Title Page

Title:

Genes2Networks: Connecting Lists of Proteins by Using Background Literature-based Mammalian Networks

Authors:

Jeremy M. Posner, Seth I. Berger, Avi Ma'ayan

E-mail addresses:

jeremy.posner@mssm.edu
seth.berger@mssm.edu
avi.maayan@mssm.edu

Address:

Department of Pharmacology and Systems Therapeutics
Mount Sinai School of Medicine, New York, NY 10029 USA

Corresponding author:

Avi Ma'ayan
avi.maayan@mssm.edu
(212) 659-1744

Abstract

Background:

In recent years, *in-silico* literature-based mammalian protein-protein interaction network datasets have been developed. These datasets contain binary interactions extracted manually from legacy experimental biomedical research literature. Placing lists of genes or proteins identified as significantly changing in multivariate experiments, in the context of background knowledge about binary interactions, can be used to place these genes or proteins in the context of pathways and protein complexes.

Results:

Genes2Networks is a software system that integrates the content of ten mammalian literature-based interaction network datasets. Filtering to prune low-confidence interactions was implemented. Genes2Networks is delivered as a web-based service using AJAX. The system can be used to extract relevant subnetworks created from "seed" lists of human Entrez gene names. The output includes a dynamic linkable three color web-based network map, with a statistical analysis report that identifies significant intermediate nodes used to connect the seed list. Genes2Networks is available at http://actin.pharm.mssm.edu/genes2networks.

Conclusion:

Genes2Network is a powerful web-based software application tool that can help experimental biologists to interpret high-throughput experimental results used in genomics and proteomics studies where the output of these experiments is a list of significantly changing genes or proteins. The system can be used to find relationships between nodes from the seed list, and predict novel nodes that play a key role in a common function.

Background

The rapid increase in binary interactions experimentally identified has brought us to a stage where on one hand we are now able to start viewing how all those interactions and components come together to form large functional regulatory networks [1]. But on the other hand, it is impossible for researchers to keep up with the literature. The emergence of multivariate experimental technologies such as yeast-2-hybrid screens [2, 3], cDNA microarrays [4, 5] and mass-spectrometry [6] as well as databases that mine legacy experimental literature [7, 8] allows for the construction of networks. Networks are a simple abstract representation of biomolecular interactions where cellular components are represented as nodes, and where interactions connect these nodes through links.

The construction of cellular network datasets has several valuable uses. Network representation allows for extraction of global topological statistical and structural properties such as connectivity distribution [9], clustering [10], and the identification of network motifs [11] or graphlets [12]. These measurements provide clues about the design principles of intracellular organization. Interaction network datasets can also be used to predict unidentified interactions [13, 14], or used as a starting point for quantitative computational modeling [15]. Additionally, interaction networks can assist in interpreting experimental results where identified lists of proteins or genes from multivariate experiments can be placed in their contextual local networks of interactions [16].

Implementation

Our aim here is to provide cell- and molecular-biologists with an easy method to create subnetworks from lists of mammalian genes or proteins by using large-scale high-quality protein-protein and signaling networks created from background literature. To develop Genes2Networks, we merged ten currently available literature-based mammalian interaction network datasets by consolidating them into one large dataset. To prune out interactions of low confidence, a simple filter was implemented. Genes2Networks is delivered as a web interface that can be used to extract relevant subnetworks based on inputted lists of gene or protein names, commonly produced by high-content experiments, using the consolidated datasets as a background. The system's input is a list of gene symbols; the system then uses the merged datasets as a background, to produce subnetworks that connect the gene names (commonly representing proteins) as the output. The output includes a statistical analysis report, and a three color network map highlighting the seed nodes in one color, the significant intermediates in another color, and the non-significant intermediates in a third color. The statistical analysis provides a list of intermediate nodes used to connect the genes, sorted based on their specificity significance. The process is illustrated in Figure 1.

Developing the background network

We used all mammalian (mouse/rat/human) interactions recorded in the following datasets: BIND [17], HPRD [18], IntAct [19], DIP [20], MINT [21], Vidal [22], Stelzl

[23], Ma'ayan [24], PDZBase [25], and PPID [19, 26]. All interactions from these databases were determined experimentally and include the PubMed reference of the research article that describes the experiments used to identify the interactions. Consolidating interactions from the different ten network databases was accomplished by combining human/mouse/rat gene symbols using information from Swiss-Prot [27]. The consolidated dataset before filtering contained 44,877 interactions among 11,033 nodes. The consolidated interactions were stored in a flat file format which is loaded into a hash data structure for fast loading and access. We did not include datasets of interactions created via *in-silico ab-initio* interaction prediction methods. Most of the above listed datasets describe binary interactions but some complexes of more than two proteins are described. We did not include those in the consolidated dataset. Nodes in the ten datasets are provided with accession codes linking them to entries describing genes and proteins in databases such as Swiss-Prot [27] and NCBI's Entrez Gene [28]. HPRD [18] and PPID [19, 26] are not included in the public web interface since these databases require a license for redistribution. Currently, these datasets are only available to internal users at Mount Sinai School of Medicine.

Filtering

Many of the interactions and components listed in the ten datasets we used are a result of high-throughput experiments which are considered low-quality since they often report high level of false positives [29]. Thus, we applied a simple filtering approach which excludes interactions if they originated from articles that provided many interactions, and/or gives more confidence to interactions reported by several different papers. We made the assumption that a research article that reports many interactions is likely to report false positives, and, alternatively, interactions that were reported in many different research articles and appear in many databases are likely to be real, and thus, have higher confidence. More sophisticated filtering techniques that would implement machine learning technologies such as support vector machines (SVM) [30], and would take into account more knowledge about interactions (i.e. experimental method used) are planned to be implemented in the future.

Web interface

To enhance accessibility to the tool, we developed a web-based interface to the software. The interface allows users to input a list of human Entrez Gene symbols, entered in a textbox or through a text file. As genes are added, the symbol is validated using NCBI-eutils (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html) by searching the NCBI gene database with the entered query restricting the organism to human. When an exact match is not found, the user is presented with a list of suggestions to choose from. Using the background merged consolidated network dataset, the program outputs subnetworks that describe all found interactions for the list of inputted gene symbols. Through the web interface, the user has full access to configure which datasets to include in the background network, and what filtering to apply to the background networks to remove low confidence interactions. This also includes the ability to upload user developed network datasets for inclusion in the background. The output is visualized

using a dynamical web-enable AJAX viewer called AVIS (http://actin.pharm.mssm.edu/AVIS2/). The viewer allows browsing, zooming and panning, and linking to interaction resources. The user can configure the colors of the outputted nodes so that the seed genes, intermediate genes above a specified z-score, and the rest of the nodes have different colors. The user needs to specify maximum number of steps/hops to use in order to connect the genes from the input list. Steps/hops are links (and nodes) needed to connect the inputted seed list. Additionally, the program outputs a statistical report that ranks intermediates used to connect the genes based on their specificity to interact with the seed list. As the user adjusts the settings, changes in the resulting network are automatically redisplayed. A representative screenshot of the system is illustrated in Figure 2.

Significant intermediates

The output subnetworks produced by Genes2Networks contain nodes, mostly proteins, which were not originally provided by the user as input. Some of those intermediate nodes may be present in the output subnetwork because the intermediates are highly connected nodes in the background network. On the other hand, intermediate nodes may be specific to interacting with components from the original seed input list. If those intermediates are specific, it may be useful for the user to identify them as potential specific regulators and specific participants in pathways and modules involving the input list of gene symbols. For this, Genes2Networks output a z-score value of the significance of intermediates in the output subnetwork. The z-score is computed using a binomial proportions test [31]:

$$z = \frac{\left(\frac{a}{c} - \frac{b}{d}\right)}{\sqrt{\frac{\frac{b}{d}\cdot\left(1-\frac{b}{d}\right)}{d}}} \qquad \text{Equation 1}$$

Where “a” equals the links from the intermediate node to nodes from the input list, “b” equals the total links for the intermediate node in the background network, “c” is the total links in the subnetwork, and “d” is the total links in the background network.

Discussion and Conclusions

Several commercial and academic initiatives have been attempting to address the need for integration, consolidation, visualization and organization of information about binary mammalian protein interactions and signaling pathways from sparse sources. For example, Cytoscape [32] and its several plug-ins allow for analysis and integration of experimental data as well as incorporation with Gene Ontology [33]. One of the plug-ins, called cPath [34] (http://cbio.mskcc.org/cpath/) is a database that joins together databases stored in PSI-MI XML format [19]. Other similar software platforms include: PIANA [35], Pathway Studio [7], ProViz [36], PATIKA [37], and Ingenuity (http://www.ingenuity.com/). Some are commercial products and some developed by

academic labs and are freely available. Genes2Networks provides several advantages over existing systems. The quality of the background datasets is high yet comprehensive, the user interface is an intuitive web-based Web 2.0 enabled application, the systems is free for academic users, the system provides predictions about intermediate components involvement with the proteins from seed lists by ranking intermediates according to their specificity to interact with the seed list. Genes2Networks is suitable for analysis of diverse high-content multivariate experimental results. The web interface and visualization provide easy access and user friendly environment eliminating the need for training.

Availability and requirements

Project name: Genes2Networks
Project home page: http://actin.pharm.mssm.edu/genes2networks/
Operating system: Platform independent
Programming language: C, AJAX, Perl, HTML
Other requirements: The HPRD and PPID dataset are only available to Mount Sinai School of Medicine users due to licensing restrictions.
License: GNU GPL
Any restrictions to use by non-academics: License needed. Users should contact technology@mssm.edu

Acknowledgements

This research is supported by NIH Grant No. GM-072853 and an advanced center grant from NYSTAR to Ravi Iyengar.

Author contributions

AM designed and supervised the study, wrote the manuscript, and implemented the significant intermediates statistical algorithm, and the initial Genes2Network prototype. JMP re-implemented and upgraded the tool that merges and filters the datasets, as well as the tool to construct subnetworks from lists of gene names. SB implemented and designed the web interface and the AVIS visualization tool.

References

1. Ma'ayan A, Blitzer RD, Iyengar R: **TOWARD PREDICTIVE MODELS OF MAMMALIAN CELLS**. *Annual Review of Biophysics and Biomolecular Structure* 2005, **34**(1):319-349.
2. Fields S, Song O-k: **A novel genetic system to detect proteinÂ–protein interactions**. 1989, **340**(6230):245-246.
3. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome**. *PNAS* 2001, **98**(8):4569-4574.
4. Brown PO, Botstein D: **Exploring the new world of the genome with DNA microarrays**. *Nat Genet* 1999, **21**:33 - 37.
5. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: **Expression profiling using cDNA microarrays**. *Nat Genet* 1999, **21**:10 - 14.
6. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR: **Direct analysis of protein complexes using mass spectrometry**. 1999, **17**(7):676-682.
7. Nikitin A, Egorov S, Daraselia N, Mazo I: **Pathway studio--the analysis and navigation of molecular networks**. *Bioinformatics* 2003, **19**(16):2155-2157.
8. Marcotte EM, Xenarios I, Eisenberg D: **Mining literature for protein-protein interactions**. *Bioinformatics* 2001, **17**(4):359-363.
9. Barabsi A-L, Albert R: **Emergence of Scaling in Random Networks**. *Science* 1999, **286**(5439):509-512.
10. Watts DJ, Strogatz SH: **Collective dynamics of `small-world' networks**. *Nature* 1998, **393**(6684):440-442.
11. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network Motifs: Simple Building Blocks of Complex Networks**. *Science* 2002, **298**(5594):824-827.
12. Przulj N, Corneil DG, Jurisica I: **Efficient estimation of graphlet frequency distributions in protein-protein interaction networks**. *Bioinformatics* 2006, **22**(8):974-980.
13. Albert I, Albert R: **Conserved network motifs allow protein-protein interaction prediction**. *Bioinformatics* 2004, **20**(18):3346-3352.
14. Yu H, Paccanaro A, Trifonov V, Gerstein M: **Predicting interactions in protein networks by completing defective cliques**. *Bioinformatics* 2006, **22**(7):823-829.
15. Eungdamrong NJ, Iyengar, R.: **Computational approaches for modeling regulatory cellular networks**. *Trends in Cell Biology* 2004, **14**(12):661-669.
16. Ulitsky I, Shamir R: **Identification of functional modules using network topology and high-throughput data**. *BMC Systems Biology* 2007, **1**(1):8.
17. Bader GD, Betel D, Hogue CWV: **BIND: the Biomolecular Interaction Network Database**. *Nucl Acids Res* 2003, **31**(1):248-250.
18. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM *et al*: **Human protein reference database--2006 update**. *Nucl Acids Res* 2006, **34**(suppl_1):D411-414.

19. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C *et al*: **The HUPO PSI's Molecular Interaction format[mdash]a community standard for the representation of protein interaction data**. 2004, **22**(2):177-183.
20. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D: **DIP: the Database of Interacting Proteins**. *Nucl Acids Res* 2000, **28**(1):289-291.
21. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database**. *FEBS Letters*
Protein Domains 2002, **513**(1):135-140.
22. Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N *et al*: **Towards a proteome-scale map of the human protein-protein interaction network**. 2005, **437**(7062):1173-1178.
23. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S: **A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome**. *Cell* 2005, **122**(6):957-968.
24. Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, Dubin-Thaler B, Eungdamrong NJ, Weng G, Ram PT, Rice JJ *et al*: **Formation of Regulatory Patterns During Signal Propagation in a Mammalian Cellular Network**. *Science* 2005, **309**(5737):1078-1083.
25. Beuming T, Skrabanek L, Niv MY, Mukherjee P, Weinstein H: **PDZBase: a protein-protein interaction database for PDZ-domains**. *Bioinformatics* 2005, **21**(6):827-828.
26. Grant SG: **Systems biology in neuroscience: bridging genes to cognition**. *Current Opinion in Neurobiology* 2003, **13**(5):577-582.
27. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I *et al*: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**. *Nucl Acids Res* 2003, **31**(1):365-370.
28. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank**. *Nucl Acids Res* 2006, **34**(suppl_1):D16-20.
29. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions**. 2002, **417**(6887):399-403.
30. Boser BE, Guyon, I. M. , Vapnik, V. N.: **A training algorithm for optimal margin classifiers**. Pittsburgh; 1992.
31. Rosner B: **Fundamentals of biostatistics**. Pacific Grove, CA: Duxbury; 2000.
32. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks**. *Genome Res* 2003, **13**(11):2498-2504.
33. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks**. *Bioinformatics* 2005, **21**(16):3448-3449.

34. Cerami EG, Bader GD, Gross B, Sander C: **cPath: open source software for collecting, storing, and querying biological pathways**. *BMC Bioinformatics* 2006, **7**:497.
35. Aragues R, Jaeggi D, Oliva B: **PIANA: protein interactions and network analysis**. *Bioinformatics* 2006, **22**(8):1015-1017.
36. Iragne F, Nikolski M, Mathieu B, Auber D, Sherman D: **ProViz: protein interaction visualization and exploration**. *Bioinformatics* 2005, **21**(2):272-274.
37. Dogrusoz U, Erson EZ, Giral E, Demir E, Babur O, Cetintas A, Colak R: **PATIKAweb: a Web interface for analyzing biological pathways through advanced querying and visualization**. *Bioinformatics* 2006, **22**(3):374-375.

Figures

Figure 1

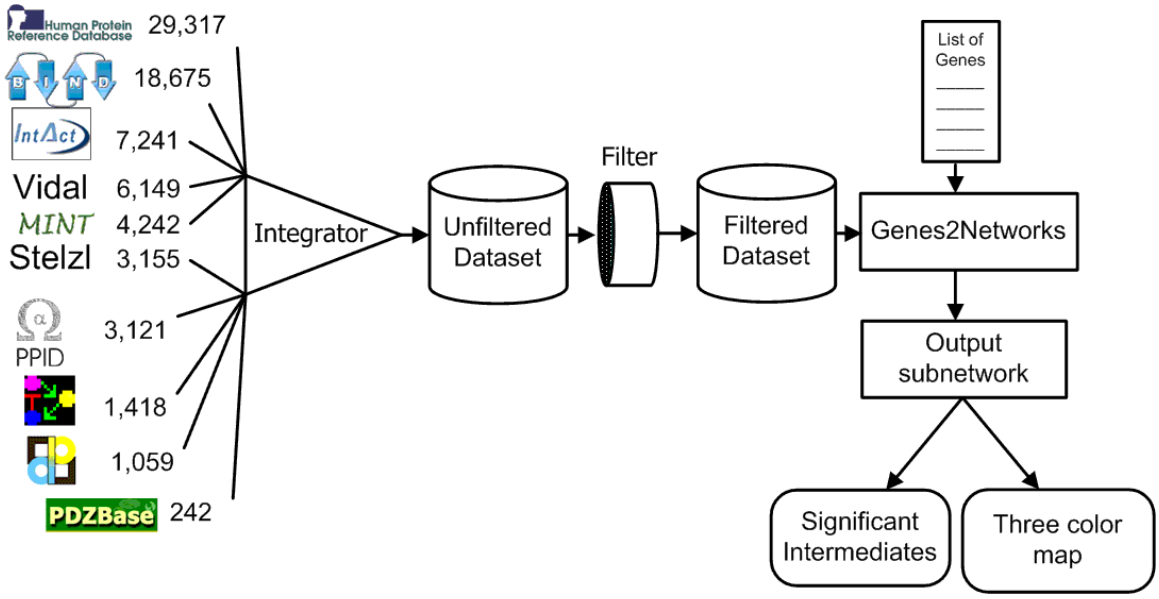

Figure 1: Ten mammalian PPI network datasets were consolidated into one dataset, and then filtered by excluding interactions originating from articles that contributed many interactions, or by excluding interactions with few references. The filtered merged dataset is then used to analyze lists of gene or protein names by outputting a subnetwork with nodes in three different colors: seed, significant, insignificant. The output also includes a statistical report that ranks intermediate nodes based on their specificity to interact with the seed list.

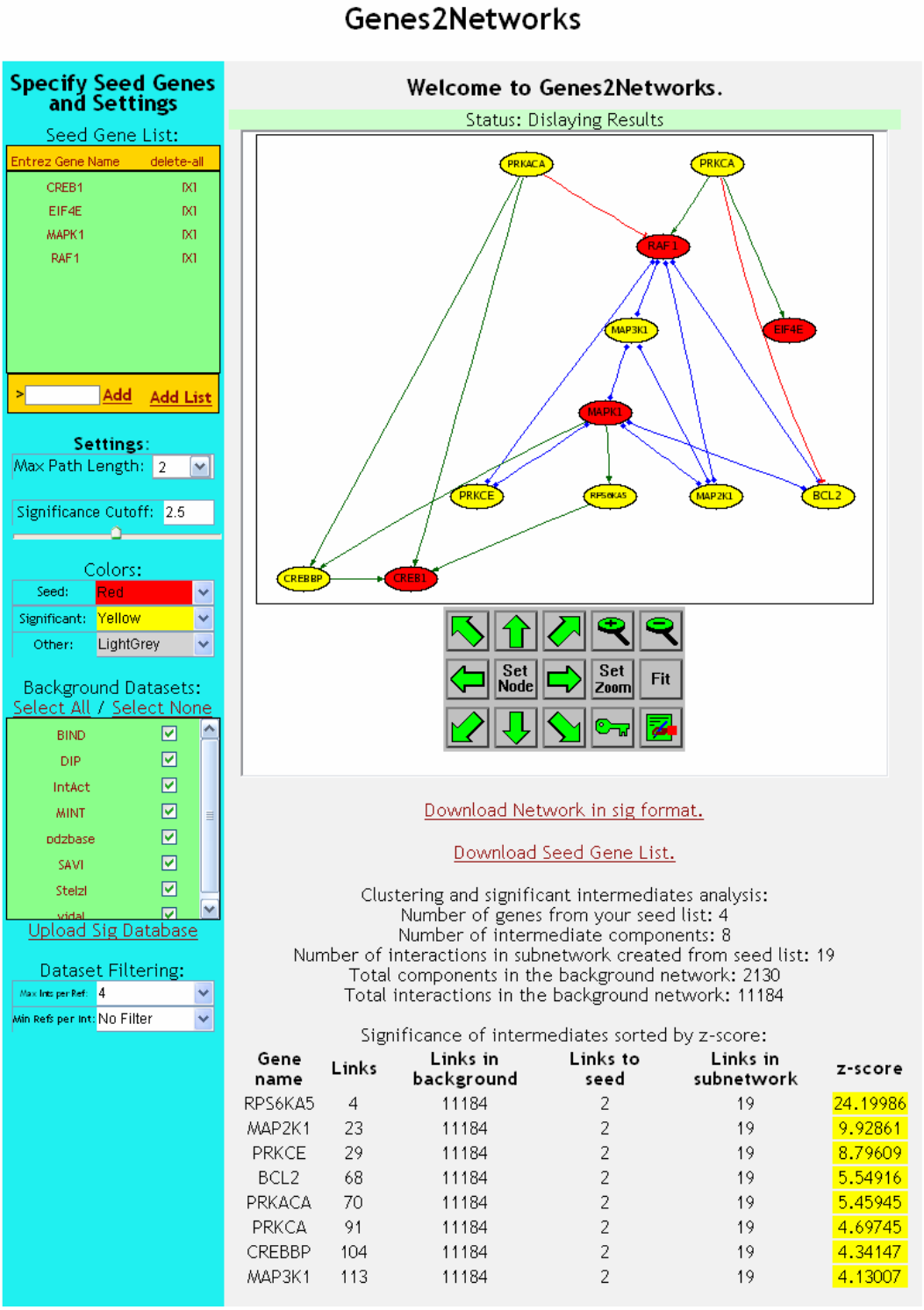

Figure 2: Genes2Networks web interface. The interface allows users to input a list of human Entrez Gene symbols, entered in a textbox or through a text file (top left). As genes are added, using the background merged consolidated network, the program outputs a network map that visualize known interactions that “connect” the list of gene symbols, and a statistical report that ranks intermediates based on their specificity to interact with the seed list.