Research article

Probabilistic prediction and ranking of human protein-protein interactions
Michelle S Scott and Geoffrey J Barton*

Address: School of Life Sciences Research, College of Life Sciences, University of Dundee, Scotland, UK

Email: Michelle S Scott - michelle@compbio.dundee.ac.uk; Geoffrey J Barton* - geoff@compbio.dundee.ac.uk

* Corresponding author

Abstract

Background: Although the prediction of protein-protein interactions has been extensively investigated for yeast, few such datasets exist for the far larger proteome in human. Furthermore, it has recently been estimated that the overall average false positive rate of available computational and high-throughput experimental interaction datasets is as high as 90%.

Results: The prediction of human protein-protein interactions was investigated by combining orthogonal protein features within a probabilistic framework. The features include co-expression, orthology to known interacting proteins and the full-Bayesian combination of subcellular localization, co-occurrence of domains and post-translational modifications. A novel scoring function for local network topology was also investigated. This topology feature greatly enhanced the predictions and together with the full-Bayes combined features, made the largest contribution to the predictions. Using a conservative threshold, our most accurate predictor identifies 37606 human interactions, 32892 (80%) of which are not present in other publicly available large human interaction datasets, thus substantially increasing the coverage of the human interaction map. A subset of the 32892 novel predicted interactions have been independently validated. Comparison of the prediction dataset to other available human interaction datasets estimates the false positive rate of the new method to be below 80% which is competitive with other methods. Since the new method scores and ranks all human protein pairs, smaller subsets of higher quality can be generated thus leading to even lower false positive prediction rates.

Conclusion: The set of interactions predicted in this work increases the coverage of the human interaction map and will help determine the highest confidence human interactions.

Background
Protein-protein interactions perform and regulate fundamental cellular processes. The comprehensive study of such interactions on a genome-wide scale will lead to a clearer understanding of diverse cellular processes and of the molecular mechanisms of disease. Although the determination of interactions by small-scale laboratory techniques is impractical for a complete proteome on the grounds of cost and time, several experimental techniques now exist to determine protein-protein interactions in a high-throughput manner [1]. High-throughput datasets have been generated for model organisms such as yeast [2-6], worm [7] and fly [8,9] as well as Escherichia coli [10]. In addition, the first broad-focus experimental datasets for the human interactome have recently been published [11,12]. Interactions determined by high-throughput

methods are generally considered to be less reliable than those obtained by low-throughput studies [13,14] and as a consequence efforts are also underway to extract evidence for interactions from the literature [15-17]. Analysis of the high-throughput datasets has shown that they overlap very little with each other, suggesting that their coverage is low. Indeed, it has been estimated recently that the current yeast and human protein interaction maps are only 50% and 10% complete, respectively [18].

The low coverage and variable quality of the experimental interaction datasets have prompted many groups to investigate computational methods to predict interactions or to determine the most likely interactions seen in the high-throughput datasets. The different approaches to predict interactions can be grouped into five main categories:

1) Predictors based on sequence and structure exploit the observation that some pairs of sequence motifs, domains and structural families tend to interact preferentially. Some methods predict interaction from sequence-motifs found to be over-represented in interacting protein pairs [19], or by considering the physico-chemical properties and the location of groups of amino acids in the sequence [20,21]. Others investigate the co-occurrence in interacting proteins of specific protein domains or their structural family classification [22,23]. When three-dimensional structures are available for both proteins thought to interact, high quality predictions and additional information such as the residues involved in the interaction and their binding affinity may be estimated (reviewed in [24]). Similarly, when two proteins show clear sequence similarity to proteins that exist in a complex for which the three-dimensional structure is known, detailed predictions of the atomic-level interactions may be made. For example, the major complexes in yeast have been predicted by this strategy [25].

2) Predictors based on comparative genomics have been exploited primarily in prokaryotes. They consider the physical location of genes, as well as their pattern of occurrence and evolutionary rate, to predict interactions or functional relationships between protein pairs. Some predictors make use of the observation that neighboring genes whose relative location is conserved across several prokaryotic organisms are likely to interact [26]. Other predictors exploit the observation that gene pairs that co-occur in related species or that co-evolve also tend to be more likely to interact [27-30]. In addition, domains that exist as separate proteins in some genomes but are also seen fused in a single protein in other genomes have been used to suggest the isolated domains may interact [31,32].

3) Predictors based on orthology work on the assumption that the orthologs of a protein pair that are known to interact in one organism will also interact. Such relationships are often referred to as interologs [33]. For example, at BLAST e-values below 10^{-10}, it has been shown that 16–30% of yeast interactions can be transferred to the worm [34] while further studies have estimated that a joint e-value below 10^{-70} is required to transfer interactions reliably between organisms [35]. Interologs have been used to predict protein-protein interactions in human [36].

4) Predictors based on functional features exploit non-sequence information to infer interactions. Some predictors exploit the observation that there is a significant correlation in the expression levels of transcripts encoding proteins that interact [37]. Since proteins must be co-localized in order to interact, protein subcellular localization has often been used to assess the quality of interaction datasets [38,39]. Similarly, interacting proteins are also often involved in similar cellular processes, so Gene Ontology "process" and "function" annotations have been exploited to predict interactions and validate high-throughput datasets [16,36,38].

5) Predictors have exploited similarities in the network topology of known interaction datasets to predict novel interactions. In one study, the local topology of small-world networks has been used to assess the quality of interaction datasets and predict novel interactions [40] while Gerstein and colleagues have investigated the prediction of interactions by the identification of missing edges in almost fully connected complexes [41].

In addition to these diverse approaches, some groups have combined concepts from several of the above categories in integrative frameworks. The first such predictor integrated co-expression data, co-essentiality as well as biological function in a naïve Bayes network to provide proteome-wide *de novo* prediction of yeast protein interactions [37]. Subsequently, the combination of many more diverse features was investigated using different frameworks to predict yeast protein-protein interactions, increasing the prediction accuracy and allowing an assessment of the limits of genomic integration [42-44]. The integration of diverse genomic features has also been useful in the investigation of the related but broader problem of predicting protein-protein associations as well as complex and pathway membership (see for example [45]).

Although, many computational methods have investigated the prediction of protein-protein interactions, few have so far been applied to the human proteome. The first large-scale prediction of the human interactome map involved transferring interactions from model organisms [36]. This resulted in over 70000 predicted physical interactions involving approximately 6200 human proteins. A second method integrated expression data, orthology,

protein domain data and functional annotations into a probabilistic framework and resulted in the prediction of nearly 40000 human protein interactions [46]. It has recently been estimated that the false-positive rates of these computational datasets as well as of available high-throughput human interaction datasets are, on average, as high as 90% and their coverage is only approximately 10%, indicating that more such efforts are needed to increase the coverage and confidence we have in current maps of the human interactome [18].

In this paper, the prediction of physical interactions between human proteins has been investigated by integrating in a Bayesian framework several different pieces of evidence including orthology, functional features and local network topology. In order to increase the accuracy and coverage of the predictions, different types of negative data (non-interacting protein pairs) were explored to train the predictor. The most accurate of the predictors was then used to assess the likelihood of pair-wise interaction for over 20000 human proteins from the IPI (International Protein Index) database. These predictions provide a likelihood of interaction for over 260 million human protein pairs and lead to the prediction of over 37000 human interactions. They should thus augment current knowledge of the human interactome as well as the understanding of the relationship between distinct cellular processes.

Results and discussion
Architecture of the predictor and training of the modules
The prediction of human protein-protein interactions was investigated in a Bayesian framework by considering combinations of individual protein features known to be indicative of interaction. The seven individual features considered are summarized in Table 1 and detailed in the Methods section. As indicated in Table 1, the different features were grouped into five distinct modules: Expression (E), Orthology (O), Combined (C), Disorder (D) and Transitive (T). Figure 1 illustrates the training scheme and architecture of the method. The Expression, Orthology, Combined and Disorder modules can calculate likelihood ratios (LR) of interaction independently and are referred to as the Group A modules (Figure 1A). The product of their likelihood ratios is referred to as the Preliminary Score. The Transitive module considers the local topology of the network predicted by the group A modules and thus requires the completion of their analysis to calculate its own likelihood ratios of interaction (Figure 1B). As such, all combinations of the Group A modules can be used to predict interaction in the presence or absence of the Transitive module. In the absence of the Transitive module, the Preliminary Score is used as the final likelihood ratio output by the predictor.

The likelihood ratios of interaction are evaluated for each module by considering the relative proportions of positive and negative training examples that have a specific state (i.e. that fall in a particular bin of a module). The datasets used to train the predictor consisted of 26896 known human protein interactions extracted from the Human Protein Reference Database (HPRD) [15] and approximately 100 times more randomly chosen protein pairs used as negative examples. The composition of the datasets and likelihood ratio calculations are explained in greater detail in the Methods section. Once the final likelihood ratio of interaction (LR_{final}) is calculated for a given protein pair as shown in Figure 1B, it is possible to estimate the posterior odds ratio of interaction by multiplying the final likelihood ratio by the prior odds ratio of interaction. Protein pairs that have a posterior odds of interaction above 1 are more likely to interact than not to interact, thus providing an obvious threshold to predict interacting proteins. Estimates for the prior odds ratio of interaction vary. Previous interaction studies on yeast and human use prior odds ratios that range from 1/600 to > 1/400 [37,43,46,47]. The evaluation of this ratio is difficult because not all true interactions are known. As detailed in Methods, the prior odds ratio for human protein interaction was explored by considering different versions and subsets of human interaction datasets. This suggested that there is insufficient data currently available to determine a reliable ratio for human. Accordingly, we selected a prior odds ratio of interaction of 1/400 which is similar to current estimates for yeast and is probably still quite conservative. Thus, the likelihood ratio threshold to predict interactions is 400.

Likelihood ratios of the modules
Figure 1 summarizes the likelihood ratios computed for the five modules. The different modules differ in the range of likelihood ratio values achieved by their different states. The Orthology and Combined modules both have states that achieve likelihood ratios above 400 (as high as 1207 for the Orthology module and 613 for the Combined module), indicating that both these modules can, on their own, predict some interacting protein pairs with a posterior odds ratio above 1.

The Expression module follows trends seen in previous studies with increasing likelihood ratios of interaction reflecting increasing expression correlation [37,46]. However, since the highest likelihood ratio for the expression datasets that we consider is 33, they are not sufficient on their own to predict interacting protein pairs with a posterior odds ratio above 1. Similarly, but in a much more pronounced way, the Disorder module is only slightly predictive of interaction, with a maximum likelihood ratio of 1.8.

Table 1: Features considered in the prediction of interactions for each module

Module abbreviation	Features considered	Data source	Description	Scoring function	Bins
E	Expression	GDS596 from the Gene Expression Omnibus [70]	Gene expression profiles from 79 physiologically normal tissues obtained from various sources [71]	Pearson correlation of co-expression over all conditions	20 of equal size covering the correlation value range (-1 to +1)
O	Orthology	InParanoid [72], BIND [66], DIP [65] and GRID [69] databases	Interactions of homologous protein pairs from yeast, fly, worm and human	Organism-based using InParanoid score	13
C (Combined)	Localization	PSLT predictions [54]	PSLT is a human subcellular localization predictor that considers nine different compartments (ER, Golgi, cytosol, nucleus, peroxisome, plasma membrane, lysosome, mitochondria and extracellular)	Qualitative score: proximity of compartments	4 (same, neighboring, different compartments, or not localized)
	Domain co-occurrence	InterPro [73] and Pfam [74]	Protein domains and motifs	Chi-square	5 covering range of Chi-square scores
	PTM co-occurrence	HPRD [15] and UniProt [76]	Post-translational modifications	$\dfrac{P(PTM[i], PTM[j]}{P(PTM[i]\mid I) * P(PTN}$	4 covering range of PTM scores
D	Disorder	VLS2 predictions [78]	Prediction of protein intrinsic disorder	Sum of the percent disorder for each protein in a pair	6 covering range of scoring function (0 to 200%)
T	Transitive	-	Module that considers local topology of underlying network predicted using combinations of above features	$T = \dfrac{\sum\limits_{e \in E_c} s_e}{1 + \mid E_i \setminus E_c \mid + \mid}$	5 covering range of scoring function

Most states of the Orthology module achieve higher likelihood ratios than the highest obtained by the Expression and Disorder modules. This is not surprising as the transfer of interacting orthologs (known as interologs [33]) from one organism to another is a popular method to predict interactions (see for example [34,48]), particularly in the case of organisms like human for which only a small proportion of interactions are known. The direct transfer of interactions to human from either yeast, fly or worm does not alone result in a posterior odds ratio above 1 (as the likelihood ratios of interaction for all yeast, fly and worm bins in the Orthology module are below 400). This is not surprising as previous studies have indicated that quite stringent joint E-values must be used to transfer interactions safely between organisms [34,35]. In contrast, the consideration of human interactions paralogous to the human protein pairs under investigation results in likelihood ratios of 431 and 1034 (depending on how close the paralogs are as described in Methods) which is much higher than those obtained for any single model organism. This agrees with a recent report that suggested protein-protein interactions are more conserved within species than across species [49].

The Combined module uses domain co-occurrence, post-translational modification (PTM) co-occurrence and subcellular localization information to predict interaction. These features were originally investigated separately, as shown in Figure 3, but their combination into one module that considers all dependencies between them achieves higher accuracy (data not shown) and higher likelihood ratios (as can be seen by comparing to Figure 1) while still being computationally feasible. Additionally, this combination circumvents possible problems of dependence between these features.

Previous methods have investigated the use of co-occurring domains to predict interaction (see for example [23,46]). Many pairs of domains co-occur in proteins known to interact. When investigated as a separate feature, the chi-square score of co-occurrence of domain pairs correlates well with the likelihood of interaction of protein pairs that contain these domains, with the highest chi-square score bin obtaining a likelihood ratio of 14, as shown in Figure 3A. Similarly, the co-occurrence of PTMs is also predictive of interaction, with its highest scoring bin obtaining a likelihood ratio of 6 as shown in Figure
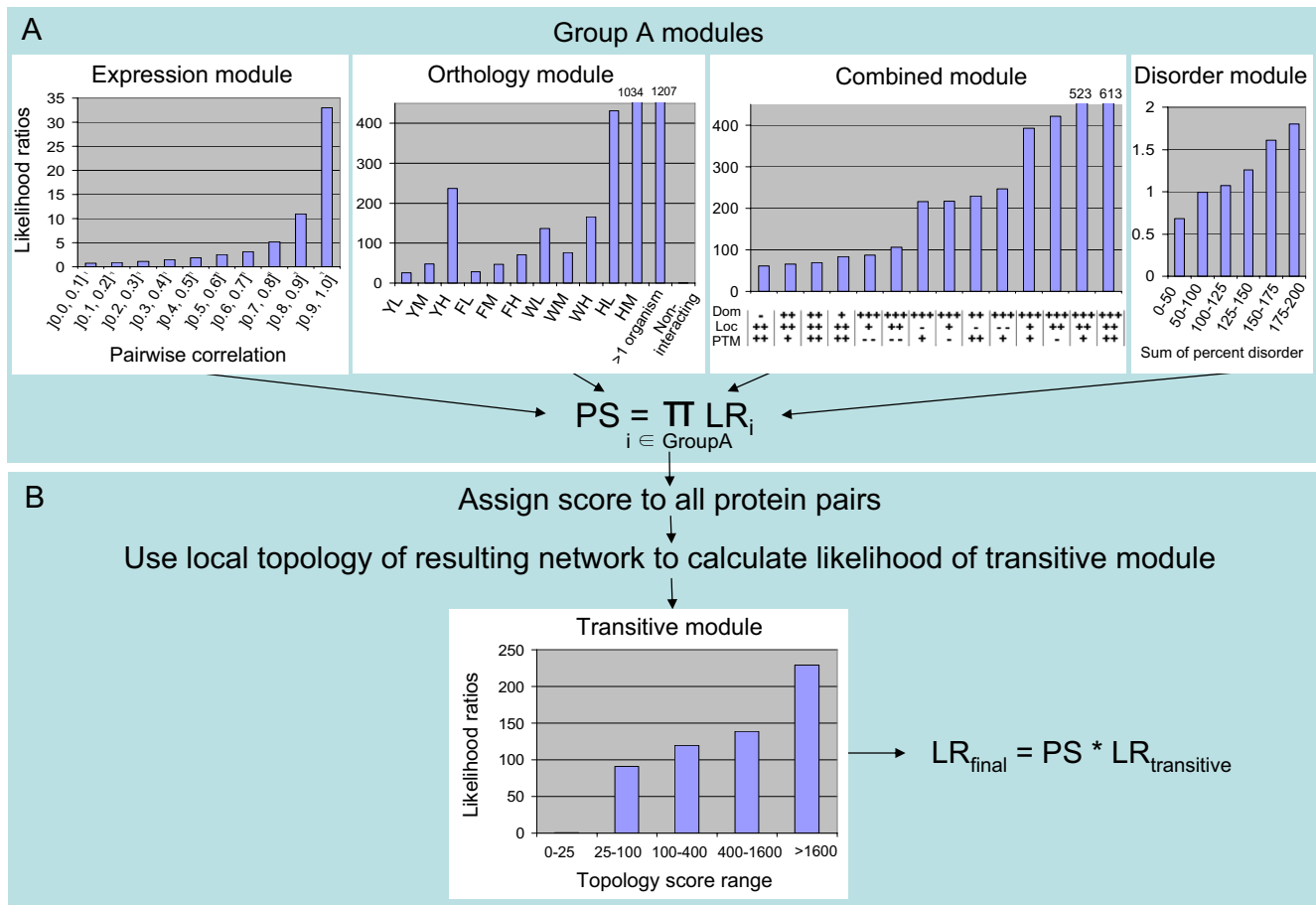
Figure 1
Architecture of the predictor and likelihoods of the modules. The predictor consists of two different parts (A and B) which are trained consecutively. The Group A modules (shown in panel A) are trained in parallel. The likelihood ratios (LR) for most of their states are shown in panel A (their complete likelihood ratios are available in Additional File 4). The product of the likelihood ratios of all Group A modules considered in a given prediction is referred to as the preliminary score (PS) and can be calculated for all human protein pairs. If the Transitive module is not considered, the final likelihood ratios assigned to all protein pairs is the preliminary score (PS). If the Transitive module is considered, the local topology of the network determined by the assignment of preliminary scores to all protein pairs considered in the training set is used to calculate the likelihood ratios for the transitive module (shown in panel B) for every protein pair in the training set. The final likelihood ratio is then the product of the preliminary score calculated in panel A and the likelihood ratio output by the transitive module in panel B. For the Orthology module: YL, YM, YH: yeast low, medium and high scoring bins; FL, FM, FH: fly low, medium and high scoring bins; WL, WM, WH: worm low, medium and high scoring bins; HM and HL: medium and low scoring bins for human protein pairs that have human paralogs; > 1 organism: bin for human protein pairs that have interologs in more than one organism. For the Combined module, —— refers to the lowest scoring bin (for the domain (Dom), post-translational modification (PTM) and subcellular localization (Loc) features), – refers to the second lowest scoring bin and +, ++, +++ refer respectively to the third highest, second highest and highest scoring bins.

3B. Lists of high scoring domain pairs and PTM pairs are shown in Additional Files 1 and 2.

Subcellular localization has been extensively used both to assess the quality of interaction datasets [11,50,51] and to generate examples of non-interacting protein pairs to use as negative datasets when training and testing predictors

[37,46]. In the present study, the use of localization was investigated as a feature predictive of interaction. Four possible localization states were considered for protein pairs: same compartment, neighboring compartments, different non-neighboring compartments and absence of localization annotation (more details are given in the Methods section). As shown in Figure 3C, the likelihood
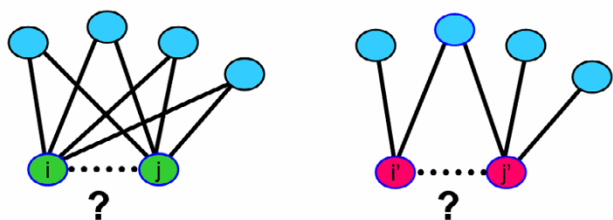
Figure 2
Transitive module hypothesis. The Transitive module investigates whether two proteins (such as i and j) that share many common interactors and have few additional interactors that are not common to both proteins are more likely to interact than two proteins (such as i' and j') that share few common interactors.

A

Domain co-occurrence

B

PTM co-occurrence

C

Subcellular co-localization

Figure 3
Likelihood ratios of the features that form the Combined module, considered separately. The Combined module considers simultaneously three distinct features: the co-occurrence of both domains and PTMs as well as the subcellular co-localization of proteins. Here the likelihood ratios of these features considered separately are shown. In panel A, all domain pairs considered were given scores and likelihood ratios were estimated for different values of these scores. Similarly, shown in panel B are the likelihood ratios for different values of PTM co-occurrence scores. Panel C shows the likelihood ratios for protein pairs localized to different sets of cellular compartments.

ratio of same compartment protein pairs was found to be twice as high as that of randomly chosen or non-annotated protein pairs whereas different non-neighboring protein pairs are more than three times less likely to interact than random protein pairs Individual localization features achieve low interaction likelihood ratios. However, when integrated into the Combined module, domain, PTM and localization information together achieve likelihood ratios that are high enough to predict interaction on their own (i.e. above 400). As expected, the highest likelihood ratio bins for the Combined module are those representing the highest combinations of the three features separately.

The transitive module enhances the preliminary likelihood score (PS) (calculated using the group A modules) by considering the local topology of the resulting network which is assessed using the neighborhood topology score as detailed in the Methods section. The likelihood ratios for different values of the neighborhood topology score are shown in Figure 1B. The Transitive module is highly predictive of interaction and achieves likelihood ratios as high as 229. This module cannot be used alone as it requires as input the output of at least one group A module. However, it can predict interacting protein pairs with a posterior odds ratio above 1.0 when used in combination with any single module in group A (as the product of the highest likelihood ratios of the transitive module and any group A module is greater than 400 as can be seen from Figure 1).

Independence of the modules
The final likelihood ratio output by the predictor is only representative of the true likelihood of interaction of a protein pair if the modules considered are independent. If the modules were not independent, some likelihood ratios would likely be overestimated, particularly for protein pairs that achieve simultaneously high likelihoods for
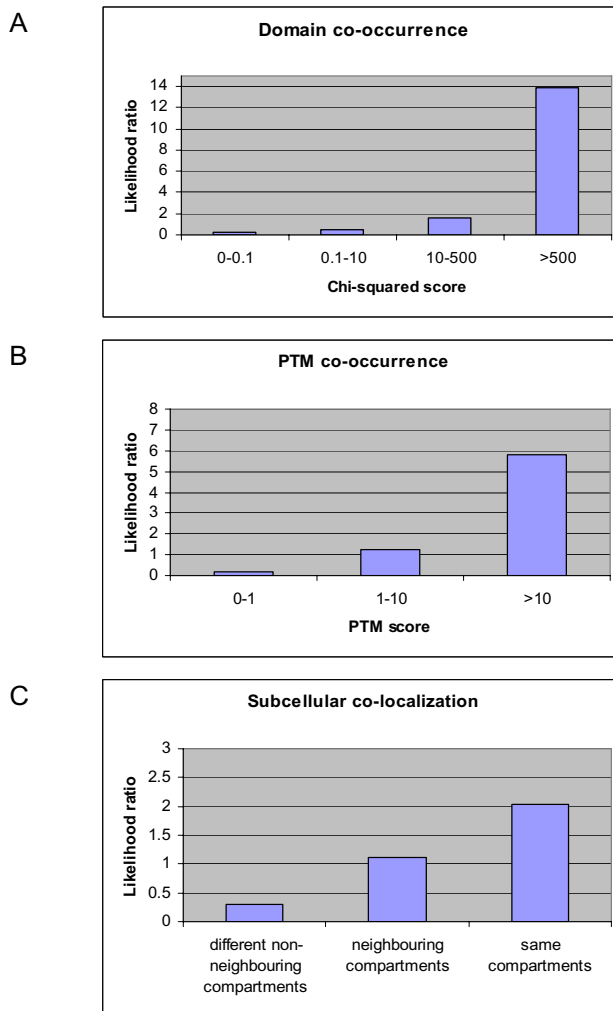
non-independent features. Conversely, some likelihood ratios would be underestimated for protein pairs achieving simultaneously low likelihoods for non-independent features. Previous studies have demonstrated that some of the features considered here are indeed independent [43]. Independence of all modules used in our predictor was verified by calculating Pearson correlation coefficients for

all pairs of modules. As shown in Table 2, all modules considered are independent, since the highest Pearson correlation coefficients computed are well below any value considered significant.

Accuracy of the predictors

All combinations of modules were examined to determine which of the resulting predictors achieved the highest prediction accuracy. In order to analyze the predictions, five-fold cross validation experiments were performed and the area under partial ROC (receiver operator characteristic) curves (partial AUCs) measured. ROC50 and ROC100 curves were selected as they consider a large enough number of positives to include all protein pairs predicted to have a posterior odds ratio above 1.0 by all the predictors investigated. Protein pairs predicted to have a posterior odds ratio below 1.0 have an estimated true positive rate below 50% and thus are more likely not to interact than to interact. These protein pairs are therefore not of interest in this context. The area under all ROCn curves considered is relatively low because of the high proportion of negatives with respect to positives in the training and test sets (100:1).

Table 3 summarizes the characteristics of 19 different predictors and shows accuracy measures. Individual modules do not achieve high scores for the areas under the ROC50 and ROC100. In fact, all ROC50 AUC values achieved by individual modules are below 0.025 and the Expression and Disorder modules do not predict any protein pairs (positive or negative) above a posterior odds ratio of 1, which is expected as the highest likelihood ratios they achieve are lower than 400 (see Figure 1A). As more Group A modules are considered within the same predictor, the ROCn AUC scores increase significantly, as would be expected since these features are independent (as shown in Table 2) and thus contribute different information to the prediction. For example, the predictor that considers both the Expression and Combined modules achieves a ROC50 AUC of 0.033 compared to 0.003 and 0.022 respectively for the individual modules. However, the Disorder module does not contribute significantly to the prediction as predictors that consider it do not, in general, do better than their counterparts that do not use it. For example, both the Expression-Orthology predictor

and the Expression-Orthology-Disorder predictor achieve a ROC50 AUC of 0.024. The Disorder module offers the advantage of increasing the coverage of the prediction as a disorder score is calculated for all protein pairs. However, this appears to add more noise to the prediction without increasing the accuracy.

As the scores of the predictors increase, so do the number of interactions predicted above different posterior odds ratio thresholds (see lower portion of Table 3). For example, the Expression-Orthology predictor achieves a ROC50 AUC of 0.024 and predicts 5670 interactions at a posterior odds ratio greater than 1 whereas the Expression-Orthology-Combined predictor achieves a ROC50 AUC of 0.044 and predicts over 15000 interactions at a posterior odds ratio above 1. The best combination of Group A modules is the predictor consisting of the Expression, Orthology and Combined modules.

The Transitive module, which can only be used in combination with other modules, increases substantially the scores and number of interactions predicted. The right-hand portion of Table 3 shows the accuracy measures for the highest scoring subset of predictors that consider the Transitive module. The Transitive module enhances the prediction by identifying among protein pairs with a relatively high preliminary score those that are most likely to interact, by considering the local topology of the network around them. For example, the ROC50 AUC rises from 0.044 to 0.075 when the Transitive module is added to the Expression-Orthology-Combined predictor, and the number of predictions above a posterior odds ratio of 1 doubles from 15330 to 34780. Once again, the Disorder module does not contribute positively to the prediction. Its inclusion does not increase any of the measures of accuracy considered. The predictor that considers the Expression, Orthology, Combined and Transitive modules is the one that achieves the highest accuracy overall. It is this predictor that is further analyzed in the next sections.

Comparison to predictions generated using alternative training sets

In this work training sets were used that comprised 100 times more negatives than positives, with the negatives

Table 2: Pairwise Pearson correlation for all modules

	Expression	Orthology	Combined	Disorder	Transitive
Expression	-	0.00460	0.01299	0.00995	0.00562
Orthology	-	-	0.01000	0.00977	0.01555
Combined	-	-	-	0.02086	0.02380
Disorder	-	-	-	-	-0.01702
Transitive	-	-	-	-	-

randomly selected and filtered to remove any known or suspected positives (see Methods). Other groups have used negative:positive ratios ranging from 1 to more than 600 (see for example [37,47,52]). In addition, several groups use localization-derived negatives (i.e. protein pairs that are not annotated as being localized to the same cellular compartment) rather than randomly chosen negatives (see for example [37,43,46]). These issues have been investigated previously [53].

Since the choice of negative training data may influence the method, the choice of different training sets in the context of the probabilistic predictor presented here was investigated to determine which type of training set offers the highest accuracy.

Table 4 compares the accuracy of predictors trained with negative:positive ratios of 1:100 and 1:1 and tested by five-fold cross validation. Ratios greater than 100 were not considered because they are computationally infeasible given the size of our datasets and the architecture of the predictor. To perform such a comparison, the EOCT predictor (Expression, Orthology, Combined and Transitive modules) was trained on datasets consisting of either equal numbers of positives and negatives or 100 times more negatives than positives and then tested on both types of datasets. As shown in Table 4, the predictors trained on datasets containing 100 times more negatives than positives perform significantly better than those trained on datasets containing equal numbers of positives and negatives. For example, the 1:1 pos:neg trained predictor achieves a ROC50 AUC of 0.0645 whereas its 1:100 pos:neg trained counterpart achieves a 0.0747 ROC50 AUC. This could be due to the fact that the number of non-interacting protein pairs outweighs greatly the number of interacting protein pairs in cells. When equal numbers of positives and negatives are used in training, the diversity that exists in the non-interacting protein pair space may not be captured, thus resulting in misleading likelihood ratios for the predictive modules. It should be noted that predictors tested on datasets consisting of equal numbers of positives and negatives achieve much higher accuracy measures than those tested on datasets containing 100 times more negatives than positives. This is because the number of positives scoring higher than the highest scoring n negatives, for a given value of n and a given predictor, will be greater if there are equal numbers of positives and negatives in the test set than if there are more negatives than positives.

The effect of localization-derived negatives rather than randomly chosen negatives was also investigated to see if it would increase the prediction accuracy. A criticism of randomly chosen negatives is that they will contain some true interactors. However, the set of interacting pairs in the full protein pair space is small and thus the contamination rate of randomly chosen negative datasets will in fact be very low. Contamination is probably below 1%, which is likely lower than the contamination rate of the positive dataset as discussed in [47]. Localization-derived negatives, on the other hand, should be free of contamination, if the localization annotations are complete and accurate, both conditions that are difficult to obtain as discussed in [54]. However, one can argue that localization-derived negatives might not be able to capture the full diversity of the non-interacting protein space since many proteins in the same cellular compartment do not interact. In addition, proteins specific to a cellular compartment may have different characteristics to proteins in other compartments. Such predictors may not generalize well when predicting on cell-wide protein pairs which consist not only of non-colocalized non-interacting pairs but also numerous protein pairs that do not interact but are present in the same cellular compartment. These issues have been discussed previously [52]. In order to see if different types of negatives could influence the accuracy of the predictors developed here we generated negative training/test sets as in [46] by identifying all pairs of human proteins for which one protein is annotated as being nuclear and the other is annotated as being localized to the plasma membrane in the HPRD database [15]. The Combined module for these predictors only considers domains and PTMs but not subcellular localization as this would result in using this feature both in the selection of the training set and as a feature predictive of interaction. The localization-derived negative trained predictor tested on sets containing localization-derived negatives achieves a lower accuracy than that of the random negative trained predictor tested on a test set containing randomly-generated negatives (0.0686 +/- 0.0010 vs 0.0747 +/- 0.0022). This is most likely due to the fact that the localization-derived negative trained predictor cannot take full advantage of the Transitive module, since the network resulting from the predictions of the Group A modules likely does not sample the whole protein pair space well.

Our predictor trained with randomly generated negatives and a negative:positive ratio of 100 performs the best out of all the combinations of training sets investigated. It is this predictor that is further analyzed in subsequent sections.

Contribution of the modules
The relative contribution of the modules to the prediction of interaction was investigated in order to gain a better understanding of the predictive power and areas of highest usefulness of the different modules. To do this, all protein pairs were considered that achieve an estimated posterior odds ratio > 1 when the EOCT predictor was trained on the full datasets without cross-validation. This

Table 3: Prediction accuracy of different combinations of modules

Modules included in prediction																			
Expression	•				•	•	•				•	•	•		•		•		•
Ortho		•			•			•	•		•	•		•	•	•	•	•	•
Combined			•			•		•		•	•		•	•	•	•	•	•	•
Disorder				•			•		•	•		•	•	•	•			•	•
Transitive																•	•	•	•
Coverage of the Informative Protein Set (%)																			
	22	23	32	99	40	43	99	49	99	99	57	99	99	99	99	90	95	99	99
Measures of accuracy																			
ROC50 AUC	0.003	0.018	0.022	0	0.024	0.033	0.003	0.042	0.021	0.030	0.044	0.031	0.024	0.038	0.042	0.071	**0.075**	0.069	0.072
ROC100 AUC	0.003	0.026	0.032	0	0.030	0.045	0.005	0.054	0.027	0.041	0.058	0.044	0.029	0.049	0.058	0.090	**0.094**	0.088	0.093
Estimation of number of interactions predicted																			
posterior odds ratio > 4	0	420	0	0	1050	630	0	2520	0	0	3780	630	0	2888	2520	14200	16590	13400	**16800**
posterior odds ratio > 2	0	630	1050	0	2520	2100	0	3780	210	2100	7980	2100	1050	4200	5460	21340	**24570**	21200	24200
posterior odds ratio > 1	0	840	4830	0	5670	7140	0	11760	1890	4200	15330	5460	3990	13125	13860	28500	**34780**	28600	33180

set consists of 37606 distinct predicted interactions and is referred to as the LR400 dataset (all these interactions are listed and ranked in Additional File 3). These protein pairs represent the most probable interactors with respect to the features considered, among all protein pairs examined by the predictor.

To investigate the individual contribution of each module, we looked at the number of interactions predicted out of all LR400 pairs as a function of the minimum likelihood ratio of each module. As shown in Figure 4A, all modules contribute positively (i.e. contribute a likelihood ratio greater than 1.0) to the prediction of a certain proportion of the interactions in the LR400 dataset. The Transitive module and to an even greater extent, the Combined module contribute positively to the prediction of a very high proportion of the LR400 protein pairs (73% and 91% of the LR400 interactions have likelihood ratios greater than 1 for the Transitive and Combined modules respectively). The Transitive module provides a likelihood ratio of 91 for the prediction of over 70% of the LR400 interactions. The Combined module provides positive evidence for the highest number of interactions of the LR400 dataset. However, the value of the likelihood ratio it contributes is below 20 for over 50% of protein pairs in the LR400 dataset (which means that for these protein pairs, the Combined module must be used in combination with other modules to achieve a total likelihood ratio above 400). The Combined module does, however, achieve likelihood ratios high enough to predict over two thousand interactions of the LR400 dataset on its own,

less than 15% of which are present in the training set. The Orthology module contributes to the prediction of only 8474 protein pairs in the LR400 dataset (23%). However, a large majority (> 75%) of these 8474 predicted interactions achieve likelihood ratios above 200 from this module. In fact, almost 40% of these LR400 interactions achieve a likelihood ratio above 400 from the Orthology module. This indicates that most interactions predicted by the Orthology module (alone or in combination with other modules) are based on the highest scoring Orthology bins (see Figure 1A) which are the most conserved yeast interactions (whose bin achieves a likelihood ratio of 237), as well as human paralogous interactions and interactions found in more than one model organism (both of which achieve a likelihood well above 400). Few interactions in the LR400 dataset are predicted on the basis of having interacting orthologs in worm or fly alone. The Expression module provides positive evidence for a little less than half the predictions in the LR400 dataset. However, as previously noted, the highest likelihood provided by this module is 33 and thus the Expression module cannot predict interaction on its own.

Figure 4B summarizes the contributions of different combinations of modules. The Combined and Transitive modules contribute the most to the prediction of interactions. They alone can predict approximately 27000 of the 37606 interactions of the LR400 dataset. When they are both present, regardless of which other modules are also present, they predict over 70% of the LR400 interactions. When either of these two modules is absent, fewer than

Table 4: Influence of the negative:positive training set ratio on the prediction accuracy

		Neg:pos testing ratio			
		1:1	1:1	100:1	100:1
		ROC50 AUC (std)[a]	ROC100 AUC (std)[a]	ROC50 AUC (std)[a]	ROC100 AUC (std)[a]
Neg:pos training ratio	1:1	0.300 (0.008)	0.385 (0.006)	0.0645 (0.0019)	0.0814 (0.0009)
	100:1	0.325 (0.004)	0.403 (0.003)	0.0747 (0.0022)	0.0944 (0.0028)

[a] The ROCn AUCs are an average of five separate experiments (each of which is itself a five-fold cross validation experiment). Their standard deviation is shown in parenthesis.

12500 interactions are predicted. In contrast, the two remaining modules (Expression and Orthology) can predict approximately 5000 interactions together. This is interesting as many of the publicly available predicted interaction datasets mentioned in the Background section use mainly orthology transfer from model organisms to identify interactions. As the majority of the LR400 interactions are derived from the Combined and Transitive modules, it is possible that the method is identifying a large subset of interactions that are not common to previous human protein interaction datasets. This is discussed further in the next section. The curve representing the full predictor (consisting of the Expression, Orthology, Combined and Transitive modules) is also represented in Figure 4B (the dark green squares). By definition, it predicts all proteins in the LR400 dataset at likelihood ratios equal to or above 400 (this is how the LR400 dataset was generated). The right side of the curve illustrates the number of interactions that are predicted above likelihood ratios of 400 and more. As shown in Figure 4B, the full predictor predicts approximately 20000 interactions at a total likelihood ratio of 1600 (which is equivalent to an estimated posterior odds ratio of 4). At a likelihood ratio of 4000, approximately 11000 interactions are predicted and at a likelihood ratio of 8000, approximately 6500 interactions are predicted. We verified that the increasing estimated posterior odds ratios translated into better predictive value. Figure 5 shows the true positive rate versus false positive rate for different posterior odds ratios as measured by five-fold cross validation. As the posterior odds ratio increases, the false positive rate decreases and the relative proportion of true positives increases when compared to the proportion of false positives. Accordingly, subsets of very high quality predictions may be generated by choosing a suitably high posterior odds ratio threshold.

Comparison to other interaction datasets

The false positive rate (FPR) of our predictor was estimated by the method of D'Haeseleer and Church [18,55] and used to compare it to other prediction datasets. The Ramani interaction dataset that was automatically extracted from the literature [16] as well as all new interactions present in the October 2006 version of the manually curated HPRD database [15] (but none of the interactions also present in earlier versions of the HPRD which were used to train our predictor) were taken as reference datasets. The D'Haeseleer and Church method compares two experimental datasets to a reference set and assumes that all intersections between the three datasets contain true positives. It is thus possible to estimate the number of true positives predicted by an experimental dataset by comparing the number of interactions present in the different intersections of the two experimental methods and the reference dataset (for details, see [18,55]). Here, we compare three human interaction prediction datasets: the Rhodes probabilistic dataset [46], the Lehner orthology-derived dataset [36] and the most accurate of our predictors (the LR400 subset of the predictor considering the Expression, Orthology, Combined and Transitive modules). We estimated false positive rates for each of the datasets by comparing them two by two to one of the reference datasets, thus generating 4 to 6 different estimates of false positive rates for each computational dataset, as shown in Figure 6A (the two Lehner datasets were not compared to each other, which is why they have fewer FPR estimates). The rates estimated for the Rhodes and Lehner datasets are similar to previous estimates [18]. The estimated false positive rates for the LR400, Rhodes and core Lehner are quite similar (an average of 76% FPR for both the LR400 and core Lehner datasets and 78% for the Rhodes dataset) and well below the overall average false-positive rate of 90% estimated for most available human high-throughput experimental and prediction interaction datasets [18]. It should be noted that the Rhodes, Lehner and Ramani datasets annotate interactions as a relationship between human genes and not their protein products directly. However, not all proteins encoded by a single gene will necessarily interact with all protein products encoded by a second gene, even if one such protein pair does. This is why we describe interactions as a relationship between two proteins, allowing for a more precise description of the interaction. To compare our predictions to these datasets, we consider that two genes interact if at least one of their respective protein products interact.
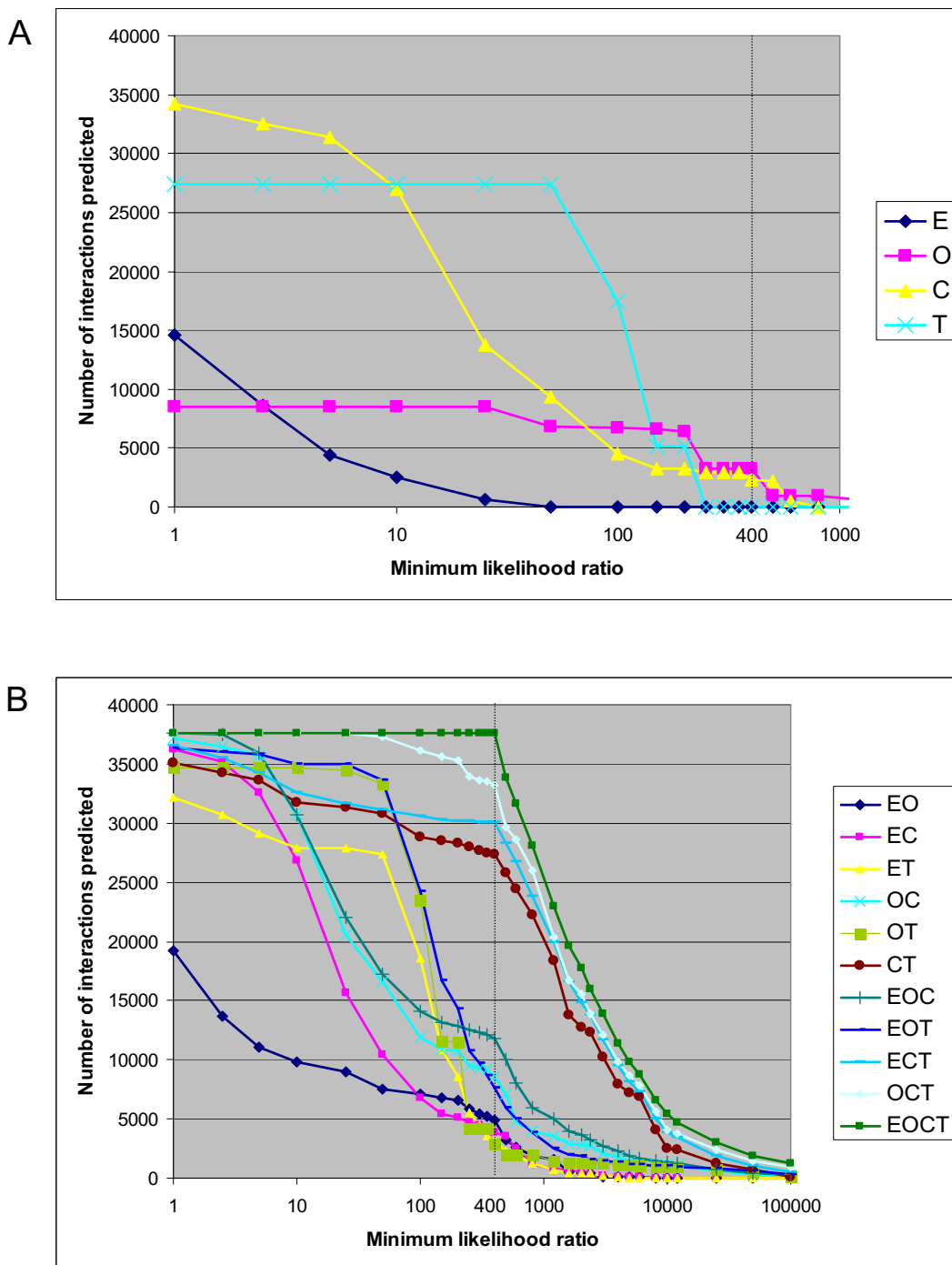
Figure 4
Contribution of the modules. To examine the contribution of the different modules, we plotted the number of interactions predicted among all LR400 interactions (all interactions predicted using the full predictor that obtain a likelihood ratio of interaction greater than 400) as a function of the minimum likelihood ratio of individual modules (in panel A) or of combinations of modules (in panel B). In the case of combinations of modules (panel B), the minimum likelihood ratio is the product of the likelihood ratios of the modules considered. Thus for example, the product of the expression and orthology ratios is greater than 1 for almost 20000 LR400 interactions and greater than 10 for approximately 10000 LR400 interactions (dark blue diamonds in panel B). E: Expression module, O: Orthology module, C: Combined module, T: Transitive module.
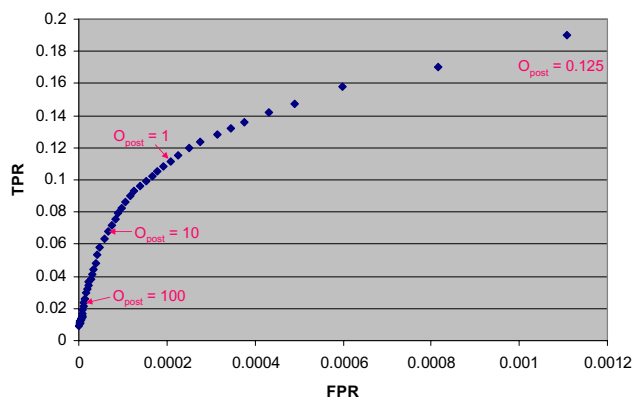
Figure 5
True positive rate versus false positive rate for different estimated posterior odds ratios. The true positive rate (TPR) versus false positive rate (FPR) is plotted for different values of the posterior odds ratio estimated for the dataset by five-fold cross-validation. As the posterior odds ratio increases, the false positive rate decreases and the ratio of the true positive rate divided by the false positive ratio increases. Thus, higher quality datasets can be generated by requiring higher posterior odds ratios. The TPR is calculated as the number of true positives predicted divided by the total number of positives in the test set. The FPR is calculated as the number of false positives predicted divided by the total number of negatives in the test set.

In Figure 6B and 6C, we compare the number of distinct proteins and distinct interactions of the LR400 dataset to those of the Rhodes prediction dataset and the June 2006 version of the HPRD which was used to train our predictor. The Rhodes dataset was trained using an earlier version of the HPRD. As can be seen in Figure 6, the intersections between the three datasets considered are low, especially when comparing the interactions. Both the Rhodes dataset and our LR400 dataset predict interactions involving many proteins that are not even present in their positive training set (the HPRD). Many of the predictions in these two datasets concern protein pairs and proteins that are not present in other datasets, suggesting that they cover different regions of the human interaction space. As suggested in [18], by making more such datasets available, it will be possible to increase our coverage of the interaction space and determine the most likely human interactions.

Another human interaction dataset has recently become available: the IntNetDB [56]. It was generated by integrating seven different features (four of which involve transferring interactions or characteristics of protein pairs from model organisms to human) in a probabilistic framework. Interactions were predicted above a TP/FP ratio (number of true positives divided by the number of false

positives in the test set) of 1. Using such a threshold, the authors claim to predict 180 010 human interactions. We do not compare our predictions to this dataset because such a threshold of TP/FP > 1 does not correspond to a posterior odds threshold > 1. Depending on the positive-to-negative ratio used in the datasets, TP/FP > 1 might correspond to an average posterior odds ratio of 1. In contrast, the average posterior odds ratio of our LR400 dataset is above 700. In comparison, by using a threshold of TP/FP > 1 in our test set, we predict over 1 000 000 human interactions. We do not believe that the quality of this large number of predictions is high enough to warrant their publication since the great majority of these protein pairs achieve a posterior odds ratio below 1.

Independent validation

Although the overlap between the LR400 dataset and the HPRD-derived positive training set is below 10% as shown in Figure 6C, the proportion of interactions common to these two sets is not equally distributed for all posterior odds ratios of interaction values. As shown in Figure 7, while less than 3% of the protein pairs predicted to interact at posterior odds ratios between 1 and 2 overlap with the HPRD dataset used for training, this value increases to over 50% for the highest scoring subsets of the LR400 dataset. These highest scoring predictions receive high likelihood ratios of interaction from all four predictive modules and represent the strongest examples of interaction as evaluated by our predictor. Such examples include interactions that allow the formation of well-known protein complexes such as the proteasome, the MCM protein complex involved in the initiation of genome replication, replication factor C, the TBP/TAF complex (TBP-associated factors) and the EIF complex (eukaryotic translation initiation factors). The highest scoring predictions in the LR400 dataset thus mainly represent interactions present in the HPRD dataset as well as interactions between proteins that have strong sequence identity to these known interacting pairs. However, as the posterior odds ratio decreases, the overlap between the predictions and the HPRD-derived training set decreases. Some subsets of quite high posterior odds have much smaller overlaps with the training set. For example, interactions predicted at posterior odds ratios between 128 and 2048 have a 20 to 30% overlap with the training set as shown in Figure 7. Although many of these novel predictions have not been previously investigated in the literature, there exists experimental evidence supporting a subset of these predictions which is not present in the June 2006 version of the HPRD used to train our predictor, thus providing independent validation of our method. Five such validated predictions are reported here:

-TCPTP was predicted to interact with STAT6 at a posterior odds ratio of 4300. It has been recently reported that
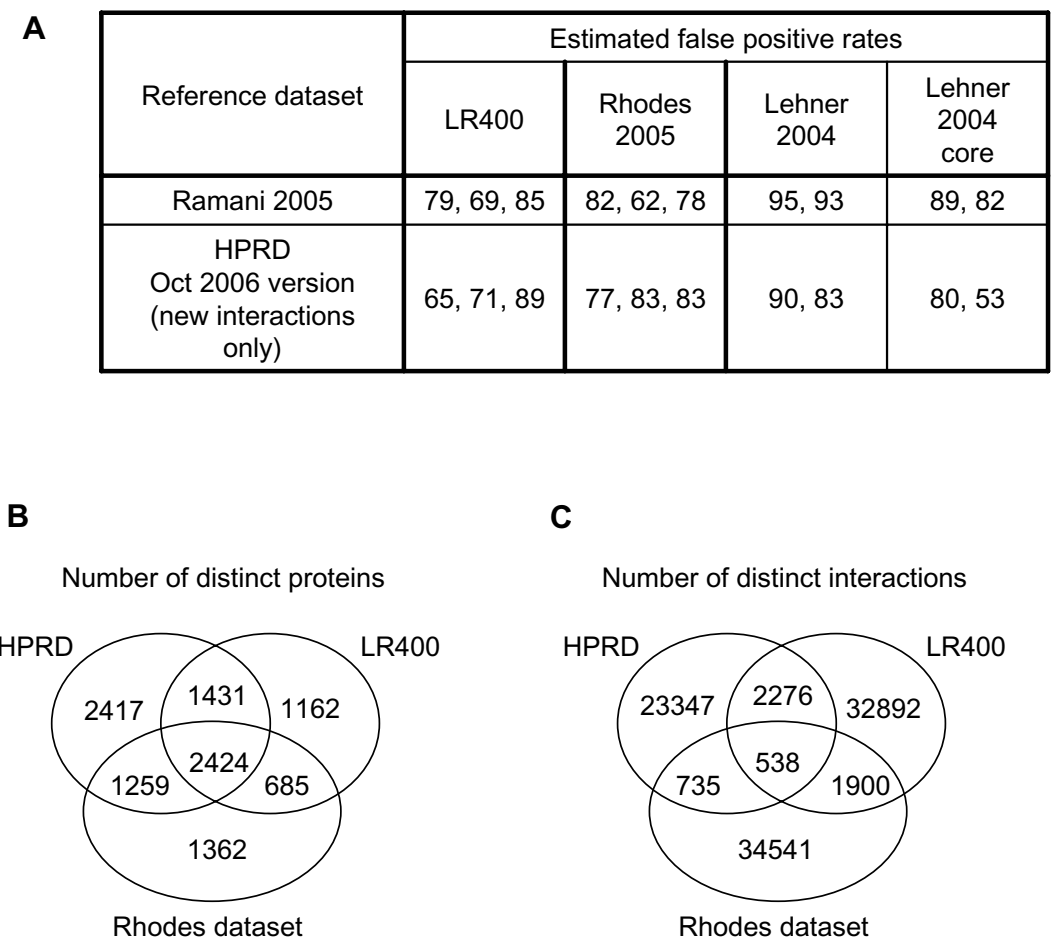
A

Reference dataset	Estimated false positive rates			
	LR400	Rhodes 2005	Lehner 2004	Lehner 2004 core
Ramani 2005	79, 69, 85	82, 62, 78	95, 93	89, 82
HPRD Oct 2006 version (new interactions only)	65, 71, 89	77, 83, 83	90, 83	80, 53

B

Number of distinct proteins

HPRD   LR400

2417   1431   1162
1259   2424   685
1362

Rhodes dataset

C

Number of distinct interactions

HPRD   LR400

23347   2276   32892
735   538   1900
34541

Rhodes dataset

Figure 6
Comparison to other interaction datasets. The false positive rates shown in panel A were estimated for the LR400 dataset as well as the Rhodes [46] and Lehner [36] predictions using the method described in [18, 55] by comparing them two-by-two to a reference dataset. The number and overlap of distinct proteins (shown in B) and distinct interactions (shown in C) are shown for the LR400 dataset, the Rhodes prediction dataset and the June 2006 version of the HPRD.

TCPTP, the only protein tyrosine phosphatase known to localize to the nucleus, dephosphorylates STAT6 in this cellular compartment, which may in turn lead to the suppression of Interleukine-4 (IL-4) induced signaling [57].

-N-WASP and ARP3 achieve a predicted posterior odds ratio of interaction of 2700. A recent report suggested that the IQGAP1 protein can activate N-WASP thus changing its conformation and allowing it to bind the ARP2/3 complex, which in turn directs the generation of branched actin filaments required for the extension of a lamellipodium [58].

-The VAMP3-VTI1A interaction was predicted with a posterior odds ratio of 1518. Both these proteins are believed to be part of the SNARE (soluble *N*-ethylmaleimide-sensitive factor attachment protein receptor) family of proteins which are involved in membrane fusion events. VTI1A is a trans-Golgi-network-localized putative t-SNARE [59] and VAMP3 is an early/recycling endosomal v-SNARE [60]. These two proteins were recently shown to interact, leading to their functional implication in the post-Golgi retrograde transport step [61].

-CDK2 and MCM4 were predicted to interact at a posterior odds ratio of 62. CDK2 has recently been shown to phosphorylate MCM4, a subunit of a putative replicative helicase essential for DNA replication, on two distinct residues, leading to a change in its affinity to chromatin and its enrichment in the nucleolus [62].
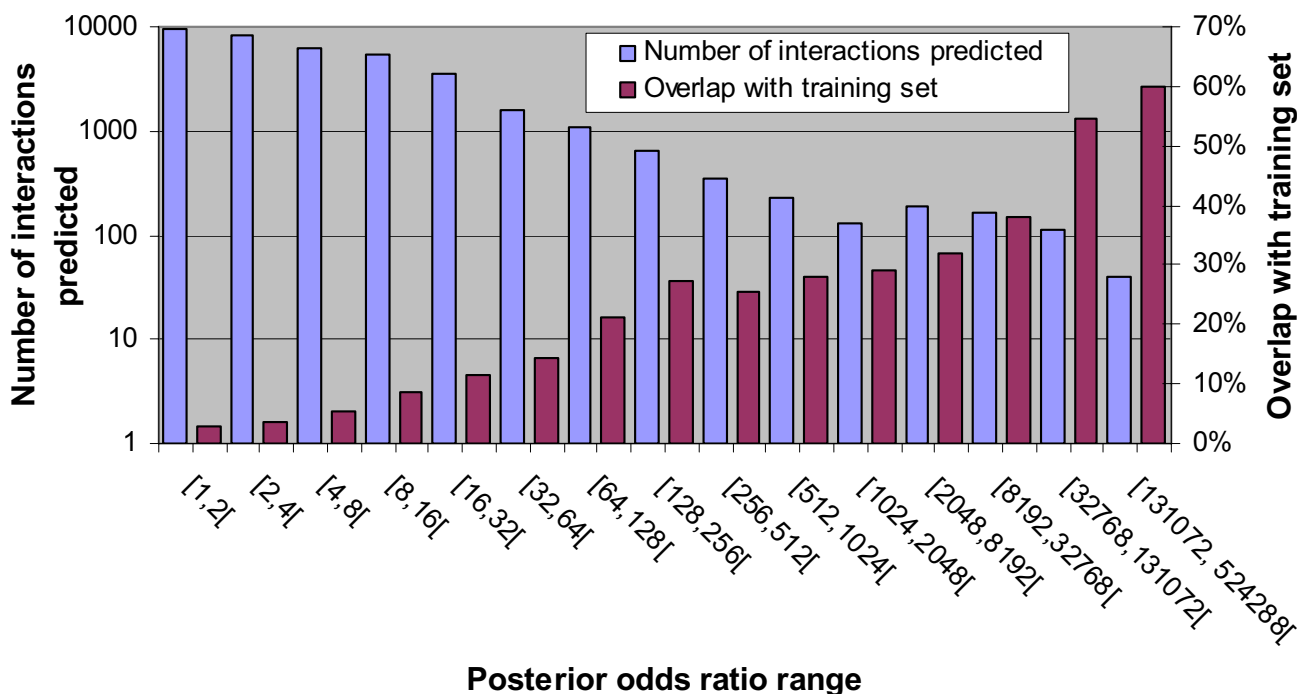
Figure 7
Overlap of different subsets of the LR400 dataset with the HPRD-derived training set. The number of interactions predicted and the proportion of overlap with the training set (which was derived from the HPRD) were calculated for subsets of the LR400 dataset of different posterior odds ratios.

-Sam68 and Smad2 achieve a predicted posterior odds ratio of 32. This interaction has been experimentally demonstrated by large-scale yeast-two-hybrid analysis of the Smad signaling system [63].

Our probabilistic predictor therefore not only reproduces and completes well-known protein complexes but also identifies novel interactions, a subset of which have been independently validated.

Conclusion
The current human protein interaction map is estimated to be only 10% complete [18]. Here, we investigated the prediction of human protein-protein interactions in an effort to increase the coverage of the human interactome while simultaneously providing high quality predictions. By considering several different types of orthogonal and quite distinct features including expression, orthology, combined protein characteristics and local network topology, we predicted over 37000 human protein interactions and explored a subspace of the human interactome that has not been investigated by previous large interaction datasets. Our investigation led us to compare the influence of different training sets on the prediction accuracy. The use of randomly generated negative training examples

and large negative-to-positive ratios in the training set generated the most accurate predictors in the context of our model. A comparison to other large human interaction datasets revealed the average false positive rate of our dataset to be 76%, which is much lower than the overall average for most large scale, currently available, human interaction datasets (experimental and computational) estimated to be 90% [18]. A subset of our novel predictions have been independently validated by identifying recent reports that experimentally investigated and confirmed that these protein pairs do interact. We provide all our predictions ranked according to the posterior odds ratio of interaction in Additional File 3. It is thus possible to restrict the dataset to the highest scoring protein pairs (and only choose for example, protein pairs that have an estimated true positive rate of interaction above 90%). By making this human interaction prediction dataset publicly available, it is our hope that it will help to identify the most high-confidence interactions, leading to a more complete and accurate human interaction map.

Methods
Datasets
In order to investigate the likelihood of interaction of human proteins, 62322 human protein sequences were

downloaded from the International Protein Index (IPI) database (version 3.16) [64]. Some of these proteins are alternative transcripts of the same gene but can have distinct interaction partners. Known interactions were downloaded from the Human Protein Reference Database (HPRD; June 2006 version) [15]. Duplicate interactions and self-interactions were not considered. Additionally, some proteins were not recovered in the conversion between different identifiers. This resulted in 26896 distinct human protein interactions involving 7531 distinct human proteins present in the initial IPI dataset. The 26896 interactions from the June 2006 version of the HPRD were used as the positive dataset in the training/testing of the predictor. Two different sets of non-interacting protein pairs were investigated: the main analysis employed a randomly-generated negative dataset but this was also compared to a localization-derived negative dataset. Both non-interacting protein datasets were cleaned by removing all protein pairs that came from the positive dataset as well as protein pairs that were annotated as interacting in other databases (DIP [65]: 679 interactions, BIND [66]: 2650 interactions), or predicted to interact in other studies (OPHID [67]: 21815 interactions).

Of the 62322 human proteins from the initial IPI dataset, 22889 were characterized by at least one of the features that we considered to predict interaction (see the Features section). These 22889 human proteins are encoded by 16904 distinct genes and are referred to as the Informative Protein Set. The randomly-generated negative dataset used for the training and testing of the predictor was created by selecting protein pairs at random from the Informative Protein Set. In contrast, the localization-derived negative dataset was created by selecting protein pairs from the Informative Protein Set for which the HPRD [15] annotates one as being primarily in the plasma membrane and the other as primarily in the nucleus. Training and testing was performed with 5-fold cross-validation. In addition, positive to negative ratios of 1:1 and 1:100 were considered.

The predictions were compared to the literature-mined Ramani dataset [16], the orthology-derived Lehner prediction dataset [36] and the probabilistic Rhodes prediction dataset [46]. All three datasets identify the interactions by stating the names and/or gene locus IDs of the genes that encode the interacting proteins. In contrast, we work directly on the protein sequences and so related the gene annotations to our protein identifiers by extracting Entrez Gene IDs corresponding to the IPI protein entries from the IPI cross-reference files (for the IPI release 3.24) [64]. Ensembl gene identifiers (Ensembl 42) were also matched to Entrez Locus IDs (NCBI36) using BioMart [68].

Some gene-gene entries were not recovered in the conversion between different identifiers, or due to the deletion or replacement of some Entrez Locus IDs. Despite this, 37714 gene-gene interactions were recovered from the Rhodes dataset and 6132 interactions from the Ramani dataset as well as 64306 and 10454 interactions from the Lehner full and core datasets respectively.

Learning method
Semi-naïve Bayes classifiers were used to measure the likelihood of interaction of two proteins given the presence of the features considered. This learning method was chosen because it allows the integration of highly heterogeneous data in a model that is easy to interpret and that can readily accommodate missing data. The transparency of the method allows the straightforward determination of which features are most predictive of interaction at the level of the whole proteome as well as for individual protein pairs.

The prediction of protein interaction is a binary problem which can be expressed in Bayesian formalism. We are interested in determining the posterior odds ratio of interaction of two proteins, given the presence of the features we are considering. This posterior odds ratio can be rewritten using Bayes rule:

$$
\begin{aligned}
O_{post} &= \frac{P(I \mid f_1,...,f_n)}{P(\sim I \mid f_1,...,f_n)} \\
&= \frac{\dfrac{P(f_1,...,f_n \mid I)*P(I)}{P(f_1,...,f_n)}}{\dfrac{P(f_1,...,f_n \mid \sim I)*P(\sim I)}{P(f_1,...,f_n)}} \\
&= \frac{P(f_1,...,f_n \mid I)*P(I)}{P(f_1,...,f_n \mid \sim I)*P(\sim I)} \\
&= \frac{P(I)}{P(\sim I)} * \frac{P(f_1,...,f_n \mid I)}{P(f_1,...,f_n \mid \sim I)} \\
&= O_{prior} * LR(f_1,...,f_n)
\end{aligned}
$$

where I is a binary variable representing interaction, $\sim$I represents non-interaction, f_1 through f_n are the features we are considering, O_{prior} is the prior odds ratio and LR is the likelihood ratio.

If the features considered are independent, the likelihood ratio LR can be calculated as the product of the individual likelihood ratios with respect to the features considered separately. If the features are not independent, all possible combinations of all states of these features must be considered, which can be computationally quite intensive. In

the independent case, the likelihood ratio can be calculated as:

$$LR(f_1,...,f_n) = \left[\frac{P(f_1,..., f_n \mid I)}{P(f_1,..., f_n \mid \sim I)} \right]$$

$$= \prod_{i=1}^{n} \left[\frac{P(f_i \mid I)}{P(f_i \mid \sim I)} \right]$$

The likelihood ratios for the different features considered can be estimated by evaluating the ratio of the proportion of interacting and non-interacting proteins for which a particular state of the feature is true in the training set (i.e. by determining to which bin of the feature the protein pair belongs, for every protein pair in the positive and negative training sets). More precisely, the training step consisted of calculating the respective proportions of positive and negative examples that fall into each bin of the feature(s) considered (i.e. that have a particular state). The likelihood ratio of interaction for a given state is simply the ratio of the proportion of all positives that have that state divided by the proportion of all negatives that have that same state. When a particular state of a feature occurs only in positive examples (known interacting proteins), the likelihoods are set to the highest non-infinite value of any state for that feature (to avoid infinite values). Additionally, when no data are available for a specific feature (for a given pair of proteins), the likelihood of the feature is set to 1.0. For a detailed calculation of the likelihoods see Additional File 4.

Prior odds ratio estimate

The prior odds ratio (O_{prior}) is difficult to estimate because we do not know all the true interactions, even for a small subset of proteins. The prior odds ratio of interaction for yeast was estimated by combining all protein-protein interactions (but only those related to direct physical interactions, and no entries derived by synthetic lethal-type experiments) from the BIND, DIP and GRID databases [65,66,69]. This subset of interactions contains 36466 distinct interactions involving 5202 distinct proteins, thus resulting in a prior odds ratio of 1/370. This is most likely a conservative estimate since a certain proportion of interactions remain unknown and so when more data become available, the prior odds ratio will increase. For human proteins, 12191 distinct interactions were recovered, involving 5164 human proteins from the September 2005 version of the HPRD [15] and 26896 distinct interactions involving 7531 human proteins from the June 2006 version, leading respectively to prior odds estimates of 1/1093 and 1/1053. However, taking the subset of 5164 proteins from the September 2005 version that are seen in the June 2006 version (20842 distinct interactions), gave a prior odds of interaction estimate of 1/639. Thus, between the two releases of the HPRD, there was a

large increase in the number of interactions for this subset of proteins and this is likely to continue for at least the next few releases. Accordingly, it is reasonable to conclude that there are not enough known human interactions to calculate a realistic and stable estimate of the prior odds ratio of interactions for human. As a consequence, a prior odds ratio of 1/400 was used for all work in the paper, which is similar to the estimate for yeast and is likely still an underestimate of the true value.

Features

Seven distinct features combined into five modules were investigated as summarized in Table 1 and described below.

1. Expression module

Expression data were downloaded from the Gene Expression Omnibus [70]. The GDS596 dataset was used which examines gene expression profiles from 79 physiologically normal tissues obtained from various sources [71]. Expression data were recovered for 10642 distinct transcripts in 158 different arrays (2 arrays per tissue). Pearson correlations were calculated for all 56620761 transcript pairs and correlation values were grouped into 20 bins of increasing co-expression.

2. Orthology module

Orthology maps between human and yeast, worm and fly were downloaded from the InParanoid database [72]. Interaction datasets for model organisms were downloaded from the BIND [66], DIP [65] and GRID [69] databases. Orthology interaction data were classified into 13 bins. High, medium and low confidence bins were defined for human protein pairs that have interacting orthologs in either yeast, fly or worm (for a total of 9 bins). The high confidence bins were populated by human protein pairs that have interacting orthologs that both achieve an InParanoid score of 1 (i.e. both proteins involved in an interaction in another organism are respectively the best orthology match for the two human proteins under consideration). The medium confidence bins were populated by human protein pairs that have interacting orthologs but only one of the interacting orthologs has an InParanoid score of 1. The low confidence bins were filled by human protein pairs that have interacting orthologs according to InParanoid but neither achieves a score of 1 (i.e. neither is the best match for the two human proteins under consideration). The orthology module has four additional bins: two bin for human pairs that have interacting paralogs in human (a medium and a low confidence bin which use the same definition as above for the model organisms), one bin for human pairs that have interacting homologs in more than one organism (these can be orthologs in yeast, worm or fly, or paralogs in

human) and one bin for human pairs that have only non-interacting orthologs.

3. Combined module
This module incorporates three distinct features in a non-naïve Bayesian framework: subcellular localization, domain co-occurrence and post-translational modification co-occurrence.

Subcellular localization
PSLT (Protein Subcellular Localization Tool) subcellular localization predictions [54] were used to classify protein pairs in one of four groups: pairs of proteins predicted to be in the same compartment, pairs of proteins predicted to be in neighboring compartments (cytosol-nucleus, endoplasmic reticulum-Golgi, Golgi-cytosol, cytosol-plasma membrane, and plasma membrane-secreted), pairs of proteins predicted in different non-neighboring compartments and pairs of proteins for which there were no localization predictions. Neighboring compartments were chosen as compartment pairs sharing a high proportion of proteins, as investigated previously [54].

Co-occurrence of domains
The chi-square test was used as a measure of the likelihood of co-occurrence of specific InterPro domains and motifs [73] in protein pairs. Chi-square scores were calculated for all pairs of domains/motifs that occurred in the training data and were then grouped into 5 bins of increasing value. Additionally, Pfam [74] domain pairs known to interact from three-dimensional structures [75] were included in the highest Chi-square score bin. When protein pairs contained more than one domain pair, the domain pair assigned to the highest Chi-square score bin was used to assign a likelihood of interaction.

Post-translational modification (PTM) pair co-occurrence
Likelihoods were assessed using a PTM pair enrichment score calculated as the probability of co-occurrence of two specific PTMs in all pairs of interacting protein pairs divided by the probability of occurrence of both of these PTMs separately:

$$\text{PTM\_score} = \frac{P\big(\text{PTM}[i],\text{PTM}[j]|I\big)}{P\big(\text{PTM}[i]|I\big) * P\big(\text{PTM}[j]|I\big)}$$

where PTM[i] and PTM[j] are distinct PTMs and I is the set of all interacting proteins that were used to train the predictor. The annotations of PTMs in human proteins were downloaded from UniProt [76] and HPRD [15]. PTM instances described as "predicted", "probable" or "possible" were excluded, leaving 3439 distinct proteins with PTM annotations in the training set. The PTM pair enrichment scores were grouped into 4 bins of increasing value.

The localization, co-occurrence of domains, and PTMs were considered simultaneously to measure their predictive power in assessing the likelihood of protein interaction. To do this, all possible combinations of the 4 localization bins, 5 chi-square domain-co-occurrence bins and 4 PTM_score bins were investigated and are referred to as the **combined module**.

4. Disorder module
It has been suggested that unstructured regions of proteins are often involved in binding interactions, particularly in the case of transient interactions [77]. Protein intrinsic disorder was predicted for all proteins considered by using the VSL2B predictor [78]. The disorder score for protein pairs was then calculated as the sum of percent disorder for each protein of the pair. Disorder scores were grouped into 6 bins of increasing value.

The Expression, Orthology, Combined and Disorder modules are referred collectively as the **Group A modules**. Likelihood ratios for each of the Group A modules are illustrated in Figure 1A (see Additional File 4 for complete likelihood ratios for every possible state of these modules and for detailed calculations of these likelihood ratios).

5. Transitive module
The transitive module works on the premise that a pair of proteins is more likely to interact if it shares interacting partners. It does this by considering the local topology of the network predicted by the integration of the Group A modules as depicted in Figure 2. Thus, the transitive module takes as input the product of the likelihood ratios of all other modules considered by the predictor (as illustrated in Figure 1B). For each pair of proteins in the training set, the product of the likelihood ratios from all other modules (referred to as the preliminary score (PS) in Figure 1) was calculated for all protein pairs neighboring the pair (i.e. all protein pairs which involve one protein from the initial protein pair under study and for which it is possible to calculate such a score). All preliminary scores above 10 were kept. This parameter was determined empirically. A neighborhood topology score T was then calculated as follows:

$$T = \frac{\displaystyle\sum_{e \in E_c} s_e}{1 + |\, E_i \setminus E_c \,| + |\, E_j \setminus E_c \,|}$$

where E_c is the set of edges that connect proteins i and j to their common interactors, E_i is the set of edges that involve protein i, s_e is the score (likelihood ratio) of edge e and $E_i \setminus E_c$ refers to the set difference of E_i and E_c. For a given set of neighbors, T increases as the interactions with these neighbors become more likely (as the sum of s_e

increases). Additionally, the topology score T of a pair of proteins increases as the proportion of likely interactors that these two proteins share increases. The topology scores were grouped into 5 bins of increasing value. It should be noted that the neighborhood topology score calculated for a given protein pair does not consider the preliminary score assigned to that protein pair. It only considers the preliminary scores of its neighbors and so is truly based on the local network topology around that protein pair. Accordingly, the likelihood ratio the transitive module outputs for a given protein pair is independent of the likelihood ratio calculated by the Group A modules for this same protein pair.

Correlation analysis

The Pearson correlation between pairs of modules was estimated by taking 150 samples of 10000 protein pairs each and calculating the Pearson correlation of the likelihood ratios for the two modules considered, for each sample. The reported correlation values are the average of the 150 experiments. Samples of the protein pair space were taken instead of considering the whole space as this was more computationally tractable.

Accuracy measurements

The accuracy of the predictors was measured by performing five-fold cross validation experiments in which the datasets were randomly divided into five non-overlapping sets, four of which were used to train the predictor while the fifth was used to test the prediction accuracy. The accuracy reported is the average measured for all combinations of training and testing sets using these five sets. Testing was done by predicting the total likelihood scores for all protein pairs in the test set using the models computed in the training phase and then counting the number of pairs that were well predicted. We used the area under partial ROC curves as a measure of accuracy. Receiver operator characteristic (ROC) curves plot the true positive rate versus the false positive rate over their full range of possible values. In some circumstances, it is more informative to use partial ROC curves (ROCn curves) which illustrate the number of true positives identified by the predictor that score higher than the n highest scoring negatives, plotted for all values from 0 to n. There are many more negatives than positives in our datasets and this is also thought to be true for the full protein interaction networks we are modeling. Since the aim is to identify the largest number of true interacting pairs while leaving out as many non-interacting pairs as possible, it is most informative to measure the performance of the predictor under conditions of very low false-positive rates. Accordingly, ROC50 and ROC100 curves were analyzed because given the size of the datasets, these curves consider all the protein pairs predicted to have a posterior odds ratio above 1.0, for all the predictors investigated. The area under ROC curves is

often used as a summary measure of accuracy. For ROCn curves, it can be calculated as

$$\text{AUC}_{\text{ROCn}} = \frac{1}{nT} * (\sum_{i=1}^{n} T_i)$$

where i takes on values from 1 to n, T is the total number of positives in the test set and T_i is the number of positives that score higher than the ith highest scoring negative.

Authors' contributions

MSS conceived and designed the study, created and implemented the predictor, analyzed the predictions and drafted the manuscript. GJB participated in the design of the study, the analysis of the predictions and the writing of the manuscript. Both authors read and approved the final manuscript.

Additional material

Additional File 1

High scoring InterPro domain pairs. List of InterPro domain pairs that achieve highest chi-square score of co-occurrence in our set of positive interactors.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-239-S1.xls]

Additional File 2

High scoring post-translational modification pairs. List of high-scoring post-translational modification pairs.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-239-S2.xls]

Additional File 3

All LR400 predicted interactions ranked. All human protein pairs predicted to have a likelihood ratio of interaction greater than 400 and thus a posterior odds ratio of interaction greater than 1.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-239-S3.xls]

Additional file 4

Additional methods. In depth description of the calculation of likelihood ratios for the modules.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-239-S4.pdf]

Acknowledgements

References

1. Xia Y, Yu H, Jansen R, Seringhaus M, Baxter S, Greenbaum D, Zhao H, Gerstein M: **Analyzing cellular biochemistry in terms of molecular networks.** *Annu Rev Biochem* 2004, **73:**1051-1087.

2. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415(6868):**141-147.

3. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.** *Nature* 2002, **415(6868):**180-183.

4. 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 U S A* 2001, **98(8):**4569-4574.

5. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, 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 MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, 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(7084):**637-643.

6. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, 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 JM: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, **403(6770):**623-627.

7. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M: **A map of the interactome network of the metazoan *C. elegans*.** *Science* 2004, **303(5657):**540-543.

8. Formstecher E, Aresta S, Collura V, Hamburger A, Meil A, Trehin A, Reverdy C, Betin V, Maire S, Brun C, Jacq B, Arpin M, Bellaiche Y, Bellusci S, Benaroch P, Bornens M, Chanet R, Chavrier P, Delattre O, Doye V, Fehon R, Faye G, Galli T, Girault JA, Goud B, de Gunzburg J, Johannes L, Junier MP, Mirouse V, Mukherjee A, Papadopoulo D, Perez F, Plessis A, Rosse C, Saule S, Stoppa-Lyonnet D, Vincent A, White M, Legrain P, Wojcik J, Camonis J, Daviet L: **Protein interaction mapping: a *Drosophila* case study.** *Genome Res* 2005, **15(3):**376-384.

9. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, 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 CA, Finley RL Jr., White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM: **A protein interaction map of *Drosophila melanogaster*.** *Science* 2003, **302(5651):**1727-1736.

10. Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N, Davey M, Parkinson J, Greenblatt J, Emili A: **Interaction network containing conserved and essential protein complexes in *Escherichia coli*.** *Nature* 2005, **433(7025):**531-537.

11. 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(7062):**1173-1178.

12. 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, Toksoz 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.

13. Sprinzak E, Sattath S, Margalit H: **How reliable are experimental protein-protein interaction data?** *J Mol Biol* 2003, **327(5):**919-923.

14. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417(6887):**399-403.

15. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HG, Nagini M, Kumar GS, Jose R, Deepthi P, Mohan SS, Gandhi TK, Harsha HC, Deshpande KS, Sarker M, Prasad TS, Pandey A: **Human protein reference database--2006 update.** *Nucleic Acids Res* 2006, **34(Database issue):**D411-4.

16. Ramani AK, Bunescu RC, Mooney RJ, Marcotte EM: **Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome.** *Genome Biol* 2005, **6(5):**R40.

17. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, Stark C, Ho Y, Botstein D, Andrews B, Boone C, Troyanskya OG, Ideker T, Dolinski K, Batada NN, Tyers M: **Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae.** *J Biol* 2006, **5(4):**11.

18. Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biol* 2006, **7(11):**120.

19. Martin S, Roe D, Faulon JL: **Predicting protein-protein interactions using signature products.** *Bioinformatics* 2005, **21(2):**218-226.

20. Chinnasamy A, Mittal A, Sung WK: **Probabilistic prediction of protein-protein interactions from the protein sequences.** *Comput Biol Med* 2006, **36(10):**1143-1154.

21. Bock JR, Gough DA: **Predicting protein--protein interactions from primary structure.** *Bioinformatics* 2001, **17(5):**455-460.

22. Park J, Lappe M, Teichmann SA: **Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast.** *J Mol Biol* 2001, **307(3):**929-938.

23. Sprinzak E, Margalit H: **Correlated sequence-signatures as markers of protein-protein interaction.** *J Mol Biol* 2001, **311(4):**681-692.

24. Szilagyi A, Grimm V, Arakaki AK, Skolnick J: **Prediction of physical protein-protein interactions.** *Phys Biol* 2005, **2(1-2):**S1-S16.

25. Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin AC, Bork P, Superti-Furga G, Serrano L, Russell RB: **Structure-based assembly of protein complexes in yeast.** *Science* 2004, **303(5666):**2026-2029.

26. Huynen M, Snel B, Lathe W 3rd, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences.** *Genome Res* 2000, **10(8):**1204-1210.

27. Gaasterland T, Ragan MA: **Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes.** *Microb Comp Genomics* 1998, **3(4):**199-217.

28. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE: **Co-evolution of proteins with their interaction partners.** *J Mol Biol* 2000, **299(2):**283-293.

29. Pazos F, Valencia A: **Similarity of phylogenetic trees as indicator of protein-protein interaction.** *Protein Eng* 2001, **14(9):**609-614.

30. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci U S A* 1999, **96(8):**4285-4288.

31. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402(6757):**86-90.

32. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285(5428):**751-753.

33. Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M: **Protein interaction mapping in *C. elegans* using proteins involved in vulval development.** *Science* 2000, **287(5450):**116-122.

34. 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 Res* 2001, **11(12):**2120-2126.

35. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M: **Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs.** *Genome Res* 2004, **14(6):**1107-1118.

36. Lehner B, Fraser AG: **A first-draft human protein-interaction map.** *Genome Biol* 2004, **5(9):**R63.

37. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science* 2003, **302(5644):**449-453.

38. Deane CM, Salwinski L, Xenarios I, Eisenberg D: **Protein interactions: two methods for assessment of the reliability of high throughput observations.** *Mol Cell Proteomics* 2002, **1(5):**349-356.

39. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A: **Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets.** *Nat Genet* 2006, **38(3):**285-293.

40. Goldberg DS, Roth FP: **Assessing experimentally derived interactions in a small world.** *Proc Natl Acad Sci U S A* 2003, **100(8):**4372-4376.

41. Yu H, Paccanaro A, Trifonov V, Gerstein M: **Predicting interactions in protein networks by completing defective cliques.** *Bioinformatics* 2006, **22(7):**823-829.

42. Ben-Hur A, Noble WS: **Kernel methods for predicting protein-protein interactions.** *Bioinformatics* 2005, **21 Suppl 1:**i38-46.

43. Lu LJ, Xia Y, Paccanaro A, Yu H, Gerstein M: **Assessing the limits of genomic data integration for predicting protein networks.** *Genome Res* 2005, **15(7):**945-953.

44. Jaimovich A, Elidan G, Margalit H, Friedman N: **Towards an integrated protein-protein interaction network: a relational Markov network approach.** *J Comput Biol* 2006, **13(2):**145-164.

45. Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, Theesfeld CL, Dolinski K, Troyanskaya OG: **Discovery of biological networks from diverse functional genomic data.** *Genome Biol* 2005, **6(13):**R114.

46. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM: **Probabilistic model of the human protein-protein interaction network.** *Nat Biotechnol* 2005, **23(8):**951-959.

47. Qi Y, Bar-Joseph Z, Klein-Seetharaman J: **Evaluation of different biological data and computational classification methods for use in protein interaction prediction.** *Proteins* 2006, **63(3):**490-500.

48. Kemmer D, Huang Y, Shah SP, Lim J, Brumm J, Yuen MM, Ling J, Xu T, Wasserman WW, Ouellette BF: **Ulysses - an application for the projection of molecular interactions across species.** *Genome Biol* 2005, **6(12):**R106.

49. Mika S, Rost B: **Protein-protein interactions more conserved within species than across species.** *PLoS Comput Biol* 2006, **2(7):**e79.

50. Jonsson PF, Cavanna T, Zicha D, Bates PA: **Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis.** *BMC Bioinformatics* 2006, **7:**2.

51. Lim J, Hao T, Shaw C, Patel AJ, Szabo G, Rual JF, Fisk CJ, Li N, Smolyar A, Hill DE, Barabasi AL, Vidal M, Zoghbi HY: **A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration.** *Cell* 2006, **125(4):**801-814.

52. Ben-Hur A, Noble WS: **Choosing negative examples for the prediction of protein-protein interactions.** *BMC Bioinformatics* 2006, **7 Suppl 1:**S2.

53. Myers CL, Barrett DR, Hibbs MA, Huttenhower C, Troyanskaya OG: **Finding function: evaluation methods for functional genomic data.** *BMC Genomics* 2006, **7:**187.

54. Scott MS, Thomas DY, Hallett MT: **Predicting subcellular localization via protein motif co-occurrence.** *Genome Res* 2004, **14(10A):**1957-1966.

55. D'Haeseleer P, Church GM: **Estimating and improving protein interaction error rates.** *Proc IEEE Comput Syst Bioinform Conf* 2004:216-223.

56. Xia K, Dong D, Han JD: **IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model.** *BMC Bioinformatics* 2006, **7:**508.

57. Lu X, Chen J, Sasmono RT, Hsi ED, Sarosiek KA, Tiganis T, Lossos IS: **TCPTP, Distinctively Expressed in ABC-Like Diffuse Large B-Cell Lymphomas, is the Nuclear Phosphatase of STAT6.** *Mol Cell Biol* 2007, **27(6):**2166-2179.

58. Le Clainche C, Schlaepfer D, Ferrari A, Klingauf M, Grohmanova K, Veligodskiy A, Didry D, Le D, Egile C, Carlier MF, Kroschewski R: **IQGAP1 stimulates actin assembly through the N-WASP-Arp2/3 pathway.** *J Biol Chem* 2007, **282(1):**426-435.

59. Xu Y, Wong SH, Tang BL, Subramaniam VN, Zhang T, Hong W: **A 29-kilodalton Golgi soluble N-ethylmaleimide-sensitive factor attachment protein receptor (Vti1-rp2) implicated in protein trafficking in the secretory pathway.** *J Biol Chem* 1998, **273(34):**21783-21789.

60. Galli T, Zahraoui A, Vaidyanathan VV, Raposo G, Tian JM, Karin M, Niemann H, Louvard D: **A novel tetanus neurotoxin-insensitive vesicle-associated membrane protein in SNARE complexes of the apical plasma membrane of epithelial cells.** *Mol Biol Cell* 1998, **9(6):**1437-1448.

61. Mallard F, Tang BL, Galli T, Tenza D, Saint-Pol A, Yue X, Antony C, Hong W, Goud B, Johannes L: **Early/recycling endosomes-to-TGN transport involves two SNARE complexes and a Rab6 isoform.** *J Cell Biol* 2002, **156(4):**653-664.

62. Komamura-Kohno Y, Karasawa-Shimizu K, Saitoh T, Sato M, Hanaoka F, Tanaka S, Ishimi Y: **Site-specific phosphorylation of MCM4 during the cell cycle in mammalian cells.** *Febs J* 2006, **273(6):**1224-1239.

63. Colland F, Jacq X, Trouplin V, Mougin C, Groizeleau C, Hamburger A, Meil A, Wojcik J, Legrain P, Gauthier JM: **Functional proteomics mapping of a human signaling pathway.** *Genome Res* 2004, **14(7):**1324-1332.

64. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R: **The International Protein Index: an integrated database for proteomics experiments.** *Proteomics* 2004, **4(7):**1985-1988.

65. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, **32(Database issue):**D449-51.

66. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, Buzadzija K, Cavero R, D'Abreo C, Donaldson I, Dorairajoo D, Dumontier MJ, Dumontier MR, Earles V, Farrall R, Feldman H, Garderman E, Gong Y, Gonzaga R, Grytsan V, Gryz E, Gu V, Haldorsen E, Halupa A, Haw R, Hrvojic A, Hurrell L, Isserlin R, Jack F, Juma F, Khan A, Kon T, Konopinsky S, Le V, Lee E, Ling S, Magidin M, Moniakis J, Montojo J, Moore S, Muskat B, Ng I, Paraiso JP, Parker B, Pintilie G, Pirone R, Salama JJ, Sgro S, Shan T, Shu Y, Siew J, Skinner D, Snyder K, Stasiuk R, Strumpf D, Tuekam B, Tao S, Wang Z, White M, Willis R, Wolting C, Wong S, Wrong A, Xin C, Yao R, Yates B, Zhang S, Zheng K, Pawson T, Ouellette BF, Hogue CW: **The Biomolecular Interaction Network Database and related tools 2005 update.** *Nucleic Acids Res* 2005, **33(Database issue):**D418-24.

67. Brown KR, Jurisica I: **Online predicted human interaction database.** *Bioinformatics* 2005, **21(9):**2076-2082.

68. **BioMart**   [http://www.biomart.org/]
69. Breitkreutz BJ, Stark C, Tyers M: **The GRID: the General Repository for Interaction Datasets.** *Genome Biol* 2003, **4(3):**R23.
70. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R: **NCBI GEO: mining millions of expression profiles--database and tools.** *Nucleic Acids Res* 2005, **33(Database issue):**D562-6.
71. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A* 2004, **101(16):**6062-6067.
72. O'Brien KP, Remm M, Sonnhammer EL: **Inparanoid: a comprehensive database of eukaryotic orthologs.** *Nucleic Acids Res* 2005, **33(Database issue):**D476-80.
73. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, Wu CH: **InterPro, progress and status in 2005.** *Nucleic Acids Res* 2005, **33(Database issue):**D201-5.
74. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A: **Pfam: clans, web tools and services.** *Nucleic Acids Res* 2006, **34(Database issue):**D247-51.
75. Jefferson ER Walsh, T., Roberts, T. and Barton, G. J.: **SNAPPI-DB: A Database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions.** *Nucleic Acids Research* 2007.
76. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B: **The Universal Protein Resource (UniProt): an expanding universe of protein information.** *Nucleic Acids Res* 2006, **34(Database issue):**D187-91.
77. Singh GP, Ganapathi M, Dash D: **Role of intrinsic disorder in transient interactions of hub proteins.** *Proteins* 2006, **66(4):**761-765.
78. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z: **Length-dependent prediction of protein intrinsic disorder.** *BMC Bioinformatics* 2006, **7:**208.