ProFeatMap
Protein Feature Map

User Guide

Description: ProFeatMap is online web interface creating highly customizable 2D representations of protein lists based on Uniprot data.

Written by: BICH Goran

Contact information: profeatmap@gmail.com

Last update: March 3, 2022

ProFeatMap version: 1.0.0

Summary: ProFeatMap creates 2D representations (maps) of elements of interest (features) for a list of proteins based on the information available in the Uniprot database following several steps:

Step 1: The user has to input a list of proteins as Uniprot accession codes that ProFeatMap will use to download from the Uniprot database.

Step 2: Extraction of the features in the downloaded files, compiling data in a single output file.

Step 3: An optional step to add numerical values that can be shown with colormaps in Step 4.

Step 4: Creation of the map itself.

Table of content

I. General comments

I.1. Description

Feature

A feature is an element of interest in a protein such as These features can be domains, repeats, post-translational modifications, variants, secondary structure,… These features appears in the "FT" category in Uniprot files.

Map

In ProFeatMap, a map is a schematic 2D representation of proteins in which features are represented according to their relative position and size on a given protein.

I.2. Components

I.2.a. Drag and Drop or Select Files

Description

ProFeatMap allow the user to upload files in several formats. You can do this either by drag and dropping your file within the dotted field, or click on it and select your file.

Compatible file formats

"***.xlsx***": 2007 and later Excel version file format.

"***.xls***": Excel file format before 2007.

"***.ods***": LibreOffice Calc file format.

"***.csv"***, *"****.tsv"***, *"****.tab"***, *"****.txt***": These file format are also recognized by ProFeatMap. If tabulations are detected in these files, ProFeatMap will consider that you are using them as separators. If not tabulations are detected, it will search for ";" in your file. If some are found it will consider you are using a format, with ";" separator and "," as decimal separator and replace "," by ".". If these two cases are not true, it will use "," as separator.

Warnings

Numerical values in columns should not be formulas and decimal separator should be ".".

I.2.b. Tables

Description

Tables are the way to give ProFeatMap information needed to process the different steps to make your maps. They are a total of 5 tables in the interface:

"***Protein list***": (*mandatory*) This table contains the list of proteins that you want to appear on the map.

"***Modifications***": (*optional*) This table contains modifications you want to make during the extraction. This table is used to "correct" data or to manage more precisely the display of your proteins.

"***Numerical values***": (*optional*) This table contains the numerical values you want to give to specific features of your proteins.

"***Shapes and colors***": (*mandatory*) This table contains the display parameters of each feature for fine tuning of the map creation. The VI.2.c. Automatic feature selection  option fills up the table automatically and could be used as base.

"***Cut regions***": (*optional*) This table contains the proteins you want to appear shorter on the map.

File name

The name of the file is up to the user, only the content is used by ProFeatMap.

Mandatory columns

Each table has a number of columns that are recognized by ProFeatMap. Therefore, it is important to make sure all these columns appear in your tables. Warning: The names of the columns are case sensitive.

Additional columns

When building these tables in your table editor, you can add additional columns that will not be used by ProFeatMap. You must not use the same column names as the mandatory ones. You should avoid spaces and special characters in the names. Avoid replicating column names.

Buttons

 Adds a row at the end of the table. Empty rows should be avoided if possible. The new row has to be saved to be taken into account.

 This button is used to clear completely the table, leaving only the template columns and an empty row. Clearing a table needs to be saved to be taken into account.

 Is used to save changes made in the table. Non saved modifications will not be taken into account when you execute the step. If you haven't saved your modifications yet and want to go back to the previous version, you can reload the page. This button will automatically light up () when modifications are made and switched off when clicked.

 This button will download the latest saved table as an .xlsx file. A template file will be provided, with headers only, if no table is displayed.

I.2.c. Map parameters buttons

Description

This section only applies the three buttons under the map parameters, which behavior is a different from the buttons below the tables.

Buttons

 Will reset all map parameters with the default values.

 This button saves the current selected map parameters. Once you close ProFeatMap and reopen it, these parameters will show the latest saved state. Saving is not needed to be taken into account in Step 4. It allows you to play with the parameters and only save them once you are happy with the result.

 Clicking this button will download currently saved map parameters as an .xlsx file.

I.2.d. Downloadable files

Description

At multiple steps, ProFeatMap will propose to download files. Downloading these files will allow the user to save locally their work.

Formats

"***.xlsx***": Table, extracted data and map parameters can be downloaded as Excel files.

"***.fa***": In Step 2, you can extract sequences of a feature. The result can be downloaded as a .fa file compatible with multiple alignment tools.

"***.png***": The maps and legend generated by ProFeatMap can be downloaded by right clicking and saving images. The file format of these images is .png.

II. Quick run guide

II.1. Description

The quick run guide shows how to obtain a map following the 4 main steps of ProFeatMap. It is intended to use to have more control on your map creation and customization.

II.2. Map only

May the user only want a quick look at the protein map, he can instead click on the "Map only" button in Step 1 after inputting his protein list. The map will appear below Step 1 and be ready to download. "VI.2.f. General feature parameters", "VI.2.g. Feature parameters" and "VI.2.h. Order of feature drawing" will affect the resulting map.

II.3. Steps

1. Upload a protein list in Step 1 section in the Drag and Drop zone. (A protein "List example" file can be downloaded if needed)
2. Click on the Normal run button and wait until finished.
3. Go to Step 2 section and click the Run button.
4. Skip Step 3 and go directly to Step 4. Under the Shapes and colors table, click on the VI.2.c. Automatic feature selection and save the changes.
5. Click the Run button in Step 4. The map will be shown once generated.

III. Step 1: Protein data gathering

III.1. Description

In this first step the user has to give ProFeatMap a list of protein he wants to see as a map. ProFeatMap will download all needed Uniprot data files for these proteins by running this step.

III.2. Components

III.2.a. Protein list table

Description

Table used to give ProFeatMap your list of protein of interest. This list contains a Uniprot accession code and a unique name for the protein.

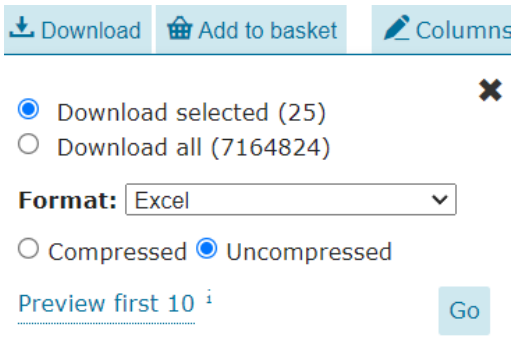
Table construction

"***code***" or "***Entry***": Column where you indicate the Uniprot accession code corresponding to the protein. The same code can be used multiple times.

"***protein***" or "***Entry name***" (optional): Column containing the name of your protein. This is the name that will be used during all the process and has to be unique and will serve as identifier to ProFeatMap. If no names are given, ProFeatMap will search for Uniprot names after downloading files.

Uniprot search compatibility

When searching your proteins of interest on the Uniprot webpage, you can export your selection or basket as an Excel file. The uncompressed version is compatible with ProFeatMap.



III.2.b. Remove organism

Description

This option removes the organism tag from your list of proteins (i.e. TRIO_HUMAN will be transformed in TRIO). It only works if all your protein names have the same tag. This feature is meant to be used on proteins names from Uniprot, when you are only interested in one organism.

Downloading troubleshooting

If the protein list contains obsolete or invalid Uniprot codes, ProFeatMap will make a section appear in which you will be able to download a file containing all codes that ProFeatMap did not manage to download properly.

"***obsolete***": The first tab contains all obsolete codes. These codes that are not used by Uniprot anymore but still recognized. By searching these codes on Uniprot you might find the new associated code.

"***unrecognized***": The second tab contains all codes that ProFeatMap could not download. The codes might contain a typo. If your whole list shows up in this section, it is most likely a format issue in your input file (see here for more information). It could also mean that the connection to Uniprot database is currently not possible. To check this, reload the page. A cross on red background will appear next to the "Uniprot website" if the connection is not possible.

IV. Step 2: Feature extraction

IV.1. Description

The main goal of this step is to extract features found in the Uniprot data files ProFeatMap gathered in Step 1. Additionally you can provide a modification table to add or remove feature during the extraction process.

This step also contains additional tools: feature sequence extraction and feature/motif search by regular expression.

IV.2. Components

IV.2.a. Modification file (optional)

Description

Table filled with features to add or remove during the extraction process.

Table construction

"***ex_type***": Column where you indicated if you want to add ("add" or "+") or remove ("remove" or "-") a feature.

"***protein***": Name of the protein you want to modify.

"***FT_type***": Column to indicate the type of feature (DOMAIN, REPEAT, HELIX, BINDING, ZN_FING...). When removing a feature, you should check for the corresponding type in the output extracted data file. When adding a feature that already exists, you should consider using the same FT_type. If you want to add a whole new feature, DOMAIN should fill you needs. Using other FT_type may be used for specific uses such as drawing order of features in the final map.

"***name***": Column to indicate the name of the feature. This name will appear on the legend.

"***start***": (only for adding features) Starting position of the feature you want to add.

"***length***": (only for adding features) Length of the feature you want to add.

IV.2.b. Structural coverage extraction

Description

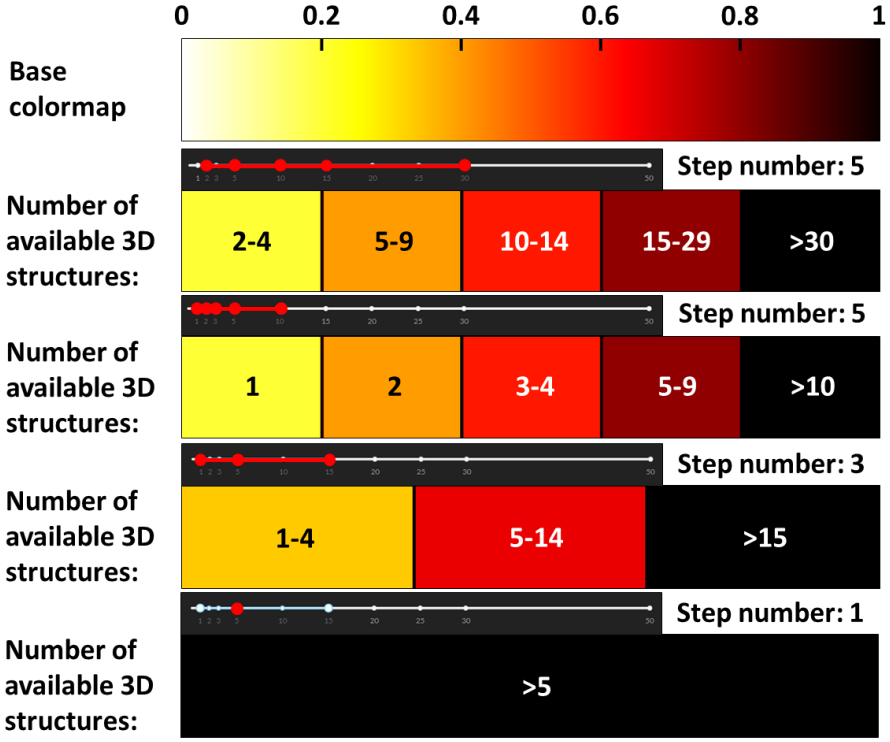
ProFeatMap searches for resolved 3D structures and calculate the number of structure available for each residue in each protein. This "coverage" can then be shown in Step 4.

Parameters

"***Colormap thresholds***": (*default*: [1, 2, 3, 5, 10]) Number of structures needed for each color in the colormap. See example below. The color of the first step is the color at 0.2 on the base colormap (and not 0!). If there is only on step, the color will be the color at 1 on the base colormap. Number of structures below the first number in the array (1 in the default), will not appear on the final map.

"***Step number***": (*default*: 5) Number of section the colormap will be cut into. See example below.

Examples



IV.2.c. Feature sequence extraction

Description

Optional tool of ProFeatMap used to get the sequences of all occurrences of a given feature in your whole protein list. This tool will only appear once Step 2 has been run once.

Parameters

"***Feature name***": The name of the feature you want to extract sequences. The name should correspond at the name found in the extracted data file. The name is case sensitive. With the "N-ter" and "C-ter" parameter set to 0, ProFeatMap will extract the sequences as defined in Uniprot.

"***N-ter ext***": (*default*: 0) Positive values will extend the sequence by the value towards the N-terminus of the protein. Negative values will shorten the sequence towards the C-terminus.

"***C-ter ext***": (*default*: 0) Positive values will extend the sequence by the value towards the C-terminus of the protein. Negative values will shorten the sequence towards the N-terminus.

IV.2.d. Feature/Motif search by regular expression

Description

Optional tool which can be used to find the localization of features or motifs based on a regular expression. The output file can be used as modification file. After rerunning Step 2 a second time, you will be able to use the feature sequence extraction if you want the sequence of the features corresponding to the regular expression. This tool will only appear once Step 2 has been run once.

Parameters

"***Feature regular expression***": To make the search you need to build a regular expression of the feature or motif you want. More info on how to build a compatible regular expression here: https://docs.python.org/3/library/re.html

"***Feature name***": The name you want to give to the extracted feature/motif. This is the name that will appear on the final figure if you add the output to your modification table.

IV.2.e. Direct download

Description

If activated, this option will directly propose you to download the extracted data instead of saving them first in your browser when Step 2 ends. This will cause ProFeatMap to not be able to go further in the steps anymore. This option should only be used in the case you want to run data extraction on a very large list of proteins (>1000), which cannot fit in your browsers memory. If the download button does not return you anything after successful ending of Step 2 :

Try to empty your browsers navigation data for ProFeatMap (it will cause everything to be removed, so save your work before making it). If the result is the same after this step, you can either split your protein list in smaller sets or use this option if you're only interested in the content of your protein list.

V. Step 3: Numerical values addition

V.1. Description

This section can be used to add numerical values to features that can then be displayed on the map.

V.2. Components

V.2.a. Numerical values table

Description

A table containing a list of features and the numerical values that should be associated.

Table construction

"***protein***": name of the protein that contains the feature you have one or more values for.

"***feature***": name of the feature that you have one or more values for. This name has to correspond to the exact name of the feature that appears in the extracted data file.

"***start***": the starting position of the feature. This value has to appear in the extracted data file. It is used to know which feature (if multiple) should have a value

"***condition_x***": multiple conditions can be added where you have values for. One column per condition is needed and must have a unique user defined name. Spaces should be avoided. These conditions can then be selected in Step 4 for the map creation.

Values

The values given by the user should be normalized (between 0 and 1). -1 will be interpreted as missing value and will appear in grey on the map. No value will result in the use of the specified "color" and "contour_color".

VI. Step 4: Map creation

VI.1. Description

This step will create the map of your protein list. It will use the shapes and colors defined automatically or by the user and the optional cut regions. Map parameters are used to change the overall map generation, the display of specific features, the order of feature drawing and the condition (if any) to show numerical values.

VI.2. Components

VI.2.a. Shapes and colors table

Description

The shapes and colors table contains the list of features you want to appear as specific shapes and colors on the protein map.

Table construction

"***shape***": see available in Shapes and colors

"***orientation***": some shapes have associated orientation you need to specify

"***height***": (*default*: vertical stretch factor) corresponds to the height, in pixels of the feature.

"***contour_color***": (*default*: black) if no other color is specified

"***contour_colormap***": (*default*: none) Used if you want to show numerical values on the contour of the shape. In the Uniform shape fill/contour is selected, a "colormap" is defined and no "contour_colormap" is selected, the contour will share the "colormap" with the inside of the shape.

"***contour_threshold***": (*default*: none) Only used if a "contour_colormap" is defined. A value between 0 and 1 can be indicated. Values below this threshold will appear in the color defined in "contour_color". Values above the threshold will use the "contour_colormap".

"***color***": (*default*: white) color used to fill the shape.

"***colormap***": (*default*: none)

"***threshold***": (*default*: none) Only used if a colormap is defined. A value between 0 and 1 can be indicated. Values below this threshold will appear in the color defined in "color". Values above the threshold will use the "colormap".

"***pensize***": (*default*: protein thickness) corresponds to the thickness of the contour of the feature.

Showing other features

By adding “Other” as feature name, you will also show all features that are not listed in the table. (Classically by a rectangle with black “contour_color” and white “color”, but can of course be changed)

VI.2.b. Protein cuts table

Description

If you want specific proteins to be shorter, you can indicate it in this table. This will hide a fraction of the protein (indicated by a “//” on the map) and all features inside.

Table construction

“***protein***”: Name of the protein you want to cut out a region.

“***start***”: Starting position of the region to hide.

“***length***”: Length of the region to hide.

VI.2.c. Automatic feature selection 

Description

Automatic feature selection will search in the extracted the features either the most represented features or features that appear more than a certain number of times in your list. Only features under the DOMAIN, REPEAT, MOTIF, REGION tag are considered. If the feature is one of the most common ones (DOMAIN and REPEAT in the human proteome), the default shape and color will be used. Other features will have a random shape and color. All features that are not represented enough will be represented by the Others category. By default “more than 2 occurrences” is used.

Parameters

“***x most represented features***”: (*default*: √number_of_proteins) automatic feature selection will choose a shape and color for the top x most represented features

“***more than x feature occurrences***”: (*default*: 2) automatic feature selection will choose a shape and color for features that have been found at least x times in your protein list

Lock seed

This option can be toggled to fix the current seed used for random picking of shapes and colors by the automatic feature selection.

VI.2.d. Sorting

Description

ProFeatMap features several sorting option that will impact the order of appearance of the proteins in the list.

Options

“***None***”: order of the proteins as defined in the table in Step 1.

“***abc***”: to sort your protein in alphabetical order.

“***feature_number_distance***”: (*default option*) to let ProFeatMap sort automatically your proteins, by regrouping proteins that share similar feature content. WARNING : This sorting is quiet intensive, please do not try on very large lists of proteins (>2,000). Once you have done this sorting and the map is drawn, you can save the resulting sorting with the Latest sorting protein list button. You can use this list as input list, and put None as sorting. This will increase dramatically the speed of the drawing. (Advised for protein list longer than 200 proteins)

"***value***": see VI.2.e. Value related below

VI.2.e. Value related

Description

Sorting by value is only available if the user has given numerical values in Step 3. This sorting will order proteins by descending values. Each protein is represented by the highest value if multiple occurrences of the target feature is found.

Parameters

"***Case to draw***": (*default*: None) all conditions defined in the numerical value file will appear in this list when Step 3 is run. Selecting a condition will affect the values displayed on the map.

"***Focus on***": (*default*: None) the name of the feature by which the map should be sorted must be inputted here.

"***Threshold***": (*default*: None) by indicating a threshold (float value), you can remove all proteins that have not at least one of the focused on feature with a value above this threshold.

VI.2.f. General feature parameters

Description

General feature parameters will affect general visual aspects of the map such as horizontal and vertical stretch, protein thickness, text sizes and the display of the protein length.

Parameters

"***Horizontal stretch factor***": (*default*: 1) multiplying factor of the horizontal length on the map. > 1 will cause the proteins and features to appear longer, whereas < 1 will make them appear smaller. Support floats.

"***Vertical height***": (*default*: 20) vertical space used by each protein. Increasing the value will make the figure bigger. It will also cause all features shapes that have no "height" specified to have this value as default value.

"***Protein thickness***": (*default*: 3) thickness of the line representing the protein. This value is the default pensize value for feature shapes (contours), if none is specified.

"***Protein name size***": (*default*: 30) size of the text showing the protein names.

"***Biased regions text size***": (*default*: 20) size of the text showing above the composition biased regions.

"***Show protein length***": (*default*: y) toggle that enables or deactivates the display of the protein length next to each protein.

"***Consistent shape fill/contour***": (*default*: n) toggle that enables or deactivates the standardization of the filling color and the contour color. When activated, all feature shapes that have no "contour_color" specified, will appear the same color as the filling color (if there is one). It also applies to colormaps.

VI.2.g. Feature parameters

Description

These parameters can be toggled on or off to make specific features appear or disappear from the created map.

Parameters

"***3D structure coverage***": (*default*: n) As defined during extraction (see Structural coverage extraction), toggling this option will show a line where resolved 3D structures have been found in the PDB database. The line is colored depending on the number of structures.

"***Secondary structure***": (*default*: n) Secondary structures (helixes, strands and turn) will be shown on the proteins.

"***Disorder***": (*default*: y) When toggled, predicted disordered regions will be shown on proteins.

"***Modified residues***": (*default*: n) When toggled, modified residues (phosphorylation, glycosylation,…) will appear on proteins.

"***Composition biased regions***": (*default*: None) Selected biases can be selected, and will appear on the proteins.

Feature parameters default

Clicking this button will add to the current shapes and colors table the default representation of each selected "Feature parameter" in the "Map parameters". The adding of these representations will overwrite the default representation. This will also cause these parameters to be always displayed whenever the corresponding parameter is toggled or not.

VI.2.h. Order of feature drawing

Description

This list of feature categories defines the order in which the features will be drawn on the proteins. In case of overlapping features, the category appearing the latest in the list will appear above. Default parameters should avoid most overlaps.

Parameters

(*default*: DISORDER, DOMAIN, CHAIN, INIT_MET, ZN_FING, DNA_BIND, REGION, ACT_SITE, METAL, SITE, LIPID, HELIX, STRAND, TURN, CONFLICT, CARBOHYD, BINDING, MOTIF, MOD_RES, COMPBIAS, REPEAT, VARIANT, PDB)