BMC Genomics

RESEARCH ARTICLE Open Access

Combining modularity, conservation, and interactions of proteins significantly increases precision and coverage of protein function prediction

Samira Jaeger[1*], Christine T Sers[2], Ulf Leser[1]

Abstract

Background: While the number of newly sequenced genomes and genes is constantly increasing, elucidation of their function still is a laborious and time-consuming task. This has led to the development of a wide range of methods for predicting protein functions in silico. We report on a new method that predicts function based on a combination of information about protein interactions, orthology, and the conservation of protein networks in different species.

Results: We show that aggregation of these independent sources of evidence leads to a drastic increase in number and quality of predictions when compared to baselines and other methods reported in the literature. For instance, our method generates more than 12,000 novel protein functions for human with an estimated precision of ~76%, among which are 7,500 new functional annotations for 1,973 human proteins that previously had zero or only one function annotated. We also verified our predictions on a set of genes that play an important role in colorectal cancer (*MLH1*, *PMS2*, *EPHB4*) and could confirm more than 73% of them based on evidence in the literature.

Conclusions: The combination of different methods into a single, comprehensive prediction method infers thousands of protein functions for every species included in the analysis at varying, yet always high levels of precision and very good coverage.

Background

Elucidating protein function is still one of the major challenges in the post-genomic era [1,2]. Even for the best-studied model organisms, such as yeast and fly, a substantial fraction of proteins is still uncharacterized [3]. As high-throughput techniques increase the availability of completely sequenced organisms, annotation of protein function becomes more and more a bottleneck in the progress of biomolecular sciences and the gap between available sequence data and functionally characterized proteins is still widening [2]. Manual annotation, using, for instance, the scientific literature, and experimental identification of protein function remains a difficult, time- and cost-intensive task [4]. Reliable methods for assigning functions to uncharacterized proteins are required to support and supplement these methods. There are various automatic approaches for the prediction of protein function. These use, for instance, protein sequences and 3D-structures [5-9], evolutionary relationships [10,11], phylogenetic profiles [12,13], domain structures [14], or functional linkages [15]. Another important class of information for function prediction are protein-protein interactions (PPIs). PPIs are a type of data that is close to the biological role of a protein within cells and therefore ideally suited to form the basis for function prediction methods [16,17]. Furthermore, more and more such data sets are becoming available (e.g. [18,19]). These data sets may be used to identify functional modules within protein networks [20], to find protein complexes [21], or to

* Correspondence: sjaeger@informatik.hu-berlin.de
[1]Knowledge Management in Bioinformatics, Humboldt-Universitat zu Berlin Unter den Linden 6, 10099 Berlin, Germany
Full list of author information is available at the end of the article

determine evolutionary conserved processes [22-25], all of which provide valuable clues to the function of a protein [3].

The approaches that use PPI for function prediction can be classified into two main classes:

1. Link-based methods predict novel functions for a protein by transferring known functions from directly or indirectly interacting proteins. This may be achieved by studying the set of neighbors [16,19,26,27], by considering the position of the protein within its neighborhood [28], or by looking at the position of the protein in the entire interaction network [29,30].
2. Module-based methods assign functions to proteins by first computing clusters (or modules) within the protein network [31]. Based on the hypothesis that cellular functions are organized in a highly modular manner [32,33], all members of a cluster are assigned annotations that are enriched within the module [23].

Both approaches have their benefits and their drawbacks. PPI-based prediction methods provide a better coverage but are sensitive to the high level of false-positives [34,35] and false negatives [36] in current PPI data sets. Module-based methods are more robust to missing or wrong interactions, but are able to predict function only within dense regions of a species network disregarding, for instance, chain-like pathways. This largely reduces their coverage [21,31]. Module-based methods have been shown to be less accurate than for example simple guilt-by-association approaches but their performance improves in networks with less functional coverage [37]. Furthermore, both methods in first place only work within a species, which disregards the wealth of information that might be available in evolutionary related other species (this is particularly true for humans). This limitation can be removed by using annotations of homologous proteins. However, purely homology-driven prediction strategies are rather imprecise [38]. Although prediction precision may be improved by using only orthology, the overall precision remains below that of most PPI-based methods [7].

In this paper, we describe a novel algorithm for protein function prediction that combines link-based and module-based prediction with orthology, thus overcoming the respective limitations of each individual approach. The key to our method is to analyze proteins within modules that are defined by evolutionary conserved processes. To this end, we first compute PPIs that are highly conserved within a given set of species. These so-called interologs [39] are assembled to highly conserved protein subnetworks. For a given protein, we then predict functions of other proteins in the same CCS using both directly interacting proteins as well as orthology relationships.

We apply our function prediction strategy to different sets of species, ranging from species pairs to groups of up to four species. We show that our approach reaches very high prediction precision, especially for three and four species. Especially due to the combination of different sources of evidence for functional similarity between proteins, our method is able to predict many functions even for uncharacterized or only weakly characterized proteins. These functions are not reflected in the recall since these functions are novel, i.e., counted as FP in the comparison against a gold standard. For instance, when combining the novel predictions from different species combinations, we suggest 7,500 new functional annotations for 1,973 human proteins that previously had only zero or one function annotated. Overall, our method produces 12,300 novel annotations for human with an estimated precision of ~76% and 5,246 for mouse with ~81% precision. These numbers by far outreach that of comparable methods. It is also remarkable that our predictions are rather specific, which is reflected in a mean GO-depth of 8 for humans and 7 for mice. To confirm our estimated precision values, we manually verified a number of predictions in the context of colon cancer. Specifically, we studied the gene products *MLH1*, *PMS2* and *EPHB4*, which received 14, 16, and 15 novel annotations through our method. Detailed literature analysis indicates that at least 73% of the novel functions actually are true predictions.

Finally, we compare our approach against three other approaches, *Neighbor Counting* [19], χ^2 [16], and *FS-Weighted Averaging* [27]. We show that our CCS-based method performs significantly better than those methods in almost all settings we studied, especially in terms of precision.

Methods

We devise an algorithm for predicting functional annotations of proteins using Gene Ontology (GO) [40] terms. Our approach is based on comparison of interaction networks from various species and utilizes orthology relationships, conserved modules and local PPI neighborhoods. It is divided into the (a) integration of PPI data from various databases, (b) detection of maximal conserved and connected subgraphs (CCS) using approximate cross-species network comparisons and (c) prediction of new annotations for proteins within functionally coherent CCS (see Figure 1).

Data

We use interaction data of the model organisms *S. cerevisiae*, *D. melanogaster* and *C. elegans*, and the mammals

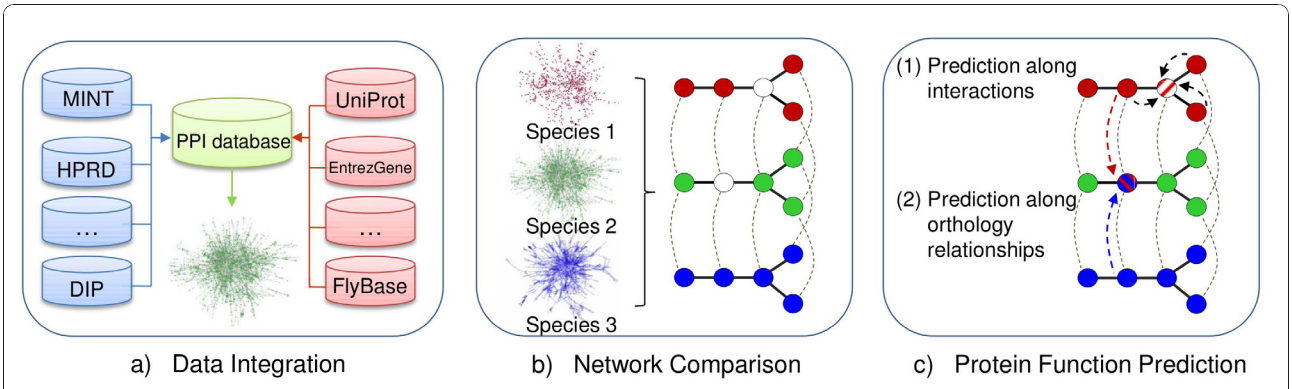

Figure 1 Flowchart summarizing the main steps of our function method. (a) We collect PPI data from several sources and integrate them with additional protein data to generate species-specific PPI networks. (b) PPI network comparisons are performed to identify CCS which (c) are analyzed afterwards for function prediction by exploiting orthology relationships and interacting neighbors.

R. norvegicus, *M. musculus* and *H. sapiens*. Corresponding PPI data were obtained from the major public PPI databases DIP [41], IntAct [42], BIND [43], MIPS-MPPI [44], HPRD [45], MINT [46] and BioGRID [47]. Since the individual coverage and overlap between the data of these resources is comparably low [34,48], we integrate PPI data from the different sources to generate comprehensive data sets for our study. For data integration we map the interacting proteins from external or database specific identifiers to unique protein identifiers from UniProt and EntrezGene [49] to enable the combination of the different data sets to one comprehensive set of interaction data for each species. From the combined data sets we generated comprehensive species-specific protein interaction networks.

Besides the interaction data we utilize protein sequences and protein domain information [50] from UniProtKb/Swiss-Prot [51]. All proteins in the protein interaction network are associated with the respective information. Additionally, proteins are annotated with GO annotations retrieved from UniProtKb/Swiss-Prot, EntrezGene and species-specific databases, such as FlyBase [52], MGD [53], RGD [54], SGD [55] and WormBase [56] (see Additional File 1, Table S1 for a detailed resource listing). Note, when annotating proteins we consider all available GO annotations except for annotations that are assigned without curatorial judgment (GO evidence code: IEA - Inferred from Electronic Annotation). Moreover, we filter for GO subontology root terms to exclude molecular function, biological process and cellular component. The annotated species-specific protein interaction networks (see Table 1) provide the basis of our protein function prediction method.

Network Comparison

We compare protein interaction networks across different species to detect subgraphs that are evolutionary conserved and likely represent functional modules. Figure 2 depicts the strategy of our network comparison approach which involves (1) the identification of orthologous proteins and (2) the detection and assembly of interologs into CCS.

(1) Orthology is a strong indicator for functional conservation. However, the presence of large protein families, typical for mammals and higher eukaryotes in general, makes it hard to distinguish between true orthologs, in-paralogs and paralogs [57]. We determine orthology relationships among multiple species by applying OrthoMCL [58] using default parameters. Previous work showed that OrthoMCL is able to discriminate between orthologs, in-paralogs and functionally unrelated (out-)paralogs at a balanced trade-off between specificity and sensitivity [59].

(2) For comparing protein networks across species, we consider all ortholog groups that comprise at least one protein of each species under consideration. We then use

Table 1 Characteristics of the generated species-specific PPI networks.

species	#proteins	#PPIs	GO terms/ protein	median PPI/ protein
R. norvegicus (*rno*)	973	1221	8	1
M. musculus (*mmu*)	3892	4670	4	1
H. sapiens (*hsa*)	13494	43637	2	2
D. melanogaster (*dme*)	10646	38723	3	3
C. elegans (*cel*)	3499	5858	1	1
S. cerevisiae (*sce*)	6578	67059	4	7

PPI networks for each species are created by integrating PPI data from DIP, BIND, IntAct, BioGrid, MIPS-MPPI, MINT and HPRD. Proteins within the networks are additionally associated with sequences, protein domains and GO annotations. For each species the number of proteins and protein interactions as well as the median number of GO terms per protein is specified.

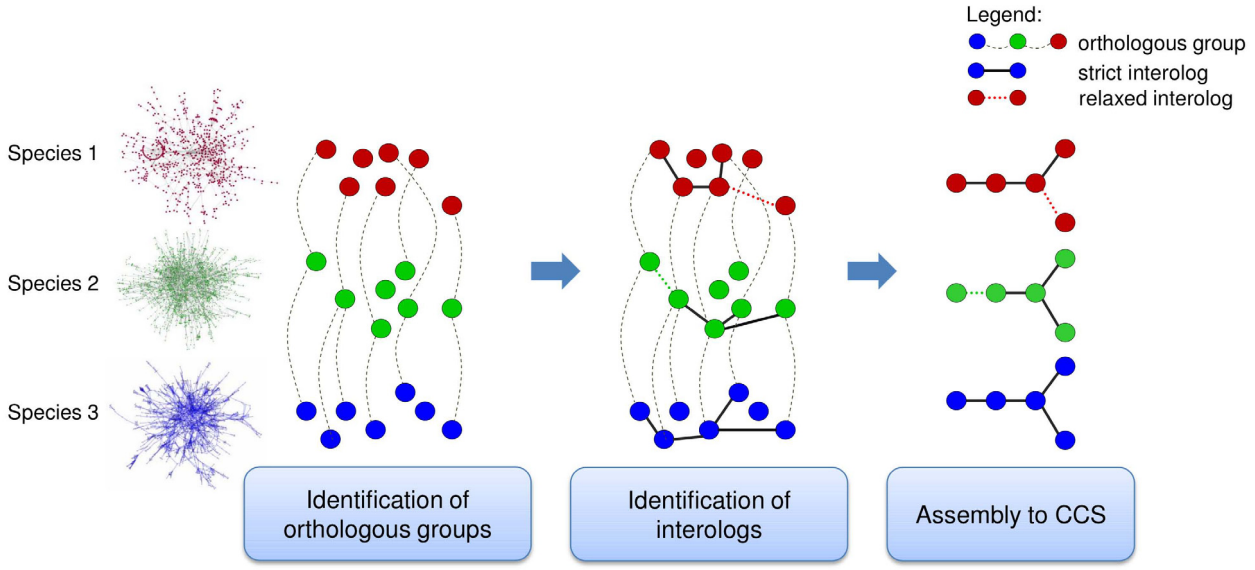

Figure 2 Illustration of the detection of CCS. Protein interaction networks are compared across different species to identify evolutionary and conserved subgraphs. First, orthology relationships across multiple species are determined by using OrthoMCL. Second, all pairs of conserved interactions (interologs) are identified between the orthologs within the species. Adjacent interologs are then assembled to CCS.

an adaption of an algorithm for frequent subgraph discovery [60] to assemble interologs into CCS. Our approach first identifies all interactions (interologs) that are conserved across the different species. For identifying interologs we use two different definitions for interologs depending on the number of species that are involved. When comparing only two species, we use the classical, strict definition considering each interaction as interolog that is present in both species. When comparing more than two species, we consider each interaction as interolog that is present in more than 50% of the species networks (see Discussion). Out of the set of interologs, one interolog is chosen as subgraph seed and all interologs adjacent to this subgraph are added recursively. If a subgraph can not be further extended we store this maximal and connected subgraph as CCS (see Figure 2).

Prediction of Functional Annotation

CCS are conserved subgraphs of interacting proteins and therefore a strong indicator for functional similarity of proteins within a CCS even across species. However, not all detected CCS are good candidates for function prediction due to the noise and incompleteness within the existing PPI and annotation data sets. Therefore, we first filter for CCS that are too heterogeneous or simply too small to be used for function prediction. We then use different methods for predicting functional annotations for all proteins in a CCS, namely transfer of annotations from other species along orthology relationships and transfer within species from all PPI neighbors. In both cases, only proteins within the same CCS are considered. Finally, special care has to be taken for the processing of large CCS which, due to their sheer size, usually are functionally heterogeneous. In the following, we give details for each of these steps.

Filtering coherent CCS

We first test all detected CCS for functional coherence using a functional similarity measure proposed by Couto *et al*. [61] that is based on semantic similarity. We compute, for each CCS, its average functional similarity within a species (Sim_{neigh} - similarity between neighbors) and across the species (Sim_{ortho} - similarity between orthologs). The formal definitions of both similarity measures are provided in the Additional File 1 (see Eq. S7 and S8 in Section S1.1).

We further only consider CCS which have (a) more than two proteins and (b) whose similarity score, either Sim_{ortho} or Sim_{neigh}, exceeds a given threshold. We applied three different thresholds (low: 0.3, medium: 0.5, high: 0.7) to study the performance of our method for different levels of functional coherence. This scheme is applied separately for each subontology of GO (molecular function (MF), biological process (BP), cellular component (CC)).

Prediction using orthology relationships

For inferring protein function from orthology relationships within a CCS, we determine orthologous groups that differ significantly in their individual functional similarity from the similarity score of the CCS by computing the standardized z-score (see Eq. S9). In groups

with significant differences (p-value <0.01) we transfer all known protein annotations to poorly annotated or uncharacterized orthologs. Note that an orthologous protein group might consist of more than one protein per species (orthologs and in-paralogs). Although all proteins within such a group in theory should be functionally highly similar, this is, probably due to missing or wrong annotations, not always reflected in the data (see Results). We define the consensus annotation of all proteins of one species in an orthologous group to be the set of all GO terms that are associated to more than half of the annotated proteins of that species in that group. When considering more than two species we combine the species-specific sets of consensus annotations and transfer them to the other proteins in the same group.

Prediction using neighboring proteins

Given a protein in a CCS, we decide for each GO term annotated to any of its direct neighbors whether it also should be annotated to the protein itself. Let G be the set of terms annotated to at least one neighbor of a protein p, and let N_g be the set of neighbors of p annotated with a term $g \in G$. We transfer g to p if there are more than f proteins in N_g whose functional similarity to p is higher than a given threshold t. For functional similarity between proteins, we again use the method from Couto *et al*. [61] (see Additional File 1, Eq. S5 in Section S1.1.2).

Because this approach cannot predict functions for proteins without any annotation (their computed similarity to other proteins is always zero), we also consider the pairwise functional relation between interaction partners, assuming that a high functional similarity between indirectly linked partners should also hold for the protein itself. Again, if the pairwise similarity scores exceed the threshold t we predict common GO annotations to the candidate protein.

Combined prediction method

We combine the two different methods to predict protein functions within a CCS (see Figure 1c). Proteins that are only weakly and incompletely characterized or not annotated at all are candidates for our prediction approach. For each candidate protein we infer novel protein function (a) within functionally coherent CCS by exploiting its (b) orthology relationship across other species as well as (c) the information shared by its neighboring proteins.

Processing large CCS

Comparing evolutionary close species (such as human and mouse) often results in very large CCS with up to several hundreds of proteins. However, biological processes typically involve only between 5 and 25 proteins [21]. Consequently, large CCS often encompass various functions (see Figure 3) which is reflected in a minor functional homogeneity. Our results confirm this fact, as large CCS always get low coherence scores (see Results). To adequately treat such CCS, we split CCS with more than 25 proteins into smaller, overlapping sub-subgraphs. Sub-subgraphs are built by considering each protein of the CCS as seed of a new, smaller CCS. Subsequently, we add all direct neighbors of this seed to the new CCS (see Additional File 1, Figure S1 for an example). Subgraphs with less than three proteins are removed. We then consider each of these subgraphs as an independent CCS.

Performance evaluation

We use a leave-one-out cross-validation to estimate the expected precision and recall of function prediction using (a) only orthology within CCS, (b) only neighbors within CCS, and (c) the combination of both methods. Precision P and recall R are defined as:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

where TP and FP denote true and false positives, respectively, and FN denotes false negatives.

For cross-validation we 'hide' selected annotations before applying our algorithm. Predicted terms are then compared to the held out annotations. We count a GO term as correctly predicted if the proposed term was an ancestor of the original term on the path to the root or the term itself (see Additional File 1, Section S3.2 and Figure S2 for an evaluation of this criterion). For all methods involving CCS, we give recall values on the basis of all annotations of proteins within qualifying CCS. We call this measure per-protein recall. It must be distinguished from the traditional per-species recall (Eq. 2) which is also used frequently, but which punishes all methods that first filter proteins. When determining the per-protein recall (R_{pp}) we consider only proteins p that are part of a CCS:

$$R_{pp} = \frac{\sum_{p \in \mathrm{CCS}} \mathrm{TP}_p}{\sum_{p \in \mathrm{CCS}} \left(\mathrm{TP}_p + \mathrm{FN}_p \right)}, \tag{3}$$

where TP_p denotes the number of correctly predicted functions for a protein p in a CCS and $(\mathrm{TP}_p + \mathrm{FN}_p)$ corresponds to the number of annotations that are originally associated with the protein p. To also give an idea of the per-species performance, we always complement

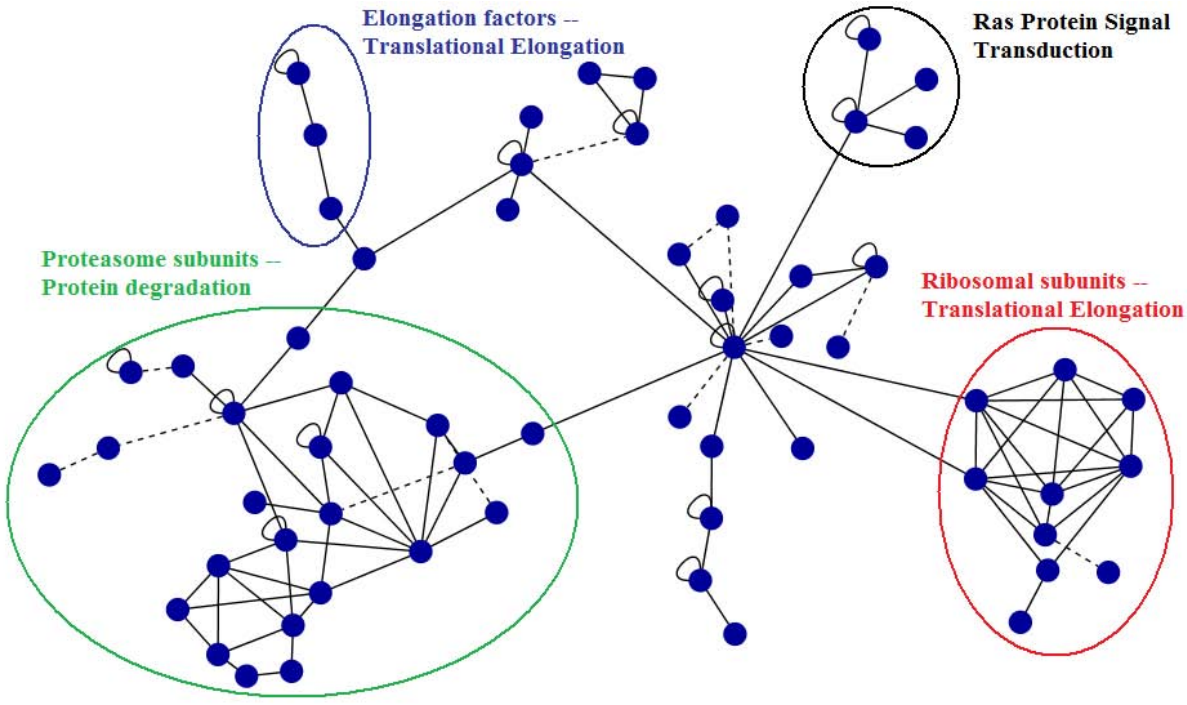

Figure 3 Different biological subprocesses within the largest CCS from human, fly, worm and yeast. This CCS consists of 61 proteins and 108 interologs and encompasses different biochemical activities, such as protein degradation, translational elongation and signal transduction.

precision and recall values with the coverage measure, which simply counts the total number of predictions.

Keep in mind that, as always when comparing to an incomplete gold standard, cross-validation inherently considers any new annotations as false, although new annotations are the primary target of function prediction. Therefore, we also performed an extensive literature evaluation to judge the correctness of selected new annotations.

Comparison to other methods

We compare our approach against a number of different techniques.

First, we use two baseline methods: The orthology baseline purely considers orthology ignoring structural network conservation. We randomly select one third of the orthologous protein groups, remove annotations from one protein in the group and predict their functions using only its orthologs. The link-based baseline takes only direct interaction partners into account, again independently of conservation of interactions. For each species we randomly choose one third of the proteins from the corresponding interaction network and exploit their direct neighbors for deriving new functions. We repeat this procedure 100 times for each baseline and compute average and standard deviation across all runs.

We also compare our results with three popular PPI-based function prediction methods. The Neighbor Counting Approach from Schwikowski *et al.* is a local prediction approach that derives new annotations for a protein based on the frequency of annotations within its direct interaction partners [19]. The χ^2 algorithm from Hishigaki *et al.* extended this idea by also considering the background frequency of a functional term [16]. Finally, the Functional Similarity Weighted Averaging method from Chua *et al.*, a weighted averaging method to predict the function of a protein based on its direct and indirect interaction partner [27,62]. Chua *et al.* demonstrate in [27] that the FS-Weighted Averaging significantly outperforms local and global network approaches, e.g. methods that are based on markov random field or functional flow [26,29]. For comparisons, we adapted a script provided by Chua *et al.* that implements these three methods (see Additional File 1, Section S1.4 for details). To enable a valid direct comparison, we evaluate the three related predictions methods only on proteins that are involved in CCS. The individual performance of each method on the entire data set is shown for completeness in the Additional File 1.

Results

We integrated PPI data for rat (*rno*), mouse (*mmu*), human (*hsa*), fly (*dme*), worm (*cel*) and yeast (*sce*) from

several public databases to generate species-specific PPI networks (see Table 1). We computed CCS for 15 combinations of two species, 20 comparison with three, and 11 with four species, and subjected them to our function prediction method. The number of detected CCS for combinations of five and six species is too low for a systematic and detailed analysis (see Additional File 1, Table S2).

In the following, we focus on four selected species combinations that cover different interactome sizes and evolutionary distances to discuss properties and results of our function prediction strategy. Complete results are given as Additional File 2, Table S2 and Additional File 3, Table S3.

Network Comparisons

We compared protein interaction networks across different species to identify evolutionary and functionally conserved subgraphs that are used as basis for function prediction. Conserved sub-networks are assembled by combining conserved interactions, called interologs, using different definitions of interologs depending on the number of species being compared. For species pairs, we use the classical, strict definition: An interolog is an interaction present in both species. We relax this demand when comparing more than two species to cater for evolutionary variation [63] and for the incompleteness [36] and noise within present PPI data sets [34]: An interolog then is defined as an interaction which is present in more than 50% of the species being compared.

We present a brief overview on the respective network comparison of *rno-dme*, *rno-hsa-sce*, *hsa-dme-sce*, *hsa-dme-cel-sce* and *mmu-hsa-dme-cel* (see Additional File 2, Table S2 for complete results). Table 2 summarizes the outcomes for the selected species combinations in terms of orthologous protein groups, identified interologs and assembled CCS. As expected, the number of orthologous protein groups, interologs and identified CCS differs depending on the number of compared species, their evolutionary distance as well as their current interactome coverage. Comparison of fly and yeast results in 17 CCS (out of 73) with at least three proteins. For more than two species we use the relaxed interolog definition which generally results in a considerable higher number of CCS. For instance, we identify 163 CCS for *hsa-dme-sce* of which 23 comprise more than two proteins. These CCS are shown in Additional File 1, Figure S3. Even combinations with four species result in a reasonable number of CCS, such as *mmu-hsa-dme-cel* producing 16 CCS with more than two proteins.

Table 2 Overview on the outcomes of the selected network comparisons.

	# OrthoMCL groups	# Interologs	# CCS (≥3)	largest CCS
dme-sce	1514	137	73(17)	17
rno-hsa-sce	151	88	31(9)	22
hsa-dme-sce	542	692	163 (23)	187
hsa-dme-cel-sce	519	300	94 (4)	61
mmu-hsa-dme-sce	325	238	73 (16)	20

For each species combination the number of orthologous groups, interologs, CCS are presented as well as the size of the largest CCS. Note, we use the strict interolog definition for two species and the relaxed criterion for multiple species (see Methods).

Function Prediction

We use orthology relationships, functionally conserved modules, and direct and indirect protein interactions for predicting functional annotations for proteins in a CCS by transferring annotations from other species along orthology relationships and within species from interaction partners. We evaluated our approach in three ways. First, we compared our combined strategy to baseline methods which disregard conservation in networks. Second, we compared it to the results obtained from using orthology and PPI neighborhood within CCS in isolation. Third, we performed a comparison to three recent function prediction methods from the literature.

We first show the performance of our two baseline methods, orthology and link-based, for function prediction. Precision for predictions based solely on orthology relationships varies between 3% and 11% (see Additional File 1, Table S4). Recall is higher (3% to 40%), but decreases steeply with the number of species being compared. Precision of the link-based baseline ranges from 3% to 17%. Contrary to the orthology baseline, recall is rather high, varying between 51% and 75% (see Additional File 1, Table S5). Thus, the link-based baseline reaches a similar precision but higher recall than the orthology baseline. Both baselines yield very low precisions. The orthology baseline indicates the challenges transferring function from ortholog templates. Although function tends to be conserved in orthologs, orthology does not guarantee conservation of function [38]. When transferring function solely based on protein sequences, more sophisticated approaches, e.g. using advanced statistical frameworks [9], are needed to ensure high prediction quality. The precision of the link-based baseline is lower than expected most likely through the strong impact of the quality of the interaction data. However, precision and recall are similar to the results of the two local prediction approaches of Schwikowski *et al.* and Hishigaki *et al.* that are applied to our data (see Discussion).

Across Orthology Relationships within CCS

We use orthology relationships underpinned by interologs to infer novel functions from multiple species. Considering only orthology relationships for transferring functions to proteins within CCS results in predictions with medium to high precision. Additional File 1, Table S6 shows precision and recall estimated using cross-validation for the selected examples. Precision reaches 88% to 97% for yeast proteins when comparing *hsa-dme-sce* and 67% to 85% for mouse proteins when comparing *mmu-hsa-dme-sce*. Precision values increase considerably with a higher coherence threshold for CCS, but this improvement comes at the cost of lower coverage. Particularly low numbers of predictions are obtained for comparisons involving species with low PPI coverage. This is especially prominent for *rno*, where comparison of *rno-hsa-sce* result in only 8 predictions - but with a precision of 100%.

Besides the coherence threshold, also the number of species being compared has a strong impact on performance. Higher average precisions are achieved when analyzing multiple species compared to species pairs. For instance, the average precision for *mmu-hsa-dme-sce* is 79% at 0.3 in comparison to *dme-sce* with 54% at 0.3 and 69.5% at 0.7. This shows that using more species implicitly selects functions that are conserved more strongly, which underlines the impact of evolutionary functional conservation for protein function prediction. This fact also shows up when comparing to the orthology baseline (see Additional File 1, Table S4): Precision and per-protein recall using orthology within CCS are much higher, but the overall coverage is much lower. This means that CCS strongly restrict the number of proteins for which predictions are made, but this restriction is done in a very sensible way removing mostly false positive predictions.

Across Neighborhood within CCS

Additional File 1, Table S7 shows precision and recall for inferring functions only from interaction partners within CCS. Compared to predicting function based on orthology within CCS, precision is higher, while per-protein recall roughly stays the same. At the same time, neighbor-based prediction has a considerable better coverage. However, there are also species combinations in which this method performs worse. Precision again correlates with the functional coherence of CCS and with the number of compared species, but the impact is less pronounced. Especially the step from coherence threshold 0.3 to 0.5 mostly makes only a small difference. Compared to the link-based baseline (see Additional File 1, Table S5), precision is much higher and coverage and per-protein recall decreases.

Combining module, orthology and link-based PPI evidences

We hypothesized that the integration of orthology relationships, evolutionary conserved functional modules, and direct and indirect protein-protein interactions into a single prediction strategy will combine the strengths of the three individual methods. Selected results from this combined strategy are shown in Table 3 (see Additional File 3, Table S3 for complete results). As before, precision varies (from 46% to 91%) depending on the species combination and the threshold for functional coherence of CCS. Best results are obtained for *rno-hsa-sce* at a threshold of 0.7, with precision of 85%, 89% and 86%, respectively.

As mentioned before, one of the major drawbacks of using only CCS orthology relationships is the low number of predictions due to the restriction to orthologous proteins with at least one known function (see Additional File 1, Table S6). In contrast to orthology-only, the combined approach creates many more predictions (2- to 50-times more). It generates hundreds or even thousands of predictions also for those cases where the orthology-only method could not predict any function.

Comparing the combined method and CCS link-based only (see Additional File 1, Table S7) shows an increase within the amount of predictions (e.g. about 2-times for *dme* from *dme-sce*), although it is less steep than observed for orthology-only. This increase has mostly only minor influence on precision and recall. Precision reaches similar levels and the recall increases slightly. Note, for few combinations the combined method yields

Table 3 Prediction results when combining module-based CCS, orthology relationships, and neighboring proteins.

	0.3			0.5			0.7		
	# terms	P	R_{pp}	# terms	P	R_{pp}	# terms	P	R_{pp}
dme	6242	0.50	0.29	5072	0.52	0.25	1522	0.73	0.32
sce	3567	0.61	0.27	2581	0.71	0.28	1303	0.83	0.40
rno	1125	0.63	0.20	485	0.67	0.27	1185	0.85	0.30
hsa	1489	0.56	0.29	368	0.85	0.34	223	0.89	0.34
sce	1870	0.60	0.25	1206	0.61	0.17	229	0.86	0.24
hsa	13975	0.46	0.35	4418	0.57	0.36	723	0.73	0.33
dme	18638	0.62	0.41	16225	0.61	0.38	3462	0.71	0.48
sce	16544	0.72	0.44	15524	0.72	0.43	4135	0.84	0.55
hsa	3314	0.47	0.25	439	0.75	0.28	160	0.91	0.41
dme	5190	0.58	0.22	4586	0.59	0.23	866	0.81	0.29
cel	2464	0.47	0.27	1796	0.56	0.27	256	0.65	0.31
sce	5361	0.70	0.31	5126	0.71	0.32	1212	0.80	0.37
mmu	1212	0.66	0.17	459	0.81	0.32	53	0.81	0.34
hsa	3301	0.48	0.28	1658	0.57	0.33	436	0.65	0.81
dme	5561	0.56	0.29	4642	0.57	0.29	1400	0.59	0.55
sce	5159	0.63	0.31	4906	0.63	0.31	2140	0.73	0.72
average	5870	0.58	0.29	4343	0.65	0.30	1160	0.77	0.42

Precision (P) and per-protein recall (Rpp) are estimated for low (0.3), medium (0.5) and high (0.7) functional similarity/conservation thresholds.

the same results as link-based-only because no predictions could be inferred through orthology relationships.

Overall, the impact of our combined approach is dominant, especially in terms of the number of predictions. Precision drops for some combinations compared to the single methods. However, the decrease of precision does not indicate a lower prediction quality. It rather indicates that the combined method derives many more novel predictions that can not be validated during cross-validation rather than successfully reproducing known function for well-characterized proteins (see Discussion of predictions). Precision is affected the least for the highest similarity threshold (0.7) fostering the most reliable precisions.

Overlap between orthology- and link-based predictions within CCS

We combined orthology- and link-based function prediction within CCS to benefit from the strengths of both methods. To study whether the predictions of the individual methods result in the same or complementary sets of predictions we determined the overlap of GO terms predicted by either strategy. For *hsa-dme-sce*, the respective numbers are shown as Venn diagrams in Additional File 1, Figure S4. In general, the major fraction of unique predictions is derived from neighboring proteins. The overlap between predictions is comparably small and decreases when increasing the similarity threshold. This shows that both methods complement each other very well as they predict rather different sets of functions. For *hsa-dme-sce*, the respective numbers are shown as Venn diagrams in Additional File 1, Figure S4. In general, the major fraction of unique predictions is derived from neighboring proteins. The overlap between predictions is comparably small and decreases when increasing the similarity threshold. This shows that both methods complement each other very well as they predict rather different sets of functions.

This behavior is also observable when predictions are analyzed separately per species (see Additional File 1, Figure S5). However, contrary to fly and yeast proteins (see Additional File 1 Figure S5(b) and S5(c)), the amount of orthology and link-based predictions is quite similar for human proteins (see Additional File 1, Figure S5(a)), which can be explained by the much denser PPI data available for the two model organisms (see Table 1). This observation clarifies that different species pro t differently from our method. Especially less characterized species, such as human, benefit strongly from the functional knowledge of model organisms.

Overlap between predictions derived from different species combinations

Not only does the neighbor-based method complement the orthology-based method, but also predictions derived from different species combinations are rather complementary. Table 4 shows the overlap between predictions for human proteins inferred from different species pairs. The overlap is determined by dividing the number of overlapping predictions through the total number of predictions of a combination (expressed as percentage). The overlap mostly is far below 50% and strongly depends on evolutionary distance between the species. For example, the overlap between predictions derived from CCS with mouse and those derived from rat is much larger than that of the sets derived from mouse and, say, fly. The same holds for combinations of three and four species (data not shown). Moreover, the more species we combine the more we focus our prediction on evolutionary conserved functions, which becomes clear when studying predictions for highly conserved housekeeping functions (see Discussion).

Table 4 Fraction of overlapping function predictions (in %) for human proteins derived from different species pairs.

	mmu-hsa	*hsa-dme*	*hsa-cel*	*hsa-sce*
rno-hsa	44.6/48.7	40.4/14.1	22.0/28.0	33.3/19.4
mmu-hsa	-	47.7/16.1	21.5/94.4	52.4/25.1
hsa-dme	-	-	4.2/17.5	39.4/42.9
hsa-cel	-	-	-	89.0/74.2

The overlap is defined as the number of overlapping predictions divided by the total number of predictions (expressed as percentage). Each cell contains two different values - i/j - that specify the overlap based on the total number of predictions of the two combinations. i presents the overlap between the non-human species from row i and column j and value j presents the overlap between non-human species from column j and row.

Large CCS

Large CCS naturally encompass various biological functions. In consequence, their functional homogeneity is often too low which excludes the entire CCS from function prediction. However, large CCS actually are strong indicators for conserved functions. For instance, Figure 3 shows the largest CCS from *hsa-dme-cel-sce* consisting of 61 proteins and 108 interologs with its different biological subprocesses. It clearly contains several functionally highly conserved clusters, probably forming discrete protein complexes. Considering such a large CCS as a whole is insufficient. Therefore, we modify our approach for large CCS by breaking them up into sub-subgraphs (see Methods). The impact on precision and recall is shown in Additional File 1, Table S8 (large CCS are split), which should be compared with entries of Table 3 (large CCS are ignored). As can be seen, processing large CCS creates many more predictions with mostly better precision. For example, the number of predictions almost triples for *hsa-dme-sce* at a similar or even better precision. When comparing split and non-split results from *hsa-dme-cel-sce* the precision decreases for human along a five-fold increase of the number of predictions, but increases for all the other species (at 0.7).

Comparing with other methods

We compare the performance of our CCS-based prediction approach against *Neighbor Counting* (NC) [19], χ^2 statistics [16] and *FS-Weighted Averaging* (FS-WA) [27] considering only proteins that are involved in CCS. The performance of the individual methods on the complete data is shown in Additional File 1, Figure S6. Figure 4 presents precision - recall graphs (based on varying thresholds) for predictions for human proteins separated by the three GO subontologies. CCS-based function prediction significantly outperforms *NC* and χ^2 statistics. Precision and recall obtained from the latter two are very low and even below our baselines. This also holds for yeast and fly (see Additional File 1, Figure S7 and S8).

When comparing *FS-WA* results with our approach, CCS-based function prediction performs consistently as well or better. Depending on species and subontology we achieve either higher precision at a similar recall or an improved precision and recall. Especially, when considering molecular function and biological process in human (see Figure 4) our method clearly outperforms *FS-WA*.

Discussion

We presented a novel approach to predict protein functions that uses data from multiple species and combines three different sources of evidences for functional similarity: Orthology relationships, evolutionary conservation of functional modules in protein networks, and direct and indirect protein-protein interactions. Integrating these evidences into a single prediction algorithm overcomes the individual weaknesses of the base methods: (1) Orthology restricts prediction to proteins that have at least one orthologous protein with known function and exhibits a very low precision. (2) Considering only protein-protein interactions disregards the power of comparative genomics, leading to low coverage in organisms where PPI data is not available in abundance. (3) Using only functional modules within protein networks yields high precision, but strongly affects recall on a species basis, as only highly conserved functions performed by dense protein clusters can be predicted. We showed that combining these methods leads to high precision predictions with very good coverage. Essentially, we achieve high precision by looking only at subgraphs conserved in multiple species without restricting them to dense modules. Furthermore, we achieve high coverage when considering multiple species, by using a relaxed definition of interologs, and by transferring function from PPI neighbors and from orthologous proteins. Altogether, our method predicts thousands of protein functions for every species included in the analysis at varying, yet always high levels of precision (see Table 5).

Network Comparison

For comparing protein interaction networks we used two definitions for determining interologs: the strict and the relaxed definition when studying either two or more than two species, respectively. We also experimented with using the strict interolog definition for multiple species, but this often results in zero or only very few

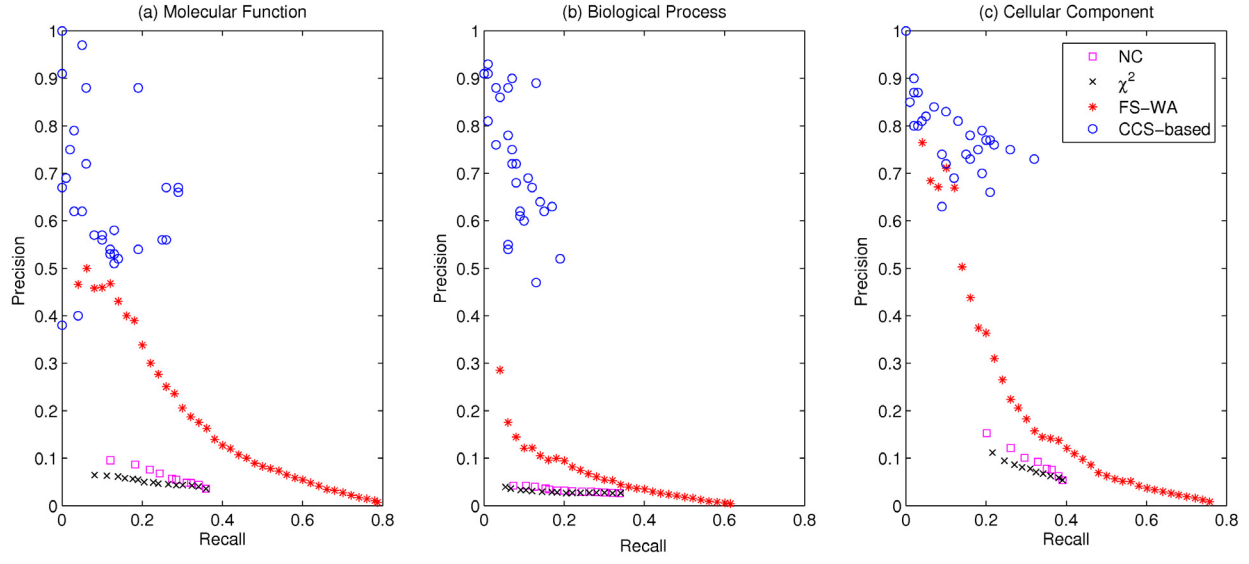

Figure 4 Direct performance comparison for human. Comparing precision and recall of function predictions for proteins involved in CCS from weighted average (WA), neighbor counting (NC), χ^2 statistics and CCS-based approach for (a) molecular function, (b) biological process and (c) cellular component. CCS-based results are retrieved from different similarity thresholds and species combinations.

Table 5 Overall function prediction statistics across all species combinations.

species	# terms	P	R_{pp}	# terms	P	R_{pp}	# terms	P	R_{pp}
rno	17430	0.50	0.12	9110	0.57	0.14	5738	0.75	0.20
mmu	26089	0.47	0.13	11441	0.59	0.16	5246	0.81	0.23
hsa	62833	0.44	0.14	22911	0.59	0.17	12317	0.76	0.39
dme	26155	0.56	0.19	19463	0.60	0.20	10455	0.75	0.30
cel	4098	0.52	0.14	1983	0.60	0.18	1332	0.71	0.24
sce	10666	0.66	0.23	9013	0.69	0.25	6335	0.81	0.35

Total number of newly derived GO annotations and their estimated average precision (P) and per-protein recall (Rpp) for each species using low, medium and high similarity thresholds.

and small CCS within species groups (see Additional File 1, Table S2 for strict vs. relaxed results). This leads to a small but highly precise set of function predictions. Being less strict leads to a significant improvement in the coverage of our prediction method at comparable precision (see Additional File 1, Section S2.1 and Table S9). In turn, we also tested the effect of applying a relaxed definition of interologs to species pairs. This leads to very few (often only one) yet very large networks, as it only creates the union of interactions between orthologous proteins of the two species. However, this does not reflect evolutionary conservation of PPIs and therefore misses the important signals of functional conservation.

Function Prediction

We evaluated or method in several ways using precision and recall, two baselines and three other function prediction methods. However, besides pure precision and recall values, an important property of any function prediction method is the specificity of its predicted terms. Clearly, predicting only very general terms is much simpler but much less useful than predicting terms close to the leaves of GO. Our methods predicts terms at a median level of 10 for cellular component, 8 for biological process, and 6 for molecular function. Thus, our method is capable of predicting quite specific functions (also see discussion of novel functions below).

Compared to other methods presented in the literature, our method has also the important property that it is not limited to so-called "informative" GO terms [64]. Many prediction methods use only GO terms that are associated to more than ten or 30 genes [26,27,62]. Such an approach implicitly disregards more specific annotations, although those are the most valuable ones. For example, in 2007 82.5% of GO annotations in human were associated to less than ten genes [65] leaving only 17.5% as annotation basis. GO-based methods have been shown to result in higher precisions when applied on a small number of frequently annotated GO terms. In contrast, we are able to generate accurate predictions also for rarely used GO terms.

Comparison to Baselines

Compared to the orthology baseline, including CCS yields precisions up to 10-times higher, confirming that information on conserved interactions is a very effective filter for avoiding false positive predictions across orthology relationships. Compared to the neighbor baseline, considering CCS also leads to a clear and significant increase in precision. This effect can be explained by the fact that using interologs (strict or relaxed) instead of single interactions largely improves reliability of PPI data [66], since false positive PPIs are unlikely to be reproduced across multiple species.

Our results show that combining various evidences into a single and comprehensive method leads to improved results. Evidently, the predictions made by different methods using information from different species complement each other quite well instead of only predicting the same functions again and again. However, the concrete approach has to be chosen with care. We showed that good results can only be achieved when using a proper definition of interaction conservation and when treating large CCS in an adequate manner. Failing to do so either restricts coverage of the method or leads to a higher false positive rates.

Comparison to other Function Prediction Methods

We compared precision and recall of our approach to Neighbor Counting, χ^2 statistics and FS-Weighted Averaging (see Methods). Our combined CCS-based approach significantly outperforms Neighbor Counting and χ^2 statistics, especially in terms of precision. Moreover, we perform comparably well or better against FS-Weighted Averaging, mostly achieving much higher precision at higher recall. Notably, our method achieves better results especially for species without comprehensive PPI coverage, such as human. However, precision for Neighbor Counting and χ^2 is significantly lower (on the entire data sets, see Additional File 1, Figure S6, and the filtered protein sets) than reported in the respective original publications [16,19]. There are three explanations for this drop (from ~70% to 15% precision). First, both methods originally were evaluated only on the functional classification scheme from YPD. This scheme covers, similar to GO, three categories of yeast protein function: biochemical function, cellular role and subcellular localization. However, categories have only 57, 41 and 22 members, respectively. Compared to our evaluation using GO, in which methods have to chose between up-to 17398 functional categories, this increases the chances to predict correct terms purely by chance. Furthermore, yeast is a particularly well-studied organism, while we applied the method also to less-well

covered species. A similar performance drop was observed by Chua *et al*. [62], which also applied both methods to GO term prediction, with precision decreasing to 60% (*NC*) and 20% (χ^2) for yeast and 20% (*NC*) and 16% (χ^2) for fly. The second point concerns the amount of interaction data. For example, results from [19] are based on only 2,709 interactions among 2,039 proteins. In contrast, we integrated six different databases, leading to, for instance, almost 70,000 interactions for 6,500 proteins in yeast. Thus, we cover many more proteins and interactions which also increases the probability of false positives. Third, many prediction methods, including the two studies compared to here, consider only annotated proteins with at least one annotated interaction partner for their studies [16,19,26,27,62]. We did not exclude those proteins because we believe that especially weakly or un-annotated proteins must be a primary target for function prediction. In our combined approach, such proteins often receive functions from orthologous proteins in other species, an option missing in Neighbor Counting and χ^2. However, functions predicted for non-annotated proteins are necessarily counted as false positives although these are truly novel findings. Thus, disregarding such proteins results in higher precisions.

We did not compare to purely module-based prediction methods, as link-based techniques have been shown to outperform those [3,37]. However, we evaluated the effect of requiring CCS to be "module-like", i.e., to exhibit a certain density of interactions between its members. CCS-density is defined as $\frac{2^*|E|}{|V|(|V|-1)}$ where E presents the edges and V denotes the nodes within a CCS. As expected, filtering CCS according to their density considerably improves precision (see Additional File 1, Figure S9), e.g. in fly from 80% without filtering to 90% for a density of 0.7 and 95% for a density of 1, but this increase is at the cost of much fewer predictions (see Additional File 1, Section S2.2 for a detailed discussion).

Effects of Size of Data Sets

Results of our prediction method vary depending on the level of available annotations and PPIs for the species that are compared (see Additional File 1, Section S3.1 for a discussion of the data). They are better when well-studied species, such as yeast or fly, are involved. This is an inherent property of methods that transfer annotations, since better annotated species provide more source functions. This property underpins the importance of comparative genomics for elucidating the function of human proteins. It is also clearly visible that prediction precision is correlated to the threshold for functional conservation (see Figure 5a) and increases with the degree of evolutionary conservation of a CCS - from pairwise to multiple network comparisons (Figure 5b). Obviously, the functional conservation threshold is an important possibility to tune or method to the specific needs of an application. The higher the functional conservation, the higher is the precision of the predictions.

Note that in any gold standard evaluation as ours, new findings are always counted as false positives, independently of their real, biological truthfulness. Consequently, prediction methods perform better on well-studied organisms than on species that are functionally less well characterized. The precision values we report therefore should be considered as lower bounds on the true precision.

Performance on Weakly and Non-Annotated Proteins

An important goal of protein function prediction is to derive novel functions for proteins without any or with only very little functional information. Thus, we analyzed how our method performs on such proteins. We define as a weakly annotated protein (WAP) any protein which has at most two terms assigned a-priori in our data. For WAP, we count annotations as new if they are more specific than the existing ones or if they belong to another sub-branch in the subontology. Note that such annotations are counted as false positives in our evaluation as they cannot be validated from our gold standard data.

Results from comparing *hsa-dme-sce* are shown in Figure 6 and Additional File 1, Figure S10. As expected, the highest number of proteins without any annotation can be found in human. Annotation coverage of fly is not as good as for yeast but still much better than in human. For example, CCS at threshold 0.3 contain ~300 human proteins without any functional annotation in biological process. By means of our method, we predict 156 annotation for 52 of those proteins. Similarly, 20 fly proteins out of 72 are annotated with 67 GO annotation in biological process. But also well-studied species still contain many WAPs and benefit from our approach. For instance, about 380 yeast proteins are only weakly characterized for cellular component and for more than a quarter of them we predict about 200 functions. Note that the fraction of WAPs receiving new annotations decreases with the increase of the similarity threshold for each species.

Predictions for Selected Human Proteins

In the following, we discuss specific predictions for proteins that are relevant for colorectal cancer. Note that these predictions were counted as false positives in our evaluation because they are not contained in the Gene Ontology annotations at all or only marked as putative (mostly "inferred from electronic annotation", IEA). However, we show that many predictions already have

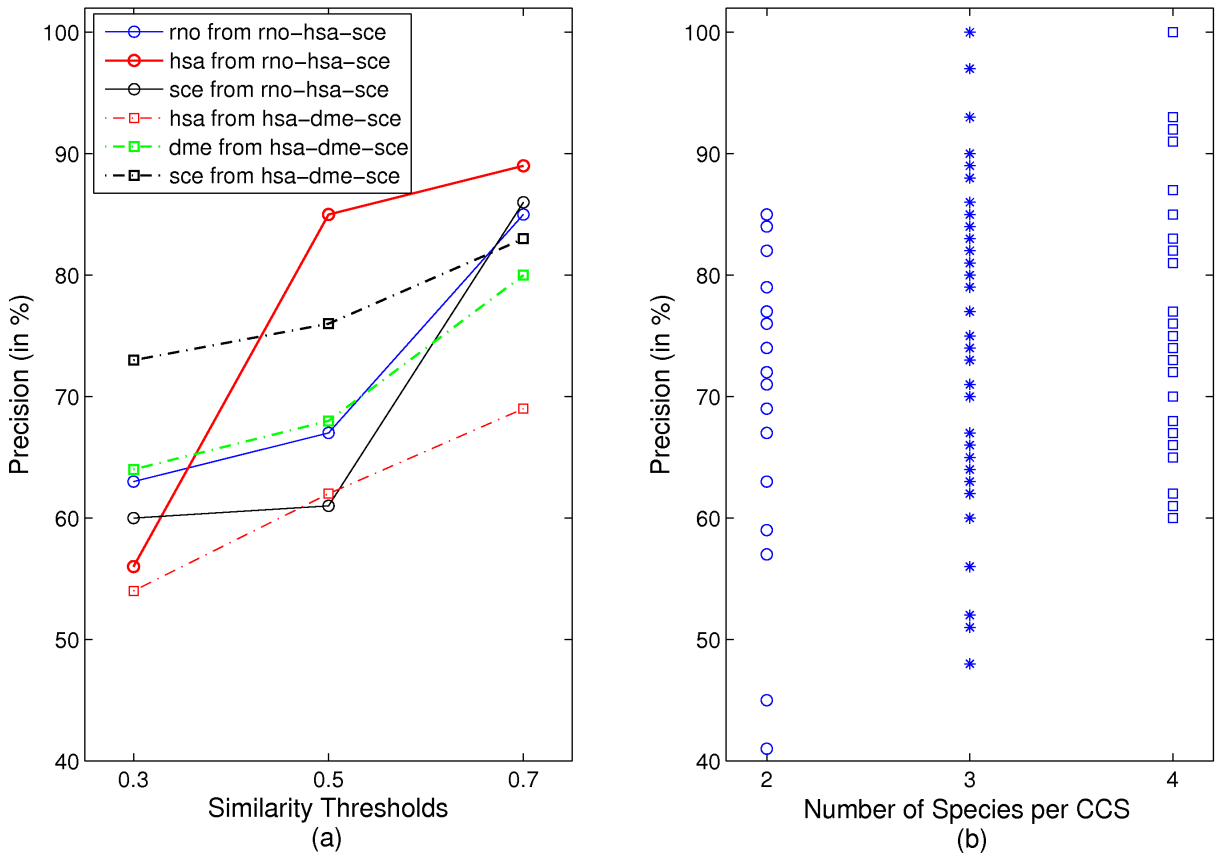

Figure 5 Correlation of the prediction precision with (a) functional and (b) evolutionary conservation, respectively. (a) Species-specific precision values for predictions derived from CCS among *rno-hsa-sce* (solid lines) and *hsa-dme-sce* (dashed lines). For each species we plot the estimated precision against the applied similarity threshold (low: 0.3, medium: 0.5, high: 0.7) that indicates the level of functional conservation. (b) Species-specific precision grouped by evolutionary conservation that is given by the increasing number of species involved in a CCS.

strong experimental support in the literature. Thus, the group of novel predictions falls into two classes - those that, given the current literature, can be considered as true but have not yet made it into the annotation databases and those for which we could not find conclusive evidence in the literature. We argue that, given the large amount of predictions that fall in the first class, predictions from the second class should be considered as promising candidates for further studies.

We discuss predicted functions for the gene products of *MLH1*, *PMS2* and *EPHB4*, all of which have an established importance for colorectal cancer [67,68]. Overall, literature curation largely confirms the predictions for these three genes by different experimental studies.

MLH1 and PMS2

The DNA mismatch repair protein MLH1 and the mismatch repair endonuclease PMS2 belong to the main components of the post-replicative DNA mismatch repair (MMR) system (see Figure 7) [69]. The MMR system is required for correcting base mismatches and insertion or deletion loops resulting from DNA replication, DNA damage, or from recombination events between non-identical sequences during meiosis [70]. Curated annotation for *MLH1* and *PMS2* from UniProt and EntrezGene and newly inferred functions are listed in Additional File 1, Table S10 and S11.

The majority of our predictions (terms are set *italics* in the following) is directly related to the functionality of the MutLα complex which is formed by *MLH1* and *PMS2*. Rich supporting evidence can be found from the respective orthologs in yeast and mouse. For instance, *PMS1*, the *PMS2* ortholog in yeast, contributes to *dinucleotide insertion or deletion binding*, *loop DNA binding* [71]. *Mlh1*, the mouse ortholog of *MLH1*, is annotated to *guanine/thymine mispair binding* [72] and likely plays a role in the formation, stabilization and/or the resolution of Holliday junction intermediates (*four-way junction DNA binding*) [73]. High and low affinity ATP binding sites have been observed for *MLH1* and *PMS1* in yeast [74] which supports the *ATP binding* and

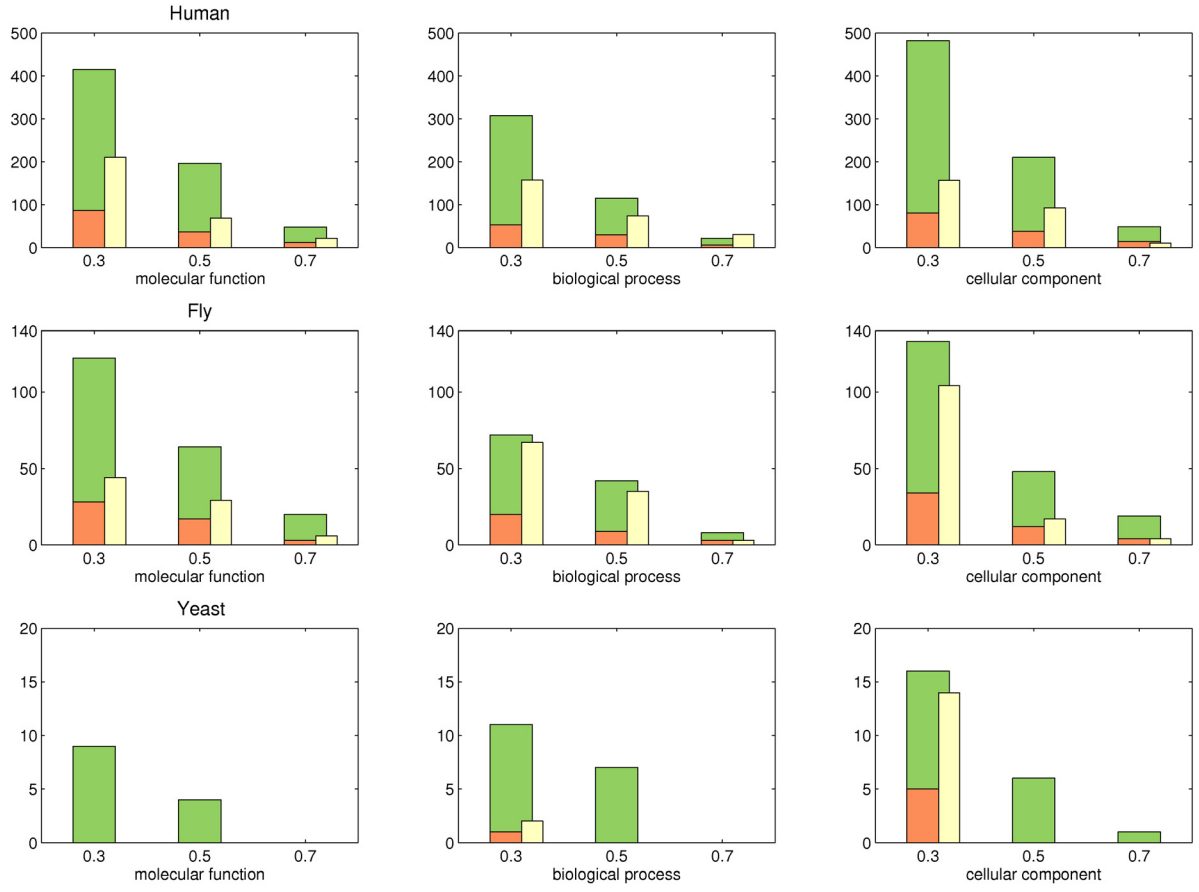

Figure 6 Number of predicted functions for proteins without annotations within CCS from hsa-dme-sce. For each subontology and similarity threshold the number of proteins without any annotation (olive), the number of proteins that receive new annotations (orange) and the total number of novel annotations are shown (yellow). Recall that a higher coherence threshold for CCS leads to less proteins being included in function predictions; thus, numbers generally decrease with higher thresholds.

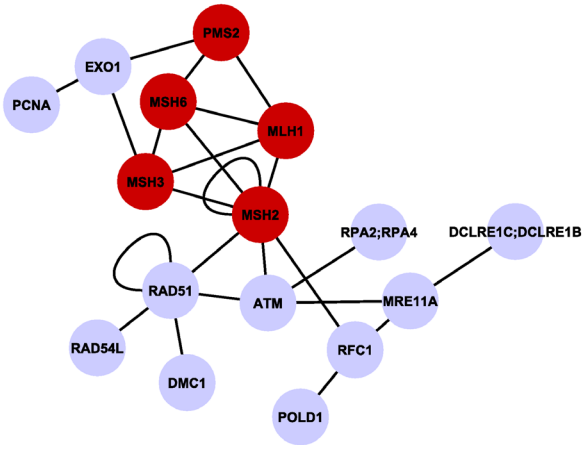

Figure 7 Components of the post-replicative DNA mismatch repair system (MMR). CCS derived from PPI network comparison between human and yeast. Subgraph clusters proteins that are involved in mismatch repair (protein names correspond to human proteins). Proteins associated to colorectal cancer are indicated in red.

ATPase activity predictions for their human orthologs [75]. Moreover, *PMS2* contains a conserved metal-binding motif that constitutes part of the active site for the endonuclease activity of the protein and might enable *magnesium ion binding* [76]. Considering *protein homodimerization activity*, the dysregulated gene expression of *PMS2*, either as a monomer or homodimer, can disrupt MMR function in mammalian cells [77]. Note that, although we support our predictions by literature evidences that are mostly based on orthologs, our algorithm actually inferred them from conserved interaction partners as the orthologs in most cases do not carry the annotation we found in the literature.

Our algorithm also generates a number of predictions that are not as clearly supported by the existing literature, such as *guanine/thymine mispair binding*, *single guanine* and *thymine insertion binding* or *oxidized DNA binding*. Moreover, we associate both proteins to *base-excision repair* as well as *postreplication repair* and

MLH1 to *maintenance of DNA repeat elements*. These are interesting hypotheses supported by recent findings from Erdeniz *et al*. who suggested that the endonuclease activity of *PMS2* in MutLα is not only important in MMR-dependent mutation avoidance but also for suppression of homologous recombination, DNA damage signaling, and damage response functions [78]. Association of yeast *PMS1* with *meiotic mismatch repair* and *DNA recombination* [79] further support these predictions. Regarding their cellular components both proteins are associated to the *MutLα complex* [67], an annotation predicted jointly from orthology and the CCS neighborhood. *MutLα complex* is a clearly sensible refinement of the existing annotation *nucleus* and only seven others genes are annotated to this term, which emphasizes the specificity of our method.

EPHB4

Ephrin type-B receptor 4 is a transmembrane receptor for the ephrin-B family. It belongs to the family of receptor tyrosine kinase (RTK) and is usually expressed in endothelial and neuronal cells. Known and predicted functional annotations are displayed in Additional File 1, Table S12.

Several predicted functions, such as *protein*, *enzyme* and *ATP binding*, *SH3/SH2 adaptor* and *enzyme regulator activity* and *protein amino acid phosphorylation*, derived both from conserved interactions and orthology, are evidently consistent with the characteristics of receptor tyrosine kinases.

Two functions inferred by orthology are *transmembrane-ephrin receptor activity* and *transmembrane receptor protein tyrosine kinase signaling pathways*. Both are supported by annotations from highly related receptors, such as *Ephb1* in mouse and *EPHB2* in human [80,81]. Less evident predictions are, for instance, *cell-cell signaling* [82], *cell migration* [83], *angiogenesis* and *behavior* [84]. These functions were not predicted by orthology alone but only in combination with the conserved interaction neighborhood of *EPHB4*. *EPHB4* participates in the axon guidance pathway and in this context predictions like *axon guidance* or *axon guidance receptor activity* can be integrated [85-87].

Conclusion

Elucidating protein function is a major challenge in the post-genomic era. We developed a method for predicting protein function based on the structural and functional conservation of PPI subnetworks in multiple species. Our approach integrates three different sources of evidences for inferring functional similarity. Altogether, we employ orthology relationships, evolutionary conservation of functional modules in protein networks, and direct and indirect protein-protein interactions for deriving novel functions for uncharacterized proteins. Using our method we derive thousands of protein functions for every species in our study at varying, yet high levels of precision. Thus, combining orthology relationships, functional modules and PPI neighborhood into a single, comprehensive prediction method yields high-quality predictions with very good coverage. In comparison against three other function prediction approaches, Neighbor Counting, ^{2} statistics, and FS-Weighted Averaging, our CCS-based prediction strategy performs comparably well or significantly better, especially in terms of precision.

Additionally, we predict a large amount of novel functions for a number of poorly or non-annotated proteins that can not be validated directly. However, this shows that our method also generates novel functional knowledge rather than only reproducing known functions for well-characterized proteins. The manual curation of predictions for three selected proteins confirms their high quality and precision as many predictions already have strong experimental support in the literature.

Apart from the promising results of our prediction approach, our method currently only provides lists of yes/no predictions. This binary behavior is implicit in the way we compute CCS and how we determine predicted terms and targets of prediction. For further improvement and applicability we plan to derive confidence scores for each prediction based on the multiple biological evidences. Predictions ranked by reliability will provide a method of selection for focusing experimental resources on hypotheses (predictions) that are more likely to be true. This is essential for experimental biologists to decide which proteins and predictions should be investigated further, e.g. in follow-up experiments.

Additional material

Additional file 1: Supplementary Material. The Supplementary Material includes supplementary figures and tables as well as additional analysis.

Additional file 2: Complete results of the strict and relaxed network comparisons. This files contains the complete results of the strict and relaxed network comparisons for pairs of species and three, four, five and six species combinations. The number of OrthoMCL groups, interologs from strict and relaxed definition as well as the total number of CCS and the size of the largest CCS are given.

Additional file 3: Complete results of the CCS-based function prediction approach. This file contains the complete results of the combined CCS-based function prediction approach for pairs of species and three, four, five and six species combinations. CCS from strict and relaxed network comparison are used depending on the species combinations.

Acknowledgements

We would like to thank Hugues Roest Crollius for critical reading of the manuscript. This work is funded by an Elsa-Neumann scholarship and the Deutsche Forschungsgemeinschaft (DFG).

Author details
[1]Knowledge Management in Bioinformatics, Humboldt-Universitat zu Berlin Unter den Linden 6, 10099 Berlin, Germany. [2]Institute of Pathology, Molecular Tumorpathology, University Medicine Charite, Chariteplatz 1, 10117 Berlin, Germany.

Authors' contributions
SJ: developed the methods to identify conserved protein interaction subgraphs and to predict protein functions, carried out the studies described in this paper and contributed to the manuscript. CS: contributed to the manuscript. UL: conceived the study and contributed to the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 6 April 2010 Accepted: 20 December 2010
Published: 20 December 2010

References

1. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: **Protein function in the post-genomic era.** *Nature* 2000, **405(6788)**:823-826 [http://dx.doi.org/10.1038/35015694].
2. Frishman D: **Protein annotation at genomic scale: the current status.** *Chem Rev* 2007, **107(8)**:3448-3466 [http://dx.doi.org/10.1021/cr068303k].
3. Sharan R, Ulitsky I, Shamir R: **Network-based prediction of protein function.** *Mol Syst Biol* 2007, **3**:88 [http://dx.doi.org/10.1038/msb4100129].
4. Reference Genome Group of the Gene Ontology Consortium: **The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species.** *PLoS Comput Biol* 2009, **5(7)**: e1000431 [http://dx.doi.org/10.1371/journal.pcbi.1000431].
5. Baxter SM, Fetrow JS: **Sequence- and structure-based protein function prediction from genomic information.** *Curr Opin Drug Discov Devel* 2001, **4(3)**:291-295.
6. Pal D, Eisenberg D: **Inference of protein function from protein structure.** *Structure* 2005, **13**:121-130 [http://dx.doi.org/10.1016/j.str.2004.10.015].
7. Lee D, Redfern O, Orengo C: **Predicting protein function from sequence and structure.** *Nat Rev Mol Cell Biol* 2007, **8(12)**:995-1005 [http://dx.doi.org/10.1038/nrm2281].
8. Hawkins T, Chitale M, Luban S, Kihara D: **PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data.** *Proteins* 2009, **74(3)**:566-582 [http://dx.doi.org/10.1002/prot.22172].
9. Chitale M, Hawkins T, Park C, Kihara D: **ESG: extended similarity group method for automated protein function prediction.** *Bioinformatics* 2009, **25(14)**:1739-1745 [http://dx.doi.org/10.1093/bioinformatics/btp309].
10. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17)**:3389-3402.
11. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A: **Protein function annotation by homology-based inference.** *Genome Biol* 2009, **10(2)**:207 [http://dx.doi.org/10.1186/gb-2009-10-2-207].
12. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE: **Protein molecular function prediction by Bayesian phylogenomics.** *PLoS Comput Biol* 2005, **1(5)**:e45 [http://dx.doi.org/10.1371/journal.pcbi.0010045].
13. Ranea JAG, Yeats C, Grant A, Orengo CA: **Predicting protein function with hierarchical phylogenetic profiles: the Gene3D Phylo-Tuner method applied to eukaryotic genomes.** *PLoS Comput Biol* 2007, **3(11)**:e237 [http://dx.doi.org/10.1371/journal.pcbi.0030237].
14. Forslund K, Sonnhammer ELL: **Predicting protein function from domain content.** *Bioinformatics* 2008, **24(15)**:1681-1687 [http://dx.doi.org/10.1093/bioinformatics/btn312].
15. Llewellyn R, Eisenberg DS: **Annotating proteins with generalized functional linkages.** *Proc Natl Acad Sci USA* 2008, **105(46)**:17700-17705 [http://dx.doi.org/10.1073/pnas.0809583105].
16. Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T: **Assessment of prediction accuracy of protein function from protein-protein interaction data.** *Yeast* 2001, **18(6)**:523-531 [http://dx.doi.org/10.1002/yea.706].
17. Chen XW, Liu M, Ward R: **Protein function assignment through mining cross-species protein-protein interactions.** *PLoS One* 2008, **3(2)**:e1562 [http://dx.doi.org/10.1371/journal.pone.0001562].
18. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, other: **A protein interaction map of Drosophila melanogaster.** *Science* 2003, **302(5651)**:1727-1736 [http://dx.doi.org/10.1126/science.1090289].
19. Schwikowski B, Uetz P, Fields S: **A network of protein-protein interactions in yeast.** *Nat Biotechnol* 2000, **18(12)**:1257-1261 [http://dx.doi.org/10.1038/82360].
20. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T: **Identifying functional modules in protein-protein interaction networks: an integrated exact approach.** *Bioinformatics* 2008, **24(13)**:i223-i231 [http://dx.doi.org/10.1093/bioinformatics/btn161].
21. Spirin V, Mirny LA: **Protein complexes and functional modules in molecular networks.** *Proc Natl Acad Sci USA* 2003, **100(21)**:12123-12128 [http://dx.doi.org/10.1073/pnas.2032324100].
22. Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, Ideker T: **Conserved pathways within bacteria and yeast as revealed by global protein network alignment.** *Proc Natl Acad Sci USA* 2003, **100(20)**:11394-11399 [http://dx.doi.org/10.1073/pnas.1534710100].
23. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T: **Conserved patterns of protein interaction in multiple species.** *Proc Natl Acad Sci USA* 2005, **102(6)**:1974-1979 [http://dx.doi.org/10.1073/pnas.0409522102].
24. Jaeger S, Leser U: **High-Precision Function Prediction using Conserved Interactions.** In *Proceedings of the German Conference on Bioinformatics, GCB 2007, September 26-28, 2007, Potsdam, Germany, Volume 115 of LNI* Edited by: Falter C, Schliep A, Selbig J, Vingron M, Walther D, GI 2007, 146-162.
25. Dutkowski J, Tiuryn J: **Identification of functional modules from conserved ancestral protein-protein interactions.** *Bioinformatics* 2007, **23(13)**:i149-i158 [http://dx.doi.org/10.1093/bioinformatics/btm194].
26. Deng M, Zhang K, Mehta S, Chen T, Sun F: **Prediction of protein function using protein-protein interaction data.** *J Comput Biol* 2003, **10(6)**:947-960 [http://dx.doi.org/10.1089/106652703322756168].
27. Chua HN, Sung WK, Wong L: **Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions.** *Bioinformatics* 2006, **22(13)**:1623-1630 [http://dx.doi.org/10.1093/bioinformatics/btl145].
28. Huynen MA, Snel B, von Mering C, Bork P: **Function prediction and protein networks.** *Curr Opin Cell Biol* 2003, **15(2)**:191-198.
29. Vazquez A, Flammini A, Maritan A, Vespignani A: **Global protein function prediction from protein-protein interaction networks.** *Nat Biotechnol* 2003, **21(6)**:697-700 [http://dx.doi.org/10.1038/nbt825].
30. Sun S, Zhao Y, Jiao Y, Yin Y, Cai L, Zhang Y, Lu H, Chen R, Bu D: **Faster and more accurate global protein function assignment from protein interaction networks using the MFGO algorithm.** *FEBS Lett* 2006, **580(7)**:1891-1896 [http://dx.doi.org/10.1016/j.febslet.2006.02.053].
31. Bader GD, Hogue CWV: **An automated method for finding molecular complexes in large protein interaction networks.** *BMC Bioinformatics* 2003, **4**:2.
32. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402(6761 Suppl)**:C47-C52 [http://dx.doi.org/10.1038/35011540].
33. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5(2)**:101-113 [http://dx.doi.org/10.1038/nrg1272].
34. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417(6887)**:399-403 [http://dx.doi.org/10.1038/nature750].
35. Deng M, Sun F, Chen T: **Assessment of the reliability of protein-protein interactions and protein function prediction.** *Pac Symp Biocomput* 2003, 140-151.
36. Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biol* 2006, **7(11)**:120 [http://dx.doi.org/10.1186/gb-2006-7-11-120].

37. Song J, Singh M: **How and when should interactome-derived clusters be used to predict functional modules and protein function?** *Bioinformatics* 2009, **25(23)**:3143-3150 [http://dx.doi.org/10.1093/bioinformatics/btp551].
38. Punta M, Ofran Y: **The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function.** *PLoS Comput Biol* 2008, **4(10)**:e1000160 [http://dx.doi.org/10.1371/journal.pcbi.1000160].
39. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M: **Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs".** *Genome Res* 2001, **11(12)**:2120-2126 [http://dx.doi.org/10.1101/gr.205301].
40. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.
41. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, , **32 Database:** D449-D451.
42. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R: **IntAct: an open source molecular interaction database.** *Nucleic Acids Res* 2004, , **32 Database:** D452-D455 [http://dx.doi.org/10.1093/nar/gkh052].
43. Bader GD, Betel D, Hogue CWV: **BIND: the Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2003, **31**:248-250.
44. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stümpflen V, Mewes HW, Ruepp A, Frishman D: **The MIPS mammalian protein-protein interaction database.** *Bioinformatics* 2005, **21(6)**:832-834 [http://dx.doi.org/10.1093/bioinformatics/bti115].
45. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, *et al*: **Development of human protein reference database as an initial platform for approaching systems biology in humans.** *Genome Res* 2003, **13(10)**:2363-2371 [http://dx.doi.org/10.1101/gr.1680803].
46. Chatr-aryamontri A, Ceol A, Montecchi-Palazzi L, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular INTeraction database.** *Nucleic Acids Research* 2007, , **35 Database:** 572-574 [http://dx.doi.org/10.1093/nar/gkl950].
47. Stark C, Breitkreutz BK, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets.** *Nucleic Acids Research* 2006, , **34 Database:** 535-539 [http://dx.doi.org/10.1093/nar/gkj109].
48. Futschik ME, Chaurasia G, Herzel H: **Comparison of human protein-protein interaction maps.** *Bioinformatics* 2007, **23(5)**:605-611 [http://dx.doi.org/10.1093/bioinformatics/btl683].
49. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, *et al*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2008, , **36 Database:** D13-D21 [http://dx.doi.org/10.1093/nar/gkm1000].
50. Mulder NJ, Apweiler R: **The InterPro database and tools for protein domain analysis.** *Curr Protoc Bioinformatics* 2008, **Chapter 2**:Unit 2.7 [http://dx.doi.org/10.1002/0471250953.bi0207s21].
51. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31**:365-370.
52. FlyBase Consortium: **The FlyBase database of the Drosophila genome projects and community literature.** *Nucleic Acids Res* 2003, **31**:172-175.
53. Bult JCarol, Eppig TJanan, Kadin AJames, Richardson EJoel, Blake AJudith, Mouse Genome Database Group: **The Mouse Genome Database (MGD): mouse biology and model systems.** *Nucleic Acids Res* 2008, , **36 Database:** D724-D728.
54. Twigger NSimon, Shimoyama Mary, Bromberg Susan, Kwitek EAnne, Jacob JHoward, Rat Genome Database Team: **The Rat Genome Database, update 2007-easing the path from disease to data and back again.** *Nucleic Acids Res* 2007, , **35 Database:** D658-D662.
55. Hong EL, Balakrishnan R, Dong Q, Christie KR, Park J, Binkley G, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Krieger CJ, Livstone MS, Miyasato SR, Nash RS, Oughtred R, Skrzypek MS, Weng S, Wong ED, Zhu KK, Dolinski K, Botstein D, Cherry JM: **Gene Ontology annotations at SGD: new data sources and annotation methods.** *Nucleic Acids Res* 2008, , **36 Database:** D577-D581 [http://dx.doi.org/10.1093/nar/gkm909].
56. Bieri T, Blasiar D, Ozersky P, Antoshechkin I, Bastiani C, Canaran P, Chan J, Chen N, Chen WJ, Davis P, Fiedler TJ, Girard L, Han M, Harris TW, Kishore R, Lee R, McKay S, Müller HM, Nakamura C, Petcherski A, Rangarajan A, Rogers A, Schindelman G, Schwarz EM, Spooner W, Tuli MA, Auken KV, Wang D, Wang X, Williams G, *et al*: **WormBase: new content and better access.** *Nucleic Acids Res* 2007, , **35 Database:** D506-D510 [http://dx.doi.org/10.1093/nar/gkl818].
57. Dolinski K, Botstein D: **Orthology and functional conservation in eukaryotes.** *Annu Rev Genet* 2007, **41**:465-507 [http://dx.doi.org/10.1146/annurev.genet.40.110405.090439].
58. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13(9)**:2178-2189 [http://dx.doi.org/10.1101/gr.1224503].
59. Chen F, Mackey AJ, Vermunt JK, Roos DS: **Assessing performance of orthology detection strategies applied to eukaryotic genomes.** *PLoS ONE* 2007, **2(4)**:e383 [http://dx.doi.org/10.1371/journal.pone.0000383].
60. Koyutürk M, Grama A, Szpankowski W: **An efficient algorithm for detecting frequent subgraphs in biological networks.** *Bioinformatics* 2004, **20(Suppl 1)**:i200-i207 [http://dx.doi.org/10.1093/bioinformatics/bth919].
61. Couto FM, Silva MJ, Pedro Coutinho PM: **Measuring semantic similarity between Gene Ontology terms.** *Data Knowl Eng* 2007, **61**:137-152.
62. Chua HN, Sung WK, Wong L: **Using indirect protein interactions for the prediction of Gene Ontology functions.** *BMC Bioinformatics* 2007, **8(Suppl 4)**:S8 [http://dx.doi.org/10.1186/1471-2105-8-S4-S8].
63. Yamada T, Bork P: **Evolution of biomolecular networks: lessons from metabolic and protein interactions.** *Nat Rev Mol Cell Biol* 2009, **10(11)**:791-803 [http://dx.doi.org/10.1038/nrm2787].
64. Zhou X, Kao MCJ, Wong WH: **Transitive functional annotation by shortest-path analysis of gene expression data.** *Proc Natl Acad Sci USA* 2002, **99(20)**:12783-12788 [http://dx.doi.org/10.1073/pnas.192159399].
65. Tao Y, Sam L, Li J, Friedman C, Lussier YA: **Information theory applied to the sparse gene ontology annotation network to predict novel gene function.** *Bioinformatics* 2007, **23(13)**:i529-i538 [http://dx.doi.org/10.1093/bioinformatics/btm195].
66. Saeed R, Deane C: **An assessment of the uses of homologous interactions.** *Bioinformatics* 2008, **24(5)**:689-695 [http://dx.doi.org/10.1093/bioinformatics/btm576].
67. Jiricny J: **MutLalpha: at the cutting edge of mismatch repair.** *Cell* 2006, **126(2)**:239-241 [http://dx.doi.org/10.1016/j.cell.2006.07.003].
68. Kumar SR, Scehnet JS, Ley EJ, Singh J, Krasnoperov V, Liu R, Manchanda PK, Ladner RD, Hawes D, Weaver FA, Beart RW, Singh G, Nguyen C, Kahn M, Gill PS: **Preferential induction of EphB4 over EphB2 and its implication in colorectal cancer progression.** *Cancer Res* 2009, **69(9)**:3736-3745 [http://dx.doi.org/10.1158/0008-5472.CAN-08-3232].
69. Li GM: **Mechanisms and functions of DNA mismatch repair.** *Cell Res* 2008, **18**:85-98 [http://dx.doi.org/10.1038/cr.2007.115].
70. Jiricny J: **Mediating mismatch repair.** *Nat Genet* 2000, **24**:6-8 [http://dx.doi.org/10.1038/71698].
71. Habraken Y, Sung P, Prakash L, Prakash S: **Enhancement of MSH2-MSH3-mediated mismatch recognition by the yeast MLH1-PMS1 complex.** *Curr Biol* 1997, **7(10)**:790-793.
72. Yoshioka K, Yoshioka Y, Hsieh P: **ATR kinase activation mediated by MutSalpha and MutLalpha in response to cytotoxic O6-methylguanine adducts.** *Mol Cell* 2006, **22(4)**:501-510 [http://dx.doi.org/10.1016/j.molcel.2006.04.023].
73. Baker SM, Plug AW, Prolla TA, Bronner CE, Harris AC, Yao X, Christie DM, Monell C, Arnheim N, Bradley A, Ashley T, Liskay RM: **Involvement of**

mouse Mlh1 in DNA mismatch repair and meiotic crossing over. *Nat Genet* 1996, **13(3)**:336-342 [http://dx.doi.org/10.1038/ng0796-336].
74. Hall MC, Shcherbakova PV, Kunkel TA: **Differential ATP binding and intrinsic ATP hydrolysis by amino-terminal domains of the yeast Mlh1 and Pms1 proteins.** *J Biol Chem* 2002, **277(5)**:3673-3679 [http://dx.doi.org/10.1074/jbc.M106120200].
75. Guarne A, Junop MS, Yang W: **Structure and function of the N-terminal 40 kDa fragment of human PMS2: a monomeric GHL ATPase.** *EMBO J* 2001, **20(19)**:5521-5531 [http://dx.doi.org/10.1093/emboj/20.19.5521].
76. Hsieh P, Yamane K: **DNA mismatch repair: molecular mechanism, cancer, and ageing.** *Mech Ageing Dev* 2008, **129(7-8)**:391-407 [http://dx.doi.org/10.1016/j.mad.2008.02.012].
77. Gibson SL, Narayanan L, Hegan DC, Buermeyer AB, Liskay RM, Glazer PM: **Overexpression of the DNA mismatch repair factor, PMS2, confers hypermutability and DNA damage tolerance.** *Cancer Lett* 2006, **244(2)**:195-202 [http://dx.doi.org/10.1016/j.canlet.2005.12.009].
78. Erdeniz N, Nguyen M, Deschenes SM, Liskay RM: **Mutations affecting a putative MutLα endonuclease motif impact multiple mismatch repair functions.** *DNA Repair (Amst)* 2007, **6(10)**:1463-1470 [http://dx.doi.org/10.1016/j.dnarep.2007.04.013].
79. Stone JE, Petes TD: **Analysis of the proteins involved in the in vivo repair of base-base mismatches and four-base loops formed during meiotic recombination in the yeast Saccharomyces cerevisiae.** *Genetics* 2006, **173(3)**:1223-1239 [http://dx.doi.org/10.1534/genetics.106.055616].
80. Ikegaki N, Tang XX, Liu XG, Biegel JA, Allen C, Yoshioka A, Sulman EP, Brodeur GM, Pleasure DE: **Molecular characterization and chromosomal localization of DRT (EPHT3): a developmentally regulated human protein-tyrosine kinase gene of the EPH family.** *Hum Mol Genet* 1995, **4(11)**:2033-2045.
81. Birgbauer E, Oster SF, Severin CG, Sretavan DW: **Retinal axon growth cones respond to EphB extracellular domains as inhibitory axon guidance cues.** *Development* 2001, **128(15)**:3041-3048.
82. Himanen JP, Nikolov DB: **Eph receptors and ephrins.** *Int J Biochem Cell Biol* 2003, **35(2)**:130-134.
83. Sturz A, Bader B, Thierauch KH, Glienke J: **EphB4 signaling is capable of mediating ephrinB2-induced inhibition of cell migration.** *Biochem Biophys Res Commun* 2004, **313**:80-88.
84. Pasquale EB: **Eph receptor signalling casts a wide net on cell behaviour.** *Nat Rev Mol Cell Biol* 2005, **6(6)**:462-475 [http://dx.doi.org/10.1038/nrm1662].
85. Brambilla R, Klein R: **Telling axons where to grow: a role for Eph receptor tyrosine kinases in guidance.** *Mol Cell Neurosci* 1995, **6(6)**:487-495 [http://dx.doi.org/10.1006/mcne.1995.0001].
86. Dickson BJ: **Molecular mechanisms of axon guidance.** *Science* 2002, **298(5600)**:1959-1964 [http://dx.doi.org/10.1126/science.1072165].
87. Huot J: **Ephrin signaling in axon guidance.** *Prog Neuropsychopharmacol Biol Psychiatry* 2004, **28(5)**:813-818 [http://dx.doi.org/10.1016/j.pnpbp.2004.05.025].

doi:10.1186/1471-2164-11-717
Cite this article as: Jaeger *et al.*: Combining modularity, conservation, and interactions of proteins significantly increases precision and coverage of protein function prediction. *BMC Genomics* 2010 **11**:717.