Questions

Introduction to Instrumentation - 15 questions

Chromatography - 5 questions -> split LC and GC parts

1. What are the basic principles of separation in liquid chromatography?
    1. Compounds have different affinity/interactions with the stationaty phase and mobile phase -> Polarity
    2. Physicochemical properties of the molecule
2. What are the principles of separation in gas chromatography?
    1. Boiling point of the chemical and therefore its vapour pressure
3. Sort compounds after polarity -> see lecture slides + period table with electronegativity
    1. Polarity depends on presence of electronegative atoms and corresponding bonds with partial charges.
4. What is the mobile phase (MP) in liquid chromatography (LC)?
    1. It carries the sample through the column.
5. In what form do we meet stationary phase (SP) in LC?
    1. Packed columns -> stationary phase is in small particles packed into the column
    2. Monolithic columns -> stationary phase as coating on the inside of the column
    3. What are the differences between these? -> packed column provides more space for molecular interactions
6. What is gas chromatography (GC) and how does it differ in principle from LC? -> split this question into smaller parts
    1. Mobile phase is a gas
    2. Separation is based more on boiling point than polarity.
    3. Capillary column vs. packed or monolithic column. GC capillary columns can be multiple meters (10 - 30) while LC packed or monolithic columns are very short.
    4. Oven to make our sample volatile.
7. How does the MP work in GC and what are the known examples of these MP?
    1. Mobile phase is an inert gas which carries the sample.
    2. Commonly used is Nitrogen, Hydrogen or Helium.

Mass Spectrometry - 5 questions

1. Briefly explain the fundamental principles of mass spectrometry. -> Slide 19 & 20
    1. Detection of charged molecules in gaseous state - based on ionization in the ion source, separation in the mass analyzer and quantification of ion abundance in the detector.
2. List the main parts of the mass spectrometer. -> Slide 22
    1. Ion source, analyzer and the detector.
3. What is the difference between soft and hard ionization? -> Slide 23
    1. Soft ionization -> molecular ion is being kept intact
    2. Hard ionization -> molecule fragments and we only detect the fragments of the original compound
4. What is the purpose of the mass analyzer? -> Slide 24
    1. To separate the ions based on their mz value for detection.

5. Give some examples of the use of mass spectrometry.
    1. Measuring of biomolecules.
    2. Detection of specific substances for example for doping, forensics or drug tests etc.
    3. In geology/archaeology to determine the age of rocks based on the C13 isotope abundances.
6. What is the molecular ion and base peak in the mass spectrum? -> Chapter 4 Slide 13
    1. Molecular ion -> peak with the mz value representing the compound of interest
    2. Base peak -> peak of highest abundance in the mass spectrum

Acquisition Methods - 5 questions

1. What is data independent acquisition (DIA)? -> Slide 39
    1. There is no known connection between the precursor ion and product ion -> all precursors get fragmented
2. What is data dependent acquisition (DDA)? -> Slide 39
    1. Precursor ions are fragmented selectively and it is known which product ion belongs to which precursor
3. What is full scan acquisition on MS1/MS2? Chapter 3, Slide 23
    1. Acquisition of all ions in specified mz range.
4. What is the goal of selective acquisition? -> Chapter 3, Slide 22
    1. Acquisition of specific ions for quantitation to have more accurate data.
5. How is data acquired from selective acquisition different from data acquired in full scan? -> Chapter 3, Slide 21
    1. Signal is only present at certain mz values which have been monitored.

Introduction to Omics - 6 questions

1. What is the central dogma of molecular biology?
    1. Genome = all inherited genetic material
    2. Transcriptome = all the transcribed genetic material i.e. all expressed genes
    3. Proteome = all the proteins/peptides of an organism i.e. all translated genetic material
2. What is "-omics"?
    1. Systematic and comprehensive analysis of a cellular/ molecular layer
    2. Genomics = study of genome, Transcriptomics = study of transcriptome, Preoteomics = study of the proteome, Metabolomics = study of the Metabolome
3. Explain the term "phenotype".
    1. Phenotype = characteristics [traits] displayed by an entity [organism]
4. What is the main difference in application between mass spectrometry and sequencing technology? -> Slide 25 -> split into technical differences and what can be measured differences
    1. Sequencing is more robust than MS.
    2. Sequencing is cheaper than MS.
    3. Sequencing shows all potential biological reactions that can occur.
    4. MS methods provide finer details about phenotype due to greater inclusion of external factors.
    5. MS shows biological and non-biological reactions that occur.
5. Briefly describe the difference between targeted, suspect screening and non-targeted analysis. -> Slide 21
    1. Non-target -> as generic sample prepearation as possible, full scan acquisition and non-target data processing with as little bias as possible.

    2. suspect screening -> non-target sample acquisition and targeted data processing for a specific group of compounds of interest
    3. targeted analysis -> specific sample preparation , selective acquisition and targeted data processing.
6. Explain the benefit of adding chromatographic separation prior to MS.
    1. More information for chemical characterization, retention time provides orthogonal information for otherwise potentially indistinguishable molecules.

Introduction to MS Data Processing - 10 questions

1. List three (3) factors related to the instrumental platform which influence the acquired data.
    1. Mass analyzer - the mass analyzer determines the resolution, which is directly visible in the data.
    2. Tandem MS - Data coming from a tandem mass spectrometer which acquires MS1 and MS2 data looks inherently different, as the data contains multiple layers.
    3. Acquisition range - the acquisition range specified on the instrument determines the m/z range for which ions are detected. The data is therefore limited to that range.
    4. Scan speed - the scan speed (in Hertz) determines the number of scans per second. A higher scan speed results in more data to be recorded over the same time compared to a lower scan speed.
    5. Ion source - the ion source determines the polarity of the data (positive or negative) and whether we will observe extensive fragmentation (hard ionization) or not.
    6. Chromatography - depending on the chromatography (gas or liquid) we will observe different background noise.
2. List three (3) factors related to the sample preparation which influence the data processing.
    1. Timing - depending on whether the samples have been prepared all in one session or in multiple batches, we might have to adjust the normalization method to normalize first all samples which have been prepared together and then to normalize based on the sample groups.
    2. Solvents and other substances - depending on the solvent and other chemical substances used during sample preparation with LC-MS, we may have to adjust the adduct table for annotation, as the substances used during sample processing might interact with our sample.
    3. People - depending on who prepared the samples (either all of them or subsets, or whether this was fully automated), we may have to adjust normalization to normalize all samples prepared by the same person in a first stage and then to normalize across people involved in a second stage.
    4. Sample cleaning - depending on how the sample was cleaned during preparation, we want to limit the chemical space we consider during annotation.
    5. Normalization - if samples have been "normalized" (diluted or concentrated etc.) during sample preparation, we might not need to perform any computational normalization.
3. What is an extracted ion chromatogram?
    1. The extracted ion chromatogram is the intensity at a specific m/z value given a certain tolerance over the whole duration of the chromatogram,
4. What is the total ion chromatogram?
    1. The total ion chromatogram is a function mapping from time to the intensity integrated over the entire m/z range at the specified time point.
5. What is the difference between centroid and profile mode data?
    1. Centroid data is more compact and expressed at a single m/z value while profile data is raw and not preprocessed and expressed continously over all measured mz values. Centroid data is more compact but loses information about the peak shape in m/z domain.

6. What types of noise can we distinguish in the data?
    1. Random noise occurs as the name says, random, and it also not reproducible or deterministic. There is no specific single source for random noise other than randomly occurring errors during data acquisition or transfer.
    2. Non-random noise is signal that doesn't represent a compound of interest. This includes background signal originating through eg. column bleed or solvent or other chemical agents added to the sample or originating from the instrumentation.
7. What are the benefits of open data formats?
    1. Open data formats allow us to process data coming from different instruments with the same software. One benefit is that users are not tied to one vendor specific software, another is that data from different vendors can eventually be compared. Open formats overall contribute to the FAIRification of research data, by making it interoperable and accessible.
8. What are the benefits of normalizing data?
    1. By normalizing data, we remove intensity deviation which are not related to differences in the sample. Intensity drift occurs naturally in the instrument over time and can be removed by normalizing the data using the drift of some background ions.
9. Explain the term "resolution" for MS data.
    1. Resolution describes how close the m/z values of two ions can be so that they can still be distinguished. The definitions of resolution are related to either the overlap between the adjacent peaks related to the difference in m/z or the peak width at half maximum at a specific m/z value.

Feature Extraction - 10 questions

1. Why do we need some form of peak detection? -> Slide 8
    1. labeling/clustering of the data
    2. distinguish signal from instrumental noise
    3. quantification of signal
    4. individual raw data points don't represent the actual "ion" as a feature = grouping of primitive features into "higher-level! features
2. Briefly describe one method for peak picking is MS data. -> Slides 10, 11 & 12
    1. XCMS -> based on continuous wavelet transform for apex detection
    2. SAFD -> rule based algorithm which works of the highest intensity point and iteratively fits 3D gaussians to the data, subtracts the points from the datatset and then repeats the procedure.
    3. GridMass -> Initialize a uniform grid of probes and then update their position by letting them gravitate towards the local maximum in a defined window around each probe.
3. How do we map from raw point intensities to feature intensity?
    1. For accurate peak integration, we can use probability distributions or mathematical functions to model the peak shape and then optimize the parameters to fit the data points assigned to this feature. We can then integrate the function to obtain the area under the curve.
4. Is profile or centroid data better for quantification?
    1. Profile mode data is better for quantification as it contains all data points recorded by the instrument. If the data is already centroided, the centroiding method is the main contributing factor for the quantification, and these methods are often not well described.
5. Which factors related to the peak integration method determine accurate feature intensity estimation?
    1. peak shape model - The function chosen to model the peak is the main factor determining the accurate quantification and is specific to the instrument.

    2. parameter optmization method - How the numerical parameters for the model function are optimized also influences the feature intensity and its accuracy.
6. List 2 properties of features which can be used for feature alignment?
    1. m/z - The m/z ratios of two features can be used to determine whether these two peaks might originate from the same compound.
    2. retention time / index - this information can also be used for alignment given that the two samples were acquired using the same chromatographic method and drifts in retention time have been corrected.
7. What are reasons for retention time drift in LC/GC-MS data?
    1. LC-MS -> The chromatographic column degrades with time which changes the separation properties. Another reason for retention time drift can be fluctuations in the pressure or other interactions of compounds within the sample.
    2. GC-MS -> While GC separation is very stable, the length of the column influences the retention time. As the column degrades mostly at the end which is connected to the mass spectrometer, the column needs to be shortened at times. This leads to shifts in retention time for all compounds.
8. How can small retention time drifts in LC-MS and GC-MS data be corrected?
    1. We can correct retention time drift by identifying corresponding features across samples and then creating an alignment. This can be done for example by choosing one sample as the reference and then mapping the other samples onto the reference.
9. How can large shifts in retention time occurring in large GC-MS based studies be corrected?
    1. Shifts in retention time in GC-MS data which occur after changes to the chromatographic column can be corrected using the retention index system. The shifts which are observed on a reference set of compounds (e.g. Alkanes) can be used to shift all other compounds based on the closes reference compounds in the chromatogram.
10. What motivates the comparison of features across samples?
    1. By comparing the intensities of corresponding features across samples, we can identify statistical correlations between features and other data, such as clinical information about the health of an individual. It therefore allows to identify which features are related to a certain condition.
11. What is the difference between weaker signal recovery and missing value imputation?
    1. Weaker signal recovery - reduces false negatives by recovering wrongly discarded signal is based on individual samples doesn't necessarily fill all gaps
    2. Missing value imputation - makes data compatible with statistical methods and tools is based on all samples all gaps are filled

Spectra Reconstruction - 7 questions

1. Explain the term "adduct".
    1. Adducts are functional groups which attach or detach from the original molecule. They often originate as reaction products of interactions with the solvent or other chemicals during sample preparation or ionization.
2. Explain the term "isotopic pattern".
    1. Atoms have multiple stable configurations with varying number of neutrons. The abundances of the stable isotopes relative to one another are constant for a specific atom type. Therefore, a compound with a specific chemical formula occurs with a fixed distribution of stable versions with different masses, depending on the atomic composition. We can observe all isotopic configurations of the compound in the mass spectrum.

3. Explain the term "multimer".
    1. Multimers are molecules made up of repeating identical subunits - therefore, molecules which tend to react with itself can form dimers and trimers which we can then detect.
4. What are the reasons for fragmentation in hard-ionization and tandem MS?
    1. Transfer of energy to the molecule which causes transfer to the excited state and eventual loss of an electron. This leads to an unfavourable energy configuration and the molecule fragments to achieve an energetically preferable configuration.
5. List 3 factors which contribute to observing multiple peaks for a single compound in MS data.
    1. Adducts
    2. Multimers
    3. Isotopes
    4. Fragmentation
6. What are the benefits and pitfalls of experiment wise deconvolution?
    1. Beneftis -> detection of low abundant features as long as they are consistent across the experiment. Pitfalls -> peaks which occur sporadically across samples are not included
7. What are the benefits and pitfalls of sample-by-sample deconvolution?
    1. Benefits -> each sample is treated individually, so we have less side effects from processing steps like alignment etc., deconvolution can be based on peak-shape correlation. Pitfalls -> low-abundant peaks might get filtered out most often as they might be below baseline or the signal-to-noise threshold.

Annotation - 12 questions

Methods - 6 questions

1. Which sources of information can be used for annotation?
    1. Chromatographic information - retention time and elution order (LC), retention index (GC)
    2. Spectrometric information - MS1 accurate mass, MS1 isotopic distribution, MS2 spectral matching, MS2 structural similarity, MS2 in silico fragmentation
    3. Non-analytical information - biotransformations, metabolic pathways
2. What is the difference between "annotation" and "identification?
    1. Annotation - annotation of a chemical with a "putative identity" = speculative
    2. Identification - comparison of the spectrum and other information with those of an analytical standard purchased and analyzed on the same instrumental platform or structure elucidation using other techniques, e.g. NMR = as conclusive as technically possible
3. Explain the "retention index" system for GC-MS data.
    1. The "retention index" system is based on the reproducibility of GC (and partially LC) separation and the use of common reference standards. For GC, the Kovats retention index system is based on recording the retention times of Alkanes and using those as references. Compounds are assigned the retention index by linear interpolation of the closest (with regard to retention time) Alkanes.
4. How can we use retention time information for annotation of LC-MS data?
    1. We can train a model to predict retention times based on analytical standards recorded on the instrument. The model can then predict the retention times of previously unseen compounds. We can then compare those predicted retention times to the observed retention time of putative annotations and exclude annotations which don't fit the predicted retention time.
5. Explain the steps involved in calculating the similarity between 2 mass spectra.

    1. Vectorization - spectral similarity functions operate on discrete vectors of same length, therefore the spectra have to be vectorized. This step can be done by "binning" the spectra based on a fixed resolution or by pairing the peaks of both spectra. If there is no corresponding peak in the other spectrum, the value gets filled as 0. The "frame of reference" can be either one of the spectra, or the union of peaks in both spectra or their intersection.
    2. Spectral similarity function - here, any function which takes two vectors of identical length can theoretically be used, though a good spectral similarity metric will perform close to structural similarity and will provide the maximum values for comparing two identical spectra.
6. Why is the resolution of MS data crucial for in silico annotation?
    1. The resolution is crucial as it limits the number of potentially correct chemical formulas to fit an m/z value. The higher the resolution, the more potential formulas can be excluded.

Resources - 6 questions

1. Why are mass spectral & compound databases important?
    1. Databases allow us to collect various information related to chemicals in a shared place and create links between individual pieces of information. They can also provide reference information with which we can compare the information obtained from our experiment. In practice, this means using the calculated m/z values of compounds in databases for annotation of high-resolution LC-MS data or by comparing the mass spectra recorded in GC-MS experiments with those spectra stored in reference mass spectral libraries.
2. Why are chemical identifiers important for chemical data processing?
    1. We need to be able to represent chemicals in a way that it is easy for a computer to interpret, parse and eventually link to other information (if we want to search for additional information). We also need identifiers to clearly communicate in our results which chemical substances we are actually talking about.
3. Which properties should an ideal chemical identifier have?
    1. Unique
    2. Machine readable
    3. Human readable
    4. Derived by exact structure and geometry
4. Write down the SMILES code for compound with the depicted structure.
    1. 2_6-Dichloro-4-nitroaniline - NC1=C(Cl)C=C(C=C1Cl)N+=O
    2. Coumarin - C1=CC=C2C(=C1)C=CC(=O)O2,
5. Which molecule is depicted by the following SMILES string?
    1. C(C(C(=O)O)O)C(=O)O - Malic Acid
    2. CC(C(=O)O)N - Alanine
6. What is the InChI identifier and what are its advantages and disadvantages?
    1. The International Chemical Identifier (InChI) is a textual identifier for chemical substances, designed to provide a standard way to encode molecular information and to facilitate the search for such information in databases and on the web. An advantage of the InChI is that it is directly related to chemical structure and uniquely encodes the chemical substance. Another advantage is that it is defined by the IUPAC, so it is universally accepted and supported. A disadvantage is that since it depicts chemical structure, it is not human readable and cumbersome to use in communication meant for humans.
7. What are potential problems when using "community curated" databases?
    1. information might be incomplete

    2. information might be wring
    3. lack of standards or changes thereof
8. Which aspects have to be considered when matching spectra with a mass spectral library?
    1. resolution
    2. acquisition range
    3. ionization method