Current Plant Biology 1 (2014) 73–82

Contents lists available at ScienceDirect

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).