DETECTION OF PHYLLOGENETIC FOOTPRINTS IN LARGE GENE CLUSTERS

S.I. Prohaska, C. Fried, C. Flamm, G.P. Wagner, P.F. Staeller

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

Tel: +49 341 194-5200  Fax: +49 341 194-5199  Email: {lonja, claustia, xtof, studia}@bioinf.uni-leipzig.de

WWW & Web: 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 gene 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 Phyllogenetic Footprinting. We present here a new method for Phylogenetic Footprinting based on comparison of the sequences surrounding orthologous genes in different species which handle a large set of sequences at once. Compared to the most recent program FootPrinter [1] our method does not require the phylogeny as input.

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

The trainer approach

The program trainer is based on local alignments of sets of orthologous intragenic regions. As before, 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 energy measure to identify and to remove repetitive sequences. 

\[ H_2(k) = \sum_{i=1}^{n} f(k) \log_2 f(k) \]

The measure is based on the nucleotide frequencies \( f(k) \) and the joint frequencies \( g(k) \) of finding the nucleotides \( a \) and \( b \) separated by a distance \( d \) along the chains. The values for \( H_2(k) \) and \( H_4(k) \) are low if the sequence composition is restricted and highly repeated.

The alignments consisting of a few highly conserved blocks separated by large sections of completely non-conserved sequences are split at regions of low sequence identity. To further constitute 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 intervals from this organism do overlap. A cluster ion graph (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 incompleteness problem is resolved by designating pairs of alignments as incompleteness and building maximum consistent cliques.

For each cluster we obtain a list of incompleteness alignments by following a path in the overlap graph that leads to non-overlapping intervals in the same species.

Furthermore we extended the analysis to include both copies of Fags HoxA clusters. The Baumann graph centered around mitostatin 3a shows the evolution of (co-linear) footprint cliques.

Statistical evaluation of the trainer output yields

We applied trainer to the comparison of footprint analyses in the HoxA clusters sequenced in Branch (HRM), human (HsA), avian (MaA), and the duplicated clusters of zebrafish (DaeA, DaeB) recently reported by Chien et al. [3]. The purpose was to study the effect of the cluster duplication on the pattern of regulatory sequences in the non-coding region. Our main result is that the automated procedure detects a more complete set of footprints.

Bibliography


References


Produced by the University of Leipzig.