Surveys of Phylogenetic Footprints in Large Gene Clusters

Claudia Fried,[1,2] Sonja J. Prohaska,[1,2] Christoph Flamm,[2] Günter P. Wagner,[3] Peter F. Stadler[1,2]
[1] Bioinformatik, Institut für Informatik, Universität Leipzig, Germany.
[2] Institut für Theoretische Chemie und Strukturbiologie, Universität Wien, Austria
[3] Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT
Email: claudia@bioinf.uni-leipzig.de www: www.bioinf.uni-leipzig.de

1. Introduction

Gene regulatory regions in noncoding genomic sequences are subject to stabilizing selection and therefore evolve much slower than adjacent non-functional DNA. The resulting *phylogenetic footprints* can be detected by comparison of the sequences surrounding orthologous genes in different species [8]. Experimental evidence from a variety of sources shows that a major mode of developmental gene evolution is based on the modification of cis-regulatory elements. Therefore the loss and acquisition of conserved non-coding sequences in some lineages, but not others, provides evidence for the evolutionary modification of cis-regulatory elements.

2. The `tracker` method

The comparative analysis of long sequences, such as complete *Hox* clusters, requires a computationally efficient and fully automatized approach. The changes in the footprint patterns are not necessarily well correlated with established phylogenetic relationsships. For example, the footprint pattern in the shark *HoxA* cluster closely resembles the human distribution, while teleost fishes deviate dramatically as a consequence of an additional genome duplication. In contrast to other footprinting algorithm such as `FootPrinter` [1] we therefore do not invoke a maximum parsimony assumption. Our new program `tracker` first generates `blast` alignments of all pairs of input sequences with a non-restrictive parameter setting. A hierarchy of filtering steps then removes insignificant matches. The resulting list of pairwise alignments is then combined into clusters of overlapping footprints. The technically demanding part of the algorithm is the resolution of various types of inconsistencies that may arise when overlapping alignments of multiple sequences are combined to a multiple alignment.

In Table 1 we compare the performance of different programs for phylogenetic footprinting by comparing the orthologous region from *hoxA4* to *hoxA3* for which four experimentally verified protein binding sites were described recently [5]. None of these four binding sites is detected by `TFsearch` [9] or `FootPrinter`. `Bayes Block Aligner` [10] in general detects fewer footprints than `tracker`; `dialign`, on the other hand, is more sensitive albeit at the expense of a more than tenfold consumption of computer time and memory. It is also worth noting that `dialign` even fails to correctly align some of exons when the complete *HoxA* clusters are used as input.

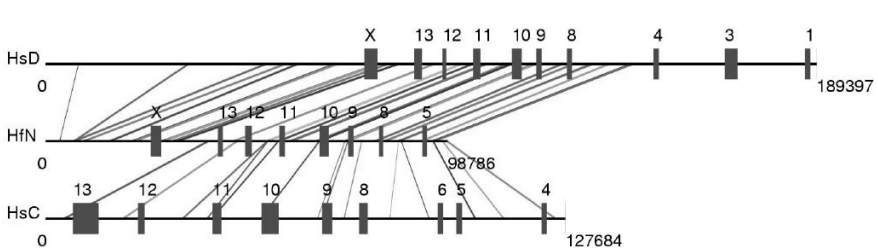
Table 1. Sensitivity of footprinting programs

We compare the recovery of four experimentally verified binding sites by different computational approaches

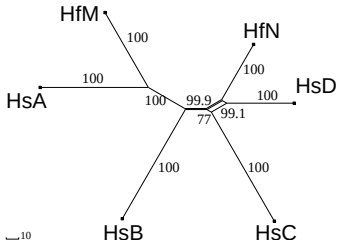
	`dialign`		`tracker`		BBA	`FootPr.`	
Binding site	Hf	Hs	Hf	Hs		Hf	Hs
KrA site	−	+	+	−	−	−	−
HOX/PBC siteA	+	+	−	−	−	−	−
HOX/PBC siteB	+	+	+	+	+	−	−
Prep/Meis	+	+	+	+	+	−	−

BBA ... `Bayes-Block-Aligner` makes pairwise comparisons only, in this case of the human and hornshark sequence.

Four binding sites in the intergenic regions between *hoxA4* and *hoxA3* were discovered experimentally [5]. Different footprinting methods detect only some of them (+) by comparing the *HoxA* cluster sequences from the hornshark (*Heterodontus francisci*, Hf), human (*Homo sapies*, Hs), bichir (*Polypterus senegalus*) and the *HoxAa* clusters from the pufferfish (*Takifugu rubripes*) and the zebrafish (*Danio rerio*).



Phylogenetic footprint cliques produced by the `tracker` program. X denotes the *Evx* genes.



Buneman graph for the parsimony splits method as implemented in `splitstree 3.1`.

Figure 1. Phylogenetic footprint patterns confirm the homology of the shark *HfN* cluster with the mammalian *HoxD* clusters.

3. Application to *Hox* gene clusters

A comparative survey of the *HoxA* clusters of hornshark, human, zebrafish and pufferfish reveals a massive loss of sequence conservation in the intergenic region, consistent with the earlier findings of Chiu *et al.* [2]. Furthermore, there is good evidence for adaptive loss of sequence conservation [6]: The rate of non-structural sequence modification is about doubled in the *HoxAb* cluster of the pufferfish compared to its *HoxAa* cluster. 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 the pufferfish, the *HoxAb* cluster experienced a higher amount of adaptive change in its cis-regulatory elements than the *HoxAa* cluster.

The distribution of footprints and the sequence conservation within footprint clusters is a useful source of phylogenetic evolution, in particular when the data from the nearby genes are hard to interpret, e.g. because of gene-loss in some species. As a first application we use the footprint pattern to resolve the relationship of the two sequenced hornshark *Hox* clusters *HfM* and *HfN* with the four mammalian clusters. It is known that the *HfM* cluster of shark is homologous to the human *HoxA* cluster[4]. The assignment of *HfN* cluster, however, remained unclear as there is evidence for homology with both the human *HoxD* and the human *HoxC* cluster [3, 4]. Statistical analysis of the footprint patterns in the *HfN* cluster shows that the shark *HfN* cluster is indeed *HoxD*-like Fig. 1. A second line of evidence is derived from concatenating the alignments of the footprint cliques (treating gaps as missing data rather than a seperate character state) [7]. Phylogenies are then reconstructed by means split-based methods which are known to be very conservative in the sense that they rather produce multifurcation than ill-supported branches. These data strongly support the homology of *HfN* to the mammalian *HoxD* clusters [7]. It follows that the most recent common ancestor of the jawed vertebrates had at least four Hox clusters, including those that are orthologous to the four mammalian Hox clusters.

References

[1] M. Blanchette, B. Schwikowski, and M. Tompa. *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. *Proc. Natl. Acad. Sci. USA*, 99:5492–5497, 2002.
[3] C. B. Kim, C. Amemiya, W. Bailey, K. Kawasaki, J. Mezey, W. Miller, S. Minosima, N. Shimizu, G. P. Wagner, and F. Ruddle. *Proc. Natl. Acad. Sci. USA*, 97:1655–1660, 2000.
[4] E. Málaga-Trillo and A. Meyer. *Amer. Zool.*, 41:676–686, 2001.
[5] M. Manzanares, S. Bel-Vialar, L. Ariza-McNaughton, E. Ferretti, H. Marshall, M. M. Maconochie, F. Blasi, and R. Krumlauf. *Development*, 128:3595–3607, 2001.
[6] S. Prohaska, C. Fried, C. Flamm, G. Wagner, and P. F. Stadler. Surveying phylogenetic footprints in large gene clusters: Applications to Hox cluster duplications. submitted; SFI preprint #03-02-011.
[7] 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.
[8] D. A. Tagle, B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess, and R. T. Jones. *J. Mol. Biol.*, 203:439–455, 1988.
[9] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prüß, I. Reuter, and F. Schacherer. *Nucl. Acids Res.*, 28:316–319, 2000.
[10] J. Zhu, J. S. Liu, and C. E. Lawrence. *Bioinformatics*, 14:25–39, 1998.