Comparative Genomics of Regulatory Sequence Elements
(Diss., Universität Leipzig, 2006**)**
Claudia Stocsits

Abstract

Transcription factor binding sites can be predicted by searching for recurring motifs in the regulatory regions of co-expressed genes. In yeast this approach is feasible at a genome-wide scale. In contrast to vertebrates, however, intergenic regions are very short in yeast. A simple search for exact string matches of experimentally verified binding sides thus leads to an large number of false positives in genome-wide surveys of vertebrates. To overcome this problem, comparative approaches, including our own program tracker, have been developed that are based on utilizing evolutionary sequence conservation. Well conserved regions are thought to be functionally relevant because they tend to evolve slower than adjacent non-functional sequences.

The analysis of patterns of these phylogenetic footprints often leads to hard combinatorial problems. Even sorting the phylogenetic footprints obtained by tracker along their order on the genome requires the solution of the NP-complete minimum feedback vertex set problem. Alternatively, we showed that the problem can also be rephrased as a combinatorial optimization problem.

The molecular evolution of phylogenetic footprints is difficult to study because their functional significance is hard to deduce from sequence information alone. Here we propose an approach for studying the rate of evolution of functional non-coding sequences at a macro-evolutionary scale.

Variation databases promise an approach towards the assessment of recent selection pressure on genomic sequence elements that cannot be obtained from phylogenetic footprinting. Evidence of recent selection suggests recent functional relevance of elements potentially important for understanding the organization of the human genome and resulting complex phenotypes and diseases. In a genome-wide study, we identified phylogenetic footprints in the vicinity of human genes. In agreement with the distribution of known regulatory sites, the density of these phylogenetic footprints was highest within two thousand base pairs upstream and downstream of genes. Stabilizing selection acting on these phylogenetic footprints was indicated by significantly reduced single nucleotide polymorphism (SNP) density.

We found a weak correlation between SNP densities of phylogenetic footprints and coding sequences which suggests that gene regulation and function often evolve independently. Decreasing diversity in human genes with increasing time of conservation suggests that most old genes have not ceased to be functionally important today.

A series of recent studies of the mammalian transcriptome have dramatically changed our perception of genome organization at least in higher eukaryotes. Experimental studies agree that a substantial fraction of the genome is transcribed and that non-protein-coding RNAs are the dominating component of the transcriptome. These RNAs fall into heterogenous classes of transcripts with very diverse functions, including tRNAs, snoRNAs, and microRNAs.

MicroRNAs have been identified as crucial regulators in both animals and plants. We report here on a comparative study of all known miRNA families in animals. We can show that significant waves of innovations map to the branch leading to the vertebrates and to placental mammals.

In summary we demonstrate that the methods of comparative genomics can be used to detect, classify and further characterize evolutionary conserved non-protein-coding DNA elements independent of their detailed function.