Evolutionary Analysis of the Protein
Domain Distribution in Eukaryotes

Arli Aditya Parikesit

Junior Professorship for Computational EvoDevo, Institute of Computer Science, University of Leipzig. Härtelstr. 16-18, 0417 Leipzig, Germany, arli@bioinf.uni-leipzig.de

Abstract:

Investigations into the origins and evolution of regulatory mechanisms require quantitative estimates of the abundance and co-occurrence of functional protein domains among distantly related genomes. The metabolic and regulatory capabilities of an organism are implicit in its protein content. Currently available methods suffer for strong ascertainment biases, requiring methods for unbiased approaches to protein domain contents at genome-wide scales. The discussion will be highlighted on large scale patterns of similarities and differences of domain contains between phylum-level or even higher level taxonomic groups. This provides insights into large-scale evolutionary trends.

Its complement of recognizable functional protein domains and their combinations convey essentially the same information and at the same time are much more readily accessible, although protein domain models trained for one phylogenetic group frequently fail on distantly related sequences. Transcription factors (TF) typically cooperate to activate or repress the expression of genes. They play a critical role in developmental processes. While Chromatin Regulation (CR) facilitates DNA organization and prevent DNA aggregation and tangling which is important for replication, segregation, and gene expression.

To compare the set of TFs and CRs between species, the genome annotation of equal quality was employed. To overcome this problem, performing gene prediction followed by the detection of functional domains via HMM-based annotation of SCOP domains were proposed. This methods was demonstrated to lead toward consistent estimates for quantitative comparison. To emphasize the applicability, the protein domain distribution of putative TFs and CRs by quantitative and boolean means were analyzed. In particular, systematic studies of protein domain occurrences and co-occurrences to study avoidance or preferential co-occurrence of certain protein domains within TFs and CRs were utilized.

Pooling related domain models based on their GO-annotation in combination with *de novo* gene prediction methods provides estimates that seem to be less affected by phylogenetic biases. It was shown for 18 diverse representatives from all eukaryotic kingdoms that a pooled analysis of the tendencies for co-occurrence or avoidance of protein domains is indeed feasible. This type of analysis can reveal general large-scale patterns in the domain co-occurrence and helps to identify lineage-specific variations in the evolution of protein domains. Somewhat surprisingly, Strong ubiquitous patterns governing the evolutionary behavior of specific functional classes were not found. Instead, there are strong variations between the major groups of eukaryotes, pointing at systematic differences in their evolutionary constraints. Species-specific training is required, however, to account for the genomic peculiarities in many lineages. In contrast to earlier studies wide-spread statistically significant avoidance of protein domains associated with distinct functional high-level gene-ontology terms were found.