Bioinformatics, 35(22), 2019, 4800–4802
doi: 10.1093/bioinformatics/btz457
Advance Access Publication Date: 4 June 2019
Applications Note

Sequence analysis

NX4: a web-based visualization of large multiple sequence alignments

A. Solano-Roman[1,2,]*, C. Cruz-Castillo[3], D. Offenhuber[2] and A. Colubri[1,]*

[1]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA, [2]College of Arts, Media + Design, Northeastern University, Boston, MA 02115, USA and [3]Department of Computer Science, Universidad Latina de Costa Rica, San José 11501, Costa Rica

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on November 20, 2018; revised on May 2, 2019; editorial decision on May 27, 2019; accepted on May 29, 2019

Abstract

Summary: Multiple Sequence Alignments (MSAs) are a fundamental operation in genome analysis. However, MSA visualizations such as sequence logos and matrix representations have changed little since the nineties and are not well suited for displaying large-scale alignments. We propose a novel, web-based MSA visualization tool called NX4, which can handle genome alignments comprising thousands of sequences. NX4 calculates the frequency of each nucleotide along the alignment and visually summarizes the results using a color-blind friendly palette that helps identifying regions of high genetic diversity. NX4 also provides the user with additional assistance in finding these regions with a 'focus + context' mechanism that uses a line chart of the Shannon entropy across the alignment. The tool offers geneticists an easy-to-use and scalable analysis for large MSA studies.

Availability and implementation: NX4 is freely available at https://www.nx4.io, and its source code at https://github.com/NX4/nx4.

Contact: solanoroman.j@husky.neu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online

1 Introduction

There is a vast array of software tools designed to support the visual analysis of genetic sequences, including identification of functional or mutation-prone regions and visualization of phylogenetic trees (Daugelaite *et al.*, 2013). These tasks typically require the help of Multiple Sequence Alignment (MSA) viewers. The visual appearance and interfaces of MSA viewers have changed little since their introduction decades ago, in spite of the exponential increase in the amount of genome sequence data. For instance, the output of the program AliView (Larsson, 2014) is very similar to that of the program SEA VIEW, released in 1996 (Galtier *et al.*, 1996).

Our proposed tool, NX4, is an open-source, web-based visualization for MSAs that proposes a non-traditional visual output. Comparable efforts include the desktop alignment tool GenomeRing (Herbig *et al.*, 2012), which offers an alternative interactive radial layout, and MSA Viewer (Yachdav *et al.*, 2016), which is a traditional MSA viewer that can be deployed on the web. NX4 is a web application that incorporates some of the concepts proposed by Roca (2014), information theory measures to quantify nucleotide diversity and interaction techniques that allow the user to navigate large genomic sequences with ease, including a novel layered visualization.

2 Materials and methods

Traditional MSA views comprise a matrix of *N* columns where every row corresponds to a sequence in the alignment and every column corresponds to a nucleotide position across all the alignments. Each cell represents a nucleotide per sequence and is colored according to its type. Because MSAs can contain up to thousands of sequences, it is hard to find patterns in the resulting large matrices, first because the color schemes demand the user to closely look at differences in many colors across columns and rows, and second because the user has to scroll both vertically and horizontally to peruse the data, thus burdening their working memory.

NX4 reduces visual clutter and focuses on relevant differences by taking a statistical approach: it displays only five rows, each corresponding to one of the nucleotides (A, G, C, T), and an additional row for missing values (NA) regardless of the number of sequences in the set. This is achieved by calculating the frequency of the nucleotides, missing values and gaps for every position. The frequency of the nucleotide *n*, ($n =$ A, C, G, T or NA) for any given position $1 \leq i \leq N$ is calculated with the formula

$$f_i(n) = \frac{C(n,i)}{N - g_i - a_i},$$

where $C(n,i)$ is the total number of sequences with nucleotide n at position i in the alignment, g_i is the number of gaps in that position and a_i is the number of sequences with an ambiguous nucleotide. Roca (2014) proposed a similar approach for the visualization of amino acid sequences, displaying a matrix of twenty-one rows coupled with the known Sequence Logos (Schneider and Stephens, 1990) visualization, which shows each letter stacked on top of each other using various heights. However, the approach used in Logos has the shortcoming of warping the shape of the letters, which decreases the readability of the chart.

Therefore, to provide the user with assistance for identifying regions of high diversity, we implemented the information design technique of 'focus + context' (Cockburn *et al.*, 2009), which allows a user to look at a specific part of a visualization in detail, while maintaining a visual reference of the whole context. In this case, we included a line chart with the calculated values of Shannon entropy as a measure of variability for every position of the sequence, with values between zero and one (see Supplementary Note for details). This line chart allows the user to quickly pinpoint sections of high variability without the need to look at the specific nucleotide frequencies.

Another usability problem in traditional MSA viewers is that the most common color palettes have highly saturated colors which can 'produce afterimages when the viewer looks away from the screen, resulting in visual stress from prolonged viewing' (MacDonald, 1999). Moreover, several of these color palettes are not colorblind-safe and force users to read the letter inside the cell of the matrix, increasing the burden on the working memory. NX4 features two important improvements: first, a custom color mapping applied to the five-row matrix that reduces the risk of visual stress by using less saturated colors (Supplementary Fig. S1A) and second, the use different, color-blind-friendly colors for the entropy charts and for the matrix, in order to better differentiate the entropy and frequency values. (Supplementary Fig. S1B).

3 Results

NX4 can be used by visiting the website of the application and uploading an aligned file in the standard FASTA format, which also makes it easy to use with popular databases and alignment software like ViPR and MAFFT. The second way to use NX4 is to download the package from its GitHub repository and run it locally. The application was implemented in JavaScript with the help of the libraries *biojs-io-fasta* (Wilzbach, 2017) and *D3* (Bostock, 2017). Once the user has uploaded a valid file, they will be able see the visualization interface (Supplementary Fig. S2).

Here, we show two practical examples. One uses the first sample dataset available from NX4's home screen: this dataset contains 101 Ebola Virus (EBOV) genomes collected during the 2014 outbreak in West Africa (Gire *et al.*, 2014). In this dataset a key SNP occurred at position 10, 218, which allowed to identify two epidemiologically relevant lineages, one of them becoming the dominant strain within Sierra Leone (Park *et al.*, 2015). After loading this dataset, the user can see the entropy chart showing peaks, some of them reaching an entropy value of almost 0.5. If the user clicks around position 10, 218 in the top chart or drags the rectangular brush, the enlarged version of the entropy below to update and visible dark purple rectangles will show up in the matrix. Finally, by hovering over the second or third chart at the exact position of the described SNP, it is possible to see a value of entropy of 0.495 (Fig. 1A). The user can obtain an exportable list of sequence identifiers for further analysis by clicking on one of the dark rectangles of the matrix (Supplementary Fig. S3).

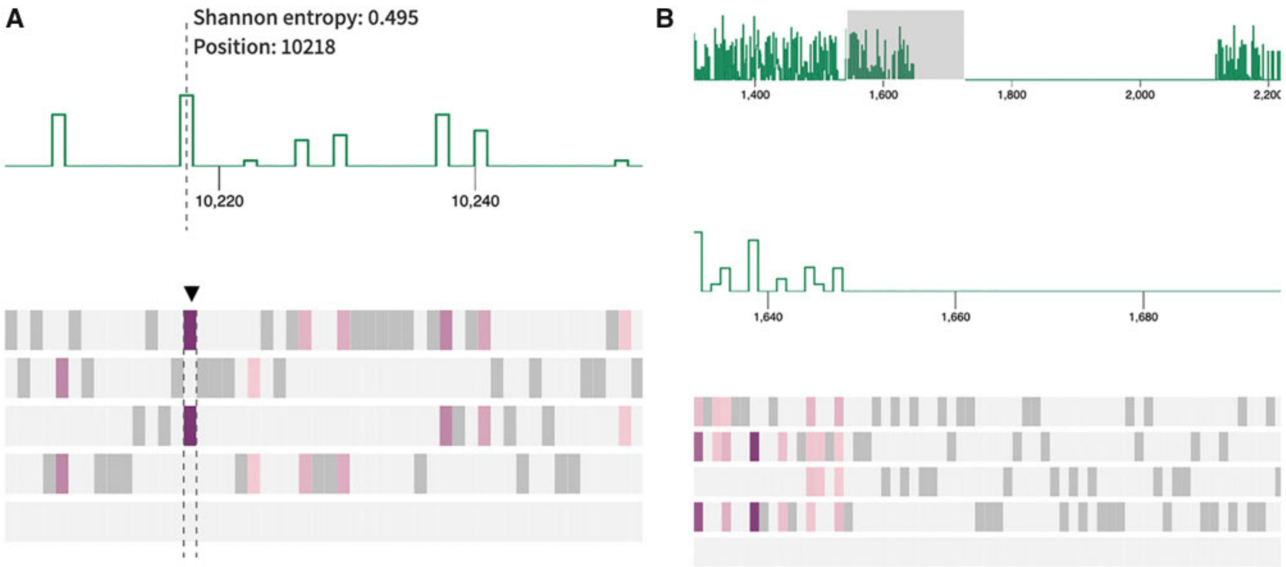

Fig. 1. (**A**) Detail view of the EBOV alignment where position 10, 218 shows high entropy and is consistent with findings of lineage divergence. (**B**) Detail of MSA for Rotavirus A displaying how areas of low/high conservation can be easily identified. Full color images are provided in the Supplementary Data (S1C and S1D)

In the second example, whose dataset is also available from the tool's home screen, we show how NX4 is well suited for the visual exploration of MSAs and it can help to easily pinpoint areas of high or low conservation. In this case (Fig. 1B), the dataset corresponds to an alignment of 5369 sequences of Rotavirus A generated using MAFFT v7.31. In the figure it is possible to see how visually distinct the regions of high conservation versus low conservation are, all within a concise representation that does not require vertically scrolling through thousands of rows, as it is the case in more traditional viewers, and allows getting all the relevant information at a quick glance.

We tested NX4 on a comprehensive collection of alignments for 272 single-stranded RNA viruses known or predicted to infect humans, some including up to 34 961 sequences. The list of alignments can be found in the Supplementary Materials.

4 Conclusions

NX4 is a flexible and intuitive web-based MSA viewer that integrates a contextual view of the Shannon entropy to localize areas of high variability, with a concise visual representation of that variability at the nucleotide level. This tool represents an improvement over classic MSA visualizations, such as sequence logos and matrix-based visualizations, particularly for large sequence alignments. The EBOV example shows that it is not necessary to inspect conventional alignment matrices to identify regions of possible clinical or epidemiological relevance. The Rotavirus A example shows the quick identification of regions of changing diversity. NX4 can accommodate alignments with thousands of sequences without increasing the visual burden for the user, and it allows to query the IDs of sequences that have variants in specific locations. Furthermore, NX4 is open-source so users can modify it to include their own customizations and integrate with larger tools. Currently, alignments of up to 20 000 nucleotides can be visualized conveniently with NX4, which makes well suited for working with viral genomic data. Future improvements to the tool include adding modifications to the layout to accommodate for much longer sequences and including visualizations of phylogenies for more advanced applications.

Acknowledgements

Special thanks to the research team of the Sabeti Laboratory, and the Design Affinity Group at the Broad Institute of MIT and Harvard.

Funding

This work was partly supported by the Northeastern University Graduate Grant from the College of Arts, Media and Design, as well as partly supported by a grant from the Bill and Melinda Gates Foundation.

Conflict of Interest: none declared.

References

Bostock,M. (2017) d3.js D3.

Cockburn,A. *et al.* (2009) A review of overview+detail, zooming, and focus+context interfaces. *ACM Comput. Surv.*, **41**, 2:1–2:31.

Daugelaite,J. *et al.* (2013) An overview of multiple sequence alignments and cloud computing in bioinformatics. *ISRN Biomathematics*, 2013, 14.

Galtier,N. *et al.* (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Bioinformatics*, **12**, 543–548.

Gire,S.K. *et al.* (2014) Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, **345**, 1369–1372.

Herbig,A. *et al.* (2012) GenomeRing: alignment visualization based on SuperGenome coordinates. *Bioinformatics*, **28**, i7–i15.

Larsson,A. (2014) AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, **30**, 3276–3278.

MacDonald,L.W. (1999) Using color effectively in computer graphics. *IEEE Comput. Graphics Appl.*, **19**, 20–35.

Park,D.J. *et al.* (2015) Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell*, **161**, 1516–1526.

Roca,A.I. (2014) ProfileGrids: a sequence alignment visualization paradigm that avoids the limitations of sequence logos. *BMC Proc.*, **8**, S6.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Wilzbach,S. (2017) biojs-io-fasta: parses fasta files biojs-io.

Yachdav,G. *et al.* (2016) MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics*, **32**, 3501–3503.