# Discovery of Transcription Factor Binding Sites

**Diplomarbeit**

zur Erlangung des akademischen Grades

**Magistra rerum naturalium**

an der Fakultät für Naturwissenschaften und Mathematik
der Universität Wien


Vorgelegt von

**Claudia Fried**


Juni 2003

Dank an meine Eltern
nicht nur dafür, dass sie mir mein Studium ermöglicht haben

**An dieser Stelle möchte ich mich herzlich bei all jenen bedanken, die mich bei meiner Arbeit unterstützt haben.**

Allen voran Peter Stadler, der mich durch seine wissenschaftliche Leitung und sein überwältigendes Wissen unterstützte.

Für die gute Kollaboration:
Sonja J. Prohaska, Chi-Hua Chiu, Christoph Flamm, Günther P. Wagner, Peter Ahnert, Wim Hordijk, Chris T. Amemiya, Frank H. Ruddle.

Allen Kolleginnen und Kollegen, für die Hilfsbereitschaft und Unterstützung insbesondere:
Ivo Hofacker, Konstantin Klemm, Gil Benkö, Roman Stocsits, Gunther Eble, Michael Wolfinger, Guido Fritzsch, Andreas Svrcek-Seiler, Kurt Grünberger, Markus Jaritz, Jörg Hackermüller.

Petra, Peter, Paul, Heidi, Christoph, Klaus, Jürgen.

Doris, Rafael, Peter, Michael, Maria, Magda, Maria, Franz, Michaela und natürlich auch dem Rest der ganzen Sippe.

Für die tägliche Berichterstattung aus Wien und für Alles Andreas Neustifter.

# Abstract

Gene regulatory regions in non-coding genomic sequences are subject to stabilizing selection and therefore evolve much more slowly than adjacent non-functional DNA. The resulting *phylogenetic footprints* can be detected by comparison of the sequences surrounding orthologous genes in different species. Experimental evidence from a variety of sources shows that a major mode of developmental gene evolution is based on the modification of cis-regulatory elements. The comparative analysis of long sequences, such as complete *Hox* clusters, requires a computationally efficient and fully automatized approach. The changes in the footprint patterns are not necessarily well correlated with established phylogenetic relationships. In contrast to other footprinting algorithms such as `FootPrinter` we therefore do not invoke a maximum parsimony assumption. Our new program `tracker` first generates `blastz` alignments of all pairs of input sequences with a non-restrictive parameter setting. A hierarchy of filtering steps then removes insignificant matches. The resulting list of pairwise alignments is then combined into clusters of overlapping footprints. The technically demanding part of the algorithm is the resolution of various types of inconsistencies that may arise when overlapping alignments of several sequences are combined to a multiple alignment.

A comparative survey of the *HoxA* clusters of hornshark, human, zebrafish, pufferfish, striped bass and bichir reveals a massive loss of sequence conservation in the intergenic region of the derived teleost, consistent with the earlier findings of Chiu *et al.*. Furthermore, our analysis suggests, that the *HoxA* cluster of bichir shows a pattern of conservation that place this basal actinopterygian in between the single *HoxA* cluster of human and shark and the duplicated *HoxA* of the teleosts zebrafish, pufferfish and striped bass.

To investigate the chronology of *Hox* cluster duplication at the emergence of vertebrates we carried out an analysis of the publicly available *Hox* gene sequences from lampreys. This study provides evidence that the *Hox* clusters in lampreys and other vertebrate species arose from independent duplications. In particular, our analysis supports the hypothesis that the last common ancestor of agnathans and gnathostomes had only a single *Hox* cluster which was subsequently duplicated independently in the two lineages.

Furthermore we applied tracker to the analysis of several components of the immune system. We found that single nucleotide polymorphisms (SNPs), i.e. alterations in the DNA that have the potential to alter protein functions and therfore cause diseases are underrepresented in phylogenetic footprints. The result suggests that SNPs are in general detrimental to the function of phylogenetic footprints.

# Zusammenfassung

Regulatorische Regionen in nicht codierenden genomischen Sequenzen sind einer stabilisierenden Selektion unterworfen, daher evolvieren sie wesentlich langsamer als nicht funktionelle DNA. Mehrere Experimente belegen, dass die Veränderungen dieser cis-regulatorischen Elementen für die Evolution von Entwicklungsgenen verantwortlich sind. Die sogenannten phylogenetischen *"Footprints"* können durch Vergleich mehrer orthologer Sequenzen aus verschiedenen Spezies entdeckt werden. Die vergleichende Analyse mehrerer langer Sequenzen, z.B. der kompletten *Hox* Cluster, benötigt eine effiziente und automatisierte Vorgehensweise. Änderungen in der Zusammensetzung der phylogenetischen Footprints korrelieren jedoch nicht notwendigerweise mit der phylogenetischen Verwandtschaft der Sequenzen. Daher verwenden wir nicht, im Gegensatz zu anderen phylogenetischen Footprinting-Algorithmen, die "maximum parsimony assumption". Unser neuentwickeltes Progamm `Tracker` basiert auf der Bildung aller paarweisen `Blastz` Alignments der Ausgangssequenzen. Nicht signifikante Ergebnisse, durch nicht restriktive Parameterwahl entstanden, werden aus den anfänglichen Ergebnissen ausgesondert. Die daraus resultierende Liste von paarweisen Alignments wird zu Clustern aus überlappenden Alignments zusammengefasst. Der darauffolgende, technisch aufwendige Teil des Algorithmus wird benötigt um Widersprüche, die durch Gruppierung der Alignments aus mehreren Sequenzen entstehen, auflösen zu können.

Eine Analyse der *HoxA* Cluster von Mensch, Hornhai, Zebrafisch, Kugelfisch, Streifenbarsch und Flösselhecht zeigt einen massiven Verlust an Konservierung in der intergenischen Region der Teleosten. Diese Ergebnisse stimmen mit den Erkenntnissen aus einer früheren Studie von Chiu *et al.* überein. Die Verteilung der phylogenetische Footprints im *HoxA* Cluster des Flösselhechtes legt eine Einordnung desselben zwischen den Clustern des Hornhais und Menschen auf der einen Seite und den duplizierten *HoxA* Clustern der Teleosten auf der anderen Seite, nahe.

Um die Chronologie der *HoxA* Cluster Duplikation beim Auftreten der Vertebraten zu untersuchen, haben wir eine Neubewertung der öffentlich zugänglichen Sequenzen von Neunaugen vorgenommen. Die Untersuchung zeigt, dass die *Hox* Cluster der Vertebraten und Neunaugen durch unabhängige Duplikation entstanden sein können. Daraus folgt, dass der Vorläufer der Agnatha und Gnathostoma einen einzelnen *Hox* Cluster besessen hat.

Weiters verwendeten wir `Tracker` für eine Untersuchung verschiedener Elemente des Immunsystem. Wir konnten sehen, dass "single nucleotide polymorphisms" — Änderungen in der DNA, die zu veränderter Proteinfunktion führen und dadurch Krankheiten auslösen können — in phylogentischen Footprints unterrepresentiert sind. Es scheint das SNPs schädlich für die Funktion der regulatorischen Elemente sind.

# Contents


**1 Introduction** **1**

**2 A New Approach to Identify Phylogenetic Footprints: `Tracker`** **7**

2.1 Initial Set of Pairwise Alignments . . . . . . . . . . . . . . . . . . . . 7
2.2 Consistent Cliques . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
2.3 Postprocessing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
2.4 Implementation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

**3 Survey of the Vertebrate Immune System** **19**

3.1 Evolution of the Immune System . . . . . . . . . . . . . . . . . . . . 19
3.2 Analysis of Different Components of the Immune System . . . . . . . . . . . . . . . . . . . . . . . . . . 22
3.2.1 *CD8A* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22
3.2.2 Interferon$\gamma$ (*IFN*$\gamma$) . . . . . . . . . . . . . . . . . . . . . . . 24
3.2.3 Interleukin-13 (*IL13*) . . . . . . . . . . . . . . . . . . . . . . 25
3.2.4 Interleukin-2 (*IL2*) . . . . . . . . . . . . . . . . . . . . . . . 26
3.2.5 Interleukin-2 Receptor (*IL2RA*) . . . . . . . . . . . . . . . . . 27
3.2.6 Interleukin-4 (*IL4*) . . . . . . . . . . . . . . . . . . . . . . . 28
3.2.7 Matrix Metalloproteinase 3 (*MMP-3*) . . . . . . . . . . . . . . . 29
3.2.8 Interferon Regulatory Factor 1 (*IRF1*) . . . . . . . . . . . . . . 30
3.2.9 Prolactin (*PRL*) . . . . . . . . . . . . . . . . . . . . . . . . . 31
3.2.10 Interferon$\alpha$1 (*IFN*$\alpha$*1*) . . . . . . . . . . . . . . . . . . . . . 32
3.2.11 Signal Transducer and Activator of Transcription 3 (*STAT3*) . . 33
3.3 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

**4 New Insights Into the Evolution of *Hox* Clusters** **39**

4.1 *Hox* Genes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 39

4.2 Surprising Trends in the Evolution of Ray-finned Fishes . . . . . . . . . . 46
4.3 Evolution of Teleosts . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 46
4.3.1 Sequence Data . . . . . . . . . . . . . . . . . . . . . . . . . . . 49
4.3.2 Gene Positions . . . . . . . . . . . . . . . . . . . . . . . . . . . 49
4.3.3 Phylogenetic Footprints . . . . . . . . . . . . . . . . . . . . . . 51
4.3.4 Footprint Cluster Summary Statistics . . . . . . . . . . . . . . . . 52
4.3.5 Footprints as Phylogenetic Signal . . . . . . . . . . . . . . . . . . 54
4.3.6 Footprint Loss Statistics . . . . . . . . . . . . . . . . . . . . . . 54
4.4 Independent *Hox* Cluster Duplication in Lampreys . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 58
4.4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 58
4.4.2 Sequence Data . . . . . . . . . . . . . . . . . . . . . . . . . . . 59
4.4.3 Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 61
4.4.4 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 62
4.4.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 65

**5 Conclusion and Outlook 69**

**A Analysis of polypterus senegalus footprints 75**

**B Recipes 83**
B.1 Baked Striped Bass with Herb Stuffing [63] . . . . . . . . . . . . . . . . 83
B.1.1 Materials . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 83
B.1.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84
B.1.3 Result . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84
B.2 Stewed Shark [12] . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 85
B.2.1 Materials . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 85
B.2.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 85
B.2.3 Result . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86
B.3 Fugu [64] . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86
B.3.1 History of Fugu Eating . . . . . . . . . . . . . . . . . . . . . . . 86
B.3.2 Preparation of Fugu . . . . . . . . . . . . . . . . . . . . . . . . 86
B.3.3 Typical Fugu menus . . . . . . . . . . . . . . . . . . . . . . . . 86

**References 87**

# Chapter 1

# Introduction

A major challenge of genomics is the understanding of the regulation of gene expression. All eukaryotes share complex and highly conserved mechanisms of transcriptional regulation. A major component of this regulation are transcription factors, which bind to specific, short DNA sequence motifs in the cis-regulatory region of a gene and activate or repress its transcription. A key feature of transcriptional regulation is that genes are often regulated by more than one transcription factor. This implies that the sequence surrounding a gene is composed of a complex pattern of transcription factor binding sites [119]. The intergenic regions of *Hox* genes contain several clusters of transcription factor binding sites. A single binding element or cluster can drive the expression of a gene in one tissue. However the expression of the same gene in other tissues or in other stages of development is may be caused by other transcription factors [2]. The different regulatory elements bind to different binding sites.

Experimental detection of regulatory regions e.g. by electrophoretic mobility shift or nuclease protection [77] is difficult and time consuming. Therefore computational methods to identify these regions become very important. In the non-coding region surrounding a gene strongly conserved segments are found, so-called phylogenetic footprints [112]. Functional sequences tend to evolve much more slowly than nonfunctional sequences due to the fact that mutation in functional sequences generally cause phenotypic effects and therefore are eliminated by natural selection. On the other hand, neutral mutations — not causing phenotypic mutations — can accumulate. Therefore mutations in non-coding and non-functional sequences accumulate and sequences of different organisms differs significantly. In contrast the sequence composition of functional regulatory elements is under stabilizing selection and therefore remains highly conserved.

Since phylogenetic footprints are located in a sequence which is not translated it is clear that they have regulatory function. Furthermore, it has been shown in a number

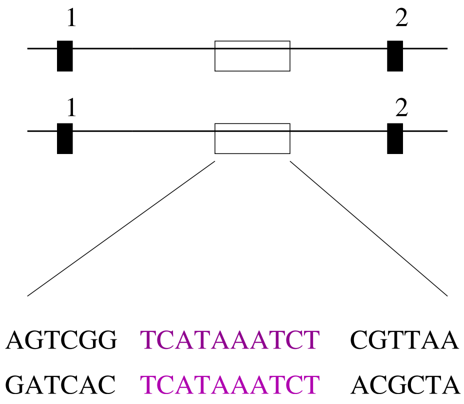



Figure 1.1. Conserved elements in the non-coding region of DNA often constitute elements that are involved in the regulation of gene expression. Usually these elements are small (5-20 bp) and closely spaced i.e. forming clusters of phylogenetic footprints.

of cases that these phylogenetic footprints are functional cis-regulatory elements [112, 73, 68, 25, 15], reviewed by Duret *et al.* [33] and Fickett *et al.* [38]. Most of these regulatory elements appear to be relatively short stretches of DNA (5-20 nucleotides). In the majority of cases the known regulatory elements are located in the 5‘ sequence of a gene, but they are also found 3‘ of the gene and in introns.

Phylogenetic footprinting is a very efficient approach to identify new unknown regulatory elements by comparative analysis of homologous or orthologous sequences [49]. The selected orthologous sequences should be from moderately diverse species so that there has been sufficient evolutionary time for mutations to accumulate in non-functional regions. Species with cumulative phylogenetic branch lengths of more than about 200 million years (e.g. orthologous genes from human, mouse and chicken) offer good candidates for such analysis.

There are two classes of approaches to identify regulatory regions. The most common methods are based on searches for common motifs in the non-coding sequences associated with related genes in the same organism, see e.g. [51, 98, 119]. Frequently occurring patterns in these sequences are presumably regulatory elements, which bind a common transcription factor. The disadvantage of these algorithms is the fact that they can determine only regulatory elements shared by many genes. This means that only common transcription factor binding sites are detected. Alternatively, orthologous non-coding sequences from a group of related species are used. Unusually well-conserved sequences then hint at a regulatory function. This approach is at present limited by the requirement of sequences of many different species. Sequences of appropriate species are often not available. Therefore the recently ongoing large-scale sequencing projects for some model vertebrate organisms will be valuable for the prediction of regulatory elements [33]. Most searches for phylogenetic footprints in the past were based on computing global alignments. The downside of these approaches is that multiple alignments are often computed in a way that the short highly conserved alignments are part of an optimal alignment but the sequences themselves are not aligned to each other. Standard motif search techniques such as `AlignAce` [55] and `ANN-Spec` [124] and segment-based alignment algorithm such as `DIALIGN` [82] have been shown to be more efficient [14]. 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 [70] such as `TRANSFAC` [123]. Most recently footprinting was expressed as a *substring parsimony problem* and an exact and rather efficient dynamic programming algorithm was proposed and implemented [14]. 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.

We developed a different algorithmic approach that appears to be more suitable for large clusters of genes with a complex regulation structure such as the *Hox* clusters. The reason is that at least in this case there appear to be substantial changes in the regulatory patterns that do not necessarily conform with established phylogenetic relationships. We therefore utilize a stepwise procedure that first extracts potentially conserved regions from pairwise sequence comparisons with `blastz` and pass these candidates through a series of filtering steps to obtain a final list of phylogenetic footprints. In Chapt. 2 we will present this new method of phylogenetic footprinting in detail.

An important mechanism of evolutionary changes are duplications of genes. It has been proposed, that for instance duplication events have played a decisive role in the evolution of vertebrates. They facilitate the achievement of the complexity of vertebrate forms through evolution of new gene functions. It is differentiated between different kinds of duplications: (1) Tandem duplication providing the embryo with an additional copy of the gene and (2) genome duplications duplicating the whole genome at once. Both of these events are leading to an increased number of genes and therefore genetic redundancy appears. Differential use of the genetic redundancy can cause higher diversity and changes in the morphology between different species. During the development of vertebrates from early deuterostoma entire genomes where duplicated through two rounds of duplication. In the actinopterygian lineage of fish a third duplication occurred after the two major lineages of fish the ray-finned fish (actinopterygian) and lobe finned fishes (sarcopterygian) had diverged [110].

Duplicated genes initially have fully redundant functions if the gene dosage is not

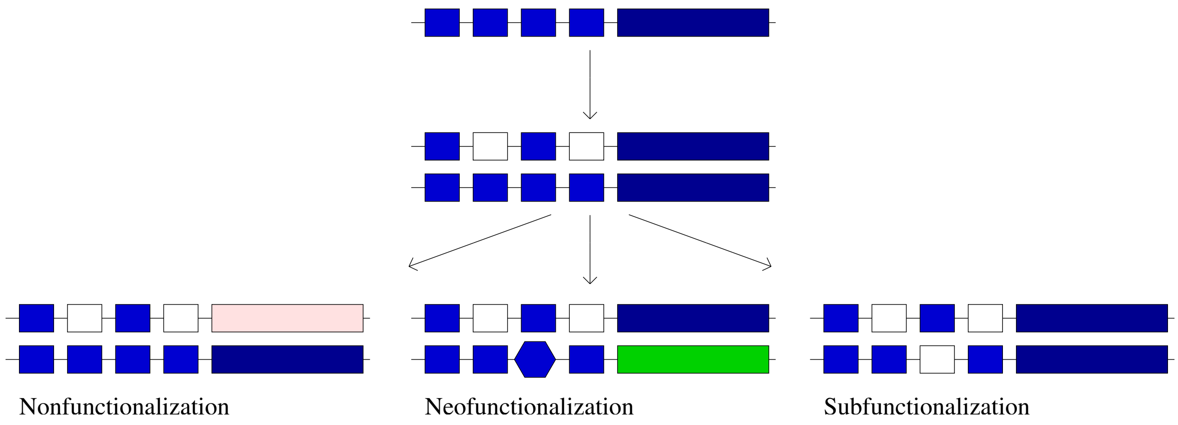



Figure 1.2. Overview of the three possible fates of genes after duplication according to the DDC model. Small boxes denote regulatory elements and large boxes symbolize genes. Blue boxes denote that the function of the gene or regulatory element is intact. Colorless or pink denote loss of function whereas green, respectively, a hexagon indicates the acquisition of a new function.

critical. The classical model [88] predicts that the most common fate for duplicated genes is the fixation of a null allele that prevents normal transcription and translation, i.e. the formation of a pseudo gene at one of the duplicated loci. In this model preservation of a gene is only possible by the fixation of rare beneficial mutations, where one copy of the duplicated genes gains a novel function and the second gene maintains the original functions. Disadvantageous mutations are far more likely than beneficial ones. Therefore it is difficult to explain the preservation of gene duplicates during evolution with the classical model. Despite the prediction of the classical model, loss of duplicated genes is not fast ongoing and many copies of the genes are retained in vertebrates [91]. The new DDC model introduced by Force *et al.* [42] takes this fact into account and instead explains gene preservation by fixation of degenerative mutations. After duplication, loss of function, loss of individual regulatory elements and gene inactivation can occur. According to the DDC model, loss of function can happen in three different ways which can be seen in Fig. 1.2:
(1) One gene copy is mutated. The mutation is fixed, leading to nonfunctionalization (gene loss). (2) All regulatory regions of one gene are destroyed leading to nonfunctionalization. (3) One copy gains a new function which eventually become fixed (neofunctionalization).
In all of these three cases one copy is altered and the other copy is preserved to maintain the function of the original single gene. Mutations in the genes can further involve changes in the regulatory region. Therefore the pattern of regulation changes after du-

plication. For example if duplicated genes lose different regulatory subfunctions they must complement each other to retain the full set of subfunctions present in the original gene. This includes the possibility that each duplicate loses or reduces the expression of different subfunctions therefore both copies are preserved and required to fulfill the function of the ancestral gene (subfunctionalization).

The duplication of whole clusters of genes is a significant factor in the evolution of vertebrates [59, 99, 53]. Clusters of particular importance are, among others, the *Hox* gene clusters, the *Dlx* bi-gene clusters and the *T-box* gene clusters. Gene duplications can be dated by determining which taxa possess the duplicated paralogs and which possess genes and regulatory elements descendant from the pre-duplication state. The study of gene clusters provides a particularly good opportunity for the study of non-coding sequence evolution, because the identity and extent of a non-coding sequence is uniquely defined by the flanking coding genes. A recent study of duplicated *Hox* clusters in zebrafish has shown that duplicated *Hox* gene clusters undergo a massive loss of non-coding sequence conservation [25]. This loss is associated with other changes in the cluster, most notably gene loss and shrinkage of inter-intergenic sequence length [4]. A systematic study of non-coding sequence conservation after duplication thus requires a stochastic model to estimate the amount of loss of sequence conservation due to the simple loss of some genes and the loss of cross-regulatory links. The purpose of such a model is to estimate the amount of conservation loss that can be attributed directly to gene loss and to determine whether other factors, such as adaptive evolution or binding site turnover, might also have played a role. In addition, the comparison of gene clusters has to be extended to include additional species in order to provide a sound comparative basis for evolutionary inferences. A model that fulfills these demands was recently established [95], a short introduction to the model is given in Sect. 4.2

In Chapt. 3 we describe an analysis of several components of the immune system that are known to be associated with different kinds of diseases. It is commonly believed that single nucleotide polymorphisms (SNPs) are responsible for the appearance of these diseases. These single nucleotide polymorphisms can be located in a regulatory region. The combination of information on SNPs with the results obtained by `tracker` has the potential to deliver insights in to the mechanism of diseases.

In Chapt. 4 we describe an application to the family of *Hox* genes. This gene family is of certain interest to study the evolution of vertebrates and the function of genome duplication during this evolution. The cluster consists of thirteen paralogous groups of genes and already all invertebrates possess one *Hox* cluster albeit not all of them consist of thirteen genes. Vertebrates, on the other hand, have multiple *Hox* gene clusters that have arisen from a single cluster in the most recent common ancestor of chordates, i.e. Amphioxus and vertebrates [44, 59]. While mammals have four *Hox* clusters an additional duplication event in the teleost lineage leads to an increased number of distinct clusters [4].

In Sect. 4.2 we investigate the timing and evolution of *Hox* cluster duplication in the ray-finned fish (actinopterygian). We used only *HoxA* clusters for our analysis due to the fact that for this cluster sufficient sequences are available. Therefore we apply our program to the investigation of the non-coding sequence of some derived actinopterygians which possess duplicated *HoxA* clusters and bichir, a basal actinopterygian. These basal species contain only one *HoxA* cluster. Our analysis supports that bichir has a single *HoxA* cluster that is mosaic in conservation of non-coding sequences between human and horn shark with single *HoxA* clusters and zebrafish, pufferfish and striped bass with duplicated *HoxA* clusters designated as *HoxAa* and *HoxAb*. The bichir *HoxA* cluster has conserved non-coding sequences that are shared uniquely with *HoxA* clusters of teleosts. This suggests that the *HoxA* clusters acquired novel cis-regulatory sequences in the actinopterygian stem lineage that are maintained after cluster duplication.

In Chapt. 4.4 we report the analysis of the publicly available *Hox* gene sequences from lampreys. This analysis provides evidence that the *Hox* clusters in lampreys and other vertebrate species arose from independent duplications. In particular, our analysis supports the hypothesis that the last common ancestor of agnathans and gnathostomes had only a single *Hox* cluster which was subsequently duplicated independently in the two lineages.

Chapter 2

# A New Approach to Identify Phylogenetic Footprints: `Tracker`

In this chapter we will describe in detail the new approach of phylogenetic footprinting. `Tracker` is suitable for the analysis of large sequences, in contrast to other footprinting algorithms. The method was developed in a joint work with S. Prohaska, C. Flamm and G. P. Wagner [95].

## 2.1 Initial Set of Pairwise Alignments

The initial step of the program `tracker` is based on `blastz` [102] that computes local gapfree alignments of two sequences using dynamic Programing. Dynamic Programming refers to a collection of algorithms that can be used to compute optimal solutions for solving combinatorial optimization problems. `Blastz` is an implementation of `blast` [3]. `Blast` is a standard tool for sequence comparison to discover sequence homology. The basic principle of `blast` is a heuristic algorithm which seeks local alignments. Therefore the program is able to detect small regions of similarity between to sequences. To measure the similarity of two sequences `blast` uses a matrix of similarity scores for all possible base pairs. Identities and conservative replacements have positive scores while unlikely replacements have negative scores. For instance the scores for mismatch in a DNA sequence is -4 and +5 for identical bases. The overall similarity score for the two sequences of an alignment is the sum of the similarity values for aligned residues. The algorithm is an modification of the Smith-Waterman [107] and Sellers [103] algorithms and determines all segment pairs whose scores can not be improved by extension or trimming of the sequences. `Blast` provides all of these so called maximal segment pairs with a score higher than a chosen threshold. `Blastz` is

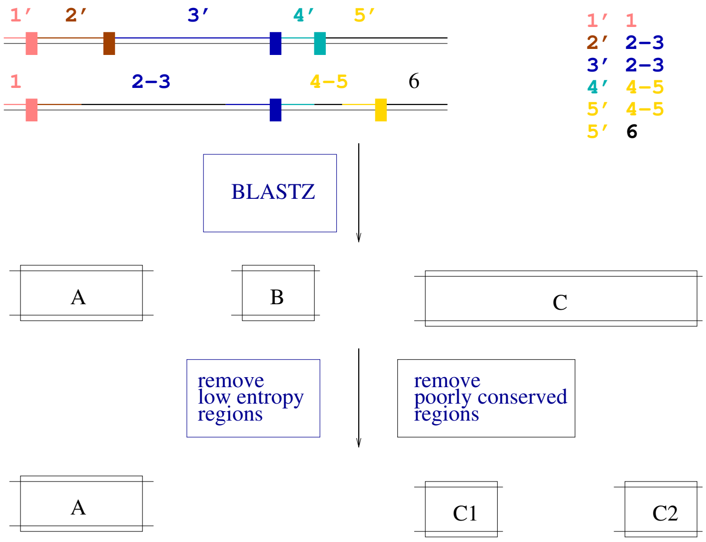



Figure 2.1. The first step of the `tracker` is to build local alignments for the intergenic regions between two orthologous genes. Blastz searches for a conserved gapless sequence of length $W$ and extends the alignment in both directions until it reaches a maximal score greater than a threshold score $K$. The next step are some filtering procedures to remove repetitive and poorly conserved regions, see text for a detailed description.

an implementation that is designed to align very long sequences. It is implemented in the `tracker` 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.

Consider the example shown in Fig. 2.2. The IGR between the *1'* and *2'* together with the region between *2'* and *3'* of Sequence 1 is compared with the region between *1* and *1* of the sequence 2. In the study of teleost *Hox* cluster evolution see Chapt. 4 we exclude introns. But they can be easily included by simply treating them analogous to IGRs, i.e., by listing individual exons instead of entire genes in the input. This method was used for the analysis of several genes of the immune system in Chapt. 3. 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` [19] 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

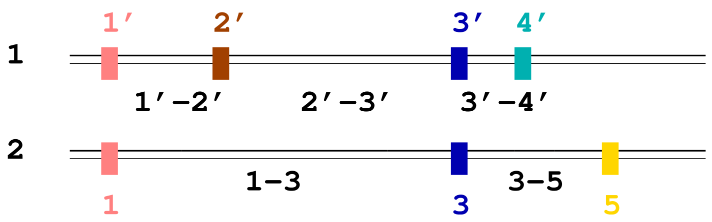



Figure 2.2. Only the intergenic regions between two homologous genes are compared. The intergenic region *1-3* of sequence 2 is compared with the regions *1'-2'* and *2'-3'* of sequence 1. The intergenic region *3'-4'* is compared with *3-5*.

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 `RepeatMasker` [106] 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 as well be functional. Local entropy measures are computed from the nucleotide frequencies $f_a(i)$ for 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 $\tau$ 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` [115] 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, see Sect. 2.4.

## 2.2 Consistent Cliques

We say that *two alignments $A_k$ and $A_l$ overlap* if at least one of the four intersections $A_k^1 \cap A_l^1$, $A_k^1 \cap A_l^2$, $A_k^2 \cap A_l^1$, and $A_k^2 \cap A_l^2$ 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 $A_k^u = x^p[i..i']$ and $A_l^v = x^p[j..j']$ 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 $\Gamma$ that has the individual `blastz`-alignments as its vertices. The edges of $\Gamma$ 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 $\Gamma$ into its connected components $\Gamma_c$, $c = 1, \ldots, n_C$, which we will refer to as "clusters". Since the clusters are independent of each other, they can be processed separately in further processing stages.

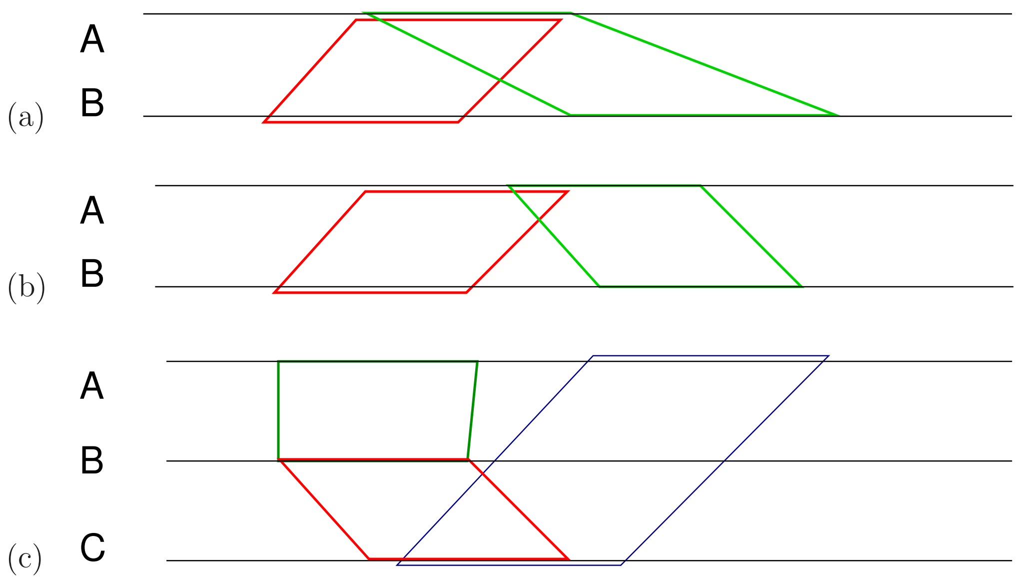



Figure 2.3. (a) Two alignments that overlap in sequence $A$ match with disjoint subsequences of $B$: clearly these to alignments are inconsistent in the sense that they cannot even be approximately part of a common alignment. (b) This situation is more subtle because the small overlap of only a few nucleotides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

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 *consistent* 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 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 *compatible* 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 situation in Fig. 2.3c,

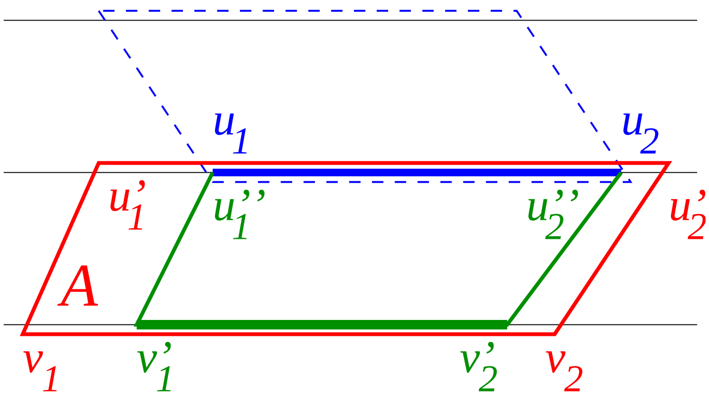



Figure 2.4. Notation for the inconsistency-finding algorithm. $[v_1', v_2']$ is trace of $[u_1, u_2]$ under the alignment $A$. See text for details.

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 $\Gamma_c$. 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 $A_1^1$ are parts of the same input sequence $x^p$ we have to check whether $\hat{A}_k^2 \subseteq A_1^1$. An inconsistency occurs if $\hat{A}_k^2 \not\subseteq A_1^1$, 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* which is associated with an interval $[u_1, u_2]$, we attach edges for each alignment 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

*with the exception of $[p_1, p_2]$, of course 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 $\alpha_j$ such that $u''_j = u'_1 + \alpha_j(u'_2 - u'_1)$ 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 if 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 situations may occur:
(i) We arrive at a trace $[p^*_1, p^*_2]$ such that there is a 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 in this case we 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 $\Gamma$. 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 incompatible alignments. Repeating this search procedure with each alignment as "root" we obtain all pairwise inconsistencies. These define the graph $\Psi_c$ that has the `blastz`-alignments of the cluster $\Gamma_c$ as its vertices and has an edge between $A$ and $B$ if and only if $A$

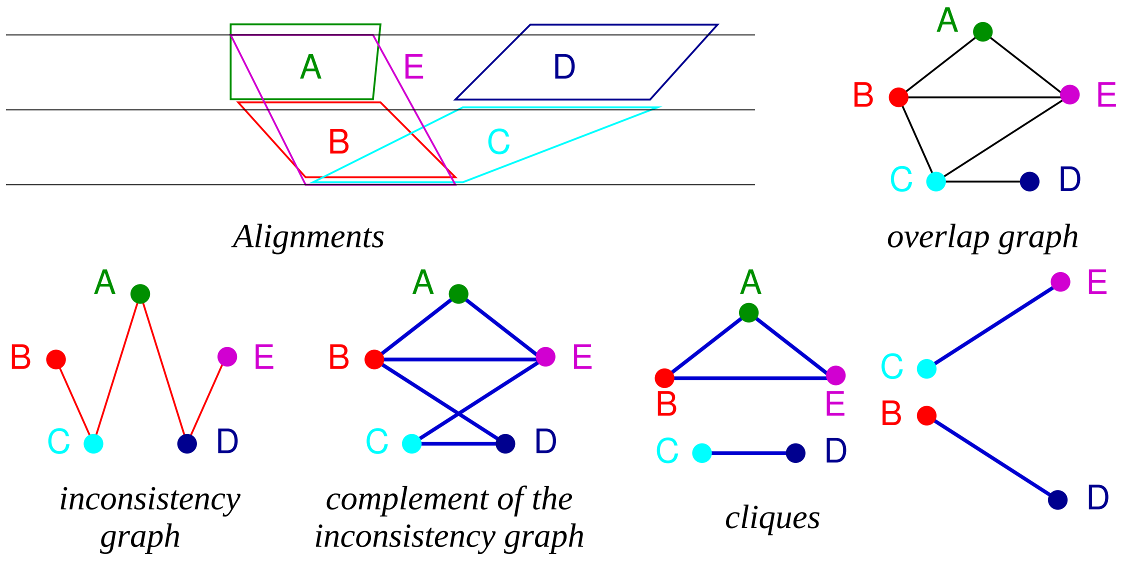



Figure 2.5. Decomposition of a cluster of alignments: First the overlap graph $\Gamma$ is computed for a set of alignments. Here we show only a single connected component ("cluster"). The incompatibility graph $\Psi$ summarized pairs of alignments that cannot be derived from a common multiple alignment. 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 $C_3$.

and $B$ are incompatible. From $\Psi_i$ we obtain the maximal sets of mutually compatible 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 $\Psi_c$). 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 compatible. The list $\mathcal{C}_c = \{C_h^c\}$ of the cliques of $\overline{\Psi_c}$ can therefore be produced efficiently by means of the Bron-Kerbosch algorithm [17] although in general even finding the maximal clique of a graph is NP-hard [45].

The induced subgraphs $\Gamma_c[C_h^c]$ are not necessarily connected. However, they might consist of alignments that do not overlap, see 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.

## 2.3 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 cases 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 [25] 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, so that it makes little difference whether PFCs or `tracker`'s footprint cliques 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` [115] or `dialign` [82]. 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 clusters are arranged according to common presence/absence patterns together with the parsimony score for the corresponding tree. In addition, overview charts are produced that summarize the locations of the footprints with a common distribution on the phylogenetic tree, see Chapt. 4 for examples.

## 2.4 Implementation

The `tracker` method is implemented as a `perl` program utilizing a number of external `ANSI C` modules e.g. for determining the inconsistency graph. Furthermore, `blastz` [102], dialign [82] and `clustalw` [115] are used as system calls. The output is provided as a LaTeX document with included Postscript figures. The tables in the appendix are, with the exception of the annotation in the last column, taken directly from the `tracker` output.

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. To test the performance of our approach on other datasets we run `tracker` on different kinds of datasets such as Interleukins which are chemical substances produced by leukocytes, that serve regulatory functions in the immune response e.g. stimulation of B cell proliferation, helper T cell activation, induction of fever. For these genes only

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 |

sequences of the closely related organisms *Homo sapiens*, *Mus musculus* and *Rattus norvegicus* are available. For such closely related sequences however, one should use more stringent values for the minimal quality of the conserved sequence blocks. To obtain suitable results we we used a threshold of $\mu_{\min} = 95\%$ in these cases.

The computations for entire *Hox* cluster sequences using a dataset of 5-8 taxa with an average length of 100kb is running below 5 minutes on a fast PC. This should be compared with many month of tedious work using web-based tools for data reported in the study of Chiu *et al.* [25]. Few of the available methods for phylogenetic footprinting can cope with multiple sequences at once. Those that can are restricted to rather short sequences in comparison to the sequences in our dataset.

A detailed comparison of the performance of different footprinting programs can be found in the MSc thesis of Sonja Prohaska [93]. In this work a detailed analysis of the orthologous region from *HoxA4* to *HoxA3* described. Four experimentally verified protein binding sites were described recently [74] for this region. None of these four binding sites is detected by `TFsearch` [122] or `FootPrinter`. The `FootPrinter` program is designed to find motifs in promoter regions or introns, where each sequence at most a few thousand nucleotides long [15]. An attempt to use this program for the complete *HoxA* cluster sequences was not successful. `Bayes block aligner` [127] in general detects fewer footprints than `tracker`; `dialign`, on the other hand, is typically more sensitive albeit at the expense of a more than tenfold consumption of computer time and memory. It is also worth noting that `dialign` even fails to correctly align

Table 2.2. Sensitivity of footprinting programs. The recovery of four experimentally verified binding sites by different computational approaches is compared. See the MSc thesis of Sonja Prohaska [93] for a more detailed analysis.
Four binding sites in the intergenic regions between *HoxA4* and *HoxA3* were discovered experimentally [74]. Different footprinting methods detect only some of them (+) by comparing the *HoxA* cluster sequences from the hornshark (*Heterodontus francisci*, Hf), human (*Homo sapiens*, Hs), bichir (*Polypterus senegalus*) and the *HoxAa* clusters from the pufferfish (*Takifugu rubripes*) and the zebrafish (*Danio rerio*).
BBA ... `Bayes block aligner` makes pairwise comparisons only, in this case of the human and hornshark sequence.

| | `dialign` | | `tracker` | | BBA | `FootPr.` | |
|---|---|---|---|---|---|---|---|
| Binding site | Hf | Hs | Hf | Hs | | Hf | Hs |
| KrA site | − | + | + | + | − | − | − |
| HOX/PBC siteA | + | + | − | + | − | − | − |
| HOX/PBC siteB | + | + | + | + | + | − | − |
| Prep/Meis | + | + | + | + | + | − | − |

some of exons when the complete *HoxA* clusters are used as input.

Chapter 3

# Survey of the Vertebrate Immune System

## 3.1 Evolution of the Immune System

Exposure to pathogens and parasites has always existed and all organisms exposed to this danger have to find mechanisms to deal with an infection. Studying the evolution of the immune system the fundamental question arises, how diverse organisms recognize and respond to non-self, especially to parasites and microbes. The elucidation of fundamental mechanisms of immune defense, and the manner in which they have changed and adapted to pathogens is basic to our understanding of the evolution of the immune system. Plants have developed primitive defenses against bacteria by producing toxic substances. Higher animals and humans fight back against an infection with more specific response mechanisms. The elements of the vertebrate immune system are complex and highly developed. Most of these components cannot be traced to their origin although it is possible to find some of the components in primitive animals. In this study we have to restrict ourselves to higher vertebrates because the analysis of non-coding sequence requires a sufficient amount of sequenced homologous genes and surrounding DNA. This information however is almost exclusively available for vertebrates.

The immune system is divided into two parts: (1) the innate and (2) adaptive immunity. Innate immunity is the protection against infections that relies on mechanisms that already exist in the body before infection. Therefore these mechanisms are capable of a rapid response to microbes and parasites. The innate immune system includes physical and chemical barriers e.g. epithelial barriers, phagocytes (neutrophils, macrophages), natural killer cells, the complement system (complement falls into both innate and adaptive categories depending on whether or not it is activated by antibodies) and cytokines produced by mononuclear phagocytes. These cytokines are able to regulate and coordinate the activity of many cell types of the innate immunity. The function of the innate immunity is to provide a first unspecific line of defense against microbes.

In contrast to the unspecific defense mechanisms of the innate immune system the adaptive immune system is highly complex. The mechanisms of the adaptive immunity are capable to adapt to the infection and increase in efficiency and defensive capabilities with each exposure to a particular microbe. The major components of the adaptive immunity are the lymphocytes and their products. The activation of lymphocytes requires two distinct signals both the recognition of antigen and either microbial products or components of the innate immunity. Following activation the lymphocytes start the synthesis of new proteins e.g. antibodies (*IG* molecules). Furthermore, the cells start to proliferate and differentiate into effector and memory cells. The adaptive immunity is further divided into two parts, the humoral immunity and the cell-mediated immunity. Humoral immunity is mediated by antibodies, molecules circulating in the blood which are produced by the so called B lymphocytes. Antibodies are capable of specifically recognizing microbial antigens, to neutralize them and target microbes for elimination by other cell types. Cell-mediated immunity or cellular immunity is mediated by T lymphocytes. This line of defense is targeted against intracellular microbes such as viruses and some bacteria that survive an proliferate inside phagocytes and other host cells where they are inaccessible to circulating antibodies. Adaptive immune response takes place in three phases. At the beginning the antigen is recognized, this recognition phase is followed by the activation of lymphocytes and initiation of the effector phase which means the elimination of the antigen [1].

All multi-cellular organisms possess multiple components of the innate immune system, e.g. phagocytic cells, antimicrobial peptides and the alternative pathway of complement. The adaptive immune system, characterized by T and B lymphocytes, rearranging genes encoding T-cell receptors (*TCRs*) and antibodies (*IGs*), and the major histocompatibility complex (*MHC*) arose early in vertebrate evolution. Adaptive immunity is lacking in hagfish and lampreys, those species that represent early states of vertebrate evolution, but it is present in all other vertebrates examined. The jawed vertebrates such as shark are the first representatives that possess an adaptive immunity [121].

It is likely that the emergence of the adaptive immune system has been closely tied to the two genome-wide duplications that occurred early in the evolutionary history of the vertebrates. In Chapt. 4.4 we report on an evaluation of the publicly available lamprey *Hox* gene cluster. We can show that the *Hox* genes cluster of the lampreys have arisen from independent duplication events [43]. This suggests that the duplication events after the branching of the agnathan lineage — whose extant representative

lamprey do not possess adaptive immunity — were responsible for the evolution of the adaptive immune system.
Furthermore, in addition to the genome duplication events, the emergence of the adaptive immune system is correlated with the integration of recombination activating enzymes (*RAG*) genes. *RAG* genes are required for rearrangements of members of the *IG* family[32]. Members of the *IG* family are the *IGs* produced by B lymphocytes and T cell receptors (*TCRs*). It has been suggested that the insertion of a transposon into an exon of an *IG* domain containing gene is tied to the emergence of recombination [60]. Both *IG* molecules and *TCRs* are required for the recognition of antigens. *IGs* are circulating in the blood stream or are bound to B lymphocytes and can bind soluble antigens whereas *TCRs* are displayed on the surface of T lymphocytes and are required for the recognition of antigens which are presented by other cells a *MHC*-antigen complex [1]. The genetic organization of the *IG* and *TCR* genes are similar and are characterized by spatial separation of sequences that must be joined together to produce functional proteins. This joining and recombining leads to a great diversity in the produced *IG* and *TCRs* and therefore the capability of the adaptive immunity to adapt specifically to different microbes and parasites. Because *RAG* genes are necessary for these rearrangements, these genes are only active in immature B or T lymphocytes, the phase of the cells in which the specificity of the receptors is determined.
The adaptive immune system of vertebrates did not replace earlier forms of innate immunity. Rather, it makes efficient use of the ancient innate immune systems. Whilst the innate immune systems can eliminate the majority of pathogens to which an animal is exposed, the adaptive immune system of vertebrates is essential for responding to and eliminating those pathogens that escape the innate immune system. One aspect of the integration of the innate and adaptive immune systems is that, upon pathogen recognition, the innate immune system communicates this information to the adaptive immune system and enhances the adaptive immune response. Once activated, the adaptive system produces antibodies and molecular signaling molecules that coordinate the adaptive system and interactions between the adaptive and innate systems. It is known that the immune recognition, signaling, and gene regulatory mechanisms are conserved throughout evolution. The analysis of control elements (promoters and enhancers) in the surrounding of immune genes therefore can give insights into the mechanisms of the immune system and the evolution of these mechanisms.

## 3.2 Analysis of Different Components of the Immune System

The collection of the genes used here comes from miscellaneous origins. Nevertheless they all have in common that they are bio-medically important components of the immune system.

Most of the analyzed genes e.g. *IFNs* and interleukins are representatives of cytokines. These are proteins secreted by cells of both innate and adaptive immune systems in response to microbes and other antigens and are adopted to mediate many of the functions of these cells. They stimulate the growth and differentiation of lymphocytes, the development of hematopoetic cells and in the effector phase of innate and adaptive immunity the activate different cells to eliminate the cause of the immune response. They are important in many immune and inflammatory diseases as therapeutic agents or targets for specific antagonists [1]. Cytokines are part of the innate immune system. It is therefore likely that cytokines and their receptors have counterparts in the invertebrates. The tend to evolve very quickly like most of the molecules of the immune system. Therefore it is difficult to isolate these genes and until now the homologs are not detected [32]. Since *MMP-3* is a component of the innate immune system detection of homologs in invertebrates should be possible. However we could not found a homolog in the databank of the recently sequenced *Ciona intestinals* genome [108, 118] nor *Caenorhabditis elegans*. We found a homolog in *Drosophila melanogaster* but this sequence was excluded from the analysis due to the consistency with the other datasets. It is produced by macrophages in response to microbes. Prolactin is a peptide hormone that regulates cells of the immune. STAT3 has also regulatory functions in the immune system.

### 3.2.1 *CD8A*

*CD8A* is a subunit of the receptor *CD8* which plays a crucial role during thymocyte development and in class I *MHC*-restricted antigen-induced T cell effector function. Its expression is developmentally regulated. *CD8* is displayed on the surface of cytotoxic T-cells it functions as an adhesion molecule that binds class I*MHC*-antigen complexes [1]. *CD8A* polymorphisms are associated with rheumatoid arthritis (RA) susceptibility [9].

Table 3.1. Presence/absence patterns of phylogenetic footprints in the *CD8A* sequences.

| *Homo sapiens* | *Mus musculus* | *Rattus norvegicus* | Parsimony score | $n$ |
|---|---|---|---|---|
| + | + | + | 0 | 43 |
| − | + | + | 1 | 20 |
| + | − | + | 1 | 14 |
| + | + | − | 1 | 23 |

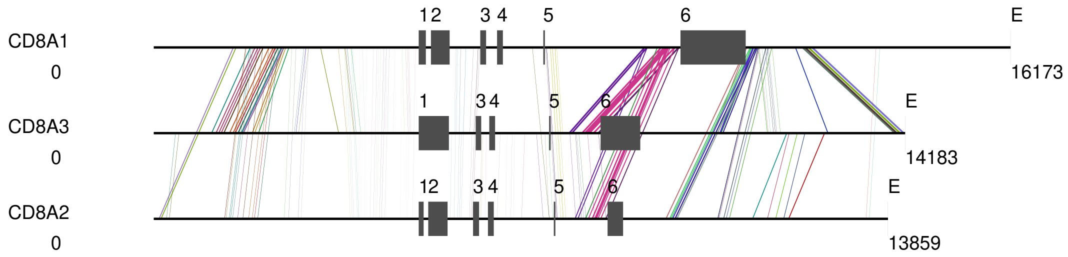



Figure 3.1. Overview of phylogenetic footprints of *CD8A* automatically generated by `tracker`. Each line corresponds to a footprint, consistent cliques are shown with the same color. Sequence data used in this study, the sequences span the entire gene and surrounding region 5000 base pairs upstream and downstream of the gene. Organisms and accession numbers are listed, $^{rc}$ denotes reverse complements of the database entries:
CD8A1 = *Homo sapiens* NT_015805$^{rc}$
CD8A2 = *Rattus norvegicus* NW_043757
CD8A3 = *Mus musculus* NW_000258

### 3.2.2 Interferon$\gamma$ (*IFN*$\gamma$)

*IFN*$\gamma$ is a multifunctional cytokine that is essential in the development of Th1 cells and in cellular responses to a variety of intracellular pathogens including human immunodeficiency virus (HIV-1). Its principal function is to activate macrophages in the innate and adaptive immune responses. It has been shown recently that the proximal promoter is invariant i.e. contains no polymorphisms [16]. Polymorphisms in the *IFN*$\gamma$ gene itself play an important role e.g. in type I diabetes [114, 58], asthma [83, 84], and multiple sclerosis [27]. Furthermore the association with RA is suspected [61].

Table 3.2. Presence/absence patterns of phylogenetic footprints in the *IFN*$\gamma$ sequences.

| *Homo sapiens* | *Mus musculus* | *Rattus norvegicus* | Parsimony score | $n$ |
|---|---|---|---|---|
| + | + | + | 0 | 50 |
| − | + | + | 1 | 33 |
| + | − | + | 1 | 5 |
| + | + | − | 1 | 40 |

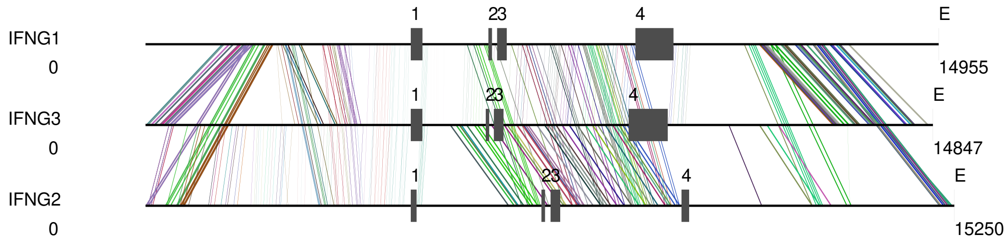



Figure 3.2. Phylogenetic footprints of *IFN*$\gamma$. Organisms and accession numbers, $^{rc}$ denotes reverse complements of the database entries:
IFNG1 = *Homo sapiens* NT_029419$^{rc}$
IFNG2 = *Rattus norvegicus* NW_044022
IFNG3 = *Mus musculus* NW_000033

### 3.2.3 Interleukin-13 (*IL13*)

*IL13* is produced by $T_H2$ i.e. the subfraction of T cells that stimulate B cell growth and regulate macrophage activation by cytokine production. The function of *IL13* is the inhibition of macrophage activation. It therefore antagonizes the action of *IFN*$\gamma$. It has been shown recently that the risk for Type 1 Diabetes (T1D) is determined, in part, by polymorphisms within the *IL4R* locus, including promoter and coding-sequence variants, and by specific combinations of genotypes at the *IL4R* and the *IL4* and *IL13* loci [20, 48]. Furthermore the association of IL13 with asthma has been shown recently [54].

Table 3.3. Presence/absence patterns of phylogenetic footprints in the *IL13* sequences.

| *Homo sapiens* | *Mus musculus* | *Rattus norvegicus* | Parsimony score | $n$ |
|---|---|---|---|---|
| + | + | + | 0 | 36 |
| − | + | + | 1 | 66 |
| + | + | − | 1 | 2 |

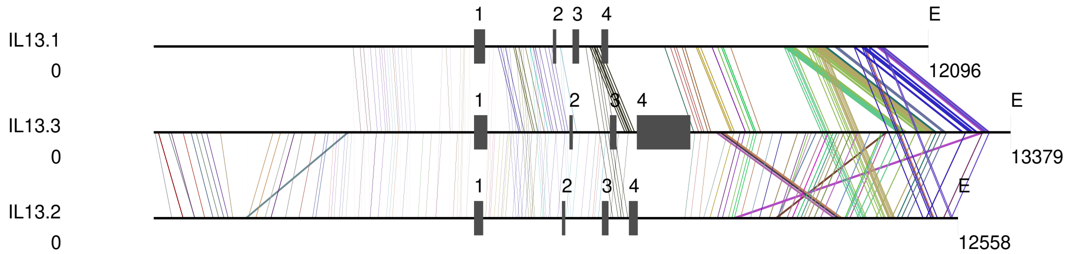



Figure 3.3. Phylogenetic footprints of *IL13*. Organisms and accession numbers, $^{rc}$ denotes reverse complements of the database entries:
IL13.1 = *Homo sapiens* NT_007072
IL13.2 = *Rattus norvegicus* NW_042654$^{rc}$
IL13.3 = *Mus musculus* NT_031405$^{rc}$

### 3.2.4 Interleukin-2 (*IL2*)

*IL2* is a cytokine produced by antigen activated T cells that stimulates T cell proliferation and also potentiates the apoptotic cell death of antigen-activated T cells. Thus *IL2* is required for the induction and selfregulation of T cell mediated immune responses. *IL2* also stimulates the proliferation and effector functions of natural killer cells and B cells [1] .

Table 3.4. Presence/absence patterns of phylogenetic footprints in the *Il2* sequences.

| *Homo sapiens* | *Mus musculus* | *Rattus norvegicus* | Parsimony score | $n$ |
|---|---|---|---|---|
| + | + | + | 0 | 11 |
| − | + | + | 1 | 17 |
| + | + | − | 1 | 27 |

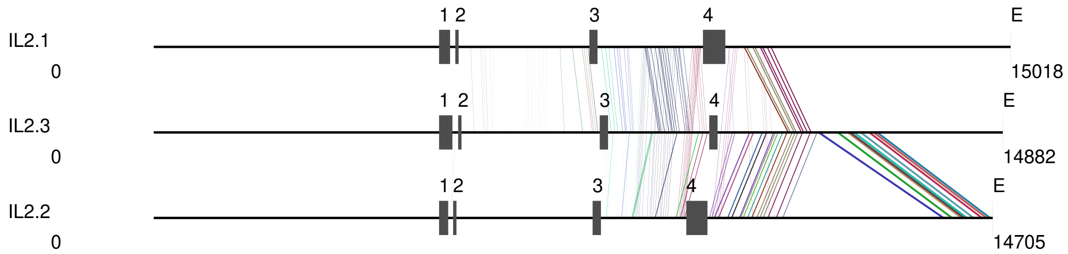



Figure 3.4. Phylogenetic footprints of *IL2*. Organisms and accession numbers, $^{rc}$ denotes reverse complements of the database entries:
IL2.1 = *Homo sapiens* NT_016354$^{rc}$
IL2.2 = *Rattus norvegicus* NW_043519$^{rc}$
IL2.3 = *Mus musculus* NW_000184$^{rc}$

### 3.2.5 Interleukin-2 Receptor (*IL2RA*)

Activation of T cells leads to the production of surface receptors for *IL2*. This Interleukin-2 receptor (*IL2RA*) is a 55-kD protein and is released into peripheral blood following T cell activation. Allergic bronchopulmonary aspergillosis is suspected to be associated with increased soluble interleukin 2 receptors [18].

Table 3.5. Presence/absence patterns of phylogenetic footprints in the *IL2RA* sequences.

| *Homo sapiens* | *Mus musculus* | *Rattus norvegicus* | Parsimony score | $n$ |
|---|---|---|---|---|
| + | + | + | 0 | 10 |
| − | + | + | 1 | 294 |
| + | + | − | 1 | 21 |
| + | − | + | 1 | 6 |

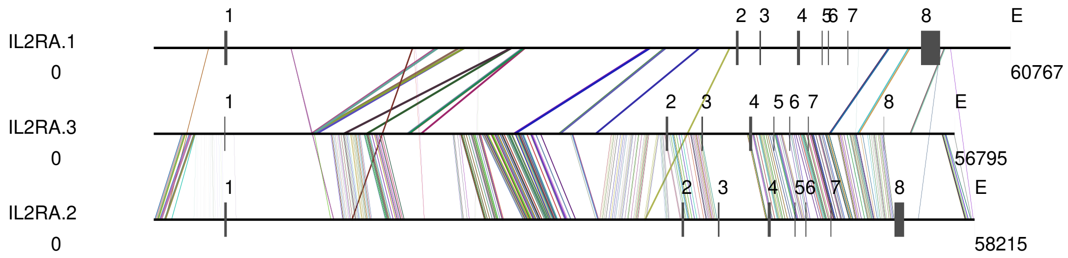



Figure 3.5. Phylogenetic footprints of *IL2RA*. Organisms and accession numbers, $^{rc}$ denotes reverse complements of the database entries:
IL2RA.1 = *Homo sapiens* NT_008705$^{rc}$
IL2RA.2 =*Rattus norvegicus* NW_043113$^{rc}$
IL2RA.3 =*Mus musculus* NT_037366$^{rc}$

### 3.2.6 Interleukin-4 (*IL4*)

*IL4* is a cytokine produced by TH2 cells and its function is the induction of the differentiation of TH2 cells from precursor cells. Furthermore it stimulates the *IgE* production of B cells and supresses macrophage functions [1]. Analysis of [20] supports the idea that the risk for Type 1 diabetes is determined, by polymorphisms within the *IL4R* locus and specific combinations of genotypes at the *IL4R* and the *IL4* and *IL13* loci. In addition, polymorphisms in the *IL4R*, *IL4*, and the *IL13* loci have been reported to be associated with atopic asthma [116, 100, 54].

Table 3.6. Presence/absence patterns of phylogenetic footprints in the *IL4* sequences.

| *Homo sapiens* | *Mus musculus* | *Rattus norvegicus* | Parsimony score | $n$ |
|---|---|---|---|---|
| + | + | + | 0 | 62 |
| − | + | + | 1 | 50 |
| + | − | + | 1 | 30 |
| + | + | − | 1 | 18 |

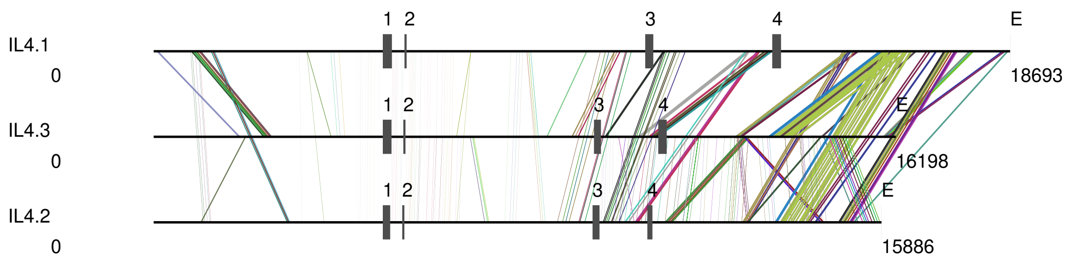



Figure 3.6. Phylogenetic footprints of *IL4* automatically generated by `tracker`. Organisms and accession numbers, $^{rc}$ denotes reverse complements of the database entries:
IL4.1 = *Homo sapiens* NT_007072
IL4.2 = *Rattus norvegicus* NW_042654$^{rc}$
IL4.3 = *Mus musculus* NT_031405$^{rc}$

### 3.2.7 Matrix Metalloproteinase 3 (*MMP-3*)

*MMP-3* is a member of the metalloproteinase family, structurally related to zinc dependent proteinases. These factors are considerd to be primarily responsible for the proper degradation and remodelling of the extracellular matrix molecules, that process takes place e.g. during adipose tissue formation [24]. *MMPs* participate in a large number of physiological processes such as embryonic development, angiogenesis, and wound repair. Furthermore, they can mediate the release and activation of growth factors and cleavage of cell surface receptors. The extracellular matrix metalloproteases produced by tumour and stromal cells are believed to play a key role in tumour cell invasion and metastasis [69]. The exposure of malignant mesothelioma cells to different growth factors increases the secretion of *MMP-3*. *MMP-3* also shows clinical significance for rheumatoid arthritis.

Table 3.7. Presence/absence patterns of phylogenetic footprints in the *MMP-3* sequences.

| *Homo sapiens* | *Mus musculus* | *Rattus norvegicus* | Parsimony score | $n$ |
|---|---|---|---|---|
| + | + | + | 0 | 40 |
| − | + | + | 1 | 27 |
| + | + | − | 1 | 13 |
| + | − | + | 1 | 2 |

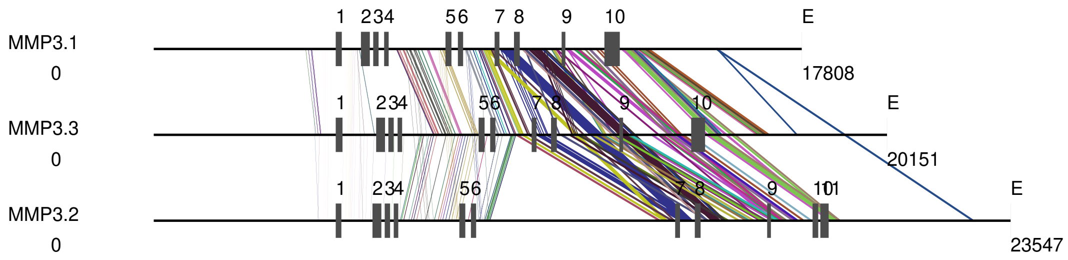



Figure 3.7. Phylogenetic footprints of *MMP-3* Organisms and accession numbers, $^{rc}$ denotes reverse complements of the database entries:
MMP3.1 = *Homo sapiens* NT_009151$^{rc}$
MMP3.2 =*Rattus norvegicus* NW_044087
MMP3.3 =*Mus musculus* NW_000350

### 3.2.8 Interferon Regulatory Factor 1 (*IRF1*)

*IRF1* is member of the interferon regulatory family. It functions as an activator of *IFN*$\alpha$ and *IFN*$\beta$ transcription. Therefore it has influence on the expression of genes that are regulated by these genes, e.g. *MHC* class I molecules. Furthermore *IRF1* plays a role in the regulation of apoptosis and tumor-suppression. It is supposed that *IRF1* is involved in RA susceptibility [31].

Table 3.8. Presence/absence patterns of phylogenetic footprints in the *IRF1* sequences.

| *Homo sapiens* | *Mus musculus* | *Rattus norvegicus* | Parsimony score | $n$ |
|---|---|---|---|---|
| + | + | + | 0 | 119 |
| − | + | + | 1 | 53 |
| + | − | + | 1 | 26 |
| + | + | − | 1 | 26 |

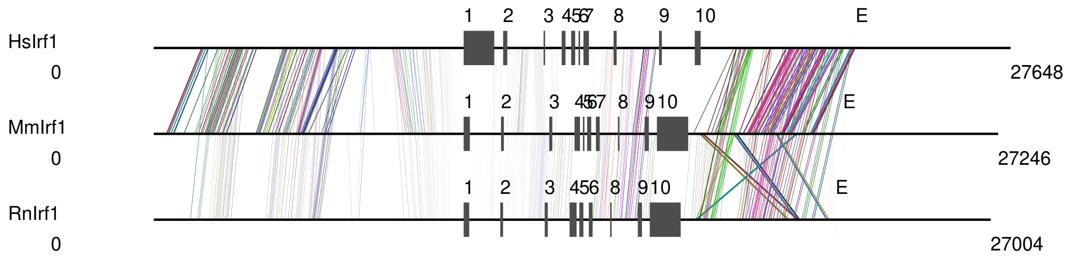



Figure 3.8. Phylogenetic footprints of *IRF1*. Organisms and accession numbers, $^{rc}$ denotes reverse complements of the database entries:
HsIrf1 = *Homo sapiens* NT_007072$^{rc}$
RnIrf1 = *Rattus norvegicus* NW_042654
MmIrf1 = *Mus musculus* NT_031405

### 3.2.9 Prolactin (*PRL*)

Prolactin is a well-known peptide hormone that regulates the growth, differentiation, maturation and apoptosis of cells of the immune system. The PRL receptor, which is a member of the hematopoietin/cytokine receptor superfamily, is widely expressed by immune cells, and subsets of lymphocytes secrete bioactive *PRL*. *PRL*, a female hormone may contribute to the prevalence of women in autoimmune diseases such as SLE (Systemic Lupus Erythematosus). *PRL* regulates the expression of one target gene, the transcription factor interferon regulatory factor-1 (*IRF-1*), which is a multifunctional immune regulator gene [76]. It appears that prolactin has a modulatory role in several aspects of immune function, but is not strictly required for these responses. Ectopic production of prolactin can be found in tumors such as renal cell, liver, uterine fibroids and mammary carcinomas.

Table 3.9. Presence/absence patterns of phylogenetic footprints in the *PRL* sequences.

| *Homo sapiens* | *Mus musculus* | *Rattus norvegicus* | Parsimony score | $n$ |
|---|---|---|---|---|
| + | + | + | 0 | 43 |
| − | + | + | 1 | 78 |
| + | + | − | 1 | 20 |
| + | − | + | 1 | 13 |

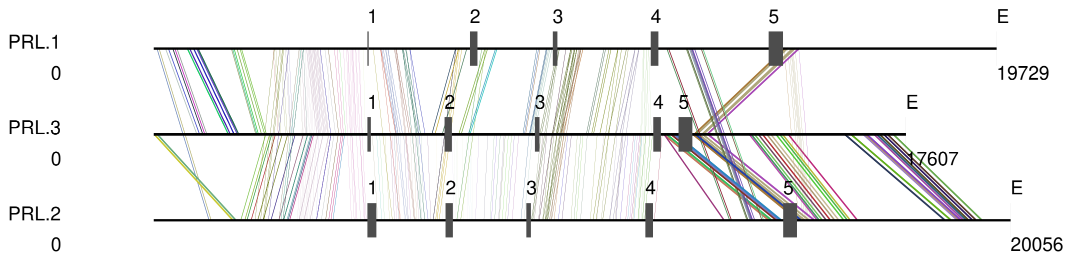



Figure 3.9. Phylogenetic footprints of *PRL*. Organisms and accession numbers, $^{rc}$ denotes reverse complements of the database entries:
PRL.1 = *Homo sapiens* NT_007592$^{rc}$
PRL.2 = *Rattus norvegicus* NW_043104$^{rc}$
PRL.3 = *Mus musculus* NT_030233

### 3.2.10 Interferon$\alpha$1 (*IFN$\alpha$1*)

*IFN$\alpha$1* is produced by monocytes and phagocytes after response to viral infections [1]. The gene product increases the expression of class I *MHC* molecules which leads to enhanced recognition of *MHC*-antigen complexes by cytotoxic T lymphocytes. Furthermore, *IFN$\alpha$* causes cells to produce enzymes that negatively affect the transcription of viral genome [1].

Table 3.10. Presence/absence patterns of phylogenetic footprints in the *IFN$\alpha$1* sequences.

| *Homo sapiens* | *Mus musculus* | *Rattus norvegicus* | Parsimony score | $n$ |
|---|---|---|---|---|
| + | + | + | 0 | 2 |
| − | + | + | 1 | 5 |
| + | − | + | 1 | 1 |
| + | + | − | 1 | 1 |

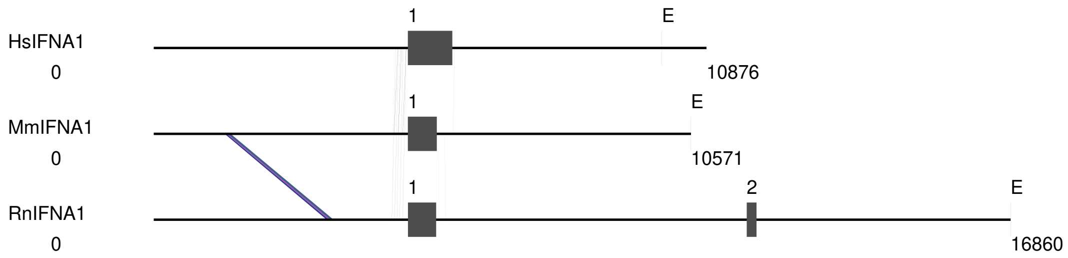



Figure 3.10. Phylogenetic footprints of *IFN$\alpha$1*. Organisms and accession numbers, $^{rc}$ denotes reverse complements of the database entries:
HsIFNA1 = *Homo sapiens* NT_037734
RnIFNA1 = *Rattus norvegicus* NW_043853$^{rc}$
MmIFNA1 = *Mus musculus* NW_000209

### 3.2.11 Signal Transducer and Activator of Transcription 3 (*STAT3*)

*STAT3* a member of a family of proteins that function as signaling molecules and transcription factors in response to cytokines binding to their receptors. *STATs* are present as inactive monomers in the cytoplasm of cells and are recruited to the cytoplasmic site of cytokine receptors, where they are phosphorylated at a tyrosin residue by Jak kinases. The phosphorylated *STAT3* protein dimerizse and moves to the nucleus, where it binds to regulatory sites in the promotor regions of various genes and therefore stimulates their transcription [1]. It is suppossed that STAT3 has an influence on RA [66].

Table 3.11. Presence/absence patterns of phylogenetic footprints in the *STAT3* sequences.

| *Homo sapiens* | *Mus musculus* | *Rattus norvegicus* | Parsimony score | $n$ |
|---|---|---|---|---|
| + | + | + | 0 | 128 |
| − | + | + | 1 | 171 |
| + | + | − | 1 | 9 |
| + | − | + | 1 | 11 |

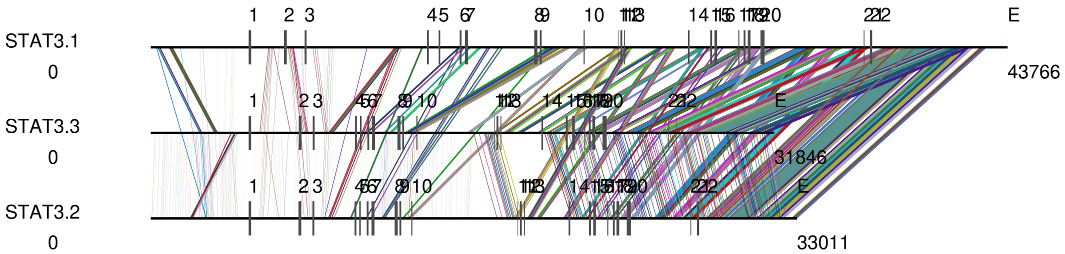



Figure 3.11. Phylogenetic footprints of *STAT3*. Organisms and accession numbers, $^{rc}$ denotes reverse complements of the database entries:
STAT3.1 = *Homo sapiens* NT_010840$^{rc}$
STAT3.2 =*Rattus norvegicus* NW_042674$^{rc}$
STAT3.3 =*Mus musculus* NT_026181

## 3.3 Discussion

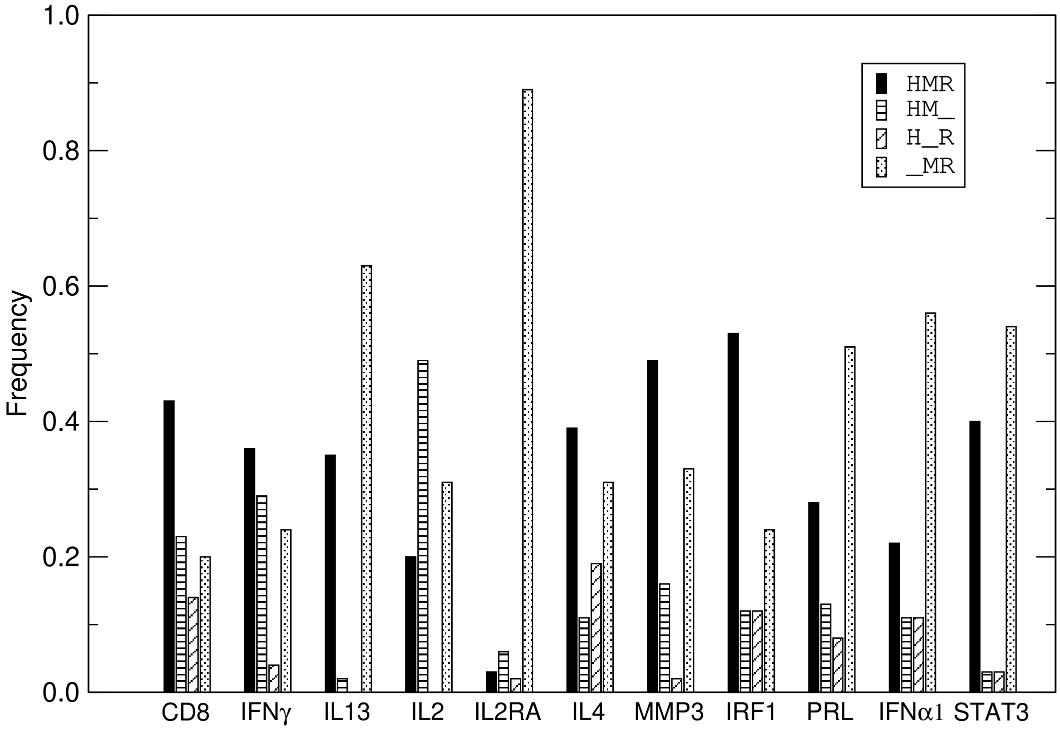



Figure 3.12. Graphical representation of the frequency of footprint presence/absence pattern in *Homo sapiens*, *Mus musculus* and *Rattus norvegicus*. Abbreviations used in this figure: H ... *Homo sapiens*
M ... *Mus musculus*
R ... *Rattus norvegicus*

In the majority of cases most conserved elements of the analysed genes can be found in all three sequences (human, mouse and rat) see Fig. 3.1-Fig. 3.11 and Fig. 3.12. In the other cases the most phylogenetic footprints are shared by the mouse and rat sequences, in some cases with a significantly higher number of footprints, e.g. *MMP-3* in Tab. 3.7. This depends to some extent on the fact that these sequences are closely related. It seems that there has not been sufficient evolutionary time for mutations to accumulate in non-functional regions. Therefore detection of phylogenetic footprints may produce false positives. Some of the footprints are highly conserved although they do not display a function. It is obvious both from the exon composition of the genes and the footprint patterns (see overview graphs Fig. 3.1-Fig. 3.11) that the rat sequences have undergone remodelling. In the footprint pattern of the rat one can see, in comparison to both of the other seqeunces that the rat sequence has sometimes big gaps between adjacent footprints indicating an inversion. Recent analysis showed that

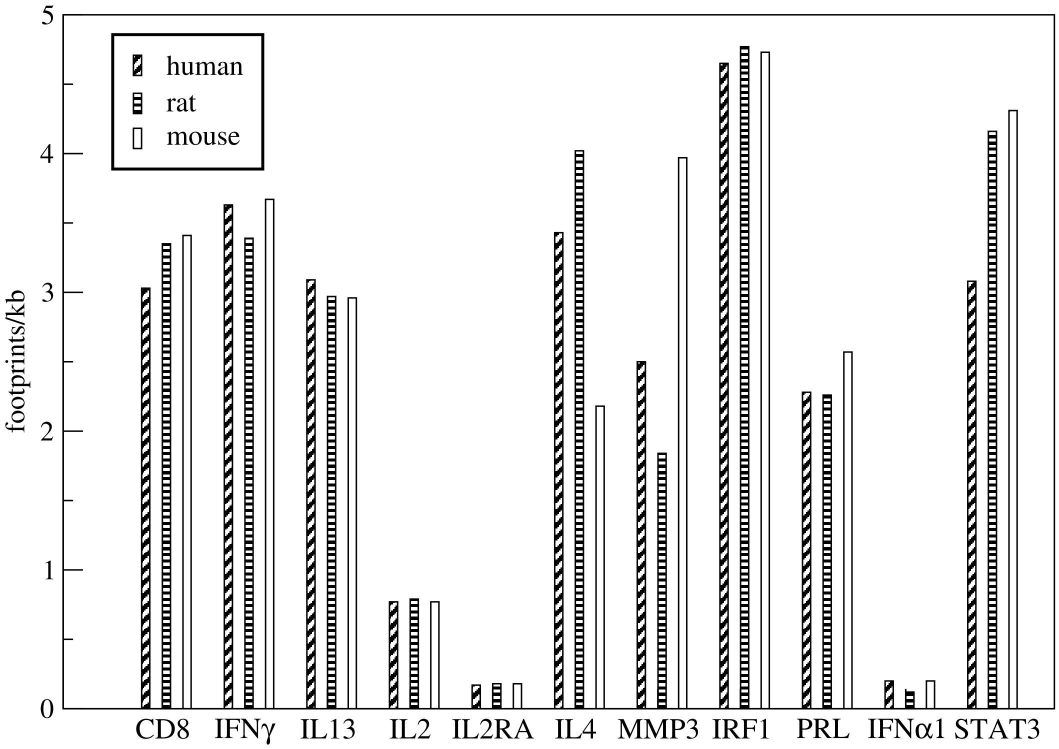



Figure 3.13. Amount of phylogentic footprints per kb in the analyzed genes for the species *Homo sapiens*, *Mus musculus* and *Rattus norvegicus*. Only footprints that are shared between all the clusters are counted. Abbreviations used in this figure:
H ... *Homo sapiens*
M ... *Mus musculus*
R ... *Rattus norvegicus*

the rat chromosome 10 differs by several chromosome rearrangements from the mouse or human homologs. For instance 6 inversions and 1 transposition event distinguish the human chromosome 17 and the homologous part of the rat 10 chromosome [13] . The rearrangements of the sequences are supported by the analysis of [80] who also demonstrate that several inversions and transpositions distinguish the rat, mouse and human X chromosome. To test the significance of these rearrangements we can compare the distances between adjacent footprints of several organisms.

Fig. 3.13 shows the average amount of footprints in 1000 basepairs. In most cases there are between 3-4 footprints per 1000 basepairs. Only *Il2*, *IL2RA* and *IFNα* have less than 1 footprint/kb. The sequences of *Il2RA* have a intron of about 35.000 basepairs between the first and the second exon. In this region the amount of footprints is small since most transcription factors bind in a close distance to the coding-region. A multiple alignment of the *Il2* sequences show that the rat sequence has large gaps in comparison to the other seqeunce. This is maybe a sequencing error since the sequence is derived by autuomatic computational analysis. For *IL2* and *IFNα* it is not clear why the have less transcriptionfactors than the other genes, although it seems that only the promotor region is conserved.

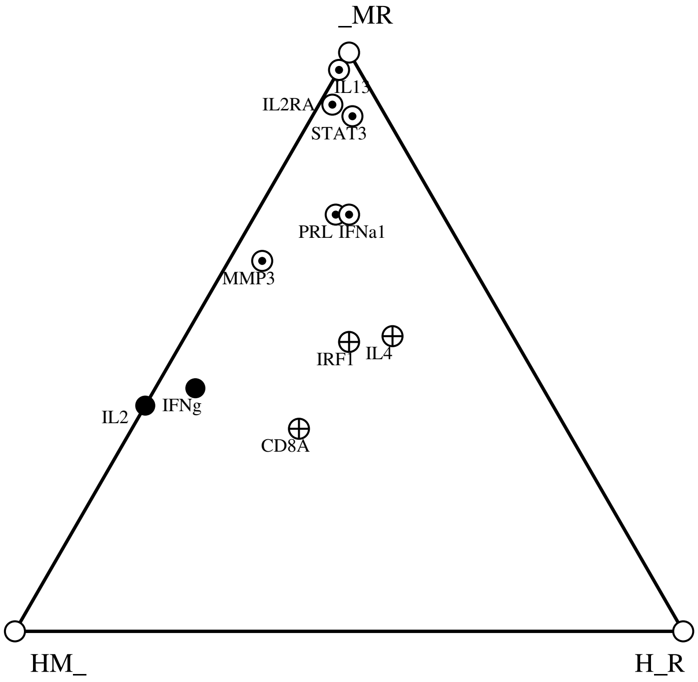



Figure 3.14. The frequency of the three different patterns *HM_*, *H_R* and *_MR*. The clusters that are conserved in all sequences are excluded.
• the most frequent pattern is *HM_*.
⊙ the most frequent pattern is *_MR*.
⊕ all patterns occurr equally in the sequences.

In Fig. 3.14 are the frequency of the three different patterns (*HM_* (footprints shared exlusively by human and mouse), *H_R* (footprints shared exlusively by human and rat) and *_MR* (footprint shared exlusively by mouse and rat) plotted. The clusters that are conserved in all sequences are excluded. The figure shows that the pattern *_MR* tend to be the most frequent. For CD8A, IRF1 and Il4 all patterns emerge approximately equal. It is obvious from Fig. 3.14 that human and rat have the fewest footprints in common followed by human and mouse. This findings are not surprising in that they correlate with the acepted phylogenetic relationship of these organisms.

Many kinds of differences among people have a genetic basis alterations in the DNA that change the way important proteins are made. Sometimes the alterations involve a single base pair and are shared by many people. Such single base pair differences are called “single nucleotide polymorphisms” (SNPs).

Polymorphisms have the potential to alter protein functions in ways that are biologically or clinically important. As increasing numbers of polymorphism are identified in regulatory regions of genes [81] it is of great interest for the unterstanding of diseases to combine the results of our `tracker` program with the available data of known SNPs. For this purpose we checked the frequency of SNPs (data about the SNP position for the chemokine receptor 5 *CCR5* gene were provided by Peter Ahnert) in the phylogenetic footprints of the *CCR5* gene. *CCR5* is a member of a chemokine receptor family expressed by T cells and macrophages with a typical seven transmebran domains. *CCR5* recruit cells of the immune system to the site of tissue damage or disease [1]. As can be seen in Fig. 3.15 and Tab. 3.12 only one SNP lies in an phylogenetic footprint. Given that functional elements are under stabilzing selection during evolution the frequency of SNPs in phylogenetic footprints should be significantly lower than in the surrounding nonfunctional DNA sequence.

Table 3.12. Frequency of SNPs in the different functional or nonfunctional sequence of *CCR5*. Abbreviations:
PF ... phylogenetic footprints
NCR ... non-coding region
FT ... result of Fisher's Exact Test

| Region | #SNPs | Nt | Nt - SNP |
|---|---|---|---|
| PF | 1 | 1143 | 1142 |
| NCR | 30 | 11256 | 11226 |
| Gen | 8 | 3655 | 3647 |
| FT PF-NCR | | 0.36 | |
| FT NCR-Gen | | 0.71 | |

We compared the frequency of SNPs on conserved elements of the human sequence to the frequency of SNPs in the noncoding regiom and in coding region. The most interesting outcome of this analysis is that SNPs may be underrepresented in phylogenetic footprints since we would expect about 3 SNPs in the footprints but find only a single one, see Tab. 3.12. The interpretation of this outcome is that SNPs seems to be detrimental to function since the footprints are under stabilizing selection. One can argue that this should be true in genes as well however the coding sequence tolerate more mutations due the redundancy of the genetic code. Unfortunately, the amount of data is not large enough for a signifikant statistical support of this finding. For statistical analysis we used Fisher's exact test [40, 39], that can be used to test if the occurence of SNPs in the sequence dependent of the function or a region or is it independent. When the P-value is less then 0.05 there is a significant relationship, on the other hand, if the P-value is larger then 0.05 the possibility that the proportion of SNPs in footprints and SNPs in non-functional region have arisen by chance is too great. To deal with this problem the analysis should be expanded larger sets of genes.

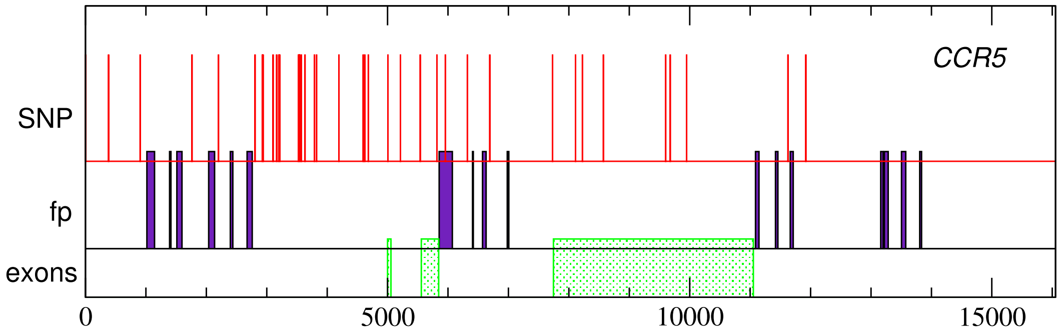


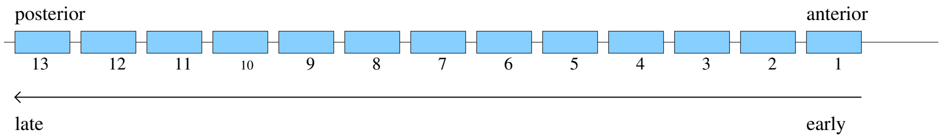

Figure 3.15. Graphical representation of the SNP distribution of the *CCR5* gene. Acccesion number: NT_0058250, the part of the sequence that we used in our analysis spans the entire gene and surrounding region 5000 basepairs upstream and downstream of the gene

Chapter 4

# New Insights Into the Evolution of *Hox* Clusters

## 4.1 *Hox* Genes

*Hox* genes were first characterized in the fruitfly *Drosophila melanogaster*. They code for homeodomain containing transcription factors which are homologous to the genes in the homeotic gene cluster of the fruitfly [78]. Characteristic for the genes of the homeobox family is a 60 amino-acid helix-turn-helix DNA binding homeodomain, see Fig. 4.1. The genes appear clustered on the genome whereby the position of a gene on the *Hox* cluster correlates with its expression in Anterior-Posterior or axial domain in many embryonic tissues [92]. This property is termed collinearity, and is conserved in arthropods and vertebrates suggesting that the regulatory mechanisms for controlling the spatially restricted domains of *Hox* expression are important features in maintaining the organization of these gene clusters. *Hox* clusters consist of genes from thirteen paralogous groups where the genes of a paralogous group are related through genome duplication. Members of a paralogous group show more similarity to paralogous genes on other clusters than to genes within the same cluster. None of the vertebrate *Hox* clusters contain representatives of all 13 paralogous groups. For example the four mammalian clusters consist only of 39 genes due to the fact that genes are lost as a result of genetic redundancies [78]. See Chapt. 1 for further details. *Hox* genes have a basic function in the specification and interpretation of positional information in the embryo by specific combinations of genes that are expressed in a certain position at a temporal level. They operate in a regulatory cascade and therefore play an important role in the determination of the identity along the anterioposterior axis, see Fig. 4.2.



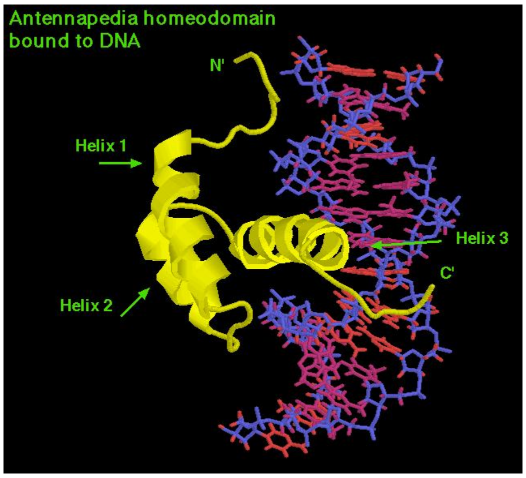




Figure 4.1. The *Hox* gene cluster consists of 13 paralogous genes which are related through genome duplication. Anterior *Hox* genes are expressed first and then the other genes are progressively expressed. Characteristic for *Hox* genes and other members of the homeobox gene family is a 180 bp helix-turn-helix motif the so called homeobox.

The regulation of *Hox* genes or the activity of *Hox* genes is modulated throughout development by local signals, hormone receptors and many any other stimuli that are able to mediate gene regulation. Small changes particularly in the structure of their regulatory elements can change the phenotype of segments and the morphology of an organism. It is likely that the increase in morphological complexity in the evolution of vertebrates is associated with a boost in the number of genes in the genome of the vertebrates compared to non-vertebrates genome. Therefore duplications of whole genome seem to have played an important role in the evolution of vertebrates [59, 99, 53].

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 [44, 59]. This ancient cluster is supposed to have arisen by the tandem gene duplication from a more ancient hypothetical protohox cluster see Fig. 4.3 [37].

It is still not resolved which of the vertebrate cluster is the most ancient and gave

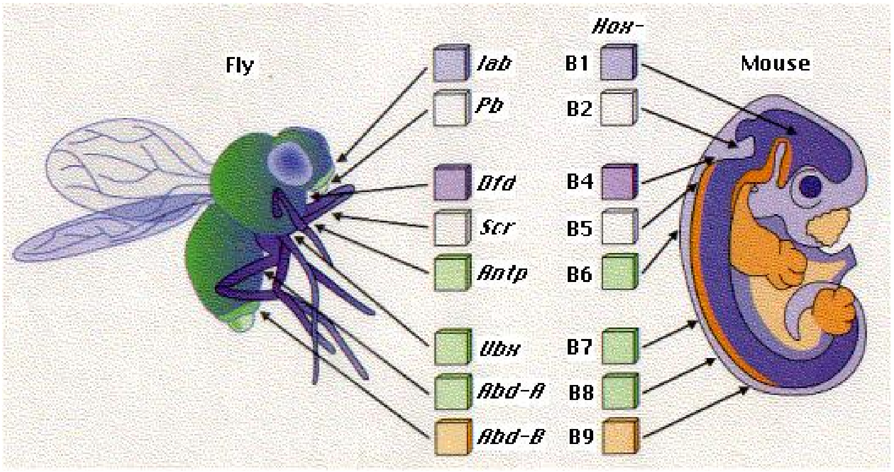



Figure 4.2. *Hox* genes specify the identity along the anterioposterior axis of an organism. Anterior *Hox* genes (*1-2*, *lab* and*bp*) pattern structures at the front end of an animal, central *Hox* genes (*3-8*, *zen*, *dfd*, *scr* ,*ftz*, *Antp*, *Ubx* and *abdA*) give rise to structures in the middle of an organism and posterior genes (*9-13*, *AbdB*) dictate posterior structures.

rise to the other clusters. One of the most accepted rules is the one-two-four rule or 2R for "two rounds" hypothesis of gene duplication. This model suggest that the the genome underwent two rounds of duplication leading from a single cluster to two clusters after first duplication and to four clusters after the second duplication event. The first duplication leads to the formation of a proto-AB cluster and a proto-CD cluster ((A,B)(C,D)). A third duplication event in the actinopterygian lineage is assumed to lead to the increased number of clusters in species of the actinopterygian lineage.
The alternative model assumes that the four mammalian clusters were generated by three rounds of sequential duplications leading to a *HoxD* and a *proto-ABC* cluster in the first place (D(A(BC))). Further, the second duplication provides a *HoxA* and a *proto-BC* cluster and the third duplication produces a *HoxC* and a *HoxB* cluster [4]. This model would imply that either four of the eight clusters arising by three duplication steps would have been lost or that only single clusters were duplicated [8]. Fig. 4.4 shows the tree representations of the 2R model and the model of sequential duplication.

The cephalochordate amphioxus possesses only a single *Hox* cluster. Therefore the duplications must have occurred after the divergence of the cephalochordate lineage and the vertebrate lineages. Two duplication events lead to four *Hox* clusters in mammals. These four *Hox* clusters are designated $A$, $B$,$C$ and $D$ and are located on four different

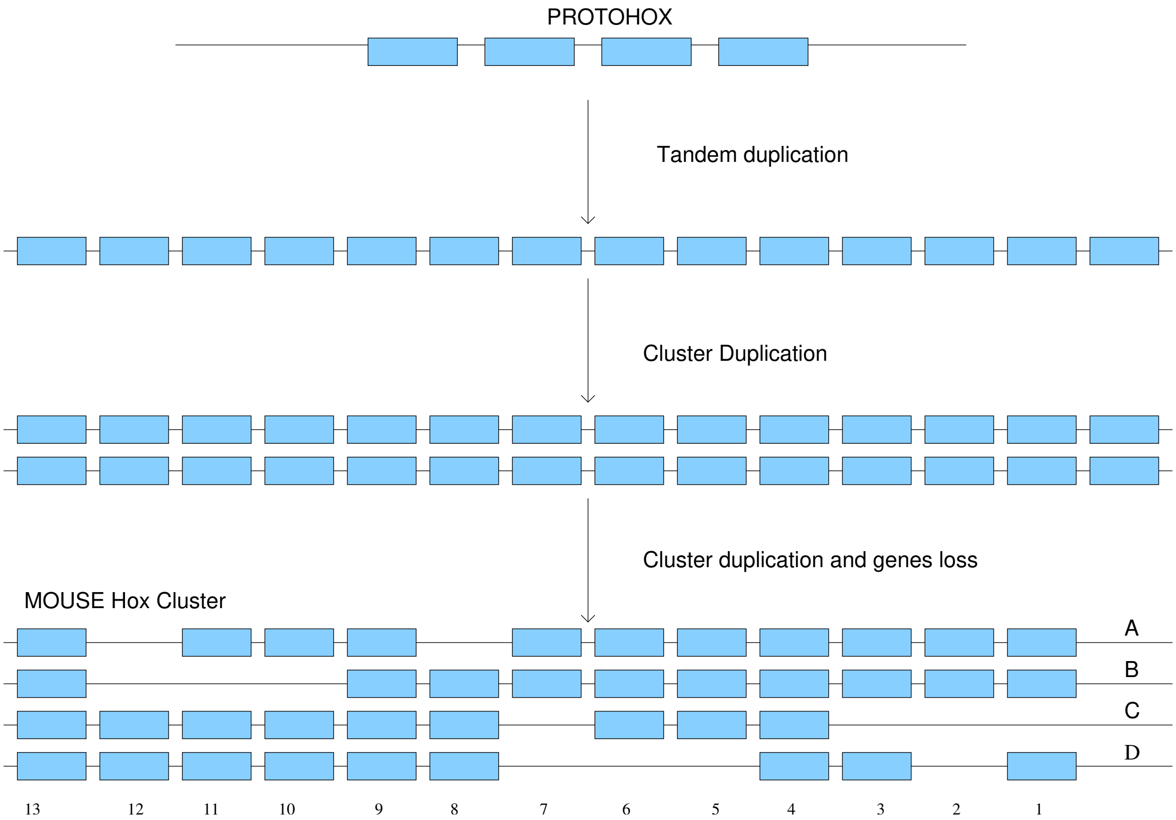



Figure 4.3. Tandem duplication of an hypothetical proto *Hox* cluster leads to the formation of the recent *Hox* clusters. The ancient *Hox* cluster of the nonvertebrate species was subject to two rounds of duplication to give rise to the mammalian *Hox* clusters. The organization of the 39 mouse *Hox* genes is shown.

genomes [78]. Furthermore, an additional duplication event in the teleost lineage leads to an increased number of distinct clusters. There are at least 7, e.g. in zebrafish *Danio rerio* [4]. The acanthopterygian teleost medaka *Oryzias latipes* also has at least 7 *Hox* clusters according to a recently generated linkage map and pufferfish which has at least six clusters [95, 6, 4]. The teleosts are the group of vertebrates that show greatest diversity in bodyplans.

Although it is known that duplications gave rise to the known *Hox* clusters it is not known at which point of time the duplication events occur. One theory postulates that the duplications leading to the four mammalian *Hox* clusters must have occurred before the divergence of the agnathan and gnathostome lineages. If this is correct, the clusters of lamprey and the vertebrates must be real homologous. Otherwise, at least one of

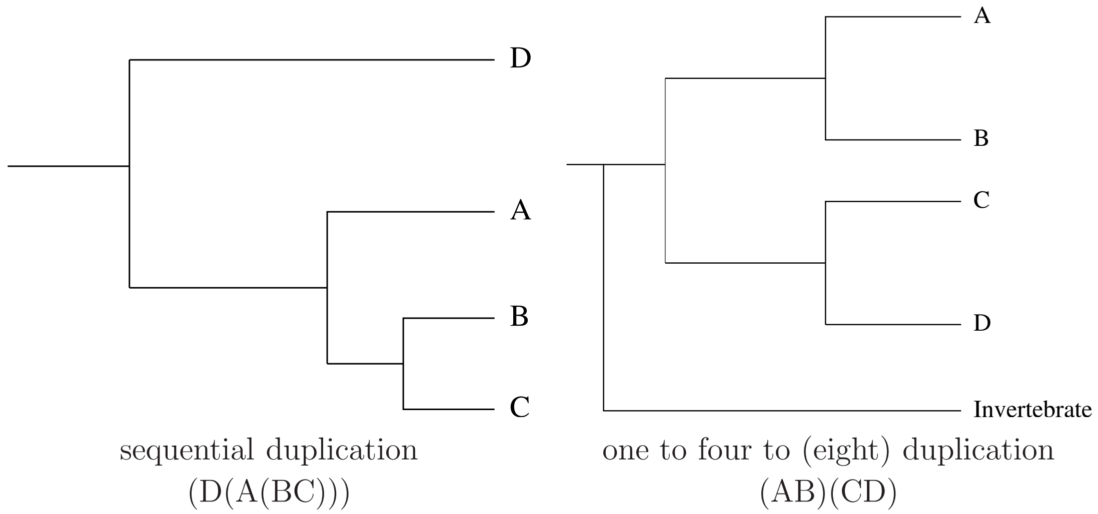



Figure 4.4. Two models used to describe the evolution of vertebrate *Hox* gene clusters. On the left side the socalled sequential duplication assuming that three sequential duplication events lead to formation of four *Hox* clusters. According to the model on the right side the mammalian clusters were formed through two rounds of duplications.

the duplications took place after the branching of the two lineages [41] see Sect. 4.4 and Fig. 4.10 for details. To test these hypotheses we investigated in this study the relationship of the lamprey *Petromyzon marinus*, a surviving member of the agnathan lineage and some representatives of the gnathostome lineage. We found no evidence for direct relationship between vertebrate *Hox* cluster and lamprey *Hox* cluster. This supports the hypothesis that the common ancestor of the agnathans and gnathostomes had only a single *Hox* cluster which was subsequently duplicated independently in both lineages. See Chapt. 4.4 for details.

The evolution of the morphological characters is basically realized through the modification of the genes that control their development. An explanation of morphological evolution thus requires the study of the molecular evolution of developmental genes such as *Hox* genes [96, 50]. Experimental evidence from a variety of sources shows that a major mode of developmental gene evolution additional to the increasing numbers of genes is based on the modification of cis-regulatory elements, see e.g. [7, 22, 28, 111]. The investigation of the molecular evolution of these cis-regulatory elements is difficult because of the absence of a reliable "genetic code for non-coding sequences". Binding sites for transcription factors are usually short and variable and are thus hard to identify unambiguously, in particular if the transcription factors involved are not known [113, 71]. It has been noted for a long time, however, that non-coding sequences can contain islands of strongly conserved segments, so-called phylogenetic footprints [112]. Hence it is possible in principle to gain insights into the extent and the phylogenetic timing of major changes in the regulation of a gene by studying the phylogenetic pattern of non-coding sequence conservation. A cluster of phylogenetic footprints which is present in an outgroup clade but not in an ingroup can be evidence for the modification or the complete loss of a cis-regulatory element. On the other hand a set of phylogenetic footprints that is uniquely shared by a nested clade can provide evidence for the acquisition and subsequent conservation of a cis-regulatory element.

It is known that there is a positive correlation between the *Hox* gene number and the morphological complexity which means that changes in these gene clusters and in the regulation of the genes play an important role during evolution. It has been shown recently that the duplication of *Hox* clusters leads to a massive loss of non-coding sequence conservation. Chiu et al. [25] showed that there is a massive change in the cis-regulatory pattern of the duplicated *Danio rerio HoxA* clusters compared to the single *HoxA* cluster of shark and human. In our study we analysed the regulation pattern shared between some groups of teleosts and bichir *Polypterus senegalus* a basal ray-finned fish in comparison with human and shark. We find a loss of footprints in the teleost lineage consistent with the results of the previous study [25]. See Sect. 4.2 for details.

*Hox* cluster duplication can lead to extensive loss of non-coding sequence conservation, as shown by Carter et al. [23], but the causes remain unclear. As result of our study we can see an extensive loss of non-coding sequences conservation which is suitable to this earlier findings. Furthermore it is known that following cluster duplication the *HoxA* clusters of the teleosts undergo further extensive remodeling including gene loss and shortening of intergenic sequences [25]. We can show in our analysis that the duplicated *HoxAa* and *HoxAb* clusters of zebrafish and pufferfish have lost approximately between 60-80% of the footprints that are present in human, shark and bichir. About 30-40% of the PFCS have no counterpart in either one duplicate cluster. There are three biologically distinct processes that can account for this phenomenon: (1) structural loss which means the loss of cis-regulatory elements due to gene loss and/or stochastic resolution of genetic redundancy after an duplication event, (2) binding site turnover where the conservation is lost although the function of the element remains, and (3) adaptive modification which means a change in the sequence of cis-regulatory sites due to directional natural selection and would thus be associated with functional differences. A systematic study of non-coding sequence conservation after duplication thus requires a stochastic model to estimate the amount of sequence conservation loss due to the simple loss of some genes and the loss of cross-regulatory links. Günther Wagner recently proposed a model in order to estimate the amount of conservation loss that can be attributed directly to gene loss and to determine whether other factors, such as adaptive evolution or binding site turnover, might also have played a role [95]. The model builds on the assumption that the evolution of footprints can be expressed in terms of retention probabilities. The total 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, $r(\mathrm{G})$:

$$r(\mathrm{F}) = r(\mathrm{F}|\mathrm{G})r(\mathrm{G})\,. \tag{4.1}$$

The footprint retention probability based on structural causes is estimated by

$$\begin{aligned} r_0 &= \left[\frac{1}{2}P(1^{\mathrm{st}}) + \big(1 - P(1^{\mathrm{st}})\big)\right]\big(1 - d\,P(\mathrm{G_{ext}})\big) \\ &= \big(1 - \frac{1}{2}P(1^{\mathrm{st}})\big)\big(1 - d\,P(\mathrm{G_{ext}})\big)\,. \end{aligned} \tag{4.2}$$

$P(\mathrm{G_{ext}})$ is the fraction of genes in the whole network that were lost
$d$ is the fraction of genes in the network which received regulatory input from these extinct genes.
*1st order paralogs* are genes that are related by the most recent gene/cluster duplication. Genes that retain 1st order paralogs are expected to resolve the genetic redundancy by, on average, losing 50% of their cis-regulatory inputs [42]. The probability that a footprint is lost because of stochastic resolution of genetic redundancy is equal to the probability that the associated gene has a 1st order paralogous times 1/2. If only one copy of the gene survives all relevant cis-regulatory elements are maintained, i.e., there is a contribution of $1 - P(1^{\mathrm{st}})$ to the retention probabilities.

The footprint loss due to non-structural causes (binding site turnover and adaptive effects) is given by the probability $\alpha$. The total retention rate of footprints is therefore:

$$\hat{\alpha} = 1 - \frac{r(\mathrm{F}|\mathrm{G})}{\big(1 - P(1^{\mathrm{st}})/2\big)\big(1 - P(\mathrm{G_{ext}})\big)} \tag{4.3}$$

$\hat{\alpha}$ ... minimal estimate for the degree of non-structural loss of phylogenetic footprints.
$r(R|G)$ ... retention rate of footprints per gene.
$P(1^{\mathrm{st}})$ ... fraction of 1st order paralogs.
$P(\mathrm{G_{ext}})$ ... gene extinction rate
$d$ ... degree of cross-regulatory connectivity.

It is assumed that $d = 1$ in the case of *Hox* genes because each gene has a cross-regulatory link to every other gene.

## 4.2 Surprising Trends in the Evolution of Ray-finned Fishes

The work reported in this section is a cooperation with Chiu *et al.* [26]. The bichir sequence was obtained by screening a PAC library. The whole experimental analysis was performed in Chi-Hua Chiu's laboratory. The computational analysis of the phylogenetic footprint data was performed in Leipzig.

## 4.3 Evolution of Teleosts

We applied `tracker` to the analysis of the *HoxA* genes clusters of the horn shark *Heterodontus francisci*, human *Homo sapiens*, striped bass *Morone saxatilis*, zebrafish *Danio rerio*, pufferfish *Takifugu rubripes* and bichir *Polypterus senegalus* to study the changes in the organization of the sequence and the regulation pattern of the *Hox* genes caused by the duplication of the whole clusters. Bichir on the one hand has morphological characters that place him him to the actinopterygians e.g. the scale structure and cranial ossification. On the other hand bichir has a lung and fleshy pectoral fins similar to the ones of the lungfish, a sarcopterygian fish [87]. Fig. 4.6 shows an overview of the vertebrate taxonomy. The phylogenetic tree displays the most popular hypothesis placing the bichir as a basal actinopterygian as inferred from independent molecular datasets [67, 87, 117]. For example phylogenetic analysis of the mitochondrial protein-coding and ribosomal RNA genes of bichir brought evidences that there is greater relationship to ray-finned fish than to lamprey or lungfish. Zebrafish, pufferfish and striped bass on the other hand are derived teleosts. All of these species and bichir are representatives of the actinopterygian lineage of the vertebrates. Shark is a representative of the jawed vertebrates. Vertebrates exhibit between three to at least seven *Hox* clusters [110]. The clusters are similar, having been duplicated during the course of evolution. Human and bichir contain four *Hox* clusters. Shark has four clusters but only two of them have been sequenced. In the lineage of the teleosts a third duplication occurred. Therefore zebrafish has seven *Hox* clusters and the pufferfish has six clusters, Fig. 4.5. Bichir is a basal actinopterygian but exhibits only one *HoxA* cluster in contrast to the other actinopterygian fish. Therefore bichir is an interesting species to include in the study of *Hox* cluster evolution. The other species were chosen because of the availability of sequences or because of their capacity as good model organism. The zebrafish has been used as a model for vertebrate development for 20 years because of its small size, relatively rapid life cycle and easy breeding [125].

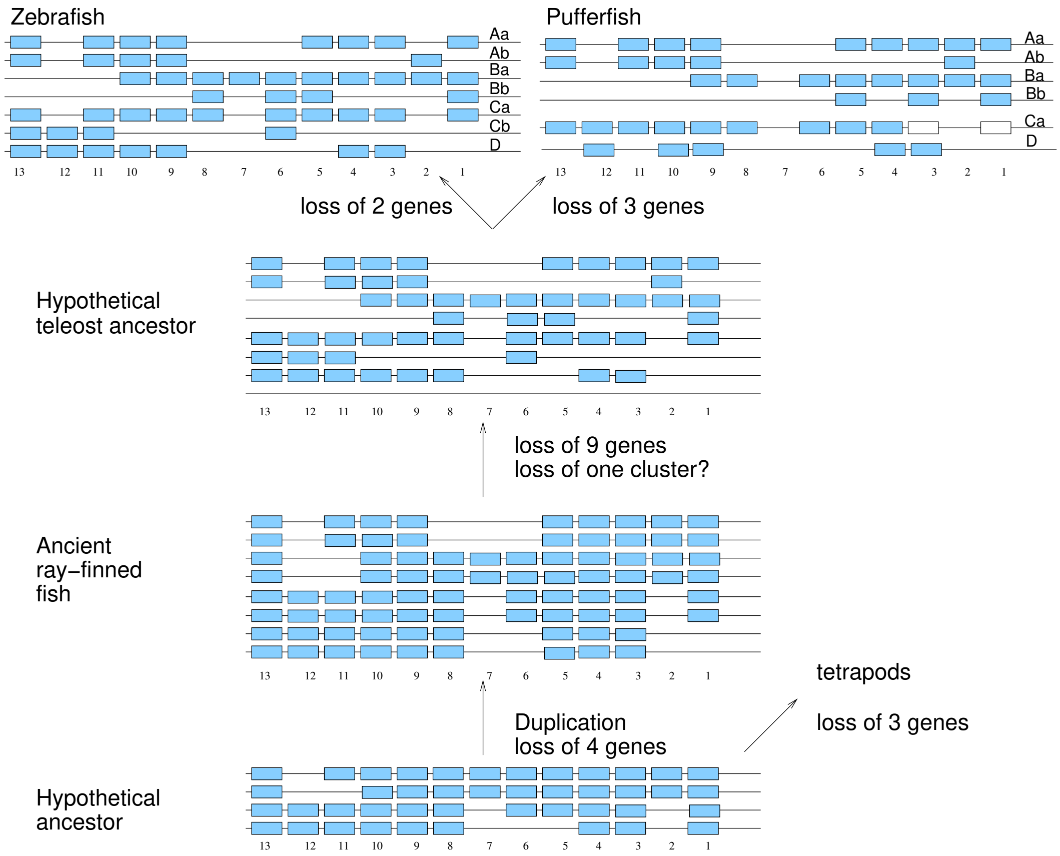



Figure 4.5. During the development of vertebrates from early deuterostoma entire genomes where duplicated through two rounds of duplication. Teleosts exhibit a third duplication which occurred after the two major lineages of vertebrates, the ray-finned fish (actinopterygian) and lobe finned fishes (sarcopterygian) diverged. Following duplication some genes can get lost or acquire new functions and therefore the distribution pattern of regulatory elements may undergo massive changes. Furthermore the *Hox* gene organization of zebrafish and pufferfish is shown in detail (unfilled rectangles denote pseudogenes).

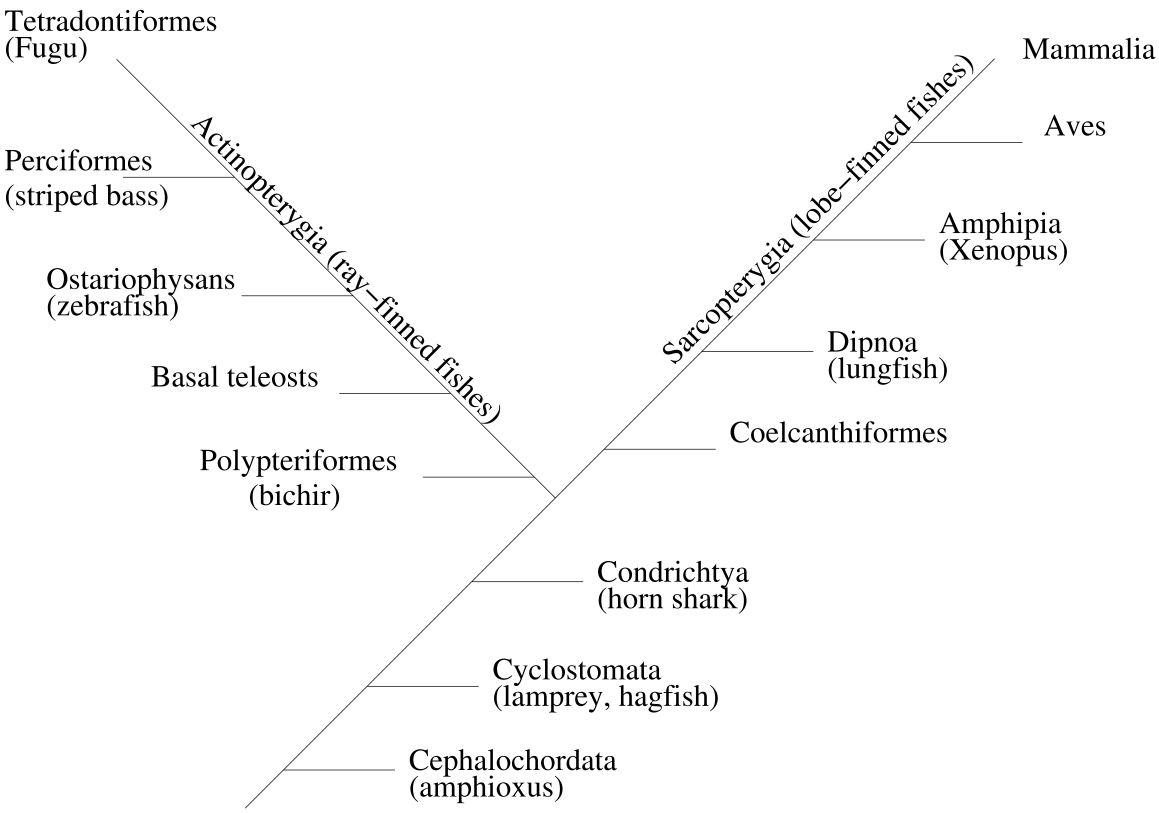



Figure 4.6. Overview over the vertebrate phylogeny. Soon after the split of the sarcopterygian and actinopterygian lineage a duplication occurred in the actinopterygian lineage. Bichir is one of the most basal actinopterygians. It has not acquired additional *Hox* clusters.

Zebrafish shares many orthologous genes with mammals which gives it considerable relevance in comparison to other models used for developmental studies. The pufferfish is a good model vertebrate for comparative genomics because its genome is small relative to other vertebrates (7.5 times smaller than the human genome) while it contains approximately the same set of genes as other vertebrates. The evolutionary distance between mammals and the bony fishes is about 450 Myrs. This timespan is sufficiently large for the accumulation of mutations but enough that regulatory elements are still remained. Therefore the distinctions that arose in the regulation pattern during evolution of this species are observable. Including bichir into the analysis gives the opportunity to determine basic principles of the massive loss in the non-coding region of the teleosts. The fundamental question is: Is the loss of conservation typical for all actinopterygians or are the changes caused only by the duplication event? The

phylogenetic position of the bichir as a basal lineage of the actinopterygians is still uncertain. Also analysis of the mitochondrial genome placed the bichir in between the actinopterygian and the sarcopterygian fishes [87]. We concentrate on the analysis of *HoxA* clusters due to the fact that only for these clusters sufficient data is publicly available.
Furthermore we developed an appetite for the investigation of edible fish species in the examined group of vertebrates. Therefore we collected some recipes, listed in Sect. B of the appendix.

### 4.3.1 Sequence Data

Sequences for *HoxA* clusters were obtained from Genbank, the Fugu database [30], and the web pages of the Zebrafish Sequencing Project [101]. Accession numbers of the sequences are listed in Tab. 4.1.

Table 4.1. Sequence data used in this study. $^{rc}$ denotes reverse complements of the database entries. AC010990 overlaps exactly 200nt with both AC004080 and AC004079.

| Sequence | Name | Source |
|---|---|---|
| *Homo sapiens* | HsA | AC004080$^{rc}$ + AC010990$^{rc}$ + AC004079[75001-end]$^{rc}$ |
| *Heterodontus francisci* | HfM | AF479755 |
| *Polypterus senegalus* | PsA | AC135508 |
| *Morone saxatilis* | MsA | AF089743 |
| *Danio rerio* A$\alpha$ | DrAa | AC107365$^{rc}$ |
| *Danio rerio* A$\beta$ | DrAb | AC107364$^{rc}$ |
| *Takifugu rubripes* A$\alpha$ | TrAa | Fugu v.3.0 scaffold_47[103001-223000]$^{rc}$ [5] |
| | `http://genome.jgi-psf.org/fugu6/fugu6.home.html` | |
| *Takifugu rubripes* A$\alpha$ | TrAb | Fugu v.2.0 scaffold_1874 [5] |
| | `http://genome.jgi-psf.org/fugu3/fugu3.home.html` | |

### 4.3.2 Gene Positions

The `tracker` program requires a list of the genes as input since only the intergenic regions between two homologous genes are compared. The information about gene positions was not available for the sequences of pufferfish since the annotation of the genes in the Fugu database
`http://genome.jgi-psf.org/fugu3/fugu3.home.html` is still incomplete. To obtain the information about gene positions, the pufferfish sequences TrAa and TrAb have

Table 4.2. Positions of *HoxA* genes in the sequence files used in this study.

| Genes | HfM | | HsA | | PsA | | MsA | | DrAa | | FrAa | | DrAb | | FrAb | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| evx1 | | | 9561 | 12955 | 26897 | 29924 | | | 8554 | 13818 | 2938 | 5152 | | | 1305 | 4340 |
| *Hox13* | 16296 | 17590 | 55909 | 57785 | 130314 | 131696 | | | 23571 | 24620 | 11576 | 12865 | 59165 | 60421 | 7498 | 8502 |
| *Hox12* | | | | | | | | | | | | | | | | |
| *Hox11* | 29986 | 32024 | 70842 | 73187 | 141825 | 144111 | | | 31267 | 32814 | 20875 | 22503 | 68178 | 69778 | 13556 | 15305 |
| *Hox10* | 43461 | 45781 | 81731 | 84084 | 149716 | 151645 | 176 | 2050 | | | 27585 | 29117 | 73569 | 75911 | 18810 | 20267 |
| *Hox9* | 53330 | 55234 | 90529 | 92380 | 157007 | 158246 | 6473 | 7675 | 44124 | 45169 | 32999 | 34150 | 78790 | 80322 | 22458 | 23453 |
| *Hox8* | | | | | | | | | | | | | | | | |
| *Hox7* | 62339 | 63947 | 99441 | 101077 | | | 11642 | 13276 | | | | | | | | |
| *Hox6* | 71938 | 72829 | 108237 | 110325 | 168721 | 169912 | | | | | | | | | | |
| *Hox5* | 74807 | 76107 | 112379 | 114148 | 171229 | 172687 | 18701 | 20200 | 53666 | 54856 | 42823 | 44199 | | | | |
| *Hox4* | 87863 | 88955 | 125253 | 126761 | 179635 | 180607 | 29109 | 30386 | 61628 | 62827 | 49044 | 50320 | | | | |
| *Hox3* | 106363 | 109634 | 145346 | 148071 | 192215 | 195930 | | | 71733 | 73572 | 59731 | 62635 | | | | |
| *Hox2* | 114242 | 115852 | 153486 | 155260 | 200129 | 201690 | | | | | 66537 | 68146 | 90839 | 92515 | 28480 | 29953 |
| *Hox1* | 120174 | 121584 | 160074 | 161546 | 205589 | 206994 | | | 80301 | 81442 | 71274 | 72563 | | | | |
| *Snex* | | | | | | | | | | | 117327 | 118025 | | | 46935 | 47348 |

been prepared from the consensus of `tblastn` alignments with as many known *Hox* protein sequences from related species as possible against version 2.0 of the Fugu database `http://genome.jgi-psf.org/fugu3/fugu3.home.html` and version 3.0 of the Fugu database `http://genome.jgi-psf.org/fugu6/fugu6.home.html`. To prove the assembly of the Takifugu *Hox* genes we performed an automatized `clustalW` comparison against different known *Hox* proteins. For this purpose we implemented a comparison tool in `perl` which takes as input a list of files containing *Hox* protein sequences. This input list is then automatically compared with the query sequences and all pairwise comparisons are computed. The output of this tool is an `html` page where all percent identities of the query sequences with the comparison group of known genes are listed. The best hit and hits that lie in a range of five percent around the best hit are listed at the end of the produced table. Hits are colored on the basis of the percent identity for better clarity, see Fig. A.1. This method can easily be extended to other protein or gene sequences as it only needs a list of input files for the comparison. In addition to the pairwise `clustalW` comparisons one can suggest which gene the query sequence seems to be and expand the input of the program with this information. In this case the program searches through the input list, extract all genes of the same paralogous group and computes a Buneman graph using `splitstree.3.1`.

It is planned to extend the program in a way that it can take a list of Accession numbers and start an automatic download of the requested sequences from Genbank. The appropriate tool for downloading genbank entries is at present time implemented as external tool, merging of both programs is an easy task but still left to be done. Furthermore, DNA sequences can be translated into protein sequences. The problem in this case is that available tools like `nt2aa.pl` used here do not exclude introns from the translation into protein sequences. Therefore the percent identity score of the alignments is inferior to the score that could be obtained by comparing only the coding sequence exclusively. The gene positions in all the sequences used in this section are compiled in Tab. 4.2.

### 4.3.3 Phylogenetic Footprints

Tab. 4.3 lists experimentally determined binding sites from the literature.

To test the efficiency of `tracker` we checked the results of our analysis for four experimentally evaluated protein binding sites lying in the orthologous region from *HoxA4* to *HoxA3* which were described recently [74, 89]. The results of `tracker` and the corresponding protein binding site are listed in Tab. 4.3.

Table 4.3. Experimentally known binding sites

| Binding site | cluster | HfM | HsA | PsA | MsA | DrAa | FrAa | DrAb | FrAb | Ref. |
|---|---|---|---|---|---|---|---|---|---|---|
| -2.9 RARE | 418 A5-4f | 84903 | 122066 | 177562 | 27265 | 59820 | 47414 | | | [89] |
| -2.9 RARE | 418 A5-4f | 84914 | 122077 | 177573 | 27275 | 59831 | 47425 | | | [89] |
| KrA | 459 | 101863 | | | | | 56856 | | | [74] |
| Hox/PBC B | 461 | 102087 | 140697 | | | | | | | [74] |
| Prep/Meis | 461 | 102112 | 140722 | | | | | | | [74] |

### 4.3.4 Footprint Cluster Summary Statistics

Tab. 4.4 lists the number of PFCs *sensu* Chiu *et al.* [25] that each pair of sequences has in common. We count only footprint clusters that contain at least one reliably detected footprint. Clusters are separated by horizontal lines in the first column of the list of all footprints (Tab. A.1). Footprints taken into account for the summary statistics are marked with • in the first column. The footprints are combined into a cluster if they are separated by less than 100nt in at least two sequences. Footprints that violate collinearity are disregarded (marked by × in the last column of the list). Several patterns of conservation are evident. (1) The *HoxA* clusters of human and horn shark share the largest number of phylogenetic footprints, 49. (2) The bichir *HoxA* clusters share 44 phylogenetic footprints with horn shark and 40 with human. Bichir as well shares 45 footprints with both *HoxA* clusters of the zebrafish and 43 footprints with the clusters of the pufferfish. Strikingly, the consensus between bichir and zebrafish or pufferfish is not greater than the consensus with human or shark.
(3) Neither shark nor human share as many footprints with the teleosts as bichir does. The analysis of footprints shared exclusively between certain sets of sequences shows a similar pattern, see Tab. 4.5. (1) Bichir shares eleven footprints with the teleosts (2) whereas human shares only two with the teleosts. (3) 24 Footprints are shared exclusively among the different teleosts. It can be assumed that some of these footprints, if not all, are derived in the stem lineage of the teleost. The results implies that bichir — although one would expect that it shows more similarity with the other species of the actinopterygian lineage — has a pattern of phylogenetic footprints that lies in between the two great vertebrate lineages.

Table 4.4. Co-occurrence of PFCs.
The last row gives the total number of PFCs that occur in each cluster.

| | HfM | HsA | PsA | MsA | DrAa | TrAa | DrAb | TrAb |
|---|---|---|---|---|---|---|---|---|
| HfM | * | 49 | 40 | 14 | 24 | 31 | 20 | 11 |
| HsA | | * | 44 | 14 | 26 | 28 | 15 | 10 |
| PsA | | | * | 13 | 28 | 30 | 17 | 13 |
| MsA | | | | * | 14 | 25 | 5 | 2 |
| DrAa | | | | | * | 37 | 14 | 9 |
| TrAa | | | | | | * | 17 | 14 |
| DrAb | | | | | | | * | 11 |
| TrAb | | | | | | | | * |
| #PFCs | 66 | 55 | 58 | (26) | 42 | 59 | 32 | 18 |

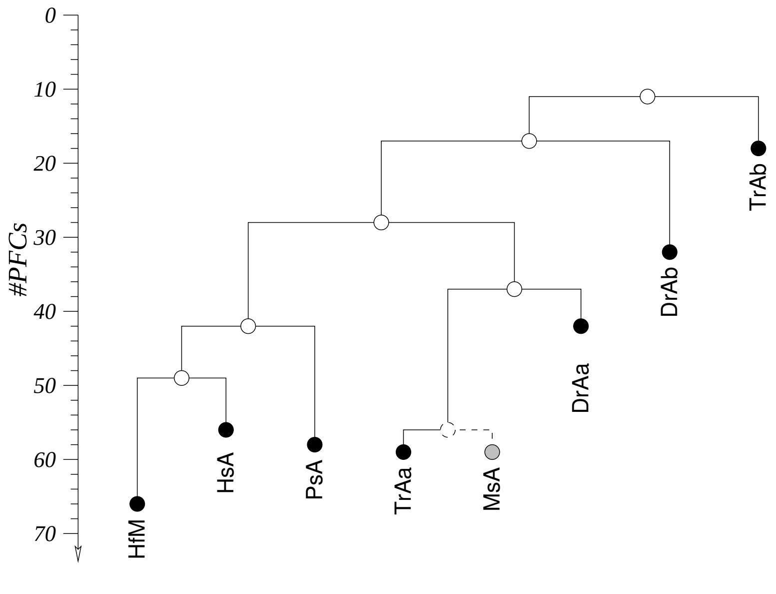



Figure 4.7. Tree representation of the co-occurrence data of Tab. 4.4 obtained using a weighted pairgroup similarity clustering: the height of each internal node is the average number of co-occurring footprints in the sequences located in the two subtrees.

### 4.3.5 Footprints as Phylogenetic Signal

For a phylogenetic analysis we used all 188 co-linear tracker hits marked with • in Tab. A.1 and essentially created a 01-string for each of the *Hox* clusters that denotes the presence/absence pattern of the tracker hits. These 0-1 strings (= presence/absence of characters data) are then used to construct phylogenetic trees Fig. 4.8. We used Canonical split decomposition [10] and parsimony splits [11]. These methods are implemented in the `splitstree` package (version 3.1) [56]. The split-based methods are particularly suitable for our purposes because they are known to be very conservative in that they tend to produce multifurcations rather than poorly supported edges [104].

Treating phylogenetic footprint cliques as presence/absence characters in a parsimony framework also supports the hypothesis obtained by the analysis of phylogenetic footprints. It shows evidence that the bichir *HoxA* cluster is an intermediate state in between the duplicated *HoxA* clusters of the derived teleost and the single *HoxA* clusters of shark and human.

### 4.3.6 Footprint Loss Statistics

Tab. 4.6 shows the footprint retention statistics after *Hox* cluster duplication based on the predictions of the structural loss model. The results of this analysis indicate for an excess loss of conservation in the non-coding sequence in the duplicated *HoxA* clusters. The predicted retention rate based on the structural loss model is consistently higher than the observed rates. The loss is higher than expected by the loss of associated genes due to other factors like adaptive modification and binding site turnover. We observe retention rates of 0.44 for zebrafish *HoxAa* cluster, 0.52 for zebrafish *HoxAb* cluster, and 0.47 for both. The overall rate for the pufferfish is 0.39, for the pufferfish *HoxAa* cluster we find a rate of 0.43, and 0.34 for *HoxAb*. Binding site turnover has an equivalent effect on either paralogous cluster. Therefore the asymmetries seen in the retention probabilities between *HoxAa* and *HoxAb* clusters indicate that the *HoxAa* cluster is more modified than expected by structural and stochastic reasons. In Tab. 4.5 footprints/PFC are counted according to three definitions: (1) *all* `tracker` cliques listed in Tab. A.1, (2) `tracker` cliques that are *co-linear* (marked by • in Tab. A.1), and (3) PFCs *sensu* Chiu [25] (blocks separated by horizontal lines that contain at least one •). Only the data between *evx-1* and *HoxA1* are taken into account. Counting method (3) is the same that was used for the co-occurrence statistics in Sect. 4.3.4. It is obvious from the results in Tab. 4.6 that counting PFCs sensu Chiu *et al.* [25] — where lots of • hits are collapsed into a single PFC — produce data that are not sensitive enough for statistical analysis due to undercounting of loss rates. Here we use only the counts between *evx-1* and *HoxA1* that are consistent with all other counts.

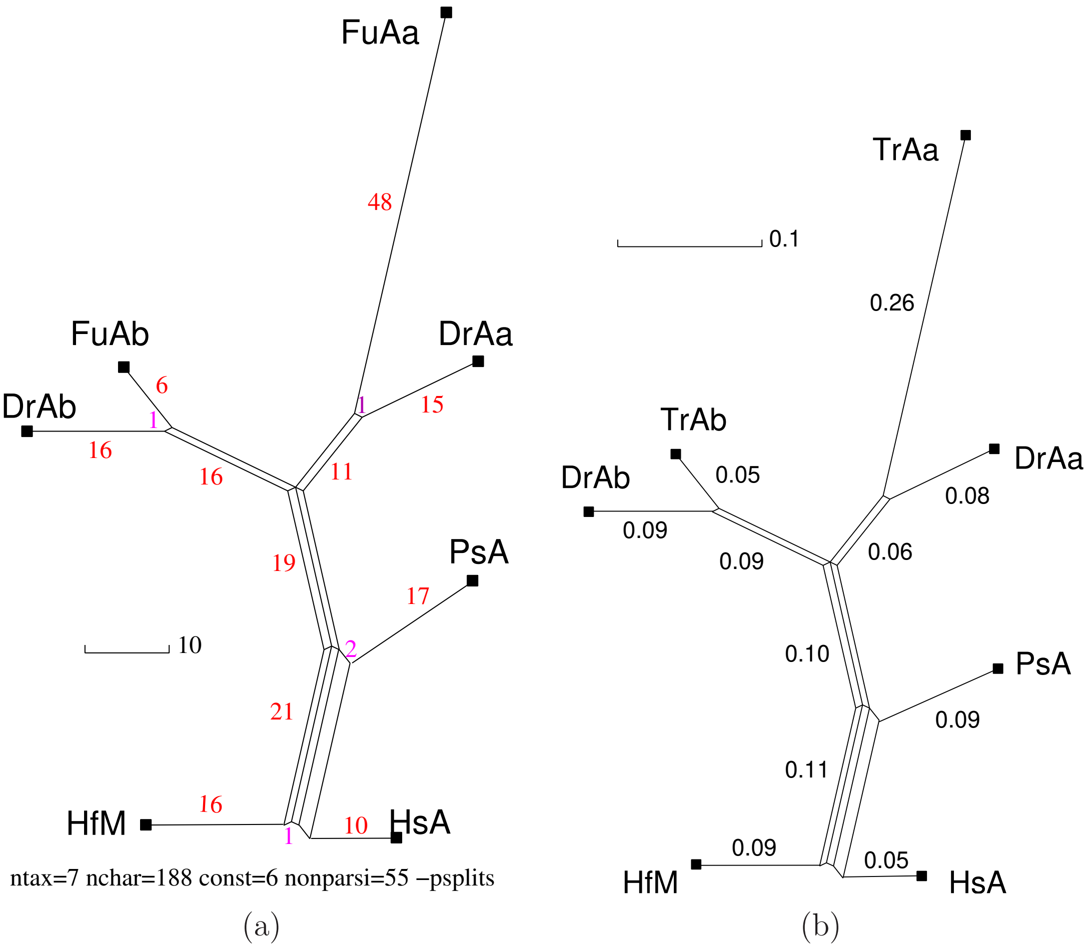



Figure 4.8. Buneman graphs of the presence/absence pattern of the 188 collinear cliques listed in Tab. A.1 obtained from two different split-based analysis methods. The drawings are obtained using `splitstree 3.1` [56].
(a) Parsimony-splits methods [11]. Numbers in red are the numbers of losses and gains that are unambiguously assigned to a tree-like edge of the graph. Number in magenta give box width indicating the number of gains/losses supporting the alternative hypothesis.
(b) Distance-based split-decomposition [10]. This method resolves 91.9% of the distance information to produce the graph displayed here. Edges are labeled by fractions of total pairwise distance.

Table 4.5. Footprint Counts. Footprints exclusively shared by different combinations of the species. Organisms in brackets separated by || denote that at least one of the listed species has this footprint. All species separated by a plus have these footprints in common.

| | all cliques | • cliques | PFCs |
|---|---|---|---|
| Shark and Human as outgroup, Bichir treated separately | | | |
| (HfM ‖ HsA)+PsA+Teleosts | 40 | 36 | 31 |
| (HfM ‖ HsA)+PsA | 37 | 28 | 16 |
| PsA+Teleosts | 15 | 12 | 11 |
| HfM+HsA | 27 | 26 | 11 |
| Teleosts only | 80 | 66 | 26 |
| (HfM ‖ HsA)+Teleosts | 33 | 20 | 16 |
| Total | 232 | 188 | 111 |
| Pufferfish clusters versus Shark Human and Bichir as outgroups | | | |
| (HfM ‖ HsA ‖ PsA) + Aa + Ab | 13 | 13 | 13 |
| (HfM ‖ HsA ‖ PsA) + Aa | 36 | 30 | 23 |
| (HfM ‖ HsA ‖ PsA) + Ab | 7 | 6 | 4 |
| (HfM ‖ HsA ‖ PsA) | 98 | 71 | 46 |
| Pufferfish + Zebrafish only | 78 | 65 | 25 |
| Total | 232 | 188 | 111 |
| Zebrafish clusters versus Shark Human and Bichir as outgroups | | | |
| (HfM ‖ HsA ‖ PsA) + Aa + Ab | 12 | 11 | 12 |
| (HfM ‖ HsA ‖ PsA) + Aa | 32 | 23 | 19 |
| (HfM ‖ HsA ‖ PsA) + Ab | 25 | 18 | 14 |
| (HfM ‖ HsA ‖ PsA) | 85 | 71 | 41 |
| Pufferfish + Zebrafish only | 78 | 65 | 25 |
| Total | 232 | 188 | 111 |

The results of the analysis of non-coding sequence evolution support the idea that the bichir *HoxA* cluster is orthologous to the single *HoxA* clusters in shark and human. However bichir exhibits a mosaic pattern of conservation of non-coding sequences with human and the derived actinopterygians. Uniquely conserved non-coding sequences between *HoxA* clusters of bichir and teleosts suggest that novel cis-regulatory elements were already acquired in the stem lineage of actinopterygians. On the other hand bichir does not show the extensive degree of remodeling the non-coding sequence observed in derived actinopterygians. Therefore the loss of conservation in the duplicated clusters of zebrafish pufferfish and striped bass is a derived pattern which can be examined only in derived actinopterygians. The phylogenetic analysis of the *HoxA* cluster coding sequences made by Chiu *et al.* support these findings. Furthermore it supports the idea that the duplication event leading to the duplicated *HoxA* clusters of the teleosts took place after bichir lineage branches off.

Table 4.6. Retention Statistics.
$g$ number of genes in cluster; $r$(G) retention rate of genes; $P(1^{st})$ fraction of first order paralogous genes; #pX number of plesiomorphic phylogenetic footprint cliques or PFCs, Tab. 4.5); $r$(X) retention rate of footprint cliques or PFCs; $r$(X|G) retention rate conditional on gene retention; $r_0$ predicted conditional retention rate assuming stochastic resolution of genetic redundancy, $\alpha$ lower bound on the rate PF loss due to non-structural reasons [95]. See text for details.

| Cluster | $g$ | $r$(G) | $P(1^{st})$ | #pX | $r$(X) | $r$(X\|G) | $r_0$ | $\hat{\alpha}$ |
|---|---|---|---|---|---|---|---|---|
| Raw `tracker` cliques excluding bichir | | | | | | | | |
| DrAa | 7 | 0.63 | 0.43 | 40 | 0.29 | 0.46 | 0.69 | 0.33 |
| DrAb | 5 | 0.45 | 0.60 | 32 | 0.23 | 0.51 | 0.62 | 0.18 |
| DrA | 12 | 0.55 | 0.50 | 72 | 0.26 | 0.47 | 0.66 | 0.29 |
| TrAa | 9 | 0.82 | 0.56 | 45 | 0.32 | 0.39 | 0.58 | 0.33 |
| TrAb | 5 | 0.45 | 1.00 | 17 | 0.12 | 0.27 | 0.40 | 0.33 |
| TrA | 14 | 0.64 | 0.71 | 62 | 0.22 | 0.34 | 0.52 | 0.35 |
| Raw `tracker` cliques *including* bichir | | | | | | | | |
| DrAa | 7 | 0.63 | 0.43 | 44 | 0.29 | 0.46 | 0.69 | 0.33 |
| DrAb | 5 | 0.45 | 0.60 | 37 | 0.24 | 0.53 | 0.62 | 0.15 |
| DrA | 12 | 0.55 | 0.50 | 81 | 0.26 | 0.47 | 0.66 | 0.29 |
| TrAa | 9 | 0.82 | 0.56 | 49 | 0.32 | 0.39 | 0.58 | 0.33 |
| TrAb | 5 | 0.45 | 1.00 | 20 | 0.13 | 0.29 | 0.40 | 0.28 |
| TrA | 14 | 0.64 | 0.71 | 69 | 0.22 | 0.34 | 0.52 | 0.35 |
| Co-linear `tracker` cliques, marked with • | | | | | | | | |
| DrAa | 7 | 0.63 | 0.43 | 34 | 0.28 | 0.44 | 0.69 | 0.36 |
| DrAb | 5 | 0.45 | 0.60 | 29 | 0.24 | 0.52 | 0.62 | 0.16 |
| DrA | 12 | 0.55 | 0.50 | 63 | 0.26 | 0.47 | 0.66 | 0.29 |
| TrAa | 9 | 0.82 | 0.56 | 43 | 0.35 | 0.43 | 0.58 | 0.26 |
| TrAb | 5 | 0.45 | 1.00 | 19 | 0.15 | 0.34 | 0.40 | 0.15 |
| TrA | 14 | 0.64 | 0.71 | 62 | 0.25 | 0.39 | 0.52 | 0.25 |
| PFCs *sensu* Chiu et al. [25] | | | | | | | | |
| DrAa | 7 | 0.63 | 0.43 | 31 | 0.36 | 0.57 | 0.69 | 0.2 |
| DrAb | 5 | 0.45 | 0.60 | 26 | 0.30 | 0.67 | 0.62 | -0.1 |
| DrA | 12 | 0.55 | 0.50 | 57 | 0.33 | 0.60 | 0.66 | 0.1 |
| TrAa | 9 | 0.82 | 0.56 | 36 | 0.42 | 0.51 | 0.58 | 0.1 |
| TrAb | 5 | 0.45 | 1.00 | 17 | 0.20 | 0.44 | 0.40 | -0.1 |
| TrA | 14 | 0.64 | 0.71 | 53 | 0.31 | 0.48 | 0.52 | 0.1 |

# 4.4 Independent *Hox* Cluster Duplication in Lampreys

## 4.4.1 Introduction

There is good evidence that the common ancestor of sharks and bony fish (which also includes the land vertebrates) had four clusters homologous to the mammalian ones [52, 94]. The agnathan vertebrates, lampreys (*Hyperoartia*) and hagfishes (*Hyperotreti*), as the most primitive extant true vertebrates, occupy a phylogenetically intermediate position between the cephalochordates, such as amphioxus, with a single *Hox* cluster [44] and the gnathostomes with four or more clusters. The lampreys share many characters with the higher vertebrates including e.g. a notochord, dorsal neural tubes, tripartite brain, segmented muscle blocks and paired sense organs, but lacks jaws and paired fins [52, 110]. PCR surveys [90, 105] and recent genomic mapping data [41, 57] indicate that lampreys have at least three and possibly four *Hox* clusters, cf. Fig. 4.9.

Despite recent efforts the evolutionary history of the lamprey *Hox* genes and their relationship with the quadruplicate mammalian *Hox* clusters is far from being resolved. Irvine *et al.* [57] conclude that their data are "insufficient data to determine with confidence the identities and evolutionary histories of the lamprey Hox clusters." Amores *et al.* [4] argue for a two-step duplication scenario with a duplication of both ancestral agnathan clusters, possibly simultaneously by genome duplication, to produce the four cluster of the ancestral gnathostome arrangement. Force *et al.* [41] report that "in general, the lamprey Hox genes do not appear to be orthologous of specific Hox genes in gnathostomes" and conclude that the most likely scenario is genome duplication in the vertebrate ancestor producing a *HoxAB* and *HoxCD* cluster with subsequent

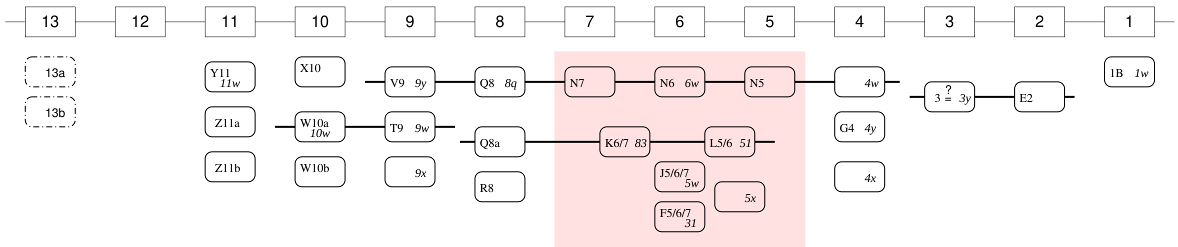



Figure 4.9. *Petromyzon marinus Hox* clusters. Summarized from Force et al. [41], and Irvine et al. [57] Tab. 4.7. *Hox13* genes identified in the PCR survey of Force et al. [41] but for which no cDNA or cosmid was reported are indicated by dashed boxes. The corresponding sequences are not publicly available. Physical linkage is indicated by a line. The sequences of the paralogous groups 5, 6, and 7 are insufficient to resolve their mutual relationships. Hence they are excluded from this study.

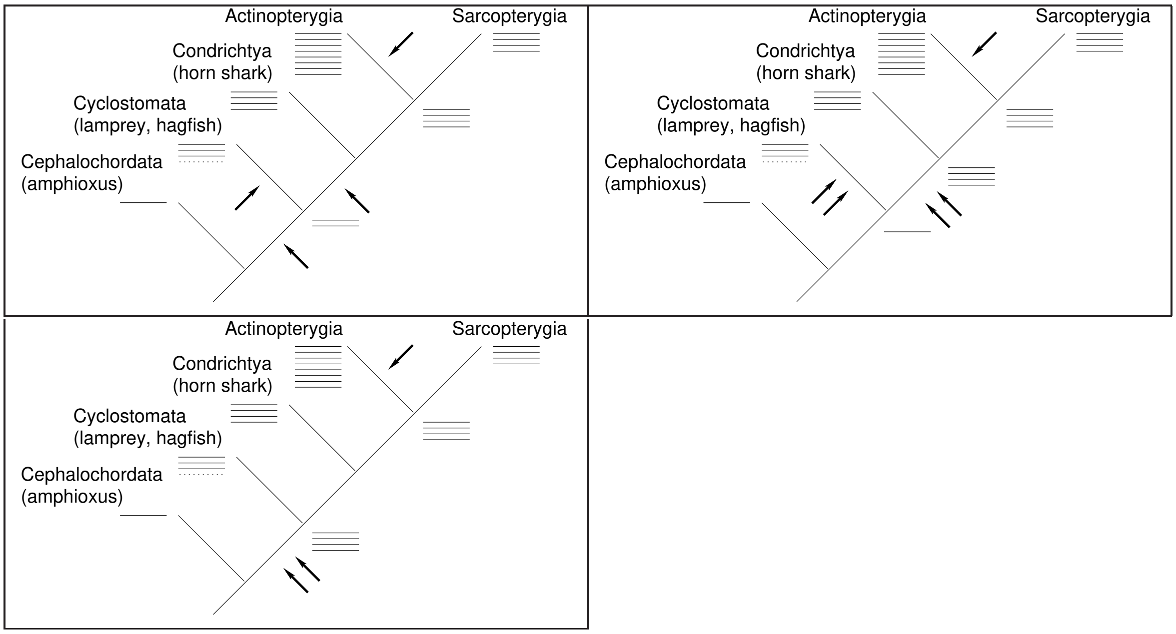



Figure 4.10. Three different duplication scenarios.
Top left: One duplication occurred in the early vertebrate lineage before the separation of the Cyclostoma lineage. In the lineage leading to the lampreys and hagfish and in the lineage of the gnathostome a second independent duplication occurred [85, 34, 46, 41].
Top right: Two independent duplications after the divergence of cyclostomata and gnatosthomes [109].
Bottom left: Both duplication events took place before the divergence of cyclostomata and gnatosthomes [4].
Arrows denote duplication events. Dotted lines denote the uncertain fourth lamprey cluster.

divergence of the agnathan and gnathostome lineages and independent duplications in each linage. Ample evidence from other gene families, including *Dlx* [85] and *Otx* [46] confirms at least one independent duplication in the agnathan and the gnathostome lineages, see also [34].

Here we report on a re-evaluation of the publicly available lamprey *Hox* sequences. This study, a joint work with S. J. Prohaska [43]. should give insights into the order of cluster duplication and whether or not cluster duplication occurred independently in different vertebrate lineages.

### 4.4.2 Sequence Data

The available lamprey *Hox* sequences are compiled (together with their accession numbers) in Tab. 4.7. In contrast to the previous studies we use the nucleic acid sequences rather than the sequences of the *Hox* proteins because of the weak phylogenetic signals in the short sequences that are available in the database. The sequence from the PCR

Table 4.7. Lamprey *Hox* sequences used in this section.

| *Hox* | *Petromyzon marinus* | | | | | *Lampetra planeri* | |
|---|---|---|---|---|---|---|---|
| | genomic clones | | | PCR surveys | | | |
| | [57] | [41, 21] | Acc. No. | [90] | Acc. No. | [105] | Acc. No. |
| 13 | | | | | | Lp13A | AF044814 |
| 11 | Y11 | 11w | AF410923 | 11.1 | | | |
| | Z11a | | AF410924 | 11.8 | | | |
| | Z11b | | AF410925 | | | | |
| | | | | 11.6 | | | |
| 10 | X10 | Hx13(9) | AF410922 | 10x | L14900 | | |
| | W10a | 10w | AF410920 | 10w | L14895 | Lp10B | AF044812 |
| | W10b | | AF410921 | | | | |
| 9 | V9 | 9y | AF410919 | 9v | L14889 | Lp9A | AF044809 |
| | | | | 9s | L14911 | | |
| | T9 | 9w | AF410918 | 9t | L14894 | Lp9B | AF044810 |
| | | 9x | | 9u | L14910 | Lp9C | AF044811 |
| 8 | R8 | | AF035588 | 8r | | Lp8A | AF044807 |
| | Q8 | | AF035591 | | | | |
| | Q8a | | AF035589 | 8q | L14901 | Lp8B | AF044808 |
| 4 | G4 | 4y | AF410911 | 4g | L14912 | | |
| | | 4w | AF434666 | 4n | L14896 | Lp4-7B | AF044803 |
| | | 4x | AY056469 | 4l | L14891 | Lp4-7A | AF044802 |
| | | | | | | Lp4-7E | AF044806 |
| | | | | (4h | L14909) | | |
| 3 | 3 | 3y | AF410909 | | | Lp3A | AF044801 |
| 2 | E2 | | AF410908 | 2e | | Lp2A | AF044800 |
| 1 | 1B | 1w | AF434665 | 1b | L14902 | Lp1B | AF044798 |
| | | | | 1a | L14893 | Lp1A | AF044797 |
| | | | | 1c | L14908 | Lp1C | AF044799 |
| | | | | (1d | L14904) | | |

The sequences shown in parentheses are not included because we could not confirm their assignment to a paralogous group based on their nucleic acid sequence.

survey of *Lampetra planeri* [105] are much shorter (82nt) than the *Petromyzon marinus* sequence reported by Pendleton *et al.* [90] (180nt) and Irvine *et al.* [57] (240nt). In almost all cases it was possible to identify the homology between the *Lampetra planeri* sequences and their *Petromyzon marinus* counterparts, see Tab. 4.7. We therefore use the data from Irvine *et al.* [57] where possible.

### 4.4.3 Analysis

Canonical split decomposition [10] implemented in the `splitstree` package (version 3.1) [56] is used for the reconstruction of the phylogeny. The split-based methods are particularly suitable for our purposes because they are known to be very conservative in that they tend to produce multifurcations rather than poorly supported edges [104]. For comparison we compute exact maximum parsimony trees using the program `dnapenny` which is part of the `phylip` package [35]. We use a variety of *Hox* genes from mammals *Homo sapiens* and *Rattus norvegicus*, shark *Heterodontus francisci*, coelacanth *Latimeria menadoensis* [65], and amphioxus *Branchiostoma floridae* for phylogeny reconstruction. Since split-based methods tend to lose resolution with increasing number of taxa we use different combinations of lamprey sequences and sequences from other taxa instead of using all sequences together.

An independent line of evidence is derived from the analysis of conserved non-coding DNA. The 30kb PAC clone Pm18 containing the *HoxW10a* region of *Petromyzon marinus* was sequenced by Irvine *et al.* [57]. Here use the `tracker` program [95] in order to search for phylogenetic footprints in the non-coding parts of this sequence (Acc. no. AF464190) by comparing it with the corresponding regions of the publicly available sequences of human, pufferfish (*Takifugu rubripes*), zebrafish (*Danio rerio*), and shark *Hox* clusters. We compare the intergenic parts of the Pm18 sequence with the homologous parts of the *Hox* clusters from pufferfish, human, and shark. In the case of the *HoxB* clusters, which lack *Hox-10*, *Hox-11* and *Hox-12* gene, we use the region from *Hox-13* to *Hox-9* for the `tracker` run. The analysis is then restricted to the region between the first and the last footprint that the lamprey sequence shares with another cluster to account for the fact that Pm18 does not span the entire range to the neighboring genes.

The `tracker` program produces alignments of the footprint cliques using `dialign` [82]. 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: (1) gene loss will cause almost certainly the loss of all the the associated regulatory sequences. In the extreme case, presence-absence data of 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 from the frequency of mutations within cliques that are shared by the two clusters. Technically, this amounts to treating "gaps" as missing data rather than as an additional character state.

### 4.4.4 Results

Only the paralogous groups 1, 2, 3, 4, 8, 9, 10, and 11 could be used for our purposes because (i) no *Hox-13* sequences for lampreys were found in the databank, (ii) there does not seem to be a *Hox-12* gene at all in lampreys, and (iii) the available sequences are too short to distinguish unambiguously between members of the paralogous groups 5, 6, and 7, see also Force *et al.* and Irvine *et al.* [41, 57].

The comparison of mammalian, shark, lamprey, and amphioxus sequences for a given paralogous group presents a striking pattern. We find that the lamprey sequences cluster together outside the gnathostome *Hox* sequences for paralogous groups 11, 10, 9, 8, and 4 according to the split decomposition analysis, Fig. 4.11. Paralogous group 1 is at least consistent with this picture. The single paralogous group 3 sequence shows affinity with the shark *HoxA* sequence but is well separated from the mammalian *HoxA-3* genes in the split data. The *PmE2* sequence, which is physically linked to *Pm3*, is more similar to the mammalian *HoxB-2* genes. Replacing the rat sequences by coelacanth sequences from the work of Koh *et al.* [65] yields very similar results, drawn in Fig. 4.12.

The same picture is obtained from maximally parsimonious trees, see Tab. 4.8, for groups 11, 10, 9, 8. In contrast to the split decomposition method, the best trees for both paralogous group 3 and 2 place the lamprey and amphioxus sequences together and as outgroup to the gnathostome clusters. Paralogous group 1 yields one tree that shows the 1w sequence outside the mammalian cluster and two alternative trees placing 1w with mammalian *A* clusters. In paralogous group 4 the lamprey sequences also lie outside the mammalian clusters but form two separate branches. In no case do we find a clear assignment of the lamprey clusters to either a single or a pair of mammalian and/or fish clusters. Furthermore, the single *Hox-13* sequence of *Lampetra planeri* from Sharman *et al.* [105] also branches outside the other vertebrate genes.

At present the genomic context of only a single lamprey *Hox* gene, *Hox-W10a* from

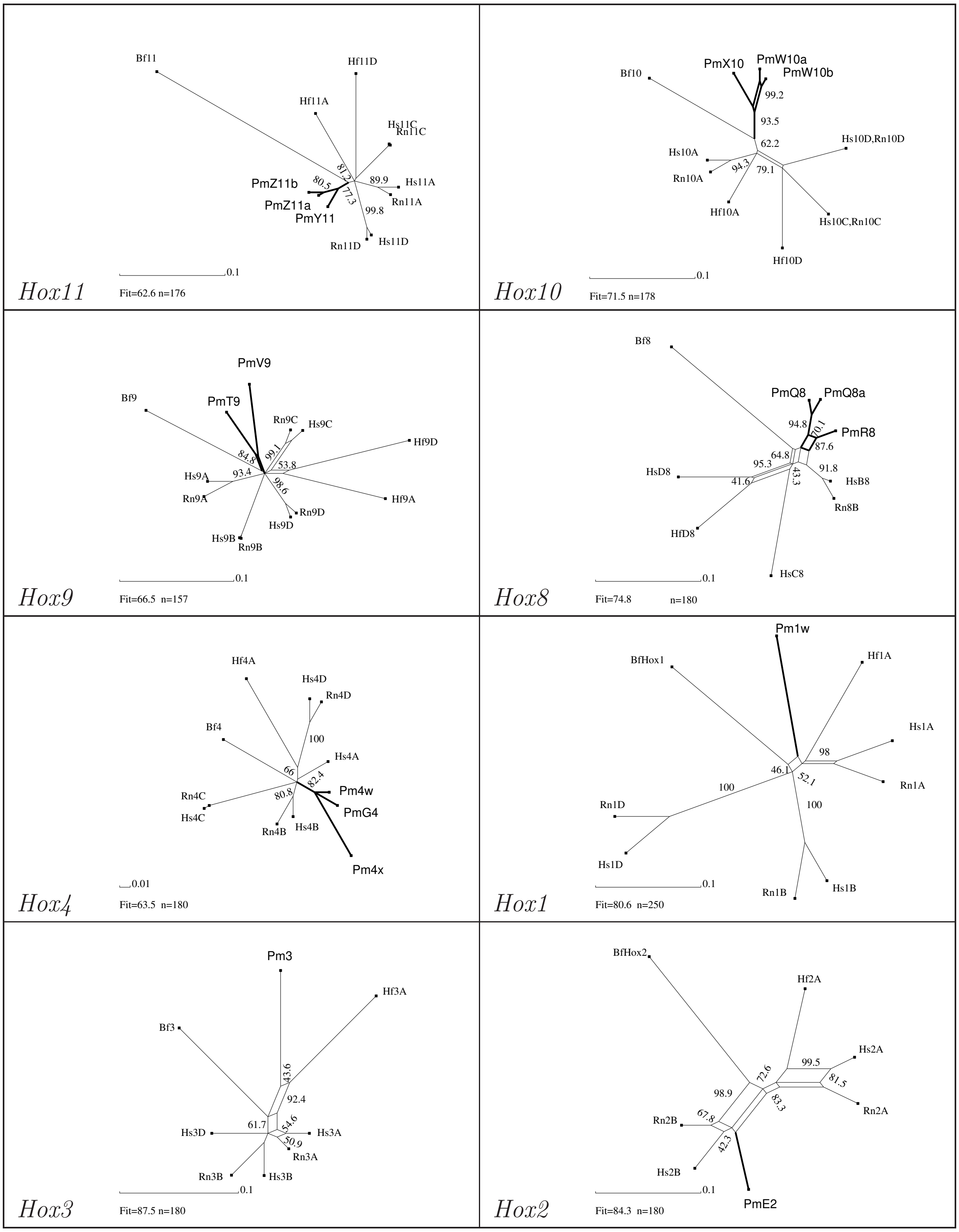



Figure 4.11. Buneman graphs of the homeobox sequences for paralogous groups 1, 2, 3, 4, 8, 9, 10, and 11. We show here the comparison with human, rat, shark, and amphioxus sequences.

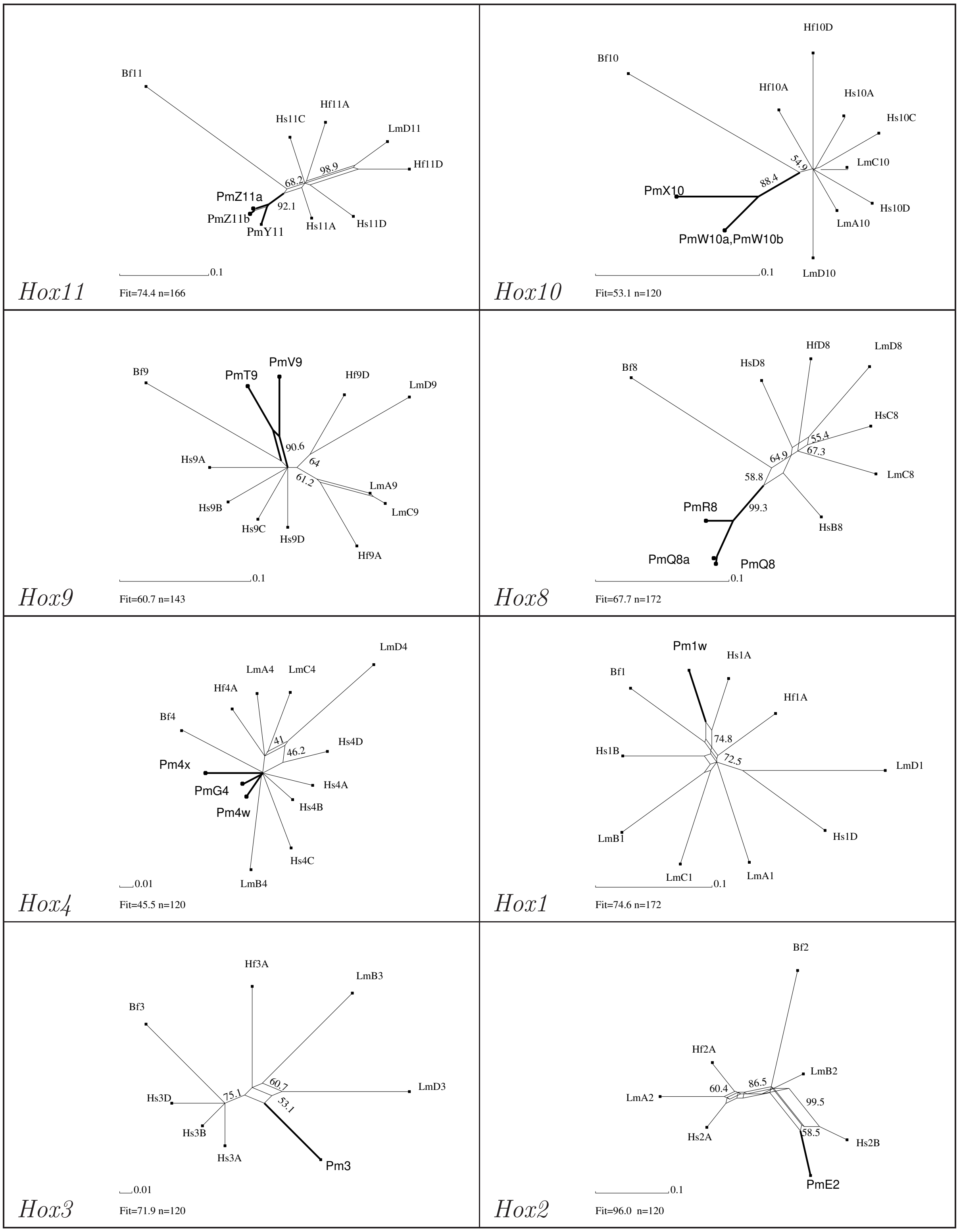



Figure 4.12. Buneman graphs of the homeobox sequences for paralogous groups 1, 2, 3, 4, 8, 9, 10, and 11. As in Tab. 4.11 we show here the comparison with human, shark, and amphioxus hovwever the rat sequence is replaced by the coelacanth a more basal vertebrate. Using Teleost fish or coelacanth sequences instead of mammalian data yield qualitatively the same results (data not shown).

*Petromyzon marinus*, has been published. Irvine *et al.* [57] report footprint clusters shared with both *HoxA* and *HoxC* clusters. The footprint cliques detected by the `tracker` program in a comparison with pufferfish, shark, human, and ciona *Hox* clusters are summarized in Tab. 4.9. Non-colinear cliques have been removed because they are most likely not homologous [95]. There is no clear evidence that the non-coding part of the Pm18 sequence is more closely related to either a particular single gnathostome cluster or pair of clusters. The total length of available footprints is unfortunately insufficient for an independent reconstruction of the phylogeny. The most significant footprint cliques are those shared with the *HoxA* and *HoxC* clusters, in particular, and an element designated ***pp*** that is most likely the proximal promotor of the *Hox-10* genes and also appears in the *HoxD* clusters. The elements ***A1***, ***A2***, ***C1***, and ***C3*** are described already in the work of Irvine et al. [57]. Both ***A1***, and ***A2*** were also detected in comparisons of *HoxA* clusters only by Chiu *et al.* [25]. It is interesting to note that both ***A2*** and the ***C1***, ***C2*** motifs also have their counterparts in the human *HoxB* cluster, even though it lacks the *HoxB-10* gene.

### 4.4.5 Discussion

Analysis of different gene familys (*Msx*, *Cdx*, *Hox*, *En*, *Wnt*) [44] showed that two duplication events occured early in the vertebrate evolution. Consistent with this hypothesis the investigation of *Dlx* genes [85], neural crest marker *AP-2* [79] suggests that a duplication occurred before the divergence of lampreys from gnathostomes, which was then followed by independent chromosomal or genome duplications and gene loss in each lineage. The assignment of the the analyzed *Hox* genes of lampreys to a paralog group of the vertebrate genes was not possible. Independent duplication events in lampreys and vertebrates implies that none of the lamprey genes has a homolog gene in the vertebrates. Therefore the phylogenetic relationship obtained by the analysis of the *Hox* gene sequences of lamprey indicate that independent duplication events lead to the three or possibly four *Hox* clusters of the lampreys.

Table 4.8. Maximum parsimony trees of the homeobox sequences obtained with the program `dnapenny` from the `phylip` package [35]. Lamprey sequences are indicated in bold. Grey boxes indicated that all available lamprey paralogs form a subtree, dark gray boxes are used when all lamprey and the amphioxus sequence are separated from the vertebrate *Hox* clusters.

| Hox | Maximum parsimony tree |
|---|---|
| 13 | (**LpHox13**,(((*Bf13_ex2*,(RnC13,HsC13)),((RnD13,HsD13),((RnA13,HsA13),(HfD13,HfA13)))),(HsB13,RnB13))) |
| | (**LpHox13**,(((*Bf13_ex2*,(RnC13,HsC13)),((RnD13,HsD13),(HfD13,((RnA13,HsA13),HfA13)))),(HsB13,RnB13))) |
| 11 | ( (*Bf11*,**(PmY11,(PmZ11a,PmZ11b))**) ,(((Hf11D,(Hs11C,Rn11C)),((Rn11A,Hs11A),Hf11A)),(Rn11D,Hs11D))) |
| 10 | ( (*Bf10*,**(PmX10,(PmW10b,PmW10a))**) ,((Rn10A,Hs10A),(Hf10A,((Rn10C,Hs10C),((Hs10D,Rn10D),Hf10D))))) |
| 9 | ( (*BfHox9*,**(HoxT9,HoxV9)**) ,(((Hs9D,Rn9D),((((Rn9A,Hs9A),Hf9A),(Hs9B,Rn9B)),((Hs9C,Rn9C),Hf9D))))) |
| | ( (*BfHox9*,**(HoxT9,HoxV9)**) ,(((((Hs9D,Rn9D),((Hs9C,Rn9C),Hf9D)),(Hf9A,(Hs9B,Rn9B))),(Rn9A,Hs9A)))) |
| 8 | ( (*Bf8*,**(PmQ8,(PmR8,PmQ8a))**) ,((HsC8,(HsB8,Rn8B)),(HfD8,HsD8))) |
| 4 | ((*Bf4*,**PmG4**),(((**Pm4x,Pm4w)**,(Hs4A,(Hf4A,(Rn4D,Hs4D)))),(Hs4B,Rn4B)),(Hs4C,Rn4C)) |
| | ((*Bf4*,**PmG4**),(**(Pm4x,Pm4w)**,((Hs4B,Rn4B),(Hs4A,(Hf4A,(Rn4D,Hs4D))))),(Hs4C,Rn4C)) |
| 3 | ( (*Bf3*,**Pm3**) ,(Hf3A,(Hs3D,((Hs3A,Rn3A),(Hs3B,Rn3B))))) |
| 2 | ( (*Bf2*,**PmE2**) ,(((Hf2A,(Rn2A,Hs2A)),(Rn2B,Hs2B)))) |
| | ( (*Bf2*,**PmE2**) ,((((Hf2A,(Rn2A,Hs2A)),Hs2B),Rn2B))) |
| 1 | ( (*Bf1*,**Pm1w**) ,((((Hs1D,Rn1D),Hf1A),(Rn1A,Hs1A)),(Hs1B,Rn1B))) |
| | (*Bf1*,(((Hs1D,Rn1D),Hf1A),((Hs1B,Rn1B),((Rn1A,Hs1A),**Pm1w**)))) |
| | (*Bf1*,(((Hs1D,Rn1D),(Hs1B,Rn1B)),(Hf1A,((Rn1A,Hs1A),**Pm1w**)))) |

Table 4.9. Summary of co-linear footprint cliques produced by the `tracker` program in the range of the *Petromyzon marinus* Pm18 sequence. Hs *Homo Sapiens*, Hf *Heterodontus fransisci*, Tr *Takifugu rubripes*. Numbers in parentheses are non-colinear with the footprints in this species. The last column marks previously described footprints. ***pp*** is the proximal promotor of the *Hox-10* gene, numbers in sans serif font are cliques listed in [95] for a comparison of *HoxA* clusters. PFC, "phylogenetic footprint cluster", names from [25] are given in normal text font.

| # | Pm18 | HsA | HsB | HsC | HsD | HFM | HfD | TrAa | TrAb | TrBa | TrD | Remark |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 53 | 1150 85 | 8426 114 | | 9641 88 | 7528 161 | 11071 223 | 8404 162 | 4980 98 | 3360 139 | | 8112 126 | ***pp*** 42 |
| | *Hox-10* | | | *Hox-10* | | | | | | | *Hox-10* | |
| 61 | 10436 38 | | 39411 38 | | | | | | | | | |
| 62 | | | | | | | 11114 37 | | | 15608 37 | | |
| 64 | | | | | | | 13213 26 | | | 19306 26 | | |
| 66 | 10538 33 | | | | | | 15207 33 | | | | | |
| 68 | 17246 19 | | 60196 19 | | | | | | | | | |
| 70 | | 12127 39 | | | | 14376 43 | | | | | | 43 10-9-a |
| 71 | | 12248 24 | | | | 14522 24 | | | | | | 44 10-9-a |
| 72 | 21911 64 | 12292 187 | | | | 14566 189 | | 7525 108 | | | | ***A1*** 45 10-9-a |
| 73 | 23635 27 | | | | | | | | | 22252 27 | | |
| 74 | | | 70272 42 | | | | | | 5159 42 | (18197) (21) | | |
| 75 | | | 70497 33 | | | | 15247 33 | | | | | |
| 80 | 25443 21 | | 80312 21 | | | | | | | | | |
| 81 | 26904 26 | | | | | | | | | 25684 26 | | |
| 90 | 27436 105 | 13224 116 | (53397) (38) | 16339 94 | | 15518 116 | | 7896 99 | | (18814) (82) | | ***A2*** 46 10-9-b |
| 82 | | | | 18063 27 | | | | | | 28309 27 | | |
| 83 | | | 84517 19 | 20176 19 | | | | | | | | |
| 86 | | | | | 10744 52 | | | | | | 10133 50 | |
| 87 | | | 91014 30 | | 11932 41 | | | 8538 41 | | | | |
| 91 | | 14160 49 | | | | 16310 77 | | 8304 99 | 6287 73 | | | 48 10-9-b |
| 92 | | 14373 80 | 84219 75 | | | | | | | | | |

Chapter 5

# Conclusion and Outlook

Evolutionarily conserved non-coding genomic sequences represent a potentially rich source for the discovery of gene regulatory regions. Since these elements are subject to stabilizing selection they evolve much more slowly than adjacent non-functional DNA. These phylogenetic footprints can be detected by comparison of the sequences surrounding orthologous genes in different species. Therefore the loss of phylogenetic footprints as well as the acquisition of conserved non-coding sequences in some lineages, but not others, can provide evidence for the evolutionary modification of cis-regulatory elements. Furthermore the evolution of development is to a large part based on changes in the cis-regulatory elements of developmental genes [29]. To study the pattern of cis-regulatory element evolution, however, requires the comparison of relatively long sequences from many species. To this end we have developed an efficient software tool for the identification of corresponding footprints in long sequences from multiple species. We presented here a novel computational method that allows the identification of partially conserved, homologous sequences in long stretches of DNA. This method opens up an alternative avenue to the study of non-coding sequence evolution. It uses the fact that, at least in vertebrates, cis-regulatory sites have been shown to evolve at a lower rate than surrounding sequences. With a sufficient number of sequences from phylogenetically well placed taxa it is then possible to study the origin, maintenance and loss of conserved sequence segments among different lineages. The method, which is based on pairwise sequence comparisons and subsequent assembly and filtering steps, is designed to deal with a moderately large number of long sequences. The survey of the eight *HoxA* clusters reported here, for instance, requires less than 5min on a modern PC. The `tracker` tool can therefore be used for much larger datasets as the resource usage scales approximately as $\mathcal{O}(L \times N^2)$ for $N$ input sequences of length $L$.

We have applied this tool to *HoxA* clusters to analyze the modifications of non-coding sequences following *Hox* cluster duplication. The study of gene clusters provides a particularly good opportunity for the study of non-coding sequence evolution, because the identity and extent of a non-coding sequence is uniquely defined by the flanking coding genes. We analyzed the *HoxA* cluster of eight vertebrate species. We used sequences of the teleosts zebrafish, pufferfish and striped bass in comparison to the single *HoxA* cluster of shark, human and the basal actinopterygian fish bichir. We were able to show that it is possible to detect differences in the pattern of cis-regulatory sequence conservation between teleost, basal fishes and tetrapods. The footprint pattern in the shark *HoxA* cluster closely resembles the human distribution, while teleost fishes deviate dramatically as a consequence of an additional genome duplication. The most striking outcome of the analysis is that bichir shows a pattern of sequence conservation that lies between the pattern of human and shark on the one hand and the footprint pattern of the duplicated *Hox* cluster of striped bass, zebrafish and pufferfish on the other hand. In addition we showed in this analysis a collaboration with Chi-Hua Chiu [26] that some of the phylogenetic footprints of teleosts are obtained at a basal stage of the actinopterygian evolution already. The acquisition of these phylogenetic footprints took place before the *Hox* cluster duplication that leads to an increased number of clusters in the derived teleosts. The derived teleosts show an extensive remodeling in the pattern of phylogenetic footprints, some of this innovations appear already in bichir. Therefore we hypothesize that bichir had already obtained changes in the pattern of non-coding sequence conservation although the major part of changes like the massive loss of footprints occur after the duplication event.

The comparative analysis of sequences is much aided by models of sequence evolution. In the case of coding sequences a large number of models can be used to detect unusual patterns of sequence change [47]. We use here a model for a comparable analysis of non-coding sequence. The purpose of this model is to estimate the amount of footprint clique loss that can be attributed to “structural” reasons, such as gene loss. The results show that the observed amount of non-coding sequence modification is in all cases higher than expected solely for structural reasons. This is consistent with the idea that *Hox* cluster duplication can facilitate the evolution of development [72, 25]. It is hard to distinguish between the two possible reasons for this excess in the loss of sequence conservation: binding site turnover and adaptive modification. The former changes sequences of cis-regulatory elements without affecting function, while the latter is the cis-regulatory trace of changes in the function of the associated genes.

In our analysis we find a higher retention rate for the pufferfish *HoxAb* compared to the *HoxAa* cluster. Since there is no reason to assume that the rate of binding site turnover should be different between paralogous *Hox* clusters, the most parsimonious interpretation is that, in pufferfish, the *HoxAb* cluster experienced a higher amount of adaptive change in its cis-regulatory elements than the *HoxAa* cluster. This suggestion can be tested by expression studies and transgenic tests of non-coding sequences.

For further investigation of the history of *Hox* cluster evolution at a more basal level we re-evaluated the available lamprey *Hox* genes. The analysis strongly supports an independent origin of the three (or four) lamprey *Hox* clusters and would suggest that the common ancestor of agnathans and gnathostomes had only a single *Hox* cluster. This is in particular consistent with the *Dlx* gene phylogeny of Neidert *et al.* [85]. These authors propose that a tandem duplication of an ancestral *Dlx* gene predated the divergence of lampreys from gnathostomes, which was then followed by independent chromosomal or genome duplications and gene loss in each lineage. Our evaluation of the *Hox* clusters supports this hypothesis. Similar patterns have been reported for other developmentally important gene families. The neural crest marker *AP-2*, for which no duplicates have been found in lampreys, also fails to group with any one gnathostome *AP-2* isoform [79]. Consistent with an independent duplication history, it is impossible to assign any one of the lamprey (and hagfish) *Otx* sequences to one of the three classes identified in gnathostomes [46]. The phylogenetic signal in the *Hox* clusters is not as strong as one would like so that a definitive result will have to await more complete sequencing of the lamprey. This will in particular allow the unambiguous identification of the genes of paralogous group 5, 6, and 7. At present, at least, the publicly available sequence information does not contain evidence for a *Hox* cluster duplication proceeding our common ancestor with the lampreys.

The distribution of footprints and the sequence conservation within footprint clusters is a useful source of phylogenetic evolution, in particular when the data from the nearby genes are hard to interpret, e.g. because of gene-loss in some species. Prohaska *et al.* [94] used the footprint pattern obtained by `tracker` to resolve the relationship of the two sequenced hornshark *Hox* clusters *HfM* and *HfN* with the four mammalian clusters. The statistical analysis of the footprint patterns in the *HfN* cluster shows that the shark *HfN* cluster is indeed *HoxD*-like as was supposed in the first place [62, 72]. A second line of evidence was derived from concatenating the alignments of the footprint cliques (treating gaps as missing data rather than as a separate character state) [94]. Phylogenies were then reconstructed by split-based methods which are known to be very conservative in the sense that they rather produce multifurcation than ill-supported branches. These data strongly support the homology of *HfN* to the mammalian *HoxD* clusters [94]. It follows that the most recent common ancestor of the jawed vertebrates had at least four *Hox* clusters, including those that are orthologous to the four mammalian *Hox* clusters.

Polymorphisms have the potential to alter protein functions in ways that are biologically or clinically important. As increasing numbers of polymorphism are identified in regulatory regions of genes [81] it is of great interest for the understanding of diseases to combine the results of our `tracker` program with the available data of known SNPs. We find that that SNPs seems to be detrimental to function since the are underrepresented in phylogenetic footprints. Given the conservation of functional elements it is an expected finding that the amount of SNPs in phylogenetic footprints is lower than in the surrounding nonfunctional DNA sequence. To test if this result is significant the comparison of SNPs and phylogenetic footprints analysis should be extended. The question arise: Are SNPs more detrimental to phylogenetic footprints that are stronger conserved i.e. are common to all of the analyzed species than to footprints that are lost in several species?

The analysis of some components of the immune system show several rearrangements between the analyzed species. Especially the rat sequence seems to have undergone different rearrangements in comparison to the sequences of human and mouse. Whether these rearrangements — mostly inversions in the rat genome — occurred in the rat genome or occurred specifically in this genes or if the rearrangements are typical for the whole rat genome may can be distinguished with an extensive study of a more complex dataset. For this purpose we would like to know if the differences in the rat genome are significant. To this end, the distances between adjacent footprints in two sequences have to be compared. For this purpose we need a perfect sorted list where co-linear footprints are disregarded. It seems to be a simple problem to sort footprints in their order along the genomes. Nevertheless it is complicated by the fact that not all footprints are co-linear. 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. Solving the footprint sorting problem requires the solution of the "Minimum Weight Vertex Feedback Set Problem", which is known to be NP-complete and APX-hard. Despite all the above mentioned problems it seems that good approximations can be obtained for datasets of interest. The remaining steps of the sorting process are straight forward: computation of the transitive closure of an acyclic graph, linear extension of the resulting partial order, and finally sorting.

In principle the analysis of phylogenetic footprints can also be used to study the cis-regulatory changes associated with other evolutionary changes. For instance, it is known that the *AbdB*-related *HoxA* and *HoxD* cluster genes acquired a novel pattern of regulation with the origin of the tetrapod limb [86, 126, 120]. It should be possible to detect differences in the pattern of cis-regulatory sequence conservation between basal fishes and tetrapods. However, no data from appropriately placed taxa is currently available. The recently ongoing sequencing projects for several organism e.g. *Xenopus laevis*, *Rattus norvegicus* will provide us in near future with a expanded set of organisms. These expanded dataset can be used to investigate the correlation of footprint patterns with morphological changes. Furthermore the data obtained by the analysis of this species together with the recently published sequences of sea squirt *Ciona intestinalis* [108, 118] and sea urchin *Strongylocentrotus purpuratus* [75] opens up the avenue of further analysis of the *Hox* cluster evolution. *Ciona intestinalis* and *Strongylocentrotus* are invertebrate species and therefore contain only a single *Hox* cluster. Using one or both of them as an outgroup organism may be suitable to decide whether the 2R duplication scenario or sequential duplication is more likely.

The comprehensive data obtained by `tracker` can help to reconstruct the regulatory network of *Hox* genes. The predictions obtained from a model of the interactions between the genes can be extended by comparison with expression data or analysis of specific mutants acquired by laboratory work. We will be able to study the effect of different motif combinations on expression patterns of *Hox* genes. Furthermore, the regulatory network would be important for the understanding how the highly ordered *Hox* expression is set up during the development in vertebrate embryos. *Hox* genes are in general expressed temporally and spatially in a co-linear manner with respect to to the gene order within the *Hox* clusters. It has been shown recently that in some animals the tight organization of the clusters has disintegrated. For instance, co-linearity has been lost during evolution in the *Hox* cluster of *Ciona intestinalis* [36]. In the skin of developing chicken embryos the expression of some *Hox* genes (*HoxD4* to *HoxD13*, *HoxA11* and *HoxC6*) is not spatially restricted whereas the expression of the remaining genes is conform with co-linearity [97]. Comparison of the regulatory networks from organisms with lost co-linearity and organisms with an intact co-linearity would bring insights into the mechanism of co-linearity in vertebrates and invertebrates.

Appendix A

# Analysis of polypterus senegalus footprints

The following long table contains the complete list of footprints detected by the `tracker` program version 0.05 with default parameter settings in run `02041944STGC` with the time-stamp

Tue Feb 4 19:44:41 CET 2003

The list has been modified from the raw data by (1) removing duplicate hits and multiple entries of the same cluster, (2) removing obvious repeated elements, (3) arranging clusters in linear order as a good as possible, and (4) adding annotation in the "remarks" column.

Table A.1. **List of all footprints** found by `tracker` with default parameter settings. Alternative hits belonging to the same cluster have been removed from this table, the unique representative is marked with "f" or "ff". Numbers in the first column are the same as in the un-edited raw output.
The remarks column gives the nomenclature from Chiu *et al.* 2002.
× denotes violations of co-linearity between Hox-A13 and Hox-A1, ? markes uncertain hits. Uncertain hits within an otherwise significant cluster are shown in parentheses. These entries are disregarded in the PFC statistics. Footprint clusters are separated by lines in the first column. Footprints that are deemed reliable are marked by •. PFC summary statistics are constructed from footprint *clusters* with at least one "reliable" footprint. Clusters that are shared by euteleosts only (♮), and by euteleosts and only one of shark (♣), human (♠), bichir (◇) are marked after the first clique in the cluster. Experimentally verified binding sites are given in ***bold italics***, binding sites inferred from `transfac` searches are annotated in sans serif. Footprints that violate colinearity are disregarded (marked by × in the last column of the list).

| # | HfM | HsA | PsA | MsA | DrAa | TrAa | DrAb | TrAb | Remarks |
|---|---|---|---|---|---|---|---|---|---|
| 2 | | | | | | 99 46 | 22925 46 | | |
| 3 | | | | | | | 33702 74 | 315 74 | |
| 10 | | | 4385 46 | | | | 3820 44 | | |
| 11 | | | | | 3246 62 | | 29283 60 | | |
| 12 | | | 14118 70 | | 4938 63 | | | | |
| 13 | | | 17715 110 | | 2287 70 | | 6101 179 | | |
| 14 | | | 19741 42 | | | | 24118 42 | | |
| 15 | | 3222 76 | 21731 77 | | | | | | |
| 16 | | 4617 249 | 23240 257 | | | 920 253 | | | |
| 17 | | 4878 57 | 23498 57 | | | | | | |
| 4 | | | | | | 1522 28 | 46640 28 | | |
| 18 | | 7024 158 | 24842 165 | | | 2280 92 | | | |
| 19 | | | 25014 13 | | | 2397 13 | | | |
| 20 | | 7213 26 | 25039 35 | | | 2422 33 | | | |
| 25 | | | | | 7143 50 | 2898 22 | 32044 46 | | |
| 21 | | 8197 31 | | | | | 33525 31 | | |
| 22 | | 8623 29 | 26090 28 | | | | | | |
| 23 | | 8688 43 | 26148 40 | | | | | | |
| 24 | | | 26308 38 | | | | 32655 38 | | |
| | | *Evx-1* | | | *Evx-1* | | | *Evx-1* | |
| • 1 ♣ | 865 23 | | | | | | 1553 23 | | |
| • 26 | | 13112 228 | 29987 229 | | 15129 146 | | | | |
| 27 | | | | | 16307 78 | | 22716 78 | | × |
| 5 | | | | | | 6078 21 | 29449 21 | | × |
| • 28 ◇ | | | 33240 44 | | 17212 41 | 6153 29 | | | |
| 6 | | | | | | 6637 27 | 45312 27 | | × |
| • 7 ♣ | 1089 36 | | | | | 7484 36 | | | |
| • 8 ♣ | 2059 22 | | | | | 10359 22 | | | |
| 9 | 2891 51 | | | | | | 39450 51 | | × |
| 29 | | 13851 60 | 35561 60 | | | | | | ? |
| • 30 | | 14115 30 | 44387 30 | | | | | | |
| 31 | | 15102 41 | 17715 110 | | 2283 76 | | 7560 183 | | × |
| • 32 | 3734 81 | 20391 84 | | | | | | | |
| 33 | 3881 23 | | | | 4203 23 | | | | × |
| • 34 | 3931 23 | | 44043 23 | | | | | | |
| 35 | 5239 44 | | 9524 44 | | | | | | × |
| • 36 ◇ | | | 47200 86 | | | | | 4824 88 | |
| 37 | | | 58295 20 | | | | 4899 20 | | × |
| • 38 ◇ | | | 67194 31 | | | | 14262 31 | | cdxA SRY |

| # | HfM | HsA | PsA | MsA | DrAa | TrAa | DrAb | TrAb | Remarks |
|---|---|---|---|---|---|---|---|---|---|
| • 40 ♠ | | 25741 38 | | | | | 15607 33 | | |
| 41 | | 27295 29 | | | | | 35475 29 | | × |
| • 42 | 5901 75 | 28483 75 | | | | | | | |
| 43 | 5949 23 | 4134 23 | | | | | | | × |
| • 44 ♮ | | | | | 18316 42 | | 29824 42 | | |
| • 45 | | | | | 18387 55 | | 29928 55 | | |
| • 46 ◇ | | | 74748 41 | | 20366 41 | | | | |
| • 47 ◇ | | | 75731 46 | | | | 31340 48 | | |
| • 49 ◇ | | | 82220 112 | | | | 45382 111 | | cdxA |
| • 50 ◇ | | | 85120 32 | | | | 48451 32 | | nkx2 cdxA AML-1a |
| • 51 | | 36430 33 | 91930 33 | | | | | | |
| • 52 | | 39828 26 | 94586 26 | | | | | | |
| • 53 | 6483 120 | 45120 121 | | | | | | | Upstream of 13-a |
| • 60 ♮ | | | | | | | 54090 84 | 5381 84 | |
| • 54 ◇ | | | 108304 44 | | | | | 6334 44 | |
| • 55 ◇ | | | 112155 24 | | | 9279 24 | | | |
| • 56 | 6775 40 | 45433 51 | 123822 41 | | | | | | Upstream of 13-b |
| 57 | 11868 70 | 6297 99 | 24412 104 | | | 1800 99 | 47489 26 | | × |
| 58 | 12706 53 | | 93728 49 | | | | | | × |
| 59 | 13165 11 | | | | | 10614 11 | | | ? |
| • 61 | 13185 137 | 53810 88 | 128728 134 | | 22603 170 | 10639 149 | 58295 121 | 6656 93 | Upstream of 13-d |
| • 62 | 13360 13 | | | | | | 58469 14 | | |
| • 63 | 16120 170 | | 130136 172 | | 23420 145 | 11378 183 | 58985 174 | 7315 176 | 13pp |
| | | *Hox-A13* | | | | *Hox-A13* | | | |
| • 64 | 19133 112 | 59484 135 | 133160 130 | | (25574) (24) | | | | 13-11-a |
| • 65 ♣ | 20828 47 | | | | | | 63519 47 | | |
| • 66 ♮ | | | | | 27080 34 | 14008 34 | | | |
| • 67 ♣ | 27207 32 | | | | 28565 32 | | | | |
| • 69 ♮ | | | | | | 14820 39 | | 8813 39 | |
| • 70 | | | | | 29386 61 | 18580 56 | | | |
| 71 | | | | | 29483 35 | | 67002 35 | | ?× |
| • 72 | | | 139572 38 | | 29521 38 | | | | |
| • 73 | 27545 177 | 68084 219 | 139764 276 | | | 18891 220 | 66363 139 | | |
| 74 | | 70181 58 | | | 28402 58 | | | | × ? |
| • 75ff | 29781 168 | 70665 161 | 141628 189 | | 31057 162 | 20662 159 | 67963 189 | 13384 158 | 13-11-pp |
| | | *Hox-A11* | | | | *Hox-A11* | | | |
| 79 | 33041 93 | | | | | 23147 89 | | | ? |
| • 80 | | | | | 33813 39 | 24159 40 | | | |
| 81 | | | | | 33862 12 | 24213 12 | | | ? |
| • 82 ♮ | | | | | 37209 54 | 25259 57 | | | |
| • 83 ♮ | | | | | 42263 64 | 25886 64 | | | |
| 84 | 34076 42 | | | | 43022 42 | | | | ? |
| • 85 | 34423 78 | 75337 82 | 145837 58 | | | | | | 11-10-a |
| • 86ff | | (75818) (88) | 146163 172 | | 33891 187 | 24243 245 | 71142 250 | 16517 224 | 11-9-a |
| • 100 | 35034 87 | 76069 52 | 146429 85 | | 34209 58 | 24565 49 | 71440 76 | 16767 84 | 11-10-c |
| 101 | 41272 55 | | | | | | 71853 55 | | ? |
| • 102 ♣ | 41390 47 | | | | | 25206 46 | | | |
| 103 | | 78189 21 | | | 32835 21 | | | | × |

| # | HfM | | HsA | | PsA | | MsA | | DrAa | | TrAa | | DrAb | | TrAb | | Remarks |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| • 104 ♮ | | | | | | | | | | | 26271 | 31 | 73298 | 31 | | | |
| • 105 | | | | | | | | | | | | | 73382 | 17 | 18624 | 17 | |
| • 106 | 43095 | 143 | 81613 | 114 | 149585 | 125 | | | | | 27418 | 163 | 73404 | 159 | 18661 | 143 | |
| | *Hox-A10* | | | | | | | | | | *Hox-A10* | | | | | | |
| • 109 ♮ | | | | | | | 2110 | 96 | | | 29223 | 97 | | | | | |
| • 110 | | | | | | | 2298 | 28 | | | 29389 | 28 | | | | | |
| 111 | | | | | | | 2340 | 13 | | | 29422 | 13 | | | | | ? |
| 112 | | | | | | | 2436 | 14 | | | 29482 | 14 | | | | | ? |
| • 113 ♮ | | | | | | | 2464 | 17 | | | 29505 | 17 | | | | | |
| • 114 | | | | | | | 2492 | 50 | | | 29536 | 56 | | | | | |
| • 115 | | | | | | | 2581 | 46 | | | 29630 | 51 | | | | | |
| • 116 | | | | | | | 2644 | 21 | | | 29698 | 20 | | | | | |
| • 117 | | | | | | | 2672 | 93 | | | 29719 | 93 | | | | | |
| • 107 | 46400 | 43 | 85314 | 39 | | | | | | | | | | | | | |
| • 108 | 46546 | 24 | 85435 | 24 | | | | | | | | | | | | | |
| • 118 | 46579 | 213 | 85479 | 187 | 153511 | 215 | 2946 | 174 | (41286) | (50) | 29966 | 175 | | | | | 10-9-a |
| • 119 | | | | | | | 3139 | 59 | | | 30165 | 59 | | | | | |
| • 120 | | | | | | | 3210 | 66 | | | 30238 | 67 | | | | | |
| • 121 | | | | | | | 3313 | 20 | | | 30336 | 19 | | | | | |
| • 122 | 47542 | 116 | 86411 | 116 | 154119 | 122 | 3348 | 155 | 41556 | 97 | 30366 | 155 | 76735 | 115 | | | 10-9-b |
| • 123 | | | | | | | 3556 | 112 | | | 30587 | 92 | 76892 | 16 | | | |
| 124 | | | | | | | 3707 | 22 | | | 30716 | 20 | | | 21531 | 10 | ? |
| • 125ff | 48333 | 116 | (87347) | (49) | | | 3742 | 349 | (41872) | (35) | 30746 | 346 | 77044 | 245 | 21592 | 212 | 10-9-c |
| • 149 | | | | | | | 3929 | 96 | | | 30941 | 84 | 77166 | 94 | 21694 | 78 | |
| • 150 ♣ | 50073 | 49 | | | | | 4812 | 172 | | | 31572 | 169 | | | | | |
| • 151 | 52969 | 35 | 90123 | 35 | | | | | | | | | | | | | 10-9-d |
| • 152 | 53030 | 45 | 90216 | 44 | | | | | | | | | | | | | |
| • 157 ♮ | | | | | | | | | 43707 | 68 | 32112 | 62 | | | | | |
| • 154 | | | | | | | 5901 | 138 | | | 32451 | 133 | | | | | |
| • 155 | | | | | | | 6051 | 15 | | | 32596 | 15 | | | | | |
| • 156 | | | | | | | 6076 | 56 | | | 32626 | 55 | | | | | |
| • 153 | 53084 | 55 | 90268 | 55 | 156745 | 53 | | | | | | | 78506 | 48 | 22196 | 31 | 10-9-pp |
| • 158 | 53229 | 77 | 90337 | 157 | 156818 | 148 | 6154 | 222 | 43987 | 68 | 32692 | 220 | 78594 | 118 | 22269 | 89 | 10-9-pp |
| • 161 | | | | | | | 6417 | 51 | | | 32953 | 41 | | | | | |
| | *Hox-A9* | | | | | | | | | | | | | | | | |
| • 159 | | | 92822 | 24 | 158492 | 24 | | | | | | | | | | | |
| • 160 | | | 92882 | 61 | 158546 | 55 | | | | | | | | | | | |
| • 162 ♮ | | | | | | | 7720 | 180 | | | 34183 | 178 | | | | | |
| • 163 | | | | | | | 7913 | 12 | | | 34366 | 12 | | | | | |
| • 164 | | | | | | | 7947 | 29 | | | 34398 | 29 | | | | | |
| • 165 | | | | | | | 7987 | 46 | | | 34438 | 45 | | | | | |
| • 166 | | | | | | | 8086 | 44 | | | 34534 | 44 | | | | | |
| • 167 | | | | | | | 8154 | 62 | | | 34584 | 61 | | | | | |
| • 168 | | | | | | | 8226 | 60 | | | 34648 | 57 | | | | | |
| • 169 | | | | | | | 8296 | 223 | 45766 | 47 | 34717 | 230 | | | | | |
| • 170 | | | | | | | 8534 | 17 | | | 34965 | 16 | | | | | |
| • 171f | 56941 | 111 | 94192 | 62 | | | 8888 | 225 | 46679 | 175 | 35174 | 225 | 81365 | 83 | | | 9-7-a |
| 174 | | | | | 158746 | 31 | | | 52101 | 37 | | | | | 27731 | 44 | ×? |
| 175 | | | | | | | 9221 | 11 | | | 35441 | 11 | | | | | ? |
| • 176 | | | | | | | 9284 | 56 | | | 35493 | 54 | | | | | |
| • 177 | 57228 | 226 | 94466 | 223 | 160242 | 226 | 9359 | 537 | 47011 | 213 | 35554 | 531 | | | | | 9-7-b |
| 178 | 57228 | 226 | 97346 | 38 | 160242 | 226 | 9359 | 537 | 47011 | 213 | 35554 | 531 | | | | | × |
| • 179 | 57682 | 31 | 94837 | 31 | | | | | | | | | | | | | |
| • 180 ♣ | 59503 | 39 | | | | | | | | | | | 87245 | 36 | | | |

| # | HfM | HsA | PsA | MsA | DrAa | TrAa | DrAb | TrAb | Remarks |
|---|---|---|---|---|---|---|---|---|---|
| • 185 ♮ | | | | 10120 16 | | 36280 16 | | | |
| • 186 | | | | 10199 67 | | 36353 68 | | | |
| • 183 ♠ | | 99196 28 | | | | | | 26430 28 | |
| • 184 | 62154 12 | 99258 12 | | | | | | | 9-7-pp |
| • 187 | 62176 159 | 99280 157 | 162084 159 | 11415 223 | 48807 58 | 37137 266 | | | 9-7-pp |
| | *Hox-A7* | | | *Hox-A7* | | | | | |
| 189 | | | 162686 79 | | | 37812 77 | | | possible remnant of Hox-A7 |
| 188 | 64887 42 | | 161405 42 | | | | | | × |
| • 190 ◇ | | | 163492 76 | 14139 73 | | | | | |
| 191 | 66397 22 | | 165572 22 | | | | | | ? |
| 173 | | | | | 49660 26 | | 88070 26 | | × |
| 192 | | | | 14518 34 | | 39351 34 | | | ? |
| • 193 | 66439 231 | 103200 218 | 165611 235 | 14565 440 | 49926 235 | 39395 379 | | | 7-6-a |
| • 194 | 66688 17 | | 165857 17 | | | | | | |
| • 195 | 66923 24 | 103654 24 | | | | | | | |
| 197 | | | | 15018 14 | | 39785 14 | | | ? |
| • 198 ♮ | | | | 15098 194 | | 39857 186 | | | |
| • 199 | | | | 15319 22 | | 40077 22 | | | |
| • 200 ♮ | | | | 15700 67 | | 40338 65 | | | |
| • 201 | | | | 16398 25 | | 40811 25 | | | |
| • 196 | 71706 57 | 108022 41 | 168484 58 | | | | | | 7-6-pp |
| • 202 | 71778 148 | 108078 147 | 168558 140 | 16526 139 | | 40909 133 | | | 7-6-pp |
| | *Hox-A6* | | | | | | | | |
| • 203 | 74397 30 | 111988 29 | 170824 32 | | | | | | |
| • 204 ♮ | | | | 16856 52 | | 41246 45 | | | |
| • 205 | | | | 16953 70 | | 41310 67 | | | |
| • 206 ♮ | | | | 17826 70 | | 41962 65 | | | |
| • 207 | | | | 18013 30 | 50568 30 | | | | |
| • 208 | | | | 18045 145 | | 42174 144 | | | |
| • 209 | | | | 18217 34 | 53081 39 | 42347 40 | | | |
| • 210ff | 74469 318 | 112037 330 | 170873 328 | 18269 366 | 53161 318 | 42403 356 | | | 6-5-pp |
| | *Hox-A5* | | | | | | | | |
| • 255 | 76119 11 | 114171 11 | | | | | | | |
| • 256 | 76145 22 | 114197 22 | | | | | | | |
| • 257 | 76181 22 | 114231 22 | | | | | | | |
| • 258 | 76215 226 | 114264 253 | 172865 191 | | | | | | 5-4-a |
| • 259 | 76451 50 | 114542 29 | 173072 51 | | | | | | 5-4-a |
| • 260 | 76530 47 | 114585 62 | 173138 72 | | | | | | |
| • 261 | 76627 83 | 114670 123 | 173231 125 | | | | | | 5-4-b |
| • 262 | 76740 88 | 114894 44 | 173382 69 | | | | | | |
| • 263 | 77370 22 | | 173881 21 | | | | | | |
| • 264ff | 77534 357 | 115509 408 | 173989 398 | 21536 276 | 55930 275 | 45361 265 | | | 5-4-c |
| • 399 | 78818 52 | 116750 54 | | | | | | | |
| • 400 | 79794 29 | | | | | | 83629 29 | | |
| 401 | | 117477 34 | | | | | | 23956 34 | ?× |
| • 403ff ♮ | | | | 21789 23 | 56180 25 | 45612 14 | | | |
| • 412 | | | | 21873 12 | 56277 12 | | | | |
| • 413 | 81937 81 | 119338 113 | 175417 111 | 23483 104 | 57507 118 | 47002 61 | | | 5-4-d |
| 414 | 82035 16 | | | 23604 16 | | | | | ? |
| • 415 | 82436 286 | 119799 284 | 175886 122 | 24139 181 | 57972 163 | | | | 5-4-e |
| • 416 | 82749 16 | 120098 15 | | | | | | | |
| • 417 | | | | 24368 67 | 58179 66 | | | | |

| # | HfM | | HsA | | PsA | | MsA | | DrAa | | TrAa | | DrAb | | TrAb | | Remarks |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| • 418f | 84824 | 233 | 121990 | 242 | 177482 | 244 | 27151 | 298 | 59797 | 180 | 47302 | 295 | | | | | 5-4-f ***RARE*** [89] |
| • 422 | | | 122238 | 27 | | | | | | | | | 86591 | 27 | | | |
| • 423 | 85596 | 41 | 122775 | 40 | | | | | | | | | 88770 | 23 | | | 5-4-g |
| • 424 | 85651 | 41 | 122822 | 41 | | | | | | | | | | | | | 5-4-g |
| 425 | 85787 | 19 | | | | | | | | | | | 85007 | 19 | | | ? |
| 426 | 85814 | 29 | | | | | | | | | | | 85029 | 31 | | | ? |
| • 427 ♮ | | | | | | | 27487 | 42 | | | 47629 | 42 | | | | | |
| • 428 | | | | | | | 27533 | 150 | | | 47679 | 146 | | | | | |
| • 429 ♮ | | | | | | | 28379 | 34 | | | 48327 | 39 | | | | | |
| • 430 | | | | | | | 28479 | 37 | | | 48460 | 38 | | | | | |
| 431 | | | | | | | 28648 | 44 | | | 48614 | 37 | | | | | ? |
| 432 | | | | | | | 28709 | 30 | | | 48665 | 27 | | | | | ? |
| • 433 | | | | | | | 28787 | 33 | | | 48728 | 29 | | | | | |
| • 434 | 87745 | 114 | 125173 | 76 | 179515 | 106 | 28831 | 274 | 61442 | 176 | 48768 | 272 | | | | | 5-4-pp |
| | *Hox-A4* | | | | | | | | | | | | | | | | |
| • 435 | 91064 | 132 | 128822 | 129 | | | | | | | | | | | | | 4-3-a |
| • 436 | 91515 | 58 | 129461 | 58 | | | | | | | | | | | | | |
| • 437 | 91602 | 30 | 129556 | 30 | | | | | | | | | | | | | |
| • 438 | 92853 | 91 | 131248 | 89 | | | | | | | | | | | | | |
| • 439 | 93227 | 73 | 131592 | 77 | | | | | | | | | | | | | |
| • 440 | 93311 | 42 | 131680 | 42 | | | | | | | | | | | | | |
| • 441 | 93372 | 81 | 131766 | 83 | | | | | | | | | | | | | |
| • 442 ♣ | 94873 | 34 | | | | | | | | | | | 88361 | 34 | | | |
| • 444 | 98246 | 67 | 136878 | 77 | 186333 | 83 | | | | | | | | | | | |
| • 445 | | | 136979 | 39 | 186421 | 41 | | | | | | | | | | | |
| • 446 | 98414 | 45 | 137066 | 37 | 186496 | 41 | | | | | | | | | | | |
| • 447 | 98476 | 62 | 137119 | 58 | | | | | | | | | | | | | |
| • 448 ♮ | | | | | | | | | 63246 | 21 | 50977 | 21 | | | | | |
| • 449 ♮ | | | | | | | | | 65895 | 19 | | | 87490 | 19 | | | |
| • 450 | | | | | | | | | 65919 | 45 | 54155 | 47 | | | | | |
| • 451 | 98855 | 161 | 137523 | 150 | 186937 | 113 | | | 66768 | 90 | 54889 | 99 | | | | | |
| • 452 | | | | | | | | | 66882 | 24 | 55005 | 27 | | | | | |
| • 453 | | | | | | | | | 67013 | 43 | 55144 | 42 | | | | | |
| • 454 | 99108 | 85 | 137815 | 83 | 187226 | 61 | | | 67086 | 81 | 55209 | 82 | | | | | |
| 455 | 99140 | 99 | 137815 | 83 | 187226 | 61 | | | 67086 | 77 | 55209 | 128 | | | | | ? |
| • 456 ♣ | 99764 | 29 | | | | | | | | | | | 89449 | 29 | | | |
| • 457 | 100021 | 112 | | | 187870 | 109 | | | | | | | | | | | |
| • 458 | 100163 | 12 | | | 188003 | 11 | | | | | | | | | | | |
| • 460 ♮ | | | | | | | | | 67923 | 141 | 55758 | 146 | | | | | |
| • 459 ♣ | 101851 | 29 | | | | | | | | | 56844 | 29 | | | | | ***KrA*** [74] |
| • 461 | 101931 | 276 | 140542 | 277 | 189075 | 65 | | | 69137 | 65 | 56932 | 85 | | | | | ***Hox/PBC*** ***Prep/Meis*** [74] |
| • 462 | 102585 | 77 | 141732 | 33 | 189590 | 118 | | | 69676 | 74 | 57684 | 111 | 82326 | 30 | 24929 | 53 | |
| • 463 | 102694 | 126 | 141955 | 93 | 189730 | 135 | | | | | 57903 | 24 | | | 25022 | 51 | |
| 464 | 102844 | 14 | | | 189892 | 14 | | | | | | | | | | | ? |
| • 465ff | 102907 | 280 | 142331 | 191 | 189952 | 253 | | | 70088 | 200 | 58074 | 252 | | | | | 4-3-b |
| • 477f | 105041 | 191 | 144086 | 205 | 191239 | 125 | | | 70908 | 76 | 59043 | 164 | | | 25595 | 41 | |
| • 479f | 106120 | 237 | 145095 | 245 | 192001 | 208 | | | 71522 | 205 | 59521 | 204 | | | | | 4-3-pp |
| | *Hox-A3* | | | | | | | | *Hox-A3* | | | | | | | | |
| • 481 | | | 148228 | 70 | 196002 | 67 | | | | | | | | | | | |
| • 482 | 109890 | 95 | 148351 | 96 | 196165 | 44 | | | | | | | | | | | |
| • 483 | 109999 | 217 | 148482 | 218 | 196384 | 62 | | | | | | | | | | | |
| • 484 ♮ | | | | | | | | | 73712 | 35 | | | 87719 | 35 | | | |
| 485 | | | | | | | | | | | 62877 | 30 | 84706 | 30 | | | × |
| • 486 | | | 148901 | 30 | 196614 | 65 | | | 74216 | 62 | 63470 | 48 | | | | | |

| # | HfM | HsA | PsA | MsA | DrAa | TrAa | DrAb | TrAb | Remarks |
|---|---|---|---|---|---|---|---|---|---|
| • 487f | 112888 123 | 151198 158 | 198551 154 | | 75155 165 | 65063 173 | 89629 170 | 27397 156 | |
| 489 | 113355 59 | | 199175 64 | | | | | | ? |
| 490 | 113534 69 | | 199366 75 | | | | | | ? |
| • 491 | 113657 137 | 152783 127 | 199544 137 | | | | | | 3-2-a |
| • 492ff | 113929 262 | 153128 277 | 199795 267 | | | 66175 300 | 90511 242 | 28162 263 | 3-2-pp |
| 524 | 114200 23 | | 200079 23 | | | | | | ? |
| | *Hox-A2* | | | | | *Hox-A2* | | | |
| • 525 | 116088 86 | 155551 84 | 202095 55 | | | | | | |
| • 526 | 116229 30 | 155683 30 | | | | | | | |
| 527 | 116301 11 | 155747 11 | | | | | | | ? |
| • 528 | 117329 209 | 156872 199 | 203123 220 | | | | | | 2-1-a |
| • 529 | 117798 31 | | 204874 31 | | | | | | |
| • 530 ♣ | 118642 32 | | | | | | | 35682 32 | |
| • 531 ♮ | | | | | | 70408 21 | 94869 21 | | |
| • 532 ◇ | | | 205312 25 | | | 70953 25 | | | |
| • 533 | 119948 59 | 159802 73 | 205344 71 | | 79953 29 | 70981 69 | | | |
| • 534 | 120009 125 | 159883 162 | 205446 93 | | 80042 58 | 71066 56 | | | |
| | *Hox-A1* | | | | *Hox-A1* | | | | |
| 535 | | | | | 83630 36 | | | 33838 36 | |
| 536 | | | | | | 73503 37 | | 30224 37 | |
| 537 | | | | | 86121 69 | 76990 70 | | | |
| 538 | | | | | 86214 25 | | | 38195 25 | |
| 539 | | 161549 39 | | | 92267 39 | | | | |
| 540 | 121736 18 | 161979 16 | | | | | | | |
| 541 | 121808 11 | 162032 11 | | | | | | | ? |
| 542 | 121838 56 | 162050 57 | | | | | | | |
| 543 | 122096 44 | | | | | | | 41172 44 | |
| 544 | 122218 155 | 162406 161 | 207648 99 | | 81903 69 | 72993 75 | | | |
| 545 | 122397 103 | 162592 88 | 207787 75 | | | | | 38283 30 | |
| 546 | | | | | | 84123 29 | 94644 29 | | |
| 547 | | | | | 101278 32 | 85496 32 | | | |
| 548 | | | | | | 102479 20 | | 37421 20 | |
| 549 | | | | | | 106652 31 | | 40486 31 | |
| 550 | | | | | 102743 35 | 106732 31 | | 40549 41 | |
| 551 | | | | | 107573 26 | 91180 26 | | | ? |
| 552 | | 162790 27 | | | 113979 27 | | | | |
| 553 | 122765 79 | 162923 79 | | | | | | | |
| 554 | 123177 27 | | 208449 27 | | | | | | |
| 555 | | | 209019 56 | | | 74627 56 | | | |
| 556 | | | 219903 152 | | 106436 149 | | | | |
| 557 | | | 221425 104 | | 107711 102 | | | | |
| 558 | | | 229097 29 | | | 87863 29 | | | |
| 559 | | | 233065 22 | | | 99467 22 | | | |
| 560 | | | 241154 31 | | | 101869 31 | | | |
| 561 | | | | | 114389 43 | | | 41890 43 | |
| 562 | | | | | 119410 22 | | | 45900 22 | |
| 563 | | | 243714 41 | | 126560 41 | | | | ? |
| 564 | 124150 27 | | 244948 27 | | | | | | ? |
| 565 | | | 256706 64 | | | | | 42362 70 | ? |
| 566 | | | 257265 49 | | | 111824 48 | | | ? |
| 567 | | | | 30397 274 | | 50335 273 | | | |

| | 138803.fa | 144849.fa | 144857.fa | 144868.fa | 155467.fa | 159072.fa | 159083.fa |
|---|---|---|---|---|---|---|---|
| HShoxA10.fa | 56 | 31 | 20 | 15 | 94 | 27 | 90 |
| HShoxA11.fa | 52 | 22 | 22 | 16 | 71 | 27 | 43 |
| HShoxA13.fa | 43 | 24 | 14 | 12 | 56 | 23 | 35 |
| HShoxA1.fa | 66 | 27 | 21 | 20 | 59 | 29 | 44 |
| HShoxA2.fa | 73 | 27 | 15 | 47 | 63 | 34 | 39 |
| HShoxA3.fa | 91 | 25 | 18 | 22 | 61 | 29 | 37 |
| HShoxA4.fa | 71 | 29 | 26 | 21 | 63 | 54 | 41 |
| HShoxA5.fa | 73 | 36 | 50 | 22 | 63 | 37 | 41 |
| HShoxA6.fa | 68 | 35 | 48 | 26 | 66 | 30 | 47 |
| HShoxA7.fa | 70 | 41 | 40 | 24 | 64 | 31 | 40 |
| HShoxA9.fa | 63 | 28 | 22 | 22 | 84 | 24 | 58 |
| HShoxB13.fa | 42 | 22 | 18 | 14 | 50 | 20 | 34 |
| HShoxB1.fa | 66 | 24 | 23 | 23 | 59 | 33 | 40 |
| HShoxB2.fa | 9 | 19 | 11 | 95 | 9 | 7 | 9 |
| HShoxB3.fa | 91 | 27 | 17 | 21 | 59 | 29 | 35 |
| HShoxB4.fa | 73 | 28 | 35 | 26 | 59 | 52 | 37 |
| HShoxB5.fa | 73 | 36 | 58 | 24 | 63 | 33 | 44 |
| HShoxB6.fa | 66 | 35 | 59 | 29 | 64 | 35 | 44 |
| HShoxB7.fa | 68 | 36 | 36 | 25 | 63 | 33 | 39 |
| HShoxB8.fa | 59 | 84 | 32 | 18 | 68 | 31 | 44 |
| HShoxB9.fa | 64 | 33 | 29 | 21 | 80 | 24 | 55 |
| HShoxC10.fa | 57 | 28 | 21 | 16 | 94 | 26 | 81 |
| HShoxC11.fa | 49 | 22 | 22 | 18 | 71 | 24 | 46 |
| HShoxC13.fa | 47 | 23 | 18 | 14 | 57 | 23 | 38 |
| HShoxC4.fa | 3 | 9 | 11 | 12 | 7 | 12 | 7 |
| HShoxC5.fa | 71 | 29 | 45 | 26 | 63 | 38 | 41 |
| HShoxC6.fa | 66 | 32 | 42 | 23 | 64 | 30 | 45 |
| HShoxC8.fa | 57 | 64 | 30 | 22 | 68 | 24 | 44 |
| HShoxC9.fa | 63 | 27 | 23 | 23 | 82 | 26 | 56 |
| | HShoxA3 | HShoxB8 | HShoxB5 | HShoxB2 | HShoxA10 | HShoxA4 | HShoxA10 |
| | HShoxB3 | | HShoxB6 | | HShoxC10 | HShoxB4 | |
| | 138803.fa | 144849.fa | 144857.fa | 144868.fa | 155467.fa | 159072.fa | 159083.fa |

Figure A.1. Clustalw comparison of pufferfish Hox genes against sequences of the human *HoxA*, *HoxB* and *HoxC*. This is only a small part of the table actually used, see supplemental material of [95] for more details. Colour scheme for the Percent Identities:
$40 \leq x < 50$ turquoise
$50 \leq x < 60$ green
$60 \leq x < 70$ yellow
$70 \leq x < 80$ ginger
$80 \leq x < 90$ orange
$90 \leq x$ red

# Appendix B

# Recipes

## B.1 Baked Striped Bass with Herb Stuffing [63]

### B.1.1 Materials

- 3-4 pounds striped bass, cut into two fillets
- Salt and freshly ground black pepper
- 3 or more tablespoons butter
- 1/2 cup chopped shallots or green onions (use some of green stems)
- 1 clove garlic, minced fine
- 1/2 cup finely chopped celery
- 1/2 cup coarsely chopped fresh mushrooms
- 1 tablespoon chopped fresh chervil (or 1 teaspoon dried)
- 1/2 teaspoon chopped fresh sage leaves (or 1/4 teaspoon dried)
- 1/2 teaspoon minced fresh summer savory (or 1/4 teaspoon dried)
- 1/2 teaspoon minced freesh basil (or 1/4 teaspoon dried)
- 1/4 Cup chopped fresh parsley (the Italian type is best) and parsley sprigs
- 1 cup dry white wine
- 5 slices whole-wheat bread, toasted and coarsely crumbled
- 1/4 Cup grated Parmesan cheese
- 1/4 Cup Olive Oil
- 4 slices salt pork, thinly sliced (optional)

- 1 teaspoon lemon juice and lemon wedges

### B.1.2 Methods

Preheat oven to 400 degrees F.

Rinse fish well under cold water, dry with paper towels, rub with salt and pepper inside and out.

Over moderate heat melt butter in a heavy skillet; when butter foam subsides, add shallots or onions, garlic, and celery. Reduce heat and saute about 5 minutes, or until vegetables are wilted. Stir occasionally. Add a little more butter if indicated. Turn heat to high, add mushrooms and cook 3 or 4 minutes more. Add chervil, sage, savory, basil, parsley, and 1/2 cup of the wine; stir well, reduce heat, and let simmer for several minutes.

Remove skillet from heat, stir in bread crumbs and grated cheese, lifting lightly with a fork to combine all ingredients. Additional salt and pepper may be added, if required. Allow mixture to cool slightly.

Place one fillet on shallow baking pan lined with greased foil. If fish slabs are thick and it looks as though the dish would be unwieldy when assembled, lay two pieces of twine on pan first and use to tie stuffed fillets together. Use larger piece of fish for bottom, if they differ. Arrange stuffing neatly on top of fish, then lay second fillet on that. Press stuffing inward if needed to help hold it in place. Rub top of fish with olive oil and dust lightly with salt and pepper if desired. Optional salt-pork slices should be added at this point. Tie with twine if needed.

In a small saucepan heat remaining wine and olive oil and the lemon juice. Pour it over fish and bake, uncovered, about 30 to 45 minutes, or until fish flakes easily when pierced with a fork. While the fish is baking, baste it three or four times with the liquids in the baking pan. Transfer fish to a heated platter and discard twine. Serve the fish very hot, garnished with lemon wedges and sprigs of parsley.

Note: For fewer diners, the stuffing can be carefully piled on a single bass fillet, the salt-pork slices (if used) placed on the stuffing, a dusting of salt and pepper added if desired, and the mixture of wine, olive oil, and lemon juice poured over the top. Adjust proportions and baking time accordingly. If salt pork is not used, a piece of foil placed over the fish during final 10 minutes or so of baking will help keep stuffing from drying out too much.

### B.1.3 Result

Makes 6 to 8 portions. “A Waunsinn, normal!” (PFS)

## B.2 Stewed Shark [12]

### B.2.1 Materials

- 2 1/2 lb New Jersey shark steak, cut in 1 inch cubes
- juice of 1 lime
- 1 teaspoon salt
- 1 tablespoon rum
- 3 tablespoons olive oil
- 1/4 lb salt pork, diced
- 2 large Bermuda onions, finely sliced
- 3 garlic cloves, crushed
- 2 sweet red peppers, seeded and finely sliced
- 1 can (16 oz) tomatoes, drained and chopped
- 1 cup white wine
- 1 hot seasoning pepper
- 1 teaspoon dried oregano
- 1 tablespoon chopped fresh cilantro
- salt and freshly ground black pepper

### B.2.2 Methods

Wash the shark cubes under cold water, rub with lime juice and salt then rinse under cold water.

Place in a bowl, pour over the rum and set aside for 10 minutes.

Heat the oil in a Dutch oven, add the diced salt pork and brown. Remove with a slotted spoon and drain on paper towels.

Add the onions to the oil and saute until soft, stir in the garlic and sweet red peppers and cook for 5 minutes.

Drain the shark, reserving the liquid, and add to the pot. Cook for 5 minutes, stirring frequently. Pour over the reserved liquid and add the tomatoes, white wine, seasoning pepper, oregano and cilantro. Return the salt pork to the pot and season to taste with salt and pepper. Bring to a boil, then lower the heat, cover the pot and simmer for 30 minutes, stirring frequently.

### B.2.3 Result

Makes 6 portions.

## B.3 Fugu [64]

### B.3.1 History of Fugu Eating

Eating fugu developed in Kyushu where the best edible kinds are found in abundance during the winter months. The name fugu comes from fuku (to swell). Because of its small fins, the fugu is a very slow-moving, comical-looking fish, which has evolved the peculiar defense mechanism of inhaling water into its stomach so as to turn its appetizing-looking body into a menacing ball twice its normal size. This characteristic led the Japanese to spell fugu with the Chinese characters for riverpig, and gave it the English names, balloonfish, pufferfish, swellfish, blowfish and globefish. Served only in wintertime to avoid increased toxic levels generated as a reproductive defense mechanism, fugu cuisine is regulated by Japanese law and can only be served by licensed chefs to ensure that the proper varieties are used. Fugu meat is a cross between crunchy and chewy, said by the Japanese to go shiko-shiko in one's mouth when absolutely fresh.

### B.3.2 Preparation of Fugu

The raw meat is sliced paper thin and arranged artistically in rosettes that reveal the pattern of the dish it is presented on. Whether dipped in the piquant soy, chive and bitter orange sauce or eaten as chowder, or with rice porridge, the fugu has a delicious taste.

### B.3.3 Typical Fugu menus

Fugu-sashi: Sliced raw pufferfish eaten with ponzu sauce-soy sauce juice of Japan's native bitter orange and green chives.
Fuguchiri: Literally, "shredded fugu" tossed into a rich vegetable chowder-perfect on a winter day
Fuguzosui: A rice porridge flavored with the broth from fugu cooking.
Mizutaki: A tableside boiling broth dip (merely boiled fugu) served instead of zosui
Hirezake: Toasted fugu fins can be dunked in hot sake, and are eaten as a crisp digestive

# Bibliography


[1] A. K. Abbas, A. H. Lichtman, and J. S. Pober. *Cellular and Molecular Immunology*. W.B. Saunders Company, Philadelphia, Pennsylvania, 2000.

[2] M. Akam. *Hox* genes, homeosis and the evolution of segment identity: no need for hopeless monsters. *Int. J. Dev. Biol.*, 42:445–51, 1998.

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

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

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

[6] S. Aparicio, K. Hawker, A. Cottage, Y. Mikawa, L. Zuo, B. Venkatesh, E. Chen, R. Krumlauf, and S. Brenner. Organization of the *Fugu rubripes Hox* clusters: evidence for continuing evolution of vertebrate *Hox* complexes. *Nat. Genetics*, 16:79–83, 1997.

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

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

[9] D. Bali, S. Gourley, D. D. Kostyu, N. Goel, I. Bruce, A. Bell, D. J. Walker, K. Tran, D. K. Zhu, T. J. Costello, C. I. Amos, and M. F. Seldin. Genetic analysis of multiplex rheumatoid arthritis families. *Genes Immun.*, 1:28–36, 1999.

[10] H. J. Bandelt and A. W. Dress. A canonical decomposition theory for metrics on a finite set. *Advances in mathematics*, 92(1):47–105, 1992.

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

[12] J. Bastyra. *Caribbean Cooking*. Exeter Books, New York, NY, 1987.

[13] A. Behboudi, E. Sjostrand, P. Gomez-Fabre, A. Sjoling, Z. Taib, K. Klinga-Levan, F. Stahl, and G. Levan. Evolutionary aspects of the genomic organization of rat chromosome 10. *Cytogenet. Genome Res.*, 96:52–59, 2002.

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

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

[16] J. H. Bream, M. Carrington, S. O'Toole, M. Dean, B. Gerrard, H. D. Shin, D. Kosack, W. Modi, H. A. Young, and M. W. Smith. Polymorphisms of the human IFN$\gamma$ gene noncoding regions. *Immunogenetics*, 51:50–58, 2000.

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

[18] J. E. Brown, P. A. Greenberger, and P. R. Yarnold. Soluble serum interleukin 2 receptors in patients with asthma and allergic bronchopulmonary aspergillosis. *Ann. Allergy Asthma Immunol.*, 74:484–488, 1995.

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

[20] T. L. Bugawan, D. B. Mirel, A. M. Valdes, A. Panelo, P. Pozzilli, and H. A. Erlich. Association and interaction of the IL4R, IL4, and IL13 loci with type 1 diabetes among filipinos. *Am. J. Hum. Genet.*, 72:1505–1514, 2003.

[21] J. L. Carr, C. S. Shashikant, B. W. J., and F. H. Ruddle. Molecular evolution of *Hox* gene regulation: cloning and transgenic analysis of the lamprey *HoxQ8* gene. *J. Exp. Zool.*, 280:73–85, 1998.

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

[23] A. J. Carter and G. P. Wagner. 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, 2002.

[24] C. Chavey, B. Mari, M. N. Monthouel, S. Bonnafous, P. Anglard, E. Van Obberghen, and S. Tartare-Deckert. Matrix metalloproteinases are differentially expressed in adipose tissue during obesity and modulate adipocyte differentiation. *J. Biol. Chem.*, 278:11888–11896, 2003.

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

[26] C.-H. Chiu, K. Dewar, G. P. Wagner, K. Takahashi, F. Ruddle, C. Ledje, P. Bartsch, J.-L. Scemama, E. Stellwag, C. Fried, S. J. Prohaska, P. F. Stadler, and C. T. Amemiya. Bichir *HoxA* cluster sequence reveals surprising trends in ray-finned fish genomic evolution. *Proc. Roy. Soc. B*, 2003. submitted.

[27] Y. Dai, T. Masterman, W. X. Huang, M. Sandberg-Wollheim, M. Laaksonen, H. F. Harbo, A. Oturai, L. P. Ryder, P. Soelberg-Sorensen, A. Svejgaard, and J. Hillert. Analysis of an interferon$\gamma$ gene dinucleotide-repeat polymorphism in nordic multiple sclerosis patients. *Mult. Scler.*, 7:157–163, 2001.

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

[29] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C.-H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, A. G. Rust, Z. jun Pan, M. J. Schilstra, P. J. C. Clarke, M. I. Arnone, L. Rowen, R. A. Cameron, D. R. McClay, L. Hood, and H. Bolouri. A genomic regulatory network for development. *Science*, 295:1669–1678, 2002.

[30] (DOE Joint Genome Institute). Fugu genome database, 2002.
version 2.0: `http://genome.jgi-psf.org/fugu3/fugu3.home.html`,
version 3.0: `http://genome.jgi-psf.org/fugu6/fugu6.home.html`.

[31] R. P. Donn, J. H. Barrett, A. Farhan, A. Stopford, L. Pepper, E. Shelley, N. Davies, W. E. R. Ollier, and W. Thomson. Cytokine gene polymorphisms and susceptibility to juvenile idiopathic arthritis. *Arthritis Rheum.*, 44:802–810, 2001.

[32] L. Du Pasquier and M. Flajnik. Origin and evolution of the vertebrate immune system. In W. E. Paul, editor, *Fundamental Immunology*, pages 605–650. Lippincott-Raven Publishers, Philadelphia, 1999.

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

[34] H. Escriva, L. Manzon, J. Youson, and V. Laudet. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. *Mol. Biol. Evol.*, 19:1440–1450, 2002.

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

[36] D. Ferrier and P. W. Holland. Ciona intestinalis *ParaHox* genes: evolution of *Hox/ParaHox* cluster integrity, developmental mode, and temporal colinearity. *Mol. Phylogenet. Evol.*, 24:412–417, 2002.

[37] D. E. Ferrier and P. W. Holland. Ancient origin of the *Hox* gene cluster. *Nat. Rev. Genet.*, 2:33–38, 2001.

[38] J. W. Fickett and W. W. Wasserman. Discovery and modeling of transcriptional regulatory regions. *Current Opinion in Biotech.*, 11:19–24, 2000.

[39] R. A. Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society Series A*, 98:39–54, 1935.

[40] R. A. Fisher. Confidence limits for a cross-product ratio. *Australian Journal of Statistics*, 4:41, 1962.

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

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

[43] C. Fried, S. J. Prohaska, and P. F. Stadler. Independent *Hox*-cluster duplications in lampreys. *J. Exp. Zool. (Mol. Dev. Evol.)*, 2003. submitted.

[44] J. Garcia-Fernández and P. W. Holland. Archetypal organization of the amphioxus *Hox* gene cluster. *Nature*, 370:563–566, 1994.

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

[46] A. Germot, G. Lecointre, J.-L. Plouhinec, C. Le Mentec, F. Girardot, and S. Mazan. Structural evolution of *Otx* genes in craniates. *Mol. Biol. Evol.*, 18:1668–1678, 2001.

[47] D. Graur and W.-H. Li. *Fundamentals of Molecular Evolution.* Sinauer Associates, Sunderland, Massachusetts, 2000.

[48] P. E. Graves, M. Kabesch, M. Halonen, C. J. Holberg, M. Baldini, C. Fritzsch, S. K. Weiland, R. P. Erickson, E. von Mutius, and F. D. Martinez. A cluster of seven tightly linked polymorphisms in the IL-13 gene is associated with total serum IgE levels in three populations of white children. *J. Allergy Clin. Immunol.*, 105:506–513, 2000.

[49] D. L. Gumucio, D. A. Shelton, W. Zhu, D. Millinoff, T. Gray, J. H. Bock, J. L. Slightom, and M. Goodman. Evolutionary strategies for the elucidation of cis and trans factors that regulate the developmental switching programs of the beta-like globin genes. *Mol. Phylogenet. Evol.*, 5:18–32, 1996.

[50] B. K. Hall. *Evolutionary Developmental Biology.* Chapman & Hall, New York, 1992.

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

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

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

[54] T. D. Howard, G. H. Koppelman, J. Xu, S. L. Zheng, D. S. Postma, D. A. Meyers, and E. R. Bleecker. Gene-gene interaction in asthma: IL4RA and IL13 in a dutch population with asthma. *Am. J. Hum. Genet*, 70:230–236, 2002.

[55] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. *J. Mol. Biol.*, 296:1205–1214, 2000.

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

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

[58] M. Jahromi, A. Millward, and A. Demaine. A CA repeat polymorphism of the IFN$\gamma$ gene is associated with susceptibility to type 1 diabetes. *J. Interferon Cytokine Res.*, 20:187–190, 2000.

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

[60] J. Kaufman. The origins of the adaptive immune system: whatever next? *Nat. Immunol*, 3:1124–1125, 2002.

[61] A. Khani-Hanjani, D. Lacaille, D. Hoar, A. Chalmers, D. Horsman, M. Anderson, R. Balshaw, and P. Keown. Association between dinucleotide repeat in non-coding region of interferon$\gamma$ gene and susceptibility to, and severity of, rheumatoid arthritis. *Lancet*, 2:820–825, 2000.

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

[63] L. T. King and J. S. Wexlar. *The Martha's Vineyard Cookbook.* The Globe Pequot Press, Old Saybrook, Conn., 1993.

[64] Kitchenwizard.
`http://www.greatworldchefs.com/fugu.html`.

[65] E. G. L. Koh, K. Lam, A. Christoffels, M. V. Erdmann, S. Brenner, and B. Venkatesh. *Hox* gene clusters in the indonesian coelacanth, *Latimeria menadoensis*. *Proc. Natl. Acad. Sci. USA*, 100:1084–1088, 2003.

[66] A. Krause, N. Scaletta, J. D. Ji, and L. B. Ivashkiv. Rheumatoid arthritis synoviocyte survival is dependent on STAT3. *J. Immunol.*, 169:6610–6616, 2002.

[67] H. Le, G. Lecointre, and R. Perasso. A 28s rRNA-based phylogeny of the gnathostomes: first steps in the analysis of conflict and congruence with morphologically based cladograms. *Mol. Phylogenet Evol.*, 2:31–51, 1993.

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

[69] Z. Liu and J. Klominek. Regulation of matrix metalloprotease activity in malignant mesothelioma cell lines by growth factors. *Thorax*, 58:198–203, 2003.

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

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

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

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

[74] M. Manzanares, S. Bel-Vialar, L. Ariza-McNaughton, E. Ferretti, H. Marshall, M. M. Maconochie, F. Blasi, and R. Krumlauf. Independent regulation of initiation and maintenance phases of *Hoxa3* expression in the vertebrate hindbrain involve auto- and cross-regulatory mechanisms. *Development*, 128:3595–3607, 2001.

[75] P. Martinez, J. P. Rast, C. Arenas-Mena, and E. H. Davidson. Organization of an echinoderm *Hox* gene cluster. *Proc. Natl. Acad. Sci. U S A*, 16:1469–1474, 1999.

[76] M. B. McAlexander and L.-y. Yu-Lee. Prolactin activation of IRF-1 transcription involves changes in histone acetylation. *FEBS Letters*, 488:91–94, 2001.

[77] L. McCue, W. Thompson, C. Carmack, M. Ryan, J. Liu, V. Derbyshire, and C. Lawrence. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nuc. Acids Res.*, 29:774–782, 2001.

[78] W. McGinnis and R. Krumlauf. Homeobox genes and axial patterning. *Cell*, 68:283–302, 1992.

[79] D. Meulemans and M. Bronner-Fraser. Amphioxus and lamprey AP-2 genes: implications for neural crest evolution and migration patterns. *Development*, 129:4953–4962, 2002.

[80] I. Y. Millwood, M. T. Bihoreau, D. Gauguier, G. Hyne, E. R. Levy, R. Kreutz, G. M. Lathrop, and A. P. Monaco. A gene-based genetic linkage and comparative map of the rat X chromosome. *Genomics*, 40:253–261, 1997.

[81] N. A. Mitchison. Polymorphism in regulatory gene sequences. *Genome Biology*, 2:2001.1–2001.6, 2000.

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

[83] R. Nagarkatti, C. B. Rao, J. P. Rishi, R. Chetiwal, V. Shandilya, V. Vijayan, R. Kumar, H. K. Pemde, S. K. Sharma, S. Sharma, A. B. Singh, S. V. Gangal, and B. Ghosh. Association of IFN$\gamma$ gene polymorphism with asthma in the indian population. *J. Allergy Clin. Immunol.*, 110:410–412, 2002.

[84] F. Nakao, K. Ihara, K. Kusuhara, Y. Sasaki, N. Kinukawa, A. Takabayashi, S. Nishima, and T. Hara. Association of IFN-$\gamma$ and IFN regulatory factor 1 polymorphisms with childhood atopic asthma. *J. Allergy Clin. Immunol.*, 107:499–504, 2001.

[85] A. H. Neidert, V. Virupannavar, G. W. Hooker, and J. A. Langeland. Lamprey *Dlx* genes and early vertebrate evolution. *Proc. Natl. Acad. Sci. USA*, 98:1665–1670, 2001.

[86] C. E. Nelson, B. A. Morgan, A. C. Burke, E. Laufer, E. DiMambro, L. C. Murtaugh, E. L. Gonzales, Tessasollo, L. Parada, and T. C. Analysis of *Hox* gene expression in the chick limb bud. *Development*, 122:1449–1466, 1996.

[87] M. A. Noack K, Zardoya R. The complete mitochondrial DNA sequence of the bichir (polypterus ornatipinnis), a basal ray-finned fish: ancient establishment of the consensus vertebrate gene order. *Genetics*, 144:1165–1180, 1996.

[88] S. Ohno. *Evolution By Gene Duplication.* Springer Verlag, Heidelberg, Germany, 1970.

[89] A. I. Packer, D. A. Crotty, V. A. Elwell, and D. J. Wolgemuth. Expression of the murine *Hoxa4* gene requires both autoregulation and a conserved retinoic acid response element. *Development*, 125:1991–1998, 1998.

[90] J. Pendleton, B. K. Nagai, M. T. Murtha, and F. H. Ruddle. Expansion of the *Hox* gene family and the evolution of chordates. *Proc. Natl. Acad. Sci. USA*, 90:6300–6304, 1993.

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

[92] V. E. Prince, L. Joly, M. Ekker, and R. K. Ho. Zebrafish *Hox* genes: genomic organization and modified colinear expression patterns in the trunk. *Development*, 125:407–420, 1998.

[93] S. J. Prohaska. Master thesis, 2003.

[94] S. J. Prohaska, C. Fried, C. T. Amemiya, F. H. Ruddle, G. P. Wagner, and P. F. Stadler. The shark *HoxN* cluster is homologous to the human *HoxD* cluster. *J. Mol. Evol.*, 2003. submitted.

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

[96] R. Raff. *The Shape of Life.* Chicago University Press, Chicago, IL, 1996.

[97] A. I. Reid and S. J. Gaunt. Colinearity and non-colinearity in the expression of *Hox* genes in developing chick skin. *Int. J. Dev. Biol.*, 46:13–23, 2002.

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

[99] F. H. Ruddle, K. L. Bentley, M. T. Murtha, and N. Risch. Gene loss and gain in the evolution of the vertebrates. *Development*, (Supplement):155–161, 1994.

[100] A. J. Sandford, T. Chagani, S. Zhu, T. D. Weir, T. R. Bai, J. J. Spinelli, J. M. Fitzgerald, N. A. Behbehani, W. C. Tan, and P. D. Pare. Polymorphisms in the IL4, IL4RA, and FCERIB genes and asthma severity. *J. Allergy Clin. Immunol*, 106:135–140, 2000.

[101] (Sanger Institute). The *Danio rerio* sequencing project, 2002. `http://www.sanger.ac.uk/Projects/D_rerio/`.

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

[103] P. H. Sellers. Pattern recognition in genetic sequences by mismatch density. *Bull. Math. Biol.*, 46:501–514, 1984.

[104] C. Semple and M. Steel. *Phylogenetics*. Oxford University Press, Oxford UK, 2003.

[105] A. C. Sharman and H. P. W. Estimation of *Hox* gene cluster number in lampreys. *Int. J. Dev. Biol.*, 42:617–620, 1998.

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

[107] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.

[108] A. Spagnuolo, F. Ristoratore, A. Di Gregorio, F. Aniello, M. Branno, and R. Di Lauro. Unusual number and genomic organization of *Hox* genes in the tunicate ciona intestinalis. *Gene*, 309:71–79, 2003.

[109] J. Spring. Vertebrate evolution by interspecific hybridisation–are we polyploid? *FEBS Lett.*, 2:2–8, 1997.

[110] E. J. Stellwag. *Hox* gene duplications in fish. *Cell Devel. Biol.*, 10:531–540, 1999.

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

[112] D. A. Tagle, B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess, and R. T. Jones. 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, 1988.

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

[114] H. Tegoshi, G. Hasegawa, H. Obayashi, K. Nakano, Y. Kitagawa, M. Fukui, S. Matsuo, M. Deguchi, M. Ohta, M. Nishimura, N. Nakamura, and T. Yoshikawa. Polymorphisms of interferon$\gamma$ gene CA-repeat and interleukin-10 promoter region (-592a/c) in japanese type i diabetes. *Hum. Immunol.*, 63:121–128, 2002.

[115] J. D. Thompson, D. G. Higgs, and T. J. Gibson. 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, 1994.

[116] T. C. Van der Pouw Kraan, A. van Veen, L. C. Boeije, S. A. van Tuyl, E. R. de Groot, S. O. Stapel, A. Bakker, C. L. Verweij, L. A. Aarden, and J. S. van der Zee. An il-13 promoter polymorphism associated with increased risk of allergic asthma. *Genes Immun.*, 1:61–65, 1999.

[117] B. Venkatesh, M. V. Erdmann, and B. S. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. *Proc. Natl. Acad. Sci. U S A*, 98:11382–11387, 2001.

[118] S. Wada, M. Tokuoka, E. Shoguchi, K. Kobayashi, A. Di Gregorio, A. Spagnuolo, M. Branno, Y. Kohara, D. Rokhsar, M. Levine, H. Saiga, N. Satoh, and Y. Satou. A genomewide survey of developmentally relevant genes in ciona intestinalisii. genes for homeobox transcription factors. *Dev. Genes Evol.*, 2003.

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

[120] G. P. Wagner and C.-H. Chiu. The tetrapod limb: a hypothesis of its origin. *J. Exp. Zool. (Mol. Dev. Evol.)*, 291:226–240, 2001.

[121] G. W. Warr, R. W. Chapman, and L. C. Smith. Evolutionary immunobiology: new approaches, new paradigms. *Dev. Comp. Immunol.*, 27:257–262, 2003.

[122] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, and T. Meinhardt. `TRANSFAC`: an integrated system for gene expression regulation. *Nucl. Acids Res.*, 28:16–319, 2000.

[123] E. Wingender, P. Dietze, H. Karas, and R. Knuppel. Transfac: a database on transcription factors and their DNA binding sites. *Nucleic. Acids. Res.*, 24:238–241, 1996.

[124] C. T. Workman and G. D. Stormo. `ANN-Spec`: a method for discovering transcription factor binding sites with improved specificity. In *Pacific Symposium on Biocomputing*, pages 467–78, 2000.

[125] J. A. Yoder, M. E. Nielsen, C. T. Amemiya, and G. W. Litman. Zebrafish as an immunological model system. *Microbes Infect.*, 4:1469–1478, 2002.

[126] J. Zákány and D. Duboule. *Hox* genes in digit development and evolution. *Cell Tissue Res.*, 296:19–25, 1999.

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

# Curriculum vitae

## Personal details

| | |
|---|---|
| Name: | Claudia Fried |
| Date of Birth: | 17 May 1978 |
| Nationality: | Austrian |

## Education

| | |
|---|---|
| 1998 – 2003 | Diploma studies in Biology (Stzw. Genetic)<br>Thesis: Discovery of Transcription Factor Binding Sites<br>University of Vienna, Austria |
| 1998-1999 | Kolleg für Chemie (Molekularbiologie und Gentechnologie)<br>Höhere Bundes-Lehr und Versuchsanstalt für chemische Industrie, Vienna, Austria |
| 1996 – 1998 | Kolleg für technische Chemie<br>Höhere Bundes-Lehr und Versuchsanstalt für chemische Industrie, Vienna, Austria |
| 1996 | Austrian Matura<br>BORG, Mistelbach, Austria |

## Practical experience

| | |
|---|---|
| 07 – 09 2000 | Molecular biology lab<br>IAEA Agriculture and Biotechnology, Seibersdorf, Austria |
| 05 – 07 1999 | Biotechnology in Plant Production<br>Institute for Agrobiotechnology, Tulln, Austria |
| 07/1998 | Optimization of phantom materials using polymer powder sintering under vacuum.<br>Institute of biomedical technology and physics, University of Vienna, Austria |
| 08/1997 | Development of Newspaper-Offset Colours<br>SunChemical, Vienna, Austria |
| 1998-2001 | Archaeological excavations<br>Schletz, Austria |

## Publications

C-H. Chiu, K. Dewar, G. P. Wagner, K. Takahashi, F. Ruddle, C. Ledje, P. Bartsch, J-L. Scemama, E. Stellwag, C. Fried, S. J. Prohaska, P. F. Stadler, C. T. Amemiya. Bichir *HoxA* cluster sequence reveals surprising trends in ray-finned fish genomic evolution. submitted, Proc. Roy. Soc. B, 2003.

S. P. Prohaska, C. Fried, C. Flamm, G. P. Wagner, P. F. Stadler. Surveying Phylogenetic Footprints in Large Gene Clusters: Applications to Hox Cluster Duplications. submitted, Mol. Phyl. Evol., 2003.

S. P. Prohaska, C. Fried, C. T. Amemiya, F. H. Ruddle, G. P. Wagner, P. F. Stadler. The Shark HoxN cluster is homologous to the Human HoxD cluster. submitted, J. Mol. Evol., 2003.

C. Fried, S. P. Prohaska, P. F. Stadler. Independent Hox-Cluster Duplications in Lampreys. submitted, J. Exp. Zool. (Mol. Dev. Evol., 2003

C. Fried, W. Hordijk, S. P. Prohaska, C R. Stadler, P. F. Stadler. The Footprint Sorting Problem. submitted, J. Chem. Inf. Comput. Sci., 2003.

## Presentations

**Talk** Surveying phylogenetic footprints: an efficient method and an application to vertebrate Hox clusters
Computational Biology in Saxony 2003, Dresden 14.03.2003

**Poster** Detection of phylogenetic footprints in large gene clusters
2.Biotechnologie-Tag 2003, Leipzig 21.05.2003

**Poster** The Footprint Sorting Problem
MATH/CHEM/COMP 2003, Dubrovnik, 21.06.-28.06.2003

So Long, and Thanks for All the Fish