INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND STUDIES

VOLUME03ISSUE06

# ‘’BIOLOGICAL PROCESS ANNOTATION OF PROTEINS ACROSS THE KINGDOM PLANTAE’’

KANJI, E.E; ODE P.O.

## ABSTRACT

Accurate annotation of protein function is vital to understanding life at the molecular level, but automated annotation of functions is challenging. We here demonstrate the mix of a way for protein function annotation that uses network information to predict the biological processes a protein is involved in, with a sequence-based prediction method. The combined function prediction relies on co-expression networks and combines the network-based prediction method BMRF with the sequence-based prediction method Argot2. the mixture shows significantly improved performance compared to every of the methods separately, additionally as compared to BlastGO. The approach was applied to predict biological processes for the proteomes of rice, barrel clover, poplar, soybean, and tomato. The novel function predictions are available at www.ab.wur.nl/bmrf. Analysis of the relationships between sequence similarity and predicted function similarity identifies numerous cases of divergence of biological processes within which proteins are involved, despite sequence similarity. this means that the combination of network-based and sequence-based function prediction is useful towards the analysis of evolutionary relationships. samples of potential divergence are identified for various biological processes, notably for processes associated with cell development, regulation, and response to a chemical stimulus. Such divergence in organic process annotation for proteins with similar sequences should be taken into consideration when analyzing plant gene and genome evolution.

Keywords

Gene function prediction Gene function divergence

## 1. Introduction

The amount of plant genome data grows disproportional to the number of accessible experimental data on these genomes [1], [2], [3], [4], [5]. to attach this ever-increasing amount of genome data to plant biology, factor annotation followed by function annotation is imperative. for instance, the identification of candidate genes involved in a very trait of interest greatly benefits from gene function annotation [6]. within the context of the study of genome evolution, gene function annotations are necessary to enable comparison between sets of genes with different evolutionary histories, e.g. those retained vs. those lost after duplication [7]. To annotate gene or protein function, experimental data, if available, will be wont to annotate gene or protein function. However, the scarcity of experimental data highlights the attractiveness of computational approaches to help in gene function annotation [8]. Indeed, newly sequenced genomes are generally amid a function annotation that heavily relies on computational predictions. Such automated annotations are delivered by a spread of approaches, often without much knowledge about their reliability. For studying plant genomes and plant genome evolution, reliable function annotation is, therefore, a significant challenge.

One way to annotate proteins without experimental data is to infer function from sequence data [3]. The actual standard to capture function annotation today is that the Gene Ontology (GO), particularly, the Molecular Function (MF) and process (BP) sub-ontologies [9]. MF describes activities, like catalytic or binding activities, that occur at the molecular level, whereas BP describes a series of events accomplished by one or more ordered assemblies of molecular functions [9]. Compared to MF, terms within the BP ontology are generally related to more conceptual levels of function; BP terms describe the execution of 1 or more molecular function instances working together to accomplish a specific biological objective. The prediction of BP terms can depend upon the cellular and organismal context [10]. Therefore, BP terms tend to be poorly predicted by methods supported sequence similarity only, like BLAST [10], [11]. The reliability of BP predictions increases with advanced approaches that employ, e.g., phylogenetic frameworks [12], [13] or network data like protein-protein interactions [14].

We recently developed a protein function prediction method for BP terms called Bayesian Markov Random Field (BMRF) [15], which uses network data as input. In BMRF, each protein is represented as a node within the network, and connections within the network indicate functional relationships between proteins. Networks may be supported, e.g., protein-protein interactions or co-expression data. BMRF uses existing BP annotations for proteins within the network to infer biological processes for unannotated proteins therein network. To do so, BMRF uses a statistical model describing how likely neighbors are to participate within the same BP; this constitutes the Markov Random Field. Existing BP annotations are used as “seed” or “training” data, providing a group of initial labels for the Markov Random Field. Parameters within the statistical model are trained employing a Bayesian approach by performing the simultaneous estimation of the model parameters and prediction of protein functions. Importantly, BMRF can transfer functional information beyond direct interactions. Therefore, it can generate function predictions for proteins that are only linked with other proteins with unknown functions.

In the Critical Assessment of Function Annotations (CAFA) protein function prediction challenge [10], BMRF obtained particularly good performance in humans (first place) and Arabidopsis (second place) for BP term prediction [10]. In these species, BMRF performance benefits from the wealth of existing function annotation, i.e. experimental data. due to its dependence on training data, function annotation for species with more sparse function annotation is challenging for BMRF. to enhance the prediction performance in sparsely annotated species, we present here a method to mix BMRF with the sequence-based function prediction method Argot2 [16]. Argot2 was among the top-performing sequence-based algorithms within the CAFA category “eukaryotic BP”. In its computational approach, Argot2 is complementary to BMRF, because it's purely sequence-based.

We demonstrate that a mix of Argot2 and BMRF encompasses a markedly better function prediction performance than each method separately. This integrated method was applied to predict BP terms for proteins in five plant species, Medicago truncatula (barrel clover), cultivated rice (rice), Western balsam poplar (poplar), soja (soybean), and Solanum Lycopersicum (tomato), using microarray co-expression networks as input. Numerous new proteins were related to specific biological processes, like seed development in rice or organic process in Medicago. By comparison between sequence divergence and predicted function divergence, numerous cases of putative neo-functionalization involving various biological processes were identified. This new method and also the resulting set of predicted gene functions are going to be of great value in capitalizing on the massive amount of plant genome data that's currently being generated for the study of the evolution of genome and gene function.

## 2. Results

### 2.1. Method development and evaluation

We previously developed the protein function prediction method BMRF and used it to annotate protein function in cress [17]. This method relies, besides on network data, on existing function annotation as input. For Arabidopsis, we demonstrated that the quantity of accessible annotation (training) data was sufficient to attain a decent prediction performance [17]. However, for crop species, much less annotation data is out there as input. to extend the function prediction performance for plants with sparse experimental data, we explored combining BMRF with the sequence-based method Argot2.

Argot2 and BMRF were tested separately (standalone setting) or in two combinations (Fig. 1). Performance assessment focussed on rice, the crop with the most important amount of annotation data available: 415 proteins with experimental evidence for a organic process. The rice network used as input for BMRF was obtained from a mixture of microarray-based co-expression data, data from STRING [18] and FunctionalNet [19] (Table S1). Of the 415 proteins with experimental evidence, 394 were present within the network and were used for validation of predicted functions.

J.W. Bargsten et al. / Current Plant Biology 1 (2014) 73–82

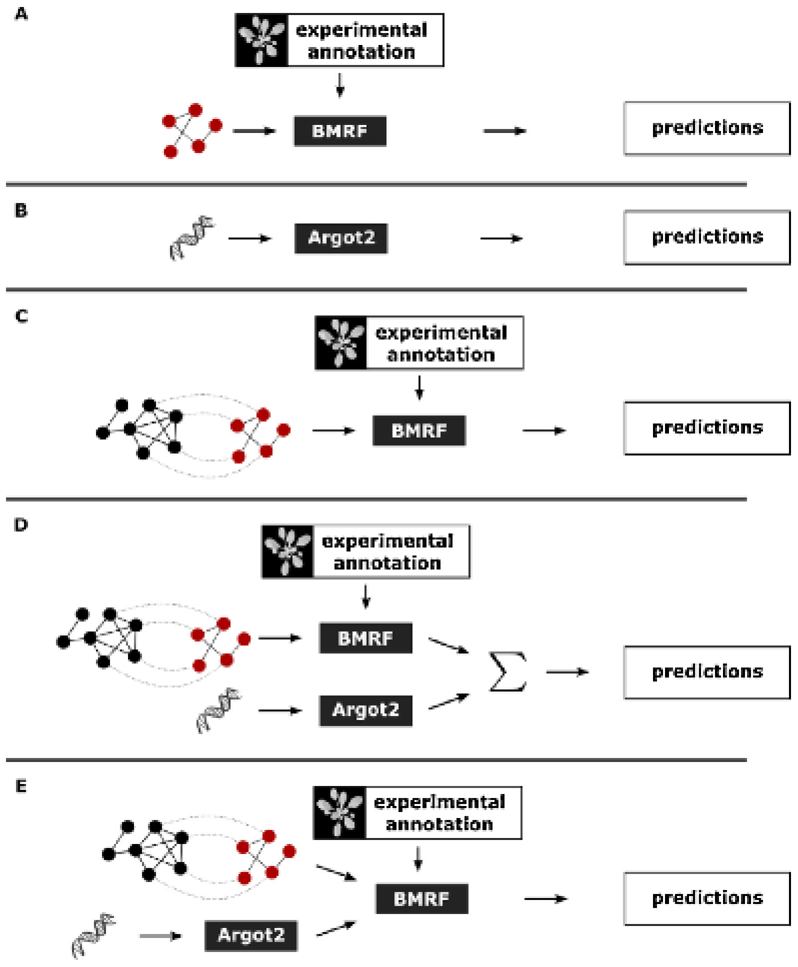

Fig. 1. Strategies for predicting protein function. BMRF (A and C) and Argot2 (B) were utilized in a standalone setting or in two different combinations (D and E). Combining BMRF and Argot2 was done by combining the results of every of the 2 methods (D), and by using Argot2 predictions as input for BMRF (E). The rice network is indicated in red, the Arabidopsis network in black and interspecies connections in grey dashed lines. Sequence-based input is indicated by a DNA-helix symbol. (For interpretation of the references to paint during this figure legend, the reader is spoken the net version of this text.)

Function prediction performance was assessed on the premise of cross-validation, leaving out randomly selected proteins with known function and comparing the predictions with those data. the world under the receiver operator graphical record (AUC) was accustomed compare the performance of the predictions that come as ordered lists of predicted proteins per process. within the standalone setting (Fig. 1A and B) with rice sequence and network data, BMRF and Argot2 both have an occasional performance, with AUC (average ± standard deviation) of 0.6 ± 0.12 and 0.67 ± 0.11, respectively (Tables 1 and S2). These values are considerably not up to the AUC previously obtained with BMRF for Arabidopsis (0.75) [17] thanks to the tiny amount of coaching data (annotated gene functions) that's available for rice. Assuming information from Arabidopsis would improve the performance of rice protein function predictions in BMRF, we connected proteins in an available Arabidopsis network (Table S1) to proteins within the rice network supported sequence similarity using BLAST. With this rice-Arabidopsis interspecies network additionally to the networks of both species separately (Fig. 1C), BMRF performed slightly better than Argot2 (AUC 0.70 ± 0.12). The precise value of the BLAST E-value cut-off wont to create the interspecies network failed to influence the performance of BMRF (data not shown).

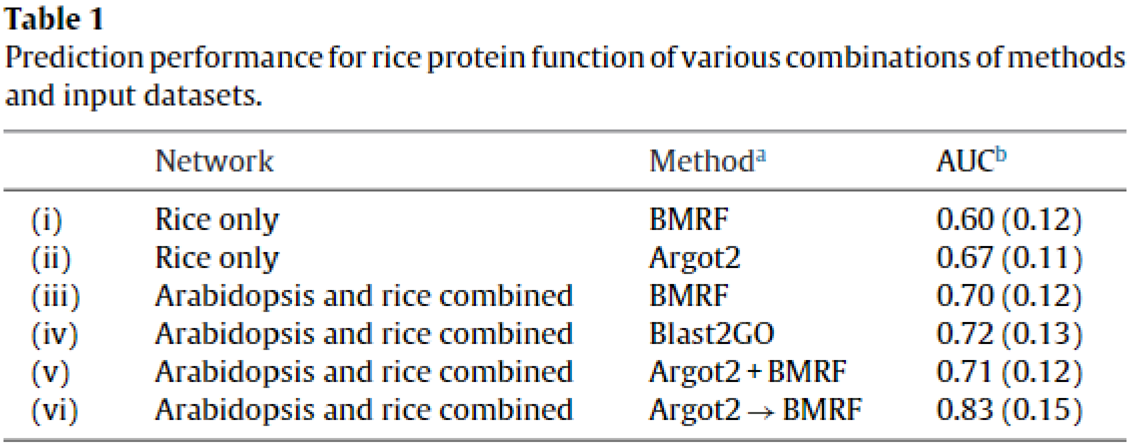

**Table 1**
Prediction performance for rice protein function of various combinations of methods and input datasets.

| | Network | Method[a] | AUC[b] |
|---|---|---|---|
| (i) | Rice only | BMRF | 0.60 (0.12) |
| (ii) | Rice only | Argot2 | 0.67 (0.11) |
| (iii) | Arabidopsis and rice combined | BMRF | 0.70 (0.12) |
| (iv) | Arabidopsis and rice combined | Blast2GO | 0.72 (0.13) |
| (v) | Arabidopsis and rice combined | Argot2 + BMRF | 0.71 (0.12) |
| (vi) | Arabidopsis and rice combined | Argot2 → BMRF | 0.83 (0.15) |

Table 1. Prediction performance for rice protein functions of assorted combinations of methods and input datasets.

Network Methods AUC

(i) Rice only BMRF 0.60 (0.12)

(ii) Rice only Argot2 0.67 (0.11)

(iii) Arabidopsis and rice combined BMRF 0.70 (0.12)

(iv) Arabidopsis and rice combined Blast2GO 0.72 (0.13)

(v) Arabidopsis and rice combined Argot2 + BMRF 0.71 (0.12)

(vi) Arabidopsis and rice combined Argot2 → BMRF 0.83 (0.15)

a).Methods analyzed were BMRF, Argot2, Blast2GO, Argot2 + BMRF (rank-sum), and Argot2 → BMRF (seeding). Rice network was used separately (rice only), or it had been connected to an Arabidopsis network supported sequence similarity (combined).

b).The area under the curve; mean (standard deviation).

Both methods use complementary information about biological processes (network input for BMRF, sequence input for Argot2). Therefore, we tested combining the 2. Argot2 and BMRF are often combined in multiple ways. We used an easy rank-based approach to predict biological processes by ordering Argot2 and BMRF results separately then combining their ranks to provide a final rank (Fig. 1D). This integration was performed for every process separately by sorting the proteins supported their score for that process and using the sum of the ranks induced by this ordering for BMRF and for Argot2. This integration of Argot2 and BMRF failed to improve results compared to standalone BMRF (Table 1). The performance was markedly improved, however, by generating initial predictions with Argot2 and supplying these to BMRF as training data (seed data; Fig. 1E). During this integration method, the initial labeling of proteins within the network (i.e. the seed data for BMRF) was supported the Argot2 predictions. Argot2 uses an algorithm-specific score to rank its results and requires a threshold for such a score. To assess the influence of various thresholds on the performance of BMRF, BMRF was seeded with 5 different output sets of Argot2 (Table S3). The most effective performance was achieved with the default threshold of 5.

The results above indicate that our integrated method performed markedly better than each of the 2 methods separately. As an extra assessment of performance, we predicted annotations with the often-used method Blast2GO [21]. The resulting AUC of Blast2GO was 0.72 ± 0.13, and also the AUC of the combined Argot2-BMRF predictions was 0.83 ± 0.15 which is significantly ($p < 10-15$; Mann–Whitney U) better than Blast2GO (Fig. 2A). the little number of experimentally verified annotations (true positives) and a high number of annotated proteins (true negatives) could introduce skew within the cross-validation sets, resulting in a bias within the AUC performance assessment [22]. The F-score (harmonic mean of precision and recall) doesn't suffer from this skew and therefore the final prediction performance was therefore also assessed with the most F-score (Fmax-score). In agreement with the AUC evaluation, the Fmax-scores of Argot2-seeded BMRF (0.56 ± 0.24) were significantly better ($p < 10-15$; Mann–Whitney U) than Blast2GO (0.51 ± 0.23). Visual inspection of a histogram of AUC values and of Fmax-score values for various BP terms in numerous cross-validation runs confirms the performance difference between the combined Argot2-BMRF predictions and Blast2GO (Fig. 2B and C).

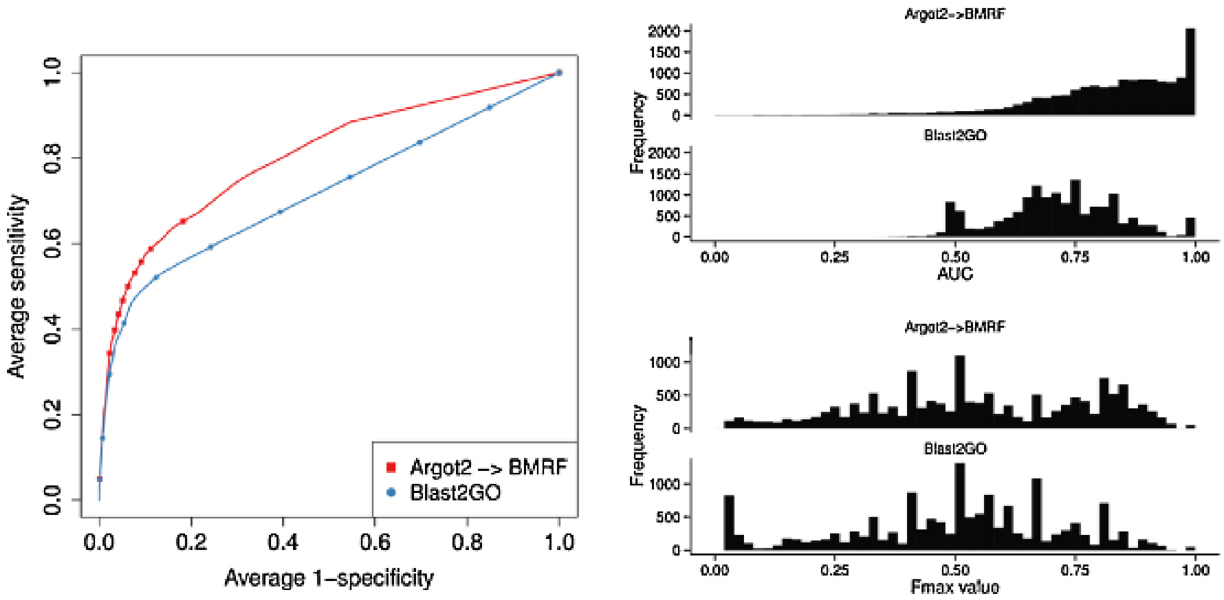

Fig. 2. Performance assessment of function prediction on rice proteins. (A) Receiver operator graphical record showing 1-specificity vs. sensitivity of the predictions of Argot2-seeded BMRF and Blast2GO. Specificity and sensitivity were averaged over all

cross-validation runs. Dots indicate evenly spaced intervals of the underlying prediction score, the road represents a whole curve. Performance is summarized as AUC which is that the area under these curves. (B) Histogram of AUC values per GO term of each cross-validation run calculated for Argot2-seeded BMRF and Blast2GO. (C) Histogram of Fmax values per GO term of each cross-validation run calculated for Argot2 seeding BMRF and Blast2GO.

To obtain independent validation additionally to the cross-validation performed above, the Argot2-seeded BMRF predictions were compared to annotations available within the Oryzabase database [23] which weren't present in our computer file (71 proteins). The AUC of 0.88 ± 0.13 we obtained was just like the AUC obtained within the cross-validation, confirming the performance assessment. Overall, the performance evaluation demonstrates that Argot2-seeded BMRF is a good thanks to predict BP protein function in sparsely annotated plant genomes.

2.2. Application to crop species

Argot2-seeded BMRF using PlaNet [24] co-expression networks as input (Table S4) was applied to predict BP protein functions in a very selection of model and crop plants comprising O. Sativa (rice), M. truncatula (barrel clover), G. max (soybean), P. trichocarpa (poplar) and S. Lycopersicum (tomato). The posterior probability of a protein related to a specific GO term was estimated for all GO terms and every one proteins within the network. so as to answer an issue like “does protein X perform process Y”, a finite set of predictions is required. to get such a finite set, an F-score-based cut-off was applied to the posterior probability. As Arabidopsis has the best coverage of experimental data, this cut-off was adjusted per GO term by comparing Arabidopsis predictions with available experimental data, as previously described [17]: for every GO term, a threshold on the posterior probability was defined that leads to the most F-score for that GO term. All predictions are available online (http://www.ab.wur.nl/bmrf/). the net resource is queried for predictions of proteins or for GO terms of interest, and therefore the results is downloaded in bulk. Queries is supported protein identifiers, organic process GO identifiers or text descriptors of biological processes (Fig. 3). By default, only the foremost detailed Gene Ontology terms (leave terms within the GO structure) are displayed, so as to concentrate on the foremost relevant predictions.

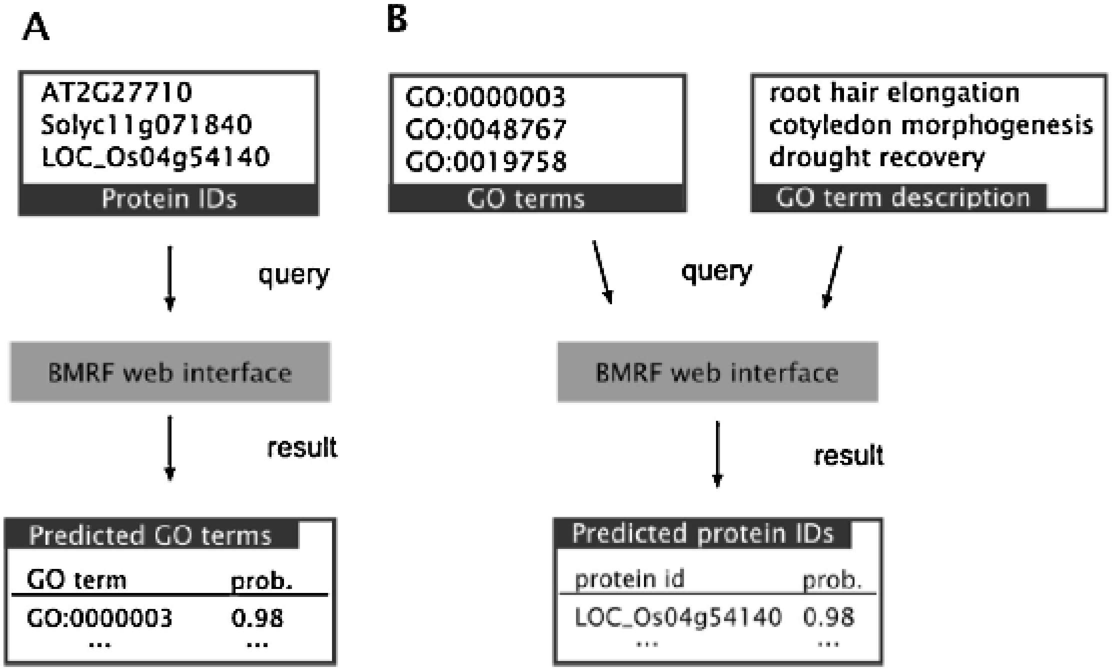

Fig. 3. Use case scenarios for the online interface. Argot2-seeded BMRF results are often queried in two ways. (A) Protein identifiers as query input. The result consists of predicted GO terms for every protein. (B) GO terms (or GO term descriptions) as query input. The result consists of predicted protein identifiers for the relevant GO term(s) and associated posterior probabilities (prob).

The fraction of proteins out of the whole proteome annotated with a minimum of one organic process (annotation coverage) varies considerably between the species: rice shows the best annotation coverage (99%), followed by poplar (77%). Soybean (43%) and barrel clover (39%) show lower coverage. Tomato has rock bottom coverage (12%). Such differences in annotation coverage can have a minimum of two reasons. First, although for each organic process every protein within the input network will have an associated posterior probability, these probabilities are often below the F-score-based cut-off. this suggests that not necessarily every protein within the input network are annotated. additionally, because BMRF only predicts functions for proteins within the input network, the utmost possible annotation coverage is restricted by the quantity of proteins within the respective network. This limit is reflected by the tomato annotation coverage, because the tomato network is that the smallest with 4355 proteins. With exception of soybean, the annotation coverage correlates with the quantity of proteins within the respective network (Table S4).

To investigate differences between available gene function annotation data and Argot2-seeded BMRF, we compared the results with existing protein function predictions from the reference genomes of barrel clover [25], poplar [26], tomato [27], rice [28], and soybean [29]. aside from tomato, the present annotations have much lower coverage than the above-mentioned coverage obtained by Argot2-seeded BMRF (Table S5). the rise of the share of variety of proteins with a minimum of one organic process predicted by our approach varied per species. the share increase ranged from ~60% for rice (24,160 in existing annotation vs. 38,998 in our annotation) to over 100% for poplar (13,682 vs. 32,119).

To complement the above-presented results on coverage, which focused on the question of what number proteins obtain a minimum of one annotation, we also compared the quantity of predicted functions per protein. the common number of GO terms per protein within the available experimental annotation data for Arabidopsis is 4.4. As additional experimental evidence is meant to accumulate, this number should be thought to be a boundary of the common complex quantity of GO terms a protein should be annotated with. Existing sets of predicted annotations for the plant species included here are considerably below this bound, whereas our set of predictions is comparatively near this bound (Table S5). Note that during this assessment, only the foremost granular level of the Gene Ontology is taken under consideration (i.e. only leaf-node terms are considered, and no more general parent terms). For those proteins that existing annotations are available, these annotations are to an oversized extent a subset of what we predict (~80% of the prevailing annotations is additionally predicted by Argot2-seeded BMRF; data not shown). the upper annotation coverage together with the nice prediction performance demonstrates the appreciable added value of the Argot2-seeded BMRF strategy for obtaining gene function annotations.

2.3. Predicted protein functions: showcases

To illustrate the potential of the functions predicted, we screened all predictions for newly annotated biological processes that are considered particularly relevant for the individual species (Table S6). Biological processes considered comprise seed development for rice and soybean; biological process for barrel clover; fruit development for tomato; and lignin related processes for poplar. Inspection of the chosen predictions shows that the functions of proteins tend to become more specific: broadly defined functions are replaced by or augmented with more specific biological processes. for instance, the rice protein LOC_Os10g38080 was previously annotated with bodily structure morphogenesis and is annotated by Argot2-seeded BMRF with seed (coat) development. LOC_Os10g38080 may be a subtilisin homolog which in keeping with available RNAseq data is expressed amongst other reproductive organs and seeds [28]. As additional evidence for the Argot2-seeded BMRF prediction, in Arabidopsis subtilisin and related proteases are involved in reproductive structure development [30]. An example of an annotation for a previously completely annotated protein is LOC_Os05g02520, a cupin domain-containing protein, which was annotated by Argot2-seeded BMRF with seed maturation.

2.4. Divergence and conservation of biological processes in ortholog groups

the set of function predictions delivered above allows us to match function annotation between different plants, a task that's much less easily performed with existing annotations that are derived from various methods which have much lower coverage than our approach. Such a comparison between orthologous genes in several plants allows for assessment of the bounds of orthology-based function prediction, and to investigate gene function evolution.

To characterize ortholog groups with functional predictions that differ from expectations supported sequence similarity, orthologs and paralogs were identified with orthoMCL [31], leading to 25,347 groups (Table S7). Group members that no functions were predicted were removed. To assess the similarity of function predictions within ortholog groups, the mean functional distance within each ortholog group (dubbed 'inner group distance') was calculated (see Section 4). just in case the anticipated biological processes in such a gaggle are different despite high sequence similarity, this is able to be indicative of evolutionary divergence by, e.g., neo-functionalization. to spot such cases, groups with a minimum of four different organisms (6073) were ranked by their largest inner group distance and therefore the most divergent groups (n = 100) were selected. In those groups, biological processes that were significantly overrepresented (more present than randomly expected) were obtained. a spread of biological processes was found (Supp. Figure S1), indicating the widespread occurrence of changes in biological processes proteins are involved in. Most prominent are processes associated with cell development, regulation, and response to a chemical stimulus. For the latter group, the biological processes involved are shown in Fig. 4A. Among the top-ranking groups (with highest 'inner group distance') involved in those processes, we chose as an example a phosphatase with existing experimental annotation in Arabidopsis, PURPLE ACID PHOSPHATASE 26 (PAP26). PAP26 plays a task in phosphate metabolism [32] and phosphate starvation [32] in Arabidopsis. the bulk of the proteins with function predictions within the orthologous group (five out of seven) are indeed predicted by Argot2-seeded BMRF to be involved in phosphate metabolism or the response to phosphate starvation. However, additional function predictions differ. Populus and soybean proteins are predominantly annotated with cell death-related terms; Arabidopsis with pollination and pollen germination processes; tomato with DNA repair and rice with microtubule cytoskeleton organization. This diversity in function isn't reflected by orthology predictions and phylogenetic relationships of the group members (Fig. 4B and C). Independent expression data indicate that Arabidopsis PAP26 is expressed in an exceedingly housekeeping-like manner, but the expression pattern varies between paralogs in other species, e.g. soybean, and to lesser extent orthologs, e.g. between tomato and soybean (Fig. 4D). the various expression patterns give credibility to the variation in function predictions of Argot2-seeded BMRF. this means that PAP26, although its molecular function presumably is invariant, is involved in various biological processes in various plant species. More generally, the analysis of functional divergence presented here highlights the potential of using our set of predicted gene functions for giant scale comparisons between various plant species.

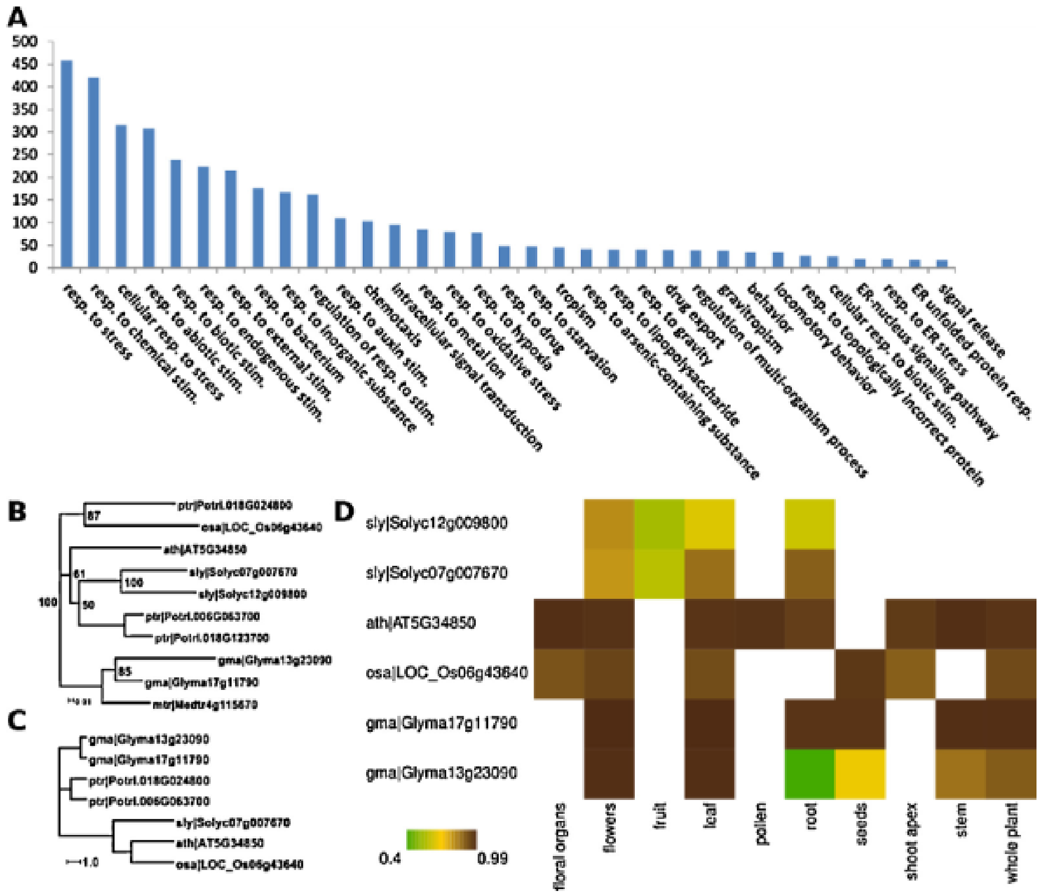

Fig. 4. Comparison between sequence divergence and functional divergence. (A) Overview of the foremost frequent GO terms within the top 100 most functionally divergent ortholog groups that are represented by “response to chemical stimulus” (Figure S1). (B and C) Phylogenetic relations of Arabidopsis PURPLE ACID PHOSPHATASE 26 orthologs. Trees contain Arabidopsis (ath), soybean (GMA), tomato (sly), Populus (ptr), and rice (OSA) PAP26 orthologs. (B) Unrooted phylogenetic tree supported sequence

data. The tree was calculated with 1000 bootstraps. Confidence values are indicated at the branches in percent. (C) Distance tree supported our function predictions. Missing identifiers weren't a part of} the co-expression network and are therefore not part of the functional distance tree. (D) Expression ranking of PURPLE ACID PHOSPHATASE 26 orthologs and paralogs in numerous tissue clusters. The heatmap color represents a mean percentile rank of normalized expression studies aggregated by averaging to 10 tissue clusters (Table S8). Missing data is indicated in white. an summary of the aggregated studies is accessible in Table S8. (For interpretation of the references to paint during this figure legend, the reader is mentioned the net version of this text.)

## 3. Discussion

Finding associations between proteins and biological processes could be a major challenge in non-model plants. Most experimental studies are aimed towards model organisms; hence experiment-based function annotation is sparse within the remainder of sequenced plant genomes. High-throughput experiments to define protein functions are overall less informative than those provided by low-throughput experiments [33]. Moreover, the experimental setup in large-scale approaches might restrict the sort of function annotation which will be obtained. An example is that the characterization of overexpressed rice genes in Arabidopsis [34] to infer function. Here, the matter is that the process of a protein is commonly absolute to the local environment or a particular condition and a special (plant) environment might change the result. Another large scale analysis of gene families in Arabidopsis used prokaryotic gene information to predict function [35]. This semi-manual approach yielded good results for conserved gene families; however, gene families with low conservation weren't covered.

Several computational approaches to protein function annotation exist, albeit mostly not targeted to plants, or model plant species only [36]. An integrated platform like Phytozome [2] provides a standardized set of Gene Ontology annotations for various plant species and hence overcomes the above-mentioned problem that annotations related to genomes are obtained by various methods. However, Phytozome only provides sequence-based predictions. The recently published MORPH algorithm ranked genes for his or her membership of Arabidopsis and tomato pathways, supported a group of known genes from the target pathway, a set of expression profiles, and interaction and metabolic networks [37]. Approaches like PlaNet construct networks supported expression data [24] but such networks don't directly result in gene function annotation. Similarly, a recently presented text mining approach generated networks in Arabidopsis and not gene function annotations [38]. Here we offer a structured approach to extract gene function information from networks and mix that with sequence-based information.

The combination of sequence- and network-based function prediction obtained by seeding BMRF with Argot2, offers a big benefit over applying these methods separately. We validated the rice method and demonstrated greatly improved performance compared to every of the methods separately and compared to Blast2GO. This performance assessment was performed using two complementary indicators, AUC and F-score, which both gave consistent results. Existing annotations provided for the plant genomes to which we applied our method are obtained by various, mostly sequence-based approaches. a transparent description of the methods and input file is usually lacking, resulting in the chance of error propagation and circular reasoning [3], [39]. Our approach has the advantage of applying a regular method to the assorted genomes. Moreover, for several proteins that to this point weren't related to any organic process, we now provide predictions of biological processes. Nevertheless, the mixture of Argot2 and BMRF is indirectly constrained by the experimental data in databases like UniProt [40] or PFAM [41], and by the proteins covered in available networks. it'll however be straightforward to integrate newly available datasets like additional co-expression networks or novel gene function annotations within the framework presented. counting on the supply of a completely unique network or annotation data, we indeed arrange to update our resource. an extra limitation of our current approach is that the structure of the Gene Ontology isn't taken under consideration within the prediction process. Most existing computational methods for gene function prediction suffer from this drawback. it's feasible to create a collection of GO term predictions per the GO-structure [42] and that we commit to apply this method to Argot2-seeded BMRF predictions within the future.

BMRF output consists of a listing of probabilities for every gene to be related to each process. this enables to rank proteins so as of their likelihood of association with a organic process of interest. However, it also can be important to possess a finite set of predictions. to produce that, we applied a cut-off to the chances, supported Arabidopsis, the sole species from which enough data was available. it's difficult to assess how valid the appliance of this cut-off in other plant species is. However, the typical number of predictions per protein that we obtain in each of the species supported the cut-off that was applied is near the observed average for Arabidopsis, giving some credibility to the present cut-off. For one species, tomato, the amount of predicted BP terms per protein is somewhat over the experimentally observed number for Arabidopsis. Hence, argot2-seeded BMRF possibly suffers from overprediction during this case. this might be caused by the upper density (number of interactions compared to the quantity of proteins) of the tomato network. However, in any case, the chances related to the predictions allow narrowing down the prediction results to the foremost reliable ones, if so desired.

With the consistent annotation of multiple plant genomes that we performed, the relation between homology and organic process predictions are often analyzed. Ortholog groups with divergent functions indicated cases where conclusions supported sequence similarity could be inappropriate. Such inappropriate conclusions is also more common than generally acknowledged. Indeed, as recently noted within the context of comparing putative orthologs between species, counting on sequence similarity alone might identify an ortholog with the right molecular function, but will more often than not fail to spot an ortholog that participates within the correct organic process [43]. during a comparison of organic phenomenon patterns between different plant species, the quantity of times that the homolog with the foremost similar pattern of expression (“express log") wasn't also the foremost similar at the sequence level, ranging between 15% and 50% [44]. Similarly, about 1/2 a set of Arabidopsis loss-of-function mutants had only low or moderate phenotypic similarity with mutants of putative orthologs in tomato, rice, or maize [45]. Large-scale evolutionary comparisons between plant species, as an example, geared toward identifying patterns within the retention of duplicated genes [46], [47], or functional biases in single-copy genes [7], are currently performed supported function annotations obtained using sequence similarity. Such studies will enjoy the gene annotations presented here, which overcome the constraints of purely sequence-based annotation of gene functions.

In the example of PAP26 homologs, homology captures the molecular function, but at the process level, there's divergence. Our integrated sequence- and network-based function annotation method allows us to predict such divergent biological processes. Differences in expression between the various PAP26 homologs in several species provide additional evidence for our function predictions. More generally, the results on process divergence are in line with the concept that evolution acts specifically by "tinkering" with genes, coopting available components of a genome for brand spanking new processes.

The combination of sequence-based and network-based predictions could be a huge improvement for sparsely annotated plant genomes. With the arrival of RNA-seq [48] coexpression network-based protein function prediction can become a preferred method. Combined with additional analysis, like genome-wide association studies (GWAS), potential candidate genes for traits-of-interest might be identified more reliably. Such candidate genes are of great help in applications associated with plant breeding. the power to associate unannotated proteins to particular biological processes will spark experimental work and be essential for the advancement of the understanding of gene function in plant genome evolution.

## 4. Materials and methods

4.1. Function prediction methods and their integration

BMRF uses network data as input. Each protein is represented as a node within the network, and connections within the network indicate functional relationships between proteins. A statistical model (Markov Random Field) describes how the involvement of a protein during a particular BP influences the probability that its neighbors within the network also are involved in this BP. The parameters within the statistical model describe for every BP how strongly neighbors influence one another. Parameter values are trained employing a Bayesian approach by performing the simultaneous estimation of the model parameters and prediction of protein functions. This strategy needs a collection of known protein functions because the initial labeling of the network. Argot2 could be a purely sequence-based prediction method, using searches of the UniProt and Pfam databases as input. to mix these two methods, two strategies were applied. within the first integration method, for every organic process, ranks for the various proteins were obtained from both BMRF and Argot2, by ordering the proteins supported their score for that process. These ranks were added to get a final ranking, which was used because the prediction score for that organic process. during a second integration strategy, initial predictions were generated with Argot2. These were supplied to BMRF as training data, meaning that the initial labeling of the nodes within the network was supported the Argot2 predictions.

4.2. Sequence and domain data

Sequence data for Arabidopsis, rice, soybean, and M. truncatula were obtained from the Phytozome database v8.0 [2]. Poplar sequence data were downloaded from the JGI (ftp://ftp.jgi-psf.org/pub/JGI_data/Poplar/annotation/v1.1), annotation version 1.1. Tomato sequence v2.4 and annotation v2.3 data [27] were retrieved from the SGN network (http://www.solgenomics.net). Arabidopsis Interpro domains were retrieved from TAIR10 [49]. Domains of transcript isoforms were merged into one set per gene.

4.3. Function annotation data

Annotations from the Gene Ontology project, version 1.1418 [9], and from Gramene

[50], were used as input for training and cross-validation. Annotations from Oryzabase version 4 [23] were used as an independent validation set. Only genes that no annotation was available within the data from the Gene Ontology project were used for validation. altogether cases, only process (BP) terms with evidence codes IDA (inferred from direct assay), IGI (genetic interaction), and IMP (mutant phenotype) were used.

### 4.4. Network data

Co-expression networks supported microarray data for Arabidopsis, rice, G. max, M. truncatula, and poplar were obtained from PlaNet [24]. For tomato, a recently published microarray-based co-expression network [51] was used. The probe of the tomato co-expression network was obtained from Affymetrix (http://www.affymetrix.com) and mapped with BLAST v2.2.26 [11] to the tomato protein sequences. Further network data for Arabidopsis and rice was obtained from FunctionalNet (http://www.functionalnet.org/) [19] and STRING [18]. Arabidopsis yeast-two-hybrid data were acquired from the literature [52]. The rice-Arabidopsis interspecies network was generated by using BLAST (cut-off on E-value of 1e−4). BMRF requires all proteins to be a part of the input network. Thus, proteins not contained within the input network were removed. altogether cases, the longest isoform of alternatively spliced variants was used.

### 4.5. Validation setup

Performance assessment was performed with rice. HMMER v3 (http://hmmer.org/) search against PFAM [41] and BLAST [11] alignment against UniProt [53] were wont to generate the input for Argot2 [16]. within the context of the validation setup, all rice proteins were far from the UniProt database to avoid Argot2 using information from those proteins.

For comparison, sequence similarity-based annotation was disbursed with Blast2GO [21]. Rice protein sequences were queried against the non-redundant a part of GenBank (NR) [54], using an E-value cut-off of 1e−4. within the context of the validation setup, hits to monocot proteins in NR were far from the BLAST results before supplying them to Blast2GO.

Prediction runs of various methods and network combinations were assessed with 100 cross-validation runs. In each run, randomly, a subset (n = 200) of proteins was chosen and also the annotation was removed (masked). for each run, predicted functions were compared with the masked ones. Only process terms with a minimum of three masked proteins were utilized in the performance assessment to permit for sufficient statistics. within the performance assessment, negative cases consisted of gene-BP associations that weren't annotated intrinsically within the experimental data.

The performance was assessed by the realm under the receiver operating graphical record (AUC) and therefore the F-score. The AUC is that the area under the curve of 1-specificity vs. sensitivity and is adequate to the probability that a classifier will rank a randomly chosen positive instance on top of a randomly chosen negative one [20]. Specificity is that the fraction of proteins experimentally known to not perform a given function that's indeed not predicted to try to to so, whereas sensitivity (or recall) is that the fraction of proteins experimentally known to perform a given function that's indeed predicted to try and do so. F-score relies on the precision-recall (precision vs. sensitivity) curve. Precision is that the fraction of proteins predicted to perform a given function that's indeed experimentally known to try to to so. The F-score is capable the mean of precision and recall, and therefore the maximum value of the F-score (Fmax-score) was used for every organic process.

To obtain a finite set of predictions, functions of a protein were assigned by using an F-score-based cut-off. The F-score was calculated per GO term and its maximum (Fmax-score), calculated with Arabidopsis data as previously described [17], was wont to set a cut-off on the posterior probability. the brink obtained with Arabidopsis data was employed in the opposite species because, in those species, too few annotations are available to get a species-specific threshold. All performance measures were calculated with the R-package ROCR [55] and custom R-scripts.

4.6. Application setup

Function annotations predicted for barrel clover, poplar, rice, soybean, and tomato were compared with existing predictions in terms of coverage of proteins and therefore the number of predicted functions per protein. Barrel clover, poplar, and rice process predictions were obtained from the official genome annotations version Mt3.5v5 [25], v1.1 [26], and v7.0 [28], respectively. Soybean annotation was obtained from Phytozome [2]. Tomato function annotation data was extracted from the ITAG annotation v2.3 [27].

To determine the entire number of proteins and also the total number of GO terms that annotations were obtained, the annotation of every protein was expanded by including the parent GO terms of all assigned GO terms. For the calculation of the quantity of annotations per protein, only the leaf-terms of the Gene Ontology were included.

4.7. Evolutionary and functional distance calculation

Groups of orthologs were predicted with OrthoMCL [31]. To calculate functional divergence, BMRF posterior probabilities for every protein were interpreted as vectors. The Euclidian distance for every combination of proteins within a gaggle of orthologs was calculated. The mean of distances within a bunch (inner group distance) was wont to rank groups of orthologs. For the PAP26 example, only groups with existing experimental annotation in Arabidopsis were taken into consideration. The PAP26 tree was estimated with RaxML version 7.2.8-ALPHA [56] using the PROTGAMMAJJTF substitution model and 1000 bootstraps. Expression data for PAP26 was obtained from the AtGenExpress developmental set [57]; publicly available RNA-seq datasets from tomato (S. Lycopersicum cv. Heinz 1706; data SRA049915) were retrieved from the SRA database (http://www.ncbi.nlm.nih.gov/sra). Reads were mapped with GSNAP [58] against the tomato reference genome (v. 2.40, Sato et al. [27]) and also the expression made up our minds with cufflinks [59] with default parameters. Soybean expression data was obtained from SoyBase [60]. Rice expression data were obtained from the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/). All expression experiment data were z-score normalized and percentile ranked to facilitate comparison. Replicates were merged by averaging over the expression for every gene.

Acknowledgments

We thank Dr. Yiannis Kourmpetis for his or her helpful discussions. This work was supported by the FP7 “Infrastructures” project TransPLANT Award 283496 and by the BioRange program of Kingdom of The Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through Kingdom of The Netherlands Genomics Initiative (NGI).

References

[1] M.C. Schatz, J. Witkowski, W.R. McCombie
Current challenges in de novo plant genome sequencing and assembly Genome Biol., 13 (2012), p. 243, 10.1186/gb4015
View Record in ScopusGoogle Scholar.

[2] D.M. Goodstein, S. Shu, R. Howson, R. Neupane, R.D. Hayes, et al. Phytozome: a comparative platform for green plant genomics Nucleic Acids Res., 40 (2012), pp. D1178-D1186, 10.1093/nar/gkr944
CrossRefView Record in ScopusGoogle Scholar.

[3] L. Du Plessis, N. Skunca, C. Dessimoz
The what, where, how, and why of gene ontology – a primer for bioinformaticians
Brief Bioinform., 12 (2011), pp. 723-735, 10.1093/bib/bbr002
CrossRefView Record in ScopusGoogle Scholar.

[4] S. De Bodt, J. Hollunder, H. Nelissen, N. Meulemeester, D. Inzé CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations, and functional annotations

New Phytol., 195 (2012), pp. 707-720, 10.1111/j.1469-8137.2012.04184.x
CrossRefView Record in ScopusGoogle Scholar.

[5] M. Van Bel, S. Proost, E. Wischnitzki, S. Movahedi, C. Scheerlinck, et al. Dissecting plant genomes with the PLAZA comparative genomics platform Plant Physiol., 158 (2012), pp. .590-600, 10.1104/pp.111.189514
CrossRefView Record in ScopusGoogle Scholar.

[6] R. Monclus, J.-C. Leplé, C. Bastien, P.-F. Bert, M. Villar, et al.
Integrating genome annotation and QTL position to spot candidate genes for productivity, architecture, and water-use efficiency in Populus spp.
BMC Plant Biol., 12 (2012), p. 173, 10.1186/1471-2229-12-173
CrossRefView Record in ScopusGoogle Scholar.

[7]R. De Smet, K.L. Adams, K. Vandepoele, M.C.E. Van Montagu, S. Maere, et al.
Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants
Proc. Natl. Acad. Sci. U. S. A., 110 (2013), pp. 2898-2903, 10.1073/pnas.1300127110
CrossRefView Record in ScopusGoogle Scholar.

[8] S.Y. Rhee, M. Mutwil
Towards revealing the functions of all genes in plants
Trends Plant Sci. (2013), 10.1016/j.tplants.2013.10.006
Google Scholar.

[9] Gene Ontology Consortium
Gene Ontology: a tool for the unification of biology
Nat. Genet., 25 (2000), pp. 25-29, 10.1038/75556
Google Scholar.

[10] P. Radivojac, W.T. Clark, T.R. Oron, A.M. Schnoes, T. Wittkop, et al. A large-scale evaluation of computational protein function prediction
Nat. Methods, 10 (2013), pp. 221-227, 10.1038/nmeth.2340
CrossRefView Record in ScopusGoogle Scholar.

[11]S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman
The basic local alignment search tool, J. Mol. Biol., 215 (1990), pp. 403-410, 10.1016/S0022-2836(05)80360-2
ArticleDownload PDFView Record in ScopusGoogle Scholar.

[12]D.M.A. Martin, M. Berriman, G.J. Barton
GOtcha: a replacement method for prediction of protein function assessed by the annotation of seven genomes
BMC Bioinform., 5 (2004), p. 178, 10.1186/1471-2105-5-178
CrossRefView Record in ScopusGoogle Scholar.

[13] W.T. Clark, P. Radivojac
Analysis of protein function and its prediction from aminoalkanoic acid sequence Proteins, 79 (2011), pp. 2086-2096, 10.1002/prot.23029
CrossRefView Record in ScopusGoogle Scholar.

[14] A. Vazquez, A. Flammini, A. Maritan, A. Vespignani
Global protein function prediction from protein-protein interaction networks, Nat. Biotechnol., 21 (2003), pp. 697-700, 10.1038/nbt825
View Record in ScopusGoogle Scholar.

[15]Y.A.I. Kourmpetis, A.D.J. van Dijk, M.C.A.M. Bink, R.C.H.J. van Ham, C.J.F. ter Braak, Bayesian Markov Random Field analysis for protein function prediction supported network data
PLoS ONE, 5 (2010), p. e9293, 10.1371/journal.pone.0009293

CrossRefView Record in ScopusGoogle Scholar.

[16]M. Falda, S. Toppo, A. Pescarolo, E. Lavezzo, B. Di Camillo, et al.Argot2: an out-sized scale function prediction tool looking forward to semantic similarity of weighted Gene Ontology terms
BMC Bioinform., 13 (Suppl. 4) (2012), p. S14, 10.1186/1471-2105-13-S4-S14
CrossRefView Record in ScopusGoogle Scholar.

[17]Y.A.I. Kourmpetis, A.D.J. van Dijk, R.C.H.J. van Ham, C.J.F. ter Braak, Genome-wide computational function prediction of Arabidopsis proteins by integration of multiple data sources
Plant Physiol., 155 (2011), pp. 271-281, 10.1104/pp.110.162164
CrossRefView Record in ScopusGoogle Scholar.

[18] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, et al.
The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored
Nucleic Acids Res., 39 (2011), pp. D561-D568, 10.1093/nar/gkq973
CrossRefView Record in ScopusGoogle Scholar.

[19]I. Lee, B. Ambaru, P. Thakkar, E.M. Marcotte, S.Y. Rhee
Rational association of genes with traits employing a genome-scale gene network for Arabidopsis thaliana
Nat. Biotechnol., 28 (2010), pp. 149-156, 10.1038/nbt.1603
CrossRefView Record in ScopusGoogle Scholar

[20]J.A. Hanley, B.J. McNeil
The meaning and use of the realm under a receiver operating characteristic (ROC) curve
Radiology, 143 (1982), pp. 29-36

CrossRefView Record in ScopusGoogle Scholar

[21]A. Conesa, S. Götz, J.M. García-Gómez, J. Terol, M. Talón, et al.
Blast2GO: a universal tool for annotation, visualization, and analysis
in genomics research
Bioinformatics, 21 (2005), pp. 3674-3676, 10.1093/bioinformatics/bti610
CrossRefView Record in ScopusGoogle Scholar

[22]J. Davis, M. Goadrich
The relationship between precision-recall and ROC curves
Proceedings of the 23rd International Conference on Machine Learning – ICML'06,
ACM Press, New York, NY, USA (2006), pp. 233-240, 10.1145/1143844.1143874
CrossRefView Record in ScopusGoogle Scholar

[23]N. Kurata, Y. Yamazaki
Oryzabase. An integrated biological and genome information database for rice
Plant Physiol., 140 (2006), pp. 12-17, 10.1104/pp.105.063008
CrossRefView Record in ScopusGoogle Scholar

[24]M. Mutwil, S. Klie, T. Tohge, F.M. Giorgi, O. Wilkins, et al.
PlaNet: combined sequence and expression comparisons across plant networks derived from seven species
Plant Cell, 23 (2011), pp. 895-910, 10.1105/TPC.111.083667
CrossRefView Record in ScopusGoogle Scholar

[25]N.D. Young, F. Debellé, G.E.D. Oldroyd, R. Geurts, S.B. Cannon, et al.
The Medicago genome provides insight into the evolution of rhizobial symbioses
Nature, 480 (2011), pp. 520-524, 10.1038/nature10625
CrossRefView Record in ScopusGoogle Scholar

[26]G.A. Tuskan, S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, et al.
The genome of black cottonwood, Western balsam poplar (Torr. & Gray)
Science, 313 (2006), pp. 1596-1604, 10.1126/science.1128691
CrossRefView Record in ScopusGoogle Scholar

[27]S. Sato, S. Tabata, H. Hirakawa, E. Asamizu, K. Shirasawa, et al.
The tomato genome sequence provides insights into fleshy fruit evolution
Nature, 485 (2012), pp. 635-641, 10.1038/nature11119
View Record in ScopusGoogle Scholar

[28]S. Ouyang, W. Zhu, J. Hamilton, H. Lin, M. Campbell, et al.
The TIGR rice genome annotation resource: improvements and new features
Nucleic Acids Res., 35 (2007), pp. D883-D887, 10.1093/nar/gkl976
CrossRefView Record in ScopusGoogle Scholar

[29]J. Schmutz, S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, et al.
Genome sequence of the palaeopolyploid soybean
Nature, 463 (2010), pp. 178-183, 10.1038/nature08670
CrossRefView Record in ScopusGoogle Scholar

[30]C. Rautengarten, B. Usadel, L. Neumetzler, J. Hartmann, D. Büssis, et al.
A subtilisin-like serine protease essential for mucilage release from Arabidopsis seed coats
Plant J., 54 (2008), pp. 466-480, 10.1111/j.1365-313X.2008.03437.x
View Record in ScopusGoogle Scholar

[31]L. Li, C.J. Stoeckert, D.S. Roos
OrthoMCL: identification of ortholog groups for eukaryotic genomes
Genome Res., 13 (2003), pp. 2178-2189, 10.1101/gr.1224503
CrossRefView Record in ScopusGoogle Scholar

[32]B.A. Hurley, H.T. Tran, N.J. Marty, J. Park, W.A. Snedden, et al.
The dual-targeted purple acid phosphatase isozyme AtPAP26 is important for efficient acclimation of Arabidopsis to nutritional phosphate deprivation
Plant Physiol., 153 (2010), pp. 1112-1122, 10.1104/pp.110.153270
CrossRefView Record in ScopusGoogle Scholar

[33]A.M. Schnoes, D.C. Ream, A.W. Thorman, P.C. Babbitt, I. Friedberg
Biases within the experimental annotations of protein function and their effect on our understanding of protein function space
PLoS Comput. Biol., 9 (2013), p. e1003063, 10.1371/journal.pcbi.1003063
CrossRefView Record in ScopusGoogle Scholar

[34]T. Sakurai, Y. Kondou, K. Akiyama, A. Kurotani, M. Higuchi, et al.
RiceFOX: a database of Arabidopsis mutant lines overexpressing rice full-length cDNA that contains a large range of trait information to facilitate analysis of gene function
Plant Cell Physiol., 52 (2011), pp. 265-273, 10.1093/PCP/pcq190
CrossRefView Record in ScopusGoogle Scholar

[35]S. Gerdes, B. El Yacoubi, M. Bailly, I.K. Blaby, C.E. Blaby-Haas, et al.
Synergistic use of plant–prokaryote comparative genomics for functional annotations
BMC Genomics, 12 (Suppl. 1) (2011), p. S2, 10.1186/1471-2164-12-S1-S2
CrossRefView Record in ScopusGoogle Scholar

[36]I. Lee, Y.-S. Seo, D. Coltrane, S. Hwang, T. Oh, et al.
Genetic dissection of the biotic stress response employing a genome-scale gene network for rice
Proc. Natl. Acad. Sci. U. S. A., 108 (2011), pp. 18548-18553, 10.1073/pnas.1110384108
CrossRefView Record in ScopusGoogle Scholar

[37]O. Tzfadia, D. Amar, L.M.T. Bradbury, E.T. Wurtzel, R. Shamir
The MORPH algorithm: ranking candidate genes for membership in Arabidopsis and tomato pathways
Plant Cell, 24 (2012), pp. 4389-4406, 10.1105/TPC.112.104513
CrossRefView Record in ScopusGoogle Scholar

[38]G. Blanc, K.H. Wolfe
Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution
Plant Cell, 16 (2004), pp. 1679-1691, 10.1105/TPC.021410
CrossRefView Record in ScopusGoogle Scholar

[39]B.E. Engelhardt, M.I. Jordan, J.R. Srouji, S.E. Brenner
Genome-scale phylogenetic function annotation of huge and diverse protein families
Genome Res., 21 (2011), pp. 1969-1980, 10.1101/gr.104687.109
CrossRefView Record in ScopusGoogle Scholar

[40]E.C. Dimmer, R.P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, et al.
The UniProt-GO annotation database in 2011
Nucleic Acids Res., 40 (2012), pp. D565-D570, 10.1093/nar/gkr1048
CrossRefView Record in ScopusGoogle Scholar

[41]R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, et al.
The Pfam protein families database
Nucleic Acids Res., 38 (2010), pp. D211-D222, 10.1093/nar/gkp985
CrossRefView Record in ScopusGoogle Scholar

[42]Y.A. Kourmpetis, A.D. van Dijk, C.J. Ter Braak
Gene Ontology consistent protein function prediction: the FALCON algorithm applied to

6 eukaryotic genomes
Algorithms Mol. Biol., 8 (2013), p. 10, 10.1186/1748-7188-8-10
CrossRefView Record in ScopusGoogle Scholar

[43]S. Netotea, D. Sundell, N.R. Street, T.R. Hvidsten
ComPlEx: conservation and divergence of co-expression networks in a very. thaliana, Populus and O. Sativa
BMC Genomics, 15 (2014), p. 106, 10.1186/1471-2164-15-106
CrossRef View Record in Scopus Google Scholar.

## AUTHOR AFFILIATION

KANJI, E.E;

ODE P.O.