Exercises and solutions for chapter 'Data Technologies'

August 11, 2008

Exercise 1
Which chromosome has the the most probe sets and which has the fewest?

Exercise 2
Use the function `ppc` *from Exercise* **??** *to create a new function that can find and return the probes that map to any chromosome (just prepend the caret to the chromosome number) or the chromosome number with a p or a q after it.*

Exercise 3
How many genes are in the homologous region shared by chromosomes X and Y.

Solutions: Basically you need to first reduce to unique Entrez Gene IDs, then find those that have just one location on the X chromosome and those that have just one location on the Y, then count the number in common, which turns out to be 15.

```
> len2 = names(mLens[mLens == 2])
> len2EG = unlist(mget(len2, hgu95av2ENTREZID))
> len2EG = len2EG[!duplicated(len2EG)]
> len2 = len2[!duplicated(len2EG)]
> mapP = mget(len2, hgu95av2MAP)
> hasX = sapply(mapP, function(x) if (length(grep("^X",
+     x)) == 1) TRUE else FALSE)
> hasY = sapply(mapP, function(x) if (length(grep("^Y",
+     x)) == 1) TRUE else FALSE)
> table(hasX & hasY)
FALSE  TRUE
  170    15
```

Exercise 4
Which chromosome band has the most probe sets contained in it? How many chromosome bands are from chromosome 2? How many are on the p-arm and how many on the q-arm?

Exercise 5
Is there a ***DBI*** *generic function that will retrieve an entire table in a single command. If so, what is its name, and what is its return value?*

Solutions: The function is called `dbReadTable` and its return value is a *data.frame*.

Exercise 6
Select all entries from the `USArrests` *database where the murder rate is larger than 10.*

Solutions: To do this, we use the `WHERE` clause in a SQL `SELECT` query.

```
> rs = dbSendQuery(con,
+     "SELECT * FROM USArrests WHERE Murder > 10")
```

Exercise 7
For each table in the `hgu95av2.db` *database, determine the type of each field.*

Exercise 8
How many GO evidence codes are there, and what are they?

Exercise 9
Use an inner join to relate GenBank IDs to GO ontology codes.

Solutions: One way to solve this problem is with an inner join, using the Affymetrix IDs, between the tables `acc` and `go_probe`.

```
> query = paste("SELECT acc_num, go_id", "FROM acc, go_probe",
+     "WHERE (acc.affy_id = go_probe.affy_id)")
> rs = dbSendQuery(con, query)
> f3 = fetch(rs, n = 3)
> f3
```

```
  acc_num      go_id
1  X13589 GO:0004497
2  X13589 GO:0005489
3  X13589 GO:0005506
> dbClearResult(rs)
[1] TRUE
```

Exercise 10
How many name space definitions are there for the XML document that was parsed? What are the URIs for each of them?

Solutions: There are two and their URIs are given below.

```
> length(nsY)
[1] 2
> sapply(nsY, function(x) x[[2]])
                          "net:sf:psidev:mi"
                                         xsi
"http://www.w3.org/2001/XMLSchema-instance"
```

Exercise 11
Carry out the first suggestion above. That is, starting with `f1`*, retrieve the element attributes and then process them via* `grep` *and* `gsub` *to find the names of the packages. Compare your results with those above.*

Solutions:

```
> pkgs = sapply(f1, xmlGetAttr, "href")
> pkg = grep("/packages/2.1/bioc/html/", pkgs, fixed = TRUE)
```

Exercise 12
What other data were returned by the call to `getGene`*?*