PSICQUIC

Paul Shannon

January 16, 2018

Contents

1 Introduction

PSICQUIC (the Proteomics Standard Initiative Common QUery InterfaCe, pronounced "psy-kick") is "an effort from the HUPO Proteomics Standard Initiative (HUPO-PSI) to standardise the access to molecular interaction databases programmatically". The Bioconductor PSICQUIC package provides a traditional R function-calling (S4) interface layered on top of the PSICQUIC REST interface, to obtain a data.frame of annotated interactions between specified proteins, each of which is typically described by the HUGO symbol of the gene which codes for the protein of interest.

PSICQUIC is loose association of web accessible databases, "providers", linked explicitly only by virtue of being listed at the central PSICQUIC web site. Each provider supports the **MIQL** (molecular interaction query language), and each of which returns standard columns in tab-delimited text. In typical use one queries for all of the interactions in which a protein participates. Equally typical are queries for all known interactions between two specified proteins. These queries are easily constrained by **provider** (e.g., BioGrid or IntAct), by **detectionMethod**, by interaction **type**, and/or by **publicationID**.

Interactions among a set of three or more genes may also be requested. The combinations of possible pairs grows non-linearly with the number of genes, so use this option with care.

PSICQUIC may therefore be best suited to the close study of a few dozen genes or proteins of interest, rather than for obtaining interactions for hundreds or thousands of genes or proteins. For bulk interactions, we recommend that you directly download databases from individual PSICQUIC (or other) providers.

Approximately thirty databases currently implement PSICQUIC. They all

- Support the molecular interaction query language (MIQL)
- Use a controlled vocabulary describing interactions and detection methods
- Communicate via SOAP or REST
- Return results in XML or a tab-delimited form
- May be interogated programmatically or via a URL in a web browser

```
> library(PSICQUIC)
> psicquic <- PSICQUIC()
> providers(psicquic)
```

```
 [1] "APID Interactomes" "BioGrid"          "bhf-ucl"
 [4] "ChEMBL"            "DIP"              "HPIDb"
 [7] "InnateDB-All"      "IntAct"           "IMEx"
[10] "mentha"            "MPIDB"            "MatrixDB"
[13] "MINT"              "Reactome"         "Reactome-FIs"
[16] "EBI-GOA-miRNA"     "I2D"              "UniProt"
[19] "MBInfo"            "BindingDB"        "BAR"
[22] "EBI-GOA-nonIntAct" "ZINC"
```

2   Quick Start: find interactions between Myc and Tp53

A simple example is the best introduction to this package. Here we discover that BioGrid, Intact, Reactome, STRING and BIND each report one or more interactions between human Myc and Tp53:

```
> library(PSICQUIC)
> psicquic <- PSICQUIC()
> providers(psicquic)
```

```
 [1] "APID Interactomes" "BioGrid"          "bhf-ucl"
 [4] "ChEMBL"            "DIP"              "HPIDb"
 [7] "InnateDB-All"      "IntAct"           "IMEx"
[10] "mentha"            "MPIDB"            "MatrixDB"
[13] "MINT"              "Reactome"         "Reactome-FIs"
[16] "EBI-GOA-miRNA"     "I2D"              "UniProt"
[19] "MBInfo"            "BindingDB"        "BAR"
[22] "EBI-GOA-nonIntAct" "ZINC"
```

```
> tbl <- interactions(psicquic, id=c("TP53", "MYC"), species="9606")
> dim(tbl)
```

```
[1] 10 16
```

Note that the several arguments to the *interactions* method are unspecified. They maintain their default values, and act as wildcards in the query.

How many of the approximately twenty-five data sources reported interactions?

```
> table(tbl$provider)
```

```
APID Interactomes           BioGrid             IMEx      InnateDB-All
                2                 1                1                 2
           IntAct      Reactome-FIs           mentha
                1                 1                2
```

What kind of interactions, detection methods and references were reported? (Note that the terms used in the controlled vocabularies used by the PSICQUIC data sources are often quite long, complicating the display of extractions from our data.frame. To get around this here, we extract selected columns in small groups so that the results will fit on the page.)

```
> tbl[, c("provider", "type", "detectionMethod")]
```

```
          provider                                 type
1  APID Interactomes                                    -
2  APID Interactomes                                    -
3          BioGrid psi-mi:MI:0915(physical association)
4      InnateDB-All psi-mi:MI:0915(physical association)
5      InnateDB-All psi-mi:MI:0915(physical association)
6            IntAct            psi-mi:MI:0914(association)
7              IMEx            psi-mi:MI:0914(association)
8            mentha            psi-mi:MI:0914(association)
9            mentha psi-mi:MI:0915(physical association)
10     Reactome-FIs                                    -
                                     detectionMethod
1        psi-mi:MI:0676(tandem affinity purification)
2  psi-mi:MI:0004(affinity chromatography technology)
3  psi-mi:MI:0004(affinity chromatography technology)
4  psi-mi:MI:0004(affinity chromatography technology)
5              psi-mi:MI:0030(cross-linking study)
6        psi-mi:MI:0676(tandem affinity purification)
```

```
7         psi-mi:MI:0676(tandem affinity purification)
8         psi-mi:MI:0676(tandem affinity purification)
9  psi-mi:MI:0004(affinity chromatography technology)
10        psi-mi:MI:0046(experimental knowledge based)
```

These are quite heterogeneous. The well-established "tandem affinity purification" proteomics method probably warrants more weight than "predictive text mining". Let's focus on them:

```
> tbl[grep("affinity", tbl$detectionMethod),
+     c("type", "publicationID", "firstAuthor", "confidenceScore", "provider")]


                            type                 publicationID
1                              -                 pubmed:21150319
2                              -                 pubmed:21150319
3 psi-mi:MI:0915(physical association)          pubmed:21150319
4 psi-mi:MI:0915(physical association)          pubmed:21150319
6        psi-mi:MI:0914(association) pubmed:21150319|imex:IM-16995
7        psi-mi:MI:0914(association) pubmed:21150319|imex:IM-16995
8        psi-mi:MI:0914(association)          pubmed:21150319
9 psi-mi:MI:0915(physical association)          pubmed:21150319
          firstAuthor     confidenceScore            provider
1 Agrawal, P. et al.(2010)                <NA> APID Interactomes
2 Agrawal, P. et al.(2010)                <NA> APID Interactomes
3         Agrawal P (2010)                <NA>           BioGrid
4    Agrawal et al.(2010) lpr:108|hpr:108|np:1      InnateDB-All
6   Agrawal et al. (2010)   intact-miscore:0.35            IntAct
7   Agrawal et al. (2010)   intact-miscore:0.35              IMEx
8                        -   mentha-score:0.236            mentha
9                        -   mentha-score:0.236            mentha
```

This result demonstrates that different providers report results from the same paper in different ways, sometimes omitting confidence scores, and sometimes using different (though related) terms from the PSI controlled vocabularies.

3   Retrieve all Myc interactions found by Agrawal et al, 2010, using tandem affinity purification

These reports of TP53/Myc interactions by detection methods variously described as "affinity chromotography technology" and "tandem affinity purification", both accompanied by a reference to the same recent paper (**"Proteomic profiling of Myc-associated proteins"**, Agrawal et al, 2010), suggests the next task: obtain all of the interactions reported in that paper.

```
> tbl.myc <- interactions(psicquic, "MYC", species="9606", publicationID="21150319")
```

How many were returned? From what sources? Any confidence scores reported?

```
> dim(tbl.myc)
```

```
[1] 2199   16
```

```
> table(tbl.myc$provider)
```

```
APID Interactomes          BioGrid           IMEx      InnateDB-All
          554              107            452              108
         IntAct           mentha
          452              526
```

```
> table(tbl.myc$confidenceScore)
```

```
 intact-miscore:0.35  intact-miscore:0.53  intact-miscore:0.56
            804                   54                   18
 intact-miscore:0.60  intact-miscore:0.64  intact-miscore:0.67
              2                    2                   10
 intact-miscore:0.69  intact-miscore:0.79  intact-miscore:0.98
              6                    2                    6
lpr:108|hpr:108|np:1   mentha-score:0.126   mentha-score:0.236
            108                  296                  166
```

3

```
mentha-score:0.309    mentha-score:0.332    mentha-score:0.416
               1                    18                    25
mentha-score:0.454     mentha-score:0.49    mentha-score:0.554
               4                     3                     4
mentha-score:0.623    mentha-score:0.731     mentha-score:0.74
               1                     1                     1
mentha-score:0.891    mentha-score:0.967       mentha-score:1
               2                     1                     3
```

4   Gene symbols for input, "native" identifers for results

PSICQUIC queries apparently expect HUGO gene symbols for input. These are translated by each provider into each provider's native identifier type, which is nearly always a protein id of some sort. The results returned use the protein identifier native to each provider – but see notes on the use of our IDMapper class for converting these protein identifiers to gene symbols and entrez geneIDs. If you submit a protein identifier in a query, it is apparently used without translation, and the interactions returned are limited to those which use exactly the protein identifier you supplied. Thus the use of gene symbols is recommended for all of your calls to the *interactions* method.

Here is a sampling of the identifiers returned by the PSICQUIC providers:

- refseq:NP_001123512

- uniprotkb:Q16820

- string:9606.ENSP00000373992—uniprotkb:Q9UMJ4

- entrez gene/locuslink:2041—BIOGRID:108355

5   Add Entrez GeneIDs and HUGO Gene Symbols

Though informative, this heterogeneity along with the frequent absence of entrez geneIDs and gene symbols limits the immediate usefulness of these results for many prospective users. We attempt to remedy this with the IDMapper class, which uses biomaRt and some simple parsing strategies to map these lengthy identifiers into both geneID and gene symobl. At this point in the development of the PSICQUIC package, this step – which adds four columns to the results data.frame – must be done explicitly, and is currently limited to human identifiers only. Support for additional species will be added.

```
> idMapper <- IDMapper("9606")
> tbl.myc <- addGeneInfo(idMapper,tbl.myc)
> print(head(tbl.myc$A.name))
```

```
[1] "MYC" "MAX" "MAX" "MAX" "MAX" "MAX"
```

```
> print(head(tbl.myc$B.name))
```

```
[1] "MYC" "MYC" "MYC" "MYC" "MYC" "MYC"
```

6   Retrieve Interactions Among a Set of Genes

If the *id* argument to the *interactions* method contains two or more gene symbols, then all interactions among all possible pairs of those genes will be retrieved. Keep in mind that the number of unique combinations grows larger non-linearly with the number of genes supplied, and that each unique pair becomes a distinct query to each of the specified providers.

```
> tbl.3 <- interactions(psicquic, id=c("ALK", "JAK3", "SHC3"),
+                       species="9606", quiet=TRUE)
> tbl.3g <- addGeneInfo(idMapper, tbl.3)
> tbl.3gd <- with(tbl.3g, as.data.frame(table(detectionMethod, type, A.name, B.name, provider)))
> print(tbl.3gd <- subset(tbl.3gd, Freq > 0))
```

```
                                  detectionMethod
37  psi-mi:MI:0004(affinity chromatography technology)
39                           psi-mi:MI:0493(in vivo)
43  psi-mi:MI:0004(affinity chromatography technology)
45                           psi-mi:MI:0493(in vivo)
154 psi-mi:MI:0004(affinity chromatography technology)
178 psi-mi:MI:0004(affinity chromatography technology)
196 psi-mi:MI:0004(affinity chromatography technology)
326       psi-mi:MI:0046(experimental knowledge based)
332       psi-mi:MI:0046(experimental knowledge based)
442 psi-mi:MI:0004(affinity chromatography technology)
466 psi-mi:MI:0004(affinity chromatography technology)
                                type A.name B.name         provider Freq
37                                 -   JAK3    ALK APID Interactomes    2
39                                 -   JAK3    ALK APID Interactomes    1
43                                 -   SHC3    ALK APID Interactomes    1
45                                 -   SHC3    ALK APID Interactomes    1
154 psi-mi:MI:0915(physical association)   ALK   JAK3          BioGrid    2
178 psi-mi:MI:0915(physical association)   ALK   SHC3          BioGrid    1
196 psi-mi:MI:0915(physical association)     -      -      InnateDB-All    3
326                                -   JAK3    ALK      Reactome-FIs    1
332                                -   SHC3    ALK      Reactome-FIs    1
442 psi-mi:MI:0915(physical association)   ALK   JAK3           mentha    2
466 psi-mi:MI:0915(physical association)   ALK   SHC3           mentha    1
```

7 References

- Aranda, Bruno, Hagen Blankenburg, Samuel Kerrien, Fiona SL Brinkman, Arnaud Ceol, Emilie Chautard, Jose M. Dana et al. "PSICQUIC and PSISCORE: accessing and scoring molecular interactions." Nature methods 8, no. 7 (2011): 528-529.

- Agrawal, Pooja, Kebing Yu, Arthur R. Salomon, and John M. Sedivy. "Proteomic profiling of Myc-associated proteins." Cell Cycle 9, no. 24 (2010): 4908-4921.