



AB Applied Biosystems | **MDS SCIEX**

Perspective and Progress in Protein and Peptide Identification

Sean L. Seymour, Ph.D.
Senior Staff Scientist

NCI Strategies for Improving Reliability in Protein and Peptide Identification Workshop

Nov. 15, 2007

Outline

- **Our perspective and approach**
 - Continuous measures, preservation of information, and handling ambiguity
 - Ease of use
- **Current proteomics community issues**
 - Reproducibility and the comparison problem
 - Error rate assessment and reporting
- **Future issues**
 - Data standards
 - Proteomics community cooperation and collective efforts

The Paragon™ Algorithm Key Idea:

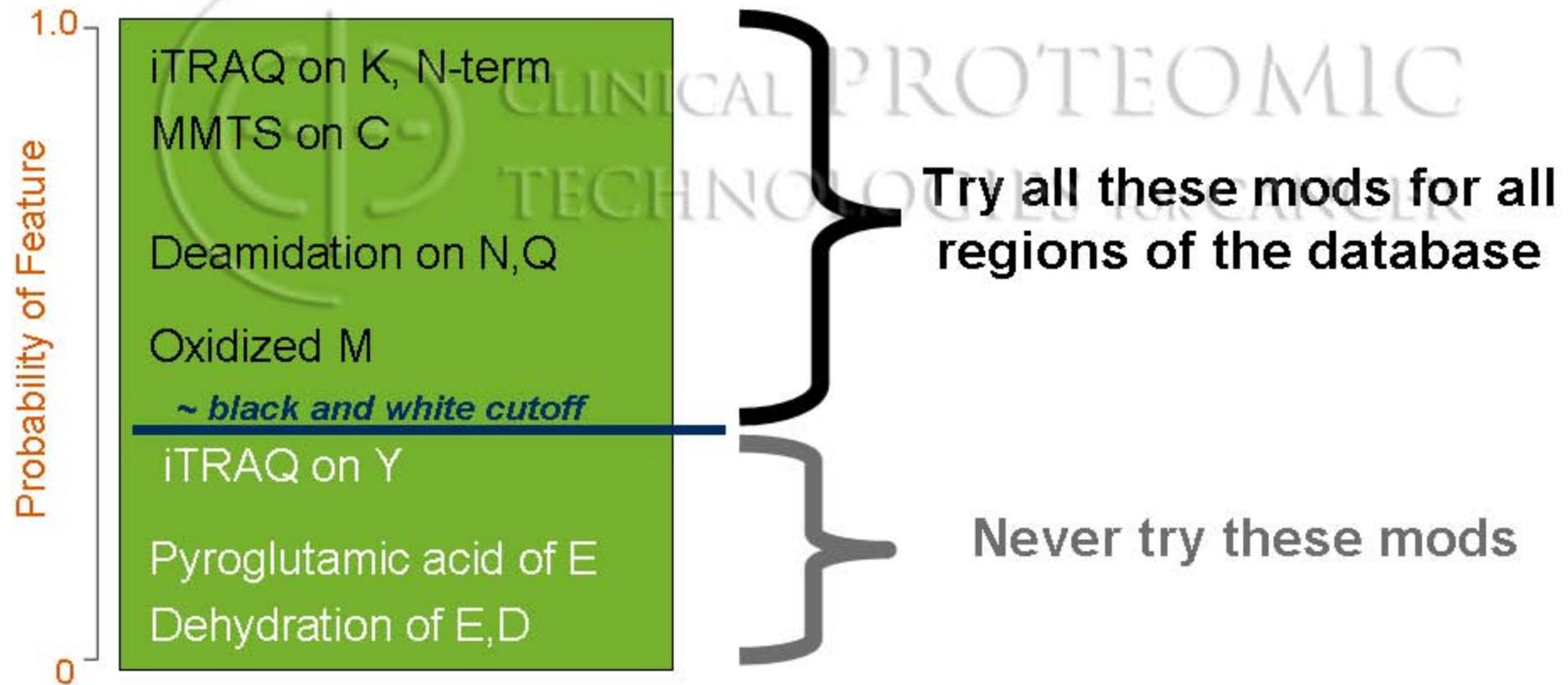
- For the search of a single spectrum, the allowed search space applied to each sequence region should be proportional to the degree of tag evidence implicating that region.



CLINICAL PROTEOMIC
TECHNOLOGIES FOR CANCER

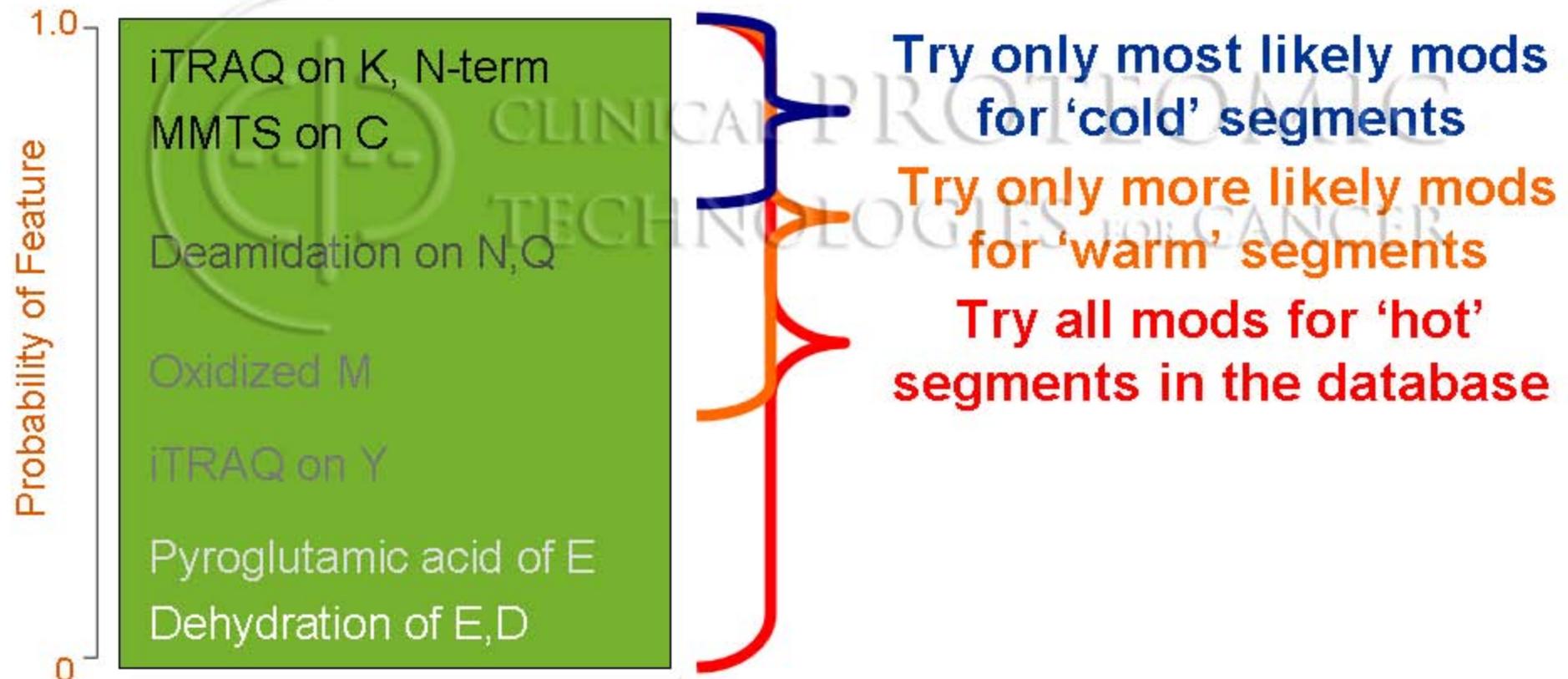
How to Model Search Space?

Conventional search engine: discrete boundaries.



How to Model Search Space?

The Paragon Algorithm: 'Feature probabilities' give a continuous description of search space elements.



Same idea with digestion features, tolerances, etc.

Simplified Informatics

The algorithm enables a new kind of user interface

- Informatics expertise is not necessary for success
 - All 'algorithmic' knobs are gone - mass tolerances, number of missed cleavages, digest settings, which modifications to choose, or multi-pass search strategies
- Describe the sample, the ID focus, and search effort

Paragon Method

Paragon Method: Standard method - EColi Project Delete

Describe Sample

Sample Type: Identification

Cys Alkylation: Iodoacetamide

Digestion: Trypsin

Instrument: QSTAR ESI

Special Factors:
 Phosphorylation emphasis
 Gel-based ID
 Urea denaturation

Species: Escherichia coli

Specify Processing

Quantitate

ID Focus:
 Biological modifications
 Amino acid substitutions
 User-defined modifications

Database: combined_KBMS5.0.20050302

Search Effort:
 Rapid ID Thorough ID

Detected Protein Threshold [Unused ProtScore (Conf)] >: 2.0 (99.0%)

Save Save As... Cancel

The Importance of Preserving Peptide ID Ambiguity

The risk of thresholding and simplification of intermediate results



- Can result in an unnecessary protein
 - There are many sources of this risk: deamidation, switched residues, mutations, tryptic vs. semitryptic context, other modification variants
 - A particular risk for multi-pass approaches that lock in proteins or spectra
- It is not enough just to track all the proteins that share a single peptide hypothesis.
- This is effectively propagation of uncertainty

Reproducibility in Protein Identification

Many examples of studies finding poor intersection

- An example from the HUPO Brain proj.
- Sample treatment issues?
- There are also informatics issues here.

Reidegeld et al. Proteomics 2006, 6, 4997-5014.

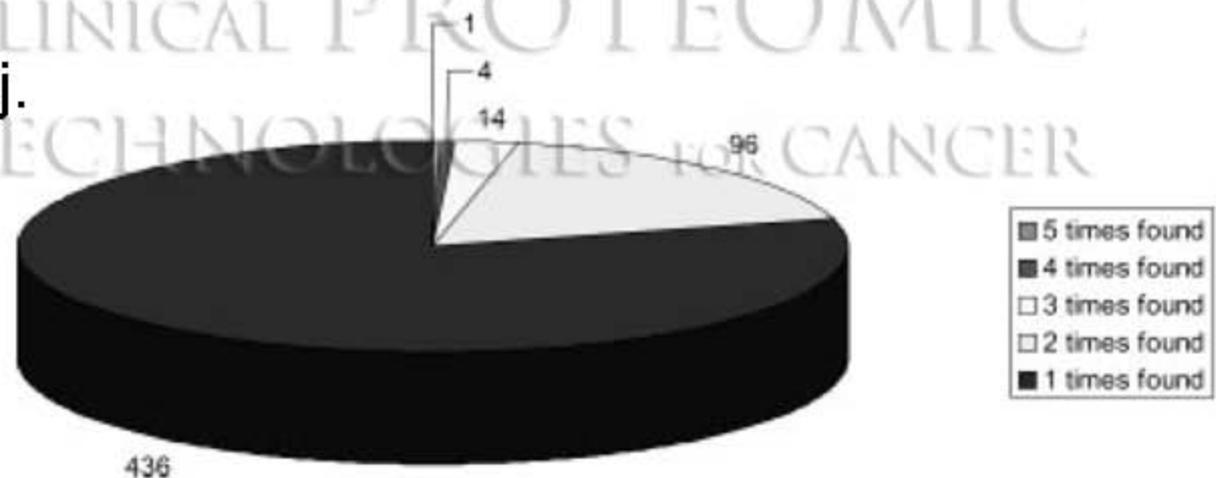


Figure 2. In total 551 non-redundant proteins from the mouse brains were identified after reprocessing and could be mapped to differential expression ratios. One protein was found in five independent analyses, four proteins in four. While 14 proteins were found in three different analyses and 96 proteins in two, 436 proteins were discovered in one single analysis.

How Do You Compare Two Protein Lists?

Part of the apparent reproducibility problem?

Data Set 1 - Protein Group #8: plectin 1

First protein (winner) IPI:00398775.3 (isoform 2)

Is this same protein being detected in both data sets or are these different?

Data Set 2 - Protein Group #3: plectin 1

First protein (winner) IPI:00186711.3 (isoform 6)

Preserving Protein Ambiguity is Key for Comparison

'Competitor' Proteins

ProteinPilot™ Software
File Configure Window Help

Result - C:\Applied Biosystems MDS Sciex\ProteinPilot Data\Results

Protein ID Spectra

Proteins Detected

N	Unused	Total	% Cov	Accession #	Name
6	54.17	54.17	64.9	IP:IP00554648.1	Keratin, type II cytoskeletal
7	53.19	53.19	40.3	IP:IP00019502.2	Myosin-9
8	52.72	52.72	30.3	IP:IP00398775.3	plectin 1 isoform 2
9	52.40	52.40	54.6	IP:IP00003865.1	Splice Isoform 1 of Heat shock cognate 71 kDa protein
10	52.26	52.26	66.4	IP:IP00011654.2	Tubulin beta-2 chain

Protein Group 8

Proteins in Group				
N	Unused	Total	Accession #	Name
8	52.72	52.72	IP:IP00398775.3	plectin 1 isoform 2
	0.00	52.72	IP:IP00398002.4	plectin 1 isoform 1
	0.00	52.72	IP:IP00186711.3	plectin 1 isoform 6
	0.00	52.72	IP:IP00420096.4	plectin 1 isoform 3
	0.00	52.72	IP:IP00398779.3	plectin 1 isoform 11
	0.00	52.72	IP:IP00398778.3	plectin 1 isoform 10
	0.00	52.72	IP:IP00398777.3	plectin 1 isoform 8
	0.00	52.72	IP:IP00398776.3	plectin 1 isoform 7

ProteinPilot™ Software
File Configure Window Help

Result - C:\Applied Biosystems MDS Sciex\ProteinPilot Data\Results

Protein ID Spectra

Proteins Detected

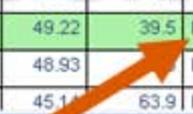
N	Unused	Total	% Cov	Accession #	Name
1	63.34	63.34	64.2	IP:IP00013808.1	Alpha-actinin-4
2	51.09	51.09	70.7	IP:IP00219018.6	Glyceraldehyde-3-phosphate dehydrogenase
3	49.22	49.22	39.5	IP:IP00186711.3	plectin 1 isoform 6
4	48.93	48.93		IP:IP00418411.4	Keratin 8 variant
5	45.14	45.14	63.9	IP:IP00003865.1	Splice Isoform 1 of Heat shock cognate 71 kDa protein

Protein Group 3

Proteins in Group					Contrib	Conf	S
N	Unused	Total	Accession #	Name			
3	49.22	49.22	IP:IP00186711.3	plectin 1 isoform 6	2.00	99	AGTLSITEP
1246	0.12	41.11	IP:IP00215943.1	Splice Isoform 3 of Plectin 1	2.00	99	APVPASELL
	0.00	49.22	IP:IP00398002.4	plectin 1 isoform 1	2.00	99	AQAEAQOPT
	0.00	49.21	IP:IP00420096.4	plectin 1 isoform 3	2.00	99	DALDGPAAE
	0.00	49.21	IP:IP00398779.3	plectin 1 isoform 11	2.00	99	DGHNLSL
	0.00	49.21	IP:IP00398778.3	plectin 1 isoform 10	2.00	99	DLLPDMAV
	0.00	49.21	IP:IP00398777.3	plectin 1 isoform 8	2.00	99	DPYTGQQIS
	0.00	49.21	IP:IP00398776.3	plectin 1 isoform 7	2.00	99	EAKELQQR
	0.00	49.21	IP:IP00398775.3	plectin 1 isoform 2			

Workflow Tasks

Workflow Tasks



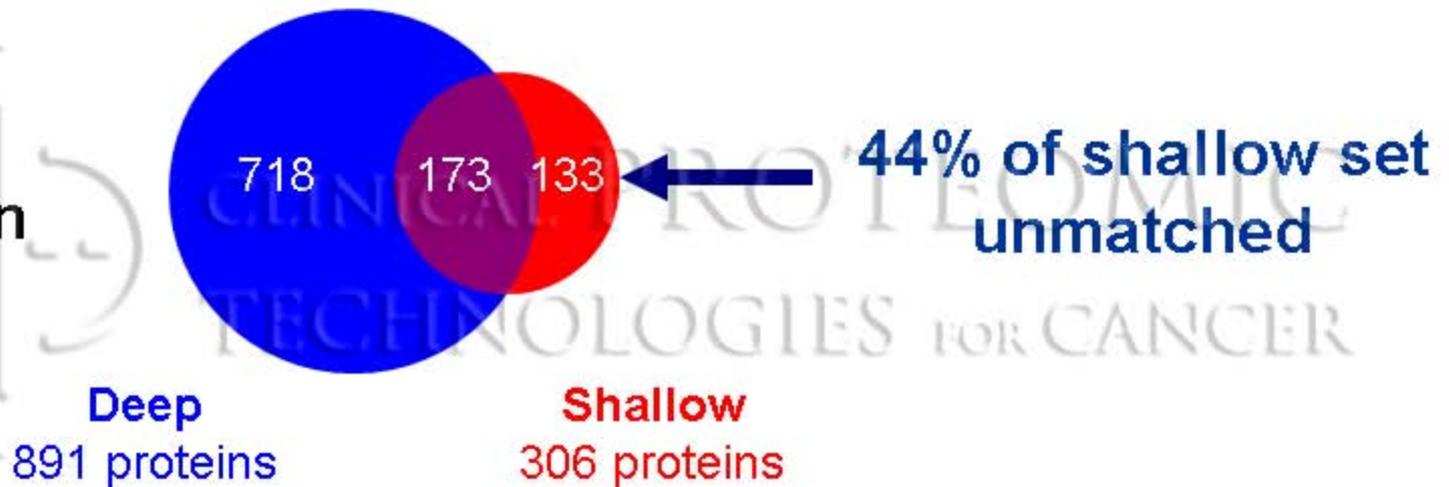
Preserving Protein Ambiguity is Key for Comparison 'Competitor' Proteins

- Conclude for this example:
 - We have detected the same molecular species in both samples
 - We cannot say specifically which of these isoforms it is
- This ambiguity should be preserved and reported, not discarded to 'simplify things'.
- This is propagation of uncertainty in proteins for comparison analysis.
- This is critical for better comparison across data sets.
- This can be especially important for single hit proteins.

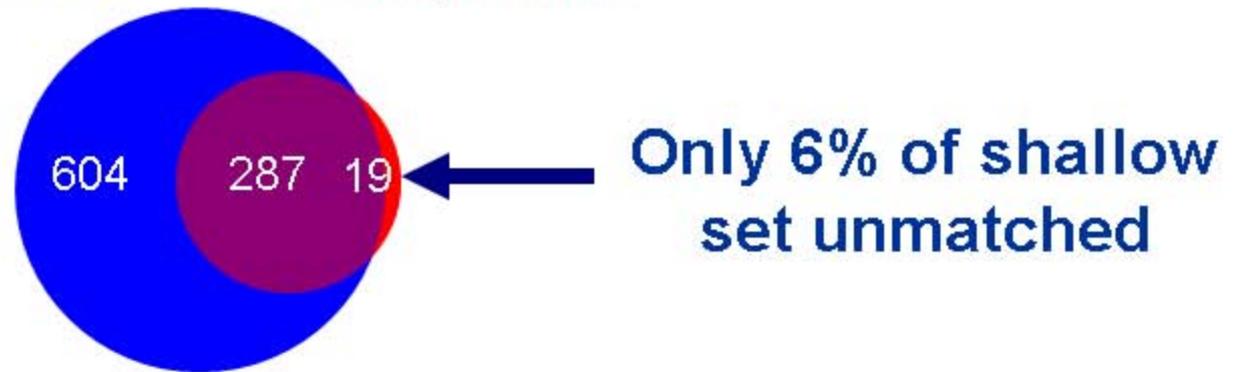
How Much Does This Improve Comparison?

An Assay for Comparison Quality: Deep vs. Shallow Acquisition

Simplistic Comparison



Competitor Comparison

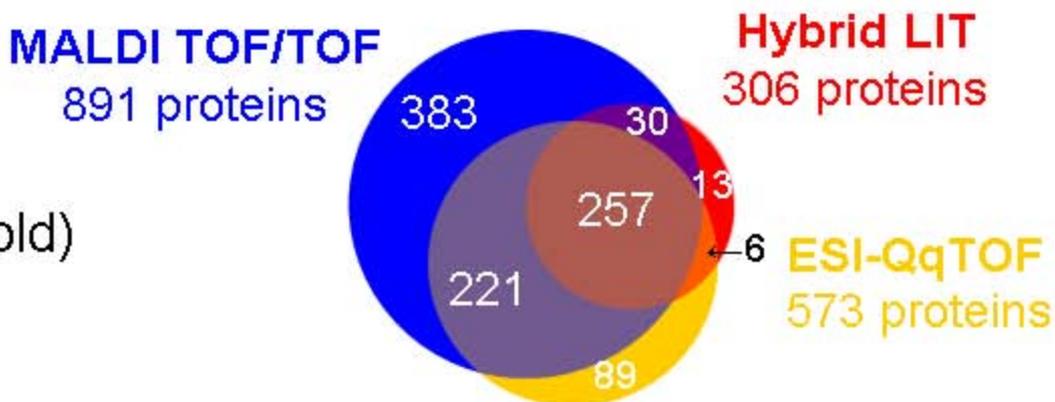
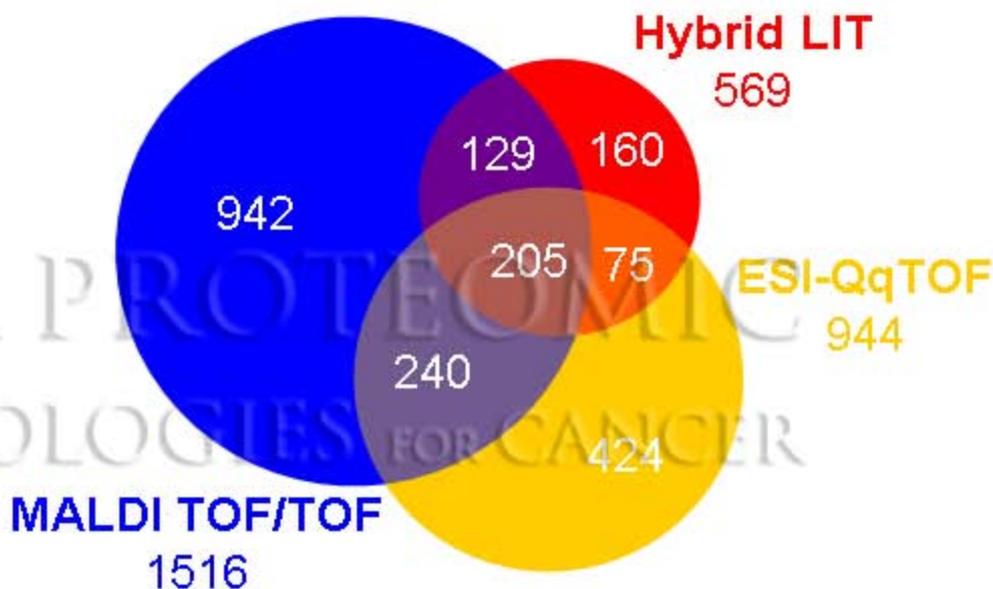


Good Comparison Requires Good Protein Inference

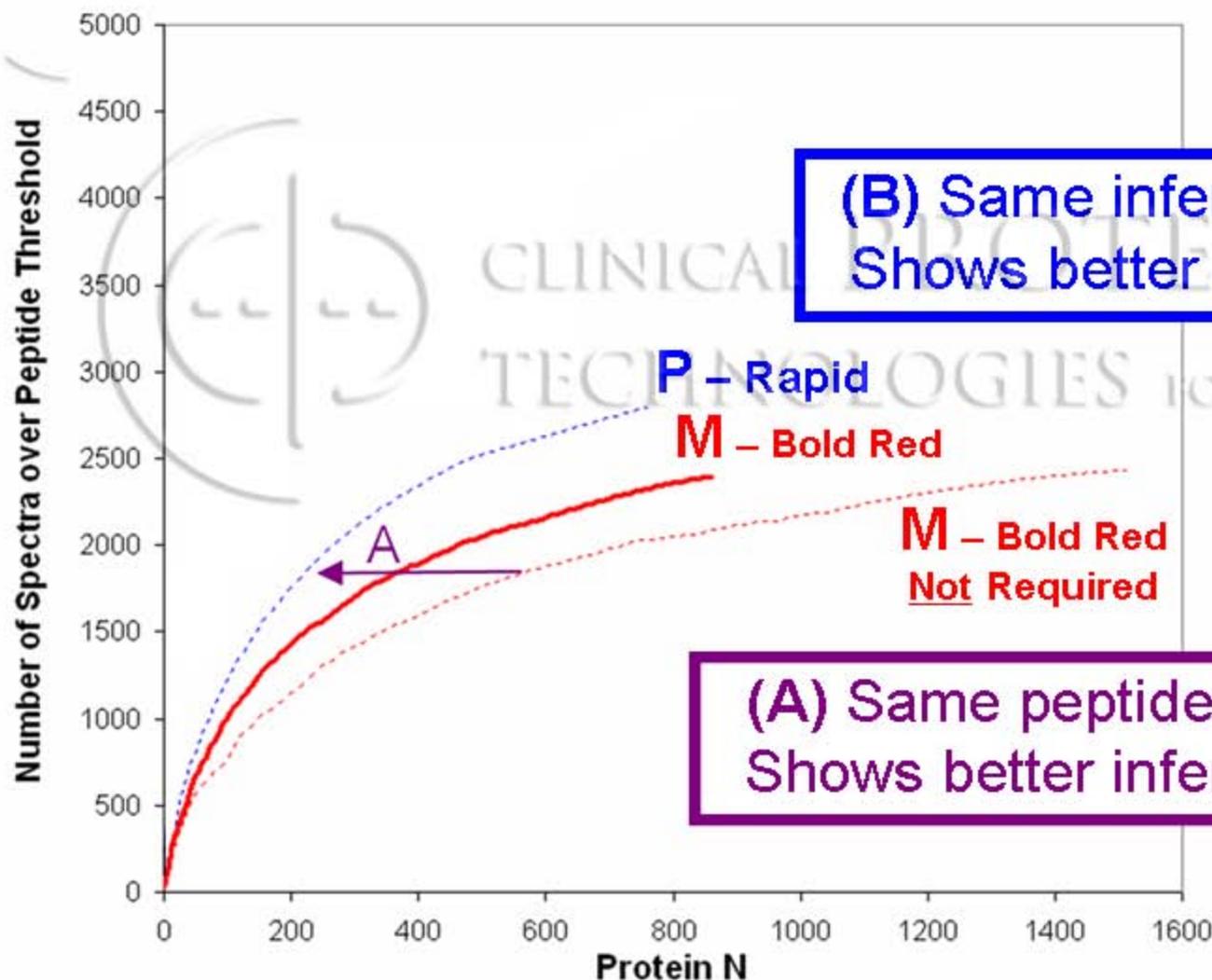
Data from 3 instruments on the same sample

Union protein inference, **2175**
 Vastly not comparable

Paragon™ Algorithm
 Union = **999** (-54%)
 Good protein inference,
 Intersection = **26%** (+2.8 fold)
 competitor comparison



How to Measure the Quality of Protein Inference?



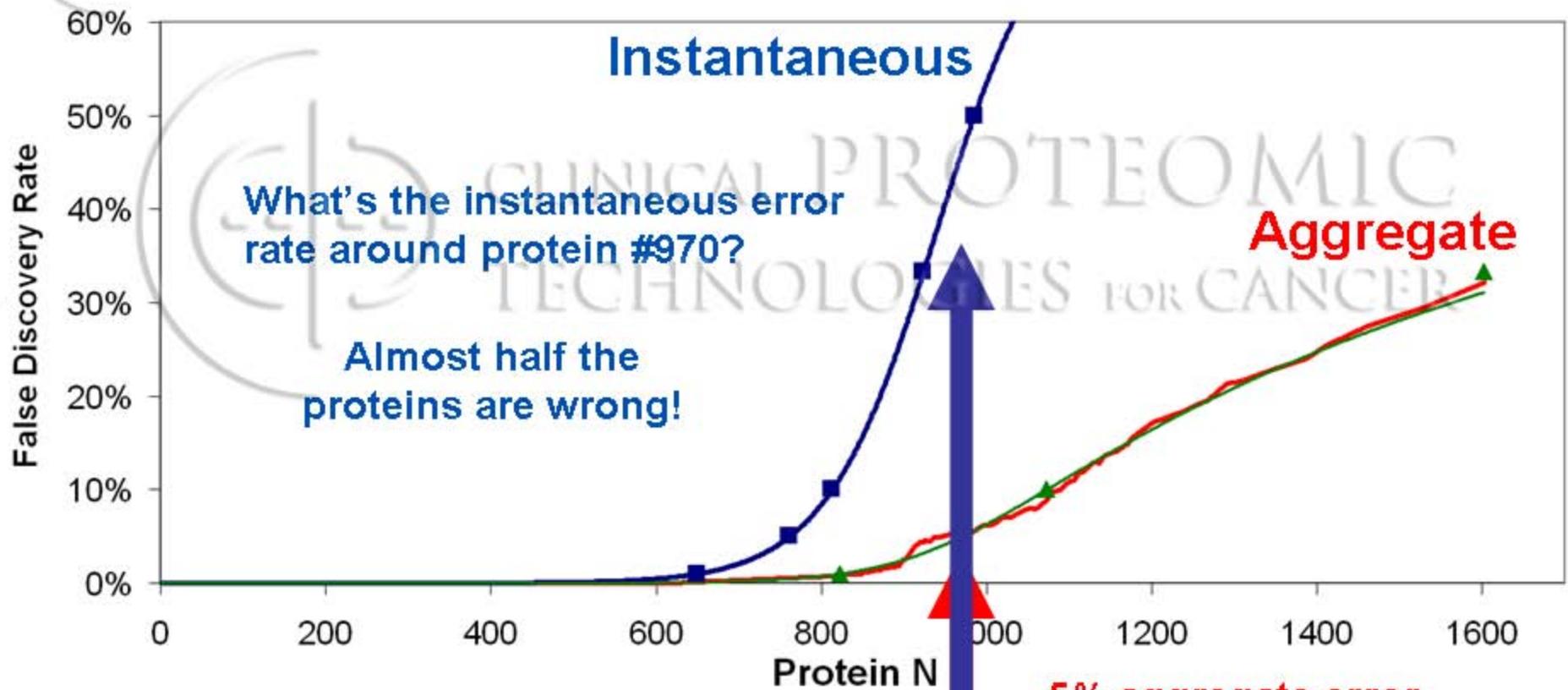
Same data in all searches

(B) Same inference algorithm. Shows better peptide ID.

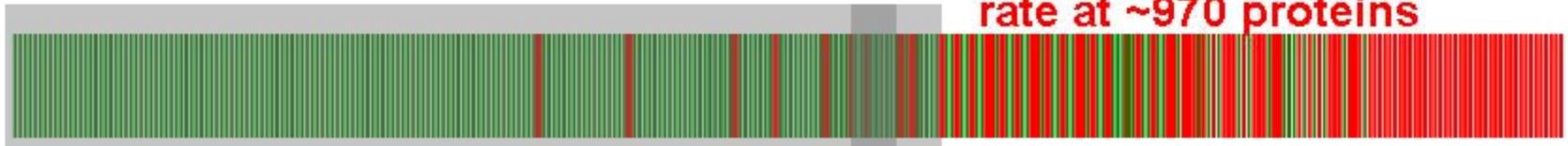
(A) Same peptide search space. Shows better inference algorithm

Measuring Error Rates

Instantaneous vs. Aggregate Error Estimates



(right)
(wrong)

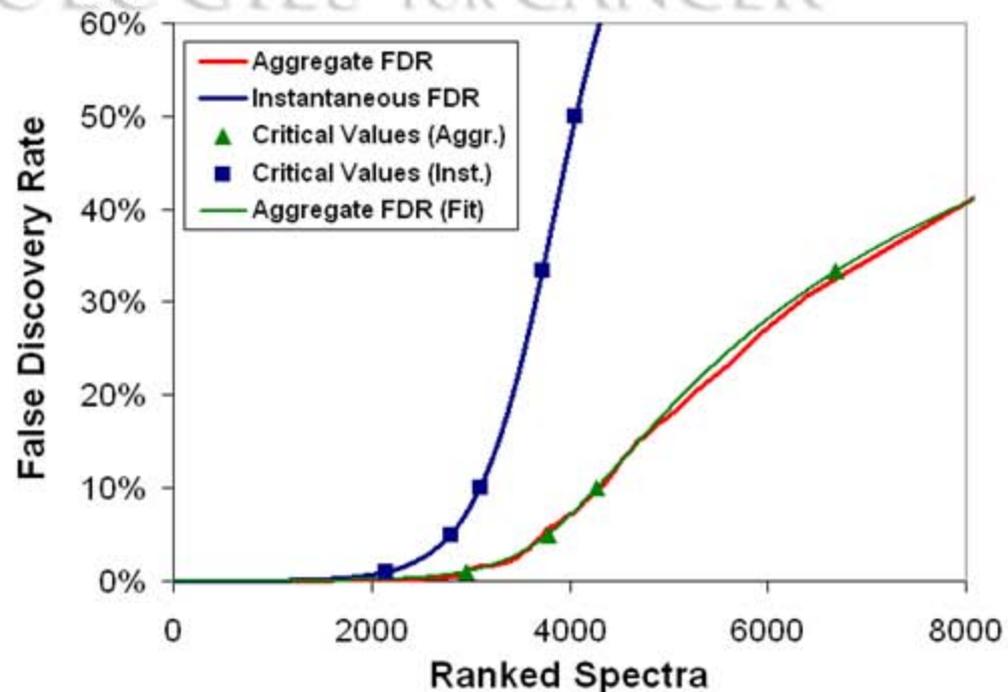
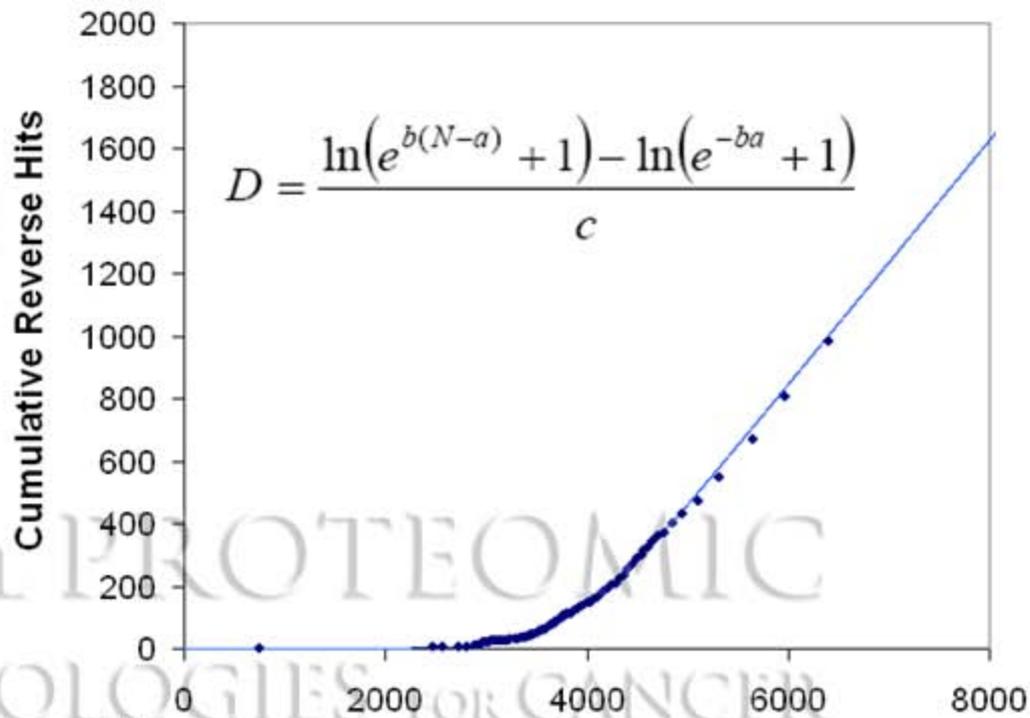


Instantaneous Error Rate by Non-Linear Fitting

Fitting Plot



Instantaneous FDR
at All Thresholds



Reporting at Critical Error Rate Values

840 proteins, **0.6%** FDR

vs.

870 proteins, **1.5%** FDR

How do you compare these?

Number of Proteins Detected at Critical False Discovery Rates

<i>Critical Value</i>	<i>Protein N Cutoff</i>	
Accepted FDR	Instantaneous FDR	Aggregate FDR
1.0%	685	851
5.0%	787	984
10.0%	834	1072
33.3%	929	1534
50.0%	976	2183

Hold something constant

Future Issues

- **When will data format standards happen?**
 - The chicken and egg problem...
 - The dream of the 'living publication'
 - The solution with funding agencies? - Timelines toward requirement of the submission of intermediate data and results using standard data formats, ontologies, and minimal content.
- **Cooperation where common interests**
 - HUPO-PSI, ABRF-PRG/sPRG/iPRG, CPTAC, ProDaC, HUPO-SPI, MCP/Paris Guidelines
 - US, Canadian, European granting agency cooperation?
 - The vendors perspective: a single standard and fewer community efforts will be supported or implemented more quickly and better than a multitude of redundant or nearly redundant efforts.

Acknowledgements

Mass Spec Informatics R&D

Ignat Shilov
Alpesh Patel
Wilfred Tang
Alex Loboda
Sean Keating
Bret Pehrson
Jim Bohannon
Robert Deutschman
Lauren Mansfield
Winnie Leung
Liliya Shilova
Vera Loboda
Lisa Schaechter
Dan Schaeffer

Applications and Informatics Research

Sean Seymour
Matt Willetts
Christie Hunter

Other AB|Sciex

Lydia Nuwaysir
Steve Tate
Christof Lenz
Frank Rooney

CLINICAL PROTEOMIC
TECHNOLOGIES FOR CANCER

Trademarks/Licensing

Applera, Applied Biosystems, and AB (design) are registered trademarks of Applera Corporation or its subsidiaries in the US and/or certain other countries.

MALDI TOF/TOF, ProteinPilot, Paragon, and Pro Group are trademarks of Applied Biosystems/MDS SCIEX, a joint venture between Applera Corporation and MDS Inc.

All other trademarks are the sole property of their respective owners.

© 2007 Applera Corporation and MDS Inc. All Rights Reserved.