ConceptGen: A Gene Set Enrichment and Concept Mapping Tool

Vasudeva Mahavisno, Zach Wright, Alla Karnovsky, Gilbert S. Omenn, Brian Athey, James Cavalcoli, and Maureen A. Sartor
National Center for Integrative Biomedical Informatics, University of Michigan; Ann Arbor, MI

Abstract

Identification of biological concepts enriched in an experimentally-derived gene list has become an integral part of the analysis and interpretation of genomic data. Of additional importance is the ability to explore *networks* of relationships among previously defined biological concepts from diverse information sources. We will present for the first time ConceptGen, a gene set enrichment and concept mapping tool that integrates gene sets from 14 biological knowledge sources totaling ~ 18,000 concepts and provides a user friendly web interface. The experimentally-derived concepts include several hundred from public microarray datasets downloaded from Gene Expression Omnibus (GEO), which we analyzed using a custom-built gene expression analysis pipeline incorporating advanced statistical methods and quality control checks. Additional concept types include Gene Ontologies, pathway databases, protein domain families, miRNA target sets, drug target sets, gene-centered protein interaction sets, MeSH-derived concepts, and metabolite-specific gene sets created using published human metabolic networks that link compounds and reactions to enzymes and genes. ConceptGen can easily be expanded to include experimental data from other technologies, such as ChIP-Seq, RNA-Seq, or high-throughput metabolomics and proteomics.

Using a modified Fisher's Exact Test, we pre-computed the significance of overlap among all concepts, and developed a state-of-the-art user interface with Flex technology. Visualizations include a network and heat-map view of significantly enriched concepts. Users are provided with private accounts for uploading gene or metabolomic datasets. We demonstrate the usefulness of ConceptGen using a bipolar disorder case study.

Introduction

A **Fisher's exact test modified** to increase stability is performed for each biological grouping (concept)

	# signif	# not signif
# in category	A - 1	B
# not in category	C	D

Hypergeometric Distribution:

$$p = \frac{(A-1+B)!(C+D)!(A-1+C)!(B+D)!}{(A-1)!B!C!D!(A-1+B+C+D)!}$$

	ConceptGen	Oncomine concept mapping	DAVID/ EASE
Freely-available?	Yes	Very limited	Yes
Performs concept mapping?	Yes	Yes	No
Contains experimental microarray data?	Yes	Limited to cancer	No
Uses modified Fisher's Exact for stability?	Yes	No	Yes
Private account ?	Yes	Yes, but not free version	No
Heatmap view of network?	Yes	No	No
Metabolite, MeSH, or MiMI protein interactions?	Yes	No	No

Concept Building

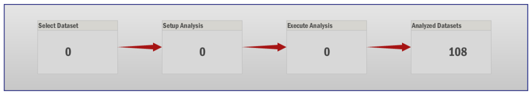
Biological Knowledge types represented in ConceptGen

Biological Knowledge Type	Concept Type(s) Name	# of Concepts	Source
Biological processes	GO biological process	1043	Gene Ontology
Molecular functions	GO molecular function	660	Gene Ontology
Cellular components	GO cellular component	292	Gene Ontology
Protein-centered interactions	Protein Interaction (MiMI)	7050	NCIBI database
Medical Literature Derived	MeSH (using Gene2MeSH)	4028	NCIBI database
Human Diseases	OMIM	52	NCBI
Drug Targets	Drug Bank	256	Drug Bank
Chromosomal location	Cytoband	1178	NCBI
Molecular Pathways	KEGG pathway; Panther pathway; Biocarta pathway	50	KEGG; Panther; Biocarta
Transcription Factor targets	Transfac	119	Transfac
Protein families	pFAM	770	pFAM
microRNA targets	MIRBase	587	MIRBase
Metabolic interactions	Metabolite	960	NCIBI database
Differential expression profiles	Gene Expression	603	Custom-built analysis pipeline for GEO data

Expression data analysis pipeline

Protocol:

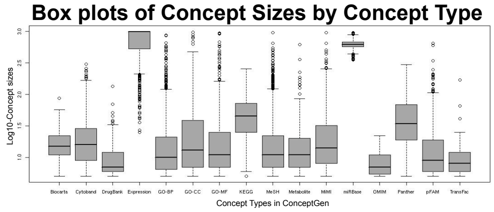
1. Affymetrix raw data downloaded directly from NCBI GEO repository
2. Probes mapped to Entrez Gene IDs using custom CDF (Dai, et al., 2005)
3. Data is normalized with RMA, quality control checked, and comparisons set up through pipeline interface
4. Differential expression is tested using an empirical Bayes method (Sartor, et al., 2006)
5. Up, Down, and differentially expressed concepts are created using genes with fold > 20% and p-value < 0.05



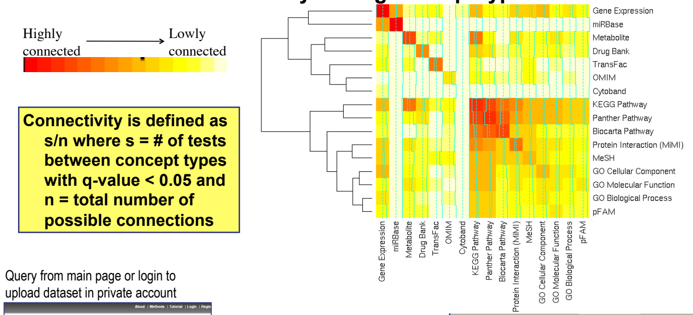
Select Dataset	Setup Analysis	Execute Analysis	Analyzed Datasets
0	0	0	108

Results

Box plots of Concept Sizes by Concept Type

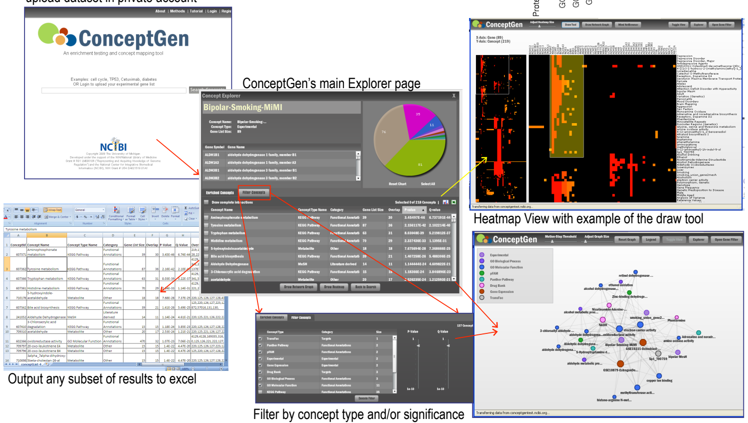


Connectivity among concept types



Connectivity is defined as s/n where s = # of tests between concept types with q-value < 0.05 and n = total number of possible connections

Query from main page or login to upload dataset in private account



Bipolar-Smoking-MiMI

Concept Name: Bipolar-Smoking-...
Concept Type: Experimental
Gene List Size: 89

Gene Symbol	Gene Name
ALDH1B1	aldehyde dehydrogenase 1 family, member B1
ALDH1A3	aldehyde dehydrogenase 1 family, member A3
ALDH3B1	aldehyde dehydrogenase 3 family, member B1
ALDH3B2	aldehyde dehydrogenase 3 family, member B2

Selected 0 of 218 Concepts

Concept Name	Concept Type Name	Category	Gene List Size	Overlap	P-value	Q-value
Aminophosphonate metabolism	KEGG Pathway	Functional Annotati	39	30	3.45497E-46	6.737191E-44
Tyrosine metabolism	KEGG Pathway	Functional Annotati	87	36	2.156117E-42	2.102214E-40
Tryptophan metabolism	KEGG Pathway	Functional Annotati	63	31	8.03048E-39	5.219812E-37
Histidine metabolism	KEGG Pathway	Functional Annotati	70	29	2.337436E-33	1.1395E-31
5-hydroxyindoleacetaldehyde	Metabolite	Other	18	18	7.675694E-28	7.368666E-25
Bile acid biosynthesis	KEGG Pathway	Functional Annotati	39	21	1.407258E-26	5.488306E-25
Aldehyde Dehydrogenase	MeSH	Literature derived	14	11	1.144444E-24	4.609822E-21
3-Chloroacrylic acid degradation	KEGG Pathway	Functional Annotati	15	15	1.18366E-24	3.846896E-23
acetaldehyde	Metabolite	Other	20	17	2.526235E-24	1.212593E-21

137 Concepts

Concept Type	Category	Size
☑ TransFac	Targets	1
☑ Panther Pathway	Functional Annotations	2
☑ pFAM	Functional Annotations	2
☑ Experimental	Experimental	2
☑ Gene Expression	Experimental	2
☑ Drug Bank	Targets	3
☑ GO Biological Process	Functional Annotations	3
☑ GO Molecular Function	Functional Annotations	11
□ KEGG Pathway	Functional Annotations	35

P-Value 1 – 1e-10
Q-Value 1 – 1e-10

X-Axis: Gene (89)
Y-Axis: Concept (219)

ConceptGen's main Explorer page

Heatmap View with example of the draw tool

Output any subset of results to excel

Filter by concept type and/or significance

Network of enriched concepts after filtering as shown to the left.

References

1. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, et al. 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res 33(20): e175.
2. Dennis G, Jr., et al. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4(5): 3.
3. Rhodes DR, Kalyana-Sundaram S, Tomlins SA, Mahavisno V, Kasper N, Varambally R, et al. 2007. Molecular concepts analysis links tumors, pathways, mechanisms, and drugs. Neoplasia 9(5): 443-454.
4. Sartor MA, et al. 2006. Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. BMCBioinformatics 7(1): 538.

Acknowledgements

This work was supported by National Institutes of Health: Grant #U54 DA021519