

Short instructions for the proteomaps workflow

1 How to prepare data files for the proteomaps workflow

Proteomaps are generated by the paver software, and input files for paver can be generated by a python script that preprocesses the original proteome data. This script expects the original data to be given in a fixed table format. The input file is a table in tab-separated .csv format. It contains two columns, one containing the protein identifiers (see below), and one containing the abundance data (numbers). For clarity, the file can also contain a header row with extra information, and a row with the column headers. These two rows are marked, respectively, by "!!" and "!". Here is an example (for *E. coli* proteins):

!!Sbtab TableType='Proteomaps'	Organism='eco'
!ProteinIdentifier	!Abundance
b0878	7
b0948	12
b3933	12
..	..

The file attributes in the first row (such as "Organism" in the example) contain additional information. Recommended possible attributes are "Organism", "Condition", "Tissue", "ValueType", "Unit", and "Title". The leading rows starting with "!" are optional and are currently not evaluated by the python script; thus, it is also possible to simply write

b0878	7
b0948	12
b3933	12
..	..

In any case, it is the order of columns that counts. The protein identifiers (in the first column) must be chosen according to the instructions given in "How to add a new organism to the proteomaps workflow". The abundance data (second column) must represent protein molecule count numbers. Do not use data that correspond to mass-weighted count numbers (mass weighting will be done anyway in the workflow). Do *not* use logarithmic values.

Several data sets can be provided together, but must be given as separate files. For each file, the following information needs to be given:

1. The organism (three-letter) shortname
2. A short name of the experiment or paper corresponding to the data set, to be used within file names (e.g., Valgepea_2013_48)
3. A human-readable name of the experiment or paper corresponding to the data set (e.g. "Escherichia coli ($\mu=0.48/h$) - Valgepea et al. (2013)")

2 How to add protein annotations

If you notice that proteins in proteomaps are wrongly placed in the functional hierarchy or not mapped at all, you can change this by adding new protein annotations. To do this, you need to find the proteins to be annotated (or reannotated, which does not make a difference here), their protein identifiers, protein names, and KO numbers, as well as the desired pathway annotations, and put all this information into a table file. The file format for your additional annotations follows the format of the file `genomic_data/KO_gene_hierarchy/KO_gene_hierarchy_changes.csv` in the proteomaps github project. The columns are as follows:

Organism shortname
Protein identifier
Protein name
KO number
Pathway name
Comment or provenance information

Example (human hemoglobin Hba2, assigned to the pathway "Hemoglobin"):

hsa	P69905	Hba2	K13822	Hemoglobin	http://ghr.nlm.nih.gov/gene/HBA2
-----	--------	------	--------	------------	---

To see our list of pathways, you can browse the hierarchy tree file (at <http://www.proteomaps.net/download.html>, link "Hierarchy tree"). A hierarchy file without the KO numbers (which is therefore more easy to browse) can be found in the github project: `/genomic_data/KO_gene_hierarchy/KO_gene_hierarchy_level_3_DO_NOT_EDIT.tms`.

When your file is ready, please send it to Wolf. If you think that a pathway should be added to the hierarchy or that the hierarchy should be restructured, please contact Wolf.

A suggestion: if you'd like to focus on not-mapped, highly expressed proteins, have a look at the preprocessed version of your proteome data file (which contains, for every protein, the current pathway annotation). If you sort the proteins by abundance, you will see which proteins deserve new annotations.

3 How to add a new organism to the proteomaps workflow

Currently the workflow can only handle the organisms from the following list. For each of them, a specific type of protein identifiers must be used (the examples shown refer to enolase proteins):

Organism	org	ID type	Example	Example URL
Mycoplasma pneumoniae	mpn	Locus tag	MPN606	www.ncbi.nlm.nih.gov/gene/?term=MPN606
Mycobacterium tuberculosis	mtu	Locus tag	Rv1023	www.genome.jp/dbget-bin/www.bget?mtu:
Escherichia coli	eco	Locus tag	b2779	www.genome.jp/dbget-bin/www.bget?eco:b2779
Synechococcus elongatus sp. PCC7942	syf	CyanoBase	Synpcc7942_0639	genome.microbedb.jp/cyanobase/SYNPCC7942/genes/slr0752
Synechocystis sp. 6803	syn	CyanoBase	slr0752	genome.microbedb.jp/cyanobase/Synechocystis/genes/slr0752
Saccharomyces cerevisiae	sce	Locus tag	YHR174W	identifiers.org/sgd/YHR174W
Schizosaccharomyces pombe	spo	PomBase	SPBC1815.01	www.pombase.org/spombe/result/SPBC1815.01
Arabidopsis thaliana	ath	Tair	AT1G74030.1	www.arabidopsis.org/servlets/TairObject?type=locus&name=AT1G74030.1
Drosophila melanogaster	dme	FlyBase	CG17654	www.genome.jp/dbget-bin/www.bget?dme:Dmel_CG17654
Mus musculus	mmu	NCBI Gene Id	13806	www.ncbi.nlm.nih.gov/gene/13806
Pan troglodytes	hsa	UniProt	P06733	identifiers.org/uniprot/P06733
Homo sapiens	hsa	UniProt	P06733	identifiers.org/uniprot/P06733

The list is given as a file in the proteomaps project on github (file `genomic_data/KO_gene_hierarchy/organisms.csv`) and on the proteomaps website (<http://www.proteomaps.net/download.html>, link “Organism information”) To add a new organism to the list, the necessary information need to be prepared, and a file with protein lengths must be provided.

As an example, assume that you wanted to add *E. coli* as a new organism (which, in fact, is already included). For the example, consider the KEGG page <http://www.genome.jp/dbget-bin/www.bget?eco:b3124> for enolase in *E. coli*.

1. Find out KEGG organism shortname (three letters)). On the KEGG page, it appears in the field **Organism**: (in the example, `eco` for *Escherichia coli* K-12 MG1655). Make sure that not only the organism, but also the strain matches the one for which you want to generate proteomaps. Choose the organism name (e.g., “*Escherichia coli*”) to be used in proteomaps (which can, but does not have to coincide with the name used in KEGG).
2. Find out the type of protein identifiers used in KEGG for this organism. On the KEGG page, it appears first in the field **Entry**: (in the example, `b2779` for enolase. Make sure that your protein data carry the same type of protein identifiers.
3. Find the protein identifier for enolase (which we use as an example case)
4. Find a reference database that provides URLs for these protein identifiers (see table above). In the case of *E. coli*, we simply use KEGG itself: www.genome.jp/dbget-bin/www.bget?eco:; in other cases, we use an organism-specific database, e.g. genome.microbedb.jp/cyanobase/ for cyanobacteriae).
5. Prepare a table with protein lengths for these proteins. Note that length information refers to **PROTEINS** and not to mRNA! Important: unlike other tables (which use protein identifiers as keys), this table must use protein shortnames (e.g., “`eno`”) as keys. The format is:

Gene names	Length
<code>eno</code>	350
..	..

Information about protein lengths can be obtained from uniprot, but due to gene name conversions, there may be some loss of information.

6. Send all this information, as well as a test protein data set (see instruction “How to prepare input data”), to Wolf; he prepares the input files for paver, and Dan makes the first picture.
7. Be aware that heavy reannotating is usually necessary after the first test pictures.

4 Contact

wolfram.liebermeister@gmail.com