Using GO for Statistical Analyses

R. Gentleman

April 25, 2006

1 Introduction

While there are a number of different definitions of an ontology we will use the notion of a restricted vocabulary as the basis for the discussions here. Ontologies and related concepts are becoming increasingly important tools for organizing and navigating information. Initiatives in biology (our main focus) as well as the semantic web are providing a variety of resources and interesting problems related to ontologies.

One of the major problems facing comprehensive searching and use of available biological information is the lack of a common set of terms and descriptions for basic biological functions, processes and entities. In fact, many genes have a variety of different names. For genes and gene products the Gene Ontology Consortium, or GO, (`www.geneontology.org`) is an initiative that is designed to address this problem. GO provides a restricted vocabulary as well as clear indications of the relationships between terms. Readers are referred to the GO web site for more specific definitions.

GO is clearly a valuable tool for data analysis, however its structure (as a DAG) and the complex nature of the relationships that it represents make appropriate use of this tool challenging. Most data analytic software is not readily able to handle these data and much of the organizational burden falls on the data analyst. We have been developing tools, as part of the Bioconductor Project, that should greatly simplify the use of GO in a data analytic context. We consider some of these different methods in the remainder of this paper. First we outline some general properties and behaviors of GO. That discussion is followed by a description of our example and finally the different methods are applied to the data.

In this paper we will use meta-data packages from the Bioconductor Project to carry out statistical analyses of gene expression data. But would like to note that the potential scope of these applications is much broader and many of the methods described here could be applied to other types of high-throughput data. To provide context we will make use of data from an investigation into acute lymphoblastic leukemia (ALL) reported in Chiaretti et al. (2004).

The outline of the paper is as follows. First we discuss and explain what GO is, how it is structured and how genes are annotated at specific terms within GO. Then we briefly describe the example data. Then we consider three specific analyses. The first uses GO to help understand sets of genes that have been selected according to specific criteria by the data analyst. The second use of GO is to reduce the set of probes to a manageable number before carrying out an analysis, while the third considers some variations on the graph theoretic themes introduced by Zhou et al. (2002) and finally we discuss our findings and make some suggestions for further exploration.

2 The Gene Ontology

An ontology is a restricted structured vocabulary of terms that represent domain knowledge. In a practical sense an ontology specifies a vocabulary that can be used to exchange queries and assertions. A commitment to the use of the ontology is an agreement to use the shared vocabulary in a consistent way. There is no commitment to completeness, the commitment is to coherence and consistency.

The Gene Ontology (GO) consortium produces three independent ontologies for gene products. The three ontologies form the basis for the description of the molecular function, biological process and cellular component of gene products. The relationships between gene products and specific terms in the three ontologies, molecular function, biological process and cellular component, are all many to many. Gene products are physical things such as proteins or RNA. In the remainder of the paper the terms *gene product* and *gene* will be used interchangeably (except in situations where it is obvious that the remark pertains only to one or the other).

The *molecular function* of a gene product is defined to be biochemical activity or action of the gene product. This describes a capability that the gene product has and there is no reference to where or when this activity or usage actually occurs. Examples of terms that fall into this ontology are: "enzyme," "transporter," or "ligand."

The term *biological process* should be interpreted as a biological objective to which the gene product contributes. A biological process is accomplished via one or more ordered assemblies of molecular functions. There is generally some temporal aspect to the process and it will often involve the transformation of some physical thing. Examples of terms in this ontology include "cell growth and maintenance" or "signal transduction". The concept of a pathway is different from that of a biological process. A pathway is more complex and has dependencies and dynamics that are not part of the concept of a biological process. It is not always easy to distinguish between molecular function and biological process. The GO consortium suggests that a process must have more than one distinct step.

A *cellular component* is a component of a cell that is part of some larger object or structure. Examples of cellular components include "chromosome", "nucleus" and "ribosome".

	Number of Terms
BP	10765
CC	1733
MF	7686

Table 1: Number of GO terms per ontology.

Table 1 presents the number of different terms associated with each of the three ontologies. For example, we see that for version 1.12.0 of the *GO* package there are 10765 terms in the biological process (BP) ontology.

2.1 The Graph Structure of GO

The GO ontologies are structured as directed acyclic graphs (DAGs) that represent a network in which each term may be a *child* of one or more *parents*. We use the expressions *GO node*

and *GO term* interchangeably. Child terms are more specific than their parents. The term "transmembrane receptor protein-tyrosine kinase" is child of both "transmembrane receptor" and "protein tyrosine kinase". For each of the three different ontologies there is a root node that has the ontology name associated with it and that node is the most general node for terms in the respective ontology. In the published hierarchy these three nodes all have a common parent node which is labeled as the "gene ontology" node, but this node is dropped from the examples presented here.

The relationship between a child and a parent can be and be either a *is a* relation or a *has a* (*part of*) relation. For example "mitotic chromosome" is a child of "chromosome" and the relationship is an *is a* relation. On the other hand, a "telomere" is a child of "chromosome" with the *has a* relation. Child terms may have more than one parent term and may have a different class of relationship with its different parents.

Each term in the ontology is associated with a unique identifier and the relationships between the GO terms (parent/child) as well as other relevant data are provided by GO. The *GO* package provides six sets of mappings, two for each ontology. For example the two cellular component mappings are provided by `GOCCPARENTS` and `GOCCCHILDREN` and the others are similarly named, with `CC` replaced by `BP` and `MF` respectively. For example the term *transcription factor* is in the molecular function ontology and has the GO label `GO:0003700`. Figure 1 shows the graph that is created by starting with this node and finding all of the less specific terms that are related to it.

In general, given a set of most specific terms of interest we can find the graph that consists of those terms and any less specific terms (parents). We will refer to this graph as the *induced GO graph* for the specific set of child nodes.

We can also obtain the children of a node, for example we can obtain those terms within the MF ontology that are more specific than *transcription factor*.

```
GO:0003705: RNA polymerase II transcription factor activity, enhancer binding
```

P.W.Lord et al. (2003) suggest that the reason that ontologies are of particular interest in computational biology is that they provide a mechanism for representing a communities domain knowledge in a form that is accessible by humans and is amenable to computation. We largely agree with that observation. GO (and potentially other ontologies) provide us with a data resource that is useful and in this paper we consider some of the different ways that this resource can be used.

2.2 Associating Genes With GO Terms

GO itself is strictly the ontology. The mapping of genes to GO terms is carried out separately. And as the ontology is being constantly updated so is this set of mappings. A set of mappings between manufacturer identifiers and GO terms is available in every Bioconductor meta-data package. These packages are available at `http://www.bioconductor.org/data/metaData.html`. Further data, regarding the relationships between GO terms, the specific term names etc. are provided in the *GO* package. The actual mappings are provided by GOA (Camon et al., 2004) and are mappings between GO terms and LocusLink IDs which are modified to account for the multiplicity of mappings between the manufacturer IDs and LocusLink IDs.

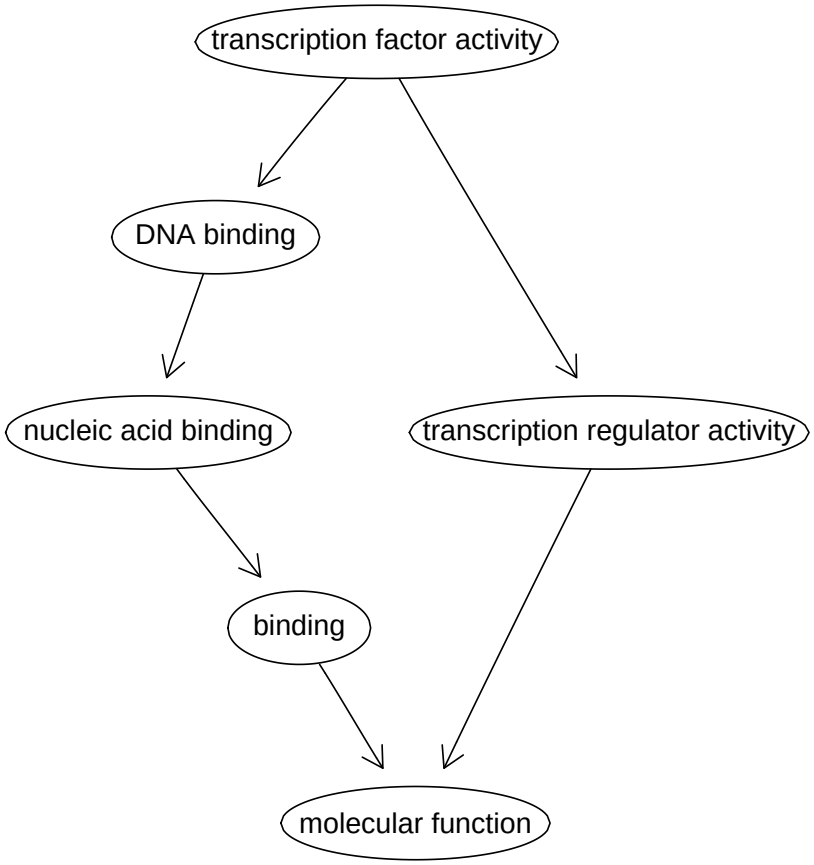

Figure 1: Graph of GO relationships for the term: transcription factor

Mappings from GO terms to specific genes is provided in the hash table (R `environment`) named `GOLOCUSID`, which maps from GO terms to EntrezGene identifiers. These are only the set of most specific mappings. To get all EntrezGene IDs associated with a specific GO term use `GOALLLOCUSID` instead. A histogram of the logarithm of the number of EntrezGene identifiers per GO term is given in Figure 2.

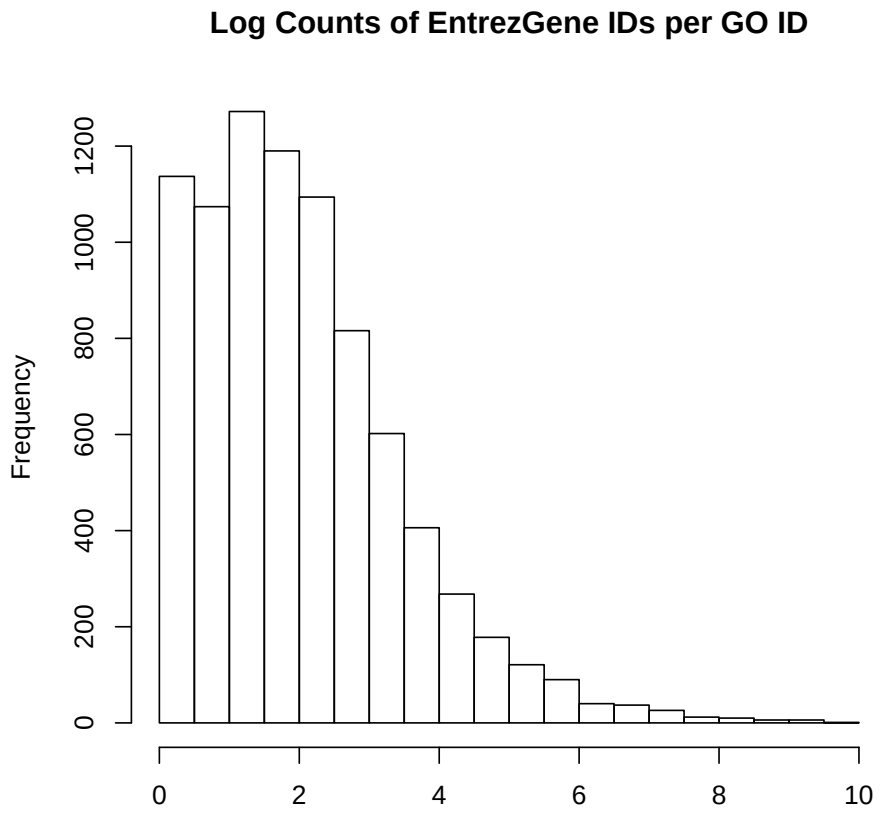

Figure 2: EntrezGene counts per GO term (log scale). .

For any particular microarray there may be many probes that are mapped to the same EntrezGene identifier. In some of the computations it will be important that this multiplicity be accounted for. If, for example, one asks whether a particular GO term is overrepresented in the probes that were selected by some particular procedure then any multiplicities should be dealt with in making that assessment. We report the set of multiplicities for the Affymetrix HGU95Av2 GeneChip below.

Multiplicity	1	2	3	4	5	6	7	8
No. EntrezGene IDs	6886	1556	507	108	21	13	8	6

We can see that the problem can be quite substantial; there are 2219 EntrezGene IDs that have more than one probe set identified with them. Making the correct inference requires that some adjustment be made for the many-to-one mapping from probes to EntrezGene identifiers. These are sometimes called *technical replicates*, but investigation suggests that the are not necessarily probing the same thing, so some caution is warranted.

For any particular microarray analysis you will need to obtain the correct meta-data package from Bioconductor (or create your own equivalent set of mappings). Each Bioconductor meta-data package follows a specific naming convention. A short name for the chip is used as a prefix and a variety of suffixes are used. So for the *hgu95av2* package the data objects that have GO mappings are named `hgu95av2GO`, `hgu95av2GO2PROBE` and `hgu95av2GO2ALLPROBES`.

There are 925 Affymetrix probes that are annotated specifically at `GO:0003700` and these correspond to 568 distinct genes.

We can also use the data in `hgu95aGO2ALLPROBES` to find all the probe sets that are annotated at the term `GO:0003700`, this is the union of those specifically annotated at that term together with those annotated at any of the child nodes. For the term `GO:0003700` there are 952 probes.

3 An Example

To demonstrate some of the tools that are included in the *GOstats* package we consider expression data from 79 samples from patients with acute lymphoblastic leukemia (ALL) that were investigated using HGU95AV2 Affymetrix GeneChip arrays (Chiaretti et al., 2004). The data were normalized using quantile normalization and expression estimates were computed using RMA (Irizarry et al., 2003). Of particular interest is the comparison of 37 samples from patients with the BCR/ABL fusion gene resulting from a chromosomal translocation (9;22) with the 42 samples from the NEG group.

To reduce the set of genes for consideration we applied two different sets of filters. Gene filtering is considered in more detail in von Heydebreck et al. (2004) and the interested reader is referred there. A non-specific filter was used to remove genes that showed little or no change in expression level across experiments. The resulting data set had 2391 probes remaining. To select genes whose expression values were associated with the phenotypes of interest (BCR/ABL and NEG) we used the `mt.maxT` function from the *multtest* package which computes a permutation based t-test for comparing two groups.

After adjustment for multiple testing there were only 19 probes (which correspond to 16 genes) with an adjusted p-value below 0.05. Using those genes we obtain the set of most-specific GO terms in the MF ontology that they are annotated at and compute the induced GO graph which is rendered in Figure 3. No labels have been added to the nodes in this plot since there is not sufficient room to provide informative ones. The most specific terms are at the top of the graph and that arrows go from more specific nodes to less specific ones. The node in the bottom center is the MF node. Clearly some sort of interactivity (e.g. tooltips) would be beneficial. We will return to this plot in the next section and use it to provide a more detailed view of the data.

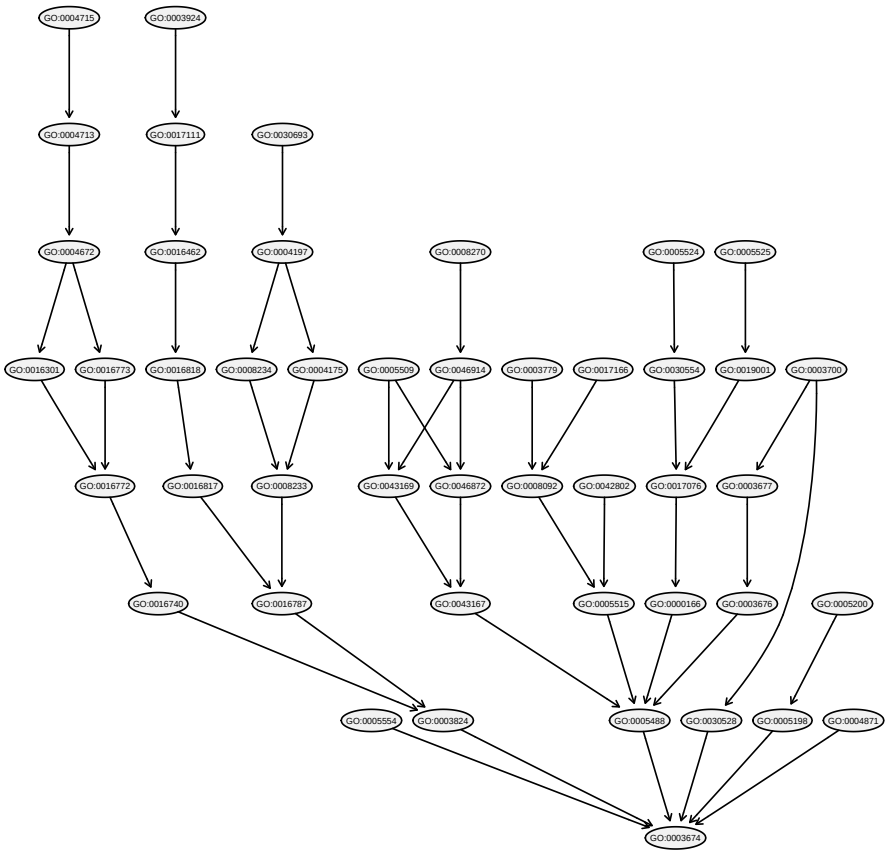

Figure 3: The induced GO graph for the selected genes.

4 Statistical Analyses

4.1 Finding Interesting GO terms

If genes have been partitioned into distinct sets, say by finding those with small p-values (as was done above) or by some form of clustering, then one of the questions that arises is whether genes that comprise a cluster have a common function, process or location in the cell. A second, related application of this idea is to provide meaning to a list or set of genes that were selected according to some criteria. For example, in our microarray experiment we selected genes that were differentially expressed between the BCR/ABL group and the NEG group. We might then wonder whether these genes have a common function, are involved in common processes, or perhaps are co-located in some region of the cell.

The actual test employed is quite straight forward. Given a set of genes and one of the three ontologies we first find the set of all unique GO terms within the ontology that are associated with one or more of the genes of interest (i.e. the induced GO graph). Next, for each term we determine how many of the interesting genes are annotated at that node and how many genes that were assayed (i.e. have probes on the chip that represent that gene) are annotated at the node. Here we must work in terms of the unique EntrezGene identifiers and not the manufacturers identifiers because of the multiple mapping issues raised in Section 2.2.

We can ask if there are more interesting genes at the node than one might expect by chance. If that is true, then that term can be thought of as being overrepresented in the data. This question can be answered using a Hypergeometric distribution. Suppose that there are N total genes annotated for the ontology of interest and that our list of interesting genes contains m distinct genes. Then we can imagine an urn with N balls in it and $N - m$ are black while m are white. If we draw k balls from the urn, where k is the number of genes annotated at a node, we are asking whether the number of white balls in that drawn sample is unusually large. Suppose that there are q white balls (interesting genes) in the drawn sample, we then ask what is the probability that $X \geq q$ where X is a Hypergeometric random variable with parameters as we have described. This probability constitutes a p-value since it is the probability of seeing something as extreme or more extreme than what was observed. This functionality is provided in the function `GOHyperG` available in the *GOstats* package.

There are some issues that arise in the interpretation of these p-values. First, we note that often very many hypotheses will have been tested and that some form of p-value correction will be needed. However, there is no simple or straightforward way to do that. The different hypotheses are not independent by virtue of the way that GO is structured and even with this difficulty addressed it is further the case that we are most likely interested in patterns of p-values that correspond to structure in GO rather than single p-values that exceed some threshold. These and other issues are considered in more detail in Gentleman (2003).

A second issue that arises is the fact that nodes with few genes annotated at them will typically have small p-values. This phenomenon occurs due to the way that we selected nodes for evaluation and the structure of GO. Recall that for each gene we have the most specific set of nodes that it is associated with and it is also annotated at all nodes that are less specific. Therefore most genes are annotated out quite far into the leaves of the GO graph and hence at nodes that have relatively few other genes annotated there. Calculation of the Hypergeometric p-values for those nodes results in very small p-values and often they are not too interesting. Nodes that are interesting are typically those with a reasonable number of genes (10 or more)

and small p-values.

In Figure 4, we reproduce the plot from Figure 3 except that we have now colored the nodes according to the p-value obtained from the Hypergeometric test described above. The nodes in Figure 4 are colored either red or blue depending on whether the unadjusted Hypergeometric p-value was less than 0.10 or not (for those viewing this document in black and white the nodes should be dark and light grey, respectively). The GO terms for the terms colored red are printed below. The relevant biology suggests that these are quite reasonable. We note that while the smallest p-values are associated with nodes that have few genes annotated at them there are some nodes with a reasonable number of genes annotated at them (counts) and small p-values.

	GO ID	Term	p-value	No. of Genes
1	GO:0017166	vinculin binding	0.005	5
2	GO:0003924	GTPase activity	0.005	108
3	GO:0042802	identical protein...	0.006	118
4	GO:0004713	protein-tyrosine ...	0.011	154
5	GO:0030693	caspase activity	0.013	12
6	GO:0017076	purine nucleotide...	0.015	1029
7	GO:0005525	GTP binding	0.015	185
8	GO:0004715	non-membrane span...	0.016	15
9	GO:0019001	guanyl nucleotide...	0.016	191
10	GO:0000166	nucleotide bindin...	0.025	1195
11	GO:0017111	nucleoside-tripho...	0.047	341
12	GO:0016462	pyrophosphatase a...	0.051	355
13	GO:0016818	hydrolase activit...	0.051	357
14	GO:0016817	hydrolase activit...	0.051	358
15	GO:0005554	molecular functio...	0.061	393
16	GO:0004672	protein kinase ac...	0.062	397
17	GO:0004197	cysteine-type end...	0.066	64
18	GO:0005198	structural molecu...	0.074	438
19	GO:0008234	cysteine-type pep...	0.076	75
20	GO:0005200	structural consti...	0.081	80
21	GO:0016773	phosphotransferas...	0.082	464

Table 2: GO terms, p-values and counts.

4.2 Selecting Genes according to GO term

GO can also be used as a method of data reduction. Here one might carry out an analysis focusing on a particular subset of genes, say those associated with the GO term `transcription factor`. Carrying out such an analysis is very straight forward. First one selects the GO term or terms that are of interest and then collects the set of genes that were assayed and that are annotated at that term. Generally one would consider all genes annotated at the term either directly or by inheritance. Given this set of genes the data are then reduced only to them and

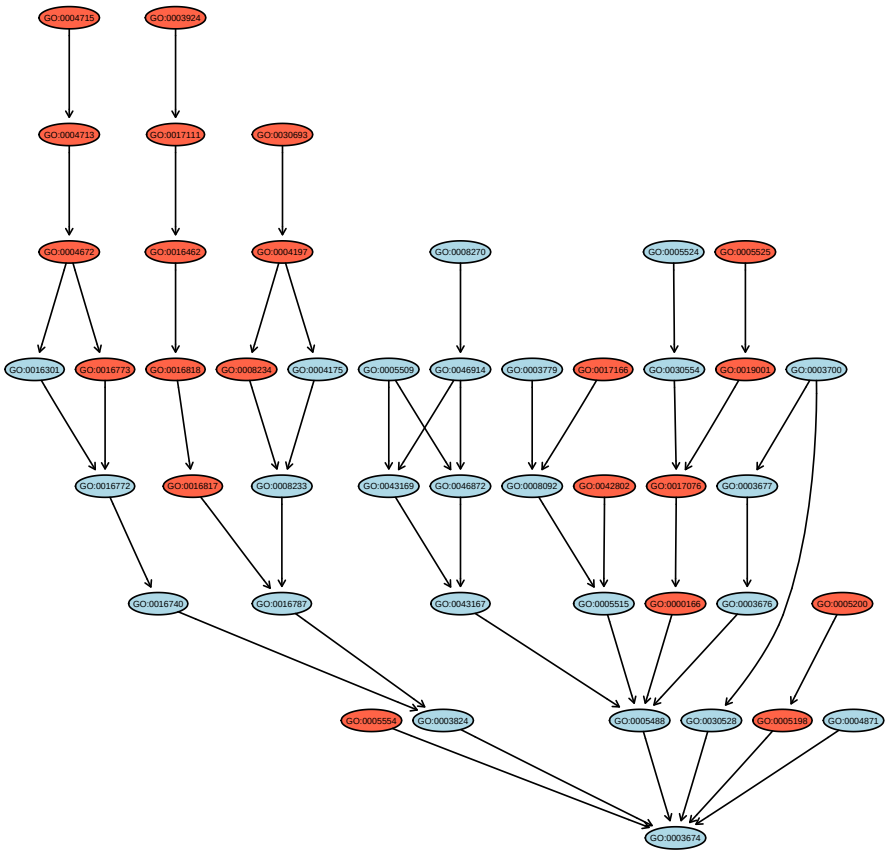

Figure 4: The induced GO graph colored according to unadjusted Hypergeometric p-values

the usual machine learning or visualization procedures can be applied.

It is often the case that these more directed approaches can be more successful (especially if determined *a priori*) than the more omnibus approach of considering all assayed genes simultaneously. The p-value corrections are much less drastic and smaller, but important, effects can be detected more often.

Many of the effects due the BCR/ABL translocation are mediated by tyrosine kinase activity. It will therefore be of interest to examine genes that are known to have tyrosine kinase activity. We examine the set of GO terms and identify the term, `GO:0004713` from the *molecular function* portion of the GO hierarchy as referring to `protein-tyrosine kinase activity`. We can then obtain all Affymetrix probes that are annotated at that node, either directly or by inheritance, using the following command.

```
> tykin <- unique(lookUp("GO:0004713", "hgu95av2", "GO2ALLPROBES"))
```

We see that 268 probe sets are annotated at this particular term. Of these only 42 were selected by the non-specific filtering step. We focus our attention on these probes and carry out a permutation t-test analysis.

In this analysis of the GO–filtered data, 7 probe sets have FWER–adjusted p–values less than 0.1. They are printed below, together with the adjusted p-values from an analysis that used all probes that passed our non-specific filter and hence involved 2391 genes.

```
[1] "GO analysis"

   1635_at  1636_g_at   39730_at 40480_s_at  2039_s_at   36643_at  2057_g_at
    0.0001     0.0001     0.0001     0.0001     0.0003     0.0254     0.0905

[1] "All Genes"

   1635_at  1636_g_at   39730_at 40480_s_at  2039_s_at   36643_at  2057_g_at
     0.001      0.001      0.001      0.001      0.018      0.473      0.823
```

Due to the reduced number of tests in the analysis focused on tyrosine kinases, we are left with more significant genes after correcting for multiple testing. For instance, the probe set `36643_at`, which corresponds to the gene DDR1, was not significant in the unfocused analysis, but would be if instead the investigation was oriented towards studying tyrosine kinases *a priori*.

4.3 Using Shortest Paths

Zhou et al. (2002) consider some interesting applications of GO in conjunction with microarray expression data. In this section we consider a related idea and apply it to the ALL data. At their most basic level the ideas of Zhou et al. (2002) consist of forming a graph between genes (which are the nodes) based on some relevant distance. This distance might be correlation distance or it could be any other relevant distance. Then all edges in the graph that correspond to distances that are larger than some threshold are removed. Next, genes are grouped according to some specific categorization (they used GO biological process terms) and the shortest paths

(using Dijkstra's algorithm) between all pairs of nodes are computed. Those shortest paths can then be examined to see whether they provide information of relevance.

Among the points that they make is that it will often be reasonable (and possibly essential) to partition the genes being analysed according to their cellular location. They provide the two following examples in order to make their point. For gene expression data consider an analysis that explores the biological process of *protein biosynthesis*. This process occurs in both the mitochondria and the cytoplasm and there is little reason to assume that there is any sort of relationship between the two. Other concepts, such as *membrane transport* are very distinct processes in the three different compartments and are unlikely to be related at the transcriptional level, and hence should be modeled separately when considering microarray data.

In the ALL experiment we are most interested in comparing patients that have the BCR/ABL defect to those that have no measured cytogenetic abnormalities. Our adaptation of the shortest path technology is as follows. We use the output of the first filtering step described previously – that is we select genes that show some level of expression and some variation in expression across samples. We then separate the data into two sets (BCR/ABL and NEG) and within each group we define the distance between two genes, u and v, as one minus the absolute value of the Pearson correlation. Other approaches that could be used are Spearman's correlation, that use by Zhou et al. (2002) or some other robust correlation estimate. We let $C_{u,v}$ denote the absolute value of the *correlation* between probes u and v and used an edge weight of $d(u,v) = (1 - C_{u,v})^k$ with $k = 1$ and $\tau = 0.6$ as the cutoff for correlations. If $C_{u,v} < \tau$ then no edge exists in our group. Zhou et. al used $k = 6$ in their analysis and some experimentation may be warranted.

Our interest in this particular example is on transcription factors. Hence we use the GO term `GO:0003700` which maps to the molecular function `transcription factor activity` to identify all genes with transcription factor activity. We used only genes for which this was a most specific annotation and obtained 952 mappings and 579 unique EntrezGene ids. Of these 171 were among those probes selected for our analysis. Of these we found that there were 3 with duplicate entries (*technical replicates*. A visual inspection (not reported) suggested that the correlation between these duplicate probes was quite high and so only one of each was used in the subsequent analysis. This left us with 162 distinct transcription factors for our study.

We note that while Zhou et al. (2002) divided the probes/genes according to the cellular component they were annotated at, we did not do the same. In our case we are attempting to understand transcriptional networks, not gene function. It does not seem likely that transcriptional regulation is heavily dependent on the cellular location of the gene products.

For every pair of transcription factors we compute two quantities. The shortest path between each pair for each of the different conditions. For example in our ALL example we compute the shortest paths between all transcription factors using a graph based only on data from those with BCR/ABL and secondly the same set of values based only on data from those without any noticeable genomic defects. Then each pair the distances are compared (plotted) and those pairs for which the distance has changed the most identified and further explored.

This approach is quite different in substance than that originally proposed in Zhou et al. (2002). Their interests were centered on providing annotation for genes whose biological annotation is still incomplete, or potentially in error. Whereas our investigation is aimed at study transcription regulatory networks.

We first consider those transcription factors that are not connected to the others in their respective graphs. There are three sets, those that are not connected in either graph, those that are not connected in one of the two graphs but not in the other. They are reported in Table 3.

	Affymetrix ID	Symbol	Which graph
1	34730_g_at	TRO	Both
2	1106_s_at	TRA@	NEG only
3	34850_at	UBE2E3	NEG only
4	1185_at	IL3RA	BCR/ABL only
5	32186_at	SLC7A5	BCR/ABL only
6	33641_g_at	AIF1	BCR/ABL only

Table 3: Genes not connected in the different graphs

We now consider the finite pairwise distances. First a simple t-test can be carried out to see if there is any difference between the distances in on graph versus the other. We took each pairwise distance in the NEG graph and subtracted from it the same pairwise distance computed on the BCR/ABL graph. The t-test is for whether the mean is zero and the test statistic was 0.078 with an extremely small p-value. So we see that distances in the NEG graph seem to be longer than those in the BCR/ABL. Further evidence of this difference comes from the observation that the proportion of values that were larger in the NEG graph was 0.566.

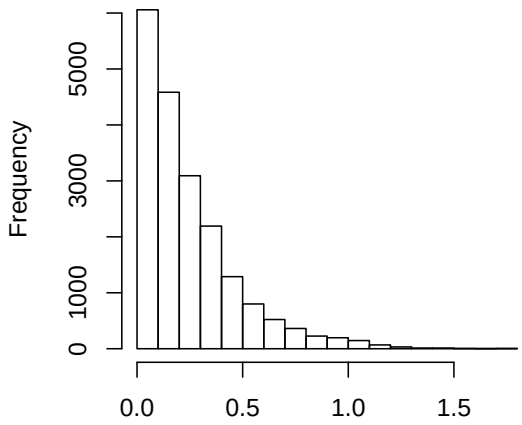

Figure 5: Histogram of the absolute value of the difference between the distance computed in the NEG graph and that on the BCR/ABL graph.

In Figure 5 a histogram of the absolute value of the pairwise differences is plotted. We see a number of interesting features. For example the large number of differences that are small in absolute value as well as a rather long right hand tail. We will focus our attention on those differences that are large in absolute value. We chose a value of 2.5 as our cut-off and found that there were 28 differences that were larger than 2.5. These corresponded to 13 distinct genes.

While all may be interesting and a particular investigator may want to expend considerable effort in study transcription factors that are of particular interest we will center our analysis on the set of genes that appear most frequently in this list. The multiplicities are reported below.

```
rowSym
C20orf14     GFI1   LILRA2   PFAAP5     PIM2     SARS  SULT1A3     TPM4
       1        1        1        1        1        1        1        1
 Unknown LOC57228      MYC      MPO  GADD45A
       1        3        3        4        9
```

And we can see that GADD45A has much higher counts than the rest. Additionally MYC, MPO and L0C57228 each have several. Very little is currently known about L0C57228 and so we concentrate our examination on the other three genes. The evidence here suggests that, perhaps, the expression patterns of these three different transcription factors are substantially different in the two phenotypes we are studying.

For each of the three transcription factors we can compute the average distance, separately within each graph, to all the other selected genes. We find that the results are quite consistent and that in all cases the path length is much shorter in the BCR/ABL group than it is in the NEG group. For MYC the means were 2.6 for NEG and 1 for BCR/ABL, for MPO they were 2.4 for NEG and 0.9 for BCR/ABL and for GADD45A the means were 2.8 for NEG and 1.2 for BCR/ABL. It is rather interesting to observe that amongst the pairwise distances that have changed the most are those between these three specific genes.

Specific paths between transcription factors can also be examined. Recall that we compute out distance between two transcription factors based on the shortest path length between them in each of the two graphs. In our examples we focus on MYC and the distances between it and MPO and GADD45A.

We print out the different shortest paths for genes connecting MYC to both MPO and GADD45A for each of the two phenotypes, respectively (first the paths for BCR/ABL, then for the NEG samples). The MYC to MPO results are:

```
BCR/ABL
```

```
[1] "MYC<->NBEAL2<->HMG20B<->MPO"
```

```
NEG
```

```
[1] "MYC<->CDC25B<->TRAP1<->POLR2H<->MSH2<->EMP3<->S100A4<->LGALS1<->MPO"
```

If we then make use of the results in Figure 6 we see that there are positive correlations between MYC and KIAA0540 and as well between KIAA0540 and HMG20B, but that for

HMG20B and MPO the correlation is negative. Positive correlations are suggestive of shared transcriptional activity while negative correlations are suggestive of transcriptional inhibition. But the reader is cautioned that these are stable values averaged over many cells and are not in any sense a time-course so direct relationships are very difficult to establish.

The results comparing MYC to GADD45A are:

```
BCR/ABL

[1] "MYC<->UBE2A<->BAZ1A<->CD53<->GADD45A"

NEG

[1] "MYC<->CDC25B<->TRAP1<->POLR2H<->MSH2<->MSH6<->HCK<->SH3BP1<->PVRL2<->GADD45A"
```

We do not have space to present the other pairwise scatterplots here but readers that are making use of the compendium version of this paper can easily explore those different plots on their own.

We notice that the path lengths for the NEG samples are longer (involve more genes) than those for the BCR/ABL samples. We might also want to ask whether the distances are also larger (that is that the correlations are smaller). To do this we need to obtain the edge weights from the respective graphs and compare them. We found that there appeared to be no difference (all averaged around a distance of about 0.3 but the number of edges is quite small and one might expect to see systematic differences if a larger study were undertaken.

We can check our results, at least to some extent, by examining pairwise scatterplots of the gene expressions. In Figure 6 the genes on the path from MYC to MPO are plotted. We see quite strong correlations along the diagonal and note that HMG20B and MPO have a negative correlation.

Finally, we finish our examination of these data by considering some of the specific paths between the different transcription factors. Comparison of paths in different graphs is problematic since the same nodes need not be connected in the two different graphs (and that is the case here). We consider two specific paths - the ones between MYC and MPO in each of the two phenotypes.

We see, in Figures 7 and 8, the actual shortest paths between the genes MYC and MPO. The two end points have been colored red, genes along the path are colored blue. While the two graphs have the same set of nodes, their layout is quite different. This is because they have different sets of edges, with different weights. We will be developing layout tools that allow the user to fix the layout of the nodes and to subsequently add edges, recolor nodes etc.

5 Discussion

GO and the mappings from genes to specific terms in each of the three ontologies provide a number of important and unique data analytic opportunities. In this paper we have considered three separate applications of these resources to the problem of analysing gene expression data and in all cases the GO related data have provided new and important insights into the data.

Using GO mappings to select certain terms for further study and reference has the possibility of providing meaning to sets of genes that have been selected according to different criteria.

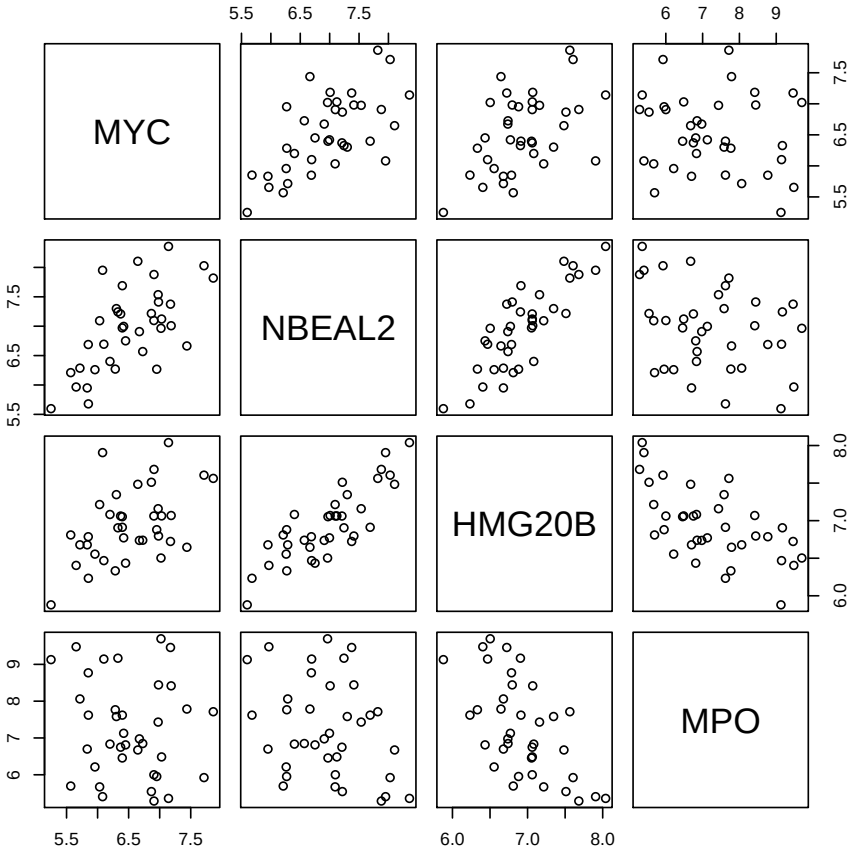

Figure 6: Pairwise scatterplots of gene expression for those genes on the shortest path between MYC and MPO from patients with the BCR/ABL translocation. .

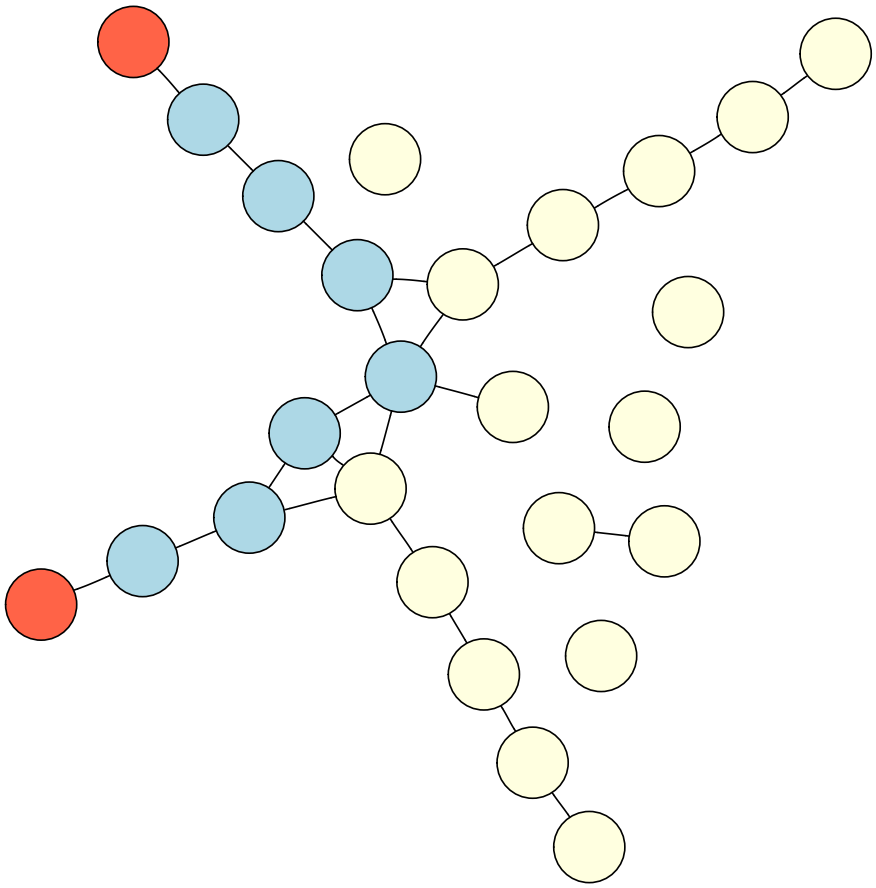

Figure 7: Shortest path between MYC and MPO in the NEG samples.

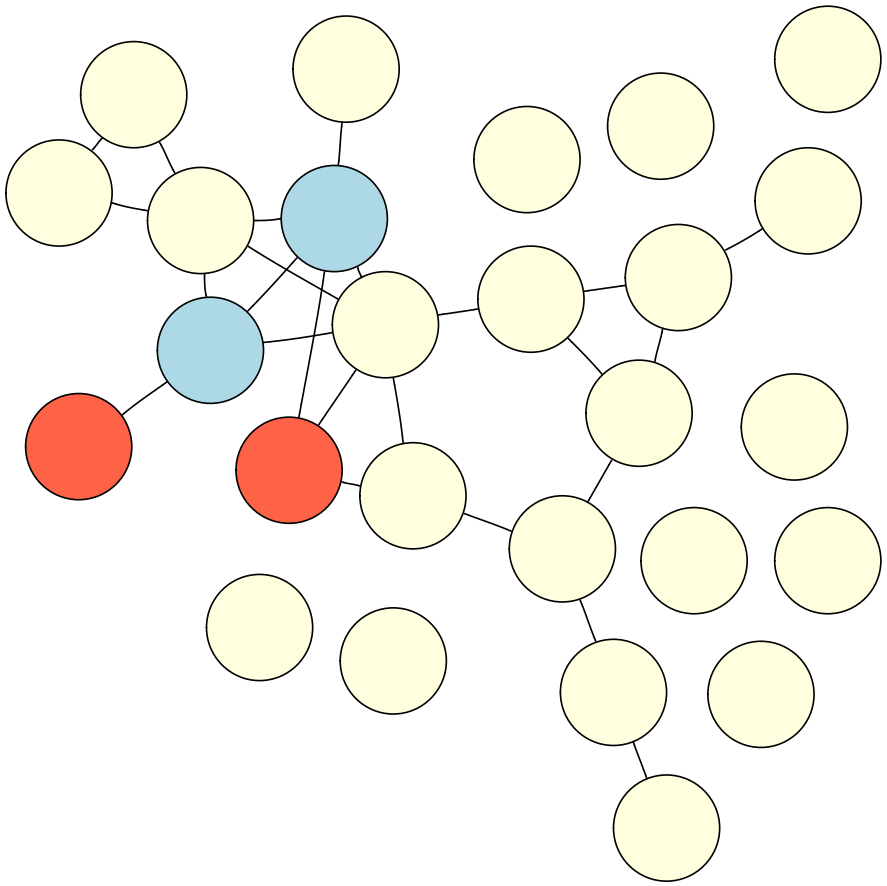

Figure 8: Shortest path between MYC and MPO in the BCR/ABL samples.

An equally important application is to use GOA mappings to reduce the set of genes under consideration. As the capacity of microarrays increase it is important that we begin developing tools and strategies that directly address specific questions of interest. P-value correction methods are at best a band-aid and do not represent an approach that has long term viability (von Heydebreck et al., 2004).

In our final example we adapt the method proposed by Zhou et al. (2002) to a different problem, one where we consider only transcription factors and where we are interested in understanding their interrelationships. The results are promising and in our example reflect a fundamental difference between those with the BCR/ABL translocation and those patients with no observed genetic abnormalities. Ideally these, and other observations will lead to better understanding of transcriptional regulation and from that to better understanding modalities of efficacy for drug treatments.

Perhaps more important than the statistical presentation is the fact that we have also provided software implementations for all tools described and discussed in this paper. They are available from the Bioconductor Project in the form of the *GOstats* package. *GOstats* makes substantial use of software infrastructure from the Bioconductor Project in carrying out this analysis. In particular the *graph*, *Rgraphviz* and *RBGL*, together with the different meta-data packages.

Finally, this document itself represents an approach to reproducible research in the sense discussed by Gentleman and Temple Lang (2003) and it can be reproduced on any users machine equipped with R and the appropriate set of R packages. We encourage the interested reader to avail themselves of the opportunity to explore the data and the methods in more detail on their own computer.

Thanks

I would like to thank Vincent Carey for many helpful discussions about these, and very many other topics. I would like to thank Drs. J. Ritz and S. Chiaretti of the DFCI for making their data available and for helping me to understand how it relates to ALL. I would like to thank J. Zhang and J. Gentry for a great deal of assistance in preparing the data and writing software in support of this research.

References

E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, D. Binns J. Maslen, N. Harte, R. Lopez, and R. Apweiler. The Gene Ontology annotation (GOA) database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 32:D262–D266, 2004.

S. Chiaretti, X Li, R Gentleman, A Vitale, M. Vignetti, F. Mandelli, J. Ritz, , and R. Foa. Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103:2771–2778, 2004.

R. Gentleman. Hypothesis testing and GO. 2003.

R. Gentleman and D. Temple Lang. Statistical analyses and reproducible research. 2003.

R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003.

P.W.Lord, R.D. Stevens, A. Brass, and C.A.Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–83, 2003.

A. von Heydebreck, W. Huber, and R. Gentleman. Differential expression with the bioconductor project. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. John Wiley and Sons, 2004.

X. Zhou, M-C J. Kao, and W. H. Wong. Transitive functional annotation by shortest-path analysis of gene expression data. *PNAS*, 99:12783–12788, 2002.