KAEA Docker tutorial

This is a tutorial intented to explain to KAEA users how to create a docker container of the pipeline using docker and git.

1) Docker Installation

The first step is to install Docker on the host system. This will allow you to retrieve an image of the pipeline. The first step to do is to create a docker account: https://id.docker.com/login/ Then download the docker installation file that corresponds to your OS and install it.

Check Docker installation with:

```
docker --version
```

2) Pull Docker Image

Create a directory where you would like to work and head in that directory. Type in:

```
docker pull mhallal/kaea:1.0
```

This might take ~5 minutes.

3) Run a docker container

```
docker run -v <local_data_path>:/home/user/KAEA/data --name kaea1 -it mhallal/kaea:1.0 /bin/bash
```

<local_data_path> being the path to the MaxQuant and NetworKIN data file on the local host.

4) Activate conda environment

```
source activate snakemake-general
```

You have set up a conda environment which is ready to run.

5) Prepare NetworKIN file

This is an optional step, but we recommend using NetworKIN input.

Paste the 'http://networkin.info/index_ht.shtml' into your browser and refresh, this will direct you to the NetworKIN platform. NetworKIN takes as input the phosphorylated sites in the following format "Protein ID, Position, Residue" (space/tab delimited), i.e:

A0AVK6 71 S

Q2QGD7 172 T

Q9UJM3 394 Y

NetworKIN accepts different formats, please refer to the web page above for mroe details.

Prepare the input data as recommended above from your MaxQuant phosphoSTY file and copy-paste the table in the assigned box. Submit the results and wait for the output which you will save as a csv file in your chosen directory above (point 3). Use the direcotry of your NetworKIN file for the next step.

6) Prepare config.yaml parameters file

Go to the config file (config.yaml) and modify according to your samples.

```
vi config.yaml
```

The KAEA needs 16 parameters to run smoothly. For an experiment using cell line CL1 with 2 replicates for control (CTRL) and 2 with drug (DRG), these parameters are represented below:

Parameter	Description	Example
fdr_cutoff	FDR cutoff value	0.05
p-value_cutoff	p-value cutoff	0.05
cell_line	Cell line/ Sample name	CL1
Species	Human, Mouse or Yeast	Human
Conditions	Conditions	CL1_CTRL,CL1_DRG
control	Control condition	CL1_CTRL
Samples	Tested condition(s)	CL1_DRG
Biological_replicates	Replicates per condition	2-2
SILAC	Is this SILAC data? (T/F)	F
Imputation	Impute missing values (T/F)	F
CWD	Working directory	/your_path
path_Input	phosphoSTY input path	/your_path/phosphoSTY.txt
NWKIN_Input	NetworKIN database path	/your_path/NetworKIN.csv
Intensity_columns	Names of intensity columns	[“Intensity.CL1_C1_INC1”,“Intensity.CL1_C1_INC2”, “Intensity.CL1_C2_INC1”,“Intensity.CL1_C2_INC2”, “Intensity.CL1_D1_INC1”,“Intensity.CL1_D1_INC2”, “Intensity.CL1_D2_INC1”,“Intensity.CL1_D2_INC2”]
Sum_Conditions	Sum columns of same sample (T/F)	T
Conditions_to_sum	[[“Final column name condition 1”, “Column 1 condition 1”,	[[“Intensity.1_CL1_CTRL”, “Intensity.CL1_C1_INC1”,

Parameter	Description	Example
	"Column 2 condition 1"], ...]	"Intensity.CL1_C1_INC2"], ["Intensity.2_CL1_CTRL", "Intensity.CL1_C2_INC1", "Intensity.CL1_C2_INC2"], ["Intensity.1_CL1_DRG", "Intensity.CL1_D1_INC1", "Intensity.CL1_D1_INC2"], ["Intensity.2_CL1_DRG", "Intensity.CL1_D2_INC1", "Intensity.CL1_D2_INC2"]]

Remark: the final name of a column that the KAEA can parse should be in the following format "Intensity.ReplicateNumber_Condition" such as :"Intensity.1_CL1_CTRL" and "Intensity.1_CL1_DRG"

7) Run snakemake pipeline

After you have modified the config file, you are ready to run the pipeline by typing in:

```
snakemake --cores 2
```

This might take ~1-10 minutes depending on the size of your dataset.

8) Copy the output file to the local host

After the pipeline is done, you will find the outputs in the results folder (/home/user/KAEA/results). To copy the results to your local host:

Ctrl+p

Ctrl+q

The container is still running now. To copy the results:

```
cp kaea1:/home/user/KAEA/results <desired_path_on_local_host>
```

9) Reconnect to the running container

```
docker exec -it kaea1 /bin/bash
```

Note that 'kaea1' is the container name suggested but you may use any name that suits your project.

10) Open the shinyApp and upload your data:

```
Paste 'https://mh-apps.shinyapps.io/ShinyApp/' into your browser and refresh.
Upload the 'results_shiny_CELLLINE_X.XXP_X.XXFDR.Rda' from your results folder.
```

Go through the tabs of the app:

1) Data Quality: barplot and venn diagram for phosphosites, phosphopeptides and phosphoproteins.

2) Phosphosites: volcano plot of the phosphosites. Hover over the desired point to get the quantification barplot.

3) Enrichment: heatmap and barplot of enrichment analysis. Press over the desired kinase to get more information.

4) KEGG pathways: mapping to the KEGG pathways. Change the pathway in the drop-down menu on the left side.