BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 8 2008, pages 1100–1101
doi:10.1093/bioinformatics/btm518

Genome analysis

Rintact: enabling computational analysis of molecular interaction data from the IntAct repository

Tony Chiang[1,2,†], Nianhua Li[2,†], Sandra Orchard[1], Samuel Kerrien[1], Henning Hermjakob[1], Robert Gentleman[2] and Wolfgang Huber[1,*]

[1]EBI-EMBL, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2]Computational Biology – FHCRC, 1100 Fairview Avenue North, M2-B876, Seattle, WA 98109, USA

Received on August 14, 2007; revised and accepted on October 10, 2007

Advance Access publication November 7, 2007

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: The *IntAct* repository is one of the largest and most widely used databases for the curation and storage of molecular interaction data. These datasets need to be analyzed by computational methods. Software packages in the statistical environment R provide powerful tools for conducting such analyses.

Results: We introduce *Rintact*, a Bioconductor package that allows users to transform PSI-MI XML2.5 interaction data files from *IntAct* into R graph objects. On these, they can use methods from R and Bioconductor for a variety of tasks: determining cohesive subgraphs, computing summary statistics, fitting mathematical models to the data or rendering graphical layouts. *Rintact* provides a programmatic interface to the *IntAct* repository and allows the use of the analytic methods provided by R and Bioconductor.

Availability: *Rintact* is freely available at http://bioconductor.org

Contact: huber@ebi.ac.uk

1 INTRODUCTION

Protein–protein interaction mapping is a widely used approach to obtain a picture of cellular protein networks. The *IntAct* (Kerrien *et al.*, 2006) database is a primary repository for the publication of molecular interaction data. There are many types of interactions, and each experiment is subject to effects that lead to error, so access to software tools for analysis and visualization is essential.

XML formats are intended for data exchange. They are usually not directly amenable for computational queries nor manipulations, and a transformation into data structures appropriate for the analysis of interest is needed.

We describe the Bionconductor package *Rintact*, which provides a programmatic interface to *IntAct*. It translates the primary data encoded in PSI-MI XML2.5 (Kerrien *et al.*, 2007) files into R graph objects (R Development Core Team, 2007), which can then be analyzed by a variety of computational methods (Barenco *et al.*, 2006; Chiang *et al.*, 2007; Gentleman *et al.*, 2004; Markowetz *et al.*, 2005; Radivoyevitch, 2004; Siek *et al.*, 2000–2001).

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

2 OBTAINING INTERACTION DATA

To illustrate the use of *Rintact*, we access the human CoIP data measured by Ewing *et al.* (2007) and the Y2H data by Stelzl *et al.* (2005). Files can either be downloaded and read from the local file system or read directly from the remote site; we construct the filename vectors for downloaded files:

```
> efiles = sprintf("human_ewing-2007-1_%02d.xml", 1:4)
> sfiles = sprintf("human_stelzl-2005-1_1_%02d.xml", 1:2)
```

and convert the files into R *intactGraph* objects.

```
> ewingG = intactXML2Graph (efiles)
> stelzlG = intactXML2Graph (sfiles)
```

Because both CoIP and Y2H use a bait/prey system, the resulting graph has directed edges from the bait to the prey.

To estimate the translation time of the function `intactXML2Graph`, we applied it to seven separate datasets from Utez *et al.* (2000) (two datasets), Cagney *et al.* (2001), Giot *et al.* (2003), Stelzl *et al.* (2005), Zhao *et al.* (2005) and Ewing *et al.* (2007). The data vary in size, and we found the general trend suggests a linear time algorithm based on the number of interactions. Thus *Rintact* provides a feasible approach in parsing the *IntAct* PSI-MI XML2.5 files.

IntAct uses internal, persistent identifiers called *IntAct* accession codes to unify the various identifier schemes of submitted datasets. The PSI-MI XML2.5 files provide translation information from the contained *IntAct* accession codes to various other commonly used molecule identifiers. This information allows the rendering of the interaction datasets using different types of molecule identifiers.

```
> ID = nodes(ewingG)[c(1, 45)]
> translateIntactID(ewingG, ID, c("geneName",
  "uniprotId"))
```

```
                geneName uniprotId
EBI-1003700 ''CENPH'' ''Q9H3R5''
EBI-1046072 ''PPP4C'' ''P60510''
```

The function `intactXML2Graph` can also be called on protein complex membership XML files, and the structure of the output is an `intactHyperGraph`. The relationship between proteins in multi-protein complexes is not binary;

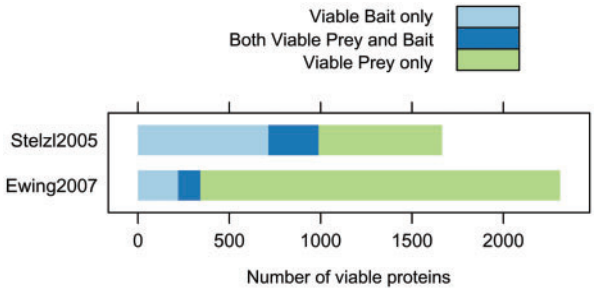

Fig. 1. The Bar chart shows the viable bait and prey distributions of the two datasets.

each protein complex can be represented as a hyperedge, and so the collection of protein complexes is a hypergraph.

3 COMPUTATIONAL ANALYSIS

After obtaining the molecular interaction data, we can exploit the various statistical methods in R and Bioconductor. For example, we can identify the densely connected subgraphs in Ewing *et al.*'s data using the `highlyConnSG` function from the *RBGL* package. Since `highlyConnSG` takes an undirected graph without self-loops, we first need to call the functions `ugraph` and `removeSelfLoops` on the directed data graph.

```
> g1 = removeSelfLoops(ugraph(ewingG))
> hc1 = highlyConnSG(g1)
```

A graph G with n vertices is *highly connected* if removal of any set of less than $n/2$ vertices does not disconnect G. Calling the `length` function on the first element of `hc1` enumerates the number of highly connected subgraphs at 328, of which the largest has 640 vertices.

We can use the package *ppiStats* to compute summary statistics. Defining a *viable prey* (VP) as a protein that was found as a prey at least once in a given dataset (*viable bait* (VB) and viable bait/prey (VBP) are defined analogously (Chiang *et al.*, 2007), we can produce the bar chart in Figure 1. It shows that Stelzl *et al.*'s (2005) Y2H data had a comparable number of viable baits to viable prey while in Ewing *et al.*'s (2007) CoIP experiment the viable prey population is larger than that of the viable baits.

We can view a subset of the CoIP data by rendering the subgraph induced by 10 baits and the group of preys they pull down in Figure 2 using *Rgraphviz*, and so we can easily see the clustering effects of the CoIP technology. *Rintact* can also work with the *STRING* database and the *Cytoscape* software via the *Gaggle* (Shannon *et al.*, 2006) Bioconductor package. Other annotations can be obtained via the *biomaRt* (Durinck *et al.*, 2005) Bioconductor package.

4 DISCUSSION

We have shown the capabilities of *Rintact*. While there are several software tools that also read PSI-MI XML2.5 files, *Rintact* has the additional benefit of being a computational conduit between *IntAct* and the analytic methods found in R and Bioconductor. *Rintact* provides an efficient and straightforward approach towards the analysis of molecular interaction data.

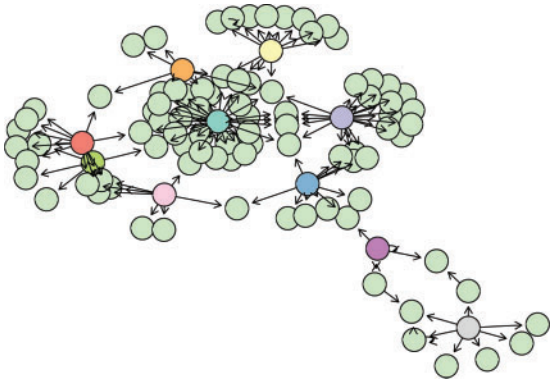

Fig. 2. The CoIP subgraph restricted to 10 baits and their pulldowns. Each selected bait is rendered in a unique color while all the prey are rendered in light green.

ACKNOWLEDGEMENTS

We would like to thank Abhishek Pratap and Li Wang for testing the *Rintact* software package. We acknowledge funding through the HFSP Grant RGP0022/2005 to W.H. and R.G. and NIH Research 1P41HG004059 to R.G.

Conflict of Interest: none declared.

REFERENCES

Barenco,M. *et al.* (2006) Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biol.,* **7**.

Cagney,G. *et al.* (2001) Two-hybrid analysis of the *Saccharomyces cerevisiae* 26s proteasome. *Physiol. Genomics,* **7**, 27–34.

Chiang,T. *et al.* (2007) Coverage and error models of protein-protien interaction data by directed graph analysis. *Genome Biol.,* **8**.

Durinck,S. *et al.* (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.

Ewing,E.M. *et al.* (2007) Large-scale mapping of protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.,* **3**.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.,* **5**, R80.

Giot,L. *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.

Kerrien,S. *et al.* IntAct – open source resource for molecular interaction data. *Nucleic Acids Res.,* **35**, D561–D565.

Kerrien,S. *et al.* (2007) Broadening the horizon – level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.*

Markowetz,F. *et al.* (2005) Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics,* **21**, 4026–4032.

Development Core Team,R. (2007) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Radivoyevitch,T. (2004) A two-way interface between limited systems biology markup language and R. *BMC Bioinformatics,* **5**, 190–190.

Shannon,P. *et al.* (2006) The Gaggle: an open-source software system for integrating bioinformatics software and data sources. *BMC Bioinformatics,* **7**.

Siek,J, *et al.* (2000–2001) *The Boost Graph Library*. Cambridge University Press, Cambridge.

Stelzl,U. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell,* **122**, 957–968.

Peter Uetz, *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature,* **403**, 623–627.

Zhao,R. *et al.* (2005) Navigating the chaperone network: an integrative map of physical and genetic interactions mediated by the Hsp90 chaperone. *Cell,* **120**, 715–727.