BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 6 2012, pages 851–857
doi:10.1093/bioinformatics/btr720

Systems biology

Advance Access publication January 4, 2012

Automatic selection of reference taxa for protein–protein interaction prediction with phylogenetic profiling

Martin Simonsen[1,2,*], Stefan R. Maetschke[2,3] and Mark A. Ragan[2,3]

[1]Bioinformatics Research Centre, Aarhus University, Aarhus 8000, Denmark, [2]Australian Research Council Centre of Excellence in Bioinformatics and [3]The University of Queensland, Institute for Molecular Bioscience, Brisbane, QLD 4072, Australia

Associate Editor: David Posada

ABSTRACT

Motivation: Phylogenetic profiling methods can achieve good accuracy in predicting protein–protein interactions, especially in prokaryotes. Recent studies have shown that the choice of reference taxa (RT) is critical for accurate prediction, but with more than 2500 fully sequenced taxa publicly available, identifying the most-informative RT is becoming increasingly difficult. Previous studies on the selection of RT have provided guidelines for manual taxon selection, and for eliminating closely related taxa. However, no general strategy for automatic selection of RT is currently available.

Results: We present three novel methods for automating the selection of RT, using machine learning based on known protein–protein interaction networks. One of these methods in particular, *Tree-Based Search*, yields greatly improved prediction accuracies. We further show that different methods for constituting phylogenetic profiles often require very different RT sets to support high prediction accuracy.

Availability: The datasets and software used in the experiments can be found at http://users-birc.au.dk/zxr/phyloprof/

Contact: zxr@birc.au.dk; somme89@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 18, 2011 ; revised on December 23, 2011; accepted on December 27, 2011

1 INTRODUCTION

With the advent of next-generation technologies, genomic data are now appearing at a rate that renders manual annotation and analysis of genome sequences infeasible, thus motivating the development of automated computational methods. Prediction of protein–protein interactions (*PPIs*) is a key problem in systems biology, and methods based on e.g. interactions between protein domains (Ng *et al.*, 2003), gene-fusion events (Enright *et al.*, 1999) and phylogenetic profiling (Gaasterland *et al.*, 1998; Pellegrini *et al.*, 1999) (*PP*) support genome-scale inference at low computational cost. The application of PP in PPI prediction is based on the hypothesis that gene products which function together are likely to exhibit a common pattern of occurrence across taxa. The *phylogenetic profile* of a gene or gene product represents its pattern of presence and absence across a set of taxa, and thus similarity between profiles can be used as an indicator of functional linkages. PPI prediction using PP has proven to be fairly accurate in prokaryotes but has been much less successful in eukaryotes, a situation attributed to the modularity of eukaryotic proteins, large-scale evolutionary trends including secondary endosymbioses and parasitism and the limited diversity of sequenced genomes (Jothi *et al.*, 2007; Ruano-Rubio *et al.*, 2009; Snitkin *et al.*, 2006).

Many studies identify the choice of reference taxa (RT) as critical for accurate prediction and suggest that accuracy can be improved by matching the RT to the interaction network under investigation. Representatives of all three domains of life (bacteria, archaea and eukaryotes) should be used for highly conserved networks, whereas close relatives are best suited for predicting more specialized functions. Furthermore, inclusion of too many closely related taxa, and taxa with derivative or biased gene complements, can diminish the prediction accuracy (Herman *et al.*, 2011; Jothi *et al.*, 2007; Karimpour-Fard *et al.*, 2007; Singh *et al.*, 2008; Snitkin *et al.*, 2006; Sun *et al.*, 2005, 2007). Balancing these considerations greatly complicates manual RT selection, thus motivating the development of automatic selection approaches.

Here we introduce three novel machine-learning methods for selection of informative RT based on known PPI networks. We compare the accuracy with which PPIs in one prokaryote and three eukaryotes are predicted with the resulting taxon sets, and demonstrate that one of the methods in particular, *Tree-Based Search* (TBS), supports highly accurate predictions. Four different PP methods are used in the PPI prediction experiments, and we show that no single taxon set achieves top accuracy with all PP methods. Finally, we show that taxon sets can be optimized using PPI networks from other taxa without substantially affecting the prediction accuracy.

2 DATA

2.1 Phylogenetic profiles

Our set of RT consists of all fully sequenced species in UniProt (Bairoch *et al.*, 2005) version 22. This set contains 52 eukaryotes, 859 bacteria and 69 archaea, including multiple strains of some species. We downloaded the complete set of protein sequences from UniProt, including plasmids, and used these to create real-valued phylogenetic profiles containing E-values from pairwise BLAST searches. For binary profiles, we used an E-value threshold of 10^{-7}.

Jothi *et al.* (2007) investigated the prediction of metabolic pathways in *Escherichia coli* and *Saccharomyces cerevisiae* using PP and a series of manually selected reference-taxon sets. In their experiments, the taxon set *BAE3a*, containing 60 taxa from all three domains, supported the best average

[*]To whom correspondence should be addressed.

prediction accuracy in both query taxa. We compiled a taxon set which closely resembles BAE3a (five strains unavailable in Uniprot were replaced by closely related strains) to enable a comparison with a manual taxa selection approach.

2.2 Phylogenetic trees

We created a multifurcating phylogenetic tree (T) over all 980 RT using the taxonomic common tree tool from the NCBI webpage (http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi). In T, taxa with multiple strains in RT are represented as internal nodes, whereas all other taxa are represented as leaf nodes. A second tree T_{bin}, based on T, was constructed as follows. Each taxon in RT represented as an internal node in T was replaced with an artificial node and then reinserted as a leaf at a random position under the artificial node. For each set of sibling leaf nodes S_i with parent p_i, where $|S_i| > 2$ a new binary tree with room for $|S_i|$ leaves was constructed and used to replace the subtree rooted at p_i in T. The nodes in S_i were then inserted as leaves at random positions in this binary tree. In the resulting tree T_{bin}, all taxa in RT are thus represented as leaves and have at most one sibling (Fig. 1). T_{bin} allows a more finely grained selection of RT compared with T when used in combination with the Tree Level Filtering (TLF) and TBS methods presented in Section 3.3, due to an increased number of hierarchical levels.

2.3 PPI networks

High confidence PPI networks were created for each of *E. coli*, *S. cerevisiae*, *Drosophila melanogaster* and *Arabidopsis thaliana* using the STRING database (Jensen *et al.*, 2009; von Mering *et al.*, 2005) version 8.3 (Table 1). These networks were created by filtering for *binding* proteins supported by experimental data, curated databases or both. Only interactions where at least one of these evidence channels has a confidence score ≥ 0.9 (the highest confidence in STRING) were used. The interactions include both proteins that interact physically and proteins that are part of the same protein complex but without direct contact.

For each of the four networks, we created a dataset containing all PPIs in the network, and twice as many non-interacting protein pairs. These were generated following Ben-Hur *et al.* (2006) by randomly pairing proteins and discarding pairs for which STRING contains evidence of interaction at any confidence level excluding evidence originating from PP methods (the co-occurrence evidence channel). Exclusion of PPIs that are supported only by co-occurrence-based methods would lead to an overestimate of the prediction accuracy of PP methods, and thus such pairs are regarded as negatives. Results of experiments with datasets containing different positive–negative ratios (see Section 3 of the Supplementary Material) indicate that the use of a ratio higher than 1:2 will not affect prediction accuracies significantly.

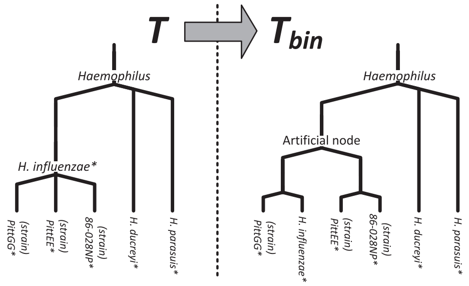

Fig. 1. An example of how a phylogenetic tree, containing multiple strains of *Haemophilus influenzae*, is transformed into a tree where all RT (marked with *) are represented as leaf nodes.

Table 1. PPI network statistics

Taxa	No. of proteins	No. of interactions
Escherichia coli	789	1752
Saccharomyces cerevisiae	3301	21 096
Drosophila melanogaster	1205	7718
Arabidopsis thaliana	1886	22 941

3 METHODS

3.1 Prediction accuracy measure

We use the area under curve (AUC) as a measure of prediction accuracy, which enables the accuracy of different combinations of PP methods, optimization methods and PPI networks to be compared in a clear way but ROC curves for key results can be found in Section 4.4 of the Supplementary Material. Given a list of protein pairs ranked by interaction confidence, the AUC was computed using the Gini coefficient:

$$\mathrm{AUC} = \frac{1}{2}\sum_{k=1}^{n}\left((X_k - X_{k-1})(Y_k + Y_{k-1})\right), \tag{1}$$

where X is the false positive rate and Y the true positive rate at pair k in the ranked list.

Tenfold cross-validation experiments are used to evaluate the performance of different taxa selection approaches as follows. Each dataset was divided into 10 random subsets of equal size and each subset was then used to evaluate a taxon set which had been optimized on the remaining nine subsets. Paired t-tests for key cross-validation experiments are provided in Section 4.5 of the Supplementary Material.

3.2 Phylogenetic profiling methods

The original PP method by Pellegrini *et al.* (1999) (referred to as the *Pellegrini* method) employs the Hamming distance between binary phylogenetic profiles to cluster similar profiles. Instead of clustering profiles, the Hamming distance between profile pairs can be used as a confidence score for interactions between the corresponding proteins, thus allowing the ranking of a list of protein pairs. The Pellegrini method considers two profile pairs with the same Hamming distance as equally significant evidence of interactions. This is problematic as, in an extreme case, all matching entries in one profile pair could contain ones while they contain zeroes in the other pair. To ensure that the Pellegrini method infers only interactions between proteins where there is significant evidence of correlated evolution, profile pairs for which the number of *matches* (Number of matching entries containing ‘1’) falls below the threshold $|BP| \cdot 0.1$ are assigned a distance of $|BP|$ where $|BP|$ is the length of the binary profiles. This threshold yielded the highest accuracy in preliminary experiments.

Date *et al.* (2003) use mutual information between pairs of real-valued profiles as a confidence score, where a high score is taken as evidence of correlated evolution (referred to as the *Mutual Information* method). We calculated mutual information values as described in Date *et al.* (2003), and use these values to rank protein pairs.

Wu *et al.* (2003) calculate confidence scores for interactions between protein pairs using binary profiles and the hypergeometric distribution (referred to as the *Hypergeometric Distribution* method). The expected number of matches between two profiles is described with a hypergeometric distribution, assuming no interaction and uniform distribution of matches. The confidence score is then computed as the *P-value* for the number of matches between two profiles being as large as the number of matches under the hypergeometric distribution. This method was further developed first by Kharchenko *et al.* (2006), who adjusted for the size of each reference genome, and later by Cokus *et al.* (2007), who showed how phylogenetic relationships among RT could be taken into account. For this method we adopt the linear coefficient from Cokus *et al.* (2007) to combine P-values. The confidence scores computed by both the hypergeometric distribution and Runs Informed methods are used to rank a list of protein pairs in our experiments.

3.3 Reference-taxon set optimization methods

Selecting the set of RT that leads to the best prediction accuracy can be formulated as an optimization problem. Given a set of RT, the objective is to identify the subset of taxa which optimizes a score function $S(RT',P,P',F)$. Here RT' is a subset of RT, P is a set of putative interacting proteins, and P' is a subset of P containing true interacting protein pairs (e.g. a PPI network). F is a PP-based method which assigns confidence scores to protein pairs in P using phylogenetic profiles computed using the taxa in RT'. The score function ranks protein pairs in P according to their confidence score and computes the corresponding AUC as described in Section 3.1. RT' is represented as a binary vector of length $|RT|$, where each entry corresponds to a reference-taxon and contains '*1*' if the taxon is in RT' and '*0*' otherwise. As there are $2^{|RT|}$ different combinations of RT, it is infeasible to identify an optimal RT' using exhaustive search, and hence heuristic methods must be used.

3.3.1 Tree-level filtering The method of Sun *et al.* (2007), referred to here as the *TLF* method, filters out closely related taxa from a reference-taxon set. A phylogenetic tree with hierarchical levels (Fig. 2) is used for the selection of taxon sets of different sizes as follows. For each level l, a set of clades C_l is constructed by first placing the leaves of each subtree rooted at level l in the same clade, and then placing all remaining leaves in their own unique clade. Given some l, one of three different filters is applied to select a single taxon from each clade defined at level l. The *exterior* and *outlier* filters select a random taxon positioned at the highest and lowest level in each clade in C_l, respectively, and the *random* filter selects at random a taxon from each clade in C_l. We applied all three filters to each level in T_{bin}, and the taxon set which gave rise to the highest score was taken as the best. If T is used instead of T_{bin}, taxa positioned as sibling leaf nodes can be selected in only one of three ways by the TLF method. (i) None of them is selected. (ii) One of them is selected. (iii) All of them are selected. T_{bin} allows such nodes to be selected in more ways as each leaf has a maximum of one sibling.

3.3.2 Iterative taxon selection In the Iterative taxon selection (ITS) method, the binary vector RT' is first initialized with random values and scored. The entries in RT' are then reversed in random order, and after each change the resulting set of taxa RT'' is scored. If $S(RT'',P,P',F) > S(RT',P,P',F)$, RT' is replaced by RT''. The optimization terminates if changing all entries in RT' once does not improve the score; otherwise another iteration of the search heuristic is performed.

3.3.3 Genetic algorithm Genetic algorithms (GA) (Holland, 1975) are a family of search heuristics applicable to binary parameter vectors. The genetic algorithm used in this work maintains a population of n parameter vectors (initialized with random values) and iteratively applies *recombination*, *mutation* and *selection* steps to create successive generations of parameter vectors. The recombination step randomly pairs parameter vectors in the population, and applies uniform crossover (Syswerda, 1989) on each pair with a probability of 0.6. The uniform crossover strategy swaps each pair of parameters in two vectors with a probability of 0.4. The mutation step introduces random mutations in parameter vectors by inserting random values in each vector element with a probability of 0.001. After the recombination and mutation steps, the score of each vector is computed and a new population of size n is created from the old population in the selection step, using *binary tournament selection* (Goldberg, 1990; Goldberg *et al.*, 1991). This selection scheme iteratively selects two random parameter vectors from the previous population, and stores the best-scoring vector in the new population (note that each vector can be selected multiple times). Our selection scheme also uses *elitist selection*, where the best-scoring vector is always included in the new population.

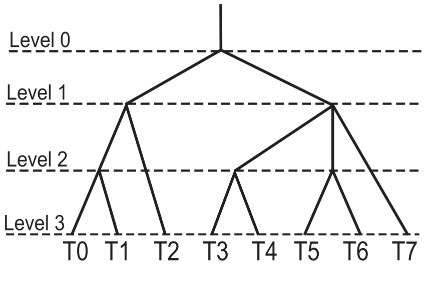

Fig. 2. An example of hierarchical levels in a phylogenetic tree. TLF defines a set of clades at level 1 as: $\{T0,T1,T2\}$ $\{T3,T4,T5,T6,T7\}$. At level 2, the clades are defined as: $\{T0,T1\}$ $\{T2\}$ $\{T3,T4\}$ $\{T5,T6\}$ $\{T7\}$.

We also investigated two other recombination strategies that take phylogenetic relationships of RT into account, but as the accuracy and rate of convergence achieved using these strategies were similar to those of uniform crossover, they were not included in the experiments. Details on these recombination strategies can be found in Section 1 of the Supplementary Material.

3.3.4 TBS Similar to TLF (above), the TBS method uses a phylogenetic tree with hierarchical levels (Fig. 2) to guide the search for an informative reference-taxon set (see Section 2 of the Supplementary Material for pseudocode). This takes place in three steps: first, RT' is initialized with all reference taxa ($RT'=RT$), and scored. A new set of taxa, RT'', is then constructed by copying RT' and removing all taxa rooted at some inner node $v \in T$. If RT'' results in a better score than RT', the pair $(v,\Delta s)$ [where $\Delta s = S(RT',P,P',F) - S(RT'',P,P',F)$] is stored in a set M, and RT' is replaced by RT''. The hierarchical levels in T_{bin} are used to define the order in which nodes are visited as follows: at each level, a node v is visited if v is an inner node and no ancestor of v is found in M. Nodes are visited in random order at each level, starting at level 0 and progressing to the highest level in the tree. This first step works as a crude filter that quickly excludes many taxa from RT'.

In the second step, a more fine-grained filter is applied to the taxa rooted at v, where $(v,\Delta s) \in M$. The pairs in M are ordered by their Δs values and removed in ascending order, i.e. the nodes which gave rise to the smallest improvements in the score are removed first. For each removed pair, the children of v are visited in random order and for each child c, RT'' is constructed as a copy of RT' where entries corresponding to taxa rooted at c are negated. Next, Δs is computed and if $\Delta s > 0$, RT' is set to RT''. For any value of Δs, M is updated by inserting the pair $(c,|\Delta s|)$. Experiments showed that exclusion of subsets of taxa rooted at a node v [where $(v,\Delta s) \in M$ and Δs is small] improves the score more often than does exclusion of taxa rooted at nodes where Δs is large. Intuitively, a set of taxa rooted at a node with a small Δs is more likely to contain both informative and uninformative taxa that cancel each other out, where removing uninformative taxa improves the score.

In the third step, all remaining taxa in RT' are excluded one by one. If an exclusion results in a better score, that taxon stays excluded; otherwise it is again included in RT'.

As with the TLF method, the use of T_{bin} in the first and second steps allows taxa represented as sibling nodes in T to be selected in more ways. However, because of the fine-grained selection performed by third step, the difference between using T and T_{bin} is often negligible with this method.

4 RESULTS AND DISCUSSION

4.1 Cross-validation experiments

Four PPI datasets (Section 2.3) were used to evaluate the 10-fold cross-validation prediction accuracy (AUC) of the different RT selection methods using the PP methods described in Section 3.2.

Because of space limitations we present only the results for the *E. coli* and *D. melanogaster* PPI networks here, whereas the results for *S. cerevisiae* and *A. thaliana* are presented in Section 4.3 of the Supplementary Material. T_{bin} was used where applicable, as preliminary experiments indicated that this tree supported the best overall prediction accuracy for TLF and TBS.

The four optimization methods described in Section 3.3 were employed to create sets of taxa for each PP method, with the exception of the Runs Informed method. As both ITS and GA require the score function to be evaluated many times where $|RT'|$ is large, it is infeasible to use them in combination with Runs Informed, as it runs in $O(|RT'|^4)$ time. In all experiments, the GA method used a population size of 32 and 5000 generations to optimize reference-taxon sets; this was sufficient to find (local) maxima for all PP methods in our experiments.

The results of cross-validation experiments for *E. coli* and *D. melanogaster* are shown in Figures 3 and 4, respectively. All optimization methods were able to identify reference-taxon sets which improved the prediction accuracy compared with both the complete set of 980 taxa and the BAE3a set. However, the increase in prediction accuracy differs considerably among the four PP methods. The Pellegrini and Runs Informed methods show a significant improvement in AUC for both query taxa, whereas the AUC of the Mutual Information method was improved only slightly. The improvement in prediction accuracy for the hypergeometric distribution method is small for the three eukaryotes, whereas a large improvement can be observed on the *E. coli* datasets. As shown in Table 2, the hypergeometric distribution method requires a large number of RT to make accurate predictions, hence with only 52 eukaryotes available as RT it is difficult to improve the accuracy of this method for an eukaryotic query taxon. The standard variations in Figure 3 are larger compared with the results for the three eukaryotes which can be explained by the relatively small size of the *E. coli* network.

Interestingly, when used with optimized reference-taxon sets the cross-validation results show that the Pellegrini method achieves prediction accuracies equal to or better than the more recent PP methods we investigated. Predictions with the Pellegrini (and Runs Informed) method have a tendency to result in either low or high AUCs depending on the reference-taxon set used. Thus, it seems that the correct choice of RT is of particular importance for the Pellegrini method.

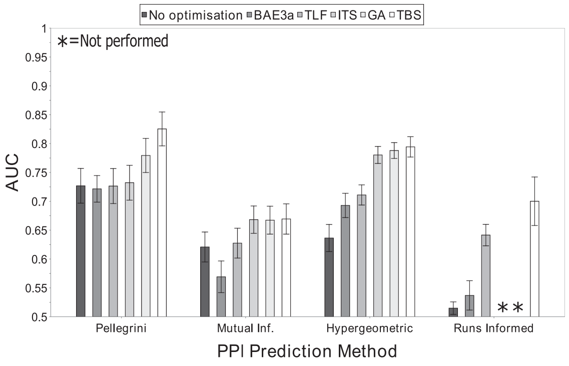

Fig. 3. *Escherichia coli* 10-fold cross-validation results. Each bar shows the average AUC and the corresponding Standard Deviation.

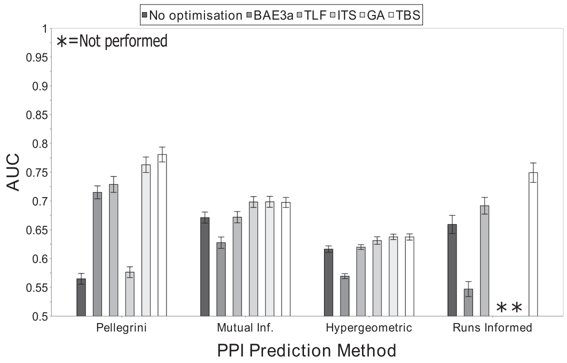

Fig. 4. *Drosophila melanogaster* 10-fold cross-validation results. Each bar shows the average AUC and the corresponding Standard Deviation.

Reference-taxon sets optimized using ITS generally lead to higher AUCs than those optimized with TLF. A notable exception is the results for the Pellegrini method with *D. melanogaster* where ITS apparently gets caught in a local maximum, leading to an AUC similar to that of the complete taxon set. The granularity by which the TLF method can select RT is limited by the number of hierarchical levels in the phylogenetic tree (25 in T_{bin}). Consequently, the simple ITS method is often able to outperform TLF as it can select taxa with a much finer granularity. TBS achieves the best overall prediction accuracy, followed closely by our GA. Compared to ITS, GA and TBS sample a larger part of the search space and are therefore less prone to getting caught in a poor local maximum. Paired *t*-tests (see Section 4.5 of the Supplementary Material) confirms that TBS achieves significantly higher AUCs compared with other optimization methods in most cases and as TBS require considerably less evaluations of S than GA, it is the preferred method for selecting RT.

In many cases, the BAE3a taxon set resulted in lower AUCs compared with the complete taxon set. This indicates that the greater number of fully sequenced genomes now available has helped to increase prediction accuracy and it is therefore unfair to compare the AUCs from BAE3a directly with those achieved by optimizing on the complete taxon set. Furthermore, BAE3a is aimed at predicting metabolic pathways and may therefore not contain an appropriate composition of taxa for predicting the proteins in our datasets. But as BAE3a was selected specifically for PPI prediction in *E. coli* and *S. cerevisiae*, where it showed a good average performance on different metabolic pathways, BAE3a is expected to be more informative regarding interactions in these query taxa compared with the full taxon set. The cross-validation results for *E. coli* and *S. cerevisiae* do not indicate that this is the case, thus illustrating the difficulty of selecting informative RT manually.

4.2 Composition of RT

Table 2 presents the distribution of taxa over the tree kingdoms of life in taxon sets optimized using TBS. The distribution of taxa in sets optimized for *S. cerevisiae* and *A. thaliana* can be found in Section 4.1 of the Supplementary Material. It is obvious that different PP methods require very different configurations of RT to achieve top prediction accuracy. Both the Pellegrini and Runs Informed methods perform well on relatively small taxon sets composed mostly of taxa that represent the same domain as the query

Table 2. Distribution of RT over the tree kingdoms of life after optimization of the taxon set with *E. coli* and *D. melanogaster* as query taxa

Prediction method	*Escherichia coli*				*Drosophila melanogaster*			
	Archaea	Bacteria	Eukaryotes	Total	Archaea	Bacteria	Eukaryotes	Total
Pellegrini	8	92	0	100	0	4	10	14
Mutual Information	25	162	9	196	8	79	18	105
Hypergeometric Distribution	44	264	3	311	14	301	28	343
Runs Informed	54	118	5	177	1	4	20	25

All taxon sets were optimised with TBS using the full dataset of each PPI network.

taxon. Conversely, the Mutual Information and Hypergeometric Distribution methods achieve top accuracies with large taxon sets in which prokaryotes predominate. In the case of *E. coli*, the best taxon set found for the Runs Informed method is relatively large compared with taxon sets found for the three eukaryotes. We therefore speculate that the low prediction accuracy of the Run Informed method with *E. coli* could be attributed to a poorly optimized reference-taxon set.

4.3 Cross-taxon prediction

Figures 5 and 6 present the results of cross-taxon PPI prediction experiments (See Section 4.3 of the Supplementary Material for additional results). In these experiments, taxon sets were optimized using TBS with the full dataset for each of the four PPI networks. The resulting reference-taxon sets were then used for PPI prediction in *S. cerevisiae* and *D. melanogaster* to investigate if optimized taxon sets can be applied across species. For each PP method, we show the result of a self-test, i.e. where training and prediction were carried out using the same dataset. As the AUCs of the self-tests are similar to the average AUCs in the 10-fold validation experiments, overfitting does not appear to be a major concern, and the self-test AUCs can therefore be used as benchmarks.

In both Figure 5 and 6, the taxon sets optimized for *E. coli* lead to low prediction accuracies when used for prediction in eukaryotes. This is unsurprising, considering the large phylogenetic distance between the three eukaryotes used here and *E. coli*. Although the three eukaryotes are not closely related either, the results in Figure 6 indicate that optimized taxon sets can be shared between these. In particular, the AUC for the Pellegrini or Runs Informed methods show a significant increase when both optimization and prediction was performed with an eukaryote. In one experiment where the Runs Informed method was used to predict PPIs in *S. cerevisiae* with a taxon set optimised for *D. melanogaster*, the prediction accuracy did not improve compared to using the full taxon set (see Section 4.3 of the Supplementary Material). An explanation for this observation could be that the relatively small size of the *D. melanogaster* PPI network. When a taxon set is optimized using a small network, the evolutionary history of taxa in the resulting set cannot be expected to encompass other interactions than those present in the network, and thus the taxon set will be more-specialized than a set optimized using a larger network. As large high-quality networks are currently available for only a few eukaryotes, cross-taxon optimization seems to be a useful approach for increasing the accuracy of PPI prediction. The results in Table 3 support this as a taxon set optimized with the large *A. thaliana* network results in high prediction accuracies when used for PPI prediction in *S. cerevisiae* and *D. melanogaster*.

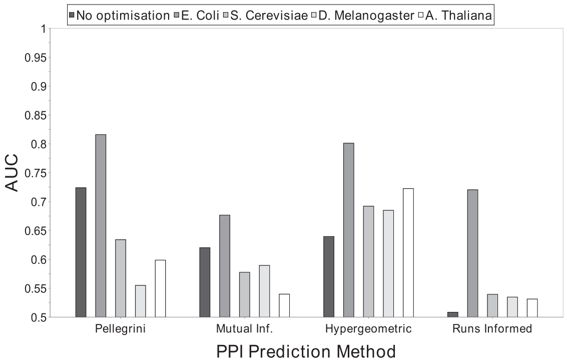

Fig. 5. *Escherichia coli* cross-taxon prediction results. The *x*-axis shows the PP method used for PPI prediction. The networks used by TBS to optimize the reference-taxon sets are shown in the legend.

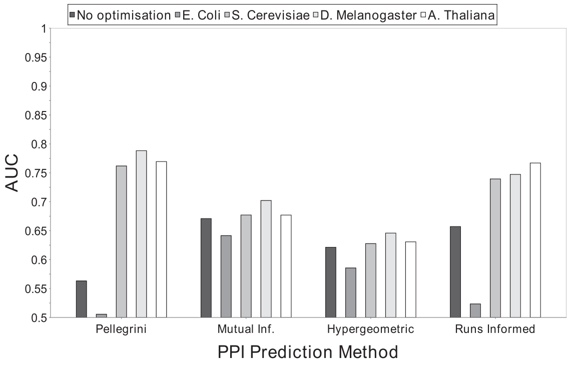

Fig. 6. *Drosophila melanogaster* cross-taxon prediction results. The *x*-axis shows the PP method used for PPI prediction. The networks used by TBS to optimize the reference-taxon sets are shown in the legend.

The number of taxa shared between each pair of taxon sets used in the cross-taxon experiments is shown in Table 4 (see Section 4.2 of the Supplementary Material for additional results). Taxon sets optimized for *E. coli* have a small or non-existent overlap with those optimized for the three eukaryotes, and result in a low prediction accuracy in eukaryotes (Fig. 5). However, there is no clear relationship between the size or extent of intersections and the prediction accuracy in three eukaryotes. The Pellegrini method achieves a high AUC when predicting PPI in *S. cerevisiae* using the taxon set optimized for *A. thaliana* although only 20% of the taxa in this set are found in the set optimized for *S. cerevisiae*.

Table 3. Comparison of AUCs from cross-validation and cross-taxon experiments using TBS optimization

Prediction method	*Saccharomyces cerevisiae*		*Drosophila melanogaster*	
	CV	CT	CV	CT
Pellegrini	0.70	0.69	0.78	0.77
Mutual Information	0.65	0.62	0.70	0.68
Hypergemetric Distribution	0.57	0.56	0.64	0.63
Runs Informed	0.72	0.70	0.75	0.76

CV = average AUCs of 10-fold cross-validation. CT = AUCs of cross-taxon experiments, where the reference-taxon set was optimized with *A. thaliana* PPI data.

Conversely, when the taxon set optimized for *D. melanogaster* is used in combination with the Runs Informed method to predict PPIs in *S. cerevisiae*, the result is a low AUC despite a much-larger overlap of 40%. Studies (Jothi *et al.*, 2007; Singh *et al.*, 2008) indicate that some taxa can be substituted or even omitted from a reference-taxon set without significantly degrading the prediction accuracy, whereas other taxa cannot. Accordingly, two taxon sets with a small overlap can support similar AUCs as long as informative taxa are present in both sets and less-informative taxa, e.g. with low-quality annotation, are left out.

4.4 Cross-method prediction

Studies of the effect of reference-taxon selection on the prediction accuracy of PP methods typically focus on a single method (Jothi *et al.*, 2007; Karimpour-Fard *et al.*, 2007; Singh *et al.*, 2008; Sun *et al.*, 2005, 2007). Our results demonstrate that the optimal composition of RT depends not only on the query taxon but also on the PP method used. Figures 7 and 8 show the cross-method prediction results where taxon sets were optimized with TBS for each of the four PP methods, and then used for PPI prediction with each of the four PP methods (see Section 4.3 of the Supplementary Material for additional results). In cases where taxon sets are optimized for one PP method and then used for prediction with another method, the prediction accuracy drops significantly. From these results, it is evident that PP methods are not interchangeable with regard to a given taxon set which further complicates the selection of RT, as selection must target a specific PP method.

4.5 Implications of results

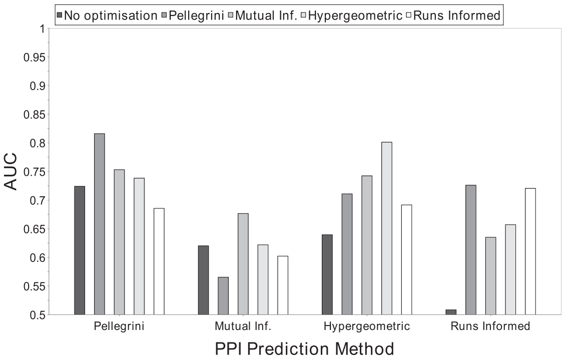

Fig. 7. *Escherichia coli* cross-method prediction results. The *x*-axis shows the PP method used for PPI prediction while the legend shows the PP method used in combination with TBS to optimize the reference-taxon sets.

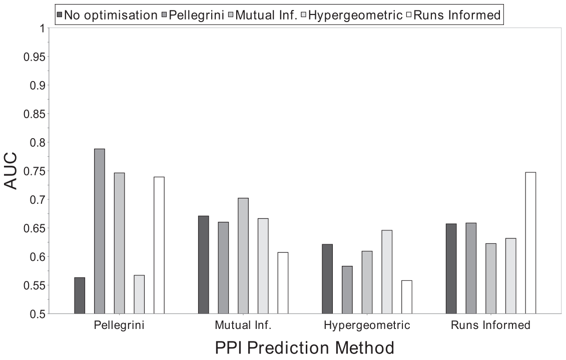

Fig. 8. *Drosophila melanogaster* cross-method prediction results. The *x*-axis shows the PP method used for PPI prediction while the legend shows the PP method used in combination with TBS to optimize the reference-taxon sets.

The optimization methods presented here reduce the complexity of constructing reference-taxon sets by automatically accounting for factors such as phylogenetic relationships among RT and quality of annotation. Guidelines such as those presented by Jothi *et al.* (2007) represent an improvement over use of all possible RT, but as the optimal composition of RT depends on not only query taxon but also PP method, manual selection of RT is an unpromising direction, particularly as more genomes are sequenced.

Previous studies have concluded that the prediction accuracy of PP methods in eukaryotes is generally poor compared with the results achieved in prokaryotes (Jothi *et al.*, 2007; Kharchenko *et al.*, 2006; Ruano-Rubio *et al.*, 2009; Snitkin *et al.*, 2006). Insufficient numbers of fully sequenced eukaryotic genomes, complex cellular histories

Table 4. The number of taxa in the intersection of reference-taxon sets optimised with TBS using the full dataset of each PPI network. Numbers in parentheses show the size of intersections relative to the smaller of the two taxon sets in percent

	Pellegrini				Runs Informed			
	EC (%)	SC (%)	DM (%)	AT (%)	EC (%)	SC (%)	DM (%)	AT (%)
E. coli (EC)	100 (100)	0 (0)	0 (0)	1 (5)	177 (100)	6 (20)	4 (16)	3 (10)
S. cerevisiae (SC)	0 (0)	20 (100)	5 (35)	4 (20)	6 (20)	30 (100)	10 (40)	16 (55)
D. melanogaster (DM)	0 (0)	5 (35)	14 (100)	8 (57)	4 (16)	10 (40)	25 (100)	11 (44)
A. thaliana (AT)	1 (5)	4 (20)	8 (57)	20 (100)	3 (10)	16 (55)	11 (44)	29 (100)

(e.g. endosymbioses), genome reduction in parasites, mismatch between breadth of the reference-taxon set and the network under investigation, and quality of annotation were put forward as possible causes of lower prediction accuracy. Here we have shown that with a correct selection of taxon sets, PPI prediction can be as accurate for eukaryotes as for *E. coli*. However, the results in Herman *et al.* (2011) indicate that co-complexed proteins are easier to predict with co-occurrence-based methods compared with transient interactions. Thus, the increase in prediction accuracy which can be achieved with the presented taxon set optimization methods may vary depending on the type of query interactions and the quality of available training data.

5 CONCLUSIONS

We have introduced three new methods for automatic selection of RT for PP which can be used in combination with existing PP methods. Our results show that the choice of RT can have a substantial effect on the accuracy of PPI prediction, and that a single reference-taxon set does not guarantee good prediction accuracy with all PP methods. Taxon sets optimized with TBS achieved consistently high prediction accuracies compared with those attained via the previously published TLF method or with a manually composed reference-taxon set. Furthermore, our experiments showed that the original Pellegrini *et al.* PP method often supported higher prediction accuracies compared with more recent alternatives when optimized taxon sets were used.

As complete genome sequences appear at an ever-increasing frequency, manual selection of informative RT becomes increasingly difficult. Here we have shown that informative sets of RT can be selected in an objective, fast and scalable manner for eukaryotes as well as prokaryotes. Our approach utilizes a known PPI network to select informative RT and we have demonstrated that where a high-quality network is unavailable for the query taxon, an informative taxon set can be constructed using a network from another taxon. As discussed by Jothi *et al.* (2007), the evolutionary history of the RT must correspond with that of the query PPI to achieve high prediction accuracy. Thus, PPI prediction accuracies could likely be improved even more by optimizing taxon sets on networks enriched in proteins homologous (or better, orthologous) to those in the set of query interactions. For example, a network enriched in membrane proteins might be used to create a reference-taxon set for predicting novel PPIs among membrane proteins.

We anticipate that in-depth examination of the composition of optimized taxon sets could lead to further improvements in taxon-selection strategies, and in the accuracy of PP methods for PPI prediction.

ACKNOWLEDGEMENTS

We thank Dr Melissa J. Davis for help with retrieval of genomic data from Uniprot and constructing PPI networks.

Funding: Australian Research Council grants (CE0348221 and DP110103384).

Conflict of Interest: none declared.

REFERENCES

Bairoch,A. *et al.* (2005) The universal protein resource (UniProt). *Nucleic Acids Res.*, **33**, 154–159.

Ben-Hur,A. *et al.* (2006) Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, **7** (Suppl. 1), S2.

Cokus,S. *et al.* (2007) An improved method for identifying functionally linked proteins using phylogenetic profiles. *BMC Bioinformatics*, **8** (Suppl. 4), S7.

Date,S. . *et al.* (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.*, **21**, 1055–1062.

Enright,A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Gaasterland,T. *et al.* (1998) Constructing multigenome views of whole microbial genomes. *Microb. Comp. Genomics*, **3**, 177–192.

Goldberg,D.E. (1990) A note on Boltzmann tournament selection for genetic algorithms and population-oriented simulated annealing. *Complex Syst.*, **4**, 445–460.

Goldberg,D.E. *et al.* (1991) A comparative analysis of selection schemes used in genetic algorithms. *Found. Genet. Algorith.*, **1**, 69–93.

Herman,D. *et al.* (2011) Selection of organisms for the co-evolution-based study of protein interactions. *BMC Bioinformatics*, **12**, 363.

Holland,J.H. (1992) *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA.

Jensen,L.J. *et al.* (2009) STRING 8 – a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, 412–416.

Jothi,R. *et al.* (2007) Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC Bioinformatics*, **8**, 173.

Karimpour-Fard,A. *et al.* (2007) Investigation of factors affecting prediction of protein-protein interaction networks by phylogenetic profiling. *BMC Genomics*, **8**, 393.

Kharchenko,P. *et al.* (2006) Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics*, **7**, 177.

Ng,S.K. *et al.* (2003) Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, **19**, 923–929.

Pellegrini,M. *et al.* (1999) Assigning protein functions by comparative genome analysis protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.

Ruano-Rubio,V. *et al.* (2009) Comparison of eukaryotic phylogenetic profiling approaches using species tree aware methods. *BMC Bioinformatics*, **10**, 383.

Singh,S. *et al.* (2008) Testing the accuracy of eukaryotic phylogenetic profiles for prediction of biological function. *Evol. Bioinformatics Online*, **4**, 217–223.

Snitkin,E.S. *et al.* (2006) Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics*, **7**, 420.

Sun,J. *et al.* (2005) Refined phylogenetic profiles method for predicting protein-protein interactions. *Bioinformatics*, **21**, 3409–3415.

Sun,J. *et al.* (2007) Phylogenetic profiles for the prediction of protein-protein interactions: how to select reference organisms? *Biochem. Biophys. Res.*, **353**, 985–991.

Syswerda,G. (1989) Uniform crossover in genetic algorithms. In *Proceedings of the Third International Conference on Genetic Algorithms*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 2–9.

von Mering,C. *et al.* (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33**, 433–437.

Wu,J. *et al.* (2003) Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, **19**, 1524–1530.