COMPUTATIONAL PROTEOMICS AND METABOLOMICS

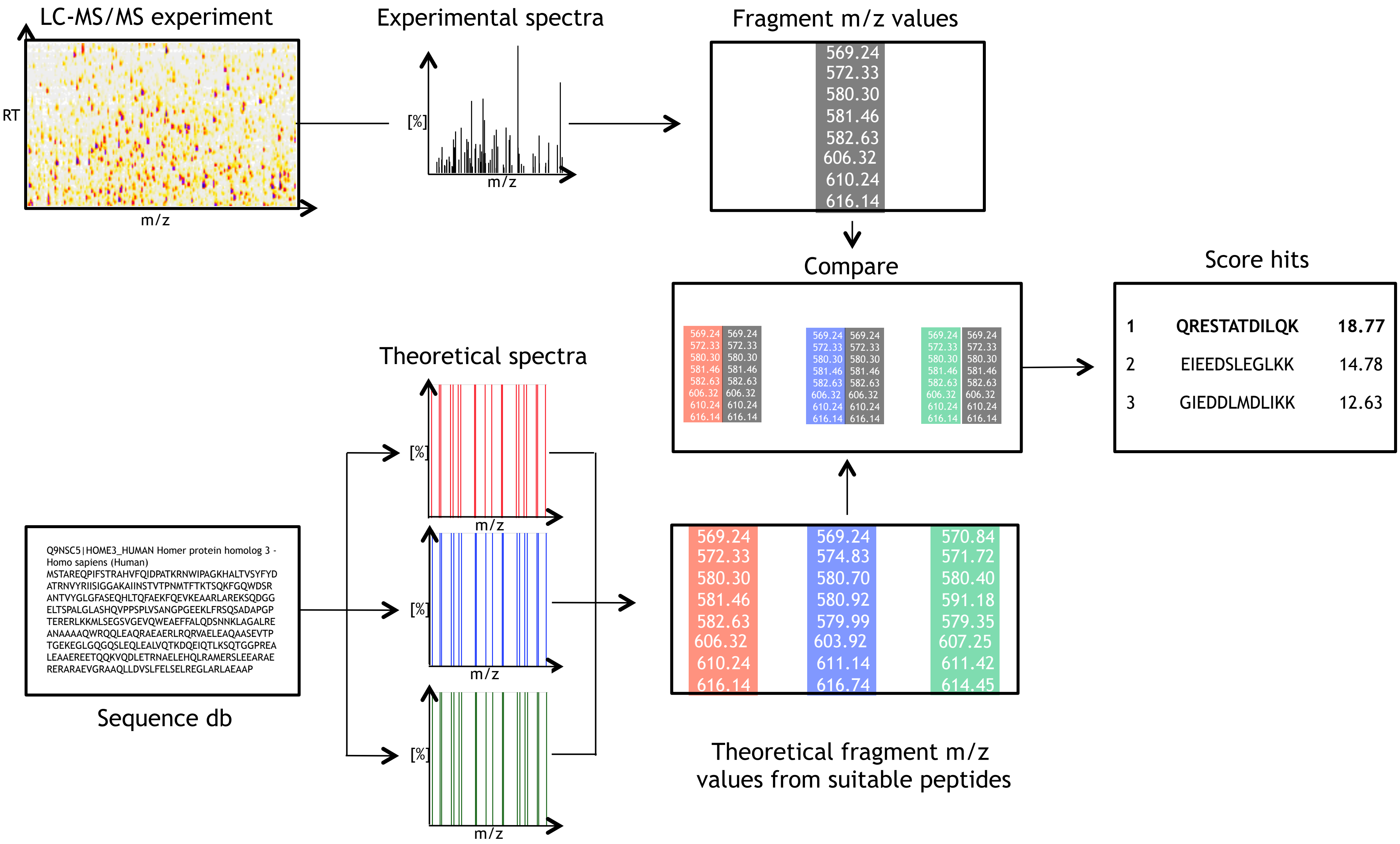
Oliver Kohlbacher, Sven Nahnsen, Knut Reinert

8. De Novo Sequencing

LEARNING UNIT 8A
CONCEPTS OF DE NOVO ID

- Difference to database search
- Problem definition
- Manual interpretation of spectra

Database Search



LC-MS/MS experiment

Experimental spectra

Fragment m/z values

569.24
572.33
580.30
581.46
582.63
606.32
610.24
616.14

Compare

569.24	569.24	569.24	569.24	569.24	569.24
572.33	572.33	572.33	572.33	572.33	572.33
580.30	580.30	580.30	580.30	580.30	580.30
581.46	581.46	581.46	581.46	581.46	581.46
582.63	582.63	582.63	582.63	582.63	582.63
606.32	606.32	606.32	606.32	606.32	606.32
610.24	610.24	610.24	610.24	610.24	610.24
616.14	616.14	616.14	616.14	616.14	616.14

Score hits

1	**QRESTATDILQK**	**18.77**
2	EIEEDSLEGLKK	14.78
3	GIEDDLMDLIKK	12.63

Theoretical spectra

Q9NSC5|HOME3_HUMAN Homer protein homolog 3 -
Homo sapiens (Human)
MSTAREQPIFSTRAHVFQIDPATKRNWIPAGKHALTVSYFYD
ATRNVYRIISIGGAKAIINSTVTPNMTFTKTSQKFGQWDSR
ANTVYGLGFASEQHLTQFAEKFQEVKEAARLAREKSQDGG
ELTSPALGLASHQVPPSPLVSANGPGEEKLFRSQSADAPGP
TERERLKKMLSEGSVGEVQWEAEFFALQDSNNKLAGALRE
ANAAAAQWRQQLEAQRAEAERLRQRVAELEAQAASEVTP
TGEKEGLGQGQSLEQLEALVQTKDQEIQTLKSQTGGPREA
LEAAEREETQQKVQDLETRNAELEHQLRAMERSLEEARAE
RERARAEVGRAAQLLDVSLFELSELREGLARLAEAAP

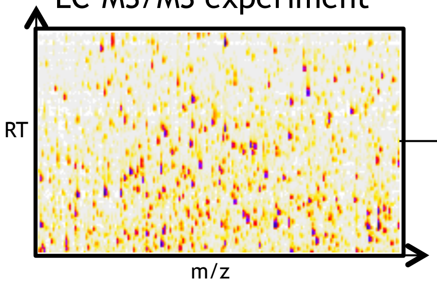
Sequence db

569.24	569.24	570.84
572.33	574.83	571.72
580.30	580.70	580.40
581.46	580.92	591.18
582.63	579.99	579.35
606.32	603.92	607.25
610.24	611.14	611.42
616.14	616.74	614.45

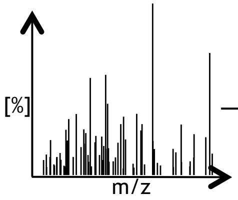
Theoretical fragment m/z
values from suitable peptides

De Novo Sequencing?

LC-MS/MS experiment



Experimental spectra



Fragment m/z values

569.24
572.33
580.30
581.46
582.63
606.32
610.24
616.14

Compare

569.24	569.24	569.24	569.24	569.24	569.24
572.33	572.33	572.33	572.33	572.33	572.33
580.30	580.30	580.30	580.30	580.30	580.30
581.46	581.46	581.46	581.46	581.46	581.46
582.63	582.63	582.63	582.63	582.63	582.63
606.32	606.32	606.32	606.32	606.32	606.32
610.24	610.24	610.24	610.24	610.24	610.24
616.14	616.14	616.14	616.14	616.14	616.14

Score hits

1	**QRESTATDILQK**	**18.77**
2	EIEEDSLEGLKK	14.78
3	GIEDDLMDLIKK	12.63

All theoretical spectra

All peptides for a given precursor mass?

De Novo Sequencing Problem

- **Given**
  - A tandem MS spectrum *s*
  - A (precursor) peptide mass *M*
  - A scoring function *f(s, p)* scoring a peptide sequence $p = a_1a_2\ldots a_n$ against the spectrum *s*

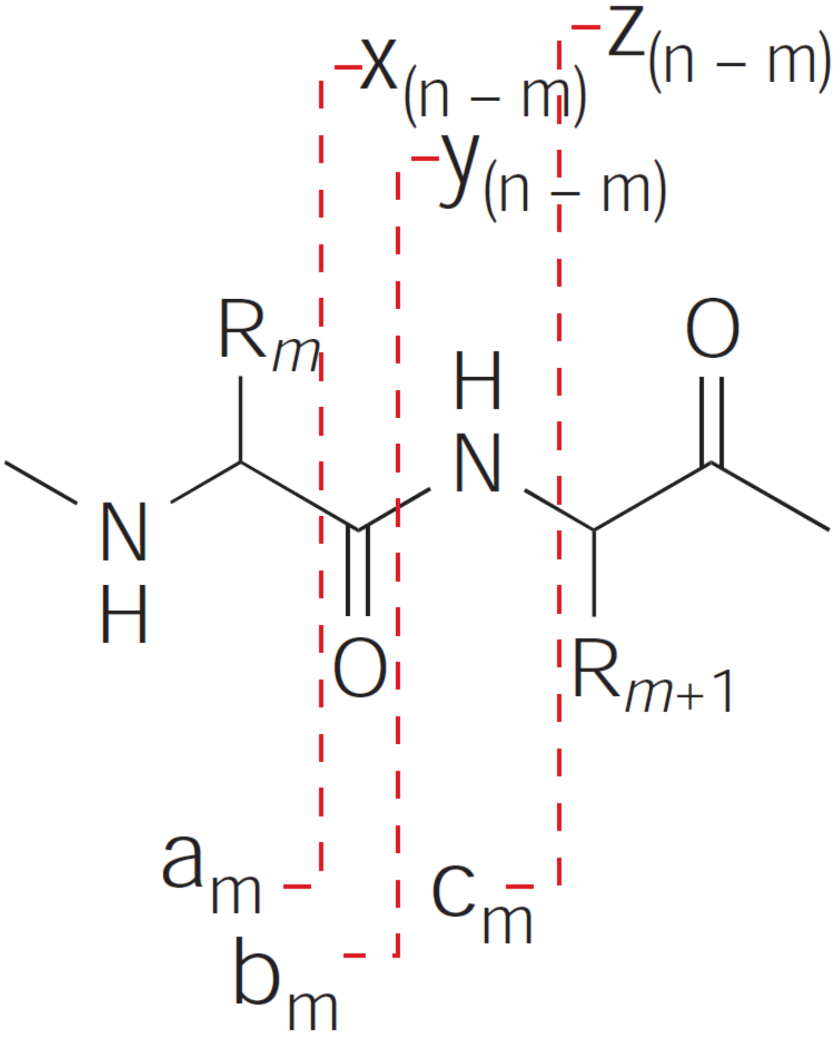
- **Find**
  - The amino acid sequence p^* with mass *M* maximizing the score $f(s, p^*)$

De Novo Statistics

- How many peptides are there for a given mass?
  - Without the restriction of the database search, all potential sequences need to be searched
  - Peptides with the same composition (i.e., same number of of residues from each amino acid) will have the same mass
  - The number of potential peptides of the same composition rises with the peptide length *n* (and thus the mass) as *n!*

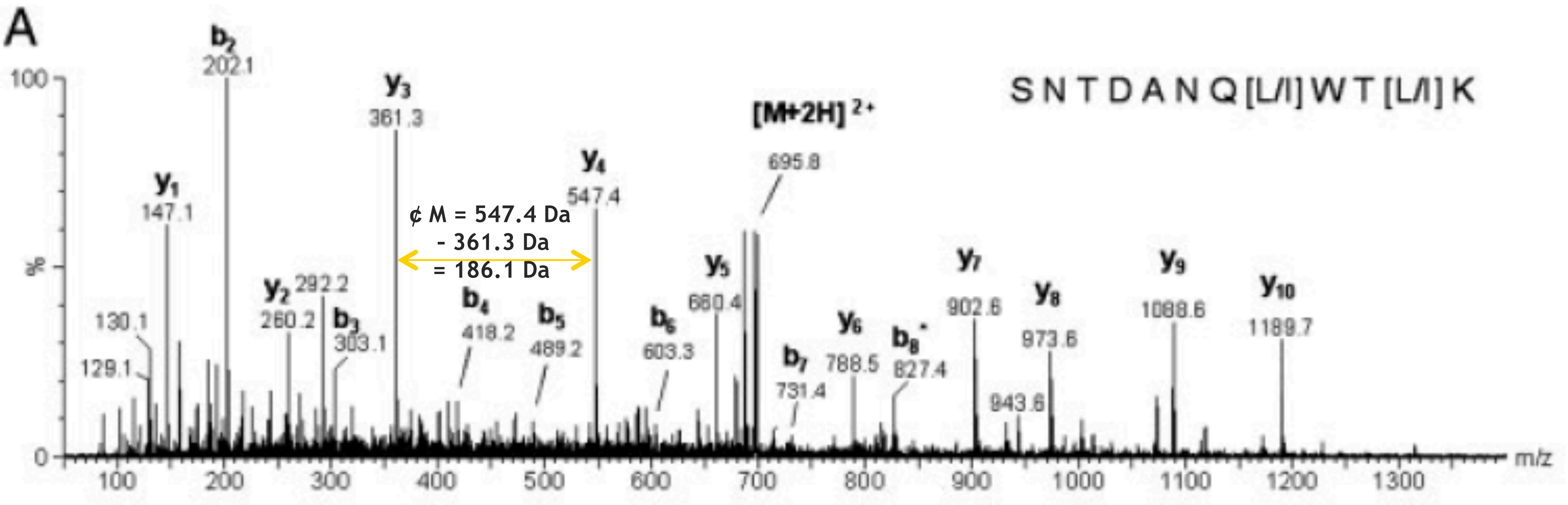
Fragmentation

- As discussed earlier, fragmentation gives rise to ion series (b, y most of all)
- De novo sequencing requires ***complete* ion series** (ladders)
- Incomplete ladders, missing peaks imply that the true sequence can usually not be identified
- Apart from the abc/xyz series, **neutral losses** and **internal fragments** play an important role as well



Steen and Mann. Nature Reviews, Molecular Cell Biology, Vol. 5 2004

Manual Sequencing

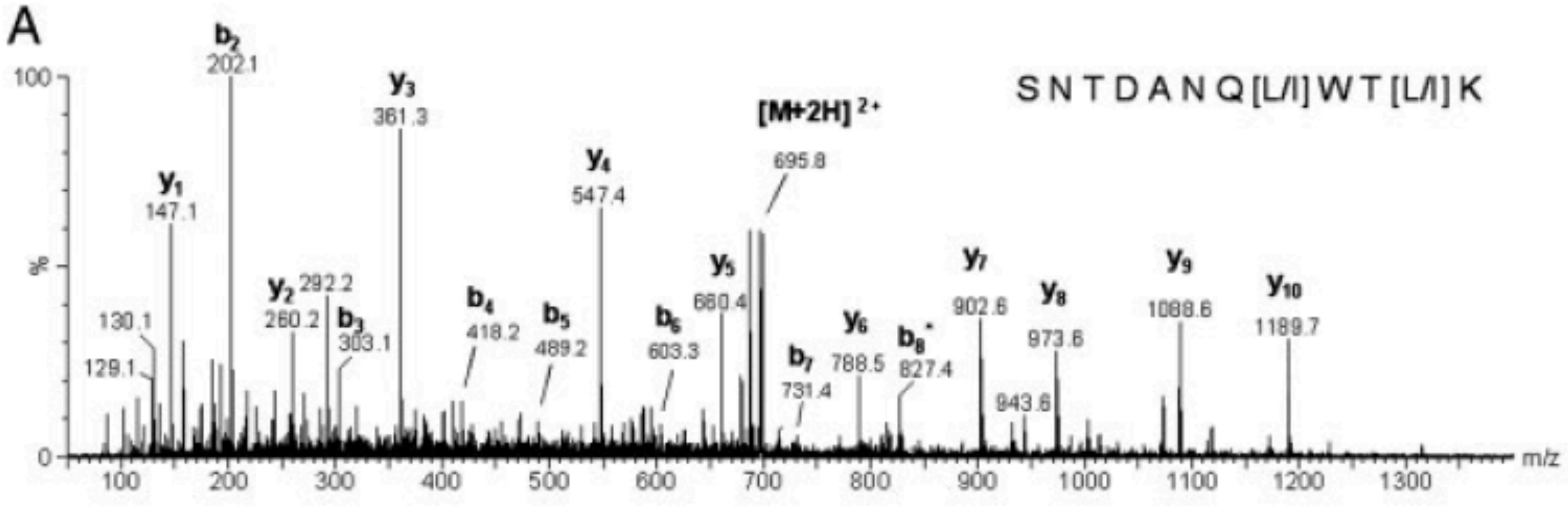


¢ M = 547.4 Da
- 361.3 Da
= 186.1 Da

- Q-TOF CID spectrum of the tryptic peptide SNTDANQ[L|I]WT[L|I]K
- The graph shows the complete spectrum with annotated b and y ion series
- Differences between the masses of adjacent ions of the same series permit the identification of the sequence at this position
- y ion series contains suffix ions (from the N terminus)
- b ion series contains prefix ions (from the C terminus)

Seidler et al., Proteomics (2010), 10:634-649

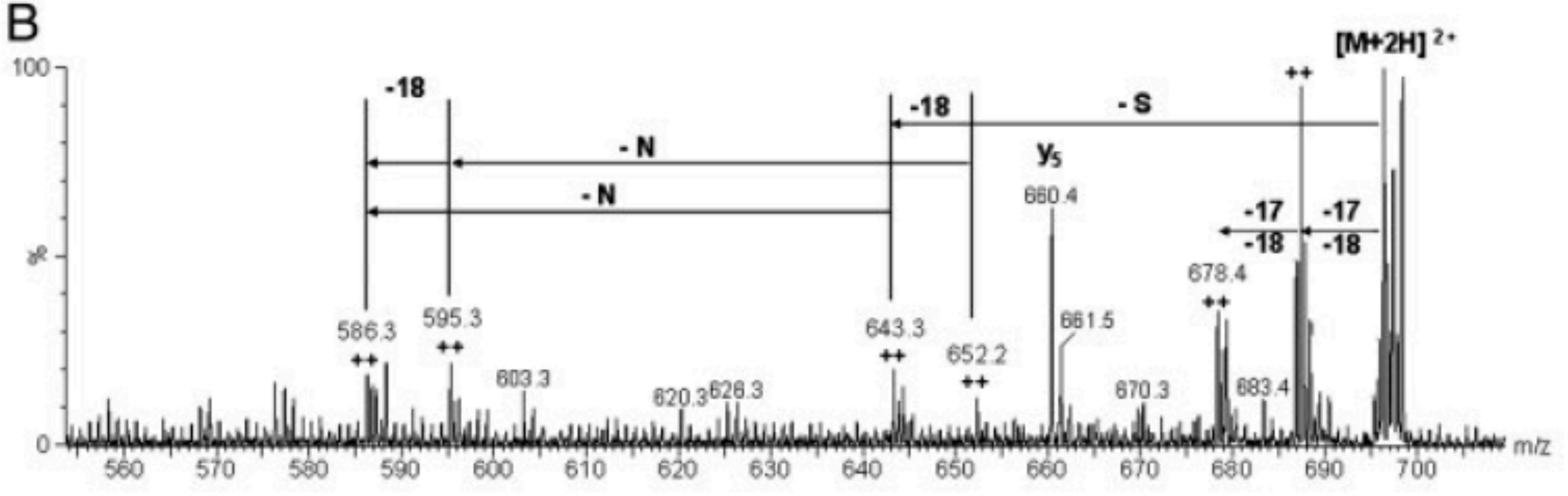
Manual Sequencing



- Corresponding b/y ion pairs should add up to the precursor mass
  - m(b_2) + m(y_{10}) = 202.1 Da + 1189.7 Da = 1391.8 Da
  - m(b_3) + m(y_9) = 303.1 Da + 1088.6 Da = 1391.7 Da
  - m(b_4) + m(y_8) = 418.2 Da + 973.6 Da = 1391.8 Da
  - …
- Absent: y_{11} and b_1 – C terminal sequence can be SN or NS from its mass, no information in b/y ion series on the order
- Theoretical mass of the sequence: 1390.6961 Da

Seidler et al., Proteomics (2010), 10:634-649

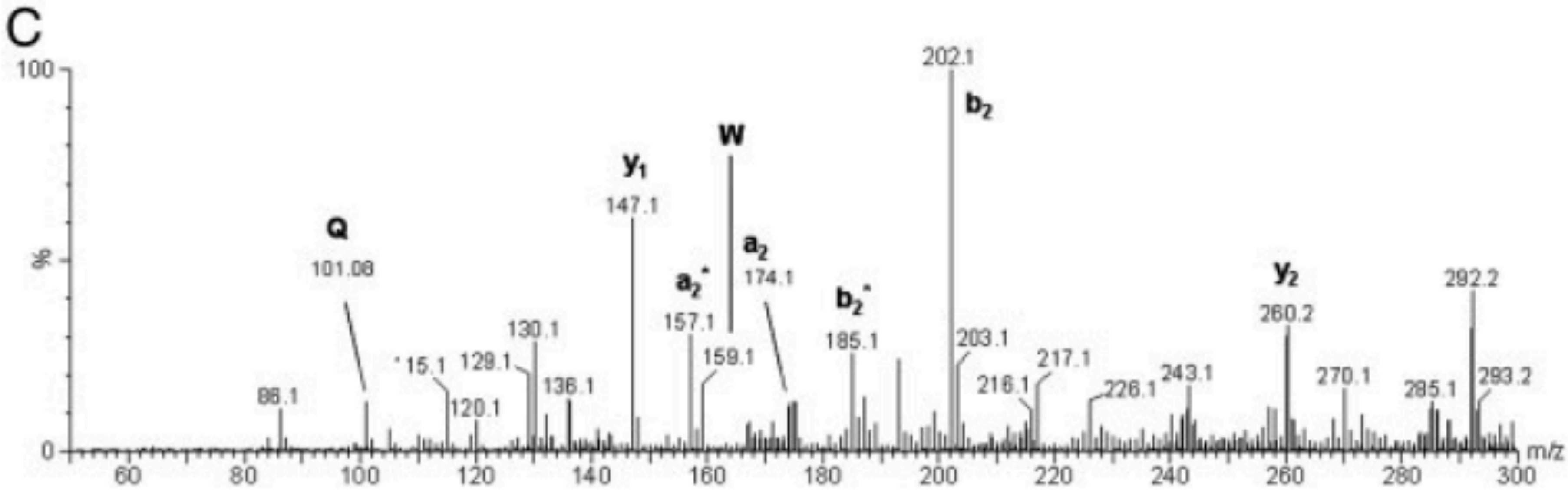
Manual Sequencing



- Central region of the spectrum showing C-terminal neutral losses
  - In this case the presence of a strong signal for the neutral loss of S and then N is present
  - Additional neutral losses for water (18.01 Da) are present, supporting the hypothesis
  - C-terminal sequence has thus to be SN…

Seidler et al., Proteomics (2010), 10:634-649

Manual Sequencing



- Low-mass region
  - contains shortest suffix/prefix ions from the C/N termini
  - Contains immonium marker ions for the amino acids present
  - In this case the sequence has to contain W and Q from the very prominent marker ions at 101.1 and 159.1 Da

Seidler et al., Proteomics (2010), 10:634-649

LEARNING UNIT 8B
ALGORITHMIC CONCEPTS

- Spectrum graphs
- Extended spectrum graph
- Antisymmetric paths
- Precursor mass correction

What’s the Problem?

- **Problems**
  - Manual annotation is a matter of hours or days per spectrum – **not high throughput**!
  - Automatic annotation is difficult
    - **Assignment** of ion series is not known in advance
    - ‘**Noise peaks**’ are present and intensities of ion series can vary widely
    - Some **ion peaks will be missing**
- In order to solve the problem, we need the following:
  - An **abstraction** permitting an efficient search
  - A **search algorithm and scoring function** that tolerate missing peaks and additional noise peaks

Formal Models

- A very popular abstraction of the de novo sequencing problem is the so-called **spectrum graph**
  - **Nodes** in this graph represent possible **interpretations of a peak** (in the simplest case: one for every b, one for every y ion)
  - Two nodes are connected by a (directed) **edge**, if they are of the same series, but **differ by an amino acid mass**

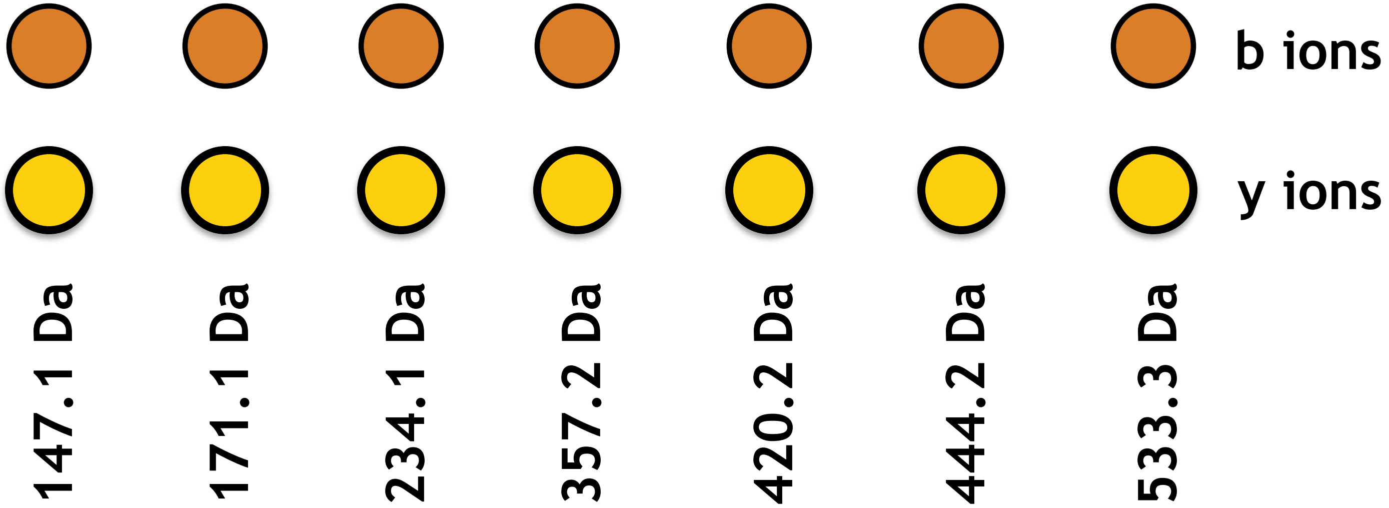
  Note: There are several, slightly different definitions of spectrum graphs in the literature

- **Construction**
  - Clean up the spectrum (remove noise peaks) and create two nodes, z_0 and z_m, on a line to represent the zero mass and the total residue mass
  - For each peak, create a pair of nodes, z_j and z_{m-j} , placed at the mass for the b and y ions.
  - For all pairs of nodes (except the b/y pairs), check whether the mass difference corresponds to an amino acid mass and add an edge if it matches

Formal Models

- A very popular abstraction of the de novo sequencing problem is the so-called **spectrum graph**
  - **Nodes** in this graph represent possible **interpretations of a peak** (in the simplest case: one for every b, one for every y ion)
  - Two nodes are connected by a (directed) **edge**, if they are from **differ by an amino acid mass**
- Construction
  - Clean up the spectrum (remove noise peaks)
  - For each peak, add a node (b/y) to the graph, color by ion series
  - For all pairs of nodes of the same series, check whether the mass difference corresponds to an amino acid mass and add an edge if it matches
  - Label each edge with the matching amino acid

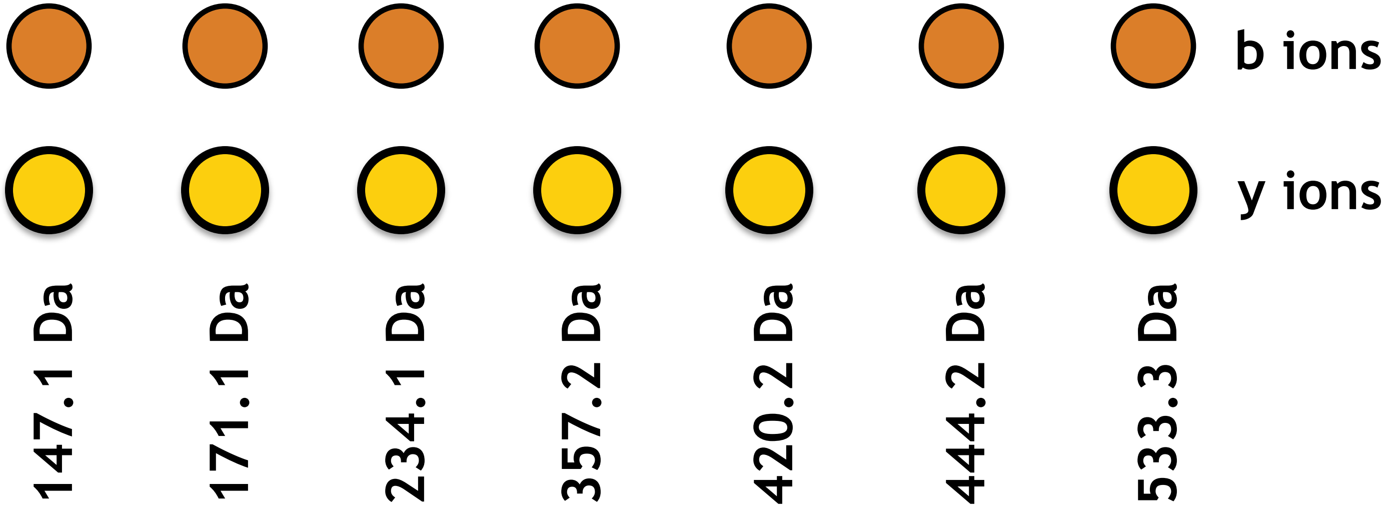
Spectrum Graph – Example



m/z [Da]
147.1128
171.1128
234.1448
357.1921
420.2241
444.2241
533.3082

- For a simple spectrum (no noise peaks, no missing peaks), we will illustrate the construction of the spectrum graph and its interpretation
- For each peak, add two nodes: brown for b ions, yellow for y ions

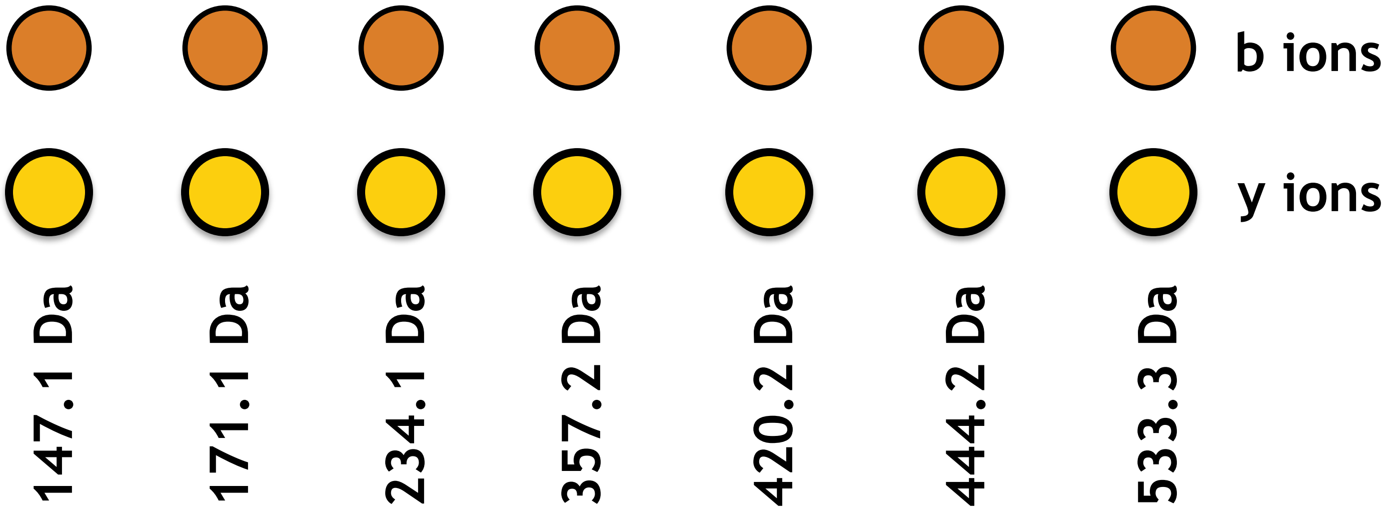
Spectrum Graph – Example



m/z [Da]
147.1128
171.1128
234.1448
357.1921
420.2241
444.2241
533.3082

- For all pairs (u, v) of nodes of the same color:
  - If |m(u) – m(v)| = m(aa) for any amino acid aa, add the edge (u,v) and label it with aa

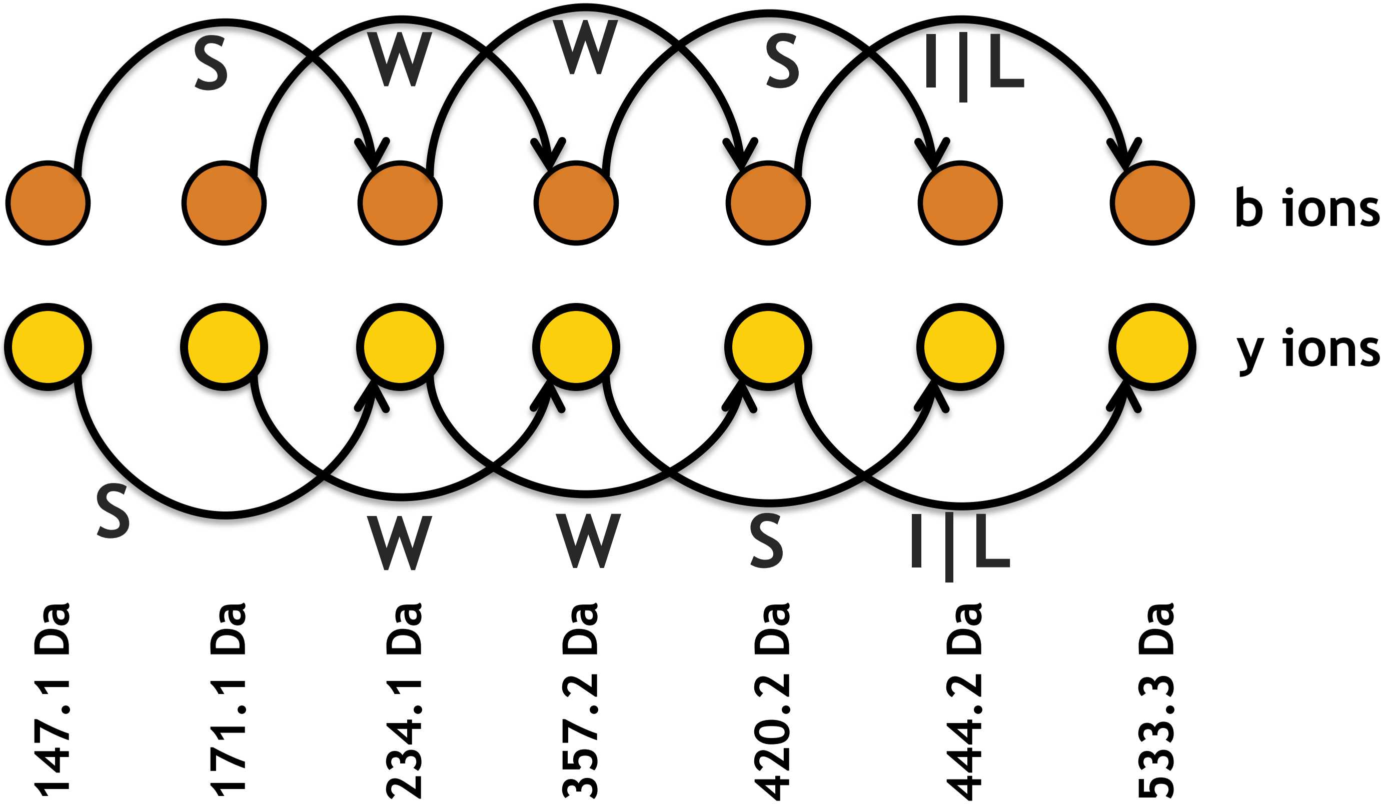
Spectrum Graph – Example



m/z [Da]
147.1128
171.1128
234.1448
357.1921
420.2241
444.2241
533.3082

- 147.1 Da – 171.1 Da = -24.0 Da – nothing
- 147.1 Da – 234.1 Da = -87.0 Da – serine!
- 147.1 Da – 357.2 Da = -210.1 Da – nothing
- …

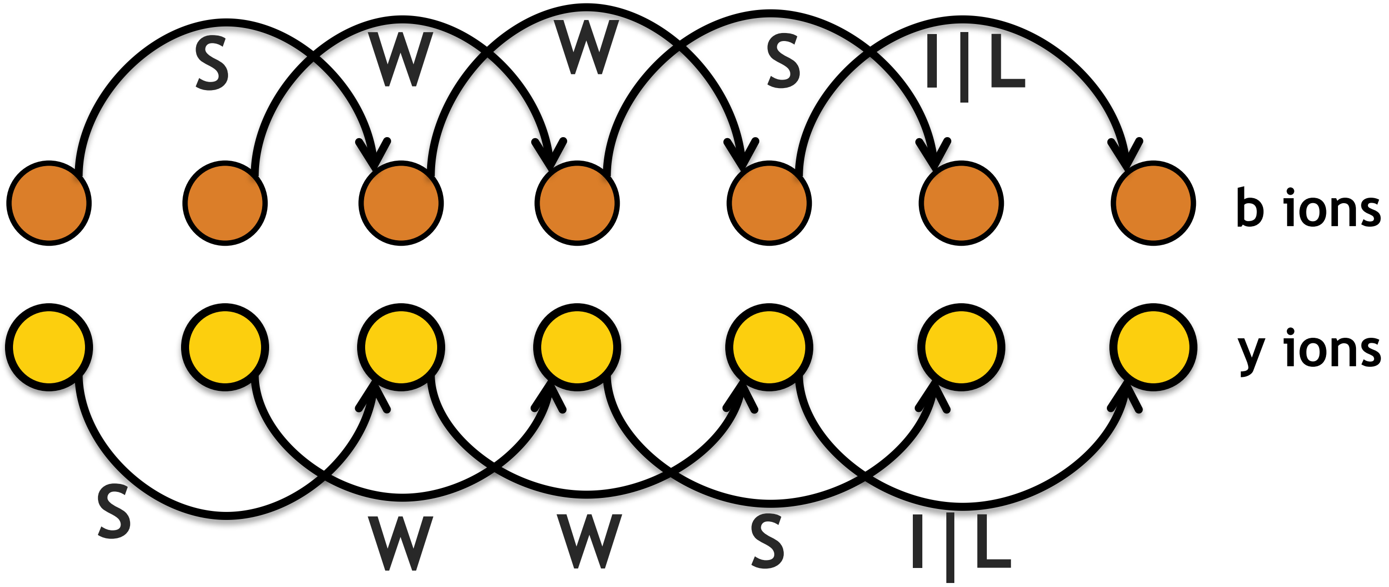
Spectrum Graph – Example



m/z [Da]
147.1128
171.1128
234.1448
357.1921
420.2241
444.2241
533.3082

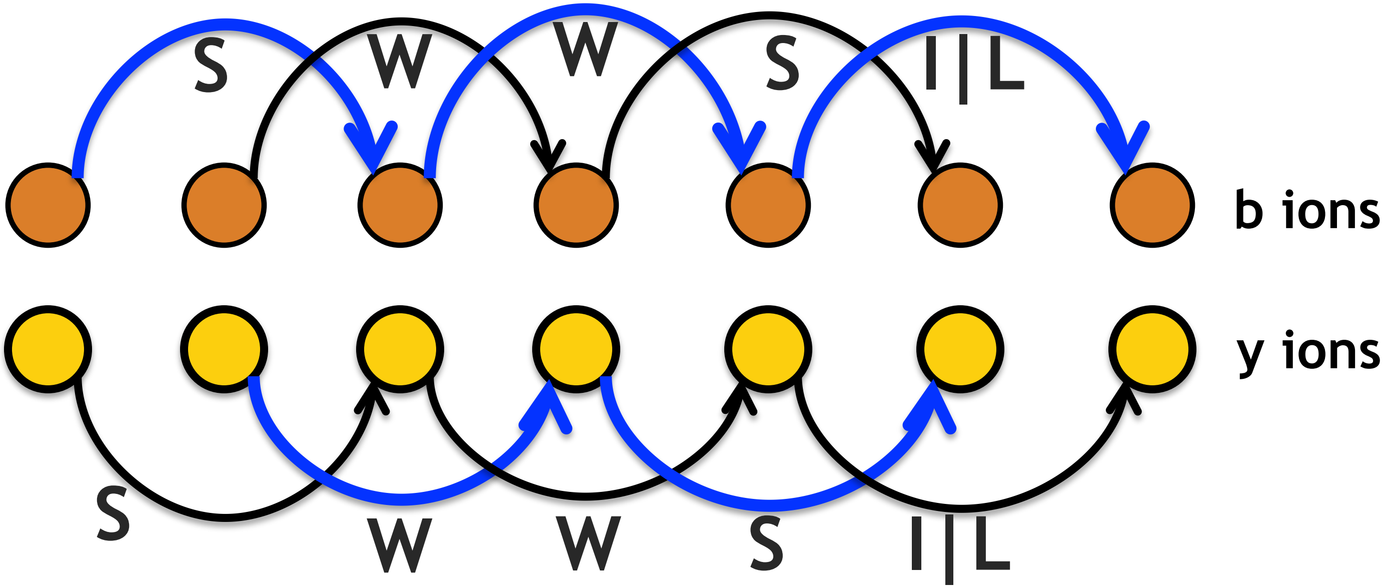
- 234.1 Da – 147.1 Da = 87.0 Da = S
- 357.2 Da – 171.1 Da = 186.1 Da = W
- 420.2 Da – 234.1 Da = 186.1 Da = W
- 444.2 Da – 357.2 Da = 87.0 Da = S
- 533.3 Da - 420.2 Da = 113.1 Da = [I|L]

Spectrum Graph – Example



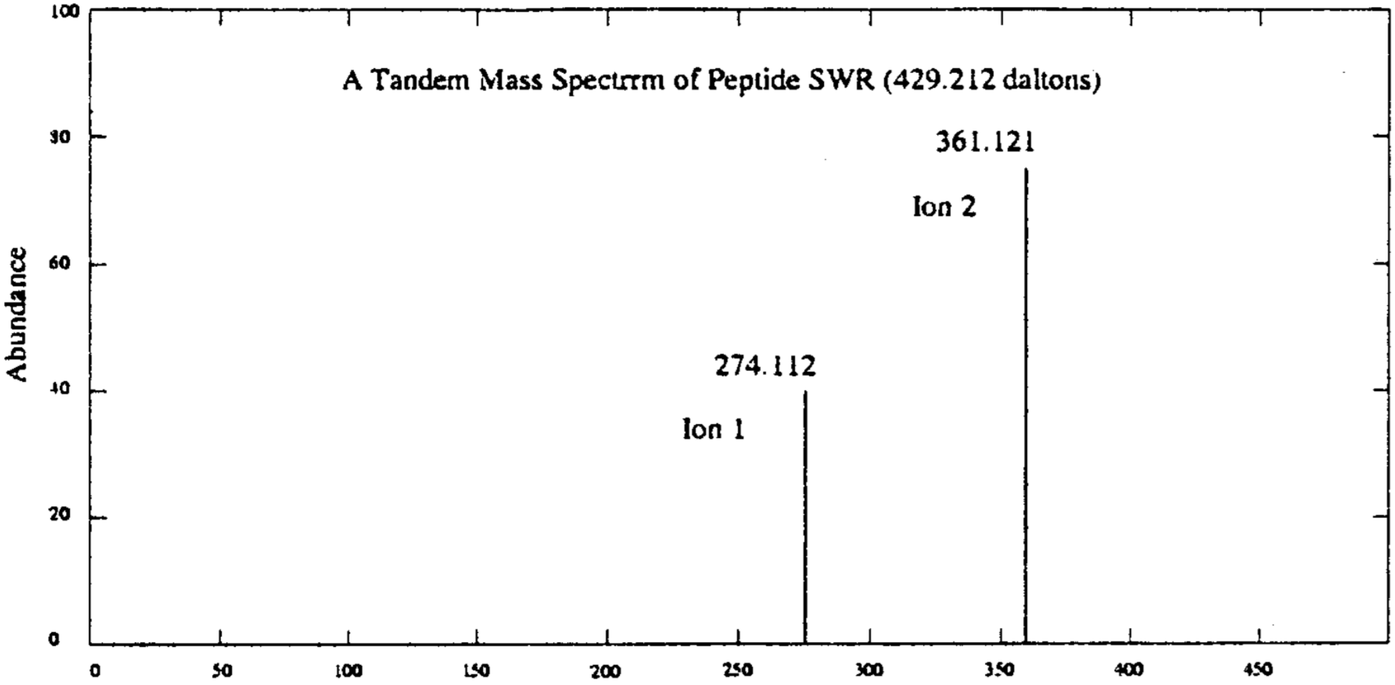
- Every mass can only come from ONE type of ion (each peak corresponds to either a b ion or a y ion, not both!)
- Missing: b_1/y_1: no difference to mass zero or parent mass present (it is however straight-forward to add these as additional nodes)
- Now what is the sequence of our peptide?

Spectrum Graph – Example

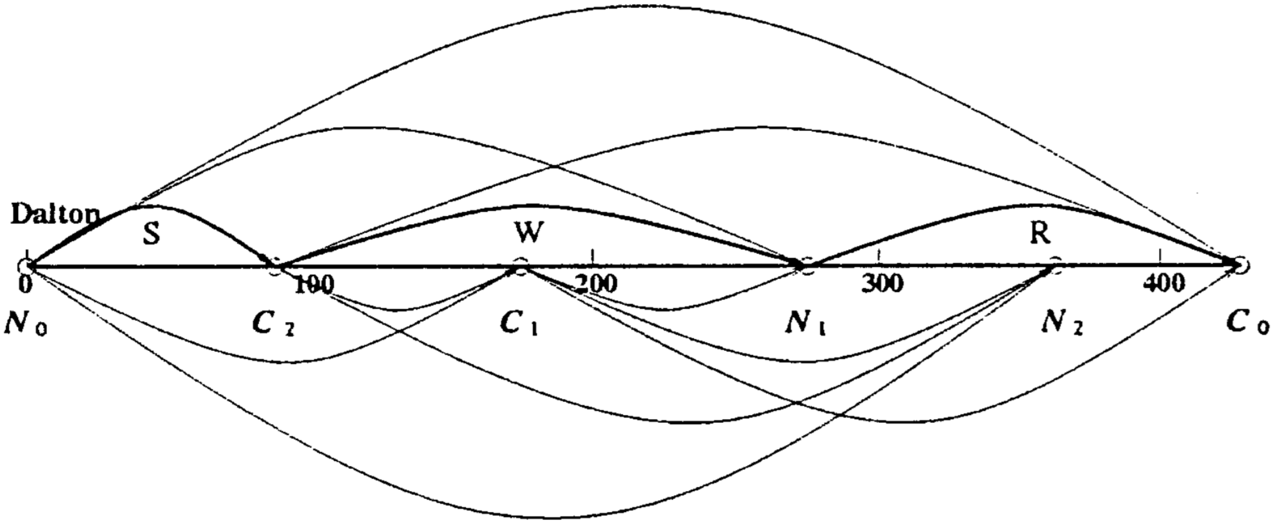


- The sequence can be read from a complete series of either b or y ions
- An ion series is a path through nodes of the same color
- Each peak can only be contained in either series (brown or yellow)
- We thus need to find a path through brown nodes or yellow nodes from very small to very large masses (or the other way round)
- This path would correspond to an ion series
- In this case: our peptide seems to contain the sequence SW[I|L] or [I|L]WS (note that we do not know the order since we do not know whether red or blue are b or y ions! Also, in our case the b_1 ion is missing)

Formal Models



NC-spectrum graph G=(V,E)



Extended Spectrum Graph

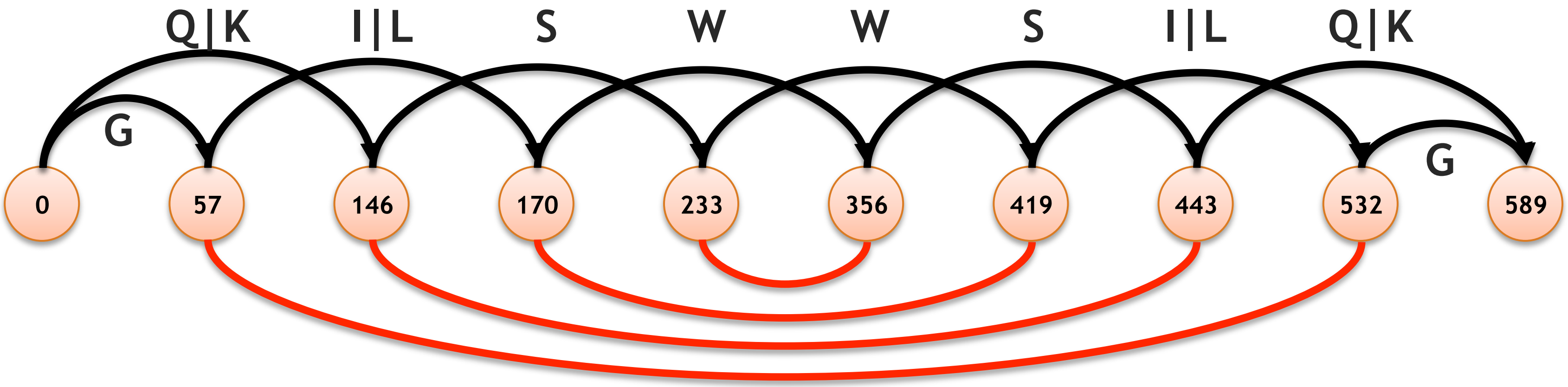
- A convenient simplification to the spectrum graph was introduced by Liu et al. in 2001: the **extended spectrum graph** (ESG)
- The ESG *G(V, E)* contains
  - Nodes for each peak (plus nodes v_0 for mass 0 and the v_M representing the total mass *M* of the peptide)
  - Directed edges (*u, v*) for each pair of nodes *u, v* where *m*(*v*)-*m*(*u*) matches a single amino acid mass
  - **Undirected edges for each pair of nodes *u, v* that are complementary**, i.e., where *m*(*u*) + *m*(*v*) = *M*
- Note that all raw m/z of the peaks have to be corrected by their charge and the proton mass subtracted from the resulting mass!

Liu et al., PSB 2005

Extended Spectrum Graph – Example

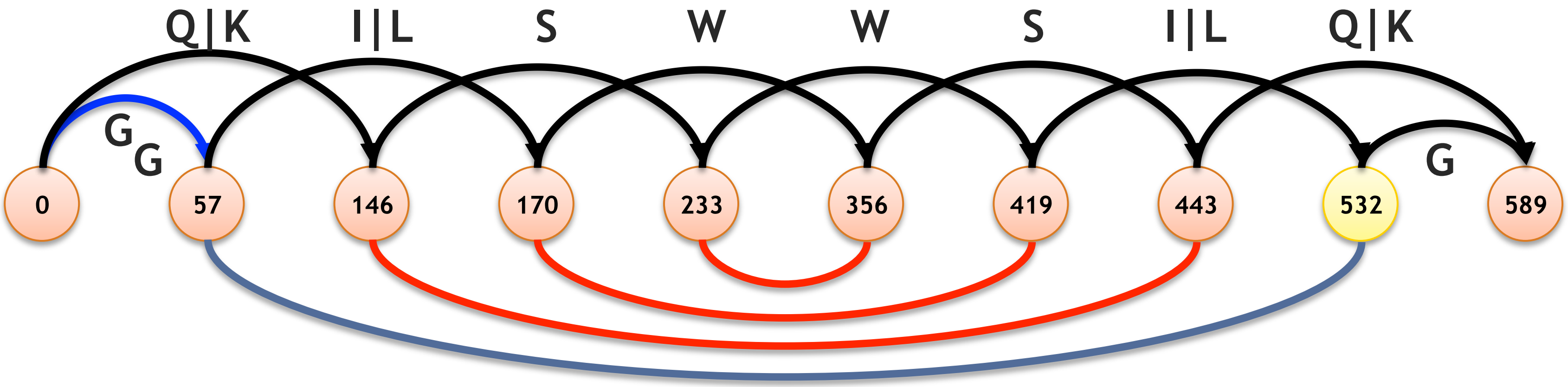
- Example from the extended spectrum graph:
  - Correct masses, add node for intact peptide (589 Da) and source node
  - For simplicity's sake, we will use only nominal masses, but add the mass of the missing b_1 ion (57 Da)
  - (note that the edges to the sink and source have to be corrected for the mass of water for C-terminal b/y ions)

m/z [Da]
147.1128
171.1128
234.1448
357.1921
420.2241
444.2241
533.3082



Extended Spectrum Graph – Example

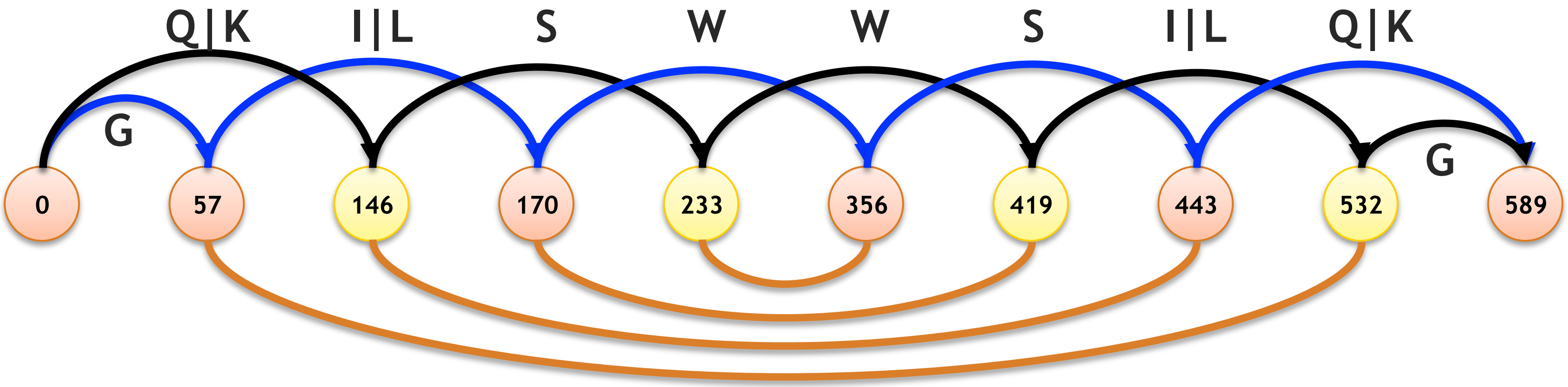
- An **antisymmetric path** is a path from source v_0 to sink v_M if it includes at most one of each of the pairs of complementary vertices
- Example
  - Path going from 0 to 57 (G) can no longer use 532 as 57 and 532 are complementary



Extended Spectrum Graph – Example

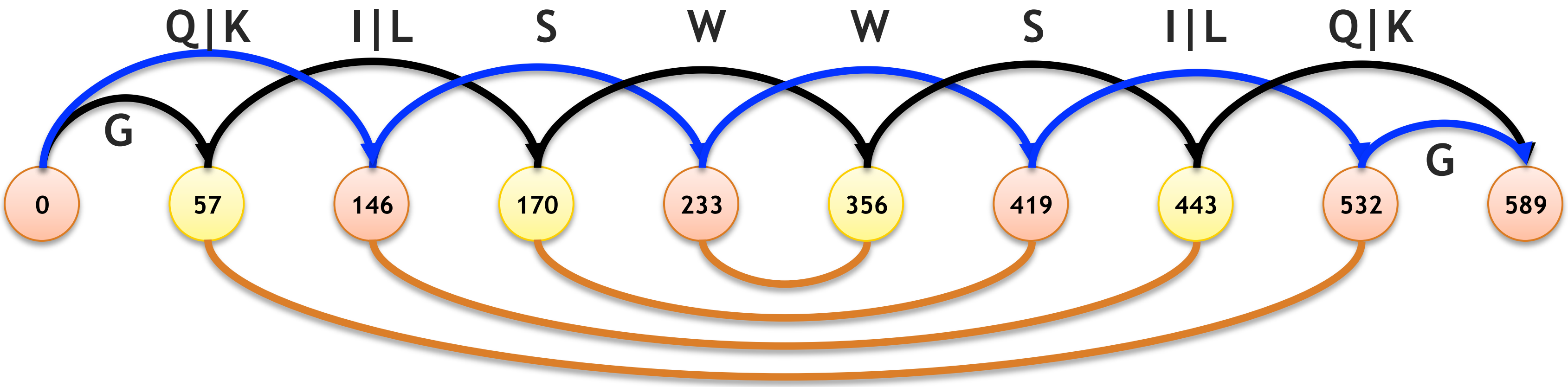
- An **antisymmetric path** is a path from source v_0 to sink v_M if it includes at most one of each of the pairs of complementary vertices
- Example:
  - Path going from 0 to 57 (G) can no longer use 532 as 57 and 532 are complementary
  - The resulting longest path contains a possible sequence of the peptide:

    G[I|L]WS[Q|K]

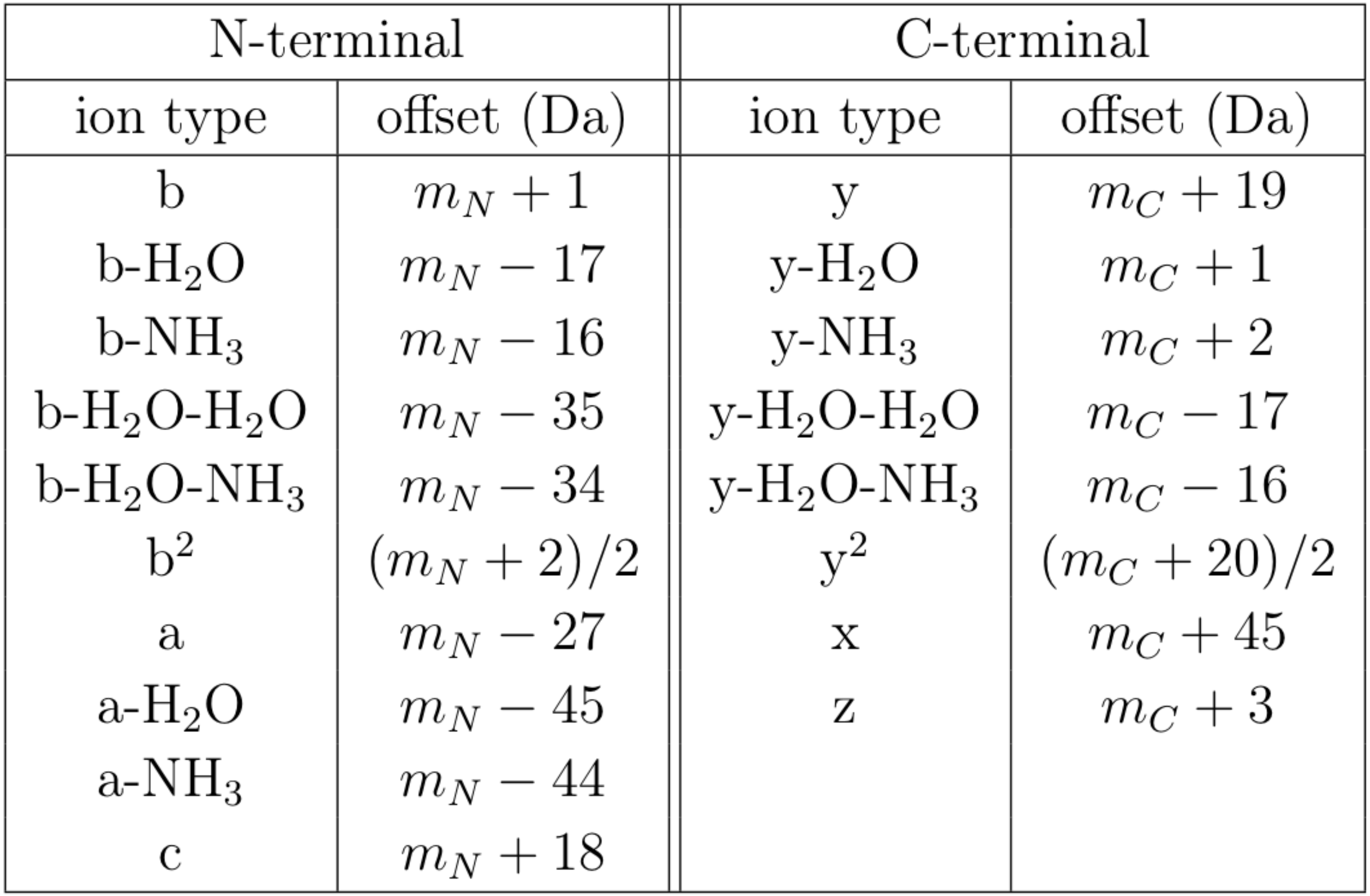


Extended Spectrum Graph – Example

- An **antisymmetric path** is a path from source v_0 to sink v_M if it includes at most one of each of the pairs of complementary vertices
- Example:
  - Another option would yield the inverse sequence [Q|K]SW[I|L]G
  - If we knew that we are dealing with a tryptic peptide, it would be obvious that the first solution is the correct one
  - In reality, the presence of noise peaks and missing peaks render the problem vastly more difficult

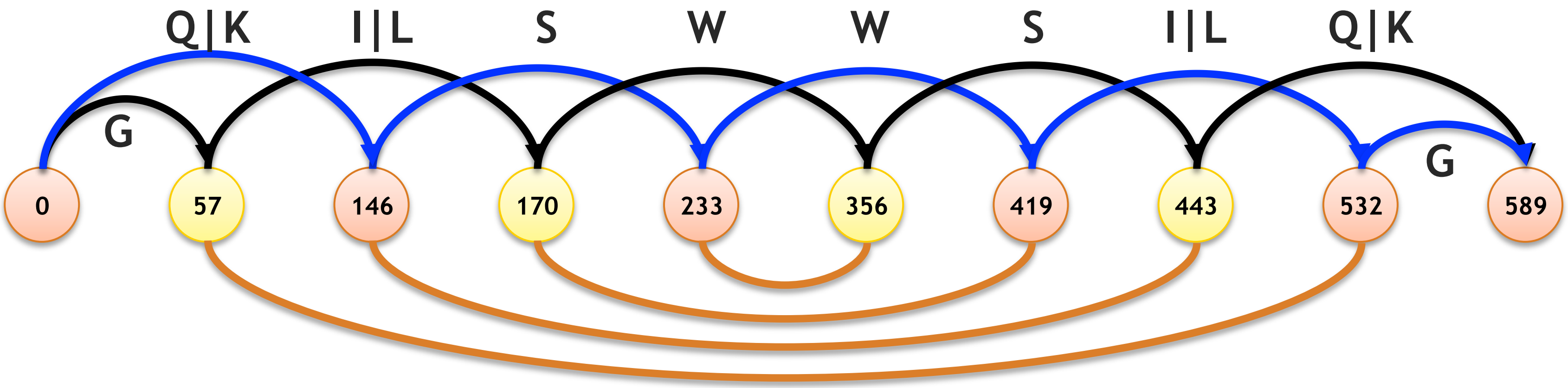


Extended Spectrum Graph – Ion types



N-terminal		C-terminal	
ion type	offset (Da)	ion type	offset (Da)
b	$m_N + 1$	y	$m_C + 19$
b-H_2O	$m_N - 17$	y-H_2O	$m_C + 1$
b-NH_3	$m_N - 16$	y-NH_3	$m_C + 2$
b-H_2O-H_2O	$m_N - 35$	y-H_2O-H_2O	$m_C - 17$
b-H_2O-NH_3	$m_N - 34$	y-H_2O-NH_3	$m_C - 16$
b^2	$(m_N + 2)/2$	y^2	$(m_C + 20)/2$
a	$m_N - 27$	x	$m_C + 45$
a-H_2O	$m_N - 45$	z	$m_C + 3$
a-NH_3	$m_N - 44$		
c	$m_N + 18$		

Scoring



- Generally, a large number of possible antisymmetric paths can exist (including noise peaks!)
- The search for a longest path is then generally replaced by the search for a heaviest path
- Node weights are introduced and usually contain peak intensity and mass deviations, but also statistical models of the likelihood of observing a certain peak type learned from experimental data

Precursor Mass Correction

- The precursor mass of a tandem MS spectrum is usually defined with low accuracy only due to the large mass selection window
- It can also be determined incorrectly by the instrument software (e.g., selecting an isotope peak of the MS spectrum instead of the monoisotopic peak, wrong charge state assignment)
- A more accurate knowledge of the precursor mass (i.e., the total peptide mass) can significantly reduce the search space (both for database search and de novo sequencing)
- Before applying de novo methods, it is thus common to obtain a more accurate estimate from the tandem spectrum
- This is known as **precursor mass correction**

Precursor Mass Correction

Definition:

Let $S = \{m_1, m_2, \ldots m_n\}$ be a mass spectrum of a peptide with mass M with peaks at m/z $m_1, .. m_n$ and charge state z.

The inverse (or reverse) spectrum S' is then defined as follows:

$$S' = \{m'_i \mid m'_i = M + z\, m_p - m_i\}$$

where m_p is the proton mass.

Since the masses of complementary ions add up to $M + z\, m_p$, the masses of b or y ions are translated to their corresponding complementary ion masses in the inverse spectrum.

Precursor Mass Correction

- **Idea**
    - The tandem spectrum contains complementary ions (b/y, a/x, c/z)
    - Complementary ion masses will add up to the correct total peptide/precursor mass
    - For the correct precursor mass M the inverse spectrum will be computed correctly and share a maximum number of peaks with the original spectrum
- This problem can be formulated as a combinatorial optimization problem
- There are various ways to solve the problem, we will look at a simple algorithm that solves the problem in cubic time in the number of peaks

Precursor Mass Correction

```
Algorithm
$max\_spc \leftarrow 0$
$best\_M_p \leftarrow 0$
FOR $1 \leq i, j \leq$ n:
    Compute potential precursor mass $M_p = m_i + m_j - z\ m_p$
    Compute $S'$ given $M_p$
    Compute shared peak count between $S$ and $S'$:
    $spc \leftarrow \{|(m_i, m'_j), 1 \leq i, j \leq n \mid |m_i - m_j| < \delta\}$
    IF $spc > max\_spc$:
        $max\_spc \leftarrow spc$
        $best\_M_p \leftarrow M_p$
RETURN $best\_M_p$
```

Eindhammer et al., Computational methods for mass spectrometry proteomics, p.137

LEARNING UNIT 8C
DE NOVO ID WITH ANTELOPE

- Key ideas
- Heaviest path search
- ILP formulation
- Performance of de novo ID

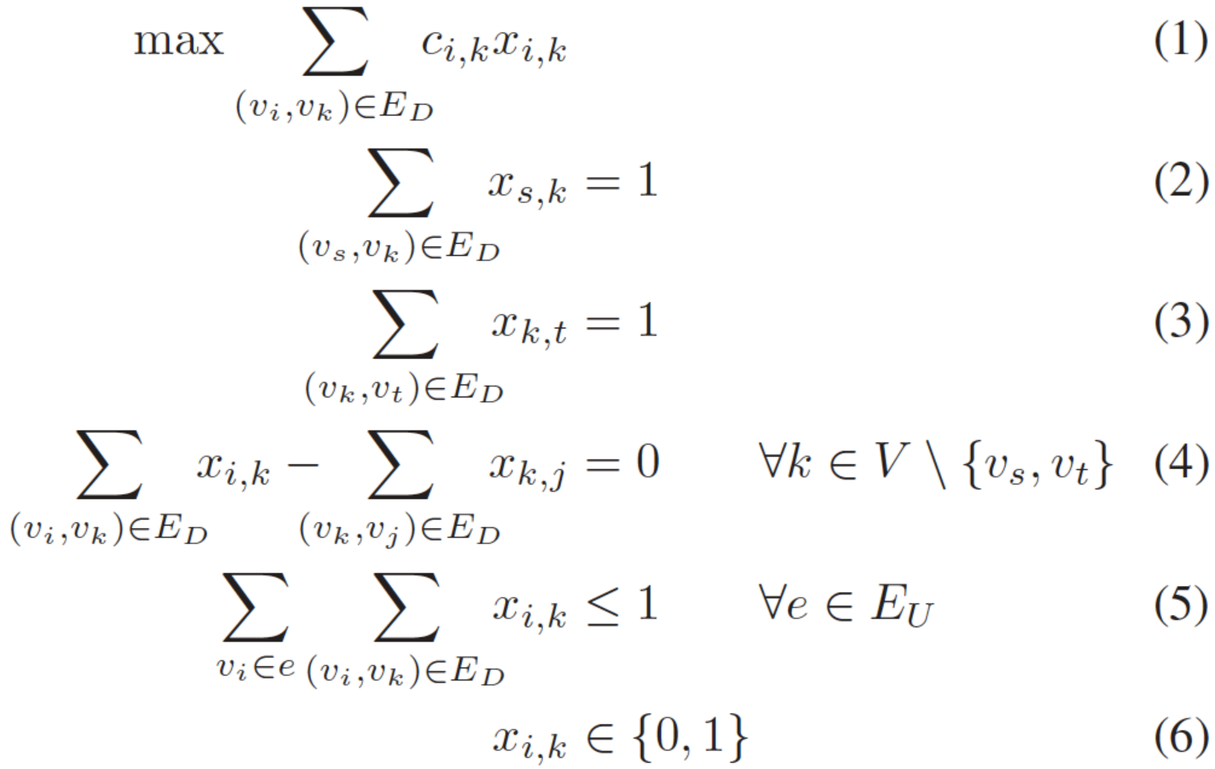
ANTILOPE

- ANTILOPE is a de novo sequencing approach based on the extended spectrum graph
- The problem of finding the longest asymmetric path is slightly modified
- It can be formulated as an integer linear program (ILP)
- This ILP formulation can then be solved using Lagrangian relaxation quite efficiently
- We will only discuss the ILP formulation for the sake of time

Andreotti, Klau, Reinert, IEEE TCCB (2011), 99, 159

ANTILOPE

- ESG $G(V, E_D, E_U)$ with directed edges E_D and undirected edges E_U and binary decision variables $x_{i,k}$ for each directed edge in G
- Assign weights $c_{i,k}$ to each edge (the weight of a node is assigned to each outgoing edge)
- Solve the following optimization problem:



$$\max \sum_{(v_i,v_k)\in E_D} c_{i,k}x_{i,k} \tag{1}$$

$$\sum_{(v_s,v_k)\in E_D} x_{s,k} = 1 \tag{2}$$

$$\sum_{(v_k,v_t)\in E_D} x_{k,t} = 1 \tag{3}$$

$$\sum_{(v_i,v_k)\in E_D} x_{i,k} - \sum_{(v_k,v_j)\in E_D} x_{k,j} = 0 \qquad \forall k \in V \setminus \{v_s, v_t\} \tag{4}$$

$$\sum_{v_i\in e} \sum_{(v_i,v_k)\in E_D} x_{i,k} \leq 1 \qquad \forall e \in E_U \tag{5}$$

$$x_{i,k} \in \{0, 1\} \tag{6}$$

Andreotti, Klau, Reinert, IEEE TCCB (2011), 99, 159

ANTILOPE

Find the heaviest path…

$$\boxed{\max \sum_{(v_i,v_k)\in E_D} c_{i,k}x_{i,k} \qquad (1)}$$

$$\sum_{(v_s,v_k)\in E_D} x_{s,k} = 1 \qquad (2)$$

$$\sum_{(v_k,v_t)\in E_D} x_{k,t} = 1 \qquad (3)$$

$$\sum_{(v_i,v_k)\in E_D} x_{i,k} - \sum_{(v_k,v_j)\in E_D} x_{k,j} = 0 \qquad \forall k \in V \setminus \{v_s, v_t\} \qquad (4)$$

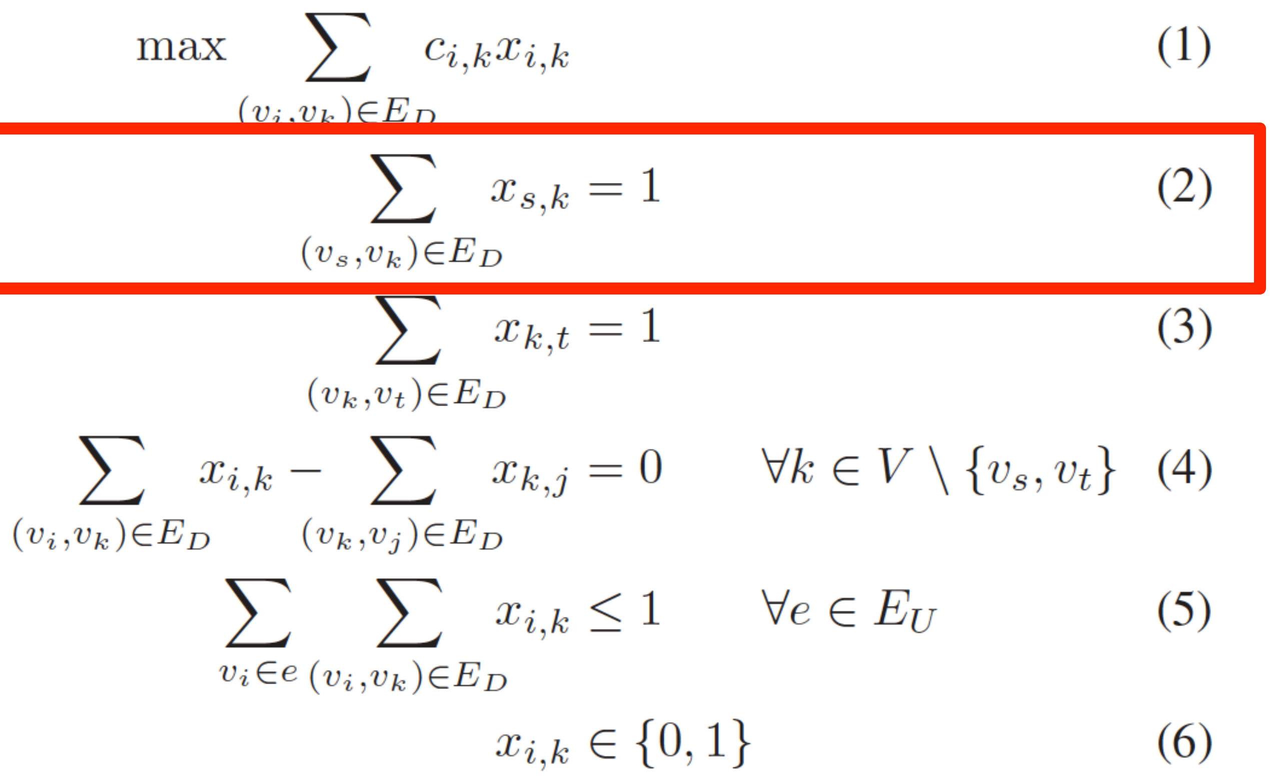
$$\sum_{v_i\in e} \sum_{(v_i,v_k)\in E_D} x_{i,k} \leq 1 \qquad \forall e \in E_U \qquad (5)$$

$$x_{i,k} \in \{0,1\} \qquad (6)$$

Andreotti, Klau, Reinert, IEEE TCCB (2011), 99, 159

ANTILOPE

…starting in the source node *s*…

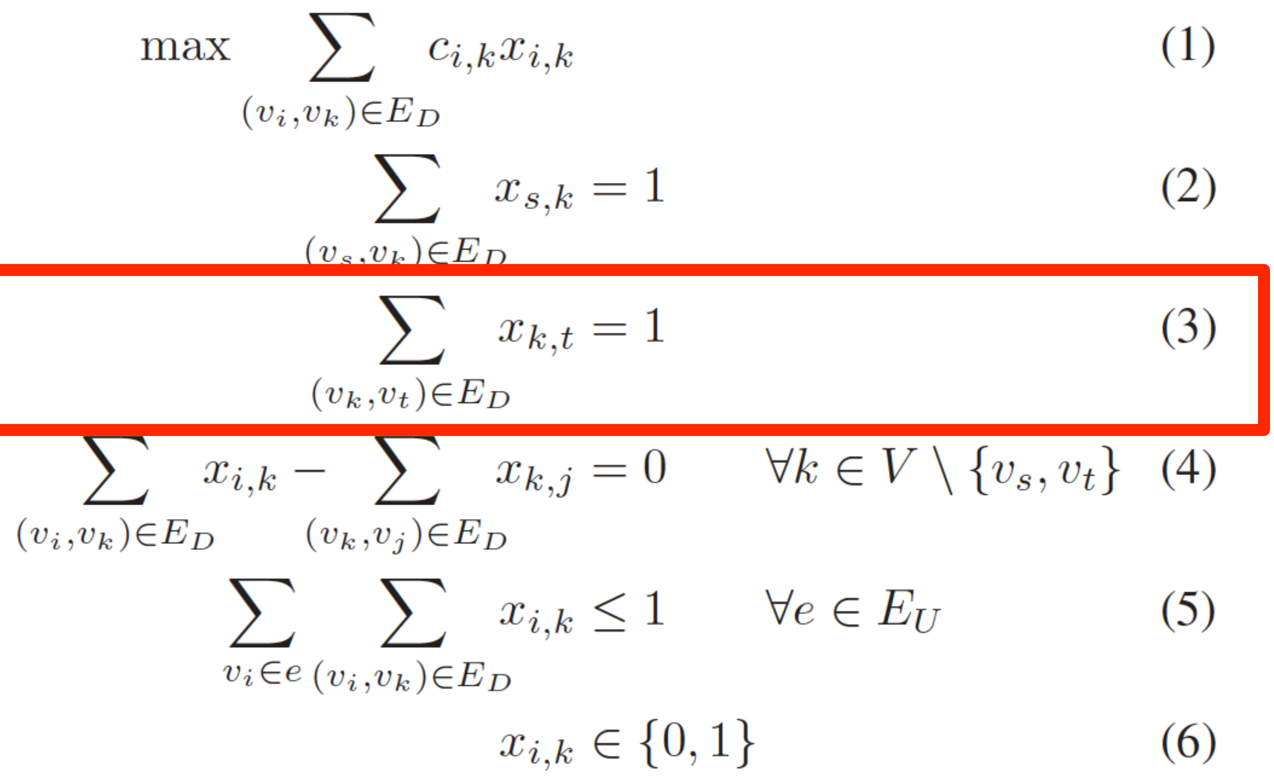


$$
\begin{aligned}
\max \sum_{(v_i,v_k)\in E_D} & c_{i,k}x_{i,k} && && \text{(1)}\\
\boxed{\sum_{(v_s,v_k)\in E_D} x_{s,k} = 1} & && && \text{(2)}\\
\sum_{(v_k,v_t)\in E_D} x_{k,t} &= 1 && && \text{(3)}\\
\sum_{(v_i,v_k)\in E_D} x_{i,k} - \sum_{(v_k,v_j)\in E_D} x_{k,j} &= 0 && \forall k \in V \setminus \{v_s, v_t\} && \text{(4)}\\
\sum_{v_i\in e}\sum_{(v_i,v_k)\in E_D} x_{i,k} &\leq 1 && \forall e \in E_U && \text{(5)}\\
x_{i,k} &\in \{0,1\} && && \text{(6)}
\end{aligned}
$$

Andreotti, Klau, Reinert, IEEE TCCB (2011), 99, 159

ANTILOPE

…and ending in the target node *t*…



$$\max \sum_{(v_i,v_k)\in E_D} c_{i,k}x_{i,k} \tag{1}$$

$$\sum_{(v_s,v_k)\in E_D} x_{s,k} = 1 \tag{2}$$

$$\boxed{\sum_{(v_k,v_t)\in E_D} x_{k,t} = 1} \tag{3}$$

$$\sum_{(v_i,v_k)\in E_D} x_{i,k} - \sum_{(v_k,v_j)\in E_D} x_{k,j} = 0 \qquad \forall k \in V \setminus \{v_s, v_t\} \tag{4}$$

$$\sum_{v_i\in e}\sum_{(v_i,v_k)\in E_D} x_{i,k} \le 1 \qquad \forall e \in E_U \tag{5}$$

$$x_{i,k} \in \{0,1\} \tag{6}$$

Andreotti, Klau, Reinert, IEEE TCCB (2011), 99, 159

ANTELOPE

…that form a path from *s* to *t*…

Internal nodes of any path from *s* to *t* need to have exactly one incoming and one outgoing edge in any node they pass through. For nodes that are not part of the path, the number of incoming and outgoing edges has to be zero.

$$\max \sum_{(v_i,v_k)\in E_D}$$

$$\sum_{(v_s,v_k)\in E_D} x_{s,k} = 1 \qquad (2)$$

$$\sum_{(v_k,v_t)\in E_D} x_{k,t} = 1 \qquad (3)$$

$$\boxed{\sum_{(v_i,v_k)\in E_D} x_{i,k} - \sum_{(v_k,v_j)\in E_D} x_{k,j} = 0 \qquad \forall k \in V \setminus \{v_s, v_t\} \qquad (4)}$$

$$\sum_{v_i\in e} \sum_{(v_i,v_k)\in E_D} x_{i,k} \leq 1 \qquad \forall e \in E_U \qquad (5)$$

$$x_{i,k} \in \{0,1\} \qquad (6)$$

ANTILOPE

…and are antisymmetric.

Two nodes that are connected by an undirected edge *e* in E_U may not be selected at the same time.

$$
\begin{aligned}
\max \sum_{(v_i,v_k)\in E_D} c_{i,k} x_{i,k} & & (1)\\
\sum_{(v_s,v_k)\in E_D} x_{s,k} = 1 & & (2)\\
\sum_{(v_k,v_t)\in E_D} x_{k,t} = 1 & & (3)\\
\sum_{(v_i,v_k)\in E_D} x_{i,k} - \sum_{(v_k,v_j)\in E_D} x_{k,j} = 0 & \quad \forall k \in V \setminus \{v_s, v_t\} & (4)\\
\sum_{v_i \in e} \sum_{(v_i,v_k)\in E_D} x_{i,k} \le 1 & \quad \forall e \in E_U & (5)\\
x_{i,k} \in \{0,1\} & & (6)
\end{aligned}
$$

Andreotti, Klau, Reinert, IEEE TCCB (2011), 99, 159

ANTILOPE – Solving the ILP

- ANTILOPE uses Lagrangian relaxation to solve the ILP

$$\max \sum_{(v_i,v_k)\in E_D} c_{i,k} x_{i,k} \tag{1}$$

$$\sum_{(v_s,v_k)\in E_D} x_{s,k} = 1 \tag{2}$$

$$\sum_{(v_k,v_t)\in E_D} x_{k,t} = 1 \tag{3}$$

$$\sum_{(v_i,v_k)\in E_D} x_{i,k} - \sum_{(v_k,v_j)\in E_D} x_{k,j} = 0 \qquad \forall k \in V \setminus \{v_s, v_t\} \tag{4}$$

$$\sum_{v_i\in e} \sum_{(v_i,v_k)\in E_D} x_{i,k} \leq 1 \qquad \forall e \in E_U \tag{5}$$

$$x_{i,k} \in \{0,1\} \tag{6}$$

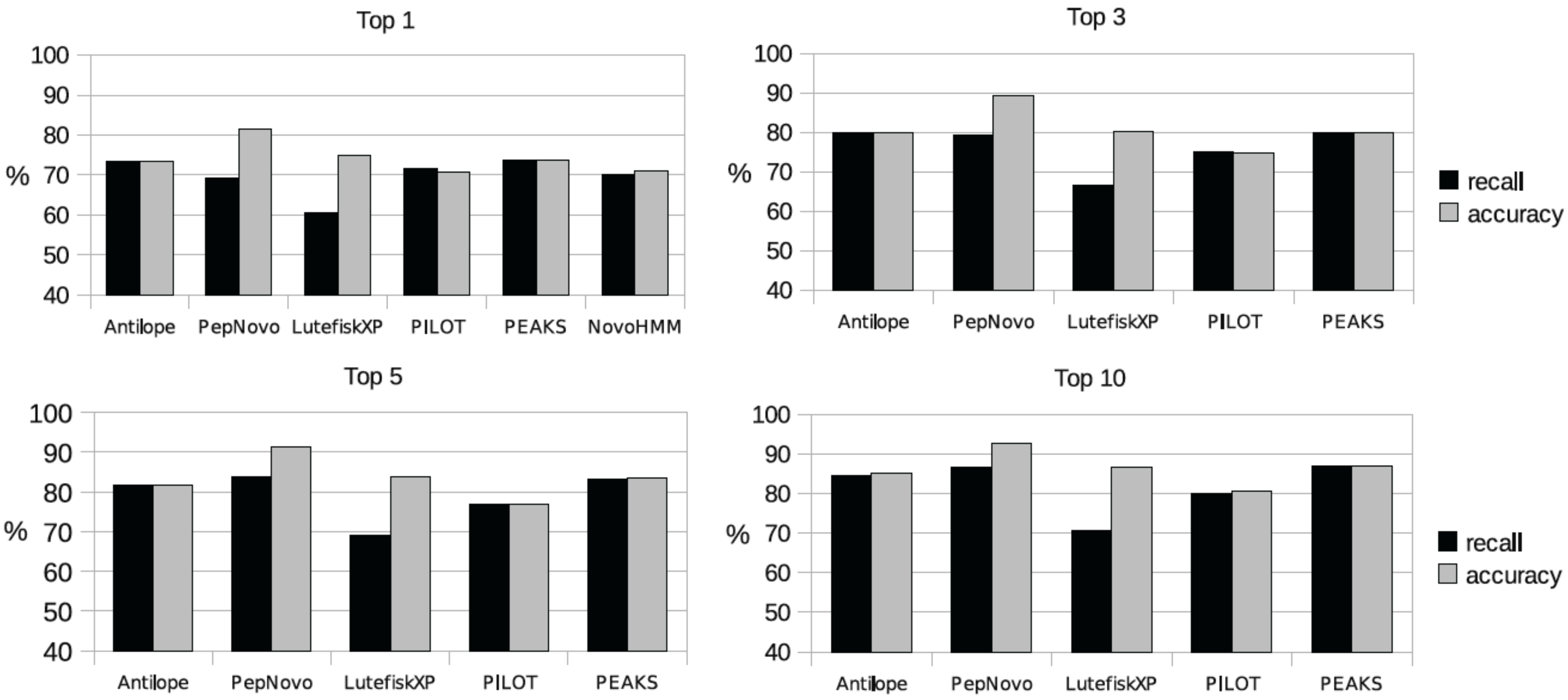
ANTILOPE – Scoring

- ANTILOPE uses a Bayesian network to score nodes
- Idea
  - Fragmentation events are not independent
  - Learn intensities for a specific ion type in the spectrum using a Bayes network (a machine learning method)
  - Learning is based on identified peptide spectra (e.g., through database search)
- Details of the scoring are beyond the scope of this lecture

Performance

- De novo peptide sequencing has still (even with high-resolution data)
  - Very low reliability
  - Large runtimes compared to database search
- It is usually employed as a method of last resort
  - If no genome/proteome sequence of an organism is known
  - For peptides that are not encoded genetically
- Top ranked hits are rarely correct, but usually contain correct subsequences

Performance of De Novo Seqencing



Andreotti, Klau, Reinert, IEEE TCCB (2011), 99, 159

Multisequences

- The lack of completeness of CID fragmentation makes de novo sequencing difficult
- In most cases, we thus obtain **multisequences** for parts with missing peaks
- Example:
  - S(GA|AG|V)K is a multisequence corresponding to one of the isobaric sequences SGAK, SAGK, or SVK
  - If no fragment ion between the second and third amino acid is observed, the three options cannot be kept apart
  - Similarly, I and L and (depending on the resolution) Q and L are isobaric

LEARNING UNIT 8D COMPLEMENTARY FRAGMENTATION FOR DE NOVO ID

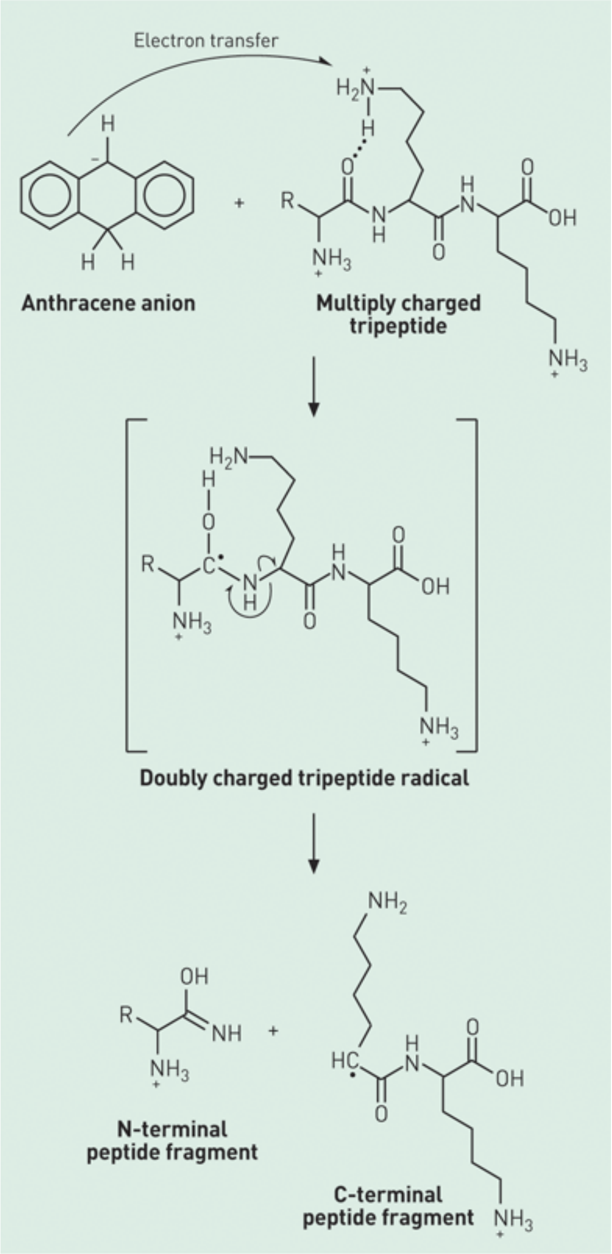
- Electron transfer dissociation (ETD)
- Comparison fragmentation statistics of ETD and CID
- CompNovo algorithm

Complementary Fragmentation

- The issue of missing information/peaks cannot be addressed by computational means
- One way to address this problem is the use of complementary fragmentation methods:
  - Fragment the eluting peptide with two different methods (e.g., CID and ETD)
  - The different methods have different preferences for fragmentation and chances are that missing peaks will be at different backbone positions in both spectra
  - The search algorithm then has to deal with two types of spectra and needs to be adapted accordingly
- Disadvantages
  - Only few mass spectrometers are equipped to record complementary fragmentation types
  - Recording twice as many spectra reduces the total number of peptides fragmented

Electron Transfer Dissociation (ETD)

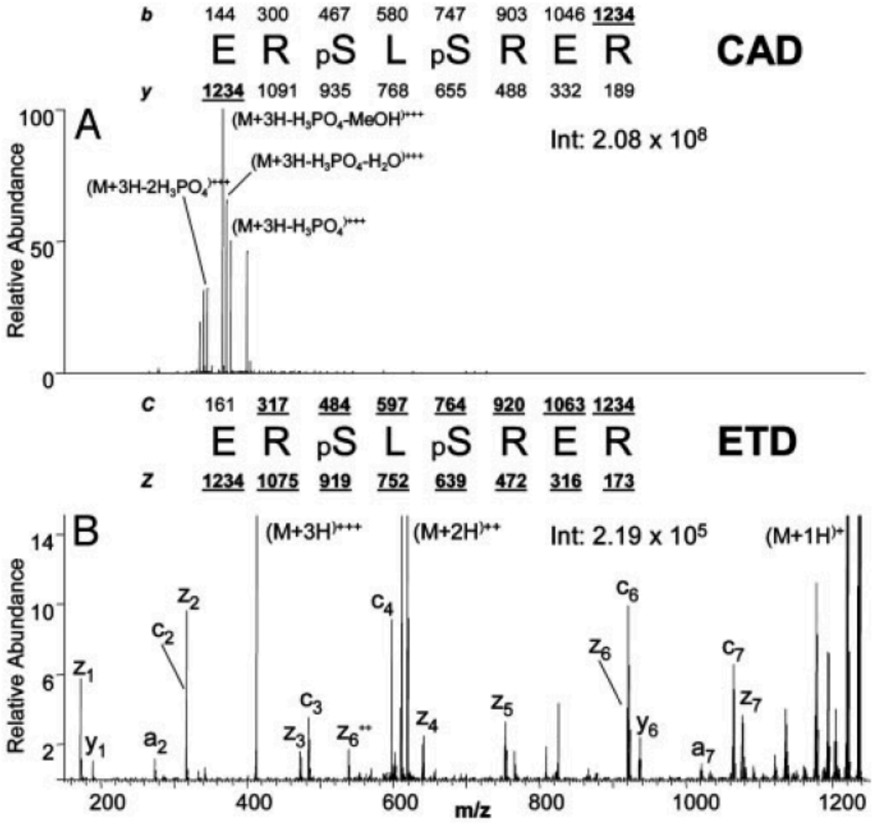
- Electron Transfer Dissociation (ETD) uses an organic compound (usually anthracene) as a charge transfer agent
- Anthracene is (negatively) charged and transfers an electron to multiply charged peptides in the collision cell
- The resulting fragmentation mechanism differs from the fragmentation observed in CID
- Consequently, different ion series are observed
- ETD produces mostly c and z ions


Electron transfer
Anthracene anion
Multiply charged tripeptide
Doubly charged tripeptide radical
N-terminal peptide fragment
C-terminal peptide fragment

Borman, C&EN (2004), 82(28):22-23

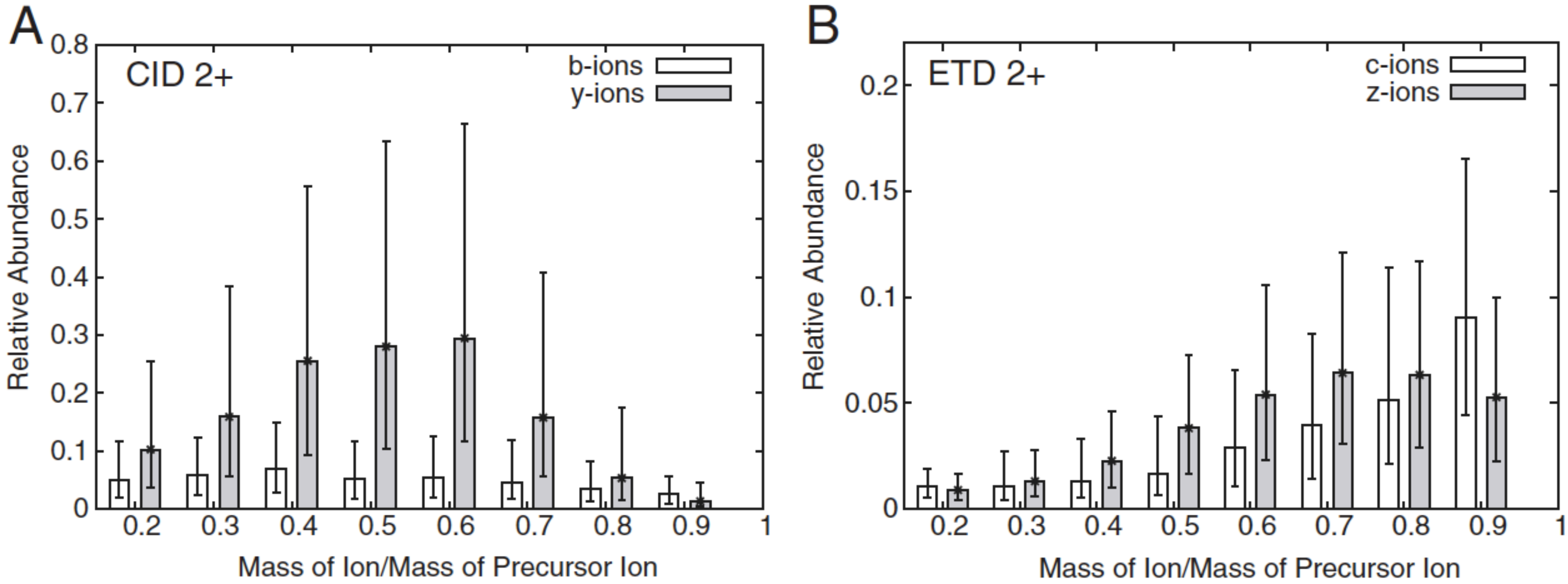
Electron Transfer Dissociation (ETD)

- ETD leads to a different fragmentation pattern as CID/CAD
- The example on the right shows fragmentation patterns of the same peptide for CAD and ETD
- In particular for modified peptides (phosphopeptides) ETD produces more complete fragmentation patterns than CID



Syka et al., Proc Natl Acad Sci U S A. (2004), 101(26): 9528–9533.

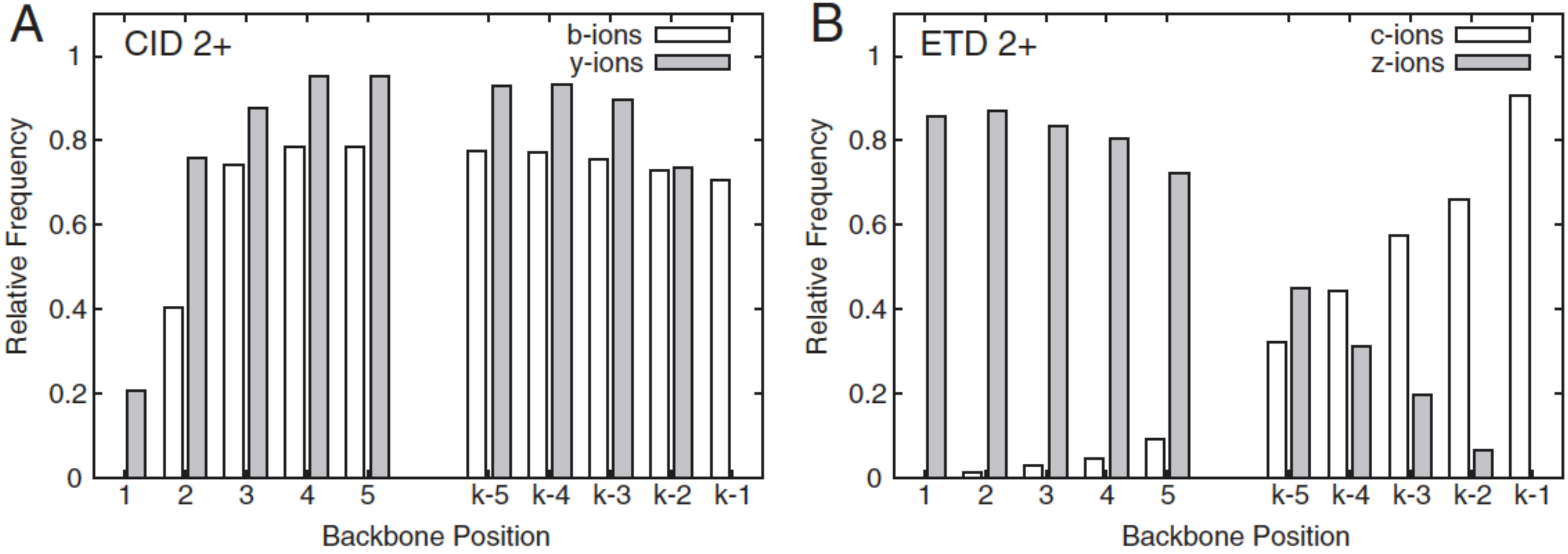
Complementary Fragmentation



- CID spectra preferentially produce b/y ions, whereas ETD spectra produce mostly c/z ions
- As can be seen on the left, CID spectra fragment preferentially around the middle of the peptide
- ETD spectra preferentially fragment asymmetrically with a higher likelihood of forming fragments towards the ends
- The two fragmentation techniques thus produce nicely complementary information

Bertsch et al., Electrophoresis (2009), 30(21), 3736-3747.

Complementary Fragmentation

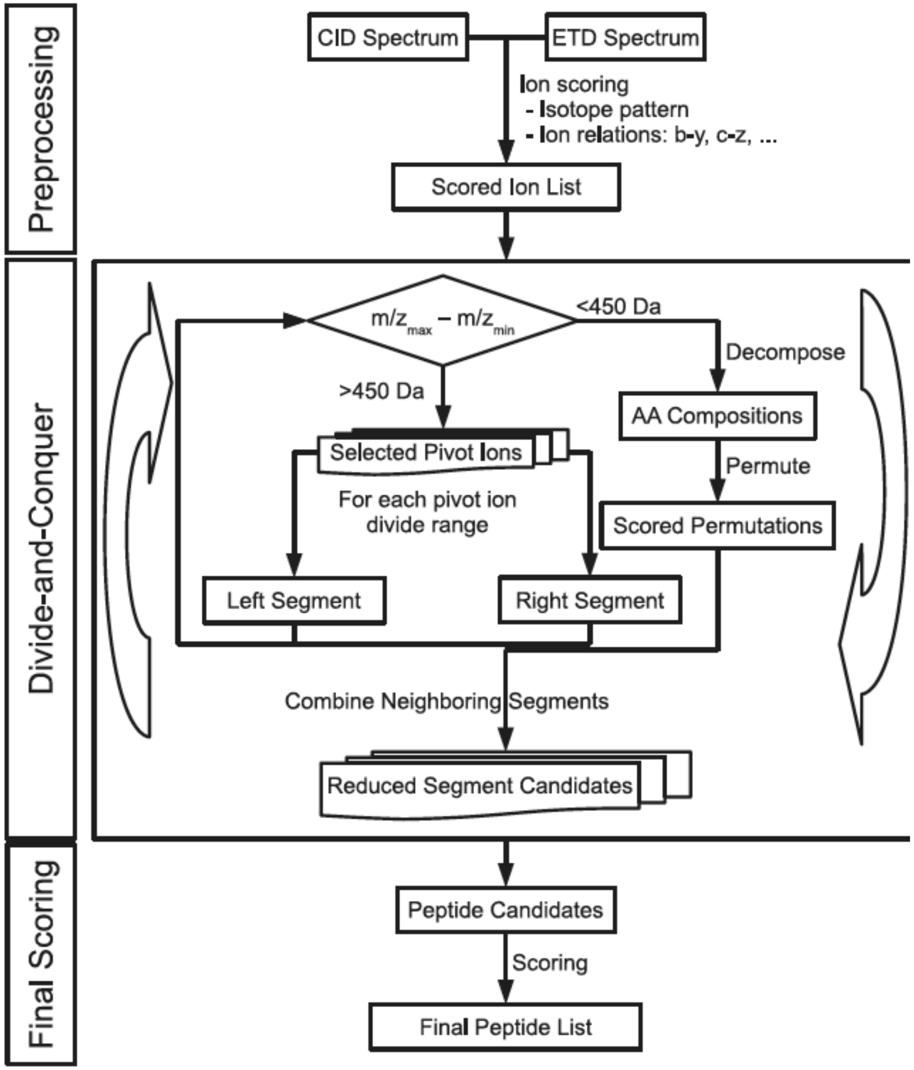


- The complementarity is also obvious when looking at fragmentation frequencies observed as a function of the backbone position (ion trap data)
- ETD yields information for the C and N terminus and CID provides more information in the middle of the peptide
- Together a larger coverage of the whole peptide sequence is achieved

Bertsch et al., Electrophoresis (2009), 30(21), 3736-3747.

CompNovo

- **CompNovo** (Bertsch et al., 2009) uses pairs of CID and ETD spectra (**Comp**lementary fragmentation methods, hence the name)
- The spectrum is decomposed in a divide-and-conquer approach into smaller parts
- For each part of the spectrum below a certain threshold (450 Da), we use a rapid mass decomposition (introduced later for metabolomics) to enumerate all possible sequences
- Possible subsequences are combined and then scored against the experimental spectra



Bertsch et al., Electrophoresis (2009), 30(21), 3736-3747.

CompNovo - Performance

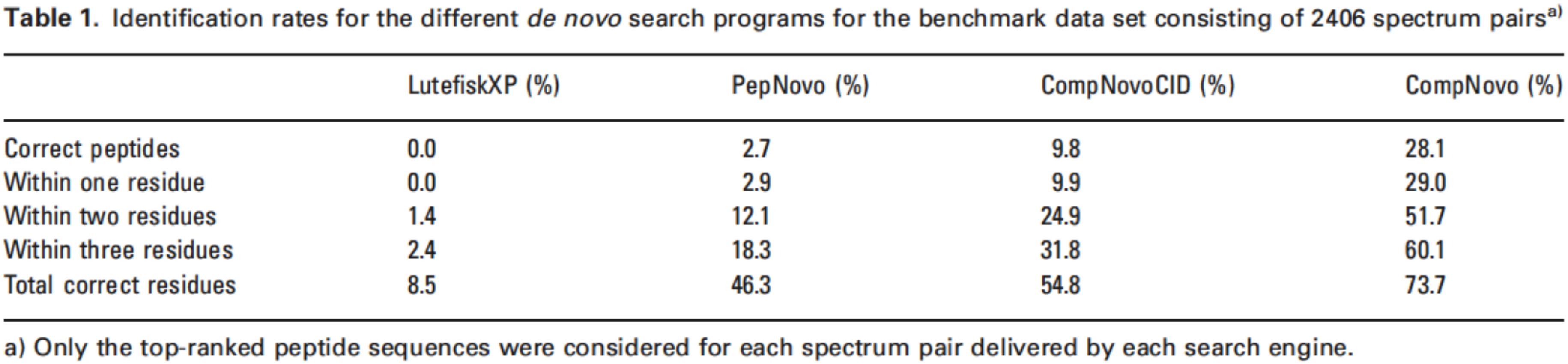


Table 1. Identification rates for the different *de novo* search programs for the benchmark data set consisting of 2406 spectrum pairs[a)]

	LutefiskXP (%)	PepNovo (%)	CompNovoCID (%)	CompNovo (%)
Correct peptides	0.0	2.7	9.8	28.1
Within one residue	0.0	2.9	9.9	29.0
Within two residues	1.4	12.1	24.9	51.7
Within three residues	2.4	18.3	31.8	60.1
Total correct residues	8.5	46.3	54.8	73.7

a) Only the top-ranked peptide sequences were considered for each spectrum pair delivered by each search engine.

- Not surprisingly, CompNovo achieves drastically improved identification rates than other de novo sequencing tools
- Note the the other tools cannot use information from the ETD spectra
- CompNovoCID is a version of CompNovo using only CID spectra
- Only inclusion of the complementary fragmentation information can yield good identification rates

Bertsch et al., Electrophoresis (2009), 30(21), 3736-3747.

CompNovo - Performance

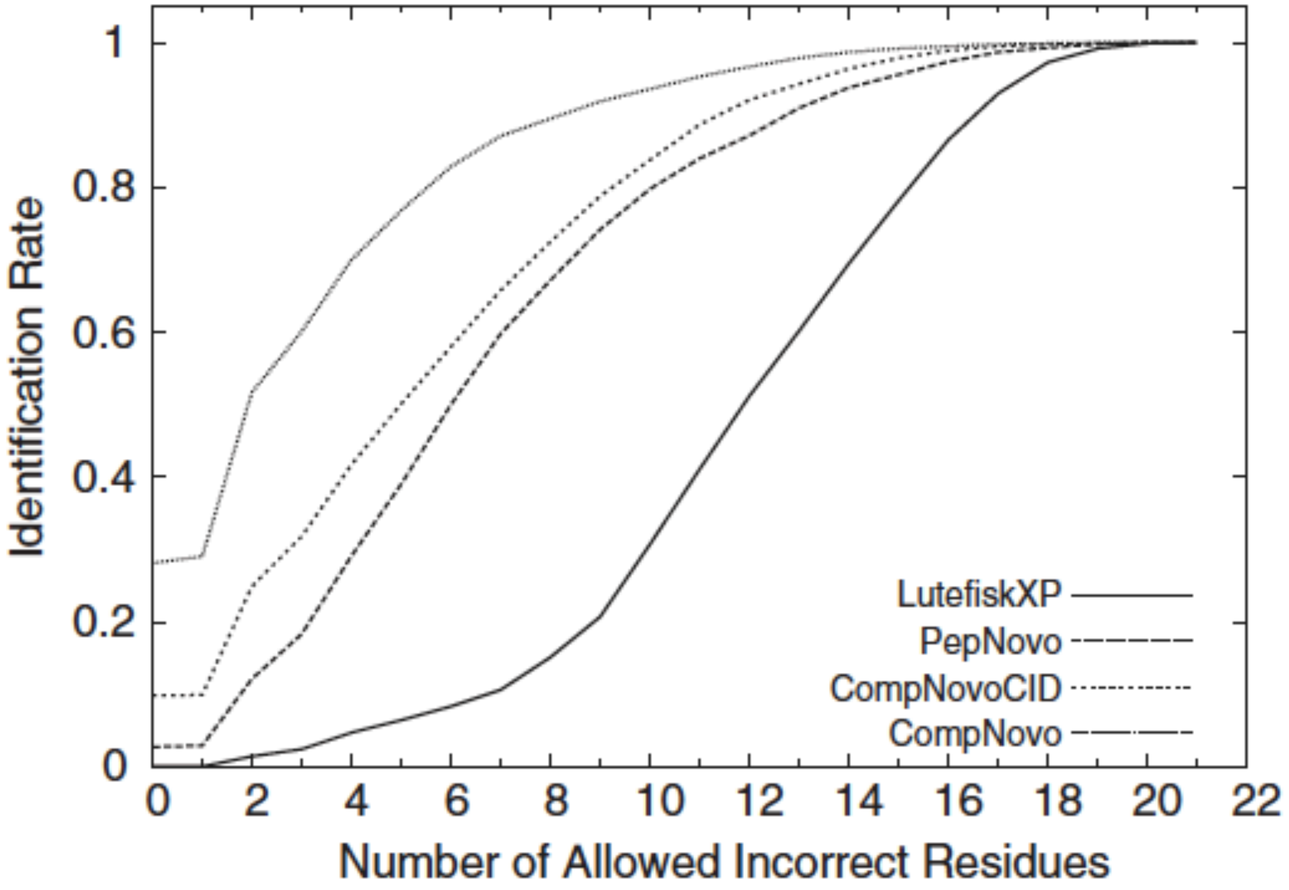


Figure 7. Identification rates of the different algorithms depending on the number of tolerated false amino acid annotations in the top-ranked peptide sequence (benchmark data set, $n = 2406$).

Bertsch et al., Electrophoresis (2009), 30(21), 3736-3747.

Online Materials

- Learning Units 8A-D

References

- **Manual interpretation**
  - Seidler et al., De novo sequencing of peptides by MS/MS, Proteomics (2010), 10:634-649
- **Definition of antisymmetric paths**
  - C. Liu, Y. Song, B. Yan, Y. Xu, and L. Cai, Fast de novo peptide sequencing and spectral alignment via tree decomposition, in Proc. 11th Pacific Symp Biocomp (PSB). World Scientific, 2006, pp. 255–266.

    http://helix-web.stanford.edu/psb06/liu.pdf
- **ANTILOPE**
  - S. Andreotti, G. Klau, and K. Reinert, "Antilope – A Lagrangian Relaxation Approach to the de novo Peptide Sequencing Problem," IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 99:159.

    http://doi.ieeecomputersociety.org/10.1109/TCBB.2011.59

    http://arxiv.org/pdf/1102.4016v1
- **CompNovo**
  - A. Bertsch, A. Leinenbach, A. Pervukhin, M. Lubeck, R. Hartmer, C. Baessmann, Y. A. Elnakady, R. Müller, S. Böcker, C. Huber, and Oliver Kohlbacher, "De novo peptide sequencing by tandem MS using complementary CID and electron transfer dissociation," Electrophoresis (2009), 30(21), 3736-3747.