Surveys of phylogenetic Footprints in Large Gene Clusters

C. Fried, S.J. Prohaska, C. Flamm[1], G.P. Wagner[2], P.F. Stadler

Lehrstuhl für Bioinformatik, Institut für Informatik, Universität Leipzig, Germany
[1]Institut für Theoretische Chemie und Strukturbiologie, Universität Wien, Austria
[2]Dept. of Ecology and Evolutionary Biology, Yale University, New Haven CT.

Tel: ++49 341 149 5120 Fax: ++49 341 149 5119 Email: {claudia, sonja, xtof, studla}@bioinf.uni-leipzig.de
WorldWideWeb: http://www.bioinf.uni-leipzig.de/

Non-coding DNA in the genome of eukaryotes contains a large amount of functional regions for the regulation of genes expression. During evolution the gene regulatory regions are subject to stabilizing selection and therefore evolve much slower than adjacent non-functional DNA. These so-called phylogenetic footprints can be detected with the technique of Phylogenetic Footprinting. We present here a new method for Phylogenetic Footprinting based on the comparison of sequences surrounding orthologous genes in different species. Compared to the most recent program FootPrinter [1] our method does not require the phylogeny as input.

The program tracker[5] is designed to survey large clusters without relying on a pre-supposed phylogeny and to handle a large set of sequences at once.

The tracker approach

The program tracker is based upon local alignments of pairs of orthologous intergenic regions. As blastz searches [7] with non stringent parameters can lead to the affiliation of alignments with repetitive sequences or low conserved blocks into the list, we have to remove the concerning stretches of these elements.

We use a local entropy measures to identify and to remove repetitive sequences.

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

The measures are based on the nucleotide frequencies $f_a(k)$ and the joint frequencies $f^\tau_{ab}(k)$ of finding the nucleotides a and b separated by a distance τ along the chain. The values for $H(k)$ and $H_\tau(k)$ are low if the sequence composition is restricted and highly repeated.

Alignments consisting of a few highly conserved blocks separated by long stretches of completely non-conserved sequences are split at regions of low sequence identity.

We further construct multiple alignments based on the list of local matches. Therefore we combine *overlapping* alignments to clusters.

Alignments overlap and are clustered if they have at least one sequence from the same organism and the two sequences intervals from this organism overlap.
A cluster is a graph Γ (overlap graph) with alignments as its vertices and edges for overlaps.

Using this definition we also get clusters that cannot be represented by a common multiple alignment (as given in the example above). This incompatibility problem is resolved by designating pairs of alignments as incompatible and building maximum consistent cliques.
For each cluster we obtain a list of incompatible alignments by following a path in the overlap graph that leads to non-overlapping intervals in the same species.

The Bron-Kerbosch algorithm [2] then produces a list of cliques or subgraphs which are not necessary connected. We thus revise the list and replace the clique by all its connected components. If it happens that such a component is a strict subset of a larger one it is removed from the list of cliques.
In the final step multiple alignments of all cliques are produced using a standard program such as DIALIGN-2.

A comparison of different computational approaches to detect known binding sites is shown below:

	dialign-2		tracker		BBA	FootPr.	
Binding site	HfM	HsA	HfM	HsA		HfM	HsA
KrA site	−	+	+	+	−	−	−
HOX/PBC A	+	+	−	+	−	−	−
HOX/PBC B	+	+	+	+	+	−	−
Prep/Meis	+	+	+	+	+	−	−

BBA …Bayes block aligner makes pairwise comparisons only, in this case of the human and hornshark sequence.

In the intergenic regions between *hoxA4* and *hoxA3* four binding sites were discovered experimentally [4]. Different footprinting methods detect only some of them (+) by comparing the ***HoxA*** cluster sequences from the hornshark (***Heterodontus francisci***, HfM), human (***Homo sapiens***, HsA), bichir (***Polypterus senegalus***) and the ***HoxAa*** clusters from pufferfish (***Takifugu rubripes***) and zebrafish (***Danio rerio***).

Application to Hox gene clusters

We applied tracker to the analysis of footprints in the HoxA cluster sequences of shark, human, stripe bass (MsA), and the duplicated clusters of zebrafish (DrAa, DrAb) recently reported by Chiu et al. [3]. The purpose was to study the effect of Hox cluster duplication on the pattern of regulatory sequences in the non-coding region. Our main result is that the automatized procedure detects a more completes set of footprints. We extended the analysis to include both copies of Fugu HoxA clusters. The Buneman graph created using splitstree 3.1 shows the evolution of (co-linear) footprint cliques:

The statistical evaluation of the tracker output yields:

Cluster	Footprint loss data	model	$\hat\alpha$ excess
DrHoxAa	0.49	0.69	0.29
DrHoxAb	0.51	0.62	0.18
DrHoxA	0.49	0.66	0.26
TrHoxAa	0.45	0.58	0.22
TrHoxAb	0.21	0.40	0.48
TrHoxA	0.37	0.52	0.29

The model introduced in [5] assumes that the loss of phylogenetic footprint can be explained entirely by the loss of genes . For our analysis of the fugu clusters we observe higher rates of phylogenetic footprint loss than predicted. The deviation from the model given by $\hat\alpha$ is indicating for additional non-structural causes of conservation loss.

Overview of the phylogenetic footprint cliques produced by tracker for the comparison of the horn shark ***HoxN*** sequence (HfN) and the human ***HoxC*** (HsC) and ***HoxD*** (HsD) sequences.

The figure shows that the ***HoxN*** sequence shares more conserved sequences with the human ***HoxD*** than the ***HoxC*** cluster [6].

Buneman graphs for the parsimony splits method

Treating phylogenetic footprint cliques as presence/absence characters in a parsimony framework also supports the hypothesis that ***HfHoxN*** is more closely related to ***HsHoxD*** than to ***HsHoxC***. Further evidence for this claim can be obtained from the phylogenetic analysis of the combined footprint cliques of the four mammalian clusters for either human or rat together with the two available shark sequences.

The results suggest that the most recent common ancestor of cartilaginous fish and the bony fish clade had at least four Hox clusters orthologous to the four mammalian Hox clusters. It is thus likely that sharks have two more clusters than those currently described.

References

[1] M. Blanchette, B. Schwikowski, and M. Tompa. Algorithms for phylogenetic footprinting. *J. Comp. Biol.*, 9:211–223, 2002.
[2] C. Bron and J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *CACM*, 16:575–577, 1973.
[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] M. Manzanares, S. Bel-Vialar, L. Ariza-McNaughton, E. Ferretti, H. Marshall, M. M. Maconochie, F. Blasi, and R. Krumlauf. Independent regulation of initiation and maintenance phases of Hoxa3 expression in the vertebrate hindbrain involve auto- and cross-regulatory mechanisms. *Development*, 128:3595–3607, 2001.
[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. *J. Mol. Biol.*, 2003. submitted; SFI preprint #03-02-011.
[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. *J. Mol. Evol.*, 2003. submitted.
[7] 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.

Printed by the Universitätsrechenzentrum Leipzig