BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 6 2009, pages 830–831
doi:10.1093/bioinformatics/btp055

Systems biology

The DICS repository: module-assisted analysis of disease-related gene lists

Sabine Dietmann[1,*], Elisabeth Georgii[2], Alexey Antonov[1], Koji Tsuda[2] and Hans-Werner Mewes[1]

[1]Institute for Bioinformatics and Systems Biology, German Research Center for Environmental Health, Ingolstädter Landstraße 1, D-85764 Neuherberg, [2]Max Planck Institute for Biological Cybernetics, Spemannstr. 38, D-72076 Tübingen, Germany

Received on April 23, 2008; revised on November 05, 2008; accepted on January 24, 2009

Advance Access publication January 28, 2009

Associate Editor: Trey Ideker

ABSTRACT

Summary: The DICS database is a dynamic web repository of computationally predicted functional modules from the human protein–protein interaction network. It provides references to the CORUM, DrugBank, KEGG and Reactome pathway databases. DICS can be accessed for retrieving sets of overlapping modules and protein complexes that are significantly enriched in a gene list, thereby providing valuable information about the functional context.

Availability: Supplementary information on datasets and methods is available on the web server http://mips.gsf.de/proj/dics

Contact: sabine.dietmann@googlemail.com

1 INTRODUCTION

Network-based approaches have recently received much attention for identifying disease-related genes and for characterizing their functional context (Chuang *et al*., 2007; Lage *et al.*, 2007). A common strategy to exploit network information is the derivation of modules, i.e. sub-networks, whose nodes are densely interconnected. Dense modules derived from protein–protein interaction networks suggest potential protein complexes or functional modules. A number of approaches for module detection have been presented (Enright *et al*., 2002; Spirin and Mirny, 2003; see Sharan *et al.*, 2007 for a review). For example, CFinder (Palla *et al*., 2005) offers tools for analyzing dense overlapping modules from networks; and CellCircuits (Mak *et al*., 2007) is a searchable repository for published module predictions.

Until recently, there was no easy-to-use database for the non-expert user to exploit predicted modules for the analysis of experimentally derived gene lists. Here, we introduce the DICS (dense modules from protein interaction networks) repository, which offers to experimental researchers carefully benchmarked modules at multiple granularities that are compiled from the human protein–protein interaction network. The web server supports the exploration of measurement data, such as those resulting from genome-wide expression, proteomics and whole genome association studies, by providing enriched modules and protein complexes, as well as disease-related annotation.

*To whom correspondence should be addressed.

2 THE DICS SERVER

At the heart of the DICS server there is an algorithm that exhaustively enumerates all modules from the human protein–protein interaction network whose density exceed a pre-specified threshold (Uno, 2007). The density of a module is defined by the number of known direct interactions between genes within the module divided by the number of interactions in a clique formed using those genes. Briefly, the algorithm adopts the reverse search paradigm to organize the modules efficiently in a search tree such that their density is monotonically decreasing (see web server for details). Human protein–protein interaction data are collected from the IntAct,[1] BIND,[2] MINT[3] and HPRD[4] databases. Confidence scores are assigned to each interaction by assessing the corresponding set of experimental techniques (Jansen *et al*., 2003, see web server for details). DICS is updated on a 3-month basis according to updates in the reference databases.

2.1 Collection of modules

To obtain modules, we need to determine cutoff values for the module density and for the interaction confidence score (Fig. 1). There are two competing goals: (i) to most accurately recover the protein complexes in the CORUM database (Ruepp *et al*., 2008) and (ii) to extend the coverage of disease-related genes as much as possible. By default, we set the density threshold to 1.0, i.e. fully interconnected modules, and remove 30% of the interactions considered to be least reliable. The resulting 9859 modules cover 598 out of 1077 protein complexes with an average reliability of 0.58 for complex prediction. Furthermore, the modules cover 40% of the disease-related genes listed in the HGMD database (Stenson *et al*., 2003). Protein complexes cover only 11% of the disease genes. Optionally, the user can choose three pre-computed module sets with selected parameter combinations, as shown in Figure 1.

[1]http://www.ebi.ac.uk/intact.
[2]http://www.bind.ca.
[3]http://mint.bio.uniroma2.int/mint.
[4]http://www.hprd.org.

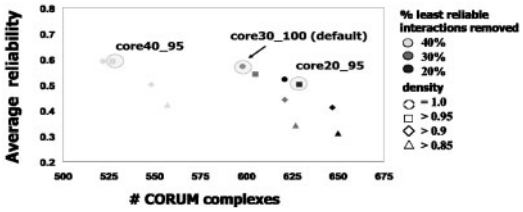

Fig. 1. Module validation. Dense modules are compared with 1077 human protein complexes from CORUM. Three module sets that are provided on the web server are indicated by circles.

2.2 Enrichment analysis

DICS can be accessed for identifying dense modules and known protein complexes that are significantly enriched in gene lists provided by high-throughput studies. The significance of association between a given gene set and each module or complex is estimated by a Monte-Carlo simulation procedure (Antonov *et al.*, 2008; see web server for details). Protein complexes and modules can be listed individually; or unions of the significant modules are provided for all pairs of modules, whose overlap score exceeds a specified threshold. The overlap score is defined as $(N \times N)/N1 \times N2$, where N, $N1$ and $N2$ are the number of proteins in the overlap, and the modules 1 and 2, respectively.

The web server provides modules that are significantly associated with the disease mutations extracted from the HGMD database and mouse phenotypes from the MPD database (Bogue *et al*., 2007; see 'Examples' section on the web server). Due to the small number of interactions experimentally determined for mouse proteins, orthologous modules are inferred in the mouse using the groups of orthologous proteins provided by the InParanoid database (O'Brien *et al*., 2005).

2.2.1 Examples To demonstrate the utility of module-assisted analysis, we automatically extracted results reported by 171 recently published proteomics studies and performed an enrichment analysis. For example, Quero *et al.* (2004) identified proteins to be differentially expressed in patients affected by the toxic oil syndrome (TOS), a disease that is characterized by severe muscle pain, polyneuropathy and skin changes. Among the proteins were several haptoglobin isoforms, which the authors concluded to be related to TOS affection.

Our analysis raises a further interesting hypothesis about a functional module underlying the observed phenotypes and the relationship of TOS to other diseases. Figure 2 shows the results provided by the web server for 12 differentially expressed proteins (Quero *et al*, 2004; Table 1, automatically extracted). One set of overlapping modules contained two input proteins (*TTR*, *APOA1*), which are associated with amyloidotic polyneuropathy. Within the same module, there are three proteins (*KRT1*, *KRT9*, *KRT16*), which are associated with keratoderma, a skin disease caused by clumped keratin filaments. Interestingly, these clinical manifestations overlap with those observed for the TOS, suggesting that the changed expression of proteins (*TTR*, *APOA1*, *IGHA1*, *IGHG1*) from this functional module is causing a related phenotype. The predicted

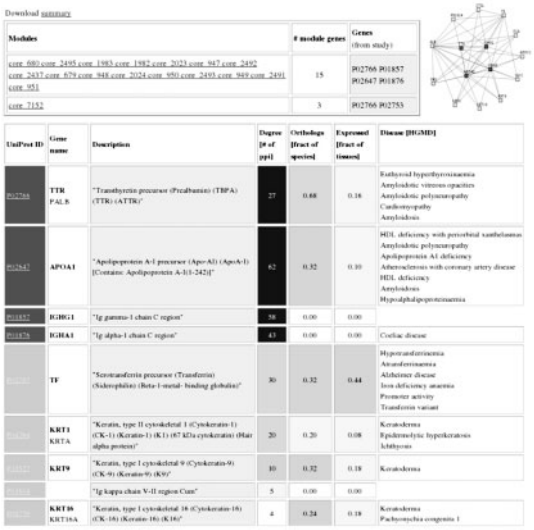

Fig. 2. Enrichment analysis results pages. A query with 12 proteins retrieves two sets of modules. The first set is a union of 16 overlapping modules containing four input genes (overlap score = 0.4). Input proteins are shown in dark blue, additional genes in light blue. Only a fraction of the module gene list is shown. Legends and color codes are described on the web server.

modules for the discussed study and 170 further proteomics studies can be accessed on the web server in the 'Examples' section.

Conflict of Interest: none declared.

REFERENCES

Antonov,A. *et al.* (2008) ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data *Nucleic Acid Res.*, **36**, W347–W351.

Bogue,M.A. *et al.* (2007) Mouse Phenome Database (MPD). *Nucleic Acid Res.*, **35**, D643–D649.

Chuang,H.Y. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.

Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acid Res.* **30**, 1575–1584.

Jansen,R. *et al.* (2003) A Bayesian networks approach for prediction protein-protein interactions from genomic data. *Science*, **17**, 449–453.

Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Mak,H.C. *et al.* (2007) CellCircuits: a database of protein network models. *Nucleic Acid Res.*, **35**, D538–D545.

O'Brien,K.P. *et al.* (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acid Res.* **33**, D476–D480.

Palla,G. *et al.* (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Science*, **435**, 814–818.

Quero,C. *et al.* (2004) Determination of protein markers in human serum: analysis of protein expression in toxic oil syndrome studies. *Proteomics*, **4**, 303–315.

Ruepp,A. *et al.* (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.*, **36**, 646–650.

Sharan,R. *et al.* (2007) Network-based prediction of protein function. *Mol. Syst. Biol.*, **3**, 88.

Spirin,V. and Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA*, **14**, 12123–12128.

Stenson.P.D. *et al.* (2003) Human Gene Mutation Database (HGMD): 2003 update. *Human Mutat.*, **21**, 577–581.

Uno,T. (2007) An efficient algorithm for enumerating pseudo-cliques. In *Proceedings of ISAAC 2007*, LNCS, Springer Verlag, Heidelberg, pp. 402–414.