Constructing and working with protein interaction graphs

Charalampos Moschopoulos[1, 2], Seferina Mavroudi[1], Spiridon Likothanassis[1] and Sophia Kossida[2]

[1] Department of Computer Engineering & Informatics, University of Patras, Rio, Patras, Greece

[2] Bioinformatics & Medical Informatics Team, Biomedical Research Foundation of the Academy of Athens, Athens, Greece

Abstract

Nowadays, the detection of protein – protein interactions is performed mainly from experimental methods that are able to record thousands of them in a single experiment. In this contribution, we present methods that are used to assess the reliability of protein -protein interactions and construct reliable protein interaction graphs. Furthermore, we present the most popular algorithms that are used to detect protein complexes or discover the functionality of unknown proteins through clustering protein interaction graphs. Finally, the most popular visualization software tools are described.

Introduction

Recent development of high-throughput methods produced enormous datasets of protein – protein interaction (PPI) data. Techniques such as yeast two – hybrid and mass spectrometry detect PPIs and give an insight of the cellular organization of an organism. However, these methods suffer from high error rate as they miss an important fraction of protein interactions and yield several protein interactions that do not exist in reality.

Due to the large number of interactions, there is a great need of computational methods and models that would make it easy to extract valuable information from them. Usually, through protein interaction data, information derives about functional modules such as protein complexes (which reveal insights into both the topological properties and functional organization of protein networks) as well as the function of uncategorized proteins. A very efficient way of summarizing these new datasets is by forming protein interaction graphs. These graphs provide a valuable tool that helps the better understanding of the functional organization of the proteome. A graph is represented as G = (V,E), where V is the set of the graph vertices and E is the set of the graph edges. In a protein interaction graph, the vertices represent the proteins and the edges the pairwise interactions between two proteins. Unfortunately, because of the unreliability of protein interaction data, algorithmic methods applied on protein interaction graphs can not produce results of high information value.

In this contribution, we present the most popular methods to assess the reliability of protein-protein interactions. Furthermore, we describe the features of a protein interaction graph and the computational methods which are used to acquire valuable conclusions from them. Additionally, this manuscript presents the most popular open source software tools which visualize PPI data.

Assessing the reliability of protein – protein interactions

Although data sets on the protein interactome, obtained by high-throughput protein interaction assays, are being accumulated rapidly, they usually come at the expense of relatively low quality, containing a high rate of spurious (false positives) and missing (false negatives) protein-protein interactions [1].

To address the problem of false positives, different confidence scores have been assigned that reflect the reliability and biological significance of each protein interaction pair derived from the experiments [2]. Confidence scores are often computed as single indices, correlating interaction pairs derived from direct experiments (e.g., two-hybrid screens and mass spectrometry) with either indirect biological data sources (e.g., gene expression, protein - DNA binding, biological function, biological process, protein localization, protein class), or sequence based data sources (domain information, gene fusion, etc.). More recently, they are derived from supervised learning methods, which are employed to integrate direct and indirect biological data sources for the prediction task. The training data sets for these methods include known true positives and true negative interactions. For both strategies different approaches have been proposed, where the data sources varied along with the implementations.

Specifically, indices have been based on the sharing of a common cellular localization or a common cellular role [3, 4]. Alternatively, ranking of the reliability of protein interactions have been based on the reproducibility and non-randomness of the observation of an interaction [5-7]. Related to the ideas of functional homogeneity, localization coherence and observational reproducibility are a large number of other approaches based on the use of additional information, such as protein annotation, or the use of information from multiple assays [8-12]. Interaction network topology is a different mean of identifying reliability of interactions relying solely on the topology of the neighborhood of an interacting pair of proteins in the interactome [13, 14].

Bayes classifiers [15] and Bayesian Networks that combine multiple data sources are among the promising machine learning schemes that have been employed to predict true and false protein-protein interactions. Nevertheless, in [16] article, a Bayes classifier has been compared to Random Forest (RF) and Logistic Regression (LR), showing the RF classifier to have the best performance among them. In a more extended comparison including a Random Forest (RF), a RF similarity based k-Nearest-Neighbor classifier, Na□ve Bayes, Decision Tree, Logistic Regression and Support Vector Machine, the superior performance of RF was confirmed along with a satisfying performance of the Suppert Vector Machines (SVMs) [17]. Alternatively, a variant kernel canonical correlation analysis, has been used for predicting pathway protein interactions [18], while in [19] a sum of likelihood ratio scores strategy was explored to predict human PPI confidence.

Even though most of the above approaches are hindered or limited due inherent difficulties (eg., not all model organisms have well annotated genomes, expression of interacting proteins may need not to be correlated over many conditions and conversely protein pairs with correlation patterns do not necessarily physically interact, the number of proteins having known paralogs is limited as well as the number of available structures, etc), nevertheless, most studies suggest that utilizing any of the confidence assignment schemes is always more beneficial than assuming all observed interactions to be true or equally likely.

The problem of false negatives is essentially related to the problem of the ab initio prediction of protein-protein interactions by computational methods. Well known methods rely on gene fusion events [20-22], interacting domains [23, 24], interacting motifs [25-27], co-evolution of proteins or residues [28-30] and the topology of protein–protein interaction networks [31, 32]. Alternatively association rules have been explored [33]. In a different approach some of the confidence scores initially designed for the reliability assignment of observed interactions, are used for the assigning probability scores to putative unobserved interactions pairs.

Protein interaction Graphs

A protein interaction graph can be weighted or unweighted. In a weighted one, each edge connecting two proteins has been characterized by a number that represents the validity of the connection between these two proteins. In an unweighted protein interaction graph, an assumption is made that this number is equal to 1 for all the edges of the graph.

Generally, the protein interaction graphs are undirected and unweighted. Some properties have been identified to be common between the protein interaction graphs of all the organisms. First of all, they are all scale free. Moreover, it is proved that similar proteins usually interact with each other and that they lie within short distance in the interaction graph. Finally, there are few vertices having many interactions and many that have few interactions. This means that if some proteins are eliminated, the topology of the protein interaction graph does not change which subsequently confirms the robustness of the organisms as they can afford to loose some proteins without jeopardizing the existence or even the normal function of the network.

In protein interaction graphs, the dense subgraphs are valuable since they provide details concerning the functionality of the proteins within the subgraph and the consistency of protein complexes. Given the mathematical representation of a graph, algorithms derived from the graph theory are well suited in order to isolate these dense areas.

The amount of data that has been derived from high-throughput approaches, automated text mining techniques, and/or manually from the scientific literature, has been stored in databases called protein-protein interaction databases. These databases are valuable resources for the researchers, where from they can easily retrieve and analyze the stored data [34]. Usually these databases include data of protein interactions obtained from many organisms. The most popular ones are BIND [35], MIPS [36], UniProt [37], IntAct [38].

It must be noted that there is a significant difference in the total number of protein-protein interactions among the various protein-protein interaction databases [39], due to the fact that data for each database were derived using different methods. Apart from the databases, where data obtained from experimental methods are stored, there are some other databases, where protein interactions predicted by computational methods are stored. The most significant one is called STRING database which has integrated known and predicted interactions from a variety of sources as well [40].

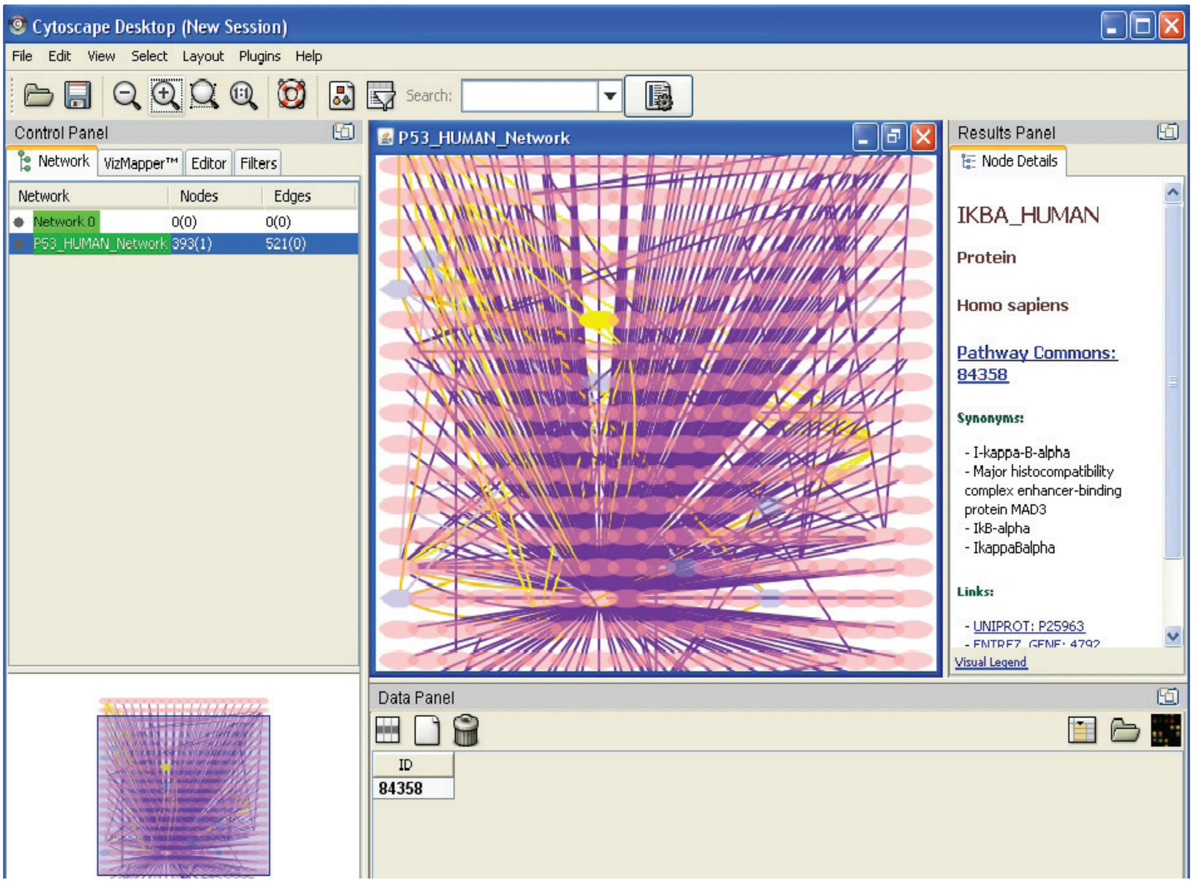
Extracting information from protein interaction graphs

Protein interaction graphs are used mainly to detect protein complexes in which individual proteins assemble into functional modules [41] or elucidate the function of uncharacterized proteins.

Various algorithms have been applied for the identification of protein complexes through protein – protein interaction networks. They can be divided in two big categories: those using a local search strategy and those using a hierarchical one. In the first category, the first introduced algorithm in the field was the Molecular Complex Detection (Mcode) [42]. A year before the appearance of Mcode, Enright et al. had introduced an algorithm called TRIBE-MCL [43] based on the Markov cluster algorithm (MCL), a previously developed algorithm for graph clustering. Besides that, King et al. suggested the RNSC algorithm [44] a cost-based local search algorithm. These two algorithms separate the whole protein interaction graph into clusters that represent protein families. This means that not even a single protein is discarded from the final results and several clusters can not be considered as protein complexes. Nevertheless, the RNSC algorithm uses a filtering strategy to achieve the identification of protein complex candidates. Another algorithm of the local search approach is the Local Clique Merging Algorithm (LCMA) [45] which locates local cliques in an interaction graph and subsequently tries to expand them.

On the other hand, all the hierarchical clustering algorithms are based on the concept of dividing the initial graph by removing the minimum set of edges. The Highly Connected Subgraph method (HCS) [46] separates a graph into several subgraphs using minimum cuts and stops when the cut is bigger or equal to the number of the graph vertices divided by 2. Koyutürk suggested the SIdeS algorithm [47] which uses the HCS algorithm philosophy; however the stopping criterion is based on the statistical significance of the derived subgraphs. More specifically, the SideS algorithm uses a framework for analyzing the occurrence of dense patterns in randomly generated graph-structured data with a view to assessing the significance of a pattern based on the statistical relationship between subgraph density and size.

The algorithms applied for the identification of protein complexes can be used to functionally annotate proteins. As it is presented in [48], identifying protein modules helps annotating uncharacterized proteins using the function shared by the majority of the module's proteins. However, these methods are outperformed by more direct "methods" which infer the function of a protein


Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help
Search:
Control Panel
Network VizMapper™ Editor Filters
P53_HUMAN_Network
Results Panel
Node Details
Visual Legend
Data Panel

Network	Nodes	Edges
Network 0	0(0)	0(0)
P53_HUMAN_Network	393(1)	521(0)

IKBA_HUMAN

Protein

Homo sapiens

Pathway Commons: 84358

Synonyms:

- I-kappa-B-alpha
- Major histocompatibility complex enhancer-binding protein MAD3
- IkB-alpha
- IkappaBalpha

Links:

- UNIPROT: P25963
- ENTREZ_GENE: 4792

ID
84358

Figure 1. Snapshot of Cytoscape tool.

based on its connections in the network. Based on the principle that proteins that lie closer to one another in the protein – protein interaction graph are more likely to have similar function, they are simpler and more effective from the clustering approaches. They can be divided in those using the neighborhood of a protein [6], the approaches which are graph theoretic [49], the probabilistic ones [50] and those that integrate multiple data sources [51].

Tools for PPI graphs visualization

Availability of large scale experimental data and numerous approaches of extracting information from PPI graphs enable the development of many software tools. The visualization of the vast volume of PPI data allows the observation of the whole proteome of an organism [52].

Among those tools being freely available for academic use, the most popular visualization tool is Cytoscape [53] in which a user can construct his own graph or import PPI graphs from online databases. Additionally, Cytoscape includes a flexible plugin architecture that enables developers to add extra functionality beyond that provided in the core. Another visualization tool is Medusa [54] which is based on the Fruchterman – Reingold algorithm [55]. However, it is less suited for the visualization of big datasets and its own text file format is not compatible with other visualization tools. 2D and 3D representations are offered by BioLayout Express 3D tool [56]. This tool is highly interactive and in the latest version, the MCL algorithm is hosted in this tool. Other visualization tools are VisANT [57] and PIVOT [58] which are best suited for visualizing protein –protein interactions and identifying relationships between them.

Future perspectives and conlusions

Graph – based model can exploit global and local characteristics of biology and more particularly PPI graphs. Various algorithmic methods and tools have been developed in order to extract information using graph theoretic approaches.

Although the above methods of assessing the protein interactions reliability are useful and some of them exhibit encouraging results, there is still room for improvement. None of the existing methods gain both a high specificity and a good sensitivity at the same time. Data integration usually improves the results. However, different biological sources represent different and apparently biased subsets of the true interactions and simply taking the union may lead to poor performance, while taking the intersection may result in a minimal overlap. New methods, able to cope with partial domain knowledge would be desirable. Supplementary to data integration, model integration could further enhance performance. Besides of accuracy issues, newly designed methods should allow for the interpretability of the results. There are different and often contradicting opinions regarding the biological evidence that should be taken into account for the computation and the evaluation of the reliability of protein-protein interactions. Researchers should be given the means to judge each feature's contribution and to extract new explainable knowledge.

Another future aspect would be the use of heterogeneous source of data to construct weighted graphs. This way, the above mentioned methods can offer better quality of information, while the results of PPI graph analysis would suffer by less error rate. Furthermore, working in the same direction, the variety of information that derives from PPI graph analysis could be retrieved by using web services tools. Web services enable programmers to build complex applications without the need to install and maintain the databases and analysis tools and without having to take on the financial overheads that accompany these. Moreover, Web services provide easier integration and interoperability among applications and the data they require. Finally, it would be interesting to apply these methods in trancsriptomics or metabolomics, sections of research that are still in their infancy.

References

1 von Mering C, Krause R, Snel B, Cornell M, Oliver S G, Fields S,Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417: (6887) 399-403.

2 Yu J,Finley R L, Jr. (2009) Combining multiple positive training sets to generate confidence scores for protein-protein interactions. Bioinformatics 25: (1) 105-11.

3 Sprinzak E, Sattath S,Margalit H (2003) How reliable are experimental protein-protein interaction data? J Mol Biol 327: (5) 919-23.

4 Chen J, Chua H N, Hsu W, Lee M L, Ng S K, Saito R, Sung W K,Wong L (2006) Increasing confidence of protein-protein interactomes. Genome Inform 17: (2) 284-97.

5 Nabieva E, Jim K, Agarwal A, Chazelle B,Singh M (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 21 Suppl 1: i302-10.

6 Chua H N, Sung W K,Wong L (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics 22: (13) 1623-30.

7 Hart G T, Lee I,Marcotte E R (2007) A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics 8: 236.

8 Bader J S, Chaudhuri A, Rothberg J M,Chant J (2004) Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol 22: (1) 78-85.

9 Patil A,Nakamura H (2005) Filtering high-throughput protein-protein interaction data using a combination of genomic features. BMC Bioinformatics 6: 100.

10 Giot L, Bader J S, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao Y L, Ooi C E, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon C A, Finley R L, Jr., White K P, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets R A, McKenna M P, Chant J,Rothberg J M (2003) A protein interaction map of Drosophila melanogaster. Science 302: (5651) 1727-36.

11 Samanta M P,Liang S (2003) Predicting protein functions from redundancies in large-scale protein interaction networks. Proc Natl Acad Sci U S A 100: (22) 12579-83.

12 Martin S, Roe D,Faulon J L (2005) Predicting protein-protein interactions using signature products. Bioinformatics 21: (2) 218-26.

13 Saito R, Suzuki H,Hayashizaki Y (2003) Construction of reliable protein-protein interaction networks with a new interaction generality measure. Bioinformatics 19: (6) 756-63.

14 Chen J, Hsu W, Lee M L,Ng S K (2006) Increasing confidence of protein interactomes using network

topological metrics. Bioinformatics 22: (16) 1998-2004.

15 Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan N J, Chung S, Emili A, Snyder M, Greenblatt J F,Gerstein M (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302: (5644) 449-53.

16 Lin N, Wu B, Jansen R, Gerstein M,Zhao H (2004) Information assessment on predicting protein-protein interactions. BMC Bioinformatics 5: 154.

17 Qi Y, Bar-Joseph Z,Klein-Seetharaman J (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins 63: (3) 490-500.

18 Yamanishi Y, Vert J P,Kanehisa M (2004) Protein network inference from multiple genomic data: a supervised approach. Bioinformatics 20 Suppl 1: i363-70.

19 Rhodes D R, Tomlins S A, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A,Chinnaiyan A M (2005) Probabilistic model of the human protein-protein interaction network. Nat Biotechnol 23: (8) 951-9.

20 Marcotte E M, Pellegrini M, Ng H L, Rice D W, Yeates T O,Eisenberg D (1999) Detecting protein function and protein-protein interactions from genome sequences. Science 285: (5428) 751-3.

21 Enright A J, Iliopoulos I, Kyrpides N C,Ouzounis C A (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402: (6757) 86-90.

22 Tsoka S,Ouzounis C A (2000) Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. Nat Genet 26: (2) 141-2.

23 Sprinzak E,Margalit H (2001) Correlated sequence-signatures as markers of protein-protein interaction. J Mol Biol 311: (4) 681-92.

24 Han D S, Kim H S, Jang W H, Lee S D,Suh J K (2004) PreSPI: a domain combination based prediction system for protein-protein interaction. Nucleic Acids Res 32: (21) 6312-20.

25 Tong A H, Drees B, Nardelli G, Bader G D, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S, Quondam M, Zucconi A, Hogue C W, Fields S, Boone C,Cesareni G (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295: (5553) 321-4.

26 Aytuna A S, Gursoy A,Keskin O (2005) Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. Bioinformatics 21: (12) 2850-5.

27 Li H, Li J,Wong L (2006) Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. Bioinformatics 22: (8) 989-96.

28 Pazos F,Valencia A (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng 14: (9) 609-14.

29 Pazos F, Ranea J A, Juan D,Sternberg M J (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J Mol Biol 352: (4) 1002-15.

30 Juan D, Pazos F,Valencia A (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc Natl Acad Sci U S A 105: (3) 934-9.

31 Yu H, Paccanaro A, Trifonov V,Gerstein M (2006) Predicting interactions in protein networks by completing defective cliques. Bioinformatics 22: (7) 823-9.

32 Pei P,Zhang A (2005) A topological measurement for weighted protein interaction network. Proc IEEE Comput Syst Bioinform Conf 268-78.

33 Kotlyar M,Jurisica I (2006) Predicting Protein-Protein Interactions by Association Mining. Information Systems Frontiers 8: (1) 37-47.

34 Suresh S, Sujatha Mohan S, Mishra G, Hanumanthu G R, Suresh M, Reddy R,Pandey A (2005) Proteomic resources: integrating biomedical information in humans. Gene 364: 13-8.

35 Bader G D, Betel D,Hogue C W (2003) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res 31: (1) 248-50.

36 Mewes H W, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S,Weil B (2002) MIPS: a database for genomes and protein sequences. Nucleic Acids Res 30: (1) 31-4.

37 Consortium U (2009) The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res 37: (Database issue) D169-74.

38 Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R,Hermjakob H (2007) IntAct--open source resource for molecular interaction data. Nucleic Acids Res 35: (Database issue) D561-5.

39 Mathivanan S, Periaswamy B, Gandhi T K, Kandasamy K, Suresh S, Mohmood R, Ramachandra Y L,Pandey A (2006) An evaluation of human protein-protein interaction data in the public domain. BMC Bioinformatics 7 Suppl 5: S19.

40 von Mering C, Jensen L J, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B,Bork P (2007) STRING 7--recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35: (Database issue) D358-62.

41 Hartwell L H, Hopfield J J, Leibler S,Murray A W (1999) From molecular to modular cell biology. Nature 402: (6761 Suppl) C47-52.

42 Bader G D,Hogue C W (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4: 2.

43 Enright A J, Van Dongen S,Ouzounis C A (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: (7) 1575-84.

44 King A D, Przulj N,Jurisica I (2004) Protein complex prediction via cost-based clustering. Bioinformatics 20: (17) 3013-20.

45 Li X-L, Tan S-H, Foo C-S,Ng S-K (2005) Interaction Graph Mining for Protein Complexes Using Local Clique Merging. Genome Informatics 16: (2) 260-269.

46 Hartuv E,Shamir R (2000) A clustering algorithm based on graph connectivity Information Processing Letters 76: (4-6) 175-181.

47 Koyuturk M, Szpankowski W,Grama A (2007) Assessing significance of connectivity and conservation in protein interaction networks. J Comput Biol 14: (6) 747-64.

48 Sharan R, Ulitsky I,Shamir R (2007) Network-based prediction of protein function. Mol Syst Biol 3: 88.

49 Vazquez A, Flammini A, Maritan A,Vespignani A (2003) Global protein function prediction from protein-protein interaction networks. Nat Biotechnol 21: (6) 697-700.

50 Letovsky S,Kasif S (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 19 Suppl 1: i197-204.

51 Tsuda K, Shin H,Scholkopf B (2005) Fast protein classification with multiple networks. Bioinformatics 21 Suppl 2: ii59-65.

52 Pavlopoulos G A G, Wegener A L A,Schneider R R (2008) A survey of visualization tools for biological network analysis. BioData Min 1: (1) 12.

53 Shannon P, Markiel A, Ozier O, Baliga N S, Wang J T, Ramage D, Amin N, Schwikowski B,Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: (11) 2498-504.

54 Hooper S D,Bork P (2005) Medusa: a simple tool for interaction graph analysis. Bioinformatics 21: (24) 4432-3.

55 Fruchterman T, M. J.,Reingold E, M. (1991) Graph drawing by force-directed placement. Softw. Pract. Exper. 21: (11) 1129-1164.

56 Freeman T C, Goldovsky L, Brosch M, van Dongen S, Maziere P, Grocock R J, Freilich S, Thornton J,Enright A J (2007) Construction, visualisation, and clustering of transcription networks from microarray expression data. PLoS Comput Biol 3: (10) 2032-42.

57 Hu Z, Snitkin E S,DeLisi C (2008) VisANT: an integrative framework for networks in systems biology. Brief Bioinform 9: (4) 317-25.

58 Orlev N, Shamir R,Shiloh Y (2004) PIVOT: protein interacions visualizatiOn tool. Bioinformatics 20: (3) 424-5.