Jafari et al. BMC Bioinformatics (2015) 16:319
DOI 10.1186/s12859-015-0755-1

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Interlog protein network: an evolutionary benchmark of protein interaction networks for the evaluation of clustering algorithms

Mohieddin Jafari[1,2*], Mehdi Mirzaie[3] and Mehdi Sadeghi[4*]

Abstract

Background: In the field of network science, exploring principal and crucial modules or communities is critical in the deduction of relationships and organization of complex networks. This approach expands an arena, and thus allows further study of biological functions in the field of network biology. As the clustering algorithms that are currently employed in finding modules have innate uncertainties, external and internal validations are necessary.

Methods: Sequence and network structure alignment, has been used to define the Interlog Protein Network (IPN). This network is an evolutionarily conserved network with communal nodes and less false-positive links. In the current study, the IPN is employed as an evolution-based benchmark in the validation of the module finding methods. The clustering results of five algorithms; Markov Clustering (MCL), Restricted Neighborhood Search Clustering (RNSC), Cartographic Representation (CR), Laplacian Dynamics (LD) and Genetic Algorithm; to find communities in Protein-Protein Interaction networks (GAPPI) are assessed by IPN in four distinct Protein-Protein Interaction Networks (PPINs).

Results: The MCL shows a more accurate algorithm based on this evolutionary benchmarking approach. Also, the biological relevance of proteins in the IPN modules generated by MCL is compatible with biological standard databases such as Gene Ontology, KEGG and Reactome.

Conclusion: In this study, the IPN shows its potential for validation of clustering algorithms due to its biological logic and straightforward implementation.

Keywords: Protein-protein interaction network, Interlog protein network, Evolution, Network module

Background

One of the important challenges in the interpretation of proteomic data is the detection of the cellular active process by exploring protein function. This newly emerging discipline; network science, has demonstrated that the majority of biological and evolutionary concepts make sense in the light of Systems Biology [1, 2]. Hence, the protein function, and consequently, cell function are more clearly demonstrated in the context of protein interaction network [3]. Protein-protein interactions make up the major branch in the study of protein interaction networks. From a biochemical view, these interactions can be divided into two categories: physical and functional [4, 5]. There are several methods employed to describe these interactions. The advantages and disadvantages of these methods have been widely reviewed [6–8]. Different limitations such as slow- and small-scale performances, inability to identify protein complexes, artificial interaction obtained from the in vitro assay and the operational restrictions, led to the discovery of methods that complement each other [9]. Some experimental methods are involved in determining physical and functional interactions [10, 11]. The phylogenetic profile, Rosetta stone, gene neighborhood and co-evolution are the most prevalent computational methods [11–13].

Biological network modules

After constructing a Protein-Protein Interaction Network (PPIN), the next step is the exploration of the protein

* Correspondence: mjafari@ipm.ir; sadeghi@nigeb.ac.ir
[1]Drug Design and Bioinformatics Unit, Medical Biotechnology Department, Biotechnology Research Center, Pasteur Institute of Iran, 69 Pasteur St, PO Box 13164, Tehran, Iran
[4]National Institute of Genetic Engineering and Biotechnology (NIGEB), Pajoohesh Blvd, 17 Km Tehran-Karaj Highway, PO Box 161-14965, Tehran, Iran
Full list of author information is available at the end of the article

tasks within this complex circuit. As Alessandro Vespignani mentioned, "evolution thinks modular" [14]; a cell's activity is a result of groups of interacting proteins, known as functional module (if they do not necessarily interact at the same time and place), in PPIN [15, 16]. Therefore, the PPIN modules should be identified and determined and then a biological function could be assigned to them based on the protein annotations. Sometimes this procedure is specifically more successful for the protein complexes that work together at the same time and place, rather than for the functional modules [6, 9].

Module detection can be divided into two approaches, namely graph clustering, and distance-based clustering. In the first approach, algorithms seek communities of the nodes in the graph that contain more intra-edges than inter-edges, e.g. Super Paramagnetic Clustering (SPC) [17], Highly Connected Subgraph (HCS) based on the Monte Carlo algorithm [18], Markov clustering (MCL) [19] and Restricted Neighborhood Search Clustering (RNSC) [20]. In the distance-based approach, the clustering algorithms e.g. the hierarchical or k-means are used so that the concept of distance and its associated measures in graph theory are applied as the similarity measures in the clustering. Some of the distances used in this approach are as follows: shortest path [21], number of edges [22, 23], shortest path profiles [24, 25] and a combination of distance and the statistical objects [26]. The detected modules are then well-characterized, biologically, based on information beyond the network topology such as gene expression, cell localization, virulence and knockout phenotypes [6, 27].

The biological data are also applied in the next step for validation assays of module finding. The validation is based on the protein annotation homogeneity in the modules and these annotations could consist of functional, structural, local and/or interactomic information. Generally, the Gene Ontology (GO) and MIPS database are used in PPIN module validation [20, 28–32]. In addition, data from the gene expression profiles, co-localization and gene phenotype are also used [9]. It should be highlighted that the major presumption of the validation is that most of the proteins in a module, i.e. a statistically significant number of proteins, should be similar in intended attribute if the modules were identified correctly. A pictogram of this procedure is summarized in Fig. 1.

Although, protein complexes consist of functional and also physical interactions at the same time, it is obvious that many of the functional interactions are not in the data of the protein complexes. Several protein interactions happen transiently and indirectly, and as such are not detectable by empirical routine tests. These are important steps in the protein interactions which are lacking in the MIPS database [33–35]. In addition, modeling and representing the protein complex acquired by some experimental techniques, such as a graph (e.g. "spoke" and "matrix" model) is a challenging issue [36]. Furthermore, one cannot ignore the fact that the inherited experimental errors inherent to these problems and many more protein complexes, have not been fully studied as yet [37]. These flaws may lead to misinterpretation in the validation step if we use only the MIPS database.

On the other side, GO is a standard glossary of biological terms known as the first and most common reference for the Biological Process (BP), Molecular Function (MF) and Cellular Compartment (CC) of the proteins [38]. Additionally, a significant correlation between the node distances in some biological networks and the semantic similarity of their GO terms has also been reported [39–42]. Although the GO contains comprehensive and organized information, it has some limitations, namely, insufficient GO annotations (35–55 % false-negatives) [43], inaccurate GO annotation (false-positives, it should be observed that most of the annotations in GO are obtained by an indirect method such as gene manipulation as well as heterogeneous experimental and computational data), the functional diversity of proteins under different conditions resulting in different and sometimes conflicting annotations for one protein (false-positives) [44] and errors due to the manual annotation approach [45]. These deficiencies lead also to misinterpretation in the validation step.

The goal of present study

Regarding the aforementioned restrictions in the validation step of module finding algorithms, we propose a network-based evolutionary benchmark as a complementary approach to solving some of the presented issues. Recently, in a companion study [46], we introduced a common network, that had low false-positives and tuned false-negatives. Using the four PPINs, a network with a high degree of conservation between four species was constructed. We call this common network the Interlog Protein Network (IPN) (Additional file 1). In the present study, the IPN, which is confirmed using experimentally proteomic data, has been suggested to be applied as a complementary benchmark in the validation of the different module finding algorithms, namely, Markov Clustering (MCL) [19], Restricted Neighborhood Search Clustering (RNSC) [20], Cartographic Representation (CR) [47], Laplacian Dynamics (LD) [48, 49] and Genetic Algorithm, to find communities in Protein-Protein Interaction networks (GAPPI) [16].

Results and discussion

Mitochondrial IPN

In the current study, the mitochondrial IPN of the four eukaryotic species was constructed. These species consisted of, human, rat, fruit fly and worm. The IPN was

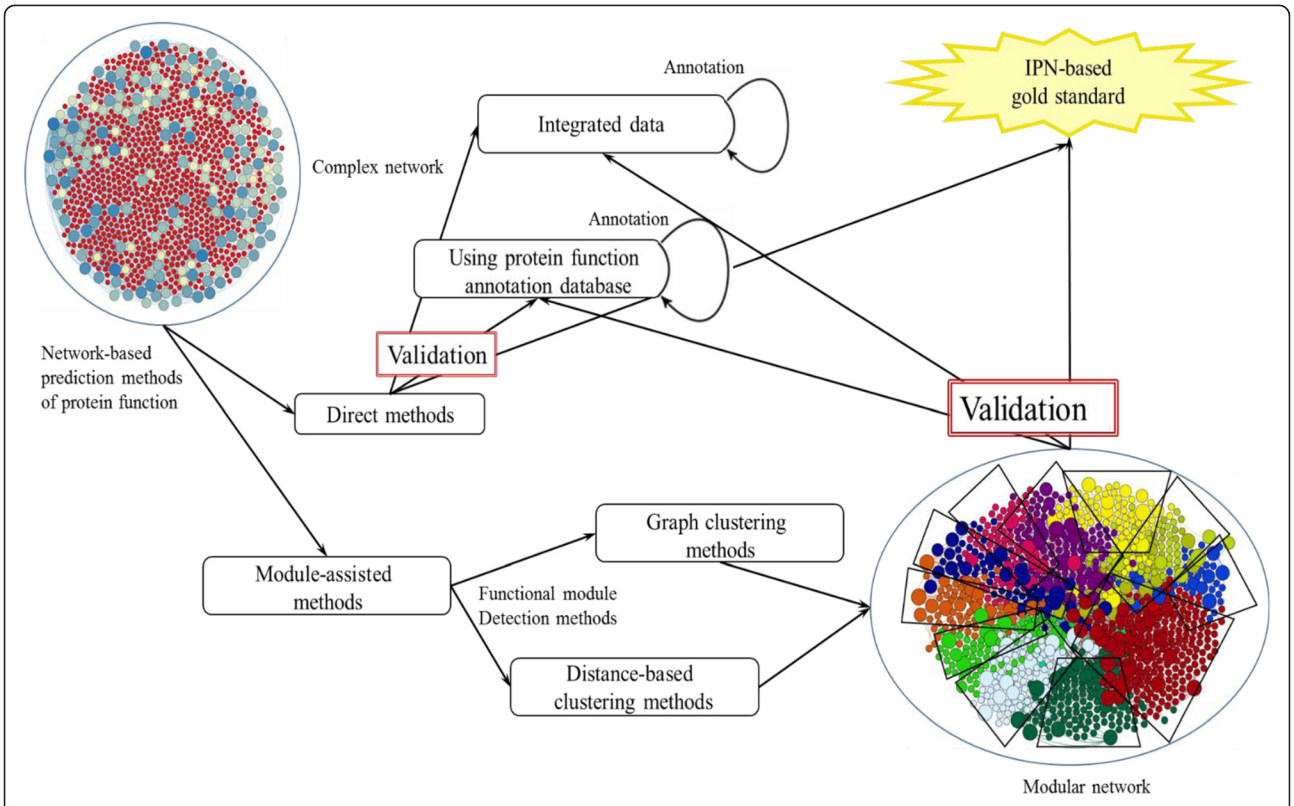

Fig. 1 Function prediction procedures in network biology. Network-based prediction methods of protein function are described schematically. There are two approaches to explore function from network in biology, direct and module-assisted methods. In the direct method which is not our subject in this study, the annotation of gene/protein neighbors are used to predict function. But in the module-assisted methods, node's community/neighborhood is principal for function prediction. These methods are also divided into two categories i.e. graph and distance-based clustering methods. The validation is the main step after finding modules in the PPIN or using direct methods. Any assignment of function based on annotation of neighbors or neighborhood should be evaluated by the different validating methods. We introduce the IPN-based validation for this purpose

achieved through the interlog finding procedure of the mitochondrial PPIN of these species. In the other words, the IPN is an evolutionarily conserved network obtained from the overlap of orthologous proteins reinforced by gene expression. By pair-wise sequence alignment (≥30 %), the 226 Orthologous Protein Sets (OPSs) were obtained. Each OPS contained four orthologous proteins from the four species (83 human, 82 rat, 83 fruit fly and 80 worm). Finally, the IPN showed 29 nodes, 61 edges, 4.34° on average, a diameter of 6, and an average clustering coefficient of 0.625 (Additional files 1 and 2). This network represents the evolutionarily conserved topological network features shared among these species.

The expression data is used to empirically validate the IPN. The significantly high correlation between the protein concentrations endorsed the edges in the IPN. In fact, the correlation or co-expression network was reconstructed based on this concentration data and this network was compared to the IPN. In the previous study, the rat mitochondrial proteins (~500) were analyzed by several electrophoresis techniques [50]. It was claimed that different electrophoresis techniques are capable of fractionating proteins with different subcellular localizations [50]. Hence, the significant correlation between the expression profiles of the proteins in different electrophoresis implies they are co-localized proteins [51–53].

Of the total 563 proteins reviewed (UniProtKB database) 82 were in the mitochondria of rats which participated in OPSs. Later, 31 proteins were involved in the IPN and 20 of the 31 proteins were detected experimentally and involved in the co-expression network. In all, there were 23 significantly high correlations among these 20 proteins that matched with 13 of the 22 links in the IPN between these 20 proteins. The hypergeometric test confirms the reasonable matching ratio of the IPN ~60 % with respect to the edge match number in the rat PPIN 12 % with p-value $3.9 \times 10–7$ (Table 1).

Comparison of network clustering methods

All the methods including RNSC, MCL, CR, LD and GAPPI were performed to discover modules in all four PPINs and the IPN. It should be noted that all of these algorithms are unsupervised and the network size affects the number of clusters. By defining the IPN as

Table 1 IPN expression data

Network name	STRING derived networks (Nodes and Edges)	Co-expression networks (Nodes and Edges)	Matched edges	Ratio of matched edges
Interlog protein network (IPN)	31, 63	20, 22	13	~60 %
Protein-protein interaction network (PPIN)	563, 2431	230, 1205	150	12 %

Using the expression proteomic data, a correlation network was constructed and compared to database derived networks i.e. IPN and PPIN. This evaluation was performed for rat mitochondrial proteins and related proteins and their links were considered. The number of nodes and edges in STRING derived (col 2) and co-expression (col 3) networks are presented in this table along with the number of matched or same edges in two corresponding networks (col 4). The edge matching ratios of these networks are represented (col 5)

the benchmark, some external measures were used to validate all the methods. Next, some comparison indices were used, including Jaccard, Rand, Fowlkes-Mallows and Minkowski for all species. Except for the Minkowski index with the range [0, +∞) (where the values near to zero indicated the greater similarity) the other indices have a range [0,1] and the values closer to zero, indicate greater inconsistency.

The Rand index (unlike the Jaccard, Fowlkes-Mallows and Minkowski indices), measures the degree of similarity between two matrices as a function of the positive and negative agreements. Some studies claimed that the n_{00} value (Number of paired entities in the similarity matrices in which both are 0, see Methods) was often larger than the other values in most of the gene clustering studies and suggested using three other indices [54]. On the other hand, the Jaccard index is recommended due to its low variance [55]. However, the mean value of all the indices and standard deviations are shown graphically in Fig. 2 (All the values are shown in Additional file 2).

As represented in Fig. 2, our defined network showed that MCL outperforms the clustering methods of GAPPI, RNSC, LD and CR in terms of external measure indices. The range of the standard deviations showed that the MCL is more dependent on the size of the graph. This superiority was evident in all the indices, even in the Rand index with the above-mentioned imperfection. However, in the case of human PPIN as a large network, MCL and GAPPI cluster with similar accuracy (Additional file 2).

Meanwhile, the superiority of the MCL is compatible with the earlier results [32]. By distinct approach, they presented a comparative assessment of clustering algorithms and showed that the MCL was remarkably robust in graphing alterations and capable of the extraction of the complexes from the PPINs. CR took a long computation time and could not specify the modularity as well as the other algorithms within a reasonable time. The RNSC, LD and CR clustering showed similar ability to find the module robustly but the LD algorithm showed the lowest standard deviation

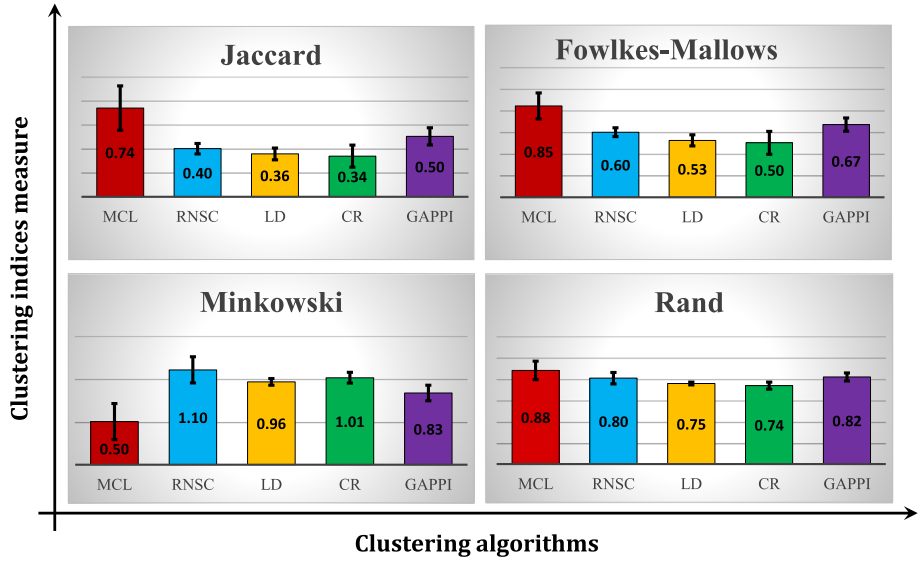

Fig. 2 External clustering indices. The average of the calculated indices with SD (Standard Deviation) bars are shown graphically in the five clustering algorithms. As shown, the MCL outperformed in all the indices including even the Rand index with argued imperfection. Note that the range of the Minkowski index is [0, +∞) and the values (here is MCL) near to zero indicate the more similarity. But the other indices range is [0, 1] and the values near to zero specify the less similarity

among all the indices calculated. The GAPPI as the most recently proposed algorithm for this problem, works better than RNSC, LD and CR. This algorithm takes second place after MCL in all comparisons and it clusters large network same as MCL. This pattern was almost repeated in the recent study [16] using MIPS as gold standard.

Biological evidence

The biological relevance of the proteins in each module detected by MCL was assessed. By using the Enrichr tool [56], three well-known biological standard databases are used namely; Gene Ontology (Biological Processes) [38], KEGG [57] and Reactome [58]. The result shows that each module enriched significantly and annotated separately (Additional file 3). Briefly, in 3 modules of this conservative IPN, the results are as follows. The first module is related to citrate/TCA cycle and oxidation phosphatase based on these ID numbers (GO:0006099, GO:0022904, GO:0022900, ko00020 and ko00190). The second module is related to Mitochondrial protein import based on these ID numbers (GO:0006626, GO:0070585, GO:0072655 and GO:0006839). The last module is related to Mitochondrial translation, ribosome and nucleoside biosynthetic process based on these ID numbers (GO:0046031, GO:0009133, GO:0046033, GO:0009135, GO:0009179, ko03010, ko00240 and ko00230). These results are compatible with our earlier study about biological meaningful communities in IPN as a pure evolutionary extract of mitochondrial PPIN [46].

Conclusion

There are several module detection methods based on different approaches. Validation assays are required to compare and select the best one for network analysis. The major prerequisite for validation is the determination of the reliable benchmark. A standard topological and functional PPIN helps us to assess and verify the PPIN modularity results. In the earlier studies, researchers used the MIPS or GO dataset as the gold standard in validation assays. As mentioned earlier, these datasets are not point-device gold standard and each one has its own particular shortcomings. In other words, these databases have been designed with specific purpose and are diverse conceptually [59].

In the current study, we used the pair-wise sequence alignment and comparative interactomics of evolutionary distant species to reconstruct a conserved and common network that can be used as the benchmark or ground truth. The proposed benchmark does not have the above-mentioned limitations. First, the edges (interaction data) in IPN and associated compared networks are generally of the same origin. This implies that if the edges of the associated compared networks are predicted and designated computationally, this benchmark is also constituted from the computational data and so on for experimentally identified interaction. In other words, the IPN edges are a result of the filtering procedure (see Additional file 4) and they do not originate from logically distinct methods. Second, the IPN reconstruction procedure most likely leads to a network with low false-positives and tuned false-negatives. This issue has a high impact on the assessment results in the validation step. Third, the reconstruction of IPN is possible for all the sequenced proteins and genes that are well-conserved across multiple species with predicted interactions. It implies that this approach does not require special expensive and time consuming techniques to generate the experimental data and evaluate the molecular networks.

Similar to the previous result [32], but dissimilar in approach, we found MCL to be the outstanding algorithm based on its performance in the comparison study. In the traditional method, the MIPS database was used to evaluate the different clustering methods. The sensitivity and accuracy of the different methods was also examined by adding and subtracting the edges i.e. artificial false-positives and negatives (Note that their tests did not contain large size changes). Our findings about GAPPI implementation are also consistent with the prior study [16], which showed the improving ability of a genetic algorithm to search modules in PPINs based on the MIPS database.

However, interaction data was retrieved from the STRING database which includes different sources of information, including various experimental, computational and even text mining methods [60, 61]. In addition, an independent set of empirical data was applied and the IPN quality was experimentally confirmed. However, our goal was to search for and introduce a method that could segregate the functional modules. It should be noted again that a functional module means a group of cell components and their interactions that do or do not promise specific biological functions at the same time and place. So, these modules also include all the protein complexes. Therefore, the validation standard should not lack the functional interaction data. MCL is the superior module detection method in exploring the protein complexes and also for the functional modules based on the previous [32] and current studies, respectively. In addition, in terms of different graph sizes, it appears that MCL is not as robust as the other algorithms based on the range of the standard deviation.

In this study, we suggest the IPN to justify the modularity results of any PPIN due to three preponderances mentioned above. The graph clustering

algorithm would be inefficient if it could not find the modules analogously in the individual PPINs and IPN as a purified, conserved and confirmed network. This approach to make a new benchmark may also help to assess and verify other biological networks e.g. gene regulatory networks or gene correlation networks and other biological network analysis methods such as network motif finding or orienting PPINs, which are subjects for further research. Again, this approach uses evolutionary concept i.e. conservedness to evaluate the biological networks. This is reminiscent of the well-known quote, "Nothing in biology makes sense, except in the light of evolution" [62].

Methods

Interlog protein network (IPN)

Construction of the common PPIN or IPN has been described earlier in detail [46]. Briefly, the mitochondrial reviewed proteins were retrieved from the four eukaryotic model species (Rattus norvegicus, Drosophila melanogaster, Caenorhabditis elegans and Homo sapiens) from the UniProtKB database (UniProt release 2013_02) [63]. Then, Using the Needleman and Wunsch algorithm, the homologous proteins were identified in the OPSs. In the next step, four distinct mitochondrial PPINs of the four species were identified from the STRING database (Ver. 9) [61]. The four PPINs were elicited with the default value in the database by all the prediction methods. Finally, by applying a stringent rule that is the existence of interlog in all four species, the mitochondrial IPN of these species was enucleated.

To explain more, an edge links two OPSs, say OPS1 = (p1h, p1r, p1f, p1w) and OPS2 = (p2h, p2r, p2f, p2w), if the protein pairs i.e. (p1h, p2h), (p1r, p2r), (p1f, p2f) and (p1w, p2w), have interaction based on STRING database (h, r, f and w indicate human, rat, fruit fly and worm respectively). Therefore, those OPSs which do not satisfy this condition will not be used in the IPN. In other word, the IPN was constructed in a way that each edge between two OPSs in the IPN indicates the six interlogs in these species (In this example, if it is a link between OPS1 and OPS2, then there are six interlogs i.e. ((p1h, p2h) and (p1f, p2f)), ((p1h, p2h) and (p1w, p2w)), ((p1h, p2h) and (p1r, p2r)), ((p1r, p2r) and (p1f, p2f)), ((p1r, p2r) and (p1w, p2w)) and ((p1f, p2f) and (p1w, p2w))). It should be noted that in each step some proteins are pretermitted to discern conserved structures (Additional file 4).

Proteomics data

The results of the mitochondrial proteomic study of rat [50] were used for the empirical evaluation. In the shotgun proteomics strategy, the rat liver proteome with different cellular compartments was detected and quantified by several gel-based fractionation techniques. In the present study, normalized peptide counts were used to estimate the protein concentrations in a label-free quantification method. Then, Pearson correlation was applied to find the correlated proteins ($|r\ \text{coefficient}| \geq 0.7$, $P\text{-value} \leq 0.05$). According to the distinction made by the electrophoresis methods, the correlated proteins are likely co-localized. And, also as discussed earlier, co-localization can confirm the protein-protein interaction. Later, the ratio of the correlated proteins in the rat PPIN and IPN was computed separately and compared with the hypergeometric test ($P\text{-value} \leq 0.001$). Thus, the IPN edges were examined by independent experimental data statistically.

Network clustering algorithms

In the present study, five well-known different clustering algorithms (MCL, RNSC, CR, LD and GAPPI) were used to cluster the PPINs and IPN. The general characteristics of each algorithm are shown in the Table 2 and the associated references [19, 20, 47–49]. We presented further details regarding these algorithms in the Additional file 5. We clustered all the PPINs of species and IPN independently.

Evaluation of the clustering results

After clustering, validation is required to confirm the results or compare the different methods. A new benchmark was introduced, i.e., IPN in the validation step, so that the modules corresponding to each of the PPINs are compared with the IPN's modules. In fact, the IPN was used as the ground truth in the standard external measures assay. Note that the clustering results on the PPIN are restricted to those proteins also in the IPN. It was expected that the successful algorithm should be able to find the modules analogously in PPIN and IPN. In order to assess the clustering results, the similarity matrices (Symmetric binary matrices) of clustering results were constructed, such that a 1 indicated placing two objects in the same cluster or module and a 0 indicated the opposite. Then, the entities of each of the PPINs and IPN matrices were compared with each other. If the corresponding entities in the two matrices were equal, the two clustering methods resulted in the same clusters. The following four conditions occurred: Agreements; n11 (Number of paired entities in the similarity matrices in which both are 1) and n00 (Number of paired entities in the similarity matrices in which both are 0), Disagreements; n10 (Number of paired entities that are 1 in the PPIN similarity matrix and 0 in the IPN similarity matrix) and n01 (Number of paired entities that are 1 in the IPN similarity matrix and 0 in the PPIN similarity matrix).

There are several benchmarking indices to measure the degree of agreement and disagreement between the

Table 2 Main features of the graph clustering methods presented in this study

	Markov clustering (MCL)	Restricted Neighborhood Search Clustering (RNSC)	Laplacian dynamics (LD)	Cartographic Representation (CR)	Genetic Algorithm to find communities in Protein-Protein Interaction networks (GAPPI)
Type	Flow simulation & Pagerank centrality	Cost-based local search	Multiscale modular structure	Inter- and intramodule connection	Search inspired by natural evolution
Allow multiple assignations	No	No	No	No	No
Allow unassigned nodes	No	No	No	No	No
Edge-weighted graphs supported	Yes	No	Yes	No	No
First application	Protein family detection	Protein complex prediction	High modularity partitions of large (more than million) networks finding	Metabolic network	Protein-protein interaction networks
Availability	http://rsat.scmbb.ulb.ac.be/rsat/index_neat.html	http://rsat.scmbb.ulb.ac.be/rsat/index_neat.html	Upon Gephi program	Upon request	http://staff.icar.cnr.it/pizzuti/codes.html
Developer (Year)	Enright A.J. et al. (2002) [19]	King A.D. et al. (2004) [20]	(1) Lambiotte R. et al. (2007) [49]; (2) Blondel V.D. et al. (2008) [48]	Guimera R. & Amaral LAN (2005) [47]	Pizzuti C. & Rombo S. E. (2014) [16]

two matrices [55]. Some of the indices used in this study are as follows:

$$Rand = \frac{(n_{11} + n_{00})}{(n_{11} + n_{10} + n_{01} + n_{00})}$$

$$Jaccard = \frac{n_{11}}{(n_{11} + n_{10} + n_{01})}$$

$$Minkowski = \sqrt{\frac{(n_{10} + n_{01})}{(n_{11} + n_{01})}}$$

$$\text{Folkes–Mallows} = \frac{n_{11}}{\sqrt{(n_{11} + n_{01}) \times (n_{11} + n_{10})}}$$

In order to perform biological evaluation of IPN modules, Enrichr software was used [56]. In this web-based tool, significantly enriched terms are extracted based on the Gene Ontology, Biological Processes [38], Kegg Orthology [57] and Reactome databases [58]. The combined score; consisting of the Z-score and adjusted *p*-value, was used to rank and define enriched terms. This validation was done for the modules defined by MCL algorithm as a superior algorithm in our comparison.

Additional files

Additional file 1: IPN. The IPN derived from the four different PPINs of four species is shown. They include human, worm, fruit fly, and rat PPINs from down left clockwise. Nodes with greater degrees are indicated as circles with larger diameters and each module is shown with the specific node color. (JPEG 4673 kb)

Additional file 2: Network Parameters. Some network parameter are included to gain clear insight about four PPINs and IPN. Also, all calculated external clustering measures and indices are presented in this table. The ranges for Rand, Jaccard and Fowlkes-Mallows are [0, 1] which are presented in percent whereas the Minkowski rang is [0, +∞). These values are computed for all species and five clustering algorithms. (DOCX 17 kb)

Additional file 3: Biological evidence. In this file, the IPN nodes, the MCL modules, the human proteins in OPSs and the enriched terms of proteins consist IPN modules are presented. The enrichment analysis of each module presented separately in different sheets. The human proteins were used to perform this analysis. (XLSX 38 kb)

Additional file 4: IPN reconstruction steps. First, the mitochondrial proteins are extracted from the *UniProt* database, and then the reviewed proteins are filtered. Using the Needleman and Wunsch algorithm, the homologous proteins are identified in the OPSs. In the next step, four distinct PPINs from four species are identified from the STRING database. Finally, the IPN is created by finding the interlog proteins in all four PPINs. In each step some proteins are pretermitted to discern conserved structures. (PDF 123 kb)

Additional file 5: Some details regarding the five clustering algorithm algorithms (MCL, RNSC, CR, LD and GAPPI) are described briefly. (DOCX 13 kb)

Competing interest
The authors declare that they have no competing interests.

Authors' contributions
MJ, MM and MS participated in the design of the study. MJ carried out acquisition, computational analysis and interpretation of the data and drafted the manuscript. MM helped computational analysis and in the writing of the draft. MS conceived of the study and participated in revising the manuscript critically. All authors read and approved the final manuscript.

Authors' information
Not applicable

Acknowledgment
The authors would like to express their gratitude to Dr. Anne-Ruxandra Carvunis for thoroughly reviewing the manuscript and to Dr. Sayed-Amir Marashi for his useful comments. Valuable contributions were also made by Zhi Wang, Jianzhi Zhang (University of Michigan) and Clara Pizzuti (National Research Council of Italy) in providing module finding programs (CR and GAPPI respectively) and by Mohammad-Reza Okhovat in graphically designing some of the Figures. The computation was performed at Math. Computing Center of IPM (http://math.ipm.ac.ir/mcc). The article processing charge has been waived by BioMed Central.

Author details
[1]Drug Design and Bioinformatics Unit, Medical Biotechnology Department, Biotechnology Research Center, Pasteur Institute of Iran, 69 Pasteur St, PO Box 13164, Tehran, Iran. [2]School of Biological Science, Institute for Research in Fundamental Sciences (IPM), Shahid Lavasani St, PO Box 19395-5746, Tehran, Iran. [3]Department of Computational Biology, Faculty of High Technologies, Tarbiat Modares University, Jalal Ale Ahmad Highway, PO Box 14115-111, Tehran, Iran. [4]National Institute of Genetic Engineering and Biotechnology (NIGEB), Pajoohesh Blvd, 17 Km Tehran-Karaj Highway, PO Box 161-14965, Tehran, Iran.

Received: 16 March 2015 Accepted: 29 September 2015
Published online: 05 October 2015

References
1. Barabási AL. The network takeover. Nat Phys. 2011;8(1):14–6.
2. Carvunis AR. From proteins and their interactions to evolutionary principles of biological systems. Université Joseph Fourier. 2011.
3. Hou J, Chi X. Predicting protein functions from PPI networks using functional aggregation. Math Biosci. 2012;240(1):63–9.
4. Srivas R, Hannum G, Ruscheinski J, Ono K, Wang P-L, Smoot M, et al. Assembling global maps of cellular function through integrative analysis of physical and genetic networks. Nat Protoc. 2011;6(9):1308–23.
5. Bandyopadhyay S, Kelley R, Krogan NJ, Ideker T. Functional maps of protein complexes from quantitative genetic interaction data. PLoS Comput Biol. 2008;4(4):e1000065.
6. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol. 2007;3(1):88.
7. Sharan R, Ideker T. Modeling cellular machinery through biological network comparison. Nat Biotechnol. 2006;24(4):427–33.
8. Futschik ME, Chaurasia G, Herzel H. Comparison of human protein–protein interaction maps. Bioinformatics. 2007;23(5):605–11.
9. Luonan C, Rui-Sheng W, Xiang-Sun Z. Biomolecular networks methods and applications in systems biology. USA: John Wiley and Sons, Inc; 2009.
10. Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, et al. An experimentally derived confidence score for binary protein-protein interactions. Nat Methods. 2009;6(1):91–7.
11. Ngounou Wetie AG, Sokolowska I, Woods AG, Roy U, Deinhardt K, Darie CC. Protein-protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. Cellular and Molecular Life Sciences: CMLS. 2013.
12. Junker BH, Schreiber F. Analysis of biological networks. USA: Wiley; 2007.
13. Yu D, Kim M, Xiao G, Hwang TH. Review of biological network data and its applications. Genomics Inform. 2013;11(4):200–10.
14. Vespignani A. Evolution thinks modular. Nat Genet. 2003;35(2):118–9.
15. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402(6761 Suppl):C47–52.
16. Pizzuti C, Rombo SE. Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods. Bioinformatics. 2014;30(10):1343–52.
17. Blatt M, Wiseman S, Domany E. Superparamagnetic clustering of data. Phys Rev Lett. 1996;76(18):3251–4.
18. Hartuv E, Shamir R. A clustering algorithm based on graph connectivity. Inf Process Lett. 2000;76(4-6):175–81.
19. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30(7):1575–84.

20. King AD, Przulj N, Jurisica I. Protein complex prediction via cost-based clustering. Bioinformatics. 2004;20(17):3013–20.
21. Arnau V, Mars S, Marín I. Iterative cluster analysis of protein interaction data. Bioinformatics. 2005;21(3):364–78.
22. Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol. 2003;21(6):697–700.
23. Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinformatics. 2006;7:207.
24. Rives AW, Galitski T. Modular organization of cellular networks. Proc Natl Acad Sci U S A. 2003;100(3):1128–33.
25. Maciag K, Altschuler SJ, Slack MD, Krogan NJ, Emili A, Greenblatt JF, et al. Systems-level analyses identify extensive coupling among gene expression machines. Mol Syst Biol. 2006;2:2006.0003.
26. Samanta MP, Liang S. Predicting protein functions from redundancies in large-scale protein interaction networks. Proc Natl Acad Sci U S A. 2003;100(22):12579–83.
27. Zhou H. Network landscape from a Brownian particle's perspective. Phys Rev E. 2003;67:041908.
28. Bader GD, Betel D, Hogue CWV. Bind: the biomolecular interaction network database. Nucleic Acids Res. 2003;31(1):248–50.
29. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci U S A. 2003;100(21):12123–8.
30. Dunn R, Dudbridge F, Sanderson CM. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC Bioinformatics. 2005;6:39.
31. Pereira-Leal JB, Enright AJ, Ouzounis CA. Detection of functional modules from protein interaction networks. Proteins. 2004;54(1):49–57.
32. Brohée S, van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics. 2006;7:488.
33. Mewes HW, Ruepp A, Theis F, Rattei T, Walter M, Frishman D, et al. MIPS: curated databases and comprehensive secondary data resources in 2010. Nucleic Acids Res. 2011;39(Database issue):D220–4.
34. Mewes HW, Frishman D, Güldener U, Mannhaupt G, Mayer K, Mokrejs M, et al. MIPS: a database for genomes and protein sequences. Nucleic Acids Res. 2002;30(1):31–4.
35. Mewes HW, Dietmann S, Frishman D, Gregory R, Mannhaupt G, Mayer KFX, et al. MIPS: analysis and annotation of genome information in 2007. Nucleic Acids Res. 2008;36(Database issue):D196–201.
36. Wang Z, Zhang J. In search of the biological significance of modular structures in protein networks. PLoS Comput Biol. 2007;3(6):e107.
37. Phan HTT, Sternberg MJE. PINALOG: a novel approach to align protein interaction networks–implications for complex detection and function prediction. Bioinformatics. 2012;28(9):1239–45.
38. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Gene. 2000;25(1):25–9.
39. Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, Martínez-Cruz LA, et al. Correlation between gene expression and GO semantic similarity. IEEE ACM Trans Comput Biol Bioinformatics. 2005;2(4):330–8.
40. Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics. 2003;19(10):1275–83.
41. du Plessis L, Skunca N, Dessimoz C. The what, where, how and why of gene ontology–a primer for bioinformaticians. Brief Bioinform. 2011;12(6):723–35.
42. Cho YR, Hwang W, Ramanathan M, Zhang A. Semantic integration to identify overlapping functional modules in protein interaction networks. BMC Bioinformatics. 2007;8:265.
43. Dotan-Cohen D, Letovsky S, Melkman AA, Kasif S. Biological process linkage networks. PLoS One. 2009;4(4):e5313.
44. Luciani D, Bazzoni G. From networks of protein interactions to networks of functional dependencies. BMC Syst Biol. 2012;6(1):44.
45. Khatri P, Drăghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics. 2005;21(18):3587–95.
46. Jafari M, Sadeghi M, Mirzaie M, Marashi SA, Rezaei-Tavirani M. Evolutionary conserved motifs and modules in mitochondrial protein interaction network. Mitochondrion. 2013;13(6):668.
47. Guimerà R, Nunes Amaral LA. Functional cartography of complex metabolic networks. Nature. 2005;433(7028):895–900.
48. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theor Exp. 2008;10:P10008.
49. Lambiotte R, Ausloos M, Hołyst JA. Majority model on a network with communities. Phys Rev E. 2007;75(3):030101.
50. Jafari M, Primo V, Smejkal GB, Moskovets EV, Kuo WP, Ivanov AR. Comparison of in-gel protein separation techniques commonly used for fractionation in mass spectrometry-based proteomic profiling. Electrophoresis. 2012;33(16):2516–26.
51. Chen L, Wang RS, Zhang XS. Biomolecular networks: methods and applications in systems biology: Wiley. 2009.
52. Zhang A. Protein interaction networks: computational analysis: Cambridge University Press. 2009.
53. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 2003;302(5644):449–53.
54. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a survey. IEEE Trans Knowl Data Eng. 2004;16(11):1370–86.
55. Campello RJGB. A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. Pattern Recogn Lett. 2007;28(7):833–41.
56. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. 2013.
57. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
58. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014;42(D1):472–7.
59. Tieri P, Nardini C. Signalling pathway database usability: lessons learned. Mol BioSyst. 2013;9(10):2401–7.
60. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005;33(Database issue):D433–7.
61. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 2011;39(Database issue):D561–8.
62. Dobzhansky T. Nothing in biology makes sense except in the light of evolution. Am Biol Teach. 1973;35:125–9.
63. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res. 2006;34(Database issue):D187.