Protein Domain Coocurrences Reveal Functional Changes of Regulatory Mechanisms During Evolution

A.A. Parikesit*, S.J. Prohaska, P.F. Stadler
Chair of Bioinformatics, University of Leipzig

Introduction
The emergence of higher organisms was facilitated by a dramatic increase in the complexity of gene regulatory mechanisms. This is achieved not only by addition of novel but also by expansion of existing mechanisms. Such an expansion is usually characterized by the proliferation of functionally paralogous proteins and the appearance of novel combinations of functional domains. Large scale phylogenetic analysis can shed light on the relative amounts of functional domains and their combinations and interactions involved in certain regulatory networks.

Methods
We performed comparative and functional analysis of three regulatory mechanisms: (1) transcriptional regulation by transcription factors, (2) post-transcriptional regulation by miRNAs, and (3) chromatin regulation across all domains of life. All of these mechanisms are evolutionarily old and passed through several major innovations. We calculated single domain distributions and domain cooccurrences from the SUPERFAMILY domain annotations [1] of about 900 genomes. Functional annotation from GeneOntology and protein domain descriptions were integrated into our comparative analysis.

Results and Discussion

Chromatin Regulation

Chromosomal architectural proteins and modification and demodification enzymes are present in all domains of life. However, demodification enzymes are less frequent. In contrast, reader domains are specific to eukaryots and co-emerge with the usage of histone modifications as signals [2].

Transcriptional Regulation

The number of transcription factors scales with the total number of proteins.

Correlation between transcription factors and chromatin related proteins

Results show massive problems with data quality: closely related species (e.g. dolphin and human) show dramatically different distributions of transcription factors and chromatin domains. This is not reasonable within mammals and contradicts biological knowledge.

SUPERFAMILY thus cannot be used for large-scale quantitative comparisons across species due to several sources of bias:
- different completeness of protein annotation for different genomes
- differences in transcript coverage
- different coverage of protein domains at kingdom level
- misannotations of functions (e.g. the chromodomain, a chromatin regulation domain is annotated as a transcription factor in SCOP)

A strategy for de novo domain annotation

We are currently testing how biases can be avoided in a de novo domain annotation. To this end we re-annotate a randomly selected subset of SCOP domains in three different sets of peptides: (1) those derived from the annotated ENSEMBL transcripts, (2) the output of the de novo gene predictor genscan, and (3) a conceptual translation of the entire genomic DNA in all six reading frames.

First data show that transcript sequences and genscan predictions correlate but show large systematic biases that, at least in part, can be explained by uneven coverage of the transcript annotation. A direct annotation on genomic DNA appears problematic since protein domains frequently overlap exon boundaries.

References

* Corresponding Author email: arli@bioinf.uni-leipzig.de