

# Surveying Phylogenetic Footprints in Large Gene Clusters: Applications to Hox Cluster Duplications

Sonja Prohaska<sup>a,b</sup>, Claudia Fried<sup>a,b</sup>, Christoph Flamm<sup>b</sup>,  
Günter P. Wagner<sup>c</sup>, Peter F. Stadler<sup>a,b,d</sup>

<sup>a</sup>*Lehrstuhl für Bioinformatik, Institut für Informatik, Universität Leipzig,  
Kreuzstraße 7b, D-04103 Leipzig, Germany.*

*{sonja,claudia,studla}@bioinf.uni-leipzig.de*

<sup>b</sup>*Institut für Theoretische Chemie und Molekulare Strukturbiologie,  
Universität Wien, Währingerstraße 17, A-1090 Wien, Austria*

*{sopr,claudia,xtof,studla}@tbi.univie.ac.at*

<sup>c</sup>*Department of Ecology and Evolutionary Biology*

*Yale University, New Haven, CT, USA*

*gunter.wagner@yale.edu*

<sup>d</sup>*The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*

---

## Abstract

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 slower than adjacent non-functional DNA. These so-called phylogenetic footprints can be detected by comparison of the sequences surrounding orthologous genes in different species. In this paper we present a new method and an efficient software tool for the identification of corresponding footprints in long sequences from multiple species. This allows the evolutionary study of the origin and loss of phylogenetic footprints if sufficient number and appropriately placed species are included. We apply this method to the published sequences of HoxA clusters of shark, human, and the duplicated zebrafish and Takifugu clusters as well as the published HoxB cluster sequences. We find that there is a massive loss of sequence conservation in the intergenic region of the HoxA clusters, consistent with the finding in [Chiu *et al.*, PNAS **99**, 5492-5497 (2002)]. We further propose a simple model to estimate the loss of sequence conservation that can be attributed to gene loss and other structural reasons. We find that the loss of conservation after cluster duplication is more extensive than expected by this model. This suggests that binding site turnover and/or adaptive modification may also contribute to the loss of sequence conservation. We conclude that this method is suitable for the large scale study of the evolution of (putative) cis-regulatory elements.

*Key words:* Phylogenetic footprints, Hox gene clusters, gene duplication

---

## 1 Introduction

Non-coding DNA in eukaryotes contains a large number functionally important signals for the regulation of gene expression. These *cis*-acting regulatory elements can be interpreted as the “hardwiring of development” at the genomic level [4]. For a recent review see [20].

Functional and non-functional parts of genomes evolve with different speeds reflecting the fact that mutations are selected against in the functional parts [17]. The technique of *Phylogenetic Footprinting* exploits these differential evolution rates for identifying regulatory elements [29].

There are two classes of approaches to identify regulatory regions. Most commonly, one searches for common motifs in the non-coding sequences associated with related genes in the same organism, see e.g. [15,27,31]. 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 was successful to identify the regulatory elements in many cases, see e.g. [18,23,29,9,6] and the review [11]. 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 [19]. Most searches for phylogenetic footprints in the past were based on computing global alignments. Standard motif search techniques such as **AlignAce** [16] and **ANN-Spec** [32] and segment-based alignment algorithm such as **DIALIGN** [24] have been shown to be more efficient [5]. Most recently footprinting was expressed as a *substring parsimony problem* and an exact and rather efficient dynamic programming algorithm was proposed and implemented [5]. 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.

In this contribution we pursue a different algorithmic approach that appears to be more suitable for large clusters of genes with 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: In [9], for example, it has been reported that — quite unexpectedly — the footprint pattern of the horn shark *Heterodontus francisci* has much more in common with the pattern in *Homo sapiens* than with other fish species (*Morone saxatilis* and *Danio rerio*). We therefore drop the maximum parsimony assumption for the evolution of regulatory sequences in large gene clusters and instead adopt a stepwise procedure that first extracts potentially conserved regions from pairwise sequence comparisons and passes these candidates through a series of filtering steps. Since our software **tracker** is intended for large-scale surveys of

large gene clusters the entire procedure has been fully automated and includes a variety of post processing and analysis steps, in particular the assembly of regions that contain footprints in various combinations of sequences to multiple sequence alignments.

As a first application of our technique we re-evaluate and extend the survey of the HoxA clusters of *Heterodontus francisci*, *Morone saxatilis*, *Danio rerio* and *Homo sapiens*. Our main result is that the automatized procedure detects a more complete set of footprints, and that it does so in less than a minute on a modern PC, in contrast to weeks of tedious analysis with web-based bioinformatics tools.

We then extend the analysis to include the two HoxA clusters of Takifugu, based on the published genomic sequence [34], to assess whether the new method provides biologically meaningful and consistent results. The purpose of the study [9] was to assess the effect of Hox cluster duplication on the structure and function of Hox genes. The qualitative results in [9] suggested that cluster duplication leads to a massive loss of non-coding sequence conservation, which could be indicative of extensive modifications in the function of Hox genes. If this is the case one would expect to find a similar degree of loss of conservation in other teleost Hox clusters. In fact we do find an even greater loss of sequence conservation in Takifugu than in zebra fish (see below).

The quantitative analysis of the retention statistic for phylogenetic footprint clusters (PFCs) has to be put into context of the other changes that happen after Hox cluster duplication. Most notably there is a tendency for gene loss after duplication [1] which can have a direct, "structural" rather than functional, influence on the retention of PFCs. To assess whether the loss of sequence conservation can be explained in its entirety by gene loss we propose a simple probabilistic model for the rate of PFC loss due to gene loss and stochastic resolution of genetic redundancy (structural causes). With this model we show that the number of PFC retained is less than what is expected from structural causes in all examined cases. This supports the idea that Hox cluster duplication can facilitate the evolution of development [22,9].

## 2 Materials

A whole-genome shotgun assembly of the genome of *Takifugu rubripes* was published recently [2]. Blast searches of the known Hox-A proteins from other species against version 3.0 of the Fugu database [34] leave little doubt that there are two Hox-A clusters. The Hox-Aa cluster is located in **scaffold 47**. It differs from its zebrafish-homologue in two features: (a) Takifugu has a Hox-10 and (b) the Hox-2 gene is retained in Takifugu. The best homologues

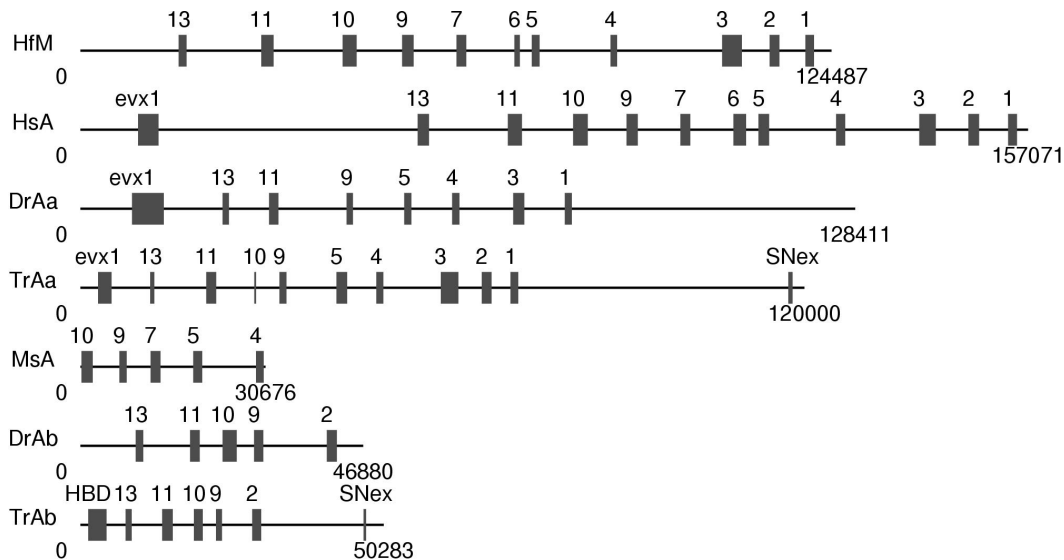


Fig. 1. Hox-A clusters used in this paper. Accession numbers and data sources:  
HsA *Homo sapiens* AC004080 (reverse complement), AC010990 (r.c., overlaps 200nt with AC004080), and AC004079 (pos. 75001-end, r.c., overlaps 200nt with AC010990), as in [9];  
HfM *Heterodontus francisici* AF479755 (as in [9]);  
DrAa *Danio rerio* HoxAa: AC107365 (reverse complement);  
DrAb *Danio rerio* HoxAb: AC107364 the first 50000nt are omitted in this drawing;  
TrAa *Takifugu rubripes* HoxAa: Fugu v.3.0 scaffold 47 positions 103001-223000 (reverse complement), contains FRU92573  
TrAb *Takifugu rubripes* HoxAb: Fugu v.2.0 scaffold 1874  
MsA *Morone saxatilis* AF089743, almost certainly an Aa cluster.

of the Hox-Ab genes of *Danio rerio* are found in scaffold 330 and scaffold 5310. In the previous release 2.0 the entire Hox-Ab cluster is contained in the single scaffold 1874. The assembly in version 2.0 is furthermore consistent with the “d-cluster” of [3]. The best `blast` hits for the proteins of the fugu v.3.0 gene model are almost exclusively Hox-A genes from human, horn shark, and teleost species. The gene inventory of scaffold 1874 is identical with the Hox-Ab cluster of the zebrafish *Danio rerio*, see also [26]. Sequences for Hox-B, Hox-C, and Hox-D clusters were obtained from Genbank, the Fugu database [34], and the web pages of the Zebrafish Sequencing Project [35], for further details see the caption of Figure 2. Prince [26] notes that Takifugu has most likely two A, one B, and one C cluster. The sequences obtained from the Fugu database [34] contain unambiguous evidence for the existence of two B-clusters in this species, see Fig. 2. Comparisons of known Hox genes with the putative hox genes of the cluster sequences are provided as supplemental material<sup>1</sup>.

<sup>1</sup> See [www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/](http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/).

### 3 Method

#### 3.1 Initial set of pairwise alignments

The program `tracker` is based upon `blastz` [28] searches of all sequences against each other. The resulting list of raw pairwise sequence alignments is then assembled into clusters of partially overlapping regions that are analyzed in detail. By default, only the intergenic regions between two homologous genes are compared. Additional (non-homologous) genes contained in one or both sequences are disregarded. For instance the IGR between Hox-A9b and Hox-A2b together with the region between Hox-A2b and SNex of Takifugu is compared with the region between Hox-A9a and SNex of the zebrafish with the exception of the exons and introns of the zebrafish Hox-A5a, Hox-A4a, Hox-A3a, and Hox-A1a genes and the Takifugu Hox-A2b gene. Formally, the combined results of all `blastz` comparisons of the input sequence  $x^1, x^2, \dots, x^N$ ,  $N \geq 3$ , form a set  $\mathfrak{A} = \{A_k | k = 1, \dots, M\}$  of alignments which is bases of all further analysis steps.

We perform the `blastz` searches with non-stringent parameters in an attempt to avoid false negative at this 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 the repetitive sequences actually connect two significantly conserved sequences. In this case we fragment the alignment into two or more shorter ones.

Local entropy measures are based on the nucleotide frequencies  $f_a(k)$  measured for a sequence window  $[k - W/2, k + W/2]$  of width  $W$  around position  $k$ . In addition, we use analogously defined joint frequencies  $f_{ab}^\tau(k)$  of finding the nucleotides  $a$  and  $b$  separated by a distance  $\tau$  along the chain. The corresponding local entropies are

$$H(k) = - \sum_a f_a(k) \log_2 f_a(k) \quad H_\tau(k) = - \sum_{a,b} f_{ab}^\tau(k) \log_2 f_{ab}^\tau(k) \quad (1)$$

Clearly,  $H(k) \leq 2\text{bit}$  and  $H_\tau(k) \leq 4\text{bit}$ . We designate a position  $k$  as having “low complexity” if both  $H(k)$  and the average mutual information measure

$$M(k) = \frac{1}{\tau_{\max}} \sum_{\tau=1}^{\tau_{\max}} H_\tau(k) - H(k) \quad (2)$$

are smaller than user-defined threshold values  $H_{\min}$  and  $M_{\min}$ , respectively.

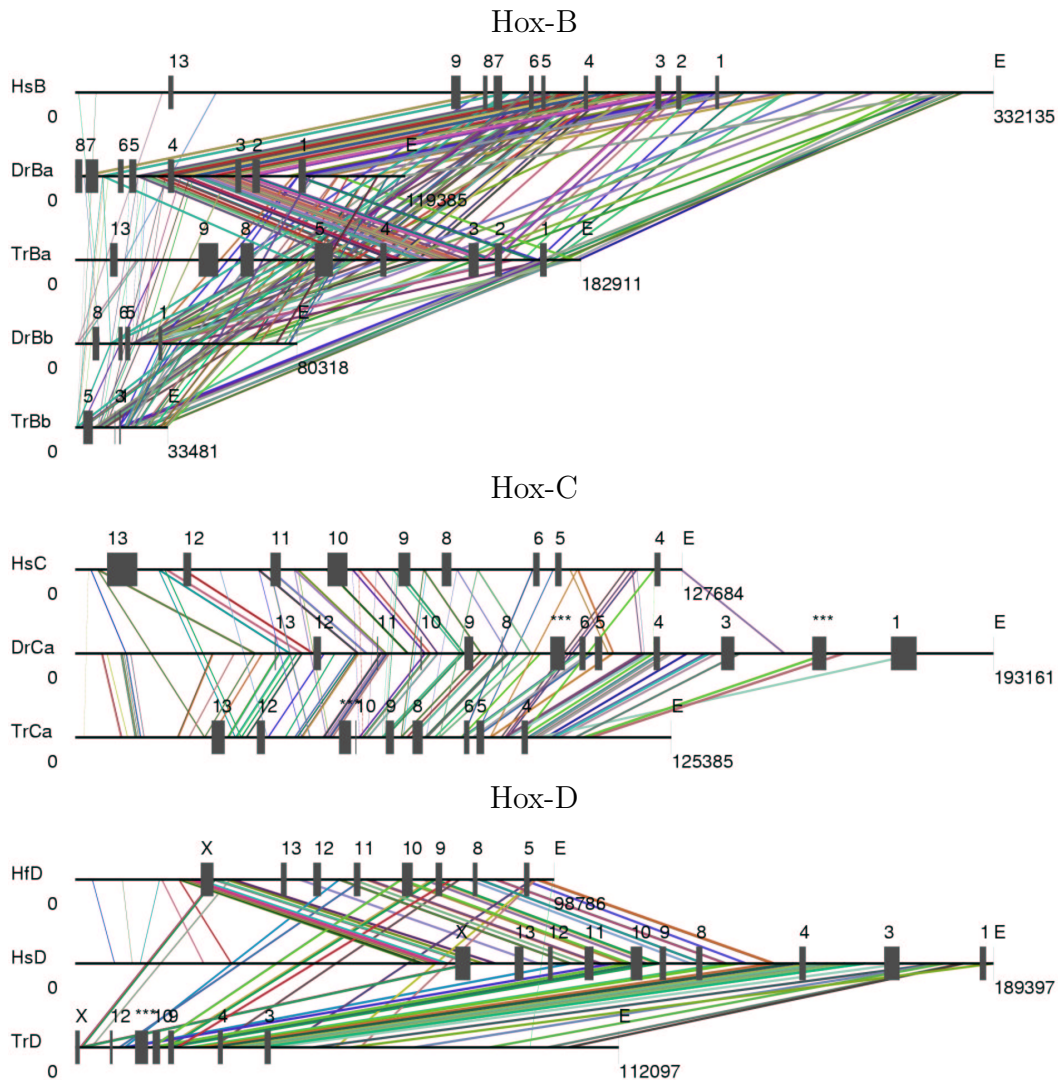


Fig. 2. Phylogenetic footprints in Hox-B, Hox-C, and Hox-D clusters. Such overviews are automatically generated by `tracker`. Each line corresponds to a footprint, consistent cliques (PFC) are shown with the same color. Input sequences were obtained as follows:

HsB = NT\_010783 [931646-1263780] reverse complement, HsC = NT\_009563 [580371-708054] r.c., HsD = NT\_037537 [4075338-end]; HfD = AF224263; DrBa = AL645782, DrBb = AL645798, DrCa is a composite of zK81P22.00296(r.c.) + 3084×N + zK81P22.01466(r.c.) + 2956×N + zK81P22.00552 from the Sanger site (download 12.1.03) with approximately 3000 Ns as spacers inserted (marked by \*\*\* in the drawing); TrBa is a composite of scaffold\_1439(r.c) + 2501×N + scaffold\_706 from version 3.0 of the Fugu DB [34], TrBb is a composite of scaffold\_1245 [59047-end] + 3020×N + scaffold\_2182 [1-19481], TrC is a composite of scaffold\_93[184545-end]+2936×N + scaffold\_285 [134158-end] (r.c.), TrD is a composite of scaffold\_3959 (r.c.) + 2645×N + scaffold\_214 [160440-end] (r.c.). All these composite sequence are consistent with a single contiguous cluster.

Table 1  
Default parameters for **tracker**.

Processing step	Parameter	Value
<b>blastz</b> 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

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. We therefore re-align the **blastz** hits using a conventional dynamic programming alignment algorithm such as **clustalw** [30] and post-process these alignments. We define the partial alignment  $\mathcal{A}[k, l]$  as sufficiently conserved if (i) contains a sequence window  $[p, p + L - 1]$  of length  $L$  in which the sequence identity is at least  $\mu_{\min}$  and (ii) if it does not contain a window of the same length  $L$  with an identity of less than  $\nu_{\max}$ .

### 3.2 Consistent Cliques

Each alignment  $A_k = \{x^p[i..j], x^q[k..l]\}$  is represented as pair of intervals  $A_k = \{A_k^1, A_k^2\} = \{x^p[i..j], x^q[k..l]\}$  where  $A_k^1 x^p[i..j]$  denotes the subsequence of input sequence  $x^p$  from positions  $i$  to  $j$ . For short, we will often write  $A_k = [p_i, p_j], [q_k, q_l]$  in the following.

We say that *two alignments A and B overlap* if there is a sequence interval  $u = A_k^1$  or  $u = A_k^2$  and a sequence interval  $v = B_l^1$  or  $v = B_l^2$  that “overlap”, i.e.,  $u \cap v \neq \emptyset$ . In the following steps it may be convenient to treat almost overlapping alignments, i.e., those that come closer than a small distance  $D_{\max}$  on one sequence, as if they were overlapping. We can view the combined results from the **blastz** scans as a graph  $\Gamma$  that has the individual **blastz**-alignments

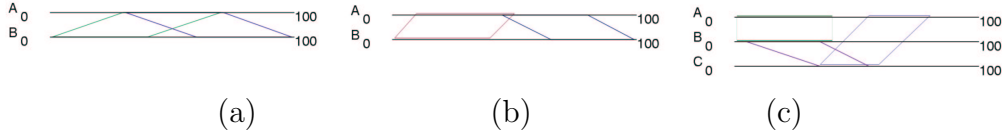


Fig. 3. (a) Two alignments that overlap in sequence  $A$  match with disjoint subsequences of  $B$ : clearly these two alignments are inconsistent in the sense that cannot even be approximately be part of a common alignment. (b) This situation on the r.h.s. is more subtle because the small overlap of only a few nucleotides might be the artifact here. In this case we might want to treat them as a single alignment with a long insertion in sequence  $B$ . (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 iff we were to disregard the  $B$ - $C$  then the  $A$ - $B$  and the  $A$ - $C$  alignments belong to different connected components.

as its vertices. The edges of  $\Gamma$  are then the (almost) 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. The second stage of a **tracker** run therefore consists of a careful analysis of the overlap graph and its constituent sequence alignments.

The first step is the decomposition of  $\Gamma$  into its connected components  $\Gamma_i$ ,  $i = 1, \dots, n_C$ , which we will refer to as “clusters”. The complicated part of the analysis is of course the further investigation of the individual clusters since they may contain mutually incompatible alignments.

From the graph-theoretical point of view it seems most natural to first consider the question whether alignments within a cluster are indeed compatible with each other, or whether they are *incompatible* in some way. Then one may define a graph  $\Psi_i$  that has the **blastz**-alignments of the cluster  $\Gamma_i$  as its vertices and has an edge between  $A$  and  $B$  if and only if  $A$  and  $B$  are incompatible. What we really want to know are 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  $\Psi_i$ ). These are efficiently computed by means of the Bron-Kerbosch algorithm [7]. It remains to specify when pairwise sequence alignments are incompatible for our purposes.

The simplest case of incompatibility involves only a 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. 3a,b. More complicated inconsistencies appear to be very rare in practical applications with few sequences. Below we describe

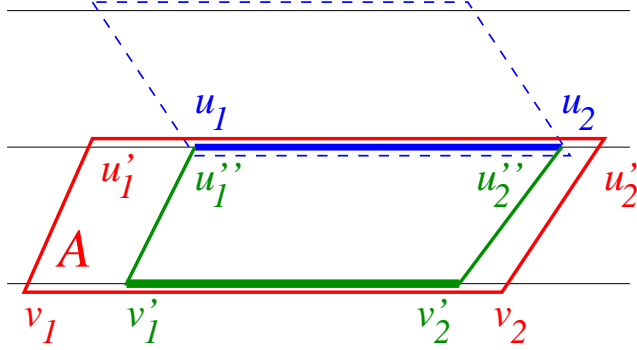


Fig. 4. Notation for the inconsistency-finding algorithm.  $[v'_1, v'_2]$  is trace of  $[u_1, u_2]$  under the alignment  $A$ . See text for details.

a general procedure for determining inconsistent alignments within a cluster which will be indispensable for larger sets of input data.

The basic idea is to consider a sequence of distinct alignments  $A_k = \{A_k^1, A_k^2\}$  such that  $A_j^2 \cap A_{j+1}^1 \neq \emptyset$ . Any such sequence corresponds to a path in the overlap graph  $\Gamma_i$ . Then we consider the image of the initial sequence interval  $A_1^1$  at each step of the sequence. Whenever  $A_k^2$  and  $A_1^1$  are parts of the same input sequence an inconsistency occurs if  $A_k^2 \not\subseteq A_1^1$ , i.e., if the image of  $A_1^1$  after a sequence of alignments is another interval on the same input sequence, see Fig. 3c. The sequences of alignments correspond to paths in the overlap graph  $\Gamma_i$ .

In order to find alignments in the cluster that are inconsistent with an alignment  $A_0 = [p_1, p_2], [q_1, q_2]$  build directed tree recursively starting with the directed edge  $[p_1, p_2] \rightarrow [q_1, q_2]$ . To each endpoint  $u$  of the growing tree (except  $[p_1, p_2]$ , of course), 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 be used already along the path from from  $[p_1, p_2]$  to  $[u_1, u_2]$ . The vertex at the endpoint of the new edge is associated with the interval  $[v'_1, v'_2]$  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. 4. We call  $[v'_1, v'_2]$  the *trace* of  $[u_1, u_2]$  under  $A_k$ . The traces can be interpreted as sequence pieces that *should* be aligned with  $[p_1, p_2]$  according to the sequence of alignments. If we arrive at a trace  $[p_1, p_2]$  such that there is an previously constructed trace  $[p'_1, p'_2]$  satisfying  $[p_1, p_2] \subseteq [p'_1, p'_2]$  that we can abandon the branch at  $[p_1, p_2]$ .

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.,  $\alpha_j = (u''_j - u'_j)/(u'_2 - u'_1)$ ;

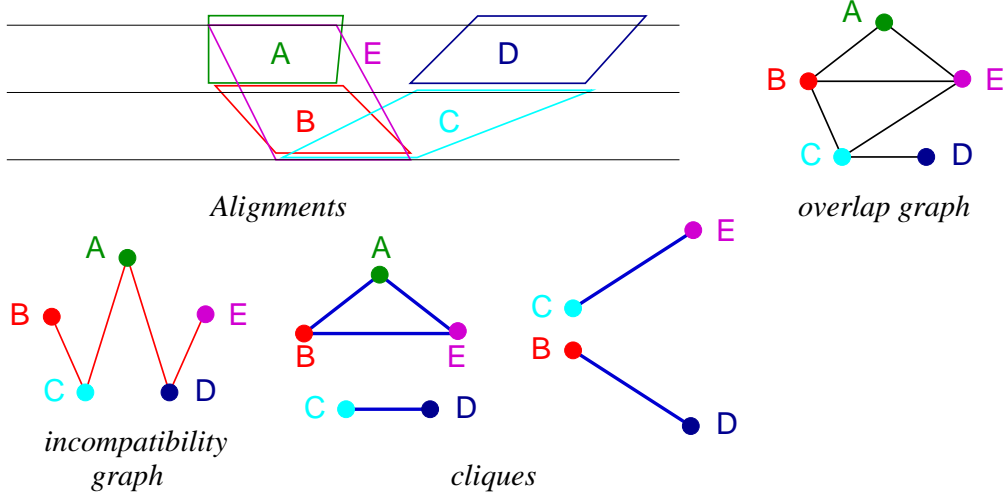


Fig. 5. Decomposition of a cluster of alignments: First the overlap graph  $\Gamma$  is computed for a set of alignments. Here we show only a single connected component (“cluster”). The incompatibility graph  $\Psi$  summarized pairs of alignments that cannot be derived from a common multiple alignment. Next cliques of its complement  $\bar{\Psi}$  are determined. Here we obtain four cliques  $C_1 = \{A, B, E\}$ ,  $C_2 = \{C, D\}$ ,  $C_3 = \{C, E\}$ , and  $C_4 = \{B, D\}$ . Only  $\Gamma[C_1]$ ,  $\Gamma[C_2]$  and  $\Gamma[C_3]$  are connected, hence obtain the revised list of cliques  $C_1, C_2, C_3, \{B\}, \{D\}$ . Neither of the two isolated points is maximal, i.e., each of them is contained in strictly larger clique, thus the final result of the decomposition are the three non-trivial cliques  $C_1, C_2,$  and  $C_3$ .

then

$$v'_j = v_1 + (u''_j - u'_1) \frac{v_2 - v_1}{u'_2 - u'_1}. \quad (3)$$

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  $b \leq d + t$ .

Now suppose that somewhere in the search tree 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. In this case we do not further extend the search tree from  $[p_1^*, p_2^*]$ .

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, and hence the graph  $\Psi_i$ . The Bron-Kerbosch algorithm [7] then produces a non-empty list  $\mathcal{C}_i = \{C_j^i\}$  of cliques. The induced subgraphs  $\Gamma_i[C_j^i]$  are not necessary connected, i.e., they might consist of alignments that do not overlap, Fig. 5. We thus revise the list of cliques by replacing  $\Gamma_i[C_j^i]$  by all its connected components. It may happen that such a component  $C'$  is a strict subset of a larger one. In this case  $C'$  is removed from the list of cliques.

Phylogenetic footprints typically appear in clusters. For the purpose of the analysis in this contribution we pragmatically define a *phylogenetic footprint cluster* (PFC) as a single consistent clique. In some case one might want to argue that two or several cliques in close proximity should only be counted as a single PFC. For example, in [9] footprints are merged into the same PFC if they are separated by less than 100nt. Since we are interested in relative abundances here this distinction is not important for our conclusions.

### 3.3 Multiple Alignments

The next step is rather straightforward. For each clique  $X$  and each sequence  $p$  we determine the minimal interval  $[p', p'']$  that contains all intervals of  $p$  appearing in alignments belonging to  $X$ . A multiple alignment of these sequence intervals is then produced using a standard program such as `clustalw` [30] or `dialign` [24]. 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.

### 3.4 Phylogenetic Distribution of Footprints

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 (not shown here).

### 3.5 Implementation

The `tracker` method is implemented as a perl program utilizing ANSI C modules e.g. for determining the inconsistency graph. Furthermore, `blastz` [28] and `clustalw` [30] as system calls. The output is provided as a L<sup>A</sup>T<sub>E</sub>X document with included Postscript figures (such as Fig. 2). The tables in the appendix are, appart from the annotation in the last column, taken directly from the `tracker` output.

## 4 A Model for the Amount of Structural Loss of PFCs

The quantitative data produced by this new algorithm for zebrafish and Takifugu is consistent with the qualitative observation in [9] for zebra fish, namely that there is a massive loss of non-coding sequence conservation associated with cluster duplication. Between 70 and 90% of the PFCs that are present in shark or human are lost after duplication.

There are three biologically distinct process that can account for this phenomenon: 1) structural, 2) binding site turnover, and 3) adaptive modification. Structural loss is the loss of putative cis-regulatory elements due to gene loss and stochastic resolution of genetic redundancy. Below we will give a more detailed account of what we think can be counted as structural loss. Binding site turnover is loss of noncoding sequence conservation due to the replacement of binding sites even though the function of the enhancer remains conserved. This was first documented in the *Drosophila* even skipped stripe 2 enhancer [21] and has since been documented for many other invertebrate taxa. In vertebrates, however, no widespread binding site turnover has been documented, which might have to do with a variety of reasons [8]. Adaptive modification would be a change in the sequence of cis-regulatory sites due to directional natural selection and would thus be associated with functional differences.

Loss of non-coding sequence conservation is associated with other structural changes, most notably gene loss. Hence the question arises whether the amount of loss observed is more than expected from the amount of gene loss. To address this question we introduce here a simple model to estimate the amount of PFC loss due to structural changes of the cluster. There are three main sources of PFC loss we consider in this model. Clearly, if a gene is lost, also the associated cis-regulatory elements will be lost, disregarding enhancer sharing. Hence the amount of loss of non-coding sequence conservation has to be calculated in relation to the number of genes which are lost in the focal clusters, in our case the *HoxA* clusters. We will express these numbers in terms of retention probabilities.

Table 2

PFC retention statistic after HoxA cluster duplication based on alignment of all seven cluster sequences

Cluster	#genes	$r(G)$	#pPFC	$r(PFC)$	$r(PFC G)$
DrHoxAa	7	0.63	39	0.31	0.49
DrHoxAb	5	0.45	29	0.23	0.51
DrHoxA	12	0.55	68	0.27	0.49
TrHoxAa	9	0.82	47	0.37	0.45
TrHoxAb	5	0.45	12	0.10	0.21
TrHoxA	14	0.64	59	0.23	0.37

Dr: zebra fish, Tr: Takifugu #genes: number of coding genes retained in cluster #pPFC: number of plesiomorphic phylogenetic footprint cluster, i.e., PFC which have a counterpart in shark or human. See text for the definition of the retention rates.

The total retention probability of an ancestral PFC,  $r(PFC)$ , depends on the retention probability assuming that the associated coding gene is retained,  $r(PFC|G)$ , and the probability that the gene is retained  $r(G)$ ,

$$r(PFC) = r(PFC|G)r(G). \quad (4)$$

In order to calculate whether the observed rate of PFC retention is larger than expected for structural reasons one thus has to consider  $r(PFC|G)$  rather than  $r(PFC)$  directly. We can estimate  $r(PFC|G)$  from the observed rate of gene  $r(PFC|G) = r(PFC)/r(G)$ . These per gene retention rates are given in Table 2 and are between 0.49 for zebrafish and 0.39 for Takifugu.

There are two other factors we need to take into account in calculating the expected loss of cis-regulatory sequence due to structural changes: the rate of loss due to (1) the loss of cross regulatory interactions among Hox genes and (2) the loss of enhancers due to stochastic resolution of genetic redundancy. The latter plays a role in cases where two paralog genes are retained. It is well known that Hox genes are cross-regulatory, i.e., a Hox gene can be the regulatory input for other Hox genes. It has been observed, both in zebrafish as well as in Takifugu that there are Hox genes that go extinct after cluster duplication, i.e. do not retain a copy of themselves in the duplicated Hox clusters. We assume that with the extinction of that gene its associated enhancer inputs to other Hox genes will be lost as well.

The expected amount of loss due to gene extinction therefore depends on the

fraction  $P(G_{\text{ext}})$  of genes in the whole Hox network that were lost and the fraction  $d$  of genes in the Hox network which received regulatory input from these extinct genes.  $P(G_{\text{ext}})$  is calculated by counting the number of paralog group members on each of the four clusters in the ancestor of bony fish, i.e. the most recent common ancestor of mouse and zebrafish, for instance. This number is compared with the number of paralog groups which are present in the two duplicated clusters of a teleost.

The number and identity of genes in the most recent common ancestor of bony fish is based on the maximal parsimony reconstruction in [1]. For instance, the ancestor of bony fish has 11 paralog group members in HoxA while zebrafish HoxAa and HoxAb only have a total of 9 paralog groups represented. In other words 18% (2) of the genes in the ancestral HoxA cluster went extinct in the zebrafish lineage, i.e. have no descendent gene copy in the zebrafish genome. In total there are 42 genes in the four ancestral Hox clusters of which only 37 have at least one descendent genes in zebrafish. This means that 12% of the genes went extinct, or  $P(G_{\text{ext}}) = 0.12$ . Similarly, in the Takifugu Hox clusters there are descendents of 34 of the 42 genes present in the ancestral Hox clusters, which means that the extinction frequency in the Takifugu lineage is  $P(G_{\text{ext}}) = 0.19$  (Chris Amemiya, pers. comm. 2003). The expected rate of PFC loss due to gene extinction is now  $d \times P(G_{\text{ext}})$ , the corresponding retention probability is therefore  $1 - dP(G_{\text{ext}})$ .

All genes in the Hox cluster arose in some time by gene duplication and are thus all paralogs. There are however different “generations” of paralogs, resulting from different gene and cluster duplication events. We call genes which are related by the most recent gene/cluster duplication 1st order paralogs. The fraction of genes which retain first order paralogs  $P(1^{\text{st}})$  differs between zebra fish and Takifugu HoxA clusters. There are six genes in zebra fish HoxA clusters which have 1st order paralogs: HoxA-13a/b, HoxA-11a/b and HoxA-9a/b. Hence the fraction of 1st order paralog genes in zebra fish is  $P(1^{\text{st}}) = 0.50$ . In Takifugu there are ten genes which have first order paralogs retained: HoxA-13a/b, HoxA-11a/b, HoxA-10a/b, HoxA-9a/b, and HoxA-2a/b; hence  $P(1^{\text{st}}) = 0.71$ . Genes which retain 1st order paralogs are expected to resolve the genetic redundancy by, on average, losing 50% of their respective cis-regulatory inputs [12]. Consequently, the larger the fraction of 1st order paralogs the larger the expected amount of PFC loss. If only one copy of the gene survives, by default one would expect that all the relevant cis-regulatory elements are maintained. Hence the probability that a PFC is lost because of stochastic resolution of genetic redundancy is equal to the probability that the associated gene has a 1st order paralog times 1/2. The

retention probability of a PFC is therefore

$$\begin{aligned} r_0 &= \left[ \frac{1}{2}P(1^{\text{st}}) + (1 - P(1^{\text{st}})) \right] (1 - dP(G_{\text{ext}})) \\ &= \left(1 - \frac{1}{2}P(1^{\text{st}})\right)(1 - dP(G_{\text{ext}})) \end{aligned} \quad (5)$$

Further, we have to consider a factor for the loss of PFCs due to non-structural causes, such as adaptation or binding site turnover. We call this probability  $\alpha$ , which means that the retention probability is  $(1 - \alpha)$ . Then the total retention rate of PFC is

$$\hat{r}(\text{PFC}|\text{G}) = r_0(1 - \alpha) = \left(1 - \frac{1}{2}P(1^{\text{st}})\right)(1 - dP(G_{\text{ext}}))(1 - \alpha) \quad (6)$$

In this model, we can determine the fraction of 1st order paralogs and the gene extinction rate, but we do not know the degree  $d$  of cross regulatory connectivity and the rate  $\alpha$  of PFC loss due to non-structural reasons. But we have the observed rate of per gene PFC retention and we can thus estimate the degree of non-structural PFC loss  $\alpha$ , by solving equ.(6) as

$$\alpha = 1 - \frac{r(\text{PFC}|\text{G})}{\left(1 - P(1^{\text{st}})/2\right)(1 - dP(G_{\text{ext}}))} \quad (7)$$

The only remaining problem is that we do not know  $d$ . We can at least obtain a lower bound estimate of the rate of non-structural PFC loss,  $\hat{\alpha} \leq \alpha$ , by assuming  $d = 1$ , i.e., that each Hox gene has a cross regulatory link to every other Hox gene:

$$\hat{\alpha} = 1 - \frac{r(\text{PFC}|\text{G})}{\left(1 - P(1^{\text{st}})/2\right)(1 - P(G_{\text{ext}}))} \quad (8)$$

In the next section we will apply this model to the analysis of the data from zebrafish and Takifugu Hox-A clusters.

An analogous analysis of the other Hox clusters (Fig. 2) is difficult at present since the sequences for Takifugu and zebrafish are incomplete and/or the corresponding outgroup sequences are not yet available.

The preliminary PFC statistics for the Hox-B clusters are compiled in Table 3. These numbers should be viewed with caution. In particular, the PFC retention rates  $r(\text{PFC})$  are upper bounds since we miss PFCs that have been lost completely in either mamalia or fish lineages. The quality of these data will improve when further outgroups, e.g., the B-cluster of bichir, become available. An additional source of uncertainty is the fact that the 3'-end of the DrBb cluster is missing in the currently available assembly, see Fig. 2.

Table 3

PFC retention statistic after HoxB cluster duplication based on alignment of all seven cluster sequences. Note that due to limited data the retention rates are only upper bounds. For the DrHoxBa cluster we count only the genes that are contained in available sequence data, see the caption of Fig. 2 for details.

Cluster	#genes	$r(G)$	#pPFC	$r(PFC)$	$r(PFC G)$
DrHoxBa	8+	0.8+	62	0.53	< 0.66
DrHoxBb	4	0.4	43	0.37	0.92
DrHoxB	12+	0.6+	105	0.45	< 0.75
TrHoxBa	8	0.8	69	0.59	0.74
TrHoxBb	3	0.3	35	0.30	1.00
TrHoxB	11	0.55	104	0.44	0.8

For the Hox-C and Hox-D clusters sequence data of duplicate clusters are currently not publicly available with sufficient data quality.

## 5 Results

There are 126 PFC that are found in either the shark or human HoxA cluster or both. In contrast, there are only 68 of those retained in at least one zebrafish clusters and 59 are retained in at least one Takifugu HoxA cluster, while only 8 and 9 PFCs, resp., survived in both paralog clusters. This corresponds to a retention rate of 27% and 23% respectively (Table 2). This confirms the qualitative observation in [9], that Hox cluster duplication is associated with a massive loss of non-coding sequence conservation. In this section we will use the model proposed above to set this rate of sequence conservation loss in relation to gene loss. But before we go into the analysis of the data we want to point out a methodological issue in scoring the rate of PFC loss in this type of data.

There are 53 PFCs in zebrafish and Takifugu that have no counterpart in shark or human; of these 14 were found only in zebrafish and 10 only in Takifugu. These PFCs most likely correspond either to cis-regulatory elements which were lost independently in the shark and human lineage or which are PFCs acquired in the stem lineage of teleost fish. These PFCs, however, cannot be used to estimate the rate of PFC retention after cluster duplication, because one can not detect the PFCs that have only been maintained in one of the paralog clusters. For that reasons we ignore the number of PFCs which have

Table 4

Conditional PFC retention statistic after HoxA cluster duplication based on the predictions of the structural loss model. Note that the predicted retention rate based on the structural loss model is consistently higher than observed rate of loss, indicating other, non-structural causes of sequence conservation loss. There is a notable asymmetry in the predicted minimal rate of non-structural conservation loss between the clusters. In zebrafish the HoxAa cluster seems to be twice as strongly modified while in Takifugu the HoxAb cluster has an exceptionally high minimal modification rate of 0.46. This pattern is consistent with rates of coding sequence evolution among paralog Hox genes in these species (Takahashi et al., in prep.).

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

no counterpart in shark or human. We have to keep in mind that the counts of PFCs are just a sample of all putative cis-regulatory elements involved. If, however, the retention rates of these PFCs are comparable to those present in shark and human, the statistics will still give valid estimates.

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

Applying our model for the structural loss of non-coding sequence conservation to the PFC data of the Hox-A clusters shows that the observed amount of retention is in all cases less than predicted as the minimal amount of retention if only structural reasons would cause loss of sequence conservation. Hence the model is consistent with the data, in the sense that we do not observe more conservation than the minimal amount predicted by this model.

Calculating the minimal probability of PFC loss, due to non-structural reasons

(binding site turnover and directional selection) shows that in zebrafish and Takifugu this rate is roughly comparable, about 26% and 29% respectively, see Table 4. The slightly higher rate in Takifugu, however, is entirely accounted for by the higher rate estimate for the HoxAb cluster. The non-structural modification rate in the HoxAa cluster is 0.22, about the same as in zebrafish, while the minimal rate of non-structural modification in the Takifugu HoxAb cluster is 48%. This suggests that there was a differential loss of non-coding sequence conservation in the Takifugu HoxAb cluster. Assuming that the probability of functionally conservative binding site turnover is about the same in the two paralog clusters, this result strongly suggests that the Takifugu HoxAb cluster experienced adaptive modification at a higher rate than both the Takifugu HoxAa cluster and either of the zebrafish clusters.

## 6 Discussion

The evolution of development is to a large part based on changes in the cis-regulatory elements of developmental genes [10]. Hence the evolutionary genetics of development requires tools for analyzing the rate and pattern of evolution of cis-regulatory sequences. This task is more difficult than the study of coding sequence evolution, because we lack a “genetic code” for the interpretation of non-coding DNA sequences. There are two approaches used in the current literature. One requires a model species for which the cis-regulatory sequences have been characterized experimentally and where the upstream factors are known. This approach provides the highest level of detail but is limited to a few well characterized genes. The other method was pioneered by Greg Wray and collaborators [14] and looks for the statistical over or under representation of known binding sites. For instance, Wray has shown that overall known binding sites are on average less frequent than expected on the basis of nucleotide frequencies in prokaryote genomes [14].

The novel computational method presented in this paper 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. The software tool presented here allows the identification of partially conserved, homologous sequences in many long 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 (very) long sequences. The survey of the seven Hox-A 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  $L$ .

Here we have applied this tool to the modifications of non-coding sequences following Hox cluster duplication in teleosts, zebrafish and Takifugu. In principle this method can also be used to study the cis-regulatory changes associated with other evolutionary changes. For instance, it is known that the AbdB related HoxD cluster genes acquired a novel pattern of regulation with the origin of the tetrapod limb [25,33]. 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 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 [13]. No comparable models are available for the analysis of non-coding sequence. In this paper we have proposed a simple model for the loss of non-coding sequence conservation after gene and cluster duplication. The purpose of this model is to estimate the amount of PFC 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. 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. It is hard to distinguish between these two factors contributing to the non-structural loss of sequence conservation. In the data set analyzed in this paper, however, we found a possible signature of adaptive loss of sequence conservation. In Takifugu the rate of non-structural sequence modification is twice as high in the HoxAb cluster than in the HoxAa cluster (see Table 4). Since there is no reason to assume that the rate of binding site turnover should be different between paralog Hox clusters, the most parsimonious interpretation is that, in Takifugu, 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.

### *Acknowledgments*

Helpful discussion with Frank Ruddle, Türker Bıyıköğlü, Chi-hua Chiu, Ivo L. Hofacker, Kazuhiko Takahashi, information on the Takifugu gene complement by Chris Amemiya, and computational assistance in the early stages of this project by Paul Perco are gratefully acknowledged.

## References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] M. I. Arnone and E. H. Davidson. The hardwiring of development: Organization and function of genomic regulatory systems. *Development*, 124:1851–1864, 1997.
- [5] M. Blanchette, B. Schwikowski, and M. Tompa. Algorithms for phylogenetic footprinting. *J. Comp. Biol.*, 9:211–223, 2002.
- [6] M. Blanchette and M. Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research*, 12:739–748, 2002.
- [7] C. Bron and J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *CACM*, 16:575–577, 1973.
- [8] 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.
- [9] 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.
- [10] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Caletani, 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.
- [11] L. Duret and P. Bucher. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, 7:399–406, 1997.

- [12] 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.
- [13] D. Graur and W.-H. Li. *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, Massachusetts, 2000.
- [14] M. W. Hahn, J. E. Stajich, and G. A. Wray. Selection against spurious transcription factor binding sites shapes genomes. *Mol. Biol. Evol.*, 2003. in press.
- [15] G. Hertz and G. D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15:563–577, 1999.
- [16] 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.
- [17] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK, 1983.
- [18] J. Y. Leung, F. E. McKenzie, A. M. Ugliarolo, 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.
- [19] 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.
- [20] M. Z. Ludwig. Functional evolution of noncoding DNA. *Curr. Op. Genet. Devel.*, 12:634–639, 2002.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] B. Morgenstern. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15:211–218, 1999.
- [25] C. E. Nelson, B. A. Morgan, A. C. Burke, E. Laufer, E. DiMambro, L. C. Murtaugh, E. L. Gonzales, T. S. Terasololo, L. Parada, and T. C. Analysis of Hox gene expression in the chick limb bud. *Development*, 122:1449–1466, 1996.

- [26] V. E. Prince. The hox paradox: More complex(es) than imagined. *Developmental Biology*, 249:1–15, 2002.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] A. Wagner. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, 15:776–784, 1999.
- [32] 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.
- [33] J. Zákány and D. Duboule. Hox genes in digit development and evolution. *Cell Tissue Res.*, 296:19–25, 1999.
- [34] 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>.
- [35] The *Danio rerio* sequencing project, 2002.  
[http://www.sanger.ac.uk/Projects/D\\_rerio/](http://www.sanger.ac.uk/Projects/D_rerio/).

## Appendix A: Tables

The analysis of the Hox-A clusters was performed in two steps. A re-evaluation of the analysis reported in [9], see Table 5ff, and a combined evaluation that uses the sequences from Takifugu as well. Tables 12ff summarize the additional footprints and has been used as the basis for the summary statistics reported in Table 2.

The **tracker** program recovers all footprints reported in [9] with the three exceptions, included in *italics* in Tables 5ff, briefly discussed below. We find that **tracker** is more sensitive, detecting about three times as many hits, some of which, however, are combined into the same PFC in [9].

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

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

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

```
HfM_up13-c      ACAGAAAACAGTTTTTGTAAAATAGTCATTTAGTATTAAAT
DrHoxAa_up13-c  -----CAAAAAAAAAAAAAAAAACACTG---
                **** * * * * *
```

**5-4-b** does not correspond to a significant match at the reported positions. The corresponding **clustalw** alignment is

```
HsA_5-4-b      --GCTGTGCTGCGATAGGGGTTGTGGGAGGGCAAAAAAAAAAAAAAAAAAGGTGATCGC--G
HfM_5-4-b      TAATTAAGAGATCGAAGCACTTTCTCCAACCTATTTAATGGAGGATGATTTATTGCCCA
                * *      **  ** * *      ** * * * * * * *
HsA_5-4-b      GGTGAGGAAAACAAGTTTCCATTCTAAACAATGGGGTGGTAGA
HfM_5-4-b      GCTAGTCAGAAAATGACCTTCTGTGCTCTCCCC----ATCTTAGA
                * * *   *** * *** * ** * *   *   ****
```

Table 5. Comparison with [9].

The last column gives the designation of the footprints from [9]. Footprints that were not found by **tracker** are listed in *italics* without numbering, + denotes novel ones. +XXX means that we found footprint also in XXX; analogously, -XXX means that the footprint was detected in the Hox cluster XXX in the previous study [9] but was not found in this sequence by the **tracker** program with the default parameter setting. Positions of footprints that are missing in some sequences in the **tracker** output are given in parentheses. Differences between the published position numbers of the DrAa sequence and our data are explained by the use of two versions of the DrAa sequence in [9].

Footprint	HfM	HsA	DrAa	DrAb	MsA	PFC 5'-3'
1	865 23			1553 23		+
2	2891 51			39450 51		+
3		8197 31		33525 31		+
4			2283 76	7560 76		+
5			2287 70	6101 70		+
6			3246 62	29283 60		+
7			7147 46	32044 46		+
8		13150 59	15129 59			+
9		13216 30	15198 31			+
10		13258 12	15241 12			+
11		15102 41		7692 47		+
12	3734 81	20391 84				+
13	3881 23		4203 23			+
14		25741 38		15607 33		+
15		27295 29		35475 29		+
16	5901 75	28483 75				+
17	5949 23	4134 23				+
18	6483 120	45120 121				upstream of 13-a
19	6775 40	45433 37				upstream of 13-b
	<i>8558 40</i>		<i>21743 19</i>			<i>upstream of 13-c</i>
20	11868 26			47489 26		+

Table 6. Table 5 continued.

Footprint	HfM		HsA		DrAa		DrAb		MsA	PFC 5'-3'
21					16307	78	22716	78		+
22					18316	42	29824	42		+
23					18387	55	29928	55		+
24	13192	120	53810	88	22652	65	58295	121		upstream of 13-d
25	13360	13					58469	14		+
26	16127	49					58996	48		13pp -DrAa +DrAb
27	16233	57					59103	56		13pp -DrAa +DrAb
28	19133	112	59505	114	25574	24				13-11-a +DrAa
29	20828	47					63519	47		+
30	27207	32			28565	32				+
31	27545	30					66363	30		+
32	27606	116	68103	118						+
33			70181	58	28402	58				+
34					29483	35	67002	35		+
35	29781	168	70665	152	31068	118	67981	132		13-11pp
36					33896	155	71142	153		11-9-a DrAa(29667)
37	34076	42			43022	42				+
38	34423	77	75337	78						11-10-a
			<i>76034</i>	<i>9</i>			<i>71322</i>	<i>9</i>		<i>11-10-b</i>
39	35043	77	76069	52	34212	31	71442	74		11-10-c +DrA
40	41272	55					71853	55		+
41			78189	21	32835	21				+
42	43143	93	81631	94			73488	75		+

Table 7. Table 5 continued.

Footprint	HfM		HsA		DrAa		DrAb		MsA		PFC 5'-3'
43	46400	43	85314	39							10-9-a
44	46546	24	85435	24							10-9-a
45	46591	188	85479	187					2977	139	10-9-a
46	47542	116	86410	116	41556	97	76755	93	3393	97	10-9-b DrAa(37297)
47							76892	16	3556	16	+
48	48333	30	87347	38	41872	35	77048	44	3791	49	10-9-c +HsA +DrAa +MsA
49	52969	35	90122	35							10-9-d
50	53030	45	90215	44							+
51	53084	55	90267	55							10-9pp -MsA(6219)
52	53229	28	90412	28							10-9pp -MsA
53	53264	42	90452	41							10-9pp -MsA
54					43987	63			6298	64	+
55					45766	47			8387	46	+
56							77140	16	3893	16	+
57							77166	94	3929	96	+
58	56953	99	94192	61	46679	175	81365	81	8912	182	9-7-a +DrAa +Drab +MsA
59	57228	219	94465	223	47016	208			9511	229	9-7-b +DrA + MsA
60	57682	31	94836	31							+
61	59503	39					87245	36			+
62			97345	38					9394	38	+
63	62154	12	99257	12							9-7-pp
64	62176	33	99279	32					11485	29	9-7-pp

Table 8. Table 5 continued.

Footprint	HfM		HsA		DrAa		DrAb		MsA		PFC 5'-3'
65	62226	107	99327	107	48807	54			11530	104	9-7-pp
66					49660	26	88070	26			+
67	66439	203	103206	206	49942	219			14805	164	7-6-a +DrAa
68	66923	24	103654	24							+
69	71720	40	108022	41							7-6-pp
70	71778	148	108078	147					16637	28	7-6-pp
71	74400	27	111988	27							+
72					53087	33			18217	34	+
73	74469	34	112053	26	53164	31			18300	31	+
74	74519	268	112101	265	53250	229			18389	228	6-5-pp
75	76119	11	114171	11							5-4-a HfM(76427)
76	76145	22	114197	22							5-4-a HfM(76427)
77	76181	22	114231	22							5-4-a HfM(76427)
78	76215	38	114264	37							5-4-a HfM(76427)
79	76266	25	114314	26							5-4-a HfM(76427)
80	76323	69	114356	70							5-4-a HfM(76427)
	<i>76648</i>	<i>63</i>	<i>114717</i>	<i>77</i>							<i>5-4-b</i>
81	76784	44	114894	44							+
82	77565	326	115543	323	55930	245			21536	250	5-4-c
83	78818	52	116743	54							+
84	79794	29					83629	29			+
85					56180	25			21789	23	+
86					56277	12			21873	12	+

Table 9. Table 5 continued.

Footprint	HfM		HsA		DrAa		DrAb		MsA		PFC 5'-3'
87	81947	71	119346	105	57520	105			23483	104	5-4-d +DrAa
88	82035	16							23604	16	+
89	82436	286	119799	284	57972	163			24139	180	5-4-e +DrAa
90	82749	16	120098	15							+
91					58177	68			24365	70	+
92	84826	231	121990	231	59802	175			27247	180	5-4-f +MsA
93			122238	27			86591	27			+
94	85596	41	122775	40			88770	23			+
95	85651	41	122822	41							5-4-g
96	85787	19					85007	19			+
97	85814	29					85029	31			+
98	87745	114	125173	76	61442	176			28922	183	+
99	91064	132	128822	129							4-3-a
100	91515	58	129461	58							+
101	91602	30	129556	30							+
102	92853	91	131248	89							+
103	93227	73	131592	77							+
104	93311	42	131680	42							+
105	93372	81	131766	83							+
106	94873	34					88361	34			+
107	98246	55	136897	58							+
108	98424	35	137066	37							+

Table 10. Table 5 continued.

Footprint	HfM		HsA		DrAa		DrAb		MsA	PFC 5'-3'
109	98476	62	137119	58						+
110	98868	148	137526	147						+
111					65895	19	87490	19		+
112	99108	85	137815	82	67086	81				+
113	99764	29					89449	29		+
114	101931	276	140542	277						+
115	102590	50			69681	56				+
116	102694	27	141968	27						+
117	102966	86	142331	46	70109	86				4-3-b
118	103058	129	142393	129	70220	50				+
119	105041	154	144063	157						+
120	105199	33	144236	32						+
121	106120	92	145095	94	71542	39				4-3-pp +HsA
122	106233	124	145205	135	71593	132				4-3-pp +HsA
123	109890	95	148351	96						+
124	109999	217	148482	218						+
125			151198	30			89631	28		+
126					73712	35	87719	35		+
127	112888	123	151235	121	75190	114	89669	117		+
128	113671	123	152783	127						3-2-a
129	113939	243	153130	247			90535	218		3-2-pp

Table 11. Table 5 continued.

Footprint	HfM		HsA		DrAa	DrAb	MsA	PFC 5'-3'
130	116088	86	155551	83				+
131	116229	30	155683	30				+
132	116301	11	155747	11				+
133	117348	99	156872	100				2-1-a
134	117460	78	156985	79				2-1-a
135	119953	54	159818	54				+
136	120009	44	159883	44				+
137	120063	69	159973	72				+
138			161549	39	92267	39		+
139	121736	18	161979	16				+
140	121808	11	162032	11				+
141	121838	56	162050	57				+
142	122218	85	162406	90				+
143	122334	39	162528	39				+
144	122397	12	162592	12				+
145	122423	25	162618	23				+
146	122483	17	162663	17				+
147			162790	27	113979	27		+
148	122765	79	162923	79				+

Table 12. Footprints in *Takifugu rubripes*.

Data from Table 5 that do not involve a *Takifugu rubripes* match are not listed. Clusters that are separated into more than one entry are sometimes merged into a single cluster here. Cluster numbers in brackets refer to Table 5.

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

Table 13. Table 12 continued.

#	HfM		HsA		DrAa		DrAb		MsA		TrAa		TrAb		Difference
71									2298	28	29389	28			
72									2340	13	29422	13			
73									2436	14	29482	14			
74									2464	17	29505	17			
75									2492	50	29536	56			
76									2581	46	29630	51			
77									2644	21	29698	20			
78									2672	93	29719	93			
79	46591	188	85479	187	41286	50			2946	174	29966	175			+TrAa +DrAa [45]
80									3139	59	30165	59			
81									3210	66	30238	67			
82									3313	20	30336	19			
83	47542	116	86411	116	41556	97	76755	95	3348	155	30366	155			+TrAa [46]
84							76892	16	3556	112	30587	92			+TrAa [47]
85									3707	22	30716	20	21531	10	
86	48333	116	87347	49	41872	35	77048	241	3742	349	30746	346	21592	212	+TrAa +TrAb [48,56,57]
87	50073	49							4812	172	31572	169			!!
93					43707	68					32112	62			
94							78511	32					22196	31	
95									5901	138	32451	133			
96									6051	15	32596	15			
97									6076	56	32626	55			
98					43987	68	78594	75	6154	222	32692	220	22274	80	+TrAa +TrAb +DrAb [54]
99									6417	51	32953	41			
100									7720	180	34183	178			
101									7913	12	34366	12			
102									7947	29	34398	29			
103									7987	46	34438	45			
104									8086	44	34534	44			
105									8154	62	34584	61			
106									8226	60	34648	57			
107					45766	47			8296	223	34717	230			+TrAa [55]
108									8534	17	34965	16			
109	56941	111	94192	62	46679	175	81365	83	8888	225	35174	225			+TrAa [58]
110									9221	11	35441	11			
111									9284	56	35493	54			

Table 14. Table 12 continued.

#	HfM		HsA		DrAa		DrAb		MsA		TrAa		TrAb		Difference
112	57228	215	97346	38	47011	213			9359	537	35554	531			!! [62,59]
113	57228	219	94466	223	47011	213			9359	537	35554	531			+TrAa [59]
115								82326	30				24929	30	
116								84706	30			62877	30		
118	59598	95										36922	95		
119			99196	28									26430	28	
121									10120	16	36280	16			
122									10199	67	36353	68			
123	62176	159	99280	157	48807	54			11415	222	37137	223			+TrAa +DrAa [64]
125									14518	34	39351	34			
126	66439	203	103206	206	49926	235			14790	215	39476	298			+TrAa [67]
127	66439	203	103206	206	49926	235			14565	97	39395	343			!! MsA(new)
130									15018	14	39785	14			
131									15098	194	39857	186			
132									15319	22	40077	22			
133									15700	67	40338	65			
134									16398	25	40811	25			
135	71778	148	108078	147					16526	139	40909	133			+TrAa [70]
137					52101	37							27738	37	
138									16856	52	41246	45			
139									16953	70	41310	67			
140									17826	70	41962	65			
141									18045	145	42174	144			
142					53081	39			18217	37	42347	43			+TrAa [72]
147	74469	318	112053	313	53164	316			18269	366	42403	356			+TrAa [73]
152	77565	326	115543	323	55930	245			21536	250	45370	239			+TrAa [82]
155			117477	34									23956	34	
156					56180	25			21789	23	45612	14			+TrAa [85]
158	81947	71	119346	105	57520	105			23483	104	47002	59			+TrAa [87]
163	84826	231	121990	231	59797	180			27151	298	47302	295			+TrAa [92]
169									27487	42	47629	42			
170									27533	150	47679	146			
171									28379	34	48327	39			
172									28479	37	48460	38			
173									28648	44	48614	37			
174									28709	30	48665	27			
175									28787	33	48728	29			
176	87745	114	125173	76	61442	176			28831	274	48768	272			+TrAa [98]

Table 15. Table 12 continued.

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