webQueries: Retrieve genes or protein annotations by querying NCBI.

Frederic Commo[1]

[1]*INSERM U981, Institut Gustave Roussy, 114 rue Edouard Vaillant, 94805 Villejuif, France*

March 17, 2015

1 Introduction

`webQueries` provides R functions for retrieving gene and protein annotations from NCBI, given HUGO symbols or ids (entrezgene and Uniprot are supported).
The two main functions, `geneQuery()` and `protQuery()`, are wrappers for the NCBI E-utilities functions.
Visit `http://www.ncbi.nlm.nih.gov/books/NBK25500/` for more details.

2 Examples

Here are some examples showing how a single, or multiple, annotation(s) can be retrieved using either official symbols or identifiers:

```
require(webQueries)

Loading required package:  webQueries

# Using HUGO symbol
geneQuery("erbb2")
ERBB2 found: 1 id(s)... Done.
                                                  query
                                                "ERBB2"
                                                 symbol
                                                "ERBB2"
                                               fullName
                    "erb-b2 receptor tyrosine kinase 2"
                                                  alias
"CD340, HER-2, HER-2/neu, HER2, MLN 19, NEU, NGL, TKR1"
                                               organism
                                           "Homo sapiens"
                                                verifId
                                                    "0"
                                                 status
                                             "Official"
                                                    chr
                                                  "17"
```

```
                                              cytoband
                                               "17q12"
                                             exonCount
                                                  "32"
                                            accVersion
                                         "NC_000017.11"
                                               chrStart
                                            "39688083"
                                                 chrEnd
                                            "39728661"
                                              genomStart
                                           "2539859947"
                                                genomEnd
                                           "2539900525"
                                           entrezgeneId
                                                 "2064"
# Using entrezgene id
data.frame(output=geneQuery(2064))
                                                                    output
query                                                                 2064
symbol                                                               ERBB2
fullName                                erb-b2 receptor tyrosine kinase 2
alias        CD340, HER-2, HER-2/neu, HER2, MLN 19, NEU, NGL, TKR1
organism                                                     Homo sapiens
verifId                                                             0
status                                                          Official
chr                                                                17
cytoband                                                         17q12
exonCount                                                           32
accVersion                                                   NC_000017.11
chrStart                                                        39688083
chrEnd                                                          39728661
genomStart                                                    2539859947
genomEnd                                                      2539900525
entrezgeneId                                                        2064
```

The same task can be applied on multiple symbols (or ids):

```
# Multiple queries
ids <- c("egfr", "erbb2", "fgfr1")
annots <- lapply(ids, function(id) geneQuery(id) )
EGFR found: 1 id(s)... Done.
ERBB2 found: 1 id(s)... Done.
FGFR1 found: 1 id(s)... Done.
annots <- do.call(rbind, annots)
annots
     query   symbol  fullName
[1,] "EGFR"  "EGFR"  "epidermal growth factor receptor"
[2,] "ERBB2" "ERBB2" "erb-b2 receptor tyrosine kinase 2"
[3,] "FGFR1" "FGFR1" "fibroblast growth factor receptor 1"
     alias
[1,] "ERBB, ERBB1, HER1, NISBD2, PIG61, mENA"
[2,] "CD340, HER-2, HER-2/neu, HER2, MLN 19, NEU, NGL, TKR1"
```

```
[3,] "BFGFR, CD331, CEK, FGFBR, FGFR-1, FLG, FLT-2, FLT2, HBGFR, HH2, HRTFDS, KAL2, N-SAM, OGD, b
     organism       verifId status     chr  cytoband         exonCount
[1,] "Homo sapiens" "0"     "Official" "7"  "7p12"           "30"
[2,] "Homo sapiens" "0"     "Official" "17" "17q12"          "32"
[3,] "Homo sapiens" "0"     "Official" "8"  "8p11.23-p11.22" "24"
     accVersion     chrStart   chrEnd     genomStart   genomEnd
[1,] "NC_000007.14" "55019031" "55207337" "1288676058" "1288864364"
[2,] "NC_000017.11" "39688083" "39728661" "2539859947" "2539900525"
[3,] "NC_000008.11" "38468833" "38411137" "1431264523" "1431206827"
     entrezgeneId
[1,] "1956"
[2,] "2064"
[3,] "2260"
```

Similarly to `geneQuery`, `protQuery` can deal with either symbols or ids - here UniProt identifiers - and can be reformated:

```
# Using UniProt id
myProt <- protQuery("P04626")
P04626 found: 4 id(s)...
as.list(myProt)
$query
[1] "P04626"

$recName
[1] "Receptor tyrosine-protein kinase erbB-2"

$altName
[1] "Metastatic lymph node gene 19 protein|Proto-oncogene Neu|Proto-oncogene c-ErbB-2|Tyrosine ki

$UniProt
[1] "P04626"

$Gi
[1] "119533"

$Extra
[1] "gi|119533|sp|P04626.1|ERBB2_HUMAN[119533]"

$lengthAA
[1] "1255"
```

Again, the function can be put into any kind of loop, e.g. `lapply`, in order to run multiple queries:

```
# Multiple queries
ids <- c("egfr", "erbb2", "fgfr1")
annots <- lapply(ids, function(id) protQuery(id) )
EGFR found: 20 id(s)...
ERBB2 found: 20 id(s)...
FGFR1 found: 20 id(s)...
annots <- do.call(rbind, annots)
annots
     query   recName
```

```
[1,] "EGFR"  "Epidermal growth factor receptor"
[2,] "ERBB2" "Receptor tyrosine-protein kinase erbB-2"
[3,] "FGFR1" NA
     altName
[1,] "Proto-oncogene c-ErbB-1|Receptor tyrosine-protein kinase erbB-1"
[2,] "Metastatic lymph node gene 19 protein|Proto-oncogene Neu|Proto-oncogene c-ErbB-2|Tyrosine k
[3,] NA
     UniProt        Gi
[1,] "P00533"       "2811086"
[2,] "P04626"       "119533"
[3,] "XP_011542754" "767950690"
     Extra                                          lengthAA
[1,] "gi|2811086|sp|P00533.2|EGFR_HUMAN[2811086]"  "1210"
[2,] "gi|119533|sp|P04626.1|ERBB2_HUMAN[119533]"   "1255"
[3,] "gi|767950690|ref|XP_011542754.1|[767950690]" "494"
```

3 Accessing R code

R source code for `webQueries` is available on github: `https://github.com/fredcommo/webQueries`