pxr: an *R* interface to the ProteomeXchange repository

Laurent Gatto
lg390@cam.ac.uk
Computational Proteomics Unit*

March 28, 2014

1 Introduction

The goal of the *pxr* package is to provide programmatic access to proteomics data from *R*, in particular to the ProteomeXchange[1] (PX) central repository (see http://www.proteomexchange.org/ and http://central.proteomexchange.org/). Additional repositories are likely to be added in the future.

2 The pxr package

`PXDataset` objects

The central object that handles data access is the `PXDataset` class. Such an instance can be generated by passing a valid PX experiment identifier to the `PXDataset` constructor.

```
library("pxr")
id <- "PXD000001"
px <- PXDataset(id)
px
```

```
## Object of class "PXDataset"
##  Id: PXD000001 with 8 files
##  [1] 'F063721.dat' ... [8] 'erwinia_carotovora.fasta'
##  Use 'pxfiles(.)' to see all files.
```

*http://cpu.sysbiol.cam.ac.uk

[1] Vizcaíno J.A. et al. *ProteomeXchange: globally co-ordinated proteomics data submission and dissemination*, Nature Biotechnology 2014, 32, 223 – 226, doi:10.1038/nbt.2839.

Data and meta-data

Several attributes can be extracted from an `PXDataset` instance, as described below.

The experiment identifier, that was originally used to create the `PXDataset` instance can be extracted with the `pxid` method:

```
pxid(px)
```

```
## [1] "PXD000001"
```

The file transfer url where the data files can be accessed can be queried with the`pxurl` method:

```
pxurl(px)
```

```
## [1] "ftp://ftp.pride.ebi.ac.uk/2012/03/PXD000001"
```

The species the data has been generated the data can be obtain calling the `pxtax` function:

```
pxtax(px)
```

```
## [1] "Erwinia carotovora"
```

Relevant bibliographic references can be queried with the `pxref` function:

```
strwrap(pxref(px))
```

```
## [1] "Gatto L, Christoforou A. Using R and Bioconductor for proteomics data analysis."
## [2] "Biochim Biophys Acta. 2013 May 18. doi:pii: S1570-9639(13)00186-6."
## [3] "10.1016/j.bbapap.2013.04.032"
```

All files available for the PX experiment can be obtained with the `pxfiles` method:

```
pxfiles(px)
```

```
## [1] "F063721.dat"
## [2] "F063721.dat-mztab.txt"
## [3] "PRIDE_Exp_Complete_Ac_22134.xml.gz"
## [4] "PRIDE_Exp_mzData_Ac_22134.xml.gz"
## [5] "PXD000001_mztab.txt"
## [6] "TMT_Erwinia_1uLSike_Top10HCD_isol2_45stepped_60min_01.mzXML"
## [7] "TMT_Erwinia_1uLSike_Top10HCD_isol2_45stepped_60min_01.raw"
## [8] "erwinia_carotovora.fasta"
```

The complete or partial data set can be downloaded with the `pxget` function. The function takes an instance of class `PXDataset` as first mandatory argument.

The next argument, `list`, specifies what files to be downloaded. If missing, a menu is printed and the user can select a file. If set as `all`, all files of the experiment are downloaded in the working directory. Alternatively, numerics or logicals can also be used to subset the relevant files to be downloaded based on the `pxfiles(.)` output.

The last argument, `force`, can be set to `TRUE` to force the download of files that already exists in the working directory.

```
pxget(px, "erwinia_carotovora.fasta")
```

```
## Downloading 1 file
```

```
dir(pattern = "fasta")
```

```
## [1] "erwinia_carotovora.fasta"
```

By default, `pxget` will not download and overwrite a file if already available. The last argument of `pxget`, `force`, can be set to `TRUE` to force the download of files that already exists in the working directory.

```
pxget(px, 8)  ## same as above
```

```
## Downloading 1 file
## erwinia_carotovora.fasta already present.
```

Finally, a list of recent PX additions and updates can be obtained using the `pxannounced()` function:

```
pxannounced()
```

```
## 14 new ProteomeXchange annoucements
```

```
##     Data.Set    Publication.Data             Message
## 1  PXD000433 2014-03-27 10:01:25                 New
## 2  PXD000090 2014-03-26 14:43:56                 New
## 3  PXD000379 2014-03-26 14:02:34                 New
## 4  PXD000324 2014-03-21 16:11:11 Updated information
## 5  PXD000323 2014-03-21 16:10:31 Updated information
## 6  PXD000322 2014-03-21 16:10:08 Updated information
## 7  PXD000321 2014-03-21 16:09:46 Updated information
## 8  PXD000320 2014-03-21 16:08:42 Updated information
## 9  PXD000854 2014-03-20 20:58:16                 New
## 10 PXD000556 2014-03-19 13:31:05                 New
## 11 PXD000847 2014-03-18 16:56:00                 New
## 12 PXD000846 2014-03-18 16:55:59                 New
## 13 PXD000845 2014-03-18 16:55:58                 New
## 14 PXD000844 2014-03-18 16:55:57                 New
```

A simple use-case

Below, we show how to automate the extraction of files of interest (fasta and mzTab files), download them and read them using appropriate Bioconductor infrastructure.

```
(mzt <- grep("F0.+mztab", pxfiles(px), value = TRUE))
```

```
## [1] "F063721.dat-mztab.txt"
```

```
(fas <- grep("fasta", pxfiles(px), value = TRUE))
```

```
## [1] "erwinia_carotovora.fasta"
```

```
pxget(px, c(mzt, fas))
```

```
## Downloading 2 files
## erwinia_carotovora.fasta already present.
```

```
library("Biostrings")
readAAStringSet(fas)
```

```
##   A AAStringSet instance of length 4499
##        width seq                                                names
##    [1]   147 MADITLISGSTLGSAEYVAEHLAELLE...EIDITQHQIPEDPAEEWLGSWVNLLK ECA0001 putative
##    [2]   153 VAEIYQIDNLDRGILSALMENARTPYA...IQTIDEIQSTETLISLQNPIMRTIAP ECA0002 AsnC-fami
##    [3]   330 MKKQYIEKQQQISFVKSFFSSQLEQLL...LQLPHIGQVQCGVWPQPLRESVSGLL ECA0003 putative
##    [4]   492 MITLESLEMLLSIDENELLDDLVVTLM...IFDHIWRFDTGLKSRLMRRWQHGKAY ECA0004 conserved
##    [5]   499 MRQTAALAERISRLSHALEHGLYERQH...PSEWLAKIEASLQQVAEQIQQSEQQD ECA0005 conserved
##    ...   ... ...
## [4495]   634 MSDKIIHLTDDSFDTDVLKADGAILVD...EWISVRRKVDPLRVFASDMARRLELL trx-rv3790 trx-rv
## [4496]    93 MTKMNNKARRTARELKHLGASIQTTSL...KPALYRELRDEFPMGYLGDYKDDDDK TimBlower TimBlowe
## [4497]   309 MFSNLSKRWAQRTLSKSFYSTATGAAS...SIWVKKFKWAGIKTRKFVFNPPKPRK sp|P07143|CY1_YEA
## [4498]   231 FPTDDDDKIVGGYTCAANSIPYQVSLN...AQKNKPGVYTKVCNYVNWIQQTIAAN sp|P00761|TRYP_PI
## [4499]   269 GVSGSCNIDVVCPEGNGHRDVIRSVAA...LSDWLDAAGTGAQFIDGLDSTGTPPV sp|Q7M135|LYSC_LY
```

```
library("MSnbase")
(x <- readMzTabData(mzt, "PEP"))
```

```
## Detected a metadata section
## Detected a peptide section
```

```
## MSnSet (storageMode: lockedEnvironment)
## assayData: 1528 features, 6 samples
##   element names: exprs
## protocolData: none
## phenoData
##   rowNames: sub[1] sub[2] ... sub[6] (6 total)
##   varLabels: abundance
##   varMetadata: labelDescription
## featureData
##   featureNames: 1 2 ... 1528 (1528 total)
##   fvarLabels: sequence accession ... uri (14 total)
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
## Annotation:
## - - - Processing information - - -
## mzTab read: Fri Mar 28 02:40:09 2014
```

```
##  MSnbase version: 1.11.12
```

```
head(exprs(x))
```

```
##     sub[1]   sub[2]   sub[3]   sub[4]   sub[5]   sub[6]
## 1 10630132 11238708 12424917 10997763  9928972 10398534
## 2 11105690 12403253 13160903 12229367 11061660 10131218
## 3  1183431  1322371  1599088  1243715  1306602  1159064
## 4  5384958  5508454  6883086  6136023  5626680  5213771
## 5 18033537 17926487 21052620 19810368 17381162 17268329
## 6  9873585 10299931 11142071 10258214  9664315  9518271
```

```
head(fData(x)[, 1:2])
```

```
##    sequence accession
## 1   DGVSVAR   ECA0625
## 2    NVVLDK   ECA0625
## 3 VEDALHATR   ECA0625
## 4 LAGGVAVIK   ECA0625
## 5  LIAEAMEK   ECA0625
## 6 SFGAPTITK   ECA0625
```

3 Session information

- R version 3.1.0 beta (2014-03-26 r65300), `x86_64-unknown-linux-gnu`
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, utils
- Other packages: Biobase 2.23.6, BiocGenerics 0.9.3, Biostrings 2.31.18, IRanges 1.21.36, MSnbase 1.11.12, Rcpp 0.11.1, XVector 0.3.7, ggplot2 0.9.3.1, mzR 1.9.7, pxr 0.99.5
- Loaded via a namespace (and not attached): BiocInstaller 1.13.3, BiocStyle 1.1.18, MASS 7.3-30, RColorBrewer 1.0-5, RCurl 1.95-4.1, XML 3.98-1.1, affy 1.41.4, affyio 1.31.0, codetools 0.2-8, colorspace 1.2-4, dichromat 2.0-0, digest 0.6.4, doParallel 1.0.8, evaluate 0.5.1, foreach 1.4.1, formatR 0.10, grid 3.1.0, gtable 0.1.2, highr 0.3, impute 1.37.1, iterators 1.0.6, knitr 1.5, labeling 0.2, lattice 0.20-27, limma 3.19.28, munsell 0.4.2, mzID 1.1.5, pcaMethods 1.53.4, plyr 1.8.1, preprocessCore 1.25.5, proto 0.3-10, reshape2 1.2.2, scales 0.2.3, stats4 3.1.0, stringr 0.6.2, tools 3.1.0, vsn 3.31.2, zlibbioc 1.9.0