Modern Visualizations for Gene and Protein Multiple Sequence Alignments

SAVD3

David Andrés Ramírez

Introduction

Multiple sequence alignment (MSA) is a process that compares genes or proteins to infer information on evolutionary history between species. This analysis is very dependent on the methods of alignment chosen because alignment errors could cause incorrect tree reconstructions. Depending on which species are being compared and which multiple sequence alignment algorithm is used, many types of trees can be constructed. Some examples of tree variations are shown below, where each leaf of the tree represents a different species.

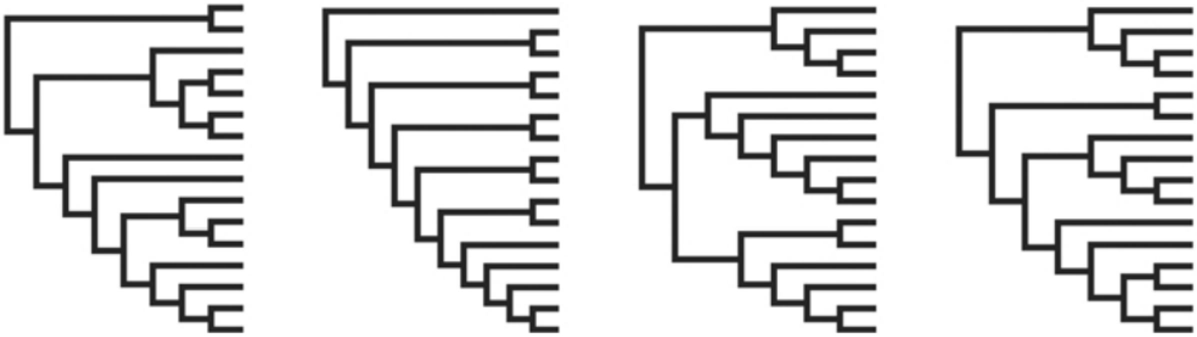


Figure 1: Multiple Sequence Alignment Accuracy and Phylogenetic Inference [7]

Although MSA is essential for biologists, it is not an easy process considering the complex algorithms that need to run to find these similarities. The number of nucleotides in a genome can reach thousands or millions, turning sequence alignment algorithms very computationally expensive. As an example, for a sequence of 100 nucleotides being aligned with a sequence of 95 nucleotides, there could be 55 million possible alignments if we add five gaps to the second sequence for the sake of comparing genomes of the same size.

Even though many sequence alignments programs and algorithms have been developed so far, most of them do not provide good visualization features. For this reason, specialized visualization applications have been developed to supply this demand and give biologists an easier way to compare and understand different genes or proteins. One of these applications is called MSAViewer. This is probably one of the most popular solutions because it is part of the BioJS group, a collection of JavaScript components with growing applications in biology. As a robust program, it provides the basic visualization facilities such as filters and sorting. The purpose of its developers is to predict disorders, functions, or locations of a protein. Being deployed on the web makes it more accessible to everyone interested in these types of topics. The application can read FASTA or CLUSTAL files, which are popular file formats to save MSA results. One problem of this application is that it does not offer a good overview where the user could see a percentage or relation between genes or proteins. An image of how MSAViewer displays the information is shown below.

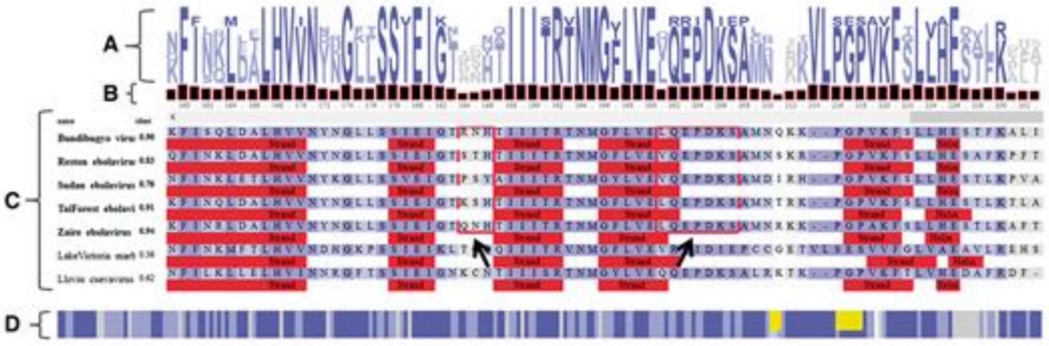


Figure 2: MSAViewer: interactive JavaScript visualization of multiple sequence alignments [2]

A more recent solution regarding sequence alignment visualizations, called NX4, was released in 2019. NX4 offers a progress compared to MSAViewer, in the sense that separate charts are displayed for the detailed and overview data. This is a great progress because seeing dense tables at first might be overwhelming for a user. Hence, NX4 implements a different way of comparing genes or proteins apart from a matrix of sequences, which makes the user scroll more than intended. NX4 offers the possibility to filter, so when the user selects a certain section, the detailed table jumps to the corresponding area afterward. One of the rows in their table represents missing values. Contrary to MSAViewer, new technologies were used such as D3, generating more complex solutions, but still being able to run on any browser. NX4 developers knew the importance of making it accessible to everyone, so they process the popular FASTA files. Contrary to MSAViewer, NX4 does not offer a table with the nucleotides of all genes or the amino acids of all proteins. The bottom table displays the entropy in each position. However, it is impossible to see the information regarding only one sequence. This is important if the user wants to see the sequence of a specific gene or protein used in the alignment. Another missing aspect is the proper use of standard colors that helps the user differentiate the nucleotides.

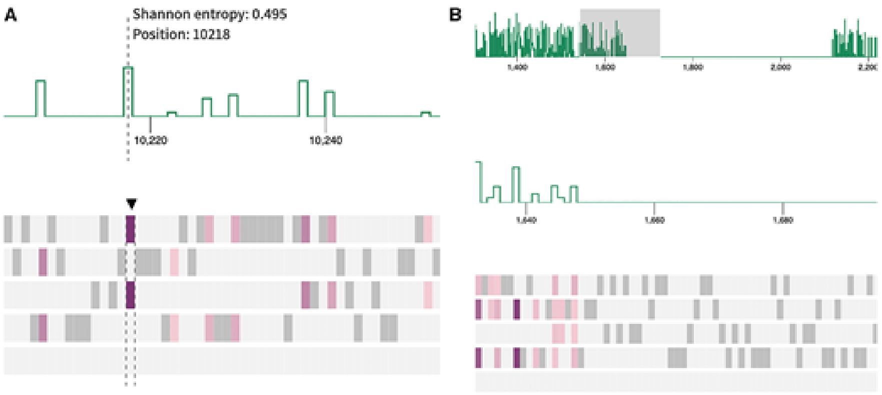


Figure 3: NX4: a web-based visualization of large multiple sequence alignments [1]

A common way to measure the variability of sequences between different genes or proteins is by using the Shannon Entropy or more commonly known as the Information Entropy. The Shannon entropy is a popular way of measuring variances because it provides a number, which represents the variety of items within a set. In this scenario, groups of equal nucleotides at the same position in all genes are assembled, and the groups are then analyzed all together to see the entropy at the given position. This method has been used before such as in the previously mentioned visualization website, NX4, and identifying genomic regions of the Papilloma virus [5].

Motivation

Multiple sequence alignment is a classical bioinformatics problem in genomics. For this reason, it is an opportunity to learn how genetics has evolved and transformed in recent decades. It is a good first approach to this topic that has always caught my attention. Within the entire field of bioinformatics, I chose visualization of multiple sequence alignments because I feel that despite being a highly developed field, there are few modern visualization solutions, making it hard for people to approach and learn about genetics and bioinformatics. What makes these previously mention visualizations not so great is the fact that they were done years ago with old technologies. Computer engineering allows me to shape multiple sequence alignments using new web technologies to make it more understandable and accessible, making this graduation project the perfect opportunity to combine two very different fields into one solution. I hope this graduation project will encourage many into learning about bioinformatics and all it has to offer.

Methodology

The application was built using React.js as the main web library for rendering all the present components on the website. React was chosen over other JavaScript frameworks due to its great support of the community and because it is the most used library for building websites. All the components were done in different files for having a more modular code, meaning it is easier to maintain over time. React helps to have modular code by allowing you to make components separate from each other.

Each chart was built using D3 because it is one of the most modern web technologies for data visualization. Not only is it modern, but it specializes in manipulating big amounts of data, like those used in genetics. Even though D3 is well optimized, the website code was done focusing on memory usage and time of response. NPM is the package manager used to install D3 in the website.

Having a clean page is important to understand more easily the information offered by the graphs. For this reason, various components were built using Bootstrap, which is also a very popular framework for web visualizations. The framework was installed in the index.html file, without using any package manager.

Tooltips were added to the right side of every title, helping the user understand what the corresponding table or graph does. Another JavaScript library called *react-popper-tooltip* was installed using NPM, which shows a tooltip when any information icon is clicked.

The covid sample file was acquired when aligning covid-like sequences with Blast. Blast is a software that searches for similar sequences to the one being analyzed. The sequences in this file are from viruses with similar sections, meaning that they could have a common ancestor. This virus was analyzed to generate more insight into the origin of the virus. The other two sample files were downloaded by searching for gene and protein alignments of different sizes and number of sequences.

Use Cases Diagram

The use case diagram shows all the actions that the user can perform on the web page, from the moment he selects an alignment until he interacts with all the graphs and tables.

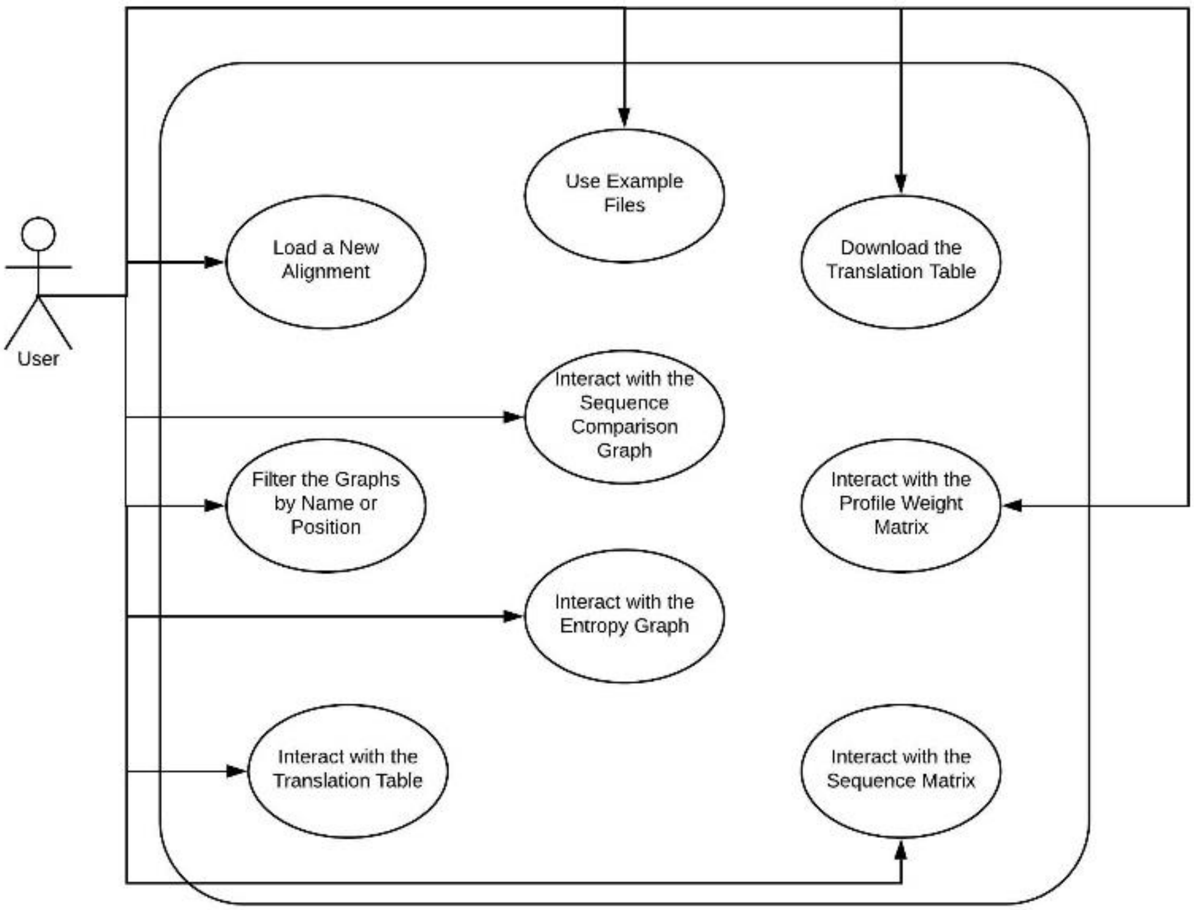

User
Load a New Alignment
Use Example Files
Download the Translation Table
Interact with the Sequence Comparison Graph
Filter the Graphs by Name or Position
Interact with the Profile Weight Matrix
Interact with the Entropy Graph
Interact with the Translation Table
Interact with the Sequence Matrix

Figure 4: Use Cases Diagram

Class Diagram

The class diagram shows how the components in the code are distributed. Additionally, the image shows what attributes and functions each component has. In the application source code, the purpose of all the functions and attributes is explained.

Index - This file has the function that tells the framework that App is the main component of the page, so it knows what to render first.

App - It is the main component of the page. Its function is to connect all the sections of the website. It is also responsible for managing the file that the user selects and sending the processed alignment to all the components that need it.

Tooltip - This component is responsible for creating a tooltip with a description for each graph and table on the page.

SequenceMatrix - This component creates and displays the sequence matrix. It also has the function of generating the modal with the sequences organized by value when the user selects a position.

Translation - This component handles everything related to the translation table if the selected file is a gene alignment. It is also responsible for generating the text file and sending it to the user when they want to download it.

EntropyAndProfile - This component generates the entropy graph and the profile weight matrix. Both graphs are built within the same component since the entropy graph handles the zoom and brush features of both visualizations.

SequenceComparison - This component builds the sequence comparison graph and the respective name inputs to update the visualization.

myObjectSecondGraphProteins - This class is used to instantiate objects with the repetition rate of each value per position used by the profile weight matrix. These objects are used in case the selected file is a protein alignment.

myObjectSecondGraphGenes - This class is used to instantiate objects with the repetition rate of each value per position used by the profile weight matrix. These objects are used in case the selected file is a gene alignment.

myObject - This class is used to instantiate objects with the entropy per position for all selected sequences. These objects are used by the entropy graph.

myObjectFifthGraph - This class is used to instantiate objects with the value per position a sequence has. These objects are used by the sequence matrix.

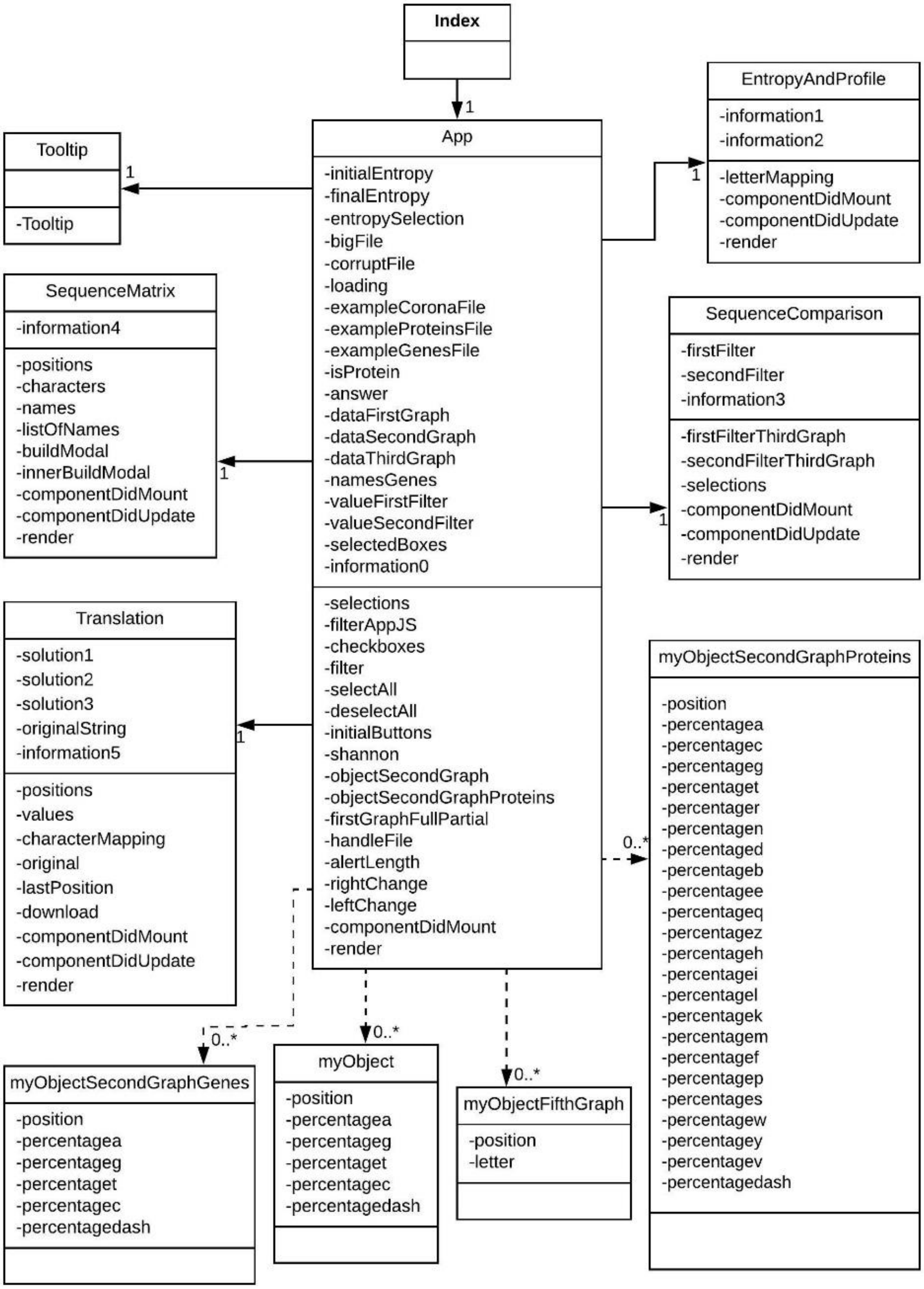


Figure 5: Class Diagram

Sequence Diagram

The sequence diagram shows the steps that a user follows throughout his interaction with the page. The diagram also shows all the responses the website can make.

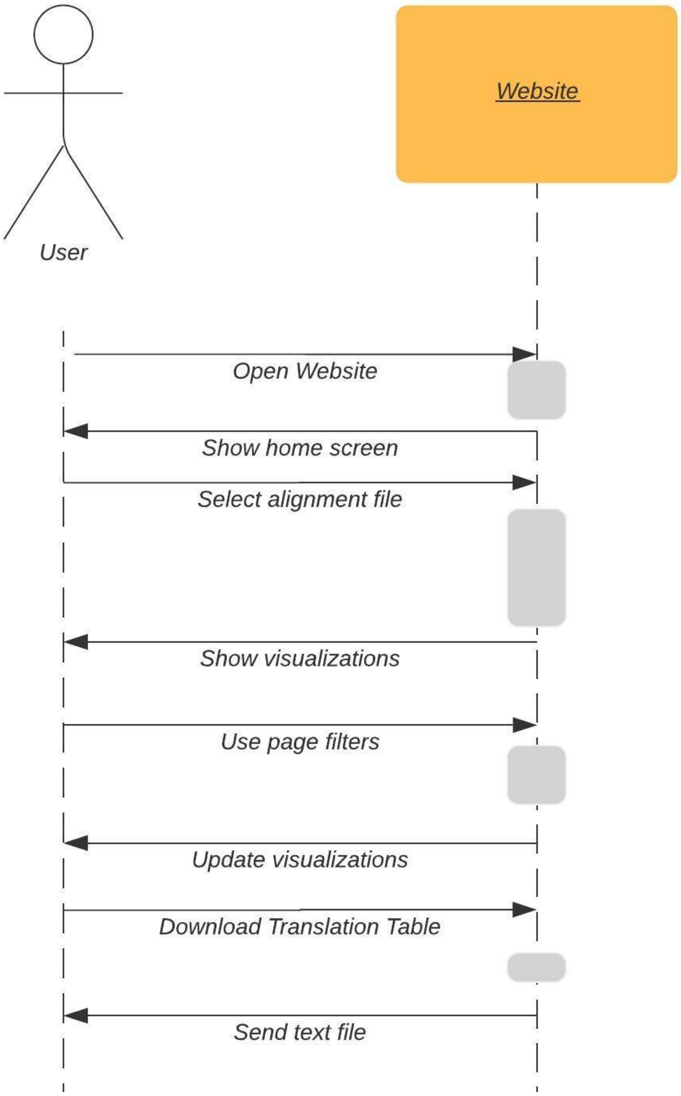
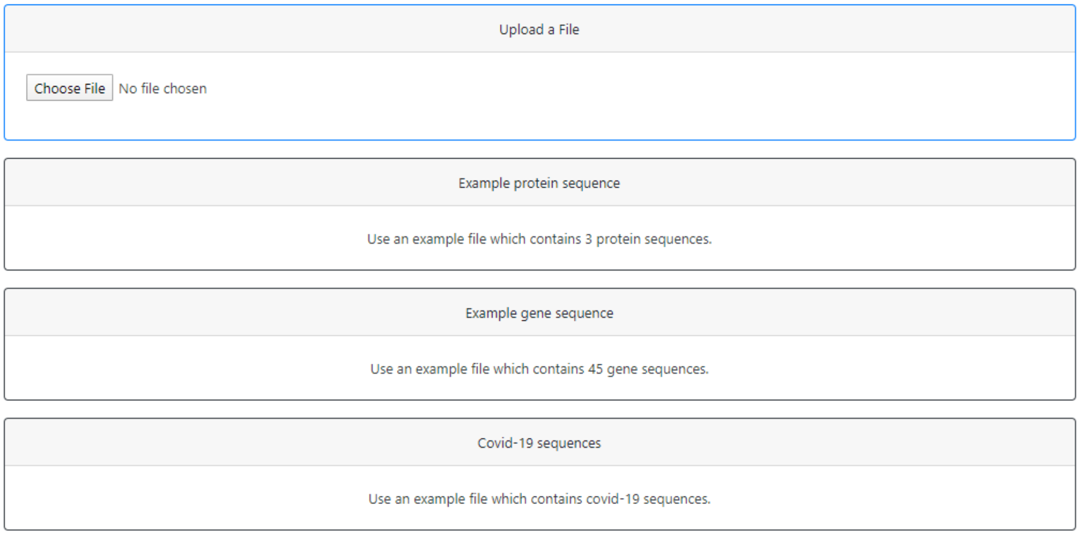


Figure 6: Sequence Diagram

Results

Figure 7 shows the front screen of SAVD3 for MSA. What the user will first encounter when opening the website is a section with four different options. Three of them use the sample files already loaded. The first option allows users to upload a custom alignment. To prevent the page from reading files with a wrong format, the website checks whether it can be processed. Otherwise, the page will ask for another file. Depending on what type of alignment is loaded, the graphs will vary in color and information. For this reason, the user is encouraged to check all sample files to learn about all the features the application offers. If the user wants to use their file, the alignment must be in FASTA format. There are many applications which provide an alignment service such as MAFFT, Clustal Omega, or MEGA, among others.



Upload a File

Example protein sequence

Use an example file which contains 3 protein sequences.

Example gene sequence

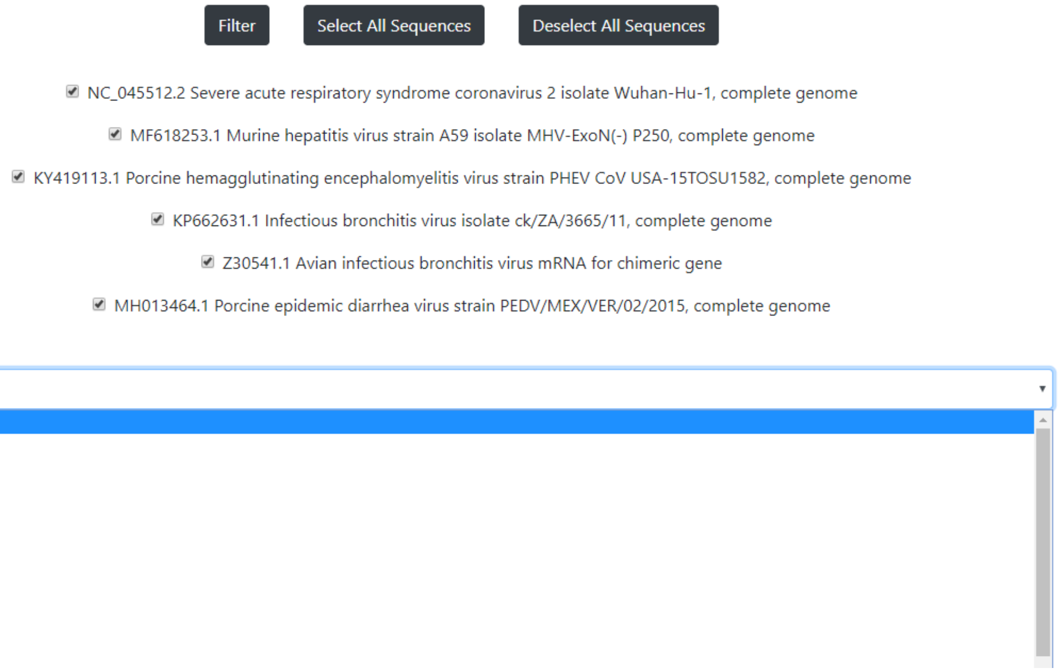
Use an example file which contains 45 gene sequences.

Covid-19 sequences

Use an example file which contains covid-19 sequences.

Figure 7: File Selection Page

At the top of the website, there is a position and name filter for all sequences contained in the selected alignment. Both filters update all the visualizations across the website without having to reload the browser tab. The purpose of the filters is to help with visualizations of big datasets, as they can be overwhelming or confusing at first. Another benefit of having filters is letting the user compare specific sequences, instead of all of them.



☑ NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

☑ MF618253.1 Murine hepatitis virus strain A59 isolate MHV-ExoN(-) P250, complete genome

☑ KY419113.1 Porcine hemagglutinating encephalomyelitis virus strain PHEV CoV USA-15TOSU1582, complete genome

☑ KP662631.1 Infectious bronchitis virus isolate ck/ZA/3665/11, complete genome

☑ Z30541.1 Avian infectious bronchitis virus mRNA for chimeric gene

☑ MH013464.1 Porcine epidemic diarrhea virus strain PEDV/MEX/VER/02/2015, complete genome

1 - 1000

Figure 8: Name and Position Filter

The objective of the first graph is to give a general notion of the similarity between the sequences. To achieve this goal, each point in the Y-Axis shows the Shannon entropy at that specific position, displaying the variability patterns across the alignment. Since there are many positions in a sequence, each point in the X-Axis will group nearby nucleotides or amino acids. Having a graph like this is important to infer conserved zones. The lower the entropy, the greater the sequence similarity within the region. More important than the exact value in a certain position, is the comparison between the average height of different sections of the graph. As mentioned in the introduction, the Shannon Entropy has been used in other genetics projects to compare genes and proteins.

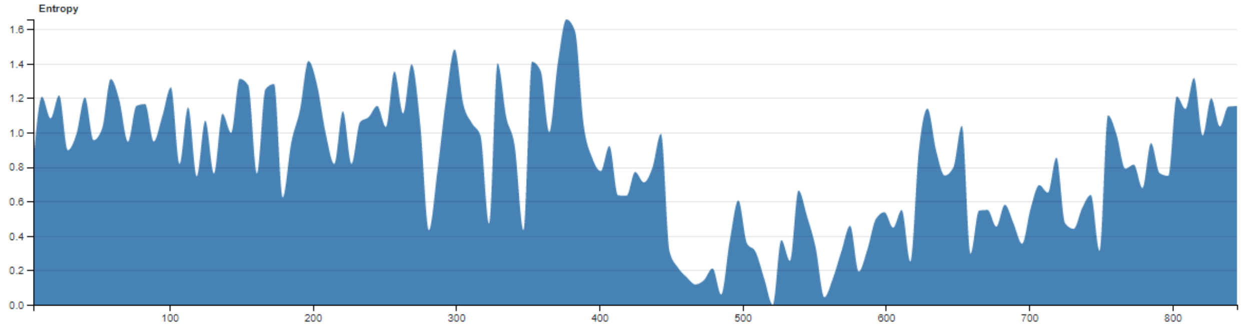


Figure 9: Entropy Graph

For calculating the Shannon Entropy, the algorithm used by the application is the traditional function. The summation will do an iteration for each existing nucleotide or amino acid at the position, where p_i is the percentage of finding that specific nucleotide or amino acid in all sequences.

$$Entropy\ (x) = -\sum_{i=0}^{N-1} p_i * \log_2 p_i$$

Figure 10: Shannon Entropy

Taking into account that a typical multiple sequence alignment can extend over thousands of characters, just below the first chart, there will be a smaller graph with an overview of the first one, where the user can zoom and brush a specific region to update the entropy graph.

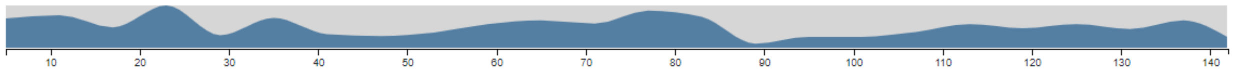


Figure 11: Entropy Graph Zoom and Brush Feature

When a file with sequences longer than a thousand positions is selected, an alert message is shown telling the user that he only can observe a thousand positions at a time. However, two buttons can change the number of positions displayed in the entropy graph. The page needs to show the full entropy graph so it can be inferred the common sections between the sequences.

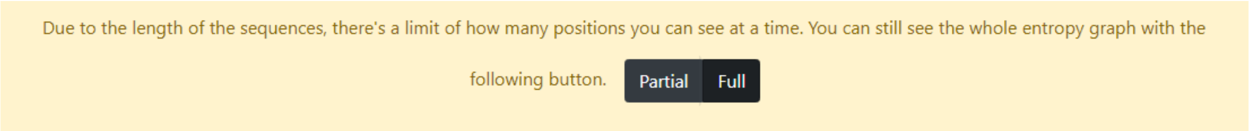

Due to the length of the sequences, there's a limit of how many positions you can see at a time. You can still see the whole entropy graph with the following button.

Figure 12: Entropy Graph Size Selection

Right beneath, there will be a profile weight matrix chart that displays the nucleotide (or aminoacid) distribution for the sequences and sections selected by the user. The objective of this graph is to show the percentage of each different value per position in all the selected sequences. This information is important as it shows how similar the sequences are in a certain section. Each point in the X-Axis represents the position, and the Y-Axis displays the percentage of each distinct value. The selected colors were chosen to help the user understand what he is analyzing. The graph will update itself as the user zoom or brush the entropy graph, because both are built together. The reason behind joining both graphs is to prevent the user from having to zoom twice when he wants to check the same section in both graphs. Depending if the input is a gene or a protein alignment, different colors and legends are shown.

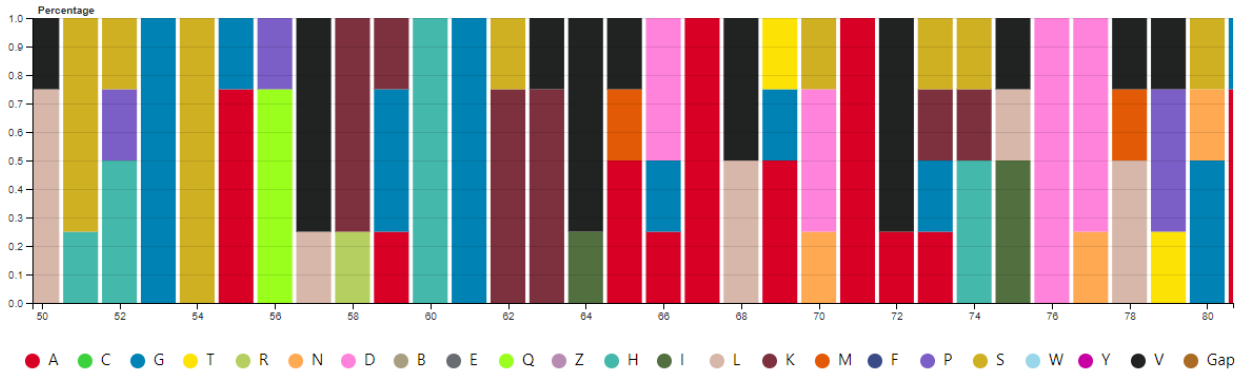


Figure 13: Profile Weight Matrix of a protein alignment

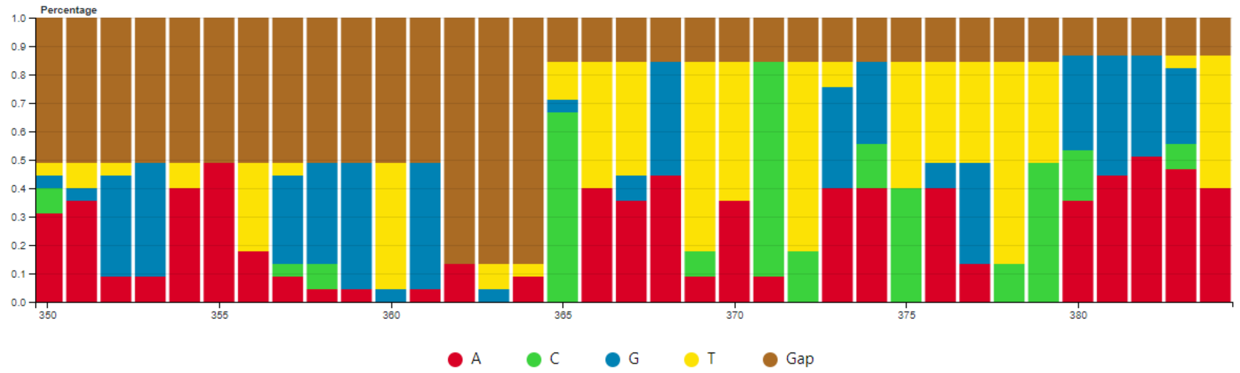
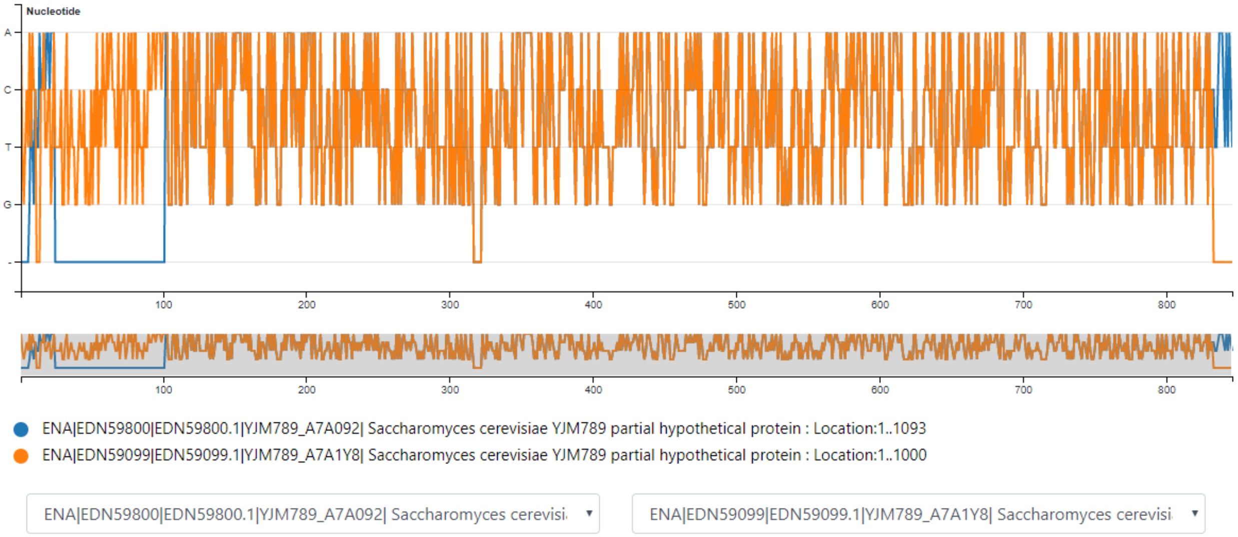


Figure 14: Profile Weight Matrix of a gene alignment

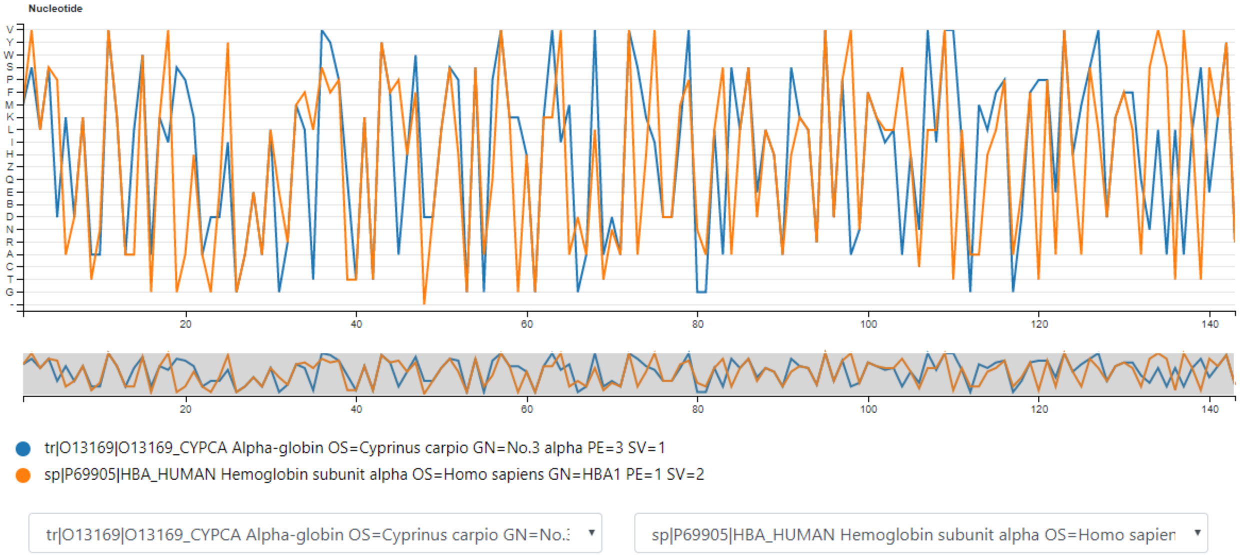
The next graph is a detailed comparison between two sequences. The purpose of this visualization is to compare how similar two sequences are, as it shows where both sequences have the same values. Each value in the Y-Axis is the corresponding nucleotide or amino acid, and the X-Axis displays the position where each value is. As the previous graphs, this visualization has a brush and zoom feature too. When only one color appears at a certain position, it means that both sequences have the same value at that position.


Nucleotide
A
C
T
G
-
100
200
300
400
500
600
700
800
● ENA|EDN59800|EDN59800.1|YJM789_A7A092| Saccharomyces cerevisiae YJM789 partial hypothetical protein : Location:1..1093
● ENA|EDN59099|EDN59099.1|YJM789_A7A1Y8| Saccharomyces cerevisiae YJM789 partial hypothetical protein : Location:1..1000

ENA|EDN59800|EDN59800.1|YJM789_A7A092| Saccharomyces cerevisi ▾

ENA|EDN59099|EDN59099.1|YJM789_A7A1Y8| Saccharomyces cerevisi ▾

Figure 15: Sequence Comparison of a gene alignment

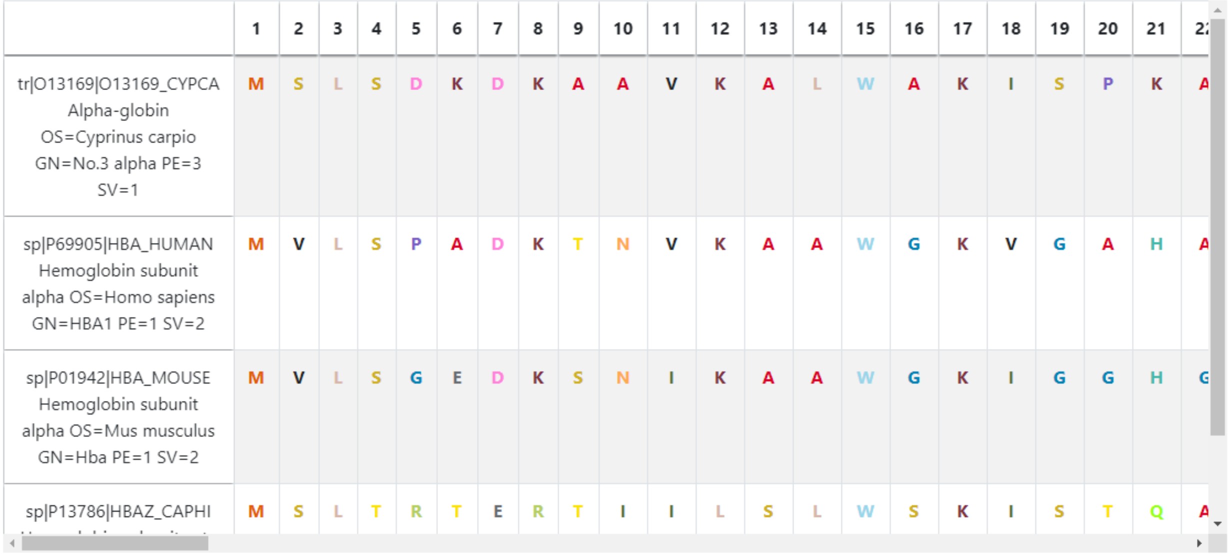

Nucleotide
V
Y
W
S
P
F
M
K
L
I
H
Z
Q
E
B
D
N
R
A
C
T
G
-
20
40
60
80
100
120
140
● tr|O13169|O13169_CYPCA Alpha-globin OS=Cyprinus carpio GN=No.3 alpha PE=3 SV=1
● sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2

tr|O13169|O13169_CYPCA Alpha-globin OS=Cyprinus carpio GN=No.: ▾

sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapien ▾

Figure 16: Sequence Comparison of a protein alignment

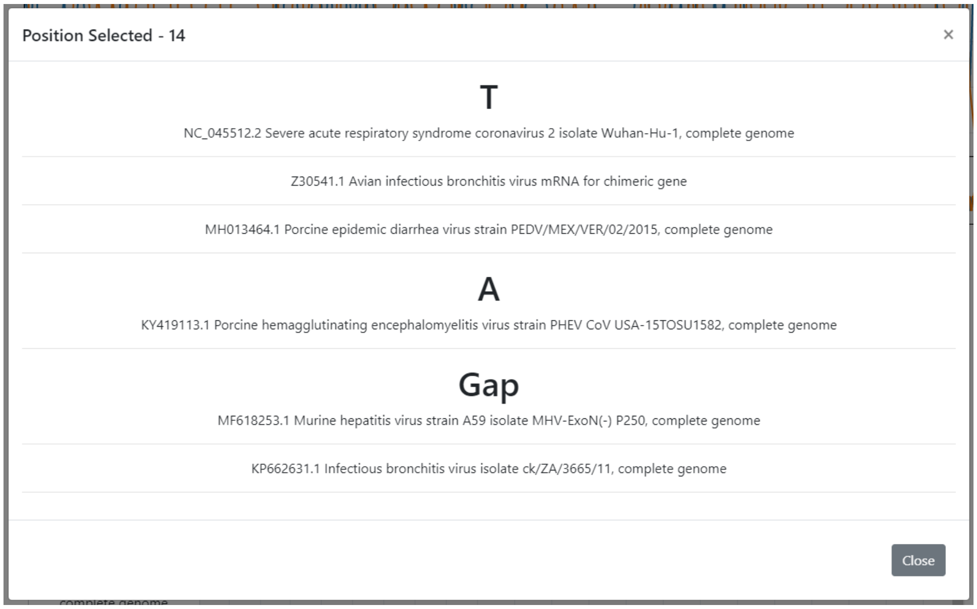
At the bottom of the page, the user will find a table containing the nucleotides or amino acids found in the selected position and sequences. This table is necessary for having the values in an organized way, different as it appears inside the FASTA file. The colors used for each character are the same colors displayed in the legend of the profile weight matrix graph. This helps to maintain the same structure across the website.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
tr|O13169|O13169_CYPCA Alpha-globin OS=Cyprinus carpio GN=No.3 alpha PE=3 SV=1	M	S	L	S	D	K	D	K	A	A	V	K	A	L	W	A	K	I	S	P	K	A
sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2	M	V	L	S	P	A	D	K	T	N	V	K	A	A	W	G	K	V	G	A	H	A
sp|P01942|HBA_MOUSE Hemoglobin subunit alpha OS=Mus musculus GN=Hba PE=1 SV=2	M	V	L	S	G	E	D	K	S	N	I	K	A	A	W	G	K	I	G	G	H	G
sp|P13786|HBAZ_CAPHI	M	S	L	T	R	T	E	R	T	I	I	L	S	L	W	S	K	I	S	T	Q	A

Figure 17: Sequence Comparison of a protein alignment

When the user selects a matrix position by clicking on the first row, a modal pop up, displaying all sequences separated by nucleotide or amino acid at that specific position. The purpose of the modal is to save the user time if he wants to visualize the sequences grouped in this way.



Position Selected - 14

T

NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

Z30541.1 Avian infectious bronchitis virus mRNA for chimeric gene

MH013464.1 Porcine epidemic diarrhea virus strain PEDV/MEX/VER/02/2015, complete genome

A

KY419113.1 Porcine hemagglutinating encephalomyelitis virus strain PHEV CoV USA-15TOSU1582, complete genome

Gap

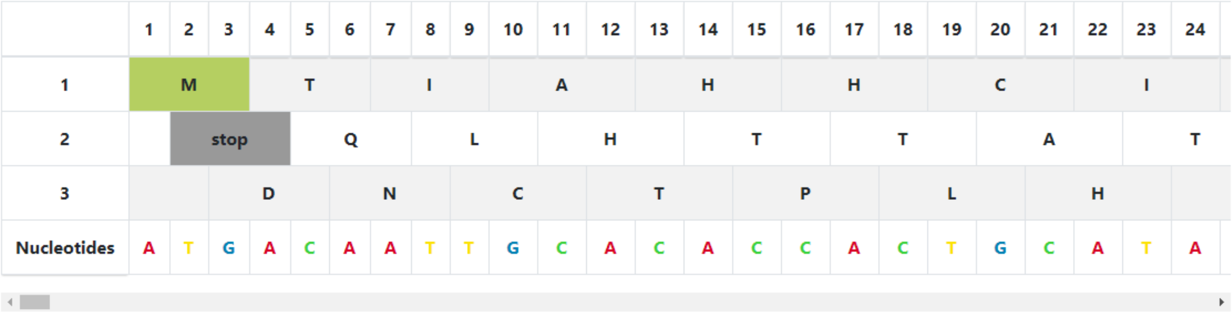
MF618253.1 Murine hepatitis virus strain A59 isolate MHV-ExoN(-) P250, complete genome

KP662631.1 Infectious bronchitis virus isolate ck/ZA/3665/11, complete genome

Figure 18: Sequence Matrix Detailed Modal

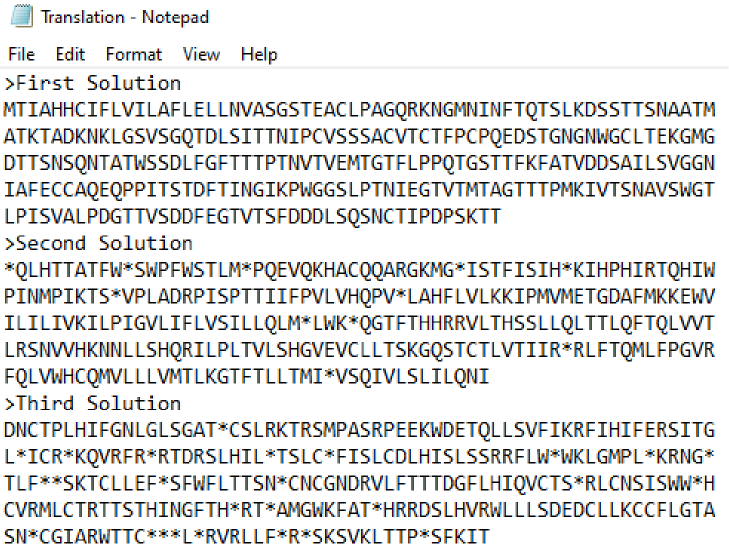
Since the purpose of the website is to be accessible for the greatest number of scientists, selected colors will be used to differentiate nucleotides or amino acids. The idea behind the chosen colors is to help everyone with vision impairment to read the graphs without having any trouble. The colors within gene alignment visualizations are also present throughout other sequence alignment applications. Colors used for protein alignments are original to this website, as there is not a pattern across multiple websites.

Finally, the last table only appears when a gene alignment is loaded. This visualization is the translation of nucleotides to amino acids using the DNA codon table. Since there are three possible translations, the table shows all the possible solutions. The start codons have a green background and the stop codons have a black one, letting the user identify which of the possible solutions is the correct one. Additionally, the website offers the option of downloading the table into a text file.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	M			T			I			A			H			H			C			I		
2		stop			Q			L			H			T			T			A			T	
3			D			N			C			T			P			L			H			
Nucleotides	A	T	G	A	C	A	A	T	T	G	C	A	C	A	C	C	A	C	T	G	C	A	T	A

Figure 19: Translation table (only shown with gene alignments)



```
>First Solution
MTIAHHCIFLVILAFLELLNVASGSTEACLPAGQRKNGMNINFTQTSLKDSSTTSNAATM
ATKTADKNKLGSVSGQTDLSITTNIPCVSSSACVTCTFPCPQEDSTGNGNWGCLTEKGMG
DTTSNSQNTATWSSDLFGFTTTPTNVTVEMTGTFLPPQTGSTTFKFATVDDSAILSVGGN
IAFECCAQEQPPITSTDFTINGIKPWGGSLPTNIEGTVTMTAGTTTPMKIVTSNAVSWGT
LPISVALPDGTTVSDDFEGTVTSFDDDLSQSNCTIPDPSKTT
>Second Solution
*QLHTTATFW*SWPFWSTLM*PQEVQKHACQQARGKMG*ISTFISIH*KIHPHIRTQHIW
PINMPIKTS*VPLADRPISPTTIIFPVLVHQPV*LAHFLVLKKIPMVMETGDAFMKKEWV
ILILIVKILPIGVLIFLVSILLQLM*LWK*QGTFTHHRRVLTHSSLLQLTTLQFTQLVVT
LRSNVVHKNNLLSHQRILPLTVLSHGVEVCLLTSKGQSTCTLVTIIR*RLFTQMLFPGVR
FQLVWHCQMVLLLVMTLKGTFTLLTMI*VSQIVLSLILQNI
>Third Solution
DNCTPLHIFGNLGLSGAT*CSLRKTRSMPASRPEEKWDETQLLSVFIKRFIHIFERSITG
L*ICR*KQVRFR*RTDRSLHIL*TSLC*FISLCDLHISLSSRRFLW*WKLGMPL*KRNG*
TLF**SKTCLLEF*SFWFLTTSN*CNCGNDRVLFTTTDGFLHIQVCTS*RLCNSISWW*H
CVRMLCTRTTSTHINGFTH*RT*AMGWKFAT*HRRDSLHVRWLLLSDEDCLLKCCFLGTA
SN*CGIARWTTC***L*RVRLLF*R*SKSVKLTTP*SFKIT
```

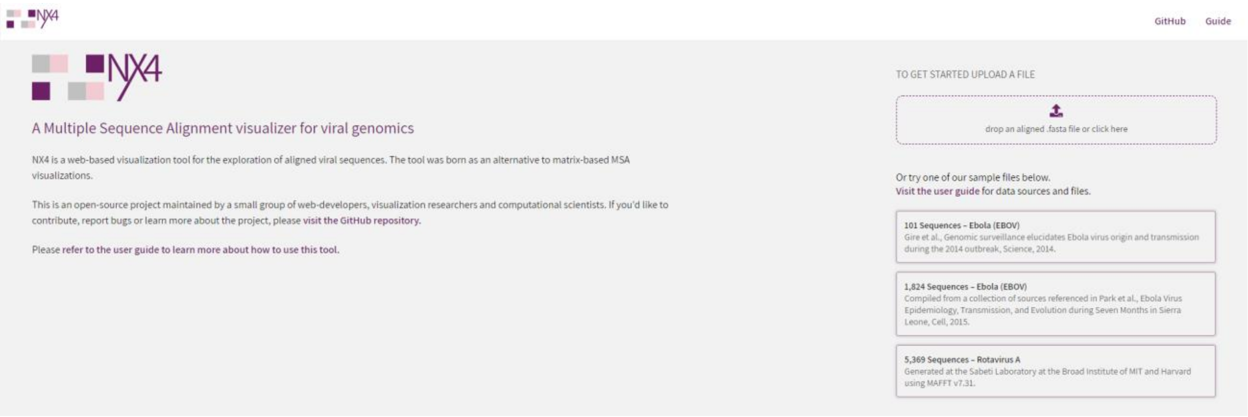
Figure 20: File generated by the previous table

Comparisons

To analyze SAVD3, a comparison was done against two other popular solutions. The purpose of this is to show why this website is a progress in the multiple sequence alignment field.

NX4

As mentioned in the introduction, NX4 is a recent solution which offers a couple of visualizations regarding sequence alignment. The first view when NX4 loads is shown below.



NX4

GitHub Guide

NX4

A Multiple Sequence Alignment visualizer for viral genomics

NX4 is a web-based visualization tool for the exploration of aligned viral sequences. The tool was born as an alternative to matrix-based MSA visualizations.

This is an open-source project maintained by a small group of web-developers, visualization researchers and computational scientists. If you'd like to contribute, report bugs or learn more about the project, please visit the GitHub repository.

Please refer to the user guide to learn more about how to use this tool.

TO GET STARTED UPLOAD A FILE

drop an aligned .fasta file or click here

Or try one of our sample files below.
Visit the user guide for data sources and files.

101 Sequences – Ebola (EBOV)
Gire et al., Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak, Science, 2014.

1,824 Sequences – Ebola (EBOV)
Compiled from a collection of sources referenced in Park et al., Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone, Cell, 2015.

5,369 Sequences – Rotavirus A
Generated at the Sabeti Laboratory at the Broad Institute of MIT and Harvard using MAFFT v7.31.

Figure 21: NX4 file selection component

Both applications offer the option to select sample files or to load one. A difference between SAVD3 and NX4 is that the latter only works with gene alignments. This narrows the possibility to analyze a vast number of alignments. NX4 has a navbar which redirects the user to its repository or site usage guide. This guide is necessary to understand what the visualizations tell. This is not needed in SAVD3 because each graph has its tooltip saying what to expect of the visualizations. An image of NX4 visualizations are shown below.

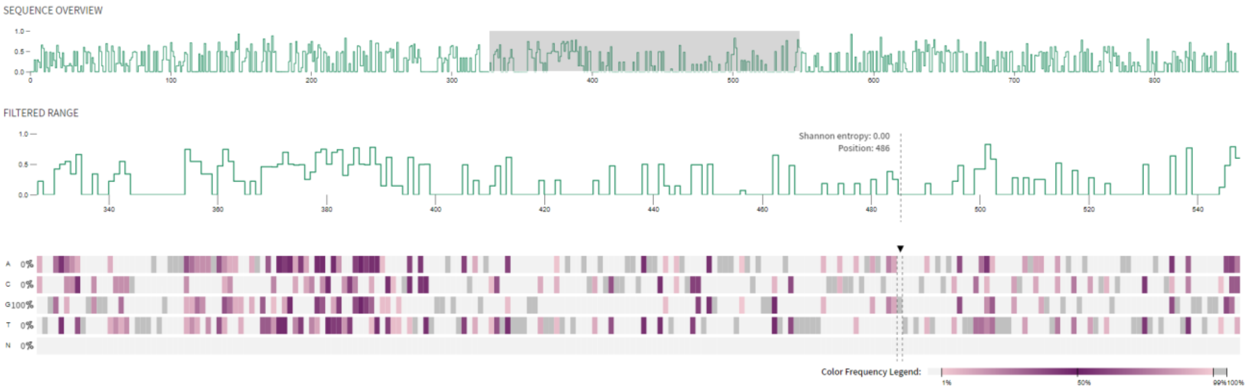
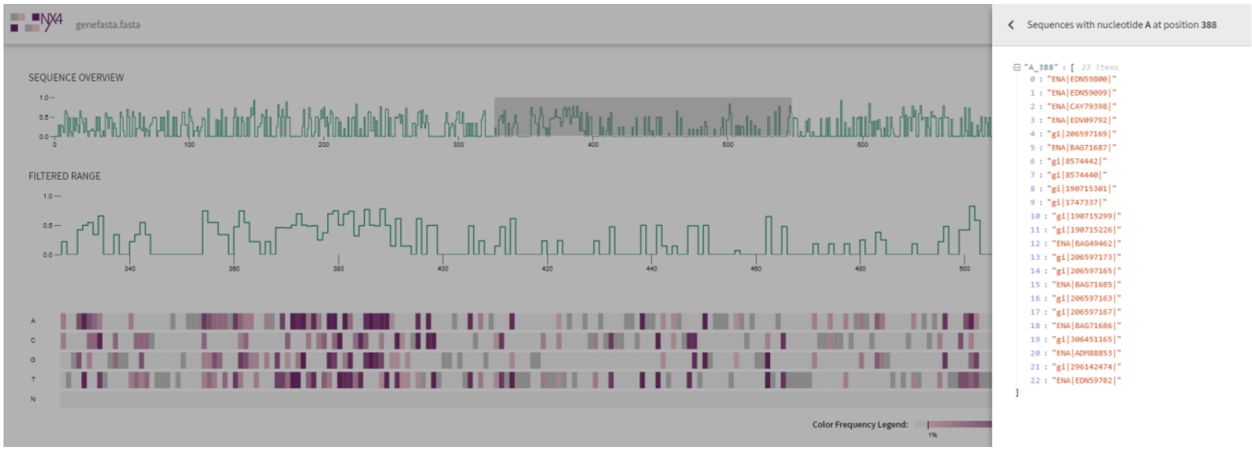


Figure 22: NX4 visualization

The first similarity between both pages is that both have an entropy graph. Because of how close the points are in NX4, it might be difficult to know if a section has a low or high entropy. SAVD3 group positions together, having a smoother visualization. A zoomed representation of the entropy graph is separated from itself in NX4, meaning that the user cannot change the default zoom. SAVD3 opts to have both representations together, and let the user zoom in if he wants to visualize a specific section.

The second graph in NX4 is a table of all possible nucleotides per position. This is like the profile weight matrix displayed in SAVD3, because the purpose of both is showing the percentage of each distinct value per position. NX4 lacks in having different colors depending on what nucleotide is being analyzed. As mentioned before, SAVD3 colors were chosen to make the website more accessible.

When the user clicks on a position in NX4 second graph, a table pops on the right side of the website, listing the names of genes having that specific nucleotide at that position. The same visualization is present SAVD3 when the user opens a specific position on the sequence matrix.



Sequences with nucleotide A at position 388

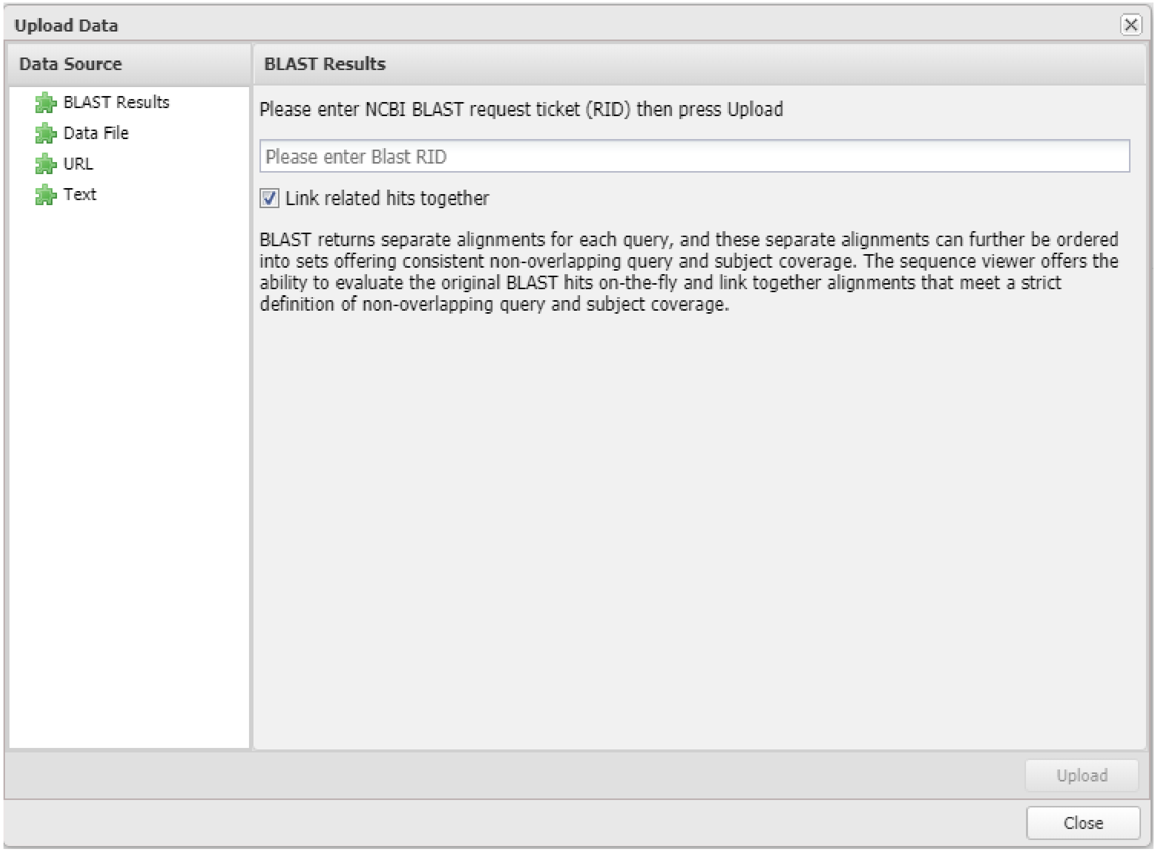
```
⊟ "A_388" : [ 23 items
    0 : "ENA|EDN59800|"
    1 : "ENA|EDN59099|"
    2 : "ENA|CAY79398|"
    3 : "ENA|EDV09792|"
    4 : "gi|206597169|"
    5 : "ENA|BAG71687|"
    6 : "gi|8574442|"
    7 : "gi|8574440|"
    8 : "gi|190715301|"
    9 : "gi|1747337|"
    10 : "gi|190715299|"
    11 : "gi|190715226|"
    12 : "ENA|BAG49462|"
    13 : "gi|206597173|"
    14 : "gi|206597165|"
    15 : "ENA|BAG71685|"
    16 : "gi|206597163|"
    17 : "gi|206597167|"
    18 : "ENA|BAG71686|"
    19 : "gi|306451165|"
    20 : "ENA|ADM88853|"
    21 : "gi|296142474|"
    22 : "ENA|EDN59782|"
]
```

Figure 23: NX4 nucleotide table

SAVD3 provides more visualizations than NX4. NX4 lacks a sequence comparison between genes. Not having a filter is a problem if the user wants to check only a subset of the loaded genes. The position filter is also important if the sequences are quite extensive and might affect the browser render speed. SAVD3 also offers the translation table, giving more insight into what the nucleotides might translate into. Finally, NX4 does not have a button for selecting a new file. The user must reload the whole browser tab in order to do this.

NCBI Multiple Sequence Alignment Viewer

NCBI also has a solution for multiple sequence alignment visualizations. Like all MSA applications, the first component the user encounter is the file selection page. NCBI offers the possibility to load files in different ways, such as providing an URL, pasting the text, loading a local file, or writing an NCBI BLAST request ticket. All these possibilities make the site accessible. NCBI MSA Viewer works with both protein and gene alignments.



BLAST Results

Please enter NCBI BLAST request ticket (RID) then press Upload

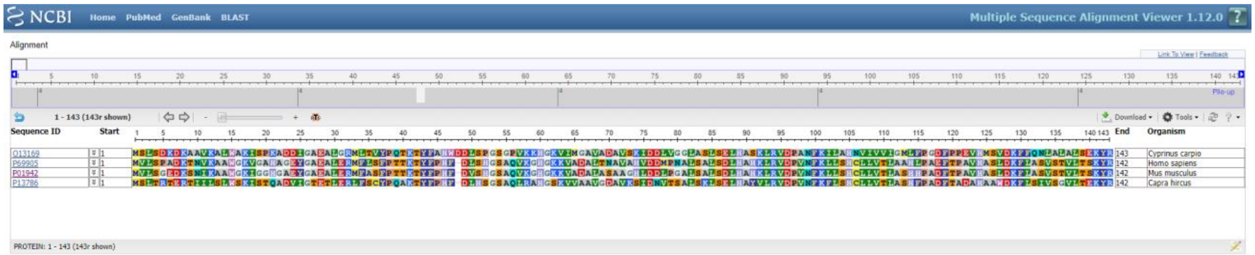
Please enter Blast RID

☑ Link related hits together

BLAST returns separate alignments for each query, and these separate alignments can further be ordered into sets offering consistent non-overlapping query and subject coverage. The sequence viewer offers the ability to evaluate the original BLAST hits on-the-fly and link together alignments that meet a strict definition of non-overlapping query and subject coverage.

Figure 24: NCBI MSA Viewer file selection component

Once loaded, only one visualization appears based on the selected file. This chart is very similar to SAVD3 sequence matrix. In both solutions, each different nucleotide or amino acid has a different color, which helps the user understand the graph. A difference between both applications is the fact that NCBI has a zoom feature. This feature was not implemented in SAVD3 because when the sequence matrix is zoom out, not relevant information can be inferred by having illegible small letters.



Sequence ID	Start	End	Organism
O13169	1	143	Cyprinus carpio
P69905	1	142	Homo sapiens
P01942	1	142	Mus musculus
P13786	1	142	Capra hircus

Figure 25: NCBI MSA Viewer - protein alignment

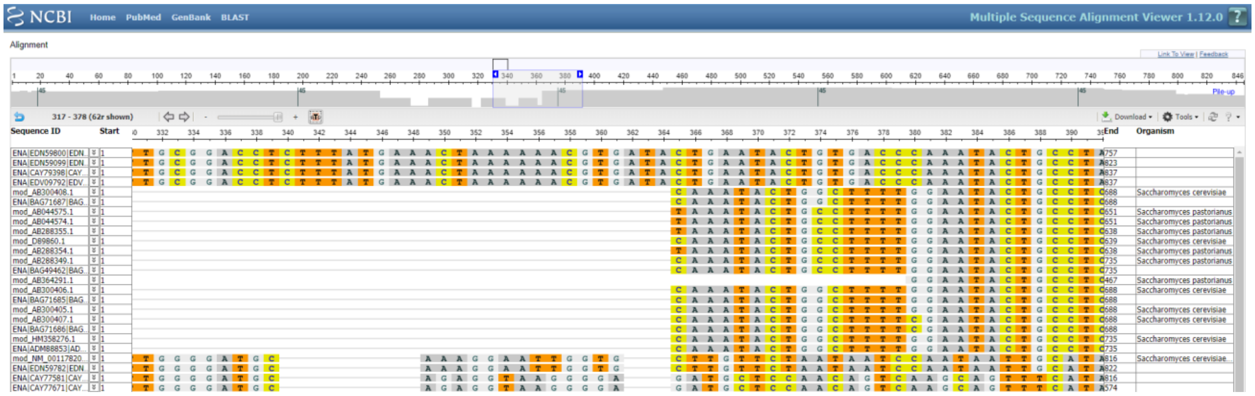
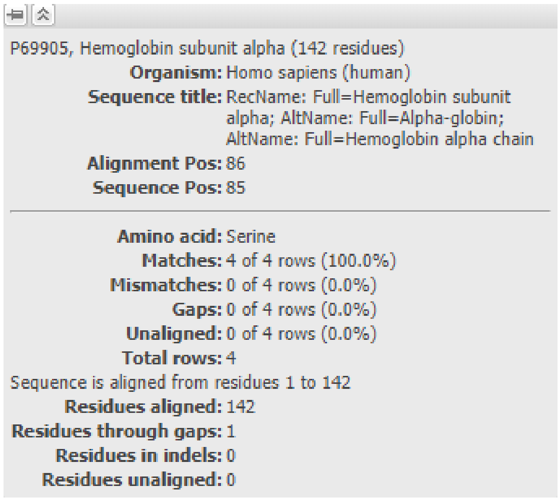


Figure 26: NCBI MSA Viewer - gene alignment

An additional feature of the NCBI website is a detailed table when the user clicks in any position within the chart. The table has relevant information such as the position, sequence name, organism name, the name of the amino acid or nucleotide, number of mismatches, gaps, and matches. In SAVD3, the names and positions are shown in the row and column headers of the sequence matrix, while the mismatches, matches, and gaps are displayed in the profile weight matrix chart.



P69905, Hemoglobin subunit alpha (142 residues)

Organism: Homo sapiens (human)

Sequence title: RecName: Full=Hemoglobin subunit alpha; AltName: Full=Alpha-globin; AltName: Full=Hemoglobin alpha chain

Alignment Pos: 86

Sequence Pos: 85

Amino acid: Serine

Matches: 4 of 4 rows (100.0%)

Mismatches: 0 of 4 rows (0.0%)

Gaps: 0 of 4 rows (0.0%)

Unaligned: 0 of 4 rows (0.0%)

Total rows: 4

Sequence is aligned from residues 1 to 142

Residues aligned: 142

Residues through gaps: 1

Residues in indels: 0

Residues unaligned: 0

Figure 27: NCBI MSA Viewer – details table

Although the NCBI solution is robust and relevant, the user needs to be able to see general information or an overview of the loaded sequences. SAVD3 offers this in all the graphs present on its website, such as the entropy or the sequence comparison graph. Additionally, when NCBI tells the number of matches or matches, they do not mention the percentage of each value, as shown in SAVD3 weight profile matrix.

References

1. Solano-Roman, A. et al. (2019) NX4: a web-based visualization of large multiple sequence alignments. *Bioinformatics, 35,* 4800–4802.
2. Yachdav, G. et al. (2016) MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics, 32,* 3501–3503.
3. Roca, A. (2013) ProfileGrids: a sequence alignment visualization paradigm that avoids the limitations of Sequence Logos. *BMC Proceedings*, 8, S6
4. Rosenberg, M. (2009) Sequence alignment: methods, models, concepts, and strategies. *Berkeley: University of California Press*
5. Batista, M. (2011) An entropy-based approach for the identification of phylogenetically informative genomic regions of Papillomavirus. *Infection, Genetics and Evolution*, 11, 2026-2033
6. Gray, R. (1990) *Entropy and Information Theory.* Springer, New York, London
7. Heath, T. (2006) Multiple Sequence Alignment Accuracy and Phylogenetic Inference. *Systematic Biology*, 55, 314-328
8. Phillips, A. (2000) Multiple sequence alignment in phylogenetic analysis. *Molecular Phylogenetics and Evolution*, 16, 317-330