rTANDEM, an R/Bioconductor package for MS/MS protein identification

Frédéric Fournier¹,*, Charles Joly Beauparlant¹, René Paradis¹ and Arnaud Droit¹,²,*
¹Proteomics Center and ²Department of Molecular Medicine, CHUQ Research Center, Laval University, Quebec G1V 4G2, Canada
Associate Editor: Igor Jurisica

ABSTRACT

Summary: rTANDEM is an R/Bioconductor package that interfaces the X!Tandem protein identification algorithm. The package can run the multi-threaded algorithm on proteomic data files directly from R. It also provides functions to convert search parameters and results to/from R as well as functions to manipulate parameters and automate searches. An associated R package, shinyTANDEM, provides a web-based graphical interface to visualize and interpret the results. Together, those two packages form an entry point for a general MS/MS-based proteomic pipeline in R/Bioconductor.

Availability and implementation: rTANDEM and shinyTANDEM are distributed in R/Bioconductor, http://bioconductor.org/packages/release/bioc/. The packages are under open licenses (GPL-3 and Artistic-1.0).

Contact: frederic.fournier@crchuq.ulaval.ca or arnaud.droit@crchuq.ulaval.ca

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on December 5, 2013; revised on March 24, 2014; accepted on March 28, 2014

1 INTRODUCTION

Protein identification and analysis by tandem mass spectrometry relies mostly on matching spectra to a database of protein sequences and scoring those matches. The resources of the R statistical language and its platform Bioconductor (Gentleman et al., 2004) have proven to be well suited for this kind of data manipulation and statistical scoring with regard to genomic data. However, those resources remain largely untapped with regard to proteomic data. This is partly explained by the lack of tools to perform protein identification directly from R. rTANDEM fills this gap by implementing in R the tandem algorithm (Craig and Beavis, 2004) and various associated scoring functions (Keller et al., 2005; Kertesz-Farkas et al., 2014; MacLean et al., 2006). rTANDEM also provides functions for conversion to/from R. This brings to proteomics the many advantages of building an analysis pipeline in the R/Bioconductor statistical platform: easy deployment on high-performance computing and cloud computing through Bioconductor Cloud Amazon Machine Image (AMI), fully open-source workflows, interconnectivity of annotation and analytic packages, full reproducibility of analysis, etc.

2 rTANDEM

rTANDEM can use the same input as the classic X!Tandem, namely parameter files in XML format, FASTA databases and spectra files in ASCII format (like DTA, PKL or mgf), and return an XML result file. However, rTANDEM aims to streamline launching analyses, so it proposes different functions to create and manipulate parameters within R. For example, functions are provided for setting default values for different types of mass spectrometers or scoring keys. It also proposes a more complex syntax that allows multiple analyses to be launched using a single parameter object. rTANDEM lends itself naturally to the use of scripts for launching analyses, making every search fully traceable and reproducible.

The data structure holding the results is an R S4 class (Chambers, 1998, 2008) named rTResult. It contains slots to retrieve the settings of the analysis (for example, the parameters and data files used) and provide general statistics on the analysis (like the total number of spectra assigned or the estimated false discovery). The most important slots of the rTResult object are four tables containing the information on protein, peptide and Post Translational Modification (PTM) identifications as well as raw spectra. Those tables are of the class ‘data.table’, which is an R class optimized for quick and memory-efficient manipulation of large datasets. The tables can be efficiently queried to make subsets according to various properties (like identification scores, protein sequences or protein names) or can be subjected to vectorized operations.

One of the main advantages of using rTANDEM is that the results are available as an R object, making it easy to create complex processing pipelines that take full advantage of the

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com
3 shinyTANDEM

The shinyTANDEM package is a web-based graphical interface that provides an easy way to visualize results from rTANDEM. It is based on the newly released ‘shiny’ package from the RStudio team. This package adds web-server functionalities to R, making it possible to visualize the results of computations in a web browser. shinyTANDEM extends the basic use of ‘shiny’ to provide a full Graphical User Interface (GUI) from a single R command. The interface will display in the user’s default web browser.

The interface comprises a series of tabs representing various aspects of the search results: overview, statistics, protein view, peptide view. The ‘Overview’ tab presents the search parameters as well as raw lists of protein and peptide identifications. The ‘Statistics’ tab plots the protein and peptide matches according to their score and expectation values (Fig. 2). The ‘Protein’ tab features filters to select specific proteins and present their coverage and associated peptides. The ‘Peptide’ tab presents filters to select specific peptide sequences and see their associated MS2 spectra and protein sequence.

shinyTANDEM also features tabs that act as entry points for R. A loading tab allows the user to change the R object that is represented. Tabs allowing the user to create parameters objects, launch analysis, start conversion process or link to other R packages directly from the interface are currently being implemented.

4 RESULTS

rTANDEM was tested on a recently published dataset of breast cancer tissues (Liu et al., 2013) and compared with the reported results obtained with MaxQuant (v.1.1.1.36). We used raw data obtained from four whole tissue lysate of breast cancer samples. rTANDEM obtained an average of 8044 unique peptide-spectrum matches at an expect value <0.01 for those samples (Supplementary Table S1). This represents a noticeable increase compared with the original results from MaxQuant, which reported an average of 6254 peptides per sample at P<0.01.

5 CONCLUSION

The Bioconductor package rTANDEM and its associated graphical interface, shinyTANDEM, form an entry point to build complete proteomics workflows in the R statistical language. They provide ways to perform protein identification directly from R and to convert search results to/from R object. The S4 result data structure makes it easy to use R/Bioconductor to build complex processing pipelines around proteomic datasets. Further statistical tests for quantification in rTANDEM are currently under development.

ACKNOWLEDGEMENTS

The authors would like to thank the Proteomics Center of the CRCHUQ and its members.

Funding: A.D. holds a Réseau de médecine génétique appliquée (RMGA) salary award.

Conflict of Interest: none declared.

REFERENCES