

# PINA v2.0: mining interactome modules

Mark J. Cowley<sup>1,2</sup>, Mark Pinese<sup>1</sup>, Karin S. Kassahn<sup>3</sup>, Nic Waddell<sup>3</sup>, John V. Pearson<sup>3</sup>, Sean M. Grimmond<sup>3</sup>, Andrew V. Biankin<sup>1</sup>, Sampsa Hautaniemi<sup>4</sup> and Jianmin Wu<sup>1,2,\*</sup>

<sup>1</sup>Cancer Research Program, <sup>2</sup>Peter Wills Bioinformatics Centre, Garvan Institute of Medical Research, Darlinghurst, Sydney NSW 2010, <sup>3</sup>Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, Brisbane QLD 4072, Australia and <sup>4</sup>Genome-Scale Biology Program, Institute of Biomedicine, University of Helsinki, Helsinki 00014, Finland

Received August 15, 2011; Revised October 10, 2011; Accepted October 16, 2011

## ABSTRACT

The Protein Interaction Network Analysis (PINA) platform is a comprehensive web resource, which includes a database of unified protein–protein interaction data integrated from six manually curated public databases, and a set of built-in tools for network construction, filtering, analysis and visualization. The second version of PINA enhances its utility for studies of protein interactions at a network level, by including multiple collections of interaction modules identified by different clustering approaches from the whole network of protein interactions ('interactome') for six model organisms. All identified modules are fully annotated by enriched Gene Ontology terms, KEGG pathways, Pfam domains and the chemical and genetic perturbations collection from MSigDB. Moreover, a new tool is provided for module enrichment analysis in addition to simple query function. The interactome data are also available on the web site for further bioinformatics analysis. PINA is freely accessible at <http://cbg.garvan.unsw.edu.au/pina/>.

## INTRODUCTION

Protein–protein interactions (PPIs) mediate biological function and play a pivotal role in many cellular processes. Different small- and large-scale experimental approaches generate ever-increasing amounts of publicly accessible data. Given the availability of vast amounts of PPI data, analysis of PPI networks has become a major challenge and considerable efforts have been undertaken.

A common type of analysis focuses on the whole network of protein interactions for a given species ('interactome') (1). A number of studies have shown that interactomes follow a power-law degree distribution, exhibit small world behavior and tend to be modular (2,3). Identification of sub-networks with special

characteristics using graphical approaches can also lead to biologically relevant insights. It is well established that densely interconnected regions of a global PPI network often correspond to functionally related groups of genes/proteins that can be identified as modules (4). Understanding how these modules are organized can lead to a better understanding of how cellular processes are coordinated in normal cells and perturbed under pathological conditions. Several efforts have been undertaken to identify modules, which might represent protein complexes or signaling pathways, from interactome networks (5–10). However, there is no unified resource for biologists to interrogate these interactome modules extracted from a regularly updated PPI database with extensive functional annotations and advanced network analysis tools.

In PINA v2.0, we generated the interactome data for six model organisms based on the existing PINA PPI integration database and applied different clustering algorithms to identify collections of modules. To improve biological interpretation, the identified modules have been comprehensively annotated by different knowledge databases. Both modules and annotations were saved in PINA v2.0 database and an advanced tool was developed for module enrichment analysis in addition to a simple query form. These new data and tools have been seamlessly integrated and can co-operate with the existing resources in PINA, which together provides a unique portal for biologists to better understand their genes of interest in the context of a PPI network.

## INTERACTOME DATASET

The first version of PINA (11) has established a non-redundant PPI database, updated quarterly based on the integration of protein interaction data from six publicly available, manually curated databases: IntAct (12), MINT (13), BioGRID (14), DIP (15), HPRD (16) and MIPS MPact (17). We exported interactome data from the PINA PPI integration database in PSI-MI

\*To whom correspondence should be addressed. Tel: +61 2 9295 8326; Fax: +61 2 9295 8321; Email: j.wu@garvan.org.au

(Proteomics Standards Initiative Molecular Interactions) tab-delimited data exchange format (18) for six model organisms (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*). Each exported interactome file includes self and binary interactions with one interaction per row. The UniProt accession number was used as the protein identifier. These files will be updated concurrently with each new release of the PINA integration database and can be freely downloaded from the PINA website for further bioinformatics analysis.

## INTERACTOME MODULES

### Module identification from interactomes

Several algorithms have been developed to identify highly interconnected groups of nodes within a network (5–10). These algorithms are mostly either agglomerative ('bottom-up') or divisive ('top-down'). We selected molecular complex detection (MCODE) method (6) and Markov clustering (MCL) method (5) as representatives from each category (19), and applied them to the interactomes from each of the six species. We selected a range of parameter settings (Supplementary Table S1) to control the properties of the resulting modules, from small and densely interconnected (protein-complex-like), to large and loosely interconnected (pathway-like). From a total of 30 analysis runs, we detected approximately 2400 modules containing at least five proteins. Modules identified from each run were saved as a module collection in the PINA module database. End users can select which collection to be used in the query or the enrichment analysis depending on whether they are looking for protein-complex-like modules, or pathway-like modules, with advice given on the PINA website. As the body of PPI data accrues over time, the interactomes will become more complete, and thus some modules identified from an interactome may change. To facilitate historical comparisons, we will timestamp each module collection and retain the last five releases.

### Module annotation and visualization

Following module identification, we annotated each module by looking for enriched terms from multiple functional databases including Gene Ontology (20), KEGG pathways (21), Pfam domains (22) and the chemical and genetic perturbations collection from MSigDB (23). Since modules often show strong functional coherence (24), the diverse set of annotations provide a complementary overview of module function. The back-end module annotation tool uses a hypergeometric test to identify the overrepresented terms, with a correction for multiple testing using false discovery rate (25). For each module, we stored at least the 10 most significant terms, and any other significant terms (adjusted *P*-value < 0.05) in the PINA annotation database. Based on approximately 25.7 million comparisons, there are approximately 270,000 significant terms saved in the PINA annotation database.

A thumbnail image is available for each module (Figure 1b), which offers a quick impression of the module's topology. Users can also launch our previously developed visualization tool to interactively visualize and manipulate the selected module. Since each module can be treated as a network, other existing PINA tools can be applied to filter and analyze the selected module, through web pages or the visualization tool.

### Module query and enrichment analysis tool

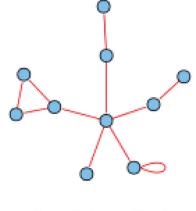
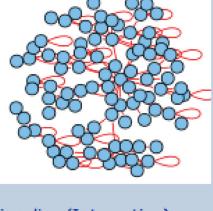
There are two ways to make use of the interactome modules in PINA. Users can either perform a simple search to find modules which have at least one protein from their query proteins or use the newly developed module enrichment tool to identify statistically enriched modules. The module enrichment tool compares a list of proteins from a user query with all the modules in a specified collection, by using a hypergeometric test to identify modules that are overrepresented in query proteins relative to the background frequency in either the interactome or the whole proteome. Fig. 1 shows the module enrichment result of a set of proteins, which contain non-synonymous coding single nucleotide variations (SNV) in two primary pancreatic adenocarcinoma tumors (APGI-1959 and APGI-1992) and one pancreatic cancer cell line (CRL-2557 Panc-05.04). These mutated genes (Supplementary Table S2) were detected by next-generation sequencing and downloaded from the International Cancer Genome Consortium (ICGC) (26) data portal (<http://dcc.icgc.org>; Pancreatic cancer AU project). The annotation summary indicates that the top module may play an important role in cancer through its influence on the cell's transcriptional machinery.

## IMPROVED USER ACCESS

In the first version of PINA, protein annotations were fetched on the fly through the UniProt web service, which was slow for construction of a large PPI network consisting of hundreds of proteins. In PINA v2.0, we have saved the UniProt annotations into the PINA annotation database, which has significantly improved the query speed for large networks. The PINA web services were also updated for easier use and quicker response, by adopting a lightweight RESTful web service, as opposed to the previously used SOAP service. In addition, PINA has been wrapped as a component of the Anduril framework (27), which is a component-based workflow framework for large-scale biological data analysis. The PINA component can be executed as either standalone or as one step of a complex workflow analyzing high-throughput screening data, such as SNP, gene expression or exon microarray, which can start from preprocessing and normalization of raw data to functional annotation of the identified genes/proteins.

## IMPLEMENTATION

*AllegroMCODE* 1.0 (<http://www.allegroviva.com/allegromcode>), which is a GPU-enabled *Cytoscape* 2.8.1

Module ID	Module Thumbnail	Annotation Summary	Sample Freq	Background Freq	p-value	adjusted p-value
PINAM00065	<p>(a) <a href="#">View annotation details</a></p> <p>(b) </p> <p><a href="#">Visualize (Interactive)</a></p>	<p><b>GO (Biological Process)</b> transcription elongation from RNA polymerase III promoter <a href="#">GO:0006385</a> termination of RNA polymerase III transcription <a href="#">GO:0006386</a> transcription elongation, DNA-dependent <a href="#">GO:0006354</a></p> <p><b>GO (Cellular Component)</b> DNA-directed RNA polymerase III complex <a href="#">GO:0005666</a> (c) DNA-directed RNA polymerase complex <a href="#">GO:0000428</a> nuclear DNA-directed RNA polymerase complex <a href="#">GO:0055029</a></p> <p><b>GO (Molecular Function)</b> DNA-directed RNA polymerase activity <a href="#">GO:0003899</a> RNA polymerase activity <a href="#">GO:0034062</a> nucleotidyltransferase activity <a href="#">GO:0016779</a></p> <p><b>KEGG</b> RNA polymerase <a href="#">KEGG:03020</a> Cytosolic DNA-sensing pathway <a href="#">KEGG:04623</a> Pyrimidine metabolism <a href="#">KEGG:00240</a></p> <p><b>PFAM</b> zf-C2HC <a href="#">PFAM:PF01530</a> KOW <a href="#">PFAM:PF00467</a> SCAN PFAM:PF02023</p> <p><b>MSigDB</b> <a href="#">KOYAMA_SEMA3B TARGETS_UP</a> <a href="#">NIKOLSKY BREAST CANCER_10Q22_AMPLICON</a> <a href="#">SUZUKI AMPLIFIED IN ORAL CANCER</a></p>				
PINAM00032	<p><a href="#">View annotation details</a></p> <p><a href="#">View interaction details</a></p> <p></p> <p><a href="#">Visualize (Interactive)</a></p>	<p><b>GO (Biological Process)</b> regulation of cell communication <a href="#">GO:0010646</a> regulation of MAPKKK cascade <a href="#">GO:0043408</a> regulation of cellular process <a href="#">GO:0050794</a></p> <p><b>GO (Cellular Component)</b> intracellular part <a href="#">GO:0044424</a> intracellular organelle <a href="#">GO:0043229</a> organelle <a href="#">GO:0043226</a></p> <p><b>GO (Molecular Function)</b> protein binding <a href="#">GO:0005515</a> MAP-kinase scaffold activity <a href="#">GO:0005078</a> protein kinase binding <a href="#">GO:0019901</a></p> <p><b>KEGG</b> mRNA surveillance pathway <a href="#">KEGG:03015</a> Insulin signaling pathway <a href="#">KEGG:04910</a> Spliceosome <a href="#">KEGG:03040</a></p> <p><b>PFAM</b> HLH <a href="#">PFAM:PF00010</a> PID <a href="#">PFAM:PF00640</a> 2-oxoacid_dh <a href="#">PFAM:PF00198</a></p> <p><b>MSigDB</b> <a href="#">BLALOCK_ALZHEIMERS DISEASE_UP</a> <a href="#">MCBRYAN_PUBERTAL BREAST_3_4WK_UP</a> <a href="#">ROY_WOUND_BLOOD VESSEL_UP</a></p>	2 / 122	10 / 42159	0.0004	0.0206

**Figure 1.** An example of the module enrichment result, showing the top two modules. (a) The top link is to the page showing the complete list of functional annotations, while the bottom link is to the page showing the list of protein interactions in the module. (b) The link underlying the thumbnail image will launch the interactive visualization tool. (c) A summary of module function, only showing the top three terms in each functional annotation categories. (d) The left number is the number of query proteins found in this module, while the right number is the total number of query proteins found in the background. (e) The left number is the total number of proteins in the module, while the right number is the total number of proteins in the background. (f) The enrichment P-value is based on a hypergeometric test. (g) The adjusted P-value for multiple hypothesis testing.

(28) plug-in for running *MCODE* (6), and *MCL* v10-201 (5) were used for identifying interactome modules. The specified parameters are listed in *Supplementary Table S1*. The output files of each tool were parsed using custom R scripts and the functional enrichment analyses were performed on an SGE cluster using *GOstats* (29) and a custom extension to the *Category* package from the Bioconductor project v2.8, using

R v2.13.1. The mappings from genes to KEGG, GO and PFAM were from the *AnnotationDbi* package, and the c2.cgp.v3.0.symbols.gmt geneset collection were from MSigDB (23). Module thumbnails were generated using *igraph* (30). The RESTful web services were implemented using a Java library *jersey*, and example code for a Java client is available on the PINA web site.

## FUTURE DIRECTION

In PINA v2.0, we seamlessly integrate interactome modules and the associated functional annotations with the existing PINA resource including the PPI integration database and a set of network-based tools, providing significant new functionalities for researchers looking to analyze PPI data at a network level. We intend to continue this effort and plan to integrate built-in network alignment tools, which will allow the comparison of two networks either generated by user queries, or selected from the interactome modules. In addition, another important model organism *Arabidopsis thaliana* will be added to PINA in the near future.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2.

## ACKNOWLEDGEMENTS

We thank Dr Warren Kaplan and Derrick Lin for their support with high performance computing infrastructure.

## FUNDING

Cancer Council New South Wales, Australia (grant SRP11-01, ICGC 09-01); National Health and Medical Research Council, Australia (grant 631701); Cancer Institute New South Wales, Australia (grant 10/CRF/1-01 to A.V.B); Academy of Finland (grant 125826); Avner Nahmani Pancreatic Cancer Foundation; R. T. Hall Trust. Funding for open access charge: Cancer Council New South Wales, Australia (grant ICGC 09-01).

*Conflict of interest statement.* None declared.

## REFERENCES

1. Cusick,M.E., Klitgord,N., Vidal,M. and Hill,D.E. (2005) Interactome: gateway into systems biology. *Hum. Mol. Genet.*, **14 Spec No. 2**, R171–R181.
2. Maslov,S. and Sneppen,K. (2002) Specificity and stability in topology of protein networks. *Science*, **296**, 910–913.
3. Milo,R., Shen-Orr,S., Itzkovitz,S., Kashtan,N., Chklovskii,D. and Alon,U. (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.
4. Rives,A.W. and Galitski,T. (2003) Modular organization of cellular networks. *Proc. Natl Acad. Sci. USA*, **100**, 1128–1133.
5. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
6. Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.
7. Adamcsek,B., Palla,G., Farkas,I.J., Derenyi,I. and Vicsek,T. (2006) CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, **22**, 1021–1023.
8. Yan,X., Mehan,M.R., Huang,Y., Waterman,M.S., Yu,P.S. and Zhou,X.J. (2007) A graph-based approach to systematically reconstruct human transcriptional regulatory modules. *Bioinformatics*, **23**, i577–i586.
9. Jiang,P. and Singh,M. (2010) SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics*, **26**, 1105–1111.
10. Rhrissorrakrai,K. and Gunsalus,K.C. (2011) MINE: module identification in networks. *BMC Bioinformatics*, **12**, 192.
11. Wu,J., Vallimius,T., Ovaska,K., Westermarck,J., Makela,T.P. and Hautaniemi,S. (2009) Integrated network analysis platform for protein-protein interactions. *Nat. Methods*, **6**, 75–77.
12. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. et al. (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.
13. Ceol,A., Chatr Aryamontri,A., Licata,L., Peluso,D., Brigandt,L., Perfetto,L., Castagnoli,L. and Cesareni,G. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.
14. Stark,C., Breitkreutz,B.J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X. et al. (2011) The BioGRID interaction database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.
15. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
16. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafran,B., Venugopal,A. et al. (2009) Human protein reference database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
17. Guldener,U., Munsterkotter,M., Oesterheld,M., Pagel,P., Ruepp,A., Mewes,H.W. and Stumpflen,V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.*, **34**, D436–D441.
18. Kerrien,S., Orchard,S., Montecchi-Palazzi,L., Aranda,B., Quinn,A.F., Vinod,N., Bader,G.D., Xenarios,I., Wojcik,J., Sherman,D. et al. (2007) Broadening the horizon—level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.*, **5**, 44.
19. Brohee,S. and van Helden,J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, **7**, 488.
20. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
21. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
22. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. et al. (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
23. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdottir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
24. Lysenko,A., Defoin-Platel,M., Hassani-Pak,K., Taubert,J., Hodgman,C., Rawlings,C.J. and Saqi,M. (2011) Assessing the functional coherence of modules found in multiple-evidence networks from *Arabidopsis*. *BMC Bioinformatics*, **12**, 203.
25. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **1165–1188**.
26. Hudson,T.J., Anderson,W., Artez,A., Barker,A.D., Bell,C., Bernabe,R.R., Bhan,M.K., Calvo,F., Eerola,I., Gerhard,D.S. et al. (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.
27. Ovaska,K., Laakso,M., Haapa-Paananen,S., Louhimo,R., Chen,P., Aittomaki,V., Valo,E., Nunez-Fontarnau,J., Rantanen,V., Karinen,S. et al. (2010) Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med.*, **2**, 65.
28. Smoot,M.E., Ono,K., Ruscheinski,J., Wang,P.L. and Ideker,T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.
29. Falcon,S. and Gentleman,R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.
30. Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJ. Complex Syst.*, **1695**, 1–9.