Further data analysis

Introduction

For further analyses of proteins with ambiguous localization (MAN, MEMBR, N with TMHMM), we exploit information deposited in the UniProtKB database for similar proteins from related bacterial genera. The proteins are searched by sequence similarity using the UniProt BLAST function. For example, when analysing ambiguous Cronobacter proteins, a very similar protein may be found for the Escherichia coli proteome (a Cronobacter relative from Enterobacteriaceae bacterial family), and this protein may have either a **Gene ontology (cellular component)** or **Subcellular localization [CC]** localization record. To facilitate and further automate the process of determining the subcellular localization of such proteins, an in-house database for BLAST-localized proteins is used.

In this procedure, UniParc identifiers (UPIs) are used rather than accession numbers (under **Entry**) because the latter numbers can be deleted from UniProtKB during database cleanup. A UPI refers to the unique sequence of a particular protein in the UniProtKB database while an accession number only indicates a unique protein within the bacterial strain. In this way, it is possible to establish a database of BLAST-localized proteins and, thus, to identify the localization in subsequent datasets. This prevents the user from having to re-use the BLAST function for the same proteins, resulting in reduced analysis time.

Data preparation

Go to the final Excel file with the analysed data (**cell\_localization.xlsx** file in the output\_data folder generated using *Guide 1*). To change accession numbers to UPIs, copy all of the accession numbers of the downloaded or identified proteins and paste them into the **Provide your identifiers** field in the **Retrieve/ID mapping** tab on <https://www.uniprot.org>. Next, select **From: UniProt AC/ID To: UniParc**, and **Submit**. In the **Download** tab, download **Target list** (uncompressed), open it in the text editor, copy all of the UPIs and paste them into **cell\_localization.xlsx** together with the accession numbers. Then, change these two columns so that the UPIs are in column A and accession numbers in column I (Fig. 11).

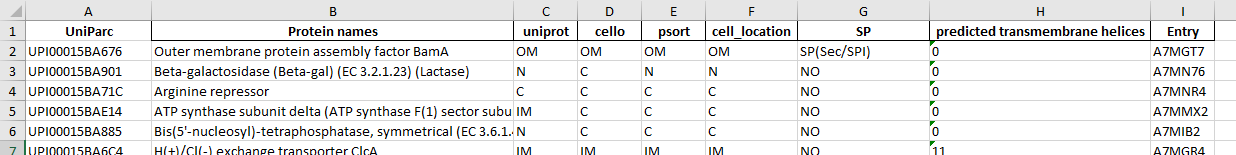


Fig. 11: cell\_localization.xlsx with UPIs

BLAST evaluation of proteins (MAN proteins used as an example)

Name the *in silico* results sheet as ***Proteins.*** When analysing the real data, it is recommended to copy the results to the corresponding real sample data Excel file to avoid mismatching proteins after the BLAST search. **Make sure that the results are copied in the same order** in which they were inserted into the script (e.g., do not mix the rows using the Filter function of MS Word). Select the first row and click on **Filter** in the **Data** tab. Next, **click on the arrow** next to the **cell\_location** column and **check only the MAN** proteins. Mark all proteins in **column A** (UniParc) and highlight them in any colour (blue in this example, Fig. 12).

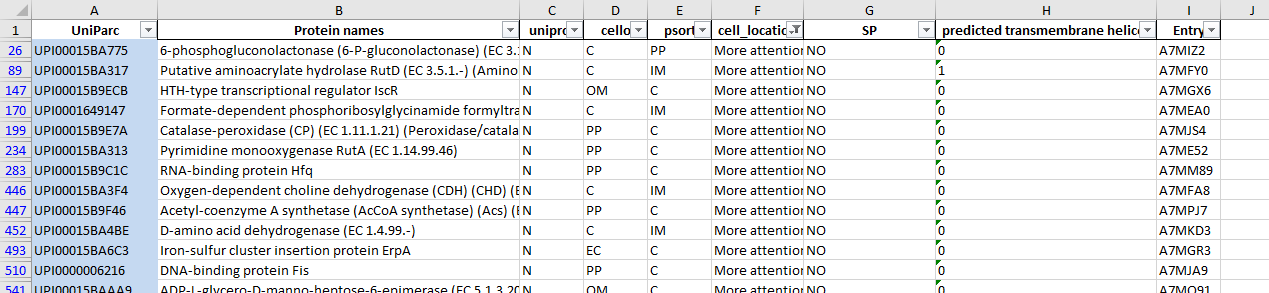


Fig. 12: Filtered and marked MAN proteins

Copy all of the information (**columns A-I**) for the filtered proteins into a new sheet and name it ***MAN***. Then name **column J**as **BLAST** and **column K** as **final\_location**.

Open <https://www.uniprot.org/blast/> and use **Target database** to select **UniprotKB/Swiss-Prot**. This database is preferred because it is the expertly curated component of UniProtKB with nonredundant protein sequences. Click on **Hits** and select 250. Copy the UPI of the first MAN protein, paste it into the BLAST window and **run BLAST**. Under **Columns**, be sure to check **Gene ontology (cellular component)** and **Subcellular localization [CC]** because these are the main categories for which information is searched. When the BLAST run has completed, display all results by clicking on **Show** and selecting **250**.

In the **Gene ontology (cellular component)** and **Subcellular localization [CC]** columns (Fig. 13), look at the list of related proteins (ordered from most related to least related). Starting at the top of the list (green square), look across at the identity score (choose a cut-off for the analyzed data of at least **identity above 70 %)** in the **Info** column (red square). If the similarity of the identified proteins is at least 70 % and is from a closely related organism (e.g., for Cronobacter spp. most often E. coli – blue square), mark and copy the **Entry** (purple square) and paste it into **column J** of the *MAN* sheet, appending the following information: related organism, subcellular localization abbreviation, and % of identity between the two proteins. In the case of the example in Fig. 13, the appended information would appear as follows: **E. coli (P37313) IM, 95%** in **column J** and **IM** in **column K**. Follow these steps for all MAN proteins.

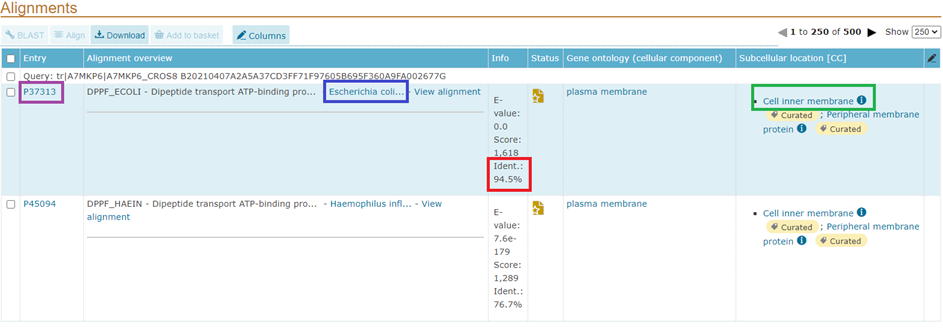


Fig. 13: BLAST results for protein with unknown localization

In the ***Proteins*** sheet, cancel the filter and then turn it on again, clicking the arrow next to **column A** (*UniParc*) and selecting **blue** for **sort by color**. Next, in the ***MAN*** sheet, select all of the localization information in **column K (final\_location)** and copy it to **column F (cell\_location)** of the ***Proteins*** sheet. In a similar vein, select all of the BLAST information in ***MAN***, **column J (BLAST)** and copy it to ***Proteins*, column J(BLAST)**. **Make sure the information is copied to the correct proteins because you cannot change the order of proteins between evaluations!**

In the same way, it is possible to specify the localization of proteins (i) evaluated as MEMBR (use a colour other than blue for MEMBR proteins, create a new sheet and name it ***MEMBR***) and (ii) that have transmembrane helices (**column H**, predicted transmembrane helices) but whose cell localization is unknown.

After the BLAST function, protein localization information can be stored and used to build a database for further use (see below)

Database creation

The proteins localized by the BLAST function can be used to create a database, meaning that in subsequent BLAST runs you will be spared evaluation for proteins already encountered.

For the next real sample dataset, create a new sheet, name it ***Database*** and copy into it the entire data file of already evaluated proteins, namely those for which information exists in **columns J and K** (Fig. 14).

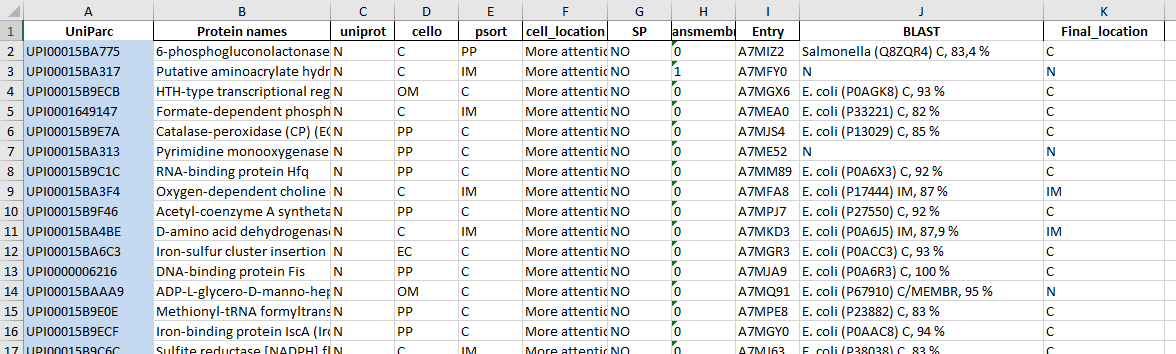


Fig. 14: Database sheet with BLAST and Final\_location information

In this new dataset, filter proteins with ambiguous localization **as described in the previous step of this guide** (sheets *MAN*, *MEMBR*, *N with TMHMM*). For example in the *MAN* sheet, copy Function 1 below (without the > < symbols), paste it into **column J**(second row), press **Enter** to confirm, and expand this function to all rows in **column J**.

Function 1

> =IFERROR(VLOOKUP($A2;Database!$A:$K;10;FALSE);"")<

Then copy Function 2 below (again without the > < symbols), paste it into **column K**(second row), confirm by pressing **Enter**, and expand this function to all rows in **column K**.

Function 2

> =IFERROR(VLOOKUP($A2;Database!$A:$K;11;FALSE);"")<

These functions copy information about the BLAST evaluation and final localization from the ***Database*** sheet to the ***MAN*** sheet. For proteins that have not yet been evaluated by BLAST (i.e., are not in the ***Database***sheet), empty cells appear (Fig. 15). Find a corresponding row for such a protein and copy all of its information (only **columns A to I**, not columns J and K) to the ***Database*** sheet. Then manually BLAST all of the newly copied proteins. In this way, the data newly added to the ***Database*** sheet will automatically be added to the *MAN* sheet because Functions 1 and 2 are still active.

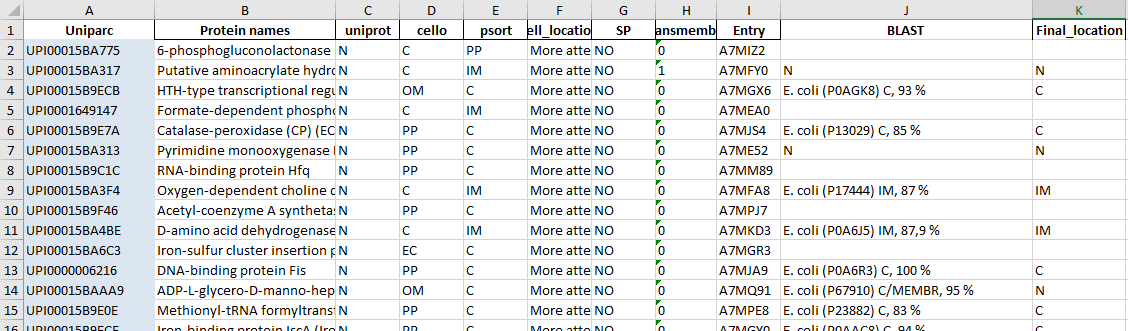


Fig. 15: *MAN* sheet of a new dataset with results partially added from the *Database* sheet via Functions 1 and 2. Rows 2, 5, 8, 10 and 12 must be copied to the *Database* sheet for their subcellular localization to be determined by the BLAST function.

In the ***Proteins*** sheet of this new dataset, cancel the filter and then turn it on again, clicking the arrow next to the first column (**Entry**) and selecting **blue** for **sort by color**. Next, in the ***MAN*** sheet, select all of the localizations in **column K (final\_location)** and copy them **as valueS** to**column F (cell\_location)** of the ***Proteins*** sheet. Then, in a similar vein, select all of the BLAST localizations in ***MAN***, **column J (BLAST)** and copy them **as valueS** to ***Proteins*, column J (BLAST)**. **Make sure the information is copied to the correct proteins because you cannot change the order of proteins between evaluations![[1]](#footnote-1)**

Final note:

This further data analysis guide has shown how it is possible to determine the protein localization information for unknown or ambiguously localized proteins. The more BLAST-determined proteins there are in the in-house database, the faster the subsequent analyses of new datasets will be. In the **BLAST** Excel sheet, we provide our current database of 2447 *Cronobacter* proteins whose localization has been specified using BLAST.

1. Be careful when editing and manipulating function-bound data when the filters are on. [↑](#footnote-ref-1)