

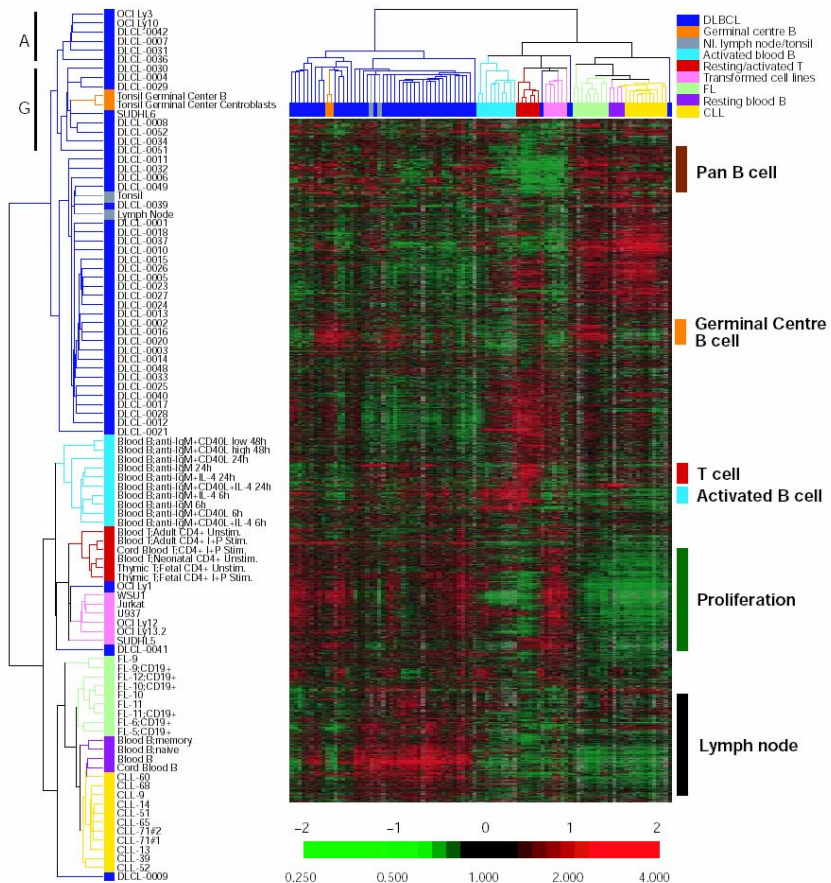


Functional Gene Clustering by Latent Semantic Indexing of MEDLINE Abstracts

Ramin Homayouni, Ph.D.
Department of Neurology
University of Tennessee Health Science Center



Gene Expression Profiling



Now What?

Alizadeh, et al., (2000) Nature 403:503.



Some Web Resources



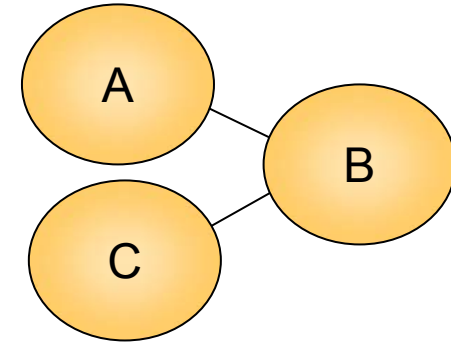
➤ NCBI Sites

- **OMIM** <http://www.ncbi.nlm.nih.gov/Literature/index.html>
- **LocusLink** <http://www.ncbi.nlm.nih.gov/LocusLink/>
- **PubMed** <http://www.ncbi.nlm.nih.gov/entrez/>

➤ Others

- **HAPI** <http://array.ucsd.edu/hapi/>
- **GenMAPP** <http://www.genmapp.org/>
- **GO Tree Machine** <http://genereg.ornl.gov/gotm/>
- **PubGene** <http://www.pubgene.org>
- **Arrowsmith** <http://arrowsmith.psych.uic.edu/>
- **Chilibot** <http://www.chilibot.net/>

Defining Functional Relationships between Genes



➤ Direct Relationship

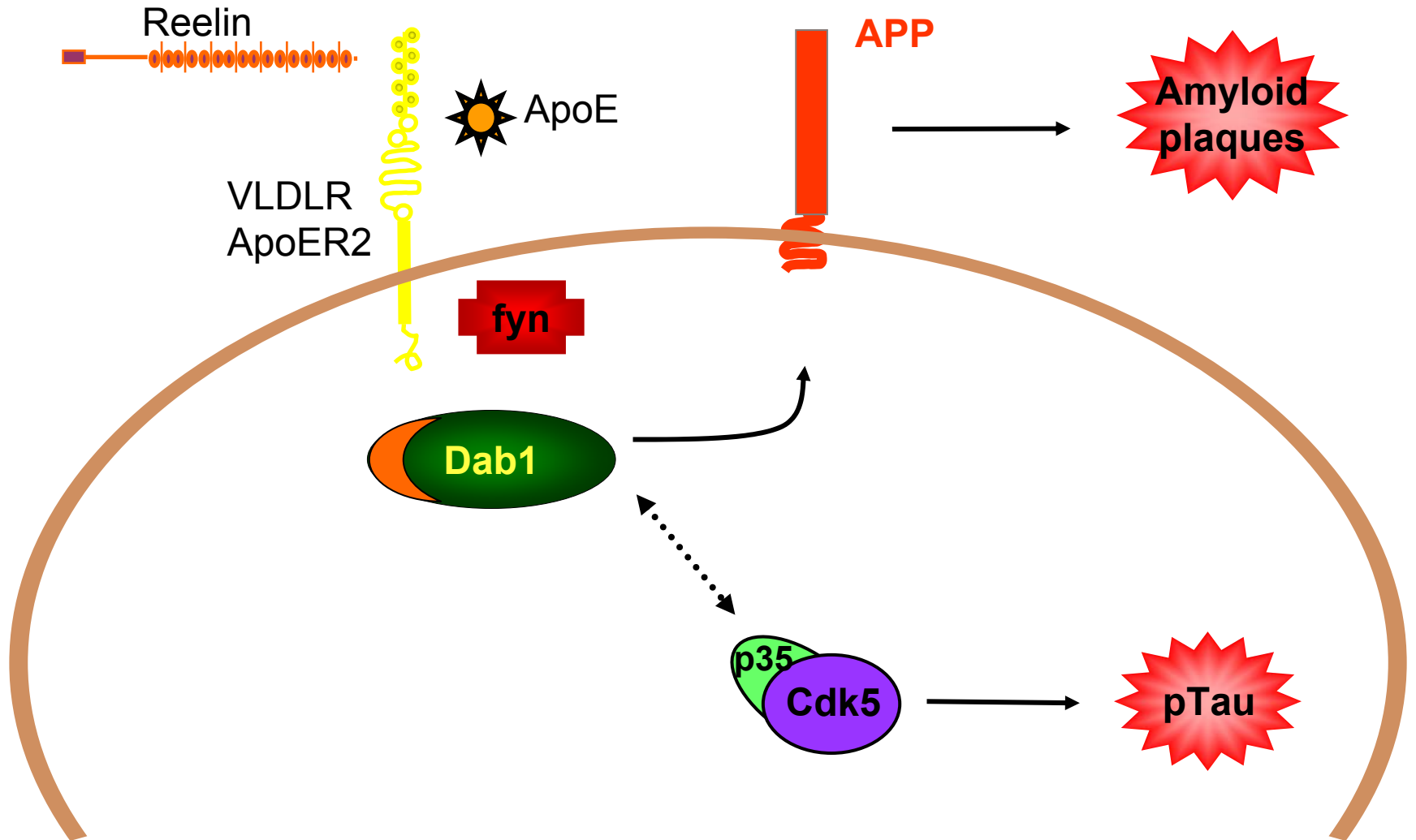
Gene relationships already known (e.g., A-B or B-C)

- Term co-occurrence
 - Gene symbol: PubGene (*Jenssen et al., Nature Genetics 2001 28:21*)
 - Gene names (synonyms and aliases) – biochemical

➤ Indirect Relationship

Gene relationships unknown (e.g., such as A-C)

Reelin Signaling Pathway



Gene Document Test Set



Reeler

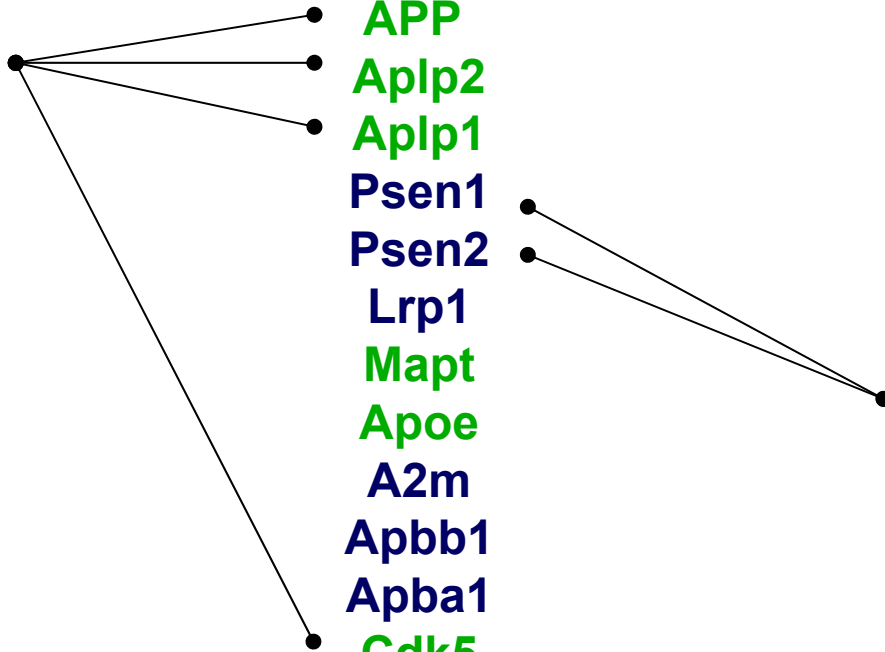
Reln
Dab1
VLDLR
Lpr8

Alzheimer Disease

APP
Aplp2
Aplp1
Psen1
Psen2
Lrp1
Mapt
ApoE
A2m
Apbb1
Apba1
Cdk5
Cdk5r
Cdk5r2

Miscellaneous

Trp53
Fos
Nras
Rasa1
Rab1
Src
Notch1
DII1
Jag1
Robo1
Ptch
Smo



PubGene Query: Dab1

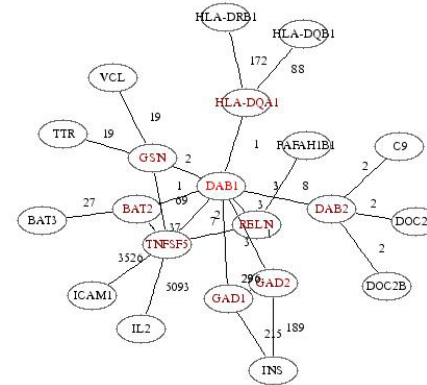
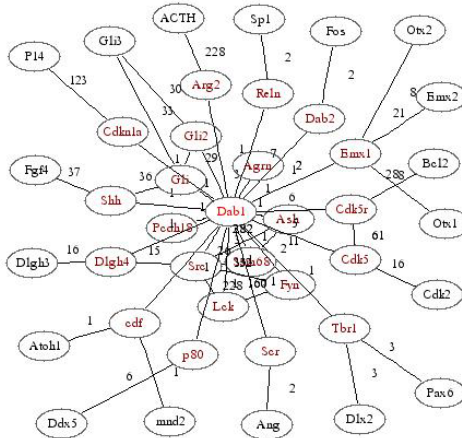
<http://www.pubgene.org/>



Mouse

Human

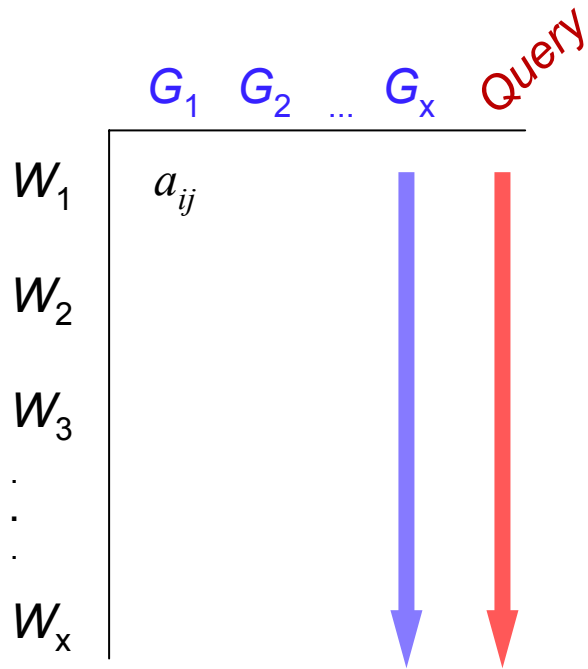
Reln 7 times
 Cdk5r 6 times
Cdk5 5 times
 Gli2 3 times
 Src 3 times
 Dab2 2 times
 Fyn 2 times
 Sam68 1 times
 Cdkn1a 1 times
 Tbr1 1 times
 Gli 1 times
 Scr 1 times
 Shh 1 times
 cdf 1 times
 Ash 1 times
 Dlg4 1 times
 p80 1 times
 Lck 1 times
 Emx1 1 times
 Pcdh18 1 times
 Agrn 1 times
 Arg2 1 times



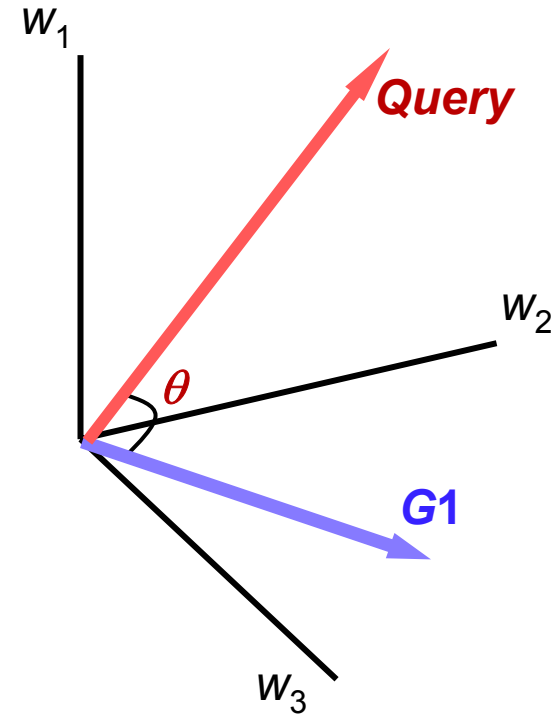
DAB2 3 times
 GAD1 3 times
 RELN 3 times
 GSN 2 times
 TNFSF5 2 times
 HLA-DQA1 1 times
 BAT2 1 times
 GAD2 1 times

PubMed Query: Dab1 AND Reln = 10
PubMed Query: Dab1 AND reelin = 57 !

Vector Space Model: Latent Semantic Indexing

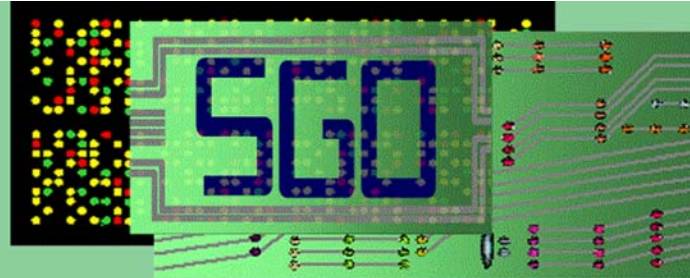


$$a_{ij} = l_{ij} g_i$$



Semantic Gene Organizer[©]

User Interface



© 2002-2003 Dr. Michael Berry, Dr. Ramin Homayouni, Kevin Heinrich, Lai Wei

Document Collection

Select document collection:

Please input a session name:

- 10_10_02
- 2_10_03PMID
- 2_10_03ST
- 3_04_03PMID
- Down_Regulated
- Regulated
- Up_Regulated

Query

Query Method:

(separate keyword queries by a blank line) (set id per line)

- Demo using Gene Accession Numbers
- Demo using Probe Set IDs
- Demo using Keywords

Work with existing session:

Reelin Accession # Query



http://shad.cs.utk.edu/sgo/cgi-bin/pub/start.cgi - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Links Google PubMed_UT UniGene LocusLink BLAST MD Consult SGO UT directory Address

http://shad.cs.utk.edu/sgo/cgi-bin/pub/create_frames.cgi - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Links Google PubMed_UT UniGene LocusLink BLAST MD Consult SGO UT directory Address

A secreted soluble form of ApoE receptor 2 acts as a dominant-negative receptor and inhibits Reelin signaling. Specialized neurons throughout the developing central nervous system secrete Reelin, which binds to ApoE receptor 2 (ApoER2) and very low density lipoprotein receptor (VLDLR), triggering a signal cascade that guides neurons to their correct position. Binding of Reelin to ApoER2 and VLDLR induces phosphorylation of Dab1, which binds to the intracellular domains of both receptors. Due to differential splicing, several isoforms of ApoER2 differing in their ligand-binding and intracellular domains exist. One isoform harbors four binding repeats plus an adjacent short 13 amino acid insertion containing a furin cleavage site. It is not known whether furin processing of this ApoER2 variant actually takes place and, if so, whether the produced fragment is secreted. Here we demonstrate that cleavage of this ApoER2 variant does indeed take place, and that the resulting receptor fragment consisting of the entire ligand-binding domain is secreted as soluble polypeptide. This receptor fragment inhibits Reelin signaling in primary neurons, indicating that it can act in a dominant-negative fashion in the regulation of Reelin signaling during embryonic brain development. Reelin gene alleles and susceptibility to autism spectrum disorders. A polymorphic trinucleotide repeat (CGG/GCC) within the human Reelin

Done

Home



© 2002-2003 Dr. Michael Berry, Dr. Ramin Homayouni, Kevin Heinrich, Lai Wei

Queries

1. AV263736 (reln)

Successfully found 1 out of 1 genes.
Queries were performed using 15 factors.

AV263736 (reln)

1 reln	LocusLink	reelin rl reeler
0.817371 lrp8		
0.799684 dab1	LocusLink	disabled homolog 1 drosophila scm scr yot scrambler
0.653108 vldlr	LocusLink	very low density lipoprotein receptor
0.531238 robo1	LocusLink	
0.498368 rab1a		
0.398951 dll1	LocusLink	delta-like 1 drosophila delta 1
0.380493 cdk5	LocusLink	cyclin-dependent kinase 5 crk6
0.322634 lrp1	LocusLink	low density lipoprotein receptor-related protein 1 lrp a2mr cd91 ai316852
0.312118 wnt5b		

Done

Internet

Reelin Keyword Query



http://shad.cs.utk.edu/sgo/cgi-bin/pub/start.cgi - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Links Google PubMed_UT UniGene LocusLink BLAST MD Consult SGO UT directory Address

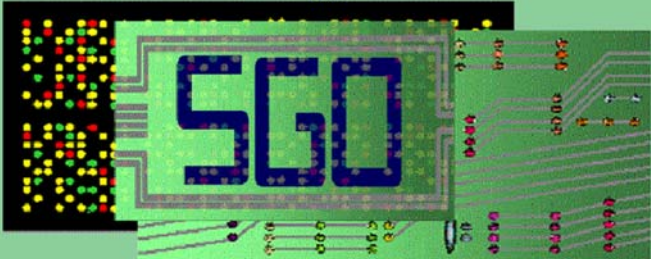
http://shad.cs.utk.edu/sgo/cgi-bin/pub/create_frames.cgi - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Links Google PubMed_UT UniGene LocusLink BLAST MD Consult SGO UT directory Address

Home

A secreted soluble form of ApoE receptor 2 acts as a dominant-negative receptor and inhibits Reelin signaling. Specialized neurons throughout the developing central nervous system secrete Reelin, which binds to ApoE receptor 2 (ApoER2) and very low density lipoprotein receptor (VLDLR), triggering a signal cascade that guides neurons to their correct position. Binding of Reelin to ApoER2 and VLDLR induces phosphorylation of Dab1, which binds to the intracellular domains of both receptors. Due to differential splicing, several isoforms of ApoER2 differing in their ligand-binding and intracellular domains exist. One isoform harbors four binding repeats plus an adjacent short 13 amino acid insertion containing a furin cleavage site. It is not known whether furin processing of this ApoER2 variant actually takes place and, if so, whether the produced fragment is secreted. Here we demonstrate that cleavage of this ApoER2 variant does indeed take place, and that the resulting receptor fragment consisting of the entire ligand-binding domain is secreted as soluble polypeptide. This receptor fragment inhibits Reelin signaling in primary neurons, indicating that it can act in a dominant-negative fashion in the regulation of Reelin signaling during embryonic brain development. Reelin gene alleles and susceptibility to autism spectrum disorders. A polymorphic trinucleotide repeat (CGG/GCC) within the human Reelin



© 2002-2003 Dr. Michael Berry, Dr. Ramin Homayouni, Kevin Heinrich, Lai Wei

Queries

1. [reelin](#)
2. [Reelin](#)

reelin

0.881131 reln	LocusLink	reelin rl reeler
0.696037 dab1	LocusLink	disabled homolog 1 drosophila scm scr yot scrambler
0.689973 lrp8		
0.473022 vldlr	LocusLink	very low density lipoprotein receptor
0.295902 cdk5	LocusLink	cyclin-dependent kinase 5 crk6
0.277224 rab1a		
0.248064 robo1	LocusLink	
0.131076 cdk5r	LocusLink	cyclin-dependent kinase 5, regulatory subunit p35 p35 d11bwg0379e
0.120507 cdk5r2	LocusLink	cyclin-dependent kinase 5, regulatory subunit 2 p39
0.112751 dlx1	LocusLink	delta like 1 drosophila delta1

Done

Internet

Evaluation of Results



➤ Recall (completeness)

$$\text{Recall} = \frac{\text{Relevant documents retrieved}}{\text{Total \# relevant documents}}$$

➤ Precision (accuracy)

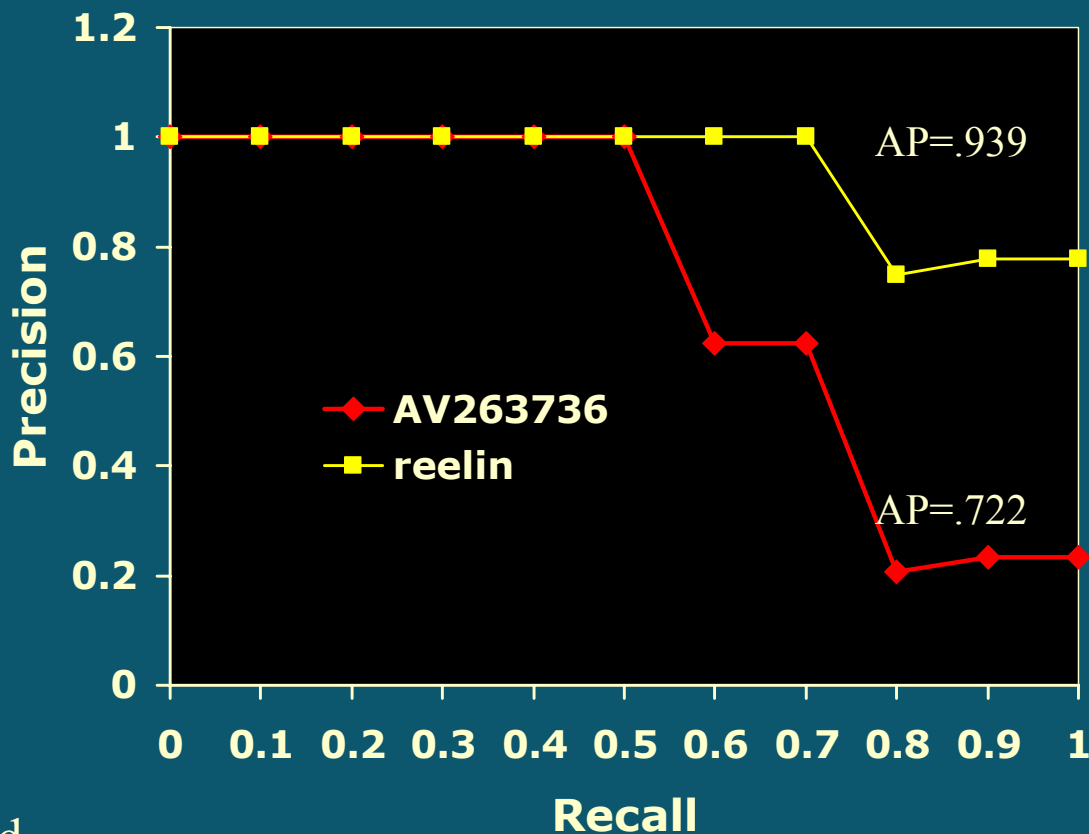
$$\text{Precision} = \frac{\text{Relevant documents retrieved}}{\text{Total \# documents retrieved}}$$

➤ Recall and precision are inversely related.



SGO Performance on Reelin

7 Primary Genes

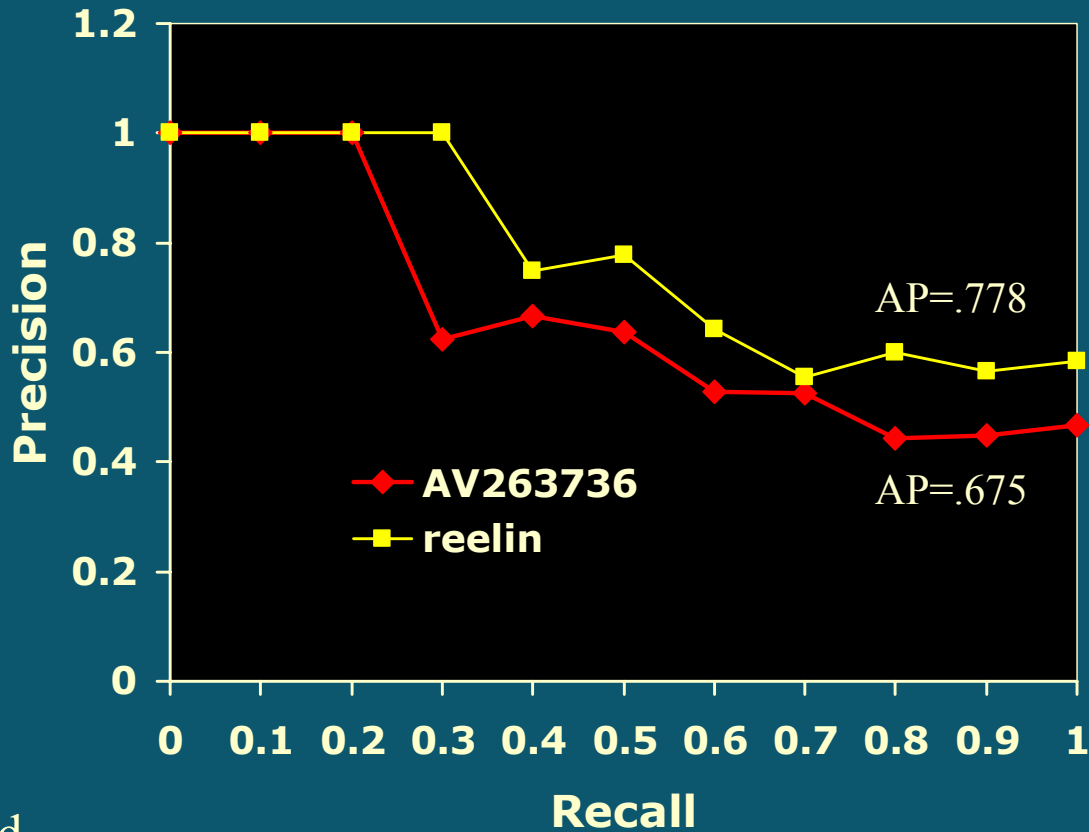


11-pt interpolated
average precision



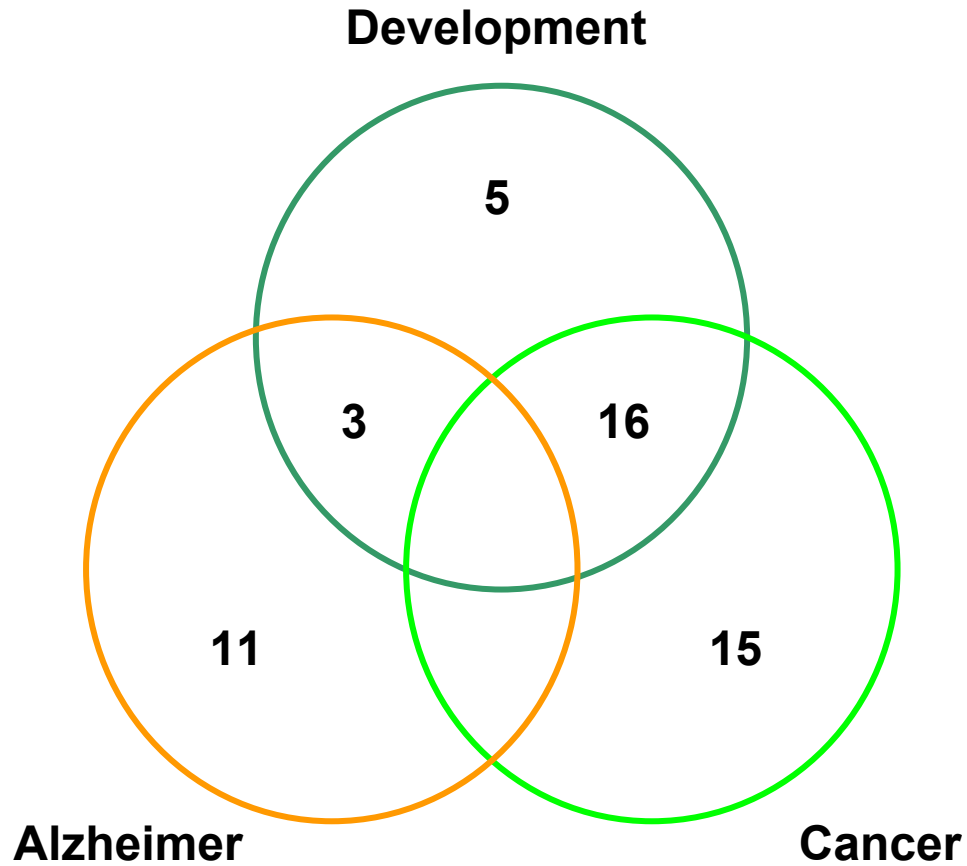
SGO Performance on Reelin

14 Primary and Secondary Genes

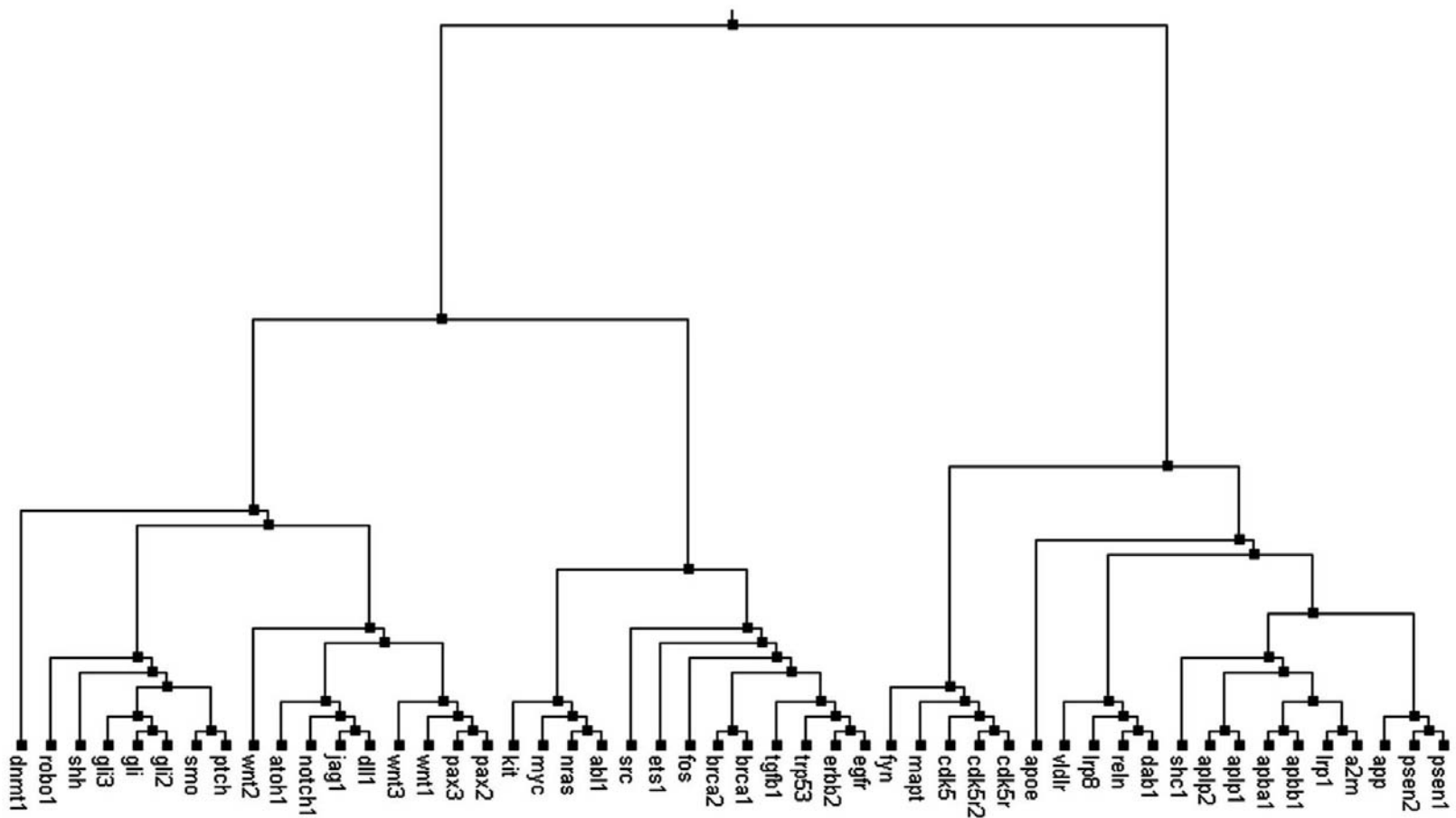


11-pt interpolated
average precision

50-Gene Document Collection



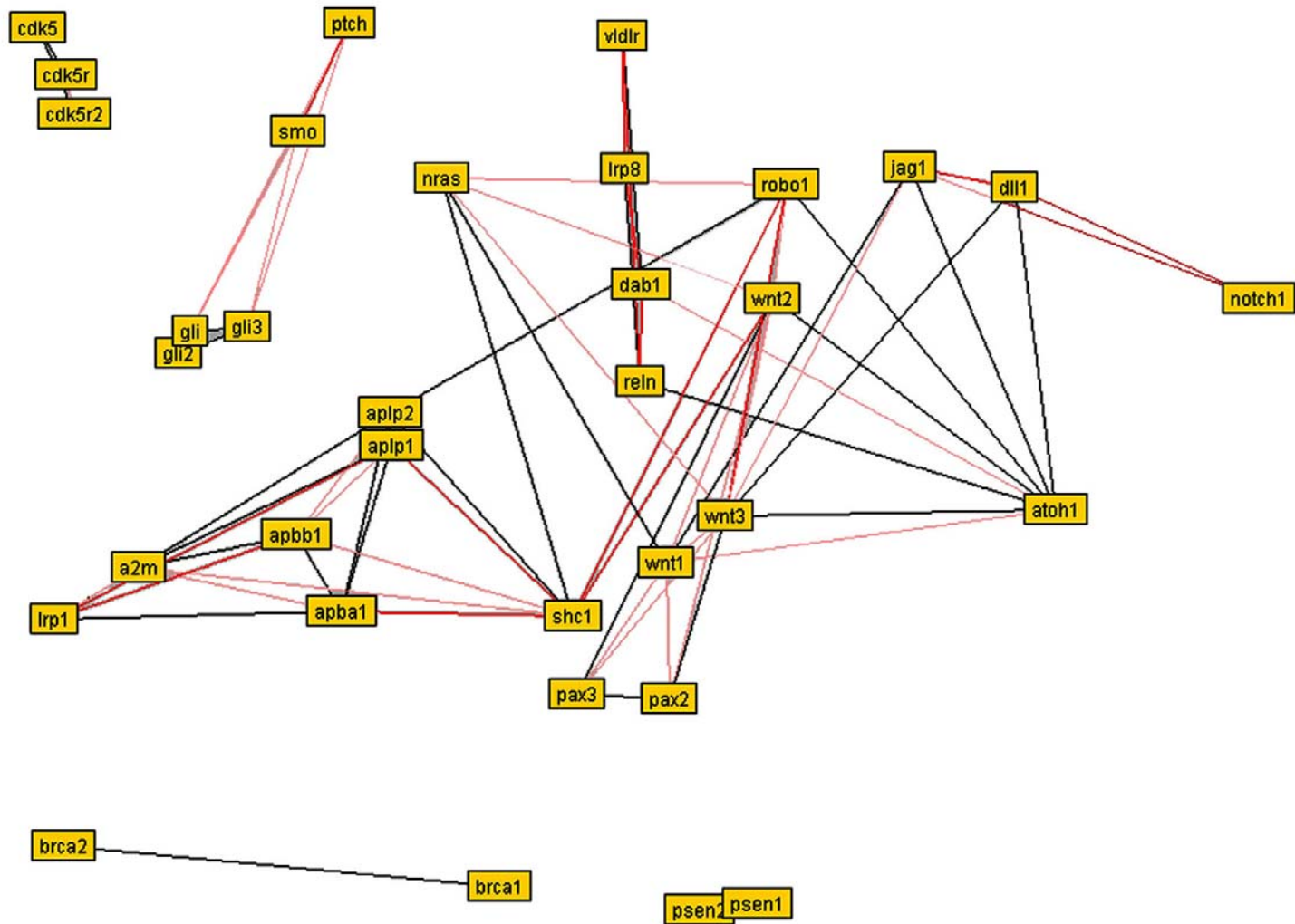
Hierarchical Tree



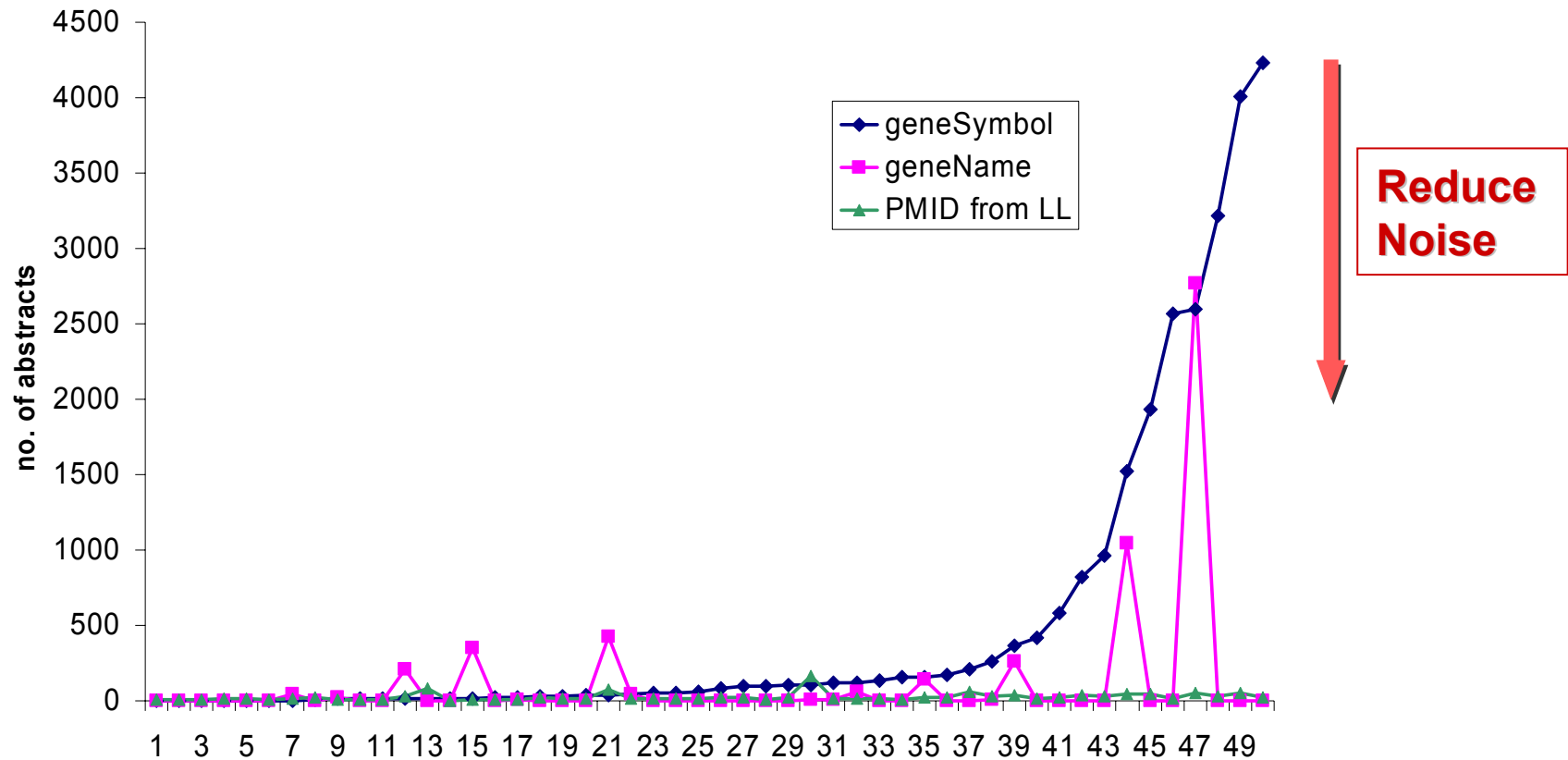
Development Cancer Development Alzheimer



Unrooted Tree



Variation in Abstract Representation



Abstract References in LocusLink



NCBI LocusLink

PubMed Entrez BLAST OMIM Map Viewer Taxonomy Structure

Search LocusLink Display Brief Organism:

All

Query: Go Clear

View Mm Cdk5r2 One of 1 Loci Save All Loci

ABCDEFGHIJKLMNOPQRSTUVWXYZ

Click to view mRNA-Genomic Alignments (spanning 1109 bps)

Gene PUB UNIGENE MAP HOMOL MGI e! UCSC

Mus musculus Official Gene Symbol and Name (MGI)

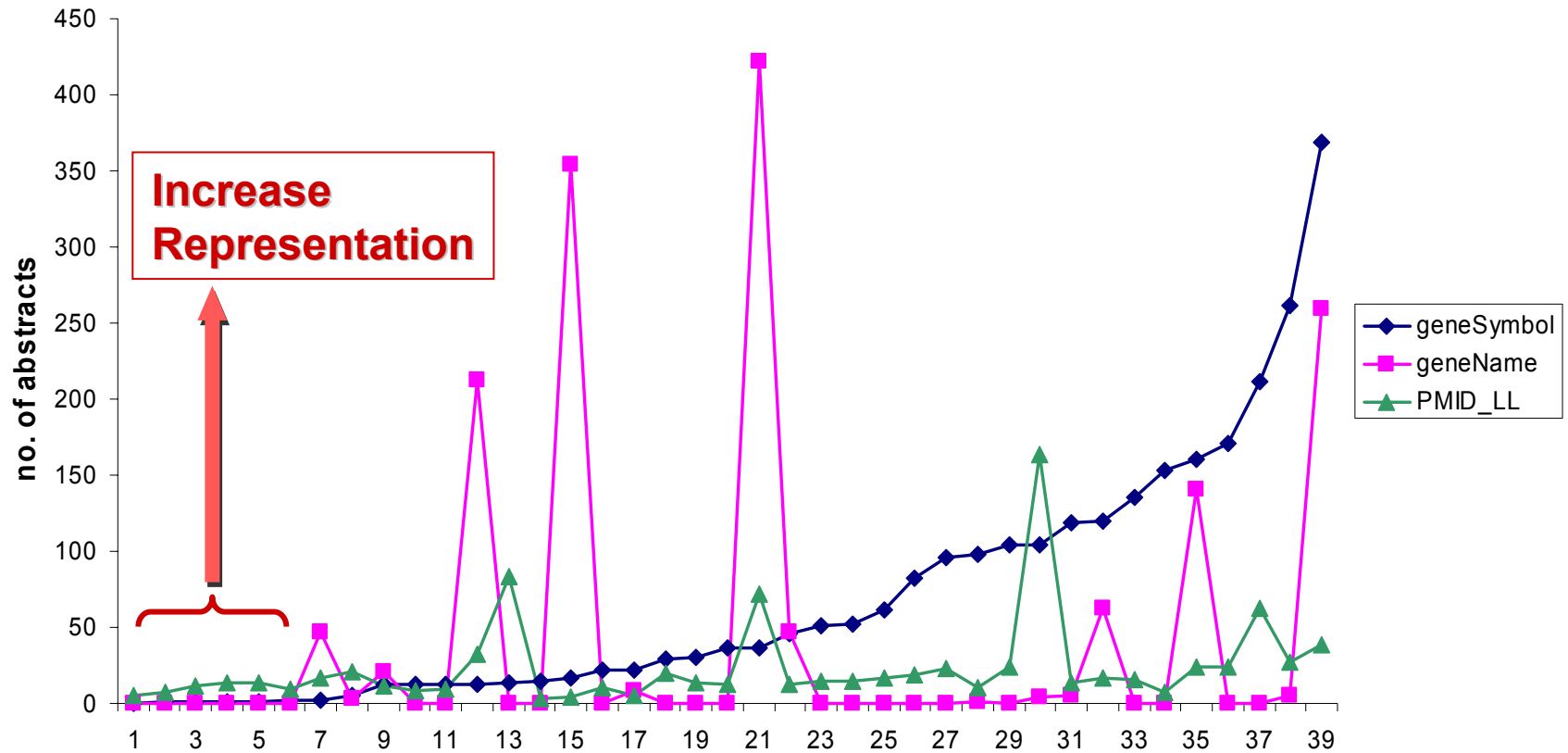
Cdk5r2: cyclin-dependent kinase 5, regulatory subunit 2 (p39)

LocusID: 12570

Overview ?

Locus Type:	gene with protein product, function known or inferred
Product:	cyclin-dependent kinase 5, regulatory subunit 2 (p39)
Alternate Symbols:	p39

Gene symbols and names that are not used in the literature



Alternate Names and Aliases



NCBI LocusLink

PubMed Entrez BLAST OMIM Map Viewer Taxonomy Structure

Search LocusLink Display Brief Organism:

All

Query: Go Clear

View Mm Cdk5r2 One of 1 Loci Save All Loci

ABCDEFGHIJKLMNOPQRSTUVWXYZ

Click to Display mRNA-Genomic Alignments (spanning 1109 bps)

Gene PUB UNIGENE MAP HOMOL MGI e! UCSC

Mus musculus Official Gene Symbol and Name (MGI)

Cdk5r2: cyclin-dependent kinase 5, regulatory subunit 2 (p39)

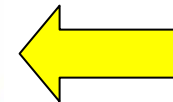
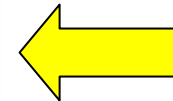
LocusID: 12570

Overview ?

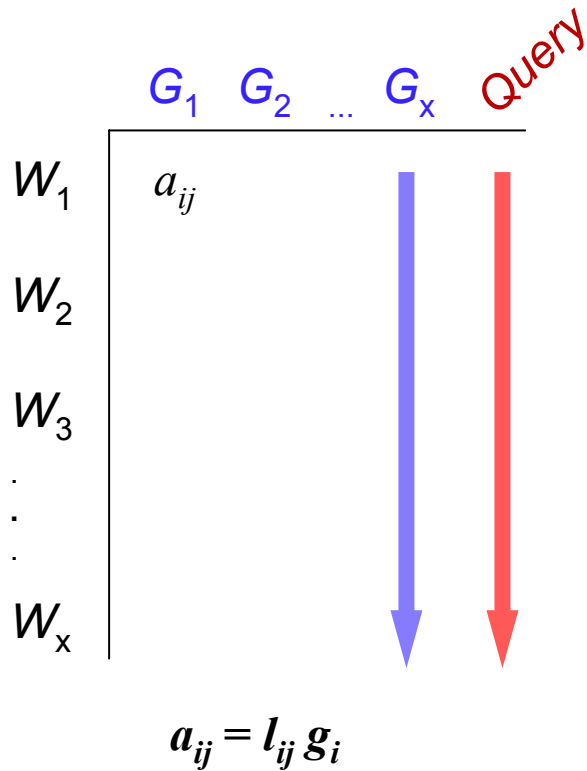
Locus Type: gene with protein product, function known or inferred

Product: cyclin-dependent kinase 5, regulatory subunit 2 (p39)

Alternate Symbols: p39



Log-entropy Term Weighting



$$l_{ij} = \log_2 (1 + f_{ij})$$

$$g_i = 1 + \left(\frac{\sum_j (p_{ij} \log_2 (p_{ij}))}{\log_2 n} \right)$$

$$p_{ij} = \frac{f_{ij}}{\sum_j f_{ij}}$$

Top Terms in Gene Document

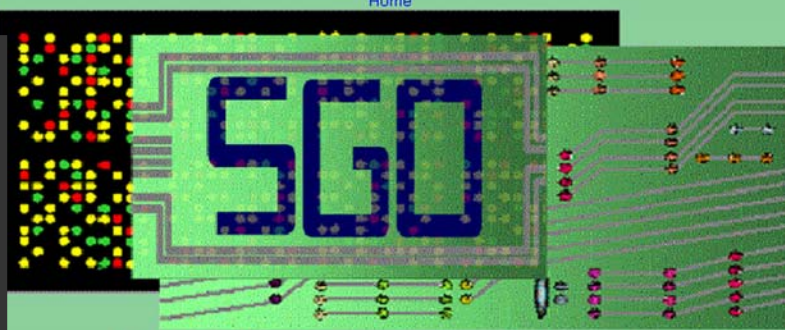


reln

Top 100 Terms

reelin (4.0323)	ex
reeler (3.7762)	se
positioning (1.9135)	cg
lissencephaly (1.8491)	all
schizophrenia (1.7113)	far
apoer2 (1.5637)	ma
cr (1.5544)	pu
esophageal (1.5339)	ma
dab1 (1.5118)	pa
vldlr (1.4973)	lar
carcinoma (1.4881)	kd
wild-type (1.4862)	all
cask (1.4288)	hip
psychiatric (1.4266)	ra
apoe (1.3739)	lob
positioned (1.3726)	de
children (1.3303)	hy
migration (1.2912)	ce
cells (1.278)	po
dendritic (1.2653)	lip
apolipoprotein (1.2565)	gr
mutant (1.2346)	bu
migrate (1.2262)	co
recessive (1.2262)	re
serine (1.2252)	ph

reelin	(4.0323)
reeler	(3.7762)
positioning	(1.9135)
lissencephaly	(1.8491)
schizophrenia	(1.7113)
apoer2	(1.5637)
cr	(1.5544)
esophageal	(1.5339)
dab1	(1.5118)
vldlr	(1.4973)
carcinoma	(1.4881)
wild-type	(1.4862)
cask	(1.4288)
psychiatric	(1.4266)
apoe	(1.3739)
positioned	(1.3726)



© 2002-2003 Dr. Michael Berry, Dr. Ramin Homayouni, Kevin Heinrich, Lai Wei

Queries

- AV263736 (reln)

Successfully found 1 out of 1 genes.
Queries were performed using 15 factors.

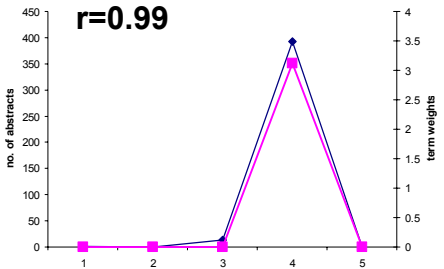
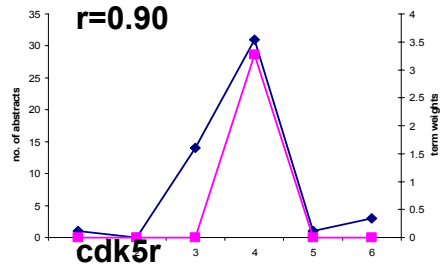
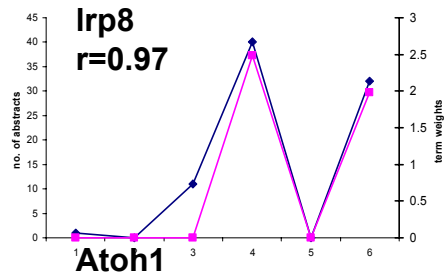
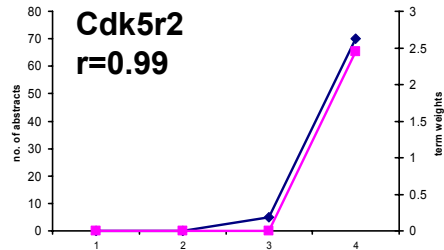
(reln)

1 reln
LocusLink
reelin rl reeler

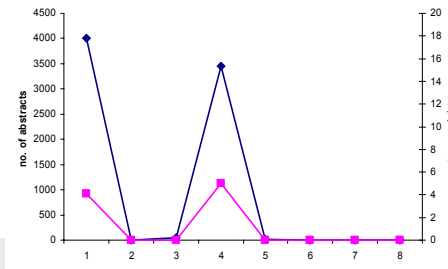
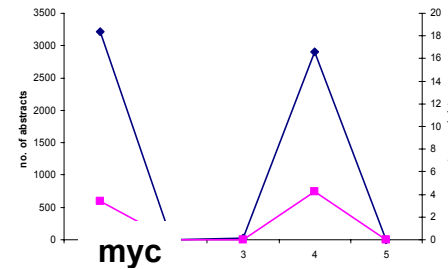
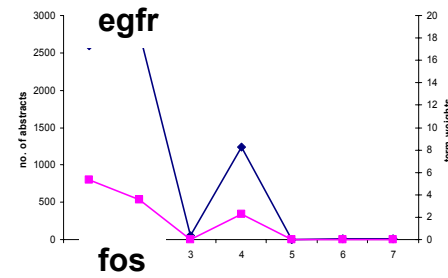
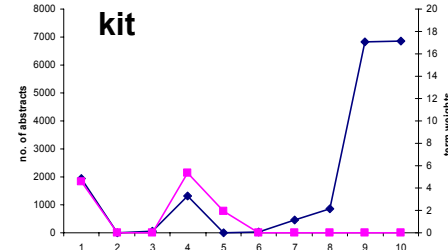
Terms with high log-entropy weights retrieve more abstracts by PubMed query



Under-represented Genes



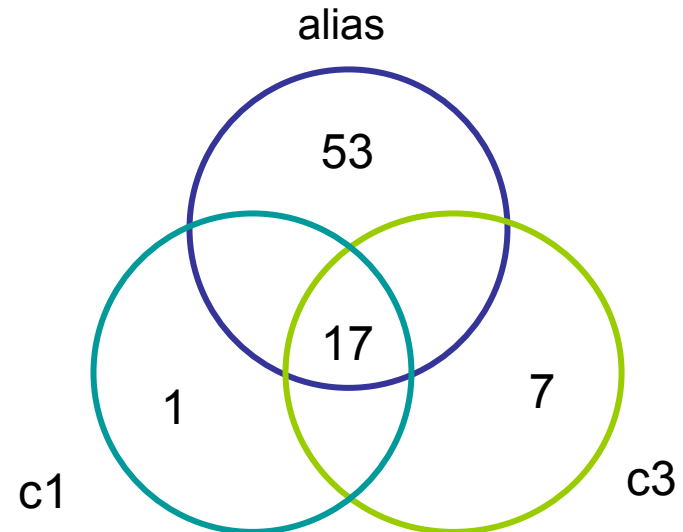
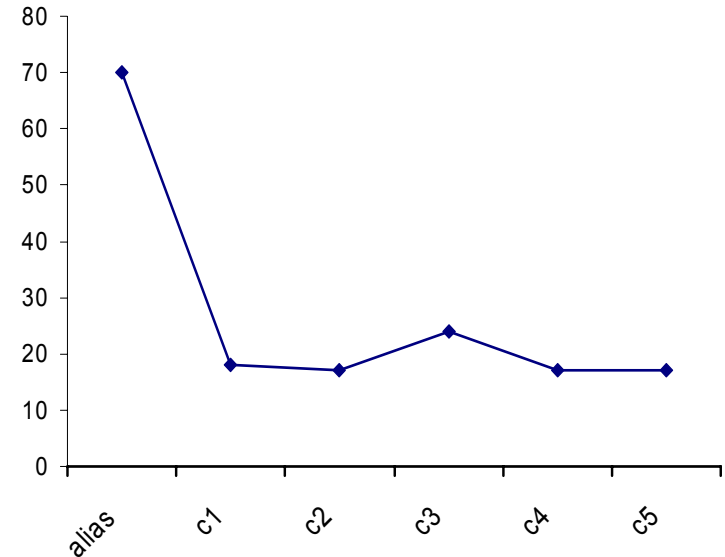
Over-represented Genes



Abstract retrieval by combining weighted terms in gene name, symbol or aliases



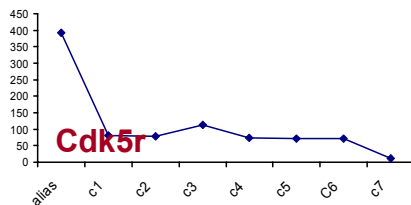
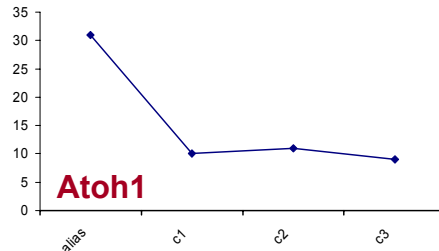
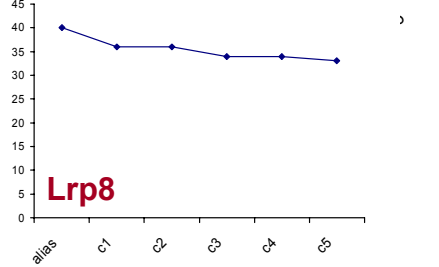
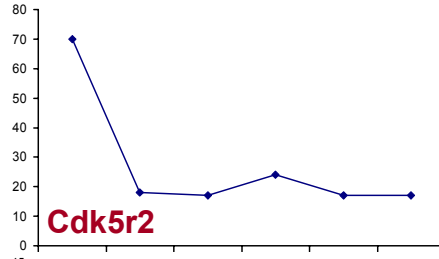
<u>Query</u>	<u>Description</u>	<u># abstracts</u>
symbol	Cdk5r2	0
alias	p39	70
name	cyclin-dependent kinase 5, regulatory subunit 2	0
c1	p39 AND cdk5	18
c2	p39 AND cyclin-dependent	17
c3	p39 AND kinase	24
c4	p39 AND cdk5 AND cyclin-dependent	17
c5	p39 AND cdk5 AND cyclin-dependent AND kinase	17



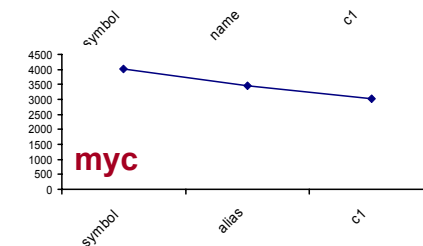
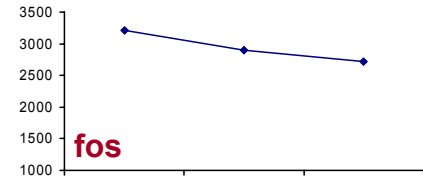
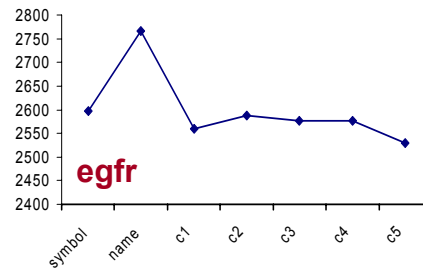
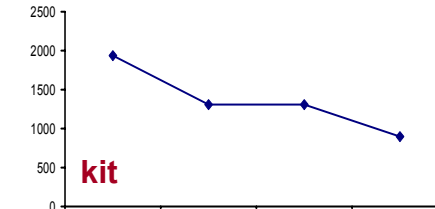
Weighted Combinatorial Queries



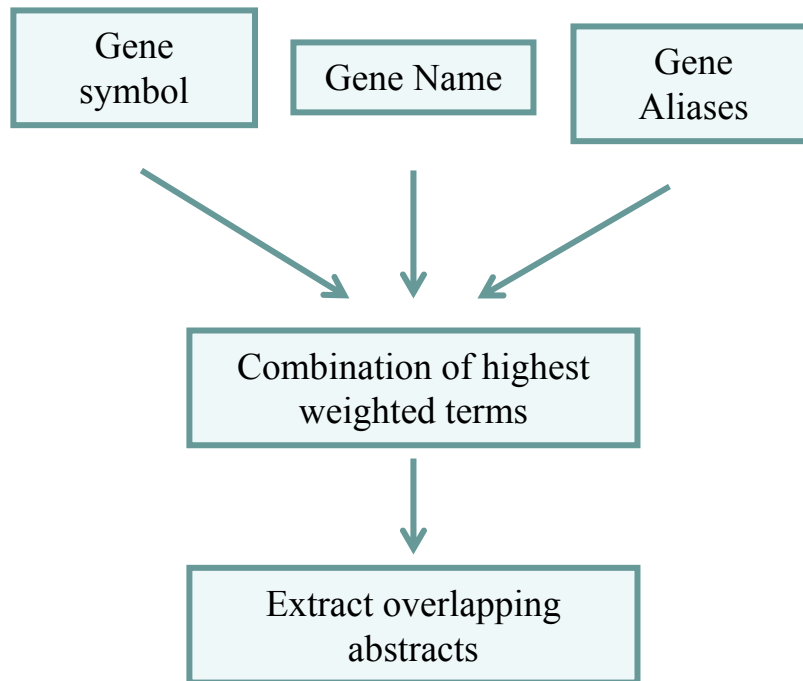
Under-represented Genes



Over-represented Genes



Combinatorial Query Algorithm



RESULTS:
2-59 fold increase in the number of abstracts associated with genes compared to those referenced in LL

Summary and Conclusions

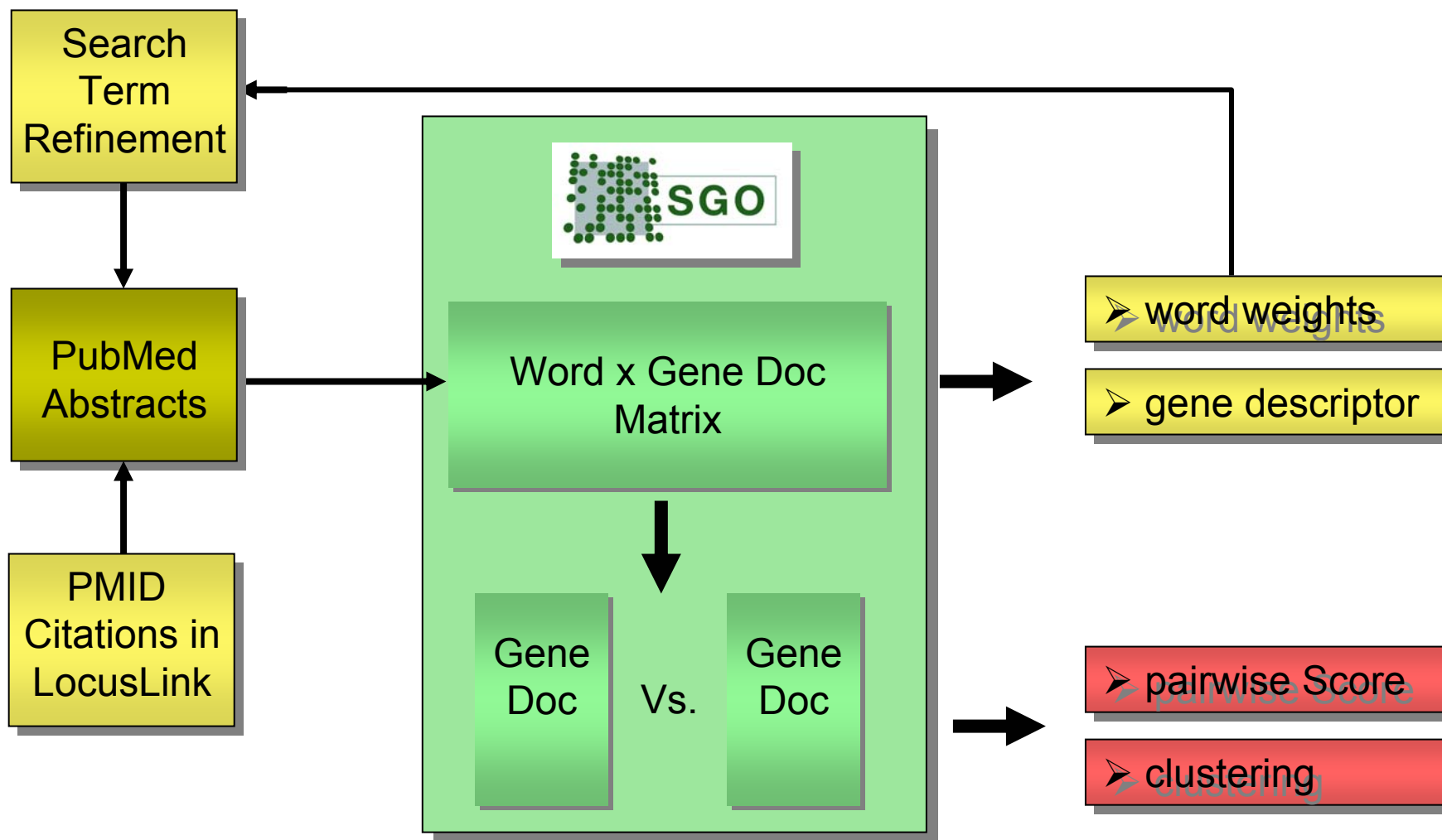


Log-entropy weighting identifies descriptive or 'useful' aliases for genes.

Weighted Combinatorial Querying increases abstracts for under-represented genes and decreases abstracts for over-represented genes with high specificity.

This automated method improves gene abstract assignment 2 to 59 fold beyond those assigned by LocusLink indexers.

SGO overview



Acknowledgments



UT Memphis

Neurology

Lijing Xu, M.S.

Lai Wei, M.D.

Molecular Sciences

Yan Cui, Ph.D.

Mi Zhou, M.S.

UT Knoxville

Computer Science

Michael Berry, Ph.D.

Kevin Heinrich

