology. For each group a graph $O = (V, E)$ with vertices $V$ and edges $E$ is constructed, which represents the species-internal overlap of CNBs on the genomic coordinate level. Each vertex $a \in V$ represents a CNB, which is a pairwise local alignment between two species. An undirected edge is placed between two vertices if the corresponding CNBs have only one species in common and show an overlap of at least 10 bp on the sequence level.

**Definition 4.1** (Clique in an undirected graph)**.** A clique in a undirected graph G, is a set of vertices V' such that for every two vertices in V, there exists an edge connecting the two. The size of a clique is the number of vertices it contains.

In our graph $O$, cliques of minimal size three are detected with an implementation of the Bron-Kerbosh algorithm (Bron and Kerbosch, 1973). Only those cliques are selected whose species count is equal to their size. This move prohibits the emergence of multiple alignments by similarity of multiple short CNBs to a single long CNB.

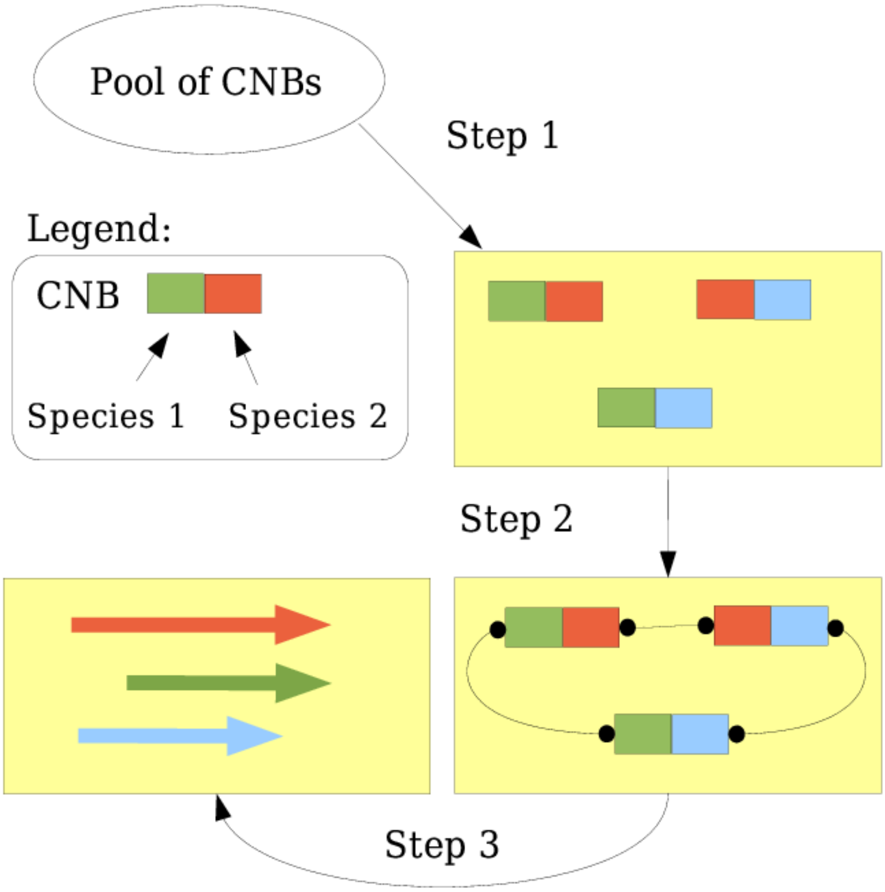


**Figure 4.4: CORG multiple alignment building.** This sketch depicts the three principal steps of multiple alignment building. Step 1: Retrieval of all CNBs that belong to one group of homologous genes. Step 2: Construction of a CNB overlap graph $O = (V, E)$ with vertices $V$ and edges $E$, which represents the species-internal overlap of CNBs on the genomic coordinate level. Step 3: Clique finding with an implementation of the Bron-Kerbosh algorithm and subsequent realignment of the CNB sequences with POA.

**Table 4.3: Multiple alignment workflow**

```
1:  $G \leftarrow$ Retrieve homologous gene groups from database
2:  foreach $g \in G$ do
3:          Construct CNB graph: $O \leftarrow$ overlap(g)
4:          Max-Clique detection: $C \leftarrow$ bron_kerbosh(O)
5:          Compute multiple alignments: $M \leftarrow$ partial_order_alignment(C)
6:          Trim multiple alignments to conserved core
7:          Output multiple alignments
8:  end
```

Multiple alignments are then computed based on all cliques that meet the outlined criteria. Alignment results are subsequently trimmed to a "core" alignment where alignment columns are mostly populated. A summary of the building procedure is shown in Table 4.3.

**CNB overlap graph, line 3.** The graph of overlapping CNBs is constructed from an exhaustive search over all CNBs within a group of homologous genes.

**Multiple alignment computation, line 5.** Multiple alignments are computed with the POA software from input sequences that are extracted from the corresponding pairwise alignments.

**Multiple alignment trimming, line 6.** This step removes segments of low conservation in border areas of the alignment. Trimming stops if an ungapped block of at least 6 nucleotides is encountered on each alignment border.

### 4.3.4 Annotation of conserved non-coding blocks and promoter regions

**Genomic mapping of validated promoter regions** Various recent experimental efforts supply information as to the position of transcriptional start sites in the human and mouse genome. We will briefly give an overview on the resources that were employed in CORG.

**Eukaryotic promoter database (EPD).** The Eukaryotic promoter database is the smallest in size, but largely consists of manually curated entries (Schmid et al., 2004)

**DataBase of Transcriptional Start Sites (DBTSS).** Suzuki et al. (2004) compiled reliable information on the transcriptional start sites for man and mouse promoters. They exploit the oligo-capping technique to enrich their pool of clones for full-length 5'-to-3' cDNAs.

**H-Invitational Database (H-InvDB).** An international effort to integrate annotation of 41,118 full-length human cDNA clones that are currently available from six high throughput cDNA sequencing projects (Imanishi et al., 2004).

**FANTOM 2 collection of full-length cDNAs (RIKEN).** Bono et al. (2002) presented the FANTOM collection of RIKEN full-length cDNA clones. FANTOM stands for Functional Annotation of Mouse cDNA clones.

**The Reference Sequence project (RefSeq).** The Reference Sequence project (Pruitt and Maglott, 2001) aims to provide a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms.

Some repositories offer genomic coordinates for their start site entries. Existing genomic mapping information was incorporated unless the corresponding CORG genome assembly build differed. The remaining data were projected onto the genome with `SSAHA` (Sequence Search and Alignment by Hashing Algorithm), a rapid near-exact alignment algorithm (Ning et al., 2001).

**Exon detection with assembled EST clusters** Promoter regions in CORG always extend upstream from the coding start (ATG) of the transcript, which is the most downstream with respect to the other transcripts. As a consequence, our promoter regions may contain exons from the same transcript that are not translated or exons from other transcripts of the same gene. Our way of detecting such exons is a similarity search of CNBs versus `GENENEST` (Krause et al., 2002), a database of assembled EST clusters. Database searches are carried out for human and mouse CNBs with the `BLASTN` program (Altschul et al., 1997). An E-value cut-off of $10^{-4}$ is applied.

**Annotation with predicted binding sites** The TRANSFAC database (Matys et al., 2003) is a repository of experimentally verified binding site sequences and representations thereof. These representations are used for querying the collection of man-mouse CNBs for known binding site patterns. Two different types of binding site descriptions can be distinguished:

**Transfac sites** Site entries give information on individual (regulatory) protein binding sites that were experimentally verified. Sequences of these sites are *bona fide* candidates for stringent binding site detection in CORG. Site searches were effected with the `FUZZNUC` software from the EMBOSS package (Rice et al., 2000).

**Transfac weight matrices** Count matrix representations are generated from aligning single binding site sequences. A multiple alignment of length $m$ can be translated into a profile matrix $F$ with $n = 4$ rows and $m$ columns:

$$F_{ij} = \frac{c_{ij} + p_{ij}}{c_{\cdot j} + p_{\cdot j}} \qquad i \in \{A, C, G, T\};\ j \in \{1, \dots, m\} \tag{4.1}$$

where $F_{ij}$ is the matrix entry at the $i$th row and $j$th column, $c_{ij}$ is the observed count of nucleotide $i$ in alignment column $j$ ($c_{\cdot j} \equiv$ sum of observed counts in column j) and $p_{ij}$ is a pseudo count that is added for regularization purposes ($p_{\cdot j} \equiv$ sum of pseudo counts in column j). To decide whether a sequence $S$ of length $m$ is an occurence of the signal described by the profile matrix $F$, we have to contrast it to a random or background model. A suitable and common choice to model a random sequence is a simple i.i.d model, that is a profile matrix $B$ where each column consists of the same probability vector $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ (see also Section 3.2.1, page 32).

A quantity to guide the decision is the *likelihood ratio* of the models, that is the ratio of probabilities of observing sequence $S$ in both models. The *log-likelihood ratio* is then defined as:

$$\texttt{Score}(S) := \log(\mathbb{P}_F(S)/\mathbb{P}_B(S)) = \sum_{i=1}^{m} \log(F_{S_i i}/\pi_{S_i}) \tag{4.2}$$

If the background distribution $\pi$ is fixed, the profile matrix $F$ can be directly translated into a position weight matrix $W$ (PWM) or position specific score matrix (PSSM) by setting $W_{ij} = \log(P_{ij}/\pi_i)$. The whole procedure is analogous to introducing an alignment score as in Section 3.1.3.

Potential binding sites are detected with Transfac weight matrices by the method of Rahmann et al. (2003). Here, the intuition is that there are two random models for a given sequence $S$: one is given by the signal profile $F$ and the other one by the background model $B$. The sum distribution of column scores (generally independent random variables), which is the distribution of weigth matrix scores, is conveniently calculated by convolution. Probability mass distributions of $\mathbb{P}_F(\texttt{Score}(S))$ as well as $\mathbb{P}_B(\texttt{Score}(S))$ can be computed by dynamic programming if column scores are reasonably discretized (Rahmann et al., 2003).

**Intra-species sequence variation** Polymorphisms within promoter regions may also affect gene regulation. That is why, we are particularily interested in annotating CNBs and putative binding sites with SNP data. The data stems from public efforts that were submitted to `dbSNP` (Sherry et al., 2001).

### 4.3.5 CORG pipeline

All subsequent steps of building process of CORG are summarized in Figure 4.5.

## 4.4 Database design

CNBs or local pairwise similarities are at the heart of the CORG database. This comparative perpective is also reflected in the design of the database structure. The database's core is formed by the tables `dna`, `dna_alignment` and `alignment`. Definitions of upstream regions (promoters) are stored in table `dna`. A pair of upstream region entries is linked to a single alignment entry in `alignment` via `dna_alignment` entries. In Figure 4.6, these tables are connected via 1-to-many relations and describe the many-to-many relation of upstream regions to CNBs. This architecture does not impose any constraints on placing local similarities onto genomic sequence. Cases where alignments fall into overlapping upstream regions or repetitive similarities occur can be easily deposited in this structure. Likewise, multiple alignments are defined by their pairwise counterparts: Tables `alignment, ali_multiple, multiple_alignment` describe the position and content of multiple alignments. Thus, CNBs as building blocks have a many-to-many relation to multiple alignments.

Most annotations are CNB-oriented as this is the essence of CORG: One class of conserved elements may simply originate from untranslated exons. Those CNBs are detected via similarity searches vs. assembled EST sequences and the corresponding annotation is stored in table `hits`. Furthermore, CNBs often encompass binding site motifs or SNPs (tables `transcript_hits, tfbs, snps`). In short, sequence features where conservation hints at the functional importance are directly linked to CNBs.

Other, more global properties of upstream regions or genes like mapped transcription start sites, sequence or homology assignment are directly linked to entries in `dna` or `gene`.

## 4.5 Web interface

In the previous section we have seen how CORG data are internally stored in a relational database. This is the method of choice for handling large amounts of data. Individual promoter studies are better supported with a graphical interface that provides a user-friendly view of the database. We have implemented a WWW service for accessing the database over the internet. The CORG database is accessible via its home page (`http://corg.molgen.mpg.de`). One can quickly jump to gene loci via EnsEMBL or other standard identifers (i.e. HUGO symbol, LocusLink identifier, etc.).

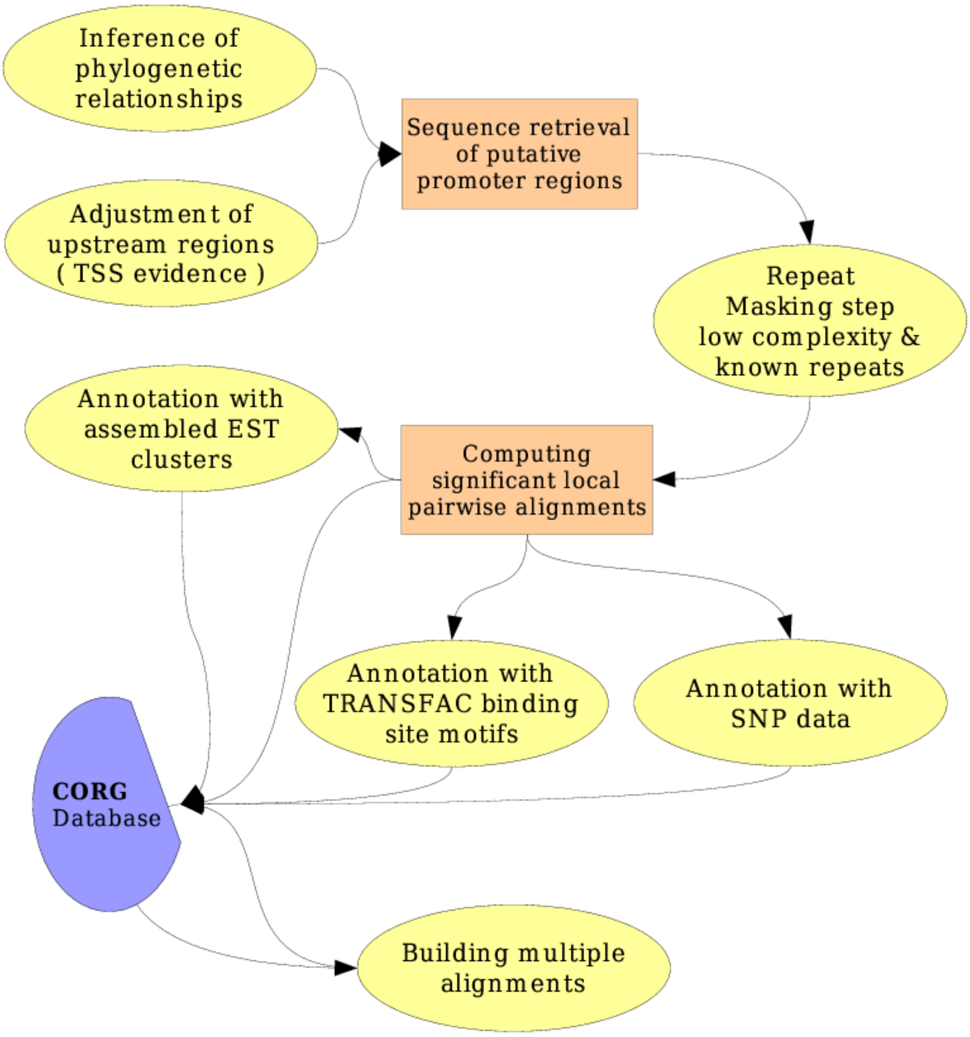

Inference of phylogenetic relationships
Adjustment of upstream regions ( TSS evidence )
Sequence retrieval of putative promoter regions
Repeat Masking step low complexity & known repeats
Annotation with assembled EST clusters
Computing significant local pairwise alignments
Annotation with TRANSFAC binding site motifs
Annotation with SNP data
CORG Database
Building multiple alignments

**Figure 4.5: CORG pipeline workflow.** Firstly, groups of putative orthologs have to be defined. Secondly, putative promoter regions are extracted from the set of EnsEMBL databases with some prior information on individual validated transcription start site. The repeat masking step is either done de novo or by employing precomputed results. Subsequent analysis is centered around the set of CNBs.

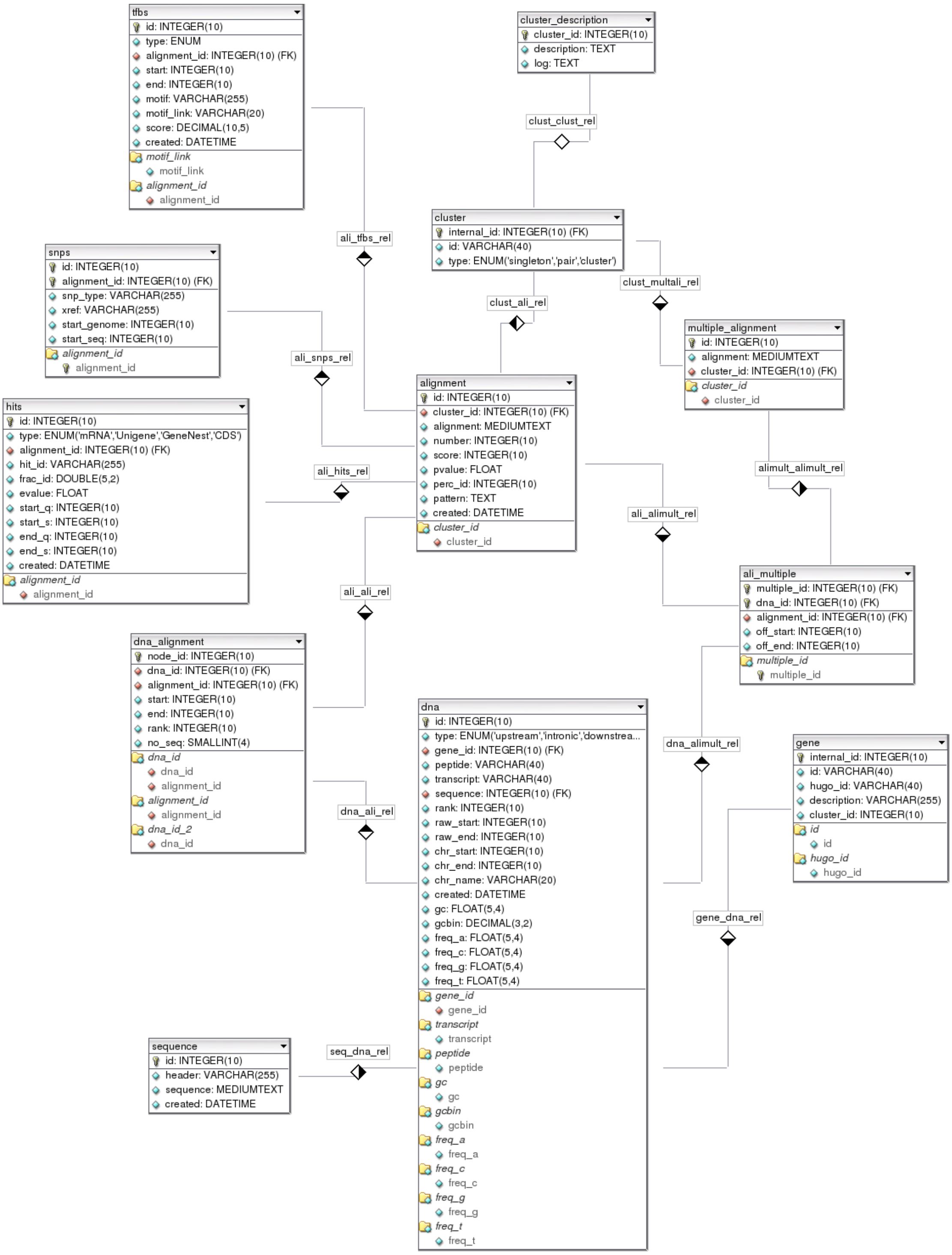
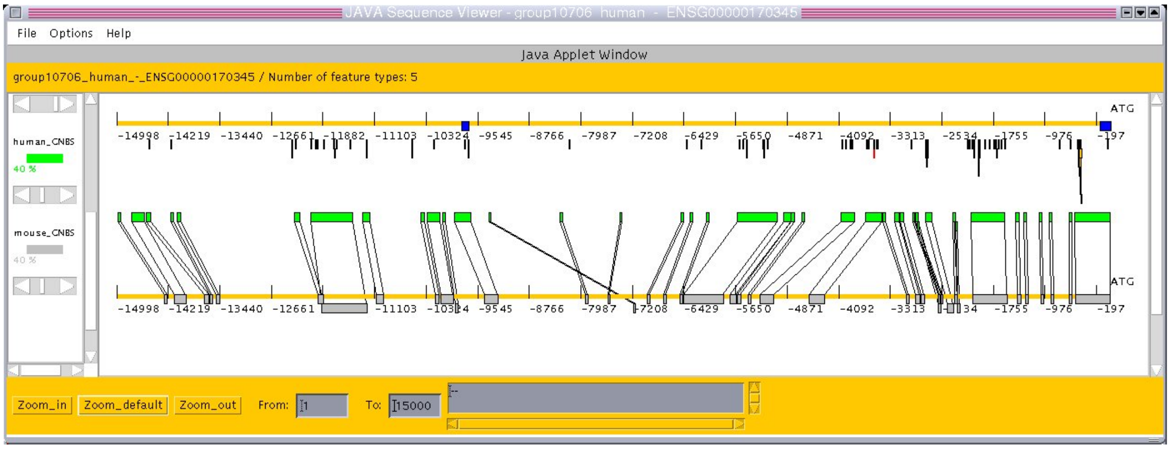


**Figure 4.6:** CORG database schema. Tables and main relations are shown. Some tables `promoter2genome`, `sequence_coordinates`, `trash_hits` and `organism` were omitted due to space constraints.

The search query is processed according to the chosen reference source and a list of all matching database entries is returned to the user. This list serves as a springboard to the visualization step. All annotation results are then visualized by a JAVA applet that complies with the JDK 1.1 standard (see Figure 4.7). Thus, the applet should run on all JAVA-compatible web browsers. Detailed information about the conserved non-coding block structure are simultaneously shown for multiple upstream regions of different species. If available, annotation information on putative binding sites of transcription factors and EST matches are displayed for the query sequence. The applet facilitates zooming into sequence and annotation. In addition, web links are assigned to sequence features that relate external data sources to the corresponding annotation.

Alternatively, CORG data may be embedded into other viewers via the distributed annotation system (*DAS*, Dowell et al. (2001)). DAS facilitates the display of distributed data sources in a common framework with respect to a reference sequence. Our DAS server (`http://tomcat.molgen.mpg.de:8080/das`) constitutes such an external data source. Position information on all conserved non-coding blocks and mapped promoters is accessible from this DAS server. Each DAS sequence feature provides a link to the corresponding CORG database entry. New DAS sources can be easily added to the ENSEMBL display. A small tutorial on installing external DAS data sources is available on our web page (`http://corg.molgen.mpg.de/DAS_tutorial.htm`).



JAVA Sequence Viewer - group10706_human_-_ENSG00000170345

Java Applet Window

group10706_human_-_ENSG00000170345 / Number of feature types: 5

From: 1 To: 15000

**Figure 4.7: JAVA technology based web interface to CORG.** The upstream region of the human *c-FOS* gene and its conservation pattern to mouse is shown. On the left, a window provides a legend to read the display. Sequence features are represented by coloured boxes. Cross-species similarity features are linked with black lines. The translation start site is depicted with the label 'ATG' to the right of the display.

## 4.6 CORG content summary

An overview on global properties of the CORG database content will be given in this section. Quantities like GC content and average degree of conservation will be discussed as well as the position of CNBs relative to the translation start.

### 4.6.1 GC content and upstream region length

Figure 4.8 is a boxplot of the distribution of GC content in the upstream regions excluding repetitive sequences of the five species under investigation. In the boxplot, the range from 1st to 3rd quantile is framed by the actual box. The median is highlighted within this box by a horizontal line. The whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box.

There are clearly a few striking observations: *Danio rerio* is the species with the lowest median GC content (35.5 %). The reason for this gap in GC contents is elusive. Median GC values for the other species are close to each other (43.9 % - 45.4 %). Median GC values are consistently higher for mammalian genomes compared to the overall genomic GC content, which is plotted as red triangles. This indicates a more important role of CpG island and methylation in mammalian promoter regions. The GC content distribution of *Homo sapiens* upstream regions is most "flattened" out and populates a broad range of GC contents. Figure 4.9 demonstrates that the length of an upstream region is largely determined by the preset promoter region cutoff of 15 Kb. Upstream regions may be longer due to the existence of alternative 5' ends. As can be seen in Figure 4.9, the genome of *Fugu rubripes* is the most condensed since it is almost devoid of repetitive sequence (Aparicio et al., 2002).

### 4.6.2 Conservation extent and localization

Overlaying repeat densities and conservation coverage as in Figure 4.10 yields an idea of the nature of the putative "average" man-mouse promoter region. The majority of upstream regions happen to be less than 10% conserved and contains 40% - 48% repetitive sequence. No simple linear corelation was observed between both measures.

It is also instructive to study the distribution of significant local alignments with respect to their localization relative to the translation start sites. Typically, in man-mouse alignments one would anticipate that local alignments would lie in phase with exonic regions of the two genes under comparison. In random alignments, aligned regions do not show a preference to cluster around the main diagonal of the alignment matrix. Figure 4.11 summarizes these findings for normalized alignment localization data.

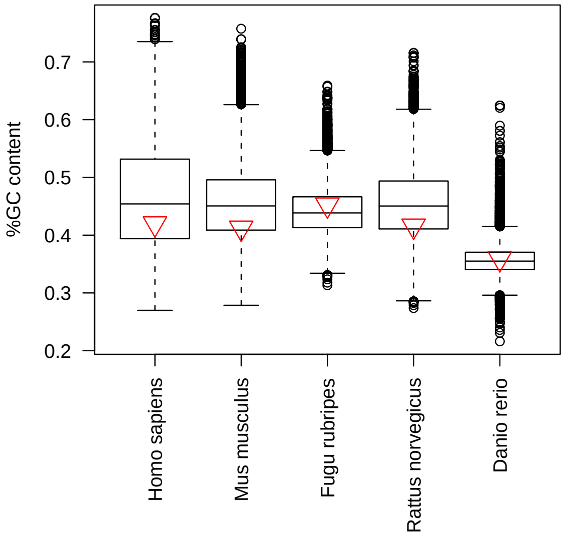


**Figure 4.8: GC content of promoter regions in five different species.**
The GC content is computed on the occurences of nucleotides A,C,G and T. Sequencing gaps and repetitive sequences are not taken into account. Red diamond symbols depict genomewide GC content. Further details in text.

Upstream region lengths were uniformly normed to 1. Thus, alignment positions `pos` are always in $[0, 1]$. The human sequence was taken as a reference for the comparison against mouse and fish. Figure 4.11 is a “smoothed” histogram of all observed distances: $\texttt{pos}_{human} - \texttt{pos}_{other}$. Remarkably, man-mouse alignments are mostly collinear. Generally, CNBs in human upstream regions tend to be further away from the translation start site than in mouse, zebrafish and fugu (median = (0.01, 0.03, 0.05)). Histograms for man-fish comparisons also show “bumps” in their off-peak flanks, which points at cases of non-colinear alignment positioning.

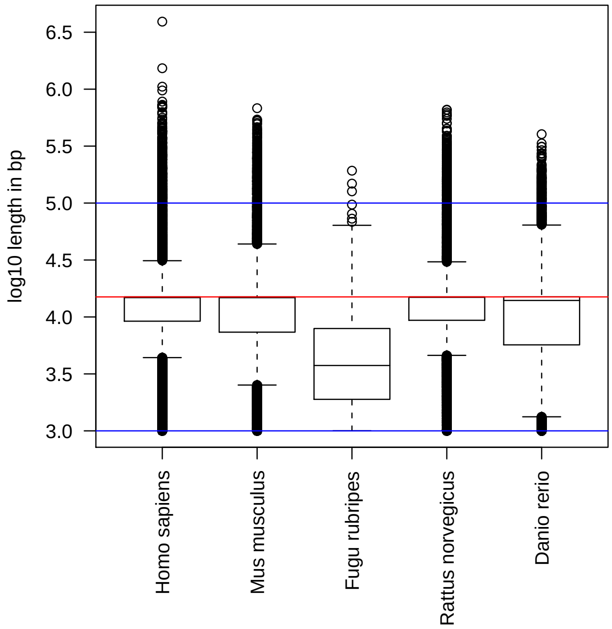


**Figure 4.9: Length distribution of upstream regions in five different species.** The upstream region length is defined by the most downstream coding start (ATG) and the most 5' promoter region. The distribution is plotted in log10 scale. The red line indicates the 15 Kb cutoff for promoter regions. Blue lines mark the upper (100 Kb) and lower (1 Kb) bound on the length of upstream regions.

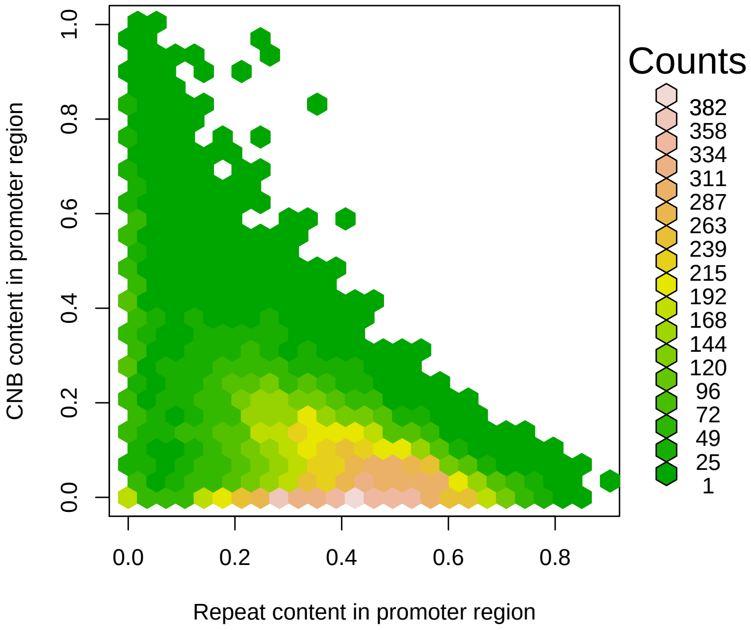


**Figure 4.10: Joint density distributions of repeat and conservation content for human upstream regions.** Proportional coverage of normalised upstream regions with repeats and CNBs is plotted as a 2D-histogram where the density is color-coded in terms of counts in hexagonal bins.

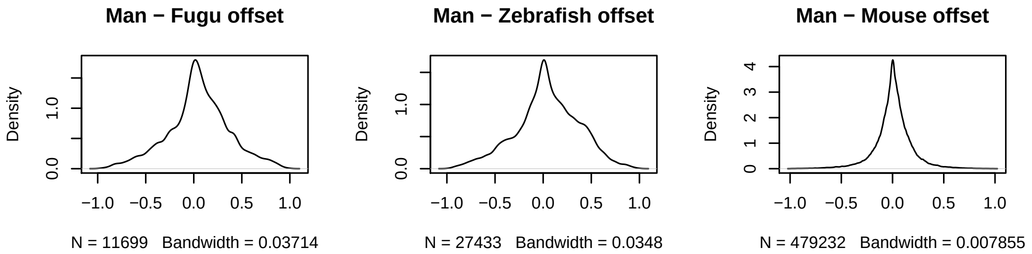


**Figure 4.11: Alignment localization comparison of man to three species.** Gaussian kernel density estimates of the difference of normalized alignment start positions are plotted for three species pairs. `N` is the sample size after elimination of missing values. `Bandwidth` is the standard deviation of the smoothing kernel. The vast majority of alignments is centered around zero. This effect is more pronounced in man-mouse comparison than in man-fish comparisons.