Bioinformatics Advance Access published February 23, 2009

*Data and Text Mining*

# PAnnBuilder: An R Package for Assembling Proteomic Annotation Data

Hong Li[1,2], Guohui Ding[1], Lu Xie[2*] and Yixue Li[1,2*]

[1]Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, P. R. China.

[2]Shanghai Center for Bioinformation Technology, Shanghai 200235, P. R. China.

Associate Editor: Prof. John Quackenbush

**ABSTRACT**

**Summary:** PAnnBuilder is an R package to automatically assemble protein annotation information from public resources to provide uniform annotation data for large-scale proteomic studies. Sixteen public databases have been parsed and fifty-four annotation packages have been constructed based on R environment or SQLite database. These ready-to-use packages cover most frequently needed protein annotation for three model species including human, mouse and rat. Several extended applications such as annotation based on protein sequence similarity are also provided. Sophisticated users can develop their own packages using PAnnBuilder. PAnnBuilder may become an important tool for proteomic research.

**Availability:** PAnnBuilder and example annotation packages are freely available from http://www.biosino.org/PAnnBuilder/

**Contact:** xielu@scbit.org and yxli@sibs.ac.cn

## 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 (Zhang, et al., 2003) in R (R Development Core Team, 2008). Protein annotation, however, has not yet been specifically addressed, despite its importance for post-genomic functional studies. The requirements for protein annotation differ from those of genomic annotation, due to protein-specific features such as subcellular location and post-translational modifications. On account of this, we developed an R package PAnnBuilder to assemble proteomic annotation data.

The PAnnBuilder package automatically obtains protein-related annotation information from diverse databases and builds proteomic annotation data packages. The products of PAnnBuilder are uniform R annotation data packages, which can be used in any operating system and make subsequent statistical analysis easy and efficient.

## 2 DESCRIPTION

PAnnBuilder is an open-source package for the programming language and statistical 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–to-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.

The user chooses suitable function in PAnnBuilder to build annotation package from the protein database and organism of his interest. PAnnBuilder downloads data files from relevant websites, parses data using Perl scripts, saves data using *environment* objects or 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).

***pBaseBuilder*** and ***pBaseBuilder_DB,*** the most important PAnnBuilder functions, map protein IPI, Uni-Prot, or RefSeq identifiers to nearly all available annotations from these three primary

*To whom correspondence should be addressed.

databases and from other functional annotation databases. ***pBaseBuilder_DB,*** an improved version of ***pBaseBuilder,*** builds SQLite-based annotation packages instead of environment-based packages. SQLite-based annotation package uses a real database as the ultimate data structure, which provides advantages in data accessing, reverse mapping, filtering, and combining.

In addition to the three widely used primary protein databases mentioned above (i.e., IPI, Uni-Prot, 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.

**Table 1.** PAnnBuilder functions.

| Annotation type | Source database | PAnnBuilder function |
|---|---|---|
| available notations from primary protein databases | IPI; Uni-Prot; RefSeq | pBaseBuilder or pBaseBuilder_DB |
| structural classification of proteins | SCOP | scopBuilder or scopBuilder_DB |
| subcellular location | DBSubLoc; BaCeLLo | subcellBuilder or subcellBuilder_DB |
| protein-protein or domain-domain interaction | IntAct; MPPI; 3DID; NCBI; DOMINE; | intBuilder or intBuilder_DB |
| modification | SysPTM (in submission) | ptmBuilder or ptmBuilder_DB |
| protein in body fluids | Sys-BodyFluid | bfBuilder or bfBuilder_DB |
| peptide | PeptideAltas | PeptideAtlasBuilder or PeptideAtlasBuilder_DB |
| gene ontology | GOA | GOABuilder or GOABuilder_DB |
| ortholog | InParanoid | InParanoidBuilder or InParanoidBuilder_DB |
| homolog | HomoloGene | HomoloGeneBuilder or HomoloGeneBuilder_DB |

### 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, RefSeq) has 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.ipi"*, to annotate proteins from the HUMAN IPI database:

***pBaseBuilder(baseMapType="ipi",organism="Homo sapiens", pkgName="org.Hs.ipi", pkgPath=tempdir(), version="1.1.0", author=list(author="Hong Li",maintainer="Hong Li <lihong@sibs.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.ipi* 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.ipi" package*, the following commands are run:

***library(org.Hs.ipi)***
***ls("package:org.Hs.ipi")***

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 metadata 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 for this research was provided by National Basic Research Program of China: 2006CB910700, 2004CB720103, 2004CB518606 and Shanghai Natural Science Foundation 08ZR1415800.

## REFERENCES

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y. and Zhang, J. (2004) Bioconductor: open software development for computational biology and bioinformatics, Genome Biol, 5, R80.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Zhang, J., Carey, V. and Gentleman, R. (2003) An extensible application for assembling annotation for genomic data, Bioinformatics, 19, 155-156.