queryMed package: annotate medicine and pathology codes for pharmaco-epidemiological studies

Y. Rivault, O.Dameron and N. Le Meur

version 0.1 - 17/04/2018

Context

In the era of massive open-data access and big data in general, pharmaco-epidemiology and public health sciences are in need for bioinformatics tools. Researchers can now query large medical information systems such as medico-administrative databases and claim databases. Although those information systems are often well strutured, their contents are highly codified with several medical terminologies and ontologies, which might be difficult to analyze by non-expert. Pharmaco-epidemiologists and public health scientists thus need batch translation of diverse medical codes like diagnostic codes (based on ICD9 or ICD10), medical procedures or drugs codes (in ATC nomenclature). In addition to the translation of those codes, their annotation can help making the most of medico-administrative and claim databases analysis by public health researchers. For instance, in a patient care trajectory it might help identifying critical drug interactions that might impair the patient safety. It might also be used to predict evitable hospitalisation.

Using ontologies in R has proven to be efficient and useful[Kurbatova et al., 2011], notably in *omic fields of research. For example, many genomes (including the Human genome) are available for downlaod through the BioConductor repository. Those genomes are annotated and statistical analyses of enrichment of standardised terms (commun or closely related) within part of living organisms have help discovering new or impaired functions.

Today the Linked Open Data and the Semantic Web provide the technical solutions for the integration of distributed data, their interrogation and their interpretation. For instance, the Ressource Description Framework (RDF) standardise the representation of data and knowledge, and thus allows their sharing and reuse. Furthermore, SPARQL, an other standard from the Semantic Web, provides a way to querring these kind of data.

So these technologies allow reasoning through knowledge, especially in the form of ontologies, classifications or medical thesauri. Essential knowledge in pharmacoepidemiology, drug interactions, indications and contraindications has thus been taken into account in clinical data analyzes [Pathak et al., 2013]. However, while works like the Drug Indication Database (DID) and Drug Interaction Knowledge Base (DIKB) have pooled different sources of medical knowledge from the Web of Data [Ayvaz et al., 2015, Sharp, 2017], the use and merging of knowledge is still well-to-do with the multitude of medical classifications and sources of knowledge.

We propose a tool to integrate medical ontologies programmatically in the R environment. The queryMed package provides functions and algorithms to query the different sources of medical knowledge representations from the Web of data and to link them to the main medical classifications, for the enrichment and the analysis of medical data. The proposed functions are of two sort: for expert and non-expert. In this vignette, we first present in the material and methods the data sources and the query tools and methods. Next, we present application for SPARQL programmers and non-expert users of that query language. Finally, we illustrate the interest of our packages in the context of pharmacovigilance.

Material and Methods

The queryMed package aims to provide to pharmaco-epidemiologists tools allowing easiest way of accessing and linking medical and pharmacological knowledges.

To interrogate Linked Open Data, queryMed propose a general function to querying the SPARQL endpoints on the Web. The use of a such function requires from the user to know both SPARQL standard and data structure available on these endpoints. Therefore, the package also aims to provide generic functions for commun SPARQL queries on medical and pharmacological specific SPARQL endpoints. For instance, Bioportal[Salvadores et al., 2012, 2013], DB-pedia and Bio2rdf[Callahan et al., 2013] contain lots of informations about drugs. Quering their SPARQL endpoints can allow retrieving definitions, formulas, labels, comments about drugs, on also mappings between drugs nomenclatures.

Medical informations from ontologies are also available through REST API. For instance, the REST api from BIOPORTAL[Whetzel et al., 2011], allows us to retrieve mappings between main international medical nomenclatures via concept unique identifier (cui).

queryMed also integrate the Drug Interaction Knowledge Base (DIKB) [Sharp, 2017] and the Drug Indication Database (DID) [Ayvaz et al., 2015], which are already inititives aiming at mutualise different sources of essential pharmacological knowledges.

Applications

Using queryMed, the platelet antiaggregant CLOPIDOGREL presents 26 known interactions with other drugs according to DIKB (Fig1A). Some interactions are reported by more than one source (represented by the thinkness of the edge). Some drugs belong to the same ATC family (represented by the vertex color). When considering more general classes of the ATC nomenclature, the interactions concern 15 and 11 classes, respectively (Fig1B, Fig1C). Again some classes are characterized by a high number of interactions (size of the vertex) and sources (thinkness of the edge), notably the anti-thrombotic and gastro-esophageal family. In the context of pharmacovigilance, one could then make assumptions about the value of expanding the study of interactions between clopidogrel and the 26 known drugs to those between clopidogrel and the families they belong to.

```
library(queryMed)
library(data.table)
```

```
DIKB <- get_DIKB(path="/tmp",mapping="ATC")
DIKB$object <- toupper(DIKB$object)
DIKB$precipitant <- toupper(DIKB$precipitant)
clop = DIKB[DIKB$object=="CLOPIDOGREL",]
clop = as.data.table(clop)

clop2 <- clop
clop2$atc2   <- substr(clop$atc2, 24, 30)
clop2$parent  <- substr(clop2$atc2,1, 5)
clop2$root  <- substr(clop2$atc2,1, 4)
clop2<- clop2[!is.na(clop2$atc2),]

statclop = clop2[, .N,by=list(precipitant, atc2, root)]
```

```
library(RColorBrewer)
mypalette<-brewer.pal(11,"Set3")
mypalette = cbind(mypalette, "root"=unique(statclop$root))
statclop<- merge(statclop, mypalette, by="root")

library(igraph)
g1 <- graph_from_edgelist(as.matrix(cbind("CLOPIDOGREL", statclop[,2])), directed=F)
statclop$N[statclop$N == 1 ] <- 0
E(g1)$weight <- statclop$N
E(g1)$width <- 1+E(g1)$weight*2
```

```
V(g1)$size <- 20
V(g1)$frame.color <- "white"
V(g1)$color <-  c("lightgrey", statclop$mypalette)
l1 <- layout_as_star(g1)
plot(g1, edge.color="grey", layout=l1)

# ancestor I
res1<-c()
res2<-c()
for (i in 1:nrow(clop2)){
temp<- get_ancestors(clop2$atc2[i], ontology="ATC", api_key="e6f2d058-f206-4ac7-a8f9-60b84c9e57dc")
res1[i] <- temp[1,2]
res2[i] <- temp[2,2]
}
clop2$anc1 <- res1
clop2$anc2 <- res2


statclop1 = clop2[, .N, by=list(anc1,root, parent)]
nbdrugclop1 = clop2[, length(unique(precipitant)), by=parent]

statclop2 = clop2[, .N, by=list(anc2,root)]
nbdrugclop2 = clop2[, .N, by=root]

colnames(nbdrugclop1) = c("parent","nbdrug")
colnames(nbdrugclop2) = c("root","nbdrug")
statclop1 = merge(statclop1, nbdrugclop1, by="parent")
statclop2 = merge(statclop2, nbdrugclop2, by="root")

statclop1<- merge(statclop1, mypalette, by="root")
statclop2<- merge(statclop2, mypalette, by="root")

g2 <- graph_from_edgelist(as.matrix(cbind("CLOPIDOGREL", statclop1[,3])), directed=F )
statclop1$N[statclop1$N == 1 ] <- 0
E(g2)$weight <- statclop1$N
E(g2)$width <- 1+E(g2)$weight
V(g2)$size <- c(8, statclop1$nbdrug)*6
V(g2)$vertex.label.cex <-2
V(g2)$frame.color <- "white"
V(g2)$color <-  c("lightgrey", statclop1$mypalette)
#V(g2)$color <- "orange"
l2 <- layout_as_star(g2)
plot(g2, layout=l2)

# ancestor II
g3 <- graph_from_edgelist(as.matrix(cbind("CLOPIDOGREL", statclop2[,2])), directed=F)
statclop2$N[statclop2$N == 1 ] <- 0
E(g3)$weight <- statclop2$N
E(g3)$width <- 1+E(g3)$weight
w= c(15,20,15,20, rep(15,7))
V(g3)$size <- c(20, statclop2$nbdrug+w)
V(g3)$vertex.label.cex <-0.5
V(g3)$frame.color <- "white"
```

```
V(g3)$color <-  c("lightgrey", statclop2$mypalette)
l3 <- layout_as_star(g3)
plot(g3, vertex.label.cex= 0.8, layout=l3)

png("queryMed/vignettes/Figure1.png")
par(mfrow=c(2,2), mar=c(1,1,1,1))
plot(g1, vertex.label.cex= 0.8, layout=l1*0.6, main="A - Drug")
plot(g2, vertex.label.cex= 0.8, layout=l2*0.6, main="B - Ancestor I")
plot(g3, vertex.label.cex= 0.8, layout=l3*0.6, main="C - Ancestor II")
dev.off()
```

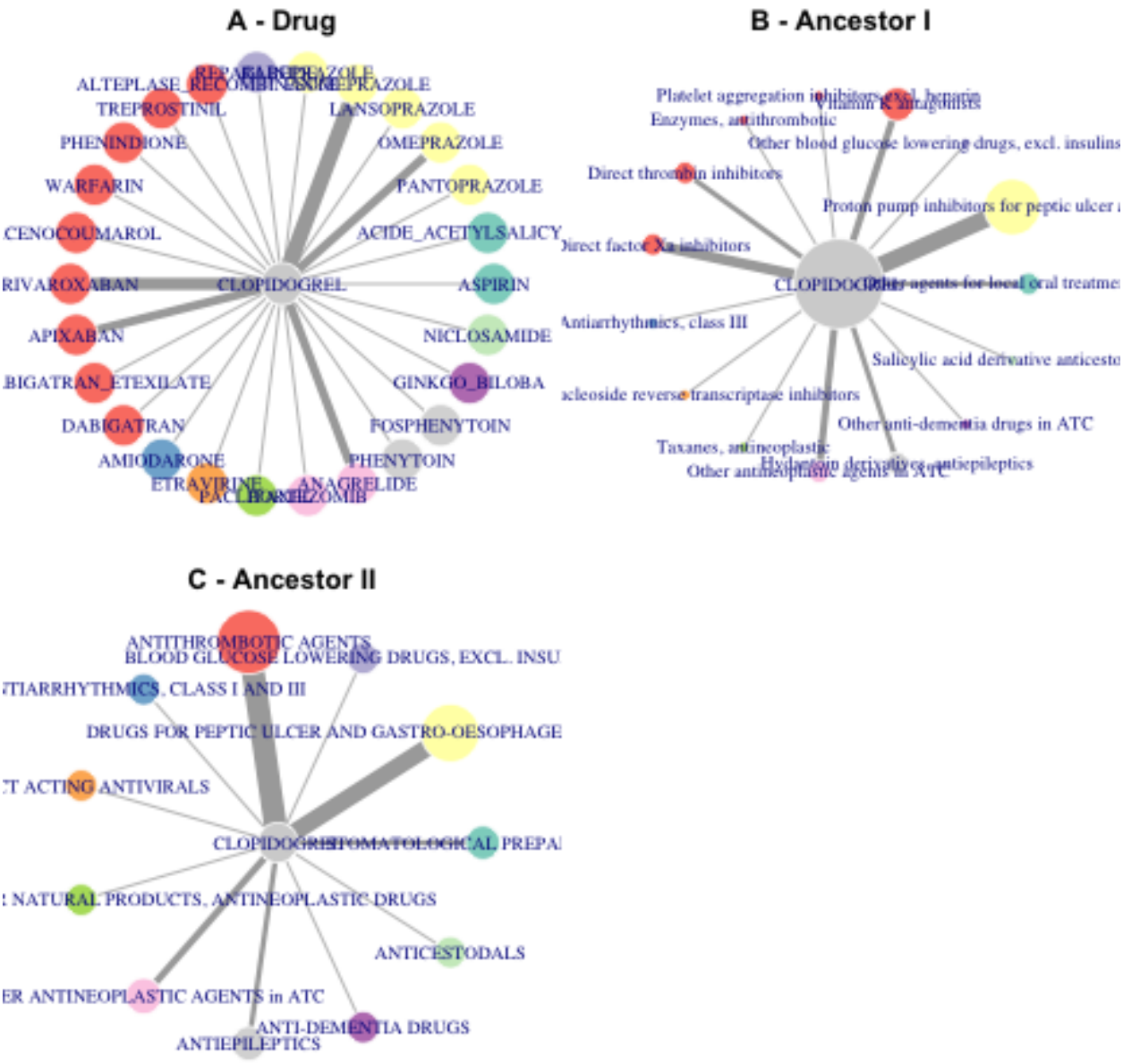

Figure 1: Clopidogrel's friends

```
sessionInfo()
```

```
## R version 3.4.3 (2017-11-30)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Debian GNU/Linux 8 (jessie)
##
## Matrix products: default
## BLAS: /usr/lib/openblas-base/libblas.so.3
## LAPACK: /usr/lib/libopenblasp-r0.2.12.so
##
## locale:
##  [1] LC_CTYPE=fr_FR.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=fr_FR.UTF-8        LC_COLLATE=fr_FR.UTF-8
##  [5] LC_MONETARY=fr_FR.UTF-8    LC_MESSAGES=fr_FR.UTF-8
##  [7] LC_PAPER=fr_FR.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=fr_FR.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_3.4.3  backports_1.1.1 magrittr_1.5    rprojroot_1.3-2
##  [5] tools_3.4.3     htmltools_0.3.6 yaml_2.1.14     Rcpp_0.12.13
##  [9] stringi_1.1.6   rmarkdown_1.8   knitr_1.17      stringr_1.3.0
## [13] digest_0.6.12   evaluate_0.10.1
```

References

Serkan Ayvaz, John Horn, Oktie Hassanzadeh, Qian Zhu, Johann Stan, Nicholas P. Tatonetti, Santiago Vilar, Mathias Brochhausen, Matthias Samwald, Majid Rastegar-Mojarad, Michel Dumontier, and Richard D. Boyce. Toward a complete dataset of drug–drug interaction information from publicly available sources. *Journal of Biomedical Informatics*, 55:206–217, June 2015. ISSN 15320464. doi: 10.1016/j.jbi.2015.04.006. URL http://linkinghub.elsevier.com/retrieve/pii/S1532046415000738.

Alison Callahan, José Cruz-Toledo, Peter Ansell, and Michel Dumontier. Bio2rdf Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data. In *The Semantic Web: Semantics and Big Data*, Lecture Notes in Computer Science, pages 200–212. Springer, Berlin, Heidelberg, May 2013. ISBN 978-3-642-38287-1 978-3-642-38288-8. doi: 10.1007/978-3-642-38288-8_14. URL https://link.springer.com/chapter/10.1007/978-3-642-38288-8_14.

N. Kurbatova, T. Adamusiak, P. Kurnosov, M. A. Swertz, and M. Kapushesky. ontoCAT: an R package for ontology traversal and search. *Bioinformatics*, 27(17):2468–2470, September 2011. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btr375. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr375.

Jyotishman Pathak, Richard C. Kiefer, and Christopher G. Chute. Using linked data for mining drug-drug interactions in electronic health records. *Studies in Health Technology and Informatics*, 192:682–686, 2013. ISSN 0926-9630.

Manuel Salvadores, Matthew Horridge, Paul R. Alexander, Ray W. Fergerson, Mark A. Musen, and Natalya F. Noy. Using SPARQL to Query BioPortal Ontologies and Metadata. In Philippe Cudré-Mauroux, Jeff Heflin, Evren Sirin, Tania Tudorache, Jérôme Euzenat, Manfred Hauswirth, Josiane Xavier Parreira, Jim Hendler, Guus Schreiber, Abraham Bernstein, and Eva Blomqvist, editors, *The Semantic Web – ISWC 2012*, pages 180–195. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-35173-0.

Manuel Salvadores, Paul R. Alexander, Mark A. Musen, and Natalya F. Noy. BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF. *Semantic Web*, 4(3):277–284, 2013. ISSN 1570-0844.

Mark E Sharp. Toward a comprehensive drug ontology: extraction of drug-indication relations from diverse information sources. *Journal of Biomedical Semantics*, 8(1), December 2017. ISSN 2041-1480. doi: 10.1186/s13326-016-0110-0. URL http://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-016-0110-0.

Patricia L. Whetzel, Natalya F. Noy, Nigam H. Shah, Paul R. Alexander, Csongor Nyulas, Tania Tudorache, and Mark A. Musen. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39 (Web Server issue):W541–545, July 2011. ISSN 1362-4962. doi: 10.1093/nar/gkr469.