Analysis of Phylogenetic Footprint Patterns in Large Gene Clusters

Sonja J. Prohaska[1,2], Claudia Fried[1,2], Christoph Flamm[2], Peter F. Stadler[1,2]

[1]Bioinformatik, Institut für Informatik, Universität Leipzig, Kreuzstraße 7b, D-04103 Leipzig, Germany Phone: ++49 341 149 5120; Fax: ++49 341 149 5119;
Email: sonja@bioinf.uni-leipzig.de.
[2]Institut für Theoretische Chemie und Molekulare Strukturbiologie Universität Wien, Währingerstraße 17, A-1090 Wien, Austria

1. Introduction

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 resulting "islands" of strongly conserved segments — known as *phylogenetic footprints* — can be detected by comparison of the sequences surrounding orthologous genes in different species [13]. Hence it is possible 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 may serve as 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. Biologically, these changes of regulatory elements account for modifications of gene expression patterns, a major mode in particular of developmental gene evolution [4].

We developed an efficient software tool for the identification of footprints in long sequences from multiple species in order to determine the distribution of phylogenetic footprints among a set of orthologous sequences. Here we briefly compare the performance of our method with other phylogenetic footprinting methods and discuss applications to the evolution of *HoxA* clusters.

2. The Tracker Methods

The `tracker` program is designed to analyze the footprint patterns of a moderately large sample of very long (≥ 100kb) genomic sequences. The stepwise procedure, which is described in detail in Ref. [9], first extracts potentially conserved regions from pairwise sequence comparisons using `blastz` and passes these candidates through a series of filtering steps. One of them splits long alignments with low sequence similarity into smaller block with high sequence identity. Another one eliminates repetitive sequences with low complexity. The remaining pairwise alignments are assembled into clusters of partially overlapping regions that are subsequently analyzed in detail: If these clusters cannot be represented by a single multiple alignment due to conflicting pairwise alignments the clusters are decomposed into all possible consistent cliques satisfying the constraint of multiple consistency (Fig. 1). A table listing the consistent cliques with their positions and lengths in all concerned sequences is compiled. Multiple alignments of the cliques are obtained using `dialign` or `clustalw`. The final processing stage consists of arranging cliques according to presence/absence patterns and summarizing the locations of the footprints with a common distribution on the phylogenetic tree in overview charts.

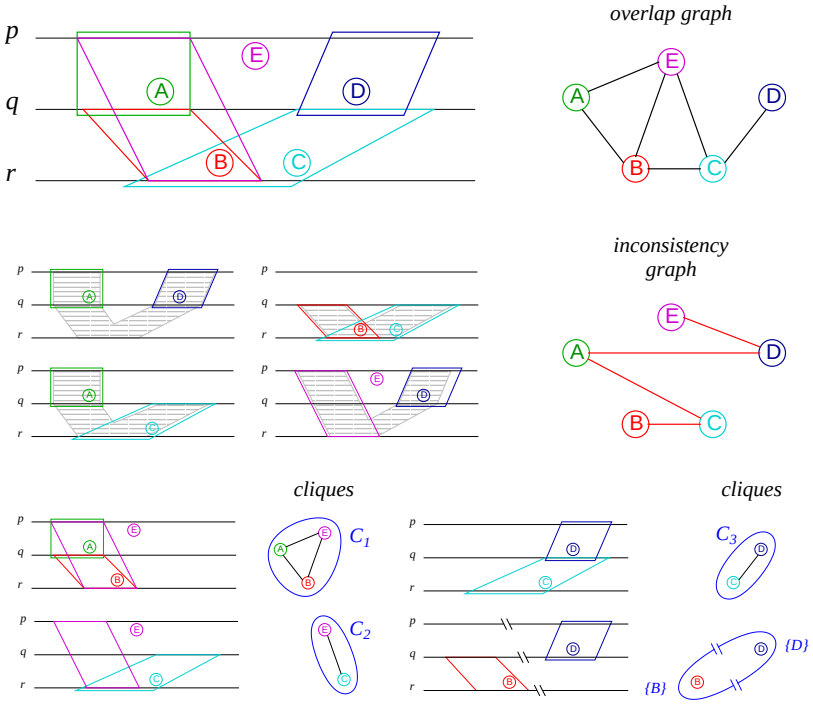

Figure 1. Decomposition of footprint clusters into consistent cliques.
Given a collection of pairwise alignments (A-E in this example) the overlap graph Γ is computed. The inconsistency graph Ψ summarizes pairs of alignments that cannot be derived from a common multiple alignment if there is a path of overlapping alignments (hatched) that connects two non-overlapping sequence fragments on the same sequence. Here we obtain four cliques $C_1 = \{\mathsf{A},\mathsf{B},\mathsf{E}\}$, $C_2 = \{\mathsf{C},\mathsf{E}\}$, $C_3 = \{\mathsf{C},\mathsf{D}\}$, 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., they are 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.

3. Performance of Tracker and Other Programs

The promising method of phylogenetic footprinting uses the search for unusually well-conserved fragments in orthologous non-coding sequences of related species. In the past, the algorithms were based on computing global alignments. Local alignments as `blastz` [12] used in `PipMaker` are more suitable. A different pairwise local alignment algorithm is implemented in `BayesAligner` [14]. Whereas standard algorithms rely on suitable scoring matrix and gap penalty parameters, `BayesAligner` returns the best alignments weighted proportional to its probability, considering the full range of gapping and scoring matrices. These methods perform pairwise comparisons and are therefore not capable of detecting multiple shared footprints without postprocessing.

Segment-based alignment algorithm such as `dialign` [8] that can cope with large sets of sequences have been shown to be more efficient. Most recently, footprinting was expressed as a *substring parsimony problem* and an exact and rather efficient dynamic programming algorithm was proposed and implemented [2]. 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 contrast, `tracker` does not rely on the phylogeny of input sequences since it was shown that changes in the footprint patterns do not necessarily correlate with established phylogenetic relationships [3].

In order to compare the performance of different footprinting programs and to assess their ability to detect potential protein binding sites, we consider the orthologous region from *hoxA4* to *hoxA3* in a variety of vertebrate species ranging from chondrichthyes (horn shark – *Heterodontus*

Table 1. Experimentally defined transcription factor binding sites [7] of the intergenic region from *hoxA4* to *hoxA3* that are correctly aligned by different methods for phylogenetic footprinting.

		KrA	Hox/PbcA	Hox/PbcB	Prep/Meis
	human	GTCAGCA	TGATTATTGAC	TCATAAATCT	TGACAA
	shark	GTCAGCA	TGCGCATTGAC	TCATAAATCT	CGACAG
TRACKER	*human / shark*	+ +	+ −	+ +	+ +
DIALIGN	*human / shark*	+ −	+ +	+ +	+ +
FOOTPRINTER	*human / shark*	− −	− −	− −	− −
BAYESALIGNER	*human / shark*	− −	− −	+ +	+ +

francisci) to acanthopterygii (zebrafish – *Danio rerio*) and sarcopterygii (human – *Homo sapiens*) since at least four experimentally determined footprints are conserved between shark and human [7], see Tab. 1.

Because of size limitations of `BayesAligner` and `FootPrinter` we restricted the comparisons to fragments of about 2000nt in length. The results are summarized in Table 1. Irrespective of the phylogenetic tree, `FootPrinter` recognized neither of the experimentally known homologous sites even though it reports a bunch of other (credible) sites. The results of `dialign` are consistent with those of `tracker`. It reports more hits at the expense of loss of specificity. An additional test on the whole *HoxA* cluster sequences demonstrated that `dialign` does not even correctly align all exons of the *Hox* genes. We therefore conclude that it can only be used to align regions of distantly related species in the range of 10000nt maximum (the length of a typical intergenic region).

4. Application to *Hox* Gene Clusters

Since `tracker` is capable of surveying footprints in large gene clusters it can be used to accumulate a sufficient amount of data for a statistical analysis of the evolution of phylogenetic footprints. We can therefore quantitatively investigate the fate of regulatory elements after *Hox* cluster duplication.

Application of `tracker` to the *HoxA* clusters of vertebrate species and the most recent *HoxA* cluster duplication in teleost fishes (pufferfish — *Takifugu rubripes* (Tr), zebrafish — *Danio rerio*, (Dr) tilapia — *Oreochromis niloticus* (On), striped bass — *Morone saxatilis* (Ms)) confirms the previous observation that horn shark and human have more footprints in common than shark and and bony fish, e.g. [3, 11]. The distribution of footprints itself may serve as an independent source of phylogenetic information, see e.g. Fig. 2. The shape of the presence/absence tree (middle) suggests that there might have been significant teleost specific modification in the footprint patterns prior to the cluster duplication.

Duplication of genes and their regulatory regions provides a rich opportunity for the modification of their function since one copy that retains the vital functions of a gene shields the second copy from negative selection. To determine whether these modifications derive from random loss of redundant cis-regulatory elements and subsequent subfunctionalisation [5] or other non-structural causes, we estimate the structural causes by the structural loss model [9] and compare the results with the data observed by `tracker`, see Tab. 2.

The comparison between `tracker` data and the predicted loss rates from structural causes shows that the footprint loss rates are two times larger than the rates expected from structural causes

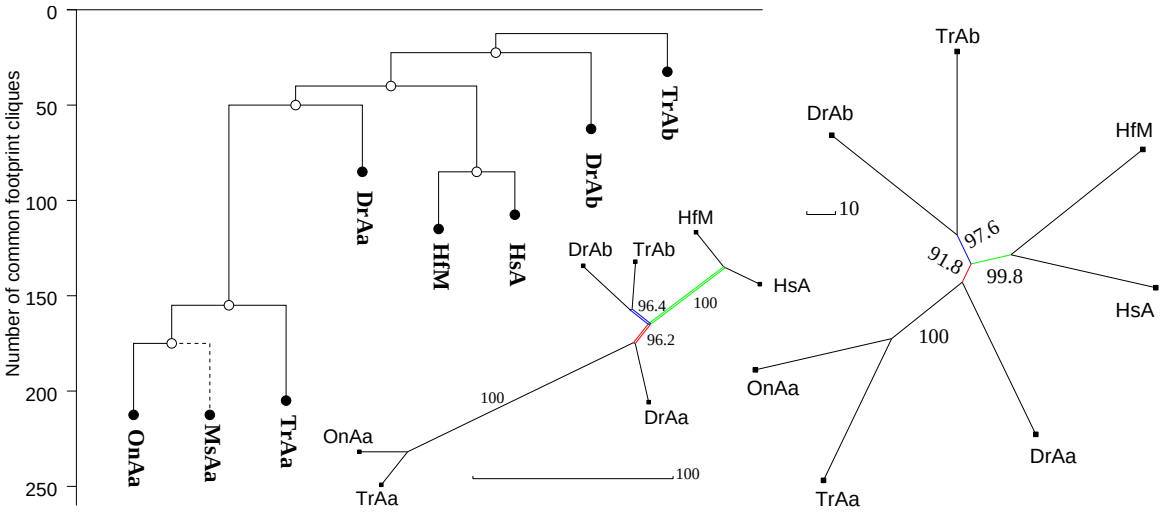

Figure 2. Co-occurrences of phylogenetic footprint cliques in *HoxA* clusters.
The leftmost tree is constructed by the weighted pairgroup clustering method using the number of footprint cliques that are shared between two clusters as similarity scores. The height of an internal node is therefore the average number of co-occurring footprints in pairs of sequences located in the two subtrees. The bass sequence (MsAa) is incomplete; we have therefore corrected the observed footprint numbers based on the assumption that the total number of cliques matches its closest neighbor tilapia in order to compute its placement in the tree. The middle tree is a parsimony split graph [1] obtained using the presence/absence of footprints as characters (339 characters, bootstrap values are shown at interior edges). The rightmost tree uses the sequences of the individual footprints treating gaps as missing characters, i.e., the phylogeny is reconstructed from the relative distance of the pairwisely conserved sequence motifs (28235 characters) using the parsimony splits method as implemented in `splitstree` [6]. Sequence data: *Heterodontus francisci* (HfM), *Homo sapiens* (HsA), *Takifugu rubripes* (TrAa and TrAb), *Danio rerio* (DrAa and DrAb), *Morone saxatilis* (MsAa) and *Oreochromis niloticus* (OnAa).

for duplicated clusters in both zebrafish and fugu (the only species for which sufficient data are available at present). This additional modification of the putative cis-regulatory elements can be explained by "binding site turnover", i.e., the co-evolution of transcription factors and their target sequence motifs, and by an enhanced rate of adaptive evolution. Binding site turnover should affect paralogous gene clusters in the same way; it is likely, therefore, that the large value for the excess probability of footprint loss $\hat{\alpha}$ observed for the fugu *Ab* cluster is a consequence of adaptive modifications during teleost phylogeny. The placement of the *HoxAb* clusters far from the clustered *HoxAa* sequences in co-occurrence tree, left part of Fig. 2, also suggests a high degree of modification after cluster duplication.

Cluster	Retention Rate		Excess
	data	model	$\hat{\alpha}$
DrAa	0.49	0.69	**0.29**
DrAb	0.51	0.62	**0.18**
DrA	0.49	0.66	0.26
TrAa	0.45	0.58	**0.22**
TrAb	0.21	0.40	**0.48**
TrA	0.37	0.52	0.29

Table 2. Conditional footprint retention statistics after *HoxA* cluster duplication based on the predictions of the structural loss model [9].
The predicted retention rate based on the structural loss model is consistently higher than the observed rate of loss, indicating other, non-structural causes of sequence conservation loss.

5. Perspectives

The novel `tracker` method for phylogenetic footprinting can handle large sets of long sequences with computational resources that bring genome-wide surveys within reach. Currently it is the only program suitable for analysis of phylogenetic footprint patterns in data sets that are large enough to provide quantitative data on non-coding sequence evolution. These data can be compared with predictions from models of gene cluster evolution. In a recent study [10] we have shown, furthermore, that footprints contain sufficient phylogenetic information to resolve questions about the homology of shark and human *Hox* clusters. Such questions are hard to tackle with other methods because of the effects of gene loss and small differences of the protein sequences. Finally, the method can be used to identify taxon-specific footprint patterns that — at least in the case of the *Hox* genes — are indicative of modification of gene expression patterns associated with important evolutionary transitions such as the innovation of the tetrapod limb.

Acknowledgments. This work is supported in part by the *DFG* Bioinformatics Initiative.

References

[1] 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.

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

[3] 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.

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

[5] 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.

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

[7] 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 phase of hoxa3 expression in the vertebrate hindbrain involve auto- and cross-regulatory mechanisms. *Development*, 128:3595–3607, 2001.

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

[9] S. Prohaska, C. Fried, C. Flamm, G. 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.

[10] 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. 2003. submitted.

[11] S. Santini, J. L. Boore, and A. Meyer. Evolutionary conservation of regulatory elements in vertebrate *Hox* gene clusters. *Genome Res.*, 13:1111–1122, 2003.

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

[13] 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.

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