COMMITTEE ON TOXICITY OF CHEMICALS IN FOOD CONSUMER PRODUCTS AND THE ENVIRONMENT (COT)

Pathway Analysis Software for the Interpretation of Complex Datasets

General Introduction

In the past decade there has been an explosion of high-content and high-throughput data associated with a large number of disease states, chemical exposures and biological species. To fully interpret this information it has become necessary to develop a range of software tools that will identify the potentially biologically important patterns within a given set of data, and present it in a context that is both understandable to non specialists, and searchable so that the data underlying the constructed networks can be viewed and assessed. To this end a number of analysis tools have been developed, and this paper will provide a short overview of the most commonly used approaches, using examples of both commercial and open access software suites.

General Concepts

Literature Mining

An important source of information available to any researcher is peer-reviewed literature. Much work has thus been undertaken to develop automated systems for analysis of this large, and varied, dataset. Such approaches have been designed for general extraction of information from biological texts (Barnickel et al., 2009) or to address targeted questions, such as compound profiling (Frijters et al., 2007) or disease-specific drug-protein interactions (Li et al., 2009). In general, these algorithms extract up to three different levels of data: First, the appearance of two keywords within an abstract will generate an association, but one that has no directionality. Second, information on the relationship between the two keywords may exist; for example ‘x induces expression of y’, which gives directionality to the interaction. Third, more in-depth descriptors for the interaction may be present, such as ‘x induces expression of y by three-fold’. Robust text mining lies at the heart of many pathway analysis suites, and as such its continued development is central to the development of analysis software as well.

Over-representation Analysis

The level of representation of any pathway within a given data set can be calculated using the hypergeometric distribution of the data, essentially calculating the probability that the number of genes contained within a given pathway that are observed within a dataset would have occurred by random chance. Usually, Fishers exact test is used to compare the number of targets from a specific pathway within the dataset, with the total number of targets contained within that pathway and the

total number of targets analysed. The resulting probability value is a function of the level of representation of that pathway in the data set, with a lower p value indicating a greater probability that the particular pathway is over-represented. Such tests are the basis for data mining within complex datasets and the identification of potentially altered biological functionality.

Gene Ontology

The GO knowledge database is recognised as the international standard for the annotation of genes [1] (Ashburner et al., 2000). The main GO terms are as follows; BP (Biological process); CC (Cellular Component); MF (Molecular Function). Each term is sub-divided into five levels, representing a hierarchy of information detail from the broadest gene list coverage with lowest specificity, to level 5 annotations, which have the lowest gene list coverage but the highest specificity of categorization. In complex GO algorithms a gene may be associated with multiple GO terms at a base level, but as the gene is classified further up the hierarchy the number of associated GO terms decreases; it is hence probable that genes with shared higher-order GO terms will integrate into a biologically relevant network. The database is constantly updated, and recently text-mining approaches have been applied to increase the coverage of GO term assignment (Van Auken et al., 2009).

Commonly Used Databases

To facilitate network construction, analysis tools use remote access to a large number of online databases. At the simplest level these allow the interpretation of codes from microarray datasets, most commonly the Affymetrix [2] or Illumina [3] platforms. However, once such data has been imported into an analysis tool it must be interpreted and networks constructed. To achieve this multiple databases are used, either as part of the initial literature mining sweep to populate the database with information, or for interrogation of networks by users once a network has been constructed, allowing visualisation of the data underneath each indicated interaction.

Commonly used databases for network analysis are presented in table 1.

---

1 http://www.geneontology.org/

2 http://www.affymetrix.com/

3 http://www.illumina.com/

Database	Description	Link
Entrez PubMed	Published Scientific Literature	http://www.ncbi.nlm.nih.gov/pubmed/
DrugBank	Drug-target interactions	http://www.drugbank.ca/
LIGAND	Small molecular-target interactions	http://ligand.info/
Entrez OMIM	Human genes and genetic phenotypes	http://www.ncbi.nlm.nih.gov/omim
KEGG	Kyoto Enyclopedia of Genes and Genomes	http://www.genome.jp/kegg/
dbSNP	Single Nucleotide Polymorphism Database	http://www.ncbi.nlm.nih.gov/projects/SNP/
miRBase	miRNA sequences and targets	http://www.mirbase.org/
Argonaute2	miRNA sequences and targets	http://www.ma.uni-heidelberg.de/apps/zmf/argonaute/
UniProt	Protein Sequence and Function	http://www.uniprot.org/
HMDB	Human Metabolome Database	http://www.hmdb.ca/
IPI	International Protein Index	http://www.ebi.ac.uk/IPI/IPIhelp.html
ClinicalTrials.gov	Clinical trial information from over 170 countries	http://clinicaltrials.gov/
BioCarta	Repository of graphic representation of pathways	http://www.biocarta.com/Default.aspx
BIND	Biomolecular Interaction Network Database	http://www.bind.ca/
MINT	Molecular Interaction Database	http://mint.bio.uniroma2.it/mint/Welcome.do

Table 1: Commonly used Databases for Pathway Analysis

Analysis Tools

Two levels of analysis for complex datasets may be envisaged; first, those that rely purely upon pre-ascribed annotation for genes (such as GO terms) and allow the identification of pathways enriched within a given dataset. Second, visualisation tools that are more free-form and use text-mining approaches to gather interaction information from a wider range of information and present it in a user-friendly network map.

Pathway Identification

DAVID

DAVID (Database for Annotation, Visualization and Integration Discovery[4]) is a web-based software suite designed to categorize complex, high content, genomic and proteomic data sets (Dennis et al., 2003; Huang et al., 2009). It comprises a set of functional annotation tools allowing scientists to examine biological meaning behind large list of genes, with common outputs from DAVID including identification of over-represented biological pathways; visualisation of gene lists overlaid onto BioCarta and KEGG pathway maps; identification of interacting proteins; identify gene-disease associations.

An advantage of DAVID is its open-source nature and relative rapidity of use. However, this advantage is offset to some degree by the less intuitive output presented by the programme when compared to pathway visualisation tools such as Cytoscape and Ingenuity Pathway Analysis. This concern has been mitigated slightly by the recent publication of a full protocol for use of the software suite (Huang et al., 2009), although it should still be viewed as a potential limitation for ease of use by the non-specialist.

Example outputs form DAVID are shown in Figure 1, including identification of GO-enriched terms, functional classification of enriched target genes and 2D heatmaps of gene-term associations,

Pathway Visualisation

Cytoscape

Cytoscape[5] is the leading open source software project for integrating biomolecular interaction networks with high-throughput expression data (Shannon et al., 2003; Cline et al., 2007). It uses a graphical front end to allow the visualisation of networks in a cellular context, incorporating expression data from microarray experiments and allowing linkage to online databases such as

4 http://david.abcc.ncifcrf.gov/

5 http://www.cytoscape.org/

PubMed. The strongest attribute of Cytoscape is its ability to produce publication quality networks from experimental data, allowing users to visualise not only how different transcripts/proteins links together to form a biological network, but also how the expression levels of these components alter under different experimental conditions through the overlay of single-source array datasets. Such an approach allows both a visual and statistical identification of pathways/networks likely to be impacted by, for example, chemical exposure (Rocke et al., 2009). Finally, interactive text-mining of databases such as PubMed can be used to annotate and expand networks using published literature.

Ingenuity Pathway Analysis

Ingenuity Pathway Analysis[6] is the leading commercial software suite for integrating biomolecular interaction networks with high-throughput expression data. It has been used for a large number of analysis', including liver (Fukushima et al., 2006; Lambert et al., 2009), dermal (Gerecke et al., 2009), lung (Abdel-Aziz et al., 2008) and gastric mucosal (Naito et al., 2007) toxicity. In addition, biological models of disease states (Sharma et al., 2009) and biomarker identification (Gunawardana et al., 2009) have also been investigated using Ingenuity Pathway Analysis Software.

As with Cytoscape, the base output of Ingenuity is a generated pathway showing potential network interactions, which is overlayed with links to all datasets/databases used to generate the network (figure 3A). Importantly, multiple datasets can be analysed at one time, for example transcript and proteomic data, increasing the robustness of any conclusions drawn (Figure 3B). An additional potential benefit of Ingenuity Pathway Analysis over Cytoscape is the ability to search for potential biomarkers within a generated network. However, it should be noted that this, relatively new, feature has not been sufficiently tested to determine how successful it is at selecting robust biomarkers.

Limitations of Analysis Software.

Although all of the abovementioned approaches will provide a wealth of information, it is important to realise that a number of limitations exist to the analysis, all of which need to be taken into account when interpreting any given analysis.

1. **GO Term Assignment:** Association into pathways based upon GO terms, such as used by DAVID, is highly dependent upon the GO term(s) associated with each target gene. Figure # shows the GO terms associated with CYP3A4, and while it can be seen that the terms do describe its major function as a mixed-function oxidase, they do not perhaps clearly describe it as a scientist might. Hence, potential associations may be missed

---

6 http://www.ingenuity.com/

2. **Other Factors Regulating Pathway Flux**: Over-representation of a pathway in a dataset is not indicative that flux through a pathway is increased. Other factors, such as post-translational modifications or control coefficients (Groen et al., 1982) for each step within a pathway are not taken into account, and these may have significant impact on the overall biological activity of a pathway.

3. **Coverage of databases:** The degree of dataset coverage within any individual database varies greatly. For a given microarray dataset, coverage of differentially expressed targets within different databases varies from good (>80%) to poor (<20%). As coverage of a dataset within any given database decreases then so does the confidence of analysis using said database. In general, coverage is best at the gene and transcript level, while coverage of proteomic and metabolomic interactions are the poorest.

4. **Accuracy of Literature Mining:** All text-mining approaches are limited by the robustness of their algorithms. The information that is generated my range from in accurate (e.g. two words in an abstract that actually are not related), to accurate (e.g. a correct identification of an interaction), to a useful description (e.g. the relationship between the interactors).

Conclusions

In the past few years the availability and robustness of pathway analysis tools has improved dramatically. It is now possible for the non-specialist scientist to freely use software and rapidly produce a network of potential interactions based around a query; as such this makes these software tools extremely powerful. However, it should also be noted that there are a number of important caveats associated with their use, meaning that any generated network is, essentially, a statistical probability rather than a certainty. As such, it is important that these tools are seen as starting points for analysis, providing investigative leads, rather than an endpoint. The putative impact of a derived network on biological functionality must ultimately be proved experimentally before it can be fully believed (Plant et al., 2009).

References

Abdel-Aziz HO, Murai Y, Takasaki I, Tabuchi Y, Zheng HC, Nomoto K, Takahashi H, Tsuneyama K, Kato I, Hsu DK, Liu FT, Hiraga K and Takano Y (2008) Targeted disruption of the galectin-3 gene results in decreased susceptibility to NNK-induced lung tumorigenesis: an oligonucleotide microarray study. *Journal of Cancer Research and Clinical Oncology* 134:777-788.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G and Gene Ontology C (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* 25:25-29.

Barnickel T, Weston J, Collobert R, Mewes H-W and Stumpflen V (2009) Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PLoS One* 4:e6393.

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T and Bader GD (2007) Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols* 2:2366-2382.

Dennis G, Sherman DT, Hosack DA, Yang J, Gao W, Lane HC and Lempicki RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* 4:R60.

Frijters R, Verhoeven S, Alkema W, Van Schaik RH and Polman J (2007) Literature-based compound profiling: application to toxicogenomics. *Pharmacogenomics.* 8:1521-1534.

Fukushima T, Kikkawa R, Hamada Y and Horii I (2006) Genomic cluster and network analysis for predictive screening for hepatotoxicity. *Journal of Toxicological Sciences* 31:419-432.

Gerecke DR, Chen M, Isukapalli SS, Gordon MK, Chang Y-C, Tong W, Androulakis IP and Georgopoulos PG (2009) Differential gene expression profiling of mouse skin after sulfur mustard exposure: Extended time response and inhibitor effect. *Toxicol Appl Pharmacol* 234:156-165.

Groen AK, Wanders RJA, Westerhoff HV, Vandermeer R and Tager JM (1982) Quantification of the contribution of various steps to the control of mitochondrial respiration. *Journal of Biological Chemistry* 257:2754-2757.

Gunawardana CG, Kuk C, Smith CR, Batruch I, Soosaipillai A and Diamandis EP (2009) Comprehensive analysis of conditioned media from ovarian cancer cell lines identifies novel candidate markers of epithelial ovarian cancer. *J Proteome Res* 8:4705-4713.

Huang DW, Sherman BT and Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* 4:44-57.

Lambert CB, Spire C, Renaud MP, Claude N and Guillouzo A (2009) Reproducible chemical-induced changes in gene expression profiles in human hepatoma HepaRG cells under various experimental conditions. *Toxicology In Vitro* 23:466-475.

Li J, Zhu XY and Chen JY (2009) Building Disease-Specific Drug-Protein Connectivity Maps from Molecular Interaction Networks and PubMed Abstracts. *Plos Computational Biology* 5.

Naito Y, Kuroda M, Mizushima K, Takagi T, Handa O, Kokura S, Yoshida N, Ichikawa H and Yoshikawa T (2007) Transcriptome analysis for cytoprotective actions of rebamipide against indomethacin-induced gastric mucosal injury in rats. *Journal of Clinical Biochemistry and Nutrition* 41:202-210.

Plant KE, Anderson E, Simecek N, Brown R, Forster S, Spinks J, Toms N, Gibson GG, Lyon J and Plant N (2009) The neuroprotective action of the mood stabilizing drugs lithium chloride and sodium valproate is mediated through the up-regulation of the homeodomain protein Six1. *Toxicology and Applied Pharmacology* 235:124-134.

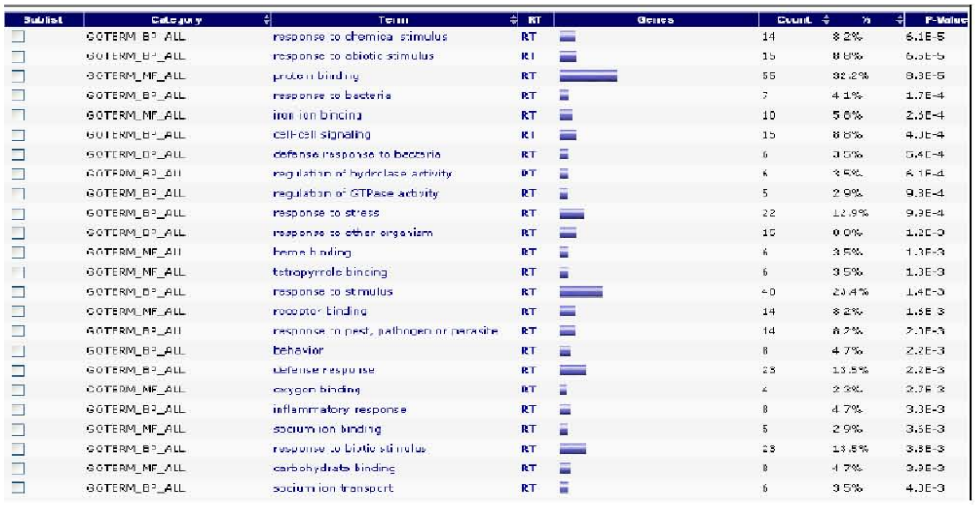
Rocke DM, Ideker T, Troyanskaya O, Quackenbush J and Dopazo J (2009) Papers on normalization, variable selection, classification or clustering of microarray data. *Bioinformatics* 25:701-702.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B and Ideker T (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research* 13:2498-2504.

Sharma AK, Searfoss GH, Reams RY, Jordan WH, Snyder PW, Chiang AY, Jolly RA and Ryan TP (2009) Kainic acid-induced F-344 rat model of mesial temporal lobe epilepsy: gene expression and canonical pathways. *Toxicol Pathol* 37:776-789.

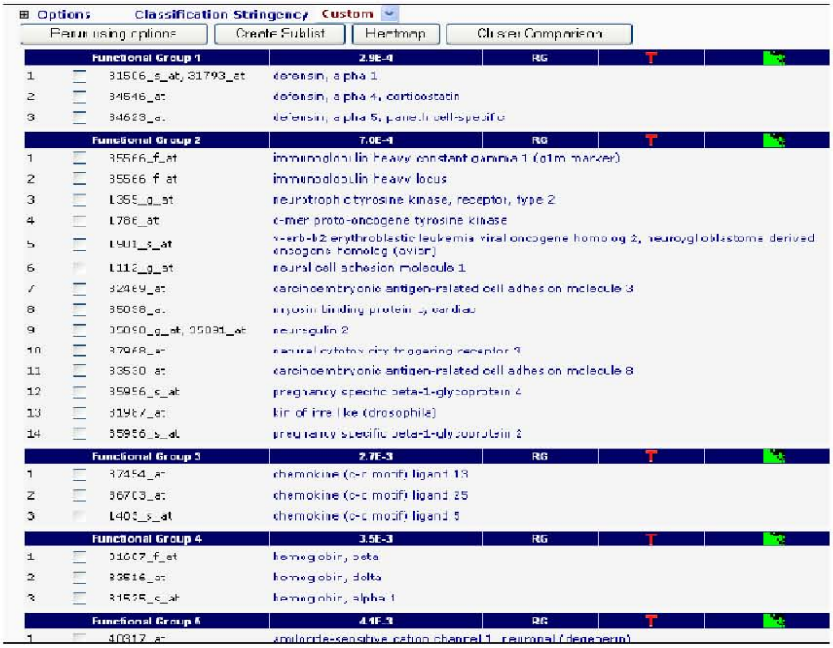
Van Auken K, Jaffery J, Chan J, Muller HM and Sternberg PW (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. *BMC Bioinformatics* 10.

A



Sublist	Category	Term	RT	Genes	Count	%	P-Value
	GOTERM_BP_ALL	response to chemical stimulus	RT		14	8.2%	6.1E-5
	GOTERM_BP_ALL	response to abiotic stimulus	RT		15	8.8%	6.5E-5
	GOTERM_MF_ALL	protein binding	RT		55	32.2%	9.3E-5
	GOTERM_BP_ALL	response to bacteria	RT		7	4.1%	1.7E-4
	GOTERM_MF_ALL	iron ion binding	RT		10	5.8%	2.6E-4
	GOTERM_BP_ALL	cell-cell signaling	RT		15	8.8%	4.1E-4
	GOTERM_BP_ALL	defense response to bacteria	RT		6	3.5%	5.4E-4
	GOTERM_BP_ALL	regulation of hydrolase activity	RT		6	3.5%	6.1E-4
	GOTERM_BP_ALL	regulation of GTPase activity	RT		5	2.9%	9.3E-4
	GOTERM_BP_ALL	response to stress	RT		22	12.9%	9.3E-4
	GOTERM_BP_ALL	response to other organism	RT		16	9.4%	1.2E-3
	GOTERM_MF_ALL	heme binding	RT		6	3.5%	1.3E-3
	GOTERM_MF_ALL	tetrapyrrole binding	RT		6	3.5%	1.3E-3
	GOTERM_BP_ALL	response to stimulus	RT		40	23.4%	1.4E-3
	GOTERM_MF_ALL	receptor binding	RT		14	8.2%	1.8E-3
	GOTERM_BP_ALL	response to pest, pathogen or parasite	RT		14	8.2%	2.1E-3
	GOTERM_BP_ALL	behavior	RT		8	4.7%	2.2E-3
	GOTERM_BP_ALL	defense response	RT		23	13.5%	2.2E-3
	GOTERM_MF_ALL	oxygen binding	RT		4	2.3%	2.7E-3
	GOTERM_BP_ALL	inflammatory response	RT		8	4.7%	3.3E-3
	GOTERM_MF_ALL	sodium ion binding	RT		5	2.9%	3.5E-3
	GOTERM_BP_ALL	response to biotic stimulus	RT		23	13.5%	3.8E-3
	GOTERM_MF_ALL	carbohydrate binding	RT		8	4.7%	3.9E-3
	GOTERM_BP_ALL	sodium ion transport	RT		6	3.5%	4.0E-3

B



C

Figure 1: Examples of DAVID 2008 Output
Screen shots from DAVID analysis of a microarray dataset, demonstrating (A) Identification of over-represented GO terms; (B) Functional classification of over-represented biological pathways; (C) He maps of gene-GO term associations

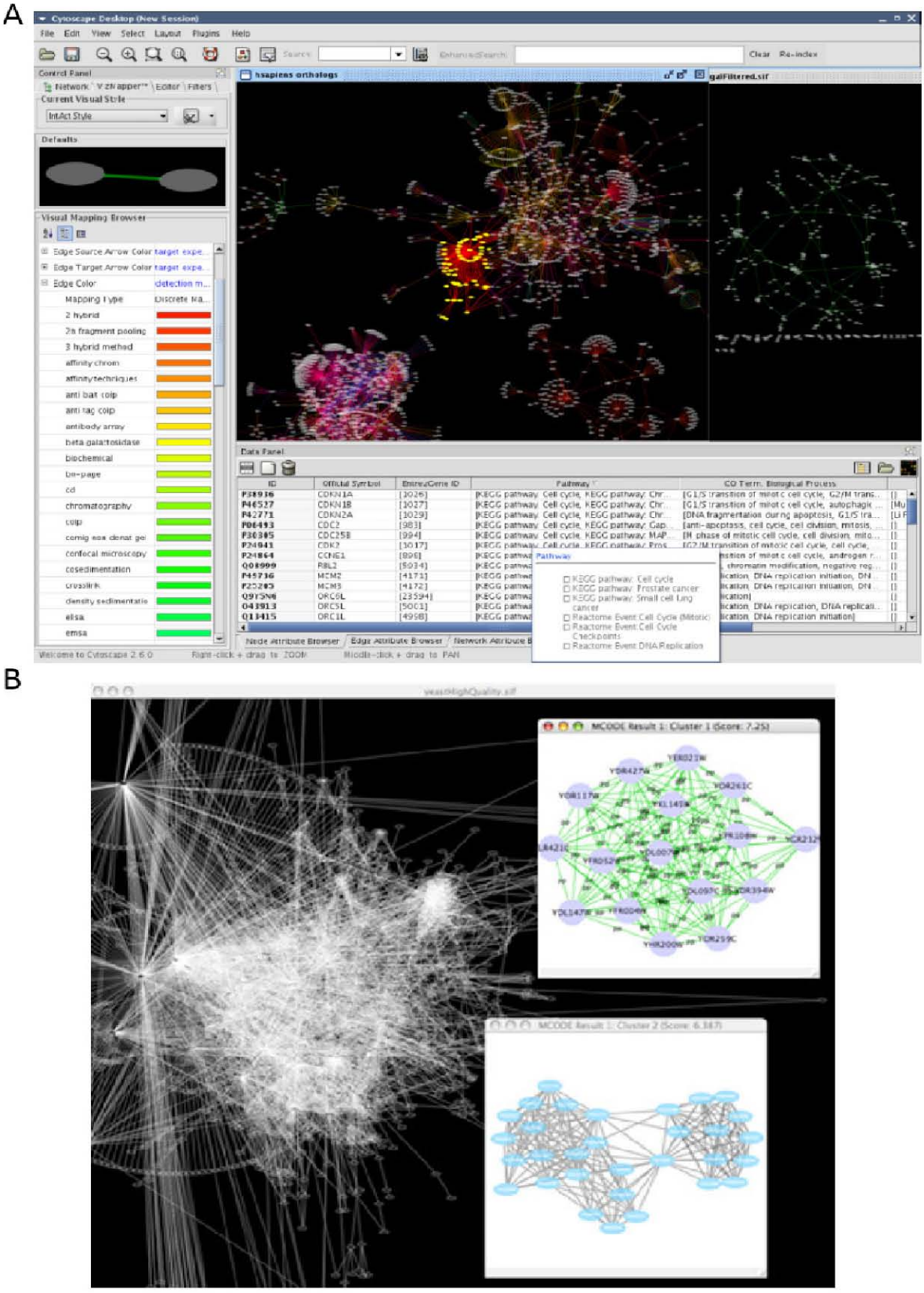
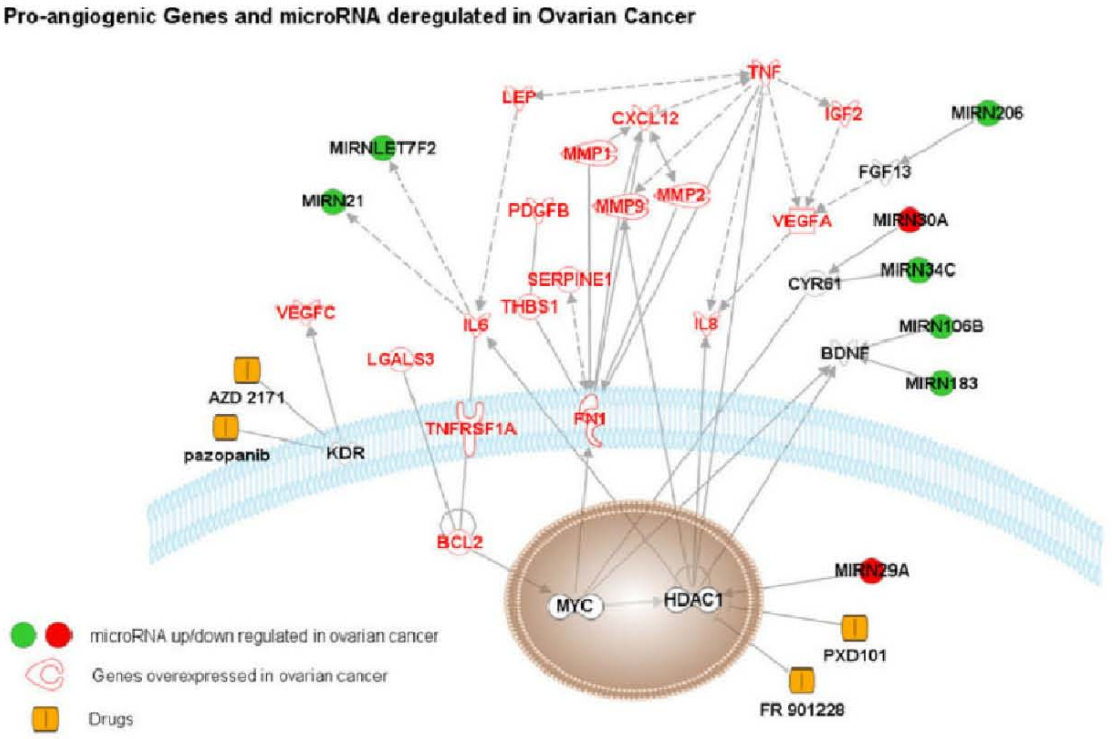


Figure 2: Examples of Cytoscape (v2.6) Output.
(A) Constructed biological network based upon microarray dataset analysis; (B) Alternate network views, providing different levels of layer data view

A



B

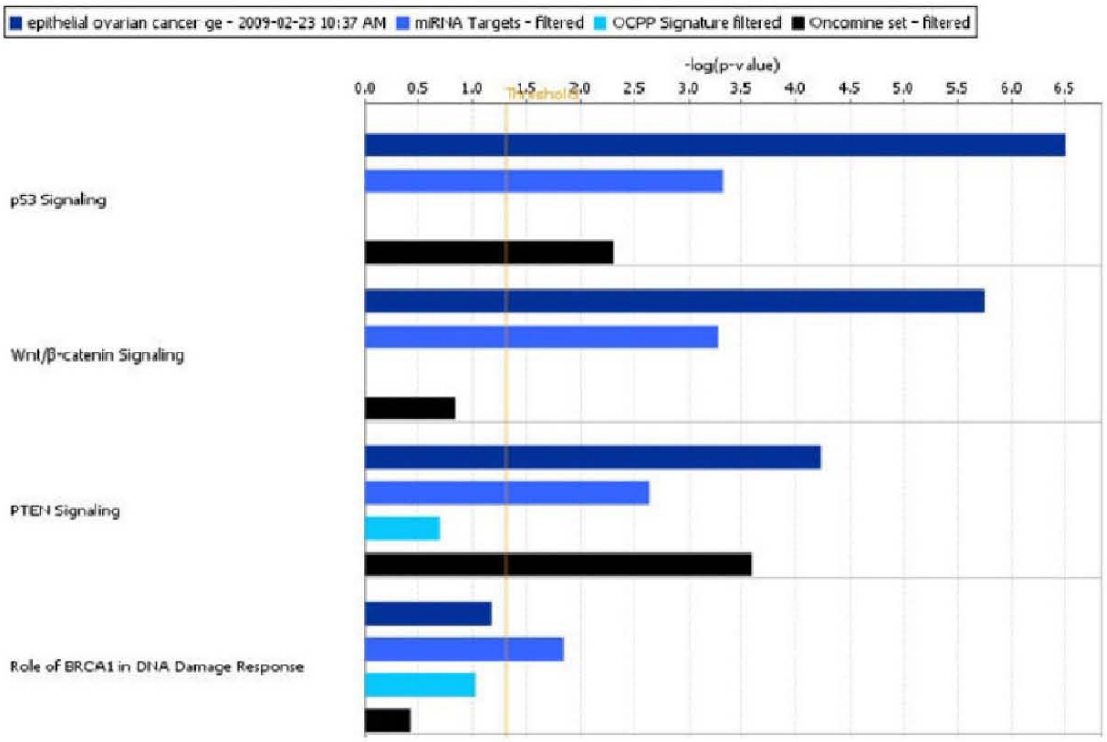


Figure 3: Examples of Ingenuity Pathway Analysis (v 7.6) Output.
(A) Constructed biological network based upon microarray dataset analysis; (B) Comparison of pathway over-representation from multiple data sources