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

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

Bioinformatics Group
Institute of Computer Science
University of Leipzig

February 11, 2011

Why is it interesting to study domain distribution of Transcription Factor?

Background

- Transcription factors (TF) typically cooperate to activate or repress the expression of target genes. They play critical roles in essentially every developmental process.
- 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, it has recently published.

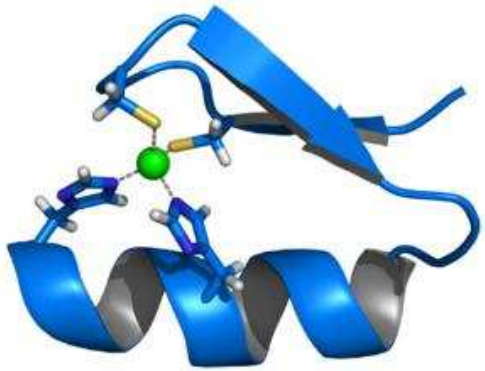
Scheme of Transcription Factor Protein

Function

Protein localizes to (sequence-specific) sites on the DNA	The protein is a transcription factor, which brings the transcription complex and (sequence-specific) sites on the DNA together. Protein is responsible for selective gene transcription.	The protein is part of the transcription complex

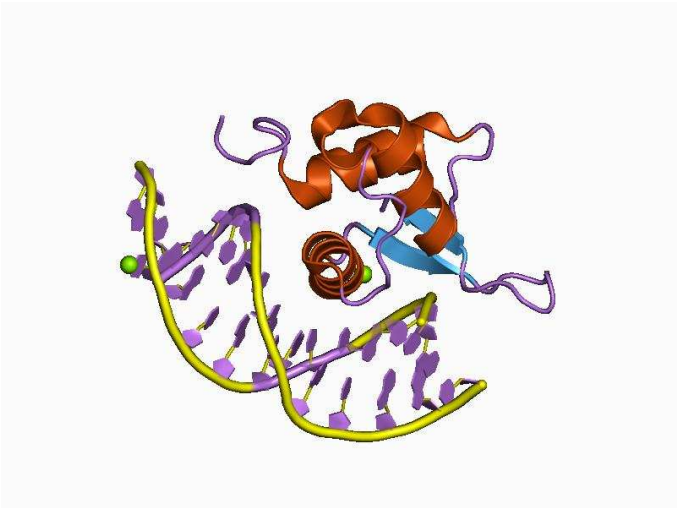
Protein domains and their combinations contain information about the functions in a cell.

Zinc Finger Domain



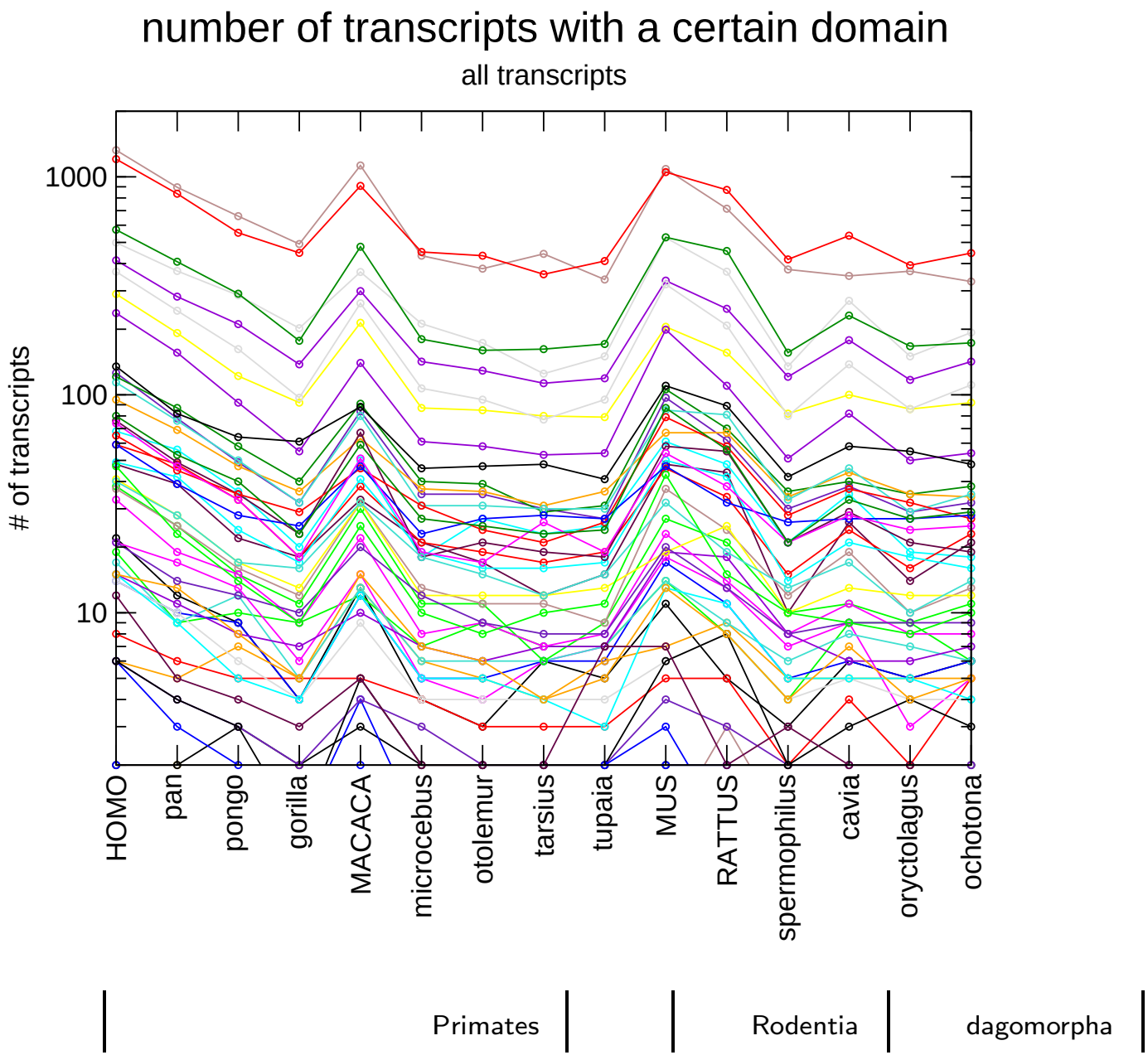
A zinc finger is a large superfamily of protein domains that can bind to DNA.
A zinc finger consists of two antiparallel β strands, and an α helix.
The zinc ion is crucial for the stability of this domain type.
In the absence of the metal ion the domain unfolds as it is too small to have a hydrophobic core.
Zinc finger is a part of Transcription Factor Regulation Domains.

Winged-helix Domain

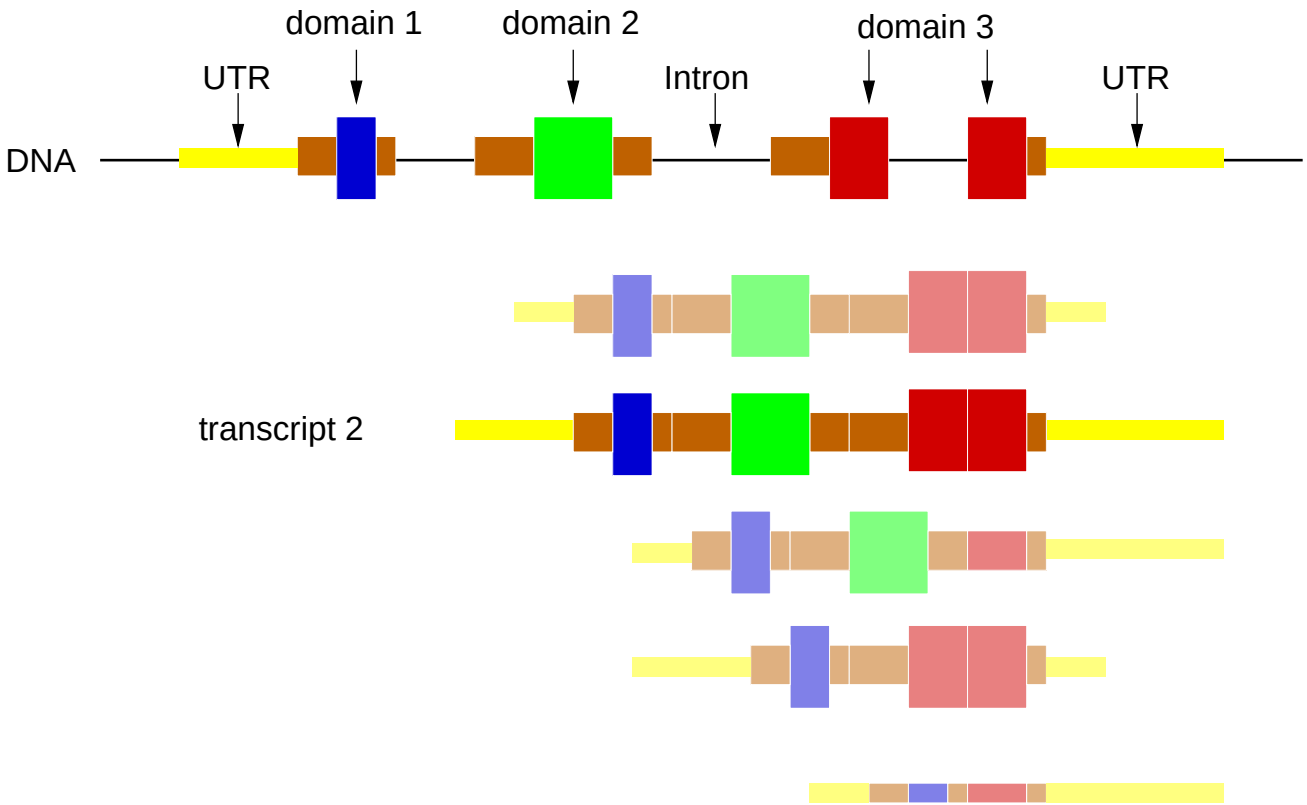


Winged-helix is a DNA-binding domain which binds to specific DNA sequences.
Consisting of about 110 amino acids, the domain in winged-helix transcription factors has four helices and a two-strand beta-sheet.
Wing-helix is a part of Transcription Factor Regulation Domains.

Genome Annotation is Biased Towards Model Organisms

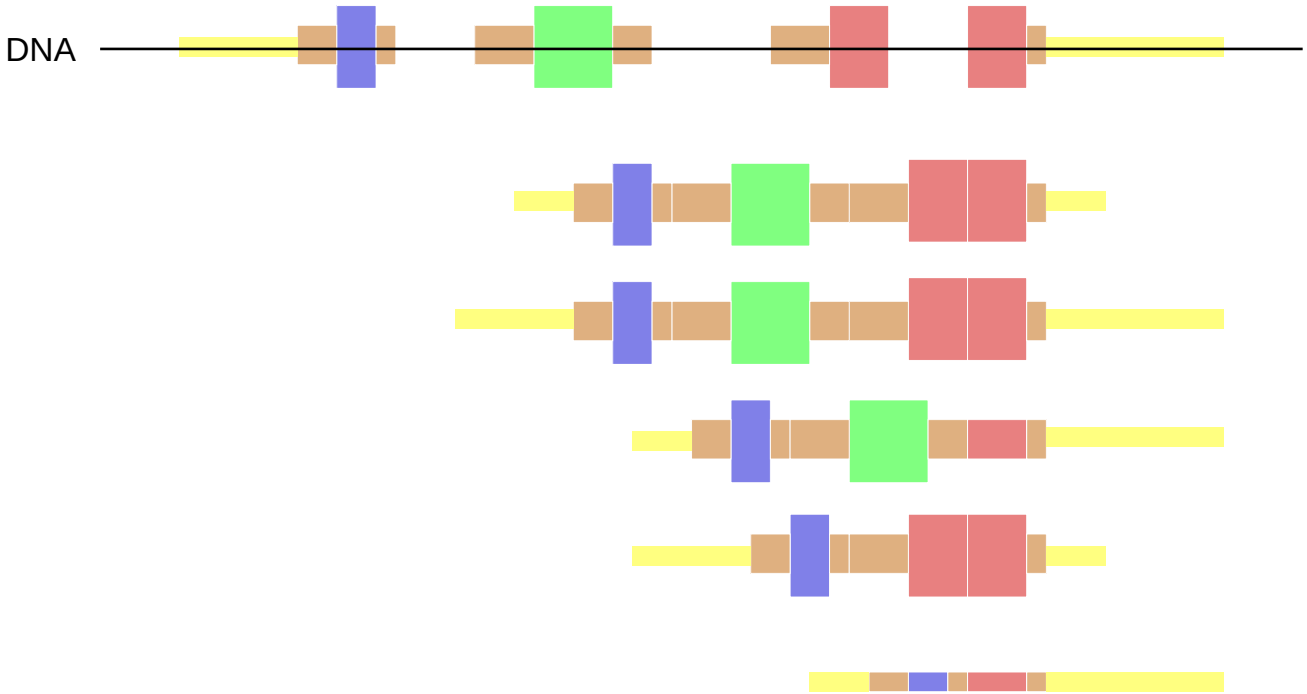


Counting Genes and Domain (Co-)Occurences



Result: Large discrepancies between the number of transcripts for orthologous loci.
This is a problem for interspecies comparisons.
We develop a new approach to overcome this difficulties.

Counting Genes and Domain (Co-)Occurences



Little gene annotation effort goes to genomes with close reference genomes.
Result: Large amounts of false negative gene annotations.

Methods on Quantitative Estimates of Protein Domain (Co-)Occurences

Required Data

1. 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)
2. We also determined the domain co-occurrence of znf with other non-DNA-binding domains, namely wd40, phd, ring, and tpr
3. We will present systematic analysis of co-occurrences and potential reasons for avoidance, by comparing the Genscan Prediction (GP) and Superfamily (SF) annotation.
4. Based on our published methodology, we investigate more domain co-occurrences for significant and biologically meaningful avoidance

Methods on Quantitative Estimates of Protein Domain (Co-)Occurences

Expectation Value Formula

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 (GP) and Superfamily (SF) 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.

Methods on Quantitative Estimates of Protein Domain (Co-)Occurences

Legend of the 18 species

We compare domain co-occurrences computed from the *de novo* predictions (GP) with domain co-occurrences recorded in the SUPERFAMILY database (SF) for the following 18 species:

Legend:

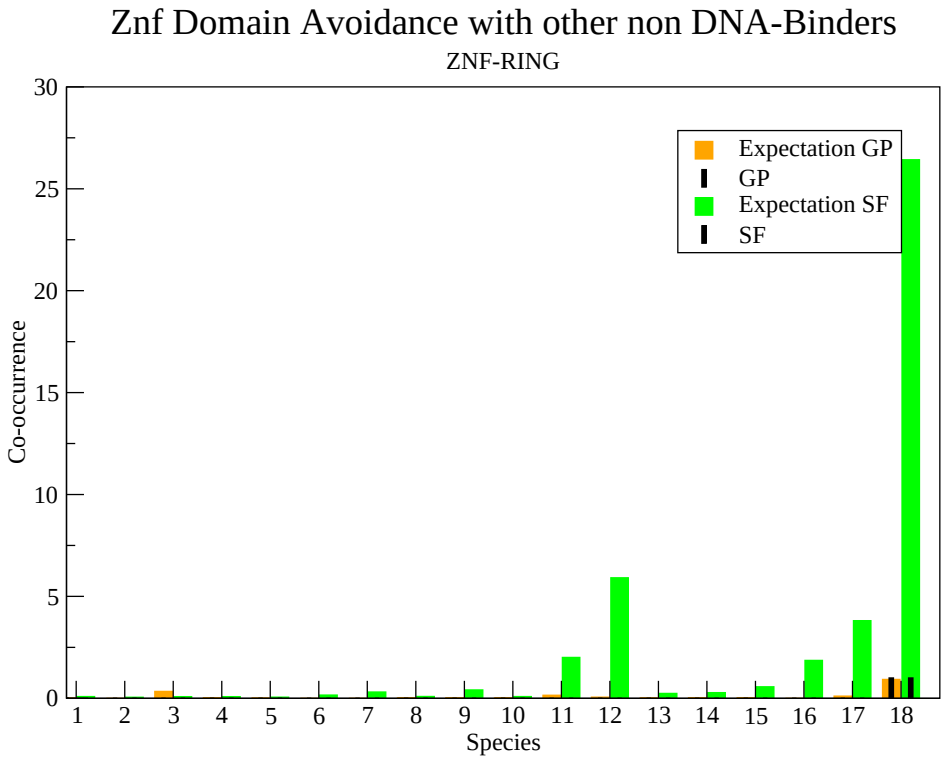
1= *Giardia intestinalis*
2= *Trichomonas vaginalis*
3= *Trypanosoma brucei*
4= *Leishmania major*
5= *Naegleria gluberi*
6= *Plasmodium falciparum*
7= *Tetrahymena*
8= *Thalassiosira pseudonana*
9= *Phytophthora ramorum*
10= *Chlamydomonas*
11= *Arabidopsis thaliana*
12= *Oryza sativa*
13= *Dictyostelium*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

Result(Domain Avoidance)

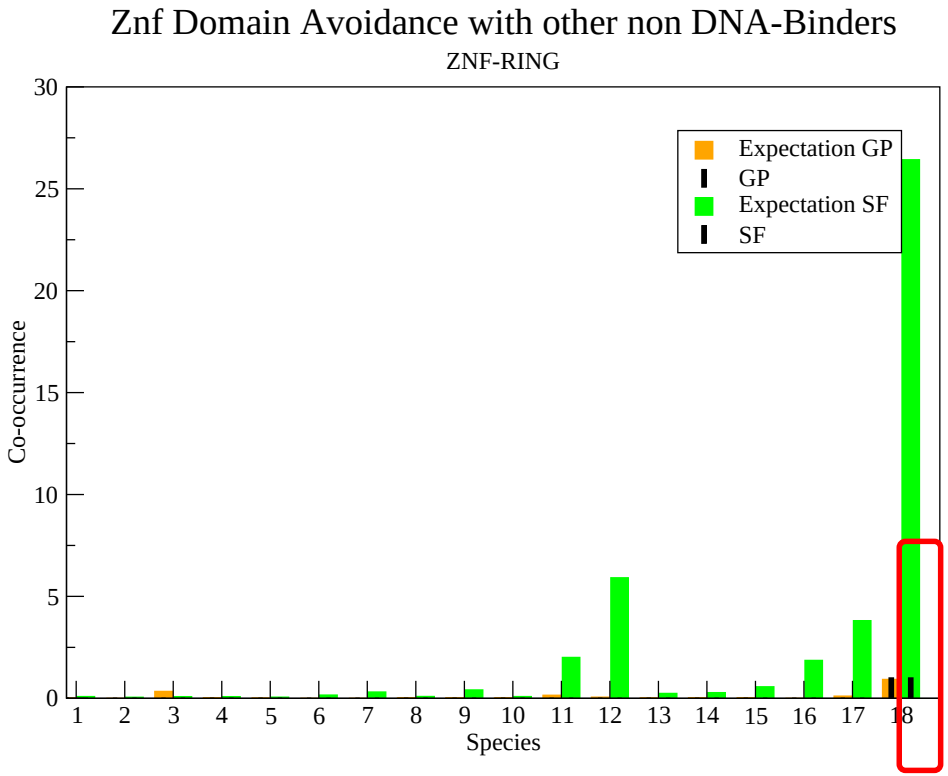
Insight

In many case we observe systematically fewer domain avoidance than expected, i.e., there is a selection pressure causing the domains to "co-occur" 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.

Result(Domain Avoidance)

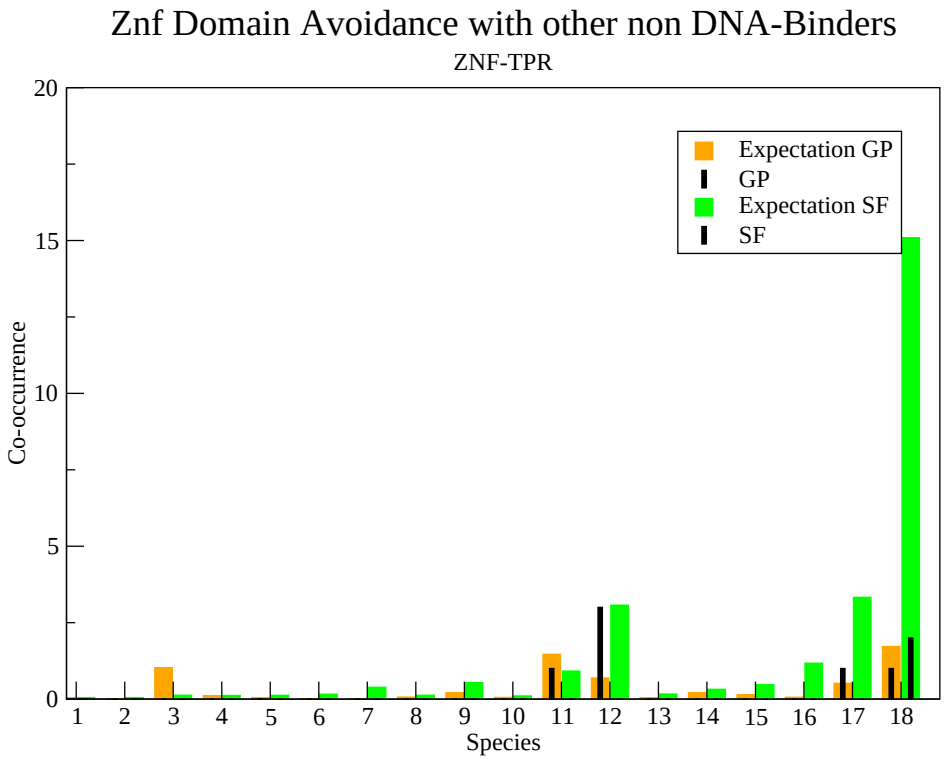


Result(Domain Avoidance)

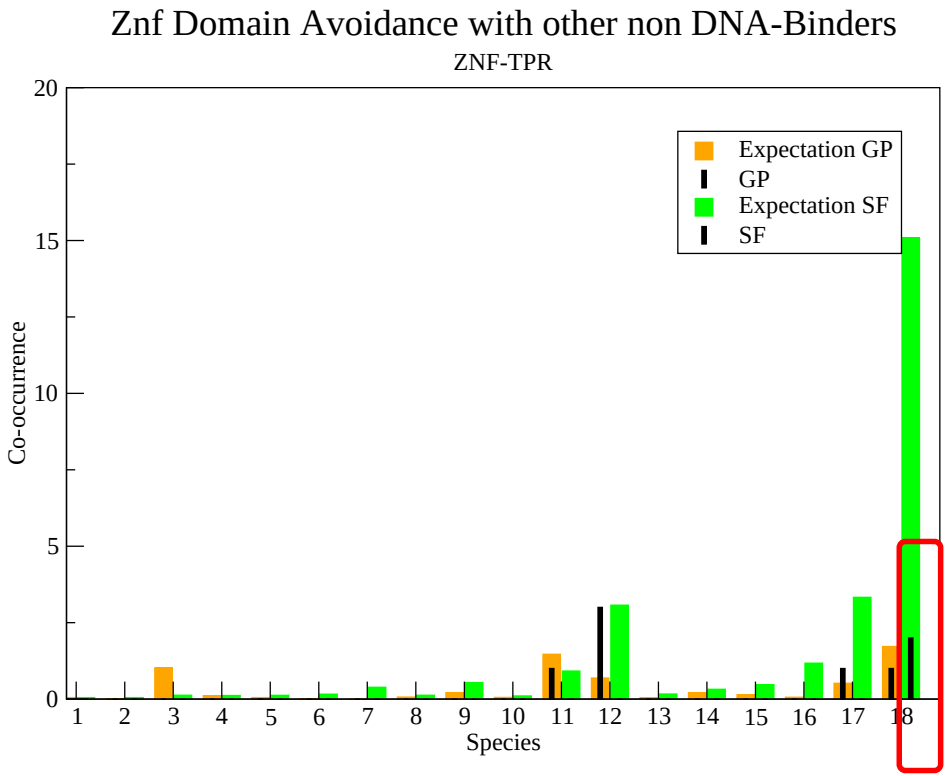


The znf-ring pairs showed an avoidance tendency.It is shown in the *Homo Sapiens* (SF), which $E \gg F$ and $P \ll 0.05$. In *Homo sapiens* and *Drosophila melanogaster*, $E(GP) \ll E(SF)$, because The GP Domain occurrences are not abundant enough. The ratio of domain occurrences/total genes in $GP \ll SF$

Result(Domain Avoidance)



Result(Domain Avoidance)



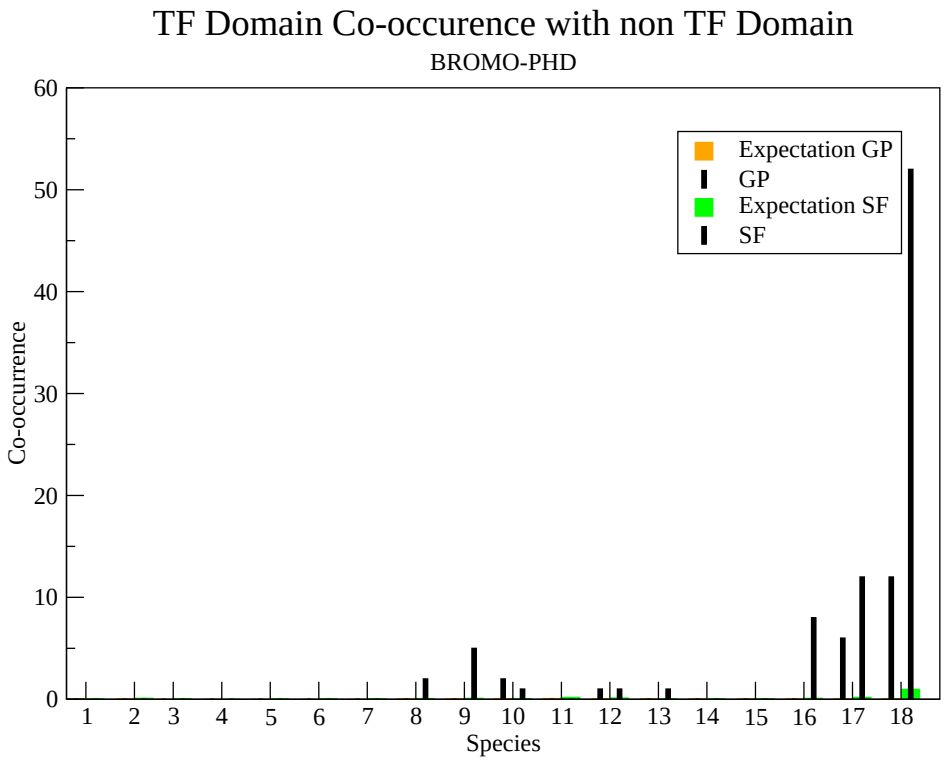
The znf-tpr pairs showed an avoidance tendency.It is shown in the *Homo Sapiens* (SF) domain co-occurrence , which $E \gg F$ and $P \ll 0.05$. The efficacy of Genscan prediction will be verified in *Homo Sapiens* (SF) because it has SF entries and fewer domain hmm co-occurrences. znf-tpr appears to show avoidance in animals, but not in plants. In *Homo sapiens, Oryza sativa* and *Drosophila melanogaster*, $E(GP) \ll E(SF)$, because The GP Domain occurrences are not abundant enough. The ratio of domain occurrences/total genes in $GP \ll SF$

Result(Domain Co-occurrence)

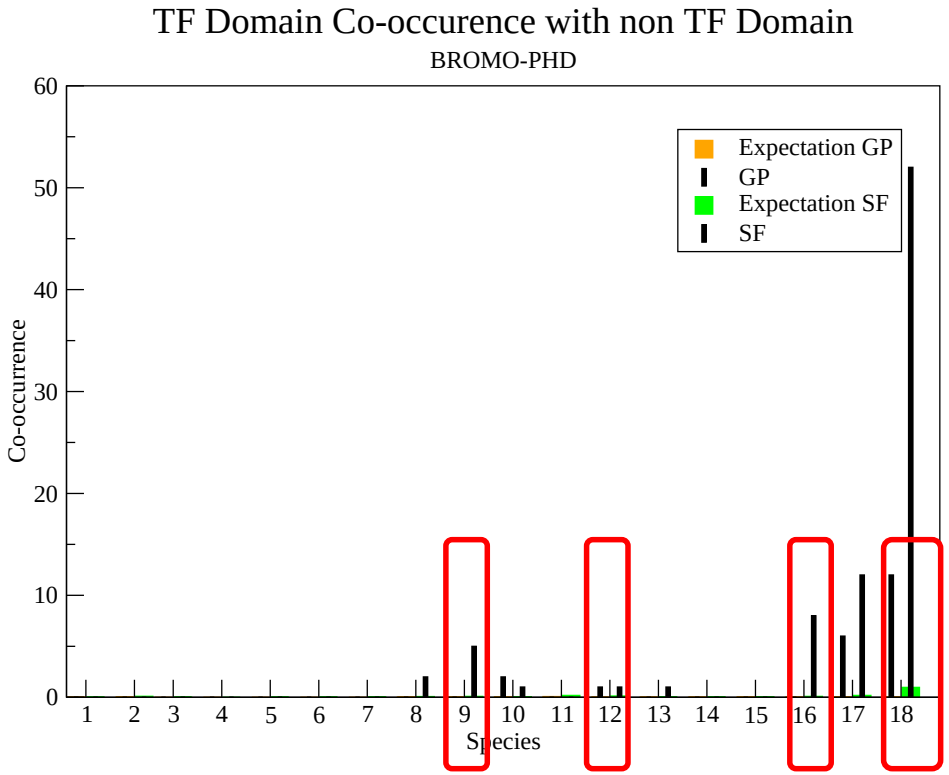
Insight

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.

Result (Domain Co-occurrence)

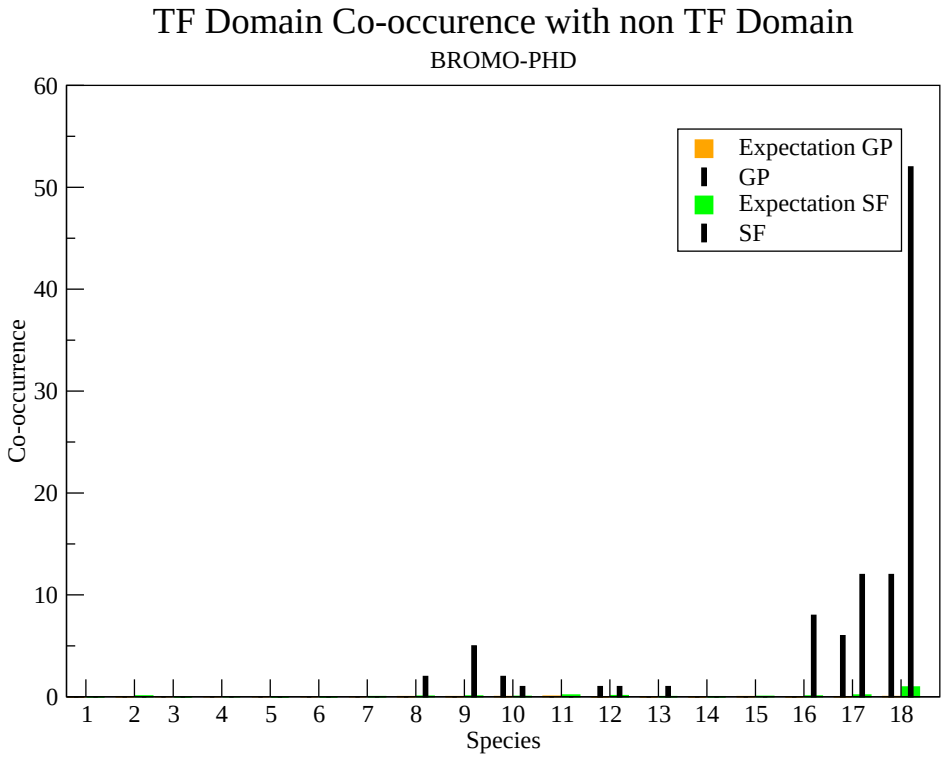


Result (Domain Co-occurrence)

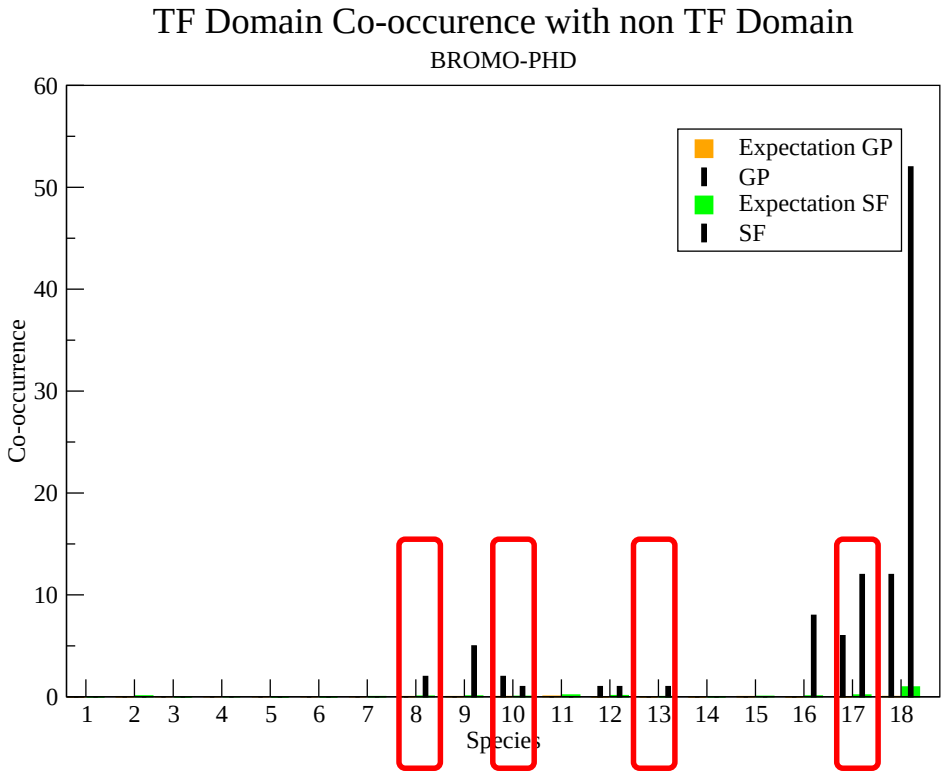


There is a co-occurrence tendency in bromo-phd pair. It shows primarily in *Phytophthora ramorum* (SF), *Oryza sativa* (GP), *Caenorhabditis elegans* (SF), *Homo sapiens* (SF and GP) which $E \ll F$ and $P \ll 0.05$. The efficacy of Genscan prediction will be verified in *Caenorhabditis elegans* (SF), *Phytophthora ramorum* (SF), and *Homo Sapiens* (SF) because they have SF entries and no or few domain hmm co-occurrences. The data is consistent, because in every species, is always $E(GP) \ll GP$ and $E(SF) \ll SF$.

Result (Domain Co-occurrence)



Result (Domain Co-occurrence)



There is a co-occurrence tendency in bromo-phd pair. It shows primarily in *Thalassiosira pseudonana* (SF), *Chlamydomonas* (GP), *Dictyostelium* (SF), and *Drosophila melanogaster* (SF and GP), which $E \ll F$ and $P \ll 0.05$. The search for Hypothetical Protein existence in *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 *Dictyostelium* (SF), *Thalassiosira pseudonana* (SF) and *Drosophila melanogaster* (SF) because they have SF entries and no or few domain hmm co-occurrences. The data is consistent, because in every species, is always $E(GP) \ll GP$ and $E(SF) \ll SF$.

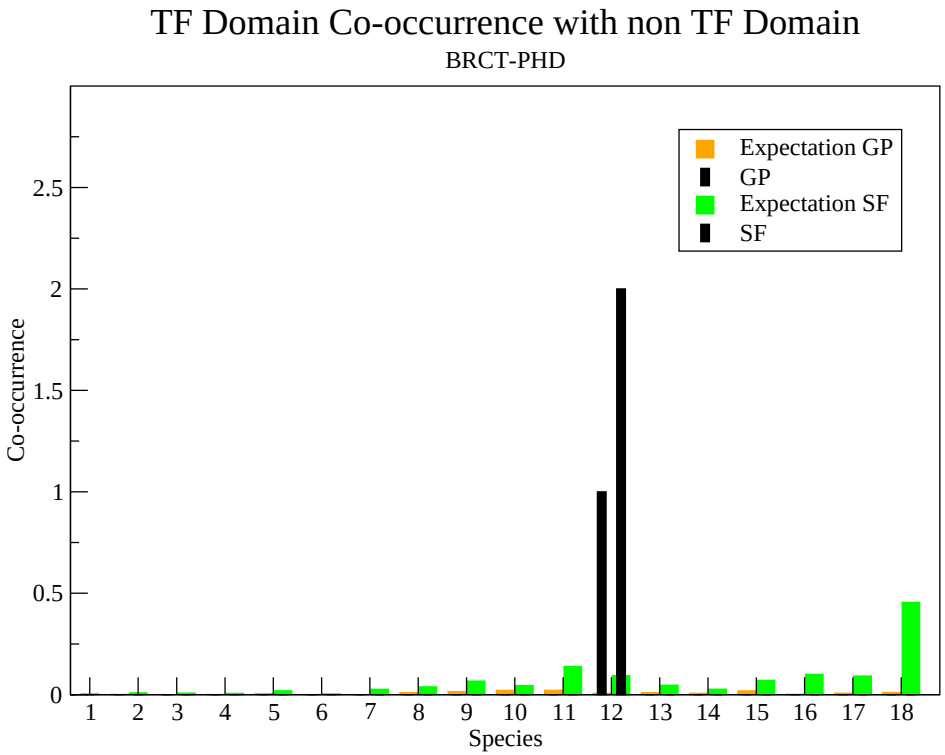
Drosophila's Bromo-Phd Genes annotation

Table 1: Overlapping annotation of Drosophila Genome's Bromo-Phd Domain Co-occurrence

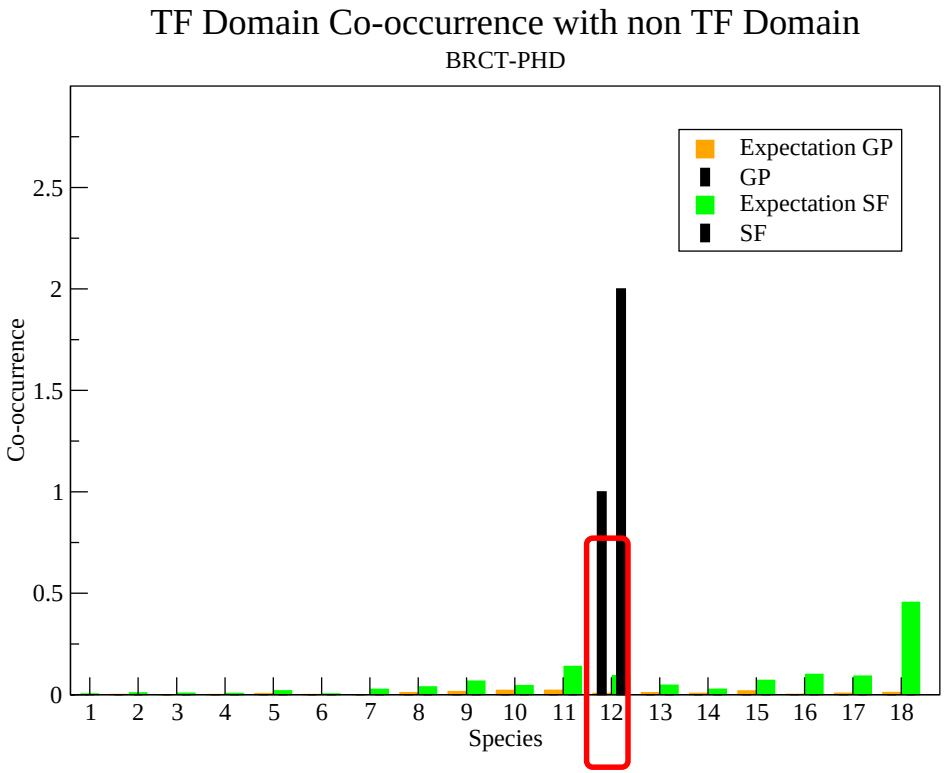
No	Overlapping Loci	SF ID	GP ID
1	chr3R:16,418,875-16,438,264	FlyBaseAnnotationIDs:CG5206-PA,REFSEQ:NP524724	FLYBASE:FBgn0023097,LOCUS=NM079985 LOCUS=AF210315
2	chr2R:3,160,780-3,165,388	FlyBaseAnnotationIDs:CG1845-PA,REFSEQ:NP610266	FLYBASE:FBgn0033155,LOCUS=FJ631261
3	chr2R:7,466,385-7,503,336	FlyBaseAnnotationIDs:CG10897-PA,REFSEQ:NP523701 FlyBaseAnnotationIDs:CG10897-PE,REFSEQ:NP001097270 FBpp0087194 FBpp0087195	FLYBASE:FBgn0033636,LOCUS=NM001103800
4	chr3R:27,619,811-27,625,533	FlyBaseAnnotationIDs:CG1815-PA,REFSEQ:NP733441 FlyBaseAnnotationIDs:CG1815-PB,REFSEQ:NP733442 FlyBaseAnnotationIDs:CG1815-PC,REFSEQ:NP651881 FlyBaseAnnotationIDs:CG1966-PA,REFSEQ:NP536734	FLYBASE:FBgn0027620,LOCUS=NM080486
5	chr3L:234,101-246,901	FlyBaseAnnotationIDs:CG32346-PA,REFSEQ:NP728507 FlyBaseAnnotationIDs:CG32346-PB,REFSEQ:NP728505	FLYBASE:FBgn0000541,LOCUS=NM206224

5 Locis are detected and overlapped.

Result(Domain Co-occurrence)

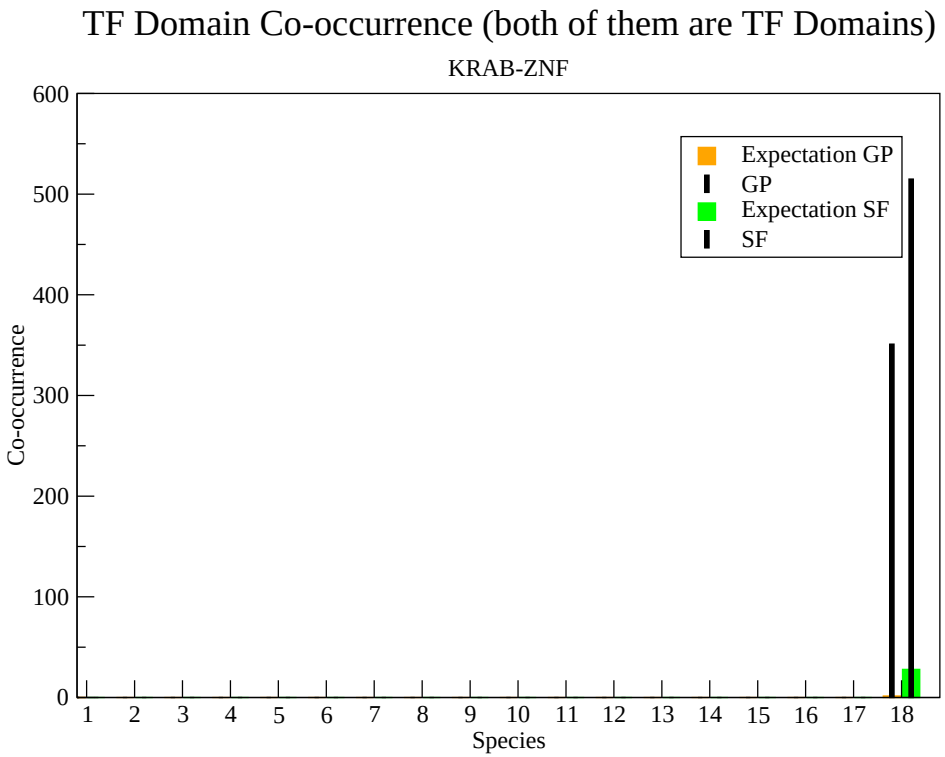


Result(Domain Co-occurrence)

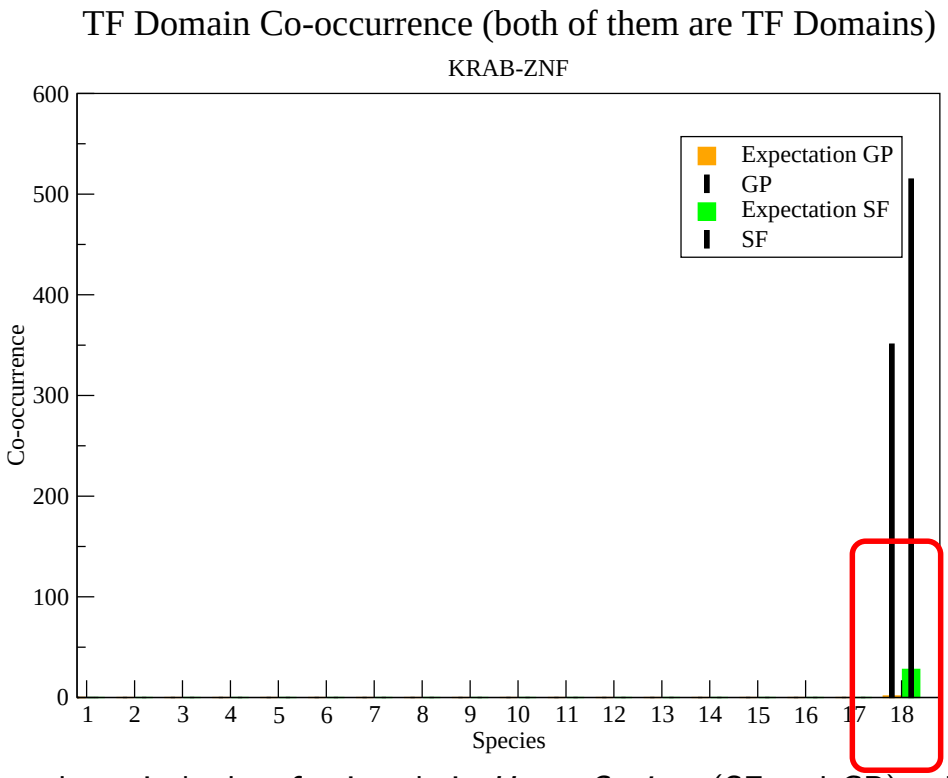


There is a co-occurrence tendency in krab-znf pair only in *Oryza sativa* (SF and GP), which $E \ll F$ and $P \ll 0.05$. The efficacy of Genscan prediction will be verified in *Oryza sativa* (SF and GP) because they have SF entries and no or few domain hmm co-occurrences. Brct domain is the most important adaptor domain involved in eukaryotic repair. It has been detected in a vast variety of proteins involved in repair and cell cycle checkpoint regulation and may provide the critical connections between these processes. Only plants has Phd domain extension.

Result(Domain Co-occurrence)

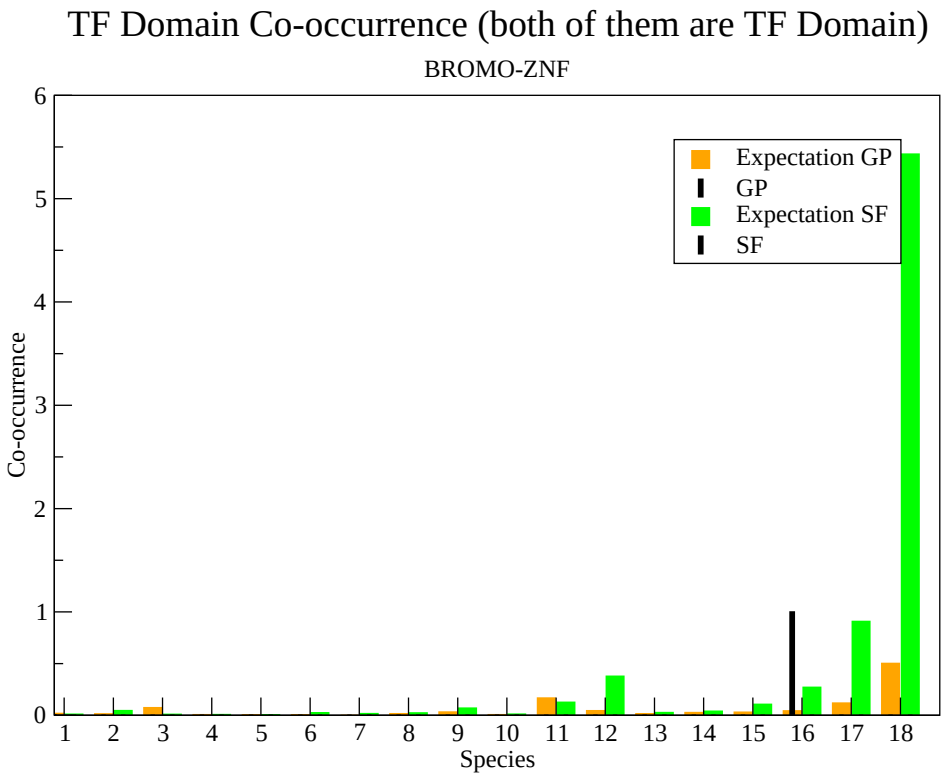


Result(Domain Co-occurrence)

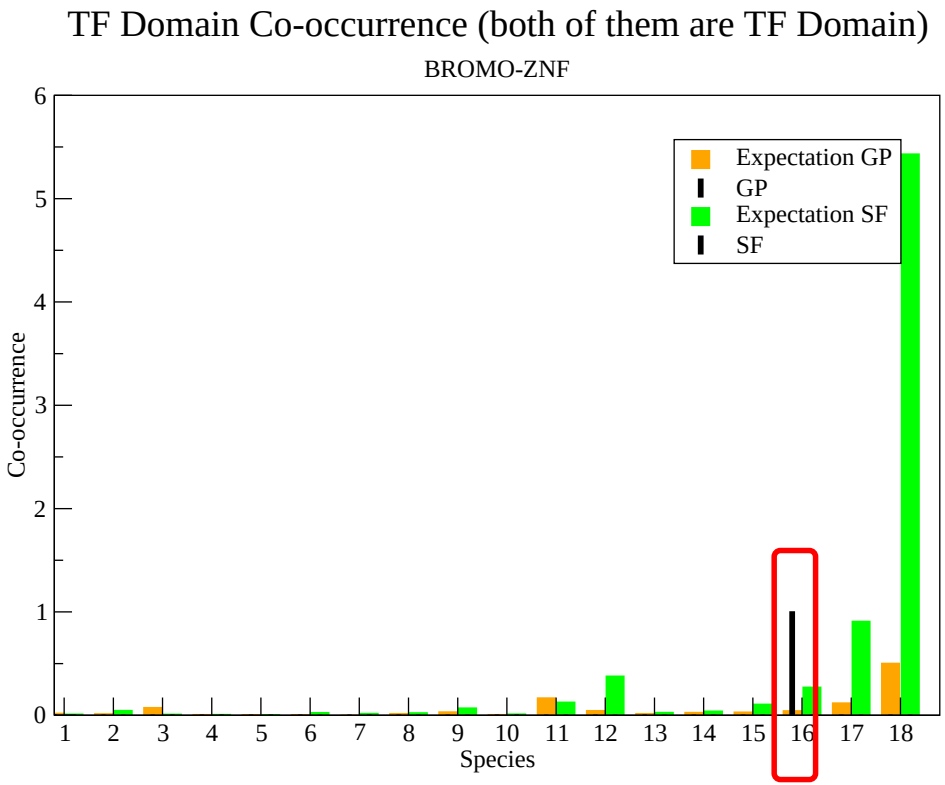


There is a co-occurrence tendency in krab-znf pair only in *Homo Sapiens* (SF and GP), which $E \ll F$ and $P \ll 0.05$. Krab-znf co-occurrence are happening primarily in *Homo sapiens* and overrepresented among genes contained within these recent human SD. Literature stated, that the human genome contains more than 400 KRAB-ZNF genes. Our finding is in accordance with it.

Result(Domain Co-occurrence)

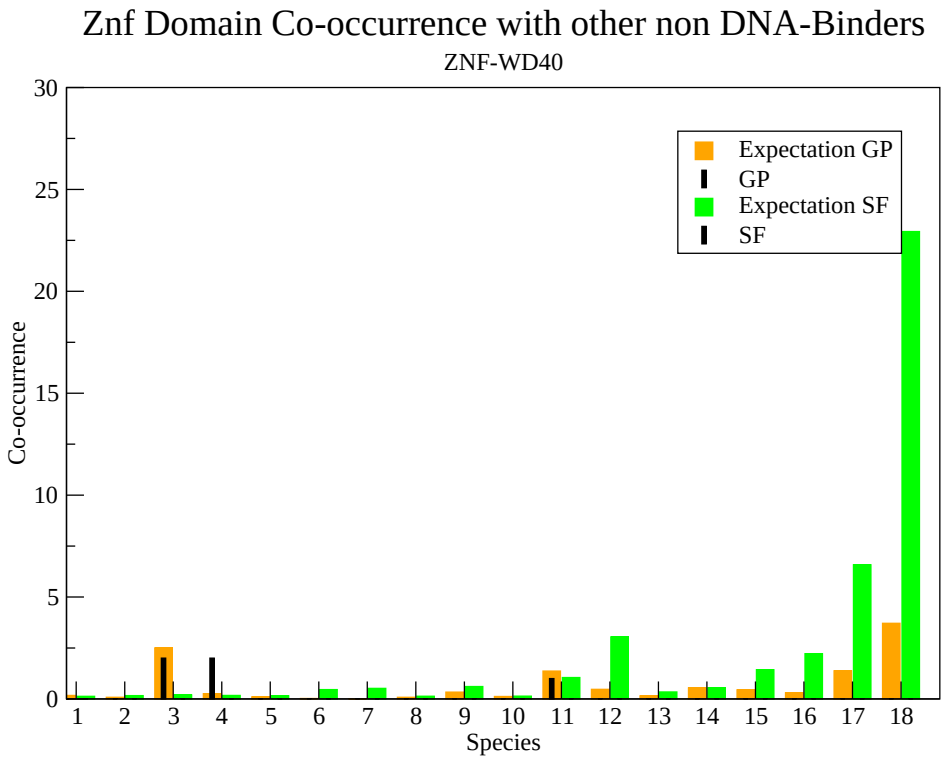


Result(Domain Co-occurrence)

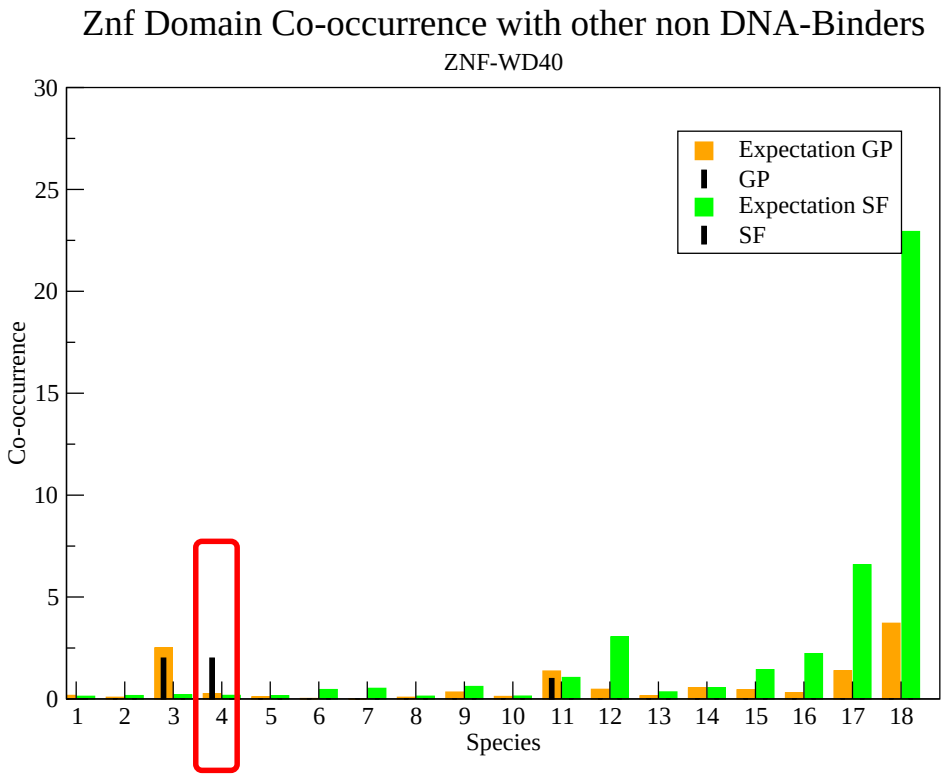


The bromo-znf pairs showed a co-occurrences tendency. It is shown in the *Caenorhabditis elegans* (GP) domain co-occurrence , which $E \ll F$ and $P \ll 0.05$. The found GP annotation is a PolyBRoMo domain containing family member (pbrm-1) protein. PBRM-1 is predicted to function in chromatin remodeling and transcriptional regulation.

Result(Domain Co-occurrence)

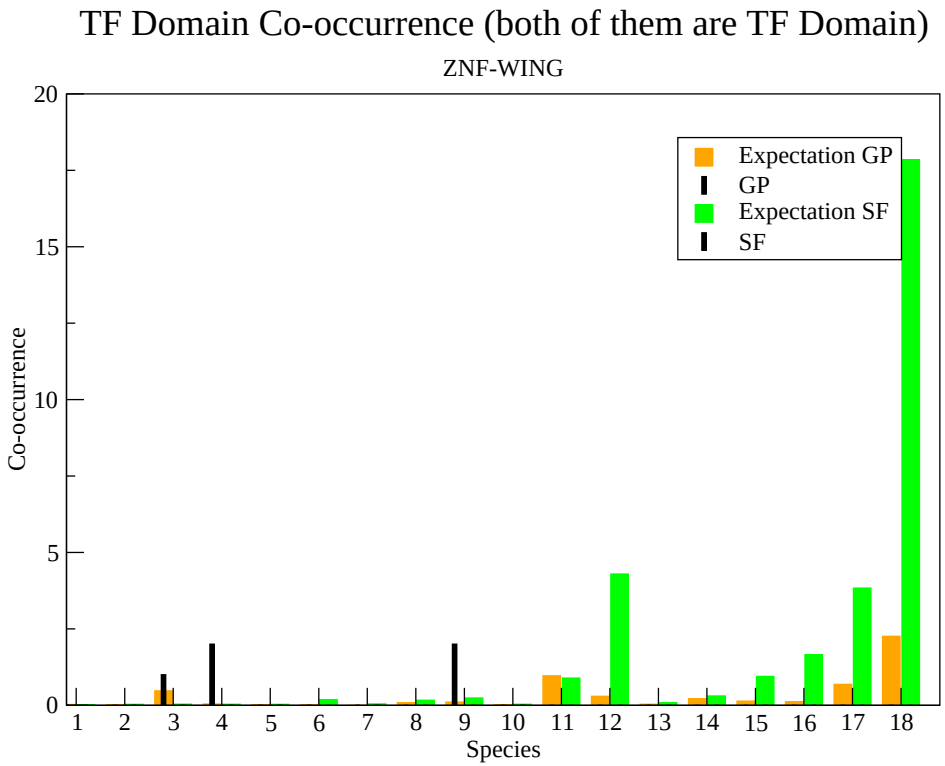


Result(Domain Co-occurrence)

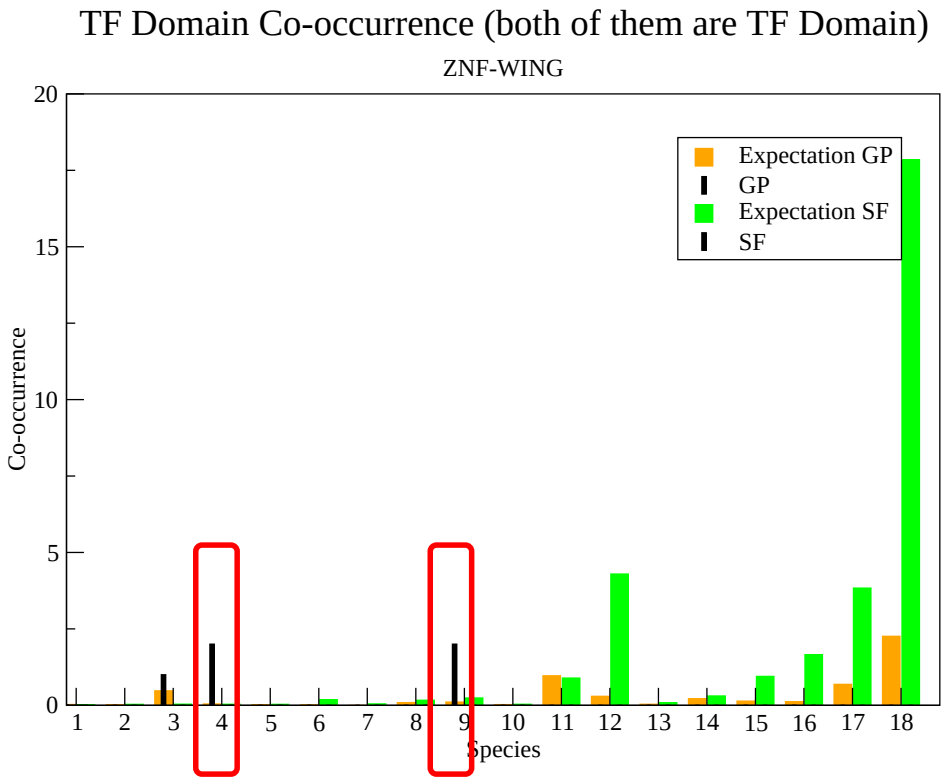


The znf-wd40 pairs showed a co-occurrence tendency.It is shown in the *Leishmania major* (GP) domain co-occurrence , which $E \ll F$ and $P \ll 0.05$. Both GP observation in *Leishmania major* are different Hypothetical protein with unknown function.

Result(Domain Co-occurrence)



Result(Domain Co-occurrence)



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.

Discussion

- 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).

Discussion

- Avoidance and preferential co-occurrence 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.

Summary

Conclusions

- 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 and co-occurrence 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
- znf-tpr appears to show avoidance in animals, but not in plants

Summary

Outlook

- Evaluate the power of quantitative comparative analysis of protein domain (co-)occurences.
- Analysis of protein domains involved in chromatin regulation with higher phylogenetic resolution.
- Annotate more overlapping locis in the selected genomes.
- Conduct a more rigorous post-processing for transplicing genomes and genomes with specific genome organizations.

Acknowledgements

- DAAD (German Academic Exchange Service)
- Peter F Stadler
- Sonja Prohaska
- Christian Arnold