Quantitative Comparison of Genomic-Wide Protein Domain Distributions

Arli A. Parikesit[1*], Peter F. Stadler[1−5], Sonja J. Prohaska[1],
[1]Bioinformatics Group, Dept. Computer Science,
and Interdisciplinary Center for Bioinformatics, University of Leipzig,
Härtelstraße 16-18, D-04107 Leipzig, Germany
[2]Max Planck Institute for Mathematics in the Sciences,
Inselstraße 22, D-04107 Leipzig, Germany
[3]Fraunhofer Institute for Cell Therapy und Immunology,
Perlickstr.1, D-04103 Leipzig, Germany
[4]Institute for Theoretical Chemistry, University of Vienna,
Währingerstrasse 17, A-1090 Vienna, Austria
[5]The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico
[*]Corresponding Author

Abstract: Investigations into the origins and evolution of regulatory mechanisms require quantitative estimates of the abundance and co-occurrence of functional protein domains among distantly related genomes. Currently available databases, such as the SUPERFAMILY, are not designed for quantitative comparisons since they are built upon transcript and protein annotations provided by the various different genome annotation projects. Large biases are introduced by the differences in genome annotation protocols, which strongly depend on the availability of transcript information and well-annotated closely related organisms.

Here we show that 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. As an application, we considered four major classes of DNA binding domains: zink-finger, leucine-zipper, winged-helix, and HMG-box. We found that different types of DNA binding domains systematically avoid each other throughout the evolution of Eukarya. In contrast, DNA binding domains belonging to the same superfamily readily co-occur in the same protein.

1 Introduction

The expression of genomically encoded information is subject to tight regulation and control in all organisms that have been studied in detail. These regulatory rules are implemented in a highly complex network of several biochemically distinct mechanism that act at multiple levels of the gene expression cascade. They include specific chromatin states, the action of transcription factors, regulated mRNA export, alternative splicing, translational control, post-transcriptional and post-translational modifications, and con-

German Conference on Bioinformatics 2010
Dietmar Schomburg, Andreas Grote (Eds.)
Lecture Notes in Informatics (LNI) Volume P-173

trolled degradation of both RNA and polypeptides. Surprisingly, it appears that different phylogenetic clades emphasize certain types of mechanisms while reducing or even abolishing others. Regulation in eubacteria, for example, appears to be dominated by transcription factors networks, trypanosomes use the post-transcriptional processing of large polycistronic transcripts, ciliates utilize extensive amplification of DNA in creating their macro-nuclei, and crown group eukaryotes have evolved an elaborates system of histone modifications. An understanding of the diversity of life thus requires the investigation of the origin(s) and evolution of these different regulatory mechanisms and their interplay.

The most direct approach towards this goal is the comprehensive reconstruction of the evolutionary histories of the many protein families that play a role in the various modes of evolution. In practice, however, this is an exceedingly difficult and tedious task, since homologies even between highly conserved proteins become hard to establish in comparisons across kingdoms or even across the three domains of life. This is not only for technical reasons: Proteins are composed of recognizable protein domains that implement well-defined functions such as catalytic activities, specific binding, and anchoring in membranes. Over large time-scales, these components have been combined in a combinatorial fashion to produce new functionalities, so that individual proteins often have multiple ancestors that contributed different domains [MBE+08, KAK00]. A more modest approach thus aims at tracing the *distribution* of protein domains comparatively. In a recent study of chromatin evolution, we demonstrated that this is indeed feasible [PSK10]. More detailed insights can be gained from considering domain combinations. For instance, Itoh *et al.* [INK+07] showed that there are many animal-specific or even vertebrate-specific domain-combinations. Network analysis of domain co-occurrences, furthermore, demonstrates a growing core of combinations in multicellular organisms [WA05].

Typically, studies of this type are based on existing annotation. For instance, the protein annotation compiled in `KEGG`, `ENSEMBL` and `Pfam` [FMSB+06] domains were used in [INK+07], ref. [PSK10] was based on the `SUPERFAMILY` database [WPZ+09], whose HMM models in turn are based on the SCOP (Structural Classification of Proteins) domain definitions [AHC+08].

We recently attempted to investigate the origins of the proteins associated with the microRNA pathway using a rather straightforward approach: For each of the most prominent proteins associated with the microRNA pathway (Drosha, Dicer, DGCR8, TRBP, and TRBP), we searched the SUPERFAMILY database for putative homologs. To this end, we collected the functional domains of these proteins from the literature and then identified the SUPERFAMILY peptide entries in which these known domains co-occurred. Somewhat surprisingly, this approach did not recover the phylogenetic distributions reported in detailed, homology-based studies [CR07, MDB08]. Apart from domains that were missing completely (such as PIWI), we observed that many domains are annotated only in a small subset of the species that are expected to contain them. We concluded from this pilot study that existing peptide annotations are a problematic data source for quantitative cross-species comparisons. The issues are twofold:

1. A comprehensive analysis of the evolution of gene *function* requires a reasonably complete collection and annotation of protein domains. Of course, the current

knowledge is not complete, and there are still novel functional domains yet to be discovered. Interestingly in that regard, co-occurrence data can help to detect undescribed and divergent protein domains [TGMB09]. Furthermore, most protein domains in well-studied model organisms are evolutionarily very old, suggesting that the innovation of protein domains is a relatively infrequent phenomenon [BBHS10]. For example, a recent study showed that the majority of "plant-specific" DNA binding domains originated much earlier then the comparably recent expansion into the diverse gene families present in higher plants [SeoopstfDbd08].

2. The annotation of protein domains is performed on protein sequences retrieved from sequence databases. For each species, these "protein models" are constructed by combining the genomic DNA sequence, EST and cDNA data, and computational predictions. Large differences in EST and/or cDNA coverage as well as in the computational procedures imply that domain annotations can be very different even for phylogenetically closely related species. For example, the current version (1.73) of SUPERFAMILY annotates 64225 domains in human, but only 45312 in chimpanzee, 21208 in gorilla and 14748 in the alpaca, although one would expect a very similar gene complement throughout the eutherian mammals.

In this contribution, we focus on the second issue and investigate strategies to construct inventories of protein domains that avoid the biases arising from gene annotation. While it would certainly be desirable to obtain a complete set of protein domains encoded in any given genome, this is not feasible at present. Our goal here is thus more moderate: we are content with estimates that are consistent between different genomes and thus allow quantitative comparisons. To this end, we re-annotate protein domains using the following three different collections of (putative) polypeptides for each genome: (1) computational translations of annotated transcripts available in sequence databases, (2) conceptual translations of the entire genomic DNA in all 6 reading frames, and (3) protein predictions generated by a *de novo* gene predictor.

2 Materials and Methods

As test system we use the genomes of three apes (human GRCh37.57, chimp CHIMP2.1.57, and gorilla gorGor3.57). The genomes were downloaded from the ENSMBL website (`www.ensembl.org`), version 57. Transcript files were downloaded from the cDNA section of the corresponding genome builds. The three ape species are so similar that we can expect a virtually identical complement of protein domains. Even in very rapidly evolving gene families, such as the KRAB-ZNF family of transcriptional repressors [NHZS10], the copy numbers differences in between primates are restricted to a few percent. The most extreme case are olfactory receptors [Nii09], where the number of functional copies differs by up to 25% between human and chimp due to massive gene loss [GN08]. This difference, however, will not be clearly detectable at domain level, since many of the very recent pseudogenes are expected to yield inconspicuous hits to the HMM domains models. In contrast to expected similarity of the great apes, their transcriptome and proteome

Table 1: Summary statistics of source data. The number of domains refers to query set of 100 randomly selected SCOP entries. n.d.: not determined.

Species	Human	Chimpanzee	Gorilla	Yeast
Data set	RCh37.57	CHIMP2.1.57	gorGor3.57	SGD1.01.57
	number of peptides investigated			
transcripts	76592	34142	27325	5885
`genscan`	118894	96615	113532	4197
	number of detected domains			
transcripts	5551	3769	3386	621
`genscan`	3392	2796	3323	614
genomic translation	23	n.d.	n.d.	409

annotations differ by nearly a factor of three, Tab. 1.

Gene predictions were performed using `genscan` [BK97, BK98]. To this end, the chromosomes were split into fragments between 500kb and 600kb since `genscan` does not accept larger input files. The sequences of the predicted genes were extracted directly from the `genscan` output. The chromosome fragments were constructed with substantial overlaps to avoid artifacts arising from incomplete gene predictions at fragment boundaries, leading to redundant predictions within the overlapping regions. These were removed before further analysis.

We also tested `GeneMark` [LTHCM05] as an alternative gene predictor and obtained comparable results. We decided to focus on `genscan` because: (1) it has been reported to perform well across distantly related species (teleost fishes, nematodes, amphioxus, and fungi) without retraining its internal model [Kor04], (2) because it is much faster than the alternatives, and (3) because it is the mostly widely used gene predictor [MMNH04].

Protein domains are represented as Hidden Markov Models (HMMs) [Edd96, DEKM98, Edd98]. In order to save computations resources we randomly selected 100 domains from the SUPERFAMILY database [WPZ+09], version 1.73 (10.01.2010) for the statistical analysis. We used `HMMER 3.0rc1` to map the HMMs to the protein sequences with the the same E-value cut-off as the SUPERFAMILY: $E \leq 10^{-4}$. In case of overlapping HMM hits, we retain only the best-scoring match.

3 Results

Scatter-plots of the number of domain occurrences measured on the set of annotated transcript and on the *de novo* gene predictions shows a significant correlation, Fig 1. In contrast, an attempt to estimates the domain numbers by running the HMMs on translated genomic DNA failed miserably: only a small fractions of the known domains can be recovered. This is not surprising. Although there is a statistically significant correlation

between protein domain boundaries and exon boundaries [LWWG05], about two thirds of the annotated protein domains domains are interrupted by at least one introns, and on average a domain contains 3 or 4 introns [BPMS09]. Thus most domains are undetectable in conceptual translations of the genomic DNA.

In the human data, the majority of domains is observed more frequently in annotated transcripts than in `genscan` predictions (Fig. 1a). This effect is less pronounced in chimpanzee (Fig. 1b). In yeast, on the other hand, the correspondence between transcript-based domain annotation and the `genscan`-based results is excellent. We can understand these differences because of dramatic differences in the quality and coverage of the transcript annotation. In the human genome, for example, a large number of annotated isoforms and alternative transcripts are annotated as a result of extensive cataloging efforts. Thus, multiple transcripts may incorporate the same genomic domain. A comparable density of data is not available for any other species, which results in an inevitable underestimation of annotated transcripts (as in the two ape genomes). Transcript annotation and `genscan` predictions agree extremely well in yeast, however. The data in Table 2 show a good overall correlation between the domain counts as reported by the SUPERFAMILY database and those computed from the `genscan` predictions, although counts can deviate largely in some species. For instance, in *Trypanosoma brucei* we detect 146 zinkfingers using gene predictions compared to only 7 annotated in SUPERFAMILY.

To investigate the suitability of gene predictions for the assessment of domain co-occurrences, we selected two very abundant classes of DNA binding domains: zink-finger domains (ZNF) and winged-helix domains. If the two domain types were distributed randomly, we would expect about 17.8 co-occurrences, estimated from the data in the SUPERFAMILY (30712 transcripts, of which 1324 contain a ZNF domain and 414 have a winged-helix domain). Surprisingly, not a single co-occurrence between these two domains is observed in the SUPERFAMILY data in any species, even though both domains are conserved throughout the Eukarya, Table 2.

In the `genscan`-based analysis, we detected co-occurrences of ZNF and winged-helix domains only in the clades Kinetoplastida (*Leishmania* and *Trypansoma*) and in *Phytophtora*. Upon closer inspection, these can can be identified as artifacts. In Kinetoplastida, the problem is caused by the unusual structure of the transcriptome of Kinetoplastida, which consists of long, polycistronic mRNAs that are processed by transsplicing [MCVdRFM$^+$10]. Our hits fall into a highly conserved polycistron of more than 10kb length, for which `genscan` predicts a "polyprotein". Interestingly, no spurious co-occurrences are found in the nematode *C. elegans*, whose polycistronic messages contain much fewer proteins. The second artifact are two hits in *Phytophtora*: one is again a putative artifact `genscan`, which here predicts a chimera of RNA polymerase III subunit C34 and a hypothetical zinkfinger protein. The second hit covers a protein annotated as homolog of the EAP30 subunit of the ELL complex containg two winged-helix domains. In the latter case, the zink-finger domain is most likely located in an additional downstream exon that is conserved between *Phytophtora sojae* and *Phytophtora ramorum*.

The exclusive usage of one of the two types of DNA binding domains is statistically highly significant. In human, for instance, we expect 11.7 co-occurrences (5090 ZNF and 274 winged-helix domains in 118894 `genscan` predictions) while none is observed

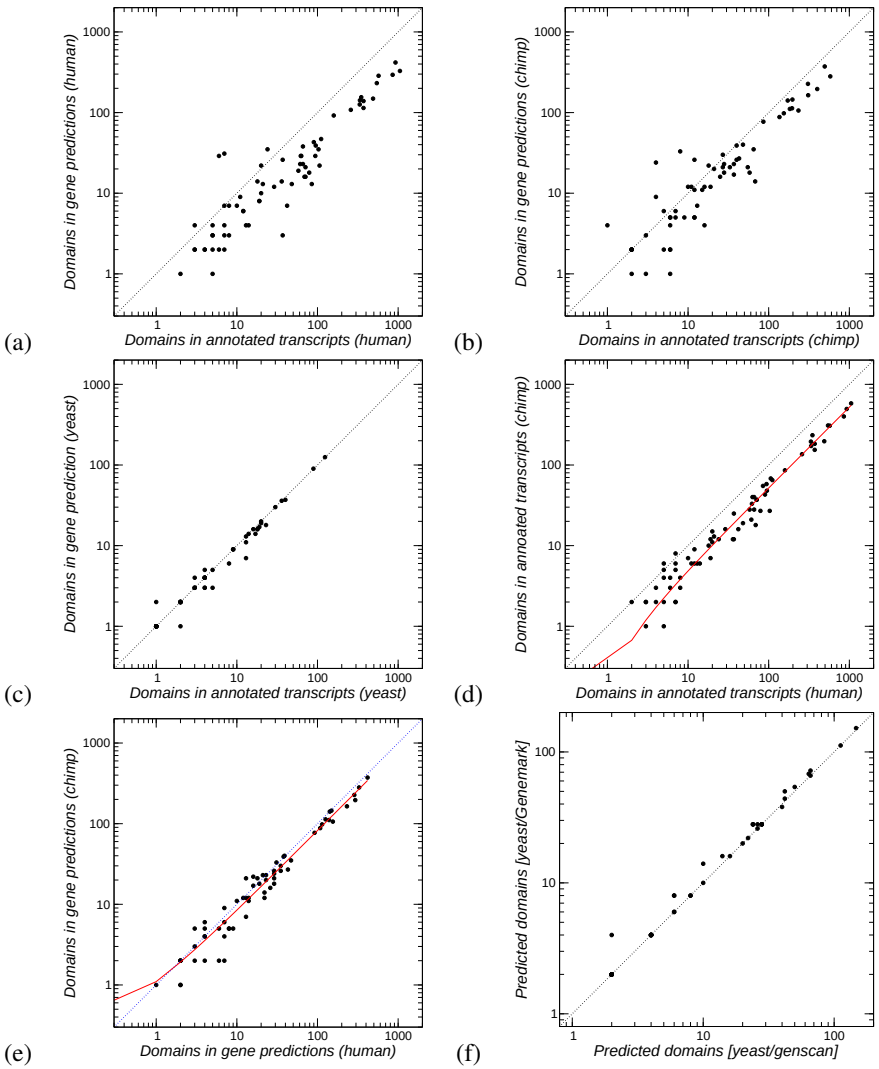

Figure 1: Correlation of the number of protein domains. Top row: Annotated transcripts compared to *de novo* predicted "genes" for (a) human, (b) chimp, and (c) yeast. Below: While domain prediction based on existing annotation yield systematic differences between human and chimp (d), congruent abundances are obtained from `genscan` predictions (e). Linear regression is shown as red line in panels (e) and (f). Different gene predictors (`genscan` and `GeneMark`) yield comparable results (f), shown here for yeast.

($p < 10^{-5}$). This indicates a selective pressure against their co-occurrences. We therefore also investigated two additional families of DNA binding domains, namely the leucine

Table 2: Domain occurrences and co-occurrences of zink-finger and winged-helix domains. The table shows the number of domains (Dom.), the number of "genes", i.e., `genscan` predictions that contain the domain (Genes), and for comparison the number of genes that contain the domain in SUPERFAMILY (SF). For species marked with *, multiple entries from different strains or variants in the SUPERFAMILY database exist, and SF values tend to over-count in these cases.

	ZNF [57667]			winged helix [46785]			co-occurrence		
Species	Dom.	Genes	SF	Dom.	Genes	SF	Dom.	Genes	SF
Giardia lamblia	7	6	4	16	13	11	0	0	0
Trichomonas vaginalis	23	14	9	100	98	89	0	0	0
Trypanosoma brucei	156	148	6	34	32	24	1	1	0
Leishmania major *	29	14	6	50	27	23	2	1	0
Naegleria gruberi	20	7	6	67	45	47	0	0	0
Plasmodium falciparum *	5	5	12	3	3	38	0	0	0
Tetrahymena	1	1	13	3	3	39	0	0	0
Thalassiosira pseudonana	15	11	8	145	138	130	0	0	0
Phytophthora ramorum	81	46	34	80	75	62	6	2	0
Clamydomonas	18	13	7	48	44	37	0	0	0
Arabidopsis thaliana *	151	115	74	186	168	241	0	0	0
Oryza sativa *	284	224	307	151	146	443	0	0	0
Dictyostelium	21	10	12	42	37	48	0	0	0
Aspergilus niger	64	51	34	68	65	47	0	0	0
Schizosaccaromyces pombe *	34	24	38	43	41	80	0	0	0
Caenoharbditis elegans *	58	27	144	15	14	165	0	0	0
Drosophila melanogaster *	853	301	322	126	122	152	0	0	0
Homo sapiens *	5090	1048	1324	274	256	414	0	0	0

zippers (SUPERFAMILY ID 57979) and the "high mobility group" (HMG) domains (SUPERFAMILY ID 47095). We again observe only very few candidate co-occurrences with other DNA binding domains in the species listed in Table 2 (our co-occurences between leucine-zipper and winged-helix and one between HMG and winged-helix). Inspection of these five cases revealed that four of them are clear artifacts of `genscan`, which predicts a fusion protein. The last candidate, human LARP1B, is predicted by `genscan` to have an additional internal exon containing a leucine-zipper domain. More likely, however, `genscan` stumbled across a retro-pseudogene deriving from FOSL1 located in an intron of LARP1B. Conversely, SUPERFAMILY, reports the co-occurrence of leucine-zipper and zink-finger in some isoforms of the paralogous human ATF2 and ATF7 genes, which are not found in our `genscan`-based approach.

We therefore conclude that the major types of DNA binding domains, and possibly other evolutionarily unrelated domains of similar function, strongly avoid each other in Eukarya. In contrast, domains with complementary functions readily co-occur with each other. A good example are zink-fingers and the "Küppel associated box" (KRAB) domain. The KRAB domain is a small (75 AA) protein domain [SUPERFAMILY ID 57667] that functions as a transcriptional repressor and is predicted to act via protein-protein interactions. It appears in a highly prolific family of evolutionarily very young transcription factors. Among the species listed in Table 2, it appears only in human. We detected 446 domains in 421 "genes", in agreement with the literature [NHZS10]. In contrast to the winged-helix domain, however, it readily combines with zink-finger domains: 351 `genscan` predictions (i.e., a third) of the 1048 ZNF proteins and 5/6 of the KRAB domain proteins belong to the KRAB-ZNF family, again in good agreement with the literature.

4 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. For example, the fraction of transcription factors increases approximately quadratically with the total number of genes in eubacteria [vN03]. A result like this requires an estimate of the total number of genes with reasonable reliability and accuracy. Similarly, investigations into lineage-specific variations of regulatory schemes require plausible statistics of protein domains and their combinations [PSK10]. For prokaryotes, this task is more or less solved by the common practice of annotating all open reading frames. The HMM models of protein domains are easily searched against the (translation of) these ORFs and included e.g. in the SUPERFAMILY database. False positives in the ORF annotation pose little problem since they are very unlikely to contain recognizable protein domains.

In Eukarya, however, the situation is different. Direct annotation of ORFs on the genome level does not work for most organisms since introns interrupt many domains. On the other hand, databases of experimentally determined transcripts are often subject to massive sampling biases. Here, we show that protein domains can be annotated with acceptable accuracy using *de novo* gene predictors such as `genscan`. This strategy also avoids methodological biases such as the enrichment of 3'-exons in poly-A ESTs.

We emphasize that it is impossible in practice to devise a fair benchmark for domain co-occurrence counts since the ground truth depends on the complete knowledge of all transcripts, even if one settles for the definition that two particular protein domains co-occur if they appear together in at least one protein-coding transcript. Therefore, we have to resort to comparing counts between closely related species for which we can plausibly expect to obtain similar numbers.

In easy cases, such as yeast, where the transcript structure is simple and data coverage is excellent, gene prediction and transcript annotation yield nearly identical results. For large mammalian genomes, on the other hand, estimates of domain numbers depend strongly on transcript coverage, while gene predictions yield numbers that are consistent among closely related species. Our investigation suggests that the biases and artifacts in the `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.

As an application of genome-wide domain counts, we investigated the co-occurrences of four major types of DNA binding domains (zink-fingers, leucine-zipper, HMG-box domains, and winged-helix domains). We found a strong and statistically highly significant anti-correlation of the four different domains. In constrast, evolutionarily related DNA binding domains readily co-occur in DNA binding proteins. It will be interesting to investigate whether a similar avoidance can be observed among other evolutionarily unrelated

protein domains that share a common molecular function.

Acknowledgment. AAP is supported by a DAAD fellowship. We thank Christian Arnold for helpful comments and for proofreading the manuscript.

References

[AHC+08] Antonina Andreeva, Dave Howorth, John-Marc Chandonia, Steven E. Brenner, Tim J. P. Hubbard, Cyrus Chothia und Alexey G. Murzin. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, 36:D419–D425, 2008.

[BBHS10] E Bornberg-Bauer, A K Huylmans und T. Sikosek. How do new proteins arise? *Curr Opin Struct Biol.*, 20:390â396, 2010.

[BK97] C. Burge und S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94, 1997.

[BK98] C. B. Burge und S. Karlin. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.*, 8:346–354, 1998.

[BPMS09] A Bhasi, P Philip, V Manikandan und P. Senapathy. `ExDom`: an integrated database for comparative analysis of the exon-intron structures of protein domains in eukaryotes. *Nucleic Acids Res.*, 37:D703–D711, 2009.

[CR07] Kevin Chen und Nikolaus Rajewsky. The Evolution of gene regulation by Transcription Factors and microRNAs. *Nature Genetics*, 8:93–103, 2007.

[DEKM98] R. Durbin, Sean Eddy, Anders Krogh und G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[Edd96] S R Eddy. Hidden Markov models. *Curr Opin Struct Biol*, 6:361–365, 1996.

[Edd98] S R Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998.

[FMSB+06] R D Finn, J Mistry, B Schuster-Böckler, S Griffiths-Jones, V Hollich, T Lassmann, S Moxon, M Marshall, A Khanna, R Durbin, S R Eddy, E L Sonnhammer und A. Bateman. `Pfam`: clans, web tools and services. *Nucleic Acids Res.*, 34:D247–D251, 2006.

[GN08] Y Go und Y. Niimura. Similar numbers but different repertoires of olfactory receptor genes in humans and chimpanzees. *Mol Biol Evol.*, 25:1897–1907, 2008.

[INK+07] Masumi Itoh, Jose C Nacher, Kei-ichi Kuma, Susumu Goto und Minoru Kanehisa. Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. *Genome Biol.*, 8:R121, 2007.

[KAK00] E Koonin, L Aravind und A Kondrashov. The impact of comparative genomics on our understanding of evolution. *Cell*, 101:573–576, 2000.

[Kor04] I. Korf. Gene finding in novel genomes. *BMC Bioinformatics*, 5:59, 2004.

[LTHCM05] A. Lomsadze, V. Ter-Hovhannisyan, Y. Chernoff und Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.*, 33:6494–6506, 2005.

[LWWG05] Mingyi Liu, Heiko Walch, Shaoping Wu und Andrei Grigoriev. Significant expansion of exon-bordering protein domains during animal proteome evolution. *Nucleic Acids Res.*, 33:95–105, 2005.

[MBE+08] Andrew D. Moore, Åsa K. Björklund, Diana Ekman, Erich Bornberg-Bauer und Arne Elofsson. Arrangements in the modular evolution of proteins. *Trends Biochem. Sci.*, 33:444–451, 2008.

[MCVdRFM+10] Santiago Martínez-Calvillo, Juan C. Vizuet-de Rueda, Luis E. Florencio-Martínez, Rebeca G. Manning-Cela und Elisa E. Figueroa-Angulo. Gene Expression in Trypanosomatid Parasites. *J. Biomed. Biotech.*, 2010. doi:10.1155/2010/525241.

[MDB08] Dennis Murphy, Barry Dancis und James R. Brown. The evolution of core proteins involved in microRNA biogenesis. *BMC Evolutionary Biology*, 8:92, 2008.

[MMNH04] Webb Miller, Kateryna D Makova, A Nekrutenko und Ross C Hardison. Comparative Genomics. *Ann. Rev. Genomics Hum. Genet.*, 5:15–56, 2004.

[NHZS10] K Nowick, A T Hamilton, H Zhang und L Stubbs. Rapid sequence and expression divergence suggests selection for novel function in primate-specific KRAB-ZNF genes. *Mol Biol Evol.*, 2010. doi:10.1093/molbev/msq157.

[Nii09] Y. Niimura. Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents. *Hum. Genomics*, 4:107–118, 2009.

[PSK10] Sonja J Prohaska, Peter F. Stadler und David C. Krakauer. Innovation in Gene Regulation: The Case of Chromatin Computation. *J. Theor. Biol.*, 265:27–44, 2010.

[SeoopstfDbd08] Structures und evolutionary origins of plant-specific transcription factor DNA-binding domains. Yamasaki, Kazuhiko and Kigawa, Takanori and Inoue, Makoto and Watanabe, Satoru and Tateno, Masaru and Seki, Motoaki and Shinozaki, Kazuo and Yokoyama, Shigeyuki. *Plant Physiol. Biochem.*, 46:394–401, 2008.

[TGMB09] Nicolas Terrapon, Olivier Gascuel Gascuel, Éric Maréchal und Laurent Bréhélin. Detection of new protein domains using co-occurrence: application to *Plasmodium falciparum*. *Bioinformatics*, 25:3077–3083, 2009.

[vN03] Erik van Nimwegen. Scaling laws in the functional content of genomes. *Trends Genetics*, 19:479–484, 2003.

[WA05] Stefan Wuchty und Eivind Almaas. Evolutionary cores of domain co-occurence networks. *BMC Evol. Biol.*, 5:24, 2005.

[WPZ+09] D Wilson, R Pethica, Y Zhou, C Talbot, C Vogel, M Madera, C Chothia und J. Gough. SUPERFAMILY — Comparative Genomics, Datamining and Sophisticated Visualisation. *Nucleic Acids Res.*, 37:D380–D386, 2009.