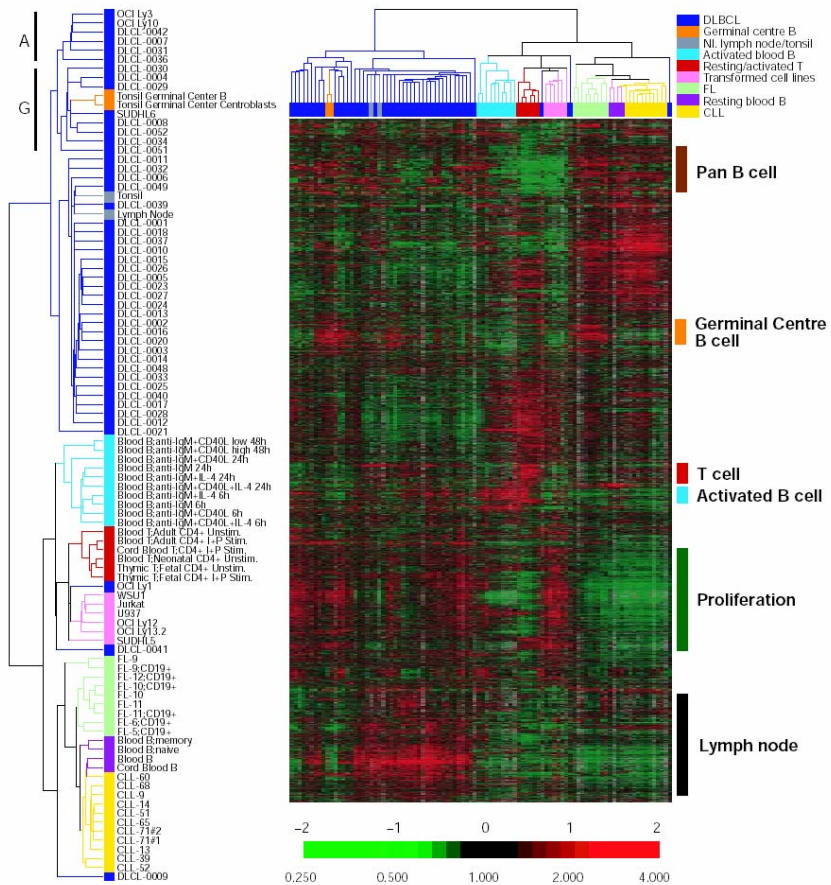


Gene clustering and annotation using GO classifications, MeSH terms and MEDLINE abstracts

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.

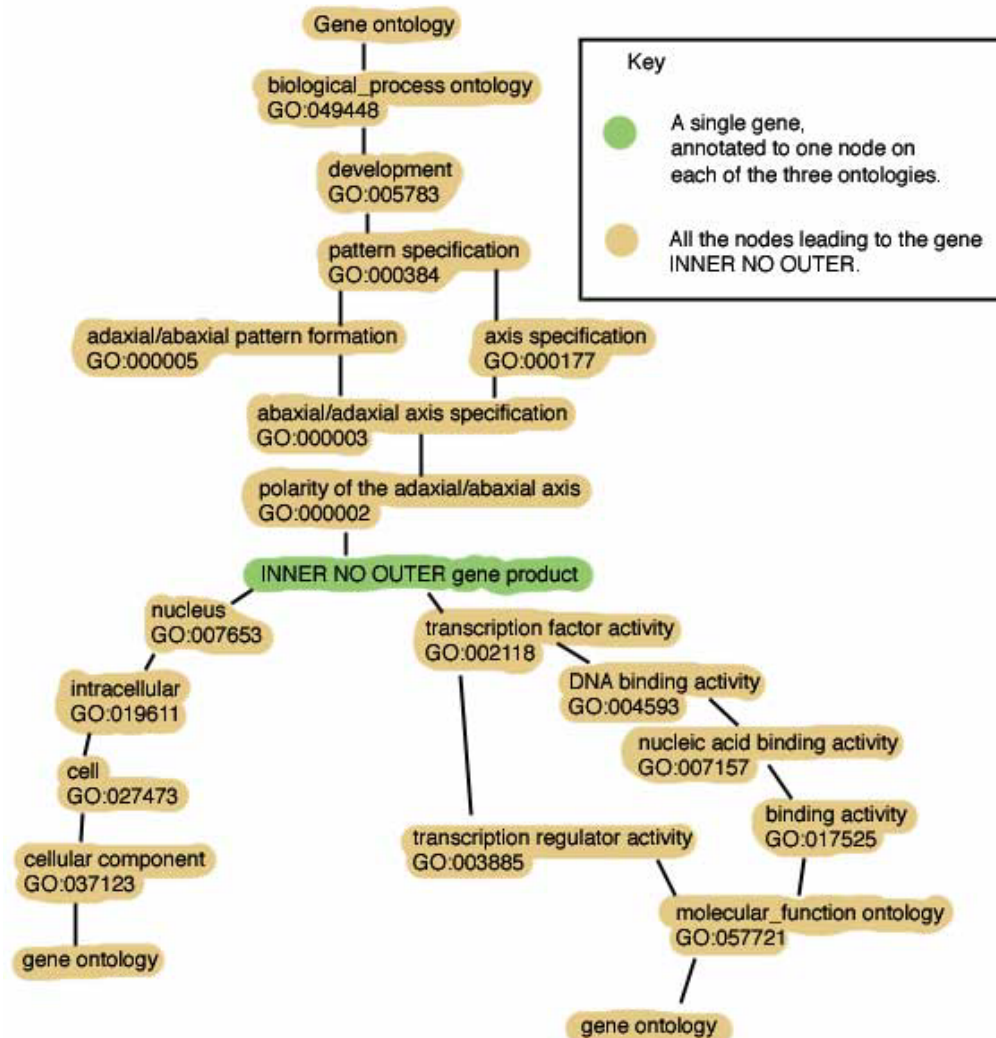
Gene Ontology Consortium

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

- A controlled vocabulary applied to genes in a variety of organisms; updated every 30 minutes!
- Established in 1998 as a collaboration between
 - FlyBase (*Drosophila*)
 - Saccharomyces Genome Database (SGD)
 - Mouse Genome Database (MGD)
- Three main classifications:
 - Molecular Function (7385 terms)
 - Biological Process (8822 terms)
 - Cellular Component (1430 terms)

Gene Ontology Consortium

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



Products of the National Library of Medicine

- **Databases**

 - GenBank, UniGene, LocusLink (Gene)

 - MEDLINE**

 - OMIM

- **Services**

 - HealthSTAR

 - Health Services Research Projects in Progress

 - HSTAT

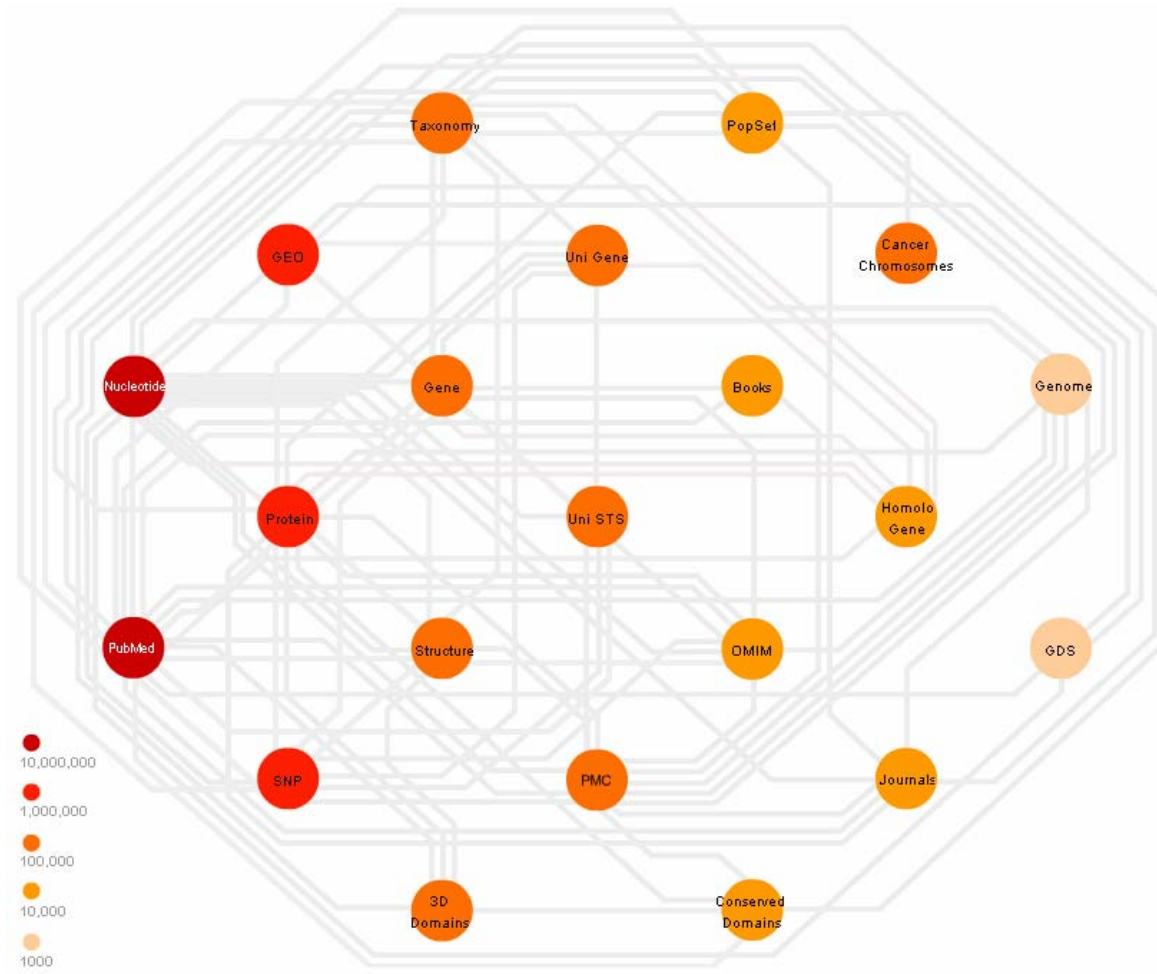
- **Vocabulary**

 - Medical Subject Headings (MeSH)**

 - NLM Classification

 - Unified Medical language Systems

NCBI's *Entrez* <http://www.ncbi.nlm.nih.gov/entrez/>



MEDLINE

Pubmed: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>

- Medical literature, analysis and retrieval system online
- Over 12 million citations from over 4,600 international journals (89% are in English)
- Covers basic biomedical research and clinical sciences dated back to 1966.
- Citations have defined structure. ([Example](#))
- Can be searched through PubMed® using MeSH terms, author names, title words, journal names, phrase, or any combination of these.

Medical Subject Heading (MeSH)

<http://www.nlm.nih.gov/mesh/meshhome.html>

- First edition 1966
- Controlled vocabulary – A Thesaurus
- Used for indexing MEDLINE and Index Medicus
- 22,568 descriptors in hierarchical and alphabetical structure

MeSH Keywords are organized in 15 Concept Hierarchies

- **Anatomy [A]**
- **Organisms [B]**
- **Diseases [C]**
- **Chemicals and Drugs [D]**
- **Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]**
- **Psychiatry and Psychology [F]**
- **Biological Sciences [G]**
- **Physical Sciences [H]**
- **Anthropology, Education, Sociology and Social Phenomena [I]**
- **Technology and Food and Beverages [J]**
- **Humanities [K]**
- **Information Science [L]**
- **Persons [M]**
- **Health Care [N]**
- **Geographic Locations [Z]**

From http://www.nlm.nih.gov/cgi/mesh/2004/MB_cgi

MeSH Hierarchies

[Nervous System Diseases \[C10\]](#)

[Neurologic Manifestations \[C10.597\]](#)

[Bladder, Neurogenic \[C10.597.200\]](#)

[Cerebrospinal Fluid Otorrhea \[C10.597.230\]](#)

[Cerebrospinal Fluid Rhinorrhea \[C10.597.267\]](#)

[Decerebrate State \[C10.597.305\]](#)

[Dyskinesias \[C10.597.350\] +](#)

► [Gait Disorders, Neurologic \[C10.597.404\]](#)

[Gait Apraxia \[C10.597.404.400\]](#)

[Gait Ataxia \[C10.597.404.450\]](#)

[Meningism \[C10.597.544\]](#)

[Neurobehavioral Manifestations \[C10.597.606\] +](#)

[Neurogenic Inflammation \[C10.597.609\]](#)

[Neuromuscular Manifestations \[C10.597.613\] +](#)

[Pain \[C10.597.617\] +](#)

[Paralysis \[C10.597.622\] +](#)

[Paresis \[C10.597.636\] +](#)

[Pupil Disorders \[C10.597.690\] +](#)

[Reflex, Abnormal \[C10.597.704\] +](#)

[Seizures \[C10.597.742\] +](#)

[Sensation Disorders \[C10.597.751\] +](#)

[Vertigo \[C10.597.951\]](#)

[Voice Disorders \[C10.597.975\] +](#)

From <http://www.nlm.nih.gov/mesh/MBrowser.html>

Useful Links

Databases:

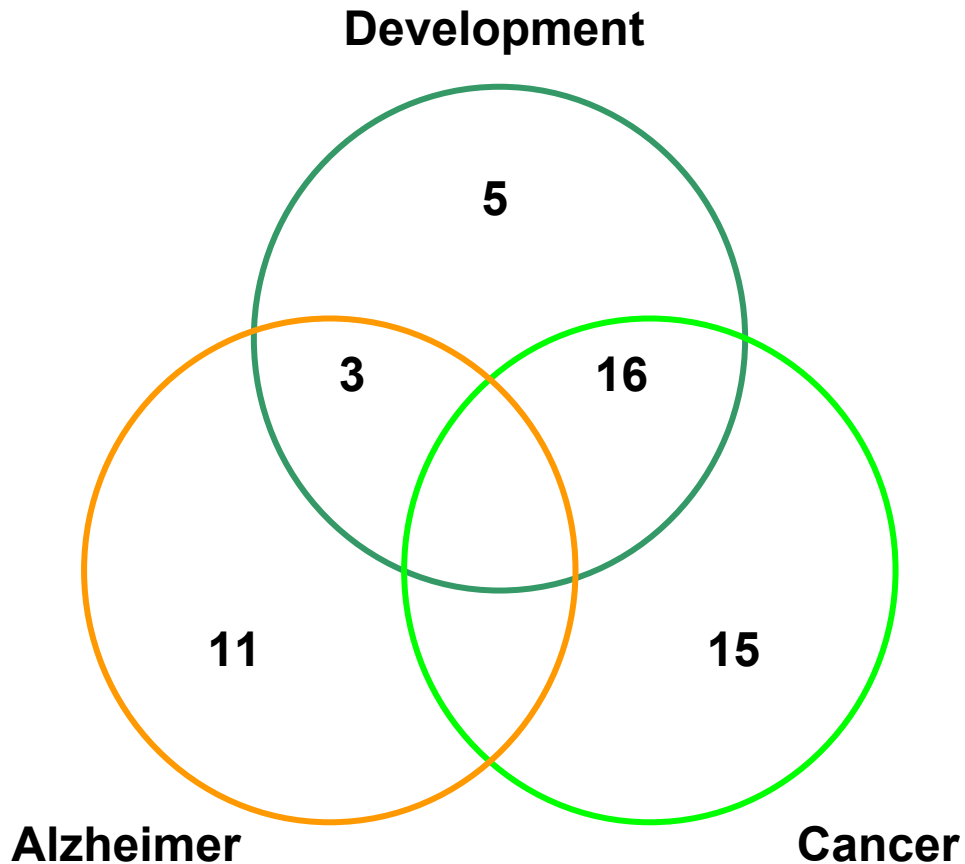
- **GO:** <http://www.geneontology.org/>
- **MeSH:** <http://www.nlm.nih.gov/mesh/meshhome.html>
- **MEDLINE:** <http://www.ncbi.nlm.nih.gov/entrez/>

Programs:

- **GOTM (GO):** <http://genereg.ornl.gov/gotm/>
- **HAPI (MeSH):** <http://array.ucsd.edu/hapi/>
- **PubGene (MEDLINE):** <http://www.pubgene.org/>
- **Chilibot (MEDLINE):** <http://www.chilibot.net/>
- **Arrowsmith (MEDLINE):** <http://arrowsmith.psych.uic.edu/>
- **PubMatrix (MEDLINE):** <http://pubmatrix.grc.nia.nih.gov/>
- **TXTGate (MEDLINE):** <http://www.esat.kuleuven.ac.be/txtgate/>

50 Gene Test Set

Development, Alzheimer's, & Cancer



GO Tree Machine (GOTM) <http://genereg.ornl.gov/gotm/>

Bing Zhang & Jay Snoddy, UTORNL



Gene Ontology Tree Machine

<http://genereg.ornl.gov/gotm>

University of Tennessee and Oak Ridge National Laboratory

[[login](#)] [[register](#)] [[retrieve password](#)]

GOTM (GOTree Machine) is a web-based platform for interpreting microarray data or other interesting gene sets using [Gene Ontology](#) hierarchies. GOTM currently works with human, mouse, rat and fly.

Key features:

- User friendly web-based interface
- Expandable tree for browsing the GO hierarchy, Fixed tree as HTML output for archive, Bar chart for publication
- Statistic analysis indicating GO terms with relatively enriched gene numbers and suggesting biological areas that warrant further study. Sub-tree and DAG visualizing enriched GO categories
- Retrieving subset of genes by GO term or keyword searching.

Screenshots: [[Schematic overview](#)] [[Input view](#)] [[Output view](#)] [[Text output](#)] [[DAG output](#)]

GOTM manual is available [here](#)

GOTM has been published in [BMC Bioinformatics](#). 2004 Feb 18;5(1):16 . If you use GOTM for your research, please cite the paper in your publication.

GOTM is free to academic users after registration.

If you are a registered user, please [login](#).

If you forget your password, [retrieve password](#).

If you are a new user, please [register](#)

GO Tree Machine

Demo GOTM

<http://genereg.ornl.gov/gotm/>

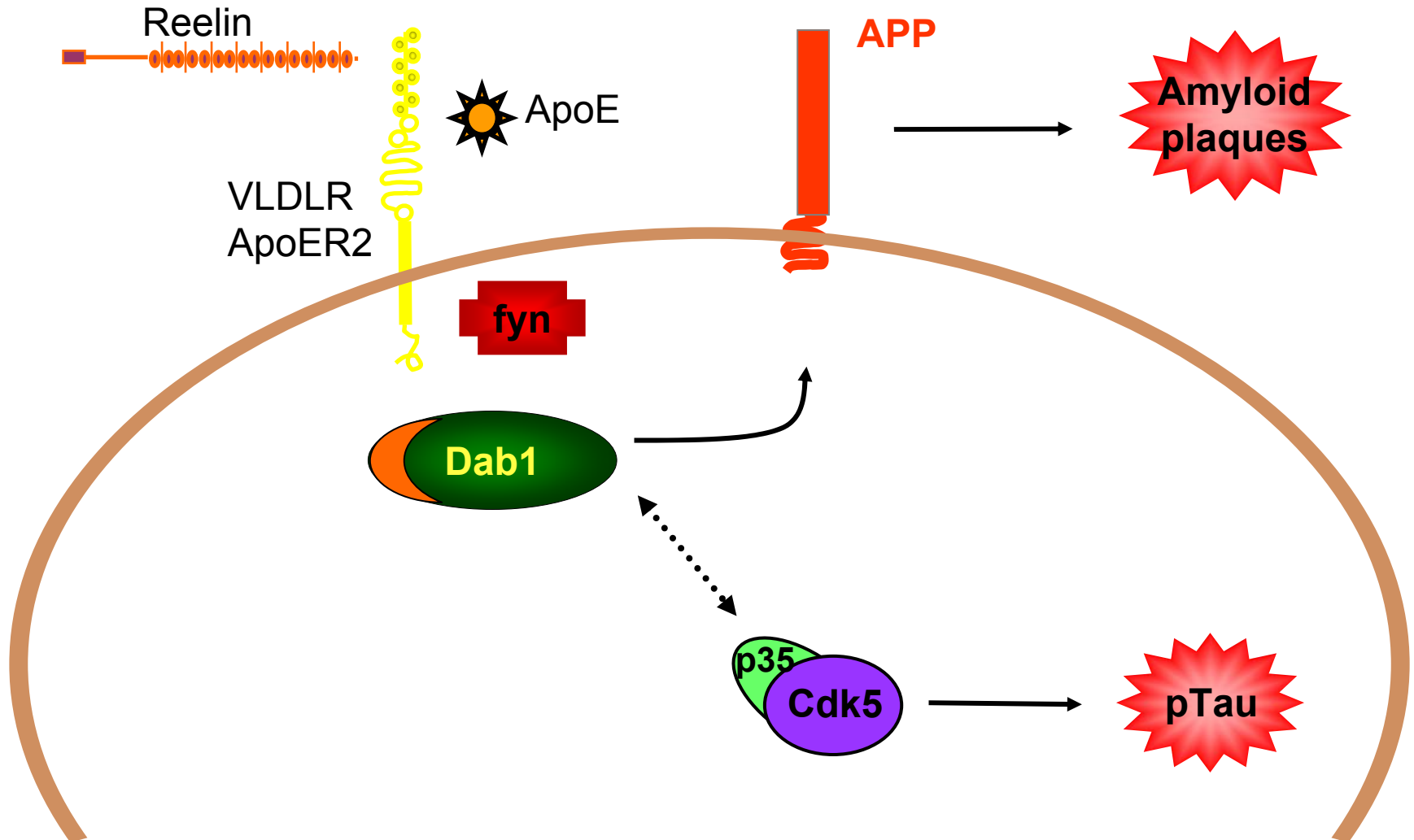
High Density Array Interpreter (HAPI)

<http://array.ucsd.edu/hapi/>



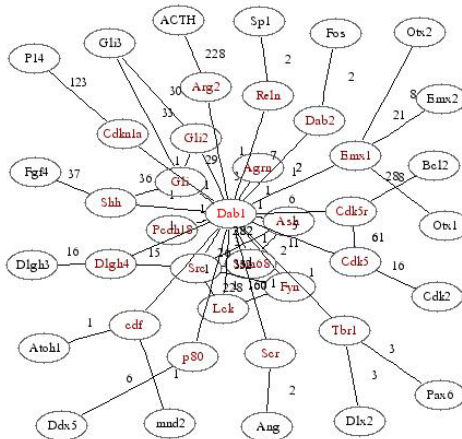
Finds similarities between genes based on co-occurrence of MeSH terms in manually assigned gene abstracts.

Reelin Signaling Pathway



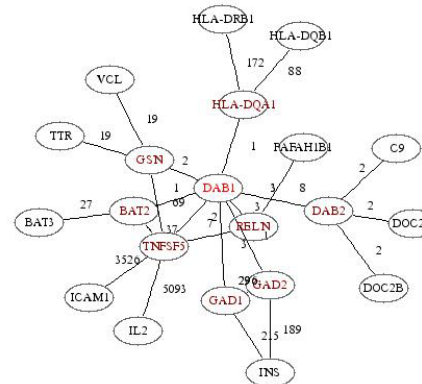
Mouse

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
Dlgh4 1 times
p80 1 times
Lck 1 times
Emx1 1 times
Pcdh18 1 times
Agrn 1 times
Arg2 1 times



Human

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

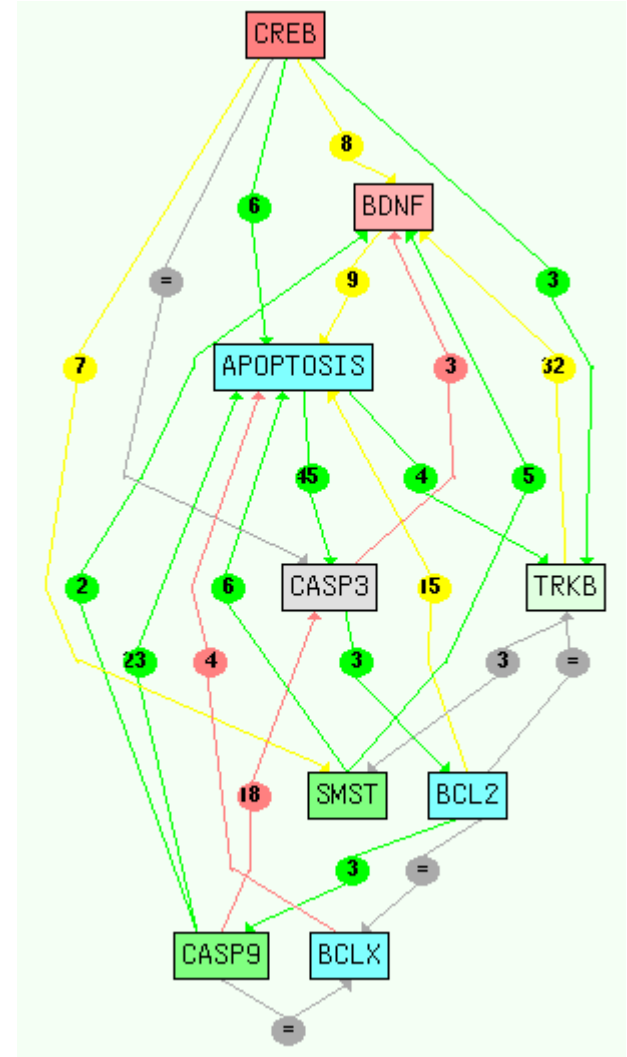


PubGene Query: Dab1

PubMed Query: Dab1 AND Reln = 10

PubMed Query: Dab1 AND reelin = 57 !

- Extracts term-term relationship from Medline abstracts.
- Differentiates interactive (e.g. stimulation or inhibition) and non-interactive (e.g. homology, co-existence, etc.) interactions.
- Color-codes gene expression values when data are provided.
- Automatically suggests new hypothesis based on the literature.

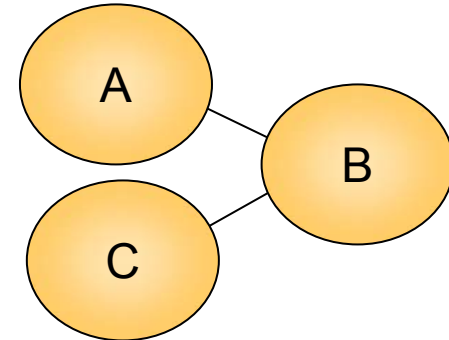


Demo

Chilibot

<http://www.chilibot.net/>

Defining Gene Relationships



➤ Direct Relationship

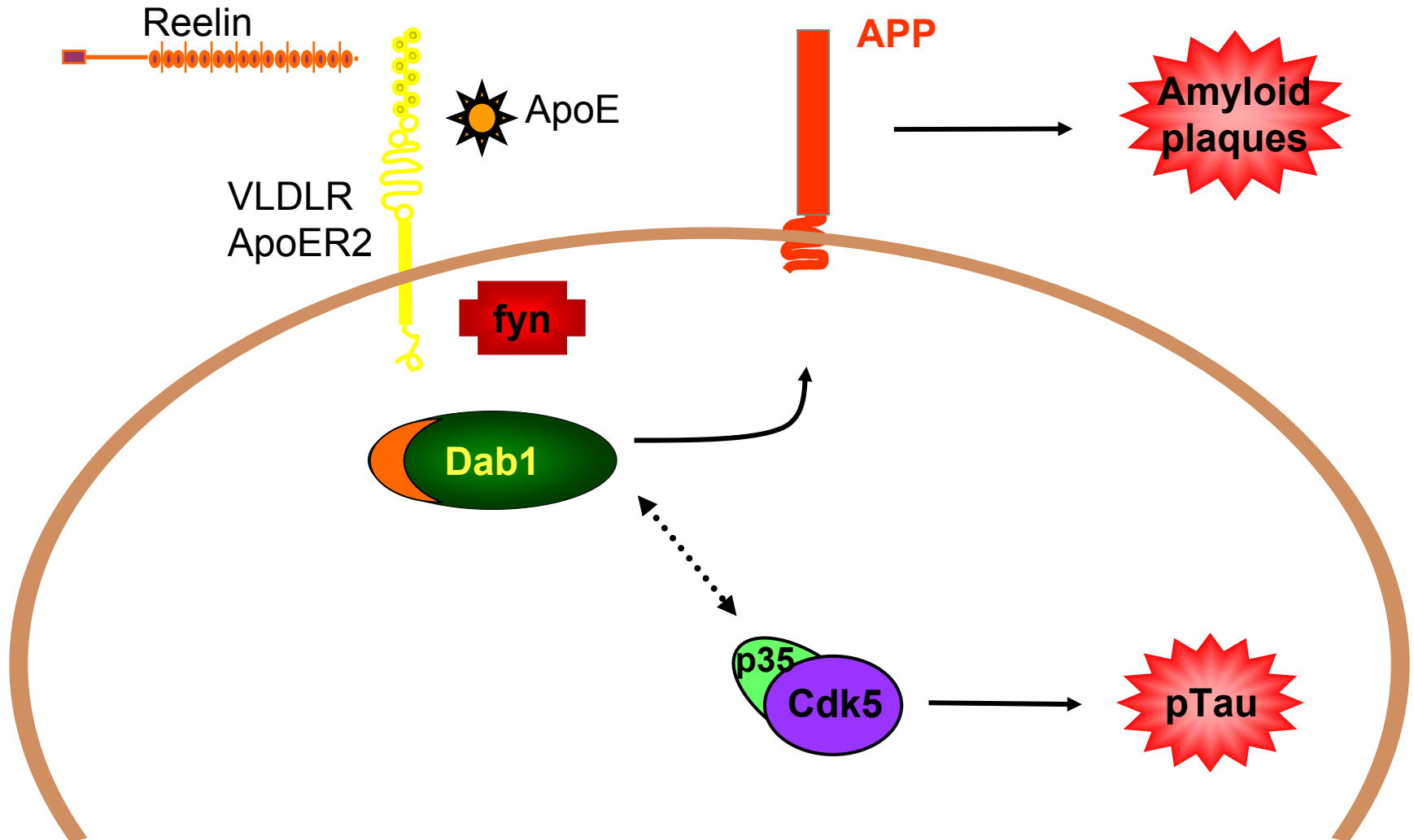
Gene relationships already known (e.g., A-B or B-C)

- 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



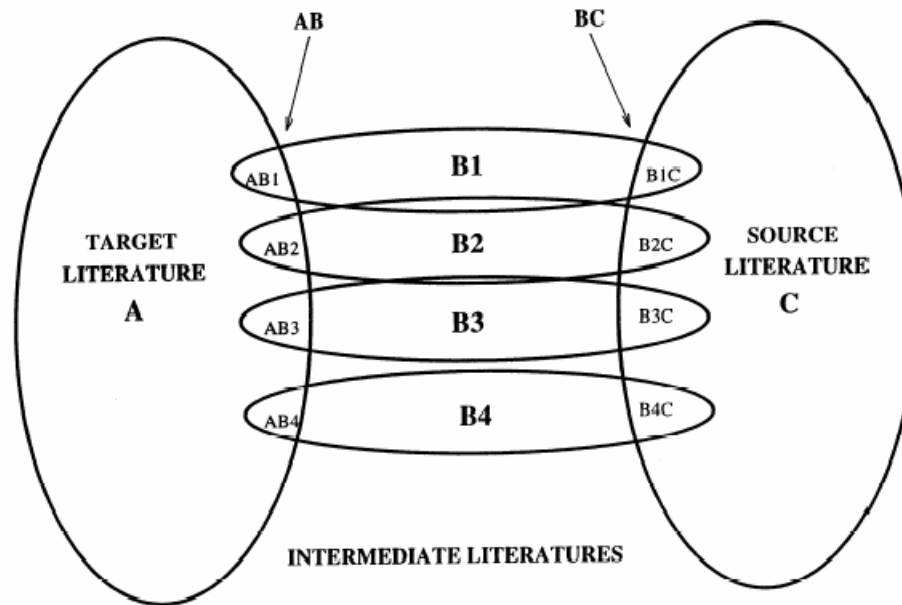


Fig. 1. A Venn diagram that represents sets of articles, or literatures, containing the words A and C in their titles. Sets A and C are linked through intermediate sets B_i ($i = 1, 2, 3, \dots$) which contain the word B_i in their titles and which overlap both A and C. By examining the articles in the pairs of intersections AB_i and B_iC , useful information may be inferred regarding possible biological linkages among A, B and C. (A and C are shown here as having no articles in common. When there is overlap between sets A and C, the articles in the direct intersection should first be identified and evaluated prior to carrying out an ARROWSMITH search.) Modified from [9] with permission.

PubMatrix <http://pubmatrix.grc.nia.nih.gov/>

Becker & Engle, NIH

- Rapid comparison of any list of terms against any other list of terms in PubMed.
- Lists of terms may be gene names, diseases, gene functions, authors, etc.
- Reports back the frequency of co-occurrence between all pairwise comparisons between the two lists as a matrix table.

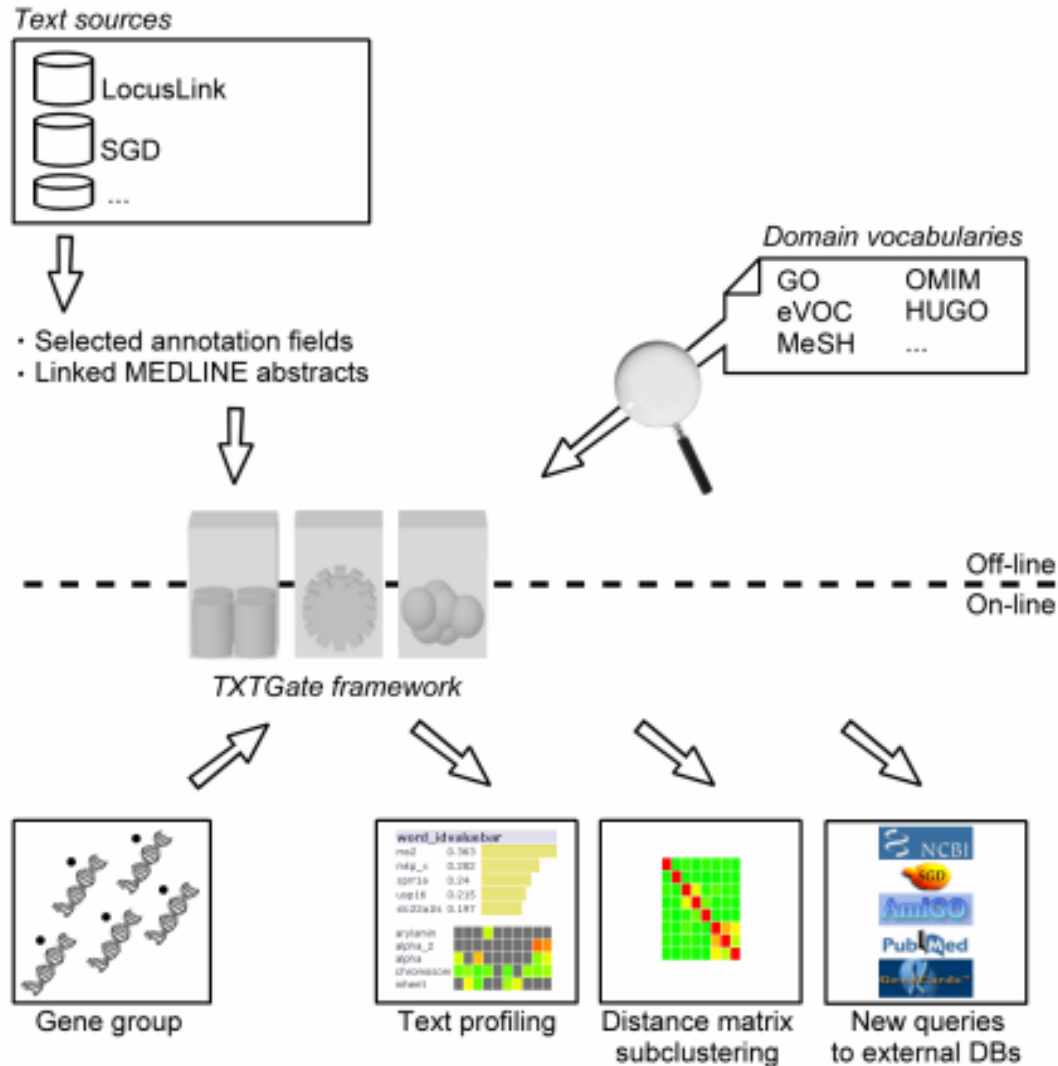


PubMatrix Results for run 3324

Reelin pathway x Search for RHOMAYOUNI						
PubMatrix	brain development	cancer	alzheimer's disease	cell migration	cerebellum	cortex
RELN	<u>0</u>	<u>1</u>	<u>2</u>	<u>7</u>	<u>12</u>	<u>20</u>
VLDLR	<u>0</u>	<u>4</u>	<u>8</u>	<u>10</u>	<u>7</u>	<u>13</u>
APOER2	<u>0</u>	<u>1</u>	<u>5</u>	<u>13</u>	<u>7</u>	<u>17</u>
CDK5R	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>	<u>0</u>
CDK5	<u>0</u>	<u>51</u>	<u>120</u>	<u>34</u>	<u>35</u>	<u>75</u>
GLI2	<u>0</u>	<u>19</u>	<u>0</u>	<u>2</u>	<u>1</u>	<u>0</u>
SRC	<u>0</u>	<u>3385</u>	<u>43</u>	<u>552</u>	<u>81</u>	<u>155</u>
FYN	<u>0</u>	<u>172</u>	<u>12</u>	<u>40</u>	<u>23</u>	<u>36</u>
DAB2	<u>0</u>	<u>19</u>	<u>0</u>	<u>2</u>	<u>0</u>	<u>2</u>
APP	<u>3</u>	<u>307</u>	<u>2307</u>	<u>18</u>	<u>82</u>	<u>468</u>
APLP1	<u>0</u>	<u>3</u>	<u>30</u>	<u>1</u>	<u>2</u>	<u>8</u>

Demo PubMatrix

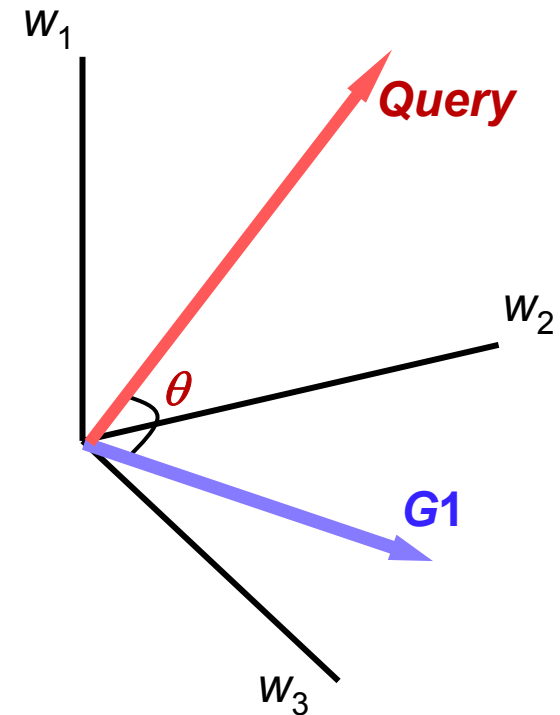
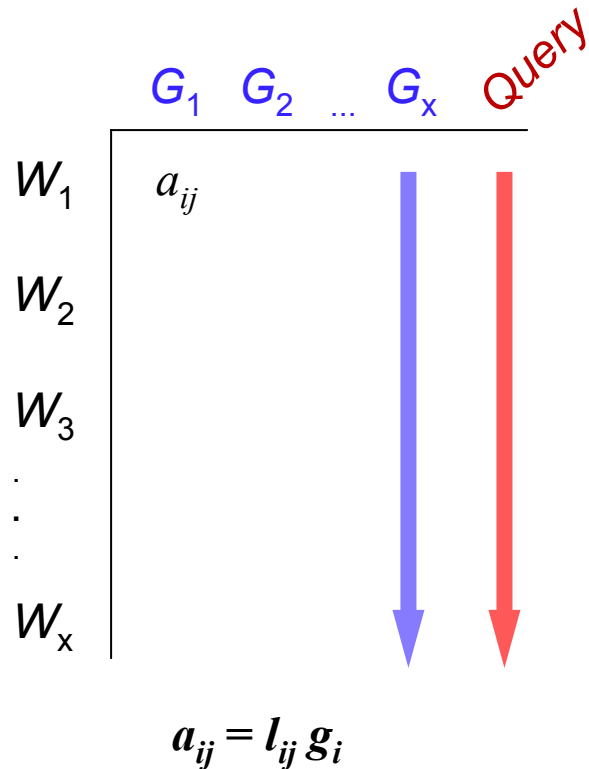
<http://pubmatrix.grc.nia.nih.gov/>



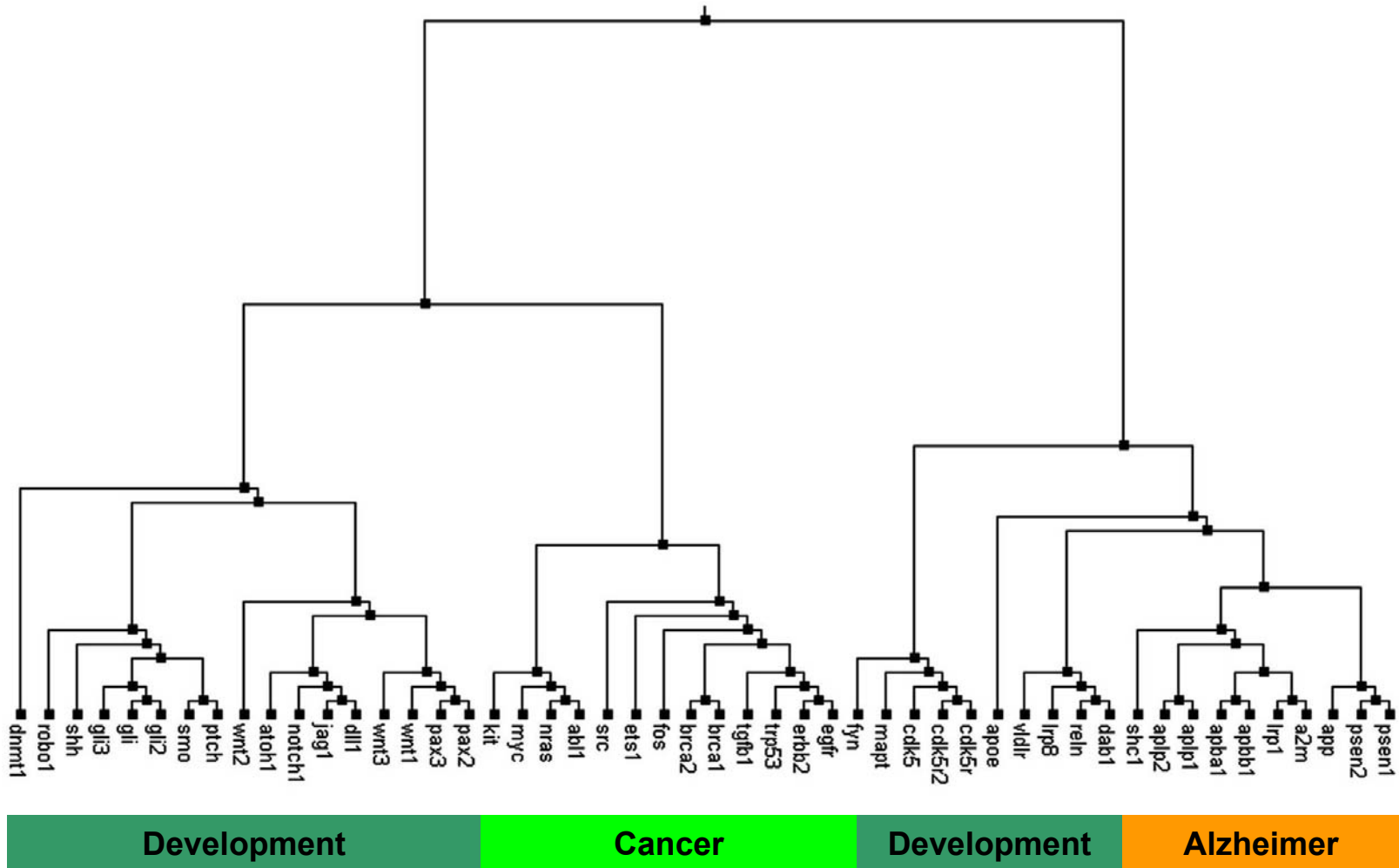
Demo TXTGate

<http://www.esat.kuleuven.ac.be/txtgate/>

Vector Space Model: Latent Semantic Indexing



Hierarchical Tree by Semantic Gene Organizer©



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