Picking Up the Trail of Phylogenetic Footprints

The Story of Tracker and Its First Tracings

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Magistra rerum naturalium

an der Fakultät für Naturwissenschaften und Mathematik
der Universität Wien

Vorgelegt von

Sonja J. Prohaska

Juli 2003

Abstract

Non-coding sequence in eukaryotes encodes functionally important signals for the regulation of gene expression. Since these elements are evolutionary conserved among related taxa due to stabilizing selection they accumulate less mutations than adjacent non-functional DNA. These conserved non-coding sequences with potential regulatory activity, are known as *phylogenetic footprints*. They can be detected by comparison of the sequences surrounding orthologous genes in (distantly) related species. Loss or acquisition of phylogenetic footprints in some of these lineages provides evidence for the evolutionary modification of cis-regulatory elements. In order to observe the distribution of phylogenetic footprints among whole gene clusters and various species we developed the software tool `tracker`. It is designed for large scale analysis and identifies corresponding footprints in long sequences from multiple species more efficiently than other available algorithms.

We apply our novel method to the published sequences of *HoxA* clusters to study the footprint evolution after the most recent cluster duplication in *Danio rerio* (zebrafish) and *Takifugu rubripes* (pufferfish). We introduce a statistical model that allows us to estimate the loss of non-coding sequence conservation that can be attributed to gene loss and other structural reasons. According to this model we observe an unexpectedly high loss of sequence conservation suggesting that binding site turnover and/or adaptive modification also contribute to the massive loss of sequence conservation.

The statistical analysis of phylogenetic footprints in the two known *Hox* clusters of *Heterodontus francisci* (horn shark) and the four mammalian *Hox* clusters *A,B,C* and *D* shows that the shark *HoxN* cluster is *HoxD*-like. From this finding we conclude that the most recent common ancestor of gnathostomes (jawed vertebrates) had at least four *Hox* clusters, including those which are orthologous to the four mammalian *Hox* clusters.

Within the intergenic region from *hoxA13* to *hoxA11* we discovered a set of footprints specifically conserved among the available representatives of the tetrapods. Since exclusive expression domains of *hoxA13* and *hoxA11* are known to determine limb development we propose that these footprints are crucial for the fin limb transition. Consequently we predict the presence of homologous footprints in amphibians and reptiles.

Zusammenfassung

In nicht-kodierender DNA von Eukaryoten finden sich funktionell wichtige Signale für die Regulation der Genexpression, die durch stabilisierende Selektion in verwandten Taxa evolutionär konserviert sind. Sie akkumulieren weniger Mutationen als benachbarte nicht-kodierende Regionen und werden gemeinhin als 'phylogenetische Fußstapfen' (engl. *"phylogenetic footprints"*) bezeichnet. Durch Vergleich nicht-kodierender Bereiche in der Umgebung von orthologen Genen verwandter Arten, werden solche phylogenetischen Fußstapfen aufgefunden. Der Verlust von konservierten Fußstapfen, sowie das Erlangen neuer Motive gibt Auskunft über evolutionäre Änderungen an *cis*-regulatorischen Sequenzen. Um die Verteilung von konservierten Regionen entlang gesamter Gen-Cluster und einer Vielzahl von Organismen zu untersuchen, entwickelten wir das Programm `tracker`. Es ist in der Lage, Analysen an einer Vielzahl langer Sequenzen durchzuführen und identifiziert zusammengehörige Fußstapfen effizienter als andere verfügbare Programme.

Eine erste Anwendung fand unser neues Programm bei der Untersuchung der rezentesten Duplication im *HoxA* Cluster von *Danio rerio* (Zebrafisch) and *Takifugu rubripes* (Kugelfisch). Zusammen mit dem vorgestellten, statistischen Modell zur Abschätzung des Verlustes nicht-kodierender Sequenzkonservierung, der durch Gen Verlust und strukturelle Änderungen erklärt wird, schlagen wir vor den unerwartet hohe Verlust von konservierten Fußstapfen durch co-evolutiven Änderung von Bindungsstellen und adaptive Modifikation zu interpretieren.

Statistische Analysen phylogenetischer Fußstapfen in den beiden bekannten *Hox* Clustern von *Heterodontus francisci* (Hornhai) und den vier *Hox* Clustern *A,B,C* und *D* in Säugetieren zeigten, daß der *HoxN* Cluster des Haies dem *HoxD* Cluster am ähnlichsten ist. Aus dieser Beobachtung schließen wir, daß der rezenteste gemeinsame Vorfahre der Gnathostoma (Kiefertragende) mindesten vier *Hox* Cluster hatte, von denen vier ortholog zu den vier Clustern in Säugetieren sind.

Die weitere Analye der intergenischen Region von *hoxA13* bis *hoxA11* ergab eine Reihe von Fußstapfen, die spezifisch für alle verfügbaren Vertreter der Tetrapoda (Vierfüßer) sind. Da die nicht überlappende Expression von *hoxA13* und *hoxA11* die Entwicklung von Extremitäten bestimmt, ist anzunehmen, daß diese Fußstapfen für die Entwicklung von Flossen zu Extremitäten entscheidend sind. Daraus schließen wir, daß die Konservierung der Sequenzstücke auch in Amphibien und Reptilien zu erwarten ist.

Acknowledgements

Co-work with Claudia Fried, helpful discussion with Günter P. Wagner, Christoph Flamm, Ivo Hofacker, Chris T. Amemiya, Chi-Hua Chiu, Frank H. Ruddle, Guido Fritzsch, amicable supervision by Peter Stadler and computational assistance by Paul Perco in the initiation phase are greatfully acknowledged.

Contents

1 **Introduction** **1**

2 **The Tracker Method** **7**

2.1 The Three Wishes 7

2.2 The Intestines of Tracker 8

2.2.1 Initial Set of Pairwise Alignments 8

2.2.2 Consistent Cliques 11

2.2.3 Postprocessing 16

2.2.4 Implementation 16

2.3 Program Performance 18

2.3.1 Detection of Regulatory Elements 18

2.3.2 The Test Set 19

2.3.3 Available Methods - how suitable are they? 20

2.4 Improvements of Tracker 36

2.4.1 Inputfiles 36

2.4.2 Automatic Detection of Orthologous Regions 36

2.4.3 Clustering Footprints 37

2.4.4 Sorting Footprints 37

2.4.5 Significanc Measure 37

2.4.6 Counting Module 38

2.4.7 System Compatibility and Weberver 38

3 **The Destiny of Duplicated Footprints** **41**

3.1 The DDC Model 41

3.2 Structural Footprint Loss 42

3.3 Footprint Retention Statistics in Teleosts 44
3.3.1 Observed Footprint Retention 44
3.3.2 Estimation of Non-Structural Causes 46
3.4 Evidence for Adaptive Modification 50

4 Assignment of Orthologous Hox Clusters 53
4.1 Models for Hox Cluster Evolution 53
4.2 Footprints as Phylogenetic Information 54
4.3 The Shark HoxN Cluster 55

5 The Fin Limb Transition 63
5.1 Origin of the Tetrapod Limb 63
5.2 Footprint Detection 64
5.3 The Tracks of Tetrapods 65
5.4 Zooming into Highly Conserved Cliques 66

6 Conclusion 71

7 Outlook 73
7.1 Detecting Protein Binding Sites 73
7.2 Biological Challenges Lying Ahead 73
7.2.1 Distances of Footprints 73
7.2.2 The Importance of Colinearity 74
7.2.3 Major Evolutionary Transitions 76

Formende Worte

Hier texturt der Text in Schwall und Wort zu makeln der Schönheit Tum. Wird er sich gefällig mit dem Opfer brechen einen neuen Weg aufzuzeilen?

gewidmet den beiden LaTeX-Experten

Roman Stocsits und Peter Stadler

CHAPTER 1

Introduction

One of the major principles in biology that still did not get out of fashion is to sloppily copy information. This may seem strange since one does not expect to introduce improvements at random. But everyone who has ever copied a homework at highschool or observed others doing so for a long enough time has learned that it is worth taking the risk of slightly unprofitable modifications to experiences the benefit of rare striking changes at least once. In general, after prove of the benefit the innovation is selected, copied and disseminated to serve as template in another round of multiplication bringing fame to its creator.

The biological term for this principle is 'evolution', even though it is valid in many situations of life it refers to the process of modification and selection of DNA sequences and the corresponding changes of the phenotype. Some of the DNA modifications do not cause changes in the phenotype and are said to be neutral. The others concern functional DNA and therefore lead to functional changes and 'visible', phenotypic differences. During the last centuries, geneticists focused on protein coding regions (genes) and their function. By means of mutagenesis they modified the coding region and analyzed the phenotypes to further assign functions to the coding regions. As the term 'non-coding DNA' implies, they thought of it as non-functional, spacing DNA. Nowadays, we know that regulatory sequences in these regions occupy a much larger fraction of genomic sequences than the protein coding counterpart [49]. Furthermore, they are at least as important as the coding regions themselves because of their dramatic effect on the function of the regulated genes. Mutations at these regulatory sites can cause the establishment of different expression patterns and tremendous phenotypic changes.

Experimental evidence from a variety of sources shows that a major mode of evolution is based on the modification of cis-regulatory elements [5, 19, 24, 70]. The most famous data concern the *Hox* clusters of developmental genes [53]. These are clusters of paralogous genes coding for homeodomain containing transcription factors with widespread conservation among bilateral metazoans. They participate in the formation of the anterior-posterior axes of the embryo and determine the positional identities of segments and matter for the development of their morphology. *Hox* genes have critical effects on nearly all functional groups in a vertebrate body.

This turns *Hox* clusters into an interesting research subject as well as an appropriate model system for various reasons:

First, changes of regulatory regions in the cluster of genes with general importance in the pattering of the body will unerringly cause visible morphological changes. This is due to the activation of the first *hox* gene before primitive streak formation a developmental stage at which structural changes can already be analyzed.

Second, the *Hox* cluster are common in a wide range of different structured organisms with still highly conserved genes and cluster composition. This attracts the attention of theoretical biologists dealing with the homology of characters [67]. Changes in the level and sequence of *hox* gene expression result in the change of archetypes. The discovery of homologous genes determining non-homologous morphological structures in 'non'-related groups is a new challenge to morphologists studying the problem of homology.

Third, to compare cluster sequences one aligns the sequences to identify regions of similarity and divergence. Alignment methods tend to work more efficiently on very similar sequences but less similar sequences are quiet more interesting from the biological point of view. Thus using conserved *hox* proteins assists the alignment of the less conserved surrounding DNA and this increases the quality of the alignment.

Forth, when many taxonomically well placed species are sequenced, it may be possible to assign regulatory changes to major evolutionary changes, for example to the fin limb transition [76] or the origin of mammals.

Finally, a particular feature of *hox* gene expression is that the genes in each cluster are expressed in a temporal and spatial order that reflects there order on the chromosome (Fig. 1.1). There is experimental evidence that 3' genes are turned on in the anterior part of the body in early stages of development regulated by retinoic acid which binds to RARE sites. During proceeding embryogenesis more 5' genes are expressed in the posterior parts (or distal parts concerning limb development) due to the influence of *cdx*. The reason for colinearity is not known yet. One plausible idea is that enhancer

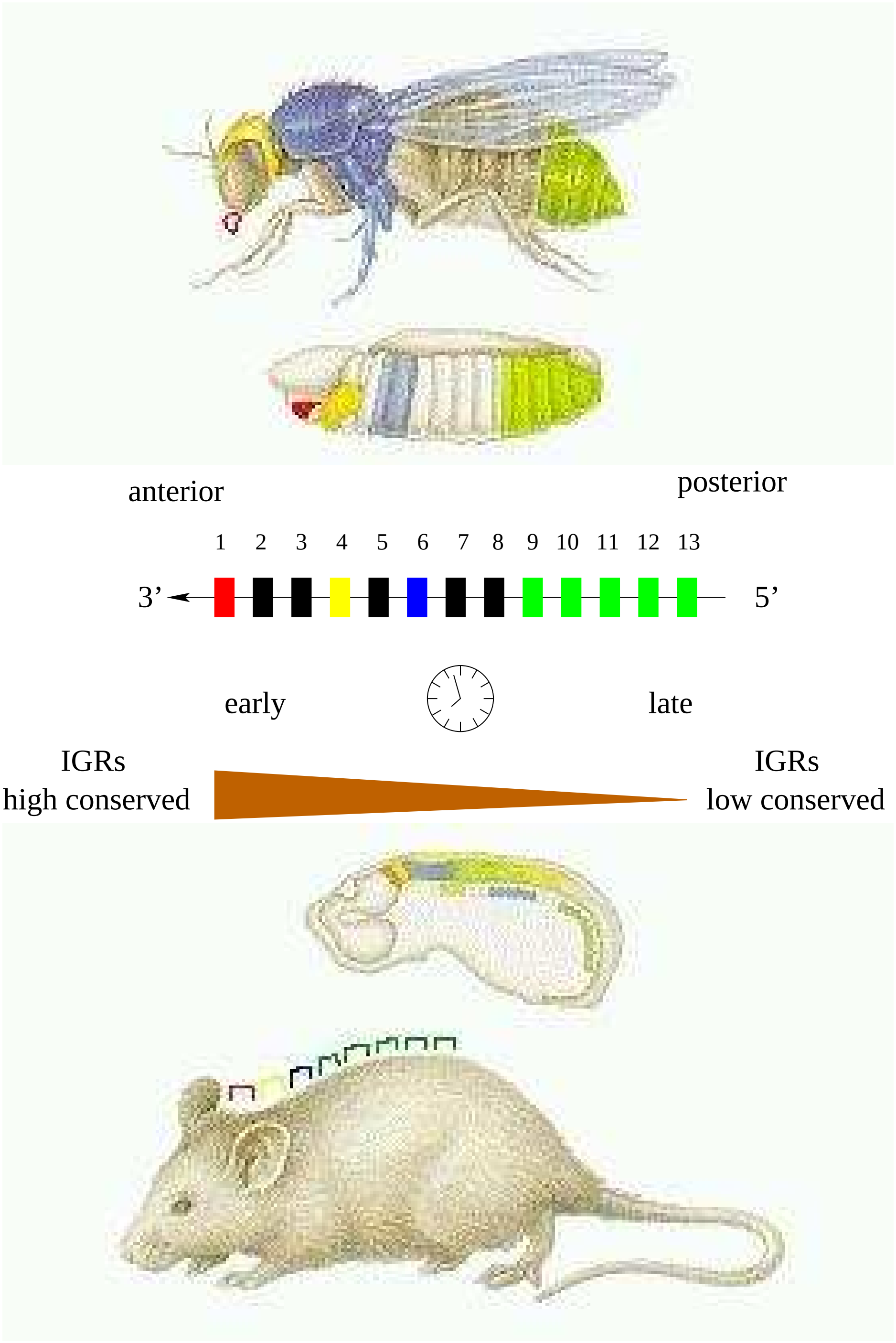


Figure 1.1: Colinearity in *Hox* clusters. The 13 genes of the different paralogous groups of a *Hox* cluster are expressed in spatial and temporal order colinear to the order on the chromosome. The 3' *hox* genes are expressed in the anterior body parts at early developmental stages. As recently reported, the intergenic regions (IGRs) located 3' in the cluster are more conserved.

shearing prevents that the cluster breaks up. For the *HoxD* cluster Kmita *et al.* show quantitative colinearity [45] which is due to a general enhancer element and its decreasing activity with rather gene distance than gene identity. Santini *et al.* found out that intergenic regions between genes located 3' in the *Hox* cluster are significantly more conserved [63]. That explains the ability of functional complementation observed for 3' genes by Manzanares *et al.*[53]. The conclusion is that the body is first roughly patterned by the highly conserved genes located 3' in the cluster which are also necessary for the correct activation of the subsequent Hox expression system. With increasing distance from the core of the body, the level of non-coding sequence conservation decreases in the vicinity of 5' genes in contrast to the evolutionary diversity of these distal body parts. The underlying mechanism is not well understood and captures the attention of developmental geneticists [45, 60].

It is difficult to investigate the molecular evolution of cis-regulatory elements because of the absence of a reliable "genetic code for non-coding sequences" [10]. Binding sites for transcription factors are usually short and variable or fuzzy and are thus hard to identify unambiguously, in particular if the transcription factors involved are not known *a priori* [73]. Moreover, there is good evidence that not only the nucleotide sequence defines the activity of a binding site [46]. Additional constraints such as their spatial distribution, relative or absolute distances and helical phasing between adjacent binding sites is often suggested to play an important role. Little work, however, has been carried out to investigate spatial constraints.

An experimental way to determine protein binding sites is called interference footprinting (Fig. 1.2a). In this case, DNA is modified and the effect on recognition by the protein is studied. The nucleotides which cannot be modified without losing the binding of the protein represent a *footprint*. Phylogenetic footprinting is the computational attempt to find regulatory elements assuming that functional important parts of non-coding sequences as well as coding sequences evolve much slower than the adjacent non-functional DNA due to selective pressure. Evolution modifies the sequences stochastically while negative selection prevents mutations in functional regions to get fixed in the population. Therefore, conserved regions detected in a carefully selected set of related species reveal functional DNA if the sequences are divergent enough so that stochastic conservation is highly unlikely (Fig. 1.2b). Phylogenetic footprints can therefore be viewed as islands of strongly conserved segments in non-coding sequences [72].

The above assumption is very important for phylogenetic footprinting. It is basic for the detection of conserved non-coding DNA after 300-450 million years of evolution [21]. Unfortunately, it is not exceptional that experimentally characterized regulatory elements fail to show sequence similarity even between

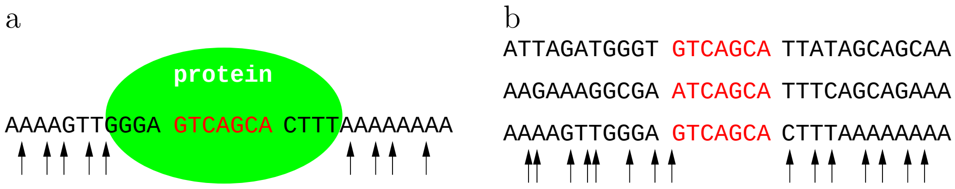

a
protein
AAAAGTTGGGA GTCAGCA CTTTAAAAAAAAA
b
ATTAGATGGGT GTCAGCA TTATAGCAGCAA
AAGAAAGGCGA ATCAGCA TTTCAGCAGAAA
AAAAGTTGGGA GTCAGCA CTTTAAAAAAAAA

Figure 1.2: Footprinting is a method to determine the region of DNA sequence covered by a bound protein. (a) One experimental approach is called interference footprinting. In this case, DNA is modified and the effect on recognition by the protein is studied. The nucleotides which must not be modified to retain binding of the protein represent a footprint. (b) Analogously evolution is modifying DNA and selection causes conservation of the binding site, whereas the surrounding non-coding DNA evolves faster. Detecting phylogenetic footprints by comparison of orthologous sequences from different species is a bioinformatics approach to define protein binding sites.

closely related species [73]. Explanations would be functional co-evolution and selection for compensatory neutral mutations rather than individual binding site specificity [50].

CHAPTER 2

The Tracker Method

2.1 The Three Wishes

The comparison of various footprinting methods described within section 2.3.3 revealed a number of shortcomings (Table 2.14) and brought up the idea of writing our own method for phylogenetic footprinting. This put us in the position to satisfy our own wishes. We want a fast method (first wish) that is able to cope with a large set of long sequences (second wish). It should be no big effort to write it (third wish) and it should automatically write a paper given the input data (forth wish). It is worth mentioning that just three of them came true.

In order to follow our own specifications we combine the parts of the programs with good performance. Therefore, we decide to take the fast pairwise local alignment tool `blastz` [64] to preselect regions which might be conserved. This alignment algorithm is used by `PipMaker`. It can cope with sequences in the order of nearly 100Mb and produces an output that describes the alignments by coordinates within the sequences. The benefit is, that the output is easy to parse. To build up multiple alignments the pairwise alignments spanning overlapping sequence fragments of the same organism are clustered. The information of sequence sharing by a certain subset of sequences can be extracted from these clusters. To regain the information about the location of gaps the sequences in a cluster are now realigned with a standard multiple alignment tool. We included `DIALIGN` due to the good results in our test and those done by *Blanchette et al.* [14]. An extra benefit was incurred by implementing our own method. We could design an output that is human readable as well as applicable to further quantitative and statistical analysis. We decided that a list of footprints should give their position in the sequences. From this list one can extract the distribution of footprints among the input sequences. It turned out that this fits our purpose very well.

2.2 The Intestines of Tracker

2.2.1 Initial Set of Pairwise Alignments

The program `tracker` is based upon the `blast` [2] implementation `blastz` [65], which is used to produce an initial list of local pairwise alignments from comparisons of all pairs of the N input sequences. This list is then assembled into clusters of partially overlapping regions that are subsequently analyzed in detail. By default, only the intergenic regions between two homologous genes are compared. Additional (non-homologous) genes contained in one or both sequences are disregarded. For instance the IGR between *Hox-A9b* and *Hox-A2b* together with the region between *Hox-A2b* and *SNex* of Takifugu is compared with the region between *Hox-A9a* and *SNex* of the zebrafish with the exception of the exons and introns of the zebrafish *Hox-A5a*, *Hox-A4a*, *Hox-A3a*, and *Hox-A1a* genes and the Takifugu *Hox-A2b* gene (Fig. 2.2). In the current study we exclude introns; they could be included easily by simply treating them analogous to IGRs, i.e., by listing individual exons instead of entire genes in the input. In the current implementation a table listing which genes (or exons) are homologous has to be provided by the user. A tool such as `lagan` [17] could easily be integrated to construct this table automatically from the input sequences.

Formally, the combined results of all `blastz` comparisons of the N input sequences $x^1, x^2, \ldots, x^N$ form a set $\mathfrak{A} = \{A_k | k = 1, \ldots, M\}$ of alignments which is the basis of all further analysis steps. Each alignment A_k is represented as pair of intervals $\{A_k^1, A_k^2\}$. More explicitly, we write $A_k = \{A_k^1, A_k^2\} = \{x^p[i..j], x^q[i..j]\}$. For instance, $x^p[i..j]$ is the substring between positions i and j of the input sequence x^p that forms first sequence in the alignment A_k.

The `blastz` searches are performed with non-stringent parameters in an attempt to avoid false negative at this early stage. As an undesirable side-effect of reducing the stringency of `blastz` we observe that some repetitive sequence elements slip into the initial set of alignments. We use the rather straightforward local entropy criterion described below to identify such sequences and to remove the corresponding *parts* of pairwise alignments from our initial list. In some cases low complexity repetitive sequences actually connect two significantly conserved sequences. In this case we fragment the alignment into two or more shorter ones.

We prefer to use a local entropy measure rather than a tool such as `Repeat-`

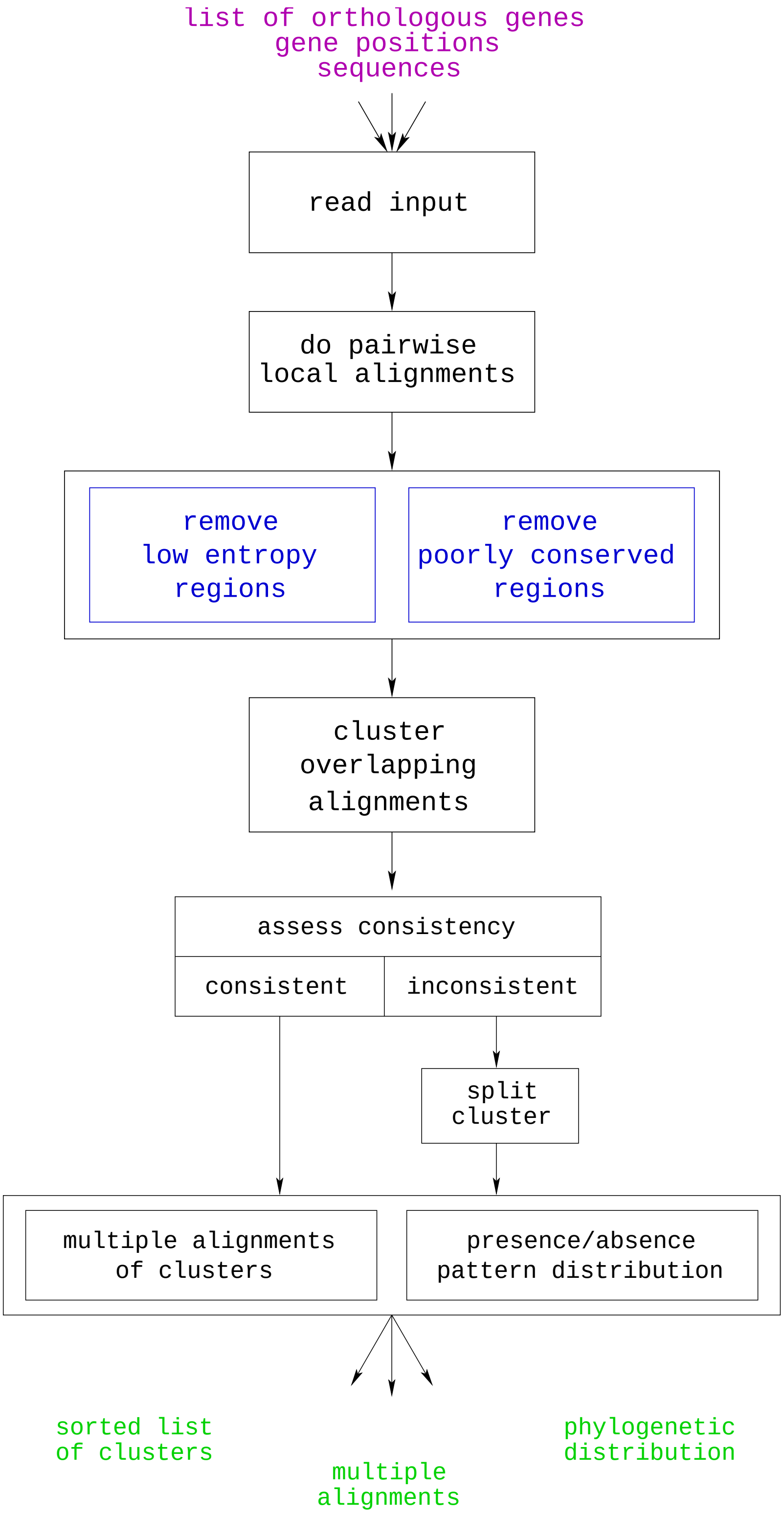

list of orthologous genes
gene positions
sequences
read input
do pairwise local alignments
remove low entropy regions
remove poorly conserved regions
cluster overlapping alignments
assess consistency
consistent
inconsistent
split cluster
multiple alignments of clusters
presence/absence pattern distribution
sorted list of clusters
multiple alignments
phylogenetic distribution

Figure 2.1: Functional parts of the `tracker` algorithm. For a detailed description see section 2.2.

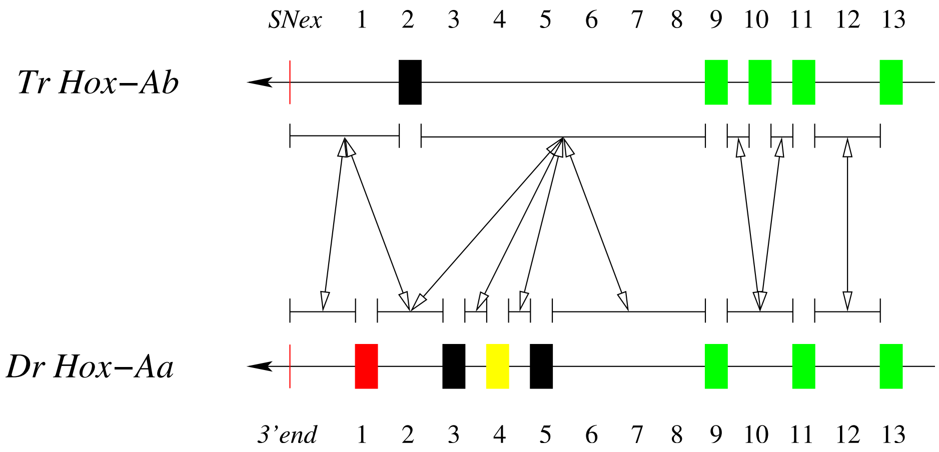


Figure 2.2: Comparison of orthologous intergenic regions. Pairs of orthologous genes are used as boundaries for the description of orthologous non-coding sequences. For a detailed description see text.

`Masker` [69] which uses a database of repetitive elements. The reason is that we only want to remove repetitive low complexity sequences, since more complex repetitive elements that are conserved between very distant species may well be functional. Local entropy measures are computed from the nucleotide frequencies $f_a(i)$ in a sequence window $[i - W/2, i + W/2]$ of width W around position i. In addition, we use analogously defined joint frequencies $f^{\tau}_{ab}(i)$ of finding the nucleotides a and b separated by a distance τ along the chain. The corresponding local entropies are

$$H(i) = -\sum_{a} f_a(i) \log_2 f_a(i) \qquad H_\tau(i) = -\sum_{a,b} f^{\tau}_{ab}(i) \log_2 f^{\tau}_{ab}(i) \tag{2.1}$$

Clearly, $H(i) \leq 2$bit and $H_\tau(i) \leq 4$bit. We designate a position i as having "low complexity" if both $H(i)$ and the average mutual information measure

$$M(i) = \frac{1}{\tau_{\max}} \sum_{\tau=1}^{\tau_{max}} H_\tau(i) - H(i) \tag{2.2}$$

are smaller than user-defined threshold values $H_{\min}$ and $M_{\min}$, respectively. The default values $H_{\min} = 1.25$ and $M_{\min} = 0.75$ have been determined by inspecting a large sample of test cases. The procedure is insensitive to small changes of these parameters.

The second problem with the initial `blastz` alignments is that in many cases they consist of a few highly conserved blocks separated by relatively long (several dozens of nucleotides) stretches of completely diverged sequences. For our purposes it is desirable to separate such hits by removing the non-conserved parts of the sequence. To this end, we re-align the `blastz` hits using a conventional dynamic programming alignment algorithm such as `clustalw` [74] and post-process these alignments in the following way: We define a partial alignment as sufficiently conserved if (i) it contains a window $[i, i+L-1]$ of length L in which the sequence identity is at least $\mu_{\min}$ and (ii) it does not contain a window of the same length L with an identity of less than $\nu_{\max}$. In other words, the `blastz` hit is divided at any sequence window of length at least L with very low conservation. Of the resulting fragments only those that contain a sufficiently conserved block of length at least L are retained for further evaluation. The values of L, $\mu_{\min}$, and $\nu_{\max}$ may have to be adjusted from their default values for sequences from very closely related species.

2.2.2 Consistent Cliques

We say that *two alignments A_k and A_l overlap* if at least one of the four intersections $A^1_k \cap A^1_l$, $A^1_k \cap A^2_l$, $A^2_k \cap A^1_l$, and $A^2_k \cap A^2_l$ is non-empty. For the construction of footprint clusters it can be useful to combine alignments that are separated only by a short intervening sequence into a single one. We thus treat $A^u_k = x^p[i..i']$ and $A^v_l = x^p[j..j']$ with $i' \leq j$ as if they were overlapping when $j - i' \leq D_{\max}$. The default is $D_{\max} = 0$, however, so that only true overlaps are considered. We can view the combined results from the `blastz` scans as a graph Γ that has the individual `blastz`-alignments as its vertices. The edges of Γ are then the overlapping alignments.

Overlapping alignments may either indicate that (parts of) footprints are conserved between more than two sequences or they arise e.g. by the duplication of a footprint pattern in one or both of the input sequences. In the first case we will attempt to construct a multiple alignment of the footprint in all sequences in which it appears. In the second case this is not possible since we have conflicting pairwise alignments between parts of the same two sequences, Fig. 2.3a. The second stage of a `tracker` run therefore consists of a careful analysis of the overlap graph and its constituent sequence alignments. We begin with a decomposition of Γ into its connected components Γ_c, $c = 1, \dots, n_C$, which we will refer to as "clusters". Since the clusters are independent of each other, they can be processed separately in further processing stages.

The complicated part of the analysis is the further investigation of the individual clusters since they may contain mutually incompatible alignments. A set $\mathfrak{U} \subset \mathfrak{A}$ of pairwise alignments is said to be *compatible* if there is a multiple alignment $\mathbf{A}$ that "contains" each pairwise alignment $A \in \mathfrak{U}$ in the following sense: If the sequence positions $x^p[i]$ and $x^q[j]$ are aligned in $\mathbf{A}$ then they are also aligned

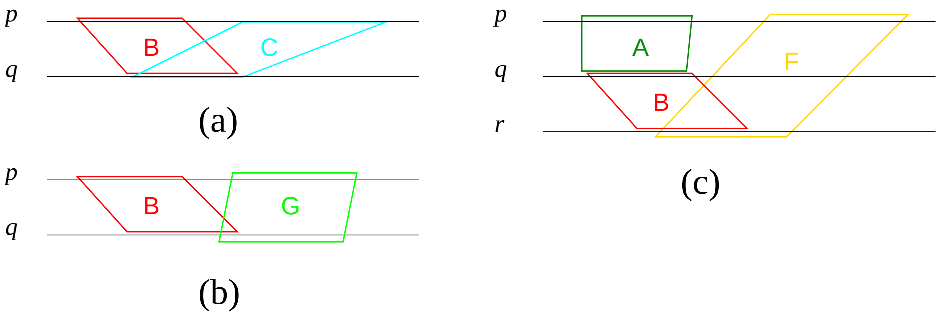


Figure 2.3: (a) Two alignments that overlap in sequence q match with disjoint subsequences of p: clearly these two alignments are inconsistent in the sense that they cannot even be approximately part of a common alignment. (b) This situation on the r.h.s. is more subtle because the small overlap of only a few nucleotides might be the artifact here. In this case we might want to treat them as a single alignment with a long insertion in sequence p. (c) In this case the alignments A and F between sequence p-q and p-r respectively are inconsistent because different subsequences of p are mapped to the same subsequence of r by means of the alignment B. Note that iff we were to disregard alignment B then the alignments A and F belong to different connected components.

in A provided A is an alignment of subsequences of x^p and x^q that contains the positions i and j, respectively. We will use here a somewhat weaker notion that allows us to avoid the explicit construction of alignments at this stage. We say that $\mathfrak{U}$ is *consistent* if $A = \{x^p[i..i'], x^q[j..j']\}$ is contained in $\mathbf{A}$ in the (weaker) sense that the sequence intervals $x^p[i..i']$ and $x^q[j..j']$ are aligned in $\mathbf{A}$, but not necessarily in the exact same way. The simplest case of incompatibility involves only one pair of alignments $A = \{x[i..i'], y[j, j']\}$ and $B = \{x[k..k'], y[l, l']\}$ between the same two input sequences x and y that overlap in one sequence but not in the other one, as in the example shown in Fig. 2.3a,b. More complicated inconsistencies, such as the situations in Fig. 2.3c, appear to be very rare in practical applications with few sequences but play an important role for larger samples. Our task is therefore to determine maximal sets of mutually consistent alignments within a cluster. Such sets of pairwise alignments can be combined to a multiple alignment which we call a *clique* of footprints.

The basic idea is to consider a series $(A_1, A_2, \ldots, A_m)$ of distinct alignments such that $A_j^2 \cap A_{j+1}^1 \neq \emptyset$. Note that any such sequence corresponds to a path in the overlap graph Γ_c. Then we consider the "image" of the initial sequence interval A_1^1 at each step of the series, i.e., the part $\hat{A}_k^2$ of the sequence A_k^2 that is aligned with (a part of) A_1^2 through the concatenation of the alignments A_j, $1 \leq j \leq k$. We call u the *trace* of the initial sequence. Whenever $\hat{A}_k^2$ and A_1^1

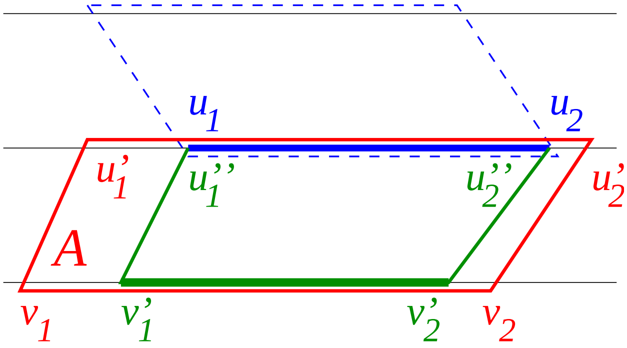


Figure 2.4: Notation for the inconsistency-finding algorithm. $[v'_1, v'_2]$ is trace of $[u_1, u_2]$ under the alignment A. See text for details.

are parts of the same input sequence x^p we have to check whether $\hat{A}^2_k \subseteq A^1_1$. An inconsistency occurs if $\hat{A}^2_k \not\subseteq A^1_1$, i.e., if the image of A^1_1 after a series of alignments is another interval on the same input sequence. Fig. 2.3c is the simplest example for this situation.

In the following paragraphs we outline the algorithm for detecting inconsistencies in more detail. It is convenient to drop the explicit reference to the sequence from the notation and to write $A = [p_1, p_2], [q_1, q_2]$ instead of $A = \{x^p[i_1..i_2], x^q[j_1, j_2]\}$. In order to find all alignments in the cluster that are inconsistent with the initial alignment $A_0 = [p_1, p_2], [q_1, q_2]$ we construct a directed tree recursively starting with the directed edge $[p_1, p_2] \rightarrow [q_1, q_2]$ by means of the following rule: To each endpoint u of the growing tree[1] which is associated with an interval $[u_1, u_2]$, we attach edges for each alignment A that overlaps with $[u_1, u_2]$ and has not been used already along the path from $[p_1, p_2]$ to $[u_1, u_2]$. The vertex at the endpoint of the new edge is associated with the trace $[v'_1, v'_2]$ of $[u_1, u_2]$ under the alignment A that is defined as the part of $[v_1, v_2]$ aligned with the overlap $[u''_1, u''_2] = [u_1, u_2] \cap [u'_1, u'_2]$, see Fig. 2.4. The traces can be interpreted as sequence pieces that *should* be aligned with $[p_1, p_2]$ according to the given series of alignments.

The preprocessed alignments do not contain large gaps in our case. We can therefore estimate the traces just from the intervals by assuming that alignments act like linear transformations on the intervals. Simply determine α_j such that

[1] with the exception of the root $[p_1, p_2]$, of course

$u''_j = u'_1 + \alpha_j(u'_2 - u'_1)$ for $j = 1, 2$, i.e., $\alpha_j = (u''_j - u'_1)/(u'_2 - u'_1)$; then

$$v'_j = v_1 + (u''_j - u'_1)\frac{v_2 - v_1}{u'_2 - u'_1}\,. \tag{2.3}$$

In this way we avoid the explicit construction of the alignments. The correction factor $(v_2 - v_1)/(u'_2 - u'_1)$ is close to 1 since gaps are rare. The inaccuracies incurred by this approximation may lead to slight displacements of the aligned intervals. This can be compensated in the computation by allowing a small tolerance t such that we accept the interval $[a, b]\dot{\subseteq}[c, d]$ iff $a \geq c-t$ and $b \leq d+t$.

After each extension of our search tree three situation may occur:
(i) We arrive at a trace $[p^*_1, p^*_2]$ such that there is an previously constructed trace $[p'_1, p'_2]$ satisfying $[p^*_1, p^*_2] \subseteq [p'_1, p'_2]$. Then we abandon the branch at $[p^*_1, p^*_2]$ since any inconsistency with $[p^*_1, p^*_2]$ is also an inconsistency with the larger trace $[p'_1, p'_2]$.
(ii) We encounter an alignment A_k with a trace $[p^*_1, p^*_2]$ at its terminal vertex that is part of the same sequence p as the "root interval" $[p_1, p_2]$. If $[p^*_1, p^*_2] \not\subseteq [p_1, p_2]$ then at least one sequence interval $[u_1, u_2]$ encountered (as trace) somewhere along the path from $[p^*_1, p^*_2]$ to $[p_1, p_2]$ would be aligned with two distinct intervals on the same sequence p. Consequently, the initial alignment A_0 and the alignment A_k are inconsistent. We store this fact and do not further extend the search tree from $[p^*_1, p^*_2]$.
(iii) Otherwise, the tree is extended along all alignments that overlap with $[p^*_1, p^*_2]$.

We remark that, more abstractly, this procedure can be understood as a depth first search on the path-graph of the overlap graph of the alignments. (The path-graph $P(\Gamma)$ of a graph has as its vertices all paths in Γ. Two paths are adjacent in $P(\Gamma)$ if one is obtained as an extension by a single edge of the other one.) The individual alignments are represented by the paths of length 0 and serve as roots of the search trees. Along each edge of the search tree (i.e., an alignment) we compute the trace (which can be regarded as a vertex label) and check for consistency with the label of the root vertex.

For each alignment we therefore obtain a (possibly) empty list of inconsistent alignments. Repeating this search procedure with each alignment as "root" we obtain all pairwise inconsistencies. These define the graph Ψ_c that has the `blastz`-alignments of the cluster Γ_c as its vertices and has an edge between A and B if and only if A and B are inconsistent. From Ψ_i we obtain the maximal sets of consistent alignments as the cliques of the complement graph $\overline{\Psi_i}$ (which has an edge between A and B if and only if there is no edge in Ψ_c). The graphs $\overline{\Psi_c}$ have sometimes dozens or even a few hundred nodes (individual pairwise alignments). In general, $\overline{\Psi_i}$ is close to a complete graph, i.e., "most" pairwise alignments are mutually consistent. The list $\mathcal{C}_c = \{C^c_h\}$ of the cliques of $\overline{\Psi_c}$ can therefore be produced efficiently by means of the Bron-Kerbosch algorithm [16]

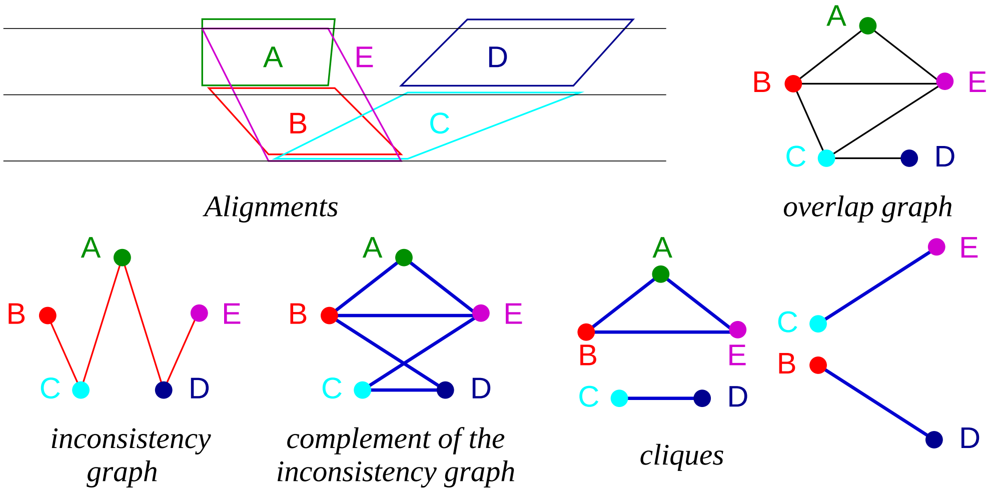


Figure 2.5: Decomposition of a cluster of alignments: First the overlap graph Γ is computed for a set of alignments. Here we show only a single connected component ("cluster"). The inconsistency graph Ψ summarized pairs of alignments that cannot be derived from a common multiple alignment. Next cliques of its complement $\overline{\Psi}$ are determined. Here we obtain four cliques $C_1 = \{\mathsf{A}, \mathsf{B}, \mathsf{E}\}$, $C_2 = \{\mathsf{C}, \mathsf{D}\}$, $C_3 = \{\mathsf{C}, \mathsf{E}\}$, and $C_4 = \{\mathsf{B}, \mathsf{D}\}$. Only $\Gamma[C_1]$, $\Gamma[C_2]$ and $\Gamma[C_3]$ are connected, hence we obtain the revised list of cliques C_1, C_2, C_3, $\{\mathsf{B}\}$, $\{\mathsf{D}\}$. Neither of the two isolated points is maximal, i.e., each of them is contained in at least one strictly larger clique, thus the final result of the decomposition are the three non-trivial cliques C_1, C_2, and C_3.

even though in general even finding the maximal clique of a graph is NP-hard [36].

The induced subgraphs $\Gamma_c[C_h^c]$ are not necessarily connected, however, i.e., they might consist of alignments that do not overlap, Fig. 2.5. We thus revise the list of cliques by replacing $\Gamma_c[C_h^c]$ by all its connected components. It is possible that such a component C' is a strict subset of a larger one. In this case C' is removed from the list of cliques.

2.2.3 Postprocessing

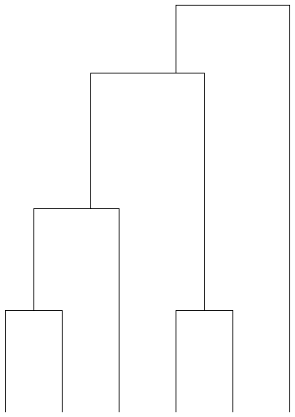
Phylogenetic footprints typically appear in clusters. For the purpose of the analysis in this contribution we pragmatically define a *phylogenetic footprint clique* as a single consistent clique. In some case one might want to argue that two or several cliques in close proximity should only be counted as a single footprint clique. For example, in [21] footprints are merged into the same "phylogenetic footprint cluster" (PFC) if they are separated by less than 100nt. This bound on the separation appears to be rather arbitrary. Furthermore, we are interested in relative abundances here, so that it makes little difference whether PFCs or footprint cliques from the tracker program are used.

The next step is rather straightforward. For each clique X and each sequence x we determine the minimal interval $[x', x'']$ that contains all intervals of x appearing in alignments belonging to X. A multiple alignment of these sequence intervals is then produced using a standard program such as clustalw [74] or DIALIGN [56]. So far our data indicate that the final outcome is essentially independent of the multiple alignment algorithm, which at this level serves mostly as a convenient method for visualization.

The final processing stage consists of relating the presence/absence pattern of the detected footprints with the established (or assumed) phylogeny of the species in question. Given a phylogenetic tree (in phylip format) as input, tracker automatically compiles an overview table in which cliques are arranged according to common presence/absence patterns together with the parsimony score for the corresponding tree (see Fig. 2.6). In addition, overview charts are produced that summarize the locations of the footprints with a common distribution on the phylogenetic tree.

2.2.4 Implementation

The tracker method is implemented as a perl program utilizing ANSI C modules e.g. for determining the inconsistency graph. Furthermore, blastz [65], clustalw [74], and DIALIGN [56] are used as system calls. The output is pro-



Fa4-3	Da4-3	Ps4-3	Rn5-3	Hs4-3	Hf4-3	Parsimony score	n	Cliques
+	+	+	+	+	+	0	7	13 14 18 19 22 23 24
+	−	+	+	+	+	1	1	20
−	−	+	+	+	+	1	3	5 10 11
−	−	−	+	+	+	1	7	2 3 4 6 7 8 12
−	−	−	+	−	+	2	1	1
−	−	+	−	−	+	2	4	9 15 16 21
+	−	−	−	−	+	2	1	17

Figure 2.6: Footprint distribution summarized by `tracker`. Given a phylogenetic tree, `tracker` assigns the cliques to the possible distributions on the tree, calculates the minimum number of mutations necessary to achieve a certain presence/absence pattern (parsimony score) and counts the number of cliques (n) matching that distribution. The concrete example shown here, is taken from a comparison of the *hoxA4-hoxA3* intergenic region of 6 sequences: *Heterodontus francisci* (Hf), *Homo sapiens* (Hs), *Rattus norvegicus* (Rn), *Polypterus senegalus* (Ps), *Danio rerio* cluster Aa (Da) and *Takifugu rubripes* cluster Aa (Fa). RunID(`tracker`) = 06292219BCSU

vided as a LaTeXdocument with included Postscript figures (such as in Fig. 2.13) and it is on the way to look like an automatically generated paper.

The `tracker` program allows the user to adjust a number of parameters, compiled in Tab. 2.1. We found that the results are relatively insensitive to the parameter settings. For closely related sequences however, one should use more stringent values for the minimal quality of the conserved sequence blocks. Interleukins, for instance, are only available for closely related species (man, mouse, and rat). In this case we used a threshold of $\mu_{\min} = 95\%$. Results for these data sets are reported in the diploma thesis of Claudia Fried [32].

Table 2.1: Default parameters for `tracker`.

Processing step	Parameter		Value
`blastz` search	Minimal Score	K	1500
Low Complexity Detection	Window Size	W	20
	Separation	$\tau_{\max}$	6
	Minimal Entropy	$H_{\min}$	1.25
	Minimal Avg. Surprisal	$M_{\min}$	0.75
Minimum Identity	Window Size	L	12
	Quality of Best Block	$\mu_{\min}$	75%
	Low Quality Cutoff	$\nu_{\max}$	35%
Cluster Construction	Maximal Distance	$D_{\max}$	0
Clique Decomposition	Tolerance	t	3

2.3 Program Performance

2.3.1 Detection of Regulatory Elements

A variety of programs exist that is used to define potential regulatory elements — mostly protein binding sites — using quite different assumptions to solve this biologically hard task (*Meeting on Systems Biology: Genomic Approaches to Transcriptional Regulation, Cold Spring Harbor Laboratory, March 6 - 9, 2003*).

A popular way to find regulatory elements is to search for known motifs in the non-coding sequences of interest. Its popularity among biologist simply arises with the obvious, low sophisticated algorithm used within this approach. The existence of large binding site databases such as TRANSFAC further increased the popularity of this method since the database specifies not only some patterns to look for but all known patterns. Such pattern detection methods are limited by a signal-to-noise problem for many eucaryotic genomes, as relatively weak sequence patterns are dispersed across large regions of potential function. Various methods have been developed to increase specificity. Some of them are based on searches for homotypic or heterotypic clustered cis-regulatory elements [75]. Blanchette *et al.* [13] have developed a method to septerate real motifs from their artifacts. They report a real motif if it's overrepresentation can explain the high counts for similar motifs. An even higher hit rate and greater precision depends on the involvement of additional constraints in enhancer organization, such as distances. Structure of similar regulated or related genes in the same organism can be used as additional constraints [38, 61].

Alternatively, orthologous non-coding sequences from a group of related species are aligned. However, the genomes to use in these comparison must be carefully selected if useful results are to be obtained. Comparison of too closely related species identifies non-functional conservation, whereas too distantly related sequences lack sufficient conservation for a meaningful comparison.

Most searches for phylogenetic footprints in the past were based on computing global alignments. Standard motif search techniques and segment-based alignment algorithm such as `DIALIGN` [56] have been shown to be more efficient [12]. The identification of unusually well-conserved sequences that hint at a regulatory function has shown to be a successful approach see [47, 52, 72, 21, 14] and the review, see [26].

In a related approach, the `rVISTA` tool uses pairwise alignments of orthologous regions to determine the significance of putative transcription factor binding sites found by comparison with a database of binding motifs [48]. Most recently footprinting was expressed as a *substring parsimony problem* and an exact and rather efficient dynamic programming algorithm was proposed and implemented [12]. This method takes the known phylogeny of the involved species explicitly into account and retrieves all common substrings with a better-than-threshold parsimony score from a set of input sequences.

2.3.2 The Test Set

In our attempt to test the performance of different programs to define potential protein binding sites, we looked at the orthologous region from *hoxA4* to *hoxA3* in a variety of vertebrate species (Table 2.2).

The region from *hoxA4* to *hoxA3* is especially useful to test the performance of different programs since four experimentally described protein binding sites are situated in this intergenic region of mammals and shark [53]. Sequences spanning the whole cluster are usually not accepted by the available programs.

Table 2.2: Source and length of the *hoxA4-hoxA3* non-coding region of *Heterodontus francisci* (Hf), *Homo sapiens* (Hs), *Rattus norvegicus* (Rn), *Polypterus senegalus* (Ps), *Danio rerio* cluster Aa (Da) and *Takifugu rubripes* cluster Aa (Fa). The lower panel gives the boundary positions for short orthologous fragments in the intergeneic region. rc = reverse complement.

organism	length	source
Hf	17407	AF479755 (as in *Chiu et al. 2002* [21])
Hs	18584	AC004080rc+AC010990rc(overlaps 200nt with flanking fragments)+AC004079[75001-end] (as in *Chiu et al. 2002* [21])
Rn	29196	NW_043751[910030-1194462]rc
Ps	11607	AC126321rc+AC132195 (overlapping 4307nt)
Da	8905	AC107365rc
Fa	9410	Fugu v.3.0 scaffold_47[103001-223000]rc

organism	length	subregion
Hf_{reg}	1750	12500 - 14250
Hs_{reg}	2000	13750 - 15750
Rn_{reg}	2100	24400 - 26500
Ps_{reg}	1500	8500 - 10000
Da_{reg}	1300	6100 - 7400
Fa_{reg}	1000	6000 - 7000

Shorter fragments (Table 2.2) can also cover the cluster of the four known binding sites and are used in cases when the intergenic regions from *hoxA4* to *hoxA3* are still too long to pass the qualifying conditions of the program. The four experimentally described sites are listed in Table 2.3.

2.3.3 Available Methods - how suitable are they?

String Search

Using the string search function of the editor emacs it is easy to map the known sites (Table 2.3) onto the sequences of shark, human and its close relative, the rat. The lack of matches to the teleost sequences may be explained by the loss of these motifs and their function or by variations in the binding site motifs similar to the differences between the human and shark HOX/PBC siteA and Prep/Meis site patterns while all four functions are retained.

This problem is due to the shortness of footprints with 6 to 7nt in length, as the KrA and the Prep/Meis site. Therefore, it is not very unlikely to find them by chance, which means that they need not to be homologous and carry out any function. To estimate the likelihood of a binding site occurring by chance, we calculate the average stochastic occurrence of a binding site bs:

$$P(bs) = ({f_{(A)}}^{n_A} {f_{(C)}}^{n_C} {f_{(G)}}^{n_G} {f_{(T)}}^{n_T})(M - l) \qquad (2.4)$$

while M is the length of the sequence, $f_{(N)}$ is the nucleotide frequency in the sequence of length M, l is the length of the conserved binding site and n_N is the total number of Ns in the binding site. For our test data set the solution of the above equation differs by a factor of 1 to 3 for the short KrA and Prep/Meis sites (Table 2.4). Thus, the occurrence of these sites could still be explained by stochastic processes.

To benefit from this method one requires the precise pattern of the site from the organisms used. Otherwise the method yields a lot of false negative due to binding site variation (turnover). This problem could be solved by summing up the individual variations of a site. A pattern is than represented as a consensus sequence or a position specific score matrix giving a statistical description of regulatory signals (as used for TFsearch).

Assuming that a footprint is more than a nucleotide pattern also a bunch of false positives is detected due to stochastic occurrence of the typically short strings. Improvement could be achieved in two ways. First, by taking into account some of the properties which may turn a nucleotide pattern into an active regulatory site. We may think of protein binding sites in the vicinity, absolute or relative distances to these sites or the regulated genes, and epigenetic effects such as chromatin structure. Second, by looking at over-representation of the sites compared to their stochastic occurrence since patterns of active sites are usually clustered and the sites and there variations are over-represented in the surrounding sequence [68].

TFsearch

TFsearch [1] makes use of the `TRANSFAC` databases [37] as source of binding sites with known function. The detection of sites is done by a simple correlation calculation with the position specific score matrix provided by `TRANSFAC` for every known site. These matrices are binding site profiles that sum up the individual variations of a site and serves for overcoming the problem with false negatives due to slight biding site variation. This increases the problem with false positives causing a huge number of predictions that are randomly and uniformly distributed (Table 2.5 and 2.6).

The data in Table 2.5 and 2.6 show that none of the experimentally known binding site is detected by TFsearch. Most likely, this is due to the lack of entries of hox regulatory elements in the last public release of the database. Two patterns for CdxA are yielding 100% identity matches, these are 5'-CATAAATCT-3' and 5'-ATTTATG-3'. The first pattern is part of the HOX/PBC siteB. With the default setting of an identity score >85.0 the Hf$_{reg}$ sequence gives 51 high scoring hits for CdxA, thereby matching the CdxA profile to any AT-rich region.

Advanced search tool relying on TRANSFAC are provided together with the licensed version of the database. The algorithm of Match uses two values to score putative hits. First, the matrix similarity score which is a weight for the quality of a match between the sequence under study and the whole matrix. Second, the core similarity weights the quality of a match between the sequence and the five most conserved consecutive positions in a matrix.

Further improvement should also consern the alignment score. It might be better to use a normalized sequence alignment score for composition bias like the z-score instade of the identity score. This would lead to lower scores for weak patterns and reduce there recurring detection. Furthermore, the position specific score matrix makes the strong assumption that each position within a binding site is independent. Correlations among positions exist in many examples of experimentally characterized binding sites as shown by *Kaplan et al.* [9]. Their study revaled that modelling the dependencies leads to a more accorate identification of the exact binding locations in the sequence.

The comparison with a large database of binding sites could also be improved by combination with comparative sequence analysis, as done by rVISTA. This procedure reduces the number of predicted transcription factor binding sites by several orders of magnitude. It simultaneously searches the major TRANSFAC matrices selected by the user and utilizes global sequence alignment to sieve through the data.

PipMaker

PipMaker computes pairwise local alignments using blastz [64]. Because of the alignment algorithm used, PipMaker is rather fast. The resulting local alignments are then summarized in an percent identity plot (Fig. 2.7). This plot correlates the sequence position in the first sequence with the percent identity of a local alignment spanning this position. Subsequent problems are (i) the strong dependency on the first sequence (in multiple sequence runs) showing different numbers of aligned regions for different runs with varying first sequences (Fig. 2.8)) and (ii) the loss of information where this alignment is located in the sequences other than the first one. Some of this information can be regained from the global alignment by tedious work.

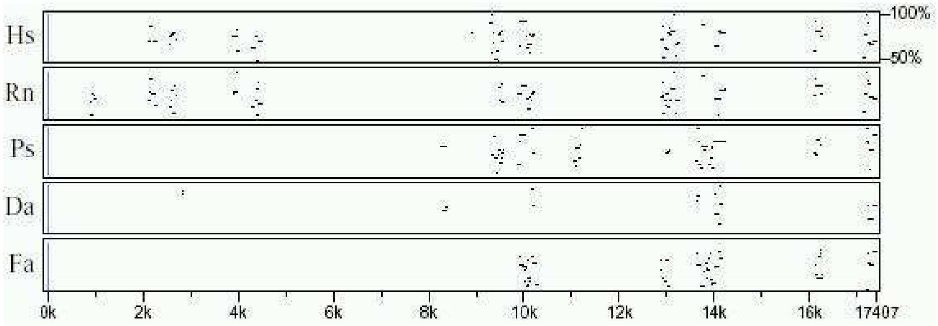


Figure 2.7: Percent identity plot of *hoxa4-hoxa3* from Heterodontus francisci (first sequence) compared to the orthologous region in Hs, Rn, Ps, Da and Fa. This plot correlates the sequence position in H. francisci (x-axis) with the percent identity (y-axis) of a local alignment (small horizontal lines) spanning this position. Information about the location of the alignments in the orthologous sequences is not provided by this representation of the data.

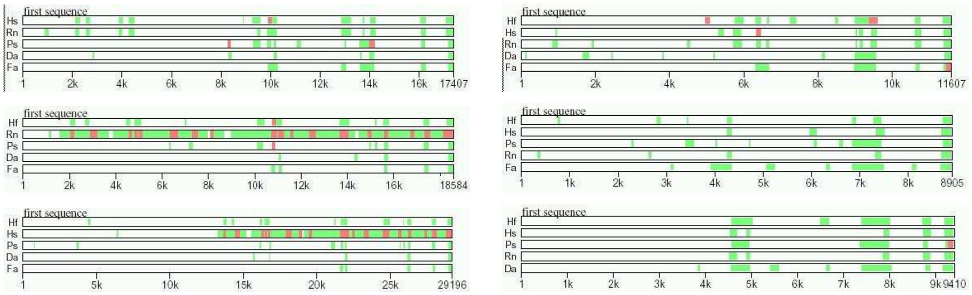


Figure 2.8: `PipMaker` overview plots. Each of the 6 sequences once used as first sequence: Hf, Hs and Rn on the left hand side, Ps, Da and Fa on the right hand side. Green bars highlight aligned regions, strongly aligned regions (at least 100 bp without a gap and more than 70% nucleotide identity) are shown in red.

FootPrinter

Utilizing a dynamic programming algorithm, `FootPrinter` is able to calculate an exact solution for the footprinting problem and promises good results on a set of multiple sequences. It was designed to find motifs in promoter regions or introns, where each sequence is of length at most a few thousands bp. Restrictions for the length of input sequences are set by the FootPrinter webserver. The downloadable version of FootPrinter2.0 does not have any constraints on the length of the input sequences but (even with short sequences, with length about 2000nt) rarely terminates without error.

For each motif the coocurrence, position, evolutionary span, significance score and number of mutations are available. A dependency of the results on the phylogentic relationship can be observed in our set of significant motifs. This is visualized in the graphical representation of a FootPrinter output. If Ps$_{reg}$ is placed neighboring to shark and tetrapods (Hf$_{reg}$, Hs$_{reg}$ and Rn$_{reg}$) 8 fooptrints within the Ps sequence are detected that would not have been found with Ps$_{reg}$ neighboring to the Teleosts (Da$_{reg}$ and Fa$_{reg}$). This is a major disadvantage in our purpose to compare *Hox* cluster sequences because of previous findings by *Chiu et al. 2002*[21] showing that substantial changes in the regulatory patterns do not necessarily conform with established phylogenetic relationships. Irrespective of the phylogenetic tree, neither of the experimentally known homologous sites were recognized by FootPrinter.

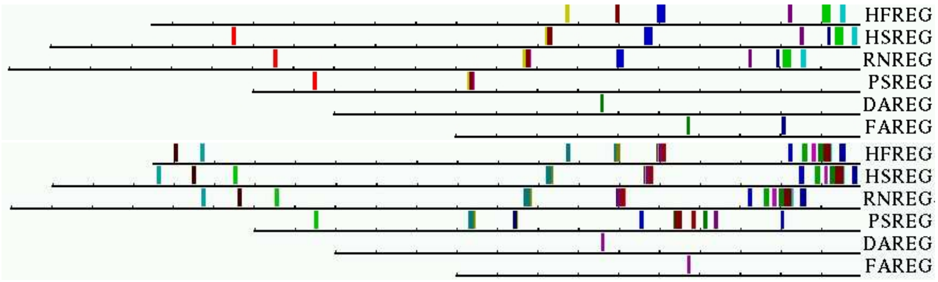


Figure 2.9: Dependency of FootPrinter results on the phylogenetic relation of the input sequences.
Upper panel: (Hfreg,((Hsreg,Rnreg),(Psreg,(Dareg,Fareg)))),
lower panel: (((Hfreg,(Hsreg,Rnreg)),Psreg),(Dareg,Fareg)).
Corresponding motifes in different sequences are highlighted with the same color. Tick marks are separated 100 nuleotides.

BayesAligner

BayesAligner is a pairwise local alignment algorithm [78]. Whereas standard algorithms rely on suitable scoring matrix and gap penalty parameters, the BayesAligner returns the best alignments weighted proportional to its probability, considering the full range of gapping and scoring matrices. This requires NMk space and time where N is the length of the query sequence, M the length of the data sequence and k the number of blocks expected to be aligned. Therefore, the Bayesian Phylogenetic Footprint Homepage accepts a total sequence length (query sequence length + data sequence length) of 4000nt maximum. 5000nt for each sequence is as large as one can go using the local Bayes Block Aligner since the number of possible alignments overflows a double precision number.

The results strongly depend on the length of the sequences (Table 2.7) and their order of input (data not shown). A footprint once found may therefore not

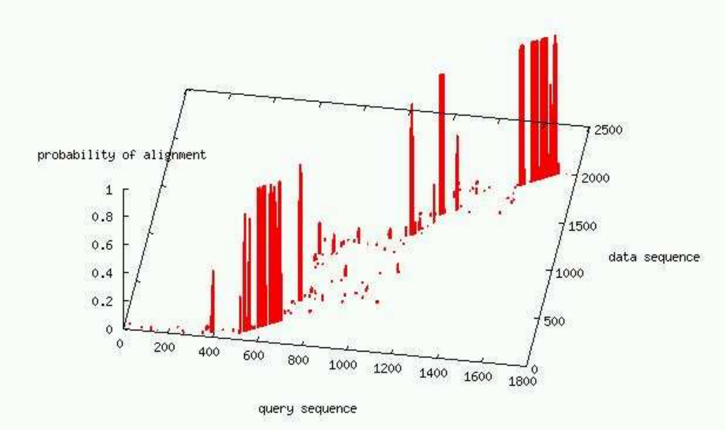


Figure 2.10: Probability distribution of the sequences Hf$_{reg}$(query sequence) and Hs$_{reg}$ (data sequence) being aligned computed with BayesAligner. The high scoring hits reported in table 2.7 appear around position 600 of the query sequence.

be retrieved when length and order of sequences or the number of aligned blocks (k) are changed. Because of the non-symmetric relation between data and query sequence one would have to sample over different possible pairwise combinations of fragments of different size to gain a meaningful result.

DIALIGN

DIALIGN is a segment based alignment algorithm able to handle a large set of long sequences. The output is a typical multiple alignment file with lower case letters that are not considered to be aligned even though they are at the same position in the alignment. Therefore, one can just 'trust' into capital letters when looking for footprints. It is obvious that this calls for post-processing when attempting to use DIALIGN for phylogenetic footprinting. Moreover, there is no easy way to check the quality or plausibility of the multiple alignment at glance. Therefore, a simple visualizing tool was implemented that extracts the conserved regions and alowes a comparison with our tracker method.

The alignment of the 6 sequences ranging from *hoxA4* to *hoxA3* took DIALIGN 18 min. The graphical overview of the DIALIGN output is compared to the overview of our own method tracker. In general, the results seem to be consistent even though DIALIGN reports more hits visualized throw the higher density of lines in Fig. 2.12. To determine if this results from a greater sensitivity and more aligned positions we calculate the fraction of aligned positions for each sequence and both methods (see Table 2.9). We observe a higher percentage of upper case letters in the single DIALIGN alignment compared to the sum of all non-overlapping DIALIGN alignments of footprints resulting from a tracker run. This shows that DIALIGN is indeed more sensitive than our method but at the expense of loss of specificity. Tracker does not report very short stretches about 4 to 6 nt due to low significance to ensure higher specificity while performing a more restrictive search. To further describe the quality of the alignment we define a measure for 'multiplarity' which is the fraction of multiple aligned positions on all aligned positions. This measure can be calculated independently for $2, 3, 4...n$ sequences part of an alignment, further called 'degree of multiplarity'. Table 2.10 and Fig. 2.11 show that the multiplarity decends exponentially with the degree of multiplarity without significant differences between DIALIGN and tracker.

Table 2.8 lists the known motifs from Table 2.3 that are aligned correctly and shows that most of the sites are found.

An additional test on the whole cluster sequences ran for 2 days on a Pentium IV with 2.4 GHz and revealed three major shortcomings. First, DIALIGN does not even correctly align all exons of the *Hox* genes. This is shown in Fig. 2.13 by the absence of lines protruding from the *hox3*, *hox2* and *hox1* of the HsA sequence. A quantitative comparison of matches in the whole cluster region and a seperat alignments of all orthologous genes underlines the fact of missmatching sequences since more aligned positions are reported for the sum of all genes that for the whole cluster. This results in negative numbers for the difference between whole cluster and gene alignements (Tables 2.11 and 2.12).Therefore, the quantified DIALIGN results cannot be directly compared to tracker. Calculations would lead to the assumption that no non-coding sequences can be found using DIALIGN. Second, one of the sequences (DrAa) is wrongly placed in the alignment and therefore not part of any aligned segment denoted by zero upper case letters in Table 2.11. Second, we expect a reduction of aligned positions with a higher degree of multiplarity (Table 2.12) in DIALIGN results due to such displacements. Third, with a runtime of 2 days DIALIGN is a very resource consuming method.

This shows that DIALIGN can only be used to align previously selected subregions in the range of 10000nt maximum (the length of a typical intergenic region).

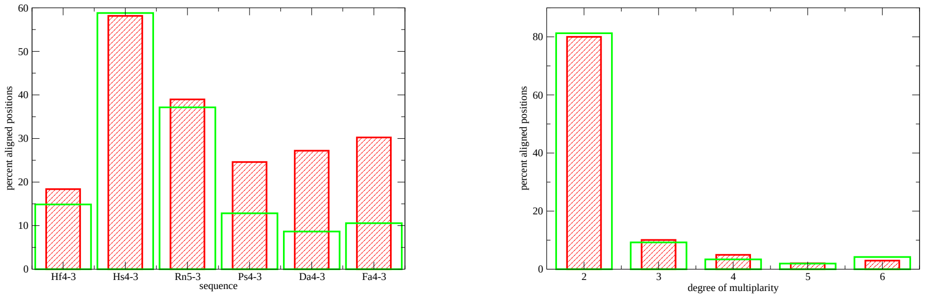


Figure 2.11: Percent of aligned positions in the *hoxA3-hoxA4*. The data resulting from `DIALIGN` are shown as red hetched bars. The values for `tracker` are overlayed as bars with green outline. Abbrevations refer to Fig. 2.12 and Fig. 2.13.

Table 2.3: Four experimentally examined footprints in the human or shark *hoxA4-hoxA3* non-coding region [53] and their positions in the sequences. Positions in bold letters are the functional sites. For Ps, Da and Fa positions of both, the shark and human patterns of a site are given. Abbreviations: Hf = *Heterodontus francisci*, Hs = *Homo sapiens*, Rn = *Rattus norvegicus*, Ps = *Polypterus senegalus*, Da = *Danio rerio* cluster Aa and Fa = *Takifugu rubripes* cluster Aa.

	binding sites	
site name	human	shark
KrA site	GTCAGCA	GTCAGCA
HOX/PBC siteA	TGATTATTGAC	TGCGCATTGAC
HOX/PBC siteB	TCATAAATCT	TCATAAATCT
Prep/Meis	TGACAA	CGACAG

	positions					
	Hf	Hs	Rn	Ps	Da	Fa
KrA site	**12907**	**13707**	**24565**	-	-	-
		12454	23322			
		14685				
HOX/PBC siteA	**12935**	**13736**	**24594**	-	-	-
HOX/PBC siteB	**13132**	**13936**	**24800**	-	-	-
Prep/Meis	**13156**	**13960**	**24824**	-	-	-
		16712	566	1393	878	4112
			6755	1446	6666	8776
			9873	3241	7828	
			14981	3590		
			16102	6938		
			22371			
			27298			

Table 2.4: Occurrence of the four binding sites (KrA, HOX/PBC site A, HOX/PBC site B, Prep/Meis (Table 2.3)) in the *hoxA4-hoxA3* non-coding region of Hf, Hs, Rn, Ps, Da and Fa. Columns containing integers show the observed occurrence of the given site, followed be the corresponding stochastic occurrence. The stochastic occurrence of the human and shark patterns in the sequences of for Ps, Da and Fa are summed up for each site. Abbreviations used are defined in Table 2.2.

		occurrence of							
organism	length	KrA		site A		site B		Prep/Meis	
Hf	17407	1	0.6218	1	0.00216	1	0.05749	1	1.9437
Hs	18584	3	1.0494	1	0.00576	1	0.02076	2	4.5447
Rn	29196	2	0.8817	1	0.00433	1	0.01713	8	4.6657
Ps	11607	0	0.2777	0	0.01005	0	0.05382	5	4.4504
Da	8905	0	0.2466	0	0.00787	0	0.03324	3	3.5021
Fa	9410	0	0.2116	0	0.00209	0	0.00633	2	2.1789

Table 2.5: TFsearch result with identity score >92.0 applied to the Hf$_{reg}$ sequence (1750nt see Table 2.2).

site	occurrence	site	occurrence
CdxA	18	CP2	1
GATA	7	STATx	1
SRY	2	v-Myb	1
C/EBP	2	Nkx-1	1
Oct-1	2	MZF1	1

Table 2.6: Number of CdxA binding sites found by TFsearch (identity score >95.0). Sequence names refer to Table 2.2; rand represents the average result over 5 random sequences (min #CdxA = 0, max #CdxA = 9).

sequence	length	#CdxA	#CdxA/length
Hf$_{reg}$	1750	10	0.571%
Hs$_{reg}$	2000	7	0.350%
Rn$_{reg}$	2100	5	0.238%
Ps$_{reg}$	1500	18	1.200%
Da$_{reg}$	1300	9	0.692%
Fa$_{reg}$	1000	2	0.200%
rand	1500	3	0.200%

Table 2.7: `BayesAligner` results for pairwise combinations with the short orthologous fragment within the *hoxA4-hoxA3* shark sequence. The left column of each column gives the distances to the upstream *hox* gene. The right lane shows the probability of the concerning alignment. All the listed matches do align with the sequence of comparison in the expected regions. A * indicates that the motif is not entirely high scoring. The Hfsm and Hssm sequences marked with $^+$ are shortened to a third of the length of Hfreg and Hsreg respectively. Comparisons between Hf$_{reg}$ and Ps$_{reg}$, as well as Fa$_{reg}$ and Ps$_{reg}$ did not yield any high scoring hits at the region of interest.

BayesAligner	positions			
	Hfreg-Hsreg		Hfsm-Hssm$^+$	
KrA site	-	0.0-0.2	12907	0.2-0.4
HOX/PBC siteA	-	0.0-0.2	-	0.0-0.2
HOX/PBC siteB	13132	**0.999**	13132	**1.000**
Prep/Meis	13156*	**0.999**	13156*	**0.995**

BayesAligner	positions			
	Hfreg-Fareg		Fareg-Hfreg	
KrA site	-	0.0-0.2	-	0.0-0.2
HOX/PBC siteA	-	0.0-0.2	-	0.0-0.2
HOX/PBC siteB	13132	**0.875**	6782	**0.874**
Prep/Meis	-	0.0-0.2	-	0.0-0.2

Table 2.8: Experimentally known motifs of the intergenic region from *hoxA4* to *hoxA3* correctly aligned by `DIALIGN`. Notice: The alignment between the Hf HOX/PBC siteA and Fa can not be validated with any other method.

DIALIGN	positions					
	Hf	Hs	Rn	Ps	Da	Fa
KrA site	-	13707	24565	-	-	-
HOX/PBC siteA	12935	-	-	-	-	6221
	-	13736	24594	-	-	-
HOX/PBC siteB	13132	13936	24800	-	-	-
Prep/Meis	13156	13960	24824	-	-	-

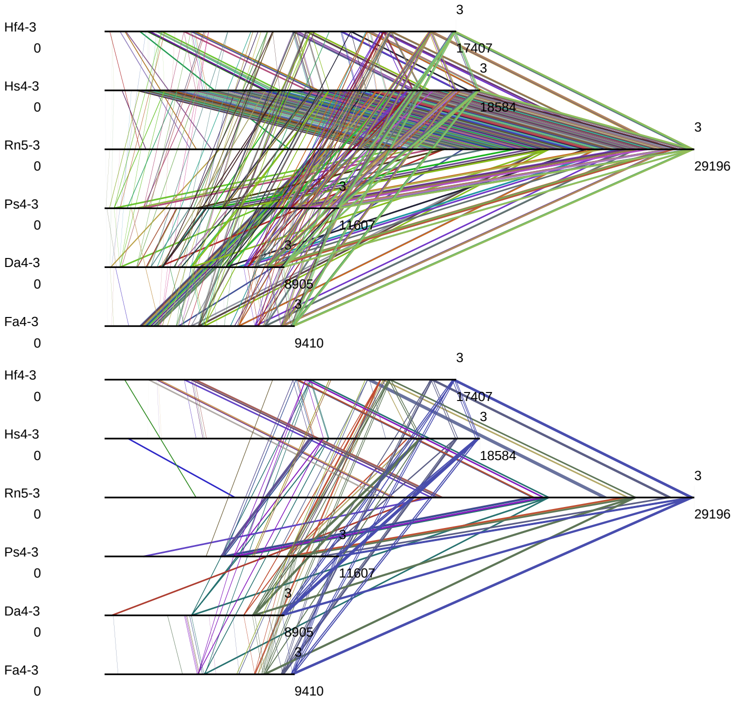


Figure 2.12: Comparison of DIALIGN (upper panel) and tracker (lower panel) output for the *hoxA4-hoxA3* intergenic region. Aligned positions are connected by a line of the same color. In the tracker run one direct match between Hs4-3 and Rn5-3 longer than 1000nt is ignored because of its overall high similarity which cannot provide any information about single binding sites or footprint clusters. Abbrevations refer to Table 2.2. RunID(tracker) = 05150002YDNE

Table 2.9: Number and percentage of aligned positions (upper case letters) in the *hoxA4-hoxA3* intergenic region observed by `DIALIGN` and `tracker`

hoxA4-hoxA3	# upper case		% upper case	
sequence	DIALIGN	tracker	DIALIGN	tracker
Hf4-3	3199	2586	18.38	14.86
Hs4-3	10804	10927	58.14	58.80
Rn5-3	11378	10846	38.97	37.15
Ps4-3	2855	1489	24.60	12.83
Da4-3	2422	770	27.20	8.65
Fa4-3	2846	994	30.24	10.56

Table 2.10: Number and percentage of multiple aligned sequence positions in the *hoxA4-hoxA3* intergenic region observed by `DIALIGN` and `tracker`. The 'degree of multiplarity' defines the number of sequences being part of the multiple aligned position.

hoxA4-hoxA3	# aligned		% aligned	
degree	DIALIGN	tracker	DIALIGN	tracker
2	11267	9401	80.00	81.25
3	1418	1068	10.07	9.23
4	696	391	4.94	3.38
5	286	224	2.03	1.94
6	417	487	2.96	4.21

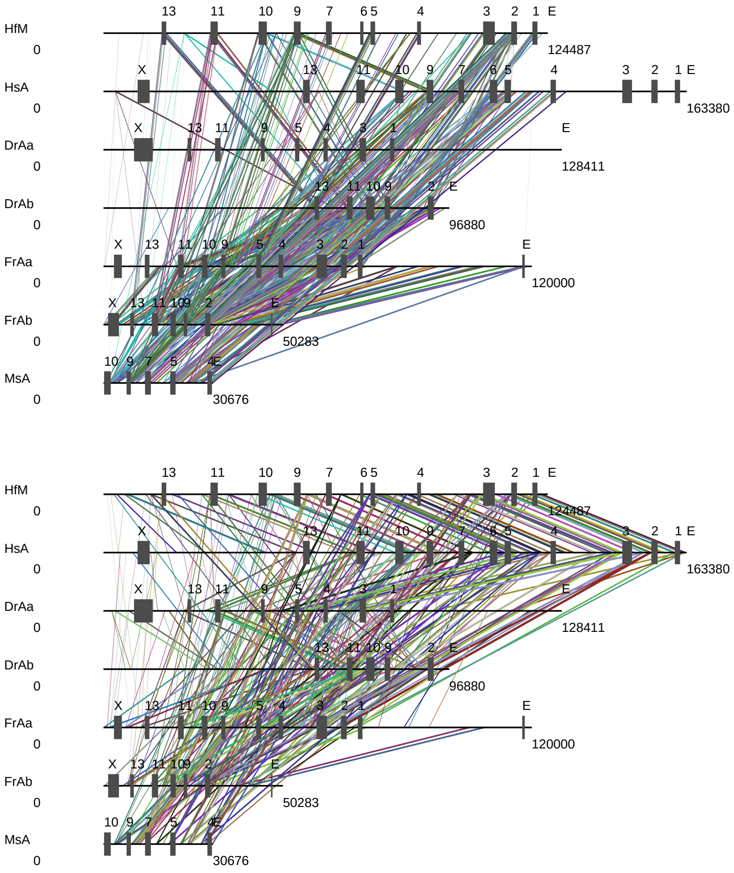


Figure 2.13: Comparison of `DIALIGN` (upper panel) and `tracker` (lower panel) output for the whole *HoxA* cluster sequence. Aligned positions are connected by a line of the same color. HfM = *Heterodontus francisci* cluster M, HsA = *Homo sapiens* cluster A, DrAa = *Danio rerio* cluster Aa, DrAb = *Danio rerio* cluster Ab, FrAa = *Fugu rubripes* cluster Aa, FrAb = *Fugu rubripes* cluster Ab, MsA = *Morone saxatilis* cluster A; RunID(`tracker`) = 06061727JGHQ

Table 2.11: Number and percentage of aligned positions (upper case letters) in the whole *hoxA* gene cluster observed by `DIALIGN` and `tracker`. Whereas `tracker` results refer to intergenic regions only the `DIALIGN` results include aligned genes. To correct the effect of aligend genes, we aligned the genes using `DIALIGN` and substracted the counts from the whole cluster alignment (column D.corr.). Negative values indicate wrong aligned exons.

cluster	# upper case			% upper case	
sequence	DIALIGN	D.corr.	tracker	DIALIGN	tracker
HfM	8888	-2896	8449	7.14	6.79
HsA	3977	-8199	7799	2.43	4.77
DrAa	0	-7909	5225	0.00	4.07
DrAb	5735	-25	3141	5.92	3.24
FrAa	9988	-1171	8146	8.32	6.79
FrAb	9301	3726	2158	18.50	4.29
MsA	4831	-537	4388	15.75	14.30

Table 2.12: Number and percentage of multiple aligned sequence positions in the whole *hoxA* cluster observed by `DIALIGN` and `tracker`. The 'degree of multiplarity' defines the number of sequences being part of the multiple aligned position. The data for `DIALIGN` are corrected by the effect of aligned genes as in Table 2.11 (column D.corr.)

cluster	# aligned			% aligned	
degree	DIALIGN	D.corr.	tracker	DIALIGN	tracker
2	13677	7521	9366	78.06	65.78
3	1411	-1246	1300	8.05	9.13
4	1579	-1390	1664	9.01	11.69
5	307	-2301	1459	1.75	10.25
6	547	-1434	428	3.12	3.00
7	0	-378	22	0.00	0.15

Tracker

`Tracker` is designed to deal with a moderately large set of (very) long sequences as whole gene clusters for example. The computations for the entire *Hox* cluster sequences of 5-8 taxa and an average length of 100kb run in well below 5 minutes on a fast PC. Since the resource usage scales approximately as $\mathcal{O}(L \times N^2)$ for N input sequences of length L it is possible to use the `tracker` tool for much larger datasets than those reported in this and the following sections. The resulting overview is shown in Fig. 2.13 and discussed in section 2.3.3. In Fig. 2.12 you can see that `tracker` is yielding less hits than `DIALIGN` for the intergenic region from *hoxA4-hoxA3*. A property of `tracker` that turnes out to be a benefit if sequences that are 10 times longer are used as input. Morover, false positive results are avoided while pointing at the most signifcant hits.

Table 2.13: Experimentally known motifs of the intergenic region from *hoxA4* to *hoxA3* detected by `tracker`. The hit concerning Fa is not supported with any other method.

tracker	positions					
	Hf	Hs	Rn	Ps	Da	Fa
KrA site	12907	-	-	-	-	6535
KrA site	-	13707	24565	-	-	-
HOX/PBC siteA	-	13936	24594	-	-	-
HOX/PBC siteB	13132	13936	24800	-	-	-
Prep/Meis	13156	13960	24824	-	-	-

In contrast to the local alignment tools presented in Table 2.14 aligning the sequences is not the major issue our method solves. It rather suveys the potential footprints bounded by local alignments and takes a close look at their distribution on one sequence and among the whole set of sequences. The table holding this information is easy to understand and easy to pars.This is imortant for biologists and downstream programs respectively. Additional outputs such as multiple alignments of all footprints can serve to construct a phylogenetic tree (section 4). TFsearch or a similar program can be applied to these multiple alignments to destruct a footprint cluster into its single footprints (section 5).

2.4 Improvements of Tracker

2.4.1 Inputfiles

At this state of development it is still hard to write the inputfiles. `Tracker` needs the positions and names of the genes given in the exonfile, and the orthology information in the form of a table (genenamefile). Later file is used to conclude from orthologous genes to the orthologous non-coding regions. This part may be automated. A tool like `lagan` [17] may be integrated to pair orthologous genes.

The fastafile and exonfile of each input sequence used in a `tracker` run must be listed in the main input file, called seqexfile, together with the genenamefile, a phylogenetic tree, the parameter settings and the run description. This results in four possibilities to spell each sequence name incorrectly and to cause `tracker` to exit with an error message. (1) The name of the fastafile, (2) the path given in the seqexfile pointing at it, (3) the leaves of the phylogenetic tree and (4) the label of the columns in the genenamefile. Analogously, this is also valid for the gene names. But there are only two possibilitys to missspell the gene names given (1) in the exonfile and (2) the genenamefile.

Even though homemade tools exist for various input processing steps, as the extraction of exon positions from genebank files (read_genbank.pl) or the task to revert and cut sequence and exonfile simultaniously (cut_seqandex.pl), we still have to facilitate the writing of input files.

2.4.2 Automatic Detection of Orthologous Regions

Orthologous regions are sequence fragments in different species that can be viewed as 'the same' sequences irrespective of accumulated mutations during the evolution from the common ancestor to there final composition. They need neither to be highly similar in sequence nor identical in function even though they usually are. Again, that's why it is difficult to destinguish orthologous sequences from simply homologous (homology - similarity in any feature due to a common origin) ones, which may arise by intragenomic duplications yielding paralogous sequences. One of the algorithms solving the task of correlating orthologous genes is `lagan` [17]. If it workes well even for long sequences including clusters of paralogous genes, an integration in our footprinting program would finally yield a tool, that could be used for screening whole sets of unanotated genomic fragments for regulatory regions.

2.4.3 Clustering Footprints

With a look at the implemented clustering method it becomes clear that the definition for 'clique' is a computational one. The biological meaning of the resulting cliques is much harder to outline. It is obvious, that a clique in the order of 15-300nt can not always be interpreted as a single footprint/protein binding site. Thinking of a clique as a cluster of footprints therefore seems to be much closer to the facts. But one has to be aware that a detected cluster of footprints is not necessarely conserved in its entirety. To underpin the interpretation of a clique as a whole footprint cluster we tried to put all cliques together into PFCs (phylogenetic footprint clusters as defind in [21]) that are separated by less than 100nt. PFCs therefore include more non-conserved DNA stretches and reduces the number of clusters and the selectivity of the detection method. This turns out to be problematic for statistical evaluations.

2.4.4 Sorting Footprints

The output written by `tracker` is an unsorted list of cliques. It seems to be a simple problem to sort footprints in their order along the genomes. But it turns out to be complicated by the fact that not all footprints are co-linear among all sequences: they may cross each other. The problem thus becomes to identify the crossing footprints, to sort the remaining co-linear cliques, and finally to insert the non-colinear ones at "reasonable" positions. This problem can be interpreted as a "Minimum Weight Vertex Feedback Set Problem" [33], which is known to be NP-complete and APX-hard, which means that it is difficult to solve. Nevertheless we think that good approximations could be obtained for our datasets by computation of the transitive closure of an acyclic graph and linear extension of the resulting partial order.

2.4.5 Significanc Measure

Significance measures for the reliability of the reported cliques may be based on the reliability of the alignment. This would involve the length, composition and complexity of the sequence and the fragment found. Such scores could be adapted from alignment programs and there significanc measure, e.g., the e-value. A similar approach could utilize `BayesAligner` and its probability score in the initial step of our program. The resulting significance score for a clique would combine the probabilities of all alignments part of the clique. The conservation of the sequences in a clique could be calculated this way, but it will not reflect the probability of containing protein binding sites. Insisting on the assumption that conserved regions are functional in terms of binding a regulatory protein we have to look at further binding site properties to varify the binding activity of a sequence fragment. To achieve a measure for the posibility of a clique being a (sum of) binding site(s), we could try to look at overrepresentation of motifs [68] or run `TRANSFAC` on the cliques. In any case, reporting a reliable significance value to our cliques would definitely make our method more attractive.

2.4.6 Counting Module

Counts of consistent cliques with sequences in special (boulean) combinations still need to be done by the 'Präzisionszählwerk' (available in every experimental lab working on genetic or microbiological issues) without computer aid. This may be suprising since we stated that the `tracker` output is machine readable as well as human readable. The concept we developed, defining a general format for phylogenetic footprints, setting up a method with XML that handles them and finally writing the counting functions was obviously deterrent. If one bears in mind that the 'Präzisioszählwerk' thinks of going on strike it becomes obvious that the implementation of a counting module is important.

2.4.7 System Compatibility and Weberver

`Tracker` is written with Perl, v5.6.1 built for i386-linux. According to numerous system calls and files that are written to disk, it will not easily run on Windows. To make our program accessible to biologists and windows users it will be necessary to write a web interface and a local version also running on non-linux computer systems.

Table 2.14: Properties of a set of programs that are used for detection of protein binding sites. TFsearch scans a sequence for known binding sites, `PipMaker` visualizes the output of the local alignment algorithm `blastz` in form of a PIP (percent identity plot), `BayesAligner` is a local alignment tool relying on Bayesian statistics and `DIALIGN` is a segment-based multiple alignment tool. The remaining programs are primarely concepted for phylogenetic footprinting. All the methods are described in more detail in the text. Restrictions concerning the sequence length refer to the webtools if possible.

	programs						
properties	TFsearch	rVISTA	PipMaker	FootPrinter	BayesAligner	DIALIGN	Tracker
webtool	+	+	+	+	+	+	–
free stand-alone tool	–	–	–	+	+	+	+
max sequence length	8000	>200000	2000000	40000/ #seqs	4000/ #seqs	1000000/ #seqs	>2000000
max number of sequences	1	2	20	>20	2	>20	>20
phylogenetic tree	no	no	no	yes	no	no	no
binding sites	yes	yes	no	no	no	no	no
important parameters	2	many	some	many	non	some	many
setting parameters	easy		easy	not easy	easy	easy	easy
runtime	–	very long	short	long	–	long	short
risk for false positives	high	low	low	–	low	low-high	low
risk for false negatives	–	–	high	high	–	high	–
significance measure	yes	yes	yes	yes	yes	no	no
human readable output	+	+	+	+	+	–	+
mashine readable output	+	+	–	+	+	+	++

CHAPTER 3

The Destiny of Duplicated Footprints

3.1 The DDC Model

Eucaryotic genomes contain multiple copies of many genes with closely related function. These copies arise from polyploidization or duplication of genome fragments. After duplication, one copy shields the second copy from natural selection. This causes accumulation of mutations that destroy the function of one of the copies since deleterious mutations occur much more frequently than beneficial ones. Finally, it should turn out that just one set of genes remains functional in most cases. But it is observed that a large proportion of duplicated genes is preserved for a long period of time.

In attempt to explain the high rate of duplicated gene preservation the duplication-degeneration-complementation (DDC) model was developed by *Force et al.* [31]. They assume that genes often have several functions, each of which may be controlled by different DNA regulatory regions. After duplication, degenerative mutations in these regulatory elements can increase the probability of duplicated gene preservation in the absence of positive selection. The usual mechanism is partitioning of ancestral functions (Fig. 3.1) leading to functional complementation of the duplicated genes which jointly retain the full set of subfunctions present in the original ancestral gene. This can be confirmed by the observation that duplicated genes with related function often show spatial and/or temporal partitioning of expression patterns.

The DDC model implies that ancestral regulatory information is distributed to both copies resulting in at least half of the genes and 'exactly' one half of the

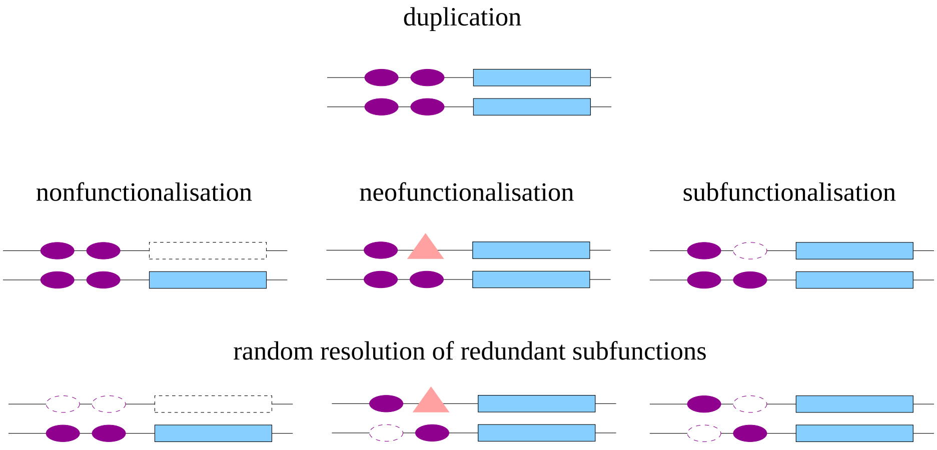


Figure 3.1: The duplication-degeneration-complementation (DDC) model by *Force et al.* After duplication the selective pressure is taken off the copies until non-neutral mutations occur. A null mutation in one copy of the gene result in nonfunctionalization and formation of a pseudogene (left). If a regulatory region acquires a new function at the expense of an essential function, both duplicated genes are retained (center). Initiation of subfunctionalization starts with a degenerative mutation in a regulatory region. This function is taken over by the second copy. Over time, the remaining regulatory regions undergo random resolution of persisting redundant subfunctions. In either case, half of the functional non-coding sequence is retained after resolution. Light blue boxes denote functional genes whereas boxes with dashed outline represent pseudogenes. The solid ovals denote regulatory regions with unique function, while dashed ovals indicate loss of function. New functional regulatory regions are represented by a triangle.

functional non-coding sequence of the duplicated region being retained (assuming unique function of the regulatory sites).

3.2 Structural Footprint Loss

Hox cluster duplication can lead to extensive loss of non-coding sequence conservation, as shown by Chiu et al. [21], but the causes remain unclear. There are three biologically distinct processes that can account for this phenomenon: (1) structural loss, (2) binding site turnover, and (3) adaptive modification. Structural loss is the loss of putative cis-regulatory elements due to gene loss and/or stochastic resolution of genetic redundancy in the aftermath of the duplication event. Binding site turnover is loss of noncoding sequence conservation due to the replacement of binding sites even though the function of the enhancer remains conserved. This was first documented in the Drosophila even skipped stripe 2 enhancer [50] and has since been documented for many other invertebrate taxa. In vertebrates, however, no widespread binding site turnover has been documented, which might have to do with a variety of reasons [20]. Adaptive modification, finally, would be a change in the sequence of cis-regulatory sites due to directional natural selection and would thus be associated with functional differences.

Loss of non-coding sequence conservation is associated with structural changes, namely gene loss. Hence the question arises whether the amount of loss observed is more than what should be expected from the changes in the gene-content. To address this question a model for estimating the amount of non-structural causes of footprint loss was developed in collaboration wit Günther Wagner [59].

To estimate the structural loss of footprints we have to take into account the three main causes: (1) Clearly, if a gene is lost, also the associated cis-regulatory elements will be lost (nonfunctionalization Fig. 3.1), disregarding enhancer sharing. (2) due to stochastic resolution of genetic redundancy half of the redundant enhancers are lost (subfunctionalization Fig. 3.1)(3) if a gene goes extinct, its cross-regulatory interactions within the gene cluster may be lost.

First, given the number of retained footprints one can estimate the amount of footprint retention. Gene-loss implies the loss of the associated cis-regulatory elements. Therefore, the amount of loss of non-coding sequence conservation has to be calculated in relation to the number of genes which are lost in the focal clusters. The retention probability of an ancestral footprint, $r(\mathrm{F})$, depends on the retention probability assuming that the associated coding gene is retained, $r(\mathrm{F}|\mathrm{G})$, and the probability that the gene is retained, $r(\mathrm{G})$:

$$r(\mathrm{F}) = r(\mathrm{F}|\mathrm{G})r(\mathrm{G})\,. \tag{3.1}$$

Implicitly, equ.(3.1) assumes that footprints are retained only when the associated gene is retained as well.

Second, all genes in the *Hox* cluster are paralogs. We call genes which are related by the most recent gene/cluster duplication *1st order paralogs*. Genes which retain 1st order paralogs ($P(1^{\mathrm{st}})$ are expected to resolve the genetic redundancy by, on average, losing 50% of their respective cis-regulatory inputs [31]. If only one copy of the gene survives ($1 - P(1^{\mathrm{st}})$, one would expect that all the relevant cis-regulatory elements are maintained. This can be written as:

$$\left[\frac{1}{2}P(1^{\mathrm{st}}) + \left(1 - P(1^{\mathrm{st}})\right)\right]\,. \tag{3.2}$$

Third, it is well known that *hox* genes are cross-regulatory. We assume that with the extinction of that gene its associated enhancer inputs to other genes will be lost as well. The expected amount of loss due to gene extinction therefore depends on the fraction $P(\mathrm{G_{ext}})$ of genes in the whole network that were lost and the fraction d of genes in the network which received regulatory input from these extinct genes. In general we do not know the degree d of cross-regulatory connectivity. We will assume that $d = 1$ which implies that each gene has a cross-regulatory link to every other gene. After elimination of cross-regulatory effects we have:

$$\big(1 - d\,P(\mathrm{G}_{\mathrm{ext}})\big)\,. \tag{3.3}$$

The expected footprint retention probability r_0 taking structural loss into account is a combination of equation 3.2 and 3.3:

$$\begin{aligned} r_0 &= \left[\frac{1}{2}P(1^{\mathrm{st}}) + \big(1 - P(1^{\mathrm{st}})\big)\right]\big(1 - d\,P(\mathrm{G}_{\mathrm{ext}})\big) \\ &= \big(1 - \frac{1}{2}P(1^{\mathrm{st}})\big)\big(1 - d\,P(\mathrm{G}_{\mathrm{ext}})\big)\,. \end{aligned} \tag{3.4}$$

Now we introduce a factor for the footprint loss due to non-structural causes. We call this probability α. The theoretical total retention rate of footprints is therefore

$$\hat{r}(\mathrm{F}|\mathrm{G}) = r_0(1 - \alpha)\,. \tag{3.5}$$

We can determine the rates for $P(1^{\mathrm{st}})$ and $P(\mathrm{G}_{\mathrm{ext}})$. The degree of cross-regulation d will be set to 1 and we can observe the rate of footprint retention per gene $r(F|G)$. Thus we can estimate the degree of non-structural footprint loss α, by solving equ.(3.5) and equ. (3.4) as:

$$\hat{\alpha} = 1 - \frac{r(\mathrm{F}|\mathrm{G})}{\big(1 - P(1^{\mathrm{st}})/2\big)\big(1 - P(\mathrm{G}_{\mathrm{ext}})\big)}\,. \tag{3.6}$$

$\hat{\alpha}$ is a minimal estimate for the degree of non-structural loss of phylogenetic footprints due to the assumption that $d = 1$.

3.3 Footprint Retention Statistics in Teleosts

3.3.1 Observed Footprint Retention

The qualitative results in [21] suggested that cluster duplication leads to a massive loss of non-coding sequence conservation, which could be indicative of extensive modifications in the function of Hox genes. If this is the case one would expect to find a similar degree of loss of conservation in other teleost *Hox* clusters. Here we extend the analysis to include the two *HoxA* clusters of Takifugu, based on the published genomic sequence [25].

Therefore, the analysis of the *HoxA* clusters was performed in two steps. A re-evaluation of the analysis reported in [21], see Table 7.1, and a combined evaluation that uses the sequences from Takifugu as well. Tables 7.2ff summarize the additional footprints and has been used as the basis for the summary statistics reported in Table 3.1. The `tracker` program recovers all footprints reported in [21] with the three exceptions:

11-9-b is a footprint of length 9. It is too short to be accepted as significant hit with the default parameter settings of `tracker`:

```
HsA_11-10-b       GTCTCTCGGCTCGGGGCTGGAACTCCGGCCC--
DrAb_11-10-b      --CTAGAAAACAACGGCTGGAACCATTGAAAGC
                                ********* 
```

up13-c does not exist at the reported location. A `clustalw` alignment yields:

```
HfM_up13-c        ACAGAAAACAGTTTTTGTAAAATAGTCATTTAGTATTAAAT
DrHoxAa_up13-c    -----------------CAAAAAAAAAAAAAAACACTG---
                                    **** *   *   *  *
```

As the footprint above, **5-4-b** does not correspond to a significant match at the reported positions. The corresponding `clustalw` alignment is:

```
HsA_5-4-b         --GCTGTGCTGCGATAGGGGGTTGTGGGAGGGCAAAAAAAAAAAAAAAGGTGATCGC--G
HfM_5-4-b         TAATTAAGAGATCGAAGCACTTTCTCCAACTTATTTAATGGAGGATGATTTATTTGCCCA
                      *  *       **    ** *   *       **   *  *  *  *  * **

HsA_5-4-b         GGTTGAGGAAAACAAAGTTTCCATTCTAAACAATGGGGTGGTAGA
HfM_5-4-b         GCTAGTCAGAAAATGACCTTCTGTGCTCTCCCC----ATCTTAGA
                  * * *    ***   *  ***  * **   *       *  ****
```

This footprints are included in *italics* in Table 7.1.

The calculation of retention rates assumes that the combination of footprints from *Heterodontus franciscii* and *Homo sapiens* can be interpreted as ancestral stage. 126 footprint cliques are found in either the shark or human *HoxA* cluster or both. In contrast, there are only 68 of those retained in at least one zebrafish clusters and 59 are retained in at least one fugu *HoxA* cluster, while only 8 and 9 footprint cliques, resp., survived in both paralog clusters. This corresponds to a retention rate of 27% and 23% respectively (Table 3.1). This confirms the qualitative observation in [21], that *Hox* cluster duplication is associated with a massive loss of non-coding sequence conservation. The per gene retention rates are listed in Table 3.1 and are between 0.49 for zebrafish and 0.37 for Takifugu *HoxA* clusters.

To test whether comparable numbers are observed for all *Hox* clusters, we went about performing analogous analysis for the other *Hox* clusters (Fig. 3.2). It turned out to be difficult at present since either the sequences for Takifugu and zebrafish are incomplete and/or the corresponding outgroup sequences are not yet available. An additional source of uncertainty is the fact that the 3'-end of the DrBb cluster is missing in the currently available assembly, see Fig. 3.2. The preliminary footprint clique statistics for the *HoxB* clusters are also compiled in Table 3.1. These numbers, which are substantially larger than for the *HoxA* clusters, should be viewed with caution. In particular, the footprint retention rates $r(\mathrm{F})$ are upper bounds since we miss footprints that have been lost completely in either mammalia or fish lineages. For the *HoxC* and *HoxD* clusters sequence data of duplicate clusters are currently not publicly available with sufficient data quality.

3.3.2 Estimation of Non-Structural Causes

Before we apply the statistical model proposed in section 3.2 to analyze the retention data outlined above, we want to point out a methodological issue in scoring the rate of footprint loss in this type of data.

There are 53 footprint cliques in zebrafish and Takifugu that have no counterpart in shark or human; of these 14 were found only in zebrafish and 10 only in Takifugu. These footprint cliques most likely correspond either to cis-regulatory elements which were lost independently in the shark and human lineage or which are footprint cliques acquired in the stem lineage of teleost fish. These footprint cliques, however, cannot be used to estimate the rate of footprint clique retention after cluster duplication, because one cannot detect the footprint cliques that have only been maintained in one of the paralog clusters. For that reasons we ignore all footprint cliques which have no counterpart in shark or human. We have to keep in mind that the counts of footprint cliques are just a sample of all putative cis-regulatory elements involved. If, however, the retention rates of these footprint cliques are comparable to those present in shark and human, the statistics will still give valid estimates.

The fraction of extinct genes in the *Hox* network, $P(\mathrm{G}_{\mathrm{ext}})$, is calculated by counting the paralog group members on each of the four clusters in the ancestor of bony fish, i.e. the most recent common ancestor of mouse and zebrafish, for instance. This number is compared with the number of representatives from different paralog groups which are present in the two duplicated clusters of a teleost.

The number and identity of genes in the most recent common ancestor of bony fish is based on the maximal parsimony reconstruction in [3]. For instance, the ancestor of bony fish has 11 paralog group members in *HoxA* while zebrafish *HoxAa* and *HoxAb* only have a total of 9 paralog groups represented. In other words 18% (2) of the genes in the ancestral *HoxA* cluster went extinct in the zebrafish lineage, i.e. have no descendant gene copy in the zebrafish genome. In total there are 42 genes in the four ancestral *Hox* clusters of which only 37 have at least one descendant gene in zebrafish. This means that 12% of the genes

Table 3.1: Footprint clique retention statistic after cluster duplication based on alignment of human, shark, pufferfish, zebrafish and striped bass sequences.

Cluster	#genes	r(G)	#pFC	r(F)	r(F|G)
	HoxA Clusters				
DrHoxAa	7	0.63	39	0.31	0.49
DrHoxAb	5	0.45	29	0.23	0.51
DrHoxA	12	0.55	68	0.27	0.49
TrHoxAa	9	0.82	47	0.37	0.45
TrHoxAb	5	0.45	12	0.10	0.21
TrHoxA	14	0.64	59	0.23	0.37
	HoxB Clusters				
DrHoxBa	8+	0.8+	62	0.53	< 0.66
DrHoxBb	4	0.4	43	0.37	0.92
DrHoxB	12+	0.6+	105	0.45	< 0.75
TrHoxBa	8	0.8	69	0.59	0.74
TrHoxBb	3	0.3	35	0.30	1.00
TrHoxB	11	0.55	104	0.44	0.8

Dr: zebra fish, Tr: Takifugu #genes: number of coding genes retained in cluster #pFC: number of plesiomorphic phylogenetic footprint cliques, i.e., footprint cliques that have a counterpart in shark or human. See section 3.2 for the definition of the retention rates.

Due to limited data the retention rates for the *HoxB* clusters are only upper bounds. For the *DrHoxBa* cluster we count only the genes that are contained in available sequence data, see the caption of Fig. 3.2 for details.

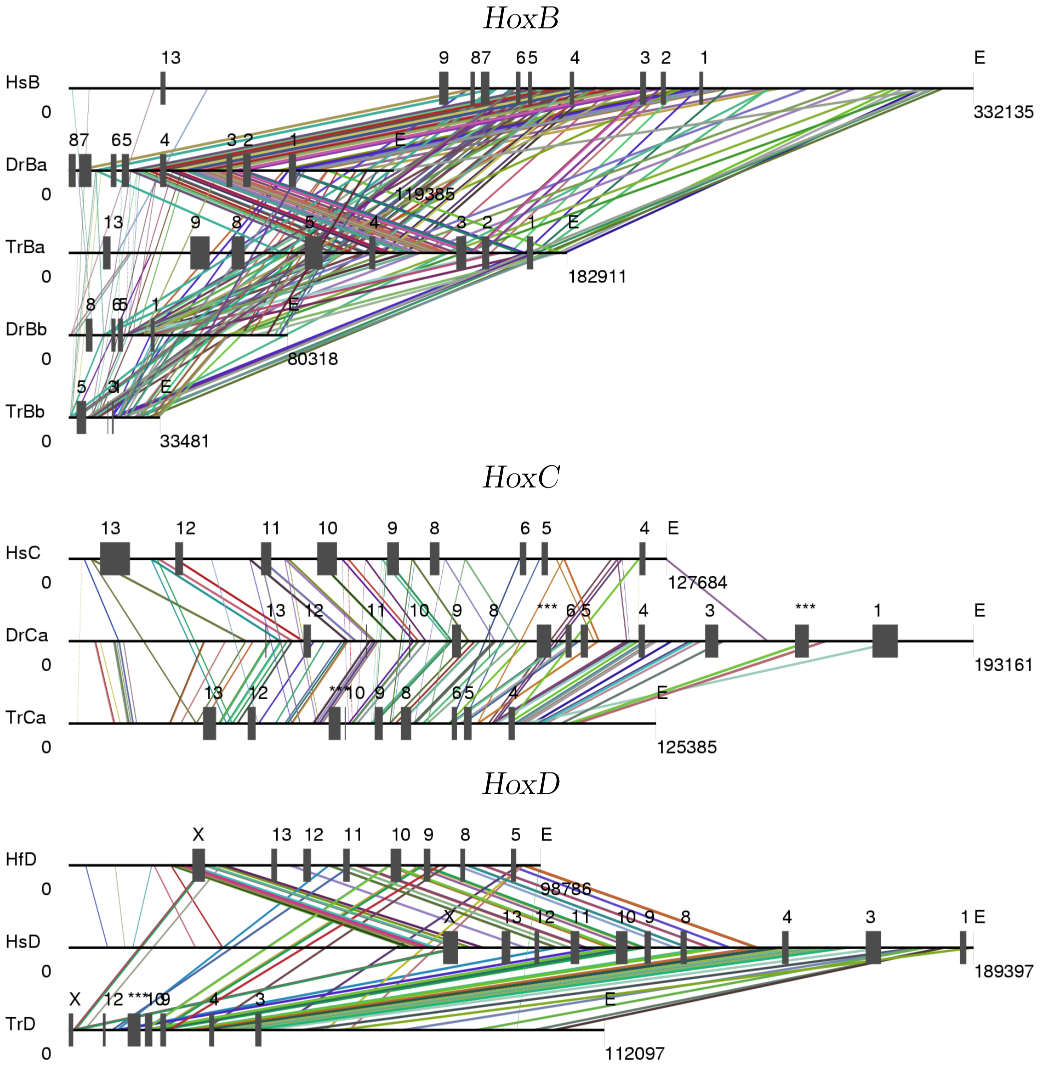


Figure 3.2: Phylogenetic footprints in *HoxB*, *HoxC*, and *HoxD* clusters. Such overviews are automatically generated by `tracker`. Each line corresponds to a footprint, consistent cliques are shown with the same color. Input sequences were obtained as follows: HsB = NT_010783 [931646-1263780] reverse complement, HsC = NT_009563 [580371-708054] r.c., HsD = NT_037537 [4075338-end]; HfD = AF224263; DrBa = AL645782, DrBb = AL645798, DrCa is a composite of zK81P22.00296(r.c.) + 3084×N + zK81P22.01466(r.c.) + 2956× N + zK81P22.00552 from the Sanger site (download 12.1.03) with approximately 3000 Ns as spacers inserted (marked by *** in the drawing); TrBa is a composite of scaffold_1439(r.c) + 2501×N + scaffold_706 from version 3.0 of the Fugu DB [25], TrBb is a composite of scaffold_1245 [59047-end] + 3020×N + scaffold_2182 [1-19481], TrC is a composite of scaffold_93[184545-end]+2936×N + scaffold_285 [134158-end] (r.c.), TrD is a composite of scaffold_3959 (r.c.) + 2645×N + scaffold_214 [160440-end] (r.c.). All these composite sequence are consistent with a single contiguous cluster.

Table 3.2: Conditional footprint retention statistics after *HoxA* cluster duplication based on the predictions of the structural loss model. Note that the predicted retention rate based on the structural loss model is consistently higher than the observed rate of loss, indicating other, non-structural causes of sequence conservation loss. There is a notable asymmetry in the predicted minimal rate of non-structural conservation loss between the clusters. In zebrafish the *HoxAa* cluster seems to be twice as strongly modified while in Takifugu the *HoxAb* cluster has an exceptionally high minimal modification rate of 0.48. This pattern is consistent with rates of coding sequence evolution among paralog *Hox* genes in these species (Takahashi et al., in prep.).

Cluster	#genes	$P(1^{\text{st}})$	$r(\text{F}\|\text{G})$		$\hat{\alpha}$
			data	equ.(3.4)	
DrHoxAa	7	0.43	0.49	0.69	0.29
DrHoxAb	5	0.60	0.51	0.62	0.18
DrHoxA	12	0.50	0.49	0.66	0.26
TrHoxAa	9	0.56	0.45	0.58	0.22
TrHoxAb	5	1.00	0.21	0.40	0.48
TrHoxA	14	0.71	0.37	0.52	0.29

went extinct, or $P(\text{G}_{\text{ext}}) = 0.12$. Similarly, in the Takifugu *Hox* clusters there are descendants of 34 of the 42 genes present in the ancestral *Hox* clusters, which means that the extinction frequency in the Takifugu lineage is $P(\text{G}_{\text{ext}}) = 0.19$ (Chris T. Amemiya, pers. comm. 2003).

The fraction of genes which retain first order paralogs $P(1^{\text{st}})$ differs between zebrafish and Takifugu *HoxA* clusters. There are six genes in zebrafish *HoxA* clusters which have 1st order paralogs: *HoxA-13a/b*, *HoxA-11a/b*, and *HoxA-9a/b*. Hence the fraction of 1st order paralog genes in zebrafish is $P(1^{\text{st}}) = 0.50$. In Takifugu there are ten genes which have first order paralogs retained: *HoxA-13a/b*, *HoxA-11a/b*, *HoxA-10a/b*, *HoxA-9a/b*, and *HoxA-2a/b*; hence $P(1^{\text{st}}) = 0.71$.

In order to account for the loss of genes in the focal *HoxA* clusters after duplication, we calculate the conditional retention rates: we find about 50% for zebrafish and 37% overall for Takifugu. This suggests that, corrected for gene loss in the *HoxA* cluster, Takifugu has a lower retention rate than zebrafish. The two paralog clusters in Takifugu have strongly different retention rates, 0.21 for the *HoxAb* cluster and 0.45 for *HoxAa* cluster. In contrast, the conditional retention rate in zebrafish is about the same for both clusters, 0.49 and 0.51 respectively.

Applying the structural loss model to the footprint loss data of the *HoxA* clusters shows that the observed amount of retention is in all cases less than predicted as the minimal amount of retention if only structural reasons would cause loss of sequence conservation. Hence the model is consistent with the data, in the sense that we do not observe more conservation than the minimal amount predicted by this model.

Calculating the minimal probability of footprint clique loss due to non-structural reasons (binding site turnover and directional selection) shows that in zebrafish and Takifugu this rate is roughly comparable, about 26% and 29% respectively, see Table 3.2. The slightly higher rate in Takifugu, however, is entirely accounted for by the higher rate estimate for the *HoxAb* cluster. The non-structural modification rate in the *HoxAa* cluster is 0.22, about the same as in zebrafish, while the minimal rate of non-structural modification in the Takifugu *HoxAb* cluster is 48%. This suggests that there was a differential loss of non-coding sequence conservation in the Takifugu *HoxAb* cluster. Assuming that the probability of functionally conservative binding site turnover is about the same in the two paralog clusters, this result strongly suggests that the Takifugu *HoxAb* cluster experienced adaptive modification at a higher rate than both the Takifugu *HoxAa* cluster and either of the zebrafish clusters.

3.4 Evidence for Adaptive Modification

While the role of *Hox* genes in animal development is well established, their role in evolution is less well understood, see e.g. [24]. A particularly intriguing problem is the role of *Hox* cluster duplications in vertebrate evolution. All invertebrates examined today have at most one cluster, including the sister taxon of vertebrates, the cephalochordates, e.g. *Branchiostoma floridae* [35]. In contrast, even the primitive jawless vertebrates have at least three separate clusters [30, 42] and teleosts, like zebra fish and fugu, have up to seven or eight *Hox* clusters [3]. It is not clear whether this accumulation of *Hox* clusters had played a biologically important role in the evolution of the various vertebrate body plans [51] or whether the retention of duplicated *Hox* clusters is a passive consequence of the resolution of genetic redundancy [31]. One approach to address these issues is to examine the molecular evolution of the *Hox* genes and *Hox* clusters after duplication. Is there evidence that the duplicated *Hox* clusters experienced lineage and cluster specific modifications by natural selection? Or is the evolution of duplicated clusters only a consequence of the resolution of genetic redundancy? An affirmative answer to the former question would suggest that duplicated *Hox* clusters provided genetic opportunities for adaptive evolution. An affirmative answer to the second question would suggest that *Hox* cluster duplication did not play a role in the evolution of the affected clade. The most recent cluster duplication event documented is that which leads to the additional *Hox* clusters in the teleost lineage. Teleost *Hox* genes are thus the best system to investigate the evolutionary forces acting on *Hox* genes after duplication.

The rate of coding sequence evolution in duplicated fish *Hox* genes has been shown to be increased compared to the unduplicated orthologs [23] and there is some evidence that duplicated *Hox* genes experienced directional selection [57]. These findings are consistent with the idea that the duplicated *Hox* genes became involved in adaptive evolutionary changes and played an active role in the evolution of the teleost disparity and diversity. For non-coding, putative cis-regulatory sequences it has also been found that massive modifications are associated with the duplications of the *HoxA* cluster [21]. These changes, however, are associated with other structural changes in the *Hox* clusters, most notably the loss of genes [3], and the shifting of functions among paralog genes [54]. In this paper we propose a simple model to predict the expected loss of non-coding sequence conservation (NCSC) due to gene loss and resolution of genetic redundancy according to the DDC model [31]. This model allows to estimate whether the loss of conservation is more or less than can be attributed to these structural reasons.

We applied the `tracker` software and the loss model to sequence data from the *HoxA* clusters of zebrafish and fugu and found that in all cases the loss of NCSC was more extensive than predicted by the model. This means that the modification of non-coding sequences after cluster duplication was more extensive than what can be explained by structural changes of the clusters. Even though the sets of genes retained in zebrafish and fugu *HoxA* clusters are somewhat different the estimated overall excess loss of NCSC is comparable. This shows that the estimates of non-structural conservation loss are consistent among different lineages of teleost fishes.

At face value, the existence of excess modification of putative cis-regulatory elements is consistent with the idea that the duplicated *Hox* clusters are affected by adaptive modifications during teleost phylogeny. This interpretation, however, is not the only compatible with that evidence. Duplicated genes and the resulting genetic redundancy could also promote the turnover of transcription factor binding sites, even though the overall function may not be affected. This would be a form of neutral drift of cis-regulatory sequence elements [49]. Another possibility is that the genomic re-arrangements following the cluster or genome duplications have caused an increase in mutation rate which leads to a higher rate of loss of sequence conservation than in the unduplicated lineages. It is known, for instance, that GC content [11, 77] and the frequency of CpG islands is correlated with increased mutation rate in mammals [27], and perhaps also in other vertebrates. Neither the CG content nor the number and size of CpG island show a phylogenetic pattern that would explain the loss of NCSC in the teleost lineages (data not shown). We thus find no factors that would predict an increased mutation rate of non-coding sequences in these lineages.

With our present methods it is not possible to distinguish between natural selection and increased binding site turnover to explain the excess loss of NCSC. We note, however, that the pronounced asymmetry in the per gene retention rate of NCSC in fugu *HoxAb* cluster could potentially be caused by natural selection differentially modifying the function of *HoxAb* type genes in the fugu lineage. It is reasonable to expect that binding site turnover among duplicated and thus redundant gene clusters is symmetrical, because of the basic symmetry of the situation immediately after duplication. As the redundancy is resolved, maybe differentially among the duplicated clusters, the TF binding site turnover should also cease in both paralog clusters because of complementation. This observation is suggestive of the effect of directional selection, but it would be desirable to find other statistical patterns linked to the action of natural selection for independent confirmation.

CHAPTER 4

Assignment of Orthologous Hox Clusters

4.1 Models for Hox Cluster Evolution

Homologous *hox* genes have been found in every bilaterian animal investigated and basically show clustered organization, although gene and cluster number vary. Early in metazoan evolution, *Hox* and *ParaHox* have resulted from a *ProtoHox* gene cluster [55]. Tandem duplications are thought to have increased the number of paralogous groups. Multiple *Hox* clusters have arisen from whole cluster duplication maybe as a result of genome duplication via polyploidization. The effect is, that vertebrates, in contrast to all invertebrates examined, have multiple *hox* gene clusters that presumably have arisen from a single ancestral cluster in the most recent common ancestor of chordates, i.e. amphioxus and vertebrates [35, 43]. The timing of the *Hox* cluster duplication events in vertebrate phylogeny is still somewhat unclear. The available data strongly suggest a 4-*Hox* cluster organization in the crowngroup tetrapods [58]. On the other hand, the cephalochordate amphioxus has a single *Hox* cluster. Two distinct models are currently likely to explain the evolutionary scenario along the "long way from amphioxus to us". One model, the 2R hypothesis, suggests two rounds of genome duplication, leading to $((AB)(CD))$ by two sequential duplication events [30]. An alternative model has been put forward by *Bailey et al.* [6]. It assumes three instead of two rounds of duplication whereby the ancestral *Hox* cluster was D-like and duplicated to create an A-like cluster from which the B and C clusters arose in turn $(D(A(BC)))$.

Discovery of an organism showing an intermediate cluster composition would rule out one of the models above. Therefore, *Hox* clusters branching off the phylogenetic tree connecting amphioxus and tetrapods were sequenced.

The most basal branch of vertebrates leads to lampreys (e.g. *Petromyzon marinus*). We demonstrated that the 3 to 4 *Hox* clusters of *Petromyzon marinus* and other vertebrate species had arisen from independent duplication since the paralogous *Hox* clusters in lampreys are more similar to each other than any of these clusters and a vertebrate *Hox* cluster [34]. This supports the hypothesis that the last common ancestor of agnathans and gnathostomes had only a single *Hox* cluster.

The least derived group within gnathostomes is the condrichthyes including horn shark (*Heterodontus franciscii*). A very popular hypothesis is that the common ancestor of shark and bony vertebrates (which includes teleost fish as well as tetrapods) had four clusters homologous to that in humans [39]. To test this idea the *Hox* clusters of the horn shark have been isolated and sequenced. Currently, two clusters, called *N* and *M* are available [44]. While the *M* cluster is clearly homologous to the human *HoxA* cluster, it was more difficult to assign the homology to the *HoxN* cluster. In the original description *HfHoxN* was identified as homologous to the human *HoxD* cluster, but there is also evidence consistent with homology to *HoxC* cluster [51]. Our analysis support the idea, that *HfHoxN* is D-like.

4.2 Footprints as Phylogenetic Information

The detection of non-coding sequence conservation between the horn shark *Hox* clusters *N* and *M* and tetrapod *Hox* clusters is carried out by `tracker`. It detects clusters of conserved footprints that are not easily decomposed into individual footprints. Our statistical analysis below is therefore based on the total length of significantly homologous non-coding sequence fragment between pairs of clusters. This measure is roughly proportional to the number of individual footprints. Homologous footprints are necessarily co-linear (disregarding the possibility of local transpositions or inversions which cannot be resolved with the present analysis method due to the highly diverged sequence outside the footprint clusters). Non-colinear `tracker`-hits are therefore disregarded (marked by × in the supplemental material).

The `tracker` program produces alignments of the footprint cliques using `dialign` [56]. These are padded with "gap" characters in those sequences that do not take part in a particular clique and then concatenated. The resulting "alignment" is sparse in the sense that the "gap" character is the most frequent letter. The reconstruction of phylogenies from such a dataset has to take three complications into account that: (1) gene loss will cause almost certainly caused the loss of all the the associated regulatory sequences. In the extreme case, presence-absence data footprints might just reflect that presence-absence pattern of the genes. (2) We cannot expect to have detected *all* footprints in all species. (3) Gain and loss of footprints are not symmetric processes: in fact footprint loss is much easier than the *de novo* creation. These complications can be circumvented by considering only mutations within conserved non-coding regions, i.e., within the footprint cliques detected by the `tracker` program. The distance of two clusters is therefore derived as the fraction of mutations within cliques that are shared by the two clusters. Technically, this amounts to treating "gaps" that arise because a certain cluster does not share a particular footprint cliques as missing data rather than as an additional character state.

We use a different distance-based and parsimony-type approaches here: Neighbor joining method [62] (implemented in the `phylip` package, version 3.6) [28], the canonical split decomposition [7], Buneman trees [18], parsimony splits and P-trees [8]. With the exception of NJ these methods are implemented in the `splitstree` package (version 3.1) [41]. The split-based methods are particularly suitable for our purposes because they are known to be very conservative it that they tend to produce multifurcations rather than poorly supported edges [66].

The following sequences are used for the analysis: Shark (*Heterodontus francisci*): M-cluster HfM = AF479755, N-cluster HfN = AF224263; Human (*Homo sapiens*): HsA = AC004080.2rc + AC010990 [201-6508]rc + AC004079 [75001-end]rc, HsB = NT_010783 [931646-1263780]rc, HsC = NT_009563 [580371-708054]rc, HsD = NT_037537 [4075338-end]. Rat (*Rattus norvegicus*): RnA = NW_043751 [910030-1194462]rc, RnB = NW_042671 [264022-581839], RnC = NW_044048 [722873-1060956] RnD = NW_042732 [1061702-1217610]rc. Here "rc" means that the reverse complement of the sequence has been used (after extracting the indicated interval).

4.3 The Shark HoxN Cluster

A comparison of the protein sequences of the shark *HoxN* cluster with mammalian Hox protein sequences is consistent with D-likeness, although the data in Table 4.1 do not show an unambiguous picture. In particular, the *HoxD* proteins are not always the ones with the highest degree of sequence identity, see Table 4.1. In a similar vein, the analysis of Hox genes and of genes linked to the Hox clusters such as collagenes does not yield an unambiguous picture for branching order of the four mammalian Hox clusters [6].

Let us now turn to the analysis of the conserved parts of the non-coding sequences. Table 4.2 summarizes the results of pairwise comparisons of shark and

Table 4.1: Best Correspondences of Hox proteins with the *HoxN* sequence of the hornshark. Number are identities in protein alignments obtained with `clustalw` [74]. *Italics* and sans serif fonts indicate that the best match is by a the *human* or rat sequence, respectively. A dash — indicates that the corresponding gene does not exist in mammalian Hox clusters.

Cluster	evx	13	12	11	10	9	8	5
A		70	—	*57*		63	—	*53*
B	—		—	—	—		68	*48*
C	—		*48*	54		63	*69*	44
D	*81*	68	*48*	*57*	*69*	*61*	*71*	—

human (or rat) Hox clusters. It should be noted that the sequence of the shark *HoxN* cluster is incomplete, spanning only the sequence from *evx* to (almost) *Hox-4*. The data show a particularly high conservation of non-coding sequences in the range from *Hox-4* to *Hox-1* between shark *HoxM* and mammalian *HoxA* sequences. As a consequence, the counts for *HoxN* are significantly smaller. In table 4.2 we therefore show also the counts for the *HfM* cluster restricted to the region between *evx* and *Hox-4*. The total length of sequences conserved between shark and mammalian clusters in this region is indeed comparable between *HfM* and *HfN*.

The homology of the shark *HoxM* and the mammalian *HoxA* clusters in obvious from these data. For the *HoxN* sequence we find little distinction when counting colinear cliques and only a moderate signal in the numbers of co-linear clusters. The total length of the conserved regions, however, is more than twice as large with *HoxD* than with *HoxC* and about 50% longer in *HoxD* compared to *HoxA*. The location and distribution of the footprint cliques, Fig. 4.1 also strongly argues for a homology with *HoxD* rather than *HoxC*.

A comparison of *HfHoxN* with the fugu *HoxCa* and *HoxD* sequences also places *HfHoxN* with the *D* rather than *C* cluster. These data must be interpreted with caution: (i) The Fugu sequences are preliminary constructs combining two or three scaffolds and hence not complete. (ii) Even though the current version 3.0 of the Fugu genome database [25] does not contain evidence of a *Cb* cluster, it is most likely that the teleost *C* cluster was duplicated since the zebrafish does have both a *HoxCa* and a *Cb* cluster [3]. The duplication event might have caused the additional loss of a substantial number of footprints.

The sensitivity of the tracker method in increased by including more sequences. In particular, homologous footprints can be identified between two sequences even if they do not yield a significant signal when the two sequences are compared directly. We have therefore run a complete analysis of both shark

Table 4.2: Pairwise comparison of non-coding sequences in the shark Hox clusters with mammalian Hox clusters. In addition we report the comparison with preliminary *HoxC* and *HoxD* cluster sequences (obtained from version 3.0 of the Fugu database [25, 4]; see [59] for details). Comparisons with the duplicated, highly diverges *HoxA* and *HoxB* clusters are meaningless.

		Shark *HoxM*				Shark *HoxN*			
	mamm.	*HoxA*	*HoxB*	*HoxC*	*HoxD*	*HoxA*	*HoxB*	*HoxC*	*HoxD*
		evx to *hox-4* only							
Cliques	Homo	51	30	15	12	21	23	21	36
	Rattus	56	19	13	7	25	22	23	27
	Fugu	*	*	*	*	*	*	12	17
Length	Homo	2955	1554	736	652	1359	935	891	2548
	Rattus	3871	1008	669	537	1633	910	1050	2468
	Fugu	*	*	*	*	*	*	508	1000
		Complete cluster							
Cliques	Homo	96	35	17	20				
	Rattus	97	25	17	15				
Length	Homo	7392	1995	791	1036				
	Rattus	7167	1525	827	868				

clusters and all four human Hox clusters. The supplemental material lists all footprint cliques in the range from *evx* to *hox-1* that appear in at least one shark and at least one human cluster. The statistics of the conserved regions between clusters is summarized in Table 4.3.

Table 4.3: Comparison of phylogenetic footprints from a `tracker` run of both shark and all four human clusters. Only co-linear cliques in range between *evx* and *hox-1* are counted. The data contain six cliques (484, 485, 486, 513, 514, 515 in the supplement) of which at most three are consistent with co-linearity. These are counted with a weight $1/2$.

Count	HsA	HsB	HsC	HsD	HsA	HsB	HsC	HsD
	Hf-M				Hf-N			
cliques	**79**	23	13	16	15	10	20	**25**
length					1728	961	1148	**1995**

These data clearly indicate that the shark *HoxN* cluster is D-like at least as far as the non-coding sequences are concerned. In fact, based on total size of the footprints that are shared between clusters, the next candidate would be the mammalian A-cluster, not the C-cluster as proposed in [51].

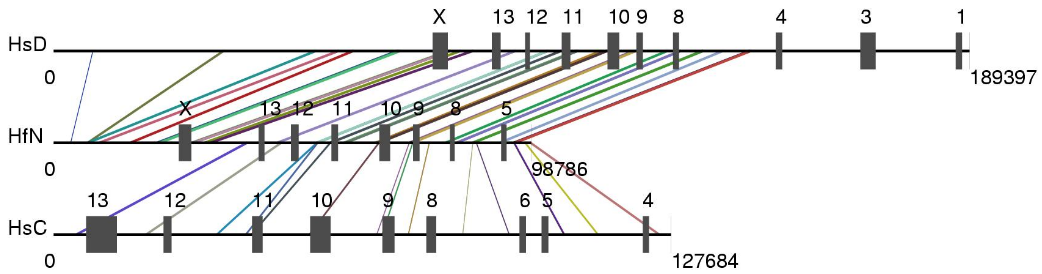


Figure 4.1: Overview of the phylogenetic footprint cliques produced by `tracker` for the comparison of the horn shark *HoxN* sequence (HfN) and the human *HoxC* (HsC) and *HoxD* (HsD) sequences, respectively. X denotes the *Evx* genes.

A phylogenetic analysis of the combined footprint cliques of the four mammalian clusters for either human or rat together with the two available shark sequences strongly suggests that the *HoxN* cluster is not only most similar to the mammalian *HoxD* clusters but in fact is a true homologue. Both distance-based, Fig. 4.2 and parsimony-based methods, Fig. 4.3, agree on this interpretation. We have chosen a variety of split-based algorithms for this analysis for this analysis because these techniques are known rather produce multifurcations than poorly supported edges. For comparison standard neighbor-joining trees are shown in Fig. 4.2.

All data presented in Figs. 4.2 and 4.3 either support the conclusion that the shark *HoxN* cluster is homologous with mammalian *HoxD* cluster or are at least consistent with this conclusion (whenever the *HfHoxN-HoxD* node is a multifurcation).

The evidence presented in this paper supports the original hypothesis, namely that the shark HoxN cluster is orthologous to the mammalian *HoxD* cluster [44]. The method employed is novel, namely to use the number and extent of non-coding sequences for phylogenetic inferences. Below we will discuss the use of this type of data as well as the implications of the present finding for our understanding of Hox cluster evolution in vertebrates.

Conserved non-coding sequences have long been used to find candidate cis-regulatory elements, see [26] for a review. Identification of putative cis-regulatory sequences requires long stretches of sequence from distantly related species [72] or a set of species which have sufficient additive divergence among them [71]. More recently this method has been used to trace the non-coding sequence divergence after *HoxA* cluster duplication in teleosts [22]. In this paper it has been shown that non-coding sequences can remain highly conserved in the absence of Hox gene cluster duplication, as documented between the shark *HoxM* and the mammalian *HoxA* cluster (see also this paper). Hence it is possible to treat

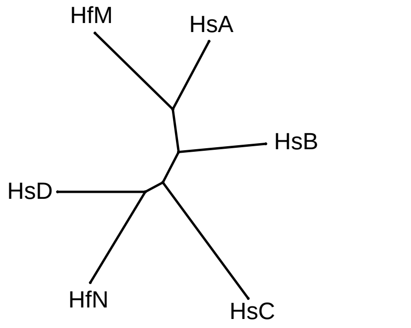
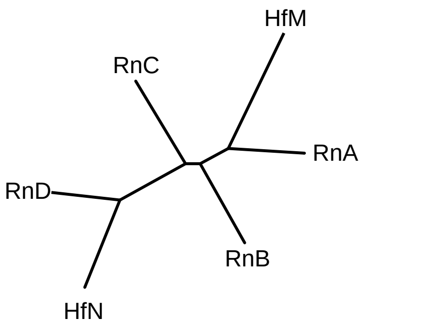
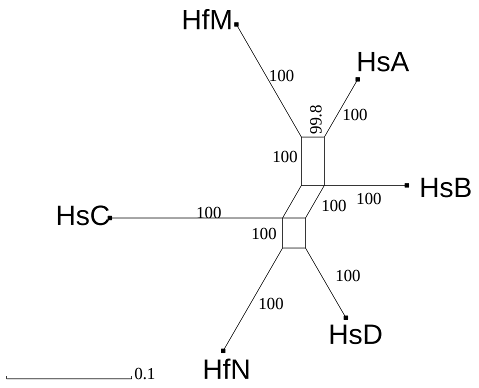
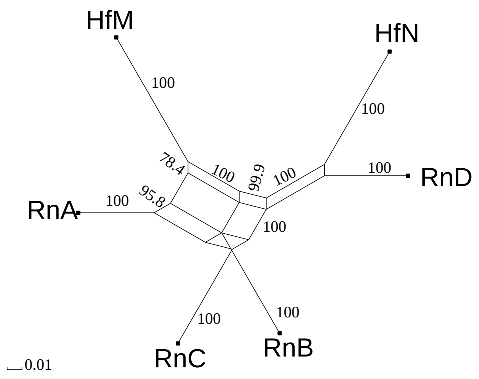
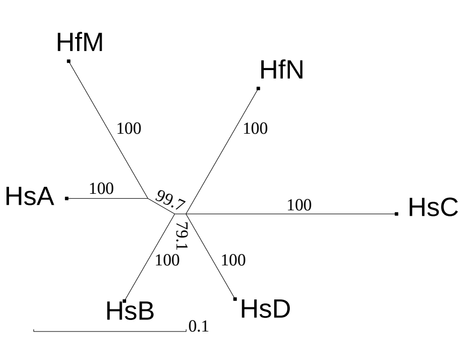
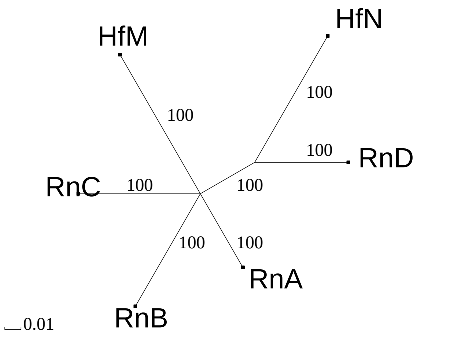
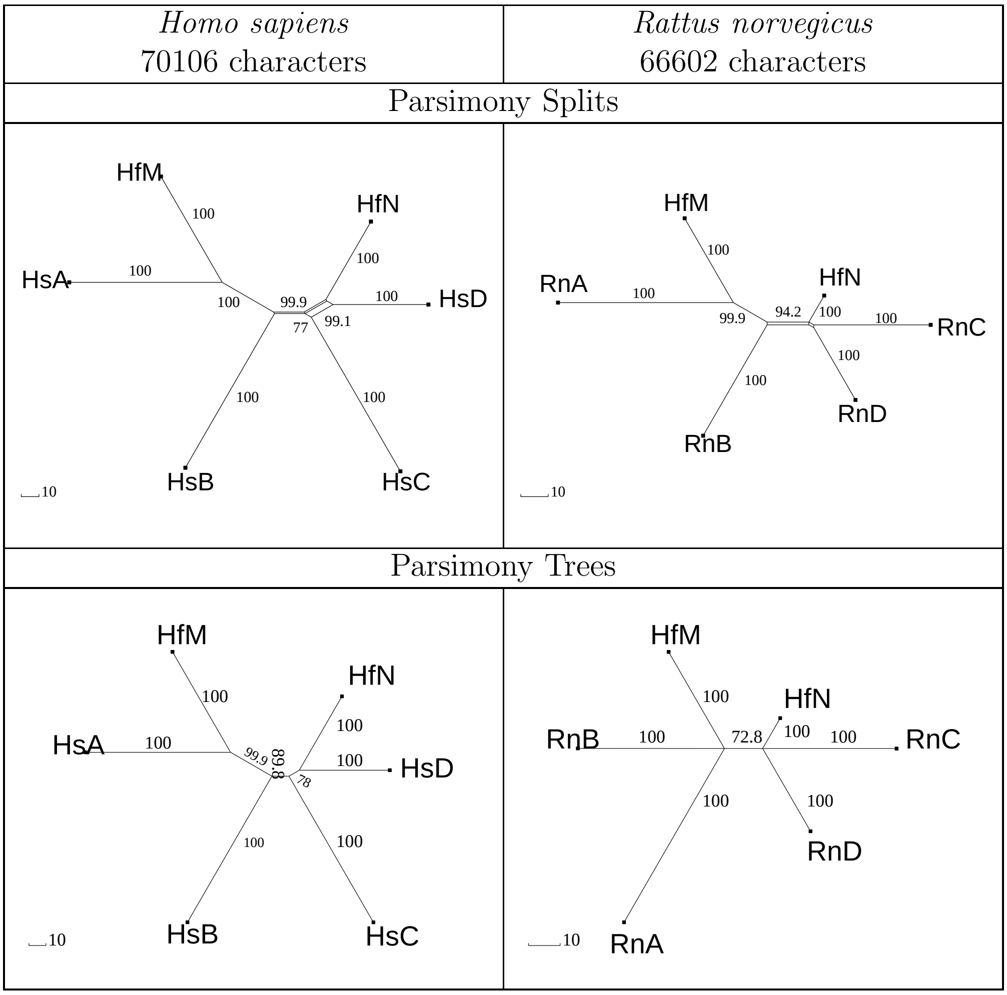
<table>
<tr><td><em>Homo sapiens</em>
70106 characters</td><td><em>Rattus norvegicus</em>
66602 characters</td></tr>
<tr><td colspan="2">Neighbor Joining Trees</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2">Buneman graphs</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2">Split-Based Trees</td></tr>
<tr><td>
</td><td>
</td></tr>
</table>

Figure 4.2: Distances based phylogenies of shark and mammalian Hox clusters. Neighbor joining trees [62] are computed using Felsenstein's `phylip` package (version 3.6). Buneman graphs representing the canonical decomposition of the distance function and the split-based Buneman trees are computed using Daniel Huson's `splitstree` package (version 3.1) [41].



Homo sapiens
70106 characters

Rattus norvegicus
66602 characters

Parsimony Splits

Parsimony Trees

Figure 4.3: Parsimony based phylogenies of shark and mammalian Hox clusters computed using `splitstree` [41].

the loss of ancestrally conserved sequences as potential apomorphies and thus as source of phylogenetic signal. The congruence between the structural and coding sequence evidence and the comparison on non-coding sequence conservation for *HoxM* and *HoxA* cluster validates this assumption. In the case of the shark *HoxN* cluster the evidence from coding sequence and structural organization is less strong and we thus rely on the evidence from non-coding sequence conservation. While the signal is still not as strong as for the HoxM each analysis is at least consistent and in many cases positively supportive of orthology between shark *HoxN* and mammalian HoxD cluster.

The conclusion that both the shark *HoxM* as well as the *HoxN* clusters are directly orthologous to the mammalian *HoxA* and *HoxD* clusters has important implications for the history of Hox cluster duplications. It follows that the most recent common ancestor of cartilaginous fishes and the bony fish clade (which includes mammals) had at least four Hox clusters orhologous to the four mammalian Hox clusters. It is thus likely that sharks have two more clusters than those currently described. This evidence also confirms the hypothesis of Peter Holland that the four cluster situation typical for most major gnathostome lineages has arisen before the most recent common ancestor of all Recent gnathostomes [35, 40]. Of course this result does not guarantee that all gnathostome lineages in fact have at least four Hox clusters since clusters can get lost. This can happen in particular soon after the duplication, which could have occurred shortly before the split between the shark and mammalian lineages.

CHAPTER 5

The Fin Limb Transition

5.1 Origin of the Tetrapod Limb

In the Devonian period various sarcopterygian fish were preadapted for moving out of water onto land. These fishes are a group of bony fishes (osteichthyes) with fleshy fins (lobe-finned fishes). They had functional lungs and two pairs of bone-strengthened muscular fins on which they could move their bodies out of the water. One of the recent sarcopterygian fishes is *Latimeria menadoensis*. The transition from fishes to crawling four-legged tetrapods occurred 370 million years ago "one sunny afternoon in the Devon" 5.1 and encompassed three or more separate lineages. However, there is strong evidence that all recent tetrapods derived from one of these lineages leading to amphibians (as *Xenopus tropicalis*) and amniotes (as *Homo sapiens* and *Gallus gallus*).

The other group within the osteichthyes are the ray-finned fishes actinopterygii (ray-finned fishes). Their fins have no specific skeletal elements in common with the tetrapod limb. The characteristic fin rays belong to the dermal skeleton and do not have homologous bones in tetrapod limbs. Derived teleost fishes as *Danio rerio* and *Takifugu rubripes*, are recent living species of this group. One of the most basal teleosts is *Polypterus senegalus*.

Morphologically, the origin of the tetrapode limb is considered to be coincidental with the origin of the autopodium (Fig. 5.2), which denotes distinct hand and feet in the paired appendages. The most distal part of the autopodium (the acropodium) arose from new elements rather than transformation of distal fin sceleton and is seperated from the zeugopodium by one or two rows of small and

Figure 5.1: One sunny afternoon in the Devon.

most often nodular elements (mesopodium). The critical developmental change underlying this morphological innovation is the origin of a genetic mechanism responsible for determining the autopodial field. One hypothesis for the basis of such a mechanism was estabished by *Wagner et al.* [76]. They observed that the expression domains of *hoxA11* and *hoxA13* are mutually exclusive in mouse and chicken but the expression domains of *hoxA11* and *hoxA13* orthologs of teleost fishes are overlapping in the fin development. Therefore, they assume that the ancestral state is one in which *hoxA11* and *hoxA13* have overlapping expression domains while in the derived state the genes have a locally exclusive expression domain determining the limit between the developing zeugopodium and autopodium.

5.2 Footprint Detection

To map the expression pattern of *hoxA11* and *hoxA13* to changes of regulatory elements, we looked for footprints in the vicinity of these genes. Therefore, we applied `tracker` to a set of sequences spanning the IGR between *hoxA13-hoxA11* from animals with fins (*Heterodontus francisci, Polypterus senegalus, Danio rerio, Takifugu rubripes* and *Latimeria menadoensis*), animals with limbs (*Homo sapiens* and *Gallus gallus*) and *Branchiostoma* (also called amphioxus), a cephalochordate, which is used as outgroup species to vertebrates. The sequences

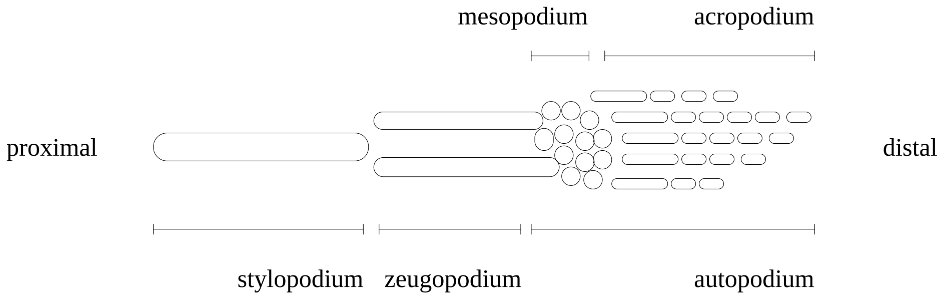


Figure 5.2: The three main segments of a tetrapod limb: stylopodium, zeugopodium and autopodium. The latter is built up by the mesopodium and the acropodium. The only consistant morphological differences between fins and limbs is the seperation of acropodial elements and the zeugopodium by one or more rows of mesopodial elements.

are listed in Table 5.1.

Clusters with interesting distribution among species where than analysed with TFsearch [37] to look at the fine scale distribution of binding sites within the clusters.

5.3 The Tracks of Tetrapods

The search for footprints using `tracker` yields 32 cliques with standard parameter settings. Eight of these footprints exclusively occure in human and chicken whereas not even one footprint is present in all of the sequences (see Table 5.2). The two largest cliques and the 8 human/chicken specific footprints are depict in Fig. 5.3.

The largest cliques concerning length and distribution among species are clique 12 and 31. Clique 12 is about 170nt long and lies 2000 - 4000nt upstream of vertebral *hoxA11* genes. It is worth noting that in case of duplicated *HoxA* clusters (as in zebrafish and fugu) only one of the duplicated clusters retained this footprint. Clique 31 is about 200nt in length and represents the proximal promotor of all vertebral *hoxA11* genes examined. These conserved footprints must have been important in regulation since the origin of vertebrates.

1000nt upstream of *hoxA11* `tracker` detects a region spanning 3000nt and 8 footprints conserved between human an chicken only. We propse that these are crucial for autopodium formation and may mediate the exclusive expression domains of *hoxA13* and *hoxA11*. With limb formation and the encounter of land some meachanical difficulties arose mediating the evolution of adaptive solutions concerning the whole body plan. Therefore, terrestrial locomotion coevolved not only with the limb formation but also with the pectoral and pelvic girdles which

Table 5.1: Source and length of the *hoxA13-hoxA11* non-coding region of *Heterodontus francisci* (Hf), *Homo sapiens* (Hs),*Gallus gallus* (Gg), *Polypterus senegalus* (Ps), *Danio rerio* cluster Aa (Da), *Danio rerio* cluster Ab (Db), *Takifugu rubripes* cluster Aa (Fa), *Takifugu rubripes* cluster Ab (Fb), *Branchiostoma floridae* (Bf) and *Latimeria menadoensis* a coelacanth (Co). The position of the genes were either taken from the annotation at genbank or tblastn searches of known *hox* proteins against the cluster sequence. rc = reverse complement.

hoxA13-hoxA11		
organism	length	source
Hf	12395	AF479755 (as in *Chiu et al. 2002* [21])
Hs	13057	AC004080rc+AC010990rc(overlaps 200nt with flanking fragments)+AC004079[75001-end] (as in *Chiu et al. 2002* [21])
Ps	10128	AC126321rc+AC132195 (overlapping 4307nt)
Da	6647	AC107365rc
Db	7756	
Fa	8009	Fugu v.3.0 scaffold_47[103001-223000]rc
Fb	5053	
Co	15184	kindly provided by *Chris Amemiya*
Bf	23261	L27515_L27516_23sept02 Hox10-14 kindly provided by *Chris Amemiya*
Gg	8471	AF327372 as in *Bodenmiller et al. 2002* [15]

support the spine at two major points along the axis. These changes could be viewed as secondary effects of limb formation that would also be regulated by human/chicken specific footprints.

In the attempt to assign the 8 footprints to the fin limb transition the following problem arises: both, human and chicken, are amniotes. Therefore, the regulatory sites conserved in these two tetrapod species may be amniote specific. To rule out this possibility one could include an amphibian (e.g. *Xenopus tropicalis*) or primitive tetrapod into the set of sequences. Appearance of human/chicken specific footprints in amphibian sequences would support our hypothesis that these are limb specific.

5.4 Zooming into Highly Conserved Cliques

Conserved cliques reported by `tracker` are thought to be clusters of functional protein binding sites. To reveal the detailed composition of these cliques, we take a closer look and apply `TFsearch` to the aligned sequence fragments. Table 5.4

Table 5.2: Footprint distribution summary table.

Bf	Hf	Co	Gg	Hs	Ps	Da	Fa	Db	Fb	Parsimony score	n	Clusters
−	+	+	+	+	+	+	+	+	+	1	2	31 32
−	+	+	+	+	+	−	+	+	−	3	1	12
−	+	−	−	−	−	−	−	+	−	2	1	6
+	−	−	−	−	−	−	−	+	−	2	1	16
−	+	+	−	−	−	−	+	−	−	3	1	10
−	−	+	−	−	−	−	+	−	−	2	1	8
+	−	−	−	−	−	−	+	−	−	2	2	15 22
−	+	+	−	+	+	+	−	−	−	4	1	5
−	−	−	−	+	−	+	−	−	−	2	1	30
−	+	−	−	−	−	+	−	−	−	2	1	7
+	−	−	−	−	−	+	−	−	−	2	2	19 20
−	−	+	−	−	+	−	−	−	−	2	1	9
+	−	−	−	−	+	−	−	−	−	2	1	14
−	−	−	+	+	−	−	−	−	−	1	8	11 23 24 25 26 27 28 29
−	−	+	−	+	−	−	−	−	−	2	1	1
+	+	−	−	+	−	−	−	−	−	2	1	21
+	−	−	−	+	−	−	−	−	−	2	1	18
−	+	−	+	−	−	−	−	−	−	2	1	4
+	−	+	−	−	−	−	−	−	−	2	3	2 3 17
+	+	−	−	−	−	−	−	−	−	1	1	13

and 5.5 summarize all possible binding sites, most of which are supported by cooccurence at the same position in more than one sequence. The high density of detected motifs including multiple CdxA binding sites and the congruent structural organization of a footprint such as clique 28 or 29 argues for a functional cluster of binding motifs. On the other hand, we also find conserved sequences, each composed of diffenrent binding sites. These observation may be interpreted as artifact of the `TFsearch` method or rather recent destruction of functional binding sites in one of the sequences (clique 27).

The distribuion of sites within the proximal promotor shows that two sites, USF/SREBP and CREB, are conserved among all sequences. Others may be missing in some of them (C/EBP site downstream of CREB) or slightly shifted

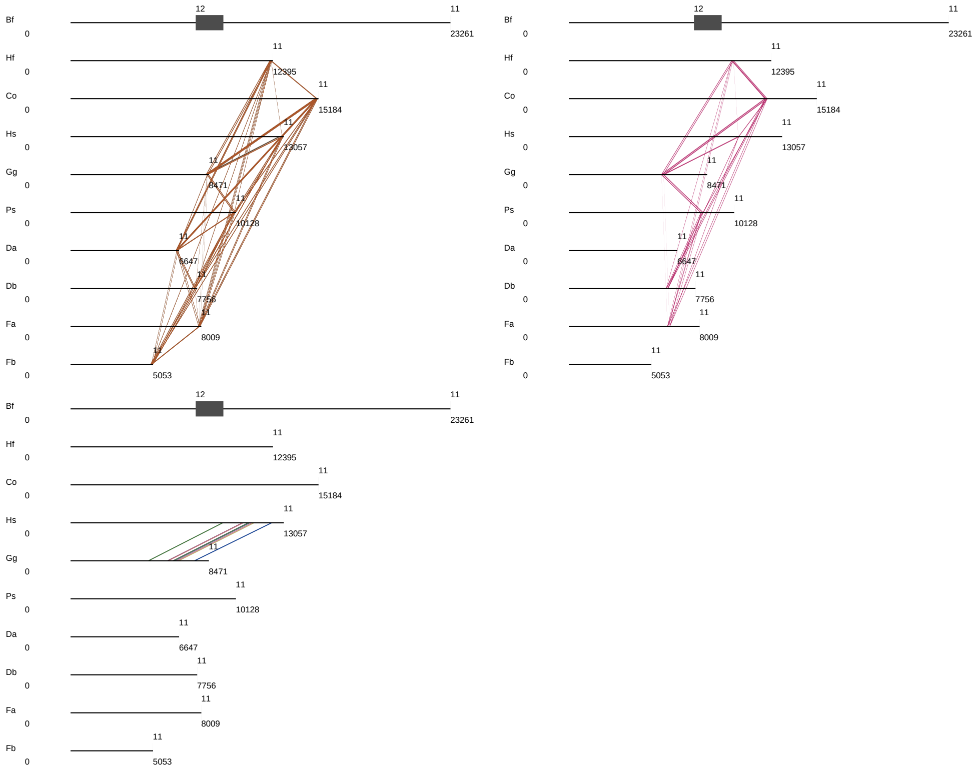


Figure 5.3: Location and pairwise similarities of clique 31, clique 12 and the eight human/chicken specific cliques 11 and 23 - 29. All sequences exept the amphioxus sequence are part of clique 31. Clique 31 is located immediately upstream of *hoxA11* and termed 'proximal promotor'. Clique 12 is present in Hf, Co, Hs, Gg, Ps, Db and Fa about 3000nt upstream of *hoxA11*. The eight footprints of Hs and Gg, spanning 3000nt are all colinear and at comparable distances in both species. Non-horizontal lines indicate pairwise similarity. Abbrevations refere to Table 5.1.

(GATA). One the level of single binding sites one could also find patterns, which destinguish tetrapods from other animals. For instance, the non-orthologous MZF1 sits between the two overall conserved motifs just occures in the human and chicken sequence, same as the second SRY site downstream of SP1 in clique 12.

We conclude that even though the sequences are highly similar they may be composed of different sites. Fine scale analysis may reveal lineage specific changes with potential regulatory effects comparable to whole clusters. Therefore, destruction of phylogenetic footprints reported by `tracker` using the information of concrete binding sites may enhance the sensitivity of predicting regulatory changes that cause major transitions in animal evolution.

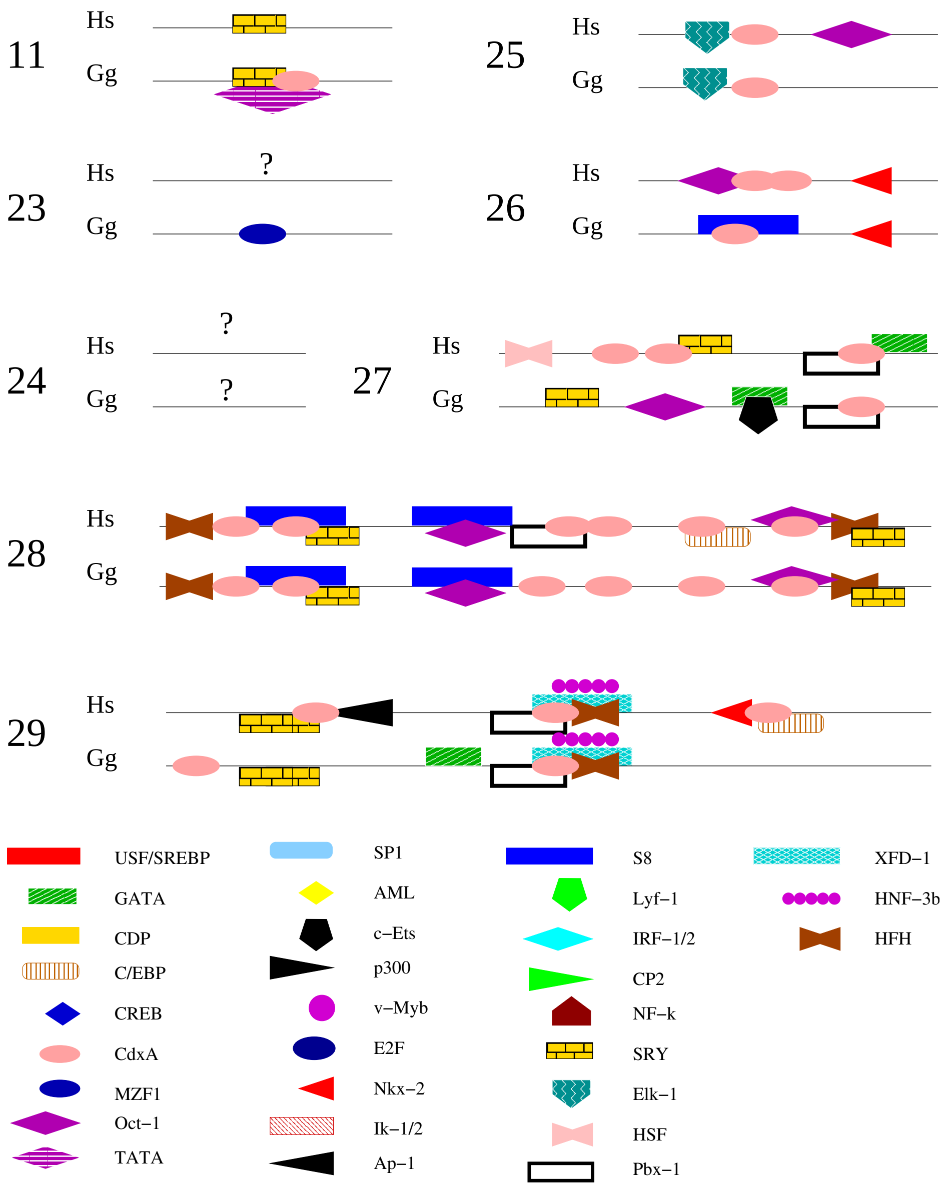


Figure 5.4: Transcription factor binding sites reported by TFsearch for the eigth tetrapod specific cliques 11 and 23 to 29 (tracker RunID = 04031410CLFR). Question marks indicate potential, unknown binding sites in highly conserved regions. Hs = Homo sapiens and Gg = Gallus gallus. Sequences and sites are roughly drawn to scale.

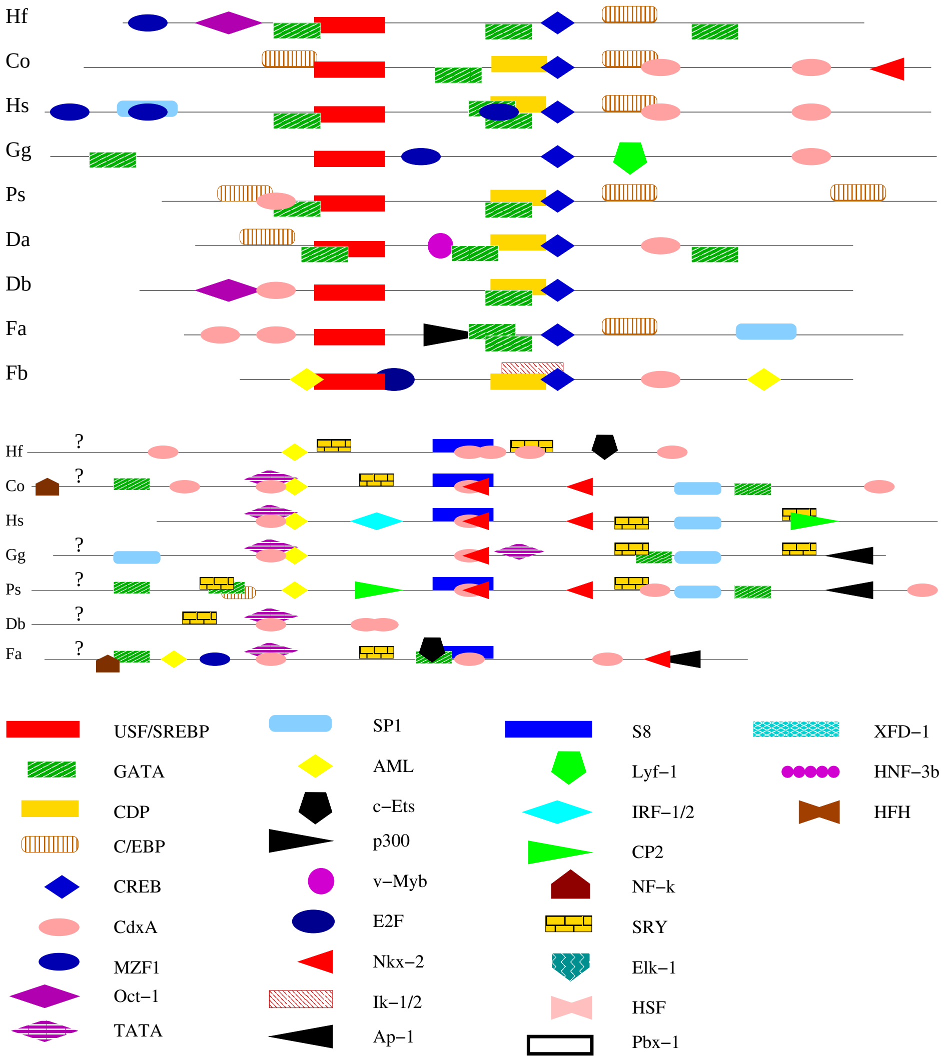


Figure 5.5: Transcription factor binding sites reported by TFsearch on clique 31 and clique 12 (tracker RunID = 04031410CLFR).Question marks indicate potential, unknown binding sites in highly conserved regions. Hf = Heterodontus francisci, Co = Latimeria menadoensis (colacanth), Hs = Homo sapiens,Gg = Gallus gallus, Ps = Polypterus senegalus, Da = Danio rerio Custer Aa, Db = Danio rerio Cluster Ab, Fa = Takifugu rubripes ClusterAa and Fb = Takifugu rubripes ClusterAb. Sequences and sites are roughly drawn to scale.

CHAPTER 6

Conclusion

In this work we have presented the novel `tracker` method for phylogenetic footprinting that is able to handle a large set of long sequences without great resource consumption. The tool is fast and runs without user intervention. It is successful in detecting footprints conserved only in a subset of sequences without relying on phylogenetic assumptions. The various outputs comprise overview pictures, a detailed list of footprints, all local multiple alignments and their distribution among the sequences. We have shown that currently it is the only suitable program to extract large amounts of quantitative data on non-coding sequence information that can be passed on to statistical analysis and the structural loss model (section 3). Furthermore, we have demonstrated that it can be used to find orthologous clusters when comparison of coding regions does not lead to an unique solution (section 4). Its ability to detect footprints with a certain distribution among input sequences can be utilized to find taxa specific footprints, which – in case of *hox* genes – may indicate evolutionary important transitions (section 5).

CHAPTER 7

Outlook

7.1 Detecting Protein Binding Sites

The obvious step following the detection of conserved regions, which usually span about 100nt and several dozens of individual protein binding sites, would be to mark concrete footprints or protein binding sites as it was done via webtools in section 5. Therefore, known motifs must be assigned to the sequences. One big database with a collection of various protein binding sites is realized in TRANSFAC. This database attempts to reach good completeness in detecting motifs by the costs of also reporting a lot of artifacts. Anyway, connecting TRANSFAC to tracker is a question of cost.

If we assume that the problem of detecting functional regulatory elements is solved, the next step would be to reconstruct the regulatory network. A lot of work will have to be done to reach this final goal that again will leave a lot of work to the hardcore biologists and the interpretation of the outcome.

7.2 Biological Challenges Lying Ahead

7.2.1 Distances of Footprints

The establishment of certain gene expression patterns is determined by the coaction of a set of regulatory regions each of which controls specific subfunctions. These regions can carry out there subfunction independently and are therefore viewed as the elementar motifs of gene regulation. Experimental evidence suggests that order, direction and distance of regulatory regions may be important. Currently, it is not examined if relative or absolute distances have to be maintained for full functionality. With our method `tracker` it is possible to locate regulatory regions onto sequences of related organisms which may have varied distances of footprints over time. With the ability to observe the changes in footprint distribution along the sequence, biological questions may be answered (Fig. 7.1). Is it the absolute or relative distance between genes and sites that is important? Is it possible to find the regulated gene by means of distance conservation in related species? Can shared enhancers be identified because of a characteristic distance patterns?

7.2.2 The Importance of Colinearity

The untouched organization of *Hox* clusters suggests a mechanism for maintaining cluster integrity. The driving force is spatial and temporal colinearity, linking gene regulation to the gene order within the cluster. But what is the mechanism of spatial and temporal colinearity? Are the mechanisms for spatial an temporal colinearity independent?

Cluster Destruction

Since the importance of colinearity and cluster integrity is a rule there are also exceptions to it. The two prominent examples are the cluster of *Drosophila melanogaster* which is split into the Antennapedia and Bithorax complexes and the broken 'cluster' of *Caenorhabditis elegans* that comprises three groups. Got anything lost together with the cluster organization? Ferrier *et al.* [29] postulated that in these destructed clusters temporal colinearity is lost. The effect, however, is not very dramatic in animals with a brief period for assignment of segmental identity. Therefore, the absence of the temporal aspect to colinearity may lead to relaxation of a selective constraint on the organization of the *Hox* cluster. But is there an influence on spatial colinearity?

One suggestion for a mechanism regulating spatial colinearity is enhancer sharing coupled with quantitative colinearity [45]. Since cluster destruction would destroy this organization it would be interesting to compare the regulatory region of homologous *hox* genes from destructed and compact clusters. One expectation would be to find a previously shared enhancer duplicated at defined distances to all cluster fragments.

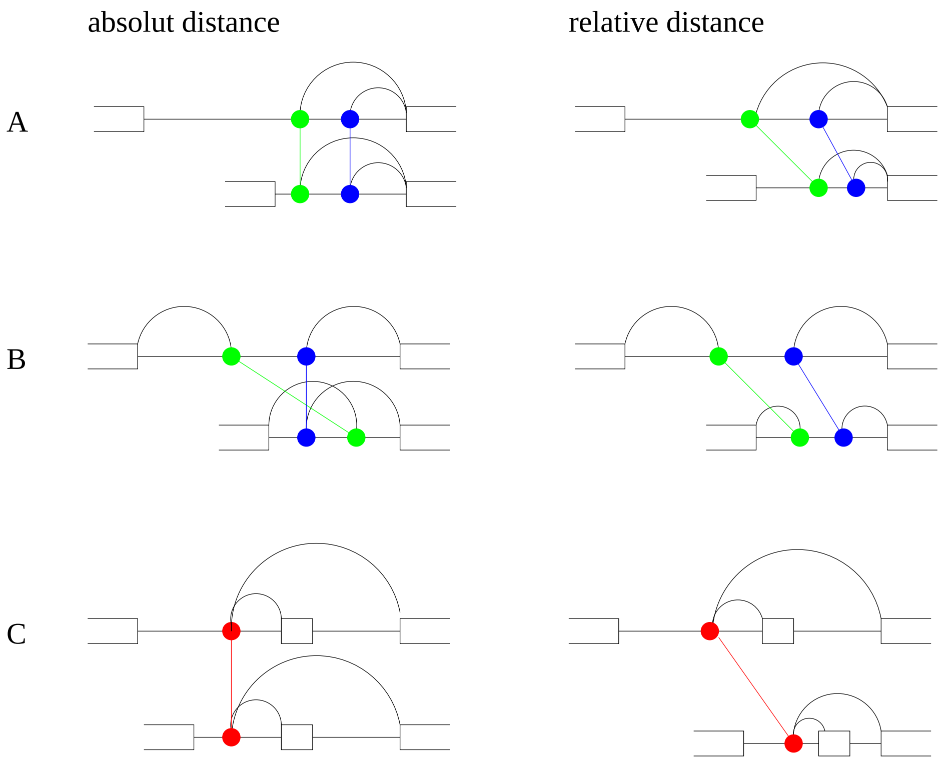


Figure 7.1: Changes in footprint distribution after shrinkage of intergenic regions. In situation A, both footprints regulate the downstream gene. In the left picture, the absolute distance to the regulated gene is maintained. Length reduction of the intergenic region occurs upstream of the footprints until the first footprint is next to the previous gene. The right picture depicts the situation of maintained distance relations. Situation B assumes that the footprints regulate different neighboring genes. The left picture shows a hardly plausible evolutionary scenario in which inversion of footprints would be necessary to maintain proper absolute distances. For the case that this does not occur the footprints would define a minimal distance of neighboring genes. Situation C assumes enhancer sharing. In the left picture the distance of the coregulated genes is fixed. Blue, green and red balls denote footprints. Arcs point at the regulated genes which are illustrated as open boxes.

Transposable Elements

The strong dependency of accurate development on the organization of the cluster causes negative selective pressure on unstable genetic elements. Since repetitive and transposable elements are thought to be one route by which genome rearrangement can occur, these elements are usually absent from vertebrate *Hox* clusters. A challenge would be to significantly prove that transposable elements are underrepresented in *Hox* clusters over hundreds of million years (Fig. 7.2).

7.2.3 Major Evolutionary Transitions

In section 5 we related a major evolutionary transition, the fin limb transition, to changes in regulatory regions upstream of *hoxA11*. Maybe a set of footprints can be defined to be sufficient to explain the origin of tetrapods. The arising question is: do other major innovations in animal body plans also map on *hox* genes and/or their regulatory sequences?

It seems possible to further assign regulatory changes to the origin of major phylogenetic groups. The origin of vertebrates an the process of cluster duplication are of special interest. But other transitions, such as the origin of mammals and one of its major characters 'hair', could be an interesting field of research. Where does hair come from? As recently investigated *hox* genes as *hoxC13* play an universal role in hair follicle development. Therefore, it is believed that subsequent to the role of specifying positional identities along the body axis subsets of *hox* genes have been co-opted for patterning functions in phylogenetically more recent structures.

However, for all these purposes whole *Hox* clusters of species phylogenetically placed around the transition of interest have to be examined first.

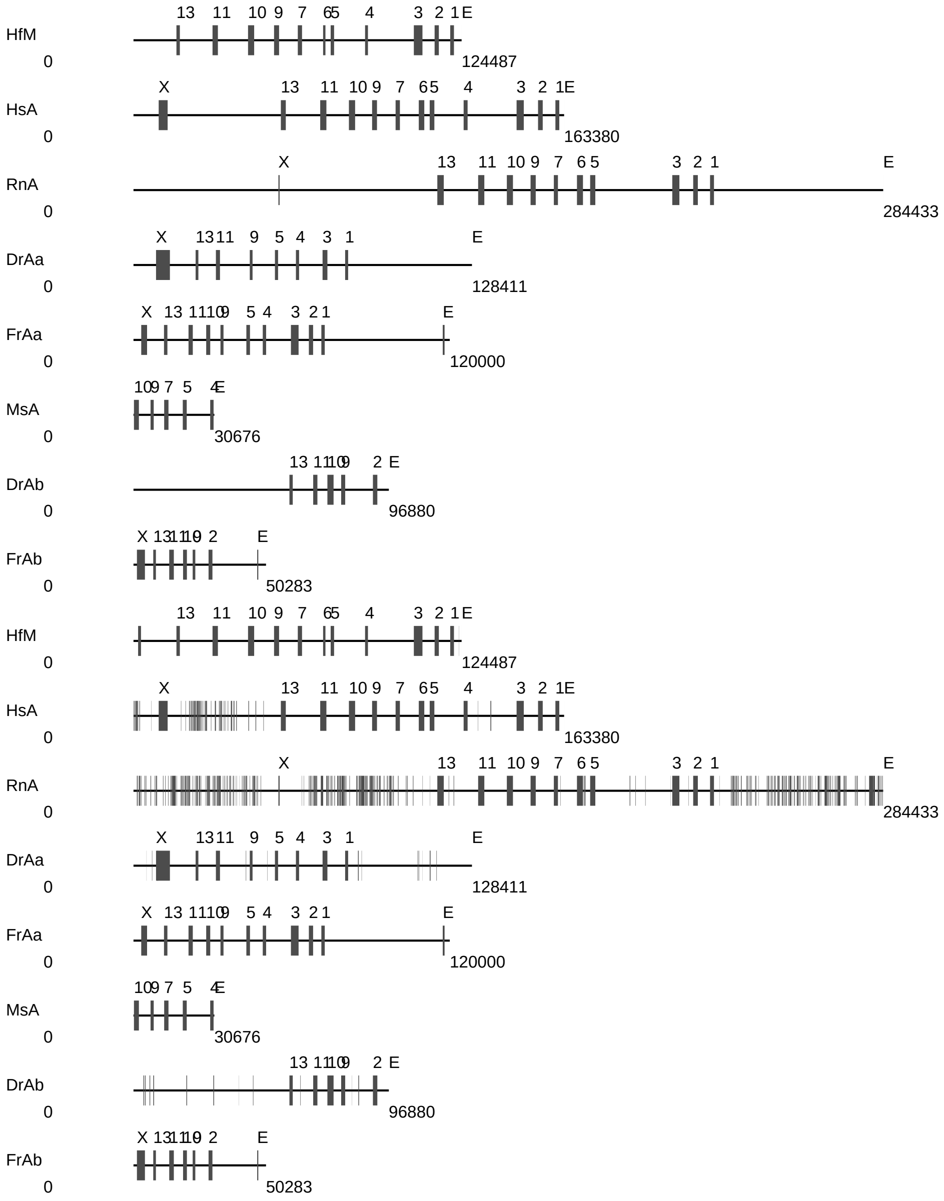


Figure 7.2: Location of repetitive and transposable elements detected by `RepeatMasker`. The left panel shows the *HoxA* clusters with their genes as gray boxes. The numbers above these boxes indicate the paralogous group. In the right panel, repeat regions are added (gray boxes without numbering) Three of the eight sequences do not show a single repeat (FrAa, MsA and FrAb).

Appendix

Reevaluation of Chiu *et al.* [21]

Chiu *et al.* [21] performed a search for phylogenetic fooprint clusters (PFCs) on the *HoxA* cluster of 4 members of the three major gnathostome lineages: *Heterodontus francisci* (HfM), *Homo sapiens* (HsA), *Danio rerio* cluster *Aa* (DrAa), *Danio rerio* cluster *Ab* (DrAb) and *Morone saxatilis* (MsA). Using `PipMaker`, `ClustalW` and `BayesAligner` they identified 36 PFCs. A reevaluation of these data using our method `tracker` resulted in detection of 148 cliques. We further extended our search onto the two *HoxA* clusters of *Takifugu rubripes* (TrAa, TrAb). For a detailed description see chapter 3.

Table 7.1: Comparison with [21]. The last column gives the designation of the "phylogenetic footprint clusters" (PFCs) from [21]. Footprints that were not found by `tracker` are listed in *italics* without numbering, + denotes novel ones. +XXX means that we found footprint also in XXX; analogously, -XXX means that the footprint was detected in the *Hox* cluster XXX in the previous study [21] but was not found in this sequence by the `tracker` program with the default parameter setting. Positions of footprints that are missing in some sequences in the `tracker` output are given in parentheses. Differences between the published position numbers of the DrAa sequence and our data are explained by the use of two versions of the DrAa sequence in [21].

Footprint	HfM		HsA		DrAa		DrAb		MsA	PFCs
1	865	23					1553	23		+
2	2891	51					39450	51		+
3			8197	31			33525	31		+
4					2283	76	7560	76		+
5					2287	70	6101	70		+
6					3246	62	29283	60		+
7					7147	46	32044	46		+
8			13150	59	15129	59				+
9			13216	30	15198	31				+
10			13258	12	15241	12				+
11			15102	41			7692	47		+
12	3734	81	20391	84						+
13	3881	23			4203	23				+
14			25741	38			15607	33		+
15			27295	29			35475	29		+
16	5901	75	28483	75						+
17	5949	23	4134	23						+
18	6483	120	45120	121						upstream of 13-a
19	6775	40	45433	37						upstream of 13-b
	8558	*40*			*21743*	*19*				*upstream of 13-c*
20	11868	26					47489	26		+
21					16307	78	22716	78		+
22					18316	42	29824	42		+
23					18387	55	29928	55		+
24	13192	120	53810	88	22652	65	58295	121		upstream of 13-d
25	13360	13					58469	14		+
26	16127	49					58996	48		13pp -DrAa +DrAb
27	16233	57					59103	56		13pp -DrAa +DrAb
28	19133	112	59505	114	25574	24				13-11-a +DrAa
29	20828	47					63519	47		+
30	27207	32			28565	32				+
31	27545	30					66363	30		+
32	27606	116	68103	118						+
33			70181	58	28402	58				+
34					29483	35	67002	35		+
35	29781	168	70665	152	31068	118	67981	132		13-11pp
36					33896	155	71142	153		11-9-a DrAa(29667)
37	34076	42			43022	42				+
38	34423	77	75337	78						11-10-a
			76034	*9*			*71322*	*9*		*11-10-b*
39	35043	77	76069	52	34212	31	71442	74		11-10-c +DrA
40	41272	55					71853	55		+
41			78189	21	32835	21				+
42	43143	93	81631	94			73488	75		+
43	46400	43	85314	39						10-9-a

Table 7.1 continued.

Footprint	HfM		HsA		DrAa		DrAb		MsA		PFCs
44	46546	24	85435	24							10-9-a
45	46591	188	85479	187					2977	139	10-9-a
46	47542	116	86410	116	41556	97	76755	93	3393	97	10-9-b DrAa(37297)
47							76892	16	3556	16	+
48	48333	30	87347	38	41872	35	77048	44	3791	49	10-9-c +HsA +DrAa +MsA
49	52969	35	90122	35							10-9-d
50	53030	45	90215	44							+
51	53084	55	90267	55							10-9pp -MsA(6219)
52	53229	28	90412	28							10-9pp -MsA
53	53264	42	90452	41							10-9pp -MsA
54					43987	63			6298	64	+
55					45766	47			8387	46	+
56							77140	16	3893	16	+
57							77166	94	3929	96	+
58	56953	99	94192	61	46679	175	81365	81	8912	182	9-7-a +DrAa +Drab +MsA
59	57228	219	94465	223	47016	208			9511	229	9-7-b +DrA + MsA
60	57682	31	94836	31							+
61	59503	39					87245	36			+
62			97345	38					9394	38	+
63	62154	12	99257	12							9-7-pp
64	62176	33	99279	32					11485	29	9-7-pp
65	62226	107	99327	107	48807	54			11530	104	9-7-pp
66					49660	26	88070	26			+
67	66439	203	103206	206	49942	219			14805	164	7-6-a +DrAa
68	66923	24	103654	24							+
69	71720	40	108022	41							7-6-pp
70	71778	148	108078	147					16637	28	7-6-pp
71	74400	27	111988	27							+
72					53087	33			18217	34	+
73	74469	34	112053	26	53164	31			18300	31	+
74	74519	268	112101	265	53250	229			18389	228	6-5-pp
75	76119	11	114171	11							5-4-a HfM(76427)
76	76145	22	114197	22							5-4-a HfM(76427)
77	76181	22	114231	22							5-4-a HfM(76427)
78	76215	38	114264	37							5-4-a HfM(76427)
79	76266	25	114314	26							5-4-a HfM(76427)
80	76323	69	114356	70							5-4-a HfM(76427)
	76648	*63*	*114717*	*77*							*5-4-b*
81	76784	44	114894	44							+
82	77565	326	115543	323	55930	245			21536	250	5-4-c
83	78818	52	116743	54							+
84	79794	29					83629	29			+
85					56180	25			21789	23	+
86					56277	12			21873	12	+
87	81947	71	119346	105	57520	105			23483	104	5-4-d +DrAa
88	82035	16							23604	16	+
89	82436	286	119799	284	57972	163			24139	180	5-4-e +DrAa
90	82749	16	120098	15							+
91					58177	68			24365	70	+
92	84826	231	121990	231	59802	175			27247	180	5-4-f +MsA
93			122238	27			86591	27			+
94	85596	41	122775	40			88770	23			+
95	85651	41	122822	41							5-4-g
96	85787	19					85007	19			+
97	85814	29					85029	31			+

Table 7.1 continued.

Footprint	HfM		HsA		DrAa		DrAb		MsA		PFCs
98	87745	114	125173	76	61442	176			28922	183	+
99	91064	132	128822	129							4-3-a
100	91515	58	129461	58							+
101	91602	30	129556	30							+
102	92853	91	131248	89							+
103	93227	73	131592	77							+
104	93311	42	131680	42							+
105	93372	81	131766	83							+
106	94873	34					88361	34			+
107	98246	55	136897	58							+
108	98424	35	137066	37							+
109	98476	62	137119	58							+
110	98868	148	137526	147							+
111					65895	19	87490	19			+
112	99108	85	137815	82	67086	81					+
113	99764	29					89449	29			+
114	101931	276	140542	277							+
115	102590	50			69681	56					+
116	102694	27	141968	27							+
117	102966	86	142331	46	70109	86					4-3-b
118	103058	129	142393	129	70220	50					+
119	105041	154	144063	157							+
120	105199	33	144236	32							+
121	106120	92	145095	94	71542	39					4-3-pp +HsA
122	106233	124	145205	135	71593	132					4-3-pp +HsA
123	109890	95	148351	96							+
124	109999	217	148482	218							+
125			151198	30			89631	28			+
126					73712	35	87719	35			+
127	112888	123	151235	121	75190	114	89669	117			+
128	113671	123	152783	127							3-2-a
129	113939	243	153130	247			90535	218			3-2-pp
130	116088	86	155551	83							+
131	116229	30	155683	30							+
132	116301	11	155747	11							+
133	117348	99	156872	100							2-1-a
134	117460	78	156985	79							2-1-a
135	119953	54	159818	54							+
136	120009	44	159883	44							+
137	120063	69	159973	72							+
138			161549	39	92267	39					+
139	121736	18	161979	16							+
140	121808	11	162032	11							+
141	121838	56	162050	57							+
142	122218	85	162406	90							+
143	122334	39	162528	39							+
144	122397	12	162592	12							+
145	122423	25	162618	23							+
146	122483	17	162663	17							+
147			162790	27	113979	27					+
148	122765	79	162923	79							+

Table 7.2: Footprints in *Takifugu rubripes*.
Data from Table 7.1 that do not involve a *Takifugu rubripes* match are not listed. Clusters that are separated into more than one entry are sometimes merged into a single cluster here. Cluster numbers in brackets refer to Table 7.1.

#	HfM		HsA		DrAa		DrAb		MsA		TrAa		TrAb		Difference
2	1089	36									7484	36			
3	2059	22									10359	22			
4							22925	46			99	46			
5							29449	21			6078	21			
6							33702	74					315	74	
8			4617	47							926	48			
9			4671	26							980	23			
10			4707	88							1013	88			
11			4838	24							1147	24			
16					7143	50	32044	46			2898	22			+TrAa [7]
29							45312	27			6637	27			
30							46640	28			1522	28			
31	11868	70	6300	91			47489	26			1808	91			+TrAa [20]
32	13165	11									10614	11			
34					17215	29					6153	29			
37							54090	84					5381	84	
38	13185	127	53810	88	22603	114	58295	121			10639	95	6656	93	+TrAa +TrAb [14]
40	16127	163			23490	69	58985	174			11378	183	7315	176	+TrAa +TrAb [26,27]
43					27080	34					14008	34			
45											14820	39	8813	39	
46	27545	30					66363	30			18891	21			+TrAa [31]
47	27606	116	68103	146							18970	141			+TrAa [32]
49					29386	61					18580	56			
51	29781	168	70665	152	31057	129	67981	132			20662	129	13385	120	+TrAa +TrAb [35]
52	33041	93									23147	89			
53					31192	27	68131	21			20795	26	13521	21	!!
54					33813	39					24159	40			
55					33862	12					24213	12			
56					33891	160	71142	176			24243	190	16517	163	+TrAa +TrAb [36]
57					37209	54					25259	57			
58					42263	64					25886	64			
61							71333	59			24477	11	16682	59	
62	35037	84	76069	52	34209	58	71441	75			24565	49	16773	78	+TrAa +TrAb [39]
64	41390	47									25206	46			
66							73382	17					18624	17	
67	43095	326	81612	48			73404	161	2	170	27418	388	18661	143	+TrAa +TrAb +MsA [42]
70									2110	96	29223	97			
71									2298	28	29389	28			
72									2340	13	29422	13			
73									2436	14	29482	14			
74									2464	17	29505	17			
75									2492	50	29536	56			
76									2581	46	29630	51			
77									2644	21	29698	20			
78									2672	93	29719	93			
79	46591	188	85479	187	41286	50			2946	174	29966	175			+TrAa +DrAa [45]
80									3139	59	30165	59			
81									3210	66	30238	67			

Table 7.2 continued.

#	HfM		HsA		DrAa		DrAb		MsA		TrAa		TrAb		Difference
82									3313	20	30336	19			
83	47542	116	86411	116	41556	97	76755	95	3348	155	30366	155			+TrAa [46]
84							76892	16	3556	112	30587	92			+TrAa [47]
85									3707	22	30716	20	21531	10	
86	48333	116	87347	49	41872	35	77048	241	3742	349	30746	346	21592	212	+TrAa +TrAb [48,56,57]
87	50073	49							4812	172	31572	169			!!
93					43707	68					32112	62			
94							78511	32					22196	31	
95									5901	138	32451	133			
96									6051	15	32596	15			
97									6076	56	32626	55			
98					43987	68	78594	75	6154	222	32692	220	22274	80	+TrAa +TrAb +DrAb [54]
99									6417	51	32953	41			
100									7720	180	34183	178			
101									7913	12	34366	12			
102									7947	29	34398	29			
103									7987	46	34438	45			
104									8086	44	34534	44			
105									8154	62	34584	61			
106									8226	60	34648	57			
107					45766	47			8296	223	34717	230			+TrAa [55]
108									8534	17	34965	16			
109	56941	111	94192	62	46679	175	81365	83	8888	225	35174	225			+TrAa [58]
110									9221	11	35441	11			
111									9284	56	35493	54			
112	57228	215	97346	38	47011	213			9359	537	35554	531			!! [62,59]
113	57228	219	94466	223	47011	213			9359	537	35554	531			+TrAa [59]
115							82326	30					24929	30	
116							84706	30			62877	30			
118	59598	95									36922	95			
119			99196	28									26430	28	
121									10120	16	36280	16			
122									10199	67	36353	68			
123	62176	159	99280	157	48807	54			11415	222	37137	223			+TrAa +DrAa [64]
125									14518	34	39351	34			
126	66439	203	103206	206	49926	235			14790	215	39476	298			+TrAa [67]
127	66439	203	103206	206	49926	235			14565	97	39395	343			!! MsA(new)
130									15018	14	39785	14			
131									15098	194	39857	186			
132									15319	22	40077	22			
133									15700	67	40338	65			
134									16398	25	40811	25			
135	71778	148	108078	147					16526	139	40909	133			+TrAa [70]
137					52101	37							27738	37	
138									16856	52	41246	45			
139									16953	70	41310	67			
140									17826	70	41962	65			
141									18045	145	42174	144			
142					53081	39			18217	37	42347	43			+TrAa [72]
147	74469	318	112053	313	53164	316			18269	366	42403	356			+TrAa [73]
152	77565	326	115543	323	55930	245			21536	250	45370	239			+TrAa [82]

Table 7.2 continued.

#	HfM		HsA		DrAa		DrAb		MsA		TrAa		TrAb		Difference
155			117477	34									23956	34	
156					56180	25			21789	23	45612	14			+TrAa [85]
158	81947	71	119346	105	57520	105			23483	104	47002	59			+TrAa [87]
163	84826	231	121990	231	59797	180			27151	298	47302	295			+TrAa [92]
169									27487	42	47629	42			
170									27533	150	47679	146			
171									28379	34	48327	39			
172									28479	37	48460	38			
173									28648	44	48614	37			
174									28709	30	48665	27			
175									28787	33	48728	29			
176	87745	114	125173	76	61442	176			28831	274	48768	272			+TrAa [98]
188					63246	21					50977	21			
190					65919	45					54155	47			
191	98868	148	137523	150	66768	90					54893	95			+TrAa +DrAa [110]
192					66882	24					55005	27			
193					67013	43					55144	42			
194	99108	131	137815	83	67086	81					55209	128			+TrAa [112]
196	101851	29									56844	29			
197					67923	141					55758	146			
198	101931	276	140542	277	69137	65					56932	85			+TrAa +DrAa [114]
199	102585	66			69676	74					57720	75			+TrAa [115]
201	102762	22									57907	20			
202	102960	227	142331	191	70088	200					58119	207			+TrAa [117,118]
203	105041	191	144063	205	70908	76					59043	164	25595	41	+TrAa +DrAa [119,120]
204	106120	237	145095	245	71522	205					59521	204			+TrAa [121,122]
208					74237	21					63478	21			
209	112888	123	151198	158	75155	165	89629	170			65064	172	27409	144	+TrAa +TrAb [127,128]
211	113939	243	153128	277			90511	242			66175	300	28162	238	+TrAa +TrAb [129]
213	113939	227	120337	29			90511	242			66175	255	28159	113	!! HsA (new)
232	118642	32											35682	32	
233	119948	59	159802	70	79953	29					70981	69			+TrAa +DrAa [135]
234	120009	123	159883	162	80042	58					71066	56			+TrAa +DrAa [136]
235					81903	69					72993	69			
236					83630	36							33838	36	
237											73503	37	30224	37	
238					86121	69					76990	70			
239					86214	25							38195	25	
244	122096	44											41172	44	
247	122397	51	162592	49									38283	30	+TrAa [144,145]
249					101278	32					85496	32			
250											102479	20	37421	20	
251											106652	31	40486	31	
252					102743	35					106732	31	40549	41	
253					107573	26					91180	26			
256					114389	43							41890	43	
257					119410	22							45900	22	
258							94644	29			84123	29			
259							94869	21			70408	21			
260									30397	169	50335	159			
261									30566	105	50500	108			

The Amphioxus Song

(This song is set to the tune of Tipperary.)

A fish-like thing appeared among the Annelids one day.
It hadn't any parapods or setae to display.
It hadn't any eyes or jaws or ventral nervous chord,
But it had a lot of gill slits and it had a notochord!

Chorus:
It's a long way from Amphioxus.
It's a long way to us.
It's a long way from Amphioxus
To the meanest human cuss.
Good-bye fins and gill slits,
Welcome lungs and hair.
It's a long, long way from Amphioxus
But we came from there.

It wasn't much to look at and it scarce knew how to swim.
And Nereis was very sure it didn't spring from him.
The Molluscs wouldn't own it and the Arthropods got sore.
So the poor thing had to burrow in the sand along the shore.

It wriggled in the sand before a crab could nip its tail.
It said, "Gill slits and myotomes are all of no avail.
I've grown some metapleural folds, and sport and oral hood,
But all these fine new characters don't do me any good.

It sulked awhile down in the sand without a bit of pep.
Then it stiffened up its notochord and said, "I'll beat 'em yet.
I've got more possibilities within my slender frame
Than all these proud invertebrates that treat me with such shame.

"My notochords shall grow into a chain of vertebrae.
As fins my metapleural folds shall agitate the sea.
This tiny dorsal nervous tube shall form a mighty brain.
And the vertebrates shall dominate the animal domain."

("Songs of Biology", copyright date 1948)

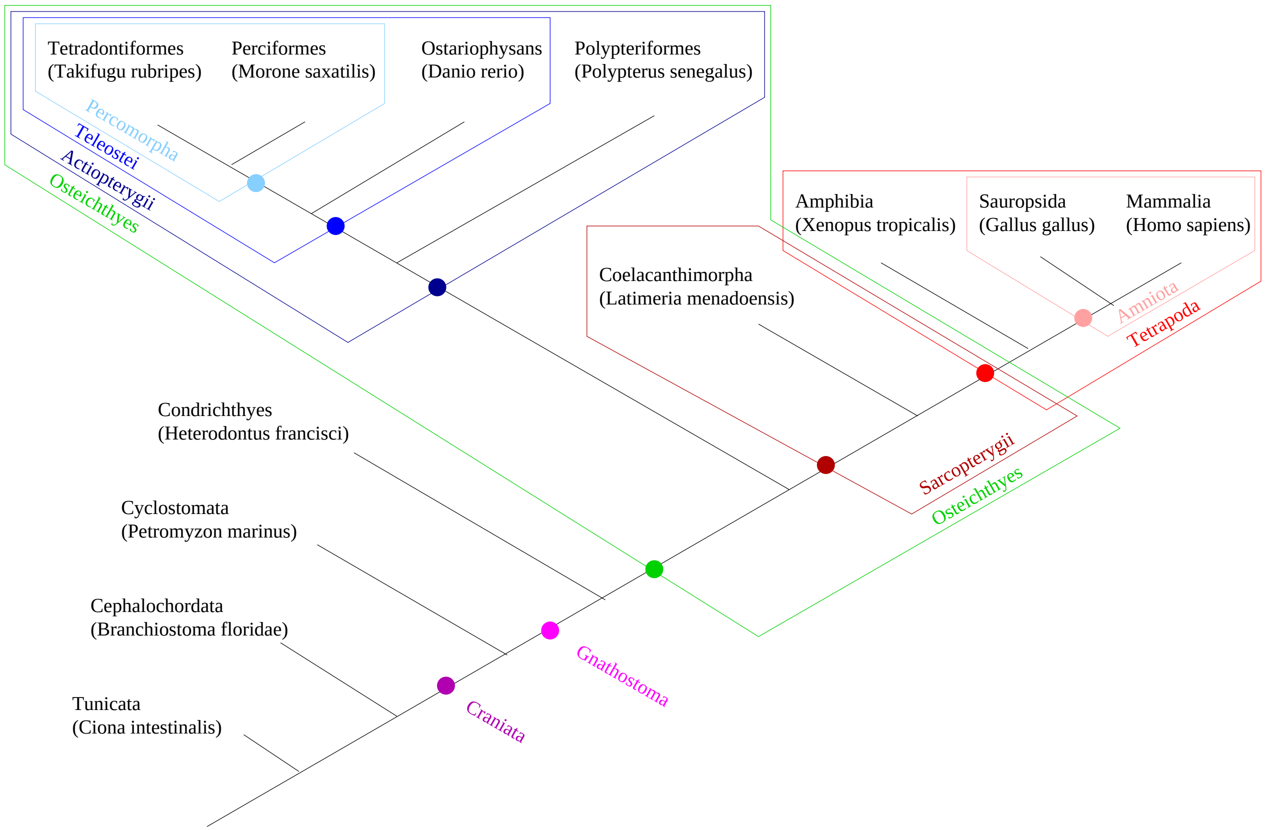


Figure 7.3: Phylogenetic relationships among chordats used in this work. In contrast to some phylogentic classifications (e.g. at the NCBI website) this tree shows that recent tetrapods do not belong to sarcopterygians or bony fish even though the devonian tetrapod has originated from sarcopterygians.

References

[1] Akiyama, Y., 1998. Tfsearch: Searching transcription factor binding sites. Http://www.rwcp.or.jp/papia/.

[2] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

[3] Amores, A., Force, A., Yan, Y. L., Joly, L., Amemiya, C., Fritz, A., Ho, R. K., Langeland, J., Prince, V., Wang, Y. L., Westerfield, M., Ekker, M., Postlethwait, J. H., 1998. Zebrafish hox clusters and vertebrate genome evolution. Science 282, 1711–1714.

[4] Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.-m., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M. D. S., Roach, J., Oh, T., Ho, I. Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S. F., Clark, M. S., Edwards, Y. J. K., Dogget, N., Zharkikh, A., Tavtigian, S. V., Pruss, D., Barstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., H., T. Y., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., Brenner, S., 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297, 1301–1310.

[5] Arnone, M. I., Davidson, E. H., 1997. The hardwiring of development: Organization and function of genomic regulatory systems. Development 124, 1851–1864.

[6] Bailey, W. J., Kim, J., Wagner, G., Ruddle, F. H., 1997. Phylogenetic reconstruction of vertebrate Hox cluster duplications. Mol. Biol. Evol. 14, 843–853.

[7] Bandelt, H. J., Dress, A. W. M., 1992. A canonical decomposition theory for metrics on a finite set. Adv. math. 92, 47.

[8] Bandelt, H.-J., Dress, A. W. M., 1993. A relational approach to split decomposition. In: Opitz, O., Lausen, B., Klar, R. (eds.), Information and Classification, Springer-Verlag, Berlin, pages 123–131.

[9] Benfey, P., Celniker, S., Eisen, M., Small, S., Zhang, M., 2003. Meeting on System Biology: Genomic Approaches to transcriptional regulation. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

[10] Benos, P., Lapedes, A., Stormo, G., 2002. Is there a code for protein-dna recognition? probab(ilistical)ly... Bioessays 24, 466–475.

[11] Bielawski, J. P., Dunn, K. A., Yang, Z.-H., 2000. Rates of nucleotide substitutions and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions. Genetics 156, 1299–1308.

[12] Blanchette, M., Schwikowski, B., Tompa, M., 2002. Algorithms for phylogenetic footprinting. J. Comp. Biol. 9, 211–223.

[13] Blanchette, M., Shina, S., 2001. Seperating real motifs from their artifacts. Bioinformatics 17, 30–38.

[14] Blanchette, M., Tompa, M., 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Research 12, 739–748.

[15] Bodenmiller, D., Baxter, C., Hansen, D., SS, P., 2002. Phylogenetic analysis of hoxa 11 sequences reveals absence of transposable elements, conservation of transcription factor binding sites, and suggests antisense coding function. DNA Seq. 13, 77–83.

[16] Bron, C., Kerbosch, J., 1973. Algorithm 457: Finding all cliques of an undirected graph. CACM 16, 575–577.

[17] Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Green, E. D., Sidow, A., Batzoglou, S., 2003. LAGAN and multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 13, 721–731.

[18] Buneman, P., 1971. The recovery of trees from measures of dissimilarity. In: Hodson, F. R., Kendall, D. G., Tautu, P. (eds.), Mathematics and the Archeological and Historical Sciences, Edinburgh University Press, Edinburgh, UK, pages 387–395.

[19] Carroll, S. B., Grenier, J. K., Weatherbee, S. D., 2001. From DNA to Diversity. Blackwell Science, Malden, MA.

[20] Carter, A. J., Wagner, G. P., 2002. Evolution of functionally conserved enhancers can be accelerated in large populations: a population-genetic model. Proc. R. Soc. Lond. B Biol. Sci. 269, 953–960.

[21] Chiu, C.-h., Amemiya, C., Dewar, K., Kim, C.-B., Ruddle, F. H., Wagner, G. P., 2002. Molecular evolution of the HoxA cluster in the three major gnathostome lineages. Proc. Natl. Acad. Sci. USA 99, 5492–5497.

[22] Chiu, C.-h., Amemiya, C., Dewar, K., Kim, C.-B., Ruddle, F. H., Wagner, G. P., 2002. Molecular evolution of the HoxA cluster in the three major gnathostome lineages. Proc. Natl. Acad. Sci. USA 99, 5492–5497.

[23] Chiu, C.-H., Nonaka, D., Xue, L., Amemiya, C. T., Wagner, G. P., 2000. Evolution of *Hoxa-11* in lineages phylogenetically positioned along the fin-limb transition. Mol. Phylogen. Evol. 17, 305–316.

[24] Davidson, E., 2001. Genomic Regulatory Systems. Academic Press, San Diego.

[25] (DOE Joint Genome Institute), 2002. Fugu genome database.

version 2.0: `http://genome.jgi-psf.org/fugu3/fugu3.home.html`,

version 3.0: `http://genome.jgi-psf.org/fugu6/fugu6.home.html`.

[26] Duret, L., Bucher, P., 1997. Searching for regulatory elements in human noncoding sequences. Curr. Opin. Struct. Biol. 7, 399–406.

[27] Ebersberger, I., Metzler, D., C., S., Pääbo, S., 2002. Genomewide comparison of DNA sequences between humans and chimpazees. Am. J. Hum. Genet. 70, 1490–1497.

[28] Felsenstein, J., 1989. Phylip – phylogeny inference package (version 3.2). Cladistics 5, 164–166.

[29] Ferrier, D., PWH, H., 2002. Ciona intestinalis parahox genes: evolution of hox/parahox cluster integrity, developmental mode, and temporal colinearity. Mol Phylogenet Evol. 24, 412–417.

[30] Force, A., Amores, A., Postlethwait, J. H., 2002. Hox cluster organization in the jawless vertebrate *Petromyzon marinus*. J. Exp. Zool. (Mol. Dev. Evol.) 294, 30–46.

[31] Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-l., Postlethwait, J., 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151, 1531–1545.

[32] Fried, C., 2003. Discovery of Transcription Factor Binding Sites. Master's thesis, University of Vienna.

[33] Fried, C., Hordijk, W., Prohaska, S., Stadler, C., Stadler, P. The footprint sorting problem .

[34] Fried, C., Prohaska, S., Stadler, P., 2003. Independet hox-cluster duplications in lampreys Submitted.

[35] Garcia-Fernández, J., Holland, P. W., 1994. Archetypal organization of the amphioxus hox gene cluster. Nature 370, 563–566.

[36] Garey, M. R., Johnson, D. S., 1979. Computers and Intractability. A Guide to the Theory of $\mathcal{NP}$ Completeness. Freeman, San Francisco.

[37] Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A. E., Kel, O. V., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Kolpakov, F. A., L., P. N., Kolchanov, N. A., 1998. Databases on transcriptional regulation: Transfac, trrd, and compel. Nucleic Acids Res. 26, 364–370.

[38] Hertz, G., Stormo, G. D., 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15, 563–577.

[39] Holland, P. W., Garcia-Fernandez, J., 1996. Hox genes and chordate evolution. Dev. Biol. 173, 382–395.

[40] Holland, P. W. H., Garcia-Fernández, J., Williams, N. A., Sidow, A., 1994. Gene duplication and the origins of vertebrate development. Development (Suppl.), 125–133.

[41] Huson, D. H., 1998. Splitstree: analyzing and visualizing evolutionary data. Bioinformatics 14, 68–73.

[42] Irvine, S. Q., Carr, J. L., Bailey, W. J., Kawasaki, K., Shimizu, N., Amemiya, C. T., Ruddle, F. H., 2002. Genomic analysis of Hox clusters in the sea lamprey, *Petromyzon marinus*. J. Exp. Zool. (Mol. Dev. Evol.) 294, 47–62.

[43] Kappen, C., Schughart, K., Ruddle, F. H., 1989. Two steps in the evolution of antennapedia-class vertebrate homeobox genes. Proc. Natl. Acad. Sci. USA 86, 5459–5463.

[44] Kim, C. B., Amemiya, C., Bailey, W., Kawasaki, K., Mezey, J., Miller, W., Minosima, S., Shimizu, N., P., W. G., Ruddle, F., 2000. Hox cluster genomics in the horn shark, *heterodontus francisci*. Proc. Natl. Acad. Sci. USA 97, 1655–1660.

[45] Kmita, M., Fraudeau, N., Herault, Y., D, D., 2002. Serial deletions and duplications suggest a mechanism for the collinearity of hoxd genes in limbs. Nature 420, 145–150.

[46] Kondo, T., Zakany, J., Duboule, D., 1998. Control of the coliniarity in abdb genes of the mouse hoxd complex. Mol. Cell 1, 289–300.

[47] Leung, J. Y., McKenzie, F. E., Uglialoro, A. M., Flores-Villanueva, P. O., Sorkin, B. C., Yunis, E. J., Hartl, D. L., Goldfeld, A. E., 2000. Identification of phylogenetic footprints in primate tumor necrosis factor-α promoters. Proc. Natl. Acad. Sci. USA 97, 6614–6618.

[48] Loots, G., Ovcharenko, I., Pachter, L., Dubchak, I., E, R., 2002. `rVISTA` for comparative sequence-based discovery of functional transcription factor binding sites. Genome. Res. 12, 832–839.

[49] Ludwig, M. Z., 2002. Functional evolution of noncoding DNA. Curr. Op. Genet. Devel. 12, 634–639.

[50] Ludwig, M. Z., Bergman, C., Patel, N. H., Kreitman, M., 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403, 564–567.

[51] Málaga-Trillo, E., Meyer, A., 2001. Genome duplications and accelerated evolution of *Hox* genes and cluster architecture in teleost fishes. Amer. Zool. 41, 676–686.

[52] Manen, J., Savolainen, V., Simon, P., 1994. The *atpB* and *rbcL* promoters in plastid DNAs of a wide dicot range. J. Mol. Evol. 38, 577–582.

[53] Manzanares, M., Bel-Vialar, S., Ariza-McNaughton, L., Ferretti, E., Marshall, H., Maconochie, M. M., Blasi, F., Krumlauf, R., 2001. Independent regulation of initiation and maintenance phase of hoxa3 expression in the vertebrate hindbrain involve auto- and cross-regulatory mechanisms. Development 128, 3595–3607.

[54] McClintock, J. M., Kheirbek, M. A., Prince, V. E., 2002. Knockdown of duplicated zebrafish hoxb1 genes reveals distinct roles in hindbrain patterning and a novel mechanism of duplicate gene retention. Development 129, 2339–2354.

[55] Minguillon, C., Garcia-Fernandez, J., 2003. Genesis and evolution of the evx and mox genes and the extended hox and parahox gene clusters. Genome Biol. 4, R12.

[56] Morgenstern, B., 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15, 211–218.

[57] Peer, Y. V. d., Taylor, J. S., I., B., Meyer, A., 2001. The ghost of selection past: rates of evolution and functinal divergence of anciently duplicated genes. J. Mol. Evol. 53, 436–446.

[58] Prince, V. E., 2002. The hox paradox: More complex(es) than imagined. Developmental Biology 249, 1–15.

[59] Prohaska, S., Fried, C., Flamm, C., Wagner, G., Stadler, P. F., 2003. Surveying phylogenetic footprints in large gene clusters: Applications to Hox cluster duplications. Mol. Phyl. Evol. Submitted; SFI preprint #03-02-011.

[60] Roelen, B., Graaff, W., Forlani, S., Deschamps, J., 2002. Hox cluster polarity in early transcriptional availability: a high order regulatory level of clustered hox genes in the mouse. Mechanisms of Developemt 119, 81–90.

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

[62] Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol. Evol. 4, 406–425.

[63] Santini, S., Boore, J., Meyer, A., 2003. Evolutionary conservation of regulatory elements in vertebrate hox gene clusters. Genome Res. 13, 1111–1122.

[64] Schwartz, S., Kent, W., Smit, A., Zhang, Z., Baertsch, R., Hardison, R., Haussler, D., Miller, W., 2003. Human-mouse alignments with blastz. Genome Res. 13, 103–107.

[65] Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., , Miller, W., 2000. PipMaker — a web server for aligning two genomic DNA sequences. Genome Research 4, 577–586.

[66] Semple, C., Steel, M., 2003. Phylogenetics. Oxford University Press, Oxford UK.

[67] Shatalkin, A., 2002. The problem of archetype and current biology. Zh Obshch Biol 63, 275–291.

[68] Sinha, S., Tompa, M., 2002. Discovery of novel transcription factor binding sites by statistical overrepresentation. Nucl. Acids Res. 30, 5549–5560.

[69] Smit, A. F. A., 1999. Interspersed repeats and other mementos of transposable elements in the mammalian genomes. Curr. Opin. Genet. Devel. 9, 657–663.

[70] Stern, D. L., 2000. Evolutionary developmental biology and the problem of variation. Evolution 54, 1079–1091.

[71] Sumiyama, K., Kim, C., Ruddle, F. H., 2001. An efficient cis-element discovery method using multiple sequence comparisons based on evolutionary relationships. Genomics 71, 260–262.

[72] Tagle, D. A., Koop, B. F., Goodman, M., Slightom, J. L., Hess, D. L., Jones, R. T., 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J. Mol. Biol. 203, 439–455.

[73] Tautz, D., 2000. Evolution of transcriptional regulation. Curr. Opin. Genet. Dev. 10, 575–579.

[74] Thompson, J. D., Higgs, D. G., Gibson, T. J., 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. Nucl. Acids Res. 22, 4673–4680.

[75] Wagner, A., 1999. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. Bioinformatics 15, 776–784.

[76] Wagner, G. P., Chiu, C.-h., 2001. The tetrapod limb: Hypothesis on its origin. J. exp. Zool. 291, 226–240.

[77] Yi, S., Ellsworth, D. L., Li, W. H., 2002. Slow molecular clocks in old world monkeys, apes and humans. Mol. Biol. Evol. 19, 2191–2198.

[78] Zhu, J., Liu, J. S., Lawrence, C. E., 1998. Bayesian adaptive sequence alignment algorithms. Bioinformatics 14, 25–39.