

MSA Class Logos – Step-by-step

MSA Class Logos is a free webserver allowing the user to group amino acid sequences into classes and to generate sequence logos for each class and the entire multiple sequence alignment. The input files are a multiple sequence alignment in FASTA format and a file that specifies the class each sequence belongs to in CSV format. After file upload, one reference sequence needs to be selected. This sequence will be displayed separately for comparison. This can be any sequence from the multiple sequence alignment and is just for convenience in case it is interesting to compare a certain sequence to the sequence logos of each class. If this is not of importance, any sequence can be selected since it does not affect the generation of the logos and therefore the results displayed. There are two tabs with different results available. The first one shows the complete multiple sequence alignment with all sequences sorted according to the classes defined by the user and a sequence logos considering all sequences. The second one shows separated sequence logos for each defined class which allows the identification of positions that are conserved only in a subset of sequences but not in all. In other words, the second view (tab) allows a more detailed analysis of a multiple sequence alignment that permits the identification of class specific conserved residues in evolutionary and/or functionally variable protein groups. Results can be reviewed visually directly on the web page or downloaded as tables in CSV format. Additionally, sections of each sequence logo can be downloaded in SVG and used for the generation of figures for publications or presentations, for example.

This step-by-step description of how to use *MSA Class Logos* uses the example files available on the web server.

File preparation

Two files are required to run *MSA Class Logos*: (1) a multiple sequence alignment file in FASTA format, and (2) a file specifying the class each sequence belongs to in CSV format. Note: *MSA Class Logos* is not able to align sequences. This step needs to be performed previously using external algorithms like MAFFT, MUSCLE or CLUSTALW among many others.

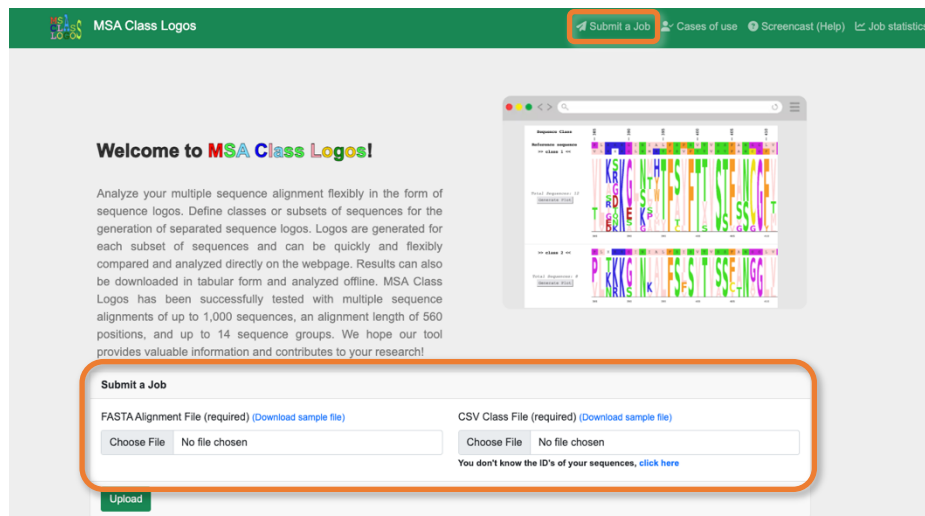


Figure 1: Submitting a job.

Files can be uploaded directly on the first page after loading the *MSA Class Logos* webpage. Should you not see the first page, you can click on 'Submit a Job' to reach the initial page (**Figure 1***Error! Reference source not found.*).

(1) FASTA Alignment File (required):

For the tutorial, click on **Download sample file** next to **FASTA Alignment File (required)**. A FASTA file will be downloaded (**Figure 2**).

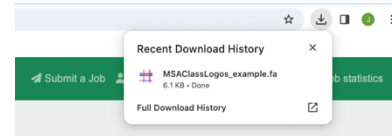


Figure 2: Download example sequence alignment.

(2) CSV Class File (required):

The class file has a simple structure of two columns. Column 1 contains the sequence IDs used in the alignment file, column 2 contains the name of the class each sequence belongs to. Sequence IDs in alignment and class file must be identical. It is recommended to extract sequence IDs using the script implemented on the webpage. Click on 'click here' in **You don't know the ID's of your sequences, click here**. A window opens where the just downloaded alignment file can be dragged and dropped into. Click **Upload** (**Figure 3A**).

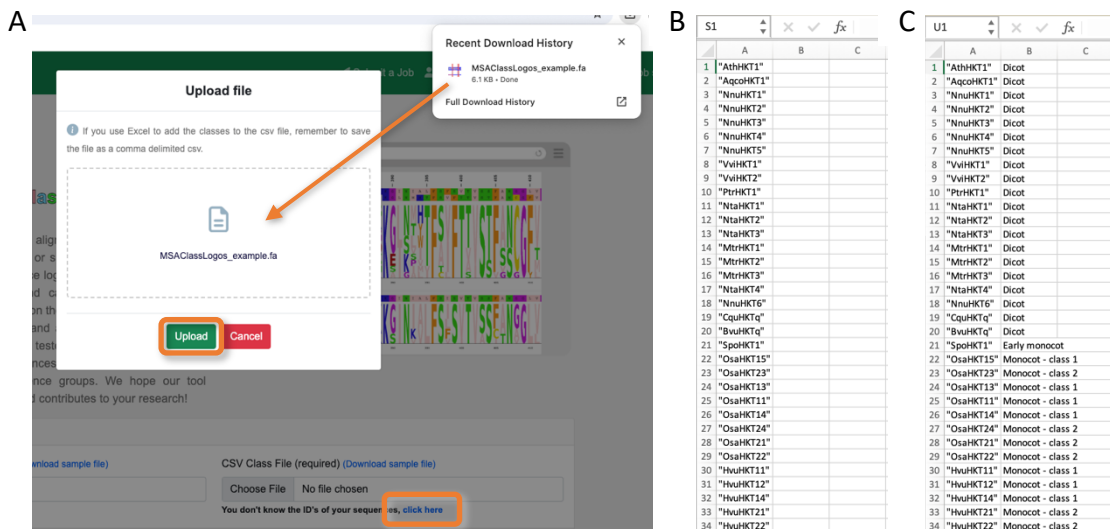


Figure 3: (A) Extracting sequence IDs using the script implemented on MSAClassLogos. (B) Sequence IDs in first column. (C) Class names associated to each sequence ID in second column.

Open the class file in, for example, Excel. The first column contains the sequence IDs extracted from the alignment file (**Figure 3B**). The second column needs to be filled by the user with specific class identifiers (**Figure 3C**). That may be a number or a word. All sequences with the same class identifier will be grouped later. Save the file as a comma delimited CSV format. For the tutorial, you can download the example class file by clicking on **Download sample file** next to **CSV Class File (required)** (**Figure 4**).

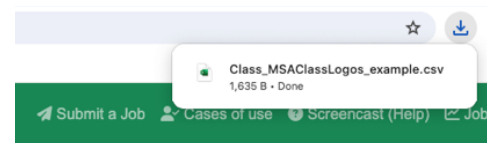
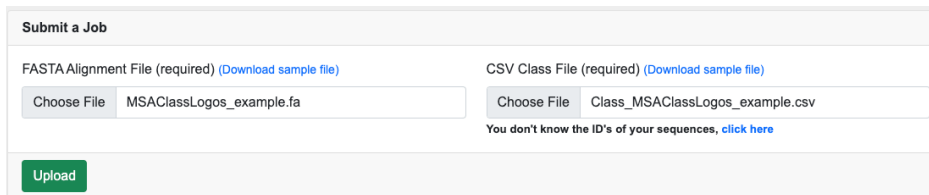


Figure 4: Download example class file.

Uploading files

On the first page of the webpage or after clicking on Submit a Job you can upload the example alignment and class file (**Figure 5**).

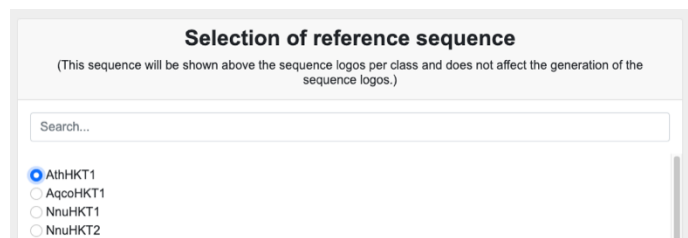


The screenshot shows a web form titled "Submit a Job". It has two main sections for file uploads. The first section is for the "FASTA Alignment File (required)" with a "Choose File" button and the filename "MSAClassLogos_example.fa". The second section is for the "CSV Class File (required)" with a "Choose File" button and the filename "Class_MSAClassLogos_example.csv". Below these sections is a green "Upload" button. There are also links for "Download sample file" and a note: "You don't know the ID's of your sequences, click here".

Figure 5: Uploading sequence alignment and class file.

Selection of reference sequence and revision

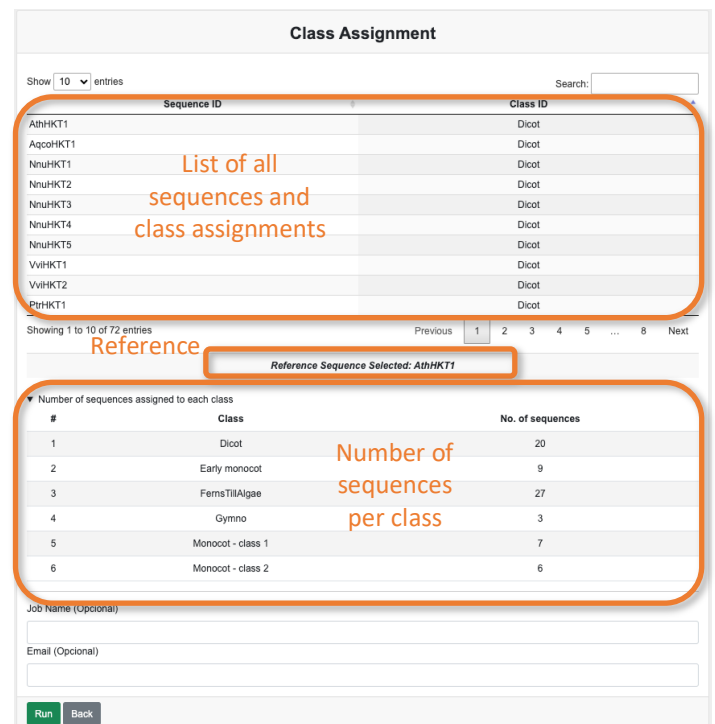
After uploading the files, one reference sequence needs to be selected (**Figure 6**). This can be any sequence from the multiple sequence alignment and is just for convenience in case it is interesting to compare a certain sequence to the sequence logos of each class. The selected sequence will be displayed separately for comparison. If this is not of importance, any sequence can be selected since it does not affect the generation of the logos and therefore the results displayed. Note the search function which is helpful in case of large alignments. Click **Select and next**.



The screenshot shows a section titled "Selection of reference sequence". Below the title is a note: "(This sequence will be shown above the sequence logos per class and does not affect the generation of the sequence logos.)". There is a search bar labeled "Search...". Below the search bar is a list of sequences with radio buttons: "AthHKT1" (selected), "AqcoHKT1", "NnuHKT1", and "NnuHKT2".

Figure 6: Selection of reference sequence.

On the following page a summary of all sequences, classes and the selected reference sequence is given (**Figure 7**). In case any error is detected you can go back and correct it. Optionally, a job title and an email address can be provided. In this case, you will be notified once the job is done. The email contains a link that leads to the results, which will be stored for one month on the server. Generally, the calculation of the sequence logos takes a few minutes dependent on the number of sequences in the alignment and the length of the alignment. *MSA Class Logos* has been successfully tested with multiple sequence alignments of up to 1,000 sequences with an alignment length of 560 positions, and up to 14 sequence groups. If everything is well selected press **Run**.



The screenshot shows a page titled "Class Assignment". It features a table with columns "Sequence ID" and "Class ID". The table lists sequences like AthHKT1, AqcoHKT1, NnuHKT1, NnuHKT2, NnuHKT3, NnuHKT4, NnuHKT5, VvHKT1, VvHKT2, and PthHKT1, all assigned to the "Dicot" class. Below the table is a "Reference" section with a box indicating "Reference Sequence Selected: AthHKT1". At the bottom, there is a table titled "Number of sequences assigned to each class" with columns "#", "Class", and "No. of sequences".

#	Class	No. of sequences
1	Dicot	20
2	Early monocot	9
3	FernsTillAlgae	27
4	Gymno	3
5	Monocot - class 1	7
6	Monocot - class 2	6

Figure 7: Summary and revision of sequences and their assigned classes.

Results based on the entire multiple sequence alignment

These results are visualized in the **All Sequences** tab.

(a) Visualization:

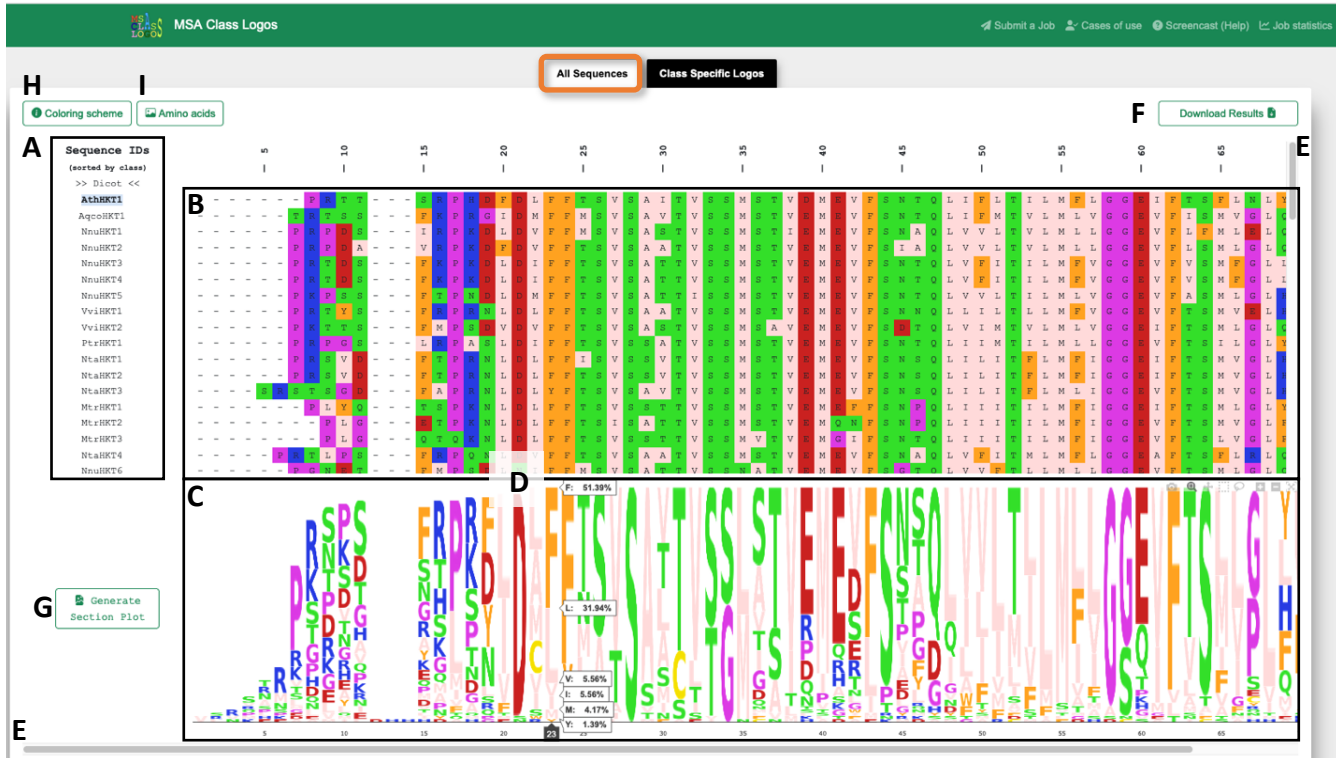


Figure 8: Results for MSA of all sequences. Sequences are sorted according to the assigned classes (A). MSA is coloured according to the Zappo colouring scheme (B, H). Sequence logos are represented below the MSA using the same colouring scheme (C). When moving the mouse over a position of the sequence logos amino acid frequencies for this position are displayed (D). MSA and sequence logos can be explored by using the scroll bars on the bottom and right (E). Results can be downloaded for offline analysis in tabular form (F). Graphical representation of sequence logos can be exported in SVG format (G). A chart of amino acid structures is available for quick structure references (I).

The first result tab shows the entire multiple sequence alignment with all sequences sorted according to the classes defined by the user (**Figure 8** (A)). Amino acids are colored according to the Zappo scheme (click on Coloring Scheme on the upper left for more details (H)) (B). Below the sequence alignment is the sequence logo considering all sequences (C). When moving the mouse over an amino acids position in the sequence logo, the frequency in which amino acids occur in this position is shown (D). Gaps are indicated as – and also considered in the calculation of frequencies. On the top left, clicking on **Amino acids** shows the 20 structures of proteinogenic amino acids for quick reference (I). Results can be explored directly on the webpage by scrolling through the sequence alignment and sequence logo (E).

(b) Download of amino acid frequencies per position:

Results that can be visually inspected on the webpage but also be downloaded as tables in CSV format from the **Download Results** button (F). The file **Amino acid frequencies complete alignment** contains all amino acid frequencies for each position of the alignment and, therefore, contains all information that is visualized on the webpage (**Figure 9A**). All other files containing a specified percentage in their name,

contain only a subset of positions applying the indicated filter. For example, the file **Positions 100% conserved** reports only those positions of the alignment that contain amino acids with a frequency of 100% (Figure 9B). In other words, 100% conserved positions. The file **Positions ≥90% conserved** reports all position where one amino acid appears with a frequency of 90% or higher (Figure 9C). And so on. The CSV files can be downloaded and further processed or filtered in, for example, Excel.

A Amino acid frequencies complete alignment (CompleteAlignment_ConservationAll.csv)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	All MSA	# 72 sequences																		
2	Position in alignment	Conservation grade (residue/%)																		
3		1 -/97.22	V/2.78																	
4		2 -/95.83	L/1.39	R/1.39	S/1.39															
5		3 -/94.44	R/2.78	N/1.39	V/1.39															
6		4 -/88.89	S/2.78	K/1.39	M/1.39	P/1.39	Q/1.39	R/1.39	V/1.39											
7		5 -/81.94	A/2.78	N/2.78	R/2.78	T/2.78	E/1.39	H/1.39	L/1.39	P/1.39	S/1.39									
8		6 -/76.39	R/11.11	M/4.17	N/2.78	E/1.39	K/1.39	P/1.39	S/1.39											
9		7 /96.11	-/33.33	R/6.94	K/5.56	L/2.78	S/2.78	T/2.78	V/2.78	A/1.39	D/1.39	E/1.39	G/1.39	M/1.39						
10		8 /20.83	K/13.89	-/13.89	S/9.72	G/8.33	T/8.33	P/5.56	D/4.17	H/4.17	L/2.78	N/2.78	Q/2.78	A/1.39						E/1.39
11		9 S/15.28	N/11.11	D/9.72	P/9.72	T/9.72	R/8.33	-/8.33	G/6.94	K/6.94	E/5.56	L/4.17	V/2.78	M/1.39						
12		10 P/15.28	K/11.11	S/9.72	D/8.33	-/6.94	G/5.56	N/5.56	T/5.56	V/5.56	A/4.17	E/4.17	H/4.17	R/4.17	L/2.78	Y/2.78	I/1.39	M/1.39	Q/1.39	
13		11 S/23.61	D/11.11	G/9.72	T/9.72	A/5.56	H/5.56	-/5.56	K/4.17	P/4.17	Q/4.17	V/4.17	E/2.78	V/2.78	N/2.78	R/2.78	L/1.39			
14		12 -/98.61	D/1.39																	
15		13 -/98.61	H/1.39																	
16		14 -/98.61	H/1.39																	
17		15 /22.22	S/9.72	G/8.33	N/8.33	-/8.33	R/6.94	A/5.56	E/4.17	K/4.17	Y/4.17	D/2.78	M/2.78	P/2.78	Q/2.78	H/1.39	I/1.39	L/1.39	T/1.39	V/1.39
18		16 R/25.0	T/12.5	H/11.11	S/9.72	K/8.33	G/6.94	-/5.56	A/4.17	M/4.17	Q/4.17	N/2.78	P/2.78	D/1.39	Y/1.39					
19		17 P/47.22	L/26.39	M/6.94	-/5.56	I/4.17	F/2.78	V/2.78	A/1.39	Q/1.39	T/1.39									
20		18 K/19.44	R/19.44	S/13.89	P/11.11	T/8.33	D/5.56	N/5.56	A/4.17	G/4.17	-/4.17	H/1.39	M/1.39	Q/1.39						
21		19 F/23.61	D/22.22	Y/18.06	N/16.67	L/4.17	-/4.17	Q/2.78	R/2.78	S/2.78	G/1.39	P/1.39								
22		20 L/44.44	I/31.94	V/12.5	F/4.17	-/2.78	E/1.39	M/1.39	T/1.39											
23		21 D/94.44	-/2.78	N/1.39	S/1.39															
24		22 L/26.39	A/23.61	M/15.28	C/13.89	V/8.33	I/6.94	S/2.78	R/1.39	W/1.39										
25		23 /51.39	L/31.94	I/5.56	V/5.56	M/4.17	Y/1.39													
26		24 F/75.0	Y/25.0																	
27		25 T/44.44	L/13.89	M/13.89	N/13.89	I/4.17	A/2.78	F/2.78	G/1.39	S/1.39	Y/1.39									
28		26 S/63.89	A/25.0	V/5.56	I/4.17	P/1.39														
29		27 V/69.44	T/22.22	A/4.17	I/2.78	L/1.39														
30		28 S/100.0																		

B Positions 100% conserved (CompleteAlignment_Conservation100.csv)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	All MSA	# 72 sequences																		
2	Position in alignment	Conservation grade (residue/%)																		
3		28 S/100.0																		
4																				

C Positions ≥90% conserved (CompleteAlignment_Conservation90.csv)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	All MSA	# 72 sequences																		
2	Position in alignment	Conservation grade (residue/%)																		
3		21 D/94.44	-/2.78	N/1.39	S/1.39															
4		28 S/100.0																		
5		55 M/91.67	S/2.78	F/1.39	L/1.39	T/1.39	V/1.39													
6		58 G/93.06	A/4.17	D/1.39	S/1.39															
7																				

Figure 9: Structure of downloadable results of amino acid frequencies per MSA position. (A) File of frequencies of all MSA position. Column A states the position of the MSA. From column B onwards, amino acids and their frequency of appearance in this particular position are indicated (amino acid / frequency in %). The more cells contain amino acid frequency information per MSA position, the higher the variability of this position or the less conserved is this position. In case of a 100% conserved position only one cell will contain information. Compare positions 15 (highly variable) and 28 (highly conserved). For simplicity only the first 28 positions of the sequence alignment are shown. (B) This file contains only the positions that are 100% conserved in the MSA. In this example, this is only one position. (C) This file contains only the positions with amino acid frequencies of 90% or higher in the MSA. In this example, this are only four positions.

(c) Download of selected sequence logo regions:

Once an interesting region is identified, the sequence logo of these positions can be downloaded as SVG by clicking on the **Generate Section Plot** button on the left to the sequence logo (Figure 8G). A window opens where the initial and end position can be specified. By pressing **Generate Plot** the SVG file is generated. This takes a short moment depending on the length of sequence range. Please, wait while the file is generated. The scalable vector graphic (SVG) file can be further processed in adequate programs to generate high resolution figures for scientific publications and presentations.

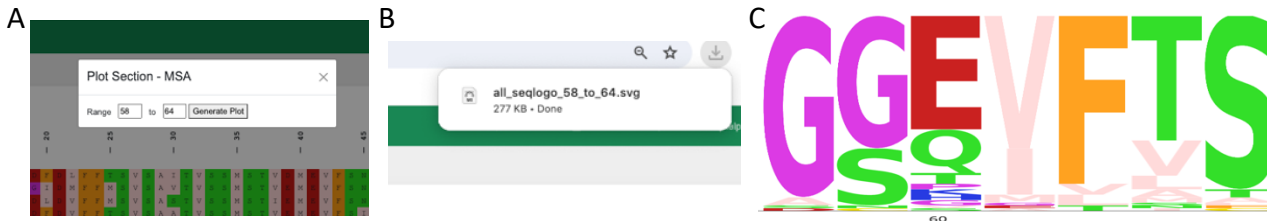


Figure 10: Generation and export of sequence logo plot. (A) The sequence logo range that shall be exported can be specified, for example position 58 to 64. (B) Within a short time (depending on the selected length of the sequence logo) an SVG file is generated. (C) The SVG file can be incorporated into publication figures or scientific presentation, for example.

Results per defined sequence class

These results are visualized in the **Class Specific Logos** tab.

The visualization (a) of the results is slightly different in this tab, while the downloadable files (b) and the generation of the sequence logo plots (c) is identical to what is described in the previous section. The initially selected reference sequence is the top sequence displayed (Figure 11A). Below the reference sequence are represented a sequence logo (Figure 11B) and a consensus sequence for each sequence class specified initially in the class file (Figure 11C). The name of the class and the number of sequences selected for this class are indicated on the left (Figure 11D). The consensus sequence shows amino acids with highest frequency for each position in case this frequency is higher than 50%. If the amino acid with



Figure 11: Results of sequence logo representation per class. Sequence logos for each class are represented based on the class assignments done during the first step (B). The initially selected reference sequence is shown on top (A). For each sequence class logo, a consensus sequence is displayed where highly variable positions (amino acid frequencies lower than 50%) are marked with an X (C). The colouring scheme, result download and sequence plot generation are equal to the All Sequences tab. The number of sequences used for the generation of sequence logos is stated on the left of each logo (D). The small square symbol with three horizontal lines in it enables hiding sequence logos and consensus sequences (E). The Table Reset button on the top reveals all hidden information (F).

the highest frequency has a value below 50% the position is marked with an X and considered as non-conserved. The sequence logos have the same features as in the previously described results page. The advantage of the visualization per class is that sequences grouped according to sequence similarity, functional similarity or other aspects can be analyzed in more detail and contrasted and compared against other groups, or the complete alignment. This can be done quick and easy directly on the webpage or by analyzing the downloadable frequency tables. For example, positions that are conserved in only some classes but not others can be easily identified in the **Class Specific Logos** tab although they appear as seemingly not highly conserved in the sequence logos based on the entire alignment (**All Sequences** tab). For example, position 34 in the example alignment shows that 50% of the sequences contain a serine, while the other 50% contain a glycine residue at this position as observed in the All Sequences tab (**Figure 12A**). When analyzing the Class Specific Logos tab it is evident that some sequence classes have a clear preference for one of the two residues (**Figure 12B**). The classes Dicot and Monocot - class 1 express only serine residues, and FernsTillAlgae only glycine. Monocot - class 2 have a preference for glycine, while the class Gymno and Early Monocot accept both residues in their sequences. To simplify the comparison of specific groups it is possible to hide consensus sequences and logos by clicking on the square symbol with three horizontal lines in it that appears on the left of each sequence logo and consensus sequence (**Figure 11E**). The **Table Reset** button on the top restores all hidden information (**Figure 11F**).

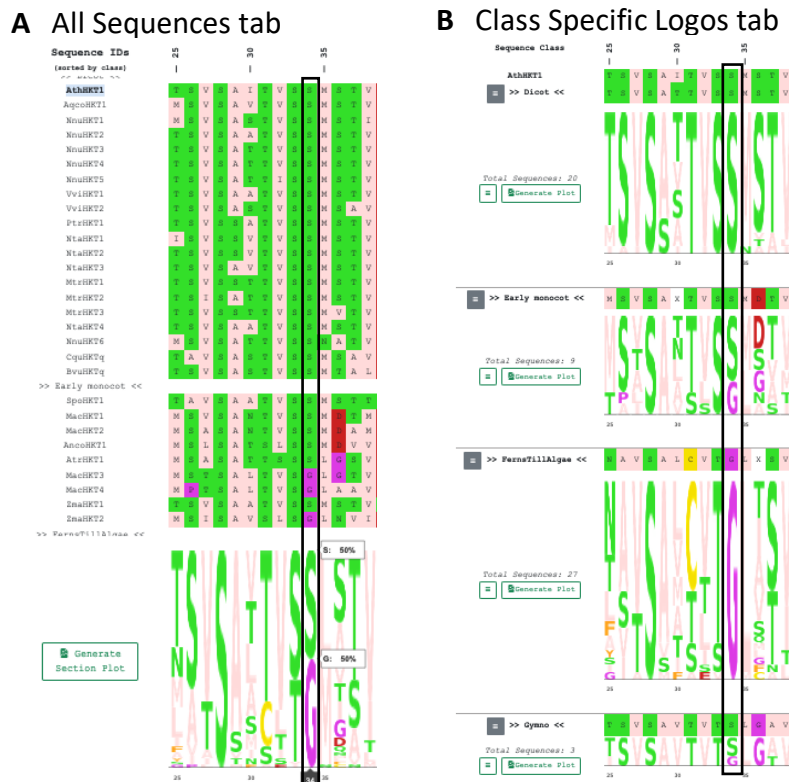


Figure 12: Comparison of All Sequence and Class Specific Logos tab on the example of position 34. (A) The All Sequence tab illustrates that 50% of the sequences contain serine residues, and the other 50% glycine residues. (B) The Class Specific Logos tab illustrates that some sequence classes conserve the serine or the glycine, while others accept both residues.