High-Precision Function Prediction using Conserved Interactions

S. Jaeger[1], U. Leser[1]

[1]Knowledge Management in Bioinformatics, Humboldt-University Berlin, Unter den Linden 6, 10099 Berlin, Germany.

Abstract: The recent availability of large data sets of protein- protein-interactions (PPIs) from various species offers new opportunities for functional genomics and proteomics. We describe a method for exploiting conserved and connected subgraphs (CCSs) in the PPI networks of multiple species for the prediction of protein function. Structural conservation is combined with functional conservation using a GeneOntology-based scoring scheme. We applied our method to the PPI networks of five species, i.e., E. coli, D. melanogaster, M. musculus, H. sapiens and S. cerevisiae. We detected surprisingly large CCSs for groups of three species but not beyond. A manual analysis of the biological coherence of exemplary subgraphs strongly supports a close relationship between structural and functional conservation. Based on this observation, we devised an algorithm for function prediction based on CCS. Using our method, for instance, we predict new functional annotations for human based on mouse proteins with a precision of 70%.

1 Introduction

Protein-protein interactions (PPIs) are fundamental for all biological processes. They are crucial to cellular function especially in signal transduction, metabolism and assembly of functional protein complexes. The development of high-throughput technologies for detecting PPI, such as mass spectrometry [AM03], co-immunoprecipitation [PF95], and yeast two-hybrid [WV01], led to an enormous increase of data in this area, which offers new opportunities and new challenges for understanding cellular functions, evolutionary changes, and diseases. For instance, network comparisons help to identify common motives of cellular interactions and may yield information about the presence and organization of functional modules [SM03].

Especially the analysis of conserved interactions in the networks of different species (also called interlogs) has recently received considerable attention [SI06]. These studies are based on the hypothesis that structural conservation correlates with functional conservation. Thus, if two orthologous proteins have interactions with the same set of other orthologous proteins, it may be speculated that these two proteins also have the same function. This idea was, for instance, used by Sharan *et al.* who compared the networks of worm, yeast, and fly and derived new annotations with an estimated precision of around 60% [Se05].

In this paper, we report on initial results on increasing the precision of function prediction based on structural conservation that would also be applicable to human genes. We present a strategy for exploiting the existence of highly conserved subgraphs within the PPI networks of multiple species. Our approach first identifies groups of orthologous proteins in multiple species using sequence similarity. Next, clusters of topologically conserved and connected subgraphs (CCSs) are detected using a fast algorithm for network comparisons. We manually analyzed some of these subgraphs, which includes a giant component of 327 proteins and 413 interactions perfectly conserved between human and mice, for their biological coherence. We could sustain the hypothesis that structural conservation of interactions strongly correlates with functional similarity. Based on this observation, we devised an algorithm which predicts, for proteins with missing annotations, GO-terms [Ae00] based on the annotation of other proteins in the same CCS. We evaluated our approach using perturbation analysis and cross-validation for all CCSs which have a certain level of functional coherence of their annotations. Using our method, we can predict, for instance, 55 new annotations for human proteins based on mouse with an estimated precision of 70%.

Our method is different from previous approaches to function prediction using structural conservation of networks in a number of points. The publicly available tool PathBLAST [Ke04] searches for high-scoring pathway alignments between pairs of interaction paths, where proteins of one path are paired with putative orthologous proteins occurring in the same order in a second path. Evolutionary variations, missing data, and experimental errors are taken into account by allowing gaps and mismatches in the alignment. This method is different to our approach in three aspects: First, PathBlast considers only linear pathways, while we also find complex subgraphs. Second, PathBlast is able to detect approximately conserved paths, while we only strive for perfectly conserved subgraphs. Finally, PathBlast can only analyze pairs of networks, while we work on up to five species. We strongly believe that our approach is more suitable to guarantee high precision for function prediction – at the cost of lower recall.

Sharan *et. al* proposed an algorithm for the detection of conserved structures among three species which they also used for function prediction [Se05]. They build a graph of nodes representing groups of sequence-similar proteins and edges representing conserved interactions between those protein groups. A search over the alignment graph results in linear paths of interacting proteins and dense cluster of interactions, modeling protein complexes. The search algorithm exhaustively detects high-scoring subnetwork seeds and expands them in a greedy fashion. The prediction of protein function is based on the assumption that a conserved subnetwork in which many proteins are of the same function suggests that the remaining proteins in the subnetwork will also have this function. They identified a high number of approximately conserved subgraphs resulting in function predictions with precisions between 33% and 63% depending on species and ontology. In contrast to Sharan *et al.*, our prediction is based on annotations of orthologous proteins and thus transfers annotation across species. Also, GO terms are transferred specifically between orthologous protein pairs, not in between entire CCSs. Again, our method results in a lower number of predictions but in a higher precision.

2 Methods

Data We used the PPI networks of *Escherichia coli*, *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae* and *Drosophila melanogaster*. Protein-protein interaction data were obtained from the Database of Interacting Proteins [Se04], BIND [BBH03], Mammalian MIPS [Pe05], IntAct [He04] (all April '06) and the Human Protein Reference Database [Pe03] (August '06). The data sets were merged based on UniProt-IDs. Information about the experimental evidence underlying an interaction was ignored (see Discussion). The total numbers of proteins and interactions of the different species are shown in Table 1 as well as the median number of GO terms per protein. The evolutionary relationships between these species, which should have an impact on the degree of conservation between interactions, are shown in Figure 1.

Table 1: Total number of proteins and interactions, and the median number of GO terms per protein for five species.

Species	# Proteins	# PPIs	GO terms/ Protein
E. coli	2235	7703	0
S. cerevisiae	5864	25527	3
H. sapiens	9695	34979	2
M. musculus	3247	3116	2
D. melanogaster	10232	41332	1

We also used protein sequences and protein domain information (for determining groups of orthologous proteins) and functional annotations (for measuring functional conservation and coherence). We obtained 31273 protein sequences, 6760 InterPro domains, and 6370 Gene Ontology annotations from the UniProt Knowledge Base [Be03].

Detection of Orthologous Protein Groups We used a rather simple two-step method for detecting putative orthology between two or more proteins (see Discussion). In the first step, we consider proteins enzyme classes, InterPro domains, and sequence alignments to detect potential orthology relationships. In a second step, we apply a global optimization algorithm for determining the optimal assignment of proteins to ortholog groups.

We first partitioned proteins into groups of potential orthologs using EC numbers and InterPro domains. EC numbers form a numerical classification system divided into six classes that characterize enzymes based on the chemical reactions they catalyze. Thus, proteins of different EC classes cannot share the same function and we considered them as non-orthologous. Furthermore, we assumed that orthologs must share at least one InterPro domain. The sequences of all proteins within one partition were compared to each other using global sequence alignment [NW70]. If the sequence identity was above 40%, we saved this value and considered the pair of proteins as potentially ortholog. We also performed the same procedure for all pairs of proteins with missing EC classes and/or InterPro information.

In the second step, we formed a graph from the proteins and their potential relationships. Edges are labeled with the percentage of sequence identify between the proteins they connect. This graph is bipartite if only two species are considered and multipartite if three or more species are considered, because the proteins of one species are never connected by edges. Thus, finding an assignment of nodes to ortholog groups which maximizes the overall sequence identity is an instance of the linear or multidimensional assignment problem (LAPs/MAPs), respectively.

For two species, the LAP (also known as bipartite matching) can be solved optimally in $O(n^3)$ using the Hungarian algorithm [Kuh55] (where n is the number of nodes of the larger partition). For more than two species, the problem is NP-complete and can only be solved approximately for non-trivial instances [BCS94, BS91]. In those cases, we used an iterative greedy heuristic to assign proteins into orthologous groups. We first order the species according to their phylogenetic relationship (see Fig. 1). We randomly chose a protein X_1 from the first species and find the protein X_2 with maximum sequence similarity to X_1 in the second species, and so on until X_5. Once such a complete group of proteins was identified, all involved proteins were removed from the graph and the group was stored as orthologous. If the search did not find an ortholog in one of the steps (because in the next species no protein with sufficient similar sequence existed), we also removed the proteins but formed no orthologous group. This was repeated until all proteins of any of the species involved had been removed.

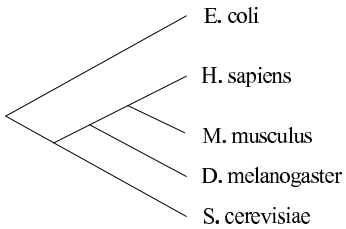


Figure 1: Phylogenetic tree of the considered species.

Detection of CCS Based on the computed groups of orthologous proteins, we used a modification of the algorithm for frequent subgraph discovery described in [KAGS04] for detecting CCSs. Our procedure consists of two steps:

1. Identification of conserved interactions.
2. Assembly of conserved interactions to maximally connected subgraphs.

For two species, we first identified all pairs of interactions between proteins X_1, Y_1 in the first species and X_2, Y_2 in the second species were X_1 and X_2 and Y_1 and Y_2 were previously identified as orthologs. Thus, only interactions contained in both PPI graphs were considered further. Out of this set, one interaction was chosen as seed edge of a subgraph and all interactions adjacent to this subgraph were added recursively. If a subgraph could

not be extended further we stored this maximal and connected subgraph as CCS. All its interactions were removed from the set of interactions and the procedure started over again, until no more interactions remained.

For more than two species, the same procedure was applied but only interactions present in all species were considered.

Assessing the Functional Coherence of CCSs To assess the functional conservation of the detected CCSs we analyzed the GO annotations of the orthologous proteins in each CCS using the method described in [Lin98], calculating a separate score for each of the three GO ontologies (processes, function and localization). Lin's approach is based on the information content IC of a GO term t defined as:

$$IC(t) = -log\left(\frac{freq(t)}{freq(root)}\right), \tag{1}$$

where the frequency of a term is defined as the number of times a term or any of its descendants occurs. Thus, less frequent terms and terms with few occurring descendants are considered more informative.

Based on this measure, the semantic similarity between two terms is defined as the ratio of the information content of their most informative common ancestor and the sum of the information contents of both terms [Lin98]. The information content of the most informative common ancestor is given by:

$$shareIC(t_1, t_2) = max\left\{IC(t)|t \in CA(t_1, t_2)\right\}, \tag{2}$$

where $CA(t_1, t_2)$ is the set of all common ancestors between terms t_1 and t_2. The similarity score between two terms is then defined as:

$$sim(t_1, t_2) = \frac{2 * shareIC(t_1, t_2)}{IC(t_1) + IC(t_2)}. \tag{3}$$

We want to apply this idea to score CCS based on the annotations of the proteins in each orthologous protein groups. Since often proteins are annotated with more than one term the similarity $Sim(p, g)$ of a protein p to a group g of terms is defined as the average similarity of its terms to their most similar terms in g [CSC07] (where $t(p)$ means the set of terms annotated to protein p):

$$Sim(p, g) = \frac{\sum\limits_{t_1 \in t(p)} max\left\{sim(t_1, t_2)|t_2 \in g\right\}}{|t(p)|} \tag{4}$$

We use the definition of Couto *et al.* [CSC07] to define the GO similarity between two proteins is the average similarity of their GO terms [CSC07] :

$$GOSim(p_1, p_2) = \frac{Sim(p_1, t(p_2)) + Sim(p_2, t(p_1))}{2} \tag{5}$$

For orthologous protein groups with more than two proteins we considered all protein pairings within the group. We added all pairwise protein similarities and divided the sum by the number n of protein comparisons within the group c ($n = \frac{k*(k-1)}{2}$ and $k = |c|$), which yields an average score for c:

$$GOSim(c) = \frac{\sum_{i,j(i<j)}^{k} GOSim(p_i, p_j)}{n} \tag{6}$$

Finally, we derive a score for a CCS, which consists of multiple orthologous groups. All single protein scores of each group are added and divided by the number m of groups in the CCS:

$$GOSim(CCS) = \frac{\sum_{c \in CCS} GOSim(c)}{m} \tag{7}$$

Depending on the similarity of the GO annotations of the groups of a CCS, this score ranges between 0 and 1, whereby 1 indicated functional equality and 0 indicates maximal functional distance. CCSs lacking protein annotations result in a decreased semantic similarity due to missing annotations.

Whenever we wanted to characterize the function of a CCS, we used a hypergeometric test to determine statistically overrepresented GO terms.

Predicting Protein Functions For function prediction, we considered all CCSs with at least three interactions and a similarity score exceeding a variable threshold. For qualifying CCSs we determined orthologous protein groups that differ significantly in their single protein scores from the average similarity score of the CCS using a standardized z-score [FPP98].

In particular, we are interested in those proteins which caused the significant difference (p-value < 0.01). The reasons might be different or missing/unknown annotations. For those proteins, we predicted the functions annotated to their orthologous partner proteins. As a baseline, we also considered transferring annotations to orthologs without filtering for conserved interactions first (see Results).

3 Results

We performed several multiple and pairwise PPI network comparisons among the species *E. coli* (EC), *H. sapiens*, *M. musculus* (MM), *S. cerevisiae* (SC) and *D. melanogaster* (DM). The detected CCSs differ considerable in number and size, depending on the number of species considered, their evolutionary distances, and the amount of PPI data available (see Tables 2 and 3). Note that especially the mouse data set is very small, compared to their closest relative (human): $\sim$3000 versus $\sim$35000 interactions.

3.1 Comparisons of multiple PPI networks

Within five or four species we detected only a very small number of orthologs and interlogs, probably due to our strict criterion for orthology. Among five species 72 groups of orthologous proteins but only one conserved interaction were detected. For the different combinations of four species the number of orthologous protein groups ranged between 72 and 407 and the number of conserved interactions varied between 0 and 5 (see Table 2). Also for three species, only few conserved interactions are found (see also Table 2), but the number increased considerably compared to the four-species cases. The numbers for HS-MM-DM, HS-MM-SC and HS-DM-SC differ clearly from the rest of the results, showing the only procaryote in our data set has a large impact.

As expected, the outcomes of the different comparisons are roughly in accordance with their evolutionary relationships (see Figure 1). This means that network comparisons of closer related species resulted in a higher number of orthologous protein groups (e.g. HH-MM-DM) than comparisons of more distant species (HS-SC-EC). However, the differences are by no means proportional to the true evolutionary distance due to the incompleteness of PPI data.

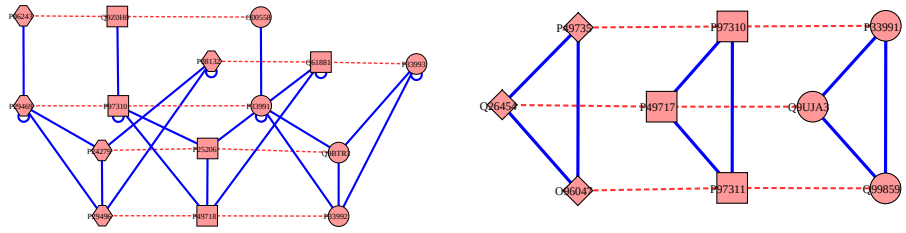
In the following, we briefly discuss results for HS-MM-DM and HS-MM-SC to show the relationship between structural and functional conservation and to hint on the usefulness of our scoring scheme.

Table 2: Results of multiple network comparisons between three and four species.

	# Orthologs	# Conserved PPIs	Largest CCS
HS-MM-DM-SC	407	4	2
HS-MM-DM-EC	87	1	1
HS-MM-SC-EC	67	2	1
HS-DM-SC-EC	139	5	1
MM-DM-SC-EC	72	0	0
HS-MM-DM	1209	22	3
HS-MM-SC	476	20	5
HS-MM-EC	95	7	1
HS-DM-SC	1001	40	3
HS-DM-EC	201	9	1
HS-SC-EC	169	10	1
MM-SC-EC	72	2	1
MM-DM-SC	432	4	2
MM-DM-EC	97	0	0
DM-SC-EC	191	10	1

Comparison between HS, SC, and MM Comparing the PPI data from human, yeast, and mouse resulted in 476 orthologous proteins and 20 conserved interactions. The largest

subgraph is composed of five proteins and eight conserved interactions (see Figure 2(a)).



(a) CCS HS-MM-SC

(b) CCS HS-MM-DM

Figure 2: Largest CCSs among *M. musculus* (rectangles), *H. sapiens* (circles) and *S. cerevisiae* (hexagons) and *D. melanogaster* (rhombus) respectively. (Solid lines represent conserved PPIs within a species. Dotted lines connect orthologous proteins.)

By manually checking the literature, we identified these proteins as four DNA licensing factors *MCM3*, *MCM5*, *MCM7*, *MCM4/MCM2* and a cell division control protein *CDC 7*. DNA licensing factors are proteins which compose protein complexes to regulate the DNA replication. The factors control the start of a replication at the origin of replication and ensure only one DNA replication per cell cycle. Thus, the CCS represents a highly conserved protein complex with a clear biological role important for all species under study. However, the functional similarity scores of the CCS are not very high (0.2, 0.35, and 0.30 for MF, BP, and CC, respectively) due to missing annotation. 3 of 5 proteins in mouse are not annotated at all. When neglecting mouse proteins, the functional similarity increase to 0.59, 0.72 and 0.65 for MF, BP, and CC, respectively. This is important to be kept in mind when our results on cross-validating function prediction are presented.

Comparison between HS, DM, and MM The comparison of human, mouse and fly yielded the largest number of orthologous protein groups among the comparisons of three species. Approximately one third (1209) of the considered mouse proteins had orthologous proteins in human and fly according to our criterion. 22 conserved interactions and one CCS with the required minimum size were detected (see Figure 2(b)).

Again, all proteins of this CCS are DNA licensing factors (*MCM2*, *MCM4*, *MCM6*). The functional coherence scores are higher than in the previous case, but still do not adequately represent to obvious strong functional conservation. Again, this effect was created from missing annotations.

Other comparisons Surprisingly, there were no CCSs detected among human, mouse, fly and yeast comprising those just described DNA licensing factors. The PPI data sets of human and mouse contained the factors *MCM2* - *MCM7*, but for fly (*MCM2*, *MCM4*,

MCM6) and yeast (*MCM2, MCM3, MCM5, CDC7, CDC7 related PK*) only subsets overlapping in only one protein (*MCM2*) were detected. We checked if the missing factors were missing in the data or if they were not identified as orthologs by our method. We found that in fly, our data contains only a short fragment of MCM3 with 35AA (instead of ∼820AA) and that the fly MCM5 homologs was not detected as ortholog. Similarly, the MCM6 factor in yeast was present in the data but was not detected as ortholog. In contrast, the MCM7 protein is missing in the PPI data of fly and yeast.

3.2 Pairwise PPI network comparisons

Table 3 shows the results of all pairwise comparisons. Clearly, the numbers of orthologous proteins are much higher except for comparisons involving *E. coli*. Accordingly, the number of detected conserved interactions and CCSs increased significantly. In particular, the pairs of HS-MM, HS-DM, HS-SC, DM-MM and DM-SC yield many and large CCSs. We discuss two comparisons in more detail.

Pairwise Comparison between HS and SC This comparison yielded 1660 orthologous proteins and 410 conserved interactions forming 129 CCSs. 21 CCSs comprise more than three interactions (see Figure 3). Besides a few linear pathway-like structures the majority of the CCSs among human and yeast are complicated structures with densely-connected components hinting on the formation of protein complexes.

Table 3: Results of pairwise comparisons. The upper-right triangle gives the number of identified orthologous protein, and the lower-left triangle shows the number of highly conserved interactions, the number of detected CCSs, and the size of the largest CCS (in parentheses).

	EC	MM	SC	HS	DM
EC	–	132	437	387	472
MM	11 → 11 (2)	–	647	2879	1580
SC	56 → 52 (3)	35 → 26 (5)	–	1660	1776
HS	65 → 63 (3)	953 → 239 (327)	410 → 129 (28)	–	3260
DM	45 → 44 (3)	57 → 48 (5)	103 → 87 (4)	269 → 210 (6)	–
# proteins in data set	2235	3247	5864	9695	10232

Two examples, chosen by their high functional similarity (scores between 0.45 and 0.74 for MF and BP), are discussed in more detail. The first one, shown in Figure 4, consists of the U6 snRNA-associated Sm-like proteins LSm1, LSm2, LSm3, LSm5/LSm8, LSm7 and a small nuclear ribonucleoprotein Sm D2 binding and stabilizing the spliceosomal U6 snRNA to facilitates the splicing process. Again, the CCS has a clear and conserved function in the cellular machinery. The second one is the largest CCS between human and yeast and involves 28 proteins and 51 conserved interactions. It represents a complex of interacting proteins whose members are DNA-directed RNA polymerases I, II and III,

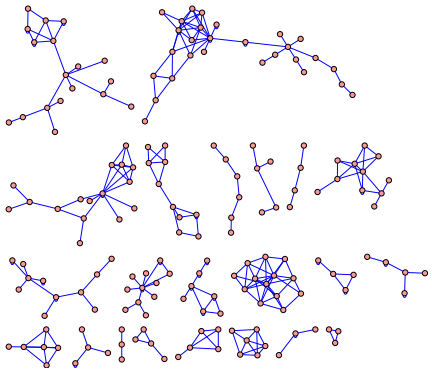


Figure 3: Detected conserved subgraphs between *H. sapiens* and *S. cerevisiae*.

TATA-box-binding proteins and associated factors, and transcription initiation/elongation factors or subunits of those. Thus, the biological relevance of the CCS is closely connected with the initiation of the transcription.

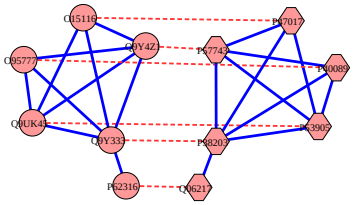


Figure 4: CCS between *H. sapiens* and *S. cerevisiae* representing a protein complex involved in mRNA splicing.

Comparison between HS and MM Among *H. sapiens* and *M. musculus*, 2879 orthologous proteins and 953 conserved interactions were detected forming 239 non-overlapping CCSs (see Figure 5). Beside many small and medium-sized CCSs we detected one astonishing large subgraph composed of 327 proteins and 413 perfectly conserved interactions. This subgraph consists of several linear and complex substructures which may represent independent biological processes. Clearly, such large CCSs are not reasonably analyzed using our method because they certainly perform more than one function. It has been shown that most biological processes involve only between 5 and 25 proteins [SM03].

However, there are also several smaller CCSs consisting of 3 to 4 proteins with high functional similarity scores. The biological processes of these CCSs are fundamental cellular processes such as co-repression of transcription, regulation of signalling pathways, and regulation mechanisms of translation and transcription. Four examples are shown in Figure 6 and their semantic similarities are specified in Table 4.

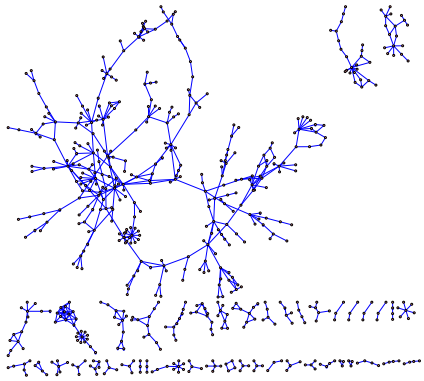


Figure 5: Detected CCSs among human and mouse.

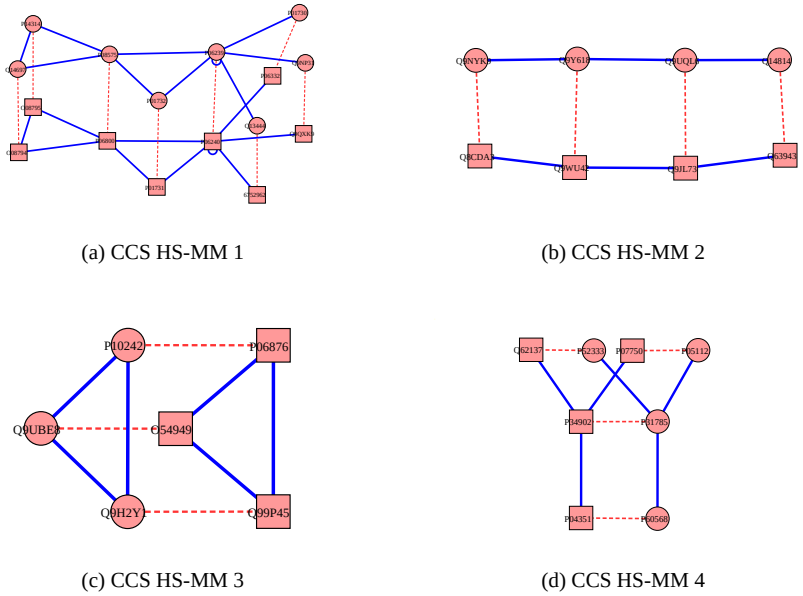
Table 4: Similarity scores of selected CCSs (Fig. 6) between human and mouse.

	MF	BP	CC
HS-MM 1	0.63	0.52	0.42
HS-MM 2	0.92	0.47	0.75
HS-MM 3	0.94	0.82	0.71
HS-MM 4	0.90	0.46	0.73

3.3 Evaluation of Function Prediction

Because of the few conserved interactions for more than two species, our method could only be tested systematically for pairs of species. Comparing three networks results in the prediction of 3 terms in HS-MM-DM and 4 terms in HS-MM-SC, with an estimated precision of 100% for human and mouse in both cases but less for yeast and fly, respectively. The results are somewhat promising, but should be treated with great care since they are based on one subgraph only. We are currently collecting more PPI data sets to evaluate our ideas on more than two species.

Baseline We defined a baseline for the precision of function prediction by using only orthology relationships, ignoring structural conservation. Therefore, we randomly selected 500 orthologous protein pairs, removed annotations from one protein, and predicted its function using only its ortholog. Table 5 shows the results averaged over 100 runs. For human-yeast we obtained a precision of 19% for human and 22% for yeast. For human-mouse, this method yielded a precision of 32% for mouse and 35% for human. Since these numbers are considerable lower than those for predicting function using also structural conservation, we conclude that considering PPI data may considerable enhance the performance of function prediction.



(a) CCS HS-MM 1

(b) CCS HS-MM 2

(c) CCS HS-MM 3

(d) CCS HS-MM 4

Figure 6: Four conserved subgraphs between *H. sapiens* and *M. musculus*.

Table 5: Baseline for function prediction using only orthology.

	Precision	Recall
HS-SC	0.21	0.3
HS	0.19	0.33
SC	0.22	0.28
HS-MM	0.33	0.47
HS	0.35	0.49
MM	0.32	0.46

Cross-Validation We evaluated the expected precision and recall of our method using cross-validation. Therefore, we first removed known annotations from randomly selected proteins and then applied our method. We selected CCS based on their (now probably artificially low) score, using three different thresholds. We predicted GO terms in the selected clusters for outliers as described in Methods. The predicted terms were compared with the original annotations to determine the amount of correctly predicted annotations. We counted an original term as correctly predicted if the proposed term was a direct ancestor or descendant of the original term or the term itself.

In Table 6, the number of predicted terms and their estimated precision and recall for

the pairs HS-SC and HS-MM are given, using similarity thresholds of 0.3, 0.5 and 0.7, respectively. In general, the higher the conservation threshold, the higher is the precision of the prediction. The precision for predicted human terms based on mouse increased from 61% for 0.3 to 70% for 0.5 to 100% for 0.7 – at the expected cost of fewer and fewer predictions. As expected, predictions are more precise for closer related species, such as human and mouse, than for human and yeast.

Table 6: Results of the cross-validation between human, mouse and yeast. The table shows the total number of predicted terms and precision and recall of the predictions for three different thresholds.

	0.3			0.5			0.7		
	# terms	P	R	# terms	P	R	# terms	P	R
total	405	0.36	0.41	200	0.37	0.44	14	0.83	0.71
HS	192	0.34	0.5	90	0.38	0.58	7	0.83	0.71
SC	213	0.37	0.33	110	0.37	0.31	7	0.83	0.71
total	198	0.62	0.67	119	0.77	0.78	6	1	1
HS	95	0.61	0.68	55	0.70	0.8	3	1	1
MM	103	0.64	0.66	64	0.87	0.76	3	1	1
total	10	0.27	0.6						
HS	4	0.15	0.75						
DM	6	0.6	0.5						

3.4 Prediction Results

All together, our method predicts 319 new annotations for 149 different proteins from all species but *E. coli*. Comparisons including the procaryote resulted in subgraphs with less than three interactions and thus do not provide the possibility for function prediction. We therefore predicted GO terms separately for each GO-ontology using a similarity threshold of 0.5 to ensure high precision. The comparisons of HS-SC (14 CCSs), HS-MM (9 CCSs) and HS-DM (3 CCSs) contributed to a different amount to the prediction of novel terms. The number of suggested GO annotations and the number of proteins receiving novel annotations are shown in Table 7, and some concrete examples are shown in Table 8. The complete list of predictions is available in the supplementary Table 9.

4 Discussion

We developed a method for predicting protein function based on structural and functional conservation. Our approach proceeds in several phases. First, putative orthologs are identified using sequence similarity, EC classification, and InterPro domains. Next, conserved and connected subgraphs are detected across multiple species. After filtering for functional

Table 7: Number of newly derived terms and the number of proteins that are annotated.

	# Terms	# Proteins
HS	184	80
MM	24	12
SC	110	52
DM	11	5

Table 8: Examples for suggested GO terms.

UniProt ID	Species	Ontology	Terms
P55010	HS	MF	GO:0005096
P55010	HS	CC	GO:0005843, GO:0043614
Q13526	HS	MF	GO:0003711, GO:0003755, GO:0042802
O76097	HS	BP	GO:0006888, GO:0006901
P33307	SC	CC	GO:0005634, GO:0005737
Q12149	SC	MF	GO:0004674, GO:0042802
P29469	SC	BP	GO:0006260, GO:0007049
P97311	MM	BP	GO:0006260, GO:0007307, GO:0009993
Q9WU42	MM	MF	GO:0003714, GO:0005112
P06876	MM	CC	GO:0016363
Q9VRT7	DM	BP	GO:0000398, GO:0006364, GO:0006402
Q9VLV5	DM	CC	GO:0005682, GO:0005685, GO:0046540
Q02748	DM	MF	GO:0005515

coherence, we can transfer annotations between proteins in the same CCS across species boundaries. Depending on the threshold for coherence, this method reaches an estimated precision between 60% and 100% between human and mouse, which is two to three times better than using only sequence conservation. Our results are consistent with our expectations in several points, such as better and more predictions for evolutionary closer species and better results for higher demands on functional coherence of subgraphs.

We also applied a text mining method to use information about protein function as mentioned in the literature as another way of confirming our predictions (data not shown here). The results are promising, and a subset of the so-predicted protein functions are now under revision of the GO team.

There are also shortcomings concerning the data and the methods. In the following, we discuss these shortcomings and suggest possible improvements.

In general, the numbers of orthologs and conserved interactions were lower than expected, especially when more than 2 species were involved. These low numbers probably have several reasons:

- We used a very strict criteria for deciding orthology. Our threshold of 40% sequence

identity that two proteins must have to be at least considered as potential orthologs is very conservative. The reason for this strictness is our goal to reach high precision in predictions rather than high recall. For test purpose, we performed multiple comparisons with a lower threshold of 20%, which yielded a much larger number of orthologous protein groups (results not shown).

- Also, our demand on CCSs to be perfectly conserved is very strict. This is a critical point, especially when more than two species are compared and when thought together with the current incompleteness of PPI data sets (see next point). A single missing interlogs might destroy networks that in reality are perfectly conserved. On the other hand, demanding perfect conservation might be too strict in itself from an evolutionary point of view. However, we have not yet performed a systematic study which would proof our assumption that more strictness (in terms of structural and sequence conservation) leads to higher precision.

- The incompleteness and the sheer difference in sizes of the PPI data sets from different species naturally hinders the detection of more and larger CCSs. This influence is especially obvious for all comparisons with mouse, for which we only used $\sim$3200 proteins and $\sim$3100 interactions. This tenfold difference compared to the human data set (in terms of interactions) clearly limits our results. It is the more surprising that we detected a gigantic and perfectly conserved subgraph of $\sim$320 proteins and $\sim$400 interlogs, which contains more than 10% of the entire mouse interaction set.

We are currently working on solutions to these shortcoming. First, we are investigating several special orthology databases, such as Inparanoid [ORS05] or COG [Te03], as a potential source for orthologs, rather than computing these ourself. Using a special database could also remove the problem of paralogs, which we essentially ignored. Second, we are scanning additional data sources, especially species-specific databases such as FlyBase [Con97] and MGD [Be99], to increase our data basis. We also strive to include more species into our studies to be able to quantify the relationship between structural conservation and evolutionary distance more precisely. Third, we are currently investigating other algorithms for the Multiple Assignment Problem that would improve on our very simple heuristic.

Apart from still being incomplete, PPI data sets are also somewhat famous for containing many false positives because interactions are often derived under very specific conditions. For example, of $\sim$80000 interactions currently known for yeast that were detected with high-throughput methods only 2400 are supported by more than one method [Me02]. It would be worthwhile to include measures for the reliability of interactions (and thus interlogs) into the scoring scheme. For instance, one could demand that a CCS for two species must completely contain all interactions in any of the two species with a reliability above a given threshold, and be less strict concerning interactions with lower thresholds. This would be a natural way of introducing more flexibility in the definition of CCS.

Finally, our method currently favors groups of proteins that are homogeneously annotated. Therefore, larger subgraphs comprising more than one biological function receive an un-

favorable score. For those, functionally conserved sub-subgraphs within the structurally conserved subgraphs should be studied instead of the entire subgraph.

Acknowledgements

We would like to thank Philip Groth for critical reading of the manuscript and the EBI/Dr. Rebholz-Schuhmann and his group for providing text mining methods to cross-check function predictions.

References

[Ae00] M Ashburner *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, May 2000.

[AM03] R Aebersold and M Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, Mar 2003.

[BBH03] GD Bader, D Betel, and CWV Hogue. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research*, 31(1):248–250, 2003.

[BCS94] HJ Bandelt, Y Crama, and FCR Spieksma. Approximation algorithms for multidimensional assignment problems with decomposable costs. *Discrete Appl. Math.*, 49(1-3):25–50, 1994.

[Be99] JA Blake *et al.* The Mouse Genome Database (MGD): genetic and genomic information about the laboratory mouse. The Mouse Genome Database Group. *Nucleic Acids Research*, 27(1):95–98, 1999.

[Be03] B Boeckmann *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1):365–370, 2003.

[BS91] E Balas and MJ Saltzman. An algorithm for the three-index assignment problem. *Operations Research*, 39(1):150–161, 1991.

[Con97] The FlyBase Consortium. FlyBase: a Drosophila database. *Nucleic Acids Research*, 25(1):63–66, 1997.

[CSC07] FM Couto, M Silva, and P Coutinho. Measuring semantic similarity between Gene Ontology terms. *Data & Knowledge Engineering*, 61(1):137–152, 2007.

[FPP98] D Freedman, R Pisani, and R Purves. *Statistics, Third Edition*. New York: W.W. Norton and Company, 1998.

[He04] H Hermjakob *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Research*, 32(Database-Issue):452–455, 2004.

[KAGS04] M Koyutürk, A Ananth-Grama, and W Szpankowski. An efficient algorithm for detecting frequent subgraphs in biological networks. In *ISMB/ECCB (Supplement of Bioinformatics)*, pages 200–207, 2004.

[Ke04] BP Kelley *et al.* PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res*, 32(Web Server issue):W83–W88, Jul 2004.

[Kuh55] Kuhn HW. The Hungarian Method of Solving the Assignment Problem. *Naval Res. Logistics Quart.*, 2:83–97, 1955.

[Lin98] D Lin. An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th ICML*, pages 296–304, Madison WI, 1998.

[Me02] C Mehring *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, May 2002.

[NW70] SB Needleman and CD Wunsch. A General Method Applicable to the Search for Similarity in the Amino Acid Sequences of Two Proteins. *J. Mol. Biol.*, 48:443–453, 1970.

[ORS05] KP O'Brien, M Remm, and ELL Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, 33(Database issue):D476–D480, Jan 2005.

[Pe03] S Peri *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10):2363–2371, Oct 2003.

[Pe05] P Pagel *et al.* The MIPS mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–834, 2005.

[PF95] EM Phizicky and S Fields. Protein-protein interactions: methods for detection and analysis. *Microbiol Rev*, 59(1):94–123, Mar 1995.

[Se04] L Salwiński *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 32(Database-Issue):449–451, 2004.

[Se05] R Sharan *et al.* Conserved patterns of protein interaction in multiple species. *PNAS*, 102(6):1974–1979, February 2005.

[SI06] R Sharan and T Ideker. Modeling cellular machinery through biological network comparison. *Nat Biotechnol*, 24(4):427–433, Apr 2006.

[SM03] V Spirin and LA Mirny. Protein complexes and functional modules in molecular networks. *PNAS*, 100(21):12123–12128, 2003.

[Te03] RL Tatusov *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, Sep 2003.

[WV01] AJ Walhout and M Vidal. High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods*, 24(3):297–306, Jul 2001.