PAnnBuilder: an R package for assembling proteomic annotation data

Hong Li1,2,3, Guohui Ding1, Lu Xie2,* and Yixue Li1,2,*

1Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, 2Shanghai Center for Bioinformation Technology, Shanghai 200235 and 3Graduate School of the Chinese Academy of Sciences, 19 Yuquan Road, Beijing 100039, P.R.China

Received on October 26, 2008; revised on January 6, 2009; accepted on February 16, 2009

1 INTRODUCTION

In the post-genomic era, large numbers of genes and proteins are linked to biological function through annotation information. As high-throughput experimental technologies become widely applied, the easy accessibility of relevant biological annotation becomes increasingly desirable. Currently, annotation information is usually available in various file formats within diverse public databases subject to constant updating. Assembling annotation for omics data thus requires downloading, parsing and formatting data files from different resources which can be error-prone and laborious. The AnnBuilder package of the Bioconductor project (Gentleman et al., 2004) set a good example for automated genomic data annotation. PAnnBuilder is an open-source package for the programming environment R. It is available from the http://www.biosino.org/PAnnBuilder/download.jsp and runs on Linux, Mac OS or MS-Windows. The aim of PAnnBuilder is to build proteomic annotation data packages. Additional step-by-step manual can be acquired from the PAnnBuilder vignette. For compatibility with the Bioconductor framework, PAnnBuilder inherits some functions from other Bioconductor packages (Gentleman et al., 2004), especially from the packages AnnBuilder (Zhang et al., 2003) and AnnotationDbi. Perl and BLAST are also required for data processing and sequence alignment.

2 DESCRIPTION

PAnnBuilder is a free software package which can be used in any operating system and make subsequent statistical analysis easy and efficient.

PAnnBuilder downloads data files from relevant websites, parses data using Perl scripts, saves data using SQLite-based database, compiles help documents and finally builds a uniform R annotation data package. Once the package has been built, annotation can be quickly accessed by protein identifiers, and various statistical analyses can be performed. PAnnBuilder automates protein-centric annotation for a uniform output.

2.1 Building packages from selected resources

In Building annotation data packages for proteins, PAnnBuilder is complementary to AnnBuilder and AnnotationDbi, which assemble genomic annotation data. PAnnBuilder currently has 20 functions to support 16 databases using R environment or SQLite technology (Table 1).

*pTo whom correspondence should be addressed.

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
Table 1. PAnnBuilder functions

<table>
<thead>
<tr>
<th>Annotation type</th>
<th>Source database</th>
<th>PAnnBuilder function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Available annotations</td>
<td>IPI, Uni-Prot, RefSeq</td>
<td>pBaseBuilder or pBaseBuilder_DB</td>
</tr>
<tr>
<td>from primary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>protein databases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structural classification of proteins</td>
<td>SCOP</td>
<td>scpBuilder or scpBuilder_DB</td>
</tr>
<tr>
<td>Subcellular location</td>
<td>DBSubLoc; BaCeLLo</td>
<td>subcellBuilder or subcellBuilder_DB</td>
</tr>
<tr>
<td>Protein–protein or</td>
<td>IntAct; MPPi; DOMINE</td>
<td>intBuilder or intBuilder_DB</td>
</tr>
<tr>
<td>domain–domain</td>
<td>3DID; NCHB; SynPDM (submission)</td>
<td>pmBuilder or pmBuilder_DB</td>
</tr>
<tr>
<td>interaction</td>
<td>Sys-BodyFluid</td>
<td>bBuilder or bBuilder_DB</td>
</tr>
<tr>
<td>Protein in body fluids</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peptide</td>
<td>PeptideAtlas or PeptideAtlasBuilder</td>
<td></td>
</tr>
<tr>
<td>Gene ontology</td>
<td>GOA or GOABuilder</td>
<td></td>
</tr>
<tr>
<td>Ortholog</td>
<td>InParanoid or InParanoidBuilder</td>
<td></td>
</tr>
<tr>
<td>Homolog</td>
<td>HomoloGene or HomoloGeneBuilder</td>
<td></td>
</tr>
</tbody>
</table>

In addition to the three widely used primary protein databases mentioned above (i.e. IPI, Uni-Prot and RefSeq), more specific secondary databases often collect protein-specific annotation information, as well. PAnnBuilder also provides functions to build packages from widely used secondary databases (Table 1). For example, subcellBuilder/subcellBuilder_DB support two subcellular location databases (DBSubLoc; BaCeLLo) and is specifically used for obtaining protein location information.

2.2 Extended functions of PAnnBuilder

In addition to providing automated and uniformed annotation, PAnnBuilder includes some extended functions. crossBuilder/crossBuilder_DB can map identifiers from different resources, which may be useful for integrative analysis among different studies. pSeqBuilder/pSeqBuilder_DB can assign annotation to poorly studied proteins based on user-defined sequence similarity with known proteins. PAnnBuilder also may be used to develop other application tools for sophisticated, researcher-specific needs.

3 EXAMPLE

To allow quick use of PAnnBuilder, protein-centric annotation packages for model species (Homo sapiens, Mus musculus and Rattus norvegicus) have been constructed and cross-mapping among multiple protein identifiers (IPI, Uni-Prot and RefSeq) have been performed. Annotation packages aimed at secondary databases also have been constructed. In total, 54 packages are available for download at http://www.biosino.org/PAnnBuilder/example.jsp.

In the following example, pBaseBuilder is employed to build an R annotation package ‘org.Hs.angi’, to annotate proteins from the HUMAN IPI database:

```r
pBaseBuilder(baseMapType="ipi",organism="Homo sapiens", pkgName="org.Hs.angi", pkgPath=tempdir(), version="1.1.0", author=list(author="Hong Li",maintainer="Hong Li <lili@stbs.ac.cn>")
```

Users should be aware that downloading, parsing and saving data may take a long time, in addition to requiring enough disk space to store temporary data files. After data have been processed, the directory org.Hs.angi will be produced in tempdir()), and the command ‘R CMD build’ can be used to build R packages. In order to list all available annotation data in ‘org.Hs.angi’ package, the following commands are run:

```r
library(org.Hs.angi)
lst("package:org.Hs.angi")
```

The output will map IPI protein identifiers to Entrez gene ID, gene symbol, KEGG pathway, gene ontology, domain and so on.

4 DISCUSSION

PAnnBuilder allows proteomic annotation data packages to be built entirely in R and can be run on any operating system. Product packages are standardized to conform to other Bioconductor meta-data annotation packages, with data saved via R environment objects or SQLite-based database and detailed document available via help pages. In the future, more functions may be added into PAnnBuilder package for other databases.

PAnnBuilder is extremely useful for large-scale annotation of given protein sets, facilitates subsequent statistical analysis, and can be used for heterogeneous proteomic data integration and meta-analysis. General proteomic projects may employ PAnnBuilder to assemble annotation or build annotation package to suit any requirements. PAnnBuilder may become an important tool for proteomic research.

ACKNOWLEDGEMENTS

We acknowledge Lynne Berry from the Vanderbilt Cancer Biostatistics Center for her editing work.

Funding: National High-Tech R & D Program of China (2006AA02Z334, 2007AA02Z3304); National Basic Research Program of China (2009CB918404, 2006CB910700, 2004CB720103, 2004CB518606); Key Research Program (CAS) KSCX2-YW-R-112; Shanghai Natural Science Foundation (08ZR141580).

Conflict of Interest: none declared.

REFERENCES

