Submitted to *Preprint*. Pages 1 - 20.

OPEN ACCESS

Article

Pitfalls of Ascertainment Biases in Genome Annotations — Computing Comparable Protein Domain Distributions in Eukarya

Arli A. Parikesit [1–3,9], Lydia Steiner [1,2], Peter F. Stadler [2–8] and Sonja J. Prohaska [1,2]*

[1] Computational EvoDevo Group, Department of Computer Science, University of Leipzig, Härtelstraße 16–18, D-04107 Leipzig, Germany; E-Mail: {arli,lydia,sonja}@bioinf.uni-leipzig.de

[2] Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16–18, D-04107 Leipzig, Germany; E-Mail: studla@bioinf.uni-leipzig.de

[3] Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany

[4] Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstraße 16–18, D-04107 Leipzig, Germany

[5] Fraunhofer Institut für Zelltherapie und Immunologie—IZI Perlickstraße 1, D-04103 Leipzig, Germany

[6] Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria

[7] Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark

[8] Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

[9] Bioinformatics Group, Department of Chemistry, Faculty of Mathematics and Science, University of Indonesia, Depok 16424, Indonesia

* Author to whom correspondence should be addressed; E-Mail: sonja@bioinf.uni-leipzig.de;
Tel.: +49-341-97-166703.

Version December 16, 2013 submitted to Computation. *Typeset by LaTeX using class file mdpi.cls*

Abstract: Most investigations into the large-scale patterns of protein evolution are based on gene annotations that have been compiled in reference databases. The use of these resources for quantitative comparisons, however, is complicated by sometimes vast differences in coverage. More importantly, however, we also observe substantial ascertainment biases that cannot be removed by simple normalization procedures. A striking example is provided by the correlations between protein domains. We observe that statistics derived from different computational gene annotation procedure show dramatic discrepancies, and even qualitative changes from negative to positive correlation, when compared to statistics obtained from annotation databases.

Keywords: protein domains; HMM models; GO classification; functional genome annotation; Eukarya

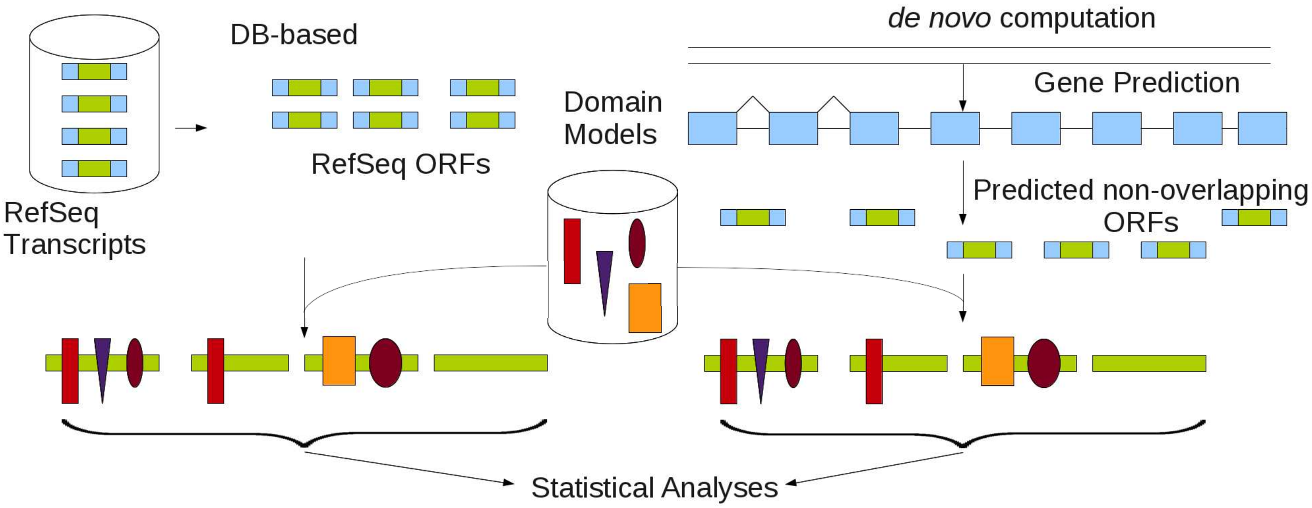
1. Introduction

Most proteins are composed of smaller building blocks. These *protein domains* typically form compact three-dimensional structures that are stable and foldable on their own. Domains convey a specific molecular function such as a particular catalytic activity or binding specificity. Protein function can indeed be inferred from domain content [1]. Protein domains are not only functional units but also constitute fundamental building blocks for protein evolution. They can be readily recombined in different arrangements leading to proteins that utilize different combinations of the same (types of) molecular interactions to fulfill different higher-level functions [2–4]. Over very large evolutionarily time scales, such as those of interest in a comparative analysis of the eukaryotic kingdom, protein domains are rearranged by fusions, fissions, and terminal loss so that larger proteins are a composite of domains deriving from several ancestral sources [5,6]. The reshuffling of domains is indeed much more frequent than the innovation of novel protein domains [7,8].

It becomes impossible at large time-scales to identify orthologous proteins [9]. One reason is that sequence similarity can degrade beyond recognition. More fundamentally, however, the re-shuffling of protein domains breaks the very concept of orthology as novel proteins arise through fusion and fission from multiple ancestors. The abundance and co-occurrence of protein domains thus becomes the most natural and promising framework to understand patterns of protein evolution, see e.g. [10–12]. In [8], for instance, it is shown that frequent gains and losses of domains lead to significant differences in functional profiles of major eukaryotic clades. Their results argue for a complex last eukaryotic common ancestor and reveal that animals are gaining increased regulatory complexity at the expense of their metabolic capabilities. Similarly, the rise of chromatin-based regulation mechanisms in crown-group eukaryotes can be traced by considering abundances and co-occurrences of the relevant protein domains [11]. A growing core of combinations in multicellular organisms was demonstrated by network analysis of domain co-occurrences [13]. The fundamental role of domains is also emphasized by the emergence of “supra-repeats”, i.e., complex and multi-domain repetitive patterns, which is of importance e.g. in nucleic acid-binding and protein-protein interaction and hence plays a key role in gene regulation [14]. The diversification of such regulatory proteins, thus translates primarily into statistically unexpected co-occurences of protein domains of the same functional type or structural family. A case in point is e.g. the rapid expansion of KRAB/zincfinger transcription factors in primates [15].

Protein domains are characterized by local amino-acid patterns and hence can be annotated computationally in protein sequences. Several databases, most notably the *SUPERFAMILY* [16] compile domain annotations for the known and predicted proteins of a wide variety of species. As we shall

Figure 1. Work flows for the estimation of domain abundance data from annotation data (l.h.s.) and starting with a *de novo* gene annotation (r.h.s). In both scenarios, protein domains compiled in databases such as *Pfam* or *SUPERFAMILY* are mapped to the known or predicted proteins and form the basis for subsequent statistical analysis.



demonstrate in this contribution, quantitative comparisons between distant species one the basis of this data, however, are plagued by large differences and biases in coverage. In principle, the most complete information about the protein complement can be inferred from the genome sequence. We therefore suggest a workflow centered around *de novo* gene predictions to obtain quantitatively comparable estimates, see Fig. 1. We observe, however, that different gene predictors still lead to qualitatively different results.

This contribution is organized as follows: In the following section we briefly outline the methods and data employed for gene prediction, protein domain annotation, and statistical evaluation. In section 3.1 we discuss differences in coverage and outline other potential sources of biases. We then proceed to show that the discrepancies between curated annotation and two distinct gene prediction tools can be large and vary substantially between genomes. In section 3.4 we investigate correlations of functionally defined classes of domains. Here the various sources of biases conspire to produce even qualitatively different results. We finally discuss the resulting limitations of the accuracy of the domain distribution data and the reliability of conclusions drawn from them.

2. Material and Methods

2.1. Sequence Data

We consider the following 18 species with sequenced genomes covering the entire phylogenetic range of the eukaryotes: H.sa: *Homo sapiens* hg19, D.me: *Drosophila melanogaster* BDGP5.13, C.el: *Caenorhabditis elegans* WS200, S.po: *Schizosaccharomyces pombe* EF1, A.ni: *Aspergillus niger* CADRE, D.di: *Arabidopsis thaliana* TAIR9.55, C.re: *Clamydomonas* Chlre4, T.th: *Tetrahymena thermophila* tta1_oct2008, P.fa: *Plasmodium falciparum* PlasmoDB-7.0, L.ma:

Leishmania major Lmj_20070731_V5.2, G.la: *Giardia lamblia* WBC6, T.va: *Trichomonas vaginalis* TrichDB-1.2, T.br: *Trypanosoma brucei* Tb927_May08_v4, N.gr: *Naegleria gruberi* Naegr1, T.ps: *Thalassiosira pseudonana* Thaps3, P.ra: *Phytophthora ramorum* Phyra1_1, O.sa: *Oryza sativa* OSV6.1, D.di: *Dictyostelium discoideum* DDB. Sources are listed in the Supplement http://www.bioinf.uni-leipzig.de/supplements/12-007.

2.2. Gene Prediction

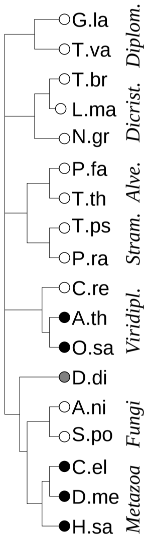
Gene prediction tools come in two flavors: homology-based and *de novo* approaches. Since homology-based methods necessarily transfer any biases from the source annotation to the newly annotated target, only *de novo* methods are applicable for our purposes. A software package that was widely used in early genome projects is GENSCAN [17,18]. It is mostly geared towards vertebrate genomic sequences. Despite the shortcomings of such a "first generation" gene predictor, good performance has been observed in the literture for a diverse set of only distantly related organism (teleost, nematodes,amphioxus, and fungi) without the need for specific training [19] and it has hence served as the most popular tool until a few years ago [20]. Furthermore, it is much less demanding computational resources. Assuming that is samples in an unbiased manner, a moderate loss in sensitivity would be acceptable for the task at hand: after all we are more interested here in statistical associations than a complete annotation. Moreover, GENSCAN it has been used extensively in the ENSMBL genome database project [21] and the fugu genome annotations are still largely based on this tool [22].

Following [23] we split long chromosomes into overlapping fragments of about 500 kb to accommodate the tool's restriction on input length. Protein sequences were extracted directly from the GENSCAN predictions. Duplicate predictions in the overlaps between fragments were removed. A potential shortcoming is that the prediction accuracy may vary substantially with differences in gene architecture.

We therefore employed a trainable gene prediction as an alternative. In this class of tools, the statistical model is trained with a set of known genes from the organism that is to be annotated. We chose AUGUSTUS [24–26] as a representative of trainable tools because it has gained popularity in recent genome annotation projects. Both "Specific" (local) and "Default" (web-based) trained models are used here. We used the tools as described in the AUGUSTUS tutorial [27]. Where available, we made use of the default training sets provided at the AUGUSTUS website. For the remaining species, we used the cDNAs available in GenBank. Redundancies were removed with a dedicated perl script. The FASTA sequences and their headers were cleaned from meta-characters and gaps. Models were trained in "Specific" mode with the pipeline downloaded from the the AUGUSTUS website. For our applications, AUGUSTUS was configured to generate only non-overlapping protein-coding genes. The predicted protein sequences are part of the AUGUSTUS output. We verified with bed-tools that no overlapping sequences were contained in the output [28].

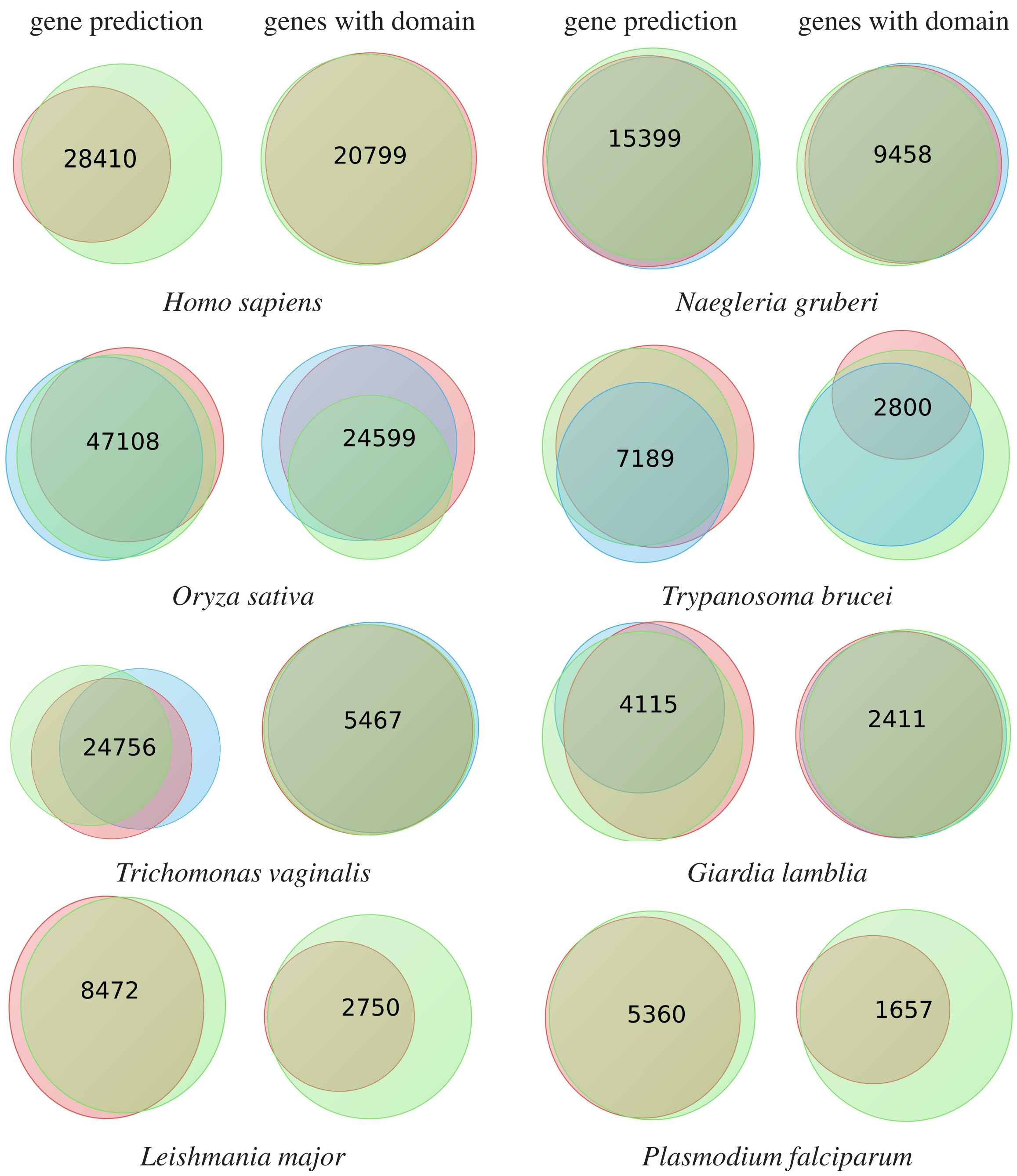
The results of the AUGUSTUS and GENSCAN predictions are compiled in Table 1 and compared with the RefSeq (release 53) genes for each of the 18 species. In order to compare the two training modes of AUGUSTUS with each other and with the RefSeq annotation we computed their overlaps with

Table 1. Summary of gene and domain annotations. The left part of the table summarized the results of the gene predictors. For the species with the equal result of both 'Default' and 'Specific' training methods, the entries are shown only once. In the right part of the table, we list the number of genes that contain at least one recognizable protein domain from the *SUPERFAMILY* and *Pfam* collections, respectively. The phylogenetic distribution [29] of the 18 investigated species is shown on the left margin.



Species	Gene Total				Gene∧*SUPERFAMILY*				Gene∧*Pfam*			
	AUGUSTUS		GENSCAN	RefSeq	AUGUSTUS		GENSCAN	RefSeq	AUGUSTUS		GENSCAN	RefSeq
	Default	Specific			Default	Specific			Default	Specific		
Giardia	4357	5178	11251	6583	3240	3265	2567	3183	2450	2672	2846	2540
Trichomonas	61750	60924	19251	60815	3278	3344	17627	6392	5872	5478	18342	28364
Trypanosoma	7874	9696	5143	10192	4626	4626	3035	4010	4939	5580	4537	2800
Leishmania	9451		4560	9155	4949		3758	4056	4451		4156	2762
Naegleria	16792	16443	10748	16620	6572	6442	6798	7091	10070	9578	8653	10070
Plasmodium	6043		1439	5512	4110		1109	2607	3338		1276	1741
Tetrahymena	21650		2011	24725	2502		1856	1003	1763		1956	952
Thalassiosira	10428	10528	8766	10988	6248	6145	5200	6264	6752	7141	6743	7500
Phytophthora	17154	16292	16701	15743	7384	7382	9159	7394	10524	10746	10478	10663
Chlamydomonas	15141		13268	14488	6852		7062	6749	9193		9632	8472
Arabidopsis	27945		20135	25498	8088		11957	9302	22521		15264	22716
Oryza	62327	63693	64109	62709	8580	7527	23659	8417	44243	45322	32678	42523
Dictyostelium	12904	12595	5323	12646	6877	6744	3468	5246	7757	7403	4945	4018
Aspergillus	9866		8112	10785	6432		5467	6275	7827		6753	6815
Schizosaccharomyces	4783		3578	4824	4259		2532	3204	4305		2834	4405
Caenorhabditis	22902		12432	21175	7418		4329	8806	14460		5378	17253
Drosophila	14217		28889	13601	7654		11618	8925	10550		13424	10283
Homo	33507		118894	36073	8908		31359	10069	20878		34283	27577

Figure 2. Comparison of gene predictions for 8 of the 18 species. (See online supplement for additional data). For each species we show a Venn diagram, drawn to scale, for both the raw output of the gene predictions and for the subset of proteins with at least one matching *Pfam* model. `RefSeq` is shown in red `AUGUSTUS` prediction with Default and Specific trained models are shown in blue and green, respectively. The numbers refer to the overlapping genes.



`bed-tools` and used `lucidchart` to compute Venn diagrams so that the displayed overlaps in Fig. 2 are drawn to scale [30].

2.3. Domain Annotation

We used the entire *Pfam* version 26.0 database, comprising 33672 domain models as well as the entire collection of 9821 Hidden Markov Models (HMMs) provided by the *SUPERFAMILY* database (version 1.75). In both cases we used `HMMER3.0rc1` [31] with an E-value threshold of $E \leq 10^{-3}$ to map the HMMs to the predicted amino acid sequences as well as the `RefSeq` proteins.

In order to test the quality of gene predictions we compared the sub-collections of protein sequences with at least one mapped *Pfam* domain between the gene prediction methods and `RefSeq` database. A representative selection of these results is shown in Fig. 2. Overall, the specific-trained `AUGUSTUS` predictions have the best coverage of the manually curated `RefSeq` and are hence used as data basis for subsequent quantitative analysis.

2.4. Functional Classification

The domain databases contain thousands of distinct domain models. Few domains thus appear with sufficient frequency in a single genome to allow a quantitative statistical analysis. Thus we pooled the occurrence data by larger functional categories. The *SUPERFAMILY* database offers a "Structural Domain Functional Ontology" providing functional and phenotypic annotations of protein domains at the ***superfamily*** and ***family*** levels [16]. The *Pfam* annotation is already integrated into GO database, providing a mapping from *Pfam* domains to GO ontology terms [32,33].

As example we use here the same high-level functional categories as in previous work [23].

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.
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 four functional groups bN, bP, rC, and rB encapsulate major modes of regulation. Both bN and bP play an important role for gene regulation by transcription factors and are among the most abundant GO classes, while rC focuses on chromatin-based epigenetic regulation. We have shown in [23] that rC group correlates well with the hand-picked collection of domain models that can act as readers, writers, and erasers of histone modification [11]. The domain groups rE and mS were intended as a form of controls that *a priori* we did not expect to correlate in a particular way with either nucleic acid or protein binding domains (bN, bP).

From the co-occurrences of domains in predicted proteins and the map of domains to functional (GO-)classes it is straightforward to obtain the number $n(C, D)$ of co-occurrences of the C and D functional classes. As in [23] we correct $n(C, D)$ for the fact that the same domain x can be a member of both C and D by counting these cases with a weight of 1/2.

2.5. 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 the domain x and y co-occur. Here x and y can be either individual domains, sets of domains belonging to the same superfamily, or the collections of domains compiled into functional classes according to their GO annotations. Denote by n_x the total number of annotated domains belonging to group x. 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 of the genome under consideration. As discussed in [23] this estimate does not account for biases arising from the non-uniform distribution of domains over genes. Let $n_d(i)$ be the number of domains predicted for protein i, and let $n_d = \sum_i n_d(i)$ be the total number of domains. Then the number of x-domains that occur in genes that also contain a y-domain can be estimated as

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

where the sum runs over all genes i that contain a domain belonging to group y. We obtain an alternative estimate by exchanging x and y in equ.(1).

We compared these expectations with the number of empirically observed co-occurrences $n(x, y)$. We speak of *co-occurrence* of domain families or groups x and y if $n(x, y) \gg \max\{E(x|y), E(y|x)\}$ and of *avoidance* if $n(x, y) \ll \min\{E(x|y), E(y|x)\}$ The statistical significance of an observed difference between $n(x, y)$ and the values of $\min\{E(x|y), E(y|x)\}$ and $\max\{E(x|y), E(y|x)\}$, respectively, is determined observing the fact that the count data $n(x, y)$ follow a Poisson distribution.

Note that the use of the $\min$ and $\max$ here ensures that the tests for avoidance and co-occurrence are conservative. In order to verify that this test is indeed not prone to detecting false positive associations but rather fails on the side of false negatives, we simulated 30 random data sets by repeatedly permuting the functional annotation of the individual domains using the `linux` command `shuf`. No result significant at the 10% level was returned for both the `SUPERFAMILY` and `Pfam` annotations (compared to an expected 3 false positives), reassuring us that the test statistic works conservatively as designed.

3. Results and Discussion

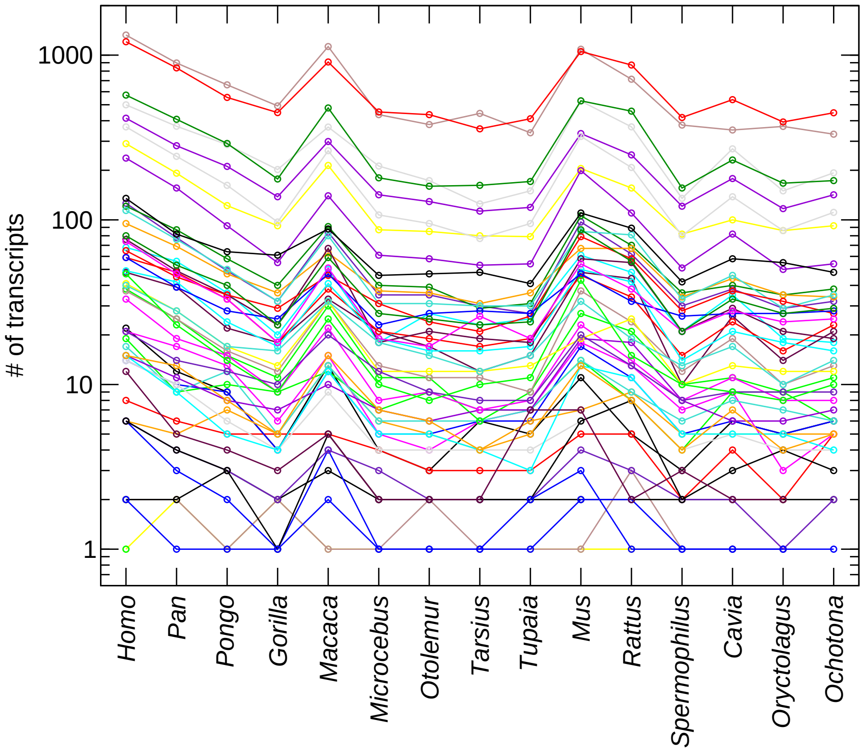
We compare the annotation data and *de novo* gene prediction at different levels: (1) w.r.t. the set of predicted transcript, w.r.t. the collection of predicted proteins that contain recognizable domains, and (3) w.r.t. to the statistics of co-occurring domains in the same annotated protein. We will see in the following, that at each level we observe large discrepancies between the data sets.

3.1. Ascertainment biases in protein annotations

In order to get an impression of the differences in coverage of protein annotation we compared the number of transcripts annotated in the genomes of Euarchontoglires, the mammalian clade comprising primates and rodents. These genomes are organized in a very similar way and are expected to have very similar or even nearly identical complements of protein coding genes. Figure 3 shows these data for `RefSeq` (version 42) subdivided as numbers of transcripts containing a particular `SUPERFAMILY` domain.

Closely related species, such as the mammals, do not differ much in their gene content. Even if they do, differences tend to be confined to a few gene families with large copy numbers that evolve very quickly, such as olfactory receptors [34]. Even in gene families with rapid losses and gains, the total

Figure 3. Differences in `RefSeq` coverage of primate and rodent genomes. Each line corresponds to number of annotated protein-coding transcripts that contain a particular class of protein domains.



numbers of family members do not change dramatically at moderate time scales. In contrast to biological wisdom, the numbers of annotated transcript, Fig. 3 vary by nearly an order of magnitude between closely related species. Instead of biological differences we clearly observe here only variations in annotation coverage. Indeed, the best-studied model species (*Homo sapiens*, *Macaca mulatta*, *Mus musculus*, and *Rattus norvegicus*) show much larger transcript numbers than the other, less well studied genomes. The lines are approximately parallel in the logarithmic plot, indicating that the relative abundances of the domains are similar and a simple difference in the completeness of the annotation can account for much of the variation.

The annotation of `RefSeq` genomes is complex composite of of experimental transcript and protein information and computational procedures including alignments as well as HMM-based *ab initio* predictions [35] complemented by manual curations. Since an independent gold standard is absent, it cannot be rigorously checked to what extent these annotations are biased. The focus on known protein-coding genes, however, that are used a starting point for the homology-based annotation steps, at least suggests the well-studied protein families are overrepresented. Differences in coverage, thus, may not be the only artefacts that need to accounted for when `RefSeq` annotation is used for statistical purposes.

As shown in [36] the use of a *de novo* gene predictor such as `GENSCAN` can alleviate the differences in coverage and provide quantitatively comparable domain annotations in closely related species. The situation is more complex, however, when a comparison of gene or protein complements for organisms from different kingdoms is required. Additional biases may arise from both the protein and the domain annotation. Homology-based annotation may be confounded by differences in sequence conservation between proteins of different functional classes. Such differences affect not only the overall rate of subsitution but also the substitution patterns [37]. The sensitivity of annotation procedure thus may show systematic dependencies on functional classes or the size of gene families. Our knowledge of protein domains is by far not complete. Although most protein domains in well-studied model organisms are evolutionarily very old and rather well conserved, the characteristic patterns of the domains slowly evolve and domain innovation is a relatively infrequent but well-documented phenomenon [7,8]. Thus, the sensitivity of the domain annotation procedure must be expected to decrease with increasing evolutionary distance from the examples that have been used to create the domain models. An important technical issue plagues in particular transmembrane regions and signal peptides. These have a hydrophobic bias leading to problematic domain models and subsequently to incorrect function assignments inherited from these domain models [38].

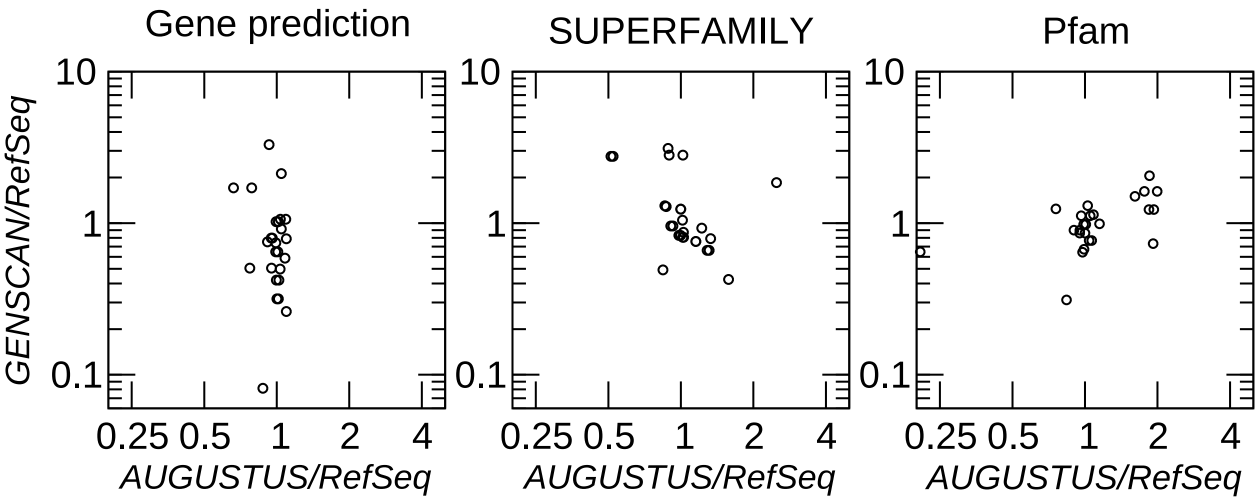
3.2. Biases in gene predictions

GENSCAN is a general purpose gene prediction tool that does not admit species specific training. Its internal model implicitly assumes a gene organization similar to that of higher plants and animals. Several lineages of the Eukarya, however, feature gene structures and a genomic organization that is very different from the situation in animals, fungi, or plants. Both *Giardia lamblia* and *Trichomonas vaginalis* are extremely intron-poor; *Trichomonas vaginalis* in addition features very large numbers of paralogs. Kinetoplastids (*Trypanosoma* and *Leishmania*) produce large polycistronic transcripts from which individual mature mRNAs are produced by trans-splicing, cis-splicing, and polyadenylation [39,40]. Trans-splicing is also prevalent in the nematodes, but absent from most other animal genomes. GENSCAN in particular has problems to detect the gene boundaries in the polycistronic transcripts. Intron-sizes differ dramatically between invertebrates and vertebrates, where intron-sizes of more than 10 kb are not at all uncommon. Another problem is posed by the extreme sequence composition as in the AT-rich genome of *Plasmodium falciparum* [41].

We therefore employed AUGUSTUS as a trainable gene predictor to test whether differences in sensitivity between species due to differences in gene structure might play a large role. A comparison of AUGUSTUS and GENSCAN predictions in Table 1 confirms our suspicion that gene predictors are also a source of biases since the numbers are not only significantly different but also deviate substantially from proportionality. While GENSCAN predicts more twice as many transcripts than AUGUSTUS in *Giardia*, *Homo*, and *Drosophila*, we find the opposite picture in *Caenorhabditis*, *Dictyostellium*, *Tetrahymena*, *Plasmodium*, *Leishmania* and *Trichomonas*.

A comparison of AUGUSTUS predictions with the RefSeq gene inventories agrees rather well in some species, while in others there are substantial differences, Fig. 2. Large discrepancies seem to be related to the degree of completeness of the gene annotation. Table 1 shows that GENSCAN predicts more than twice as many genes than RefSeq in human and *Drosophila*, while RefSeq has the more inclusive annotation in those species where AUGUSTUS is more sensitive than GENSCAN. Overall, AUGUSTUS conforms better to the RefSeq annotation than GENSCAN. The similarity of the results obtained with RefSeq and AUGUSTUS might be explained in part be the use of similar HMM-based

Figure 4. Discrepancies between *de novo* gene predictions and RefSeq annotations.



annotation methods in RefSeq [35]. Strictly speaking, therefore, one cannot conclude that AUGUSTUS performs "better" than GENSCAN because its output is more similar to RefSeq, although this is a plausible hypothesis.

3.3. Gene predictions with domains annotations

Since we are interested primarily in the distributions of protein domains we also compared RefSeq data with gene predictions restricted to only those genes in which at least one *SUPERFAMILY* or *Pfam* domain was annotated. For most species this improves the congruence between the gene sets. In a few cases, however, the differences persist, as in the case of *Trypanosoma* and human, Fig. 2. In *Trypanosoma*, most of the difference is explained by annotated RefSeq proteins without recognizable domains. In human, the discrepancy is in part explained by RefSeq isoforms and in part by AUGUSTUS prediction without domains.

Figure 4 shows that significant differences between gene predictions and RefSeq persist even when the data are restricted to predicted transcript with annotated domains. Note, furthermore, that annotation of the same gene predictions with domains from SUPERFAMILY and Pfam also yield substantially different results: in fact, the point clouds show very little correlation.

Among the predictions with annotated domains, we find e.g. for *Leishmania*, *Tetrahymena*, and *Plasmodium* that both the default and the specific trained gene predictions have a much larger coverage than the RefSeq data. For *Trichomonas* and *Giardia*, the situation is reversed. This can probably be explained in part by the large number of paralogs and possible pseudogenes included in RefSeq in *Trichomonas*, but also indicated as lack of sensitivity of the gene predictor for the two parabasalids with their extremely intron-poor genomes. At the domain level, AUGUSTUS and RefSeq agree nearly perfectly e.g. in human and *Naegleria*. In general, the RefSeq entries missed by the gene predictor are frequently putative pseudogenes and ORFs lacking further annotation. Since the AUGUSTUS 'specific' predictions overall yield the most inclusive data set, these predictions are used in the following section for all statistical analyses of domains compositions.

Surprisingly, we observe very little variation in the number of domains per protein. A significant increase is found in human and fruitfly only. It is unclear, however, whether this a true effect or an artifact arising from a bias in *Pfam* database. In [42], a difference in the complexity of chromatin proteins between Diplomonads and Dicristates on the one hand, and Alveolates and Stramenopiles on the other hand does exist. Our data do not show such a systematic difference for proteins containing domain classified as "rC: regulator of chromatin" according to GO.

3.4. Correlations in domain occurrences

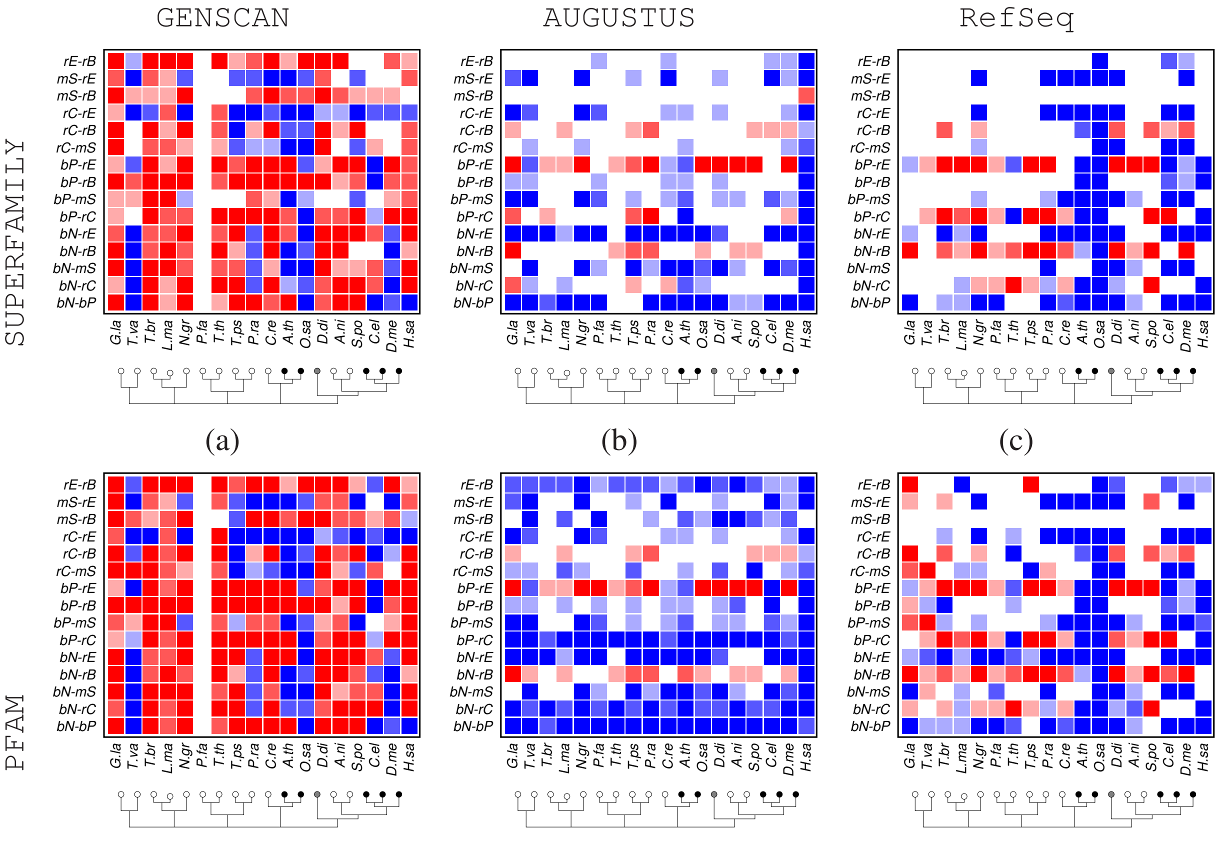
In [36] we observed regularities in co-occurrences of domains in the same protein. For instance, transcription factors often contain multiple DNA binding domains. In the human genome, these domains predominantly belong to the same (very frequent) domain class such as zinc fingers or winged-helix domains. Combinations of different domain classes, on the other hand, are observed much less frequently than predicted from a random background model. Since most domain families have only a moderate number of occurrences in most genomes this phenomenon of "domain avoidance" cannot be statistically supported for most pairs of domain types even in the large mammalian genomes. In a subsequent study [36] we therefore aggregated domain classes to groups defined by their biochemical function. This aggregation, however, is susceptible to yet another source of biases: different domains types now contribute to the functional classes in the vastly different organism studied here. There is no guarantee, of course, that coverage and quality of functional domain annotations, as provided by the `Pfam` and `SUPERFAMILY` databases are uniform across all classes of domains. The phylogenetic distribution of the individual domain families, on the other hand, is of course far from uniform.

A comparison of the panels of Figure 5 shows that the qualitative results on the domain co-occurrence are largely independent of the choice of `SUPERFAMILY` or `Pfam` as domain database. The main difference is the number of distinct domains in the two databases. Since `Pfam` is much larger, the absolute domain counts are larger and hence a larger number of the statistical tests reaches the required levels of significance.

On the other hand, there are striking differences between the gene annotation. Based on `GENSCAN` we predicted a general tendency towards domain co-ocurrences [23]. Using `RefSeq` annotation, it appears that avoidance dominates (in particular in multicellular animals and plants), while co-ocurrences appear be prevalent in unicellular protozoans. The `AUGUSTUS`-based gene predictions, however, point at a generic trend towards domain avoidance, with exceptions only for a few pairs of functional classes. The most prominent cases are bN-rB ('binding of nucleic acids' and 'regulation of binding'), bP-rE ('protein binding' and 'regulation of enzymatic activity'), and rC-rB ('regulation of chromatin' and 'regulation of binding'). For bP and rE a positive correlation is not unexpected, since regulators of enzymatic activity (rE) can be expected to act by protein-protein binding (bP). The positive correlations between nucleic acid binding domains (bN) and chromatin associated domains (rC) with domains involved in the regulation of binding deserved further investigation. It is consistent with intimate link of both DNA and RNA binding with chromatin regulation reported in [11].

In `AUGUSTUS` annotation, the domain co-occurrences of *Tetrahymena* in *SUPERFAMILY* and *Pfam* are scarce. This could be possibly due to the short scaffold in the *Tetrahymena* genome that could lead

Figure 5. Summary of co-occurrences patterns of major functional classes of protein domains across the Eukaryotes. The columns represent the gene prediction (GENSCAN and AUGUSTUS) with the DB-based RefSeq annotations. The rows represents the utilized protein domain databases (*PFAM* and *SUPERFAMILY*) for computing the domain co-occurrences, while their global patterns were computed for counting the average co-occurences. Blue rectangles indicate statistically significant avoidance between functional classes of protein domains, red indicates co-occurrence. The saturation of the color denotes the significance levels $p < 0.001$ (saturated color), $0.001 \leq p < 0.01$ (intermediate), and $0.01 \leq p < 0.1$ (pale). Entries that show neither avoidance or co-occurrence at a significance level of at least 10% remain white.



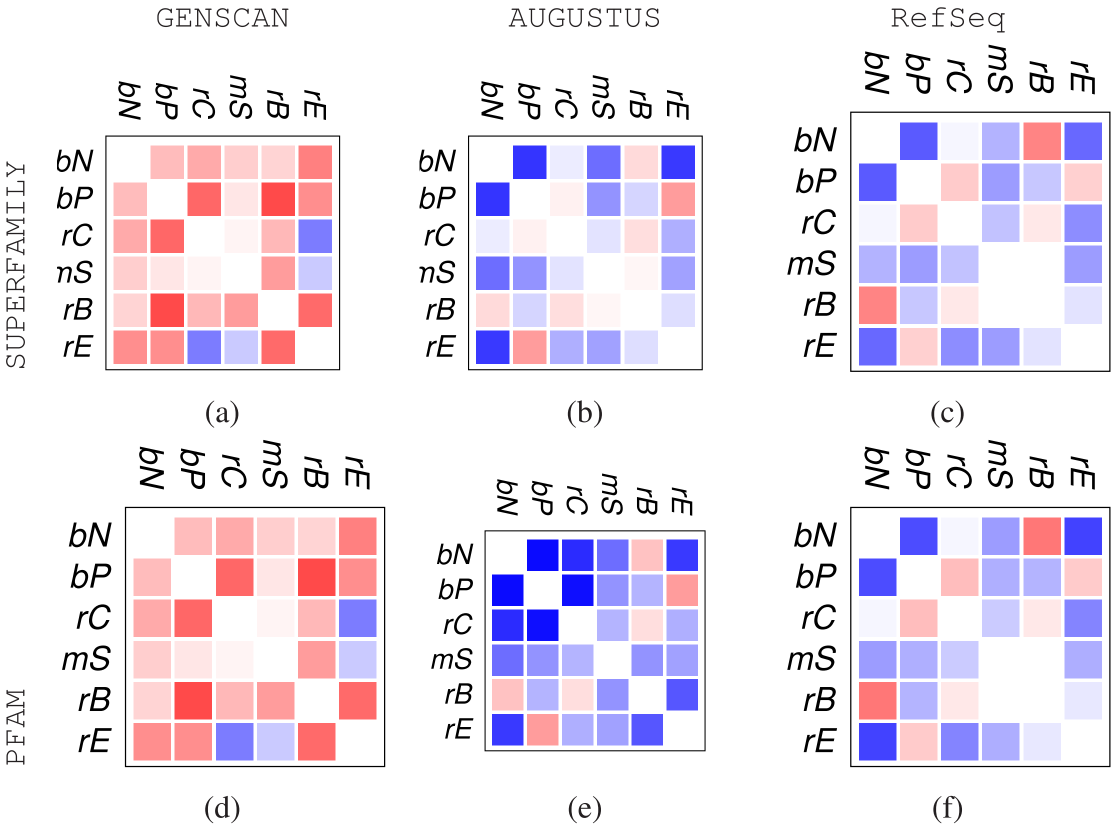
(a) (b) (c)

(d) (e) (f)

into underestimates [43]. Interestingly, in AUGUSTUS annotation we also observe intersting pattern in human genome. The strong co-occurrences of bN-rB, rC-rB, and bP-rE that exist in other genomes are missing, and replaced with a strong tendecy of avoidance. The possible explanation for the emergence of the avoidance pattern is the complex functionality diversification of the domains that tend not to co-occur in complex organisms. One other noticable pattern in human that the mS-rB co-occurrence pattern exists in *SUPERFAMILY* while it disappears completely in *Pfam*.

Figures 6 further aggregates these data and shows the overall tendency in the correlation between domains of different functional classes. The figure emphasizes the large, qualitative mutual differences between the gene prediction methods and the curated protein annotations. Nevertheless, some consistent patterns emerge, such as the positive correlation between protein binding proteins (bP) and regulators

Figure 6. Graphical summary of the correlations between the six functional domain classes computed with different protein annotations and domain databases.



(a) (b) (c)

(d) (e) (f)

of enzymatic activity (rE). This observation is not surprising although it does not derive from double memberships of domains in both groups. A less obvious signicant co-occurrence is that of nucleic acid binders (bN) with regulators of binding (rB), perhaps hinting a wide-spread involvement of transcription factors in regulatory processes that involve large protein complexes.

In the multi-cellular organisms with large genomes and large gene families, however, there is a strong signal of avoidance between several functional groups of protein domains. This may be a result of the expansion and diversication of large families of paralogous genes and their use for specific tasks in the regulation of cellular processes. In most cases, we have not been able to trace this effect to excessive duplication events in one or a few gene families, however. The most extreme example in our data is the expansion of the variant-specific surface protein (VSP) gene in *Giardia lamblia* [44], which fall into the bP (binding of proteins) and rC (regulation of chromatin) categories and features more than 200 paralogs. Still, the signal for co-occurrence of rC and bP is not consistently observed in all combinations gene finding and domain annotation. It appears, therefore, that the observed consistent patterns are not the consequence of an extreme expansion of a single gene family but must be related to the collective effect of gene families with the same functional domain classes.

The most rapidly expanding protein family in human, the KRAB/zinc finger proteins [15], one the other hand, contain only domains annotated as nucleic acid binding, and hence do not contribute to co-occurrence or avoidance patterns. This seems to be the case with many transcription factor families.

It is worth noting in this context that in earlier work we observed that the is also a strong pattern of avoidance among different SUPERFAMILY classes of nucleic acid binding domains [36].

4. Conclusion

Proteins embody a wide variety of functions in a cell, ranging from enzymatic activity to structural scaffolding. The function of a protein is reflected in its domain composition. The range of an organism's biochemical capabilities, both metabolic and regulatory, is thus largely encoded in its protein domain content. Even the presence of RNA-based modes of regulations such as the RNAi pathway are reflected by the associated protein components [45]. Large-scale trends in evolution such as an increased complexity of transcriptional regulation [46,47] or the diversification of chromatin modification throughout the eukaryotic kingdoms [11] thus can be traced by a quantitative comparison of protein and protein domain complements.

Present-day annotations for most genomes as well as the currently available collection of protein models are far from complete. Quantitative cross-species comparisons thus implicitly rely on the assumption that the available data are a fair, essentially unbiased sample. This is in general not the case. We therefore investigated to what extent *de novo* gene prediction methods can be used to generate comparable domain distribution data. The large differences between untrained (`GENSCAN`) and trained (`AUGUSTUS`) indicates that this not a straightforward endevour. Using the trained models we obtain results that in general are closer to annotation-based numbers. It is unclear, however, whether this speaks for a better quality of the computational results, or whether this simply reflects that `AUGUSTUS` has been trained with species-specific gene models using transcript data that also underlie the `RefSeq` annotation. Furthermore, gene predictors are usually tested and benchmarked against curated annotation such as `RefSeq`. Thus it may be the case that trained gene predictors and protein annotation databases only share the similar biases. However, it is reasonable to assume that the annotation of protein-coding exons is fairly complete at least in the human genome [48], suggesting that qualitatively similar results from `AUGUSTUS` and `RefSeq` are closer to the truth.

The second source of major ascertainment biases in the analysis of large scale evolutionary patterns of functional domains are the protein domain databases themselves. Recent studies reported the innovation of a large number of domain innovation events within both the green plants [49] and the animals [50]. The number of identified clade-specific domains must be expected to depend on the depths in which the clade is studied. The domain inventory is thus probably more complete in animals, fungi, and plants animals compared to most protozoan lineages.

We find that domain annotation data in particular in less well-annotated genomes may suffer from significant ascertainment biases that are uniform neither across functional classes of proteins nor across phylogeny. Data such as those in Fig. 4 indicate that errors can reach a factor of two (or even larger in extreme cases) in the number of counts cannot be rules out. This is large enough to seriously confound for instance the enrichment analyses ubiquitously used in many genomics publications. Nevertheless the reannotation of domains based on modern gene predictors instead of protein sequence databases appears at least to help reduce the ascertainment bias, although we cannot strictly rule out that manually curated data sets are strongly biased as well.

Large numbers unannotated domains, on the other hand, could further undermine the analysis presented here since the lead to a systematic under-estimation of an organisms metabolic or regulatory capability. Recent reports such as [49] seem to indicate that the innovation of novel domains is prevalent in particular in conjunction with stress response and developmental innovations, and hence a particular functional classes. A more systematic survey of so-far undescribed protein domains thus constitutes a natural next step towards a comprehensive understanding of functional evolution in the eukaryotes. Accurate domain inventories are not only of interest in their own right but also constitute an important source of phylogenetic information [51], in particular in "deep phylogeny" applications. The presence/absence patterns of protein domains were recently used for instance to place the Strepsiptera as a sister group of beetles in insect phylogeny [52]. Improved pipelines to estimate the protein domain content directly from genomic data thus have the potential to greatly facilitate phylogenomic investigations.

In the analysis of protein domain combinations at least two sources of biases interact: the dependence of the sensitivity of the protein annotation and an uneven coverage of protein domains that might reflect e.g. different levels of interest in different protein families. When domains distributions are summarized in terms of functional annotations, i.e., GO classes, the uneven interests in different protein families is again a likely source of ascertainment biases. The influence of the same or similar confounding factors in different data sources can conspire to produce even qualitatively different results such as the ones observed here for domain co-occurrence data in Fig. 5.

We conclude that ascertainment biases in current annotation databases as well as computational annotation tools fundamentally limit the accuracy with which domain distributions can be estimated. Statistical significance measures estimated via the commonly employed permutation tests often cannot account for these biases. Hence statistical significance may not not sufficient in such cases to make statements about biological reality. Large effects thus will in general be necessary to draw reliable conclusions from domain distribution data unless biases can be ruled out e.g. based on near completeness of data. We showed that at least some of the biases can be reduced by using predictions instead of manual curated annotation. Results show, however, that the prediction has to be choosen carefully so as to avoid introducing new biases.

Acknowledgment We thank Katharina Hoff (Greifswald) for technical assistance with the AUGUSTUS online gene prediction. A.A.P. was supported by the the DAAD. LS is supported by the MAGE project, BMBF grant number 50500541.

References

1. Forslund, K.; Sonnhammer, E.L.L. Predicting protein function from domain content. *Bioinformatics* **2008**, *24*, 1681–1687.
2. Apic, G.; Gough, J.; Teichmann, S.A. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* **2001**, *310*, 311–325.
3. Orengo, C.A.; Thornton, J.M. Protein families and their evolution – a structural perspective. *Annu Rev Biochem* **2005**, *74*, 867–900.

4. Buljan, M.; Bateman, A. The evolution of protein domain families. *Biochem Soc Trans* **2009**, *37*, 751–755.
5. Moore, A.D.; Björklund, Å.K.; Ekman, D.; Bornberg-Bauer, E.; Elofsson, A. Arrangements in the modular evolution of proteins. *Trends Biochem. Sci.* **2008**, *33*, 444–451.
6. Koonin, E.; Aravind, L.; Kondrashov, A. The impact of comparative genomics on our understanding of evolution. *Cell* **2000**, *101*, 573–576.
7. Bornberg-Bauer, E.; Huylmans, A.K.; Sikosek, T. How do new proteins arise? *Curr. Opin. Struct. Biol.* **2010**, *20*, 390–396.
8. Zmasek, C.M.; Godzik, A. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. *Genome Biol.* **2011**, *12*, R4.
9. Mahmood, K.; Webb, G.; Song, J.; Whisstock, J.; Konagurthu, A. Efficient large-scale Protein Sequence Comparison and Gene Matching to Identify Orthologs and Co-orthologs. *Nucleic Acids Res* **2012**, *40*, e44.
10. Kim, K.M.; Caetano-Anollés, G. The proteomic complexity and rise of the primordial ancestor of diversified life. *BMC Evol Biol* **2011**, *11*, 140.
11. Prohaska, S.J.; Stadler, P.F.; Krakauer, D.C. Innovation in Gene Regulation: The Case of Chromatin Computation. *J. Theor. Biol.* **2010**, *265*, 27–44.
12. Yang, S.; Bourne, P.E. The Evolutionary History of Protein Domains Viewed by Species Phylogeny. *PLoS ONE* **2009**, *4*, e8378.
13. Stefan, W.; Eivind, A. Evolutionary cores of domain co-occurence networks. *BMC Evol. Biol.* **2005**, *5*, 24.
14. Moore, A.D.; Bornberg-Bauer, E. Footprints of modular evolution in a dense taxonomic clade. GCB2012 proceedings; Schloss Dagstuhl: Jena, 2012; p. 5.
15. Nowick, K.; Hamilton, A.T.; Zhang, H.; Stubbs, L. Rapid sequence and expression divergence suggests selection for novel function in primate-specific KRAB-ZNF genes. *Mol Biol Evol.* **2010**, *27*, 2606–2617.
16. de Lima Morais, D.A.; Fang, H.; Rackham, O.J.; Wilson, D.; Pethica, R.; Chothia, C.; Gough, J. SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res.* **2011**, *39*, D427–D434.
17. Burge, C.; Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **1997**, *268*, 78–94.
18. Burge, C.B.; Karlin, S. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **1998**, *8*, 346–354.
19. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **2004**, *5*, 59.
20. Miller, W.; Makova, K.D.; Nekrutenko, A.; Hardison, R.C. Comparative Genomics. *Ann. Rev. Genomics Hum. Genet.* **2004**, *5*, 15–56.
21. Hubbard, T.; Barker, D.; Birney, E.; Cameron, G.; Chen, Y.; Clark, L.; Cox, T.; Cuff, J.; Curwen, V.; Down, T.; Durbin, R.; Eyras, E.; Gilbert, J.; Hammond, M.; Huminiecki, L.; Kasprzyk, A.; Lehvaslaiho, H.; Lijnzaad, P.; Melsopp, C.; Mongin, E.; Pettett, R.; Pocock, M.; Potter, S.; Rust, A.; Schmidt, E.; Searle, S.; Slater, G.; Smith, J.; Spooner, W.; Stabenau, A.; Stalker, J.; Stupka,

E.; Ureta-Vidal, A.; Vastrik, I.; Clamp, M. The Ensembl genome database project. *Nucleic Acids Research* **2002**, *30*, 38–41.

22. Korenaga, H.; Kono, T.; Sakai, M. P071 Characterization of Interleukinl-17 signaling molecules in teleost. *Cytokine* **2012**, *59*, 541 – 542. ¡ce:title¿10th Joint Meeting of International Cytokine Society and International Society for Interferon and Cytokine Research¡/ce:title¿ ¡xocs:full-name¿10th Joint Meeting of International Cytokine Society and International Society for Interferon and Cytokine Research¡/xocs:full-name¿.
23. Parikesit, A.A.; Stadler, P.F.; Prohaska, Sonja, J. Evolution and Quantitative Comparison of Genome-Wide Protein Domain Distributions. *Genes* **2011**, *2*, 912–924.
24. Stanke, M.; Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **2003**, *19*, ii215–ii225.
25. Stanke, M.; Schöffmann, O.; Morgenstern, B.; Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **2006**, *7*, 62.
26. Stanke, M.; Diekhans, M.; Baertsch, R.; Haussler, D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **2008**, *24*, 637–644.
27. Stanke, M. Lab Session on Gene Prediction with AUGUSTUS, 2011. http://bioinf.uni-greifswald.de/augustus/binaries/tutorial/training.html.
28. Quinlan, A.; Hall, I. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26(6)*, 841–2.
29. Baldauf, S.L. An overview of the phylogeny and diversity of eukaryotes. *J. Syst. Evol.* **2008**, *46*, 263–273.
30. Dilts, B. United States Patent Application Publicatin No: US 2012/0081389 (LucidChart). Technical report, LucidChart, LLC, 2012.
31. Eddy, S.R. Accelerated profile HMM searches. *PLoS Comp. Biol.* **2011**, *7*, e1002195.
32. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat Genet* **2000**, *25*, 25–29.
33. Punta, M.; Coggill, P.C.; Eberhardt, R.Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; Heger, A.; Holm, L.; Sonnhammer, E.L.; Eddy, S.R.; Bateman, A.; Finn, R.D. The Pfam protein families database. *Nucleic Acids Res* **2012**, *40*, D290–D301.
34. Niimura, Y.; Nei, M. Extensive Gains and Losses of Olfactory Receptor Genes in Mammalian Evolution. *PLoS ONE* **2007**, *2*, e708.
35. Pruitt, K.D.; Tatusova, T.; Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **2005**, *33*, D501–D504.
36. Parikesit, A.A.; Stadler, P.F.; Prohaska, S.J. Quantitative Comparison of Genomic-Wide Protein Domain Distributions. German Conference on Bioinformatics 2010; Schomburg, D.; Grote, A., Eds.; Gesellschaft für Informatik: Bonn, 2010; Vol. P-173, *Lecture Notes in Informatics*, pp. 93–102.
37. Conant, G.C.; Wagner, G.P.; Stadler, P.F. Patterns of amino acid substitution in orthologous and paralogous genes. *Mol. Phylog. Evol.* **2007**, *42*, 298–307.

38. Wong, W.C.; Maurer-Stroh, S.; Eisenhaber, F. More Than 1,001 Problems with Protein Domain Databases: Transmembrane Regions, Signal Peptides and the Issue of Sequence Homology. *PLoS Comput. Biol.* **2010**, *6*, e1000867.
39. Michaeli, S. Trans-splicing in trypanosomes: machinery and its impact on the parasite transcriptome. *Future Microbiol.* **2011**, *6*, 459–474.
40. Thomas, S.; Green, A.; Sturm, N.R.; Campbell, D.A.; Myler, P.J. Histone acetylations mark origins of polycistronic transcription in *Leishmania major*. *BMC Genomics* **2009**, *10*, 152.
41. Lu, F.; Jiang, H.; Ding, J.; Mu, J.; Valenzuela, J.G.; Ribeiro, J.M.C.; Su, X.z. cDNA sequences reveal considerable gene prediction inaccuracy in the *Plasmodium falciparum* genome. *BMC Genomics* **2007**, *8*, 255.
42. Iyer, L.M.; Anantharaman, V.; Wolf, M.Y.; Aravind, L. Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. *Int J Parasitol* **2008**, *38*, 1–31.
43. Eisen, J.; et al.. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol* **2006**, *4(9)*, e286.
44. Adam, R.D.; Nigam, A.; Seshadri, V.; Martens, C.A.; Farneth, G.A.; Morrison, H.G.; Nash, T.E.; Porcella, S.F.; Patel, R. The *Giardia lamblia* VSP gene repertoire: characteristics, genomic organization, and evolution. *BMC Genomics* **2010**, *11*, 424.
45. Drinnenberg, I.A.; Weinberg, D.E.; Xie, K.T.; Mower, J.P.; Wolfe, K.H.; Fink, G.R.; Bartel, D.P. RNAi in budding yeast. *Science* **2009**, *326*, 544–550.
46. Melzer, R Theissen, G. MADS and more: transcription factors that shape the plant. *Methods Mol Biol* **2011**, *754*, 3–18.
47. Shelest, E. Transcription factors in fungi. *FEMS Microbiol Lett* **2008**, *286*, 145–151.
48. Bánfai, B.; Jia, H.; Khatun, J.; Wood, E.; Risk, B.; Gundling Jr., W.E.; Kundaje, A.; Gunawardena, H.P.; Yu, Y.; Xie, L.; Krajewski, K.; Strahl, B.D.; Chen, X.; Bickel, P.; Giddings, M.C.; Brown, J.B.; Lipovich, L. Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **2012**, *22*, 1646–1657.
49. Kersting, A.R.; Bornberg-Bauer, E.; Moore, A.D.; Grath, S. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. *Genome Biol Evol* **2012**, *4*, 316–329.
50. Moore, A.D.; Bornberg-Bauer, E. The dynamics and evolutionary potential of domain loss and emergence. *Mol Biol Evol* **2012**, *29*, 787–796.
51. Yang, S.; Doolittle, R.F.; Bourne, P.E. Phylogeny determined by protein domain content. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 373–378.
52. Niehuis, O.; Hartig, G.H.; Garth, S.; Pohl, H.; Lehmann, J.; Tafer, H.; Donath, A.; Krauss, V.; Eisenhardt, C.; Hertel, J.; Petersen, M.; Mayer, C.; Meusemann, K.; Peters, R.S.; Stadler, P.F.; Beutel, R.G.; Bornberg-Bauer, E.; McKenna, D.D.; Misof, B. Genomic and Morphological Evidence Converge to Resolve the Enigma of Strepsiptera. *Current Biol.* **2012**.

Supplemental Data: http://www.bioinf.uni-leipzig.de/supplements/12-007