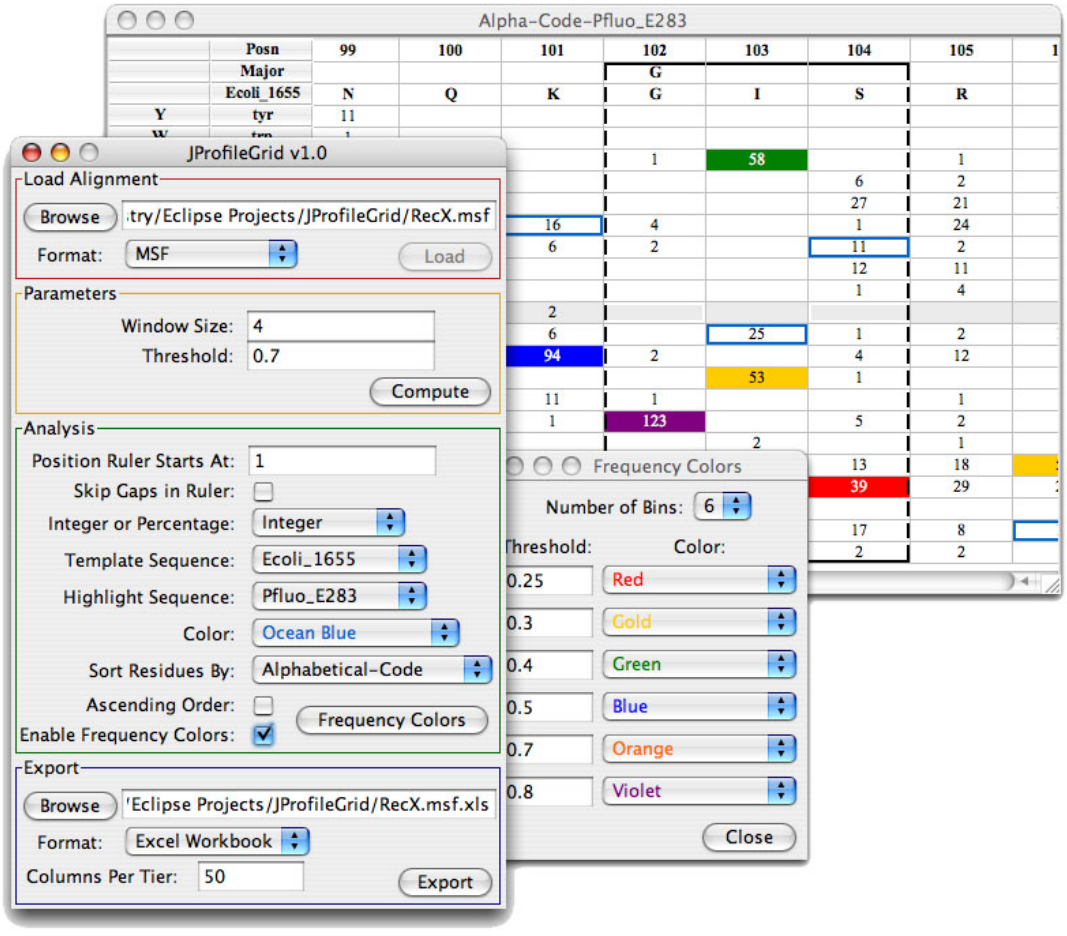
JProfileGrid Manual

visualization and analysis of large multiple sequence alignments

Alberto I. Roca and Aaron C. Abajian

www.profilegrid.org • Software Version 1.2.2 • Manual Version August 19, 2009

Citation: A.I. Roca, A.E. Almada, and A.C. Abajian (2008) ProfileGrids as a new visual representation of large multiple sequence alignments: a case study of the RecA protein family, *BMC Bioinformatics* **9:** 554.



Legal Notice

Software Copyright © 2007 The Regents of the University of California. All Rights Reserved.

The ProfileGrid *representation* is in the public domain if others wish to code their own implementations.

Permission to use, copy, modify, and distribute this software and its documentation for educational, research and non-profit purposes, without fee, and without a written agreement is hereby granted, provided that the above copyright notice, this paragraph and the following three paragraphs appear in all copies.

Permission to use this software for commercial purposes may be obtained by contacting profilegrid.org.

This software program and documentation are copyrighted by The Regents of the University of California. The software program and documentation are supplied ″as is″, without any accompanying services from The Regents. The Regents does not warrant that the operation of the program will be uninterrupted or error-free. The end-user understands that the program was developed for research purposes and is advised not to rely exclusively on the program for any reason.

IN NO EVENT SHALL THE UNIVERSITY OF CALIFORNIA BE LIABLE TO ANY PARTY FOR DIRECT, INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES, INCLUDING LOST PROFITS, ARISING OUT OF THE USE OF THIS SOFTWARE AND ITS DOCUMENTATION, EVEN IF THE UNIVERSITY OF CALIFORNIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE. THE UNIVERSITY OF CALIFORNIA SPECIFICALLY DISCLAIMS ANY WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE AND ANY STATUTORY WARRANTY OF NON-INFRINGEMENT. THE SOFTWARE PROVIDED HEREUNDER IS ON AN ″AS IS″ BASIS, AND THE UNIVERSITY OF CALIFORNIA HAS NO OBLIGATIONS TO PROVIDE MAINTENANCE, SUPPORT, UPDATES, ENHANCEMENTS, OR MODIFICATIONS.

Introduction

Large multiple sequence alignments (MSAs) are taxing current visualization methods such as the traditional stacked sequence depiction. ProfileGrids are a new paradigm for visualizing and analyzing MSAs in a concise representation. ProfileGrids represent a multiple sequence alignment as a matrix color-coded according to the residue frequency occurring at each column position. This allows all the sequence information from a large MSA to be visualized more effectively than other representations such as stacked sequences, consensus sequences, major components, Sequence Logos, partial order graphs, Base-By-Base summaries, and similarity plots. ProfileGrids were invented in 2005 as part of an updated comparative nanoanatomy analysis of the bacterial RecA protein family (Roca, *et al*, 2008; Roca and Cox, 1997).

JProfileGrid is a Java application for computing and analyzing ProfileGrids. For example, inspection of a ProfileGrid allows one to identify conserved motifs and sequence-specific residues within the context of a MSA. Furthermore, the JProfileGrid graphical user interface enables the interactive analysis of structural patterns by using residue physical properties. See the "Examples" page of the website profilegrid.org for more descriptions of ProfileGrid use.

Getting Started

Download software

JProfileGrid is distributed as a single Java JAR file available from the "Downloads" page at:

http://www.profilgrid.org

A PDF documentation file (this document) is also available at the website.

The Java Runtime Environment version 5.0 or later is required. For the most recent version visit:

http://java.sun.com/javase/downloads/index.jsp

Running JProfileGrid

JProfileGrid provides both a graphical user interface (GUI) and command-line functionality. To launch the GUI under MacOS or most Windows environments, double-click on the "JProfileGrid.jar" file icon. Alternatively launch the GUI from a terminal, using the command:

```
% java -jar JProfileGrid.jar
```

The command-line version is invoked by including parameters in the general form:

```
% java -jar JProfileGrid.jar msaFilename [-parameter value(s)]
```

The input filename is a required parameter. Table 1 lists optional parameters and their values.

If not specified, default command-line parameter values are assumed to be the following:

```
-w 5 -t 0.7 -c 50 -h 1 -p 1 -ps 1 -d 0 -v 0 -s 0
```

The default output filename is the input filename appended with ".xls".

Note that there is currently no way to change the color schemes from the command-line. Also, PyMOL script output currently can not be invoked.

Table 1. Command-line parameters

Code	Name	Values
-help	Help	(minimal description of parameters)
-w	WindowSize	odd integers, *e.g.* 1, 3, ...
-t	Threshold	real number from 0 to 1
-c	ColumnsPerTier	integer from 1 to alignment length
-h	HighlightSequence	denote by integer from 1 to (number of sequences)
-a	AscendingSort	(no value specified)
-p	TemplateSequence	denote by integer from 1 to (number of sequences)
-ps	PositionRowStart	integer from 1 to alignment length

Code	Name	Values
-g	SkipPositionGaps	(no value specified)
-d	AlignmentType	0 for protein, 1 for nucleic acid
-v	ShowValuesAs	0 for integers, 1 for frequency, 3 for none
-s	SortType	0 = Alphabetical-Code 1 = Alphabetical-Name 2 = Age 3 = Flexibility 4 = Frequency-EcoliK12 5 = Hydropathy 6 = Hydrophobicity 7 = Helix-Propensity 8 = Mutability-Dayhoff 9 = Mutability-Grantham 10 = Surface-Area 11 = Volume

For **very large alignments**, one should include the Java "Xmx" parameter to allocate more memory:

```
% java -Xmx512M -jar jprofilegrid.jar
```

Here, 512 MB of memory is specified. Examples of large alignments include the following:

- 1000's of sequences of average length 300 residues, or
- a few, long sequences such as titin (> 10,000 residues)

Input File formats

Valid input file formats are MSF (*i.e.* GCG's PILEUP) and aligned FASTA. Both protein and nucleic alignments are supported although the data type must be specified. Sequence names are taken from MSF file headers or FASTA file comments. For the latter, the sequence name is interpreted to start after ">" and *end* at the first non-"0-9", "A-Z", "a-z", or "_" (underscore) character. JProfileGrid reads sequence-weighting information from MSF files for the similarity plot calculations. Since FASTA files do not specify weights, values of 1.0 are assigned to the sequences. The command-line interface expects input filename extensions of ".msf" and ".afa" for MSF and aligned FASTA formats, respectively.

The protein characters recognized are the following:

A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y

The nucleic acid characters recognized are the following:

A, B, C, D, G, H, K, M, N, R, S, T, U, V, W, Y

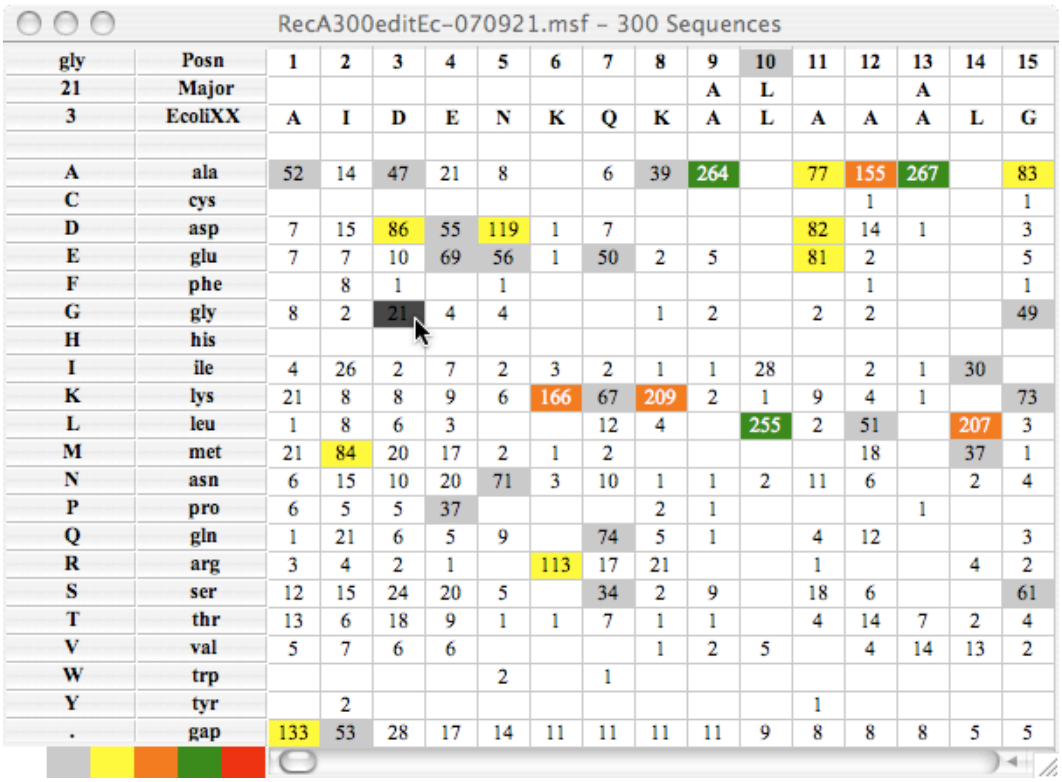
The gap characters recognized are the following:

asterisk (*), period (.), dash (-)

Any other characters are "flagged" and listed in a separate worksheet ("Flags") in the spreadsheet output. The output is listed as flagged character, sequence name, and position(s) of the flagged character. Two examples from the recent RecA alignment (Roca, *et al*, 2008) are "X Bsacc 154 183 232" and "X Lmono_51 265" from the GenBank records CAD79373.1 and AAN06665.1, respectively. Note that currently flag information is exported only if the similarity boxes are also calculated. Keep in mind that some alignment programs (such as *muscle*) will convert a "J" character in a sequence to "X" in the generated output file.

Using the Graphical Interface

The ProfileGrid viewer window (Figure 1) contains the results of the JProfileGrid calculations. In this figure a protein alignment of 300 bacterial RecA sequences was analyzed. The ProfileGrid matrix begins with three rows consisting of a position ruler, a majority consensus sequence, and a template sequence (here the *Escherichia coli* RecA homolog). The 21 rows under the template sequence represent the frequency of the amino acid and gap characters at the corresponding MSA column position. Each cell is colored according to a Frequency Colors Legend (described in detail below). In the lower left-hand corner of the ProfileGrid window is a "mini-legend" of the color scheme that is read from left to right as low to high conservation, respectively. The top left-hand corner displays the currently selected (dark grey) cell's residue, frequency value, and position for a particular column. See Figure 4 for an example of this "cell value identification" function. The features of the ProfileGrid are controlled from the Parameter window.



gly	Posn	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
21	Major									A	L			A		
3	EcoliXX	A	I	D	E	N	K	Q	K	A	L	A	A	A	L	G
A	ala	52	14	47	21	8		6	39	264		77	155	267		83
C	cys												1			1
D	asp	7	15	86	55	119	1	7				82	14	1		3
E	glu	7	7	10	69	56	1	50	2	5		81	2			5
F	phe		8	1		1							1			1
G	gly	8	2	21	4	4			1	2		2	2			49
H	his															
I	ile	4	26	2	7	2	3	2	1	1	28		2	1	30	
K	lys	21	8	8	9	6	166	67	209	2	1	9	4	1		73
L	leu	1	8	6	3			12	4		255	2	51		207	3
M	met	21	84	20	17	2	1	2					18		37	1
N	asn	6	15	10	20	71	3	10	1	1	2	11	6		2	4
P	pro	6	5	5	37				2	1				1		
Q	gln	1	21	6	5	9		74	5	1		4	12			3
R	arg	3	4	2	1		113	17	21			1			4	2
S	ser	12	15	24	20	5		34	2	9		18	6			61
T	thr	13	6	18	9	1	1	7	1	1		4	14	7	2	4
V	val	5	7	6	6				1	2	5		4	14	13	2
W	trp					2		1								
Y	tyr		2									1				
.	gap	133	53	28	17	14	11	11	11	11	9	8	8	8	5	5

Figure 1. ProfileGrid window

The Parameter window is organized in the steps necessary for using JProfileGrid (Figure 2).

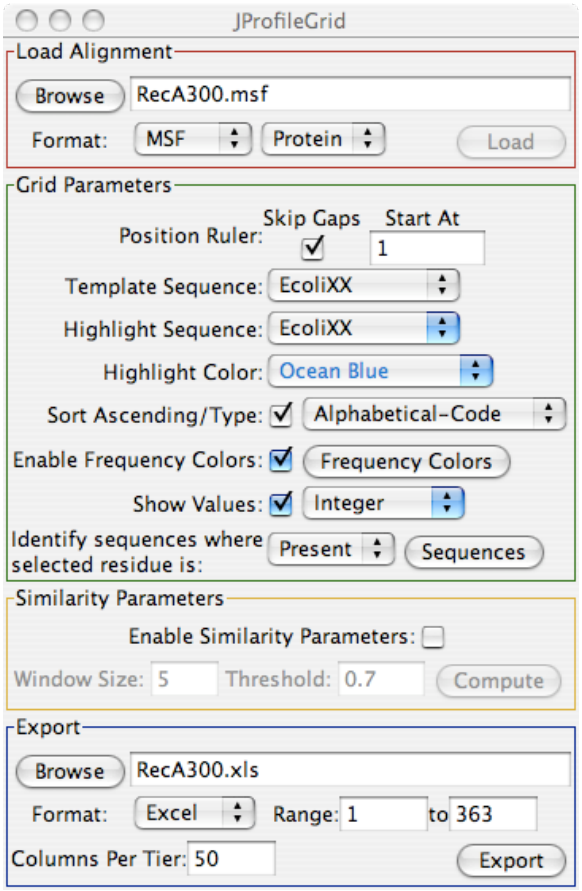


Figure 2. Parameter window

Load Alignment

Use the "Browse" button to open a file dialog window for specifying the alignment file. Indicate the file format (MSF or FASTA) and sequence type (Protein or DNA/RNA) and then "Load" the file.

Grid Parameters

Position Ruler. The position ruler numbers the ProfileGrid columns and is dependent upon the chosen template sequence. Numbering will start using the indicated value at the first non-gap character of the template sequence. The "Skip Gaps" option will account for insertions in the template sequence.

Template Sequence. By default, the first sequence in the alignment is the "template sequence". It specifies which sequence is displayed at the top of the ProfileGrid. If desired, use the menu to select another sequence as the template. For long sequence lists, type the first few letters of the sequence name to jump to that position in the menu.

Sort Ascending/Type. JProfileGrid provides a number of parameters for sorting the ProfileGrid rows. The default value is the alphabetical order of the residue one-letter code abbreviations. Use the menu to choose among the following residue (amino acid) constants: Alphabetical-Code (one-letter), Alphabetical-Name (three-letter), Age (Trifonov, 2004), Flexibility (Zhao, 2001), Frequency-EcoliK12 (Nakamura, 2000), Hydropathy (Kyte, 1982), Hydrophobicity (Sweet, 1983), Helix-Propensity (Rohl, 1996), Mutability-Dayhoff (1978), Mutability-Grantham (1974), Surface-Area (Chothia, 1975), Volume (Richards, 1974). Certain constants can not be chosen for nucleic acid sequences. See Table II for the values of the constants. The check-box toggles between ascending and descending order. This feature allows the user to search for structural patterns within the context of the sequence information from the alignment. See the "Examples" page of the website profilegrid.org for illustrations of this feature.

Table II. Amino acid physical constants. See References for abbreviations and modifications (*). Many more constants are available for those coding their own programs (Kawashima, 2000).

Code1	Code3	Age	Flex	Freq	HPA	HPO	Helix	MutD	MutG	S.A.	Vol
A	ala	2	6*	9.5	1.8	-0.4	-0.27	100	0.75	115	67
C	cys	13	9.5	1.2	2.5	0.17	0.64	20	0.31	135	86
D	asp	3	8	5.1	-3.5	-1.31	0.52	106	0.93	150	91
E	glu	7	13	5.8	-3.5	-0.91	0.21	93	0.82	190	114
F	phe	14	6.4	3.9	2.8	1.92	0.73	41	0.83	210	135
G	gly	1	20*	7.3	-0.4	-0.67	1.7	49	0.66	75	48
H	his	11	7.1	2.3	-3.2	-0.64	0.57	66	0.53	195	118
I	ile	10	8.7	6	4.5	1.25	0.44	96	0.89	175	124
K	lys	12	14.7	4.4	-3.9	-0.67	0.019	56	0.76	200	135
L	leu	8	19.6	10.7	3.8	1.22	0.095	40	0.92	170	124
M	met	16	17.7	2.8	1.9	1.02	0.25	94	0.58	185	124
N	asn	10	8.7	3.9	-3.5	-0.92	0.69	134	0.87	160	96
P	pro	5	6*	4.4	-1.6	-0.49	3.8	56	0.67	145	90
Q	gln	10	10.7	4.4	-3.5	-1.22	0.28	102	0.79	180	109
R	arg	9	13.9	5.5	-4.5	-0.59	-0.052	65	0.68	225	148
S	ser	6	16.1	5.8	-0.8	-0.55	0.52	120	0.76	115	73
T	thr	8	9.3	5.4	-0.7	-0.28	0.95	97	0.63	140	93
V	val	4	10	7.1	4.2	0.91	0.77	74	0.86	155	105
W	trp	17	6.7	1.5	-0.9	0.5	0.69	18	0.58	255	163
Y	tyr	15	6.2	2.9	-1.3	1.67	0.42	41	0.64	230	141

Frequency Colors. Every ProfileGrid cell is fill-in colored according to the frequency (count) value computed for that residue at that column position in the alignment. The check-box toggles on or off the current coloring scheme. The "Frequency Colors" button will open a window listing the default 6 frequency/color bins (Figure 3). A cell is colored by the bin that has the largest threshold value greater than or equal to that cell's residue frequency: <10% (white), ≥10% (gray), ≥25% (yellow), ≥50% (orange), ≥70% (green), and ≥90% (red). This color scheme was chosen to maximize the visual *differences* between bins during the inspection of ProfileGrids for patterns. By contrast, a color ramp (*i.e.,* shades of one color) would not facilitate such analysis. However, the user is able to define their own frequency color scheme by choosing the number of bins and the desired colors (Figure 3).

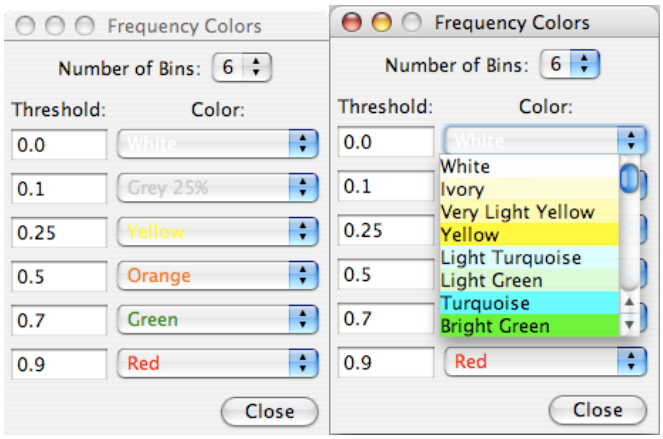


Figure 3. Frequency Colors window with default settings

Show Values. This menu lets the user choose what is displayed within each amino acid cell. The default option is a simple integer count of each residue per alignment column position. The "Percentage" option is the integer count divided by the total number of sequences. The check-box toggles on or off the display of the values. However, the user can still identify the currently selected cell's integer count from the top-left corner of the ProfileGrid window, *i.e.* the "cell value identification" feature (see Figure 4).

Highlight Sequence. This option allows one to detect and to represent unique features of one sequence with respect to a multiple sequence alignment (MSA). Such a feature may indicate specialization with respect to function or activity. The feature is not active by default because the first sequence in the MSA is chosen by default as the highlight sequence. Since this is also the same default selection as the template sequence, then no cells in the ProfileGrid are highlighted. By contrast, when the menu is used to select a sequence *different* from the template sequence, then the highlight feature is turned on. Specifically, the

highlight sequence will appear immediately below the template sequence in the ProfileGrid. Furthermore, a pairwise comparison is made such that the corresponding residue is boxed if the highlight sequence *differs* from the template sequence (Figure 4). The default highlight color is "Ocean Blue" although the user may choose other colors. The ProfileGrid cell value identification will also reflect the current cell's position with respect to the Highlight sequence.

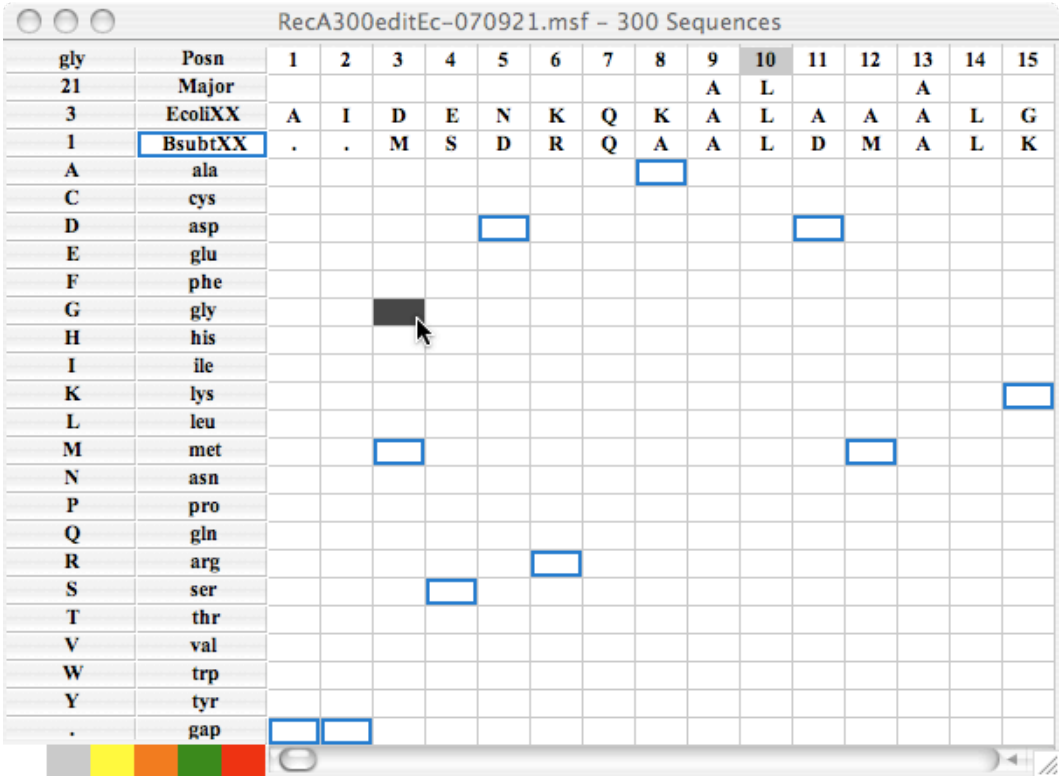


Figure 4. Highlight Sequence example of the *Bacillus subtilis* RecA homolog. ProfileGrid frequency colors and values are turned off for the sake of clarity. However, the cell value identification feature (top left-hand corner) still allows one to know the frequency value for the currently selected cell (dark grey).

Identify sequences... When a particular cell of the ProfileGrid is selected with the mouse and the "Sequences" button is activated, a new window lists the sequences that contain the indicated residue, *i.e.,* selected residue is "Present" (Figure 5). By contrast, sequences where the selected residue is "Absent" can be shown by selecting the corresponding menu item. This selection is useful when the user wants to know the list of sequences from all cells *except* the selected one from within the same column.

Figure 5. Sequences window

Activating the "Show Alignment" button opens a new window showing a MSA of the template sequence and the identified alignment members from the sequence window (Figure 6).

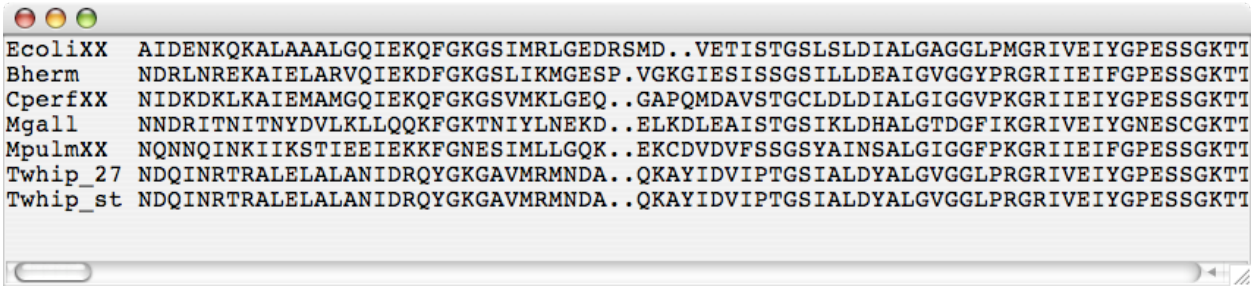


Figure 6. Alignment window

Similarity Parameters

JProfileGrid makes similarity plot calculations based on the `plotcon` algorithm (Rice, *et al.*, 2000) with the modification that the values are normalized between 0 and 1. A sliding window is used to calculate a pair-wise conservation at each residue position using the BLOSUM62 (for proteins) or EDNAFULL (for nucleic acids) scoring matrices. Weights for each sequence are taken from the MSF files to correct for over-represented sequences. However, similarity plot calculations based upon FASTA files may be biased toward over-represented sequences. The calculation results are the positional similarity values centered at the midpoint of the sliding window. All values are saved in a separate worksheet in the spreadsheet file. Note that a *graph* of the similarity values is *not* displayed by JProfileGrid. The only graphical display of the similarity calculations is via the "Similarity Boxes" outlined in black in the ProfileGrid. A threshold value determines the endpoints of these boxes.

This is a *slow* calculation since all sequence pair-wise comparisons are made. JProfileGrid supports CPU threading so that the similarity plot calculation will occur in the background while allowing one to continue examining the ProfileGrid window. Use multi-processor computers to improve performance.

Enable Similarity Parameters. Since the calculation can takes so long, the user is given the option of using this feature. After enabled, the calculation will be performed after the "Compute" button is clicked.

Window Size. This parameter specifies the length of the sliding window (default 5). Due to a programming limitation, we suggest only using odd integers, *i.e.*, 1, 3, 5, *etc*. A rule-of-thumb for proteins is to use half the average length of the secondary structural elements, if known.

Threshold. This cutoff value determines the endpoints of the Similarity Boxes in the ProfileGrid where similarity values are equal to or greater than the threshold value (default 0.7). Note that sometimes separate Similarity Boxes appear to be immediately adjacent to one another due to close peaks from the similarity plot. Final touchups of the box endpoints can be made in the exported spreadsheet file by manually editing the cell borders. The threshold value also determines the contents of the majority consensus sequence at the top of the ProfileGrid below the position ruler.

Known issue. Sometimes the similarity calculations are not normalized properly. Then the threshold value may be too large resulting in no similarity boxes being drawn. Check the spreadsheet output to examine the similarity values. The problem may be due to different alignment programs calculating sequence weight values that vary by orders of magnitude. We are working on a software fix.

Export Parameters

JProfileGrid saves the ProfileGrid value and graphical output. The file is written to the indicated folder after the "Export" button is clicked.

Range. Indicate which part of the ProfileGrid to export. Note that these position values are absolute (independent of the template sequence) beginning at the first column of the MSA.

Columns Per Tier. Specify how many residues to appear in each tier of the ProfileGrid. The maximum columns per tier is 253 (accounting for the residue name and symbol columns) since Excel has an internal limit of 256 columns per worksheet.

Format. The "Excel" selection saves the ProfileGrid as an Excel Workbook spreadsheet file (.xls) using the JExcel API (http://jexcelapi.sourceforge.net). The ProfileGrid matrix and similarity values are stored in separate worksheets. The ProfileGrid worksheet is labeled with the sort order. If the similarity values were calculated, then they are saved in a worksheet labeled "PlotSim". A third worksheet labeled "Flags" lists MSA outlier characters (such as "X") that JProfileGrid flags for verification. The character, sequence name, and MSA position(s) of the outlier character(s) are listed. See "Input File formats" above for a list of characters which will *not* be flagged.

The "PyMOL Script" selection writes a script file (.pml) for the PyMOL 3D molecular visualization program (http://pymol.sourceforge.net). Residue position numbering is determined by the template sequence chosen and the value of the position row start value. In the script file, the user must manually change the "filename.pdb" text to the name of the actual PDB file to be read in by PyMOL.

Figure 7 shows an example of the JProfileGrid output mapped on to the *E. coli* RecA protein crystal structure (Story, *et al.*, 1992). Residues that are completely conserved (identical) in the MSA are saved as a PyMOL selection named "ident" in the script file. These residues were easily chosen by clicking on the "ident" selection in the PyMOL interface (Fig. 7, right side); and, in this figure the sidechains were rendered as black spheres. Residues that pass the highest threshold value in conservation (default bin of ≥90%) are saved as a selection named "bin90". This selection is shown in the PyMOL interface, but the residues themselves are not highlighted in the figure.

The similarity plot calculations and the threshold value determine the endpoints of the motifs and variable regions linking motifs. In Fig. 7, a window size of 9 and a threshold value of 80% similarity was used. In the script file, the motifs as PyMOL selections are listed starting from the N-terminus and labeled nu-

merically starting with "mot0". JProfileGrid calculates a linear ramp from red to green to color the motifs from N-terminus to C-terminus, respectively. The non-conserved variable regions are named selections starting with "var0" and are all colored gray.

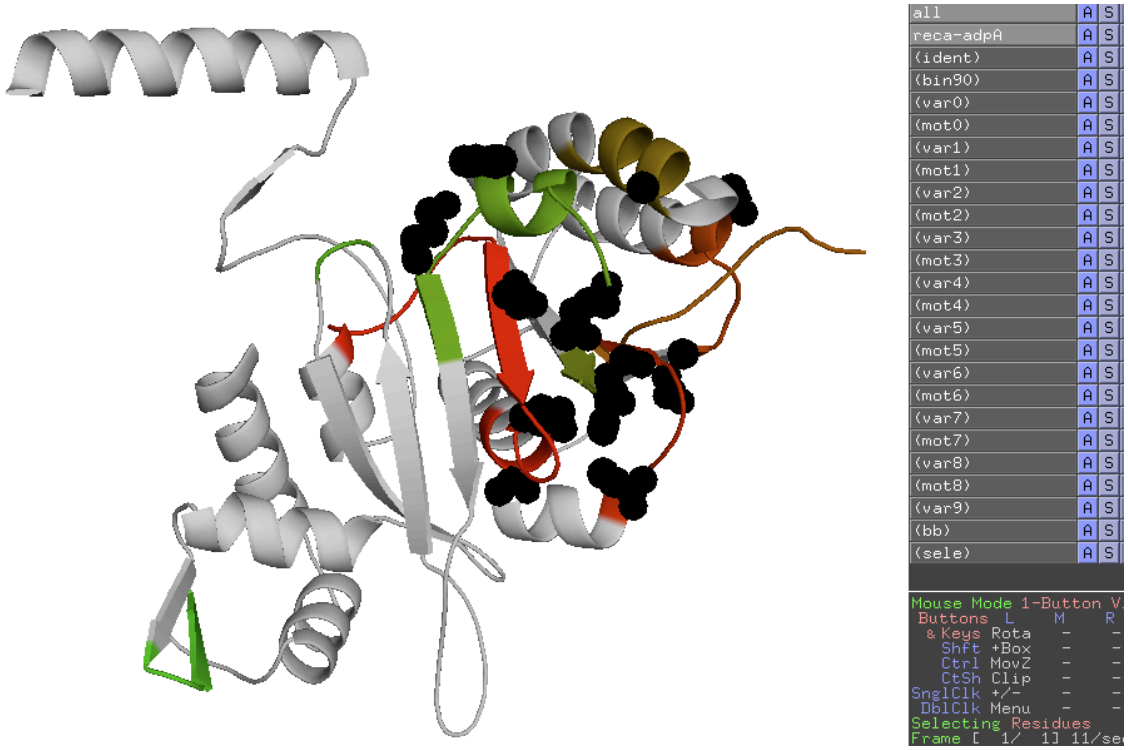


Figure 7. Visualization of JProfileGrid PyMOL script output on the RecA protein crystal structure. See text for explanation.

Menus

File. Quit JProfileGrid.

Edit. Copy sequence list. Only available when the Sequences window is active.

Window. Select among the various windows of the program.

About. Show the version and license information.

JProfilegrid Version history

ver. 1.2.2

In correcting command-line version (ver. 1.2.1), we inadvertantly introduced a bug that broke GUI version on large MSAs. This has now been fixed. Also, corrected a calculation error where non-standard amino acid code characers (B, J, O, M, X) where counted as gaps in the ProfileGrid.

ver. 1.2.1

Fixed bug that prevented command-line version from working after threading feature was implemented in ver. 1.1.

ver. 1.2

Added following features: "Window" menu, "highlight" selection to PyMOL export, reorganized parameters panel, ProfileGrid window cell identification values placement rearranged.

Fixed bugs in Excel export when Highlight sequence turned on.

ver. 1.1.1

Fixed bugs in cell value identification feature.

ver. 1.1

Added following features: threading, cell value identification, and PyMOL script export.

ver. 1.0

First version of JProfileGrid.

www.profilegrid.org website went live on December 28, 2007.

REFERENCES

Kawashima, S. and Kanehisa, M. (2000) AAindex: amino acid index database, *Nucleic Acids Research* **28:** 374.

Rice, P., Longden, I., and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite, *Trends in Genetics* **16:** 276-277.

Roca, A.I., Alamda, A.E., and Abajian, A.C., (2008) ProfileGrids as a new visual representation of large multiple sequence alignments: a case study of the RecA protein family, *BMC Bioinformatics* **9:** 554.

Roca, A.I. and Cox, M.M. (1997) RecA protein: structure, function, and role in recombinational DNA repair, *Progress in Nucleic Acid Research and Molecular Biology* **56:** 129-223.

Story, R.M., Weber, I.T. and Steitz, T.A. (1992) The structure of the *E. coli* RecA protein monomer and polymer, *Nature*, **355:** 318-325.

Age: Trifonov, E.N. (2004) The triplet code from first principles, *J Biomol Struct Dyn* **22:** 1-11.

Flex = buried sidechain flexibility: Zhao, S., Goodsell, D.S. and Olson, A.J. (2001) Analysis of a data set of paired uncomplexed protein structures: new metrics for side-chain flexibility and model evaluation, *Proteins: Structure, Function, and Genetics* **43:** 271-279. No values given for alanine, glycine, and proline so estimated for purposes of sorting.

Freq = Frequency in *E. coli* K12 genome: Nakamura, Y., Gojobori, T. and Ikemura, T. (2000) Codon usage tabulated from international DNA sequence databases: status for the year 2000, *Nucleic Acids Research* **28:** 292.

HPA = Hydropathy: Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein, *Journal of Molecular Biology* **157:** 105-132.

HPO = Hydrophobicity: Sweet, R.M. and Eisenberg, D. (1983) Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure, *Journal of Molecular Biology* **171:** 479-488.

Helix: Rohl, C.A., Chakrabartty, A. and Baldwin, R.L. (1996) Helix propagation and N-cap propensities of the amino acids measured in alanine-based peptides in 40 volume percent trifluoroethanol, *Protein Science* **5:** 2623-2637.

MutD = Mutability-Dayhoff: Schwartz, R.M. and Dayhoff, M.O. (1978) Matrices for detecting distant relationships. In: Dayhoff, M.O. (ed), Atlas of Protein Sequence & Structure. Natl. Biomed. Res. Found., Washington, D. C., 353-358.

MutG = Mutability-Grantham: Grantham, R. (1974) Amino acid difference formula to help explain protein evolution, *Science* **185:** 862-864.

S.A. = Surface Area: Chothia, C. (1976) The nature of the accessible and buried surfaces in proteins, *Journal of Molecular Biology* **105:** 1-12.

Vol = Volume: Richards, F.M. (1974) The interpretation of protein structures: total volume, group volume distributions and packing density, *Journal of Molecular Biology* **82:** 1-14.