BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 10 2004, pages 1651–1652
doi:10.1093/bioinformatics/bth133

matchprobes: a Bioconductor package for the sequence-matching of microarray probe elements

Wolfgang Huber[1,*] *and Robert Gentleman*[2]

[1]*Department of Molecular Genome Analysis, German Cancer Research Center, INF 580, Heidelberg, 69120, Germany and* [2]*Department of Biostatistical Science, Dana Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, USA*

Received on July 27, 2003; revised on December 29, 2003; accepted on January 5, 2004
Advance Access publication February 26, 2004

ABSTRACT

Summary: The nucleotide sequences of the probes on a microarray can be used for a variety of purposes in the analysis of microarray experiments. We describe software and a paradigm for the creation of data packages for curating, distributing and working with probe sequence data in a uniform, across-types-of-microarrays manner. While the implementation is specific to the Bioconductor project, the ideas and general strategies are more general and could be easily adopted by other projects.

Availability: The R package matchprobes is available under LGPL at http://www.bioconductor.org

Contact: w.huber@dkfz.de

Supplementary information: The package contains documentation in the form of a vignette and manual pages.

MOTIVATION

The probes on a DNA microarray may be short oligonucleotides (e.g. 25mers on Affymetrix genechips), long oligonucleotides (e.g. 60mers on Agilent Oligo Microarrays) or PCR-amplified fragments (spotted cDNA arrays). One use of probe sequence information is analyses of hybridization intensities that take into account the physico-chemical properties of different probe sequences (Wu *et al*., 2003; Zhang *et al*., 2003). Furthermore, it is possible to combine data sets obtained on different types of microarrays based on probe sequence similarity. For short oligonucleotide arrays, the probe sequence information can be used to query for matches to transcripts of interest that were not intentionally arrayed. For example, these could be new genes or alternative splice forms. A reasonable number of matches, or near matches, may be sufficient for detecting changes in the abundance of the transcript of interest.

Our motivation for developing *matchprobes* was providing a set of easy-to-use tools to perform calculations and comparisons on probe sequence information and to integrate these aspects of microarray data analysis with other, more numerical aspects, such as normalization and statistical analysis. We wanted to provide this functionality in a relatively uniform format to make it easier to incorporate with other software packages or data processing steps. This makes the process as transparent as possible with respect to the type of microarray being used.

We also felt that the solution must provide tools for detecting and downloading available data, or updates, with little or no direct human intervention. As the models and manipulations of data required for analysing genomic data become increasingly complex, it is essential that the need for human intervention be reduced, and methodology that simplifies the data delivery and acquisition process is becoming increasingly important. Providing data in a web services paradigm is a step in this direction.

To varying extents, the relevant sequence information is provided by the chip manufacturer. For example, Affymetrix Corporation provides extensive information on the sequences of short oligonucleotide probes and the intended mapping on the transcriptome on their website. For spotted cDNA arrays, the availability and format of such information depends on the protocols used by the manufacturer.

DESIGN

Our basic design decision was to structure the meta-data into packages that are relevant for specific chips or instruments used for the data collection. A package contains the meta-data themselves, methods to access and manipulate them, documentation and provenance information. Compared to a global approach, this has great advantages: chip- or technology-specific aspects can be dealt with on a per-package basis, unnecessary complexity is avoided, and the size of the meta-data is substantially reduced. Versioning information can be included, standard mechanisms for distribution and documentation of software packages can be used, and data formats can be standardized.

Potentially, this approach could also have disadvantages: because there are many different chips, there will be a large number of different packages, and these could diverge with respect to user interface, functionality and quality of implementation. To ensure standardization, it is preferable to have these data packages generated by software. We are proposing the development of software that writes software. The benefits

*To whom correspondence should be addressed.

of this approach are immediate and manifold. First, the modules are uniform to an extent that would not be possible if the packages were human written. Users of this technology need only become acquainted with one package to be acquainted with all such packages. Second, this leads to substantial labor savings, and we can create many packages very quickly. Third, this strategy naturally deals with the progressive nature of the biological meta-data. The automated package generation process can proceed in a regular and timely fashion. The complete data repository can be updated either on demand or on a scheduled basis.

IMPLEMENTATION

The Bioconductor project (Gentleman *et al*., 2003) is a software project for bioinformatics that is largely based on the R language (Ihaka and Gentleman, 1996). R's standard package creation and testing tools provide the means for creating and testing the data packages. The facilities provided by the *reposTools* (Gentry and Gentleman, 2003, http://www.bioconductor.org) package allow distribution, versioning and automatic updating in a web services fashion.

There are two aspects to the software provided in *matchprobes*. One is the technology for taking a specific data source, containing probe sequence information, and transforming it into an R package that can be distributed and widely used. The second aspect is a basic set of string matching and manipulation tools. The package currently contains fast algorithms for string and regular expression matching, for reversing sequences and for obtaining complementary sequences. External sequence alignment methods (e.g. a Smith–Waterman algorithm) can be plugged in.

Data package structure

One data package is produced for each type of microarray. A user needs to obtain both the appropriate data package and *matchprobes* before carrying out computations. The data are kept in a rectangular table in which rows correspond to the different probes on the array and columns to the different kinds of information that are provided about the probes. The table is wrapped into an R package, which allows versioning, documentation and remote distribution using well-established mechanisms. For the most common Affymetrix genechips, such data packages are available for download by following the link 'MetaData' on the Bioconductor website.

Data package creation

Data packages for an array type that is not available from the Website can be created easily by the user herself. In the simplest case, the R command looks like

```
makeProbePackage("HG-U95Av2",
 maintainer = "wh <w.huber@dkfz.de>",
 version    = "1.0")
```

This assumes that a file named `HG-U95Av2_probe_tab`, which contains the probe sequence information for one of Affymetrix's genechips, has been downloaded from the manufacturer's Website, http://www.affymetrix.com/support, and is available in the working directory. To deal with different file formats and additional types of probe annotation data from public or in-house databases, the function `makeProbePackage` offers a great deal of flexibility. The user can specify her own import function through the argument `importfun`. By default, its value is `getProbeDataAffy`, a function that reads tabular Affymetrix genechip sequence files. Import functions for other types of arrays can be adapted from this prototype.

The help pages and R code contained in the packages produced are derived from a template directory that obeys the usual R package conventions (R Foundation, 1999, http://www.r-project.org). A prototype for such a directory is provided within the package *matchprobes*. To facilitate the automated production of large numbers of similar packages, we provide a text substitution mechanism similar to the one used in the GNU configure system. A more detailed description of the automatable package creation process is given in the vignette.

Matching functions

There are functions for finding complement, reverse and mismatch sequences and for counting nucleotide content. There is a fast string matching function that searches for matches of the probe sequences on the microarray to a query sequence which may, e.g. represent a transcript of interest. In addition, R's built-in regular expressions can be used.

CONCLUSION

We have described a method for employing probe sequence information in the context of microarray analyses. While the ideas are designed for the Bioconductor project, they are easily adapted to other projects and needs. In the process, we have also developed a mechanism for the mass production of data packages for genomic research, such as probe sequence information and other meta-data for microarrays.

REFERENCES

Gentleman,R., Carey,V., Dudoit,S., Ellis,B., Gautier,L., Gentry,J., Huber,W., Irizarry,R.A., Rossini,A.J., Smyth,G.K. and Zhang,J. (2003) The Bioconductor Project. *Technical Report*, Dana Farber Cancer Institute, Boston, MA, USA.

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

R Foundation (1999) Writing R Extensions.

Gentry,J. and Gentleman,R. (2003) The *reposTools* package: repository tools for R.

Wu,Z., Irizarry,R., Gentleman,R., Murillo,F.M. and Spencer,F. (2003) A model based background adjustment for oligonucleotide expression arrays. *Technical Report*, Johns Hopkins University, Baltimore, MD, USA.

Zhang,L., Miles,M.F. and Aldape,K.D. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.*, **21**, 818–821.