**Introduction into data analysis for MS-based proteomics *https://www.youtube.com/watch?v=ZgwNWRul98o***

* Protein introduction
  + 1°, 2°, 3° structure, modifications (can be taken on or off), processing (e.g. trypsin, platelet activity)
  + MS studies 🡪 mostly 1° structure & modifications & processing
* AAs, peptides, proteins
  + AAs
    - 20 common AAs
    - C- & N- terminus
    - ≠ groups: nonpolar (G, A, V, L, I, M); aromatic (F, Y, W); polar (S, T, C, P, N, Q); +-charged (K, R, H); --charged (D, E)
    - Molecular weight
    - pKa & pKb 🡪 at pH = 7; C-terminus will be -, N-terminus will be + (Zwitter-ion)
    - pKx 🡪 only in charged AAs (RMK: Y has also a pKx)
    - pI 🡪 iso-electric point 🡪 pH at which net charge of AA = 0
  + Protein backbones
    - Peptide bonds (C- & N-terminus bind, H2O = ejected)
      * Quite weak bond 🡪 easy in- & unplugging
* MS basics
  + Build
    - Sample
    - Mass spectrometer
      * Ion source 🡪 Ionizes sample molecules + bringing these in gas phase
        + MALDI (matrix assisted laser desorption & ionization)

Analyte = put on metal plate (target), matrix molecules = added (acidic molecules, good at absorbing UV light), in vacuum

UV laser 🡪 onto target 🡪 matrix molecule absorbs UV 🡪 matrix molecules & analyte = blown of plate (desorption)

Ionization 🡪 matrix molecules transfers proton to analyte molecules

* + - * + ESI (electrospray ionization)

Electric field 🡪 between m/z analyzer inlet to needle (3-5 kV)

Sample = pushed through needle (buffer flow with N2, inert) 🡪 nebulisation (droplet forming)

Some droplets = neutral, some = + (formic acid = added as proton donor)

Neutral droplets don’t feel electrical field 🡪 will fly straight forward and hit barrier (no effect)

+ droplets respond to electrical field 🡪 start turning 🡪 charge-driven fission (+ droplets will repel + charges 🡪 split into 2 new droplets) or ion expulsion (if droplet becomes to small 🡪 eject charge)

More popular than MALDI

Sample input = easier, automatable ↔ MALDI 🡪 replacing plate

Can be coupled to LC 🡪 creates continuous flow on sample

* + - * Mass analayzer 🡪 Uses EM field 🡪 manipulate gas-phase ions 🡪 plotted on spectrum 🡪 x = m/z & y = ion intensity
        + Mass resolution = important for identification & quantification

Width of peak at given height, mostly 50% (FWHM)

* + - * + Variance on measurement wanted as small as possible
        + Distance between 2 peaks

If double charged; 0.5, if triple charged; 0.33, if quadruple charged; 0.25

Distance between peaks can tell what original charge was

* + - * + Info about mass & charge = given with peak spectra
        + RMK: big peptide with many C’s 🡪 chances of finding a peptide with C13 > finding peptide with no C13 at all
      * Detector 🡪 Records presence of ions
        + Electron multiplier system 🡪 e- impact 🡪 sets other e-‘s loose 🡪 impact 🡪 … 🡪 cascade
        + Not all peptides ionise equally 🡪 signal strength cannot be compared across peptides

≠ peptides may not be present in = amounts in protein or may not have the same charge in the original protein

Better to compare peptide against itself!

* + - * + Detector only gives signal when certain threshold = reached

After that linear respons 🡪 until saturation (detector cannot output more however more input = given)

* + - * + Detectors always follow sigmoid curve

For least variance, you want to be in the middle part of the curve

* + - Digitizer
      * Transforms continuous, analog signal into digital, discretized spectrum
      * Can be upgraded to get better results
  + Raw data = imprecise, expected errors around 10%
    - Technical errors due to fitting curve
  + Peak ≠ only m/z, also time factor can be taken into account
    - Volumes of peaks need to be calculated
    - Less abundant isotopes will show up later
      * Harder to overcome threshold & disappear more quicky into noise
  + Serum proteins = degraded over time
    - E.g. samples from blood 🡪 proteases in blood 🡪 degrades proteins
    - A lot of variability based on when proteins in blood are measured
    - PTMs 🡪 signal will shift due to higher/lower molecular weight
      * Can occur due to sample preparation, or biologically
* MS/MS spectra & identification
  + Breaking peptides into smaller bits 🡪 fragmentation (2nd run of MS 🡪 tandem-MS)
  + Build
    - Source
    - Ion selector
      * 1st analyzer 🡪 ion selector 🡪 only ions of given m/z may pass
    - Fragmentation
      * Peptide can be fragmented in ≠ ways
        + Break close to N-terminus

a1 (C-CO), b1 (CO-N), c1 (N-C) 🡪 1 because it contains 1 AA

* + - * + Break close to C-terminus

z1 (C-N), y1 (N-CO), x1 (CO-C)🡪 1 because it contains 1 AA

* + - * Mostly b1 & y1 ions = present
        + Weakest bond (peptide bond, CO-N/N-CO) = broken
    - Fragment analyzer
      * 2nd analyzer 🡪 after fragmentation = triggered 🡪 mass analyzer for fragments
        + Dependant where fragmentation is, the molecular waight of the fragment ≠
        + ∆mass between 2 fragments can reveal which AA differs
        + BUT…

Fragments from the other side (z1, y1 & x1) are present!

Also some AAs have = mass

Ambiguity in interpretation

Whole range of possible sequences can be extracted from MS spectra

* + - Detector
* Database search algorithms in 3 phases
  + Experimental spectra = matched to known peptide sequences
  + Only allowing amino acids which were found in the knows prot. seq.
    - Creating theoretical spectra based on peptide seq.s
  + Take experimental spectra & compare
  + Calculate score for each match
  + Peptide with highest score = best
  + 3 popular algorithms illustrate 3 types of scoring systems
    - Sequest
      * Taking all the peaks 1 by 1 & multiply intensity with peak from experimental spectra & peak from database spectra given there is a match (if no match = found 🡪 multiply with 0)
      * Sum of matched intensities
      * Create some “nonsense”/randomnesss to compare spectra with 🡪 Ri with I the value you move the peaks with (+ to right, - to left)
        + Result = histogram

take average of distribution

calculate distance between this average and R0 score (theoretical spectra)

* + - * + BUT.. we have more than 1 theoretical spectra!

Compairing against ≠ theoretical spectra

Using ∆Cn 🡪 to calculate ≠ between best spectra & runner op scores

High ambiguous 🡪 low score

Clear 🡪 high score

* + - * +1, +2 or +3 spectra 🡪 the higher you go, the more the theoretical spectra rise up compared to the actual spectra
        + +2 spectra

fragment

y1 & b1 will often have the +2 charge split between the 2 ions

Noise level = risen

* + - * + Algorithms may not be powerful enough anymore
    - Mascot
      * Based on peak counting instead of intensity sums
      * Theoretical function that predicts having a random match without a score
    - X!Tandem
      * Hybrid score
      * Based on peak counting & ion intensity
        + Intensity \* probability (= 0 (no match) or 1 (match))
        + Probabilistic model 🡪 model of matched ions
      * Either 1 or 0 real hits
      * Hyperscore 🡪 uses regression
        + 1 theoretical spectrum = used
        + Using all other spectra to calculate how special the theoretical spectrum is
      * What happens is database grow too big/small?
        + Too big 🡪 regression will expand and “eat up” theoretical spectrum
        + Too small 🡪 curve will collapse 🡪 regression line goes flat
* Sequential search algorithms
  + Sequence tags
    - but databases became too big + not using all peaks
    - “lost approach”
  + GutenTag, DirecTag, TagRecon
    - Generate list of seq. tags
    - Search in DB for matches
    - Coupling back to original spectra
    - Score DB sequences
  + De novo sequencing
    - Tries to read the entire peptide sequence from the spectrum
* Decoys & false discovery rate calculation
  + How many mistakes can be controlled?
  + All hits (good & bad) together, form a distribution score
    - BUT… where/how to distinguish good from bad scores?
      * With empirical null distributions!
        + Place threshold e.g. on 5% (only 5% bad results = accepted)
  + Limiting space of where negative results could be
    - Calculate ratio of good VS bad on every score step
  + Local false discovery rate
    - Probability of finding a false positive at every score step
  + 3 main types of decoy DB’s
    - Flipping the sequence around
      * Patterns remain
      * Reproducible
    - Shuffling sequences
      * Patterns = destroyed
      * Not reproducible
  + Now if you compare target hits to decoy hits
    - Decoy hits start going down at the right end of the curve
  + RMK: of course there will be false negatives/positives
    - Usually false positives are tried to be limited
* Protein inference: bad, ugly, and not so good
  + Occam set 🡪 only take proteins that explain all the peptides in sample
  + Anti-Occam set 🡪 include every protein that has at least 1 peptide from sample
  + Minimal set with maximal annotation 🡪 eliminate hypothetical evidence in favor of real known solid entities (fewest possible hypothetical entities)
    - List of proteins in al 3 scenarios ≠
  + Best scanarios 🡪 many peptides that can be linked to 1 protein (but often many peptides can be linked to ≠ proteins)