Galaxy

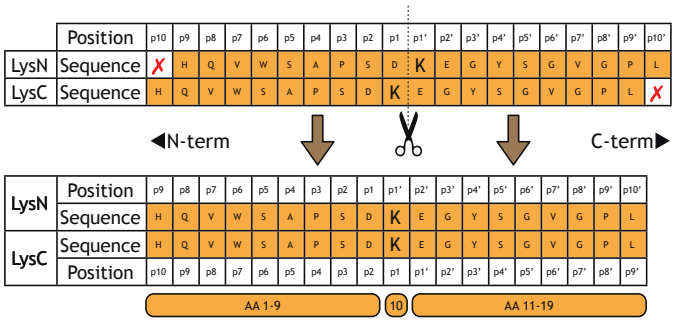
This tutorial will introduce you to the Galaxy framework, which allows you to quickly wrap command line tools into a web-based portal. Galaxy is not the first framework designed for this purpose as similar functionality is also provided for example by wEMBOSS, Pise and SRS. Galaxy is the most recent / modern one though taking functionality a bit further compared to previous projects.

Galaxy is developed at the universities of Penn State & Emory. For more info: ***http://galaxy.psu.edu/***

In this tutorial we'll analyse the cleavage specificity of 2 proteases: LysC and LysN. Both cleave proteins preferably at a Lysine (K) residue with LysC cleaving c-terminal of K and LysN on the n-terminal side of K (see top part of figure 1).

Sometimes these enzymes will cleave at other residues or cleavage may be inhibited by specific amino acids surrounding K. To analyse the cleavage specificity we'll analyse pre-aligned sequence contexts of a large number of cleavage sites (see bottom part of figure 1). These sites were derived from peptides which in turn were derived from human cell lysates treated with LysN or LysC. The peptides were separated by LC and identified by MS/MS followed by a database search.

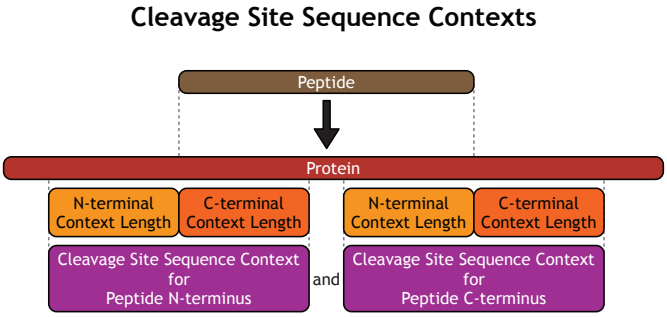
Figure 1



In Galaxy we will:

- Upload the peptides as well as the FASTA file used for the database search to identify the peptides
- Map the peptides back to the protein they were derived from (see figure 2)
- Get a fixed-width sequence context for each cleavage site (see figure 2)
- Remove redundancy from the sequence contexts
- Make a "sequence logo" to see if certain amino acids are enriched / depleted on the positions surrounding the cleavage sites

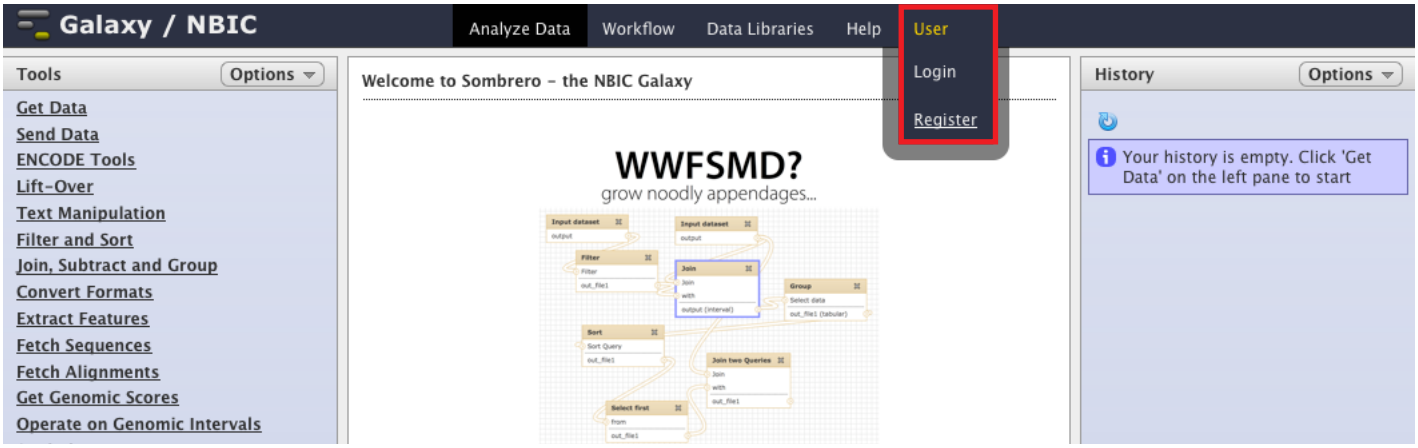
Figure 2



1. For this tutorial we'll use the public NBIC Galaxy server "Sombrero": http://galaxy.nbic.nl

Click *Click here to start analysing your data*. You should get a start page like the one below. There is a menu on the top followed by 3 sections:

- On the left is a list of all tools available in this Galaxy
- On the right is your *history,* which contains all the data sets (inputs, intermediate outputs and final results) of your analysis
- In the middle is a section where you can either view a data set or specify the details to run a tool



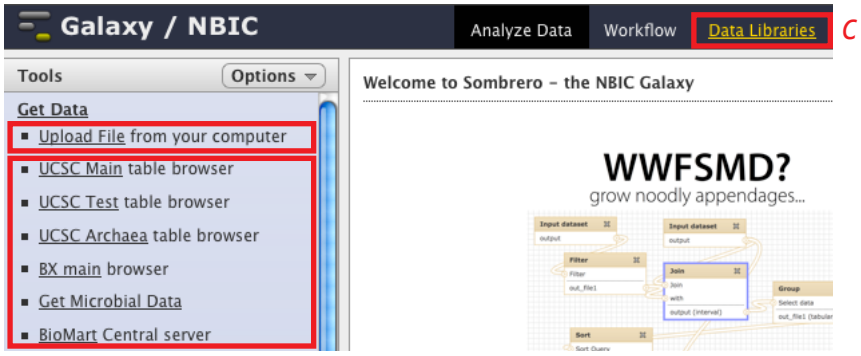
2. Login

Go to the *user* menu and login or - if you don't have an account yet - create one.

3. Fetch the input data

There are 3 ways to get data into your history:

A. Upload a file using *Tools* -> *Get Data* ->*Upload file from your computer*
B. Fetch data from an external resource (website) shown in *Tools* -> *Get Data* -> all the ones below *Upload file*
C. Fetch a data set from the *Data Libraries* menu



We'll do the latter. When the same data set like a file with reference sequences is used repeatedly by different users or for different types of analysis, an admin user can make these available as a data library. The obvious advantage for the users is that they don't have to upload the same file over and over again. In addition the advantage for the admin is that the file is only stored once: each user that imports data from a data library receives only a reference to the data in their history to prevent wasting disk space on redundant data.

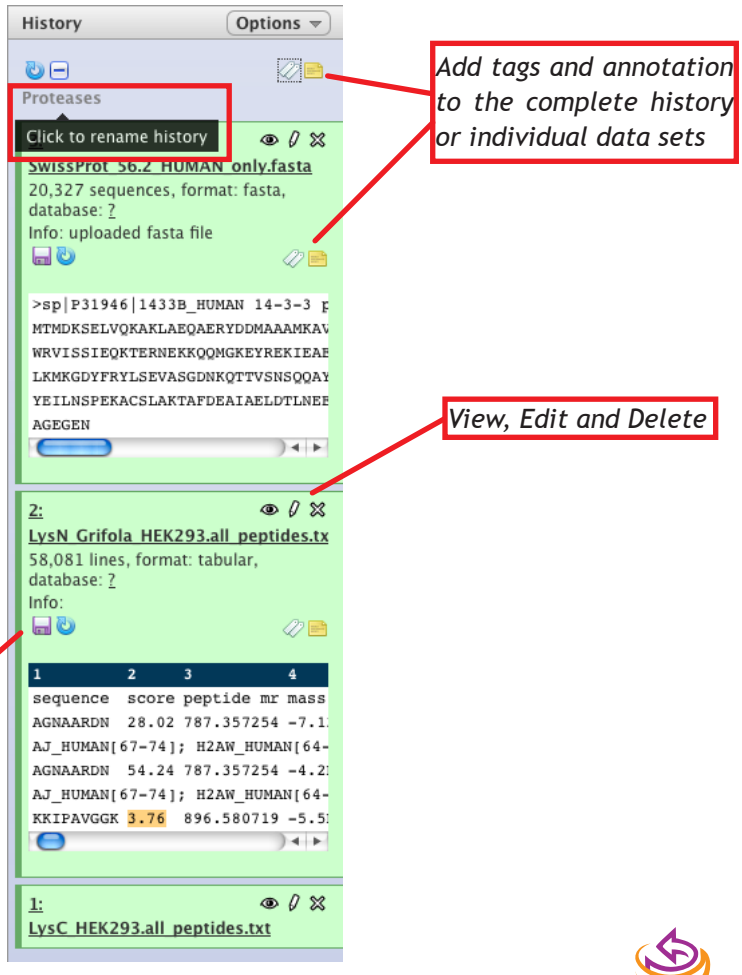
When you click on the *Data Libraries* menu, you'll see a library called *LysN versus LysC*. Select all 3 files from this library and add them to your history.

4. Inspect the data in your history

Click *Analyze Data* in the top menu to go back to the main page. You should now see the 3 data sets in your history panel as shown on the right.

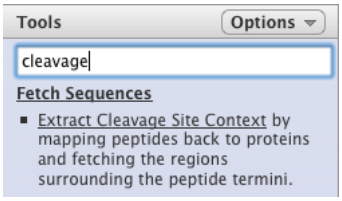
Clicking a data set will expand it to show you a preview, some meta data and the download (floppy disk) & re-run buttons. For now all we need to do is give the history a more meaningful name so we can easily find it back later on: click the history's name, type a new one and hit <enter>.

In the top right corner of each history item, you'll see 3 icons. With these you can view the complete data set in the middle part of the browser window (eye icon), edit the meta data (pencil icon) and delete the history item (cross icon).



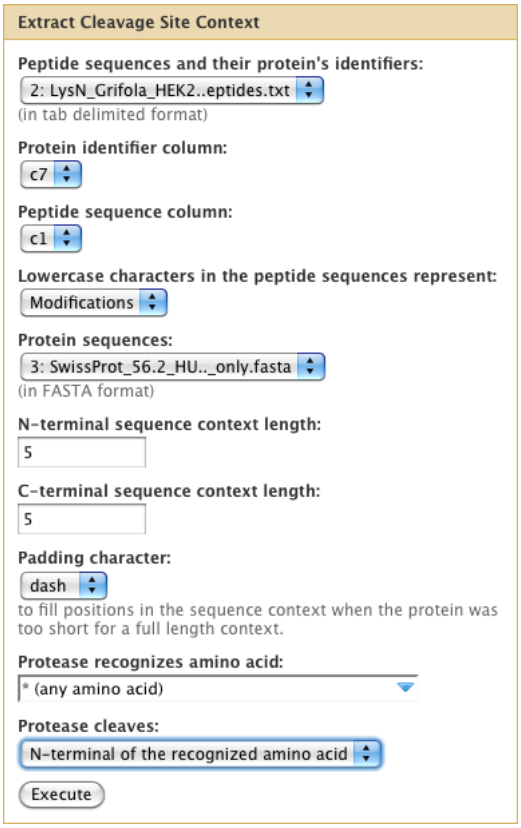
5. Fetch cleavage site sequence context

Let's go to the tools pane and search for a tool to map the peptides back to the protein they were derived from and get a fixed-width sequence context for each cleavage site. If the tool search is not visible click *Options -> Tool Search*.



We need to run this tool twice: once for the LysN and once for the LysC data set.

- Inspect the data sets to see which columns contain the peptide sequences and protein IDs (the latter column is named *all protein matches*). Adjust the *Protein identifier column* and *Peptide sequence column parameters* in the tool config accordingly.
- Change the Protease recognizes amino acid parameter to * (any amino acid), so we'll also get the non-specific cleavage sites.
- When you process the LysN data set, change the last parameter *Protease cleaves* to *N-terminal of the recognized amino acid*.
- Set the *sequence context lengths* to:
  - For LysC:
    - *N-terminal*: 10
    - *C-terminal*: 9
  - For LysN:
    - *N-terminal*: 9
    - *C-terminal*: 10
- Use defaults for the rest and hit execute.

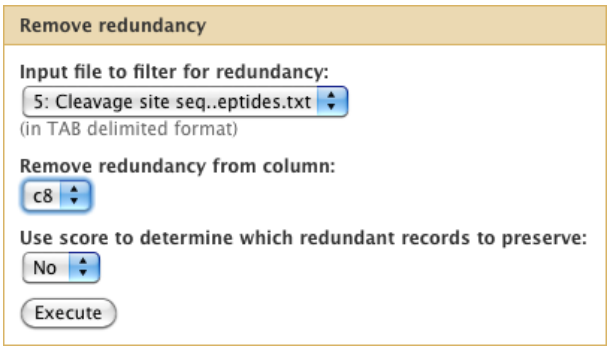


Note 1: The asymmetric sequence context lengths makes the amino acid recognized by the protease the central one in the sequence context.

Note 2: The SwissProt FASTA file was the only FASTA file in your history and hence the only compatible input for the *Protein sequences* parameter. Therefore it is the only one in the popup menu for this param. Similarly the FASTA file is not a compatible input for the *Peptide sequences and their protein's identifiers* param and therefore it is not available in that param's popup menu.

You'll see 2 new data sets in your history: one for LysN and one for LysC. These contain an additional 8th column with the sequence context of the cleavage sites.

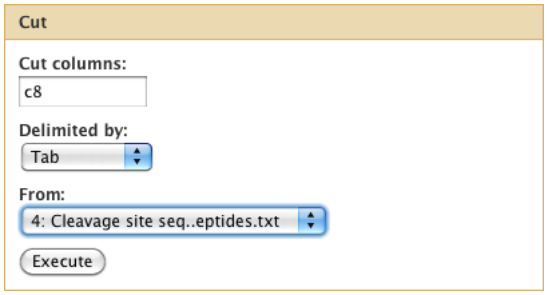
6. Remove redundancy



We want to see if certain amino acids are overrepresented in cleavage sites for LysN as compared to in cleavage sites for LysC. In a Proteomics experiment usually only a fraction of the peptides is identified. Therefore some cleavage sites may be identified twice if both the peptide to the left and the one to the right of the site were identified, while for others the cleavage site context will be present only once in our data. This will cause artificial over- and under representation, so we have to remove redundancy from the data set.

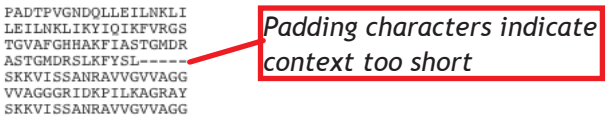
Use the *Remove redundancy from tabular data sets* tool to remove redundancy from the column with the sequence contexts (number 8). Do this for both the LysN and the LysC cleavage site contexts.

7. Cut sequence context column



Next we simply cut only the last columns with the sequence contexts from the data sets of the previous step. We do this for both LysN and LysC with the *Text Manipulation -> Cut columns from a table* tool.

8. Filter for too short contexts

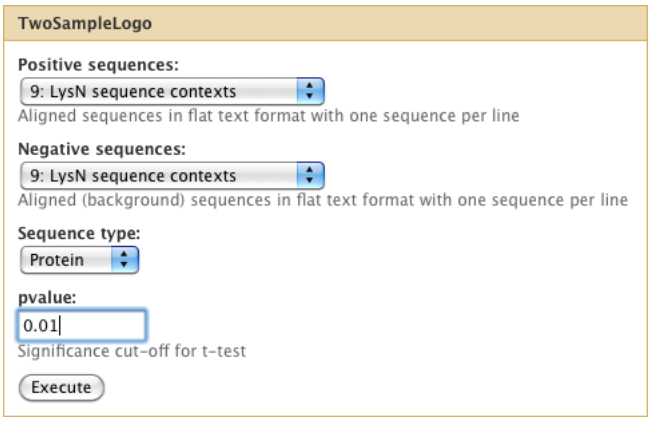


Some of the cleavage sites are too close to the protein termini to get a full length context. If you view the data sets by clicking the eye icon, you'll see some truncated contexts where the sequences are supplemented with a padding character ("-" by default). These too short sequences are problematic for further analysis, so we'll have to remove them. We do this for both LysN and LysC with the *Filter and Sort -> Select lines that match an expression* tool by selecting lines *Not matching* a - character.

Note: you'll notice that the *Cut* and *Select lines that match an expression* tools produce output that is named after the tool and the history item number of their input. Using the history item number you can trace back which history items correspond to our LysN and LysC sequences, but for clarity you may want to rename the history items by clicking the pencil icons and add LysN or LysC to their names

How many sequence contexts did we loose for LysN and LysC?

9. Make a sequence logo



To visualize over- and under represented amino acids in the LysN cleavage sites versus the LysC cleavage sites, we'll make a sequence logo. One of the tools to make sequence logos is TwoSampleLogo and this one is available on the NBIC Galaxy.

Which protease is more specific?

Can you spot a trend in the bias towards under- or over representaion of certain amino acids near the cleavage sites?

10. Make a workflow

If we would have to repeat this analysis for other proteases or other peptide data sets, we can create a workflow from our current history by clicking *Options -> Extract Workflow* in the top right corner of the history pane. You can now select *Workflows* from the top menu and from there select your new workflow to run or modify it. For each step (tool) in a workflow you can specify for each parameter whether it should be hard coded to the value you used in your history or whether it should be possible to select a different value. Hence for the workflow we just generated we could fix all parameters based on the ones we just used except for the input files. Give it a try!