BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 14 2008, pages 1625–1631
doi:10.1093/bioinformatics/btn249

Systems biology

InteroPORC: automated inference of highly conserved protein interaction networks

Magali Michaut[1,2,*], Samuel Kerrien[2], Luisa Montecchi-Palazzi[2], Franck Chauvat[1], Corinne Cassier-Chauvat[1,3], Jean-Christophe Aude[1], Pierre Legrain[1] and Henning Hermjakob[2]

[1]CEA, IBITECS, Gif sur Yvette, F-91191, France, [2]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [3]CNRS, URA 2096, Gif sur Yvette, F-91191, France

Received on February 18, 2008; revised and accepted on May 26, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Protein–protein interaction networks provide insights into the relationships between the proteins of an organism thereby contributing to a better understanding of cellular processes. Nevertheless, large-scale interaction networks are available for only a few model organisms. Thus, interologs are useful for a systematic transfer of protein interaction networks between organisms. However, no standard tool is available so far for that purpose.

Results: In this study, we present an automated prediction tool developed for all sequenced genomes available in Integr8. We also have developed a second method to predict protein–protein interactions in the widely used cyanobacterium *Synechocystis*. Using these methods, we have constructed a new network of 8783 inferred interactions for *Synechocystis*.

Availability: InteroPORC is open-source, downloadable and usable through a web interface at http://biodev.extra.cea.fr/interoporc/

Contact: michaut.bioinfo@gmail.com

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

1 INTRODUCTION

Biological organisms live in complex interaction with their constantly fluctuating environment. Changes in regulatory networks have been observed in a number of organisms when they are under specific conditions or stressed by some environmental alterations, leading to modifications of their metabolism. The understanding of these phenomena depends not only on the knowledge of the numerous molecular effectors involved such as genes and proteins but also on the understanding of the functional relationships between them.

Experimental approaches used to decipher protein–protein interaction (PPI) networks are described in Shoemaker and Panchenko (2007a). To complement these experimental techniques, a number of computational methods have been developed to predict PPIs (Shoemaker and Panchenko, 2007b). Large-scale PPI networks are only available for a limited number of model organisms, thus systematic inference of PPIs has become a central task of functional genomics. Consequently, we have investigated network inference using the interolog concept originally introduced by Walhout *et al.* (2000) which combines known PPIs from one or more source species and orthology relationships between the source and target species to predict PPIs in the target species.

Since the original introduction of the interolog concept, such interactome transfers have been performed for different species and using different ortholog identification methods. Matthews *et al.* (2001) have transferred two large-scale, two-hybrid interaction maps of *Saccharomyces cerevisiae* onto *Caenorhabditis elegans*. PPI maps have been constructed for various organisms (Yu *et al.*, 2004) based on the *S.cerevisiae* interactome. Based on the InParanoid (Remm *et al.*, 2001) algorithm to identify orthologs, human networks have been inferred from several model organisms (Huang *et al.*, 2004, 2007; Lehner and Fraser, 2004; Persico *et al.*, 2005). Brown and Jurisica (2005) have developed the web-based database OPHID containing human PPIs using BLASTP and the reciprocal best hit approach (Jordan *et al.*, 2002). Maps have also been generated for *Plasmodium falciparum* (Wuchty and Ipsaro, 2007) or *Helicobacter pylori* (Wojcik *et al.*, 2002).

Such transfers have been done only for a limited number of species and no standard method or software seems to emerge. Each study was based on a combination of selected species and orthology computation methods. Yet a common tool usable for a large number of species would greatly facilitate comparative studies, leading to a better understanding of the extent of evolutionary conservation of PPI networks. Such a method would be of great help to decipher PPI networks in the wealth of organisms with a newly sequenced genome or still lacking identified PPIs. It usually takes several years to carry out genome-wide detection of PPIs. Consequently, we have developed an automated tool, interoPORC, to predict PPIs for all organisms present in the Integr8 database (Kersey *et al.*, 2005). This database systematically provides all deciphered genomes and their corresponding proteomes (655 organisms in release 75).

Through a multidisciplinary approach, we have investigated the biological responses to environmental stresses using the model cyanobacterium *Synechocystis* PCC6803. Cyanobacteria are the most abundant photosynthetic organisms on Earth and their living conditions are frequently challenged by changes in nutrient

[*] To whom correspondence should be addressed.

availability and exposure to pollutants. *Synechocystis* is a unicellular prokaryote with a small fully sequenced genome (3600 genes) (Kaneko *et al*., 1996) easily manipulable with replicating plasmids (Domain *et al*., 2004; Mazouni *et al*., 2004). It shares a wealth of homologous proteins with plants. Thus, lessons learned from stress responses in *Synechocystis* should greatly facilitate the understanding of how plants face environmental challenges. We used our interoPORC prediction tool based on orthologous protein clusters to predict PPIs for *Synechocystis*. In addition, we developed a second prediction method which was more flexible but required more computational resources. This method, called interoBH, was based on pairwise sequence comparisons. It was also applied to *Synechocystis*, starting with a limited set of source species. We selected several model organisms whose interactomes have already been investigated, namely *S.cerevisiae*, *Escherichia coli*, *Homo sapiens*, *Arabidopsis thaliana*, *C.elegans*, *Drosophila melanogaster* and *H.pylori*, which are representative of the overall biodiversity of living organisms. The use of both methods enabled us to construct a new network of 8783 PPIs for *Synechocystis*.

2 METHODS

2.1 Data sources

All selected genomes and proteomes were collected from Integr8 (Kersey *et al*., 2005). For *E.coli* and *H.pylori*, all the proteomes from the various sequenced strains were merged to generate a global proteome for each single species because PPIs are sometimes reported at the species rather than the strain taxonomic level. Furthermore, for proteins having multiple splice variants, we have only considered the longest product of the genes encoding them. The experimental PPI datasets from the three manually curated databases DIP (February 19, 2007) (Xenarios *et al*., 2002), IntAct (April 13, 2007) (Kerrien *et al*., 2007a) and MINT (April 5, 2007) (Chatr-Aryamontri *et al*., 2007), which provide tabular data files in the MITAB25 tabular format (Kerrien *et al*., 2007b), were downloaded. We merged all PPIs and removed duplications. Finally, we extracted physical interactions occurring in each of the seven species selected, without considering self-interactions. A total of 139 325 PPIs were included in our investigation (Table 1). We collected sequence similarities from the CluSTr database (Petryszak *et al*., 2005). The PORC orthology data were available from Integr8 (Kersey *et al*., 2005). The functional annotation of *Synechocystis* proteins is described in the GOA (Camon *et al*., 2004) file available from Integr8.

2.2 Prediction methods

For both prediction methods, we used the sequence similarity to identify putative orthologous proteins between species. Based on the interolog concept, we combined interaction datasets with orthology information to transfer PPIs from different species onto *Synechocystis*.

2.2.1 InteroBH We considered homologous proteins as putative orthologs between *Synechocystis* and each source organism. This method was called interoBH since it was based on a best hit approach. Homology predictions were derived from pairwise Smith–Waterman similarities with an *E*-value for each sequence comparison (Saebo *et al*., 2005). To select the best sequence homologies, sequence comparisons with an *E*-value less than 1E-10, a standard cutoff value, were considered (Martin *et al*., 2002; Yu *et al*., 2004). For each protein, we selected in each of the other organisms the best matching sequence as a homolog. In addition, if the former protein was the best matching sequence of another protein in the same species, we added the latter as another homolog. In this way, we modified the reciprocal best-hit approach in such a way that a given protein could have several homologs, considered as putative orthologs, in a single organism.

Table 1. Source interactions

Relevant organisms	*Proteins*	*Interactions*
S.cerevisiae	5780	54560
A.thaliana	758	1406
E.coli	3853	22023
H.sapiens	9234	26587
D.melanogaster	8636	27476
C.elegans	3275	5636
H.pylori	783	1637

For each organism, the number of interacting proteins and the number of PPIs collected in the databases IntAct, MINT and DIP are indicated.

We then investigated each species separately. Let us take a transfer of *S.cerevisiae* interactome onto *Synechocystis* as an example. For each binary interaction occurring in yeast, we considered each interacting protein and looked for putative orthologs in *Synechocystis*. If both interacting proteins had a putative ortholog in *Synechocystis*, we transferred the PPI to these two putative orthologs. If a protein had several putative orthologs in *Synechocystis*, then we predicted all possible PPIs as putative PPIs. We used the joint *E*-value (Yu *et al*., 2004) to assess the quality of the predicted PPIs. The joint *E*-value was defined as the geometric mean of the individual *E*-values of both putative orthologs.

2.2.2 InteroPORC InteroBH was generalized leading to a new method called interoPORC since it was based on the new PORC data (putative orthologous clusters) defined as orthologous families from Integr8. These clusters are of paramount interest since, unlike previously defined clusters (Koonin *et al*., 1998), they contain all sequenced organisms (556 bacteria, 59 eukaryota and 50 archaea in the release 75). Each entry in PORC represents a cluster of genes grouped by the similarity of their longest protein product. We used 215 733 clusters, containing 1 548 235 proteins. According to the PORC construction process, a cluster contains at most a single protein from a given species and a protein can be assigned only to a single cluster. In other words, it is impossible to find several proteins of the same species in a single cluster. When sequence comparisons were insufficient, the PORC algorithm did not attempt to resolve potential ambiguity using phylogenetic trees or network comparison (Bandyopadhyay *et al*., 2006). The clusters were split according to the weakest sequence similarity in order to respect the ‘one gene per species per cluster’ rule (Kersey *et al*., 2005).

The inference process was similar to that of interoBH, orthologous groups of proteins were used instead of binary orthology relationships between two species. The process was broken down into two steps. In the first step, called up-casting, we abstracted PPIs onto orthologous cluster links. For a given source PPI, if both proteins belonged to a cluster, we constructed a link between these two clusters. In the second step, called down-casting, we projected these cluster links onto target species to predict new PPIs. Practically, for a given link between two clusters, if both clusters contained a protein from the target species, we inferred a PPI between these proteins unless this PPI had been used as a source PPI to construct the cluster link.

2.3 Supporting evidence analysis

In order to support some of the predicted PPIs, we explored the following approaches: (i) PPI explanation on the basis of interacting domain annotation; (ii) sharing of functional annotations for both interacting proteins; (iii) prediction by several species; (iv) identification of source PPI by several experimental techniques; (v) comparison with experimentally identified interactions.

2.3.1 Domain–domain interaction score Interacting domain annotation was used to identify the predicted PPIs associated with domain pairs indicative of true interactions. We retrieved the Pfam domain composition of all proteins of *Synechocystis* from UniprotKB (Bairoch *et al.*, 2005). Then we collected a list of domain–domain interactions (DDI) derived from iPFAM structures (Finn *et al*., 2005). We combined both types of information to compute a list of *Synechocystis* proteins that potentially interact together based on DDIs. Since we wanted to favor domain pairs that rarely occur in all protein pairs, we defined a score S_P, calculated using Equation (1). It is noteworthy that 78% of *Synechocystis* proteins were annotated with a single domain. Nevertheless, when a protein had several domains, we calculated the PPI score as the maximum score of all possible domain pairs.

$$S_p = \underset{d \in D}{\mathrm{Max}} \left(\frac{1}{\mathrm{count}(d)} \right) \tag{1}$$

D is the set of domain pairs constructed with the domain lists of both proteins of the PPI, *p* and count(*d*) is the number of occurrences of this domain pair *d* in all protein pairs of *Synechocystis*. We generated a set of 5000 random PPIs (Supplementary Material). Given that 95% of the scores were below 0.5, we considered all scores above this threshold as highly relevant.

2.3.2 Common Gene Ontology annotation Since interacting proteins either share similar functions or operate in the same biological process (Huang *et al*., 2007), we considered both molecular function (MF) and biological process (BP) ontologies of the Gene Ontology (GO) (Ashburner *et al.*, 2000) and calculated semantic similarities between both proteins using the measure defined in Lubovac *et al*. (2006). We generated a set of 5000 random PPIs. Since 95% of random PPIs had a similarity below 0.23 for both MF and BP ontologies (Supplementary Material), we used this cutoff to identify PPIs with high semantic similarities.

2.3.3 Conserved interologs Some interactions were predicted several times from different source species using the interoBH method, which led us to think that they were meaningful (Lehner and Fraser, 2004). Similarly, different source interactions enabled us to construct a unique link between two clusters during the interoPORC process. The PPIs predicted from several source species were thus more likely to be valid. We therefore extracted the interactions predicted by several organisms.

2.3.4 Multiple experimental identification methods We examined the different kinds of experimental techniques which have been used to identify each source interaction. All experimental identification methods have different weaknesses and biases (Hakes *et al*., 2008; von Mering *et al*., 2002). Nevertheless, when an interaction has been detected by different methods, it is more likely to be genuine. In the MITAB25 format, a detection method is associated with each interaction using the PSI controlled vocabularies (Kerrien *et al*., 2007b). We defined groups of methods as all children terms of the following controlled vocabulary terms: MI:0401 (biochemical), MI:0090 (Y2H), MI:0013 (biophysical), MI:0428 (imaging), MI:0254 (genetic), MI:0255 (transcription), MI:0063 (prediction), MI:0362 (inference), MI:0686 (unspecified). The list of all terms and their associated group is available in Supplementary Material file 1. Some of them do not appear in the source interaction data for the specific source (e.g. unspecified, transcription, genetic).

2.3.5 Comparison with experimental data On the one hand, we identified the predicted interactions that were present in the source dataset constructed from IntAct, MINT or DIP from low-throughput experiments. On the other hand, we analyzed all predicted interactions that overlap with the large-scale study recently published (Sato *et al*., 2007).

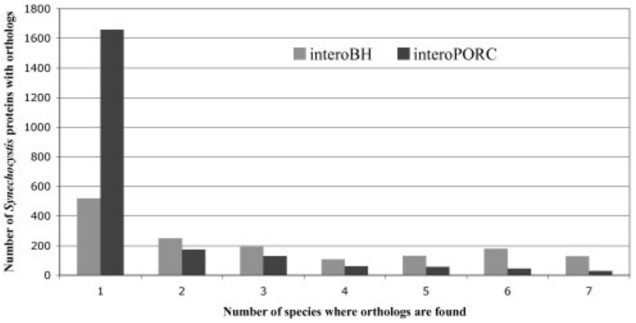

Fig. 1. Conservation of orthologous proteins in the seven selected species as predicted by each method.

3 RESULTS

To construct a PPI network for *Synechocystis*, we have developed two prediction methods, which combined known interactions from source species with putative orthology relationships. These two methods mainly differed in the way orthology relationships were constructed (see Methods), using either a pairwise approach based on best hits, interoBH or putative orthologous clusters, interoPORC. After showing some results about the orthology calculation, we will present the results of both prediction and interaction analysis methods, to end with the tool developed.

3.1 Orthology calculation

For each *Synechocystis* protein, putative orthologs were identified in the seven other species selected as reference organisms. We noted that the interoPORC method predicted more proteins with putative orthologs in only one species compared to the interoBH approach (Fig. 1). A smaller number of proteins were found to have putative orthologs across several species using interoPORC. This can be explained by the fact that the selected species are evolutionary distant not only from *Synechocystis* but also from each other. Thus only highly conserved proteins were found in a cluster with several of the selected species.

3.2 Interactions derived with InteroBH

We combined interaction dataset and orthology information to transfer interactions from seven source species onto *Synechocystis* separately (Table 2). Combining all results, we obtained a global set of 8586 interactions among 998 proteins (28% of the proteome of *Synechocystis*). This network was called interoBH_LOW. To assess the quality of the resulting interologs, we used the joint *E*-value defined in Yu *et al*. (2004). Since all sequence comparisons considered had an *E*-value less than the standard cutoff value of 1E-10, the joint *E*-value of each interolog was greater than 1E-10. Furthermore, it has been shown that a threshold of 1E-70 for the joint *E*-value enables a transfer of interactions with greater confidence (Yu *et al*., 2004). Thus we considered all interologs with a joint *E*-value less than 1E-70 as a specific dataset called interoBH_HIGH. It should be noted that the orthology construction process used in the interoPORC method only considered sequence comparisons with a joint *E*-value less than 1E-40. Another dataset called interoBH_MEDIUM was considered for all interactions with a joint *E*-value less than this value.

Table 2. Number of PPIs predicted in *Synechocystis* by interoBH

	interoBH_HIGH		interoBH_MEDIUM		interoBH_LOW	
Source species	Inter	Prot	Inter	Prot	Inter	Prot
S.cerevisiae	955	299	1826	360	3558	438
A.thaliana	0	0	5	7	10	11
E.coli	1775	61	3183	744	4894	825
H.sapiens	26	26	69	74	194	150
D.melanogaster	14	16	30	37	97	95
C.elegans	1	2	3	6	21	35
H.pylori	199	75	164	117	251	160
Total	2870	1031	5280	1345	9025	1714
Non-redundant	2748	741	5070	884	8586	998

For each source species, the number of predicted interactions (Inter) and the number of proteins (Prot) involved in these predicted interactions are indicated. The Total line indicates the sum of all line values, whereas the Non-redundant line indicates the numbers of distinct interactions or proteins.

Not surprisingly, the number of predicted interactions was highly dependent on the number of available interactions in the source organism (Tables 1 and 2). It was also dependent on the evolutionary proximity to *Synechocystis*. Indeed, with almost the same number of source interactions, we transferred many more PPIs between the bacterium *E.coli* and the cyanobacterium *Synechocystis* than between *H.sapiens* and *Synechocystis*. It confirmed the recent result of (Brown and Jurisica, 2007) who showed that the number of interactions predicted by the interolog concept depends on the evolutionary distance between the organisms studied.

3.3 Interactions derived with InteroPORC

Using the interoPORC method, we predicted a dataset of 1446 interactions between 384 proteins in *Synechocystis*. In some cases, different source PPIs have been used to construct a single link between two clusters. In such a case, the predicted PPI was inferred from several source species.

3.4 Supporting evidence

In order to support some of the predicted interactions, we explored different methods based on interacting domain annotation, functional annotation, conservation across organisms, experimental techniques and experimentally identified interactions (see Methods).

3.4.1 Interacting domain annotation Within the union of interoPORC and interoBH_LOW, 177 interacting proteins shared a pair of domains from the set of known domain interactions. This set of PPIs included 39 associated with DDIs that had a highly relevant score (see Methods) increasing our confidence in these predicted interactions (Table 3). Itzhaki *et al.* (2006) have shown that DDIs frequently occur in protein complexes and are evolutionary conserved. Indeed, we observed some interconnected subgraphs representing complexes such as the RNA polymerase or the ATP synthase (Supplementary Fig. 1). Furthermore, Itzhaki *et al.* found that the number of PPIs explained by DDIs in the different PPI networks ranged from 6% to 20% only. Consequently, PPIs supported by DDIs are strengthened but PPIs not supported by DDIs are not necessarily weakened.

Table 3. Predicted PPIs associated with DDIs

Prediction sets	H	M	L	P
Total PPIs	2748	5070	8586	1446
PPIs with known domains	2689	4939	8197	1399
PPIs associated with DDI(s)	60	100	172	37
PPIs with highly relevant score	18	27	38	16

The DDI annotation is described for H: interoBH_HIGH, M: interoBH_MEDIUM, L: interoBH_LOW and P: interoPORC in terms of number of PPIs.

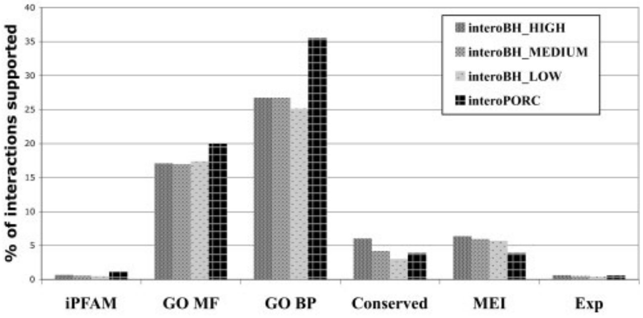

Fig. 2. Percentages of PPIs supported by different methods. iPFAM: PPIs explained by DDIs; GO MF (GO BP): interactions with similar terms in the MF (BP) ontology of GO; Conserved: interactions predicted by different source species; MEI: PPIs predicted from source PPIs identified by multiple experimental identification techniques; Exp: PPIs experimentally identified.

3.4.2 Functional annotation For the MF ontology, all networks derived with interoBH contained 17% of the interactions with similar annotation between their interacting proteins. The interoPORC network resulted in a slightly higher percentage (Fig. 2). For the BP ontology, interoBH networks resulted in about 26% of interactions with a similar annotation compared to 35% with the interoPORC network.

3.4.3 Conserved interologs The different datasets contained about 5% of conserved interologs (Fig. 2). We identified more conserved interologs in the interoBH_HIGH network. This was consistent with the fact that this network had been defined using a more stringent sequence comparison cutoff. There were seven highly conserved interactions transferred from three and four organisms. All 258 conserved interactions are represented in Supplementary Figure 2 between 167 proteins, including two highly connected chaperone proteins. This stressed the fact that interactions with a chaperone are detected with high-throughput identification techniques. This could be due to chaperone function that is used to bind to a number of proteins to assist their folding. However, non-specific binding cannot be ruled out. We noticed that 75% of all partners of the first chaperone groL1 (slr2076) do not share any GO term with it. These interactions were transferred from high-throughput detection methods such as two-hybrid or co-immunoprecipitation. We also examined existing data on the other highly connected chaperone protein dnaK2 (sll0170), and found that these predicted interactions were derived from interactions detected by different assays such as X-ray crystallography, molecular sieving, blue native

PAGE (polyacrylamide gel electrophoresis) or enzyme linked immunosorbent assay. This is consistent with the much higher rate of interacting partners sharing GO terms (40%).

3.4.4 Multiple experimental identification methods A total of 491 PPIs were transferred from source interactions identified by different experimental methods. The interoBH networks had 6% of interactions coming from several methods whereas the interoPORC network had 4% of such interactions (Fig. 2).

3.4.5 Comparison with experimental data Among all predicted interactions, 10 were among the 185 *Synechocystis* PPIs reported in the experimental datasets obtained from IntAct, MINT and DIP. To evaluate the significance of this overlap, we computed the probability of finding randomly an overlap greater than the one observed. We found according to a hypergeometric model that the probability was less than 1E-4 (Supplementary Material). Thus, the experimental results corroborated our predictions.

A further experimental study led to a new large-scale dataset of 3236 interactions between 1920 proteins (Sato *et al*., 2007). When we considered only the proteins included in this study, we had 3904 predicted PPIs instead of the 8783 PPIs predicted by interoPORC or interoBH_LOW. Among this predicted subset, Sato *et al.* identified 25 interactions, which was significant (P-value <1E-18). It is important to note that large-scale experimental datasets obtained with the same technique have an overlap smaller than 10% of the total number of interactions (Arifuzzaman *et al*., 2006), emphasizing the high false negative rate. We are currently investigating this comparison more in depth. Together, 35 predicted interactions have been experimentally identified (Supplementary Fig. 3).

Among the 8783 PPIs predicted by interoPORC or interoBH_LOW, we identified a core set of 3495 interactions supported either by interacting domain annotation, functional annotation, conservation across species, multiple experimental techniques or experimental identification (Supplementary Material file 2).

3.5 A tool of use for all sequenced genomes

Since the quality of the predictions depends on the quality of the source data, it was important to separate the prediction process and the source data used. We developed a stand-alone tool that can be applied to different source data, for example to high-quality PPIs and private datasets. In addition, interoPORC can be run on all platforms since it has been developed in Java (the source code is also available). Moreover, we have provided result files in standard formats (PSI25-XML and MITAB25) in order to interface easily with existing tools.

We also wanted to provide a tool that was fast and easy to use. Consequently, we have set up a web interface where predictions can be run just by giving a species identifier. We use source PPIs from IntAct, MINT and DIP as well as PORC data from Integr8. All source data are updated as soon as a new version is publicly available for any database.

As an illustration, we applied interoPORC to several representative organisms (Table 4). For example, we predicted 1678 interactions in the widely studied cyanobacterium *Anabaena* PCC7120, a model organism without any large-scale interaction dataset so far. All predicted PPIs are available on the web interface. First, we noted that we obtained 1% more PPIs for *Synechocystis* (1463 instead of 1446) than with the previous interoPORC prediction involving only the seven species with the largest PPI networks being involved (Table 3). Furthermore, the number of predicted PPIs was higher for eukaryota as compared to both archeae and bacteria. This can be due to the larger size of their proteome but also to the smaller evolutionary distance between source and target organisms since 83% of the source PPIs used occur in eukaryota (data not shown). This corroborates the results of Brown and Jurisica (2007) showing that the number of predicted PPIs decreases as the evolutionary distance increases.

Table 4. New PPIs for several organisms representative of the biodiversity predicted with interoPORC

Superkingdom	Species	*Taxid*	*Proteome*	*Curated*	*New*
Archaea	*P.kodakaraensis*	69014	2301	0	221
Archaea	*T.volcanium*	273116	1523	0	208
Eukaryota	*R.norvegicus*	10116	12028	2178	13469
Eukaryota	*A.fumigatus*	330879	9629	0	17225
Eukaryota	*P.falciparum*	36329	5283	2737	4026
Bacteria	*B.subtilis*	224308	4105	0	2160
Bacteria	*Synechocystis sp.*	1148	3506	185	1463
Bacteria	*Anabaena sp.*	103690	6070	1	1678

For each species, the superkingdom, the name (Species), the taxonomic identifier (taxid), the size of the proteome (Proteome), the number of PPIs in the source databases (Curated) and the number of new predicted PPIs (New) are indicated.

Consequently, interoPORC is of great interest for every organism with a newly sequenced genome for which no large-scale interactome has been determined yet. It provides a raw picture of possible PPIs, which can be experimentally validated. For most species, a global PPI dataset is yet to be determined, emphasizing the value of this tool that quickly and easily gives new insights into PPI networks in a large number of organisms.

3.6 Discussion

In this study we have developed two new prediction methods, interoPORC and interoBH to infer PPIs. They are based on the interolog concept, combining source PPIs in several species with orthology relationships. The interoPORC method can be used to predict PPIs in the ever-increasing number of organisms with a newly sequenced genome where large-scale analyses remained to be carried out. In these organisms, it is now possible to quickly get a raw picture of possible interactions using the open-source automated tool interoPORC which can be run through a web interface or downloaded for stand-alone use. Moreover, with the increasing availability of PPI networks, recent studies have shown the benefit of PPI network comparison across evolution (Kalaev *et al*., 2008). However, large PPI networks are available for only a few model organisms so far. Therefore, interoPORC is of great interest for constructing new networks, leading to improved comparative studies.

The interoBH datasets tended to contain the interoPORC dataset when the cutoff on the joint E-value decreased. The overlaps with the interoPORC amounted to 580 (40%), 1069 (74%) and 1249 (86%) interactions for interoBH_HIGH, interoBH_MEDIUM and interoBH_LOW, respectively. We noted that the two methods differed in the way orthology was calculated. Several putative orthologs were detected with the interoBH approach while only one protein was selected as a putative ortholog with the interoPORC method. To understand better the differences between the two approaches and assess to what extent the results were comparable, we investigated further the interoBH approach using the common reciprocal best-hit approach (Jordan *et al*., 2002), noted as interoRBH. Only one protein could be selected as a putative ortholog akin to the interoPORC approach. The proportion of interactions predicted only by the interoBH_MEDIUM method was highly reduced when considering interoRBH_MEDIUM, whereas the intersection between interoRBH_MEDIUM and interoPORC was only slightly reduced compared to interoBH_MEDIUM and interoPORC (data not shown). This confirmed that the additional interactions predicted by interoBH came from the choice to keep several putative orthologs. Moreover the interoPORC interactions have a lower joint E-value than the interoBH_LOW interactions (P-value <0.008, Supplementary Material). Consequently, the interoPORC method can be seen as a way to obtain a highly conserved interaction dataset.

It is worth noting that some interactions predicted by interoBH but not by interoRBH have been experimentally observed (Sato *et al*., 2007) and are thus relevant. Nevertheless, interoBH led to a higher number of predicted interactions than interoRBH. Thus we may expect a higher number of false positives as previously discussed in Yu *et al*. (2004). The interoPORC method proved to be a more stringent automated approach for all sequenced organisms. This raises the question of the tradeoff between the general and automatic nature versus the coverage and sensitivity of the different approaches. We propose here an automated tool of use for all species and we completed these stringent results with a more sensitive method for the particular species we were investigating.

The combined use of interoPORC and interoBH_LOW, enabled us to predict a global new network of 8783 PPIs for *Synechocystis* among which 3495 have been supported by different methods. Among these, 25 predicted PPIs have been identified in a new recently published large-scale dataset (Sato *et al*., 2007). Both experimental and computational approaches have weaknesses and miss lots of interactions. Thus it is highly interesting to have such computational methods at one's disposal to complete experimental datasets and identify interactions that may have escaped the experimental detection with high-throughput methods.

ACKNOWLEDGEMENTS

The authors thank Arnaud Martel for setting up the interoPORC web interface and Raphaël Guerois for insightful comments.

Funding: This work was funded by the European Commission (FELICS 021902 RII3) within the Research Infrastructure Action of the FP6 'Structuring the European Research Area' Programme (to S.K., L.M.-P., H.H.); French National Agency of Research (ANR Biosys06_134823 SULFIRHOM); European Commission, Marie Curie Fellowship (to M.M.); French Atomic Energy Commission grant (to M.M.).

Conflicts of interest: none declared.

REFERENCES

Arifuzzaman,M. *et al*. (2006) Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res.*, **16**, 686–691.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bairoch,A. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**(Database issue), D154–D159.

Bandyopadhyay,S. *et al*. (2006) Systematic identification of functional orthologs based on protein network comparison. *Genome Res.*, **16**, 428–435.

Brown,K.R. and Jurisica,I. (2005) Online predicted human interaction database. *Bioinformatics*, **21**, 2076–2082.

Brown,K.R. and Jurisica,I. (2007) Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.*, **8**, R95.

Camon,E. *et al*. (2004) The Gene Ontology Annotation (GOA) Database–an integrated resource of GO annotations to the UniProt Knowledgebase. *In Silico Biol.*, **4**, 5–6.

Chatr-Aryamontri,A. *et al*. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**(Database issue), D572–D574.

Domain,F. *et al*. (2004) Function and regulation of the cyanobacterial genes lexA, recA and ruvB: lexA is critical to the survival of cells facing inorganic carbon starvation. *Mol. Microbiol.*, **53**, 65–80.

Finn,R.D. *et al*. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.

Hakes,L. *et al*. (2008) Protein-protein interaction networks and biology-what's the connection? *Nat. Biotechnol.*, **26**, 69–72.

Huang,T.W. *et al*. (2004) POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics*, **20**, 3273–3276.

Huang,T.W. *et al*. (2007) Reconstruction of human protein interolog network using evolutionary conserved network. *BMC Bioinformatics*, **8**, 152.

Itzhaki,Z. *et al*. (2006) Evolutionary conservation of domain-domain interactions. *Genome Biol.*, **7**, R125.

Jordan,I.K. *et al*. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.*, **12**, 962–968.

Kalaev,M. *et al*. (2008) NetworkBLAST: comparative analysis of protein networks. *Bioinformatics*, **24**, 594–596.

Kaneko,T. *et al*. (1996) Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.*, **3**, 109–136.

Kerrien,S. *et al*. (2007a) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**(Database issue), D561–D565.

Kerrien,S. *et al*. (2007b) Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.*, **5**, 44.

Kersey,P. *et al*. (2005) Integr8 and genome reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res.*, **33**(Database issue), D297–D302.

Koonin,E.V. *et al*. (1998) Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.*, **8**, 355–363.

Lehner,B. and Fraser,A.G. (2004) A first-draft human protein-interaction map. *Genome Biol.*, **5**, R63.

Lubovac,Z. *et al*. (2006) Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins*, **64**, 948–959.

Martin,W. *et al*. (2002) Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl Acad. Sci. USA*, **99**, 12246–12251.

Matthews,L.R. *et al*. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or 'interologs'. *Genome Res.*, **11**, 2120–2126.

Mazouni,K. *et al*. (2004) Molecular analysis of the key cytokinetic components of cyanobacteria: FtsZ, ZipN and MinCDE. *Mol. Microbiol.*, **52**, 1145–1158.

Persico,M. *et al*. (2005) HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics*, **6** (Suppl. 4), S21.

Petryszak,R. *et al*. (2005) The predictive power of the CluSTr database. *Bioinformatics*, **21**, 3604–3609.

Remm,M. *et al*. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.

Saebo,P.E. *et al*. (2005) PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology. *Nucleic Acids Res.*, **33**(Web Server issue), W535–W539.

Sato,S. *et al*. (2007) A large-scale protein protein interaction analysis in Synechocystis sp. PCC6803. *DNA Res.*, **14**, 207–216.

Shoemaker,B.A. and Panchenko,A.R. (2007a) Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput. Biol.*, **3**, e42.

Shoemaker,B.A. and Panchenko,A.R. (2007b) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.*, **3**, e43.

von Mering,C. *et al*. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.

Walhout,A.J. *et al.* (2000) Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*, **287**, 116–122.

Wojcik,J. *et al*. (2002) Prediction, assessment and validation of protein interaction maps in bacteria. *J. Mol. Biol.*, **323**, 763–770.

Wuchty,S. and Ipsaro,J.J. (2007) A draft of protein interactions in the malaria parasite P. falciparum. *J. Proteome Res.*, **6**, 1461–1470.

Xenarios,I. *et al*. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.

Yu,H. *et al*. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.*, **14**, 1107–1118.