

# NCIBI Cores 1 and 2 Report

H. V. Jagadish



**National Center for**  
**Integrative**  
**Biomedical Informatics**



*Fourth Annual National Center for Integrative Biomedical Informatics Meeting 2009*



# Integration story

- Many individual tools and data resources from previous years
  - Progress continued this year
    - Improvements to previously reported resources
    - Brand new resources
- Putting these together for seamless use
  - Our own NCIBI resources
  - Other NCBC resources
  - Other national resources

# Individual tools and data resources

- Improvements to previously reported resources
  - Natural Language Processing Pipeline (NLP)
  - MiMI
  - GIN
- New resources
  - ConceptGen
  - ConSig-Score

# NLP Progress

- Structured queriable information extracted from all abstracts in PubMed and full text in PMCOA.
- Foundational resource for many downstream tools.
- Started work on a public API for access to this resource. Expected to be ready later this year.

# PubMed09

- Document Count: 18,075,637 (12 months of 2008) -> 18,334,510 (4 months of 2009)
- Number of Tables: 66 -> 72
- Contains word-level NLP data including part-of-speech, stem, and lemma
- Parsed with a more accurate parser using geneR2 symbols
- Includes SciMiner gene and metabolite tags
- Uses a full MeSH representation including supplemental concepts and mapped ICD9 codes
- Maps between Substances and KEGG IDs
- Includes GIN-IE derived protein interactions and interaction words

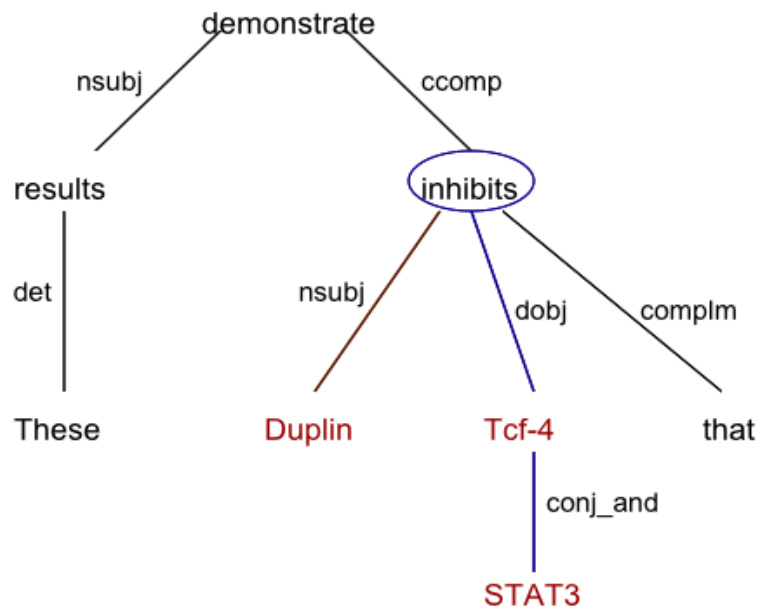
# GIN

- GIN has been divided into two modules:
  - GIN-IE for Information Extraction
  - GIN-NA for Network Analysis
- The GIN-IE processing pipeline has been improved and added to the PubMed processing pipeline and extracted interactions and interaction types are being added to the PubMed database.

# GIN-IE:

## Interaction Extraction from the Literature

- Dependency Tree Rules: Predicate-argument structures, non-local dependencies among the entities in the sentences.



"These results demonstrate that Duplin inhibits not only Tcf-4 but also STAT3."

- Determine:
  - Type of relationship: e.g. "inhibition"
  - Directionality: "Duplin->Tcf-4"
- High precision (94%) at the expense of recall (18%)
- Can handle negation and speculation:
  - The **lack** of cooperative interaction between E5 and the epidermal growth factor receptor...
  - Like RAD9, RAD9B associates with HUS1, **suggesting** that it is a RAD9 paralog (Speculative)



# RSS Feed: Daily Processing of Pubmed Updates

GIN Interactions Extracted from the Literature (PubMed Update: 11/12/2008) 165 Total

**Agent: aurora-c Target: borealin Interaction Type: Interacts (PMID: 18239465)** Dec 4, 09:48 AM  
We have found that Aurora-C interacts with Borealin in addition to the other known members of the Aurora-B chromosomal passenger complex (CPC). [Read more...](#)

**Agent: aurora-c Target: borealin Interaction Type: phosphorylates (PMID: 18239465)** Dec 4, 09:48 AM  
We have also found that Aurora-C, like Aurora-B, phosphorylates the centromeric histone Centromere Protein-A (CENP-A) and Borealin in vitro. [Read more...](#)

**Agent: nsp3 Target: eif4g Interaction Type: Interacts (PMID: 18799579)** Dec 4, 09:12 AM  
Rotavirus nonstructural protein NSP3 interacts specifically with the 3' end of viral mRNAs, with the eukaryotic translation initiation factor eIF4G, and with RoXaN, a cellular protein of yet-unknown function. [Read more...](#)

**Agent: nsp3 Target: roxan Interaction Type: interacts (PMID: 18799579)** Dec 4, 09:12 AM  
Rotavirus nonstructural protein NSP3 interacts specifically with the 3' end of viral mRNAs, with the eukaryotic translation initiation factor eIF4G, and with RoXaN, a cellular protein of yet-unknown function. [Read more...](#)

**Agent: c-abl Target: yap1 Interaction Type: phosphorylated (PMID: 18701449)** Dec 4, 09:02 AM  
In response to DNA damage, Yap1 is phosphorylated by c-Abl at the position Tyr-357. [Read more...](#)

**Agent: fhod1 Target: rac1 Interaction Type: interacts (PMID: 18694941)** Dec 4, 09:02 AM  
The DRF FHOD1 interacts with the Rho-GTPase Rac1 and mediates formation of actin stress fibers in its deregulated form; the physiologically relevant activities and molecular mechanisms of endogenous FHOD1, however, are still unknown. [Read more...](#)

**Agent: mbd1 Target: fgf-2 promoter Interaction Type: bound (PMID: 18689796)** Dec 4, 09:01 AM  
Mbd1 bound to the Fgf-2 promoter and regulates its expression in adult NSPCs. [Read more...](#)

**Agent: nes Target: crm1 Interaction Type: interacts (PMID: 18687685)** Dec 4, 09:00 AM  
Localization, reporter gene, and co-immunoprecipitation assays demonstrate that the identified NES interacts with CRM1 in a phosphorylation-sensitive manner. [Read more...](#)

**Search Articles:**

**Article Length:**  
--  --

**Sort By:**  
[Date](#)  
[Title](#)  
[Source](#)  
[New](#)

**Recent Articles:**  
[All](#)  
[Today](#)  
[Yesterday](#)  
[Last Seven Days](#)  
[This Month](#)  
[Last Month](#)

**Source:**  
GIN Interactions Ext...

**Actions:**  
[Mail Link to This Page](#)  
[Add Bookmark...](#)

# PubOnto: Open Biomedical Ontologies-Based Medline Exploration

The screenshot displays the PubOnto web application interface. At the top, there is a menu bar with 'File', 'Ontology', 'Tools', and 'Help'. Below this, there are tabs for 'FMA Ontology \*', 'Gene On...', 'Phenotype Ontology', and 'Environment Ontology'. The main content area is titled 'Exploring literature search result with Foundational Model of Anatomy (FMA) ontology'. It features a tree view of ontologies: 'Body [188]' (expanded to 'Alimentary system[51]' and 'Cardiovascular system[16]'), with 'Gastrointestinal system[50]' selected. A context menu is open over 'Gastrointestinal system[50]', showing options like 'FMA ID: 71132', 'Map Result to Other Ontologies', 'Goto Wikipedia', 'Goto Pubmed', 'Show Redraw Regions', 'Debugger', 'Settings...', and 'About Adobe Flash Player 9...'. To the right is a 'Search Panel' with a 'Keyword Search' field containing 'phylogenetic', a 'Search' button, an 'Abstract limit' of 500, a 'Search service link' field with the URL 'http://eutils.ncbi.nlm.nih.gov/entrez/eutils/', and an 'Inter-Ontology Filter' checkbox which is checked. Below the search panel is a 'User Profile' section. At the bottom, there are tabs for 'Citations', 'MeSH Profiling', and 'Authors'. The 'Citations' tab is active, showing a table with columns for 'PMID', 'Year', 'Journal', 'Title', and 'Abstract'. The table contains four rows of citation data.

	PMID	Year	Journal	Title	Abstract
<input type="checkbox"/>	17107469	2006	Mol Ecol	Comprehensive genetic analyses reveal evolutionary distincti	Zapus hudsonius preblei, listed a
<input type="checkbox"/>	17107472	2006	Mol Ecol	Phylogeography of the Western Lyresnake (Trimorphodon bis	The Western Lyresnake (Trimorp
<input type="checkbox"/>	17107556	2006	Environ Microbiol	Study of bacterial communities in Antarctic coastal waters by .	An ecological study on distributio
<input type="checkbox"/>	17107559	2006	Environ Microbiol	A new alpha-proteobacterial clade of Bdellovibrio-like predatc	Bdellovibrio-and-like organisms (

# PubAnatomy: Integrated Data and Literature Exploration for Neurobiology

The screenshot displays the PubAnatomy software interface. At the top, there are menu options: File, Anatomy, Tools, and Help. Below the menu, there are three main tabs: Brain Map, Gene Expression Correlation, and Image Section. The central area shows two brain maps with a color scale from 0.00 (light blue) to 1.79 (dark purple). The left map is labeled 'Brain Map' and the right map is labeled 'Image Section'. A yellow box highlights the 'Hippocampal region' on the left map. The right map shows a similar view with a different color scale. On the right side, there is a 'Search Panel' with a search bar containing 'bipolar disorder'. Below the search bar, there are options for 'Search Engine Options' (set to 'PubAnatomy search engine'), 'Abstract search range' (set to 500), and 'Mapping Options' (set to 'Brain Structure Term Mapping'). Below the search panel, there are 'Map Options' and 'User Input' sections. At the bottom, there is a table with columns: Citation, Gene, MeSH Profiling, Author, Disease, In Situ Image, MiMI Interac..., Expression C..., Gene Express..., Significant G..., and Import. The table is filtered by 'MESH:Hippocampus' and shows 130/130 citations. The table contains four rows of data with columns for PMID, Year, Journal, Title, and Abstract.

Citation	Gene	MeSH Profiling	Author	Disease	In Situ Image	MiMI Interac...	Expression C...	Gene Express...	Significant G...	Import	
Filter	MESH:Hippocampus										
	PMID	Year	Journal	Title	Abstract					Reset	130/130 citation
<input type="checkbox"/>	18191545	2008	Schizophr Res	Dysregulation of glutamate carboxypeptidase II in psychiatric	Experimental evidence is beginning to converge on an important role f						
<input type="checkbox"/>	19022630	2008	Psychiatry Res	The volumes of the fornix in schizophrenia and affective diso	Structural and functional pathology of limbic structures including the hij						
<input type="checkbox"/>	18980734	2008	Curr Psychiatry Rep	Limbic changes identified by imaging in bipolar patients.	The hippocampus and amygdala are key limbic regions for memory fo						
<input type="checkbox"/>	18590792	2008	Prog Neuropsychop	Dose-dependent effect of intracerebroventricular injection of	Intracerebroventricular (ICV) injection of ouabain, a specific Na-K ATPa						

# MiMI (Michigan Molecular Interactions)

- Started out with integration of public molecular interaction databases.
- Added gene level data
- Added pathway information from KEGG and Reactome
- Added metabolic reactions
- Created a vertically integrated data resource that goes across levels.
- Can be accessed programmatically, via the web, or from Cytoscape.

# MimiWeb

- A new version of MimiWeb was released, changes include:
  - Usability updates from User Studies
  - Reactome Pathway mapping
  - Metabolomic data including:
    - Compounds
    - Reactions
  - Integration with Application Data Sharing Service
- Links to new tool and service offerings from NCIBI (Metscape, GIN-IE, GIN-NA, and others)

# MimiWeb Gene Details Extensions

Added to Gene Details

- Reactome Pathways
- Compounds
- Reactions

 Pathways (3 pathways found) - [show/hide](#)

3 pathways found, displaying all pathways.

Pathway	Description	Genes Related to Pathway
<a href="#">KEGG:hsa00565</a> <a href="#">Image</a>	Ether lipid metabolism	<a href="#">View Related</a>
<a href="#">Reactome:REACT_682</a>	Mitotic Prometaphase	<a href="#">View Related</a>
<a href="#">Reactome:REACT_910</a>	M Phase	<a href="#">View Related</a>

 Compounds associated with Gene [show/hide](#)

Id	Name	MF	Mol. Weight	CASNUM	Smile
<a href="#">C00001</a>	water	H2O	18	7732-18-5	
<a href="#">C00033</a>	acetate	C2H4O2	60	64-19-7	CC(=O)O
<a href="#">C04317</a>	1-alkyl-sn-glycerol-3-phosphocholine	C8H20NO6PR	257		
<a href="#">C04598</a>	1-alkyl-2-acetyl-sn-glycerol-3-phosphocholine	C10H22NO7PR	299		

 Enzyme Reactions [show/hide](#)

Enzyme Name: 1-alkyl-2-acetyl-glycerophosphocholine esterase.

EC Number: 3.1.1.47

Id	Description	Reversible?	Equation
<a href="#">R04452</a>	Glycerophospholipid metabolism	false	<a href="#">1-alkyl-2-acetyl-sn-glycero-3-phosphocholine</a> + <a href="#">water</a> = <a href="#">1-alkyl-sn-glycerol-3-phosphocholine</a> + <a href="#">acetate</a>

# MimiWeb Compound Details

MIMI MICHIGAN MOLECULAR INTERACTIONS NCIBI

Free Text Search Gene List Search Query Interactions Browse Database About MIMI Help

## Compound Details

### Compound:

acetate [View MeSH Terms](#)

### CompoundID:

C00033 [View in KEGG](#) [View in Metcape](#)

### MF:

C2H4O2

### Molecular Weight:

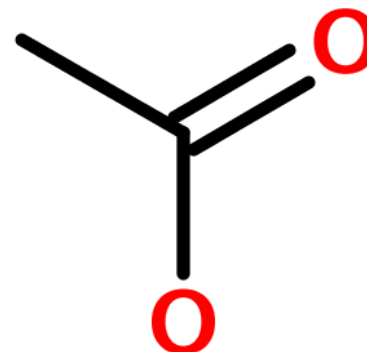
60

### CASNUM:

64-19-7

### Smile:

CC(=O)O



Reactions compound participates in [show/hide](#)

Pathways compound is found in [show/hide](#)

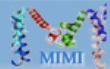


National Institutes of Health grant #J54 DA021519.  
Copyright 2008 by The University of Michigan

[www.ncibi.org](http://www.ncibi.org) - For support and questions email: [mimi-help@umich.edu](mailto:mimi-help@umich.edu)



# MimiWeb Reaction Details



MICHIGAN MOLECULAR INTERACTIONS



[Free Text Search](#)

[Gene List Search](#)

[Query Interactions](#)

[Browse Database](#)

[About MIMI](#)

[Help](#)

## Reaction Details

### Reaction Description:

Glycolysis / Gluconeogenesis

### ReactionID:

R00235 [View Reaction in KEGG](#)

### Reversible:

false

### Reaction Text:

C00002+C00033+C00010=C00020+C00013+C00024

### Equation:

ATP+Acetate+CoA=AMP+Pyrophosphate+Acetyl-CoA

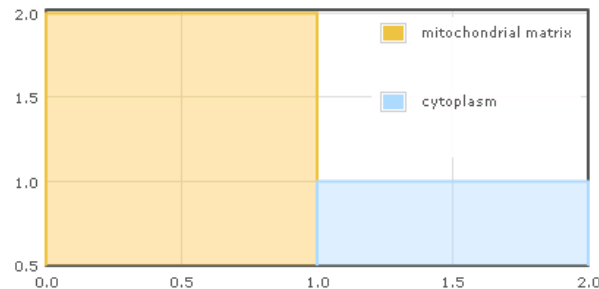
### Enzymes for Reaction:

6.2.1.1

### Genes for Reaction:

[ACSS2](#) [ACSS1](#)

### Subcellular Locations:



Compounds [show](#)/[hide](#)



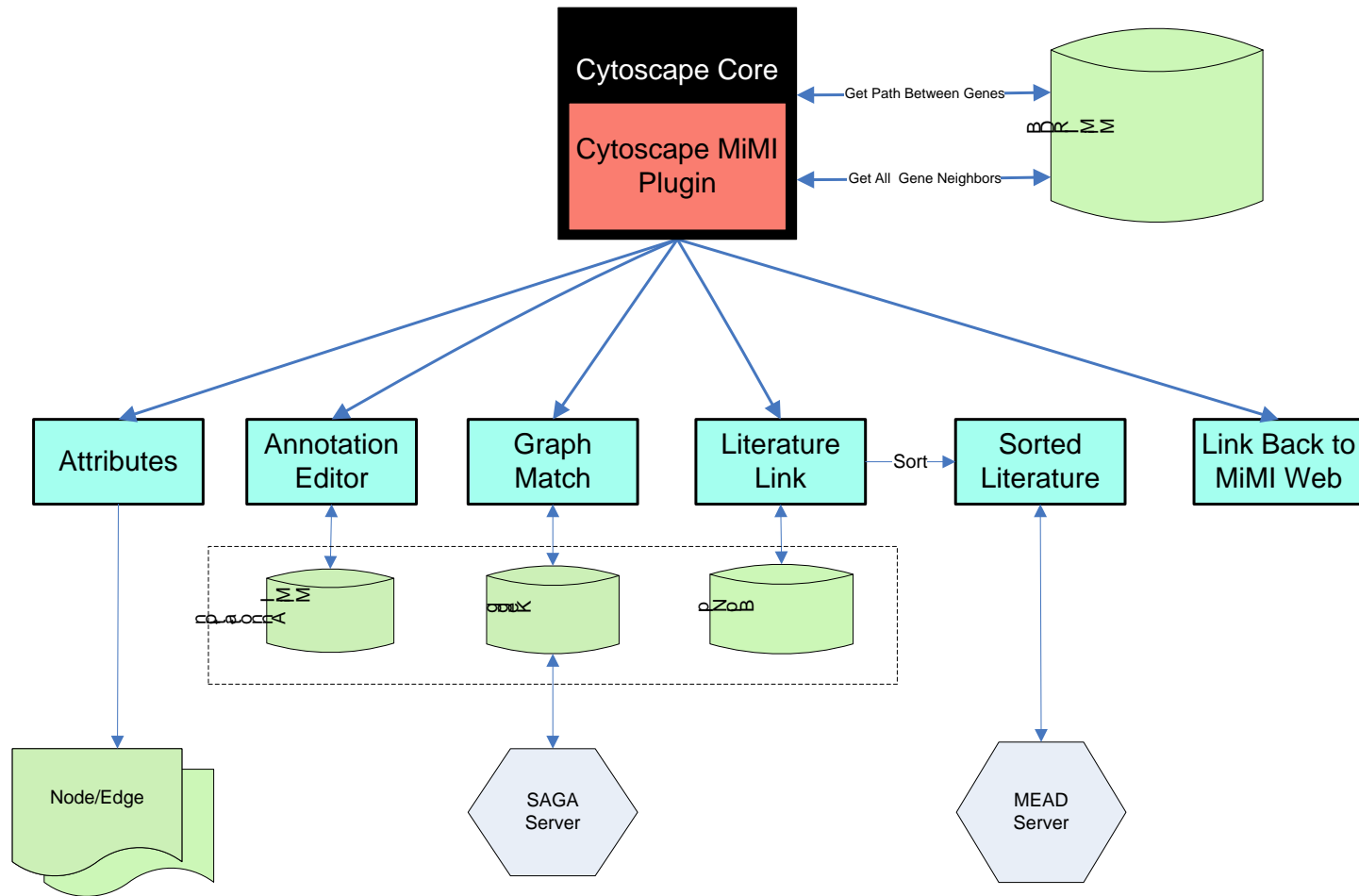
National Institutes of Health grant #U54 DA021519.  
Copyright 2008 by The University of Michigan

[www.ncibi.org](http://www.ncibi.org) - For support and questions email: [mimi-help@umich.edu](mailto:mimi-help@umich.edu)





# MiMI Cytoscape Plugin



# MiMI Cytoscape Plugin Features

**1. Annotating Editor**

Type: protein  
 Taxonomy Name: Homo sapiens  
 Provenance: HPRD; EBI; BioGRID  
 Description: tyros; member of the CSF1/PSGFR receptor family. Involved in macrophage differentiation and Cellular Component: Plasma membrane  
 Molecular Function: Transmembrane receptor protein tyrosine kinase activity  
 Biological Process: Cell communication; Signal transduction  
 Post Translational Modifications: [ENZYME] [GENERAL] in-vitro in-vivo [EFFECT] [DESCRPT]; [ENZYME] HPRD\_01269 [GENERAL] Orthology: [Y2524] JmCust80; [73368] JmCust90; [79952] JmCust90; [25798] JmCust80; [896000719] Interf BioSite: TM; Ig\_LIKE; Ig\_LIKE; IGG2; Ig\_LIKE; Ig\_LIKE; Ty\_Kinase; SP  
 Annotation: Colony stimulating factor 1 receptor precursor; Macroph colony stimulating factor 1 receptor precursor; Fms prot CD115 antigen; vFMS McDonough feline sarcoma viral onco

**2. Cytoscape Desktop (New Session)**

Control Panel: Network Nodes Edges  
 1. Network: 1000(1024) 5366(5362)  
 2. csf1r: Homo sapiens pr: (414) 30(30)

**3. Sentence List Related To [INPP5D] And [SHC1] From BioNlp**

PubMedID	Section	Symbol	Symbol	Sentence
8713348	ABSTRACT	INPP5D	SHC1	The SHP-130 and SHP-155 proteins and inositol polyphosphate 5-phosphatase activity associated with SHC1 in response to B-cell activation.
9670944	ABSTRACT	inpp5d	shc1	Although IL-4R alpha signals Shc and SH2-containing inositol phosphatase (SHIP) phosphorylation, we could not establish an association between their activation and protection from apoptosis.

**4. Interaction Details**

Interaction ID	First Interactor	Second Interactor	Description
324817	socs1_human	csf1r_human	socs1-csf1r; MINT-576193; unspecified, unspecified, unspecified, unspecified, unspecified, over-expressed, over-expressed level, over-expressed, MINT-576196; unspecified, unspecified, prey, prey, over-expressed, over-expressed level, over-expressed, physical interaction, physical interaction, aggregation

**5. DB graph: Insulin signaling**

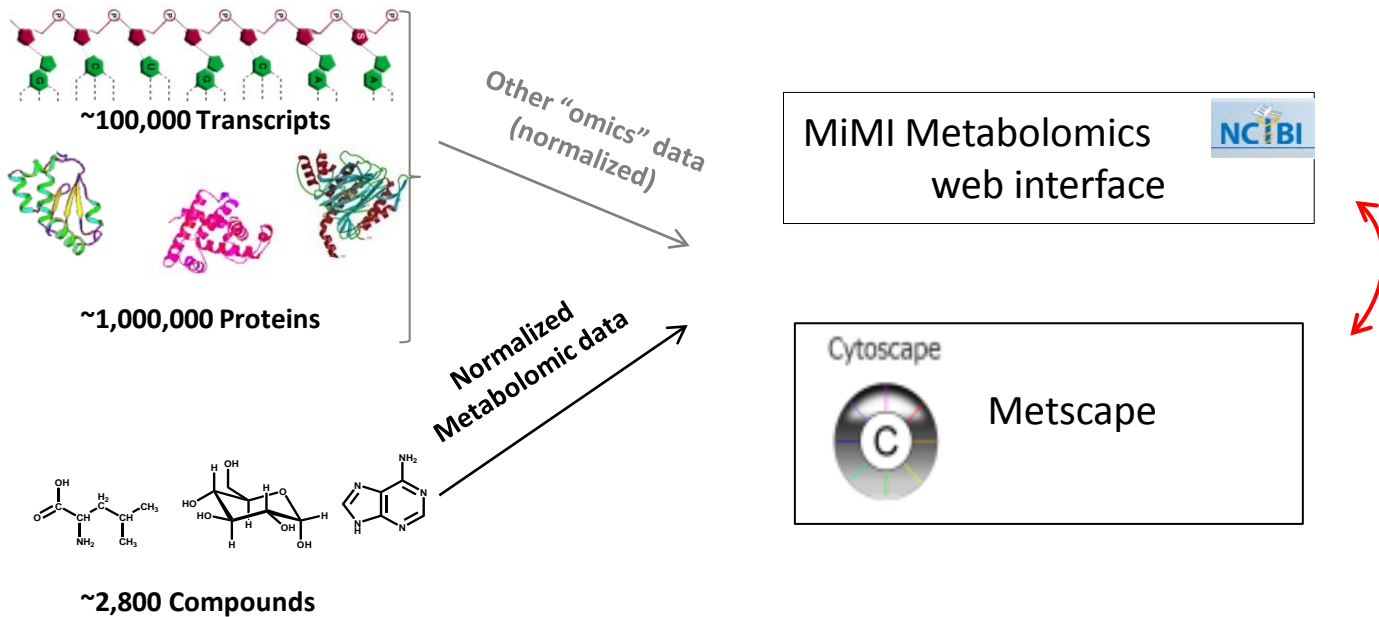
query: csf1rHomo sapiens protein in HPRD level 1

Graph showing interactions between genes: rasal, inpp5d, shc1, cb1, lyn, inpp1, csf1r, socs3, yes1, grap2, pik3r2, csf1, socs1, INSR, SOCS1..., [SORBS1|CBL...]

# Current Status

- Beginning with Cytoscape version 2.5.1, the MiMI Plugin is bundled in the default installation and is installed using the Cytoscape plugin manager.
- In addition, since Aug 2008, there have been 1017 downloads from 742 unique IP addresses.

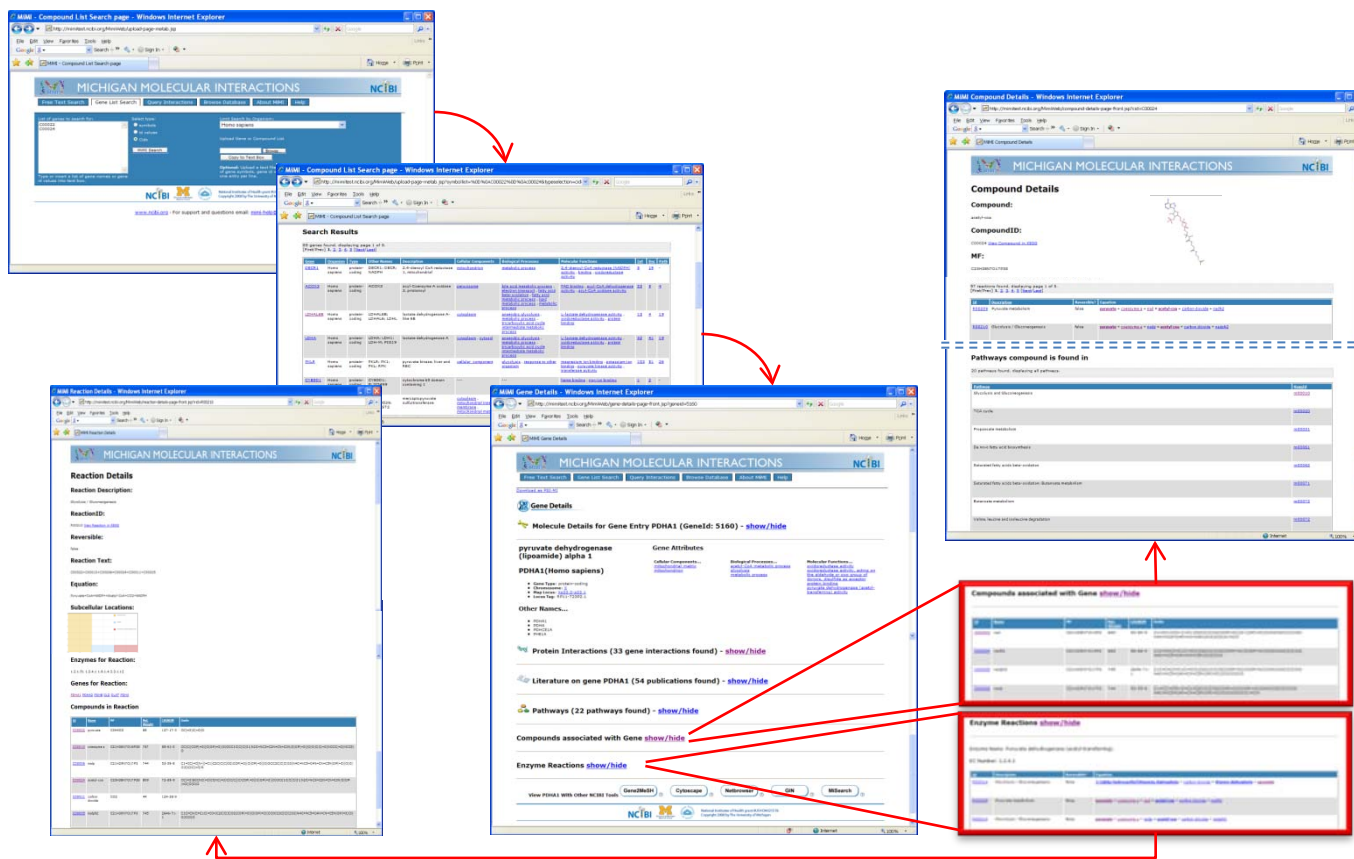
# Bioinformatics Framework for the Analysis and Interpretation of Metabolomic Data



<http://mimi.ncibi.org/MimiWebBeta/upload-page-metab.jsp>

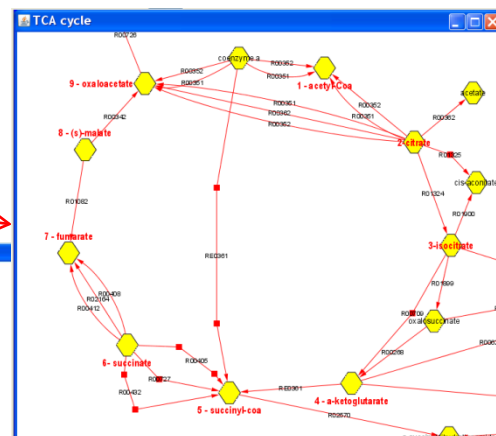
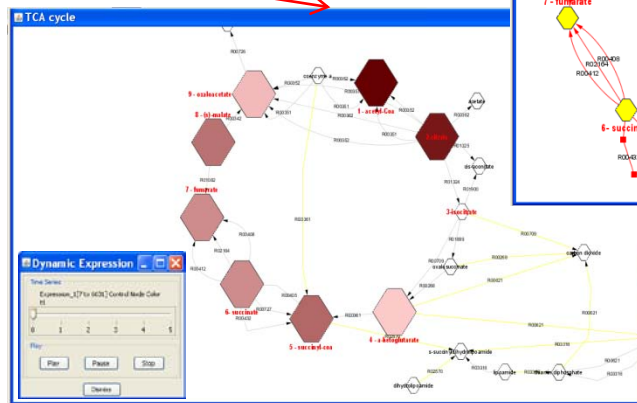
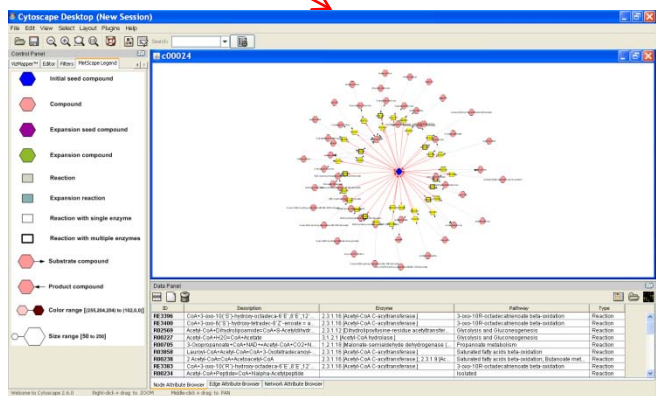
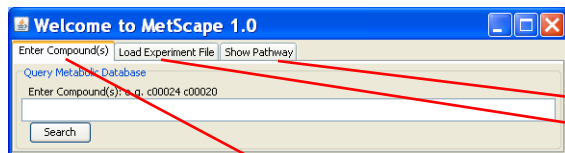
<http://www.cytoscape.org/> (A beta version of Metscape can be downloaded via Cytoscape plug-in manager under Network and Attribute I/O category)

# Extending MiMI-Web to Include Metabolites



[KEGG](#) (Kanehisa et al., 2008) and [EHMN](#) (Ma et al., 2007) databases were used to trace the connections between metabolites and genes, and provide the framework for data analysis

# Visualizing Metabolomic data in MetScape



Metscape is a new plug-in for Cytoscape (<http://www.cytoscape.org/>) that allows users to explore and visualize the networks of metabolites, reactions and pathways

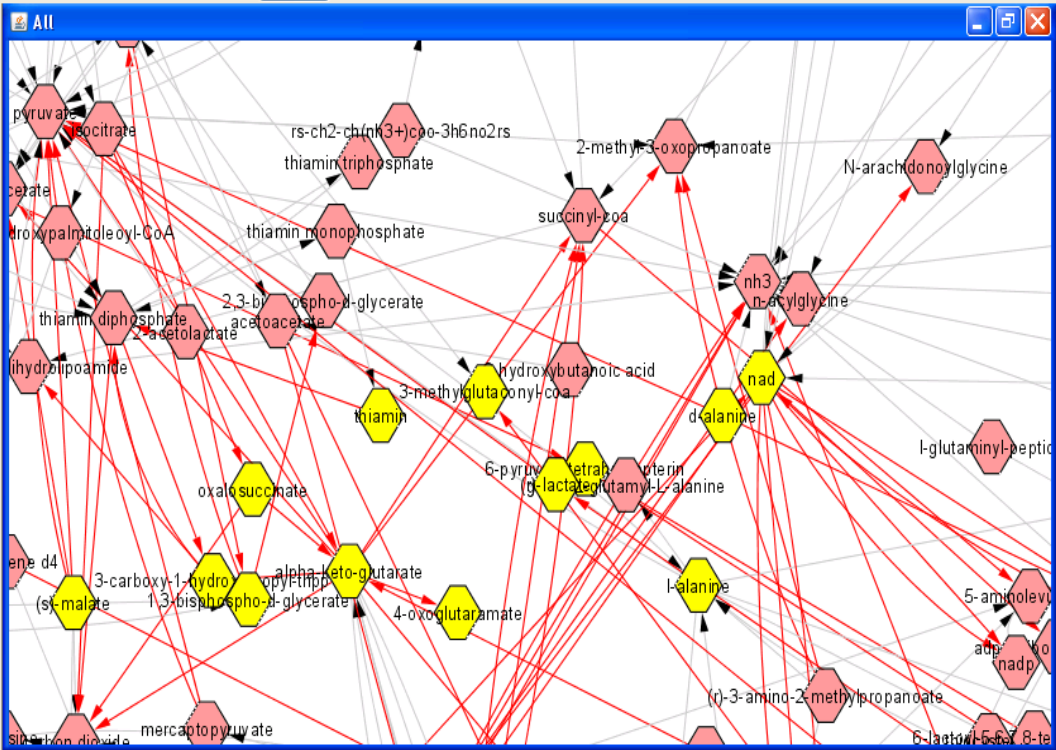
Search:

Control Panel

Network VizMapper™ Editor Filters

Network	Nodes	Edges
All	2621(13)	3603(61)

Main compound number



Results Panel

MetScape Legend

- Initial seed compound
- Compound
- Expansion seed compound
- Expansion compound
- Reaction
- Expansion reaction
- Reaction with single enzyme
- Reaction with multiple enzymes
- Substrate compound

Birds eye view of the whole compound network



Enzyme and Pathway

ID	Description	Enzyme	Pathway	c
C00149 (R00342) C00036	(S)-Malate+NAD+=Oxaloacetate+NADH+H+	1.1.1.37 [Malate dehydrogenase.]	TCA cycle	C00149
C00311 (R00709) C00011	Isocitrate+NAD+=2-Oxoglutarate+CO2+NADH+H+	1.1.1.41 [Isocitrate dehydrogenase (NAD(+))]	TCA cycle	C00311
C00037 (R00366) C00014	Glycine+H2O+Oxygen=Glyoxylate+NH3+H2O2	1.4.3.19 [Glycine oxidase.]; 1.4.3.3 [D-amino-acid o...]	Glycine, serine, alanine and threonine metabolism	C00037
C00197 (R01512) C00236	ATP+3-Phospho-D-glycerate=ADP+3-Phospho-D-g...	2.7.2.3 [Phosphoglycerate kinase.]	Glycolysis and Gluconeogenesis	C00197
C05951 (RE0596) C00037	H(,2)O+leukotriene D(,4) = glycine+leukotriene E(,4)	3.4.13.19 [Membrane dipeptidase.]	Leukotriene metabolism	C05951
C00311 (R00709) C00026	Isocitrate+NAD+=2-Oxoglutarate+CO2+NADH+H+	1.1.1.41 [Isocitrate dehydrogenase (NAD(+))]	TCA cycle	C00311
C00003 (R00112) C00006	NADPH+NAD+=NADP++NADH	1.6.1.1 [NAD(P)(+) transhydrogenase (B-specific).]; ...	Vitamin B3 (nicotinate and nicotinamide) metabolis...	C00003
C00647 (R00277) C00014	Pyridoxamine phosphate+H2O+Oxygen=Pyridoxal ...	1.4.3.5 [Pyridoxal 5'-phosphate synthase.]	Vitamin B6 (pyridoxine) metabolism	C00647

Node Attribute Browser Edge Attribute Browser Network Attribute Browser

Pathway  
Vitamin B6 (pyridoxine) metabolism

Welcome to Cytoscape 2.6.2 Right-click + drag to ZOOM Middle-click + drag to PAN

# GIN-NA:

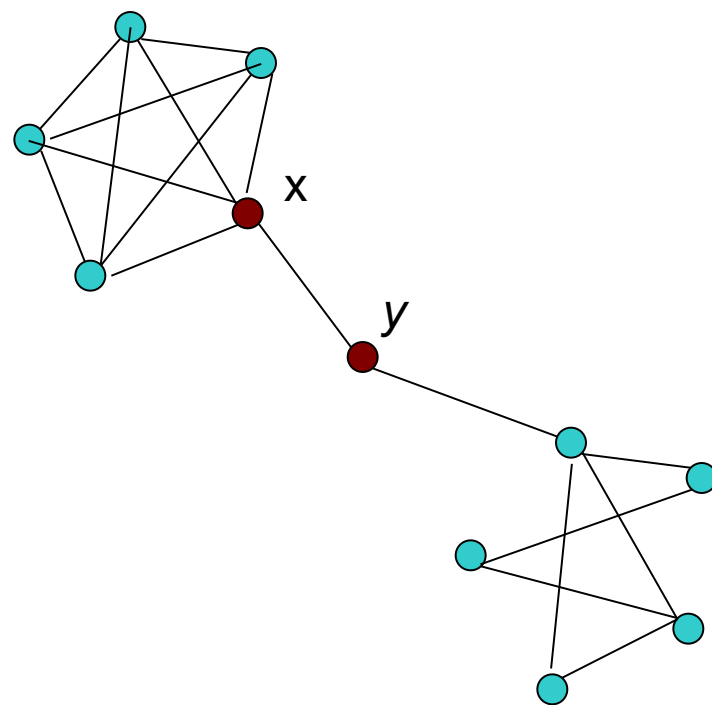
## Gene Interaction Network Analysis

- A system for analyzing molecule interaction networks (Interactions retrieved from the MiMI database)
  - Molecule-Specific Networks:
    - Network of interactions in the neighborhood of a molecule
  - Disease-Specific Networks:
    - Network of interactions “near” known disease genes
- General network statistics:
  - Degree statistics (e.g. average degree)
  - Clustering (e.g. shortest path statistics)
- Ranking Molecules
  - Centrality Measures and Second Neighbors



# GIN-NA: Ranking Molecules

- **Centrality:** Importance of a node in the graph
  - **Degree:** The more neighbors a node has, the more important it is. (most central node: x)
  - **Betweenness:** The more shortest paths pass through the node, the more important it is. Control of a node over the information flow of the network. (most central node: y)
  - **Closeness:** The closer a node to the other nodes, the more important it is. (most central node: y)
- **Second Neighbors:** Rank based on the number of common immediate neighbors





## Molecule-Specific Network: CSF1R

### Network Parameters

- Species: Homo sapiens
- Molecule Type: All Molecule Types
- Data Source: All Data Sources

### General Statistics

- Nodes: 53
- Edges: 124
- Diameter: 2
- Average degree: 4.68
- Degree statistics:
  - Power law exponent: 2.17 r-squared: 0.93
  - Newman power law exponent: 2.27 Error: 0.23
- Clustering:
  - Watts Strogatz clustering coefficient: 0.5136
  - Newman clustering coefficient: 0.1816
  - Clarifit average undirected shortest path: 1.91
  - Ferrer average undirected shortest path: 1.87
  - Harmonic mean geodesic distance: 1.63

### Most Central 10 Nodes

Degree Centrality		Betweenness Centrality		Closeness Centrality	
Molecule	Score	Molecule	Score	Molecule	Score
<a href="#">CSF1R</a>	1.000	<a href="#">CSF1R</a>	0.880	<a href="#">CSF1R</a>	1.000
<a href="#">SHC1</a>	0.288	<a href="#">SHC1</a>	0.020	<a href="#">SHC1</a>	0.584
<a href="#">PK3R1</a>	0.212	<a href="#">PK3R1</a>	0.009	<a href="#">PK3R1</a>	0.559
<a href="#">GRB2</a>	0.192	<a href="#">GRB2</a>	0.005	<a href="#">GRB2</a>	0.553
<a href="#">FYN</a>	0.173	<a href="#">SOCS1</a>	0.004	<a href="#">FYN</a>	0.547
<a href="#">CBL</a>	0.173	<a href="#">NPP5D</a>	0.004	<a href="#">CBL</a>	0.547
<a href="#">PK3R2</a>	0.154	<a href="#">FYN</a>	0.004	<a href="#">PK3R2</a>	0.542
<a href="#">NPP5D</a>	0.154	<a href="#">RASA1</a>	0.003	<a href="#">NPP5D</a>	0.542
<a href="#">SOCS1</a>	0.135	<a href="#">GLYAT1</a>	0.003	<a href="#">SOCS1</a>	0.536
<a href="#">RASA1</a>	0.135	<a href="#">CBL</a>	0.003	<a href="#">RASA1</a>	0.536

### Top 10 Second Neighbors Sorted based on their Common Immediate Neighbors

Second Neighbor	Number of Common Neighbors
<a href="#">TAF1</a>	29
<a href="#">E2F4</a>	16
<a href="#">RBL2</a>	15
<a href="#">HNF4A</a>	15
<a href="#">PLCG1</a>	13
<a href="#">PDGFBR</a>	13
<a href="#">MYC</a>	12
<a href="#">ABL1</a>	12
<a href="#">KIT</a>	11
<a href="#">EGFR</a>	11

### Links

#### MIMI

[Information about CSF1R on MIMI](#)

#### Cytoscape

[Visualize CSF1R network on Cytoscape](#)



## Disease-Specific Network: Prostate Cancer

### General Statistics

- [Nodes: 613](#)
- [Edges: 6034](#)
- [Diameter: 5](#)
- [Average degree: 19.69](#)
- [Degree statistics:](#)
  - [Power law exponent: 2.36](#) [r-squared: 0.85](#)
  - [Newman power law exponent: 1.52](#), [Error: 0.02](#)
- [Clustering:](#)
  - [Watts Strogatz clustering coefficient: 0.2520](#)
  - [Newman clustering coefficient: 0.1864](#)
  - [Clairlib average undirected shortest path: 2.48](#)
  - [Ferrer average undirected shortest path: 2.47](#)
  - [Harmonic mean geodesic distance: 2.31](#)

### Most Central 10 Nodes (Inferred Disease Genes)

Degree Centrality		Betweenness Centrality		Closeness Centrality	
Molecule	Score	Molecule	Score	Molecule	Score
<a href="#">TAF1</a>	0.384	<a href="#">TAF1</a>	0.159	<a href="#">TAF1</a>	0.594
<a href="#">HNF4A</a>	0.255	<a href="#">AR</a>	0.087	<a href="#">HNF4A</a>	0.545
<a href="#">AR</a>	0.245	<a href="#">HNF4A</a>	0.080	<a href="#">MYC</a>	0.535
<a href="#">MYC</a>	0.229	<a href="#">PTEN</a>	0.074	<a href="#">AR</a>	0.534
<a href="#">PTEN</a>	0.217	<a href="#">MYC</a>	0.035	<a href="#">TP53</a>	0.524
<a href="#">E2F4</a>	0.181	<a href="#">BRCA2</a>	0.035	<a href="#">PTEN</a>	0.514
<a href="#">MAX</a>	0.178	<a href="#">E2F4</a>	0.031	<a href="#">E2F4</a>	0.512
<a href="#">E2F1</a>	0.150	<a href="#">MAD1L1</a>	0.024	<a href="#">RBL2</a>	0.507
<a href="#">TP53</a>	0.145	<a href="#">TP53</a>	0.020	<a href="#">EP300</a>	0.506
<a href="#">RELA</a>	0.142	<a href="#">MAX</a>	0.018	<a href="#">ESR1</a>	0.504

### Seed Disease Genes

- [AR](#)
- [BRCA2](#)
- [CHEK2](#)
- [CD82](#)
- [PTEN](#)
- [MXI1](#)
- [KLF6](#)
- [ELAC2](#)
- [HIP1](#)
- [HPC1](#)
- [PCAP](#)
- [MSR1](#)
- [RNASEL](#)
- [MAD1L1](#)
- [EPHB2](#)

### Links

#### MIMI

[Information about the seed Prostate Cancer genes on MIMI](#)

#### Cytoscape

[Visualize the Prostate Cancer specific network on Cytoscape](#)

# Univ. of Wisconsin

- Madison (Jignesh Patel)
  - SAGA: Query biological graph database (e.g. KEGG) to retrieve approximate matches with given subgraphs of molecular interactions.
  - TALE: Align graphs, e.g. of protein interactions across species.
- Milwaukee (Hong Yu)
  - Natural Language Processing of full text from PubMed Central.

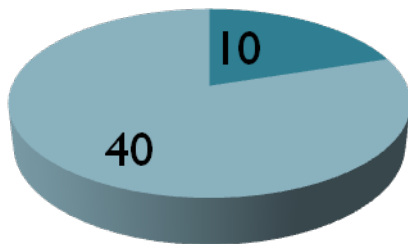
# ConceptGen

- What *biologically related sets of genes* are enriched with the differentially expressed genes from my microarray experiment?
- What pathways or biological processes were affected in my experiment?
- Is there a significant relationship between
  - disease A and any other biological condition?
  - The protein interactions of gene A and a drug or transcription factor target?

# Enrichment testing

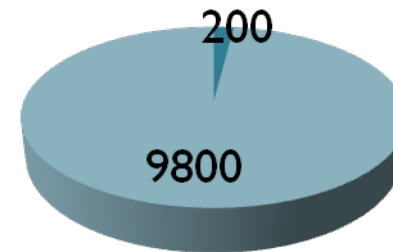
## In Gene Set

- Changed
- Not Changed



## Not In Gene Set

- Changed
- Not Changed



## Modified Fisher's Exact Test

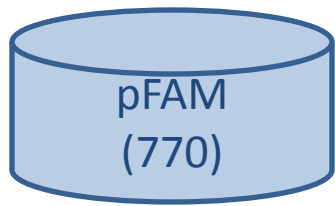
	In Gene Set	Not In Gene Set
Changed	A-1 (9)	B (200)
Not Changed	C (40)	D (9800)

# Concept mapping

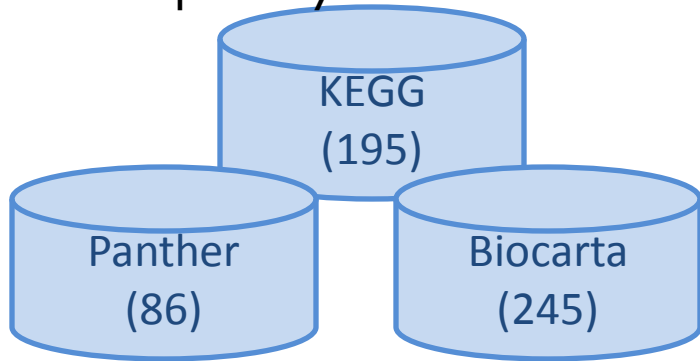
- A step beyond enrichment testing
- Gene sets = *concepts*
- Gene set database/ontology = *concept type*
- Molecular Concept
  - an aspect of biology represented by a molecular signature
  - Diseases, pathways, mechanisms, drugs
- Test enrichment AMONG all concepts
- Visualize results in a graphical network map
- Can identify interesting relationships *among* concepts in diverse concept types (e.g. significant overlap between an miRNA target list and a KEGG disease pathway.)

# Concept Types

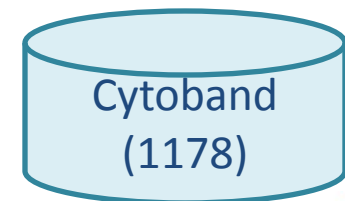
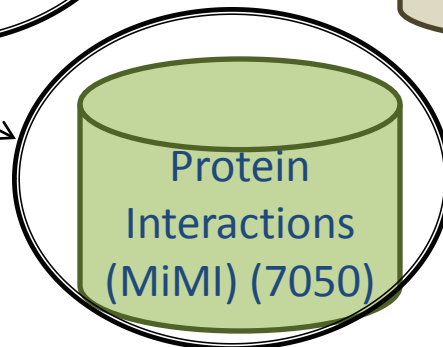
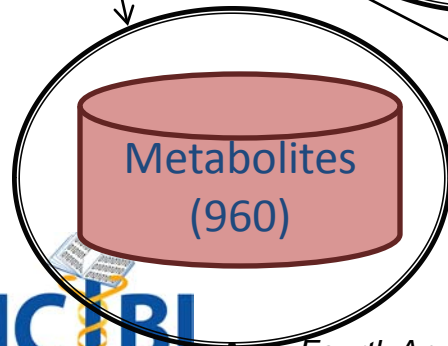
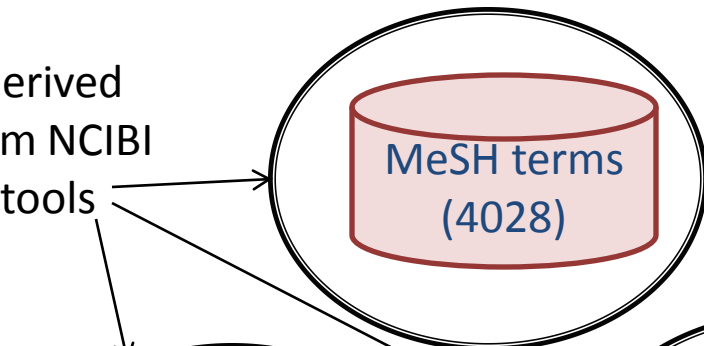
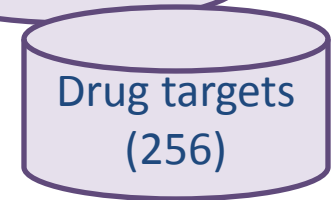
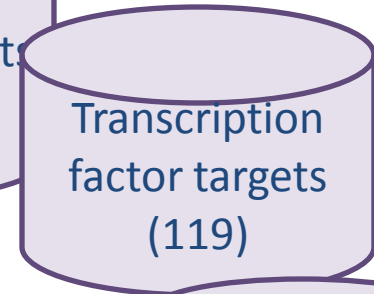
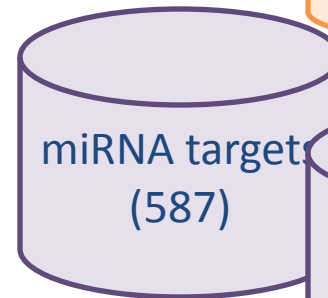
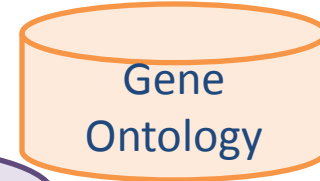
(# concepts)



pathways



- biological process (1043)
- molecular function (660)
- cellular component (292)





# ConceptGen Explorer Window

Concept Explorer

## Bipolar-Smoking-MiMI

Concept Name: Bipolar-Smoking-...  
 Concept Type: Experimental  
 Gene List Size: 87

**Genes on your list, By name**

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

**Filter tab**

**Concepts**  
 Slices = proportionate #

Protein-protein interactions: 323

Reset Chart    Select All

**Enriched Concepts**    **Filter Concepts**    **Details on Overrepresented Concepts for you List**

-- Enriched Concept Tab View --    Selected 0 of 500 Concepts |

Concept Name	Concept Type Name	Category	Gene List Size	Overlap	P-value	Q-value
Receptor, trkC	MeSH	Literature derived	10	5	3.628903E-9	1.617263E-5
Serotonin Plasma Membrane Transporters	MeSH	Literature derived	8	4	4.429032E-7	8.553347E-4
Aldehyde Dehydrogenase	MeSH	Literature derived	14	11	8.979317E-25	5.202257E-20
Alcohol Dehydrogenase	MeSH	Literature derived	12	8	1.049744E-16	3.040898E-12
Brain-Derived Neurotrophic Factor	MeSH	Literature derived	8	4	4.429032E-7	8.553347E-4
Aldehyde Oxidoreductases	MeSH	Literature derived	13	8	2.514599E-16	4.856193E-12
Sex Characteristics	MeSH	Literature derived	16	4	5.679469E-6	7.312127E-3

**Concept Type**

**Test Statistics**

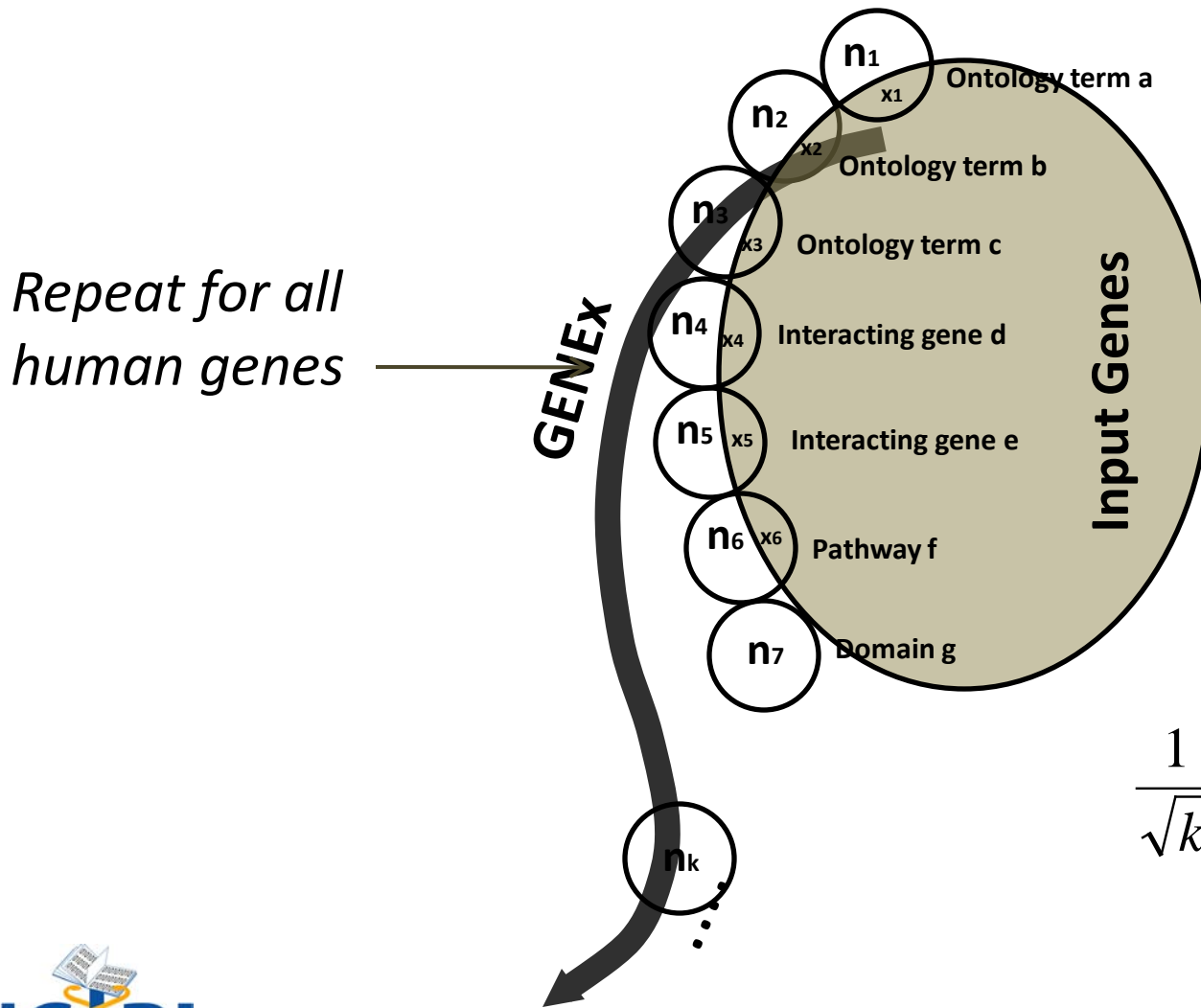
Draw Network Graph    Draw Heatmap    Back to Search

# Concept Signature (ConSig) Analysis – moving beyond concept mapping

## **Applications include:**

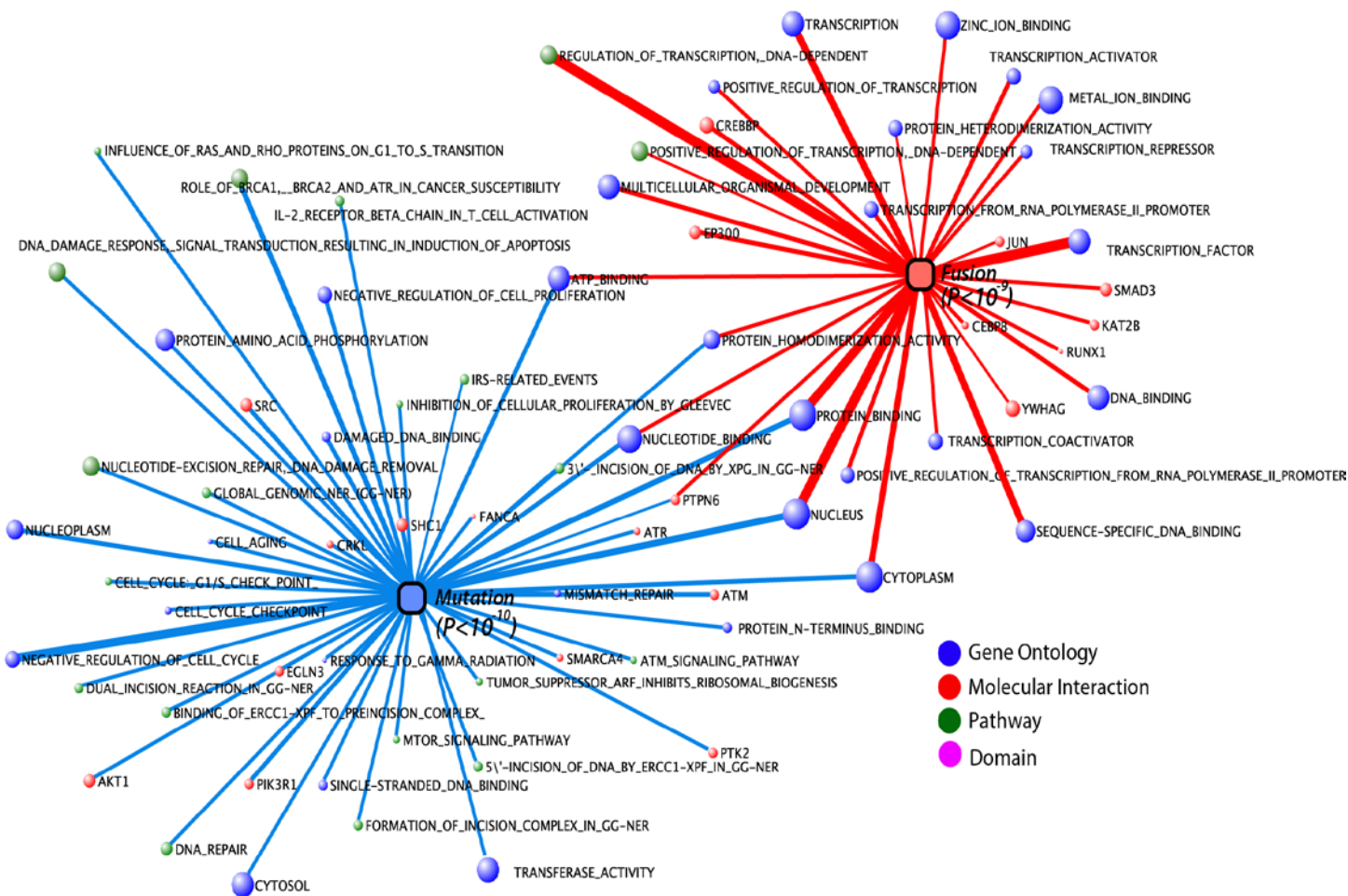
- Nomination of potential phenotype or disease associated genes
- Filtering MiMI interactions by functional importance
- Identification of interacting hubs in a gene list
- Finding common hubs across expression signatures
- Comparing the functional similarity between different gene groups

# The ConSig-Score Algorithm



$$\frac{1}{\sqrt{k}} \sum_{i=1}^k \log_{10} \left( 1 + \frac{x_i}{\sqrt{n_i}} \right)$$

# Fusion Genes Share Functional Signature



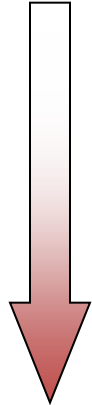
Novel fusion genes predicted from ConSig analysis have been confirmed

# Other Tools

- MiSearch – personalized PubMed
- Gene2MeSH, Metab2MeSH – relate MeSH terms to genes/metabolites and back
- BioSearch2D – relate publications to biological concepts
- Markit – biomarker prediction using topologically constrained bayesian networks.

# Levels of Integration

## User Interface:



- Pencil and paper
- Copy, paste
- URL passing IDs, single authentication
- History of searches, persistence, memory
- Stored private datasets, workspace

## Core Databases:

- Federated with common linkage of data
  - GeneID, ProteinID (Uniprot), PMID, *ConceptID*, *MoleculeID*

# NCIBI Integration Needs

- Integration of toolset into single framework
  - Tools need access to each others functionality
  - Tool interface separation and isolation (e.g. to embed one tool GUI in the GUI of another)
- Unified identification (via web service)
  - Use identification for user-tuned profiles
  - Authentication/authorization
- Protected but shared resource repository
  - Shared repository with user and web interface
  - Input/output to/from tools

# NCIBI Data Challenges

- NCIBI has large amounts of data
  - Protein, Gene, Interaction, Metabolites, Pathways, SNPs, Literature and more!
- Users need many separate entry points to get at the data through specialized tool offerings.
- Goal: Integrate services together to allow users to easily make use of NCIBI tools and data.
  - Move data easily across specialized tools
  - Locate information quickly
  - Allow users to contrast and compare multiple views of data



# Biolets – Solving the user/data divide

- Pluggable architecture that allows new biological services to be added.
- Allows users to move their findings easily across tools.
- Enhanced search and filtering capabilities brings disparate data together allowing the user to focus and drill in on items of interest.
- Three plugins in the works, more on the way.
- Integrates with outside services (NCIBI Hive TagMapper service, GenePattern)

# Application Data Sharing Services

- Allows users to save data in one tool and easily use it in another.
- Users can choose their entry point into NCIBI services moving their findings to more specialized tools as needed.
- Pre-existing data sets can be uploaded and used as the starting point.
- Secure architecture means users can reliably and safely store and access their data.

# Application Data Sharing Services

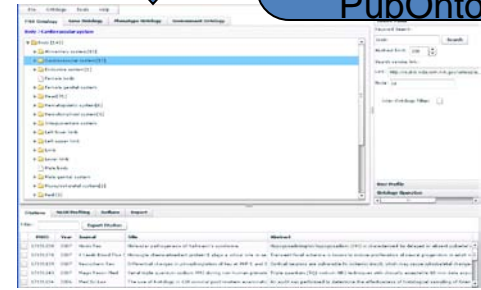
1. Users chooses starting tool (here uses starts in Gene2Mesh).



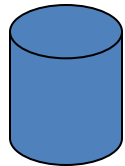
2. After selecting data they move it to MiMI.



3. The user then moves the data to PubOnto.



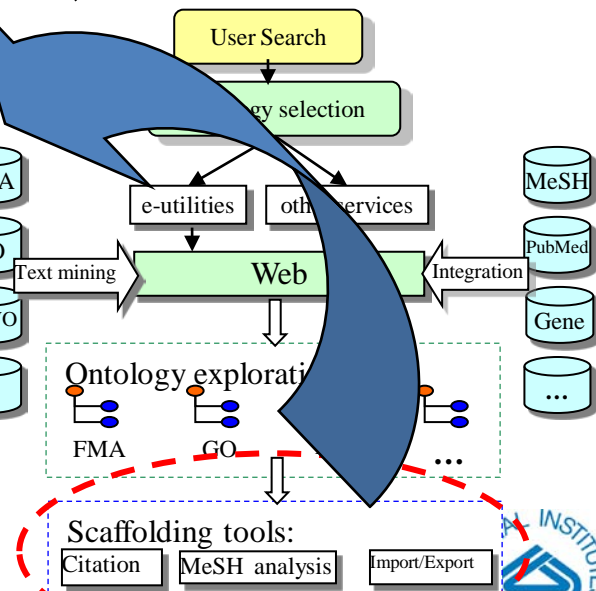
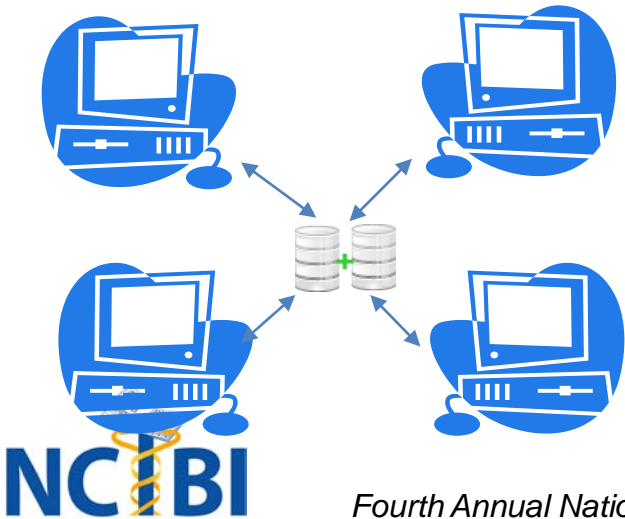
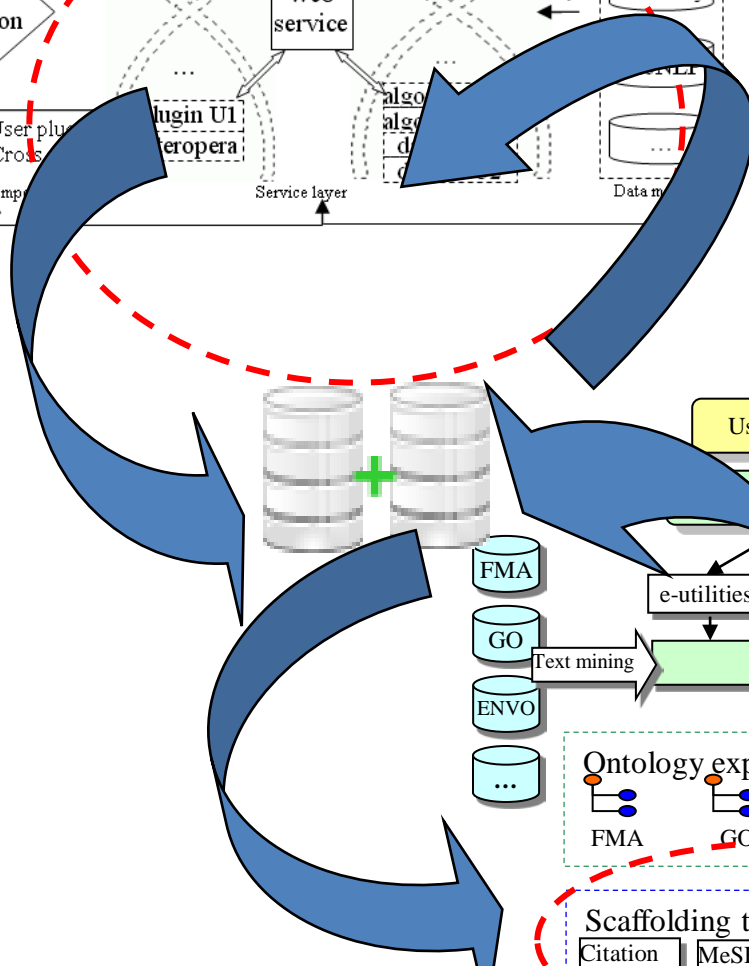
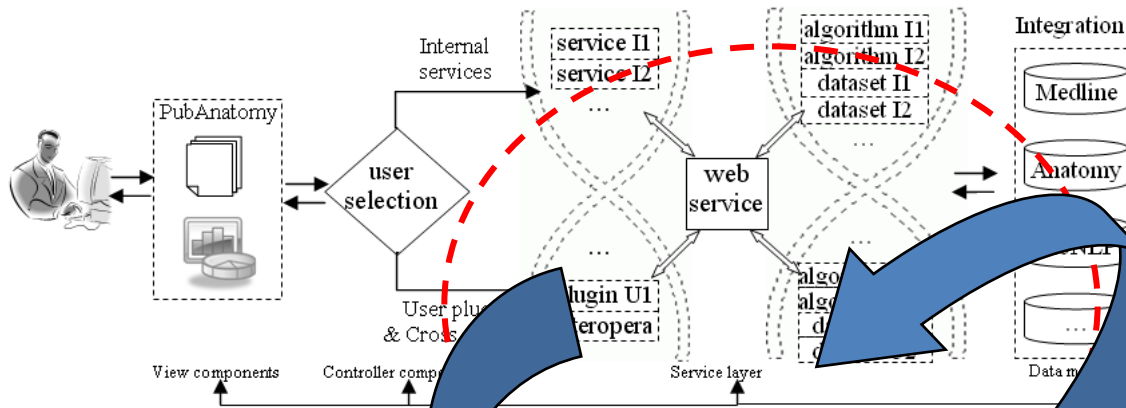
4. Finally the user saves all the interesting information they have found.



5. Next time the user can start with the saved data and choose a different specialized tool for their investigations.



# Integration Architecture: Example



# Implementation

## Principles:

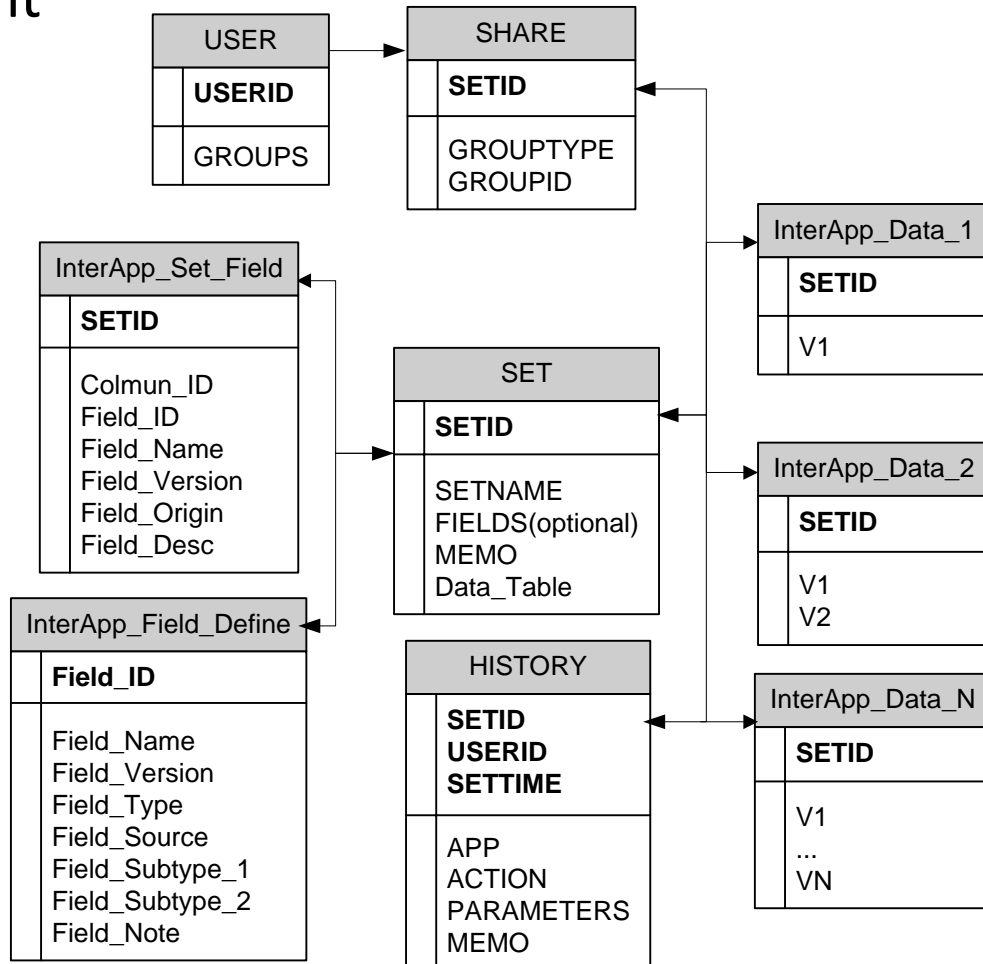
- ✓ Agreement on data sharing approach
- ✓ Loose Coupling among apps
- ✓ Encapsulation of core function
- ✓ Interoperability: app pipeline
- ✓ Composability: built larger system
- ✓ Abstraction of underlying services
- ✓ Performance: minimize overhead
- ✓ Usability: simple service calls
- ✓ Flexibility: adaptive to various NCIBI data sharing requirements

## Components:

- ❖ Database schema
- ❖ Syntactic/semantic interoperability among NCIBI web applications
- ❖ Centralized database repository
- ❖ Application-independent dataset operations
- ❖ Shared core dataset operating web services, e.g., read/write/operate
- ❖ Dataset sharing authorization and user authentication

# Data Integration Schema

- Consideration, scenarios, limitations=>adaptations
- Incremental, data tracing, format variations, complex requirement



# Demo of Example Pipeline

**Gene2MeSH – Gene Annotation with MeSH Terms**

Search Gene2MeSH | About Gene2MeSH

Search by: Gene Symbol | MeSH Term | Limit Search by Organism: All Organisms

Search:   examples: Arc2, "Prostatic Neoplasms"

History: bipolar disorder

745 genes found matching MeSH heading "bipolar disorder"

= lookup gene or MeSH heading:  = view interactions in MiMi

Gene Symbol	MeSH Heading	TaxID	Fisher's Exact	MeSH Qualifier	Gene Description
<input checked="" type="checkbox"/> APOE	<a href="#">Carotid Atherosclerosis</a>	9606	1.7e-2333	genetics	apolipoprotein E
<input checked="" type="checkbox"/> JAK2	<a href="#">Myeloproliferative Disorders</a>	9606	1.1e-2077	genetics	Janus kinase 2 (a protein tyrosine kinase)

File Anatomy Tools Help

Brain Map | Gene Expression Correlation | Image Section

0.10 N/A

11.30

CTX SUB DG OP MPT MPT OP DG CTX

CA IGL LgD LP ApHPBT NPC PAG MBmot DORpm LCA ICA

LQv DORpm TH VP Brain IB BS ZI PH PH

Search Panel

Keywords: bipolar disorder (And/Or) Gene ID or Symbol

Search Engine Options: PubAnatomy search engine

Abstract search range: 500 (Large limit may take minutes)

Mapping Options: Brain Structure Term Mapping

Image Resolution(micron): 50

Current image section 156

Current structure

Citation	Gene	MeSH Profiling	Author	Disease	In Situ Image	MiMi Interac...	Expression C...	Gene Express...	Significant G...	Import	
1) Retrieve Data Sets											
<input type="checkbox"/>	Set ID	Set Name	Time	4) Import!							
<input checked="" type="checkbox"/>	143	literature result set	21-APR-09	Import PMID as citation set							
<input checked="" type="checkbox"/>	51	literature result set	12-FEB-09	PubAnatomy	PMID		Fields	Import and execute...	et Description		
<input checked="" type="checkbox"/>	52	my search	12-FEB-09	PubAnatomy	PMID				test data update		

Cytoscape Desktop (New Session)

File Edit View Select Layout Plugins Help

Control Panel

Network | MiMiApp

Initial seed node

Neighboring node

Expansion seed

Expansion neighbor

User annotation

Normal interaction

User annotation

Data Panel

4951

Node Attribute Browser | Edge Attribute Browser | Network Attribute Browser

Welcome to Cytoscape 2.8.3 | Copyright © 2004-2009 | Middle-click + drag to PAN

File Help

Choose Action | Upload Dataset | Operate Dataset

Actions To Take:

Operate selected datasets

Select for further process

File Help

Choose Calculation Method: MiMi Interaction

Source Genes:

GeneID	Symbol	Description	
<input type="checkbox"/>	11816	APOE	apolipoprotein E
<input type="checkbox"/>	11829	AQP4	aquaporin 4
<input type="checkbox"/>	12326	CAMK4	calcium/calmodulin-dependent protein kinase II gamma
<input type="checkbox"/>	12359	CAT	catalase
<input type="checkbox"/>	12753	CLOCK	circadian locomotor output cycles kapap

Target Genes:

GeneID	Symbol	Description	
<input type="checkbox"/>	28240	TRPM2	transient receptor potential cation channel subfamily M member 2
<input type="checkbox"/>	216343	TPH2	tryptophan hydroxylase 2
<input type="checkbox"/>	21955	TNNT1	troponin T1, skeletal, slow
<input type="checkbox"/>	20977	SYP	synaptophysin
<input type="checkbox"/>	20655	SOD1	superoxide dismutase 1, soluble

Relations among given genes | Click on a gene to see its details (Found total direct path record: 18; Mapped source gene:10; Mapped target gene: 10)

Source Gen	Target Gene	Source Symbol	Target Symbol	Strength	Source Gene Description	Target Gene Description		
<input type="checkbox"/>	<input type="checkbox"/>	11816	17756	2	APOE	MTAP2	apolipoprotein E	microtubule-associated protein 2
<input type="checkbox"/>	<input type="checkbox"/>	11816	19122	2	APOE	PRNP	apolipoprotein E	prion protein
<input type="checkbox"/>	<input type="checkbox"/>	12359	20655	1	CAT	SOD1	catalase	superoxide dismutase 1, soluble
<input type="checkbox"/>	<input type="checkbox"/>	12912	20111	2	CREB1	RPS6KA1	cAMP responsive element binding protein 1	ribosomal protein S6 kinase polypeptide chain alpha
<input type="checkbox"/>	<input type="checkbox"/>	12912	56637	3	CREB1	GSK3B	cAMP responsive element binding protein 1	glycogen synthase kinase-3 beta

Gene2Mesh ⇒ PubAnatomy ⇒ PubPath ⇒ PubIO ⇒ MiMi-Cytoscape

# IBIS – Integrated Searching And Filter across data sources

Integration with Data Sharing Services allows users to easily bring findings into other tools

Filtering allows for deeper exploration of results

Search results across many different types of data

IBIS Integrated Bioinformatics Search

New Save As Search Results Filters Clipboard Preview Help

Search Filter GO MESH Gene

Search

Search Choices:

- Genes (example: brca2)
- Concepts (example: diabetes)
- Free Text

Output:

- Genes
- Concepts
- Interactions
- Metabolites
- Literature
- Pathways

Genes Interactions Literature Pathways

GeneId	Gene	Organism	Type	Other Names	Description	Cellular Compon...	Biological Proce...	Molecular Functi...	Int	Doc
--------	------	----------	------	-------------	-------------	--------------------	---------------------	---------------------	-----	-----



# IBIS Demo

- Problem: Queries to the NCIBI Databases results in long and indigestible lists of information.
- IBIS - Integrated Search and Filter Service at NCIBI. Currently in development...
  - Search services across multiple data sources.
  - Filtering of results based on search context.
  - Allows for searching and filtering associations between data.
- Search across 4 different repositories: Gene DB, MiMI for interactions, Gene to Mesh mapping, and Tagging of PubMed Abstracts.
- Demo: search 'Prostate Cancer' and filter

# MiSearch

*(before authentication)*



## Adaptive PubMed Search Tool



[LOGIN for Personalized Result Ranking](#)

Query   [MiSearch Help](#)

examples: *gab2* or *prostate cancer AND androgen receptor*



National Institutes of Health grant #U54 DA021519.  
Copyright 2008 by The University of Michigan



# Federated Login leveraging Shibboleth



Please login with your email address and password (you only need to login once for all NCIBI Applications):

Email Address:

Password:

Login

[Registration](#) is easy and allows the NCIBI tools to provide features such as search history and preferences.



# MiSearch

*(after authentication)*



## Adaptive PubMed Search Tool



Profile for Beth K

Query

Submit

MiSearch

[Help](#)

examples: *gab2* or *prostate cancer AND androgen receptor*



National Institutes of Health grant #U54 DA021519.  
Copyright 2008 by The University of Michigan

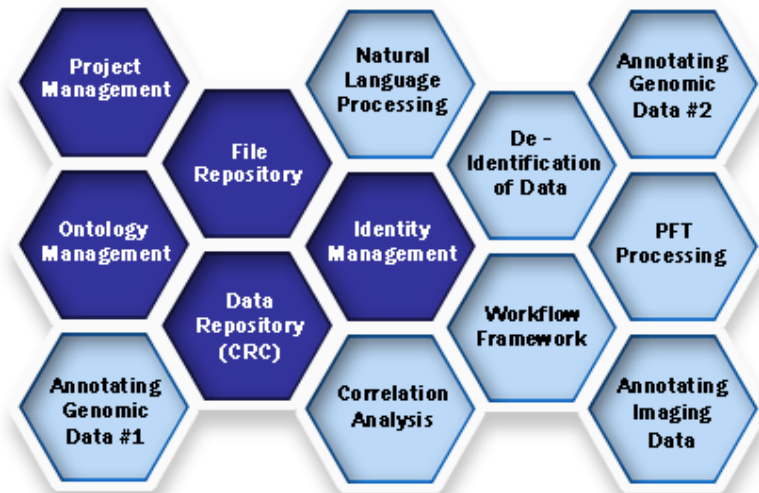


# Integration with Others' Tools

- GenePattern
  - Have worked with them for many years
  - Jill Mesirov keynote coming up
- Cytoscape
  - Have been working with them
- I2B2/Hive
  - Current effort is on-going

# HIVE - I2B2 Framework

- <https://www.i2b2.org/> (Informatics for Integrating Biology and the Bedside)
- HIVE = Framework + Workbench
- Community Contributions
- Build on open-source technologies



# TagMapper Service for Hive

- TagMapper plugin for Hive integrates the NCIBI Biolet web services with I2B2 Hive.
- Pluggable nature of Hive and Biolet's allows for easy integration of services.
- NCIBI TagMapper provides mapping services between ICD9 codes, MeSH terms and genes.
- Additional mapping services are planned.

# TagMapper: Client and Demo

- User drag/drop from diagnostic category to plugin
- Plugin/Server converts ICD9 code to Mesh Terms and retrieves related Genes (additional gene information can be displayed)
- Adding support to navigate through gene information
- Adding support to view additional mappings

The screenshot shows the i2b2 Workbench interface. A callout box at the top center points to the 'Diagnostic Categories in I2B2' section. A callout box on the left points to the 'in Term Navigator...' section. A callout box in the middle points to the '... in other contexts...' section. A callout box on the right points to the '... in patient data' section. A callout box at the bottom center points to the 'Related genes at plugin' section. A callout box on the right side points to the 'TagMapper plugin:' section, which lists 'integrated into Hive' and 'maps ICD9 to Genes'. The interface also shows a 'Gene List' with genes like Fcgr1, Il12, and Aem2, and an 'ICD9 Code' field with the value 283.0.



# Biositemaps

- Biositemaps are a collaborative venture between all NCBC centers. NCIBI has actively worked with NCBO and CCB to develop and deploy the following tools:
- The **Biositemap Search and Browse** tools provide a web interface for simple or complex queries against all resources defined in all published biositemaps across the internet.
- The **Biositemaps Editor** provides an authoring web interface to fill in the information about the resources on your site and generate a biositemap.rdf file.
- The **BioPortal** supports queries based on the Biomedical Resource Ontology (BRO).
- The **iTools** application and iTools services provide the functions of (i) authoring, (ii) locating and (iii) querying biositemaps.
- Additional information is available at <http://www.biositemaps.org>

# Biositemap Basic Search



Use the form below to search biositemap resources.

- Text will be searched in all properties unless a specific property is selected.
- Search is case-insensitive (see [Advanced Search](#) for more options).

Search in Property

Enter Search Text

Submit Search

National Center for Integrative Biomedical Informatics  
Supported by NIH Grant # U54-DA021519  
Copyright 2009 The University of Michigan  
Version 1.0



# Biositemap Advanced Search



Your [last search](#) found **92 resources**. You may search within these results or start a [New Advanced Search](#).

Use the form below to search biositemap resources or return to the [Basic Search](#).

Select Organization (Optional):

Select Center (Optional):

Select Resource Type (Optional):

Search in Property:

Enter Search Text:

Exact Match?

National Center for Integrative Biomedical Informatics  
Supported by NIH Grant # U54-DA021519  
Copyright 2009 The University of Michigan  
Version 1.0



# Integrated Biositemap Browser



Resource Descriptions	Resource Properties
<p>Name ▲</p> <ul style="list-style-type: none"><li>PeptideAtlas</li><li>X_Tandem</li><li>ProteinIdDatabase</li></ul>	<p>Resource Name: Peptide Atlas</p> <p>Description: Multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments, some published, many still unpublished. All results of sequence searching are processed through PeptideProphet to derive a probability of correct identification for all results in a uniform manner ensuring a high quality database. All peptides are mapped to Ensembl and can be viewed as custom tracks on the Ensembl genome browser.</p> <p>Resource Type: <b>BRO (1 Item)</b> BRO:Proteomics</p> <p>URL: <a href="http://www.peptideatlas.org">http://www.peptideatlas.org</a></p> <p>Keywords: Proteomics, mass spectrometry, peptides, Ensembl, PeptideProphet, uniform re-analysis of raw data</p> <p>Organization: Institute for Systems Biology</p> <p>Center: NCIBI</p> <p>Contact Person: Eric Deutsch</p> <p>Contact Person Email: <a href="mailto:edeutsch@systemsbiology.org">edeutsch@systemsbiology.org</a></p> <p>Contact Person Phone:</p> <p>Language:</p> <p>License Model: free, but registration required</p> <p>Platforms:</p>

The Biositemap Browser is a component of [The National Center for Biomedical Ontology](#)  
The National Center for Biomedical Ontology is one of the [National Centers for Biomedical Computing](#) supported by the [NIH Roadmap](#).