Global networks of functional coupling in eukaryotes from comprehensive data integration

Andrey Alexeyenko and Erik L. L. Sonnhammer

Genome Res. published online February 25, 2009
Access the most recent version at doi:10.1101/gr.087528.108

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

Abstract

No single experimental method can discover all connections in the interactome. A computational approach can help by integrating data from multiple, often unrelated, proteomics and genomics pipelines. Reconstructing global networks of functional coupling (FC) faces the challenges of scale and heterogeneity - how to efficiently integrate huge amounts of diverse data from multiple organisms, yet ensuring high accuracy?

We developed FunCoup, an optimised Bayesian framework, to resolve these issues. Because interactomes comprise functional coupling of many types, FunCoup annotates network edges with confidence scores in support of different kinds of interactions – physical interaction, protein complex member, metabolic or signalling link. This capability boosted overall accuracy. On the whole, the constructed framework was comprehensively tested to optimise the overall confidence and ensure seamless, automated incorporation of new data sets of heterogeneous types. Using over 50 datasets in seven organisms, and extensively transferring information between orthologs, FunCoup predicted global networks in eight eukaryotes. For the *Ciona intestinalis* network only orthologous information was used, and it recovered a significant number of experimental facts. FunCoup predictions were validated on independent cancer mutation data.

We show how FunCoup can be used for discovering candidate members of the Parkinson and Alzheimer pathways. Cross-species pathway conservation analysis provided further support to these observations. The networks, which are the largest interactome reconstructions to date, are freely available for download and query at http://FunCoup.sbc.su.se. The site allows detailed graphical and tabular analysis of subnetworks around query genes, as well as comparative analysis of orthologous networks in multiple species.

Introduction

The high-throughput functional analysis of genes and proteins is producing vast data resources that, if integrated into interaction networks, will be key to unraveling the function of all genes in an organism (Sonnhammer, 2005). While no single data set provides enough confidence and coverage, much experimental evidence from e.g. protein-protein interactions, mRNA co-expression have been integrated into interaction networks in such organisms as *S. cerevisiae*, *C. elegans*, and *H. sapiens* (Jansen et al., 2003; Troyanskaya et al., 2003; Bader et al., 2004; Lee et al., 2004; Li et l., 2004; Beyer et al., 2007). Srinivasan et al. (2006) also used sequence-derived interaction evidences such as correlated evolution/inheritance and chromosomal co-location to integrate interaction networks in 11 microbes.

However, as of today, using data from one organism alone is insufficient to reconstruct its interaction networks completely. It is possible to expand the data pool by transferring functional information between species via homologs (von Mering et al., 2005; Hahn et al., 2005) or orthologs (Matthews et al., 2001; Rhodes et al., 2005; Zhong and Sternberg, 2006; Hulsen et al., 2006). Furthermore, functional coupling (FC) between two proteins can have several guises: direct physical interaction (PI), protein complex members (CM), links in metabolic pathways (ML), or links in regulatory/signaling pathways (SL).

The data integration is thus multi-dimensional – using multiple evidence types from multiple species for predicting multiple classes of links. This puts high demands on the process, both in terms of computation and automatic parameter optimization for each new dataset. It is thus necessary to develop a universal, fast, and sustainable methodology in order to discover functional connections in many eukaryotic organisms at the global scale.

To achieve this, we adopted the naïve Bayesian network framework (NBN), and advanced many procedures for supervised NBN training in order to make it optimally suited for data integration in FC network reconstruction. The innovations concerned significance testing, continuous score discretisation, orthologous evidence usage, and phylogenetic profiling. A major challenge was to efficiently use data not originally produced for FC discovery. Particular attention was given to data transfer between species via orthologs – the best way to enrich sparse data sources (Rhodes et al., 2005; Hulsen et al., 2006).

These advancements allowed us to efficiently reconstruct comprehensive FC networks for eight of the most important eukaryotes: human, mouse, rat, fly, worm, yeast, *A. thaliana*, and *C. intestinalis*. For these species, they are the largest reconstructed interactomes to date. For *C. intestinalis*, we managed to reconstruct a network in absence of any own large-scale datasets, by transferring such information from coupled orthologs only. We here present both the method and the resulting FunCoup database for discovery and analysis of functional coupling in gene networks. To demonstrate the usefulness of network analysis in FunCoup, we applied it to discover new candidate members of important pathways, including those for Alzheimer's and Parkinson's diseases.

Results

Data integration

The flow of the data integration process in FunCoup is outlined in Figure 1. To infer functional coupling between gene pairs we collected large-scale data of a number of different types:

- mRNA co-expression (MEX)
- phylogenetic profile similarity (PHP)
- 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)

These types of data are most abundantly available for human and the most important eukaryotic model organisms: *M. musculus* (mouse), *R. norvegicus* (rat), *D. melanogaster* (fly), *C. elegans* (worm), *S. cerevisiae* (yeast), and *A. thaliana* (*Arabidopsis*). In total, 51 data sets were collected for these species (see Fig. 1, Supplemental Table 1 and http://FunCoup.sbc.su.se/statistics.html). To avoid direct data redundancy, we made sure that the same piece of experimental information would never appear in different sets.

For each data type, we either adopted, optimized, or invented a metric to reflect the support for FC between gene pairs (Supplemental Methods). Examples of such metrics are the Pearson correlation coefficient (used for MEX), weighted mutual information (SCL), and a new probabilistic score accounting for multiple experimental reports of the same interaction (PPI). This score extracts additional FC evidence from so-called prey-prey interactions (i.e. pairs of proteins coupled indirectly via the bait), which significantly augmented available raw PPI input (Supplemental Table

6). We constructed a novel phylogenetic profile metric based on ortholog conservation in eukaryotes, which has not been done successfully before (Snitkin et al., 2006).

We compiled FC class–specific gold standard training sets (TS) of protein-protein pairs known to belong to the respective kind of coupling. In preliminary training/prediction tests, we successfully predicted links of respective classes by integrating input data of the same collection. However, contribution to the prediction of FC classes differed across input sets. For example, mRNA co-expression usually delivered more information on CM, and experimentally reported interactions were usually more important for SL than ML. (Figure 3).

The naïve Bayesian network (NBN) was trained using TS of four FC classes and input datasets from seven organisms (Fig. 1). First, our novel discretisation algorithm split each continuous metric's range into optimally defined bins for each combination of species, input dataset, and TS. Then, the TS was used to assign probabilistic scores (*likelihood ratios*) to each bin. They were calculated by dividing the evidence occurrence in the positive training set with its background frequency. We predicted functional coupling for every gene pair within a species by summing up the log likelihood ratios of their input data. This summary score, or Final Bayesian Score (FBS), reflects the overall chance of being functionally coupled. However to render the coupling score intuitively clear, we converted it to a probabilistic confidence value *pfc* ranging from 0 to 1 (*Methods*).

Each TS produced a separate prediction model. These dedicated models enabled overall accuracy improvement (Supplemental Table 2) and help to reveal the nature of a predicted link. Supplemental Figure 9 shows that the predicted class agrees well with the GO functional annotation (The Gene Ontology Consortium, 2000). Figure 2 illustrates how FunCoup integrates multiple evidences, and how this leads to substantially increased confidence in the functional coupling. Expectedly, the classes were not mutually exclusive. For example, protein kinases often produced high scores for both PI and SL links. We did not aim at a unique classification, and stored all scores above FBS=3, which constituted overall ~1% of the processed gene-gene pairs (This fraction varies per species and functional class, and can be derived from the detailed on-line statistics page under "Release notes" at http://FunCoup.sbc.su.se).
How much evidence was contributed from other species and from different input data types? As shown in Figure 3, mouse, human, and yeast were most influential. Strikingly, without exception the majority of the evidence came from other species, indicating the great value of transferring FC data between organisms. Even counting the links' unique origin (according to the criteria in Supplemental Table 7), more links came from other species than from the same in all cases but yeast. Among the input data types, mRNA co-expression played a dominating role, followed by phylogenetic profiles, PPI, and sub-cellular localisation.

Optimising NBN performance

To find the optimal combination of parameters and procedures and to make the framework universally applicable, the impact of the introduced features was tested statistically. As the influence of algorithmic details was often dataset-, species-, and FC class-specific, we only accepted modifications that significantly improved the overall performance. This was verified under holdout cross-validation in ANOVA experimental designs, always examining 3-5 species and usually two FC classes.

A problem with classical Bayesian predictors is the lack of a significance criterion, which can lead to non-zero likelihood ratios even in the absence of significant support by the data. We devised a simple significance test to ensure that this does not happen, which noticeably increased the performance (Supplemental methods).

The NBN receives the input data in a discrete form, which is very practical if the likelihood of FC is irregularly distributed over the evidence data (Supplemental Fig. 1). To optimally discretise continuous data, we developed a new dynamic algorithm that is substantially superior to the flat binning (Supplemental Fig. 2). We defined optimal bin borders by comparing the frequencies of positive versus background examples in different score regions. This follows the approach of Butterworth et al. (2004), but avoids a laborious preliminary adjustment of parameters.

One of the most important contributions to infer functional coupling came from orthologs in other species. We used orthologs defined by InParanoid (Berglund et al., 2007), which has been shown to be most accurate for identifying functional counterparts (Hulsen et al., 2006). Nonetheless, the transfer can be done in many ways in case of multiple co-orthologs (inparalogs) stemming from species-specific gene duplication, which is commonplace in eukaryotes. As shown in Figure 4, the functional coupling may evolve in different ways after the duplication. On one extreme, the interaction is limited to single inparalogs, while on the other extreme all inparalogs may interact. By benchmarking alternative methods for using orthologs, we found that the best results are achieved by treating all alternative inparalog pairs from the same cluster equally and by using the best FC score among them (Supplemental Fig. 3). In other words, interactions are generally not considered specific to one gene copy after a duplication. However, after integrating all available FC data, either from the same (or a closely related) species, a particular gene pair may receive much higher confidence.

Independent validation tests

We developed and tested FunCoup under hold-out cross-validation (Supplemental Methods). However, the procedure employed data from the same sources for both training and testing, and thus might have involved higher order train/test dependencies. Looking for an independent proof of the predicted links' validity, we selected published research articles where authors presented local, relatively compact sub-networks related to signaling or regulatory processes.

It was of great interest to reconstruct gene networks in organisms that themselves lack interaction data. FunCoup was particularly well suited to this task because of the integrated transfer of functional coupling via orthologs. To demonstrate this ability, we generated a network in *Ciona intestinalis*, for which no large proteomics or genomics dataset is yet available. As a positive training set, we used pathway members inferred via orthology. The input data came from the seven eukaryotes listed above, and the resulting *Ciona* network contained 38445 links with a confidence *pfc*>0.5, connecting 2683 genes. 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, referred to as the "RBP" network.

We expected to only recover a small fraction of RBP because regulatory links are indirect and very hard to find, and FunCoup was not specifically trained to find such links. Despite this, FunCoup recovered 22 (at *pfc*>0.05) of the maximum achievable 116 RBP links between the 54 RBP genes in the reconstructed FunCoup network. The true discovery rate was actually much higher than 5% (Supplemental Materials). In total, the 54 RBP genes were interconnected by 180 links, which is significantly higher (p_0<10^{-8}) than expected by chance (39.5 links are expected for 54 randomly chosen genes in the reconstructed network).
FunCoup also identified many novel links to *Ciona* RBP. At *pfc*>0.5, 350 new genes were coupled to RBP, and 30 of these had >30% of the evidence from both vertebrates and invertebrates, representing an evolutionarily well-supported set of novel genes in this pathway. Functions in this group included G-proteins, GPCRs, protein kinases, phosphatases etc., i.e. signalling proteins. Most

of the novel genes were coupled to *c-Jun* (*MAPK10*), emphasizing its importance during embryonic development (Supplemental Fig. 4A).

The relatively low FunCoup sensitivity to RBP links emphasizes the elusive nature of the regulatory links related to embryonic development. In fact, they are equally hard to find in the better studied eukaryotes. For comparison, we took a set of "core transcriptional circuitry in human embryonic stem cells" (Boyer et al., 2005). It yielded a similar ratio: FunCoup recovered 7 out of 82 regulatory links shown in the paper (Supplemental Fig. 4B). Hence for this context, links predicted in *Ciona* had competitive confidence.

As a contrast, sensitivity of the FunCoup predictor to cancer-related pathways turned out to be very high. Such circuits are based on physical interactions and complexes produced by GTPases, kinases, and DNA binding/processing proteins, and FunCoup abounded in respective input data. As a test set, we considered a map of three critical signaling pathways in the development of glioblastoma released by The Cancer Genome Atlas Research Network (TCGARN, 2008). Figure 5 in this article depicted mutationally vulnerable segments of RTK/RAS/PI(3)K, p53, and RB pathways. Querying FunCoup with the listed genes retrieved 29 links from this map, and only missed 7 (Supplemental Fig. 4C). 25 more FunCoup links between the mapped genes were not drawn by the TCGARN authors, but may be true novel links relevant to either cancer or normal biology. In total, FunCoup predicted as many as 896 links (*pfc*>0.5) between the pathway members and other, not mapped, genes.

We also wanted to validate FunCoup on a raw, independently chosen, set of functionally related genes, in order to exclude any possibility of overlap between the test set and FunCoup evidence. For this purpose, we retrieved 145 somatic mutation lists, each derived for an individual patient tumour sample of glioblastoma multiforme from The Cancer Genome Atlas. The mutations were found and experimentally validated (genotype arrays, PCR amplification, etc), and made public according to TCGA protocols [http://tcga.cancer.gov/about/index.asp]. A gene list from one tumour will generally contain two kinds of mutations: drivers and passengers. The former drive the tumour towards malignancy while the latter passively accumulate mutations due to impaired DNA repair (Ding et al., 2008). One *a priori* expects the driver genes to be functionally coupled to each other in a given tumour, and passengers to not be. The 9 sufficiently large (10 or more genes) mutation lists were analysed in FunCoup to examine the connectivity. Six of these lists were significantly ($p_0<10^{-3}$) enriched with internal connections compared to random networks. Thus, using completely raw and uncurated datasets, relationships predicted by FunCoup were validated in a majority of somatic mutation cases. Details are described in Suppl. Material, as well as an example network generated from a somatic mutation list. Investigating sub-networks of patient mutation sets with the help of FunCoup can be an important step towards individualized cancer treatment.

Apart from the successful validation on an orthogonal data set, the test also confirmed that our formal *pfc* score was a very conservative, lower-bound estimate of the true discovery rate (TDR), i.e the fraction of true facts among the predictions. In reality, it was probably much higher, e.g. in the TCGA-02-0114-01A-01W subnetwork *TDR*=0.51 at *pfc*>0.02 and *TDR*=0.59 at *pfc*>0.25. Respectively in the *Ciona* RBP case, TDR exceeded 0.20 at *pfc*>0.05 (Supplemental Materials).

Web site

We created a web resource with networks for eight eukaryotic species, from yeast to human: http://FunCoup.sbc.su.se. Large (up to genome-wide) sets of interaction predictions are available for download in CytoScape-compatible (Shannon et al., 2003) or XML format. However, the user is typically interested in smaller sub-networks around a set of query genes.
The user can control the sub-network retrieval and, thus, its information content. The evidence base can be limited to particular species or data types (e.g. "mammalian" or "co-expression" only). The

query's neighborhood can be specified by subnetwork size, confidence threshold, network radius, and neighbor-search algorithm.
The subnetwork is shown with a specially designed Java applet jSquid (Klammer et al., 2008) that allows flexible user-controlled rendering of the network graph, including node grouping by pathway, organelle, or connectivity. Nodes are given a shape and colour according to functional categories. The results are also shown as a table where the support from each evidence data type or species is listed for each link. This decomposition of link support is also possible in the network graph, whereby each category is shown as a distinctly coloured line. Each category can be individually enabled/disabled, and the user can switch between viewing evidence data type, evidence species, and predicted FC class. Moreover, a subnetwork may be retrieved using only evidence of particular types or species, e.g. simulating a pure PPI network or a pure yeast network. It is possible to add pre-defined gene groups from particular functional categories (GO), diseases (OMIM), or pathways (KEGG) to highlight a network context of interest. A multi-species view, retrieved via orthologs, provides across-species network comparisons (Fig. 5).

Cross-species network analysis using FunCoup

We used FunCoup to investigate the protein network around genes known to cause Alzheimer's disease (AD). Starting with human presenilin-1 and -2 as queries, we asked for the subnetworks of functionally coupled genes in human, mouse, and fly (Fig. 5A). The subnetworks in the three species were strikingly similar, sharing many of the known gamma-secretase associated proteins. In all three species we detected two genes that to our knowledge have not previously been associated to AD. These were *BET1* (Swiss-Prot *O15155*) and *LFNG* (Swiss-Prot *Q8NES3*). *BET1* is a SNARE protein involved in the docking ER vesicles with Golgi (Zhang et al., 1997), and *LFNG* is glucosaminyltransferase found in the Golgi membrane (Haines and Irvine, 2003). The functional coupling between *BET1*, *LFNG*, and AD-related genes is supported by data from all three species, and by the fact that gamma-secretase is also known to associate with Golgi. The most prominent evidence data types linking *BET1* and *LFNG* were SCL and PPI, while no interaction class was clearly favoured.

FunCoup augmented our knowledge of the Parkinson's disease (PD). We started from 22 genes reported by Cooper et al. (2006) to be modifiers of alpha-synuclein toxicity in a yeast model for PD. We queried FunCoup for human genes that were coupled to both the PD pathway (KEGG05020) and to orthologs of the yeast modifiers, and analysed the corresponding human and yeast subnetworks (Fig. 5B). This revealed 12 human genes that have not previously been associated with PD to our knowledge. Their functions include vesicle trafficking, ubiquitination, and toxic substance removal, all compatible with a role in the PD pathway. An example is the metalloprotease *YME1L1*, whose best coupling (*pfc*=0.97) is to *RAB1A*, ortholog to one of the strongest yeast alpha-synuclein modifiers, *YPT1*. Most of the evidence for this relationship comes from mRNA co-expression in mouse. Interestingly, *YME1L1* has been reported to interact with presenilin (Pellegrini et al., 2001), supporting a commonality between the processes that lead to PD and AD (Suh and Checler, 2002).

Another example from Figure 5B is the heat shock protein *HSPA8*. It had many strong couplings to the subnetwork (nine with *pfc*>0.8). One of the strongest (*pfc*=0.99) is to *PARK7*, mostly inferred from human mRNA co-expression. Further support came from the fact that its links to *YME1L1* and ACTB were conserved in yeast. We noted that *RAB7*, *RAB11A*, and *RAB11B* were strongly coupled to the PD pathway, even though the Cooper screen only found evidence in the yeast model for *RAB1A* and *RAB1B*. Yet, the yeast model may not capture all connections. The 5 *RAB* genes were strongly interconnected in FunCoup; remarkably, they all shared transcription factors targeting their fly and plant orthologs.

Discussion

We have presented FunCoup, a general method for integrating heterogeneous data in order to reconstruct networks of functional coupling between genes. FunCoup introduces prediction of multiple functional classes in parallel, which boosts accuracy and helps annotating of the nature of the predicted interactions. Another crucial feature of FunCoup is integration of individually weak pieces of evidence and converting them into substantially high confidence (Fig. 2).

The novel methods for adaptive data discretisation and likelihood score assignment were indispensable for differential FC class prediction from the same data, and enabled high performance of the FC class-specific predictors (Supplemental Fig. 9). The ability to make different quantitative conclusions from the same data was accumulated piecewise from small differences between discrete bins and their likelihood values (Fig. 1 "Bayesian framework"). Often, different training sets suggested different signs of log likelihood ratio, e.g. a negative correlation of mRNA expression might have delivered positive evidence of SL but negative evidence of ML. Negative evidence can indeed be informative, e.g. protein localization in different compartments (strongly negative), or uncorrelated expression (weakly negative). Systematic accounting for both positive and negative evidence is thus of great value for FunCoup.

Orthologous data, i.e. FC between orthologs of a given gene pair, were treated as any other data source. This way, the FC score of orthologous data depended on its performance on the training sets in the target species. Despite reports of low conservation of protein-protein interactions between species (Mika and Rost, 2006), we found that the overlap nowadays is often substantial. For instance, between the yeast IntAct set and the human high-confidence training set, 431 of the 723 shared orthologous pairs have been observed to interact in both species (including 363 with a higher PPI score, i.e. >0.5).

Input data sets of eight major types were used to generate comprehensive networks in eight eukaryotic species. Throughout the framework, we have followed the principle of automatic parameter tuning to optimise the predictive power of each data source. It is important for FunCoup as an ongoing project that evidence is transferred between organisms to predict FC with minimal human intervention yet ensuring high coverage and quality. In principle, any kind of data can be accepted, continuous or discrete, with either original or FunCoup-calculated pairwise metrics. New data sources, training sets, and species can be easily added, as their relevance to FC is estimated automatically.
How much can more data improve FunCoup's performance? We measured the accuracy as a function of adding model organisms and data sources to the system in a random order for five species and four evidence types. After addition of about 30 evidence sources, not much improvement was seen (Supplemental Fig. 5). In the yeast – an organism without close relatives in FunCoup but with much own data –other species' datasets had a lower impact. Thus, the current data collection provides an amount of information close to practical maximum. Of course, new high-throughput platforms that cover significantly more genes or entirely novel data types may breach the "ceiling" and enable even higher accuracy.

Potentially, evidence from different sources, given the same likelihood score, could differ in "quality". We attempted to discover any potential bias, and found that the "evidence quality" was uniform across data types and species, and independent of evidence magnitude (Supplemental Table 5). This analysis showed that data of different types, or from different species, could be added up in any proportion, yet the only factor that mattered was the value of log likelihood ratio – in full agreement with the NBN assumptions. Whether the main evidence came from a single species, or spread among two or more, did not matter as well. Hence, for currently available genomics/proteomics data, the Bayesian framework based on summing up likelihood ratio scores is

robust and practically unbiased. In other words, the FunCoup data integration was overall correct and efficient on the heterogeneous, sparse, and noisy data landscape. We show that the networks generated by FunCoup are scale-free (Barabasi and Albert, 1999) even at the lowest confidence level, which agrees with the properties of so far known biological networks (Supplemental Fig. 11).

How different is FunCoup from STRING (von Mering et al., 2007)? The databases are built with different methodologies and only partly the same input data. An importantly difference is the treatment of orthologous data. To predict FC in species *X* with evidence E_M from a model species *M*, STRING evaluates likelihood $P(FC|E_M)$ against training data in the same species *M*. In other words, E_M predicts FC in any other species with the same score, although corrected for sequence dissimilarity. Thus, E_M is transferred irrespective of its relevance to FC in *X*. On the contrary, FunCoup already at the training stage finds orthologs of *M* in *X* and evaluates $P(FC_X|E_M)$ against training sets in *X*.
In terms of input data, STRING relies heavily on annotated resources such as co-membership in KEGG pathways. FunCoup does not use curated data from e.g. GO or KEGG as evidence of FC, as we do not think that such knowledge should be re-evaluated by the predictor, and chose to only use it for training datasets and as add-on links shown on the web site. Conversely, 6 out of 8 major FunCoup data types (PEX, MIR, TFB, DOM, SCL, and the ortholog-based eukaryotic PHP) are not used in STRING. Thus, both input data and their evaluation into scores differ between STRING and FunCoup. The Spearman rank correlation between scores of links found both in STRING and FunCoup was 0.19. The highest-ranking links in both resources often do not agree (10-20% overlap). However, the most confident FunCoup predictions are usually (e.g. close to 70% in the human interactome) found, at some confidence, in STRING.

FunCoup represents an efficient large-scale multi-species reconstruction of global gene networks from genomics and proteomics data. As we illustrated with examples, the future *modus operandi* of gene function discovery is to first search a gene network resource such as FunCoup with genes of interest. This will expand the set of new genes predicted to be functionally coupled, which becomes a manageable subset of genes to investigate experimentally for their role in the studied process. By making the networks available via a powerful and user-friendly website we enable biologists to accelerate discovery of gene function.

Methods

The choice of statistical tool for integrating heterogeneous data depends on many factors. From a wide range of possible approaches, such as discriminant or regression analyses, support vector machines, or neural networks (Qi et al., 2006; Huttenhower and Troyanskaya, 2006), we chose the naïve Bayesian network (NBN) because it:

1) tolerates missing values well.
2) has been successfully applied to genome-wide data integration (e.g. Troyanskaya et al., 2003; Lee et al., 2004; Rhodes et al., 2005), and has been formally justified as optimal under certain assumptions (Zhang, 2004) which hold in our case (Supplemental Methods).
3) gives straightforward and interpretable scores, which makes it attractive to both developer and end user.

Potential drawbacks of NBNs include addition of redundant evidences and assignment of scores that are statistically insignificant due to few data points. The redundancy problem, i.e. violation of the requirement of independence between input datasets, is however not an issue for classification accuracy using diverse and sparse data (Zhang, 2004). To assure confident assignment of scores, we applied a statistical test (Supplemental Methods).

Training and input datasets

NBNs require training datasets with positive and negative examples, and the quality of these very much determine the resulting performance. We compiled the positive, "gold standard", sets of the four FC classes of interest (FC-PI, FC-ML, FC-SL, and FC-CM) from IntAct (Kerrien et al., 2007), HPRD (Mishra et al., 2006), BIND (Bader et al., 2005), KEGG (Kanehisa et al., 2002), and UniProt (Boutet et al., 2007). For each class, a set of filtering criteria (Supplemental Table 4) was applied to the raw datasets to extract highly confident functional couplings.

In our case, the negative training set is very difficult to obtain, as one cannot experimentally disprove an interaction. Several authors have attempted to produce negative training sets by e.g. selecting proteins from different sub-cellular localizations (Jansen et al., 2003; Rhodes et al., 2005), membership in different pathways (Li et al., 2004), or mismatching expression. However, none of these criteria guarantee the absence of interaction, and they may introduce a bias in the probabilistic space. We therefore used Bayes' rule in such a way that it employs the background evidence probability rather than using a negative set. The difference is actually small because only about 0.001 of the total gene pairs are expected to be functionally coupled (assuming $2*10^5$ interactions of $2*10^8$ possible in the human genome). A big advantage with our approach is that the background reference sample can always be made sufficiently large, which makes the training less vulnerable to errors from small sample size.

The input data may come as pairs of proteins/genes with a binary or continuous score (PPI, TFB, MIR), individual protein/gene profiles (MEX, PEX, SCL), or annotation features for protein/domain profiles (PHP, DOM). For all types, pair-wise metrics were calculated as described in Supplementary Methods. This is done first with data from the same organism. Then the other organisms are searched for orthologs, and if these are found for both genes, orthologous data from other species are treated at separate input data types. All homologous gene pairs were removed from the training datasets.

Naïve Bayesian network

The naïve Bayesian network (NBN) is trained as follows. First, the discretisation algorithm finds the bins in each metric's range that produce the highest contrast in respect of FC (Supplementary Methods). In our framework, we analysed the set ε of evidence features $E_i \in \varepsilon$ to estimate the integrated support for FC given all non-empty evidences. Starting with an individual evidence E_i that falls in the bin j, the probability that a particular gene pair is functionally coupled is defined by Bayes' rule:

$$P(FC \mid E_{ij}) = \frac{P(FC)P(E_{ij} \mid FC)}{P(E_{ij})} \qquad (1)$$

The probabilities corresponding to 4 different FC classes are here collapsed into one *FC* for brevity. It is not possible to determine $P(FC)$ exactly, but since this is constant we may leave it out. Thus, without losing predictive efficiency, we integrate over all evidences by summing the logarithms of the remaining ratio in order to obtain a simplified classifier called *final Bayesian score, FBS*:

$$FBS(\varepsilon) = \sum_{i=1}^{|\varepsilon|} \log \frac{P(E_{ij} \mid FC\ )}{P(E_{ij})} \qquad (2)$$

where $P(E_{ij} \mid FC\ )$ is estimated from occurrence of E_{ij} in the positive training set. The background probabilities $P(E)$ were estimated from the general population of gene pairs. Only couplings with FBS>3 were kept.

When summing log likelihood ratios to an FBS, there is a potential danger that the particular combination of evidence types could lead to overestimation of the FBS due to redundancy. Using general regression models we examined this effect by measuring how prediction accuracy depended on, in addition to FBS, evidence combinations of factors. The results invariably showed that FBS alone was the only significant predictor (Supplementary Table 5). In other words, the accuracy depended only on the absolute value of FBS and was independent of the evidence configuration, such as the number of distinct supporting evidences. One explanation for this robustness could be that we are treating negative evidence equal to positive evidence.

FBS is convenient to store and analyse evidence components. Some components may be negative despite an overall positive FBS. However, the FBS score does not have strict bounds, and is not intuitively interpretable. We therefore used an approximation of an alternative form of Bayes' theorem (MacKay, 2003) that gives an intuitive and user-friendly FunCoup confidence score between 0 and 1:

$$pfc(\varepsilon) = \frac{P(FC)\prod_{i=1}^{|\varepsilon|} P(E_{ij} \mid FC)}{P(FC)\prod_{i=1}^{|\varepsilon|} P(E_{ij} \mid FC) + \prod_{i=1}^{|\varepsilon|} P(E_{ij})} \qquad (3)$$

pfc is a probability estimate that the pair is functionally coupled, similar to P used in Green and Karp (2004). $P(FC)$, the prior probability that "two random proteins are functionally coupled", is unknown. However, some expert estimates have been given by, e.g. Grigoriev (2003) and Rhodes et al. (2005), and we conservatively set $P(FC)$ to 10^{-3}. Note that *pfc* is in the interval 0…1 and monotonic with respect to $P(FC)$. Another approximation we made was $\prod_{i=1}^{|\varepsilon|} P(E_{ij})$ instead of $P(\neg FC)\prod_{i=1}^{|\varepsilon|} P(E_{ij} \mid \neg FC)$. This numerically close substitution made the estimate even more conservative. We use *pfc* as a confidence value in the FunCoup database.

Discretisation

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 (Supplemental 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 (Supplemental Methods).

Significance testing

Each evidence category (bin) was subjected to a χ^2 test based on the observed positive and background frequencies, and was discarded if p_0>0.001 (Supplemental Methods).

Phylogenetic profiles

For most datasets a continuous score, or metric, was derived that reflects the strength of the FC (Supplemental Methods). However, for phylogenetic profiles, we chose a different strategy. 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 'bin' 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 (Supplemental Fig. 6).

Estimating performance

To measure the predictor performance, we used the common *receiver operating characteristic* (ROC) curves. They map sensitivity (correctly classified fraction of the positive test set, $TP/TP+FN$) to specificity (effectively represented as "predicted non-FC" fraction of the background reference set, $TN+FP/TN+FP+TP+FN$) at varying cutoffs. To estimate hundreds of ROC curves in the course of testing, the *area under curve* (AUC) was used to measure the overall performance, as its convexity reflects the quality of the predictor. The tests were performed in regions of practical importance, i.e. when the predicted interactome reasonably compact. To ensure this we used a cutoff such that the fraction of predictions was fixed to e.g. 1% or 4% of the total $N(N\text{-}1)/2$ pairs between N proteins (Supplemental Fig. 7).

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 Supplemental 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.

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., 2006). In each ortholog cluster, we assigned EC numbers to *Ciona* proteins considering the KEGG assignments to human, fly, and worm cluster members (Supplemental Methods).

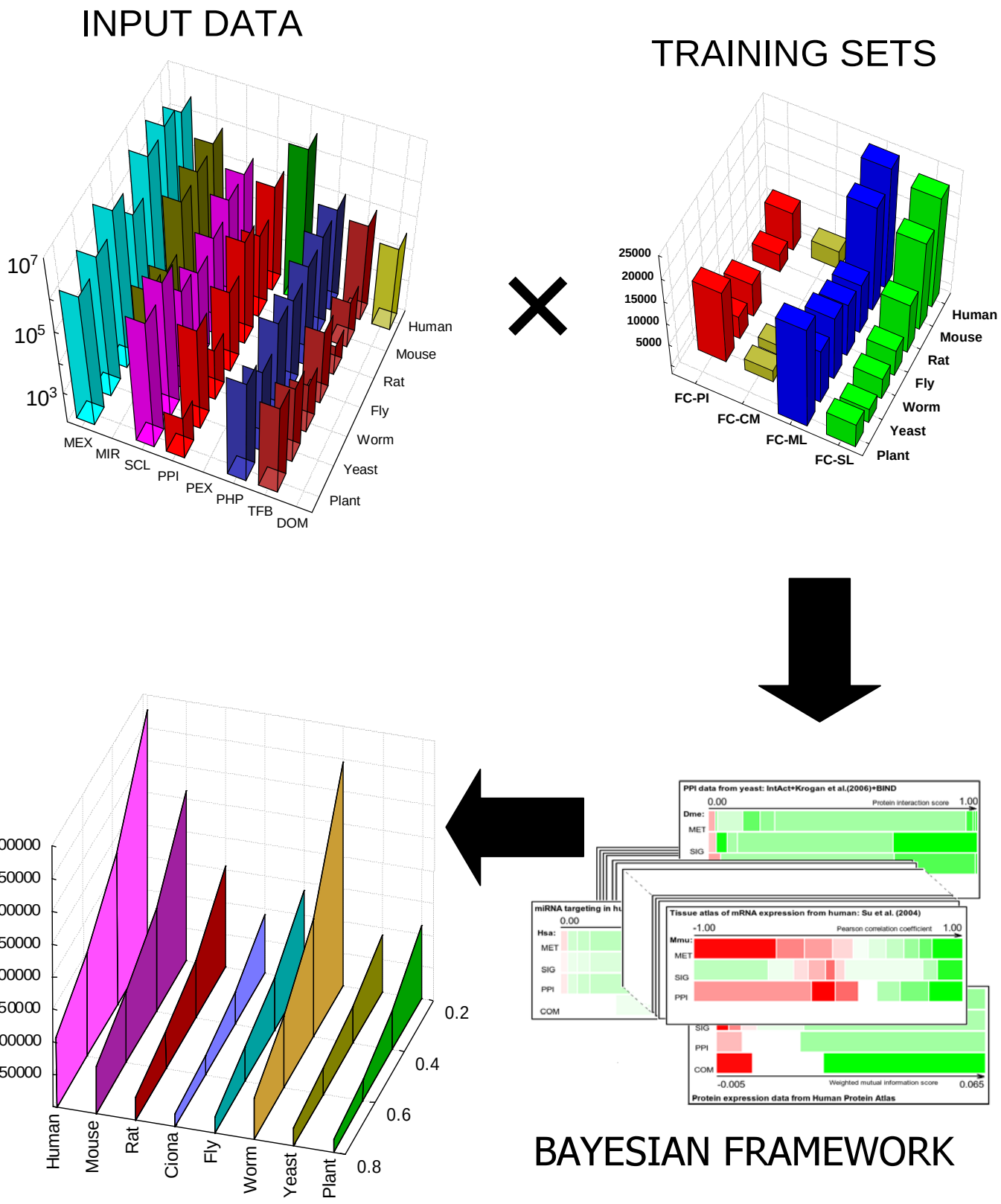


Figure 1. Outline of the FunCoup network reconstruction process. Amounts of input data and sizes of training sets are shown for each species in FunCoup version 1.0.

Input data: MEX: mRNA co-expression; PHP: phylogenetic profile similarity; PPI: protein-protein interactions; SCL: sub-cellular co-localization; PDI: protein-DNA interactions; PEX: protein co-expression: MIR: miRNA targeting of transcripts; DOM: domain associations.

Training sets: ML: links between proteins from the same metabolic pathways; SL: links between proteins from the same signaling pathways; PI: experimentally observed protein-protein interactions; CM: pairs of proteins-members of the same complex.

The Bayesian framework processes the input data using the training sets. The input datapoints are converted into raw interaction scores, which are grouped into discrete regions. Each such bin is assigned an FC score using the training sets. The ‘cards’ illustrate the results of this process, showing the raw interaction score along the horizontal axis. For each training set, or functional class, the resulting bins are shown as colored rectangles: green: positive evidence of FC, white: either close to neutral or insignificant, red: negative evidence of FC.

Finally, the FC scores are calculated for all possible gene pairs in each species. For brevity, the predicted links of different functional classes have been combined into one network per species.

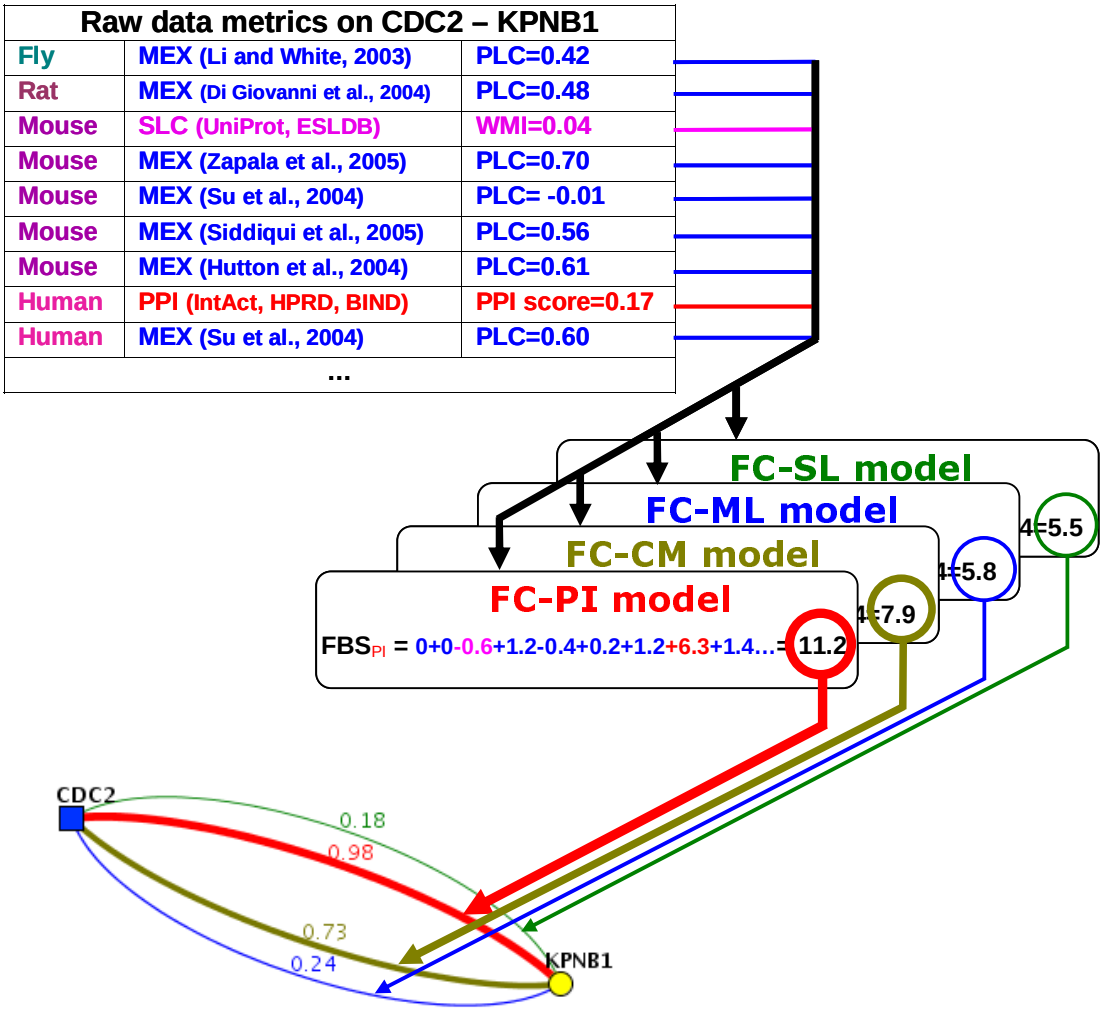
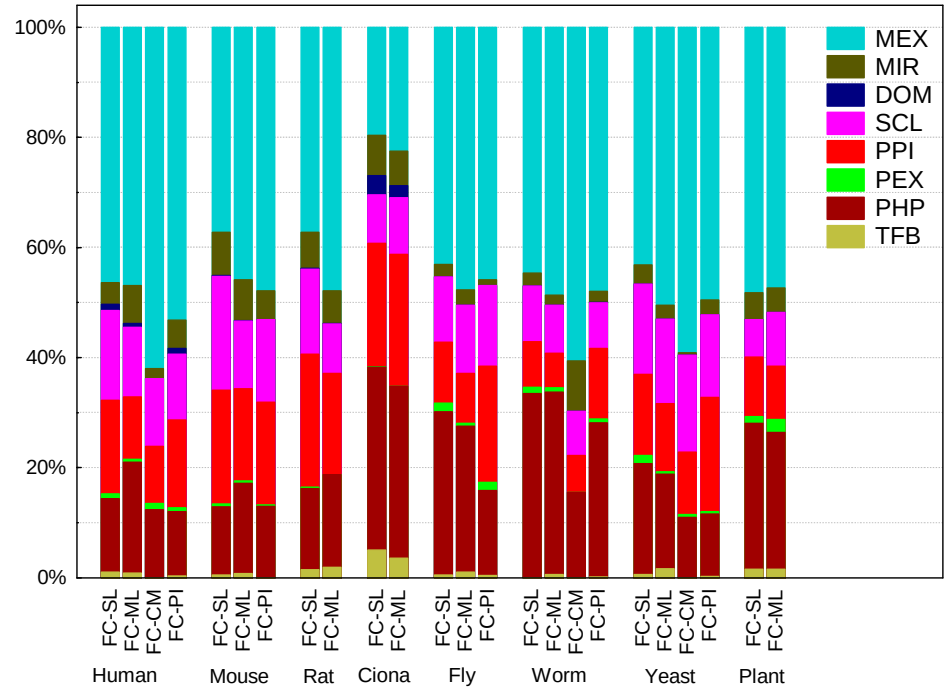


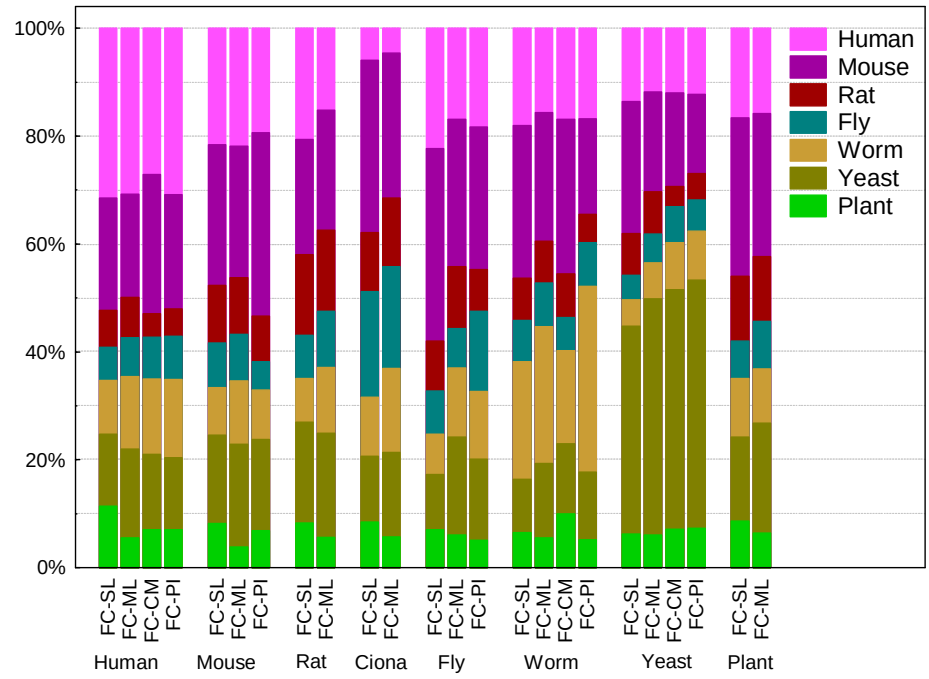
Figure 2. Example of predicting functional coupling by FunCoup's Naïve Bayesian Network. To evaluate the cumulative likelihood of the link *CDC2 – KPNB1*, several (only 9 shown) evidences from a number of species were available. The metrics computed from each evidence were scored by the 4 models of functional coupling and summed up to a total likelihood value, the *Final Bayesian Score* (FBS). In this example, a physical protein interaction (FC-PI) is more likely than membership in the same protein complex (FC-CM), a metabolic (FC-ML), or a signaling (FC-SL) link. The web interface to the database of predictions (http://FunCoup.sbc.su.se) displays likelihoods of the FC classes as 4 coloured lines with FBS values transformed into the *pfc* confidence scores.

This coupling was supported by many evidences from several species (A). The strongest came from the IntAct database that reported two experiments (Thelemann et al., 2005; Koch et al., 2007) where CDC2 and KPNB1 were mentioned as preys interacting with the same bait (in parallel with 219 and 68 other proteins, respectively). On its own this is rather weak, and would only yield *pfc*=0.35 to prove a physical interaction (FBS=6.3). However, combined with other evidence such as co-expression in human and mouse, the score was substantially strengthened (FBS=11.2; *pfc*=0.98). Note that the website can also display the coupling in terms of the individual evidence or species contributions.

MEX: mRNA expression; SLC: sub-cellular localization; PPI: known protein-protein interaction; PLC: Pearson's linear correlation coefficient; WMI: weighted mutual information score; PPI score: see SupplementalMethods.

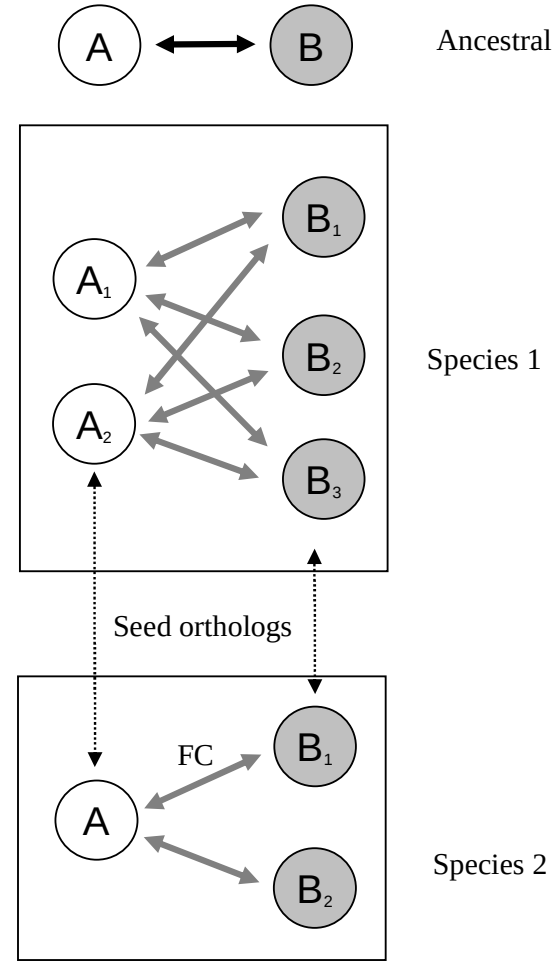


A

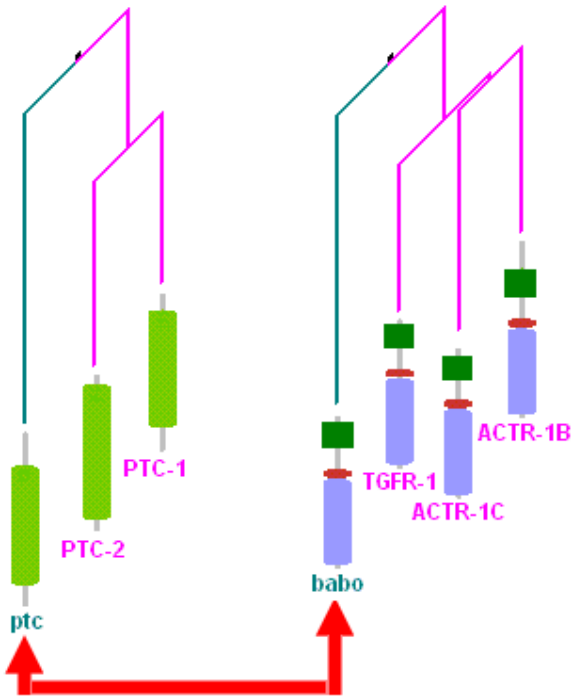


B

Figure 3. Relative evidence contribution to FunCoup networks viewed by input data type (A) or species (B). For each evidence category in each species and FC class, we calculated the sum of its contributions to the final Bayesian score (FBS) over the whole predicted network, i.e. the set of links with FBS > 3. Then, these sums were normalized by dividing with the sum of FBS scores over the same set. Note that no input data from *Ciona* was used. Data type legends as in Figure 1.



A

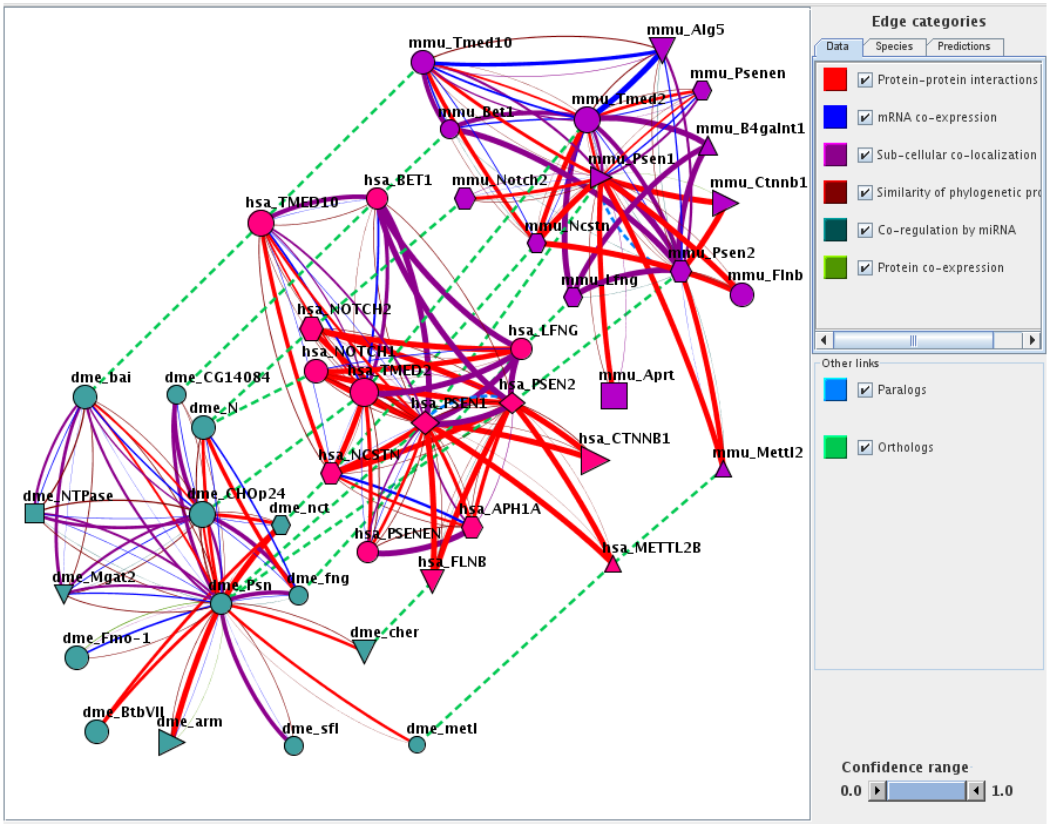


B

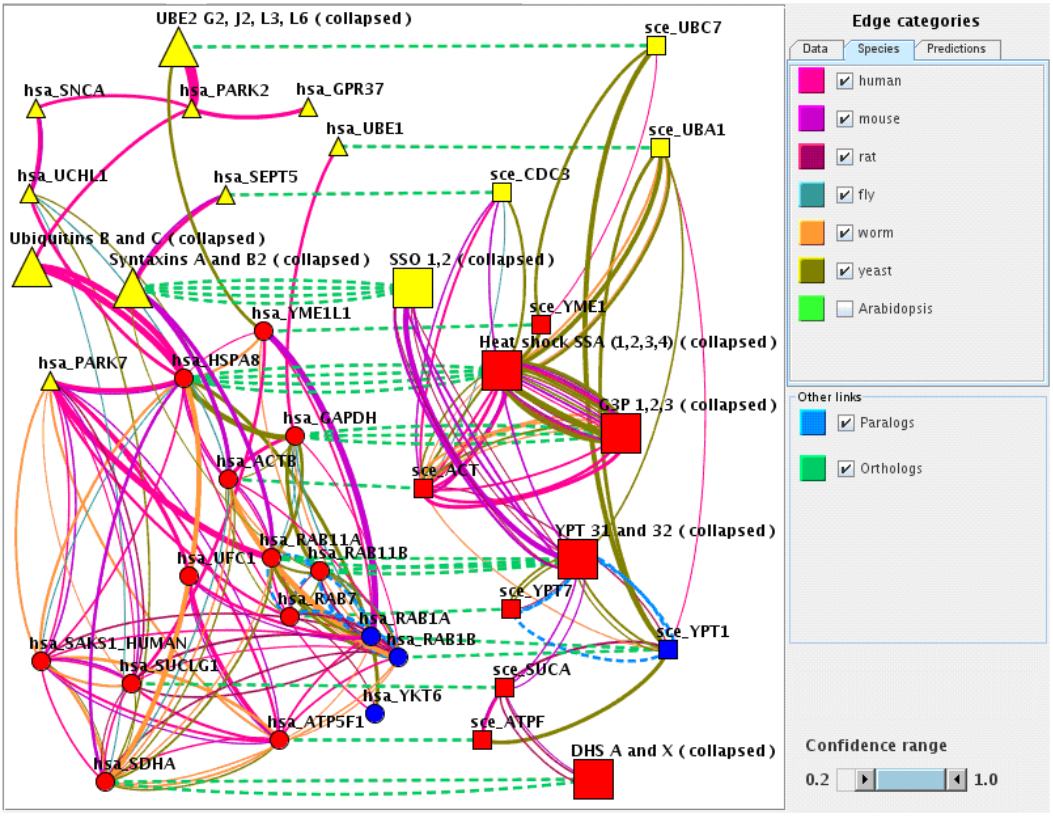
Figure 4. Scenarios for interaction inheritance.

A: Scenarios for interaction inheritance. Interacting genes A and B in an ancestral species are duplicated into 2 and 3 genes in species 1, while B is duplicated into 2 genes in species 2. These duplicated genes are clusters of inparalogs in relation to the other species, meaning that they are co-orthologous to the corresponding A and B genes in that species. When transferring functional coupling (arrow marked FC) between orthologs, one can either consider the interaction to be valid for all inparalogs in a cluster, or to be specific for a particular inparalog pair (e.g. the seed orthologs, i.e. the two most similar ones, dotted arrows). The transfer of interaction information can thus be done either from an arbitrary inparalog in a cluster, which maximises the coverage, or only be transferred between e.g. seed orthologs. Benchmarking showed that transferring the best interaction in a set of inparalogs to all inparalogs in the other species yields the best results (Supplemental Material).

B: Real example of this situation: An interaction in fly *ptc* – *babo* (Shyamala and Bhat, 2002) can either be considered to be valid for all six pairs of mouse inparalogs {*PTC-1*, *PTC-2*} vs. {*TGFR-1*, *ACTR-1B*, *ACTR-1C*} or to be specific for a particular inparalog pair. Seed orthologs in this case are *ptc*/*PTC-1* and *babo*/*TGFR-1*. The gene tree branches are coloured by species lineage; fly in deep cyan/dark and mouse in magenta/bright. The arrow denotes the known protein-protein interaction between *ptc* (UniProt:*P18502*) and *babo* (UniProt:*A1Z7L8*). Protein domains are shown as coloured segments.



A.



B.

Figure 5. Comparative network analysis in FunCoup.
A. Subnetworks in human (middle), mouse (top), and fly (left) were generated by submitting human presenilin 1 and 2 (*PSEN1* and *PSEN2*) 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. At the lower right are two newly predicted interactors of the gamma secretase complex, *BET1* and *LFNG*. On the right, the colour legend for the links is shown in terms of evidence type. Species are indicated by gene symbol prefixes: hsa – human, mmu – mouse, and dme – fly. Supplemental Figure 8 presents alternative views by species source or predicted class. The jSquid XML source to this figure is available as supplementary file psen_lfng_bet.xml.
B. Using FunCoup to identify novel candidate genes in Parkinson's disease. We here employ orthologous networks in human (left) and yeast (right, squares). Novel candidates were extracted by looking for human genes that were coupled to both known PD genes and to orthologs of yeast alpha-synuclein toxicity modifiers. This resulted in 12 novel candidate PD genes (red circles). PD and alpha-synuclein toxicity modifier genes not connected to these 12 genes were omitted from the network for clarity. Edge categories are shown in the figure. Node categories:
Yellow: known human PD genes from the pathway KEGG05020 (triangles) and their yeast orthologs (squares). 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). Larger shapes labeled "collapsed": grouped genes – inparalogs, except for *UBE2* that represents 4 ubiquitin-conjugating enzymes with similar molecular function. See Supplemental Methods for details on how the subnetwork was generated and Supplemental Figure 10 for a large scale prospective of the presented sub-interactome.

References

1. Alexeyenko, A., Tamas, I., Liu, G., Sonnhammer, E.L.L. 2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22:** e9-e15.
2. Bader, G.D., Betel, D., Hogue, C.W. 2003. BIND: the biomolecular interaction network database. *Nucleic Acids Res.* **31(1):** 248-250.
3. Bader, J.S., Chaudhuri, A., Rothberg, J.M., Chant, J. 2004. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol.* **1:** 78-85.
4. Barabasi, A.L. and Albert, R. 1999. Emergence of scaling in random networks *Science* **286(5439):** 509-12.
5. Beyer, A., Bandyopadhyay, S., Ideker, T. 2007. Integrating physical and genetic maps: from genomes to interaction networks. *Nat Rev Genet.* **8(9):** 699-710.
6. Bono, H., Goto, S., Fujibuchi, W., Ogata, H., Kanehisa, M. 1998. Systematic prediction of orthologous units of genes in the complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* **9:** 32-40.
7. Boutet, E., Lieberherr, D,. Tognolli, M., Schneider, M., Bairoch, A. 2007. UniProtKB/Swiss-Prot: the manually annotated section of the UniProt knowledgebase. *Methods Mol Biol.* **406:** 89-112.
8. Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122(6):** 947-956.
9. Butterworth, R., Simovici, D.A., Santos, G.S., Ohno-Machado, L. 2004. A greedy algorithm for supervised discretization. *J. Biomed. Informatics* **37:** 285-292.
10. Cooper, A.A., Gitler, A.D., Cashikar, A., Haynes, C.M., Hill, K.J., Bhullar, B., Liu, K., Xu, K., Strathearn, K.E., Liu, F., Cao, S., et al. 2006. Alpha-synuclein blocks ER-Golgi traffic and *Rab1* rescues neuron loss in Parkinson's models. *Science* **313(5785):** 324-328.
11. Green, M.L., Karp, P.D. 2004. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* **5:** 76.
12. Ding, L., Getz, G., Wheeler, D.A., Mardis, E.R., McLellan, M.D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D.M., Morgan, M.B., et al. 2008. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455(7216):** 1069-1075.
13. Grigoriev, A. 2003. On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res.* **31(14):** 4157-4161.
14. Hahn, A., Rahnenfuhrer, J., Talwar, P., Lengauer, T. 2005. Confirmation of human protein interaction data by human expression data. *BMC Bioinformatics* **6(1):**112.
15. Haines, N., Irvine, K.D. 2003. Glycosylation regulates Notch signalling. *Nat Rev Mol Cell Biol.* **4(10):** 786-797.
16. Hulsen, T., Huynen, M.A., de Vlieg, J., Groenen, P.M. 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.* **7(4):** R31.
17. Huttenhower, C. and Troyanskaya, O.G. 2006. Bayesian data integration: a functional perspective. *Comput Syst Bioinformatics Conf.* 341-351.
18. Imai, K.S., Levine, M., Satoh, N., Satou, Y. 2006. Regulatory blueprint for a chordate embryo. *Science* **312(5777):** 1183-1187.
19. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M.A. 2003. Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302(5644):** 449-53.
20. Kanehisa, M., Goto, S., Kawashima, S., Nakaya, A. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30(1):** 42-6.
21. Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., et al. 2007. IntAct – open source resource for molecular interaction data. *Nucleic Acids Res.* **35:** D561-5.
22. Klammer, M., Roopra, S., Sonnhammer, E.L. 2008. jSquid: a Java applet for graphical on-line network exploration. *Bioinformatics* **24(12):** 1467-1468.
23. Koch, H.B., Zhang, R., Verdoodt, B., Bailey, A., Zhang, C.D., Yates, J.R. 3rd, Menssen, A., Hermeking, H. 2007. Large-scale identification of c-MYC-associated proteins using a combined TAP/MudPIT approach. *Cell Cycle.* **6(2):** 205-17.
24. Lee, I., Date, S.V., Adai, A.T., Marcotte, E.M. 2004. A probabilistic functional network of yeast genes. *Science* **306(5701):** 1555-1558.
25. Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., et al. 2004. A map of the interactome network of the metazoan C. elegans. *Science* **303(5657):** 540-543.
26. Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., Garrels, J., Vincent, S., Vidal, M. 2001. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res.* **11(12):** 2120-2126.

27. MacKay, D.J.C. 2003. Information theory, inference, and learning algorithms. *Cambridge University Press.*
28. Mika, S., Rost, B. 2006. Protein-protein interactions more conserved within species than across species. *PLoS Comput Biol.* **2(7):** e79.
29. Mishra, G.R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T.M., et al. 2006. Human protein reference database--2006 update. *Nucleic Acids Res.* 2006 **34:** D411-414.
30. Qi, Y., Bar-Joseph, Z., Klein-Seetharaman, J. 2006. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins* **63(3):** 490-500.
31. Rhodes, D.R., Tomlins, S.A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., Chinnaiyan, A.M. 2005. Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol.***23(8):** 951-959.
32. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13(11):** 2498-2504.
33. Snitkin, E.S., Gustafson, A.M., Mellor, J., Wu, J., DeLisi, C. 2006. Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics* **7:** 420.
34. Sonnhammer E.L.L. 2005. Genome informatics: taming the avalanche of genomic data *Genome Biology* **6:** 301.
35. Srinivasan, B.S., Novak, A.F., Flannick, J.A., Batzoglou, S., McAdams, H.H. 2006. Integrated protein interaction networks for 11 microbes. *RECOMB 2006*, 1-14.
36. StatSoft, Inc. 2005. STATISTICA (data analysis software system), version 7.1. *www.statsoft.com*.
37. Suh, Y.H., Checler, F. 2002. Amyloid precursor protein, presenilins, and alpha-synuclein: molecular pathogenesis and pharmacological applications in Alzheimer's disease. *Pharmacol Rev.* **54(3):** 469-525.
38. Thelemann, A., Petti, F., Griffin, G., Iwata, K., Hunt, T., Settinari, T., Fenyo, D., Gibson, N., Haley, J.D. 2005. Phosphotyrosine signaling networks in epidermal growth factor receptor overexpressing squamous carcinoma cells. *Mol Cell Proteomics*. **4:** 356-76.
39. The Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455(7216):** 1061-1068.
40. The Gene Ontology Consortium. 2000. Gene Ontology: tool for the unification of biology. *Nature Genet.* **25:** 25-29.
41. Troyanskaya, O.L., Dolinski, K., Owen, A.B., Altman, R.B., Botstein, D.A. 2003. Bayesian network for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc Natl Acad Sci USA* **100(14):** 8348-8353.
42. von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Krüger, B., Snel, B., Bork, P. 2007. STRING 7 – recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* **35:** D358-362.
43. von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., Bork, P. 2005. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*. **33:** D433-437.
44. Zhang, H. 2004. The optimality of naive Bayes. *Proceedings of the 17th International FLAIRS conference (FLAIRS2004). AAAI Press.*
45. Zhang, T., Wong, S.H., Tang, B.L., Xu, Y., Peter, F., Subramaniam, V.N., Hong, W. 1997. The mammalian protein (rbet1) homologous to yeast Bet1p is primarily associated with the pre-Golgi intermediate compartment and is involved in vesicular transport from the endoplasmic reticulum to the Golgi apparatus. *J Cell Biol*. **139(5):** 1157-1168.
46. Zhong, W., Sternberg, P.W. 2006. Genome-wide prediction of *C. elegans* genetic interactions. *Science* **311(5766):** 1481-1484.