Evolution of domain co-occurrences: some striking results

Arli A. Parikesit[1,3], Peter F. Stadler[1,2], Sonja J. Prohaska[1,3]

[1]Interdisciplinary Center for Bioinformatics, University of Leipzig
[2]Bioinformatics Group, Inst. of Computer Science, University of Leipzig
[3]Computational EvoDevo, Inst.of Computer Science, University of Leipzig

Investigation of the origin and evolution of regulatory mechanisms requires comparable estimates for the abundance and co-occurrence of functional protein domains of distantly related genomes. Currently available methods suffer for strong ascertainment biases, requiring methods for unbiased approaches to protein domain contents at genome-wide scales [1,4]. We will discuss domain distribution patterns between taxonomic groups providing insights into large-scale evolutionary trends.

Since version 1.75, the SUPERFAMILY database provides functional information for protein domains using the gene ontology annotation terms [2,3]. To use the SUPERFAMILY database for comparison of GO annotation terms between species, one relies on existing, heavily biased gene annotation. To overcome this problem, we propose to perform gene prediction followed by the detection of protein domains via HMMs for SUPERFAMILY domains and subsequent analysis of the abundance and co-occurence of functionally related groups of domains. In this contribution, we will demonstrate that this methods leads to consistent estimates for quantitative comparison. In particular, we systematically study avoidance and preferential co-occurrence of domains associated with certain GO terms [3].

We analyze domain distributions from eight eukaryotic taxa: basal eukaryots, Kinetoplastida, Alveolata, Chromista, Viridiplantae, Amoebozoa, Fungi and Metazoa. We observed that C2H2 zinc finger domains significantly co-occure with nucleic acid binding domains in almost all taxa, even though individual DNA binding domains avoid to co-occure with C2H2 zinc fingers, e.g. with P loop containing nucleoside triphosphate and ribonuclease H-like domains (see Figure 1) [4].

Using a classification of domains involved in chromatin regulation we observe significant co-occurrence with zinc finger domains only for Chromista, Viridiplantae, Amoebozoa, Fungi, and Metazoa. Suggesting that zinc finger proteins were recruited into the role as chromatin regulators. In particular, domains capable of writing histone modifications significantly co-occure with zinc finger domains in 11 of 18 examined species.

Furthermore, protein binding domains can be roughly devided into two groups with respect to their intracellular location, namely the nuclear and cytoplasmatic group. Most of the co-occurences result are not significant. It is noteworthy, that Kinetoplastids, Chromista and Viridiplantae show significant co-occurence of zinc finger domains with potentially nuclear located protein binding domains. However, the metazoan genomes examined show a tendency for avoidance of nuclear located protein binding domains.

As a negative control, we calculated co-occurrence of zinc finger domains with domains with the functional annotation "catalytic activity" and "cellular polysaccharide metabolic process". As we expect, we observe that neither co-occurence nor avoidance is significant in most cases. Only in some cases, we see significant avoidance.

An interesting phylogenetic distribution of avoidance and co-occurence patterns is observed for the GO term "regulation of binding": while we observed significant co-occurence in Chromista and most of the Metazoa, we discovered significant avoidance in *Trypanosoma brucei* reflecting the trend of other Kinetoplastida.

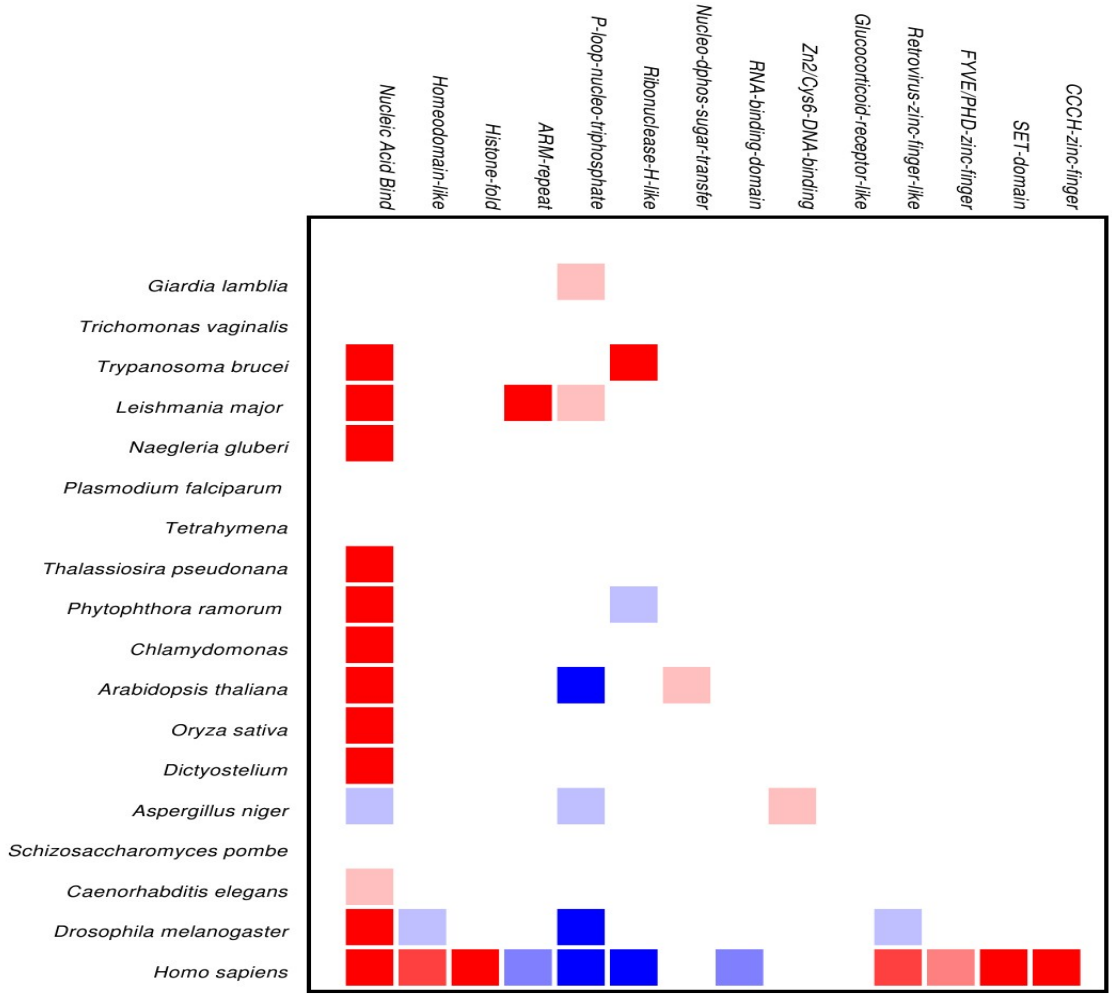

Figure 1: Co-occurrence of C2H2 zinc finger domains with nucleic acid binding domains (GO:0003676) and individual superfamily domains thereof with significant values for co-occurence (red) or avoidance (blue). The color intensity indicates significance at 1%, 5%, and 10% level for dark, midtone, and light, respectively. The absence of colored squares indicates that co-occurrence is either not significant or less then 2 co-occurences were observed.

References

[1] Prohaska S J, Stadler P F, and Krakauer D C. Innovation in Gene Regulation: The Case of Chromatin Computation. J. Theor. Biol., 265:27-44, 2010.

[2] Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C and Gough J. SUPERFAMILY Comparative Genomics, Datamining and Sophisticated Visualisation. Nucleic. Acids. Res. 37:D380-D386. 2009.

[3] De Lima Morais D A, Fang H, Rackham O J L, Wilson D, Pethica R, Chothia C, and Gough J. SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic. Acids. Res. 39: D427-D434. 2011

[4] Parikesit A A, Stadler P F, and Prohaska S J. 25th German Conference on Bioinformatics 2010. Quantitative Comparison of Genomic-Wide Protein Domain Distributions. GCB2010 conference proceeding. Vol P-173: pp 93-102. 2010