Describing the orthology signal in a PPI network at a functional, complex level

Pavol Jancura[1], Eleftheria Mavridou[2], Beatriz Pontes[3], and Elena Marchiori[1]

[1] Institute for Computing and Information Sciences, Radboud University Nijmegen, Postbus 9010, 6500 GL Nijmegen, The Netherlands
{jancura,elenam}@cs.ru.nl
[2] Department of Medical Microbiology, Radboud University Medical Center, Postbus 9101, 6500 HB Nijmegen, The Netherlands
[3] Department of Computer Science, University of Seville, Avda. Reina Mercedes s/n 41012 Seville, Spain

Abstract. In recent work, stable evolutionary signal induced by orthologous proteins has been observed in a Yeast protein-protein interaction (PPI) network. This finding suggests more connected subgraphs of a PPI network to be potential mediators of evolutionary information. Because protein complexes are also likely to be present in such subgraphs, it is interesting to characterize the bias of the orthology signal on the detection of putative protein complexes. To this aim, we propose a novel methodology for quantifying the functionality of the orthology signal in a PPI network at a protein complex level. The methodology performs a differential analysis between the functions of those complexes detected by clustering a PPI network using only proteins with orthologs in another given species, and the functions of complexes detected using the entire network or sub-networks generated by random sampling of proteins. We applied the proposed methodology to a Yeast PPI network using orthology information from a number of different organisms. The results indicated that the proposed method is capable to isolate functional categories that can be clearly attributed to the presence of an evolutionary (orthology) signal and quantify their distribution at a fine-grained protein level.

1 Introduction

In general, two proteins are orthologous if they originated from a common ancestor, having been separated in evolutionary time only by a speciation event. Orthologous proteins have high amino acid sequence similarity and usually retain the same or very similar function, which allows one to infer biological information between the proteins. Obviously, orthology as such is very important in studying evolution. Therefore, the problem of establishing proper orthology relations has been under the wide investigation in comparative genomics (see for instance [1]) and many databases and public resources of orthologs have been made available, such as Inparanoid [2] and OrthoMCL-DB[3].

Recent studies used this form of evolutionary information to analyse protein modules and PPI networks, for instance [4–12]. In particular, in a study by Wutchy et al. [6] stable evolutionary signal was found to be present in a Yeast PPI network as examined by its pairwise orthologs with respect to various different species. They observed that a high local clustering around protein-protein interactions correlates with evolutionary conservation of the participating proteins. This means that highly connected proteins and protein pairs embedded in a well clustered neighbourhood tend to be evolutionary conserved and therefore retain their evolutionary signal. These findings suggest also that more connected areas of a PPI network are potential mediators of evolutionary information.

Because more connected regions of PPI networks contain protein modules or complexes, in this paper we focus on the explicit use of orthology to see whether there are functional complexes that can be clearly attributed to this evolutionary signal. To this aim, we try to characterize those functions of complexes predicted by clustering the subgraph of a PPI network induced by all proteins with orthologs in another given species, but not predicted (or predicted for a smaller

fraction of proteins) when clustering the entire network. We consider these functions as strong characterization of the underlying evolutionary signal of orthologs, since they are suppressed or not observed when clustering using the entire network.

Specifically, we examine highly functionally coherent putative protein complexes as detected by two state-of-the-art clustering techniques in the Yeast PPI network using only proteins with orthologs in another given organism. Our target clusters should contain a function which can be genuinely attributed to the orthology signal and exclude the case that it could be attributed by chance. Therefore we consider three classes of clusters, consisting of putative complexes as detected by these clustering techniques applied to the Yeast PPI network with (1) all proteins, (2) only proteins with ortholog in the considered other organism, and (3) randomly selected proteins. The latter class of clusters is the collection of cluster sets produced by the application of clustering to the PPI network induced by a random selection of a set of proteins (of size equal to that of the set of proteins used to generate the class (2)) repeated for dozens times. For all clusters in each class we infer putative functions by measuring their gene ontology (GO) functional enrichment [13].

In general, protein functions belong to certain functional categories. Hence, we map all putative functions inferred from the clusters to these categories. For a set of clusters and a certain category, we compute the fraction of proteins contained in the clusters and having functions mapped to that category. This fraction quantifies (at protein level) the presence of that functional category in a given cluster set. This allows us to identify functional categories whose proteins' fraction is higher in clusters from the class (2) than in clusters from the other two classes. We consider the corresponding clusters in class (2) as describing the orthology signal (with respect to considered species). Furthermore, we analyse those clusters of class (2) having a predicted function for its proteins that is not inferred when using clusters of class (1). Finally we discuss the new meaningful functions for well-defined as well as for unknown proteins that are present in the compilation of putative complexes.

2 Other Related Work

In previous works on phylogenetic analysis of protein networks and complexes evolutionary information was usually used as a mean for evaluating the preservation of orthology information in functional modules [5, 8–10]. Here, however, we incorporate evolutionary information beforehand for detecting evolutionary signal at complex, functional level. Our identification of protein complexes uses only the topology of the network of the considered species and orthology information from another species, without requiring knowledge on the interactome of the other species.

In general, our approach differs from comparative network methods [14], as the latter aim to find evolutionary conserved modules across species, and exploit both orthology and network topology of the considered organisms. The clusters we obtain are in one species and are related to the orthology signal with respect to another species, but are not required to be evolutionary conserved through species (we do not enforce any type of similarity at the graph-structure level). Furthermore, comparative methods mostly do not use 'known' orthologs in available databases but rather they rely on sequence similar proteins, where the level of required similarity is determined by a minimal similarity score threshold. Instead, our method exploits the orthology information available in existing databases.

3 Method

The following terminology is used in the sequel. A PPI network is represented by means of a graph $G(V, E)$, where V is the set of nodes (proteins) and E is the set of edges (binary interactions). Let X be a subset of nodes V (e.g. ortholog set). The set X induces a subgraph $G[X] = (X, E_X)$ of G, with set X of nodes and set E_X of those edges of E that join two nodes in X. For a set S, we denote by $|S|$ the number of its elements.

We are interested in quantifying the orthology signal by means of a set of functions of putative protein complexes detected by applying clustering to a PPI network. To this end, we directly exploit

evolutionary information of proteins as described by the presence of orthologs in another, given species. We call these proteins 'true orthologs'. Specifically, we propose the following methodology.

Given a PPI network $G = (V, E)$ and a given species s, apply the following steps.

1. *Retrieve from a database the set O of 'true orthologs' of V with respect to s, with $|O| = n$.*
2. *Generate the following three classes of clusters, using a given clustering algorithm.*
   (a) *Class 1 clusters (GC). Apply clustering to the whole PPI network G.*
   (b) *Class 2 clusters (OC). Apply clustering to the sub-network induced by O.*
   (c) *Class 3 clusters (RC). Apply clustering to the sub-network induced by a randomly selected subset of V of size n. Repeat the process a number N of times. Consider all sets of clusters detected across these runs ($RC = \{RC_1, RC_2, \ldots, RC_N\}$).*
3. *For each class of clusters,*
   (a) *Infer putative complexes and their functional categories.*
   (b) *For each functional category, compute the fraction of those proteins in the detected complexes which have been assigned to that category.*
4. *Select the set of those functional categories derived using clusters from class 2 and whose fraction are higher than those of the same category derived using clusters from class 1 or from class 3.*
5. *Output the set of clusters having at least one of the selected functional categories.*

In this study we consider as putative protein complex only a group of proteins of a higher complexity than just a single protein-protein interaction. Therefore, after applying any clustering method we retain only clusters of size greater or equal than 3.

In the sequel we describe in more detail the main steps of the proposed methodology.

Inferring Putative Complexes and their Functional Categories. In order to infer the the putative functions of a cluster, we measure the enrichment of functional annotations of the corresponding protein set, as entailed by the gene ontology (GO) annotation [13], using one of the well-established tools, the Ontologizer[1] [15]. The Ontologizer offers various algorithms for measuring GO enrichments. Here, we apply the standard statistical analysis method based on the one-sided Fisher's exact test [15], which measures the statistical significance of an enrichment and assigns to the cluster a p-value for each enriched function. The p-value is further corrected for multiple testing by means of a Bonferroni correction procedure.

The GO is known to have a hierarchical structure (directed acyclic graph) which can be used to define the level of an annotation. Specifically, the level of an annotation is equal to the length of the shortest path from the root of GO hierarchy to the annotation. The GO terms closer to the root of GO give more general description of biological functions while terms closer to the leaves of GO have granular and very specific biological definitions.

Each detected cluster is a potential protein complex. The quality of a protein cluster is given by the coherence of biological functions of proteins contained in the cluster. If a certain subset of proteins in a cluster has a significantly coherent function, a prediction of that function for all proteins in the cluster can be made. We may obtain more than one protein function prediction if we find more significantly coherent functions in the cluster. We say that proteins of a cluster have a *significantly coherent function* or functional GO annotation if the following criteria are satisfied:

1. the GO annotation is significantly enriched by the proteins in the cluster (p-value < 0.001).
2. more than half of the proteins in the cluster has this significant annotation.
3. the annotation is at least at the GO level four from the root of GO hierarchy.

In such a case the cluster can be used as protein function predictor and the significantly enriched GO annotation of the cluster is used to predict protein function of each of the proteins in that cluster. If a cluster does not satisfy the above conditions, no prediction can be made. Similar criteria were used by, e.g. [16, 17]. The condition on GO hierarchy guarantees that the prediction about biological functions is sufficiently specific and informative [18]. Each cluster which is a predictor defines a putative protein complex and the set of significantly coherent functions defines the set of inferred functions.

[1] http://compbio.charite.de/index.php/ontologizer2.html

Estimating the Frequency of a Functional Category. After identifying putative protein complexes, we use them to quantify, at a fine-grained, protein level, the frequency with which a functional category was detected: for each functional category inferred using the putative protein complexes, we count the fraction of those proteins in the putative complexes assigned to that category.

Specifically, functional categories are determined by GO slim functional terms, defined in the GO hierarchy as a subset of the higher level GO terms. Each GO slim characterizes a certain type of biological functions which have some features and tasks in common. As a result, each fine-grained term can be mapped to these GO slims' terms.

The GO also consists of three main independent domains, *biological process*, *molecular function* and *cellular component*, and each of them has its own GO slim terms and hence functional categories. Given a GO domain and proteins of a cluster group of interest one can map all inferred functions of each protein to their closest GO slims in the GO hierarchy. Then for every functional category we can count the number of proteins being mapped to the category. In this framework we define the frequency of a functional category as follows.

Let C be a set of putative complexes and $P(f)$ denote the set of proteins contained in C and being mapped to a functional category f. Let B be a set of background functional categories (functional background). Then *the frequency of a functional category* f in C with respect to the background B is

$$\phi_C(f) = \frac{|P(f)|}{|P(B)|}, \text{ where } P(B) = \bigcup_{\forall b \in B} P(b). \tag{1}$$

Notice that in our definition we consider an individual background for each GO domain.

The frequency of a functional category can be viewed as the expectation that a protein in a given set of putative complexes has that functional category. This results in a distribution of functional categories associated to a set of complexes.

Identifying Orthology-Related Categories. Since we are interested in analysing the evolutionary (orthology) signal, we use as the functional background the set of functional categories enriched by the class 2 putative complexes. Therefore we compute the functional frequencies for each class of putative complexes using this background. For class 3 (random sampling), for each functional category, we average the frequencies over all random simulations as follows

$$\overline{\phi_{RC}}(f) = \frac{\sum_{i=1}^{N} \phi_{RC_i}(f)}{N}. \tag{2}$$

Once the functional frequencies are computed, we isolate functional categories related to the orthology signal by the following simple rule:

A *functional category* f is *orthology-related* iff

$$\phi_{OC}(f) > \max\{\phi_{GC}(f), \overline{\phi_{RC}}(f)\}. \tag{3}$$

4 Experimental Settings

4.1 Data Collection

We chose to perform our analysis on the widely used and well-studied species Saccharomyces cerevisiae (yeast), which PPI network is one of the best established and information on functionality of its proteins are one of the most explored. This makes yeast as a good standard model species for protein network analysis.

We used the same yeast interaction data as in [19] which combines interaction data from DIP [20] and MPact [21], and interactions from the core datasets of the TAP mass spectrometry experiments [22, 23]. This yeast interaction data are weighted by the method proposed by Jansen et al. [24] to measure the confidence of interactome. As a result, the low confidence interactions are ignored and the final yeast PPI network consists of 3545 proteins and 14354 interactions.

For obtaining orthology information we used the Inparanoid Database of Pairwise Ortholog[2] [2]. This database contains clusters of ortholog groups (COGs) constructed by the Inparanoid program, which is a fully automatic method for finding orthologs and in-paralogs between two species. Ortholog clusters in the Inparanoid are seeded with a two-way best pairwise match (the seed ortholog pair), after which an algorithm for adding in-paralogs is applied. The Inparanoid was found as one of the best performing algorithms for orthology detection with respect to its false negative and false positive rates [25].

Because in-paralogs are homologs that arise when duplication occurs after speciation, and the duplicated gene often still retains the function of the ortholog [26], they should be likely found in one protein complex. Therefore we consider all proteins present in COGs for inducing an orthology PPI sub-network and, for simplicity, we consider all proteins in a COG as orthologs. Specifically, further in this study we call orthologous protein or ortholog a protein which is a part of an orthologous cluster produced by the Inparanoid when comparing two species.

In our analysis, COGs were obtained for the following pairs of organisms:

- *Saccharomyces cerevisiae* vs. *Escherichia coli*
- *Saccharomyces cerevisiae* vs. *Caenorhabditis elegans*
- *Saccharomyces cerevisiae* vs. *Drosophila melanogaster*
- *Saccharomyces cerevisiae* vs. *Homo sapiens*

Escherichia coli (E.coli), Caenorhabditis elegans (worm), Drosophila melanogaster (fly) and Homo sapiens (human) are standard organisms used in protein network and genome comparative studies (e.g [27, 28]) and represent the diverse life-forms from a prokaryote (E.coli) to the highly complex eukaryote (Human). Yeast proteins in the derived ortholog groups are called yeast orthologs. We considered the following 4 sets of yeast orthologs (present in the yeast PPI data), namely *Yeast-E.coli, Yeast-Worm, Yeast-Fly, Yeast-Human*, consisting of 451, 1664, 1724, and 1850 number of proteins.

4.2 Yeast Protein Function Annotations and Gene Ontology Files

In order to measure functional enrichments of clusters we used only experimentally verified annotations as reported in the yeast gene association file of Saccharomyces Genome Database[3] (SGD), available at the GO database[4]. We excluded all computationally assigned annotations to yeast proteins to avoid introducing a possible bias, because many of these techniques use protein structural or sequence similarity which may often refer to orthology. GO slims and terms are also available at GO database.

4.3 Clustering

In this study we used two clustering techniques: `SiDeS` and `MCL`. We briefly address their properties:

`MCL` [29] computes clusters based on simulation of stochastic flow in graphs and it is widely used on many domains. It is able to use information on weights of edges of a given network if available. A first successful application of this algorithm on biological networks was presented in [30]; `MCL` was also modified for detecting orthologous groups [31]. A recently published comparative study [32] indicated that `MCL` outperforms other algorithms for clustering PPI networks. The inflation parameter of the algorithm was set to 1.8 as suggested in [32].

`SiDeS` [33], in contrast, is not able to use information on weights of edges. However, the main advantage of `SiDeS` is that it directly addresses the problem of statistical significance of cluster density, based on the topological structure of a PPI network, during computation.

[2] http://inparanoid.sbc.su.se/
[3] http://www.yeastgenome.org/
[4] http://www.geneontology.org/GO.downloads.shtml, SGD version: 1.1523 date: 11/13/2010, GO version: 1.1.1602 date: 16/11/2010, GO Slim version: 1.1.1543 date: 19/10/2010

Thus, all clusters isolated by SiDeS have statistically significant density and therefore the resulting clusters tend to be more biologically relevant than those produced by other methods, albeit fewer in number. SiDeS modifies an existing state-of-the-art graph clustering algorithm, HCS [34], based on recursive partitioning of a graph and incorporating the computation of statistical significance of clusters.

The above two clustering algorithms are different in their basic concepts and combining their results for identifying orthology-related functional categories should effectively minimize the possibility of finding an artefact. Therefore we applied both clustering algorithms on each yeast PPI sub-network induced by every set of yeast orthologs as well as on all yeast PPI sub-networks induced by repeated random protein selection of the same number of proteins as the protein count of a particular yeast ortholog set. We labelled each resulting cluster group as follows:

- OYC-E - yeast clusters found in the sub-network induced by the Yeast-E.coli ortholog set.
- OYC-W - yeast clusters found in the sub-network induced by the Yeast-Worm ortholog set.
- OYC-F - yeast clusters found in the sub-network induced by the Yeast-Fly ortholog set.
- OYC-H - yeast clusters found in the sub-network induced by the Yeast-Human ortholog set.

These groups are of the class (2) and we generally refer to them by the common name OYC. Analogically we also marked cluster groups induced by randomly sampled proteins as follows:

- RYC-E - yeast clusters found in the sub-network induced by random sampled proteins of the same number as the number of proteins in the Yeast-E.coli ortholog set.
- RYC-W - yeast clusters found in the sub-network induced by random sampled proteins of the same number as the number of proteins in the Yeast-Worm ortholog set.
- RYC-F - yeast clusters found in the sub-network induced by random sampled proteins of the same number as the number of proteins in the Yeast-Fly ortholog set.
- RYC-H - yeast clusters found in the sub-network induced by random sampled proteins of the same number as the number of proteins in the Yeast-Human ortholog set.

These groups belong to the class (3) and we generally refer to them by the common name RYC.

When MCL or SiDeS applied on the whole yeast network, we get clusters of the class (1) and we refer to them by the name GYC (general yeast clusters).

We randomly sampled proteins 1000 times for each given number of orthologs. Recall that every run produces one particular RYC group. In order to compare these clusters with GYC or OYC, we always report average values of RYC groups computed over all 1000 simulations according to a given ortholog set (average RYC values).

Tables 1 contains the number of all clusters and corresponding cluster predictors for GYC, all four OYC and average RYC, as identified by MCL and SiDeS.

Table 1. Numbers of Clusters.

	MCL		SiDeS	
Cluster Group	#Clusters	#Predictors	#Clusters	#Predictors
GYC	365	147	122	93
OYC-E	37	14	5	3
RYC-E	34.31 (±3.82)	12.69 (±2.96)	4.71 (±2.08)	3.8 (±1.76)
OYC-W	181	80	66	46
RYC-W	175.22 (±7.21)	67.85 (±5.87)	55.04 (±5.57)	40.32 (±4.54)
OYC-F	191	80	64	51
RYC-F	181.97 (±7.51)	70.32 (±6.01)	57.71 (±5.57)	42.25 (±4.49)
OYC-H	203	90	82	62
RYC-H	196.38 (±7.80)	75.71 (±6.21)	63.42 (±5.67)	46.12 (±4.68)

5 Results

The detected cluster predictors are considered as putative protein complexes and used to identify orthology-related functional categories. For each cluster group of predictors we compute the functional frequencies with respect to the categories enriched by OYC, as explained in Section 3. Figure 1 shows the frequency distribution of GYC, OYC-W and RYC-W clusters as detected by `SiDeS`.

In Tables 4 and 5 (see the Appendix) frequencies are reported as measured by `MCL` and `SiDeS` OYC-W clusters, respectively. Observe that not all orthology-related functional categories are shared by `MCL` or `SiDeS` cluster groups. To minimize the possibility of false positives, we employed a conservative approach and considered as orthology-related functional categories only those identified by both clustering techniques. The results are listed in Table 2.

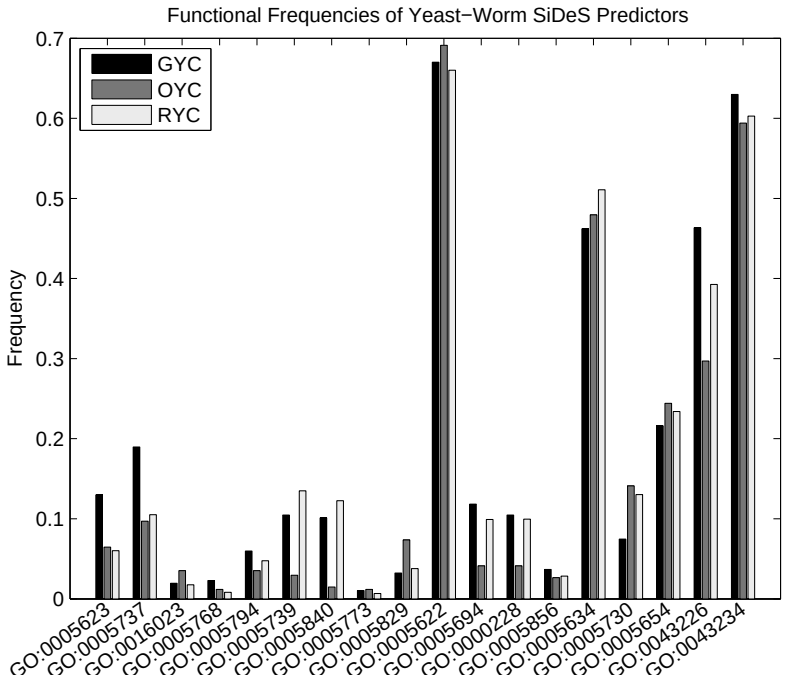

Fig. 1. Functional frequencies for Yeast-Worm orthologs as estimated by `SiDeS` predictors. On x-axis GO ids of GO slim functional categories are reported.

5.1 Orthology-related functional categories

For Yeast-E.coli orthologs, the identified clusters have higher frequencies of ribosomal and mitochondrial proteins. Indeed, it has been shown that the ribosomes in the mitochondria of eukaryotic cells resemble those in bacteria, reflecting the likely evolutionary origin of this organelle [35].

Since worm, fly and human all belong to eukaryotes, we looked at which common functional categories have yeast clusters containing orthologs with respect to these species (reported in Table 2 in boldface). Considering molecular functions, we observed that protein binding proteins and kinases activity proteins are more frequently present in OYC clusters than in GYC clusters or in RYC clusters. Thus these functional categories might be considered as orthology related. This is true in particular for proteins of protein kinase activity, which have been found conserved among eukaryotes: these kinase' functional conservations were investigated for yeast, worm, fly

and human when studying their evolution [36]. Moreover, kinases' proteins are known to regulate the majority of cellular pathways, especially those involved in signal transduction. As we may see, signal transduction is also identified as orthology-related functional category. Regarding orthology-related protein binding, many functions of this category also showed high sequence conservation among eukaryotes (e.g [37, 38]).

The next functional category which is orthology-related is translation. Many machineries involved with translation are expected to be evolutionary conserved as supported, e.g., by the evidence of finding a conserved protein family involved in translation [39], or by the presence of an evolutionary conserved mechanism for controlling the efficiency of protein translation [40].

Finally, we also observed OYC complexes containing vacuole proteins to be orthology-related. This is again supported by works which investigated yeast vacuole's proteins and function of their orthologs in other species. In particular, mammalian orthologs of yeast vacuolar protein sorting have been found to participate in early endosomal fusion and to interact with the cytoskeleton [41], and a very recent study of the same protein group revealed homologous genes and pathways that promote ageing in organisms ranging from yeast to mammals [42].

5.2 Orthology-related putative protein complexes

We consider orthology-related clusters those clusters whose proteins perform at least one function of an orthology-related functional category. In Table 3 we report the number of orthology-related clusters found by the generated predictors. We call *unique* `MCL` *or* `SiDeS` *clusters* those orthology-related clusters whose proteins have a predicted function that is not inferred for those proteins by any GYC cluster identified by `MCL` or by `SiDeS`, respectively. These are the complexes that are new and derived using (the protein complex composition present in) the orthology sub-network, that is, uniquely linked to the orthology signal.

Given a unique cluster and its protein having a novel predicted function not inferred by any GYC cluster containing the protein. Then, if the function prediction is experimentally or computationally annotated in SGD, this prediction is verified. Analogously, if we find the novel predicted function has not been experimentally or computationally annotated in SGD, then this prediction is a new one. Observe that one cluster can have verified as well as new predictions at the same time. The number of clusters that produce verified and/or new protein predictions are reported in Table 3.

Examples of these novel complexes are given in the Appendix (Table 6): they demonstrate that by examining different set of orthologs we found specific putative complexes, most of them crucial for a living cell.

For instance, proteins of Cluster 1. are predicted to be involved in mitochondrial proton-transporting ATP synthase, catalytic core. While ATP1 and ATP2 are indeed the part of the catalytic core, ATP3 is part of the central stalk of mitochondrial proton-transporting ATP synthase. Cluster 1., however, gives a proper suggestion for the mechanism of the ATP3. Moreover, as ATP3 interacts with ATP2 it may be involved also in the catalytic core.

In Cluster 2. polyadenylation-dependent r-,t- and m-RNA catabolic process is newly predicted for NRD1 protein. This complies with recent findings that NRD1 is RNA-binding protein functioning in the poly(A) independent termination, in which binding to the combined and/or repetitive termination elements elicits efficient termination [43].

Cluster 3. is a predictor for INO80 complex. Three proteins, SWR1, IES6 and VPS72, have not yet been found to be part of this complex, however all of them associate with chromatin, where IES6 directly associates with the INO80 chromatin remodelling complex. This predictor has been found by both clustering methods independently.

In Cluster 4. ERR3 is a protein of unknown function, which has similarity to enolases. The predictor was found for Yeast-Worm as well as for Yeast-Fly orthologs, and it suggests that ERR3 is part of the ubiquitin conjugating enzyme complex.

Cluster 5. predicts COPII vesicle coat proteins. This cellular component was not predicted by any GYC predictor. Newly associated proteins with COPII are HIP1 and BUG1. These predictions seem to correctly suggest their functioning in a cell, as BUG1 is cis-golgi localized protein involved

Table 2. Orthology-related functional categories. The spacing reflects the tree structure of GO slims in GO hierarchy. Functional categories in boldface are those shared by all OYC groups of eukaryotic orthologs.

Cluster Group	GO ID	Name	GO Domain
OYC-E	GO:0005739	mitochondrion	Cellular Component
	GO:0005840	ribosome	Cellular Component
OYC-W	GO:0005622	intracellular	Cellular Component
	GO:0005730	nucleolus	Cellular Component
	GO:0005773	**vacuole**	Cellular Component
	GO:0016023	cytoplasmic membrane-bounded vesicle	Cellular Component
	GO:0006412	**translation**	Biological Process
	GO:0007165	**signal transduction**	Biological Process
	GO:0009056	**catabolic process**	Biological Process
	GO:0019538	protein metabolic process	Biological Process
	GO:0005515	**protein binding**	Molecular Function
	GO:0003824	catalytic activity	Molecular Function
	GO:0004721	phosphoprotein phosphatase activity	Molecular Function
	GO:0016301	**kinase activity**	Molecular Function
	GO:0004672	**protein kinase activity**	Molecular Function
OYC-F	GO:0005730	nucleolus	Cellular Component
	GO:0005773	**vacuole**	Cellular Component
	GO:0043234	protein complex	Cellular Component
	GO:0006412	**translation**	Biological Process
	GO:0007165	**signal transduction**	Biological Process
	GO:0009056	**catabolic process**	Biological Process
	GO:0019538	protein metabolic process	Biological Process
	GO:0006139	nucleobase,-side,-tide and nucl. acid metab. proc.	Biological Process
	GO:0005515	**protein binding**	Molecular Function
	GO:0003677	DNA binding	Molecular Function
	GO:0008135	translation factor activity, nucleic acid binding	Molecular Function
	GO:0016301	**kinase activity**	Molecular Function
	GO:0004672	**protein kinase activity**	Molecular Function
OYC-H	GO:0005654	nucleoplasm	Cellular Component
	GO:0005773	**vacuole**	Cellular Component
	GO:0005829	cytosol	Cellular Component
	GO:0016023	cytoplasmic membrane-bounded vesicle	Cellular Component
	GO:0006412	**translation**	Biological Process
	GO:0007165	**signal transduction**	Biological Process
	GO:0009056	**catabolic process**	Biological Process
	GO:0005488	binding	Molecular Function
	GO:0005515	**protein binding**	Molecular Function
	GO:0016740	transferase activity	Molecular Function
	GO:0016301	**kinase activity**	Molecular Function
	GO:0004672	**protein kinase activity**	Molecular Function

Table 3. Numbers of orthology-related clusters.

Cluster Group	Method	#Predictors	#Ort-related.	#Unique	#Verified	#New
OYC-E	MCL	14	4	4	3	4
	SiDeS	3	2	1	0	1
OYC-W	MCL	80	37	32	29	31
	SiDeS	46	29	20	19	15
OYC-F	MCL	80	57	40	37	38
	SiDeS	51	44	34	32	28
OYC-H	MCL	90	33	24	20	23
	SiDeS	62	31	26	17	21

in endoplasmic reticulum to Golgi transport, and HIP1 is a high-affinity histidine permease, also involved in the transport of manganese ions.

Protein predictions for COPI vesicle coat are inferred by Cluster 6., where novel ones are for ARF1, ARF2 and ERV41 proteins. ARF1 and ARF2 are ADP-ribosylation factors involved in regulation of coated vesicle formation in intracellular trafficking within the Golgi. Because vesicles with COPI coats are found associated with Golgi membranes at steady state [44], it suggests that these predictions might be correct. ERV41 is a protein localized to COPII-coated vesicles, but again our clusters at least properly predicts a possible role of protein in a cell.

Clusters 5. and 6. were partially also discovered by Yeast-Worm and Yeast-Human orthologs. Interestingly, each of them was discovered by a different clustering technique.

Cluster 7. consists of mostly DNA-directed RNA polymerase II proteins. Although proteins DST1, TFG2 and RPA135 have not been found to be directly part of this complex, the predictor properly associates these proteins with RNA polymerase system functioning. DST1 is a general transcription elongation factor TFIIS and it enables RNA polymerase II to read through blocks to elongation. TFG2 is a Transcription Factor II middle subunit involved in both transcription initiation and elongation of RNA polymerase II. Finally, RPA135 is RNA polymerase I second largest subunit A135. Thus, the protein is correctly associated with RNA polymerases and additionally our prediction also suggests that it may play a role in formation of RNA polymerase II.

6 Conclusions

We proposed a novel methodology for quantifying the functionality of the orthology signal in a PPI network at a protein complex level. The methodology performs a differential analysis between the functions of those complexes detected by clustering a PPI network using only proteins with orthologs in another given species, and the functions of complexes detected using the entire network or a sub-network generated by random sampling of proteins.

Results of our experimental analysis indicated the usefulness of the proposed methodology to identify functional categories clearly attributed to the presence of an evolutionary (orthology) signal. The distribution of these categories was described by means of protein functions inferred from those putative complexes detected by clustering a PPI network using an explicit orthology bias incorporated in the search space.

Acknowledgements. We are grateful to Elisabeth Georgii and Koji Tsuda for sharing the protein interaction data used in [19].

References

1. Kuzniar, A., van Ham, R.C., Pongor, S., Leunissen, J.A.: The quest for orthologs: finding the corresponding gene across genomes. Trends in Genetics **24**(11) (2008) 539 – 551

2. Remm, M., Storm, C.E., Sonnhammer, E.L.: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of Molecular Biology **314**(5) (2001) 1041 – 1052
3. Chen, F., Mackey, A.J., Stoeckert, C.J., Roos, D.S.: OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Research **34**(suppl 1) D363–D368
4. Vespignani, A.: Evolution thinks modular. Nature Genetics **35**(2) (2003) 118–119
5. Wuchty, S., Oltvai, Z.N., Barabási, A.L.: Evolutionary conservation of motif constituents in the yeast protein interaction network. Nature Genetics **35**(2) (2003) 176–179
6. Wuchty, S., Barabási, A.L., Ferdig, M.: Stable evolutionary signal in a yeast protein interaction network. BMC Evolutionary Biology **6**(1) (2006) 8
7. Brown, K., Jurisica, I.: Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biology **8**(5) (2007) R95
8. Campillos, M., von Mering, C., Jensen, L.J., Bork, P.: Identification and analysis of evolutionarily cohesive functional modules in protein networks. Genome Research **16**(3) (2006) 374–382
9. Fokkens, L., Snel, B.: Cohesive versus flexible evolution of functional modules in eukaryotes. PLoS Comput Biol **5**(1) (01 2009) e1000276
10. Erten, S., Li, X., Bebek, G., Li, J., Koyuturk, M.: Phylogenetic analysis of modularity in protein interaction networks. BMC Bioinformatics **10**(1) (2009) 333
11. Yosef, N., Kupiec, M., Ruppin, E., Sharan, R.: A complex-centric view of protein network evolution. Nucleic Acids Research **37**(12) (2009) e88
12. Woźniak, M., Tiuryn, J., Dutkowski, J.: MODEVO: exploring modularity and evolution of protein interaction networks. Bioinformatics **26**(14) (2010) 1790–1791
13. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. the gene ontology consortium. Nature genetics **25**(1) (May 2000) 25–29
14. Sharan, R., Ideker, T.: Modeling cellular machinery through biological network comparison. Nature Biotechnology **24**(4) (April 2006) 427–433
15. Bauer, S., Grossmann, S., Vingron, M., Robinson, P.N.: Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics **24**(14) (2008) 1650–1651
16. Liang, Z., Xu, M., Teng, M., Niu, L.: Comparison of protein interaction networks reveals species conservation and divergence. BMC Bioinformatics **7**(1) (2006) 457
17. Jancura, P., Marchiori, E.: Dividing protein interaction networks for modular network comparative analysis. Pattern Recognition Letters **31**(14) (2010) 2083 – 2096
18. Yon Rhee, S., Wood, V., Dolinski, K., Draghici, S.: Use and misuse of the gene ontology annotations. Nat Rev Genet **9**(7) (07 2008) 509–515
19. Georgii, E., Dietmann, S., Uno, T., Pagel, P., Tsuda, K.: Enumeration of condition-dependent dense modules in protein interaction networks. Bioinformatics **25**(7) (2009) 933–940
20. Xenarios, I., Salwínski, Ł., Duan, X.J., Higney, P., Kim, S.M., Eisenberg, D.: Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Research **30**(1) (January 1 2002) 303–305
21. Guldener, U., Munsterkotter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H.W., Stumpflen, V.: MPact: the MIPS protein interaction resource on yeast. Nucl. Acids Res. **34**(suppl_1) (2006) D436–441
22. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415** (1 2002) 141–147
23. Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrín-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A., Greenblatt, J.F.: Global landscape of protein complexes in the yeast saccharomyces cerevisiae. Nature **440**(7084) (March 2006) 637–643

24. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M.: A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. Science **302**(5644) (2003) 449–453
25. Chen, F., Mackey, A.J., Vermunt, J.K., Roos, D.S.: Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE **2**(4) (2007) e383
26. Dolinski, K., Botstein, D.: Orthology and functional conservation in eukaryotes. Annual Review of Genetics **41**(1) (2007) 465–507
27. Bhardwaj, N., Lu, H.: Correlation between gene expression profiles and proteinprotein interactions within and across genomes. Bioinformatics **21**(11) 2730–2738
28. Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M., Ideker, T.: From the Cover: Conserved patterns of protein interaction in multiple species. Proceedings of the National Academy of Sciences **102**(6) (2005) 1974–1979
29. van Dongen, S.: Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht (May 2000)
30. Enright, A.J., Van Dongen, S., Ouzounis, C.A.: An efficient algorithm for large-scale detection of protein families. Nucl. Acids Res. **30**(7) (2002) 1575–1584
31. Li, L., Stoeckert, C.J., Roos, D.S.: OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research **13**(9) (2003) 2178–2189
32. Brohee, S., van Helden, J.: Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics **7**(1) (2006) 488
33. Koyuturk, M., Szpankowski, W., Grama, A.: Assessing significance of connectivity and conservation in protein interaction networks. Journal of Computational Biology **14**(6) (2007) 747–764 PMID: 17691892.
34. Hartuv, E., Shamir, R.: A clustering algorithm based on graph connectivity. Inf. Process. Lett. **76**(4-6) (2000) 175–181
35. Benne, R., Sloof, P.: Evolution of the mitochondrial protein synthetic machinery. Biosystems **21**(1) (1987) 51 – 68
36. Manning, G., Plowman, G.D., Hunter, T., Sudarsanam, S.: Evolution of protein kinase signaling from yeast to man. Trends in Biochemical Sciences **27**(10) (2002) 514 – 520
37. Sedeh, R.S., Fedorov, A.A., Fedorov, E.V., Ono, S., Matsumura, F., Almo, S.C., Bathe, M.: Structure, evolutionary conservation, and conformational dynamics of homo sapiens fascin-1, an f-actin crosslinking protein. Journal of Molecular Biology **400**(3) (2010) 589 – 604
38. Capra, J.A., Laskowski, R.A., Thornton, J.M., Singh, M., Funkhouser, T.A.: Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3d structure. PLoS Comput Biol **5**(12) (12 2009) e1000585
39. Frolova, L., Le Goff, X., Rasmussen, H.H., Cheperegin, S., Drugeon, G., Kress, M., Arman, I., Haenni, A.L., Celis, J.E., Phllippe, M., Justesen, J., Kisselev, L.: A highly conserved eukaryotic protein family possessing properties of polypeptide chain release factor. Nature **372** (12 1994) 701–103
40. Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., Pilpel, Y.: An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell **141**(2) (4 2010) 344–354
41. Richardson, S.C.W., Winistorfer, S.C., Poupon, V., Luzio, J.P., Piper, R.C.: Mammalian late vacuole protein sorting orthologues participate in early endosomal fusion and interact with the cytoskeleton. Mol. Biol. Cell **15**(3) (2004) 1197–1210
42. Fabrizio, P., Hoon, S., Shamalnasab, M., Galbani, A., Wei, M., Giaever, G., Nislow, C., Longo, V.D.: Genome-wide screen in saccharomyces cerevisiae identifies vacuolar protein sorting, autophagy, biosynthetic, and trna methylation genes involved in life span regulation. PLoS Genet **6**(7) (07 2010) e1001024
43. Hobor, F., Pergoli, R., Kubicek, K., Hrossova, D., Bacikova, V., Zimmermann, M., Pasulka, J., Hofr, C., Vanacova, S., Stefl, R.: Recognition of Transcription Termination Signal by the Nuclear Polyadenylated RNA-binding (NAB) 3 Protein. Journal of Biological Chemistry **286**(5) (2011) 3645–3657
44. Kirchhausen, T.: Three ways to make a vesicle. Nature reviews. Molecular cell biology **1**(3) (12 2000) 187–198

Appendix

Table 4. Frequencies of functional categories for Yeast-Worm `MCL` predictors. Orthology-related functional categories are in boldface.

GO ID	GYC	OYC-W	RYC-W	Name
GO:0005623	0.162	0.083	0.103 (±0.032)	cell
GO:0005737	0.120	0.037	0.086 (±0.035)	cytoplasm
GO:0016023	0.026	**0.031**	0.025 (±0.015)	**cytoplasmic membrane-bounded vesicle**
GO:0005783	0.034	0.031	0.018 (±0.012)	endoplasmic reticulum
GO:0005768	0.034	**0.035**	0.014 (±0.010)	**endosome**
GO:0005794	0.052	**0.057**	0.050 (±0.017)	**Golgi apparatus**
GO:0005739	0.108	0.044	0.114 (±0.021)	mitochondrion
GO:0005773	0.008	**0.013**	0.008 (±0.006)	**vacuole**
GO:0005829	0.015	0.009	0.013 (±0.012)	cytosol
GO:0005622	0.629	**0.657**	0.587 (±0.056)	**intracellular**
GO:0005694	0.107	0.044	0.105 (±0.027)	chromosome
GO:0000228	0.077	0.031	0.098 (±0.025)	nuclear chromosome
GO:0005856	0.034	0.026	0.033 (±0.020)	cytoskeleton
GO:0005634	0.447	**0.510**	0.461 (±0.057)	**nucleus**
GO:0005730	0.059	**0.191**	0.111 (±0.044)	**nucleolus**
GO:0005815	0.017	0.006	0.007 (±0.008)	microtubule organizing center
GO:0005635	0.013	**0.020**	0.018 (±0.012)	**nuclear envelope**
GO:0005654	0.140	0.172	0.183 (±0.033)	nucleoplasm
GO:0043226	0.605	0.470	0.412 (±0.079)	organelle
GO:0005886	0.003	**0.009**	0.002 (±0.005)	**plasma membrane**
GO:0043234	0.418	**0.539**	0.536 (±0.053)	**protein complex**

Table 5. Frequencies of functional categories for Yeast-Worm `SiDeS` predictors. Orthology-related functional categories are in boldface.

GO ID	GYC	OYC-W	RYC-W	Name
GO:0005623	0.130	0.065	0.060 (±0.028)	cell
GO:0005737	0.190	0.097	0.105 (±0.045)	cytoplasm
GO:0016023	0.020	**0.035**	0.018 (±0.015)	**cytoplasmic membrane-bounded vesicle**
GO:0005768	0.023	0.012	0.008 (±0.009)	endosome
GO:0005794	0.060	0.035	0.048 (±0.021)	Golgi apparatus
GO:0005739	0.105	0.029	0.135 (±0.023)	mitochondrion
GO:0005840	0.101	0.015	0.123 (±0.029)	ribosome
GO:0005773	0.010	**0.012**	0.007 (±0.008)	**vacuole**
GO:0005829	0.032	**0.074**	0.038 (±0.030)	**cytosol**
GO:0005622	0.670	**0.691**	0.660 (±0.065)	**intracellular**
GO:0005694	0.118	0.041	0.100 (±0.031)	chromosome
GO:0000228	0.105	0.041	0.100 (±0.030)	nuclear chromosome
GO:0005856	0.037	0.026	0.028 (±0.020)	cytoskeleton
GO:0005634	0.462	0.479	0.511 (±0.054)	nucleus
GO:0005730	0.075	**0.141**	0.130 (±0.034)	**nucleolus**
GO:0005654	0.216	**0.244**	0.234 (±0.036)	**nucleoplasm**
GO:0043226	0.463	0.297	0.393 (±0.075)	organelle
GO:0043234	0.630	0.594	0.603 (±0.045)	protein complex

Table 6. Orthology-related clusters.

Cluster ID	Proteins	Prediction	Cluster Group	Method
Cluster 1.	ATP1	mitochondrial proton-transporting ATP synthase, catalytic core	OYC-E	MCL
	ATP2		OYC-E	MCL
	ATP3		OYC-E	MCL
Cluster 2.	MTR4	nuclear polyadenylation-dependent r-,t-and m-RNA catabolic process	OYC-W	MCL
	TRF5		OYC-W	MCL
	PAP2		OYC-W	MCL
	NRD1		OYC-W	MCL
Cluster 3.	RVB1	INO80 chromatin remodelling complex	OYC-F	MCL, SiDeS
	RVB2		OYC-F	MCL, SiDeS
	ARP5		OYC-F	MCL, SiDeS
	ARP8		OYC-F	MCL, SiDeS
	INO80		OYC-F	MCL, SiDeS
	IES6		OYC-F	MCL, SiDeS
	SWR1		OYC-F	MCL, SiDeS
	VPS72		OYC-F	MCL, SiDeS
Cluster 4.	MMS2	ubiquitin conjugating enzyme complex	OYC-F,OYC-W	MCL
	UBC13		OYC-F,OYC-W	MCL
	ERR3		OYC-F,OYC-W	MCL
Cluster 5.	SEC23	COPII vesicle coat	OYC-F,OYC-W,OYC-H	MCL
	SEC24		OYC-F,OYC-W,OYC-H	MCL
	SFB2		OYC-F,OYC-W,OYC-H	MCL
	HIP1		OYC-F,OYC-W,OYC-H	MCL
	GRH1		OYC-F,OYC-W	MCL
	BUG1		OYC-F	MCL
Cluster 6.	RET2	COPI vesicle coat	OYC-F,OYC-H,OYC-W	SiDeS
	RET3		OYC-F,OYC-H,OYC-W	SiDeS
	SEC21		OYC-F,OYC-H,OYC-W	SiDeS
	SEC26		OYC-F,OYC-H,OYC-W	SiDeS
	SEC27		OYC-F,OYC-H,OYC-W	SiDeS
	ARF1		OYC-F,OYC-H,OYC-W	SiDeS
	ARF2		OYC-F,OYC-H,OYC-W	SiDeS
	COP1		OYC-F,OYC-H	SiDeS
	ERV41		OYC-F	SiDeS
Cluster 7.	SPT5	DNA-directed RNA polymerase II	OYC-H	SiDeS
	RPB2		OYC-H	SiDeS
	RPB3		OYC-H	SiDeS
	RPB4		OYC-H	SiDeS
	RPB7		OYC-H	SiDeS
	RPB8		OYC-H	SiDeS
	RPB9		OYC-H	SiDeS
	RPB11		OYC-H	SiDeS
	RPO21		OYC-H	SiDeS
	RPO26		OYC-H	SiDeS
	RPC10		OYC-H	SiDeS
	RPA135		OYC-H	SiDeS
	TFG2		OYC-H	SiDeS
	DST1		OYC-H	SiDeS