SOFTWARE TOOL ARTICLE

REVISED AlignmentViewer: Sequence Analysis of Large Protein Families [version 2; peer review: 2 approved]

Roc Reguant¹⁻³, Yevgeniy Antipin⁴, Rob Sheridan⁵, Christian Dallago¹⁻³,
Drew Diamantoukos^{2,3}, Augustin Luna^{2,3,6}, Chris Sander^{2,3,6},
Nicholas Paul Gauthier^{2,3,6}

¹Department of Informatics, Technical University of Munich, Munich, Germany

²cBio Center, Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA

³Department of Cell Biology, Harvard Medical School, Boston, MA, USA

⁴Icahn School of Medicine, Mount Sinai, New York, NY, USA

⁵Knowledge Systems Group, Computational Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA

⁶Broad Institute of MIT and Harvard, Cambridge, MA, USA

V2 First published: 27 Mar 2020, 9:213
<https://doi.org/10.12688/f1000research.22242.1>

Latest published: 15 Oct 2020, 9:213
<https://doi.org/10.12688/f1000research.22242.2>

Abstract

AlignmentViewer is a web-based tool to view and analyze multiple sequence alignments of protein families. The particular strengths of AlignmentViewer include flexible visualization at different scales as well as analysis of conservation patterns and of the distribution of proteins in sequence space. The tool is directly accessible in web browsers without the need for software installation. It can handle protein families with tens of thousands of sequences and is particularly suitable for evolutionary coupling analysis, e.g. via EVcouplings.org.

Keywords

alignment viewer, MSA, JavaScript, protein alignments, web-based, tool,

This article is included in the **Bioinformatics** gateway.

Open Peer Review

Approval Status

	1	2
version 2 (revision) 15 Oct 2020		 view
version 1 27 Mar 2020	 view	 view

1. **Erik Larsson Lekholm**, University of Gothenburg, Gothenburg, Sweden

2. **Alex Bateman** , European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Chris Sander (sander.research@gmail.com), Nicholas Paul Gauthier (nicholasp_gauthier@dfci.harvard.edu)

Author roles: **Reguant R:** Methodology, Project Administration, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Antipin Y:** Conceptualization, Methodology, Project Administration, Software, Supervision; **Sheridan R:** Software, Writing – Review & Editing; **Dallago C:** Software, Writing – Review & Editing; **Diamantoukos D:** Software; **Luna A:** Project Administration, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Sander C:** Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing; **Gauthier NP:** Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The project was supported by the Human Frontier Science Program (HFSP) (RGP0055/2015), the National Resource for Network Biology (NRNB) from the National Institute of General Medical Sciences (NIGMS) (P41 GM103504), and the Department of Cell Biology at Harvard Medical School.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Reguant R *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Reguant R, Antipin Y, Sheridan R *et al.* **AlignmentViewer: Sequence Analysis of Large Protein Families [version 2; peer review: 2 approved]** F1000Research 2020, 9:213 <https://doi.org/10.12688/f1000research.22242.2>

First published: 27 Mar 2020, 9:213 <https://doi.org/10.12688/f1000research.22242.1>

REVISED Amendments from Version 1

We have updated the manuscript and the software according to reviewer comments. We have added a new feature to select the coloring scheme of the alignment, addressed cross-browser compatibility issues, and improved the usability. We have also addressed comments from the reviewers meant to improve the clarity of the manuscript.

Any further responses from the reviewers can be found at the end of the article

Introduction

Multiple sequence alignment (MSA) analysis (e.g., analysis of sequence patterns, subfamilies, specificity residues, evolutionary couplings) and visualization allows researchers to extract information and gain a better understanding of protein families. MSA is a basic step in many protein analysis workflows, including 3D structure prediction (Marks *et al.*, 2011), structure detection in flexible ('disordered') domains (Toth-Petroczy *et al.*, 2016), function prediction (Tamames *et al.*, 1998) and intracellular localization (Goldberg *et al.*, 2014).

A number of useful tools exist for the visualization of protein MSAs, such as, MView, Wasabi, AliView, MSASviewer and Jalview (Brown *et al.*, 1998; Larsson, 2014; Veidenberg *et al.*, 2016; Waterhouse *et al.*, 2009; Yachdav *et al.*, 2016). MView was one of the first online browser-based MSA viewers, with alignments formatted as an HTML document. Wasabi is a web-based tool particularly useful for phylogenetic analysis and incorporates phylogeny-aware alignment methods. Another desktop application, AliView, has features such as sorting, viewing, removing, editing and merging sequences from large nucleotide sequence datasets. MSASviewer is an interactive MSA visualizer in JavaScript that implements basic features of viewing, scrolling and motif selection. Jalview is a Java-based desktop tool accessible through websites using an embeddable applet, but unfortunately the technology for these applets is no longer supported in most browsers.

AlignmentViewer complements these MSA tools and provides the following features: (i) in-browser and serverless execution, (ii) visualization of very large MSAs, (iii) visualization of conservation patterns, (iv) sequence filtering, (v) logo display, (vi) pairwise sequence identity map, (vii) sequence space exploration by UMAP dimensionality reduction, and (viii) display of top-ranked evolutionary couplings (Hopf *et al.*, 2019).

An earlier version of this article can be found on bioRxiv (DOI: <https://doi.org/10.1101/269720>); additional features have been implemented since the earlier version.

Methods

Operation

AlignmentViewer is a web-based tool written in JavaScript with minimal system requirements. AlignmentViewer works best on Chrome regardless of operating system. AlignmentViewer

is developed with the D3 library (d3js.org) to produce dynamic and interactive data visualizations, with performance (speed) for large alignments a major consideration. The tool is entirely client-based, running inside a web browser without the need for server-side computation.

Implementation

Users can access AlignmentViewer and all its features directly from alignmentviewer.org, but its serverless execution enables anyone to quickly start a local copy for online or offline use. Hyperlinks for lookup in background databases, such as Uniprot or Pfam, are made directly from the client. Alignments can be passed to AlignmentViewer also via a URL query parameter that is served by https and is properly encoded (e.g., https://alignmentviewer.org/?url=https://alignmentviewer.org/example/1bkr_A.1-108.msa.txt), enabling seamless integration from external web services via a simple link (e.g. the EVcouplings, evcouplings.org, web server (Hopf *et al.*, 2019) offers visualization of computed alignments via a link to AlignmentViewer). The tool has been thoroughly tested with many large alignments. An alignment with, e.g., 50,000 sequences (about 13 MB of memory) loads in the Safari browser within one minute; further speedup is planned.

Use case

Figure 1 shows the main functionalities from AlignmentViewer explained in more detail in the next subsections. The top sub-figure shows the *msa view* with the sequence logo and the alignment capturing most of the attention. This view lets the user examine in depth the alignment. Each amino acid position is represented in sequence logo and the height shows users the information content of each position, in bits. Then, from left to right we show the *pixel view*, a part of the *stats view*, and the sequence space (with annotations). The *pixel view* gives an overview display of the alignment to enable a coarse view of the alignment for better visualization and pattern identification. The all versus all sequence identity sub-figure in Figure 1 (part of the *stats view*) displays allows users to identify possible clusters in the alignment based on sequence identity. The bottom right sub-figure of Figure 1 displays the sequences clustered by similarity (see section *Sequence space*) highlighted by user-provided annotations to aid in interpretation of the clusters.

MSA view

Alignment details. The *msa view* page has summary information: number of sequences, conservation and gap counts for each position, a sequence logo, and the residues in one letter code. By default, columns with gaps in the reference sequence (first row) are omitted in order to facilitate visual focus on sequence patterns relative to a protein of interest and to avoid extremely gapped alignment views typical of many MSA presentations. The amino acids are colored using a conventional coloring scheme, adopted from Mview, based on amino acid properties.

Sequence attributes and sorting. Sequences in the alignment can be sorted using one of four different methods: (i) the original order provided by the user, (ii–iii) by % sequence identity between a particular sequence and the reference (top) sequence,

Figure 1. AlignmentViewer visualization of the beta-lactamase protein domain family. Bars above the sequence alignment quantify residue conservation. The alignment consensus logo (just below the bar chart) is based on the amino acid frequencies. Lower left: *pixel view* of the alignment especially useful for large families; lower middle: protein-protein sequence similarity matrix graded by percentage identity; lower right: distribution of sequences in sequence space (UMAP projection), colored by species groups.

relative to the first or the second (gaps not counted), and (iv) by user-provided (upload annotations tab) sequence weights or other attributes, such as alignment profile scores (e.g., HMM bit scores). Sequences can be filtered by sequence identity relative to a reference sequence or by percentage of gaps.

Pixel view (suitable for large families)

The *pixel view* (*image view* website tab) leads to an overview of the entire depth and breadth of an MSA. The amino acid letters are represented by small rectangles of pixels, retaining the amino acid type coloring (*image view* tab). This striking visual impression can reveal patterns of conservation and variation, especially for large alignments. This is very useful to gain an intuitive view of sequence properties, noise at the uncertain edges of a protein family, as well as subfamily distributions. The coloring scheme can be by (1) amino acid properties, (2) hydrophobicity (red to blue) or (3) mutational difference (stronger color) in a sequence relative to the reference (first row) sequence.

Stats view

The *stats view* tab leads to plots of statistical properties of the set of sequences in the alignment, including (i) sequence identity relative to the reference sequence, and (ii) min, max, and average of (i); and (iii) a pairwise sequence identity matrix in which each pixel represents the degree of similarity between two sequences, such that a block-diagonal structure of the matrix is indicative of distinct subfamilies, given, e.g., a tree-derived sequence order as user input. The ordering of sequences by phylogeny is (currently) not part of the tool and can be performed using external tools, e.g., Wasabi (Veidenberg *et al.*, 2016).

Annotations and evolutionary couplings

Users can upload custom numerical attributes or labels for the sequences in the MSA (*upload annotations*) or evolutionary couplings between residue positions (*load couplings*). Adding these attributes allows users to use sequence weights, compare different measures of sequence fitness (e.g., bitscore, sequence

identity, statistical energy) or visualize evolutionary coupling constraints for pairs of positions.

Sequence space

Users can view representations of the MSA sequences in two- or three-dimensional space under the “sequence space” tab. These representations are generated using the Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction algorithm (McInnes *et al.*, 2018), which has been adapted for Javascript using the, [umap-js library \(https://github.com/PAIR-code/umap-js\)](https://github.com/PAIR-code/umap-js). The alignmentviewer.org implementation uses the number of amino acid differences between pairs of sequences (the Hamming distance) as the distance metric parameter. The algorithm then iteratively calculates an embedding in two- or three-dimensional space, which is displayed in real time for the end users. UMAP hyperparameters are set to reasonable defaults, but can also be configured via the settings panel. Sequences can be colored by user provided annotations (“upload annotations” tab).

Conclusion

AlignmentViewer is a lightweight online viewer for biological multiple sequence alignments that focuses on usability and performance. Written in JavaScript, this tool can be used in many browsers. The architecture of AlignmentViewer allows its use without software installation and without an internet connection. The visualization capabilities, analysis features and metrics in AlignmentViewer are useful in many areas of biology, especially evolutionary, structural, synthetic and

chemical biology. In the future we plan to add a visualization of species diversity, predicted contact maps, and organization by sequence subfamilies with specificity residues. A standalone version of AlignmentViewer is available at alignmentviewer.org and is in use by external services including EVcouplings.org. AlignmentViewer is an open source project hosted on GitHub, which welcomes engagement of interested members of the community.

Data availability

All data underlying the results are available as part of the article and no additional source data are required.

Software availability

AlignmentViewer website and demo can be found at: <https://alignmentviewer.org/>.

Source code available at: <https://github.com/sanderlab/alignmentviewer>.

Archived source code at time of publication: <https://doi.org/10.5281/zenodo.4063551> (Reguant, 2020).

License: MIT license.

Acknowledgments

We thank Debora Marks for constructive discussions.

References

- Brown NP, Leroy C, Sander C, *et al.*: **MView: a web-compatible database search or multiple alignment viewer.** *Bioinforma.* 1998; **14**(4): 380–381.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Goldberg T, Hecht M, Hamp T, *et al.*: **LocTree3 prediction of localization.** *Nucleic Acids Res.* 2014; **42**(Web Server issue): W350–W355.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hopf TA, Green AG, Schubert B, *et al.*: **The EVcouplings Python framework for coevolutionary sequence analysis.** *Bioinformatics.* 2019; **35**(9): 1582–1584.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Larsson A: **AliView: a fast and lightweight alignment viewer and editor for large datasets.** *Bioinformatics.* 2014; **30**(22): 3276–3278.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Marks DS, Colwell LJ, Sheridan R, *et al.*: **Protein 3D structure computed from evolutionary sequence variation.** *PLoS One.* 2011; **6**(12): e28766.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- McInnes L, Healy J, Melville J: **Umap: Uniform manifold approximation and projection for dimension reduction.** arXiv preprint. arXiv: 1802.03426. 2018.
[Reference Source](#)
- Reguant R: **Alignmentviewer (Version v2.1).** *Zenodo.* 2020.
<http://www.doi.org/10.5281/zenodo.4063551>
- Tamames J, Ouzounis C, Casari G, *et al.*: **EUCLID: automatic classification of proteins in functional classes by their database annotations.** *Bioinformatics.* 1998; **14**(6): 542–3.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Toth-Petroczy A, Palmedo P, Ingraham J, *et al.*: **Structured States of Disordered Proteins from Genomic Sequences.** *Cell.* 2016; **167**(1): 158–170.e12.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Veidenberg A, Medlar A, Löytynoja A: **Wasabi: An Integrated Platform for Evolutionary Sequence Analysis and Data Visualization.** *Mol Biol Evol.* 2016; **33**(4): 1126–1130.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Waterhouse AM, Procter JB, Martin DM, *et al.*: **Jalview Version 2—a multiple sequence alignment editor and analysis workbench.** *Bioinformatics.* 2009; **25**(9): 1189–91.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Yachdav G, Wilzbach S, Rauscher B, *et al.*: **MSAViewer: interactive JavaScript visualization of multiple sequence alignments.** *Bioinformatics.* 2016; **32**(22): 3501–3503.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 2

Reviewer Report 16 October 2020

<https://doi.org/10.5256/f1000research.30135.r73164>

© 2020 Bateman A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Alex Bateman

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK

The authors have addressed my concerns fully.

The authors may wish to double check the alignment of the example 1KR:A on the website. It does not look well aligned to my eyes. This suggestion is not required for approval of the manuscript.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational biology with specialism in biological databases and analysis of protein families.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 16 April 2020

<https://doi.org/10.5256/f1000research.24533.r61763>

© 2020 Bateman A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Alex Bateman

1

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK

² European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK

Overall there is a need for Javascript based alignment viewers which can handle large numbers of sequences. So this software is rather timely. However, I think the current implementation seems to still contain significant bugs and the web page requires a round of user experience testing to make it easier to use and interpret the results. Once these improvements are made then I think this has the potential to be a very useful tool.

Manuscript changes:

Please fix capitalisation of PFAM to Pfam

Software/Website changes:

The computing conservation box is annoyingly placed. It does not go away and covers the UMAP view significantly.

I really need an estimate of how long it will take for the pairwise identity to load. Some time passes...I didn't realise I had to click on calculate to get the pairwise identity to show. Please UX test this page to make it easier/clearer for the user to understand what to do.

For the top graph on the stats view the three colours used for max/average and min identity are hard to distinguish. Please use a broader colour palette. Please add axis legends for this graph and make it clearer what the title of this graph is. The title is very clear for the bottom plot.

Please add alternative colouring schemes for viewing the alignments. It makes a surprisingly big difference for experienced users ability to interpret an alignment. I personally like the ClustalX colouring scheme that is widely adopted, but there are other popular schemes it would be nice to incorporate as options.

"Based on this and other small tests we can safely state that any computer that is able to run any modern web browser will be able to run any alignment that requires visualization". This is overstating the case. Needs to be toned down. Try testing the software on the ABC transporter family (PF00005) alignment in Pfam. The NCBI alignment contains 2.6 million sequences. If it works seamlessly for that then you can safely state.

I tried to hook the viewer up to a Pfam alignment which is not very large (840 sequences). I got a box that said Fetching file...and no response beyond that.

<https://alignmentviewer.org/?url=http://pfam.xfam.org/family/PF00571/alignment/seed/format?format=fasta&>

also tried Stockholm format with no luck.

<https://alignmentviewer.org/?url=http://pfam.xfam.org/family/PF00571/alignment/seed/format?format=stockh>

I also tried with a Pfam family with just 3 sequences. Still with no luck.

I got the viewer to work by uploading a Pfam alignment in fasta format with 3 sequences. Then I tried uploading the seed alignment for the CBS domain downloaded from Pfam in fasta format. I then got the following error:

Parsing error: sequence #3 has different length.

The first few lines of the alignment are included below and the third sequence appears to be the same length as all the others.

```
>Q8EHI4_SHEON/79-134
I.....H.Q...V.....MTR.....NP.....V.....TVAPY.
...VS.LDVAS...RT..L....L....E.....HN..I....G..CL.P.V.L...
.....ENG....D.....LV.GIVTWKDLLRAYCA
>Q97V95_SULSO/193-248
V.....L.D...A.....G.TK.....NP.....I.....TINRY.
...YS.ILNAA...KL..M.....I....E.....KR..I....G..TL.L.V.M...
.....E..NQ...K.....LV.GIVTERDLMYAYIN
>Q6M020_METMP/269-321
V.....K.E...I.....MS.....PP.....V.....MVSPE.
...AA.LNELI...KG..M....A....N.....T.....D..RV.Y.V.V...
.....DNG....N.....IL.GIISKTDIVRTL SI
>Q72KI1_THET2/367-424
V.....E.G...F.....LA.....RA.....V.....VLPPS.
...TP.LSQVE...PR..L.....R....EA..G....G..RV.L.V.GE..
.....RAGEG...WR.....LL.GIYTRTDLYRSAPK
>Q8TX96_METKA/244-298
A.....R.N...Y.....L..Q.....EM.....V.....VWPPE.
...TP.LHEAL...WE..V.....ID....KM..S....D..RI.Y.V.M...
.....DG....R....KLT.GVVPLIDAVYTLAK
>Q836T2_ENTFA/77-132
V.....Q.E...I.....MSP.....PL.....MVAQD.
...TS.IRDAI...TN..L.....FMYDV....G..SL.Y.V.M...
.....DEAK....E.....LL.GVLSRKDLLRASLN
>Q8XIV1_CLOPE/77-131
V.....K.D...I.....MS.....KP.....V.....TVCEE.
...TM.LHDAI...VH.LF....L...ND.....V....G..TM.F.VEN...
.....GG....V.....LT.GAVSRKDFLKVAIG
```

If I delete the third sequence and re-upload I get exactly the same error.

For info I carried out tests with Firefox for Mac version 74.0 and Safari Version 13.1

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational biology with specialism in biological databases and analysis of protein families.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 09 Oct 2020

Roc Reguant

Manuscript changes:

Please fix capitalisation of PFAM to Pfam

Response: The manuscript has been updated to Pfam.

Software/Website changes:

The computing conservation box is annoyingly placed. It does not go away and covers the UMAP view significantly.

Response: This bug has now been fixed - the pop-up will disappear after the alignment has been loaded.

I really need an estimate of how long it will take for the pairwise identity to load. Some time passes...I didn't realise I had to click on calculate to get the pairwise identity to show. Please UX test this page to make it easier/clearer for the user to understand what to do.

Response: We have added a button into the middle of the plot to indicate that an action needs to be taken by the user to begin calculating pairwise sequence identity.

For the top graph on the stats view the three colours used for max/average and min identity are hard to distinguish. Please use a broader colour palette. Please add axis legends for this graph and make it clearer what the title of this graph is. The title is very clear for the bottom plot.

Response: We have updated the color palette to increase the contrast, added axis labels and added a clear title.

Please add alternative colouring schemes for viewing the alignments. It makes a surprisingly big difference for experienced users ability to interpret an alignment. I personally like the ClustalX colouring scheme that is widely adopted, but there are other popular schemes it would be nice to incorporate as options.

Response: We agree alternative coloring schemes are useful and have added the clustal colors as a drop-down option on the main page as well as in the image view.

"Based on this and other small tests we can safely state that any computer that is able to run any modern web browser will be able to run any alignment that requires visualization". This is overstating the case. Needs to be toned down. Try testing the software on the ABC transporter family (PF00005) alignment in Pfam. The NCBI alignment contains 2.6 million sequences. If it works seamlessly for that then you can safely state.

Response: We agree with the reviewer that the phrasing was optimistic and have removed the sentence from the main manuscript.

I tried to hook the viewer up to a Pfam alignment which is not very large (840 sequences). I got a box that said Fetching file...and no response beyond that.

<https://alignmentviewer.org/?url=http://pfam.xfam.org/family/PF00571/alignment/seed/format?format=fasta&alnType=seed&order=t&format=stockholm>

also tried Stockholm format with no luck.

Response: There are two technical limitations inherent to how requests happen in the browser. We take as an example your suggested URL:

<https://alignmentviewer.org/?url=http://pfam.xfam.org/family/PF00571/alignment/seed/format?format=stockholm&alnType=seed&order=t&format=stockholm>

Limitation 1: The alignmentviewer.org website is hosted as a secure site under https. However, the location of the requested alignment (from Pfam) is passed via an insecure connection (http). For security reasons mixing different protocols (also called "mixed content", see https://developer.mozilla.org/en-US/docs/Web/Security/Mixed_content) is not allowed.

Limitation 2: the string passed to the query parameter URL (aka. what follows "?url=") needs to be sanitized before being served. More information about this can be found here:

https://en.wikipedia.org/wiki/Query_string#URL_encoding. For clarity, we have updated the example in the main text to be sanitized / encoded.

In your example, the alignment lives at "

<http://pfam.xfam.org/family/PF00571/alignment/seed/format?format=stockholm&alnType=seed&order=t&format=stockholm>

". Passing the string as-is will confuse the browser that will interpret e.g. "&gaps=" as query for alignmentviewer rather than for pfam.xfam.org. Many languages offer built-in functions to sanitize strings to be passed to URL queries (e.g. in a JavaScript frontend application, using the `window.encodeURIComponent()` function; an interactive website can be found here: <https://www.url-encode-decode.com>). By doing so, the string proposed for the alignment would result in: "http%3A%2F%2Fpfam.xfam.org%2Ffamily%2FPF00571%2Falignment%2Fseed%2Fformat%3Fformat%3Dstockholm%26alnType%3Dseed%26format%3Dstockholm"

Conclusion: considering the two inherit browser limitations mentioned above, your requests

can be sanitized and fixed as follows:

<https://alignmentviewer.org/?url=https%3A%2F%2Fpfam.xfam.org%2Ffamily%2FPF00571%2Falignment%2F>

Again, note that for both URLs the protocol has been changed to "https" instead of "http", and that the string that follows "&url=" has been sanitized using Chrome's built-in console and the "window.encodeURIComponent(x)" function, where x is the string to be sanitized. We have added some text to the manuscript to highlight these requirements.

I also tried with a Pfam family with just 3 sequences. Still with no luck.

I got the viewer to work by uploading a Pfam alignment in fasta format with 3 sequences. Then I tried uploading the seed alignment for the CBS domain downloaded from Pfam in fasta format. I then got the following error:

Parsing error: sequence #3 has different length.

The first few lines of the alignment are included below and the third sequence appears to be the same length as all the others.

>Q8EH14_SHEON/79-134

*I.....H.Q...V.....MTR.....NP.....V.....TVAPY.
...VS.LDVAS....RT..L....L....E.....HN..I....G..CL.P.V.L...
.....ENG....D.....LV.GIVTWKDLLRAYCA*

>Q97V95_SULSO/193-248

*V.....L.D...A.....G.TK.....NP.....I.....TINRY.
...YS.ILNAA....KL..M.....I....E.....KR..I....G..TL.L.V.M...
.....E..NQ...K.....LV.GIVTERDLMYAYIN*

>Q6M020_METMP/269-321

*V.....K.E..I.....MS.....PP.....V.....MVSPE.
...AA.LNELI....KG..M....A....N.....T....D..RV.Y.V.V...
.....DNG....N.....IL.GIISKTDIVRTL SI*

>Q72KI1_THET2/367-424

*V.....E.G...F.....LA.....RA.....V.....VLPPS.
...TP.LSQVE....PR..L.....R....EA..G....G..RV.L.V.GE..
.....RAGEG...WR.....LL.GIYTRTDLYRSAPK*

>Q8TX96_METKA/244-298

*A.....R.N...Y.....L..Q.....EM.....V.....VVPPE.
...TP.LHEAL....WE..V.....ID.....KM..S....D..RI.Y.V.M...
.....DG....R....KLT.GWVPLIDAVYTLAK*

>Q836T2_ENTFA/77-132

*V.....Q.E..I.....MSP.....PL.....MVAQD.
...TS.IRDAI....TN..L.....FMYDV....G..SL.Y.V.M...
.....DEAK....E.....LL.GVLSRKDLLRASLN*

>Q8XIV1_CLOPE/77-131

*V.....K.D...I.....MS.....KP.....V.....TVCEE.
...TM.LHDAI....VH.LF....L...ND.....V....G..TM.F.VEN...
.....GG....V.....LT.GAVSRKDFLKVAIG*

If I delete the third sequence and re-upload I get exactly the same error.

For info I carried out tests with Firefox for Mac version 74.0 and Safari Version 13.1

Response: Thank you for pointing out this issue. This bug occurred when refreshing an internal variable, and has now been fixed.

Competing Interests: No competing interests were disclosed.

Reviewer Report 09 April 2020

<https://doi.org/10.5256/f1000research.24533.r61767>

© 2020 Lekholm E. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Erik Larsson Lekholm

¹ Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

² Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

This Software Tool article, describing a web-based sequence alignment viewer (AlignmentViewer), is well-written and very clearly presented. The introduction nicely motivates the need for this tool, given the many other available alternatives. The main features of the software are clearly explained in the article. The tool in itself seems useful and ran quickly and efficiently when tested on sample data. I have only a few minor comments that may be considered by the authors:

Minor:

- A little bit of detail may be added to the "Sequence space" section. "Two- or three-dimensional sequence space" is at first a bit confusing, as no information has been given at that stage as to what it means. How is UMAP applied to the Hamming distances? Is each sequence first represented by a vector of Hamming distances, with each element corresponding to one of the included sequences? Maybe this can be clarified with a few additional words.
- The following sentence is somewhat confusing: "Figure 1 shows the main functionalities from AlignmentViewer (Reguant, 2020) explained in more detail in the next subsections."
- "This view lets the user examine in depth the alignment." - should it be "examine the alignment in depth"?
- When loading an alignment using a somewhat older version of Safari (12.1.2), the "computing conservation" window does not disappear after the operation finishes. This problem was absent in Chrome.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: I was a postdoc in Chris Sander's lab at MSKCC, NYC, during 2009-2011. I confirm that this competing interest has not affected my ability to write an objective and unbiased review of the article.

Reviewer Expertise: Genomics, bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 29 Apr 2020

Chris Sander

Thank you for your comments and suggestions - they are spot on.
We will work on taking these into account on the web site and in the next version of the manuscript.
Chris Sander - DFCI and HMS

Competing Interests: No competing interests were disclosed.

Author Response 09 Oct 2020

Roc Reguant

A little bit of detail may be added to the "Sequence space" section. "Two- or three-dimensional sequence space" is at first a bit confusing, as no information has been given at that stage as to what it means. How is UMAP applied to the Hamming distances? Is each sequence first represented by a vector of Hamming distances, with each element corresponding to one of the included sequences? Maybe this can be clarified with a few additional words.

Response: We agree that a bit more detail is warranted and have updated the manuscript that describes using the hamming distance between individual sequences as the distance metric (parameter) for UMAP.

The following sentence is somewhat confusing: "Figure 1 shows the main functionalities from AlignmentViewer (Reguant, 2020) explained in more detail in the next subsections."

Response: We have updated the manuscript to make it more understandable. The citation to Reguant 2020 was also confusing (it points to the codebase) so we moved that codebase DOI to a single instance at the end of the manuscript.

"This view lets the user examine in depth the alignment." - should it be "examine the alignment in depth"?

Response: The manuscript has been corrected.

When loading an alignment using a somewhat older version of Safari (12.1.2), the "computing conservation" window does not disappear after the operation finishes. This problem was absent in Chrome.

Response: Thank you for noticing this bug. We have now corrected the issue and the dialog disappears when using Safari.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research