Basic GO Usage

R. Gentleman

May 18, 2005

Introduction

In this vignette we describe some of the basic characteristics of the data available from the Gene Ontology (GO), (The Gene Ontology Consortium, 2000) and how these data have been incorporated into Bioconductor. We assume that readers are familiar with the basic DAG structure of GO and with the mappings of genes to GO terms that are provide by GOA (Camon et al., 2004). We consider these basic structures and properties quite briefly.

GO, itself, is a structured terminology. The ontology describes genes and gene products and is divided into three separate ontologies. One for cellular component (CC), one for molecular function (MF) and one for biological process (BP). We maintain those same distinctions were appropriate. The relationship between terms is a parent-child one, where the parents of any term are less specific than the child. The mapping in either direction can be one to many (so a child may have many parents and a parent may have many children). There is a single root node for all ontologies as well as separate root nodes for each of the three ontologies named above. These terms are structured as a directed acyclic graph (or a DAG).

GO itself is only the collection of terms; the descriptions of genes, gene products, what they do, where they do it and so on. But there is no direct association of genes to terms. The assignment of genes to terms is carried out by others, in particular the GOA project (Camon et al., 2004). It is this assignment that makes GO useful for data analysis and hence it is the combined relationship between the structure of the terms and the assignment of genes to terms that is the concern of the *GO* package.

The basis for child-parent relationships in GO can be either an *is-a* relationship, where the child term is a more specific version of the parent. Or, it can be a *has-a*, or *part-of* relationship where the child is a part of the parent. For example a telomere is a part-of a chromosome.

Genes are assigned to terms on the basis of their LocusLink ID. For this reason we make most of our mappings and functions work for LocusLink identifiers. Users of specific chips, or data with other gene identifiers should first map their identifiers to LocusLink before using *GOstats*.

A gene is mapped only to the most specific terms that are applicable to it (in each ontology). Then, all less specific terms are also applicable and they are easily obtained by traversing the set of parent relationships down to the root node. In practice many of these mappings are precomputed and easily obtained from the different hash tables provided by the *GO* package.

Mapping of a gene to a term can be based on many different things. GO and GOA provide an extensive set of evidence codes, some of which are given in Table 1, but readers are referred to the GO web site and the documentation for the *GO* package for a more comprehensive listing. Clearly for some investigations one will want to exclude genes that were mapped according to some of the evidence codes.

IMP	inferred from mutant phenotype
IGI	inferred from genetic interaction
IPI	inferred from physical interaction
ISS	inferred from sequence similarity
IDA	inferred from direct assay
IEP	inferred from expression pattern
IEA	inferred from electronic annotation
TAS	traceable author statement
NAS	non-traceable author statement
ND	no biological data available
IC	inferred by curator

Table 1: GO Evidence Codes

In some sense TAS is probably the most reliable of the mappings. IEA is a weak association and is based on electronic information, no human curator has examined or confirmed this association. As we shall see later, IEA is also the most common evidence code.

The sets of mappings of interest are roughly divided into three parts. First there is the basic description of the terms etc., these are provided in the `GOTERMS` hash table. Each element of this hash table is named using its GO identifier (these are all of the form `GO:` followed by seven digits). Each element is an instance of the `GOTerms` class. A complete description of this class can be obtained from the appropriate manual page (use `class?GOTerms`). From these data we can find the text string describing the term, which ontology it is in as well as some other basic information.

There are also a set of hash tables that contain the information about parents and children. They are provided as hash tables (the `XX` in the names below should be substituted for one of `BP`, `MF`, or `CC`.

- `GOXXPARENTS`: the parents of the term
- `GOXXANCESTOR`: the parents, and all their parents and so on.

- GOXXCHILDREN: the children of the term
- GOXXOFFSPRING: the children, their children and so on out to the leaves of the GO graph.

For the GOXXPARENTS mappings (only) information about the nature of the relationship is included.

```
> GOTERM$"GO:0003700"
```

```
GOID = GO:0003700
Term = transcription factor activity
Secondary = GO:0000130
Definition = Any protein required to initiate or regulate
     transcription; includes both gene regulatory proteins as well as
     the general transcription factors.
Ontology = MF
```

```
> GOMFPARENTS$"GO:0003700"
```

```
         isa          isa
"GO:0003677" "GO:0030528"
```

```
> GOMFCHILDREN$"GO:0003700"
```

```
[1] "GO:0003705"
```

Here we see that the term GO:0003700 has two parents, that the relationships are is-a and that it has one child. One can then follow this chains of relationships or use the ANCESTOR and OFFSPRING hash tables to get more information.

The mappings of genes to GO terms is contained in the GOLOCUSID2GO hash table. This contains mappings from a LocusLink ID to the most specific applicable GO terms. Each entry in the GOLOCUSID2GO hash table is a list of lists. The actual entries are named lists, with three names:

- GOID: the GO identifier
- Evidence: the evidence code for the assignment
- Ontology: the ontology the GO identifier belongs to (one of BP, MF, or CC).

We note that this same structure is used for the GO mappings in all chip-based metadata packages such as *hgu95av2*.

Some genes are mapped to a GO identifier based on two or more evidence codes. Currently these appear as separate entries. So you may want to remove duplicate entries if you are not interested in evidence codes. However, as more sophisticated use is made

of these data it will be important to be able to separate out mappings according to specific evidence codes.

In this next example we consider the gene with LocusLink ID 25310. FIXME: what is its symbol etc.

```
> ll1 = GOLOCUSID2GO[["4121"]]
> length(ll1)

[1] 11

> sapply(ll1, function(x) x$Ontology)

GO:0005975 GO:0006487 GO:0005624 GO:0005783 GO:0005794 GO:0016020 GO:0016021
      "BP"       "BP"       "CC"       "CC"       "CC"       "CC"       "CC"
GO:0004571 GO:0005509 GO:0015923 GO:0016798
      "MF"       "MF"       "MF"       "MF"
```

We see that there are 11 different mappings. We can get only those mappings for the BP ontology by using getOntology. We can get the evidence codes using getEvidence and we can drop those codes we do not wish to use by using dropECode.

```
> getOntology(ll1, "BP")

[1] "GO:0005975" "GO:0006487"

> getEvidence(ll1)

GO:0005975 GO:0006487 GO:0005624 GO:0005783 GO:0005794 GO:0016020 GO:0016021
     "IEA"      "IEA"      "TAS"      "TAS"      "IEA"      "IEA"      "IEA"
GO:0004571 GO:0005509 GO:0015923 GO:0016798
     "TAS"      "TAS"      "TAS"      "IEA"

> zz = dropECode(ll1)
> getEvidence(zz)

GO:0005624 GO:0005783 GO:0004571 GO:0005509 GO:0015923
     "TAS"      "TAS"      "TAS"      "TAS"      "TAS"
```

A Basic Description of GO

We now characterize GO and some of its properties. First we list some of the specific GO IDs that might be of interest (please feel free to propose even more).

- GO:0003673 is the GO root.

- GO:0003674 is the MF root.
- GO:0005575 is the CC root.
- GO:0008150 is the BP root.
- GO:0000004 is biological process unknown
- GO:0005554 is molecular function unknown
- GO:0008372 is cellular component unknown

We can find out how many terms are in each of the different ontologies by:

```
> zz = eapply(GOTERM, function(x) x@Ontology)
> table(unlist(zz))

  BP   CC   MF
9529 1536 7220
```

Or we can ask about the number of is-a and partof relationships in each of the three different ontologies.

```
> BPisa = eapply(GOBPPARENTS, function(x) names(x))
> table(unlist(BPisa))

    isa part_of
  13600    2400

> MFisa = eapply(GOMFPARENTS, function(x) names(x))
> table(unlist(MFisa))

    isa part_of
   8519       2

> CCisa = eapply(GOCCPARENTS, function(x) names(x))
> table(unlist(CCisa))

    isa part_of
   1332    1026
```

Finally, we can look at the frequency with which the different evidence codes are used to map LocusLink Identifiers to GO terms.

```
> LLec = eapply(GOLOCUSID2GO, getEvidence)
> table(unlist(LLec))

    IC    IDA    IEA    IEP    IGI    IMP    IPI    ISS    NAS     ND     NR
   567  23800 165351   1004   2239  17880   4899  43010  16136  22734   1908
   TAS
 36163
```

Working with GO

Finding terms that have specific character strings in them is easily accomplished using `grep`. In the next example we first convert the data from `GOTERM` into a character vector to make it easier to do multiple searches.

```
> goterms = unlist(eapply(GOTERM, function(x) x@Term))
> whmf = grep("molecular_function", goterms)
```

So we see that there are 2 terms with the string "molecular_function" in them in the ontology. They can be accessed by subsetting the `goterms` object.

```
> goterms[whmf]
```

```
                  GO:0005554                   GO:0003674
"molecular_function unknown"         "molecular_function"
```

Working with chip specific meta-data

In some cases users will want to restrict their attention to the set of terms etc that map to genes that were assayed in the experiments that they are working with. To do this you should first get the appropriate chip specific meta-data file. Here we demonstrate some of the examples on the Affymetrix HGu95av2 chips and so use the package *hgu95av2*. Each of these packages has a data environment whose name is the name of the package with a `GO` suffix, so in this case `hgu95av2GO`. And these data environments have exactly the same internal structure as the `GOLOCUSID2GO` environment except that the keys are the manufacturers identifiers and not LocusLink identifiers. Note that if there are many manufacturer ids that map to the same LocusLink identifier then these will be duplicate entries (with different keys).

The software tools for working with these packages is then exactly the same as we described above for working with `GOLOCUSID2GO`.

We can get all the `MF` terms for our Affymetrix data.

```
> affyGO = eapply(hgu95av2GO, getOntology)
> table(sapply(affyGO, length))
```

```
   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14
2205 2954 2956 2179 1284  544  290  140   51    8    4    6    2    1    1
```

How many of these probes have multiple GO terms associated with them? What do we do if we want to compare two genes that have multiple GO terms associated with them?

What about evidence codes? To find these we apply a similar function to the affyGO terms.

```
> affyEv = eapply(hgu95av2GO, getEvidence)
> table(unlist(affyEv, use.names = FALSE))

   IC   IDA   IEA   IEP   IGI   IMP   IPI   ISS   NAS    ND    NR   TAS
   91  2073 36498   204    19   261   641  2119  6700   886  2409 22068

> test1 = eapply(hgu95av2GO, dropECode, c("IEA", "NR"))
> table(unlist(sapply(test1, getEvidence), use.names = FALSE))

   IC   IDA   IEP   IGI   IMP   IPI   ISS   NAS    ND   TAS
   91  2073   204    19   261   641  2119  6700   886 22068
```

These functions make is somewhat straightforward to select subsets of the GO terms that are specific to different evidence codes.

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.

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.