MSAViewer: interactive JavaScript visualization of multiple sequence alignments

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1093/bioinformatics/btw474</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:29738979">http://nrs.harvard.edu/urn-3:HUL.InstRepos:29738979</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Sequence analysis

MSAViewer: interactive JavaScript visualization of multiple sequence alignments

Guy Yachdav,1,2 Sebastian Wilzbach,1 Benedikt Rauscher,1 Robert Sheridan,3 Ian Sillitoe,4 James Procter,5 Suzanna E. Lewis,6 Burkhard Rost1,2 and Tatyana Goldberg1,*

1Bioinformatik - I12, TUM, Garching, 85748, Germany, 2Biosof LLC, New York, NY 10001, USA, 3Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA, 4Institute of Structure and Molecular Biology, University College London, London, UK, 5Biological Chemistry and Drug Discovery, University of Dundee, Dundee, UK and 6Lawrence Berkeley National Laboratory, Berkeley, USA

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on April 16, 2016; revised on June 3, 2016; accepted on June 29, 2016

Abstract

Summary: The MSAViewer is a quick and easy visualization and analysis JavaScript component for Multiple Sequence Alignment data of any size. Core features include interactive navigation through the alignment, application of popular color schemes, sorting, selecting and filtering. The MSAViewer is ‘web ready’: written entirely in JavaScript, compatible with modern web browsers and does not require any specialized software. The MSAViewer is part of the BioJS collection of components.

Availability and Implementation: The MSAViewer is released as open source software under the Boost Software License 1.0. Documentation, source code and the viewer are available at http://msa.biojs.net/.

Supplementary information: Supplementary data are available at Bioinformatics online.

Contact: msa@bio.sh

1 Introduction

Multiple Sequence Alignment (MSA) is a fundamental procedure to capture similarities between sequences of nucleotides (DNA/RNA) or of amino acids (protein). Biologically meaningful MSAs highlight and capture sites with significant evolutionary conservation. MSAs are essential to predict aspects of protein structure (e.g. secondary structure (Rost, 2001) or protein disorder (Schlessinger et al., 2011)) and function (e.g. binding sites (Ofran et al., 2007) or localization (Goldberg et al., 2014)). MSAs can be used to understand genomic rearrangements (Darling et al., 2004), to derive sequence homology (Altschul et al., 1990) and to identify evolutionary rates (Pupko et al., 2002).

MSAs are widely used to display complex annotations relating to structure and function, and to transfer those annotations to sequences that lack annotations (Waterhouse et al., 2009). Many tools are available to view and analyze MSAs, including standalone applications (Larsson, 2014; Waterhouse et al., 2009) and web applets (Waterhouse et al., 2009). With the recent widespread adoption of JavaScript as the leading programming language for interactive web applications, new MSA viewing tools compatible with modern web browsers have been developed and made available (Martin, 2014). BioJS is one particular collection of JavaScript components with growing applications in biology (Corpas et al., 2014); it is interoperable with many other data visualization tools.

Here, we describe the BioJS MSAViewer. It is readily loaded into web pages to visualize and analyze MSA datasets of arbitrary sizes. MSAViewer implements most features commonly available in other popular MSA viewing software, including scrolling, selecting, highlighting, cross-referencing with protein feature annotations and phylogenetic trees (a detailed comparison of features is in Supplementary Table S1).
2 Visualization

The MSAViewer loads MSA data in FASTA (Pearson, 2000) or CLUSTAL (Larkin et al., 2007) formats from a user’s local computer or a web server. It then draws two main Canvas panels—the main MSA panel with residues in the alignment colored according to the Percentage Identity coloring scheme (Waterhouse et al., 2009). Percentage sequence identities relative to consensus (consensus sequence not shown) are listed for each sequence. Filled red rectangles indicate sequence annotations provided by the user (here: secondary structure predictions of PredictProtein (Yachdav et al., 2014)). Red frames (indicated by two black arrows) are clusters of residues of ebolavirus VP24 responsible for binding to karyopherin alpha nuclear transporters to suppress antiviral defense mechanism in human (Xu et al., 2014). (D) A compact overview MSA showing a bird’s eye view of the full alignment. Yellow rectangles are two highlighted clusters in the main MSA panel. Despite the overall sequence homology, the highlighted regions in ebolavirus (EBOV) proteins binding karyoprotein differ from those in Llovo cuevavirus (LLOV) and Lake Victoria marburgvirus (MARV). MARV is known to not suppress host’s antiviral defense mechanism (Xu et al., 2014). Though no experimental information on binding of LLOV VP24 to karyoproteins is available to date, the comparison of its sequence and structural features suggests a mechanism that is more similar to EBOV than to MARV.

The alignment is sortable by unique identifiers, sequence labels and sequences. The percentage of gaps and the sequence identity to the consensus sequence, which is calculated from most frequent bases at each position in the alignment, can also be used for sorting. Users can select sequences, columns, or arbitrary regions for analysis, and hide them from the main panel based on their selection, conservation to the consensus sequence or the percentage of gaps. An alignment can also be searched for motifs using a regular expression (e.g. K[K|R]RK for a nuclear localization signal), which are highlighted with a red frame if matched. Sequence position annotations, such as binding sites, are provided by a user and displayed as filled rectangles below the corresponding sequence in the alignment. Users can switch between 15 predefined color schemes. The MSAViewer exports alignments and annotations as ASCII files, and the visual representation as a publication-quality figure.

3 Summary

MSAViewer is a lightweight viewer for MSAs of arbitrary size. Being JavaScript-based, it can be used on any modern web browser without installing any specialized software or add-ons. As part of BioJS, the MSAViewer can interoperate with a growing set of biological data viewers. The phylogenetic tree and the sequence logo viewers are already integrated in the current release. Integration with new MSA visualization techniques, e.g. sequence bundles (Kultys et al., 2014) is planned. The MSAViewer has already been found useful and became part of Galaxy (Giardine et al., 2005). Access is free. The MSAViewer is available at http://www.jalview.org/help/html/features/biojsmsa.html.

Funding

This work was supported by a grant from the German Federal Ministry for Education and Research (BMBF), Ernst Ludwig Ehrlich Studienwerk and the Google Summer of Code program sponsored by Google Inc.

Conflict of Interest: none declared.

References


Fig. 1. A simplified view of the MSAViewer for the sequence alignment of protein VP24 within seven viruses of Filoviridae family. (A) Sequence logo (Schneider and Stephens, 1990) representation with conservation patterns at each position in the MSA. (B) Bar chart showing amino acid conservation per position. (C) Main MSA panel with residues in the alignment colored according to the Percentage Identity coloring scheme (Waterhouse et al., 2009). Percentage sequence identities relative to consensus (consensus sequence not shown) are listed for each sequence. Filled red rectangles indicate sequence annotations provided by the user (here: secondary structure predictions of PredictProtein (Yachdav et al., 2014)). Red frames (indicated by two black arrows) are clusters of residues of ebolavirus VP24 responsible for binding to karyopherin alpha nuclear transporters to suppress antiviral defense mechanism in human (Xu et al., 2014). (D) A compact overview MSA showing a bird’s eye view of the full alignment. Yellow rectangles are two highlighted clusters in the main MSA panel. Despite the overall sequence homology, the highlighted regions in ebolavirus (EBOV) proteins binding karyoprotein differ from those in Llovo cuevavirus (LLOV) and Lake Victoria marburgvirus (MARV). MARV is known to not suppress host’s antiviral defense mechanism (Xu et al., 2014). Though no experimental information on binding of LLOV VP24 to karyoproteins is available to date, the comparison of its sequence and structural features suggests a mechanism that is more similar to EBOV than to MARV.


