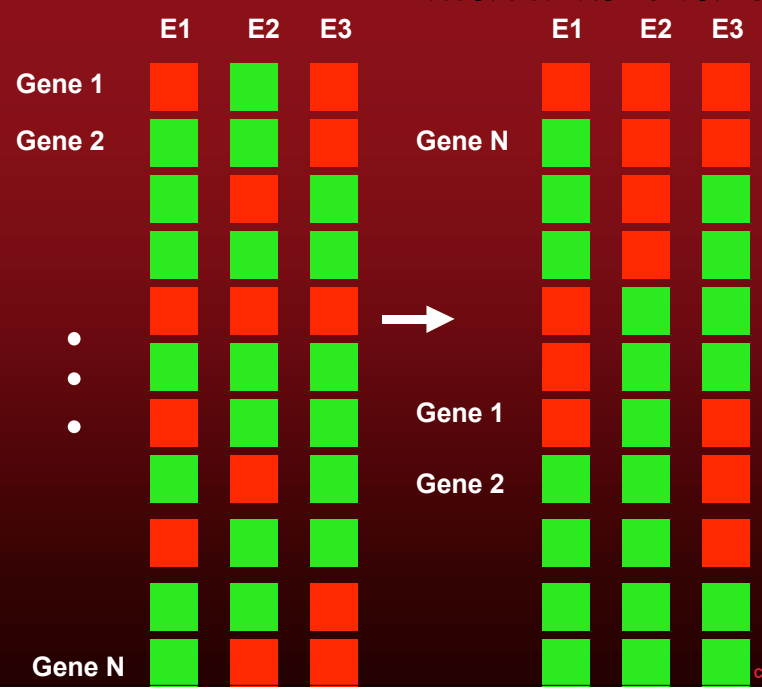
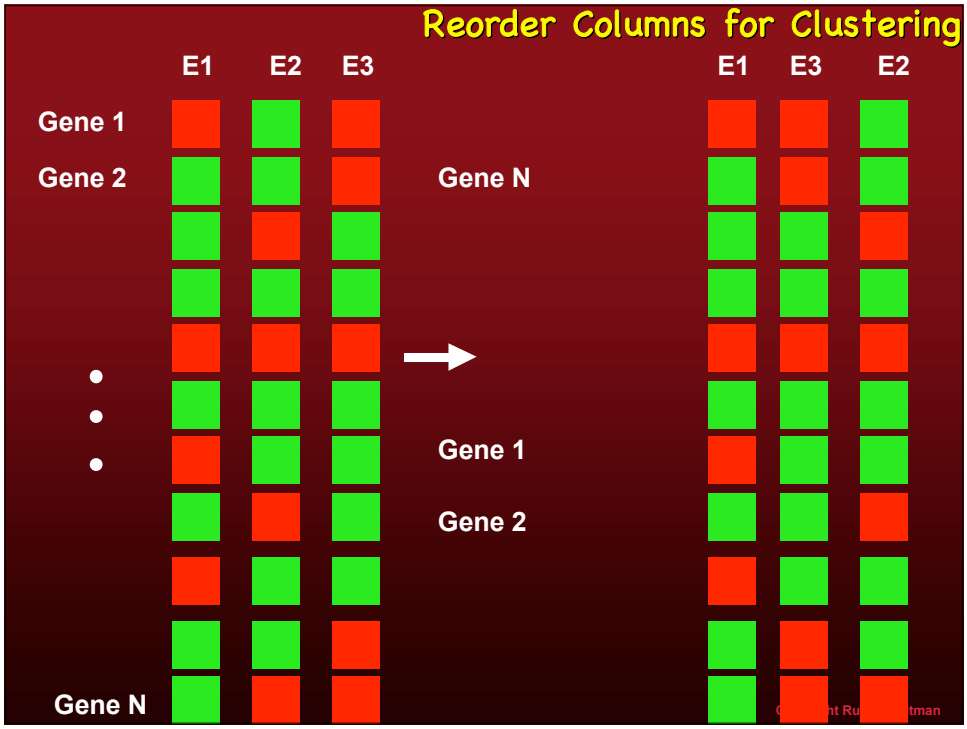
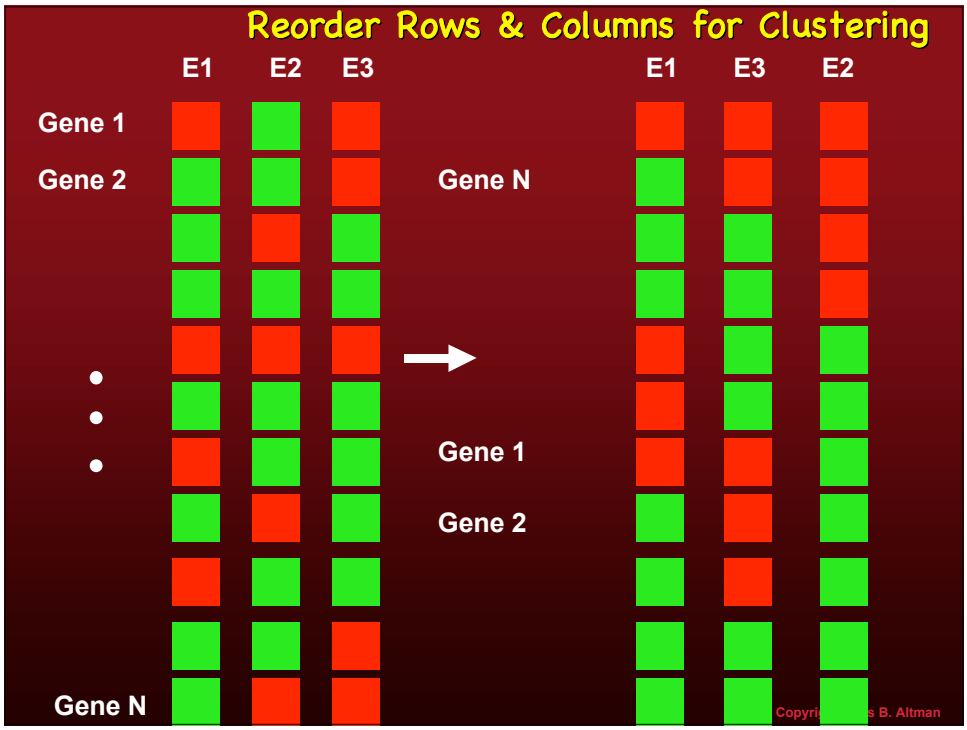
Microarray data analysis II

Russ B. Altman
BMI 214
CS 274

Reorder Rows for Clustering




Reorder Columns for Clustering
E1
E2
E3
Gene 1
Gene 2
Gene N
E1
E3
E2
Gene N
Gene 1
Gene 2


Reorder Rows & Columns for Clustering
E1
E2
E3
Gene 1
Gene 2
Gene N
E1
E3
E2
Gene N
Gene 1
Gene 2

Topics today

1. Analyzing clusters for significance
   - Gene Ontology Significance
   - Literature Coherence
2. Classification Algorithms
3. Looking for other signals in expression data.

What do we do with clusters?

We have a cluster of genes.

Besides expression behavior, do they share any known similarities?

Wouldn't it be great if there was a way to tell if these genes are known to perform similar functions?

There is.

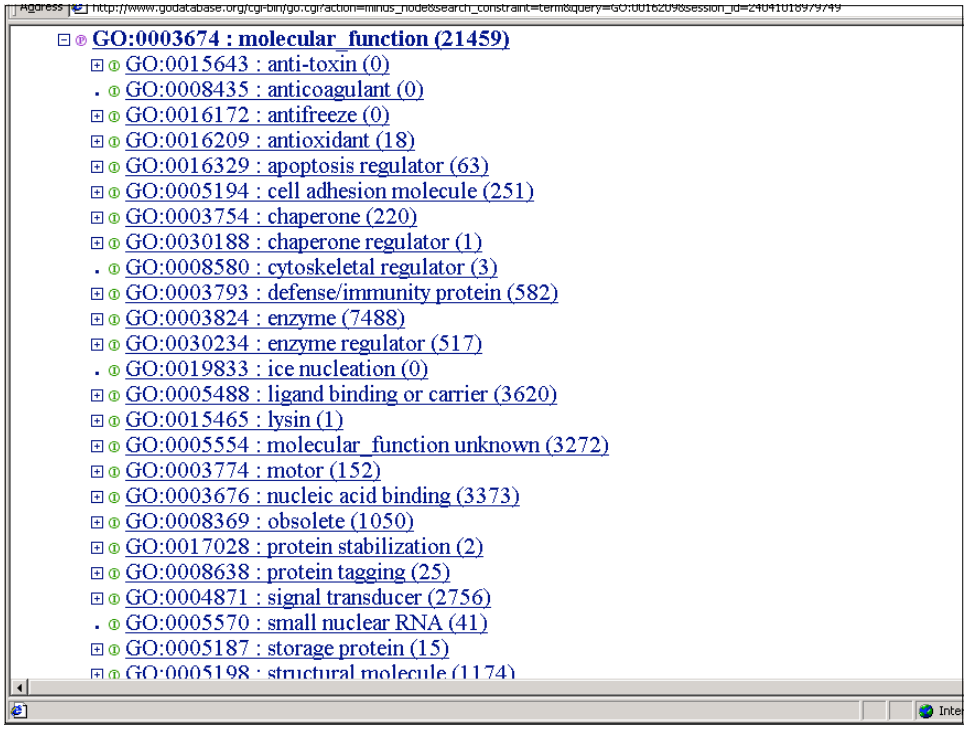
Gene Ontology
(http://www.geneontology.org/)

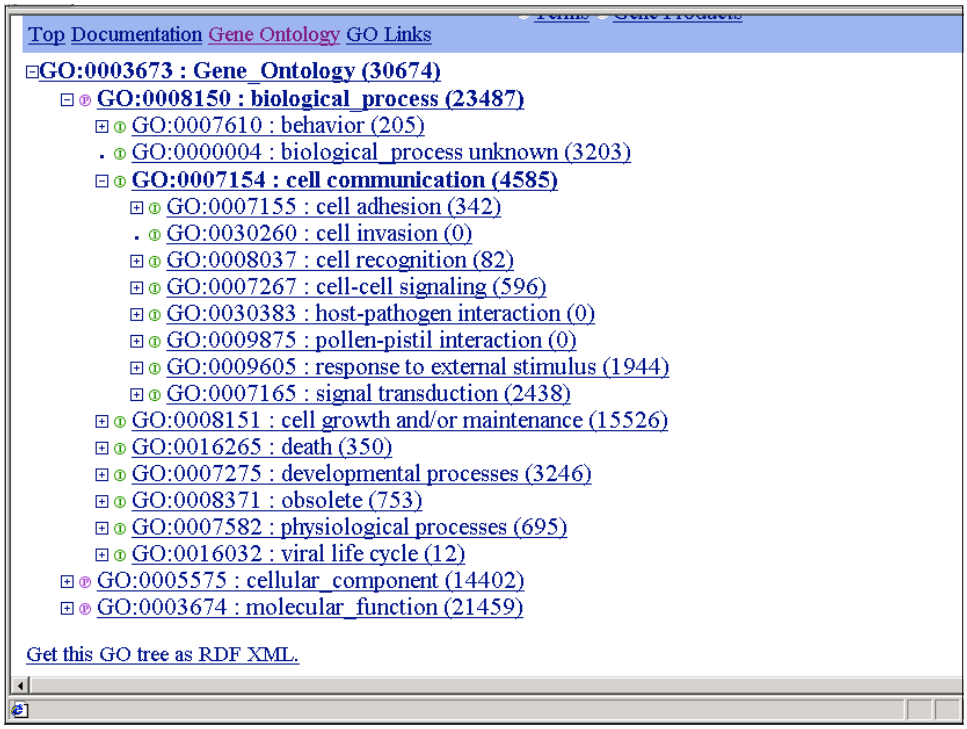
Used to classify function in human genome draft.

A controlled listing of three types of function:

- Molecular Function
- Biological Process
- Cellular Component

Represented as a Directed Acyclic Graph
See http://www.geneontology.org/


GO:0003674 : molecular_function (21459)
GO:0015643 : anti-toxin (0)
GO:0008435 : anticoagulant (0)
GO:0016172 : antifreeze (0)
GO:0016209 : antioxidant (18)
GO:0016329 : apoptosis regulator (63)
GO:0005194 : cell adhesion molecule (251)
GO:0003754 : chaperone (220)
GO:0030188 : chaperone regulator (1)
GO:0008580 : cytoskeletal regulator (3)
GO:0003793 : defense/immunity protein (582)
GO:0003824 : enzyme (7488)
GO:0030234 : enzyme regulator (517)
GO:0019833 : ice nucleation (0)
GO:0005488 : ligand binding or carrier (3620)
GO:0015465 : lysin (1)
GO:0005554 : molecular_function unknown (3272)
GO:0003774 : motor (152)
GO:0003676 : nucleic acid binding (3373)
GO:0008369 : obsolete (1050)
GO:0017028 : protein stabilization (2)
GO:0008638 : protein tagging (25)
GO:0004871 : signal transducer (2756)
GO:0005570 : small nuclear RNA (41)
GO:0005187 : storage protein (15)


Top Documentation Gene Ontology GO Links
GO:0003673 : Gene_Ontology (30674)
GO:0008150 : biological_process (23487)
GO:0007610 : behavior (205)
GO:0000004 : biological_process unknown (3203)
GO:0007154 : cell communication (4585)
GO:0007155 : cell adhesion (342)
GO:0030260 : cell invasion (0)
GO:0008037 : cell recognition (82)
GO:0007267 : cell-cell signaling (596)
GO:0030383 : host-pathogen interaction (0)
GO:0009875 : pollen-pistil interaction (0)
GO:0009605 : response to external stimulus (1944)
GO:0007165 : signal transduction (2438)
GO:0008151 : cell growth and/or maintenance (15526)
GO:0016265 : death (350)
GO:0007275 : developmental processes (3246)
GO:0008371 : obsolete (753)
GO:0007582 : physiological processes (695)
GO:0016032 : viral life cycle (12)
GO:0005575 : cellular_component (14402)
GO:0003674 : molecular_function (21459)
Get this GO tree as RDF XML.

Terms Gene Products

Top Documentation Gene Ontology GO Links

- **GO:0003673 : Gene_Ontology (30674)**
  - GO:0008150 : biological_process (23487)
  - **GO:0005575 : cellular_component (14402)**
    - **GO:0005623 : cell (11523)**
      - GO:0005627 : ascus (4)
      - GO:0030424 : axon (0)
      - GO:0005933 : bud (53)
      - GO:0000267 : cell fraction (819)
      - GO:0030425 : dendrite (2)
      - GO:0019861 : flagellum (26)
      - GO:0005622 : intracellular (10239)
      - GO:0016020 : membrane (3909)
      - GO:0030496 : midbody (0)
      - GO:0030428 : septum (0)
      - GO:0005936 : shmoo (12)
      - GO:0030427 : site of polarized growth (43)
    - GO:0008372 : cellular_component unknown (1726)
    - GO:0030312 : external protective structure (55)
    - GO:0005576 : extracellular (1102)
    - GO:0008370 : obsolete (147)
    - GO:0005941 : unlocalized (147)
  - GO:0003674 : molecular_function (21459)

ubiquitin-specific protease activity

Accession: GO:0004843
Aspect: molecular_function
Synonyms:
UBP
UCH2
Definition:
Catalysis of the hydrolysis of various forms of polymeric ubiquitin sequences. Will remove ubiquitin from larger leaving groups.

Term Lineage

Graphical View

all : all (183091)
GO:0003674 : molecular_function (116868)
GO:0003824 : catalytic activity (37487)
GO:0016787 : hydrolase activity (12113)
GO:0008233 : peptidase activity (3062)
GO:0008234 : cysteine-type peptidase activity (593)
GO:0004843 : ubiquitin-specific protease activity (185)

External References

None.

http://www.geneontology.org/

Filter Associations

Datasource	Evidence Code	Species
All	All Curator Approved	All
FlyBase	IMP	A. aeolicus
SGD	IGI	A. fulgidus

Submit Query

Gene Symbol	Datasource	Evidence	Full Name
CG12082 ATGCC / GOst	FlyBase	ISS	None
CG14619 ATGCC / GOst	FlyBase	TAS	None
CG1490 ATGCC / GOst	FlyBase	ISS	None
CG15817 ATGCC / GOst	FlyBase	ISS	None
CG3016 ATGCC / GOst	FlyBase	ISS	None
CG30421 ATGCC / GOst	FlyBase	ISS	None
CG32479 ATGCC / GOst	FlyBase	ISS	None
CG4165 ATGCC / GOst	FlyBase	ISS	None

	Biological Process		Molecular Function		Cellular Component		Total Gene Products Associated	Total References Included as Evidence	TAB Delimited File of Associations & Last Update
	All codes	non-IEA codes	All codes	non-IEA codes	All codes	non-IEA codes			
SGD *Saccharomyces cerevisiae* README	6454	6454	6437	6437	6437	6437	6454	5122	Download Apr 5, 2005
FlyBase *Drosophila melanogaster* README	9143	5835	9277	7696	6447	5106	10374	7022	Download Mar 24, 2005
MGI *Mus musculus*	12773	8377	13701	8781	12998	9773	16219	5020	Download Apr 1, 2005
TAIR *Arabidopsis thaliana* README	11650	11647	6234	6234	20835	10319	24298	2677	Download Apr 5, 2005
WormBase *Caenorhabditis elegans* README	9289	4200	9298	643	4980	597	11812	755	Download Apr 5, 2005
RGD *Rattus norvegicus*	5686	3406	5933	4055	5293	2489	6542	3645	Download Feb 23, 2005
Gramene *Oryza sativa* README	15944	8173	13746	2411	34271	31404	38273	2381	Download Mar 23, 2005
ZFIN *Danio rerio* README	7670	3922	8051	3558	7267	4352	8423	489	Download Apr 5, 2005
DictyBase *Dictyostelium discoideum*	3978	1593	4717	1597	2863	1497	5384	326	Download Apr 4, 2005

Annotation of Human Genome

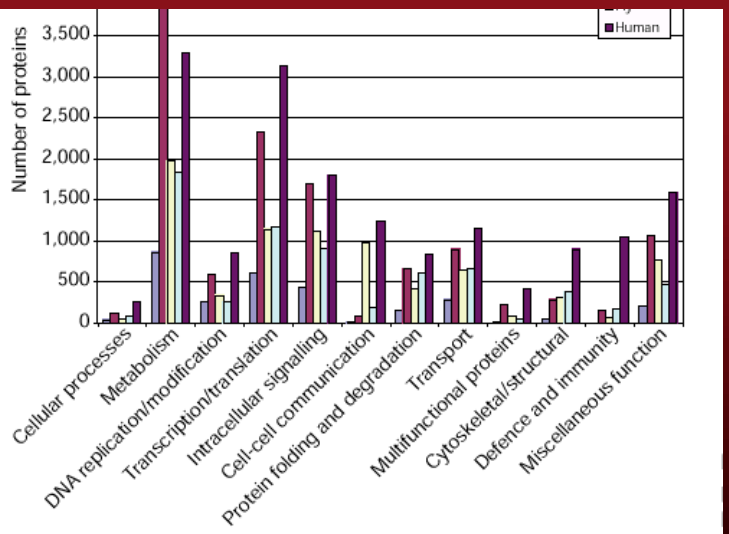
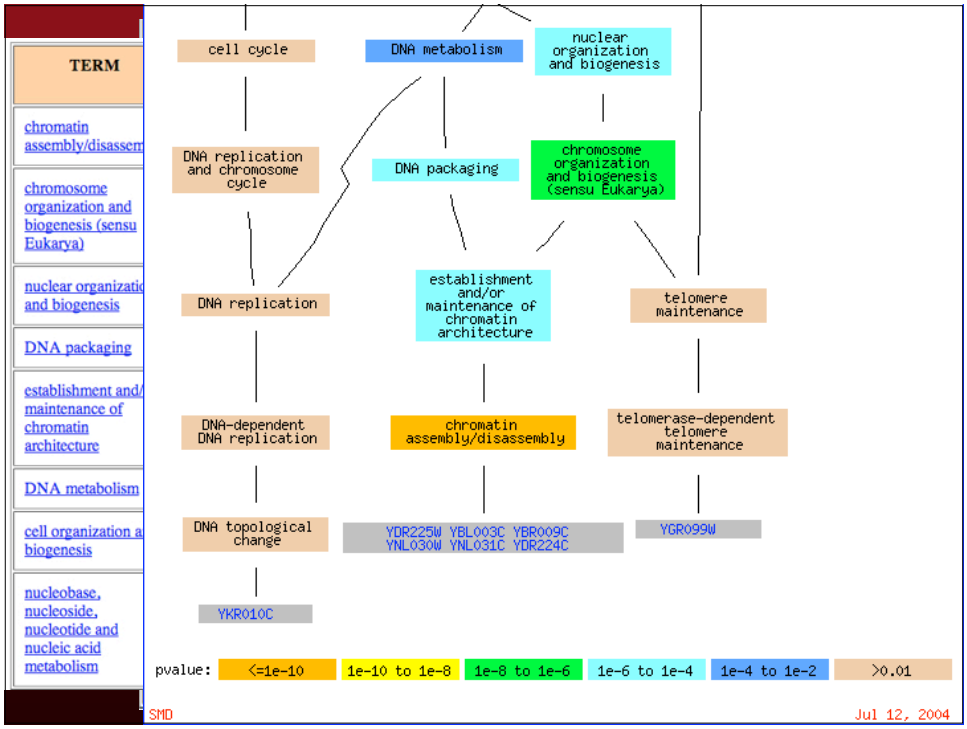


Figure 37 Functional categories in eukaryotic proteomes. The classification categories were derived from functional classification systems, including the top-level biological function category of the Gene Ontology project (GO; see http://www.geneontology.org).

Assessing clusters for presence of GO clusters

(GOOGLE: GO::Termfinder)

1. Grab a cluster of genes
2. Enter into website text box
3. Find out which GO terms are over-represented in the gene cluster.
4. Use this to focus in on likely/possible function of cluster.



How does it work?

(from GO:Termfinder help pages)

Model expected distribution of GO terms using a hypergeometric distribution
(http://mathworld.wolfram.com/HypergeometricDistribution.html)

Given: Population of N genes
Subset of M have a particular GO annotation
We sample n genes
We observe x genes with that annotation

The probability of seeing those x annotated genes is:

$$p = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

How does it work?

Model expected distribution of GO terms using a hypergeometric distribution

(rgeometricDistribution.html)

Number of ways that x annotated genes are selected from M possible annotated genes

nes

a pa

We sample n genes

We observe x genes wi

Number of ways that n-m genes without the GO code can be selected from N-M total genes without the GO code

The probability of seeing those x annotated genes is:

$$p = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Number of ways n genes can be sampled from a total of N genes

r)!

How to estimate significance of this probability?

What is probability of seeing x *or more* of an annotation out of n samples, given that M of N have that annotation?

$$p\_value = \sum_{j=x}^{n} \frac{\binom{M}{j}\binom{N-M}{n-j}}{\binom{N}{n}}$$

Probability that x or more have the annotation.

OR

$$p\_value = 1 - \sum_{j=0}^{x-1} \frac{\binom{M}{j}\binom{N-M}{n-j}}{\binom{N}{n}}$$

Probability that at least x have the annotation.

But we are not asking about the occurrence of a specific GO code, but about all possible GO codes

This raises an important issue of "multiple hypothesis testing."

When testing a hypothesis, we often look for a probability of being wrong < 0.05, then 1/20 will be false positives.

If we test 20 hypotheses, then 1 of them is likely to be wrong by chance, so we need to "correct" for the large number of tests.

How to correct for multiple hypothesis testing?

Bonferroni says "Divide the p-value by the number of hypotheses tested."

This is VERY conservative, but if something is still significant, it is likely to be true.

There are many other methods for correction (e.g. False Discovery Rate), not discussed here.

Bonferroni assumes that all GO hypotheses are independent which is not true, because GO terms are arranged in a tree, and some are more closely related than others.

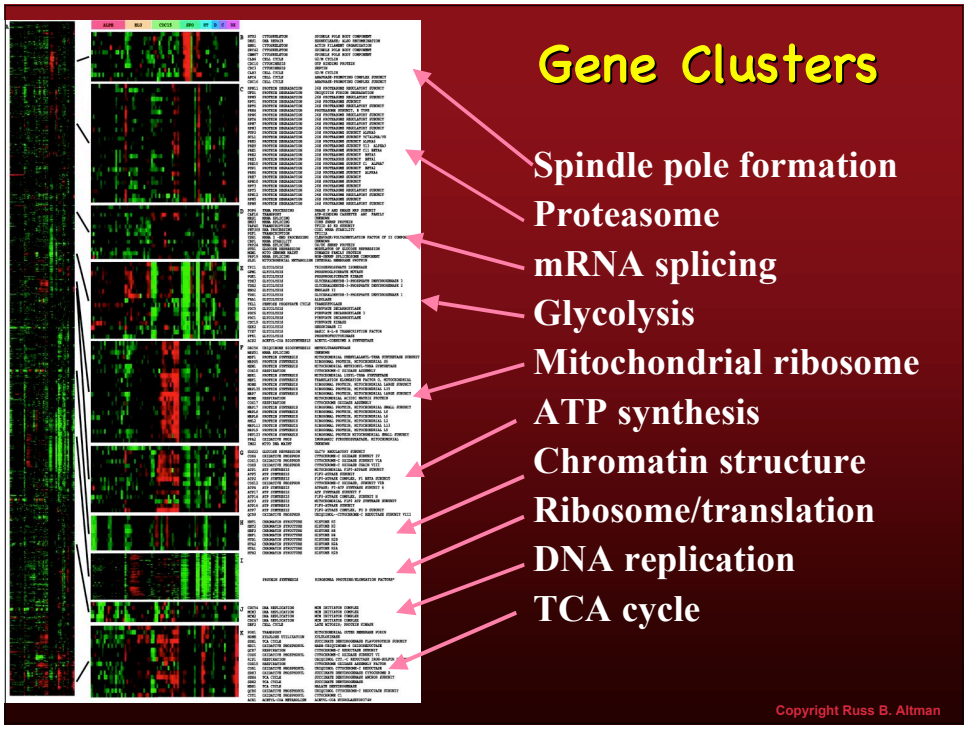
TERM	CORRECTED P-VALUE	UNCORRECTED P-VALUE	NUM_ANNOTATIONS / TOTAL_NUM_ANNOTATIONS	
chromatin assembly/disassembly	4.12076404033908e-11	9.36537281895246e-13	6 of 24	YDR225W,YBL003C
chromosome organization and biogenesis (sensu Eukarya)	9.90660015082729e-07	2.25150003427893e-08	7 of 219	YDR225W,YBL003C
nuclear organization and biogenesis	3.72275987438334e-06	8.46081789632578e-08	7 of 265	YDR225W,YBL003C
DNA packaging	1.44362962471306e-05	3.28097641980242e-07	6 of 186	YDR225W,YBL003C
establishment and/or maintenance of chromatin architecture	1.44362962471306e-05	3.28097641980242e-07	6 of 186	YDR225W,YBL003C
DNA metabolism	0.000352007286653523	8.00016560576188e-06	7 of 516	YDR225W,YBL003C
cell organization and biogenesis	0.0058908255605899	0.000133882399104316	8 of 1117	YBL003C,YBR009C
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	0.0301448651299023	0.000685110571134142	8 of 1395	YBL003C,YBR009C

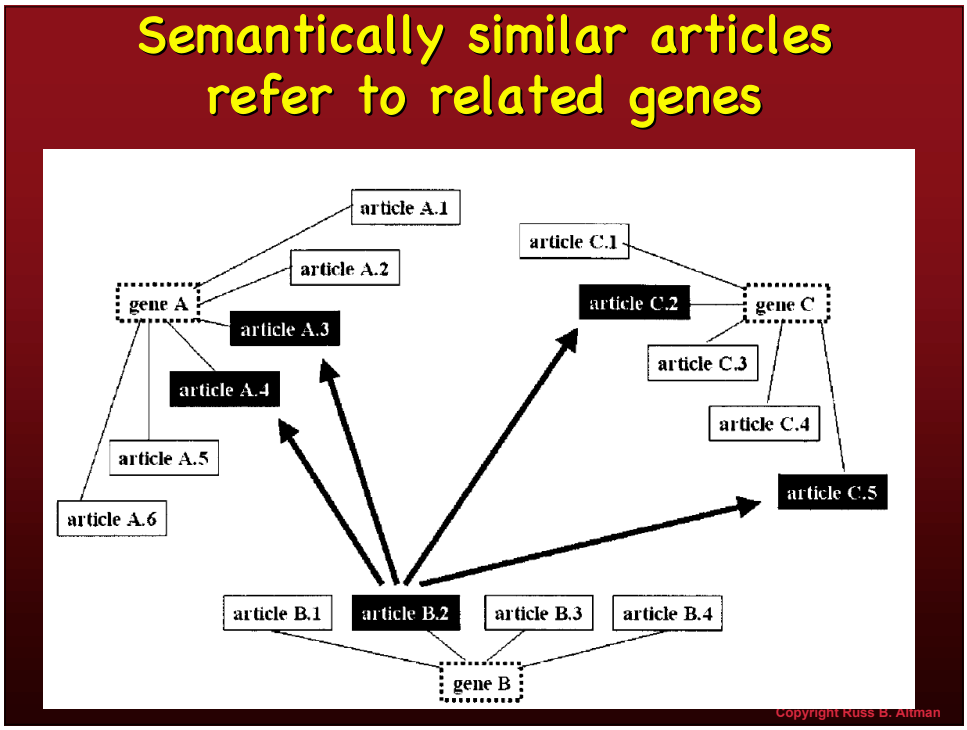
Another method for evaluating clusters.

GO codes are coarse and depend on human annotation.

How about looking at published literature for genes (as the humans do) directly?

Evaluate whether word/concept usage in literature is similar across family of genes in order to evaluate "functional coherence."


Gene Clusters
Spindle pole formation
Proteasome
mRNA splicing
Glycolysis
Mitochondrial ribosome
ATP synthesis
Chromatin structure
Ribosome/translation
DNA replication
TCA cycle
Copyright Russ B. Altman


Semantically similar articles refer to related genes
article A.1
article A.2
gene A
article A.3
article A.4
article A.5
article A.6
article C.1
article C.2
gene C
article C.3
article C.4
article C.5
article B.1
article B.2
article B.3
article B.4
gene B
Copyright Russ B. Altman

Analysis of Eisen Clusters

Function Label assigned to Expression Cluster (by Eisen et al)	Number of Genes	Neighbor Divergence Score	Score Percentile
ATP Synthesis	14	0.1358	99.9%
Chromatin Structure	8	0.1456	100.0%
DNA Replication	5	0.1867	100.0%
Glycolysis	17	0.2118	100.0%
Mitochondrial Ribosome	22	0.0269	53.3% ●
mRNA Splicing	14	0.0248	48.3% ●
Proteasome	27	0.3007	100.0%
Ribosome and Translation	125	0.2224	100.0%
Spindle Pole Body Assembly and Function	11	0.0272	53.8% ●
Tricarboxylic Acid Cycle and Respiration	16	0.1249	99.8%

Clustering vs. Classification

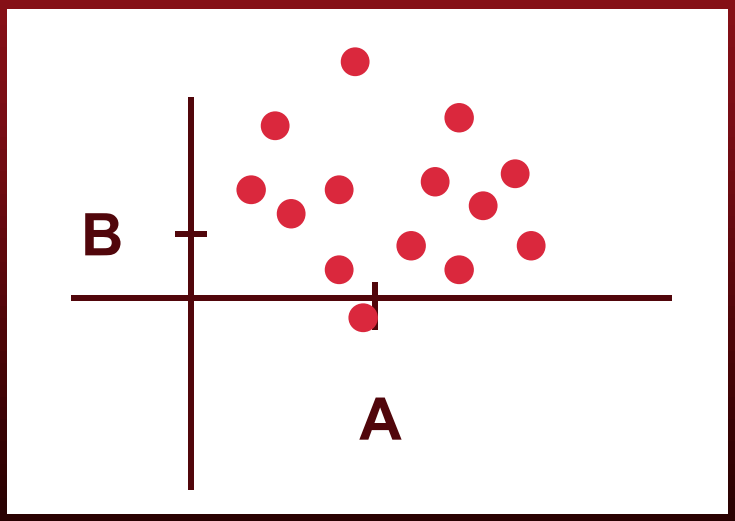
Clustering uses the primary data to group together measurements, with no information from other sources. Often called "unsupervised machine learning."

Classification uses known groups of interest (from other sources) to learn the features associated with these groups in the primary data, and create rules for associating the data with the groups of interest. Often called "supervised machine learning."

Classification Algorithms

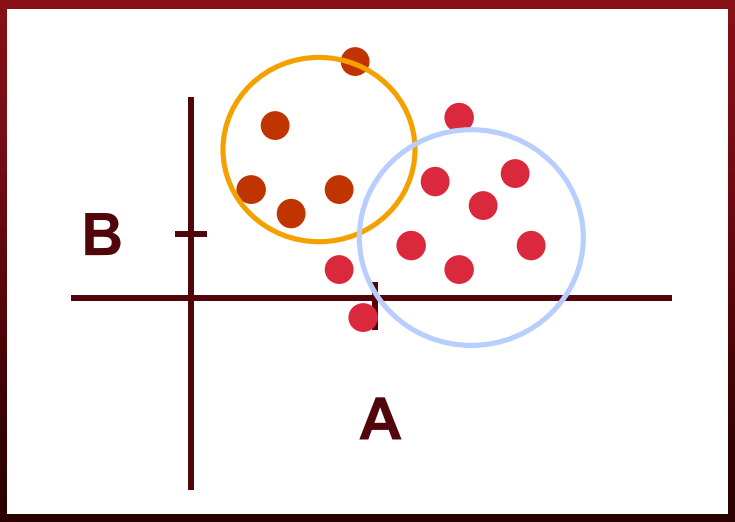
Graphical Representation

Two features f1 (x-coordinate) and f2 (y-coordinate)



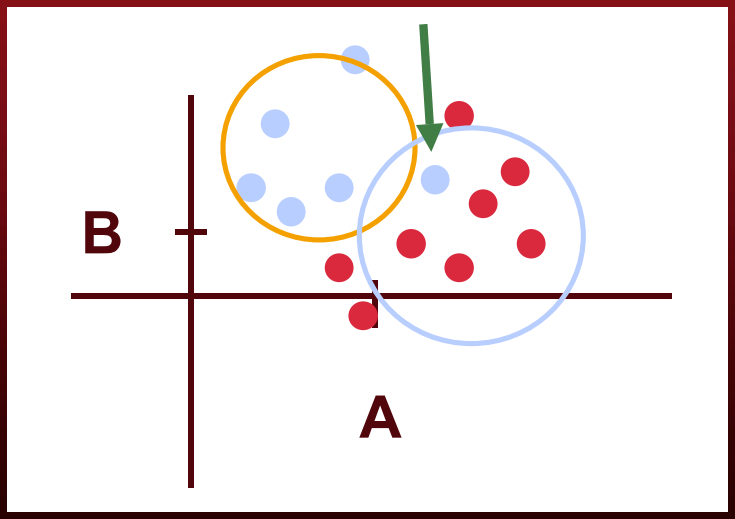
Clusters

Two features f1 (x-coordinate) and f2 (y-coordinate)

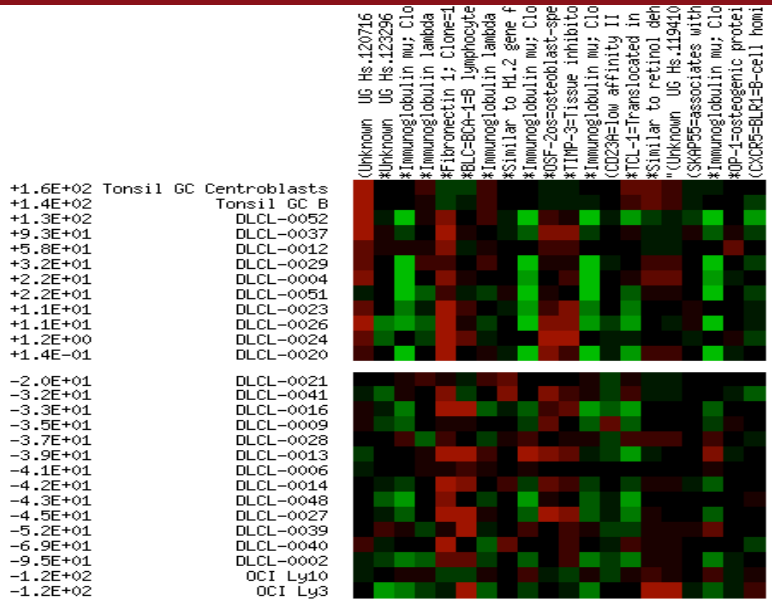


Apply external labels for classification

RED group and BLUE group now labeled



Classifying Lymphomas



Tradeoffs

Clustering is not biased by previous knowledge, but therefore needs stronger signal to discovery clusters.

Classification uses previous knowledge, so can detect weaker signal, but may be biased by WRONG previous knowledge.

Methods for Classification

- Linear Models
- Logistic Regressian
- Naïve Bayes
- Decision Trees
- Support Vector Machines

Linear Model

Each gene, g, has list of n measurements at each condition, [f1 f2 f3…fn].

Associate each gene with a 1 if in a group of interest, otherwise a 0.

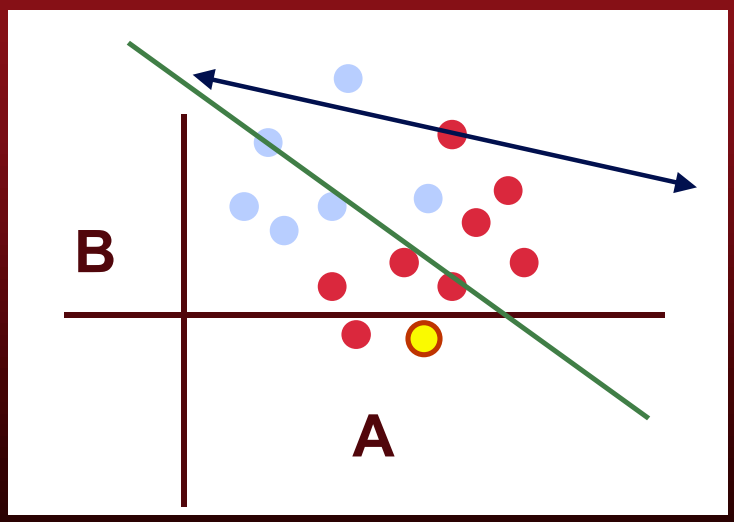
Compute weights to optimize ability to predict whether genes are in group of interest or not.

Predicted group = SUM [weight(i) * fi]

If fi always occurs in group 1 genes, then weight is high. If never, then weight is low.
Assumes that weighted combination works.

Linear Model

PREDICT RED if high value for A and low value for B, (high weight on x coordinate, negative weight on y)



Logistic Regression

(intro http://personal.ecu.edu/whiteheadj/data/logit/)

p = probability of being in group of interest
f = vector of expression measurements

$\mathrm{Log}[p/(1-p)] = a+\beta f$

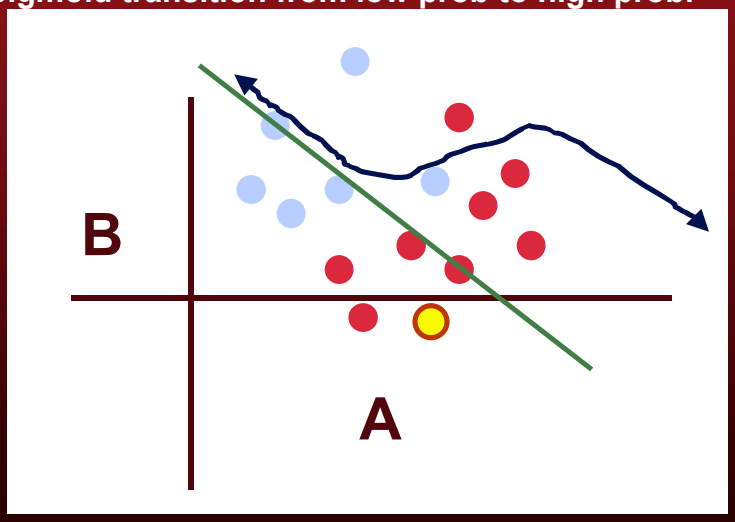
or

$p = e^{\beta f+a}/(1+e^{\beta f+a})$

Use optimization methods to find β (vector) that maximizes the difference between two groups. Then, can use equation to estimate membership of a gene in a group.

Logistic Model

PREDICT RED if high value for A and low value for B, (high weight on x coordinate, negative weight on y), but with Sigmoid transition from low prob to high prob.



Bayes Rule for Classification

Bayes' Rule: p(hypothesis|data) = p(data|hypothesis)p(hypothesis)/p(data)

p(group 1| f) = p(f|group1) p(group1)/p(f)

p(group 1|f) = probability that gene is in group 1 give the expression data

p(f) = probability of the data

p(f|group 1) = probability of data given that gene is in group 1

p(group 1) = probability of group 1 for a given gene (prior)

Naïve Bayes

Assume all expression measurements for a gene are independent.

Assume p(f) and p(group1) are constant.

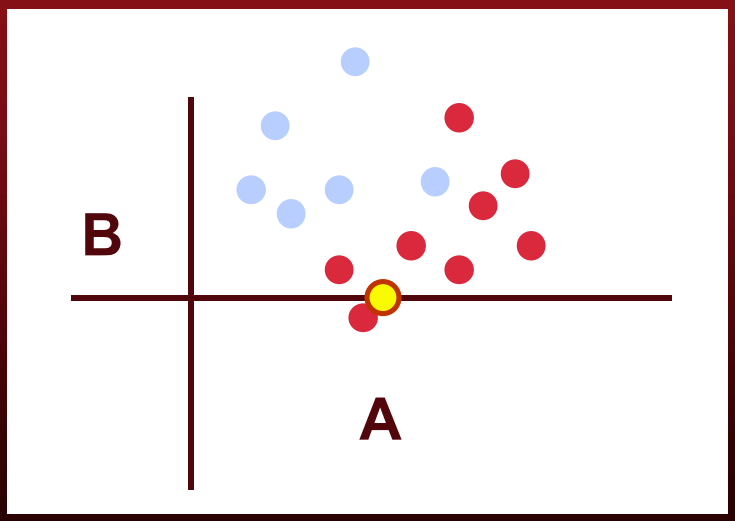
P(f|group 1) = p(f1&f2…fn|group1)
= p(f1|group1) * p(f2|group1)…* p(fn|group1)

Can just multiply these probabilities (or add their logs), which are easy to compute, by counting up frequencies in the set of "known" members of group 1.

Choose a cutoff probability for saying "Group 1 member."

Naïve Bayes

If P(Red|x=A) * P(Red| y = 0) = HIGH, so assign to RED



Decision Trees

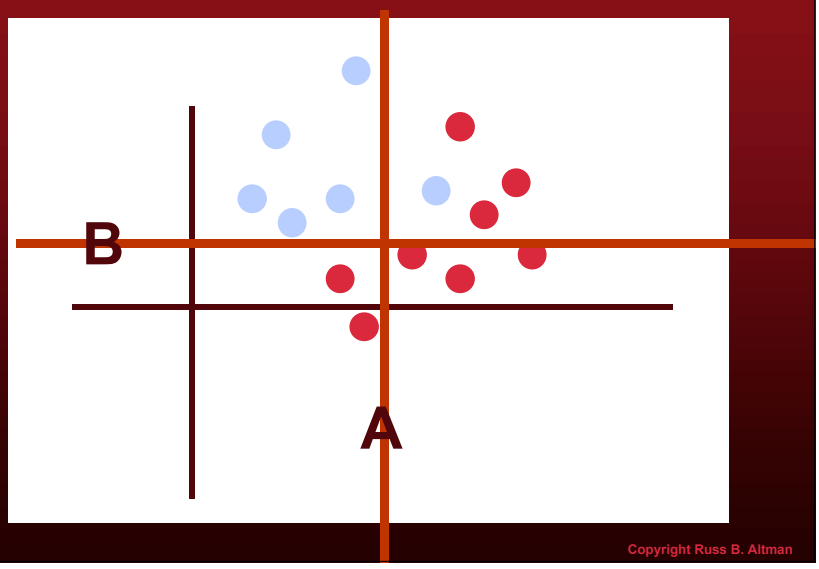
Consider an n-dimensional graph of all data points (f, gene expression vectors).

Try to learn cutoff values for each fi that separate different groups.

Decision Trees

If x < A and y > B => BLUE

If Y < B OR Y >B and X > A => RED



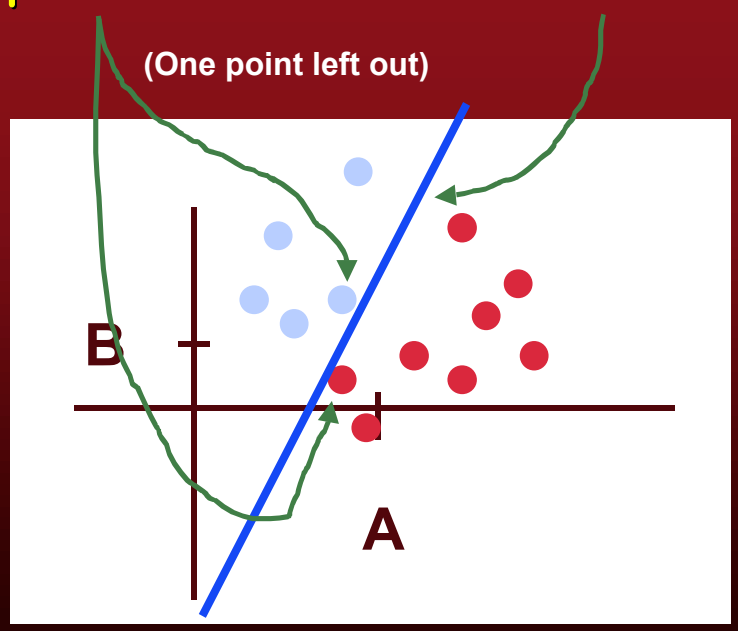
Support Vector Machines

Draw a line that passes close to the members of two different groups that are the most difficult to distinguish.

Label those difficult members the "support vectors." (Remember, all points are vectors).

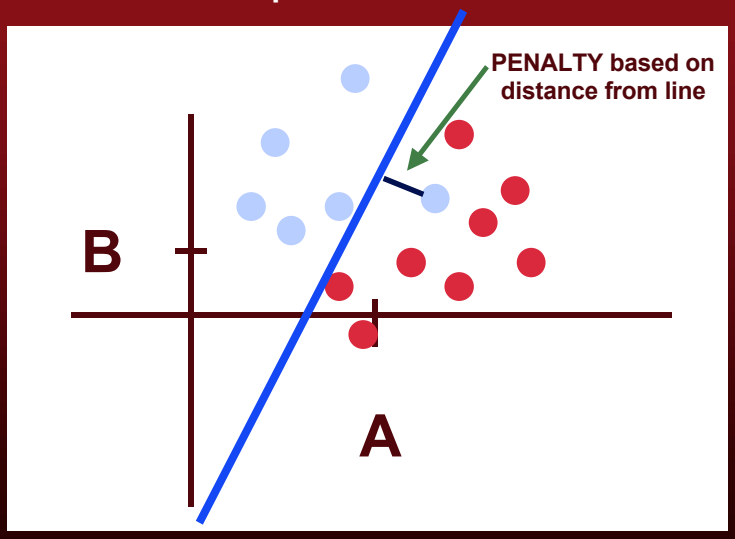
For a variety of reasons (discussed in the tutorial, and the Brown et al paper to some degree), this choice of line is a good one for classification, given many choices.

Support Vectors and Decision Line

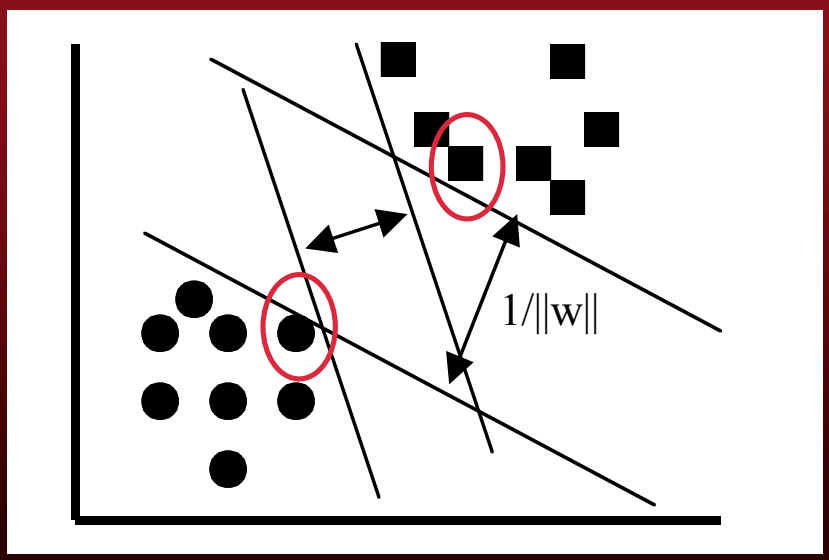


Support Vectors and Decision Line

(Bad point put back in…Can penalize boundary line for bad predictions



Choose boundary line that is closest to both support vectors



Notes about SVMs

If the points are not easily separable in n dimensions, can add dimensions (similar to how we mapped low dimensional SOM grid points to expression dimensions).

Dot product is used as measure of distance between two vectors. But can generalize to an arbitrary function of the features (expression measurements) as discussed in Brown and associated Burges tutorial.

Evaluating Yes/No Classifiers

True Positives
False Positives
True Negatives
False Negatives

Sensitivity = TP/(TP + FN)
Specificity = TN/(TN + FP)
Positive Predictive Value = TP/(TP + FP)

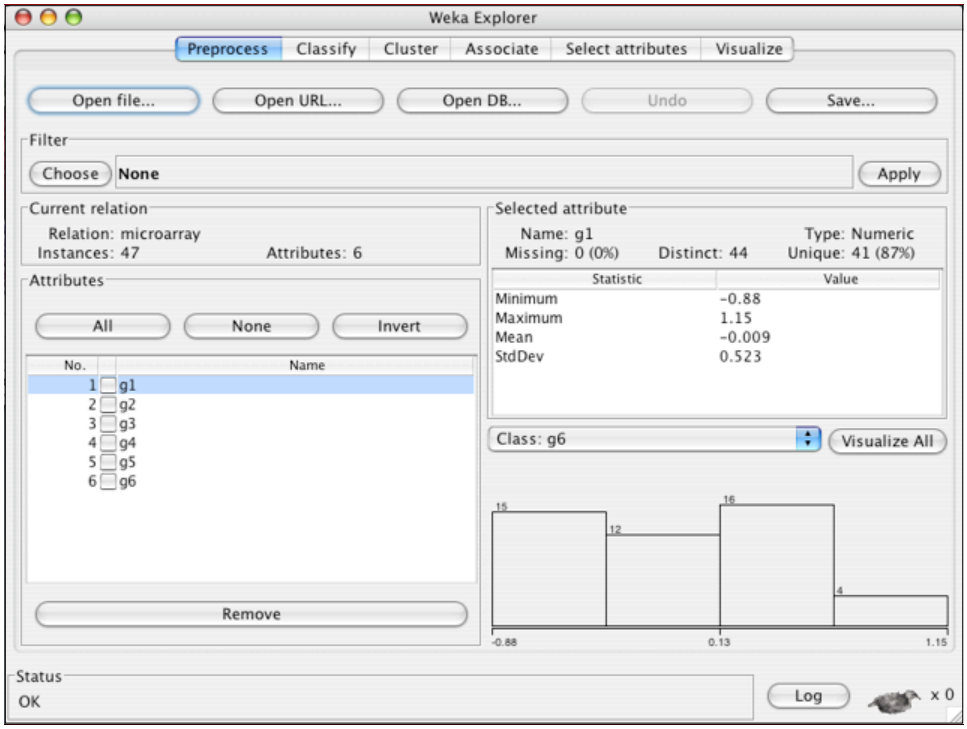
ROC Curve = Plot Sensitivity vs. Specificity
(or Sensitivity vs. 1-Specificity)

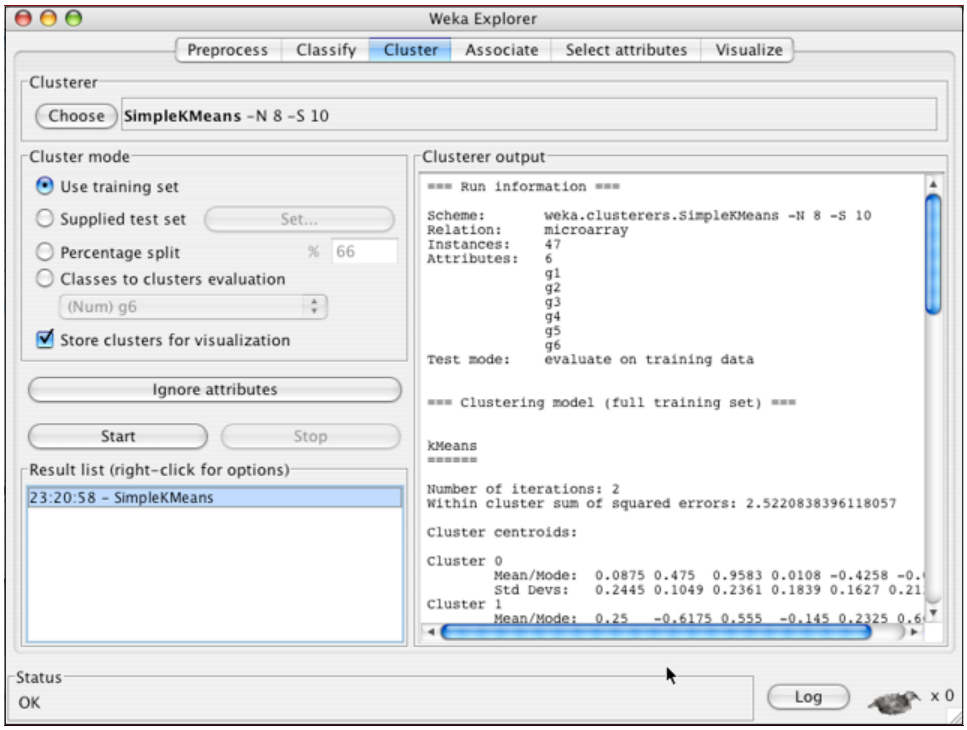
Interested in playing with cluster and classification methods?

http://www.cs.waikato.ac.nz/~ml/weka/

.ARFF format

```
%Weka test for russ.
%
@relation'microarray'
@attribute g1 real
@attribute g2 real
@attribute g3 real
@attribute g4 real
@attribute g5 real
@attribute g6 real
@data
0.23,-0.53,0.59,-0.13,0.22,0.68
0.42,-0.59,0.66,-0.20,0.20,0.65
0.21,-0.61,0.60,-0.02,0.35,0.78
0.14,-0.74,0.37,-0.23,0.16,0.55
0.32,-0.42,0.57,0.66,0.31,-0.03
0.55,-0.66,0.57,0.72,0.21,0.11
0.40,-0.55,0.61,0.70,0.22,0.13
0.29,-0.69,0.57,0.57,0.15,-0.11
-0.41,1.05,0.69,0.17,1.16,0.82
-0.46,0.80,0.88,0.33,1.09,0.83
-0.30,0.92,0.71,0.34,1.17,0.76
-0.54,0.75,0.67,0.09,0.96,0.57
-0.50,-0.19,-0.79,1.29,0.07,0.85
-0.30,-0.27,-0.59,1.14,0.22,0.85
-0.43,-0.43,-0.67,1.14,0.04,0.78
-0.29,-0.22,-0.66,1.02,0.05,0.77
-0.34,-0.19,-0.78,1.12,0.11,0.65
-0.52,-0.44,-0.89,1.00,-0.06,0.58
-0.02,0.61,1.03,0.13,-0.37,-0.25
-0.14,0.55,1.18,0.04,-0.48,-0.04
0.09,0.59,1.10,0.19,-0.65,-0.22
-0.12,0.38,1.12,0.23,-0.41,-0.29
0.11,0.49,1.21,0.17,-0.39,-0.30
```

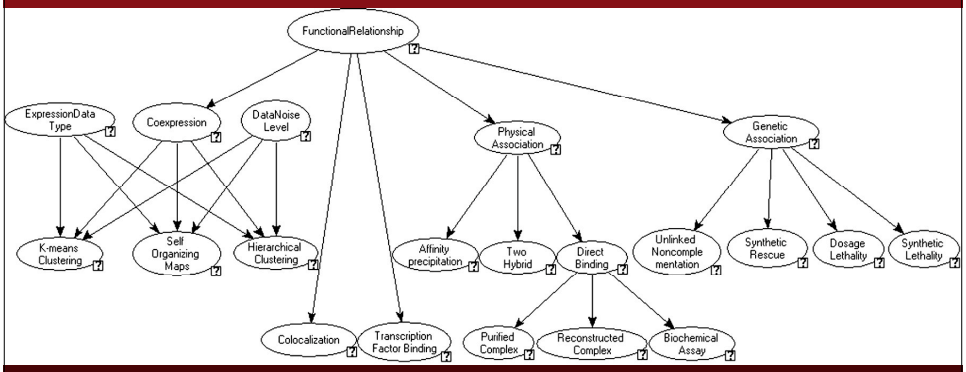

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: microarray
Instances: 47
Attributes: 6
Selected attribute
Name: g1
Type: Numeric
Missing: 0 (0%)
Distinct: 44
Unique: 41 (87%)
Statistic
Value
Minimum
-0.88
Maximum
1.15
Mean
-0.009
StdDev
0.523
Attributes
All
None
Invert
No.
Name
1 g1
2 g2
3 g3
4 g4
5 g5
6 g6
Remove
Class: g6
Visualize All
15
12
16
4
-0.88
0.13
1.15
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -N 8 -S 10
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
% 66
Classes to clusters evaluation
(Num) g6
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
23:20:58 - SimpleKMeans
Clusterer output
=== Run information ===
Scheme: weka.clusterers.SimpleKMeans -N 8 -S 10
Relation: microarray
Instances: 47
Attributes: 6
g1
g2
g3
g4
g5
g6
Test mode: evaluate on training data
=== Clustering model (full training set) ===
kMeans
Number of iterations: 2
Within cluster sum of squared errors: 2.5220838396118057
Cluster centroids:
Cluster 0
Mean/Mode: 0.0875 0.475 0.9583 0.0108 -0.4258
Std Devs: 0.2445 0.1049 0.2361 0.1839 0.1627
Cluster 1
Mean/Mode: 0.25 -0.6175 0.555 -0.145 0.2325
Status
OK
Log
x 0

Other Analysis of Microarray Data

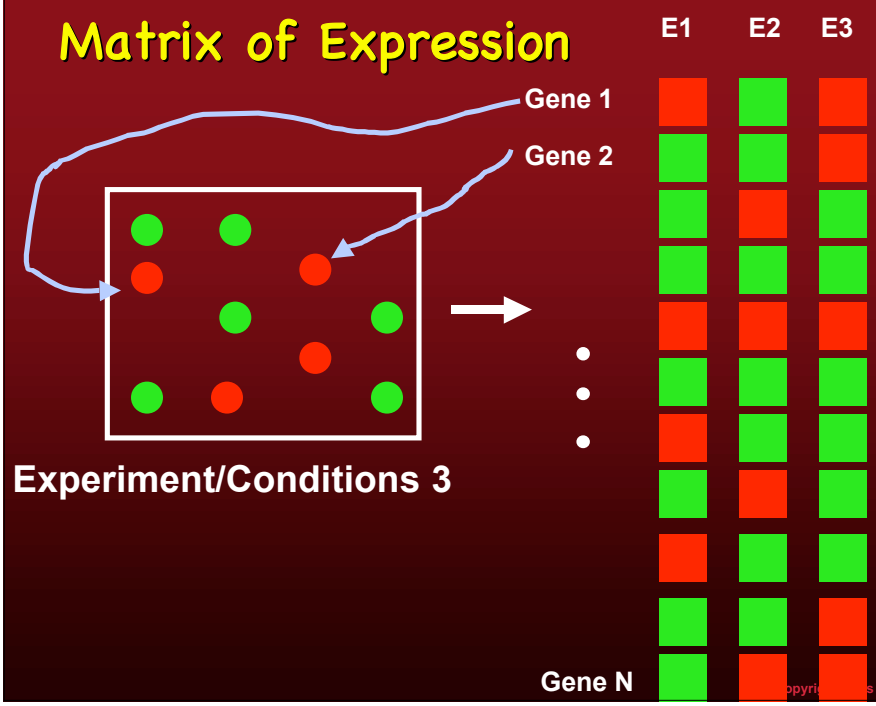
1. Reducing the number of genes/conditions that need to be considered
   - Principal Component Analysis (Raychaudhuri et al, 2000, Alter et al, 2000)
   - Independent Component Analysis (Lee & Batzoglou, 2003)
2. Combining microarray expression experiments with other sources of data to generate more robust hypotheses. (See "joint learning" session of http://helix-web.stanford.edu/psb05/)

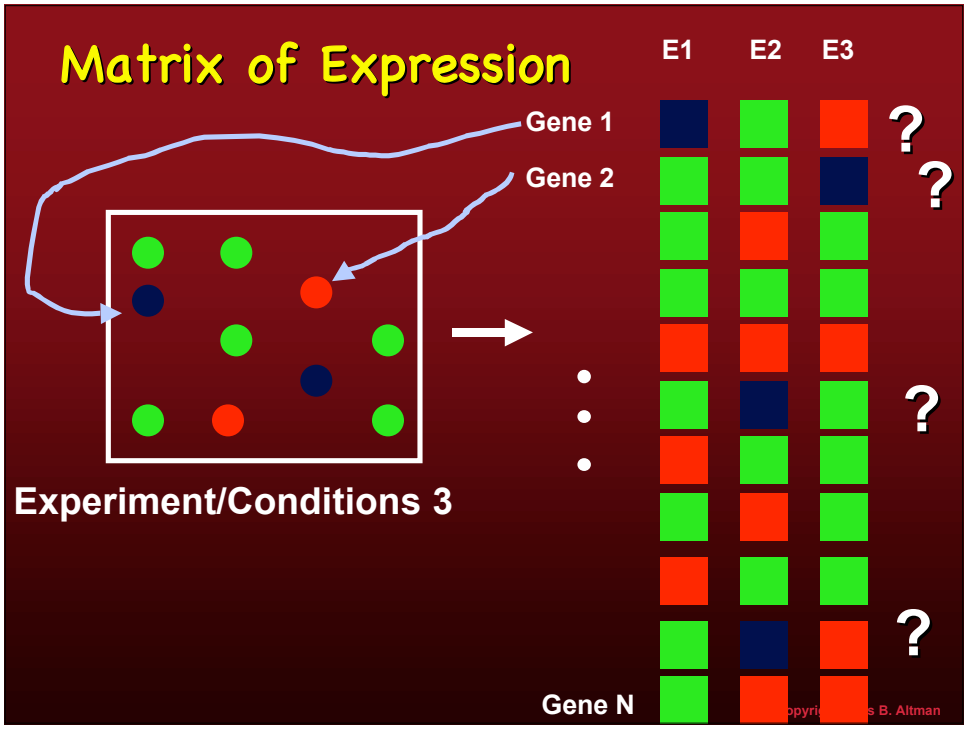
MAGIC (Troyanskaya et al, 2003, PMID 12826619)



Missing values in microarray data?

Matrix of Expression



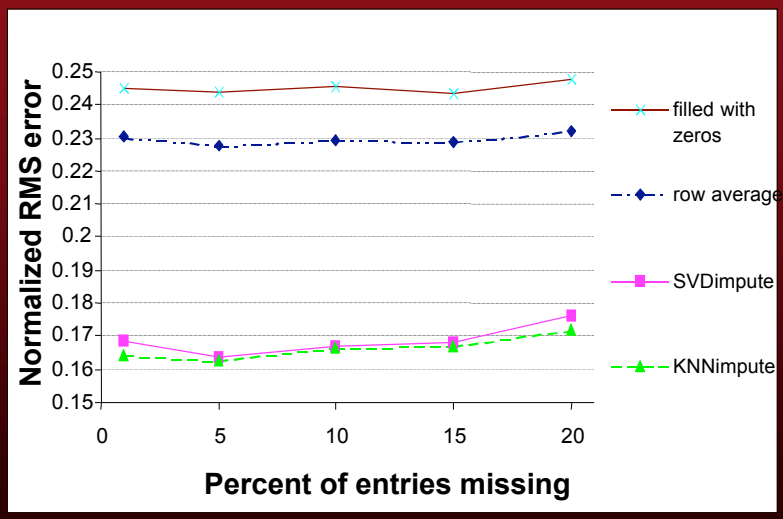


Most algorithms do not work if there are missing values (e.g. need to compute distances)

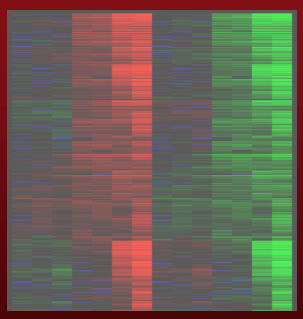
POTENTIAL SOLUTIONS:

1. Put zeros in all missing values
2. Put average of all values that are available = row average or column average
3. Estimate values based on nearest neighbor, or group of K nearest neighbors
4. Estimate value in others ways (e.g. Singular Value Decomposition)

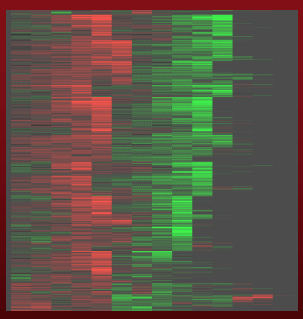
Troyanskaya O et al. Missing value estimation meth...[PMID: 11395428]



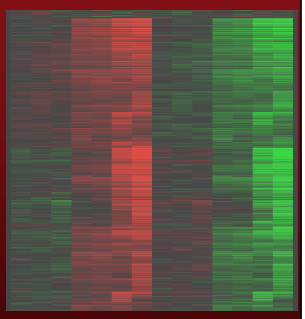
Estimate Missing Values.



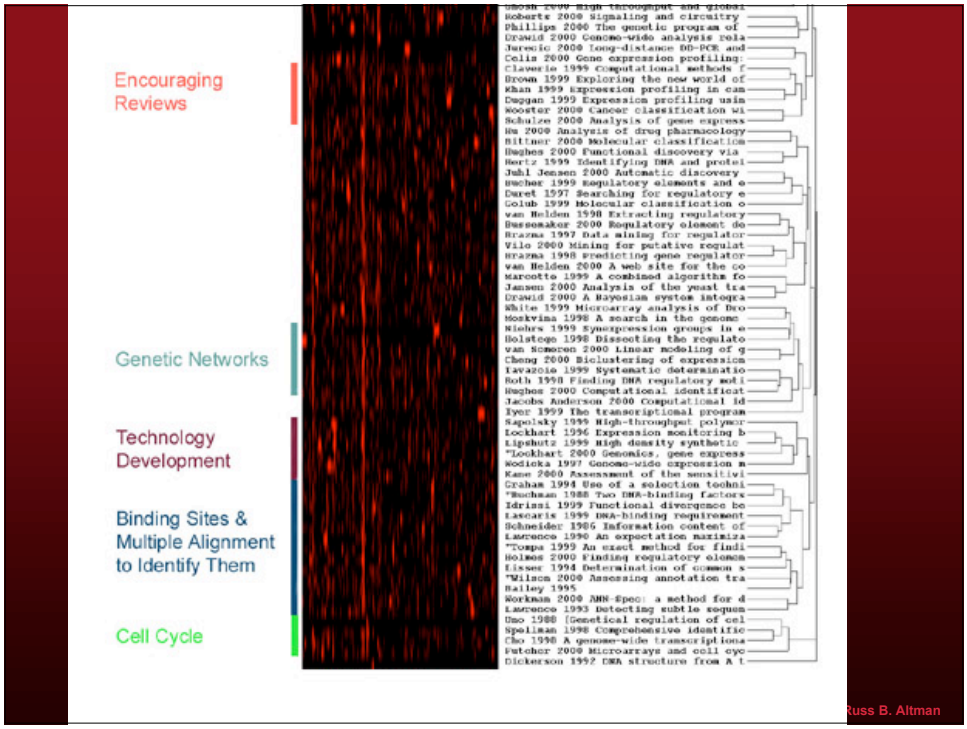
Complete data set

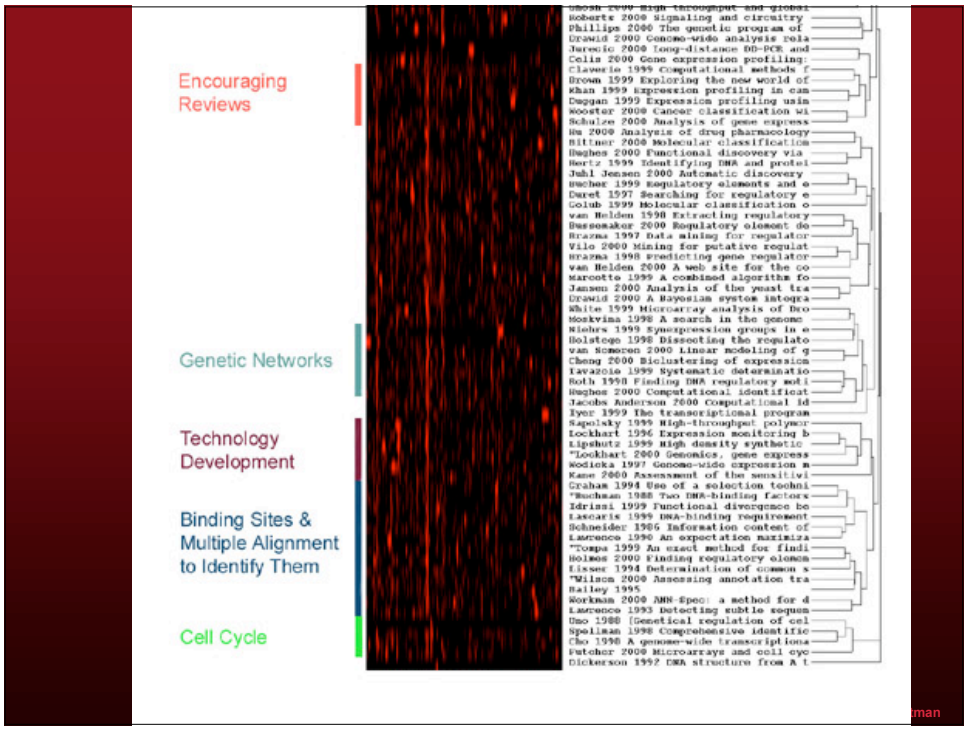


Data set with 30% entries missing (missing values appear black)



Data set with missing values estimated by KNNimpute algorithm


Encouraging Reviews
Genetic Networks
Technology Development
Binding Sites & Multiple Alignment to Identify Them
Cell Cycle
Russ B. Altman


Encouraging Reviews
Genetic Networks
Technology Development
Binding Sites & Multiple Alignment to Identify Them
Cell Cycle

Conclusions

1. Methods exist (and are still needed) for characterizing clusters that emerge from high throughput data, such as microarrays.
2. Gene Ontology is a useful way to gauge significant trends.
3. Classification methods are useful, and easily available.
4. Missing data can be imputed, but be careful about over-imputing!