Evolution and Quantitative Comparison of Genome-Wide Protein Domain Distributions

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

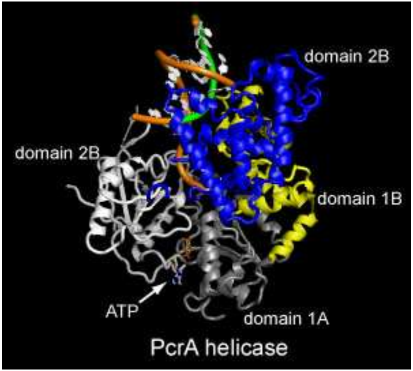
Computational EvoDevo
Institute of Computer Science
University of Leipzig

February 17, 2012

Background

- It should be possible to identify large-scale trends in evolution such as the increased complexity of gene regulation, by comparing the proteomes among species.
- We have 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.
- In particular, it can be utilized for systematic studies of the evolution of protein domain occurrences and co-occurrences.

Background



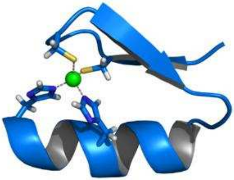
- Proteins are composed of recognizable protein domains that can be re-combined in a combinatorial fashion to produce new functionalities over large time-scales.

Background

- As most proteins contain more than a single domain, domain combinations are of particular interest when aiming a more detailed understanding of the novel functions
- In a study of chromatin evolution, we demonstrated that it is indeed feasible to determine large-scale trends in regulatory capabilities based on domain content.

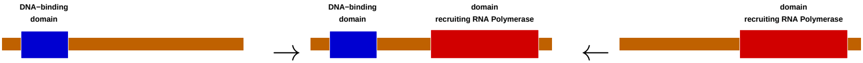
Sonja J. Prohaska, Peter F. Stadler, David C. Krakauer.2010. Innovation in gene regulation: The case of chromatin computation, Journal of Theoretical Biology, Volume 265, Issue 1, Pages 27-44.

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.

Scheme of Transcription Factor Protein


DNA–binding domain
→
DNA–binding domain
domain recruiting RNA Polymerase
←
domain recruiting RNA Polymerase

Function

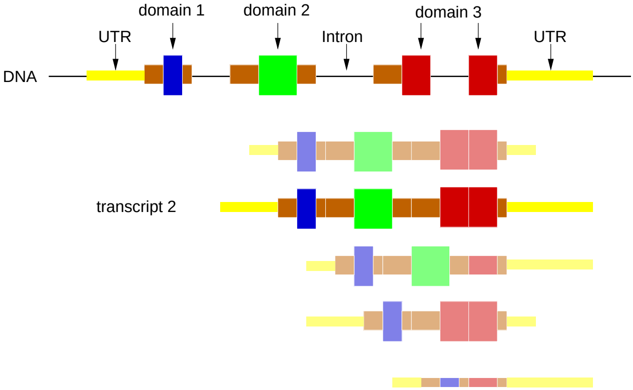
Protein localizes to (sequence-specific) sites on the DNA

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.

Counting Genes and Domain (Co-)Occurences

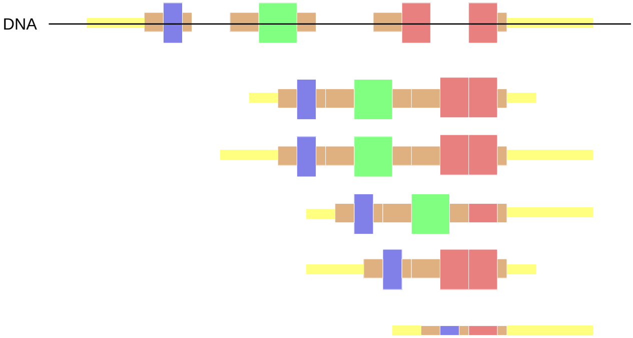


Result: Large discrepancies between the number of transcripts for orthologous loci.

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.

Background

- As a first application of this approach, we investigated the co-occurrences of four major types of DNA binding domains (zinc fingers, leucine-zipper, HMG-box domains, and winged-helix domains), and observed highly significant anti-correlation of the four different domains.
- In contrast, evolutionarily related DNA binding domains readily co-occur in DNA binding proteins.
- In many genomes, in particular the compact genomes of simple unicellular eukaryotes, the total number of genes and domains that can be annotated is too small for a meaningful statistical evaluation.

Arli A. Parikesit , Peter F Stadler, and Sonja J Prohaska. 2010. *Quantitative Comparison of Genomic-Wide Protein Domain Distributions*. GCB2010 conference proceeding. Vol P-173: pp 93-102

Background

- Our last publication has shown that this limitation can be overcome by pooling domains first in terms of domain families or even at the level of functional classes of domains.
- Global pattern of domain co-occurrence and avoidance have been elucidated by `GENSCAN` annotation. However, some genomes are under-represented, such as in *P.falciparum*.
- Here, we are replacing `GENSCAN` with `AUGUSTUS`.

Arli A. Parikesit , Peter F Stadler, and Sonja J Prohaska 2011. *Evolution and Quantitative Comparison of Genome-Wide Protein Domain Distributions*. MDPI Genes 2 no. 4: 912-924.

Phylogenetic distribution of the species

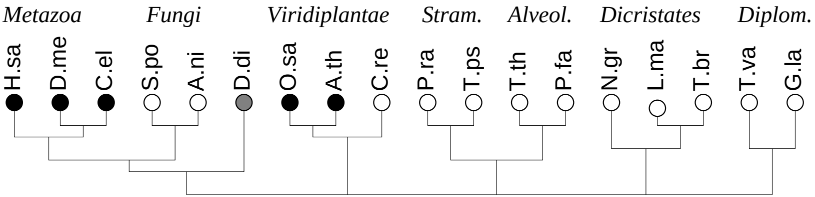


Figure: Phylogenetic distribution of the species considered in this work. Abbreviations: H.sa: *Homo sapiens* (hg19), D.me: *Drosophila melanogaster*, C.el: *Caenorhabditis elegans*, S.po: *Schizosaccharomyces pombae*, A.ni: *Aspergillus niger*, D.di: *Dictyostelium discoideum*, O.sa: *Oryza sativa*, A.th: *Arabidopsis thaliana*, C.re: *Chlamydomonas reinhardtii*, P.ra: *Phytophthora ramorum*, T.ps: *Thalassiosira pseudonana*, T.th: *Tetrahymena thermophila*, P.fa: *Plasmodium falciparum*, N.gr: *Naegleria gruberi*, L.ma: *Leishmania major*, T.br: *Trypanosoma brucei*, T.va: *Trichomonas vaginalis*, G.la: *Giardia lamblia*; *Stram.*: Stramenopiles, *Alveol.*: Alveolata, *Diplom.*: Diplomonada.

Material and Methods

AUGUSTUS Annotation

- Gene predictions were performed using AUGUSTUS
- Protein sequences were extracted directly from the AUGUSTUS predictions.
- Duplicate predictions in the overlaps between fragments were removed

Material and Method

HMM

- In order to obtain comparable domain predictions across the widely different eukaryotic genomes we took all Hidden Markov Models (HMMs) provided by the SUPERFAMILY database
- We used `HMMER 3.0rc1` to map the HMMs to amino acid sequences predicted by `AUGUSTUS` with the cut-off $E \leq 1$. Only the best scoring domain from a set of overlapping domains is considered further.
- Genes were classified as zinc finger genes if they contained at least one C2H2 domain (SCOP family 57668).
- The C2H2-like fold group is by far the best-characterized class of zinc fingers and are extremely common in mammalian transcription factors
- The result is, for each predicted protein, a list of non-overlapping domains.

Material and Method

GO Annotation

Version 1.75 of the SUPERFAMILY database offers a “Structural Domain Functional Ontology” providing functional and phenotypic annotations of protein domains at the **superfamily** and **family** levels. Since any protein can be annotated by multiple functions, it is clear that membership in GO annotation classes does lead to a partition of the set of protein domains into functional groups.

Material and Methods

GO Annotation

bN *binding of nucleic acids*: GO:0003676 at superfamily level.

bP *binding of proteins* with potential nuclear localization: GO:0005515 superfamily level.

rC *regulation of chromatin* GO:0016568 at superfamily level.

rC* *regulation of chromatin* as determined in our previous publication, comprising a combination of family and superfamily level.

rB *regulation of binding*: GO:0051098 at superfamily level.

rE *regulators of enzymatic activity*: GO:0050790 at superfamily level.

mS *metabolism of saccharides*: GO:0005976 at superfamily level.

The five functional groups bN, bP, rC, rC* and rB were chosen due to expected preferential co-occurrence with zinc finger genes. The groups rE and mS were chosen to serve as negative control

Material and Methods

Co-occurrence Analysis

- For each of the 18 species, we separately evaluated the number of domain co-occurrences and the number of genes in which two domains x and y co-occur.
- Let n_x be the total number of x-domains that are annotated.
- The simplest estimate for the expected number of domain co-occurrences is $E(x, y) = n_x n_y / n_g$, where n_g is the number genes in the genome under consideration.
- This estimate does not account for biases arising from the non-uniform distribution of domains over genes.

Material and Methods

Co-occurrence Analysis

- Thus, let n_g and n_d be the numbers genes and domains, respectively.
- Furthermore, let $n_d(i)$ be the number of domains in predicted gene i, and denote by n_x the total number of domains of functional group x. Then the number of x-domains that occur in genes that also contain a y domain is

$$E(x|y) = (n_x/n_d) \sum_{i:y \in \boldsymbol{I}} (n_d(i) - 1) \qquad (1)$$

where the sum runs over all genes that contain domain y.

Material and Methods

Co-occurrence Analysis

- These expectations are then compared with the number of empirically observed co-occurrences $n(x,y)$.
- We speak of *co-occurrence* of domain families or groups if $n(x,y) \gg E(x|y)$ and of *avoidance* if $n(x,y) \ll E(x|y)$.
- The statistical significance of an observed difference between $n(x,y)$ and $E(x|y)$, is determined under the assumption that $n(x,y)$ is drawn from a Poisson distribution.

Results

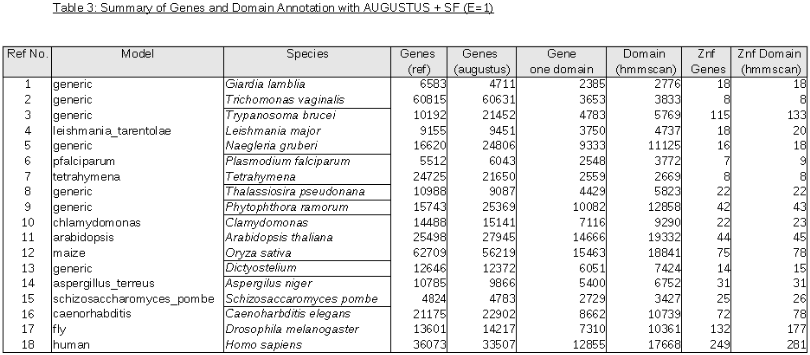


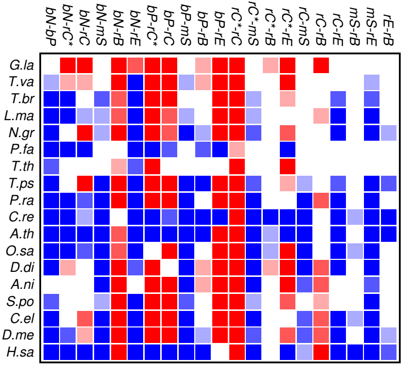
Table 3: Summary of Genes and Domain Annotation with AUGUSTUS + SF (E=1)

Ref No.	Model	Species	Genes (ref)	Genes (augustus)	Gene one domain	Domain (hmmscan)	Znf Genes	Znf Domain (hmmscan)
1	generic	*Giardia lamblia*	6583	4711	2385	2776	18	18
2	generic	*Trichomonas vaginalis*	60815	60631	3653	3833	8	8
3	generic	*Trypanosoma brucei*	10192	21452	4783	5769	115	133
4	leishmania_tarentolae	*Leishmania major*	9155	9451	3750	4737	18	20
5	generic	*Naegleria gruberi*	16620	24806	9333	11125	16	18
6	pfalciparum	*Plasmodium falciparum*	5512	6043	2548	3772	7	9
7	tetrahymena	*Tetrahymena*	24725	21650	2559	2669	8	8
8	generic	*Thalassiosira pseudonana*	10988	9087	4429	5823	22	22
9	generic	*Phytophthora ramorum*	15743	25369	10082	12858	42	43
10	chlamydomonas	*Clamydomonas*	14488	15141	7116	9290	22	23
11	arabidopsis	*Arabidopsis thaliana*	25498	27945	14666	19332	44	45
12	maize	*Oryza sativa*	62709	56219	15463	18841	75	78
13	generic	*Dictyostelium*	12646	12372	6051	7424	14	15
14	aspergillus_terreus	*Aspergilus niger*	10785	9866	5400	6752	31	31
15	schizosaccharomyces_pombe	*Schizosaccaromyces pombe*	4824	4783	2729	3427	25	26
16	caenorhabditis	*Caenoharbditis elegans*	21175	22902	8662	10739	72	78
17	fly	*Drosophila melanogaster*	13601	14217	7310	10361	132	177
18	human	*Homo sapiens*	36073	33507	12855	17668	249	281

Table: Overlaps between the 7 functional groups defined in the text.

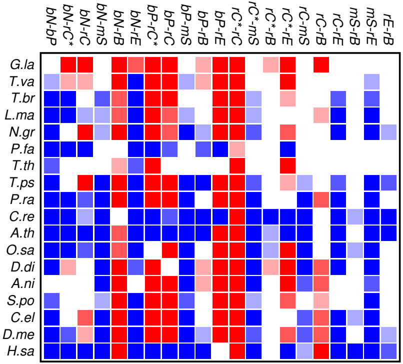
	bN	bP	rC	rC*	mS	rB	rE
bN	112	4	4	4	0	8	6
bP		118	6	7	0	4	21
rC			25	11	0	1	0
rC*				27	0	1	2
mS					14	0	0
rB						15	1
rE							55

Results



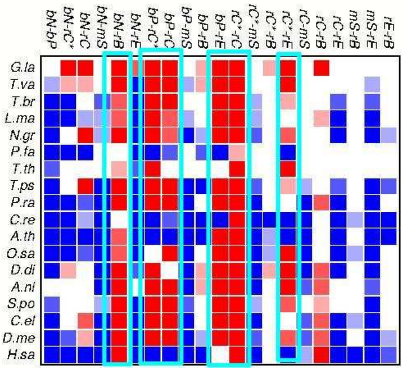
Summary of domain co-occurrences in 18 eukaryotic genomes. Colors indicate the statistical significance of co-occurrence $n(\mathcal{C},\mathcal{D}) \gg E(\mathcal{C}|\mathcal{D})$ (red) and of avoidance $n(\mathcal{C},\mathcal{D}) \ll E(\mathcal{C}|\mathcal{D})$ (blue). Significance levels on individual comparisons are shown in three levels of color saturation for $p < 0.001$, $0.001 \leq p < 0.01$, and $0.01 \leq p < 0.1$, respectively.

Results



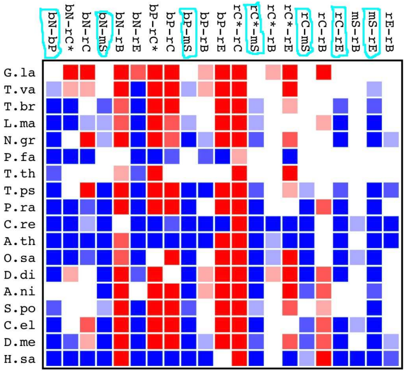
- The data shows many instance of avoidance. However, avoidance tendecy are strongly observed in multicelluar organism.This is in accordance to our previous publication.
- The annotation data of *Plasmodium* and *Tetrahymena* are available, although the occurrences are rather scarce.

Results

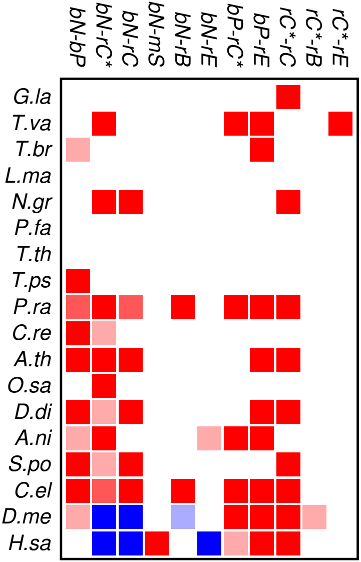


- rC*-rC pair co-occurrence is conserved in every species, which is an interesting phenomenon for chromatin regulation domain.
- While the c-occurrence of bN-rB, bP-rC, and bP-rC* are conserved in most species.

Results

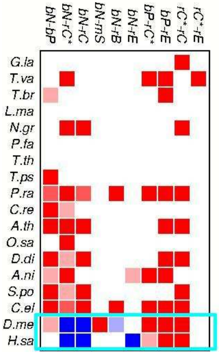


- Avoidance conservation was seen mostly in bN-bP, bN-mS, bP-mS, rC*-mS, rC-mS, rC-rE, mS-rE
- There is a strong tendency of non-occurrence in mS-rB and rE-rB.

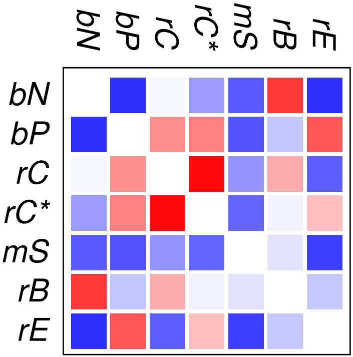


Summary of domain co-occurrences of functional classes of protein domains in zinc

Results

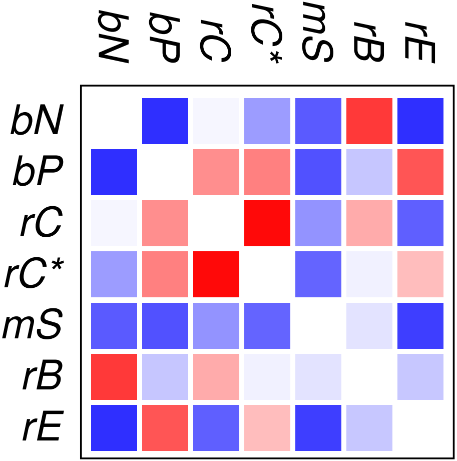


- we investigate to what extent the occurrence and co-occurrence of other domains is influenced by the additional presence of a zinc finger domain
- In this respect, there are no detected co-occurrences in *P.falciparum*, *Tetrahymena*, and *L.major* genome. However, there are strong tendency of domain avoidance in higher organism, in this respect was shown in *D.melanogaster* and Human



Summary of co-occurrence data

- Due to the large differences in genome size and domain numbers it makes little sense to compute a summary statistic by adding up the counts of occurrences across species: such data would be dominated by the large, gene-rich multicellular organisms.
- Instead we employ a simple voting procedure, associating scores of $+3$, $+1$, -1, and -3 only with the two most significant levels of co-occurrence and avoidance, respectively. Fig above displays these scores averaged over the 18 species.



Summary of co-occurrence data

- We find that slight majority of the domain GO-classes are at least weakly negatively correlated. However, some classes are positively correlated, such as strong correlation at rC-rC*.

Discussion

- Despite obvious shortcoming of the gene finding procedure in organisms with unusual genome structure or extreme sequence composition and the unavoidable limitations of the domain annotation, some global patterns nevertheless become visible in this pilot study.
- The classes of protein domains investigated here are all involved in binding and/or regulation.

Conclusion

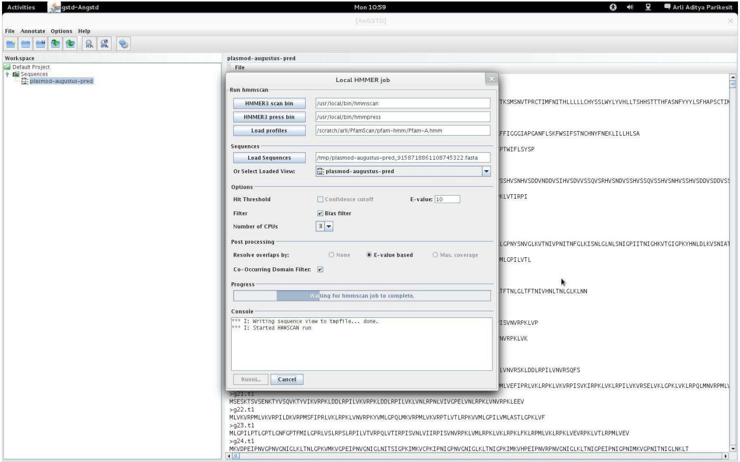
- In the multi-cellular organisms with large genomes and large gene families, however, we observe a strong signal of avoidance between several functional groups of protein domains.
- This may be a result of the expansion and diversification of large families of paralogous genes and their use for specific regulatory task.
- Furthermore, we observe substantial differences in the domain co-occurrence patterns of distant lineages, emphasizing the importance of lineage-specific histories and constraints.

Summary

Outlook

- Train Genomes without provided AUGUSTUS model for improving annotation accuracy.
- Using more sophisticated features of ANGSTD domain-annotation pipeline for annotating PFAM domain.

ANGSTD



Run hmmscan

HMMER3 scan bin: /usr/local/bin/hmmscan

HMMER3 press bin: /usr/local/bin/hmmpress

Load profiles: /scratch/arli/PfamScan/pfam-hmm/Pfam-A.hmm

Sequences

Load Sequences: /tmp/plasmod-augustus-pred_9158718861108745322.fasta

Or Select Loaded View: plasmod-augustus-pred

Options

Hit Threshold: □ Confidence cutoff E-value: 10

Filter: ☑ Bias filter

Number of CPUs: 3

Post processing

Resolve overlaps by: ○ None ● E-value based ○ Max. coverage

Co-Occurring Domain Filter: ☑

Progress

Waiting for hmmscan job to complete.

Console

```
*** I: Writing sequence view to tmpfile... done.
*** I: Started HMMSCAN run
```

```
TKSMSNVTPRCTIMFNITHLLLLLCHYSSLWYLYVHLLTSHHSTTTHFASNFYYYLSFHAPSCTIN
FFIGGGIAPGANFLSKFWSIFSTNCHNYFNEKLILLHLSA
PTWIFLSYSP
SSHVSNHVSDDVNDDVSIHVSDVVSSQVSRHVSNDVSSHVSSQVSSHVSNHVSSHVSDDVSDDVSS
KLVTIRPI
LGPNYSNVGLKVTNIVPNITNFGLKISNLGLNLSNIGPIITNIGHKVTGIGPKYHNLDLKVSNIAT
MLGPILVTL
TFTNLGLTFTNIVHNLTNLGLKLNN
ISVNVRPKLVP
NVRPKLVK
LVNVRSKLDDLRPILVNVRSQFS
MLVEFIPRLVKLRPKLVKVRPISVKIRPKLVKLRPILVKVRSELVKLGPKLVKLRPQLMNVRPMLV
>g21.t1
MSESKTSVSENKTYVSQVKTYVIKVRPKLDDLRPILVKVRPKLDDLRPILVKLVNLRPNLVIVGPELVNLRPKLVNVRPKLEEV
>g22.t1
MLVKVRPMLVKVRPILDKVRPMSFIPRLVKLRPKLVNVRPKYVMLGPQLMKVRPMLVKVRPTLVTLRPKVVMLGPILVMLASTLGPKLVF
>g23.t1
MLGPILPTLGPTLGNFGPTFMILGPRLVSLRPSLRPILVTVRPQLVTIRPISVNLVIIRPISVNVRPKLVMLRPKLVKLRPKLFKLRPMLVKLRPKLVEVRPKLVTLRPMLVEV
>g24.t1
MKVDPEIPNVGPNVGNIGLKLTNLGPKVMKVGPEIPNVGNIGLNITSIGPKIMKVCPKIPNIGPNVGNIGLKLTNIGPKIMKVHPEIPNVRPNVGNIGLKLTNIGPEIPNIGPNIMKVGPNITNIGLNKLT
```

AnGSTD shows you a tree with corresponding domain arrangements of the proteins at the same time.

Acknowledgements

- DAAD (German Academic Exchange Service)
- Peter F Stadler
- Sonja J Prohaska
- Hai Fang (Bristol)
- Andrew Moore (Münster)
- Winterseminar Organizer