Using GO

Robert Gentleman

Harvard School of Public Health

Outline

- What is an ontology
- The Gene Ontology (GO) and GOA
- Statistical Problems
- Examples
- Conclusions

Comments

- all the graphs, graphics and statistical methods discussed can be produced with BioConductor Software
- the paper this talk is based on is titled

  Using GO for Statistical Analyses
- it is available in both PDF and compendium format

  http://bioconductor.org/Docs/Papers/2003/Compendium

BioConductor Software

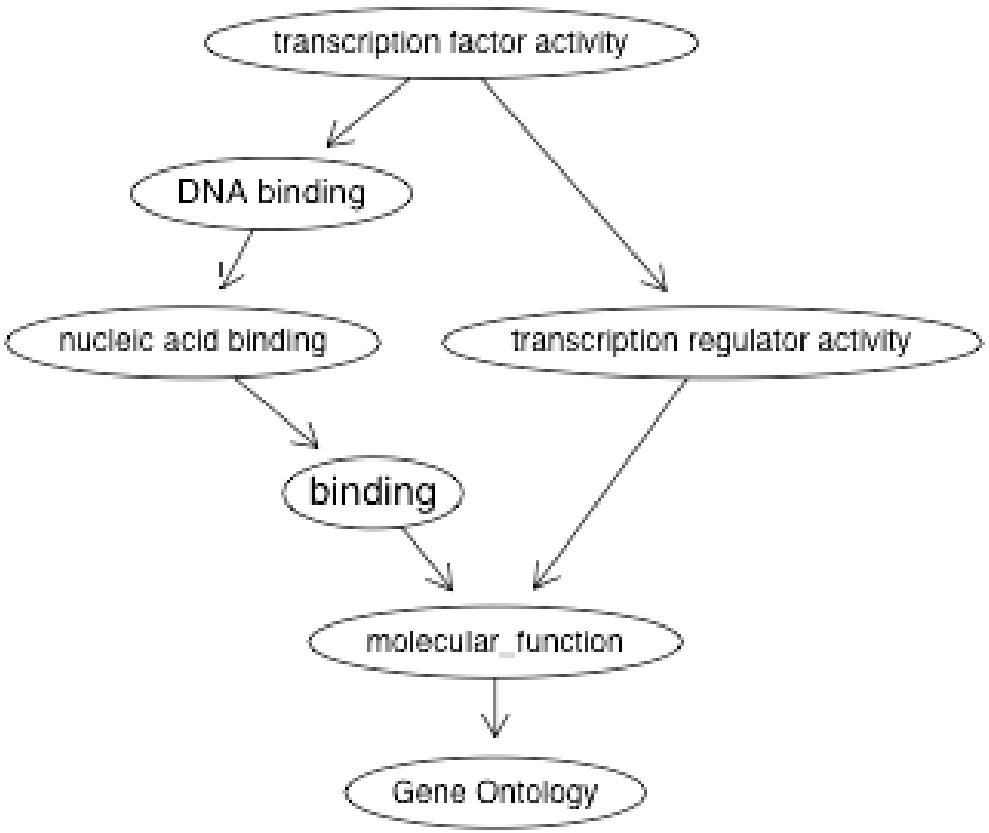
- rely on ***graph***, ***Rgraphviz***, ***RBGL***, ***GOstats***
- meta-data packages: ***hgu95av2***, ***GO***
- made use of: ***Biobase***, ***genefilter, multtest, xtable, Sweave (tools package)***

Ontology

- for our purposes an ontology is a restricted vocabulary
- keywords are another restricted vocabulary
- the purpose of these restricted vocabularies is to encourage a common usage that while not necessarily either entirely correct nor comprehensive does aid in both human and computer searching

Gene Ontology

- the Gene Ontology (GO) is a collection of ontologies (currently three) that describe genes and gene products
- these ontologies are restricted vocabularies that have the structure of directed acyclic graphs (DAGS)
- the most specific terms are the leaves of the graph
- there are edges from more specific terms (children) to less specific (parents)


transcription factor activity
DNA binding
nucleic acid binding
transcription regulator activity
binding
molecular_function
Gene Ontology

GO

- relationships between parent and child can be either “is-a” or “has-a”
- GOA (and others) provide mappings between terms and genes
- each gene is mapped to the most specific terms that are appropriate
- the other gene-term mapping are determined from the GO graph
- for a given gene and ontology the set of all appropriate terms is called the induced GO graph

The Ontologies

- MF: molecular function, terms describe what the gene/gene product does
  - 7280 terms
- BP: biological process, terms describe the biologic objectives (smaller than a pathway bigger than a function)
  - 8172 terms
- CC: cellular component, terms describe where in the cell the gene product resides (works)
  - 1388 terms

Why are Ontologies Useful

- P. W. Lord suggests:

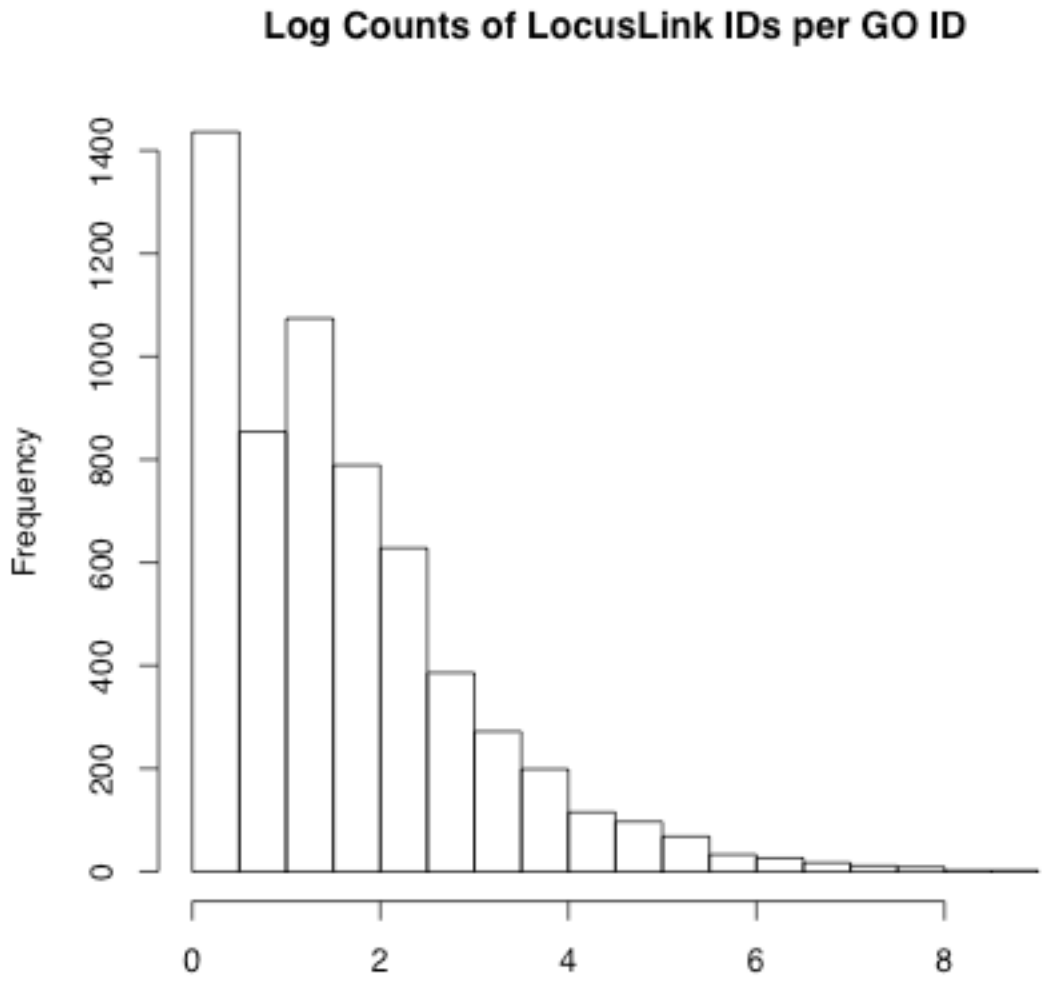
  Ontologies represent a communities domain knowledge in a form that is accessible by humans and amenable to computation

- making ontologies more complex does make them more descriptive, but at a cost, as complexity goes up we can no longer compute easily

Practicalities

- GO terms are mapped to LocusLink identifiers by GOA
- they use a variety of evidence codes (basically reasons for the mapping)
- GO:0004715 is in the MF ontology and its description is: *non-membrane spanning tyrosine kinase activity*

```
> get("GO:0004715", GOGO2LL)
   IDA    IEA      ISS      ISS      ISS      ISS      ISS      ISS      ISS      NAS
"2064" "25255" "13548" "14302" "35524" "36442" "37233" "38489" "45821"  "2534"
   NAS      NAS    TAS    TAS      TAS    TAS    TAS    TAS    TAS    TAS
"32080" "44353"  "1195"  "2185"  "2444"  "3702"  "5753"  "7006"  "7294"  "7297"
   TAS
"8711"
```


Log Counts of LocusLink IDs per GO ID
Frequency
0
200
400
600
800
1000
1200
1400
0
2
4
6
8

Microarray Experiments

- mappings from manufacturer's IDs to LL IDs can be many to one
- the data analyst must adjust for this multiplicity
- for HGU95av2, has 12625 probe sets

1	2	3	4	5	6	7	8	9
6756	1581	498	117	30	17	11	8	1

Statistical Problems

- data reduction
  - use GO to reduce the set of genes of interest
- semantic associations
  - use GO to provide meaning for genes or a basis for a specific investigation
  - location of interesting terms
  - division by cellular location, biological process, molecular function

Experiment

- S. Chiaretti of the Ritz Lab (DFCI) carried out a comprehensive study of gene expression in ALL (Acute Lymphoblastic Leukemia)
- we look at the 37 patients with BCR/ABL (9-22 translocation) and 42 with no known chromosomal abnormalities
- HGu95Av2 chips were run (12625 probes), and genes filtered for expression and variation (non-specific filtering)
- we are left with 2031 genes

Experiment

- BCR/ABL is known to mediate some of its effect through tyrosine kinase activity
- one approach might be to restrict attention to genes that are annotated at the *tyrosine kinase* node of the MF ontology (GO:0004713)
- there were 230 different probes annotated there and of these only 32 were selected by our non-specific filtering procedure

t-tests

- two permutation *t*-tests were applied to compare expression betweenn BCR/ABL and NEG (*p*-values FDR corrected)

IDs	40480_s_at	2039_s_at	56643_at	2057_g_at
Ty K	0.0001	0.0004	0.0206	0.0712
All	0.001	0.018	0.473	0.8223

- this is no surprise - if you test more things the correction is more severe

Tests

- you will generally be better off by testing fewer, more relevant, hypotheses
- *p*-value corrections are really a *band-aid* **not** a solution
- they adjust the cut-off so that those above it are enriched for truly false hypotheses but that is at the expense of rejecting more truly false hypotheses whose *p*-values fail to attain the level of the cut-off

Shortest Paths

- for microarray data we are examining gene (mRNA) expression levels
- this is typically an average over thousands of cells
- correlated expression can be related to similarity of function
  - Ge *et al* relate expression and protein complex co-membership (use a time-course experiment)
  - Zhou *et al* use shortest paths in gene expression data for annotation (used Rosetta Compendium - consists mainly of knock-outs and drug treatment)
  - here we are looking at a cohort study

Shortest Path

- the premise for our investigation was that transcription factors are largely self-regulatory
- we compare paths between transcription factors computed separately for the two subgroups of interest (BCR/ABL and NEG)
- let C_{uv} denote the absolute value of the Pearson correlation between genes *u* and *v*
- and edge exists in the graph, between nodes *u* and *v* if C_{uv} is larger than 0.6
- and in that case the distance between the nodes is 1- C_{uv}

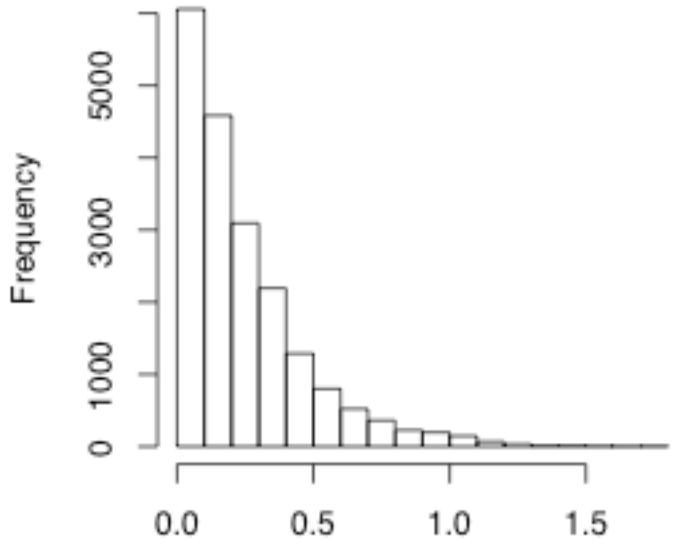
Shortest Path

- GO:003700 is the term for the molecular function of transcription factor activity
- for the U95av2 chip there are 814 probes mapped to this term, 531 unique LL IDs
- for our genes (those that passed the non-specific filter) there were 152 probes and 146 unique LL IDs (dropped the six as their correlations with the retained ones were high)
- so we have a graph on 2391 nodes and from that we compute a 146 by 146 distance matrix

Shortest Path

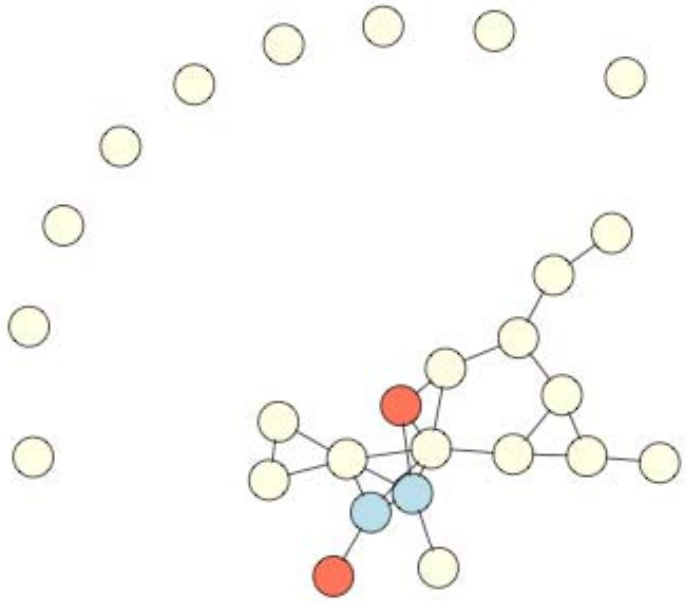
- both graphs have the same number of nodes (but not the same connected components)
- the BCR/ABL graph has 78182 edges
- the NEG graph has 87936 edges
- the between transcription factor distances in the NEG graph tended to be larger than those in the BCR/ABL graph
- preliminary investigations suggest that it is due to fewer steps in the path, not higher correlation

Histogram of the absolute value of the pairwise difference in distances between transcription factors (NEG vs BCR/ABL)

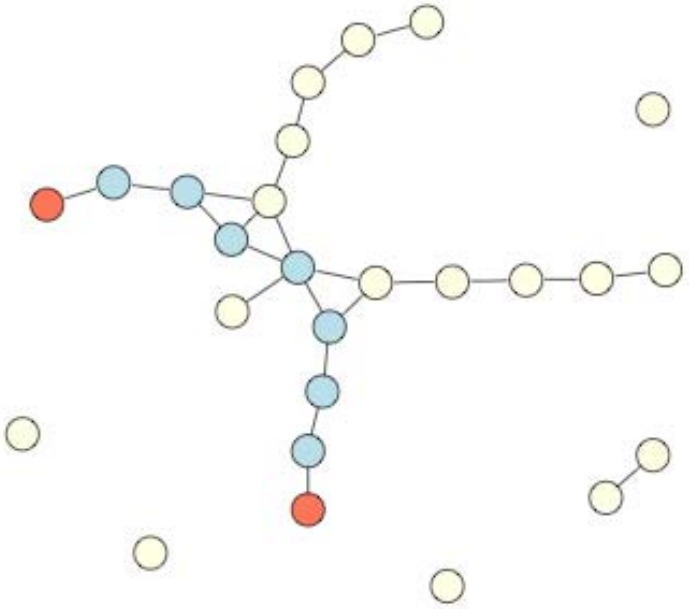


Path from MYC to MPO

Path in BCR/ABL Graph



Path in Neg Graph



MYC-MPO

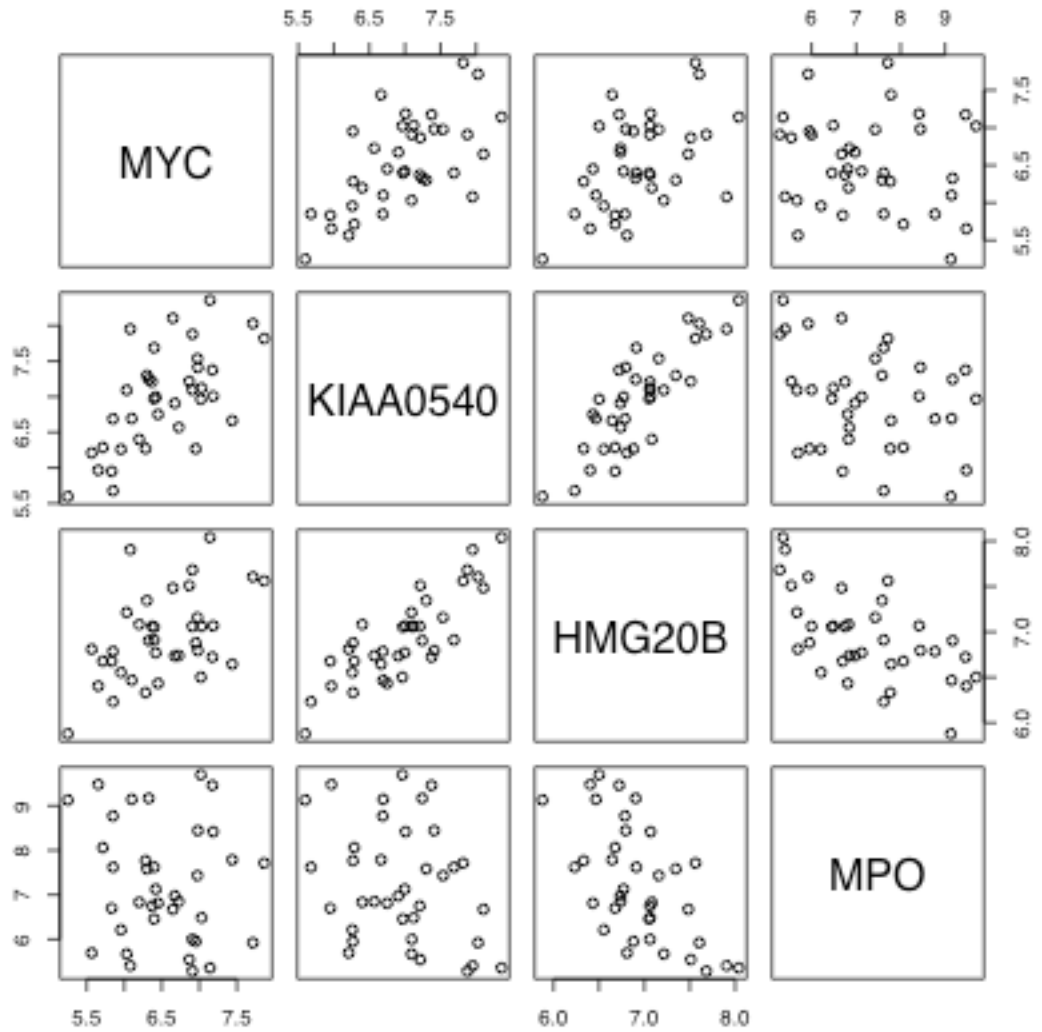
- For BCR/ABL samples:

  "MYC->KIAA0540->HMG20B->MPO”

- For NEG samples:

  "MYC->CDC25B->TRAP1->POLR2H->MSH2->EMP3->S100A4->LGALS1->MPO"

Pairwise scatter plot for BCR/ABL path from MYC to MPO



Transitive expression

- notice that the relationship (if real) between MYC and MPO seems to be mediated by other genes
- Zhou *et al* refer to this as transitive co-expression
- note there is nothing temporal in my graphic (these are representations of some sort of steady state, averaged across many cells)

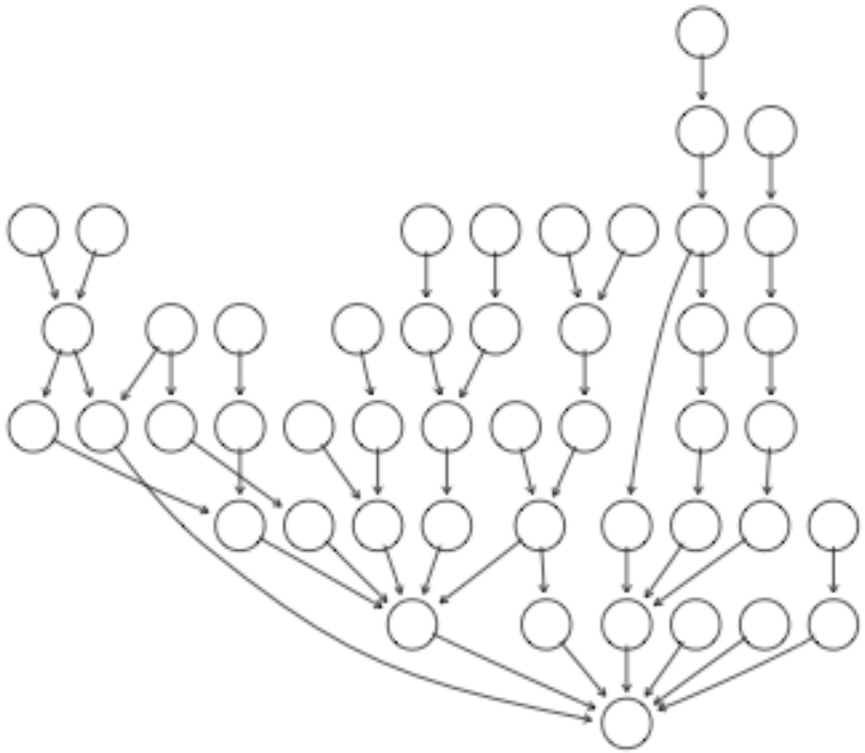
Finding Interesting GO terms

- GO has often been used to detect sets of terms that are *overrepresented* in a set of selected genes
- suppose that you have selected a set of *interesting* genes
- in our case we can choose those genes which differentiate BCR/ABL from NEG by *t*-test

Finding Interesting GO terms

- find all genes that are differentially expressed (you can decide what that means)
- select an ontology of interest
- find the set of mappings from the interesting genes to the most specific applicable terms
- from these terms and the GO structure find the *induced GO graph*

The induced GO graph for the MF ontology



The Test: two-way tables

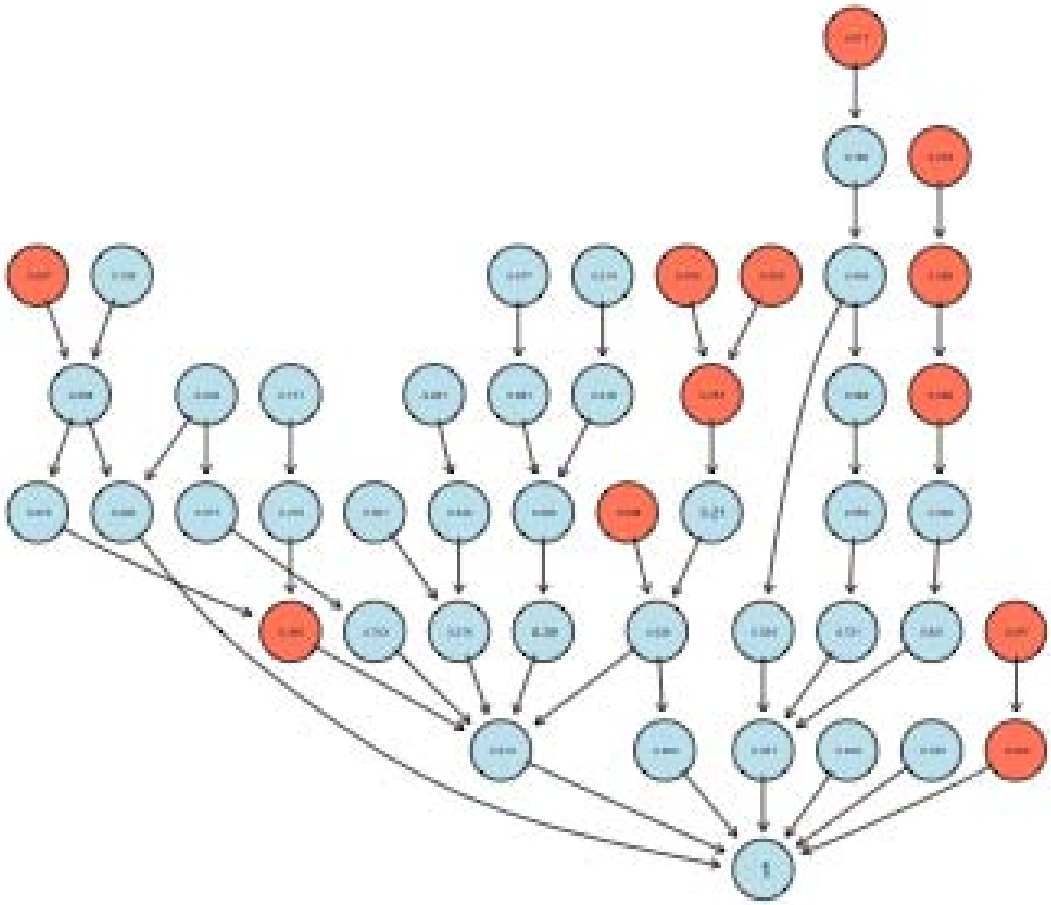
- the test most often employed is the Hypergeometric
- There are a total of N balls (LocusLink IDs)and each can either be annotated at the node of interest (or not) and each can be interesting or not.
- so we can do two way table testing (Fisher's exact test and the Hypergeometric sampling are the same thing)

Equivalent: Hypergeometric

- our urn contains N balls
- our gene list has *m* balls (we can think of these as the white balls in the urn) and hence N-*m* black balls
- a node in the graph has *k* genes annotated at it
- we think of there being *k* draws from the urn and ask if we got too many white balls

The GO graph

- we can now label (and add color) to the induced GO graph
- if the Hypergeometric *p*-value is less than 0.1 then the node is colored red, otherwise it is blue



Exploration

- which terms are associated with the extreme nodes?
- are there lots of genes there or only a few (node size)
- are there nodes that are underrepresented?
- should we further subdivide:
  - we could use either CC or BP categories to further categorize the genes here
  - and hence investigate not just what the gene is doing (MF) but additionally where, or for what reason

Testing - Issues

- there seems to be a parent-child pattern of significance
- not surprising given the construction
- there is also a problem of size
  - what should the size of the test be
  - number of significant nodes divided by the total number of nodes?

Testing - Issues

- if we agree that size should be number of significant nodes divided by the number of nodes then there is a problem
- in general the estimated proportion does not equal the nominal *p*-value (too many rejections)

Testing issues

- part of the problem is due to the method of construction
- every node in the induced graph has one or more interesting genes annotate at it and if very few genes are annotated there then the node is significant
- if there are 100 interesting genes then min size is 4 (any node in the induced graph of size less than 4 must be significant at the 0.05 level)
- and max size is 6034 (any node of this size or more will never reject-even if all genes are there)

Testing Issues

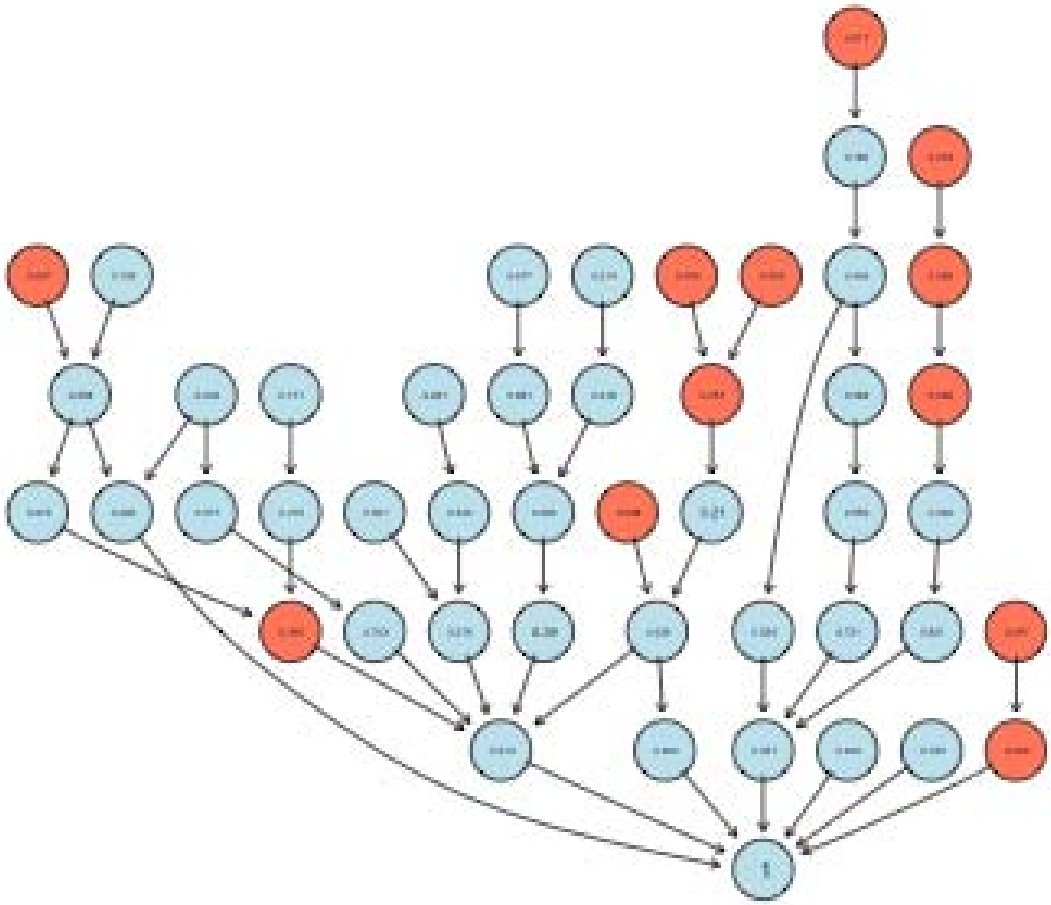
- in a simulation study (select K=100 genes at random, find the induced GO graph and carry out the significance test)
- below we have probability of rejection as a function of node size (log number of genes)

(-0.00872,2.18]	(2.18,4.36]	(4.36,6.54]	(6.54,8.73]
0.378	0.029	0.000	0.000

- nominal *p*-value was 0.05; overall empirical rejection rate was 0.123
- more extensive investigations are warranted

Testing Issues

- in addition it is not clear how to carry out any adjustment
- the tests at different nodes are typically not independent (every gene annotated at a child node is annotated at the parent)
- preliminary investigations indicate that there are issues of parent-child significance that need to be addressed



GO and Distances

- GO can be used to define, or describe, between gene distances
- the more similar the annotation the more similar the genes are
- many ideas have been put forward, most are not entirely satisfactory

Distances

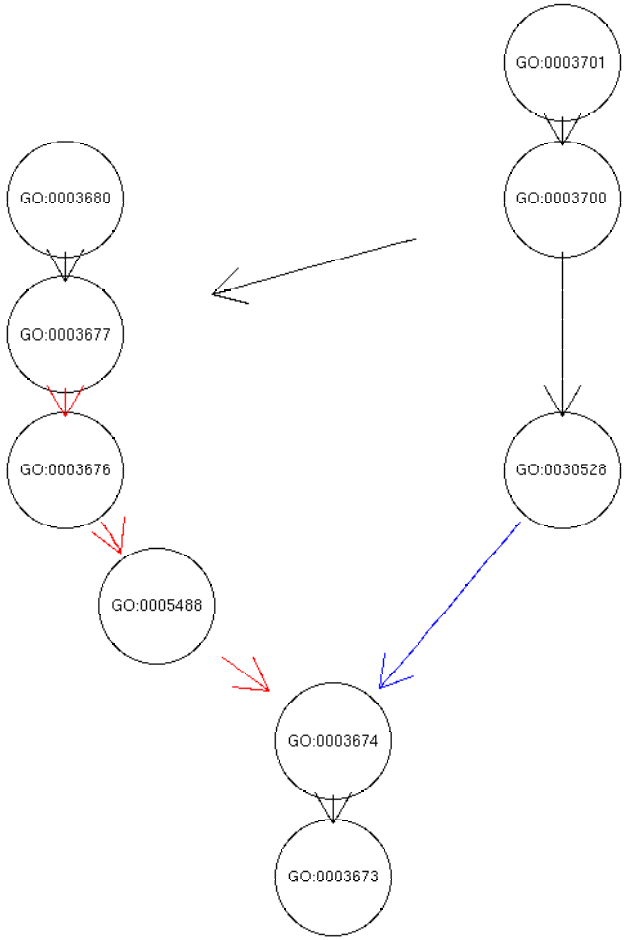
- distances between terms
- distances between genes
  - expression data
  - use the terms they are annotated at and the distances between terms

Distances Between Terms

- Cheng et al suggest that a similarity measure could be based on the number of edges in common on the shortest path to the root
- they claim that this relates to biological similarity – bigger is better with similarities
- they also state that edges close to the root have more biological importance and deserve greater weight

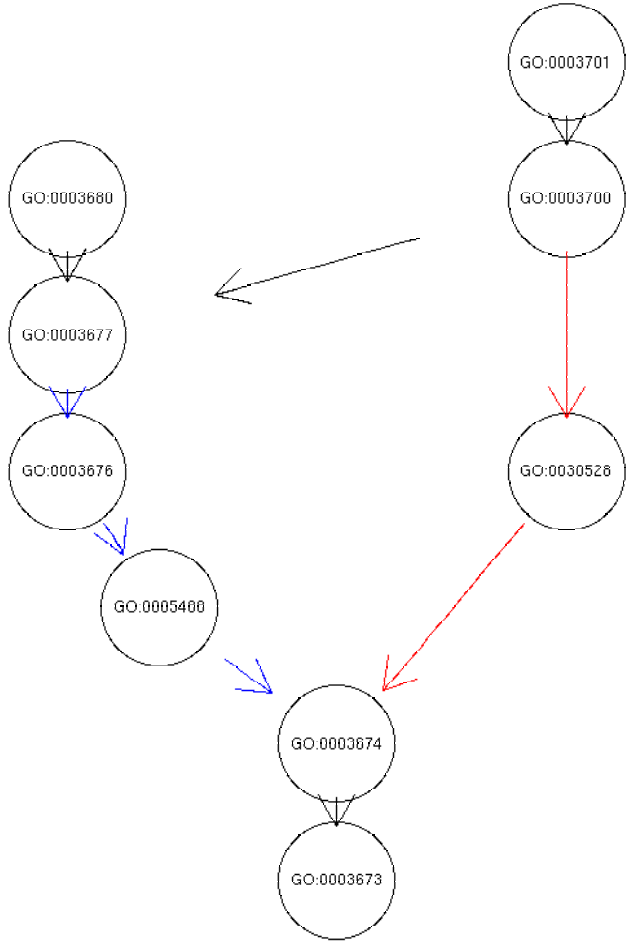
Distances

- for GO:0003700 the similarity with GO:0030528 is 1 and the similarity with GO:0003677 is 3
- they are both parents of GO:0003700



Distances

- GO:0003700 is more similar to GO:0003680 (3) than to its own child, GO:0003701 (2)



Distances

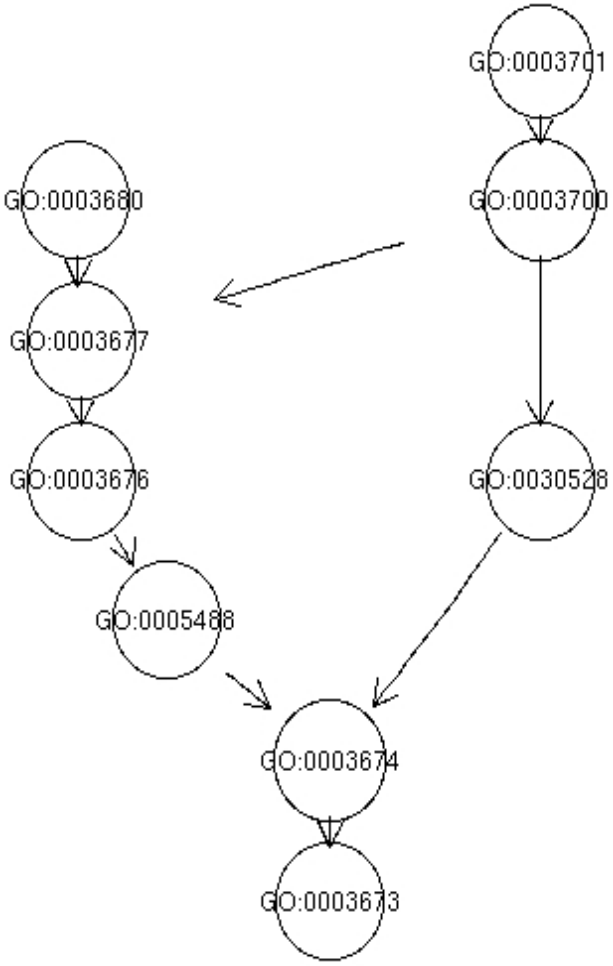
- B. Ding and I have been looking at a different measure of similarity between terms
- for any two terms, T_1 and T_2, find the induced GO graphs
- define $S_D(T_1, T_2)$ to be the set of nodes that they have in common divided by the number of nodes in total

Distances

- i.e. the cardinality of their intersection divided by the cardinality of their union
- larger values correspond to more similarity
- S_D can be criticized on the grounds that it does not account for the complexity of the graphs being considered
- the terms GO:0005488 and GO:0030528 have the same similarity as GO:0003700 and GO:0030528 (1/3 and 2/6)

Distances

- a comparison with S_Y
  - $S_Y(3700, 30528) = 2/6$
  - $S_Y(3700, 30677) = 4/6$
- but
  - $S_D(3700, 3701) = 6/7$
  - $S_D(3700, 3680) = 4/7$
- for S_D GO:0003700 is more similar to its child than to GO:0003680



Another Distance

- use the suggestion of Cheng to create a between gene similarity measure
- for any two genes, g_i, g_j,
  - find the set of common annotations (the intersection of their induced GO graphs)
  - find the depth of each term; distance to the root node
  - the similarity is the maximum depth

Distances

- we can define distances between genes using the distances between the GO terms that they are annotated at
- many genes are annotated at multiple terms within the same ontology
- we could define the distance to be the minimum (or the maximum) of all pairwise distances

Distances

- we might also want to consider why a gene is annotated at multiple terms
- in some cases this shows that the gene is related to many biological processes, performs several molecular functions, is a part of many cellular components
- in other cases (according to GO documentation) it demonstrates uncertainty

Distances: Information Content

- Lord *et al* consider similarities or distances on the basis of the information content at a node
- this measure of distance requires some set of annotated genes to provide data on information content

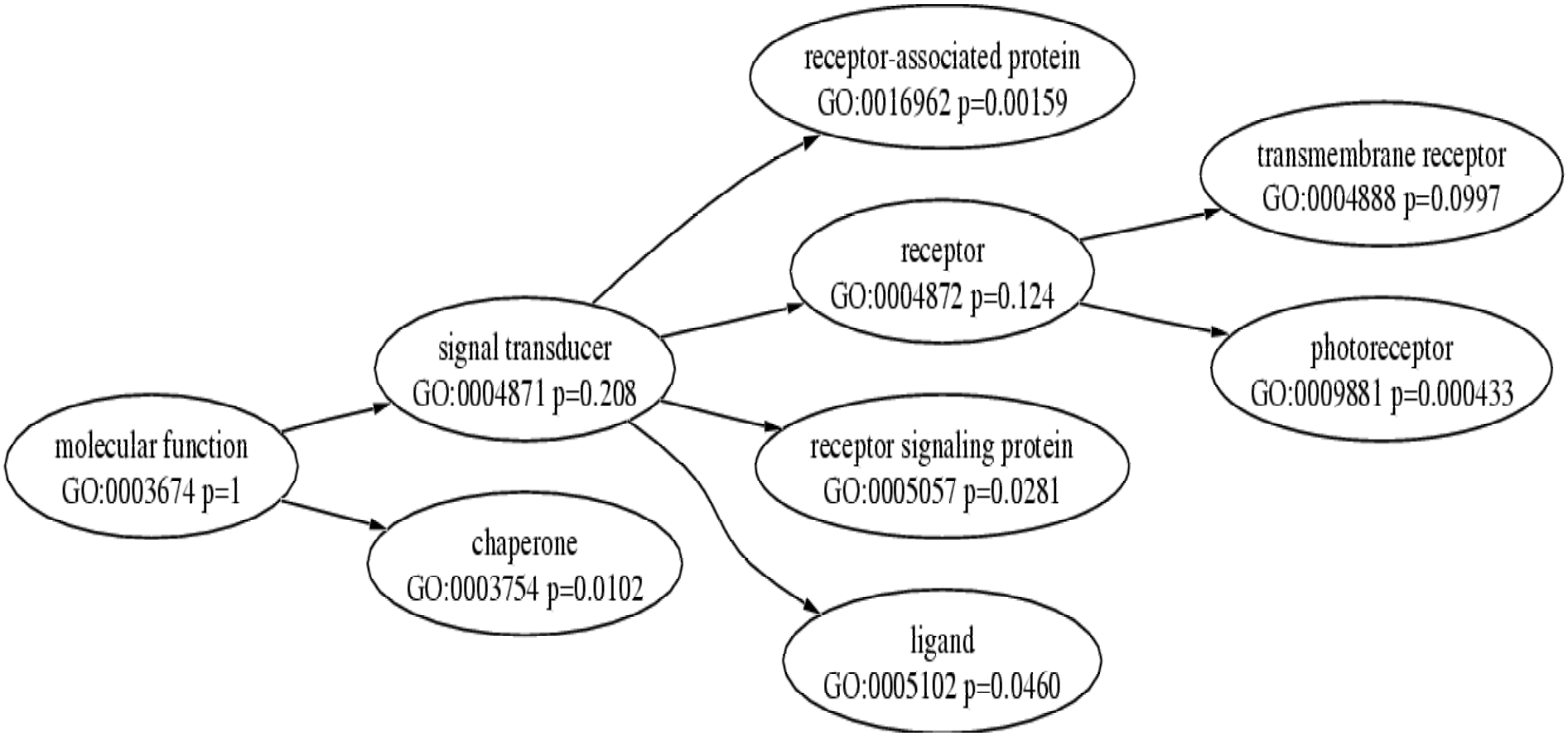
Information Content

- take all LocusLink genes for humans
- find all annotations from human genes to GO terms in the LocusLink corpus
- a term is considered to have high information content if it appears relatively infrequently in the data
- two terms are considered similar if they share annotation at a term that appears infrequently

Information Content

- for each GO term we compile the number of times that that term, or any child, is found in the data source
- one might only want to use the *is-a* children or the *has-a* children
- we then form proportions by dividing each count by the total number of references
- thus, the root node has proportion 1

Adapted from Lord *et al*



Some Uses

- suppose that we have a microarray experiment and we want to choose an appropriate metric for the gene expression data
- one way of choosing a metric is to select the one that provides the best agreement with distances based on MF (or BP or CC)

Some Issues

- for microarray data we generally obtain expression data on a per probe basis
- there is a many to one mapping from probe data to LocusLink ids and it is the LocusLink ids that map to GO terms
- some accounting of this multiplicity should be taken

Issues

- it will be important in some contexts to account for and adjust for the evidence on which an annotation was based
- for example if exploring sequence similarity as it relates to function all ISS based annotations should be excluded

Issues

- at each node we can consider splitting the annotated genes into categories
  - inherited from child i (i = 1, …n_c) and those annotated at the node
  - this might not be a partition
  - do we want to consider an analysis on these data?
  - esp. those annotated at this node

Conclusions

- GO provides a rich resource for both exploratory data analysis and hypothesis testing
- but we lack the tools to adequately exploit this rich data resource
- software tools, such as those provided by the BioConductor Project (and others FatiGO, AmiGO etc) greatly facilitate the use and understanding of these data

Thanks

- Vince Carey
- Sandrine Dudoit
- Sabina Chiaretti
- Jerry Ritz
- Tom LaFramboise
- Raji Balasubraminian
- Jeff Gentry
- Jianhua Zhang
- Elizabeth Whalen
- Beiying Ding
- Denise Scholtens
- Wolfgang Huber