The *mgsa* package

Sebastian Bauer, Julien Gagneur

9 June 2010

1 Introduction

Model-based Gene Set Analysis (MGSA, Bauer et al. [1]) is a Bayesian modeling approach for gene set enrichment. The package *mgsa* implements MGSA and tools to use MGSA together with the Gene Ontology [2].

2 Quick start

We start with a small simulated dataset which contains `example_go`, a random subset of yeast gene ontology annotations with 20 terms and `example_o`, a simulated set of observed positive genes. These genes could for example be the "hits" of some screen or a set of differentially expressed genes. In the simulation, the terms GO:0006109 and GO:0030663 were active, implying that genes annotated to these terms were more likely to be observed positives than other genes.

```
> library(mgsa)
```

```
[1] "/project/mgsa.Rcheck mgsa"
[1] "Package mgsa initialized"
```

```
> data("example")
> example_go
```

```
Object of class MgsaSets
10 sets over 158 unique items.

Set annotations:
                              term              definition
GO:0046292 formaldehyde metabol... The chemical reactio...
GO:0006109 regulation of carboh... Any process that mod...
GO:0008113 peptide-methionine-(... Catalysis of the rea...
GO:0016849 phosphorus-oxygen ly... Catalysis of the cle...
GO:0046527 glucosyltransferase ... Catalysis of the tra...
... and  5  other sets.

Item annotations:
                           name
SFA1    Bifunctional enzyme ...
YJL068C Non-essential intrac...
```

```
ADR1    Carbon source-respon...
CAT8    Zinc cluster transcr...
FYV10   Protein of unknown f...
... and  153  other items.
```

```
> example_o
```

```
 [1] "SFA1"    "ADR1"    "CAT8"    "FYV10"   "GCR1"    "GCR2"    "GID7"
 [8] "HAP2"    "HAP3"    "HAP4"    "HAP5"    "PCL10"   "PCL6"    "PCL7"
[15] "PCL8"    "PFK26"   "PFK27"   "PHO85"   "PIG1"    "PIG2"    "REG1"
[22] "SIP4"    "SNF1"    "SNF4"    "TYE7"    "UBC8"    "UBP14"   "VID28"
[29] "YLR345W" "GSC2"    "CCT5"    "CPR6"    "CPR7"    "HSC82"   "PET100"
[36] "TIM9"    "COP1"    "GLO3"    "RET2"    "RET3"    "SEC21"   "SEC26"
[43] "SEC27"
```

The method `mgsa` fits the MGSA model. It returns a `MgsaMcmcResults` object whose `print` method displays the most likely active terms. On this example, `mgsa` correctly reports largest posterior probabilities for the terms GO:0006109 and GO:0030663. The call to `set.seed()`, which sets the seed of the random number generator, simply ensures the example of this vignette to be reproducible. It is not required for `mgsa()` to work.

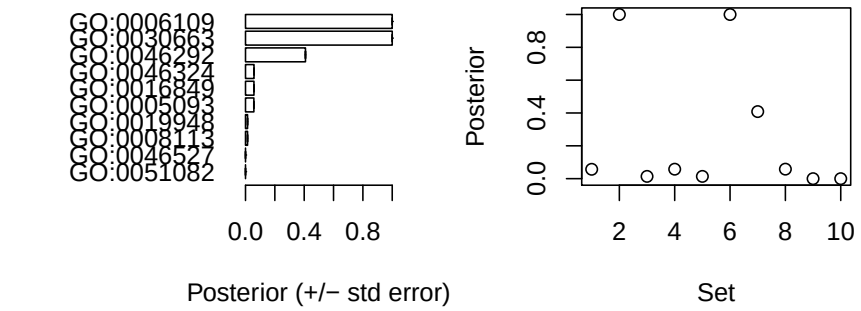
```
> set.seed(0)
> fit = mgsa(example_o, example_go)
> fit
```

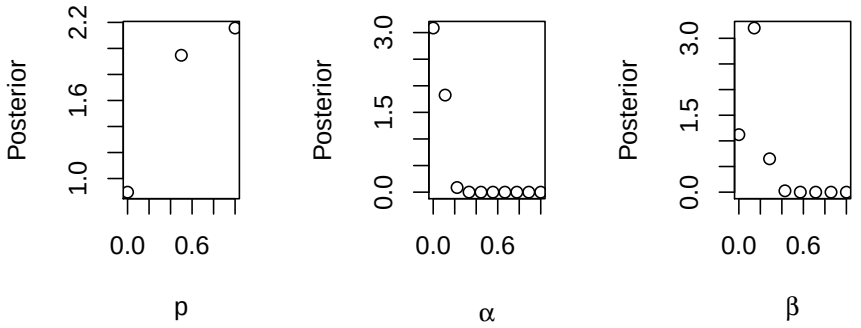
```
Object of class MgsaMcmcResults
158 unique elements in population.
43 unique elements both in study set and in population.
'data.frame':        10 obs. of  4 variables:
 $ inPopulation: int  1 34 2 1 2 8 2 1 21 86
 $ inStudySet  : int  0 28 0 0 0 7 1 0 1 6
 $ estimate    : num  0.057 1 0.0138 0.0576 0.0138 ...
 $ std.error   : num  6.77e-04 6.96e-06 1.90e-04 6.01e-04 2.79e-04 ...
NULL

Posterior on set activity (decreasing order):
           inPopulation inStudySet  estimate    std.error
GO:0006109           34         28 0.9999794 6.961322e-06
GO:0030663            8          7 0.9999464 1.683033e-05
GO:0046292            2          1 0.4086602 1.695734e-03
GO:0046324            1          0 0.0579632 6.940961e-04
GO:0016849            1          0 0.0576434 6.006926e-04
GO:0005093            1          0 0.0570154 6.773733e-04
GO:0019948            2          0 0.0138470 2.788480e-04
GO:0008113            2          0 0.0138426 1.898030e-04
GO:0046527           21          1 0.0000000 0.000000e+00
GO:0051082           86          6 0.0000000 0.000000e+00
```

The method `plot` provides a graphical visualization of the fit.

```
> plot(fit)
```





3 Using the Gene Ontology

The Gene Ontology [2]

References

[1] S. Bauer, J. Gagneur and P. N. Robinson. GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic acids research*, 2010.

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