Refining Peptide Fragmentation Models for Improved Confidence in Sequence/Spectrum Matching

Karl Clauser
Broad Institute of MIT and Harvard
Cambridge, MA

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Strategies for Improving Reliability in Protein/Peptide Identification Workshop

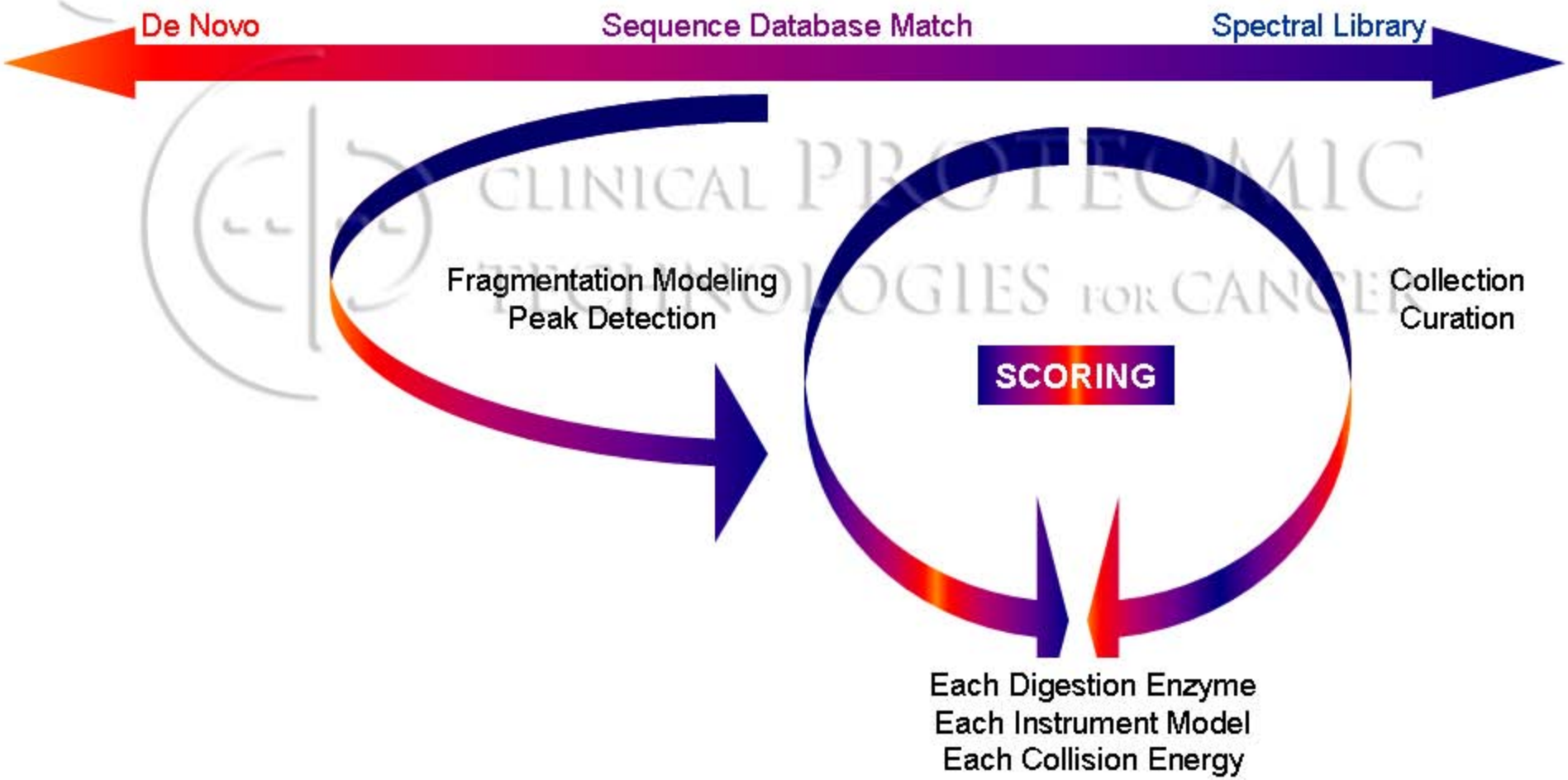
National Cancer Institute's (NCI)
Clinical Proteomic Technologies for Cancer (CPTAC) Initiative
and the
National Institute of Standards and Technology (NIST)
Gaithersburg, MD
November 15, 2007

Continuum of MS/MS Interpretation Strategies

1. Unsequenced Genome
2. Unexpected PTM
3. Multiple Mutations
4. Validate DB match
5. Spectral Quality

1. Sequenced Genome
2. Expected PTM
3. Expected Chem Mods
4. Single Mutations

1. Repeat Sample Analysis
2. Related Spectra
   a. Unexpected Partial Chem Mods
   b. Overlapping Peptides



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Are Spectral Similarity Scoring Functions Appropriate for Scoring Database Search Sequence/Spectrum Matches?

Dot product
Sokolow, 1978

$$\frac{(\Sigma(A_i * B_i))^2}{\Sigma A_i^2 * \Sigma B_i^2}$$

Spectral Contrast Angle
Wan, 2001

$$\frac{\Sigma(A_i * B_i)}{(\Sigma A_i^2 * \Sigma B_i^2)^{1/2}}$$

Similarity
Zhang, 2004

$$\frac{\Sigma(A_i * B_i)^{1/2}}{(\Sigma A_i * \Sigma B_i)^{1/2}}$$

Difference of the 3 Scores

Extent of Emphasis on low abundance ions

General Features of the Scores

Normalize Intensities & Number of Peaks

Score contribution for each ion scales with its intensity

Max Value is 1.0

Numerator

Bonus for Matched ions

Denominator

Penalty for Unmatched ions A Experimental

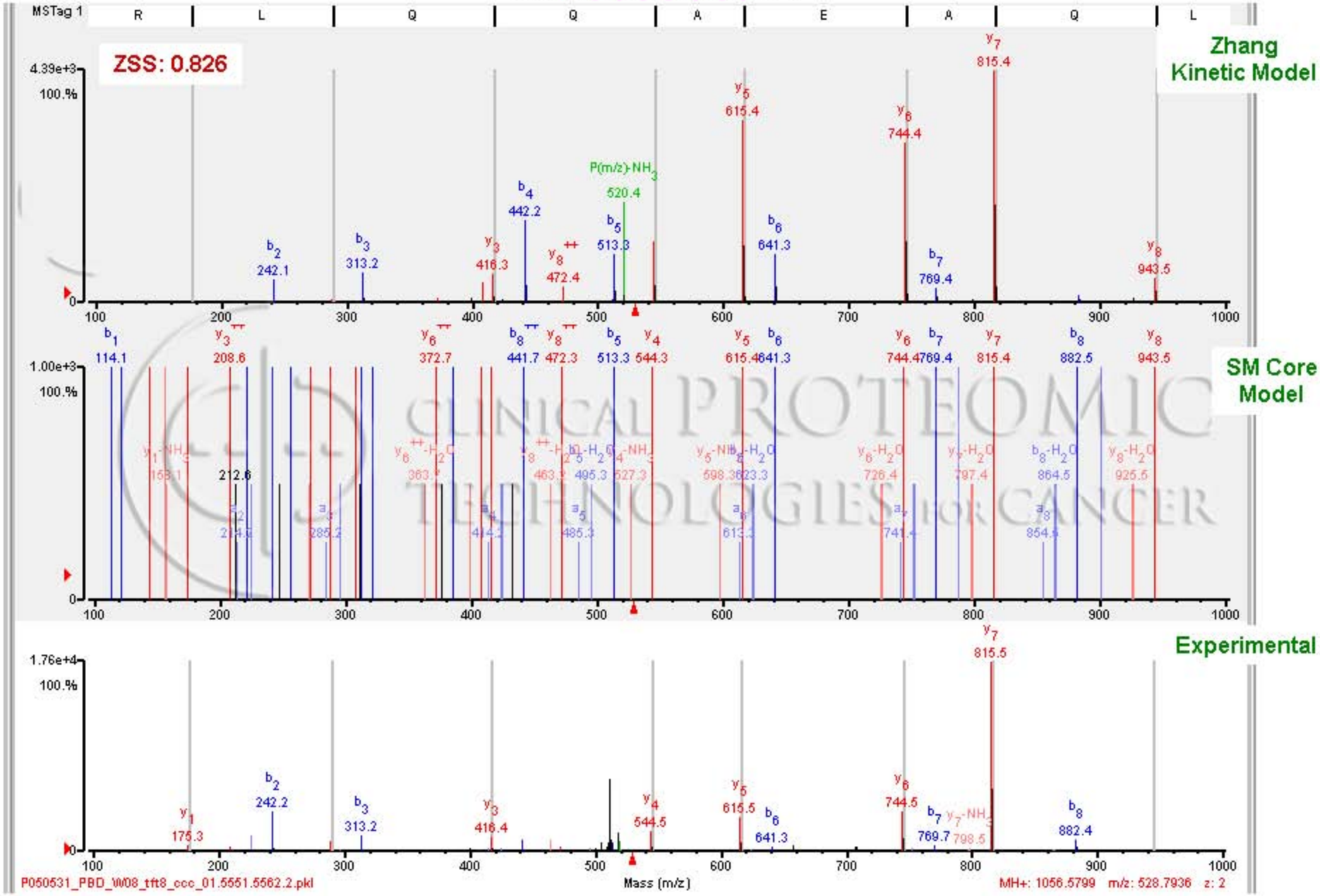
Penalty for Unmatched ions B Theoretically Possible

The naïve fragment ion models in current database search programs tend to substantially over predict the number of fragment ions.

I.e. allow for all possible ion types with similar intensities

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

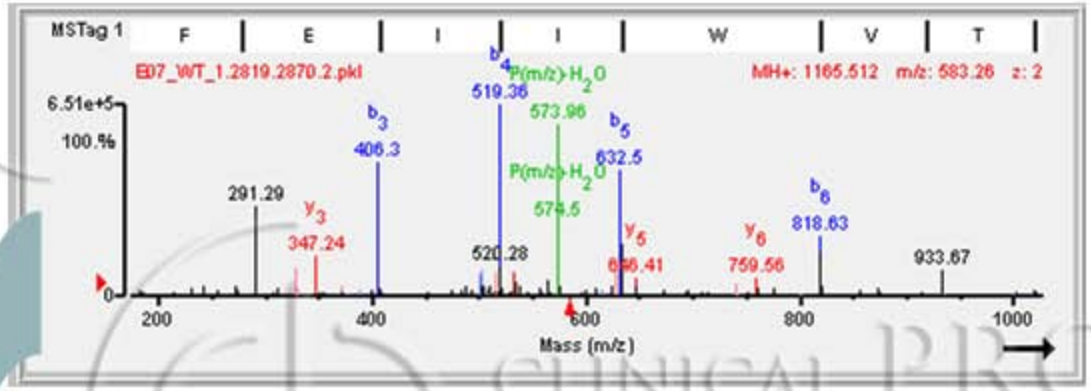
(R)L/Q|A|E|A|Q|Q|L|R(K)



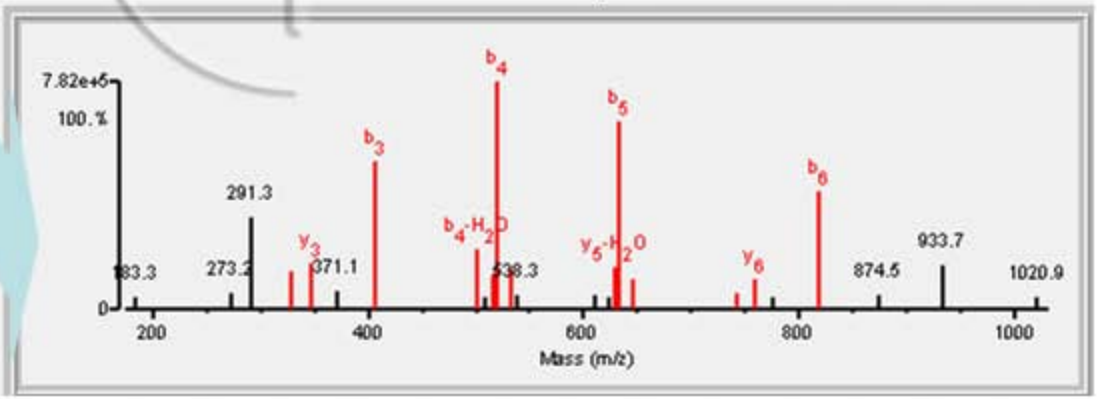
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

SpectrumMill Core Scoring of MS/MS Interpretations

(R)E F E|I|I|W|V T K(H)



Peak Selection: De-Isotoping, S/N thresholding,
Parent - neutral removal, Charge assignment
Match to Database Candidate Sequences



9.24
78.1%
ΔEBS: -3
ΔFwdRev: 1.75

Score
=
Assignment Bonus
(Ion Type Weighted)
-
Non-assignment Penalty
(Intensity Weighted)

SPI (%)
Scored Peak Intensity

Fragment-ion (m/z)	183.27	273.19	291.29	329.16	347.24	371.12	406.30	501.83	533.27	538.29	610.50	624.14	628.26	632.50	646.41	741.46	759.56	776.45	818.63	874.54	933.67	1020.91
Frac. Inten.(% of TIC)	1.04	1.36	7.12	3.01	3.51	1.48	11.31	4.65	3.10	1.11	1.19	0.99	3.24	14.31	2.33	1.27	2.42	1.00	8.96	1.13	3.44	1.03
Rel. Inten.(% of BP)	6.02	7.85	41.15	17.42	20.27	8.57	65.38	26.88	17.90	6.39	6.87	5.71	18.71	82.67	13.44	7.34	13.97	5.76	51.78	6.51	19.90	5.94
Score	-0.06	-0.08	-0.41	0.50	1.00	-0.09	1.00	0.50	1.00	-0.06	-0.07	-0.06	0.50	1.00	1.00	0.50	1.00	-0.06	1.00	-0.07	-0.20	-0.06
Ion-type				y_3-H_2O	y_3		b_3	b_4-H_2O	y_4				y_5-H_2O	b_5	y_5	y_6-H_2O	y_6		b_6			
Delta Da				-0.06	0.01		0.14	0.60	-0.04				-0.12	0.17	0.02	-0.01	0.08		0.22			

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Refining Peptide Fragmentation Models to Account for Dominant Peaks

Goal

- Increase gap between score of correct and incorrect sequences for each spectrum
  - Increase score of correct sequence
  - Decrease score of incorrect sequence

Principles to Capture

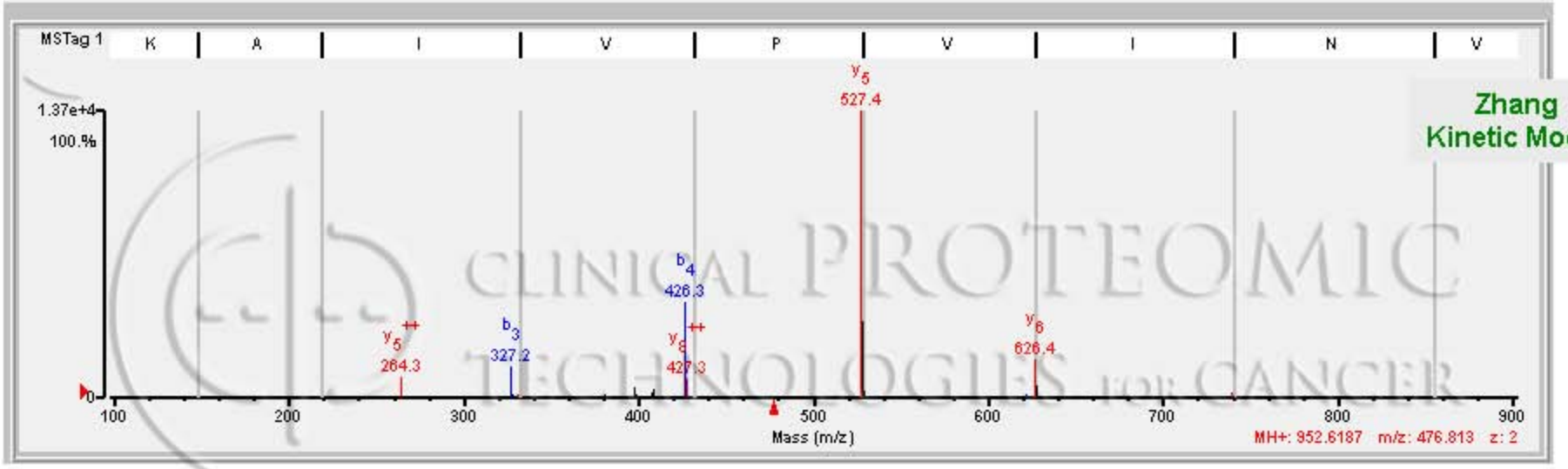
- Bonus - dominant peaks should score a bonus for ion types and AA adjacencies that should be dominant, i.e. at N-term of Pro.
- Penalty - dominant peaks should be penalized if assigned to sites that should not be dominant.
  - Correct peak assignments should rarely be penalized (<1%)
- Charge location - At a dominant site, formation of a b or y type ion depends primarily on the location of basic residues in the peptide, not the fragmentation site.
  - Scoring scheme should still work if an enzyme besides trypsin is used.
  - Do not try to predict which of b, y, b++, y++ will dominate
- Frequency - dominant fragmentation does not *always* occur at particular AA adjacencies.
  - Instead there are tendencies, additional factors not yet understood.
- How dominant - It is very difficult to predict how dominant a peak will be when it occurs at a cleavage that is expected to be dominant.
  - Do not try to predict the extent of dominance.
  - Scale a peak's score with the extent of dominance that occurs.

Dominant Ions – Mobile b2/yn-2

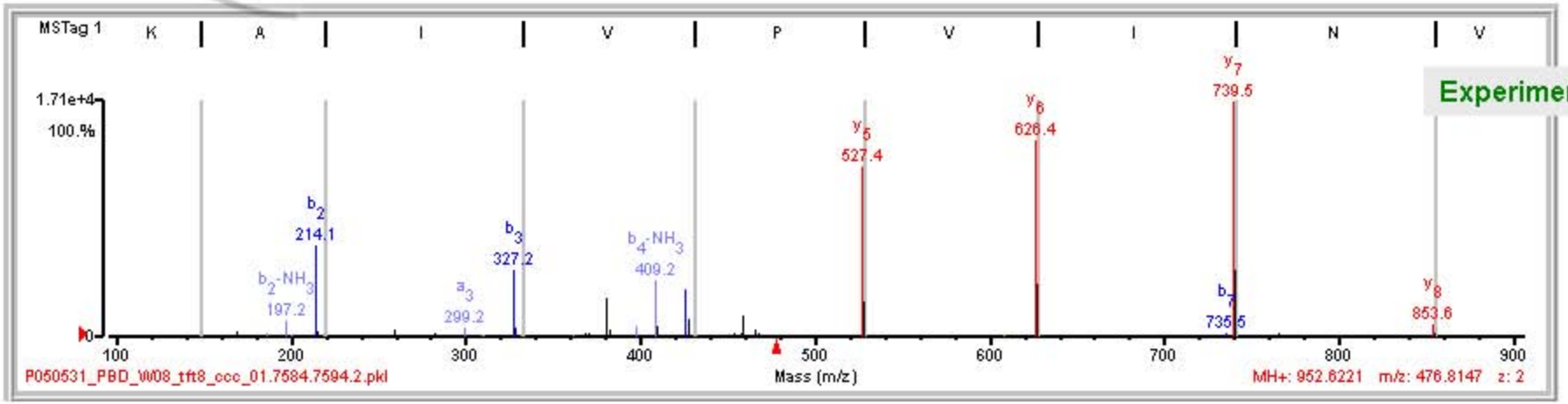
(K)V/N|I|V|P V I\A K(A)

ZSS: 0.697

ptide: VNIVPVIAK instrument: 2 numModeledIons: 64 collision energy: 30 reactionTime: 0.030 isolationWidth: 3.0 resolution: 800 startMass: 95 endMass: 2000



Zhang Kinetic Model



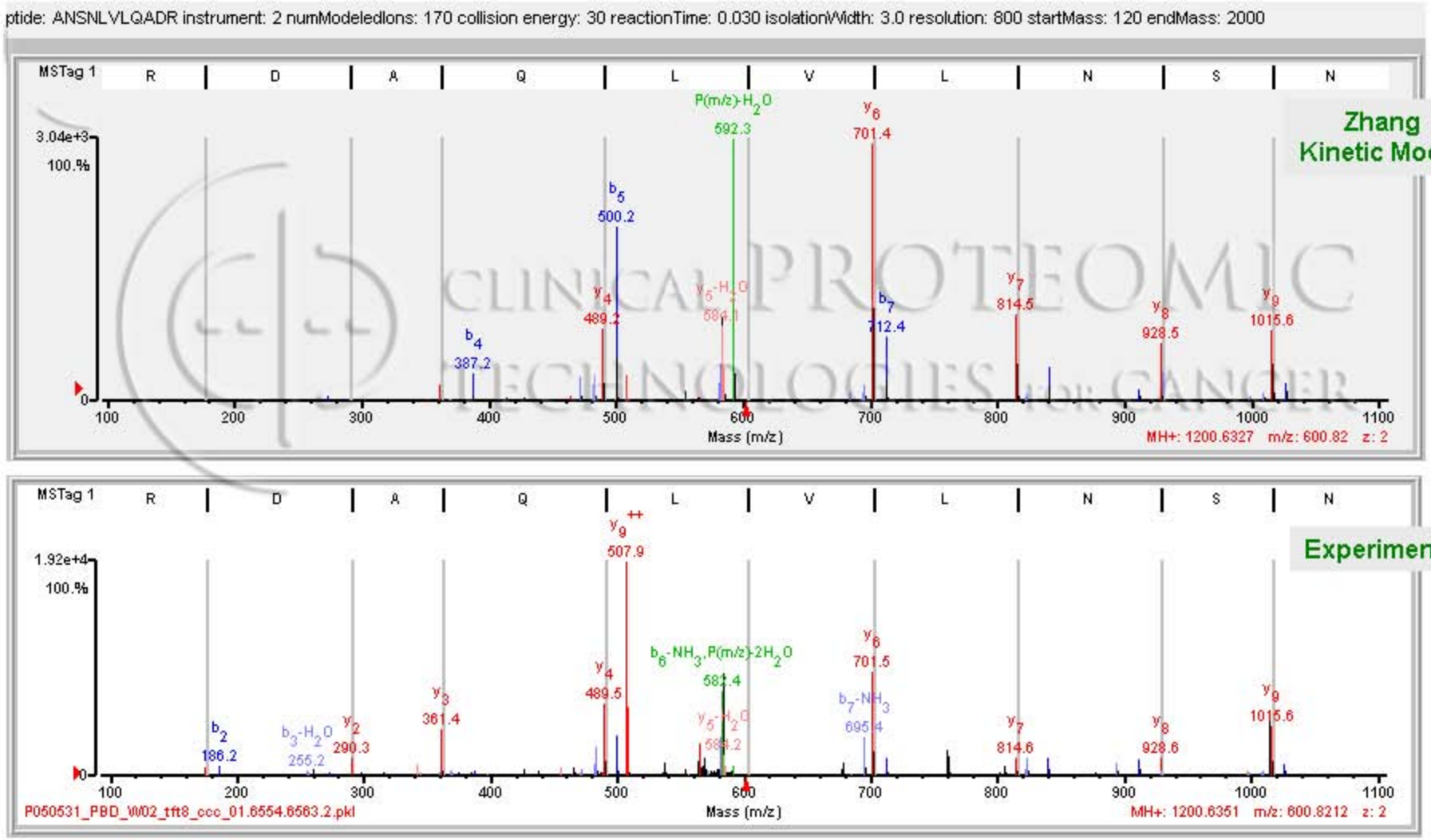
Experimental

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Dominant Ions – Mobile b2/yn-2

(K)A N|S/N/L/V L|Q|A|D\R(S)

ZSS: 0.534

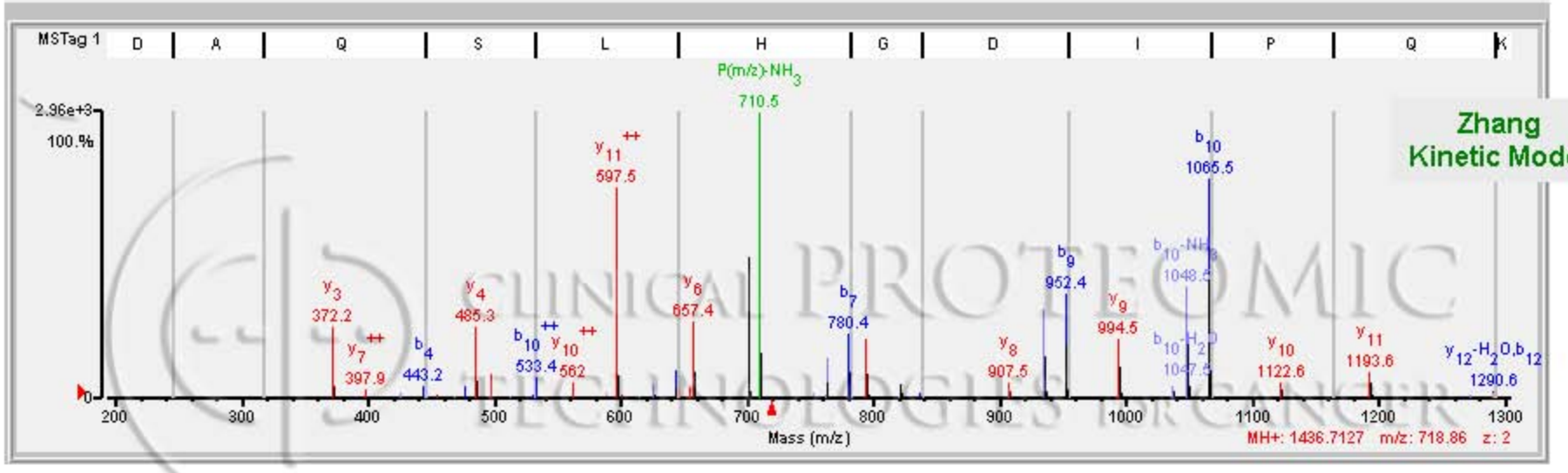


Dominant Ions – PM nterm Q b-nh3

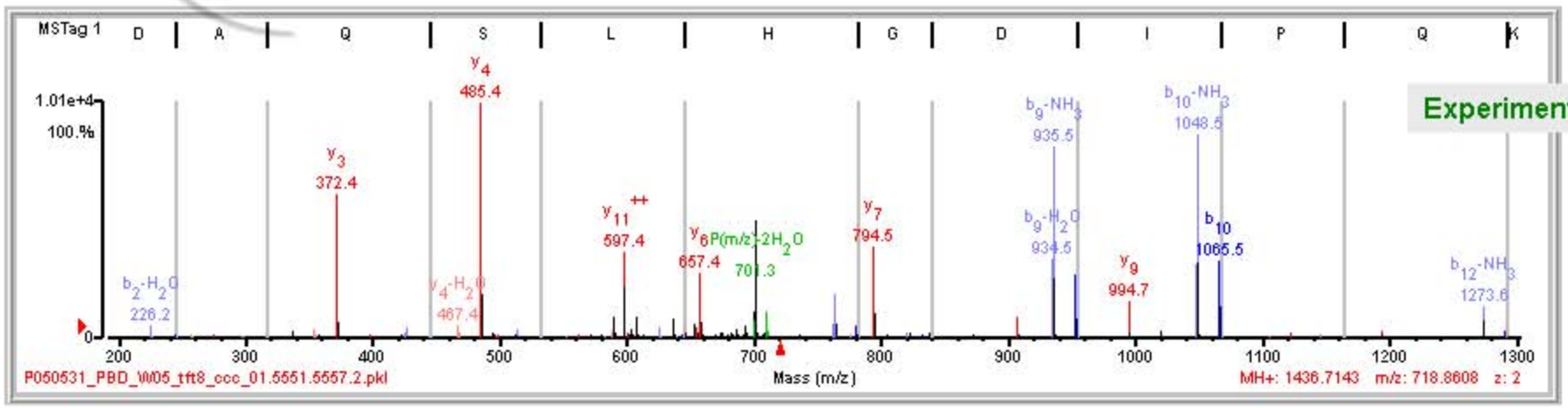
(K)Q D/A Q/S/L/H|G D|I|P Q\K(Q)

ZSS: 0.808

otide: QDAQSLHGDIPQK instrument: 2 numModeledIons: 212 collision energy: 30 reactionTime: 0.030 isolationWidth: 3.0 resolution: 800 startMass: 144 endMass: 2000



Zhang Kinetic Model



Experimental

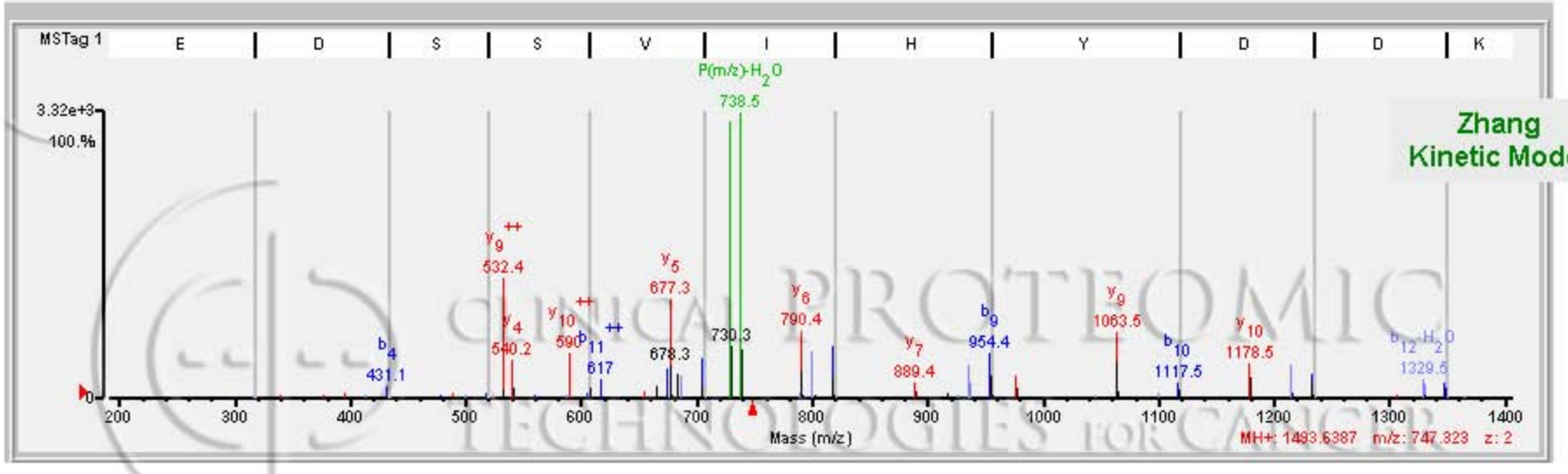
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Dominant Ions – PM nterm E b-h2o

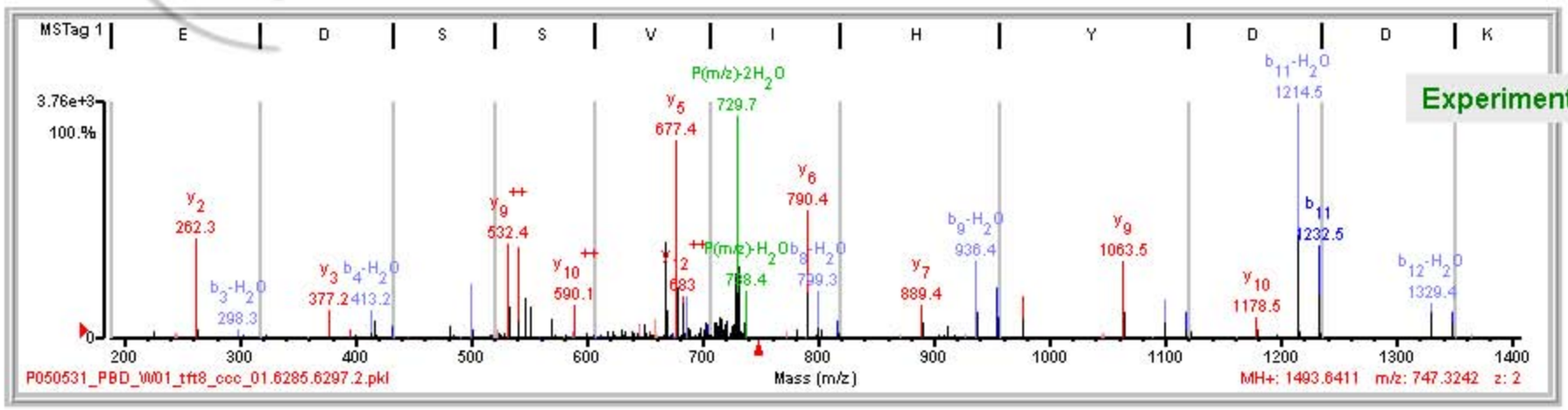
(K)E/G E/D/S/S/V/I/H|Y\D|D\K(A)

ZSS: 0.753

ptide: EGEDSSVIHYDDK instrument: 2 numModeledIons: 217 collision energy: 30 reactionTime: 0.030 isolationWidth: 3.0 resolution: 800 startMass: 149 endMass: 2000



Zhang Kinetic Model



Experimental

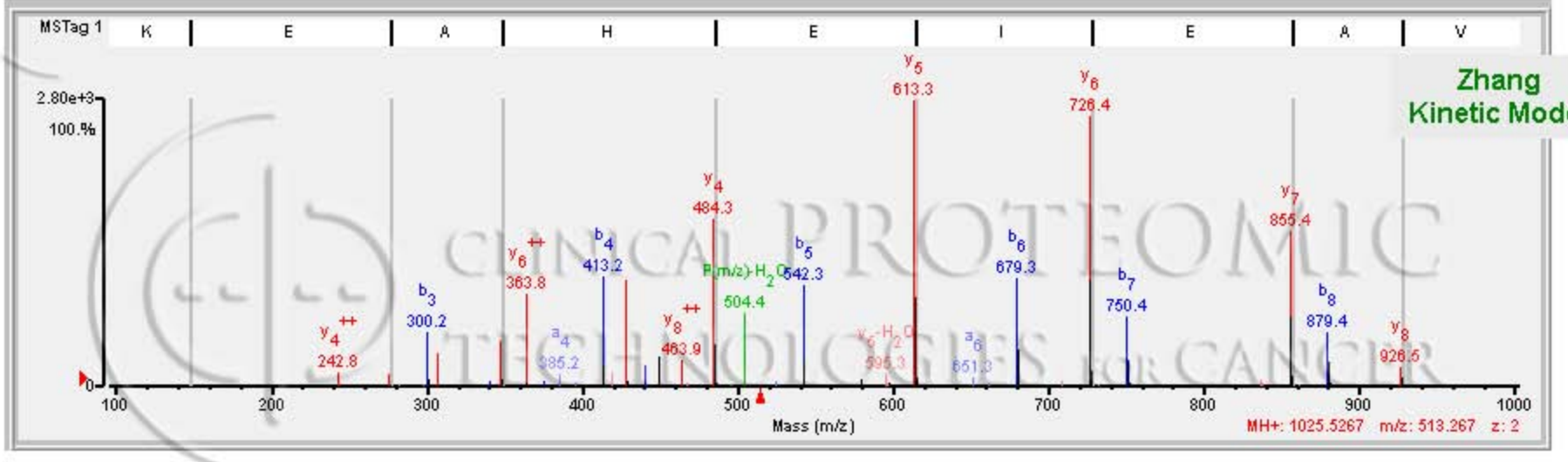
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

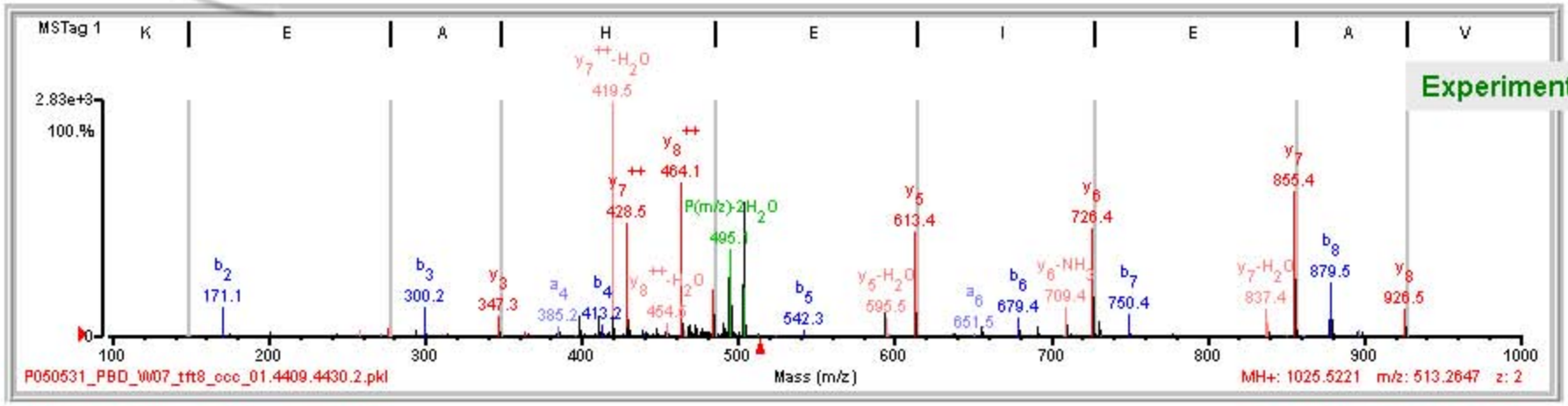
Dominant Ions – PM y++-h2o @ yn-2 DnEST

(K)V/A|E|I/E|H|A\E\K(E)

ZSS: 0.724

ptide: VAEIEHAEK instrument: 2 numModeledIons: 121 collision energy: 30 reactionTime: 0.030 isolationWidth: 3.0 resolution: 800 startMass: 103 endMass: 2000





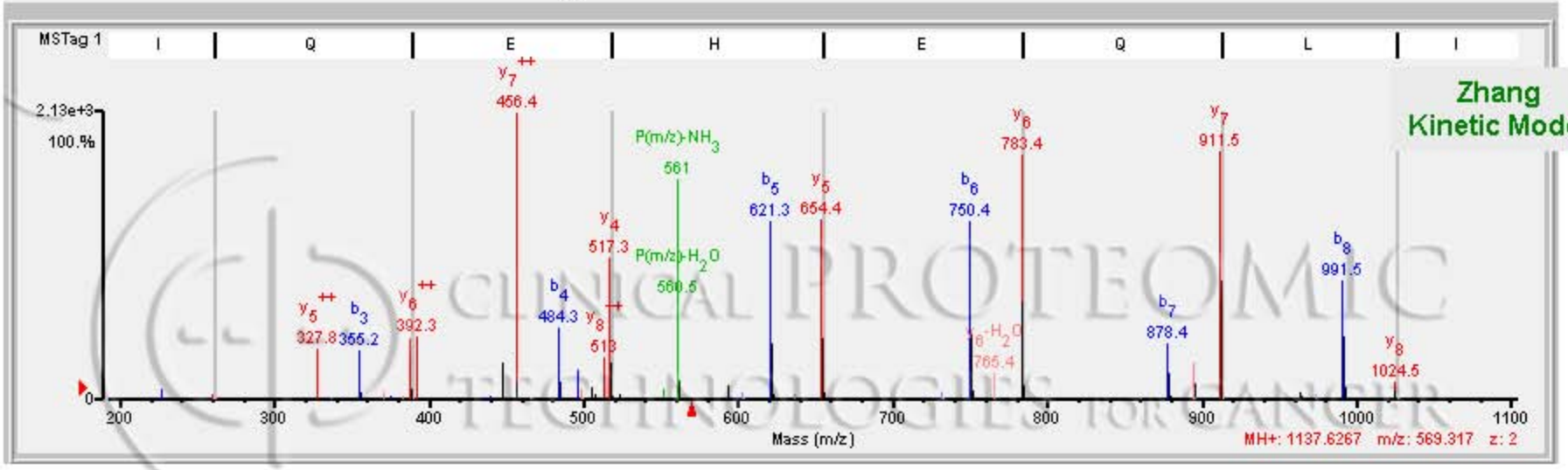
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

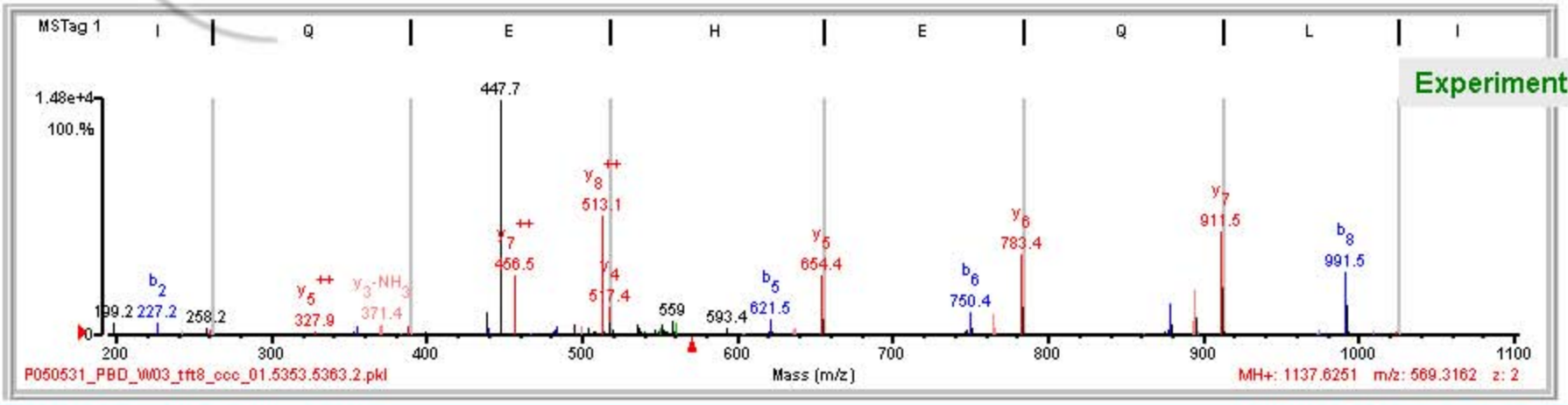
Dominant Ions – PM y++-nh3 @ yn-2 QH

(R)I/L|Q|E/H|E|Q\I\K(K)

ZSS: 0.779

ptide: ILQEHEQIK instrument: 2 numModeledIons: 131 collision energy: 30 reactionTime: 0.030 isolationWidth: 3.0 resolution: 800 startMass: 114 endMass: 2000





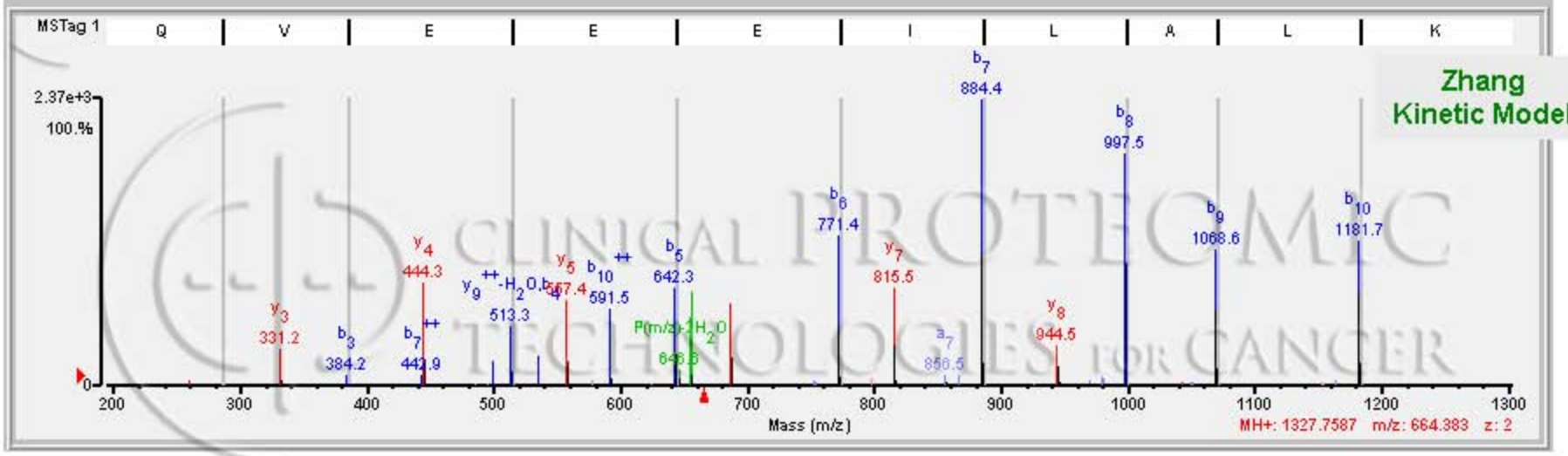
CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

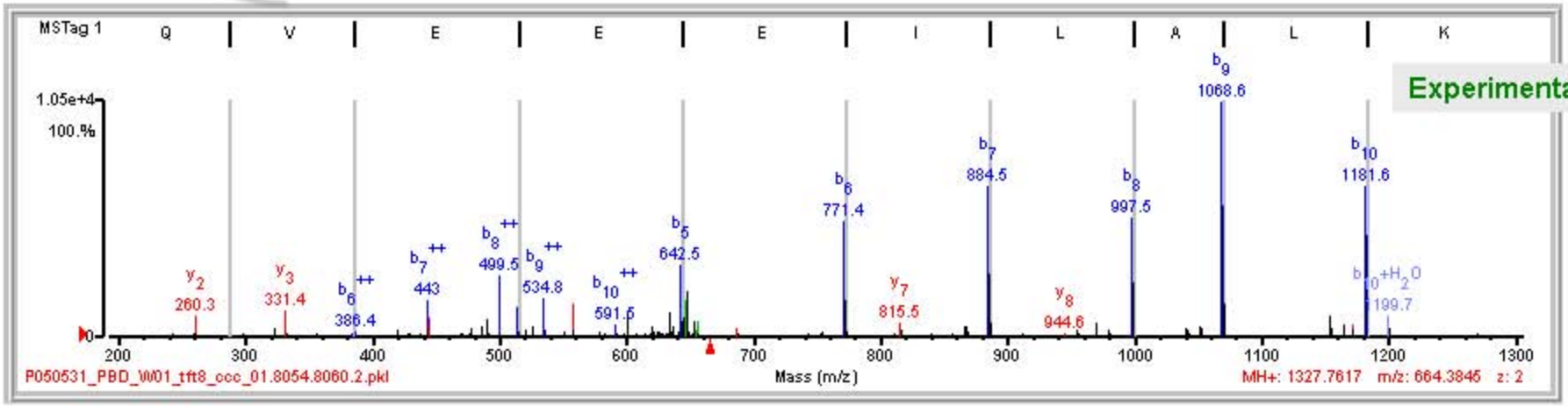
Dominant Ions – PM bn-2 contains RHK

(R)R/Q V E|E\E|I\L|A|L\K(A)

ZSS: 0.799

otide: RQVEEEILALK instrument: 2 numModeledIons: 175 collision energy: 30 reactionTime: 0.030 isolationWidth: 3.0 resolution: 800 startMass: 133 endMass: 2000

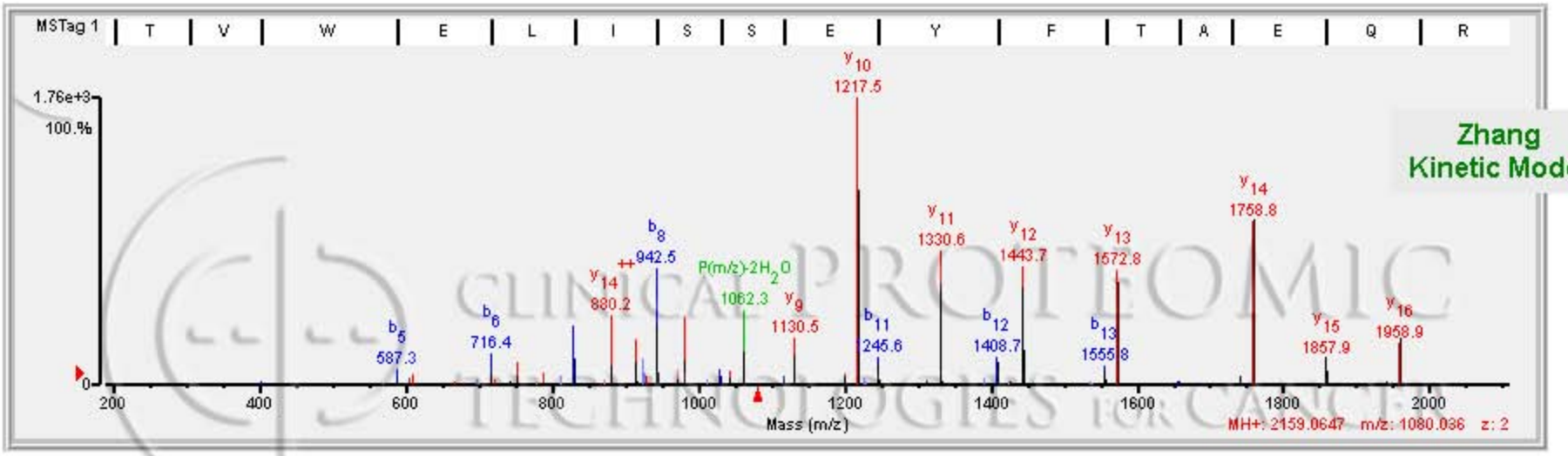


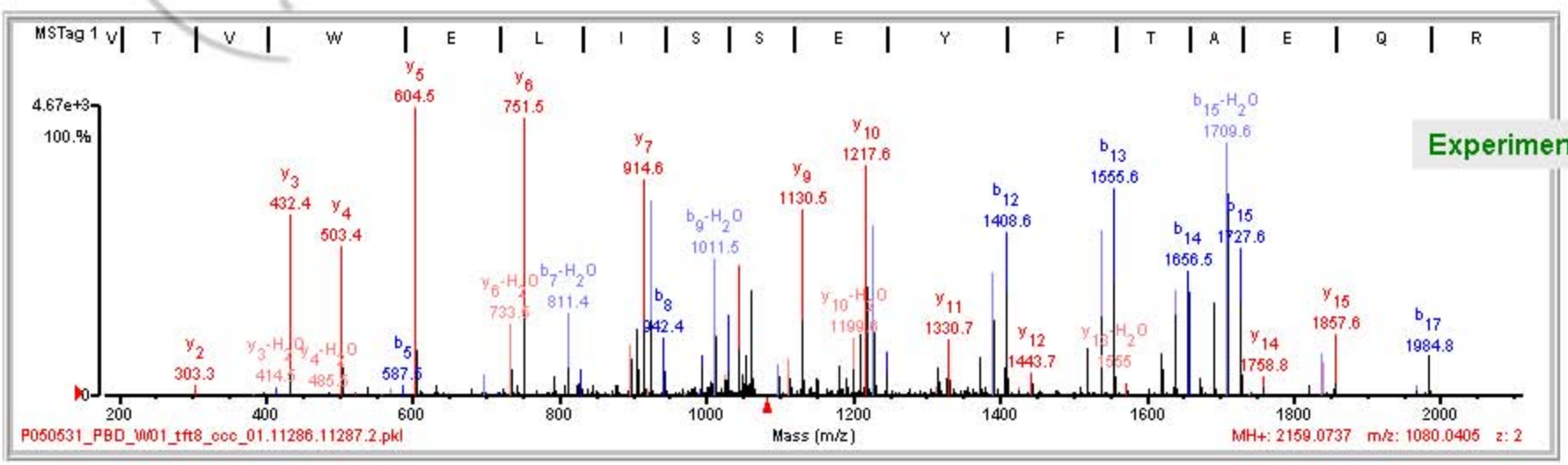


Non-dominant Ions – Mobile

(R)T/V T V W E L I/S|S/E/Y|F|T|A|E\Q R(Q)

ZSS: 0.427





Acknowledgements

Broad Institute

Steve Carr
Jake Jaffe

MIT

Michael Yaffe
Majbrit Hjerrld
Drew Lowery

Agilent Technologies

Joe Roark
David Horn

Amgen

Zhongqi Zhang

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER