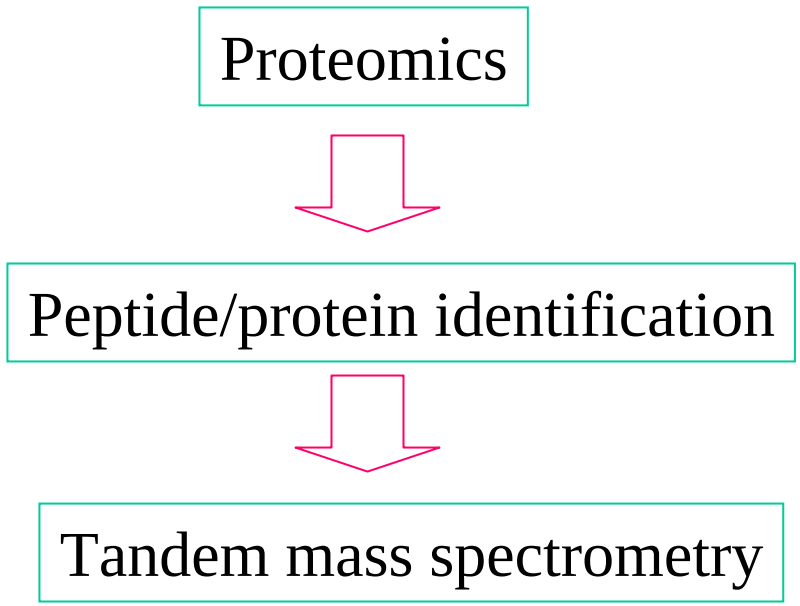
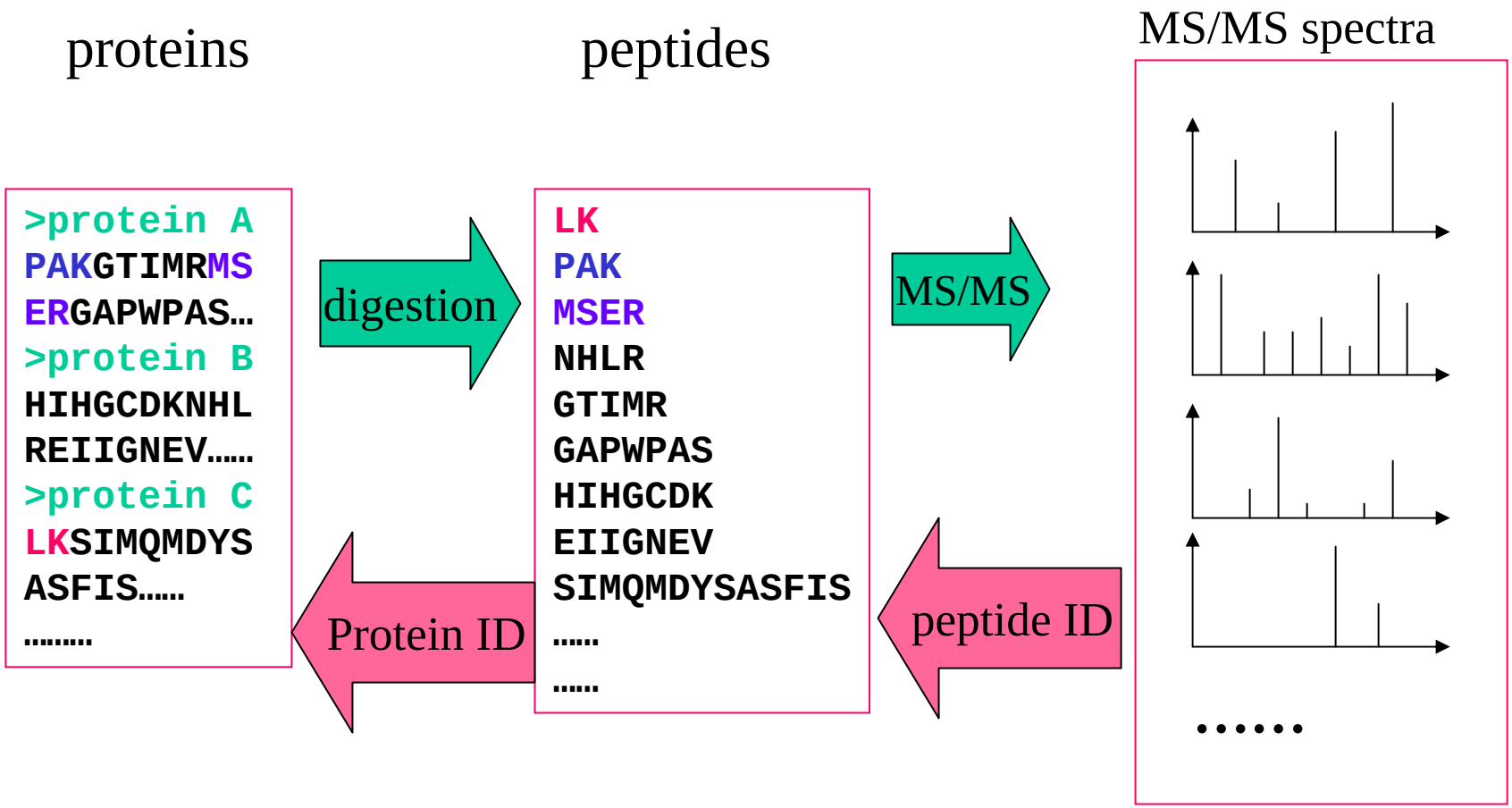
Algorithms and Score Functions Used in PEAKS *De Novo* Sequencing Software

Bin Ma
Dept. of Computer Science
University of Western Ontario
London, Ontario, Canada

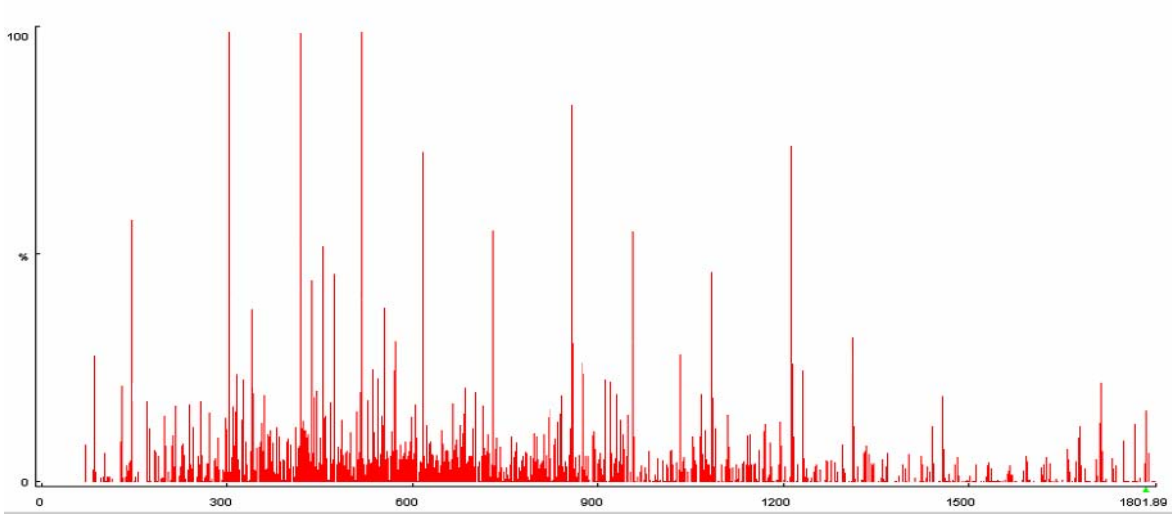
Background


Proteomics
Peptide/protein identification
Tandem mass spectrometry

Procedure


proteins
peptides
MS/MS spectra
>protein A
PAKGTIMRMS
ERGAPWPAS…
>protein B
HIHGCDKNHL
REIIGNEV……
>protein C
LKSIMQMDYS
ASFIS……
……
digestion
LK
PAK
MSER
NHLR
GTIMR
GAPWPAS
HIHGCDK
EIIGNEV
SIMQMDYSASFIS
……
……
MS/MS
Protein ID
peptide ID
……

MS/MS spectrum



de novo *sequencing:*

database

LGSSEVEQVQLVVDGVK

PEAKS (Bioinformatics Solutions)
Lutefisk (free software)

Mascot (Matrix Science)
Sequest (Thermo)

Score function! Score function! Score function!

- A score function evaluates how well a sequence fits the spectrum
- An ***ideal*** score function gives the highest score to the real sequence
- Database search can be done
  - by scoring every peptide in the database
- *De novo* sequencing can be done
  - by scoring every possible sequence (brute-force)

Algorithm! Algorithm! Algorithm!

the brute-force algorithm

Sequence length	# possible sequences	Time (1 billion per sec)
10	20^{10}	3 hours
15	20^{15}	1000 years
20	20^{20}	10000000000 years

- Efficient algorithms exist to compute the sequence with highest score without enumerating all sequences.
- For some scoring functions at least

You need not only statisticians, but also computer scientists.

Dilemma! Dilemma! Dilemma!

- Simple score functions
  - efficient algorithms to compute the best-scoring sequence
  - the best-scoring sequence may not be the real sequence.
- Sophisticated score function
  - the best-scoring sequence is more likely the real sequence
  - no efficient algorithms to compute the best-scoring sequence

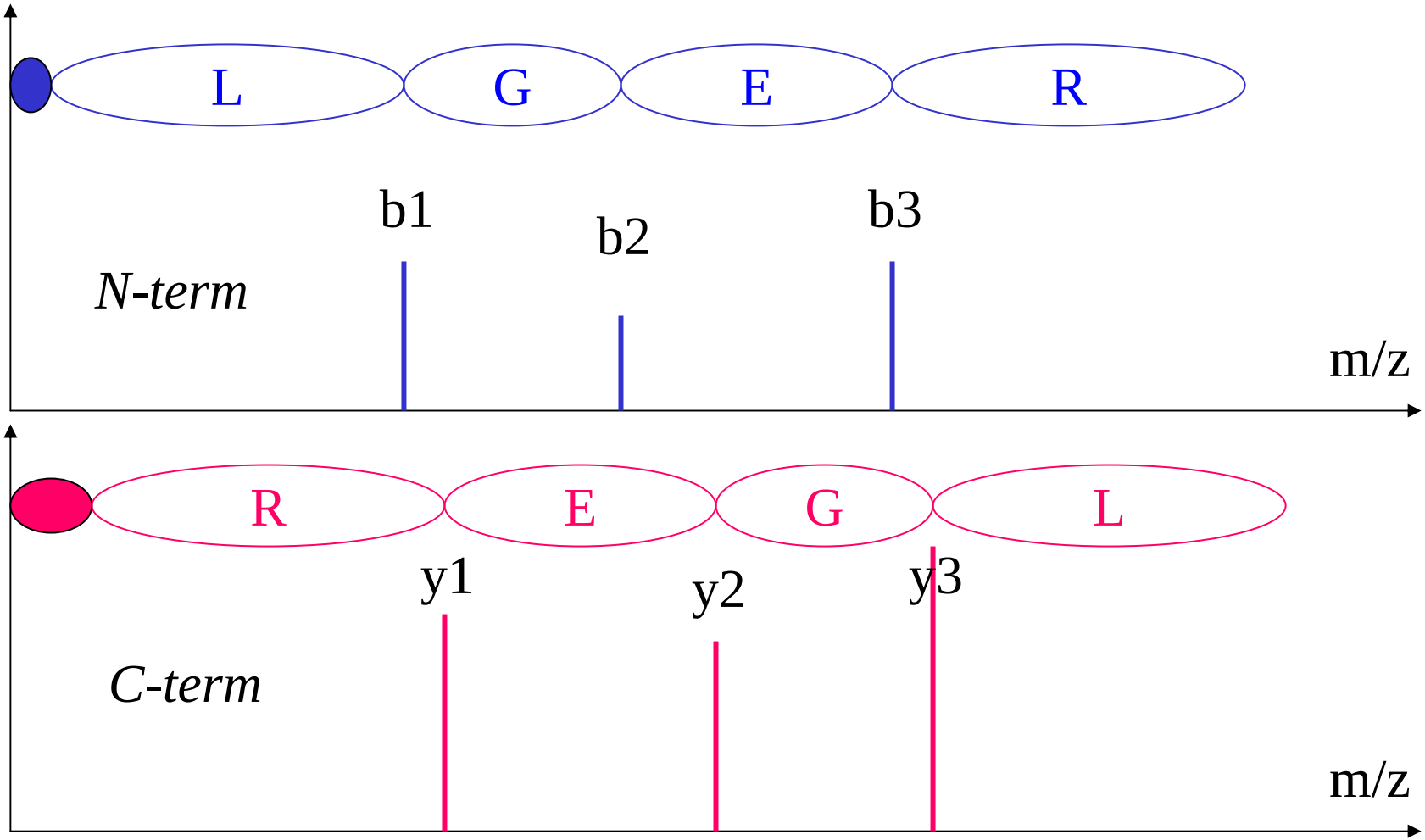
Our Solution

1. Use a simple score function to compute 10000 best-scoring candidate sequences
2. Use a sophisticated score function to evaluate each of the 10000 candidates

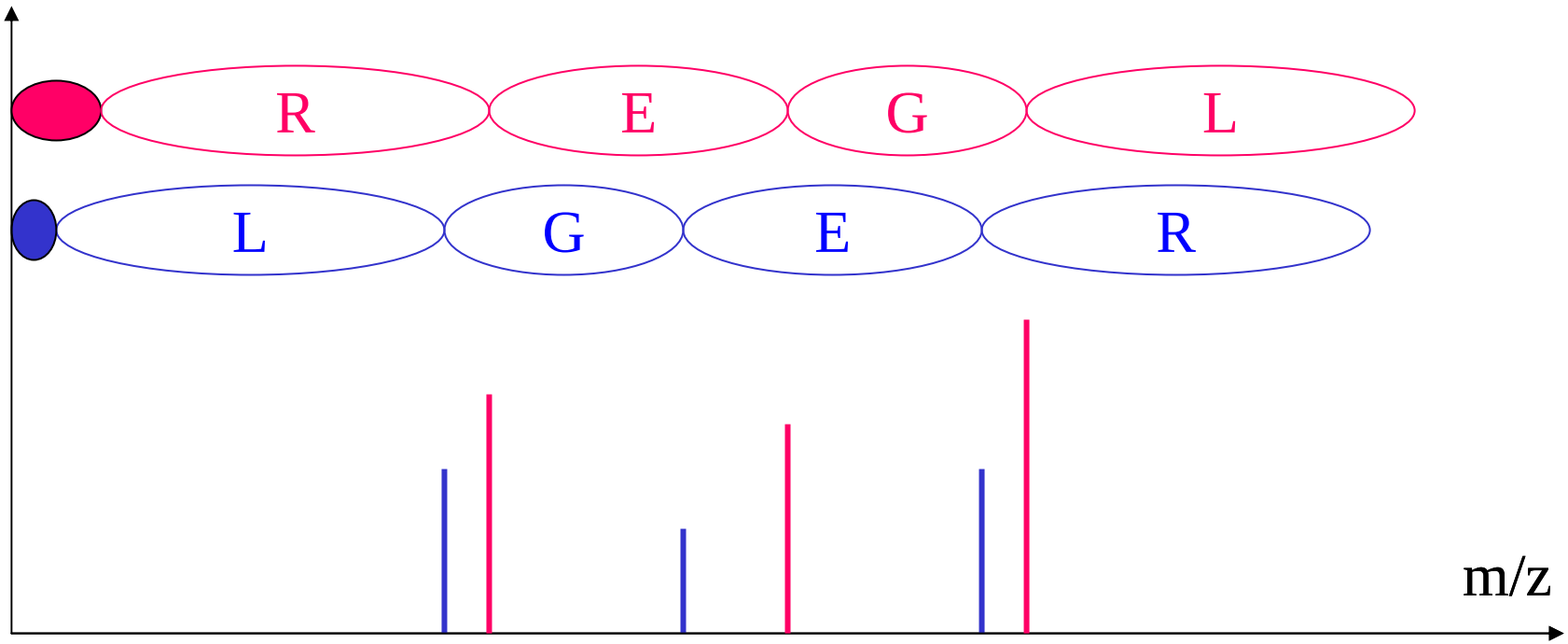
PEAKS algorithm

- A novel algorithm based on dynamic programming.
  - B. Ma, K. Zhang, C. Liang, Journal of Computer and System Sciences. (main algorithm)
  - B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, G. Lajoie, Rapid Comm. Mass Spec. (software feature, score function, experiments)
  - Downloadable from my home page.

How MS/MS corresponds to peptide

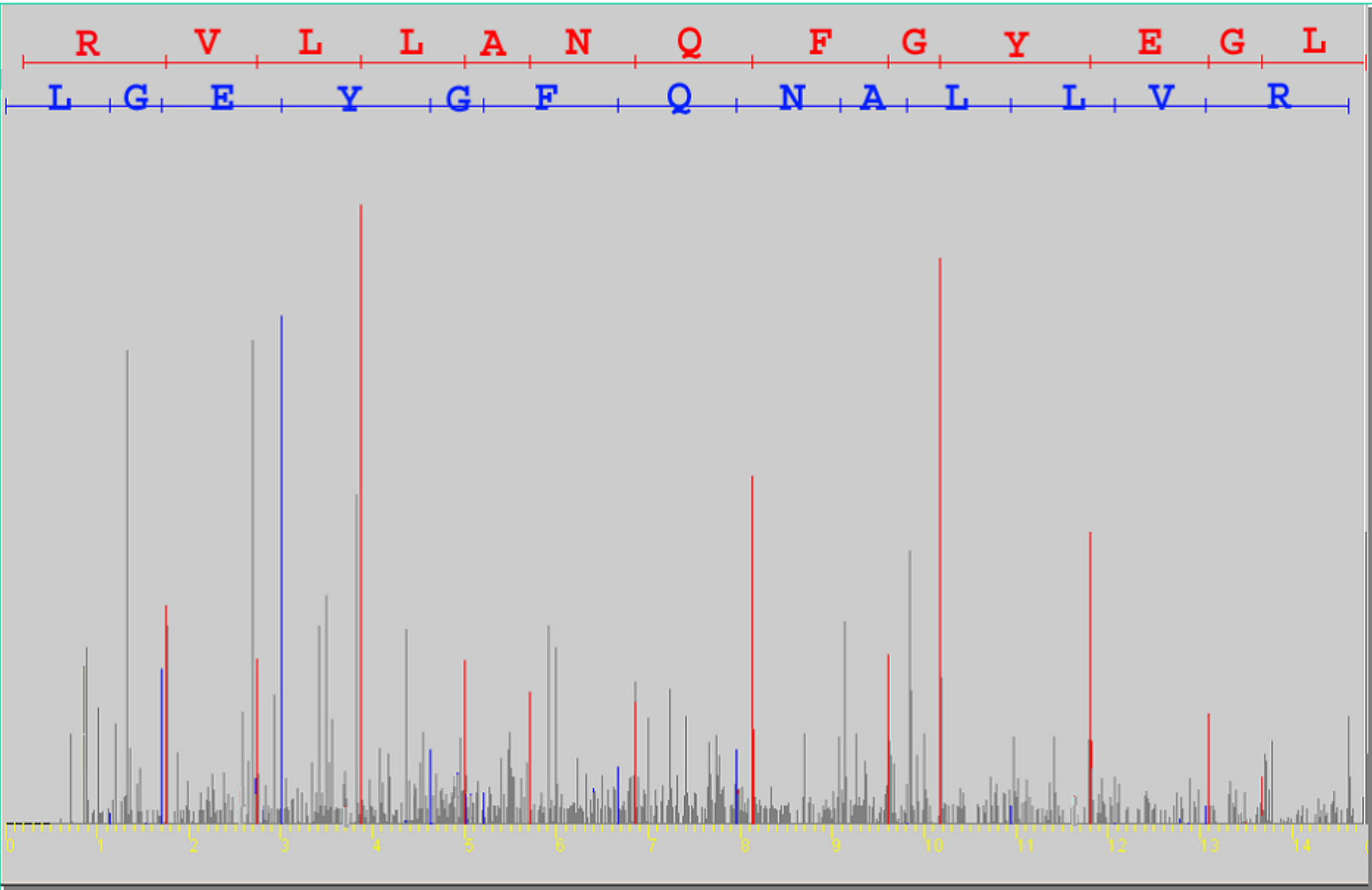


Put both together



In practice, there are many more peaks other than b and y peaks
Many b and y peaks may disappear.

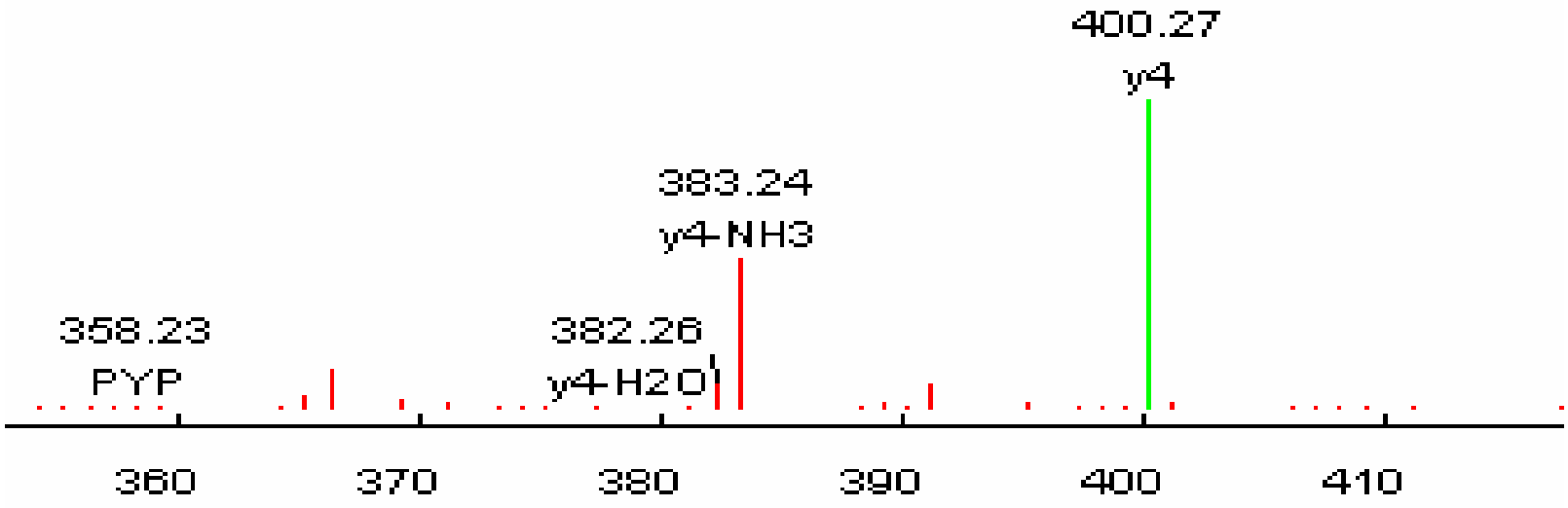
A real peptide and its spectrum



PEAKS' score function basic

- Score of one y-type matching

$$\log(\text{intensity}) \times e^{-(\text{error/tolerance})^2} \times \text{supp (y-17, y-18, x)}$$



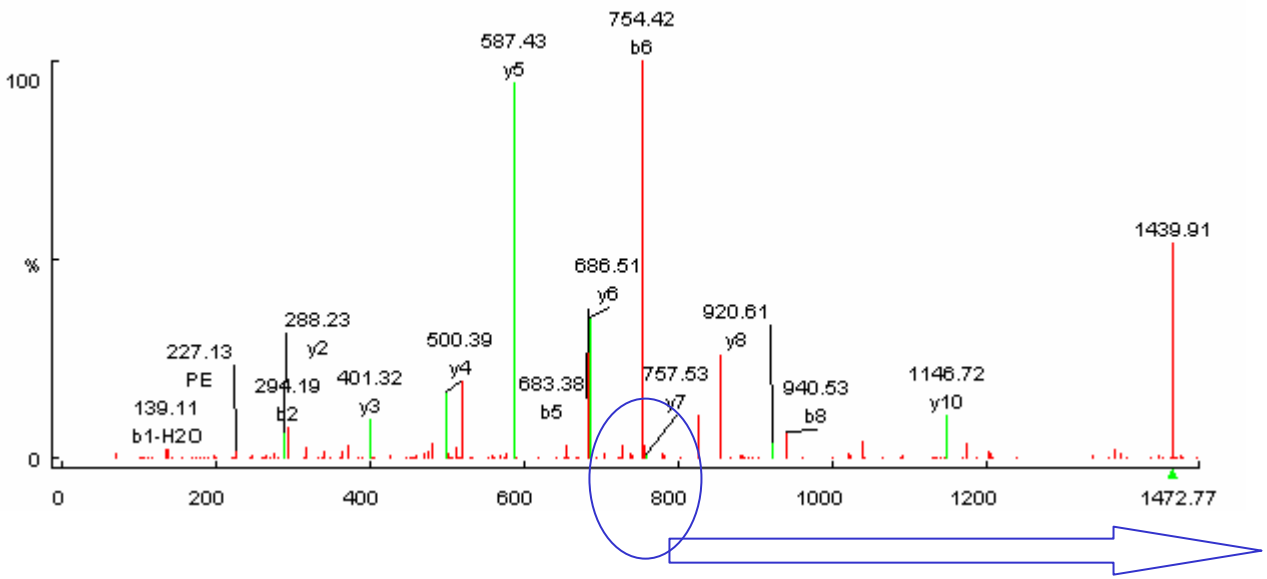
- PEAKS score is equal to the weighted sum of all peaks in consideration.

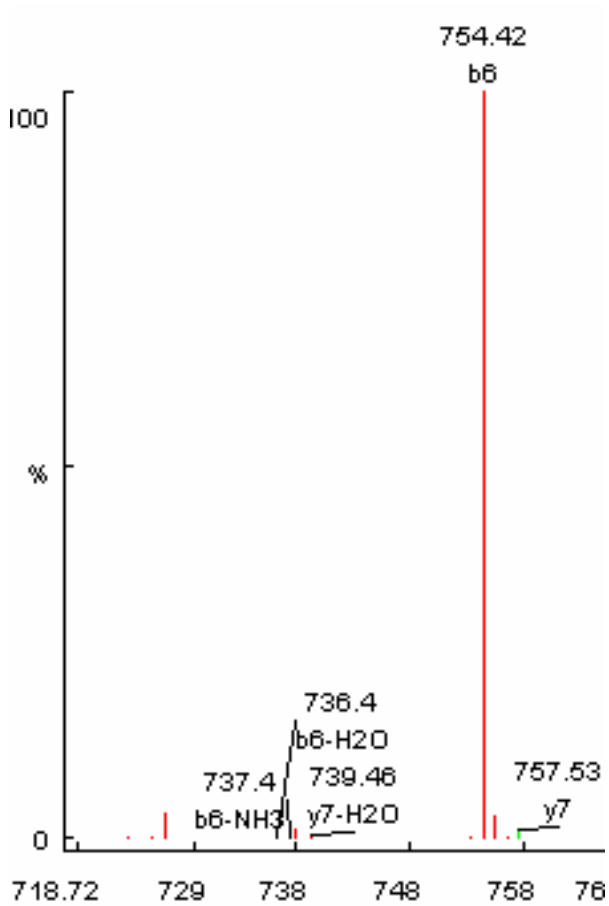
PEAKS' score

- The "simple" score is called raw score, it uses all the "terminal" ions
  - a, b, c, x, y, z ions.
  - y, b losing H2O and NH3 ions
- The "sophisticated" score is called fine score, which uses
  - the terminal ions
  - internal cleavage ions, and losing H2O and NH3 forms
  - immonium ions

Better scoring since PEAKS 2.4

- Fragmentation is usually not perfect.
- Some signal peaks are low
  - But still higher than most peaks.





A Second Independent Score Function

- We propose to use the “rank” of the peak instead of intensity.
- Statistics were done on the rank distribution of different types of ions.
- Scores were given for a y-ion being ranked at 1^{st}, 2^{nd}, 3^{rd}, ……
- The summation of all the y,b, internal,… ions is the total “rank score”.
- PEAKS 2.4 adds this “rank score” with the original score.

Comparison

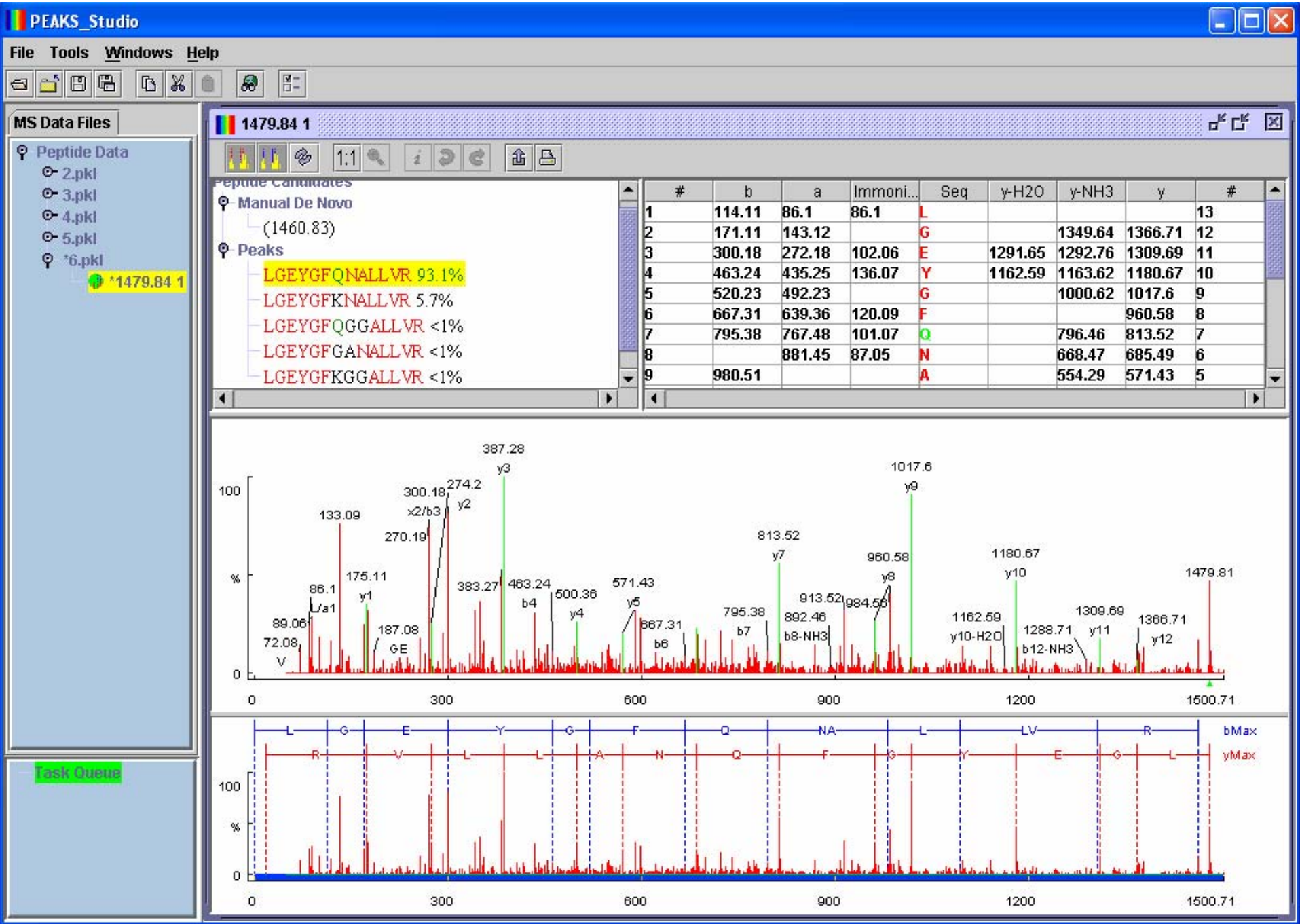
- Testing data: 61 spectra/sequences, total 764 amino acids.

	Typical other software	PEAKS 2.3	PEAKS 2.4
Correct sequence	7	13	**23**
Correct tags of length >= 5	24	38	**50**
Correct AA	233	457	**559**

Absolute score v.s. confidence score

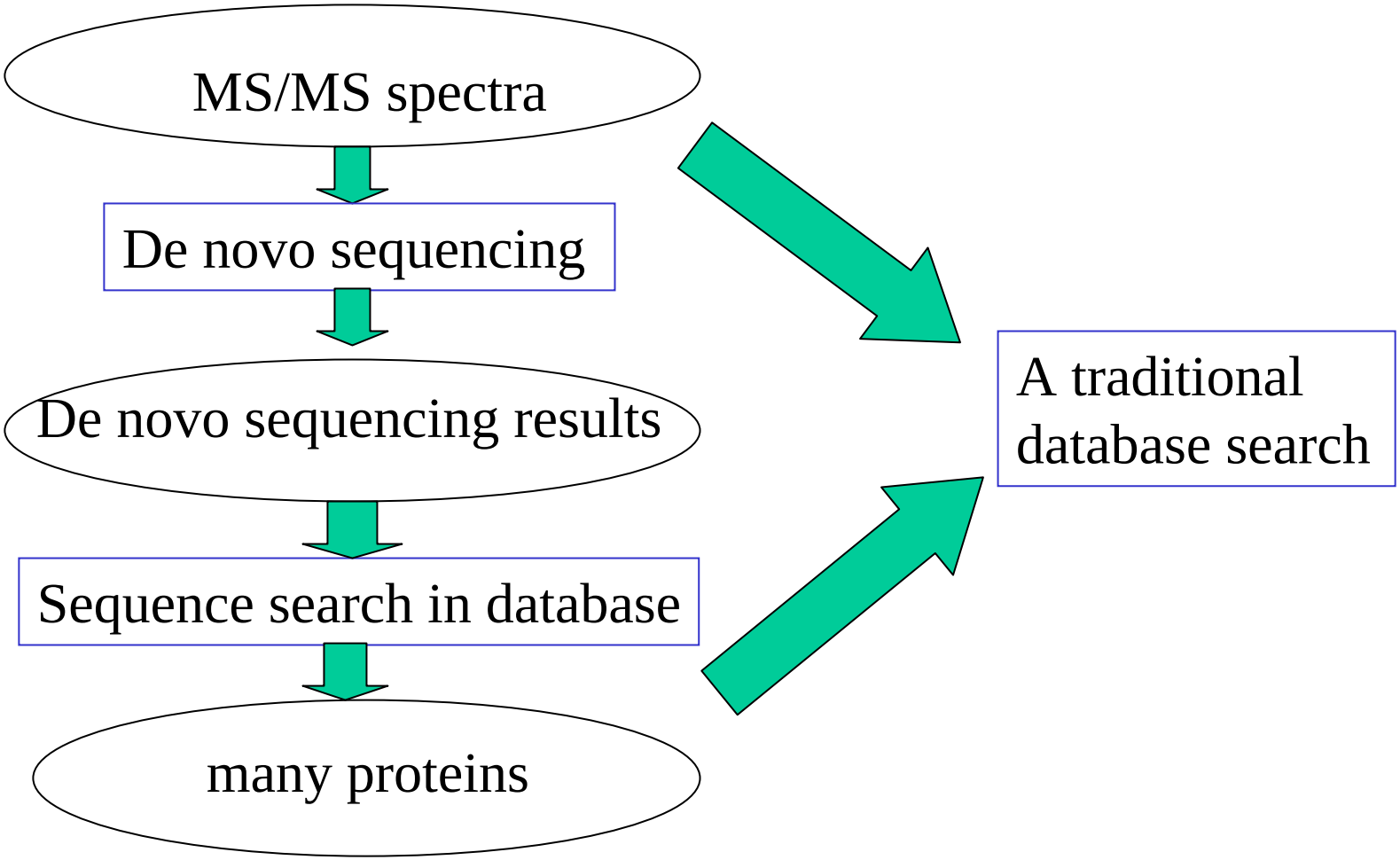
- An absolute score is meaningless to a human.
  - E.g. longer peptides have larger score even if the sequence is incorrect.
- For each spectrum, PEAKS compares the scores of its best 1000 sequences, and estimate a confidence value for each of them.
  - The summation of the confidence is equal to 1.
  - The bigger the difference between the first and the second sequences, the larger the confidence for the first sequence.

PEAKS' positional confidence



#	b	a	Immoni...	Seq	y-H2O	y-NH3	y	#
1	114.11	86.1	86.1	L				13
2	171.11	143.12		G		1349.64	1366.71	12
3	300.18	272.18	102.06	E	1291.65	1292.76	1309.69	11
4	463.24	435.25	136.07	Y	1162.59	1163.62	1180.67	10
5	520.23	492.23		G		1000.62	1017.6	9
6	667.31	639.36	120.09	F			960.58	8
7	795.38	767.48	101.07	Q		796.46	813.52	7
8		881.45	87.05	N		668.47	685.49	6
9	980.51			A		554.29	571.43	5

Protein identification since PEAKS 2.0



Advantages

- Compare score of best DB sequence with best de novo sequence
  - If the DB sequence's score is much smaller than the de novo sequence, then it is no good.
- Smaller protein "database" allows to spend more time to try different PTMs.

Clustering proteins

- A protein has many homologues in a database.
- If protein X has many peptides matched, then its homologues also have some/many peptides matched.
- PEAKS clusters all the proteins into groups of homologous proteins.

Other important factors to consider

- Preprocess the MS/MS spectra
  - Deconvolution, noise reduction, and signal enhancement.
- Recalibration
  - compress/stretch the spectrum for calibration error
- File formats
  - pkl, mgf, dta, mzxml,wiff, raw, ……

Acknowledgement

Bin Ma
Kaizhong Zhang

Gilles Lajoie
Amanda Doherty-Kirby
Cunjie Zhang

Ming Li

L. Guo, C. Hendrie, C. Liang, Z. Wang, W. Zhang, W. Yang, W. Chen, I. Rogers, C. Wigmore.