Phylogenetic Footprint Patterns in Large Gene Clusters

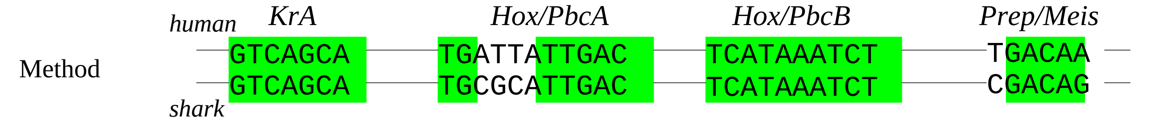
Sonja J. Prohaska[1,2], Claudia Fried[1,2], Christoph Flamm[2], Peter F. Stadler[1,2]

[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

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. 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.

We recently developed `tracker`, an efficient software tool for the identification of footprints in long sequences from multiple species, whose main purpose is to provide data for the statistical analysis of the evolution of non-coding DNA sequence in large ($\geq 100kb$) gene clusters. The stepwise procedure, which is described in detail in Ref. [5], 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, then the clusters are decomposed into all possible cliques of mutually consistent footprints using the Bron-Kerbosch algorithm. Multiple alignments of the cliques are obtained using `dialign` or `clustalw`. In this short contribution we briefly compare the performance of our method with other phylogenetic footprinting methods and discuss applications to the evolution of *HoxA* clusters.

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.



Method	KrA		Hox/PbcA		Hox/PbcB		Prep/Meis	
TRACKER	H	S	H	–	H	S	H	S
DIALIGN	H	–	H	S	H	S	H	S
FOOTPRINTER	–	–	–	–	–	–	–	–
BAYESALIGNER	–	–	–	–	H	S	H	S

Figure 1. Experimentally verified transcription factor binding sites [3] in the intergenic region from *hoxA4* to *hoxA3*. H and S indicates whether the motif is detected in human and shark, resp.

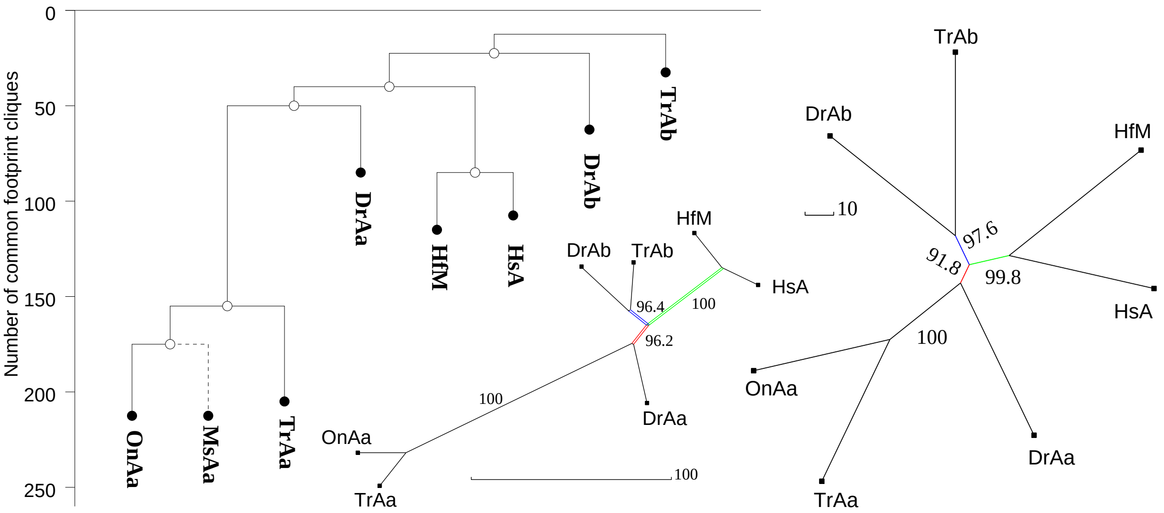


Figure 2. Co-occurrences of phylogenetic footprint cliques in *HoxA* clusters.
Left: WPGMA tree using the number of footprint cliques that are shared between two clusters as similarity scores. The height of an internal node is therefore the average number of co-occurring footprints in pairs of sequences 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. Middle: parsimony split graph obtained using the presence/absence of footprints as characters (339 characters, bootstrap values are shown at interior edges). Right: Parsimony splits tree constructed from the sequences of the individual footprints treating gaps as missing characters using `splitstree`. The phylogeny is thus reconstructed from the *relative* distances of the pairwisely conserved sequence motifs (28235 characters). 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).

Local alignments such as `blastz` used in `PipMaker` [8] are more suitable. A different pairwise local alignment algorithm is implemented in `BayesAligner` [9]. 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 algorithms such as `dialign2` [4] 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 [1]. 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 [2].

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 francisci*) to acanthopterygii (zebrafish – *Danio rerio*) and sarcopterygii (human – *Homo sapiens*) since at least four experimentally determined footprints are conserved between shark and human [3], see Fig. 1. Because of size limitations of `BayesAligner` and `FootPrinter` we restricted the comparisons to fragments of about 2000nt in length. 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).

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. 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 bony fish, e.g. [2, 7]. The distribution of footprints itself may serve as a source of phylogenetic information that is independent of protein sequences, see e.g. Fig. 2. The shape of the presence/absence tree (middle) suggests that there might have been significant teleost specific modifications in the footprint patterns prior to the cluster duplication.

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, see [5] for details. In another recent study [6] we have shown 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] M. Blanchette, B. Schwikowski, and M. Tompa. Algorithms for phylogenetic footprinting. *J. Comp. Biol.*, 9:211–223, 2002.

[2] 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.

[3] 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.

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

[5] 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. in press.

[6] 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.

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

[8] 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.

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