BMC Bioinformatics

BioMed Central

Research article

Open Access

Phylogeny-guided interaction mapping in seven eukaryotes

Janusz Dutkowski* and Jerzy Tiuryn

Address: Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland

Email: Janusz Dutkowski* - januszd@mimuw.edu.pl; Jerzy Tiuryn - tiuryn@mimuw.edu.pl

* Corresponding author

Published: 30 November 2009

BMC Bioinformatics 2009, **10**:393 doi:10.1186/1471-2105-10-393

Received: 1 July 2009
Accepted: 30 November 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/393

Abstract

Background: The assembly of reliable and complete protein-protein interaction (PPI) maps remains one of the significant challenges in systems biology. Computational methods which integrate and prioritize interaction data can greatly aid in approaching this goal.

Results: We developed a Bayesian inference framework which uses phylogenetic relationships to guide the integration of PPI evidence across multiple datasets and species, providing more accurate predictions. We apply our framework to reconcile seven eukaryotic interactomes: *H. sapiens*, *M. musculus*, *R. norvegicus*, *D. melanogaster*, *C. elegans*, *S. cerevisiae* and *A. thaliana*. Comprehensive GO-based quality assessment indicates a 5% to 44% score increase in predicted interactomes compared to the input data. Further support is provided by gold-standard MIPS, CYC2008 and HPRD datasets. We demonstrate the ability to recover known PPIs in well-characterized yeast and human complexes (26S proteasome, endosome and exosome) and suggest possible new partners interacting with the putative SWI/SNF chromatin remodeling complex in *A. thaliana*.

Conclusion: Our phylogeny-guided approach compares favorably to two standard methods for mapping PPIs across species. Detailed analysis of predictions in selected functional modules uncovers specific PPI profiles among homologous proteins, establishing interaction-based partitioning of protein families. Provided evidence also suggests that interactions within core complex subunits are in general more conserved and easier to transfer accurately to other organisms, than interactions between these subunits.

Background

Protein-protein interactions are essential to most cellular processes. Thus large-scale PPI networks can greatly contribute to our understanding of the cellular machinery at systems level. Experimental techniques such as yeast two-hybrid assays [1-4] and TAP-MS [5,6] have generated large amounts of binary PPIs and protein complex data, providing the first snapshots of eukaryotic interactomes. Unfortunately, the available experimental techniques are far from perfect, both in terms of their accuracy, as well as coverage. For instance, the yeast interactome has recently been estimated to contain from around 37,000 up to even 75,500 protein interactions between approximately 6,000 proteins [7]. Although already over 80,000 yeast PPIs have been reported, given the estimated false positive rates of the experiments, the yeast interactome is suggested to be roughly 50% complete [7]. Using a more conservative definition and omitting indirect co-complex associations, the authors of [8] estimate the number of yeast interactions to be ~18,000 and conclude that three idependent Y2H assays cover only around 20% of this amount. In case of human, the entire interacome is estimated to be covered in roughly 10% [7,9]. Furthermore, many doubts and criticism have been expressed in the literature regarding the low overlap between independent screens - originally attributed to a high false-positive rate of these experiments [10-12]. More recent studies (e.g. [8]) suggest that the low overlap can largely be explained by low sampling sensitivity and differences in assay types. Considering all mentioned limitations, none of the existing experimental systems can provide a complete and error-proof interaction map of a complex organism within reasonable time and respecting budget limitations. As recently estimated, around 20 independent proteome-scale screens would be required to reliably identify each mappable interaction in a moderately-sized interactome of *Drosophila melanogaster* [13].

Simultaneously with the development of experimental techniques, computational methods for predicting PPIs have emerged [14-16]. These approaches complement experimental methods and can be used to validate noisy data, as well as to select new targets for screening experiments [15]. Available computational techniques exploit various sources of evidence. Among them are ones based on genomic data [17,18], protein sequences [19,20], phylogenetic profiles [21], and classification-based approaches [22-24]. Other methods explore the premise that interacting proteins often co-evolve and thus similarity of phylogenetic trees can be used to infer interactions [25-27]. Approaches using maximum likelihood estimation (MLE) for inferring the probability of domain-domain interactions have been presented. The first of such analysis was performed in [28], where the authors used yeast PPI data to estimate the probability of domain-domain interactions, and subsequently predict the interactions between proteins. Finally, multiple data sources have been integrated in a Bayesian framework in [29]. The last concept was further extended and applied to a wide range of heterogeneous data types from multiple species to construct comprehensive databases of functional associations [30,31].

In this study we are specifically interested in techniques which integrate and transfer PPI evidence across species. In its simplest form, this idea is implemented in the interolog (the term interlog is also used) mapping approach [32], which predicts an interaction between a pair of proteins (a, b) if in another species there exists a known interaction between a pair (a', b'), where a' and b' are orthologs of a and b, respectively. The transfer of PPI evidence across species can also be achieved at the level of conserved domains. In [33] the authors devised a maximum likelihood method, similar to [28], but using data from multiple organisms. In summary, the method estimates the probability of interactions between each pair of considered domains, based on the PPI evidence from multiple species. Inferred domain-domain interactions constitute integrated evidence, which is in turn used to predict protein-protein interactions. A similar method, but using heterogeneous data sources (including protein fusion and Gene Ontology annotations), was used in [34]. In general, combining interaction evidence from different species makes PPI predictions more robust to experimental noise. False positive observations are unlikely to be reproduced across multiple species [35]. Furthermore, evolutionarily conserved interactions are expectedly biologically significant. Evolutionary pressures are more likely to constrain functional units such as protein complexes rather than single interactions [36]. Hence, if an interaction has experimental support in datasets from diverse species, it is likely to be part of a significant functional module. Highly probable interactions identified in a subset of species can be transferred to other species [37], as was done in [38] to predict missing interactions within conserved protein modules.

We present an approach which uses protein family phylogenies to accurately map PPI evidence between homologous proteins. Contrary to previous studies [25-27], the phylogenies are not used to assess protein co-evolution, but to account for evolutionary relationships when integrating data from different organisms. Our current work builds on previously proposed CAPPI framework for comparing PPI networks across species [39]. CAPPI is based on a duplication and divergence model which mimics the processes by which most protein interactions are formed i.e. by copying from ancestral interactions during protein duplication and subsequently being sustained or lost over time. Using this model we can naturally incorporate interdependencies between PPIs and study the available data in evolutionary context. The only previous works based on these principles are [39] and [40] both of which concentrated on inferring ancestral states of the protein interaction network (the analysis in [40] was limited to a single protein family). Our current work presents the first application of the duplication and divergence model towards genome-scale inference of PPIs in extant species.

We use our framework to integrate and infer new PPIs in seven eukaryotes: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. We perform a comprehensive validation of our predictions using a GO-based functional similarity measure and assessment based on reference datasets of binary and co-complex PPIs. The obtained results demonstrate CAPPI's ability to identify a large percentage of known interactions in a blind test and provide new hypothesis for experimental verification when all known data is integrated. Our method shows a significant advantage over the standard interlog mapping approach and a maximum-likelihood domain-oriented method. We also analyze specific examples of valid PPI predictions in well-characterized complexes in yeast and human (proteasome, endosome and exosome), and show that core subcomplexes can be accurately recovered based solely on the data from the other species (i.e. without any use of the experimental data from the species of interest). Many of the between-module interactions (possibly species-specific) are harder to transfer from distant organisms. Finally, based on our predictions, we present hypothesis on new proteins interacting with the putative SWI/SNF chromatin remodeling complex in *A. thaliana*. Our results are freely available at http://bioputer.mimuw.edu.pl/cappi.

Results and Discussion

We develop a comprehensive framework for integrating and transferring PPI evidence across species. Our approach combines and extends the concepts of interlog mapping and Bayesian data integration. As opposed to the interlog approach, we employ PPI evidence from all homologous proteins, instead of using only the best-matching sequences in each case. This strategy is advantageous given the sparseness of the source datasets from which new interactions could be inferred. It is also motivated by the fact that the role of an individual protein in one species may be distributed over several homologous proteins in another species. Further, we use a Bayesian modeling framework to integrate PPI evidence from diverse experimental sources, taking into account their reliabilities and coverage. The evidence is accounted for in the context of the families' phylogenetic trees and under an assumed model of network evolution, which assigns probability scores to events of interaction loss or gain following a duplication or a speciation event (duplication and divergence model). Intuitively, the closer a given pair of proteins is to another pair, the more impact the evidence for one pair has on predicting the interaction of the other pair. The amount and reliability of the evidence, as well as the evolutionary proximity of the observed interactions to the pair of proteins in question, determines the posterior probability of interaction computed by our framework. We apply CAPPI to infer protein-protein interactions in seven eukaryotic species: human (*H. sapiens*), mouse (*M. musculus*), rat (*R. norvegicus*), fly (*D. melanogaster*), worm (*C. elegans*), yeast (*S. cerevisiae*), and thale cress (*A. thaliana*). The initial steps of our analysis preprocess the data and gather experimental evidence for interaction between members of distinct protein families. To this end, we identify groups of homologous proteins by clustering all non-redundant protein sequences downloaded from the Integr8 database [41] and pull relevant PPI data from IntAct [42], MINT [43] and DIP [44] databases (see Additional file 1 for details). The family-oriented view of the overlap of available PPI evidence for four best-represented interactomes is shown in Figure 1.

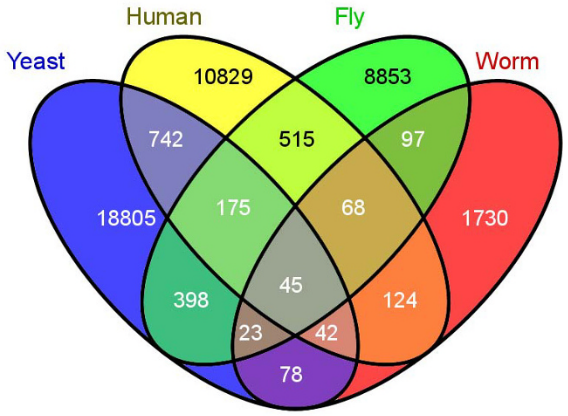

Figure 1
A family-oriented overview of the input PPIs. A 4-way Venn diagram illustrating the overlap of PPI evidence between four of the considered seven species: human, yeast, fly and worm. Each cell in the diagram is labeled with the number of pairs of protein families for which members interact in the corresponding species. For example, there are 742 pairs of protein families such that in both yeast and human there exists at least one interaction between members of the two families and no such interactions exist for fly and worm. Only about 0.5% (42514 of 8280415) of possible family pairs we consider have any evidence for interaction in any of the four species. Of these only 0.1% (45 of 42514) have evidence in all four species, which seems small, given that all considered families are evolutionarily conserved. However, the size of the overlap presumably corresponds to the fraction of the interactomes sampled experimentally, rather than to the actual level of conservation. For example, while there is a significant size difference between the overlap of the relatively best sampled yeast and human interactomes (742+175+45+42 = 1004 family pairs) and the overlap between yeast and worm interactomes (23+45+42+78 = 188 family pairs), the fraction of family pairs with PPI evidence from human and worm overlapping with such pairs in yeast is of the same magnitude (8% and 9%, respectively). It is highly probable that many of the homologous interactions in yeast and human have, yet unidentified, counterparts in worm and similarly in the other species. CAPPI uses phylogenetic information and probabilistic modeling to identify the most probable interactions in each species given the joint evidence from all input datasets and considering their reliability.

We consider two modes of application of our framework. First, the integration mode which gathers all available input data to provide a reconciled interactome view for each species. Second, the prediction mode which predicts the interactions for each species only based on the evidence from the other species (blind test). To demonstrate the different aspects of our method and enable a straightforward comparison to the previous approaches we use different combinations of the input datasets and different reliability values, yielding the following sets of inferred interactions (for details see Additional file 1):

CAPPI-Integ: interactions for all seven species inferred using all available experimental datasets.

CAPPI-Integ-3sp: yeast, fly and worm interactions inferred based on experimental datasets from Ito *et al.* [2], Uetz *et al.* [1], Giot *et al.* [45] and Li *et al.* [46], with reliability parameters set according to [33].

CAPPI-Pred: interactions inferred for each species using experimental datasets only from the other six species.

We compare the results of CAPPI with the following methods:

Domain-ML: a maximum likelihood domain-oriented method [33]. Yeast interaction predictions, based on experimental datasets of Ito, Uetz, Giot and Li, were provided by the authors.

Interlog: an interlog-based method implemented in [47]. The program was downloaded from the InteroPORC website http://biodev.extra.cea.fr/interoporc/Default.aspx and ran for each species using experimental datasets only from the other six species (same datasets as in CAPPI-Pred).

In the following sections, we investigate the performance of our method on large-scale data, as well as in small-scale experiments focused on specific functional modules.

Integration of interactions in seven eukaryotes

CAPPI-Integ provides an integrated and reconciled view of seven eukaryotic interactomes. Our ultimate goal is to provide a higher quality interactome for each input species. To assess the potential improvement, we perform two separate evaluations using a GO-based functional similarity measure and gold standard reference datasets.

GO-based scoring

Gene Ontology (GO) annotations are often used as indirect evidence for interaction. Intuitively, the more similar are the annotations of two proteins, the more confident we are in predicting an interaction between them. We first consider the biological process (BP) annotations and score our predictions, as well as the interactions from the input datasets, using the functional similarity measure from [48]. Mean *BP* scores for the input datasets and for the equal in size prediction datasets are summarized in Table 1. The scores of self interactions (present both in the input and in the inferred datasets) are excluded as they could introduce bias to the results (the GO annotations are identical in this case). Also, to avoid possible bias caused by the specific choice of proteins, input datasets are limited to interactions between members of conserved protein families used by CAPPI (see Additional file 1). For each CAPPI version in Table 1 we indicate the mean *BP* scores for the input dataset and the inferred output dataset of equal size. For example, in case of CAPPI-Integ the input yeast dataset contains 28590 interactions, for which the average *BP* score is 0.377. The corresponding CAPPI-Integ score of 0.412 was computed by taking the mean *BP* score of the 28590 best predictions in yeast (i.e. interactions with the highest probability). For each of the species CAPPI predictions receive significantly higher mean *BP* scores than the datasets used for training. The most significant improvement over the input datasets is achieved in case of the fly, worm and rat predictions. The mean *BP* score for the entire fly input dataset is 0.295, while the CAPPI-Integ dataset of the same size achieves a mean score of 0.425 (44% higher). In case of worm and rat prediction we observe a 29% and 30% increase in the *BP* score, respectively. Our results show that CAPPI is able to produce reconciled interactomes which significantly outperform the input interactomes (see also Wilcoxon test *p*-values in Table 1). A detailed view of the distributions of *BP* scores for experimental and predicted datasets of protein interactions in *D. melanogaster* is presented in Figure 2A. The predicted datasets (both CAPPI-Integ and CAPPI-Integ-3sp) contain a lot more high-scoring interactions than are present in the input datasets. Interestingly, while the Input-3sp for fly is almost as good as the Input-7sp, CAPPI-Integ-3sp is significantly outperformed by CAPPI-Integ. This is largely due to the integration of additional high quality datasets from other species, from which CAPPI-Integ can transfer new evidence when inferring the fly interactome.

The improvement in mean *BP* score described above is achieved for relatively large predicted datasets (as large as the initial inputs). As we show in Figure 2B, *BP* scores are actually higher for our top predictions. Figure 2B plots mean similarity scores according to all three ontologies: biological process (BP), molecular function (MF) and cellular component (CC), as functions of the number of predicted interactions. The mean scores for both CAPPI versions are negatively correlated with the size of the output dataset. This enables the user to trade size for quality, obtaining a smaller dataset, but of greater reliability.

Testing against gold standard datasets

We further survey the performance of our method using a set of gold standard binary PPIs pulled from [49] and [8], as well as co-complex data from the MIPS [50] and CYC2008 [51] complex catalogues (see Additional file 1 for details). Once again, we score CAPPI predictions and compare them to the scores of the input datasets. The results are presented in Figure 2C. The figure plots the

Table 1: *BP* score improvement over the input datasets.

Species	CAPPI-Integ				CAPPI-Integ-3sp			
	Data Size	Input Score	Output Score	Wilcoxon *p*-value	Data Size	Input Score	Output Score	Wilcoxon *p*-value
Yeast	28590	0.377	0.412	1.21e-31	1890	0.320	0.381	8.03e-06
Fly	12107	0.295	0.425	1.26e-113	4049	0.255	0.303	4.78e-05
Worm	2604	0.364	0.469	1.50e-21	856	0.374	0.485	2.02e-09
Arabidopsis	1349	0.596	0.623	0.02	NA			
Rat	1271	0.296	0.384	9.07e-06	NA			
Mouse	2456	0.417	0.463	1.53e-06	NA			
Human	17672	0.353	0.395	1.38e-31	NA			

The mean *BP* score of the input dataset and the inferred dataset of the same size are given for CAPPI-Integ and CAPPI-Integ-3sp. In all cases the inferred interaction set receives a significantly higher score than its input counterpart.

ratio of true positive and false positive interactions present among a subset of a given size. The true positive interactions are either confirmed by binary PPIs or known to participate in a characterized complex. Unfortunately negative gold standard sets of non-interacting protein pairs are not available. We take a standard heuristic approach and consider pairs of proteins with different subcellular localization as putative negative examples. We note that in certain situations, e.g. signalling pathways, it is possible that interacting proteins are in fact in different cellular compartments. Note also that in general true interactions constitute only a very small fraction of all possible protein pairs - at most 0.5% in yeast based on recent estimates from [7]. This is reflected in our reference datasets. The positive reference used in this case contains 22480 PPIs and co-complex pairs while the negative set contains 4857065 differencially localized pairs (see also Additional file 1). It is unlikely to identify a true interaction by pure chance alone. Results presented in Figure 2C confirm the previous observation that reliable interactions are generally ranked high by our method. It is comforting that both CAPPI datasets contain more confirmed interactions than differentially localized pairs among the top ranked predictions (TP/FP >> 1). Note that a reference interaction can only be identified if a relevant evidence interaction is present in the input experimental evidence for one of the species. Given that the gold standard datasets generally do not have a large overlap with the input high-thoughput datasets, many of the reference interactions will not be inferred by any integration procedure. Importantly as shown in Figure 2C CAPPI-Integ-3sp has a much higher TP/FP ratio than the input yeast datasets (Ito and Uetz) used for its training. CAPPI-Integ integrates four more high-throughput yeast datasets and consistently scores higher than three out of four of these inputs - Gavin (2002) dataset has a higher score, but for a smaller number of interactions.

Prediction of interactions in a blind test

We continue the performance evaluation by testing CAPPI's ability to predict interactions in a blind test. To this end, we compute the CAPPI-Pred dataset by iteratively leaving out PPI data of one of the seven species and predicting its interactions based only on the data from the other six species. We discuss the assessment of yeast and human predicted interactomes based on the two scoring frameworks.

Figure (3A and 3B) shows multiple histograms summarizing the *BP* score distribution among yeast and human predictions, respectively. The sizes of the predicted dataset (1576 for yeast and 17105 for human) have been selected to allow comparison with the interlog mapping predictions (see next section for details). Interestingly, we observe that while the performance of CAPPI-Pred is lower than CAPPI-Integ in case of yeast predictions, the opposite is true for the predicted human interactome. This suggests that while the yeast input interactions are necessary for good prediction results, human input datasets, on average, bring a less notable contribution.

In Figure 3C we plot the ratio of true positives and false positives as a function of the number of yeast PPIs returned by CAPPI-Pred. We evaluate the predictions sep-

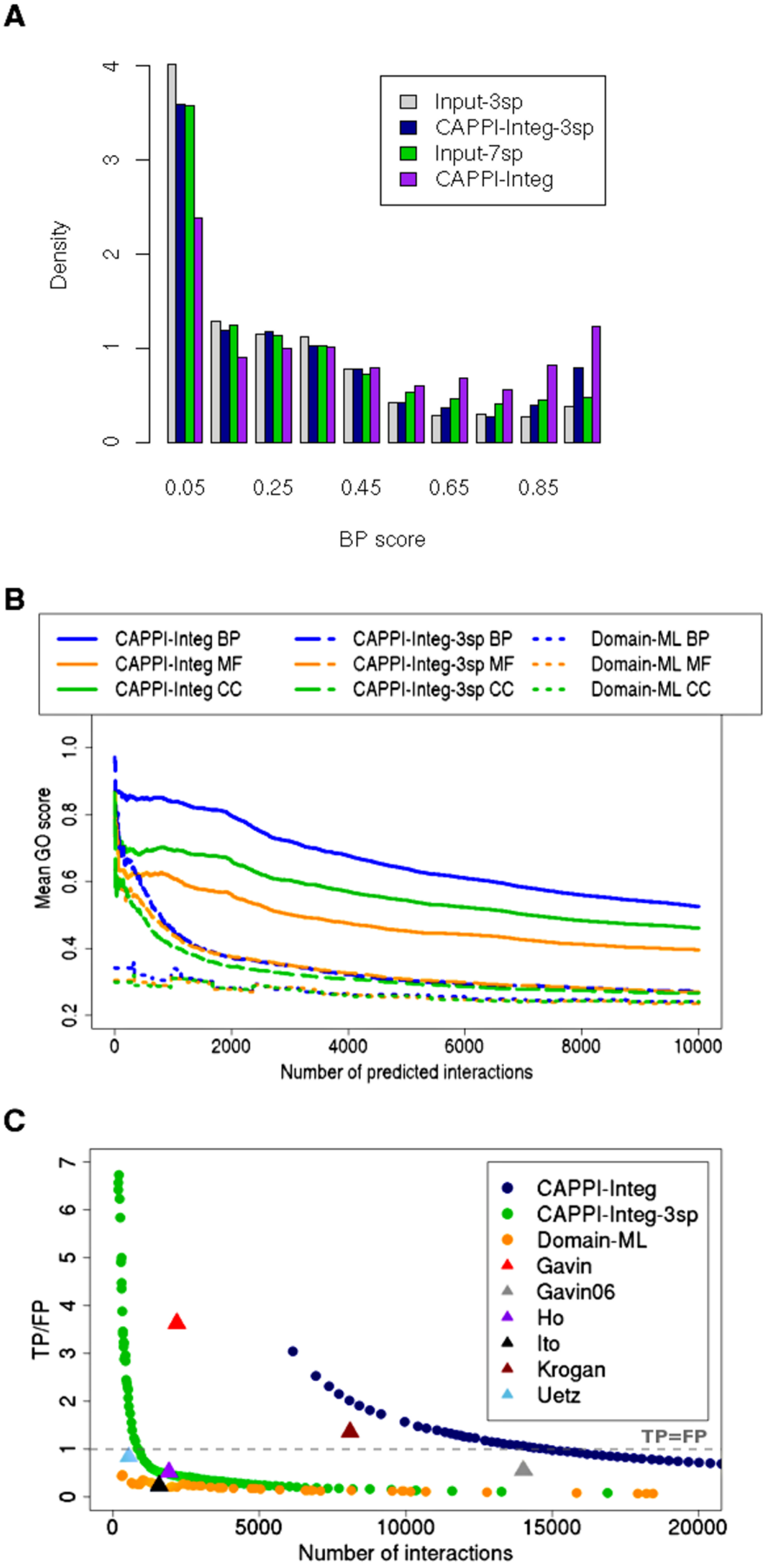

Figure 2 *(see legend on next page)*

Figure 2 *(see previous page)*
Assessment of CAPPI-Integ predictions. Assessment of CAPPI-Integ predictions. (A) Histogram of BP scores for the fly input datasets (combined) and corresponding inferred datasets of the same size (4049 PPIs in case of Input-3sp and CAPPI-Integ-3sp and 12107 PPIs in case of Input-7sp and CAPPI-Integ). Both CAPPI-Integ and CAPPI-Integ-3sp provide higher-scoring interactomes compared to their input datasets demonstrating the method's ability to use the interactions from distant species to make high quality predictions in other species. (B) Assessment of predicted yeast interactions using the three *GO scores*. The similarity of GO annotations of each pair of interacting proteins is measured in each ontology: biological process (BP), molecular function (MF) and cellular component (CC). CAPPI and Domain-ML predictions are ranked by their probabilities and the average GO score for the top *n* predictions is shown. CAPPI-Integ-3sp outperforms the domain based approach trained on the same experimental data. CAPPI-Integ integrates all available data from the seven species and further improves the predictions for yeast. (C) The ratio of true positives (TP) and false positives (FP) as a function of the number of yeast interactions. An interaction is deemed true positive if it is found in the reference dataset comprising co-complex and binary PPIs, and false positive, if the two proteins are assigned different localizations in the MIPS sub-cellular localization catalog (see text). The TP/FP ratios for the CAPPI-Integ, CAPPI-Integ-3sp and Domain-ML predictions are compared with the scores of the input experimental datasets. The gray dashed line marks the level at which the number of true positive predictions is equal to the number of false positive predictions.

arately using co-complex datasets (CAPPI-Pred Complex), gold standard binary PPI datasets (CAPPI-Pred PPI), as well as all available reference data (CAPPI-Pred All) - see Additional file 1 for details. An analogous study is performed for the predicted human interactome using the HPRD (complex and binary PPI) catalogues as reference (see Figure 3D). Note that similarly as for yeast, also for human the positive reference set is significantly smaller than the negative reference set. The joint human reference set (All) contains 57,093 protein pairs, which is less than 0.2% of the number of differentially localized pairs - consistent with the expected ratio of true interactions to all protein pairs in human, as estimated in [52]. The results show that CAPPI is able to infer high-scoring PPIs also in the case when no interactions from the predicted interactome are included in the training set. Most of the top predictions are confirmed by experimental data. We observe that while more yeast predictions are confirmed by co-complex pairs than by binary PPI data, the opposite is true in case of the human predictions. This can be explained by the differences in size of the respective reference datasets for the two species (see Additional file 1). When all available reference data is considered (CAPPI-Pred-All), the TP/FP ratios for the top 5,000 interactions in yeast and human are comparable (~0.8).

Filtering co-complex predictions

Evolutionary pressures are more likely to constrain essential functional units than individual interactions [36]. Thus co-complex PPIs should be easier to map accurately across species. This premise was previously explored in [38], where the authors showed that screening PPI predictions against conserved clusters improves prediction specificity. In an attempt to increase the percentage of co-complex PPIs in our predictions, we filtered the CAPPI-Pred output dataset, leaving only the predicted PPIs placed within conserved dense network regions. To this end, an ancestral interaction network was computed as in [39], and clustered using the MCL algorithm [53] to identify dense clusters. Each cluster was projected onto the network of the extant species (yeast or human) and CAPPI-Pred predictions within the projected regions were identified as a result. As shown in Figure (3C and 3D), this procedure significantly boosts the TP/FP ratio for both yeast and human data (see "Filtered Complex" plots). Interestingly, while the fraction of co-complex PPIs was increased, the fraction of confirmed binary PPIs was in general lowered by the filtering (except for the top ranked human predictions), suggesting that many binary PPIs placed outside or between protein complexes are filtered out in this case. This is in line with the observations made in [8] that binary and co-complex datasets are of complementary nature and often have small overlap.

Comparison with previous high-throughput multi-species approaches

Numerous existing computational approaches for predicting protein associations in multiple species can be loosely divided into three categories. The first group of methods contains approaches for predicting interactions *de novo* from protein sequence. These methods often utilize evolutionary information such as phylogenetic profiles or gene fusion events, but they do not explicitly transfer pre-identified interactions from one species to another. The second group of methods takes as input experimentally identified PPIs, integrates them and transfers the evidence to other species. The third group of studies is directed towards integration of heterogeneous experimental evidence such as PPI, mRNA co-expression, phylogenetic profile similarity, co-localization, domain associations, etc., and attempts to predict various types of functional associations, not limited strictly to protein-protein interactions. CAPPI was specifically designed as a model-based approach for integrating and transferring protein-protein

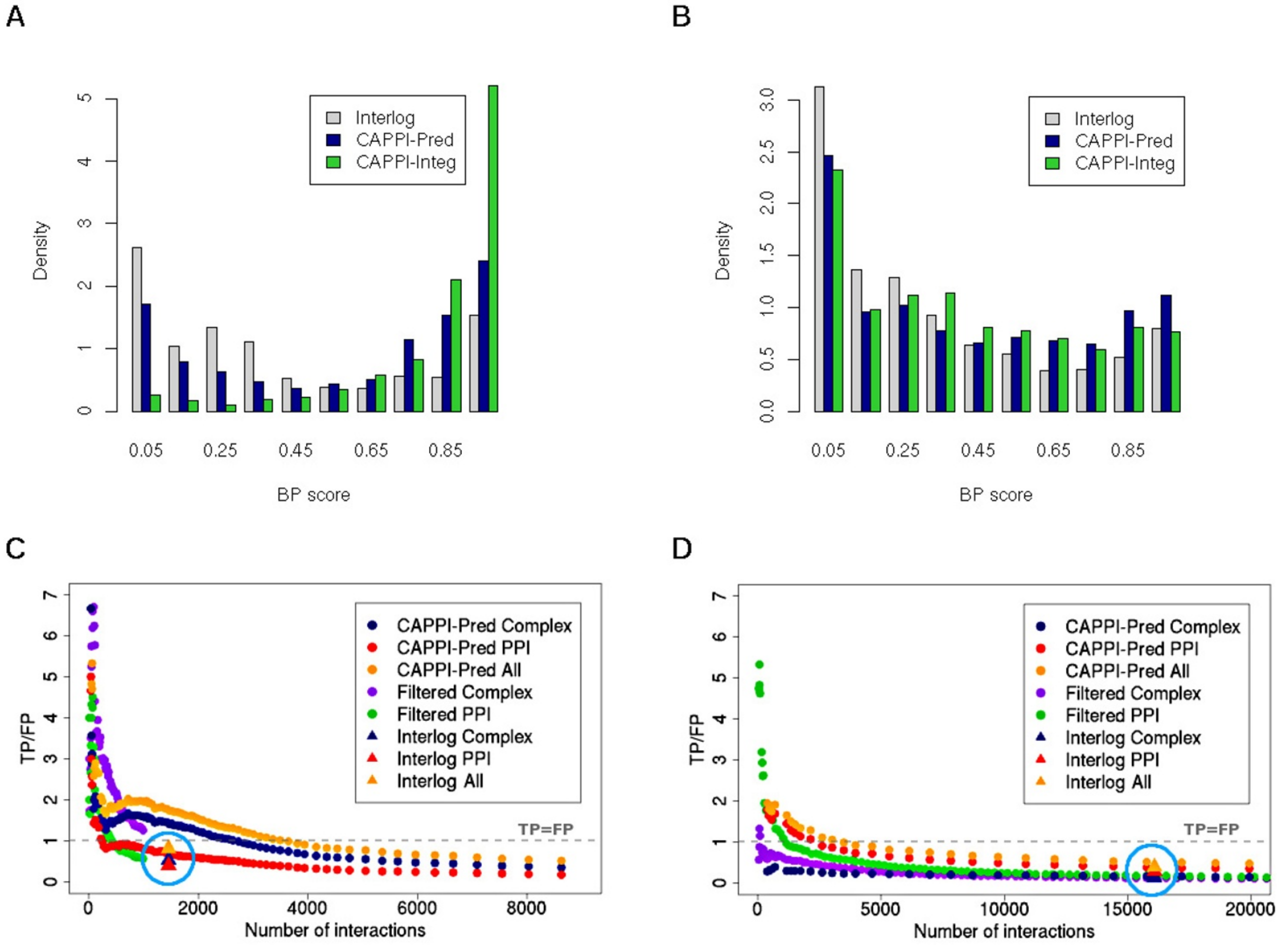

Figure 3

Assessment of CAPPI-Pred predictions. Assessment of CAPPI-Pred predictions. (A) and (B): Histograms of BP scores for the predicted yeast (A) and human (B) PPI datasets of the same size (1576 yeast PPIs and 17105 human PPIs) from the Interlog method, CAPPI-Pred and CAPPI-Integ. (C) and (D): The ratio of true positives and false positives as a function of the number of interactions in yeast (C) and human (D). An interaction is deemed true positive if it is found in the reference dataset of either co-complex interactions (Complex) or binary PPIs (PPI), or in any available reference set (ALL) and false positive, if the proteins are assigned different cellular localization (see text). Plots labeled as "Filtered Complex" and "Filtered PPI" show the results of selected CAPPI predictions which are part of dense clusters tested against either the co-complex reference (Complex) or binary reference (PPI). The gray dashed line marks the level at which the number of true positive predictions is equal to the number of false positive predictions.

interactions across species and as such it falls into the second category. Here we compare the performance of our method and two well-established frameworks for mapping PPIs: the interlog approach and the domain-based maximum likelihood method.

Comparison with the domain-based maximum likelihood approach

In [33] the domain-domain interaction prediction method was generalized to multiple species and applied to infer interactions in yeast, worm and fly (we refer to this method as the Domain-ML approach). As a final output, this approach predicts protein-protein interactions based on inferred interactions between conserved domains. Liu *et al*. trained their method using Ito, Uetz, Giot and Li experimental datasets, so the their results can be directly compared to CAPPI-Integ-3sp. Note that only the yeast interaction predictions were provided by the authors. The mean *GO* scores for Domain-ML and CAPPI are shown in Figure 2B. CAPPI-Integ-3sp significantly outperforms

Domain-ML in terms of all three *GO* scores. The performance evaluation using gold standard data (Figure 2C) also indicates a higher accuracy of CAPPI compared to the domain-based approach.

Comparison with the interlog-based approach

Next, we compare our results with a popular method of interlog mapping. This approach, similarly to CAPPI, relies on protein sequence similarity to transfer the interaction evidence across species. We choose for comparison the interlog mapping implementation from [47] and use the same input data for predicting our CAPPI-Pred dataset (for details see Additional file 1). Figure (3A and 3B) provides the distributions of *GO* scores for the Interlog and CAPPI datasets of the same size: 1576 (yeast) and 17105 (human), respectively. CAPPI predictions also contain a larger fraction of highest-scoring interactions (those with *GO* score > 0.8) and obtain a higher average score. The mean score for the CAPPI-predicted yeast dataset is noticeably higher than that of the Interlog method (0.57 vs. 0.39). CAPPI's advantage is also apparent in case of the human predictions (mean score 0.42 vs. 0.33). To assess the significance of the difference in score distributions we performed the Wilcoxon test which returned *p*-values < 2.2×10^{-16} in all cases.

Figure (3C and 3D) shows the mean scores for the Interlog output (in blue circles), which can be compared with the CAPPI rankings. In all cases CAPPI achieves a higher fraction of true positive interactions: 0.88 vs. 0.47 for the yeast co-complex predictions, 0.72 vs. 0.40 for the yeast binary PPI prediction, 0.16 vs. 0.14 for the human co-complex predictions, and 0.38 vs. 0.28 for the human binary PPI predictions. As we show in the next section, CAPPI recovers many known interactions within essential functional modules enabling the reconstruction of module subunits. The InteroPORC method is too restrictive in most of the studied cases (see Additional file 1: Table S1), suggesting that a less stringent ortholog search is needed. In fact this is recognised in [47] where more sensitive methods are considered for predicting interactions in cyanobacterium *Synechocystis*. An additional advantage of our method lies in the provided ranking (induced by the posterior probabilities), which enables the user to easily identify the most reliable interactions. As an example, for the purpose of selecting human PPI targets for verification, one could make a heuristic decision to consider only around 3,500 top predictions for which the TP/FP ratio is greater than 1 (see Figure 3D).

Case studies: mapping interactions within conserved functional modules

We now zoom-in on specific examples of functional units in the interactomes of human, yeast and thale cress, and analyze co-complex interactions inferred by CAPPI-Pred. In all described cases we demonstrate that the general topological features and organization of these complexes, as well as many known pairwise PPIs, can be recovered by our method based solely on data from the other species. We verify the inferred interactions against previously reported experimental data and assess the significance of our predictions. For an example of how the threshold selection impacts the number of interactions and the resulting *p*-value see Additional file 1: Figure S1. Note that in the following discussion gene names are used to denote corresponding proteins.

Human and yeast proteasome subnetworks

The ubiquitin-proteasome pathway is essential for eliminating damaged proteins and for regulation of intra-cellular level of proteins involved in wide spectrum of cellular functions [54]. It is conserved in eukaryotes, from yeast to human. The 26S proteasome complex contains a 20S catalytic core particle (CP), which is capped on each side by a 19S regulatory particle (RP). The structure of the 20S proteasome from yeast has been resolved [55]. It consists of 28 protein subunits: two α-rings ($\alpha1,...,\alpha7$) and two β-rings ($\beta1,...,\beta7$). The 19S proteasome can be further decomposed into two subcomplexes: the base (Rpt1-Rpt6, Rpn1, Rpn2, Rpn10 and Rpn13 - the last one probably not present in human) that binds directly to the 20S proteasome, and the lid (Rpn3, Rpn5-Rpn9, Rpn11, Rpn12 and Sem1), which is a peripheral subcomplex. In addition there is a number of transiently associated factors like p27 and S5b (the latter is apparently not present in yeast). We discuss our predictions of the 26S proteasome interactions from yeast and from human separately.

Predicted interactions in the yeast 26S proteasome are depicted in Figure 4. Overall, at the selected threshold we identify 177 confirmed interactions and 66 unconfirmed ones. The graph inferred by CAPPI is split into four parts that correspond to the four subcomplexes of the proteasome: α-ring, β-ring, lid and base. The α-ring and the β-ring have a dense set of interactions. Both of them together form a clique (i.e. every two proteins are predicted to interact), with most of the interactions being supported by experimental data. The lid and base are also very well represented and connected by 16 interactions, all of which are confirmed by previous experiments. We observe also the central role of Rpn7, which is predicted to interact with every subunit in the α- and and in the β-ring, as well as with six proteins in the lid subcomplex and eight in the base. Another hub protein identified is Rpn1, which has twelve interaction partners among the alpha and beta proteins (four of which are confirmed), seven partners in the base and seven in the lid (all having experimental support). On the other hand, the transiently associated NAS2 (p27) is predicted to interact only with the AAA-ATPase subunits (Rpt1-Rpt6) of the base subcomplex. In general, interactions within the core subcomplexes of the yeast 26S interactome are accurately recovered based solely on data from other six species, demonstrating a high level of conservation of these PPIs. The vast majority of the 66 unconfirmed predictions are localized between the characterized subcomplexes. In fact only 7 of the 44 predicted interactions between the 20S catalytic core and and the 19S regulatory particles are backed by experimental evidence in yeast. The absence of experimental data for these PPIs in *S. cerevisiae* might be explained by insufficient coverage of the yeast interactome or by possible rewiring events which changed the topology of interactions between the conserved core subunits across species. The discussion of human proteasome PPI predictions is presented in Additional file 1.

Human and yeast endosome subnetworks

The ESCRT complexes comprise a major pathway for the lysosomal degradation of transmembrane proteins (see [56]). We investigate the predicted interactions for the ESCRT complexes in human and yeast and compare the obtained results with the interactions reported in the literature. The list of proteins involved in these complexes was taken from [56].

Human ESCRT co-complex interactions as predicted by our method are depicted in Figure 5. CAPPI-Pred was able to recover all five complexes discussed in [56]. These complexes are: ESCRT-3 (well represented as a dense connected component with most edges reported in previous experiments), ESCRT-1, ESCRT-0, the Vps4 complex, and the ESCRT-2 complex. Interestingly, our results suggest that proteins CHMP1B and CHMP5 should be assigned to the ESCRT-3 complex. This association of CHMP1B and CHMP5 (consistent with the so called 'CHMP nomenclature') has been recently proposed in [57]. Moving on to the right side of the graph, we notice that the VPS4 proteins together with protein VTA1 form a triangle comprising of three reported interactions. A similar observation can be made for the ESCRT-0 complex (HGS, STAM1 and STAM2), except that the interaction STAM-STAM2 is not supported by previous experimental data. Also, the topology of interactions presented in Figure 5 suggests an important role of the TSG101 (mammalian VPS23) protein, which joins ESCRT-1 with three other complexes

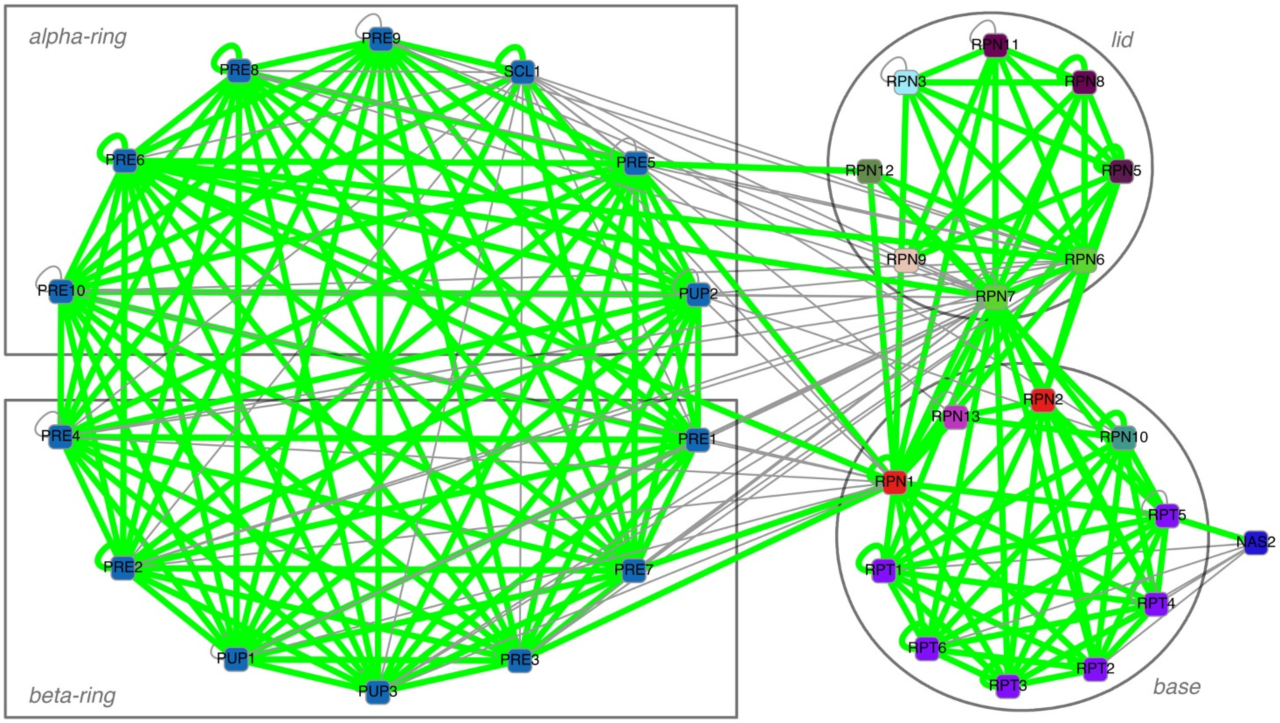

Figure 4

Inferred PPIs within the yeast proteasome complex. Interaction network of the yeast 26S proteasome complex as inferred by CAPPI-Pred. Nodes represent gene products and node colors represent protein families identified by sequence clustering. 177 of the predicted interactions which have been previously detected experimentally are denoted by green edges. 66 other PPI predictions are denoted by gray edges. The *p*-value of the predicted network is 4.348×10^{-16}. The networks are visualized using the Cytoscape software [73].

(ESCRT-3, ESCRT-0 and Vps4). TSG101 also takes part in five identified interactions within the ESCRT-1 complex, all of which have backing experimental evidence in human. Please refer to Additional file 1 for the discussion of yeast ESCRT complex predictions.

Human mRNA decay complexes

Next we investigated CAPPI's interaction predictions between proteins involved in human mRNA degradation (see [58]). The subgraph of predicted interactions is presented in Figure 6. We have a very good coverage of the human exosome complex represented by six RNase PH domain subunits (EXOSC4 (Rrp41), EXOSC5 (Rrp46), EXOSC6 (Mtr3), EXOSC7 (Rrp42), EXOSC8 (Oip2), EXOSC9 (PMScl-75)), three S1 RNA-binding domain subunits (EXOSC1 (Csl4), EXOSC2 (Rrp4), EXOSC3 (Rrp40)), and an RNase D-like subunit EXOSC10 (PMScl-100). This complex comes out as a complete subgraph (a clique) with no interactions with the other two complexes. The role of most of the subunits of the complex, in terms of interacting partners, is quite comparable. One of the exceptions is the EXOSC9 (PMScl-75) protein which is the only RNase PH domain subunit predicted to interact with DIS3 and two helicases (SKI2W and SKIV2L2). Other exosome complex members interacting with DIS3 are S1 RNA-binding subunits EXOSC1 (Csl4) and EXOSC3 (Rrp40), as well as EXOSC10. EXOSC1 and EXOSC10 also have predicted interactions with helicases SKI2W and SKIV2L2. In general, data on interactions of the peripheral subunits with the exosome complex are scarce, as reported in [58], which makes our predictions a potentially valuable target for experimental verification. The second complex which comes out as a dense subgraph in our network is the LSM complex. It consists of eight proteins (LSM1-8), forming a clique of predicted interactions, many of which are confirmed experimentally (see [58] Figure 3A). The two proteins with the largest number of confirmed interactions within the complex are LSM3 and LSM7. Both of these proteins have confirmed PPIs with six out of seven other LSM members (additional PPIs predicted by our method are LSM3-LSM4 and LSM7-LSM1). The third complex which can be retrieved from the network in Figure 6 consists of two AU-rich element ARE-binding proteins (ELAVL1 (Hur) and HNRPD (Auf1)). All three interactions predicted inside this complex are confirmed by recent experimental data (see [59]). Among the unver-

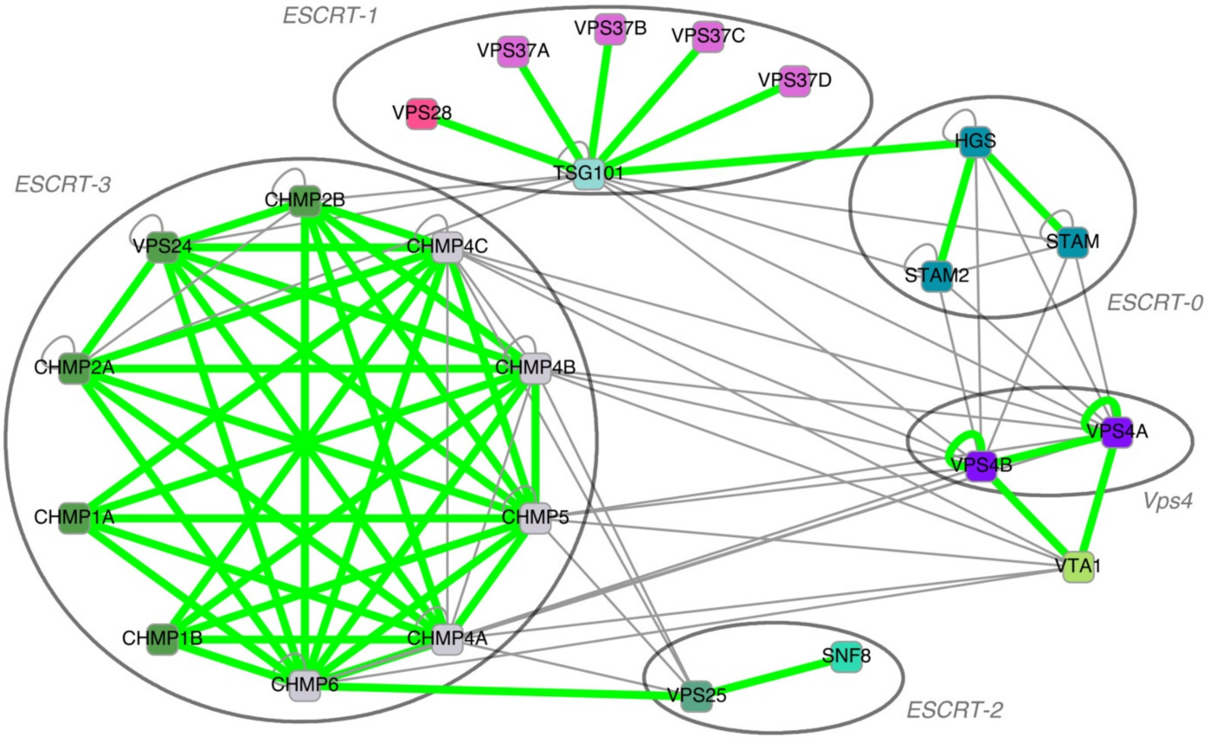

Figure 5

Inferred PPIs within the human endosome complex. Interaction network of the human endosome complexes as inferred by CAPPI-Pred. Nodes represent gene products and node colors represent protein families identified by sequence clustering. 49 predicted interactions which have been previously detected experimentally are denoted by green edges. 49 other PPI predictions are denoted by gray edges. The *p*-value of the predicted network is 3.977 × 10^{-9}.

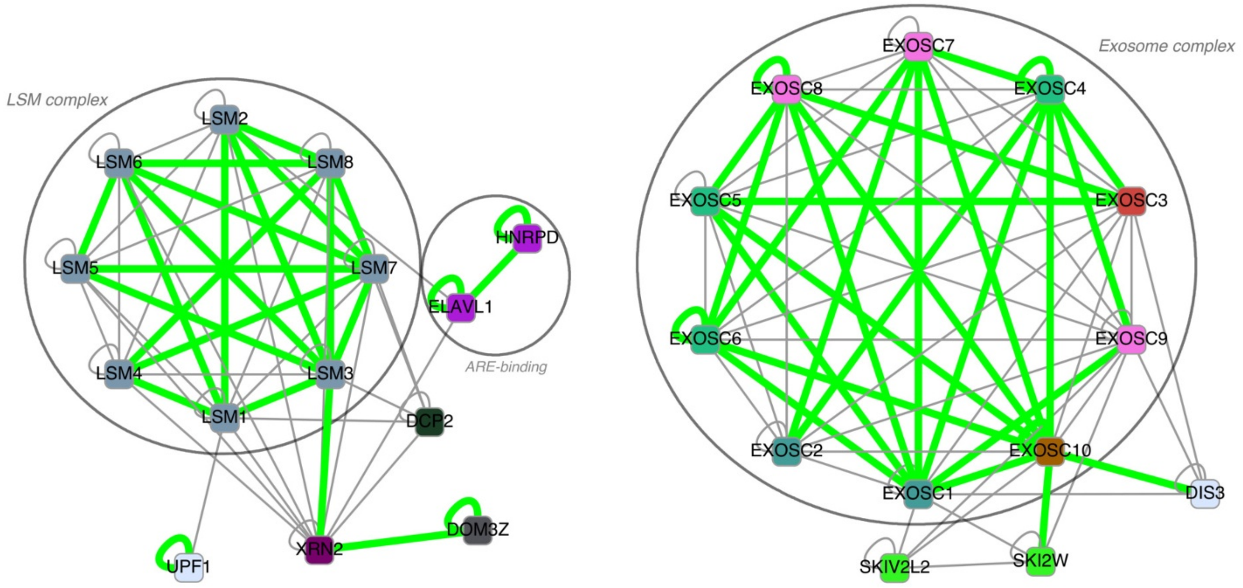

Figure 6
Inferred PPIs within the human mRNA decay complexes. Interaction network of the human mRNA decay complexes as inferred by CAPPI-Pred. Nodes represent gene products and node colors represent protein families identified by sequence clustering. 53 of the predicted interactions which have been previously detected experimentally are denoted by green edges. 76 other PPI predictions are denoted by gray edges. The *p*-value of the predicted network is 1.868 × 10^{-15}.

ified predictions is an interaction of this complex with the LSM complex (via LSM2) and with another mRNA decay factor XRN2.

A. thaliana SWI/SNF chromatin remodeling complex

In yeast and mammals, ATP-dependent chromatin remodeling complexes belonging to the SWI/SNF family play an essential role in the regulation of transcription. In Arabidopsis chromatin remodeling complexes are known to a much smaller extent. No plant SWI/SNF complex has been established and characterized to date, but it is highly probable that such complexes exist in plants (see [60]). For this reason it seems desirable to employ a computational approach for predicting interactions in the plant SWI/SNF putative complex and generate plausible working hypothesis. We present a zoom-in view of the SWI/SNF putative complex in Figure 7. A larger zoom-out view containing other homologs of the putative SWI/SNF complex members is presented in Additional file 1: Figure S2.

The graph in Figure 7 contains the core SWI/SNF proteins - the SWI3-type proteins: At2g47620 (SWI3A), At2g33610 (SWI3B), At1g21700 (SWI3C), At4g34430 (SWI3D), together with the SNF5-type protein At3g17590 (BSH). This core is presented at the bottom of the graph. In addition to the above proteins we considered four groups of Arabidopsis proteins which are reported to play a putative role in chromatin remodeling in this plant (see [60]). These are: four ATPases which are reported in [60] as potential members of the SWI/SNF complex (At2g46020 (BRM), At2g28290 (SYD), At3g06010 (Chr 12), At5g19310 (Chr 23)); two SWP73-type proteins (At3g01890 (SWP73A), and At5g14170 (SWP73B)); nine actin-related proteins (At3g27000 (ARP2), At1g13180 (ARP3), At1g18450 (ARP4), At1g73910 (ARP4A), At3g12380 (ARP5), At3g33520 (ARP6), At3g60830 (ARP7), At5g56180 (ARP8) and At5g43500 (ARP9)); and three OSA-type proteins (At1g04880, At1g76110, and At3g13350). We excluded from the graph proteins which did not show any predicted interactions. Altogether we identified 13 of 14 known interactions between the proteins visualized in Figure 7 - the missing one is At3g01890-At1g21700 (see [60]). We notice some interesting peculiarities of the presented network. Three of four of the SWI3-type proteins, are predicted to interact with the four ATPases. Only one actin-type protein (At1g18450) has a predicted interaction with the SWI/SNF core and only two more (At3g60830 and At5g56180) can be associated with the complex through member ATPases. The ability to make distinctions within homologous groups is an important feature of our approach. While methods mapping interactions to highly similar orthologs usually make very specific predictions and avoid false-positives, they are also likely to miss many true interactions which can be inferred from slightly less similar proteins. As summarised in Additional file 1: Table S1, the restrictive search applied in InteroPORC fails to map the known interactions in the SWI/SNF complex in *A. thaliana*. In fact according to the PORC ortholog clusters, only two proteins (SWI3C and SWP73A) have orthologs in any of the other six eukaryotic species considered here. In this case, a less stringent method is clearly needed. On the other hand, CAPPI bases its prediction on evidence from all homologs and thus is in danger of loosing sensitivity and assigning the same interactions to all family members. The above examples demonstrate that we can avoid these potential pitfalls by considering family members in phylogenetic context when integrating and distributing the interaction evidence.

These observations are strengthened when we consider the larger family-oriented view of the SWI/SNF-related network in Additional file 1: Figure S2. This graph was obtained from the one in Figure 7 by expanding the set of proteins to all members of the considered protein families (once again, proteins without any interactions were removed). Interestingly, the four peripheral families represented in the graph can be divided into smaller subfamilies based on the interactions partners of their members. Specifically, of the 14 ATPases presented in the larger graph only the four above described are predicted to interact directly with the core of the SWI/SNF complex. Two of them (At2g46020 (BRM) and At2g28290 (SYD)) have confirmed interactions while for the other two (At3g06010 (Chr 12), At5g19310 (Chr 23)) interaction hypothesis based on sequence similarity were formulated

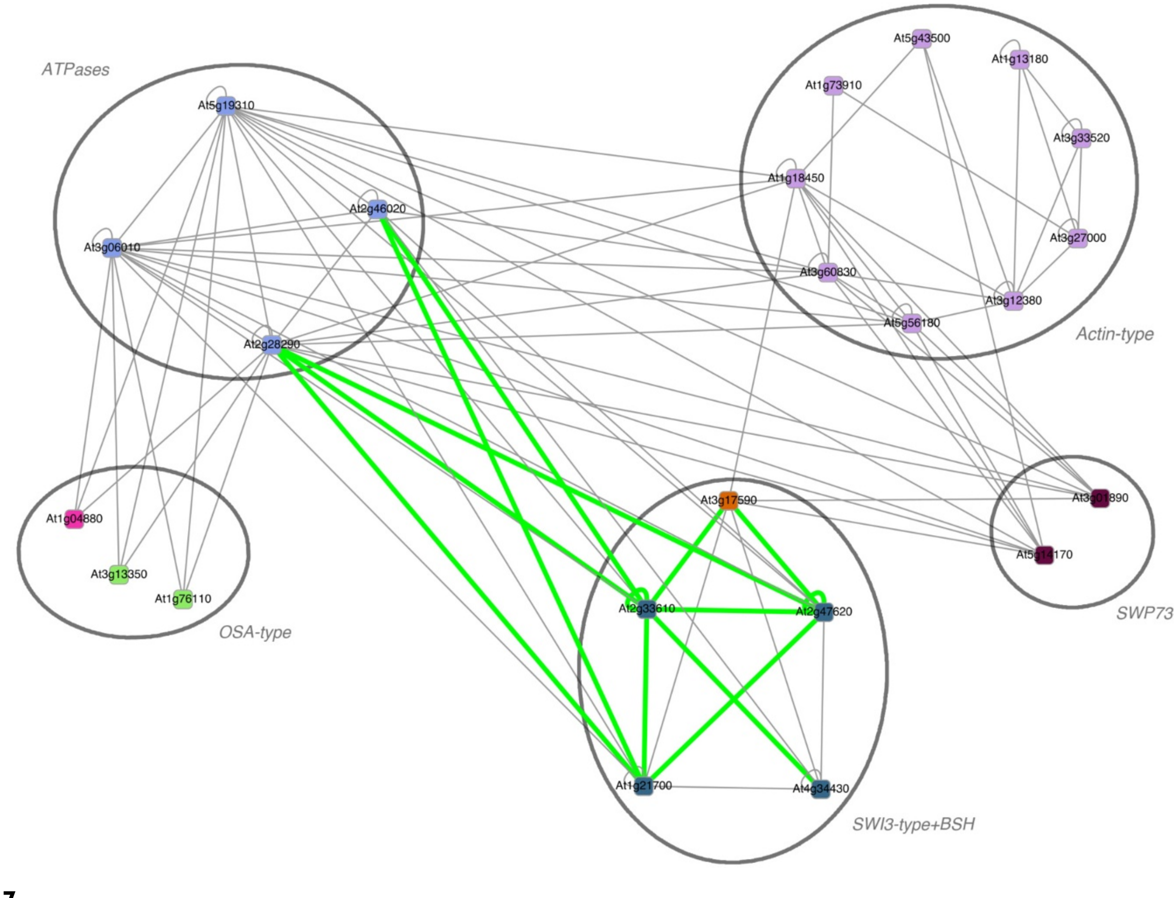

Figure 7

Inferred PPIs within the A. thaliana SWI/SNF chromatin remodeling complex. Interaction network of the putative SWI/SNF complex in Arabidopsis as inferred by CAPPI-Pred. Nodes represent gene products and node colors represent protein families identified by sequence clustering. 13 of the predicted interactions which have been previously detected experimentally are denoted by green edges. 83 other PPI predictions are denoted by gray edges. The *p*-value of the predicted network is 6.381 × 10^{-10}.

[60]. In fact the entire ATPase family, as detected by our method, contains 48 Arabidopsis proteins (a vast majority not having any predicted interactions to other proteins in the SWI/SNF subnetwork), which makes the presented predictions even more significant. These specific cases of confirmed predictions let us suggest that some of the distinctive members of the other protein families predicted to interact with the putative SWI/SNF complex (At1g18450 and six OSA family members interacting with At3g17590, five SWP73 family members interacting either with At3g17590 or at least one of the SWI3-type proteins, as well as five other actin family members interacting with ATPases At2g46020 and At2g28290), may pose valuable targets for future experimental validation.

Conclusion

We have presented a systematic phylogeny-based framework for reconciling PPI datasets across species and inferring missing interactions. Our method naturally incorporates interaction evidence from different species and experimental sources. It considers the reliability of each source and the evolutionary relationships between protein pairs. The approach was successfully applied to compute integrated interactomes for seven eukaryotic species, providing confidence scores for each possible edge in each network. Detailed analysis of our predictions indicates that we can accurately recover known interactions within conserved protein complexes. Confirmed interactions identified in a blind test provide a strong case for our top-ranked predictions, many of which await experimental verification. We also find that while core subcomplexes can be accurately recovered based solely on the data from distant species, many of the between-module interactions are harder to identify this way, suggesting possible rewiring events. One natural direction for future research is to extend our framework to include other kinds of data which may serve as indirect evidence of interaction. The integration of heterogeneous experimental sources with account of the phylogenetic model may possibly improve existing catalogues of functional associations.

Methods

Bayesian model of network evolution

We start by briefly recapitulating the network growth model from [39] which, given the ancestral network $G_{1,0}$ determines the probability of interaction between proteins at every stage of evolution. The model has four parameters: p_d, δ_d, p_s and δ_s. It assumes that starting from the ancestral graph $G_{1,0}$ a sequence of duplications and speciations is performed where these events are determined by reconciled phylogenetic trees precomputed for each protein family. We denote by $G_{i,j} = (V_{i,j}, E_{i,j})$ the graph representing the protein network of s_i after the j-th duplication event occurring in this species. In case of a node duplication event we replace the node by two copies. For each copy we retain each of its edges with probability p_d and insert edges adjacent to the copy with probability δ_d (independently for each copy and each edge). In case of a speciation event we make two copies of the network. In each network copy we retain each edge with probability p_s and insert each non-existent edge with probability δ_s (independently for each network and each edge). Assuming this model and the provided phylogeny of each protein family we construct a Bayesian network (BN) model of protein interactions at all levels of evolution. In this BN model the probability $P(X_{n_x,n_y}^{G_{i,j}} = 1)$ of interaction between a pair of nodes n_x, $n_y \in V_{i,j}$, depends on the existence or lack of an edge between the protein pair being the direct evolutionary predecessor (either before speciation or duplication) of the pair (n_x, n_y) (see Figure 8). A detailed description of the model is available in [39].

Integrating diverse experimental data

The above-described model captures the basic notions of protein network evolution. We previously assumed that the PPI data is free of error and complete and we used the model to make inferences about the ancestral interaction networks. However, due to experimental errors and incomplete sampling, the real interactions and non-interacting protein pairs are not certain. This implies that the experimental data should only be used as supporting evidence of putative interactions. To model this accurately in our framework we keep the random variables corresponding to extant interactions unknown and add another level of random variables corresponding to experimental evidence (see Figure 8A). The evidence in each experimental dataset is weighted by the dataset's reliability.

Let $G_{i,m_i} = (V_{i,m_i}, E_{i,m_i})$ be the extant protein interaction network of a present-day species s_i (we assume that m_i is the final duplication occurring in s_i). Let $O_i = \{o_1^{(i)}, \dots, o_{k_i}^{(i)}\}$ be the set of experimental datasets for species s_i, where each $o_h^{(i)}$ is the set of protein pairs confirmed to interact in the h-th experiment. Let $Rel(o_h^{(i)})$ be the fraction of elements in $o_h^{(i)}$ believed to be true positives. Let $E'_{i,m_i} = \{(n_x, n_y) : n_x, n_y \in V_{i,m_i} \wedge (n_x, n_y) \notin E_{i,m_i}\}$ be the set of non-interacting protein pairs in the graph G_{i,m_i}. For

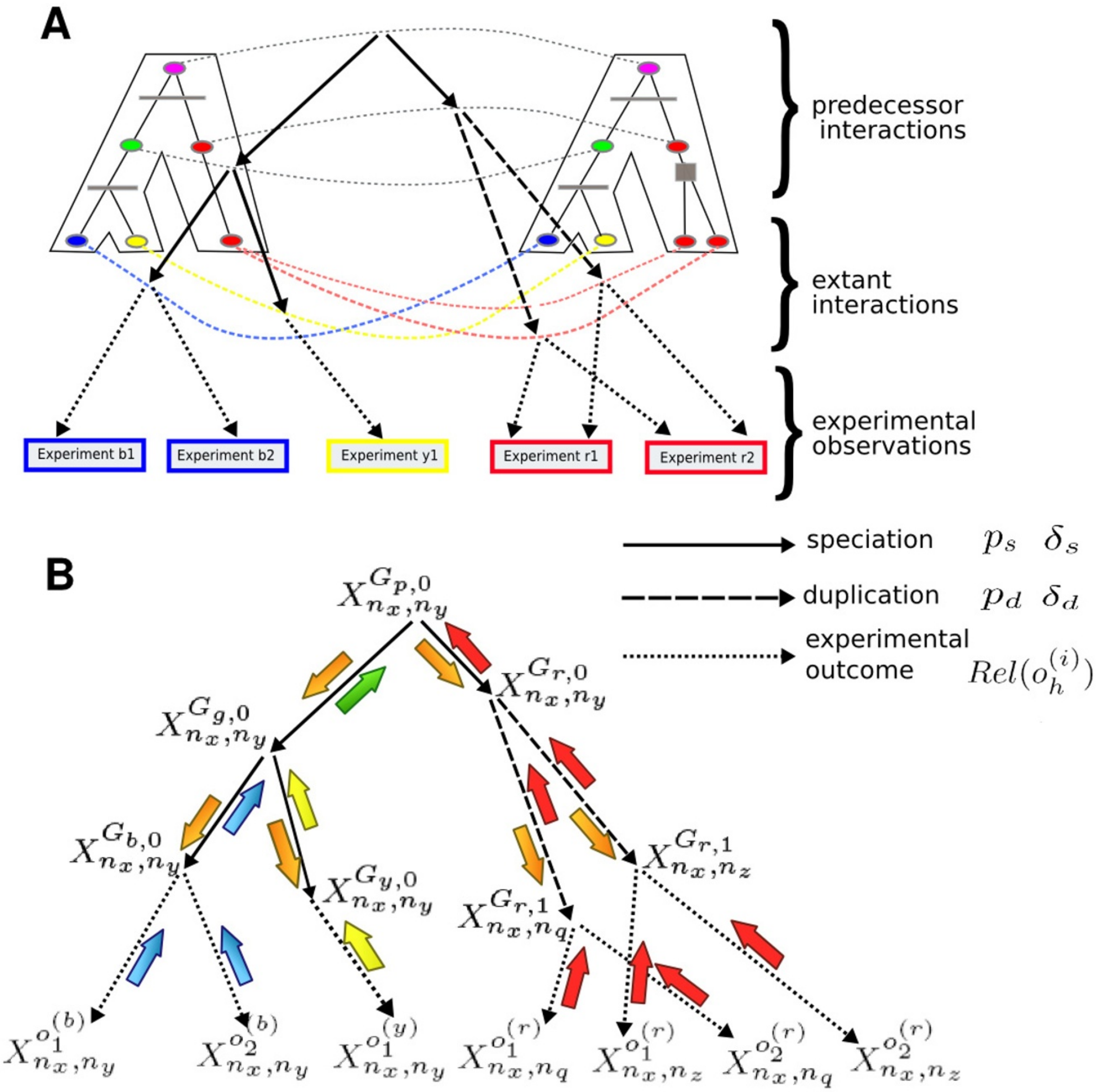
A
predecessor interactions
extant interactions
experimental observations
Experiment b1
Experiment b2
Experiment y1
Experiment r1
Experiment r2
B
speciation p_s δ_s
duplication p_d δ_d
experimental outcome $Rel(o_h^{(i)})$
$X^{G_{p,0}}_{n_x,n_y}$
$X^{G_{g,0}}_{n_x,n_y}$
$X^{G_{r,0}}_{n_x,n_y}$
$X^{G_{b,0}}_{n_x,n_y}$
$X^{G_{y,0}}_{n_x,n_y}$
$X^{G_{r,1}}_{n_x,n_q}$
$X^{G_{r,1}}_{n_x,n_z}$
$X^{o_1^{(b)}}_{n_x,n_y}$
$X^{o_2^{(b)}}_{n_x,n_y}$
$X^{o_1^{(y)}}_{n_x,n_y}$
$X^{o_1^{(r)}}_{n_x,n_q}$
$X^{o_1^{(r)}}_{n_x,n_z}$
$X^{o_2^{(r)}}_{n_x,n_q}$
$X^{o_2^{(r)}}_{n_x,n_z}$

Figure 8

Inferring protein interactions via message passing. A toy example of the Bayesian tree model of evolution of interactions between members of two protein families for three species: <u>b</u>lue, <u>y</u>ellow and <u>r</u>ed. For each species a certain number of experimental datasets is given: two for blue and red and one for yellow. Part (A) shows two reconciled trees for the considered families together with putative protein interactions at each level of evolution. The proteins in the trees are represented by ellipses (with color corresponding to their species). The speciation events are marked by horizontal lines and the duplication events are marked by filled squares. The evolution of the ancestral interaction between the root proteins (purple) can be traced down the trees to the extant interactions. Evidence for the extant interactions can be found in the experimental datasets. In (B) a random variable is associated with each putative interaction. A solid arrow indicates a dependence between two random variables which comes from the speciation event. Similarly, a dashed arrow indicates a dependence for the duplication event. Finally, dotted arrows represent an interface between the true interactions in extant species and the observed experimental evidence. The parameters p_s, δ_s, p_d and δ_d determine the probability of retaining or gaining an interaction during evolution, while the reliability of each dataset (*Rel*($o_h^{(i)}$)) determines the probability of identifying a true interaction or a false positive one. Arrows colored blue, yellow, red and green represent messages corresponding to interaction evidence coming from each of the species. These messages are passed up the tree in the first phase of the MP algorithm. In the second phase, messages containing aggregated evidence from one side of the tree are passed down to the other side (orange arrows).

each experimental dataset $o_h^{(i)}$ we denote by $X_{n_x,n_y}^{o_h^{(i)}}$ a random variable which takes value 1 if interaction (n_x, n_y) is present in this dataset and 0 otherwise. For each pair of proteins (n_x, n_y) and each dataset $o_h^{(i)}$ we set the probability of observing a true interaction to be equal the true positive rate of the experiment, and the probability of observing a false positive interaction equal the false positive rate of the experiment, as follows:

$$Pr(X_{n_x,n_y}^{o_h^{(i)}} = 1 \mid X_{n_x,n_y}^{G_{i,m_i}} = 1) = \frac{Rel(o_h^{(i)})|o_h^{(i)}|}{\left|E_{i,m_i}\right|}$$

$$Pr(X_{n_x,n_y}^{o_h^{(i)}} = 1 \mid X_{n_x,n_y}^{G_{i,m_i}} = 0) = \frac{(1-Rel(o_h^{(i)}))|o_h^{(i)}|}{\left|E'_{i,m_i}\right|},$$

where by $|A|$ we denote the number of elements in the set A. Now each experimentally observed interaction can be naturally incorporated into the BN framework. Similarly each pair not observed to interact in the considered experiment ((n_x, n_y) $\notin$ $o_h^{(i)}$) can be incorporated into the model with conditional probabilities corresponding to the false negative rate and true negative rate of the experiment (see Additional file 1 for details). The model can also be easily generalized to incorporate distinct reliability values for each single interaction.

Inferring extant protein interactions via message passing

The integrated BN model, comprising all PPI edges from every level of evolution and from the experimental datasets, is used to infer protein interactions in the input species. Each random variable corresponding either to a possible interaction, or to a single experiment outcome, depends on exactly one random variable which denotes an edge (or non-edge) in the direct evolutionary predecessor in the first case, and in the network of an extant species in the second case. The considered BN model is a set of Bayesian trees, where each tree represents the joint distribution of the random variables corresponding to putative interactions (which descended from a single edge in the ancestral graph) and the associated experimental evidence (an example of such tree is shown in Figure 8B). The tree structure allows us to apply Pearl's message passing (MP) algorithm [61] to compute the exact posterior probability of interaction between proteins in extant species, in time linear to the number of random variables (see Figure 8B for an example and [61] or [62] for details). Specifically we determine the posterior probability of interaction $P(X_{n_x,n_y}^{G_{i,m_i}} = 1|O)$ for each pair of nodes (n_x, n_y) in each extant network G_{i,m_i}, where O denotes all experimental datasets for all species.

Assessing PPI predictions in large-scale studies

In general, the assessment of PPI predictions posses problems due to the limited number of "gold standard" interactions and the lack of negative test cases. Motivated by previous studies, we employ two scoring schemes to assess the quality of predicted PPIs, as well as those from the input datasets. The first one compares Gene Ontology (GO) annotations [63] of adjacent gene products and measures their functional similarity. Functional similarity is used as an indirect measure of interaction: the more similar the annotations of the two proteins are, the more confident we are in deeming an interaction between them. We apply a recent information content method [48], implemented in the SemSim R package by Xiao Gou: http://www.bioconductor.org/packages/2.0/bioc/html/SemSim.html, which extends the measures previously proposed by [64] and [65]. For each pair of proteins we individually measure the similarity of annotations in each of the three ontologies: biological process (BP), molecular function (MF) and cellular component (CC). This results in a *BP* score, *MF* score and *CC* score, respectively, each ranging from 0 (no similarity) to 1 (maximum similarity). When the context allows, we refer to each of these scores as a *GO* score of a pair of proteins.

Our second kind of quality assessment is based on a comparison with a reference dataset. We estimate the ratio of true positive interactions (predictions which are confirmed in a reference dataset) and putative false positive interactions (unconfirmed predictions for which the two proteins have disjoint cellular localizations). A similar procedure was applied in [29]. We use separate reference datasets for binary PPIs (direct physical interactions) and for co-complex PPIs (pairs of proteins co-occurring within the same complex). For details on the reference datasets and the localization data see Additional file 1. Note that the proper sensitivity and specificity measures are hard to estimate because the reference sets of positive interactions and negative protein pairs are not comprehensive. Due to interdependencies between interactions, implied by our model, cross-validation cannot be easily applied. Instead, we perform a blind test in which we leave out the data of one species and predict its interactions only based on the data from the other species.

Assessing predictions in functional module case-studies

For small-scale functional module case studies we report all interactions predicted among a determined set of proteins for a selected threshold value. To assess the statistical significance of interaction predictions we compute a *p*-value based on the hypergeometric distribution, where confirmed interactions are regarded as successes and unconfirmed interactions are regarded as failures (Fisher's exact test). As the predictions are made by CAPPI-Pred which is trained without the use of the input datasets for the predicted species, we use the held out input data as a reference. Note that it is possible that some of the reference interactions are in fact false-positives - an inherent risk of using high-throughput data. In this particular test, however, we are interested in assessing the possibility to predict a significant portion of known PPIs (of which many are from high-throughput studies) by a mapping from other organisms. The reference set is further extended in each case by PPIs curated from specific publications characterizing interactions within the studied complexes. These are as follows: [66,67] for the 26S proteasome PPIs, [56,57] for the endosome-related PPIs, [58] for the exosome-related PPIs, and [68-72] for the SWI/SNF-related PPIs. Note that for *A. thaliana* there are no high-throughput datasets available, so all reference data for this species come from small-scale studies.

Authors' contributions

The authors together conceived the study and analysed the results. JD implemented the framework, computed the integrated interactomes and analyzed the high-throughput experiments. JT provided initial analysis of the small-scale case studies. Both authors contributed to writing the manuscript and approved its final version.

Additional material

Additional file 1

Supplementary material*. This file contains supplementary text (describing data acquisition and applied methods) as well as supplementary table and figures.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-393-S1.PDF]

Acknowledgements

We would like to thank Andrzej Jerzmanowski, Maciej Kotlieski and Marlena Roszczyk for many helpful discussions. This work was supported by the Polish Ministry of Science grants No 0652/B/P01/2009/36 and PBZ-MNiI-2/1/2005, and by the CoE BioExploratorium project WKP 1/1.4.3/1/2004/44/44/115.

References

1. Uetz P, Giot L, Cagney G, Mansfield T, Judson R, Knight J, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg J: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403:**623-627.
2. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci USA* 2001, **98:**4569-4574.
3. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M: **Towards a proteome-scale map of the human protein-protein interaction network.** *Nature* 2005, **437:**1173-1178.
4. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksöz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122(6):**957-968.
5. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G: **Proteome survey reveals modularity of the yeast cell machinery.** *Nature* 2006, **440:**631-636.
6. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregran-Alvarez JaM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MHY, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles JC, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF: **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.** *Nature* 2006, **440:**637-643.
7. Hart TG, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biology* 2006, **7:**120.
8. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual JF, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan C, de Smet AS, Motyl A, Hudson ME, Park J, Xin X, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, Barabasi AL, Tavernier J, Hill DE, Vidal M: **High-Quality Binary Protein Interaction Map of the Yeast Interactome Network.** *Science* 2008, **3:**104-110.
9. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet AS, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone C, Braun P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabási AL, Vidal M: **An empirical framework for binary interactome mapping.** *Nature Methods* 2009, **6:**83-90.
10. Mrowka R, Patzak A, Herzel H: **Is There a Bias in Proteome Research?** *Genome Research* 2001, **11:**1971-1973.
11. von Mering C, Krause R, Snel B, Cornell M, Oliver S, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417:**399-403.
12. Bader G, Hogue C: **Analyzing yeast protein-protein interaction data obtained from different sources.** *Nature Biotechnology* 2002, **20:**991-997.
13. Schwartz AS, Yu J, Gardenour KR, Finley RL Jr, Ideker T: **Cost-effective strategies for completing the interactome.** *Nature Methods* 2008, **6:**55-61.
14. Valencia A, Pazos F: **Computational methods for the prediction of protein interactions.** *Current Opinion in Structural Biology* 2002, **12:**368-373.
15. Shoemaker B, Panchenko A: **Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners.** *PLoS Computational Biology* 2007, **3:**e43.
16. Valencia A, Pazos F: **Computational methods to predict protein interaction partners.** In *Protein-protein Interactions and Networks: Identification, Computer Analysis, and Prediction* Edited by: Panchenko A, Przytycka T. Springer; 2008:67-81.

17. Enright A, Iliopoulos I, Kyrpides N, Ouzounis C: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402:**86-90.

18. Marcotte E, Pellegrini M, Ng H, Rice D, Yeates T, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285:**751-753.

19. Gomez S, Noble W, Rzhetsky A: **Learning to predict protein-protein interactions from protein sequences.** *Bioinformatics* 2003, **19:**1875-1881.

20. Burger L, van Nimwegen E: **Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method.** *Mol Syst Biol* 2008, **4(165):**.

21. Pellegrini M, Marcotte E, Thompson M, Eisenberg D, Yeates T: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96:**4285-4288.

22. Chen XW, Liu M: **Prediction of protein-protein interactions using random decision forest framework.** *Bioinformatics* 2005, **21:**4394-4400.

23. Bader J, Chaudhuri A, Rothberg J, Chant J: **Gaining confidence in high-throughput protein interaction networks.** *Nature Biotechnology* 2004, **22:**78-85.

24. Bleakley K, Biau G, Vert JP: **Supervised reconstruction of biological networks with local models.** *Bioinformatics* 2007, **23:**57-65.

25. Pazos F, Valencia A: **Similarity of phylogenetic trees as indicator of protein-protein interaction.** *Protein Engineering* 2001, **14:**609-614.

26. Jothi R, Kann M, Przytycka T: **Predicting protein-protein interaction by searching evolutionary tree automorphism space.** *Bioinformatics* 2005, **21:**241-250.

27. Juan D, Pazos F, Valencia A: **High-confidence prediction of global interactomes based on genome-wide coevolutionary networks.** *Proc Natl Acad Sci USA* 2008, **105(3):**934-939.

28. Deng M, Mehta S, Sun F, Chen T: **Inferring Domain-Domain Interactions From Protein-Protein Interactions.** *Genome Research* 2002, **12:**1540-1548.

29. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan N, Chung S, Emili A, Snyder M, Greenblatt J, Gerstein M: **A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data.** *Science* 2003, **302:**449-453.

30. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C: **STRING 8-a global view on proteins and their functional interactions in 630 organisms.** *Nucleic acids research* 2009:D412-D416.

31. Alexeyenko A, Sonnhammer EL: **Global networks of functional coupling in eukaryotes from comprehensive data integration.** *Genome research* 2009, **19(6):**1107-16.

32. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M: **Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs".** *Genome Research* 2001, **11(12):**2120-2126.

33. Liu Y, Liu N, Zhao H: **Inferring protein-protein interactions through high-throughput interaction data from diverse organisms.** *Bioinformatics* 2005, **21:**3279-3285.

34. Lee H, Deng M, Sun F, Chen T: **An integrated approach to the prediction of domain-domain interactions.** *BMC Bioinformatics* 2006, **7:**269.

35. Sharan R, Ideker T: **Modeling cellular machinery through biological network comparison.** *Nature Biotechnology* 2006, **24(4):**427-433.

36. Beltrao P, Serrano L: **Specificity and evolvability in eukaryotic protein interaction networks.** *PLoS Computational Biology* 2007, **3(2):**e25.

37. Bork P, Jensen L, von Mering C, Ramani A, Lee I, Marcotte E: **Protein interaction networks from yeast to human.** *Current Opinion in Structural Biology* 2004, **14:**292-299.

38. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T: **Conserved patterns of protein interaction in multiple species.** *Proc Natl Acad Sci USA* 2005, **102(6):**1974-1979.

39. Dutkowski J, Tiuryn J: **Identification of functional modules from conserved ancestral protein-protein interactions.** *Bioinformatics* 2007, **23:**i149-i158.

40. Pinney JW, Amoutzias GD, Rattray M, Robertson DL: **Reconstruction of ancestral protein interaction networks for the bZIP transcription factors.** *Proc Natl Acad Sci USA* 2007, **104:**20449-20453.

41. Kersey P, Bower L, Morris L, Horne A, Petryszak R, Kanz C, Kanapin E, Das U, Michoud K, Phan I, Gattiker R, Kulikova T, Faruque N, Duggan K, Mclaren P, Reimholz B, Duret L, Penel S, Reuter I, Apweiler R: **Integr8 and Genome Reviews: integrated views of complete genomes and proteomes.** *Nucleic Acids Research* 2005, **33:**297-302.

42. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R: **IntAct: an open source molecular interaction database.** *Nucleic Acids Reseach* 2004:D452-D455.

43. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular INTeraction database.** *Nucleic Acids Research* 2007:D572-D574.

44. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Research* 2004:D449-D451.

45. Giot L, Bader J, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao Y, Ooi C, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, Dasilva A, Zhong J, Stanyon C, Finley JR, White K, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets R, McKenna M, Chant J, Rothberg J: **A protein interaction map of Drosophila melanogaster.** *Science* 2003, **302:**1727-1736.

46. Li Siming, Armstrong C, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg D, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong S, Zhang L, Berriz G, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel H, Elewa A, Baumgartner B, Rose D, Yu H, Bosak S, Sequerra R, Fraser A, Mango S, Saxton W, Strome S, Heuvel S van den, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus K, Harper J, Cusick M, Roth F, Hill D, Vidal M: **A map of the interactome network of the metazoan C. elegans.** *Science* 2004, **303:**540-543.

47. Michaut M, Kerrien S, Montecchi-Palazzi L, Chauvat F, Cassier-Chauvat C, Aude JCC, Legrain P: **InteroPorc: Automated Inference of Highly Conserved Protein Interaction Networks.** *Bioinformatics* 2008, **24(14):**1625-1631.

48. Schlicker A, Domingues F, Rahnenführer J, Lengauer T: **A new measure for functional similarity of gene products based on Gene Ontology.** *BMC Bioinformatics* 2006, **7:**302.

49. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon G, Myers C, Parsons A, Friesen H, Oughtred R, Tong A, Stark C, Ho Y, Botstein D, Andrews B, Boone C, Troyanskya O, Ideker T, Dolinski K, Batada N, Tyers M: **Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae.** *Journal of Biology* 2006, **5(11):**.

50. Mewes HW, Frishman D, Mayer KF, Münsterkötter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Sümpflen V: **MIPS: analysis and annotation of proteins from whole genomes in 2005.** *Nucleic Acids Research* 2006:D169-D172.

51. Pu S, Wong J, Turner B, Cho E, Wodak SJ: **Up-to-date catalogues of yeast protein complexes.** *Nucleic Acids Research* 2009, **37:**825-831.

52. Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C: **Estimating the size of the human interactome.** *Proc Natl Acad Sci USA* 2008, **105:**6959-6964.

53. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Research* 2002, **30(7):**1575-1584.

54. Glickman MH, Ciechanover A: **The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction.** *Physiol Rev* 2002, **82(2):**373-428.

55. Groll M, Huber R: **Purification, crystallization, and X-ray analysis of the yeast 20S proteasome.** *Methods in Enzymology* 2005, **398:**329-336.

56. Hurley JH, Emr SD: **The ESCRT complexes: structure and mechanism of a membrane-trafficking network.** *Annual Review of Biophysics and Biomolecular Structure* 2006, **35:**277-298.

57. Shim S, Merrill SA, Hanson PI: **Novel interactions of ESCRT-III with LIP5 and VPS4 and their implications for ESCRT-III disassembly.** *Molecular Biology of the Cell* 2008, **19:**2661-2672.

58. Lehner B, Sanderson CM: **A protein interaction framework for human mRNA degradation.** *Genome Research* 2004, **14:**1315-1323.
59. David PS, Tanveer R, Port JD: **FRET-detectable interactions between the ARE binding proteins, HuR and p37AUF1.** *RNA* 2007, **13:**1453-1468.
60. Jerzmanowski A: **SWI/SNF chromatin remodeling and linker histones in plants.** *Biochimica et Biophysica Acta* 2007, **1769:**330-345.
61. Pearl J: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* Morgan Kaufmann; 1988.
62. Neapolitan RE: *Learning Bayesian Networks* Prentice Hall; 2003.
63. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature Genetics* 2000, **25:**25-29.
64. Resnik P: **Using Information Content to Evaluate Semantic Similarity in a Taxonomy.** *Proceedings of the 14th International Joint Conference on Artificial Intelligence* 1995:448-453.
65. Lin D: **An information-theoretic definition of similarity.** *Proc 15th International Conf on Machine Learning, Morgan Kaufmann, San Francisco, CA* 1998:296-304.
66. Chen C, Huang C, Chen S, Liang J, Lin W, Ke G, Zhang H, Wang B, Huang J, Han Z, Ma L, Huo K, Yang X, Yang P, He F, Tao T: **Subunit-subunit interactions in the human 26S proteasome.** *Proteomics* 2008, **8:**508-520.
67. Cagney G, Uetz P, Fields S: **Two-hybrid analysis of the Saccharomyces cerevisiae 26S proteasome.** *Physiological Genomics* 2001, **7:**27-34.
68. Sarnowski TJ, Swiezewski S, Pawlikowska K, Kaczanowski S, Jerzmanowski A: **AtSWI3B, an Arabidopsis homolog of SWI3, a core subunit of yeast Swi/Snf chromatin remodeling complex, interacts with FCA, a regulator of flowering time.** *Nucleic Acids Research* 2002, **30:**3412-3421.
69. Farrona S, Hurtado L, Bowman JL, Reyes JC: **The Arabidopsis thaliana SNF2 homolog AtBRM controls shoot development and flowering.** *Development* 2004, **131:**4965-4975.
70. Sarnowski TJ, Ríos G, Jásik J, Swiezewski S, Kaczanowski S, Li Y, Kwiatkowska A, Pawlikowska K, Kozbial M, Kozbial P, Koncz C, Jerzmanowski A: **SWI3 subunits of putative SWI/SNF chromatin-remodeling complexes play distinct roles during Arabidopsis development.** *The Plant Cell* 2005, **17:**2454-2472.
71. Hurtado L, Farrona S, Reyes JC: **The putative SWI/SNF complex subunit BRAHMA activates flower homeotic genes in Arabidopsis thaliana.** *Plant Molecular Biology* 2006, **62:**291-304.
72. Bezhani S, Winter C, Hershman S, Wagner JD, Kennedy JF, Kwon CS, Pfluger J, Su Y, Wagner D: **Unique, shared, and redundant roles for the Arabidopsis SWI/SNF chromatin remodeling ATPases BRAHMA and SPLAYED.** *The Plant Cell* 2007, **19:**403-416.
73. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Research* 2003, **13(11):**2498-2504.