NANYANG
TECHNOLOGICAL
UNIVERSITY

Principles of proteomics algorithms

Wilson Wen Bin Goh
School of Biological Sciences, Nanyang Technological University

16 March 2017

Learning objectives

- Describe what an algorithm is, and how it differs from a heuristic or a computer program
- Describe the various levels of spectra data and their derivations in MS-based proteomics
- Describe the steps of a library search algorithm
- Describe the steps of a *de novo* sequencing algorithm
- Describe how peptides are assembled to proteins and associated problems
- Describe and evaluate the various levels and data representation formats in proteomics

Algorithms

- https://www.youtube.com/watch?v=6hfOvs8pY1k

What is an algorithm?

- An algorithm is an ordered series of steps for solving the problem
- It is very exact and unambiguous
- It can be expressed by a programming language (where it becomes a program)
- Can also be expressed in semi-human readable form (in pseudocode)

What is an algorithm? (Cont'd

- Example: an "algorithm" to come to this lecture
  1. Walk to bus stop
  2. Catch bus
  3. Get off near Uni
  4. Walk to lecture theatre
- Is this enough as directions to give someone?
- What about: (supposing you live near Prime)
  1. Walk to bus stop in Nanyang Drive, in front of Prime
  2. Catch bus number 1 or 2
  3. Get off at bus stop code 123456
  4. Walk up the first staircase you see and turn to the right on the second floor. Enter LT 25

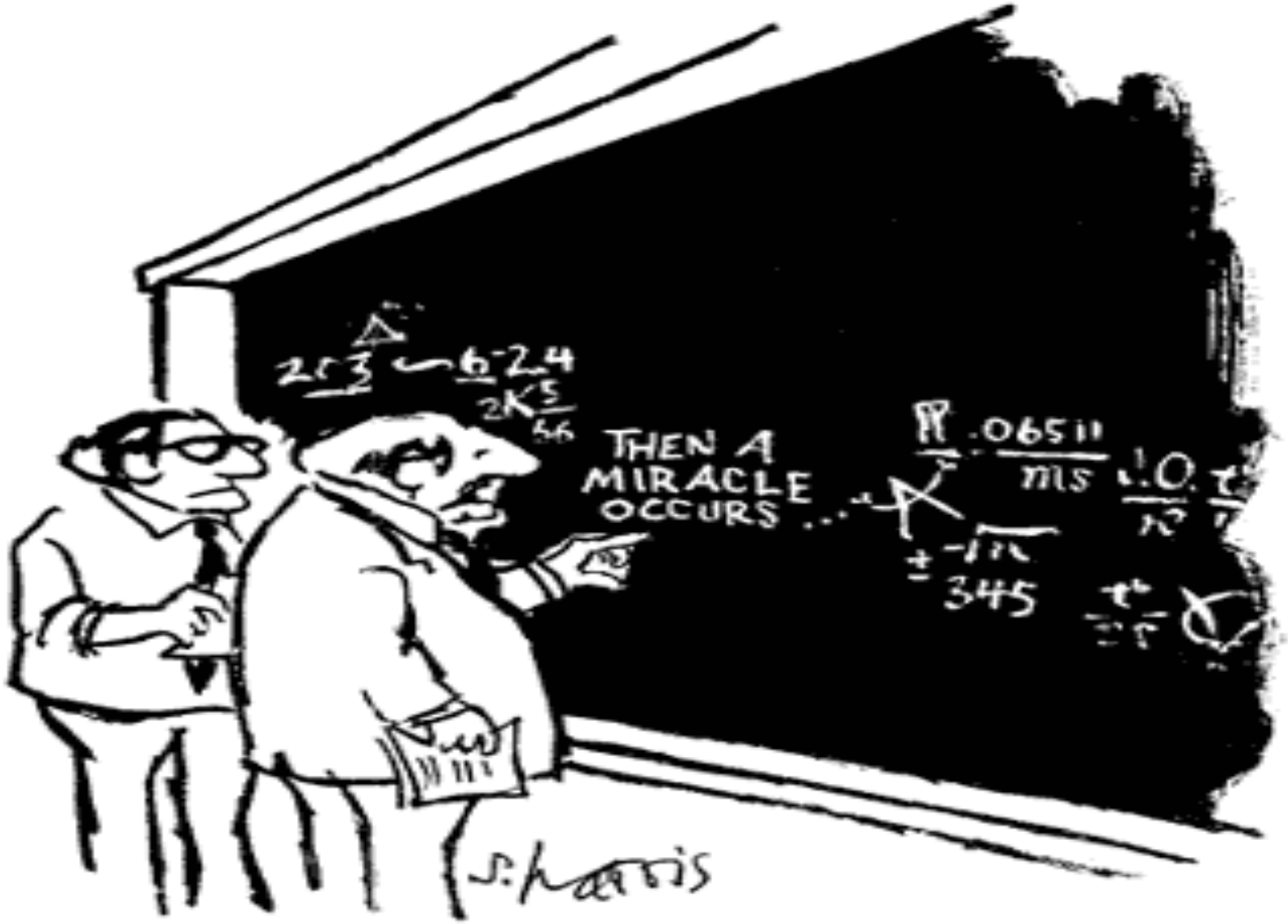
See an algorithm in action --- The bubble sort algorithm

- Objective is to sort objects using an iterative window.
- https://www.youtube.com/watch?v=lUqFelc84XE

How detailed should an algorithm be?

- First answer: “there should be enough detail”
  - How “enough” is “enough”?
  - Depends on who will read the algorithm (audience)
  - Depends on what the algorithm is for (problem to solve)
  - Depends on why someone will read algorithm (purpose)
- Second answer: “it should have enough detail so as to allow someone to
  - Understand each and every step
  - Follow the algorithm (manually) to work out a solution
  - Implement it, adapt it, extend it, embed it,…”

How detailed should an algorithm be?



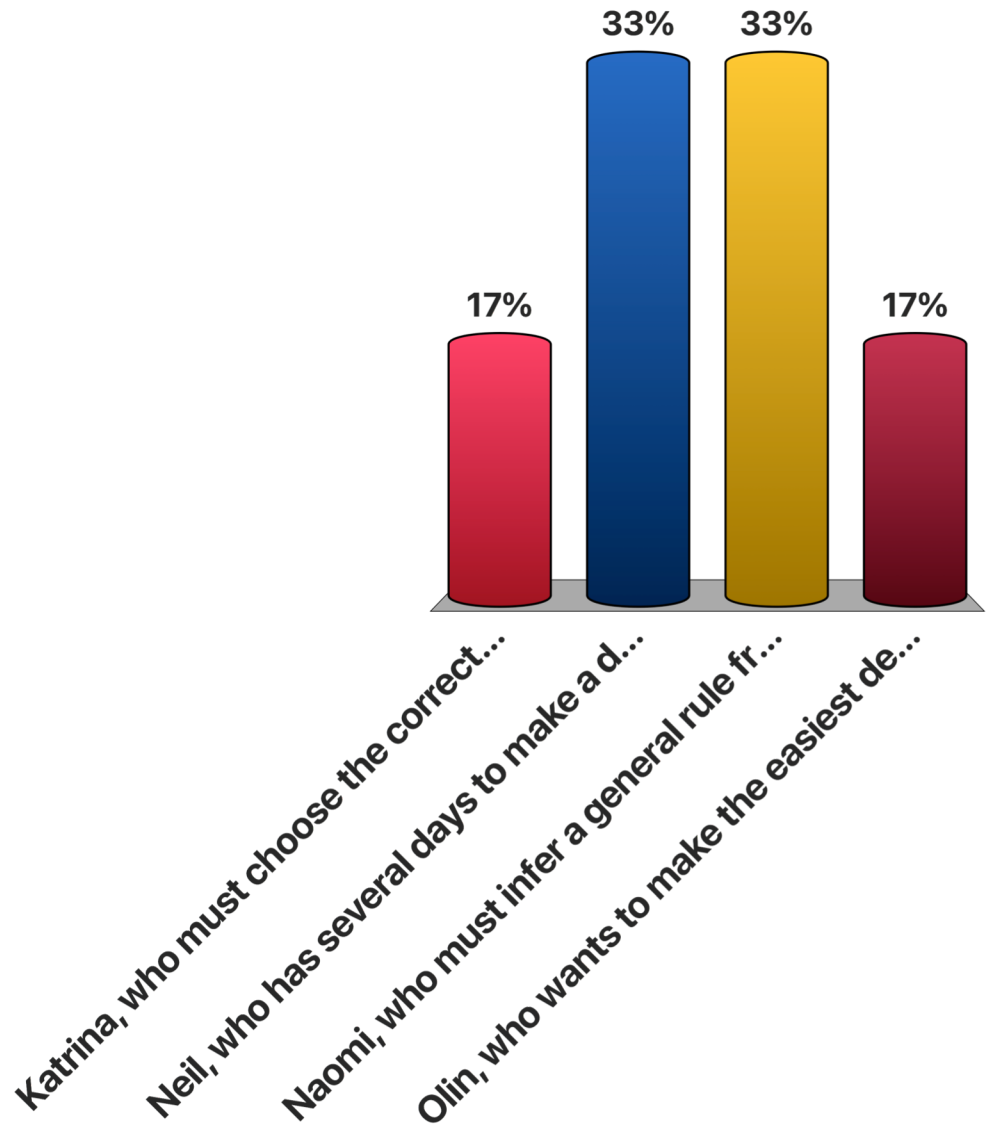
"I think you should be more explicit here in step two."

Is any process an algorithm?

- What is the difference between an *algorithm, heuristic* and a program?
  - Algorithm – a method to problem solving which is guaranteed to give the correct output for a given input, e.g. adding two numbers. Might not be optimally efficient (space/time).
  - Heuristic – a practical method to problem solving which is not guaranteed to give the correct output for a given input, but it is likely to be correct. Might not be optimally efficient (space/time). Predicting winner in a football match, e.g. score is 5-1, 5 minutes before end of game.
  - An algorithm is not the same as the computer program. The computer program is an implementation of the algorithm or a heuristic in a particular programming language.

Heuristics will be most useful for which of the following people?

A. Katrina, who must choose the correct answer on a multiple-choice exam.
B. Neil, who has several days to make a decision about which company to join.
C. Naomi, who must infer a general rule from a set of propositions.
D. Olin, who wants to make the easiest decision about which job to take.



A recap…Typical proteomic MS expt

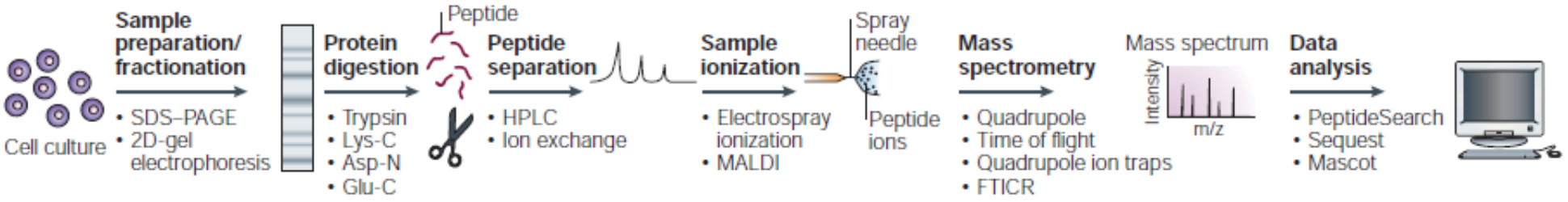


Figure 1 | **The mass-spectrometry/proteomic experiment.** A protein population is prepared from a biological source — for example, a cell culture — and the last step in protein purification is often SDS–PAGE. The gel lane that is obtained is cut into several slices, which are then in-gel digested. Numerous different enzymes and/or chemicals are available for this step. The generated peptide mixture is separated on- or off-line using single or multiple dimensions of peptide separation. Peptides are then ionized by electrospray ionization (depicted) or matrix-assisted laser desorption/ionization (MALDI) and can be analysed by various different mass spectrometers. Finally, the peptide-sequencing data that are obtained from the mass spectra are searched against protein databases using one of a number of database-searching programmes. Examples of the reagents or techniques that can be used at each step of this type of experiment are shown beneath each arrow. 2D, two-dimensional; FTICR, Fourier-transform ion cyclotron resonance; HPLC, high-performance liquid chromatography.

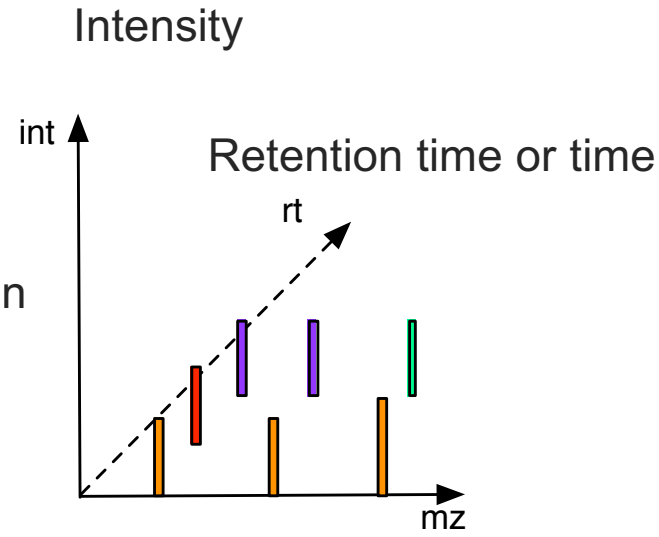
Steen & Mann. The ABC's and XYZ's of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5:699-711, 2004

NANYANG TECHNOLOGICAL UNIVERSITY

What we are really measuring…

3 dimensions of information

Unlike genomics, MS does not give direct information on sequence information



The mass-to-charge ratio (also referred to as m/z)

An example of real spectra

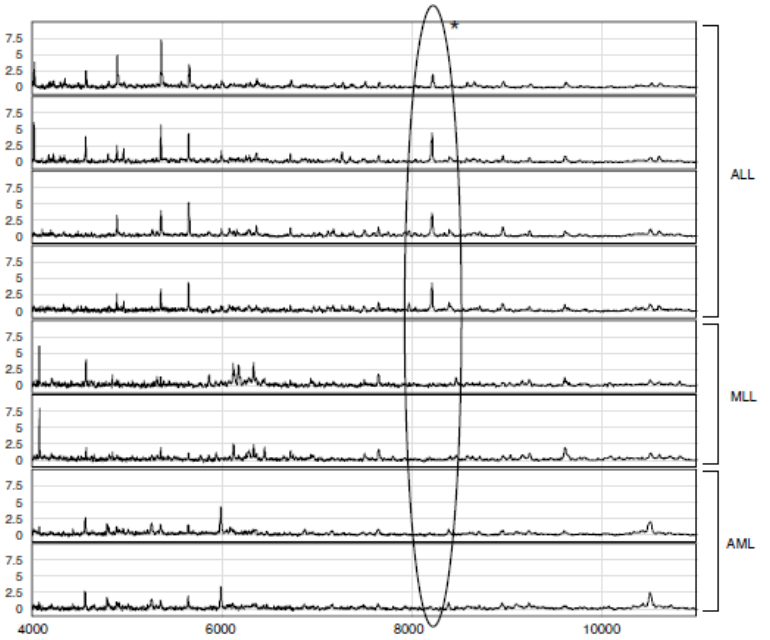


Figure 1 Spectra from SELDI-TOF MS analysis of REH, 697, MV4;11, and Kasumi cell lines. Protein (4 μg) from each cell type was analyzed on SAX2 ProteinChip® Arrays. ALL cell lines shown are REH and 697, the MLL cell line is MV4;11, and the AML cell line is Kasumi. The asterisk indicates the differentially expressed protein at 8.3 kDa.

Does this look like biological information?

Note: Although we can measure 3 dimensions, normally we only use mz and int for identification. rt is usually for effecting separation

NANYANG TECHNOLOGICAL UNIVERSITY

Comparing proteomics and transcriptomics

Similarities

- Proteomic profile
  - Which protein is found in the sample
  - How abundant it is

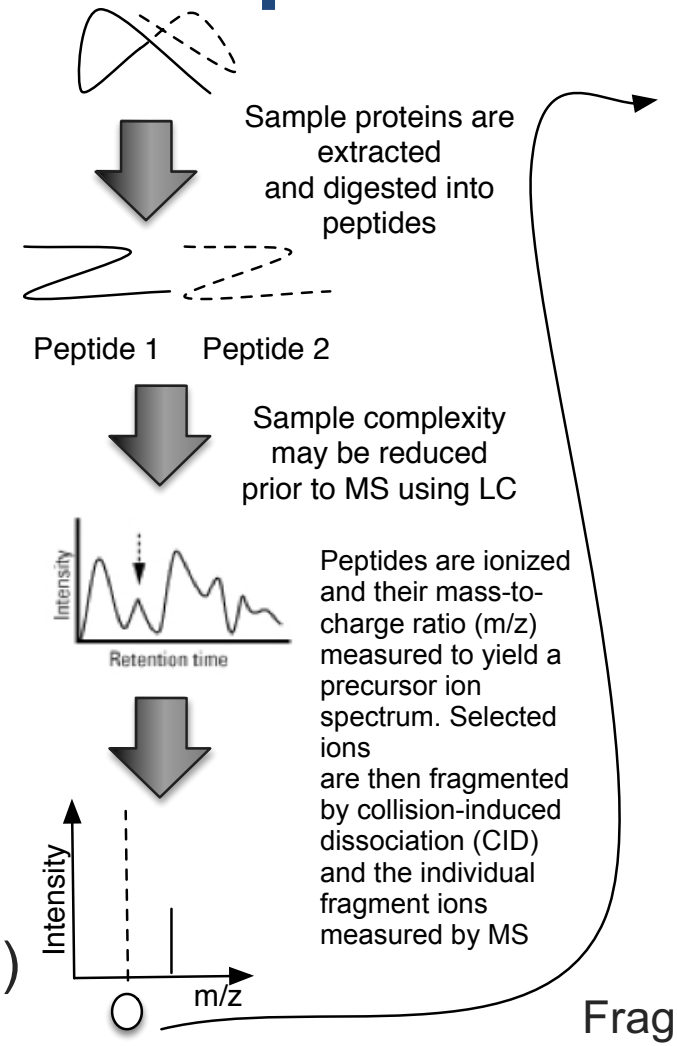
- Similar to gene expression profile. So typical gene expression profile analysis methods can be applied in theory…

Key differences

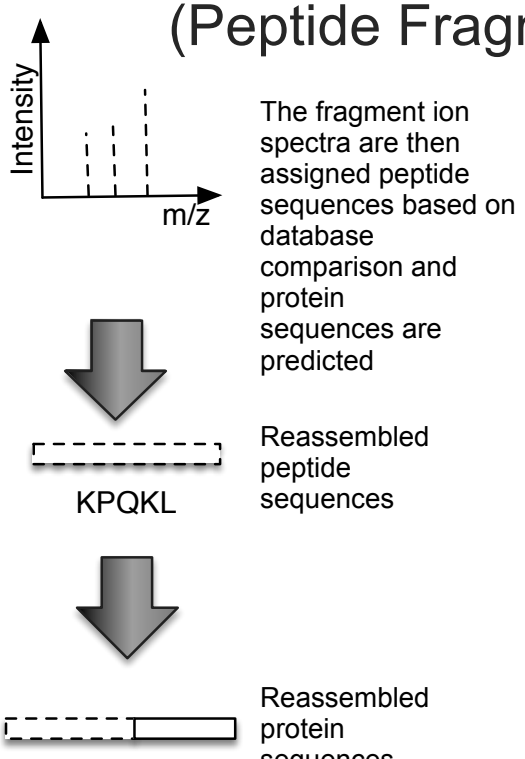
- Profiling
  - Complexity: 20k genes vs 500k proteins
  - Dynamic range: > 10 orders of magnitude in plasma. Proteins cannot be amplified
- Analysis
  - Much fewer features
  - Difficult to reproduce
  - Much fewer samples
  - Unstable quantitation

A look at the important steps of tandem mass spectrometry

The setup is tandem **because** you have 2 mass specs set up one after the other



MS^2 or MS/MS (Peptide Fragments)



Goh and Wong. Network-Based Proteomics, Imperial College Press

NANYANG TECHNOLOGICAL UNIVERSITY

Important concepts:

- Proteases, e.g. trypsin, break a protein into peptides
- Tandem mass spectrometer further breaks the peptides down into fragment ions and measures the mass of each piece
- Mass spectrometer accelerates the fragmented ions; heavier ions accelerate slower than lighter ones
- Mass spectrometer measures signal from the mass/charge ratio of an ion

Relating the various levels of data in MS-based proteomics

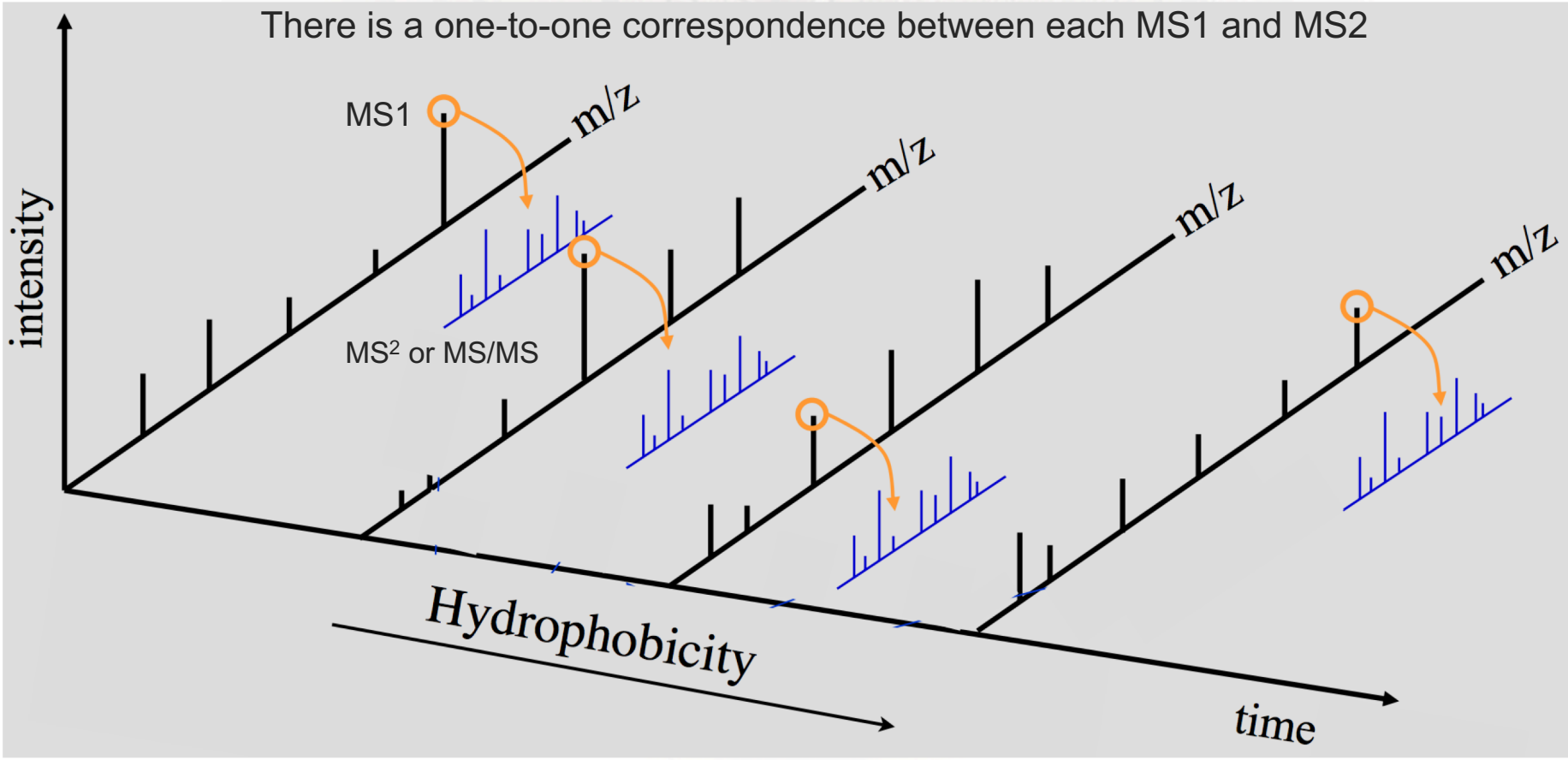


Image from Vitek O, Purdue University

NANYANG TECHNOLOGICAL UNIVERSITY

The problem

- We only have a series of peaks to tell us whether a protein exists or not.
- The peaks relate to masses obtained from the fragmentation of a peptide
- Random fragmentation of peptide along any bond will result in a large number of masses

Some help from biology and chemistry

- Trypsin has defined cleavage sites (so we know the ends of each peptide)
- Collision-induced dissociation tends to break peptides along the peptide bond, resulting in partial amino acids
- We know the masses of each of the 20 amino acids
- What further information do we need?

An idea

- The ends of trypsin-cleavage are “deterministic”. Bond breaking is also “deterministic”*.
- We know the sequences of proteins using *in silico* translation of mRNA sequences.
- What if.. We simulate spectra from known sequences and compare it against our observed spectra?

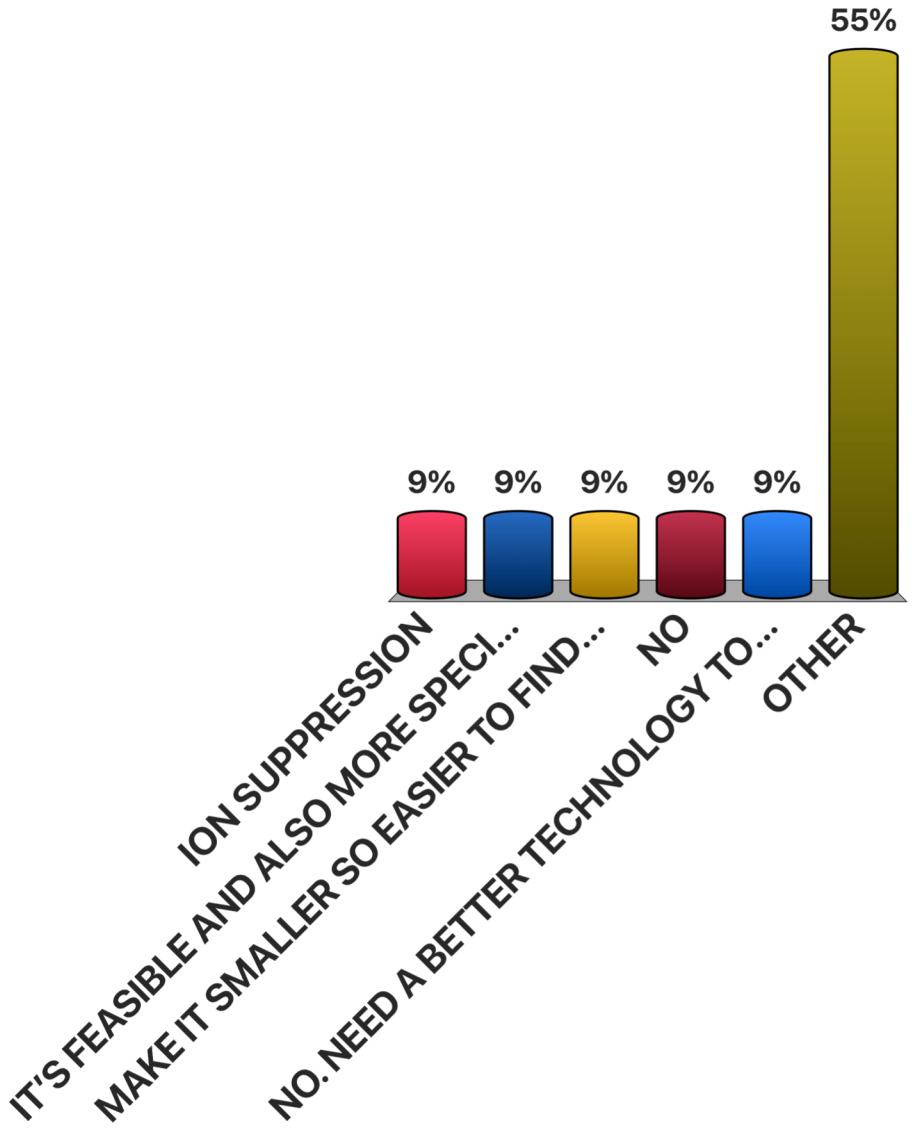
*When I say deterministic, I mean it is conserved behavior when it does happen. It does not mean that it always happens.

Is it feasible to use whole proteins for sequence matching? Would it make the matching more specific? Can it cause problems?

Rank	Responses
1	ION SUPPRESS...
2	IT'S FEASIBL...
3	MAKE IT SMAL...
4	NO
5	NO. NEED A B...
6	OTHER

Keyword:

Keyword Matches: 0

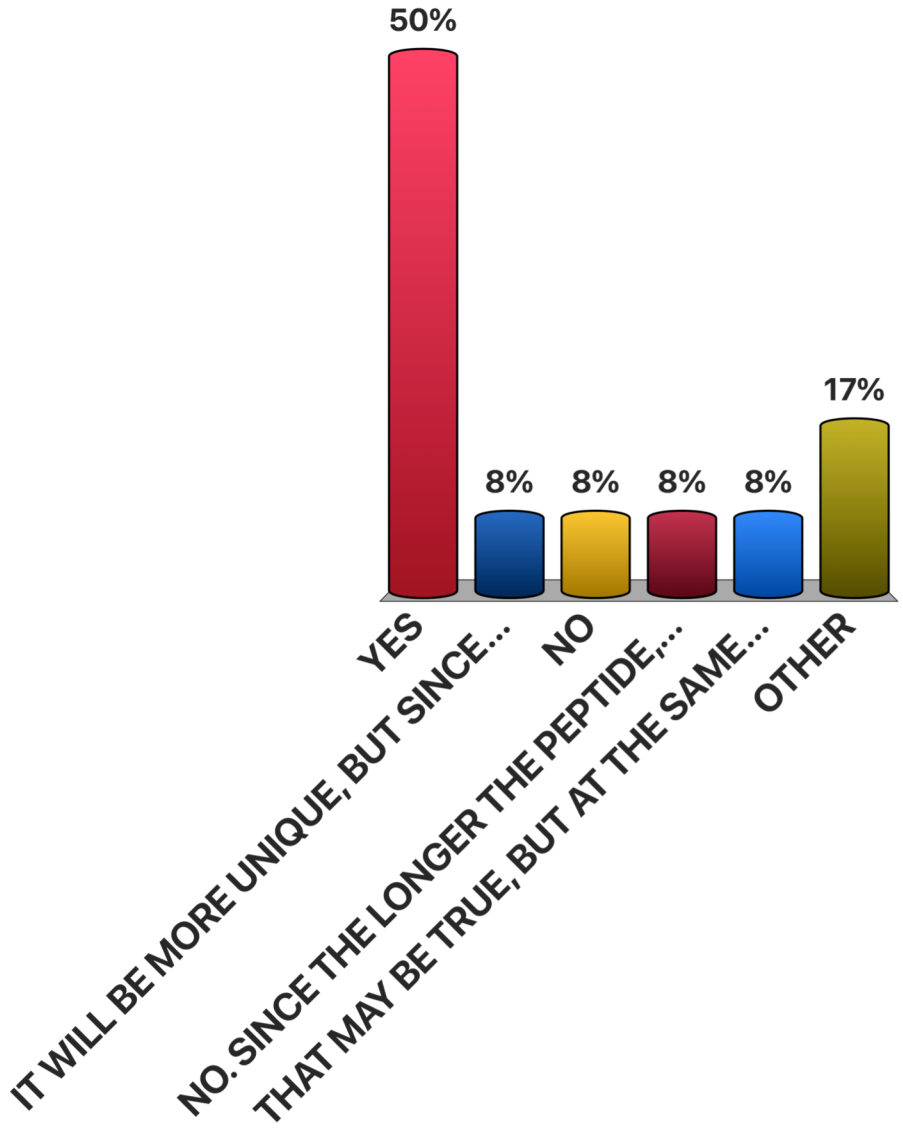


Do you agree that the longer the peptide sequence, the more unique its pattern, and therefore the more specific the match?

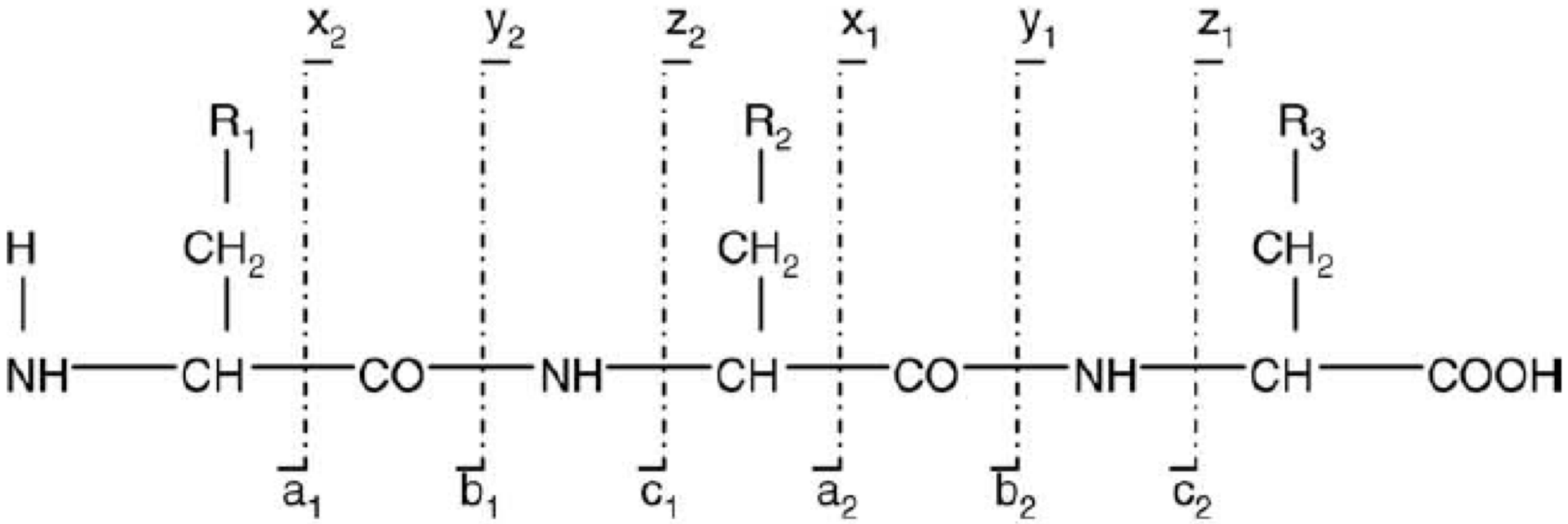
Rank	Responses
1	YES
2	IT WILL BE M...
3	NO
4	NO. SINCE TH...
5	THAT MAY BE ...
6	OTHER

Keyword:

Keyword Matches: 0



Revisiting peptide fragmentation



Generation of beta and gamma ions are more common so they may be more informative during analysis (more completeness)

Fragments can also lose neutral chemical groups like NH3 and H2O

NANYANG TECHNOLOGICAL UNIVERSITY

Peptide fragmentation (mass calculations)

Ion type	K	V	P	Q	V	S	T	P	T	L	R
b+		228.2	325.2	453.3	552.4	639.4	740.4	837.5	938.5	1051.6	
b++		114.6	163.1	227.1	276.7	320.2	370.7	419.3	469.8	526.3	
y+		1097.6	998.6	901.5	773.5	674.4	587.4	486.3	389.3	288.2	175.1
y++		549.3	499.8	451.3	387.2	337.7	294.2	243.7	195.1	144.6	88.1

$M_V = 99.1$
$M_K = 128.1$
$M_P = 97$

Peptide fragmentation (mass shifts)

Ion type	K	V	P	Q	V	S	T	P	T	L	R
b ⟶		228.2	325.2	453.3	552.4	639.4	740.4	837.5	938.5	1051.6	
b++		114.6	163.1	227.1	276.7	320.2	370.7	419.3	469.8	526.3	
y		1097.6	998.6	901.5	773.5	674.4	587.4	486.3	389.3	288.2	175.1
y++		549.3	499.8	451.3	387.2	337.7	294.2	243.7	195.1	144.6	88.1
							+80	+80	+160	+160	
Ion type	K	V	P	Q	V	S	T*	P	T*	L	R
b ⟶		228.2	325.2	453.3	552.4	639.4	820.4	917.5	1098.5	1211.6	
y		1257.6	1158.5	1061.4	933.4	834.3	747.3	566.3	469.2	288.2	175.1

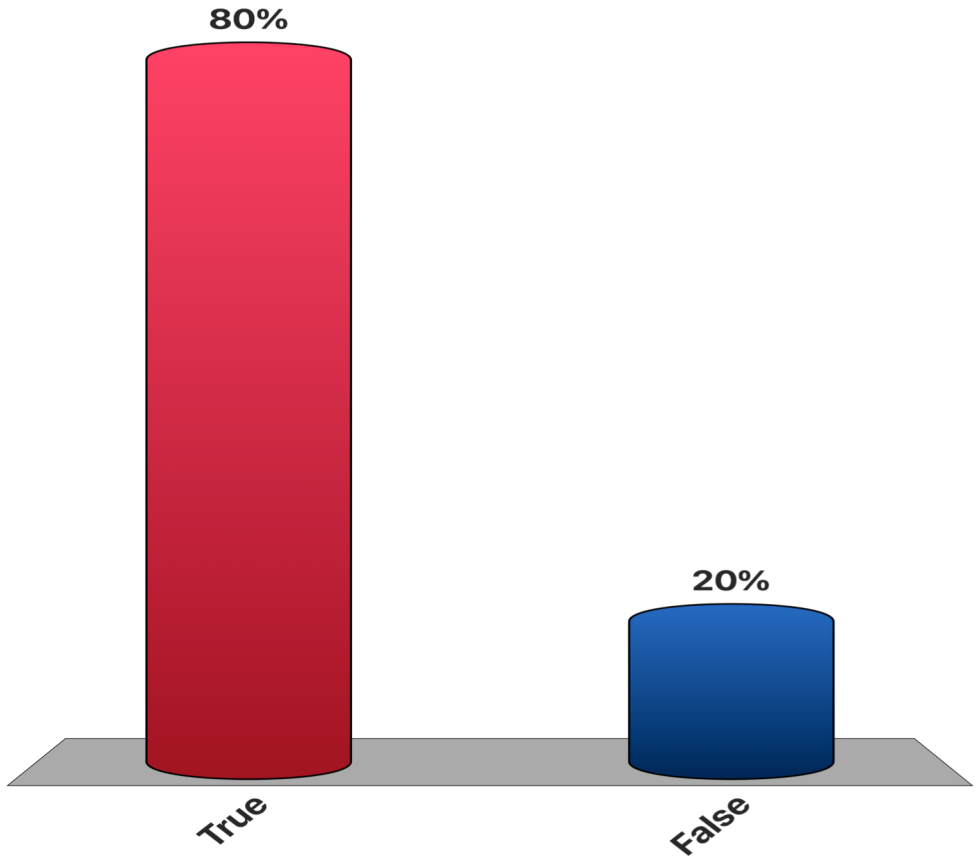
Phosphorylated threonine residues (.79.9663 Da). Note that all fragment ions including the ion with one or two threonine residues are shifted in mass once or twice, respectively.

NANYANG TECHNOLOGICAL UNIVERSITY

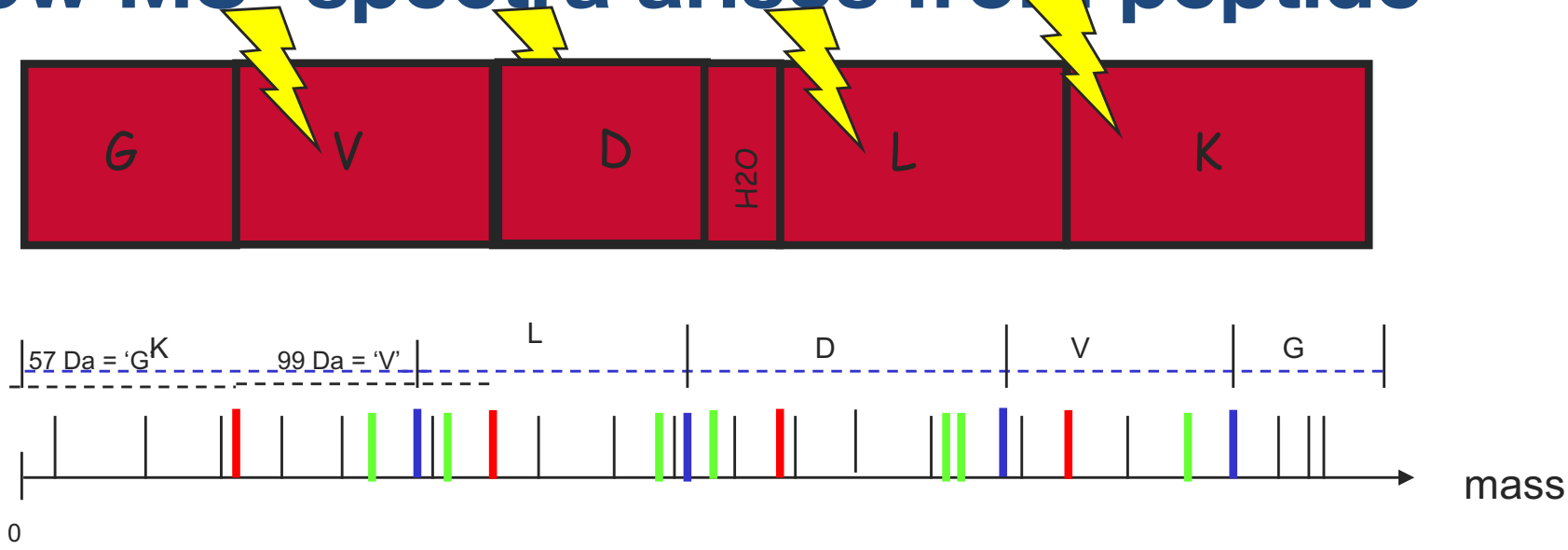
Food for thought --- where are these PTM sites supposed to be anyway?

It is important to use bioinformatics tools to predict sites for posttranslational modifications based on specific protein sequences. However, prediction of such modifications can often be difficult because the short lengths of the sequence motifs associated with certain modifications.

A. True
B. False



How MS^2 spectra arises from peptide

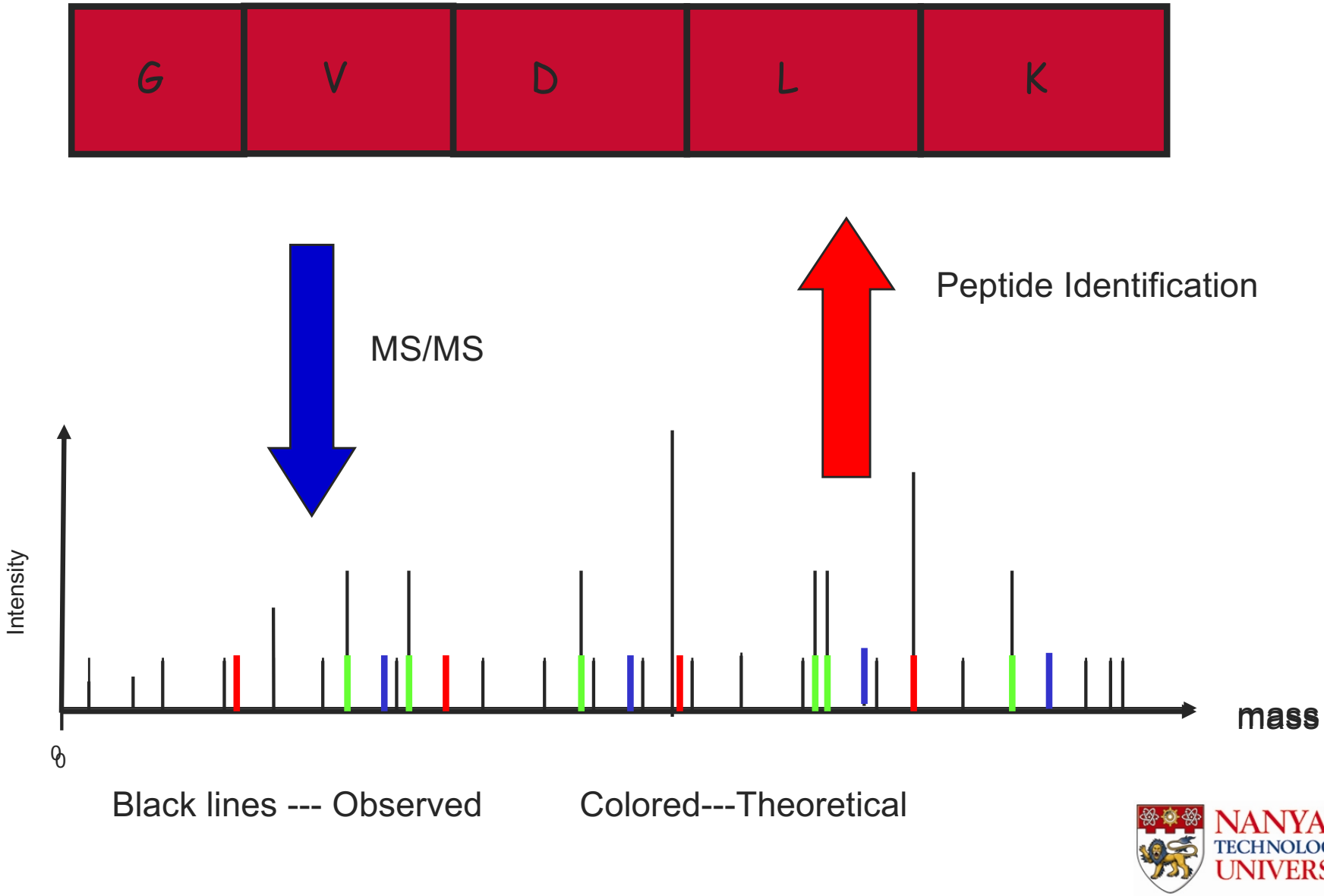


- The peaks in the mass spectrum:
  - Prefix and Suffix Fragments
  - Fragments with neutral losses ($-H_2O$, $-NH_3$)
  - Noise and missing peaks

Real spectra is complex…

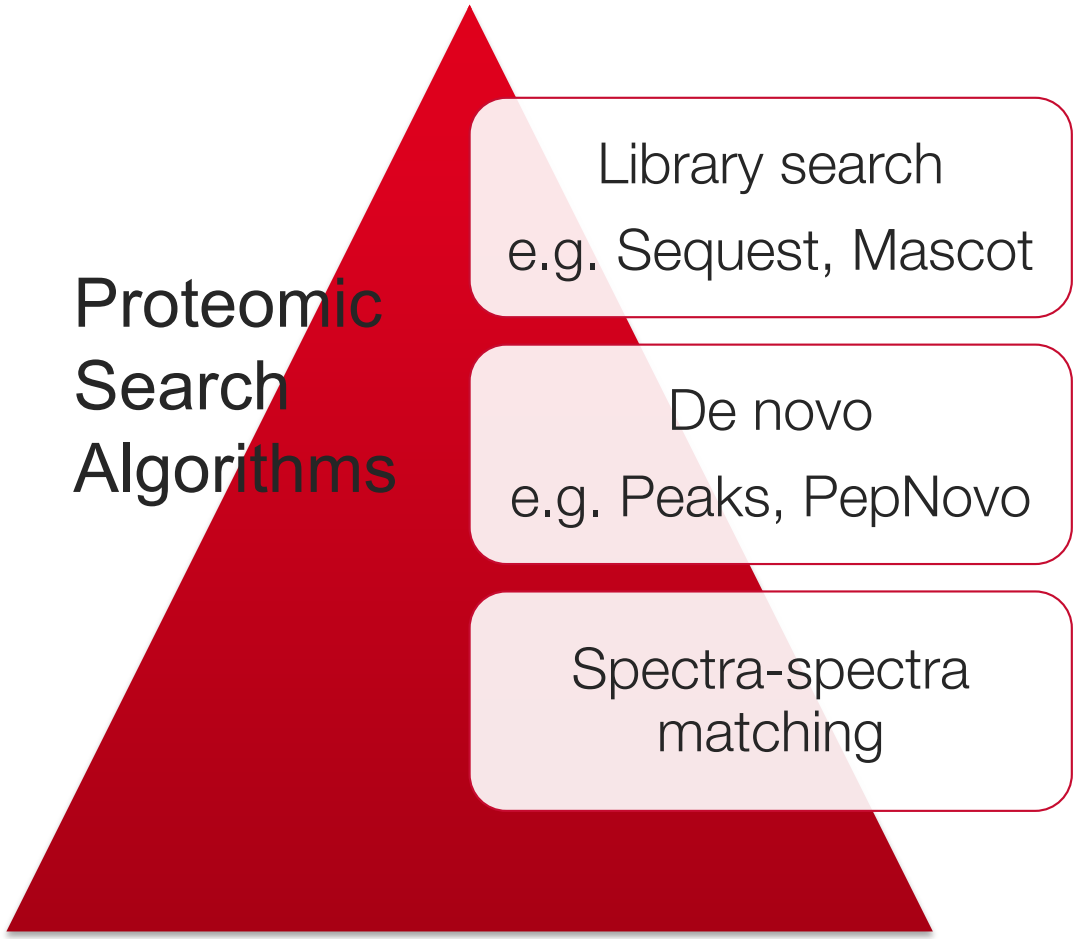
NANYANG TECHNOLOGICAL UNIVERSITY

Protein identification with MS/MS



NANYANG TECHNOLOGICAL UNIVERSITY

3 main approaches for peptide identification

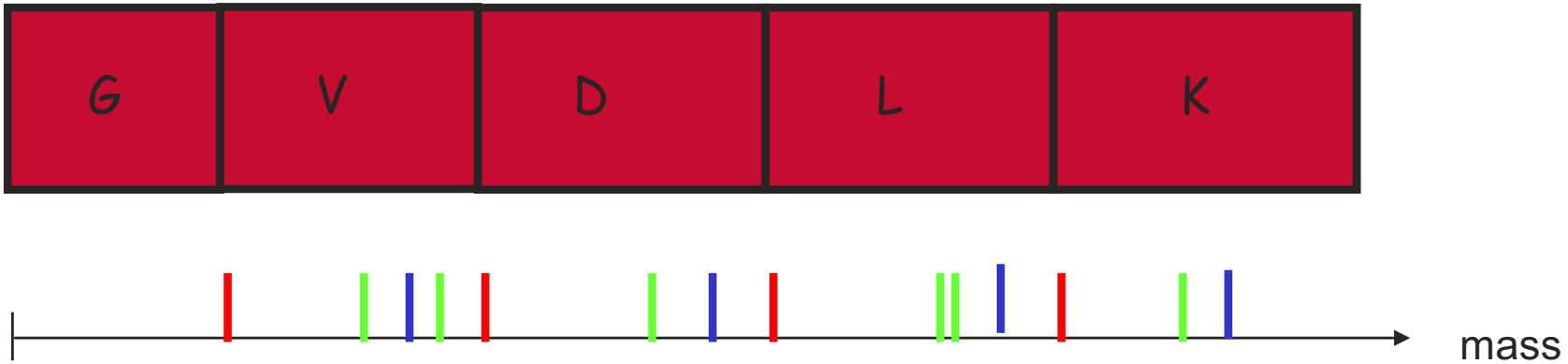


Library search algorithms

- Library (Database) search
  - Used for spectrum from known peptides
  - Rely on completeness of database
- General Approach
  - Match given spectrum with known peptide
  - Enhanced with advanced statistical analysis and complex scoring functions
- Methods
  - SEQUEST, MASCOT, InsPecT, Paragon

Theoretical spectrum for a peptide

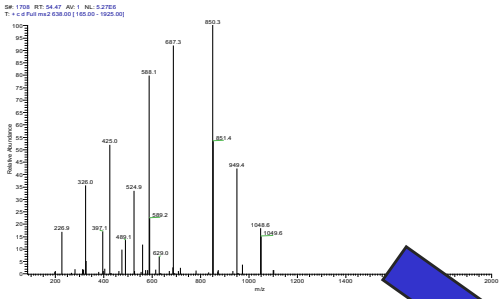
- Given this peptide



- It's theoretical spectrum is
- Theoretical spectrum is dependent on
  - Set of ion-types considered (beta-gamma)
  - Larger if multi-charge ions are considered
  - (+2, +3, +4)

The steps of a library search algorithm

Database Search



Match

Database of known peptides

MDERHILNM, KLQWVCSDL, PTYWASDL, ENQIKRSACVM, TLACHGGEM, NGALPQWRT, HLLERTKMNVV, GGPASSDA, GGLITGMQSD, MQPLMNWE, ALKIIMNVRT, AVGELTK, HEWAILF, GHNLWAMNAC, GVFGSVLRA, EKLNKAATYIN..

0

Theoretical spectrum

Matching Score for this peptide

Repeat for all the peptides in the Database

NANYANG TECHNOLOGICAL UNIVERSITY

What is the database?

Advances in high-throughput and advanced technologies allow researchers to routinely perform whole genome and proteome analysis. For this purpose, they need high-quality resources providing comprehensive gene and protein sets for their organisms of interest. Using the example of the human proteome, we will describe the content of a complete proteome in the UniProt Knowledgebase (UniProtKB). We will show how manual expert curation of UniProtKB/Swiss-Prot is complemented by expert-driven automatic annotation to build a comprehensive, high-quality and traceable resource. We will also illustrate how the complexity of the human proteome is captured and structured in UniProtKB.
Database URL : http://www.uniprot.org/

The expert curation of all functional protein isoforms produced by **alternative splicing** will require more resources and time. However, a complementary pipeline for the import of predicted human protein sequences in UniProtKB/TrEMBL has been developed in collaboration with Ensembl to complete the set of human isoform sequences. >49 000 additional predicted alternative products are currently available in UniProtKB/TrEMBL.

Read: https://academic.oup.com/database/article/doi/10.1093/database/bav120/2630095

NANYANG TECHNOLOGICAL UNIVERSITY

The steps of a library search algorithm

Pseudocode example for peptide mass finger printing

```
Get the experimental mass list L
For each sequence s in the database do
  digest s and obtain a set of peptides P
  for each peptide p in P do
    compute mass(p)
    push mass(p) in M
  x <- score(M,L)
  store score x for protein s
compute p-values for each score
return the n best proteins  /* highest score or lowest p-value */
```

Is this suitable for helping us do library search as we have looked at earlier?

Colinge and Bennett, 2007

Is this suitable for helping us do library search as we have looked at earlier? What is missing?

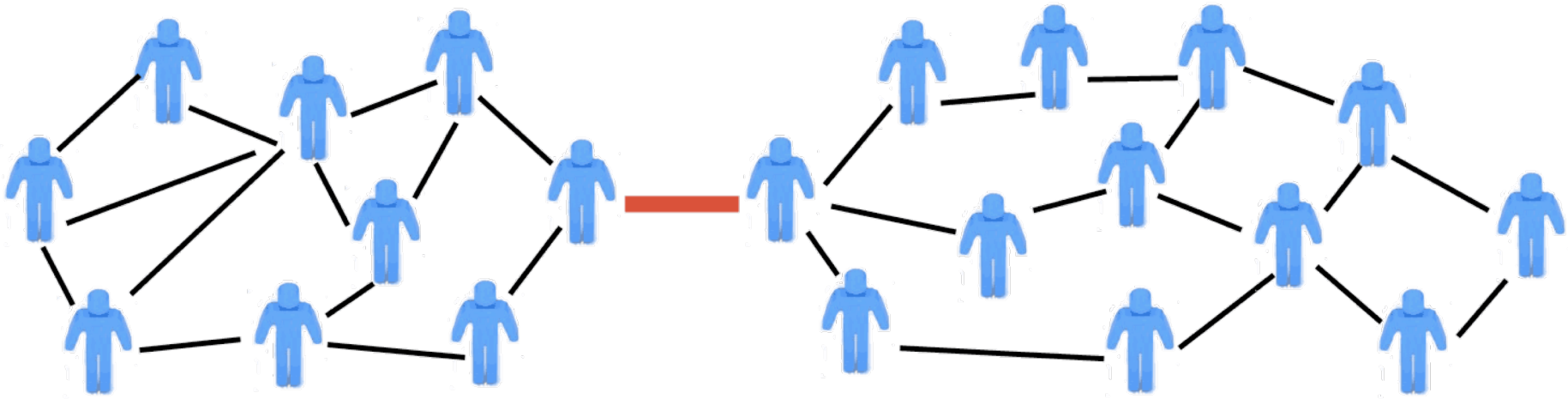
Rank	Responses
1	
2	
3	
4	
5	
6	Other

The steps of a *de novo* sequencing algorithms

- Given a spectrum
  - Build a spectrum graph (A spectrum graph is simply a graph that is built from your Mass spectra.)
  - Peptides are paths in this graph (A path is simply a directed walk from point A to point B or C)
  - Find the best path

A graph

- A generalization of the simple concept of a set of items and the relationships between them
- Representation: Graph G =(V, E) consists set of vertices denoted by V, or by V(G) and set of edges E, or E(G)



Connectivity

- Basic Idea: In a Graph Reachability among vertices by traversing the edges

  Application Example:

  - In a city to city road-network, if one city can be reached from another city.

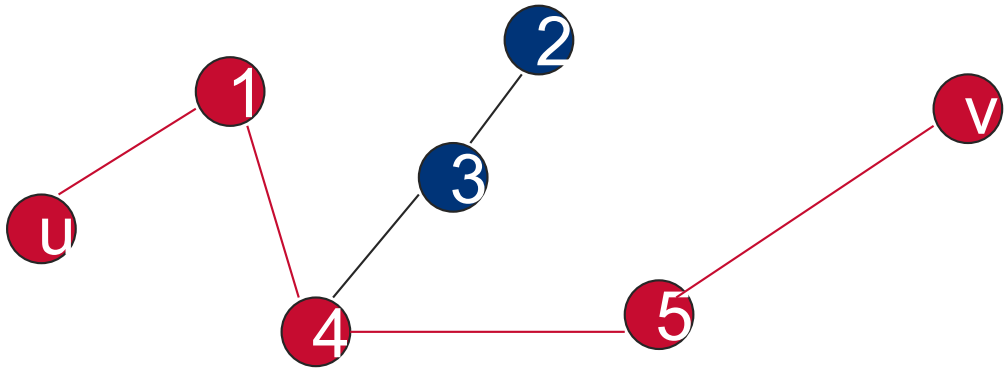
  - Problems if determining whether a message can be sent between two computer using intermediate links

  - Efficiently planning routes for data delivery in the Internet

Connectivity and paths

A **Path** is a sequence of edges that begins at a vertex of a graph and travels along edges of the graph, always connecting pairs of adjacent vertices.

Representation example: G = (V, E), Path P represented, from u to v is {{u, 1}, {1, 4}, {4, 5}, {5, v}}



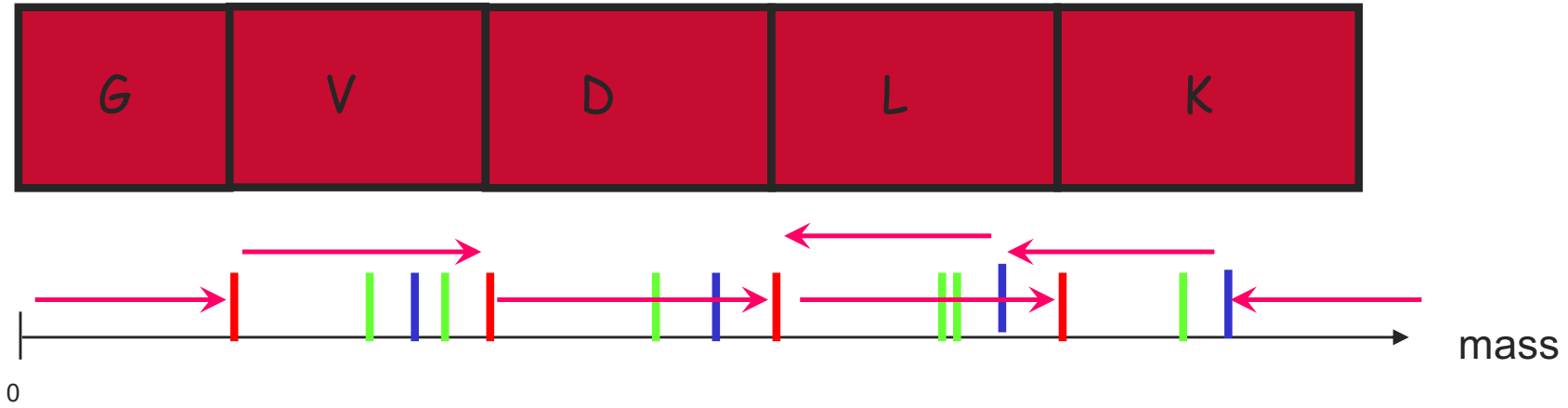
Paths do not cross the same vertex twice.
There are no loops in a path

Amino acids have specific masses

1-letter code	3-letter code	Chemical formula	Monoisotopic	Average
A	Ala	C_3H_5ON	71.03711	71.0788
R	Arg	$C_6H_{12}ON_4$	156.10111	156.1875
N	Asn	$C_4H_6O_2N_2$	114.04293	114.1038
D	Asp	$C_4H_5O_3N$	115.02694	115.0886
C	Cys	C_3H_5ONS	103.00919	103.1388
E	Glu	$C_5H_7O_3N$	129.04259	129.1155
Q	Gln	$C_5H_8O_2N_2$	128.05858	128.1307
G	Gly	C_2H_3ON	57.02146	57.0519
H	His	$C_6H_7ON_3$	137.05891	137.1411
I	Ile	$C_6H_{11}ON$	113.08406	113.1594
L	Leu	$C_6H_{11}ON$	113.08406	113.1594
K	Lys	$C_6H_{12}ON_2$	128.09496	128.1741
M	Met	C_5H_9ONS	131.04049	131.1926
F	Phe	C_9H_9ON	147.06841	147.1766
P	Pro	C_5H_7ON	97.05276	97.1167
S	Ser	$C_3H_5O_2N$	87.03203	87.0782
T	Thr	$C_4H_7O_2N$	101.04768	101.1051
W	Trp	$C_{11}H_{10}ON_2$	186.07931	186.2132
Y	Tyr	$C_9H_9O_2N$	163.06333	163.1760
V	Val	C_5H_9ON	99.06841	99.1326

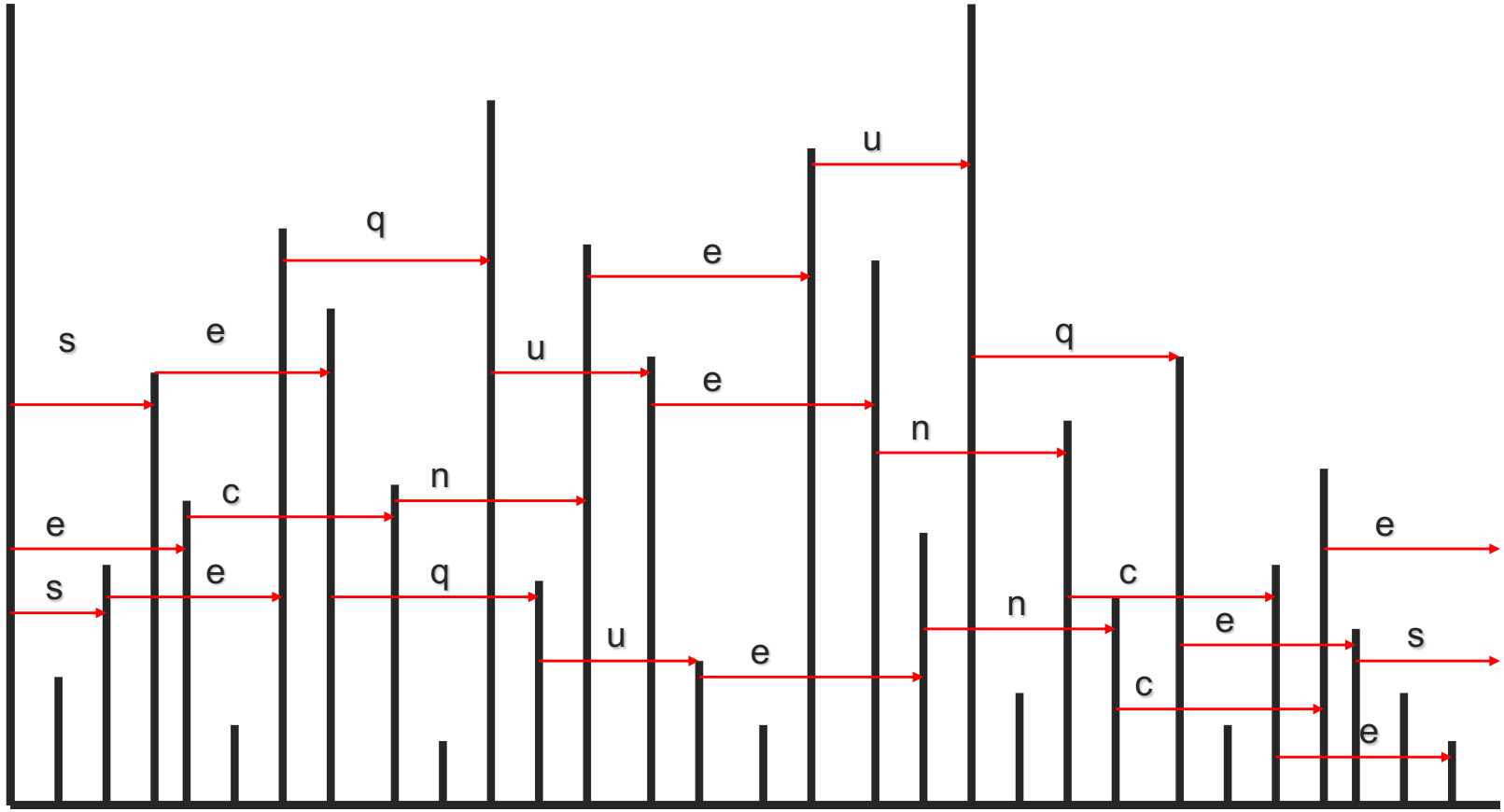
We can use this mass info to reconstruct the peptide sequence! Which amino acid cannot we not distinguish?

Spectrum graph for a peptide

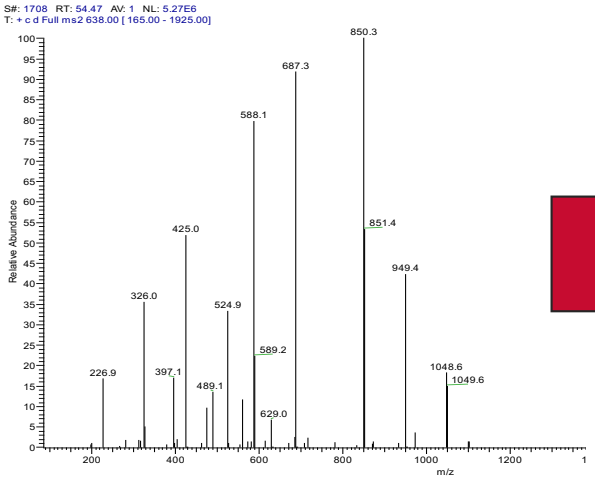


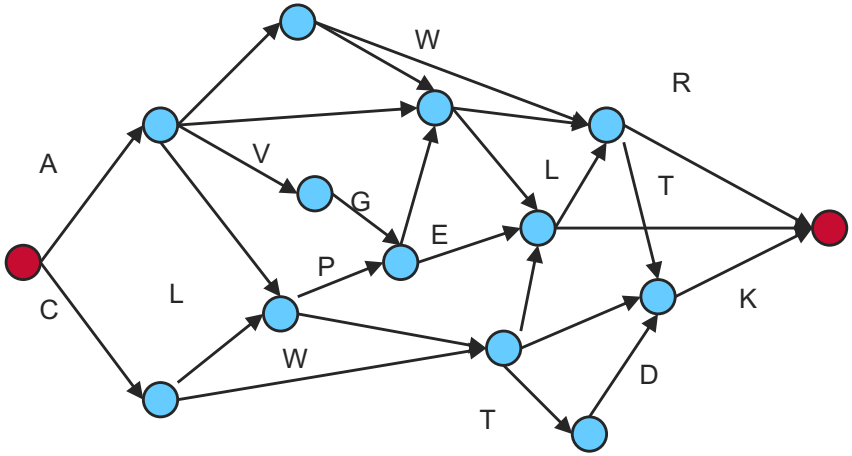
- Connect peaks together
  - If their mass difference = mass of an amino acid
- Theoretical spectrum is dependent on
  - Set of ion-types considered
  - Larger if multi-charge ions are considered

Building a graph from a spectrum



De novo sequencing algorithms





Given a series of peaks…

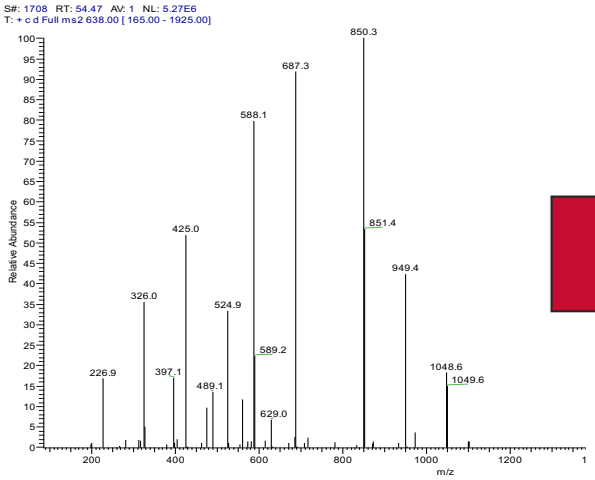
We generate a DAG (Directed Acyclic Graph) from the spectra

Assuming we know that A is the start, and the K is the terminal
Try working out the possible sequences (paths) from start to end.

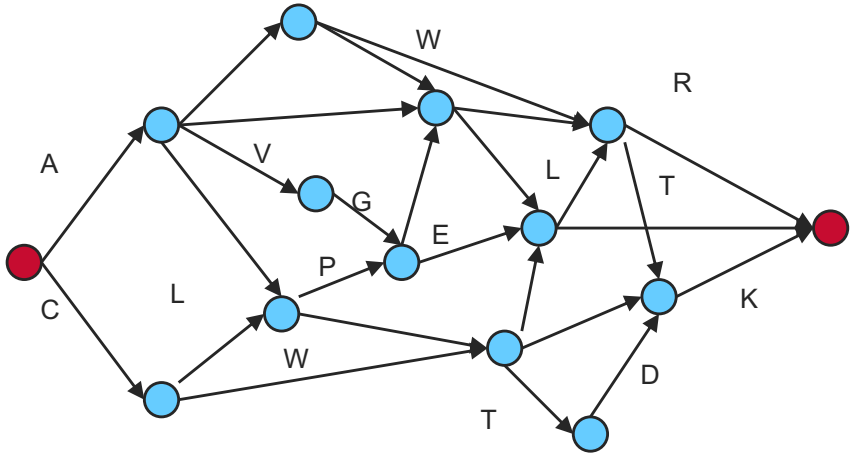
NANYANG TECHNOLOGICAL UNIVERSITY

Frank, et al. “De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry”.
J. Proteome Res. 6:114-123, 2007

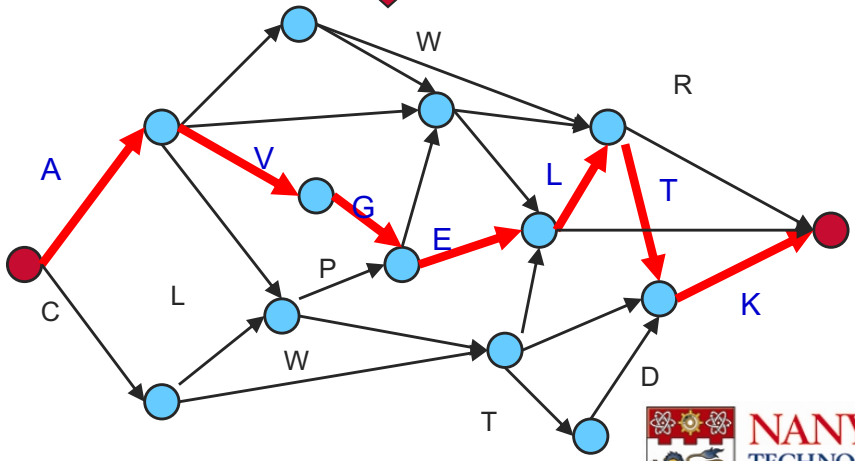
De novo sequencing algorithms



Given a series of peaks…



Find longest directed acyclic path



AVGELTK

What is the rationale for the longest possible path being correct? (The best explanation is the one that can explain the presence of all the peaks)

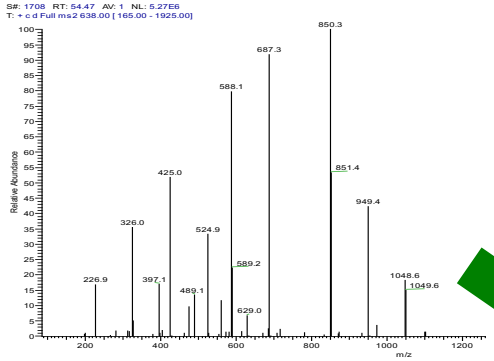
NANYANG TECHNOLOGICAL UNIVERSITY

Is the longest path necessary the correct sequence? What could go wrong?

Rank	Responses
1	
2	
3	
4	
5	
6	Other

2 paths to the same answer ---De novo vs. database search

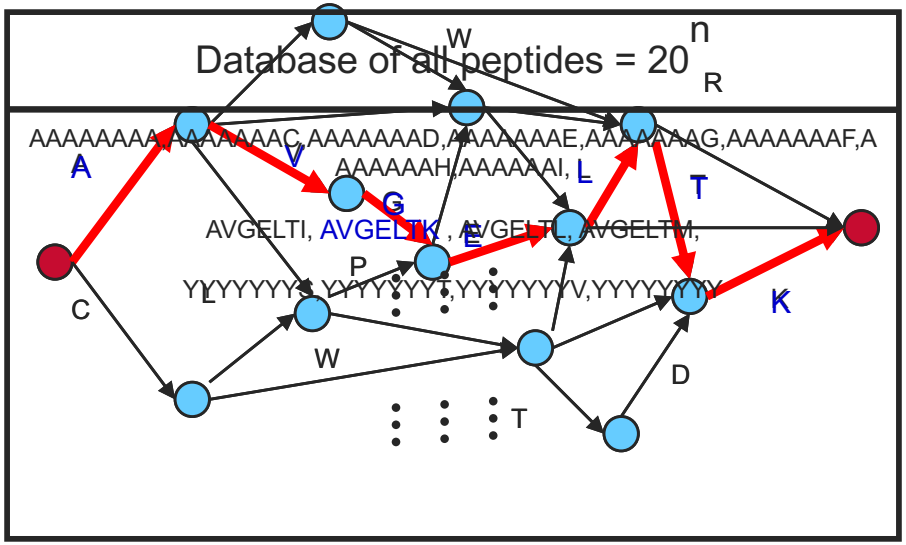
Database Search



De Novo

Database of known peptides

MDERHILNM, KLQWVCSDL, PTYWASDL, ENQIKRSACVM, TLACHGGEM, NGALPQWRT, HLLERTKMNVV, GGPASSDA, GGLITGMQSD, MQPLMNWE, ALKIIMNVRT, AVGELTK, HEWAILF, GHNLWAMNAC, GVFGSVLRA, EKLNKAATYIN..



AVGELTK

De novo vs. database search: A paradox

- The database of all peptides is huge ≈ $O(20^n)$
- The database of all known peptides is much smaller ≈ $O(10^8)$

- However, de novo algorithms can be much faster, even though their search space is much larger!
  - A database search scans all peptides in the search space to find best one
  - De novo eliminates the need to scan all peptides by modeling the problem as a graph search

Break

Steen & Mann. The ABC's and XYZ's of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5:699-711, 2004

Protein identification

- After all the peptides have been identified, they are grouped into protein identifications
- Peptide scores are added up to yield protein scores
- Confidence of a particular peptide identification increases if other peptides identify the same protein and decreases if no other peptides do so
- Protein identifications based on single peptides should only be allowed in exceptional cases

Notice we haven't said anything about quantitation yet

General rules for confident protein identification

- At least 2 unique peptides *
- Coverage of at least 50-70% total protein sequence
- High abundance

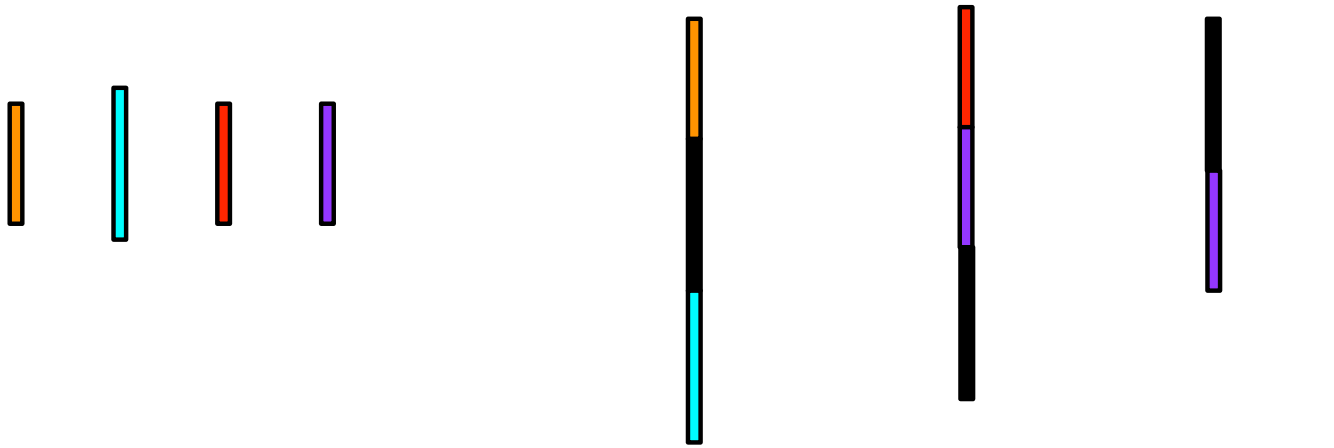
* Commonly used rule.

NANYANG TECHNOLOGICAL UNIVERSITY

Peptides to proteins

Identified peptides

known protein sequences



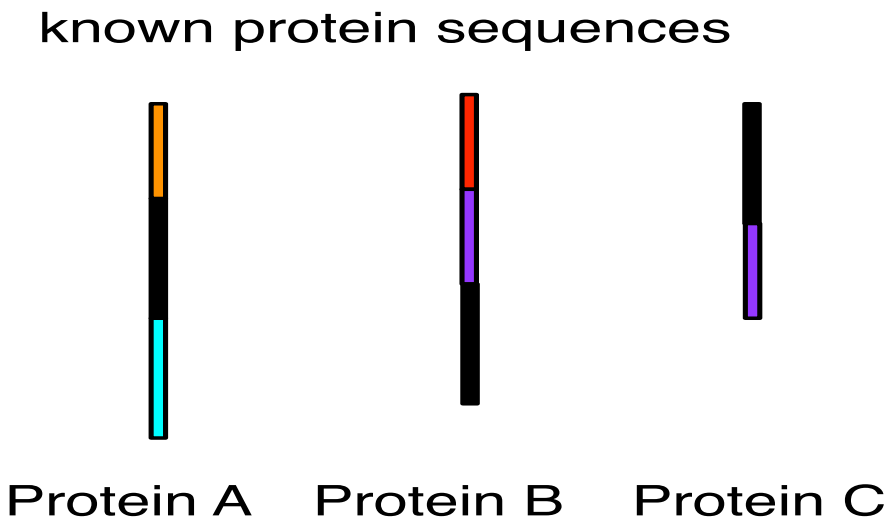
Protein A Protein B Protein C

Sequence that is not supported by any spectra

Order the proteins from highest to lowest confidence ?

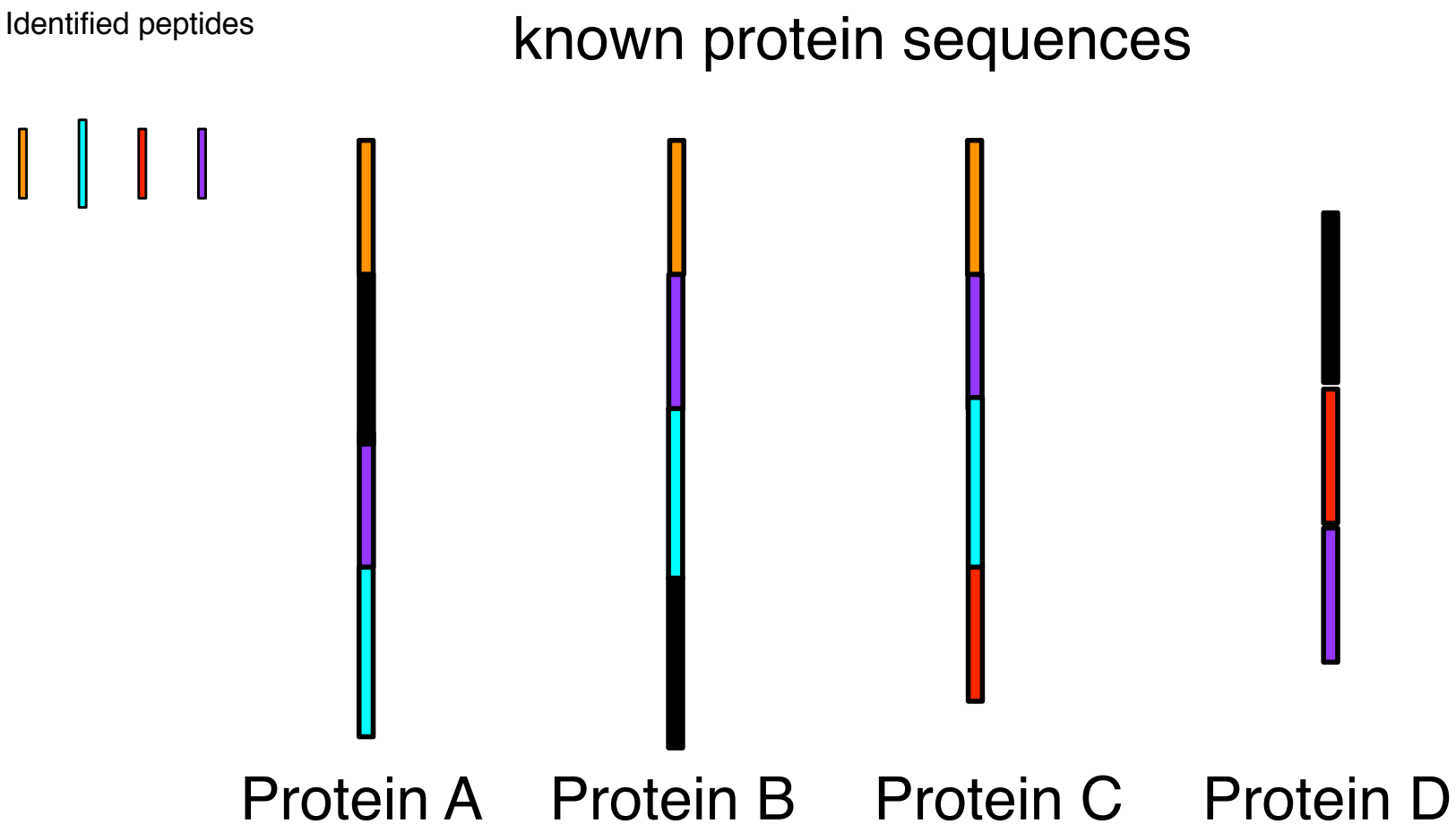
NANYANG TECHNOLOGICAL UNIVERSITY

Order the proteins from highest to lowest confidence ?



A. B -> C -> A
B. A -> B -> C
C. A -> C -> B
D. There is no clear answer

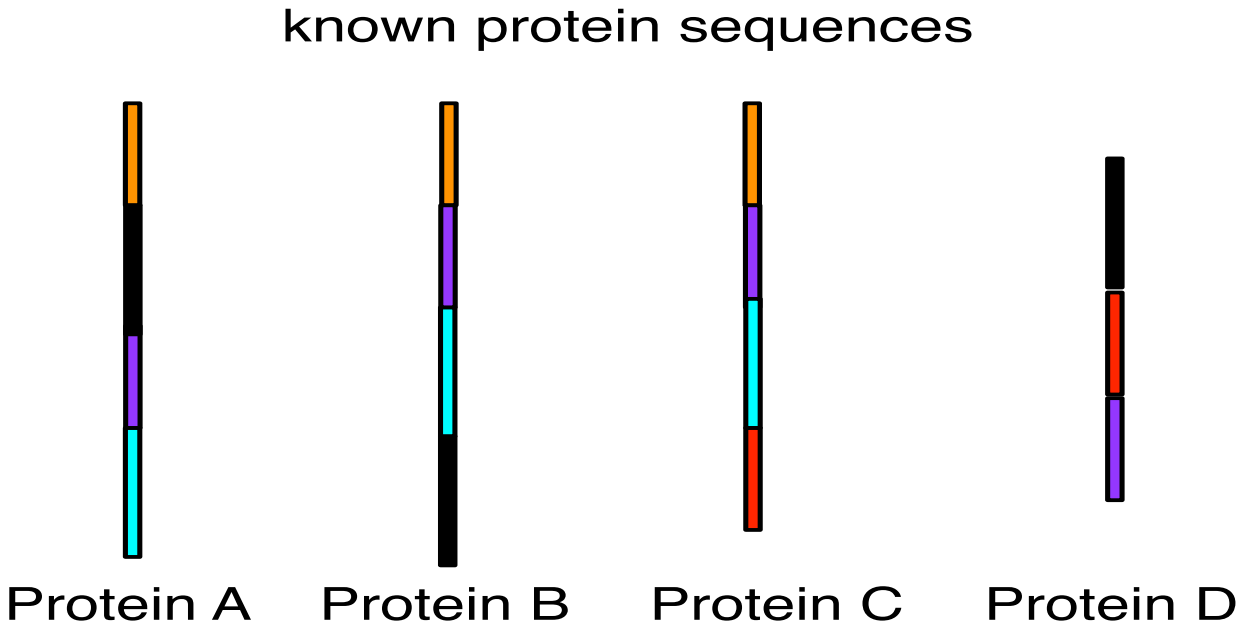
Food for thought --- meaningful ambiguity?



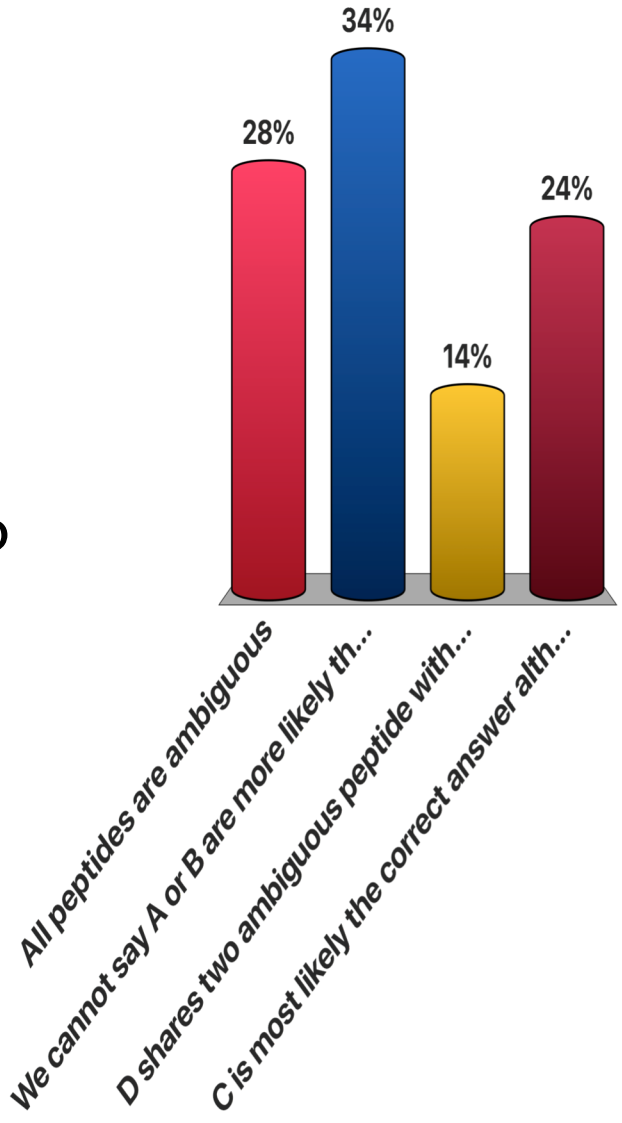
Try to order the proteins from highest to lowest confidence ?

NANYANG TECHNOLOGICAL UNIVERSITY

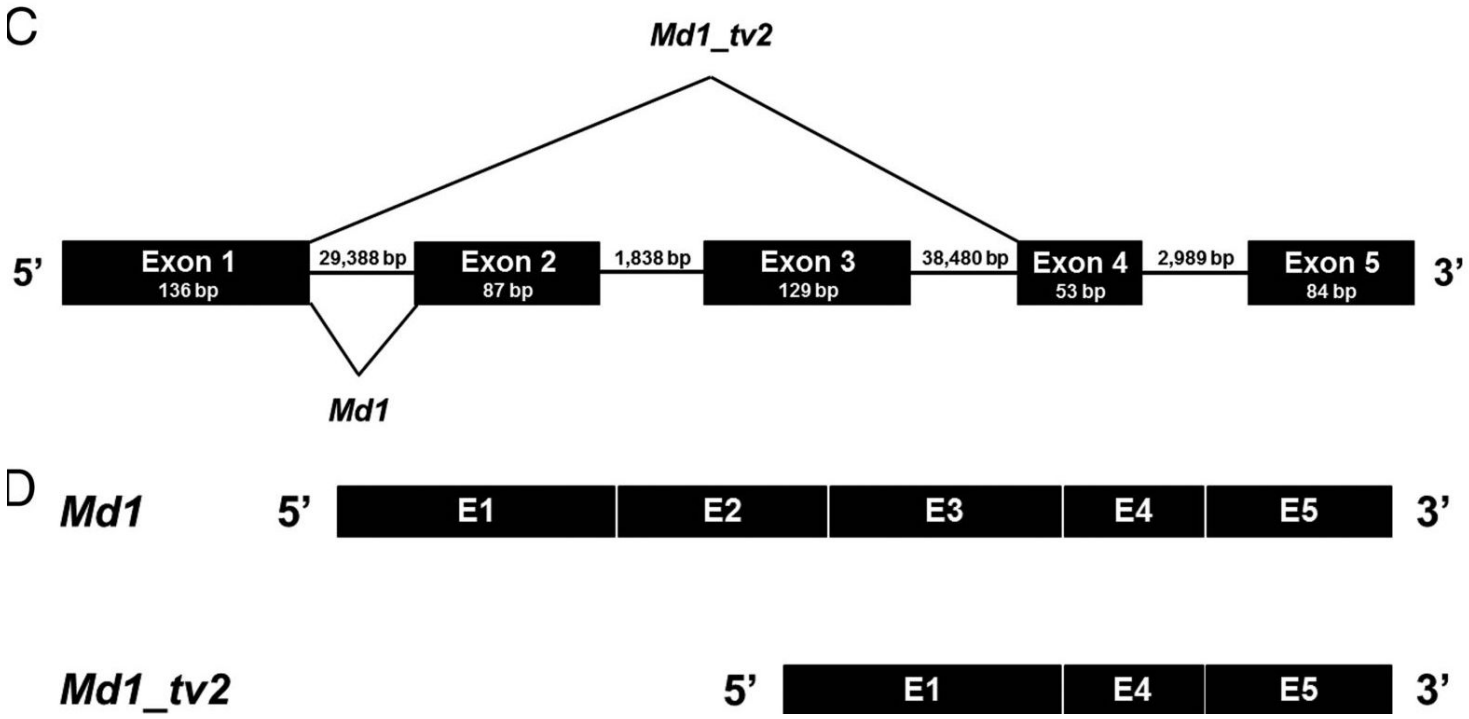
Which of the following statements regarding ambiguous peptides could potentially be true



A. All peptides are ambiguous
B. We cannot say A or B are more likely than the other because they share all same peptides
C. D shares two ambiguous peptide with C, and one with A/B.
D. C is most likely the correct answer although it comprises only ambiguous peptides

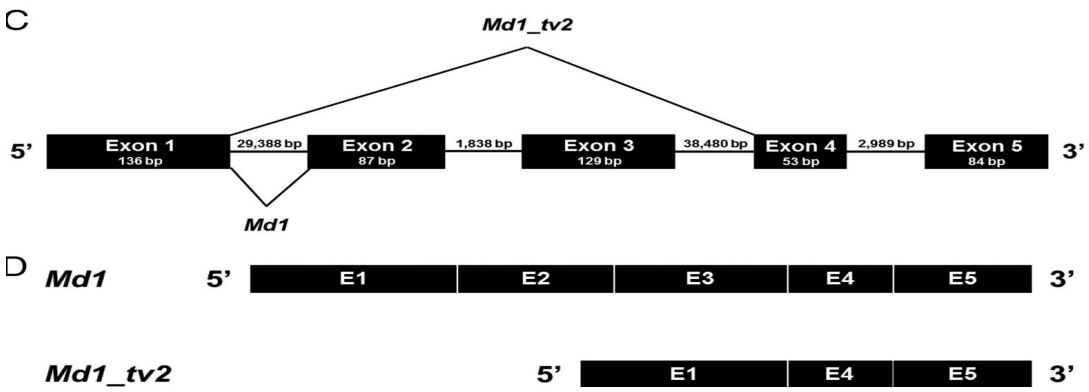


Representative sequences and splice variants

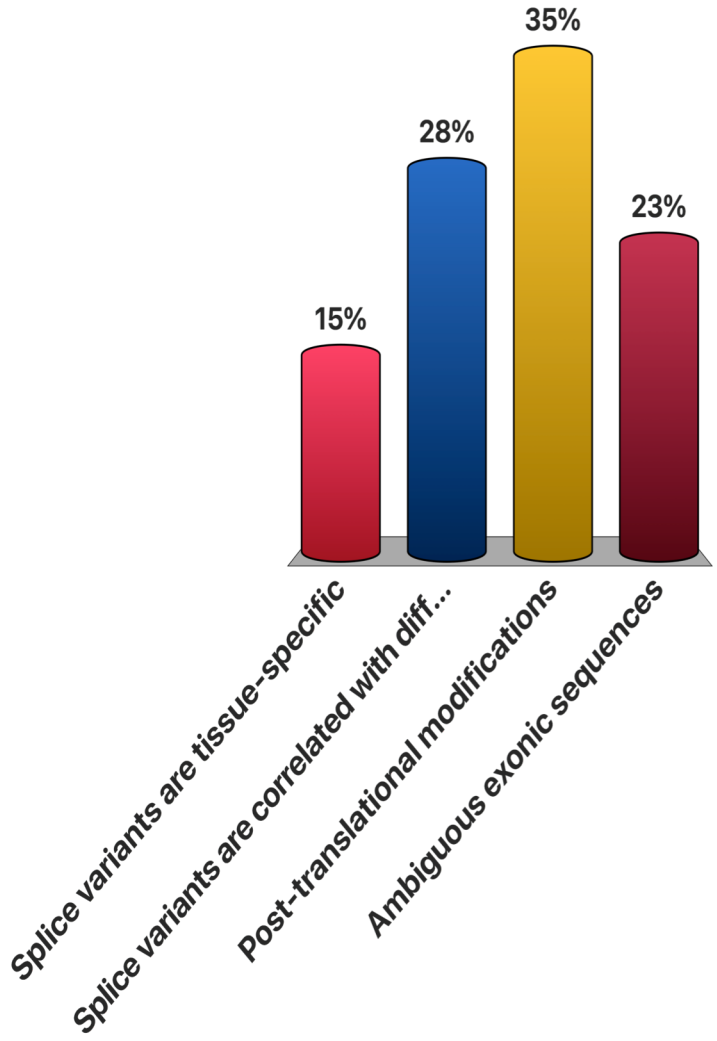


Splice variants are subsets of the full sequence
In proteomics, we usually take the full sequence as the “search sequence”

The following are potential problems when using full sequence for library search

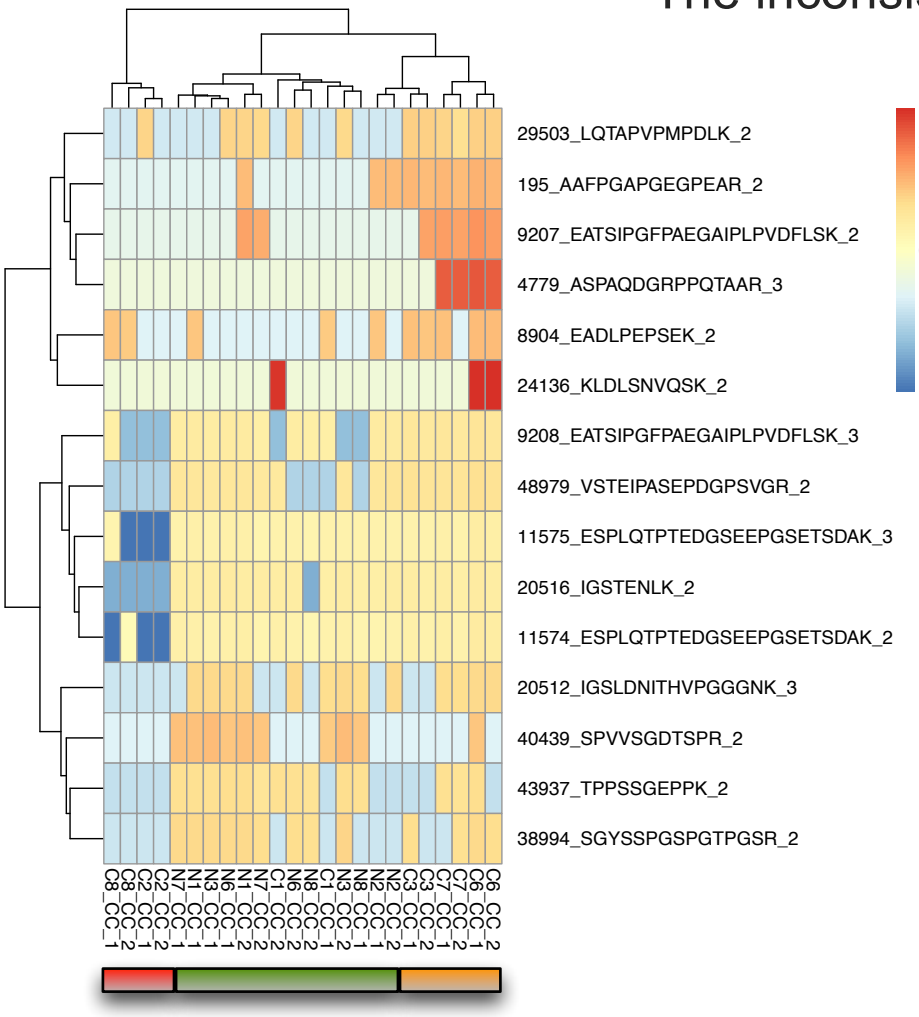


A. Splice variants are tissue-specific
B. Splice variants are correlated with different phenotypes
C. Post-translational modifications
D. Ambiguous exonic sequences



Goh and Wong Spectra-first feature analysis in clinical proteomics—A case study in renal cancer, JBCB, 2016

Food for thought --- splice variants



MAPT, P10636

The inconsistencies corresponded to splice variants

Exon4 - AEEAGIGDTPSLEDEAAGHVTQ^

Exon5 -
EPESGKVVQEGFLREPGPPGLSHQLMSGMPGAPLLPEGPREATRQPSGTGPEDTEGGR
HAPELLKHQLLGDLHQEGPPLKGAGGKERPGSKEEVDEDRDVDESSPQDSPPSKASPA
QDGRPPQTAAREATSIPGFPAEGAIPLPVDFLSKVSTEIPASEPDGPSVGRAKGQDAPL
EFTFHVEITPNVQKEQAHSEEHLGRAAFPGAPGEGPEARGPSLGEDTKEADLPEPSEK
QPAAAPRGKPVSRVPQLK^

Exon6 - ARMVSKSKDGTGSDDKKAK^

Exon7 -
TSTRSSAKTLKNRPCLSPKHPTPGSSDPLIQPSSPAVCPEPPSSPKYVSSVTSRTGSSGA
KEMKLK^

Exon8 -
GADGKTKIATPRGAAPPGQKGQANATRIPAKTPPAPKTPPSS^

Exon9 -
GEPPKSGDRSGYSSPGSPGTPGSRSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRLQ
TAPVPMPDLKNVKSKIGSTENLKHQPGGGK^

Exon10 -
VQIINKKLDLSNVQSKCGSKDNIKHVPGGGS^

Exon11 -
VQIVYKPVDLSKVTSKCGSLGNIHHKP^

Exon12 -
GGGQVEVKSEKLDFKDRVQSKIGSLDNITHVPGGGNKK^

Exon13 -
IETHKLTFRENAKAKTDHGAEIVYKSPVVSGDTSPRHLSNVSSTGSIDMVDSPQLATLA

In cancer, a gene marker MAP was reported to be down-regulated in severe cancer but this gene has 11 splice variants

NANYANG TECHNOLOGICAL UNIVERSITY

Which of the following is most similar to piecing protein information from spectra

A. Putting together one jigsaw puzzle
B. Putting together many jigsaw puzzles with the pieces mixed up
C. Putting together many jigsaw puzzles with the pieces mixed up, and some pieces missing
D. Putting together many jigsaw puzzles with the pieces mixed up, some pieces missing, and some reference picture boxes are wrong

Which of the following is incorrect regarding the Protein Identification through library Search?

A. MS characterization of proteins is highly dependent on bioinformatic analysis
B. Bioinformatics programs can be used to search for the identity of a protein in a database of theoretically digested proteins
C. Even in reality, the protease digestion is always perfect in MS
D. The purpose of the database search is to find exact or nearly exact matches

Representation of proteomics data

- Many format types
- Very complex
- But can be categorized

Format	Vendor/Creator	Comments
Wiff	Applied Biosystems	Proprietary
RAW	Thermo Fisher	Proprietary
raw	Waters	Proprietary
d	Agilent	Proprietary
dta	-	Text-based format
MGF	Matrix Science	Text-based format
mzData	PSI	Markup language; Superseded by mzML
mzML	PSI	Markup language; Current
mzXML	ISB	Markup language; Superseded by mzML

Based on what you know… how many different data formats exist for genomics data?

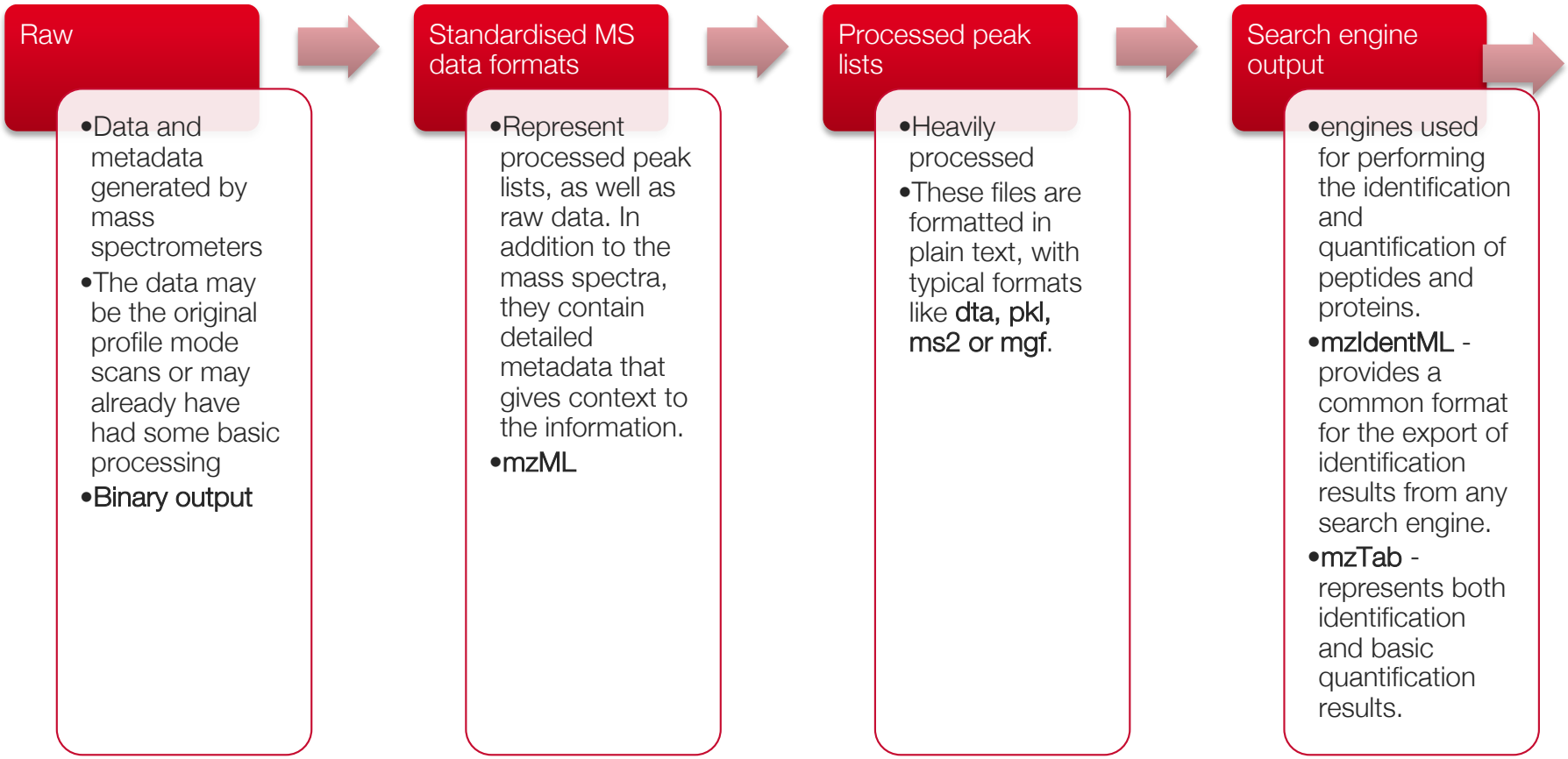
Let's see what happens when we are messy

https://www.youtube.com/watch?v=1LfgbB0Mcqo

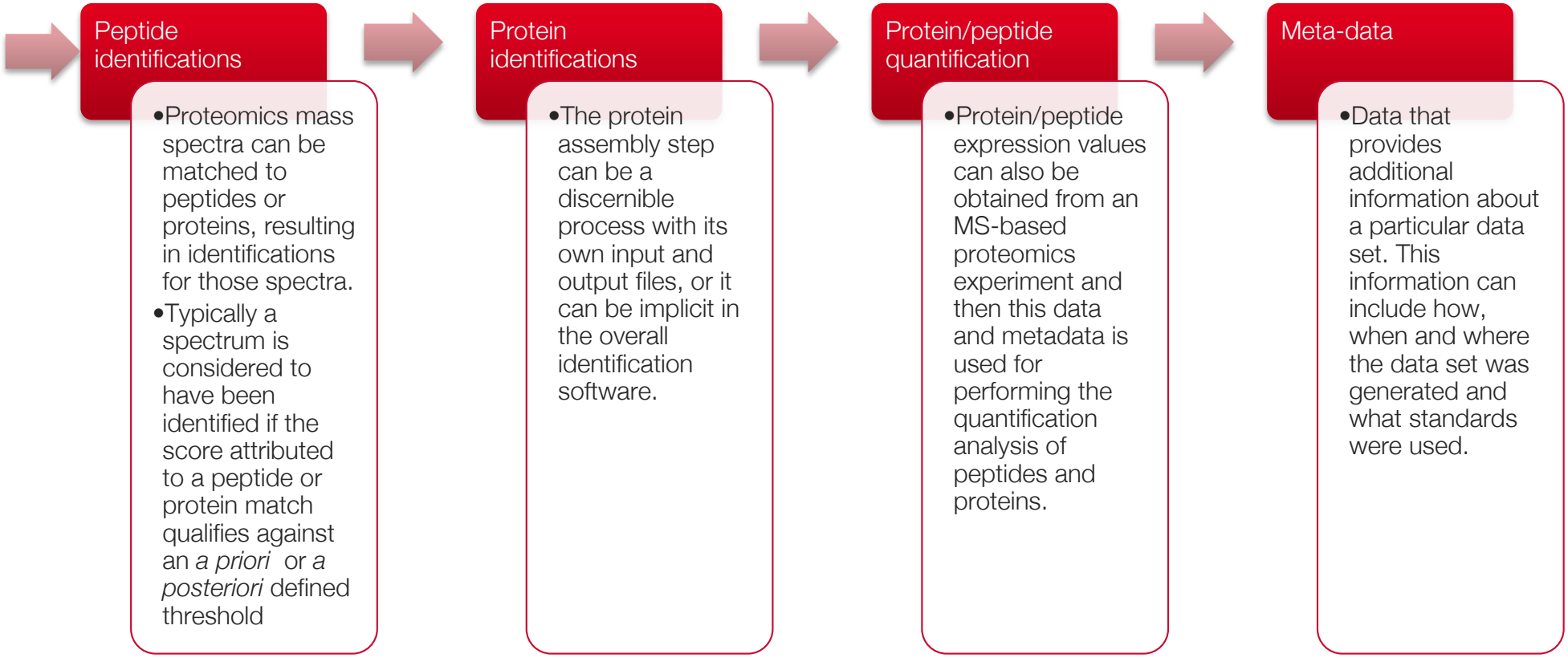
Organizing and formatting proteomics data

- Proteomics data is complex
- We need to organize data so that others can easily read and get the information they need

Proteomics data has many layers



Proteomics data has many layers



NANYANG TECHNOLOGICAL UNIVERSITY

List-based formats

Some examples:

mgf (mascot generic file)

```
BEGIN IONS
PEPMASS=406.283
CHARGE=2+,3+
TITLE=Experiment_1
145.119100 8
217.142900 75
409.221455 11
438.314735 46
567.400183 24
714.447552 31
116.113400 72
91.2165000 32
405.288933 94
39.3021000 12
549.379462 21
715.466300 81
15.1098000 62
45.1358430 28
```

Parameters

mz intensity

pkl (peak list)

```
814.27      22673800    1
221.06      2529.3
223.84      220.9
226.91      1026.9
227.97      1037.9
231.06      110.6
239.05      7193.1
239.74      2513.3
240.27      363.4
240.79      1314.7
241.45      629.9
254.85      332.5
259.71      200.5
260.93      2437.7
```

mz intensity

Which list-based format Is more informative? What information can you tell from these lists?

Non-rich descriptions
Lacking meta-data

List-based formats are easy to read?

A. True
B. False

List-based formats are an efficient storage of information?

A. True
B. False

List-based formats are information-rich (i.e., allow me to store a lot of data on running parameters, machines used, etc etc?)

A. True
B. False

Open-source formats

- Proprietary formats locks proteomics analysis to packaged software with instrument platform.
- Not all software are available
- Solution: Standardize formatting to open-source format

- All open source formats are **XML-based**.

What is XML?

- Markup language
- Controlled vocabulary
- Have you heard of HTML?

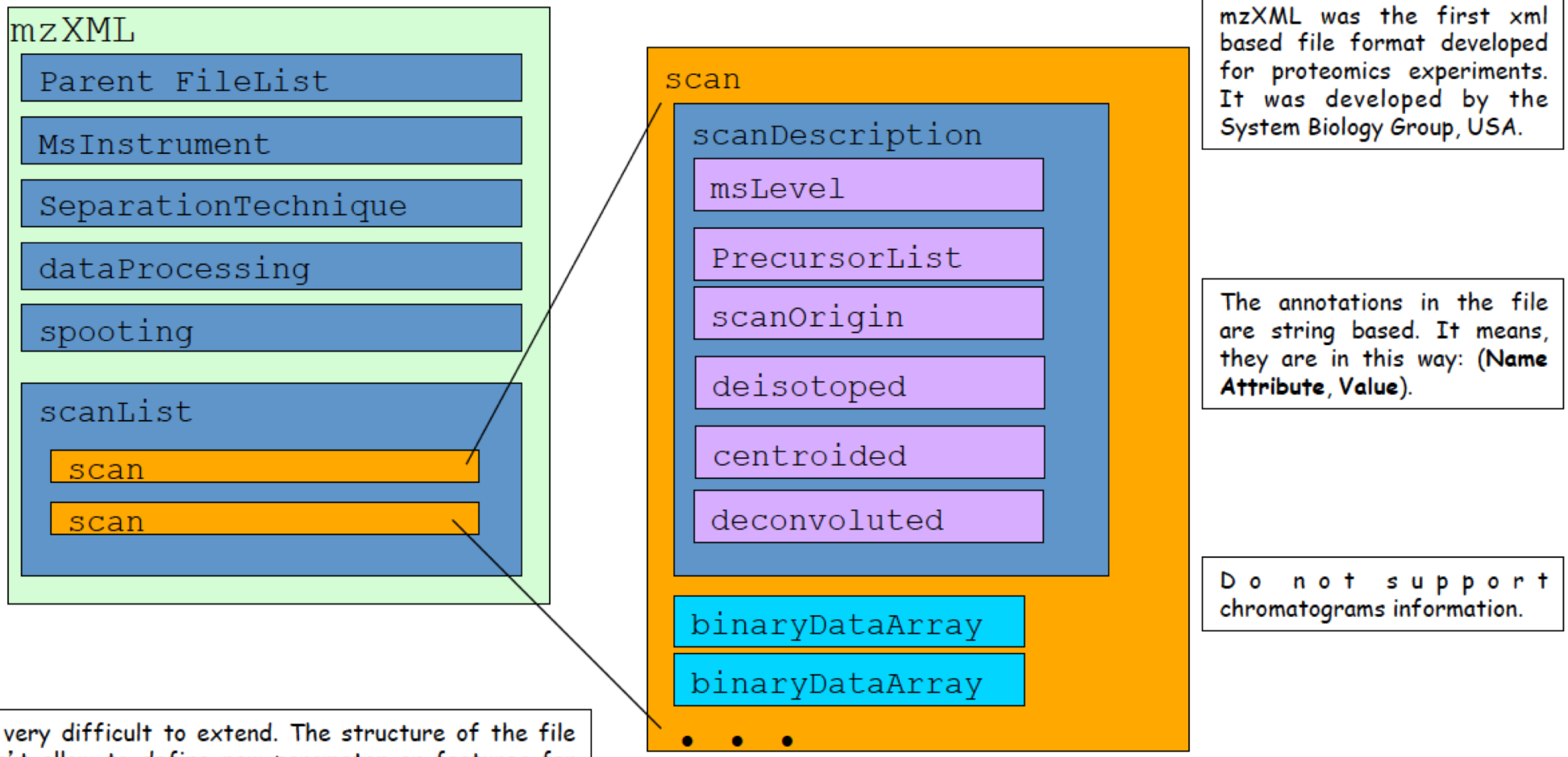
- Features you need to know:
  - Controlled structure
  - Nesting (multi-layered information)
  - Meta-data
  - Machine and human-readable syntax

Do not memorize the formats!
You only need to appreciate why it is better than list-based formats and basic features

XML-based

mzXML

Rich meta-data



mzXML was the first xml based file format developed for proteomics experiments. It was developed by the System Biology Group, USA.

The annotations in the file are string based. It means, they are in this way: (**Name Attribute**, **Value**).

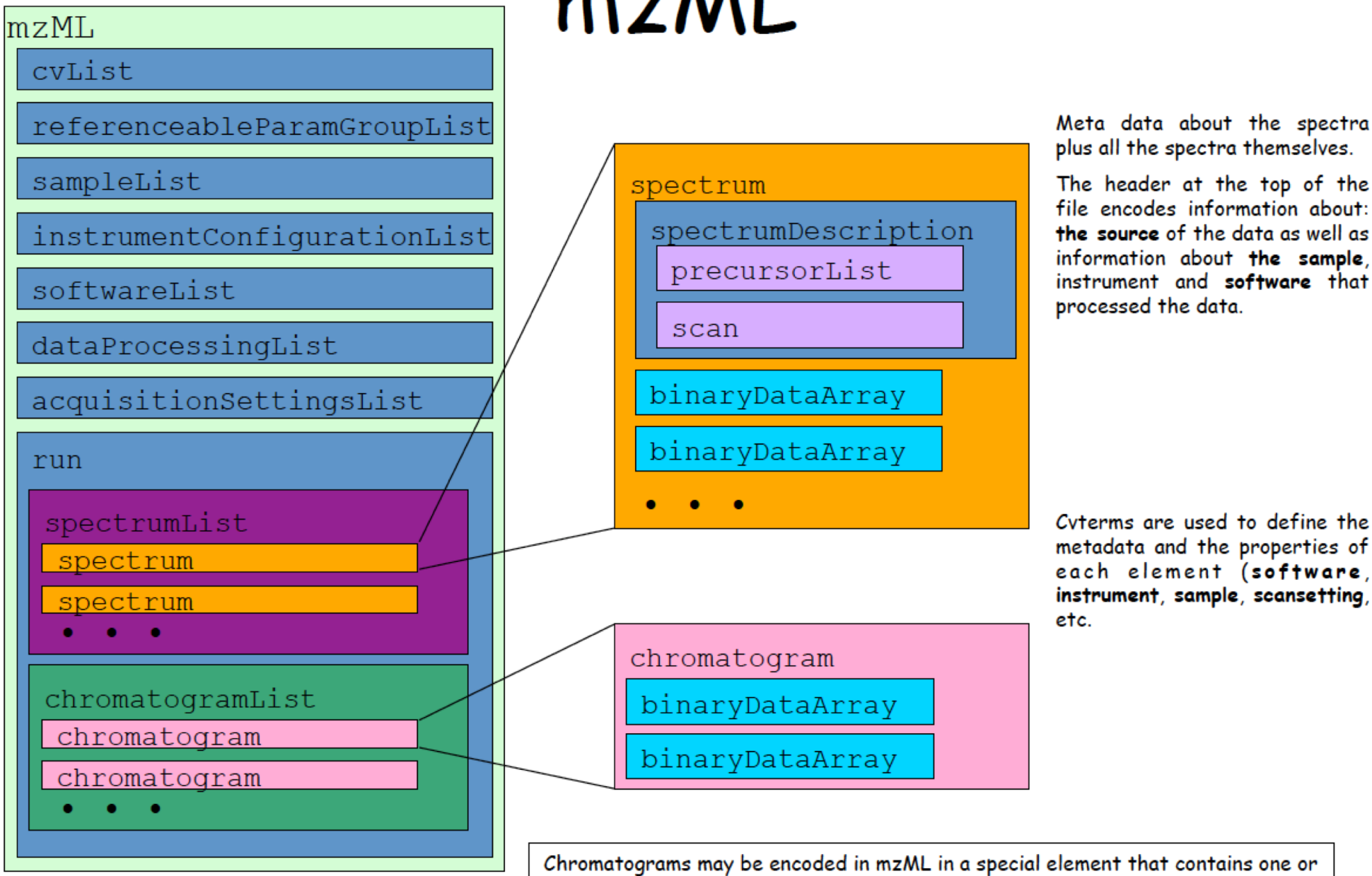
Do not support chromatograms information.

Is very difficult to extend. The structure of the file don't allow to define new parameter or features for each elements. For example, msInstrument are defined only by the name of the instrument. Also, if the spectrum is preprocessing with any program, is difficult to incorporate the information.

Actually exist more than 4 versions of the schema. The schema is supported by the System Biology Group, USA-Zurich.

XML-based

mzML



Meta data about the spectra plus all the spectra themselves.

The header at the top of the file encodes information about: **the source** of the data as well as information about **the sample**, instrument and **software** that processed the data.

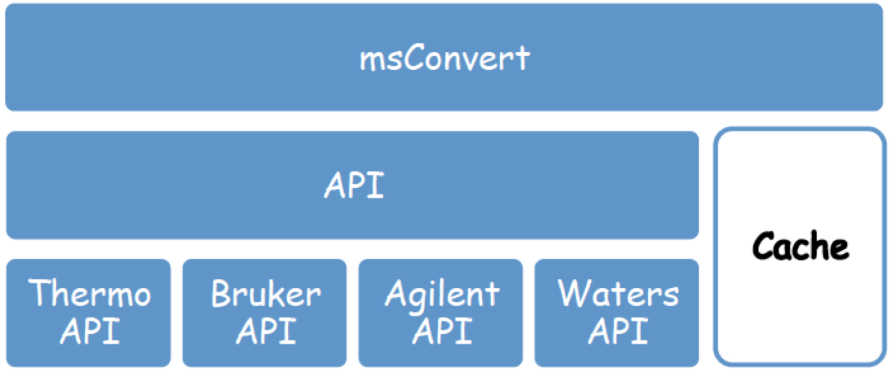
Cvterms are used to define the metadata and the properties of each element (**software**, **instrument**, **sample**, **scansetting**, etc.

Chromatograms may be encoded in mzML in a special element that contains one or more cvParams to describe the type of chromatogram, followed by two base64-encoded binary data arrays.

http://www.psidev.info/

Format conversion

If you want to use open source or analyze data from different platforms,
Then need to standardize format (to MZXML or MZML). Use ProteoWizard



File Input Supported:
- –Thermo
- –Bruker
- –Agilent
- –Waters
- –Pkl
- –mgf,
- –dta
- –ms2

Includes both proprietary and open formats

File Output Supported:
- –mzML
- –mzXML
- –mzData
- –Pkl
- –mgf

Cross-platform !!!!

BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 21 2008, pages 2534–2536
doi:10.1093/bioinformatics/btn323

Genome analysis

ProteoWizard: open source software for rapid proteomics tools development

Darren Kessner[1,*], Matt Chambers[2], Robert Burke[1], David Agus[1] and Parag Mallick[1,3,*]

[1]Spielberg Family Center for Applied Proteomics, Cedars-Sinai Medical Center, [2]Department of Biochemistry, Vanderbilt University, Nashville, TN and [3]Department of Chemistry & Biochemistry, University of California, Los Angeles, CA, USA

Received on April 18, 2008; revised on May 21, 2008; accepted on June 18, 2008

Advance Access publication July 7, 2008

Associate Editor: John Quackenbush

http://proteowizard.sourceforge.net/

Open analysis platforms

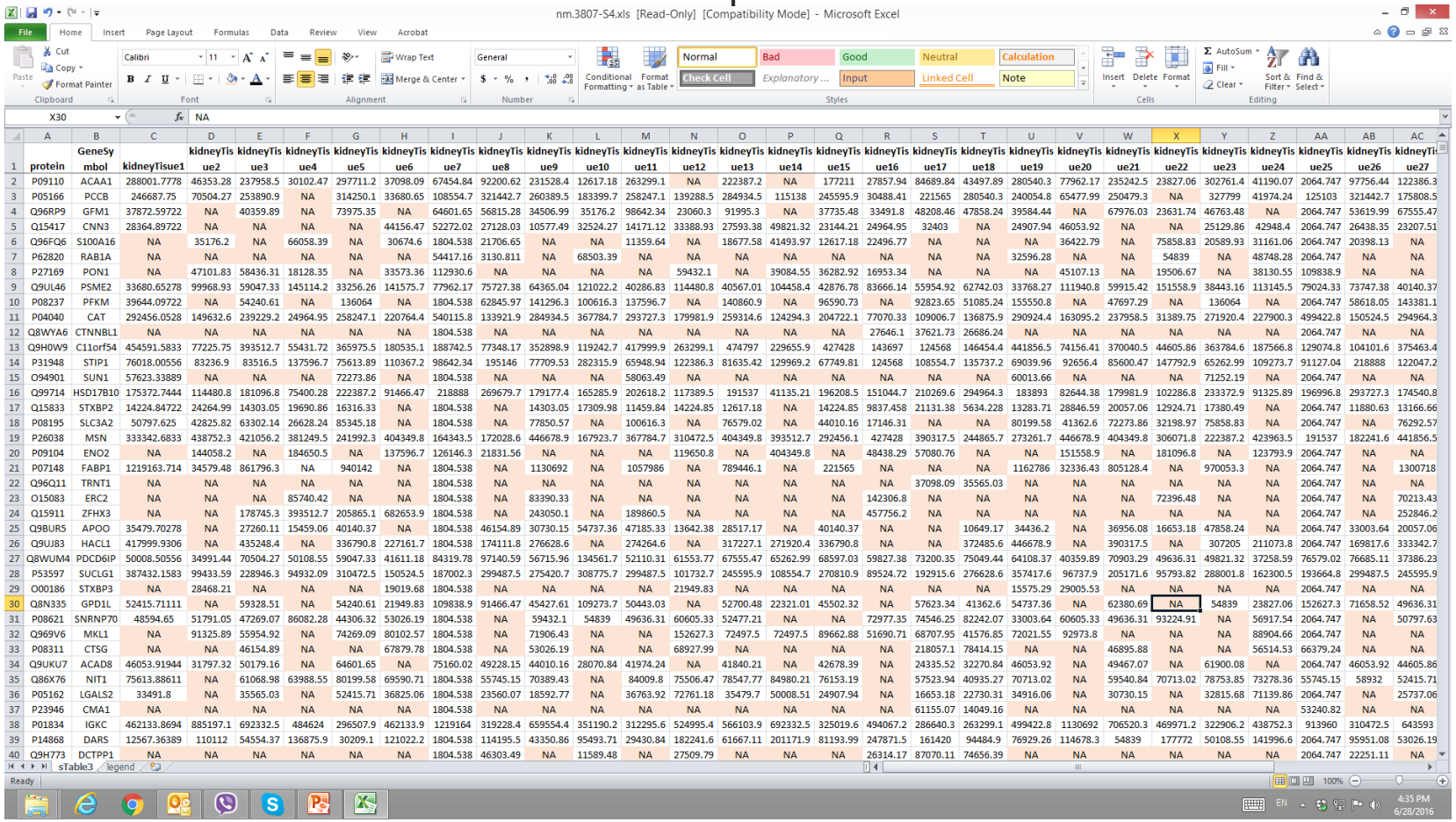
- OpenMS (https://www.openms.de/)
- Galaxy (https://galaxyproject.org/proteomics/)
- ExPASY (https://www.expasy.org/)
- TransProteomics Pipeline (http://tools.proteomecenter.org/)
- Maxquant (http://www.biochem.mpg.de/5111795/maxquant)

NANYANG TECHNOLOGICAL UNIVERSITY

Not quite there yet… what are we missing?

Samples

Proteins



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	protein	GeneSymbol	kidneyTisue1	kidneyTis ue2	kidneyTis ue3	kidneyTis ue4	kidneyTis ue5	kidneyTis ue6	kidneyTis ue7	kidneyTis ue8	kidneyTis ue9	kidneyTis ue10	kidneyTis ue11	kidneyTis ue12	kidneyTis ue13	kidneyTis ue14	kidneyTis ue15	kidneyTis ue16	kidneyTis ue17	kidneyTis ue18	kidneyTis ue19	kidneyTis ue20	kidneyTis ue21	kidneyTis ue22	kidneyTis ue23	kidneyTis ue24	kidneyTis ue25	kidneyTis ue26	kidneyTis ue27
2	P09110	ACAA1	288001.7778	46353.28	237958.5	30102.47	297711.2	37098.09	67454.84	92200.62	231528.4	12617.18	263299.1	NA	222387.2	NA	177211	27857.94	84689.84	43497.89	280540.3	77962.17	235242.5	23827.06	302761.4	41190.07	2064.747	97756.44	122386.3
3	P05166	PCCB	246687.75	70504.27	253890.9	NA	314250.1	33680.65	108554.7	321442.7	260389.5	183399.7	258247.1	139288.5	284934.5	115138	245595.9	30488.41	221565	280540.3	240054.8	65477.99	250479.3	NA	327799	41974.24	125103	321442.7	175808.5
4	Q96RP9	GFM1	37872.59722	NA	40359.89	NA	73975.35	NA	64601.65	56815.28	34506.99	35176.2	98642.34	23060.3	91995.3	NA	37735.48	33491.8	48208.46	47858.24	39584.44	NA	67976.03	23631.74	46763.48	NA	2064.747	53619.99	67555.47
5	Q15417	CNN3	28364.89722	NA	NA	NA	NA	44156.47	52272.02	27128.03	10577.49	32524.27	14171.12	33388.93	27593.38	49821.32	23144.21	24964.95	32403	NA	24907.94	46053.92	NA	NA	25129.86	42948.4	2064.747	26438.35	23207.51
6	Q96FQ6	S100A16	NA	35176.2	NA	66058.39	NA	30674.6	1804.538	21706.65	NA	NA	11359.64	NA	18677.58	41493.97	12617.18	22496.77	NA	NA	NA	36422.79	NA	75858.83	20589.93	31161.06	2064.747	20398.13	NA
7	P62820	RAB1A	NA	NA	NA	NA	NA	NA	54417.16	3130.811	NA	68503.39	NA	NA	NA	NA	NA	NA	NA	NA	32596.28	NA	NA	54839	NA	48748.28	2064.747	NA	NA
8	P27169	PON1	NA	47101.83	58436.31	18128.35	NA	33573.36	112930.6	NA	NA	NA	NA	59432.1	NA	39084.55	36282.92	16953.34	NA	NA	NA	45107.13	NA	19506.67	NA	38130.55	109838.9	NA	NA
9	Q9UL46	PSME2	33680.65278	99968.93	59047.33	145114.2	33256.26	141575.7	77962.17	75727.38	64365.04	121022.2	40286.83	114480.8	40567.01	104458.4	42876.78	83666.14	55954.92	62742.03	33768.27	111940.8	59915.42	151558.9	38443.16	113145.5	79024.33	73747.38	40140.37
10	P08237	PFKM	39644.09722	NA	54240.61	NA	136064	NA	1804.538	62845.97	141296.3	100616.3	137596.7	NA	140860.9	NA	96590.73	NA	92823.65	51085.24	155550.8	NA	47697.29	NA	136064	NA	2064.747	58618.05	143381.1
11	P04040	CAT	292456.0528	149632.6	239229.2	24964.95	258247.1	220764.4	540115.8	133921.9	284934.5	367784.7	293727.3	179981.9	259314.6	124294.3	204722.1	77070.33	109006.7	136875.9	290924.4	163095.2	237958.5	31389.75	271920.4	227900.3	499422.8	150524.5	294964.3
12	Q8WYA6	CTNNBL1	NA	NA	NA	NA	NA	NA	1804.538	NA	NA	NA	NA	NA	NA	NA	NA	27646.1	37621.73	26686.24	NA	NA	NA	NA	NA	NA	2064.747	NA	NA
13	Q9H0W9	C11orf54	454591.5833	77225.75	393512.7	55431.72	365975.5	180535.1	188742.5	77348.17	352898.9	119242.7	417999.9	263299.1	474797	229655.9	427428	143697	124568	146454.4	441856.5	74156.41	370040.5	44605.86	363784.6	187566.8	129074.8	104101.6	375463.4
14	P31948	STIP1	76018.00556	83236.9	83516.5	137596.7	75613.89	110367.2	98642.34	195146	77709.53	282315.9	65948.94	122386.3	81635.42	129969.2	67749.81	124568	108554.7	135737.2	69039.96	92656.4	85600.47	147792.9	65262.99	109273.7	91127.04	218888	122047.2
15	O94901	SUN1	57623.33889	NA	NA	NA	72273.86	NA	1804.538	NA	NA	NA	58063.49	NA	NA	NA	NA	NA	NA	NA	60013.66	NA	NA	NA	71252.19	NA	2064.747	NA	NA
16	Q99714	HSD17B10	175372.7444	114480.8	181096.8	75400.28	222387.2	91466.47	218888	269679.7	179177.4	165285.9	202618.2	117389.5	191537	41135.21	196208.5	151044.7	210269.6	294964.3	183893	82644.38	179981.9	102286.8	233372.9	91325.89	196996.8	293727.3	174540.8
17	Q15833	STXBP2	14224.84722	24264.99	14303.05	19690.86	16316.33	NA	1804.538	NA	14303.05	17309.98	11459.84	14224.85	12617.18	NA	14224.85	9837.458	21131.38	5634.228	13283.71	28846.59	20057.06	12924.71	17380.49	NA	2064.747	11880.63	13166.66
18	P08195	SLC3A2	50797.625	42825.82	63302.14	26628.24	85345.18	NA	1804.538	NA	77850.57	NA	100616.3	NA	76579.02	NA	44010.16	17146.31	NA	NA	80199.58	41362.6	72273.86	32198.97	75858.83	NA	2064.747	NA	76292.57
19	P26038	MSN	333342.6833	438752.3	421056.2	381249.5	241992.3	404349.8	164343.5	172028.6	446678.9	167923.7	367784.7	310472.5	404349.8	393512.7	292456.1	427428	390317.5	244865.7	273261.7	446678.9	404349.8	306071.8	222387.2	423963.5	191537	182241.6	441856.5
20	P09104	ENO2	NA	144058.2	NA	184650.5	NA	137596.7	126146.3	21831.56	NA	NA	NA	119650.8	NA	404349.8	NA	48438.29	57080.76	NA	NA	151558.9	NA	181096.8	NA	123793.9	2064.747	NA	NA
21	P07148	FABP1	1219163.714	34579.48	861796.3	NA	940142	NA	1804.538	NA	1130692	NA	1057986	NA	789446.1	NA	221565	NA	NA	NA	1162786	32336.43	805128.4	NA	970053.3	NA	2064.747	NA	1300718
22	Q96Q11	TRNT1	NA	NA	NA	NA	NA	NA	1804.538	NA	NA	NA	NA	NA	NA	NA	NA	NA	37098.09	35565.03	NA	NA	NA	NA	NA	NA	2064.747	NA	NA
23	O15083	ERC2	NA	NA	NA	85740.42	NA	NA	1804.538	NA	83390.33	NA	NA	NA	NA	NA	NA	142306.8	NA	NA	NA	NA	NA	72396.48	NA	NA	2064.747	NA	70213.43
24	Q15911	ZFHX3	NA	NA	178745.3	393512.7	205865.1	682653.9	1804.538	NA	243050.1	NA	189860.5	NA	NA	NA	NA	457756.2	NA	NA	NA	NA	NA	NA	NA	NA	2064.747	NA	252846.2
25	Q9BUR5	APOO	35479.70278	NA	27260.11	15459.06	40140.37	NA	1804.538	46154.89	30730.15	54737.36	47185.33	13642.38	28517.17	NA	40140.37	NA	NA	10649.17	34436.2	NA	36956.08	16653.18	47858.24	NA	2064.747	33003.64	20057.06
26	Q9UJ83	HACL1	417999.9306	NA	435248.4	NA	336790.8	227161.7	1804.538	174111.8	276628.6	NA	274264.6	NA	317227.1	271920.4	336790.8	NA	NA	372485.6	446678.9	NA	390317.5	NA	307205	211073.8	2064.747	169817.6	333342.7
27	Q8WUM4	PDCD6IP	50008.50556	34991.44	70504.27	50108.55	59047.33	41611.18	84319.78	97140.59	56715.96	134561.7	52110.31	61553.77	67555.47	65262.99	68597.03	59827.38	73200.35	75049.44	64108.37	40359.89	70903.29	49636.31	49821.32	37258.59	76579.02	76685.11	37386.23
28	P53597	SUCLG1	387432.1583	99433.59	228946.3	94932.09	310472.5	150524.5	187002.3	299487.5	275420.7	308775.7	299487.5	101732.7	245595.9	108554.7	270810.9	89524.72	192915.6	276628.6	357417.6	96737.9	205171.6	95793.82	288001.8	162300.5	193664.8	299487.5	245595.9
29	O00186	STXBP3	NA	28468.21	NA	NA	NA	19019.68	1804.538	NA	NA	NA	NA	21949.83	NA	NA	NA	NA	NA	NA	15575.29	29005.53	NA	NA	NA	NA	2064.747	NA	NA
30	Q8N335	GPD1L	52415.71111	NA	59328.51	NA	54240.61	21949.83	109838.9	91466.47	45427.61	109273.7	50443.03	NA	52700.48	22321.01	45502.32	NA	57623.34	41362.6	54737.36	NA	62380.69	NA	54839	23827.06	152627.3	71658.52	49636.31
31	P08621	SNRNP70	48594.65	51791.05	47269.07	86082.28	44306.32	53026.19	1804.538	NA	59432.1	54839	49636.31	60605.33	52477.21	NA	NA	72977.35	74546.25	82242.07	33003.64	60605.33	49636.31	93224.91	NA	56917.54	2064.747	NA	50797.63
32	Q969V6	MKL1	NA	91325.89	55954.92	NA	74269.09	80102.57	1804.538	NA	71906.43	NA	NA	152627.3	72497.5	72497.5	89662.88	51690.71	68707.95	41576.85	72021.55	92973.8	NA	NA	NA	88904.66	2064.747	NA	NA
33	P08311	CTSG	NA	NA	46154.89	NA	NA	67879.78	1804.538	NA	53026.19	NA	NA	68927.99	NA	NA	NA	NA	218057.1	78414.15	NA	NA	46895.88	NA	NA	56514.53	66379.24	NA	NA
34	Q9UKU7	ACAD8	46053.91944	31797.32	50179.16	NA	64601.65	NA	75160.02	49228.15	44010.16	28070.84	41974.24	NA	41840.21	NA	42678.39	NA	24335.52	32270.84	46053.92	NA	49467.07	NA	61900.08	NA	2064.747	46053.92	44605.86
35	Q86X76	NIT1	75613.88611	NA	61068.98	63988.55	80199.58	69590.71	1804.538	55745.15	70389.43	NA	84009.8	75506.47	78547.77	84980.21	76153.19	NA	57523.94	40935.27	70713.02	NA	59540.84	70713.02	78753.85	73278.36	55745.15	58932	52415.71
36	P05162	LGALS2	33491.8	NA	35565.03	NA	52415.71	36825.06	1804.538	23560.07	18592.77	NA	36763.92	72761.18	35479.7	50008.51	24907.94	NA	16653.18	22730.31	34916.06	NA	30730.15	NA	32815.68	71139.86	2064.747	NA	25737.06
37	P23946	CMA1	NA	NA	NA	NA	NA	NA	1804.538	NA	NA	NA	NA	NA	NA	NA	NA	NA	61155.07	14049.16	NA	NA	NA	NA	NA	53240.82	NA	NA	NA
38	P01834	IGKC	462133.8694	885197.1	692332.5	484624	296507.9	462133.9	1219164	319228.4	659554.4	351190.2	312295.6	524995.4	566103.9	692332.5	325019.6	494067.2	286640.3	263299.1	499422.8	1130692	706520.3	469971.2	322906.2	438752.3	913960	310472.5	643593
39	P14868	DARS	12567.36389	110112	54554.37	136875.9	30209.1	121022.2	1804.538	114195.5	43350.86	95493.71	29430.84	182241.6	61667.11	201171.9	81193.99	247871.5	161420	94484.9	76929.26	114678.3	54839	177772	50108.55	141996.6	2064.747	95951.08	53026.19
40	Q9H773	DCTPP1	NA	NA	NA	NA	NA	NA	1804.538	46303.49	NA	11589.48	NA	27509.79	NA	NA	NA	26314.17	87070.11	74656.39	NA	NA	NA	NA	NA	NA	2064.747	22251.11	NA

We haven't talked about quantitation!

What have we learnt?

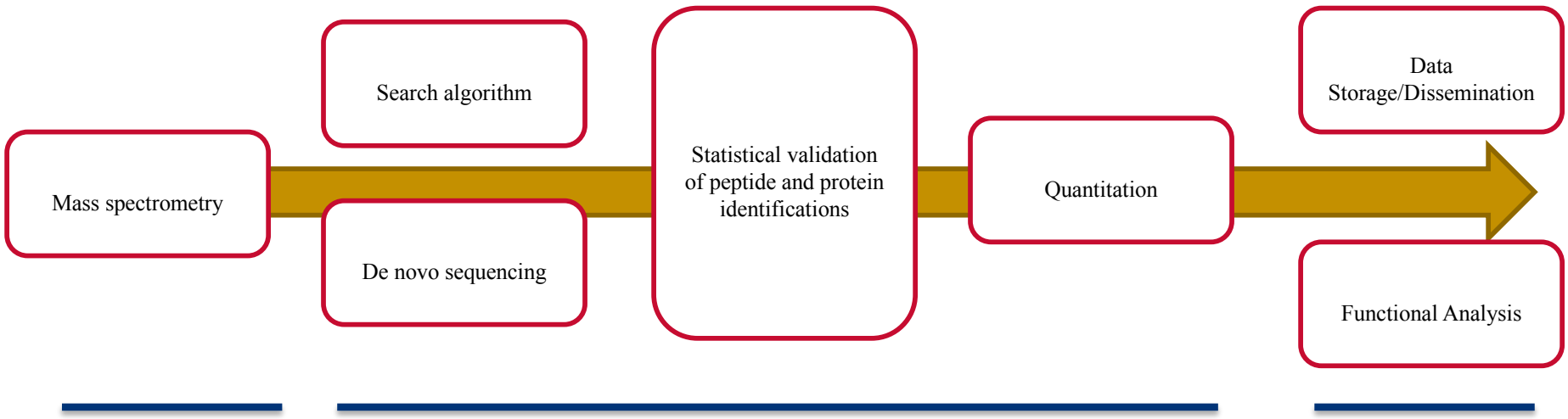
- Library search algorithms match theoretical spectra with observed spectra
- De novo sequencing algorithms use graph-theory methods to join observed spectra -> find longest path
- In proteomics, we observe peptides first. Proteins are inferred based on which peptides are detected
- A lot of useful information is lost during peptide to protein transition
- Proteomics data formats are diverse

You should now be able to

- Describe what an algorithm is, and how it differs from a heuristic or a computer program
- Describe the various levels of spectra data and their derivations in MS-based proteomics
- Describe the steps of a library search algorithm
- Describe the steps of a *de novo* sequencing algorithm
- Describe how peptides are assembled to proteins and associated problems
- Describe and evaluate the various levels and data representation formats in proteomics

NANYANG TECHNOLOGICAL UNIVERSITY

Putting things into perspective



Data storage/ representation formats

Peptide-spectra matching
Protein identification

Functional analysis

Peptide & protein identification by MS is still far from perfect

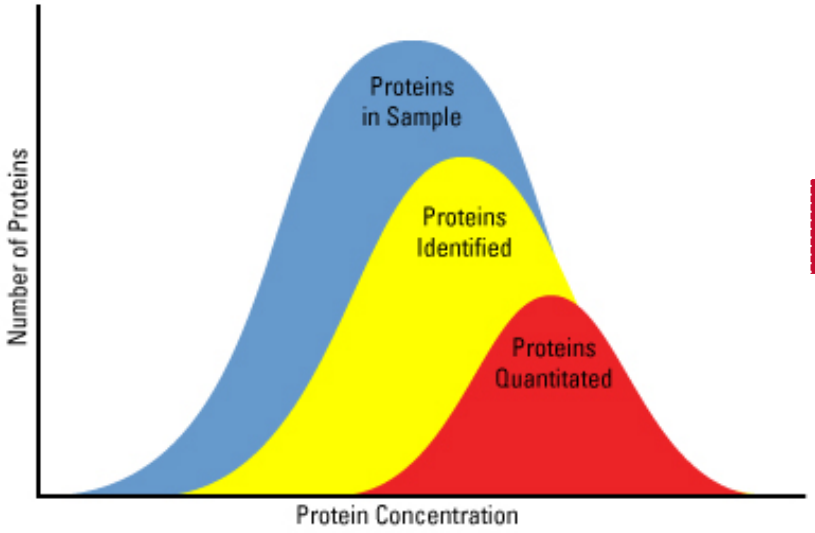
- "… peptides with low scores are, nevertheless, often correct, so manual validation of such hits can often 'rescue' the identification of important proteins."

Steen & Mann. **The ABC's and XYZ's of peptide sequencing**. *Nature Reviews Molecular Cell Biology*, 5:699-711, 2004

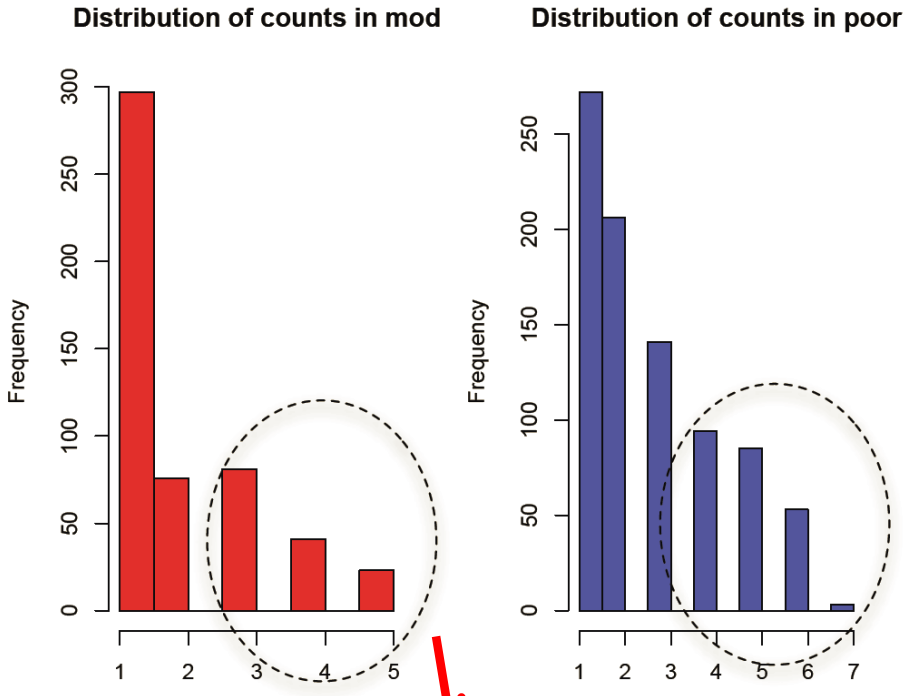
Issues in proteomics: Coverage and consistency

Technical incompleteness

How it affects real data



Per sample



Only 25 out of 800+ proteins are common to all 5 mod-stage HCC patients!

NANYANG TECHNOLOGICAL UNIVERSITY

Going further

- There is a kind of proteomics (Data-Independent Acquisition) where there is no one-to-one correspondence between MS1 and MS2. What kind of problems do you think can happen. And why do you think they avoided collecting MS1?

Hint: In traditional proteomics, MS1 peaks are semi-randomly selected for MS2 to ensure one-to-one correspondence. But what is the unintended consequence of this procedure?

Readings (Encouraged)

- Steen & Mann. The ABC's and XYZ's of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5:699-711, 2004
- Cottrell. Protein identification using MS/MS data. *Journal of Proteomics*, 74:1842-1851, 2011

Readings (Additional)

- Goh and Wong Spectra-first feature analysis in clinical proteomics—A case study in renal cancer, JBCB, 14 (05), 1644004, 2016
- Frank, et al. De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry. *J. Proteome Res*. 6:114-123, 2007
- Sung. Chap. 12: Peptide sequencing. *Algorithms in Bioinformatics: A Practical Introduction*. CRC Press, 2010
- Käll & Vitek. Computational mass spectrometry–based proteomics. *PLoS Comput Biol ,* 7(12): e1002277, 2011