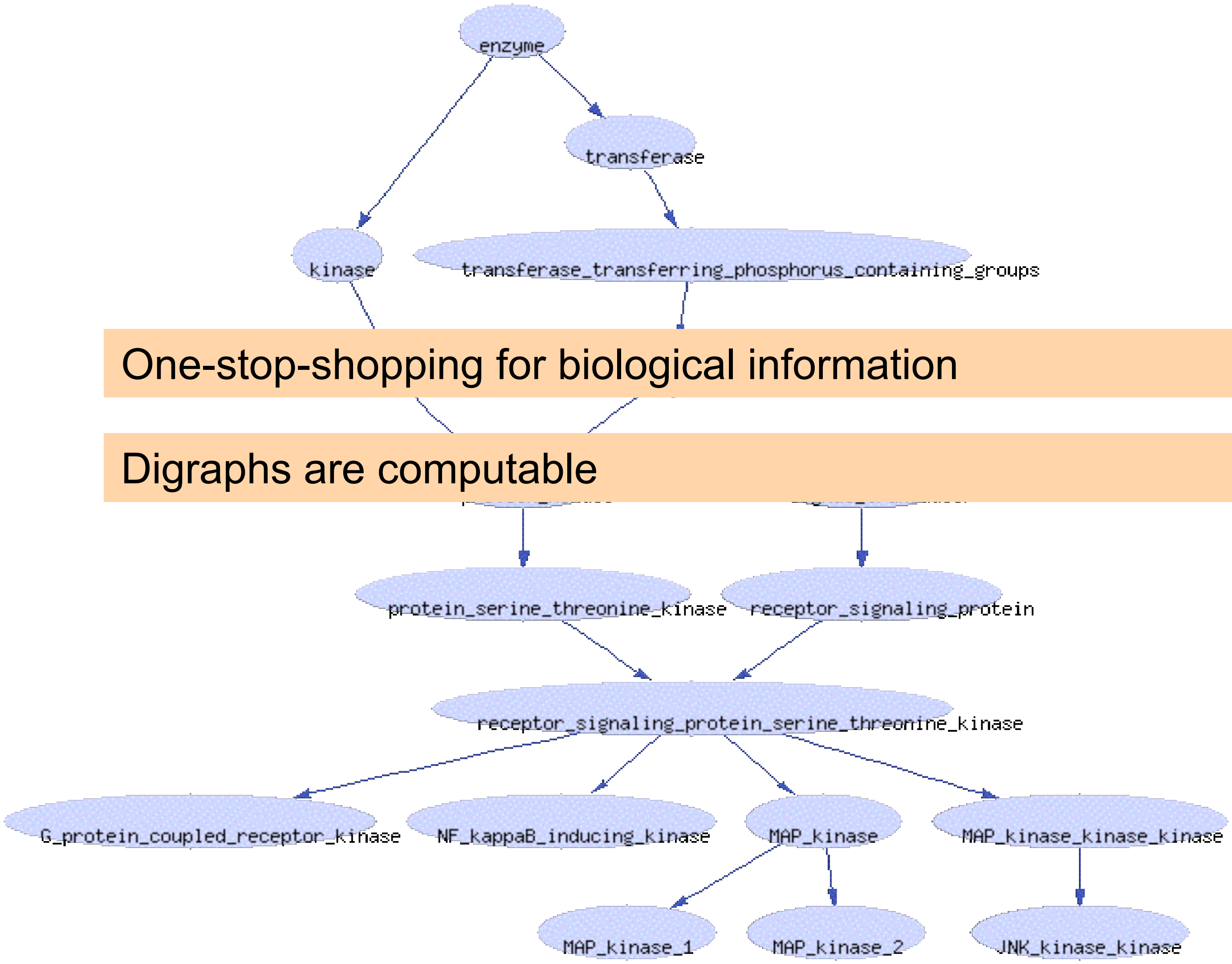
# A Knowledge-Based Clustering Algorithm Driven by Gene Ontology

Jill Cheng

Affymetrix, Inc.

Jan 15, 2004

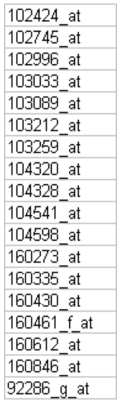
# The DAG structure of Gene Ontology


enzyme
transferase
kinase
transferase_transferring_phosphorus_containing_groups
protein_serine_threonine_kinase
receptor_signaling_protein
receptor_signaling_protein_serine_threonine_kinase
G_protein_coupled_receptor_kinase
NF_kappaB_inducing_kinase
MAP_kinase
MAP_kinase_kinase_kinase
MAP_kinase_1
MAP_kinase_2
JNK_kinase_kinase

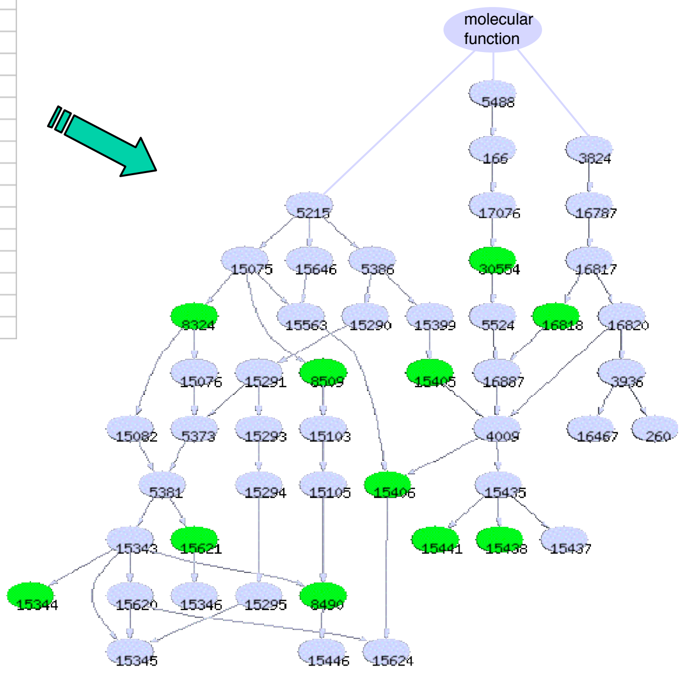
One-stop-shopping for biological information

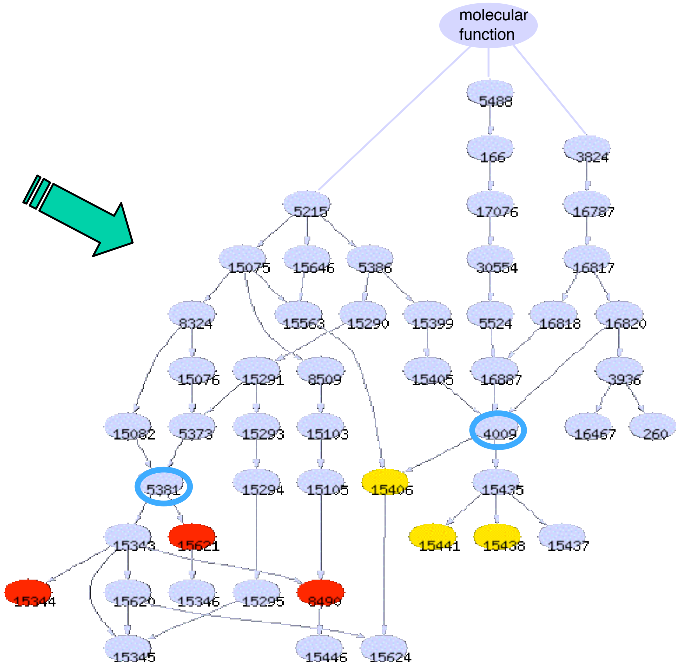
Digraphs are computable

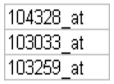
# Goal



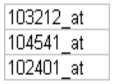
| |
|---|
| 102424_at |
| 102745_at |
| 102996_at |
| 103033_at |
| 103089_at |
| 103212_at |
| 103259_at |
| 104320_at |
| 104328_at |
| 104541_at |
| 104598_at |
| 160273_at |
| 160335_at |
| 160430_at |
| 160461_f_at |
| 160612_at |
| 160846_at |
| 92286_g_at |



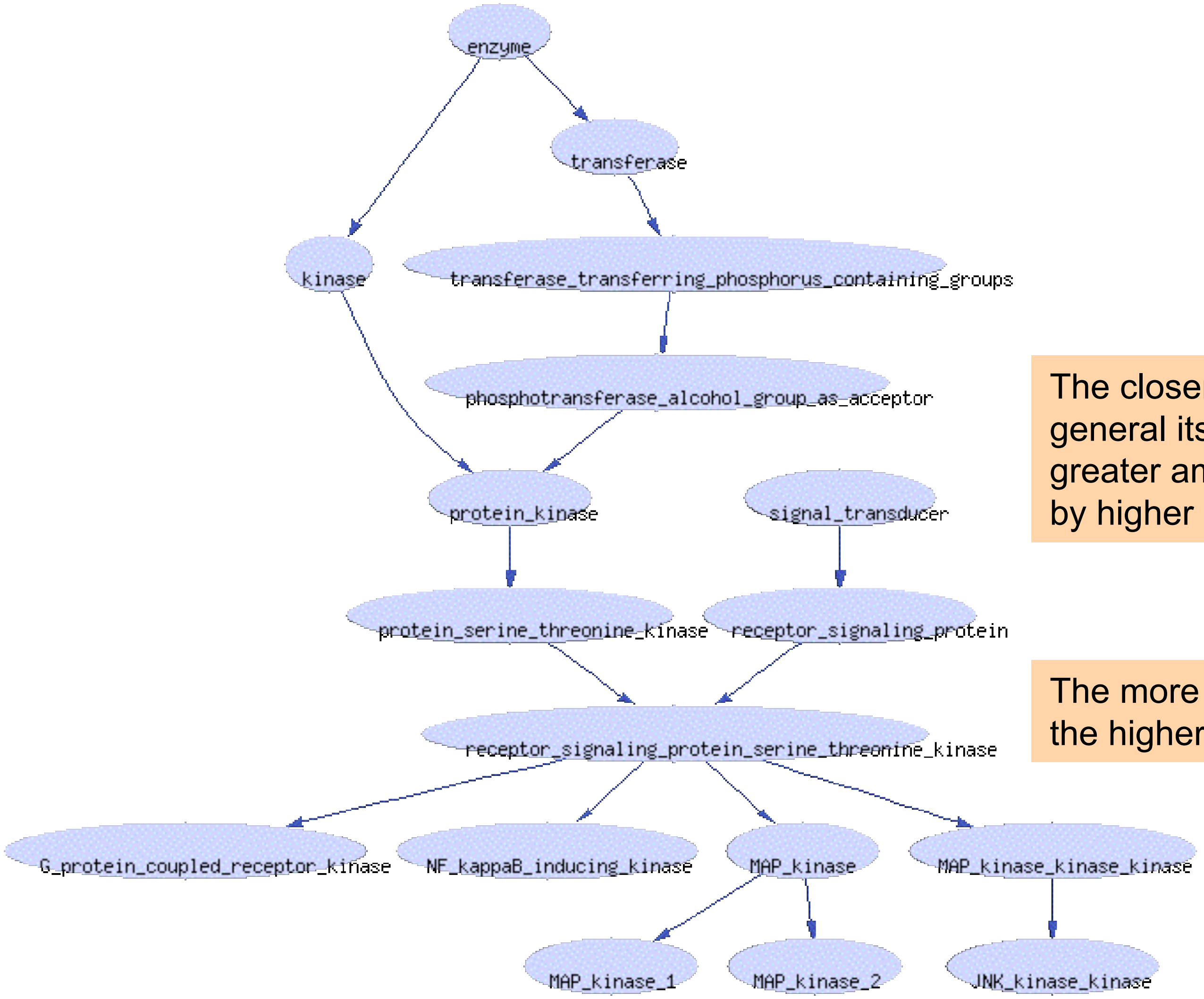




| |
|---|
| 104328_at |
| 103033_at |
| 103259_at |



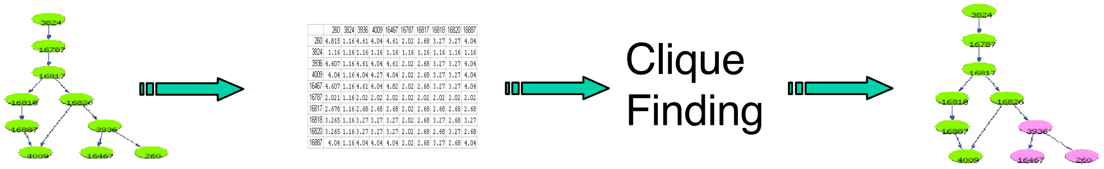
| |
|---|
| 103212_at |
| 104541_at |
| 102401_at |


enzyme
transferase
kinase
transferase_transferring_phosphorus_containing_groups
phosphotransferase_alcohol_group_as_acceptor
protein_kinase
signal_transducer
protein_serine_threonine_kinase
receptor_signaling_protein
receptor_signaling_protein_serine_threonine_kinase
G_protein_coupled_receptor_kinase
NF_kappaB_inducing_kinase
MAP_kinase
MAP_kinase_kinase_kinase
MAP_kinase_1
MAP_kinase_2
JNK_kinase_kinase

The closer a node is to the root, the more general its biological classification, thus a greater amount of information is conveyed by higher level edges

The more common parent nodes shared the higher the degree of similarity

# Pair-wise similarity score between GO terms



| | 260 | 3824 | 3936 | 4009 | 16467 | 16787 | 16817 | 16818 | 16820 | 16887 |
|---|---|---|---|---|---|---|---|---|---|---|
| 260 | 4.815 | 1.16 | 4.61 | 4.04 | 4.61 | 2.02 | 2.68 | 3.27 | 3.27 | 4.04 |
| 3824 | 1.16 | 1.16 | 1.16 | 1.16 | 1.16 | 1.16 | 1.16 | 1.16 | 1.16 | 1.16 |
| 3936 | 4.607 | 1.16 | 4.61 | 4.04 | 4.61 | 2.02 | 2.68 | 3.27 | 3.27 | 4.04 |
| 4009 | 4.04 | 1.16 | 4.04 | 4.27 | 4.04 | 2.02 | 2.68 | 3.27 | 3.27 | 4.04 |
| 16467 | 4.607 | 1.16 | 4.61 | 4.04 | 4.82 | 2.02 | 2.68 | 3.27 | 3.27 | 4.04 |
| 16787 | 2.021 | 1.16 | 2.02 | 2.02 | 2.02 | 2.02 | 2.02 | 2.02 | 2.02 | 2.02 |
| 16817 | 2.678 | 1.16 | 2.68 | 2.68 | 2.68 | 2.02 | 2.68 | 2.68 | 2.68 | 2.68 |
| 16818 | 3.265 | 1.16 | 3.27 | 3.27 | 3.27 | 2.02 | 2.68 | 3.27 | 2.68 | 3.27 |
| 16820 | 3.265 | 1.16 | 3.27 | 3.27 | 3.27 | 2.02 | 2.68 | 2.68 | 3.27 | 2.68 |
| 16887 | 4.04 | 1.16 | 4.04 | 4.04 | 4.04 | 2.02 | 2.68 | 3.27 | 2.68 | 4.04 |

Clique Finding

$$W_p = \sum_{n=0}^{p} (wt)^n, p > 0; W_0 = 0$$

A weighting factor (*wt*) was assigned to each edge as a function of the depth (*n*) in the digraph, I chose a value of 0.815 to maximize (*wt*6 – *wt*3).

Determining the longest partial path shared by two nodes, *Wp* is the sum of weights for edges from root to level *p.*

$$C = \sum_{n=0}^{\max-1} (wt)^n$$

A partial normalization scheme was applied to factor in the unevenness of the GO digraph.

$$Nf_p = \frac{W'_p}{W_p}$$

Calculate the average length for all paths that go through the shared partial path (*p*), followed by the weight for a hypothetical path with *p* edges (*Wp*).

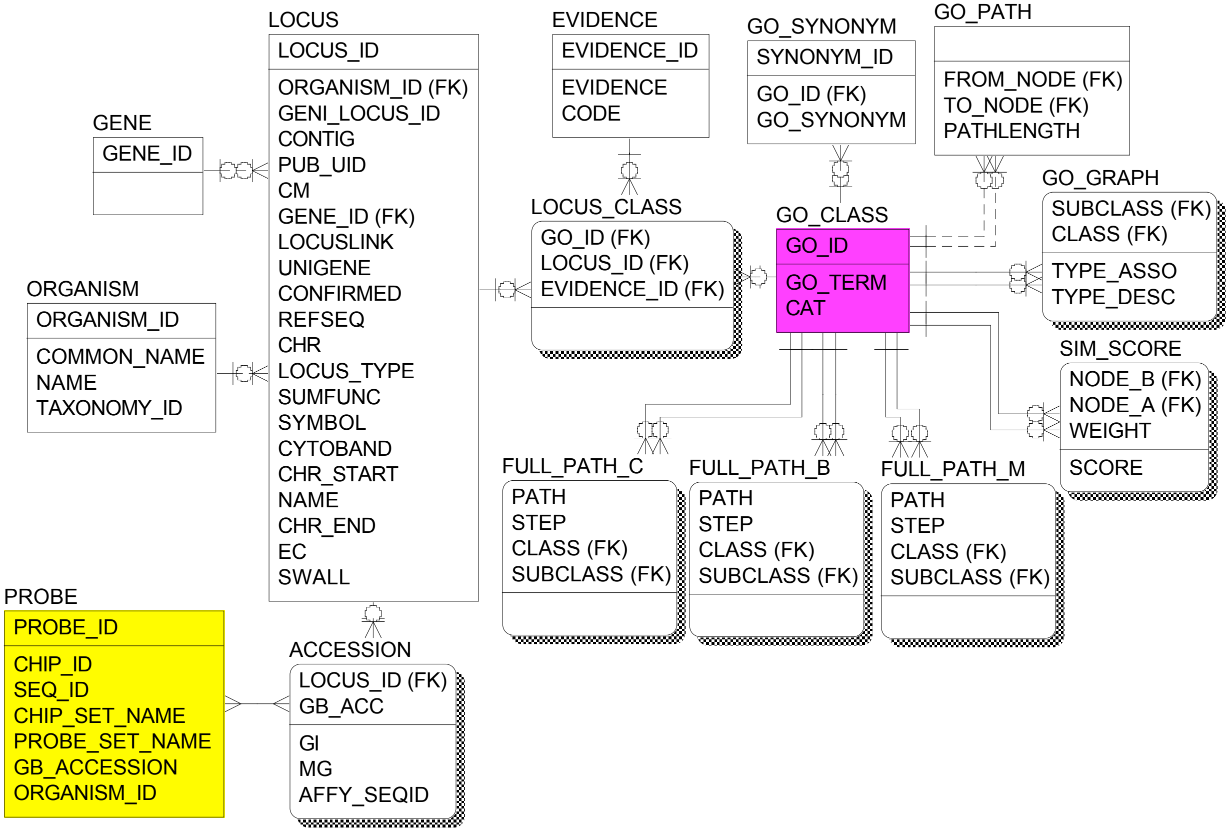
$$W_m = Nf_p \sum_{n=0}^{m} (wt)^n, m > 0$$

*Wp* is transformed to *W'p*, the mean of *Wp* and *C.*

The normalization factor (*Nfp*) is the ratio of *W'p* and *Wp*

The value for a partial path with m edges (*Wm*) is normalized by applying *Nfp*.

# Annotation database schema



LOCUS
LOCUS_ID
ORGANISM_ID (FK)
GENI_LOCUS_ID
CONTIG
PUB_UID
CM
GENE_ID (FK)
LOCUSLINK
UNIGENE
CONFIRMED
REFSEQ
CHR
LOCUS_TYPE
SUMFUNC
SYMBOL
CYTOBAND
CHR_START
NAME
CHR_END
EC
SWALL

GENE
GENE_ID

ORGANISM
ORGANISM_ID
COMMON_NAME
NAME
TAXONOMY_ID

PROBE
PROBE_ID
CHIP_ID
SEQ_ID
CHIP_SET_NAME
PROBE_SET_NAME
GB_ACCESSION
ORGANISM_ID

ACCESSION
LOCUS_ID (FK)
GB_ACC
GI
MG
AFFY_SEQID

EVIDENCE
EVIDENCE_ID
EVIDENCE
CODE

LOCUS_CLASS
GO_ID (FK)
LOCUS_ID (FK)
EVIDENCE_ID (FK)

GO_SYNONYM
SYNONYM_ID
GO_ID (FK)
GO_SYNONYM

GO_PATH
FROM_NODE (FK)
TO_NODE (FK)
PATHLENGTH

GO_CLASS
GO_ID
GO_TERM
CAT

GO_GRAPH
SUBCLASS (FK)
CLASS (FK)
TYPE_ASSO
TYPE_DESC

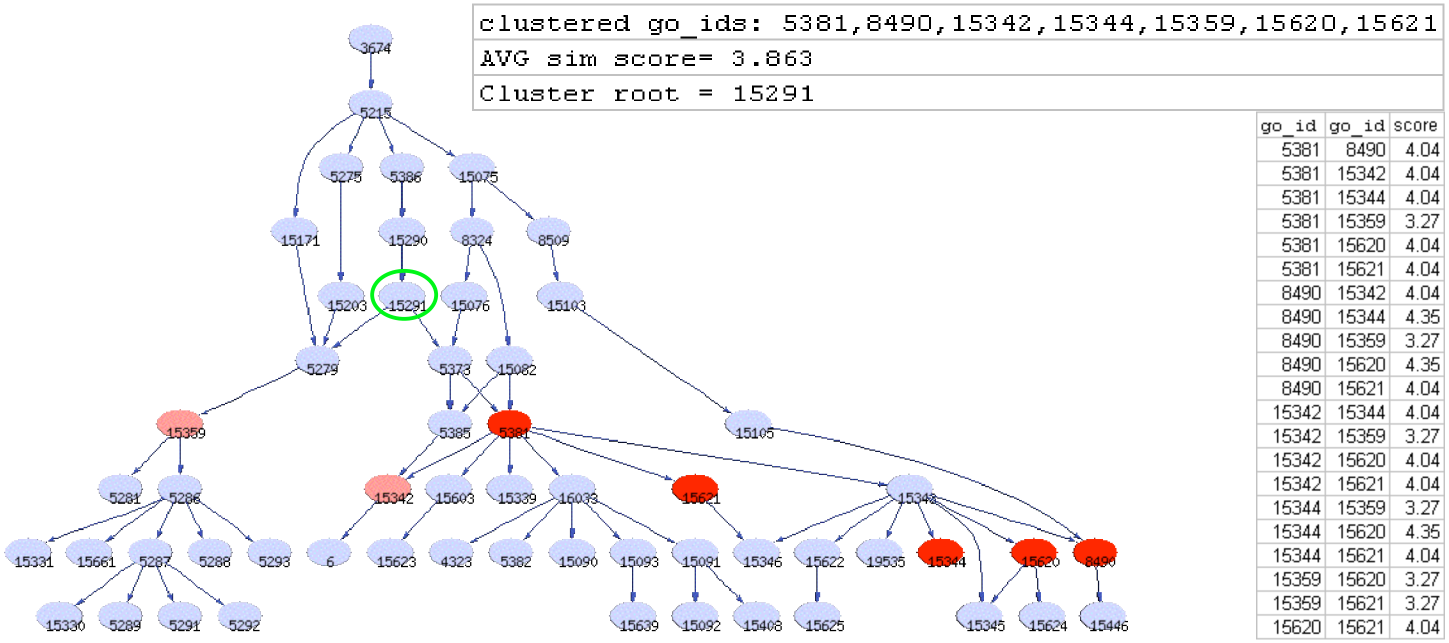
SIM_SCORE
NODE_B (FK)
NODE_A (FK)
WEIGHT
SCORE

FULL_PATH_C
PATH
STEP
CLASS (FK)
SUBCLASS (FK)

FULL_PATH_B
PATH
STEP
CLASS (FK)
SUBCLASS (FK)

FULL_PATH_M
PATH
STEP
CLASS (FK)
SUBCLASS (FK)

# Spike-in experiment



```
clustered go_ids: 5381,8490,15342,15344,15359,15620,15621
AVG sim score= 3.863
Cluster root = 15291
```

| go_id | go_id | score |
|---|---|---|
| 5381 | 8490 | 4.04 |
| 5381 | 15342 | 4.04 |
| 5381 | 15344 | 4.04 |
| 5381 | 15359 | 3.27 |
| 5381 | 15620 | 4.04 |
| 5381 | 15621 | 4.04 |
| 8490 | 15342 | 4.04 |
| 8490 | 15344 | 4.35 |
| 8490 | 15359 | 3.27 |
| 8490 | 15620 | 4.35 |
| 8490 | 15621 | 4.04 |
| 15342 | 15344 | 4.04 |
| 15342 | 15359 | 3.27 |
| 15342 | 15620 | 4.04 |
| 15342 | 15621 | 4.04 |
| 15344 | 15359 | 3.27 |
| 15344 | 15620 | 4.35 |
| 15344 | 15621 | 4.04 |
| 15359 | 15620 | 3.27 |
| 15359 | 15621 | 3.27 |
| 15620 | 15621 | 4.04 |

Five related GO nodes with GOids 5381, 8490, 15344, 15620, and 15621; labeled **red**; were spiked into a randomly selected pool of 20 nodes and subjected to GO clustering. The similarity analysis successfully re-created the set of related GO nodes. Column 1and 2 in the table shows a pair of GO nodes and column 3 shows the pair-wise similarity scores. Nodes colored **pink** (15342, 15359) are from the randomly selected 20 Go nodes and were clustered with the spiked GO nodes. **Green circle** indicates the cluster root (15291), which is the lowest level common ancestor node.

# RA stimulated MPRO cell differentiation time-series experiment

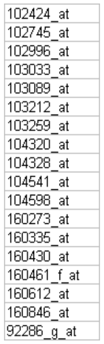
Transgenic Myeloid Progenitor (MPRO) cells transgenic for the dominant negative Retinoic Acid (RA) receptor were induced to differentiate into Neutrophils with high doses of RA.

Gene expression at 0, 1, 2, 4, and 8 hours post RA induction was analyzed with Affymetrix U74Av2 mouse microarray.

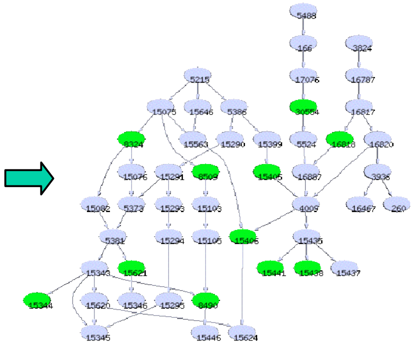
Genes showing significant changes in their expression level across a series of time points are modulated by retinoic acid stimulation and cell differentiation.

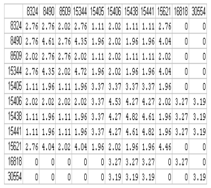
We arbitrarily took the top 80 genes based on the F-score ranking.

# GO clustering



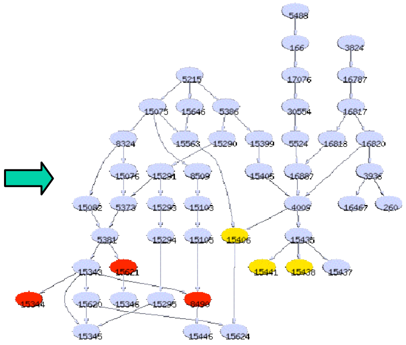
| |
|---|
| 102424_at |
| 102745_at |
| 102996_at |
| 103033_at |
| 103089_at |
| 103212_at |
| 103259_at |
| 104320_at |
| 104328_at |
| 104541_at |
| 104598_at |
| 160273_at |
| 160335_at |
| 160430_at |
| 160461_f_at |
| 160612_at |
| 160846_at |
| 92286_g_at |

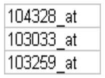




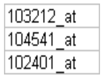
| | 8324 | 8490 | 8509 | 15344 | 15405 | 15406 | 15438 | 15441 | 15621 | 16818 | 30554 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 8324 | 2.76 | 2.76 | 2.02 | 2.76 | 1.11 | 2.02 | 1.11 | 1.11 | 2.76 | 0 | 0 |
| 8490 | 2.76 | 4.61 | 2.76 | 4.35 | 1.96 | 2.02 | 1.96 | 1.96 | 4.04 | 0 | 0 |
| 8509 | 2.02 | 2.76 | 2.76 | 2.02 | 1.11 | 2.02 | 1.11 | 1.11 | 2.02 | 0 | 0 |
| 15344 | 2.76 | 4.35 | 2.02 | 4.72 | 1.96 | 2.02 | 1.96 | 1.96 | 4.04 | 0 | 0 |
| 15405 | 1.11 | 1.96 | 1.11 | 1.96 | 3.37 | 3.37 | 3.37 | 3.37 | 1.96 | 0 | 0 |
| 15406 | 2.02 | 2.02 | 2.02 | 2.02 | 3.37 | 4.53 | 4.27 | 4.27 | 2.02 | 3.27 | 3.19 |
| 15438 | 1.11 | 1.96 | 1.11 | 1.96 | 3.37 | 4.27 | 4.82 | 4.61 | 1.96 | 3.27 | 3.19 |
| 15441 | 1.11 | 1.96 | 1.11 | 1.96 | 3.37 | 4.27 | 4.61 | 4.82 | 1.96 | 3.27 | 3.19 |
| 15621 | 2.76 | 4.04 | 2.02 | 4.04 | 1.96 | 2.02 | 1.96 | 1.96 | 4.46 | 0 | 0 |
| 16818 | 0 | 0 | 0 | 0 | 0 | 3.27 | 3.27 | 3.27 | 0 | 3.27 | 0 |
| 30554 | 0 | 0 | 0 | 0 | 0 | 3.19 | 3.19 | 3.19 | 0 | 0 | 3.19 |

Clique Finding





| |
|---|
| 104328_at |
| 103033_at |
| 103259_at |

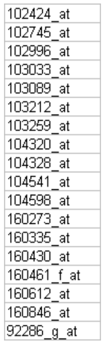


| |
|---|
| 103212_at |
| 104541_at |
| 102401_at |

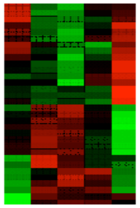
# GO clustering on Leukocyte differentiation time-series experiment

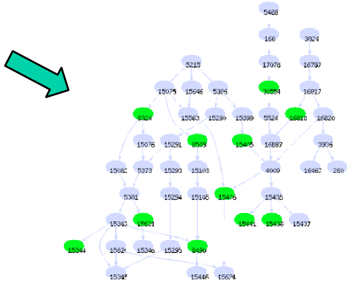
| Rank | Title | Score | probe sets | Genes | X1 | n1 | X2 | n2 | Bootstrap p-val (10000) btstrap | Enrichment |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | defense response | 3.403 | 93454_at<br>102424_at<br>92286_g_at<br>102401_at<br>103033_at<br>102745_at | lymphocyte antigen 68<br>small inducible cytokine A3<br>interleukin 4<br>interferon regulatory factor 1<br>complement component 4<br>T-cell receptor gamma | 6 | 29 | 221 | 3163 | 0.0148 | Yes |
| 2 | transcription regulation | 4.04 | 102996_at<br>103259_at<br>102401_at<br>100156_at<br>92644_s_at<br>94325_at<br>93264_at<br>97994_at<br>101502_at | eleven-nineteen lysine-rich leukemia gene<br>growth factor independent 1<br>interferon regulatory factor 1<br>mini chromosome maintenance deficient 5<br>myeloblastosis oncogene<br>pre B-cell leukemia transcription factor 1<br>sterol regulatory element binding factor 1<br>transcription factor 7, T-cell specific<br>TG interacting factor | 9 | 29 | 486 | 3163 | 0.0287 | Yes |
| 3 | steroid metabolism, steroid biosynthesis | 3.862 | 93264_at<br>94325_at | sterol regulatory element binding factor 1<br>pre B-cell leukemia transcription factor 1 | 2 | 29 | 38 | 3163 | 0.0478 | Yes |
| 4 | cell cycle control | 3.05 | 99956_at<br>92644_s_at | kit oncogene,<br>myeloblastosis oncogene | 2 | 29 | 113 | 3163 | 0.2757 | No |
| 5 | cytoskeleton organization and biogenesis | 3.592 | 103212_at<br>96747_at | dynein, axon, heavy chain 11, Wnt1<br>responsive Cdc42 homolog | 2 | 29 | 119 | 3163 | 0.2972 | No |
| 6 | protein modification | 3.265 | 93274_at<br>99956_at<br>104598_at | CDC-like kinase,<br>kit oncogene,<br>PTP non-receptor type 16 | 3 | 29 | 282 | 3163 | 0.4871 | No |

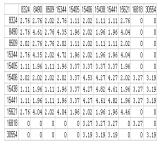
# GO-guided expression clustering



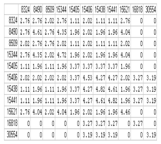
| |
|---|
| 102424_at |
| 102745_at |
| 102996_at |
| 103033_at |
| 103089_at |
| 103212_at |
| 103259_at |
| 104320_at |
| 104328_at |
| 104541_at |
| 104598_at |
| 160273_at |
| 160335_at |
| 160430_at |
| 160461_f_at |
| 160612_at |
| 160846_at |
| 92286_g_at |





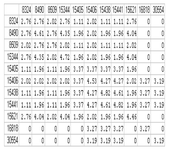


| | 8324 | 8490 | 8509 | 15344 | 15405 | 15406 | 15438 | 15441 | 15621 | 16818 | 30554 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 8324 | 2.76 | 2.76 | 2.02 | 2.76 | 1.11 | 2.02 | 1.11 | 1.11 | 2.76 | 0 | 0 |
| 8490 | 2.76 | 4.61 | 2.76 | 4.35 | 1.96 | 2.02 | 1.96 | 1.96 | 4.04 | 0 | 0 |
| 8509 | 2.02 | 2.76 | 2.76 | 2.02 | 1.11 | 2.02 | 1.11 | 1.11 | 2.02 | 0 | 0 |
| 15344 | 2.76 | 4.35 | 2.02 | 4.72 | 1.96 | 2.02 | 1.96 | 1.96 | 4.04 | 0 | 0 |
| 15405 | 1.11 | 1.96 | 1.11 | 1.96 | 3.37 | 3.37 | 3.37 | 3.37 | 1.96 | 0 | 0 |
| 15406 | 2.02 | 2.02 | 2.02 | 2.02 | 3.37 | 4.53 | 4.27 | 4.27 | 2.02 | 3.27 | 3.19 |
| 15438 | 1.11 | 1.96 | 1.11 | 1.96 | 3.37 | 4.27 | 4.82 | 4.61 | 1.96 | 3.27 | 3.19 |
| 15441 | 1.11 | 1.96 | 1.11 | 1.96 | 3.37 | 4.27 | 4.61 | 4.82 | 1.96 | 3.27 | 3.19 |
| 15621 | 2.76 | 4.04 | 2.02 | 4.04 | 1.96 | 2.02 | 1.96 | 1.96 | 4.46 | 0 | 0 |
| 16818 | 0 | 0 | 0 | 0 | 0 | 3.27 | 3.27 | 3.27 | 0 | 3.27 | 0 |
| 30554 | 0 | 0 | 0 | 0 | 0 | 3.19 | 3.19 | 3.19 | 0 | 0 | 3.19 |



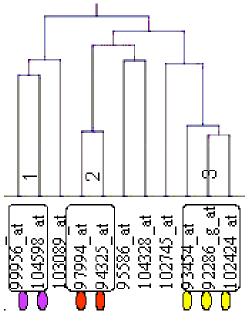
| | 8324 | 8490 | 8509 | 15344 | 15405 | 15406 | 15438 | 15441 | 15621 | 16818 | 30554 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 8324 | 2.76 | 2.76 | 2.02 | 2.76 | 1.11 | 2.02 | 1.11 | 1.11 | 2.76 | 0 | 0 |
| 8490 | 2.76 | 4.61 | 2.76 | 4.35 | 1.96 | 2.02 | 1.96 | 1.96 | 4.04 | 0 | 0 |
| 8509 | 2.02 | 2.76 | 2.76 | 2.02 | 1.11 | 2.02 | 1.11 | 1.11 | 2.02 | 0 | 0 |
| 15344 | 2.76 | 4.35 | 2.02 | 4.72 | 1.96 | 2.02 | 1.96 | 1.96 | 4.04 | 0 | 0 |
| 15405 | 1.11 | 1.96 | 1.11 | 1.96 | 3.37 | 3.37 | 3.37 | 3.37 | 1.96 | 0 | 0 |
| 15406 | 2.02 | 2.02 | 2.02 | 2.02 | 3.37 | 4.53 | 4.27 | 4.27 | 2.02 | 3.27 | 3.19 |
| 15438 | 1.11 | 1.96 | 1.11 | 1.96 | 3.37 | 4.27 | 4.82 | 4.61 | 1.96 | 3.27 | 3.19 |
| 15441 | 1.11 | 1.96 | 1.11 | 1.96 | 3.37 | 4.27 | 4.61 | 4.82 | 1.96 | 3.27 | 3.19 |
| 15621 | 2.76 | 4.04 | 2.02 | 4.04 | 1.96 | 2.02 | 1.96 | 1.96 | 4.46 | 0 | 0 |
| 16818 | 0 | 0 | 0 | 0 | 0 | 3.27 | 3.27 | 3.27 | 0 | 3.27 | 0 |
| 30554 | 0 | 0 | 0 | 0 | 0 | 3.19 | 3.19 | 3.19 | 0 | 0 | 3.19 |

Linear combination



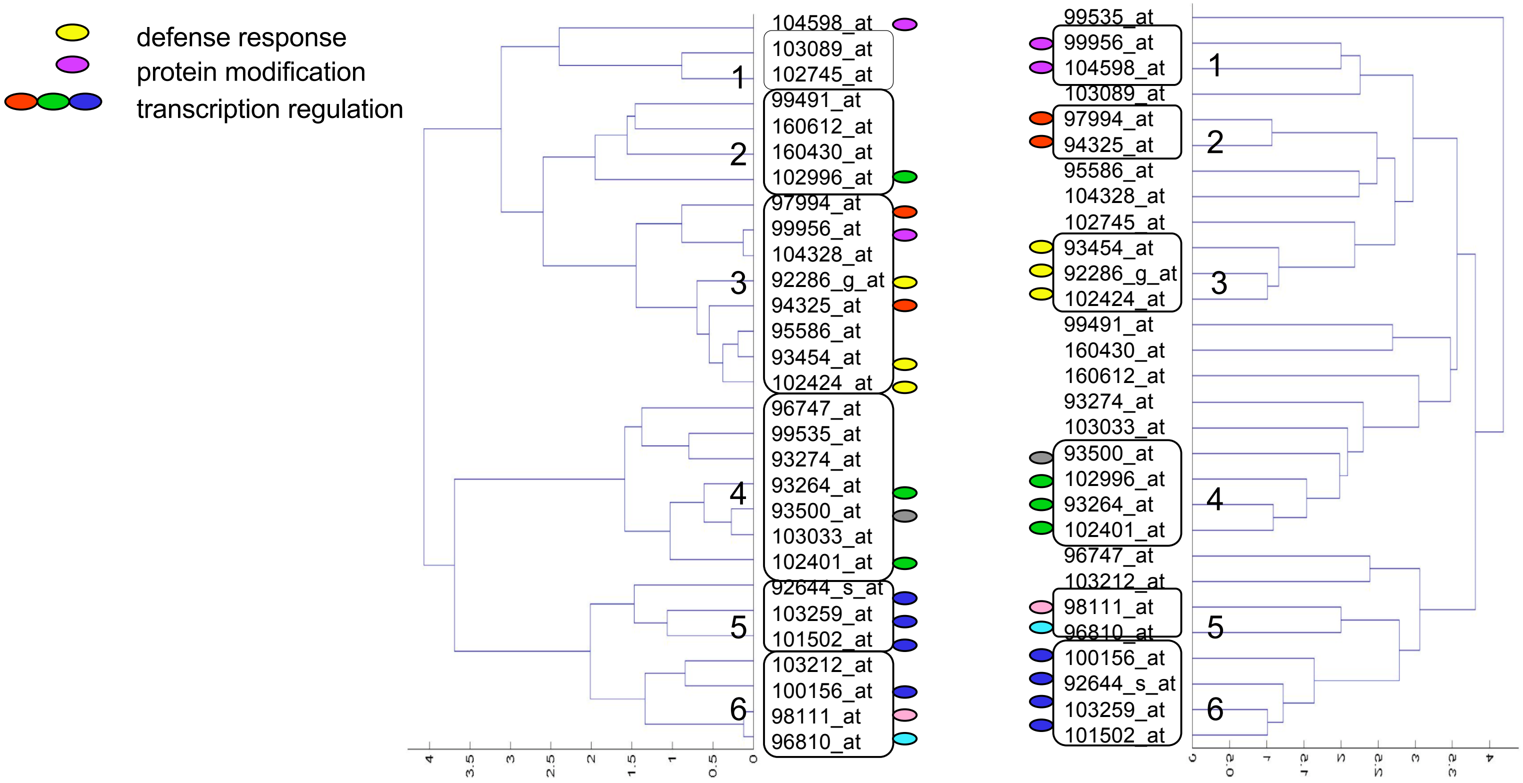
| | 8324 | 8490 | 8509 | 15344 | 15405 | 15406 | 15438 | 15441 | 15621 | 16818 | 30554 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 8324 | 2.76 | 2.76 | 2.02 | 2.76 | 1.11 | 2.02 | 1.11 | 1.11 | 2.76 | 0 | 0 |
| 8490 | 2.76 | 4.61 | 2.76 | 4.35 | 1.96 | 2.02 | 1.96 | 1.96 | 4.04 | 0 | 0 |
| 8509 | 2.02 | 2.76 | 2.76 | 2.02 | 1.11 | 2.02 | 1.11 | 1.11 | 2.02 | 0 | 0 |
| 15344 | 2.76 | 4.35 | 2.02 | 4.72 | 1.96 | 2.02 | 1.96 | 1.96 | 4.04 | 0 | 0 |
| 15405 | 1.11 | 1.96 | 1.11 | 1.96 | 3.37 | 3.37 | 3.37 | 3.37 | 1.96 | 0 | 0 |
| 15406 | 2.02 | 2.02 | 2.02 | 2.02 | 3.37 | 4.53 | 4.27 | 4.27 | 2.02 | 3.27 | 3.19 |
| 15438 | 1.11 | 1.96 | 1.11 | 1.96 | 3.37 | 4.27 | 4.82 | 4.61 | 1.96 | 3.27 | 3.19 |
| 15441 | 1.11 | 1.96 | 1.11 | 1.96 | 3.37 | 4.27 | 4.61 | 4.82 | 1.96 | 3.27 | 3.19 |
| 15621 | 2.76 | 4.04 | 2.02 | 4.04 | 1.96 | 2.02 | 1.96 | 1.96 | 4.46 | 0 | 0 |
| 16818 | 0 | 0 | 0 | 0 | 0 | 3.27 | 3.27 | 3.27 | 0 | 3.27 | 0 |
| 30554 | 0 | 0 | 0 | 0 | 0 | 3.19 | 3.19 | 3.19 | 0 | 0 | 3.19 |

Hierarchical Clustering



# GO guided clustering on Leukocyte differentiation time-series experiment

Gene clusters where correlations between biological function and expression profile are both evident were identified by GO guided clustering.



Hierarchical clustering

GO-guided hierarchical clustering

# Acknowledgements

John Martin

Melissa Cline

David Finkelstein

Tarif Awad

Michael Stewart

Michael Siani-Rose

David Kulp

**Thank you!**