# DETECTION OF PHYLOGENETIC FOOTPRINTS IN LARGE GENE CLUSTERS

S.J. Prohaska, C.Fried, 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.

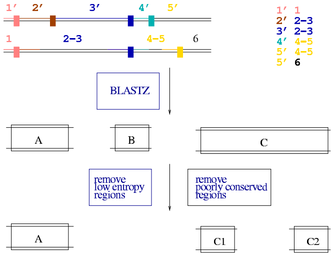
Tel: ++49 341 149 5120  Fax: ++49 341 149 5119  Email: {sonja, claudia, xtof, studla}@bioinf.uni-leipzig.de
WorldWideWeb: http://www.bioinf.uni-leipzig.de/

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 adjected non -functional DNA. These so-called phylogenetic footprints can be detected with the technique of Phylogentic Footprinting. We present here a new method for Phylogenetic Footprinting based on comparison of the sequences surrounding orthologous genes in different species which handls a large set of sequences at once. Compared to the most resent program FootPrinter [1] our method does not require the phylogeny as input.

The program tracker is designed to survey large clusters without relying on a pre-supposed phylogeny.

## The tracker approach

The program tracker is based upon local alignments of pairs of orthologous intergenetic regions. As blastz searches [5] 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.

$$H(k) = -\sum_a f_a(k)\log_2 f_a(k)$$

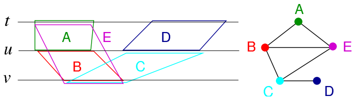
$$H_\tau(k) = -\sum_{a,b} f^\tau_{ab}(k)\log_2 f^\tau_{ab}(k)$$

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 $\tau$ along the chain. The values for $H(k)$ and $H_\tau(k)$ are low if the sequence composition is restricted and highly repeated.

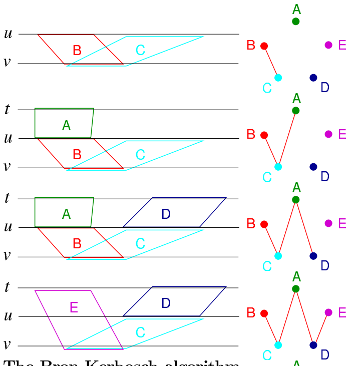
The alignments consisting of a few highly conserved blocks separated by long stretches of completly non-conserved sequences are split at regions of low sequence identity.

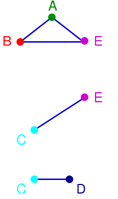
To further construct multiple alignments based on the list of local matches we combine *overlapping* alignments to clusters .

Alignments overlap and are clustered if they have at least one sequence from the same organism and the two sequences intervalls from this organism do overlap.
A cluster is a graph $\Gamma$ (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 intervalls in the same species.



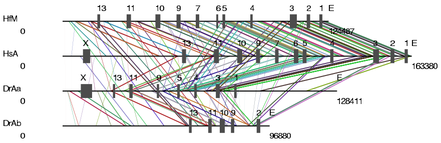


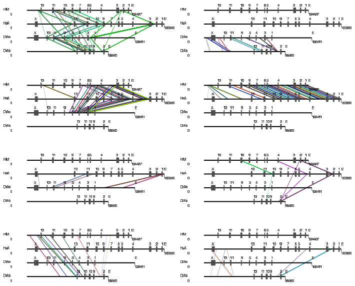
The Bron-Kerbosch algorithm [2] then produces a list of cliques or subgraphs which are not necessary connected. We thus revise the list and replace the clique by all its connected components. If it happen that such a component is a strict subset of a larger one it is removed from the list of cliques.

In the final step multiple alignments of all cliques are produced using a standard program such as DIALIGN.

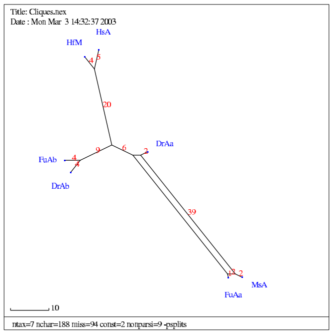
## Biological application

We applied tracker to the reeavaluation of footprint analyses in the HoxA cluster sequences of shark (HfM), human (HsA), 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.





Furthermore 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:


Title: Cliques.nex
Date : Mon Mar 3 14:32:37 2003
HfM
HsA
FuAb
DrAb
DrAa
FuAa
MsA
30
39
10
ntax=7 nchar=188 miss=94 const=2 nonparsi=9 -psplits

Statistical evaluation of the tracker output yields:

| Cluster | Footprint loss | | $\hat\alpha$ |
|---|---|---|---|
| | data | model | 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 observed rate of phylogenetic footprint loss is higher than predicted by the model introduced in [4]. It assumes that the loss of phylogenetic footprint can be explained entirely by the loss of genes. The deviation from the model given by $\hat\alpha$ is indicating for additional non-structural causes of conservation loss.

## 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] 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.

[5] 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