Bioinformatics Advance Access published August 13, 2016

Data and Text Mining

Extensive Complementarity between Gene Function Prediction Methods

Vedrana Vidulin[1], Tomislav Šmuc[1] and Fran Supek[1,2,*]

[1]Division of Electronics, Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia, [2]EMBL/CRG Systems Biology Unit, Centre for Genomic Regulation, Dr. Aiguader 88, 08003 Barcelona, Spain

*E-mail: fran.supek@irb.hr

Associate Editor: Prof. Alfonso Valencia

Abstract

Motivation: The number of sequenced genomes rises steadily, but we still lack the knowledge about the biological roles of many genes. Automated function prediction (AFP) is thus a necessity. We hypothesize that AFP approaches which draw on distinct genome features may be useful for predicting different types of gene functions, motivating a systematic analysis of the benefits gained by obtaining and integrating such predictions.

Results: Our pipeline amalgamates 5,133,543 genes from 2,071 genomes in a single massive analysis that evaluates five established genomic AFP methodologies. While 1,227 Gene Ontology terms yielded reliable predictions, the majority of these functions were accessible to only one or two of the methods. Moreover, different methods tend to assign a GO term to non-overlapping sets of genes. Thus, inferences made by diverse AFP methods display a striking complementary, both gene-wise and function-wise. Because of this, a viable integration strategy is to rely on a single most-confident prediction per gene/function, instead of enforcing agreement across multiple AFP methods. Using an information-theoretic approach, we estimate that current databases contain 29.2 bits/gene of known *E. coli* gene functions. This can be increased by up to 5.5 bits/gene using individual AFP methods, or by 11 additional bits/gene upon integration, thereby providing a highly-ranking predictor on the CAFA2 community benchmark. Availability of more sequenced genomes boosts the predictive accuracy of AFP approaches and also the benefit from integrating them.

Contact: fran.supek@irb.hr

Supplementary information: Supplementary materials are available at Bioinformatics online.

1 INTRODUCTION

Even though the number of sequenced genomes rises steadily, we still lack the knowledge about the biological roles of many genes. Gene function may be determined experimentally, for instance by observing a phenotype of a mutant organism with an altered or deleted gene of interest (Brochado and Typas, 2013), allowing curators to annotate the gene with Gene Ontology (GO) terms (Ashburner et al., 2000) or with other controlled vocabularies. Experimental essays coupled to manual curation result in high quality function assignments, but are costly, time consuming, and cannot keep up with the deluge of new genome sequences. Reliable automated function prediction (AFP) methods are, therefore, of key importance for functional annotation of newly sequenced genomes and metagenomes (Radivojac et al., 2013, The CAFA Consortium, 2016).

The most common approach to AFP is transferring functions from homologs - genes with shared ancestry - estimated by sequence similarity using BLAST (Altschul et al., 1990) or other tools. In addition to homology, there exist many AFP methods that exploit additional information extracted from the genome sequence, e.g., conserved gene neighborhoods (Ling et al., 2009), phylogenetic distribution (Pellegrini et al., 1999), protein motifs and biophysical properties (Ofer and Linial, 2015), codon usage biases (Kriško et al., 2014), remote homology (Hawkins et al., 2009; Sokolov and Ben-Hur, 2010), and composition of protein domains (Hunter et al., 2011; Punta et al., 2011). Moreover, inference using genomic information can be further supplemented by experimental data: gene expression (Tian et al., 2008), protein-protein interactions (Cao and Cheng, 2015) or protein structure (Wass et al., 2012), and also by text-mining the scientific literature (Cozzetto et al., 2013).

Combining diverse AFP models leads to higher accuracy. This was made evident in the analyses of gene/protein functional association networks, constructed using various sources of large-scale data. Integrating the individual networks resulted in gene modules that were more functionally consistent (Lee et al., 2004; von Mering et al., 2005) and could thus more accurately predict gene function (Troyanskaya et al., 2003; Hu et al., 2009) or phenotypic effects of gene perturbation (Lee et al., 2010).

One explanation for the benefits of integration is that random error from individual data sources cancels out, enabling the signal of gene function to surface. In addition, different sources of genomic or experimental data may be intrinsically better suited for predicting some gene functions than for others. For instance, physical protein-protein interactions more directly correspond to the 'Cellular component' domain of the GO, while genetic interaction experiments relate to the 'Biological process' GO domain. Such rules may, however, also extend to the deeper, more informative levels of the GO. A known example is the contrast between ribosomal proteins and membrane proteins in yeast, where the former are predictable from gene co-expression, while in the latter case, protein compositional features are more relevant (Lanckriet et al., 2004). More generally, assigning function-specific weights to integrated gene networks inferred from biological experiments improves AFP accuracy (Myers and Troyanskaya, 2007; Mostafavi and Morris, 2010). Thus, different high-throughput experimental assays appear to be better suited for predicting different aspects of a gene's role in the cell. Given that AFP methods often draw on analysis of genome sequences to predict gene function, it is thus important to systematically characterize the benefits to combining genomic methods.

We therefore investigate to what extent five well-known sequence-based methodologies differ in their ability to assign particular gene functions across many organisms. One known example are stress response genes, where phylogenetic profiling was shown to be accurate for heat, osmotic and DNA damage responses, but codon usage biases were superior for starvation and oxidative stresses (Škunca et al., 2013; Kriško et al., 2014). We search for broader trends of this sort by examining the overlap and complementarity between purely genome-based AFP methods. An advantage of these approaches is that they apply to any organism with a genome sequence of sufficient quality and do not require costly and time-consuming large-scale experimentation that is restricted to a handful of model organisms.

Relying exclusively on genomic data enabled us to perform AFP on a massive scale, considering >2,000 bacterial and archaeal genomes with >5 million genes in a single analysis, assigning 4,145 different GO functions. Since the amount of sequenced genomes will continue to rise rapidly, there is a need to characterize the contribution of various genomic AFP methodologies towards resolving particular functions of poorly described genes. Crucially, we investigate to what extent the methods will benefit from future availability of more genomes. Using information-theoretic measures, we quantify the current knowledge on gene function in model microorganisms, and suggest that common AFP methods applied to the already-available genome sequences can provide very high-confidence predictions that increase this knowledge by at least 20%. The results of our analysis provide guidelines to researchers on how to best integrate predictions of diverse AFP methods. In particular, one simple but surprisingly accurate strategy is to rely on a single most confident prediction for a given gene and function, thus best exploiting the complementarity between individual genomic predictors.

2 METHODS

2.1 Representing gene families using diverse sets of genomic features

Our pipeline includes five well-established AFP methods relying on genomic data, which we examined in terms of complementarity of their predictions (Fig. 1; implementation details in Sec. S1).

First, the phyletic profiles (PP) method represents the COG/NOG gene families (OGs; see below) by the presence/absence patterns of their member genes across 2,071 genomes, and then makes inferences about gene functions by comparing such patterns via pairwise similarity (Fig. S1a; Pellegrini et al., 1999; Kensche et al., 2008; de Vienne and Azé, 2012) or by machine learning (Tian et al., 2008;

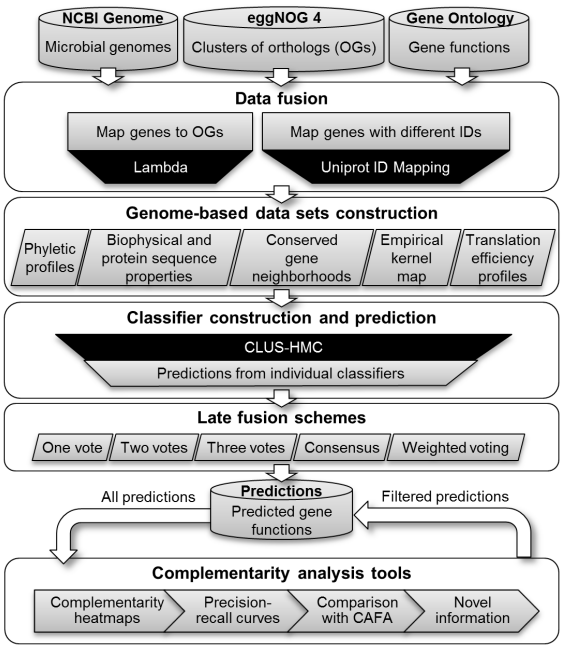

Fig. 1. A pipeline for automated function prediction from genomic data.

Škunca et al., 2013).

Second, biophysical and protein sequence properties (BPS) method includes 1,170 features representing amino acid composition, particular motifs or periodicities (King et al., 2001; Jensen et al., 2003; Lanckriet et al., 2004; Minneci et al., 2013) and various sequence statistics (summary in Sec. S1). Features were extracted using ProFET (Ofer and Linial, 2015).

Third, evolutionarily conserved gene neighborhoods (CGN) may reflect co-regulated genes (Rogozin et al., 2002; Lemay et al., 2012) and can thus be used to infer gene function (Ling et al., 2009). Here, the data consists of the average log-distance (in bp) between genes from individual gene families, measured across all genomes (Fig. S1d). For computational efficiency reasons, the feature set encompasses the 5,891 most common gene families (occurring in ≥100 genomes).

Fourth, signal from remote homologs may predict gene function, because such (individually unreliable) hits may be collectively enriched with correct gene functions (Wass and Sternberg, 2008; Hawkins et al., 2009). We employ the empirical kernel map (EKM) (Tsuda, 1999; Lanckriet et al., 2004; Sokolov and Ben-Hur, 2010) approach, wherein sequence similarity between pairs of OGs is considered by performing searches towards a reference set of genomes, in our case encompassing six genomes (Sec. S1) and 8,447 OGs therein.

Fifth, evolution of codon usage biases relates to phenotypic divergence (Man and Pilpel, 2007) and can be used to predict the roles of genes in environmental adaptation (Supek et al., 2010; Kriško et al., 2014). The translation efficiency profiles (TEP) (Kriško et al., 2014) measure codon biases associated to gene expression; similarity of such profiles suggests co-evolution of expression levels (Fraser et al., 2004). The profiles are represented with 2,071 features indicating OG's predicted expression levels throughout genomes, and additionally 5,891 features that capture OGs predicted co-expression patterns (Sec. S1, Fig. S1c).

2.2 Integrating across genomes in a single massive AFP analysis

Importantly, prior to making inferences with each method, we amalgamate 5,133,543 genes from the 2,071 bacterial and archaeal genomes using COG/NOG gene families, here collectively referred to as OGs. In particular, we selected 21,626 OGs from the EggNOG 4 database (Powell et al., 2013) that were represented in at least 20 genomes. These OGs form examples in our data sets, each described by the five distinct groups of features, as described above.

Having a single, cross-genome set of training examples facilitates unbiased comparisons between the AFP methods, with conclusions valid for many organisms. Such a gene family-based representation is moreover orders of magnitude computationally more efficient than treating thousands of organisms separately (the typically employed 'focal species' approach).

Using OGs as examples bears an implicit assumption that the genes within an OG share functions, and thus can be represented by a single data point obtained by integrating over all genes in the OG. In practice, the GO term labels of the OG were obtained by propagating the known functions of individual genes across the OG, if a specific function is initially assigned to at least 50% of the OG member genes that had known functions (as in Škunca et al., 2013). This yields 15,318 OGs with at least one non-root GO term assigned, which constitute the training set of examples; the remaining 6,308 OGs were initially unlabelled but could receive predictions. Thus, our pipeline first propagates GO annotations via sequence similarity within the OGs, and then transfers GO functions across the OGs using machine learning on five genomic representations, which are orthogonal to the homology transfer employed in the first step.

A classification model is constructed for each of the five AFP methods using the supervised learning algorithm CLUS-HMC, a Random Forest classifier adapted for hierarchical multi-label classification. CLUS-HMC can exploit the hierarchical relationships in GO to achieve higher predictive performance (Blockeel et al., 2006; Vens et al., 2008) and was previously used for AFP tasks (Schietgat et al., 2010; Slavkov et al., 2010; Škunca et al., 2013). For each OG and GO term pair, the classifier outputs a score ranging between 0 and 1 that indicates confidence in assignment of that function to the OG.

Predictions from the individual classifiers are then combined. One approach to this is 'early fusion', which would imply joining the five sets of features together before having constructed classification models (Snoek et al., 2005; Dong et al., 2014). Here, we employ the 'late fusion' approach, wherein each set of features was used to train a separate classifier and the outcomes were subsequently combined using different schemes.

The 'one vote' scheme requires the support of only a single classifier, meaning it reports the maximum confidence observed among the individual classifiers. On the other hand, 'two votes' and 'three votes' schemes require independent support of more classifiers at a given level of confidence, meaning they report the second-highest and third-highest score. Next, 'weighted voting' reports the mean of individual classifiers' confidences weighted by classifiers' accuracy (as the area under precision-recall curve (AUPRC) score; explained below). Finally, 'consensus' considers support of ≥1 classifier, reporting confidence at least equal to the maximum confidence among the individual classifiers, which can be further increased with calls from additional classifiers and was computed as:

$$C_{\text{consensus}}(OG_i, GO_j) = 1 - \prod_{p \in P} \left(1 - C_p(OG_i, GO_j)\right) \quad (1)$$

C_p is a confidence of an individual predictor p that GO_j is assigned to an OG_i. P is the set of five classification models, each trained on a set of features derived from a single AFP method.

2.3 Complementarity analysis and evaluation measures

First, a visual estimate of overall complementarity between methods was provided by clustered heatmaps revealing groups of GO functions well-predicted by each of the methods. Second, precision-recall (P-R) curves and the corresponding area-under-P-R-curve (AUPRC) score quantify the accuracy of individual predictors and of the combination schemes. Third, the choice of the scheme(s) is validated on the external Critical Assessment of Functional Annotation 2 (CAFA 2) benchmark (The CAFA Consortium, 2016). Finally, selected scheme(s) are evaluated in terms of the proportion of genes in model microorganisms that received new GO functions, amount of novel information brought by that functions and the extent to which the scheme(s) may benefit from additional genomes.

In the P-R analysis, the predictors' generalization ability is estimated using out-of-bag cross-validation (Breiman et al., 2001) performed on 15,318 OGs with available GO annotations. For OG-GO pairs, the confidence scores given by the classifier are converted into the precision (Pr) scores using P-R curves obtained from cross-validation. Importantly, unlike the confidence score, Pr has a probabilistic interpretation and is equivalent to 1 - false discovery rate (FDR). Upon combining the confidence scores of the five classification models, this integrated score is also converted to a Pr score using the joint P-R curve. In this setting, the individual fusion schemes are not inherently more permissive or more stringent, but the tradeoff between the two extremes can be adjusted by choosing a Pr threshold for the fused predictions.

AUPRC represents the area under P-R curve, summarizing the precision vs. recall tradeoff at various Pr levels. It is computed separately for each GO term by varying a Pr threshold from one to zero, thus gradually relaxing stringency of the predictions and consequently

increasing the number of OGs that receive a GO label. Classifier AUPRC is an area under P-R curve averaged over individual GO curves.

Further, both the training set and unlabeled OGs are classified with each of the five classifiers, following the rationale that the sets of known functions assigned to OGs are incomplete (Dessimoz et al., 2013).

We validate our predictions using CAFA 2, an AFP community challenge where organizers publish a benchmark set with unknown function (Critical Assessment of Functional Annotation; The CAFA Consortium, 2016). After the submission closes, the experimentally-verified annotations for these genes are collected during a certain period of time and later used to evaluate the competing methods. We benchmarked our results against CAFA 2 *Escherichia coli* set of annotations, following rules of the challenge (the 'no-knowledge' benchmark in full evaluation mode). The evaluations of accuracy of the 129 CAFA 2 participating methods and the BLAST baseline were downloaded from the CAFA web page. The F_{max} measure is computed as the maximum F-measure (harmonic mean of the precision and recall scores), and its standard deviation using bootstrapping (Supplementary Methods).

We estimated the total information in gene function annotations contributed by different predictors using the information accretion (IA) measure (Clark and Radivojac, 2013). IA of a GO term quantifies the increase in specialization in the set of genes assigned to that GO term, compared to its parent in the GO graph. In particular, IA equals zero when the information content of a GO term is equal to its parent. It was computed as:

$$IA(GO_i) = -\log_2 P(GO_i|T) \quad (2)$$

T is a set of parent terms of GO_i and P denotes conditional probability. We summed the IA of assigned annotations on the gene level and expressed it in bits per gene, both for known GO annotations and also for newly-predicted ones in several representative genomes.

3 RESULTS

3.1 Extensive complementarity between AFP methods

Two methods that predict gene functions are complementary if one draws on a set of features strongly associated with genes having a certain function, while the features used by the other method are uninformative in the context of that specific function.

A simple measure of complementarity is to consider whether a GO function is learnable by a certain method, here defined as the method being able to provide at least one prediction at Pr≥50% (equivalent to ≤50% FDR) measured in cross-validation. In other words, the features considered by this method can be used to consistently recover one or more genes with that GO function from the entire dataset. Out of 4,145 GO functions considered in our analyses, 1,227 are learnable by either of the five methods or some combination thereof. Remarkably, 30% of these GO functions are only learnable by a single classifier and inaccessible to the other four. A further 25% are only learnable by two out of five classifiers (Fig. 2a). In other words, almost half of the learnable GO terms are not accessible to the majority of the AFP methods. On the other hand, only 16% of the GOs are learnable by all of the five classifiers, and moreover these disparities become even more pronounced with a more stringent threshold for learnability (≤20% FDR; Fig. 2b). This reveals a considerable complementarity between the different methods: if methods are applied individually, some gene functions may be predicted highly accurately while the others not at all. A combination of genome sequence-based predictors is able to reach many different GO functions, consistent with the success of past approaches that integrate across large-scale experimental data sources (Troyanskaya et al., 2003; Lee et al., 2004; Lanckriet et al., 2004; von Mering et al., 2005; Hu et al., 2009; Lee et al., 2010).

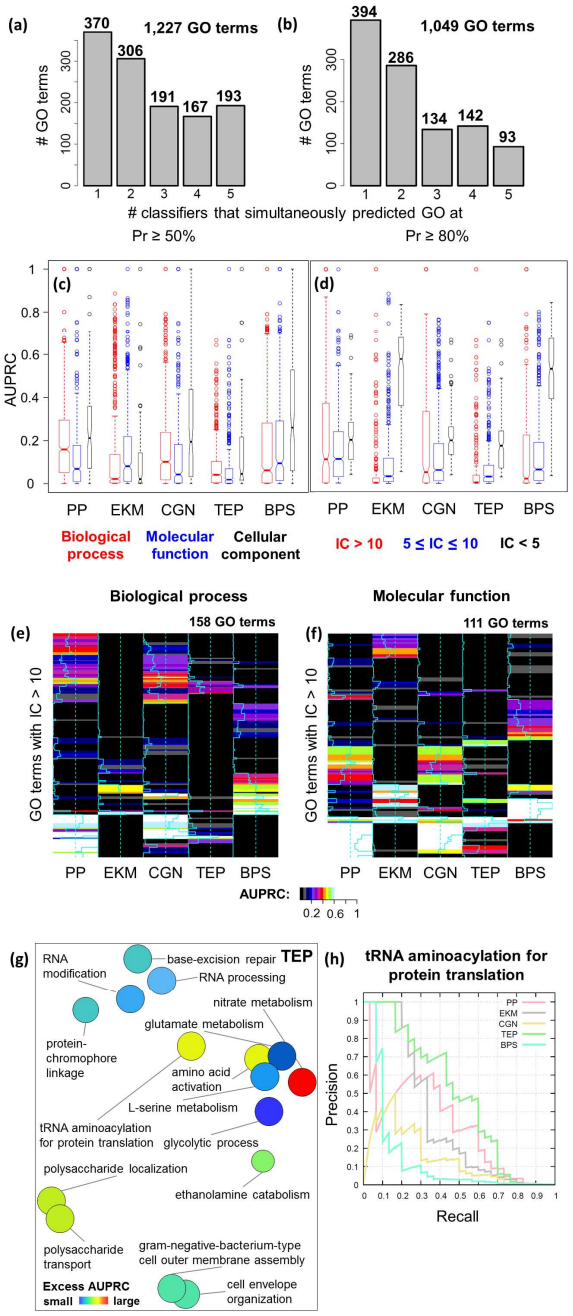

Fig. 2. Complementarity between the AFP methods. (**a**, **b**) The number of GO terms learnable by one, two or more classifiers. (**c**, **d**) Distributions of classifiers' AUPRCs scores for GO terms, stratified by GO domain and by information content (IC); lower IC scores denote more general terms. (**e**, **f**) Complementarity patterns in high-IC GO terms in different GO domains, where rows represent GO terms, columns represent prediction methods and brighter colors relate to higher accuracy, as AUPRC score (in crossvalidation). (**g**) Examples of GO terms learned by TEP better than by the rest of the classifiers. Excess AUPRC for a GO term (color) is computed by subtracting the AUPRC of the best-performing other classifier from the TEP AUPRC, for a particular GO. (**h**) Precision-recall curve for the selected GO term where the TEP method performs well. PP, phyletic profiles; EKM, empirical kernel map; CGN, conserved gene neighborhoods; TEP, translation efficiency profiles; BPS, biophysical and protein sequence properties.

Next, we compare the accuracy of individual classifiers measured by the cross-validation AUPRC score (Methods) for each individual GO category (Fig. 2c). A broad trend can be observed when comparing the three GO domains: the two sequence-based methods (EKM and BPS) are generally better at predicting Molecular function GOs than the

Biological process GOs ($p=3.6*10^{-7}$ and 0.02 for EKM and BPS, respectively; Mann–Whitney test). This is consistent with their ability to capture protein sequence motifs and general structural features informative of enzymatic activity. On the contrary, the three 'genomic context' methods (PP, TEP and CGN) are better at predicting the Biological process GOs ($p=10^{-13}$, $2*10^{-4}$ and $3*10^{-8}$, respectively). This is consistent with their ability to capture the signal emanating from genetic interactions, thus describing the context of a protein in a functional association network.

The methods' relative performance also broadly differs between the generality levels of GO functions (Fig. 2d). The sequence motif-based EKM and BPS methods are more adept at capturing broader, more general functional categories with information content (IC) <5 than the more specific GOs ($p<2*10^{-16}$). In contrast, the genomic context PP and CGN methods have higher overall performance for the more specific GOs with IC≥5, in comparison to the more general GOs ($p=10^{-9}$ and $p<2*10^{-16}$, respectively).

These broad trends notwithstanding, the predictive accuracy of individual methods varies widely even between GOs in the same domain and of similar information content (Fig. 2e,f, S2a,c). Importantly, such patterns are also to a great extent different between the individual methods, and we next examine the comparative strengths and weaknesses of each AFP method with regard to the specific GO categories they predict.

Of note, the overall ability to predict GO functions differs between methods: BPS has the highest AUPRC out of the five methods for 33% of the 1,227 learnable GO terms, and PP in 25% of the GO terms (example GO terms in Fig. S2d-g). Nevertheless, the other three methods prove valuable when predictions for particular GO terms are sought. For instance, TEP is the method with highest cross-validation AUPRC scores for the functions 'tRNA aminoacylation for protein translation' and for 'photosynthesis' (Fig. 2h, S2b), and it exhibits comparable overall performance to other methods across a set of other GO terms (Fig. 2g; trends across GO terms for other methods visualized in Fig. S2). Crucially, even two apparently equally performing methods – exhibiting similar AUPRC for a GO term – may provide complementary predictions in practice, assigning the function to disjoint sets of genes. We further examine to what extent this occurs and how can it be exploited to boost predictive power by combining classifiers.

3.2 Method complementarity and prediction fusion

We quantified the complementarity of the five predictors described above by testing the accuracy of combined predictions. In particular, we evaluated five different fusion schemes in a cross-validation test and additionally on the CAFA 2 benchmark that served as an independent validation, while stratifying by GO domain and information content (IC) of GO terms (Fig. 3, Fig. S3).

Overall, integration schemes perform substantially better than the individual predictors, regardless of the GO term generality (Fig. 3a,b) or of the GO domain analyzed (Fig. 3a,b; S3a-d). For example, the AUPRC scores for the most specific (IC>10) GO terms range between 0.04 and 0.28 for the five individual predictors, and between 0.18 and 0.40 for the five fusion schemes (Fig. 3a, S3d). Therefore, the methods indeed do cover different sets of genes with their predictions, raising the combined accuracy far above the individual methods.

With respect to strategies to integrate predictions, an appealing approach is to require that an annotation be made by more than one independent methodology. Intuitively, enforcing consistency across the methods should imply more confidence in the call. We tested this approach using 'two votes' and 'three votes' schemes that conservatively annotate functions only if supported in >1 predictor (Methods). However, such schemes were routinely outperformed by the two commonly employed fusion schemes (Fig. 3b, S3a-d) that integrate

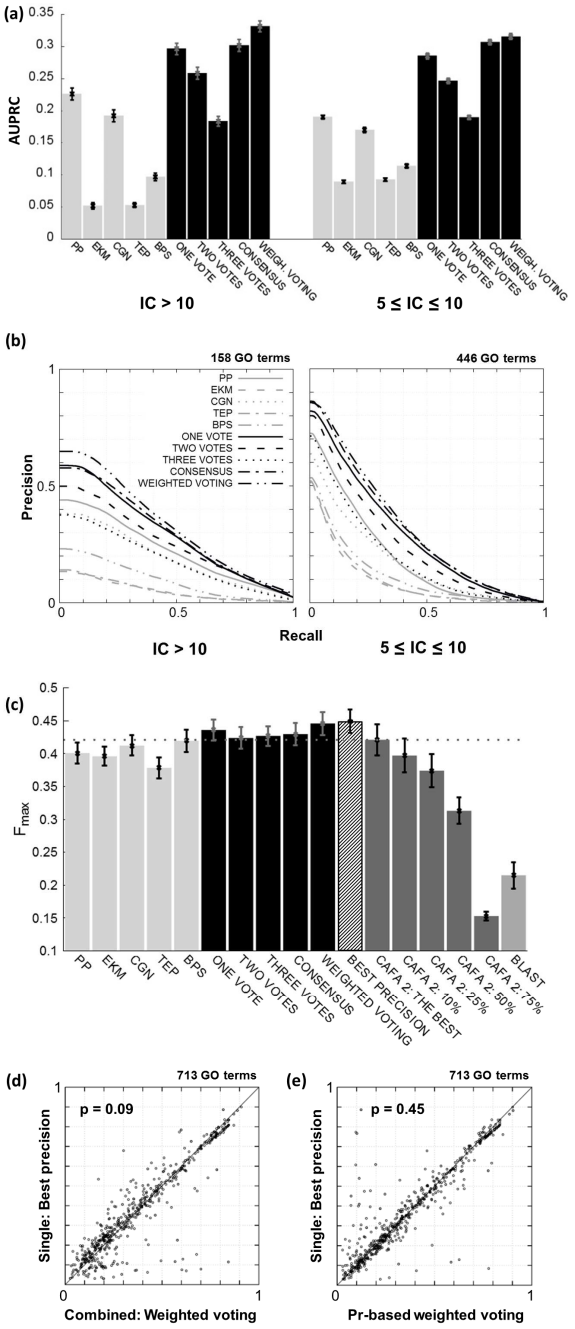

Fig. 3. Comparison of predictive performance between individual methods and integration schemes. (**a**) Average AUPRC scores for the Biological process GO domain computed from precision-recall (PR) curves, obtained in cross-validation. Error bars are standard error of the mean. IC, information content; lower IC denotes more general GO terms. (**b**) PR curves computed by averaging individual Biological process GO term PR curves, stratified by IC. (**c**) The F_{max} accuracy measure on the CAFA 2 *E. coli* validation set. Error bars are standard deviation by bootstrapping the set of benchmark genes. (**d**, **e**) Cross-validation AUPRC scores of the individual Biological process GO terms, while comparing the 'best precision' *vs.* the 'weighted voting' scheme (**d**) and a Pr-based weighted voting scheme in (**e**). p-values are by Wilcoxon test. Acronyms are explained in legend of Fig. 2.

the predictions across all methods using weights ('weighted voting' and 'consensus'; Methods). These integration schemes allow the overall result to stem only from a single confident prediction, even if it is not consistent across individual methods. This motivated us to test a simplified strategy where we simply take the prediction of the single most confident model as the final prediction ('one vote'). Somewhat counterintuitively, this approach appears to perform equally well to

the 'weighted voting' for the general GO terms (Fig. 3b, S3a-d), and similarly well even for the more specific GO terms (Fig. 3b, S3c, d). This observation can be explained by the very high complementarity between the methods – if the majority of reliable annotations are predicted only by a single method and there is little overlap, even sophisticated methods to combine them will not improve much over the 'one vote' approach, and might even be counterproductive in some instances.

We further refined the 'one vote' scheme to first compute Pr scores separately for each of the five methods and then to take the highest Pr score among the methods as an integrated prediction ('best precision' scheme; highlighted bar in Fig 3c). This implicitly incorporates information, via Pr scores, on the accuracy of classifiers in making each individual prediction. Such a scheme that considers only a single prediction with highest Pr (or, equivalently, lowest FDR) performs indistinguishably from the commonly 'weighted voting' scheme that combines many classifiers (Fig. 3d; p=0.09, Wilcoxon signed-rank test). Notably, a Pr-based weighted voting does not outperform the 'best precision' scheme either (Fig. 3e, p=0.45).

The results from cross-validation were further validated on the *E. coli* predictions from the CAFA 2 benchmark. Of note, the two tests are on a rather different scale: cross-validation results are obtained from multiple genomes (15,318 OGs) compared to a single benchmark genome (70 genes). Furthermore, the number of GOs available for testing is reduced from 713 to 232 in Biological process and 409 to 139 in Molecular function domain. The choice of optimal strategy is confirmed on the largest 'Biological process' part of the benchmark: the conservative 'two votes' and 'three votes' perform worse than the other schemes. Moreover, weighted voting does not outperform the simple 'best precision' scheme (Fig. 3c). In addition, all types of integration are beneficial since all schemes performed equally or better than the best CAFA 2 competitor on the Biological process domain (Fig. 3c). These trends are broadly confirmed on Molecular function domain (Fig. S3c, d): best precision and consensus outperform other schemes and methods. In addition, these two schemes are in the top 25% of the CAFA 2 competitors for the Molecular function *E. coli* benchmark (Fig. S3e).

The individual and integrated GO predictions for the complete set of OGs and genes therein are available from http://gorbi.irb.hr/.

3.3 The tally and overlap of newly predicted functional annotations

We examined the genomes of several model microorganisms in terms of how many genes could be covered with novel GO predictions at a certain Pr (or equivalently FDR) threshold, given a certain annotation method or a fusion scheme to combine them (Fig. 4a, S4). The individual methods could annotate roughly ~1/6 of the genes in the genomes at Pr≥50%, e.g., 9-19% for the different methods in *E. coli* and 8-16% for *Staphylococcus aureus*. Strikingly, using the combination schemes can achieve at least twice the coverage at the same FDR (36-43% and 28-34%, for *E. coli* and *S. aureus*, respectively). Alternatively, combining classifiers can increase the precision while achieving similar coverage as the individual methods. The various fusion schemes perform similarly in this test, with the consensus and weighted voting having an edge at very stringent (Pr≥90%) thresholds. Of note, genes not included in OGs cannot be annotated in our setup and contribute towards the uncovered part of the genome.

Next, we quantified overlaps between methods in terms of particular genes in model microorganisms that received predictions at various Pr thresholds. We observe that the overlap is very low at high stringency thresholds: at Pr≥90%, 98% genes that received any annotation in *E. coli* did so only by a single method (Fig. 4b); this percentage is 100%, 99% and 96% for *S. aureus* (Fig. 4b), for *Bacillus subtilis* and *Streptomyces coelicolor* (Fig. S5), respectively. However, as the stringency is relaxed, the overlap between the covered genes increases, where at Pr≥50% many of the same genes receive predictions from multiple methods (Fig. 4b, Fig. S5, Table S1). Crucially, this does not imply that the same GOs are assigned to those genes by the different methods. Indeed, when quantifying the GO terms that were annotated to at least one OG at Pr≥50%, we observe considerable differences between methods: 36% of the GO terms are assigned to at least one OG only by a single method, an additional 30% by two methods and only 6% by all five methods (Table S2). The complementarity is also evident (at any Pr threshold) in the increased number of GO terms assigned to any one OG upon applying a scheme to combine the annotations (Fig. S6). Overall, the high accuracy of the combined predictors stems both from the complementary in gene functions each method can predict and in the sets of genes that it assigns a particular gene function to.

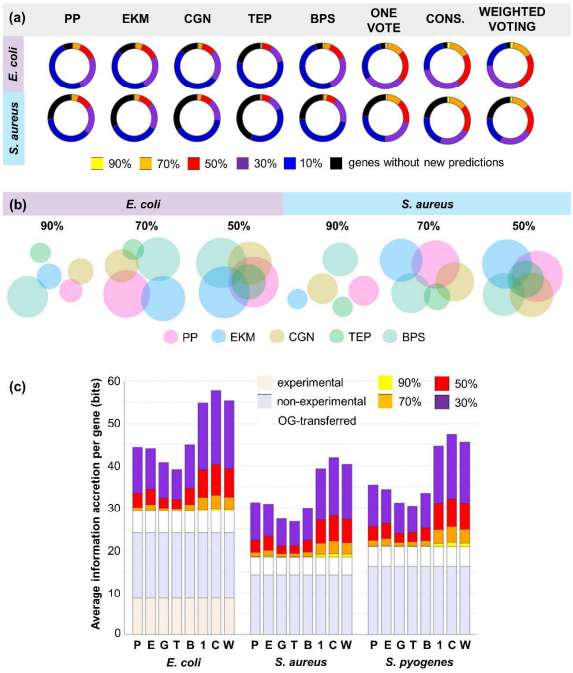

Fig. 4. Coverage of genes for predicted functions in two example microbes (a,b) and average information accretion per gene of known *vs.* newly predicted functional annotations (c). Proportions of (**a**) and overlaps in (**b**) *E. coli* and *S. aureus* genes that received at least one novel specific prediction (IC≥5) at several precision (Pr) thresholds. Pr is equivalent to 1-FDR. Venn diagrams show approximate overlap; complete data in Table S1. (**c**) Three example genomes are shown. Colors show an increase in predictions with decreasing stringency (numbers denote Pr threshold). Method/scheme name is denoted by the first letter of its name, except CGN=G. Acronyms are explained in legend of Fig. 2.

A part of newly predicted annotations can be validated using CAFA 2 *E. coli* benchmark and so can the overlap of the annotated GO terms for particular genes. We searched for examples thereof, in terms of *E. coli* genes that received validated annotations at Pr thresholds corresponding to F_{max} (Supplementary Methods). While the CAFA 2 *E. coli* set is not large enough to quantitate the overlap between validated predictions made by particular methods, we found individual examples that support the trends observed previously in cross-validation tests. For instance, the *fruA* / *fruB* genes had received correct predictions from multiple methods simultaneously, and the predictors for assigned GO terms had low levels of complementarity in our analyses (max. excess AUPRC=0.05; Fig. S7). On the other extreme, there are multiple examples of pronounced complementarity for method-specific GOs (max. excess AUPRC=0.39) correctly predicted for genes but unreachable to other methods, e.g., EKM-specific to *mobB*, TEP-specific to *ung* and BPS-specific to *yciS* (Fig. S7).

3.4 The present and future potential in function prediction methods

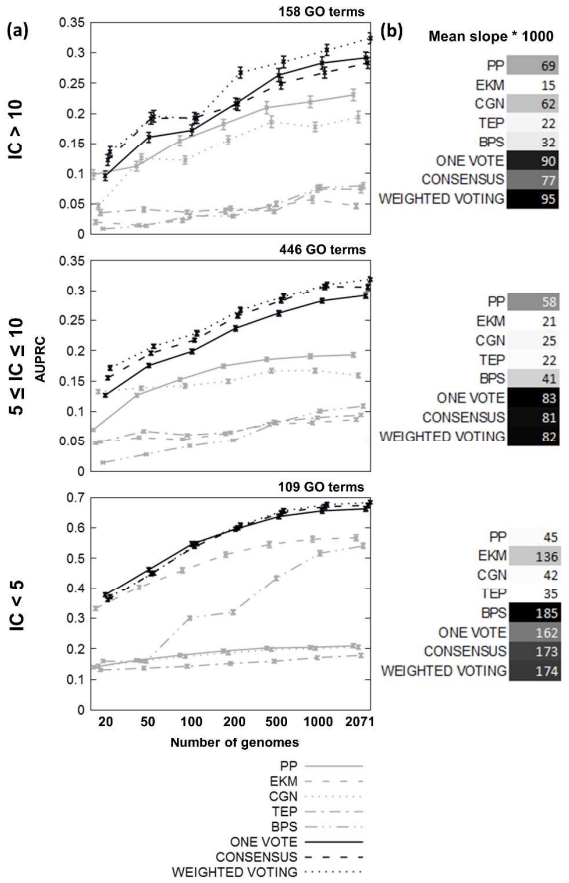

Fig. 5. Accuracy of classifiers increases with addition of genomic data. (a) X-axes represent the number of randomly sampled genomes (of the 2,071 total); approx.. log scale. Y-axes represents classifiers' AUPRCs (in cross-validation) averaged over the selected subset of GO terms from the Biological process domain and error bars standard error of the mean. IC, information content. (b) Slopes of the regression lines for a prediction method/integration scheme, as average over the slopes of segments connecting points in plot; complete table with slopes in Table S3. Acronyms are explained in legend of Fig. 2.

Next, we turn to address the issue of how well the genes are covered by novel annotations using different methodologies. In particular, we measure the total amount of information accretion (IA; Methods; Clark and Radivojac, 2013) that was contributed by different predictors. We estimate that the *E. coli* genome has on average 29.2 bits/gene of currently known functional annotations spanning all three GO domains (Fig. 4c). Of that, 8.7 bits/gene is assigned directly from experimental data, and the other 15.4 bits/gene is assigned using the commonly-applied electronic annotation methods, per the Uniprot-GOA database (Camon et al., 2005); many of these annotations derive from InterPro (Jones et al., 2014). We supplement this by a further 5.1 bits/gene obtained by transferring GO annotations across OG groups (by sequence similarity). Given that the GO electronic annotations are of comparable quality to the manually curated annotations (Škunca et al., 2012), they are used as input to our function prediction algorithms. At a permissive threshold of Pr≥50%, the individual prediction methods can assign between 2.8 bits (TEP) and 5.5 bits (BPS) for E. coli genes, on average. Integrating the predictions raises this to a total of 11 bits/gene of newly predicted functions (Fig. 4c, 'consensus' scheme) at Pr≥50%. At a more stringent threshold of Pr≥70%, 3.9 new bits/gene are still available (Fig. 4c, consensus). Interestingly, the novel annotations apply similarly well to both the poorly and the well-annotated *E. coli* genes (10.7 *vs.* 11.8 additional bits/gene for genes in the lower *vs.* upper quartile by the known bits/gene; Fig. S9). This suggests that there are still many undiscovered biological roles even in the currently well-annotated genes. This trend is also observed consistently across the three GO domains (Fig. S9). For instance, in addition to the existing 108 bits of annotations to *ftsI* gene, we predict a further 23 bits, 13 of which were from the Molecular function domain and the remainder from the other two domains. The trends above hold also for other organisms, meaning that AFP methods can also afford great gains in medically important microbes: *S. aureus* has 18.4 bits/gene of known annotations but 9.7 additional bits/gene are readily available from predictions; for *Streptococcus pyogenes* this is 20.8 plus 11.3 bits/gene, and for *Mycobacterium tuberculosis* (Fig. S8) this is 17.7 plus 8.2 bits/gene (all given at Pr≥50%).

Therefore, the established genome-based AFP methods can immediately extend our knowledge of gene function using current data. An important question is also how much of this knowledge remains to be gained in the future, as more genomes are sequenced. We address this by sampling from our full set of 2,071 genomes and examining how the accuracy changes with the number of available genome sequences. Interestingly, for the most of the tested methods and integration schemes, the average AUPRC scores increase approx. linearly with the logarithm of the number of genomes. Some saturation is evident in the individual methods with the current set of ~2000 genomes, particularly in the Cellular component GO domain (Fig. 5a, S10 and Table S3). Crucially, the fusion schemes display very little saturation in all but the very general GO terms (IC<5) of the Biological process and Molecular function domains; Fig. 5a, S10 and Table S3). In summary, many AFP methodologies stand much to gain from increases in size of genomic databases. Importantly, the integrated predictions generally exhibit steeper slopes than individual classifiers (Fig. 5b). This suggests that with more genomes, the complementarities between methods grow more pronounced and the relative benefit of integrating across many AFP methods increases.

4 DISCUSSION

Automated gene function prediction is a necessity: the numbers of sequenced genomes are growing rapidly, but the known functional annotations are not keeping up. The methods that transfer known biological roles to homologous genes *via* sequence similarity searches are well-established and appear quite successful in community evaluations (Hamp *et al.*, 2013). Thus, they present a baseline that future methods must build and improve upon, aiming to provide predictions complementary to the commonly employed methods such as PSI-Blast or Pfam searches. To this end, we have evaluated five existing methodologies that produce novel GO annotations from data orthogonal to standard sequence similarity searches, while being based exclusively on genome sequences. We find that the methods are highly complementary: more than half (676/1,227) of examined GO functions are inaccessible to the majority of the AFP methods, but only to one or two individual predictors. In particular, the protein sequence-based methods tend to be more adept at capturing general GO terms and those in the Molecular function domain, while genomic context methods better capture the specific GO terms and the Biological process GO domain. Thus, the output of various comparative genomics-based AFP approaches needs to be combined to find functionally coherent groups of genes.

We find that, due to the pronounced complementarity, a simple yet viable strategy for integrating predictions is to take the prediction of the single most confident model, which performs similarly to weighted voting schemes. Recent research in machine learning explored various classifier combination techniques, concluding that the simple late fusion schemes – not unlike the ones employed in this work – can double the recall at high precision, if the near-independence of feature families is properly exploited (Madani *et al.*, 2013). Consistently, our data also suggests that the benefit gained

from applying and subsequently integrating multiple AFP methods increases when a higher stringency of predictions is desired (Fig. 4b).

The scientific community has been painstakingly accumulating knowledge about protein function by performing experiments in model organisms, such as *E. coli*, throughout the past decades. We estimate that current knowledge – to the extent it can be described by the GO and covered in the available databases – amounts to 29.2 bits per *E. coli* gene, on average. The established AFP methods operating only on genomic data, if combined properly, can increase this by a further 11 bits/gene. Finally, we showed that various integration schemes profit more than the individual methods from inclusion of additional genomes, highlighting the increasing importance of considering multiple complementary genomic AFP methods in future work.

Funding

This work was supported by the European Commission via projects MAESTRA [ICT-2013-612944], InnoMol [316289] and MULTIPLEX [317532], and by the Croatian Science Foundation via grants DescriptiveInduction [HRZZ-9623] and Multicast [HRZZ-5660].

Author contributions

VV prepared genomic data sets and performed all statistical analyses. FS conceived and supervised the project. TS supervised the project. All authors wrote the manuscript.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990) Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... & Sherlock, G. (2000) Gene Ontology: tool for the unification of biology. Nature genetics, 25(1), 25-29.

Blockeel, H., Schietgat, L., Struyf, J., Džeroski, S., & Clare, A. (2006) Decision trees for hierarchical multilabel classification: A case study in functional genomics (pp. 18-29). In Knowledge Discovery in Databases: PKDD 2006, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, Proceedings. Lecture Notes in Computer Science, Springer Berlin Heidelberg.

Breiman, L. (2001) Random forests. Machine learning, 45(1), 5-32.

Brochado, A. R., & Typas, A. (2013) High-throughput approaches to understanding gene function and mapping network architecture in bacteria. Current opinion in microbiology, 16(2), 199-206.

Camon, E. B., Barrell, D. G., Dimmer, E. C., Lee, V., Magrane, M., Maslen, J., ... & Apweiler, R. (2005) An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC bioinformatics, 6(Suppl 1), S17.

Cao, R., & Cheng, J. (2015) Integrated protein function prediction by mining function associations, sequences, and protein-protein and gene-gene interaction networks. Methods.

Clark, W. T., & Radivojac, P. (2013) Information-theoretic evaluation of predicted ontological annotations. Bioinformatics, 29(13), i53-i61.

Cozzetto, D., Buchan, D. W., Bryson, K., & Jones, D. T. (2013) Protein function prediction by massive integration of evolutionary analyses and multiple data sources. BMC bioinformatics, 14(Suppl 3), S1.

de Vienne, D. M., & Azé, J. (2012) Efficient prediction of co-complexed proteins based on coevolution. PloS one, 7(11), e48728.

Dessimoz, C., Škunca, N., & Thomas, P. D. (2013) CAFA and the open world of protein function predictions. Trends in genetics: TIG, 29(11), 609-610.

Dong, Y., Gao, S., Tao, K., Liu, J., & Wang, H. (2014) Performance evaluation of early and late fusion methods for generic semantics indexing. Pattern Analysis and Applications, 17(1), 37-50.

Enault, F., Suhre, K., & Claverie, J. M. (2005) Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC bioinformatics, 6(1), 247.

Fraser, H. B., Hirsh, A. E., Wall, D. P., & Eisen, M. B. (2004) Coevolution of gene expression among interacting proteins. Proceedings of the National Academy of Sciences of the United States of America, 101(24), 9033-9038.

Hamp, T., Kassner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., ... & Heron, M. (2013) Homology-based inference sets the bar high for protein function prediction. BMC bioinformatics, 14(3), 1.

Hawkins, T., Chitale, M., Luban, S., & Kihara, D. (2009) PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. Proteins: Structure, Function, and Bioinformatics, 74(3), 566-582.

Hu, P., Janga, S. C., Babu, M., Díaz-Mejía, J. J., Butland, G., Yang, W., ... & Chandran, S. (2009) Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS biology, 7(4), 929.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., ... & Yong, S. Y. (2011) InterPro in 2011: new developments in the family and domain prediction database. Nucleic acids research, gkr948.

Jensen, L. J., Gupta, R., Staerfeldt, H. H., & Brunak, S. (2003) Prediction of human protein function according to Gene Ontology categories. Bioinformatics, 19(5), 635-642.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., ... & Pesseat, S. (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics, 30(9), 1236-1240.

Kensche, P. R., van Noort, V., Dutilh, B. E., & Huynen, M. A. (2008) Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. Journal of The Royal Society Interface, 5(19), 151-170.

King, R. D., Karwath, A., Clare, A., & Dehaspe, L. (2001) The utility of different representations of protein sequence for predicting functional class. Bioinformatics, 17(5), 445-454.

Kriško, A., Copić, T., Gabaldón, T., Lehner, B., & Supek, F. (2014) Inferring gene function from evolutionary change in signatures of translation efficiency. Genome Biol, 15(3), R44.

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., & Noble, W. S. (2004) A statistical framework for genomic data fusion. Bioinformatics, 20(16), 2626-2635.

Lee, I., Date, S. V., Adai, A. T., & Marcotte, E. M. (2004) A probabilistic functional network of yeast genes. science, 306(5701), 1555-1558.

Lee, I., Lehner, B., Vavouri, T., Shin, J., Fraser, A. G., & Marcotte, E. M. (2010) Predicting genetic modifier loci using functional gene networks. Genome research, 20(8), 1143-1153.

Lemay, D. G., Martin, W. F., Hinrichs, A. S., Rijnkels, M., German, J. B., Korf, I., & Pollard, K. S. (2012) G-NEST: a gene neighborhood scoring tool to identify co-conserved, co-expressed genes. BMC bioinformatics, 13(1), 253.

Ling, X., He, X., & Xin, D. (2009) Detecting gene clusters under evolutionary constraint in a large number of genomes. Bioinformatics, 25(5), 571-577.

Madani, O., Georg, M., & Ross, D. (2013) On using nearly-independent feature families for high precision and confidence. Machine learning, 92(2-3), 457-477.

Man, O., & Pilpel, Y. (2007) Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nature genetics, 39(3), 415-421.

Minneci, F., Piovesan, D., Cozzetto, D., & Jones, D. T. (2013) FFPred 2.0: improved homology-independent prediction of gene ontology terms for eukaryotic protein sequences. PLoS ONE 8(5): e63754.

Mostafavi, S., & Morris, Q. (2010) Fast integration of heterogeneous data sources for predicting gene function with limited annotation. Bioinformatics, 26(14), 1759-1765.

Myers, C. L., & Troyanskaya, O. G. (2007) Context-sensitive data integration and prediction of biological networks. Bioinformatics, 23(17), 2322-2330.

Ofer, D., & Linial, M. (2015) ProFET: Feature engineering captures high-level protein functions. Bioinformatics, btv345.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., & Yeates, T. O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences, 96(8), 4285-4288.

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., ... & Bork, P. (2013) eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic acids research, gkt1253.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., ... & Finn, R. D. (2011) The Pfam protein families database. Nucleic acids research, gkr1065.

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., ... & Pandey, G. (2013) A large-scale evaluation of computational protein function prediction. Nature methods, 10(3), 221-227.

Rogozin, I. B., Makarova, K. S., Murvai, J., Czabarka, E., Wolf, Y. I., Tatusov, R. L., ... & Koonin, E. V. (2002) Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Research, 30(10), 2212-2223.

Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Kocev, D., & Džeroski, S. (2010) Predicting gene function using hierarchical multi-label decision tree ensembles. BMC bioinformatics, 11(1), 1.

Slavkov, I., Gjorgjioski, V., Struyf, J., & Džeroski, S. (2010) Finding explained groups of time-course gene expression profiles with predictive clustering trees. Molecular BioSystems, 6(4), 729-740.

Snoek, C. G., Worring, M., & Smeulders, A. W. (2005) Early versus late fusion in semantic video analysis. In Proceedings of the 13th annual ACM international conference on Multimedia (pp. 399-402).

Sokolov, A., & Ben-Hur, A. (2010) Hierarchical classification of Gene Ontology terms using the GOstruct method. Journal of bioinformatics and computational biology, 8(02), 357-376.

Supek, F., Škunca, N., Repar, J., Vlahoviček, K., & Šmuc, T. (2010) Translational selection is ubiquitous in prokaryotes. PLoS Genet, 6(6), e1001004.

Škunca, N., Altenhoff, A., & Dessimoz, C. (2012) Quality of computationally inferred gene ontology annotations. PLoS Comput. Biol, 8(5), e1002533.

Škunca, N., Bošnjak, M., Kriško, A., Panov, P., Džeroski, S., Šmuc, T., & Supek, F. (2013) Phyletic profiling with cliques of orthologs is enhanced by signatures of paralogy relationships. PLoS Comput Biol, 9(1), e1002852.

The CAFA Consortium. (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biology (in press). arXiv:1601.00891

Tian, W., Zhang, L. V., Tasan, M., Gibbons, F. D., King, O. D., Park, J., ... & Roth, F. P. (2008) Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function. Genome Biol, 9(Suppl 1), S7.

Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., & Botstein, D. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proceedings of the National Academy of Sciences, 100(14), 8348-8353.

Tsuda, K. (1999) Support vector classifier with asymmetric kernel functions. In in European Symposium on Artificial Neural Networks (ESANN).

Vens, C., Struyf, J., Schietgat, L., Džeroski, S., & Blockeel, H. (2008) Decision trees for hierarchical multi-label classification. Machine Learning, 73(2), 185-214.

Von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., ... & Bork, P. (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic acids research, 33(suppl 1), D433-D437.

Wass, M. N., & Sternberg, M. J. (2008) ConFunc—functional annotation in the twilight zone. Bioinformatics, 24(6), 798-806.

Wass, M. N., Barton, G., & Sternberg, M. J. (2012) CombFunc: predicting protein function using heterogeneous data sources. Nucleic acids research, 40(W1), W466-W470.