Bioinformatics, 31(24), 2015, 3997–3999
doi: 10.1093/bioinformatics/btv494
Advance Access Publication Date: 26 August 2015
Applications Note

Sequence analysis

msa: an R package for multiple sequence alignment

Ulrich Bodenhofer*, Enrico Bonatesta, Christoph Horejš-Kainrath and Sepp Hochreiter

Institute of Bioinformatics, Johannes Kepler University, Linz, Austria

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 23, 2015; revised on August 11, 2015; accepted on August 15, 2015

Abstract

Summary: Although the R platform and the add-on packages of the Bioconductor project are widely used in bioinformatics, the standard task of multiple sequence alignment has been neglected so far. The msa package, for the first time, provides a unified R interface to the popular multiple sequence alignment algorithms ClustalW, ClustalOmega and MUSCLE. The package requires no additional software and runs on all major platforms. Moreover, the msa package provides an R interface to the powerful LaTeX package TeXshade which allows for flexible and customizable plotting of multiple sequence alignments.

Availability and implementation: msa is available via the Bioconductor project: http://bioconductor.org/packages/release/bioc/html/msa.html. Further information and the R code of the example presented in this paper are available at http://www.bioinf.jku.at/software/msa/.

Contact: bodenhofer@bioinf.jku.at or msa@bioinf.jku.at

1 Introduction

Multiple sequence alignment is one of the most fundamental tasks in bioinformatics. It serves as the basis for the detection of homologous regions, for detecting motifs and conserved regions, for detecting structural building blocks, for constructing sequence profiles, and as an important prerequisite for the construction of phylogenetic trees. Since exact methods scale exponentially with the number of aligned sequences, approximative methods have been introduced that try to obtain decent, yet not necessarily optimal, alignments with reasonable computational effort. ClustalW is a classic, but still very commonly used, method (Larkin *et al.*, 2007; Thompson *et al.*, 2004). T-Coffee (Notredame *et al.*, 2000) and MUSCLE (Edgar, 2004a, b) are other methods that are widely used. ClustalOmega (Sievers *et al.*, 2011) is a more recent top-notch method that has been designed to align large numbers of sequences in relatively short time. For a more comprehensive overview of multiple sequence alignment methods, we refer to Wallace *et al.* (2005), Edgar and Batzoglou (2006) and Notredame (2007).

All methods mentioned above are available as command line programs or via Web interfaces (e.g. on the EMBL-EBI server http://www.ebi.ac.uk/Tools/msa/). However, for the R platform that is widely used in bioinformatics, the possibilities to perform multiple sequence alignment are limited. The `Biostrings` package (Pages *et al.*, 2015), which is the standard package for sequence analysis within the Bioconductor project, provides data structures for storing and manipulating multiple sequence alignments, but not methods for computing them. Only recently, the `muscle` package that provides a simple interface to `MUSCLE` has been released via Bioconductor. For other methods, users still need to resort to external stand-alone programs.

This article presents `msa`, an R package released as part of Bioconductor 3.1 in April 2015. The package provides a unified interface to the three most common multiple sequence alignment methods and further integrates TeXshade (Beitz, 2000) for customizable pretty printing of multiple sequence alignments.

2 Package description

2.1. Multiple sequence alignment

The `msa` package provides interfaces to the three multiple sequence alignment methods namely ClustalW, ClustalOmega and MUSCLE. All three are available as R functions with a unified interface. The

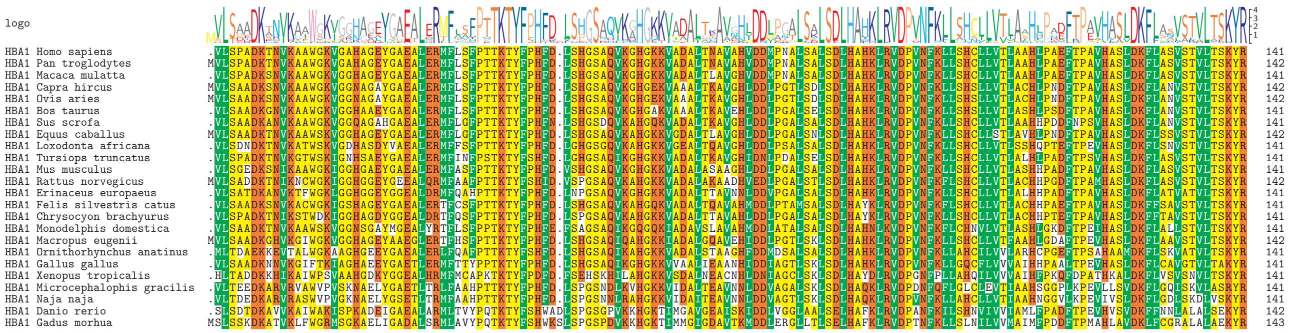

Fig. 1. Example of a multiple sequence alignment of Hemoglobin alpha 1 (HBA1) sequences of 24 vertebrates. The sequences have been downloaded from UniProt and have been aligned using ClustalW via msa with default parameters. The resulting multiple sequence alignment has been rendered using `msaPrettyPrint()` using custom parameters. The alignment is shaded in 'functional/structure' mode. A sequence logo is shown at the top of the alignment using the color mode 'rasmol'. The sequence logo and the alignment itself demonstrate that HBA1 is a highly conserved protein

functions accept DNA, RNA and amino acid sequences that can be provided as objects of classes '`DNAStringSet`', '`RNAStringSet`', or '`AAStringSet`', respectively. Sequences can also be provided as standard R character vectors. In this case, the user must specify the type of the sequences explicitly. All three methods allow for customizing substitution matrices. To this end, users can specify matrices by identifiers or they can also provide custom matrices (previously, this was not possible for ClustalOmega). For ClustalW and MUSCLE, it is also possible to specify custom gap penalties. The user can further decide whether the sequences in the resulting multiple sequence alignment should be in input order or in the order that is returned by the respective method (which typically groups similar sequences together). Results are returned as S4 objects, the classes of which have been derived from the classes provided by the `Biostrings` package. Therefore, all methods for processing multiple sequence alignments, such as the computation of consensus matrices and sequences, are inherited from the `Biostrings` package.

2.2. Pretty printing

The `Biostrings` package only offers a relatively basic function for printing multiple sequence alignments as plain text. The `msa` package integrates the powerful LaTeX package TeXshade (Beitz, 2000) which allows for flexible and customizable plotting of multiple sequence alignments. This interface is available via the function `msaPrettyPrint()`. The most common functionalities of TeXshade are controllable via the R-only interface of `msaPrettyPrint()`, so users can pretty-print multiple sequence alignments without the need to know the details of LaTeX or TeXshade. For more advanced presentations of the results, users can supply custom LaTeX code to `msaPrettyPrint()`.

There are basically two scenarios how to use the function: (1) By default, the function generates a LaTeX source file from which a PDF file is built. This is the standard use case for interactive R sessions. Note that `msaPrettyPrint()` does not allow for plotting multiple sequence alignments via R graphics devices. (2) The function also has an output mode that writes the generated LaTeX code to the output stream as it is. This is the perfect choice in code chunks of Sweave or knitr documents (Leisch, 2002; Xie, 2014), since it allows for embedding the generated TeXshade code directly into the resulting LaTeX source file that is generated by Sweave or knitr. Needless to say, this works only for knitr documents based on the LaTeX/Sweave syntax.

Figure 1 shows an example of a multiple sequence alignment produced by the ClustalW interface of msa that has been rendered with `msaPrettyPrint()` using some custom parameters.

2.3. Cross-platform availability

The package is available for all major platforms, Linux/Unix, Windows and Mac OS X. It does not require any external software or libraries for performing multiple sequence alignments. However, to generate PDF output with the `msaPrettyPrint()` function, a TeX/LaTeX system is required.

`msa` does not consist of re-implementations of the three methods, but includes the original source codes of their corresponding standalone programs along with some modifications to facilitate cross-platform compatibility. This approach ensures the easy integration of future releases of the methods. Moreover, it also allows for potentially maintaining the use of OpenMP (Dagum and Menon, 1998) to speed up ClustalOmega's computations. On Linux/Unix systems, on which the packages are usually built from source, OpenMP support depends on the availability and configuration of OpenMP on the respective system. For Mac OS, OpenMP support has been disabled entirely due to technical difficulties, but we hope to overcome this limitation in future releases.

2.4. Future extensions

One of the major goals for future releases is the integration of more multiple sequence alignment methods, in particular, T-Coffee. Previous versions of the T-Coffee source code made use of threads that were not portable to Windows in an easy way. Nevertheless, we will explore possibilities to integrate T-Coffee and other methods (Brudno *et al.*, 2003; Löytynoja *et al.*, 2012; Morgenstern, 1999; Szalkowski, 2012). Moreover, ClustalW offers phylogeny methods which are currently not accessible via the R interface of msa. Where possible and meaningful, we will strive for making all functionalities of the included methods available.

Conflict of Interest: none declared.

References

Beitz,E. (2000) TeXshade: shading and labeling of multiple sequence alignments using LaTeX2e. *Bioinformatics*, **16**, 135–139.

Brudno,M. *et al.* (2003) Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, **4**, 66.

Dagum,L. and Menon,R. (1998) OpenMP: an industry standard API for shared-memory programming. *IEEE Comput. Sci. Eng.*, **5**, 46–55.

Edgar,R.C. (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Edgar,R.C. (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Edgar,R.C. and Batzoglou,S. (2006) Multiple sequence alignment. *Curr. Opin. Struct. Biol.*, **16**, 368–373.

Larkin,M.A. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Leisch,F. (2002) Sweave: dynamic generation of statistical reports using literate data analysis. In: Härdle,W. and Rönz,B. (eds.) *Compstat 2002—Proceedings in Computational Statistics*. Physica-Verlag, Heidelberg, pp. 575–580.

Löytynoja,A. *et al.* (2012) Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics*, **28**, 1684–1691.

Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.

Notredame,C. (2007) Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput. Biol.*, **3**, e123.

Notredame,C. *et al.* (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Pages,H. *et al.* (2015) Biostrings: string objects representing biological sequences, and matching algorithms. R package version 2.36.1.

Sievers,F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

Szalkowski,A.M. (2012) Fast and robust multiple sequence alignment with phylogeny-aware gap placement. *BMC Bioinformatics*, **13**, 129.

Thompson,J.D. *et al.* (2004) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Wallace,I.M. *et al.* (2005) Multiple sequence alignments. *Curr. Opin. Struct. Biol.*, **15**, 261–266.

Xie,Y. (2014) *Dynamic Documents with R and knitr*. Chapman & Hall/CRC, Boca Raton.