RDAVIDWebService: a versatile R interface to DAVID

Cristóbal Fresno[1,2] and Elmer A. Fernández[1,2]

[1]Bio-science Data Mining Group, Catholic University of Córdoba, Córdoba, Argentina.
[2]CONICET, Argentina

April 24, 2017

Abstract

Summary: the RDAVIDWebService package provides a class-based interface from R programs/scripts to fully access/control the Database for Annotation, Visualization and Integrated Discovery (DAVID), without the need for human interaction on its website (david.abcc.ncifcrf.gov). The library enhances DAVID capabilities for Gene Ontology analysis by means of GOstats-based direct acyclic graph conversion methods, in addition to the usual many-genes-to-many-terms visualization.
Availability: RDAVIDWebService is available as an R package from the Bioconductor project (www.bioconductor.org) and on the authors' website (www.bdmg.com.ar).
Contact: cfresno@bdmg.com.ar or efernandez@bdmg.com.ar

1 Introduction

One of the most accessed systems for functional genomics/proteomics analysis is the database for annotation, visualization and integrated discovery (DAVID), a web-based online bioinformatics resource (http://david.abcc.ncifcrf.gov) that aims to provide tools for functional interpretation of large lists of genes/proteins (Huang et al., 2009). Its access is carried mainly through a website. There is also a uniform resource locator (URL)-based application programming interface (API), to query DAVID programmatically, accessible through DAVIDQuery R package (Day and Lisovich, 2010). However, the URL-API has limited capabilities, such as URL length and only works with the default settings. In the year 2012, a web service interface was made available allowing full access and control on all its functionalities except visualization (Jiao et al., 2012). Although, it is possible to handle DAVID web services (DWS) through R, it requires high programming skills. In addition, query results are very difficult to manage since they are XML (SOAP package, Lang (2012)) or Java objects (rJava package, Urbanek (2013)).
Here we provide a versatile class-based R interface to access DAVID. It is an R wrapper to all DWS functionalities, with several new features such as off-line

processing (allows using previously queried saved reports) and native R class data types. Additionally it overcomes DWS visualization constraints, providing the usual many-genes-to-many-terms feature, and enhances DAVID capabilities for Gene Ontology (GO, Ashburner et al. (2000)) analysis by means of GOstats-based (Falcon and Gentleman, 2007) direct acyclic graph (DAG) conversion methods. Therefore, it expands DAVID features by allowing new developments through one of the most used computer languages in Bioinformatics, R (R Development Core Team, 2012).

2 Implementation

The package implements *RDAVIDWebService*, a reference class object by means of R5 paradigm, for DWS communication through a Java client (*RDAVIDWebServiceStub*). This allows the establishment of a unique user access point (see Figure 1).

Figure 1: R-to-Java class diagram. `RDAVIDWebService` uses rJava to access/control DAVID web service through `RDAVIDWebServiceStub` and parse the reports back to R by `DAVIDParser`. Note that operation signatures have been omitted for simplicity.

In order to reduce Java-to-R handshaking due to parsing data structures (a time consuming computational task), the provided Java-based file report client was enhanced (*DAVIDParser*) to allow formatting of all the DAVID outputs into appropriate RDAVIDWebService S4 R classes(see Figure 2, and section 4). This speeds up the bottleneck data importation process. In addition, allows locally saving DAVID query to file for further analysis, as well as using website generated reports. Thus, permits using web services and website query results interchangeably.

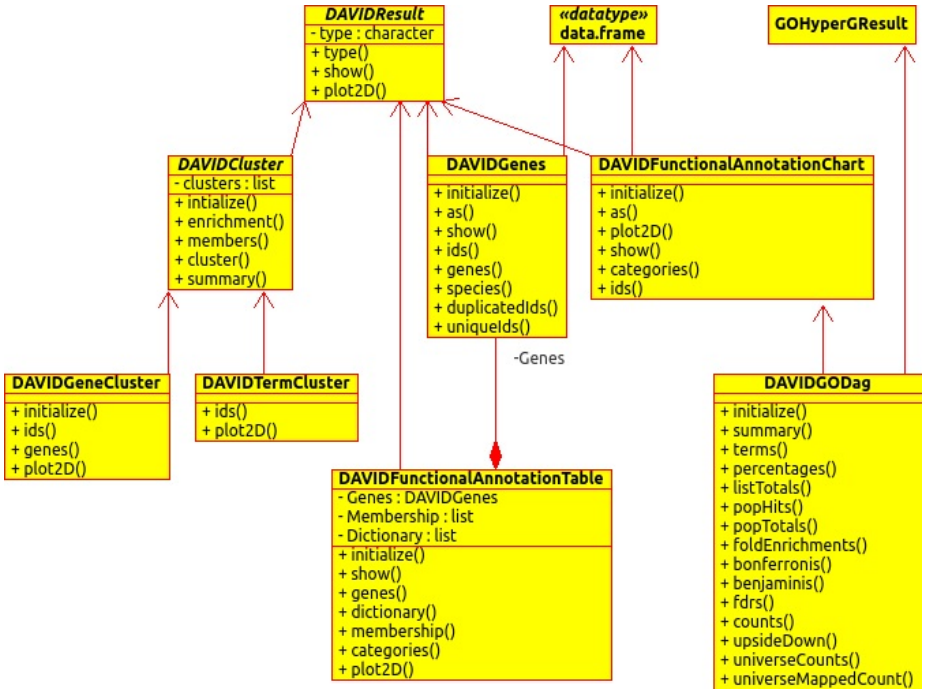

Figure 2: Hierarchy R class diagram. Note that operation signatures have been omitted for simplicity.

3 Features

1. *Ease of Use*: it provides a uniform framework to access DAVID analysis straight from R without the need of ad hoc parsing queried reports.

2. *Data import/export*: results from DAVID can be accessed through R or also generated on the website. In both cases they are stored in the same format for later use. This permits ON/OFF-line processing capabilities within R. Hence, empowering queried reports generated anytime and anywhere for processing, without the need to redo the uploading to DAVID.

3. *Visualization*: customizable many-genes-to-many-terms 2D relationship views are also available with the ggplot2 package (Wickham, 2009).

4. *Gene Ontology structure*: DAVID set enrichment analysis (SEA) or modular enrichment (MEA) results can be mapped into GOstats-based direct acyclic graphs. This enables visualization of EASE score-based enriched biological process (BP), molecular function (MF) and cellular component (CC) GO terms in the DAGs. Thus, the exploration and analysis of blurred pattern presence is facilitated, compared to the usual tabular format.

4 Package overview

The package can be conceptually divided into two parts:

- **Connectivity**: a wrapper to DAVID web service for the basic work flow: (1) upload of gene/protein ids as gene/background list/s; (2) check DAVID's status for mapped/unmapped genes in the uploaded list/s or lookup the available categories, etc. (3) select the background/species and categories to use in the present analysis, and (4) get the different reports which includes Functional Annotation Chart/Table/Clustering and so on.

- **Exploration**: on/off-line report import of the different results into R objects hierarchy (see Figure 2), ready to use them with the use favourite CRAN (Hornik, 2012) or Bioconductor (Gentleman et al., 2004) package/s. In addition, DWS capabilities are enhanced with the incorporation of the usual many-gene-to-many terms 2D relationship visualization available at DAVID's website, and the new feature which generates the induced Gene Ontology GOstats direct acyclic graph, in order to get the big biological picture, as we will show in section 4.2.

4.1 Connectivity example

RDAVIDWebService requires a registered DAVID user (this is a prerequisite to use DWS). By means of the registered institutional e-mail, the user can build a `DAVIDWebService` object and establish a connection. Then, a gene list should be uploaded providing a name and type of list. Here, the one provided in the DAVID website is used (`demoList1` with Affymetrix® identifiers).
Note: the following code will not run unless you change the `"user@inst.org"` e-mail by the user registered DAVID account.

```
R> library("RDAVIDWebService")
R> david<-DAVIDWebService$new(email="user@inst.org")
R> data(demoList1)
R> result<-addList(david, demoList1,
+ idType="AFFYMETRIX_3PRIME_IVT_ID",
+ listName="demoList1", listType="Gene")
R> result

$inDavid
[1] 0.9695122

$unmappedIds
[1] "34902_at"   "1937_at"    "35996_at"   "32163_f_at" "32407_f_at"
```

The `result` output shows that 96.95% of the complete `demoList1` are recognized `$inDavid`. In addition, this object also contains the five `$unmappedIds`. On the other hand, the status of the connection is saved in `david` object.

```
R> david

DAVIDWebService object to access DAVID's website.
User email: user@inst.org
Available Gene List/s:
       Name Using
1 demoList1     *
Available Specie/s:
              Name Using
1 Homo sapiens(155)     *
Available Background List/s:
          Name Using
1 Homo sapiens     *
```

In this example, 155 genes corresponding to *Homo sapiens* are present in `demoList1`. The complete genome is selected as the default background but, the user can upload their own by modifying `listType="Background"`. If required, the user can select which annotation category to use, e.g. `GOTERM_BP_ALL`, `GOTERM_MF_ALL` and `GOTERM_CC_ALL`.

```
R> setAnnotationCategories(david, c("GOTERM_BP_ALL",
+ "GOTERM_MF_ALL", "GOTERM_CC_ALL"))
```

Now, that everything is in order, the user can get the different reports to use right away or to save into a file for future recall. For example the Functional Annotation Clustering can be obtained on-line on `termCluster` object, or as `termClusterReport1.tab` file by invoking:

```
R> termCluster<-getClusterReport(david, type="Term")
R> getClusterReportFile(david, type="Term",
+ fileName="termClusterReport1.tab")
```

4.2 Exploration example

Hereafter, we asume that at some point `demoList1` has been used and every report saved into files (data available in the package).
A user can obtain the functional annotation cluster report of `demoList1` and inspect the results using the following code:

```
R> library("RDAVIDWebService")
R> fileName<-system.file("files/termClusterReport1.tab.tar.gz",
+    package="RDAVIDWebService")
R> untar(fileName)
R> termCluster<-DAVIDTermCluster(untar(fileName, list=TRUE))
R> termCluster

DAVID Result object
Result type:  AnnotationCluster
Number of cluster:  28

R> head(summary(termCluster))
```

```
  Cluster Enrichment Members
1       1      2.904      14
2       2      2.135       4
3       3      2.059      10
4       4      1.977      14
5       5      1.501       4
6       6      1.347       4
```

Here, `termCluster` is an object from class `DAVIDTermCluster` with the corresponding `AnnotationCluster` report data of `demoList1`, where 28 clusters are found. Then, `head(summary(termCluster))` can provide a superficial view of the `Enrichment` Score reached in each cluster and how many `Members` are present. The user can visually explore the 2D view of a particular cluster (e.g. the second on Figure 3).

```
R> clustNumber<-2
R> plot2D(termCluster, clustNumber)
```

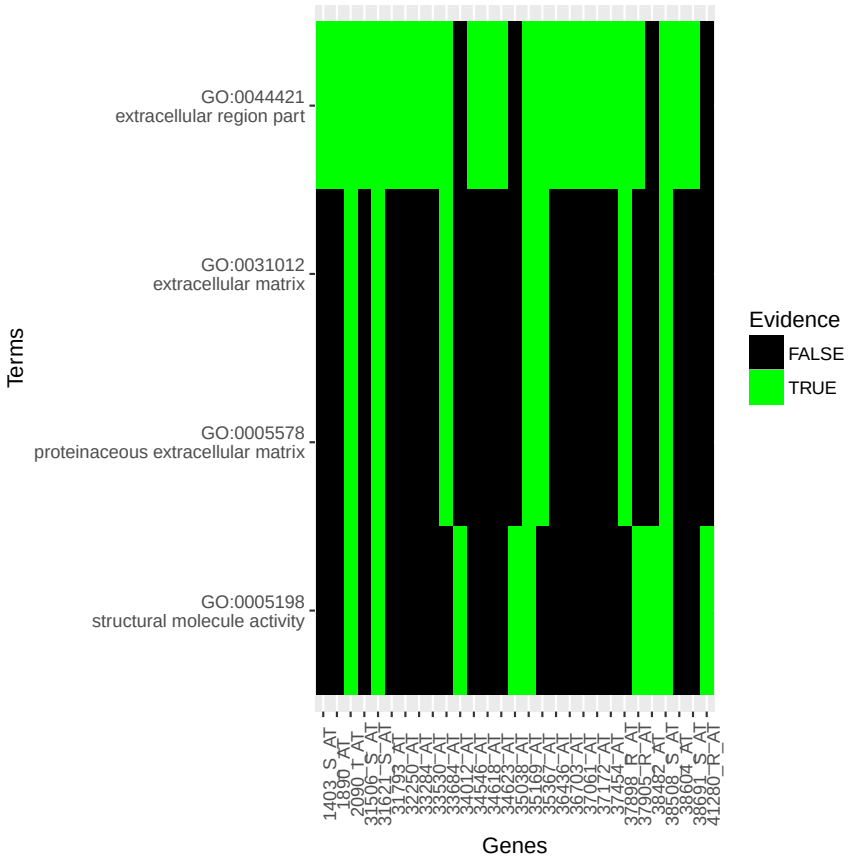

Figure 3: Functional annotation cluster exploration, using a 2D-view of the evidence of the four term/category members present in the second cluster.

However, in Figure 3, the four term/category members of this cluster share

all the ids at "extracellular region part" (upper row). But, as we go down towards the bottom row (structural molecule activity) only nine ids have evidence related to it. In this view, the hierarchical structure of GO is not considered nor the members that are enriched or not (default option). Therefore, the user can extend DAVID's features obtaining the associated induced DAG structure of the cluster (`DAVIDGODag`) and contextualize it using GOstats functionalities (`plotGOTermGraph`, see Figure 4).

```
R> davidGODag<-DAVIDGODag(members(termCluster)[[clustNumber]],
+    pvalueCutoff=0.1, "CC")
R> plotGOTermGraph(g=goDag(davidGODag),
+    r=davidGODag, max.nchar=40, node.shape="ellipse")
```

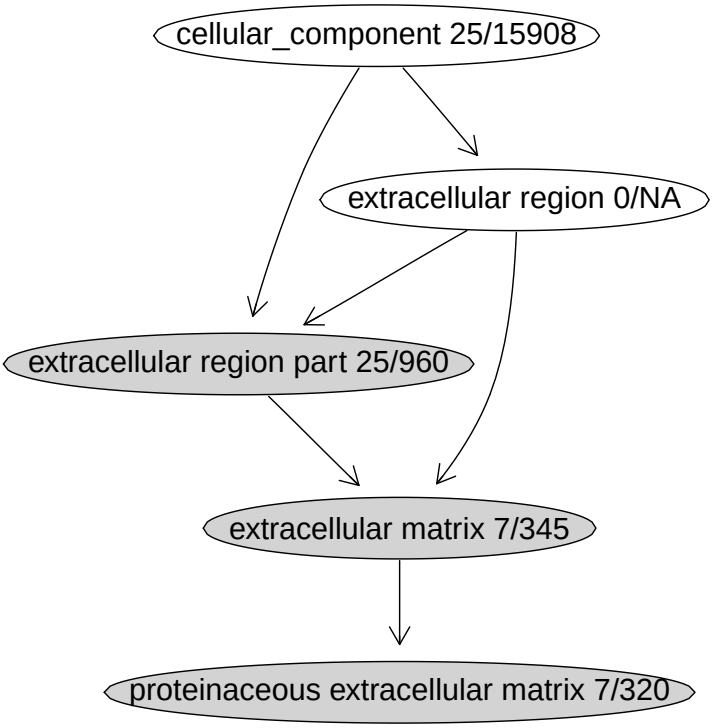

Figure 4: Gene Ontology direct acyclic graph induced by cluster two in figure 3. Terms with an EASE score < 0.1 are shown in grey. In addition, the ratio between genes on the list vs. background reference is displayed. Since no information is available regarding the other terms into the mapped GO structure from the cluster, NAs (not available) are introduced when required.

5 Trouble shooting

Sometimes apache Axis' default parameters needs to be changed in order to use RDAVIDWebService. For example if DAVID web service is bussy the default timeout would not be enought. Then you can inspect it and change it calling:

```
R> getTimeout(david)

[1] 30000

R> setTimeout(david, 50000)
R> getTimeout(david)

[1] 50000
```

Other reported parameter that might need to be changed is the transport protocol if you get the following error when uploading a gene list:

```
org.apache.axis2.AxisFault: Transport error: 501 Error: Not Implemented
```

Then you can change on the client side the appropiate parameter value:

```
R> setHttpProtocolVersion(david, "HTTP/1.0")
R> getHttpProtocolVersion(david)

[1] "HTTP/1.0"
```


Acknowledgements

Funding: this work was supported by the National University of Villa Maria [31/0186 to E.F. and 31/0187 to E.F.] and Catholic University of Córdoba, Argentina.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29.

Day, R. and Lisovich, A. (2010). *DAVIDQuery: Retrieval from the DAVID bioinformatics data resource into R*. R package version 1.20.0.

Falcon, S. and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–258.

Gentleman, R. C., Carey, V. J., Bates, D. M., and others (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.

Hornik, K. (2012). The comprehensive r archive network. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(4):394–398.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat Protoc*, 4(1):44–57.

Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2012). David-ws: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, 28(13):1805–1806.

Lang, D. T. (2012). *SSOAP: Client-side SOAP access for S*. R package version 0.9-1.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Urbanek, S. (2013). *rJava: Low-level R to Java interface*. R package version 0.9-4.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

Session Info

```
R> sessionInfo()

R version 3.4.0 (2017-04-21)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows Server 2012 R2 x64 (build 9600)

Matrix products: default

locale:
[1] LC_COLLATE=C
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
 [1] grid      stats4    parallel  stats     graphics  grDevices
 [7] utils     datasets  methods   base

other attached packages:
 [1] Rgraphviz_2.20.0        GO.db_3.4.1
 [3] RDAVIDWebService_1.14.0 ggplot2_2.2.1
 [5] GOstats_2.42.0          Category_2.42.0
 [7] Matrix_1.2-9            AnnotationDbi_1.38.0
 [9] IRanges_2.10.0          S4Vectors_0.14.0
[11] Biobase_2.36.0          graph_1.54.0
[13] BiocGenerics_0.22.0
```

```
loaded via a namespace (and not attached):
 [1] Rcpp_0.12.10           compiler_3.4.0
 [3] plyr_1.8.4             bitops_1.0-6
 [5] tools_3.4.0            digest_0.6.12
 [7] annotate_1.54.0        RSQLite_1.1-2
 [9] memoise_1.1.0          tibble_1.3.0
[11] gtable_0.2.0           lattice_0.20-35
[13] DBI_0.6-1              rJava_0.9-8
[15] genefilter_1.58.0      GSEABase_1.38.0
[17] XML_3.98-1.6           RBGL_1.52.0
[19] survival_2.41-3        scales_0.4.1
[21] splines_3.4.0          AnnotationForge_1.18.0
[23] xtable_1.8-2           colorspace_1.3-2
[25] RCurl_1.95-4.8         lazyeval_0.2.0
[27] munsell_0.4.3
```