Rintact analysis

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

Tony Chiang1,2,†, Nianhua Li2,†, Sandra Orchard1, Samuel Kerrien1, Henning Hermjakob1, Robert Gentleman2 and Wolfgang Huber1,*

1EBI-EMBL, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and 2Computational 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

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 Bioconductor package Rintact, which provides a programmatic interface to IntAct. It translates the primary data encoded in PSI-MI XML2.5 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; Markowitz et al., 2005; Radivojевич, 2004; Siek et al., 2000–2001).

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:

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

and convert the files into R `intactGraph` objects.

```r
> ewingG = intactXML2Graph (efiles)
> stelzG = 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.

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

```r

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;

*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.
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.

```r
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 RBGL with the clustering effects of the CoIP technology.

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

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

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

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


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


