Journal of
Computer Science & Systems Biology

Fechete et al., J Comput Sci Syst Biol 2013, 6:2
DOI: 10.4172/jcsb.1000102

Research Article Open Access

Using Information Content for Expanding Human Protein Coding Gene Interaction Networks

R Fechete[1], A Heinzel[1], J Söllner[1], P Perco[1], A Lukas[1] and B Mayer[1,2*]

[1]*Emergentec Biodevelopment GmbH, Vienna, Austria*
[2]*Institute for Theoretical Chemistry, University of Vienna, Austria*

Abstract

Molecular interaction networks have emerged as central analysis concept for Omics profile interpretation. This fact is driven by the need for improving hypothesis generation beyond the mere interpretation of molecular feature lists derived from statistical analysis of high throughput experiments. A number of human gene and protein interaction networks are available for such task, but these differ with respect to biological nature of interactions represented, and vary with respect to coverage of molecular feature space on the gene, transcript, protein and metabolite level. Naturally, both elements impose major impact on hypothesis generation. We here present a methodology for deriving expanded interaction networks via consolidating available interaction information and further adding computationally inferred interactions.

Integrating interaction data as provided in the public domain repositories IntAct, BioGrid and Reactome resulted in a core interaction network representing 11,162 human protein coding genes (out of a total of 19,980 protein coding genes) and 145,391 interactions. Utilizing annotation from ontologies on involvement in specific molecular pathways and function, combined with structural (domain) information as gene/protein node parameterization allowed computation of probabilities for additional interactions resting on the information content of individual sources. Utilizing topological information as degree centrality, global clustering coefficient and characteristic path length allowed defining a cutoff for interaction probabilities, resulting in an expanded interaction network holding 13,730 protein coding genes and 830,470 interactions. Evaluating such hybrid network against established interaction networks as KEGG showed significant recovery of evident interactions, indicating the validity of the expansion methodology.

Integrating available interaction data, further enlarged by inferred interactions, provided an expanded human interactome regarding both, number of represented molecular features as well as number of interactions, thereby promising improved Omics profile interpretation.

Keywords: Graph; Protein interaction; Graph measure; Receiver operator characteristics; Omics profile interpretation

Introduction

Molecular interaction networks representing the human protein coding gene universe have become widely used for analyzing Omics profiles. The intention is to traverse the descriptive set of features identified as statistically relevant into the context of pathways and processes being effectively amenable for results interpretation and hypothesis generation. This need is even more pronounced for cross-Omics data interpretation, where the challenge is the combined analysis of heterogeneous feature types essentially spanning from the genome to the metabolome level [1]. Such integrated analysis strategies rest on expanding knowledge regarding molecular feature catalogues, on their biological role and interaction specifics, altogether resembling a core of systems biology approaches [2], in the clinical context contributing to systems medicine [3,4].

Repositories for molecular catalogues are maintained by major institutions such as the National Center for Biotechnology Information (NCBI) or ENSEMBL operated by the European Bioinformatics Institute together with the Wellcome Trust Sanger Institute. The same is true for specific molecular interaction networks with the Kyoto Encyclopedia of Genes and Genomes [5] (KEGG) or PANTHER [6] as prominent representatives. Further databases focus explicitly on protein-protein interaction data, as the Online Predicted Human Interaction Database [7] (OPHID), IntAct [8], or BioGRID [9]. Each such repository exhibits specific characteristics regarding type of interaction represented, coverage of molecular catalogues, as well as evidence and relevance of interactions in the biological context. KEGG for example offers various types of interactions ranging from protein complex formation to enzyme-substrate interactions at high level of evidence, but falls short on completeness regarding the human molecular catalog, as 6,198 (version as of January 2012) human protein coding genes (compared to the total set of 19,980 reported in ENSEMBL [10]) are represented. OPHID on the other hand provides a considerable set of protein interactions covering 14,612 UniProt/SwissProt identifiers which can be translated to a roughly similar number of protein coding genes, but lacks evidence regarding biological relevance of listed interactions.

Protein interactions represented in such repositories span heterogeneous types as physical interactions or procedural dependencies [11], resting on different experimental methods such as affinity chromatography, yeast-2-hybrid screens, or being predicted based on cross-species analogies. Consequently, the data hold false-positives [12-14], as well as an undetermined false negative rate [15-

***Corresponding author:** Bernd Mayer, Emergentec Biodevelopment GmbH, Gersthofer Strasse 29-31, 1180 Vienna, Austria, Tel: +43 1 4034966; Fax: +43 1 4034966-19; E-mail: bernd.mayer@emergentec.com

Received March 04, 2013; **Accepted** April 04, 2013; **Published** April 08, 2013

Citation: Fechete R, Heinzel A, Söllner J, Perco P, Lukas A, et al. (2013) Using Information Content for Expanding Human Protein Coding Gene Interaction Networks. J Comput Sci Syst Biol 6: 073-082. doi:10.4172/jcsb.1000102

17]. Further uncertainty arises if e.g. identified binding affinities exhibit relevance in the biological context, altogether only becoming amenable via functional studies. On top the total number of protein-protein interactions in the human proteome is not known, with estimates ranging from 130,000 to 650,000 [18,19].

Obviously, the type of specific interaction network used for Omics profile interpretation influences hypothesis generation, driven by mere quantity of represented features, and biological nature of encoded interactions. For respecting these facts hybrid interaction networks have been developed, aiming at integrating diverse sources for obtaining networks showing improved coverage on the level of molecular feature catalogs, and at the same time providing a more complete representation of biologically relevant interactions at the various levels. Alexeyenko et al. [1], for example, delineated a network by inferring edge probabilities using diverse high-throughput data and achieved high gene coverage through orthology-based integration of different model organisms. Tyagi et al. [13] described a framework for delineating a human protein interactome including experimental details on complex structures and their binding interfaces together with evolutionary conservation. Kuchaiev et al. [16] presented a technique using geometric graphs to assess the confidence levels of interactions in protein-protein interaction networks obtained from experimental studies in order to predict new interactions.

We in this work attempt to expand the set of human protein-protein interactions as provided by IntAct, BioGrid and Reactome [20] with inferred interactions being computed based on pathway membership (Reactome, PANTHER), ontology membership (Gene Ontology [21]) and protein domain data (InterPro [22]). Goal of such approach is to improve coverage of the human protein coding gene set represented in a combined network, and at the same time expanding on putative interactions. Having such an expanded network in hand promises an improved representation of features identified as relevant in Omics profiling, additionally providing expanded information on interactions, in combination promising improved hypothesis generation.

Materials and Methods

Gene and protein identifier cross-referencing

For gene and protein cross-referencing the BioMart interface of ENSEMBL [10] was used. Identifiers included ENSEMBL gene (19,980) and protein (86,934) IDs, NCBI gene symbols (18,981) and IDs (18,994), NCBI protein identifiers (31,628), TrEMBL [23] (42,399) and SwissProt IDs and accessions (37,864). The NCBI gene symbols and summaries were imported directly from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/), and the lists of deprecated identifiers were imported from NCBI and UniProt.

IDs referring to the same biological entity (gene or protein level, respectively) were interlinked using abstract hyperstructures. Each such structure resembles a protein sequence as retrieved from ENSEMBL. All gene identifiers and symbols of genes coding for this sequence, and all protein identifiers were linked to such protein sequence. These hyperstructures encoded the nodes of the hybrid interaction network.

Data sources on experimentally identified and literature curated protein interactions

Human protein-protein interaction data were retrieved from the public domain data sources IntAct, BioGrid and Reactome, all in their versions as of October 2011. IntAct provided 35,634, BioGrid provided 41,496, and Reactome contributed with 91,002 interactions, respectively. Reactome offers two types of interactions, namely physical associations for all interactors occurring in the same complex, polymer, or in the same reaction, as well as associations for interactors involved in neighboring reactions or being associated via (positive or negative) regulation. Consolidation of these interaction data sources provided a core interaction network. All interactions were retrieved and stored with their characterization provided as Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI) Molecular Interaction (MI) ontology terms [24]. The PSI-MI ontology is rooted in the term molecular interaction (MI:0000) and the two branches relevant for this work are the interaction detection method (MI:0001) and the interaction type (MI:0190).

Gene and protein parameterization

For biological annotation of genes and proteins three categories of biological data were explored: biological pathway assignment, ontology assignment, and protein domain information. No explicit interaction information was used from such sources for annotation. Pathway membership information was retrieved from Reactome and PANTHER. Reactome provided 1,128 pathways covering 5,292 genes, while PANTHER contributed 140 pathways covering 2,120 genes. Ontology information was retrieved from the Gene Ontology Consortium for the branches biological process and molecular function. The process category contributed 21,433 terms covering 14,110 genes, while the function category contributed 9,087 terms covering 14,693 genes. Protein domain information was retrieved from InterPro, which provided 6,041 specific domains covering 38,275 protein sequences. These annotation data on pathway membership, ontology membership as well as protein domain specifics were subsequently used for inferring interactions.

Information consolidation

For each gene/protein entity the respective hyperstructures were populated with the retrieved annotation information. This information was subsequently used for computing relation scores (edge weights) between entities. A major implication of this information inheritance approach was that for each gene information became redundantly associated to all its splice variants. Considering the large number of splice variants, i.e. 86,934 sequences for 19,980 genes, but mostly not having splice variant specific annotation available, only one representative protein sequence per protein coding gene was further considered. For this, the hyperstructure representing the protein with the longest sequence of each ENSEMBL gene was marked as the canonical sequence. Only such canonical hyperstructures were then used as nodes in the interaction network. Protein domain information for non-canonical sequences (if available) was also linked to the canonical sequence.

Interaction scoring

Based on the consolidated information, interaction scores were computed for all pair-wise combinations of canonical hyperstructures. For each pair, five individual scores were computed, one for each of the available annotation types: GO biological process, GO molecular function, Reactome pathway membership, PANTHER pathway membership, protein domains. Additionally, for each pair interactions as retrieved from the three interaction databases (IntAct, Biogrid, Reactome) were included as applicable. These individual scores together with available explicit interaction information were subsequently combined into an overall score, expressed as edge weight.

All parameters were computed based on the information content (*IC*) of the terms, pathways and domains shared by two hyperstructures. For pathways and domains, the information content was computed as depicted in equation 1.

$$IC(E) = -\ln(P(E)) = -\ln(\frac{n_E}{N}) \quad (1)$$

E is the entity (pathway or domain), the information content is being computed for. *N* is the total number of nodes with this type of annotation. n_E is the number of nodes on this particular pathway or having this particular domain.

For computing the information content of the GO terms, this general calculation was extended to accommodate for the hierarchical structure of the ontology: each node was not only associated to the terms explicitly mentioned by GO, but also to all parent terms in the hierarchy following the "is-a" relationship. Further, the information content of a term is not only a function of the number of genes associated to it and its children [25-27], but also a function of its specificity given by the relative position in the ontology (number of its child terms) [28]. Based on these considerations two separate IC functions needed to be applied: one for the genes and one for the terms (equation 2). We added a value of 1.0 to the number of genes and child terms to enable computation of information content values also for terms which were either empty or which had no children.

$$IC_{genes}(E) = -\ln(\frac{n_{genes,E}+1}{N_{genes}})$$
$$IC_{terms}(E) = -\ln(\frac{n_{terms,E}+1}{N_{terms}}) \quad (2)$$

Subsequently, the overall information content of a term was computed by using the Root Mean Square (RMS) function (equation 3).

$$IC(E) = RMS(IC_{genes}(E), IC_{terms}(E)) \quad (3)$$

For a given node pair (X,Y), each of the individual parameter scores was computed as the sum of the information content values of the entities *E* (domains, pathways and terms) common to both nodes (equation 4).

$$f(X,Y) = \sum_{\substack{E \in X \\ E \in Y}} IC(E) \quad (4)$$

By construction, all parameter values were equal to or greater than zero. However, the individual parameter distributions were found to be strongly right-skewed. A high number of zero values resulted due to a limited set of parameter overlaps. In order to ensure a similar impact on the final edge weight obtained by utilizing distance functions on the individual parameter level for node pairs, we performed score adjustment. Due to the semantic contrast between the zero and the non-zero values, i.e. definitively inexistent versus given level of relation, we ignored the zero values during the normalization process. Each parameter distribution (as resulting from equation 4) was then scaled with an individual factor *alpha* and the logarithm was computed. The alpha values were chosen such as to ensure a maximum curve overlap after log-transformation and rescaling on the interval (0,1). The 0.5 quantiles of the resulting distributions were situated around 0.35 (Figure 1).

For computation of the final edge weight as a proxy for the aggregate relation between two nodes, we distinguished between the

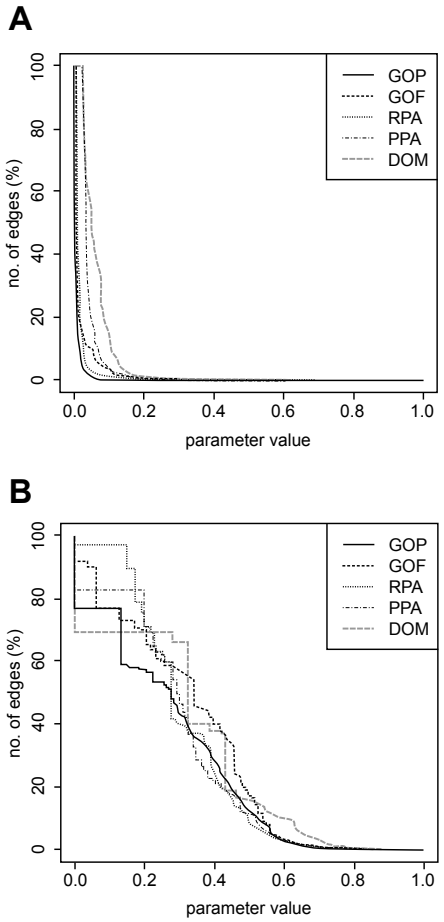

Figure 1: Distributions of parameter values of the five annotation sources included in computing pair-wise edge weights before (A) and after (B) rescaling: Gene Ontology process (GOP), Gene Ontology function (GOF), Reactome pathway assignment (RPA), PANTHER pathway assignment (PPA) and protein domains (DOM). For each parameter value in the interval [0,1] the number of edges holding at least such value are given.

interaction type "procedural" (composed of GO process, Reactome pathway membership and PANTHER pathway membership) and "functional" (composed of protein domain and GO molecular function) parameters. For each node pair, the average relation score of each category, i.e. procedural and functional, was computed separately utilizing the assigned annotation. The effective weight of a relation between two nodes (i.e. the edge weight) was given by the maximum of the procedural and the functional parameter, being in the interval (0,1). In case an experimentally described interactions (INT1) was present for a given pair the edge weight was set to 1.0 irrespective of computed procedural or functional relations (equation 5).

$$f(X,Y) = \max\begin{Bmatrix} avg(f_{Reactome}, f_{PANTHER}, f_{GO\,Process}) \\ avg(f_{GOFunction}, f_{Domains}) \\ 1(X,Y) \in INT1 \end{Bmatrix} \quad (5)$$

Individual parameters missing due to incomplete node annotation were omitted from the computation, with the constraint of having at least one shared parameter for effectively computing a score. Consequently, node pairs not sharing a single parameter could not be further taken into consideration.

Results

Data sets and parameter characteristics

The number of gene/protein nodes taken into consideration for the protein interaction graph construction corresponded to the total number of protein coding genes represented in ENSEMBL, being 19,980. Of these, 14,212 had GO process and 14,773 had GO function annotation, 5,340 were present in Reactome pathways, 2,138 in PANTHER pathways, and 13,681 in InterPro. Characteristics of data sources covering experimentally derived interactions (number of features and number of interactions) are depicted in table 1. BioGrid had with close to 9,000 genes the best node coverage, while Reactome showed with close to 91,000 interactions the most extensive edge coverage. Next to evidently varying size is the varying overlap of the data sources, clearly indicating different background and scope of given sources. IntAct and BioGrid for example provide roughly the same number of interactions, with only about one third being shared.

Consolidating the available annotation information from GO, PANTHER pathway membership, Reactome pathway membership, and InterPro for each gene provided 17,022 of the 19,980 nodes with information from at least one data source. This set of nodes was further taken into consideration for delineating the inferred interactions, neglecting the 2,958 protein coding genes not holding any annotation. The level of annotation, given as the number of data sources per node, is depicted in figure 2A. 1,199 node pairs showed annotation on the level of all five data sources, 6,682 nodes held annotation from three sources. Computing all-to-all interactions rested on annotation levels as depicted in figure 2B. Based on the 17,022 nodes annotated with at least one source a complete undirected graph holding in total 144.8 million edges (n* (n-1)/2) could theoretically be computed. Of these, 5 million edges had no basis (i.e. node pairs not sharing a single common annotation), and were therefore omitted from further processing.

Of the remaining about 140 million theoretical interactions 145,391 rested on experimentally derived/manually curated interactions as provided by IntAct, BioGrid or Reactome. Such edges are subsequently denoted as INT1 (in contrast to INT0 edges not having such experimental background on interaction).

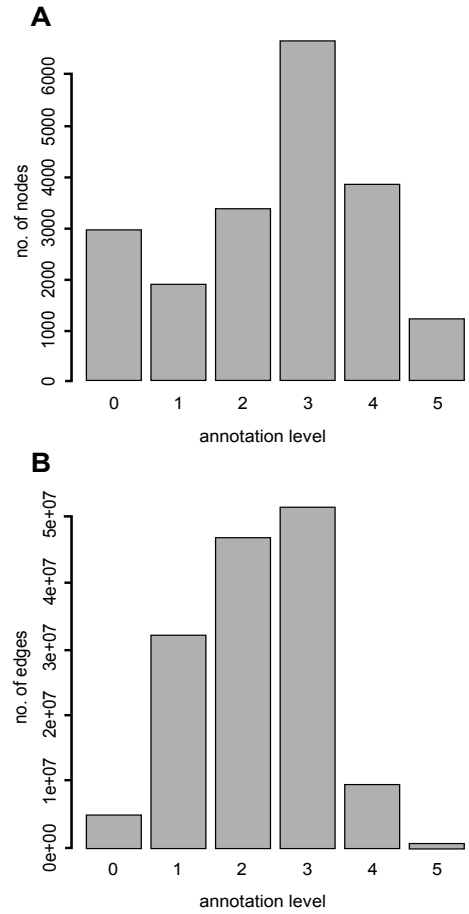

Figure 2: Overview of the number of data sources available per node (A) and shared by node pairs (B), with a maximum annotation level of five according to the annotation sources (GO process, Reactome pathway membership, PANTHER pathway membership, protein domain, GO molecular function). About 3,000 nodes hold no annotation, about 1,000 nodes hold annotation from all five sources. For about 50 million node pairs the number of shared annotation sources is three out of five.

General graph characteristics

The cumulative edge weight distribution in the interval (0,1) of the 140 million edges, together with completeness, i.e. the number of nodes holding at least one edge with a weight above a certain cutoff value, is depicted in figure 3A. All 145,391 INT1 interactions (with an edge weight set to 1.0, i.e. being considered as true positive interactions, see equation 5) provided 11,162 nodes. The maximum number of nodes, on the other hand, is only reached at a cutoff value of 0.0. This is due to the about 23 million edges holding a weight of zero, as becoming evident in the weight dependent number of edges shown in figure 3A. Edge weights of zero result from node pairs holding values in at least one common parameter (i.e. being valid for computing a weight) but showing no overlap in the annotation of that particular parameter (see equation 4).

A: number of nodes			
	IntAct	BioGrid	Reactome
IntAct	8,419	5,348	1,248
BioGrid		8,988	1,925
Reactome			4,458
B: number of edges			
	IntAct	BioGrid	Reactome
IntAct	35,634	13,928	2,258
BioGrid		41,496	3,987
Reactome			91,002

Number of unique gene identifiers (A: number of nodes) and interactions (B: number of edges), as well as pair-wise overlap, for the data sources used for retrieving interaction information.

Table 1: Interaction data source overview.

Plotting the number of nodes and edges in dependence of edge weight provided for a given number of nodes the number of edges being necessary to include the given nodes (i.e. holding at least one edge per node, figure 3B). Covering eg. additional 3,000 nodes on top of the 11,162 nodes included on the basis of INT1 resulted in an increase of the number of additionally required edges to about 1 million. Further increasing the number of nodes for including 15,000 protein coding genes already needs more than 3.0 million edges total. Adding additional edges by lowering the edge weight cutoff tends to

link nodes already being part of the graph rather than adding additional nodes. Apparently, for indeed covering all protein coding genes in the network an implausible number of edges would be needed.

In order to investigate potential bias in deriving edge weights as a consequence of specific level of annotation, the association of edge weight and gene characterization index (GCI, indicating to what extent a protein-encoding gene is functionally described) [29] was calculated. The Pearson correlation coefficient between each node's highest edge weight and its GCI was 0.43, thus a weak positive correlation between the two parameters could be observed. No substantial correlation (Pearson R=0.15) could be determined for comparing each node's highest edge weight to the number of papers associated to the respective gene based on NCBI's gene2pubmed [30] links. A third analysis regarding eventual bias in node or edge annotation focused specifically on the node annotation level (Figure 2A). An increase in the maximum edge weight per node with rising annotation level was identified (Pearson correlation score of 0.53). To estimate the impact of this bias, we performed a complementary analysis involving the edge annotation basis (Figure 2B). The Pearson correlation coefficient between edge weights and edge annotation basis was found to be-0.2.

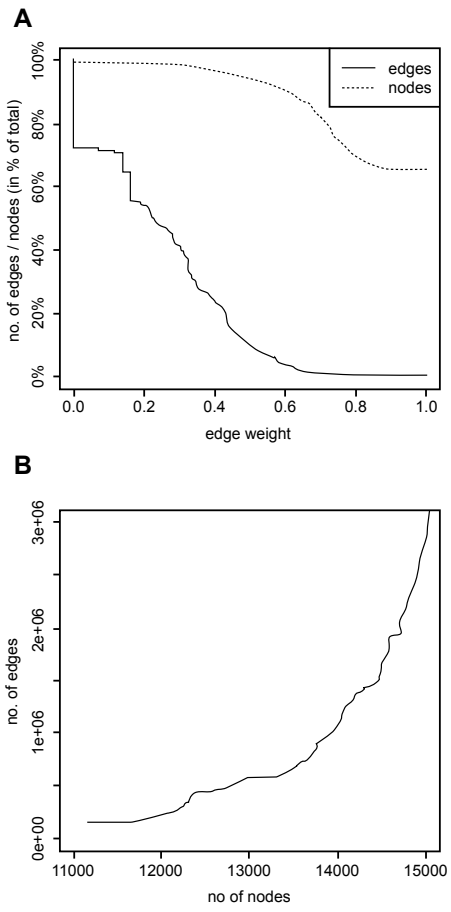

Figure 3: (A) Edge weight distribution and node coverage as a function of edge weight. Node and edge count are presented for each weight cutoff in % with respect to the entire set of nodes and theoretical edges, with a start number of nodes of 11,162 at an edge weight of 1.0 (145,391 edges as extracted from experimentally verified interaction data sources), and a maximum number of nodes being 17,022 with a maximum number of edges when considering all node pairs being about 140 million. (B) Relation of node and edge count at the different edge weight cutoff values, starting at an edge weight of 1.0 (with 11,162 nodes).

Topological graph characteristics

Already at the maximum edge weight of 1.0 (essentially resembling INT1, as no computed IC value reached a value of 1.0) the Index of Aggregation (IoA) was close to 1.0, i.e. paths between virtually all 11,162 nodes represented in IntAct, BioGrid or Reactome were found. Nodes with inferred interactions are added to this given graph when lowering the edge weight cutoff levels. The node degree distribution of the graph at different cutoff levels, expressed as median, upper and lower quartiles and outliers, is provided in figure 4. The median number of neighbors remained below 1% of included nodes for the weight interval [0.7, 1.0]. Connectivity outliers at each cutoff provided an explanation for the high IoA, i.e. a relatively small number of hub nodes is responsible for building the giant component. To further investigate this finding the five nodes with the highest degrees were exemplarily extracted at various cutoffs. At high edge weight cutoff levels of 0.8-1.0, where the graph is mainly composed of INT1 edges, genes such as ubiquitin and the MYC oncogene [31] were identified as strongly connected to other nodes. This finding may be explained in the light of the biological function of e.g. ubiquitin [32], or reflecting the substantial association of e.g. MYC in a wide range of molecular processes [33]. At lower edge weight cutoff levels of 0.5-0.7 a different effect contributed to connectivity outliers. Here mainly groups of genes jointly annotated by highly specific ontology terms, as e.g. members of the sirtuin [34] family or genes such as the P2RX4 [35] (purinergic receptor P2X, ligand-gated ion channel 4) are present. Following our procedure for the computation of edge weights, highly specific ontology terms result in higher edge weights for all genes sharing at least one of such specific terms.

The global clustering coefficient (GCC) in relation to edge weights in the interval [0.5, 1.0] is depicted in figure 5A. The graph provided on the basis of INT1 alone (i.e. at an edge weight of 1.0 holding 11,162 nodes and 145,391 edges) served as the start network, showing a GCC of 0.33. Decreasing the edge weight cutoff, i.e. gradually adding computed relations as edges together with further nodes not holding an INT1 edge to the network, resulted in an increase of the GCC, reaching a plateau at edge weights of 0.6. To contrast this graph behavior in terms of the GCC to the respective curve obtained when adding edges randomly two additional analyses starting from the same INT1 network were performed. In the first approach edges were added to the given INT1 graph in a random manner in the same pace as for adding computed edges when decreasing the edge weight cutoff. In the second approach the same procedure was performed but additionally respecting the node degree distribution seen when adding the computed edges. The results showed a strong GCC divergence for the three networks with increasing number of edges, with the two reference networks generated by adding edges in a random fashion reaching GCC values of about 0.4 (when respecting the node degree distribution) and 0.1 (adding edges entirely random), in contrast to a GCC value of 0.6 reached for the relations network. Clearly, the inferred graph differs on the level of the GCC significantly from networks populated by edges instantiated randomly between nodes.

For the three graphs also the Characteristic Path Length (CPL) was computed (Figure 5B). The initial characteristic path length of the INT1 network was found to be 3.5, decreasing with lower edge weight cutoffs to a length of 2.0 at an edge weight of 0.5. The divergence of the computed network in comparison to adding random effects also

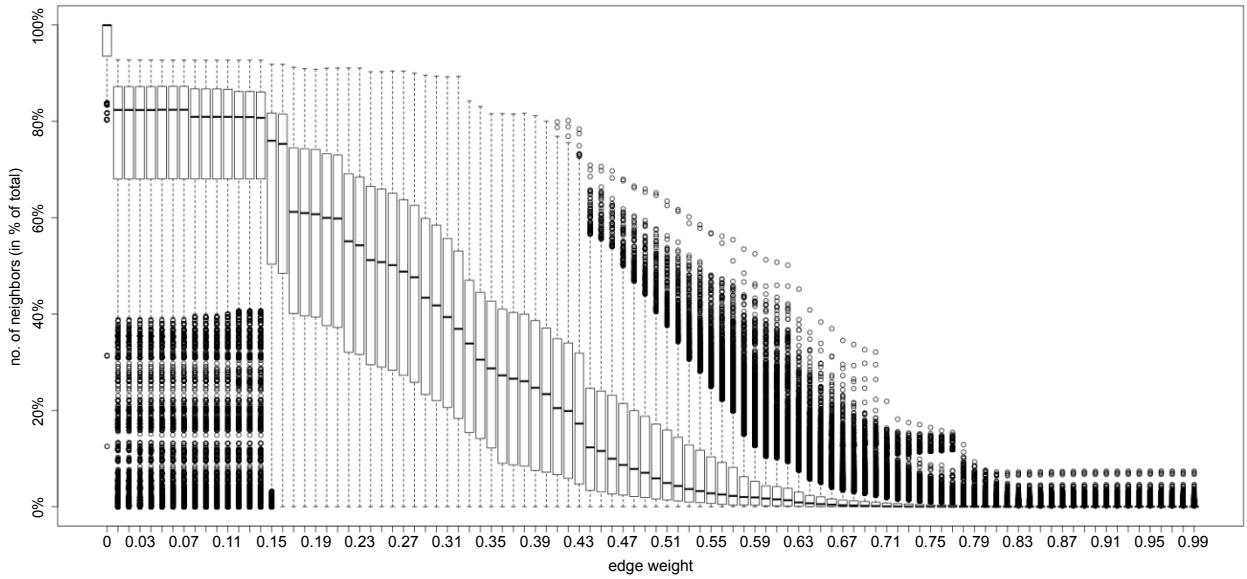

Figure 4: Degree centrality distribution as a function of edge weight for the interval [0,1]. The Box Plots (median, upper and lower quartile, outliers) show the node degree distributions (number of neighbors of a node) for all edges being present in the graph above a specific edge weight cutoff.

became clear on the level of the CPL, with smallest CPL values reached for the graph holding randomly added edges. At an edge weight of about 0.5 the three graphs converged on the level of the CPL.

Inferred edges as surrogate for experimentally determined interactions

To investigate the suitability of the inferred edge weights for predicting interactions resting on experimental evidence as provided in databases, the inferred interactions were investigated with respect to their prediction performance when using experimentally determined interactions (INT1) as reference. For this all computed edges were split into two groups, namely with (INT1) and without (INT0) experimental backup, subsequently comparing their computed edge weights. As no explicit information on interactions was included in computing the edge weights such testing against experimentally determined interactions can be considered as independent validation. Weight distributions of the two edge sets showed a highly significant ($p<10^{-16}$, student's *t*-test) right shift for INT1, indicating higher computed scores for interactions also holding experimental background (mean edge weight of 0.52) than for edges only holding computed scores (mean edge weight of 0.29). To assess the impact of this shift in terms of prediction accuracy, as well as regarding the specific composition of a relation being of type "procedural" resting on an integration of GO process, Reactome and PANTHER pathway membership annotation, or "functional" via integrating protein domain and GO molecular function annotation, the receiver operator characteristic (ROC, plotting true positive rate versus false positive rate at various edge weight threshold settings) for the two groups of edges with INT1 as the prediction target were computed. INT1 edges above a specific computed edge weight cutoff were therefore interpreted as true positive interactions, while edges above the same value but not holding experimental evidence were interpreted as false positive interactions. The latter assumption needs to be seen with caution, as intention of the approach is to expand given interactions, in this verification setting being per definition interpreted as false positives.

Considering the entire INT1 set of 145,391 edges in the analysis, and using the computed edge weights for these interactions for deriving the AUC (area under the ROC curve) as composite expression regarding sensitivity and specificity of computed relations in contrast to INT1 provided an AUC value of 0.82 (Figure 6A). However, Reactome was used both for deriving interactions contributing to INT1 but also independently for pathway category annotation utilized in computing edge weights. Consequently, the computed edge weights and the target edges being compared against cannot be considered as fully independent. To identify a potential bias resting on this assignment, the computation of the ROC curves was also performed on an INT1 data set omitting interaction data from Reactome (contributing 3,872 interactions to the total set of 145,391 edges with experimental backup), resulting in a minor decrease of the AUC to 0.78 (Figure 6B). Performing the same procedures separately for edge weights for the edge classes "procedural" and "functional" (see equation 5) indicated improved performance for the "procedural" score regarding correct recovery of INT1 (Figure 6A), and seeing the expected decrease for the INT1 data set omitting Reactome (Figure 6B), as Reactome annotation was only used for computation of weights of type "procedural".

As second quantification strategy for evaluating the correctness of computed edges with respect to recovering edges also being reported experimentally the precision (the percentage of true positives from all positives) was computed. Precision values started off at 100% at an edge weight of 1.0, and dropped to 13.5% at a cutoff of 0.92, being mainly due to noise as only a small number of edges are included in this edge weight range. Precision peaked again at 63% at a cutoff of 0.89 and then decreased continuously with decreasing cutoff (data not shown). For estimating the test accuracy the F1 score, computed as the harmonic mean of the precision and recall rates, was investigated to identify the best precision/recall ratio, being identified at an edge weight cutoff of 0.74.

An equivalent analysis, but now utilizing a completely independent interaction dataset, was performed using 31,996 interactions derived from KEGG as the set of true positive interactions. For each cutoff in the edge weight interval (0,1) the true positive rate (the number of KEGG interactions correctly identified as such with respect to the total number of KEGG interactions) and the false positive rate (the number of non-KEGG interactions above that cutoff with respect to the total number of non-KEGG interactions) was computed. The resulting ROC showed an AUC of 0.79 (Figure 7). Similarly to evaluation of INT1 edges as shown in figure 6 the "procedural" score performed equally well as the effective edge weight (AUC of 0.77), while the "functional" score showed a lower AUC of 0.69.

Evaluation of graph characteristics as discussed above rested on the absolute edge weight above a certain cutoff. Consequently, nodes holding edges with low weights were neglected, hampering the completeness of the node set effectively represented in the graph derived at a specific edge weight cutoff.

As alternative edge selection strategy a relative order relationship rather than an absolute one may be applied, e.g. by picking the top ranked x % edges of each node irrespective of the absolute weight, by this naturally maximizing coverage of nodes. Validation of such edge selection strategy was performed, again using the set of INT1 interactions, as well as KEGG interactions. Prediction performance with respect to INT1 decreased only marginally (AUC of 0.77 compared

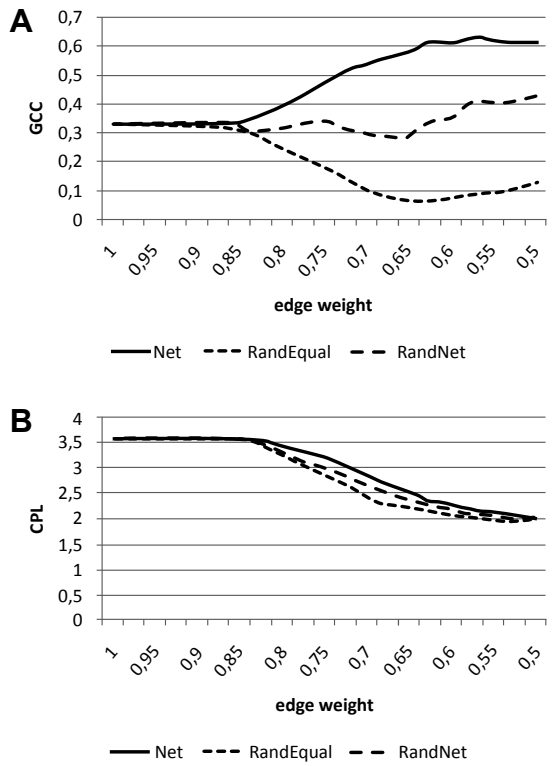

Figure 5: Global clustering coefficient (GCC) (A) and characteristic path length (CPL) (B) in relation to edge weight computed for the edge weight interval [0.5,1.0]. GCC and CPL are provided for the INT1 graph (only including edges with experimental background) extended by inferred edges according to equation 5 (Net), by extending with the same number of edges being randomly distributed between nodes (RandEqual), and by the adding the same number of edges as for Net but in addition taking the graph degree distribution of the computed graph (Net) into account (RandNet).

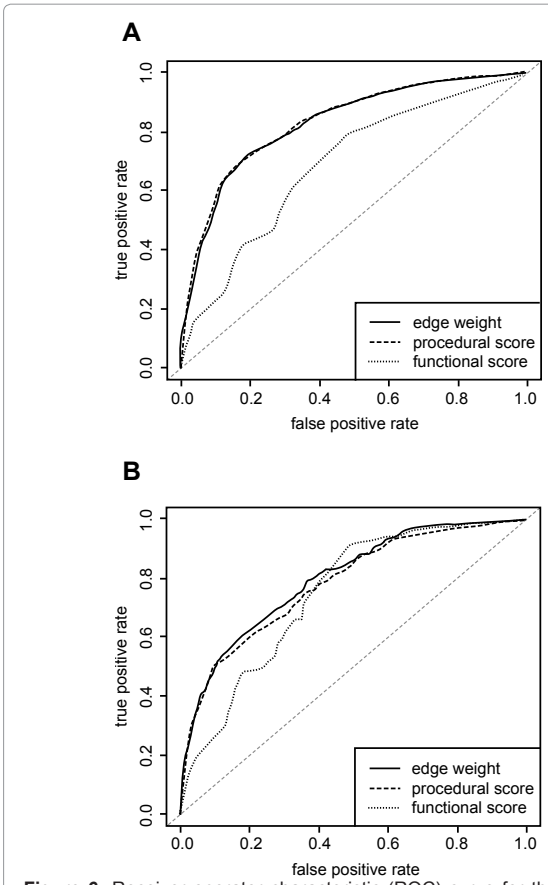

Figure 6: Receiver operator characteristic (ROC) curve for the entire set of INT1 (A) and for MI:0407 direct interactions only (B). For each weight cutoff in [0,1] the true positive rate (percentage of INT1 correctly identified from total INT1) is plotted against the false positive rate (percentage of computed interactions falsely assigned as interactions with respect to the experimental reference). ROC curves are plotted for the entire set of interactions, as well as separately for the interaction classes "procedural" (resting on annotation from GO process, Reactome and PANTHER pathway membership) and "functional" (resting on annotation from protein domain and GO molecular function).

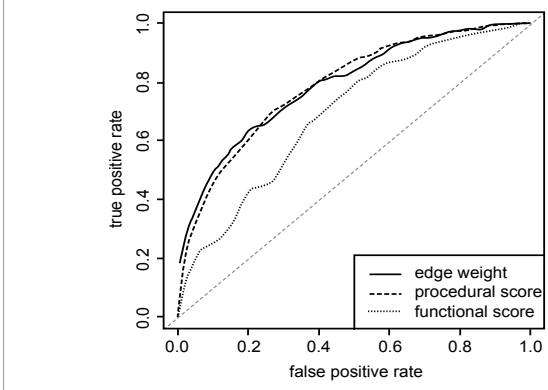

Figure 7: Receiver operator characteristic (ROC) curve for classifying KEGG interactions. For each weight cutoff in [0,1] the true positive rate (percentage of KEGG edges correctly assigned as interaction on the basis of the computed interaction) is plotted against the false positive rate (percentage of edges not represented in KEGG but predicted as being present). ROC curves are plotted for the entire set of interactions, as well as separately for the interaction classes "procedural" (resting on annotation from GO process, Reactome and PANTHER pathway membership) and "functional" (resting on annotation from protein domain and GO molecular function).

to an AUC of 0.82 when using the absolute cutoff criterion), while predicting KEGG edges was basically identical to the former approach (data not shown). While this alternative edge selection approach raised the node coverage at apparently little cost on accuracy, other factors need to be considered: Highly weighted edges were omitted if they do not fit in the top x % of their adjacent nodes, obviously increasing the false negative count.

Discussion

Correct molecular network representations address a major leap in understanding cellular processes, specifically when utilizing feature sets derived from Omics profiling. Various types of networks exist, including protein-protein interaction networks [36], metabolic [37] or regulatory networks as well as RNA networks [38]. A major shortcoming of available interaction data is their lack of completeness on both molecular feature level as well as interaction information, frequently coupled with a significant number of false positive interactions when considered in the context of biological relevance.

A number of approaches have become available utilizing alternative data sources and algorithmic approaches for computing probabilities for interactions. Examples include e.g. HumanNet, offering a probabilistic functional gene network of human protein-encoding genes utilizing a modified Bayesian integration of 21 types of 'omics' data from multiple organisms [39], or specific interaction networks e.g. for identification of disease genes [40,41]. In a previous work [42], we presented an approach to construct a dependency network based on adding biological information next to protein-protein interaction datasets. These additional data sources included pathways and ontologies, gene expression profiles and subcellular localization information utilizing predicted cellular location information as computed by WPSORT [43]. In the current work we present a generalized methodology based on a core network embedding interactions backed by experimental evidence or literature curation, and then expanding this core network with computed interactions utilizing information content to increase coverage both on a node (gene/protein) as well as on the edge (interaction) level. Datasets used for delineating such interactions included pathway membership information from Reactome and PANTHER, ontology information from Gene Ontology, and protein domain data from InterPro. Data were mapped to all human protein sequences as provided in ENSEMBL (86,934 entries), subsequently reduced to the canonical sequence of each gene/protein for which also data were available in the annotation sources used (17,022 entries).

Cutoff values for optimal coverage and accuracy

Inferring relations between these 17,022 nodes resulted in a complete graph with interaction weights for edges between 0.0 and 1.0, imposing the need for identifying an edge weight cutoff showing optimal accuracy at maximum node coverage. While a higher edge weight implies a higher probability regarding a true positive interaction, such a cutoff also leads to a drop in the number of edges included, in turn reducing the number of nodes represented in the graph. Accordingly, lowering the cutoff increased node coverage, going in hand with loss of precision on the edge level. The process of choosing an edge weight cutoff implies finding a balance between node coverage and edge accuracy.

We evaluated two methods for defining a cutoff, one where a single cutoff value was defined for all edges (absolute ranking method), and a second where for each node the top ranked x % edges were considered as true positive interactions (relative ranking method). Following the absolute ranking identified an edge weight of 0.65 as intrinsic lower boundary, as the network selected at an edge weight cutoff below 0.65 did not differ from a network generated by adding interactions randomly when tested on the level of global clustering coefficient (GCC) and characteristic path length (CPL). Such edge weight cutoff provided approximately 2,5 million edges for 14,872 nodes.

As upper edge weight boundary a value of 0.8 may be perceived. At this cutoff the network holds 204,128 edges and 11,921 nodes, i.e. providing only a minor increase with respect to the core (INT1) network holding 145,391 edges and 11,162 nodes.

As further criterion the characteristics of the degree centrality may be considered, being at steady values in the interval [0.71, 1.0], and seeing a significant increase below an edge weight cutoff of 0.71 (at this point providing 830,470 edges for 13,730 nodes). A further supportive factor for setting an edge weight cutoff in this range is the computed precision, seeing a maximum at an edge weight of 0.74, at this point providing 12,891 nodes and 533,020 edges.

Result graph characteristics

Stumpf et al. [18] estimated the size of the human interactome to hold about 650,000 interactions. In this context, the size of our resulting network at an edge weight cutoff of 0.74 is the same order of magnitude. Hart et al. [44] estimated the number of human protein-protein interactions to be situated somewhat lower between 154,000 and 369,000, while Venkatesan and colleagues [19] speculated on approximately 130,000 interactions. Certainly, with respect to node coverage there is still a gap of about 7,000 protein coding genes with respect to the 19,980 entries provided in ENSEMBL. However, utilizing the annotation approach presented in this work excluded 2,958 nodes due to lack of any annotation data for given sources, leaving 3,292 protein coding genes not exhibiting a single interaction scoring with an edge weight of at least 0.71. Contrasting the hybrid network with given reference networks provides additional 2,568 nodes when compared to the consolidated data from IntAct, Reactome, and BioGrid, naturally showing significantly increased coverage when e.g. compared to high evidence interaction networks as KEGG.

Notably, the hypothesis that the inferred edges exhibit a strong bias regarding general annotation level proved unfounded (Pearson R for comparing edge evidence level and edge weight of -0.2). Positive correlation, however, was identified for each node's strongest edge and the node's Gene Characterization Index (Pearson R=0.43) and specific level of node annotation (Pearson R=0.53), respectively.

The network at an edge weight cutoff of 0.71 is found to be more compact than the INT1 network, with a clustering coefficient of 0.51 as compared to 0.32 for the INT1 graph. The characteristic path length is found at about 3.0 compared to 3.5 for the INT1 graph. New edges added to the graph tend to link already included nodes in contrast to adding further nodes.

Graph validation

Validating such hybrid network is an essential step towards assessing the quality of the underlying methodology for inferring interactions. This is, however, in practice difficult to perform. One of the main challenges to address is finding an appropriate dataset to validate against. As the method aims at extending the presently referenced set of interactions, it is sensible to address the quality of the INT1 prediction. This is achieved by assessing the discriminative power of the method with respect to the INT0/INT1 classification.

An inherent feature of the INT1 dataset, however, is that it is by itself a heterogeneous collection of interactions. For the complete INT1 dataset, the receiver operating characteristic curve showed an AUC of 0.82 for different cutoff values in the interval between 0.0 and 1.0. While this value is indicative for good prediction quality with respect to INT1 detection, this is still subjected to debate as the goal of the method lies not in the sole prediction but in the extension of the INT1 dataset. In this context it can be argued that a high prediction value adds little novelty to the existing network, whereas a low prediction value may add novelty, but presumably also a significant fraction of false positive interactions.

Additional difficulty is added to the validation by the strongly asymmetric character of the two interaction set sizes, i.e. 145,391 edges for INT1 and 60 million for INT0 (only considering the set of INT0 edges between nodes present in INT1). While an AUC of 0.82 stands both for good precision and still significant novelty, the disproportionately large number of INT0 edges adds a great absolute number of INT0 edges already at a small false positive rate. At a cutoff at 0.71 the true positive and false positive rates were 11% and 0.7%, respectively. This stands for approx. 16,000 INT1 edges and 685,000 INT0 edges. By comparison, a 50% true positive rate with an 8% false positive rate was reached at a cutoff of 0.55. Utilizing KEGG as independent interaction source provided an AUC value of 0.79, being very much in the range of AUC values seen for INT1.

Conclusions

Molecular networks have become a central ingredient in Omics profile interpretation and hypothesis generation, consequently demanding networks with significant coverage of molecular entities combined with a comprehensive representation of interactions. The latter see various types of interactions, together with different levels of evidence regarding biological relevance. Hybrid networks aiming at integrating diverse data sources are a straightforward approach for expanding both, node and edge count, and provide a single reference network for Omics data mapping and interpretation. Although information on interactions of protein coding genes expands on a continuous basis, computational inference of interactions on top of database information, as introduced in this work, adds to a more complete representation of the interactome, expanding opportunities for Omics profile-based hypothesis generation.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme under the grant agreement n° 241544.

References

1. Alexeyenko A, Schmitt T, Tjärnberg A, Guala D, Frings O, et al. (2012) Comparative interactomics with Funcoup 2.0. Nucleic Acids Res 40: D821-828.
2. Fechete R, Heinzel A, Perco P, Mönks K, Söllner J, et al. (2011) Mapping of molecular pathways, biomarkers and drug targets for diabetic nephropathy. Proteomics Clin Appl 5: 354-366.
3. Barabasi AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 12: 56-68.
4. Heinzel A, Fechete R, Soellner J, Perco P, Heinze G, et al. (2012) Data Graphs for Linking Clinical Phenotype and Molecular Feature Space. International Journal of Systems Biology and Biomedical Technologies 1: 11-25.
5. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40: D109-114.
6. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13: 2129-2141.
7. Brown KR, Jurisica I (2005) Online predicted human interaction database. Bioinformatics 21: 2076-2082.
8. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, et al. (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res 40: D841-846.
9. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. (2011) The BioGRID Interaction Database: 2011 update. Nucleic Acids Res 39: D698-704.
10. Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. (2012) Ensembl 2012. Nucleic Acids Res 40: D84-90.
11. Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, et al. (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods 9: 345-350.
12. Liu G, Li J, Wong L (2008) Assessing and predicting protein interactions using both local and global network topological metrics. Genome Inform 21: 138-149.
13. Tyagi M, Hashimoto K, Shoemaker BA, Wuchty S, Panchenko AR (2012) Large-scale mapping of human protein interactome using structural complexes. EMBO Rep 13: 266-271.
14. Bader JS, Chaudhuri A, Rothberg JM, Chant J (2004) Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol 22: 78-85.
15. Bjorkelund H, Gedda L, Andersson K (2011) Avoiding false negative results in specificity analysis of protein-protein interactions. J Mol Recognit 24: 81-89.
16. Kuchaiev O, Rasajski M, Higham DJ, Przulj N (2009) Geometric de-noising of protein-protein interaction networks. PLoS Comput Biol 5: e1000454.
17. Sprinzak E, Sattath S, Margalit H (2003) How reliable are experimental protein-protein interaction data? J Mol Biol 327: 919-923.
18. Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, et al. (2008) Estimating the size of the human interactome. Proc Natl Acad Sci U S A 105: 6959-6964.
19. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, et al. (2009) An empirical framework for binary interactome mapping. Nat Methods 6: 83-90.
20. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37: D619-622.
21. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
22. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, et al. (2011) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40: D306-312.
23. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, et al. (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32: D115-119.
24. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, et al. (2004) The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat Biotechnol 22: 177-183.
25. Mistry M, Pavlidis P (2008) Gene Ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics 9: 327.
26. Lord PW, Stevens RD, Brass A, Goble CA (2003) Semantic similarity measures as tools for exploring the gene ontology. Pac Symp Biocomput: 601-612.
27. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF (2007) A new method to measure the semantic similarity of GO terms. Bioinformatics 23: 1274-1281.
28. Resnik P (1999) Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11: 95-130.
29. Kemmer D, Podowski RM, Yusuf D, Brumm J, Cheung W, et al. (2008) Gene characterization index: assessing the depth of gene annotation. PLoS One 3: e1440.
30. Maglott D, Ostell J, Pruitt KD, Tatusova T (2011) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 39: D52-57.
31. Paul I, Ahmed SF, Bhowmik A, Deb S, Ghosh MK (2012) The ubiquitin ligase

CHIP regulates c-Myc stability and transcriptional activity. Oncogene 32: 1284-1295.

32. Mukhopadhyay D, Riezman H (2007) Proteasome-independent functions of ubiquitin in endocytosis and signaling. Science 315: 201-205.

33. Soucek L, Whitfield J, Martins CP, Finch AJ, Murphy DJ, et al. (2008) Modelling Myc inhibition as a cancer therapy. Nature 455: 679-683.

34. North BJ, Verdin E (2004) Sirtuins: Sir2-related NAD-dependent protein deacetylases. Genome Biol 5: 224.

35. North RA (2002) Molecular physiology of P2X receptors. Physiol Rev 82: 1013-1067.

36. Hou J, Chi X (2012) Predicting protein functions from PPI Networks Using Functional Aggregation. Math Biosci 240: 63-69.

37. De Martino D, Figliuzzi M, De Martino A, Marinari E (2012) A scalable algorithm to explore the Gibbs energy landscape of genome-scale metabolic networks. PLoS Comput Biol 8: e1002562.

38. Sumazin P, Yang X, Chiu HS, Chung WJ, Iyer A, et al. (2011) An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. Cell 147: 370-381.

39. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res 21:1109-1121.

40. Wu G, Feng X, Stein L (2010) A human functional protein interaction network and its application to cancer data analysis. Genome Biol 11: R53.

41. Zhang W, Sun F, Jiang R (2011) Integrating multiple protein-protein interaction networks to prioritize disease genes: a Bayesian regression approach. BMC Bioinformatics 12: S11.

42. Bernthaler A, Mühlberger I, Fechete R, Perco P, Lukas A, et al. (2009) A dependency graph approach for the analysis of differential gene expression profiles. Mol Biosyst 5: 1720-1731.

43. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, et al. (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Res 35: W585-587.

44. Hart GT, Ramani AK, Marcotte EM (2006) How complete are current yeast and human protein-interaction networks? Genome Biol 7: 120.