W452–W459 Nucleic Acids Research, 2008, Vol. 36, Web Server issue Published online 7 May 2008
doi:10.1093/nar/gkn230

GraphWeb: mining heterogeneous biological networks for gene modules with functional significance

Jüri Reimand[1,2], Laur Tooming[1], Hedi Peterson[3,4], Priit Adler[3] and Jaak Vilo[1,4,*]

[1]University of Tartu, Institute of Computer Science, Liivi 2, Tartu, Estonia, [2]EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [3]University of Tartu, Institute of Molecular and Cell Biology, Riia 23a and [4]QureTec Ltd. Ülikooli 6a, Tartu, Estonia

Received January 31, 2008; Revised April 3, 2008; Accepted April 11, 2008

ABSTRACT

Deciphering heterogeneous cellular networks with embedded modules is a great challenge of current systems biology. Experimental and computational studies construct complex networks of molecules that describe various aspects of the cell such as transcriptional regulation, protein interactions and metabolism. Groups of interacting genes and proteins reflect network modules that potentially share regulatory mechanisms and relate to common function. Here, we present GraphWeb, a public web server for biological network analysis and module discovery. GraphWeb provides methods to: (1) integrate heterogeneous and multispecies data for constructing directed and undirected, weighted and unweighted networks; (ii) discover network modules using a variety of algorithms and topological filters and (iii) interpret modules using functional knowledge of the Gene Ontology and pathways, as well as regulatory features such as binding motifs and microRNA targets. GraphWeb is designed to analyse individual or multiple merged networks, search for conserved features across multiple species, mine large biological networks for smaller modules, discover novel candidates and connections for known pathways and compare results of high-throughput datasets. The GraphWeb is available at http://biit.cs.ut.ee/graphweb/.

INTRODUCTION AND BACKGROUND

One of the greatest challenges of biomedical research is to understand the organization and function of living organisms at the molecular level. Experimental and computational data reveal complex networks that consist of genes and proteins as nodes and associations as edges (1–3). While describing different aspects of the cell, these networks appear to share universal structural properties like log-linear distribution of connections and small-world reachability (4,5). Within networks, modules of tightly interacting genes and proteins are believed to make up functional units responsible for processes in the cell (6). For instance, collections of protein–protein interactions (PPI) form networks of physically binding proteins, where modules reflect protein complexes or signalling pathways (7,8). Gene expression measures, transcription regulator binding data, *cis*-regulatory motif discovery and conservation information are combined to uncover transcription regulatory networks with modules of transcription factors (TFs) and target genes (9–12). From a slightly different angle, text-mining methods extract knowledge-based webs and co-occurring modules of genes and proteins from scientific literature (13).

Biological network analysis proposes the following computational challenges. The strategies need to take into account the myriad of cellular interactions that may be directed (e.g. TF–gene interaction) or undirected (e.g. PPI), involve quantitative values (e.g. gene expression correlation) or appear in multiple datasets (e.g. co-expression and physical interaction) (14). Combining different cellular domains requires data integration to deal with various biomolecules and experimental measurements (15). Module detection involves algorithms that identify nodes with special topological features or search for densely connected areas (16). Biological interpretation of modules comprises functional analysis using resources such as the Gene Ontology (GO) (17) and detection of significantly enriched biological processes, functions and cellular locations (18).

The growing interest in networks and systems biology has increased the need for computational and visual methods for network analysis, and as a result, several

*To whom correspondence should be addressed. Tel: +372 50 49 365; Fax: +372 737 5468; Email: vilo@ut.ee; vilo@quretec.com

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://nar.oxfordjournals.org/ at Bora Laskin Law Library on May 30, 2012

useful tools have been published. Notable software libraries include AT&T Graphviz for visualization and C++ Boost for graph structures and algorithms, packaged into Bioconductor by Carey and collegues (19). Cytoscape is a popular software for visual analysis of biological networks (20). A number of plugins complement Cytoscape with analytical features such as microarray data integration, dense subgraph detection (21) and GO-term enrichment analysis (22). Osprey focuses on visualization (23), while VisANT also provides topological analysis and functional annotation of nodes (24). MATISSE is useful for mapping high-throughput datasets onto network topologies and detecting gene modules using a number of algorithms (25). BiologicalNetworks is a network retrieval, construction and visualization tool with an emphasis on microarray data (26). BioPIXIE provides a gene-based query engine and GO analysis for a precomputed heterogeneous network for *Saccharomyces cerevisiae* (27). NetworkBLAST allows the user to align and compare two networks of different species through user-provided sequence similarity measures to discover conserved protein complexes (28).

We have identified open questions in the field of biological network analysis. There is a lack of simple 'point-and-click' web servers that allow biological data integration and discovery of modules. Some of the available tools involve no biological background information and force the user to put great effort in integrating datasets, linking molecules and retrieving functional annotations, while others constrain the analysis to some pre-calculated network of a specific model organism. Module detection is frequently limited to neighbourhood search of gene lists or topological analysis such as node connectivity. Both Cytoscape and VisANT implement functionality for analysing high-throughput networks, detecting modules and enriched biological features. However, we believe that there is a need for web-based resources that analyse heterogeneous datasets with mixed collections of genes and proteins, detect various types of modules and and provide a rich interface for functional annotation. Moreover, there is little support for the analysis and integration of multispecies data using automatic orthology mapping. With the development of the GraphWeb server, we wish to contribute to the network challenge and propose new solutions to the above questions.

THE GraphWeb SERVER

GraphWeb (http://biit.cs.ut.ee/graphweb, Figure 1) is a public web server for graph-based analysis of cellular networks that:

(1) analyses directed and undirected, weighted and unweighted heterogeneous networks of genes, proteins and microarray probesets for 35+ eukaryotic genomes;
(2) integrates multiple diverse datasets into global networks;
(3) incorporates multispecies data using gene orthology mapping;
(4) filters nodes and edges based on dataset support, edge weight and node annotation;
(5) detects gene modules from networks using a collection of algorithms;
(6) interprets discovered modules using GO, pathways and *cis*-regulatory motifs.

Networks in GraphWeb

The primary input of GraphWeb is a combined biological network of a selected species, consisting of genes, proteins or microarray probesets as nodes and corresponding associations as edges. The user may upload the input data as a file or type it into the webform. Genes, proteins and microarray probesets of various databases and platforms are automatically mapped to gene IDs of the Ensembl database (29) using the g:Profiler software (30). Unrecognized and ambiguous IDs may be optionally removed, but remain unchanged by default in order to keep the input networks intact. Associations between nodes may be represented as directed or undirected edges, and weights may be assigned to edges to convey quantitative relations between corresponding nodes. A collection of pre-defined datasets is available for immediate analysis, including PPI from IntAct (31) and HPRD (32), and the *S.cerevisiae* transcription regulatory network by MacIsaac *et al.* (33).

Data integration

GraphWeb allows the user to insert and combine different data sources and align these into a global network. Besides native plaintext format, Graphweb supports the import of other network files such as SIF, GML, XGMML and BioPAX through the Cytoscape BiNoM plugin (34). *Labels* can be used to distinguish associations of different sources, and a *network score* may be assigned to each label to denote the predictive power of corresponding associations. For example, TF-binding networks from ChIP-chip experiments may be combined and aligned with motif discovery results, and scored with predictive values learned from gene expression data.

The integration process first creates a global network that permits several connecting edges between a pair of nodes. This is followed by a label-wise weight normalization that makes associations of different networks comparable. Finally, a linear combination of edge weights $w_{h,i,j}$ and network scores s_h for different labels h is used to rank all connected nodes i, j:

$$S_{i,j} = \sum_h s_h \cdot w_{h,i,j}.$$

The score $S_{i,j}$ is designed to highlight associations with strong evidence from several sources. The user may also choose to create network scores automatically and assign proportionally more power to smaller datasets. This option provides a direct measure for preferring smaller, assumably high-quality networks. GraphWeb only supports the alignment of unambiguous known IDs, since the alignment of ambiguous entities may lead to erroneous networks. Proteins or probes that map to several base gene IDs are treated as independent nodes and corresponding edges are not aligned.

Downloaded from http://nar.oxfordjournals.org/ at Bora Laskin Law Library on May 30, 2012

Multispecies networks

GraphWeb provides means to incorporate data from different organisms in order to improve network construction. When the user selects a target organism in the GraphWeb interface the nodes and corresponding associations of the input are automatically mapped to orthologous genes in the target. The orthology mapping information is retrieved from Ensembl via g:Profiler software. Resulting ortholog networks can be combined with other datasets of the target organism to highlight conserved associations. Similarly to single-species data integration, GraphWeb ignores ambiguous orthologs in network alignments to avoid noise and misleading results. Such a solution retains the cleanest possible network but undoubtedly results in a certain loss of information.

Graph filtering

GraphWeb filters help the user detect network areas with strong associations. Three types of filters may be used for selecting edges: minimum number of supporting datasets (i.e. labels), lower threshold on edge weights and selection of top-ranking edges. Node filtering excludes unrecognized or ambiguous genes and proteins, while module filtering limits the result to larger modules or those with significant functional enrichments. Filtering techniques are especially useful when incorporating edges from different datasets or species.

Gene module discovery

GraphWeb provides a number of methods and algorithms for detecting gene modules in directed and undirected networks. Resulting gene modules may easily be saved for later use or redirected to input for further analysis. GraphWeb identifies the following types of modules.

Connected components. A connected component (Figure 2A) is a group of genes, where every pair of genes, (g_i, g_j) is connected either directly ($g_i \smile g_j$) or indirectly via a *path* of length n, ($g_1 \smile g_2 \smile \ldots \smile g_n \smile g_{n+1}$). GraphWeb also supports two extensions of the above: a *strongly connected component* relates to directed networks and requires connections in both directions, and a *biconnected component* requires at least two non-overlapping paths. Connected component detection is the first step in studying network structure.

Neighbourhood modules. A neighbourhood module (Figure 2D) is based on a user-defined list of genes and proteins $\{G\}$ and on a distance d. If $d = 0$, GraphWeb retrieves modules that consist of nodes G with internal associations inside the list. If $d \geq 1$, modules consist of the initial list $\{G\}$ and nodes connected to the latter via paths of maximum length d. Neighbourhood modules allow the user to study her focus list in a network context, and retrieve related nodes and associations to propose new hypotheses.

Hub-based modules. A hub-based module (Figure 2B) consists of a central *hub* (a node with many connections) and related genes and proteins within distance d. GraphWeb extracts a list of hub-based modules ranked by the central hub *degree* (number of connections). Hubs in PPI networks have been described in the context of lethality (35), and proteins linking to the same hub often refer to similar function (36). Hub-based modules may also reflect systems of TFs and target genes.

Cliques. A clique (Figure 2C) is a fully connected module where every pair of nodes is directly connected. Cliques in PPI networks have often been related to protein complexes and common functions (36). Fully connected modules also reflect clusters of co-expressed genes.

Cluster modules. A cluster module corresponds to a tightly connected group of nodes. GraphWeb provides two network clustering algorithms: the Markov Cluster (MCL) algorithm (37) and Betweenness Centrality Clustering (BCC) (38). These algorithms break networks down into separate modules by removing certain edges, and have been successfully applied in a number of studies, such as protein family detection (39) and essentiality assessment (40). MCL constructs modules of edges that are frequently visited during random walks, while BCC removes paths that act as bridges between separate tightly connected modules. Graph clustering is successful in integrative network analysis since it prefers associations with evidence from multiple datasets, and allows the detection of hybrid modules that combine the characteristics of different module types.

Empirical comparisons show that the time complexity of the above algorithms is generally linear to the number of edges. The NP-complete clique detection algorithm is the most computationally expensive method in GraphWeb and is especially sensitive to dense networks, where a network of 30 nodes and 300 edges requires a computation of nearly 10 min. MCL clustering, on the other hand, takes 10 min to handle a network of nearly 8000 nodes and 300 000 edges using GraphWeb default values. Hub-based modules and connected components are detected even faster.

Module interpretation and evaluation

Interpretation and evaluation is an integral process of module detection in GraphWeb. Once a module has been identified, GraphWeb automatically assesses its biological importance through the known properties of its members using the g:Profiler software. Functional profiling of the module involves statistically enriched annotations of biological processes (bp), cellular locations (cc) and molecular functions (mf) from the GO (17), and related pathways (pw) from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (41) and Reactome (42). Besides functional annotations, the analysis takes into account *cis*-regulatory motif enrichments from TRANSFAC (43) and miRNA target site enrichments from miRBase (44).

First, g:Profiler applies the Fisher's test to evaluate the enrichments of all biological annotations in the module:

$$p_\alpha = \sum_{x=k}^{\min(n,K)} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Downloaded from http://nar.oxfordjournals.org/ at Bora Laskin Law Library on May 30, 2012

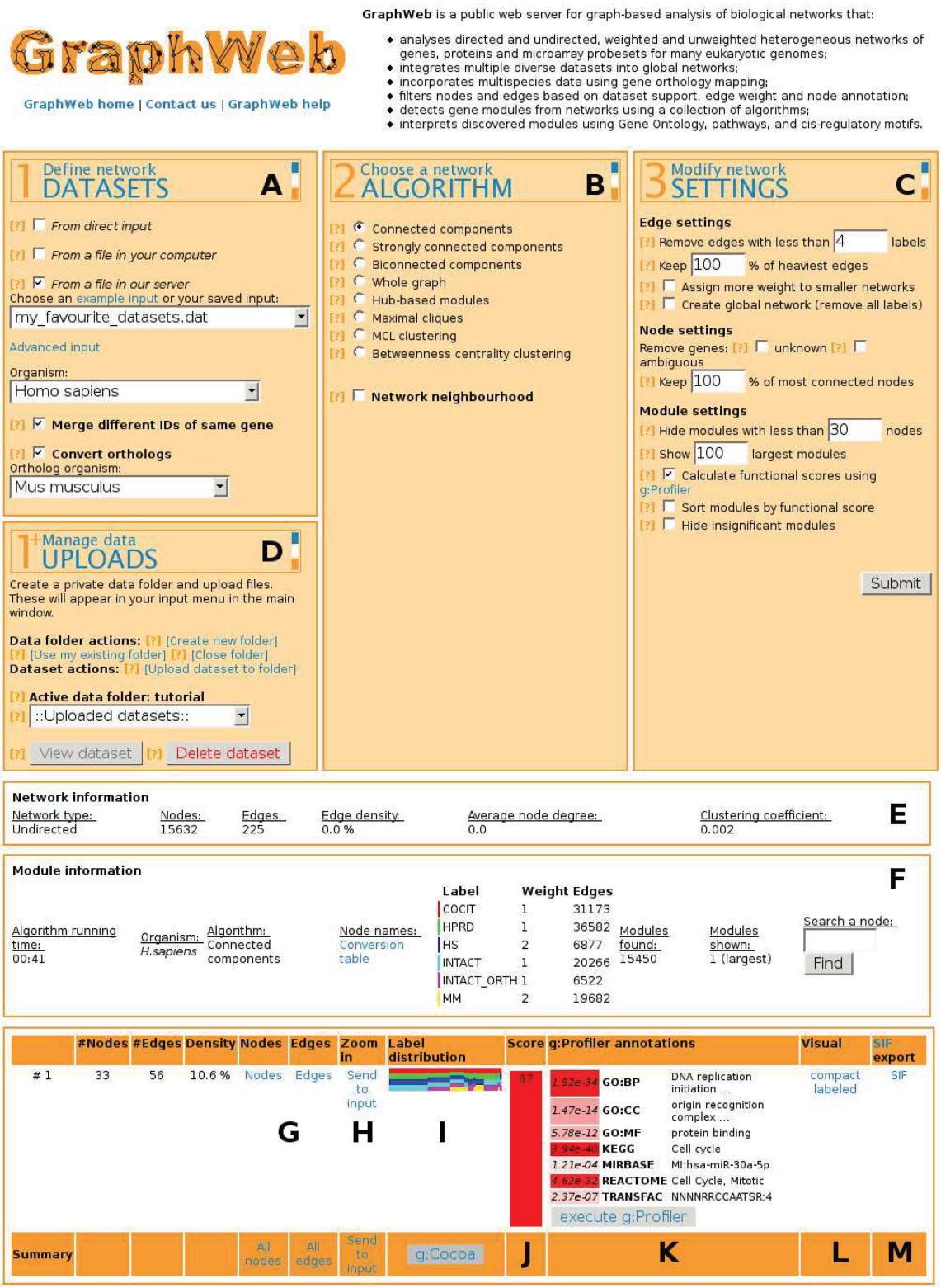

Figure 1. GraphWeb user interface with data from the case study of human PPI and gene expression (see Results Section for a detailed description). The first module of 33 nodes is shown in Figure 2. User interface legend: (**A**) data upload, (**B**) module detection algorithms, (**C**) options and filters, (**D**) user data storage, (**E**) network information and labels, (**F**) module information and gene search, (**G**) module export, (**H**) module zoom-in analysis, (**I**) module label distribution, (**J**) module annotation score, (**K**) best functional enrichments and link to g:Profiler, (**L**) links to module visualization and (**M**) export to SIF format.

Downloaded from http://nar.oxfordjournals.org/ at Bora Laskin Law Library on May 30, 2012

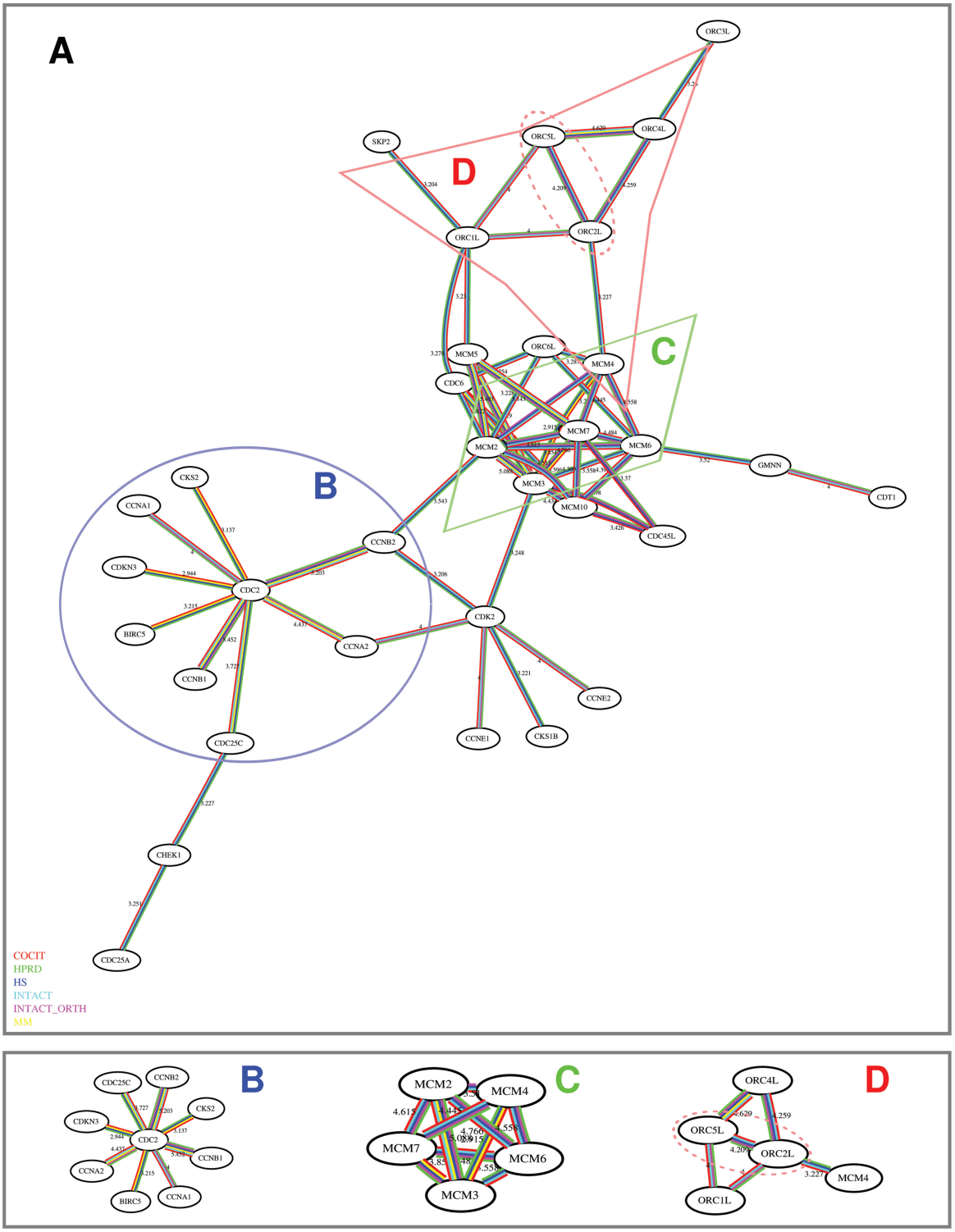

Figure 2. The case study: a connected component (**A**) detected from the combined network for protein interactions and gene expression similarity. The discovered module describes a fragment of the human cell cycle and consists of several smaller modules. Two cyclin-dependent kinases (CDC2, CDK2) are hubs regulating different cyclins [e.g. CDC2 module (**B**)]. MCM2-7 proteins form a helicase and five of these connect into a clique (**C**). The network neighbourhood module of ORC2L and ORC5L (**D**) contains origin recognition complex proteins.

The test computes the cumulative hypergeometric probability of randomly observing at least k genes with some common annotation α out of the n genes in the module, given the total number of genes N and the total number of genes having the annotation K. The g:Profiler uses a 5% multiple testing threshold g:SCS that applies a simulation procedure to retrieve only the significant enrichments from a hierarchical annotation structure like GO (45).

Downloaded from http://nar.oxfordjournals.org/ at Bora Laskin Law Library on May 30, 2012

Once all enrichments for the module are known, GraphWeb computes an *annotation score* that sums the total significance relative to module size n:

$$P = \frac{\sum_{\alpha} -log_{10}(p_{\alpha})}{n}.$$

The score is designed to highlight modules with strong size-independent enrichment of functions and regulatory features.

GraphWeb executes on-the-fly functional profiling and scoring of detected modules, displaying the names and *P*-values of most important discovered features from all the covered functional domains (GO:bp, GO:cc, GO:mf, KEGG:pw, Reactome:pw, TRANSFAC, miRBase). Hyperlinks to g:Profiler allow the user to access related terms and pathways, ortholog mapping and expression similarity search for related genes. In addition, a hyperlink to g:Cocoa at the bottom of the GraphWeb interface sends all discovered modules to comparative functional enrichment analysis.

RESULTS: A CASE STUDY

We present an example case study that demonstrates a possible data integration and module detection pipeline. The analysis concentrates on human cellular networks and involves six high-throughput datasets comprising gene expression values and PPI from public databases. Human PPI data originate from the study by (46) and the databases HPRD (32) and IntAct (31), and are interpreted as three separate networks. Human expression data are presented as an expression similarity network, computed using Multi Experiment Matrix (MEM) (Adler *et al.*, manuscript in preparation) across nearly 3700 tumour-related samples of 89 public datasets, originating from GEO (47) and ArrayExpress (48). Besides human data, we use orthology mapping to incorporate two datasets for mouse: a MEM gene expression similarity network across 28 datasets and 1700 samples, and the PPI data from IntAct.

Unweighted PPI datasets and weighted expression similarity datasets are aligned into a global-weighted network. Integration of the above datasets reveals frequently co-expressed protein complexes such as ribosome and proteasome. We applied a strong edge filter of minimum dataset support 4, and queried for connected components. The largest resulting component consists of 33 nodes and four notable submodules, is included in known pathways of Reactome and KEGG, and involves strong GO enrichments.

The module plays a significant role in cell cycle and is well described with PPI as well as gene expression similarity. The two hubs denote cyclin-dependent kinases 1 (CDC2/CDK1) and 2 (CDK2), see Figure 2B for the former module. These kinases control the cell cycle entry to S-phase, while CDK1 also controls the entry to mitosis (49). MCM2-7 proteins form a helicase and five of these connect into a clique (Figure 2C). The neighbourhood of ORC2L and ORC5L partly reveals the origin recognition complex (ORC) (Figure 2D), that temporarily interacts with CDT1 and CDC6 and binds to the helicase to initiate replication in S-phase. Other connected proteins include cell cycle checkpoint controllers (e.g. CHEK1 kinase), inhibitors (GMNN, BIRC5) and cyclins (CCNE1, CCNE2, CCNB1).

The thorough common-knowledge description of the detected module provides support for the techniques proposed in GraphWeb. The rather strong filters applied above naturally extracted a well-studied result out of a large collection of public data. The GraphWeb case study provides a simple example of the possibilities and potential results of analysing novel data or combining it with existing public repertoires.

DISCUSSION

The core data structures and algorithms in GraphWeb render the myriad of molecular entities and corresponding relations, physical connections and regulatory events into a uniform collection of network nodes and connecting edges. On the one hand, this simplification creates an intuitive view of the cellular networks. GraphWeb analysis methods allow the researcher to approach a number of interesting tasks, for example proposing novel members of known pathways by strong 'guilt by association' evidence, comparing the results of multiple high-throughput datasets, or finding associations and modules of genes that are conserved in diverse species. On the other hand, looking at topological features, weighted edges and tightly connected groups of nodes may admittedly fail to deliver crucial aspects of biological systems, such as quantitative dependencies and dynamics over time. The greatest advantage of GraphWeb analysis is its relative simplicity and speed in handling complex objects as networks. We therefore believe that GraphWeb also proves useful in detailed network studies, since it allows the user to reduce the complexity of the whole network to the complexity of modules. Such a reduction may then provide access to more elaborate methods of mathematical modelling that are inapplicable to systems larger than a handful of variables.

CONCLUSION

GraphWeb is a publicly available web server for analysing and interpreting complex cellular networks. The server provides methods for integrating heterogeneous datasets into networks of interactions, means to incorporate multispecies data using gene orthology information, algorithms and methods for discovering network modules and functional enrichment analysis for biological interpretation. With the creation of the GraphWeb server, we wish to contribute to the difficult task of deciphering and understanding complex biological networks, and provide a tool with an emphasis on ease of use.

IMPLEMENTATION

The GraphWeb web server is implemented in Perl as a CGI application. Graph structures and algorithms are written in C++ and Perl and are partly based on the Boost Graph Library (http://www.boost.org/). GraphWeb applies the MCL algorithm implementation by

Downloaded from http://nar.oxfordjournals.org/ at Bora Laskin Law Library on May 30, 2012

van Dongen (37) (http://micans.org/mcl/). Visualization is provided by the AT&T Graphviz graph drawing package (http://www.graphviz.org/) and the SWOG graphical programming language (http://biit.cs.ut.ee/SWOG/).

ACKNOWLEDGEMENTS

The authors wish to thank Dr Nicholas Luscombe and the anonymous reviewers for valuable remarks on the articles and software. This work has been supported by the EU FP6 grants ENFIN LSHG-CT-2005-518254 and COBRED LSHB-CT-2007-037730, and Estonian Science Foundation grant ETF7437. J.R. has recieved funding from the Marie Curie Biostar program and the Tiger University program of the Estonian Information Technology Foundation. Funding to pay the Open Access publication charges for this article was provided by the European Commission (COBRED) project.

Conflict of interest statement: None declared.

REFERENCES

1. Jeong,H., Tombor,B., Albert,R., Oltvai,Z.N. and Barabasi,A.L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.
2. Oltvai,Z.N. and Barabasi,A.L. (2002) Life's complexity pyramid. *Science*, **298**, 763–764.
3. Maslov,S. and Sneppen,K. (2002) Specificity and stability in topology of protein networks. *Science*, **296**, 910–913.
4. Strogatz,S.H. (2001) Exploring complex networks. *Nature*, **410**, 268–276.
5. Barabasi,A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nature*, **5**, 101–113.
6. Hartwell,L.H., Hopfield,J.J., Leibler,S. and Murray,A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.
7. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.
8. Gavin,A.C., Aloy,P., Grandi,P., Krause,R., Boesche,M., Marzioch,M., Rau,C., Jensen,L.J., Bastuck,S., Dmpelfeld,B. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.
9. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al.* (2003) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
10. Segal,E., Shapira,M., Regev,A., Pe'er,D., Botstein,D., Koller,D. and Friedman,N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, **34**, 166–176.
11. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
12. Tanay,A., Regev,A. and Shamir,R. (2005) Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *PNAS*, **102**, 7203–7208.
13. Jensen,L.J., Saric,J. and Bork,P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, **7**, 119–129.
14. Carter,G.W. (2005) Inferring network interactions within a cell. *Brief. Bioinform.*, **6**, 380-389.
15. Troyanskaya,O.G. (2005) Putting microarrays in a context: integrated analysis of diverse biological data. *Brief. Bioinform.*, **6**, 34–43.
16. Aittokallio,T. and Schwikowski, B. (2006) Graph-based methods for analysing networks in cell biology. *Brief. Bioinform.*, **7**, 243–255.
17. 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. *Nat. Genet.*, **1**, 25–29.
18. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
19. Carey,V.J., Gentry,J., Whalen,E. and Gentleman,R. (2003) Network structures and algorithms in Bioconductor. *Bioinformatics*, **21**, 135–136.
20. Cline,M.S., Smoot,M., Cerami,E., Kuchinsky,A., Landys,N., Workman,C., Christmas,R., Avila-Campilo,I., Creech,M., Gross,B. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protocols*, **10**, 2366–2382.
21. Ideker,T., Ozier,O., Schwikowski,B. and Siegel,A.F. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**, S233–S240.
22. Maere,S., Heymans,K. and Kiper,M. (2005) BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.
23. Breitkreutz,B.J., Stark,C. and Tyers,M. (2003) Osprey: a network visualization system. *Genome Biol.*, **4**, R22.
24. Hu,Z., Ng,D.M., Yamada,T., Chen,C., Kawashima,S., Mellor,J., Linghu,B., Kanehisa,M., Stuart,J.M. and DeLisi,C. (2007) VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Res.*, **W35**, W625–W632.
25. Ulitsky,I. and Shamir, R. (2007) Identification of functional modules using network topology and high-throughput data. *BMC Systems Biol.*, **1**, 8.
26. Baitaluk,M., Sedova,M., Ray,A. and Gupta,A. (2006) Biological Networks: visualization and analysis tool for systems biology. *Nucleic Acids Res.*, **W34**, W466–W471.
27. Myers,C.L., Robson,D., Wible,A., Hibbs,M.A., Chiriac,C., Theesfeld,C.L., Dolinski,K. and Troyanskaya,O.G. (2005) Discovery of biological networks from diverse functional genomic data. *Genome Biol.*, **6**, R114.
28. Kalaev,M., Smoot,M., Ideker,T. and Sharan,R. (2008) NetworkBLAST: comparative analysis of protein networks. *Bioinformatics*, **4**, 594–596.
29. Hubbard,T.J.P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **D35**, D610–D617.
30. Reimand,J., Kull,M., Hansen,J., Peterson,H. and Vilo,J. (2007) g:Profiler – a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, **W35**, W193-W200.
31. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A. *et al.* (2007) IntAct – open source resource for molecular interaction data. *Nucleic Acids Res.*, **D35**, D561–D565.
32. Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K.B., Gronborg,M. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.
33. MacIsaac,K.D., Wang,T., Gordon,D.B., Gifford,D.K., Stormo,G.D. and Fraenkel,E. (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinform.*, **7**, 113.
34. Zinovyev,A., Viara,E., Calzone,L. and Barillot,E. (2008) BiNoM: a cytoscape plugin for manipulating and analyzing biological networks. *Bioinformatics*, **6**, 876–877.
35. Jeong,H., Mason,S.P., Barabsi,A.L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
36. Przulj,N., Wigle,D.A. and Jurisica,I. (2004) Functional topology in a network of protein interactions. *Bioinformatics*, **20**, 340–348.
37. van Dongen,S. (2000) Graph clustering by flow simulation. *Ph.D. Thesis*. University of Utrecht.
38. Dunn,R., Dudbridge,F. and Sanderson,C.M. (2005) The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinform*, **6**, 39.

Downloaded from http://nar.oxfordjournals.org/ at Bora Laskin Law Library on May 30, 2012

39. 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.
40. Yu,H., Kim,P.M., Sprecher,E., Trifonov,V. and Gerstein,M. (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.*, **3**, e59.
41. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T., *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–D484. (http://nar.oxfordjournals.org/cgi/content/abstract/36/suppl_1/D480).
42. Vastrik,I., D'Eustachio,P., Schmidt,E., Joshi-Tope,G., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B., Lewis,S. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.* **8**, R39.
43. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–110.
44. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34**, D140–D144.
45. Reimand,J. (2006) Gene Ontology mining tool GOSt. *Master's Thesis*. University of Tartu, Estonia.
46. Ramani,A.K., Bunescu,R.C., Mooney,R.J. and Marcotte,E.M. (2005) Consolidating the set of known human protein–protein interactions in preparation for large-scale mapping of the human interactome *Genome Biol.*, **6**, R40.
47. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, **D35**, D760–D765.
48. Parkinson,H., Kapushesky,M., Shojatalab,M., Abeygunawardena,N., Coulson,R., Farne,A., Holloway,E., Kolesnykov,N., Lilja,P., Lukk,M. *et al.*. (2007) ArrayExpress – a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **D35**, D747–D750.
49. Bashir, T. and Pagano, M. (2005) Cdk1: the dominant sibling of Cdk2. *Nat. Cell Biol.*, **7**, 779-781.

Downloaded from http://nar.oxfordjournals.org/ at Bora Laskin Law Library on May 30, 2012