Protein Identification: Algorithmic Challenges

Vineet Bafna, Ari Frank, Pavel Pevzner
Stephen Tanner, and Dekel Tsur

Protein Identification: Algorithmic Challenges

Pavel Pevzner

(joint work with Vineet Bafna, Ari Frank, Stephen Tanner, and Dekel Tsur)

UCSD

Three Algorithmic Problems

- **Searching for a million words in a text.** Suppose it takes 1 sec to find a word in a text. How much time would it take to find 1 million words in the text?

- **Searching for a word without even looking at 99.999% of the text.** Suppose you search for a word in a text. Would it be possible to ignore 99.999% of the text, scan only the remaining part and guarantee that the word you are looking for will be found?

- **Correcting Spelling Errors.** Given a book (in an unknown language) and a misspelled word, correct spelling errors in the word by finding a word in the book that looks “almost” like the misspelled word (with insertions/deletions/substitutions).

Problems Solved.

- **Searching for a million words in a text.**

  Aho-Corasik algorithm takes roughly the same time with million words as it takes with a single word.

- **Searching for a word without even looking at 99.999% of the text.**

  Filtration algorithms (like FASTA or BLAST) ignore 99.99999% of the text.

- **Correcting Spelling Errors.**

  Sequence alignment algorithms (like Smith-Waterman) do it in quadratic time

Three Unsolved Problems in Computational Mass-Spectrometry

- **Comparing a million spectra against a database.** Suppose it takes 1 sec to interpret a spectrum. How much time would it take to interpret 1 million spectra?
- **Mass-spectrometry database search without even looking at 99.999% of the database.** Suppose you compare a spectrum against a database. Would it be possible to ignore 99.999% of the database, scan only the remaining part and guarantee that you still can identify a peptide of interest?
- **Blind PTM search and discovery of new PTM types.** Given a spectrum of a peptide with *unknown* PTM types, find this peptide in the database. Discover new PTM types by data mining of large MS/MS datasets.

Three Solutions

- **Comparing a million spectra against a database.**

  InsPecT (Anal. Chem, 2005)

- **MS/MS database search without even looking at 99.999% of the database.**

  PepNovoTag+InsPecT (J. Proteome Res., 2005)

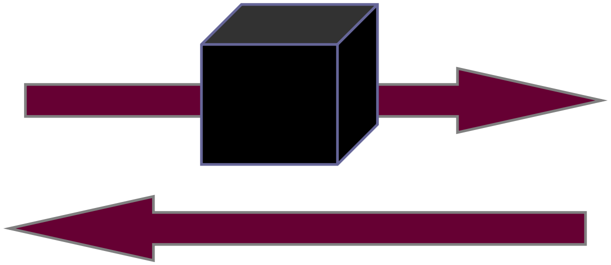
- **Blind PTM search and discovery of new PTM types.** Given a spectrum of a peptide with unknown PTM types, find this peptide in the database. Discover new PTM types by data mining of large MS/MS datasets.

  MS-Alignment (Nature Biotech., 2005)

Protein Identification by Mass Spectrometry

Sequence

MS/MS instrument

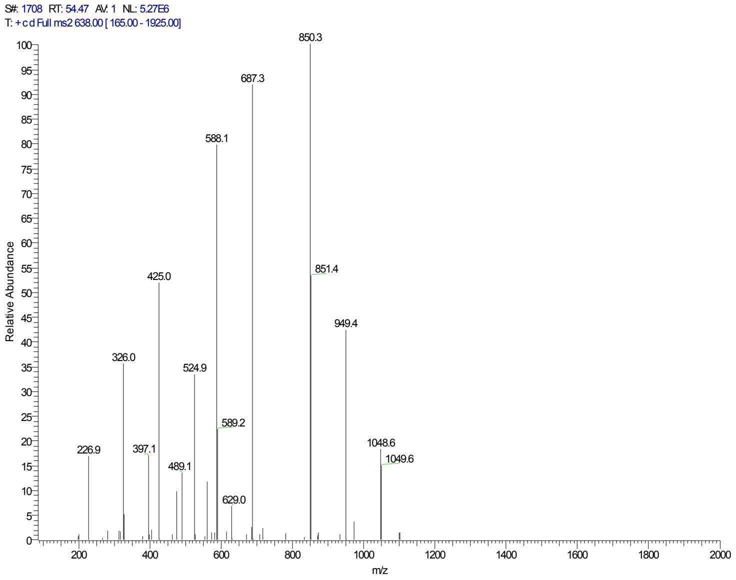


Database search

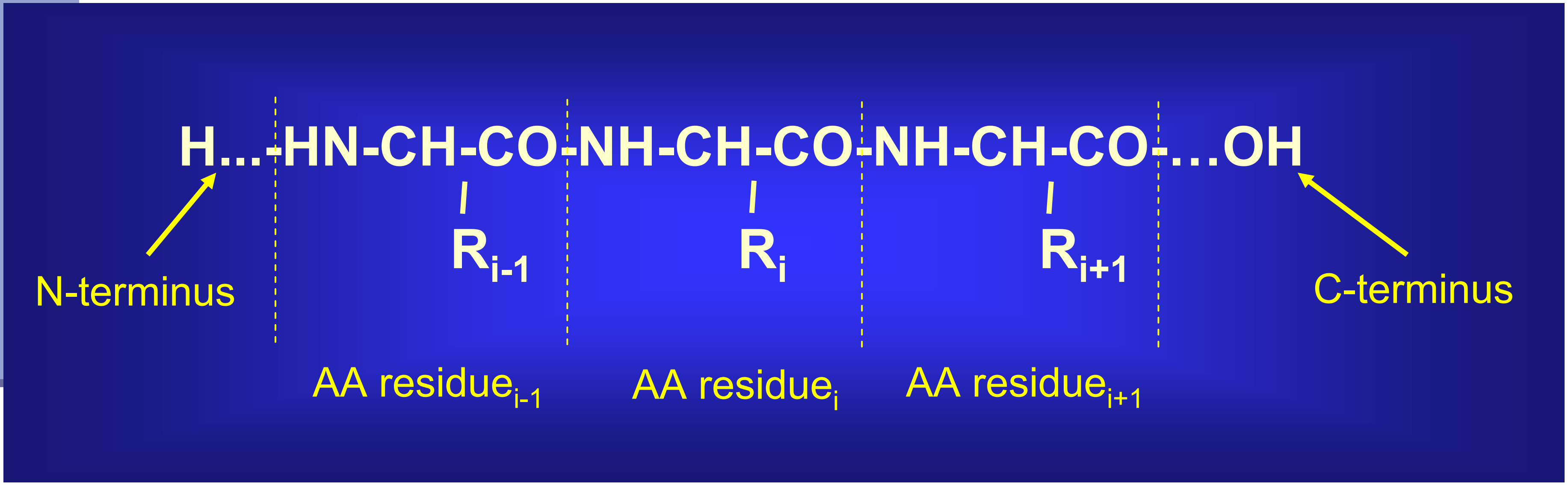
- Sequest, Mascot

de Novo interpretation

- Lutefisk, Peaks



Protein Backbone



H...-HN-CH-CO-NH-CH-CO-NH-CH-CO-…OH

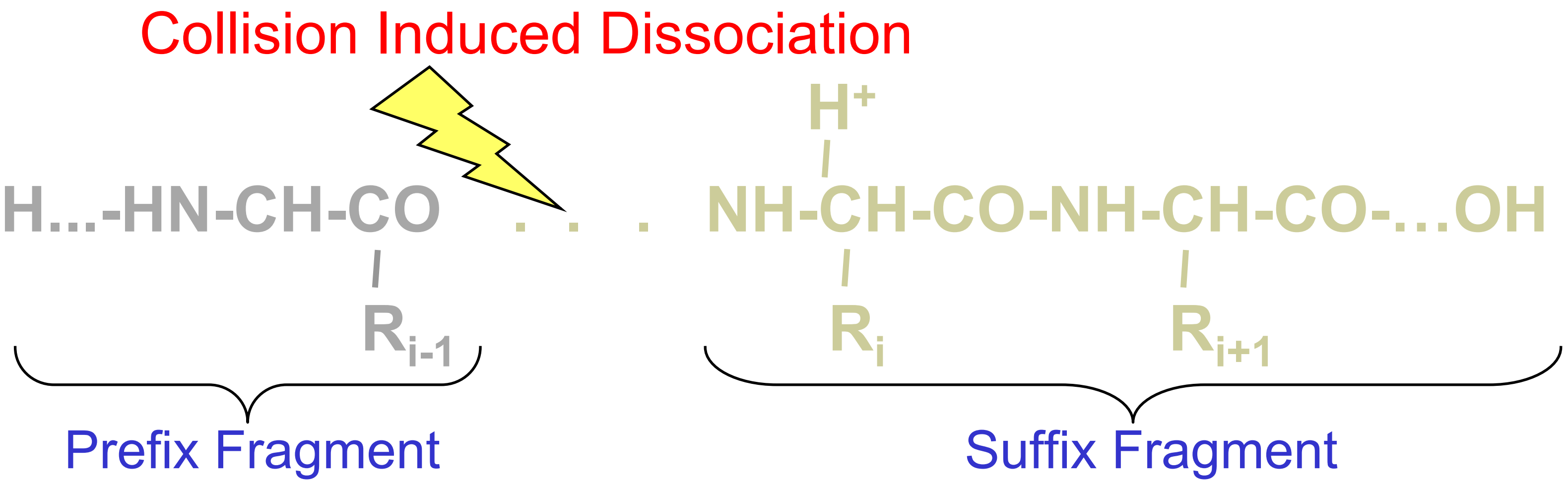
R_{i-1} R_i R_{i+1}

N-terminus

C-terminus

AA residue_{i-1} AA residue_i AA residue_{i+1}

Peptide Fragmentation



Collision Induced Dissociation

H^+

H...-HN-CH-CO . . . NH-CH-CO-NH-CH-CO-…OH

R_{i-1} R_i R_{i+1}

Prefix Fragment

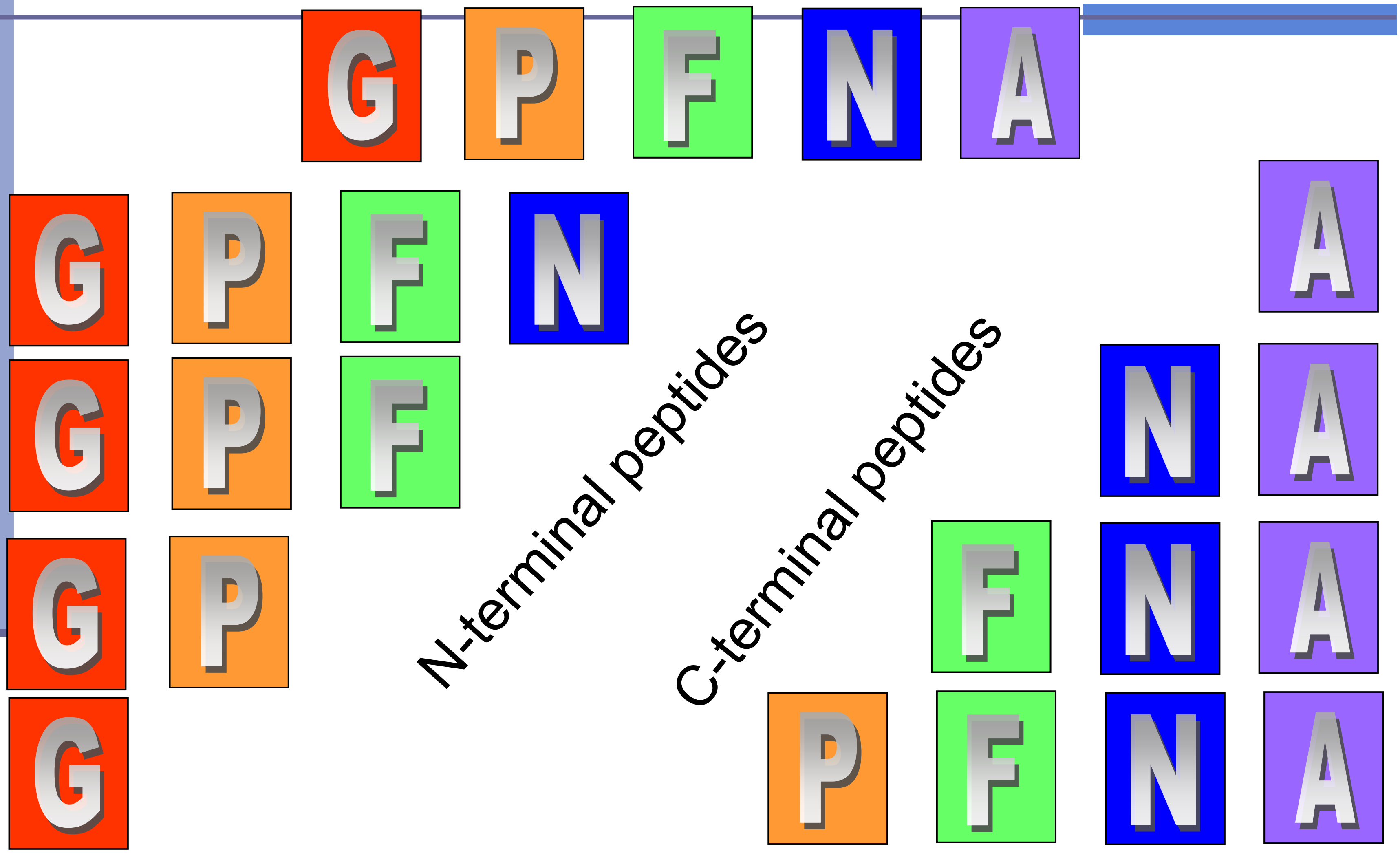
Suffix Fragment

- Peptides tend to fragment along the backbone.
- Fragments can also loose neutral chemical groups like NH_3 and H_2O.

Breaking Protein into Peptides and Peptides into Fragment Ions

- Proteases, e.g. trypsin, break protein into *peptides*.
- A 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 measure *mass/charge* ratio of an ion.

N- and C-terminal Peptides



N-terminal peptides

C-terminal peptides

Terminal peptides and ion types

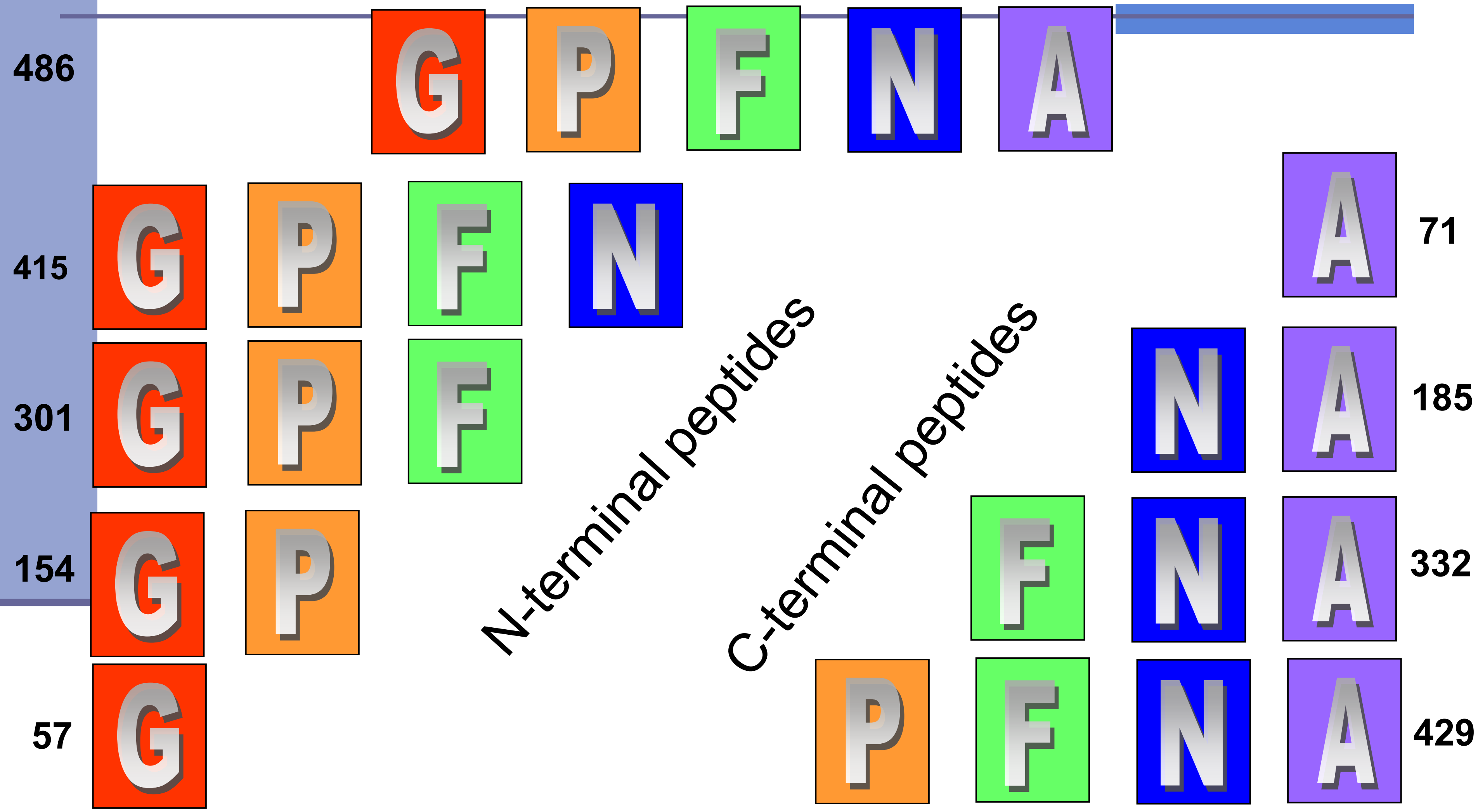
Peptide G P F N H20

Mass (D) 57 + 97 + 147 + 114 = 415

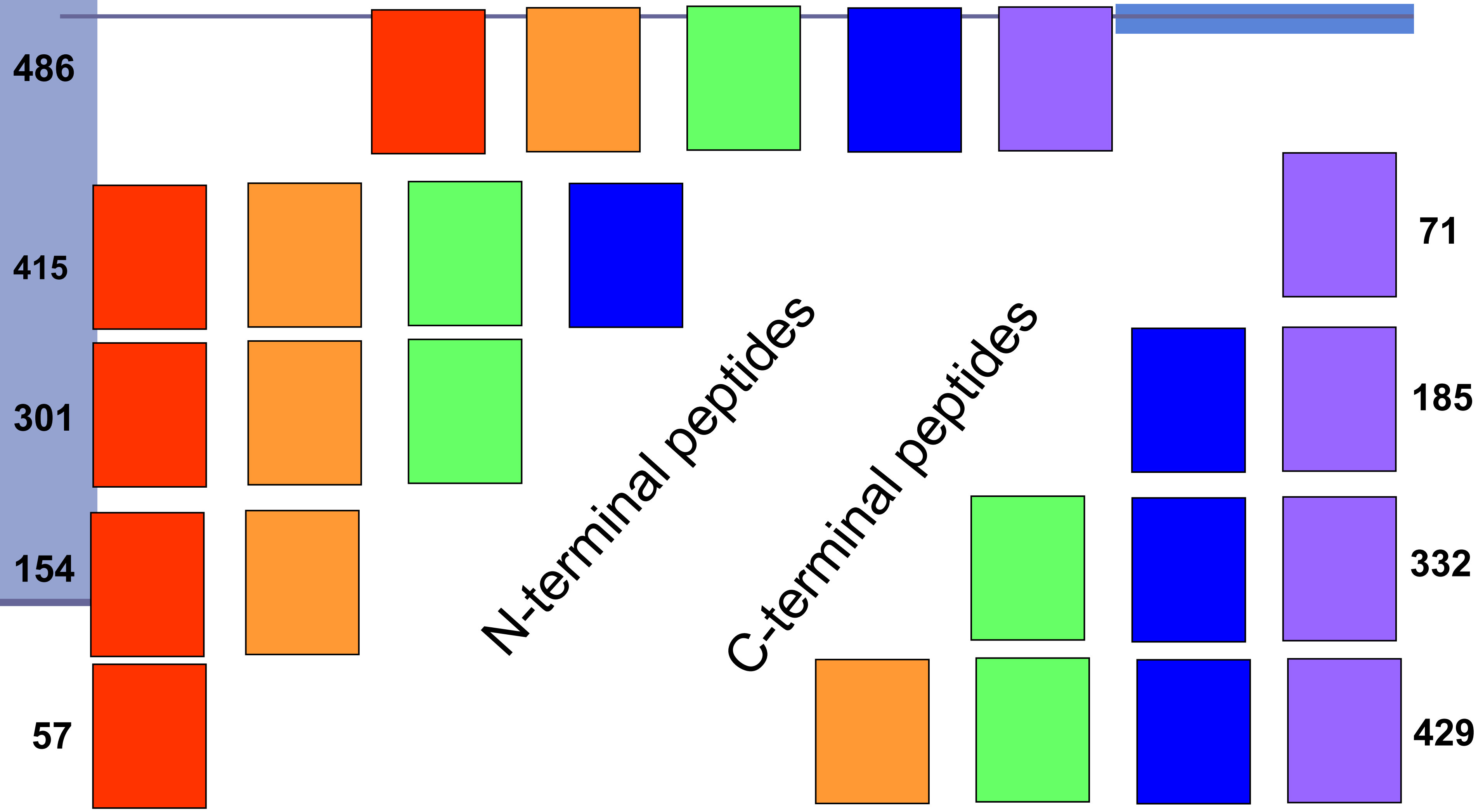
Peptide G P F N without H20

Mass (D) 57 + 97 + 147 + 114 – 18 = 397

N- and C-terminal Peptides



N- and C-terminal Peptides



N- and C-terminal Peptides

486

415

301

154

57

71

185

332

429

N- and C-terminal Peptides

486

71

415

Reconstruct peptide from the set of masses of fragment ions

(mass-spectrum)

185

301

154

332

57

429

N- and C-terminal Peptides

486

71

415

Reconstruct peptide from the set of masses of fragment ions

(mass-spectrum)

301

185

57 71 154 185 301 332 415 429 486

154

332

57

429

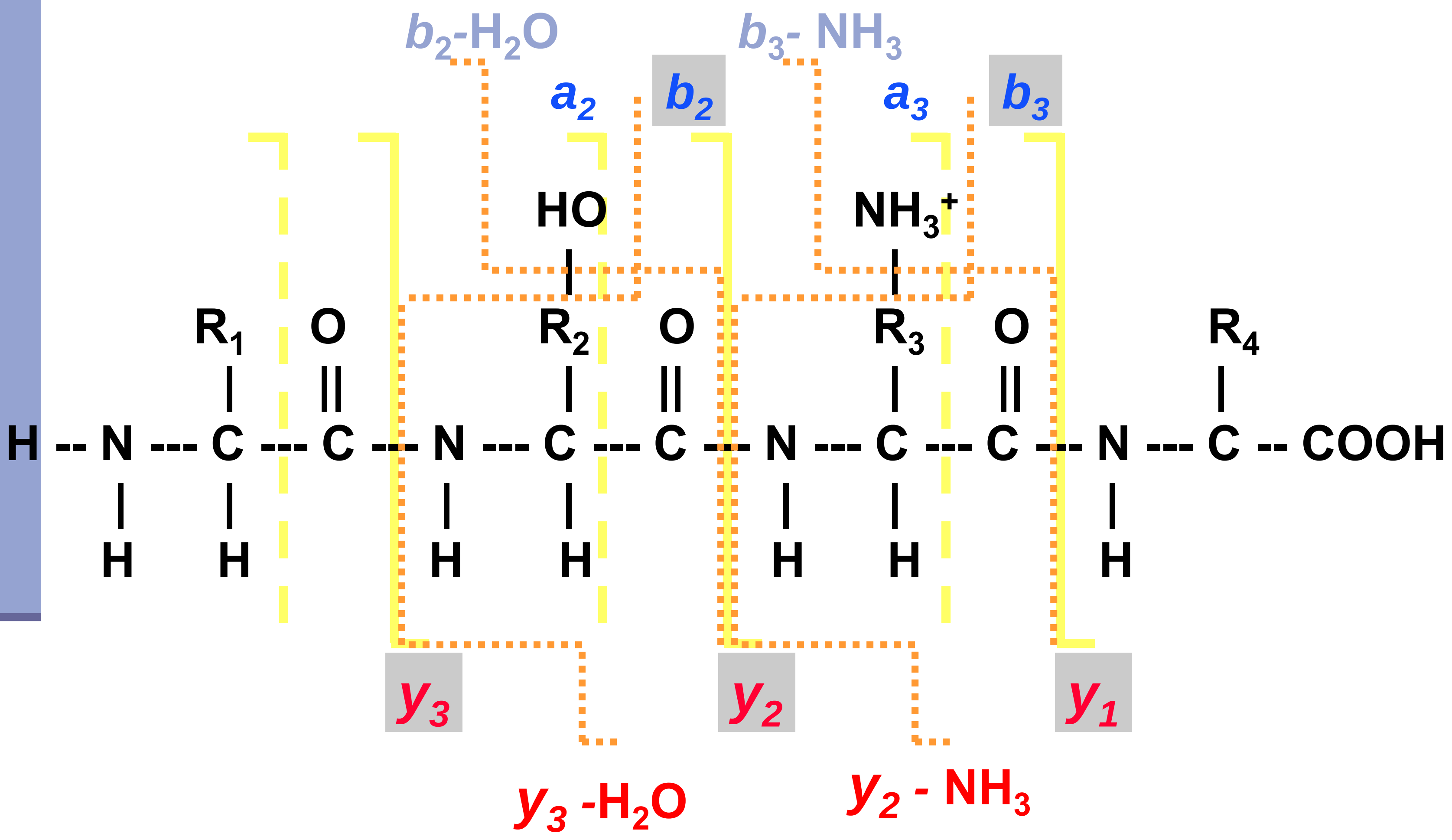
N- and C-terminal Peptides

Reconstruct peptide from the set of masses of fragment ions

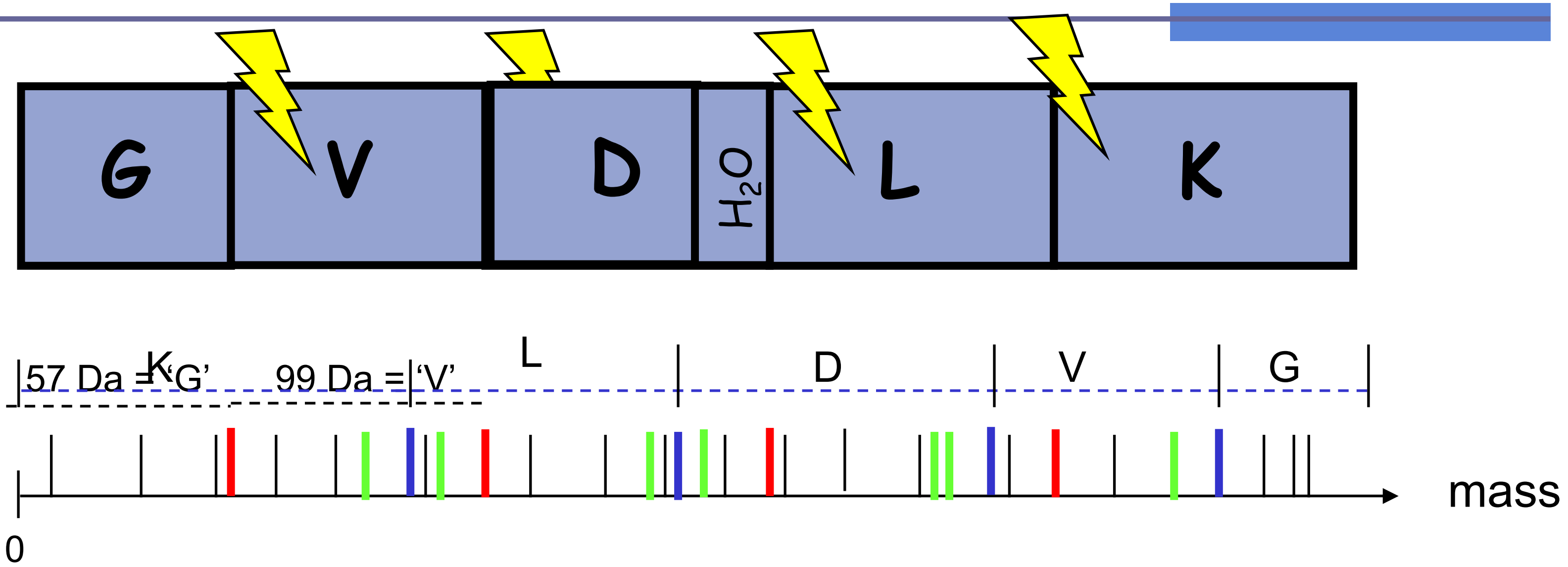
(mass-spectrum)

57 71 154 185 301 332 415 429 486

Peptide Fragmentation

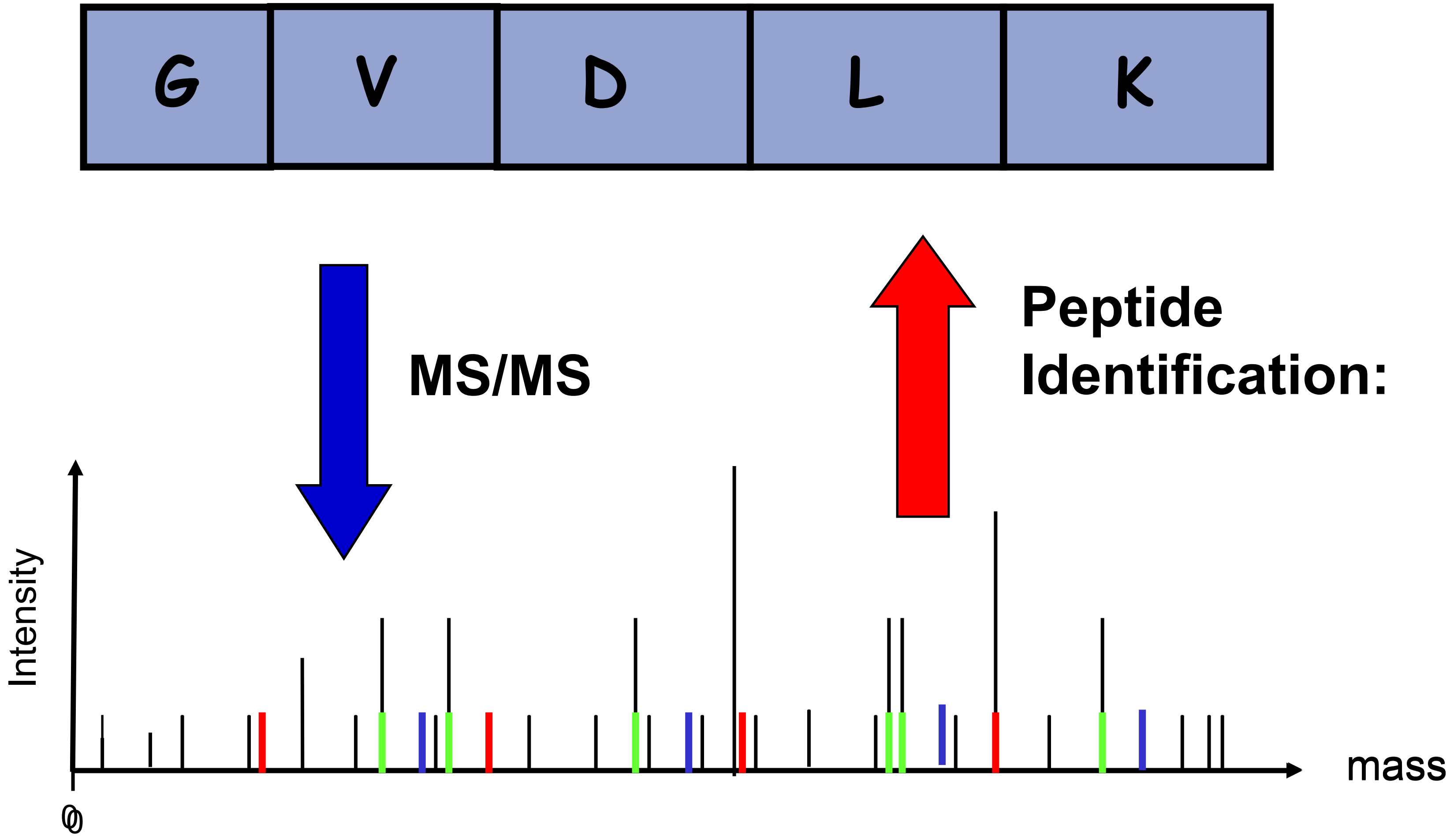


Mass Spectra

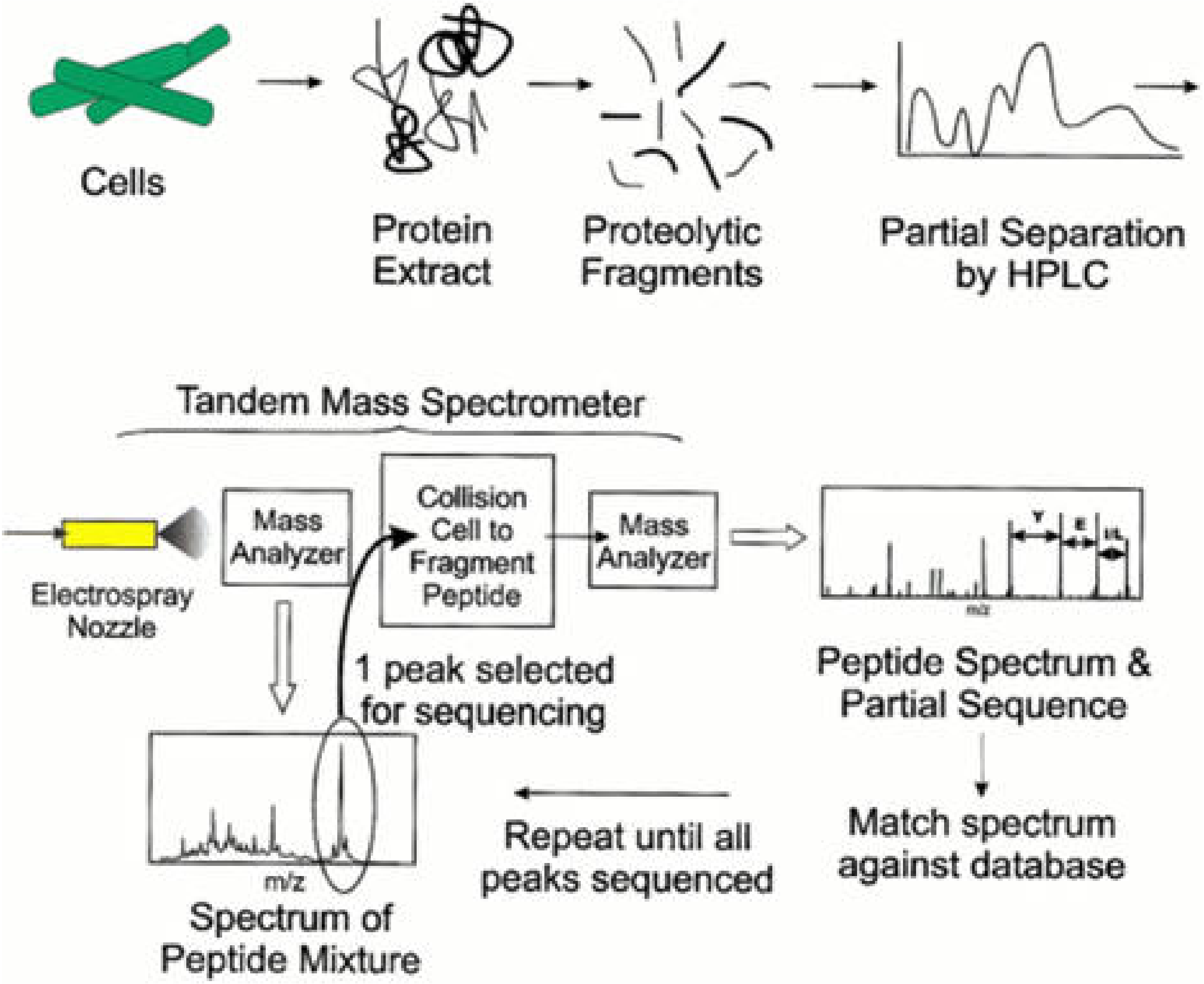


- The peaks in the mass spectrum:
  - Prefix and Suffix Fragments.
  - Fragments with **neutral losses** ($-H_2O$, $-NH_3$)
  - Noise and missing peaks.

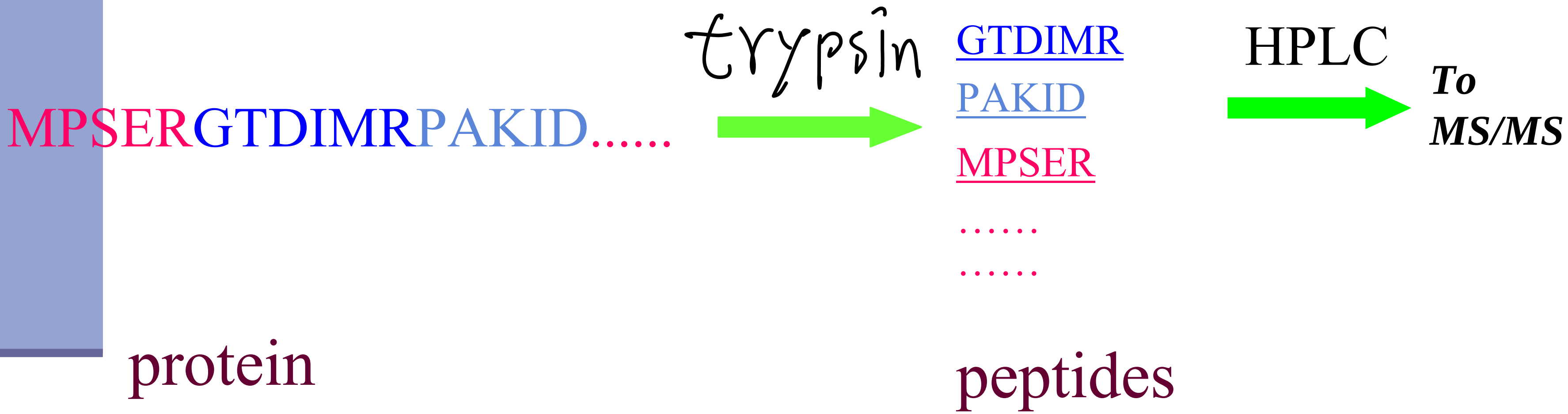
Protein Identification with MS/MS



Tandem Mass-Spectrometry


Cells
Protein Extract
Proteolytic Fragments
Partial Separation by HPLC
Tandem Mass Spectrometer
Electrospray Nozzle
Mass Analyzer
Collision Cell to Fragment Peptide
Mass Analyzer
m/z
Peptide Spectrum & Partial Sequence
1 peak selected for sequencing
m/z
Spectrum of Peptide Mixture
Repeat until all peaks sequenced
Match spectrum against database

Breaking Proteins into Peptides



MPSERGTDIMRPAKID……

trypsin

GTDIMR

PAKID

MPSER

……

……

HPLC

To MS/MS

protein

peptides

Mass Spectrometry

Matrix-Assisted Laser Desorption/Ionization (MALDI)

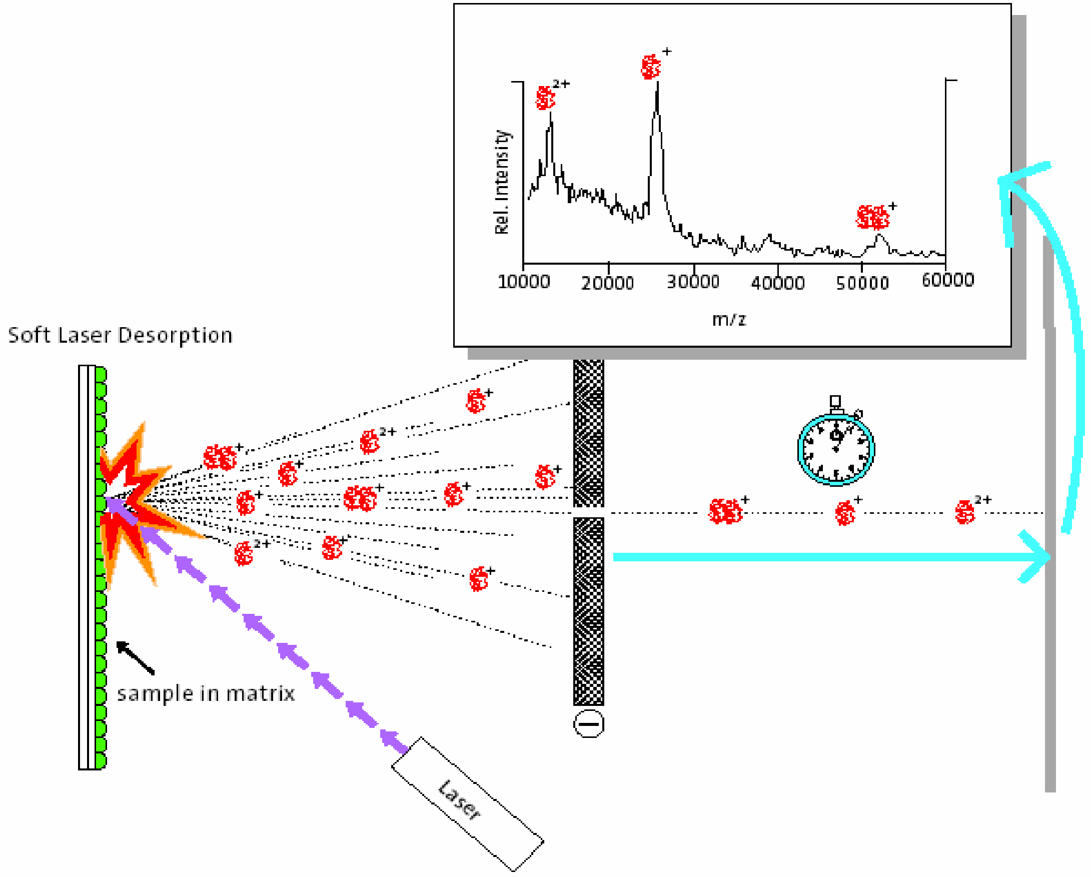
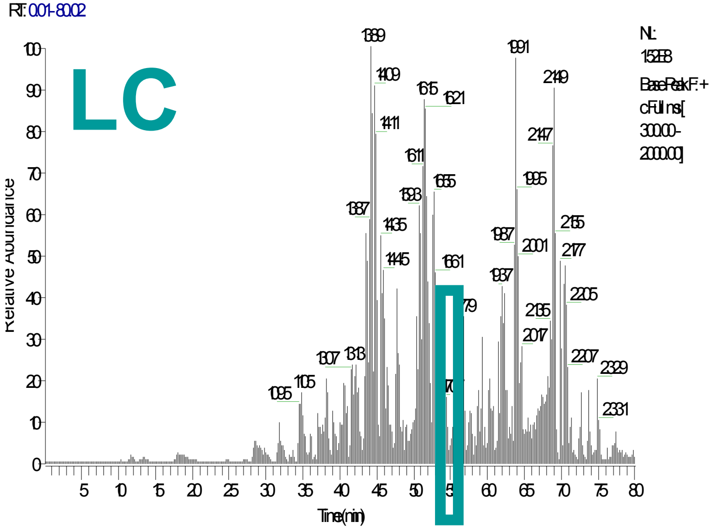
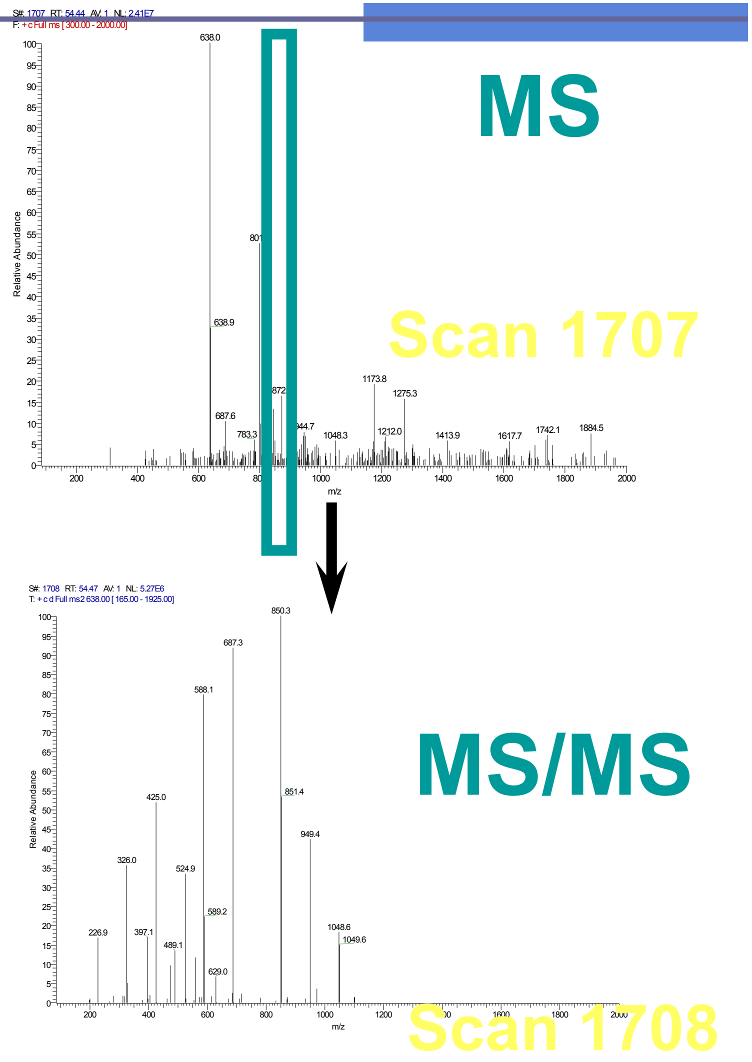


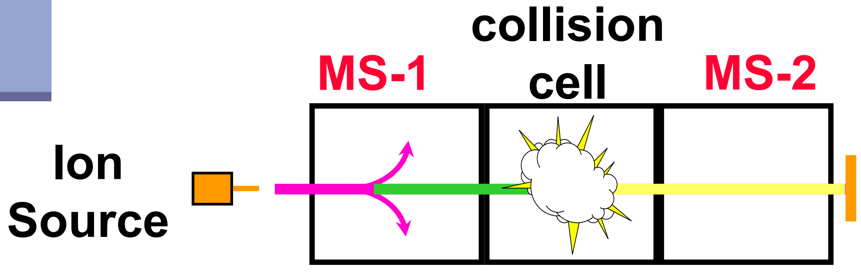
Figure 2. The soft laser desorption process.

From lectures by Vineet Bafna (UCSD)

Tandem Mass Spectrometry



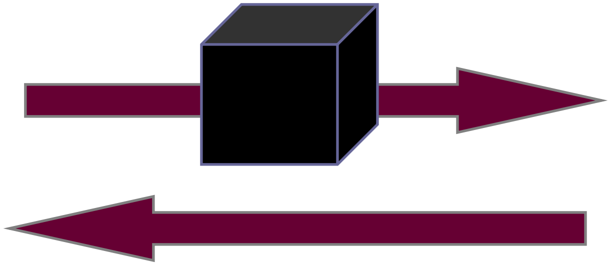


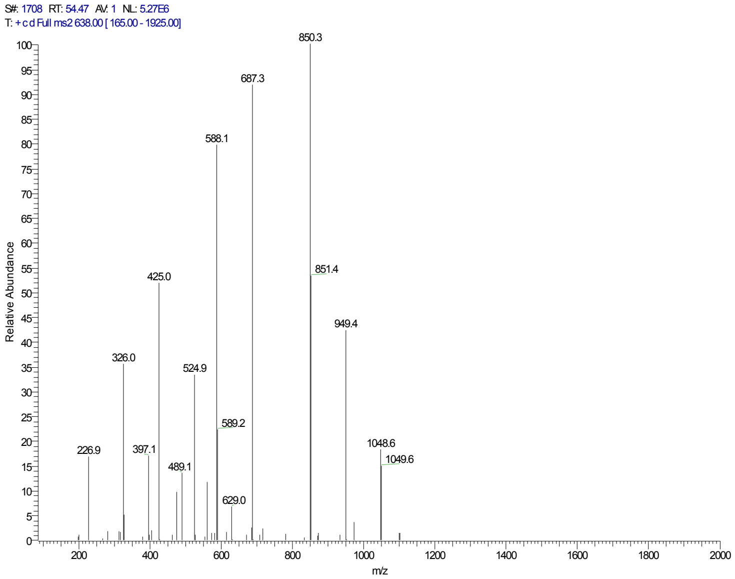


Protein Identification by Mass Spectrometry

Sequence

MS/MS instrument



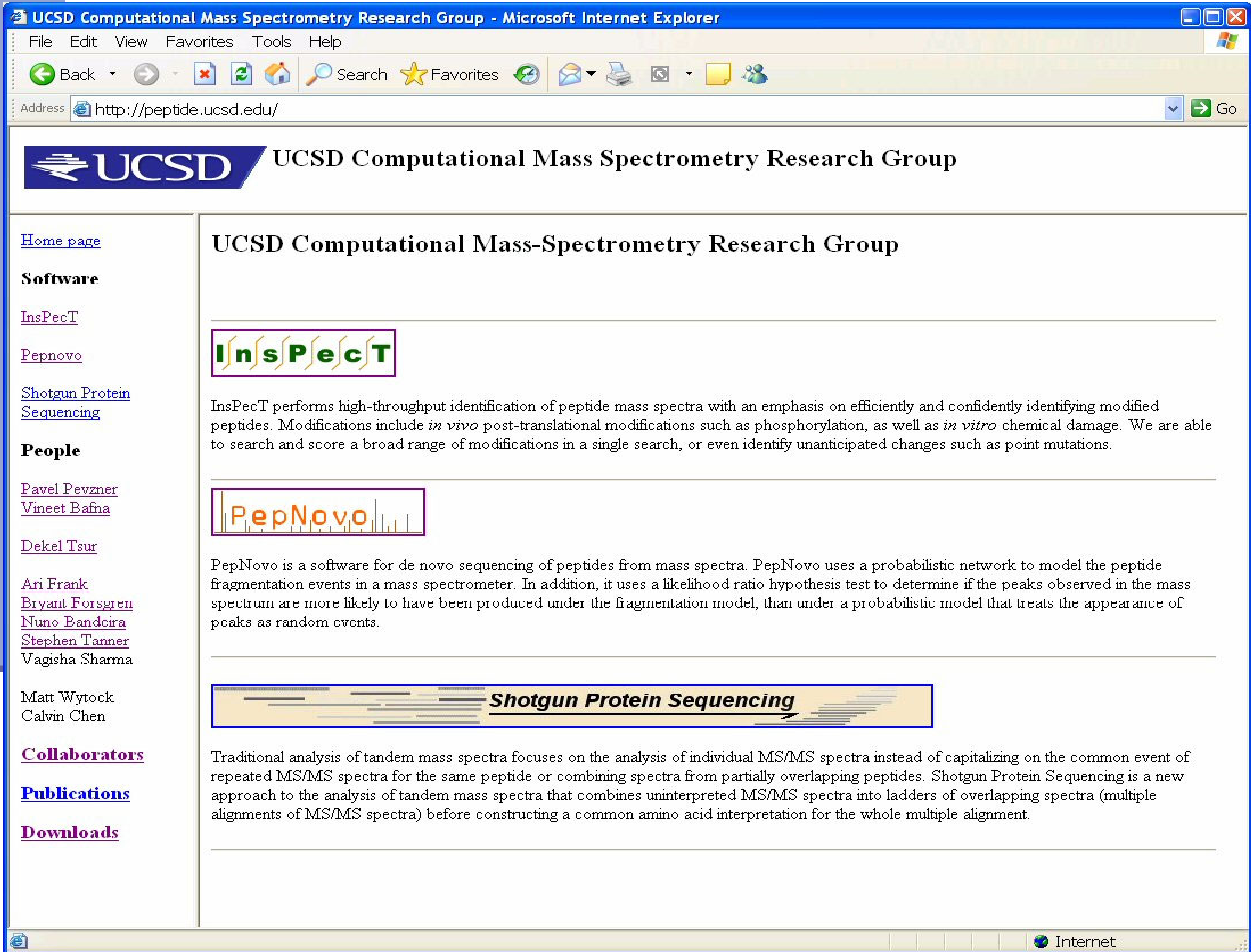


Database search

- Sequest, Mascot, ***InsPecT***

de Novo interpretation

- Lutefisk, Peaks, ***PepNovo***
- PTM Analysis and discovery
- ***MS-Alignment***



UCSD Computational Mass Spectrometry Research Group - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://peptide.ucsd.edu/
Go
UCSD
Internet

UCSD Computational Mass Spectrometry Research Group

Home page

Software

InsPecT

Pepnovo

Shotgun Protein Sequencing

People

Pavel Pevzner
Vineet Bafna

Dekel Tsur

Ari Frank
Bryant Forsgren
Nuno Bandeira
Stephen Tanner
Vagisha Sharma

Matt Wytock
Calvin Chen

Collaborators

Publications

Downloads

UCSD Computational Mass-Spectrometry Research Group

InsPecT performs high-throughput identification of peptide mass spectra with an emphasis on efficiently and confidently identifying modified peptides. Modifications include *in vivo* post-translational modifications such as phosphorylation, as well as *in vitro* chemical damage. We are able to search and score a broad range of modifications in a single search, or even identify unanticipated changes such as point mutations.

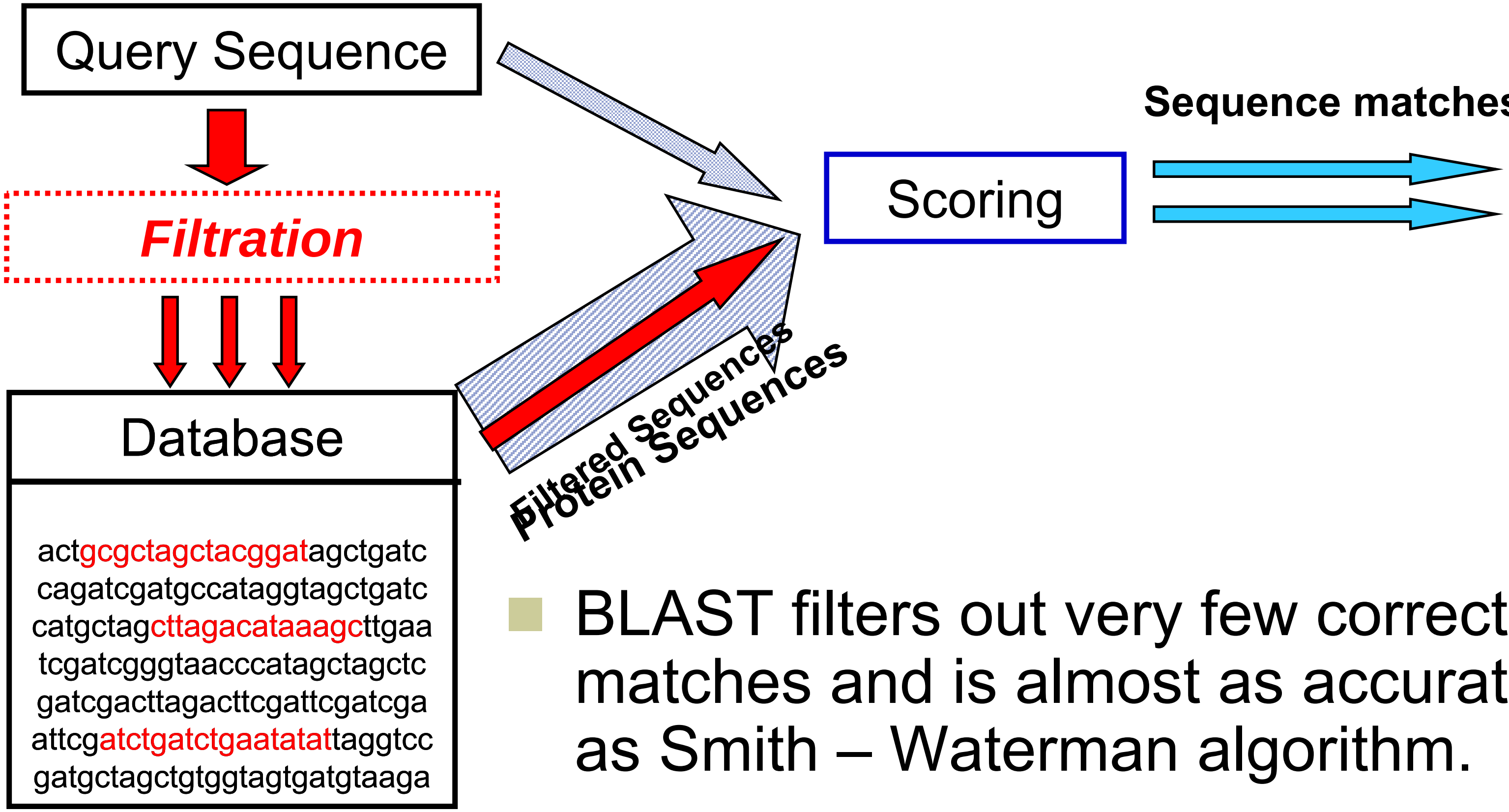
PepNovo is a software for de novo sequencing of peptides from mass spectra. PepNovo uses a probabilistic network to model the peptide fragmentation events in a mass spectrometer. In addition, it uses a likelihood ratio hypothesis test to determine if the peaks observed in the mass spectrum are more likely to have been produced under the fragmentation model, than under a probabilistic model that treats the appearance of peaks as random events.

Traditional analysis of tandem mass spectra focuses on the analysis of individual MS/MS spectra instead of capitalizing on the common event of repeated MS/MS spectra for the same peptide or combining spectra from partially overlapping peptides. Shotgun Protein Sequencing is a new approach to the analysis of tandem mass spectra that combines uninterpreted MS/MS spectra into ladders of overlapping spectra (multiple alignments of MS/MS spectra) before constructing a common amino acid interpretation for the whole multiple alignment.

Genomics: from SW Algorithm to BLAST

Sequence Alignment – BLAST

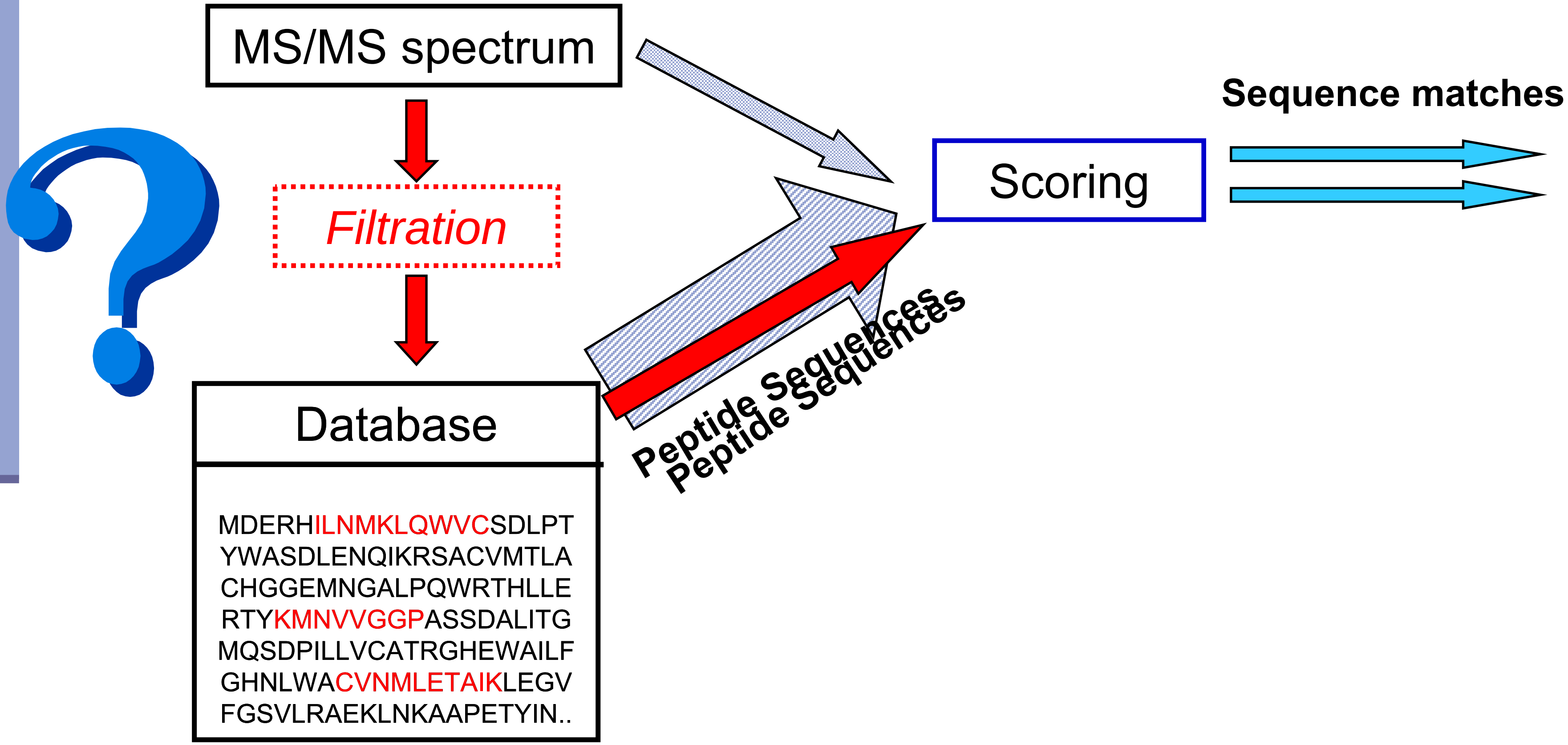
Sequence Alignment – Smith Waterman (SW) Algorithm


Query Sequence
Filtration
Database
actgcgctagctacggatagctgatc
cagatcgatgccataggtagctgatc
catgctagcttagacataaagcttgaa
tcgatcgggtaacccatagctagctc
gatcgacttagacttcgattcgatcga
attcgatctgatctgaatatattaggtcc
gatgctagctgtggtagtgatgtaaga
Filtered Sequences
Protein Sequences
Scoring
Sequence matches

- BLAST filters out very few correct matches and is almost as accurate as Smith – Waterman algorithm.

Proteomics: from SEQUEST to ???

Protein identification – SEQUEST, Mascot,…



MS/MS spectrum

Filtration

Database

MDERHILNMKLQWVCSDLPT
YWASDLENQIKRSACVMTLA
CHGGEMNGALPQWRTHLLE
RTYKMNVVGGPASSDALITG
MQSDPILLVCATRGHEWAILF
GHNLWACVNMLETAIKLEGV
FGSVLRAEKLNKAAPETYIN..

Peptide Sequences

Peptide Sequences

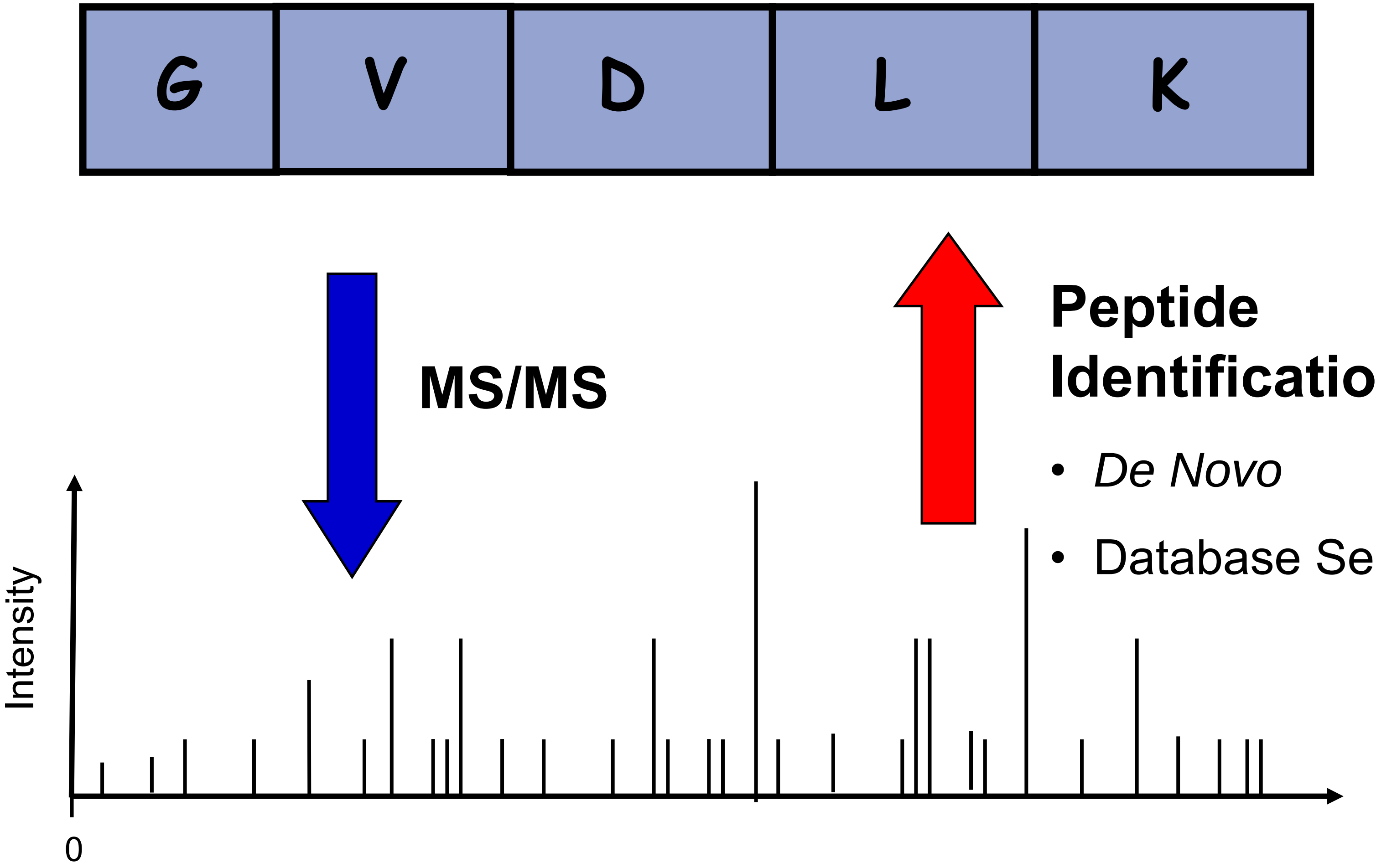
Scoring

Sequence matches

Filtration in Tandem Mass Spectrometry

- Filtration in MS/MS is more difficult than in BLAST.
- The approaches based on Peptide Sequence Tags were not able to substitute the complete database search and are mostly used to generate additional identifications rather than replace the database search.
- **InsPecT** (Tanner et al., *Anal. Chem*. July 2005) - filtration-based search that replaces the complete database search and is orders of magnitude faster.

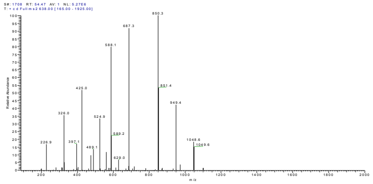
Protein Identification with MS/MS



Peptide Identification:

- *De Novo*
- Database Search

De Novo vs. Database Search



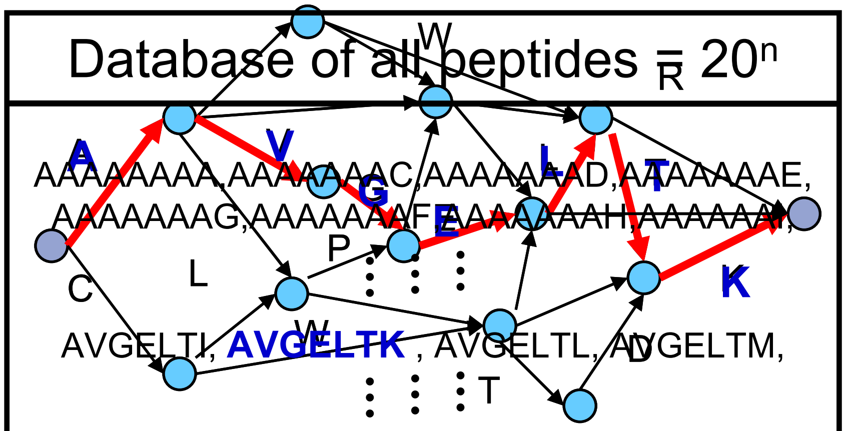
Database Search

Database of known peptides

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

De Novo

Database of all peptides = 20^n



AAAAAAAA,AAAAAAAC,AAAAAAAD,AAAAAAAE,
AAAAAAAG,AAAAAAAF,AAAAAAAH,AAAAAAAI,
AVGELTI, **AVGELTK** , AVGELTL, AVGELTM,
⋮ ⋮ ⋮
YYYYYYYS,YYYYYYYT,YYYYYYYV,YYYYYYYY

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.

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!*
- **PepNovo** (Frank and Pevzner, *Anal. Chem*., 2005) – fast and accurate *de novo* algorithm (0.1 sec to sequence a peptide, at least an order of magnitude faster than other approaches).

Why Not Sequence De Novo?

Algorithm	Avg. Predicted Length	Amino Acid Accuracy	Completely Correct Predictions
Lutefisk (Taylor and Johnson, 1997)	8.8	0.56	0.19
SHERENGA (Dancik et al., 1999)	8.7	0.69	0.29
Peaks (Ma et al., 2003)	10.3	0.67	0.25
PepNovo (Frank and Pevzner, 2005)	10.3	**0.73**	**0.30**
EigenMS (Bern and Goldberg 2005)	…	…	…

- *De novo* sequencing is still not accurate enough!

So What Can be Done with De Novo?

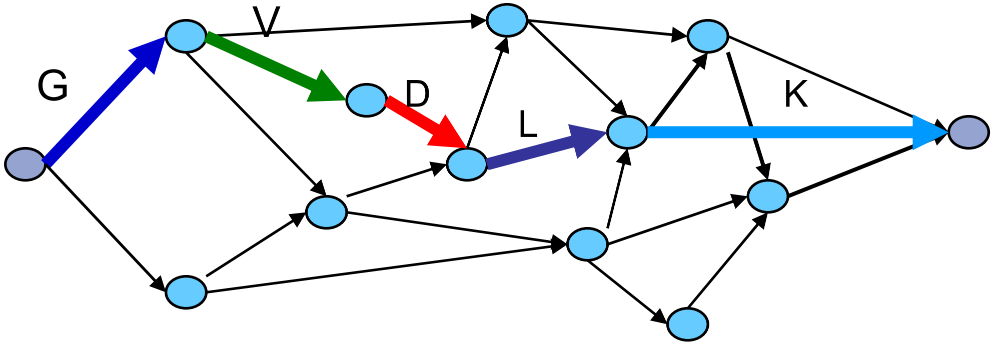
- Given an MS/MS spectrum:
  - Can *de novo* predict the entire peptide sequence? **- No!**

    (accuracy is less than 30%).
  - Can *de novo* predict a correct tag? **- No!**

    (accuracy less than 50% - GutenTag [Tabb et al. 2003] , only 80% - PepNovo)
  - Can *de novo* predict a **small** set of tags that, with high probability has at least one correct tag? **- Yes!**

A Covering Set of Tags

Peptide Sequence Tags



- A Peptide Sequence Tag is a short substring of a peptide path.

Example: G V D L K

Tags:
- G V D at mass 0.
- V D L at mass 57.
- D L K at mass 161.1

Filtration with Peptide Sequence Tags

- **The Filtration:** Consider only database peptides that contain the tag (in its correct relative mass offsets).
- First suggested by Mann and Wilm (1994).
- Similar concepts also used by:
  - GutenTag - Tabb et al. 2003.
  - MultiTag - Sunayev et al. 2003.
  - OpenSea - Searle et al. 2004.
- **PepNovoTag (**Frank et al., *J. of Proteome Res.* 2005) – provides a getaway to filtration-based MS/MS analysis by generating **covering sets** of tags (with high probability).

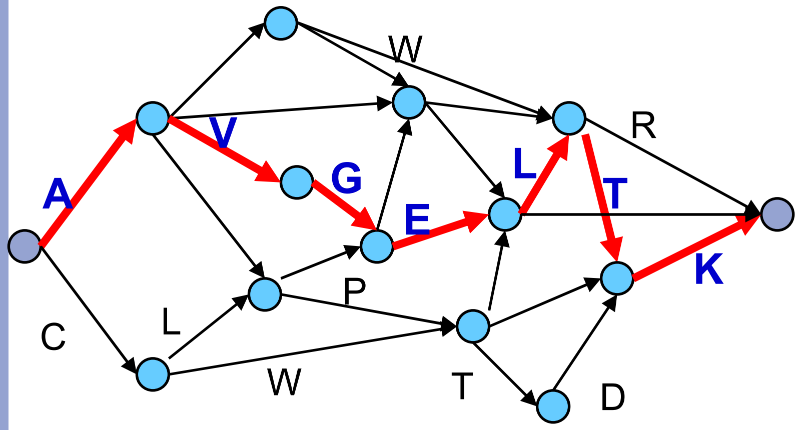
Why Filter Database Candidates?

- Database programs such as SEQUEST or Mascot are slow.
- Only simple filtration techniques are used:
  - parent mass
  - tryptic ends
  - two phase protein filtration (X! tandem)

- Effective filtration can greatly speed-up the process, enabling expensive searches involving post-translational modifications.

- Our Goal:
  To *generate a small set of covering tags and use them to filter the database peptides.*

Tag Generation - Global Tags

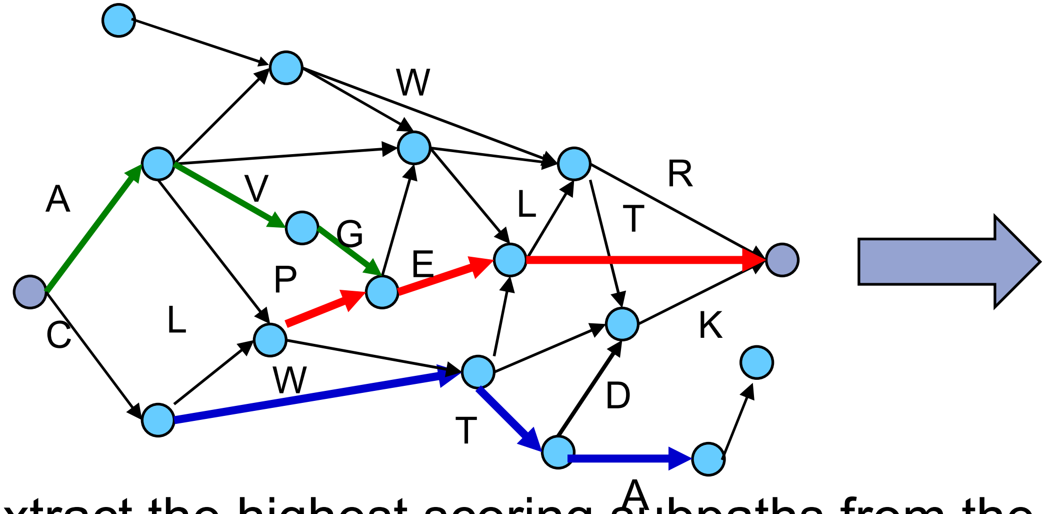


AVGELTK

TAG	**Prefix Mass**
AVG	0.0
VGE	71.0
GEL	170.1
ELT	227.1
LTK	356.2

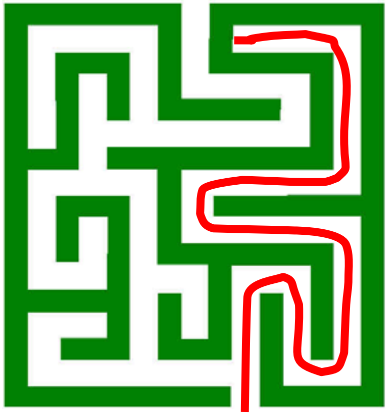
- Parse tags from PepNovo’s *de novo* sequence.
- If the *de novo* sequence is completely incorrect, none of the tags will be correct.
- Only a small number of tags can be generated.

Tag Generation



TAG	Prefix Mass
AVG	0.0
WTA	120.2
PET	211.4

- Extract the highest scoring subpaths from the spectrum graph.
- Each additional tag increases the number of database hits and slows down the database search. Therefore, tags should be ranked (tricky)
- Sometimes gets misled by locally promising-looking “garden paths”.



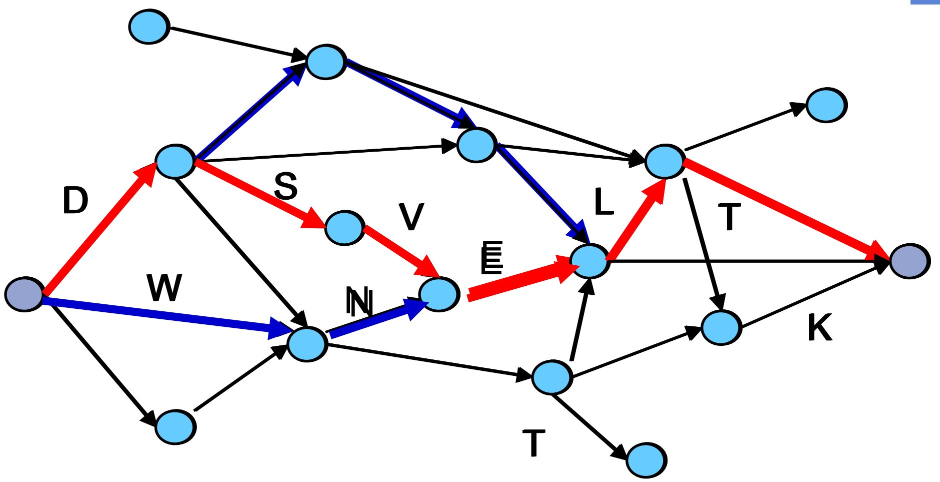
Ranking Tags

- Each additional tag used to filter increases the number of database hits and slows down the database search.
- Tags can be ranked according to their scores, however this ranking is not very accurate.
- It is better to determine the probability that each tag is correct, and choose the most probable tags.

Reliability of Amino Acids in Tags

- For each amino acid in a tag we want to assign a probability that it is correct.
- Each amino acid, which corresponds to an edge in the spectrum graph, is mapped to a feature space that consists of the following features:
  - Score Reduction due to edge removal
  - The edge's vertex scores
  - Presence of consecutive fragment ions
  - more..
- We use a logistic regression model to predict the probability that an amino acid is correct.

Removing Edges from the Spectrum Graph



- The removal of an edge corresponding to a genuine amino acid usually leads to a reduction in the score of the *de novo* path.
- The removal of an edge that *does not* correspond to a genuine amino acid tends to cause a smaller reduction.

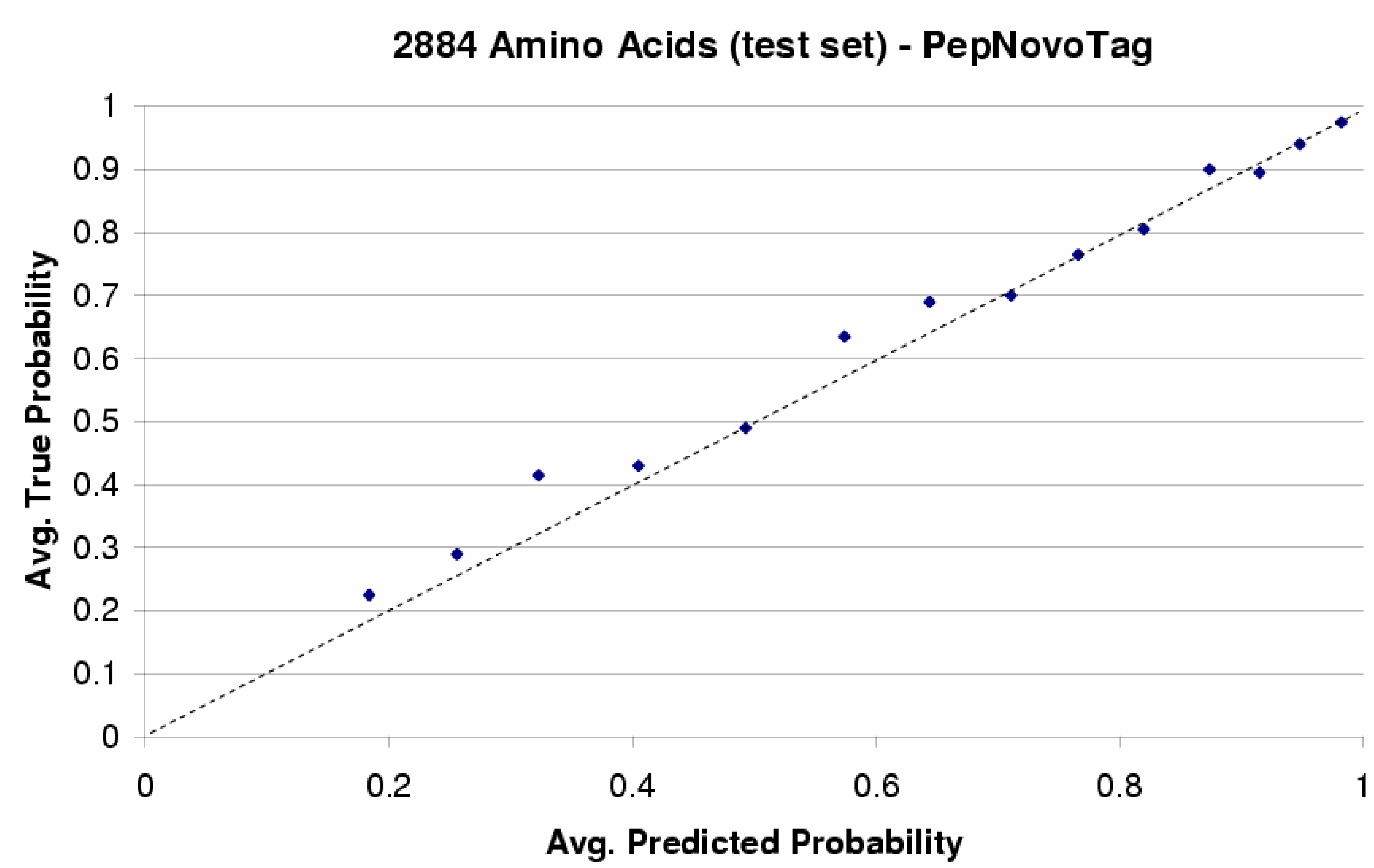
Logistic Regression Models

- Each amino acid instance x is mapped into an *n*-dimensional feature space, and can belong to one of two classes (correct, incorrect).

$$p(\mathrm{correct} \mid \mathrm{x}) = \frac{\exp\left(\lambda_0 + \sum_{i=1}^{n} \lambda_i \cdot x_i\right)}{1 + \exp\left(\lambda_0 + \sum_{i=1}^{n} \lambda_i \cdot x_i\right)}$$

- The weights λ_i are learned from the training data.

Probability of Amino Acids



- The amino acids were sorted according to their predicted probability, and grouped in bins of 200.

Probabilities of Tags

- How do we determine the probability of a predicted tag?

- We use the predicted probabilities of its amino acids for features in an additional logistic regression model.

- We follow the concept that *“a chain is only as strong as its weakest link”.*

Comparing GutenTag and PepNovoTag

	Length 3		Length 4		Length 5	
Algorithm \ #tags	**1**	**10**	**1**	**10**	**1**	**10**
PepNovoTag	0.804	**0.961**	0.732	0.900	0.664	0.803
GutenTag	0.493	0.893	0.418	0.782	0.318	0.643

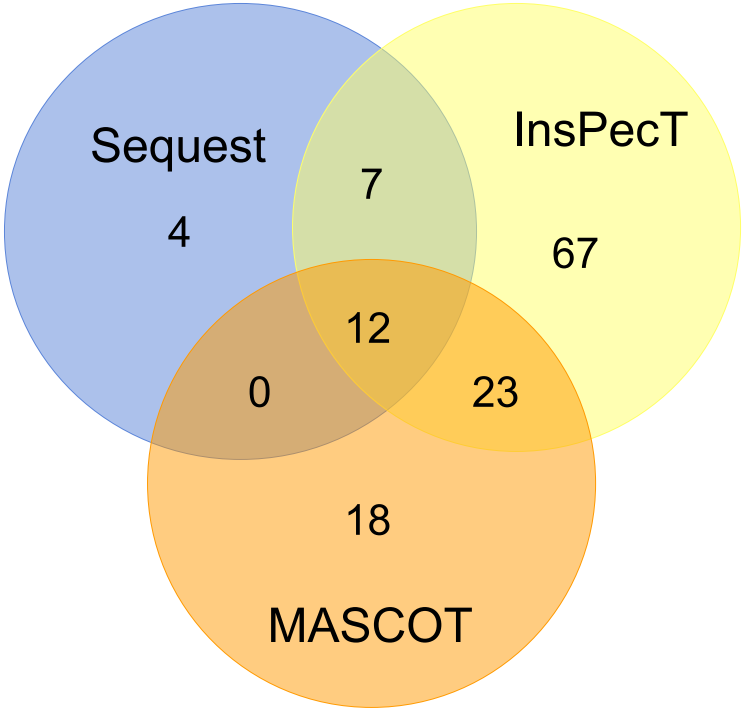
- Results are for 280 spectra of doubly charged tryptic peptides from the ISB and OPD datasets.
- The table shows the proportion of spectra for which at least one correct tag was generated.
- GutenTag is a tag generation algorithm developed in John Yates' group (Tabb et al. 2003).

Comparing Sequest with InsPecT

PTMs	**Tag Length**	**No. Tags**	**No. Candidates**	**InsPecT Runtime**	**SEQUEST Runtime**
None	3	1	181	0.17 sec	~ 1 minute
	3	10	888	**0.27 sec**	
Phosphory -lation	3	1	311	0.21 sec	~ 2 minutes
	3	10	1480	0.38 sec	

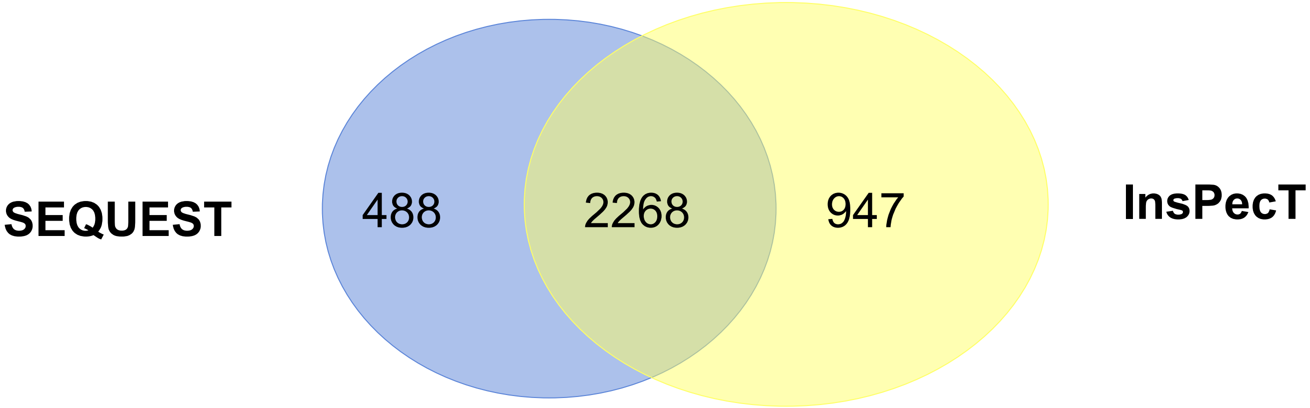
- InsPecT was used to determine filtration efficiency and runtime (run on a 3GHz desktop PC).
- The search was done against SWISS-PROT (54Mb).
- **A reminder**: many labs generate more than 100,000 spectra per day. It would take SEQUEST 2 months to analyze this data on a desktop.

Comparing Sequest, Mascot, and InsPecT



Phosphopeptides identified over 50,000 mouse spectra
(collaboration with Mark Mumby at Alliance for Cell. Signalling)

More Search Results



Spectra accurately annotated on the ISB data-set, a collection of 22,000 spectra from a known protein mixture

- Searching with a set of 7 PTMs allowed annotation of 16% more spectra, and 20% more distinct peptides.

Advantages of Filtration in MS/MS Searches

- Inspect with10 tags of length 3:
  - The filtration is **1500** times more efficient than using only the parent mass as a filter (SEQUEST).
  - Less than **4%** of the positive peptides are filtered out.
  - The search is 150 times faster than SEQUEST (per spectrum).

Advantages of Filtration in MS/MS Searches

- Inspect with10 tags of length 3:
  - The filtration is **1500** times more efficient than using only the parent mass as a filter (SEQUEST).
  - Less than **4%** of the positive peptides are filtered out.
  - The search is more than 150 times faster than SEQUEST (per spectrum).
- ***Tags from different spectra can be pooled together to take advantage of the Aho-Corasik algorithm***
- Since runtime is dramatically reduced InsPecT can perform more complex searches for post translational modifications that were not possible in the past

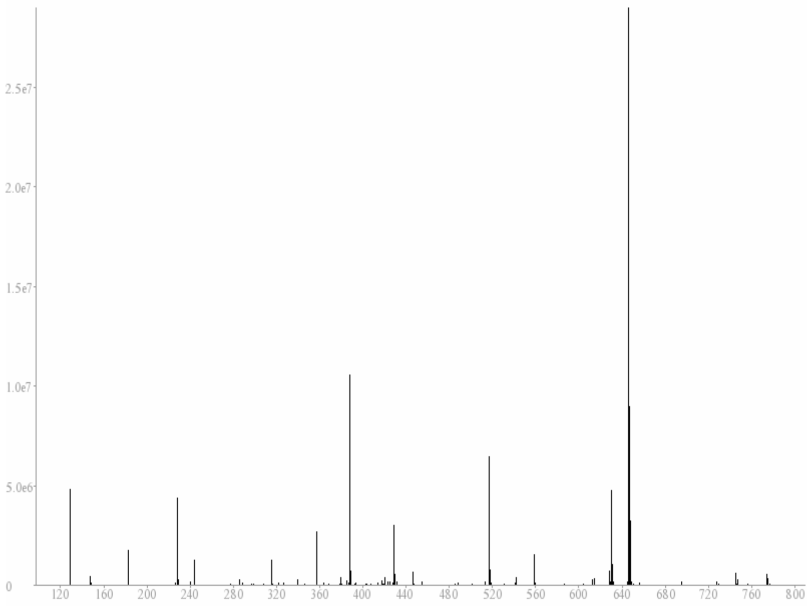
Peptide Identification Problem

Input:

- A protein database
- A *Spectrum*
- A function *SCORE(Spectrum, Peptide)* evaluating how well a *Peptide* ‘explains’ a *Spectrum.*

QDKIHPFAQTQSLVYPFPGPIPN
SLPQNIPPLTQTPVVVPPFLQPE
VMGVSKVKEAMAPKHKEMPFP
KYPVEPFTESQSLTLTDVENLHL
PLPLLQSWMHQPHQPLPPTVMF
PPQSVLSLSQSKVLPVPQK...

Database



Spectrum

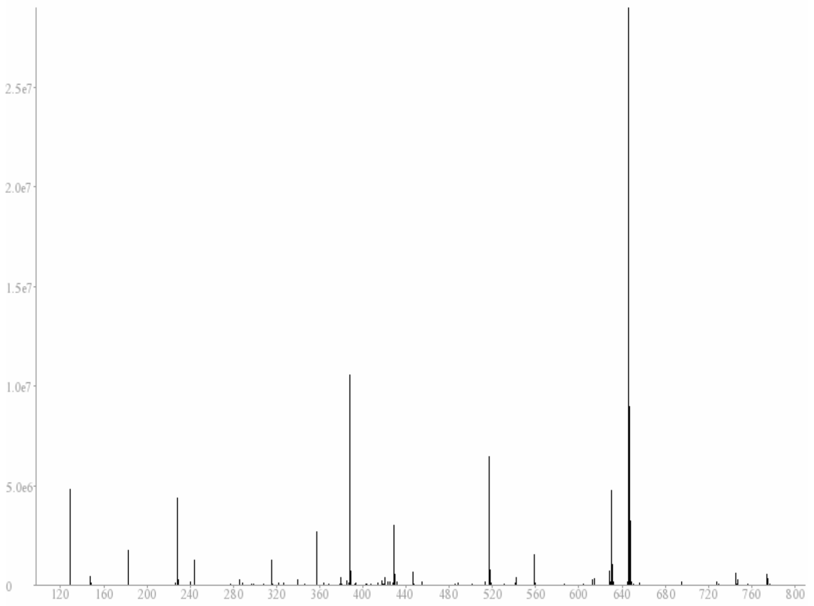
Peptide Identification Problem

Output:

- A *Peptide* in the database which maximizes *SCORE(Spectrum, Peptide)*

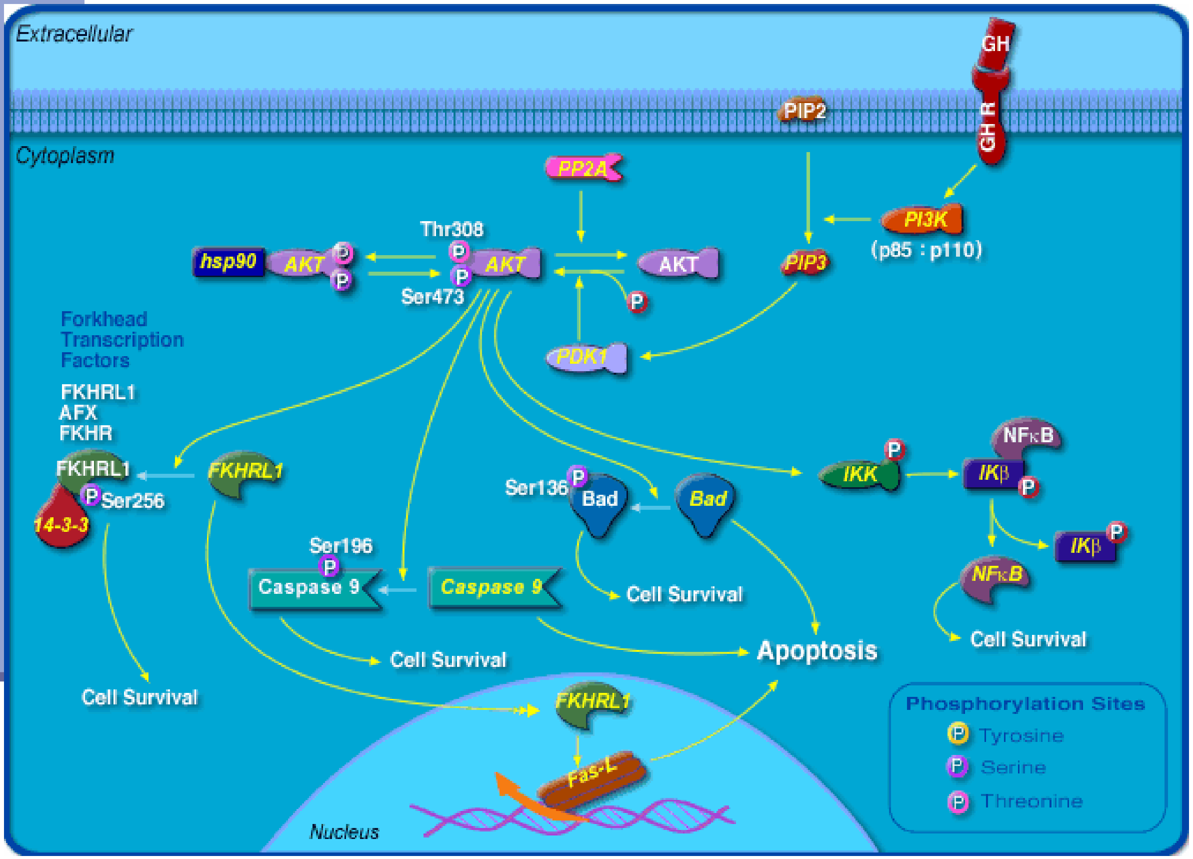
QDKIHPFAQTQSLVYPFPGPIPN
SLPQNIPPLTQTPVVVPPFLQPE
VMGVSK**VKEAMAPK**HKEMPFP
KYPVEPFTESQSLTLTDVENLHL
PLPLLQSWMHQPHQPLPPTVMF
PPQSVLSLSQSKVLPVPQK...

Database



Spectrum

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.

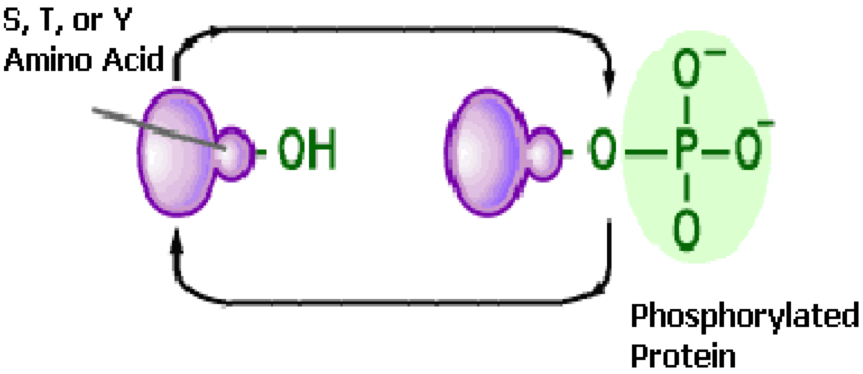
⇩

They are subject to a large number of biological modifications.

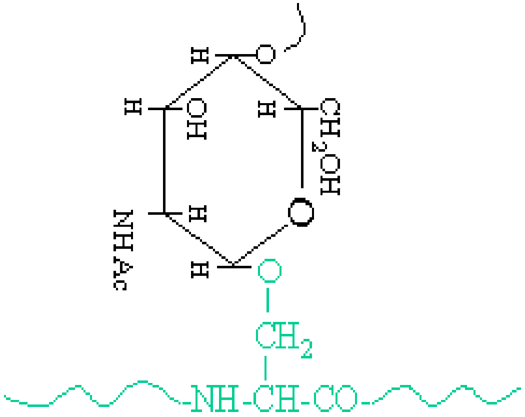
Almost all protein sequences are post-translationally modified and **200 types of modifications** of amino acid residues are known.

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.

Sequencing of Modified Peptides

De novo peptide sequencing is invaluable for identification of **unknown** proteins:

However, *de novo* algorithms are designed for working with high quality spectra with good fragmentation and without modifications.

Another approach is to compare a spectrum against a set of known spectra in a database.

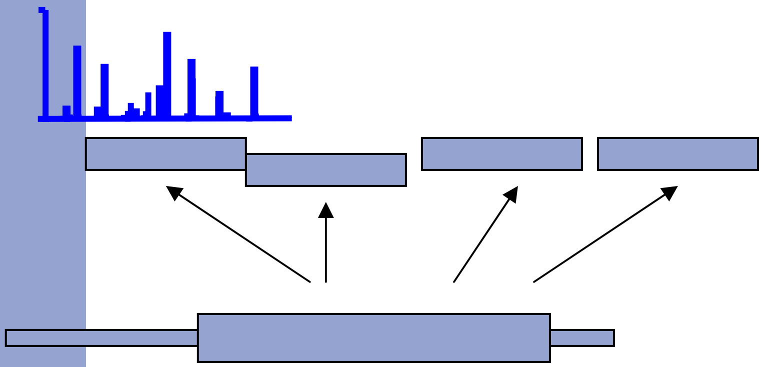
Search for Modified Peptides: Virtual Database Approach

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

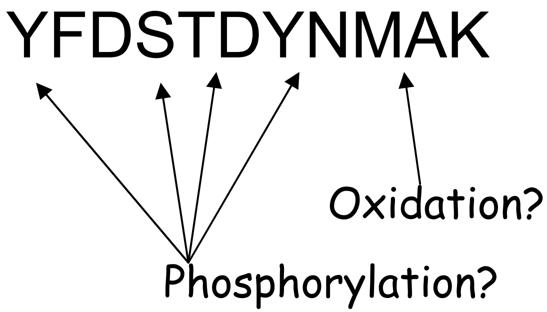
Exhaustive search leads to a large combinatorial problem, even for a small set of modifications types.

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!

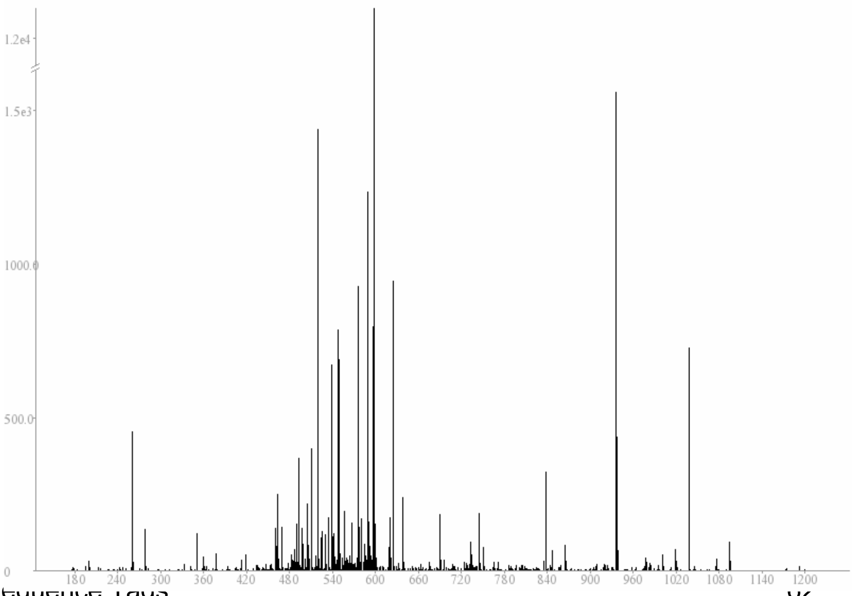
Identification of Modified Peptides

Input:

- A protein database
- A *Spectrum*
- A function *SCORE(Spectrum, Peptide)* evaluating how well a *Peptide* ‘explains’ a *Spectrum*
- **Maximum number of modifications, *k***

VDIVVSEDLNGTVKFSSSLPYPN
NLNSVLAERLEKWLQLMLMWH
PRQRGTDPTYGPNGCFKALDDI
LNLKLVHILNMVTGTIHTYPVTED
ESLQSLKARIQQDTGIPEEDQEL
LQEAGLALIPDKPATQCISDGKL
NEGHTLDMDLVFLFDNSKITYET
QISPRPQPESVSCILQEPKRN...

Database



Spectrum

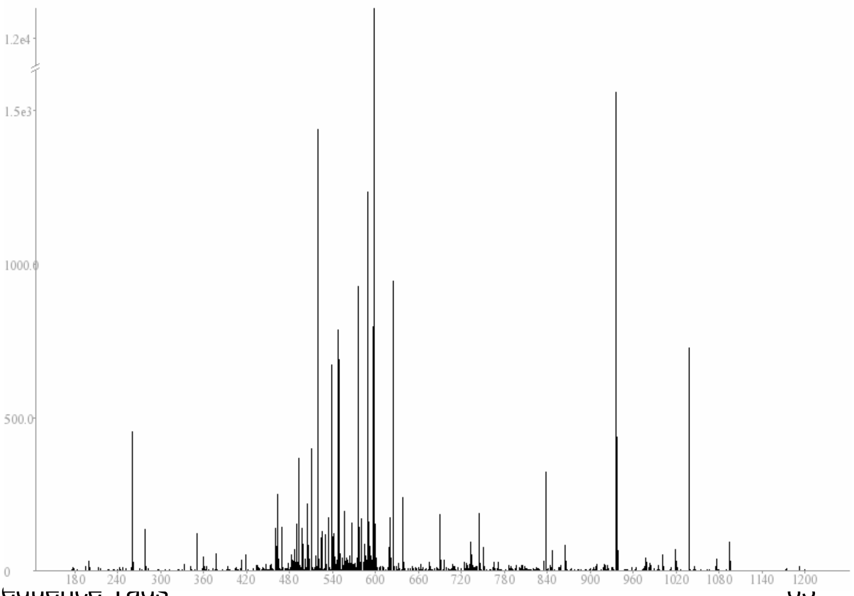
Identification of Modified Peptides

Output:

- A *Peptide* **with up to *k* modifications** which maximizes *SCORE(Spectrum, Peptide)*

VDIVVSEDLNGTVKFSSSLPYPN
NLNSVLAERLEKWLQLMLMWH
PRQRGTDPTYGPNGCFKALDDI
LNLK**LVHILNM#VTGT**IHTYPVTE
DESLQSLKARIQQDTGIPEEDQE
LLQEAGLALIPDKPATQCISDGK
LNEGHTLDMDLVFLFDNSKITYE
TQISPRPQPESVSCILQEPKRN...

Database



Spectrum

Search for Modified Peptides: Virtual Database Approach

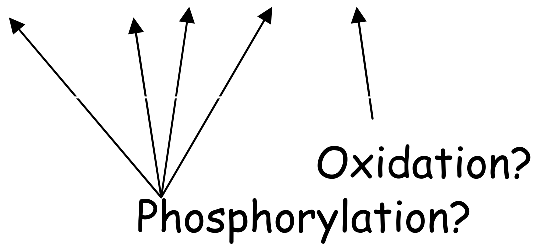
Yates et al.,1995: an exhaustive search in a virtual database of all modified peptides.

Combinatorial explosion, even for a small set of modifications types.

A larger set of spurious matches must be filtered out. It's much more likely that incorrect matches will have high scores.

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

- YFDSTDYNMAK



Oxidation?

Phosphorylation?

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

Restrictive vs Unrestrictive (Blind) Search for Modified Peptides

- **Restrictive** search (conventional tools) requires the researcher to guess which modification types are present in the sample
- **MS-Alignment** (Tsur et al., 2005, *Nature Biotech*) performs an **unrestrictive** (blind) search for *all* possible modification offsets at once.
- MS-Alignment for all possible modification offsets is about as fast as SEQUEST (in the *k=1* mode)
- Although MS-Alignment becomes slower than SEQUEST in k>1 mode, it still can be run on databases representing complex protein mixtures.

Sequence Analysis vs. MS/MS Analysis

Sequence analysis:

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

MS/MS analysis:

similar peptides (a few mut/mod 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.

Sequence Alignment=Path in a Grid

Finding similarities between

two peptides

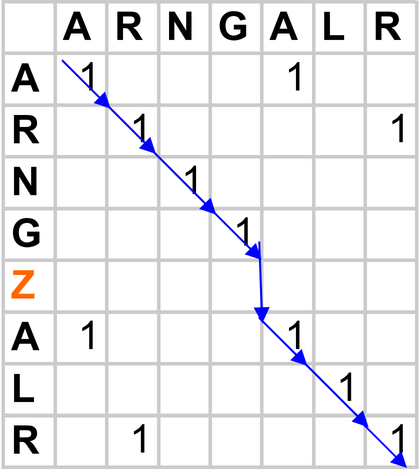
is equivalent to finding an optimal path in a Manhattan-like grid (**sequence alignment**).

	A	R	N	G	A	L	R
A	1				1		
R		1					1
N			1				
G				1			
Z							
A	1				1		
L						1	
R		1					1

Sequence Alignment=Path in a Grid

Finding similarities between

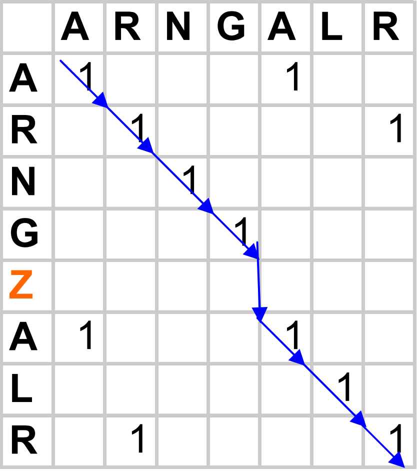
two peptides



	A	R	N	G	A	L	R
A	1				1		
R		1					1
N			1				
G				1			
Z							
A	1				1		
L						1	
R		1					1

is equivalent to finding an optimal path in a Manhattan-like grid (**sequence alignment**). Every horizontal/vertical segment in this path corresponds to insertion/deletion of an amino acid.

Sequence Alignment=Path in a Grid



	A	R	N	G	A	L	R
A	1				1		
R		1					1
N			1				
G				1			
Z							
A	1				1		
L						1	
R		1					1

Finding similarities between

two peptides

is equivalent to finding an optimal path in a Manhattan-like grid (**sequence alignment**). Every horizontal/vertical segment in this path corresponds to insertion/deletion of an amino acid.

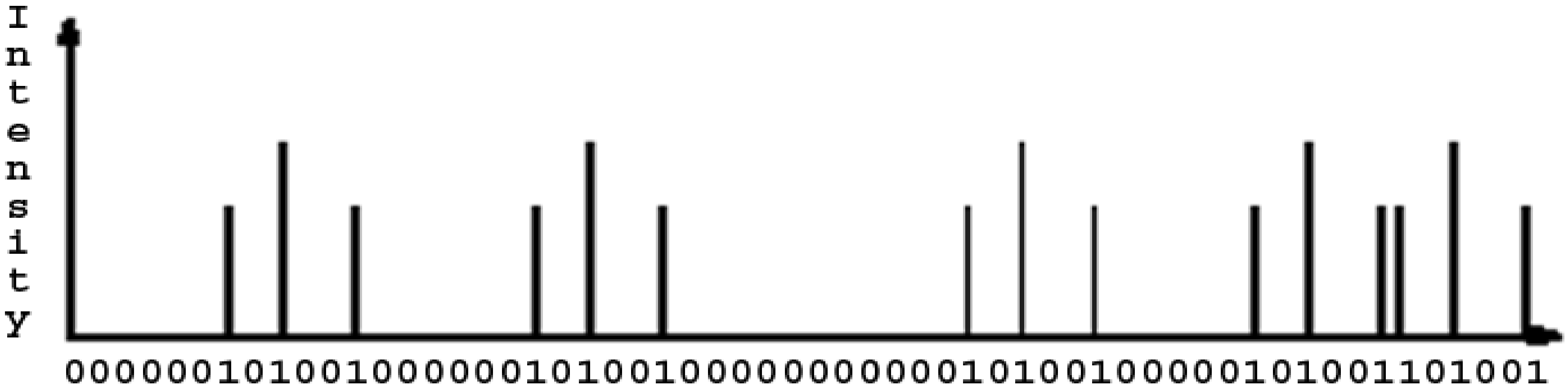
Can we find similarities between

a spectrum and a peptide

using a similar approach (**spectral alignment**)?

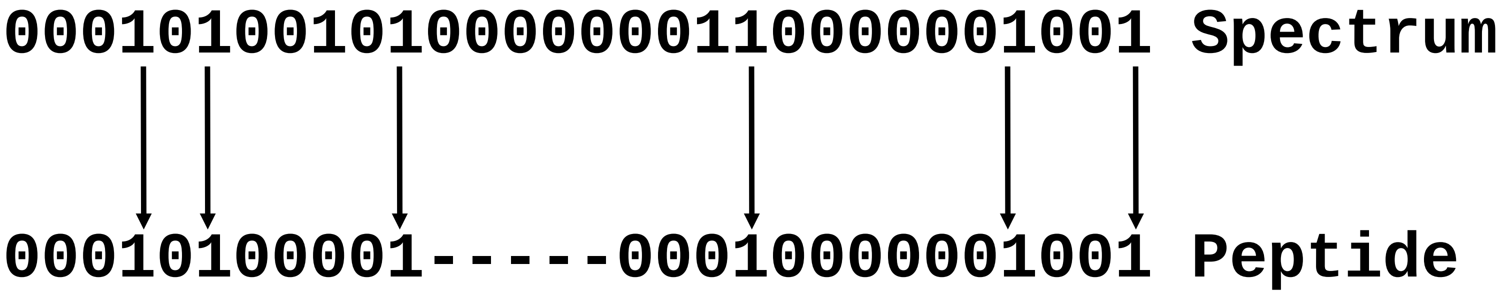
Converting Spectra into 0-1 Sequences

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



Modified peptide

Modifications are modeled as insertion (or deletions) of blocks of zeroes



000101001010000000110000001001 Spectrum

00010100001-----00010000001001 Peptide

A modification with positive offset - *inserting* a block of 0s

A modification with negative offset - *deleting* a block of 0s

Spectra Comparing vs. String Comparison

- Comparison of theoretical and experimental spectra (represented as 0-1 strings) corresponds to a (somewhat unusual) **edit distance/alignment** between 0-1 strings where elementary edit operations are insertions and deletions of blocks of 0s
- **Use sequence alignment algorithms!**

Spectral Alignment Graph

		0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
	0																									
	0																									
	0																									
	0																									
A	1				1					1						1					1					1
	0																									
	0																									
	0																									
	0																									
B	1				1					1						1					1					1
	0																									
	0																									
C	1				1					1						1					1					1
	0																									
	0																									
	0																									
	0																									
D	1				1					1						1					1					1
	0																									
	0																									
	0																									
	0																									
E	1				1					1						1					1					1

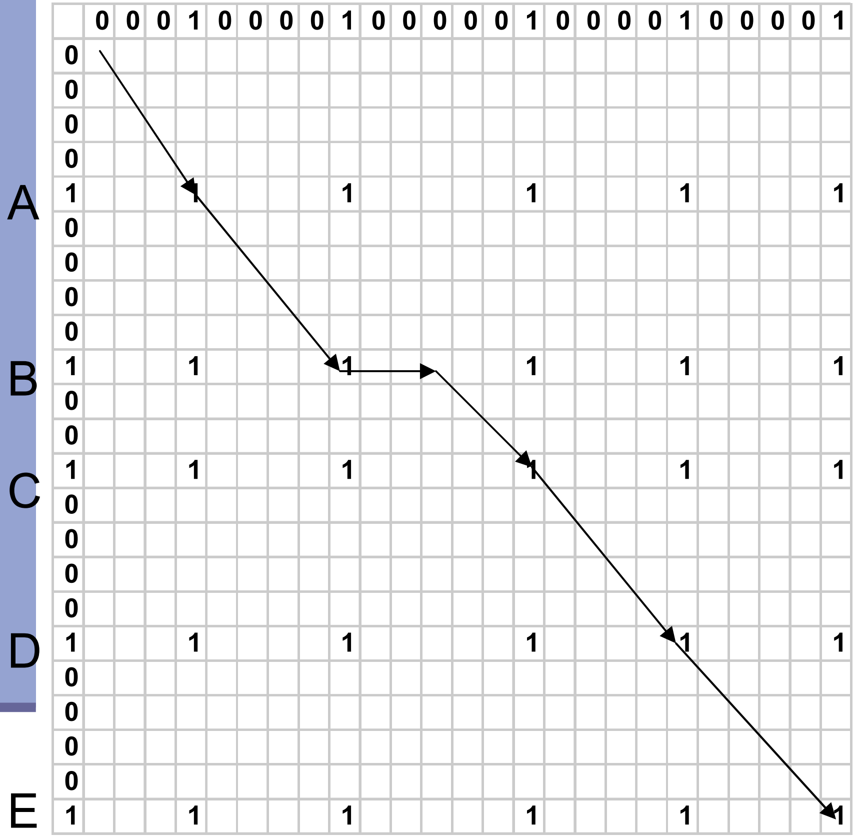
Horizontal axis:

Experimental spectrum

Vertical axis:

Theoretical spectrum of entire database

Spectral Alignment Graph



		0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
	0																									
	0																									
	0																									
	0																									
A	1				1					1						1					1					1
	0																									
	0																									
	0																									
	0																									
B	1				1					1						1					1					1
	0																									
	0																									
C	1				1					1						1					1					1
	0																									
	0																									
	0																									
	0																									
D	1				1					1						1					1					1
	0																									
	0																									
	0																									
	0																									
E	1				1					1						1					1					1

Like in SW alignment algorithm, every **path** in the spectral alignment graph represents a possible interpretation of a spectra.

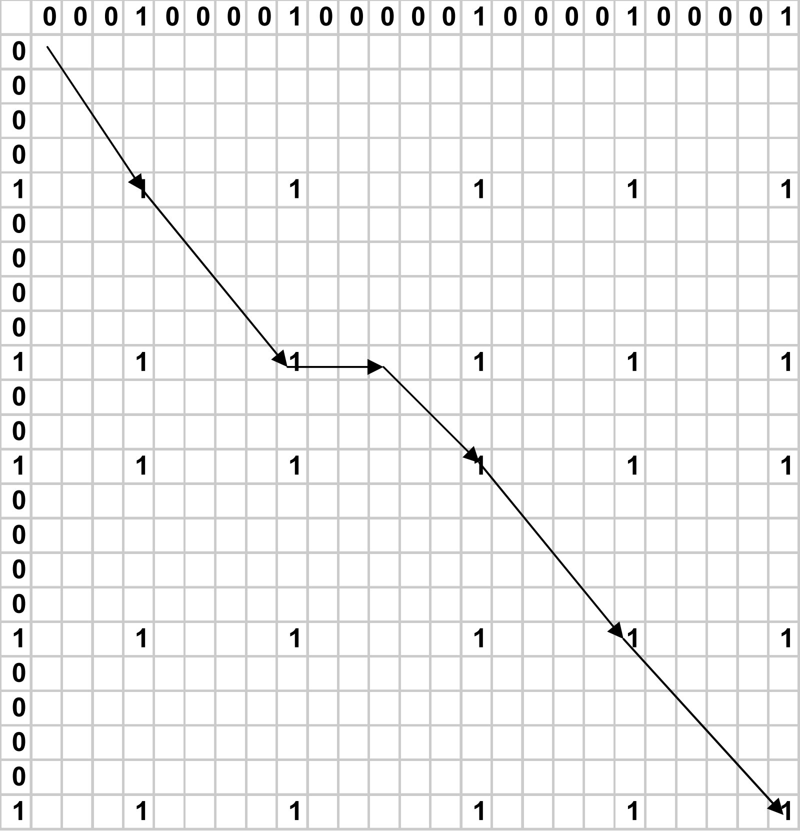
A path covering maximal number of 1s is the “best” interpretation of the spectrum.

Vertical / horizontal segment in the optimal path are **modifications**

Spectral Alignment vs. Sequence Alignment

- Alignment 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 Alignment Algorithm



	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
0																									
0																									
0																									
0																									
1				1					1						1					1					1
0																									
0																									
0																									
0																									
1				1					1						1					1					1
0																									
0																									
1				1					1						1					1					1
0																									
0																									
0																									
0																									
1				1					1						1					1					1
0																									
0																									
0																									
0																									
1				1					1						1					1					1

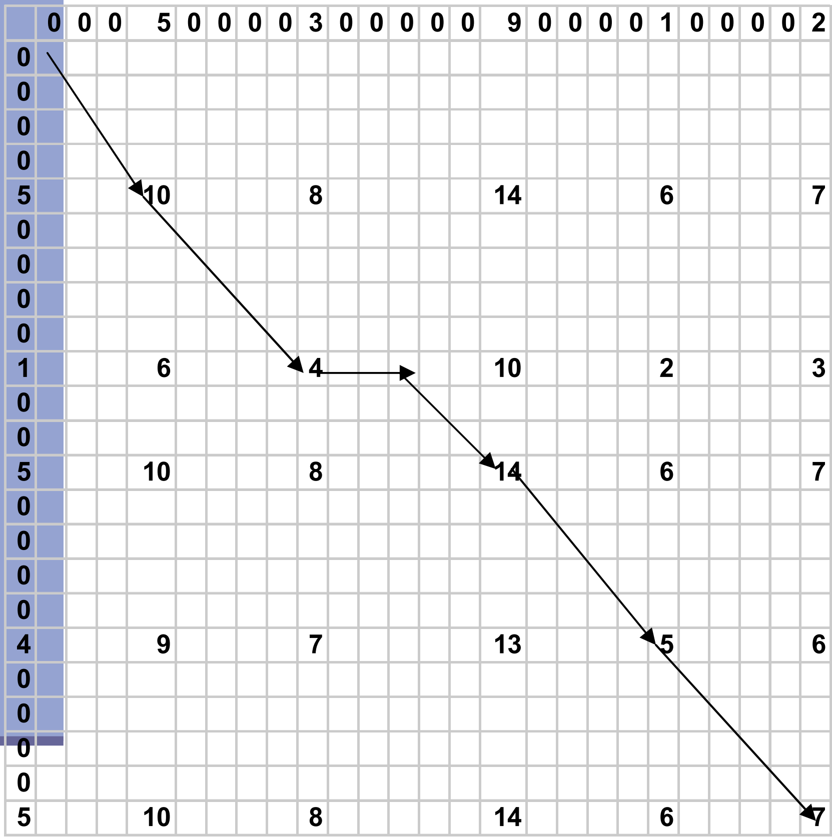
Spectral alignment was introduced in Pevzner et al.,2000.

MS-Alignment addresses a number of open problems in Pevzner et al.,2000:

Simultaneous analysis of N- and C-terminal ions
Taking into account the intensities and charges
Analysis of neutral losses
Speed
………

These improvements led to a fast algorithm that, for the first time, made blind PTM search in complex mixtures practical

Enriching the model



	0	0	0	5	0	0	0	0	3	0	0	0	0	0	9	0	0	0	0	1	0	0	0	0	2
0																									
0																									
0																									
0																									
5				10					8						14					6					7
0																									
0																									
0																									
0																									
1				6					4						10					2					3
0																									
0																									
5				10					8						14					6					7
0																									
0																									
0																									
0																									
4				9					7						13					5					6
0																									
0																									
0																									
0																									
5				10					8						14					6					7

Fitting, seeded alignment

Masses are prefix residue masses (PRMs) supported by b and/or y peaks and neutral losses

Masses need not be integers

Vertices have arbitrary scores: MassScore(v)

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

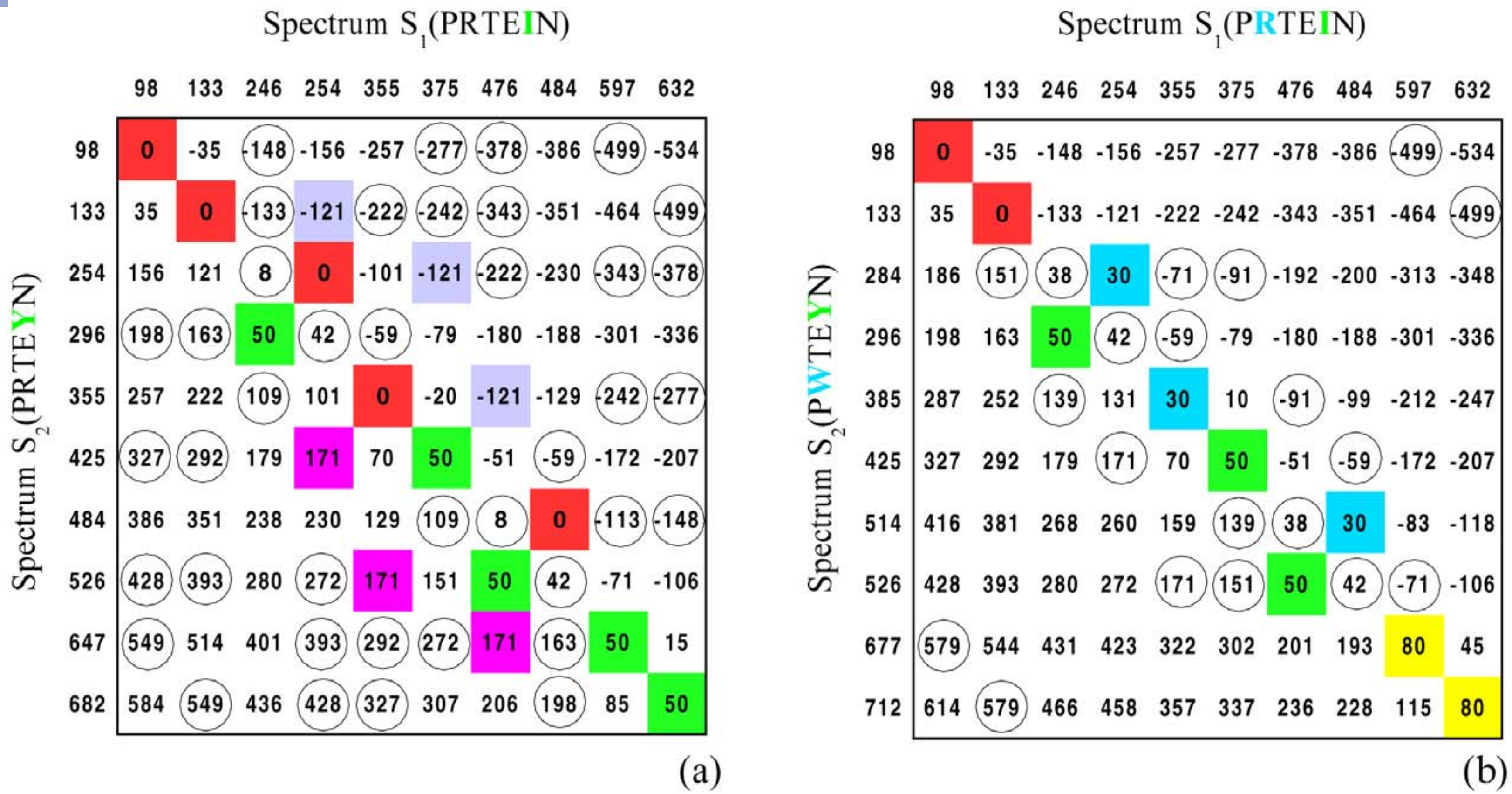
Goal: Find a modified peptide from the database with maximal match between an experimental and theoretical spectrum.

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



Spectrum S_1(PRTEIN)

Spectrum S_2(PRTEYN)

	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(PRTEIN)

Spectrum S_2(PWTEYN)

	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

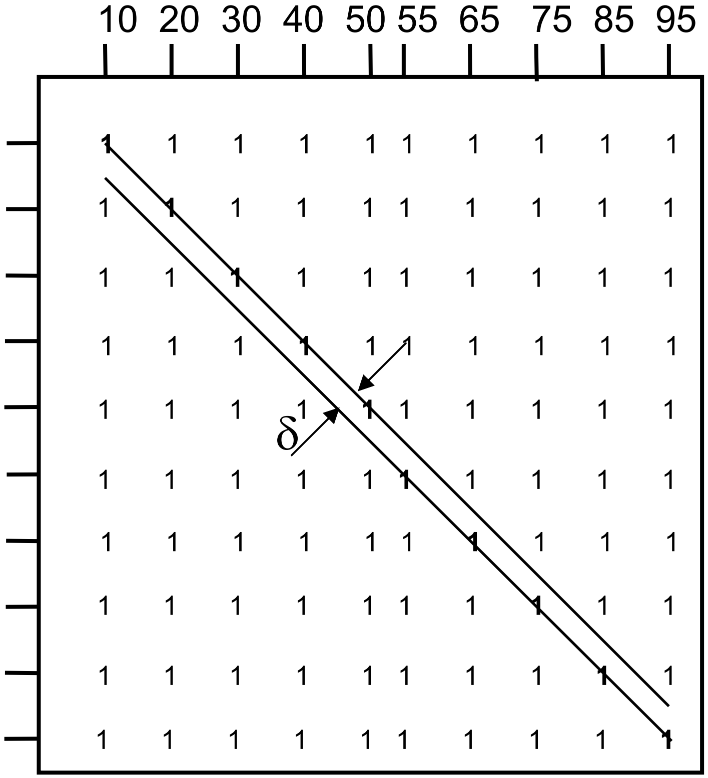
Spectral Product

$A=\{a_1, \ldots., a_n\}$ and $B=\{b_1,\ldots., b_n\}$

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 at the main diagonal.

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



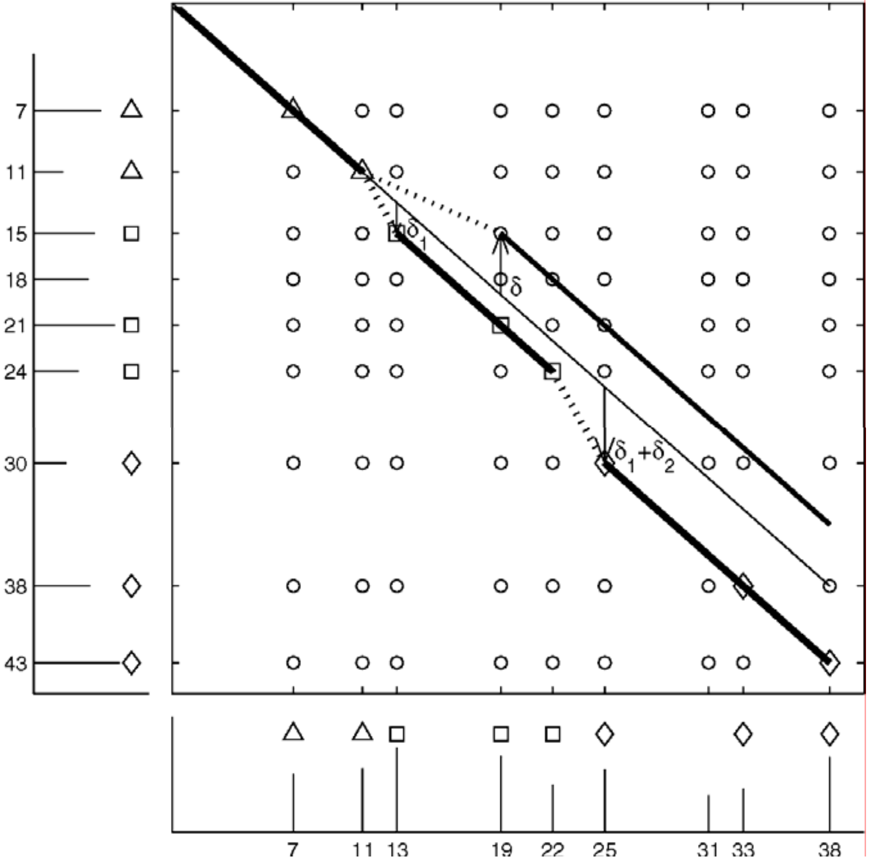
$$\begin{array}{cccccccccc}
10 & 20 & 30 & 40 & 50 & 55 & 65 & 75 & 85 & 95 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}$$

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

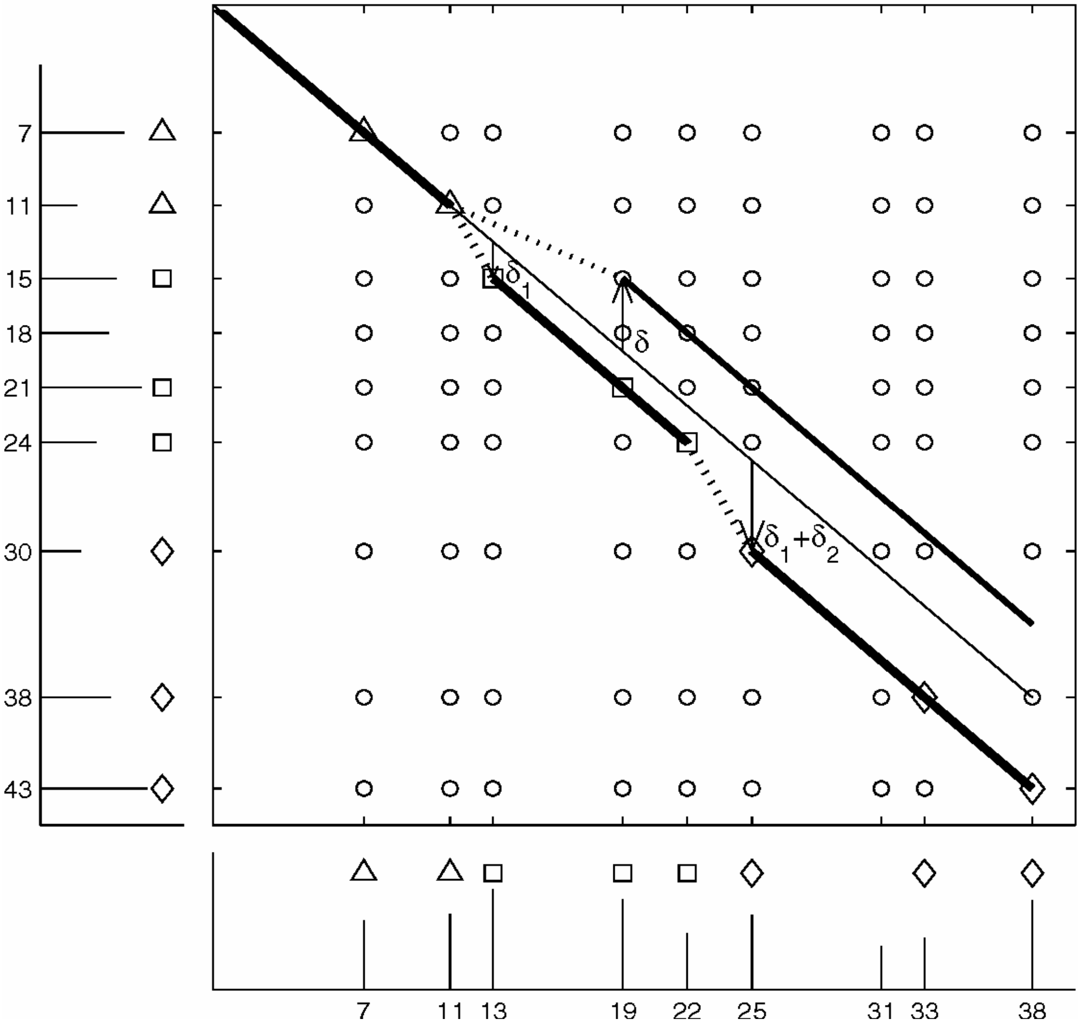
***k*-similarity between spectra**: the maximum number of 1s on a path through this graph that uses at most *k+1* diagonals.

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

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



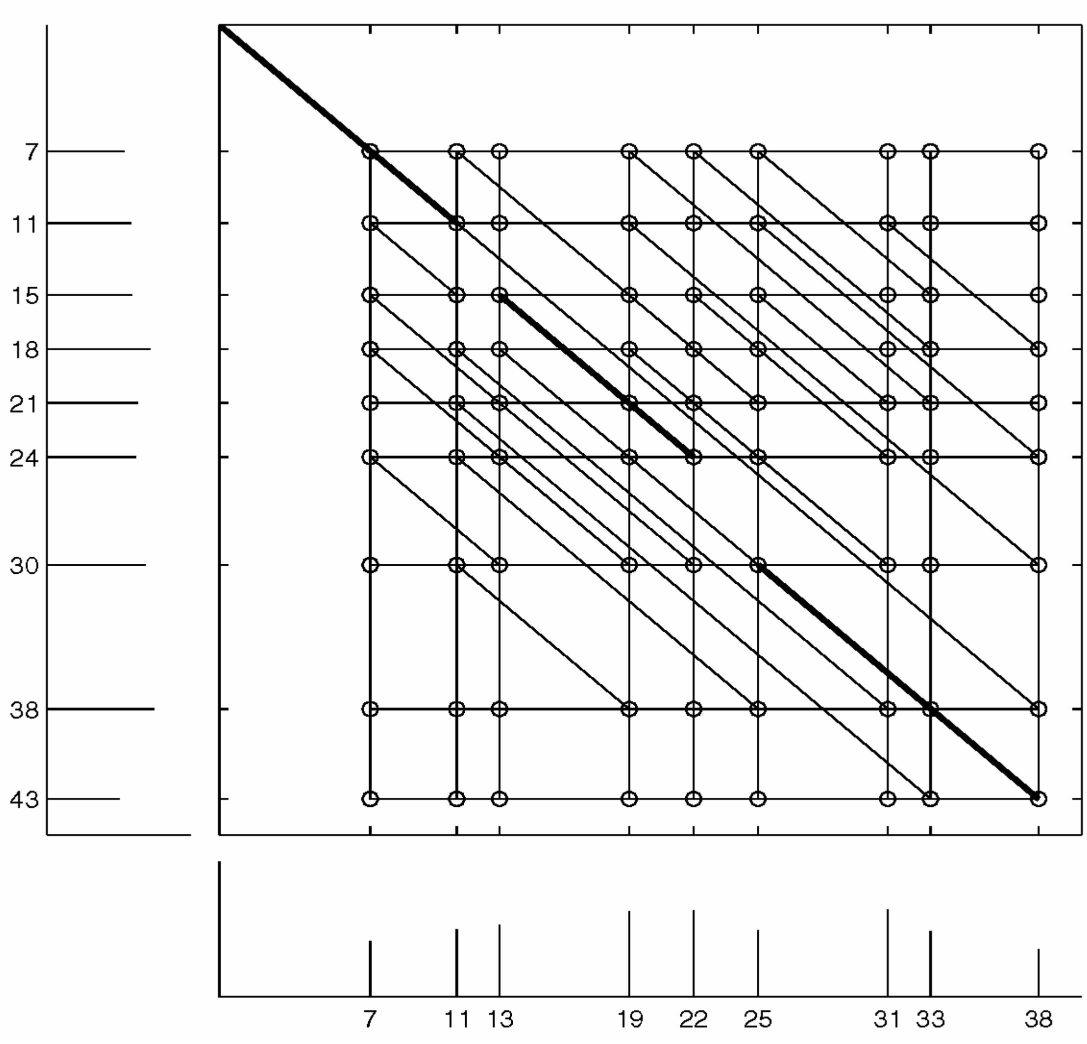
Finding Peptides with Multiple Modifications



By changing parameter *k (#modifications)* spectral alignment reveals more and more subtle similarities between the spectrum and the peptide.

MS-Alignment found a number of spectra with 3 modifications that are rarely reported in the literature

Edit Graph for Fast Spectral Alignment



diag(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.

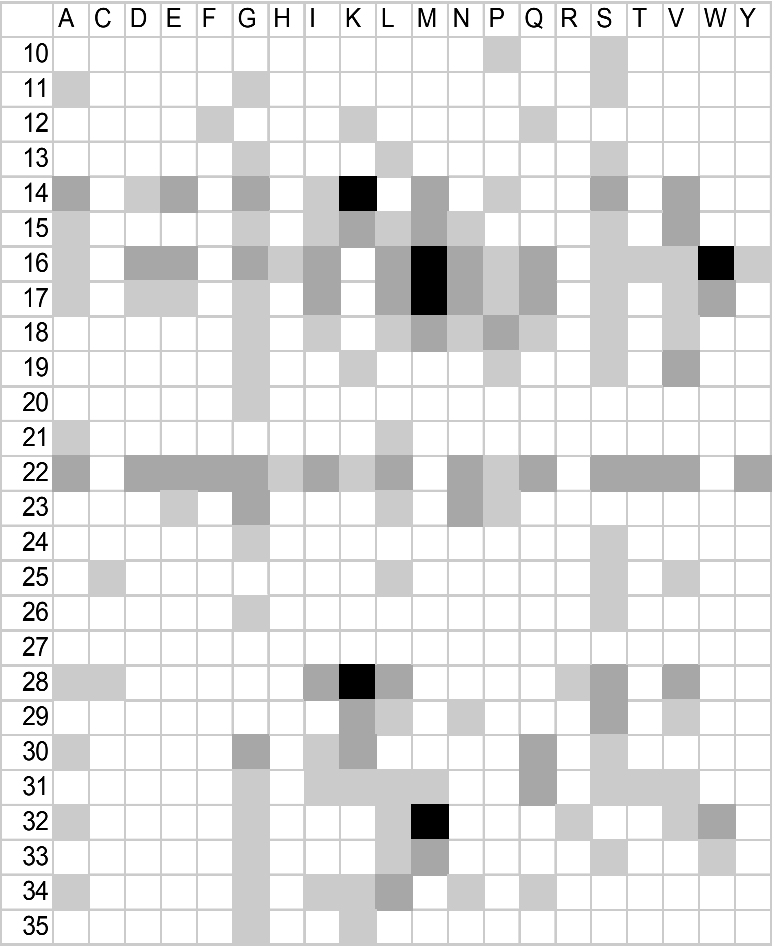
These series form two diagonals in the spectral product, the main diagonal and the perpendicular diagonal.

The described algorithm deals with the main diagonal only.

Spectral Alignment: Complications

- Simultaneous analysis of N- and C-terminal ions
- Taking into account the intensities and charges
- Analysis of minor ions

PTM Frequency Matrix



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
10													░			░				
11	░					░										░				
12					░				░					░						
13						░				░						░				
14	▓		░	▓		▓		░	█		▓		░			▓		▓		
15	░					░		░	▓	░	▓	░				░		▓		
16	░		▓	▓		▓	░	▓		▓	█	▓	░	▓		░	░	░	█	░
17	░		░	░		░		▓		▓	█	▓	░	▓		░		░	▓	
18						░		░		░	▓	░	▓	░		░		░		
19						░			░				░			░		▓		
20						░														
21	░									░										
22	▓		▓	▓	▓	▓	░	▓	░	▓		▓	░	▓		▓	▓	▓		▓
23				░		▓				░		▓	░							
24						░										░				
25		░								░						░		░		
26						░										░				
27																				
28	░	░						▓	█	▓					░	▓		▓		
29									▓	░		░				▓		░		
30	░					▓		░	▓					▓		░				
31						░		░	░	░	░			▓		░	░	░		
32	░					░				░	█				░			░	▓	
33						░				░	▓					░			░	
34	░					░		░	░	▓		░		░						
35						░			░											

50,000 spectra (IKKb sample) were searched in blind mode, and identifications with p-value <0.05 were retained

Shading of the cell *(x,y)* reflects the number of annotations with modification:

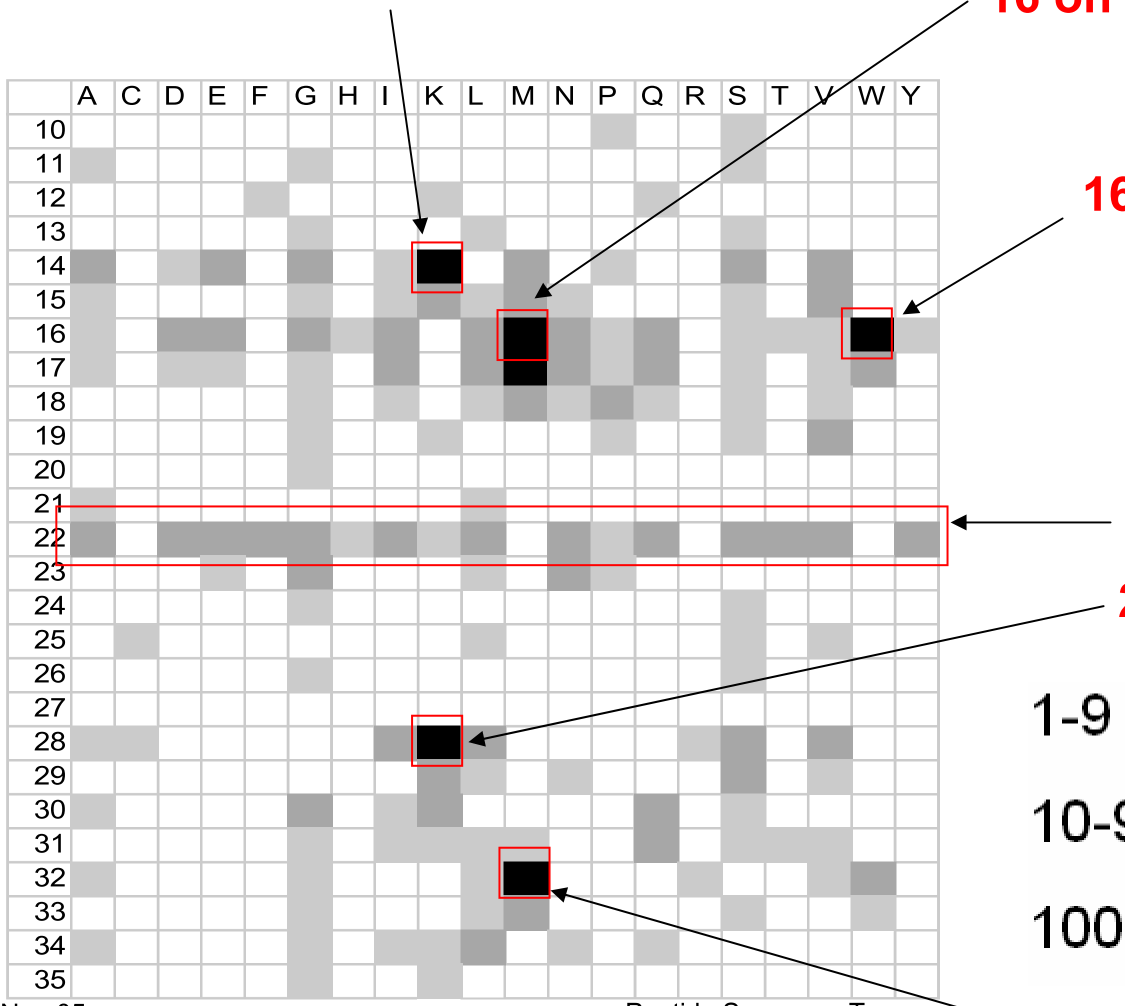
(offset x, amino acid y)

1-9 light gray ░

10-99 dark gray ▓

100+ black █

PTM Frequency Matrix



14 on K - Methylation

16 on M - Oxidation

16 on W - Oxidation

22 - Sodium

28 on K - Dimethylation

32 on M - Double oxidation

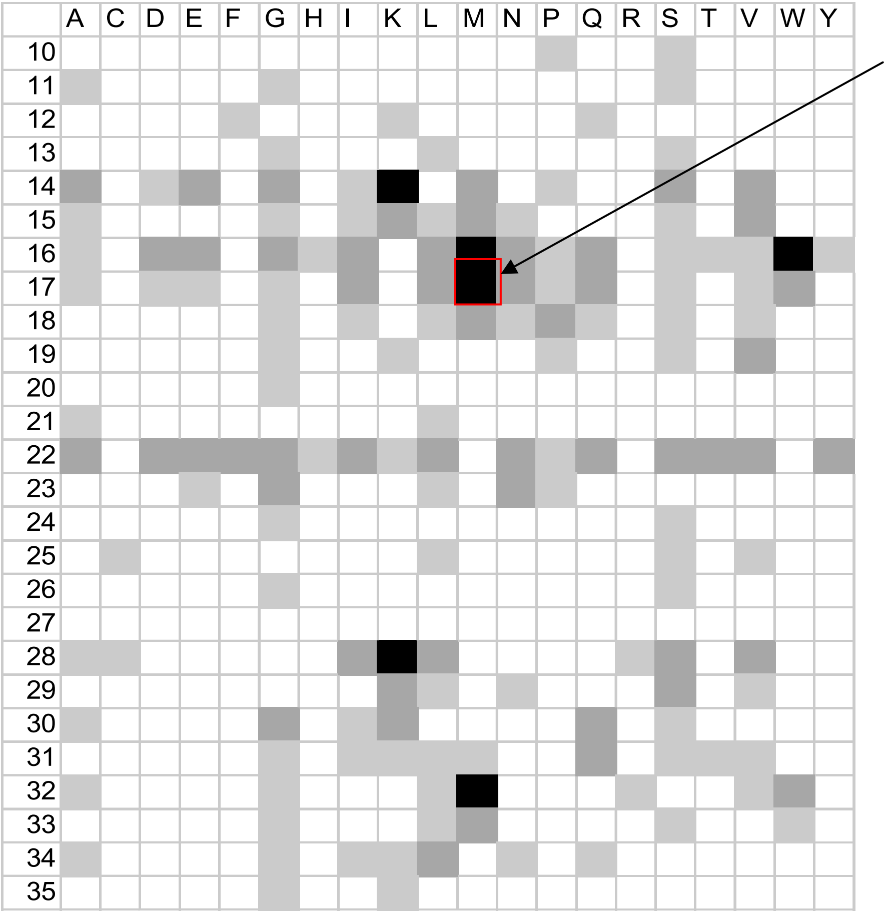
1-9 light gray ░

10-99 dark gray ▒

100+ black ■

PP1

Shadows in PTM Frequency Matrix



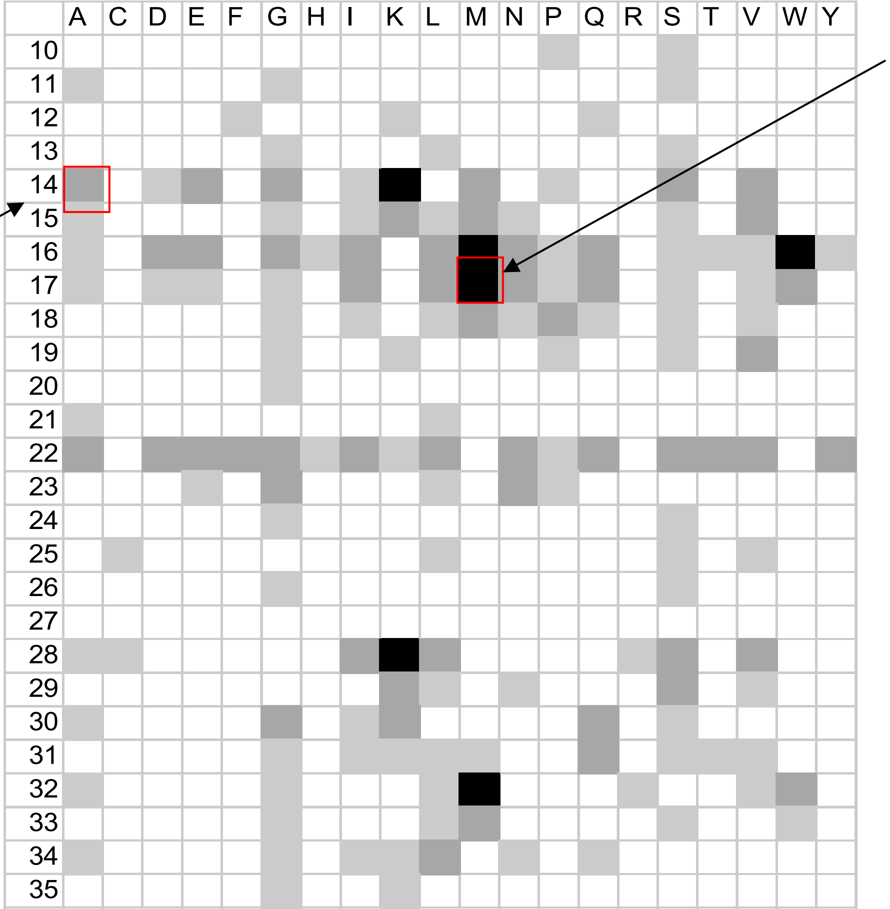
17 on M – ???
oxidation with shift off by 1 (possible error in parent mass and/or wrong assignments of isotopic peaks)

Slide 91

PP1 Pavel Pevzner, 04-Sep-05

PP2

Shadows in PTM Frequency Matrix



14 on A ??? incorrectly placed methylation (A instead of closely located M)

17 on M – ??? oxidation with offset off by 1 (possible error in parent mass and/or misassignment of isotopic peaks

Slide 92

PP2 Pavel Pevzner, 04-Sep-05

Removing Shadows

- Annotation is Δ-correct if it correctly predicts the offset but places it incorrectly on one of the neighboring amino acids (happens if fragmentation near the PTM site is poor).
- Shadows are removed by dealing with Δ-correct annotations in such a way that they are ‘explained away’ by the most frequent PTM

PTM selection: Output

a	Δ	Spectra
M,W	16	803
non-specific	1	355
C	71	332
M,W	32	248
N	1	225
K	28	184
non-specific	22	176
K,M	14	154
E,D,P	53	130
T,E,D	-18	117
L	156	92
V	28	56
I	16	49
K	-57	46
S	28	30
L	17	27
M,W	38	23
C	76	22
non-specific	2	22
M	-2	21
I	44	20
L	54	19

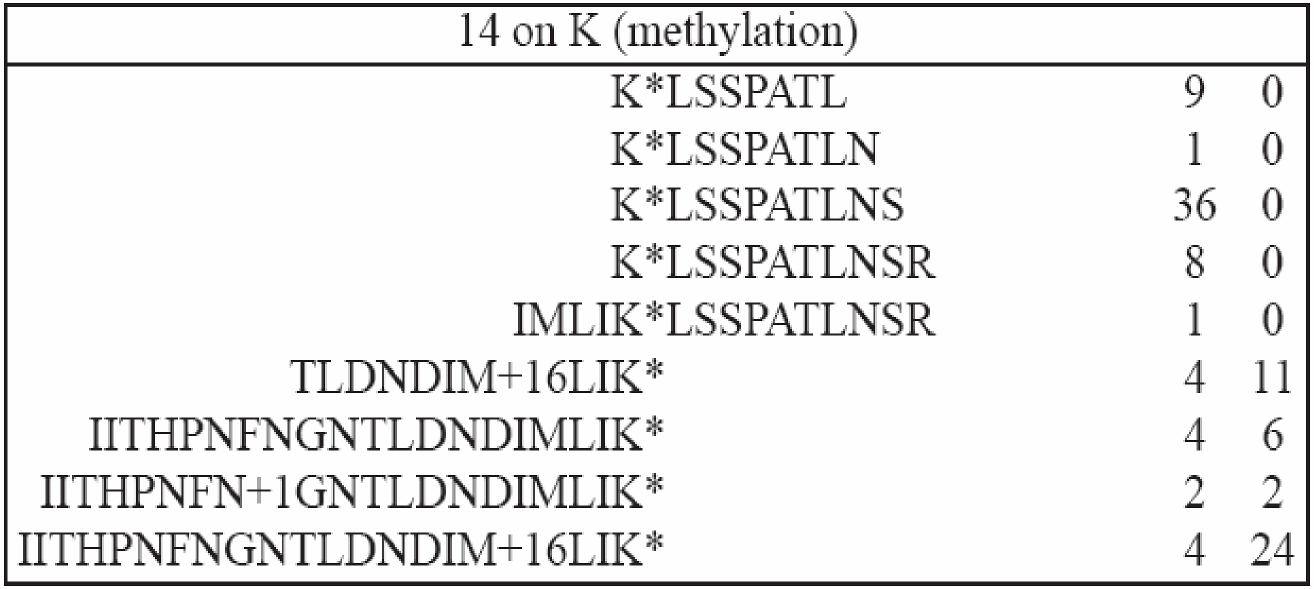
PTM selection: Curated

a	Δ	Spectra	Putative annotation
M,W	16	803	oxidation
non-specific	1	355	isotopic peaks
C	71	332	PAM-cys
M,W	32	248	double oxidation
N	1	225	deamidation
K	28	184	dimethylation
non-specific	22	176	sodium
K,M	14	154	methylation
non-specific	53	130	Fe(III) adduct
T,E,D	-18	117	dehydration
L	156	92	Truncated K+28L
V	28	56	dimethylation
I	16	49	misplaced oxidation
K	-57	46	mutation to alanine
S	28	30	mutation to aspartate
L	17	27	misplaced oxidation
M,W	38	23	potassium
C	76	22	beta-mercaptoethanol
non-specific	2	22	isotopic peaks
M	-2	21	mutation to glutamate
I	44	20	misplaced K+28,M+16
L	54	19	shadow of +53

PTM selection: Curated

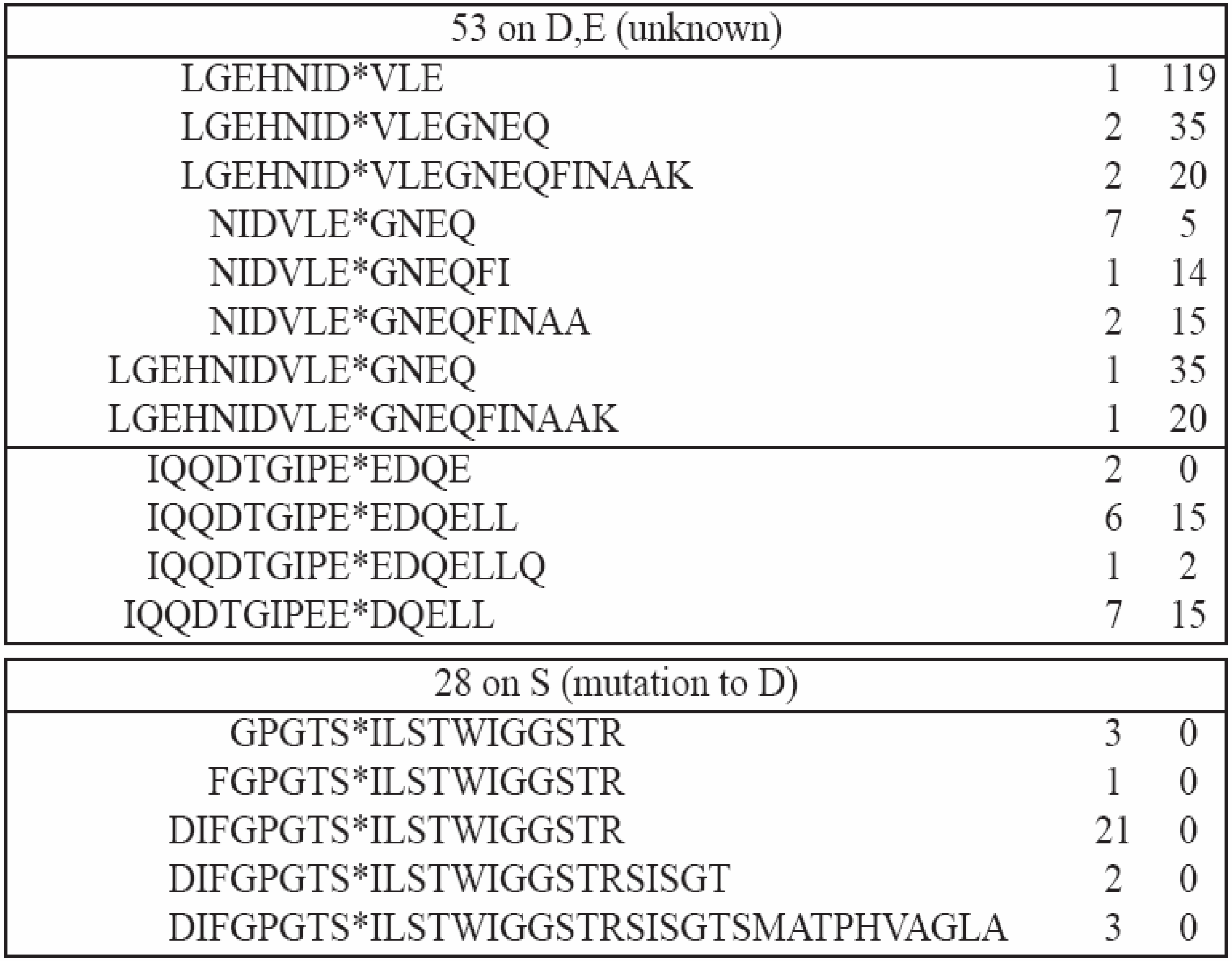
a	Δ	Spectra	Putative annotation
M,W	16	803	oxidation
non-specific	1	355	isotopic peaks
C	71	332	PAM-cys
M,W	32	248	double oxidation
N	1	225	deamidation
K	28	184	dimethylation
non-specific	22	176	sodium
K,M	14	154	methylation
non-specific	53	130	**Fe(III) adduct**
T,E,D	-18	117	dehydration
L	156	92	Truncated K+28L
V	28	56	dimethylation
I	16	49	misplaced oxidation
K	-57	46	mutation to alanine
S	28	30	mutation to aspartate
L	17	27	misplaced oxidation
M,W	38	23	potassium
C	76	22	beta-mercaptoethanol
non-specific	2	22	isotopic peaks
M	-2	21	mutation to glutamate
I	44	20	misplaced K+28,M+16
L	54	19	shadow of +53

Overlapping peptides



14 on K (methylation)		
K*LSSPATL	9	0
K*LSSPATLN	1	0
K*LSSPATLNS	36	0
K*LSSPATLNSR	8	0
IMLIK*LSSPATLNSR	1	0
TLDNDIM+16LIK*	4	11
IITHPNFNGNTLDNDIMLIK*	4	6
IITHPNFN+1GNTLDNDIMLIK*	2	2
IITHPNFNGNTLDNDIM+16LIK*	4	24

Overlapping peptides



53 on D,E (unknown)		
LGEHNID*VLE	1	119
LGEHNID*VLEGNEQ	2	35
LGEHNID*VLEGNEQFINAAK	2	20
NIDVLE*GNEQ	7	5
NIDVLE*GNEQFI	1	14
NIDVLE*GNEQFINAA	2	15
LGEHNIDVLE*GNEQ	1	35
LGEHNIDVLE*GNEQFINAAK	1	20
IQQDTGIPE*EDQE	2	0
IQQDTGIPE*EDQELL	6	15
IQQDTGIPE*EDQELLQ	1	2
IQQDTGIPEE*DQELL	7	15

28 on S (mutation to D)		
GPGTS*ILSTWIGGSTR	3	0
FGPGTS*ILSTWIGGSTR	1	0
DIFGPGTS*ILSTWIGGSTR	21	0
DIFGPGTS*ILSTWIGGSTRSISGT	2	0
DIFGPGTS*ILSTWIGGSTRSISGTSMATPHVAGLA	3	0

MS-Alignment Test Case

	1 PTM	2 PTMs
Correct	57%	16%
Δ-correct	36%	67%
Incorrect	7%	17%

Spectra from the ISB data-set were searched against a database mutated to 90% identity.

A match which reverses the mutation(s), recovering the original sequence exactly is **correct**

A match to the correct locus with incorrect modification(s) is **Δ-correct**.

Selecting modification sites

- A ‘strength in numbers’ approach: The more spectra, the better
- Overlapping peptides are strong evidence (incorrect matches unlikely to overlap)
- Overlapping peptides help pinpoint the modification site (tricky for modifications near the edge of a peptide)
- We like to see ‘rungs’ of the b and y ladders on either side of the modified residue

Blind PTM Search in Lens Proteins

- Mass spectra derived from cataractous lens proteins
- Some data is from the Larry David lab (93 year old patient), the other is from the John Yates lab (early onset cataract from a few children)
- Both data-sets were searched in blind mode against a database of human lens protein

PTMs in Lens Proteins: Validation

- MS-Alignment produced the largest set of PTMs ever reported in lens
- All spectra with found modifications were manually validated in Larry David’s lab using stringent criteria
- Manual validations were performed independently by Phil Wilmarth and Surendra Dasari and only spectra that passed both validation tests were accepted
- Many previously unknown modification sites were found:

Wilmarth, Dasari, Tanner, Bafna, Pevzner, David. *Identification of carboxymethyl modified lysine residues in aged cataractous human lens (in preparation)*

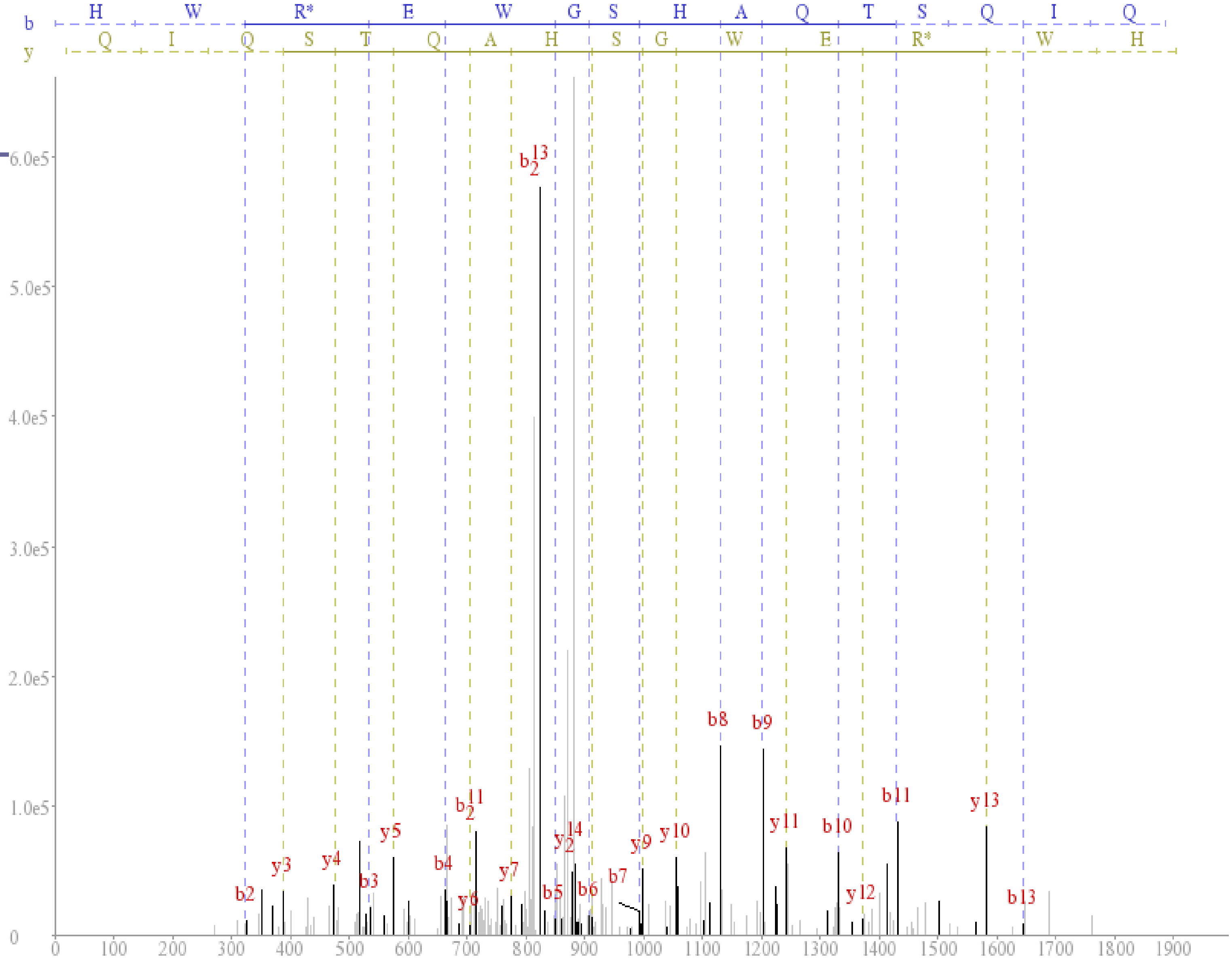
Lens: Three Unknown Modifications

- Three found modifications (R+55, K+58, and K+72) are not present in ABRF database.
- They are confirmed by multiple overlapping peptides and manually validated by both Larry David’s postdocs (Phil Wilmarth and Surendra Dasari) and Kati Medzihiradszky at UCSF

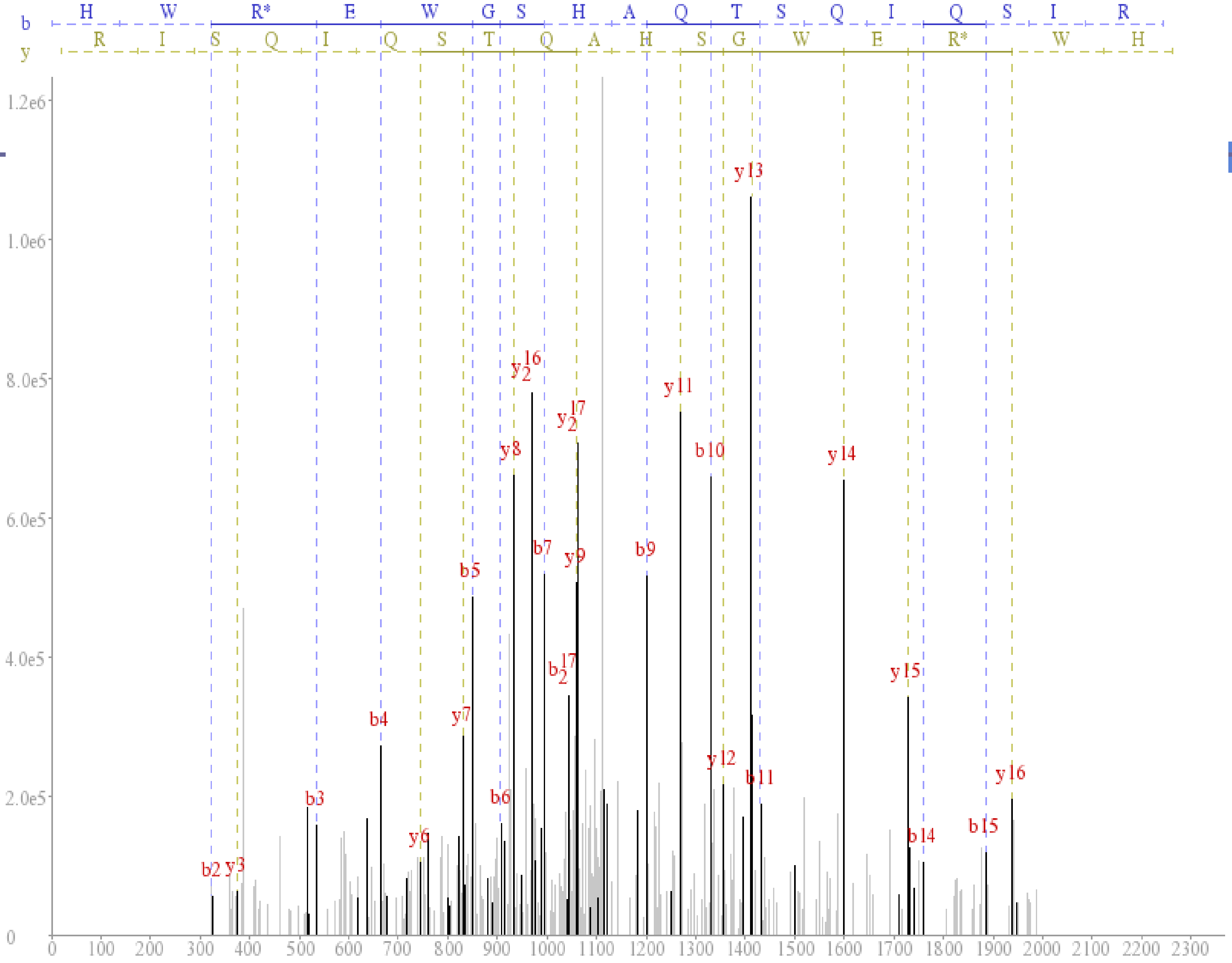
Lens: Three Unknown Modifications

- Three found modifications (R+55, K+58, and K+72) are not present in ABRF database.
- They are confirmed by multiple overlapping peptides and manually validated by both Larry David's postdocs and Kati Medzihiradszky at UCSF
- It turned out that K+58 was discovered before (but is not present in ABRF yet). Moreover, recently it was reported in a lens protein (Crabb et al., PNAS, 2002)!

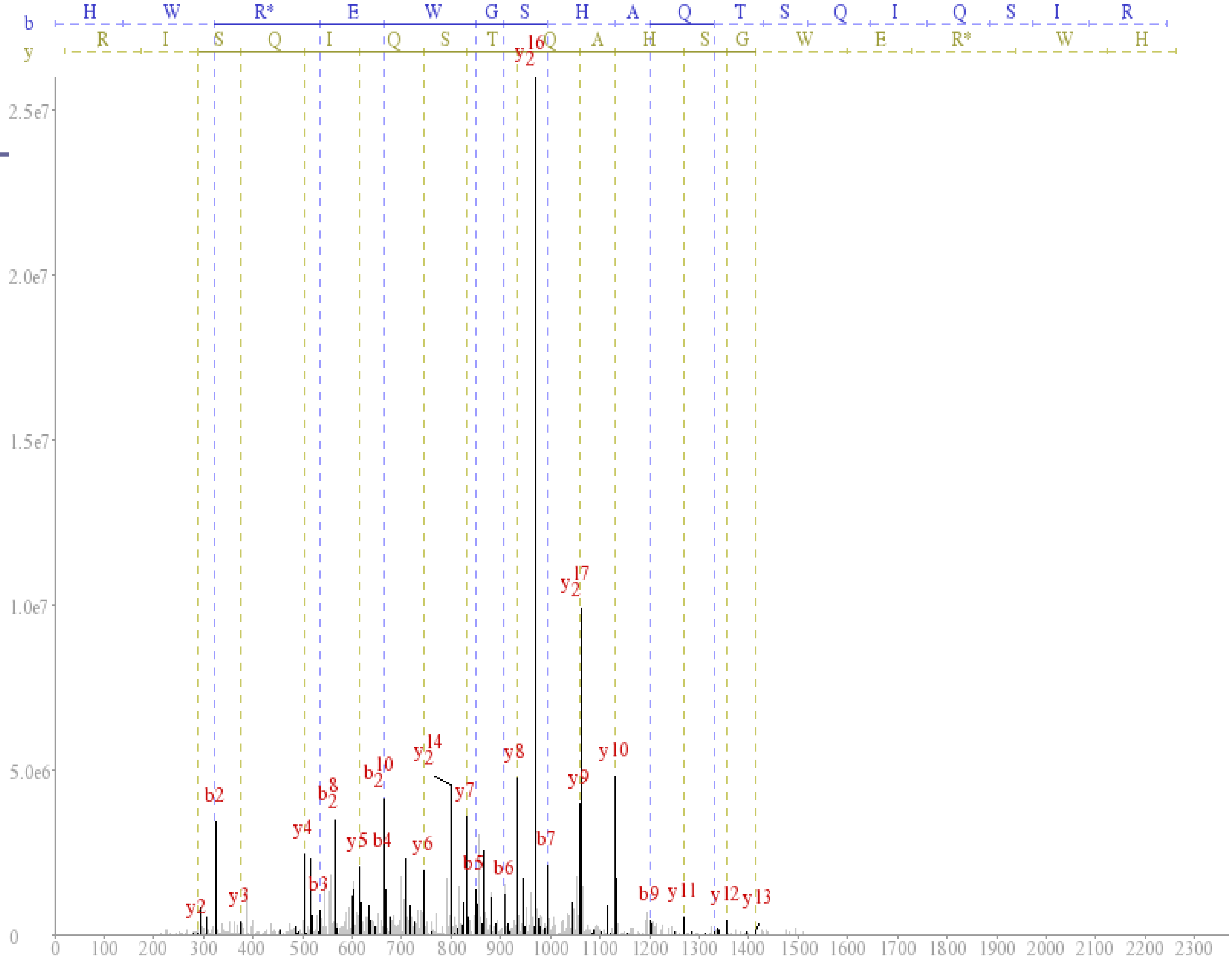
Spectra for putative R+55 modification



Spectra for putative R+55 modification



Spectra for putative R+55 modification



Lens: Common modifications

Many known modifications were found in David’s and Jates’ data-sets on the same residues.

- Phosphorylation (S+80,T+80)
- Cysteine methylation (C+14)
- Methionine oxidation (M+16)
- Carbamylation (K+43, N-termini +43)
- Deamidation (Q+1, N+1, -17 if N-terminal)

Lens: Differences

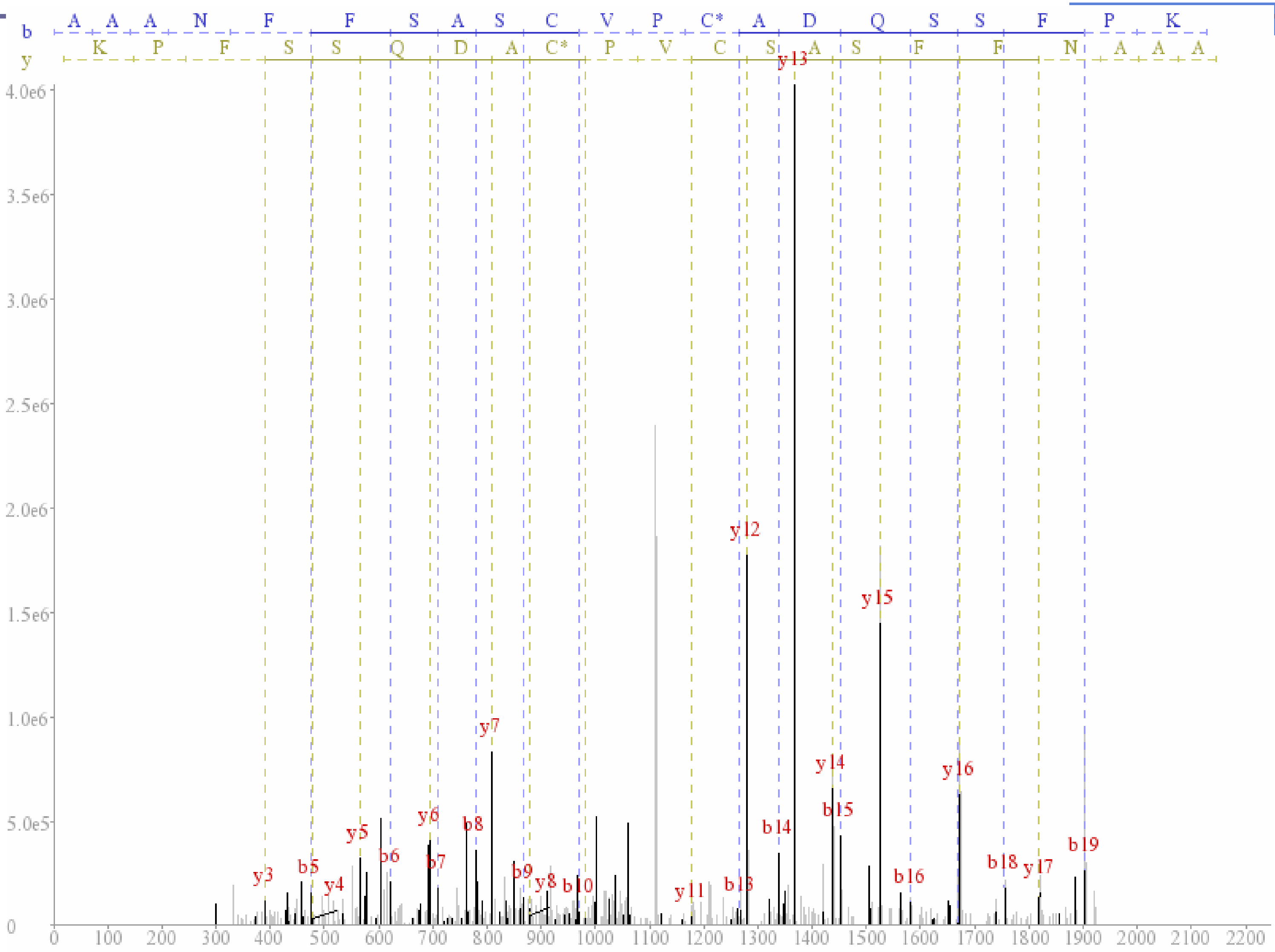
David data only:

- Potassium (+38)
- N-terminal acetylation (+42)
- Putative formylation (S+28)

Yates data only:

- Sodium (+22)
- CAM on Histidine, N-termini (+57)
- Lysinoalanine (C-34)
- Decomposed oxidized methionine (M-48)
- Putative deamidated CAM (+40 on N-terminus)

ISB Dataset: Disulfide bridges



TRFE_BOVIN, from ISB data-set (**modification -2 on C**)

Yet Another Problem

- **MS/MS database search … without ever comparing a spectrum against a database.**

  Popular database search tools (Sequest/Mascot) interpret spectra by comparing every spectrum with a database

  New database search tools (X!Tandem/InsPecT) interpret spectra by comparing every spectrum with a (somewhat smaller) database

Yet Another Problem

- **MS/MS database search … without ever comparing a spectrum against a database.**

  Popular database search tools (Sequest/Mascot) interpret spectra by comparing every spectrum with a database

  New database search tools (X!Tandem/InsPecT) interpret spectra by comparing every spectrum with a (somewhat smaller) database

  Can you interpret 1 million spectra without ever comparing a single spectrum against a peptide?

Related work

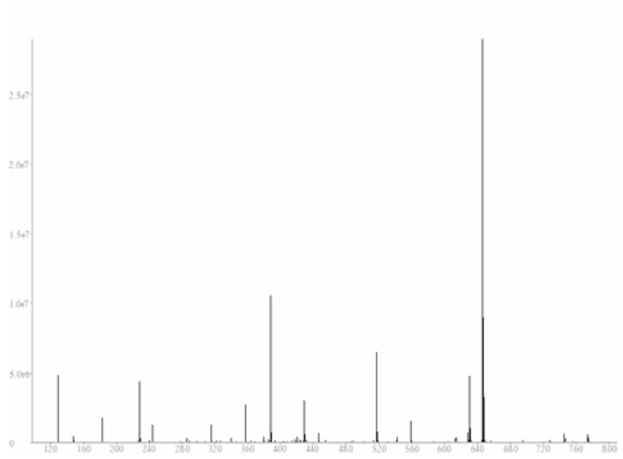
- OpenSea (Searle, 2004) and SPIDER (Han, 2004) search for unanticipated modifications
- Both tools require a starting *de novo* interpretation

Related work

- OpenSea (Searle, 2004) and SPIDER (Han, 2004) search for unanticipated modifications
- Both tools require a starting *de novo* interpretation
- In practice, such reconstruction is prone to errors, particularly around modifications

VKEAMAPK →

??? ◂··············



References

- For more information on our algorithms see:
  - Frank A., Pevzner P. "*PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling*", Analytical Chemistry, 77 : 964-973, 2005.
  - Tanner S., et al. “*Inspect: identification of post-translationally modified peptides from tandem mass spectra”*. Analytical Chemistry,77 : 4626-4639, 2005.
  - A journal version of this paper:
    Frank A. et al. “Peptide Sequence Tags for Fast Database Search in Mass-Spectrometry”, Journal of Proteome Research (ASAP articles).
- PepNovo and InsPecT can be run on a web-server at : http://peptide.ucsd.edu

Sequencing With Unknown Genomes

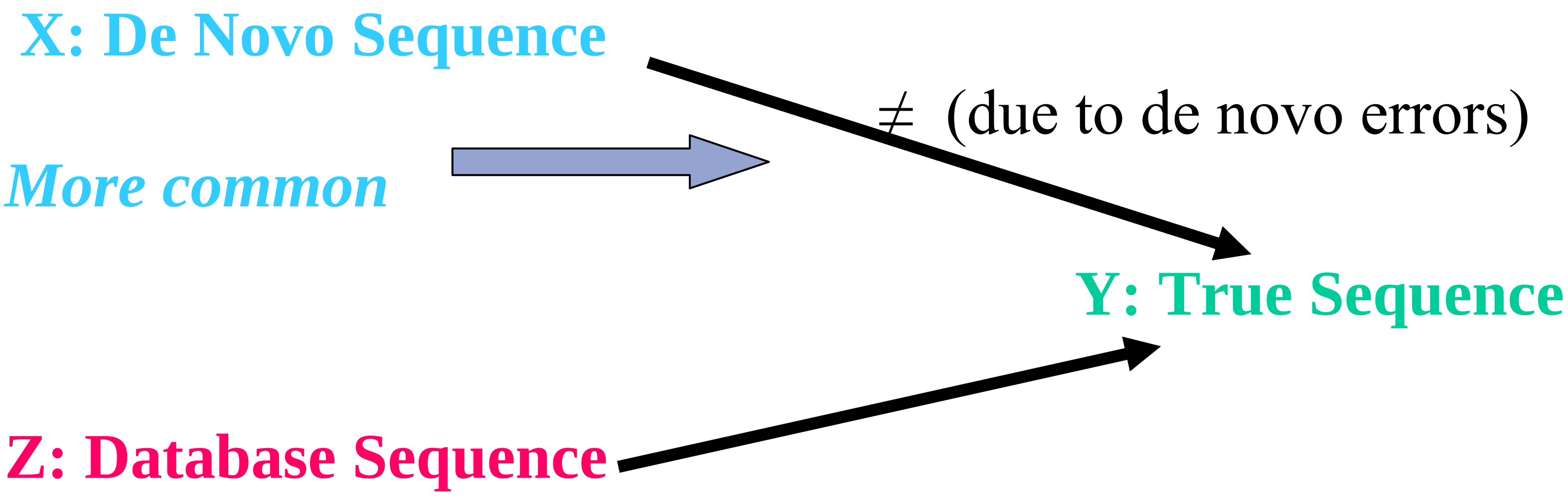
- Database search relies on having genomic sequences.
- However, many organisms have not been sequenced yet.
- How can we identify proteins from their proteome?

- Identification can be done with Homology based search using de novo as a seed.

Homology Based Search Algs.

- OpenSea [Searle et al. 2004]
- Spider [Han et al. 2004]

Key Idea of SPIDER:



X: De Novo Sequence

≠ (due to de novo errors)

More common

Y: True Sequence

Z: Database Sequence

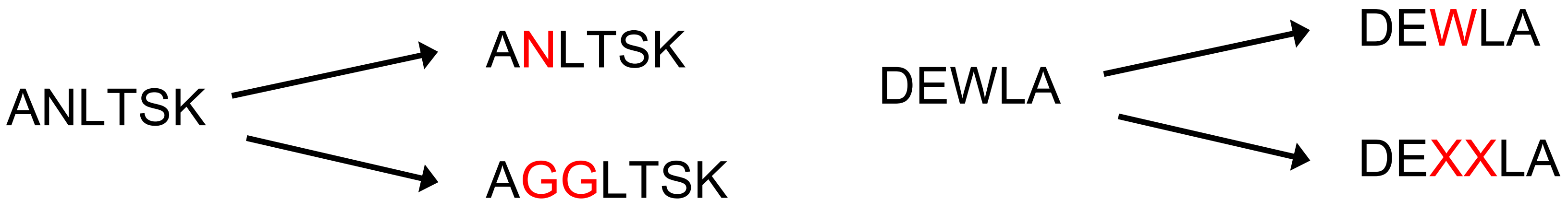
MS-BLAST [Shevchenko et al. ’01]

- Uses De Novo results to perform ungapped BLAST similarity searches.
  - Identifies proteins rather than peptides.
  - Has established statistical methods to measure significance.
  - Uses biologically driven Matrix to score mutations (Blosum / PAM).

− Does not handle the common de novo errors well.

Additional De Novo Candidates

- MS-Blast can accept several variants of the same sequence, and only choose the best scoring match.
- Common de novo errors can be accounted for by creating redundant sequences:
- Replacing problematic amino acids: N, W, Q



ANLTSK

ANLTSK

AGGLTSK

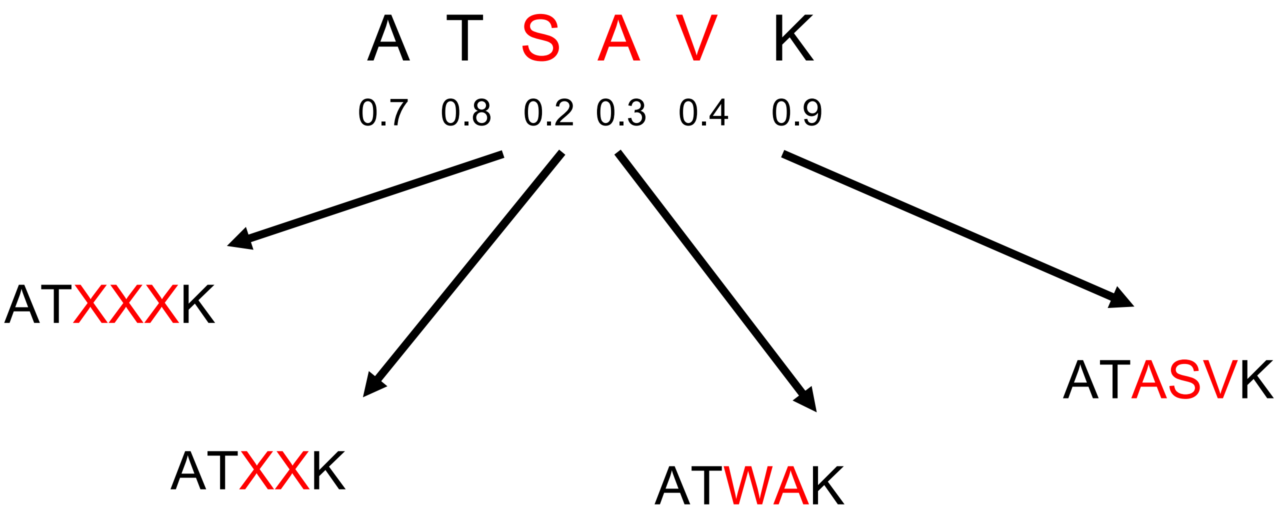
DEWLA

DEWLA

DEXXLA

Candidate Generation Cont.

- Replace low probability amino acids with alternative sequences or gaps.



Candidate Generation Cont.

- The candidates can be modified several times, until a sufficiently large and high scoring set is obtained.

- This method has been applied successfully to samples from the Dead Sea alga *Dunaliella salina* [Waridel et al. , to appear in HUPO 2005].

Collaborators

- UCSD Computational Mass Spectrometry Group (**Vineet Bafna** and P.P labs):
  - **Nuno Bandeira** (de novo sequencing of entire proteins)
  - **Ari Frank** (PepNovo, PepNovoTag)
  - **Stephen Tanner** (InsPecT, MS-Alignment)
  - **Dekel Tsur** (MS-Alignment)
- **Larry David, Phil Wilmarth, Surendra Dasari, OHSU** (lens proteins)
- **John Yates, Scripps** (lens proteins)
- **Andrey Shevchenko, Max Planck Institute** (using PepNovo in MS-BLAST)
- **Marc Mumby, Southwestern Medical School** (phosphoproteins)
- **Ebi Zandi, Tim Chen, USC** (IKKb)
- **Karl Clauser, Broad** (assembly of snake venom proteins)
- **Kati Medzihiradszky, UCSF** (new PTM types)

Support: 5R01RR016522 NIH (NCRR)