

Tandem Mass Spectrometry:

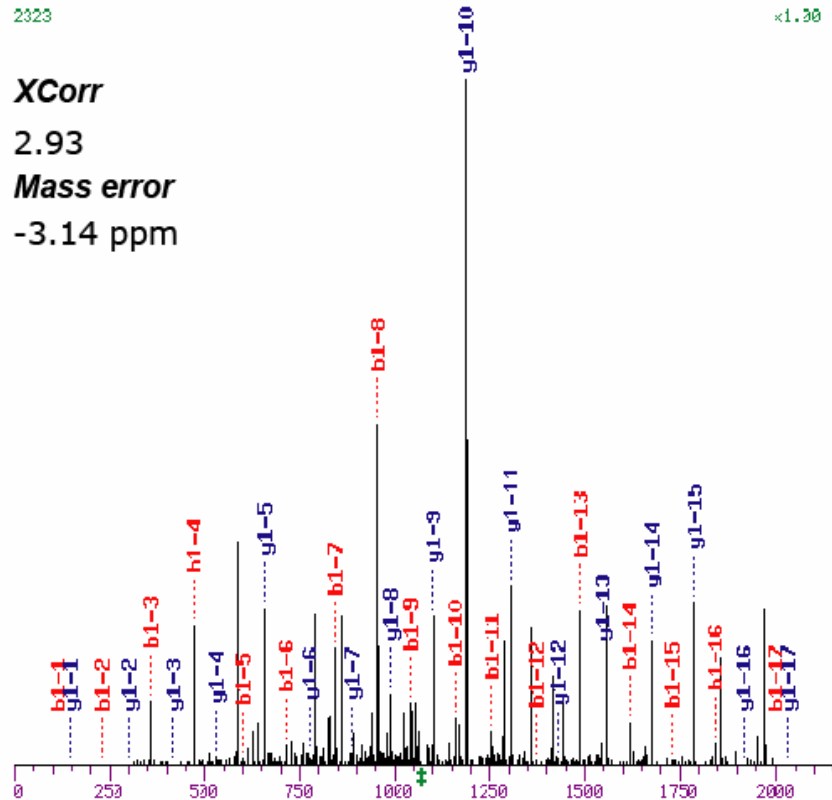
Generating function, alignment and assembly

With slides from Sangtae Kim and from Jones & Pevzner 2004

Determining reliability of identifications

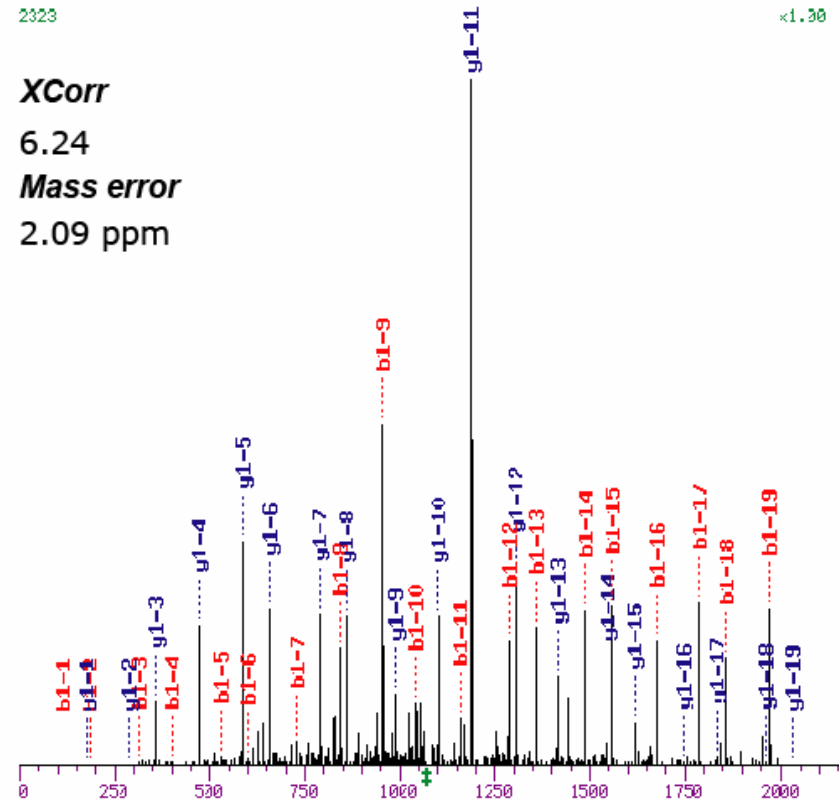
- Can we use Target/Decoy to estimate quality of de novo?
- Can we use de novo to improve Target/Decoy?

a



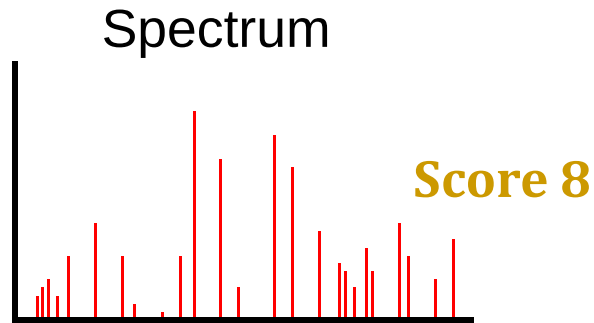
From Elias'07 LDELQDKLSNVDNELNRK

b



IAQSEAEISDEAQLALR

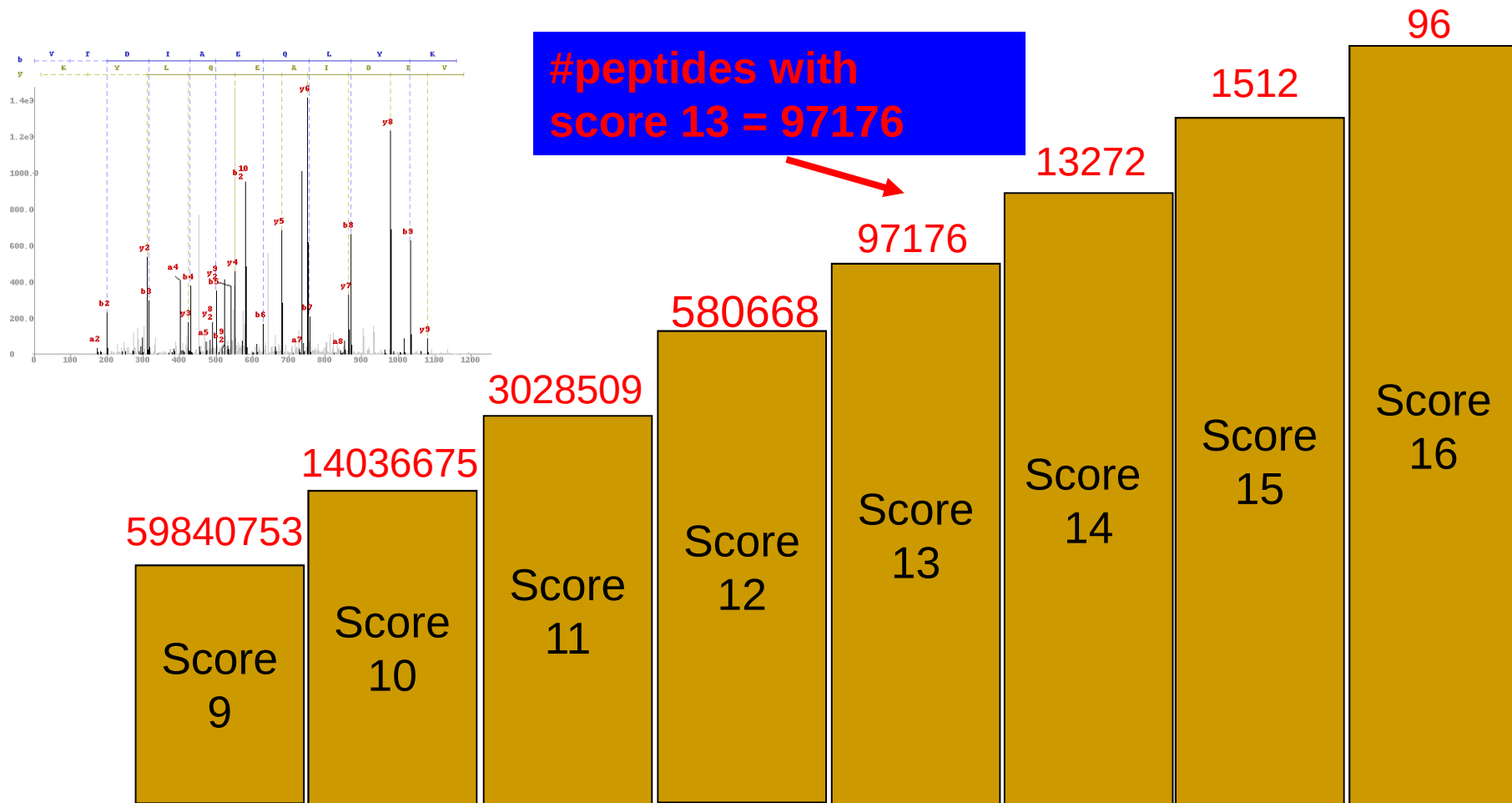
Generating function: Scoring *all* peptides



All peptides

Score<7 Score=7 Score=8 Score=9 Score=10

Score Histogram of All Peptides

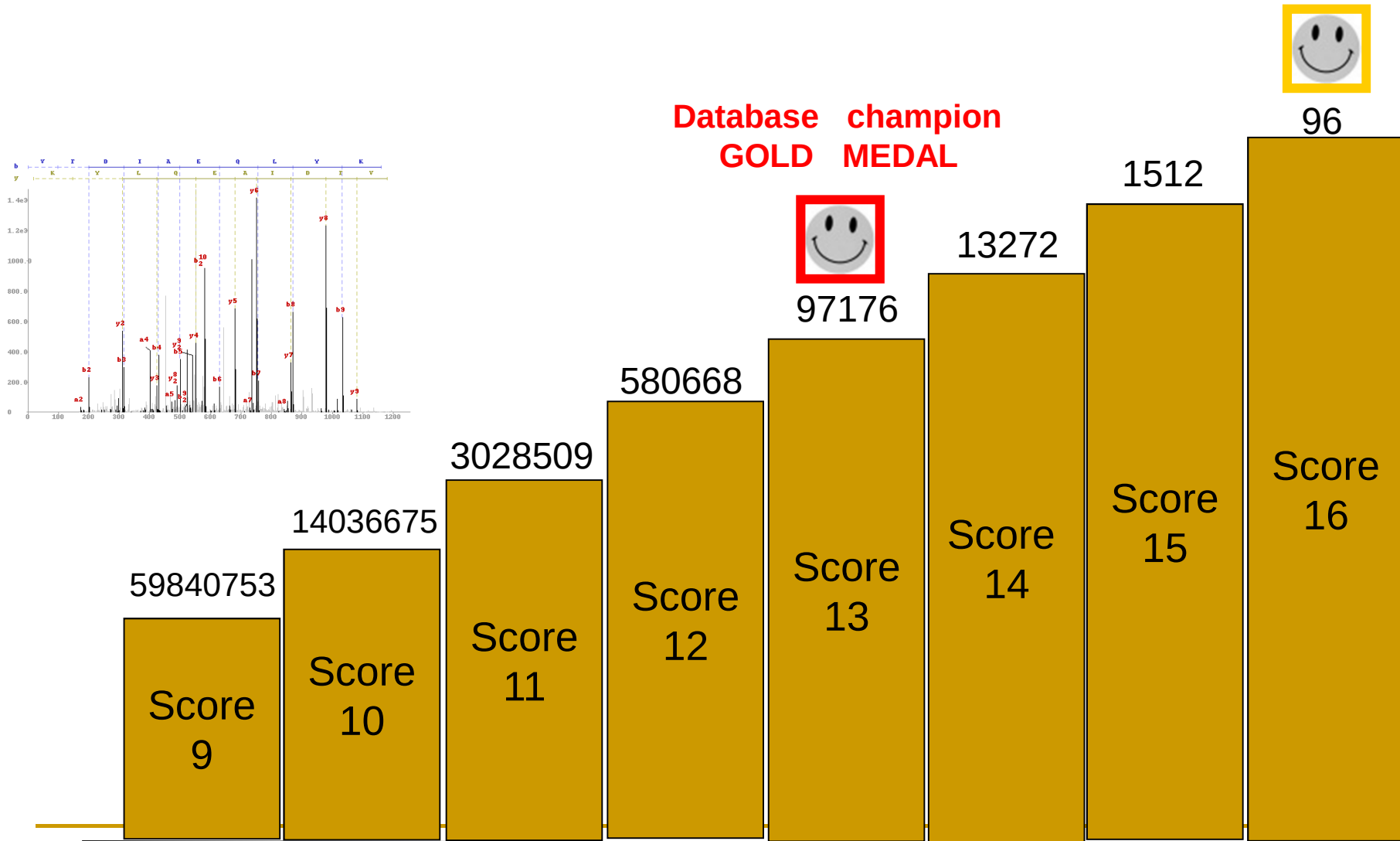


The generating function for the simplified peptide-spectrum score:

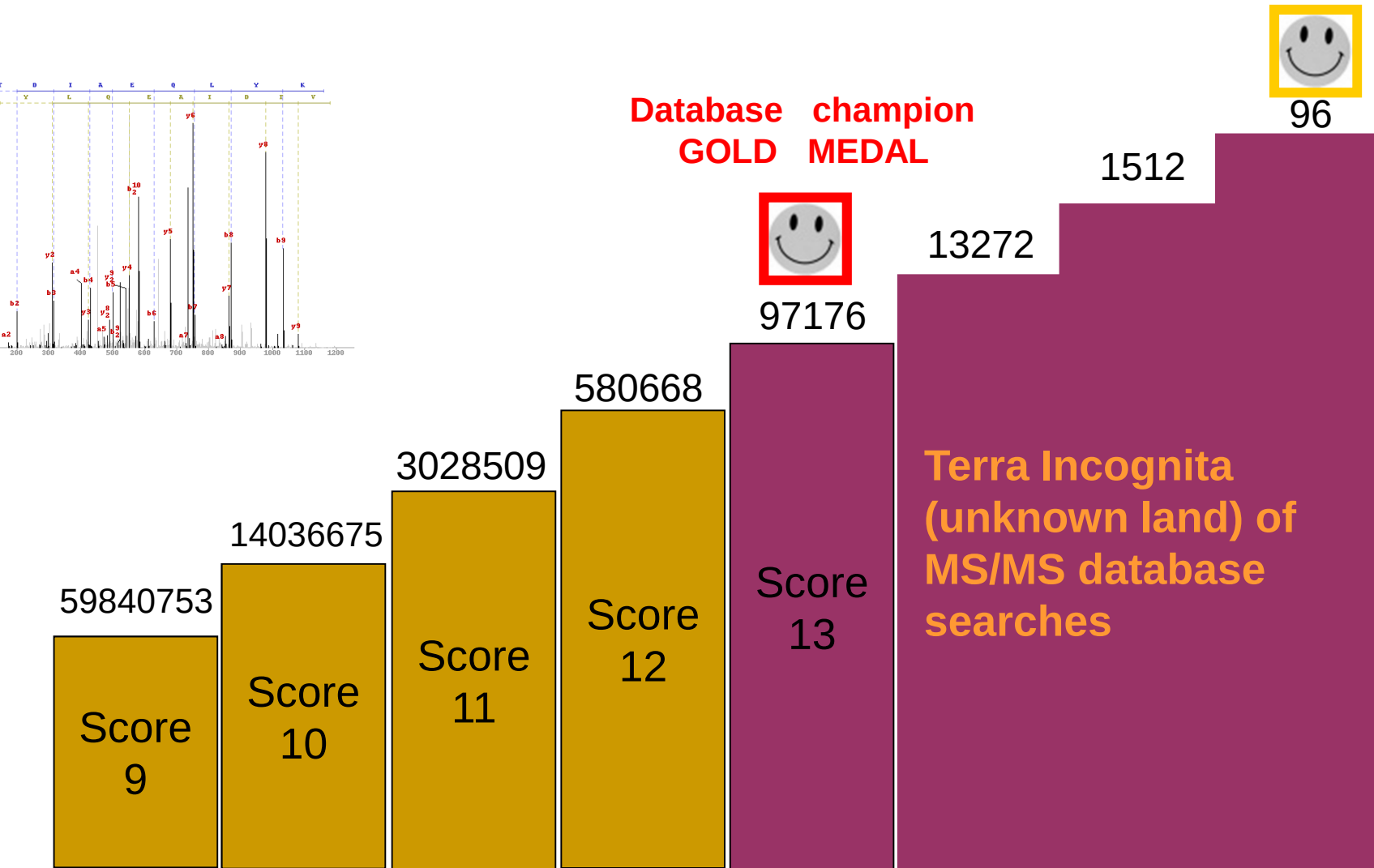
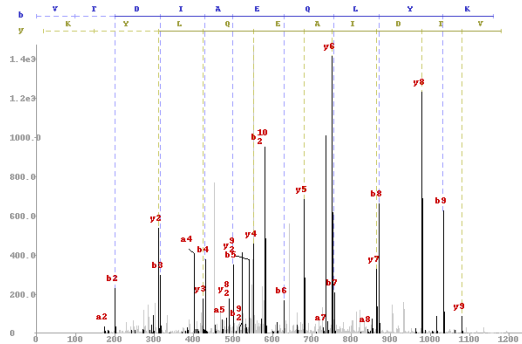
$$\text{Score}(\text{Peptide}, \text{Spectrum}) = \#b/y \text{ ions in the } \text{Spectrum} \text{ explained by the } \text{Peptide}$$

Peptide Identification: A Very Crowded Race

WORLD CHAMPION



MS-GF – New Scoring for Peptide-Spectrum Matches (PSM)

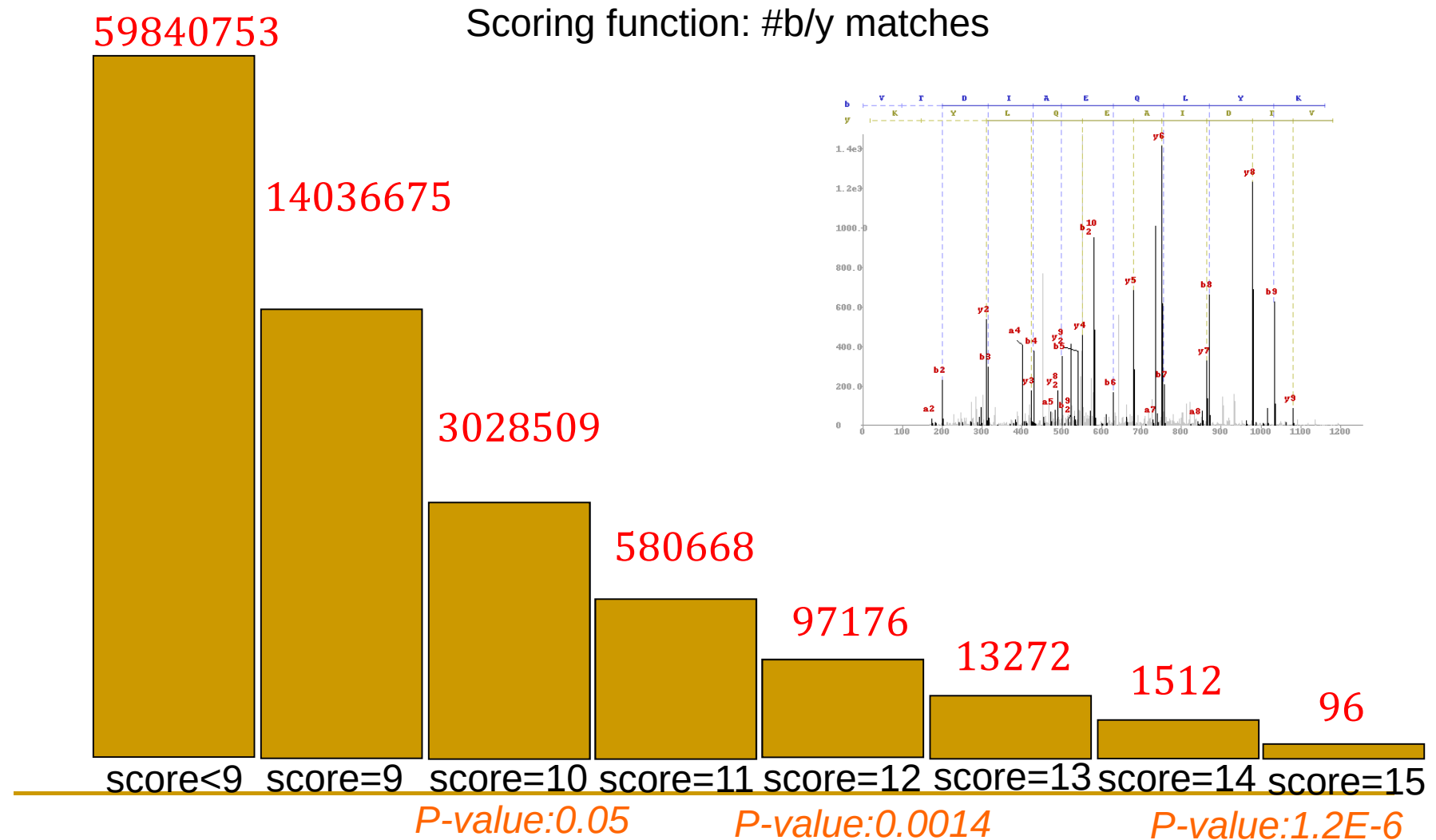


MS-GF scoring for Peptide-Spectrum Matches:

score = total **WEIGHTED** number of peptides in Terra Incognita (p-value of a PSM)

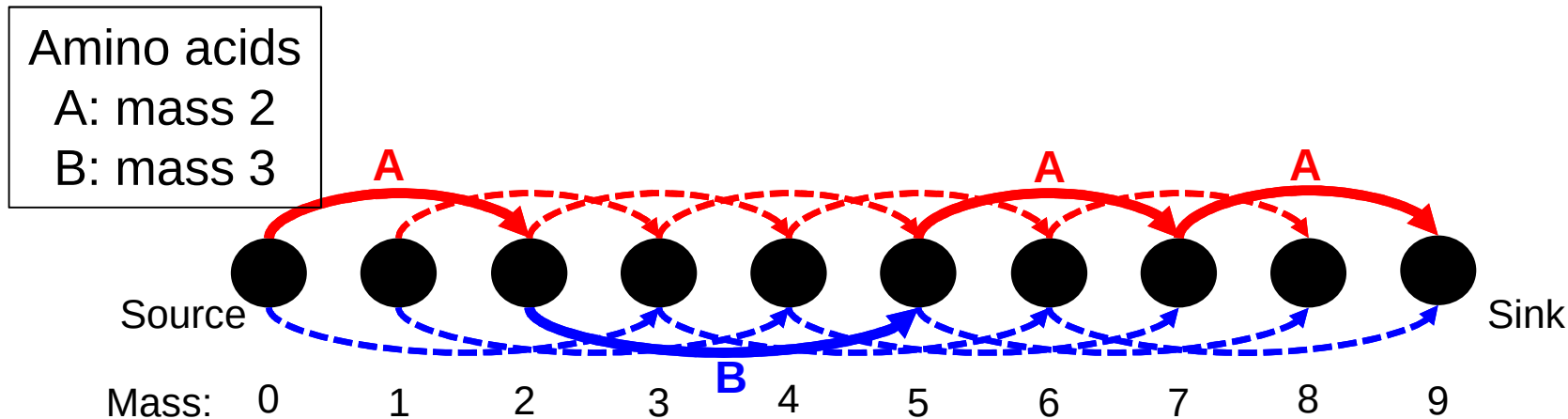
weight of a peptide of length n equals to $(1/20)^n$

Statistical Significance of DB Matches



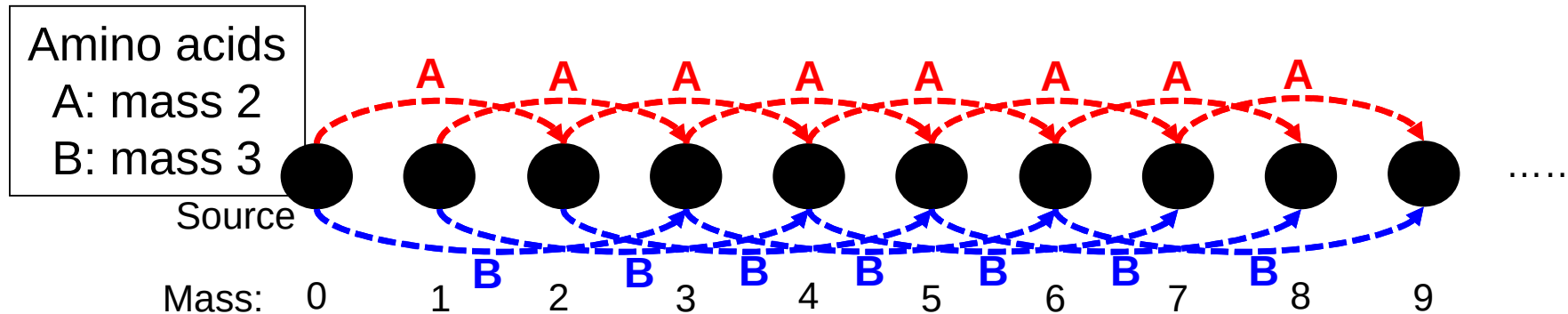
How to Compute the Generating Function?

Amino Acid Graph



- **Vertex: every mass** (spectrum graph: vertex per peak)
- Edge: two vertices are connected iff their masses differ by an amino acid mass
- **Every peptide has a corresponding path in the amino acid graph.**
- Proposed by Ma et al. (RPMS, 2003).
- PEAKS (Ma et al., RCMS 2003), MS-Novo (Mo et al., JPR 2007), MS-Dictionary (Kim et al., MCP 2009a)

De Novo Sequencing Using Graphs

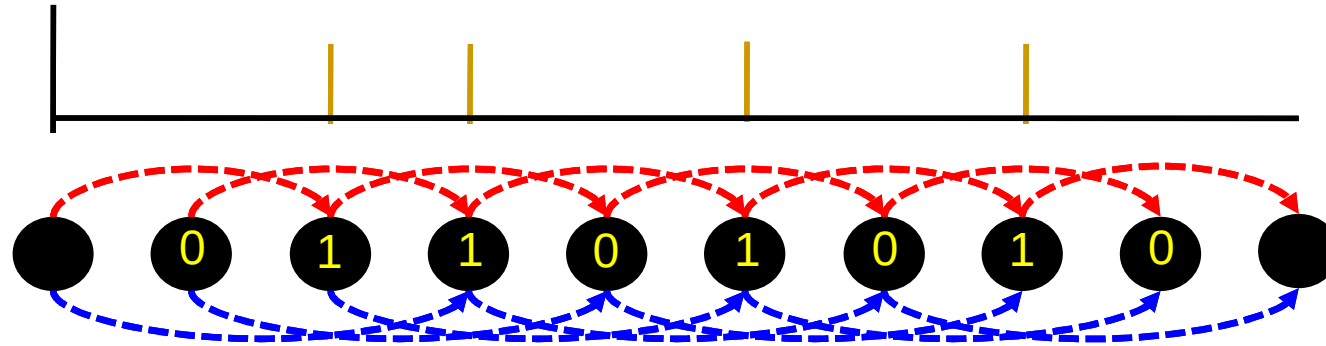


- All *paths* in amino acid graphs: *all peptides*
- Scores are embedded in vertices (and/or edges), not paths
 - Scoring function must be additive
- The Longest Path Algorithm explores the exponential number of paths in linear time.
- Time complexity?

Computing Score Distribution of All Peptides

Amino acids
A: mass 2
B: mass 3

NodeScore:



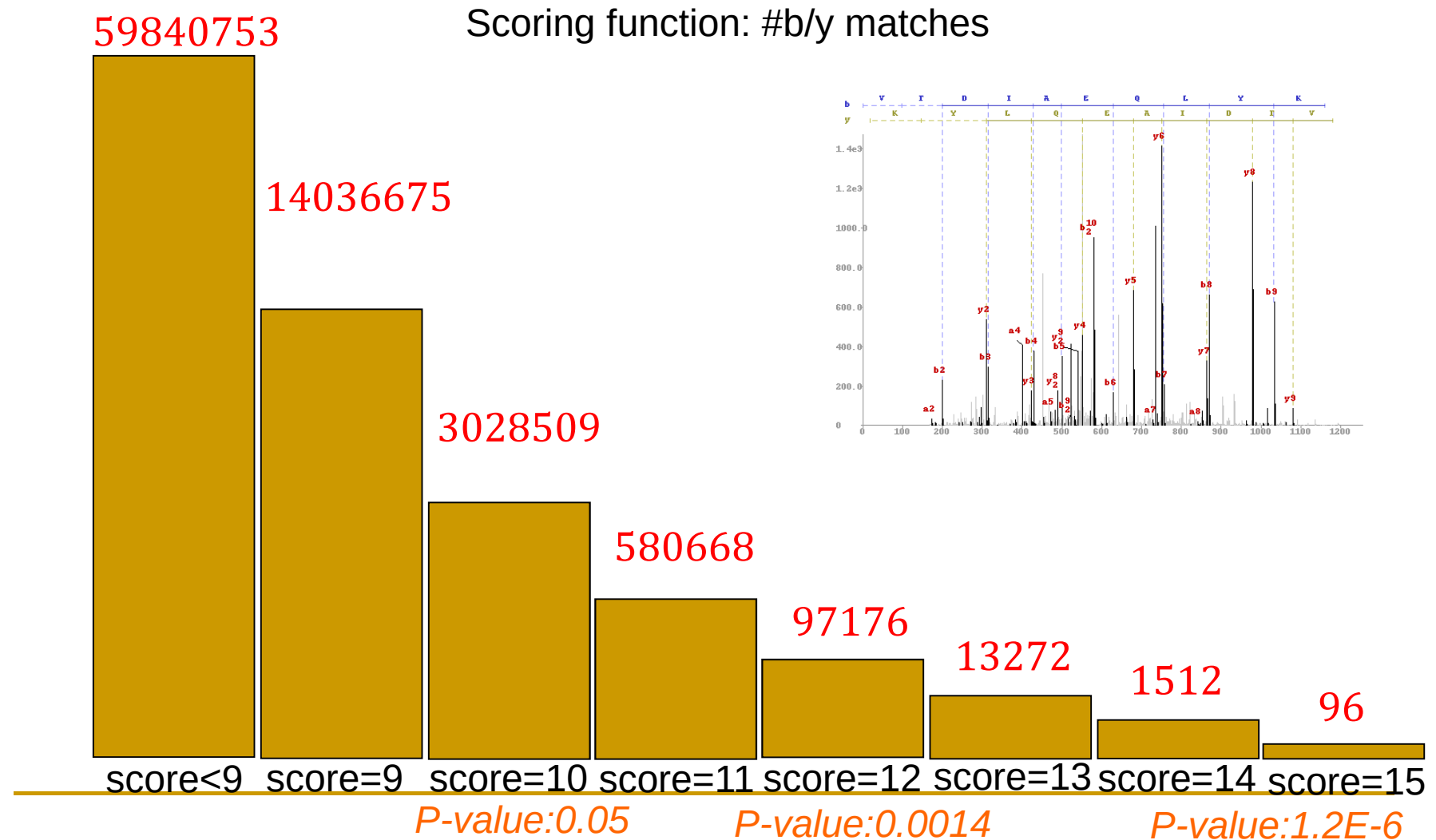
NodeScore	0	0	1	1	0	1	0	1	0	0
Score=0	1	0	0	0	0	0	0	0	0	0
Score=1	0	0	1	1	1	0	2	0	2	2
Score=2	0	0	0	0	0	2	0	1	2	1
Score=3	0	0	0	0	0	0	0	2	0	2

- Compute the score distribution of all peptides.
- Each node stores a score distribution instead of a maximum score.
- Can set edge weights to 0.05 (1/20) to determine probabilities instead of peptide counts

Recursion?

Kim et al., JPR 2008

FPR Statistical Significance of DB Matches



Assessing significance of DB matches

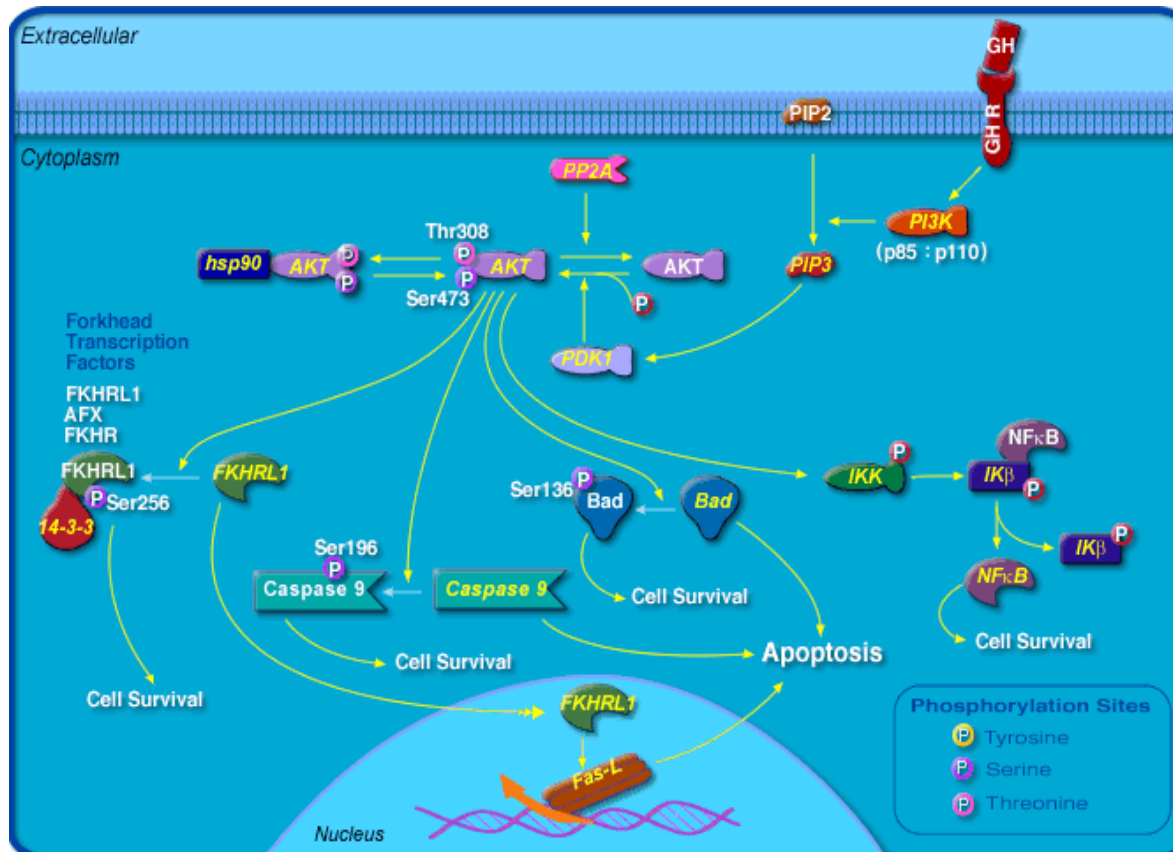
■ Generating Function

- ❑ Main purpose is to determine the significance of a peptide match to a *single* spectrum in the context of all other possible peptides
- ❑ In practice, used to make match scores comparable across peptide-spectrum matches

■ False Discovery Rate (FDR)

- ❑ Main purpose is to correct for multiple hypothesis testing – select significant Peptide-Spectrum matches for a *set* of spectra
-

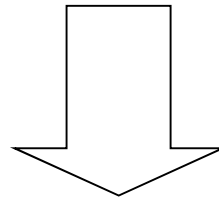
The dynamic nature of the proteome



- The proteome of the cell is changing
- Various extra-cellular, and other signals activate pathways of proteins.
- A key mechanism of protein activation is **post-translational modification (PTM)**
- These pathways may lead to other genes being switched on or off
- Mass spectrometry is key to probing the proteome and detecting PTMs

Post-Translational Modifications

Proteins are involved in cellular signaling and metabolic regulation.

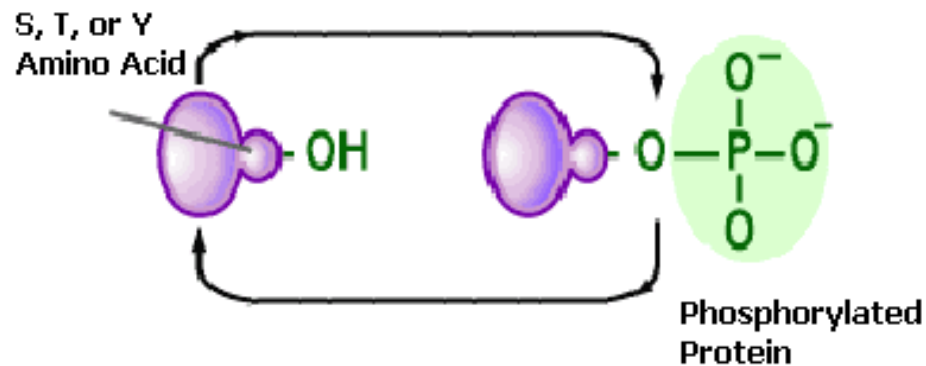


They are subject to a large number of biological modifications.

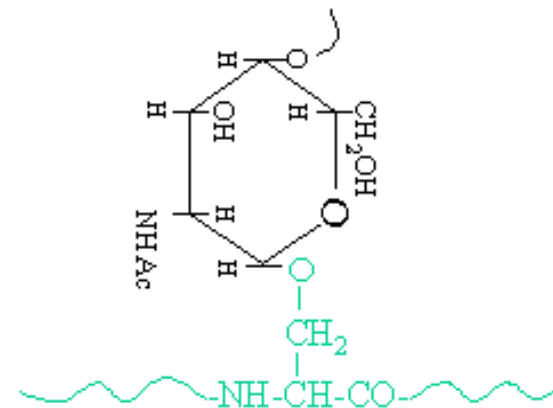
Almost all protein sequences are post-translationally modified and **500+ types of modifications** of amino acid residues are known.

Examples of Post-Translational Modification

Phosphorylation



Glycosylation



Post-translational modifications increase the number of “letters” in amino acid alphabet and lead to a combinatorial explosion in both database search and de novo approaches.

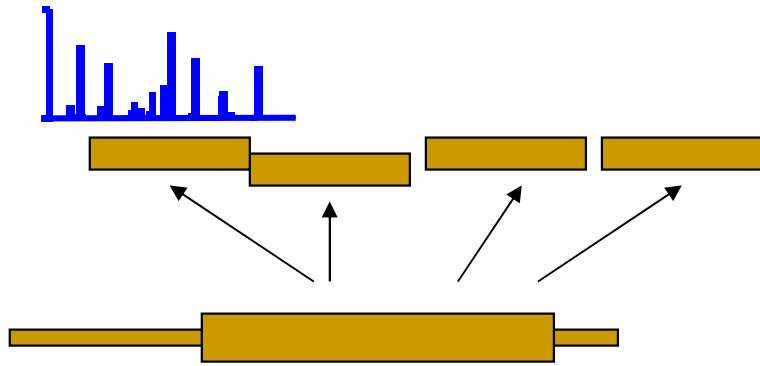
Search for Modified Peptides: Virtual Database Approach

Yates et al.,1995: an exhaustive search in a virtual database of all modified peptides.

Exhaustive search leads to a large combinatorial problem, even for a small set of modifications types.

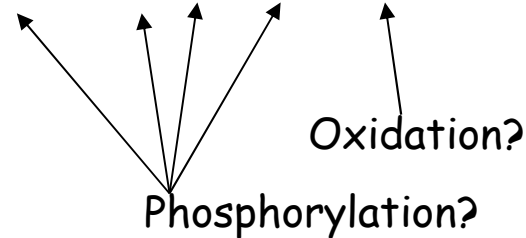
Problem (Yates et al.,1995). Extend the virtual database approach to a large set of modifications.

Exhaustive Search for modified peptides.



- For each peptide, generate all modifications.
- Score each modification.

YFDSTDYNMAK



$2^5=32$ possibilities, with 2 types of modifications!

Peptide Identification Problem Revisited

Goal: Find a peptide from the database with maximal match between an experimental and theoretical spectrum.

Input:

- ❑ S : experimental spectrum
- ❑ database of peptides
- ❑ Δ : set of possible ion types
- ❑ m : parent mass

Output:

- ❑ A peptide of mass m from the database whose theoretical spectrum matches the experimental S spectrum the best
-

Modified Peptide Identification Problem

Goal: Find a **modified** peptide from the database with maximal match between an experimental and theoretical spectrum.

Input:

- ❑ S : experimental spectrum
- ❑ database of peptides
- ❑ Δ : set of possible ion types
- ❑ m : parent mass
- ❑ **Parameter k (# of mutations/modifications)**

Output:

- ❑ A peptide of mass m that is **at most k mutations/modifications apart from** a database peptide and whose theoretical spectrum matches the experimental S spectrum the best
-

Database Search:

Sequence Analysis vs. MS/MS Analysis

Sequence analysis:

similar peptides (that a few mutations apart) have **similar** sequences

MS/MS analysis:

similar peptides (that a few mutations apart) have **dissimilar** spectra

Peptide Identification Problem: Challenge

Very similar peptides may have very different spectra!

Goal: Define a notion of spectral similarity that correlates well with the sequence similarity.

If peptides are a few mutations/modifications apart, the spectral similarity between their spectra should be high.

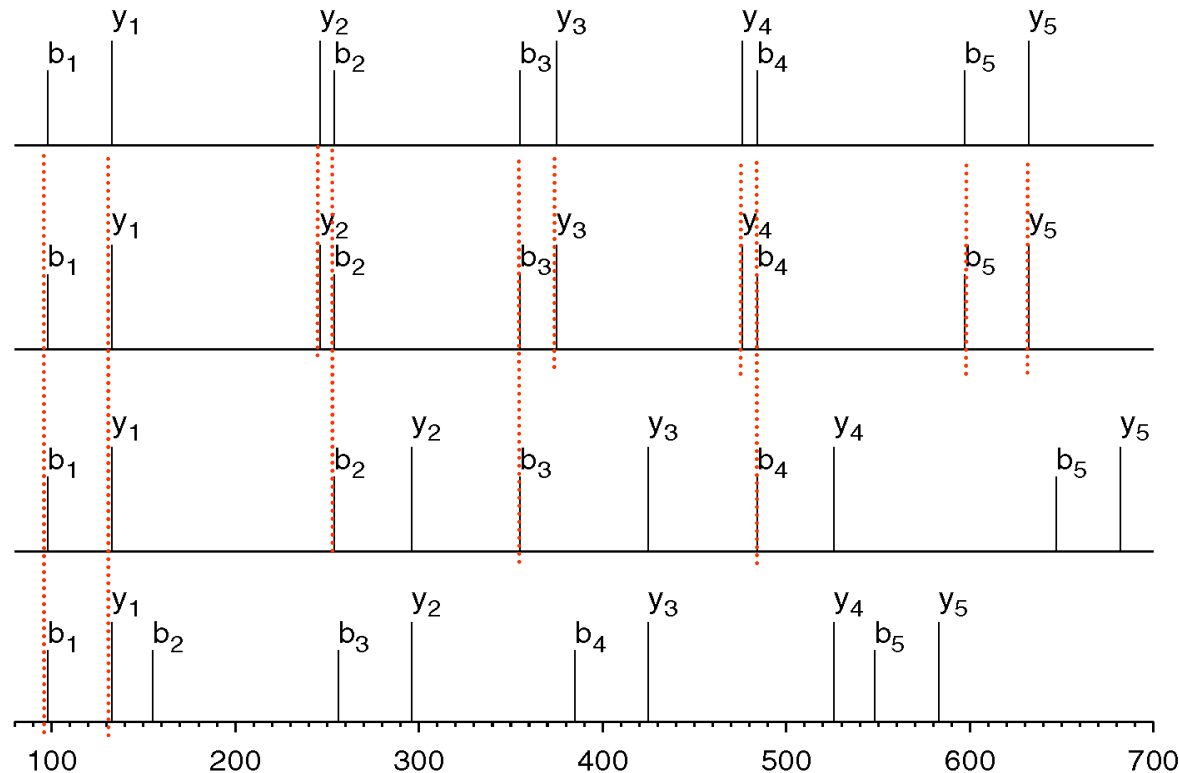
Deficiency of the Shared Peaks Count

Shared peaks count (SPC): intuitive measure of spectral similarity.

Problem: SPC diminishes very quickly as the number of mutations increases.

Only a small portion of correlations between the spectra of mutated peptides is captured by SPC.

SPC Diminishes Quickly



no mutations
SPC=10

1 mutation
SPC=5

2 mutations
SPC=2

$S(\text{PRTEIN}) = \{98, 133, 246, 254, 355, 375, 476, 484, 597, 632\}$

$S(\text{PRTEYN}) = \{98, 133, 254, 296, 355, 425, 484, 526, 647, 682\}$

$S(\text{PGTEYN}) = \{98, 133, 155, 256, 296, 385, 425, 526, 548, 583\}$

Spectral Convolution

$$S_2 \ominus S_1 = \{s_2 - s_1 : s_1 \in S_1, s_2 \in S_2\}$$

Number of pairs $s_1 \in S_1, s_2 \in S_2$ with $s_2 - s_1 = x$:

$$(S_2 \ominus S_1)(x)$$

The shared peaks count (SPC peak):

$$(S_2 \ominus S_1)(0)$$

Spectrum S_1 (PRTEIN)

	98	133	246	254	355	375	476	484	597	632
98	0	-35	-148	-156	-257	-277	-378	-386	-499	-534
133	35	0	-133	-121	-222	-242	-343	-351	-464	-499
254	156	121	8	0	-101	-121	-222	-230	-343	-378
296	198	163	50	42	-59	-79	-180	-188	-301	-336
355	257	222	109	101	0	-20	-121	-129	-242	-277
425	327	292	179	171	70	50	-51	-59	-172	-207
484	386	351	238	230	129	109	8	0	-113	-148
526	428	393	280	272	171	151	50	42	-71	-106
647	549	514	401	393	292	272	171	163	50	15
682	584	549	436	428	327	307	206	198	85	50

(a)

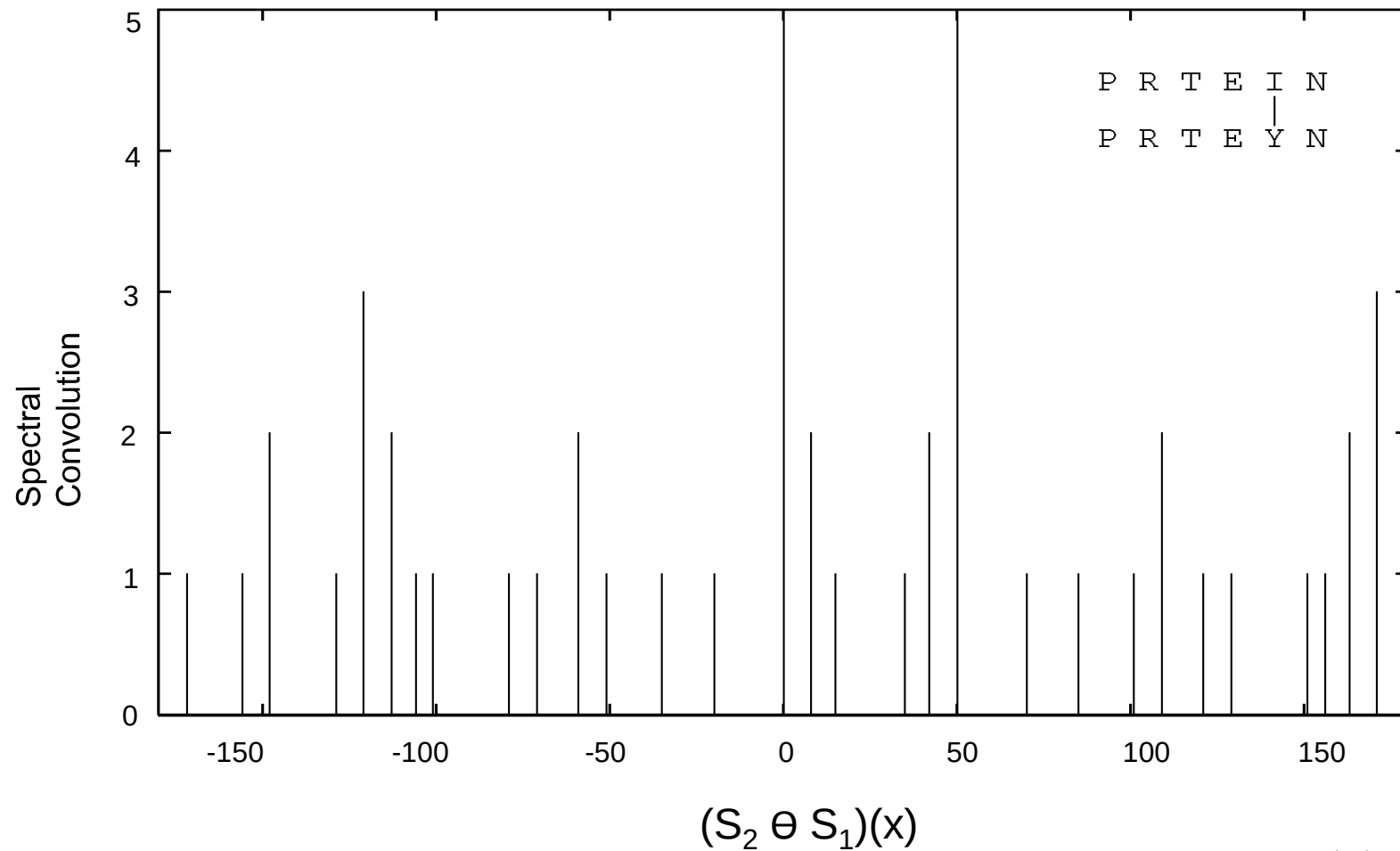
Spectrum S_1 (PTEIN)

	98	133	246	254	355	375	476	484	597	632
98	0	-35	-148	-156	-257	-277	-378	-386	-499	-534
133	35	0	-133	-121	-222	-242	-343	-351	-464	-499
284	186	151	38	30	-71	-91	-192	-200	-313	-348
296	198	163	50	42	-59	-79	-180	-188	-301	-336
385	287	252	139	131	30	10	-91	-99	-212	-247
425	327	292	179	171	70	50	-51	-59	-172	-207
514	416	381	268	260	159	139	38	30	-83	-118
526	428	393	280	272	171	151	50	42	-71	-106
677	579	544	431	423	322	302	201	193	80	45
712	614	579	466	458	357	337	236	228	115	80

(b)

Elements of $S_2 \ominus S_1$ represented as elements of a **difference matrix**. The elements with multiplicity >2 are colored; the elements with multiplicity $=2$ are circled. The SPC takes into account only the red entries

Spectral Convolution: An Example



$Mass(Y) = 163$

$Mass(I) = 113$

Spectral Comparison: Difficult Case

$$S = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$$

Which of the spectra

$$S' = \{10, 20, 30, 40, 50, 55, 65, 75, 85, 95\}$$

or

$$S'' = \{10, 15, 30, 35, 50, 55, 70, 75, 90, 95\}$$

fits the spectrum S the best?

SPC: both S' and S'' have 5 peaks in common with S .

Spectral Convolution: reveals the peaks at 0 and 5.

Spectral Comparison: Difficult Case

$S \ominus S'$	S'	S									
		0	-10	-20	-30	-40	-50	-60	-70	-80	-90
		10	0	-10	-20	-30	-40	-50	-60	-70	-80
		20	10	0	-10	-20	-30	-40	-50	-60	-70
		30	20	10	0	-10	-20	-30	-40	-50	-60
		40	30	20	10	0	-10	-20	-30	-40	-50
		45	35	25	15	5	-5	-15	-25	-35	-45
		55	45	35	25	15	5	-5	-15	-25	-35
		65	55	45	35	25	15	5	-5	-15	-25
		75	65	55	45	35	25	15	5	-5	-15
		85	75	65	55	45	35	25	15	5	-5
$S \ominus S''$	S''	S									
		0	-10	-20	-30	-40	-50	-60	-70	-80	-90
		5	-5	-15	-25	-35	-45	-55	-65	-75	-85
		20	10	0	-10	-20	-30	-40	-50	-60	-70
		25	15	5	-5	-15	-25	-35	-45	-55	-65
		40	30	20	10	0	-10	-20	-30	-40	-50
		45	35	25	15	5	-5	-15	-25	-35	-45
		60	50	40	30	20	10	0	-10	-20	-30
		65	55	45	35	25	15	5	-5	-15	-25
		80	70	60	50	40	30	20	10	0	-10
		85	75	65	55	45	35	25	15	5	-5

Limitations of the Spectrum Convolutions

Spectral convolution does not reveal that spectra S and S' are similar, while spectra S and S'' are not.

Clumps of shared peaks: the matching positions in S' come in clumps while the matching positions in S'' don't.

This important property was not captured by spectral convolution.

Shifts

$A = \{a_1 < \dots < a_n\}$: an ordered set of natural numbers.

A *shift* (i, Δ) is characterized by two parameters, the position (i) and the length (Δ) .

The shift (i, Δ) transforms

$$\{a_1, \dots, a_n\}$$

into

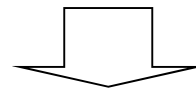
$$\{a_1, \dots, a_{i-1}, a_i + \Delta, \dots, a_n + \Delta\}$$

Shifts: An Example

The shift (i, Δ) transforms $\{a_1, \dots, a_n\}$
into $\{a_1, \dots, a_{i-1}, a_i + \Delta, \dots, a_n + \Delta\}$

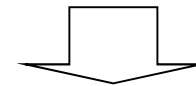
e.g.

10 20 30 40 50 60 70 80 90



shift (4, -5)

10 20 30 35 45 55 65 75 85



shift (7, -3)

10 20 30 35 45 55 62 72 82

Spectral Alignment Problem

- Find a series of k shifts that make the sets

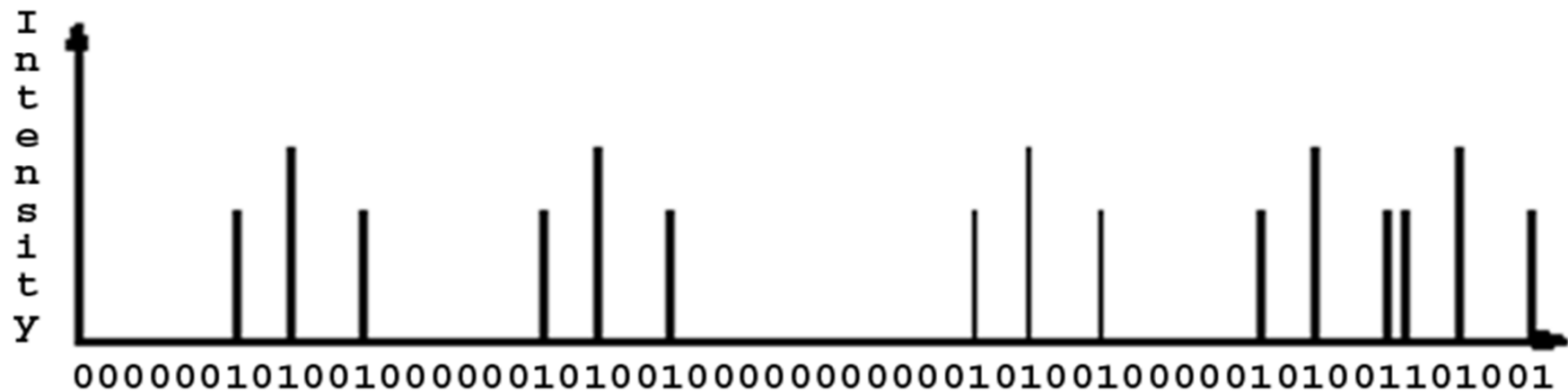
$$A = \{a_1, \dots, a_n\} \text{ and } B = \{b_1, \dots, b_n\}$$

as similar as possible.

- **k -similarity between sets**
 - $D(k)$ - the maximum number of elements in common between sets after k shifts.
-

Representing Spectra in 0-1 Alphabet

- Convert spectrum to a 0-1 string with 1s corresponding to the positions of the peaks.



Comparing Spectra=Comparing 0-1 Strings

- A modification with positive offset corresponds to inserting a block of 0s
 - A modification with negative offset corresponds to deleting a block of 0s
 - Comparison of theoretical and experimental spectra (represented as 0-1 strings) corresponds to a (somewhat unusual) **edit distance/alignment** problem where elementary edit operations are insertions/deletions of blocks of 0s
 - **Use sequence alignment algorithms!**
-

Spectral Alignment vs. Sequence Alignment

- Manhattan-like graph with different alphabet and scoring.
 - Movement can be diagonal (matching masses) or horizontal/vertical (insertions/deletions corresponding to PTMs).
 - At most k horizontal/vertical moves.
-

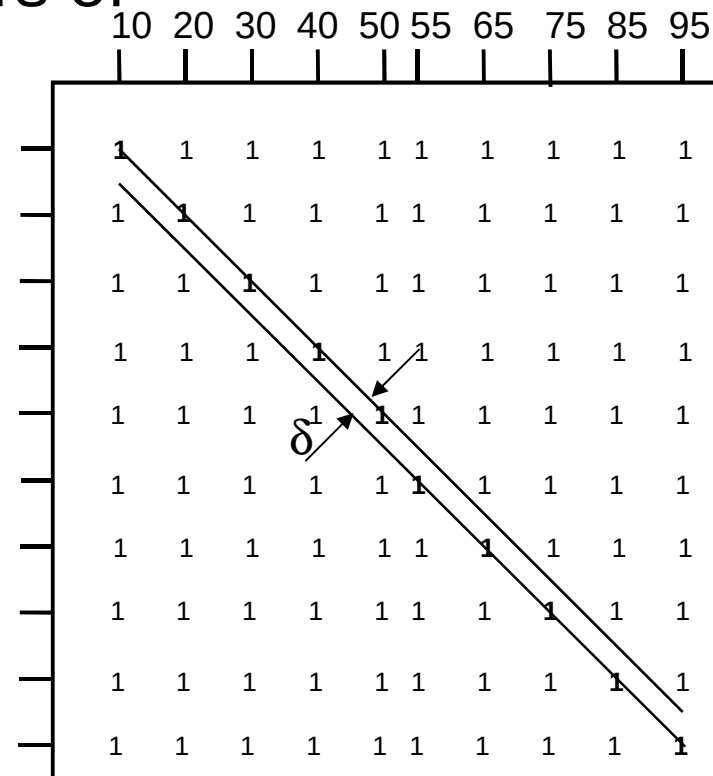
Spectral Product

$$A=\{a_1, \dots, a_n\} \text{ and } B=\{b_1, \dots, b_n\}$$

Spectral product $A \otimes B$: two-dimensional matrix with nm 1s corresponding to all pairs of indices (a_i, b_j) and remaining elements being 0s.

SPC: the number of 1s in the main diagonal.

δ -shifted SPC: the number of 1s on the diagonal $(i, i + \delta)$

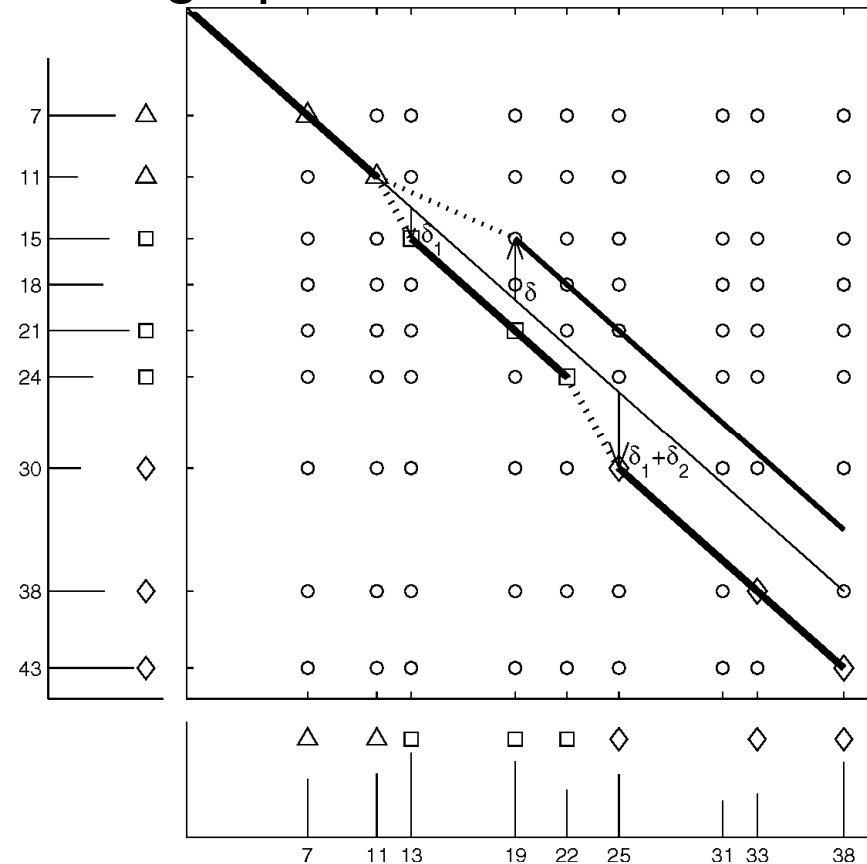


Spectral Alignment: ***k*-similarity**

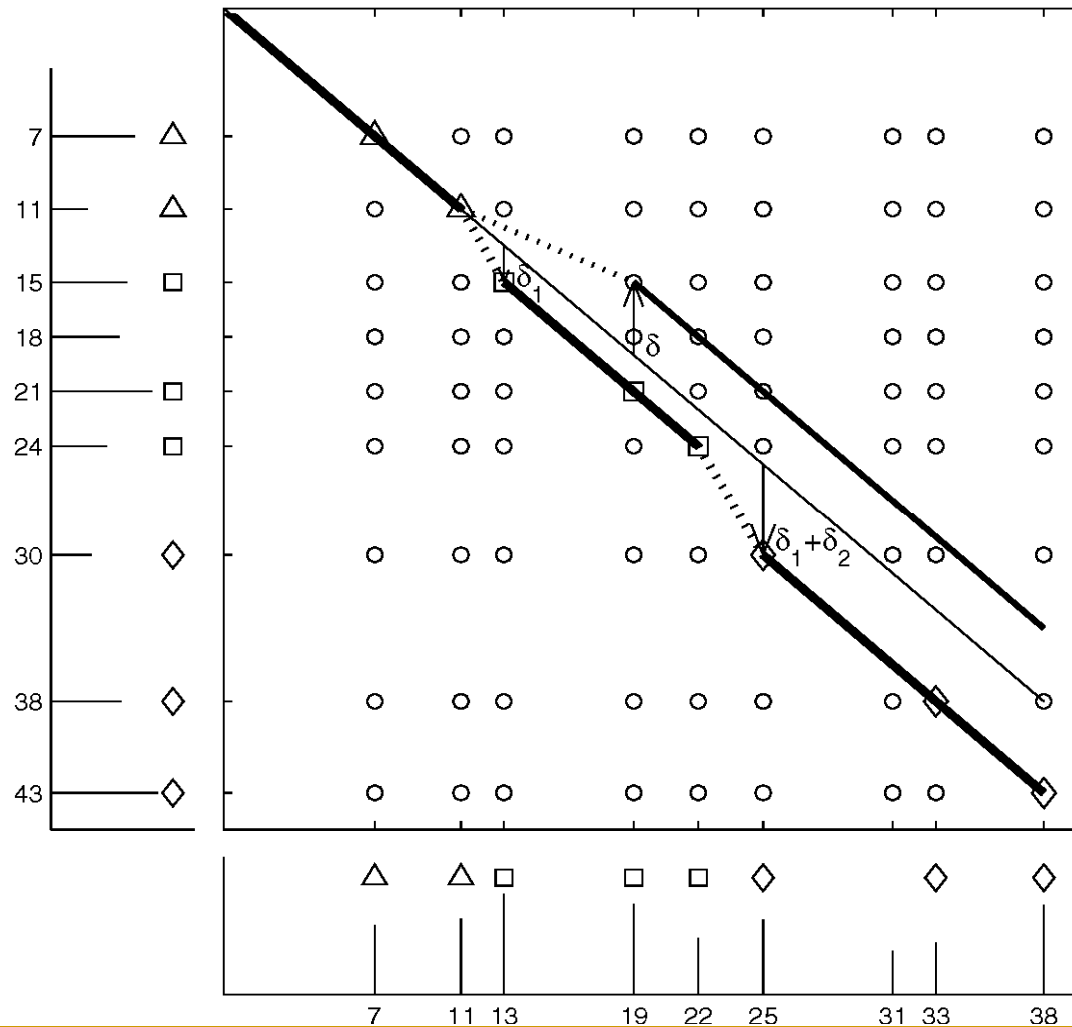
***k*-similarity between spectra:** the maximum number of 1s on a path through this graph that uses at most $k+1$ diagonals.

***k*-optimal spectral alignment = a path.**

The spectral alignment allows one to detect more and more subtle similarities between spectra by increasing k .

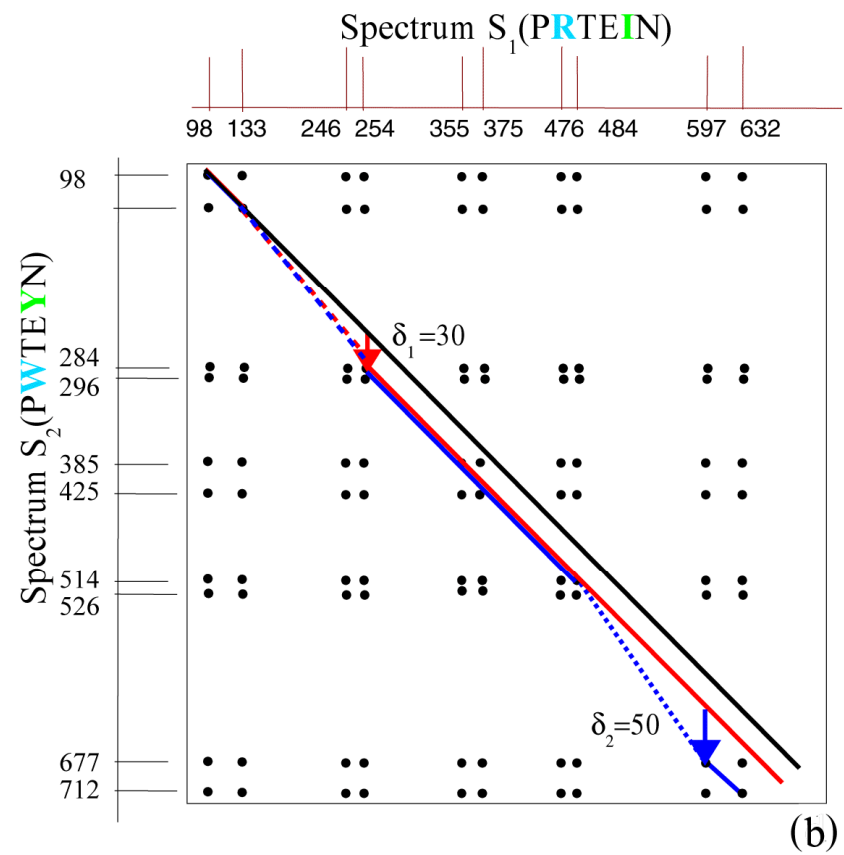
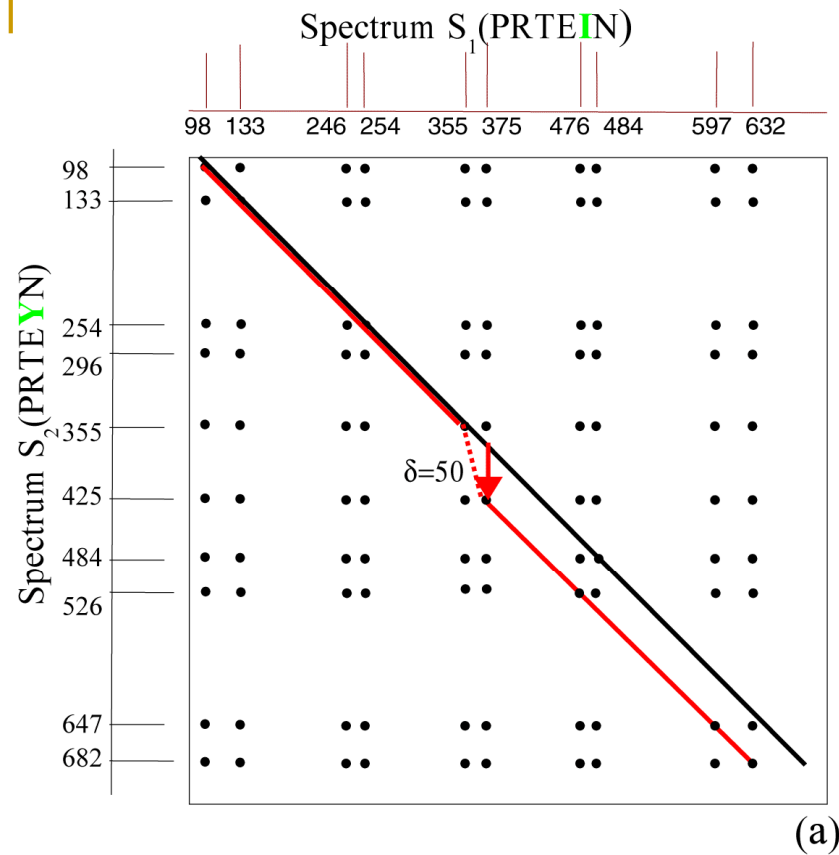


Use of k-Similarity



SPC reveals only $D(0)=3$ matching peaks.

Spectral Alignment reveals more hidden similarities between spectra: $D(1)=5$ and $D(2)=8$ and detects corresponding mutations.



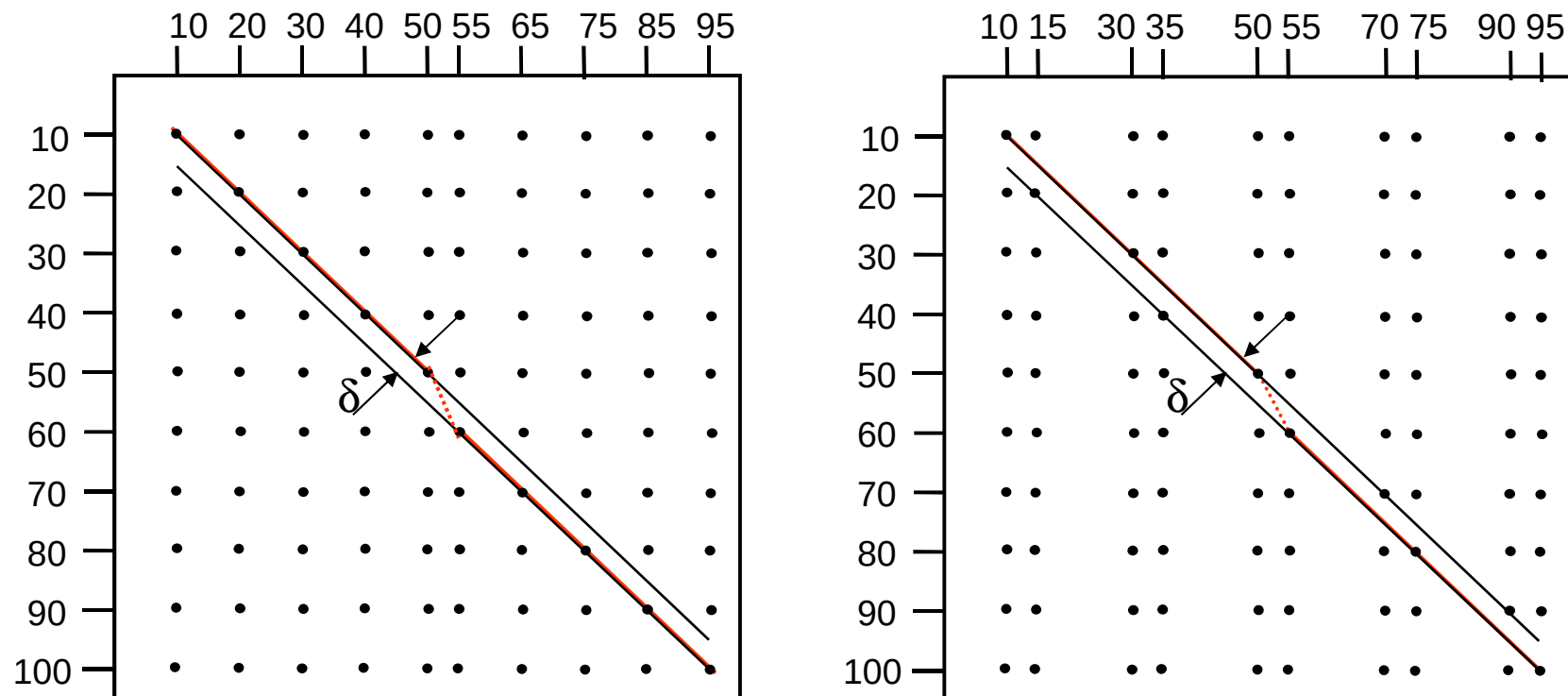
Black line represent the path for $k=0$

Red lines represent the path for $k=1$

Blue lines (right) represents the path for $k=2$

Spectral Convolution' Limitation

The spectral convolution considers diagonals separately without combining them into feasible mutation scenarios.



$D(1) = 10$

shift function score = 10

$D(1) = 6$

Dynamic Programming for Spectral Alignment

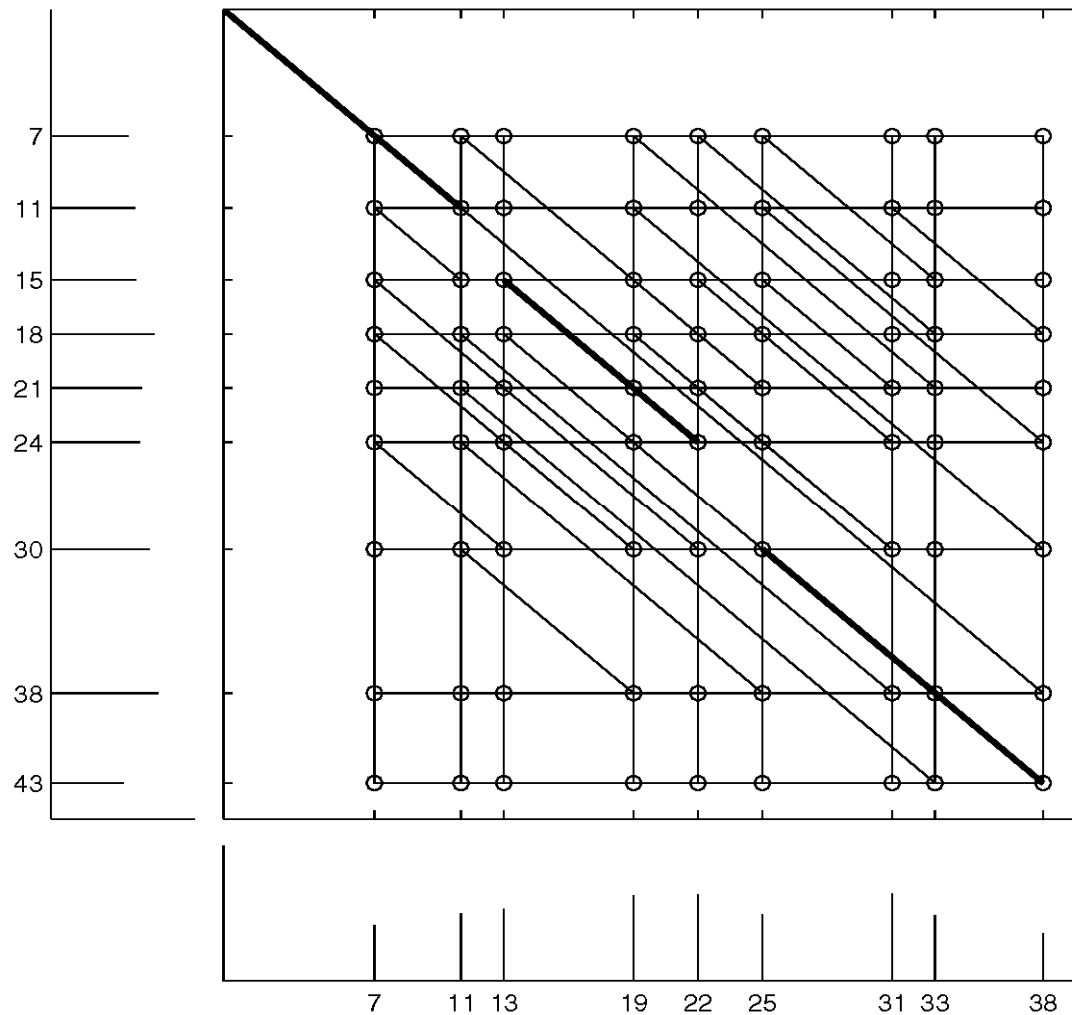
$D_{ij}(k)$: the maximum number of 1s on a path to (a_i, b_j) that uses at most $k+1$ diagonals.

$$D_{ij}(k) = \max_{(i', j') < (i, j)} \begin{cases} D_{i'j'}(k) + 1, & \text{if } (i - i' = j - j') \\ D_{i'j'}(k-1) + 1, & \text{otherwise} \end{cases}$$

$$D(k) = \max_{ij} D_{ij}(k)$$

Running time? $O(n^4 k)$

Edit Graph for Fast Spectral Alignment



$M(i,j)$ – the position of previous 1 on the same diagonal as (i,j)

Fast Spectral Alignment Algorithm

$$M_{ij}(k) = \max_{(i', j') < (i, j)} D_{i' j'}(k)$$

$$D_{ij}(k) = \max \begin{cases} D_{diag(i, j)}(k) + 1 \\ M_{i-1, j-1}(k-1) + 1 \end{cases}$$

$$M_{ij}(k) = \max \begin{cases} D_{ij}(k) \\ M_{i-1, j}(k) \\ M_{i, j-1}(k) \end{cases}$$

Running time: $O(n^2 k)$

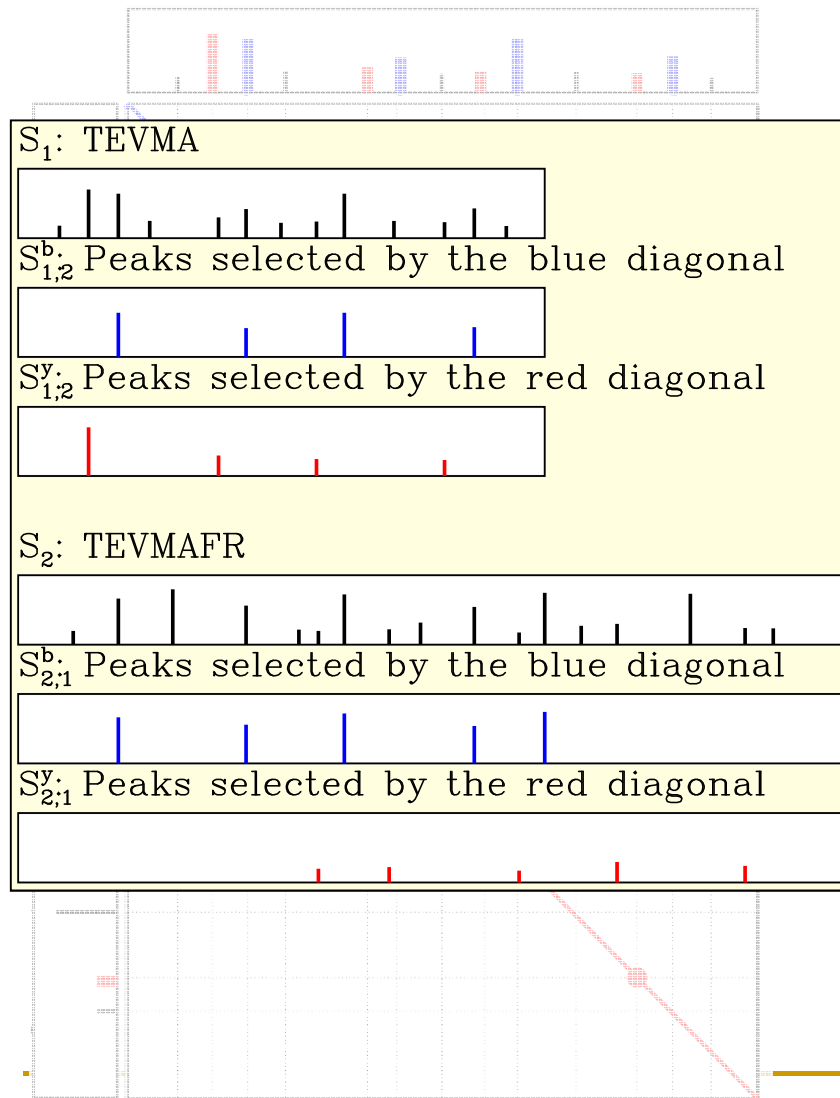
Spectral Alignment: Complications

Spectra are combinations of an increasing (N-terminal ions) and a decreasing (C-terminal ions) number series.

These series form two diagonals in the spectral product, the main diagonal and a complementary diagonal.

The described algorithm deals with the main diagonal only.

Spectral alignment



Peptide pair

TEVMA/TEVMAFR

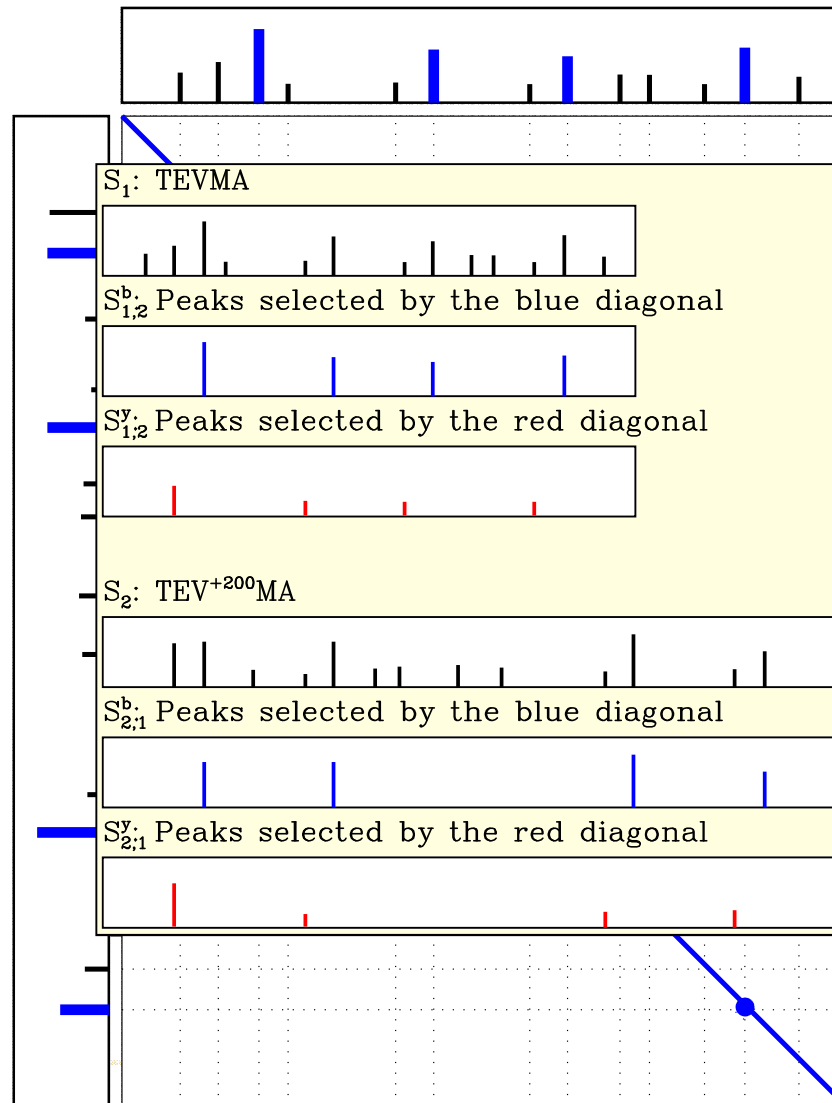
1. Find the matching points on the main diagonals:

- From top-left corner
(b: prefix masses)
- From right-bottom corner
(y: suffix masses)

Note that colors are not known.

2. Select matching masses from the aligned spectra.

Modified/Unmodified peptides



Peptide pair

TEVMA/TEV⁺²⁰⁰MA

Selecting matches on the diagonals *does not* work

Need to extend algorithm to allow modifications:

mass
insertions/deletions

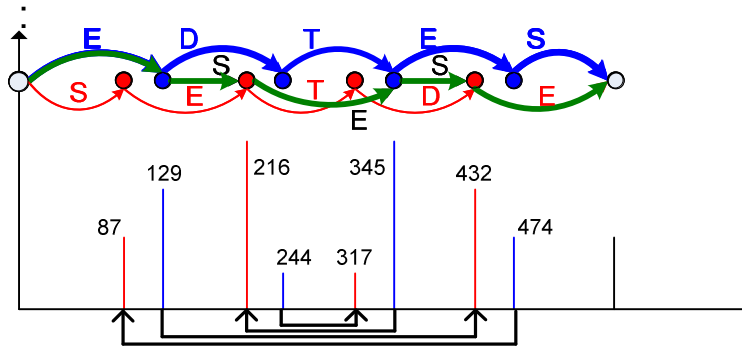
Algorithmically equivalent to computing edit distances between sequences

Computing the spectral alignment

De-novo problem

Input: One spectrum graph

Output: Longest path with no blue/red pairs



Solution:

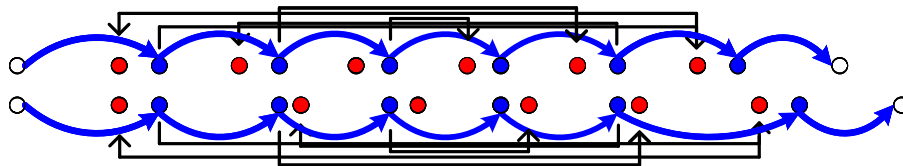
1. Jump from the blue mass closest to the start/end of the spectrum
2. Avoid reusing the pairing red mass

We saw that this ordering is always possible

Spectral alignment problem

Input: Two spectrum graphs

Output: Longest common path with no blue/red pairs in either spectrum and at most one unmatched edge



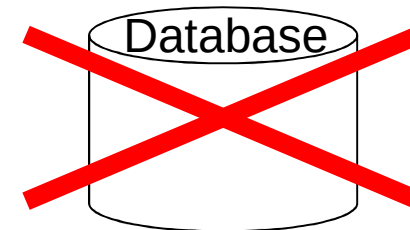
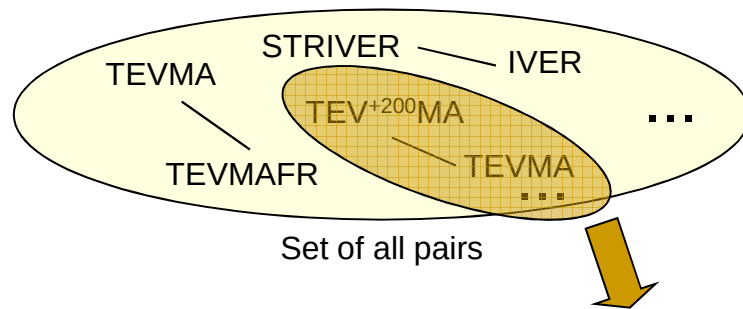
- ⇒ Ordering is not unique
- ⇒ Any choice generates multiple red masses

Alignment algorithm proceeds as above but:

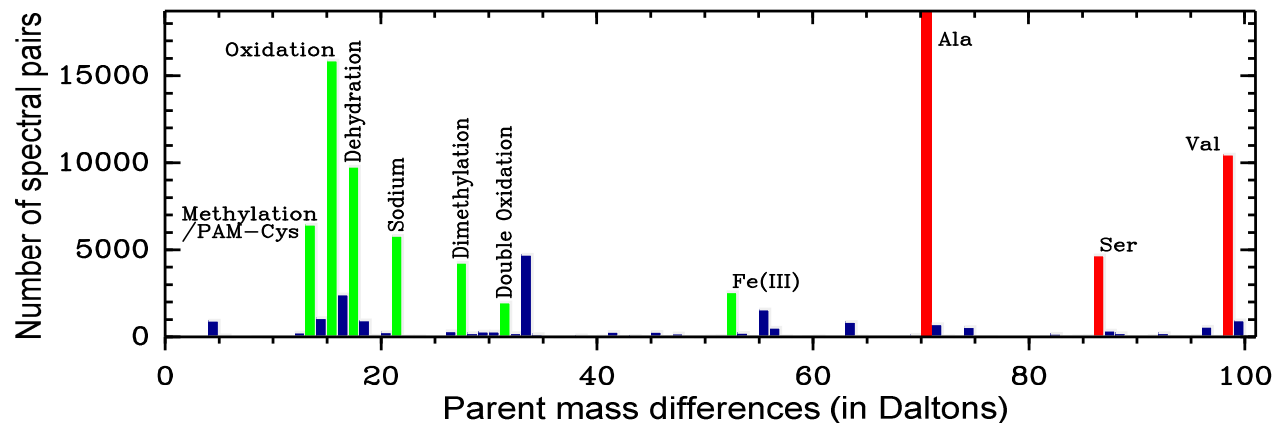
1. Imposes the order based on the smaller spectrum graph
2. Keeps a small log of all red masses
3. Worst-case exponential but works well in practice

Spectral pairs

Each spectral pair (S_1, S_2) selects a subset of masses from S_1 and another from S_2 :



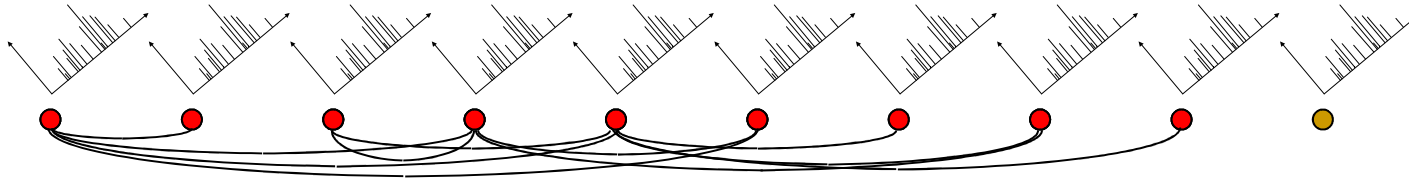
No database required



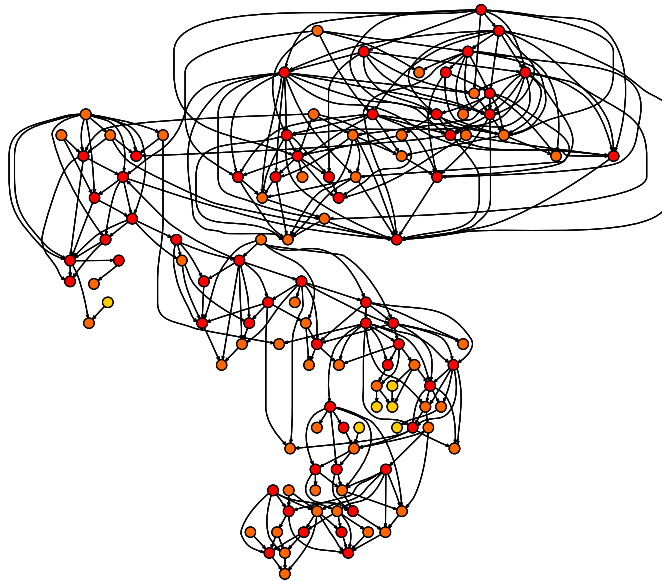
Rediscovered the most common modifications in the dataset

Combining spectral pairs

Set of
MS/MS
spectra



The set of detected spectral pairs defines a *Spectral Network*

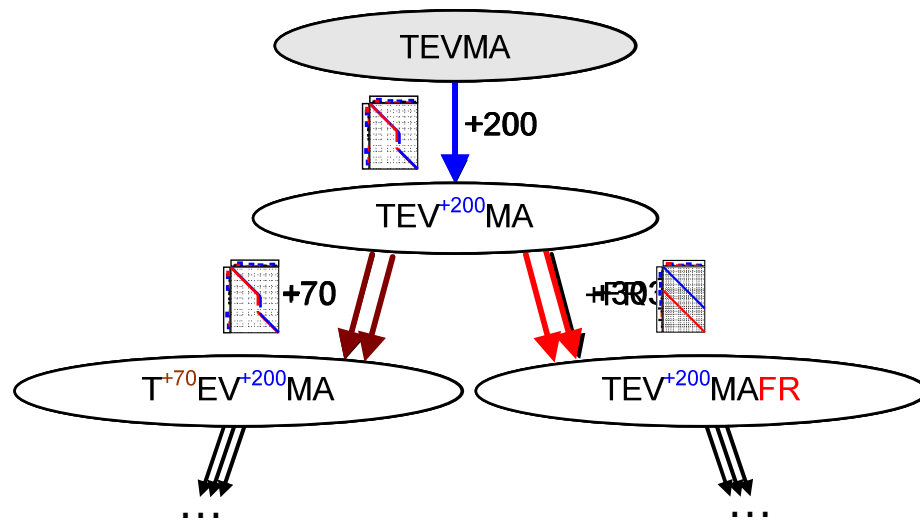


● Some spectra identified by
database search

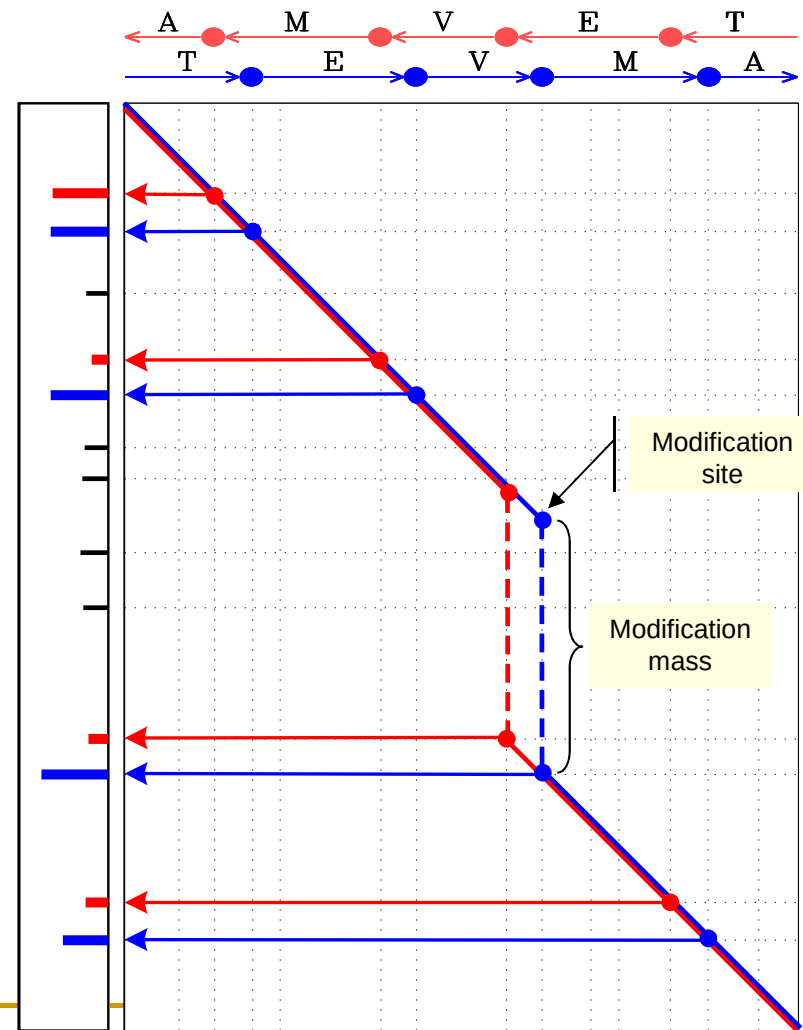
● ● Most modified spectra
identified by propagation

Propagation of peptide identifications

TEVMA identified by tag database search



Iterate until no more nodes can be annotated.



Propagation algorithm

Simple propagation algorithm:

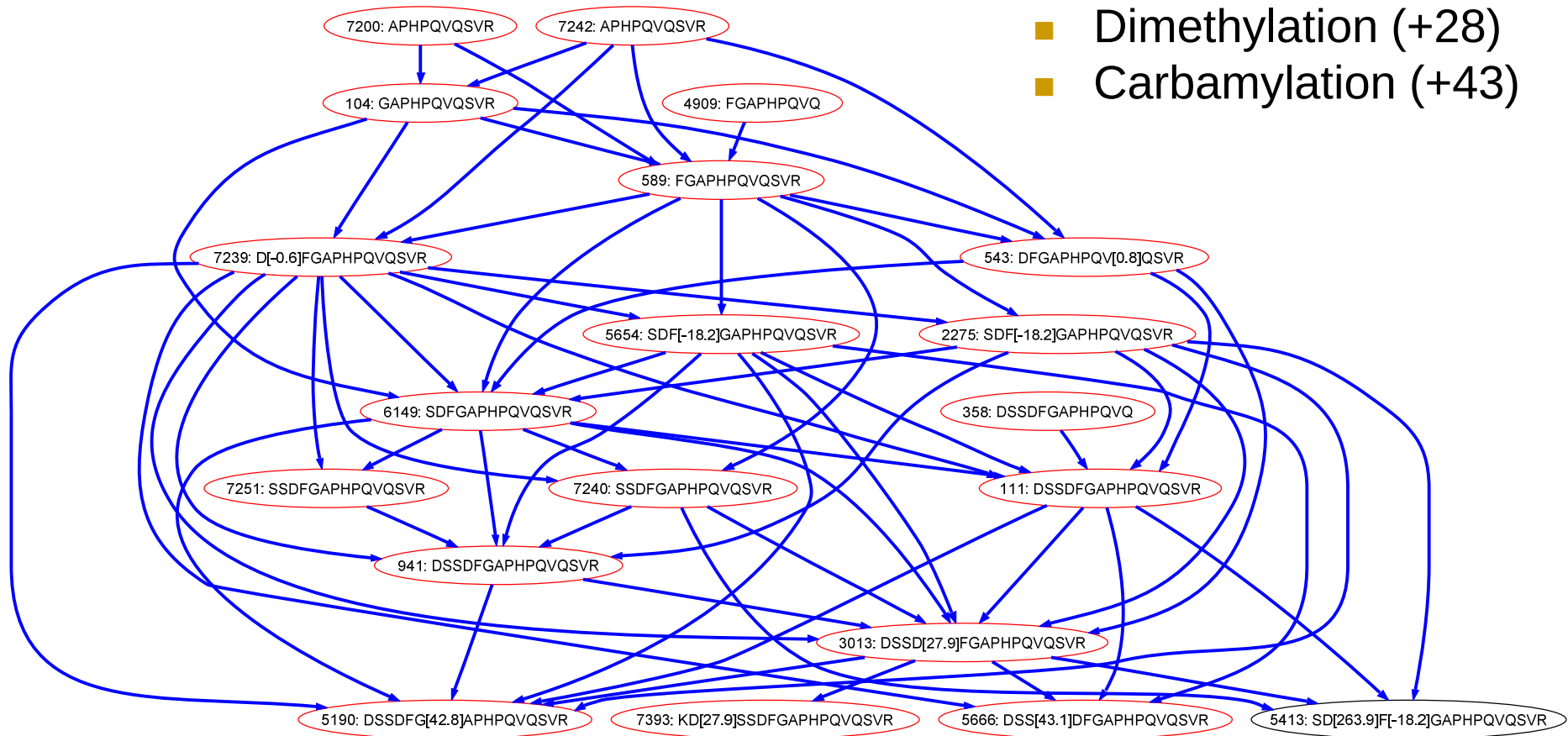
- i. Every annotated spectrum S propagates its annotation to every non-annotated neighbor S_{neigh}
- ii. Spectral alignment of (S, S_{neigh}) is used to determine the mass and location of the modification
- iii. S_{neigh} is marked as annotated
- iv. Iterate from i) until there are no more annotated spectra with non-annotated neighbors

Note that:

- Sometimes a spectrum S_{neigh} may receive different annotations from 2+ annotated neighbors
 - In these cases, S_{neigh} keeps the annotation that best explains the spectrum
-

Spectral networks output

- Dehydration (-18)
- Dimethylation (+28)
- Carbamylation (+43)



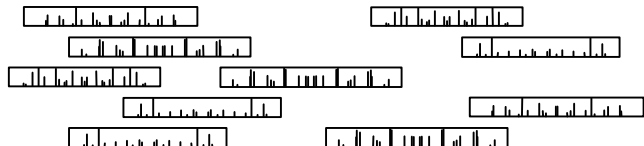
Assembling spectra into proteins

Genomes are sequenced from overlapping DNA reads.

CGTCCGCGTTTGGTCGTGGTGGCCGCACTTCGAACATGTTAC	<i>Consensus</i>
CGTC GTCGTGGTG AACATGT	
CGTTTGG GCACT TGTTAC	<i>Reads</i>
CGTCCG TTCGAACA	
GTTTGG GCCGCACTTCGA	
CCGCGTTT GTGGCCGCA CATGTTAC	

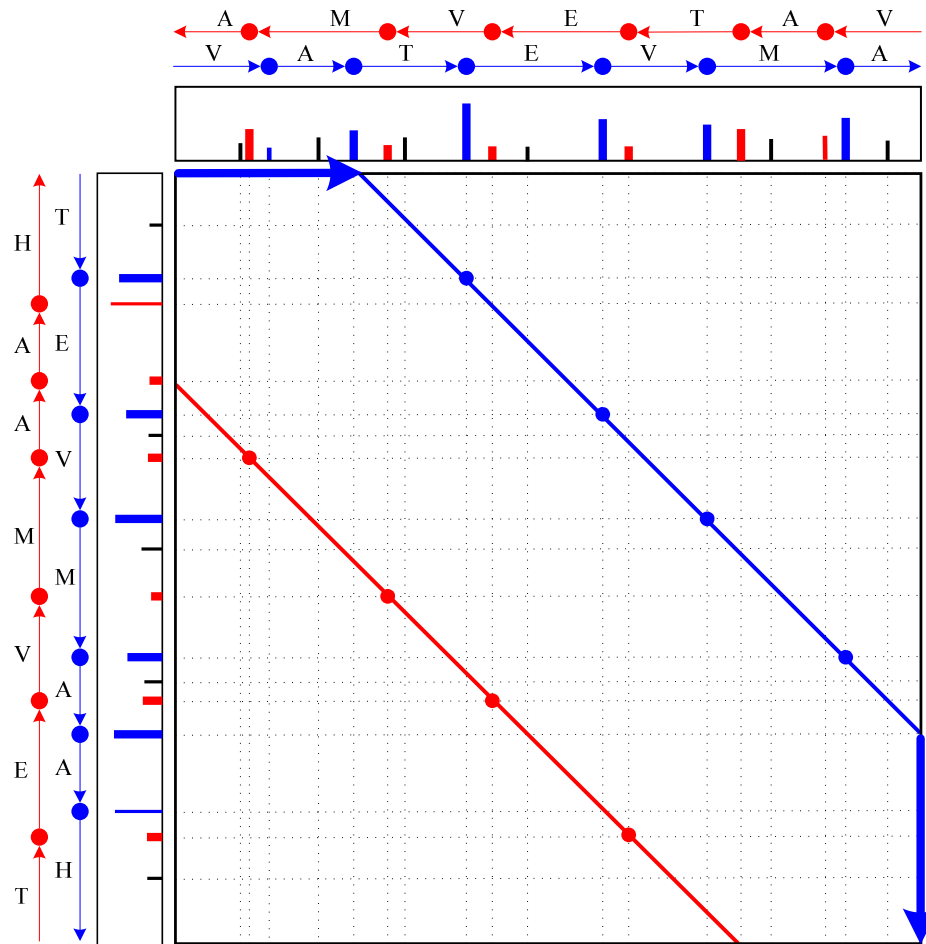
Shotgun Genome Sequencing

Now sequence *proteins* from *spectra* of overlapping peptides:

...VIQVLLVTICLAVFPYQGSSIILESGNIDDYEVVYPRKVTAL...	<i>Consensus</i>
	<i>MS/MS spectra</i>

Shotgun Protein Sequencing

Partial-overlap alignment



Peptide pair

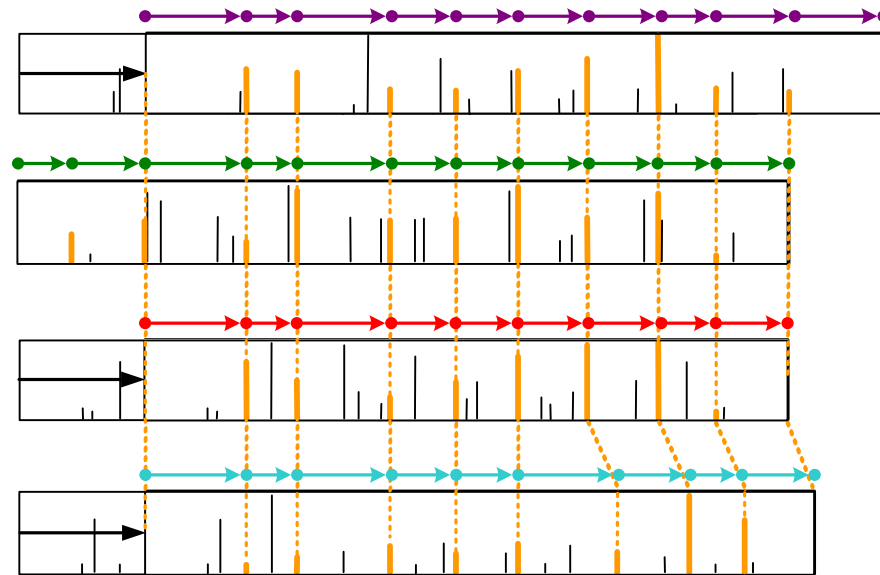
AVTEVMA/TEVMAAH

1. Very similar to prefix/suffix pairs *but*
2. Here we allow 2 jumps: one at the start and another at the end (→).

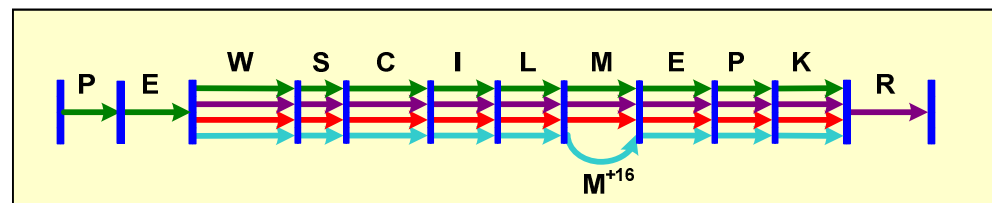
Shotgun Protein Sequencing

Assembling MS/MS spectra from overlapping peptides into protein sequences:

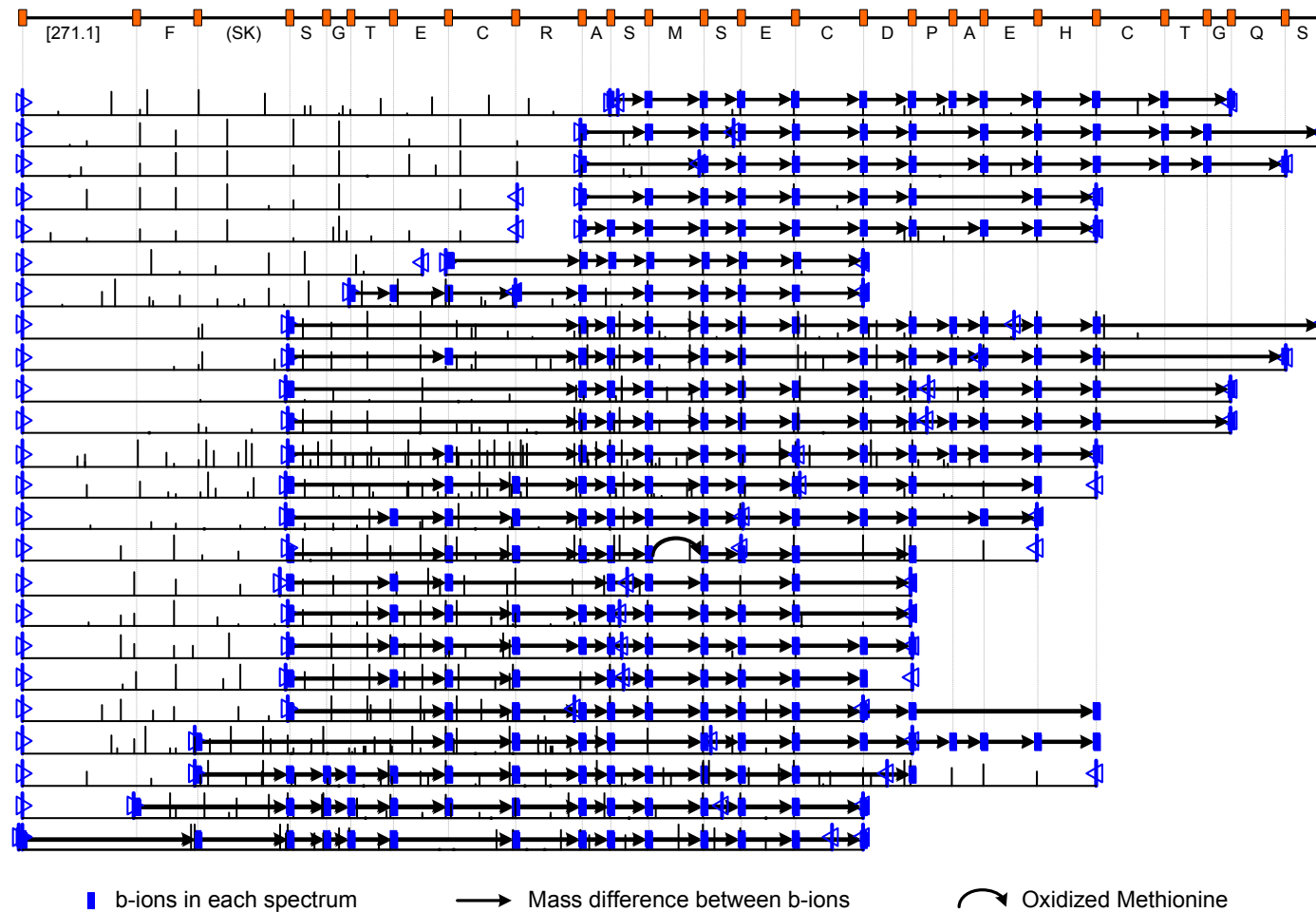
WSCILMEPKR
PEWSCILMEPK
WSCILMEPK
WSCILM⁺¹⁶EPK



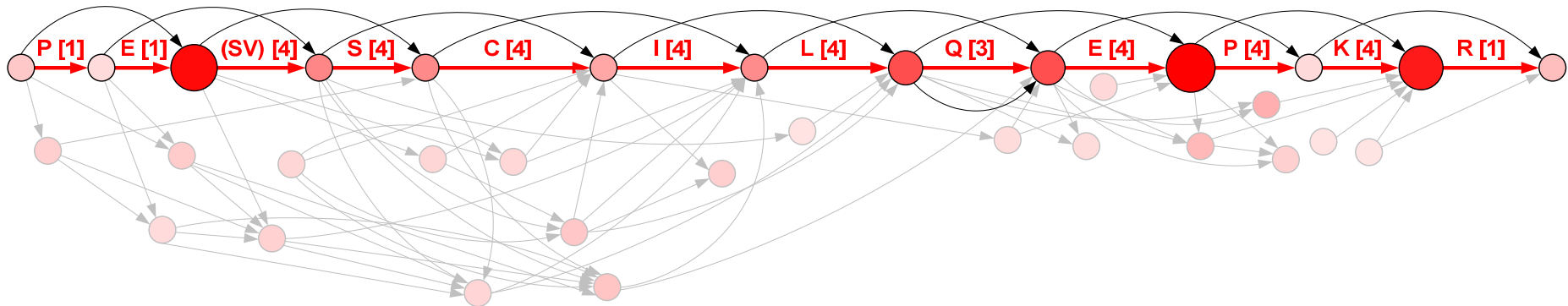
1. Find the spectral alignments
2. Select matching peaks
3. Collect mass differences between matched peaks
4. Determine the consensus sequence for *all* aligned spectra



28 aa protein contig, 24 spectra



Real graphs are more complicated



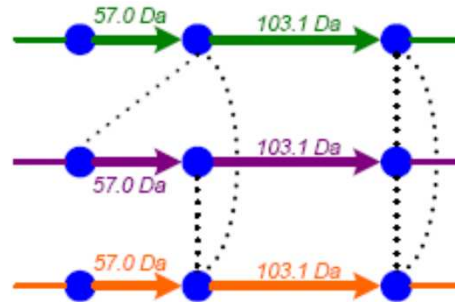
- Each spectrum is converted to a spectrum graph
- Vertices have scores proportional to peak intensities
- Must allow for missing peaks
- Ambiguities in amino acid masses,
e.g. $mass(G) + mass(A) \approx mass(Q) \approx mass(K)$

The score of a path is the summed score of all visited vertices
How to find a maximal-score path?

A-Bruijn difficulties

Difficulties caused by spectral alignment errors

- Incorrect glues



- Cycles make finding the heaviest path a harder problem

