Title
ShareMSA: real time collaboration on multiple sequence alignments

Author
Yunes, Jeffrey

Publication Date
2017

Peer reviewed|Thesis/dissertation

ShareMSA: real time collaboration on multiple sequence alignments

by

Jeffrey Yunes

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Acknowledgements

I am very thankful for the guidance and support of my thesis advisor and chair, Patricia C. Babbitt. I thank my thesis committee members Hao Li and Ian Holmes. I am grateful to Elaine Meng for feedback on user experience and Eric Pettersen for help with Chimera plugins. This work was funded by NIH GM60595 to Patricia C. Babbitt.

Table of Contents

List of Figures

Abstract

A Multiple Sequence Alignment (MSA) is a widely used data structure in bioinformatics for applications ranging from evolutionary analyses to protein engineering. ShareMSA is a website that allows users to visualize, share, and collaborate on MSAs. ShareMSA makes MSAs more accessible by allowing collaborators one-click access to the MSA and annotations without the need to install software. In addition, when those with requisite permissions change the annotations or legend, all viewers of the MSA see the changes in real time. A Chimera plugin is available via the ShareMSA website, allowing Chimera users to publish MSAs from within this sequence-structure visualization software.

Availability and Implementation ShareMSA is freely available at https://www.sharemsa.com, and does not require registration. The front end is implemented in HTML and JavaScript and runs on all modern browsers that support WebSockets.

Introduction

Multiple Sequence Alignments (MSAs) are an essential tool in bioinformatics. Computational biologists use MSAs to compare protein sequences when studying relationships between sequence, structure, and function, and as input to bioinformatics software to generate phylogenetic trees and hidden Markov models. During collaborations, experimental and computational biologists often communicate research insights through MSAs. Sharing MSAs typically requires that collaborators install an alignment viewer, download the alignment file, and view the alignment offline (Pettersen *et al.*, 2004; Waterhouse *et al.*, 2009; Hillary *et al.*, 2011; Okonechnikov *et al.*, 2012; Milne *et al.*, 2013), which is inefficient for effective communication. Although some software can format an alignment as (static or interactive) HTML for viewing in a web browser, for example (Brown *et al.*, 1998; Skinner *et al.*, 2009; Löytynoja and Goldman, 2010; Gille *et al.*, 2014), there is no software designed for easy, real time collaborative analysis of MSAs.

Features

Copy, paste, and share Copy and paste your alignment in FASTA format into the form, and click Share. You'll be taken to your alignment and provided a shareable URL. Your collaborators will not need to download any special software to view the alignment.

Annotations in real time You (or others with permission) can annotate the alignment simply by clicking on residues. When users modify the annotations or visual filters (coloring), the changes propagate immediately to everyone viewing the alignment. This feature enables collaborators to seamlessly transmit and visualize annotation data without relying on rounds of e-mail or other forms of communication. More sophisticated annotations, such as those that associate a real number with each residue, can also be uploaded.

User-defined legend You can build a list of visual filters to apply to your alignment. For example, you can change the background color of each residue to its Clustal coloring (Thompson *et al.*, 1997), set the text color illustrating differences between sequences to one color, and outline columns of active site residues in another color.

No login required You can share your MSA without creating an account. You'll still be able to protect your anonymously owned MSA with a passcode. If you create an account, or log in with Google, we'll keep track of your multiple sequence alignments for you.

Chimera plugin Chimera users can easily publish their MSAs to ShareMSA. (Chimera is a program for interactive visualization and analysis of molecular structures and related data, including sequence alignments. (Pettersen *et al.*, 2004)) Simple instructions for installation and usage of the Chimera plugin can be found on the website.

Examples of Use

An example similar to that shown in Figure 1 is demonstrated in a video screencast on the home page of the ShareMSA website.

MSA sharer You navigate to https://www.sharemsa.com and click "Share an MSA." You type in a name, paste FASTA alignment data, and click Share. (Resources such as Pfam (Punta *et al.*, 2012), Clustal Omega (Sievers and Higgins, 2014), and the Structure-Function Linkage Database (Akiva *et al.*, 2014) provide FASTA-formatted alignments.) You are brought to a page showing your alignment and a Share Link. You send this link to your collaborators.

MSA sharee You click on a link provided to you by an MSA sharer, for example,
https://www.sharemsa.com/alignment/2692/?passcode=example.
You can see the MSA and a legend explaining the user-defined coloring scheme. If the sharer annotates the alignment or modifies the legend, your browser updates the display.

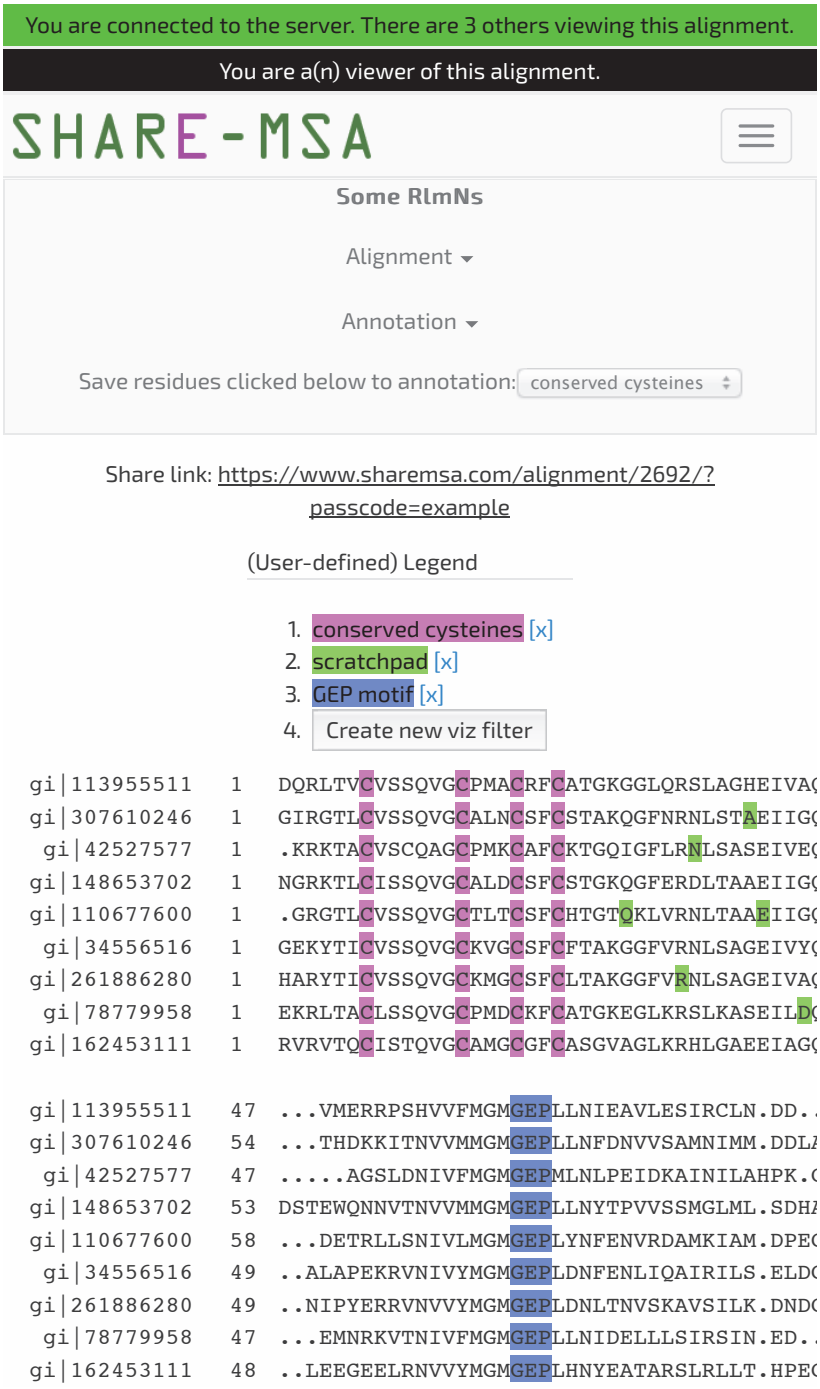


Figure 1: A screenshot from the mobile version of the site shows the menu and number of users that are connected, user permissions, and the MSA legend along with a portion of the alignment.

In this example, the sharer initially set restrictive permissions on this alignment for guests: you can view, but not change the alignment. You can make your own editable copy of the alignment by clicking Alignment → Copy. Had you been logged in to your ShareMSA account, you would have been designated the "creator" of the copy of the alignment, and it would be saved to your account.

One way that you can highlight residues for your collaborators is by clicking on the residues in the alignment. Before you begin, in the menu bar, choose to save residues clicked in the alignment to annotation "scratchpad." Then, under the legend, click the button to create a new visual filter for the residues in scratchpad and choose your color preferences. When you click on the residues in the alignment, the changes are shown on your screen. Again, the legend and annotations are propagated immediately to other viewers.

The software automatically generates some annotations, such as the differences between sequences.

Had you logged in or created an account during this process, the system would have stored the alignment to your profile, and you would have control over the permissions of the alignment.

Conclusion

ShareMSA brings next-generation, real time collaborative tools to one of the most widely used types of biological data on modern mobile and desktop web browsers.

References

Akiva, E. et al (2014). The Structure-Function Linkage Database. *Nucleic Acids Research*, **42**(D1), D521–D530.

Brown, N.P. et al (1998). MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics*, **14**(4), 380–381.

Gille, C. et al (2014). Sequence alignment visualization in HTML5 without Java. *Bioinformatics*, **30**(1), 121–122.

Hillary, W. et al (2011). Base-By-Base version 2: single nucleotide-level analysis of whole viral genome alignments. *Microbial Informatics and Experimentation*, **1**(1), 2.

Löytynoja, A. and Goldman, N. (2010). webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics*, **11**(1), 579.

Milne, I. et al (2013). Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, **14**(2), 193–202.

Okonechnikov, K. et al (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, **28**(8), 1166–1167.

Pettersen, E.F. et al (2004). UCSF chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**(13), 1605–1612.

Punta, M. et al (2012). The Pfam protein families database. *Nucleic Acids Research*, **40**(D1), D290–D301.

Sievers, F. and Higgins, D.G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.*, **1079**(Chapter 6), 105–116.

Skinner, M.E. et al (2009). JBrowse: a next-generation genome browser. *Genome Research*, **19**(9), 1630–1638.

Thompson, J.D. et al (1997). The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, **25**(24), 4876–4882.

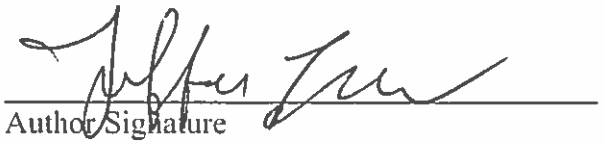
Waterhouse, A.M. et al (2009). Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**(9), 1189–1191.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature

2017-03-28

Date