Current Plant Biology 1 (2014) 73–82

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

Biological process annotation of proteins across the plant kingdom

Joachim W. Bargsten[a,c,e], Edouard I. Severing[b], Jan-Peter Nap[a,c], Gabino F. Sanchez-Perez[a,d], Aalt D.J. van Dijk[a,f,*]

[a] Applied Bioinformatics, Bioscience, Plant Sciences Group, Wageningen University and Research Centre, Wageningen, The Netherlands
[b] Laboratory of Genetics, Plant Sciences Group, Wageningen University and Research Centre, Wageningen, The Netherlands
[c] Netherlands Bioinformatics Centre (NBIC), Nijmegen, The Netherlands
[d] Laboratory of Bioinformatics, Plant Sciences Group, Wageningen University and Research Centre, Wageningen, The Netherlands
[e] Laboratory for Plant Breeding, Plant Sciences Group, Wageningen University and Research Centre, The Netherlands
[f] Biometris, Wageningen University and Research Centre, Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 20 February 2014
Received in revised form 15 July 2014
Accepted 26 July 2014

Keywords:
Gene function prediction
Gene function divergence

ABSTRACT

Accurate annotation of protein function is key to understanding life at the molecular level, but automated annotation of functions is challenging. We here demonstrate the combination of a method 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 is based on co-expression networks and combines the network-based prediction method BMRF with the sequence-based prediction method Argot2. The combination shows significantly improved performance compared to each of the methods separately, as well as compared to Blast2GO. 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 in which proteins are involved, in spite of sequence similarity. This indicates that the integration of network-based and sequence-based function prediction is helpful towards the analysis of evolutionary relationships. Examples of potential divergence are identified for various biological processes, notably for processes related to cell development, regulation, and response to chemical stimulus. Such divergence in biological process annotation for proteins with similar sequences should be taken into account when analyzing plant gene and genome evolution.

DATA: All gene functions predictions are available online (http://www.ab.wur.nl/bmrf/). The online resource can be queried for predictions of proteins or for Gene Ontology terms of interest, and the results can be downloaded in bulk. Queries can be based on protein identifiers, biological process Gene Ontology identifiers, or text descriptors of biological processes.

1. Introduction

The amount of plant genome data grows disproportional to the amount of available experimental data on these genomes [1–5]. To connect this ever increasing amount of genome data to plant biology, structural gene annotation followed by function annotation is imperative. For example, the identification of candidate genes involved in a trait of interest greatly benefits from gene function annotation [6]. In the context of the study of genome evolution, gene function annotations are necessary in order 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, can be used to annotate gene or protein function. However, the scarcity of experimental data highlights the attractiveness of computational approaches to assist in gene function annotation [8]. Indeed, newly sequenced genomes are in general accompanied by a function annotation which heavily relies on computational predictions. Such automated annotations are delivered by a variety of approaches, often without much knowledge about their reliability. For studying plant genomes and plant genome evolution, reliable function annotation is therefore a major challenge.

One way to annotate proteins without experimental data is to infer function from sequence data [3]. The *de facto* standard

[*] Corresponding author at: Applied Bioinformatics, Bioscience, Plant Sciences Group, Wageningen University and Research Centre, Wageningen, The Netherlands. Tel.: +31 317480994.
E-mail address: aaltjan.vandijk@wur.nl (A.D.J. van Dijk).

http://dx.doi.org/10.1016/j.cpb.2014.07.001

to capture function annotation today is the Gene Ontology (GO), in particular, the Molecular Function (MF) and Biological Process (BP) sub-ontologies [9]. MF describes activities, such as 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 in the BP ontology are generally associated with more conceptual levels of function; BP terms describe the execution of one or more molecular function instances working together to accomplish a certain biological objective. The prediction of BP terms can depend on the cellular and organismal context [10]. Therefore, BP terms tend to be poorly predicted by methods based on sequence similarity only, such as BLAST [10,11]. The reliability of BP predictions increases with advanced approaches that employ, e.g., phylogenetic frameworks [12,13] or network data such as 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 in the network, and connections in the network indicate functional relationships between proteins. Networks can be based on, e.g., protein–protein interactions or co-expression data. BMRF uses existing BP annotations for proteins in the network to infer biological processes for unannotated proteins in that network. To do so, BMRF uses a statistical model describing how likely neighbors are to participate in the same BP; this constitutes the Markov Random Field. Existing BP annotations are used as “seed” or “training” data, providing a set of initial labels for the Markov Random Field. Parameters in the statistical model are trained using a Bayesian approach by performing simultaneous estimation of the model parameters and prediction of protein functions. Importantly, BMRF can transfer functional information beyond direct interactions. Therefore, it is able to generate function predictions for proteins that are only linked with other proteins with unknown function.

In the Critical Assessment of Function Annotations (CAFA) protein function prediction challenge [10] BMRF obtained particularly good performance in human (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. Because of its dependence on training data, function annotation for species with more sparse function annotation is challenging for BMRF. To improve the prediction performance in sparsely annotated species, we present here a strategy to combine BMRF with the sequence-based function prediction method Argot2 [16]. Argot2 was among the top performing sequence-based algorithms in the CAFA category “eukaryotic BP”. In its computational approach Argot2 is complementary to BMRF, because it is purely sequence-based.

We demonstrate that a combination of Argot2 and BMRF has 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), *Oryza sativa* (rice), *Populus trichocarpa* (poplar), *Glycine max* (soybean) and *Solanum lycopersicum* (tomato), using microarray co-expression networks as input. Numerous new proteins were associated with specific biological processes, such as seed development in rice or nitrogen fixation 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 the resulting set of predicted gene functions will be of great value in capitalizing on the large amount of plant genome data that is 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 *Arabidopsis thaliana* [17]. This method relies, besides on network data, on existing function annotation as input. For Arabidopsis, we demonstrated that the amount of available annotation (training) data was sufficient to achieve a good prediction performance [17]. However, for crop species, much less annotation data is available as input. To increase the overall 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 largest amount of annotation data available: 415 proteins with experimental evidence for a biological process. The rice network used as input for BMRF was obtained from a combination 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 in the network, and were used for validation of predicted functions.

Function prediction performance was assessed on the basis of cross-validation, leaving out randomly selected proteins with known function and comparing the predictions with those data. The area under the receiver operator characteristic curve (AUC) was used to compare the performance of the predictions that come as ordered lists of predicted proteins per biological process. In the standalone setting (Fig. 1A and B) with rice sequence and network data, BMRF and Argot2 both have a low 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 lower than the AUC previously obtained with BMRF for Arabidopsis (0.75) [17] due to the small amount of training data (annotated gene functions) that is 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 in the rice network based on sequence similarity using BLAST. With this rice-Arabidopsis interspecies network in addition 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 used to create the interspecies network did not influence the performance of BMRF (data not shown).

Both methods use complimentary information about biological processes (network input for BMRF, sequence input for Argot2). Therefore, we tested combining the two. Argot2 and BMRF can be combined in multiple ways. We used a simple rank-based approach

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)

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

[b] Area under the curve; mean (standard deviation).

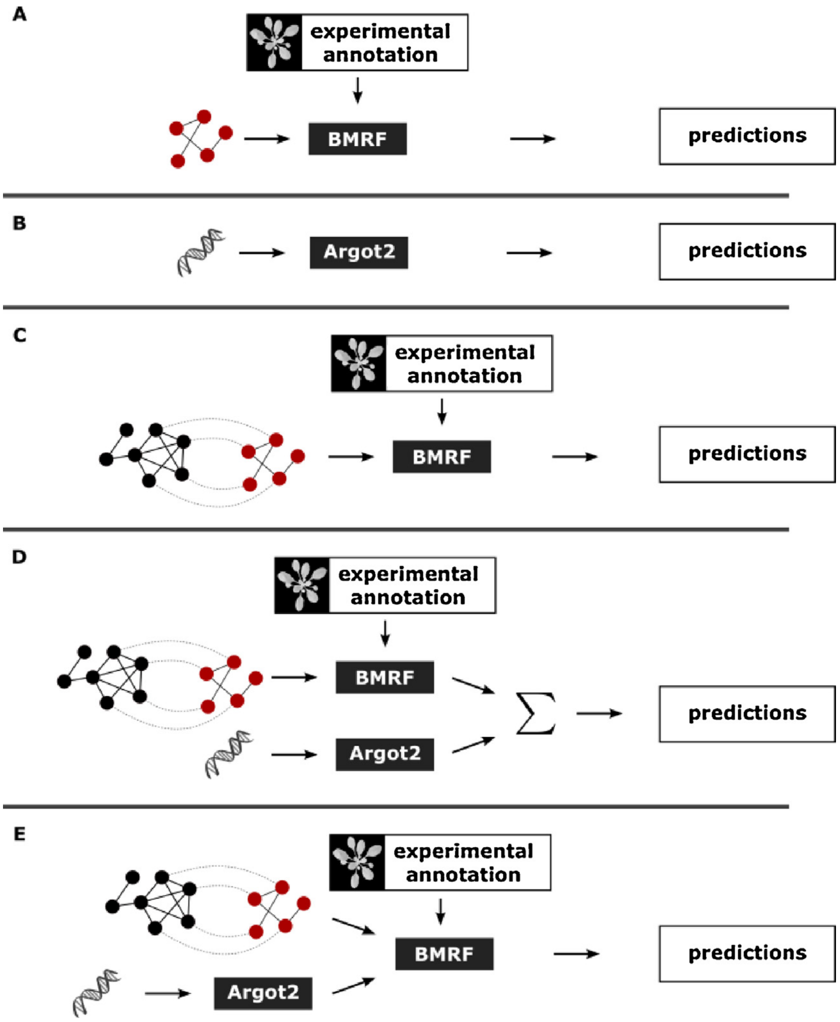

Fig. 1. Strategies for predicting protein function. BMRF (A and C) and Argot2 (B) were used in a standalone setting or in two different combinations (D and E). Combining BMRF and Argot2 was done by combining the results of each of the two methods (D), and by using Argot2 predictions as input for BMRF (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 color in this figure legend, the reader is referred to the web version of this article.)

to predict biological processes by ordering Argot2 and BMRF results separately and then combining their ranks to produce a final rank (Fig. 1D). This integration was performed for each biological process separately by sorting the proteins based on 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 did not improve results compared to standalone BMRF (Table 1). Performance was markedly improved, however, by generating initial predictions with Argot2 and supplying these to BMRF as training data (seed data; Fig. 1E). In this integration method, the initial labelling of proteins in the network (i.e. the seed data for BMRF) was based on 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 different thresholds on the performance of BMRF, BMRF was seeded with 5 different output sets of Argot2 (Table S3). The best performance was achieved with the default threshold of 5.

The results above indicate that our integrated method performed markedly better than each of the two methods separately. As additional assessment of performance, we predicted annotations with the often-used method Blast2GO [21]. The resulting AUC of Blast2GO was 0.72 ± 0.13, and 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 small number of experimentally verified annotations (true positives) and high number of unannotated proteins (true negatives) could introduce a skew in the cross-validation sets, leading to a bias in the AUC performance assessment [22]. The *F*-score (harmonic mean of precision and recall) does not suffer from this skew and the final prediction performance was therefore also assessed with the maximum *F*-score (F_{max}-score). In agreement with the AUC evaluation, the F_{max}-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 F_{max}-score values for different BP terms in different cross-validation runs confirms the performance difference between the combined Argot2-BMRF predictions and Blast2GO (Fig. 2B and C).

To obtain independent validation in addition to the cross-validation performed above, the Argot2-seeded BMRF predictions were compared to annotations available in the Oryzabase database [23] which were not present in our input data (71 proteins). The AUC of 0.88 ± 0.13 we obtained was similar to the AUC obtained in the cross-validation, confirming the performance assessment. Overall, the performance evaluation demonstrates that Argot2-seeded BMRF is an effective way to predict BP protein function in sparsely annotated plant genomes.

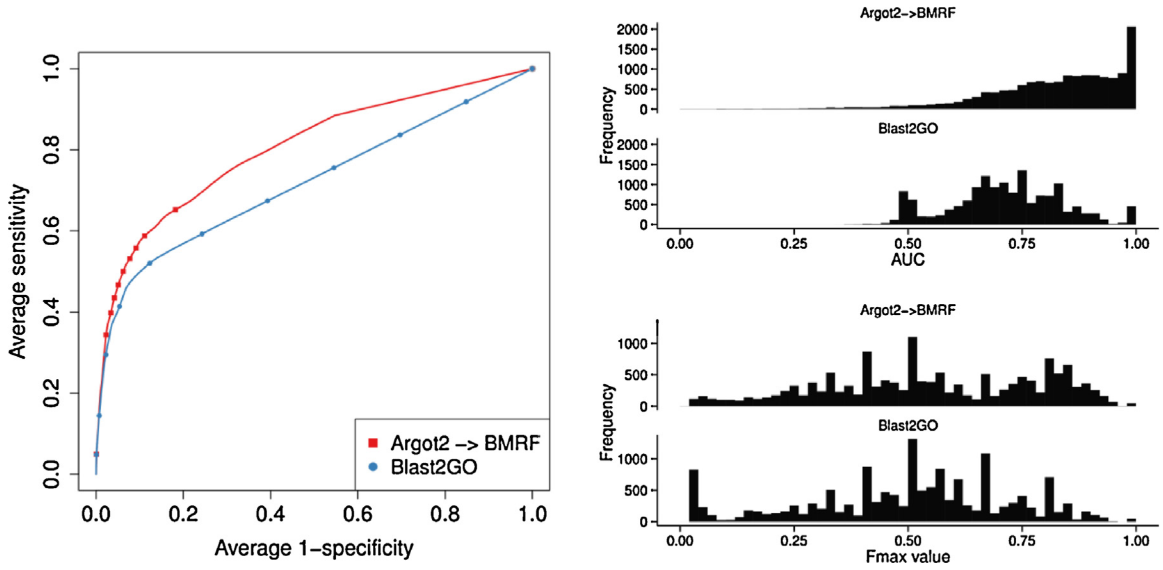

Fig. 2. Performance assessment of function prediction on rice proteins. (A) Receiver operator characteristic curve showing 1-specificity vs. sensitivity of the predictions of Argot2-seeded BMRF and Blast2GO. Specificity and sensitivity were averaged over all cross-validation runs. Dots indicate evenly spaced intervals of the underlying prediction score, line represents complete curve. Performance is summarized as AUC which is the area under these curves. (B) Histogram of AUC values per GO term of every cross-validation run calculated for Argot2-seeded BMRF and Blast2GO. (C) Histogram of $F_{\max}$ values per GO term of every cross-validation run calculated for Argot2 seeding BMRF and Blast2GO.

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 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 associated with a certain GO term was estimated for all GO terms and all proteins in the network. In order to answer a question such as "does protein X perform biological process Y", a finite set of predictions is needed. To obtain such finite set, an *F*-score-based cut-off was applied to the posterior probability. As Arabidopsis has the highest 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 each GO term, a threshold on the posterior probability was defined that results in the maximum *F*-score for that GO term. All predictions are available online (http://www.ab.wur.nl/bmrf/). The online resource can be queried for predictions of proteins or for GO terms of interest, and the results can be downloaded in bulk. Queries can be based on protein identifiers, biological process GO identifiers, or text descriptors of biological processes (Fig. 3). By default, only the most detailed Gene Ontology terms (leave terms in the GO structure) are displayed, in order to focus on the most relevant predictions.

The fraction of proteins out of the complete proteome annotated with at least one biological process (annotation coverage) varies considerably between the species: rice shows the highest annotation coverage (99%), followed by poplar (77%). Soybean (43%) and

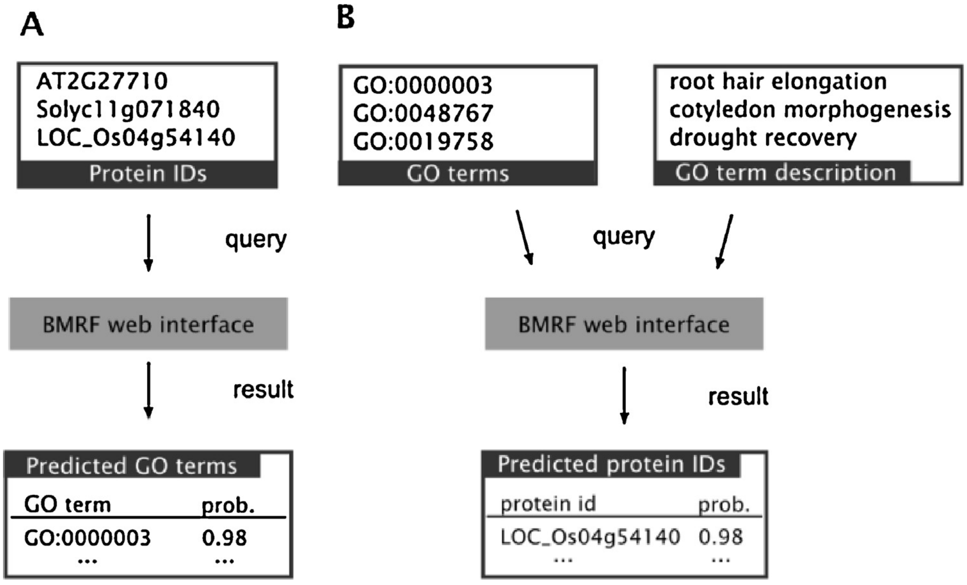

Fig. 3. Use case scenarios for the web interface. Argot2-seeded BMRF results can be queried in two ways. (A) Protein identifiers as query input. The result consists of predicted GO terms for each protein. (B) GO terms (or GO term descriptions) as query input. The result consists of predicted protein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

barrel clover (39%) show lower coverage. Tomato has the lowest coverage (12%). Such differences in annotation coverage can have at least two reasons. First, although for every biological process every protein in the input network will have an associated posterior probability, these probabilities can be below the *F*-score-based cut-off. This means that not necessarily every protein in the input network will be annotated. In addition, because BMRF only predicts functions for proteins in the input network, the maximum possible annotation coverage is limited by the number of proteins in the respective network. This limit is reflected by the tomato annotation coverage, as the tomato network is the smallest with 4355 proteins. With exception of soybean, the annotation coverage correlates with the number of proteins in 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]. Except for tomato, the existing annotations have a much lower coverage than the above mentioned coverage obtained by Argot2-seeded BMRF (Table S5). The increase of percentage of number of proteins with at least one biological process predicted by our approach varied per species. The percentage 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 how many proteins obtain at least one annotation, we also compared the number of predicted functions per protein. The average number of GO terms per protein in the available experimental annotation data for Arabidopsis is 4.4. As additional experimental evidence is supposed to accumulate, this number should be regarded as a lower bound of the average real number 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 relatively close to this bound (Table S5). Note that in this assessment, only the most granular level of the Gene Ontology is taken into account (i.e. only leaf-node terms are considered, and not more general parent terms). For those proteins for which existing annotations are available, these annotations are to a large extent a subset of what we predict (~80% of the existing annotations is also predicted by Argot2-seeded BMRF; data not shown). The higher annotation coverage in combination with the good 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; nitrogen fixation for barrel clover; fruit development for tomato; and lignin related processes for poplar. Inspection of the selected 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 example, the rice protein *LOC_Os10g38080*, was previously annotated with anatomical structure morphogenesis, and is annotated by Argot2-seeded BMRF with seed (coat) development. *LOC_Os10g38080* is a subtilisin homologue which according to available RNAseq data is expressed in 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 seed coat development [30]. An example for an annotation for a previously completely unannotated 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 to compare function annotation between different plants, a task which is much less easily performed with existing annotations that are derived from various methods and that have a much lower coverage than our approach. Such comparison between orthologous genes in different plants allows to assess the limits of orthology-based function prediction, and to analyze gene function evolution.

To characterize ortholog groups with functional predictions that differ from expectations based on sequence similarity, orthologs and paralogs were identified with orthoMCL [31], resulting in 25,347 groups (Table S7). Group members for which 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). In case the predicted biological processes in such a group are different despite high sequence similarity, this would be indicative of evolutionary divergence by, e.g., neofunctionalization. To identify such cases, groups with at least four different organisms (6073) were ranked by their largest inner group distance and 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 variety 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 related to cell development, regulation, and response to chemical stimulus. For the latter group, 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 example a phosphatase with existing experimental annotation in Arabidopsis, PURPLE ACID PHOSPHATASE 26 (PAP26). PAP26 plays a role in the phosphate metabolism [32] and phosphate starvation [32] in Arabidopsis. The majority of the proteins with function predictions in 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 is not reflected by orthology predictions and phylogenetic relationships of the group members (Fig. 4B and C). Independent expression data indicates that Arabidopsis PAP26 is expressed in a housekeeping-like manner, but the expression pattern varies between paralogs in other species, e.g. soybean, and to a lesser extent orthologs, e.g. between tomato and soybean (Fig. 4D). The different expression patterns give credibility to the variation in function predictions of Argot2-seeded BMRF. This indicates 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 large scale comparisons between various plant species.

3. Discussion

Finding associations between proteins and biological processes is a major challenge in non-model plants. Most experimental studies are aimed towards model organisms; hence

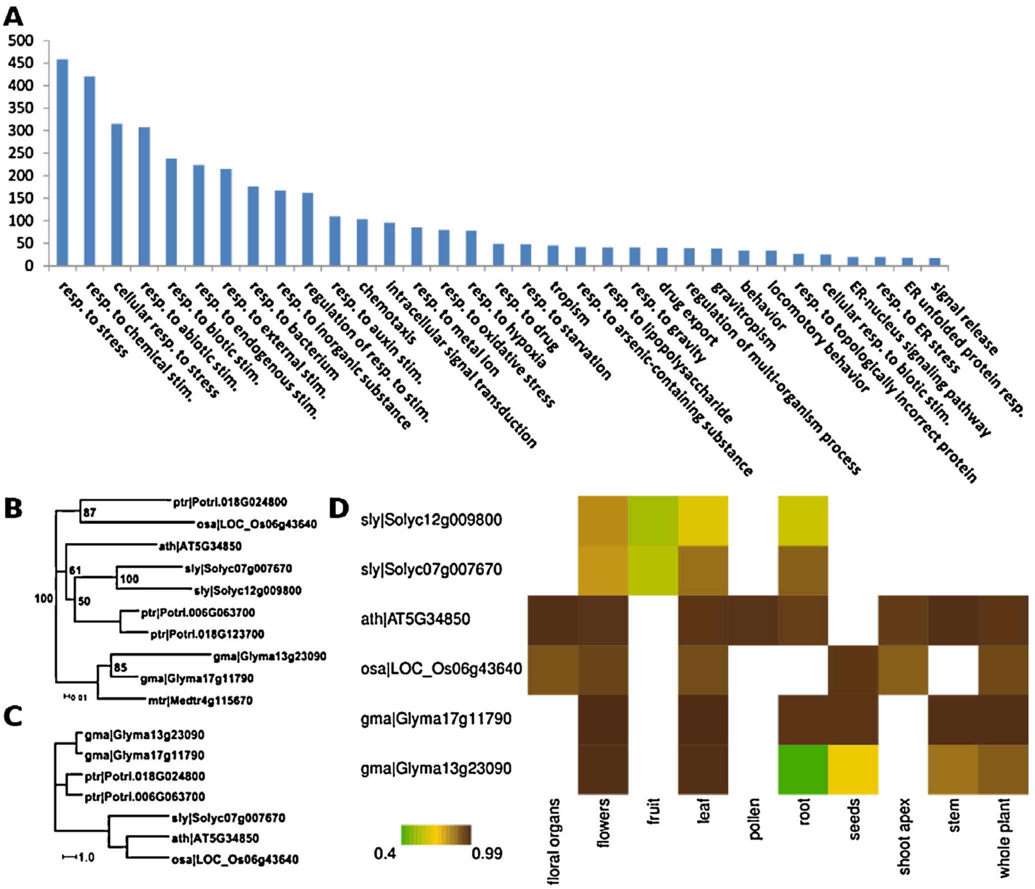

Fig. 4. Comparison between sequence divergence and functional divergence. (A) Overview of the most frequent GO terms in the top 100 most functionally divergent ortholog groups that are represented by "response to chemical stimulus" (Figure S1). (B and C) Phylogenetic relations of Arabidopsis PURPLE ACID PHOSPHATASE 26 orthologs. Trees contain Arabidopsis (ath), soybean (gma), tomato (sly), Populus (ptr) and rice (osa) PAP26 orthologs. (B) Unrooted phylogenetic tree based on sequence data. The tree was calculated with 1000 bootstraps. Confidence values are indicated at the branches in percent. (C) Distance tree based on our function predictions. Missing identifiers were not 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 different tissue clusters. The heatmap color represents a mean percentile rank of normalized expression studies aggregated by averaging to ten tissue clusters (Table S8). Missing data is indicated in white. An overview of the aggregated studies is available in Table S8. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

experiment-based function annotation is sparse in 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 type of function annotation that can be obtained. An example is the characterization of overexpressed rice genes in Arabidopsis [34] to infer function. Here, the problem is that the biological process of a protein is often bound to the local environment or a specific condition and a different (plant) environment might change the outcome. 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 were not covered.

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

The combination of sequence- and network-based function prediction obtained by seeding BMRF with Argot2, offers a significant benefit over applying these methods separately. We validated the method in rice and demonstrated greatly improved performance compared to each 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 have been obtained by various, mostly sequence-based approaches. A clear description of the methods and input data is often lacking, leading to the risk of error propagation and circular reasoning [3,39]. Our approach has the benefit of applying a standard method to the various genomes. Moreover, for many proteins which so far were not associated with any biological process, we now provide predictions of biological processes. Nevertheless, the combination of Argot2 and BMRF is indirectly constrained by the experimental data in databases such as UniProt [40] or PFAM [41], and by the proteins covered in available networks. It will however be straightforward to integrate newly available datasets such as additional co-expression networks or novel gene function annotations in the framework presented. Depending on availability of novel network or annotation data, we indeed plan to update our resource. An additional limitation of our current approach is that the structure of the Gene Ontology is not taken into account in the prediction process. Most existing computational methods for gene function prediction suffer from this drawback. It is feasible to make a set of GO term predictions consistent with the GO-structure [42] and we plan to apply this method to Argot2-seeded BMRF predictions in the future.

BMRF output consists of a list of probabilities for each gene to be associated with each biological process. This allows to rank proteins in order of their likelihood of association with a biological process of interest. However, it can also be important to have a finite set of predictions. To provide that, we applied a cut-off to the probabilities, based on Arabidopsis, the only species from which enough data was available. It is difficult to assess how valid the application of this cut-off in other plant species is. However, the average number of predictions per protein that we obtain in each of the species based on the cut-off that was applied is close to the observed average for Arabidopsis, giving some credibility to this cut-off. For one species, tomato, the number of predicted BP terms per protein is somewhat higher than the experimentally observed number for Arabidopsis. Hence, argot2-seeded BMRF possibly suffers from overprediction in this case. This could possibly be caused by the higher density (number of interactions compared to number of proteins) of the tomato network. However, in any case, the probabilities associated with the predictions allow narrowing down the prediction results to the most reliable ones, if so desired.

With the consistent annotation of multiple plant genomes that we performed, the relation between homology and biological process predictions can be analyzed. Ortholog groups with divergent functions indicated cases where conclusions based on sequence similarity might be inappropriate. Such inappropriate conclusions may be more common than generally acknowledged. Indeed, as recently noted in the context of comparing putative orthologs between species, relying on sequence similarity alone might identify an ortholog with the correct molecular function, but will more often than not fail to identify an ortholog that participates in the correct biological process [43]. In a comparison of gene expression patterns between different plant species, the number of times for which the homolog with the most similar pattern of expression (“expressolog”) was not also the most similar at the sequence level, ranged between 15% and 50% [44]. Similarly, about half of a collection 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, for example aimed at identifying patterns in retention of duplicated genes [46,47] or functional biases in single-copy genes [7], are currently performed based on function annotations obtained using sequence similarity. Such studies will benefit from the gene annotations presented here, which overcome the limitations of purely sequence-based annotation of gene functions.

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

The combination of sequence-based and network-based predictions is a huge improvement for sparsely annotated plant genomes. With the advent of RNA-seq [48] coexpression network-based protein function prediction can become a preferred method. Combined with additional analysis, such as genome-wide association studies (GWAS), potential candidate genes for traits-of-interest could be identified more reliably. Such candidate genes will be of great help in applications related to plant breeding. The ability to associate unannotated proteins to particular biological processes will spark experimental work and be essential for the advancement of 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 in the network, and connections in the network indicate functional relationships between proteins. A statistical model (Markov Random Field) describes how involvement of a protein in a particular BP influences the probability that its neighbors in the network are also involved in that BP. The parameters in the statistical model describe for each BP how strongly neighbors influence each other. Parameter values are trained using a Bayesian approach by performing simultaneous estimation of the model parameters and prediction of protein functions. This strategy needs a set of known protein functions as initial labelling of the network. Argot2 is a purely sequence-based prediction method, using searches of the UniProt and Pfam databases as input. To combine these two methods, two strategies were applied. In the first integration method, for each biological process, ranks for the different proteins were obtained from both BMRF and Argot2, by ordering the proteins based on their score for that process. These ranks were added to obtain a final ranking, which was used as the prediction score for that biological process. In a second integration strategy, initial predictions were generated with Argot2. These were supplied to BMRF as training data, meaning that the initial labelling of the nodes in the network was based on 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 was 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 for which no annotation was available in the data from the Gene Ontology project were used for validation. In all cases, only Biological 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 based on 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 ids of the tomato co-expression network were 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 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 part of the input network. Thus, proteins not contained in the input network were removed. In all 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 used to generate the input for Argot2 [16]. In the context of the validation setup, all rice proteins were removed from the UniProt database to avoid Argot2 using information from those proteins.

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

Prediction runs of different method and network combinations were assessed with 100 cross-validation runs. In each run, randomly, a subset ($n = 200$) of proteins was chosen and the annotation was removed (masked). For every run, predicted functions were compared with the masked ones. Only biological process terms with at least three masked proteins were used in the performance assessment in order to allow for sufficient statistics. In the performance assessment, negative cases consisted of gene-BP associations which were not annotated as such in the experimental data.

Performance was assessed by the area under the receiver operating characteristic curve (AUC) and the *F*-score. The AUC is the area under the curve of 1-specificity vs. sensitivity, and is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one [20]. Specificity is the fraction of proteins experimentally known not to perform a given function which are indeed not predicted to do so, whereas sensitivity (or recall) is the fraction of proteins experimentally known to perform a given function which are indeed predicted to do so. *F*-score is based on the precision–recall (precision vs. sensitivity) curve. Precision is the fraction of proteins predicted to perform a given function which are indeed experimentally known to do so. The *F*-score is equal to the harmonic mean of precision and recall, and the maximum value of the *F*-score ($F_{\max}$-score) was used for each biological 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 ($F_{\max}$-score), calculated with Arabidopsis data as previously described [17], was used to set a cut-off on the posterior probability. The threshold obtained with Arabidopsis data was used in the other species, because in those species, too few annotations are available to obtain 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 number of predicted functions per protein. Barrel clover, poplar and rice biological 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 total number of proteins and total number of GO terms for which annotations were obtained, the annotation of each protein was expanded by including the parent GO terms of all assigned GO terms. For the calculation of the number 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 each protein were interpreted as vector. The Euclidian distance for each combination of proteins within a group of orthologs was calculated. The mean of distances within a group (inner group distance) was used to rank groups of orthologs. For the PAP26 example, only groups with existing experimental annotation in Arabidopsis were taken in to account. The PAP26 tree was estimated with RaxML version 7.2.8-ALPHA [56] using the PROTGAMMA-JJTF 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 the expression was determined with cufflinks [59] with default parameters. Soybean expression data was obtained from SoyBase [60]. Rice expression data was 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 each gene.

Acknowledgements

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

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.cpb.2014.07.001.

References

[1] M.C. Schatz, J. Witkowski, W.R. McCombie, Current challenges in de novo plant genome sequencing and assembly, Genome Biol. 13 (2012) 243, http://dx.doi.org/10.1186/gb4015.

[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) D1178–D1186, http://dx.doi.org/10.1093/nar/gkr944.

[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) 723–735, http://dx.doi.org/10.1093/bib/bbr002.

[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) 707–720, http://dx.doi.org/10.1111/j.1469-8137.2012.04184.x.

[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) 590–600, http://dx.doi.org/10.1104/pp.111.189514.

[6] R. Monclus, J.-C. Leplé, C. Bastien, P.-F. Bert, M. Villar, et al., Integrating genome annotation and QTL position to identify candidate genes for productivity, architecture and water-use efficiency in *Populus* spp., BMC Plant Biol. 12 (2012) 173, http://dx.doi.org/10.1186/1471-2229-12-173.

[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) 2898–2903, http://dx.doi.org/10.1073/pnas.1300127110.

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

[9] Gene Ontology Consortium, Gene Ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29, http://dx.doi.org/10.1038/75556.

[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) 221–227, http://dx.doi.org/10.1038/nmeth.2340.

[11] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410, http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

[12] D.M.A. Martin, M. Berriman, G.J. Barton, GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes, BMC Bioinform. 5 (2004) 178, http://dx.doi.org/10.1186/1471-2105-5-178.

[13] W.T. Clark, P. Radivojac, Analysis of protein function and its prediction from amino acid sequence, Proteins 79 (2011) 2086–2096, http://dx.doi.org/10.1002/prot.23029.

[14] A. Vazquez, A. Flammini, A. Maritan, A. Vespignani, Global protein function prediction from protein–protein interaction networks, Nat. Biotechnol. 21 (2003) 697–700, http://dx.doi.org/10.1038/nbt825.

[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 based on network data, PLoS ONE 5 (2010) e9293, http://dx.doi.org/10.1371/journal.pone.0009293.

[16] M. Falda, S. Toppo, A. Pescarolo, E. Lavezzo, B. Di Camillo, et al., Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms, BMC Bioinform. 13 (Suppl. 4) (2012) S14, http://dx.doi.org/10.1186/1471-2105-13-S4-S14.

[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) 271–281, http://dx.doi.org/10.1104/pp.110.162164.

[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) D561–D568, http://dx.doi.org/10.1093/nar/gkq973.

[19] I. Lee, B. Ambaru, P. Thakkar, E.M. Marcotte, S.Y. Rhee, Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*, Nat. Biotechnol. 28 (2010) 149–156, http://dx.doi.org/10.1038/nbt.1603.

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

[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 functional genomics research, Bioinformatics 21 (2005) 3674–3676, http://dx.doi.org/10.1093/bioinformatics/bti610.

[22] J. Davis, M. Goadrich, The relationship between precision–recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning – ICML'06, ACM Press, New York, NY, USA, 2006, pp. 233–240, http://dx.doi.org/10.1145/1143844.1143874.

[23] N. Kurata, Y. Yamazaki, Oryzabase. An integrated biological and genome information database for rice, Plant Physiol. 140 (2006) 12–17, http://dx.doi.org/10.1104/pp.105.063008.

[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) 895–910, http://dx.doi.org/10.1105/tpc.111.083667.

[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) 520–524, http://dx.doi.org/10.1038/nature10625.

[26] G.A. Tuskan, S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, et al., The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), Science 313 (2006) 1596–1604, http://dx.doi.org/10.1126/science.1128691.

[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) 635–641, http://dx.doi.org/10.1038/nature11119.

[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) D883–D887, http://dx.doi.org/10.1093/nar/gkl976.

[29] J. Schmutz, S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, et al., Genome sequence of the palaeopolyploid soybean, Nature 463 (2010) 178–183, http://dx.doi.org/10.1038/nature08670.

[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) 466–480, http://dx.doi.org/10.1111/j.1365-313X.2008.03437.x.

[31] L. Li, C.J. Stoeckert, D.S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes, Genome Res. 13 (2003) 2178–2189, http://dx.doi.org/10.1101/gr.1224503.

[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 essential for efficient acclimation of Arabidopsis to nutritional phosphate deprivation, Plant Physiol. 153 (2010) 1112–1122, http://dx.doi.org/10.1104/pp.110.153270.

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

[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 wide range of trait information to facilitate analysis of gene function, Plant Cell Physiol. 52 (2011) 265–273, http://dx.doi.org/10.1093/pcp/pcq190.

[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) S2, http://dx.doi.org/10.1186/1471-2164-12-S1-S2.

[36] I. Lee, Y.-S. Seo, D. Coltrane, S. Hwang, T. Oh, et al., Genetic dissection of the biotic stress response using a genome-scale gene network for rice, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 18548–18553, http://dx.doi.org/10.1073/pnas.1110384108.

[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) 4389–4406, http://dx.doi.org/10.1105/tpc.112.104513.

[38] G. Blanc, K.H. Wolfe, Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution, Plant Cell 16 (2004) 1679–1691, http://dx.doi.org/10.1105/tpc.021410.

[39] B.E. Engelhardt, M.I. Jordan, J.R. Srouji, S.E. Brenner, Genome-scale phylogenetic function annotation of large and diverse protein families, Genome Res. 21 (2011) 1969–1980, http://dx.doi.org/10.1101/gr.104687.109.

[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) D565–D570, http://dx.doi.org/10.1093/nar/gkr1048.

[41] R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, et al., The Pfam protein families database, Nucleic Acids Res. 38 (2010) D211–D222, http://dx.doi.org/10.1093/nar/gkp985.

[42] Y.A. Kourmpetis, A.D. van Dijk, C.J. Ter Braak, Gene Ontology consistent protein function prediction: the FALCON algorithm applied to six eukaryotic genomes, Algorithms Mol. Biol. 8 (2013) 10, http://dx.doi.org/10.1186/1748-7188-8-10.

[43] S. Netotea, D. Sundell, N.R. Street, T.R. Hvidsten, ComPlEx: conservation and divergence of co-expression networks in *A. thaliana*, *Populus* and *O. sativa*, BMC Genomics 15 (2014) 106, http://dx.doi.org/10.1186/1471-2164-15-106.

[44] R.V. Patel, H.K. Nahal, R. Breit, N.J. Provart, BAR expressolog identification: expression profile similarity ranking of homologous genes in plant species, Plant J. 71 (2012) 1038–1050, http://dx.doi.org/10.1111/j.1365-313X.2012.05055.x.

[45] J. Lloyd, D. Meinke, A comprehensive dataset of genes with a loss-of-function mutant phenotype in Arabidopsis, Plant Physiol. 158 (2012) 1115–1129, http://dx.doi.org/10.1104/pp.111.192393.

[46] W. Jiang, Y. Liu, E. Xia, L. Gao, Prevalent role of gene features in determining evolutionary fates of whole-genome duplication duplicated genes in flowering plants, Plant Physiol. 161 (2013) 1844–1861, http://dx.doi.org/10.1104/pp.112.200147.

[47] H. Guo, T.-H. Lee, X. Wang, A.H. Paterson, Function relaxation followed by diversifying selection after whole genome duplication in flowering plants, Plant Physiol. (2013), http://dx.doi.org/10.1104/pp.112.213447.

[48] S. Marguerat, J. Bähler, RNA-seq: from technology to biology, Cell. Mol. Life Sci. 67 (2010) 569–579, http://dx.doi.org/10.1007/s00018-009-0180-6.

[49] P. Lamesch, T.Z. Berardini, D. Li, D. Swarbreck, C. Wilks, et al., The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools, Nucleic Acids Res. 40 (2012) D1202–D1210, http://dx.doi.org/10.1093/nar/gkr1090.

[50] K. Youens-Clark, E. Buckler, T. Casstevens, C. Chen, G. Declerck, et al., Gramene database in 2010: updates and extensions, Nucleic Acids Res. 39 (2011) D1085–D1094, http://dx.doi.org/10.1093/nar/gkq1148.

[51] A. Fukushima, T. Nishizawa, M. Hayakumo, S. Hikosaka, K. Saito, et al., Exploring tomato gene functions based on coexpression modules using graph clustering and differential coexpression approaches, Plant Physiol. 158 (2012) 1487–1502, http://dx.doi.org/10.1104/pp.111.188367.

[52] Arabidopsis Interactome Mapping Consortium, Evidence for network evolution in an Arabidopsis interactome map, Science 333 (2011) 601–607, http://dx.doi.org/10.1126/science.1203877.

[53] The UniProt Consortium, Reorganizing the protein space at the Universal Protein Resource (UniProt), Nucleic Acids Res. 40 (2012) D71–D75, http://dx.doi.org/10.1093/nar/gkr981.

[54] D.A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D.J. Lipman, et al., GenBank, Nucleic Acids Res. 41 (2013) D36–D42, http://dx.doi.org/10.1093/nar/gks1195.

[55] T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, ROCR: visualizing classifier performance in R, Bioinformatics 21 (2005) 3940–3941, http://dx.doi.org/10.1093/bioinformatics/bti623.

[56] A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, Bioinformatics 22 (2006) 2688–2690, http://dx.doi.org/10.1093/bioinformatics/btl446.

[57] M. Schmid, T.S. Davison, S.R. Henz, U.J. Pape, M. Demar, et al., A gene expression map of *Arabidopsis thaliana* development, Nat. Genet. 37 (2005) 501–506, http://dx.doi.org/10.1038/ng1543.

[58] T.D. Wu, S. Nacu, Fast and SNP-tolerant detection of complex variants and splicing in short reads, Bioinformatics 26 (2010) 873–881, http://dx.doi.org/10.1093/bioinformatics/btq057.

[59] C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, et al., Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat. Biotechnol. 28 (2010) 511–515, http://dx.doi.org/10.1038/nbt.1621.

[60] A.J. Severin, J.L. Woody, Y.-T. Bolon, B. Joseph, B.W. Diers, et al., RNA-seq atlas of *Glycine max*: a guide to the soybean transcriptome, BMC Plant Biol. 10 (2010) 160, http://dx.doi.org/10.1186/1471-2229-10-160.