Shotgun Proteomics: How Confident are you in that Identification?

or

Statistical Evaluation of Shotgun Proteomic Data

Ron Orlando

Complex Carbohydrate Research Center

University of Georgia

Athens, GA 30602

What is Proteomics?

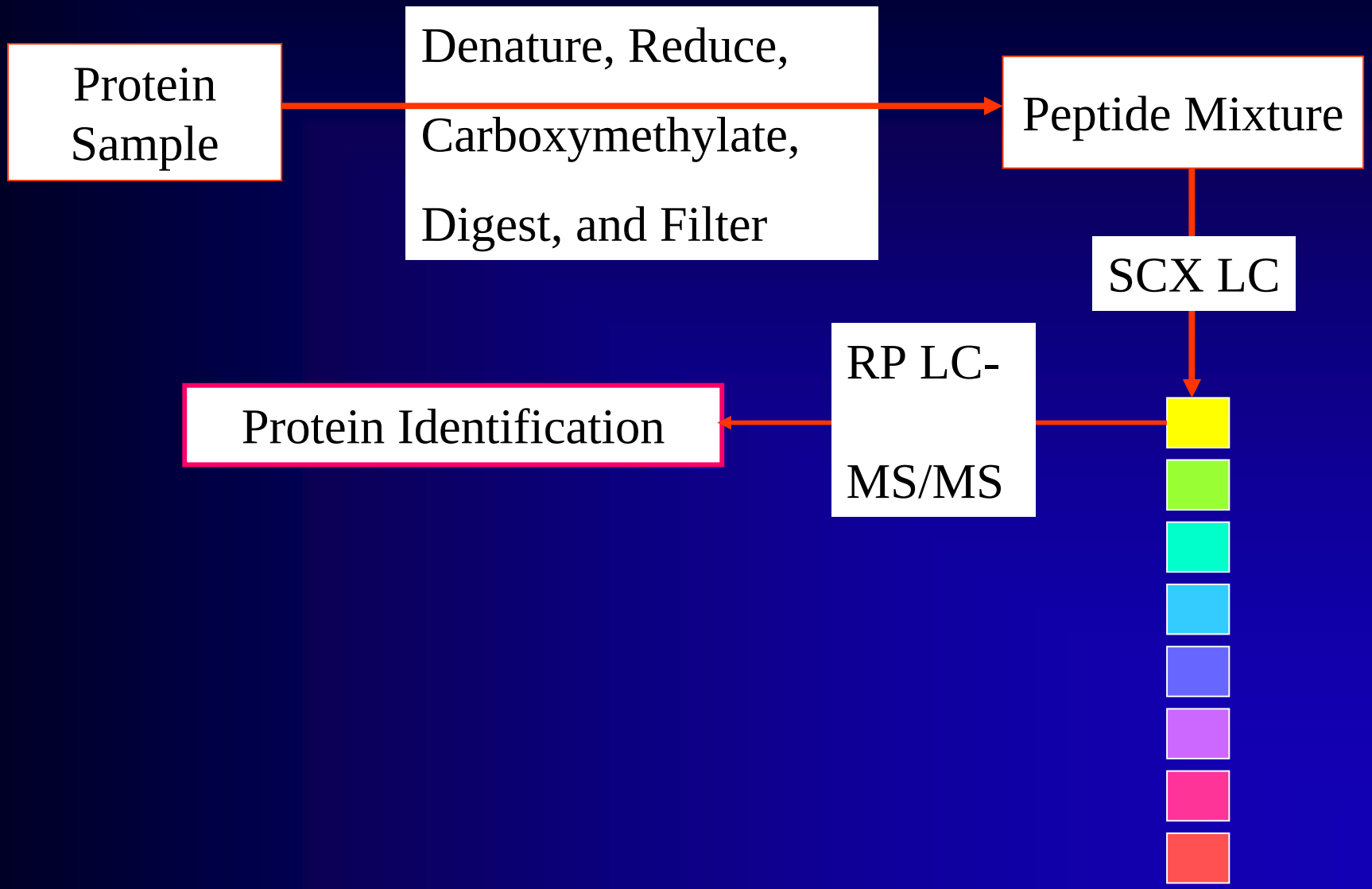
- A **proteome** is the entire protein complement of a given genome.

- **Proteomics** is the study of proteomes from two (or more) differentially treated cell (or tissue) lines.

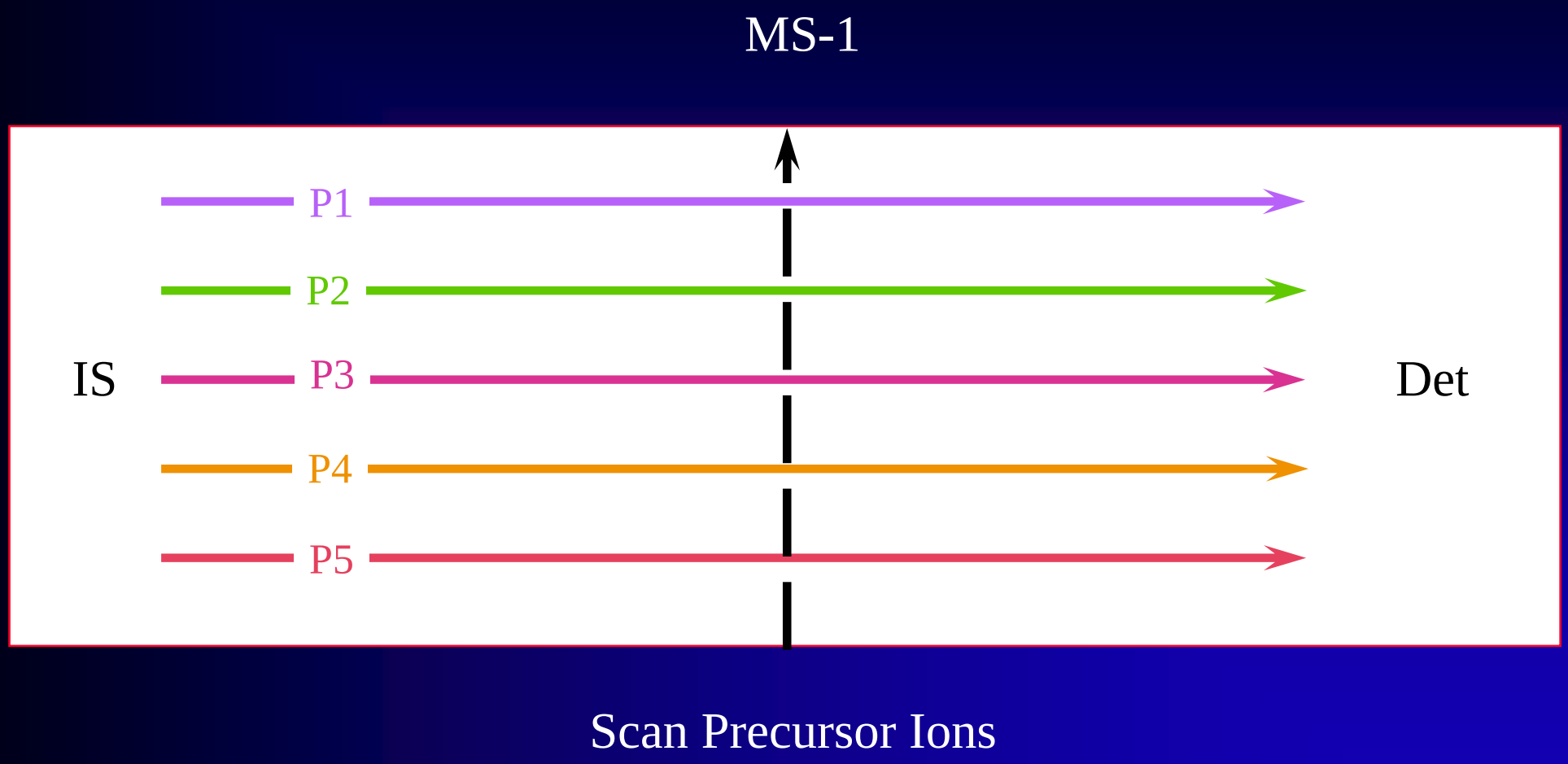
One Genome - Different Proteomes

F. Lottspeich, *Agnew. Chem. Int. Ed.*, **1999**, **38**, 2476-2492

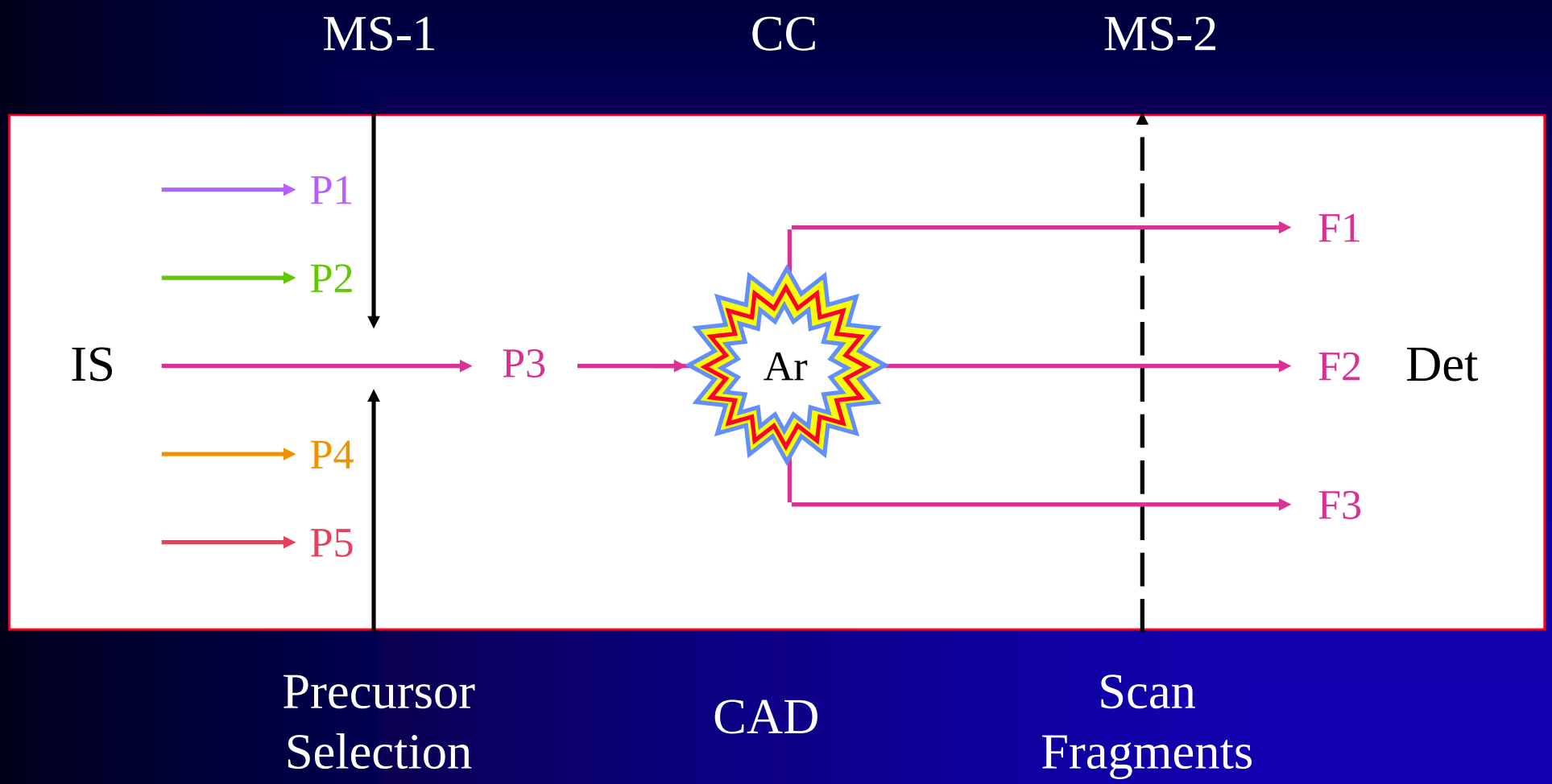
Multidimensional LC-MS/MS Proteomics



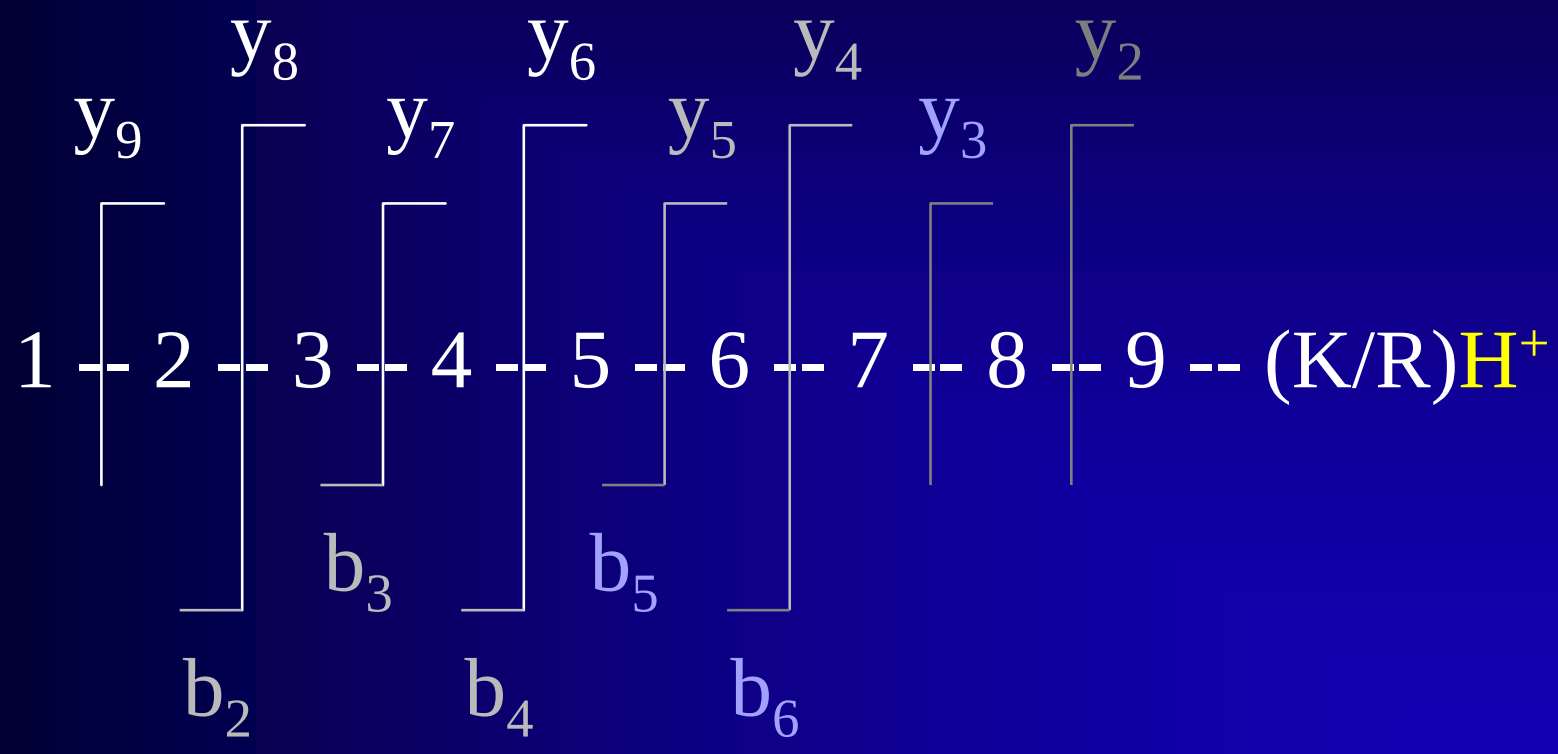
MS Scan



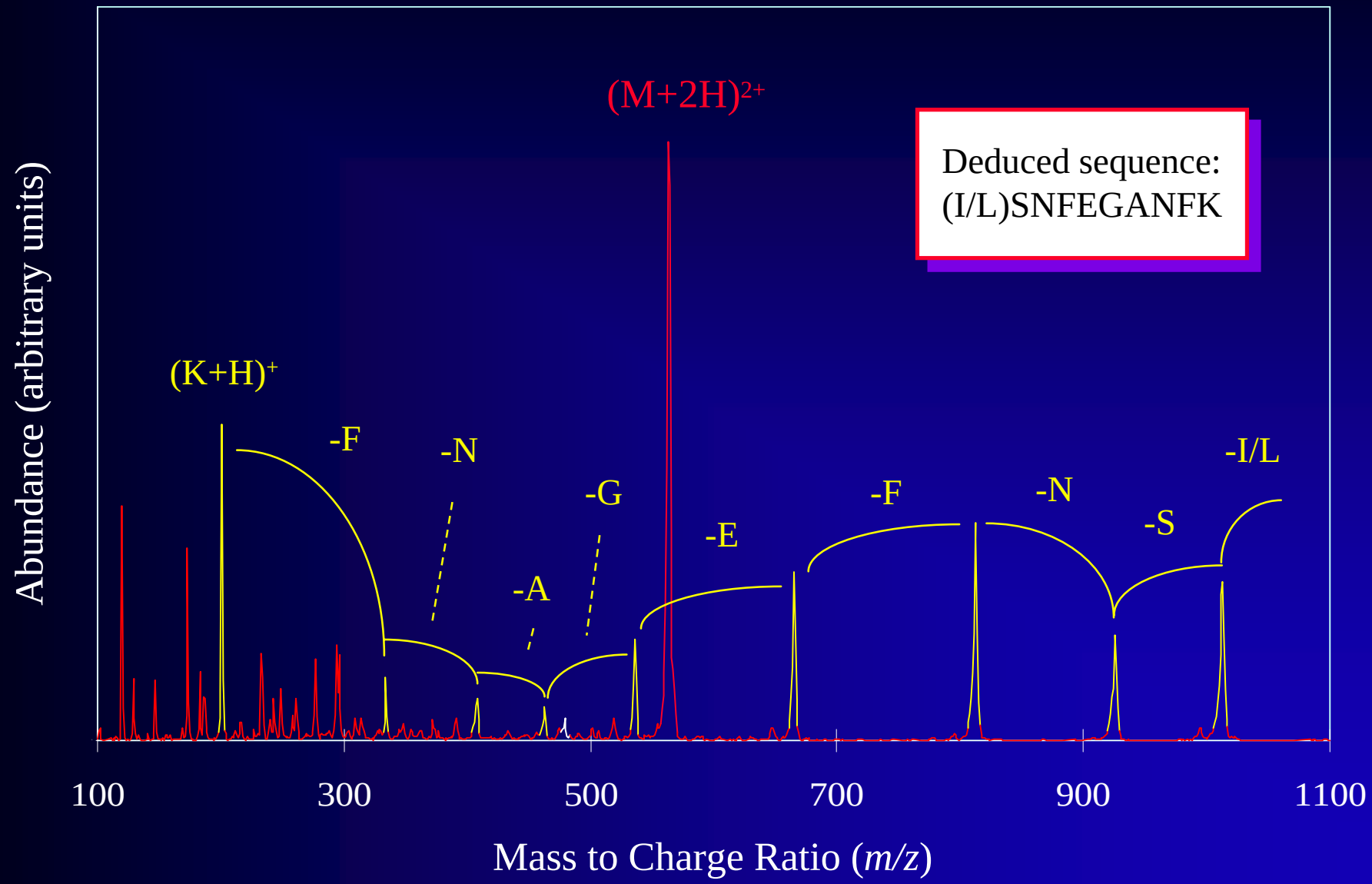
MS/MS


MS-1
CC
MS-2
P1
P2
IS
P3
Ar
F1
F2
F3
Det
P4
P5
Precursor Selection
CAD
Scan Fragments

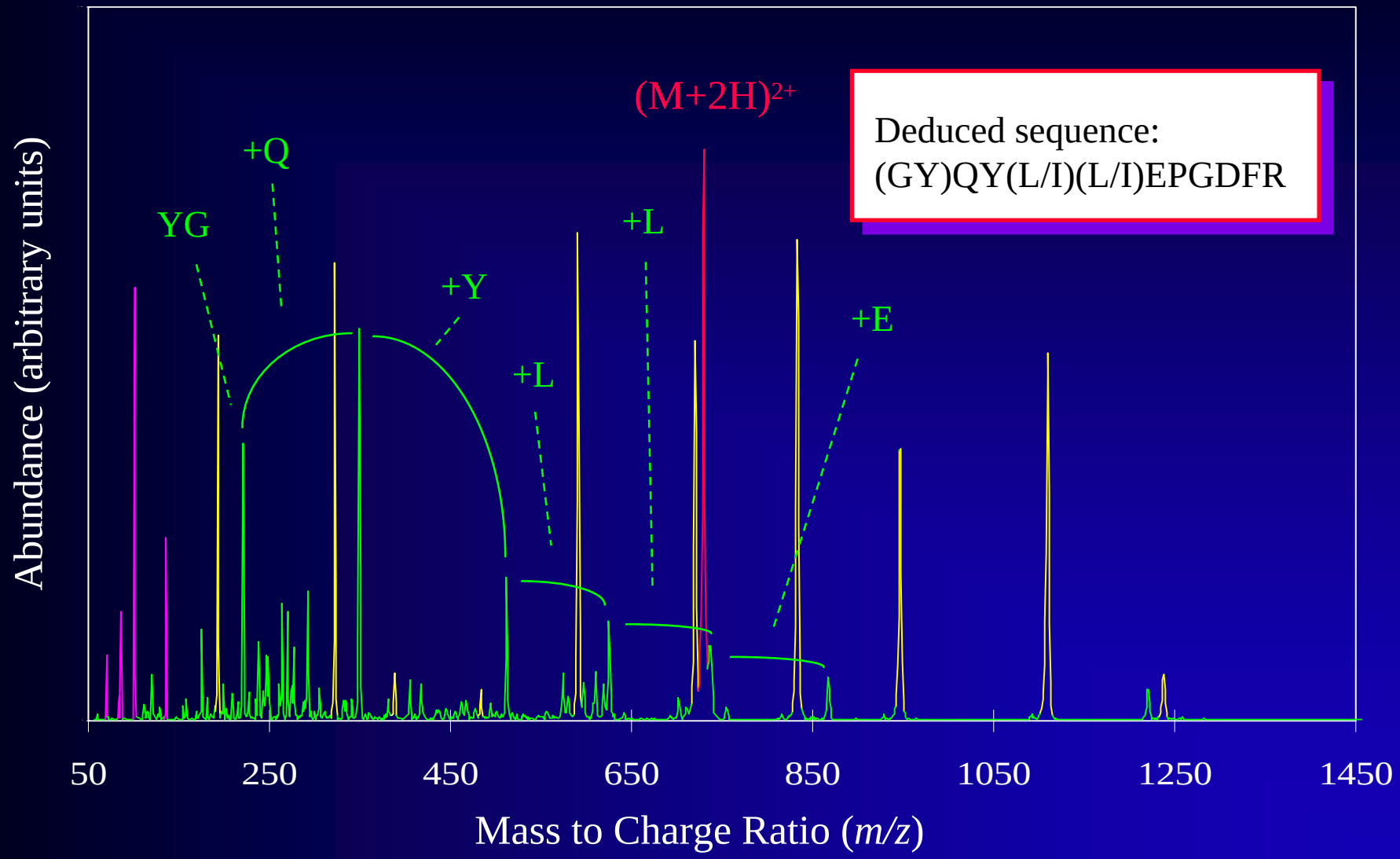
Fragment Ions Observed upon MS/MS Analysis of Tryptic Peptides

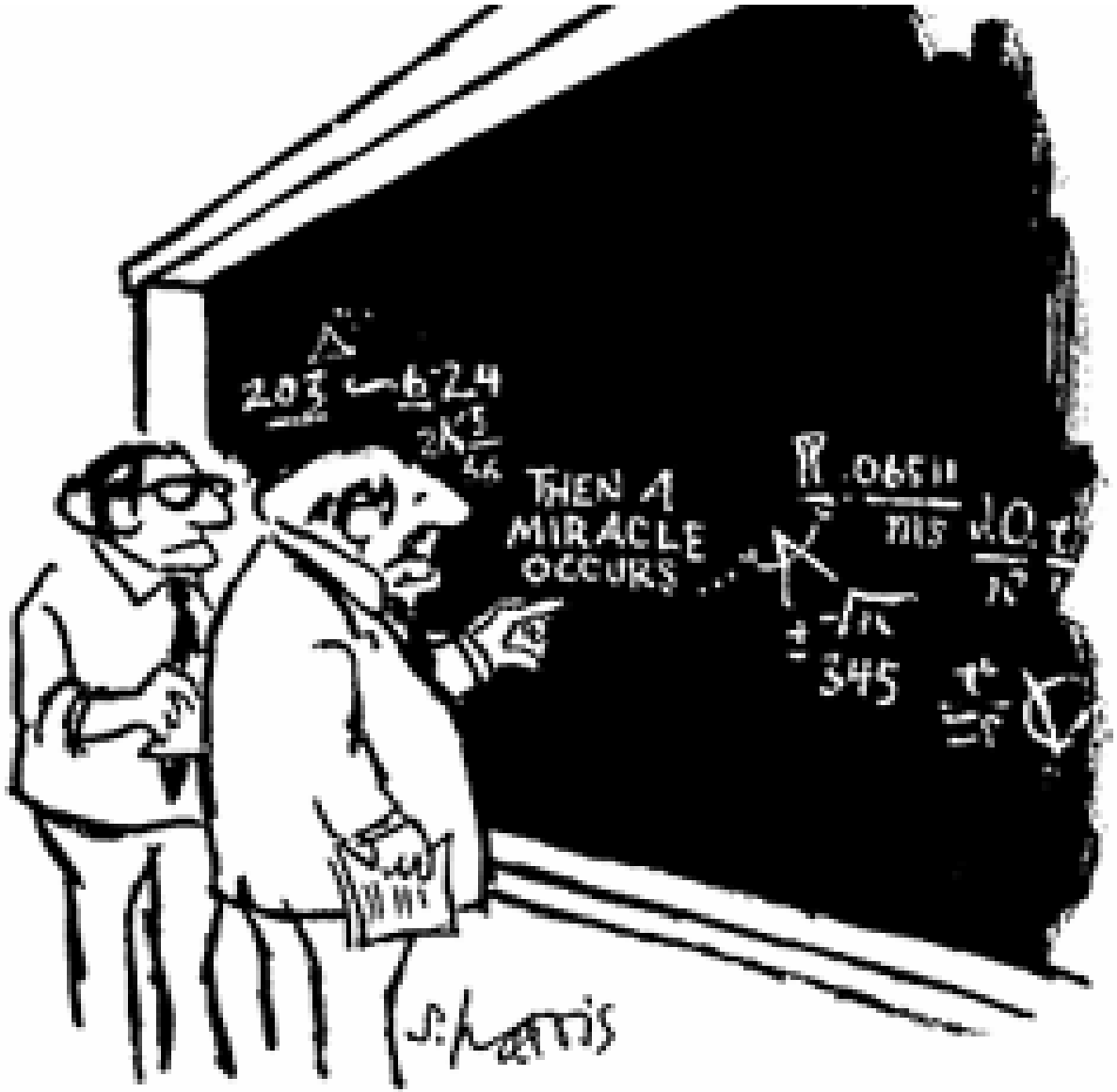


LC/MS/MS Analysis of the tryptic βB1 peptide at 1,126 Da



LC/MS/MS Analysis of the tryptic βB1 peptide at 1,457 Da

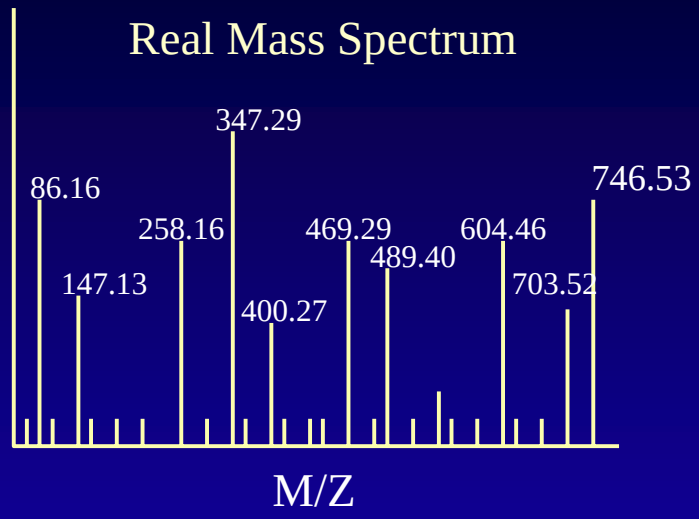


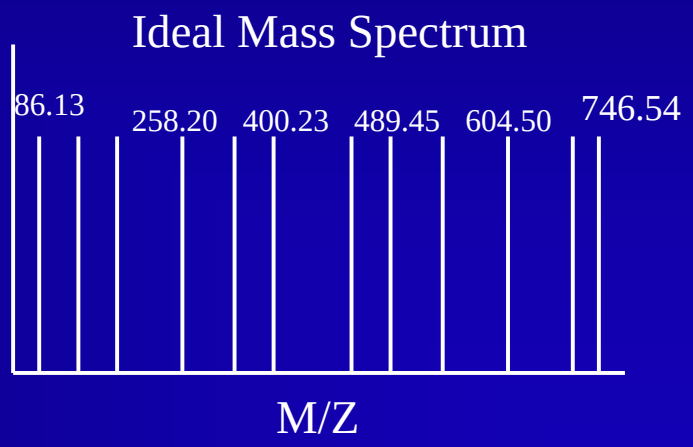


"I think you should be more explicit here in step two."

Spectrum Correlation

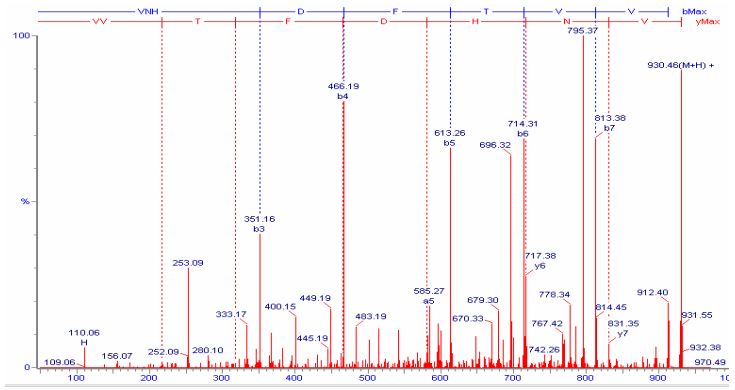
- Scan every peptide sequence in the database for a matching parent ion mass +/- error
- Construct a theoretical MS/MS spectrum for each matching peptide
- Attempt to overlay the real and theoretical mass spectrum
- Assign a score based on similarities
- Two of the major parameters are peptide mass error and fragment mass error





LC-MS/MS Identification of SP-1

MS/MS Spectrum



List of Fragment Ions

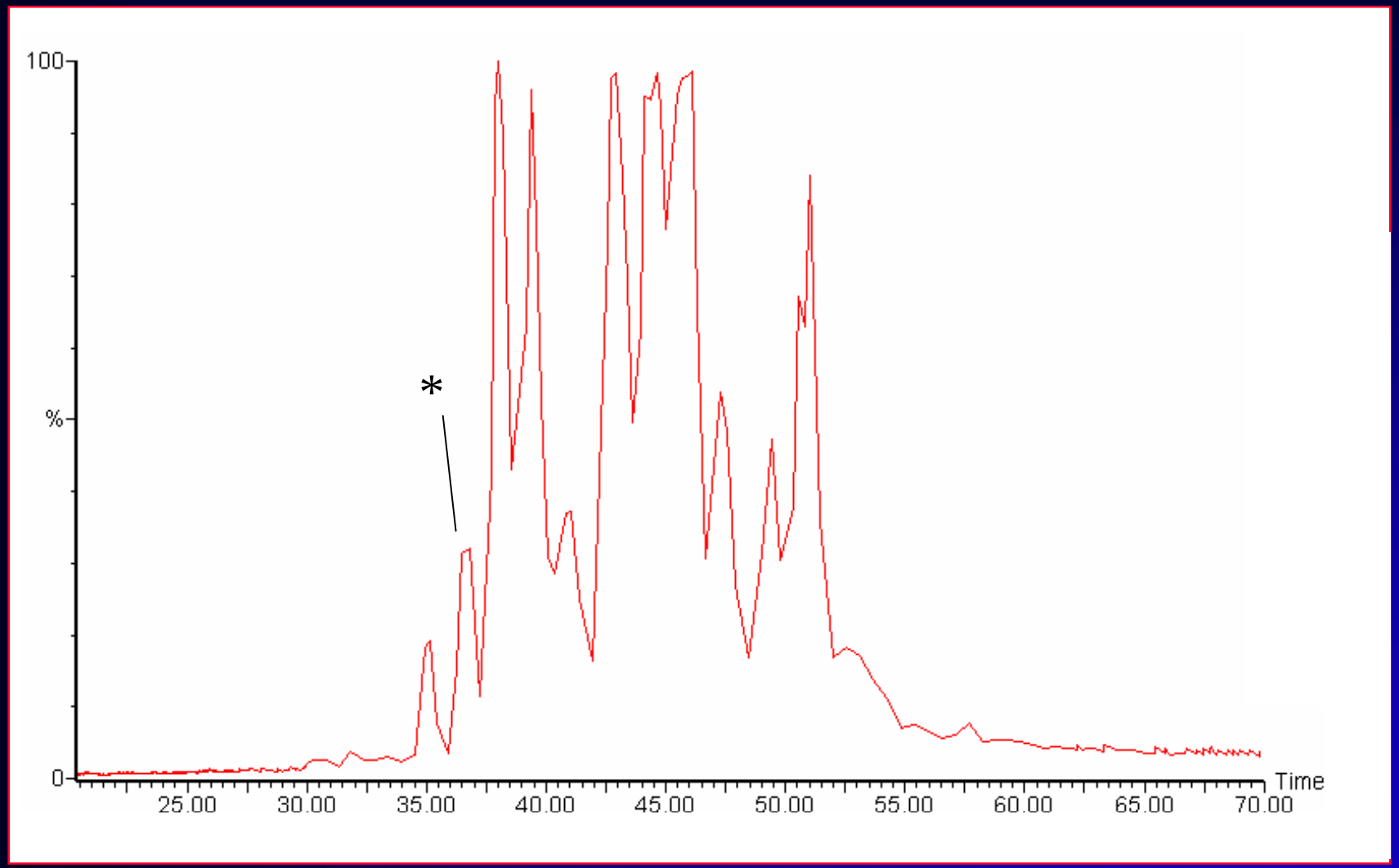
MATRIX SCIENCE Mascot Search Results

User	:	Ron orlando
Email	:	orlando@ccrc.uga.edu
Search title	:	
Database	:	NCBInr 20010107 (601500 sequences; 190198580 residues)
Timestamp	:	25 Feb 2001 at 23:18:18 GMT
Top Score	:	99 for gi|1658195, (U74494) surface protein-1 [Trypanosoma cruzi]

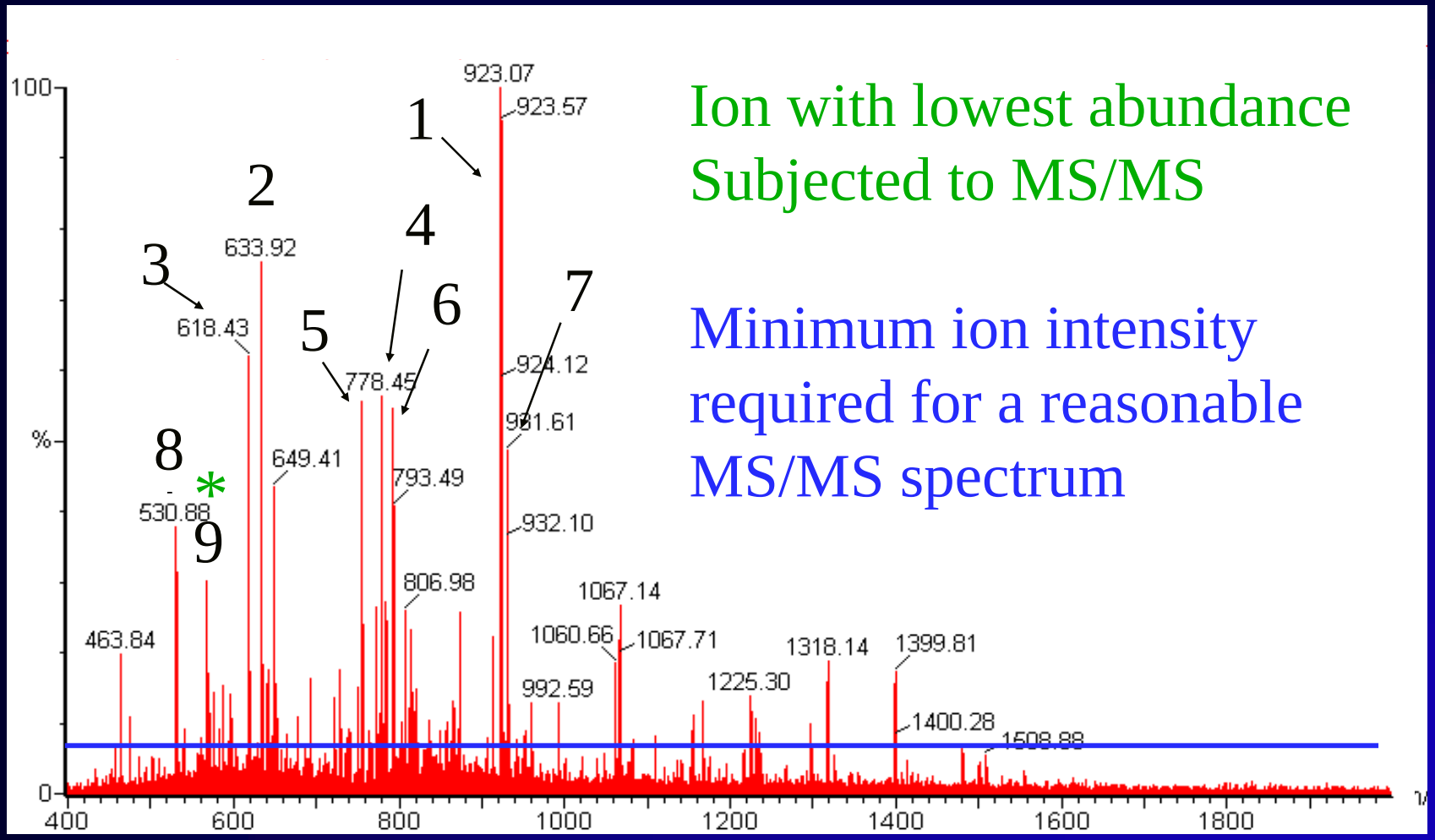
Probability Based Mowse Score

Score is -10*Log(P), where P is the probability that the observed match is a random event.
Protein scores greater than 70 are significant (p<0.05).

RPLC-MS Analysis of SCX Fraction 2



MS of a 2D-LC "Peak"



Ion with lowest abundance Subjected to MS/MS

Minimum ion intensity required for a reasonable MS/MS spectrum

Quotes from a Post I made to ABRF listserve September 11, 2002

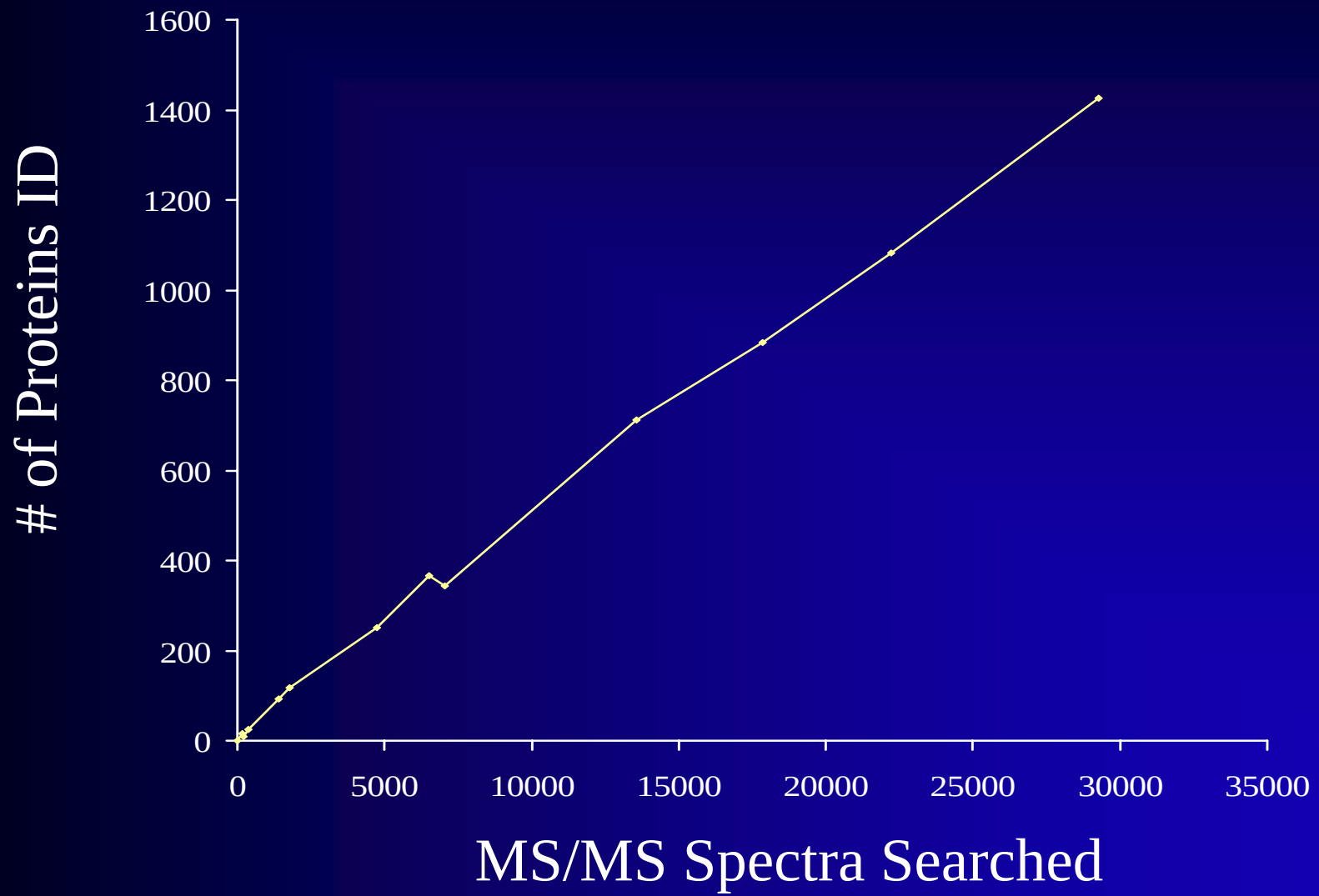
"LC-MS: …This works out to be about 100 proteins per hour of MS time. (This number is going to be important later on so remember the number.)"

"LC/LC-MS: … We collected 10 fractions from the SCX column, and performed ten 1 hour LC/MS runs, and were able to ID about 1,000 proteins, or 100 proteins ID per hour just like in the LC/MS example."

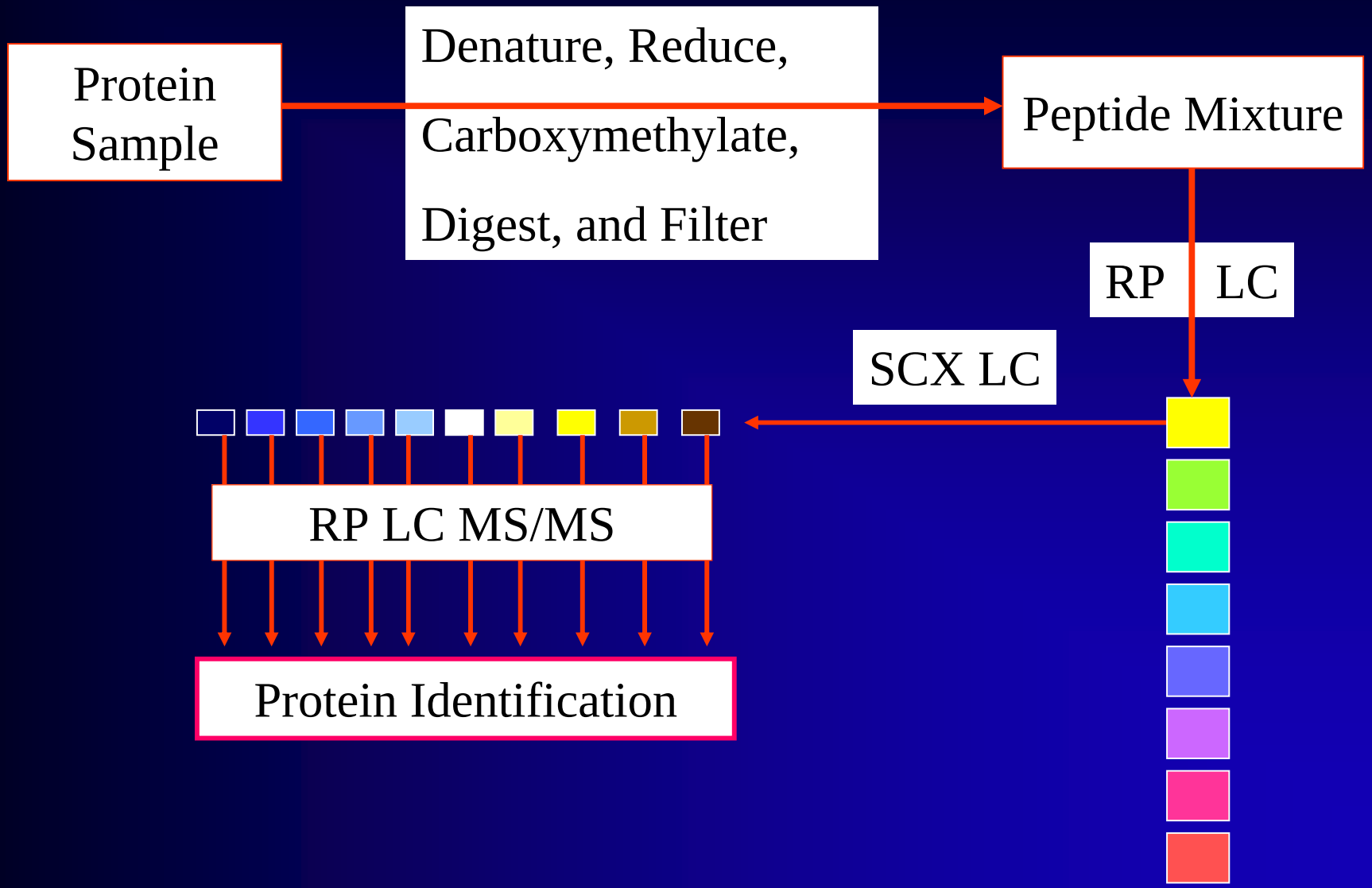
"LC/LC/LC-MS: … Results seem to fit my 100 proteins ID per hour LC/MS time theory

Proteins Identified vs. Spectra Searched

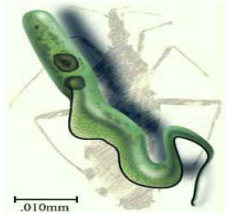
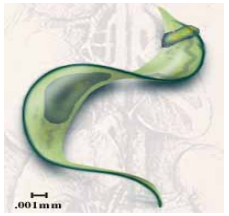
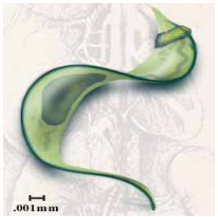
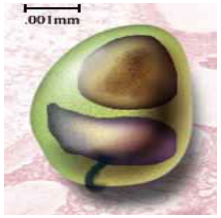
Human proteins identified with 95% probability



Multidimensional LC-MS/MS Proteomics



MS/MS Sampling by Life-cycle Stage

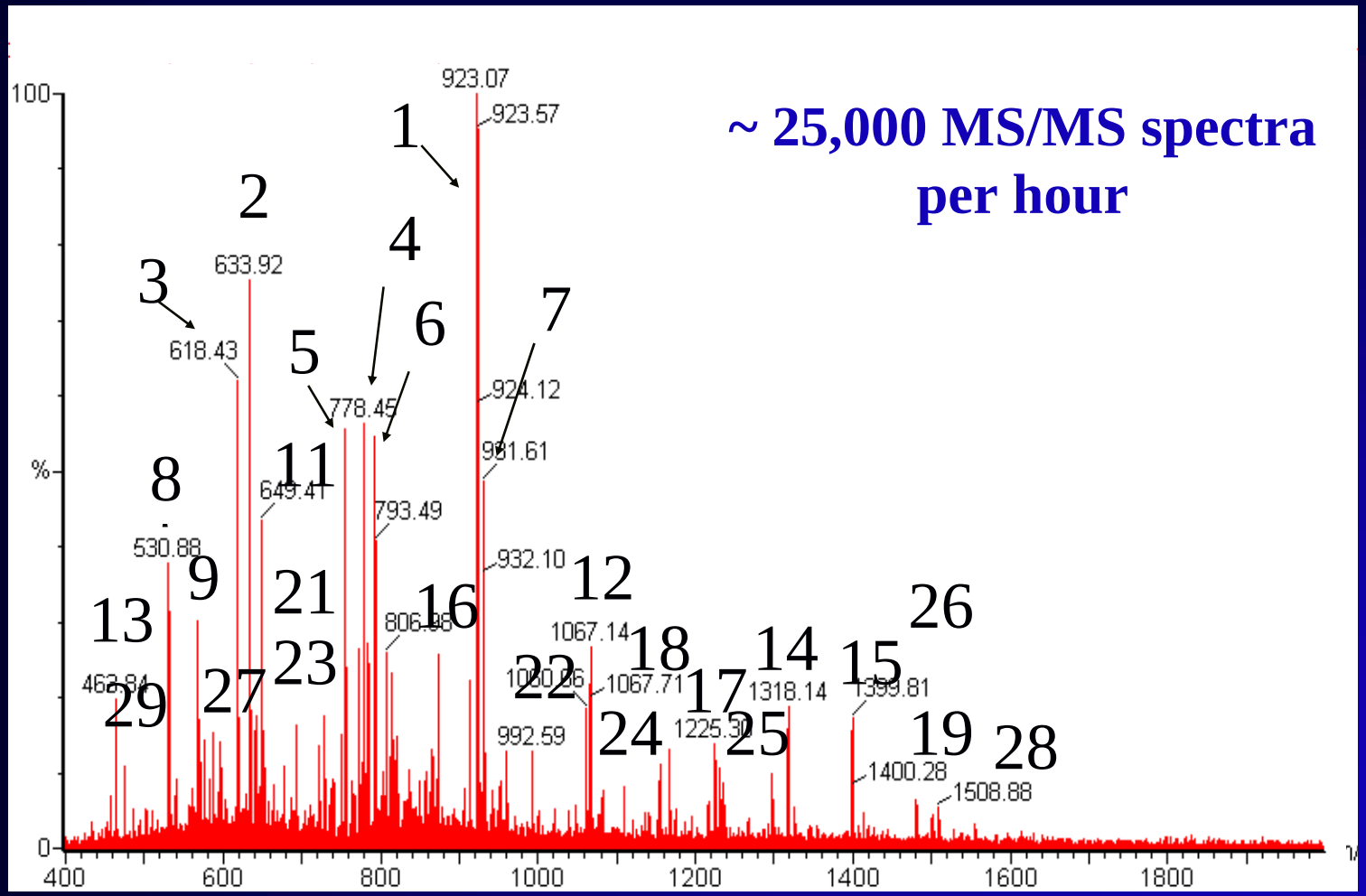
					
	Epimastigote	**Trypomastigote**	**Metacyclic**	**Amastigote**	**Total**
Spectra collected	54149	20585	38979	25434	**139,147**
Spectra matched	5857	2143	5737	3488	**17,225**
Unique peptides	2456	1453	3202	1911	**5792**
Unique proteins	1573	1194	2064	1576	**2784**

Epimastigotes most highly sampled

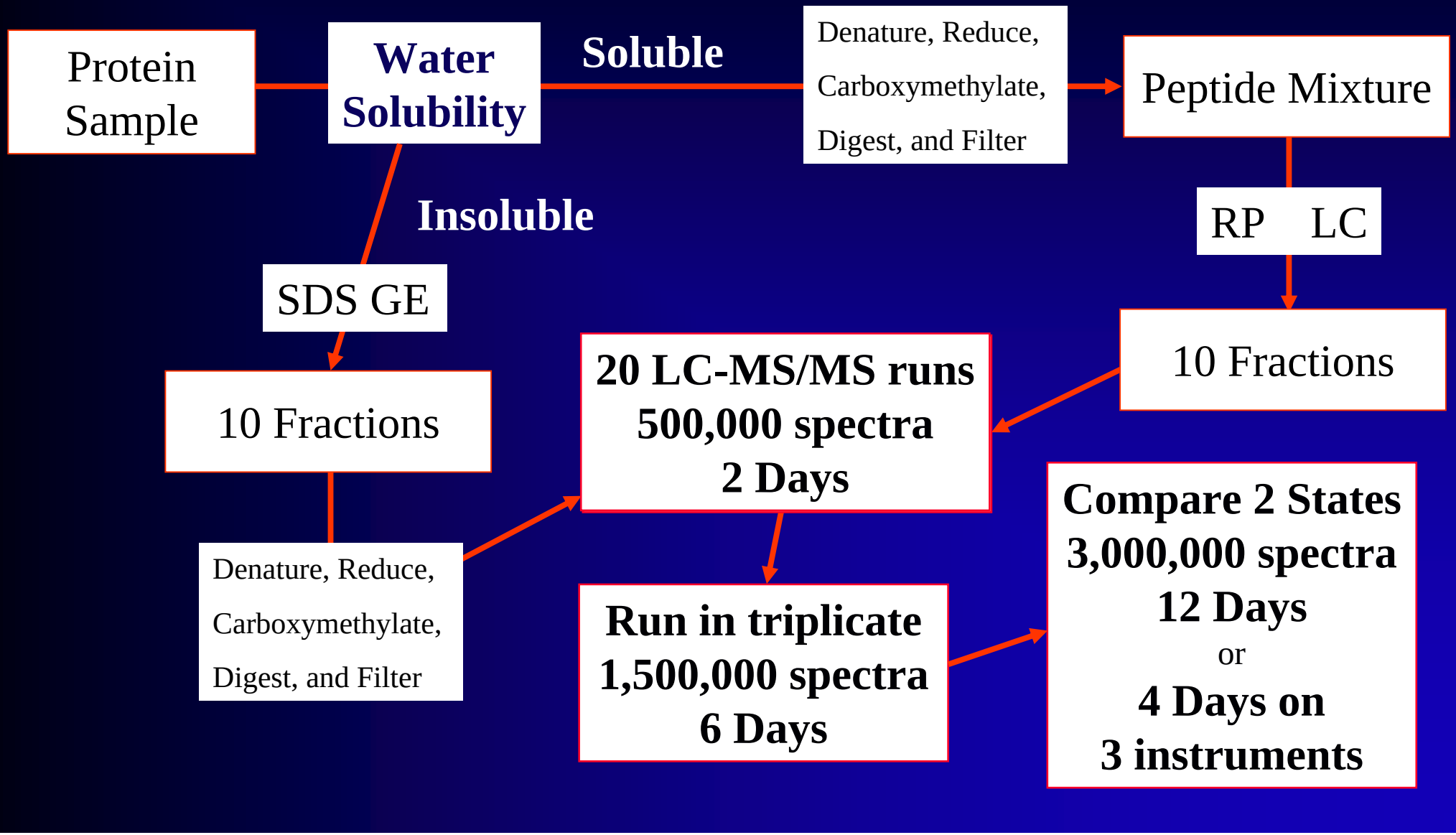
Trypomastigotes most under sampled – important for identifications only in this stage

The 2784 proteins were further sorted into 1168 protein groups “families”

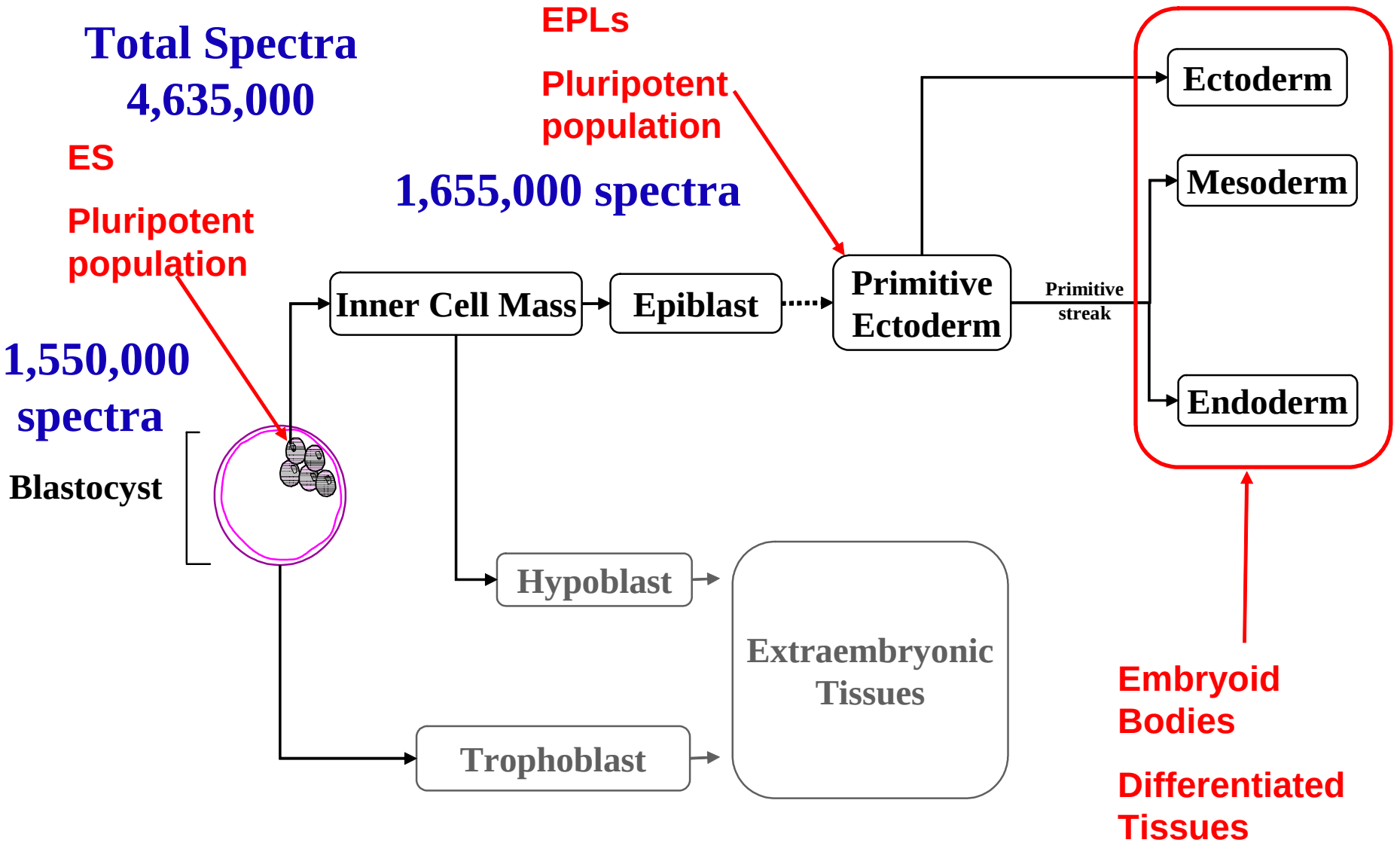
MS of a 2D-LC “Peak”: 2D Ion Trap



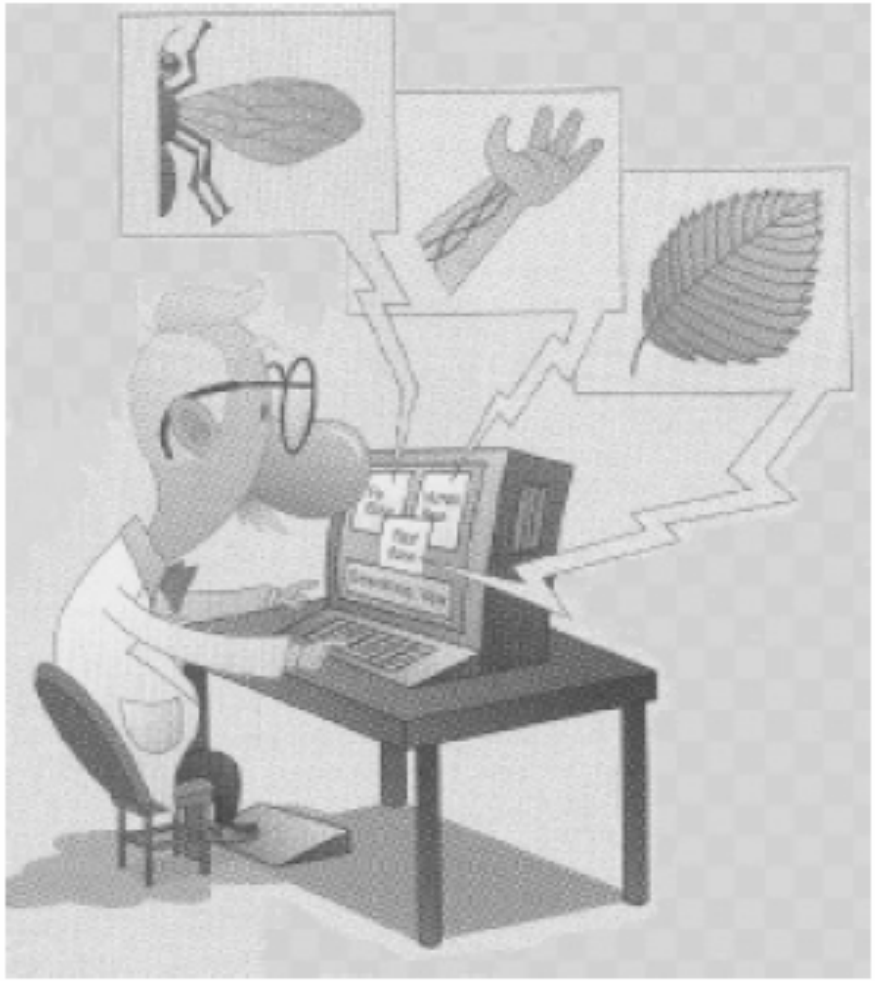
A Typical Proteomics Experiment



Embryonic Stem Cells as an *In vitro* model for Embryogenesis[1,2]



1. Gilbert, S.F. *Developmental Biology*. 4th Ed., Sinauer Associates: Sunderland, MA, 1994.
2. Gardner, R.L., *J Cell Science* Suppl. 10, 1988, 11-27.



Searching in "vain."
Computers can't distinguish when a word has different meanings, confounding data quests.

Science, 15 Oct 99

Calculating Probabilities in Proteomics

Determining the probability of an match is difficult.

It is relatively easy to compute the probability of a random hit.

Search Real Data against a Random Data Base or

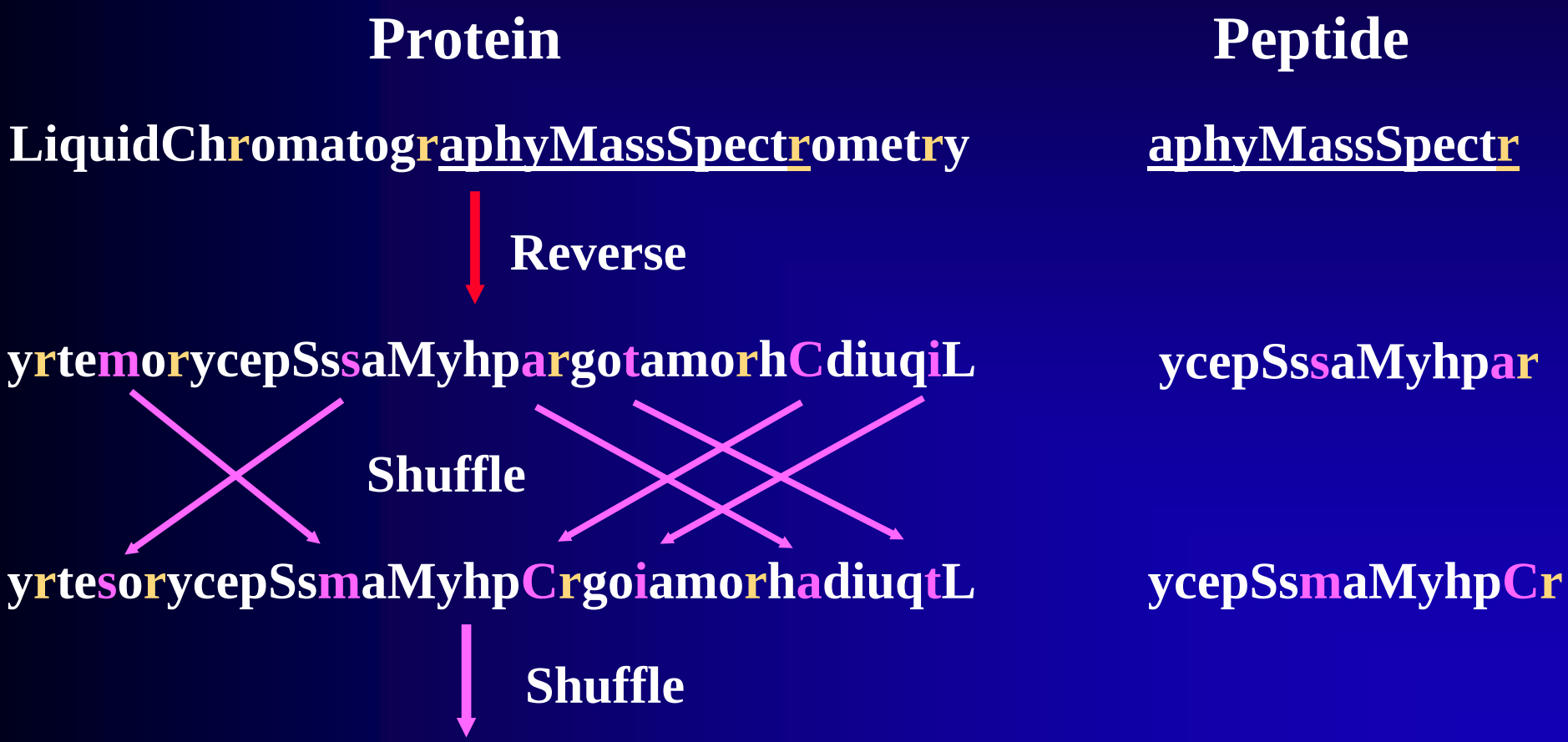
Random Data against a Random Data Base or …

These two values are related by

Probability (Random) = 1 – Probability (Not Random)

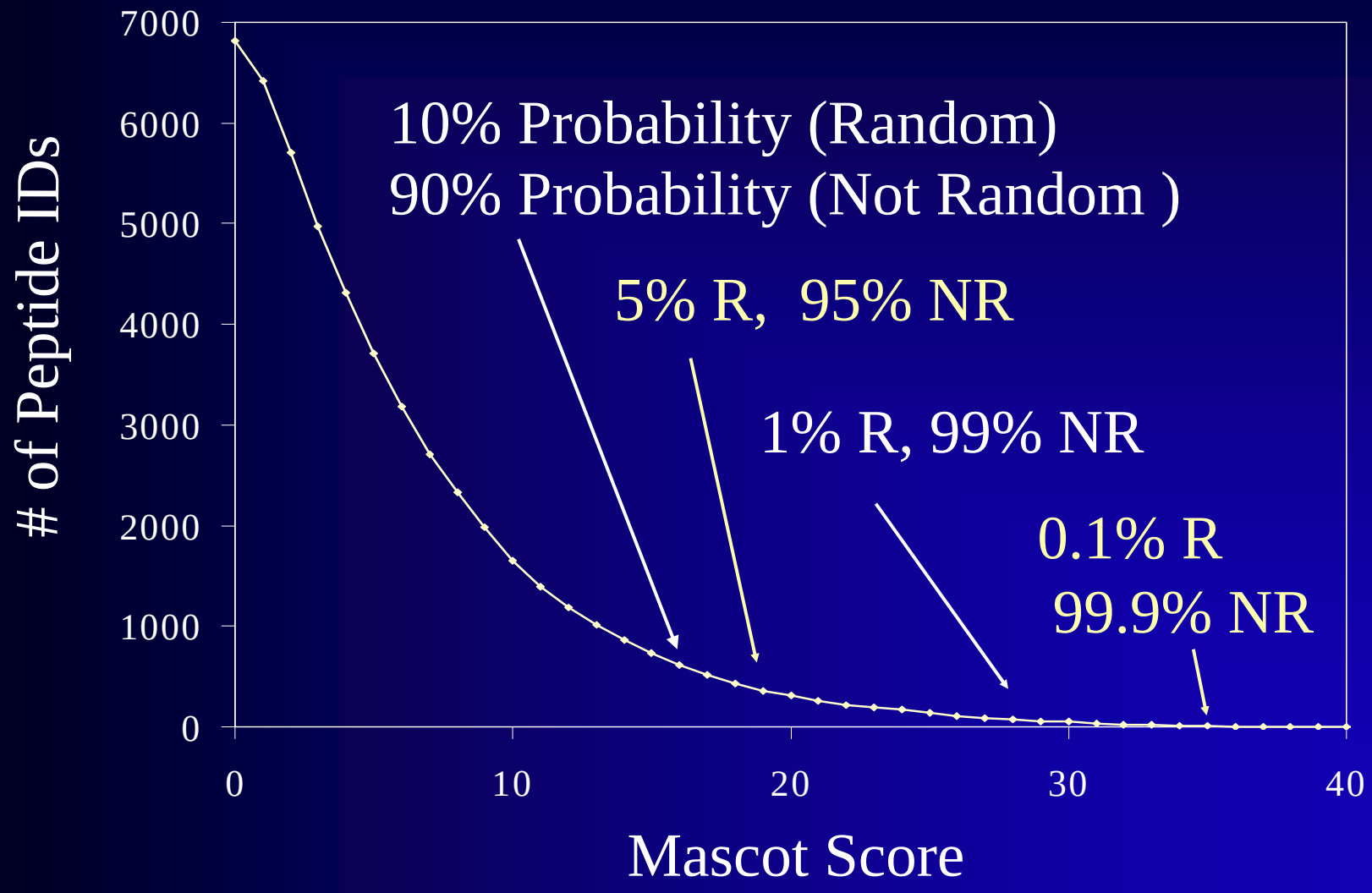
Probability (Random) is often called False Positive Rate

Random Database Creation

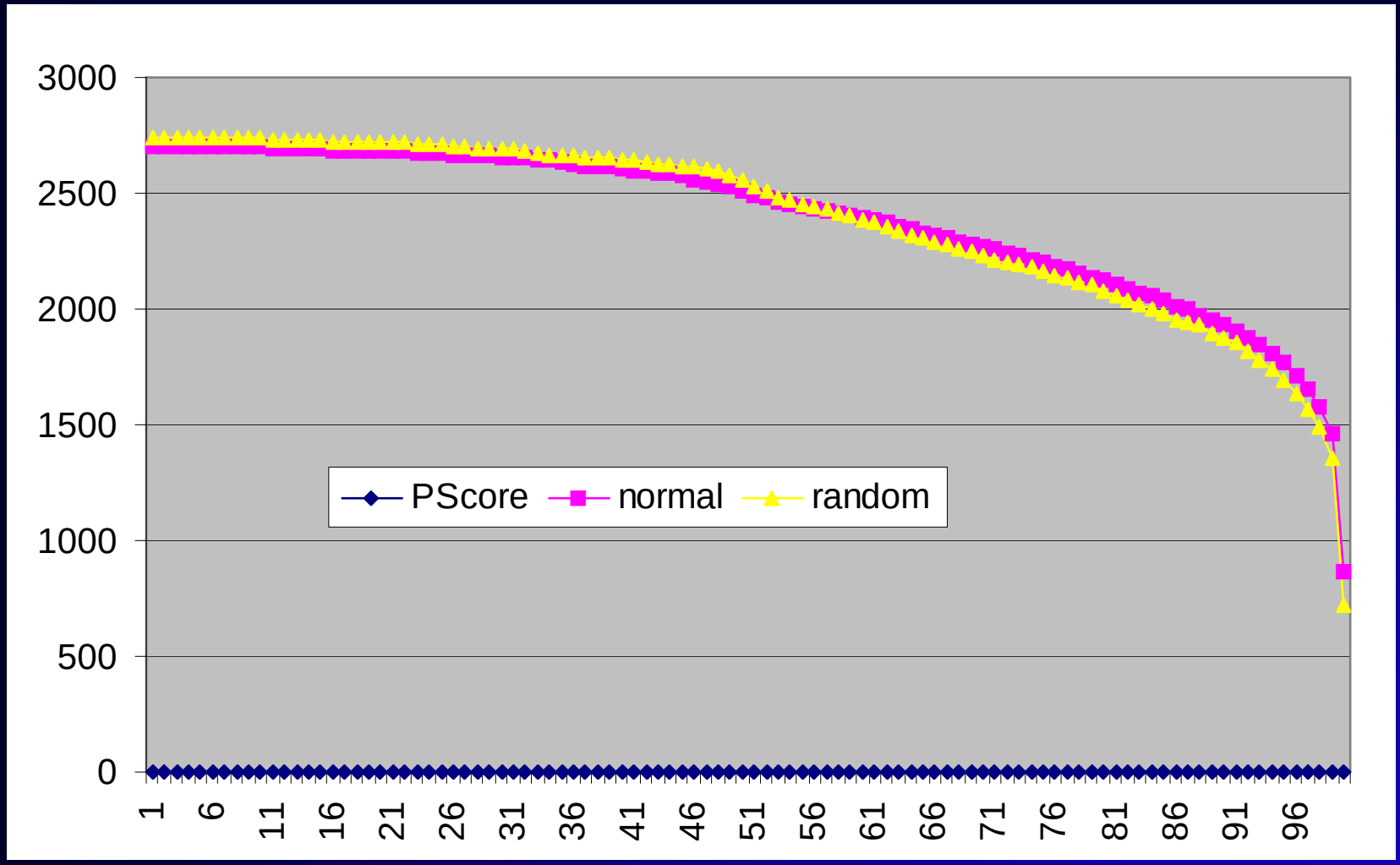


Calculating the Probability that an ID is Not Random

Real Data searched against a Random Data Base



Results from Another Search Program



How Many of my “Hits” are Random?

MS/MS Spectra 48,402

Peptide IDs (99% probability) 2,964

Invalid Calculation of Random Events

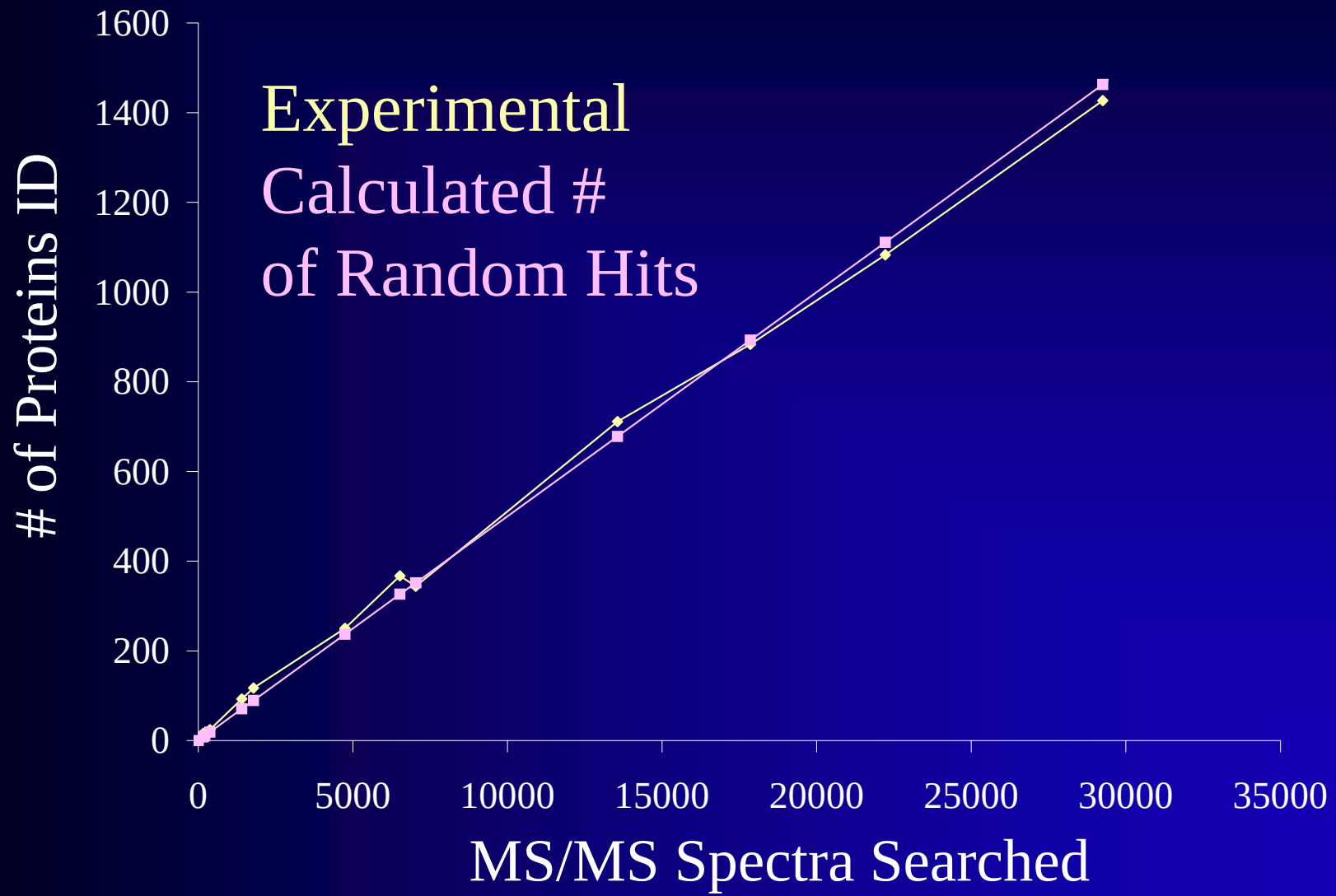
2,964 x 0.01 = 29.6

Valid Calculation of Random Events

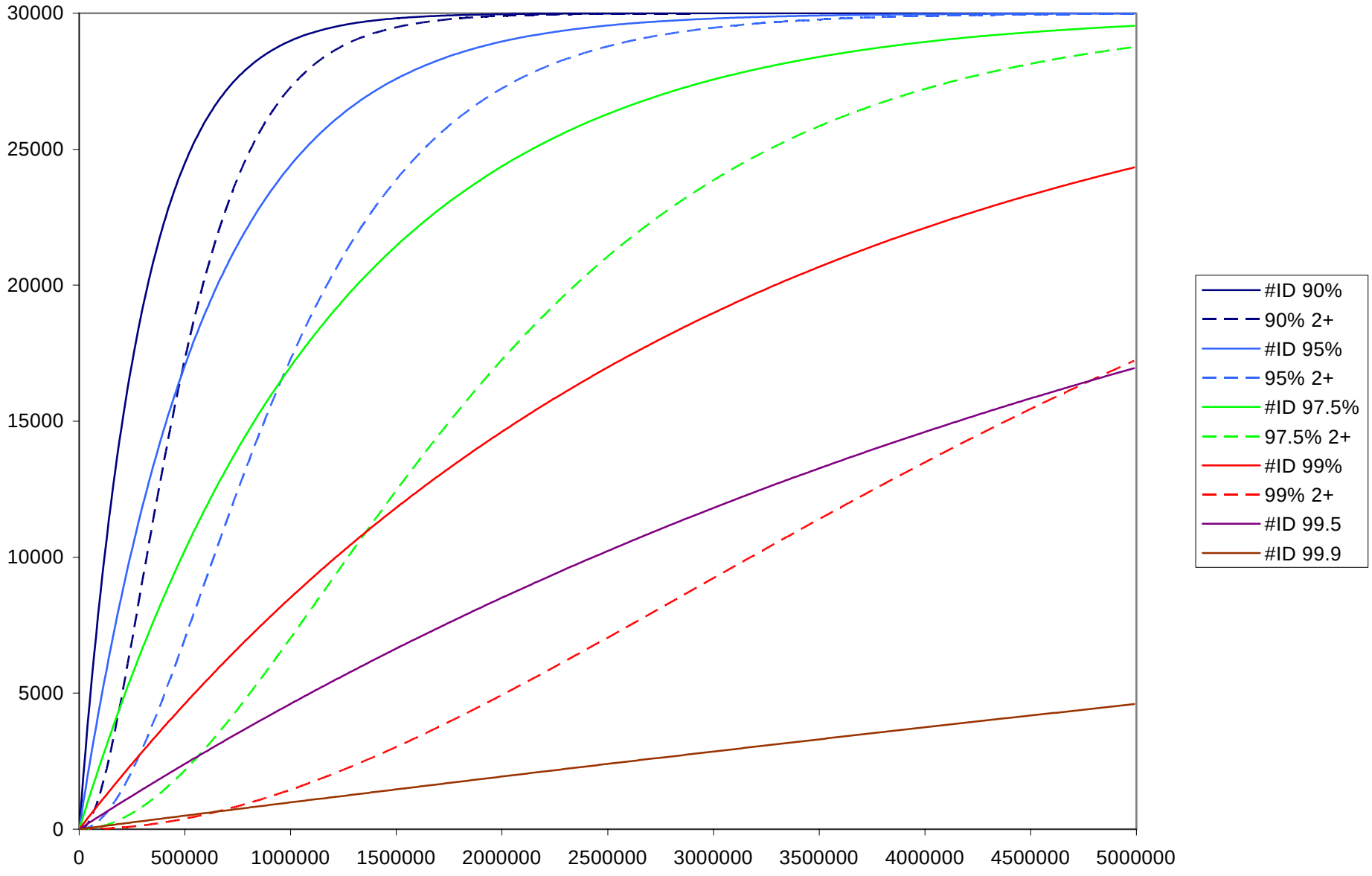
48,402 x 0.01 = 484

Proteins Identified vs. Spectra Searched

Human proteins identified at with 95% probability



Random Proteins Identified vs. Spectra Searched



The "law of large numbers" refers to the principle that unlikely outcomes become likely when an event is repeated a large number of times.

The odds that you will win the lottery are very low;

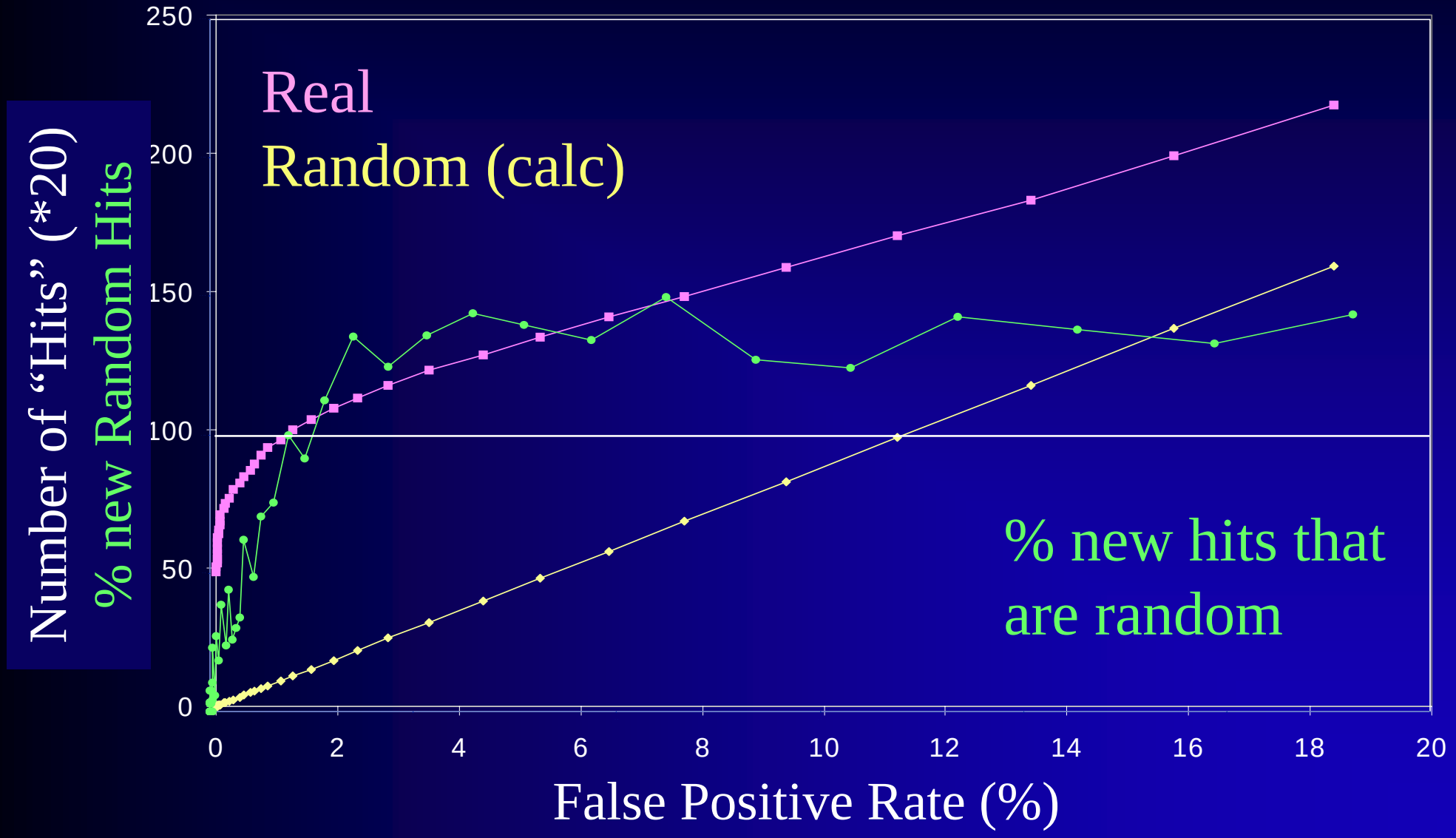
Power Ball Grand Prize Probability:
1 in 120,526,770

However, the odds that someone will win the lottery are quite good, provided that a large number of tickets are purchased.

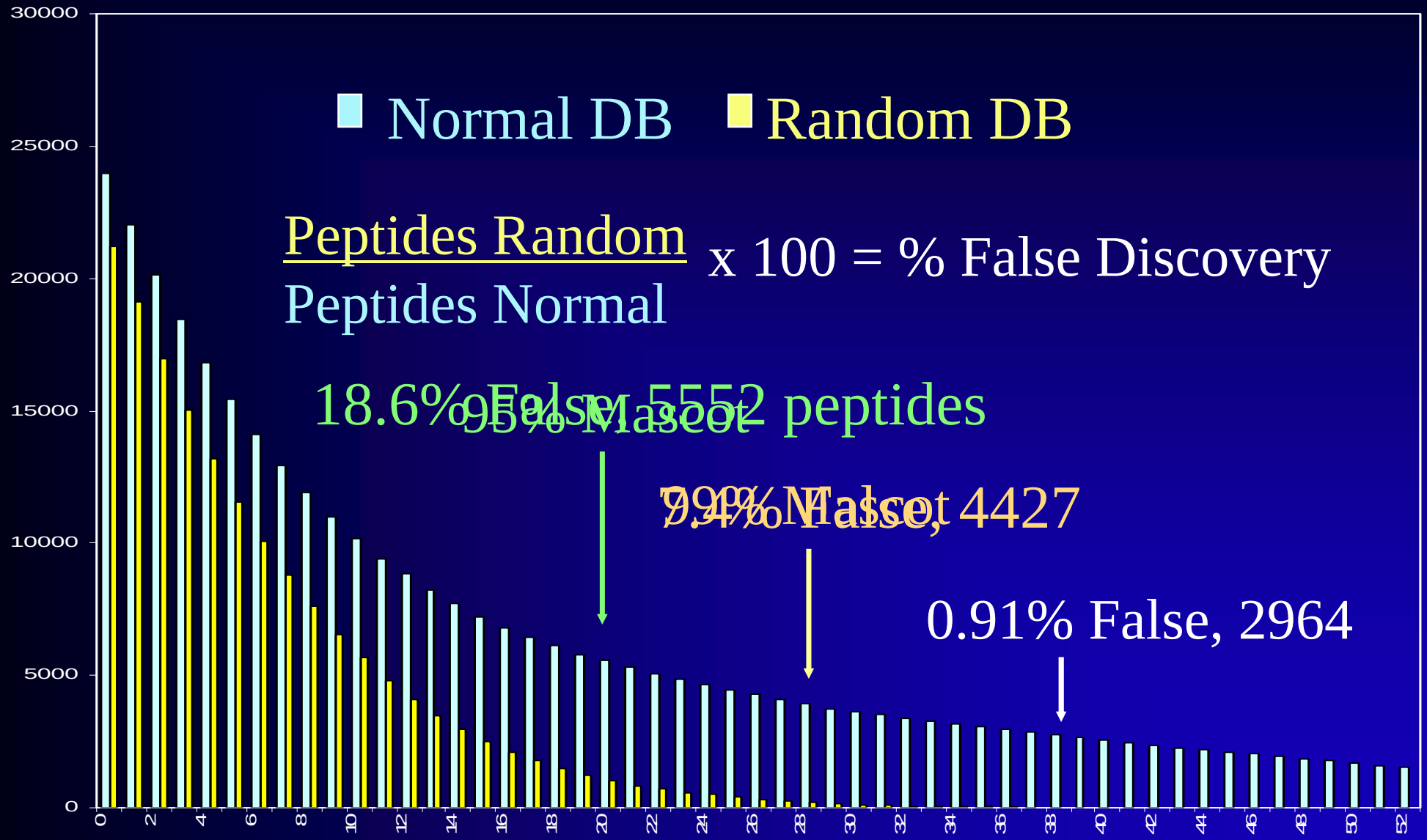
Times won in 2005: **14**

Diminishing Returns

What FPR should I use?



False Discovery Rate



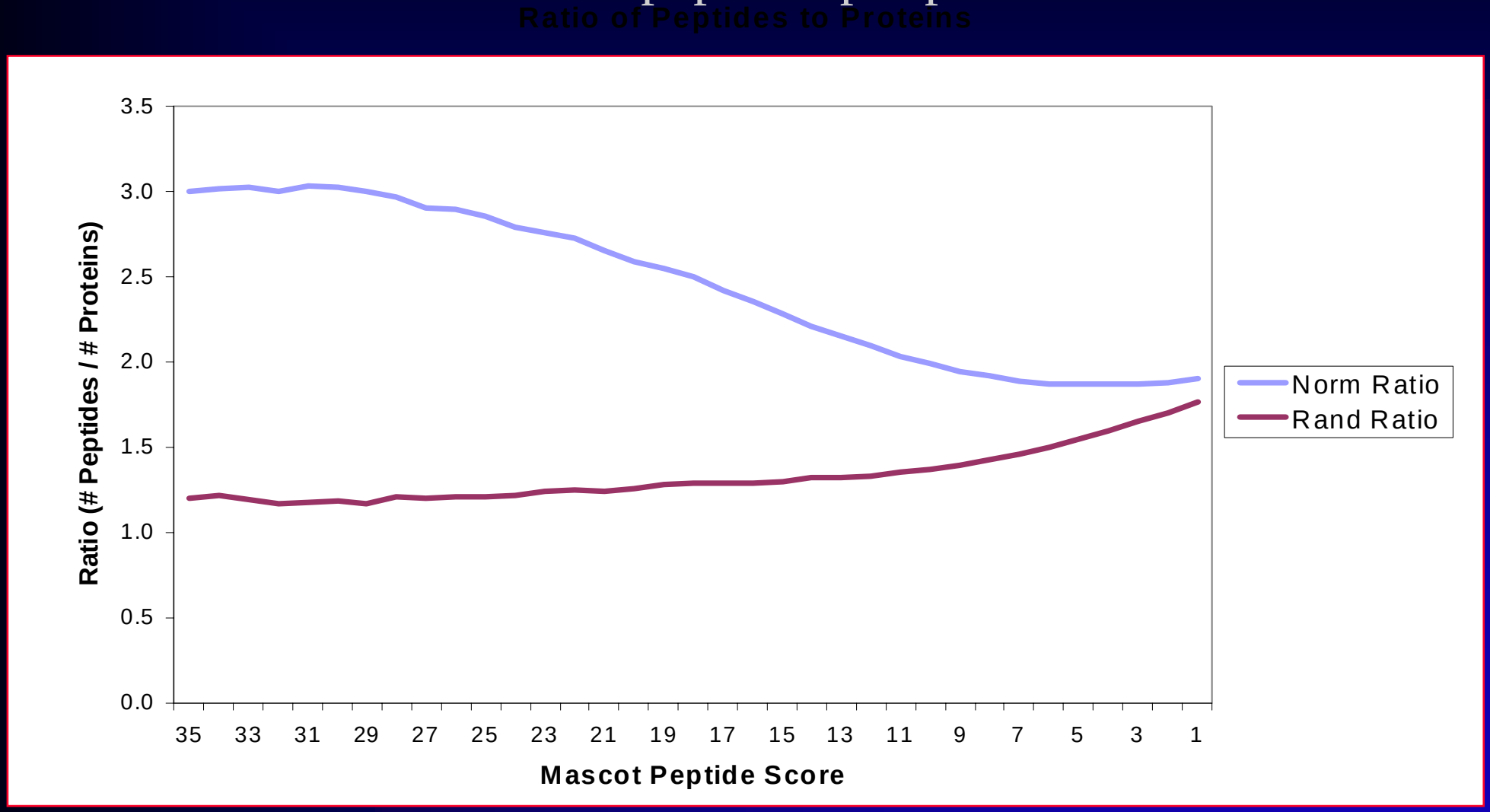
Protein ID/Data Base Search Strategies

- Score per Peptide Cut off – Protein is ID if 1 or more peptides has a score > X
- Cumulative Score for Protein – Protein is ID if all the peptide scores add to a score > Y
- Combination

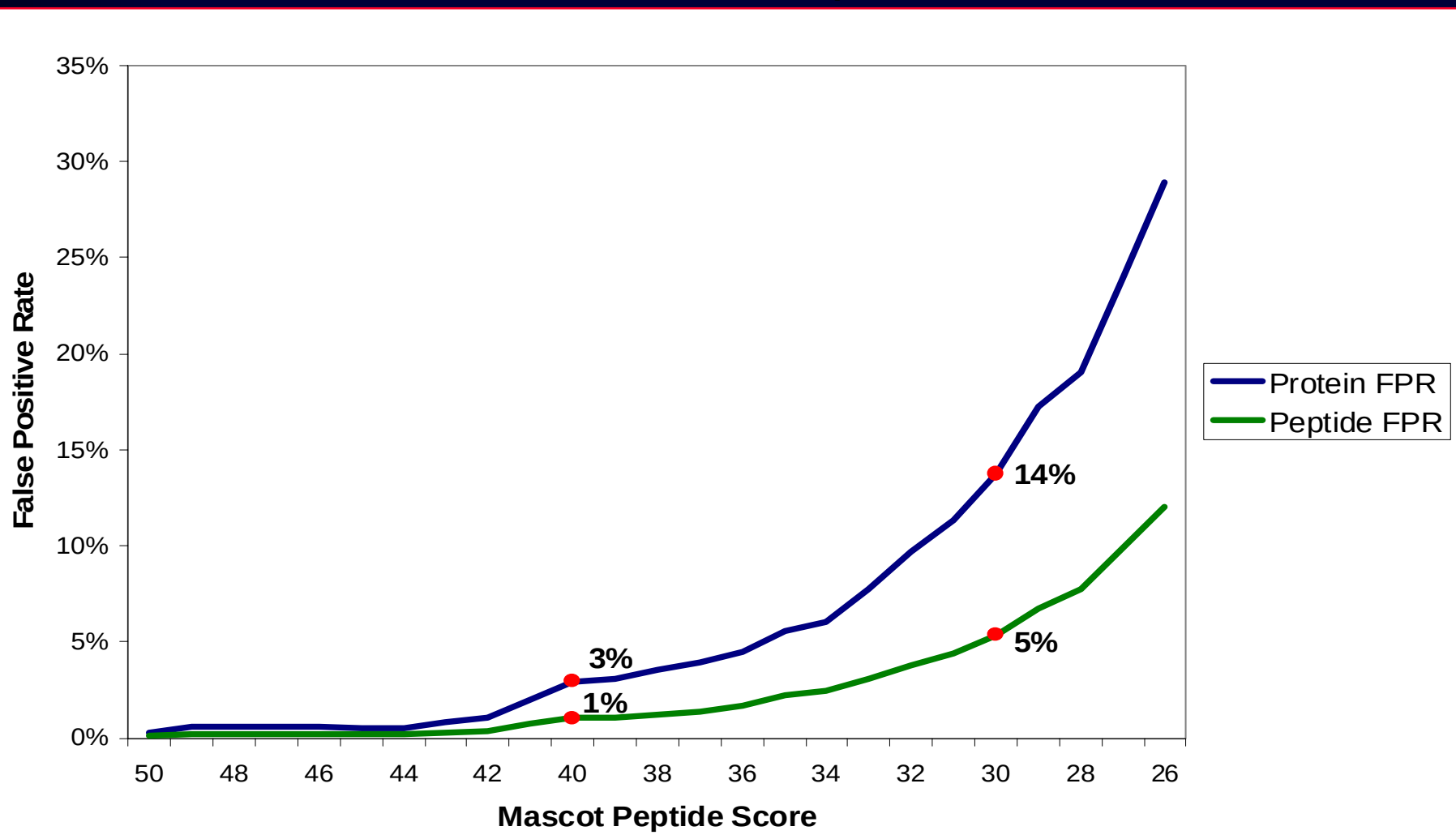
Visual inspection to validate all Protein IDs

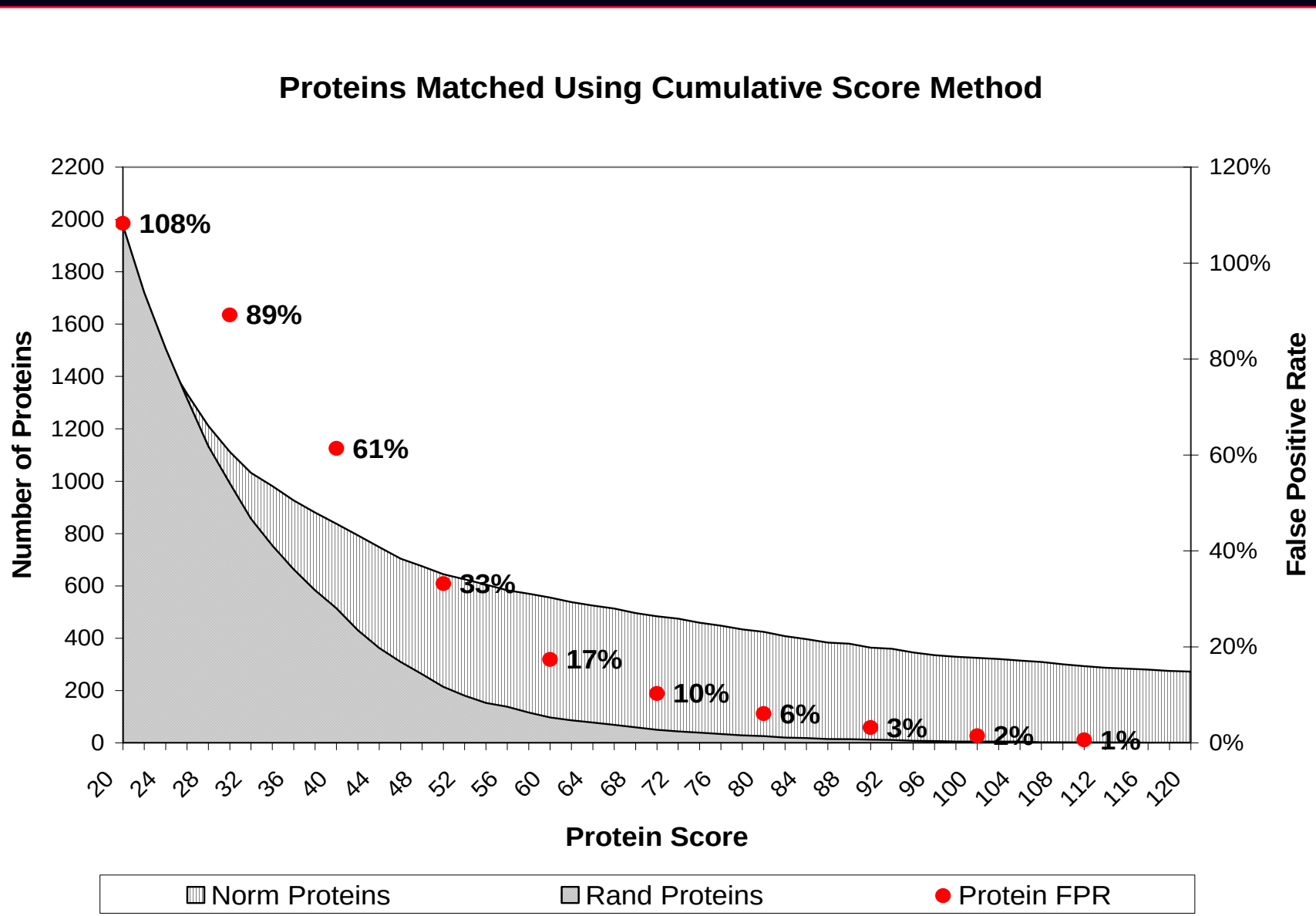
Effect of False Positive Peptides on Protein ID

number of peptides per protein



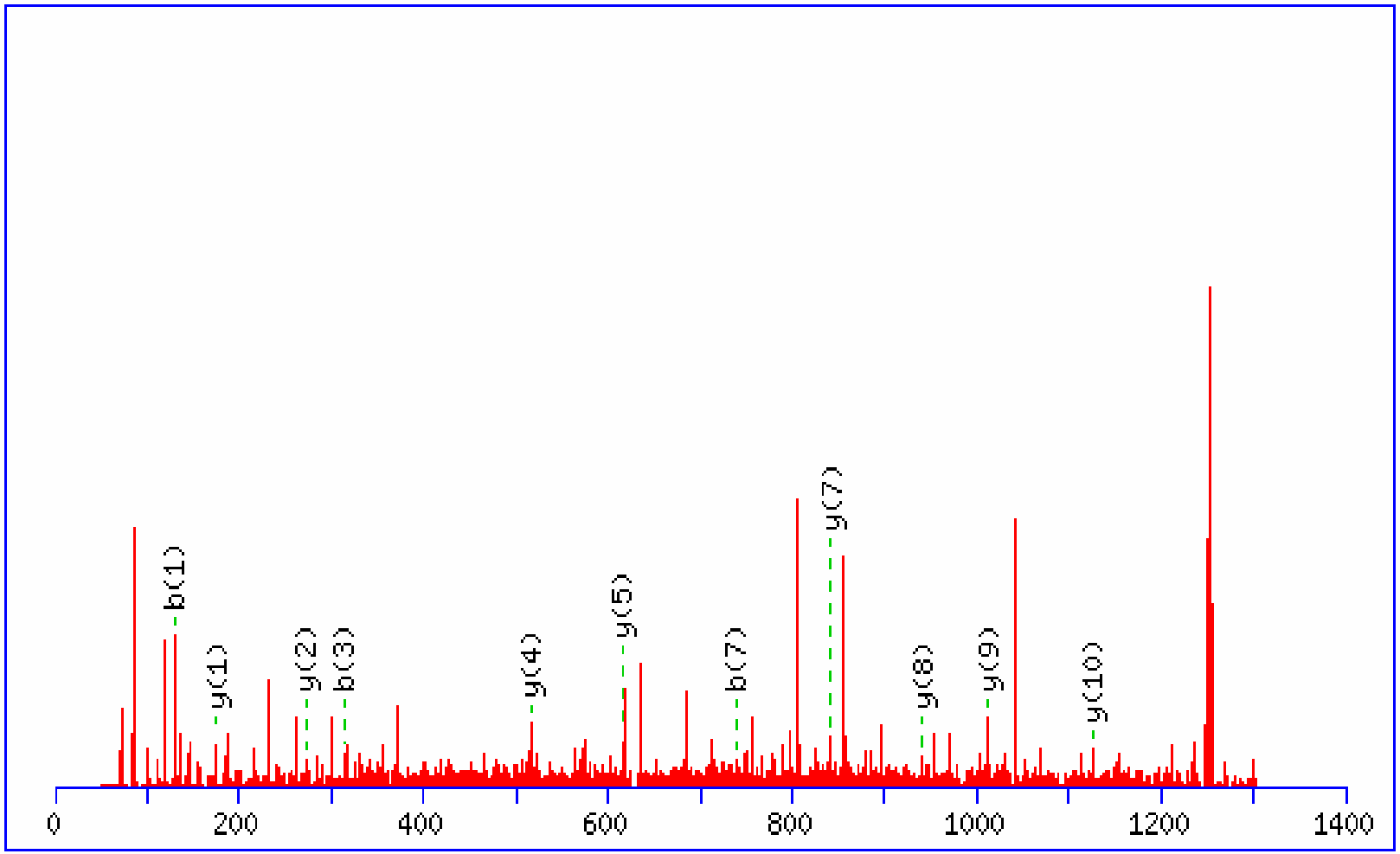
Effect of False Positive Peptides on Protein ID




Proteins Matched Using Cumulative Score Method
Number of Proteins
False Positive Rate
Protein Score
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
120%
100%
80%
60%
40%
20%
0%
20
24
28
32
36
40
44
48
52
56
60
64
68
72
76
80
84
88
92
96
100
104
108
112
116
120
108%
89%
61%
33%
17%
10%
6%
3%
2%
1%
Norm Proteins
Rand Proteins
Protein FPR

Is Manual Validation Possible?

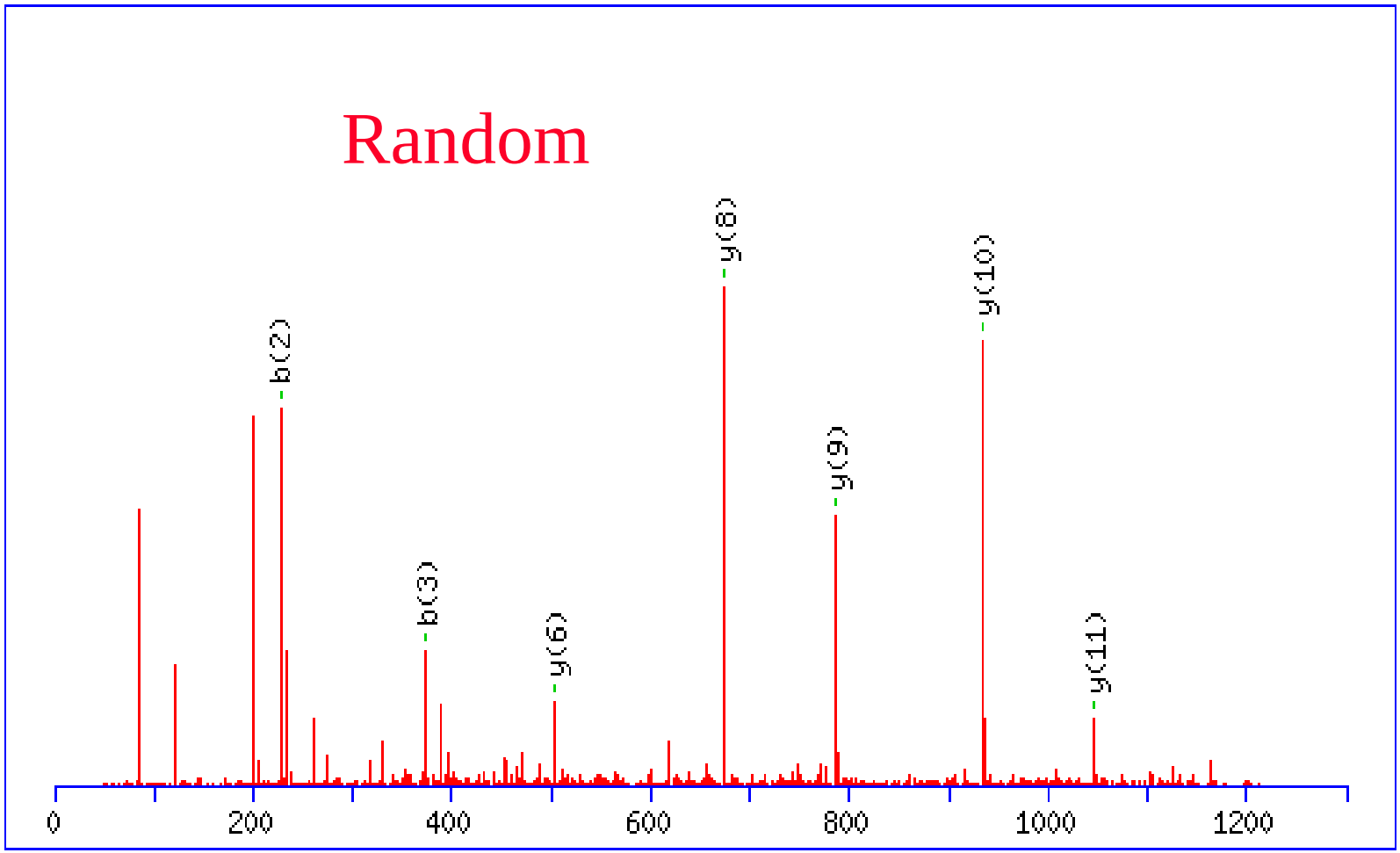
- LTQ - ~ 25,000 MS/MS spectra per hour
- 5% need to be visually inspected
- Instrument operates 24/7
- You work for 40 hours a week
- 1.4 seconds per spectrum



Monoisotopic mass of neutral peptide (Mr): 1252.77

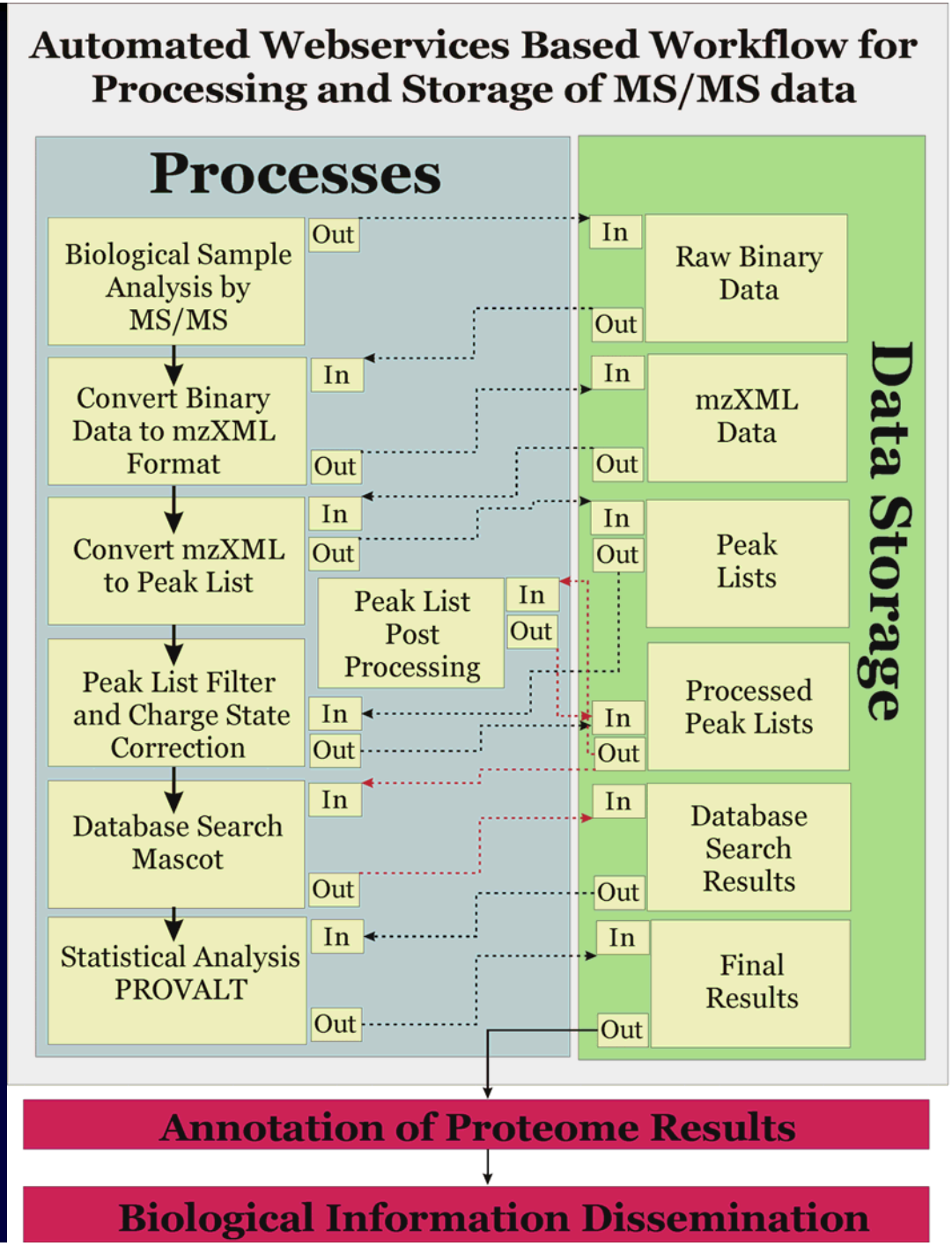
Ions Score: 9 **Matches (Bold Red):** 11/80 fragment ions using 92 most intense peaks

How Many are Valid?

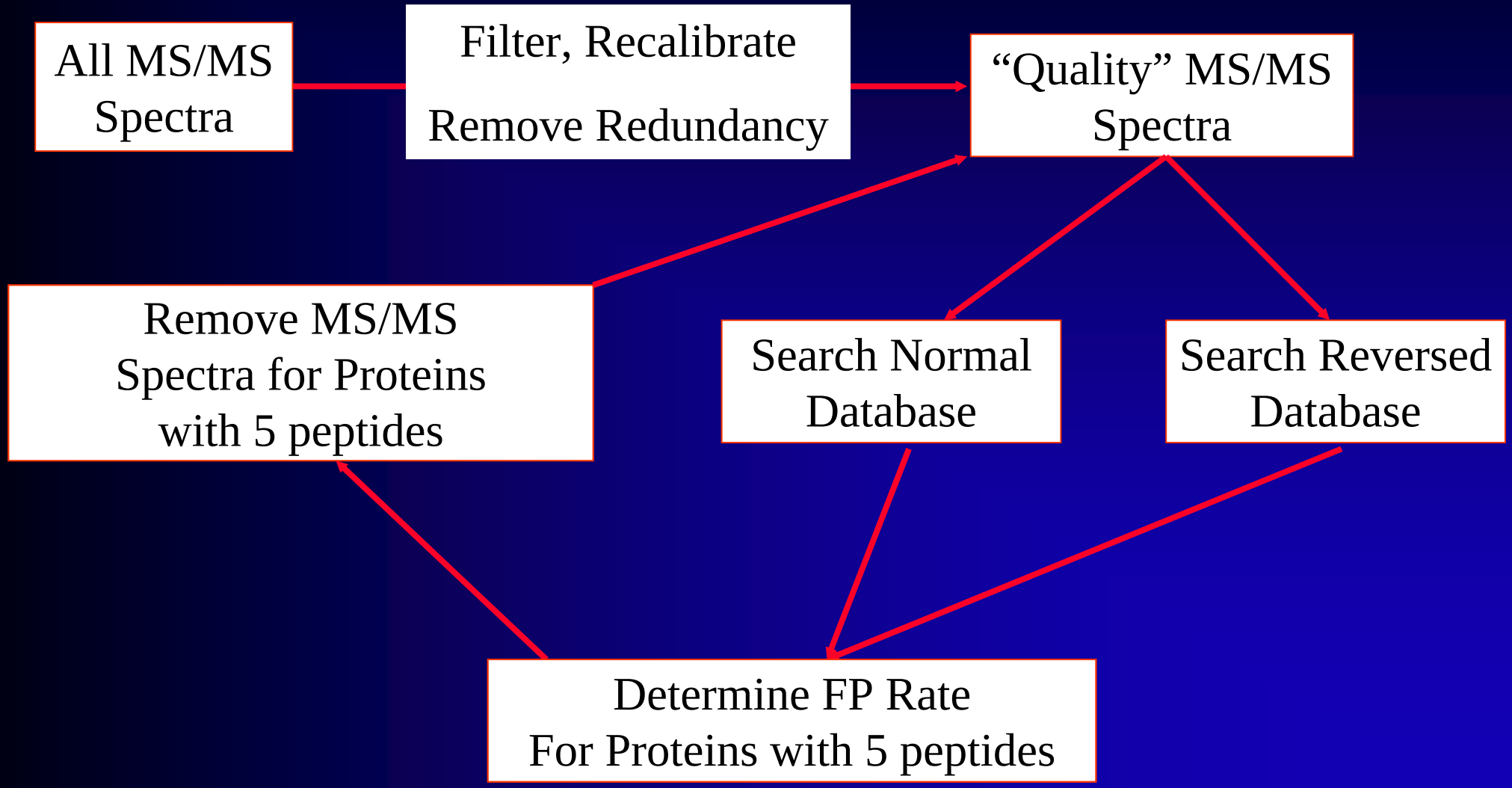


Monoisotopic mass of neutral peptide (Mr): 1159.67

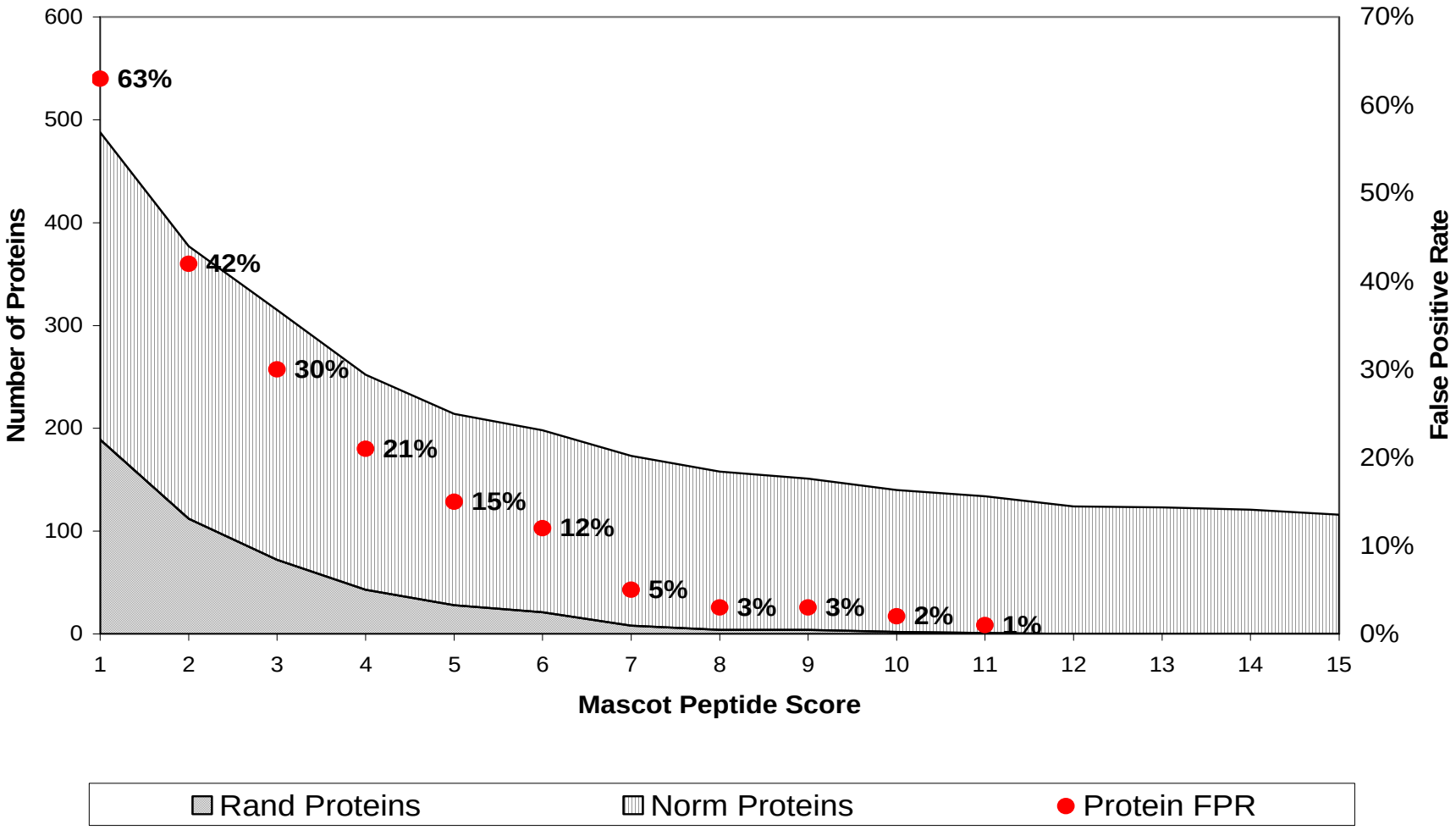
Ions Score: 34 **Matches (Bold Red):** 7/88 fragment ions using 33 most intense peaks


Automated Webservices Based Workflow for Processing and Storage of MS/MS data
Processes
Biological Sample Analysis by MS/MS
Convert Binary Data to mzXML Format
Convert mzXML to Peak List
Peak List Post Processing
Peak List Filter and Charge State Correction
Database Search Mascot
Statistical Analysis PROVALT
In
Out
Data Storage
Raw Binary Data
mzXML Data
Peak Lists
Processed Peak Lists
Database Search Results
Final Results
Annotation of Proteome Results
Biological Information Dissemination

Protein Validation Tools (Pro-ValT)



Comparison of Protein Identification With 6 or More Peptides



Pro-ValT results

Minimum Number of Peptides	Mascot Pep. Score to achieve 1% FPR
1	42
2	30
3	22
4	18
5	14
6	11

Conclusions

- Scoring algorithms predict peptide, **not protein**, probability
- Probability does not provide insight into False Discovery Rate
- Pro-ValT is an attempt to determine protein False Discovery Rates
- A different vantage point is needed to evaluate proteomic data

Acknowledgements

- James Atwood, Lin Lin, Lei Cheng, Art Nuccio, Fernanda Ludolf, Peggi Angel
- Daniel B. Weatherly, Todd Minning, Rick Tarleton
- NIH