BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 1 2003
Pages 155–156

An extensible application for assembling annotation for genomic data

Jianhua Zhang*, Vincent Carey and Robert Gentleman

Department of Biostatistical Science, Dana-Farber Cancer Institute, Harvard School of Public Health, and Channing Laboratory, Boston, MA, USA

Received on May 2, 2002; revised on July 19, 2002; accepted on July 24, 2002

ABSTRACT

Summary: AnnBuilder is an R package for assembling genomic annotation data. The system currently provides parsers to process annotation data from LocusLink, Gene Ontology Consortium, and Human Gene Project and can be extended to new data sources via user defined parsers. AnnBuilder differs from other existing systems in that it provides users with unlimited ability to assemble data from user selected sources. The products of AnnBuilder are files in XML format that can be easily used by different systems.

Availability: (http://www.bioconductor.org). Open source.

Contact: jzhang@jimmy.harvard.edu

INTRODUCTION

One of the major challenges in the post–genomic era is assembling and providing useful access to biological annotation of genomic data. The described software builds such assemblies using different sources. Our need for such a tool arose from interest in DNA microarray experiments. The analysis of such experiments is limited by both the availability of relevant biological annotation and the difficulties inherent in associating that data with the relevant experimental data. AnnBuilder can be used to assemble annotation data for many different purposes.

DESCRIPTION

Design

Assembling gene annotation data into a useful format for programmatic manipulation is difficult for many reasons. Annotation data, though typically accessible via WWW, are stored in various files in various formats, including HTML, XML, plain text files or spreadsheets. The data consist of associations between one (or more) systematic identifiers and the biological data. To process such data a mapping from a set of known identifiers into those associated with the data of interest is required. While most mappings are of a fairly standard nature (such as one to one, one to many etc.) there are others with less standard structures such as a directed acyclic graph (DAG) for gene ontology terminologies. Additionally, when mapping information is available from multiple sources it will usually be beneficial to consider all of them. Finally, since these meta data change and are enhanced constantly the system should facilitate frequent assemblage of data. AnnBuilder was designed to be run frequently with different data sources.

The first step in our process is a user specified reduction of the data to that which is deemed relevant for a set of analyses. Database technology (Postgres) is then applied to the issues that it is designed for and we rely on the more flexible processing capabilities of Perl and R (Ihaka and Gentleman, 1996) for other data handling and for communication between the tools being used.

Manufacturers of chips or other primary experimental components typically use specific identifiers for the genes or ESTs that have been used as probes. The supplier generally supplies a mapping to at least one publicly available system. We chose to map the supplied identifiers, initially, to LocusLink (Pruitt and Maglott, 2001) identifiers since LocusIDs are widely used and hence provide a good point of linkage.

The mapping to LocusLink identifiers is a crucial step and may be done using several different sources. We have implemented a system that takes advantage of the existence of multiple sources and provides a unified mapping based on the available sources. This approach helps ensure comprehensive and reliable mappings.

When multiple data sources are available, mappings from the source identifier (for example an Affymetrix probe label) to the destination identifier are made separately using each data source. These mappings are then stored in a table where the columns represent the different data sources. The mapping selected depends on the concordance of the different sources. If only one source provides a mapping or all sources concur, then that mapping is selected. When sources are discordant the mapping can be established by a possibly weighted vote.

Once that unified mapping to primary identifiers has

*To whom correspondence should be addressed.

been obtained mappings to other sets of identifiers can be made. These other mappings may also be available from multiple sources and the same procedures described above may be beneficially applied to them.

Non-standard Data Structures

The Gene Ontology Consortium (2000) provides a standardized nomenclature for subcategories of the basic ontological categories *molecular function, biological process* and *cellular component*. GO terms form a directed acyclic graph (DAG) with one root node. A gene (or EST) is associated with a specific node in the DAG but the implication is that it is also a member of all parent nodes.

We felt that it was likely that researchers would like either to concentrate on nodes which were of particular interest or on nodes that had a reasonable prevalence in their data. For these reasons we provide two mechanisms for node selection. With *number-based selection* only nodes with EST counts above a given threshold value will be selected from the DAG. For *term-based selection* a user-specified set of GO terms and their children will be selected. Data contained by each node are then output.

An example

We now consider the application of this methodology to mapping from Affymetrix identifiers from the U95v2 A chip to LocusLink identifiers. The mapping was carried out in early March 2002 so the results are likely different by the time this is being read. Affymetrix provides a mapping from the ESTs used to GenBank accession numbers or TIGR identifiers that can be mapped to GenBank accession numbers. We then identified three data sources for mapping from the GenBank accession numbers to LocusIDs. They are the UniGene data file (ftp://ftp.ncbi.nih.gov/repository/UniGene/Hs.data.gz), the LocusLink data file (ftp://ftp.ncbi.nih.gov/refseq/LocusLink/LL_tmpl.gz), and UMich. data file (http://dot.ped.med.umich.edu:2000/ourimage/microarrays/Affy_annot/Unigene/index.html) for the target chip.

UniGene provided mappings for 10525, LocusLink provided mappings for 8569, and UMich. provided mappings for 11292 of the 12625 ESTs on the chip. The three sources agreed on 7341 of the mappings. UniGene provided 119 mappings, LocusLink provided 108 mappings, and UMich provided 426 mappings that were not provided by the other two sources. The remaining 3754 constituted 3751 mappings agreed by two of the three sources and 3 *disagreements* among the three sources.

The unified mapping was then used as the base to process the LocusLink and Gene Ontology data and generate the XML files (http://www.bioconductor.org/datafiles/dtds/annotate.dtd) for annotation. Parsing of the XML files can be done using the existing functions in R or other XML parsers.

Usage

AnnBuilder provides a step by step manual on how to build an annotation file using AnnBuilder.

Discussion

AnnBuilder provides a tool for reliably assembling data related to the annotation of genomic data. The power of utilizing multiple sources to associate data in an exhaustive and reliable way well suits the situation of genomic data with scattered distribution across various sources. In this paper we used microarray data annotation as an example, but the tool is potentially applicable to any situation where annotation is sought for a given set of identifiers that can be linked to other known identifiers in public databases.

The end products of AnnBuilder are assemblies of genomic meta-data. These are exported as XML files that can be processed easily with existing tools. We further process these data into an R format suitable for use with the *annotate* package that is part of the Bioconductor project. The constructed database can be accessed directly. This option will provide greater flexibility in usage.

Annbuilder can perform a variety of roles. It can be used by an institution or lab to assemble data from a variety of sources into a more suitable format for the specific analyses being performed. In particular such labs may want to incorporate local (non–public) data for internal use. It can also be used by projects such as dChip (http://www.dchip.org), Gifi Array Analyzer (http://biowww.dfci.harvard.edu/∼ycui/Gifi.html)) to assemble annotation suitable for their needs. Manufacturers of data such as Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) and NetAffx (http://www.affymetrix.com/products/netaffx.html) could also employ Annbuilder as a tool to help assemble the data they present.

ACKNOWLEDGEMENTS

Support for the development of this software came from the High Tech Industry Multidisciplinary Research Fund at the Dana-Farber Cancer Institute.

REFERENCES

Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graphic Statist.*, **5**, 299–314.

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.

The Gene Ontology Consortium (2000) Gene Ontology:tool for the unification of biology. *Nature Genet.*, **25**, 25–29.