Problem set 6

This problem set covers the content from week 8: proteomics.

Tips and rules:

- You can answer in English or in Thai.
- There can be more than one correct answer. What I am looking for from you is not just the correct answer but the rationale for your answer.
- Please provide evidence of how you think and what sources of information you used.
- AI such as ChatGPT may be used. You can also work together with friends. But you must write the answer in your own words.
- Any incidence of plagiarism and copying of another student's work will be reported to the Graduate Affairs.

Proteomics

Explore the proteomics dataset **PXD000674** deposited on the PRIDE repository and extract the following details for **Q1**-**Q6**:

Q1: Is this a top-down or bottom-up proteomics experiment? If it is a bottom-up experiment, which enzyme was used to digest the proteins? If not, answer N/A.	
Q2: Is there a step to prevent Cysteine from forming di-sulfide bonds? If yes, what is the modification of Cysteine?	
Q3: How long was the liquid chromatography run?	
Q4: Which mass spectrometer was used? Which mass analyzer was used?	
Q5: Were the mass spectra data acquired in data-independent or data-dependent mode?	
Q6: Is this a label-free or labeled comparative proteomics experiment?	

Use the following scenario to answer **Q7-Q10**. You want to discover molecular biomarkers for diagnosing a certain kidney disease and you are deciding between whether to use transcriptomics (RNA-sequencing) or proteomics (mass spectrometry) to study this disease.

Factors or scenarios	**Is transcriptomics preferred? Why?**	**Is proteomics preferred? Why?**
Q7: You want plasma biomarkers to enable blood-based assay		
Q8: You have limited tissue sample from each patient		
Q9: The disease is caused by post-transcriptional regulation		

When studying plasma proteomics, it is highly recommended that you deplete abundant proteins in blood (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2948641/).

Q10: Explain why this process is helpful for the proteomics study.

Q11: Is there a similar consideration in transcriptomics?

Q12: Select a paper that used top-down proteomics and explain why bottom-up proteomics cannot address that paper's research question.

Q13: Database search is the technique for decoding the amino acid sequence of a mass spectrum by comparing the m/z profiles of the spectrum to either a database of theoretical m/z from known peptides (sequence database) or a database of annotated mass spectra (spectral library). What's the pros and cons of using sequence database or spectral library?

	Sequence database	Spectral library
Pros		
Cons		

Q14: If you want to study the proteome of a novel species, such as a Thai herbal plant, what information you do need to prepare in order to analyze the mass spectrometry data with a database search approach?

Q15: In order to minimize technical variation, multiple proteome samples can be analyzed in a single mass spectrometry analysis by uniquely labeling the samples (SILAC or TMT). However, as there are only a handful of different labels available, you cannot analyze >10 samples at once. Let's say you have 30 patient samples (which have to be analyzed in at least 3 batches). How would you design your mass spectrometry runs to ensure that the batch effect can be removed from the resulting data?

[Extra] Q16: You developed a new chemical drug X whose preliminary trial indicated that it can considerably slow down the progression of the Alzheimer's disease.

However, your chemist friend warned you that under certain laboratory conditions X can produce a derivative compound Z that is very toxic to human cells. You have confirmed in your lab that Z can indeed be produced from X under those suggested conditions.

In theory, these conditions should not occur in the human body, but you cannot risk it. So, you want to determine whether Z is produced in any cell treated with X.

You gather hundreds of cell lines, treat them with X, and analyze the cell lysate with mass spectrometry. However, you suddenly realize that there are no known mass spectra for the compound Z, which is needed to interpret the data. Unfortunately, there is also no AI for *de novo* interpretation of chemical compound.

What will you do?