Correlation of SNPs with Phylogenetic Footprints

Claudia Fried [1], Peter Ahnert[3], Peter F. Stadler[1,2]

[1]Lehrstuhl für Bioinformatik, Institut für Informatik, Universität Leipzig, Germany
[2]Institut für Theoretische Chemie und Strukturbiologie, Universität Wien, Austria
[3]IKIT/BBZ, Medizinische Fakultät, Universität Leipzig,

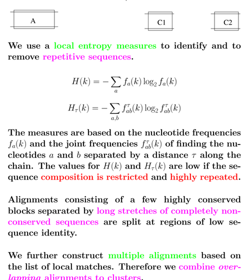
Tel: ++49 341 149 5120 Fax: ++49 341 149 5119 Email: {claudia, studla}@bioinf.uni-leipzig.de
WorldWideWeb: http://www.bioinf.uni-leipzig.de/

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 [6] 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.

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

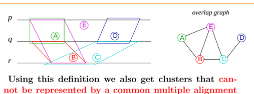
$$H_\tau(k) = -\sum_{a,b} f^\tau_{ab}(k)\log_2 f^\tau_{ab}(k)$$

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.

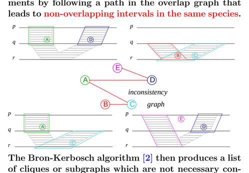
Alignments consisting of a few highly conserved blocks separated by long stretches of completely non-conserved sequences are split at regions of low sequence identity.

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.

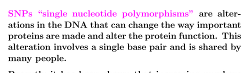


The Bron-Kerbosch algorithm [2] then produces a list of cliques or subgraphs which are not necessary connected. We thus revise the list and replace the clique by all its connected components. If it happens that such a component is a strict subset of a larger one it is removed from the list of cliques.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

SNPs versus Phylogenetic Footprints

SNPs "single nucleotide polymorphisms" are alterations in the DNA that can change the way important proteins are made and alter the protein function. This alteration involves a single base pair and is shared by many people.

Recently it has been shown that increasing numbers of polymorphism can be identified in regulatory regions of genes [5].

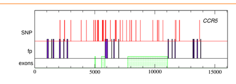


Intuitively one would expect that the rate of occurrance of SNPs in the non-functional DNA is largest since no selective pressure acts on this region, while mutations should be selected against in coding and regulatory sequences. Therefore, the frequency of SNPs in phylogenetic footprints should be significantly lower than in the surrounding non-functional DNA sequence.

For the purpose of the present study we use 2503 orthologous sequences retrieved from the EBI databases using ENSMART, which was also used to localize genes consistently with the SNP localization. We retrieved sequences for human *homo sapiens*, pufferfish *fugu rubripes* and zebrafish *danio rerio*.

For this sequences we distinguish between the frequency of SNPs in the exons, in phylogenetic footprints, and in the remaining DNA that presumably is non-functional.

Graphical representation of the SNP distribution of a typical data set:



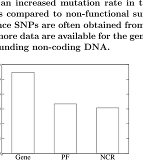
The most surprising result is that SNPs are *overrepresented* in genes. We use Fisher's exact test [3, 4] to assess the statistical signficance of the differences in SNP density in genes and footprints compared to the remaining non-functional DNA. A p-value of $p = 0.05$ or smaller indicates a significant difference in the distribution.

Distribution of SNPs in functional and non-functional sequences of 2503 genes:

		NCR	
		length	SNP
Total		49432098	15165

		Genes		Footprints	
		length	SNP	length	SNP
Total		28687928	15723	602022	201
Fisher's-	↑	0		0.111	
p-value	↓	1		0.903	

Almost certainly the overrepresentation of SNPs in genes is a bias in the SNP database: If the SNP overrepresentation in the genes were real we would have to postulate an increased mutation rate in the coding sequences compared to non-functional surrounding DNA. Since SNPs are often obtained from cDNA comparison more data are available for the genes than for the surrounding non-coding DNA.



The distribution of SNPs in tracker-predicted phylogenetic footprints does not significantly differ from distribution in non-functional DNA on average. As methods for SNP detection do not distinguish between functional and non-functional non-coding DNA we should expect that differences in SNP distribution between phylogenetic footprints and non-functional DNA are biological rather than caused by database biases.

The picture is the same for all database entries and for the small subset of validated SNPs only.

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] R. A. Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society Series A*, 98:39–54, 1935.

[4] R. A. Fisher. Confidence limits for a cross-product ratio. *Australian Journal of Statistics*, 4:41, 1962.

[5] N. A. Mitchison. Polymorphism in regulatory gene sequences. *Genome Biology*, 2:2001.1–2001.6, 2000.

[6] S. J. Prohaska, C. Fried, C. Flamm, G. P. Wagner, and P. F. Stadler. Surveying phylogenetic footprints in large gene clusters: Applications to *Hox* cluster duplications. *Mol. Phyl. Evol.*, 2003. in press.

[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