Objectives:

A. To understand the basic principles of quantitative proteomics methodologies

B. To explore online resources for „omics“ data mining
Homework #5 (Lectures April 27 & 29)
A) Problem set "Quantitative Proteomics"

1) In lecture, you heard about three basic methods to identify proteins via mass spectrometry: "Top down", "Bottom up" and "Middle down". Try to think about samples from different sources for which the application of each of these methods might be of advantage and briefly give reasons.
   a) "Top down"

   b) "Bottom up"

   c) "Middle down"

2) The detection of low-abundance proteins in gel-based proteomics approaches can be considerably improved by fractionation techniques, such as:
   a)

   b)

3) Two methods of relative protein quantification by MS after in vivo sample labelling were discussed in class.
   => List these two methods and briefly describe their main difference.
   a.

   b.

4) In a few keywords, describe the DIGE technique and its key advantage over traditional comparative 2D-GE:
5) The sequence of proinsulin (after cleavage of the signal peptide) and its three disulfide bonds is shown below. *In vivo*, Proinsulin Protease cleaves amide bonds between two basic amino acid residues, thus the products of this protease reaction yield:
- the two chains of active insulin (A-chain (N-terminal) and B-chain (C-terminal))
- an inactive peptide called C-chain.

a) On the structure below, circle and name the chains after Proinsulin Peptidase cleavage.

b) Describe how the mature insulin structure would look like in case Trypsin was used for digestion?

6) Optional exercise (taken from the Lehninger online interactive quiz section, where you could look up the solution - of course *after* the exercise ;-)  

A polypeptide is cleaved into peptides by treatment with trypsin and cyanogen bromide, and the peptides are purified and sequenced. The sequences of the peptides (from N to C-terminus) are shown below. (Note: Trypsin cleaves after K and R residues; cyanogen bromide cleaves after M.) Based on sequences of the overlapping peptides generated by treatment with trypsin and cyanogen bromide (shown below), which of the peptides could represent the N-terminus of the polypeptide?

**Trypsin peptides**
- T-1 GASMALIK
- T-2 EGAAYHDFEPIDPR
- T-3 DCVHSD
- T-4 YLIACGPMTK

**Cyanogen bromide peptides**
- C-1 EGAAYHDFEPIDPR.GASM
- C-2 TKDCVHSD
- C-3 ALIKYLICGPM
B) Problem set: "Exploring Omics data"

=> Answers to questions (marked in bold throughout this problem set) can be turned in on a separate sheet of paper or sent via email to avoid too many print-outs

1) If you are not yet feeling comfortable with using online tools like KEGG (http://www.genome.jp/kegg/) or BioCyc (http://biocyc.org/), it is a good idea to spend a few minutes and start familiarizing yourself using KEGG's and BioCyc's extensive introductory pages and web tutorials.

=> For our homework problem set, we will focus on BioCyc.

!!!!! Biocyc is not recommended for Internet Explorer (running too slowly with BioCyc's Omics Viewer)

-> in case you haven't done so already, you might want to install alternative browser(s) in addition, e.g. Firefox, Safari, Opera, etc.

- Introduction and overview:
  http://biocyc.org/intro.shtml
- Guided tour: http://biocyc.org/samples.shtml
- Podcasts & webinars: http://biocyc.org/webinar.shtml

2) Play with BioCyc's basic functions, such as searching for enzymes, reactions or pathways, etc..

3) Learn how to use the Cellular Overview Diagram ("Omics Viewer"):
   Change the database to Agrobacterium tumefaciens C58:
After a few seconds (I have seen it take up to 1 minute, depending on how busy the server is), the Cellular Overview Diagram for *Agrobacterium tumefaciens* will be displayed:
=> In the pull-down menu ”Cellular Overview” go to ”Show Legend”; this will give you an explanation what the symbols on the diagram mean.

=> Now slowly mouse over the diagram: Each time you touch a symbol for an enzyme or a metabolite, an explanatory sub-menu will appear. By clicking on symbols, a separate page with information for that compound or enzyme / pathway will open.

=> To highlight a particular pathway:

In the pop-up window, type the pathways name, for our problem set type ”gluconeogenesis”.  
- the pathway will be highlighted  
- click on any enzyme in the pathway; you will be directed to the gluconeogenesis-pathway page in a separate tab-window.
As you might recall from class, **Agrobacterium is missing one gene in this pathway**.

**Q1:** Which conversion would the corresponding enzyme catalyze?
Please name the metabolites.
(You might find a hint to the answer by looking at the little Pathway Evidence Glyph at the bottom of the page).

You become curious and want to find out how Agrobacterium might be able to make up for that missing enzyme's function. You figure you might find a hint in a proteomic data set which you measured under normal physiological conditions (called "AT_pH7.tab"); please download it from the course web page.

=> Return to the Cellular Overview page / tab and choose "Overlay Experimental Data".
Omics Viewer File Data and Parameters

Local file containing experimental data (not a URL):  
AT_ph7.tab

Do you want to display absolute or relative data values? (Absolute)

If displaying relative data values, use:
- a single data column
- the ratio of two data columns

Data values use a:
- 0-centered scale (e.g. log scale)
- 1-centered scale (negative values will be discarded)

The items in the first (zeroth) column of your datafile are (Gene names and/or identifiers)

Note: By selecting Any of the above, you can combine, for example, gene expression and metabolomics data into a single display. There are some dangers inherent in this approach, however. Some names may be ambiguous if it is not known if they refer to genes, proteins or metabolites. In addition, data values from different kinds of experiments may not be directly comparable, so the resulting diagram may be

---

Single Experiment Time Step or Animated Time Series

To display a single experiment time step, enter a single column number in one or both of the column number fields below.

To display an animated time series, enter a list of column numbers (with each column number corresponding to a single timepoint), separated by spaces, a range of integers (e.g., 2-4), or both, in the first column number field below. If you wish to include a denominator column for a ratio calculation, you can enter either a single column number (in which case the same data column will be used as the denominator for all timepoints), or one column number for each numerator column number.

Data column (numerator in ratios): If using two columns, denominator data column:
Note: For column numbering purposes, the first column, which contains the gene name, is column number 0. The first potential data column is column number 1.

**Color Scheme**

Data values are divided into color bins, with the highest value bins displayed in red, the lowest value bins displayed in green or yellow, and the middle bins displayed in blue. Three color scheme options are available:
- By default, the color bins range over the entire spectrum, and the cutoff values for the color bins are derived from the data itself. This means that different experiments could be displayed using different color schemes, making it difficult to directly compare them.
- Alternatively, you may specify a value for the maximum value cutoff bin. All displays that use the same maximum value cutoff will use the same color scheme (assuming other settings, such as relative vs absolute, or log format vs non-log format are the same), and are therefore directly comparable. The maximum cutoff value should be a number, e.g. 2 or 10, etc. All data values greater than the maximum cutoff value will be displayed in red.
- The third alternative is to use only three color bins: red for data values that exceed some threshold, yellow for data values less than the inverse of that threshold, and blue for values in between. The threshold value should be a number, e.g. 2 or 10.

Choose a color scheme:
- Full color spectrum, computed from data provided (default)
- Full color spectrum with a maximum cutoff: [ ]
- Three color display with specified threshold: [ ]

Submit Note that this request might take more than 30 seconds to complete. If your browser warns you about an unresponsive script in a popup window, please let the script complete by clicking the "Continue" button.
You should be getting something looking like this. The color coding in the little control panel on the lower right hand tells you to what level of abundance a particular protein was found in the sample, i.e. the more reddish an enzyme is colored, the more abundant. One example of such a highly abundant pathway is the Pentose-Phosphate Pathway. Try to localize it of the map using the highlight function as before. Which are the two most abundant enzymes in this pathway?

Q2: Identify 3 other enzymes and their corresponding pathways which are represented as highly abundant (i.e. at least pink color) in this sample.
   => Please do NOT list multiple versions of the same pathway
   => Do not include the enzymes of the Pentose-Phosphate Pathway.

Now go back to the Pentose Phosphate Pathway on this diagram:
Try to find the enzyme “Transaldolase” (TalB) -> click on it.
Information about the TalB-catzlyzed reaction will be displayed in a separate window-tab.
Q3: What metabolites are taking part in the conversion catalyzed by TalB? (No structures, just names).

- Is the reaction reversible?

Q4: Please explain schematically how *Agrobacterium* might make use of the TalB-catalyzed reaction within the Pentose-Phosphate-Pathway to make up for its missing enzyme in gluconeogenesis. (Are there any common metabolites?)

Now let's look at this pathway and the TalB protein a bit more closely:
To view the whole pathway, click on the pathway name above the reaction schematic. You should now see a pathway scheme.
On the top of the page, click on "species comparison".

=> On the Organism selection page, choose
- *Agrobacterium tumefaciens* C58
- *Mycobacterium tuberculosis* CDC 1551
- *Homo sapiens*
- *Plasmodium falciparum* 3D7

=> then press submit (at the very bottom of that page)

Q5: Which of these organisms are lacking one or more of the enzymes UNIQUE to this pathway?
(Hint: Look at the little pathway evidence glyphs and their color key at the bottom of the page).

<table>
<thead>
<tr>
<th>Organism</th>
<th>Evidence Glyph</th>
<th>Enzymes and Genes for pentose phosphate pathway (nor)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. tumefaciens</em> C58</td>
<td>![Glyph]</td>
<td>EC# 5.3.1.6 ribose 5-phosphate isomerase: rpiA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EC# 5.1.3.1 ribulose-phosphate 3-epimerase: rpe</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EC# 2.2.1.1 transketolase: tktB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transketolase: tkt</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transketolase: tktA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EC# 2.2.1.2 transaldolase: talB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EC# 2.2.1.1 transketolase: tktB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transketolase: tkt</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transketolase: tktA</td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>![Glyph]</td>
<td>EC# 5.3.1.6 Ribose-5-phosphate isomerase: RPIA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EC# 5.1.3.1 Ribulose-phosphate 3-epimerase: RPE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EC# 2.2.1.1 Transketolase-like protein 1: TKTL1</td>
</tr>
</tbody>
</table>