Quantitative Measurement of Genome-wide Protein Domain Co-occurrence of Transcription Factors

Arli A. Parikesit, Sonja J. Prohaska, Peter F. Stadler

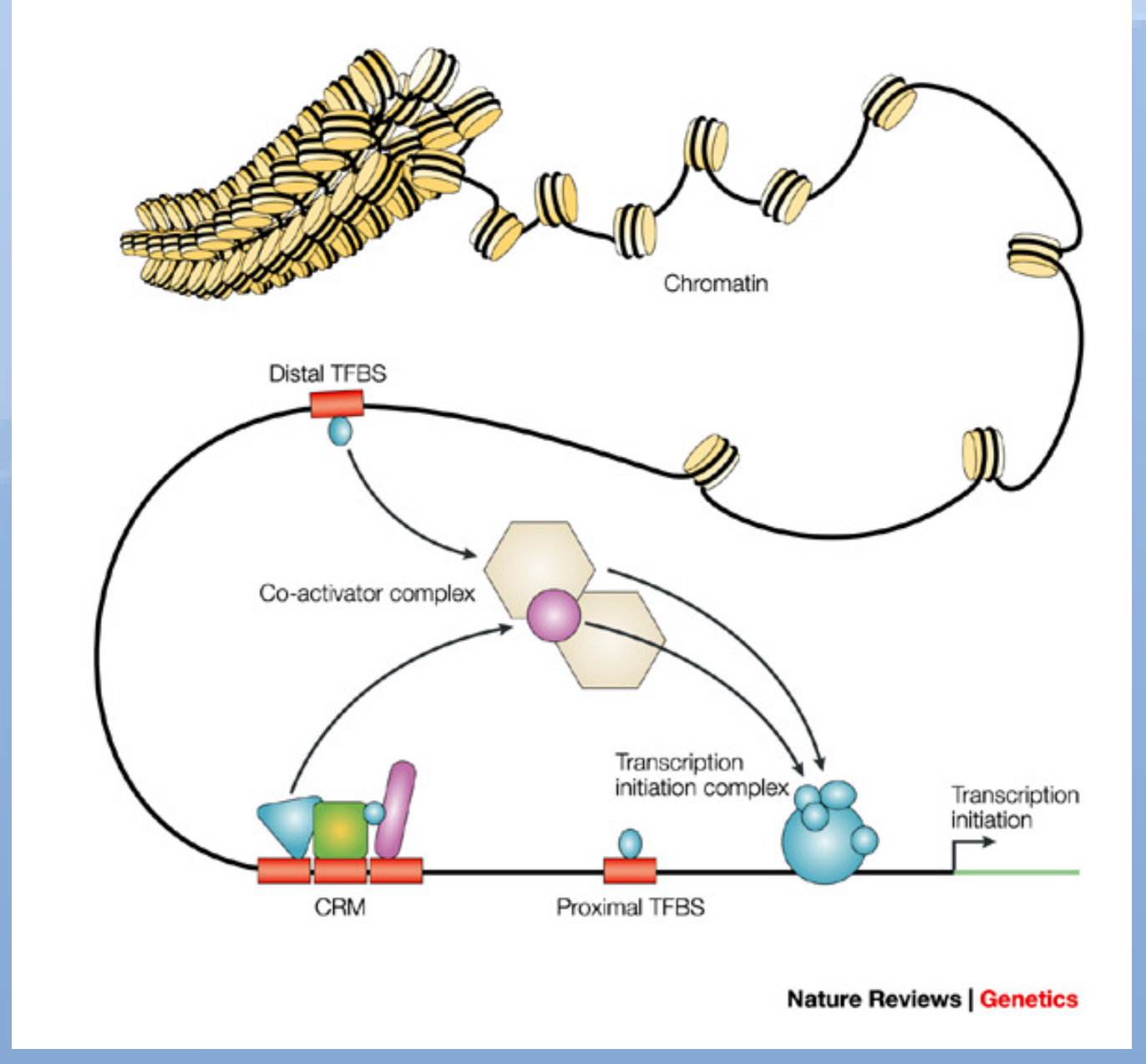
Chair of Bioinformatics, Department of Computer Science, University of Leipzig
Härtelstr. 16-18, D-04017 Leipzig
arli@bioinf.uni-leipzig.de

Abstract

WE present a methodology for quantitative comparison of protein domain co-occurrences among Eukaryotes. The research focuses on over- and under-representation of domain co-occurrences in transcription factors. [1]

1. Background

TRANSCRIPTION factors (TF) typically cooperate to activate or repress the expression of target genes. They play critical roles in essentially every developmental process, from the proliferation and differentiation of stem cells to the maintenance of differentiated cells in adult organisms. In our contribution, we analyzed the protein domain distribution in TFs. The combination of *de novo* gene prediction and subsequent HMM-based annotation of SCOP domains in the predicted peptides leads to consistent and comparable estimates of co-occurrences with acceptable accuracy. In particular, it can be utilized for systematic studies of the evolution of protein domain occurrences and co-occurrences, recently published in [PSP10].



Nature Reviews | Genetics

Figure 1: *Transcription factors (TFs) bind to specific sites (transcription-factor binding sites; TFBS) that are either proximal or distal to a transcription start site. Sets of TFs can operate in functional cis-regulatory modules (CRMs) to achieve specific regulatory properties.*

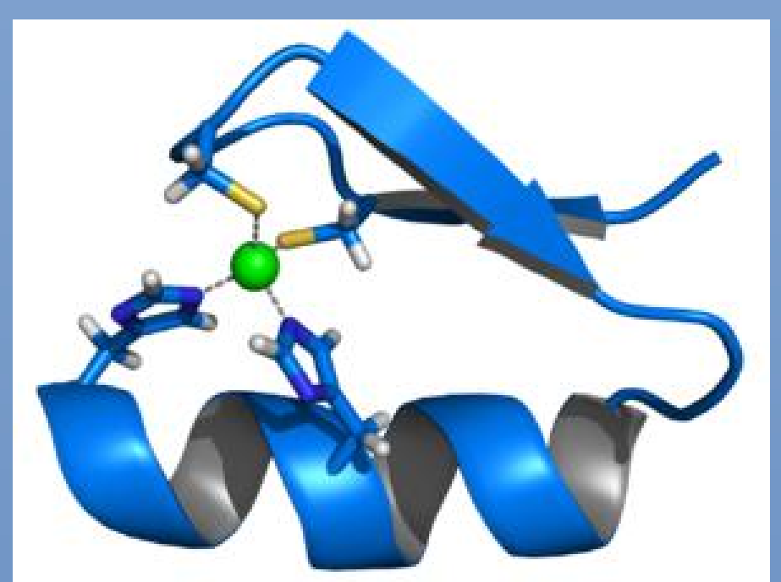


Figure 2: *Zinc fingers are a large superfamily of protein domains that can bind to DNA (and occasionally single- or double-stranded RNA and proteins).*

2. Method

AS an application, we have considered seven major classes of DNA-binding domains of TFs: zinc-finger (znf), leucine-zipper, winged-helix, bromo, brct, krab and hmg-box (hmg). Znf, leucine-zipper, winged-helix, and hmg are DNA-binding domain. We have found that different types of DNA-binding domains systematically avoid each other throughout the evolution of Eukaryote. In contrast, DNA binding domains belonging to the same superfamily readily co-occur in the same protein. We also determined the domain co-occurrence of znf with other non-DNA-binding domains, namely wd40, phd, ring, and tpr. In these cases, we also expected high numbers of co-occurrences but observed significantly fewer than expected. This also indicates avoidance. We will present systematic analysis of co-occurrences and potential reasons for avoidance. Based on our published methodology, we investigate more domain co-occurrences for significant and biologically meaningful avoidance. The expectation values for each pairwise co-occurrence was calculated with the following formula:

$$E(x,y) = \frac{X \times Y}{n}$$

where:
X is the number of genes with domain x
Y is the number of genes with domain y
n is the total number of genes.
This can be computed for Genscan predictions and Superfamily annotation. The Expectation value is then compared with the number of genes F in which x and y co-occur. If $E > F$ then we observe avoidance of domains, on the other hand, if $F > E$ then co-occurrence is preferred. A Poisson distribution with mean E is used to determine whether the observed counts F significantly deviate from the expectation.

WE compare domain co-occurrences computed from the *de novo* predictions (GP) with domain co-occurrences recorded in the SUPERFAMILY database [WPZ+09] (SF) for the following 18 species:

Legend:

1=	*Giardia intestinalis*
2=	*Trichomonas vaginalis*
3=	*Trypanosoma brucei*
4=	*Leishmania major*
5=	*Naegleria gluberi*
6=	*Plasmodium falciparum*
7=	*Tetrahymena*
8=	*Thalassiosira pseudonana*
9=	*Phytophthora ramorum*
10=	*Chlamydomonas*
11=	*Arabidopsis thaliana*
12=	*Oryza sativa*
13=	*Dictyostelium*
14=	*Aspergillus niger*
15=	*Schizosaccharomyces pombe*
16=	*Caenorhabditis elegans*
17=	*Drosophila melanogaster*
18=	*Homo sapiens*

3. Domain Avoidance

IN many case we observe systematically fewer domain co-occurrences than expected, i.e., there is a selection pressure causing the domains to "avoid" each other. In fact, this is the case with most — but not all — combinations of distinct DNA binding domains. In *Oryza sativa* $E(GP) \ll E(SF)$, because SF has more annotated individual domain than GP.

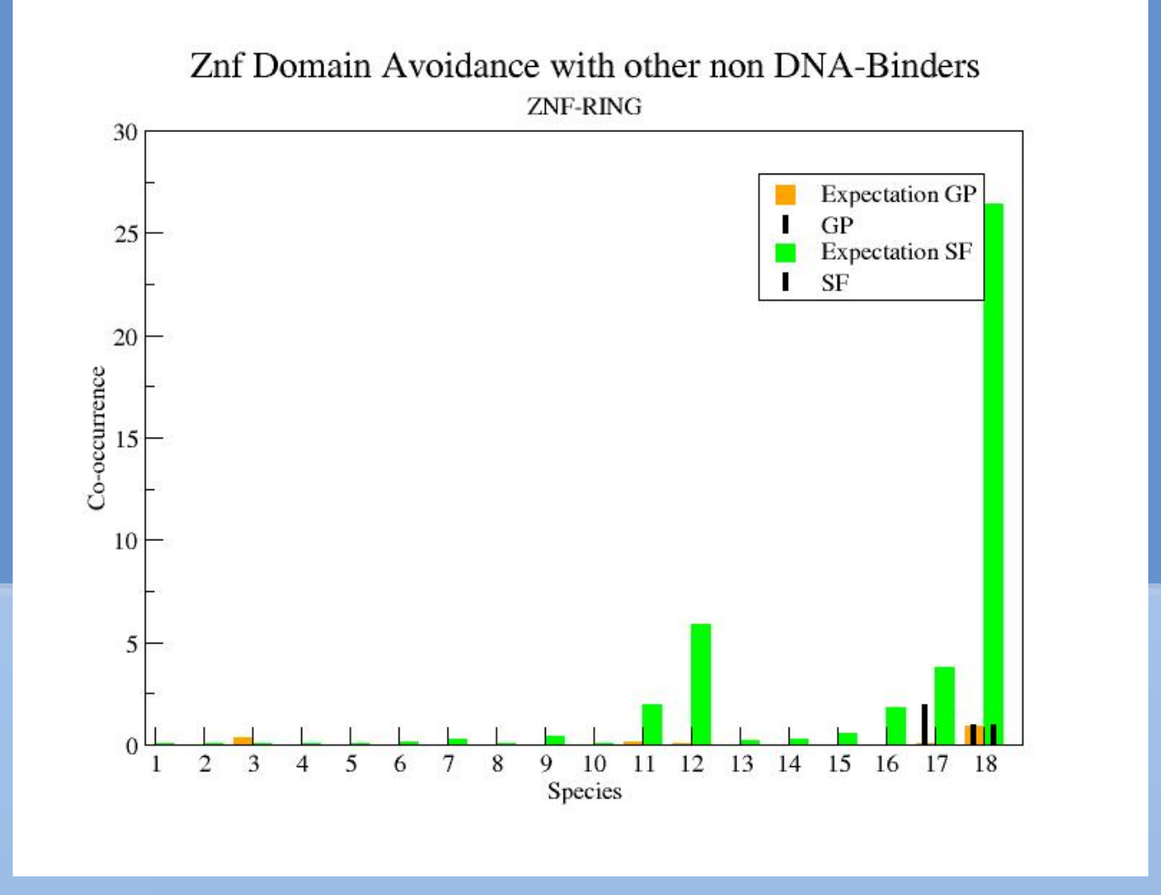


Figure 3: *The znf-ring pairs showed a strong avoidance tendency.It is shown in the Homo Sapiens (SF), which $E \gg F$ and $P \ll 0.05$.*

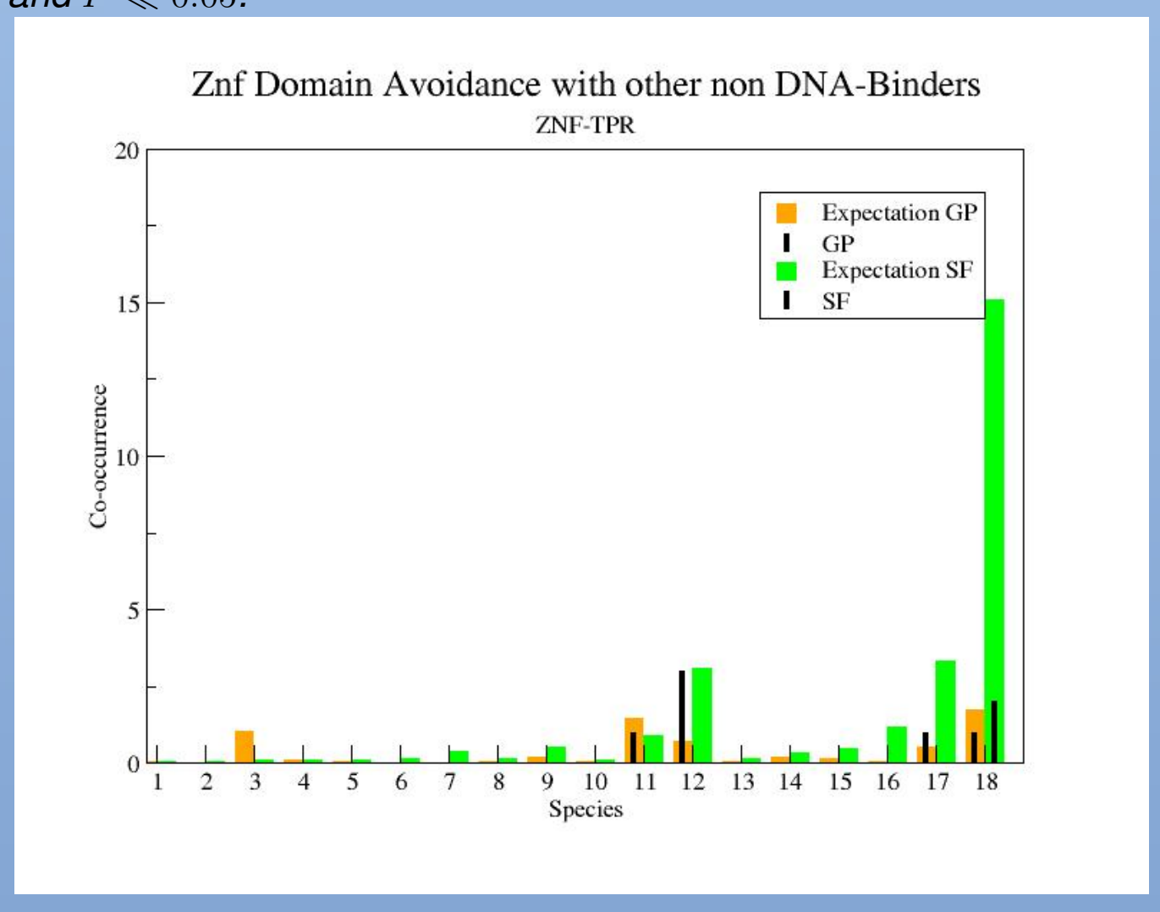


Figure 4: *The znf-tpr pairs showed a strong avoidance tendency.It is shown in the Homo Sapiens (SF) domain co-occurrence , which $E \gg F$ and $P \ll 0.05$. The efficacy of Genscan prediction will be verified in Homo Sapiens (SF) because it has SF entries and fewer domain hmm co-occurrences*

4. Domain Co-occurrences

IN some cases, however, a positive correlation between distinct DNA binding domains is observed. A well-studied example is the co-occurrence of KRAB domain and ZNF domains in a large group of primate-specific transcription factors [NHZS10].

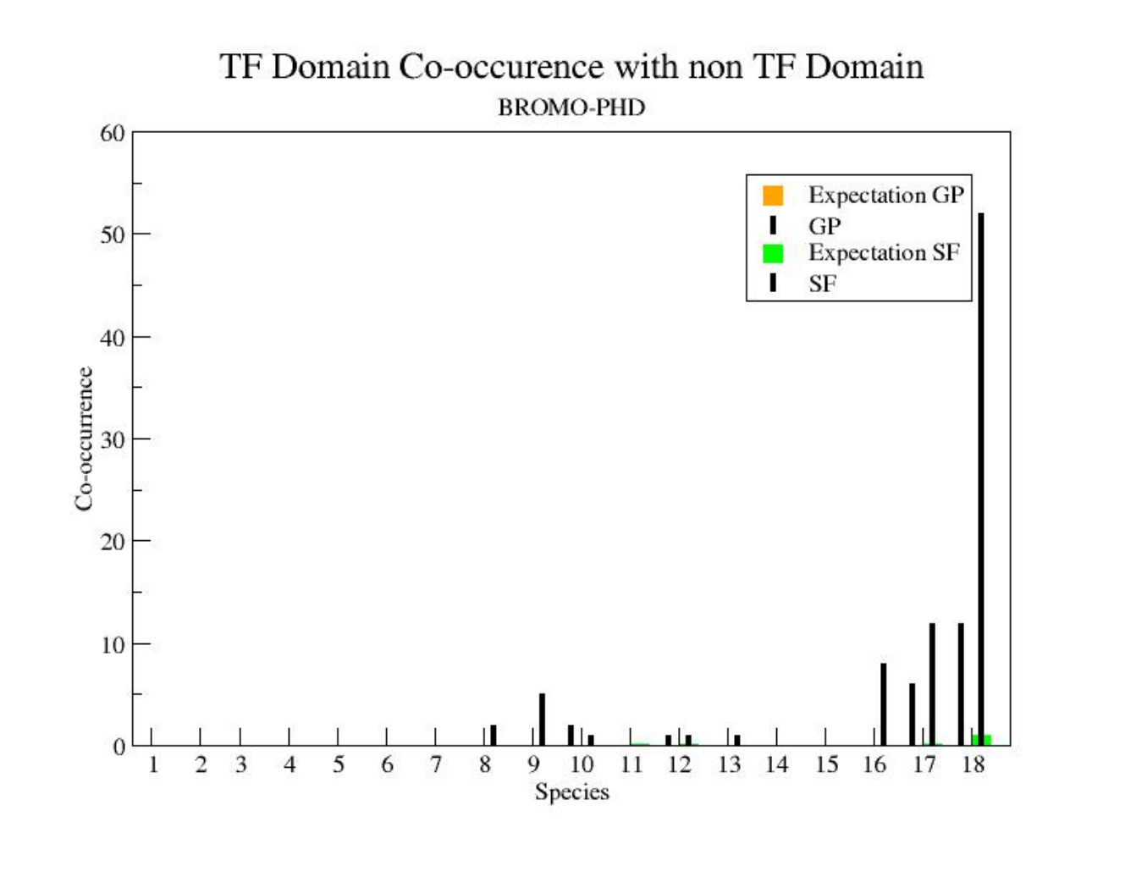


Figure 5: *There is a co-occurrence tendency in bromo-phd pair. It shows primarily in Homo sapiens (SF and GP), Drosophila melanogaster (SF and GP), Oryza sativa (GP), Caenorhabditis elegans (SF), Dictyostelium (SF), Phytophthora ramorum (SF), Thalassiosira pseudonana (SF), and Chlamydomonas (GP) which $E \ll F$ and $P \ll 0.05$. The search for Hypothetical Protein existence in Oryza sativa (GP), and Chlamydomonas (GP) are on the way, because it has no or few SF co-occurrence, and abundant domain hmm co-occurrences. The efficacy of Genscan prediction will be verified in Caenorhabditis elegans (SF), Dictyostelium (SF), Phytophthora ramorum (SF), and Thalassiosira pseudonana (SF) because they have SF entries and no or few domain hmm co-occurrences*

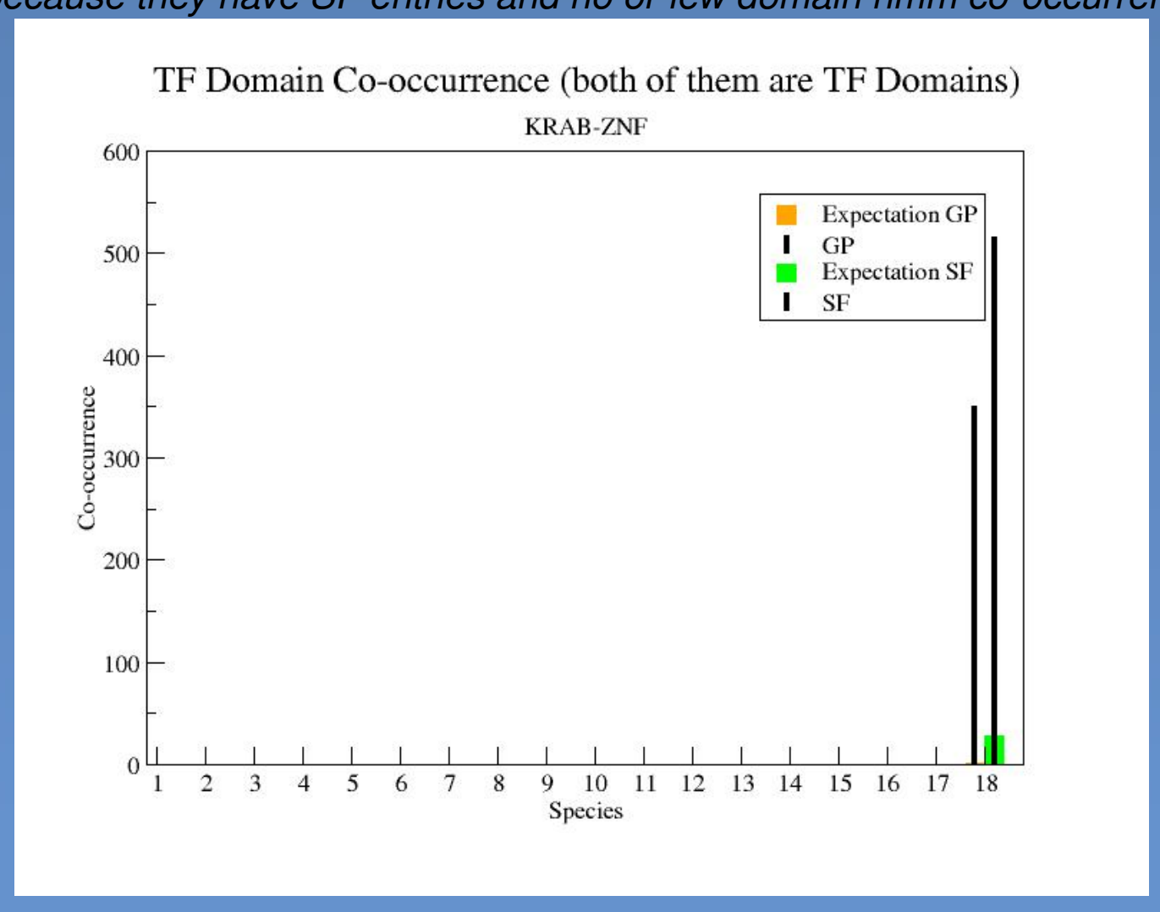


Figure 6: *There is a co-occurrence tendency in krab-znf pair only in Homo Sapiens (SF and GP), which $E \ll F$ and $P \ll 0.05$. Krab-znf co-occurrence are happening primarily in Homo sapiens [NHZS10]*

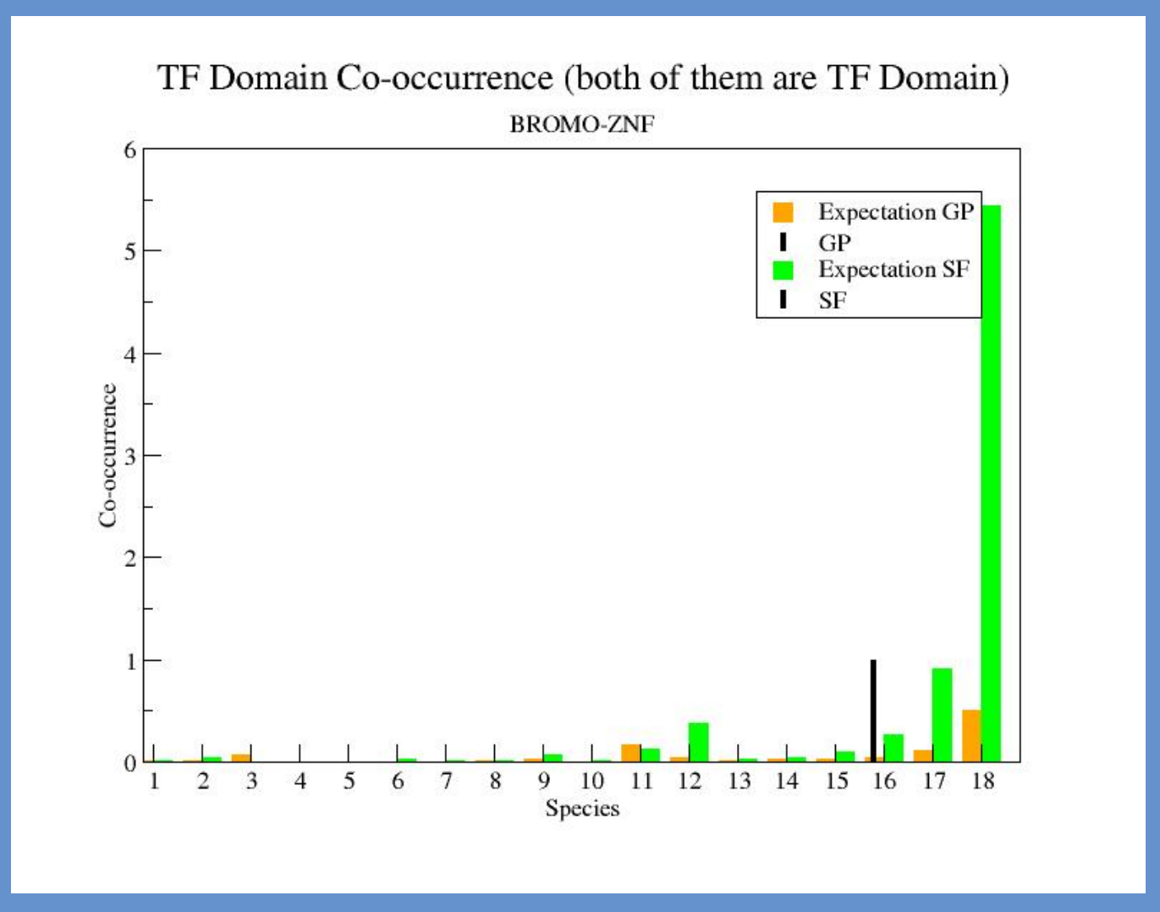


Figure 7: *The bromo-znf pairs showed a co-occurrences tendency. It is shown in the Caenorhabditis elegans (GP) domain co-occurrence , which $E \ll F$ and $P \ll 0.05$. The search for Hypothetical Protein existence in Caenorhabditis elegans are on the way, because it has no SF co-occurrence, and abundant domain hmm co-occurrences.*

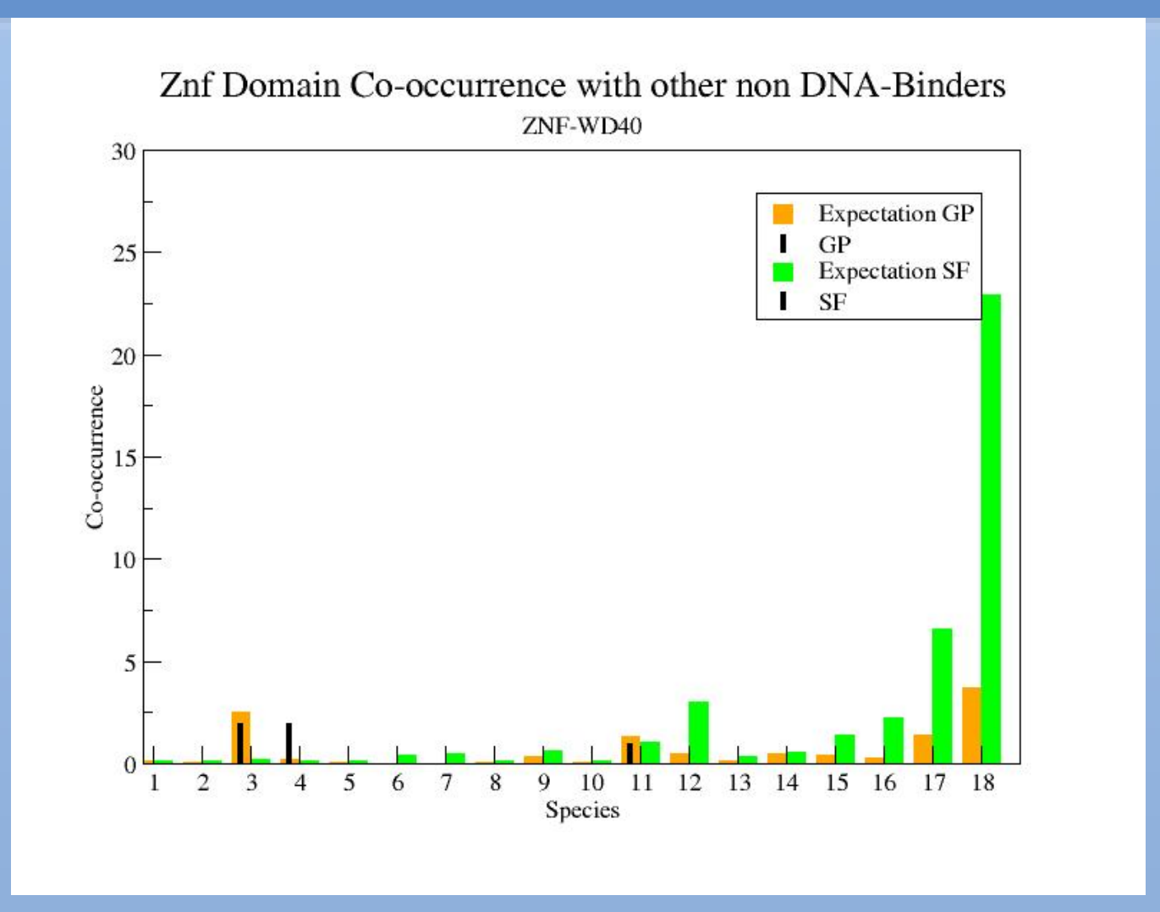


Figure 8: *The znf-wd40 pairs showed a co-occurrence tendency.It is shown in the Leishmania major (GP) domain co-occurrence , which $E \ll F$ and $P \ll 0.05$. The search for Hypothetical Protein existence in Leishmania major (GP) are on the way, because it has no SF co-occurrence, and abundant domain hmm co-occurrences.*

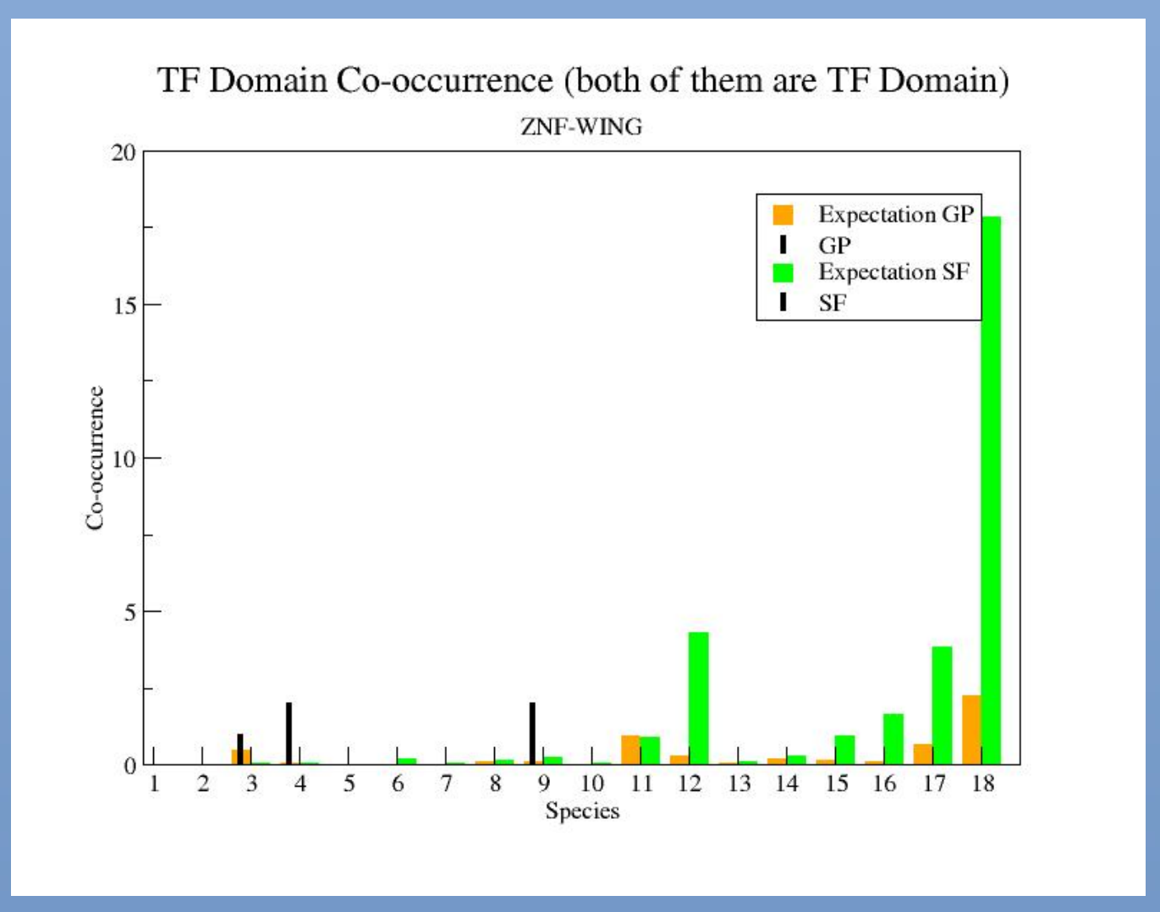


Figure 9: *The znf-wing pairs showed a co-occurrence tendency. It is shown in Phytophthora ramorum (GP) and Leishmania major (GP) domain co-occurrence , which $E \ll F$ and $P \ll 0.05$. Information about gene fragments existence are already published [PSP10]*

5. Discussion

ALTHOUGH a plethora of annotation data are available in publicly accessible databases for most of the published genomes, quantitative comparisons remain difficult due to dramatic differences in annotation methodology and data coverage. Consequently, comparative studies typically resort to testing for relative enrichment rather than considering absolute numbers of domains. In studies focusing on the evolution of regulatory mechanisms and regulatory complexity, however, absolute gene counts play an important role. Our previous investigations suggested that the biases and artifacts cause by *de novo* gene prediction methods such as `genscan` are small compared to the numerous problems of annotation-based approaches. In particular, we observe very a small number of false positive co-occurrences arising from the incorporation of additional introns and the erroneous prediction of fusion proteins. The combination of *de novo* gene predictors and subsequent HMM-based annotation of SCOP domains in the predicted peptides leads to consistent estimates with acceptable accuracy that in particular can be utilized for systematic studies of the evolution of protein domain occurrences and co-occurrences [PSP10].

PROTEIN DOMAINS are not randomly combined in functional proteins. We observe statistically significant avoidance if the TF domain paired with other non DNA-Binders (znf-ring, and znf-tpr). On the other hand, we find more co-occurrences than expected for certain combinations of TF and non-TF domains (e.g. bromo-phd), between distinct types of TF domains (e.g. in the combinations bromo-znf and znf-wing) and well as for combinations of DNA binding domains (e.g. krab-znf). The general trends are in most cases detected consistently based on *de novo* genome predictions (GP) and from annotation databases (SF).

AVOIDANCE and preferential co-occurrence, however, are only observable in genomes with sufficiently large numbers of proteins, in particular multicellular plants and animals. In most species with small genomes the expected numbers of domain co-occurrences is already below 1 so that a selection pressure for domain avoidance cannot be detected.

6. Conclusion

Several combinations of protein domains show specific tendencies to either systematically avoid each other or to co-occur preferentially in proteins. In the examples studied so far, avoidance appears to be conserved among those major Eukaryotic clades where the effect is detectable. Signals for preferential co-occurrence can arise from recent proliferation by gene duplication as in the case of the primate-specific krab-znf family of transcription factors.

References

[NHZS10] Katja Nowick, Aaron T. Hamilton, Huimin Zhang, and Lisa Stubbs. Rapid sequence and expression divergence suggests selection for novelfunction in primate-specific KRAB-ZNF genes. *Mol. Biol. Evol.*, 27(11):2606–2617, 2010.

[PSP10] Arli A. Parikesit, Peter F. Stadler, and Sonja J. Prohaska. Quantitative comparison of genomic-wide protein domain distributions. *Ger. Conf. Bioinform.*, P-173:93–102, 2010.

[WPZ+09] D Wilson, R Pethica, Y Zhou, C Talbot, C Vogel, M Madera, C Chothia, and J. Gough. SUPERFAMILY — comparative genomics, datamining and sophisticated visualisation. *Nucleic Acids Res.*, 37:D380–D386, 2009.

[1] Leipzig Research Festival for Live Science, University of Leipzig, December 17, 2010

Gedruckt im Universitätsrechenzentrum Leipzig