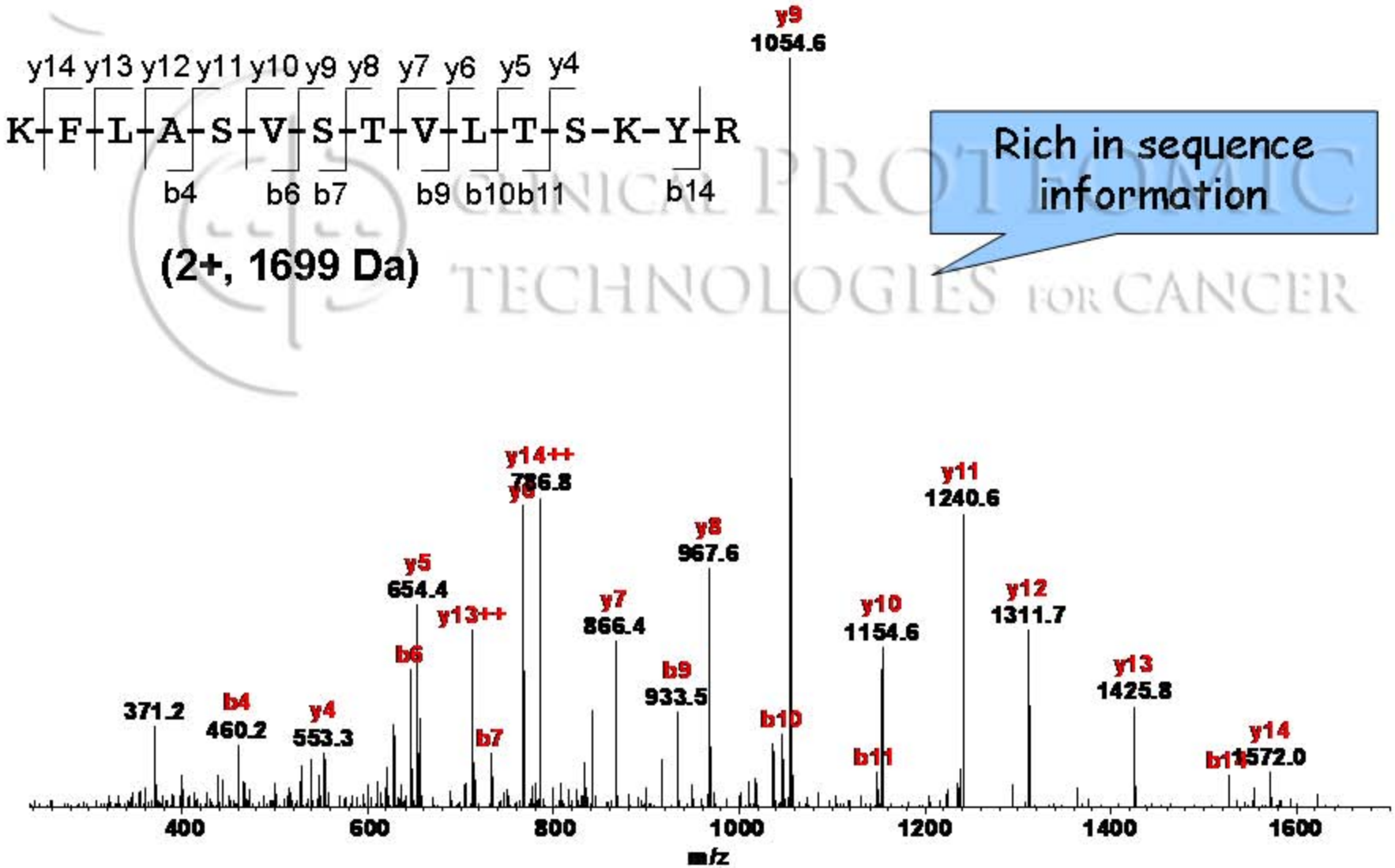
AMGEN

Physical Models for Predicting Peptide Fragmentation Spectra

Zhongqi Zhang
Amgen

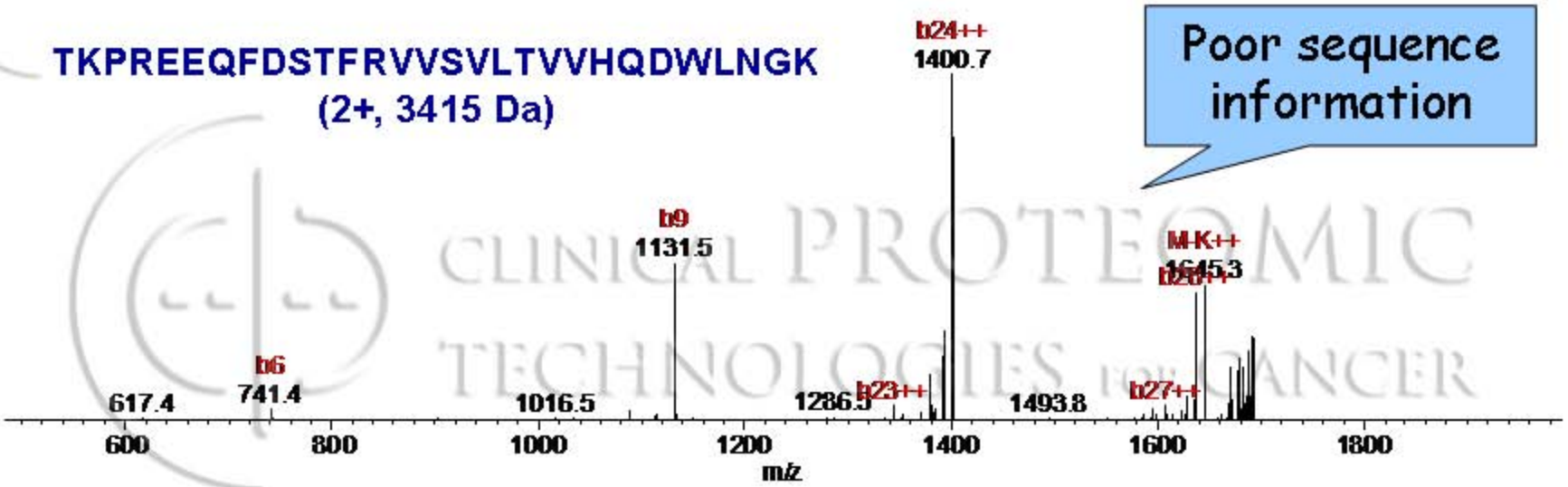
Strategies for Improving Reliability in Protein/Peptide Identification Workshop
NCI-NIST, November 15, 2007

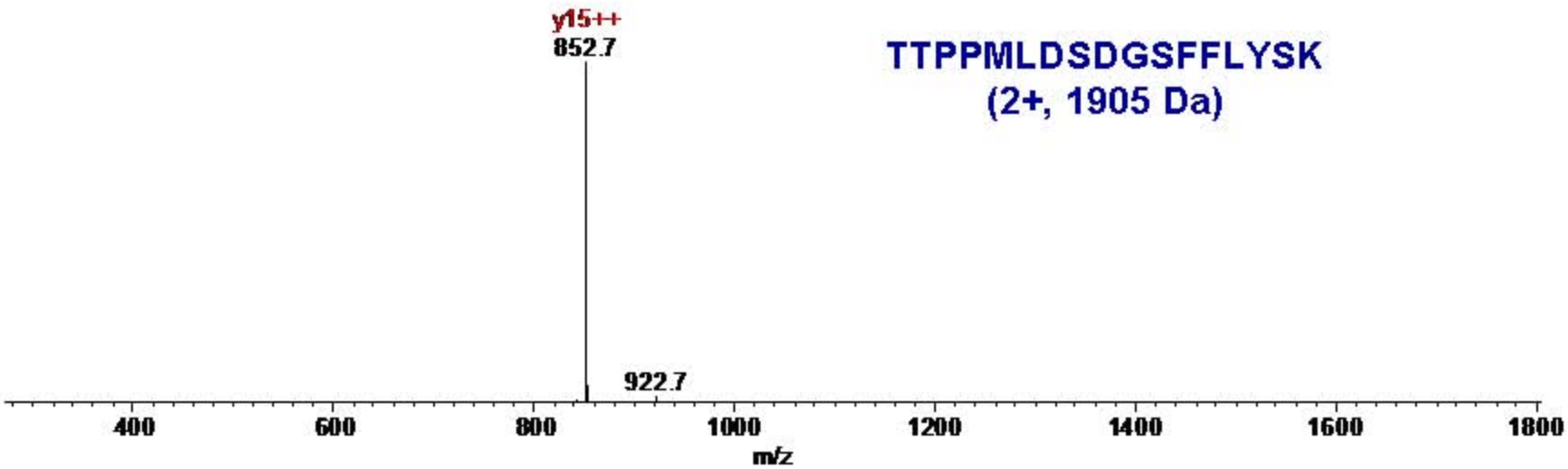
Peptide Identification by MS/MS



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

All b/y Fragments are not Created Equal!





CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

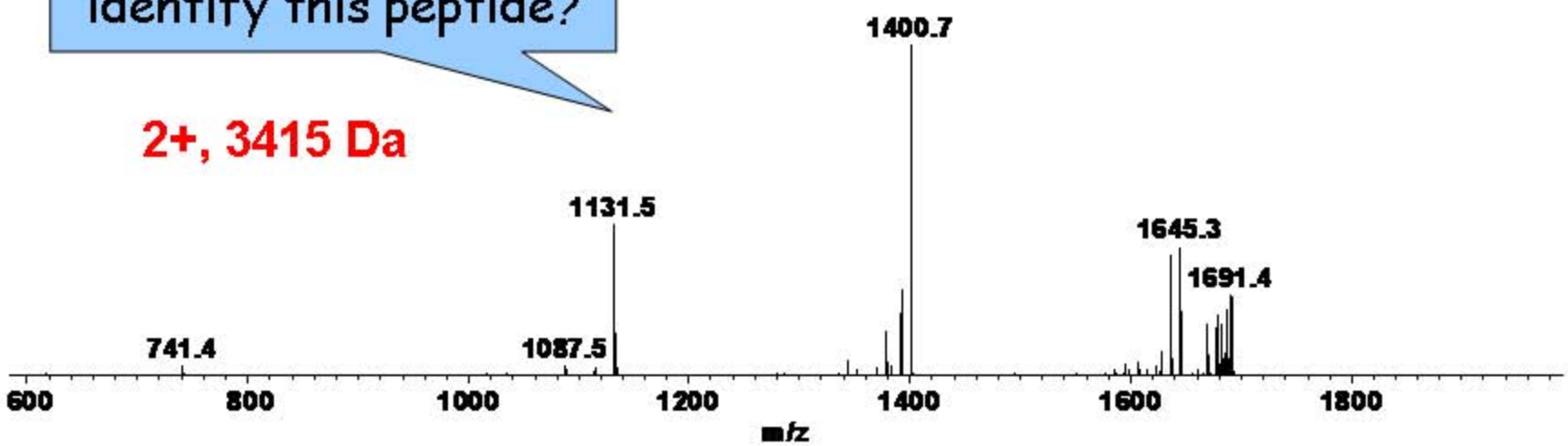
Protein Identification vs. Protein Characterization

- Protein Identification
  - Identify all proteins in a sample
- Protein Characterization
  - Identify all peptides in a sample

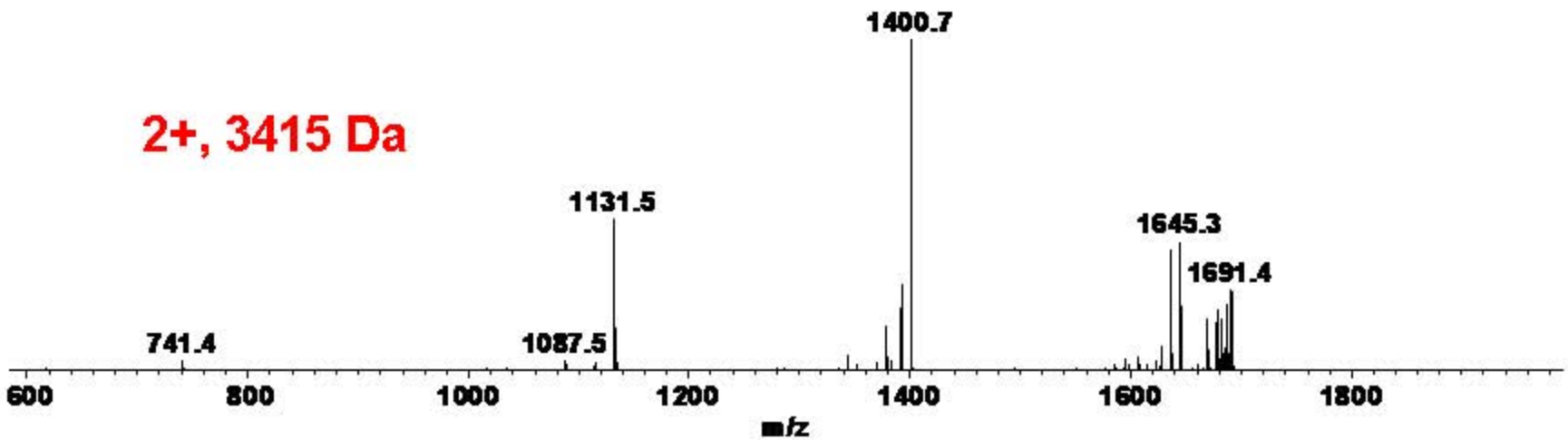
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Can we reliably identify this peptide?

2+, 3415 Da

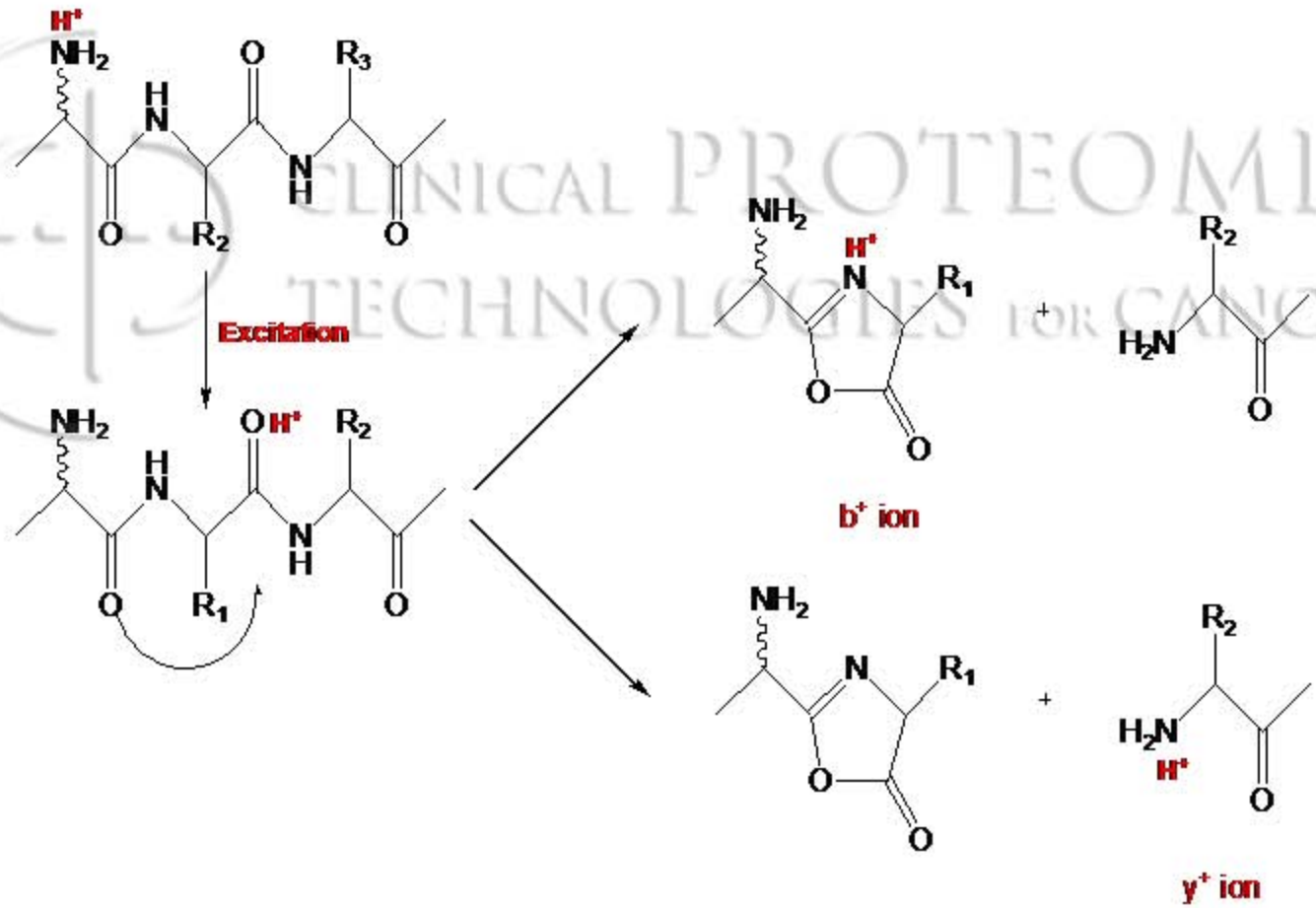


A Predictive Model for Peptide ID?



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

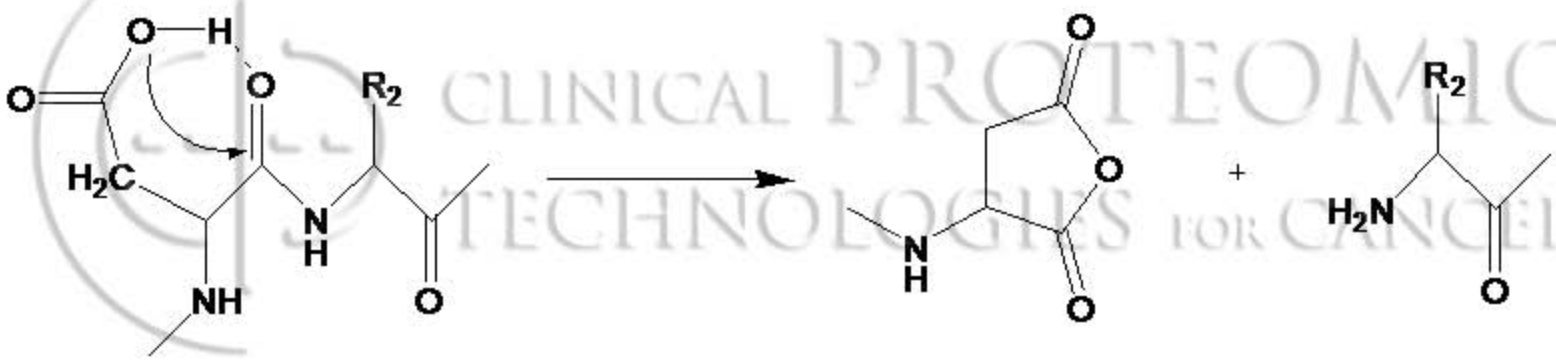
"Mobile Proton" Model
Charge-Directed Fragmentation



Wysocki et al., *J. Mass Spectrom.* (**2000**), 35, 1399-1406

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Charge-Remote Fragmentation



-D-X- b ion y ion

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Wysocki et al., *J. Mass Spectrom.* (**2000**), 35, 1399-1406

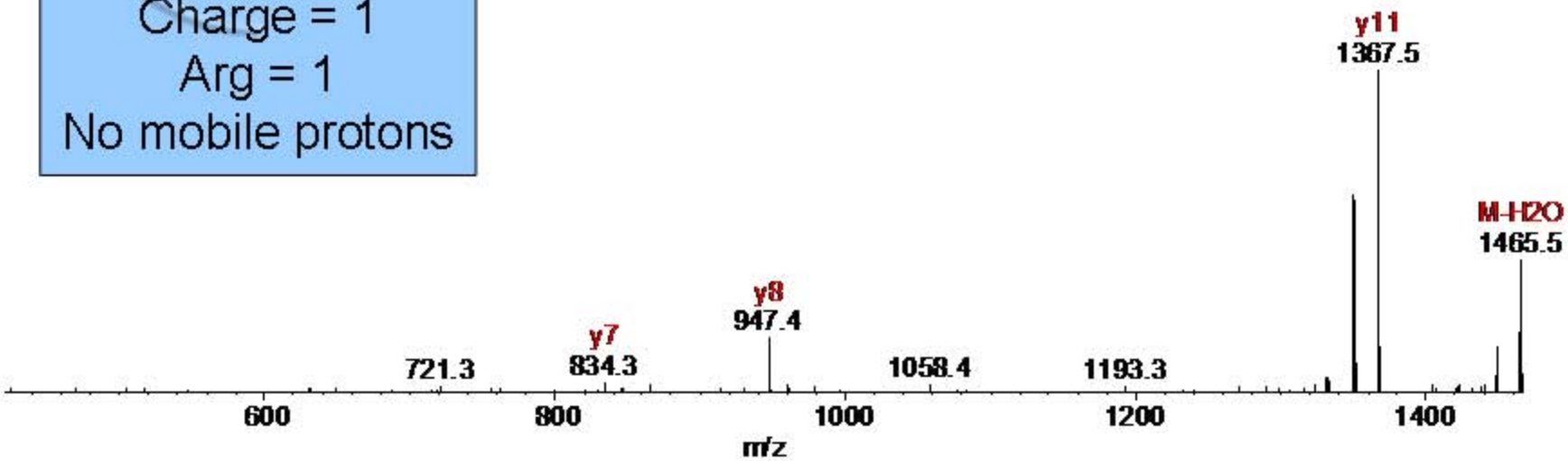
Charge-Remote Fragmentation

y11 y8

D-Q-Y-E-L-L-C-L-D-N-T-R

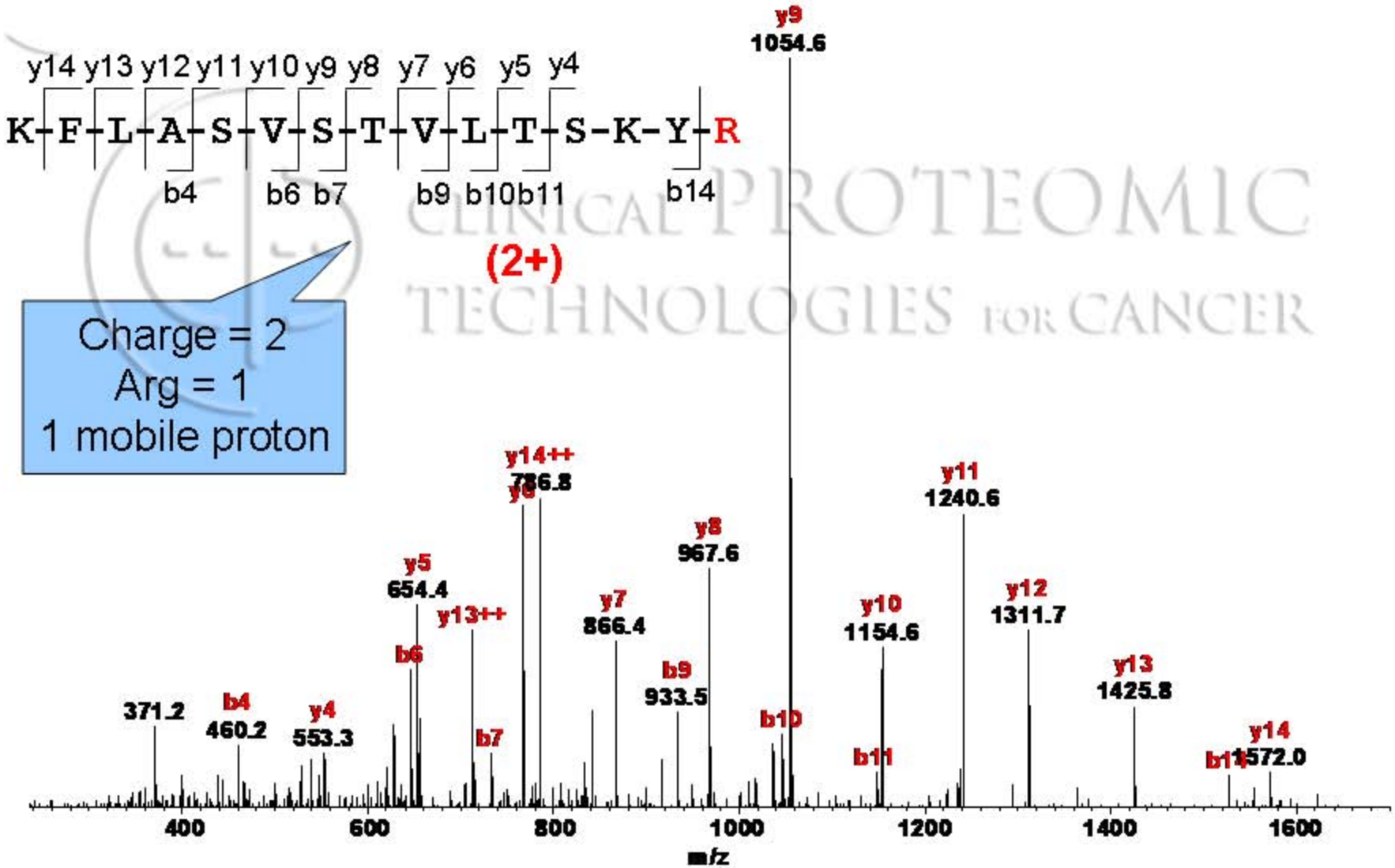
(H^+)

Charge = 1
Arg = 1
No mobile protons



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Charge-Directed Fragmentation



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Peptide Fragmentation

- Competition of many different pathways, including charge-directed and charge remote.
  - *Classical Chemical Kinetics*
- Knowledge of charge distribution is crucial in determining the peptide fragmentation behavior.
  - *Boltzmann Distribution*

PROTEOMIC TECHNOLOGIES FOR CANCER

The Kinetic CID Model

Boltzmann Distribution

$$p_j = \frac{\exp(-E_j / RT_{eff})}{\sum_i \exp(-E_i / RT_{eff})}$$

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Arrhenius Equation

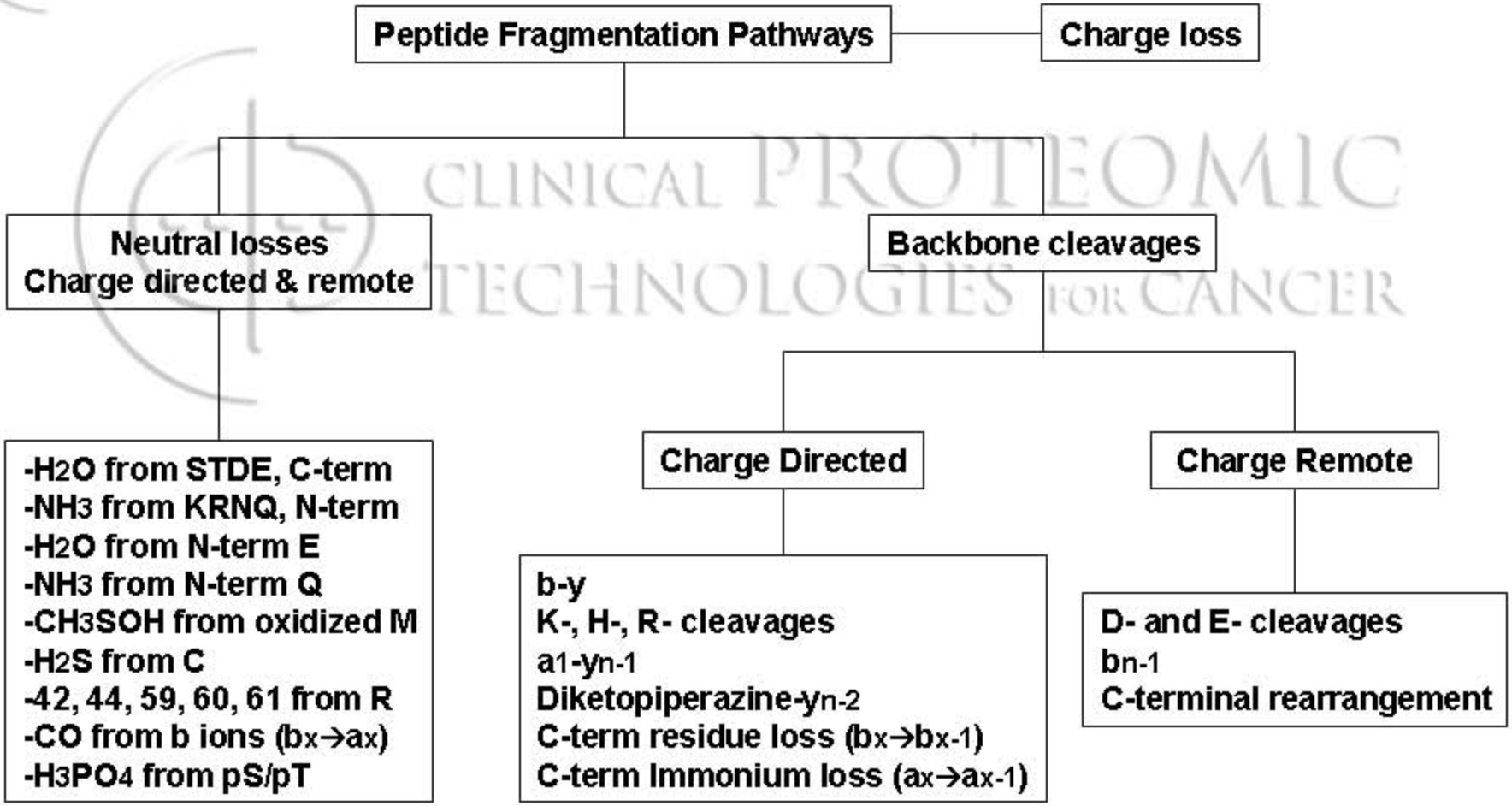
$$k = A \exp(-E_a / RT_{eff})$$

First-Order Kinetics

$$\frac{d[F_i]}{dt} = k_i [P]$$

Zhang, *Anal. Chem.* (**2004**), 76, 3908-3922
Zhang, *Anal. Chem.* (**2005**), 77, 6364-6373

Fragmentation Channels



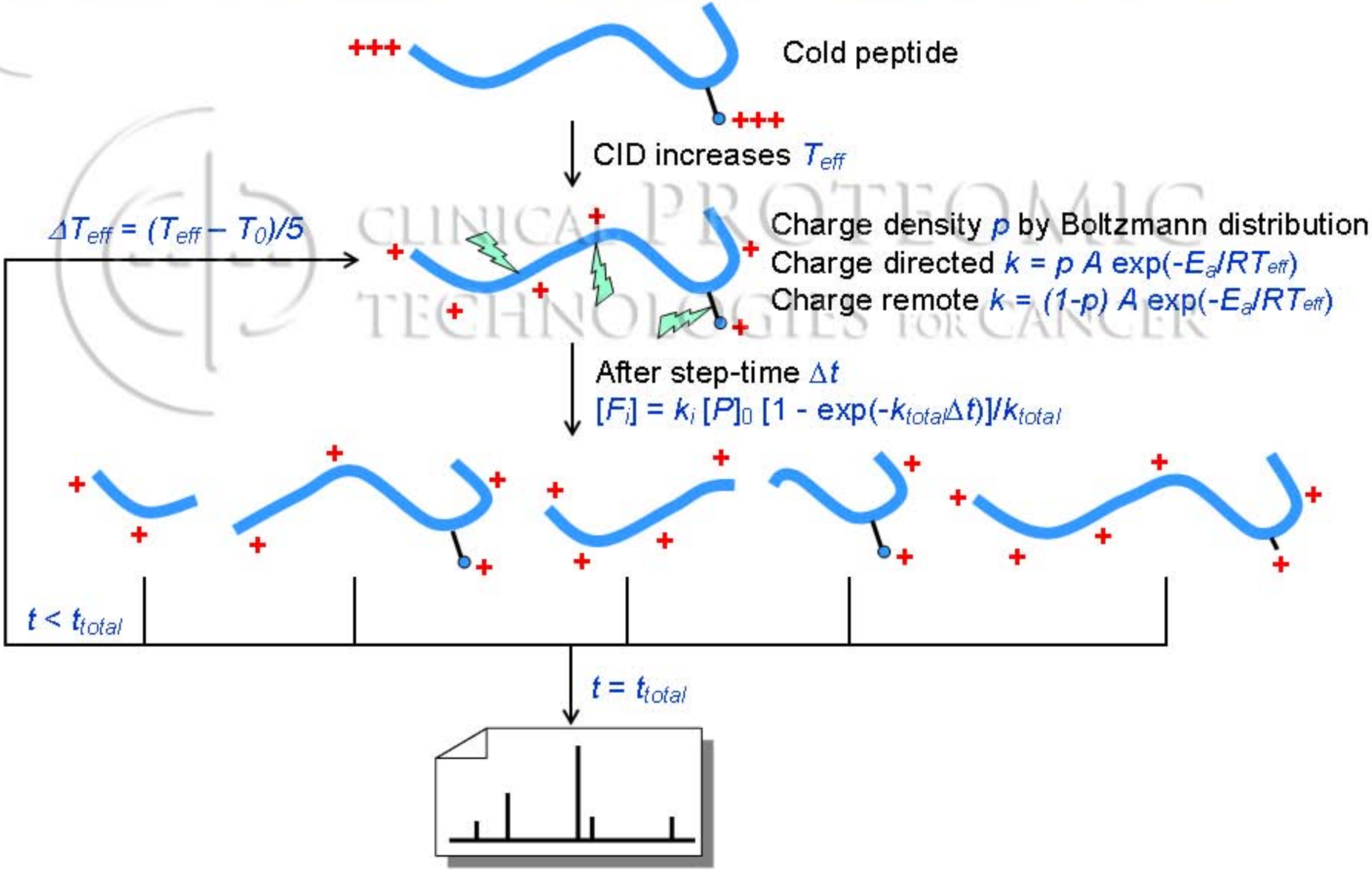
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Major Assumptions

- T_{eff} is related to the mass, charge and collision energy.
- Side chain of each residue has a constant gas-phase basicity (GB).
- GB of a backbone amide depends on the neighboring residues, and the effects are additive.
- Intra-molecular proton transfer is a fast process.
- E_a for each backbone cleavage depends on the neighboring residues, and the effects are additive.
- Same fragmentation mechanism has the same *A* factor.

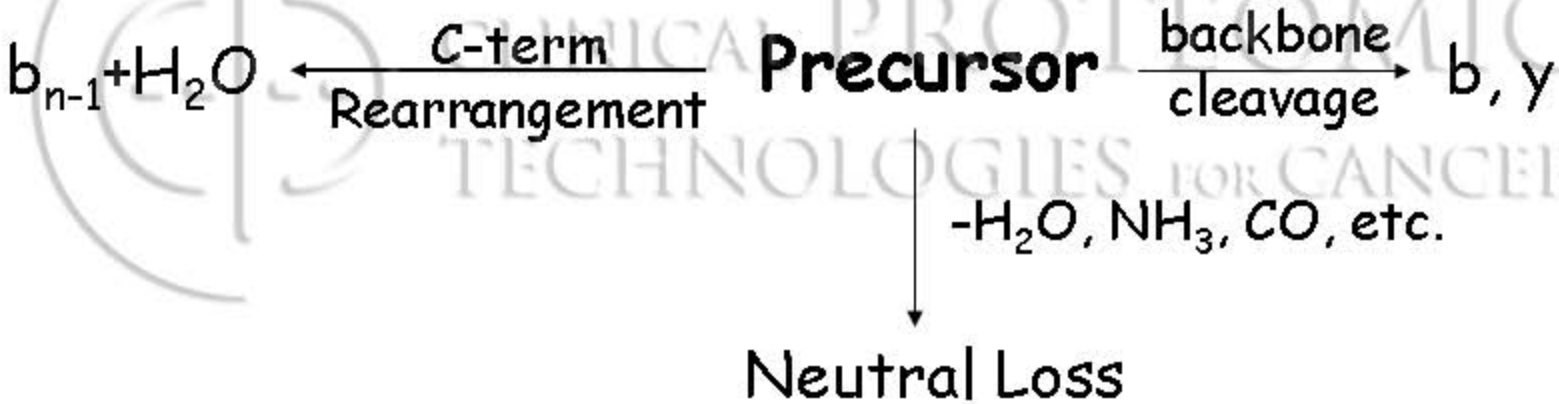
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

The Model



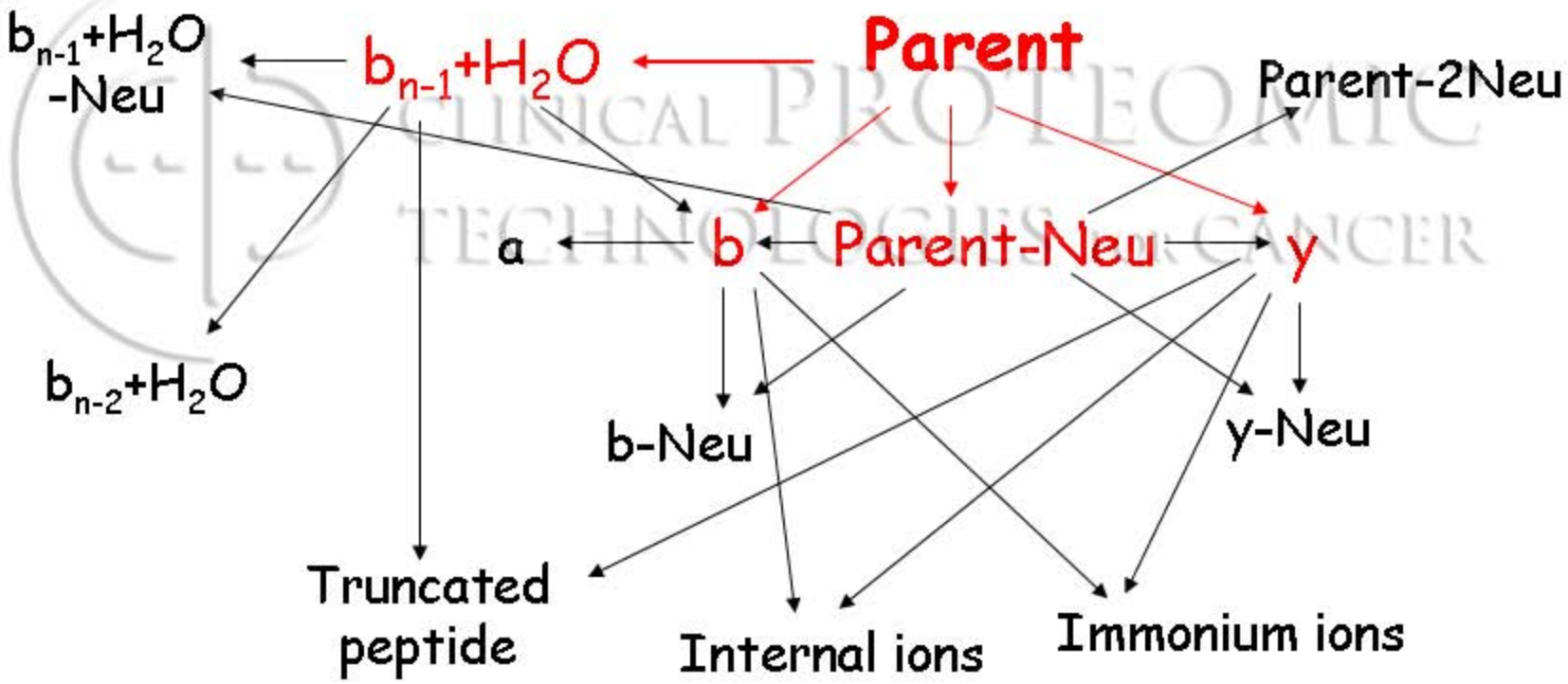
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

What Ions Are Predicted?



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Ions Predicted within Two Generations



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Optimizing the Model

- 336 parameters (15 instrument dependent)
- Initial values from best estimates based on literature values
- >10,000 high-quality ion-trap spectra used in the training

$$Similarity = \frac{\sum \sqrt{I_m^1 I_m^2}}{\sqrt{(\sum I_m^1)(\sum I_m^2)}}$$

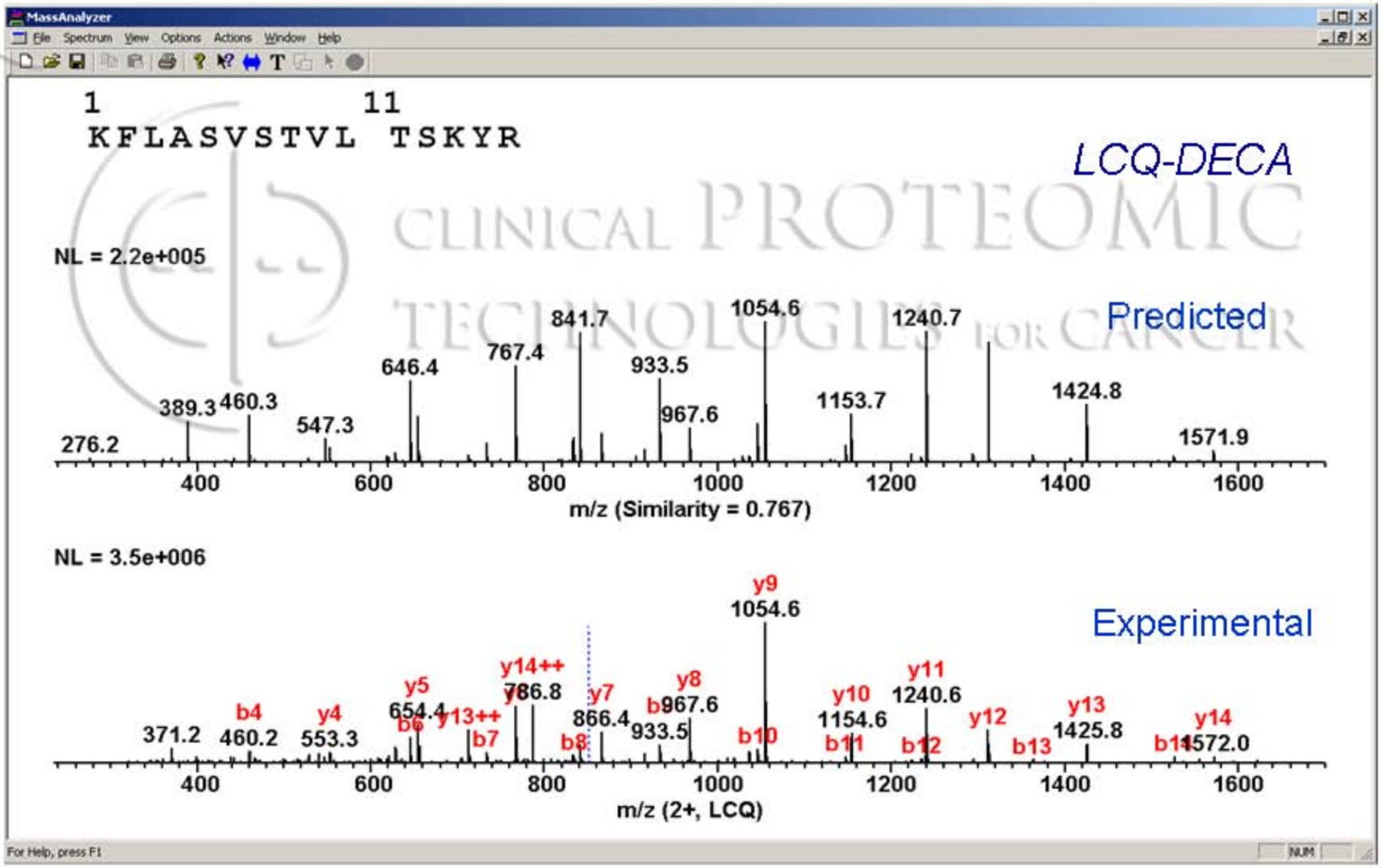
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Some Optimized Parameters

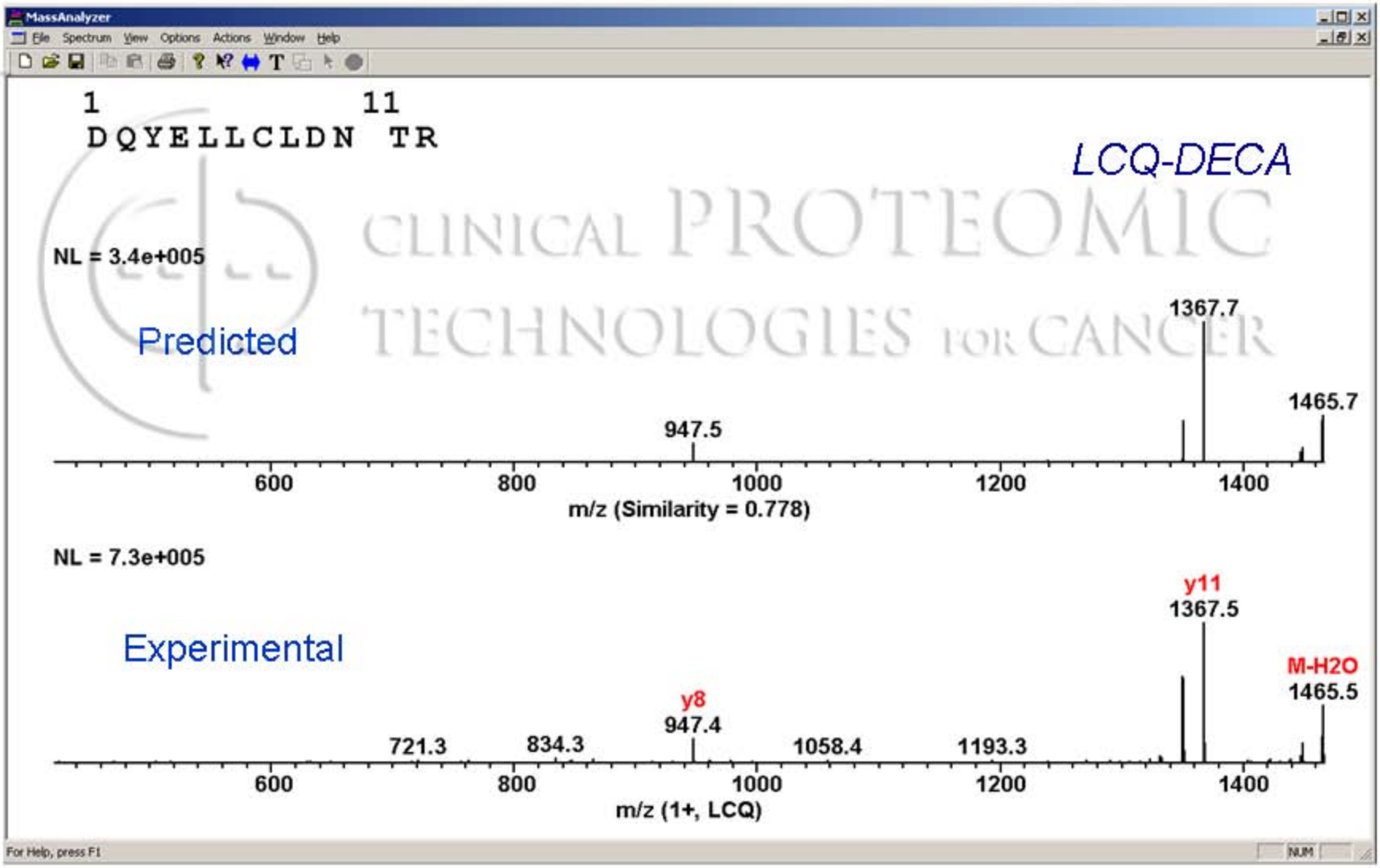
Residue	GB(SC)	GB(L)	GB(R)	EA(L-1)	EA(L)	EA(R)	EB(L)	EC(R)	EC(L)	Eb->a	ENL(d)	ENL(r)	Epyro(d)	Epyro(r)
A		883.4	0.0	0.0	168.5	0.0		185.0	0.0	130.2				
C	770.0	882.0	-1.2	4.9	171.7	-1.0		189.0	-3.2	135.6	152.5	121.6		
D	784.0	880.3	-1.0	4.4	170.4	0.3	152.9	175.6	1.1	148.3	77.4	118.9		
E	790.0	881.7	2.9	2.7	167.7	2.9	167.6	183.8	-1.3	132.2	86.5	144.4	36.7	184.8
F		881.8	0.7	0.3	165.2	-2.3		184.4	0.7	123.0				
G		883.2	1.9	2.4	181.2	0.0		206.8	-1.2	146.2				
H	929.7	883.6	-0.1	-3.1	166.4	-2.9	165.2	169.6	0.6	118.9				
I		881.1	-1.1	-1.6	159.4	-3.3		176.1	4.4	120.0				
J	830.0	880.6	-0.1	4.0	168.4	-0.8		189.7	-0.2	131.8	210.7	122.3		
K	927.6	880.4	-2.2	1.6	165.3	-4.7	161.1	169.5	2.0	187.0	206.6	221.9		
L		882.9	-0.2	-1.4	164.1	-1.3		180.4	0.9	123.4				
M	830.0	882.8	-0.5	0.4	166.8	-2.6		183.6	-3.5	125.2				
N	863.5	881.4	1.5	2.1	171.9	1.8		188.0	0.1	134.0	127.8	111.2		
O	854.0	882.8	0.4	6.2	173.9	3.8		187.0	1.0	137.4	205.4	91.2		
P		885.5	16.4	5.2	187.8	-0.4		174.3	2.1	126.7				
Q	864.6	882.9	6.2	2.9	168.4	7.1		189.1	-1.1	138.9	138.1	125.6	208.9	169.7
R	964.0	887.6	9.2	4.4	166.7	7.9	178.7	167.8	-6.9	133.9	197.4	111.2		
S	775.0	881.4	1.8	3.5	172.4	0.0		189.5	-1.2	138.7	37.8	124.0		
T	780.0	882.2	1.7	0.3	169.9	1.8		188.3	0.8	134.2	37.7	130.8		
U	860.0	882.4	0.9	5.0	171.4	1.4		188.3	-3.7	138.2	202.4	122.2		
V		882.9	-2.0	-0.3	161.9	-2.8		178.4	2.9	123.8				
W	897.6	883.1	3.0	-1.1	165.0	-2.0		184.0	2.5	123.3				
Y	790.0	882.3	0.0	0.3	166.5	-2.5		181.4	2.7	122.2				
N-term		911.9		30.4							297.5	127.7		
C-term			-97.4				169.6				51.5	125.5		
b ion			38.3			-13.6					177.7	101.6		
a ion			-200.0								246.5	81.9		
A(1/s)					6.4E+13		3.2E+10	1.6E+11	2.1e+09(directed)				1.4E+12	1.4E+12
									6.8e+08(remote)					

CENTRAL PROTEOMIC INSTITUTE FOR CANCER

A "Good" Spectrum

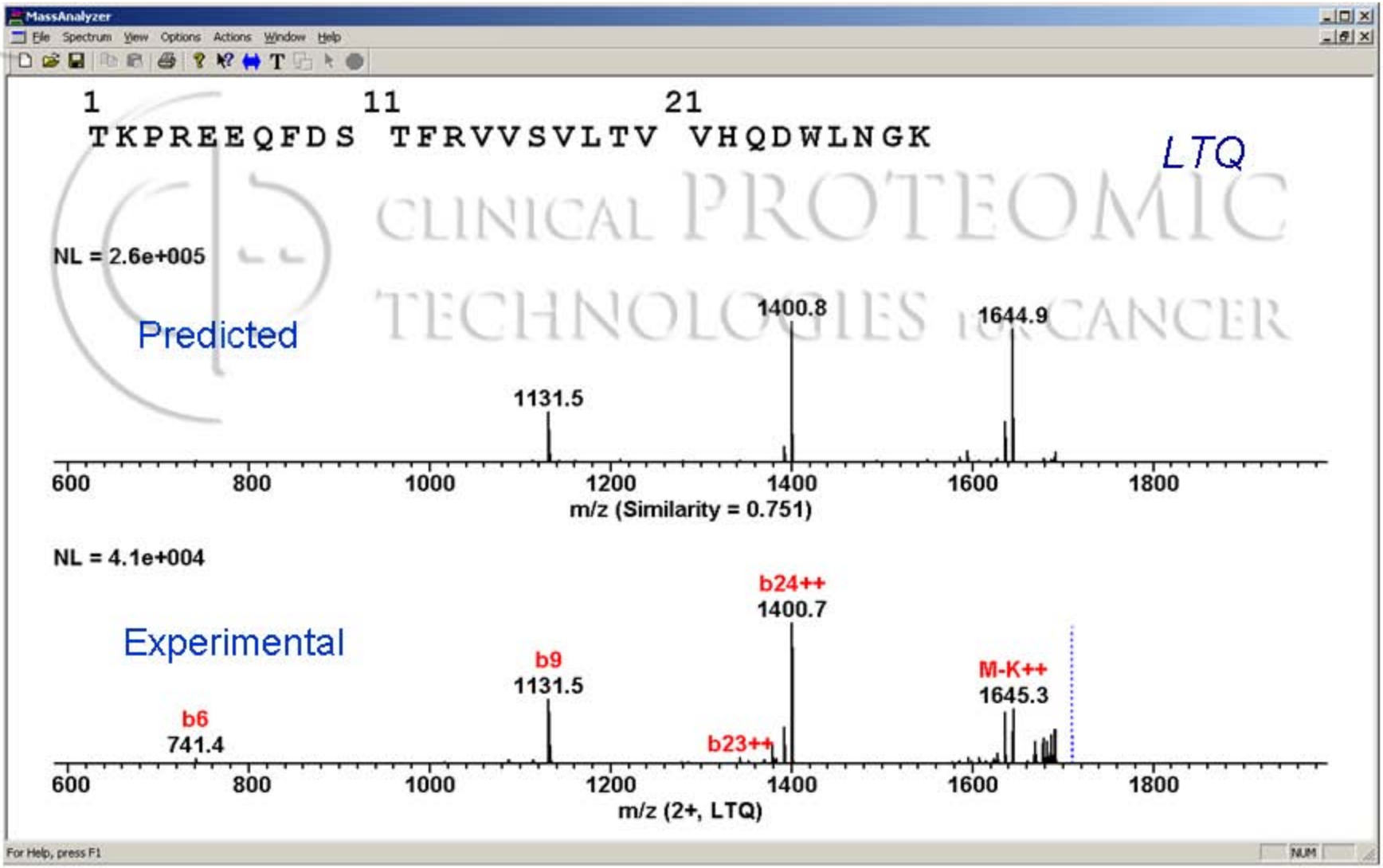


A "Bad" Spectrum?

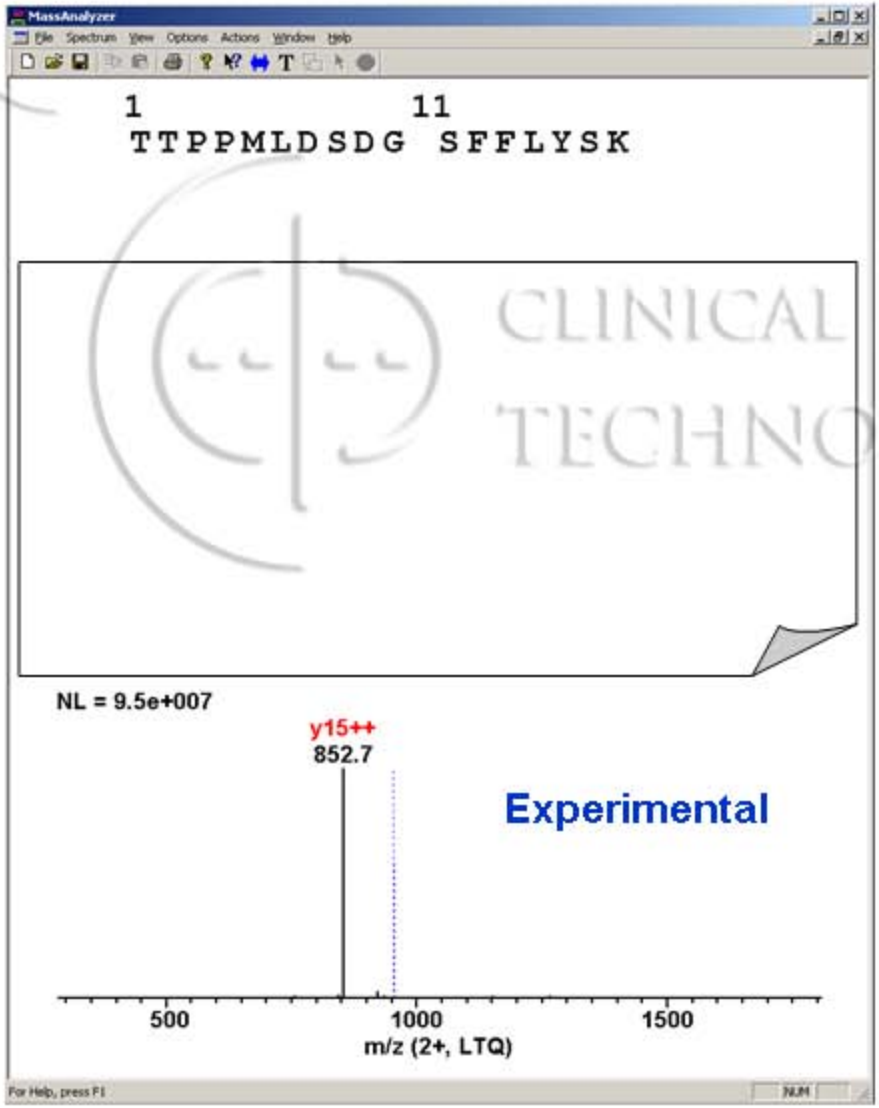


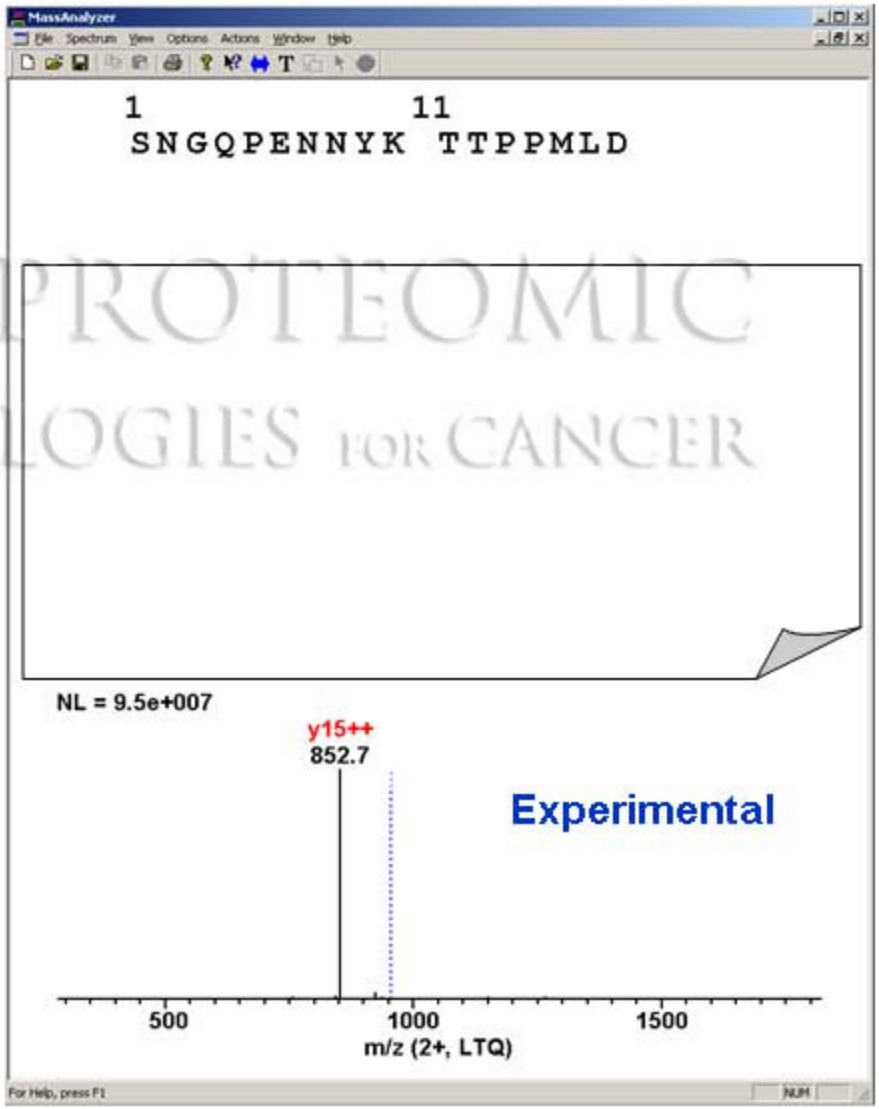
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

“Lack of information” is also information !



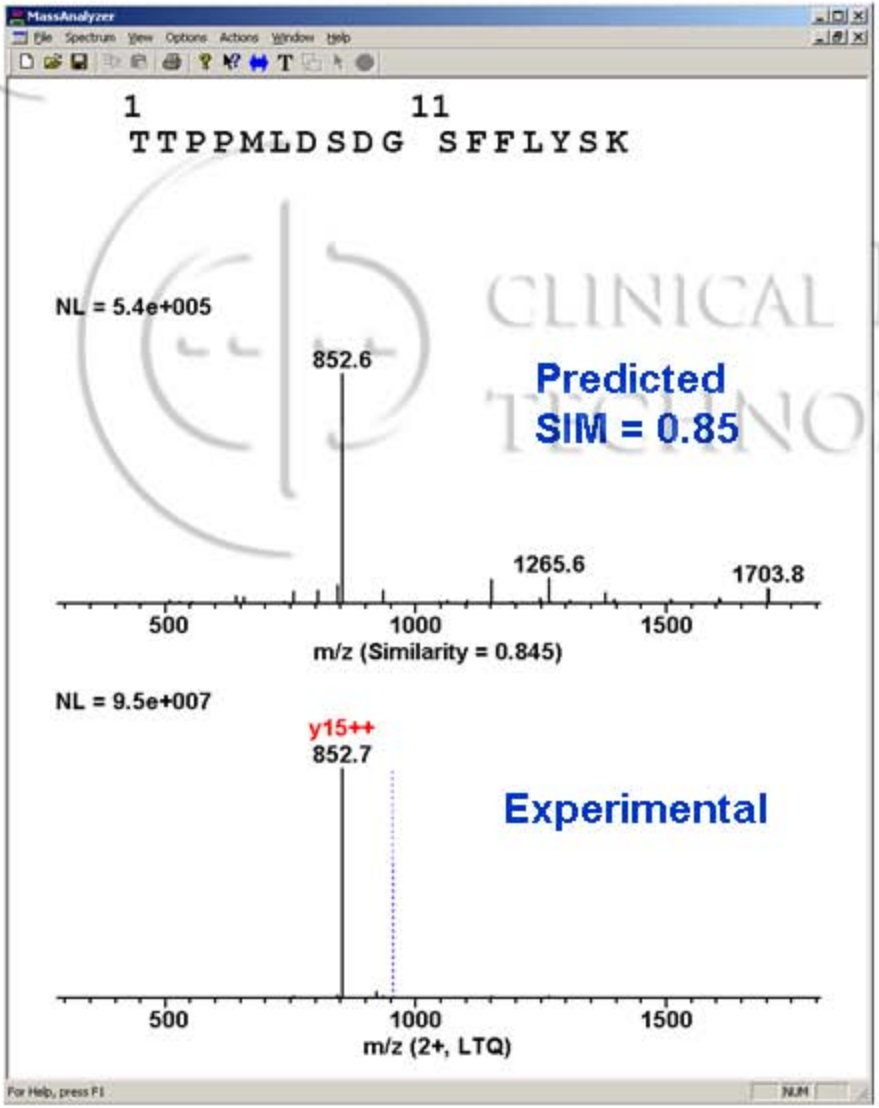
Can you distinguish these two peptides?

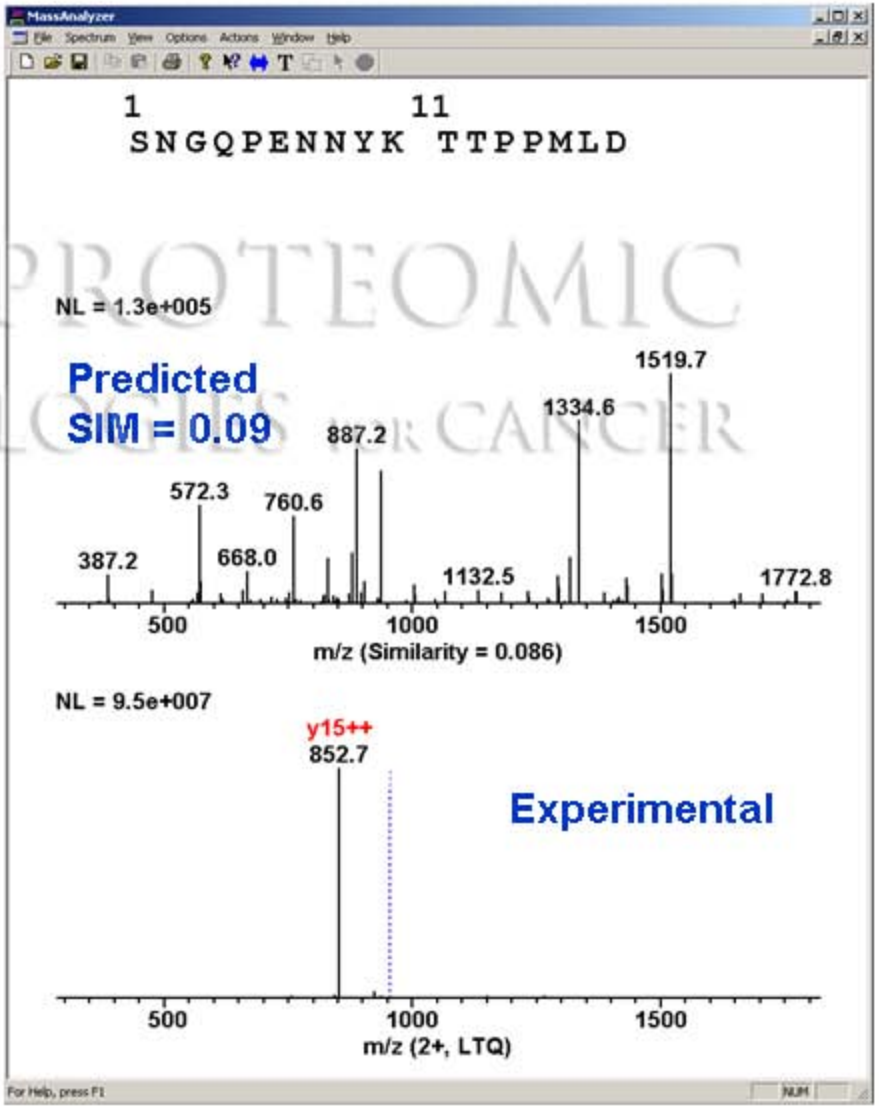




CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

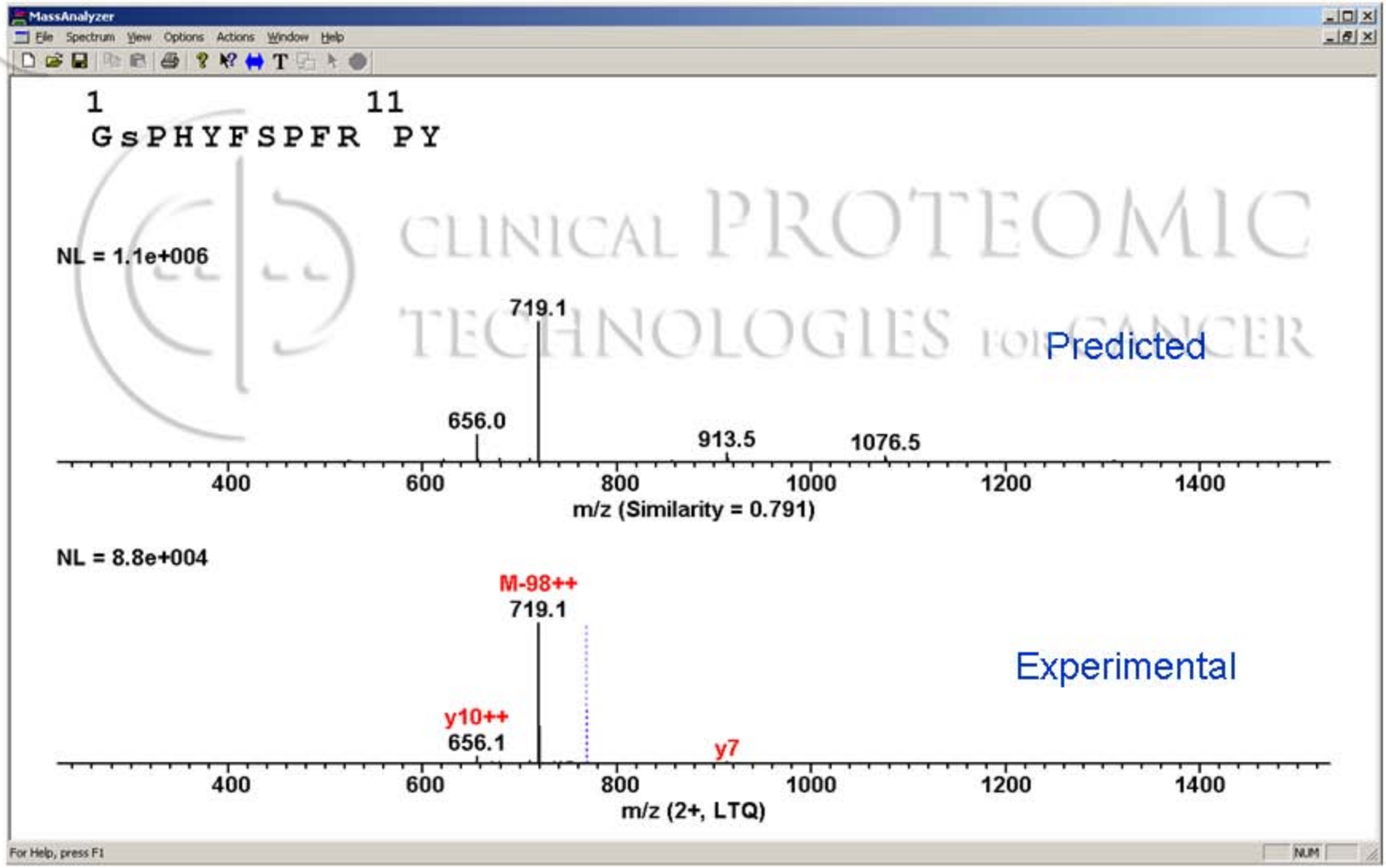
Peptide Distinguished





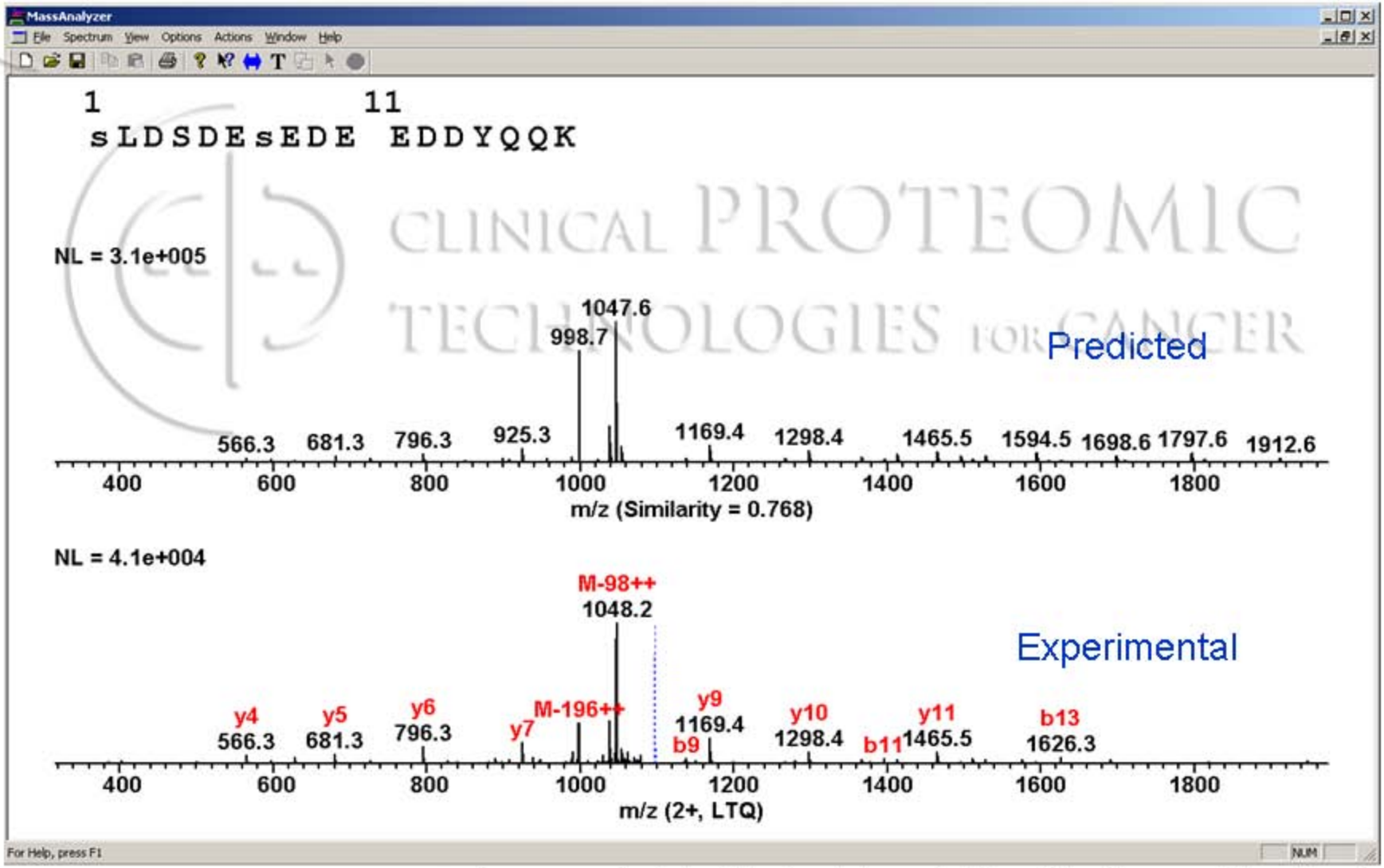
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Phosphopeptide



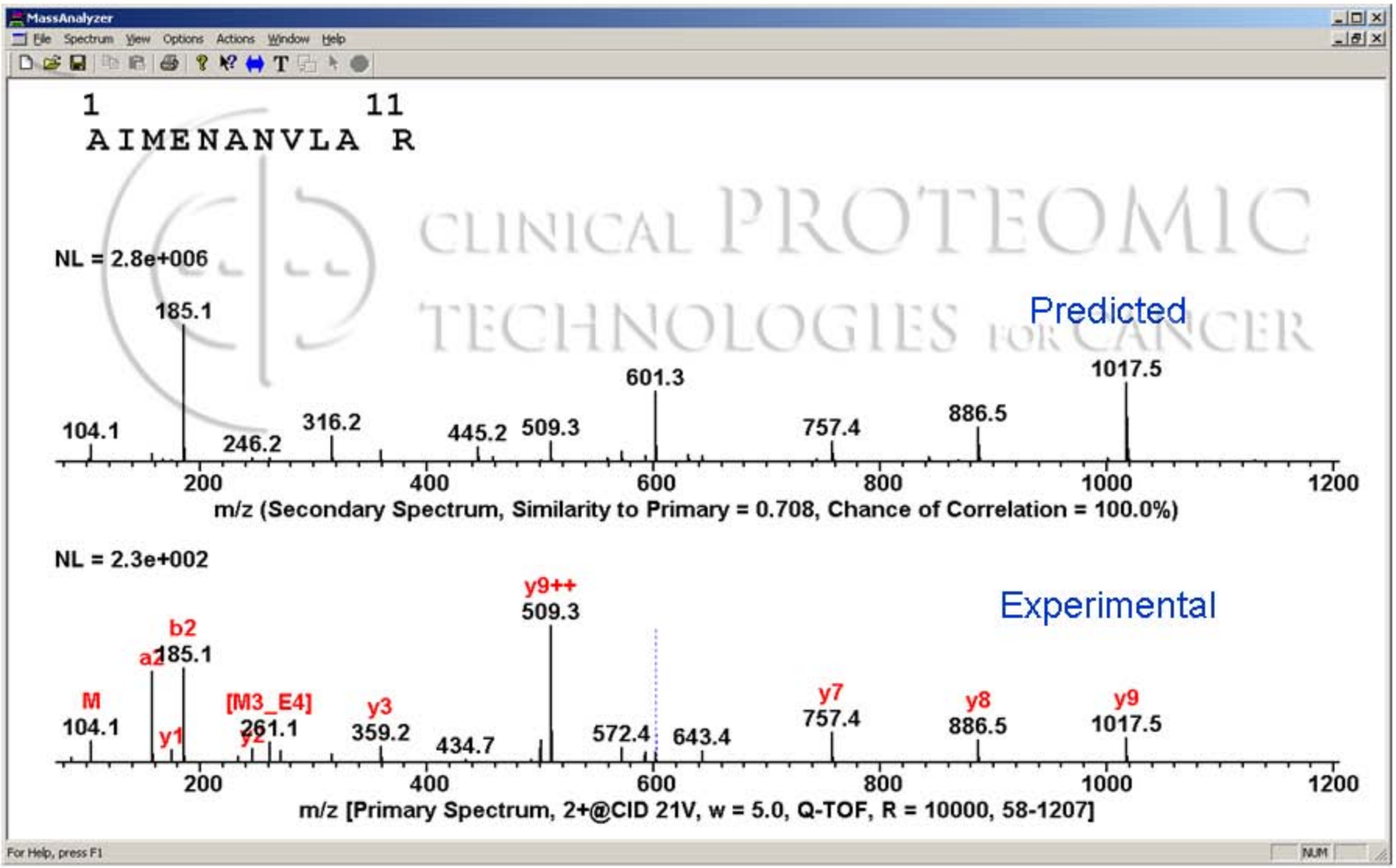
Data generated by Kimberly Lee & Wen Yu (Amgen Washington)

Phosphopeptide



Data generated by Kimberly Lee & Wen Yu (Amgen Washington)

Q-TOF



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Issues in the Model

- Computation speed
  - A long step time has to be used for computation speed, with sacrifice of accuracy

$$\frac{d[F_i]}{dt} = k_i[P]$$

$$[F_i]_{\Delta t} = \frac{k_i[P]_0}{k_{total}}[1 - \exp(-k_{total}\Delta t)]$$

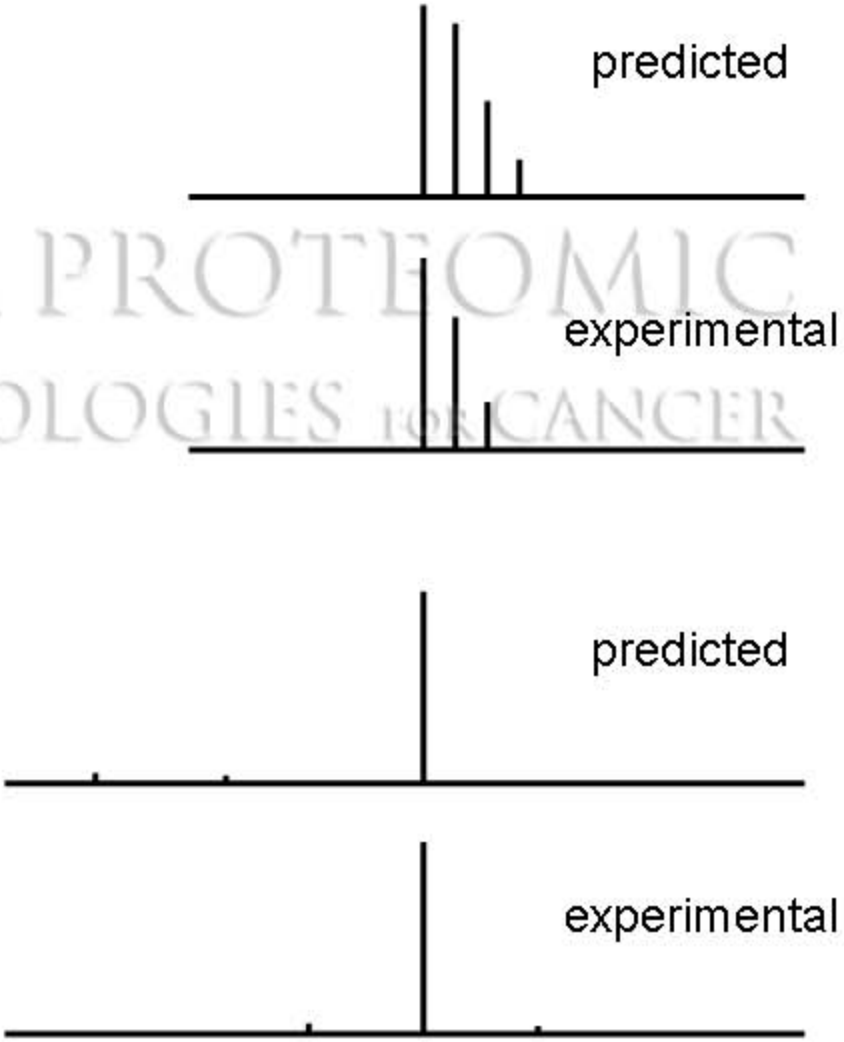
Step time Δt -

- Theoretically infinitely small
- Practically set as time required for temperature to decrease ~120K

CANCER PROTEOMIC TECHNOLOGIES FOR CANCER

Issues in the Model

- Computation speed
  - A long step time has to be used for computation speed, with sacrifice of accuracy
- Similarity score
  - Isotope peaks weigh the same as any other peaks
  - High intensity peaks weight too much



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Issues in the Model

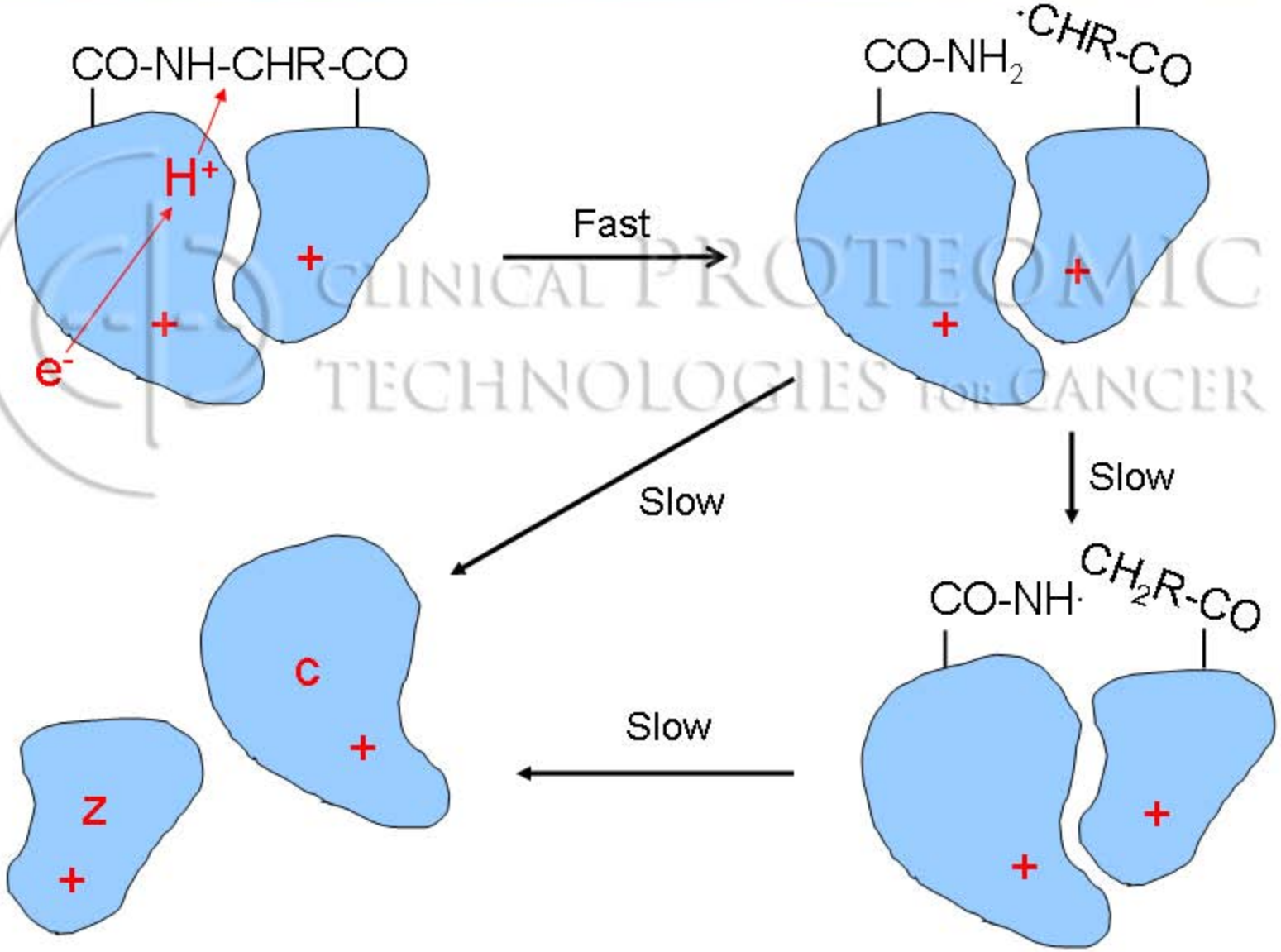
- Computation speed
  - A long step time has to be used for computation speed, with sacrifice of accuracy
- Similarity score
  - Isotope peaks weigh the same as any other peaks
  - High intensity peaks weight too much

$$\frac{\sum \sqrt{I_m^1 I_m^2}}{\sqrt{(\sum I_m^1)(\sum I_m^2)}}$$

Need a better score function

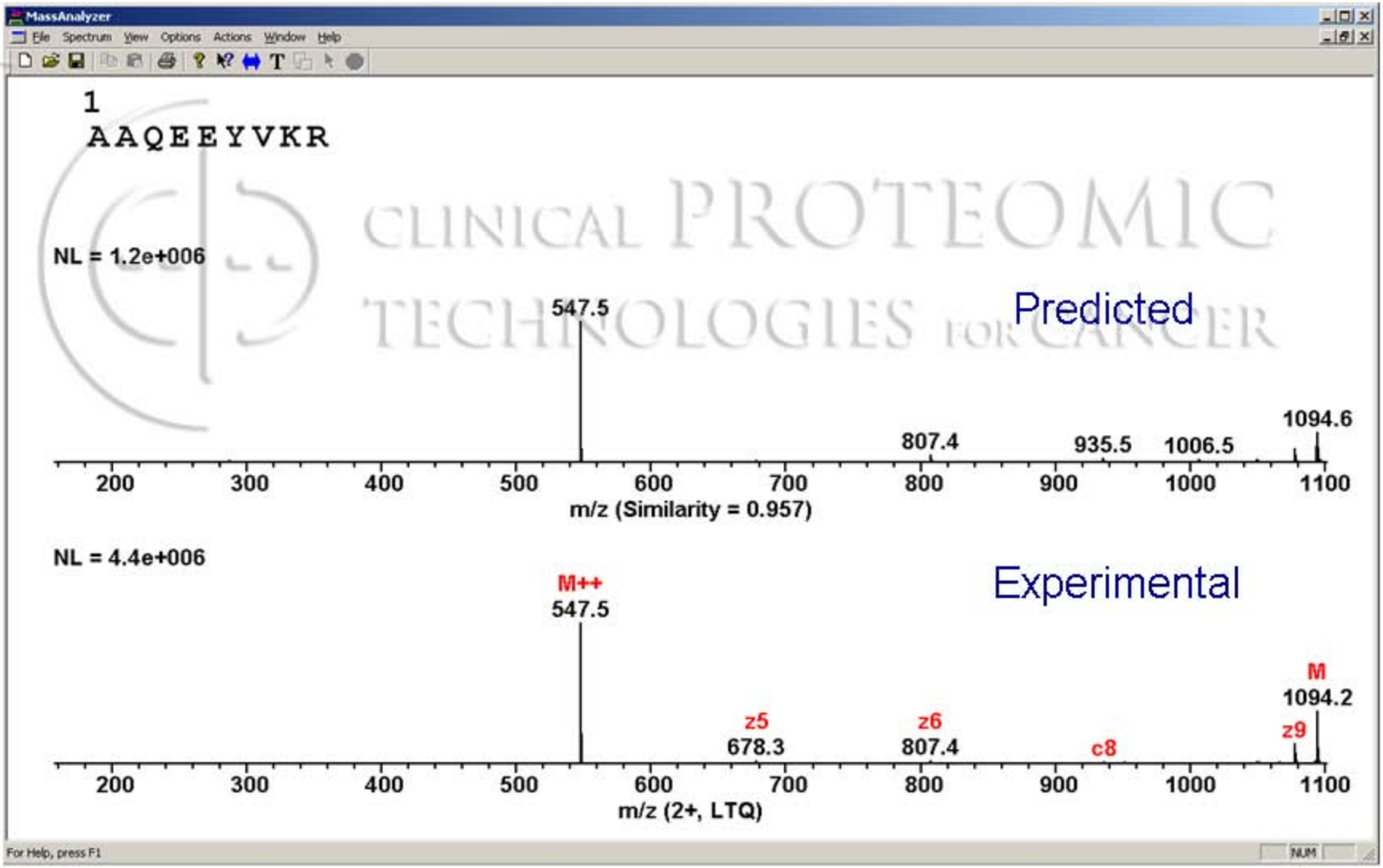
CLINICAL PROTEOMIC TECHNOLOGIES for CANCER

Electron-Transfer Dissociation



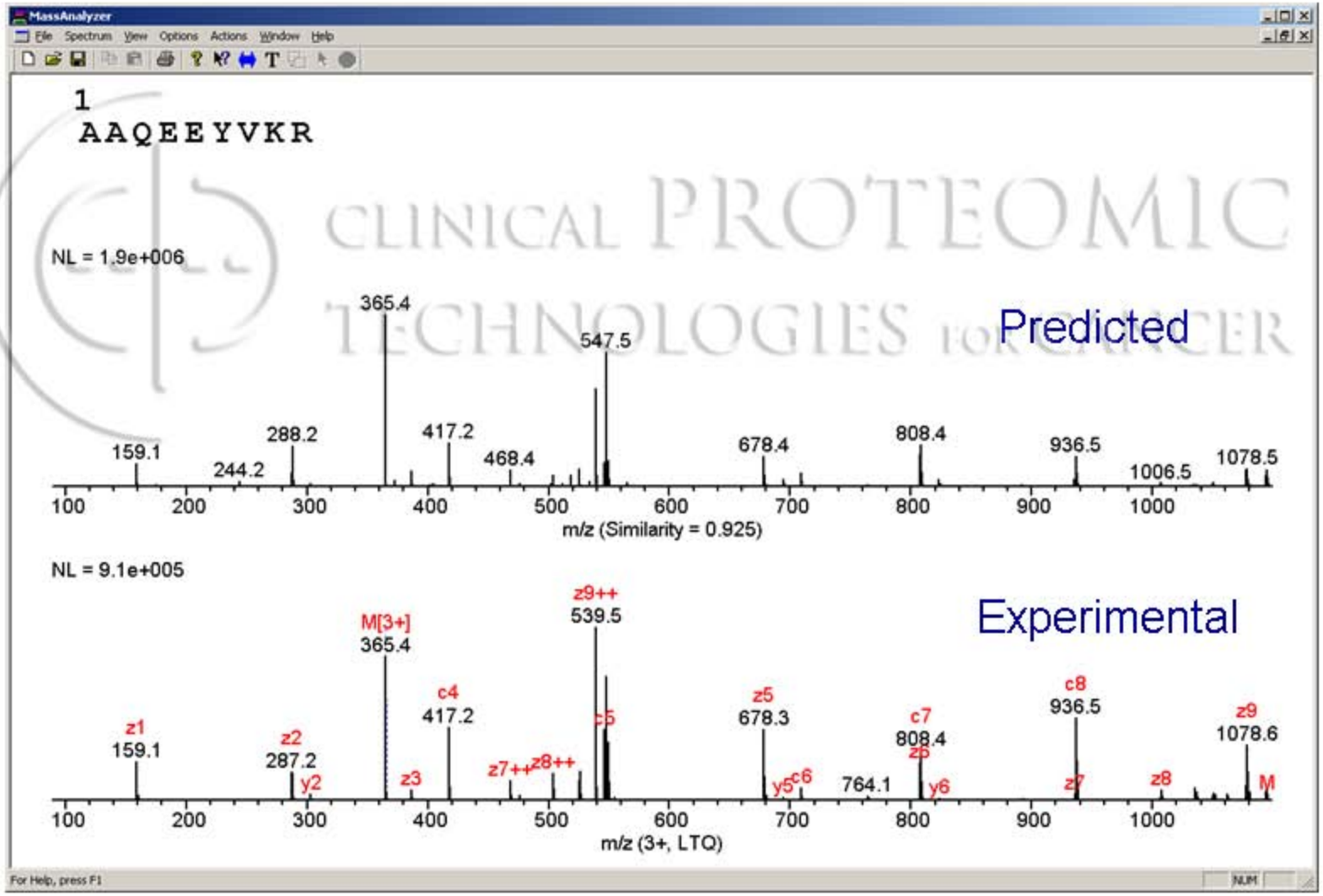
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

ETD of a 2+ Peptide



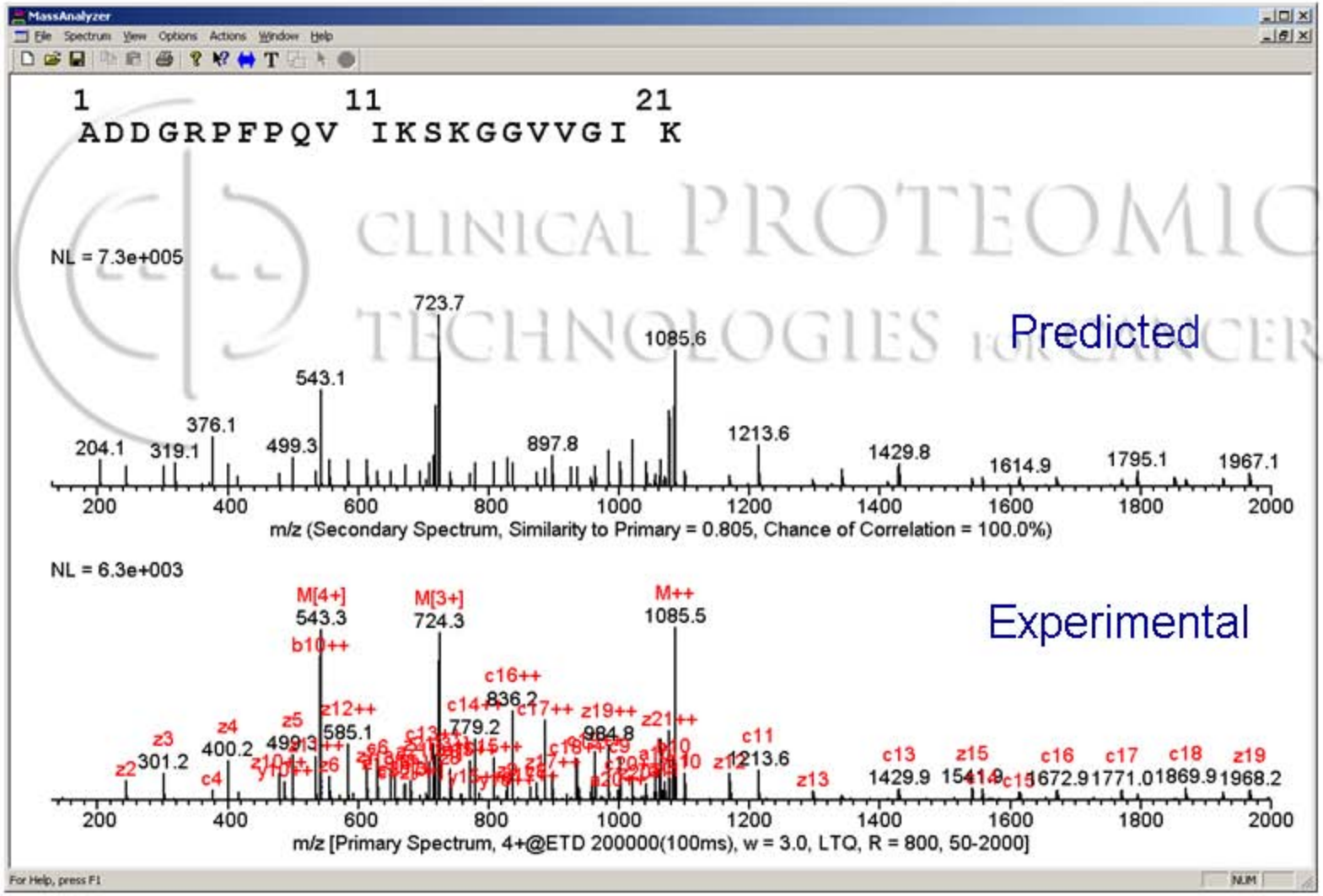
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

ETD of a 3+ Peptide



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

ETD of a 4+ Peptide



AMGEN

Potential Applications in Proteomics for Peptide Identification?

CENTER FOR PROTEOMIC TECHNOLOGIES FOR CANCER

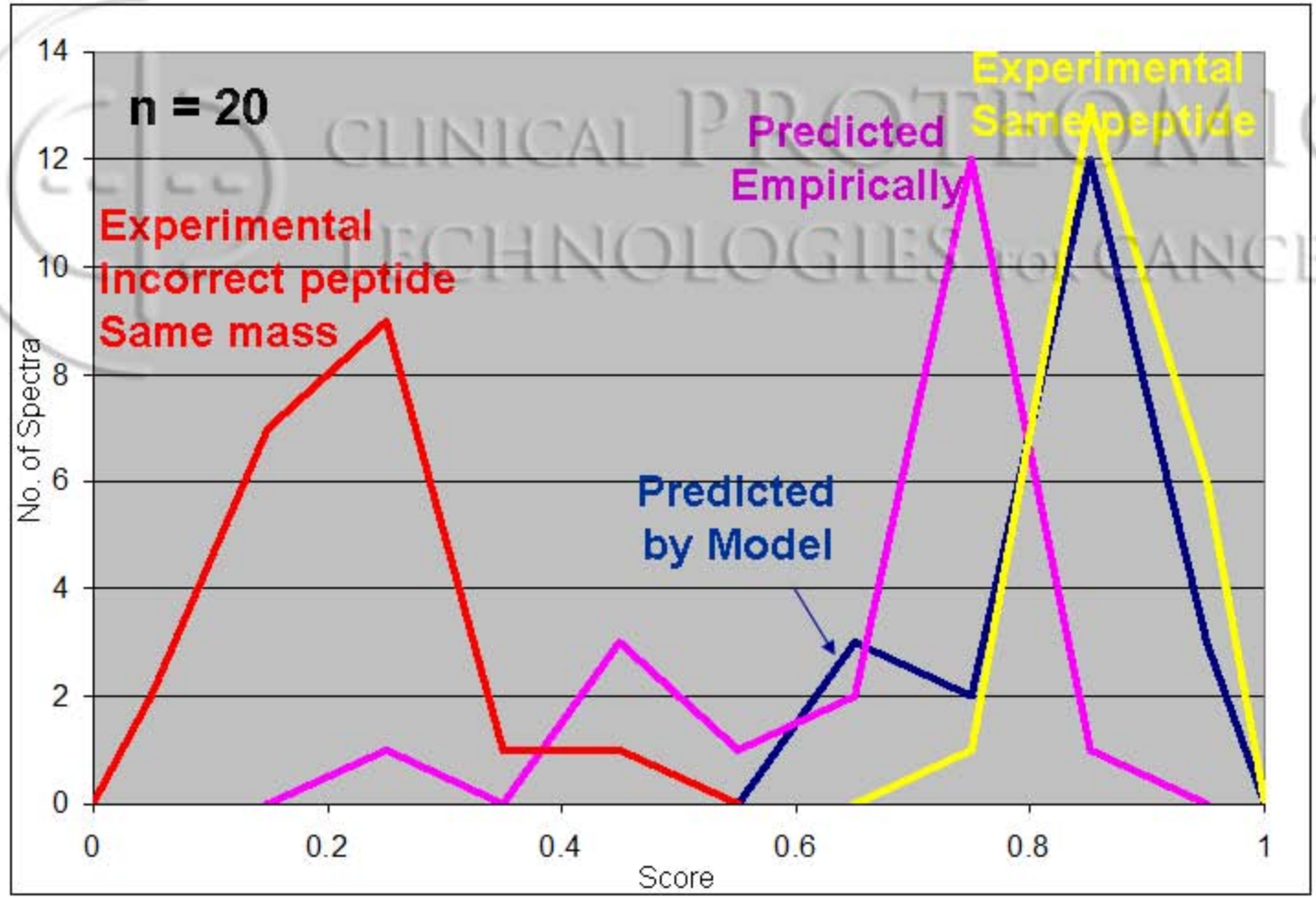
Applications in Proteomics

- Challenge
  - Too slow for database search with on-the-fly predictions (~100 spectra/second)
- Potential solutions
  - Used for validating results after a loose-criteria database search
  - Pre-calculate theoretical mass spectral library

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

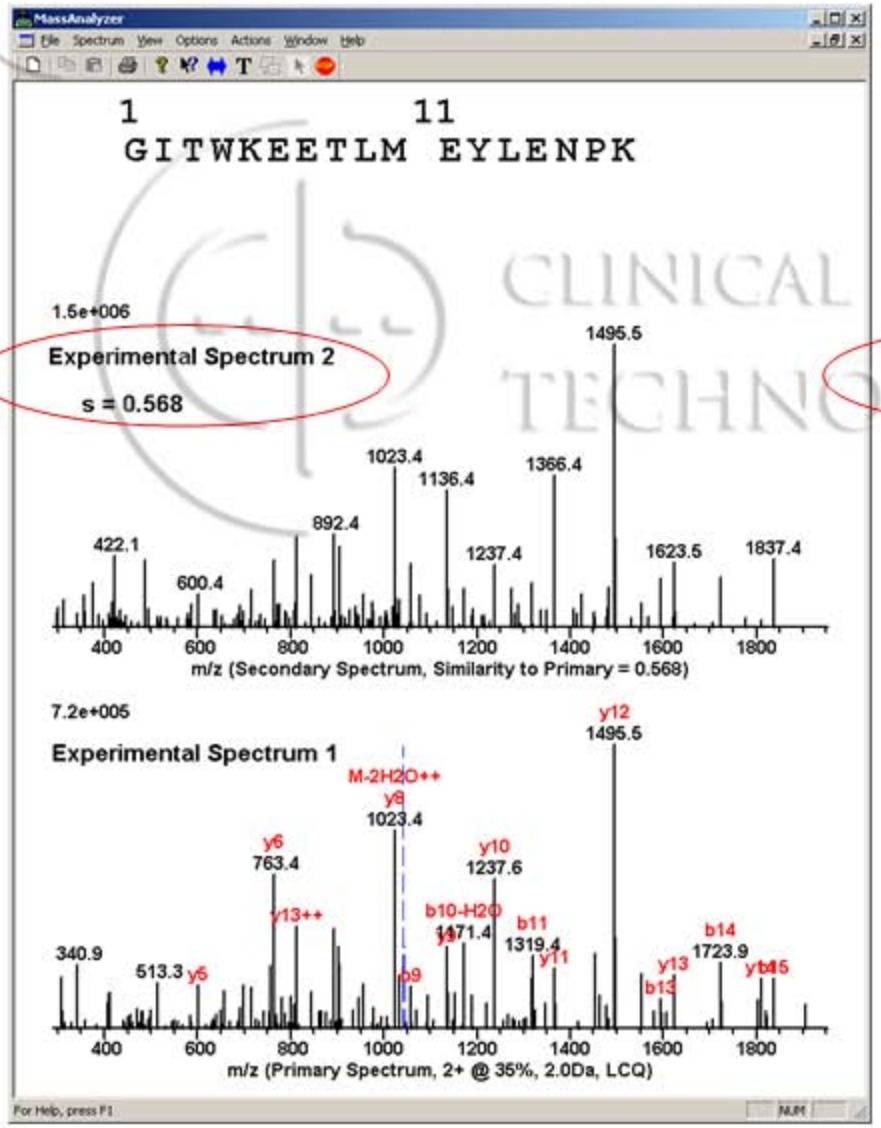
What as Reference Spectra in a Spectral Library?

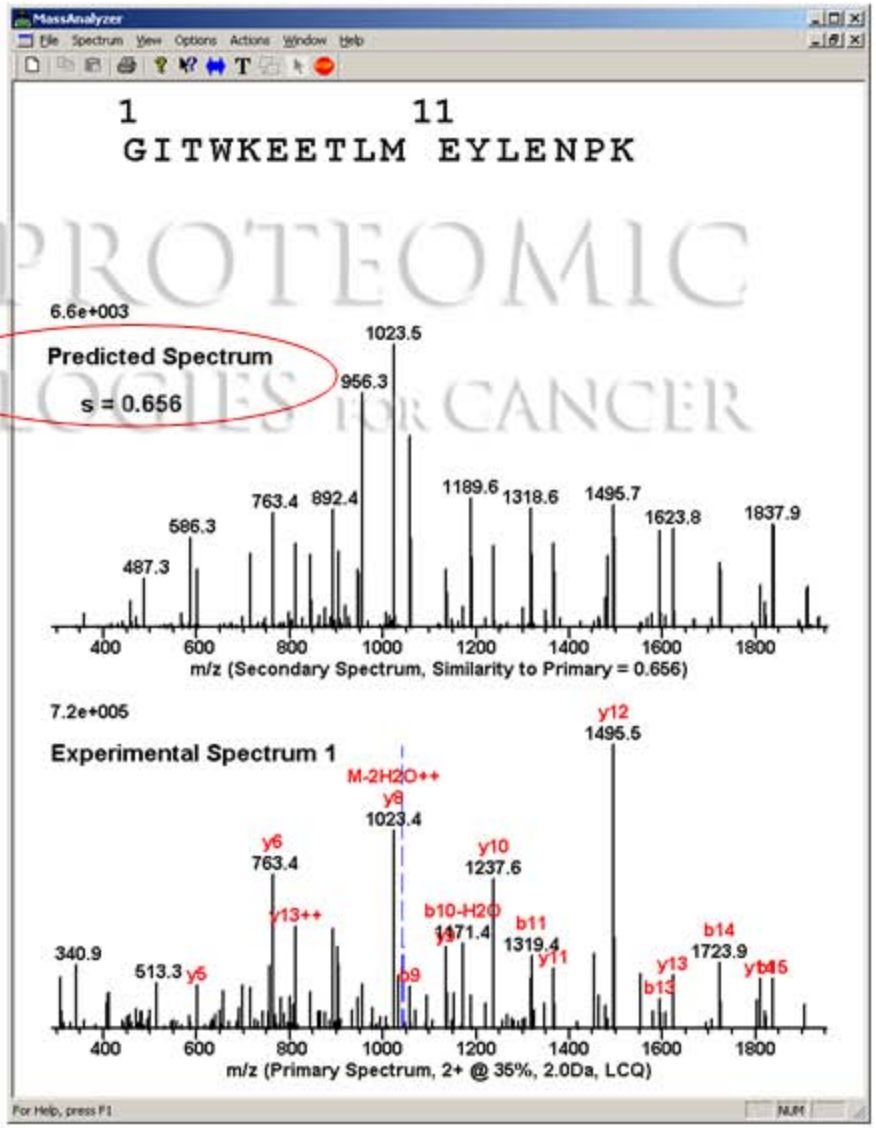
Distribution of similarity scores for different reference spectra



CLINICAL PROTEOMIC TECHNOLOGIES for CANCER

Theoretical or Experimental Spectra?





CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Theoretical Spectra as Reference Spectra?

- Not perfect but
  - Noise free!
  - No incorrect peptide assignment.
  - Fast way to get large amount of reference spectra.
  - Include spectra with few sequence ions.
  - May be better than low-quality experimental spectra such as those from shotgun experiments.
  - Definitely better than no experimental spectra at all.

CANCER PROTEOMIC TECHNOLOGIES for CANCER

Conclusions

- Physical modeling is a practical method to predict peptide fragmentation spectra, including CID and ETD.
- The CID model is successfully used for reliable peptide identification, especially for spectra with limited sequence information.
- The model can be used for more reliable peptide identification in proteomics experiments.

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER