Protein Domain Coocurences 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 increases 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 methods are evolutionarily old and passed through several major innovations. We calculated single domain distributions and domain coocurrences from the SUPERFAMILY domain annotations [1] of about 900 genomes. Functional annotation from GeneOnthology and protein domain descriptions were integrated into our comparative analysis.

Results and Discussion

Chromatin Regulation

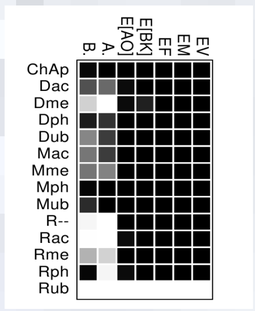

Figure 1 : Phylogenetic distribution of chromatin regulatiory components. Gray value: fraction of species in a clade that have at least one protein of the given type.

List of Abreviation

ChAp : Chromosomal Architectural Protein
Dac : De-acetylation
Dme : De-methylation
Dph : De-phosphorylation
Dub : De-ubiquitination
Mac : acetylation modifiyer
Mme : Methylation modifiyer
Mph : Phosphorylation modifiyer
Mub : ubiquitination modifiyer
R-- : Reader of an unmodified side chain
Rac : acetylation reader
Rme : Methylation reader
Rph : Phosphorylation reader
Rub : ubiquitination reader

A : Arachea
B : Bacteria
E[BK]: Basal eukaryots and kinetoplastids
E[AO]: Chromalveolata
EF : Fungi
EM : Metazoa
EV : Viridiplantae

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]. The entry for methylation reader domains in bacteria and archaea above can be identified as artifacts. Single BRCT domains can be found in bacteria, but only tandem BRCT domains have function as phosphorylation readers.

Transcriptional Regulation

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

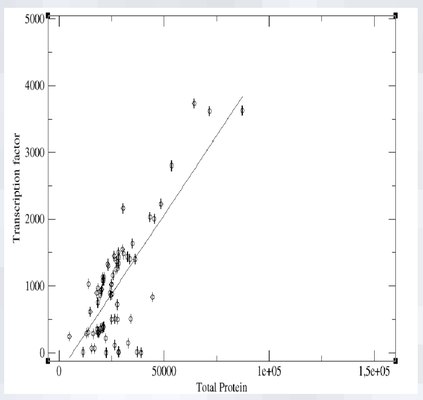

Figure 2 : Transcription Factor versus Total Domain Plot. Shown in the linear regression plot, that more total protein correlates with more transcription factor

Correlation between transcription factors and chromatin related proteins

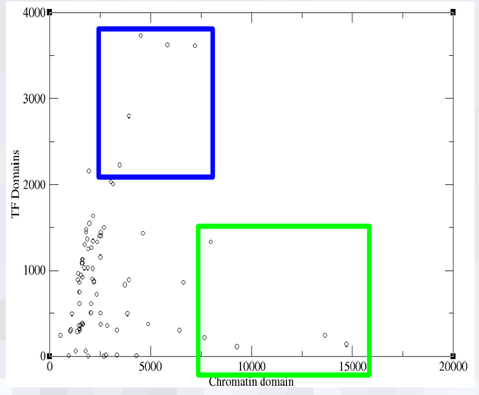

Figure 3 : Correlation of the number of transcription factors and chromatin domains.

Blue box. Species with few chromatin-related domains but many transcription factors: human, kangaroo rat, mouse, opossum, fugu.

Green box: species with many chromatin-related domains but few transcription factors: seq squirt, medaka, dolphin, yeast

There are no organisms (known) that have a lot of both.

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.

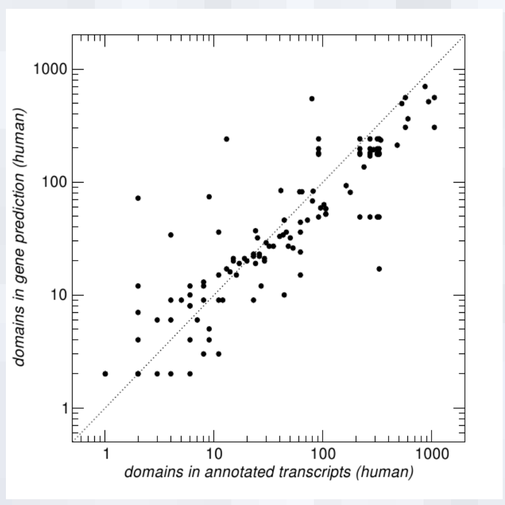

Figure 4 : Human transcript versus human genprediction plot

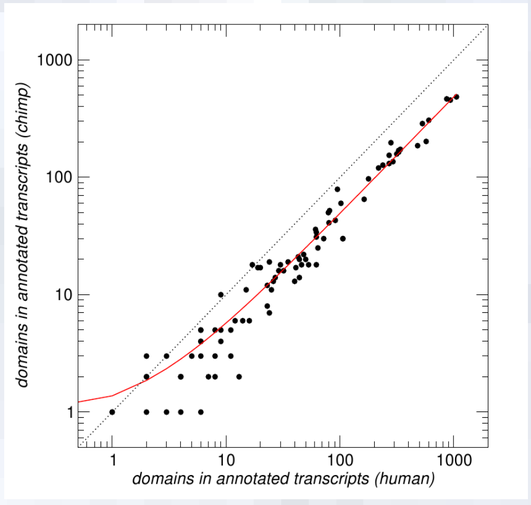

Figure 5 : Human transcript versus Chimp transcript plot

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

[1] Julian Gough and Cyrus Chothia.SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Research*, 2002, **30**, 268-272.
[2] Sonja J. Prohaska, Peter F. Stadler, David C. Krakauer. Innovation in gene regulation: The case of chromatin computation. *Journal of Theoretical Biology*, 2010.

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

Printed by the Universitaetsrechenzentrum Leipzig