SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
BIOINFORMATICS
SYSTEMS BIOLOGY
COMPUTATIONAL BIOLOGY
GENOMICS

Received
2 November 2011

Accepted
14 December 2011

Published
30 January 2012

Correspondence and requests for materials should be addressed to A.G. (ashkan_golshani@carleton.ca)

Short Co-occurring Polypeptide Regions Can Predict Global Protein Interaction Maps

Sylvain Pitre[1], Mohsen Hooshyar[2], Andrew Schoenrock[1], Bahram Samanfar[2], Matthew Jessulat[2], James R. Green[3], Frank Dehne[1] & Ashkan Golshani[2]

[1]School of Computer Science, Carleton University, Ottawa, Canada. [2]Department of Biology, Carleton University, Ottawa, Canada. [3]Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada.

A goal of the post-genomics era has been to elucidate a detailed global map of protein-protein interactions (PPIs) within a cell. Here, we show that the presence of co-occurring short polypeptide sequences between interacting protein partners appears to be conserved across different organisms. We present an algorithm to automatically generate PPI prediction method parameters for various organisms and illustrate that global PPIs can be predicted from previously reported PPIs within the same or a different organism using protein primary sequences. The PPI prediction code is further accelerated through the use of parallel multi-core programming, which improves its usability for large scale or proteome-wide PPI prediction. We predict and analyze hundreds of novel human PPIs, experimentally confirm protein functions and importantly predict the first genome-wide PPI maps for *S. pombe* (~9,000 PPIs) and *C. elegans* (~37,500 PPIs).

Protein-protein interactions (PPIs) represent an essential aspect of all biological pathways and signaling mechanisms within a cell, and are reliable indicators of functional associations between proteins. Consequently, a major goal of the post-genomics era has been to elucidate a detailed global map of PPIs within a cell. To date, high throughput attempts, which are both time and resource demanding, have been utilized to study the global PPI networks of only a few model organisms. Computational methods provide an attractive alternative. However, due to computational limitations, elucidating the PPI networks of complex organisms such as human has not been possible. Furthermore, with the exception of *S. cerevisiae*[1], other organisms are yet to be studied at an all-to-all level where all proteins are analyzed for their abilities to interact with all other proteins, again, due to the computational complexity of most approaches.

Several computational methods for predicting PPIs require in depth knowledge regarding the proteins including structure, sub-cellular location, function, interacting domains, etc[2]. The obvious drawback of such methods is their limited applicability for predicting interactions in organisms which have limited information available. This is apparent by the use of only model organisms (*e.g. S. cerevisiae*) or widely studied organisms (*e.g.* human) used in the testing and application of such methods.

Some PPI prediction methods, however, are based on sequence data only[1,3–10]. A growing body of evidence supports the usefulness of short co-occurring polypeptide sequences (interaction codes) in predicting PPIs in yeast[1,7,11]. As discussed in[1,7,11], there appear to exist a finite number of interaction codes of length around 20 amino acids that mediate a subset of PPIs. However, there is no information on the global applicability of this approach, as it has been until now computationally infeasible to apply these techniques to more complex organisms with larger proteomes. Establishing the conservation of co-occurring polypeptide codes will not only provide further evidence for their activities in mediating PPIs but, more importantly, it would highlight the applicability of this approach to detect genome-wide PPIs in other organisms such as humans. In addition, it allows for the prediction of PPIs in newly sequenced organisms for which limited or no experimental PPI data is available.

Through significant computational acceleration of our previous approach[1,7] via massive parallelization of our previous software, it is now possible to scan larger and more complex proteomes. In this paper, we will investigate the possibility of predicting PPIs based on sequence data only in various organisms including *S. pombe*, *C. elegans*, *E. coli*, and most importantly human. In addition to our previously published genome-wide scan of the *S. cerevisiae* interactome, we will present two new genome-wide scans (*S. pombe* and *C. elegans*) along with validation and analysis of the resulting predicted interactions. We will also explore part of the human interactome.

Table 1 | Summary of information gathered for different organisms tested (number of protein sequences, number of interactions and the source of the interactions)

Organism	Number of Proteins	Number of Interactions	Interaction Database
C. elegans	23,684	6,607	BioGRID[15]
E. coli	4,290	16,235	EciD[19]
H. sapiens	22,513	41,678	HPRD[20] and BioGRID[15]
S. cerevisiae	6,716	43,591	BioGRID[15]
S. pombe	5,024	2,951	BioGRID[15]

Table 2 | Definitions of performance measures used in this paper. TP = True positives, FP = False positives, TN = True negatives and FN = false negatives

Measure (abbr.)	Equation
Sensitivity (Sens.)	TP/(TP+FN)
Specificity (Spec.)	TN/(TN+FP)
Precision (Prec.)	TP/(TP+FP)
Accuracy (Acc.)	(TP+FN)/(TP+TN+FP+FN)
F-measure (Fm)	(2 Prec. Sens.)/(Prec.+Sens.) = 2 TP/(2 TP+FP+FN)

Results

Conservation of interaction codes. Having previously demonstrated the utility of interaction codes for the prediction of PPIs in yeast[1,7], here we explore the conservation of interaction codes by investigating whether PPIs in organisms other than yeast may also be predicted from their primary sequences. In the current study, we use co-occurring short polypeptide regions identified from previously reported PPIs to predict novel PPIs. Interaction codes were computationally identified from the available PPI data (see Table 1) in four model organisms (*E. coli*, *S. pombe*, *C. elegans*, *S. cerevisiae*) and in humans. Our massively parallel, accelerated software, running on a large-scale parallel computer using MPI (Message Passing Interface) and Intel Cilk Plus, is the sole reported experimental technique that is capable of exhaustively scanning all potential PPIs in complex organisms such as *C. elegans* (approximately 23,000 proteins) whose potential interactome contains 2.8×10^8 protein pairs. Unlike other reported techniques that rely on a set of known domains or structures[2], our approach is capable of predicting interactions between completely uncharacterized proteins based solely on primary sequence (and a starting PPI dataset). The performance measures used in the paper are described in Table 2.

We conducted leave-one-out (LOO) *in silico* experiments to characterize the sensitivity (i.e. relative size of the subset of the true interactions that can be detected by our method) and specificity (i.e. 1 – false positive rate) of our approach for each organism (Figure 1) as well as the impact of the positive: negative ratio on the precision (i.e. the proportion of predicted PPIs that will represent true verifiable interactions) (see Supplementary Figure S1 and discussion below). The results in Figure 1 are reported using receiver operator characteristic (ROC) curves, which plot the achievable sensitivity at a given false positive rate. Improved performance is reflected in curves with a stronger bend towards the upper-left corner of the ROC graph (i.e. high sensitivity is achieved with a low false positive rate), while a random decision rule results in the diagonal line shown in Figure 1. The ROC curves in Figure 1 show that our approach can successfully detect PPIs, from the interaction codes, in all five organisms that we investigated. It is critical to operate at extremely high

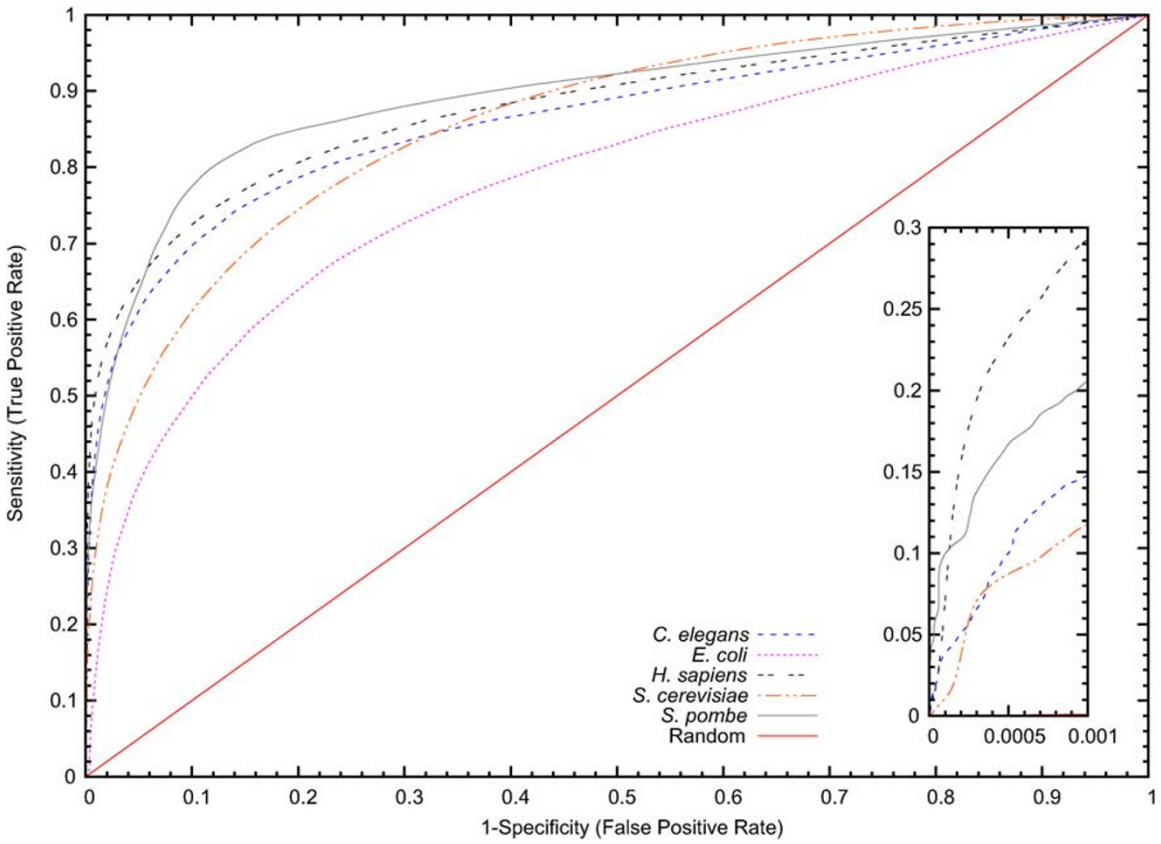

Figure 1 | ROC (Receiver Operating Characteristic) curve for *C. elegans*, *E. coli*, *H. sapiens*, *S. cerevisiae* and *S. pombe*. The curve presents the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity). Inset: performance at very high specificity (99.95%). *H. sapiens* has greater sensitivity than all other organisms tested. Note that due to the scaling of the axes, the diagonal random curve appears flat in the inset.

Table 3 | Chosen operating points for the various organisms tested. In order to reduce the number of false positives, a specificity of 99.95% is typically chosen (* for *E. coli* a lower specificity of 99.0% was chosen due to the small size of the known interaction set). The positive-to-negative ratios (PNR) of the test sets are also given

Organism	Spec. (%)	Sens. (%)	PNR
C. elegans	99.95	9.97	1 : 15.1
E. coli	99.0*	14.48	1 : 6.2
H. sapiens	99.95	23.22	1 : 2.4
S. cerevisiae	99.95	8.77	1 : 2.3
S. pombe	99.95	16.89	1 : 33.9

specificities when applying such a method to a genome-wide analysis due to the expected sparsity of the true interactome. Otherwise, in a genome-wide analysis, true predicted PPIs will be vastly outnumbered by false positives. Therefore, insets are provided for each ROC curve highlighting the expected prediction accuracy at very high specificities (99.9 – 99.95%). From Figure 1, it appears that our method has the highest accuracy for human proteins, followed by *S. pombe*, *C. elegans*, and *S. cerevisiae*. For example, at a specificity of 99.95%, 23.8% of human PPIs can be predicted. When operating at lower specificities, the accuracy for *S. pombe* surpasses that for humans. The sensitivity for *E. coli* at 99.95% specificity is zero (see inset of Figure 1), however sensitivity surpasses 60% when operating at 80% specificity which is typically useful when analyzing a smaller set of preselected protein pairs. Table 3 lists the specificity and sensitivity achieved with our method for each organism at the threshold value used for all subsequent analysis. Although the positive-to-negative ratio (PNR) ratio does not affect either sensitivity or specificity (see discussion below), we have listed the PNR for each organism's test set in Table 3. Except where listed otherwise the PNRs in Table 3 are applicable. The impact of homologous sequences in our interaction databases had an inconsistent effect on performance as discussed in detail below.

Prediction of cross-species PPIs using interaction codes. Next we investigated whether PPIs in a target organism can be predicted from cross-species PPIs (Figure 2). Specifically, can we predict PPIs in one organism using known PPIs from a different organism? Note that Park[12] also attempted to answer this question using our previous method[1] in conjunction with others. However[12], used the parameter settings for *S. cerevisiae* in all of his experiments which leads to reduced performance (see discussion below). To answer the above question, we first investigated *H. sapiens* and *S. cerevisiae* because they have the highest number of previously reported interactions. As indicated in Figure 2, using known *S. cerevisiae* PPIs to predict *H. sapiens* interactions (SC-HS) is a weaker predictor than using known *H. sapiens* PPIs to predict *H. sapiens* interactions (HS-HS). However, it is surprising that the cross-species predictors (i.e. SC-HS and HS-SC) can still predict meaningful interactions at 80% specificity (46–48% sensitivity) for both organisms. This illustrates that known PPIs in one organism may be successfully used to predict novel interactions in another.

We then investigated the applicability of this approach towards predicting PPIs in one organism from a collection of independent PPIs from an ensemble of several other organisms. For all eukaryotes tested, our ROC curves (Figure 3) suggest again the ability of interaction codes from different species to predict interactions in an independent species. The efficiencies of these predictions are lower than the experiments in Figure 1 but similar to the results in Figure 2 for *H. sapiens* and *S. cerevisiae*. In fact for *H. sapiens* we see a slight sensitivity improvement at 80% specificity (54–55% sensitivity as

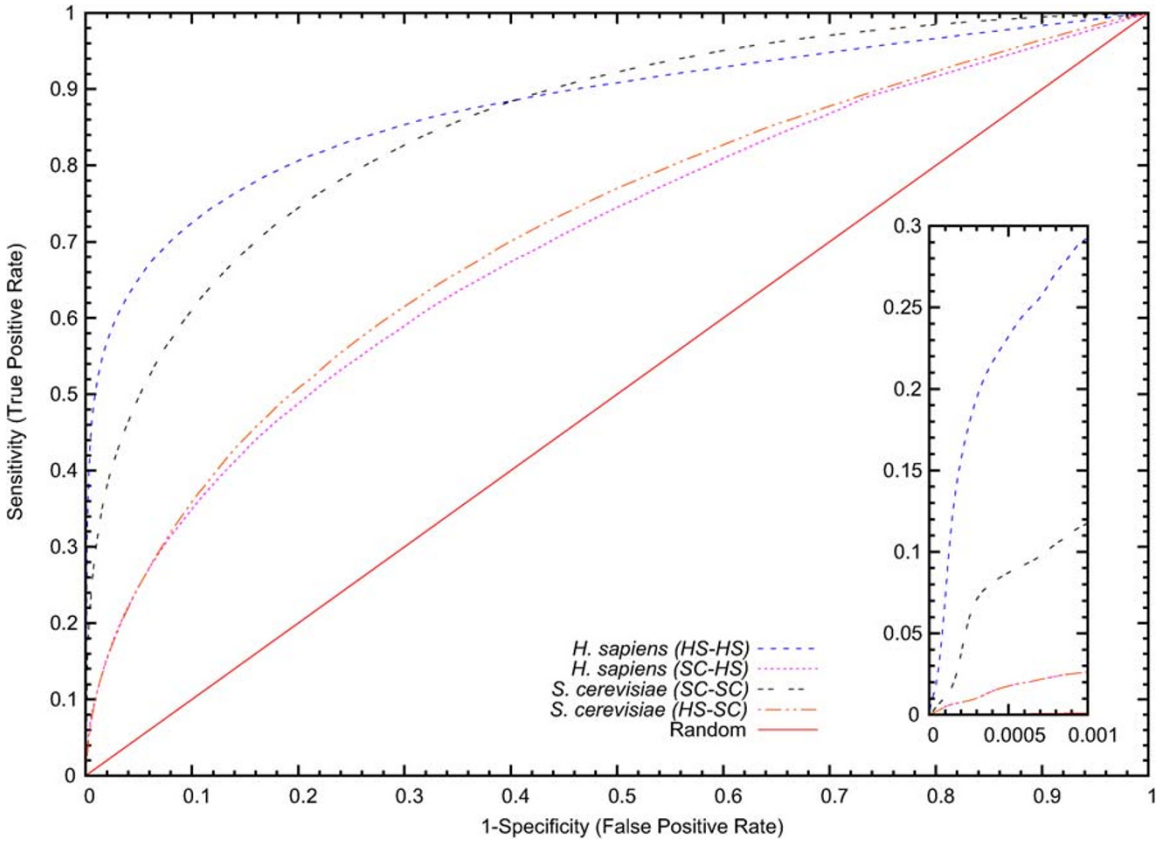

Figure 2 | Cross-species prediction. Here we see the results of using known *H. sapiens* interactions to predict *S. cerevisiae* interactions (HS-SC) and *known S. cerevisiae* interactions to predict *H. sapiens* interactions (SC-HS) compared to same-species prediction (HS-HS and SC-SC).

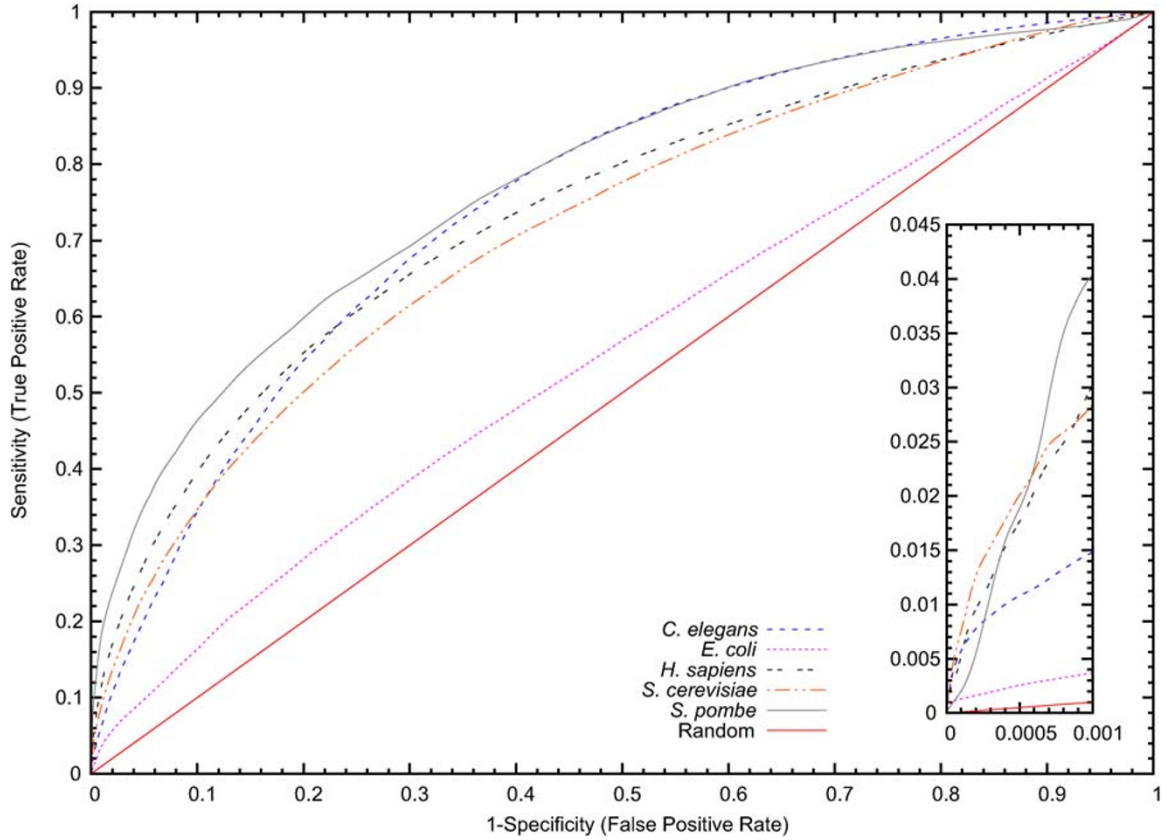

Figure 3 | ROC curve illustrating the accuracy of cross-species PPI predictions. Here, predictions are based on the union of known interactions from several organisms excluding the test organism. Inset: performance at very high specificity (99.95%). Note that due to the scaling of the axes, the diagonal random curve appears flat in the inset.

opposed to 48–49%) when predictions are based on known PPIs from multiple organisms rather than just *S. cerevisiae*. PPI prediction in *E. coli* seems to be an exception, suggesting differences between conservation of protein codes in *E. coli*, a prokaryote, and the eukaryotic organisms examined here.

Independent comparison of our approach and the impact of class imbalance. Our method has been independently evaluated by Park in[12] and compared against three other approaches[4,6,8] using a PNR of 1 : 100. Park compared the four methods using four test categories[12]: (a) *H. sapiens* data to predict *H. sapiens* PPIs, (b) *S. cerevisiae* data to predict *S. cerevisiae* PPIs, (c) *H. sapiens* data to predict *S. cerevisiae* PPIs and finally (d) *S. cerevisiae* to predict *H. sapiens* PPIs. Tests (a) and (b) evaluate how well each method predicts PPIs in an organism given data about that same organism, while (c) and (d) assess cross-species predictions. Since we already explained the need for high specificities for genome-wide predictions, Figure 4 presents the sensitivity results of the experiments in[12] for specificities higher than 95%. We can see that, apart from a close tie with[6] in Figure 4(a) at 95% specificity, our method achieves the highest sensitivity across all four experiments (Figure 4(a), (b), (c) and (d)) at high specificities. In fact at 99.95% specificity, our method attains 3.6 to 17.2 times the sensitivity of the other methods. Also note that Park[12] used the parameter settings for *S. cerevisiae* in all of his experiments even to predict *H. sapiens* PPIs. Amino acid distributions within proteins differ significantly between different organisms; see Supplementary Figure S8. Therefore, it is not optimal to run the same PPI prediction method for different organisms. Supplementary Figure S9 shows the difference between simply applying the unmodified PPI prediction method using yeast parameters to human proteins and the adapted human PPI prediction method using the PPI prediction adaptation algorithm described in Materials and Methods.

Global interactomes are expected to be highly imbalanced, where the vast number of protein pairs are not expected to form true interactions. A recent paper[9] compares their method to two other sequence-based prediction approaches[4,8] using an unbalanced dataset where there are significantly more negative pairs than positives pairs used for training. The paper evaluates balanced datasets (1 : 1 PNR) up to a mildly unbalanced 1 : 15 PNR comparing methods using a combined sensitivity and precision measure called the *F-measure*. Yu *et al* reported that the *F-measure* for all methods tested decreased as the PNR increased. Using the same datasets published in[9], we tested our method using the most unbalanced dataset examined in the paper (1 : 15 PNR). Note that the real PNR in *H. sapiens* is expected to be as high as 1 : 500 or more, but Yu *et al* only examined up to 1 : 15 ratio due to technical limitations. Table 4 presents the results for the three methods published in Yu *et al* and also our own method. Using this unbalanced dataset our approach achieves a significantly higher *F-measure* of 62.9 $\pm$ 1.1 while the next-best method[9] yields a value of 43.6 $\pm$ 1.3. It is noteworthy that our method also recorded better accuracy, precision, sensitivity and specificity that any other method tested. Supplementary Figure S10 presents the precision-recall curve for our method using Yu *et al*'s 1 : 15 ratio dataset.

We have emphasized sensitivity and specificity as performance measures in this paper (as opposed to precision or *F-measure*) since they are independent of the ratio of positive to negative protein pairs in the test data. As illustrated in Table 2, precision depends directly on this ratio, and furthermore, the true ratio of interacting protein pairs is unknown for most organisms. Supplementary Figure S1 illustrates the precision of our method over a range of PNRs when operating at the threshold values given in Table 3. As expected, the precision is negatively impacted by a decreasing ratio of positive-to-negative test samples as the positive predictions become increasingly dominated by false positives. Supplementary Figure S2 compares our

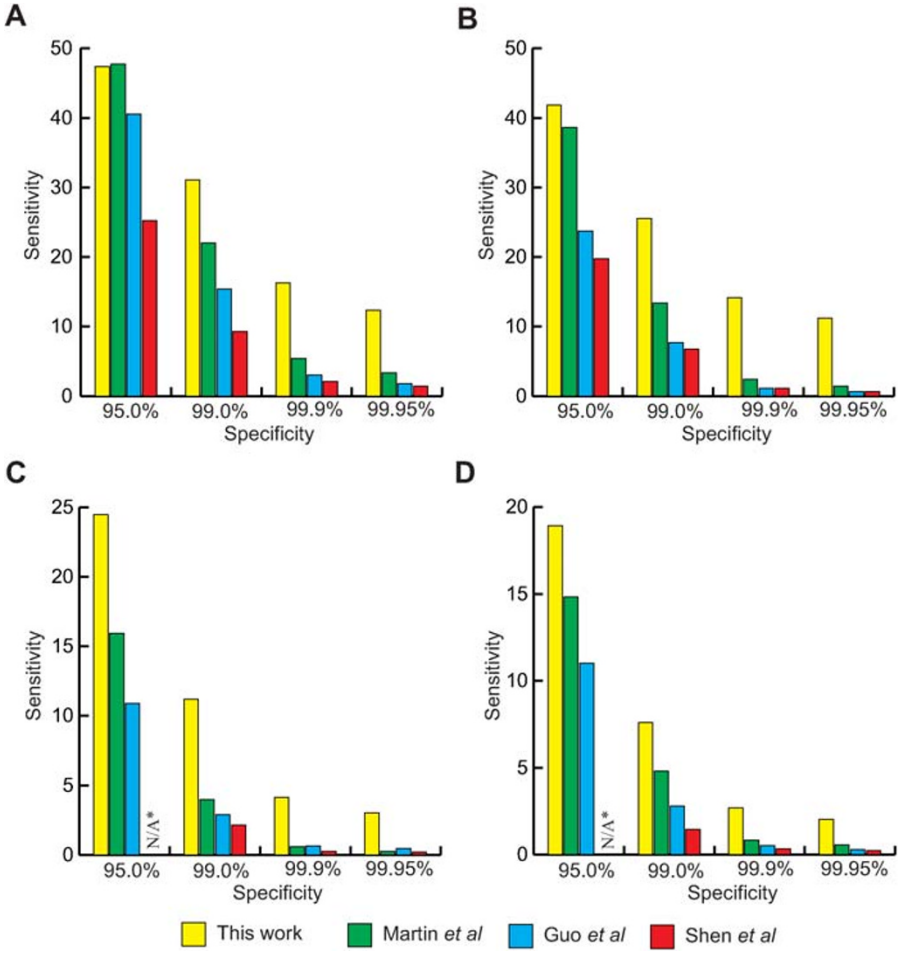

Figure 4 | Accuracy vs. specificity for this work compared to previous work[4,6,8] as found in[12] using a PNR of 1 : 100. (a) Using *H. sapiens* data to predict *H. sapiens* interactions. (b) Using *S. cerevisiae* data to predict *S. cerevisiae* interactions. (c) Using *H. sapiens* data to predict *S. cerevisiae* interactions. (d) Using *S. cerevisiae* data to predict *H. sapiens* interactions. *The experiment by Park in[12] did not have results for 95.0% specificity for Shen *et al* in (c) and (d): the results for this method jumped from 2–3% specificity to 97–98% with no intermediate values in both experiments.

precision over human proteins compared to three competing methods examined in Yu et al[9] over a range of ratios from 1 : 1 to 1 : 1000. Note that this far exceeds the range of ratios examined in[9]. Our method clearly outperforms all other methods at all ratios. It should be noted that, unlike other algorithms based on machine learning, our method does not require training *per se*. The only parameter which must be tuned is the PAM sequence window similarity score threshold, which is determined using random sequences drawn from the target organism's proteome (i.e. not from the database of known PPIs; see Methods for details). All performance results are computed using a stringent leave-one-out protocol which ensures that the test data is independent from the database of known interactions on which the predictions are based.

Effects of homologous sequences. It may be expected that homologous sequences in our evaluation data sets may lead to overfitting of the method and a corresponding overestimate of prediction performance. We therefore investigated the impact of homologous sequences in our data. Following the approach taken by Park[12], for each organism all homologous sequences were removed from the databases such that the remaining proteins share less than 40% sequence identity. The most dramatic effect was observed for human, where removing homologs reduced our sequence set from 22,513 to 14,867 proteins. This resulted in a corresponding decrease in our known interaction set from 41,678 to 19,588 pairs (47% of the original set). LOO analysis was repeated for each organism to compare the performance with and without homologous sequences. Results from these experiments are illustrated using ROC curves in Supplementary Figures S3–S7. For human, at lower specificities we notice a drop in sensitivity of approximately 7–8%, however, at high specificities (required for all-to-analysis as discussed below) the reduction in sensitivity becomes less pronounced. *S. cerevisiae* was

Table 4 | Comparison of this work with previous works[4,8,9] using the evaluation datasets with 1 : 15 PNR compiled by Yu *et al*[9]. The cutoff parameter was chosen to maximize the *F-measure* in order to compare results with those in Yu *et al*[9]

	Acc. (%)	Fm (%)	Prec. (%)	Sens. (%)	Spec. (%)
Shen *et al*[8]	92.5 ± 0.1	33.1 ± 1.4	37.5 ± 1.3	29.7 ± 1.5	96.7 ± 0.1
Guo *et al*[4]	91.7 ± 0.2	36.6 ± 1.5	35.1 ± 1.5	38.3 ± 1.9	95.3 ± 0.2
Yu *et al*[9]	93.7 ± 0.2	43.6 ± 1.3	49.5 ± 1.7	39.0 ± 1.3	97.3 ± 0.1
This work	**95.7 ± 0.1**	**62.9 ± 1.1**	**73.7 ± 2.8**	**55.0 ± 1.6**	**98.6 ± 0.2**

Table 5 | Percentages of *S. pombe, C. elegans* and *H. sapiens* pairs in which both partners share the same GO SLIM annotation as well as third party interactions. (a) Results for 100,000 random *S. pombe* pairs. (b) Results for 2,951 previously known *S. pombe* interactions from BioGRID. (c) Results for our 9,009 *S. pombe* predicted interactions. (d) Results for 100,000 random *C. elegans* pairs. (e) Results for 6,607 previously known *C. elegans* interactions from BioGRID. (f) Results for our 37,572 *C. elegans* predicted interactions. (g) Results for 100,000 random *H. sapiens* pairs. (h) Results for 41,678 previously known *H. sapiens* interactions from BioGRID. (i) Results for our 1,056 *H. sapiens* predicted interactions

	Derived from GO annotation				
	Cellular Component (CC)	Molecular Function (MF)	Biological Process (BP)	CC & MF & BP	Third Party Interaction
(a) Random *S. Pombe* pairs	32.6%	2.4%	7.3%	1.0%	0.032%
(b) Previously detected *S. Pombe* interactions	79.0%	53.9%	54.2%	29.9%	76.5%
(c) Predicted *S. Pombe* Interactions	61.8%	39.9%	34.5%	18.9%	36.2%
(d) Random *C. elegans* pairs	1.1%	2.4%	4.8%	0.2%	0.08%
(e) Previously detected *C. elegans* interactions	9.7%	44.6%	48.5%	4.2%	28.8%
(f) Predicted *C. elegans* Interactions	4.7%	30.2%	30.2%	1.8%	6.8%
(g) Random *H. sapiens* pairs	31.7%	32.7%	28.2%	11.3%	0.3%
(h) Previously detected *H. sapiens* interactions	82.5%	90.1%	82.2%	67.7%	59.2%
(i) Predicted *H. sapiens* interactions	91.2%	92.9%	88.9%	83.2%	45.3%

affected to a lesser degree (4% sensitivity decrease at 99.95% specificity) and in the other organisms (*C. elegans* and *S. pombe*) the results actually improved slightly at high specificities. Again, PPI prediction in *E. coli* seems to be an exception, suggesting differences between conservation of protein interaction codes in *E. coli*, a prokaryote, and eukaryotic organisms. Even with homologous sequences removed from our human interaction database we continue to perform as well as the methods presented in Table 4. This is despite the fact that the data sets used to train those methods[9] contain a large percentage of homologs (estimated to represent 55% of their overall positive data set when following the same homology analysis used here[12]).

Computational prediction of *S. pombe* and *C. elegans* global interactome. At this point, the run time for our approach still precludes analysis of the full human global protein interaction map (requiring ≈6.3 million CPU hours). We therefore targeted *S. pombe*, which has 5,024 proteins and has the second best ROC curve in our initial small scale LOO analysis. Large scale genome-wide analysis was conducted on all possible *S. pombe* protein pairs took approximately 110 CPU hours compared to ~3700 CPU hours for the genome-wide scan of *S. cerevisiae* reported in[1]. We detected a total of 9,009 possible interactions, 6,058 of which are novel and suitable for experimental validation (see Supplementary Table S1). Since currently there are only 2,951 known interaction pairs in *S. pombe*, our predictions have potentially increased our knowledge of the *S. pombe* interactome by over three fold. To examine the quality of the predicted interaction pairs we classified them according to their molecular function, biological process, and sub-cellular location (Table 5). As indicated in Table 5(a) to (c), a significant portion of protein pairs predicted via interaction codes have similar functions (39.9%), occur in the same cellular component (61.8%) and participate in the same cellular process (34.5%). This is comparable to levels of agreement for previously reported protein pairs, which are 53.9%, 79.0%, 54.2%, respectively, and significantly higher than those for random pairs (2.4%, 32.6% and 7.3%, respectively). Third party interactions (where both partners interact with another common protein) were also investigated to further assess the quality of our predicted interactions. Again 36.2% of protein pairs predicted via interaction codes had a common third protein partner, compared to 76.5% for previously reported and only 0.032% for random pairs. It should be noted however that for these analyses (especially third party interaction), the previous experimentally detected interactions might have an unfair advantage since proteins known to interact are often prioritized for further analysis and characterization.

Following our *S. pombe* experiments, *C. elegans* was targeted as another model organism of interest. It has nearly five times more confirmed proteins than *S. pombe* (23,684 compared to 5,024 respectively). A genome-wide scan of *C. elegans* was completed in approximately 150,000 CPU hours and resulted in 37,572 possible interactions, 31,056 of which are novel predictions. We again classified those predictions according to their molecular function, biological process, and location inside the cells as we did with our *S. pombe* predictions (Table 5(d) to (f)). For our predictions, the percentage of pairs that simultaneously have similar function, occur in the same cellular component, and also participate in the same cellular process is 1.8%, which is consistent with the percentage for previously reported protein pairs (4.2% for 6,607 pairs). In contrast, for randomly selected protein pairs, the percentage of pairs that have similar function, occur in the same cellular component and participate in the same cellular process is only 0.2% (for 100,000 tested random pairs). The results for similar function and cellular process are analogous to those of *S. pombe.* However, the values for cellular component do not appear to follow the same trend. This might be due to a lack of reliable data since even the experimentally determined PPIs do not show any enrichment. For proteins pairs predicted in *C. elegans,* 6.8% had a common third protein partner compared to 28.8% in previously reported pairs and only 0.08% in random pairs.

Note that our PPI prediction software is based solely on conserved interaction codes and does not use any information about a protein's molecular functions, biological processes or sub-cellular location. Therefore, such additional information about the biological activities of protein pairs can potentially also be utilized to determine an independent confidence level for a predicted interaction. Since true interactors are thought to be functionally related, participate in the same cellular process, occur in the same sub-cellular location, and interact with a same interacting partner (third partner), such information can be used to determine a confidence level for a predicted interacting pair. For example, a predicted protein pair "A" with both partners sharing the same 1. molecular function, 2. cellular process and 3. interacting partner, may have a higher confidence level than protein pair "B" with partners having the same 1. molecular function and 2. sub-cellular location, which may have a higher confidence level than protein pair "C" with partners only being co-localized in the same sub-cellular location. It should be noted however, that this approach may discriminate against uncharacterized proteins for which limited information is available.

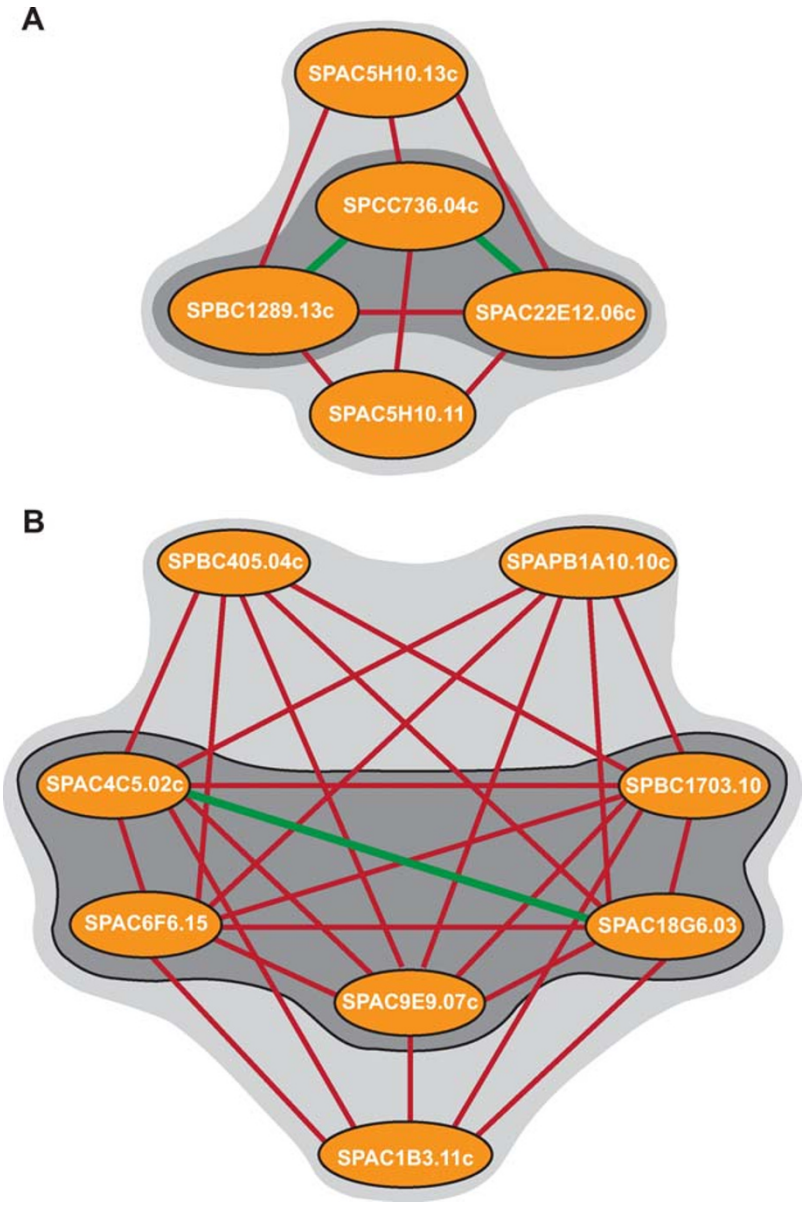

Figure 5 | (a) A novel five member complex for *S. pombe* identified using interaction codes.Previously reported PPIs are shown in green; novel PPIs are shown in red. Previous annotations of key proteins suggest that the complex may play a role in galactosylation of glycoproteins. The dark grey shade represents core proteins. (b) A novel eight member complex for *S. pombe* formed by 5 fully-connected core proteins (shaded) and three peripheral proteins. Predicted PPIs combined with previous annotation of some proteins suggest putative function in protein transport and vesicular trafficking. The dark grey shade represents core proteins.

Prediction of novel protein complexes. PPI data can also be used to determine protein complexes. From our predicted novel PPIs we identified two novel complexes for *S. pombe* membrane proteins. Membrane proteins often provide a challenge for classical PPI detection methods. The first is a five member complex shown in Figure 5A with SPBC1289.13c, a putative galactosyltransferase, as a core protein that interacts with four other proteins, some of which also interact with each other. Two of these proteins, SPAC22E12.06c and SPCC736.04c, are known to have alpha-1,2-galactosyltransferase activities involved in N-linked, and both O-linked and N-linked oligosaccharide modification of proteins, respectively[13], and have previously been reported to interact with each other (previously reported PPIs are shown in green). All five proteins are membrane proteins that are associated with the Golgi apparatus. Therefore it is likely that this complex has a role in galactosylation of glycoproteins. The second complex shown in Figure 5B consists of 8 members, with 5 proteins forming the core, and 3 additional proteins that interact with the core proteins but not with each other. They are all thought to be membrane proteins. Five of these proteins (SPBC1703.10, SPAC4C5.02c, SPAC6F6.15, SPAC9E9.07C and SPAC18G6.03) have been linked to protein transport and vesicular trafficking. This suggests a role for the complex in this process.

Investigation of human dsDNA break repair PPI network. To evaluate the effectiveness of interaction codes to predict PPIs in human we investigated the interactions for 29 proteins, with established roles in the efficiency of double stranded (ds) DNA break repair, against all human proteins. This represents an analysis of more than 650,000 possible protein pairs (29 DNA break repair proteins against all 22,500 reviewed *H. sapiens* proteins from Uniprot[14]). dsDNA breaks represent a severe case of DNA damage which can lead to cancer development if left unrepaired.

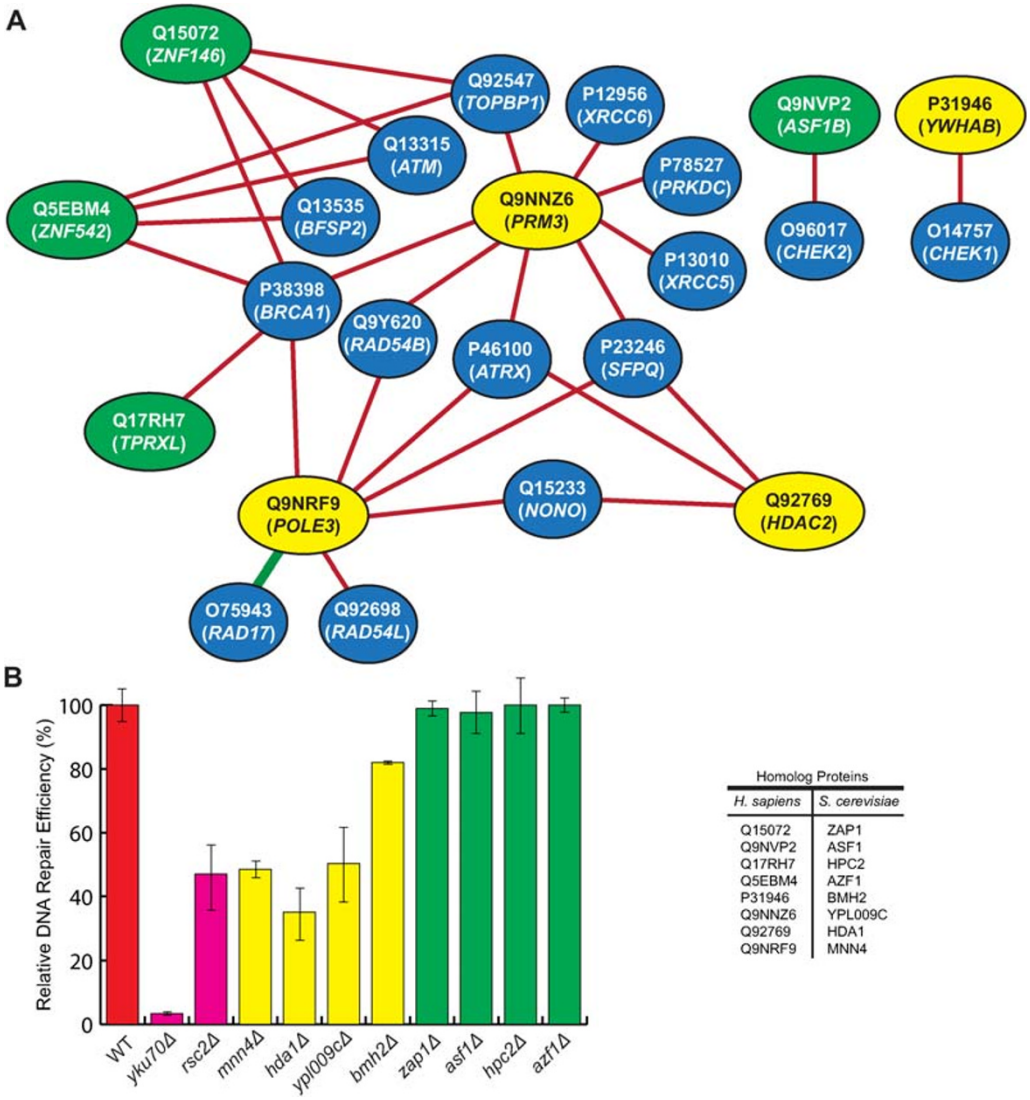

Figure 6 | Inferring protein function from predicted PPIs. Biological process may be inferred based on the annotations of predicted interaction partners. (A) Predicted PPIs for 8 novel human proteins (yellow and green nodes) against known dsDNA break repair proteins (blue nodes). Red edges represent novel predicted interactions and the green edge represents a previously reported interaction. (B) Yeast deletion mutant strains for genes homolog to human novel genes above are subjected to plasmid repair assay. The number of colonies formed after strain transformation with linearized plasmid is normalized to that of intact plasmid, and related to the wild type (red bar) strain (set at 100%). *YKU70* and *RSC2* are known players in non-homologous repair of double stranded DNA breaks and are used as positive controls (pink bars). Yellow bars represent strains with statistically significant (P-value $<$ 0.05) reduction efficiency in plasmid repair. The strains that did not show reduction in plasmid repair efficiency are represented by green bars. Each experiment was repeated at least four times. Error bars in this figure represent the standard deviation between experiments.

We detected a total of 620 interactions (at 99.95% specificity), 349 of which were previously reported, and 271 of which represented novel interactions (see Supplementary Table S2). This represents an expansion of the known dsDNA break repair interactome by an additional 75%.

Following the PPI prediction verification analysis used above, we again observed (Table 5 (g) to (i)) that a significant portion of our predicted partners have the same molecular function (92.9%), are involved in the same cellular processes (91.2%) and are co-localized (88.9%), highlighting the quality of our predictions. Interestingly our predicted interacting partners had more in common with each other than previously reported partners (90.1%, 82.5% and 82.2%, respectively). This is consistent with the ROC curves which show better prediction abilities in human than in *S. pombe*. Predicted PPIs are further verified when third party interactions are examined: only 0.3% of random protein pairs had third party interactions, while this number rose to 45.3% in predicted PPI pairs – much closer to the 59.2% observed among previously reported PPIs.

An interaction between two proteins often infers a functional relationship between the two. We next examined how our predicted PPIs can be used to infer novel protein function. We hypothesized that 8 proteins Q9NNZ6 (PRM3), Q92769 (HDAC2), Q9NRF9 (POLE3), P31946 (YWHAB1), Q17RH7 (TPRXL), Q15072 (ZNF146), Q9NVP2 (ASF1B), and Q5EBM4 (ZNF542) may be involved in double stranded DNA (dsDNA) break repair due to their novel interactions with proteins known to be involved in this process (Figure 6A). To test this hypothesis we experimentally investigated the activity of these proteins. This was done by subjecting the yeast gene deletion mutants for corresponding human homologs to a plasmid repair assay (Figure 6B). It was observed that deletion of 4 (*YPL009C*, *HDA1, MNN4,* and *BMH2*) of the 8 genes (or 50% of the tested genes) resulted in a statistically significant (P-value $<$ 0.05) decrease in the ability of the mutant cells to repair dsDNA breaks. This suggests an involvement for their corresponding human genes *PRM3*, *HDAC2*, *POLE3*, and *YWHAB1*, respectively in the efficiency of DNA damage repair. Due to technical limitations associated with

our plasmid repair approach, it is likely that more of the tested candidate genes may, in fact, be involved in dsDNA break repair. Their deletion may have subtle (or no) phenotypic consequences that cannot be detected by our current experimental method.

Some of the novel proteins in Figure 6A, for example, Q9NNZ6 and Q5EBM4 form numerous interactions with the known dsDNA break repair proteins, whereas others, for example, Q17RH7 and P31946 form a single interaction each. It is generally thought that the degree of connectivity of a protein within a PPI network may represent the significance of that protein within the system. Consequently, the "hub" proteins are thought to be essential for the integrity of a PPI network. In agreement with this, 3 of the 4 identified novel proteins are highly connected and only one novel protein, P31946, forms a single interaction. Interestingly, deletion of its homolog in yeast (*bmh2*Δ stain) had the lowest effect on the efficiency of dsDNA break repair among the 4 novel positives.

Discussion

Through computational acceleration of our PPI prediction algorithm based on conserved interaction codes, we have performed the first genome-wide PPI analysis for *S. pombe* and *C. elegans*. The new computational abilities enabled us to demonstrate the conservation of interaction codes among multiple species and that prediction of PPIs in an uncharacterized proteome can be performed based on known PPIs from other species. Furthermore, the human dsDNA interactome was considerably expanded through our method. This analysis led to the assignment of novel protein functions which was confirmed experimentally for four proteins.

The current algorithm works on the basis of available PPI data. In addition to being sparse, the available PPI data contain numerous false positives and hence are considered "noisy" data. One may expect that the availability of additional high confidence PPI data may help increase the performance of the current approach. Advances in protein 3D structures can be related to the growing number of structures for interacting complexes to predict novel binding partners. Growing databases of computationally predicted PPIs can reveal new information about novel properties of interacting partners. In future, some of these properties may be applied to the current algorithm to reduce false positives and hence increase specificity. GO terms can also be applied to eliminate false positives. These improvements however, may come in expense of novel predictions since they can discriminate against uncharacterized proteins.

The origin of the interaction codes used here for PPI prediction is not clear. They seem to be present in both eukaryotes as well as in *E. coli*, a prokaryote. However, the fact that eukaryotic codes cannot accurately predict *E. coli* PPIs may suggest a difference between pro and eukaryotic codes. A possible explanation for the origin of these short polypeptide regions is that they may have evolved from longer interaction mediating domains in order to maximize sequence usage. These codes represent alternative polypeptide signals that merit further attention.

The ability to investigate PPIs at a proteome level sets path to better study the topology of different PPI networks. In this context, "centrality" of a network can be better studied by examining hub and betweenness centralities[21,22]. These values can indicate proteins thought to be more relevant to the integrity of a network. They can provide significant clues for disease progression. In this way, important and novel drug targets can be elucidated[23,24].

In addition to the applicability of the new algorithm to study arbitrary species, the new algorithm is also considerably faster. The increased speed associated with the current algorithm however, is not sufficient for studying human genome wide PPI analysis. Using Massively Parallel (MP) computing approaches, we are currently in the process of generating a new method amenable to studying human PPI network.

Methods

PPI prediction via interaction codes. In brief, to predict whether two query proteins are likely to interact, our method examines sliding windows of primary sequences to determine if both query proteins share similarity with pairs of proteins that have been previously reported to interact. For more details, please see[1,7]. To measure sequence similarity, the PAM120 substitution matrix is used. The method requires a similarity threshold applied to the PAM120 score to determine whether sequence windows are, in fact, similar. This threshold must be tuned for each organism, as described below. Once we have completed this analysis over all possible sequence windows in each query proteins, a decision must be made whether there is sufficient evidence to support the predicted interaction. A second score threshold is then applied to make the decision whether the interaction should be predicted or not. As discussed below, this second threshold is tunable to achieve the required level of specificity or precision.

PPI prediction adaptation algorithm for other organisms. Amino acid distributions within proteins differ significantly between different organisms; see Supplementary Figure S8. Therefore, it is not optimal to run the same PPI prediction method for different organisms. Supplementary Figure S9 shows the difference between simply applying the unmodified PPI prediction method using yeast parameters to human proteins and the adapted human PPI prediction method using the organism-specific PPI prediction adaptation algorithm described in the remainder of this section.

Since amino acid sequence window comparison operations in our software are based on the PAM120 substitution matrix, a different amino acid distribution will cause a shift in the expected window score when two random sequences are compared. We plot the probability of scores when comparing two fragments of length 20 and set our threshold such that two sequence windows are declared to be 'similar' only if their PAM120 score is significantly above that expected by chance (i.e. there remains only a probability of 10^{-6} of obtaining this score by comparing random fragments). We have previously determined[7] this cutoff to be 35 in *S. cerevisiae* and that remains a valid cutoff for most of the other organisms tested except one: *H. sapiens*. For *H. sapiens* the expected score when comparing fragments increased, and a new cutoff of 40 was used exclusively for this organism. Changing the fragment window length (20 AA) and substitution matrix did not offer any significant improvement and they were therefore not modified.

For any new organism, the steps for automatically tuning the window matching score used by our PPI prediction method, precomputation of similar windows, evaluating the method's performance and selecting an operating point are as follows:

Step 1: Window matching score tuning

- (a) Part of the input is all the protein sequences for the organism (usually from an online source). We can limit the proteins to those involved in known interactions; however the entire proteome is preferred.
- (b) Our method then calculates the amino acid distribution for the organism. This is used to evaluate the expected score for comparing two random fragments of length 20 using this distribution. A score is picked to represent a 10^{-6} chance that two windows of length 20 are matched at random.

Step 2: Precomputation of similar windows

- (a) Once the correct window matching score setting have been determined, the database pre-computation is run. That is, the similarity between each unique 20 AA window in the proteome is precomputed and cached.

Step 3: Performance evaluation and picking an operating point

- (a) The second part of the input is a set of known interactions for the organism. This can be from one experiment, or from a repository such as BioGRID[15]. It can also be a gold-standard set or the union of all the databases online for example. Only physical interactions should be used since this is what the algorithm is meant to predict.
- (b) The method then needs to evaluate the sensitivity and specificity of the predictions for this organism. This is done by LOO cross-validation using the known interaction set collected earlier. The negative set is generated from 100,000 random pairs[16]. Plotting the ROC and recall vs. precision curves for the LOO cross validation experiment will indicate what cutoff to use in order to achieve a given specificity, precision, or sensitivity. For high-throughput experiments, operating at a high specificity is strongly suggested in order to avoid a large number of expected false-positives and maximize the precision of the classifier.

Step 4: PPI predictions

- (a) After completing the automatic steps listed above, the method is ready to start predicting new PPIs for that organism. For small genomes this can be done by an all-to-all experiment but for large genomes it is often preferable to only run pairs of interest due to the runtime involved.

Improved speed of PPI prediction using parallel multi-core processing. An ongoing concern with PPI prediction is performance. For example, for a single human

protein pair such as O43345 and Q05481, sequential PPI prediction took 1894.04 seconds on a single processor core. We therefore used fine-grained parallelism to accelerate our method using the Intel Cilk Plus library. For the same protein pair, our new, parallel multi-core PPI prediction method took only 321.82 seconds on a quad core Core i7 860 processor. This represents a 5.9 fold speed improvement.

Characterizing sensitivity, specificity, and precision of PPI predictions via in silico experiments. To determine the sensitivity and specificity of our PPI predictions we conducted LOO cross-validation experiments by first obtaining interactions databases (Table 1) to use as positive sets. For each individual organism, a negative set composed of 100,000 randomly chosen[16] pairs was created. The LOO experiments were conducted as per Step 5 of the "Sensitivity and Specificity Measure for Other Organisms" section discussed previously. The results are plotted as a ROC curve representing sensitivity against 1-specificity (Figures 1–3). While the ROC curve displays the entire range of specificity (0–100%), high-specificity values are of greater interest due to the relatively low number of true-positives expected for PPI maps, which can easily be outnumbered by false-positives unless the method operates at high specificity. Prediction confidence can also be measured using precision, which measures the proportion of positive predictions which are likely to be true positive interactions. The problem with precision is that it depends directly on the actual ratio of positive-to-negative protein pairs, and this is typically unknown. Previous studies have fixed this ratio at very low values (e.g. Yu et al explored ratios as high as 1 : 15, but admit that true ratios are likely on the order of 1 : 1000 for some species). For any given ratio Positive(*P*):Negative(*N*), the precision can be directly computed from sensitivity and specificity (which are unaffected by this ratio) using the following equation:

$$\text{Precision(Prec.)} = (\text{Sens.}\cdot\alpha)/((\text{Sens.}\cdot\alpha)+(1-\text{Spec.})\cdot(1-\alpha)) \quad \text{where } \alpha = \text{P}/(\text{N}+\text{P})$$

The precision of our method for each species is illustrated for a range of ratio values in Supplementary Figure S1.

Sensitivity and specificity measure for cross-species predictions. To determine if PPIs in a target organism can be predicted from cross-species PPIs using the union of multiple species, all known interactions from our select organisms (*C. elegans, E. coli, H. sapiens, C. cerevisiae* and *S. pombe*) are used except the interaction from the organism used for testing. For example to test our prediction in *H. sapiens* we would use all known interactions for the organisms above except interactions for *H. sapiens*. This simulates predictions in a new organism or for one which has few known interactions.

Yeast manipulations. The collection of yeast gene knockouts is described in[17]. Plasmid repair analysis was performed as before[18].

1. Pitre, S. *et al.* Global investigation of protein-protein interactions in yeast Saccharomyces cerevisiae using re-occurring short polypeptide sequences. *Nucleic Acids Res* **36**, 4286–4294 (2008).
2. Pitre, S. *et al.* (2008) Computational methods for predicting protein-protein interactions. Seitz, H (ed), Advances in Biochemical Engineering/Biotechnology, Springer-Verlag.
3. Zaki, N., Lazarova-Molnar, S., El-Hajj, W. & Campbell, P. Protein-protein interaction based on pairwise similarity. *BMC Bioinformatics* **10**, 150 (2009).
4. Guo, Y., Yu, L., Wen, Z. & Li, M. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res* **36**, 3025–3030 (2008).
5. Guo, Y. *et al.* PRED_PPI: a server for predicting protein-protein interactions based on sequence data with probability assignment. *BMC Res Notes* **3**, 145 (2010).
6. Martin, S., Roe, D. & Faulon, J. L. Predicting protein-protein interactions using signature products. *Bioinformatics* **21**, 218–226 (2005).
7. Pitre, S. *et al.* PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics* **7**, 365 (2006).
8. Shen, J. *et al.* Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci USA* **104**, 4337–4341 (2007).
9. Yu, C. Y., Chou, L. C. & Chang, D. T. Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinformatics* **11**, 167 (2010).
10. Betel, D. *et al.* Structure-templated predictions of novel protein interactions from sequence information. *PLoS Comput Biol* **3**, 1783–1789 (2007).
11. Neduva, V. *et al.* Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol* **3**, e405 (2005).
12. Park, Y. Critical assessment of sequence-based protein-protein interaction prediction methods that do not require homologous protein sequences. *BMC Bioinformatics* **10**, 419 (2009).
13. Yoko-o, T., Roy, S. K. & Jigami, Y. Differences in in vivo acceptor specificity of two galactosyltransferases, the gmh3+ and gma12+ gene products from Schizosaccharomyces pombe. *Eur J Biochem* **257**, 630–637 (1998).
14. Jain, E. *et al.* Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* **10**, 136 (2009).
15. Stark, C. *et al.* The BioGRID interaction database: 2011 update. *Nucleic Acids Res* **39**, D698–704 (2011).
16. Ben-Hur, A. & Noble, W. S. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics* **7**, Suppl 1:S2 (2006).
17. Winzeler, E. A. *et al.* Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
18. Jessulat, M. *et al.* Interacting proteins Rtt109 and Vps75 affect the efficiency of non-homologous end-joining in Saccharomyces cerevisiae. *Arch Biochem Biophys* **469**, 157–164 (2007).
19. Andres Leon, E., Ezkurdia, I., Garcia, B., Valencia, A. & Juan, D. EcID. A database for the inference of functional interactions in E. coli. *Nucleic Acids Res* **37**, D629–635 (2009).
20. Prasad, T. S. K. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**, D767–772 (2009).
21. Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
22. Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* **3**, e59 (2007).
23. Jessulat, M. *et al.* Recent advances in protein-protein interaction prediction: experimental and computational methods. *Expert Opinion on Drug Discovery* **9**, 921–935 (2011).
24. Nussinov, R., Tsai, C. J. & Csermely, P. Allo-network drugs: harnessing allostery in cellular networks. *Trends Pharmacol Sci* **32**, 686–93 (2011).

Acknowledgements

Our sincere thanks go to Dr. Y. Park for sharing his data with us. This research was supported by the Natural Science and Engineering Research Council (NSERC) of Canada. This work is dedicated to Minoo Golshani who dedicated her life to helping her community and touched everyone's heart on the way.

Authors' contributions

SP, MH, JG, FD and AG contributed to the conceptual development of the manuscript. SP and AS contributed to the implementation of the computational tool. SP, AS, JG, FD and AG contributed to computational data analysis. MH, BS, and MJ contributed to biological data collection. MH, BS, MJ and AG contributed to biological data analysis and interpretation of the data. All authors read and approved the final manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Pitre, S. *et al.* Short Co-occurring Polypeptide Regions Can Predict Global Protein Interaction Maps. *Sci. Rep.* **2**, 239; DOI:10.1038/srep00239 (2012).