SCIENCE MEETS LIFE

Algorithms for automatic spectra interpretation

Ralf Gabriels

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

UNIVERSITEIT GENT

Comp omics

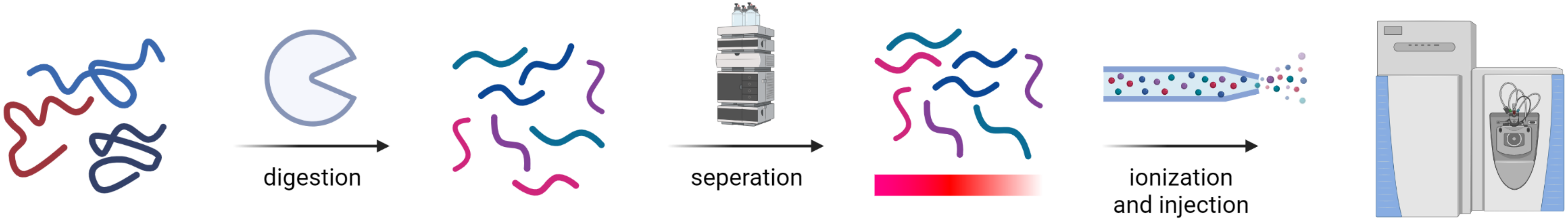
14/09/2021

1st PROTrEIN Summer School

Summary

- LC-MS/MS recap
- Database search engines: Replicating the LC-MS/MS workflow *in silico*
- Specialized methods
    - Open-modification searching
    - *De novo* and sequence tag-based searching
    - Spectral library searching

LC-MS/MS recap

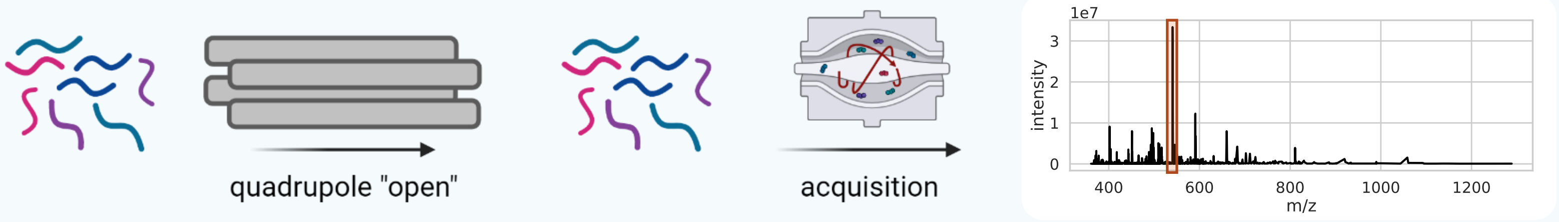


digestion

seperation

ionization and injection

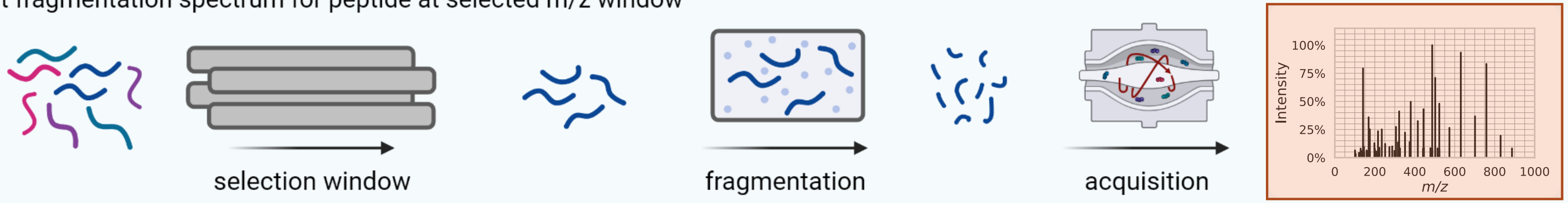
MS1 survey scan:
Find precursors (= full peptides)



quadrupole "open"

acquisition

MS2 scan:
Get fragmentation spectrum for peptide at selected m/z window



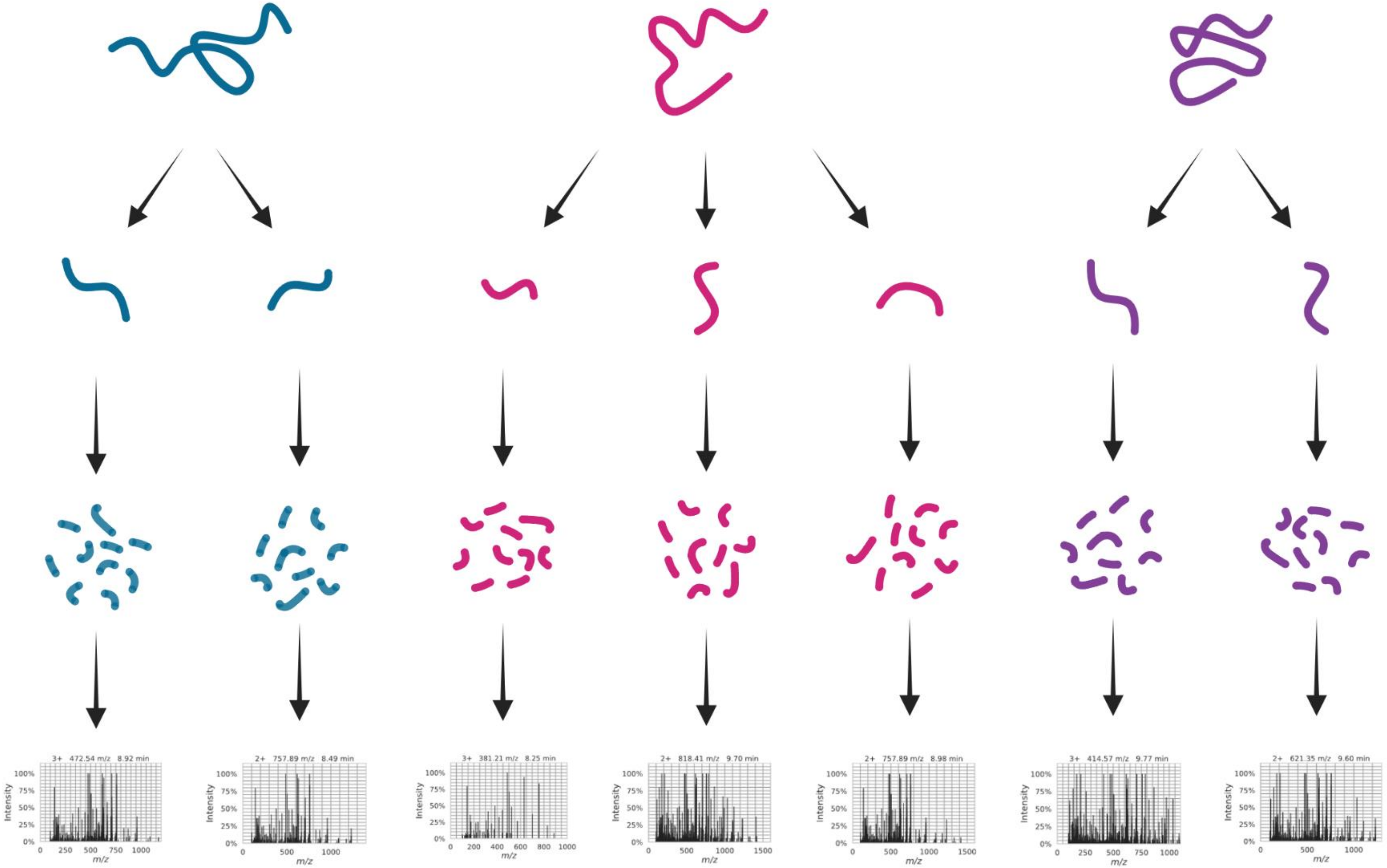
selection window

fragmentation

acquisition

Figures created in BioRender.com

LC-MS/MS recap



proteins

digestion

peptides

fragmentation

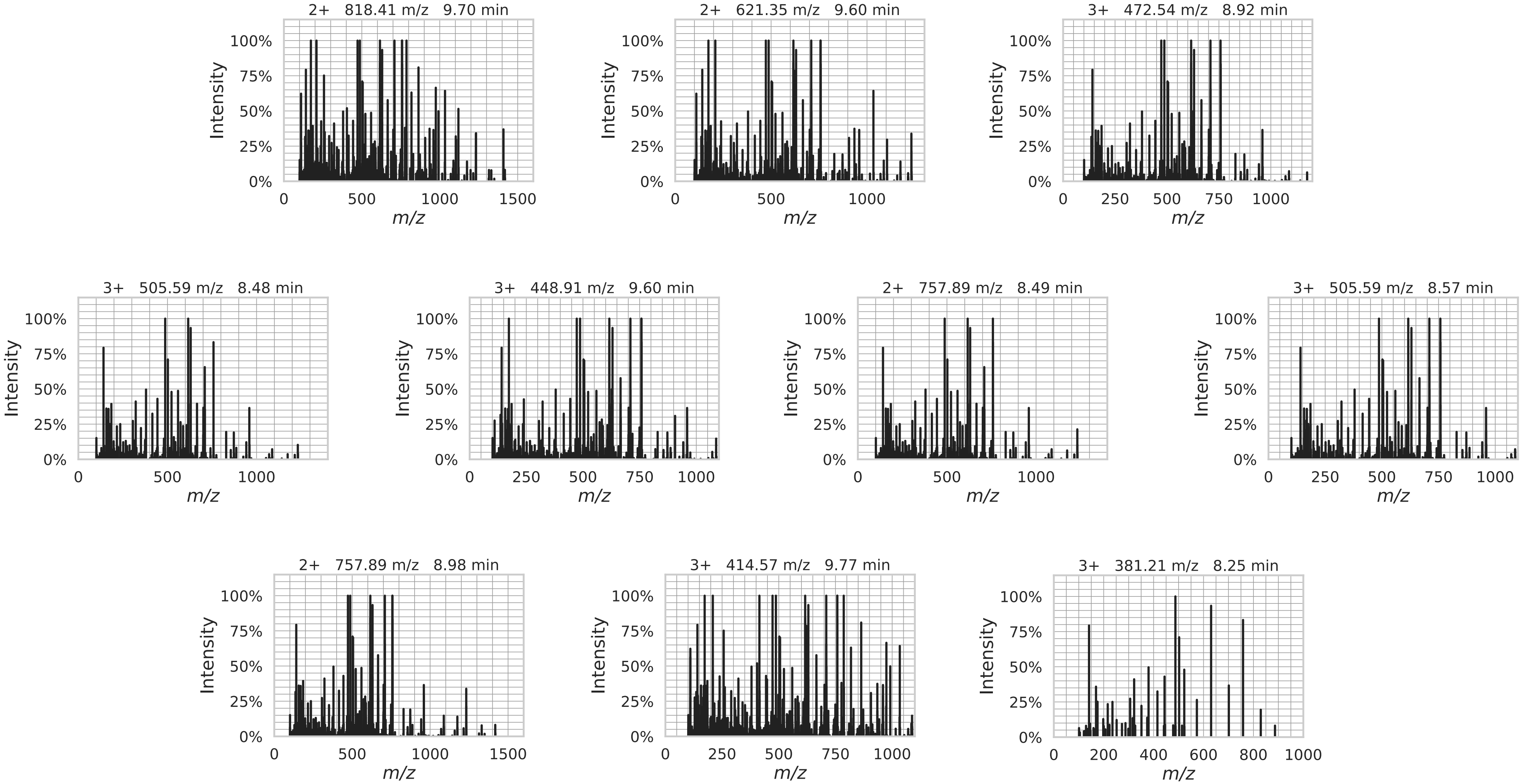
fragments ions

acquisition

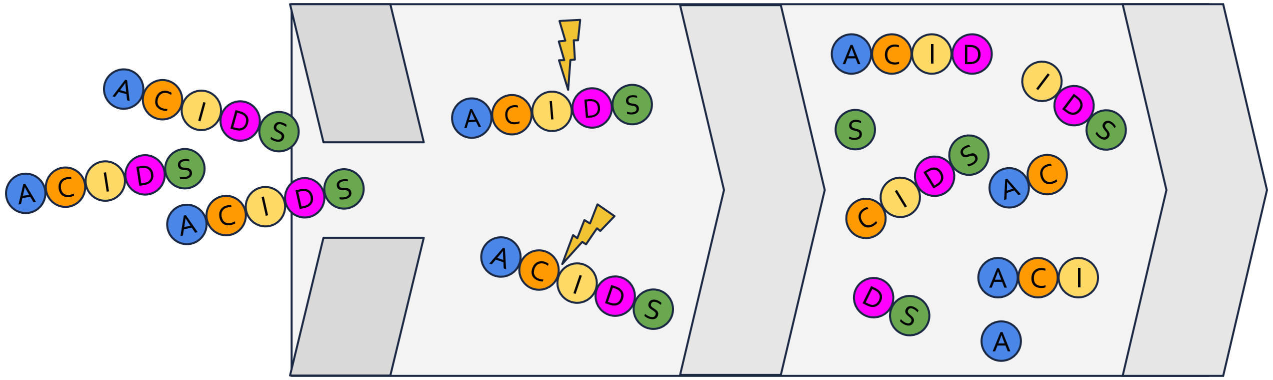
MS2 spectra

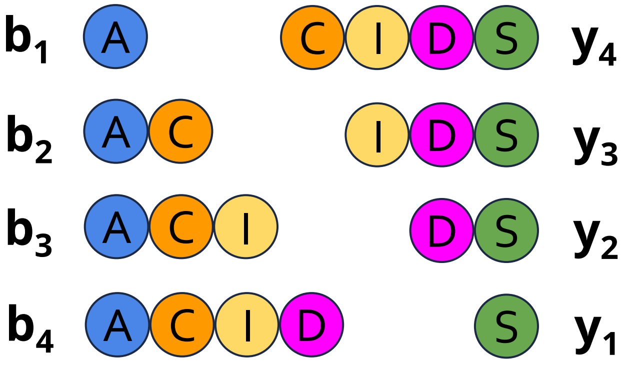
Figures created in BioRender.com

How do we link MS2 spectra back to the original proteins?

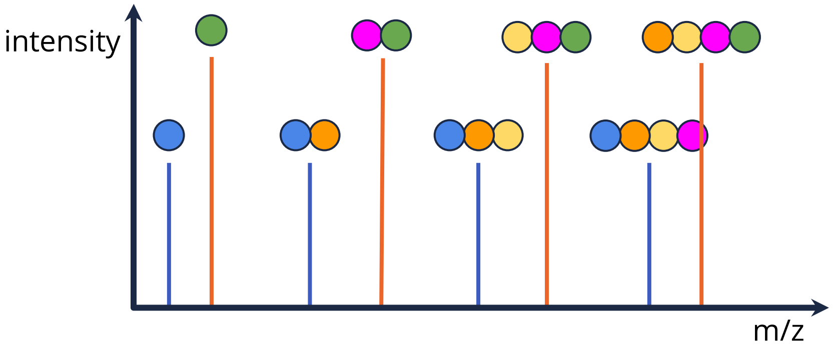


Peptides can be identified by their fragmentation spectra

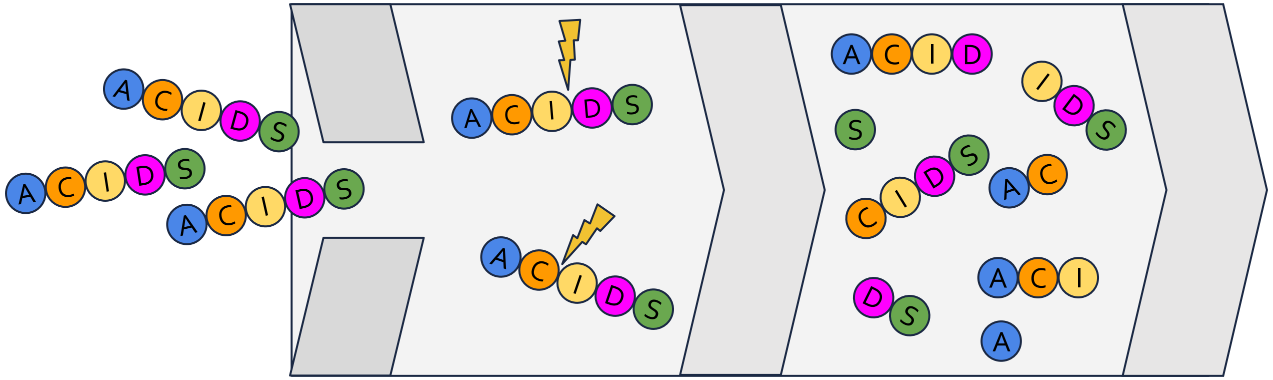


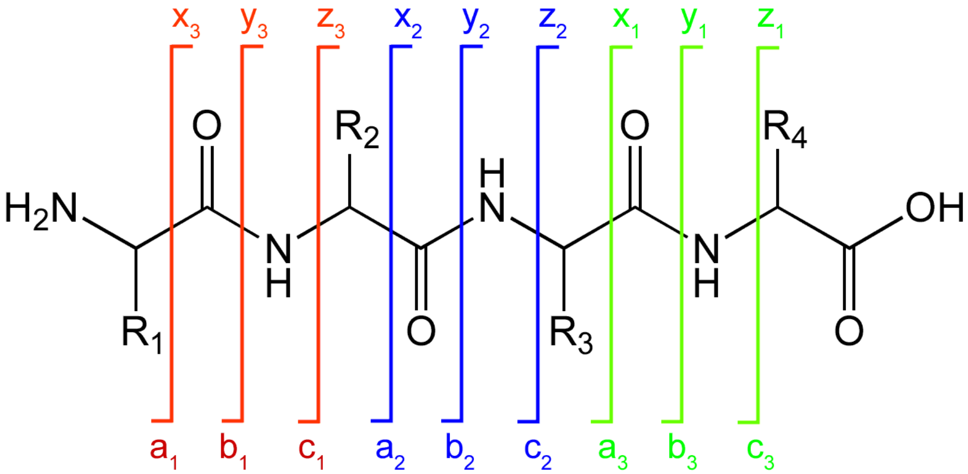


Mass calculation →

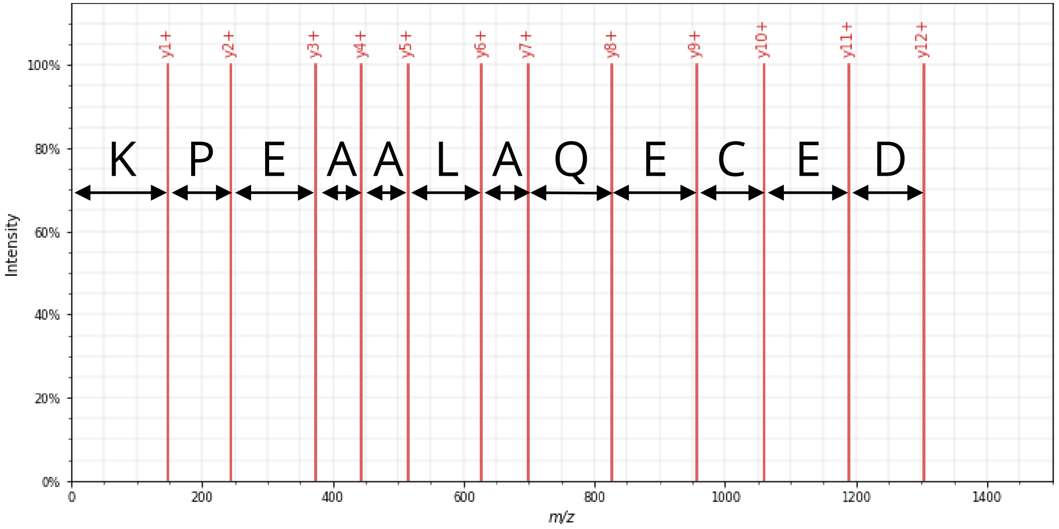


Peptides can be identified by their fragmentation spectra

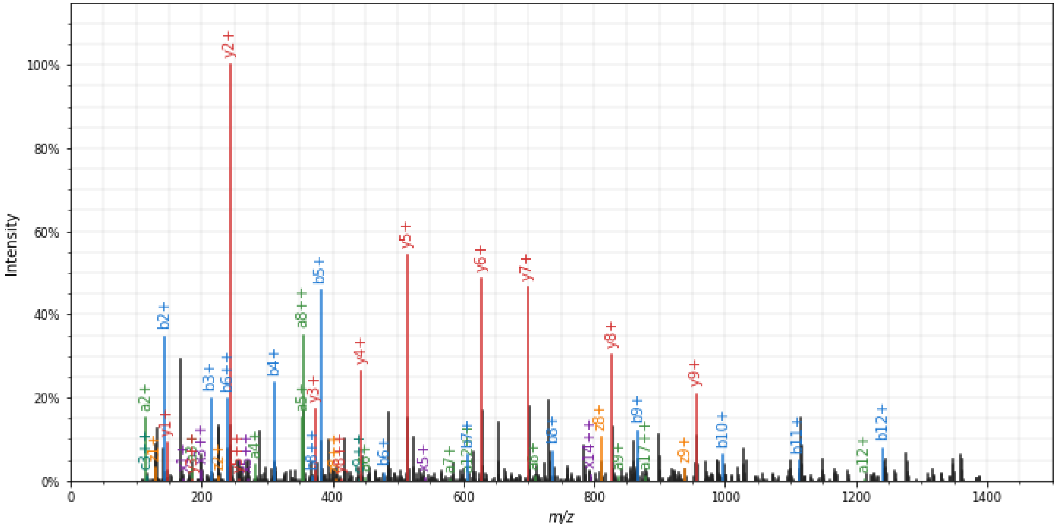




De novo peptide spectrum identification is not straightforward

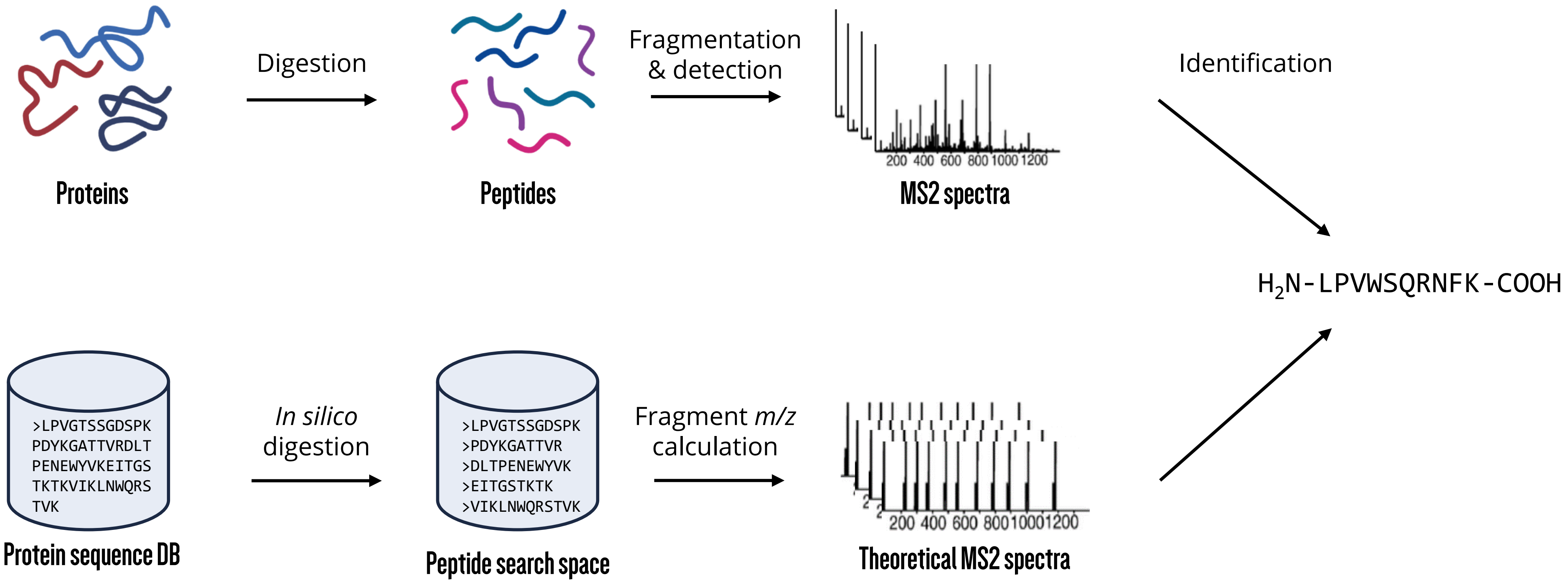


theory



reality

Database search engines replicate the LC-MS steps *in silico*



Proteins

Digestion

Peptides

Fragmentation & detection

MS2 spectra

Identification

H$_2$N-LPVWSQRNFK-COOH

```
>LPVGTSSGDSPK
PDYKGATTVRDLT
PENEWYVKEITGS
TKTKVIKLNWQRS
TVK
```

Protein sequence DB

In silico digestion

```
>LPVGTSSGDSPK
>PDYKGATTVR
>DLTPENEWYVK
>EITGSTKTK
>VIKLNWQRSTVK
```

Peptide search space

Fragment *m/z* calculation

Theoretical MS2 spectra

General proteomics search workflow

1. Define search space
2. *In silico* digestion of proteins to peptides
3. For each MS2 spectrum
    1. Select candidate peptide-spectrum matches (PSMs)
    2. Calculate theoretical peptide spectra
    3. Match candidate theoretical spectra to observed spectrum
    4. Select best match
4. PSM score post-processing

The search space should include ALL proteins that are expected to be present in the sample

UniProt

Reference proteome

- Swiss-Prot
- Swiss-Prot + TrEMBL

+

gpm

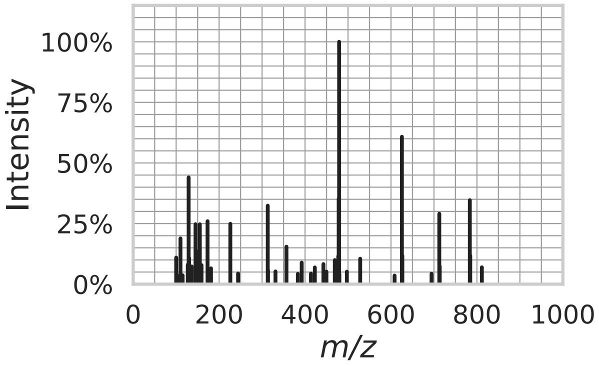
common **R**epository
of **A**dventitious **P**roteins

Common contaminants

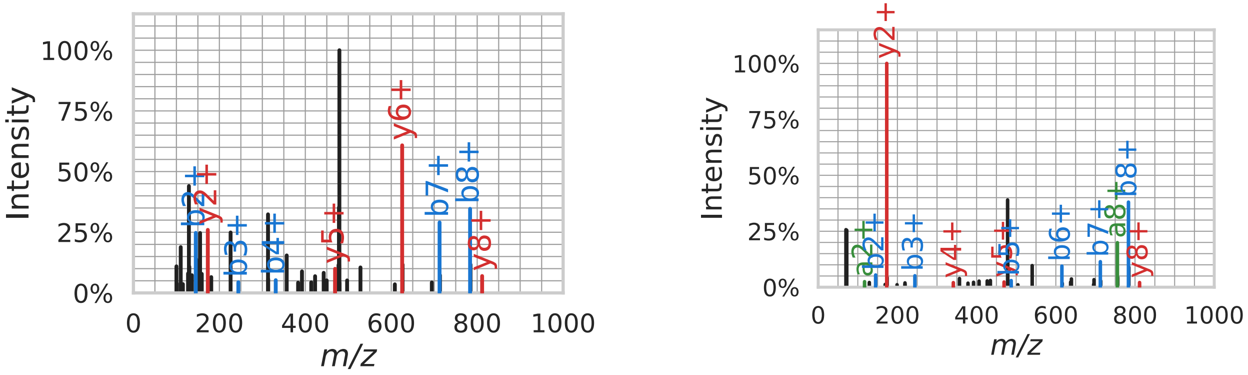
Search parameters
- Database selection

Sticker et al. 2017 (doi:10.1038/nmeth.4338)

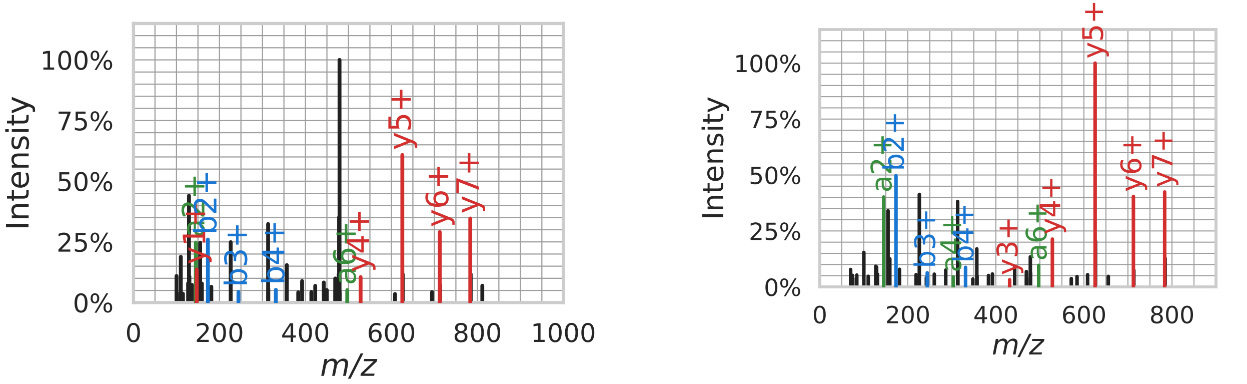
A mismatched search space can lead to false identifications



SGVSRKPAPG/2



ATASPPRQK/2



Publication: doi.org/10.1074/mcp.RA118.000877, Spectrum: mzspec:PXD000394:20130504_EXQ3_MiBa_SA_Fib-2:scan:4234

In silico digestion replicates the cleavage rules of the enzyme

```
>protein_sequence
LPVGTSSGDSPKPDYKGATTVRDLTPENEWYVKEITGSTKTKVI
KLNWQRSTVK
```

0 missed cleavages

```
LPVGTSSGDSPKPDYK
GATTVR
DLTPENEWYVK
EITGSTK
TK
VIK
LNWQR
STVK
```

1 missed cleavage

```
LPVGTSSGDSPKPDYKGATTVR
GATTVRDLTPENEWYVK
DLTPENEWYVKEITGSTK
EITGSTKTK
TKVIK
VIKLNWQR
LNWQRSTVK
```

2 missed cleavages

```
LPVGTSSGDSPKPDYKGATTVRDLTPENEWYVK
GATTVRDLTPENEWYVKEITGSTK
DLTPENEWYVKEITGSTKTK
EITGSTKTKVIK
TKVIKLNWQR
VIKLNWQRSTVK
```

Search parameters

- Allowed missed cleavages

In silico digestion replicates the cleavage rules of the enzyme

```
>protein
LPVGTSSGDSPKPDYKGATTVRDLTPENEWYVKEITGSTKTKVI
KLNWQRSTVK
```

0 missed cleavages

```
LPVGTSSGDSPKPDYK
GATTVR
DLTPENEWYVK
EITGSTK
TK
VIK
LNWQR
STVK
```

1 missed cleavage

```
LPVGTSSGDSPKPDYKGATTVR
GATTVRDLTPENEWYVK
DLTPENEWYVKEITGSTK
EITGSTKTK
TKVIK
VIKLNWQR
LNWQRSTVK
```

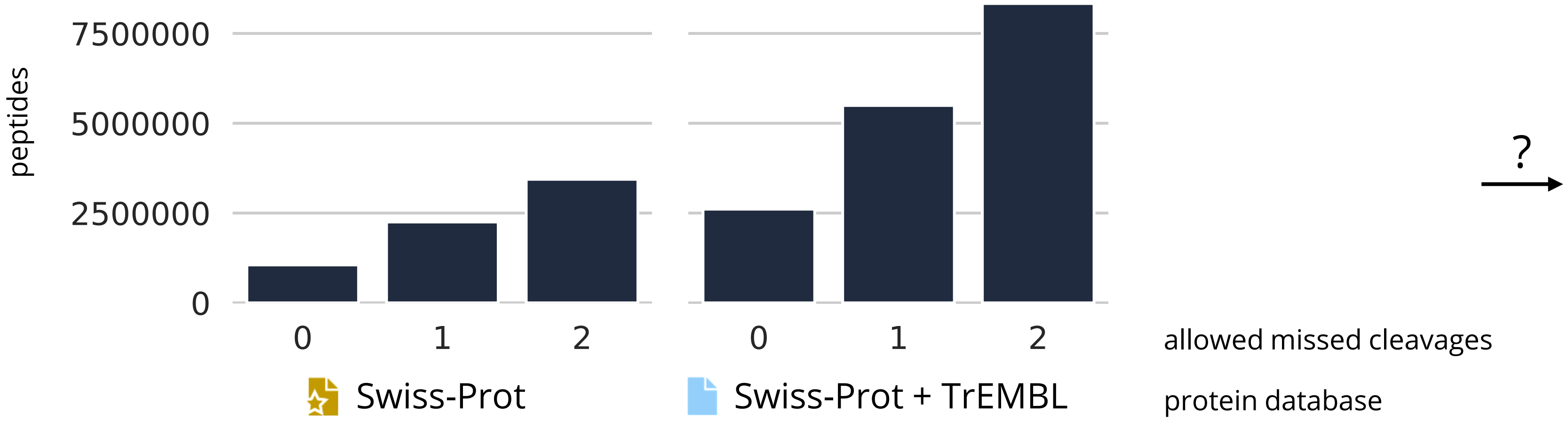
2 missed cleavages

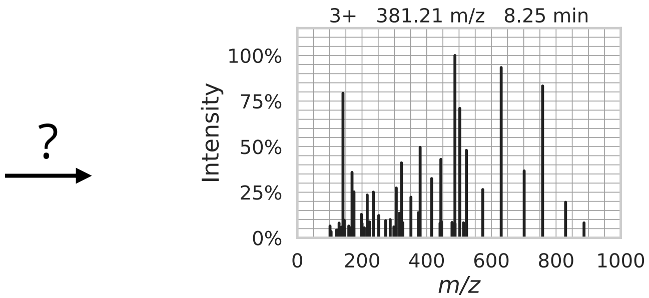
```
LPVGTSSGDSPKPDYKGATTVRDLTPENEWYVK
GATTVRDLTPENEWYVKEITGSTK
DLTPENEWYVKEITGSTKTK
EITGSTKTKVIK
TKVIKLNWQR
VIKLNWQRSTVK
```

Search parameters

- Allowed missed cleavages
- Minimum peptide length / mass
- Maximum peptide length / mass

Very large peptide search space needs to be filtered to candidate peptide-to-spectrum matches





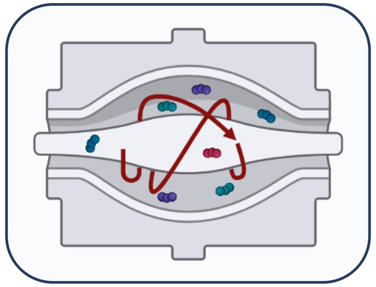
Computationally expensive
High probability of false positives

→ Filter to a limited set of PSMs

Candidate PSMs are usually selected by their precursor mass

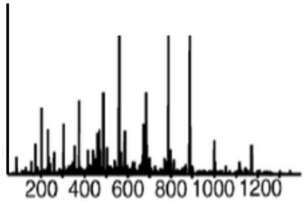
Peptide	Mass	+1 *m/z*	+2 *m/z*	+3 *m/z*
LPVGTSSGDSPKPDYK	1646.81	1647.82	824.41	549.49
GATTVR	603.33	604.34	302.67	202.11
DLTPENEWYVK	1392.65	1393.66	697.33	465.22
EITGSTK	734.38	735.38	368.19	245.80
…				

MS1

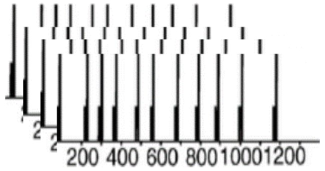


Search parameters
- Precursor mass window

For each candidate PSM, the theoretical spectrum is matched against the observed spectrum



Observed spectrum



Theoretical spectra for candidate PSMs

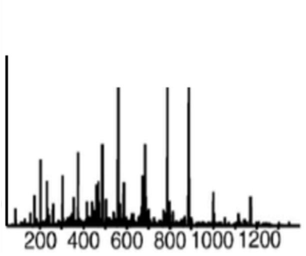
Scoring functions

- Explained intensity (e.g. SEQUEST)
- Peak counting (e.g. Mascot, Andromeda)
- Combinations (e.g. X!Tandem)

PSM scores for each candidate PSM

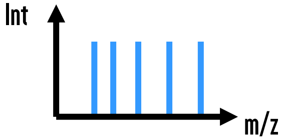
Select best-scoring PSM

The SEQUEST scoring function calculates the total intensity that can be explained by the theoretical spectrum



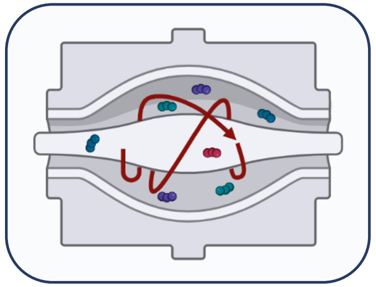
Observed spectrum

$$R_i = \sum_{j=1}^{n} x_j \cdot y_{(j+i)}$$



R_0

MS2



Search parameters

- Fragment mass window

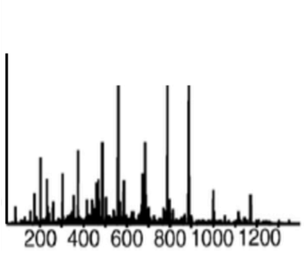
Resolution and accuracy: Fragment peak error tolerance

ion trap 0.5 Da

orbitrap 0.02 Da

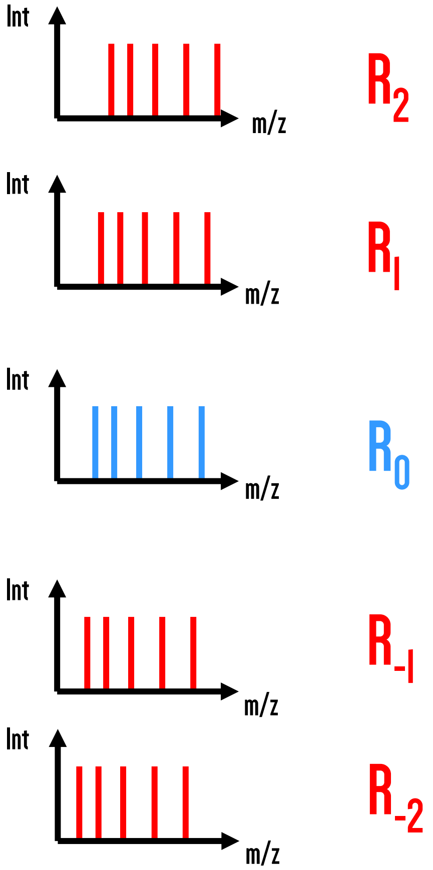
Eng, JASMS 1994

The SEQUEST scoring function calculates the total intensity that can be explained by the theoretical spectrum



Observed spectrum

$$R_i = \sum_{j=1}^{n} x_j \cdot y_{(j+i)}$$

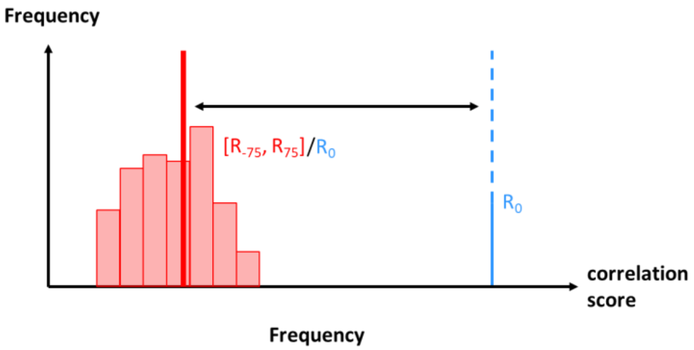


$$\sum_{i=-75}^{+75} Ri$$

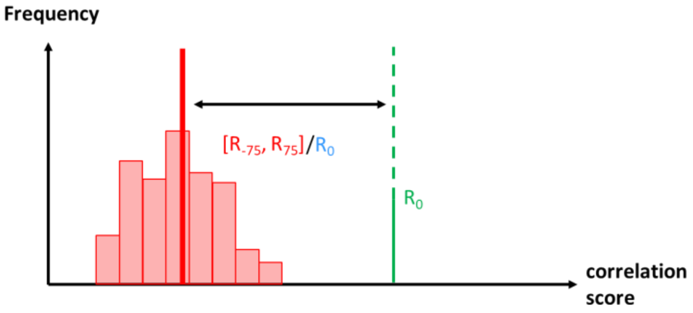
Eng, JASMS 1994

This score is then calibrated to the distribution of random matches, then the difference between the best and second-best score is calculated

XCorr_1



XCorr_2



$$\text{deltaCn} = \frac{XCorr_1 - XCorr_2}{XCorr_1}$$

$\downarrow$

Resulting PSM score for best PSM

Eng, JASMS 1994

The raw output of a search engine is a list of scores for the best-scoring PSM for each spectrum

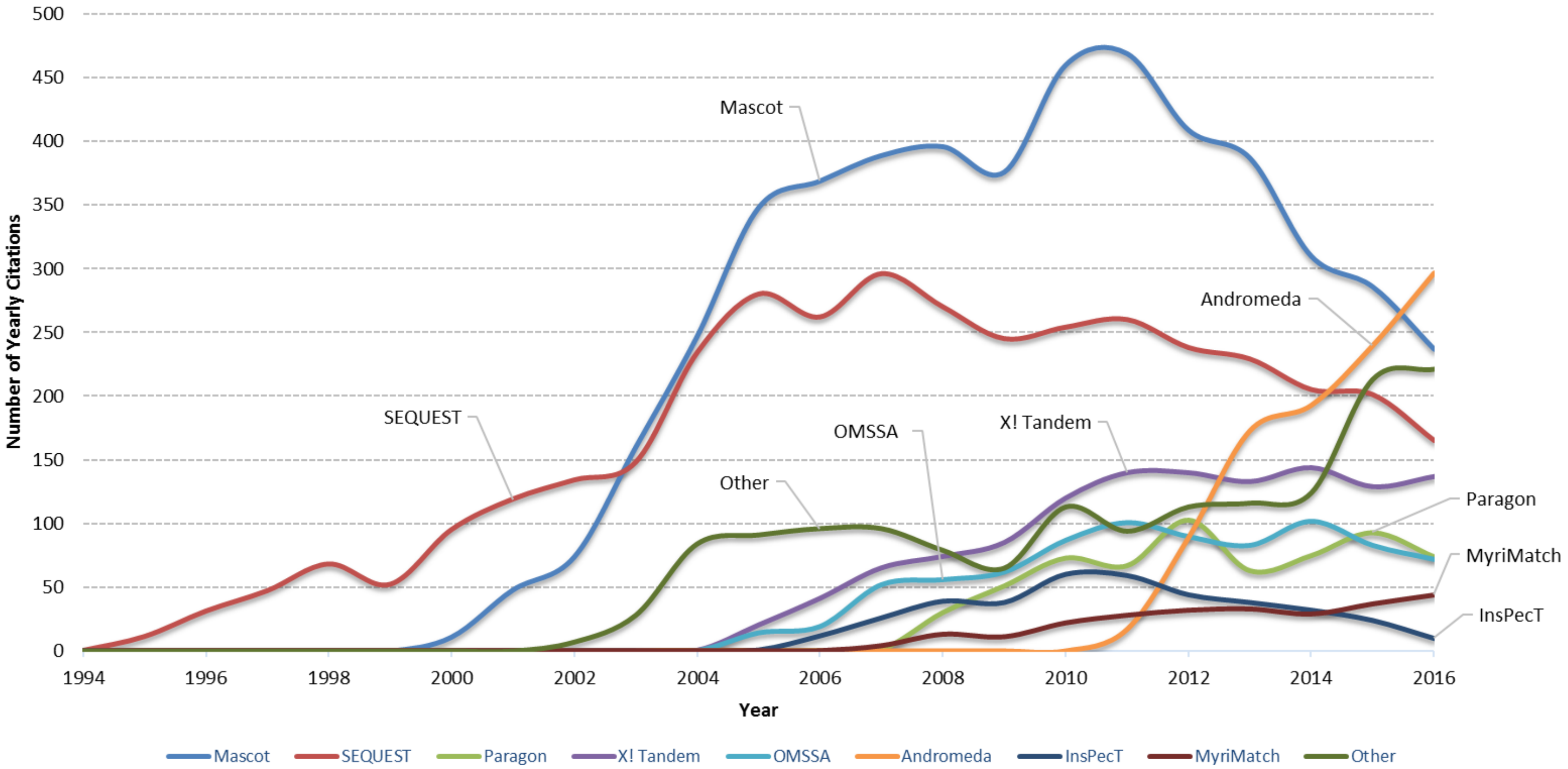
Spectrum	Best PSM	Score
1	LPVGTSSGDSPKPDYK	34.12
2	GATTVR	95.43
3	DLTPENEWYVK	134.87
4	EITGSTK	12.9
…	…	245.67

Now what? Which scores are good? Which are bad?

General proteomics search workflow

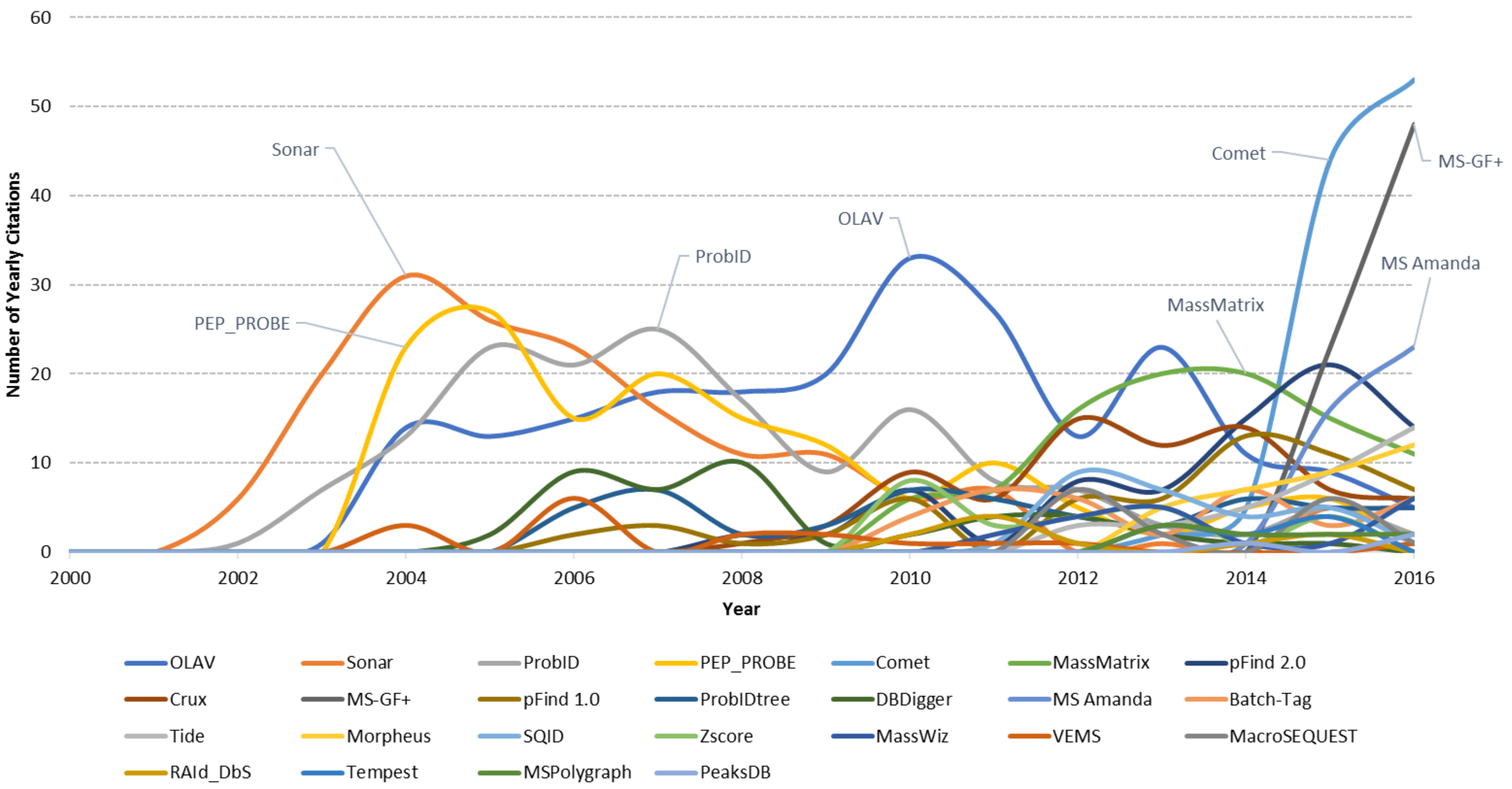
1. Define search space
2. *In silico* digestion of proteins to peptides
3. For each spectrum
    1. Select candidate peptide-spectrum matches (PSMs)
    2. Calculate theoretical peptide spectra
    3. Match candidate theoretical spectra to observed spectrum
    4. Select best match
4. PSM score post-processing: Next lecture

Historically, Mascot and SEQUEST were heavily used; now Andromeda, Comet, MS-GF+, and MS-Amanda are more favored.



Verheggen, Mass Spec Reviews, 2017

Historically, Mascot and SEQUEST were heavily used; now Andromeda, Comet, MS-GF+, and MS-Amanda are more favored.

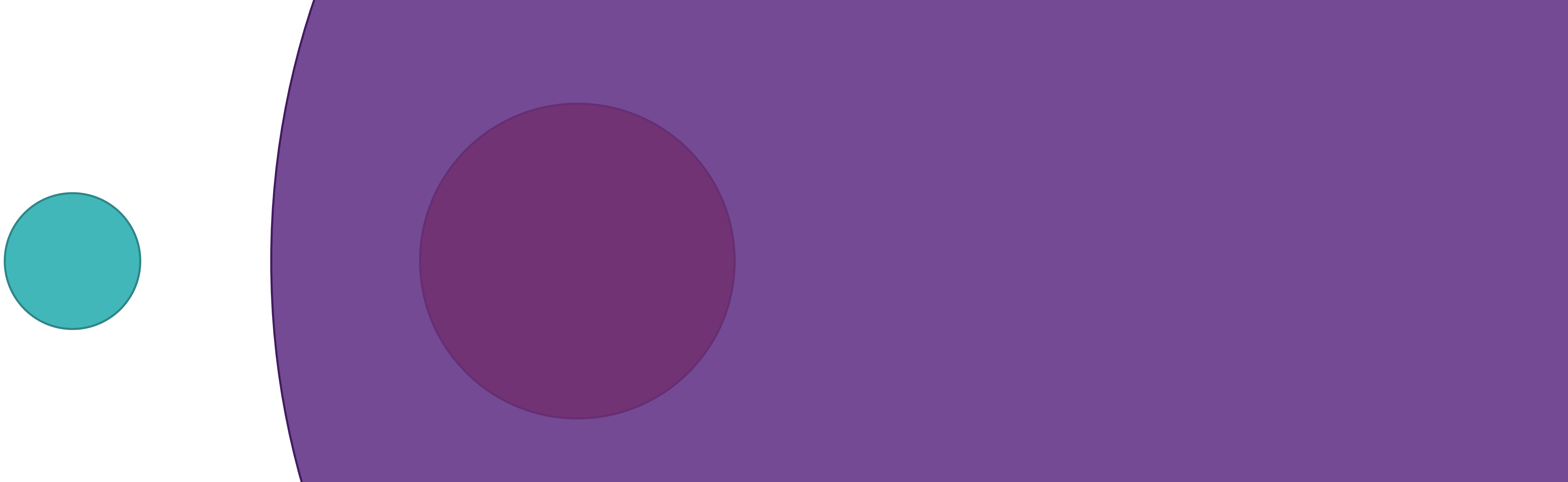


Verheggen, Mass Spec Reviews, 2017

Summary

- LC-MS/MS recap
- Database search engines: Replicating the LC-MS/MS workflow *in silico*
- Specialized methods
  - Open-modification searching
  - *De novo* and sequence tag-based searching
  - Spectral library searching

A few modifications can be added to the search space. To identify more PTMS, open-modification search engines are required.



No variable modifications

Common modifications

All known modifications

MSFragger
open-pFind
TagGraph
ionbot
...

Combinatorial explosion

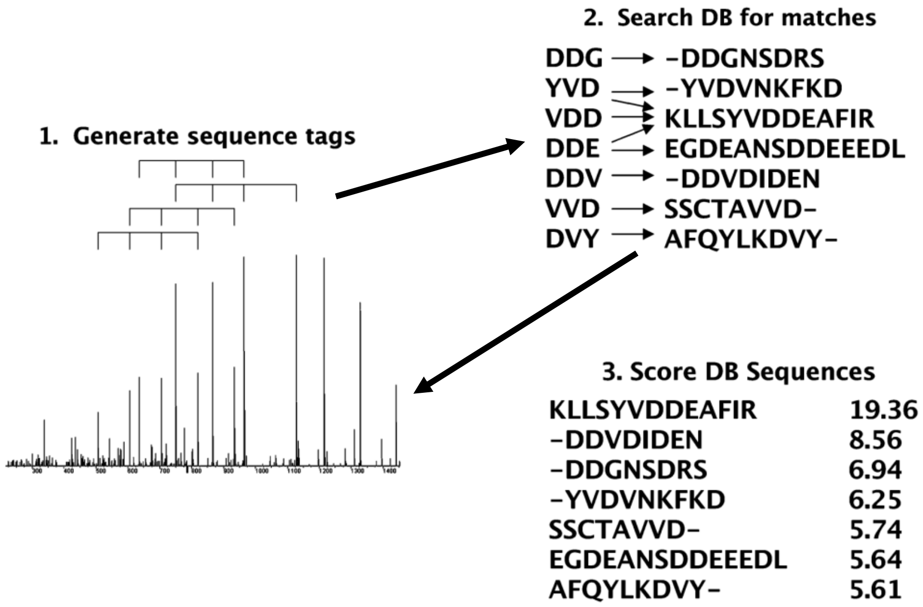
Sequence tag-based and *de novo* searching

Tag-based
DirecTag
TagGraph
ionbot
…

Find partial sequences in spectra
Usually to limit search space
Open modification searching!

De novo
Novor
PepNovo
PEAKS
…

No database at all
Try to elude full sequence
Rare / unknown species

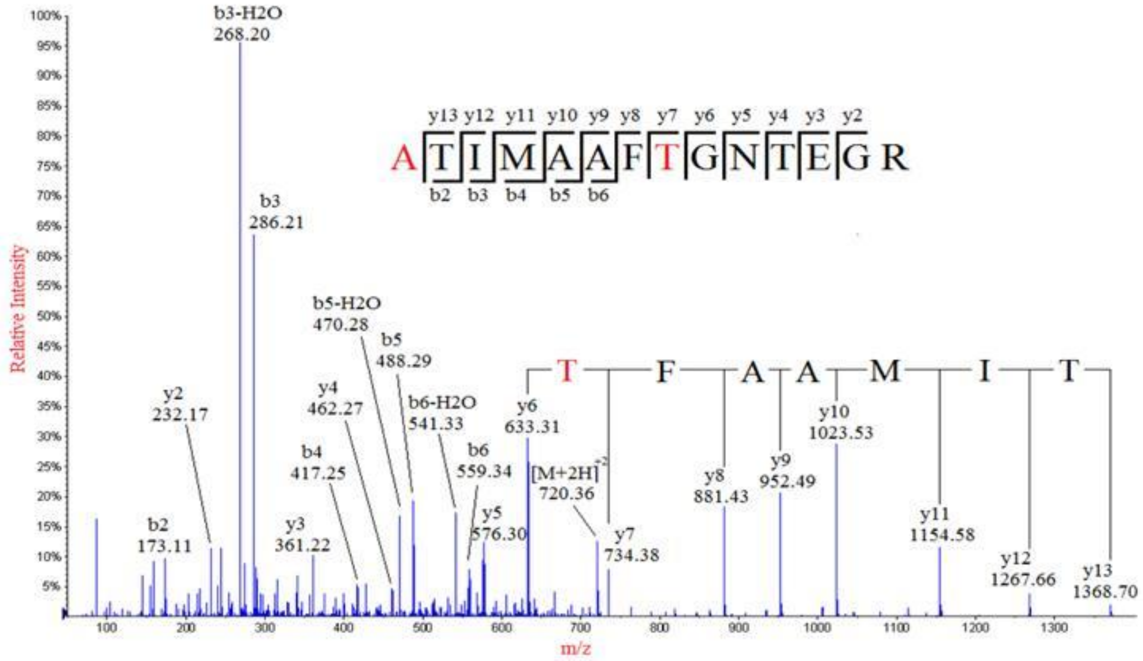

1. Generate sequence tags

2. Search DB for matches

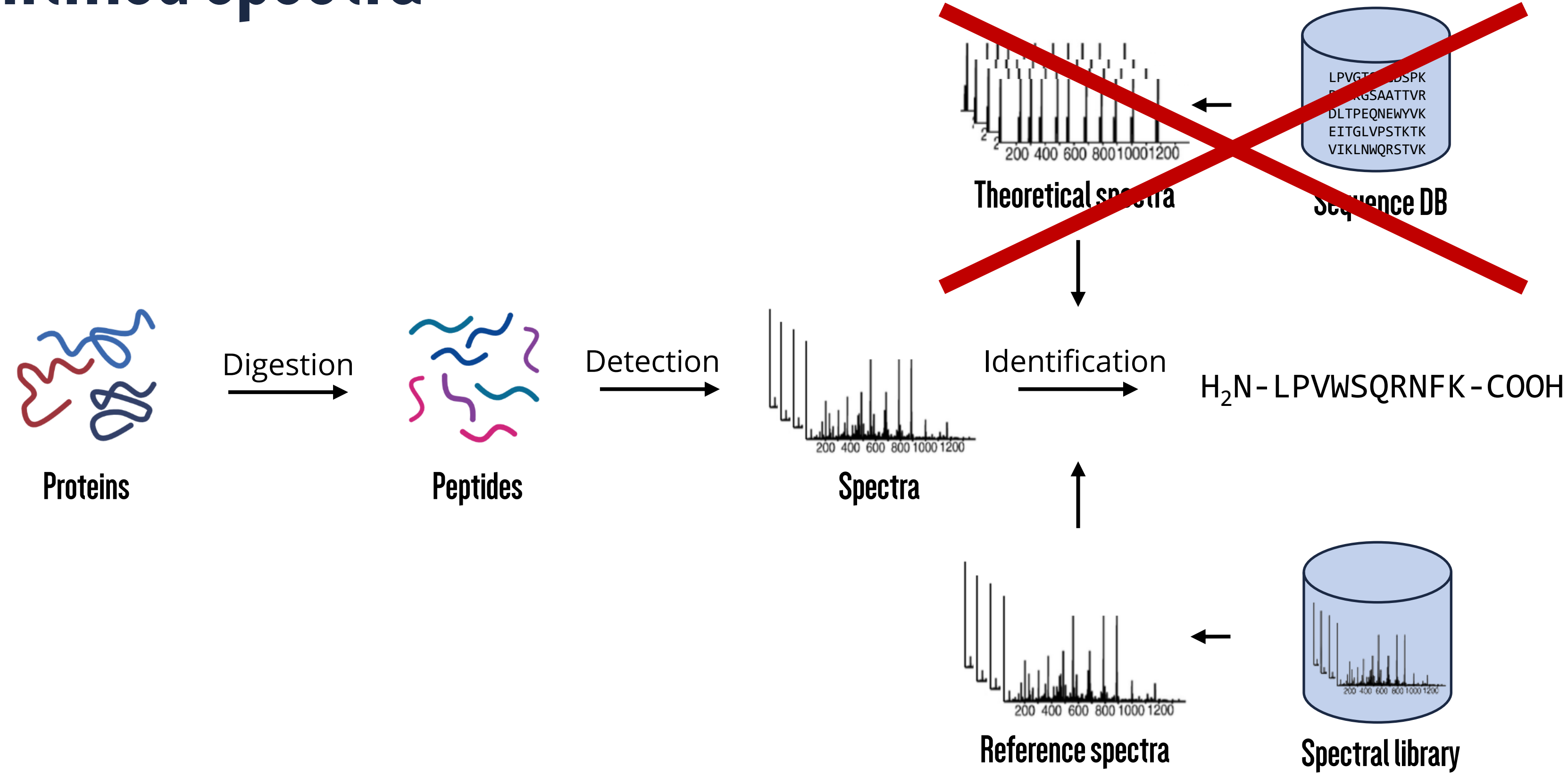
DDG → -DDGNSDRS
YVD → -YVDVNKFKD
VDD → KLLSYVDDEAFIR
DDE → EGDEANSDDEEEDL
DDV → -DDVDIDEN
VVD → SSCTAVVD-
DVY → AFQYLKDVY-

3. Score DB Sequences

KLLSYVDDEAFIR	19.36
-DDVDIDEN	8.56
-DDGNSDRS	6.94
-YVDVNKFKD	6.25
SSCTAVVD-	5.74
EGDEANSDDEEEDL	5.64
AFQYLKDVY-	5.61



Spectral library search engines match spectra against previously identified spectra



+ more sensitive scoring

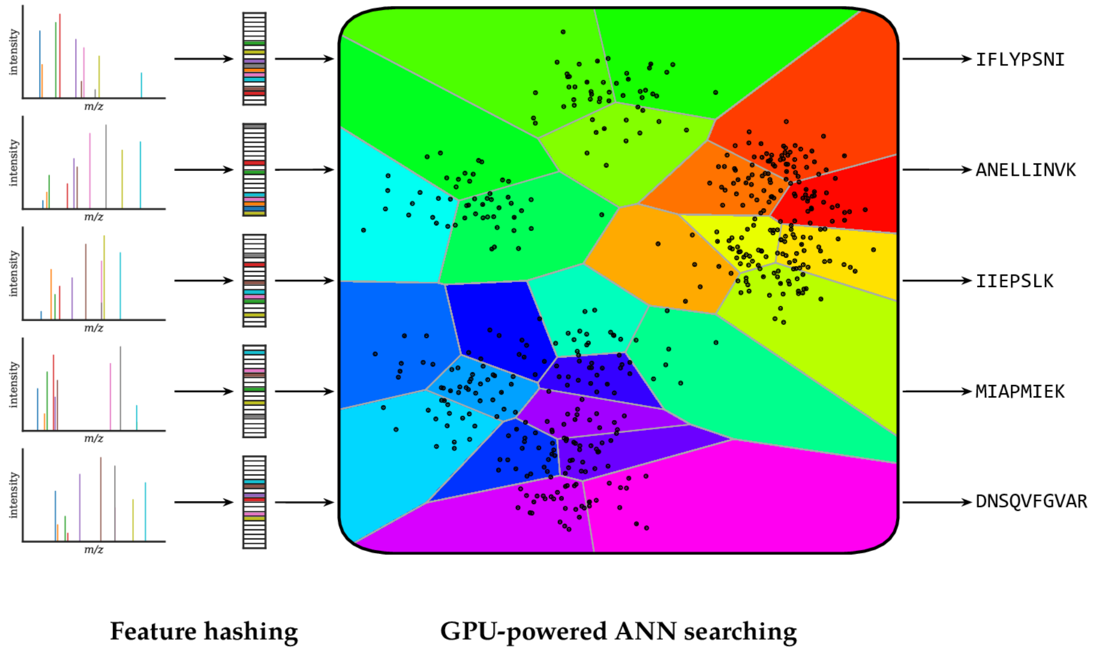
- only identifies what has been seen before

Spectral library search engines match spectra against previously identified spectra

SpectraST
Trans-proteomic pipeline (TPP)

COSS
CompOmics Spectral Library Searching

ANN-SoLo



Lam, Proteomics 2007, doi:10.1002/pmic.200600625; Shiferaw, JPR 2020 doi:10.1021/acs.jproteome.9b00743; Bittremieux, JPR 2018, doi:10.1021/acs.jproteome.8b00359

Summary

- LC-MS/MS recap
- Database search engines: Replicating the LC-MS/MS workflow *in silico*
- Specialized methods
  - Open-modification searching
  - *De novo* and sequence tag-based searching
  - Spectral library searching

VIB-UGENT CENTER
FOR MEDICAL
BIOTECHNOLOGY

GHENT
UNIVERSITY

Comp
omics

www.compomics.com
@CompOmics

Comp
omics

VIB-UGENT CENTER
FOR MEDICAL
BIOTECHNOLOGY

UNIVERSITEIT
GENT

Ralf Gabriels

Ralf.Gabriels@UGent.be

@RalfGabriels | @CompOmics

www.compomics.com