Supplementary Online Material for "**Global networks of functional coupling in eukaryotes from comprehensive data integration**"

Andrey Alexeyenko and Erik L.L. Sonnhammer

Stockholm Bioinformatics Center, Albanova, Stockholm University, 10691 Stockholm, Sweden

Methods

Statistical tests of NBN configuration

We analysed the framework configuration parameters such as "maximal number of bins in discretization", "way to use ortholog information", "choice of a co-expression metric" etc., for magnitude and significance under ANOVA general linear models (StatSoft, Inc., 2005). All accepted NBN modifications were significantly efficient (p_0<0.01). The improvements were quantified in terms of AUC. For example, introducing likelihood value check by confidence augmented AUC by 12% in the specificity region 96-100% compared to the default configuration, i.e. "using any non-zero likelihoods". The effects of single factors and their interactions are shown in **Supplementary Figure 2**. Complete balanced orthogonal ANOVA designs assured that all combinations were systematically tested. Replicates, necessary to estimate the within-combination variance, were obtained from multiple (3…10) holdout bootstraps. For each bootstrap, we randomly split the positive set and the random instance of the general population in two equal parts: one for training, and the other retained for validation.

1. Metrics of pairwise similarity

1.1. mRNA co-expression

Introduction

Although mRNA expression data from microarrays are relatively noisy, they provide very high coverage as up to 100% of all protein pairs can be analysed. mRNA expression profiles have been used extensively for predicting functional associations between genes (e.g. Bergmann et al., 2004; Fraser et al., 2004; Tirosh and Barkai, 2005).

To employ mRNA expression data in FunCoup, we tried to resolve the issues listed below. We leaned on studying the bulk FunCoup results, when non-zero likelihood values of respective bins and better overall performance of FunCoup would answer the questions:

1. *Do low (moderate) expression values contribute to the FC discovery, or only the highest (e.g. top 10%, 5%, or 1% of the range) ones shall be considered, be it in the master species or across the model organisms?*

The design "Optimal discretisation algorithm" (**Supplementary Table 2** online) proved the superiority of the multi-bin discretisation over the binary one.

2. *When some conditions of micro-array experiments are more informative on FC than others, is it feasible to select an optimal set of conditions for discriminating between FC and non-FC gene pairs?*

This problem is discussed in details in the *Data* section of this article.

3. *In some conditions, mRNA abundances of two genes can be coincidentally high, thus producing an outlier. Pearson linear correlation (PLC; Weisstein, 1999b) is notoriously sensitive to such aberrations. Does using PLC seriously compromise the discovery of FC or, on the contrary, such aberrations may deliver valuable biological information?*

Although a great deal of the micro array data analysis was dedicated to eliminating outliers, Hahn et al. (2005) found PLC better than mutual information (MI) (Weisstein, 1999d). Kemmeren et al. (2002) tested the *cosine correlation distance* in verifying yeast PPI data with co-expression and showed its minor superiority over PLC. The cosine correlation and PLC are alike, but the former does not refer to the distribution mean and is thus potentially flawed in the area of negative values. However, the mission of FunCoup was to transfer information over distant phyletic clades, with complex patterns of co-expression, which do include negative values. We thus rejected using cosine correlation. Instead, we considered MI, PLC, and Spearman rank correlation (SRC; Weisstein, 1999e). These three could potentially tackle the whole spectrum of possible problems. Indeed, while PLC is sensitive to outliers, using SRC may address the issue. If the outliers do witness FC, PLC would become advantageous. On the other hand, MI reports non-linear dependencies.

Results

A preliminary study showed that, e.g., both MI and PLC can be spuriously high when the other metric is low. FC likelihood of these cases was lower than when both MI and PLC were high.

Thus, if any of the three metrics is compromised, or not sufficient to disclose FC with co-expression, then another one can compensate for the flaw. We tested the three metrics in all the 8 possible presence/absence combinations (design ”Optimal use of co-expression metrics” of **Supplementary Table 2** online). The factorial ANOVA showed that using more than one metric is optimal. Still, no particular combination was overall superior across species, FC definitions, and specificity regions. PLC was slightly better (seldom significant) over SRC, both alone and together with MI. MI alone was usually significantly inferior to both PLC and SRC. In principle, one could use the results to choose the best option. Although, we treated the 5 tested species and the 2 FC definitions as instances of the general population of the possible data sets (*random factor effects* in ANOVA). Thus, our recommendation is to discover FC with PLC alone or, alternatively, with a combination of PLC and MI. The latter has a potential for improvement, as one can change the number of bins - which we set to 4 by default, splitting each range of expression values into quartiles.

Data preparation

All micro-array data was normalized. The absolute mRNA abundance values in Affymetrix sets were usually biased towards higher values, and applying PLC would be wrong. We rendered such data sets normal by dividing the values with the gene-specific means. Values in most of the published Affymetrix data sets are labeled as either “absent” (low confidence), “marginal”, or “present” (higher confidence). We did not exclude any of them for uniformity with dual-channel data. Higher noise content in low confidence values was thus treated uniformly with that in other data types.

The values of dual-channel arrays are ratios of “normal” vs. “treated” conditions. Such values are usually normally distributed, and we did not change it.

Co-expression metrics
PLC and SRC were computed without modifications. Mutual information (Weisstein, 1999d) addressed more complex non-linear relations between the pairs of gene expression profiles. Applying this metric needs profile values to be discretized (by classifying in a number of bins). As we knew that the distributions of particular expression values are often biased from zero, a version of adaptive partitioning (Steuer et al., 2004) was implemented: the whole range of a gene expression profile was sorted and split into four equally sized parts. Note that this binning procedure was applied to single gene expression values of rather than to pair-wise metric values such as PLC, SRC, MI (which were discretized in the BN by the procedure described in the *Discretisation* section).

Cross-hybridization
Different microarray chips can be prone to cross-hybridization: transcripts of homologous genes have an affinity to each other's probes on the chip. Hence, a higher co-expression value of two homologs might be an artifact. To avoid this factor in testing FunCoup, our tests did not include pairs of genes with a unidirectional BLAST similarity score above 100 bits. Otherwise, the results of FunCoup might be compromised by confusing inparalog pairs. How this possibility was ruled out is described in *Using orthologs* section.

1.2. Protein-Protein Interactions (PPI)

PPI-based evidence worked well across species and FC classes. In fact, the likelihood values were much stronger affected by the quality of training and test sets then by the evolutionary distance.

The protein interactions are often reported in the binary form, i.e. listing pairs of proteins. However, there have been multiple attempts to differentiate them in respect of confidence (Giot et al., 2003; von Mering et al., 2005; Suthram et al., 2006). Many approaches re-evaluated PPI confidence *post hoc*, matching them with independent evidence (equivalent to final score assignment from multiple data sources in FunCoup). Others employed "insider" information: number of tested baits / preys, their relative output, produced network topology etc. Meanwhile, times have changed and the analysis faced novel challenges:

> 1) it is feasible now to compile large multi-source datasets and evaluate them against gold standard sets in several organisms;
>
> 2) many PPIs have been reported more than once in peer-reviewed publications, and these independent confirmations became the major evidential factor.

The modern release of IntAct database keeps track of multiple replicates, experiments, methods, and publications that support a particular interaction. While even two independent experimental reports on an interaction drastically raise its confidence, many PPIs have been confirmed by 3 or more publications with in total >10 experiments.

Having compiled data from IntAct, HPRD, GRID, BIND databases plus some published datasets not yet there, we accumulated 183983 distinct protein-protein interactions in human, 38717 in *D. melanogaster*, 40732 in *M. musculus*, 183546 in *R. norvegicus*, 8887 in *C. elegans*, 173178 interactions in *S. cerevisiae*, and 3000 in *A. thaliana*. 38738 of these interactions were reported more than once. Moreover, one article might have reported validation of the same interaction in more than one assay/experiment. Normally, yeast-2-hybrid and some other techniques employ a binary approach and discover PPI as asymmetric

interactions of a "bait" protein with a "prey" protein. A bait might be then tested positive against multiple preys. Although the latter are not observed in direct interactions with each other, e.g. IntAct database presents them as members of the same interaction. We accounted for this information and included "prey-prey" interactions into our compilation. To account for such cases as well as multiplicity of reports and experiments, we introduced a novel score. It favors support from *different* articles in PubMed. The details are parsed out of the current release of the IntAct database, stored in XML format PSI-MI 2.5. This data source (October 2007) became the principal part of our compilation. Interactions from HPRD, GRID, BIND, and the large list of yeast complex members reported by Krogan et al. (2006) were added when they were traced to other PubMed publication IDs, or simply unique.

The score was similar to the main formula of Bayesian probability of functional coupling (see *Methods*), and combined the probabilistic scores S_+ (for being coupled) and S_- (for *not* being coupled):

$$S_{PPI} = \frac{S_+}{S_+ + S_-},$$

where $S_+ = P(PPI) \prod_{p=1}^{|Papers|} \prod_{a=1}^{|Assays_p|} \frac{pc_+}{\sqrt{|Assays_p(A,B)|} \cdot \log_2 |IP_a(A,B,..)|}$

and $S_- = P(PPI) \prod_{p=1}^{|Papers|} \prod_{a=1}^{|Assays_p|} pc_-$

Thus, the score grew with the number of individual experiments (|*Assays*|) reported the interaction between proteins *A* and *B* (pairwise or in a group/complex). The total number of partners $|IP_a|$ in the interaction involving simultaneously *A* and *B* penalized multi-member interactions. On the other hand, an article might present positive results on the same interaction from multiple assays. While this was a useful feature, we also wanted to distinguish such reports from ones coming from different papers, as the latter are even more reliable. Hence, the number of assays in the same article *p* was square-rooted.

The score was not intended to deliver the exact probability of interaction. Hence, the probabilities:

- $P(PPI)$, "an interaction exists between a pair of proteins", 0.001
- pc_+, "a single positive report is published given the interaction is true", 0.1
- pc_-, "a single positive report is published given the interaction is false" 0.001

were assigned roughly, equaling, and, respectively.

For example, two proteins reported to interact only in one paper, one assay, and without a third partner, received a score 0.091. If the assay had listed hundreds of interaction partners, the score fell down to 0.011. The highest score reached in the current IntAct dataset is 0.99999899. It was assigned when e.g. an interaction was tested in 10 assays reported in 2 papers or in 6 assays each published in a separate paper. After the discretisation, the range of score S_{PPI} was split into several bins with distinct LLR ranging between 1.5 and 8.5.

1.3. Phylogenetic profiling

Introduction

For phylogenetic profiles, we chose a strategy different from other datasets. Here, each gene

pair was classified into a discrete category describing its phylogenetic signature. For instance, the signature "mammals_insects_fungi" may characterise human genes that both have InParanoid orthologs in mouse and/or rat, fly, and yeast, but not in other species. Each signature is treated as a discrete evidence 'bins' during training. We benchmarked this method against a number of earlier proposed metrics, as well as against several novel potentially useful metrics, and found it superior (**Supplementary Fig. 6**). The background for this study is given below.

A number of metrics for phylogenetic profiling (PP) have been proposed and evaluated in respect of FC discovery (Pellegrini et al., 1999; Marcotte, 2000; Zheng et al., 2002; Glazko and Mushegian, 2004; Date and Marcotte, 2005). Sun et al. (2005) undertook a comprehensive study of the effect of evolutionary distance on the performance of PP in *Prokaryota*, while orthology relationships were ignored. However, FunCoup was focussed on eukaryotic species with stable evolutionary relations and only a handful completely sequenced genomes (unlike of hundreds in *Prokaryota*). Snitkin et al. (2006) reported a failure to employ PP in *Eukaryota*. We attempted to develop and evaluate our own approach that would:

1. Account for inequality of evolutionary distance between the genomes. Co-occurrence of two genes in very distant genomes (say, human and worm) has more biological meaning than in a number of closer relatives (e.g., in all mammal genomes).

2. Consider only orthologs that are most relevant in respect of conservation.

3. Focus on simultaneous presence of two genes rather than co-absence or asynchronous patterns.

Which homologs to use?

The common approaches are to build PP from all homologs found with BLAST (e.g., Marcotte et al., 2000) or to use clusters of orthologs such as COG (e.g., Wu et al., 2003). Sun et al. (2005) minutely evaluated using BLAST E-values between homologs in prokaryotic profiles and recommended optimal cutoffs. However, it is well known that orthologs are much more likely to have the same function than non-orthologous homologs. Neither a loose homology-based approach (Snitkin et al., 2006) nor using KOGs (a eukaryotic version of COG clusters by Tatusov et al., 2003; we have shown that they contain many outparalogs - Alexeyenko et al., 2006a) can reliably discriminate between outparalogs and orthologs. In contrast, a gene clustered by InParanoid (Remm et al., 2001) is most functionally conserved in the other species (Hulsen et al., 2006). We thus used the InParanoid clusters for all our PP metrics. If a gene G_X of genome X belonged to an InParanoid cluster between species X and Y, it was counted as present in Y in the PP.

Evolutionary distance

Complicated tree-based methods of weighting co-occurrence data with phylogenetic distance were recently proposed and tested on microbial (Zheng et al., 2005) and eukaryotic genomes (Barker and Pagel, 2005). We still wanted a simplified, easy-to-compute metrics for FunCoup. Marcotte et al. (1999) weighted conservation between genomes A and B with some "real number" reflecting the evolutionary distance. For our *PhyloCoup* metrics (which tested superior to other approaches except the other our method *PhyloSign*), we used the number of A genes in A-B InParanoid clusters. Thus for human gene pairs, having yeast orthologs to both proteins costs 3.35 more than having them in mouse. In fact, the ratio of the evolutionary distances $D_{\text{human-mouse}}$ / $D_{\text{human-yeast}}$ should be much larger than 3.35. But that nonlinearity was biologically justified: there is a minimal set of proteins that human and yeast

shall share to remain living eukaryotes.

Numeric metrics of pattern similarity

The more synchronous phylogenetic profiles are across the genomes, the more likely is their functional coupling. The problem was to find the optimal measure of profile similarity. Glazko and Mushegian (2004) presented a critical review of the binary data similarity metrics – from Pearson correlation to information theoretic tools – applied to the phyletic profiles. Many of them were demonstrated irrelevant with a characteristic test case. We noticed that even the best evaluated metrics, such as mutual information (MI) and the modifications of Pearson correlation coefficient (PLC) ignored the inequality of evolutionary distance and under-estimated asynchronous occurrence of two genes.

Wu et al. (2003) introduced a probabilistic formula to estimate systematic co-occurrence of genes across clades. This pure combinatorial estimate showed a good correspondence with functional coupling, but is not designed for use in a Bayesian framework and was not necessarily relevant to specific FC classes. For comparison, our *PhyloCoup* metrics were designed in a somewhat relative but simplified manner. They accounted for matches of two gene profiles and weighted each with the probability to have an ortholog. The probability of a gene of species *A* to have an ortholog in another species *G* ($P(A \rightarrow O(G))$) was estimated from the percentage of *A* genes found in InParanoid clusters between *A* and *G*.

Applying the Bayesian training in FunCoup, we re-tested a number of previously suggested metrics, often modified to a certain degree. The list included the positively defined Jaccard distance (Teknomo, 2005) which Glazko and Mushegian (2004) mentioned among the most sensitive metrics. Our original metrics were tested as well (full descriptions are given below).

One of our metrics, "*PhyloCoup, inverse weighting*", was the best out of the 13 tested on four species. The second best was the "*PhyloCoup, match only*" metric. Despite the inferiority of MI and Pearson-like coefficients in this comparison, other works showed their good performance. Probably, the disagreement is due to specificity of analysis in prokaryotes, with hundreds genomes available and the evolutionary relationships often obscured with horizontal gene transfer. Our task seemed much harder in this respect: we tested FunCoup on a very compact set of genomes with unequal evolutionary distances, and this could be the reason for the efficacy of the weighting procedures.

For each master species' genes, the phylogenetic profiles were built across 7-8 eukaryotic genomes: the series of {human, mouse, fly, worm, yeast, *A.thaliana*} plus 1-2 evolutionary close species (*Rattus norvegicus* for human, *Anopheles gambiae* for fly, *Schizosaccharomyces pombe* and *Candida albicans* for yeast, *Caenorhabditis briggsae* for worm, *Oryza sativa* for *Arabidopsis*).

Testing on KEGG-derived metabolic or signaling links would not be fair, as the KEGG pathways were enriched in genes with easily reconstructed phylogeny and thus had advantageous profiles. Therefore, we tested the metrics on PPI gold standard sets.

$a \rightarrow O(G)$ denotes an event when a particular gene *a* has an ortholog in the genome *G*;
$\neg$: NOT;
$\vee$: OR;
$\wedge$: AND.
The names used in **Supplementary Fig. 6** for each metric are given in *italics* below.

Mutual information

The common formula of *mutual information* is (Weisstein, 1999d):

$$I(a_i;a_j)=\sum_{x}\sum_{y}P(a_i=x,a_j=y)\log\frac{P(a_i=x,a_j=y)}{P(a_i=x)P(a_j=y)} \qquad \text{(MI-1)},$$

where x and y are the states of genes a_i and a_j, respectively, i.e. denote absence/presence of orthologs in the genome G:

$$x=\{\neg a_i \rightarrow O(G); a_i \rightarrow O(G)\}$$
$$y=\{\neg a_j \rightarrow O(G); a_j \rightarrow O(G)\} \qquad \text{(MI-2)},$$

Usually the numerators for the probability expressions of (MI-1):

$$P(x)=\frac{n(x)}{N_G},\ P(y)=\frac{n(y)}{N_G},\ \text{and}\ P(x,y)=\frac{n(x,y)}{N_G} \qquad \text{(MI-3)}$$

are simply counts of presence/absence of orthologs to a_i, a_j across the genomes:

$$n(x)=\mid x \mid,\ n(y)=\mid y \mid,\ \text{and}\ n(x,y)=\mid x \wedge y \mid \qquad \text{(MI-4)}.$$

Reasoning that, say, mouse and worm orthologs have different meaning for human phylogeny, we modified (MI-1) by *weighting* (Guiasu, 1977) the terms (MI-3) with fraction of species' genes that have orthologs in model organism G:

$$n(x)=\sum_{G}^{N_G}(1-P(A\rightarrow O(G)) \quad , \qquad \text{(MI-5)}$$

$$n(y)=\sum_{G}^{N_G}(1-P(A\rightarrow O(G)) \quad , \qquad \text{(MI-6)}$$

$$n(x,y)=\sum_{G}^{N_G}(1-P(A\rightarrow O(G)) \quad , \qquad \text{(MI-7)}$$

where G is a genome with the respective combination of presence/absence of orthologs to a_i and a_j.

Thus, the input from a human gene having an ortholog in mouse was 0.356, whereas a gene with an ortholog in worm gave 0.609. The difference is less than it seems from the evolutionary point of view, but it does reflect the probabilistic orthology relationships between the genomes.

Jaccard distance

Simple matching coefficient (Teknomo, 2005) is the fraction of matches (when orthologs to both genes are either present or absent) in the phyletic profile:

$$S=\frac{\mid a_i \rightarrow O(G) \wedge a_j \rightarrow O(G) \mid + \mid \neg a_i \rightarrow O(G) \wedge \neg a_j \rightarrow O(G) \mid}{N} \qquad \text{(J-1)},$$

where N is the complete number of matches plus mismatches:

$$N=$$
$$\mid (a_i \rightarrow O(G) \wedge \neg a_j \rightarrow O(G)) \vee (\neg a_i \rightarrow O(G) \wedge a_j \rightarrow O(G)) \mid +$$
$$\mid (\neg a_i \rightarrow O(G)) \wedge (\neg a_j \rightarrow O(G)) \mid +$$
$$\mid (a_i \rightarrow O(G) \wedge a_j \rightarrow O(G)) \mid \qquad \text{(J-2)}.$$

Jaccard's coefficient is the simple matching coefficient without co-absence cases (which we believe are less relevant to FC conservation):

$$JC=\frac{\mid (a_i \rightarrow O(G) \wedge a_j \rightarrow O(G)) \mid}{\mid a_i \rightarrow O(G) \wedge a_j \rightarrow O(G) \mid + \mid (a_i \rightarrow O(G) \wedge \neg a_j \rightarrow O(G)) \vee (\neg a_i \rightarrow O(G) \wedge a_j \rightarrow O(G)) \mid} \qquad \text{(J-3)}.$$

The numerator of *Jaccard's distance* counts mismatches instead of matches:

$$JD = \frac{|(a_i \to O(G) \wedge \neg a_j \to O(G)) \vee (\neg a_i \to O(G) \wedge a_j \to O(G))|}{|a_i \to O(G) \wedge a_j \to O(G)| + |(a_i \to O(G) \wedge \neg a_j \to O(G)) \vee (\neg a_i \to O(G) \wedge a_j \to O(G))|} \quad \textbf{(J-4).}$$

A *modified* version of Jaccard's distance had the mismatch count *only* in the numerator, and we took a log of it for convenience:

$$JM = -\log \frac{|(a_i \to O(G) \wedge \neg a_j \to O(G)) \vee (\neg a_i \to O(G) \wedge a_j \to O(G))|}{|a_i \to O(G) \wedge a_j \to O(G)|} \quad \textbf{(J-5);}$$

When either the numerator or the denominator of the log ratio equalled zero, it was replaced with a small constant.

Hamming distance
counts mismatches (when *G* has an ortholog to one of the genes, and not to the other):
$H = |(a_i \to O(G) \wedge \neg a_j \to O(G)) \vee (\neg a_i \to O(G) \wedge a_j \to O(G))|$.

PhyloCoup metrics
A weighted version of the Hamming distance (*PhyloCoup, mismatch only*) penalized mismatches:

$$PhC_{Mismatch} = \sum_{G}^{N_G} -1 + P(A \to O(G))(1 - P(A \to O(G)))$$

Another version assigned, on the contrary, a bonus for each match (*PhyloCoup, match only*):

$$PhC_{Match} = \sum_{G}^{N_G} \begin{cases} 1 - P(A \to O(G))^2, if\,(\neg a_i \to O(G) \wedge \neg a_j \to O(G)) \\ 1 - (1 - P(A \to O(G)))^2, if\,(a_i \to O(G) \wedge a_j \to O(G)) \end{cases}$$

And a union of the two (*PhyloCoup*) was:

$$PhC_{Complete} = \sum_{G}^{N_G} \begin{cases} 1 - P(A \to O(G))^2, if\,(a_i \to O(G) \wedge a_j \to O(G)) \\ 1 - (1 - P(A \to O(G)))^2, if\,(\neg a_i \to O(G) \wedge \neg a_j \to O(G)) \\ -1 + P(A \to O(G))(1 - P(A \to O(G))), otherwise \end{cases}$$

For a comparison, an inverted version of the latter (with reversed treating of similar vs. distant genomes) was tested (*PhyloCoup, inverse weighting*):

$$PhC_{Complete,inverted} = \sum_{G}^{N_G} \begin{cases} 1 - (1 - P(A \to O(G)))^2, if\,(a_i \to O(G) \wedge a_j \to O(G)) \\ 1 - P(A \to O(G))^2, if\,(\neg a_i \to O(G) \wedge \neg a_j \to O(G)) \\ -1 + P(A \to O(G))(1 - P(A \to O(G))), otherwise \end{cases}$$

Discrete metric
Eventually, we reasoned that because the number of different co-presence profiles is small, they can be expressed as distinct *phylogenetic signatures.* For example, {`mammal, fly, worm`} denotes presence of both genes in mammals, *D. melanogaster*, and *C. elegans*. Then likelihoods could be assigned to such individual strings. This approach removed any chance of a score peculiarity – which is still possible even after a careful design. However, the total number of distinct signatures should not be too large – this would lead to insufficient number of observations of each category in the training procedure. For this reason, the situations "both absent" and "one present, one absent" were collapsed together, and the number of species was reduced to that of major clades (*Fungi, Plantae, Animalia*) plus closer relatives.

The discrete *PhyloSign* values proved to be more discriminative (**Supplementary Fig. 6** online). The LLR ranged from –2 ({`human, worm`} in human FC-PI) to +3.5 ({`human, mouse, rat, fly, worm, yeast, plant`} in human FC-CM).

1.4. Sub-cellular co-localization

Co-localization is a binary (two proteins are either present *together* or not) feature. However, we wanted to account for:

1) occurrence of a protein in more than one compartment and

2) sizes of the compartments' sub-proteomes.

Indeed, co-localization in a big compartment which thousands of proteins (cytoplasm, nucleus) and in a small organelle (polysome) should differ in importance. Using a mutual information (MI) score addressed the first issue. The second was tackled by applying weights (Guiasu, 1977) to the MI terms: the more proteins in the compartment, the less meaning have co-occurrence of two proteins in it.

The common formula for MI was modified in a way similar to the MI formula adapted for measuring phylogenetic profiles' similarity:

$$WMI_{SLC} = \sum_{a=\{0,1\}} \sum_{b=\{0,1\}} P(i=a, j=b) \log \frac{P(i=a, j=b)}{P(i=a)P(j=b)}$$

where *a, b* are the "presence/absence" indicators of proteins *i* and *j* in locations *L*. The absence/presence observations in each sub-cellular location *l* were counted as complement to the relative size of *l*:

$c(l) = 1 - n_l / N_L$.

Therefore

$$P(i=a) = \frac{\sum_{l=1}^{N_L} c(l)}{N_L},$$

where N_L is the total number of protein localizations mentioned for the organism. Identically to the phylogenetic mutual information score, the fewer proteins assigned to *l*, the more informative the co-localization.

1.5. Protein-DNA binding

As a source of protein-DNA interactions, we collected data from 3 major datasets

- Genome-wide predicted binding sites of transcription factors (TF) GLI1, GLI2, GLI3, TCF-4 conserved between human and one or more vertebrate (Hallikas et al., 2006);
- MPromDB database of mammalian TF binding sites (Sun et al., 2006);
- RegulogDB database of yeast TF binding sites conserved in *C. albicans*, worm, fly, *A. thaliana* – plus the respective pairs of orthologs (*regulogs*) in the 3 latter species (Yu et al., 2004).

A "TF -> target" pair is a link between two genes and can be considered as input for FunCoup *per se*. However, such links would not overlap with the training sets used in FunCoup (FC-ML, FC-PI, FC-CM), and only rarely would with FC-SL. Hence, a confident LLR could not be derived in the training procedure. We thus provide such couplings as additional information. On the other hand, the FC between genes – targets of the same (sets of) TFs can be tested with any of our training sets. The score for overlap was calculated as:

$$S_{TF-ctgt} = \frac{|\{BS_i\} \cap \{BS_j\}|^2}{|\{BS_i\} + \{BS_j\}|},$$

i.e. the shared fraction of the binding site sets {*BS*} between genes *i* and *j* was multiplied by the cardinality of the shared subset. (identically to the S_{miRNA_ctgt} score).

1.6. miRNA-gene targeting

miRBase targets database version 5 (November 2007) (Griffiths-Jones et al, 2007) served input for this data type in FunCoup. The database contained predicted binding sites of known miRNA positionally conserved in 2 or more animals. The majority of genes in each species (human, mouse, rat, fly, worm) was predicted as a target to between 90 (fly) to 624 (human) distinct miRNAs, with species-average number of binding sites per gene from 3.5 (worm) to 40 (human). Most of these predictions were obviously either computationally false or biologically irrelevant. We thus tried to retrieve meaningful evidences from the sets.

For each pair of genes *i* and *j* with sets of predicted miRNA target sites having the complementarity score > 15 (about 80% of the database content), a functional coupling score for overlap was calculated as:

$$S_{miRNA-ctgt} = \frac{|\{BS_i\} \cap \{BS_j\}|^2}{|\{BS_i\} + \{BS_j\}|},$$

i.e. the shared fraction of the binding site sets {*BS*} between genes *i* and *j* times the cardinality of the shared subset (identically to the S_{TF_ctgt} score).

1.7. Protein co-expression

The Human Protein Atlas (Hober and Uhlen, 2008) provided data on staining 1400 cell line and tissue samples with antibodies to about 3000 human proteins. Each sample had been analyzed for dye intensity of the staining antibody and received a grade ("white": negative, "yellow": weak, "orange": moderate, and "red": strong). The protein co-expression score in FunCoup had thus to deal with quantitatively ordered coarse-resolution data. We tested a number of opportunities and found an optimal score.

The common formula for mutual information (Weisstein, 1999e) is:

$$WMI_{PEX} = \sum_{a=\{w,y,o,r\}} \sum_{b=\{w,y,o,r\}} P(i=a, j=b) \log \frac{P(i=a, j=b)}{P(i=a)P(j=b)}$$

where *a*, *b* are the color indicators of the staining of genes *i* and *j*. To account for the respective color abundance in each sample, the sample-specific weighting coefficients rather then unities were summated along the staining profiles. A weight coefficient was the fraction of color grade *a* (one of {"w", "y", "o", "r"}) in sample *s*:

$w(s,a) = n_{s,a} / N_s$,

where N_s is the total number of successfully stained genes in sample *s*, i.e. the sum of all "w", "y", "o", and "r"′s. Hence,

$$P(i=a) = \frac{\sum_{s=1}^{N_{i\cap j}} w(s,a)}{N_{i\cap j}}, P(j=b) = \frac{\sum_{s=1}^{N_{i\cap j}} w(s,b)}{N_{i\cap j}}, \text{ and } P(i=a, j=b) = \frac{\sum_{s=1}^{N_{i\cap j}} w(s,a) \cdot w(s,b)}{N_{i\cap j}}$$

rather than actual probabilities of *i*=*a* etc. in the common formula. $N_{i\cap j}$ is the number of cell/tissue samples which both *i* and *j* stained successfully.

The "*cell line*" and "*tissue*" subsets of HPA were processed as separate datasets and yielded

distinct likelihood values.

2. Using orthologs for inferring functional coupling

Information from model organisms is very useful in functional coupling (FC) discovery. A number of approaches have been tried to transfer pairwise functional links between genomes: best reciprocal hits (Brown and Jurisica, 2005), all homologs above a threshold (Wojcik and Schächter, 2001), InParanoid clusters of orthologs (Lehner and Fraser, 2004). von Mering et al. (2005) tested two methods:

1. *Pooling evidences across whole COG clusters of orthologs (Tatusov et al., 1997): the members of an orthologous group equally contribute to the prediction.*

2. *Weighting high scoring homologs by the sequence similarity: more distant homologs less affect the prediction.*

While the latter approach agrees with the theory of functional divergence in paralogs (and showed the best results in the test by von Mering et al., 2005), the proposed weighting system is an *ad hoc* one and uses arbitrary ("expert") estimates of sequence-function divergence ratios. Moreover, the algorithm of finding orthologs frequently includes out-paralogs, which are not orthologs (Sonnhammer and Koonin, 2002), in the clusters..

To implement using orthologs in FunCoup, we addressed following questions:

A. *In case of alternative ortholog pairs, how is the useful information distributed among them?*

B. *What kinds of genomic information (if not all) one can transfer between the orthologs to predict functional coupling?*

C. *What grouping of orthologs is optimal for transferring information on functional coupling?*

We used InParanoid as the source of orthologs (Remm et al., 2001) seemed more relevant. Homologs originating before the speciation are automatically ignored, which sets a natural, evolutionary and thus functionally relevant cutoff.

For technical reasons, InParanoid classifies orthologs into two groups: seed orthologs (reciprocally best hits) and additional inparalogs (that are closer to the seed orthologs than to any gene from the other genome). Thus, it was possible to test ortholog links classified by four types (looking from the master species of interest): "Seed-to-seed", "Seed-to-additional", "Additional-to-seed", "Additional-to-additional" (see the Table below). Option 1 corresponds to the reciprocally best hit approach.

	Type of ortholog links with examples from Fig. IllAlternativesInLinkTransfer	
	target species	**source species**
1.	Seed ortholog(s)	Seed ortholog(s)
2.	Seed ortholog(s)	Additional inparalogs

3.	Additional inparalogs	Seed ortholog(s)
4.	Additional inparalogs	Additional inparalogs

FunCoup can assign specific likelihood values to the data sources and functional classes based on their relevance to FC. It is possible to calculate likelihoods in respect of these four link types, too. The comparison clearly showed that the likelihood values of FC differed between the four types. Hence, it seemed prospective to employ them separately.

Thus, we designed an experiment to compare three ways of using ortholog data. For each evidence feature, the FunCoup framework produced one of:

1. A column with data available only via "seed-to-seed" links.

2. A column with data pooled for all possible pairs between the respective InParanoid cluster members (sum of lines 1-4 in the above Table).

3. Four separate data columns, each estimated for one of the four link types separately.

The respective levels of factor "Mode of using orthologs" were included in the ANOVA design "Optimal FunCoup configuration" (**Supplementary Table 2** online).

Using clusters of inparalogs poses another question: how to combine the information from multiple alternative ortholog pairs. Two ways seemed reasonable: choosing either the best or the mean value of the available values (with metrics-specific definitions of the "best": e.g. the maximum absolute value of PLC, the maximum for MI, or minimum values for a metric that counts phylogenetic mismatches). We tested the two ways as levels of the factor "Estimating alternative inparalog pairs" in the design "Optimal FunCoup configuration" (**Supplementary Table 2** online).

Analysis of the factorial ANOVA design "Optimal FunCoup configuration" (**Supplementary Table 2** online) clearly showed that using the option "All pooled" was either better or equal to both "Seed-to-seed" and "All separated" options in every combination of "Estimating alternative inparalog pairs" and "Likelihood confidence check" over the tested species and the classes of FC. This held at all the ranges of specificity.

The same was true about the option "Best" of "Estimating alternative inparalog pairs" in all the respective combinations (**Supplementary Table 2** online shows only main factors rather than combinations). We concluded that picking up the best values from all inparalog pairs is optimal.

There was still one more problem about using multiple ortholog data as evidence. Micro-array probes can bind transcripts similar but not identical to those they were designed for. This problem (*cross-hybridization*) has been addressed by the chip designers and users (e.g., Sartor et al., 2006). Nowadays, different technologies, chips, and experiments are differently vulnerable to this potential drawback. Sometimes there is no correlation between expression patterns of two genes with very high BLAST score and percent identity. In other cases, they do correlate. In the latter case, it can be not an artifact but rather an effect of a promoter mechanism shared by the recent duplicates, or another kind of a biologically meaningful co-regulation. However, it seemed impossible to distinguish between such cases and unspecific hybridization. It was easy to filter out such cross-hybridization in the protein pairs used for training and testing: homologous pairs were completely excluded (see *Methods*). But the expression profiles may well be confused between inparalogs of a model organism. Say, gene A can have an expression profile similar to that of B. But the differential expression of B in

some or all conditions was in fact induced by homologous transcripts cross-hybridized to the B probes. Then all the six gene pairs between the two clusters of orthologs in species Y are compromised, and we observe a value averaged across the inparalog pairs. As a consequence, the "Best" and "Mean" approaches should be equally efficient on expression data.

We did show superiority of the "Best" option in the design "Optimal FunCoup configuration", but only on a mix of evidences. Hence, a contrast experiment might help in the investigation. If there were a significant cross-hybridization in pure expression data sets, then using only mRNA expression data should diminish the effect of the "Best". In contrast, when all but expression evidence is used (PPI, sub-cellular co-localization, and phylogenetic profiling), the "Best" option should be more advantageous than on the mix of data. The tested null-hypothesis was that both sets respond to the switch from "Best" to "Mean" option identically. If one of the two sets is more prone to confusion of inparalogs, the hypothesis would be rejected.

The experiment did not expose any significant difference between the "only expression" and "non-expression" variants (not shown): the null-hypothesis was not rejected. Thus, both the expression and the non-expression evidences are either equally affected by the cross-hybridization or not affected significantly. The former possibility seems unlikely, as it would mean systematic confusion of paralogous genes in the data sources obtained with Y2H, co-precipitiation, co-citation, and sub-cellular localization techniques. In principle, PPI is shown to be more likely when interacting paralogs are known to both genes (Deane et al., 2002). But it was also demonstrated that interacting pairs can be discerned among the pairs of inparalogous proteins (Baudot et al., 2004), and thus the modern proteomics is sufficiently homolog-specific. We thus could rule out the possibility that the information flow via InParanoid clusters was significantly corrupted by micro-array cross-hybridization, and considered the results inparalog-specific. However, we did not assess here any particular data sources and genes, and the cross-hybridization still might occur in some of them.

It would have been also of interest to compare the above mentioned method of weighting homologs implemented in the STRING database (von Mering et al., 2005). A major difference between our method and STRING – which evaluates evidence datasets against the gold standard in the same species – is that we employ Bayesian estimation against a gold standard in the species for which the predictions are made. Moreover, the exact reproduction of the STRING benchmark scores for each our dataset was impossible: the FunCoup scores are of a different nature. Indeed, they are produced by comparing positive training sets vs. random protein pairs. As the former have very different ratios of seed orthologs / additional inparalogs (up to 6 times, when comparing KEGG links to random pairs), further weighting would strongly bias the result.

3. Discretisation

3.1. Summary

The discretization algorithm that we developed for FunCoup is similar to the one by Butterworth et al. (2004), but because it is based on the Pearson χ^2-statistic rather than the conditional entropy it does not require setting a parameter (power index = 1.8…2.2) as an additional step. With a χ^2-score it tests all prospective cutpoints, i.e. ones where
1) sample counts are sufficient,
2) χ^2 values are significant (p_0<0.001), and
3) the class label swaps between the positive and background FC.

The maximally scored point splits the metric range in two initial bins. Further partitions are iteratively sought while any prospective points remain. We tested the method against the default quantile-based partitioning and found the novel method significantly superior (**Supplementary Fig. 2**). The algorithm usually stops at 5-10 bins, and we introduced a practically justified limit of 10 bins. When data deliver little information on FC, fewer bins are created. No splits means that positive and background labels cannot be separated signficantly, and that the dataset is not useful. The advantage of this procedure is that it is insensitive to a metric's distribution shape and the position of local optima.

3.2. Background and reasoning

Bayesian networks were traditionally built from discrete events (Heckerman,1995; Friedman et al., 1997). Generally, they might accept continuous variables (while these are highly desirable in the discriminant analysis): each parent node of the BN assumes a function to transform input events into the probabilities of the child nodes' states. But the joint probability density over the nodes in the whole network is hard to estimate in a continuous manner. Semi-parametric density models were proposed (e.g., by Heckerman and Geiger, 1995) but proved to be very sensitive to chosen parameters. Instead, Friedman et al. (2000) introduced a discretisation of the gene co-expression values. The probabilities were then estimated for a number of intervals over the range of the continuous value. Following this approach, many genomics applications presented data in a multinomial (often binary) form. The binarisation of continuous data (e.g. Xia et al., 2004) had drawbacks. In such an approach, the continuous value range (e.g., PLC of gene expression profiles) is split in two parts. Those above a chosen cut-off usually served a positive evidence of FC. The ones below were either ignored or treated as negative evidences. The task is to find the optimal border between the two regions. Some authors accepted a once established cut-off value to be used throughout the framework (Brown and Jurisica., 2005, Gunsalus et al., 2005), others customized them in respect of dataset and species. In any case, the binarization run counter to the progress of the micro-array technology: while expression measurement increases in quality and precision, the data integration renders it binary. The waste of information was obvious. It was intuitively clear that, say, a PLC = 0.8 may be a stronger evidence than *PLC* = 0.4. Even lower values should not be discarded. von Mering et al. (2003) had shown that the dependency between raw value of an evidence feature and the probability of FC is not linear. Our experience showed that sometimes it is not monotonic either. A negative correlation of expression (at least, in some remote orthologs) can well be an evidence of FC in two proteins.

Naïve BNs do not need joint probability over nodes, hence it is possible to fit dependencies between continuous variables with regression curves. Imoto et al. (2002) developed a detailed method of using non-linear regression in building BNs of genes as nodes (to fit the joint expression distributions of gene pairs). Lee et al. (2004) fitted the dependences of FC likelihoods on co-expression (PLC) values. Very detailed series of points were produced in "co-expression vs. agreement with known FC pairs" space, and the resulting equations of non-linear regression served FC likelihood values to be summed up into final Bayesian scores.

However, we could not adopt a regression approach in FunCoup for a number of reasons. Firstly, such fine-grained structure of empirical distributions is computationally hard when processing several eukaryotic genomes. Secondly, a "co-expression-to-likelihood" application still implies a kind of discretisation: single estimates of the likelihood are made on small co-expression intervals (Lee et al. have used intervals 0.01 long; thus range of PLC [-1...1] splits into 200 bins). With data points not equally spread, or missing values prevailing, or a training

set small, the regression line might get unreliable. Lee et al. (2004) considered only the monotonic S-shaped part of the PLC-FC distribution (at best [0.3…1.0], often [0.6...1.0]), having ignored the negative values. However, the negative co-expression values acquire positive likelihoods (chiefly as evidences from model organisms). Hence, our problem should be considered as specific to multiple species applications. And finally, any fitting algorithm did not seem universal and trustworthy to let it work without human supervision with any kind of data, size of dataset, and curve shape – thus not applicable in an automated solution such as FunCoup.

We thus aimed for a compact and universal solution at the cost of some loss of continuity, yet keeping compatible with the Bayesian technique. In general, learning Bayesian likelihoods from data using multinomial features has been well studied (Heckerman, 1995). Drawid and Gerstein (2000) divided the single mRNA's abundance profiles into multiple bins (which is synonymous to discretisation), but did not mention a special procedure for finding bin borders. Splitting the co-expression range into multiple bins to discover FC was mentioned by Jiang and Keating (2005) and Myers et al. (2005) but no technical details were provided. Myers and Troyanskaya (2007) employed integer Z-scores to define bin borders. However with the expansion of the framework, an arbitrary setting of the bin borders can loose the relevance and escape the curator's notice.

Note that declaring a variable *multinomial* assumes no preliminary ordering of the particular values (*bins*). Thus, both leftmost and rightmost (and even intermediate) bins can receive a higher likelihood value. In our experience, this was sometimes the case, e.g. for PLC and one of the tested phylogenetic profiling metrics, which had parabola-shaped likelihood distributions with lowest or highest values in the middle.

Thus, no prior assumption about positive and negative regions should be made; each part should be estimated objectively. The framework should keep flexibility and be capable of accepting novel features/metrics without manual fine-tuning.

In principle, maximal precision should be achieved with as many partitions as possible. But the number of observations per bin would then drop below a critical level, and make most of the likelihoods unreliable. Also, many neighbor bins would have almost the same likelihood, and keeping them separated means lost of robustness, and no gain in the prediction ability. A flexible (*adaptive*) partitioning would be highly desirable to place bin borders only at some "optimal" points. Although it is intuitively clear that the binning should be aware of FC, Fayyad (1991) showed that the discretisation is most efficient when the bin borders coincide with the functional differences – at the contrast points between FC-rich and FC-poor regions. A simple example in the Table below shows how the class labels and the feature values may be co-distributed. Probably, the optimal cuts should be made between points 3-4 and 8-9.

Case no.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
FC	+	–	+	–	–	–	–	–	+	+	–	+

PLC	-0.7	-0.6	-0.5	-0.1	0.0	0.0	0.1	0.2	0.3	0.5	0.6	0.8

A brief illustration to the problem of discretising continuous features with regard to FC. The second row indicates if the protein pair is functionally coupled. The third row: Pearson linear correlation coefficient of the protein pair.

A review of adaptive partitioning applied to MI of gene pairs expression is given in Steuer et al. (2004). Butterworth et al. (2004) considered the problem of partitioning that is aware of class labels. They suggested a new method of adaptive discretisation based on conditional entropy (the entropy contained in the feature bins on condition of the class). Unfortunately, a parameter β needs to be adjusted empirically (Butterworth et al. tested $\beta = \{1.5, 1.8, 2.0\}$; we often found the optimum for our data at $\beta = 2.2$). Thus, the quality of a solution depends on the value of β which is learned only after the whole procedure. This makes the once optimized framework sensitive to future changes. We did not include Butterworth's method in the regular study described below: optimizing β was beyond our computational capacity, and β left under-optimized would make the comparison unfair. Other drawbacks of this method were that the cutpoints were sometimes too close to each other and the new cutpoints are accepted only when an entropy gain is achieved (which is not identical to a significantly non-zero likelihood).

3.3. Implementation

We developed a discretisation algorithm based on the well-known Pearson χ^2-test. It is not aware of entropy before and after the split, and uses very simple limitations. We stabilized the solution by fixing the maximum number of bins, and accepted the cutpoints if the difference between the partitions had $p_0 < 0.001$ by the χ2-test. Fayyad (1991) and Butterworth et al. (2004) have shown that the optimal split is always found at a point where the series of identical class labels interrupts (points 1-2, 2-3, 3-4, 5-6, 6-7, 8-9, 10-11, 11-12 in the Table). Narrowing down the scope of the search to only such cases (here 8 points instead of 11) reduced the amount of computations (practically, up to 10-fold). Still, it remained quite high in big datasets, as in our case (~2000 potential cutpoints; this part of the procedure takes several times longer than all the others together). Thus, another modification was in testing initially every n-th potential ($n = 50$) cutpoint. Falling into local minima is not particularly dangerous here, as the feature surface is gently sloping and has few inflexion points.

The χ^2-based algorithm was implemented as follows for each "feature - FC class" pair:

1) The whole series of the feature values is sorted (the positive and the random training sets merged). Functional class labels ("+": positive; "–": random, or "not known") are retained at each value.
2) All potential cutpoints are marked up. These are defined as points were the class label series interrupts (series of one or more "–" switches to "+", or vice versa).
3) Each n-th potential cutpoint is tested. If condition A (below) holds, the ratios of class labels to the left and to the right from the cutpoint (n_{+Left}/n_{-Left} vs. $n_{+Rightt}/n_{-Right}$) are compared with χ^2 metric. For the cutpoint with a maximal χ^2 value, a better position is sought in its neighborhood ($\pm n/2$ potential cutpoints). Thus, the first two partitions are split at the point of the maximal difference between the ratios n_+/n_-. A cutpoint is accepted only if conditions B and C hold.
4) The procedure is repeated for each new partition iteratively, until the limit of condition C is achieved.

Conditions of the algorithm:

A. Number of points of each class in each potential partition >9.
B. χ^2 score > 10.83 (p_0 = 0.001; 1 degree of freedom)
C. No. of already accepted partitions does not exceed the limit (10).

Formula of χ^2 –test (modified version of Weisstein, 1999a):

$$\chi^2 = \sum_{i=\{LEFT,RIGHT\}}^{P} \sum_{j=\{+,-\}}^{C} \frac{(n_{ij} - m_{ij})^2}{m_{ij}};$$

where:
i, (the tested) partition,
j, functional class label (FC pairs vs. random pairs),
n_{ij}, observed no. of labels *j* in partition *i*,
m_{ij}, expected no. of labels *j* in partition *i*, calculated as:

$m_{ij} = \frac{n_{i.} \cdot n_{.j}}{N}$, where $n_{i.}$ and $n_{.j}$ are the marginal sums over all *j* and *i*, respectively.

For example, 25% of the positive set pairs and 3% of the random set ones have Pearson correlation coefficient r_c > 0.7. This is a point of the strongest contrast, because the neighbor values r_c = 0.68, r_c = 0.71 produce lower ratios: 26% vs. 3.5%, 23% vs. 2.9% etc. Hence, r_c = 0.7 is chosen as a cutpoint. The regions r_c < 0.7 and r_c > 0.7 are attempted to be split into more bins.
The procedure stops when the maximal number of bins is found.

3.4. Results and conclusion

We tested the χ^2-based algorithm versus the simple equal bin partitioning and found it overall superior (**Supplementary Fig. 2** online). The most important for FunCoup region of [99…100]% specificity was studied, with the region [96…100]% taken for comparison.

Based on this assessment, and having considered also specific plots for each of the species (not shown), we set the maximal number of bins to 10. The χ^2-statistic should be above 10.83 (p_0<0.001; 1 *d.f.*). The number of bins may be less than the limit when the conditions A or B of the algorithm (**Methods**) hold. It usually happens to flatter distributions.

3.5. Likelihood confidence check

Some likelihoods might receive very low likelihood values but still be useful. We compared the FunCoup performance with two alternatives:

1) when all learned likelihood values, irrespective of significance, are used for prediction

2) insignificant ones are coerced to zero.

To test for significance, we counted occurrence of each bin in the positive training set vs. that in a random set of sufficient size (200.000 protein pairs for abundant data, such as mRNA expression, and all possible proteome pairs for a PPI set). These two counts, matched to the sizes of the respective sets, produced a χ^2–score. We only accepted bins that had >9 non-empty protein pairs in each set, and yielded χ^2–score > 10.83 (p_0<0.001). Otherwise, the likelihood for the bin was set to zero. In the worst case (all bins zeroed), the whole dataset is excluded.
Insignificant bins did deteriorate the FunCoup performance (effect of "Likelihood confidence check" in the ANOVA design "Optimal FunCoup configuration", **Supplementary Fig. 3** online), and we set the significant test the default.

4. Prediction of multiple classes of functional coupling

We suggested using specialized, separately trained predictors for different FC classes (*multinets* – Friedman et al., 1997). Evidence nodes of such NBN have separate outputs to class-specific predictor nodes.

Tests had shown that FunCoup does assign different likelihoods of FC when trained in respect of different FC definitions (metabolic links vs. PPI). Moreover, using specific training was more accurate.

We tested the relevance of the multinet concept to FC problem in ANOVA design "Differential FC type prediction" (**Supplementary Table 2** online). In each test, FunCoup was trained on two functional classes (PPI and metabolic) simultaneously. A half of the subset was used for the testing by cross-validation. Each test was performed twice: first, FunCoup tried to find links of the relevant class, and second, of the other class. In the latter case, the AUC was 15…25% lower (which meant 15…30% lower specificity at the same level of sensitivity; i.e., FunCoup after specific training found more relevant links).

This feature became default, and FunCoup is capable of simultaneous training in respect of several functional classes.

5. Reconstructing the *C. intestinalis* interactome

Because of FunCoup's strong reliance on orthologs for inferring functional coupling, it can be used to reconstruct protein FC networks in one species using only information transferred from others. To demonstrate this ability, we generated a network in *Ciona intestinalis*, for which no large proteomics or genomics dataset was available. As a positive training set, we used pathway members inferred via orthology. The input data came from the seven eukaryotes listed above. To validate the predicted *Ciona* network, we compared it to the "regulatory blueprint for a chordate embryo" (Imai et al., 2006). This is a set of 226 experimentally established functional links (mostly regulatory) between 80 genes in the ascidian embryo ("RBP" network).

At a cutoff *FBS*=4 (in total, it yielded 306650 *Ciona* links), FunCoup recovered 180 links in the cluster among the genes of RBP, and 22 of them (13%; *pfc*>0.05) matched the links of Imai et al. We modelled a random sampling of such a gene set (54 genes – as many as had any links with *FBS*>4). The FunCoup network in general (Supplemental Figure 11) and the Ciona network in particular (not shown) are scale-free, i.e. there is no "typical" node connectivity. Each network comprises a number of hubs, i.e. genes connected to extremely many other genes. If the RBP set were enriched in hubs, it could have biased the validation. However, the non-parametric Mann-Whitney test did not discover any difference in the connectivity distribution between RBP and the whole network (p_0=0.606). We then modelled the set of 54 RBP genes with randomly generated sets of the network genes taken in a sufficient number of samples to determine the mean and standard deviation. The expected mean number of links between 54 random genes was 39.5, thus the actually observed count 180 ($p_0<10^{-8}$) was a good result on this difficult test set. To discover by chance the 22 known links was also unlikely (modelled in a similar fashion: mean=14; p_0=0.015).

Furthermore, we assumed that, among the 158 other links, there were novel true ones – which remain to be validated in experimental research. Hence, the real p-value should be lower (e.g. it would be $p_0<10^{-6}$ already at 15 novel plus 22 known links) and the true discovery rate

$TDR = \frac{22}{180} > 0.122$ is anyway higher then the formally declared *pfc*>0.05 (*pfc* served a substitute for TDR when the needed parameters were generally unknown).

5.1.1. Deriving pathway members in a uncharacterized organism

At the time we generated the *C. intestinalis* network (December 2007), this organism was not yet present in the KEGG ortholog table. Hence, unlike the other organisms, we did not have a set of organism-specific pathway members to create a training set. We found putative *Ciona* pathway members in a way similar to the KEGG inference by orthology (Bono et al., 1998). Our method employs multi-species clusters of orthologs available from the MultiParanoid database (Alexeyenko et al., 2006b). In each ortholog cluster, we assigned EC numbers to *Ciona* proteins considering the KEGG assignments to human, fly, and worm cluster members.

These species are well studied and had 3828, 1526, and 1062 ENSEMBL genes annotated in KEGG, respectively (as of June, 2007). Inference by sequence similarity is one of the main approaches of KEGG database to map known pathways onto as many novel organisms as possible (Bono et al., 1998). The MultiParanoid clusters seemed convenient and relevant, as each contained only most likely orthologs, (delineated by InParanoid – Remm et al., 2001) between pairs of species. Potentially, members of one cluster might receive multiple EC/pathway assignments, either between genes of different species, or different genes of the same species, or even for the same gene. Moreover, genes with identical assignment might belong to different MultiParanoid clusters. As the sequence similarity was the only metric we could use (while the KEGG annotators might consider other information sources such as publications on expression patterns, interactions, etc.), we base the assignments on two scores:

1) The fraction of annotated organisms in the cluster confirming the annotation. For example, if an annotation was given for human genes but not to both *D. melanogaster* and *C.elegans* genes of the cluster, the score would equal 1/3 (one out of three possible).
2) The fraction of genes with a given annotation relative to the total number of genes with any annotation in the cluster. Each annotation was normalized by the relative evolutionary distance of the species from *Ascidia* (approximated as 1 / sqrt(number of clusters in the MultiParanoid cluster set where the species was present and Ciona was not).

To assign putative pathway roles, we empirically set the cutoff of the first score to 1/3, and the second score to 0.4. In combination with the strict clusters by MultiParanoid, this allowed reasonably reliable inference of metabolic gene roles.

The procedure for extracting FC was identical to the other species – any two genes with EC numbers from the same pathway were assumed to be coupled.

Supplementary references

1. Alexeyenko A, Lindberg J, Perez-Bercoff A, Sonnhammer ELL. Overview and comparison of ortholog databases. Drug Discovery Today: Technologies 2006a; 3:137-143
2. Alexeyenko A, Tamas I, Liu G, Sonnhammer ELL. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics, 2006b, 22: e9-e15.
3. Barker D, Pagel M. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput Biol. 2005 Jun;1(1):e3. Epub 2005 Jun 24.
4. Baudot A, Jacq B, Brun C. A scale of functional divergence for yeast duplicated genes revealed from analysis of the protein-protein interaction network. Genome Biol. 2004;5(10):R76. Epub 2004 Sep 15.
5. Berglund AC, Sjölund E, Ostlund G, Sonnhammer EL. InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res. 2008 Jan;36(Database issue):D263-6
6. Bergmann S, Ihmels J, Barkai N. Similarities and differences in genome-wide expression data of six organisms. PLoS Biol. 2004 Jan;2(1):E9. Epub 2003 Dec 15.
7. Bono H, Ogata H, Goto S, Kanehisa M. Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. Genome Res. 1998 Mar;8(3):203-10. Review.
8. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase. Methods Mol Biol. 2007;406:89-112.
9. Brown K., Jurisica I. Online predicted human interaction database. Bioinformatics, 2005; 21(9):2076-2082.
10. Butterworth R, Simovici DA, Santos GS, Ohno-Machado L. A greedy algorithm for supervised discretization. J. Biomed. Informatics. 2004. 37: 285-292.
11. Date S.V., Marcotte E.M. Protein function prediction using the protein link explorer (PLEX). Bioinformatics. 2005; 21(10):2558-9.
12. Deane CM, Salwinski L, Xenarios I, Eisenberg D. Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics. 2002 May;1(5):349-56.
13. Drawid A, Gerstein M. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. J Mol Biol. 2000 Aug 25;301(4):1059-75.
14. Fayyad UM. On the induction of decision trees for multiple concept learning. Ph.D. thesis, University of Michigan, 1991.
15. Fraser HB, Hirsh AE, Wall DP, Eisen MB. Coevolution of gene expression among interacting proteins. Proc Natl Acad Sci U S A. 2004 Jun 15;101(24):9033-8. Epub 2004 Jun 2.
16. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol. 2000;7(3-4):601-20.
17. Friedman N., Geiger D., Goldszmidt M. Bayesian network classifiers. Machine Learning, 29, 131–163 (1997).
18. Gene Ontology Consortium. The Gene Ontology project in 2008. Nucleic Acids Res. 2008 Jan;36(Database issue):D440-4.
19. 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, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL Jr, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM. A protein interaction map of Drosophila melanogaster. Science. 2003 Dec 5;302(5651):1727-36. Epub 2003 Nov 6.

20. Glazko GV, Mushegian AR. Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. Genome Biol. 2004;5(5):R32. Epub 2004 Apr 27.
21. Green M.L., Karp P.D. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. BMC Bioinformatics. 2004 Jun 9;5:76.
22. Griffiths-Jones S, Saini HK, Dongen SV, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res. 2007 Epub Nov 8.
23. Guiasu S. Information Theory with Applications, McGraw-Hill, New York (1977).
24. Hahn A, Rahnenfuhrer J, Talwar P, Lengauer T. Confirmation of human protein interaction data by human expression data. BMC Bioinformatics. 2005 May 6;6(1):112.
25. Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell. 2006 Jan 13;124(1):47-59.
26. Heckerman D. 1995 A tutorial on learning Bayesian networks. Technical report MSR-TR-95-06, Microsoft Research.
27. Hober S. and Uhlen M. Human protein atlas and the use of microarray technologies. Curr Opin Biotechnol. 2008.
28. Hulsen T, Huynen MA, de Vlieg J, Groenen PM. Benchmarking ortholog identification methods using functional genomics data. Genome Biol. 2006;7(4):R31. Epub 2006 Apr 13.
29. Huttenhower C, Troyanskaya OG. Bayesian data integration: a functional perspective. Comput Syst Bioinformatics Conf. 2006;:341-51.
30. Imai KS, Levine M, Satoh N, Satou Y. Regulatory blueprint for a chordate embryo. Science. 2006 May 26;312(5777):1183-7.
31. Imoto S, Sunyong K, Goto T, Aburatani S, Tashiro K, Kuhara S, Miyano S. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. Proc IEEE Comput Soc Bioinform Conf. 2002;1:219-27.
32. Jiang T., Keating A.E. AVID: an integrative framework for discovering functional relationships among proteins. BMC Bioinformatics 2005;6:136.
33. Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A, Holstege FC. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. Mol Cell. 2002 May;9(5):1133-43.
34. Klammer M, Roopra S, Sonnhammer EL. jSquid: a Java applet for graphical on-line network exploration. Bioinformatics, 2008 24:1467-1468
35. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrín-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF.Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature. 2006 Mar 30;440(7084):637-43. Epub 2006 Mar 22.

36. Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. Science. 2004 Nov 26;306(5701):1555-8.
37. Lehner B, Fraser AG. A first-draft human protein-interaction map. Genome Biol. 2004;5(9):R63. Epub 2004 Aug 13
38. Lu LJ, Xia Y, Paccanaro A, Yu H, Gerstein M. Assessing the limits of genomic data integration for predicting protein networks. Genome Res. 2005 Jul;15(7):945-53.
39. Marcotte E.M. Computational genetics: finding protein function by nonhomology methods. Curr Opin Struct Biol. 2000 Jun;10(3):359-65.
40. Maslov S, Sneppen K. Specificity and stability in topology of protein networks. Science. 2002 May 3;296(5569):910-3.
41. Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, Theesfeld CL, Dolinski K, Troyanskaya OG. Discovery of biological networks from diverse functional genomic data. Genome Biol. 2005;6(13):R114.
42. Myers CL, Troyanskaya OG. Context-sensitive data integration and prediction of biological networks. Bioinformatics. 2007 Sep 1;23(17):2322-30. Epub 2007 Jun 28.
43. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999 Apr 13;96(8):4285-8.
44. Remm M., Storm C.E., and Sonnhammer E.L.L. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 314: 1041-1052.
45. Sartor MA, Zorn AM, Schwanekamp JA, Halbleib D, Karyala S, Howell ML, Dean GE, Medvedovic M, Tomlinson CR. A new method to remove hybridization bias for interspecies comparison of global gene expression profiles uncovers an association between mRNA sequence divergence and differential gene expression in Xenopus. Nucleic Acids Res. 2006 Jan 5;34(1):185-200. Print 2006.
46. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498-504.
47. Snitkin ES, Gustafson AM, Mellor J, Wu J, DeLisi C. Comparative assessment of performance and genome dependence among phylogenetic profiling methods. BMC Bioinformatics. 2006 Sep 27;7:420.
48. Sonnhammer EL, Koonin EV. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet. 2002; 18(12):619-20.
49. Steuer R, Kurths J, Daub CO, Weise J, Selbig J. The mutual information: detecting and evaluating dependencies between variables. Bioinformatics. 2002;18 Suppl 2:S231-40.
50. Sun H, Palaniswamy SK, Pohar TT, Jin VX, Huang TH, Davuluri RV. MPromDb: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-chip experimental data. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D98-103.
51. Sun J, Xu J, Liu Z, Liu Q, Zhao A, Shi T, Li Y. Refined phylogenetic profiles method for predicting protein-protein interactions. Bioinformatics. 2005 Aug 15;21(16):3409-15. Epub 2005 Jun 9.
52. Suthram S, Shlomi T, Ruppin E, Sharan R, Ideker T. Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci U S A. 2005 Feb 8;102(6):1974-9. Epub 2005 Feb 1.
53. Tatusov R.L., Fedorova N.D., Jackson J.D., Jacobs A.R., Kiryutin B., Koonin E.V., Krylov D.M., Mazumder R., Mekhedov S.L., Nikolskaya A.N., Rao B.S., Smirnov S., Sverdlov A.V., Vasudevan S.,

Wolf Y.I., Yin J.J., Natale D.A. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. Sep 11;4(1):41.

54. Tatusov R.L., Koonin E.V. and Lipman D.J. 1997. A genomic perspective on protein families. Science, 278, 631–637.
55. Teknomo K. Similarity Measurement. 2005, http://people.revoledu.com/kardi/tutorial/Similarity/Jaccard.html
56. Tirosh I, Barkai N. Computational verification of protein-protein interactions by orthologous co-expression. BMC Bioinformatics. 2005 Mar 2;6(1):40.
57. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 2003 Jan 1;31(1):258-61.
58. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D433-7.
59. Weisstein E.W. "Correlation Coefficient." From MathWorld--A Wolfram Web Resource. 1999b, http://mathworld.wolfram.com/CorrelationCoefficient.html
60. Weisstein E.W. "Hamming Distance." From MathWorld--A Wolfram Web Resource. 1999c, http://mathworld.wolfram.com/HammingDistance.html
61. Weisstein E.W. "Mutual Information." From MathWorld--A Wolfram Web Resource. 1999d, http://mathworld.wolfram.com/MutualInformation.html
62. Weisstein E.W. "Spearman Rank Correlation Coefficient." From MathWorld--A Wolfram Web Resource. 1999e, http://mathworld.wolfram.com/SpearmanRankCorrelationCoefficient.html
63. Weisstein E.W."Chi-Squared Test." and “Fisher's Exact Test.” From MathWorld--A Wolfram Web Resource. 1999a, http://mathworld.wolfram.com/Chi-SquaredTest.html, http://mathworld.wolfram.com/FishersExactTest.html
64. Wojcik J. and Schächter V. Protein-protein interaction map inference using interacting domain profile maps. Bioinformatics. 2001; 17(Suppl.1):S296-S305.
65. Wu J, Kasif S, DeLisi C. Identification of functional links between genes using phylogenetic profiles. Bioinformatics. 2003 Aug 12;19(12):1524-30.
66. Xia Y, Yu H, Jansen R, Seringhaus M, Baxter S, Greenbaum D, Zhao H, Gerstein M. Analyzing cellular biochemistry in terms of molecular networks. Annu Rev Biochem. 2004;73:1051-87. Review.
67. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res. 2004 Jun;14(6):1107-18.
68. Zheng Y, Anton BP, Roberts RJ, Kasif S. Phylogenetic detection of conserved gene clusters in microbial genomes. BMC Bioinformatics. 2005 Oct 3;6:243.
69. Zheng Y, Roberts RJ, Kasif S. Genomic functional annotation using co-evolution profiles of gene clusters. Genome Biol. 2002 Oct 10;3(11):RESEARCH0060. Epub 2002 Oct 10.

Supplementary Tables

Data type	Species	Data set reference	PubMed ID / **Gene Expression Omnibus sample ID (GSE#)**	Title
MEX	Human	Su et al., 2004	15075390	Human tissue atlas
	Mouse	Murray et al., 2004	**GSE4301**	Cultured cells under stress
		Su et al., 2004	15075390	Mouse tissue atlas
		Aiba et al., 2006	**GSE4082**	mRNA expression: embryonic and adult stem/progenitor cells
		Bruce J. Aronow	**GSE1701**	mRNA expression: tissues
		Hanlon and Jefcoate	**GSE1123**	Adipogenesis: effects of dioxin
		Hovatta et al., 2005	**GSE3327**	Strain and brain region mRNA
		Hutton et al., 2004	**GSE2168**	mRNA expression: the immune system
		Siddiqui et al., 2005	**GSE4726**	Mouse Atlas of Gene Expression: brain tissues
		Siddiqui et al., 2005	**GSE4726**	Mouse Atlas of Gene Expression: non-brain tissues
		Thackaberry et al., 2005	**GSE2812**	Fetal heart, dioxin treatment
		Zapala et al., 2005	**GSE3594**	Adult brain mRNA expression
	Fly	Li and White, 2003	12852852	Fly-248
	Worm	Kim et al., 2001	11557892	Gene expression map
		Reinke et al., 2003	14668411	Germline-enriched and sex-biased expression profiles
	Yeast	Gasch et al., 2000	11102521	Expression in environmental
		Hughes et al., 2000	10929718	Rosetta expression compendium
	Arabidopsis	Schmid et al., 2005	15806101	Expression map of development
PPI	Human, mouse, rat, fly, worm, *Arabidopsis*, yeast	Kerrien et al., 2007	17145710	IntAct
		Donaldson et al., 2003	12689350	PreBIND
		Bader et al., 2003	12519993	BIND
		Bairoch et al., 2005	15608167	UniProt complex membership annotations
	Human	Mishra et al., 2006	16381900	HPRD
	Yeast	Krogan et al., 2006	16554755	Global landscape of protein complexes

SCL	Human, fly, worm, yeast	Bairoch et al., 2005	15608167	UniProt localization
	Human, mouse, rat, fly, worm, yeast, *Arabidopsis*	Pierleoni et al., 2006	17108361	eSLDB
		Huh et al., 2005	14562095	Yeast localizations
PEX	Human	Hober and Uhlen, 2008	18187316	Human protein atlas, tissues
				Human protein atlas, cell lines
TFB	Human	Hallikas et al., 2006	16413481	EEL: mammalian enhancer prediction
	Human	Sun et al., 2006	16381984	MPromDb
	Fly, worm, yeast, *Arabidopsis*	Yu et al., 2004	15173116	Protein-DNA regulogs
MIR	Human, mouse, rat, fly, worm	MicroCosm	http://microrna.sanger.ac.uk/targets/v5/	miRBase targets, v5
DOM	Human	Rhodes et al., 2005	16082366	Domain interactions

Supplementary Table 1. Data sources used as input in the FunCoup database.
3-letter codes stand for:

- mRNA co-expression (MEX)
- protein-protein interaction (PPI)
- sub-cellular co-localization (SCL)
- protein co-expression (PEX)
- shared transcription factor binding (TFB)
- co-miRNA regulation by shared miRNA targeting (MIR)
- domain associations (DOM)

Design	Data sources included in the test[1]	No. of replicates	Factors	Factor levels	Significant effects of interest	Measured outcome
Optimal use of profile similarity metrics on mRNA expression (MEX) data	PLC, SRC coefficients, and MI value of mRNA co-expression in the tested species and other organisms via orthologs	3	A, organism	hsa, dme, cel, sce	C**, D**, E**, CxD**, CxE**, DxE**	$AUC_{Spec[75…90\%]}$, $AUC_{Spec[90…96\%]}$, $AUC_{Spec[96…100\%]}$
			B, Type of FC	FC-PI, FC-ML		
			C, Pearson LC	On, off		
			D, Spearman RC	On, off		
			E, mutual information	On, off		
Optimal discretisation algorithm	All available[2], but discretization only applied to co-expression and co-localization	9	A, organism	hsa, cel, sce	C**, D**, CxD**	$AUC_{Spec[75…90\%]}$, $AUC_{Spec[90…100\%]}$, $AUC_{Spec[99…100\%]}$
			B, Type of FC	FC-PI, FC-ML		
			C, Discretization algorithm	chi-square, quantile		
			D, Allowed no. of bins	2, 3, 4, 6, 10, 15		
Optimal FunCoup configuration	All available[2]	7	A, organism	hsa, dme, cel, sce	A**, B**, C**, D**, E**, AxB**, CxD**, AxCxD**, BxCxD**	$AUC_{Spec[75…90\%]}$, $AUC_{Spec[90…96\%]}$, $AUC_{Spec[96…100\%]}$
			B, Type of FC	FC-PI, FC-ML		
			C, Mode of using othologs	1) seed orthologs only ("Seed to seed") 2) inparalogs pooled ("All pooled"), 3) inparalogs separated in 4 groups: a) seed-to-seed, b) seed-to-additional, c) additional-to-seed, d) additional-to-additional ("All separated")		
			D, Estimating alternative inparalog pairs	best, mean		
			E, Likelihood confidence check	enabled, disabled		
Likelihood values by different types of ortholog	MEX	9	A, organism	hsa, dme, cel	B**, C**, D**, BxC**, BxD**, BxCxD**	Log-likelihood of FC given the evidence
			B, FC type	FC-PI, FC-ML		

			C, Types of orthologs in the link	1) seed-to-seed, 2) seed-to-additional, 3) additional-to-seed , 4) additional-to-additional, 5) the organism's own data (no orthologs)		
			D, Bins after discretization	1, 2, 3, 4		
Assessing limits of data integration to predict ML type of FC	All available for FC-ML class	10	A, organism	hsa, dme, cel, sce	B**, AxB*	$AUC_{Spec[75…90\%]}$, $AUC_{Spec[90…96\%]}$, $AUC_{Spec[96…100\%]}$
			B, set of evidence features	a) Random set of 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44 features b) All features for the tested species + all features for a random set of 1, 2, 3, or 4 model organisms		
Differential FC type prediction	All available[2]	6	A, organism	hsa, dme, cel, sce	C**, AxC**, BxC**	$AUC_{Spec[75…90\%]}$, $AUC_{Spec[90…96\%]}$, $AUC_{Spec[96…100\%]}$
			B, FC type	FC-PI, FC-ML		
			C, NBN model applied	"To the proper FC type", "To the wrong FC type"		

Supplementary Table 2. Experimental designs used to define the optimal configuration of FunCoup.

See aims and interpretations of the particular experiments in *Methods* and *Suppl. Methods*.

The designs were orthogonal combinations of factors where every combination of factor levels was tested in the same number of replicates (column 3). The column 2 shows the set of evidence features fed to FunCoup. Main effects of the factors as well as effects of their interactions were estimated by measuring the outcome. In ANOVA, an effect is significant if the variability inferred by it is significantly exceeds a reference term (usually the residual, or "error" one).

To test the factors, particular option levels were chosen. For the tested organism (e.g. levels human, fly, worm, yeast) and FC class (FC-PI vs. FC-ML), the levels were considered *random* – to determine the variability that the choice species or FC definition may infer in general, i.e. in other cases. For the other factors, the levels were considered *fixed*, thus exactly evaluating the effects of the tested features (e.g., the four modes of using orthologs were fixed levels, as we wanted to analyze *specifically these* options). Being aware of which factors are fixed / random allowed choosing the proper statistical model in ANOVA.

[1] alternative inparalogs were pooled - unless another is explicitly stated.

[2] for prediction of FC-PI in any species, its own PPI data of any kind were not employed.

* effect is significant at $p_0<0.05$.

** effect is significant at $p_0<0.01$.

Species	Genomic locations, amino acid sequences	Original no. of entries
Homo sapiens	http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeId=25	29355
Mus musculus	http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeId=59	32833
Rattus norvegicus	http://www.ebi.ac.uk/integr8/FtpSearch.do?orgProteomeId=122	28682
Drosophila melanogaster	ftp://flybase.net/genomes/Drosophila_melanogaster/dmel_RELEASE3-1/GFF/whole_genome_annotation_dmel_RELEASE3-1.GFF.gz, ftp://ftp.fruitfly.org/pub/download/dmel_RELEASE3-1/FASTA/whole_genome_translation_dmel_RELEASE3-1.FASTA.gz	18498
Caenorhabditis elegans	http://www.wormbase.org/CHROMOSOME?.GFF, ftp://ftp.sanger.ac.uk/pub/databases/wormpep/old_wormpep114/wormpep114	22221
Caenorhabditis briggsae	ftp://ftp.ensembl.org/pub/current_cbriggsae/data/fasta/pep/Caenorhabditis_briggsae.CBR25.pep.fa	14233
Anopheles gambiae	ftp://ftp.ensembl.org/pub/current_mosquito/data/fasta/pep/Anopheles_gambiae.MOZ2a.pep.fa	16148
Saccharomyces cerevisiae	ftp://genome-ftp.stanford.edu/pub/yeast/data_download/chromosomal_feature/s_cerevisiae.gff3, ftp://ftp.ebi.ac.uk/pub/databases/SPproteomes//fasta_files/proteomes/4932.FASTAC	6017
Schizosacharomyces pombe	http://www.ebi.ac.uk/proteome/index.html?http://www.ebi.ac.uk/proteome/YEAST/, ftp://ftp.sanger.ac.uk/pub/yeast/pombe/Protein_data/pompep	5408
Candida albicans[1]	ftp://ftp.pasteur.fr/pub/GenomeDB/CandidaDB/FlatFiles/CALBI.embl	5892
Arabidopsis	ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/PSEUDOCHROMOSOMES/	27436
Oryza sativa	ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_3.0/all_chrs/	61250

Supplementary Table 3. The genome versions used in FunCoup.

[1] Mandatory citation: "Nucleotide sequence data for *Candida albicans* were obtained from the Stanford Genome Technology Center website at http://www-sequence.stanford.edu/group/candida. Sequencing of *C. albicans* was accomplished with the support of the NIDR and the Burroughs Wellcome Fund. Information about coding sequences and proteins was obtained from CandidaDB available at http://www.pasteur.fr/Galar_Fungail/CandidaDB/ which has been developed by the Galar Fungail European Consortium (QLK2-2000-00795)."

FC class	FC-SL								FC-ML								FC-PI					FC-CM		
Conditions	Be a member of a KEGG *signaling* pathway AND (*either* (Be a member of ≥ ***m*** *any* KEGG pathways) OR (the KEGG signaling pathway has < ***n*** members))								Be a member of a KEGG *metabolic* pathway AND (*either* (Be a member of ≥ ***m*** *any* KEGG pathways) OR (the KEGG metabolic pathway has < ***n*** members))								Be reported in one PPI experiment (source: BIND, HPRD, IntAct etc.) AND (be co-members of a KEGG pathway OR be reported in another PPI experiment)					Be named members of the same protein complex in the UniProt annotations.		
	m / ***n*** / *#produced links*																*#produced links*							
Species	hsa	mmu	rno	cin	dme	cel	sce	ath	hsa	mmu	rno	cin	dme	cel	sce	ath	Has	mmu	dme	cel	sce	hsa	cel	sce
	2/20/22145	2/20/17563	2/36/9064		2/42/3841	2/48/ 2811	2/27/1286	3/44/3499	2/45/24191	2/45/20881	2/45/9017		2/40/9892	2/30/6001	2/36/10229	3/40/20772	6405	1971	5424	3160	9875	2861	301	677

Supplementary Table 4. Compilation of high-confidence training sets for FunCoup.
To find a trade-off between sample size and the quality, a specific set of condition parameters were used for each FC class and species. Parameters ***m***, ***n***, and the outcome (the number of protein-protein links selected under the conditions) are shown in the last line.

	Components	Yeast		Human	
		FC-PI	FC-ML	FC-PI	FC-ML
	FBS × {M_spec, M_type}	–	–	–	–
	FBS × <evidence per data type>	–	–	–	–
	FBS × <evidence per species>	–	–	–	–
	FBS × {`M_spec`, `M_type`} × `Max_evid`	–	±	–	–
	FBS × <evidence clade profile>	–	–	–	–
	FBS × <evidence clade configuration>	–	–	–	–

Supplementary Table 5. Search for potential bias in the naïve Bayesian network output after multiple data sources integrated.

Indicators:
–: no effect.
±: a numerically positive but statistically insignificant effect.
In each option the final Bayesian score was decomposed to, and respective indicators were produced for:
`M_spec`: number of species with cumulative BS > FBS / n_s.
`M_type`: number of data sources with cumulative BS > FBS / n_t;
n_s : number of non-empty species;
n_t : number of non-empty data types.
`Max_evid`: *max*(partial BS scores).
<evidence per data type>: {`co-expression, PPI data, sub-cellular co-localization, phylogenetic score`}.
<evidence per species>: {`human, M. musculus, R. norvegicus, D. melanogaster, C.elegans, S. cerevisiae, A.thaliana`}.
<evidence clade profile> : a 3-digit binary indicator of presence/absence of evidence from {`the species itself, a close relative, a distant relative`}. For example, when evidence for human FC was available only from human and mouse, the code was `110`.
<evidence clade configuration> : an integer indicator of the number of {`the master species itself` (can only be `0` or `1`), `a close relative` (e.g. for human: mouse and rat; for yeast: always `0` – as no relatives with data), `a distant relative` (e.g. for human: not mammals)}. For example, when evidence for human FC was available only from human and mouse, rat, and yeast: `121`.

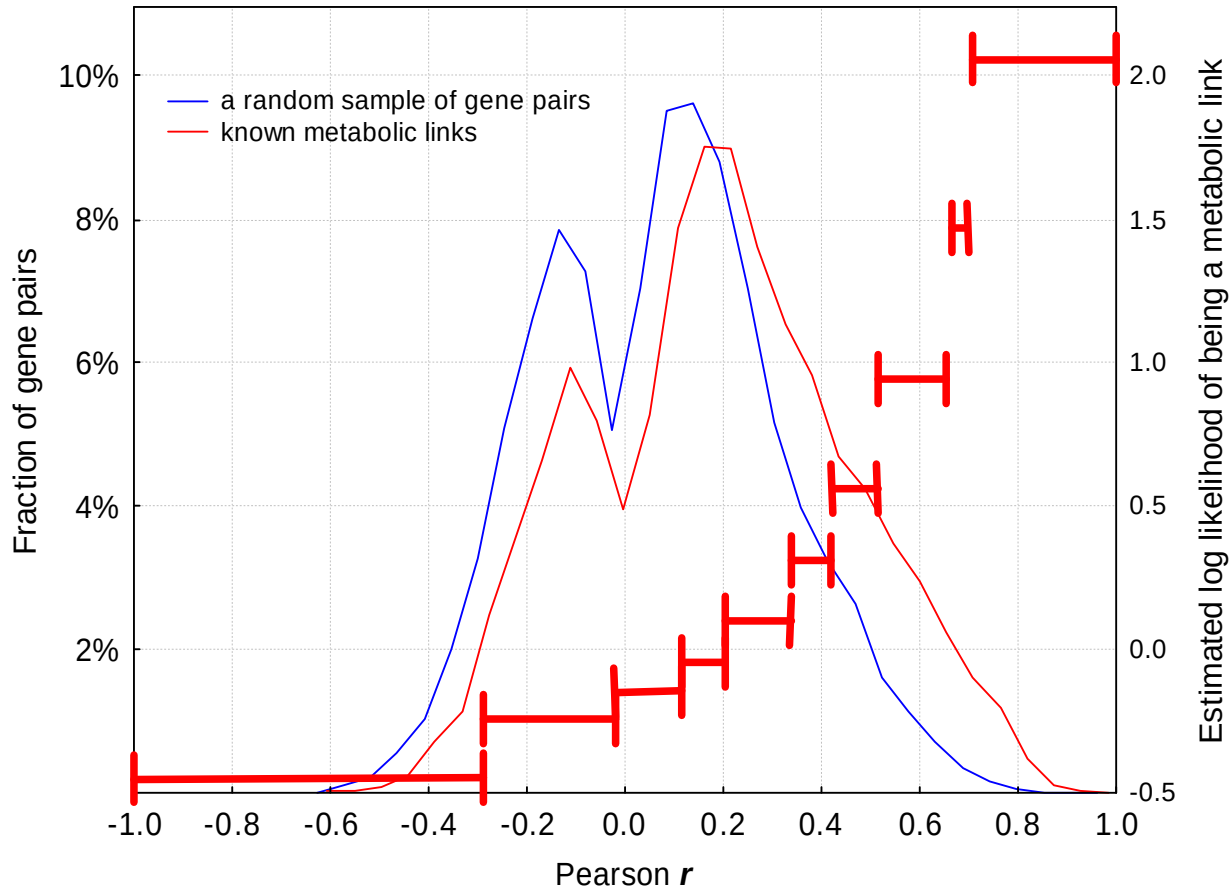
Source	Interactor roles		#pairs
BIND2005	N/A	N/A	34702
HPRD2006	N/A	N/A	32475
IntAct2007	bait	Bait	13
	bait	Prey	13858
	prey	Prey	125392
	prey	Unspecified role	7
	bait	Unspecified role	7
	ancillary	Ancillary	9
	ancillary	Bait	9
	ancillary	Prey	8
	ancillary	Unspecified role	33
	fluorescence accept	Fluorescence donor	27
	neutral component	neutral component	5286
	unspecified role	unspecified role	2858

Supplementary Table 6. Breakdown of sources for human protein-protein interactions as input data in the FunCoup database, categorized by interactor roles.

	Fraction of links, %			
Species	***pfc*(same species)>0.25 and *pfc*(other species)<0.02**	***pfc*(other species)>0.25 and *pfc*(same species)<0.02**	**mixed**	**Total no. of links**
Human	25.4	39.2	35.4	398966
Mouse	21.5	58.2	20.3	229005
Rat	2.9	93.9	3.2	135639
Ciona	0	100.0	0	75965
Fly	19.4	66.5	14.1	118522
Worm	7.8	59.5	32.7	287178
Yeast	52.2	23.6	24.1	77229
Arabidopsis	4.8	74.0	21.2	94285

Supplementary Table 7. Proportion of links “uniquely” predicted by data from the same versus from other species. As most links have weak evidence from most sources, the cutoffs were chosen to approximate uniqueness as well as possible.

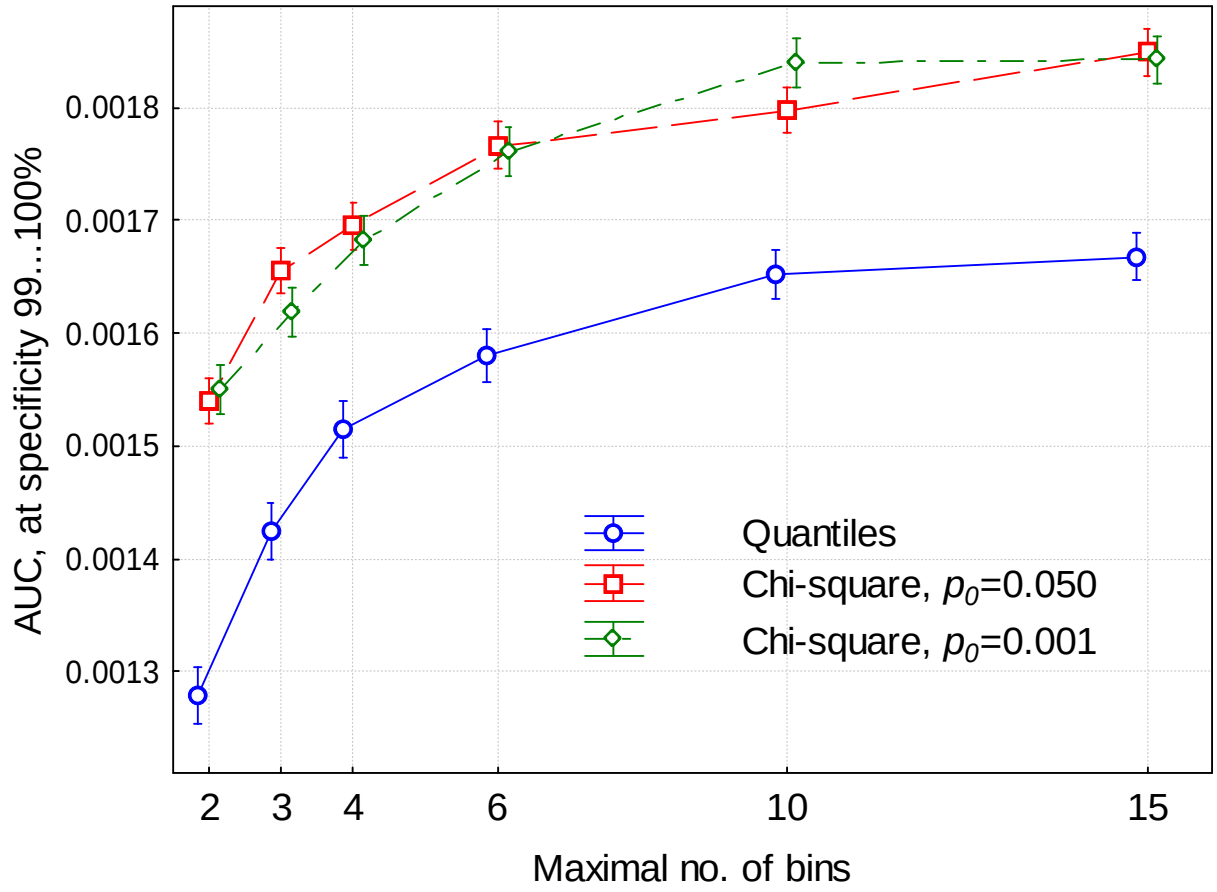
Supplementary Figures



Supplementary Figure 1. Distribution of co-expression metric and estimated likelihood of functional coupling.

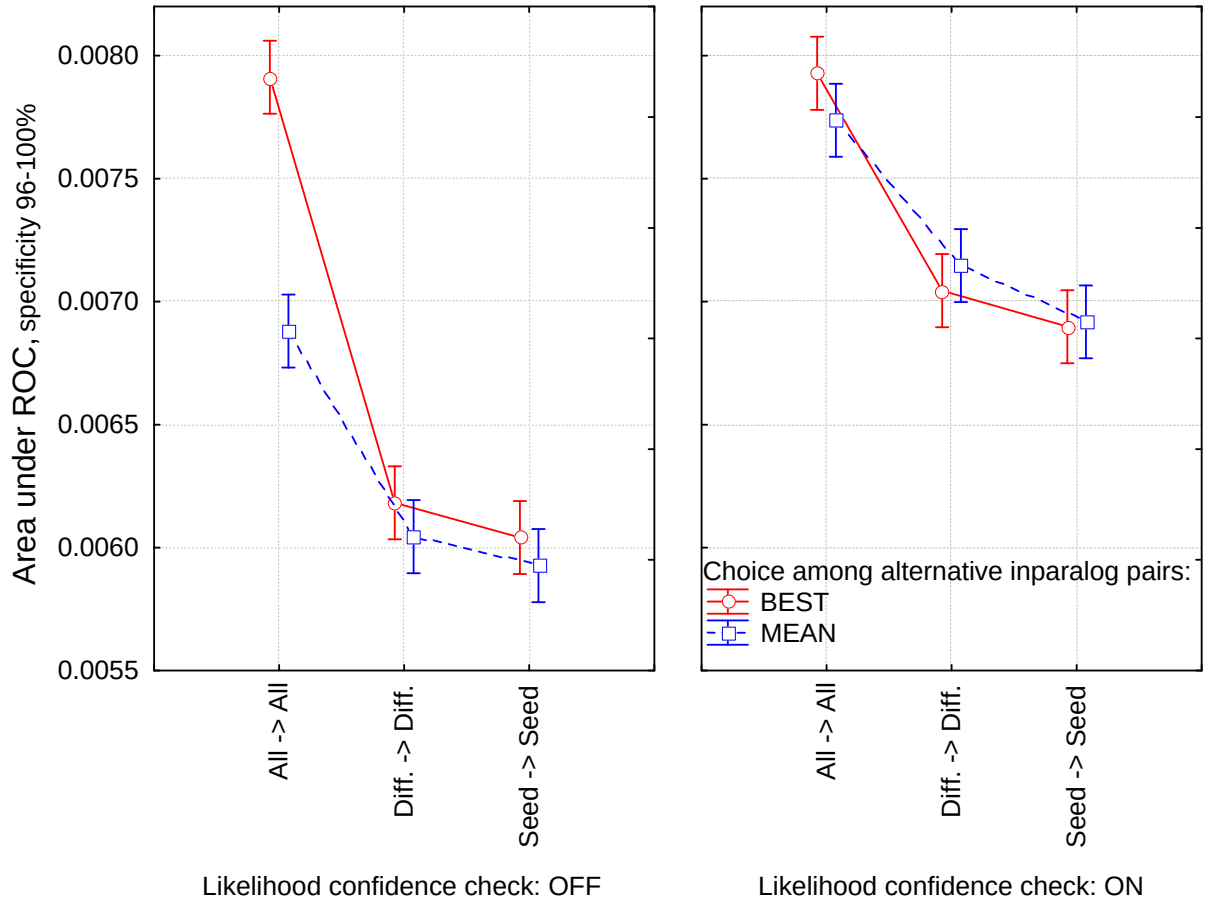
The distribution of Pearson linear correlations (PLC) in functionally coupled gene pairs differs from pairs randomly sampled. Although both tend to follow the Gaussian pattern*, interpreting the differences as evidence of functional coupling is not trivial. In this case (Human Tissue Atlas – Su et al., 2004), our discretisation algorithm established 10 bins in the range of PLC, shown as thick red interval markers. Furthermore, the bins were assigned likelihood values for a novel gene pair to be metabolically linked (either in favour or against it for positive and negative likelihood ratios, respectively).

* The local minimum at PLC≈0 is caused by the policy to take the best (maximally +1 for positive and minimally -1 for negative regions, respectively) of alternative pairs of microarray probes that might occur for some genes.



Supplementary Figure 2. Testing two alternative algorithms for discretisation of continuous features: quantile vs. χ2-based.

In the quantile case, the feature range was split at *n* equal quantity intervals (bins). In the χ2 case, the intervals were found using ratios of FC class presence/absence labels to the left and to the right of the putative cutpoints (see *Methods*). The algorithm was tested on 3 species (yeast, worm, and human), 2 types of FC (PI vs. ML), and varying the number of bins from 2 to 15. The results were processed under general linear model in 3-way ANOVA. The shown curves present the pooled (collapsed species and FC types) result. AUC values for specificity beyond the [99...100] % region produced similar results (not shown for their practical unimportance).

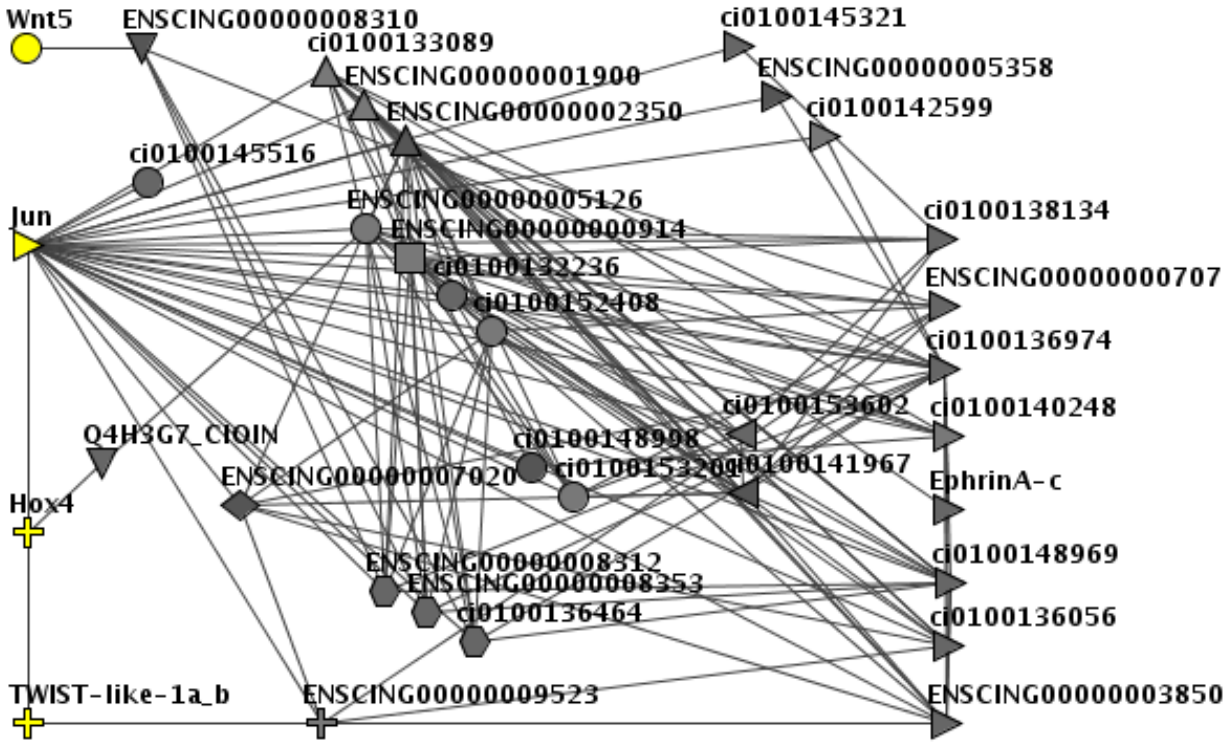


Supplementary Figure3. Optimising FunCoup performance.

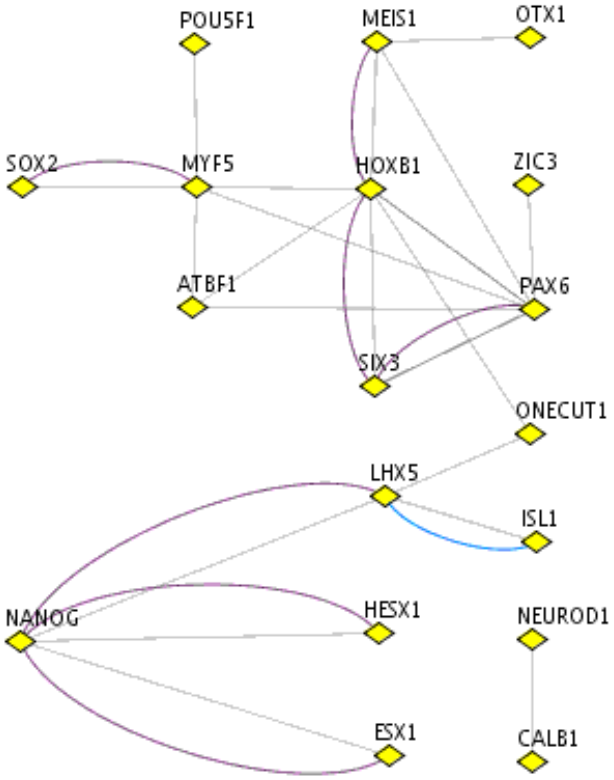
This example shows how performance was analysed for variants of three algorithmic steps in FunCoup:

– Likelihood confidence check (including removal of insignificant nodes from the Bayesian network) - On/Off
– Choice of alternative inparalog pairs - Best/Mean
– way of using orthologs - All-All/Diff-Diff/Seed-Seed

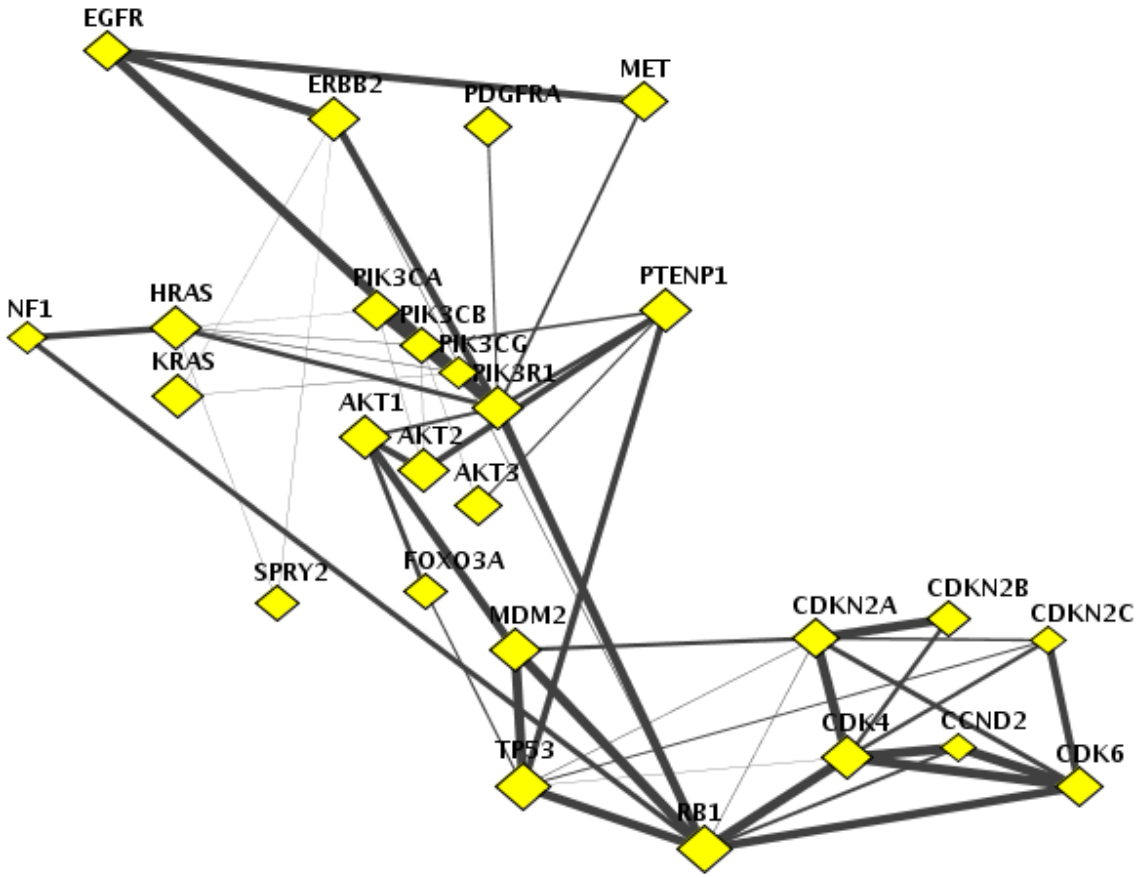
These are partial results of a larger experimental design analyzed with factorial ANOVA, see Supplementary Table 2. The performance was measured as area under the ROC curve. The design allowed estimating both the main effects of the five factors (the three shown here plus FC class and species) and their interactions. The not shown categories were employed as random factors to infer variability due to species and FC class. The plot visualizes the effects of the three first factors (species and FC class were pooled). The whiskers denote 0.95 confidence intervals for the respective group means computed via least squares.



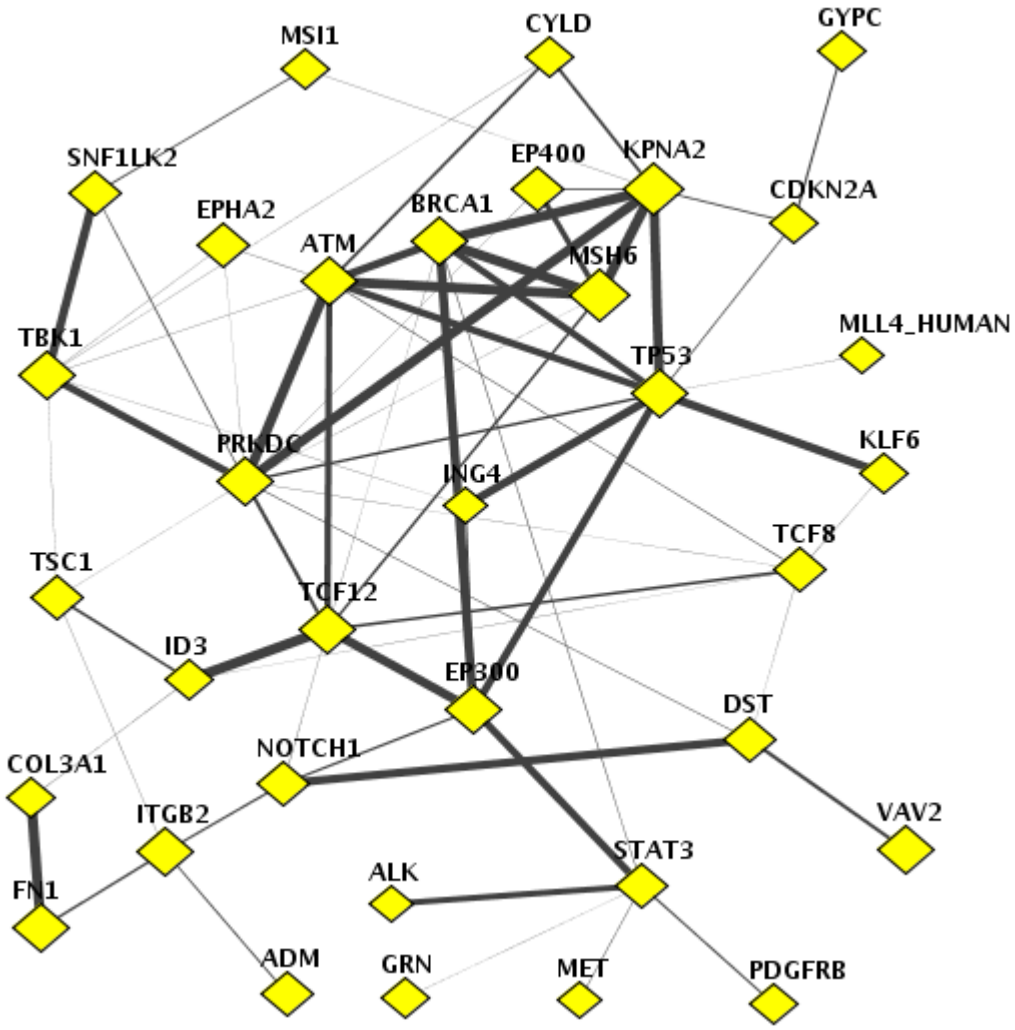
Supplementary Figure 4A



Supplementary Figure 4B



Supplementary Figure 4C



Supplementary Figure 4D

Supplementary Figure 4. Independent validation of FunCoup links.
A. Novel links for genes of the Ciona "Regulatory Blueprint".
These links were discovered as being coupled to a gene of the Ciona Regulatory Blueprint (Imai et al., 2006) under the conditions:

1. confidence *pfc*>0.5;
2. >30% of evidence from vertebrate species;
3. >30% of evidence from invertebrate species.

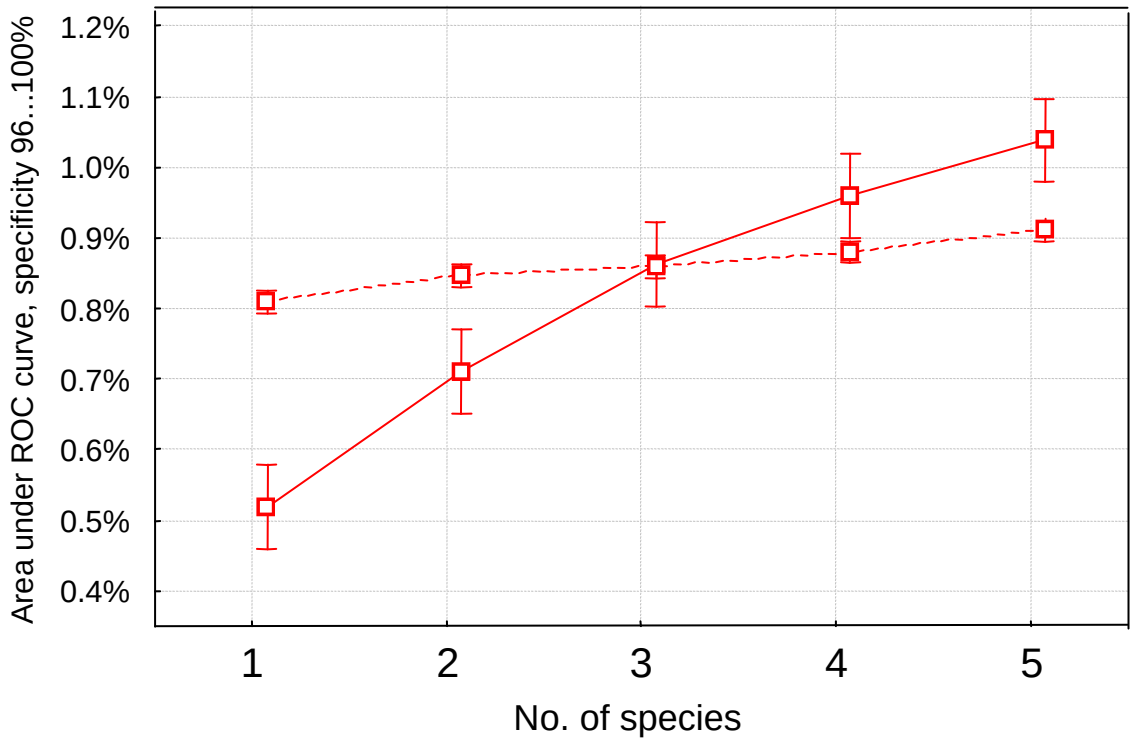
Node legend (functions derived from orthologs in *H. sapiens* by ENSEMBL annotations):
Yellow: Ciona "Regulatory Blueprint" genes;
Grey: other genes;
▽ Receptor;
△ G-protein;
▷ Protein kinase;
◁ Phosphatase;
✚ Transcription factor;
◇ β-catenin;
□ Heat shock protein;
○ Other.

B. Genes of the core transcriptional regulatory network in human ES cells (Boyer et al., 2005) and links between them found by FunCoup.
Violet: Links from the article (Boyer et al., 2005);
Grey: links found by FunCoup;
Blue: pairs of functionally coupled paralogs.

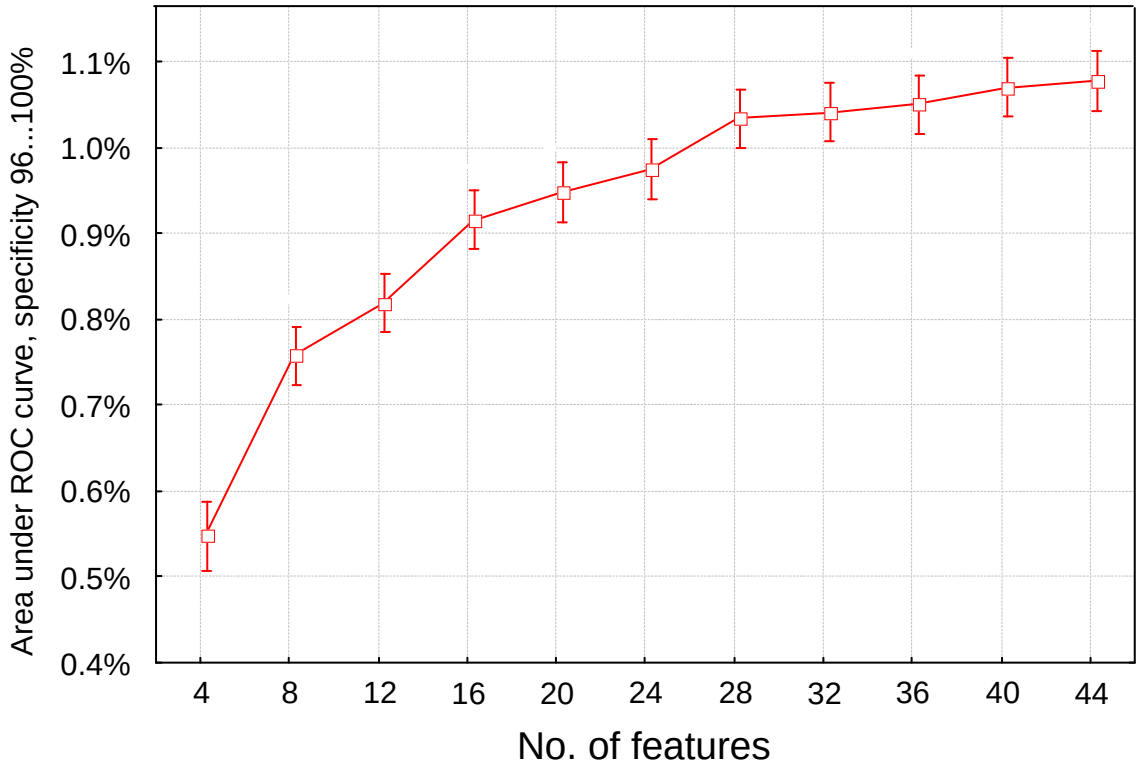
C. FunCoup's perspective on the three critical signaling pathways in glioblastoma development presented by The Cancer Genome Atlas Research Network (TCGARN, 2008). All the nodes from the pathway map in Figure 5 of that paper, including protein families and protein complexes, such as PI(3)K, RAS, AKT, were submitted as *individual genes* in a query to FunCoup (allowing only links *between* the queried genes). The retrieved sub-network recovered all of the connections from the three pathways (29 FunCoup gene-gene links), except 7. Using the jSquid applet at the FunCoup web site, we manually adjusted the layout in accordance with node positions in the original publication.

D. A typical example of FunCoup network of a set of mutated genes found in one *glioblastoma multiforme* tumor sample (downloaded from The Cancer Genome Atlas Data Portal at http://cancergenome.nih.gov/dataportal/data/access/). In total, 145 sets were available. 9 sets (requiring 10 or more genes in each) were selected for the analysis. The example set (barcode TCGA-02-0114-01A-01W) consisted of 53 genes with somatic mutations. 69 links connected 34 genes with each other in the FunCoup human network at *pfc*>0.02 (we ignored links going outside the set). To observe an average expected number of links between 34 genes, a randomization procedure was applied to the network that completely re-wired the network while keeping the nodes' connectivities constant (Maslov and Sneppen, 2002). The link counts from this randomization were normally distributed. We used means and standard deviations to calculate Z-scores and respective p-values. The probability of observing 69 links by chance was, according to the one-sample z-test, $p_0<10^{-12}$ (expected $\hat{x} = 33.3$ links).
From the difference between the expected means and the observed values, we could calculate

the true discovery rate (the fraction of truly existing links among the FunCoup predictions) The point estimate (i.e. ignoring the confidence interval of the sampled mean $\hat{x}$) $TDR = \frac{x - \hat{x}}{x}$ for this particular sub-network equalled 0.51 and 0.59 at *pfc*>0.02 and *pfc*>0.25, respectively. These values are much higher then the formally declared *pfc*, which substituted TDR when the needed parameters were generally unknown.



Supplementary Figure 5A



Supplementary Figure 5B

Supplementary Figure 5. Limits of data integration in interactome discovery.
In this test, we benchmarked the performance of recovering true FC links in the test sets of each species {human, fly, worm, *Arabidopsis*, yeast} with a variable subset of all available evidence data.

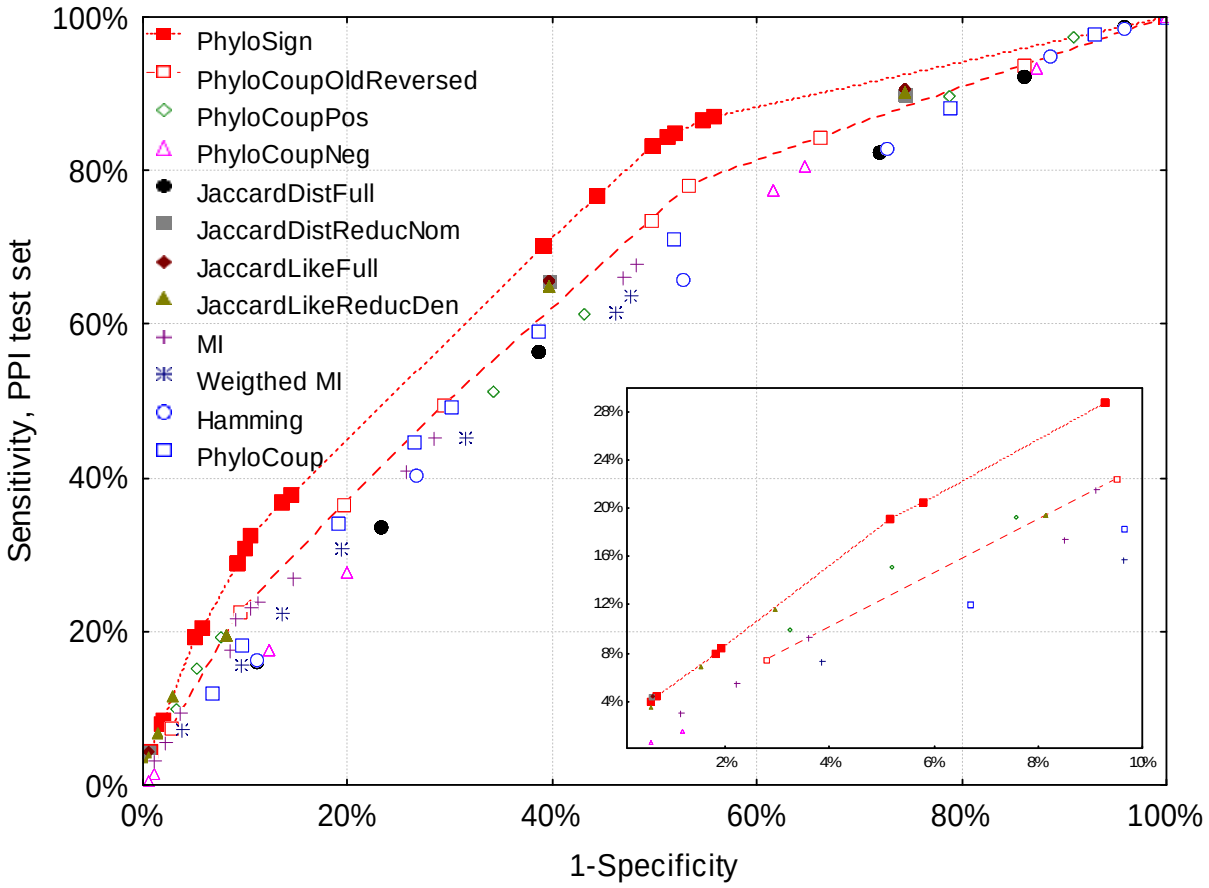
A. By limiting species evidence. The first point is the test result with evidence from only the species which the predictions were made for. Then, data from N other organisms were added, randomly pulled of the 4 available (thus, N = {0, 1, 2, 3, 4}.

B. By limiting datasets. N evidence datasets were randomly selected from the total pool of 45 without regard to the organism (N = {4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44}).

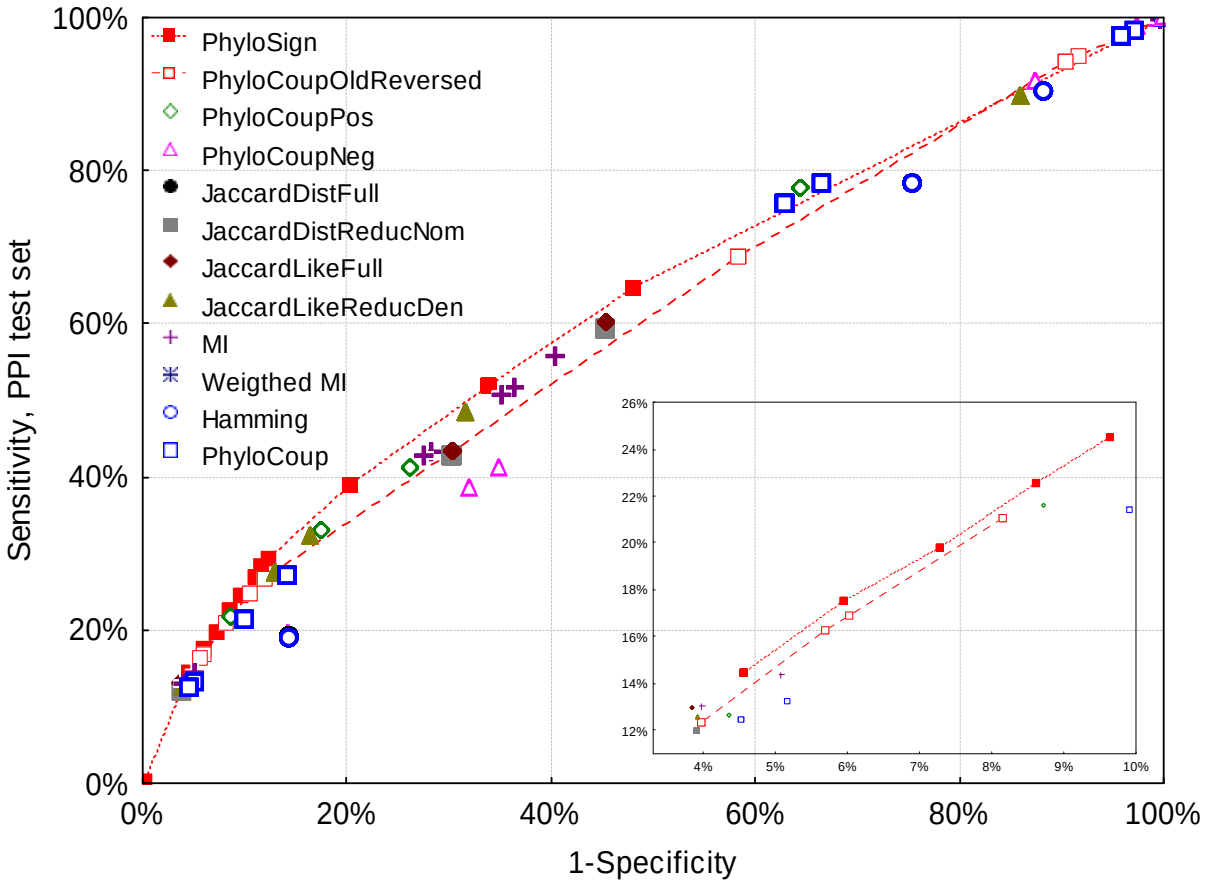
The procedure was repeated 9 times both in A and B.

Dotted line: yeast (pane A only);
Solid lines: pooled results over {human, fly, worm, *Arabidopsis*} (pane A), and over {human, fly, worm, *Arabidopsis*, yeast} (pane B).
Square markers: group means;
Whiskers: 95% confidence intervals of the group means.

Supplementary Figure 6A: *H. sapiens*

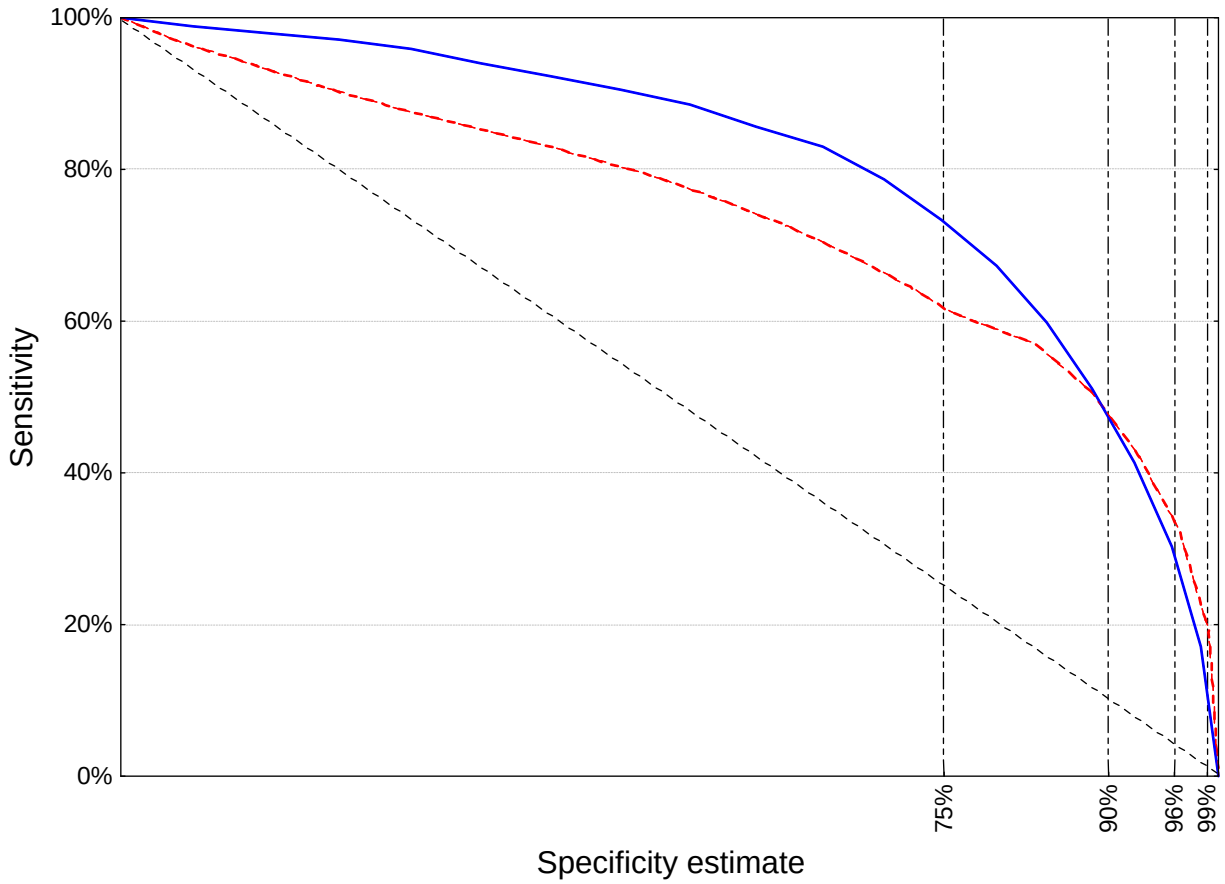


Supplementary Figure 6B: *S. cerevisiae*



Supplementary Figure 6. Relative performance of different phylogenetic profiling scores measured with ROC curves for human and yeast.

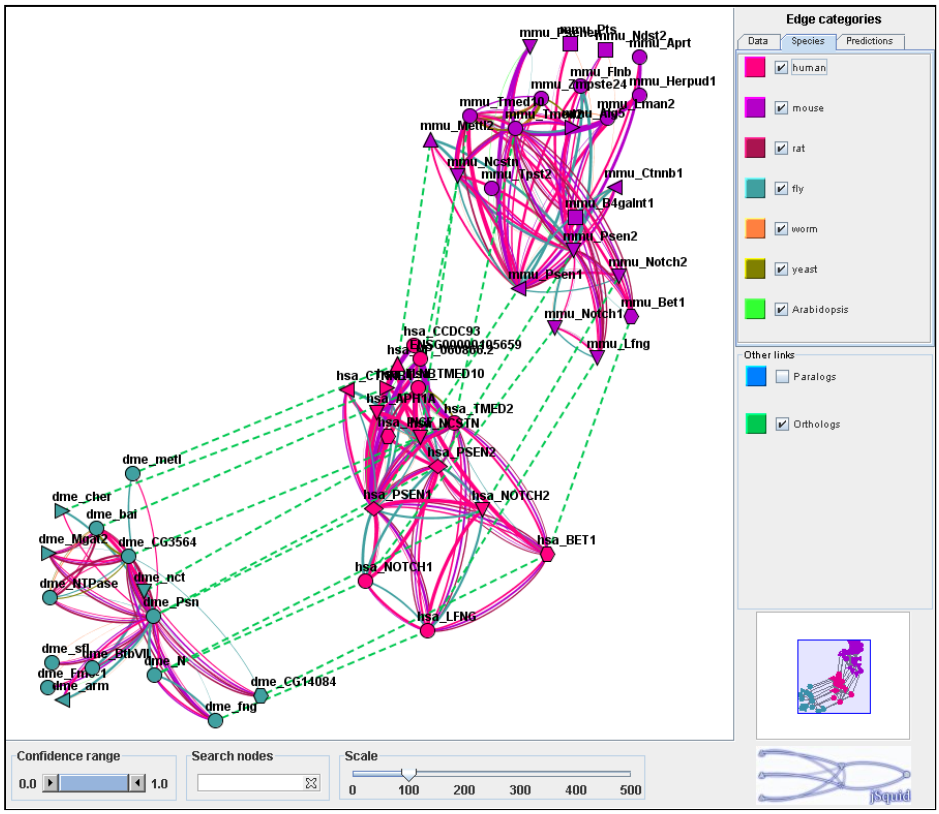
The insets focus on the most relevant areas. Two more earlier suggested scores, Pearson linear correlation PLC and PLC^2 (Glazko and Mushegian, 2004), were also tested but are not shown here due to very poor performance.



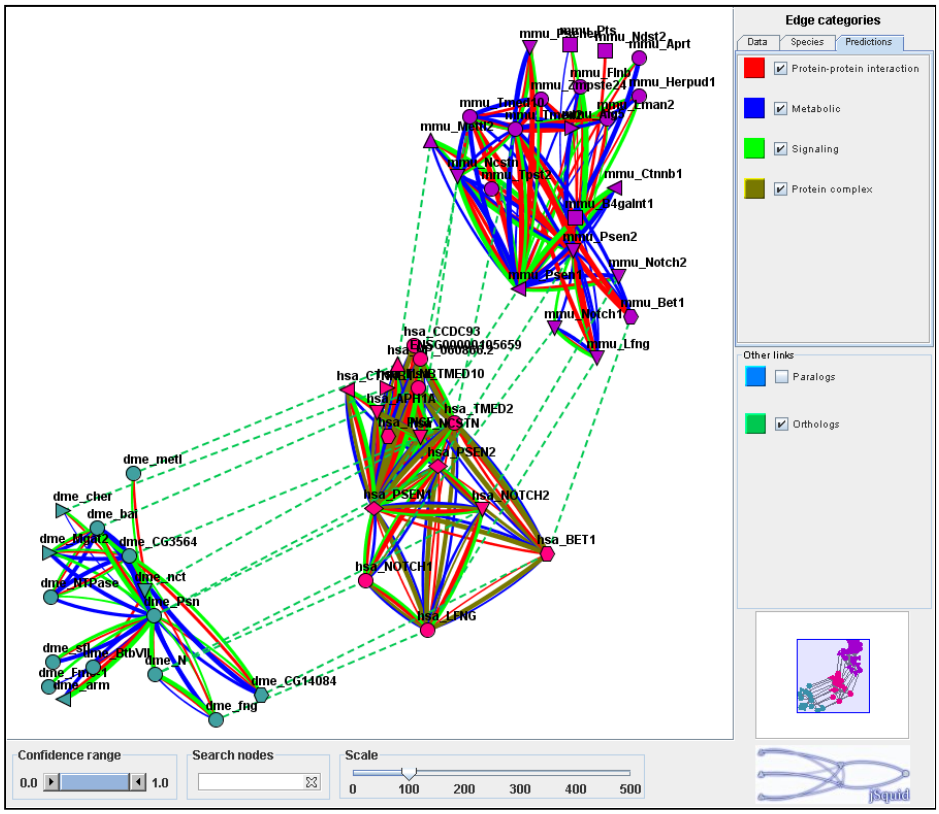
Supplementary Figure 7. Benchmarking FunCoup by receiver operating characteristic (ROC) curves.
The quality of FunCoup predictions was automatically quantified in the "specificity" – "sensitivity" space to enable massive comparison of various effects on the FunCoup predictor. The area under each curve served a performance measure of the predictor. For different purposes, sub-areas delineated with 75%, 90%, 96%, and 99% were considered separately.

Sensitivity: the fraction of known functional links in a test set that the predictor recovered.
Specificity estimate: the fraction of ALL links (unspecified in terms of functional coupling) that the predictor did not find coupled. In our experience, it this is a reliable substitute of a true negative set, which is never perfectly known in practice.

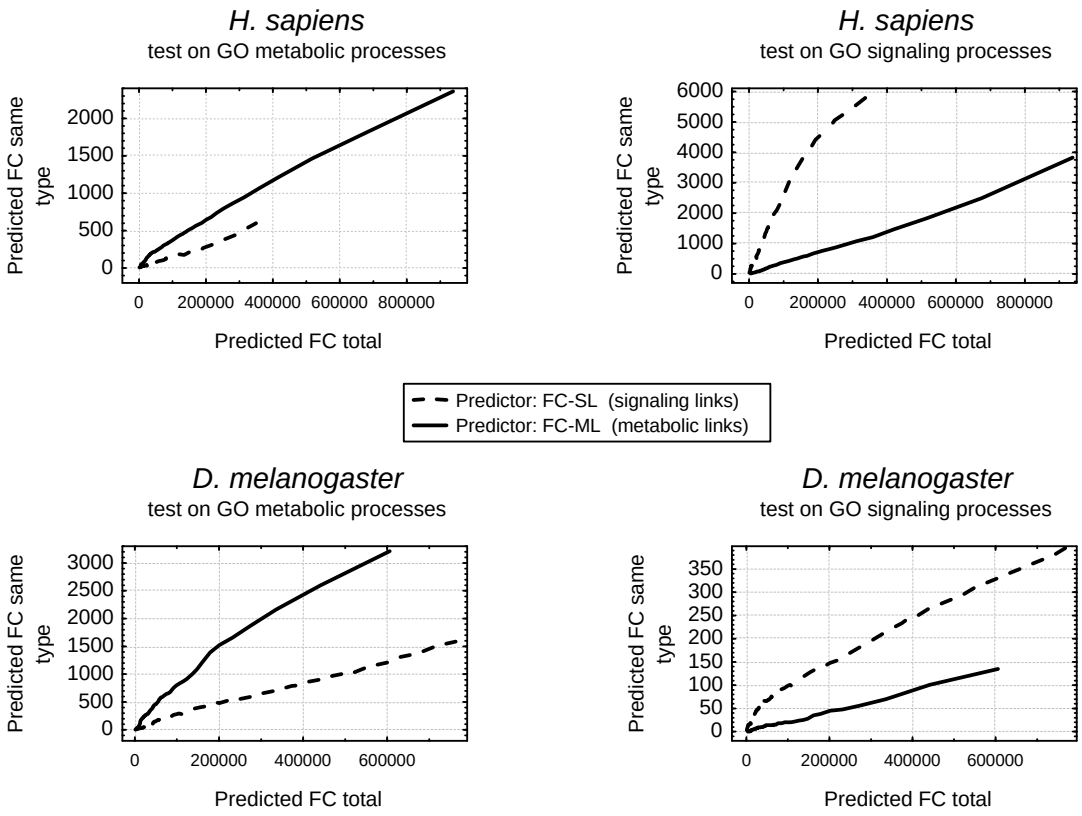
A



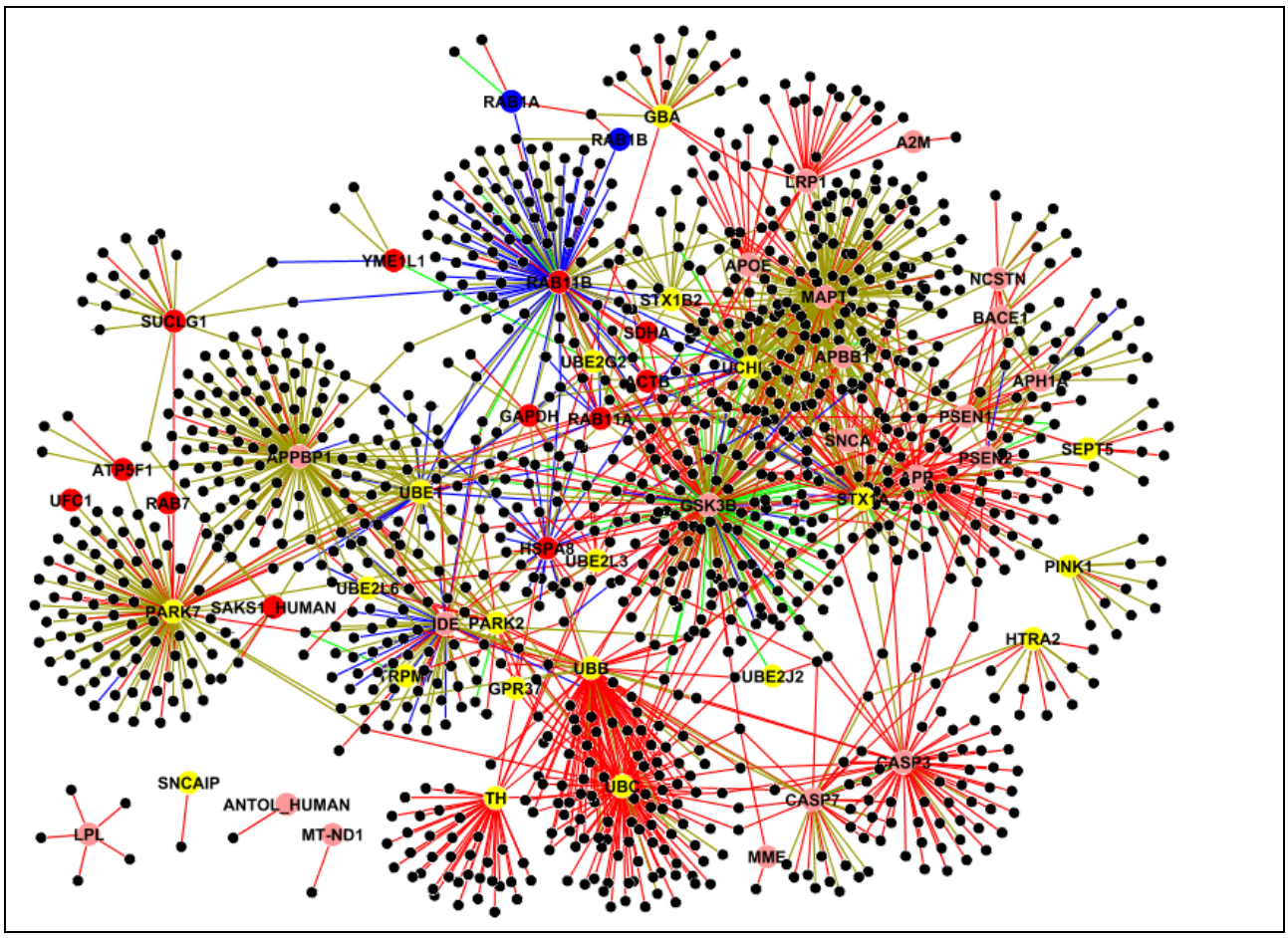
B



Supplementary Figure 8. Comparative network analysis in FunCoup. Subnetworks in human (middle), mouse (top), and fly (left) were generated by submitting human presenilin-1 and –2 (hsa_PSEN1 and 2) to FunCoup, asking for one step of network expansion keeping the 20 strongest links with P>0.5, and inclusion of orthologous subnetworks in mouse and fly. On the right, the colour legend for the links is shown in terms species source (A) or predicted class (B). At the lower right are two newly predicted interactors of the gamma secretase complex, BET1 and LFNG.



Supplementary Figure 9. Validation of predicted functional coupling class for signalling versus metabolic. Links predicted by FunCoup were checked whether both genes belong to the same biological process, here either signalling (GO:0007165 "Signal transduction" or GO:0007267 "Cell-cell signalling") or metabolic (GO:0008152 "Metabolic process") from the GO database (Gene Ontology Consortium, 2008). We trained FunCoup on reference sets from KEGG, either signalling or metabolic pathways. A FunCoup predictor trained on e.g. signalling pathways should thus preferentially find links where both genes are annotated as signalling in GO. As seen in the plots, this ability was clearly superior for predictors trained on the same class compared to those trained on the other class. The separation was also found to grow with confidence (see cross-validation test in Suppl. Table 2 "Differential FC type prediction"). Curves present counts at variable confidence cutoff: from 0.02 (upper right corner) to 1.00 with increment steps=0.01.



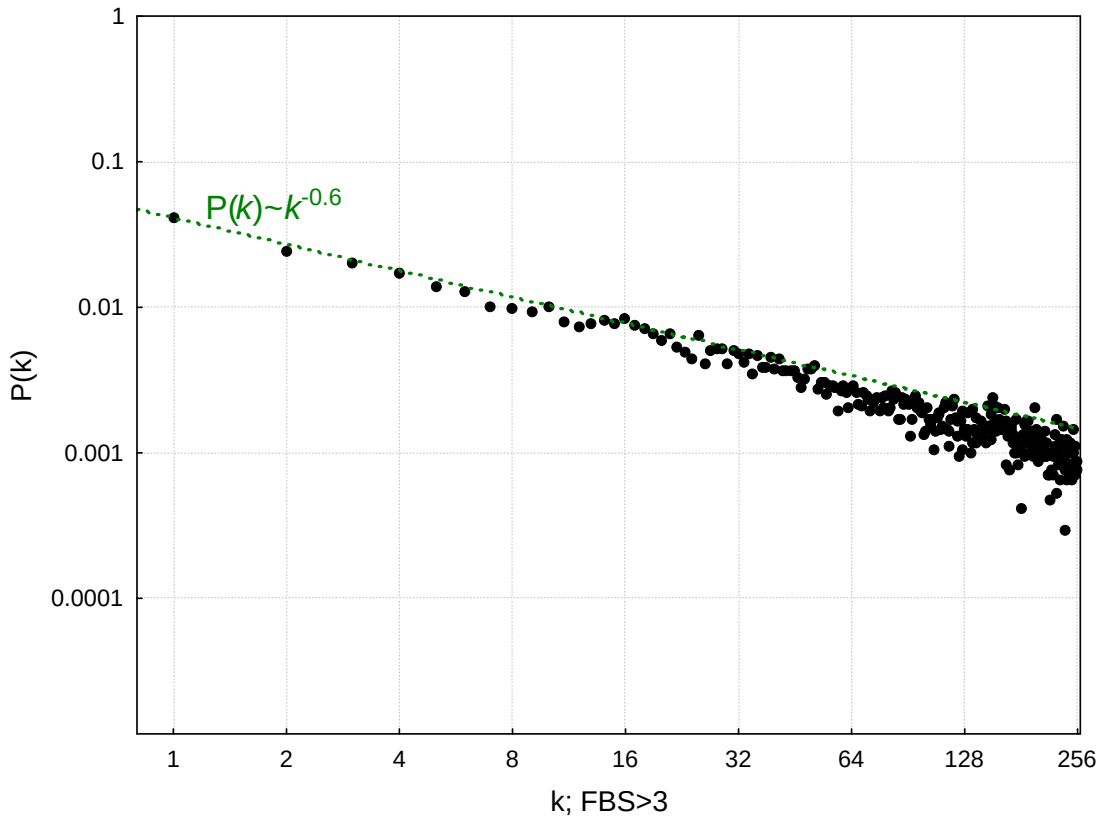
Supplementary Figure 10. View of the novel genes in the context of Parkinson and Alzheimer pathways. The retrieved human sub-network contains all genes linked at confidence >0.5 to:

- Genes of Parkinson and Alzheimer pathways according to KEGG;
- Genes associated with Parkinson and Alzheimer disease according to MIM database;
- The 3 orthologs of a-synuclein toxicity;
- The 12 novel genes.

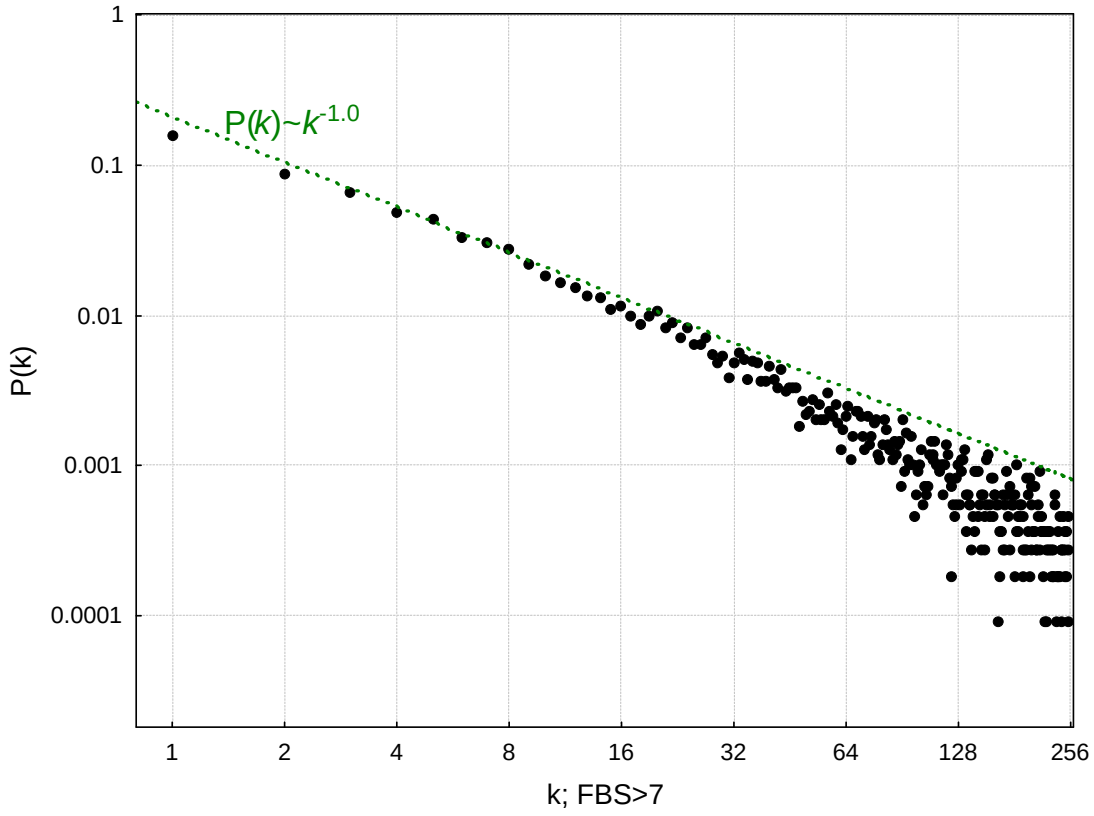
The view was prepared in CytoScape software with the force-directed network layout (weighting for link confidence). Hence, genes placed in the middle tend to be more interconnected with others in the sub-network.

Edge categories (for each protein-protein pair, only the maximally scoring line is shown):
Blue: metabolic links.
Green: signaling links.
Red: protein-protein interactions.
Olive: protein complex members.

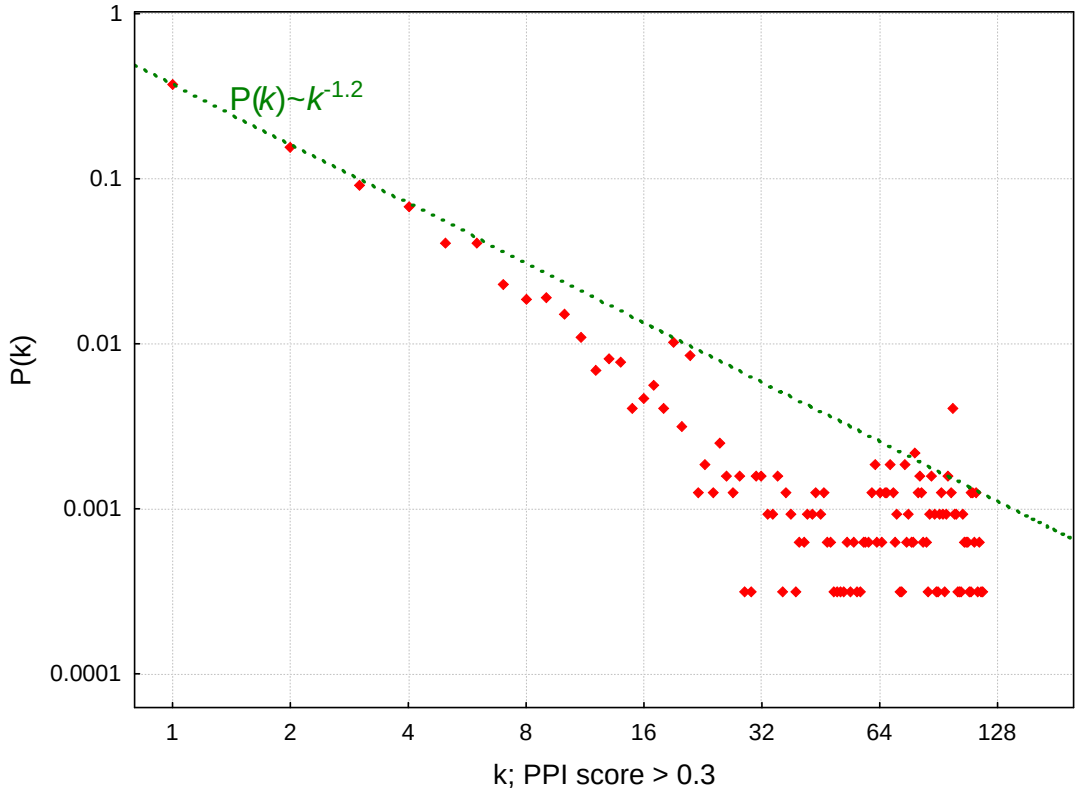
Node categories:
Yellow: pathway KEGG05020 (Parkinson's disease).
Pink: pathway KEGG05010 (Alzheimer's disease).
Orange: Parkinson's disease genes by MIM database.
Magenta: Parkinson's disease genes by MIM database.
Blue: Yeast modifiers of a-synuclein toxicity (squares) and their human orthologs (circles).
Red: 12 novel human PD candidate genes (circles) and their yeast orthologs (squares).
Black: other.


1
0.1
0.01
0.001
0.0001
P(k)
P(k)~k-0.6
1
2
4
8
16
32
64
128
256
k; FBS>3

A


1
0.1
0.01
0.001
0.0001
P(k)
P(k)~k-1.0
1
2
4
8
16
32
64
128
256
k; FBS>7

B



C

Supplemental Figure 11. FunCoup-predicted networks are scale-free.
A and B: Distribution of connectivity (number of network edges per node) in the human network predicted by FunCoup at confidence cutoffs FBS=3 and FBS=7, respectively.
C: Distribution of connectivity in the network of experimentally known human interactions (union of BIND, HPRD, and IntAct databases). The PPI score cutoff = 0.3 guarantees validation in more than one experiment.
The green lines approximate the probability $P(k)$ that a node has k links with a power law function (Barabasi and Albert, 1999).