Comparative promoter region analysis powered by CORG.

Christoph Dieterich*[1], Steffen Grossmann[1], Andrea Tanzer[2,3], Stefan Röpcke[1], Peter F. Arndt[1], Peter F. Stadler[2,3]and Martin Vingron[1]

[1] Computational Molecular Biology Department, Max Planck Institute for Molecular Genetics,Ihnestrasse 73, 14195 Berlin, Germany
[2] Institute for Theoretical Chemistry and Structural Biology, University of Vienna, Währingerstrasse 17, A-1090 Wien, Austria
[3] Bioinformatics Group, Department of Computer Science, University of Leipzig, Kreuzstraße 7b, D-04103 Leipzig, Germany

Email: Christoph Dieterich*- christoph.dieterich@molgen.mpg.de; Martin Vingron - martin.vingron@molgen.mpg.de;

*Corresponding author

Abstract

Background: Promoters are key players in gene regulation. They receive signals from various sources (e.g. cell surface receptors) and control the level of transcription initiation, which largely determines gene expression. In vertebrates, transcription start sites and surrounding regulatory elements are often poorly defined. To support promoter analysis, we present CORG (http://corg.molgen.mpg.de), a framework for studying upstream regions including untranslated exons.

Methods: The automated annotation of promoter regions integrates information of two kinds. First, it detects cross-species conservation within upstream regions of orthologous genes. Pairwise as well a multiple sequence comparisons are computed. Second, binding site descriptions (position-weight matrices) are employed to predict conserved regulatory elements. Assembled EST sequences and verified transcription start sites are incorporated to distinguish exonic from other sequences.

Results: As of now, we have included 5 species in our analysis pipeline (man, mouse, rat, fugu and zebrafish). We characterized promoter regions of 16,127 groups of orthologous genes. All data are presented in an intuitive way via our website or can be directly accessed via our DAS server http://tomcat.molgen.mpg.de:8080/das. The benefits of our framework are exemplarily shown in the context of phylogenetic profiling of transcription factor binding sites and detection of microRNAs close to transcription start sites of our gene set.

Conclusions: The CORG platform is a versatile tool to support analyses of gene regulation in vertebrate promoter regions. Applications for CORG cover a broad range from studying evolution of DNA binding sites and promoter constitution to the discovery of new sequence elements (e.g. microRNAs and binding sites).

Motivation

Comparative sequence analysis has been a powerful tool in bioinformatics for addressing a variety of issues. Applications range from grouping of sequences (e.g. protein sequences into families) to *de novo* pattern discovery of functional signatures.

Speaking of gene regulation, it has been known for a long time that there is considerable sequence conservation between species in non-coding regions of the genome. A comprehensive explanation of

this observation is still elusive. However, sequence conservation within promoter regions of genes often stems from transcription factor binding sites that are under selective pressure (see [1] for a review and [2] for a systematic assessment of binding site conservation in man and mouse comparisons).

Conserved sequence elements of other types have recently caught much attention. Not all non-coding conserved DNA in the vicinity of a gene necessarily functions at the level of transcriptional regulation. For example, most known methylation-guide snoRNAs are intron-encoded and processed from transcripts of housekeeping genes [3]. A few microRNAs are apparently linked to protein coding genes, most notably *mir-10* and *mir-196* which are located in the (short) intergenic regions in the *Hox* gene clusters of vertebrates [4–7].

A second class of conserved sequence elements exert their function as regulatory motifs on the primary transcript or the mature mRNA. The `UTRsite` database [8], for example, lists about 30 distinct functional motifs including the Histone 3'UTR stem-loop structure (HSL3) [9], the Iron Responsive Element (IRE) [10], the Selenocysteine Insertion Sequences (SECIS) [11], and the Internal Ribosome Entry Sites (IRES) [12]. Most of these elements do not contain long well-conserved sequence motifs, however, and thus they are typically not detectable as single Conserved Noncoding Blocks (CNBs). Others, such as the Amyloid Precursor Protein mRNA Stability Control Element (APP-SCE) [13], have conserved sequence patterns but do not seem to have strong structural constraints. Hence, they remain indistinguishable from elements that act at DNA level such as transcription factor binding sites.

Phylogenetic footprinting

The CORG framework aims at detecting and describing regulatory elements that are proximal to the transcription start site. In this context, the comparison of upstream regions of orthologous genes is particularily valuable. This concept is called "phylogenetic footprinting" and an excellent overview of this approach can be found in [14].

Phylogenetic footprinting in a strict sense is carried out on orthologous promoter regions. Local sequence similarities can then be directly interpreted as related regions harbouring conserved functional elements. We denote these similarities as Conserved Non-coding Blocks (CNBs). Selecting a suitable set of species is crucial to the footprinting approach since too closely or too distantly related species may show too much or too little sequence conservation.

Multi-species sequence conservation

Comparative approaches gain power from the inclusion of sequences from more than two species [15]. Multi-species comparisons help to increase specificity at the expense of intra-species sensitivity since supporting evidence (conservation) stems from many observations. In CORG, we consider cross-species conservation between promoter regions from 5 vertebrate genomes, namely *Homo sapiens, Mus musculus, Rattus norvegicus, Danio rerio and Fugu rubripes*. Multiple alignments are built from pairwise CNBs as described in the subsequent section. In a pioneering study [16], some light was shed on the extent and abundance of extremly conserved sequence motifs across vertebrate species. MicroRNAs constitute one class of such elements that are found to be conserved across vertebrate species.

Analysis pipeline

Groups of orthologous genes.

In this work, we take a gene-centered view of phylogeny. Homology among proteins and thus genes is often concluded on the basis of sequence similarity. The EnsEMBL database [17] offers an improved way to detect phylogenetic relationships by taking information on conserved synteny into account. We employed single linkage clustering on the graph of EnsEMBL orthologous gene pairs to define the CORG gene groups.

Genomic mapping of validated promoter regions.

Various recent experimental efforts supply information about the position of transcriptional start sites in the human and mouse genome. Table 1 gives an overview on the resources that were employed in CORG.

Some repositories offer genomic coordinates for their start site entries. Existing genomic mapping information was incorporated unless the underlying 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 [18].

Sequence retrieval

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 transcription start site of a gene. Since we do not know the precise location of the start of transcription for each and every gene, we chose to compare the sequence regions upstream of the start of translation from orthologous genes. If verified transcription start sites are known, we define a sequence window that is large enough to hold both, translation and transcription start sites, plus 5kB upstream sequence. In case we lack this information, our observations on known transcription start sites indicate that most promoter regions should be captured in a sequence window of 10 kb size (Supplementary data, Figure 1). 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 is shortened to exclude exonic sequence.

Detection of pairwise local sequence similarities.

Significant local sequence similarities (phylogenetic footprints) in two sequences are computed with an implementation of the Waterman-Eggert algorithm. We have already given an account of the algorithm and statistics in [19, 20]. The underlying alignment scoring scheme is the general reversible model [21]:

$$Q = \begin{pmatrix} \cdot & \alpha\pi_T & \beta\pi_C & \gamma\pi_G \\ \alpha\pi_A & \cdot & \rho\pi_C & \sigma\pi_G \\ \beta\pi_A & \rho\pi_T & \cdot & \tau\pi_G \\ \gamma\pi_A & \sigma\pi_T & \tau\pi_C & \cdot \end{pmatrix} \qquad (1)$$

where we left out the elements on the diagonal, which are constrained by the requirement that the sum of all elements in a row equals zero.

The π_i are the stationary nucleotide frequencies, their sum is constrained to be one. Although the two genomes under consideration are in general not in their stationary state with respect to the substitutional process we take the mean $\pi_i = (\pi_i^{(1)} + \pi_i^{(2)})/2$ of the two observed nucleotide frequencies, $\pi_i^{(1/2)}$, to be the best estimate of the stationary base composition.

From other studies we have further knowledge about the relative rates between transversions, the transition A:T→G:C, and the transition G:C→A:T, which occur in roughly in the ratio 1:3:5 along vertebrate lineages [22]. These ratios of rates would generate sequences with 40% GC in their stationary state. To accomodate the observed nucleotide frequencies π_i we have to allow for deviation from those ratios. We do this by choosing for example $\alpha \propto (R(A \to T)/\pi_T + R(T \to A)/\pi_A)/2$, where $R(i \to j)$ is either 1, 3, or 5 depending on the process under considereation. At the end we scale the matrix Q, such that the PAM distance [23] of the substitution model equals the observed degree of divergence between the two species under comparison.

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 between the species under comparison.

The score for aligning two nucleotides i and j is then $s(i,j) = \log(P(i,j)/(\pi_i\pi_j))$ where $P(i,j)$ is the probability of finding the pairing of i and j under the above substitution model [21].

Joining pairwise into multiple alignments

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 footprint, 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.

In our graph O, cliques of minimal size three are detected with an implementation of the Bron-Kerbosh algorithm [24]. 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. Multiple alignments are then computed based on all cliques that meet the outlined criteria. We chose to employ the multiple alignment method of [25] who applies partial order graphs (POG) to the multiple alignment problem.

Partial order graphs belong to the class of directed acyclic graphs (DAGs). A DAG is a graph consisting of a set of nodes N and edges E, which are one-way edges and form no cycles.

The multiple alignment problem is then reduced to to subsequent alignment steps of individual se-

quences to a growing multiple alignment graph. If the sequences to be aligned share substantial sequence similarity, the number of bifurcation points within the POG stays low and allows rapid computation of the multiple alignment. However, alignment results are sensitive to the input order of sequences.

Alignment results are subsequently trimmed to encompass the leftmost and rightmost ungapped block of at least 6 nucleotides.

Annotation of promoter regions

Exon detection with assembled EST clusters.

Promoter regions in CORG always extend upstream from the most downstream coding start (ATG). As a consequence, promoter regions may contain exons that are not translated. Our way of detecting such exons is a similarity search of man-mouse footprints versus *GENENEST* [26], a database of assembled EST clusters. Database searches are carried out for human and mouse footprints with the `BLASTN` program [27]. An E-value cut-off of 10^{-4} is applied.

Annotation with predicted binding sites.

The TRANSFAC database [28] 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.

Potential binding sites are detected with Transfac weight matrices by the method of [29]. 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. Under both models the distribution of weigth matrix scores can conveniently be calculated by convolution, since the score is a sum of independent random variables. 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. All details are given in [29].

Results

We now present an overview of the web interface to the database and several example applications.

Interface

Individual promoter studies are supported by a graphical interface that provides a user-friendly view of the database. 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, ...). 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 a summary page where the genomic context of the selected gene and its similarities to other upstream regions is visualized.

Pairwise as well as multiple comparisons are displayed on demand at this stage with a JAVA applet that complies with the JDK 1.1 standard. 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 or programs via the distributed annotation system (*DAS*, [30]). 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`).

Additionally, tools for on-site batch retrieval of CORG data will be added to the web portal in the near future.

Phylogenetic profiling of binding sites.

One potential application of CORG is phylogenetic profiling of promoter regions. We define phylogenetic profiling in the context of gene regulation as comparative analysis of presence/absence patterns of binding sites in promoter regions. Here, we consider conserved predicted binding sites and contrast them with validated ones.

Serum Response Factor (SRF) promoter.

SRF, a MADS-box transcription factor, regulates the expression of immediate-early genes, genes encoding several components of the actin cytoskeleton, and cell-type specific genes, e.g. smooth, cardiac and skeletal muscle or neuronal-specific genes [31, 32]. Mouse embryos lacking SRF die before gastrulation and do not form any detectable mesoderm [33, 34]. SRF mediates transcriptional activation by binding to CArG box sequences (Consensus pattern: `CC(AT)`$_6$`GG`) in target gene promoters and by recruiting different co-factors. SRF regulates transcription downstream of MAPK signaling in association with ternary complex factors (TCFs) (for a review see [35]). TCFs bind to ets binding sites present adjacent to CArG boxes in many SRF target gene promoters.

Figure 1 gives an overview of the genomic context of human SRF. As expected, the upstream region of SRF shows substantial conservation to its rodent orthologs. Additionally, significant alignments were found in comparisons with fish homologs (one from zebrafish and two from fugu). The same data ispresented in the multiple alignment view of the JAVA applet in Figure 2a. This view gives a better idea on the location of alignments in the corresponding source sequences. Note, that the spacing between translation start and alignment is greater in fish than in mammals, which hints at different extension of the promoter region in the two subgroups.

We get a better idea on the cause of sequence conservation by browsing the multiple alignment. Textual information can be obtained by clicking on the alignment. Then, the alignment appears in a pop-up window and may be copied to another destination. In Figure 2b, we used CLUSTAL X ([36]) to render the conservation structure on to the nucleotide level. Here, a striking observation is the conservation of the regulatory feedback loop of SRF to its own promoter in all species under consideration.

Non-coding RNAs

Non-coding RNA can be classified as transcribed regulatory elements. Non-coding RNAs are also accessible to the user via the CORG database. Since we were primarily interested in non-coding RNAs rather than small mRNA motifs we restricted our search here to long CNBs. A `blast` search of our multiple alignments with length $L \geq 50$ against the `Rfam` database [37] and the `microRNA Registry` [38] identifies 21 alignments as 7 distinct microRNAs and a single snoRNA, Table 2.

The snoRNA U93 is an unusual mammalian pseudouridinylation guide RNA which accumulates in Cajal (coiled) bodies and it is predicted to function in pseudouridylation of the U2 spliceosomal snRNA [39]. It appears to be specific for mammals. The genomic copy of the human U93 RNA is located in an intron of a series of reported spliced expressed sequence tags (ESTs); furthermore, it has been verified experimentally that U93 is indeed spliced from an intron [39]. It was detectable in the CORG footprint dataset because of its location upstream of a conserved putative gene C14orf87 with unknown function.

The known microRNAs belong to four different groups. The *mir10* and the *mir196* precursors are located at specific positions in the *Hox* gene clusters [4–7]. The *mir-196* family regulates *Hox8* and *Hox7* genes, the function of *mir10* is unknown.

Substitution pattern of non-coding RNAs

For a microRNA we expect a subsequence of about 20nt that is almost absolutely conserved among vertebrates (the mature miRNA) and a well-conserved complementary seqence forming the other side of the stem from which the mature microRNA is excised. In contrast, the substitution rate should be much larger in the loop region of the hairpin [40]. *mir10* is a good example of this typical substitution pattern, which gives rise to a hairpin structure. The pairwise correlation structure of nucleotides is depicted on top of the multiple alignment in Figure 3. A different pattern is observed for the Iron Responsive element in the 5'UTR of *SLCA1*, a member of the sodium transporter family. This time the substitution pattern does not meet the minimal length of the microRNA definition above. Nevertheless, it is conserved across all vertebrate species as shown in Figure 3.

Summary

We have improved and extendend our framework of comparative analysis and annotation of vertebrate promoter regions over previous releases [20]. Our resource contains pairwise and multiple alignments of homologous promoter regions from five vertebrate species. Local promoter regions are annotated with experimental evidence. The prediction of transcription factor binding site was further improved as the rates of false positives and negatives are both taken into account. The CORG database is accessible via a web site. The user is guided step-by-step through the process of selecting and analyzing her promoter region of choice. CORG features an interactive viewer based on JAVA technology, which is tailored to detailed promoter analysis. Large-scale studies make direct use of our DAS service or the MySQL implementation of CORG in conjunction with an application interface (contact authors for details).

We presented selected application examples from the realm of vertebrate gene regulation. Conserved regulatory elements of different kinds (binding sites, microRNAs and UTR elements) are readily accessible to CORG users.

Authors contributions

Christoph Dieterich built the entire pipeline and some parts of the web interface. Steffen Grossmann annotated transcription factor binding sites and provided parts of the web interface. Andrea Tanzer analyzed known and novel RNA elements in the multiple alignments of the CORG database. Stefan Röpcke set up our database of binding site descriptions. Peter F. Arndt worked on an appropriate alignment scoring scheme. Peter F. Stadler and Martin Vingron initiated this work and provided all necessary infrastructure.

References

1. Hardison RC: **Conserved noncoding sequences are reliable guides to regulatory elements.** *Trends Genet* 2000, **16**(9):369–72.
2. Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S: **Eukaryotic regulatory element conservation analysis and identification using comparative genomics.** *Genome Res* 2004, **14**(3):451–458.
3. Bachellerie JP, Cavaillé J, Hüttenhofer A: **The expanding snoRNA world**. *Biochimie* 2002, :775–790.
4. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T: **Identification of novel genes coding for small expressed RNAs**. *Science* 2001, **294**:853–858.
5. Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T: **New microRNAs from mouse and human**. *RNA* 2003, **9**:175–179.
6. Yekta S, Shih Ih, Bartel DP: **MircoRNA-directed cleavage of *HoxB8* mRNA**. *Science* 2004, **304**:594–596.
7. Tanzer A, Amemiya CT, Kim CB, Stadler PF: **Evolution of MicroRNAs Located Within *Hox* Gene Clusters**. *J. Exp. Zool.: Mol. Dev. Evol.* 2004. [Submitted].
8. Pesole G, Liuni S, Grillo G, Licciulli F, Mignone F, Gissi C, Saccone C: **UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs: Update 2002**. *Nucl. Acids Res.* 2002, **30**:335–340.
9. Williams AS, Marzluff WF: **The sequence of the stem and flanking sequences at the 3'end of histone mRNA are critical determinants for the binding of the stemm-loop binding protein**. *Nucl. Acids Res.* 1995, **23**:654–662.
10. Hentze MW, Kuhn LC: **Molecular control of vertebrate iron metabolism: mRNA based regulatory circuits operated by iron, nitric oxide, and oxidative stress**. *Proc. Natl. Acad. Sci. USA* 1996, **93**:8175–8182.
11. Walczak R, Westhof E, P C, Krol A: **A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs**. *RNA* 1996, **2**:367–379.
12. Le SY, Maizel Jr JV: **A common RNA structural motif involved in the internal initiation of translation of cellular mRNAs**. *Nucl. Acids Res.* 1997, **25**:362–369.
13. Zaidi SHE, Malter JS: **Amyloid precursor protein mRNA stability is controlled by a 29-Base element in the 3'-Untranslated region**. *J. Biol. Chem.* 1994, **269**:24007–24013.
14. Duret L, Bucher P: **Searching for regulatory elements in human noncoding sequences.** *Curr Opin Struct Biol* 1997, **7**(3):399–406.
15. McCue LA, Thompson W, Carmack CS, Lawrence CE: **Factors influencing the identification of transcription factor binding sites by cross-species comparison.** *Genome Res* 2002, **12**(10):1523–32.
16. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome.** *Science* 2004, **304**(5675):1321–1325.
17. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, Down T, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz HR, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey

R, Ureta-Vidal A, Woodwark KC, Cameron G, Durbin R, Cox A, Hubbard T, Clamp M: **An overview of Ensembl.** *Genome Res* 2004, **14**(5):925–928.

18. Ning Z, Cox AJ, Mullikin JC: **SSAHA: a fast search method for large DNA databases.** *Genome Res* 2001, **11**(10):1725–1729.

19. Dieterich C, Cusack B, Wang H, Rateitschak K, Krause A, Vingron M: **Annotating regulatory DNA based on man-mouse genomic comparison.** *Bioinformatics* 2002, **18 Suppl 2**:S84–90.

20. Dieterich C, Wang H, Rateitschak K, Luz H, Vingron M: **CORG: a database for COmparative Regulatory Genomics.** *Nucleic Acids Res* 2003, **31**:55–57.

21. Lio P, Goldman N: **Models of molecular evolution and phylogeny.** *Genome Res* 1998, **8**(12):1233–44.

22. Arndt PF, Petrov DA, Hwa T: **Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation.** *Mol Biol Evol* 2003, **20**(11):1887–96.

23. States D, Gish W, Altschul S: **Improved sensitivity of nucleic acid database searches using application-specific scoring matrices.** *Methods: A companion of Methods in Enzymology* 1991, **3**:66–70.

24. Bron C, Kerbosch J: **Algorithm 457. Finding all cliques of an undirected graph**. *Commun. ACM* 1973, **16**:575.

25. Lee C, Grasso C, Sharlow MF: **Multiple sequence alignment using partial order graphs.** *Bioinformatics* 2002, **18**(3):452–64.

26. Krause A, Haas SA, Coward E, Vingron M: **SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein.** *Nucleic Acids Res* 2002, **30**:299–300.

27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389–402.

28. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel A, Kel-Margoulis O: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Research* 2003, **31**:374–378.

29. Rahmann S, Mueller T, Vingron M: **On the Power of Profiles for Transcription Factor Binding Site Detection**. *Statistical Applications in Genetics and Molecular Biology* 2003, **2**:7.

30. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L: **The distributed annotation system.** *BMC Bioinformatics* 2001, **2**:7.

31. Miano JM: **Serum response factor: toggling between disparate programs of gene expression.** *J Mol Cell Cardiol* 2003, **35**(6):577–93.

32. Treisman R: **Journey to the surface of the cell: Fos regulation and the SRE.** *EMBO J* 1995, **14**(20):4905–13.

33. Arsenian S, Weinhold B, Oelgeschlager M, Ruther U, Nordheim A: **Serum response factor is essential for mesoderm formation during mouse embryogenesis.** *EMBO J* 1998, **17**(21):6289–99.

34. Weinhold B, Schratt G, Arsenian S, Berger J, Kamino K, Schwarz H, Ruther U, Nordheim A: **Srf(-/-) ES cells display non-cell-autonomous impairment in mesodermal differentiation.** *EMBO J* 2000, **19**(21):5835–44.

35. Buchwalter G, Gross C, Wasylyk B: **Ets ternary complex transcription factors.** *Gene* 2004, **324**:1–14.

36. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs.** *Nucleic Acids Res* 2003, **31**(13):3497–500.

37. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy S: **Rfam: an RNA family database**. *Nucl. Acids Res.* 2003, **31**:439–441.

38. Griffiths-Jones S: **The microRNA Registry**. *Nucl. Acids Res.* 2004, **32**:D109–D111. [Database Issue].

39. Kiss AM, Jády BE, Darzacq X, Verheggen C, Bertrand E, Kiss T: **Cajal body-specific pseudouridylation guide RNA is composed of two box H/ACA snoRNA-like domains**. *Nucl. Acids Res.* 2002, **30**:4643–4649.

40. Lai EC, Tomancak P, Williams RW, Rubin GM: **Computational identification of *Drosophila* microRNA genes**. *Genome Biol.* 2003, **4**:R42 (20 pages).

41. Spencer JA, Major ML, Misra RP: **Basic fibroblast growth factor activates serum response factor gene expression by multiple distinct signaling mechanisms.** *Mol Cell Biol* 1999, **19**(6):3977–88.

42. Washietl S, Hofacker IL, Stadler PF: **A fast, efficient, and reliable method for the detection of structurally conserved RNA on a genome-wide scale** 2004. [In preparation].

43. Schmid CD, Praz V, Delorenzi M, Perier R, Bucher P: **The Eukaryotic Promoter Database EPD: the impact of in silico primer extension.** *Nucleic Acids Res* 2004, **32 Database issue**:D82–5.

44. Suzuki Y, Yamashita R, Sugano S, Nakai K: **DBTSS, DataBase of Transcriptional Start Sites: progress report 2004.** *Nucleic Acids Res* 2004, **32 Database issue**:D78–81.

45. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, Yura K, Miyazaki S, Ikeo K, Homma K, Kasprzyk A, Nishikawa T, Hirakawa M, Thierry-Mieg J, Thierry-Mieg D, Ashurst J, Jia L, Nakao M, Thomas MA, Mulder N, Karavidopoulou Y, Jin L, Kim S, Yasuda T, Lenhard B, Eveno E, Suzuki Y, Yamasaki C, Takeda J, Gough C, Hilton P, Fujii Y, Sakai H, Tanaka S, Amid C, Bellgard M, Mde FBM, Bono H, Bromberg SK, Brookes AJ, Bruford E, Carninci P, Chelala C, Couillault C, Souza SJ, Debily MA, Devignes MD, Dubchak I, Endo T, Estreicher A, Eyras E, Fukami-Kobayashi K, o p i n a t h R G, Graudens E, Hahn Y, Han M, Han ZG, Hanada K, Hanaoka H, Harada E, Hashimoto K, Hinz U, Hirai M, Hishiki T, Hopkinson I, Imbeaud S, Inoko H, Kanapin A, Kaneko Y, Kasukawa T, Kelso J, Kersey P, Kikuno R, Kimura K, Korn B, Kuryshev

V, Makalowska I, Makino T, Mano S, Mariage-Samson R, Mashima J, Matsuda H, Mewes HW, Minoshima S, Nagai K, Nagasaki H, Nagata N, Nigam R, Ogasawara O, Ohara O, Ohtsubo M, Okada N, Okido T, Oota S, Ota M, Ota T, Otsuki T, Piatier-Tonneau D, Poustka A, Ren SX, Saitou N, Sakai K, Sakamoto S, Sakate R, Schupp I, Servant F, Sherry S, Shiba R, Shimizu N, Shimoyama M, Simpson AJ, Soares B, Steward C, Suwa M, Suzuki M, Takahashi A, Tamiya G, Tanaka H, Taylor T, Terwilliger JD, Unneberg P, Veeramachaneni V, Watanabe S, Wilming L, Yasuda N, Yoo HS, Stodolsky M, Makalowski W, Go M, Nakai K, Takagi T, Kanehisa M, Sakaki Y, Quackenbush J, Okazaki Y, Hayashizaki Y, Hide W, Chakraborty R, Nishikawa K, Sugawara H, Tateno Y, Chen Z, Oishi M, Tonellato P, Apweiler R, Okubo K, Wagner L, Wiemann S, Strausberg RL, Isogai T, Auffray C, Nomura N, Gojobori T, Sugano S: **Integrative annotation of 21,037 human genes validated by full-length cDNA clones.** *PLoS Biol* 2004, **2**(6):E162.

46. Bono H, Kasukawa T, Furuno M, Hayashizaki Y, Okazaki Y: **FANTOM DB: database of Functional Annotation of RIKEN Mouse cDNA Clones.** *Nucleic Acids Res* 2002, **30**:116–118.

47. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29**:137–40.

Figures

Figure 1 - Genomic context of human SRF

This image is displayed after the user selected a gene identifier on the search page. It provides the user with the genomic context of the selected gene. Known and predicted transcription start sites are shown as labelled red dots. Local similarities to homologous regions from other species are shown as connected purple boxes. Blue bars depict all upstream regions as contained in CORG. The structure of the corresponding EnsEMBL transcripts as well as the extent of RefSeq transcripts is show in the bottom track.

Figure 2 - Graphical and textual multiple alignment view.

(a) Multiple alignment view of 6 sequences from 5 species.

All consistent local similarities in the upstream region of SRF homologs are placed relative to the species-specific translation start sites. The distance of the aligned segment to the translation start site is almost equal for all mammals and larger for the fish. The extent of each upstream region is shown as orange bar. Regions covered by flanking genes would be shown in red.

(b) Multiple alignment as rendered by CLUSTAL X.

The largest multiple alignment was retrieved from the JAVA applet by a cut and paste operation and rendered in CLUSTAL X [36]. Conserved binding sites are highlighted by red or blue boxes. Known sites as given in TRANSFAC are marked with a dollar sign [41]. Note that the validated Egr-1 site is only conserved in mammals. This site is bound by the serum-inducible Krox-24 zinc finger protein.

Figure 3 - Alignments and predicted RNA structure of two transcribed regulatory elements.

The *mir-10b* CNB (top) shows the typical pattern of substitutions in a microRNA precursor hairpin: There are two well-conserved arms, of which the mature microRNA is almost absolutely conserved, and a much more variable loop region. The Iron Responsive Element (bottom) shows a very different substitution pattern. Additional orthologous sequences from the frog *Xenopus tropicalis* (xtr), the chicken *Gallus gallus* (gga) and the pufferfish *Tetraodon nigroviridis* are included. [42].

Tables

Table 1 - Resources for validated transcription start sites

Resources for validated transcription start sites	
Database name	**Features**
Eukaryotic promoter database (EPD) [43]	The Eukaryotic promoter database is the smallest in size, but largely consists of manually curated entries.
DataBase of Transcriptional Start Sites (DBTSS) [44]	The DBTSS contains 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) [45]	H-InvDB is 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.
FANTOM 2 collection of full-length cDNAs (RIKEN) [46]	The RIKEN consortium presented the FANTOM collection of RIKEN full-length cDNA clones. FANTOM stands for Functional Annotation of Mouse cDNA clones.
The Reference Sequence project (RefSeq) [47]	The Reference Sequence project aims to provide a comprehensive, integrated, non-redundant set of sequences, including full-length transcripts (mRNA)

Table 2 - Rfam non-coding RNAs in CORG

A + sign indicates that a sequence fragment from the corresponding species (hsa *Homo sapiens*, mmu *Mus musculus*, rno *Rattus norvegicus*, dre *Danio rerio*, tru *Takifugu rubripes*) is contained in the CORG CNB; $\emptyset$ indicates that a `blast` search for an orhologous sequence in the Ensemble database was unsuccessful; n.d. mean no descriptive Ensemble gene annotation. The CNBs containing *mir-196a-2* are shifted compared to the known microRNA sequences, preventing the detection of the correct stem-loop structure. The `B` columns marks whether a candidate was identified by a `blast` search against the `Rfam` or `microRNA Registry`, the `A` column shows whether a hairpin structure was identified by `RNAalifold`.$p_{\texttt{RNAz}}$ is the p-value for being an evolutionary conserved RNA secondary structure element returned by `RNAz`.

CNB	B	A	$p_{\texttt{RNAz}}$	ncRNA	hsa	mmu	rno	dre	tru	gene
119596	+	+	0.995	mir-34c	+	+	+	+	∅	n.d. (BCT-4)
119607	+	+	0.938							mir-34b in hsa
119658	+	+	0.985							
159914	+	+	0.998	mir-138-2	+	+	+	+	∅	SLC12A3, n.d. in teleosts
159932	+	+	0.999							
159939	+	+	0.998							
194777	+	+	0.998	mir-196b	+	—	+	+	+	HOXA9, dre: HOXA9a and HOXA9b
194820	+	+	0.999							
194839	+	+	0.999							
194941	+	+	0.999							
226470	+	+	0.999	mir-10a	+	+	+	+	+	HOXB4, dre: HOXB4a and HOXB4b
226514	+	+	0.999							
226555	+	+	0.999							
226677	+	—	0.004							
238163	+	+	0.992	mir-10b	+	+	+	+	+	HOXD4, dre: HOXD4a, n.d. in tru
238188	+	+	0.984							
238265	+	+	0.994							
391314	+	—	0.125	mir-196a-2	+	+	—	+	+	HOXC9, dre: HOXC9a
391315	+	—	0.999							
391318	+	—	0.511							
470004	+	—	0.218	U93	+	+	+	∅	+	n.d.
110374	—	+	0.995	IRES ?	+	+	+	+	+	DGCR8
146100	—	+	0.891		+	+	+	+	∅	Ptf1a
393794	—	+	0.999	IRE	+	+	+	+	+	SLCA1

Table 3 - Candidates in UTRdb
The column "pos.(element)" indicates the position of the annotated element in the `UTRdb` entry, pos.(candidate) is the position of the CORG CNB in the same `UTRdb` entry determined by a `blast` alignment.

CNB	type	UTRdb ID	EMBL GeneID	pos.(element)	pos.(candidate)
110374	IRES	BB277285	BC009984	244..350	134..220
393794	IRE	BB236186	BC037733	203..229	182..254

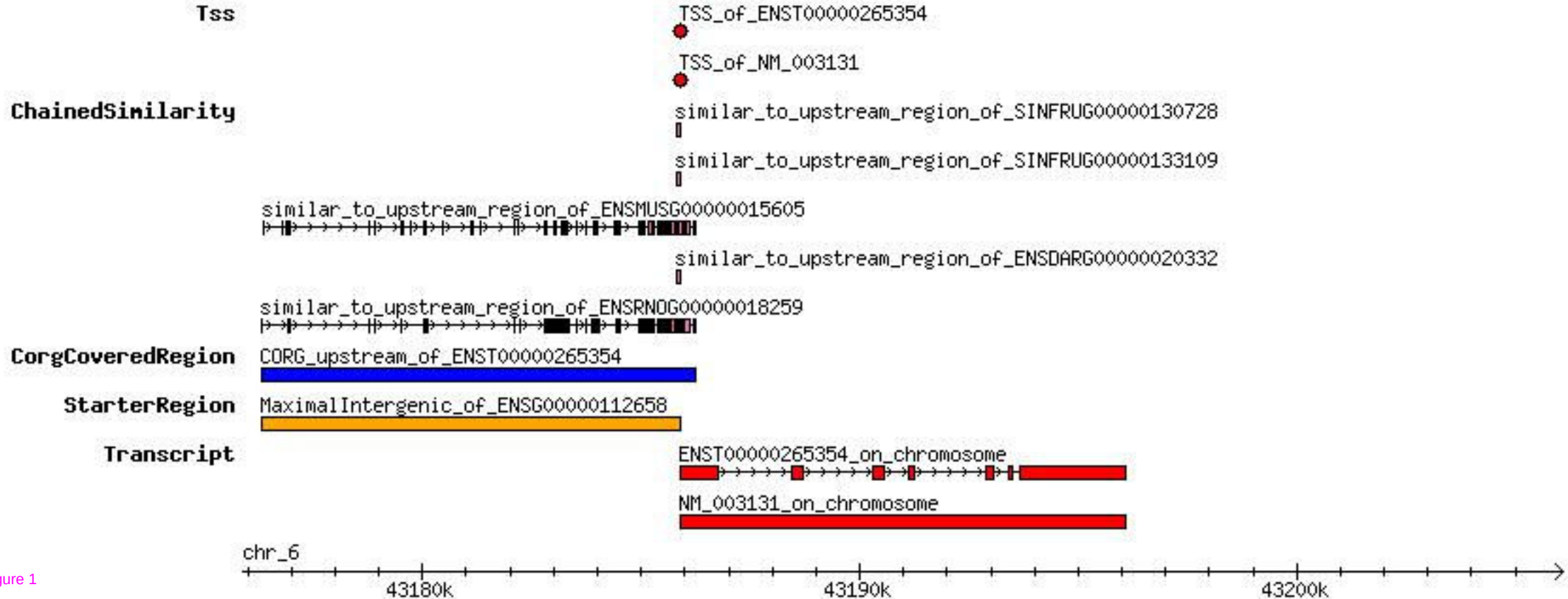
Tss
TSS_of_ENST00000265354
TSS_of_NM_003131
ChainedSimilarity
similar_to_upstream_region_of_SINFRUG00000130728
similar_to_upstream_region_of_SINFRUG00000133109
similar_to_upstream_region_of_ENSMUSG00000015605
similar_to_upstream_region_of_ENSDARG00000020332
similar_to_upstream_region_of_ENSRNOG00000018259
CorgCoveredRegion
CORG_upstream_of_ENST00000265354
StarterRegion
MaximalIntergenic_of_ENSG00000112658
Transcript
ENST00000265354_on_chromosome
NM_003131_on_chromosome
chr_6
43180k
43190k
43200k

Figure 1

JAVA Sequence Viewer - group15588_zebrafish_-_ENSDARG00000020332

File Options Help

Java Applet Window

group15588_zebrafish_-_ENSDARG00000020332 / Number of feature types: 6

zebrafis
100 %

mouse_CNBS
100 %

rat_CNBS
100 %

human_CNBS
100 %

fugu_CNBS
100 %

fugu_CNBS
100 %

-1142 -974 -806 -638 -470 -302 -134

Zoom_in Zoom_default Zoom_out From: 7283 To: 9938

Figure 2

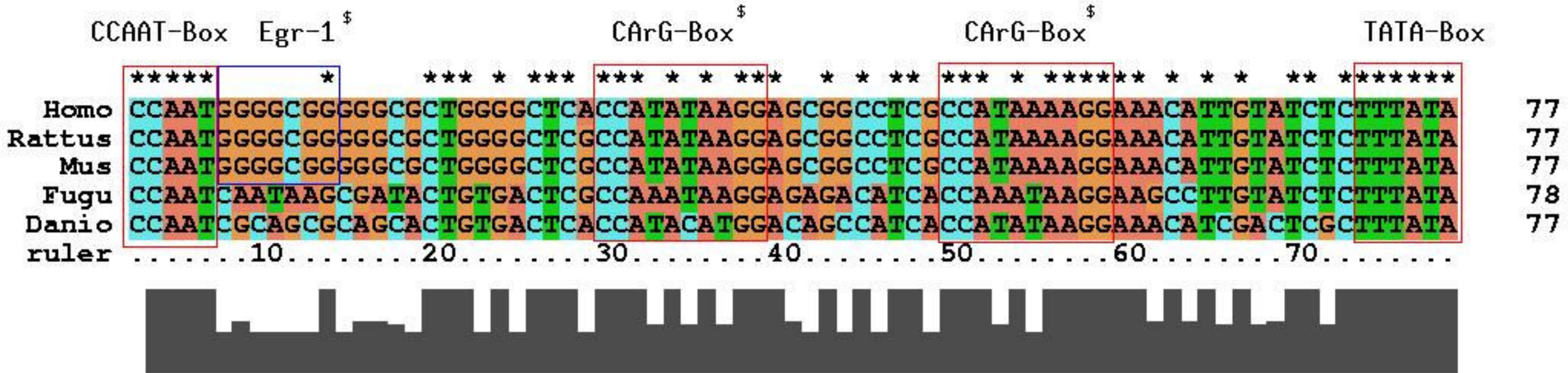
CCAAT-Box
Egr-1 $
CArG-Box $
CArG-Box $
TATA-Box
Homo CCAATGGGGCGGGGGCGCTGGGGCTCACCATATAAGGAGCGGCCTCGCCATAAAAGGAAACATTGTATCTCTTTATA 77
Rattus CCAATGGGGCGGGGGCGCTGGGGCTCGCCATATAAGGAGCGGCCTCGCCATAAAAGGAAACATTGTATCTCTTTATA 77
Mus CCAATGGGGCGGGGGCGCTGGGGCTCGCCATATAAGGAGCGGCCTCGCCATAAAAGGAAACATTGTATCTCTTTATA 77
Fugu CCAATCAATAAGCGATACTGTGACTCGCCAAATAAGGAGAGACATCACCAAATAAGGAAGCCTTGTATCTCTTTATA 78
Danio CCAATCGCAGCGCAGCACTGTGACTCACCATACATGGACAGCCATCACCATATAAGGAAACATCGACTCGCTTTATA 77
ruler .....10.....20.....30.....40.....50.....60.....70.....

Figure 3

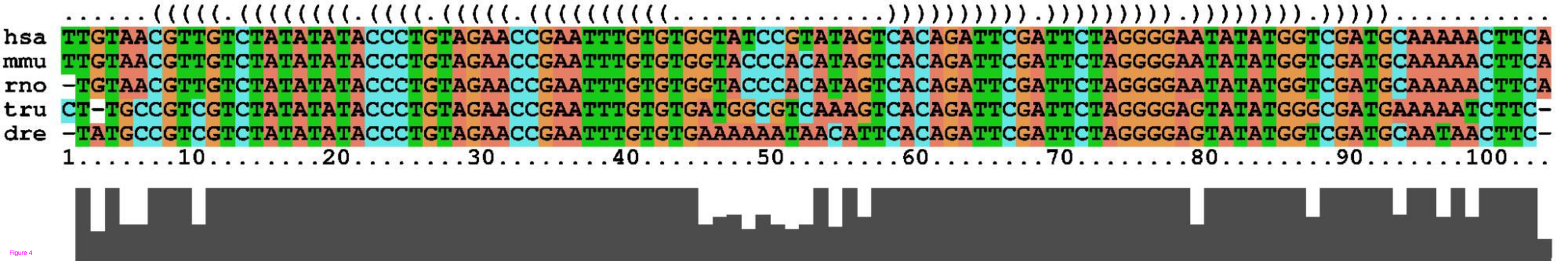
.....((((.(((((((.((((.((((.(((((((((.....)))))))))))))))))))))))))))).....
hsa TTGTAACGTTGTCTATATATACCCTGTAGAACCGAATTTGTGTGGTATCCGTATAGTCACAGATTCGATTCTAGGGGAATATATGGTCGATGCAAAAACTTCA
mmu TTGTAACGTTGTCTATATATACCCTGTAGAACCGAATTTGTGTGGTACCCACATAGTCACAGATTCGATTCTAGGGGAATATATGGTCGATGCAAAAACTTCA
rno -TGTAACGTTGTCTATATATACCCTGTAGAACCGAATTTGTGTGGTACCCACATAGTCACAGATTCGATTCTAGGGGAATATATGGTCGATGCAAAAACTTCA
tru CT-TGCCGTCGTCTATATATACCCTGTAGAACCGAATTTGTGTGATGGCGTCAAAGTCACAGATTCGATTCTAGGGGAGTATATGGGCGATGAAAAATCTTC-
dre -TATGCCGTCGTCTATATATACCCTGTAGAACCGAATTTGTGTGAAAAAATAACATTCACAGATTCGATTCTAGGGGAGTATATGGTCGATGCAATAACTTC-
1.....10.....20.....30.....40.....50.....60.....70.....80.....90.....100...

Figure 4

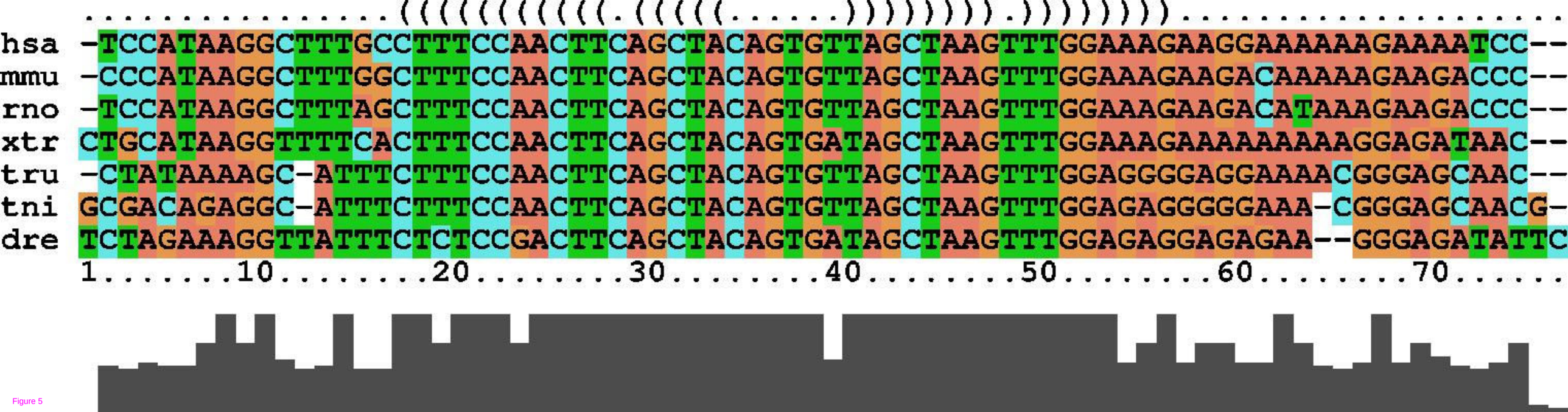
.....(((((((((((.(((((.....)))))))))).)))))))...................
hsa -TCCATAAGGCTTTGCCTTTCCAACTTCAGCTACAGTGTTAGCTAAGTTTGGAAAGAAGGAAAAAAGAAAATCC--
mmu -CCCATAAGGCTTTGGCTTTCCAACTTCAGCTACAGTGTTAGCTAAGTTTGGAAAGAAGACAAAAAGAAGACCC--
rno -TCCATAAGGCTTTAGCTTTCCAACTTCAGCTACAGTGTTAGCTAAGTTTGGAAAGAAGACATAAAGAAGACCC--
xtr CTGCATAAGGTTTTCACTTTCCAACTTCAGCTACAGTGATAGCTAAGTTTGGAAAGAAAAAAAAAGGAGATAAC--
tru -CTATAAAAGC-ATTTCTTTCCAACTTCAGCTACAGTGTTAGCTAAGTTTGGAGGGGAGGAAAACGGGAGCAAC--
tni GCGACAGAGGC-ATTTCTTTCCAACTTCAGCTACAGTGTTAGCTAAGTTTGGAGAGGGGGAAA-CGGGAGCAACG-
dre TCTAGAAAGGTTATTTCTCTCCGACTTCAGCTACAGTGATAGCTAAGTTTGGAGAGGAGAGAA--GGGAGATATTC
1.....10.....20.....30.....40.....50.....60.....70.....

Figure 5