Introduction

The Subcelloc script was created to automate the evaluation of the subcellular localization of a large number of proteins. It combines localization data obtained from the UniProKB database, as well as from PSORTb v.3.0.2 and Cello v.2.5. Additional information is provided by SignalP v.5.0 and TMHMM v.2.0, which predict the presence of a signal peptide and transmembrane helix, respectively. This script can be used to predict the subcellular localization of i) in silico proteomes (downloaded from UniProt) and ii) mass spectrometric data from bottom-up proteomics approaches. It was created in Anaconda (Python distribution) and optimized for use in MS Windows and Google Chrome.

Please cite:….

Preprocessing

Downloading FASTA sequences and subcellular localizations from UniProt

To obtain proteomes from UniProt, go to <https://www.uniprot.org> and use the **Proteome** tab to select the desired proteome whose subcellular localization you want to determine (e.g., *Cronobacter sakazakii*). Click on the **proteome ID**, select the desired components and click **View all proteins**. Next, select the protein entries for the desired components and press **Go**. On your first run, click on the **Columns** tab and check (i) that the **Gene ontology (cellular component)** column is present in the **Gene Ontology (GO)** tab and (ii) that the **Subcellular location [CC]** column is present in the **Subcellular location** tab. **Save** the selections (on the top or bottom right of the page). It is mandatory to have these two categories checked and saved, along with the **Entry** and **Protein name**; the other categories are redundant in terms of script functionality. Finally, in the **Download** tab, download both **FASTA canonical** (Fig. 1) and **Excel** (Fig. 2)as uncompressed files. Open the Excel file and save it as ‘uniprot’ in CSV (comma delimited) format. In other words, the file must be saved as ‘**uniprot.csv**’ exactly.

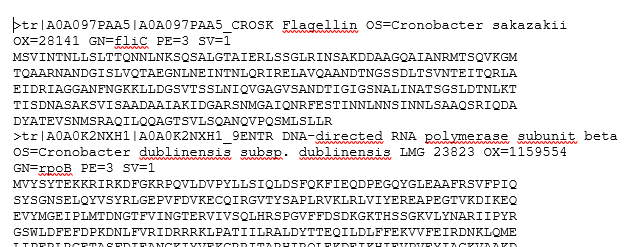


Fig. 1: Screenshot of downloaded FASTA sequences (canonical)

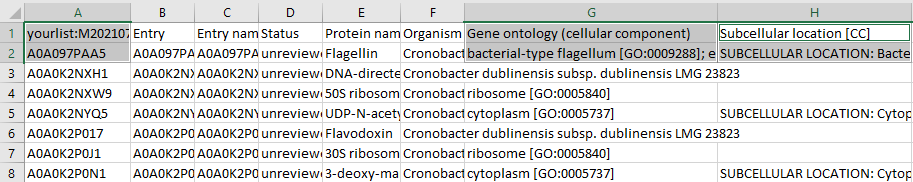


Fig. 2: Screenshot of downloaded Excel file

For the real sample proteins, copy the accession (**Entry**) numbers of the identified proteins and paste them into the **Provide your identifiers** field in the **Retrieve/ID mapping** tab on <https://www.uniprot.org>. Then, select the **From: UniProt AC/ID To: UniProtKB** option and **Submit**.

The following procedures for PSORTb, CELLO, SignalP and TMHMM are the same as for the UniProt proteomes (see above). Make sure that the **Gene ontology (cellular component)** and **Subcellular localization [CC]** columns are present in **uniprot.csv** (see above).

Downloading subcellular localizations from PSORTb

On <https://www.psort.org/psortb/>, check the following options accordingly: Choose an organism type – **bacteria**; Choose Gram stain - **negative**; Output format – **normal**; Show results - **Send by email** (enter your email address). Click **Select file**, find the path to the downloaded UniProt FASTA sequences file, and **Submit**[[1]](#footnote-1). In your email, you will receive a .txt file (Fig. 3) that must be saved as ‘**psort.txt**’.

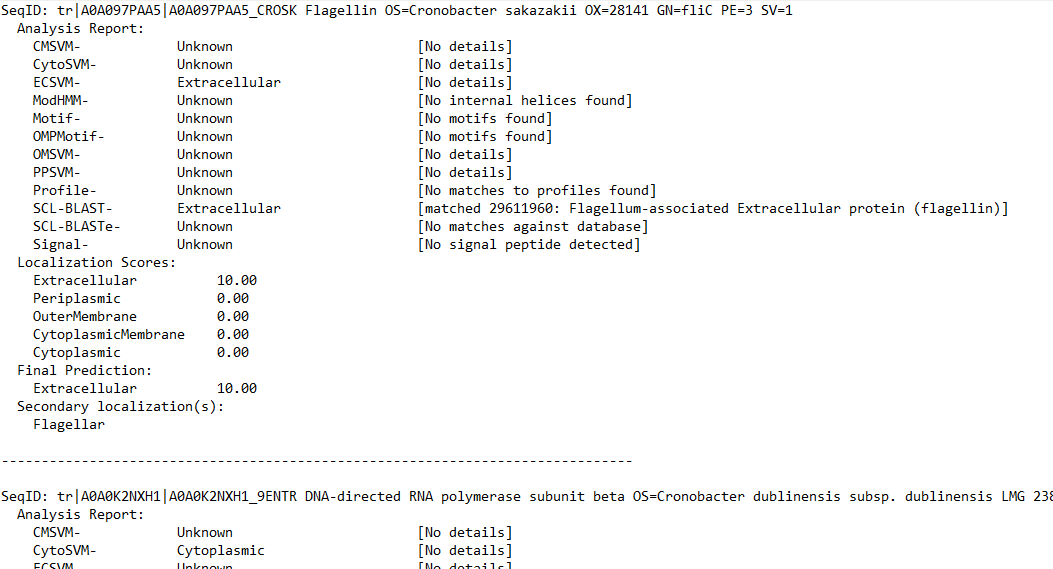


Fig. 3: Screenshot of downloaded PSORTb results

Downloading subcellular localizations from CELLO

On <http://cello.life.nctu.edu.tw/>, check ORGANISMS as **Gram negative** and SEQUENCES as **Protein**. Click **Select file**, find the path to the downloaded UniProt FASTA sequences file, and **Submit**. When the results are shown, right-click on the blue label **CELLO RESULTS**, and save and download the text file[[2]](#footnote-2) (Fig. 4) as ‘**cello.txt**’ exactly.

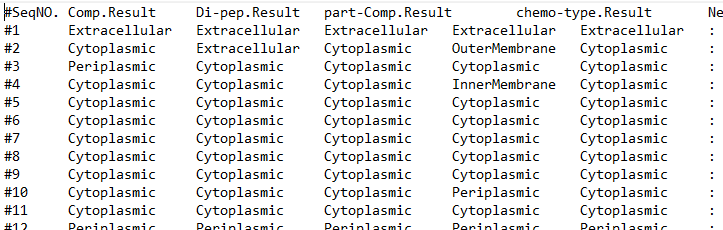


Fig. 4: Screenshot of downloaded CELLO results

Downloading signal sequences from SignalP

On <https://services.healthtech.dtu.dk/service.php?SignalP-5.0>, check the Organism group as **Gram-negative** and Output format as **Short output (no figures)**. Click **Upload FASTA File**, find the path to the downloaded UniProt FASTA sequences file, and **Submit**. When the results are displayed, download **Prediction summary** (Fig. 5) and save it as ‘**SignalP.txt**’ exactly.

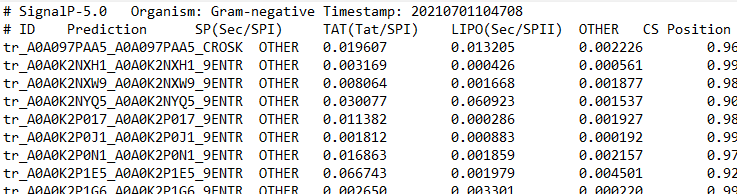


Fig. 5: Screenshot of downloaded SignalP results

Download TMHMM transmembrane helix prediction

On <https://services.healthtech.dtu.dk/service.php?TMHMM-2.0>, check Output format as **One line per protein**. Click **Choose FASTA File**, find the path to the downloaded UniProt FASTA sequences file, and **Submit**. When the results are shown, copy only the results (without the header) (Fig. 6) into the text editor and save the file as ‘**TMHMM.txt**’ exactly (Fig. 7).

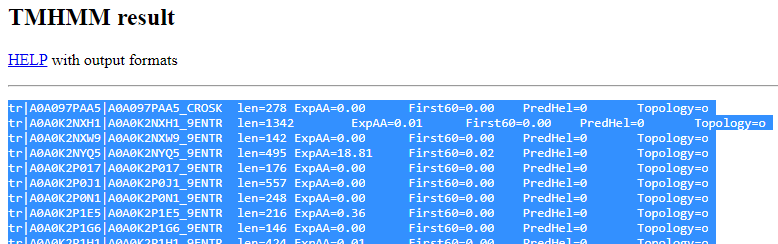


Fig. 6: Marked results must be copied to an empty text editor and saved as ‘TMHMM.txt’

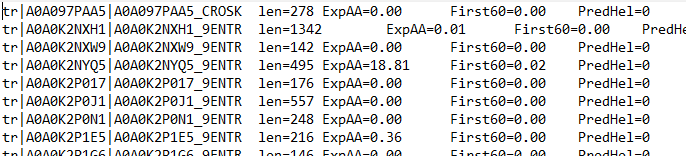


Fig. 7: Screenshot of downloaded TMHMM results

Localization prediction by Subcelloc script

Installation of Anaconda and necessary packages on MS Windows

Download Anaconda at <https://www.anaconda.com/products/individual#windows> and follow the recommended instructions. Then, open the **Anaconda Powershell Prompt** program and copy (without the > < symbols) the following to the command line prompt:

>conda install -c anaconda -c conda-forge numpy pandas jupyter xlrd openpyxl matplotlib<

When the packages have been installed, confirm the installation by pressing **Y**.

Working with the script

Download and open the **cell\_localization\_pipeline** folder. Then copy the prepared files (**uniprot.csv**, **psort.txt**, **cello.txt**, **SignalP.txt, TMHMM.txt)** to the **data** subfolder, making sure that they exactly match the naming conventions (lowercase and uppercase) shown in Fig. 8.



Fig. 8: Screenshot of the prepared files in data subfolder; C\_1 is an example of a FASTA sequences file name

Open **Anaconda Powershell Prompt** and type (without the > < symbols):

**>**jupyter notebook**<**

Then press **enter** on your keyboard to open the browser window. Next, enter the path to the *cell\_localization\_pipeline* folder (Desktop/*cell\_localization\_pipeline* if the folder was downloaded to the desktop) and select **cell\_localization.ipynb**. When the script opens in a web browser (Microsoft Edge is not compatible with this script; we recommend Google Chrome), click on the **Cell** tab and select **Run all**. In a matter of seconds, the operation will be executed, as signalled by the fact that bar graphs start to appear directly in the script without error messages (Fig. 9). In the **cell\_localization\_pipeline** folder, the Excel file **cell\_localization.xlsx** with the localization results will appear in the **output\_data** subfolder (Fig. 10).

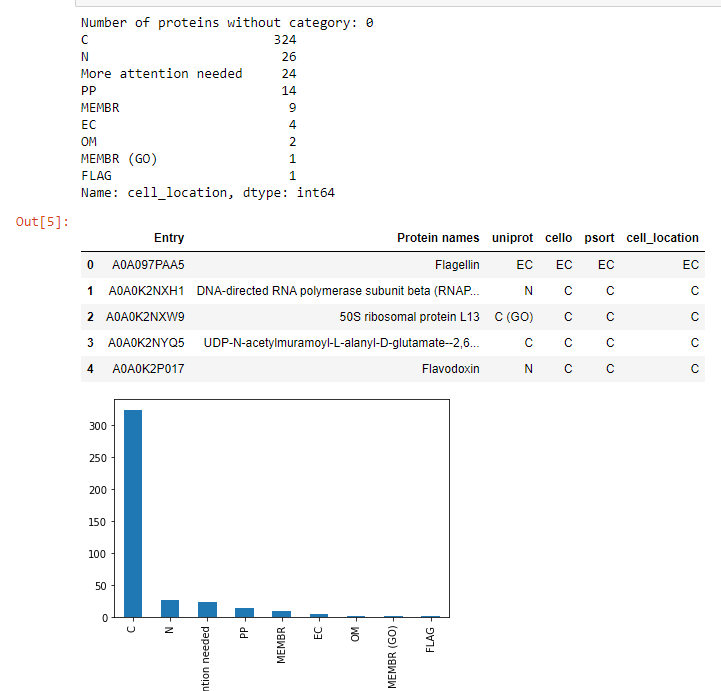


Fig. 9: Bar graphs indicate that the process is running properly

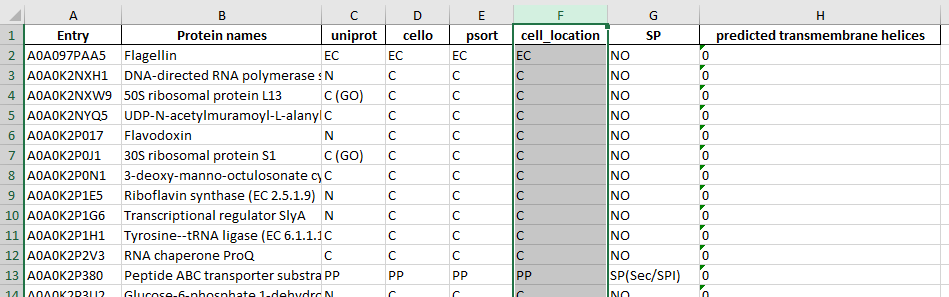


Fig. 10: Screenshot of the results in a **cell\_localization.xlsx** file

The script adds the most probable localization to the **cell\_location** column based on the obtained information:

* If at least two software tools agree on the localization, this localization is determined as the final one
* If the localization from at least two software tools is marked as unknown (N), then the final localization is marked as N
* If the software tools do not agree, but determine some localization (not N), the final localization is marked as more attention needed (MAN)
* If the software tools do not agree, but at least two results are associated with the membrane (OM/IM/PP/MEMBR), MEMBR is marked as the final localization
* If UniProt links the localization of the protein to the locomotor system of the bacterium (flagellum), the final localization is marked as FLAG
* SignalP and TMHMM provide additional information on whether a protein contains a signal sequence and how many transmembrane helices it has. The abbreviations for the specified localizations are listed in Table 1.

Table 1: Description of abbreviations in final **cell\_localization.xlsx** file

|  |  |
| --- | --- |
| **Abbreviation** | **Description** |
| EC | extracellular protein |
| FLAG | flagellar protein |
| OM | outer membrane protein |
| PP | periplasmic protein |
| IM | inner membrane protein |
| MEMBR | membrane protein; the localization software does not agree, but at least two results are associated with the membrane (OM/IM/PP/MEMBR) |
| C | cytosolic protein |
| MULTI | protein can have multiple cell localizations |
| MAN | more attention needed; the localization software does not provide an exact match, but determines some localization (not N) |
| N | unknown localization |
| (GO) | gene ontology - protein localization was found in UniProt - Gene ontology (cellular component) category |
| numbers in predicted TMH | number of membrane transitions |

If there is an error message in the script, no **cell\_localization.xlsx** file appears in the **output\_data** folder and the final data are incomplete (see Troubleshooting 2).

Final note

After localization prediction by the Subcelloc script, the MAN and MEMBR are present in the dataset, and can be further analysed. For this purpose, find details in *Guide 2 – Further data analyses*.

Troubleshooting

1. CELLO does not respond

Open the Excel file downloaded from UniProt and copy the first 1000 proteins (only the accession numbers of the proteins). On <https://www.uniprot.org>, select the **Retrieve/ID mapping** tab and copy the first 1000 proteins. Then in the **Select options** tab set **From: UniProt AC/ID To: UniProtKB**, and **Submit**. In the **Download** tab, download **FASTA (canonical)**, naming the file **1-1000.FASTA**. Then, repeat for the next 1000 proteins until N subfiles have been prepared in this way. Gradually upload these files to <http://cello.life.nctu.edu.tw/>. When the results are displayed, right-click the blue label **CELLO RESULTS**, and save and download the text file as **1-1000**. After repeating this for the other subfiles, create a new **cello.txt** file, copying into it the results from the **1-1000** file. Next, below these results, copy the results from the **1001-2000** file, and so on until all CELLO results have been appended to **cello.txt**. Then scroll up to the first protein at 1001-2000 in **cello.txt** and delete the entire header line (Fig. A). There must be no header or blank line between proteins 1000 and 1001 (Fig. B), and so on for the remainder of the incorporated subfiles.

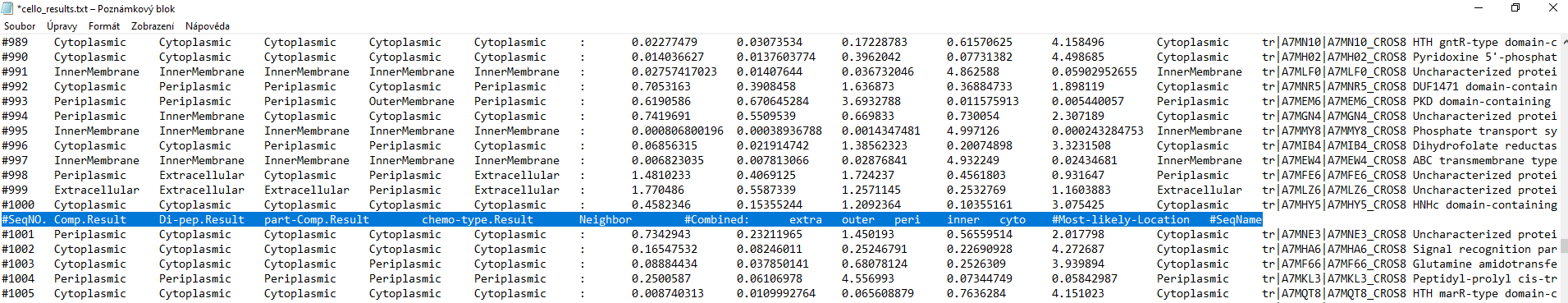


Fig. A: Marked header line between proteins 1000 and 1001 (blue color)

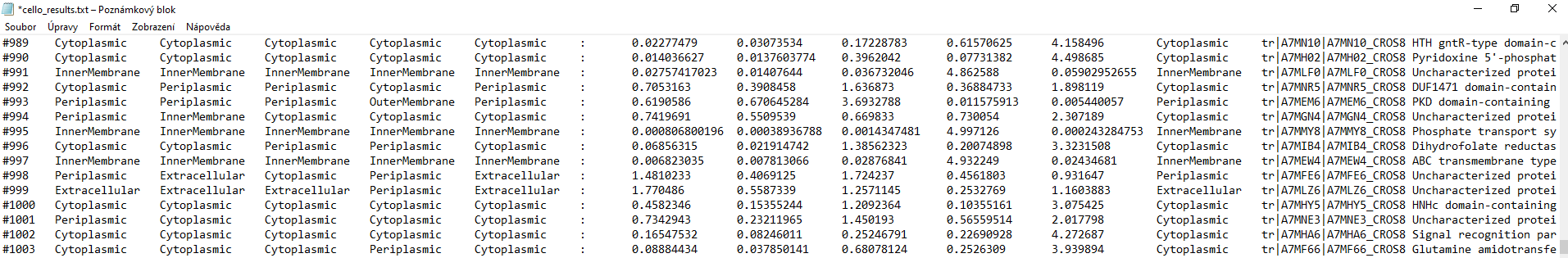


Fig. B: Two merged subfiles without a header or blank line

2: Make sure that:

* your prepared data are in the **data** subfolder and that the files name exactly match the following, paying particular attention to uppercase and lowercase letters: **uniprot.csv**, **psort.txt**, **cello.txt**, **SignalP.txt** and **TMHMM.txt** (Fig. 8; for the relevant error message, see Fig. C)
* each file (**uniprot.csv**, **psort.txt**, **cello.txt**, **SignalP.txt**, **TMHMM.txt)** is in the correct format and contains text in the form in which it appears in Figs. 2, 3, 4, 5 and 7 (for the relevant error message, see Fig. D)
* each file (**uniprot.csv**, **psort.txt**, **cello.txt**, **SignalP.txt**, **TMHMM.txt)** contains the same numbers of proteins (for the relevant error message, see Fig. E)

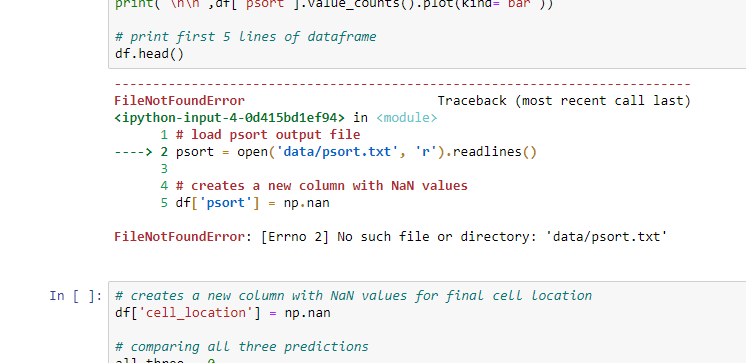


Fig. C: Error message when the name of a data file is incorrect or the file is missing (in this example, **psort.txt**)

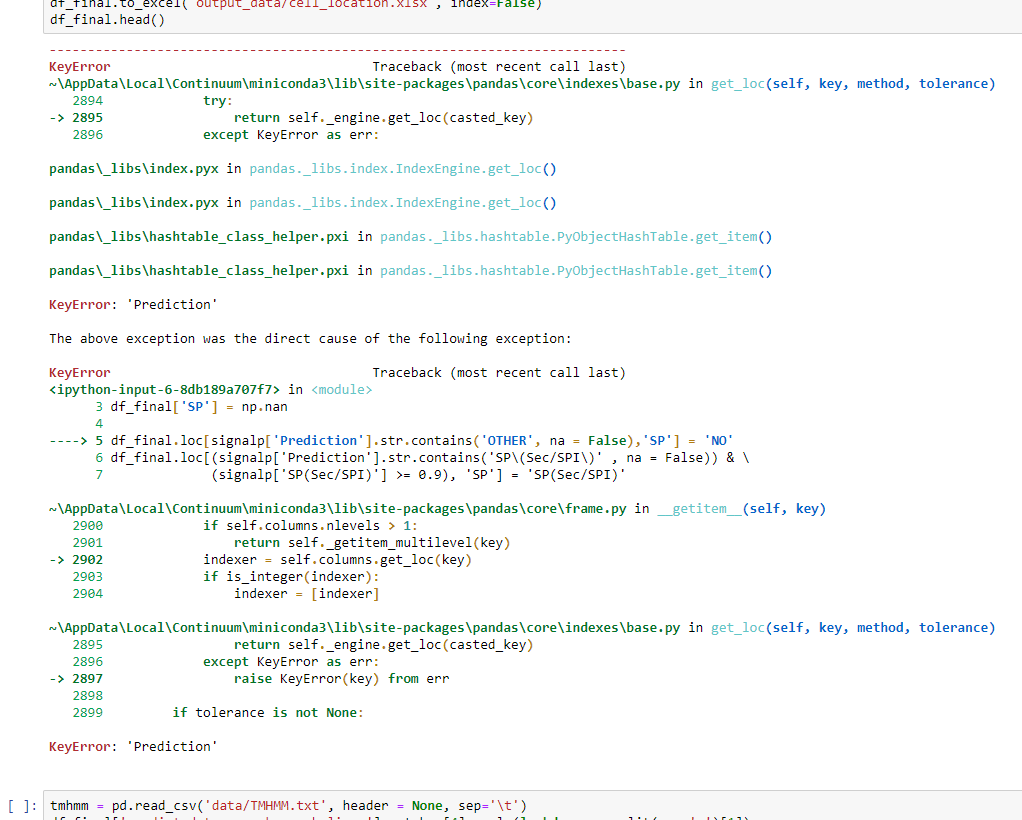


Fig. D: Error message when there is no header line in SignalP.txt (the entire.txt file is one line shorter)

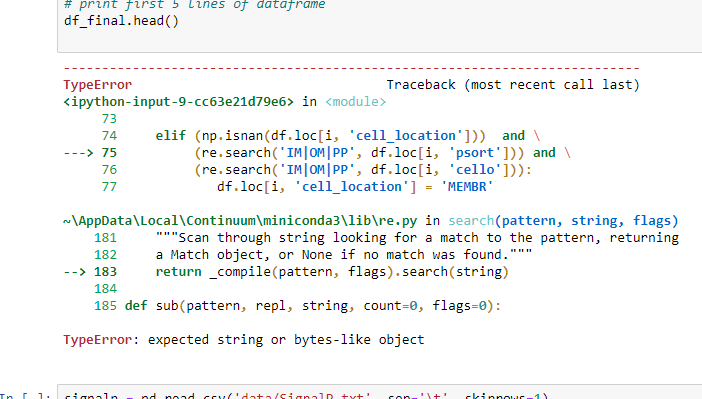


Fig. E: Error message when one data file has a different number of proteins than the others

1. 1 Even if the page suggests that the process will take a long time (for example 4500 min), within (approx.) 2 h the resulting .txt file will arrive at the given email address. [↑](#footnote-ref-1)
2. 2 Sometimes CELLO cannot handle a large number of proteins and the site does not respond. In this situation, it is necessary to divide the large set of proteins into subsets (see Troubleshooting 1). [↑](#footnote-ref-2)