COPAL Tool -- User Guide

Input file(s) details

The file(s):

- should contain slices in separate columns, and proteins in rows.
- should contain a row of unique headers for each column
- should contain at least one column with identifiers for the proteins
- can contain any number of extra columns or sheets
- can contain multiple samples in one sheet
- cannot contain information in rows below the protein migration data
- In the case of multiple files/sheets containing samples, protein identifiers should match, so the datasets can be combined.

Input instructions

Input frames

filename:

- enter filename with file path, or select a file using the choose file button.
- Files can be either excel or csv files. Select the appropriate file format from the dropdown menu.
- In the case of an excel file: enter the name of the sheet containing the data in the sheetname entry. When using a text file, this entry will be ignored.

File details:

- Skip rows: in the case of more than one header row, skip all but one.
  (ie: 3 header rows → use skip rows: 2). The 3rd header row would in this case be used for the analysis, and should match the header names specified below
- Protein identifier column: enter header of column containing protein identifiers.

sample names and columns:

- In the first box, enter names for samples contained in this file/sheet
- In the first and last column boxes, enter first and last column header for the corresponding samples on the same line.
- Take care not to leave any unnecessary spaces, tabs or empty lines in these text boxes

Add another file:

- Use this option If samples from other sheets or files are to be added
- A new window will pop up

Proceed to output:

- If all input files and samples are specified, use the proceed to output button

Output frame

Job name: Enter preferred name of this analysis job, output folder and files will have this name

Output folder: Select folder location for output files

Data normalization:

- select type of normalization
- None: no normalization will be performed (only recommended if data has been normalized already)
- Using all Proteins: normalization will be based on complete data
- From column: A column in the input file with True/False specifies which proteins will be used for normalization. Enter column header in ‘if from column’ entry box
- From file: a separate plain text file containing a list of proteins is used for normalization. Enter file name and path in ‘if from file’ entry box (or use Choose file button). The file should contain one protein identifier on each line (that correspond with specified identifier column in input).

Alignment:

- Check the box if samples are to be aligned
  - **Note:** if not selected, score analysis cannot be performed! It is recommended to align samples if score analysis is to be performed. For now, the GUI version of COPAL does not support scoring without alignment.
- Select method of warping the data. The interpolate option will fill gaps by interpolating based on values adjacent to the gap(s). The repeat option will fill gaps in the alignment by repeating the abundance value from the previous fraction. Interpolation is recommended for normal use cases.

score analysis:

- check the box if hausdorff scores should be determined
- if checked, enter sample names in group 1 and 2 boxes, one sample name per line.
- Hausdorff scores will be determined between these groups
- Take care to not leave any unnecessary enters spaces or tabs in the entry boxes
- The hausdorff factor determines the protein abundance/molecular mass axes ratio of the plane in which hausdorff distances are calculated. Taking the default value of one will consider both dimensions equally. A higher value will weigh differences in protein abundance more heavily, where a value smaller than 1 will weigh shifts in molecular mass more heavily.

Provide rank ordered protein list:

- This option can provide a ranked list of proteins with combined hausdorff scores. This list is in the .rnk format, which can be readily used in Gene Set Enrichment Analysis.
- If this option is checked, enter the header of the column containing protein identifiers to be used in rank ordered file in the entry box (for example, the “gene symbols” column if this is what is required for further analysis).

Back to start:

goes back to first frame, all data will have to be entered again. To be used in when a second dataset will be aligned, or when the input was wrong.

Save and run:

If all input is entered correctly, pressing save and run will start the complexome alignment process.

Status bar:

- Text box at the bottom will indicate the status of the analysis.
- If the analysis is complete, or if an error occurs, it will be displayed here.

COPAL Output

Copal provides an output folder, containing several files depending on the parameters of your analysis run. This section will discuss the content of the various output files in detail.

Align info file

This is the file ending with '_align_info.txt', and will always be present when performing a COPAL analysis. It is a plain text file containing detailed information about the various analysis steps performed. It contains the following information (in order):

- The source file(s), a list of the files used as input for this analysis job. It contains the full path (location) of each file.
- Number of proteins in input datasets
- In the case of multiple input files: number of proteins in matched dataset
- If normalization is performed: The factor by which all protein abundances in each sample are multiplied to normalize the abundances between samples.

The following information is present if alignment is performed in this analysis job:

- A list with the number of fractions/slices of each sample before alignment
- The number of fractions/slices after alignment of all samples
- A list with the assigned sample number and the corresponding sample names provided in the COPAL input frame
- The global dtw distance/cost between all sample combinations, sorted from lowest cost to highest. The dtw global cost or distance is a measure of difference between samples. Samples with a low dtw distance are more similar to each other
- A list of all sample pairs, ordered by global dtw cost in an ascending order
- A list with the order of progressive alignment as performed by COPAL, based on the dtw distance between samples. Structure of each step in the list: (left sample(s), right sample(s), dtw distance of this alignment step)
- A visualization of the final alignment of the fractions/slices of all samples. X indicates the location of a 'gap' introduced between fractions at this location.
- If hausdorff scoring is performed in this analysis job: the two sample groups between which hausdorff effect size scores are calculated

Alignment result excel file

This is the main output file from COPAL containing the main output data, in excel .xlsx format. It can contain a number of worksheets, depending on the type of analysis job that was run.

Data sheet

The ´data´ sheet contains the aligned complexome profiles. If normalization is performed by COPAL the data in this sheet is also normalized. The aligned complexome profiles are present as columns formatted as a heatmap. COPAL tries to preserve any additional columns present in the first input file and include them alongside the aligned abundance profiles in this sheet.

Normed not aligned sheet

If normalization and alignment is performed, this worksheet contains the normalized data without alignment. This intermediate data could for example be useful to quantify total protein abundances per sample after normalization, as alignment introduces extra fractions/slices, which affects total protein abundance.

Score sheet

If hausdorff effect size scoring is performed, this worksheet contains its result. It contains the hausdorff distances between protein migration patterns for all sample pairs. Additionally, it contains the between, within and combined columns, which are dependent on the sample groups chosen for hausdorff effect size scoring. The between column contains the mean hausdorff distance between the sample groups. The within column contains the mean hausdorff scores within sample groups. The combined column contains the final hausdorff effect size score, which is calculated by dividing the between column by the within column. The combined score, or hausdorff effect size, is also added as a column to the ´data´ sheet if scoring is performed, to allow for sorting of the profiles based on this score.

Graph sheet

If scoring is performed, the ´graphs´ sheet visualizes protein migration in graphs. It creates 200 graphs, corresponding to the top 200 proteins in the ‘data’ sheet. The data used in the graph corresponds to the protein abundances in the ´data´ sheet heatmap. By changing the order of proteins in the ´data´ sheet you can control the proteins shown in the ´graphs´ sheet. You could for example sort the proteins based on the hausdorff effect size, to visualize the 200 proteins that score the highest.

Data frame csv file

The data that is present in the excel file is also present in plain text (.csv) format. The data frame csv file contains the same information as the ‘data’ sheet of the excel output file.

Normalised data csv file

The normalized data csv file contains the information present in the ‘normed not aligned’ sheet of the excel output file in comma separated plain text format.

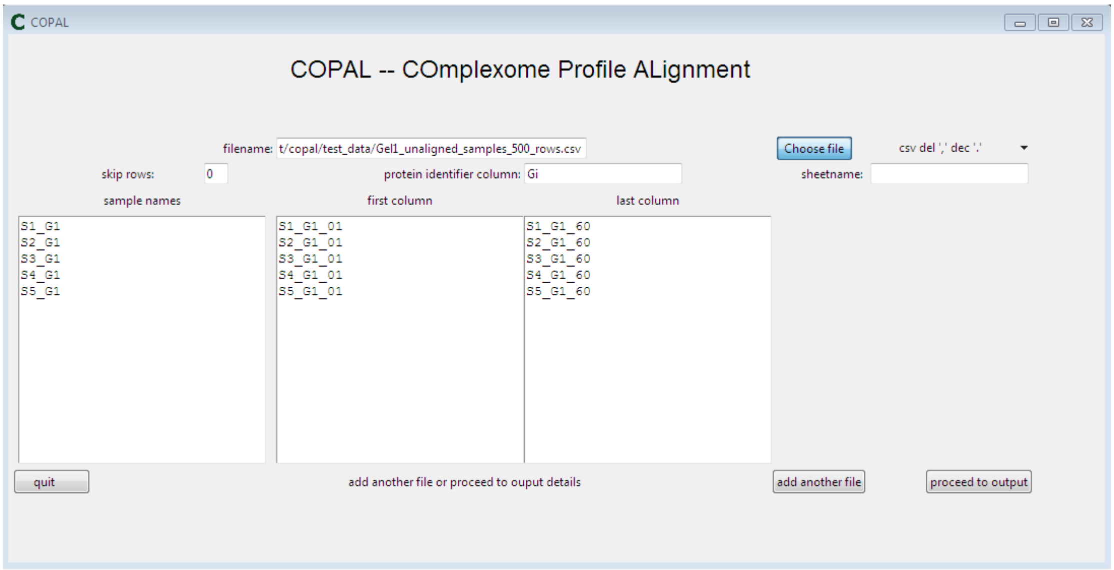
Score frame csv file

The score frame csv file contains the information present in the score sheet of the excel output file in comma separated plain text format.

Example walkthrough

This walkthrough shows how to run complexome profiling alignment of the test data provided. In this example 7 samples spread over 2 input files will be aligned and scored. Normalisation of the samples is performed based on the complete set of proteins.

When you start the GUI the file input frame shows (figure 1,2). The first file containing a set of complexome profiles is entered here. Pressing the “add another file” button will create another file input frame. Once the last input file is entered, press the “proceed to output” button. If all information is entered correctly, pressing this button will open the output details frame.



COPAL -- COmplexome Profile ALignment

filename: t/copal/test_data/Gel1_unaligned_samples_500_rows.csv

csv del ',' dec '.'

skip rows: 0

protein identifier column: Gi

sheetname:

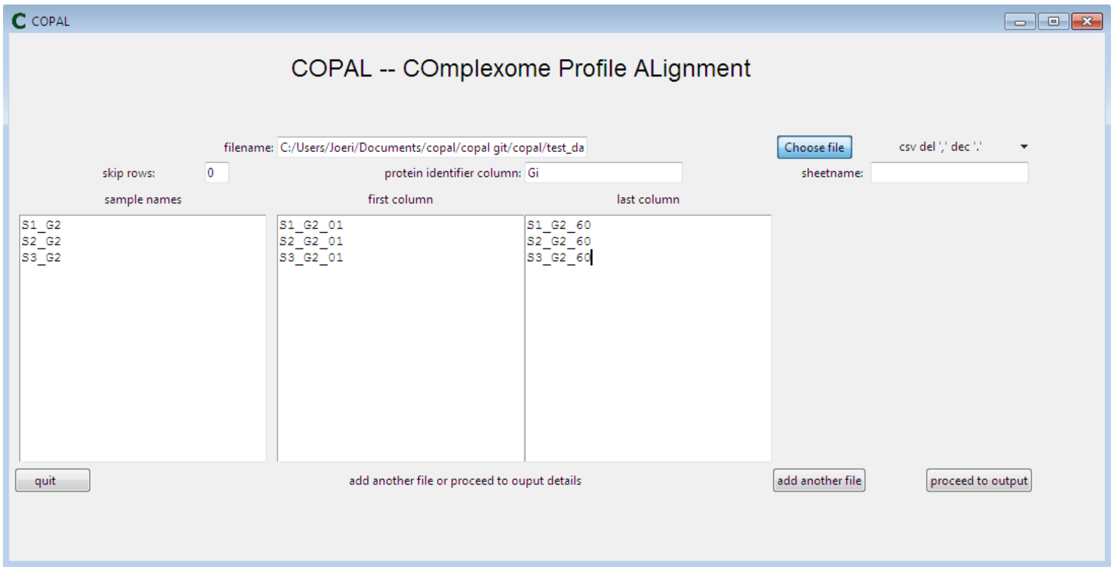
sample names
S1_G1
S2_G1
S3_G1
S4_G1
S5_G1

first column
S1_G1_01
S2_G1_01
S3_G1_01
S4_G1_01
S5_G1_01

last column
S1_G1_60
S2_G1_60
S3_G1_60
S4_G1_60
S5_G1_60

add another file or proceed to ouput details

Figure 1: First input frame



COPAL -- COmplexome Profile ALignment

filename: C:/Users/Joeri/Documents/copal/copal git/copal/test_da

csv del ',' dec '.'

skip rows: 0

protein identifier column: Gi

sheetname:

sample names
S1_G2
S2_G2
S3_G2

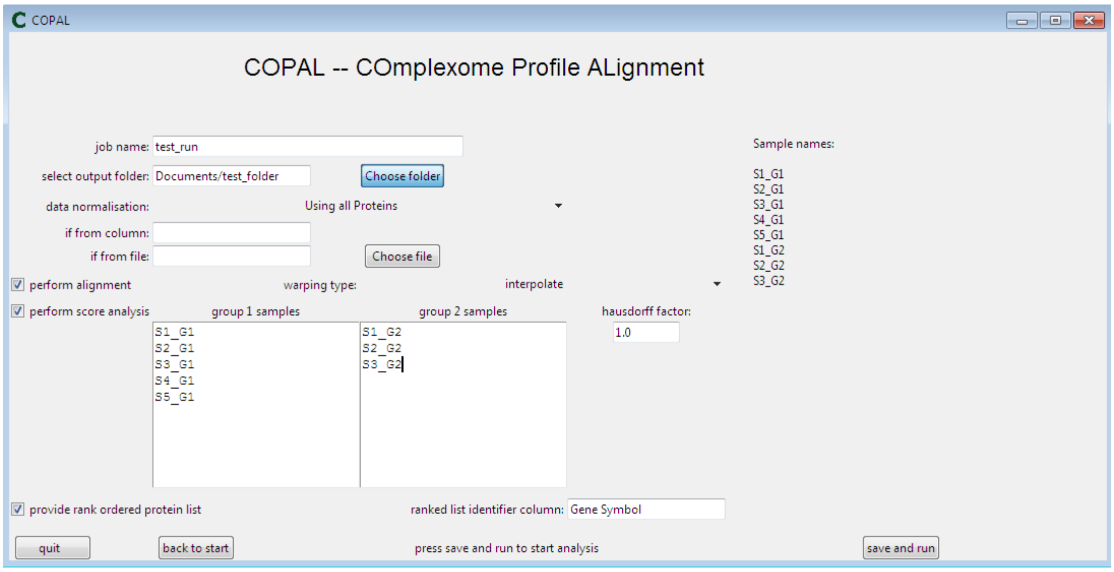
first column
S1_G2_01
S2_G2_01
S3_G2_01

last column
S1_G2_60
S2_G2_60
S3_G2_60

add another file or proceed to ouput details

Figure 2: Second input frame

The output frame displays a list of all entered samples. Specify a name for the analysis job, and then specify a preferred output folder. Use the drop-down menu to specify a normalization type. If using normalization based on a subset of proteins from a file or a column in the provided data, specify this as explained in the user guide. If the samples are to be aligned, check the box next to perform alignment. Then, select the preferred method of warping. Check the perform score analysis box if hausdorff effect sizes should be calculated after alignment. Enter sample names in the right groups for hausdorff effect sizes. If a rank ordered file with hausdorff effect sizes should be included in the output, check the “perform score analysis” box. Enter the name of the column to be added as protein identifier to the rank ordered file. To start the analysis, click save and run. This saves all input parameters and starts the analysis. The message box at the bottom displays the progress, or displays an error message if there is a problem. When the analysis is finished, a message will display the location of the folder containing the analysis results.



COPAL

COPAL -- COmplexome Profile ALignment

job name: test_run

select output folder: Documents/test_folder

data normalisation: Using all Proteins

if from column:

if from file:

☑ perform alignment

warping type: interpolate

☑ perform score analysis

group 1 samples:
S1_G1
S2_G1
S3_G1
S4_G1
S5_G1

group 2 samples:
S1_G2
S2_G2
S3_G2

hausdorff factor: 1.0

Sample names:
S1_G1
S2_G1
S3_G1
S4_G1
S5_G1
S1_G2
S2_G2
S3_G2

☑ provide rank ordered protein list

ranked list identifier column: Gene Symbol

press save and run to start analysis

Figure 3: Output details frame