ebCTC environmental bioinformatics and Computational Toxicology Center

Optimization Tools for *in silico* Proteomics

Peter A. DiMaggio Jr. and Christodoulos A. Floudas
Department of Chemical Engineering
Princeton University
March 14th, 2007

Computer Aided Systems Laboratory

Interface

- Chemical Engineering
- Applied Mathematics
- Operations Research
- Computer Science
- Computational Chemistry
- Computational Biology

Research Areas

Product & Process Systems Engineering

Computational Biology & Genomics

Unified Theory and Research Philosophy

- address fundamental problems and applications via mathematical modeling of microscopic, mesoscopic and macroscopic level
- rigorous optimization theory and algorithms
- large scale computations in high performance clusters

Themes

Mathematical Modeling, Optimization Theory & Algortihms

Discovery at the Macroscopic Level

Discovery at the Microscopic Level

Computer Aided Systems Laboratory

Discovery at the Microscopic Level

Computational Biology and Genomics

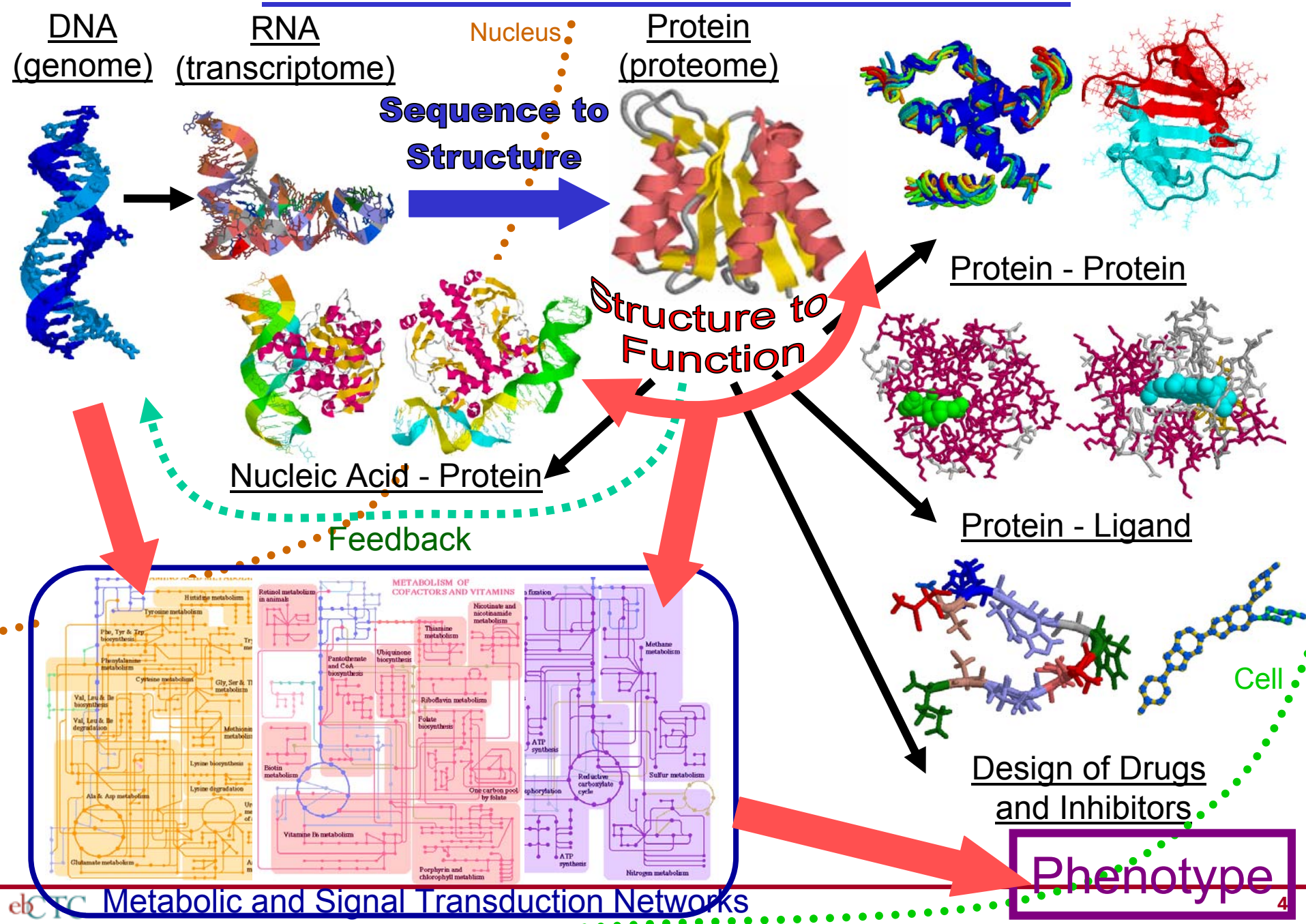
- Structure Prediction in Protein Folding
  - Secondary Structure
  - Tertiary Structure
- Structure Refinement for NMR
- Dynamics of Protein Folding
- Protein-Protein Interactions
- De Novo Protein Design
- Topology of Signal Transduction Networks and Metabolic Pathways
- Proteomics: Peptide & Protein Identification

Computational Tools

ASTRO-FOLD

In Silico ProtDIS

REVOLUTION OF GENOMICS



Computational Biology and Genomics

Structure Prediction in Lennard-Jones Clusters & Acyclic Molecules (90-95)

Structure Prediction in Protein Folding (95-)





Dynamics in Protein Folding (96-00)



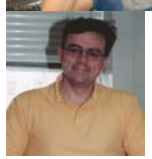
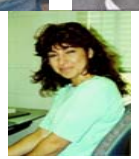
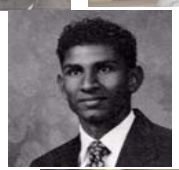
Force Field Development (01-)




De Novo Protein Design (01-)




Protein-Peptide Interactions (95-03)

Metabolic and Signal Transduction Networks (95-)



Proteomics: Peptide & Protein Identification (05-)

Computer Aided Systems Laboratory

Professor Christodoulos A. Floudas – PI

CASL Members:

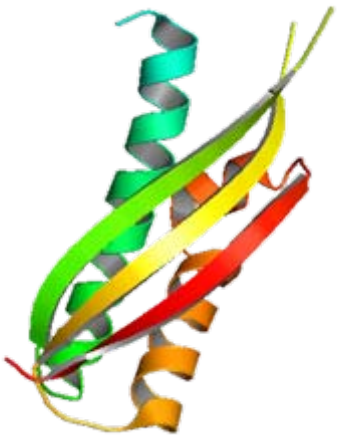
Scott R. McAllister, Ashwin Subramani – protein structure prediction

Ho Ki Fung, Meghan Bellows – *de novo* protein design

Rohit Rajgaria – high-resolution force field development

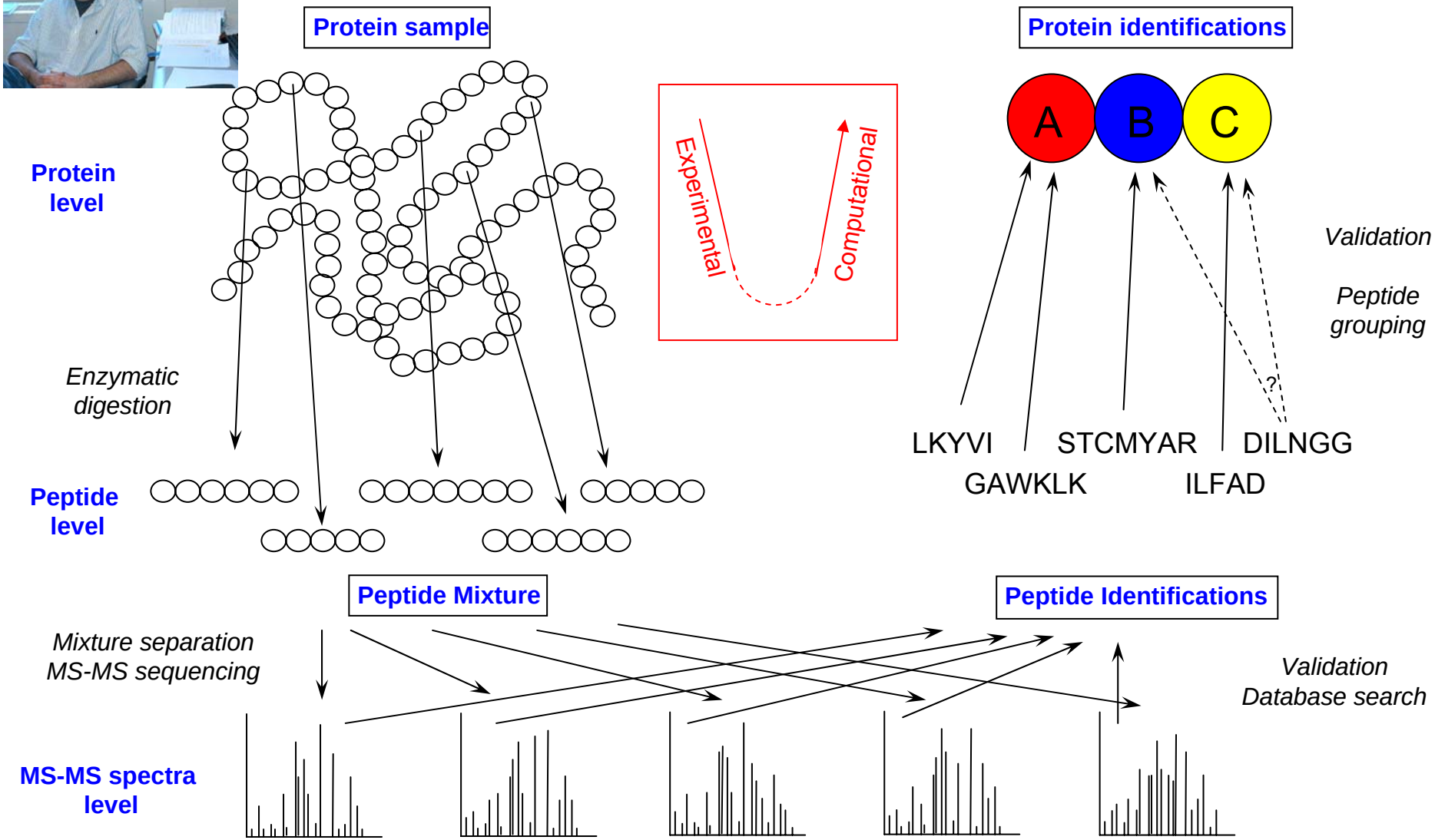
Peter A. DiMaggio – peptide and protein identification

Meng Piao Tan – signal transduction pathways



All components are aimed towards the development of optimization tools for in silico proteomics

Proteomics: Peptide and Protein Identification via Tandem Mass Spectroscopy



Proteomics

- **Specific Aim 1:** Investigate and develop a de novo computational approach for peptide identification based exclusively on information of the ion peaks in the peptide spectrum
- **Specific Aim 2:** Study and develop a new hybrid in silico method which will combine the de novo approach of Specific Aim 1 with database techniques for peptide identification
- **Specific Aim 3:** Incorporate uncertainty into the de novo framework to address experimental uncertainty in problem parameters
- **Specific Aim 4:** Study and develop computational methods for protein identification given the de novo prediction and/or hybrid prediction of the individual peptides
- **Specific Aim 5:** Research and develop computational methods and experimental protocols for protein quantification

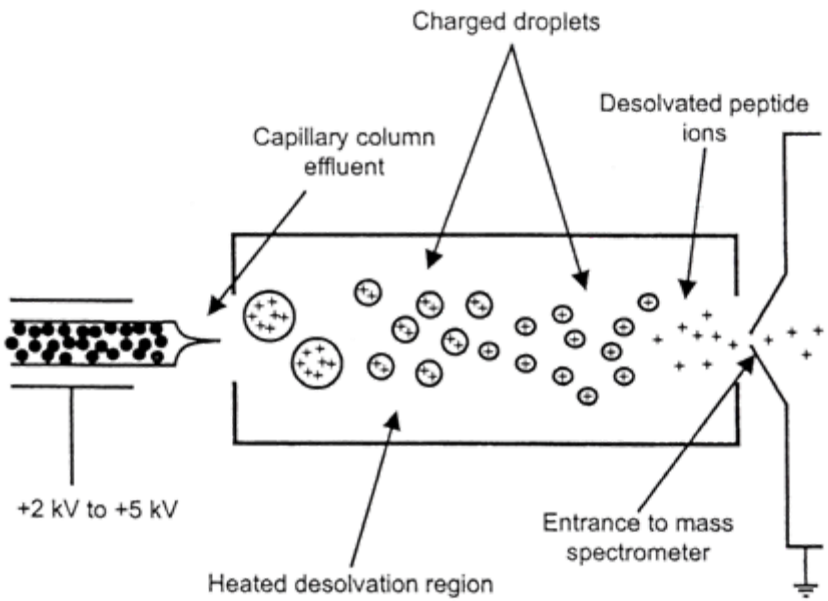
Problem Definition and Introduction

- Fundamental problem in proteomics:

  Protein and peptide identification and quantification

- Advances in high-throughput experimentation

  High-performance liquid chromatography (HPLC) coupled with tandem mass spectrometry (MS/MS)



Mass spec facilities at Princeton University

- Need for rigorous computational tools for peptide/protein identification

Peptide & Protein Identification via Tandem MS

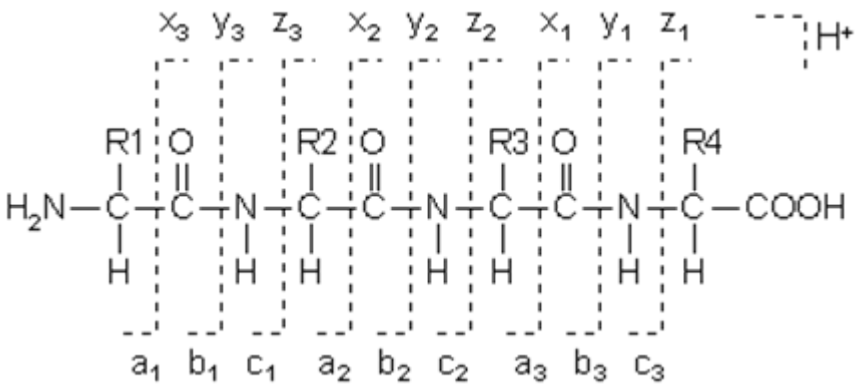
- **Database-based methods**
  - **Correlate the experimental spectra with spectra of peptides/proteins which exist in the databases**
  - SEQUEST – Eng et al. (1994), Mascot – Perkins et. al (1999), SCOPE – Bafna and Edwards (2001), MS-CONVOLUTION and MS-ALIGNMENT – Pevzner et. al (2001), Poptiam – Hernandez et. al (2003)
- **De Novo Methods**
  - **Predict peptides without sequence databases**
  - **Exhaustive listing; sub-sequencing; graphical**
  - **Graph theory and shortest path algorithms**
  - **Graph theory and dynamic programming**
  - **Bayesian scoring of random peptides**
  - Lutefisk – Taylor and Johnson (1997,2001), SHERENGA – Dancik et. al (1999), PEAKS – Ma et al. (2003), NovoHMM – Fischer et al. (2005), PepNovo – Frank and Pevzner (2005), EigenMS – Bern and Goldberg (2006)

Challenges

- Tandem MS are missing ion peaks due to incomplete fragmentation and/or instruments with low mass-to-charge ratio (m/z) cutoff (i.e., ion trap mass analyzers)
- Incorporating parametric uncertainty in the measured values for ion peaks during peptide identification
- Existing de novo techniques enumerate an exhaustive number of candidate sequences from the tandem mass spectrum
- No straightforward method for including post-translational modifications into existing frameworks

Tandem MS/MS

- Collision-induced dissociation (CID) causes a positively-charged peptide to fragment along its backbone and results in many types of fragment ions in the tandem mass spectrum (i.e., a, b, c, x, y, x, etc.)



Hypothetical parent peptide*

- Objective:

  Use these fragment ions to predict the amino acid sequence of the parent peptide

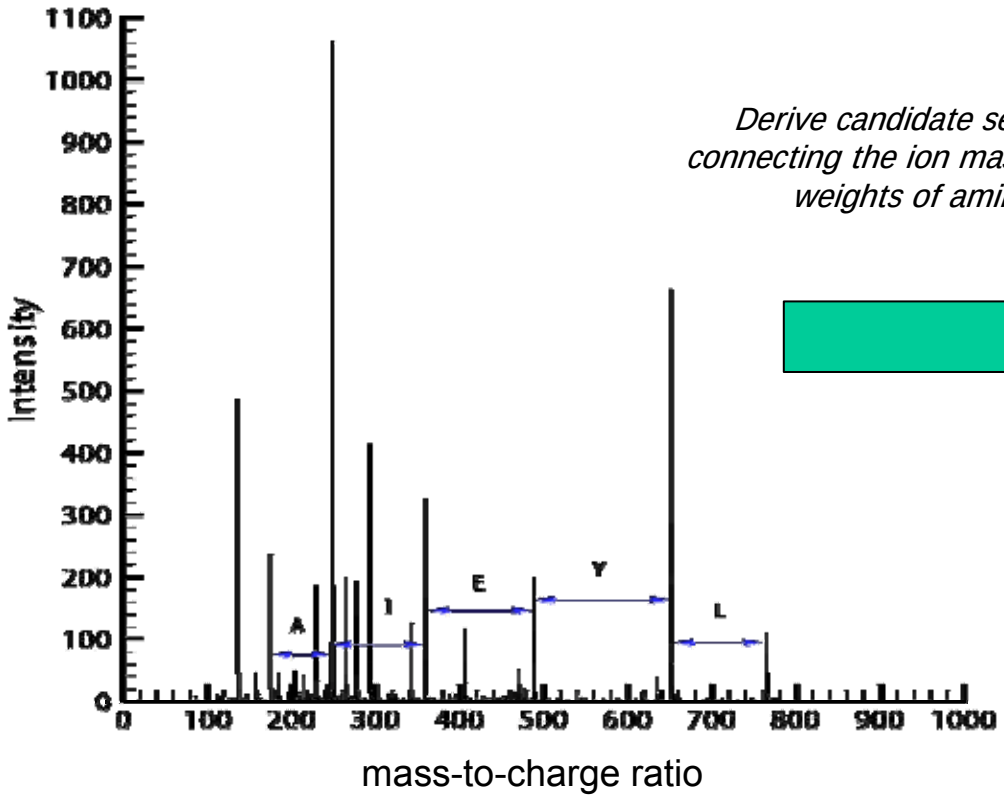
- Issues: Identifying ion types in the mass spectrum is nontrivial

*Adapted from http://www.matrixscience.com/help/fragmentation_help.html

Introduction to De Novo Peptide Identification

The De Novo Peptide Identification Problem:

Given the tandem mass spectrum (MS/MS) of a peptide, derive the primary sequence of the peptide without consulting other sources of information (i.e., protein databases)



Derive candidate sequences by connecting the ion mass peaks by the weights of amino acids

Q: Which of these possible primary sequences corresponds to the correct peptide?

YLYKNAR

YLFPMTR

YLYELAR

YFEELAR

YEYLLAR

YLYKKGR

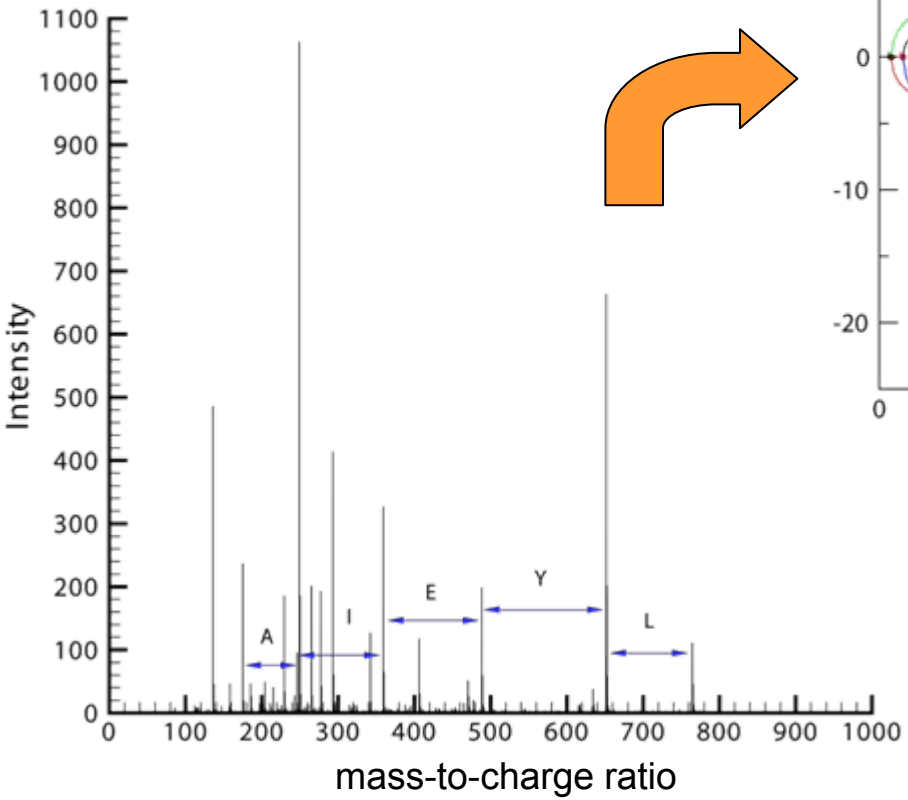
YFEKNAR

YLY[171.06]AAR

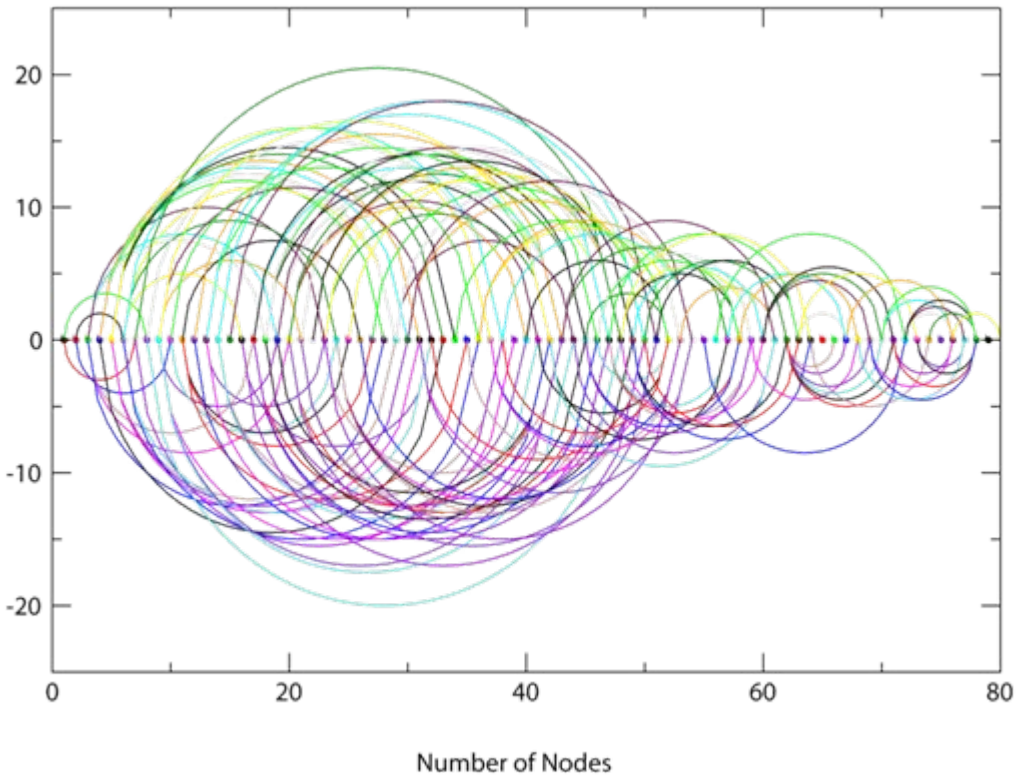
Traditional De Novo Methods

Transform tandem MS/MS into a spectrum graph, where:

paths on the graph = amino acid sequences



Spectrum Graph Approach*



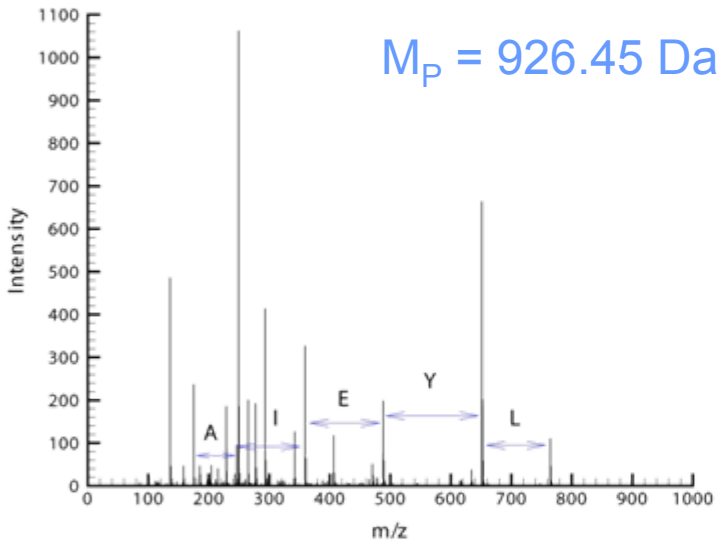
- ❑ Solve via dynamic programming
- ❑ Nodes assigned probabilistic weights
- ❑ Highest scoring path is selected

* Taylor and Johnson (1997,2001), Dancik et. al (1999), Fernandez de Cossio et. al (2000), Chen et. al (2001), Lubeck et. al (2002), Cannon and Jarman (2003), Chen and Bingwen (2003), Jarman et. al (2003), Frank and Pevzner (2005), Bern and Goldberg (2006)

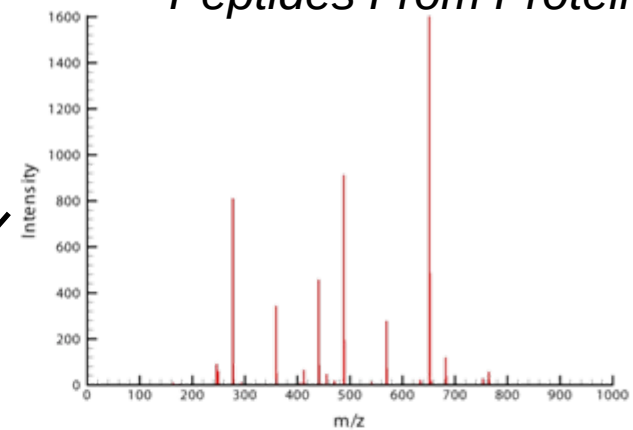
Database Methods

Peptides From Protein Database

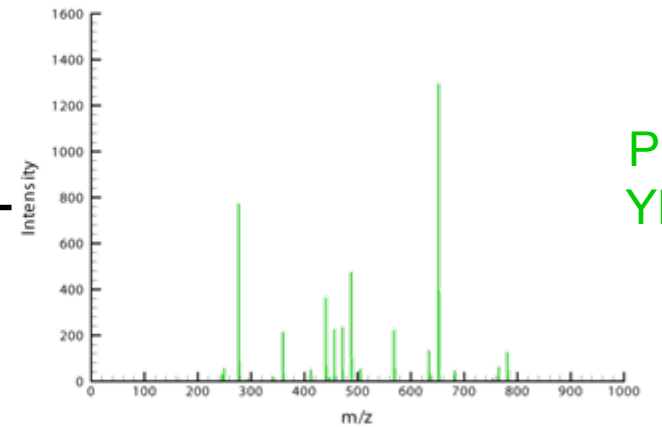
Raw Tandem MS/MS for YLYEIAR



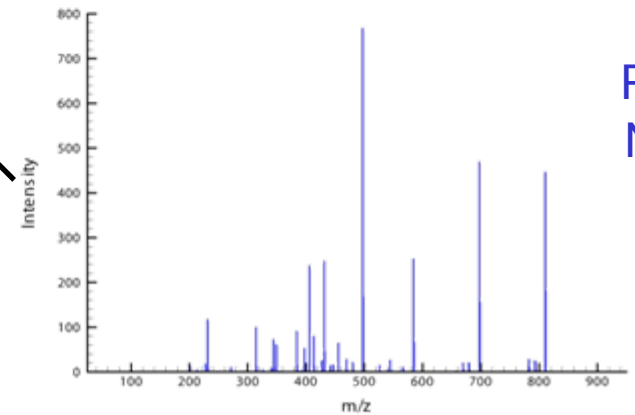
M_P = 926.45 Da



Predicted* YLYEIAR



Predicted* YLYQNVK



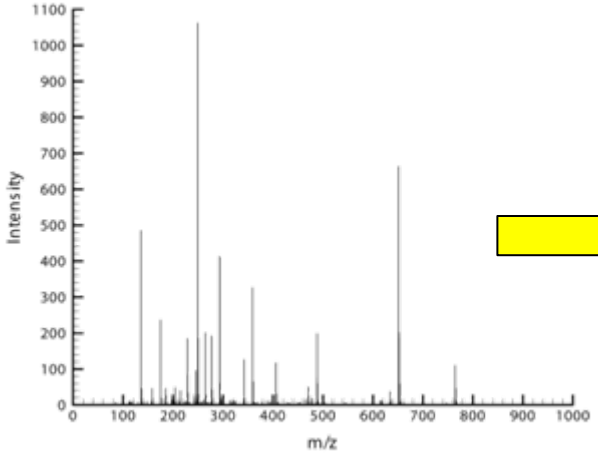
Predicted* NRIISLLV

Which *predicted* spectrum matches the experimental spectrum under question?

*Predicted spectra generated with MassAnalyzer 1.03

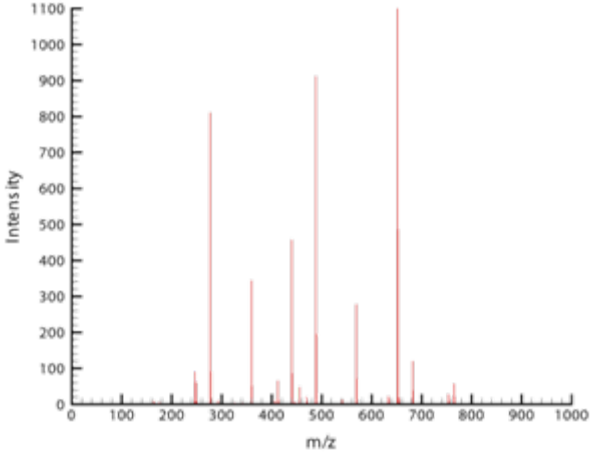
Database Methods (cont'd)

- **Cross-Correlation** (e.g., SEQUEST*)

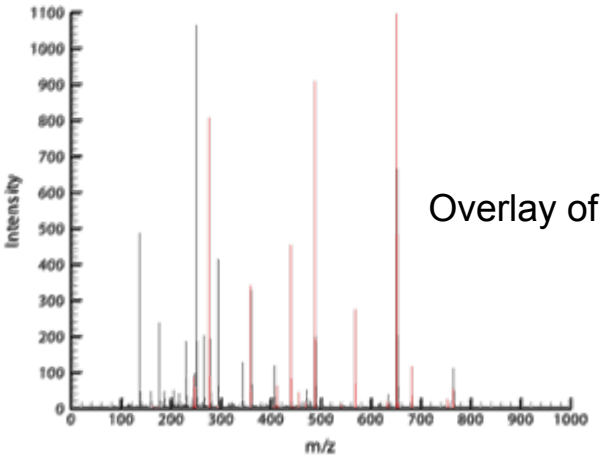


Experimental Spectrum → x

Determine "mathematical overlap"



Predicted Spectrum → y



Overlay of x & y

$$R_\tau = \sum_{i=0}^{n-1} x[\,i\,]\, y[\,i + \boxed{\tau}\,]$$

Displacement value

$$R_\tau \leftrightarrow \boxed{X_r Y_r^*}$$

Discrete Fourier Transforms

*Eng et. al (1994)

Database Methods (cont'd)

- **Probabilistic Matching*** (e.g., Mascot, SCOPE)

Predict primarily y- and b-ions, and their offsets, based on the following formulae:

$$b_i = \sum_{j=1}^{i} mass(AA_j) + 1 \qquad y_{n-i} = 19 + \sum_{j=i+1}^{n} mass(AA_j)$$

Q: Is ion match with experimental spectrum → Actual? / Random?

"A":

- Likelihood ratio hypothesis test (Bafna and Edwards (2001), Havilio et. al (2003))
- Null hypothesis (Sadygov and Yates (2003))
- Integration of spectral dependencies into model (Bafna and Edwards (2001), Havilio et. al (2003))
- Empirically estimated probabilities

*Perkins et. al (1999), Bafna and Edwards (2001), Pevzner et. al (2001), Havilio et. al (2003), Hernandez et. al (2003), Sadygov and Yates (2003)

Drawbacks of Existing Methods

❑ De Novo Methods

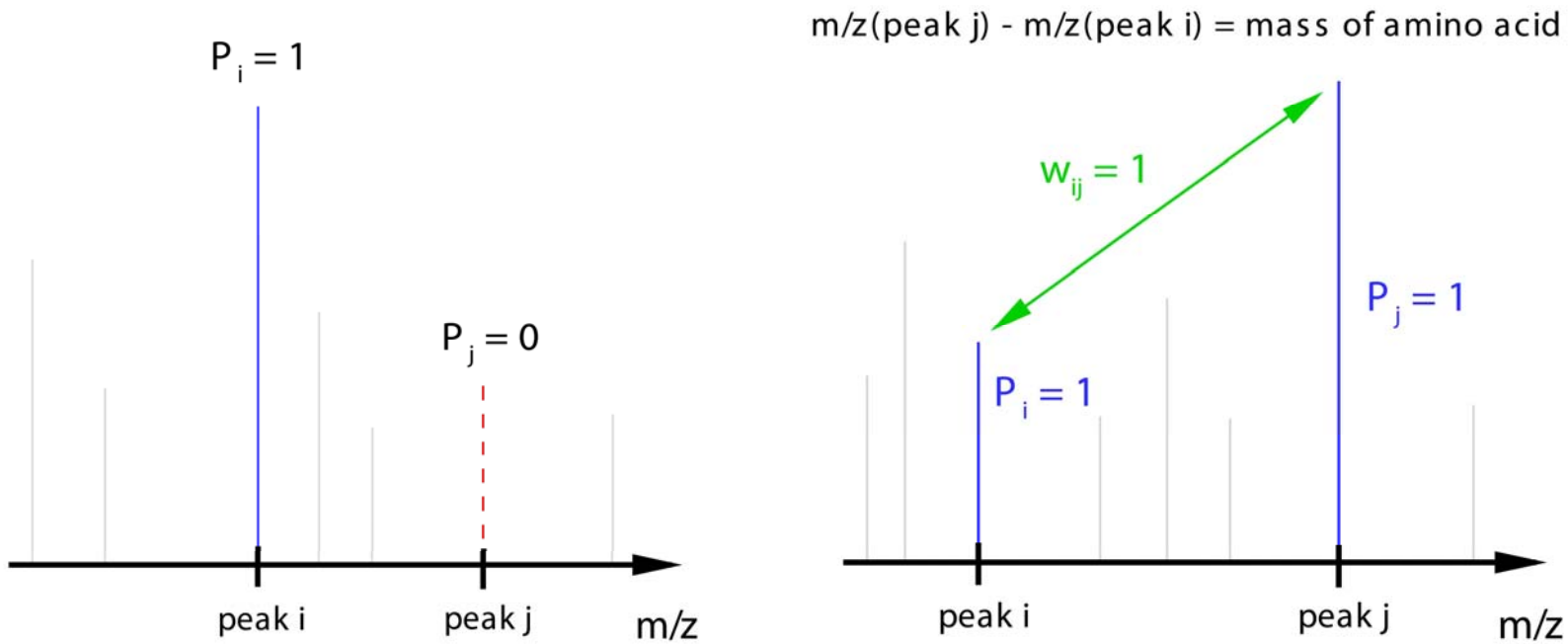
- Exhibit variable prediction accuracies
- Computationally intensive → exhaustive enumeration
- Many are instrument dependent

❑ Database Methods

- False predictions if missing protein in database
- Difficult to identify post-translational modifications / mutations
- Often exhibit dependencies on training data sets and databases

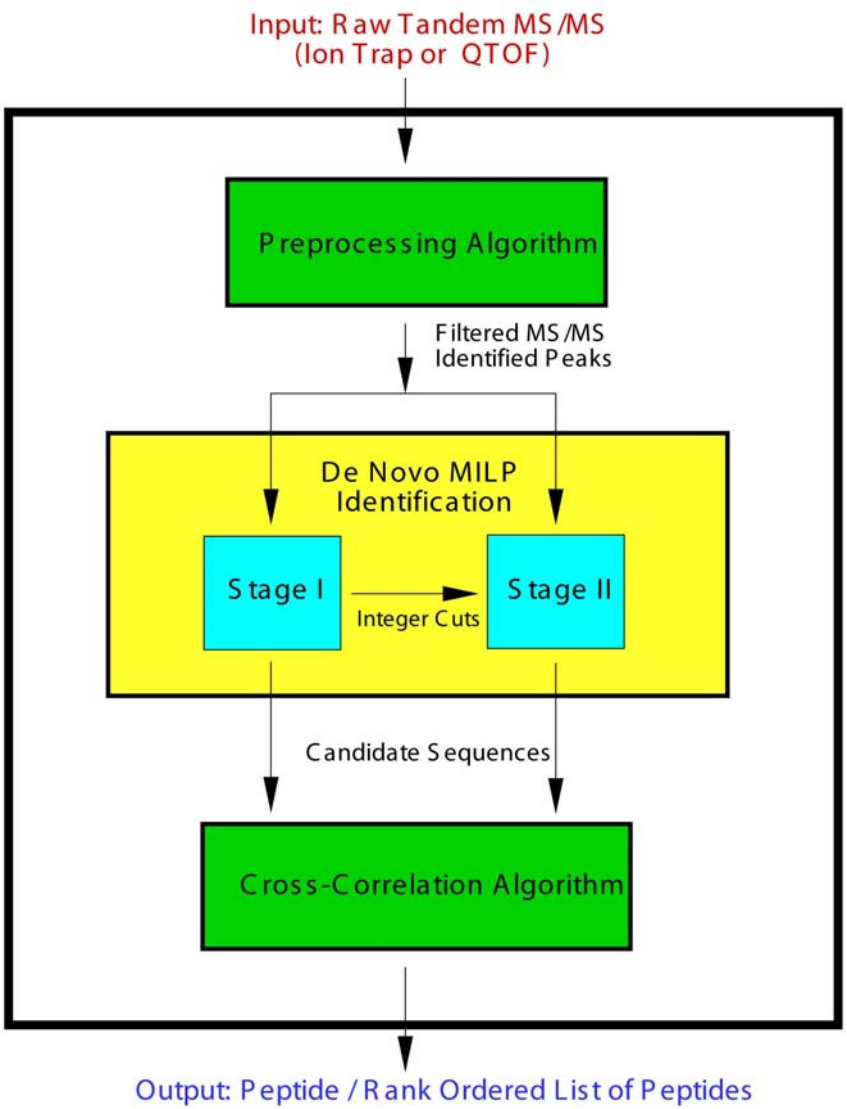
Our Approach to Solve Peptide the Identification Problem

Novel Technique: Using Mixed-Integer Linear Optimization (MILP) to formulate the peptide sequencing problem



Binary variables {0-1 variables} define whether or not **peaks** (p_i) and **paths** between peaks (w_{ij}) are used in the construction of the candidate sequence, where **1** indicates **yes** and **0** indicates **no**

Algorithmic Overview



Components of Framework:

I. Preprocessing of Tandem MS Data

II. Mathematical Model for Peptide Identification

III. Postprocessing of Candidate Sequences

I. Preprocessing Algorithm

- Determine the boundary condition (BC^{tail}) for the N-terminus of the y-ion series

- For tryptic peptides,

  C-terminus amino acid is

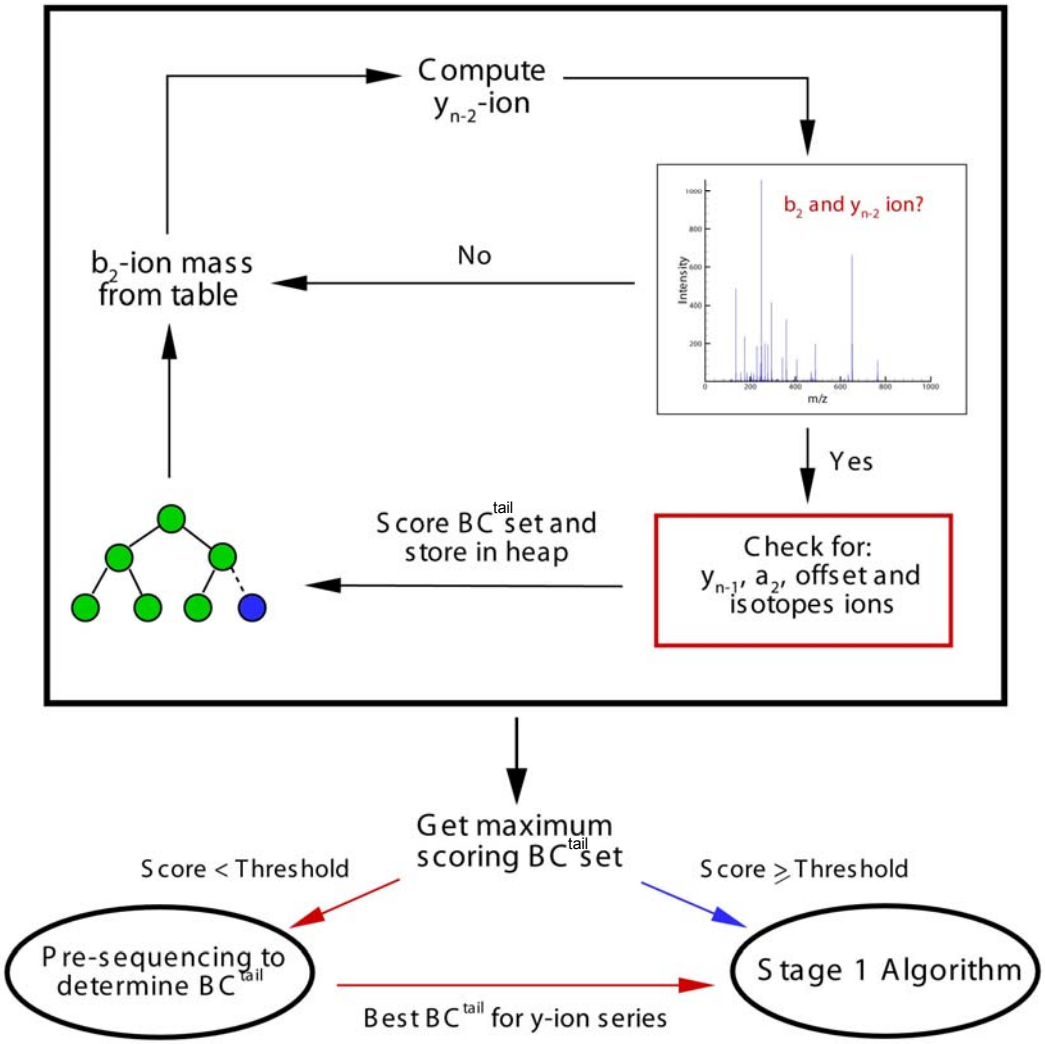
  K → 147 Da

  R → 175 Da

- Identify multiply-charged ions
  - High-resolution instrument?
  - Measure distance between isotopes

- Identify neutral losses of small molecules

  i.e., -H2O, -NH3, etc.



Threshold = 0.1* Maximum intensity peak in spectrum

II. Mathematical Model: Objective Function

$$\boxed{\underset{p_k,\, w_{i,j}}{MAX} \sum_{(i,j)\in S_{i,j}} \lambda_j \cdot w_{i,j}}$$

λ_j = intensity of ion peak j

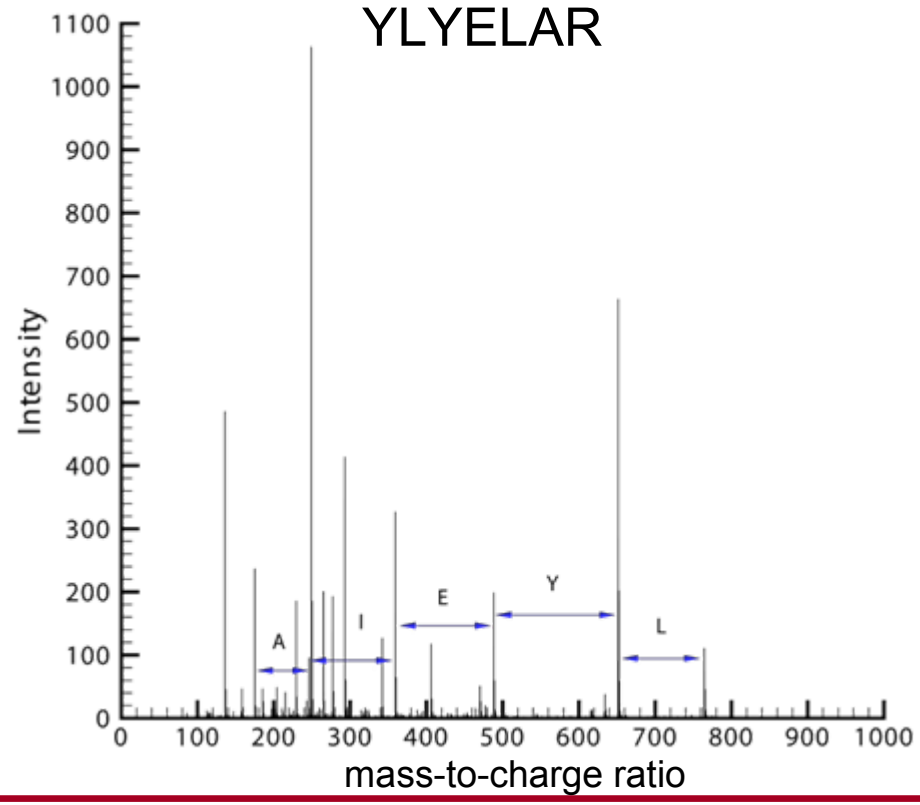
$$S_{i,j} = (i,j) : m/z(\text{ion peak j}) - m/z(\text{ion peak i}) = mass(\text{amino acid})$$

$$p_i = \begin{cases} 1, & \text{if peak (i) is selected} \\ 0, & \text{otherwise} \end{cases}$$

$$w_{i,j} = \begin{cases} 1, & \text{if peaks (i) and (j) are connected} \\ & \text{by a path} \quad (\text{i.e., } p_i = p_j = 1) \\ 0, & \text{otherwise} \end{cases}$$

- Maximize the use of *high intensity peaks* in constructing the candidate sequence
- Based on the observation that y- and b-ions are consistently the most abundant peaks in intensity in MS/MS

Illustration using the y-ion series for YLYELAR



II. Mathematical Model: Constraints

Conservation of Mass

$$\sum_{(i,j)\in S_{i,j}} M_{i,j} \cdot w_{i,j} \leq m_P + tolerance$$

$$\sum_{(i,j)\in S_{i,j}} M_{i,j} \cdot w_{i,j} \geq m_P - tolerance$$

tolerance "relaxes" equality

Boundary Conditions (BC)

$$\sum_{i\in BC_i^{head}} \sum_{j\in S_{i,j}} w_{i,j} = 1$$

$$\sum_{j\in BC_j^{tail}} \sum_{i\in S_{i,j}} w_{i,j} = 1$$

- *BC* elements are dependent on ion type
- *BC* elements are checked in a preprocessing algorithm
- If elements missing then BC set is adjusted

Complementary Ions

$$p_i + p_j \leq 1 \qquad \forall\,(i,j) \in C_{i,j}$$

b ⟷ y

a ⟷ x

c ⟷ z

Eliminates different ions of different type

II. Mathematical Model: MILP

$$\underset{p_k,\, w_{i,j}}{MAX} \sum_{(i,j)\in S_{i,j}} \lambda_j \cdot w_{i,j}$$

$$s.t. \quad \sum_{(i,j)\in S_{i,j}} M_{i,j} \cdot w_{i,j} \leq m_P + tolerance$$

$$\sum_{(i,j)\in S_{i,j}} M_{i,j} \cdot w_{i,j} \geq m_P - tolerance$$

$$p_i + p_j \leq 1 \qquad \forall\,(i,j) \in C_{i,j}$$

Relationship between p_i & $w_{i,j}$

$$\sum_{j\in S_{i,j}} w_{i,j} = p_i \qquad \forall i \in BC_i^{head}$$

$$\sum_{j\in S_{i,j}} w_{j,i} = p_i \qquad \forall i \notin BC_i^{head}$$

$$\sum_{i\in BC_i^{head}} \sum_{j\in S_{i,j}} w_{i,j} = 1$$

$$\sum_{j\in BC_j^{tail}} \sum_{i\in S_{i,j}} w_{i,j} = 1$$

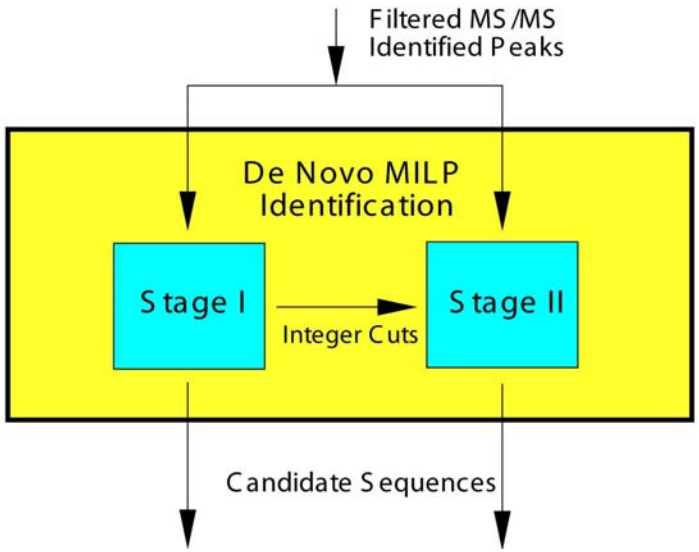
Flow conservation law

$$\sum_{j\in S_{j,i}} w_{j,i} - \sum_{k\in S_{i,k}} w_{i,k} = 0 \qquad \forall i,\ i \notin BC_i^{head},\ i \notin BC_i^{tail}$$

$$w_{i,j},\ p_k = 0 - 1 \qquad \forall (i,j),(k)$$

P.A. DiMaggio and C.A. Floudas, AIChE Journal, 53(1), 160-173 (2007).

II. Two-Stage Framework



- During the Stage I calculations, the derivation of the candidate sequence is done using only single amino acid weights

- It is common that tandem MS are missing ion peaks due to incomplete fragmentation and/or instruments with low m/z cutoff (i.e., ion trap mass analyzers)

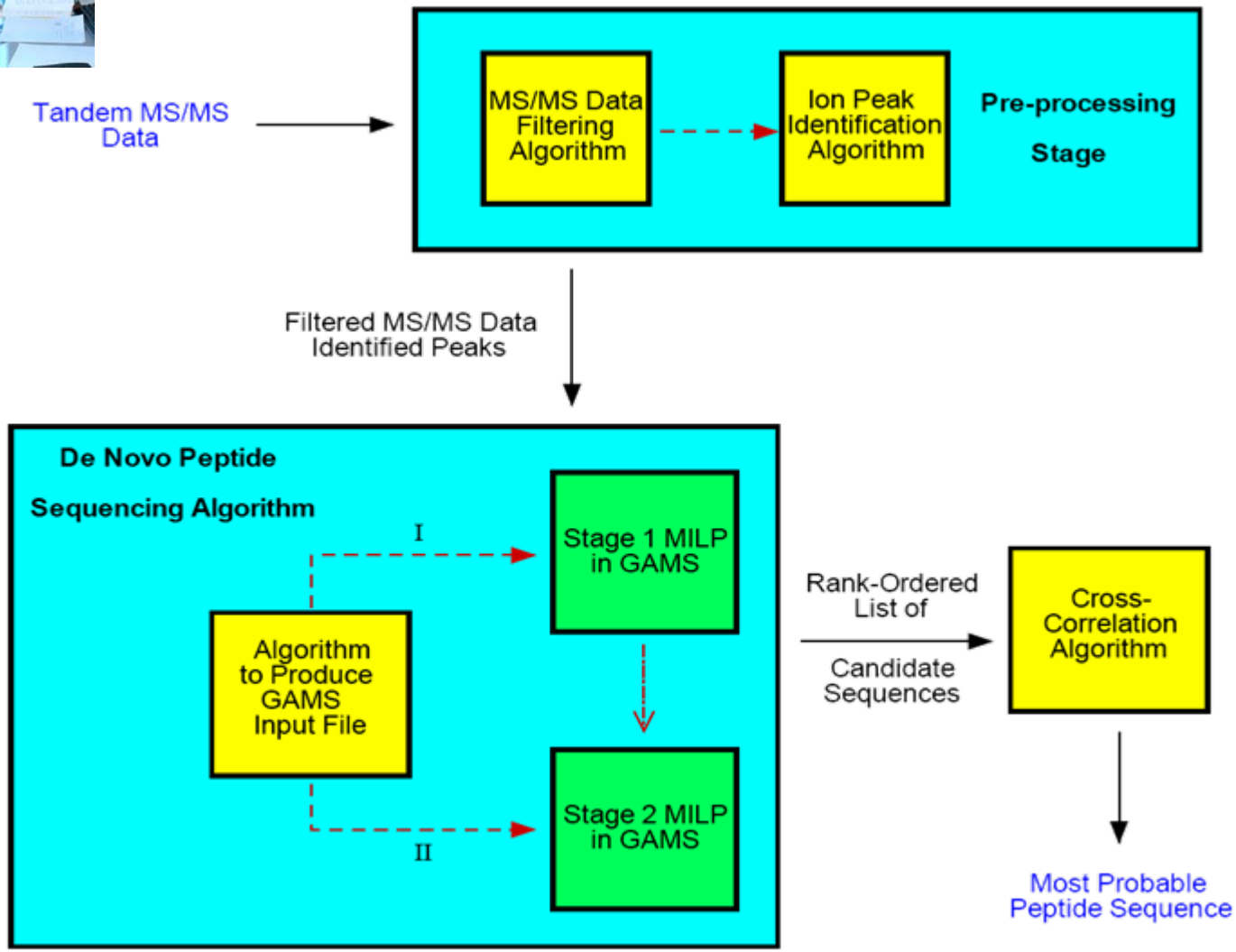
- Stage II calculations allow for combinations of amino acids to connect ion peaks

- Combinations of amino acids are penalized in objective function to bias use of single amino acid weights in derivation of candidate sequences

III. Post-Processing Algorithm

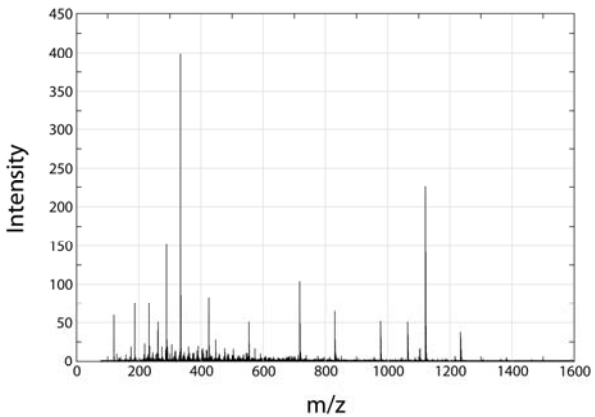
- ❑ Amino acid permutations substituted for weights in candidate sequences from Stage II calculations
- ❑ No current models exist for accurate prediction of ion intensity trends as a function of peptide composition for generalized mass analyzers
- ❑ Assume normalized intensity distribution + reward / penalty based on observation/absence of supporting ions
- ❑ Cross-correlate of all theoretical mass spectra of candidate peptide sequences with experimental tandem mass spectrum

De Novo Framework: PILOT



Peptide identification via **I**nteger **L**inear **O**ptimization and **T**andem mass spectrometry

Illustrative Example for PILOT: DAFLGSFLYEYSR



Raw MS/MS spectrum

Preprocessing Algorithm

C-terminal amino acid	R → peak at 175 Da
N-terminus boundary conditions, BC^{tail}	DA, AD, SV, VS, EG, or GE no supporting y_{n-1} ion

Adjust BC^{tail} → m/z(y_{n-2} ion) = 1381.69 Da

Filtered spectrum

Identified peaks

Stage I Sequences

Candidate Sequence	Objf
F(L/I)GSF(L/I)YHGANR	2.9850
F(L/I)GSF(L/I)YHAGNR	2.9674
F(L/I)GSF(L/I)YQHNR	2.9499
F(L/I)GSF(L/I)YHQNR	2.9374
F(L/I)GSF(L/I)YEYSR	2.8547
F(L/I)GSF(L/I)YEH(L/I)R	2.7544
F(L/I)GSF(L/I)YE(L/I)HR	2.7444
F(L/I)GSF(L/I)YYESR	2.6968
F(L/I)GSF(L/I)YYTDR	2.6391

Integer cuts

Stage I sequences

Stage II Sequences

Candidate Sequence	Objf
F(L/I)GSF(L/I)Y[194.06]ANR	2.8323
F(L/I)GSF(L/I)Y[208.10]GNR	2.8148
F(L/I)GSF(L/I)Y[265.12]NR	2.7847
F(L/I)GSF(L/I)[172.04]QGANR	2.6501
F(L/I)GSF(L/I)[171.04]EGANR	2.6501
F(L/I)GSF(L/I)[171.04]SVANR	2.6351
F(L/I)GSF(L/I)[171.04]GEANR	2.6351
F(L/I)GSF(L/I)[171.04]DAANR	2.6351
F(L/I)GSF(L/I)[172.04]NAANR	2.6351

All sequences

X = high confidence residue

X = low confidence residue

PostProcessing: *DAFLGSFLYEYSR*

P.A. DiMaggio and C.A. Floudas, AIChE Journal, 53(1), 160-173 (2007).

Comparative Study

To benchmark the performance of **PILOT**, we tested it on several tandem mass spectra from

- Quadrupole time-of-flight spectra, QTOF (higher resolution)
- Ion trap spectra (lower resolution, low m/z cutoff)

and compared the predictions to other state-of-the-art *de novo* methods, namely:

- **Lutefisk, LutefiskXP** – J.A. Taylor and R.S. Johnson, Anal. Chem., 73, 2594-2604 (2001).
- **PEAKS** – B. Ma et al., Rapid Commun. Mass Spec., 17, 2337-2342 (2003).
- **NovoHMM** – B. Fischer et al., Anal. Chem., 77, 7265-7273 (2005).
- **PepNovo** – A. Frank and P. Pevzner, Anal. Chem., 77, 964-973 (2005).
- **EigenMS** – M. Bern and D. Goldberg, J. Comp. Biol., 13(2), 364-378 (2006).

Comparative Study: Ion Trap MS/MS

- **Open Proteomics Database***: contains MS/MS spectra for 5 different organisms recorded with ESI-Ion Trap mass spectrometers

- Mass spectra accompanied with predictions from SEQUEST

Which identifications are correct?

- Assignments examined on individual basis for quality

1. Xcorr > 2.2 and ΔCn > 0.1 for +2 charge state
2. Consistent identification with Mascot
3. $\frac{\text{Number of observed b and y ions}}{\text{Number of predicted b and y ions}}$

- Organism studied: Mycobacterium smegmatis

Xcorr = cross correlation score computed by SEQUEST

ΔCn = normalized difference in cross-correlation value between #1 and #2 hit in the search

*http://bioinformatics.icmb.utexas.edu/OPD/

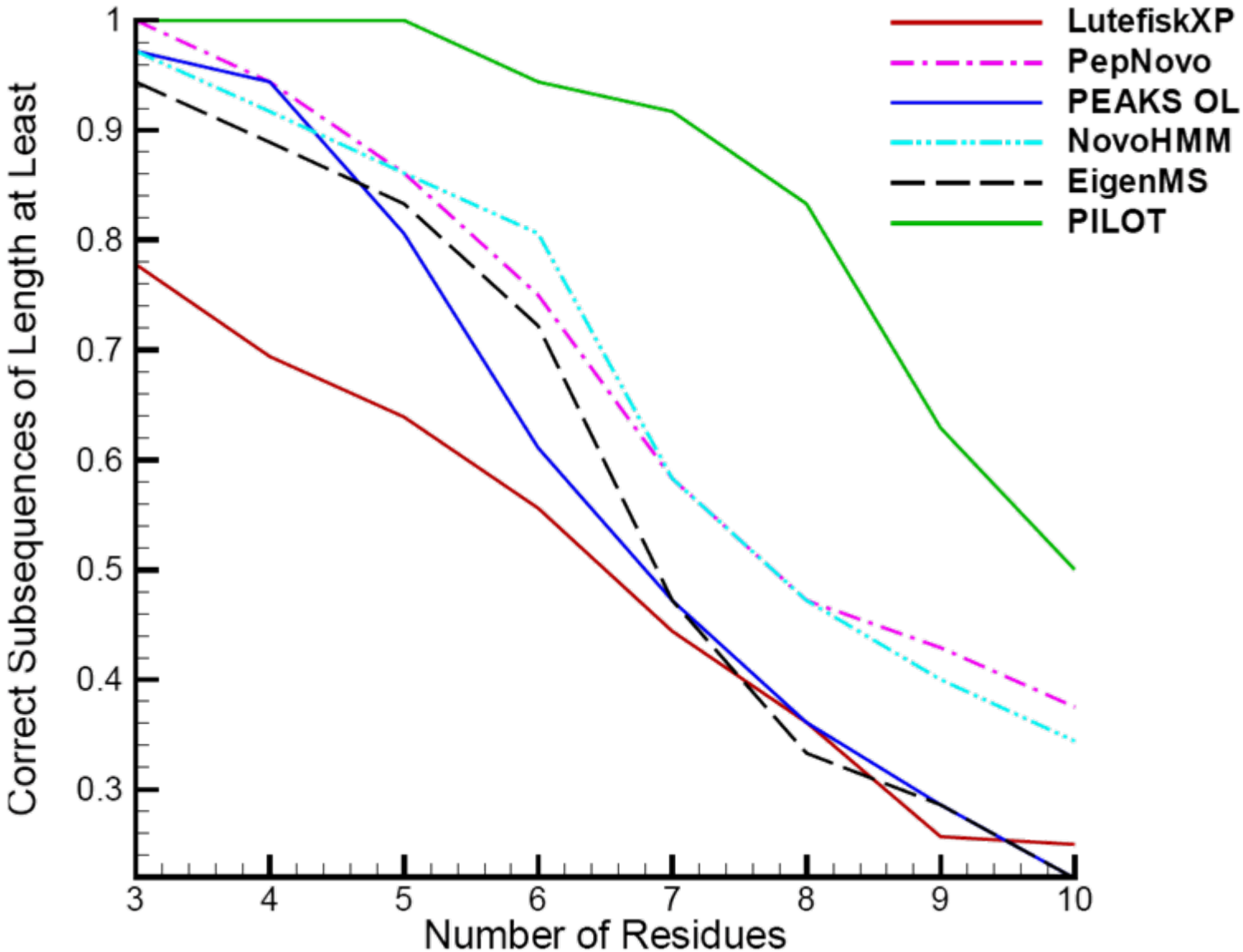
Comparative Study: Ion Trap MS/MS

	LutefiskXP	PepNovo	PEAKS Online	NovoHMM	EigenMS	PILOT
Correct Identifications	2 (0.056)	8 (0.222)	6 (0.167)	9 (0.250)	6 (0.167)	17 (0.472)
with in 1 Residue	3 (0.083)	9 (0.250)	7 (0.194)	10 (0.278)	8 (0.222)	17 (0.472)
with in 2 Residue	11 (0.306)	20 (0.556)	12 (0.333)	18 (0.500)	18 (0.500)	29 (0.806)
with in 3 Residue	17 (0.472)	23 (0.639)	17 (0.472)	25 (0.694)	19 (0.528)	32 (0.889)
Total Correct Residues	222 (0.544)	310 (0.760)	281 (0.689)	309 (0.757)	289 (0.708)	359 (0.880)

Subsequence Length	**x = 3**	**x = 4**	**x = 5**	**x = 6**	**x = 7**	**x = 8**	**x = 9**	**x = 10**
Number of Peptides of Length ≥ x	**36**	**36**	**36**	**36**	**36**	**36**	**35**	**32**
LutefiskXP	28 (0.778)	25 (0.694)	23 (0.639)	20 (0.556)	16 (0.444)	13 (0.361)	9 (0.257)	8 (0.250)
PepNovo	36 (1.000)	34 (0.944)	31 (0.861)	27 (0.750)	21 (0.583)	17 (0.472)	15 (0.429)	12 (0.375)
PEAKS Online	35 (0.972)	34 (0.944)	29 (0.806)	22 (0.611)	17 (0.472)	13 (0.361)	10 (0.286)	7 (0.219)
NovoHMM	35 (0.972)	33 (0.917)	31 (0.861)	29 (0.806)	21 (0.583)	17 (0.472)	14 (0.400)	11 (0.344)
EigenMS	34 (0.944)	32 (0.889)	30 (0.833)	26 (0.722)	17 (0.472)	12 (0.333)	10 (0.286)	7 (0.219)
PILOT	36 (1.000)	36 (1.000)	36 (1.000)	34 (0.944)	33 (0.917)	30 (0.833)	22 (0.629)	16 (0.500)

P.A. DiMaggio and C.A. Floudas, Anal. Chem., 79, 1433-1446 (2007).

Comparative Study: Ion Trap MS/MS



P.A. DiMaggio and C.A. Floudas, Anal. Chem., 79, 1433-1446 (2007).

Comparative Study: QTOF MS/MS

- *Quadrupole time-of-flight* (QTOF) spectra have better resolution that ion trap spectra
- Examined QTOF data for a mixture of 4 known proteins*:
  - ✓ Alcohol dehydrogenase (yeast)
  - ✓ Myoglobin (horse)
  - ✓ Albumin (horse, BSA)
  - ✓ Cytochrome C (horse)
- Spectra were assessed for quality based on the metric:

$$\frac{s}{m} = \frac{\sum_{\{i\,:\lambda_i > 2\}} \lambda_i}{\text{Peptide Mass}}$$

(λ_i = intensity of ion peak i)

*http://www.csd.uwo.ca/~bma/peaks/

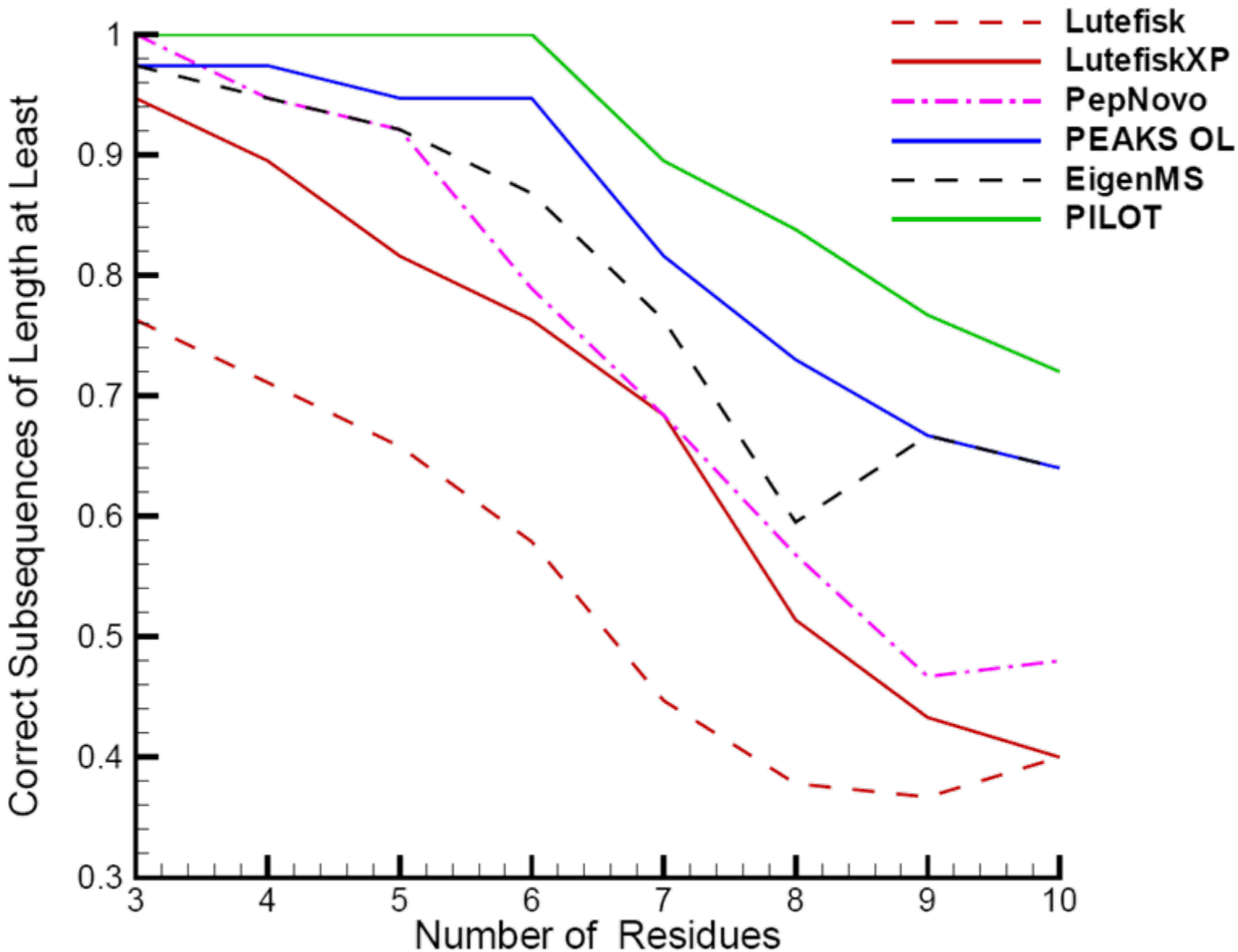
Comparative Study: QTOF MS/MS

	Lutefisk	LutefiskXP	PepNovo	PEAKS Online	EigenMS	PILOT
Correct Identifications	10 (0.263)	9 (0.237)	16 (0.421)	21 (0.553)	20 (0.526)	25 (0.658)
with in 1 Residue	11 (0.290)	10 (0.263)	17 (0.447)	22 (0.579)	21 (0.553)	25 (0.658)
with in 2 Residue	23 (0.605)	22 (0.579)	25 (0.658)	29 (0.763)	29 (0.763)	33 (0.868)
with in 3 Residue	23 (0.605)	25 (0.658)	27 (0.711)	32 (0.842)	30 (0.790)	35 (0.921)
Total Correct Residues	245 (0.586)	294 (0.703)	337 (0.806)	366 (0.876)	353 (0.845)	381 (0.912)

Subsequence Length	**x = 3**	**x = 4**	**x = 5**	**x = 6**	**x = 7**	**x = 8**	**x = 9**	**x = 10**
Number of Peptides of Length ≥ x	**38**	**38**	**38**	**38**	**38**	**37**	**30**	**25**
Lutefisk	29 (0.763)	27 (0.711)	25 (0.658)	22 (0.579)	17 (0.447)	14 (0.378)	11 (0.367)	10 (0.400)
LutefiskXP	36 (0.947)	34 (0.895)	31 (0.816)	29 (0.763)	26 (0.684)	19 (0.514)	13 (0.433)	10 (0.400)
PepNovo	38 (1.000)	36 (0.947)	35 (0.921)	30 (0.789)	26 (0.684)	21 (0.568)	14 (0.467)	12 (0.480)
PEAKS Online	37 (0.974)	37 (0.974)	36 (0.947)	36 (0.947)	31 (0.816)	27 (0.730)	20 (0.667)	16 (0.640)
EigenMS	37 (0.974)	36 (0.947)	35 (0.921)	33 (0.868)	29 (0.763)	22 (0.595)	20 (0.667)	16 (0.640)
PILOT	38 (1.000)	38 (1.000)	38 (1.000)	38 (1.000)	34 (0.895)	31 (0.838)	23 (0.767)	18 (0.720)

P.A. DiMaggio and C.A. Floudas, Anal. Chem., 79, 1433-1446 (2007).

Comparative Study: QTOF MS/MS



P.A. DiMaggio and C.A. Floudas, Anal. Chem., 79, 1433-1446 (2007).

Hybrid Approach for Peptide Identification

- In de novo predictions – incomplete fragmentation can yield regions of ambiguity in peptide sequence

??
HPEYAVEGLLR mass(EG) = mass(SV) = mass(DA) = mass(W) = 186 Da

- Utilize protein database to validate amino acid assignments to subsequences of low confidence

- **FASTA*** – tool for locally aligning a peptide query with the protein sequences in a database

301 CCDKPVLEKS HCIAEVDKDA VPENLPPLTA DFAEDKEVCK NYQEAKDVFL GSFLYEYSRR

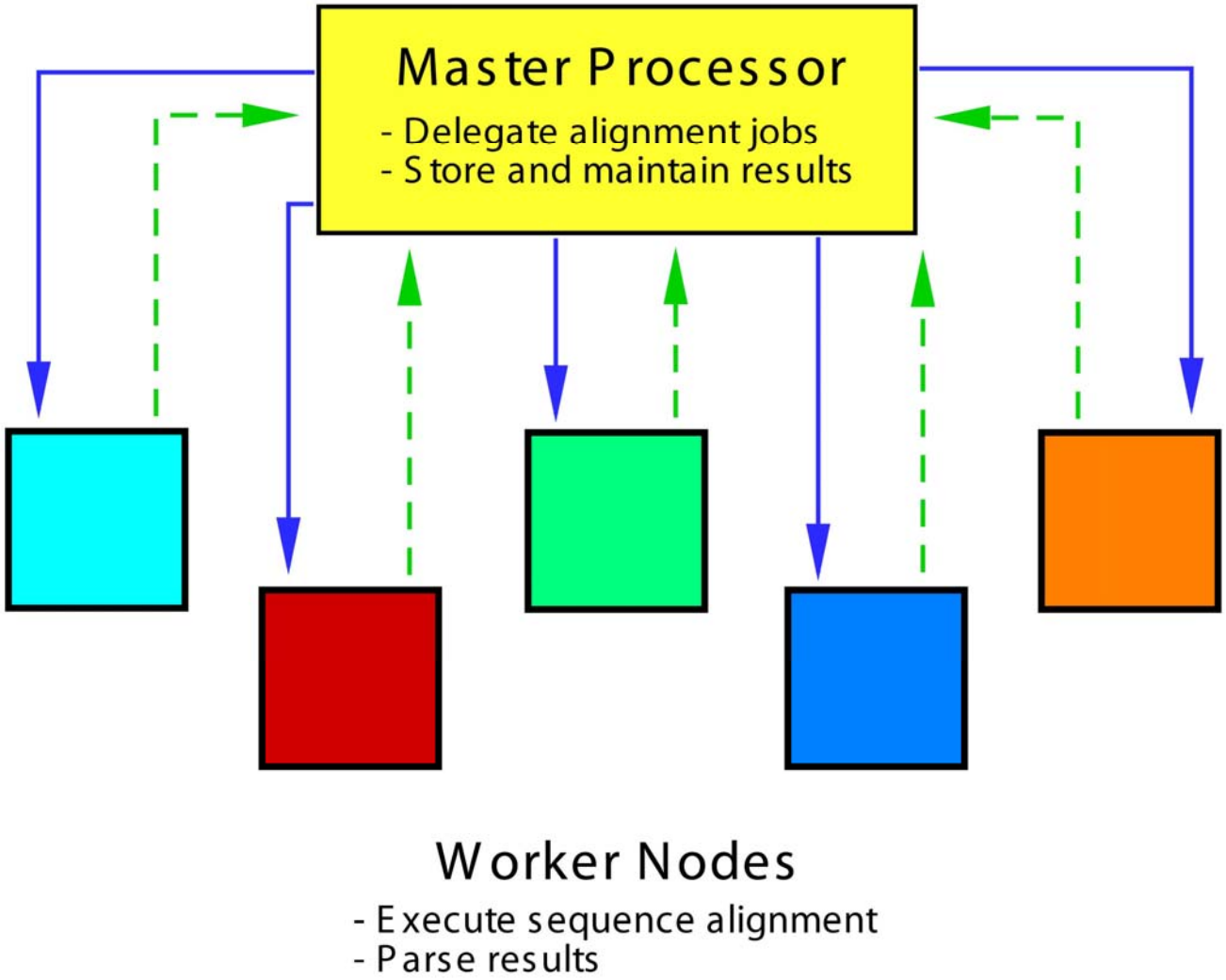
DAFL GSFLYEYSR

- Modify traditional scoring matrices (i.e., BLOSUM or PAM) to emphasize mass conservation instead of evolutionary distance

*W.R. Pearson and D.J. Lipman, PNAS, 85, 2444-2448 (1988).

Hybrid Approach for Peptide Identification

Parallel Implementation



Beowulf Cluster : 80 nodes with dual Intel Xeon 3.0 GHz processors

Hybrid Results for QTOF Spectra

Peptide	De Novo Predictions	Hybrid Predictions
AEFVEVTK	AEFEVTK	AEFEVTK
YLYELAR	YLYELAR	YLYELAR
LKAWSVAR	LKAWSVAR	LKAWSVAR
ALKAWSVAR	ALKAWSVAR	ALKAWSVAR
QTALVELLK	QTALVELLK	QTALVELLK
KQTALVELLK	KQTALVELLK	KQTALVELLK
LVNELTEFAK	LVNELTEFAK	LVNELTEFAK
HPEYAVSVLLR	HPEYAVEGLLR	HPEYAVSVLLR
HLVDEPQNLIK	HLVDEPQNLIK	HLVDEPQNLIK
SLHTLFGDELCK	AEHTLFGDELCK	SLHTLFGDELCK
YICDNQDTISSK	YICDNQDTISSK	YICDNQDTISSK
LGEYGFQNALIVR	LGEYGFQNALIVR	LGEYGFQNALIVR
VPQVSTPTLVEVSR	VPQVSTDPVVEVSR	VPQVSTPTVEVSR
DAFLGSFLYEYSR	DAFLGSFLYEYSR	DAFLGSFLYEYSR
KVPQVSTPTLVEVSR	KVPQVSTPTLVEVSR	KVPQVSTPTLVEVSR
IGDYAGIK	IGDYAGIK	IGDYAGIK
DIPVPKPK	EVPVPKPK	DIPVPKPK / PMPVPKPK
EALDFFAR	EALDFFAR	EALDFFAR
TLPEIYEK	LTPELYEK	TLPELYEK / LTPELYEK
ANELLINVK	ANELLLNVK	ANELLLNVK
SIVGSYVGNR	EAVGSYVGNR	SIVGSYVGNR
EKDIVGAVLK	KEDIVGAVLK	EKDIVGAVLK
STLPEIYEK	STLPEIYEK	STLPEIYEK
VSEAAIEASTR	VSEAAIEASTR	VSEAAIEASTR
DGGEGKEELFR	DGGEGKEELFR	DGGEGKEELFR
SISIVGSYVGNR	AESIVGSYVGNR	SISIVGSYVGNR

Table continued

SISIVGSYVGNR	AESIVGSYVGNR	SISIVGSYVGNR
GAAGGLGSLAVQYAK	QAGGLGSLAVQYAK	GAAGGLGSLAVQYAK
ANGTTVLVGMPAGAK	[444.18]TVLVGMPAGAK	ANGTTVLVGMPAGAK
GIDGGEGKEELFR	GIDGGEGKEELFR	GIDGGEGKEELFR
ADTREALDFFAR	[443.20]EALDFFAR	ADTREALDFFAR
VLGIDGGEGKEELFR	VIGIDGGKSVEELFR	VLGIDGGEGKEELFR
EDLIAYLK	EDLLAYLK	EDLLAYLK
PNLHGLFGR	PNLHGLFGR	PNLHGLFGR
GLSDGEWQQVLNVWGK	[558.55]WEQVLNVWGK	GLSDGEWQQVLNVWGK
EETLMEYLENPK	EETLMEYLENPK	EETLMEYLENPK
TGPNLHGLFGR	TGPNLHGLFGR	TGPNLHGLFGR
TGQAPGFTYTDANK	TGQAPGFTYTDANK	TGQAPGFTYTDANK
EETLMoEYLNPK	EETLMoEYLENPK	EETLMoEYLENPK

- ✓ Presented are the 38 QTOF spectra accompanied by the best de novo and hybrid predictions
- ✓ Recall the de novo method correctly identified 25 peptides
- ✓ The hybrid method correctly identifies 36 peptides

Conclusions

- Developed accurate de novo and hybrid framework, **PILOT**, for the identification of peptides via tandem mass spectrometry (MS/MS)
- **PILOT** outperformed several state-of-the-art de novo methods in a comparative study for ion trap and QTOF tandem mass spectra.
- Key elements of proposed method:
  - Novel mixed-integer linear optimization (MILP) formulation for peptide identification
  - Preprocessing algorithm for filtering spectra and identifying important ion peaks
  - Post-processing algorithm for cross-correlating theoretical tandem mass spectra with experimental tandem mass spectrum

Future Directions…

- Preliminary studies validate prediction enhancements by integrating parametric uncertainty into de novo framework

- Enhance performance of hybrid peptide identification method which combines the strengths of the proposed de novo method and protein database search algorithms

- Incorporation of post-translational modifications into current framework

- Create workbench to make PILOT available to the scientific community

Acknowledgements

Financial Support

- US Environmental Protection Agency, EPA (R 832721-010)*
- National Institutes of Health

Relevant Publications

- P.A. DiMaggio and C.A. Floudas, A mixed-integer optimization framework for de novo peptide identification, *AIChE Journal*, 53(1), 160-173 (2007).
- P.A. DiMaggio and C.A. Floudas, De novo peptide identification via tandem mass spectrometry and integer linear optimization, *Anal. Chem.*, 79, 1433-1446 (2007).

*This work has not been reviewed by and does not represent the opinions of this funding agency.