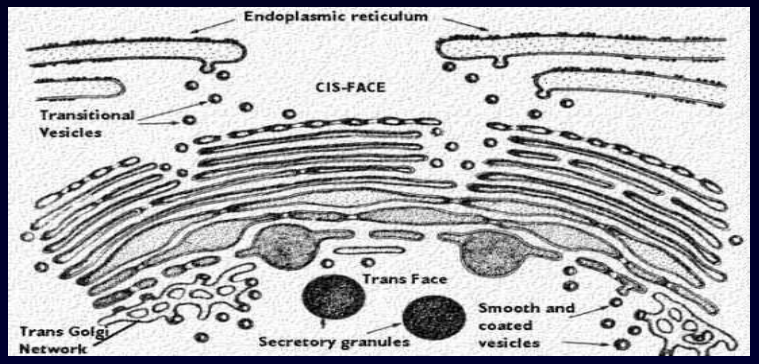
Shotgun Proteomic Analysis

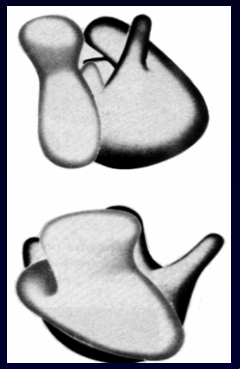
Department of Cell Biology
The Scripps Research Institute

Biological/Functional Resolution of Experiments

Multiprotein Complex

Organelle





Cells/Tissues

Function

Information Content

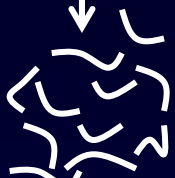
Expression Analysis

Interaction Analysis

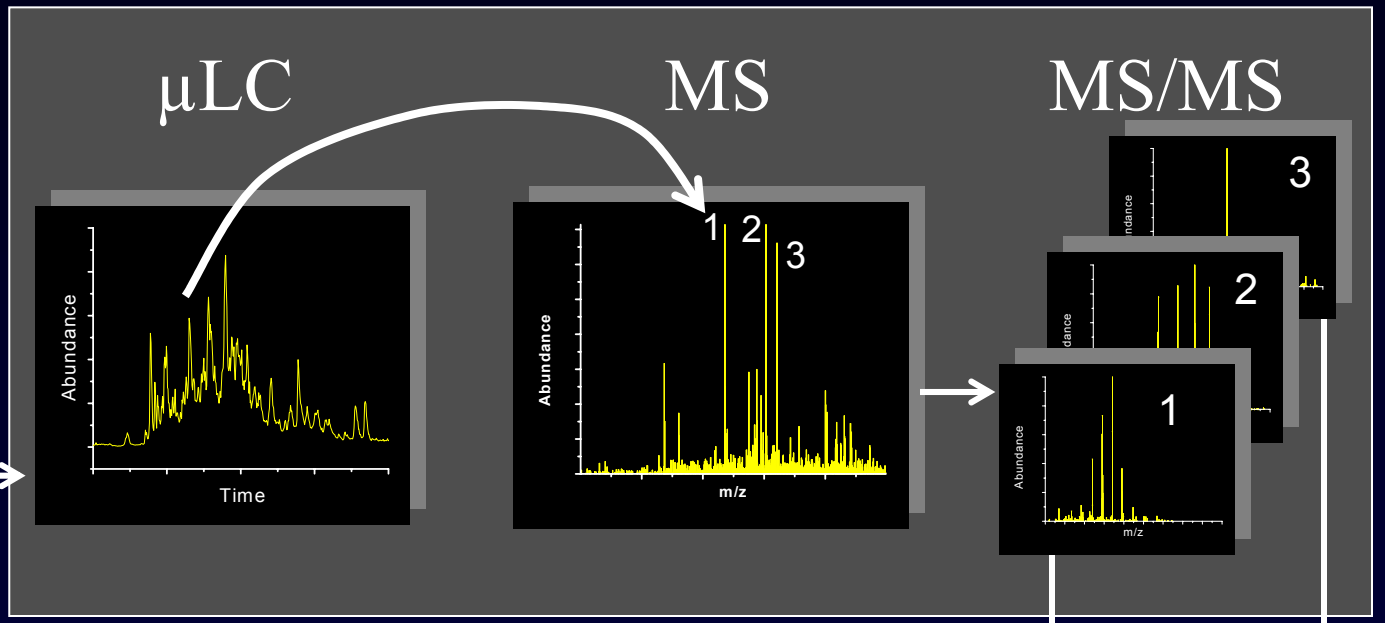
“Shotgun Proteomics”

Protein Mixture

proteolysis



Peptide Mixture

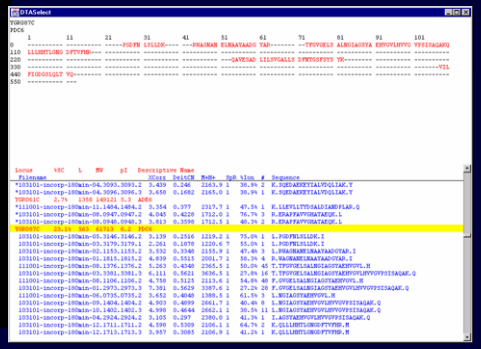


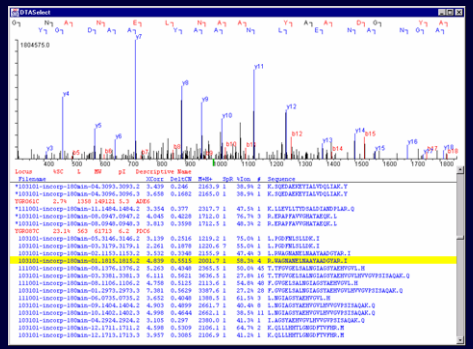
80 node Beowulf Computer Cluster

SEQUEST

Output Filtering and Re-Assembly

DTASelect





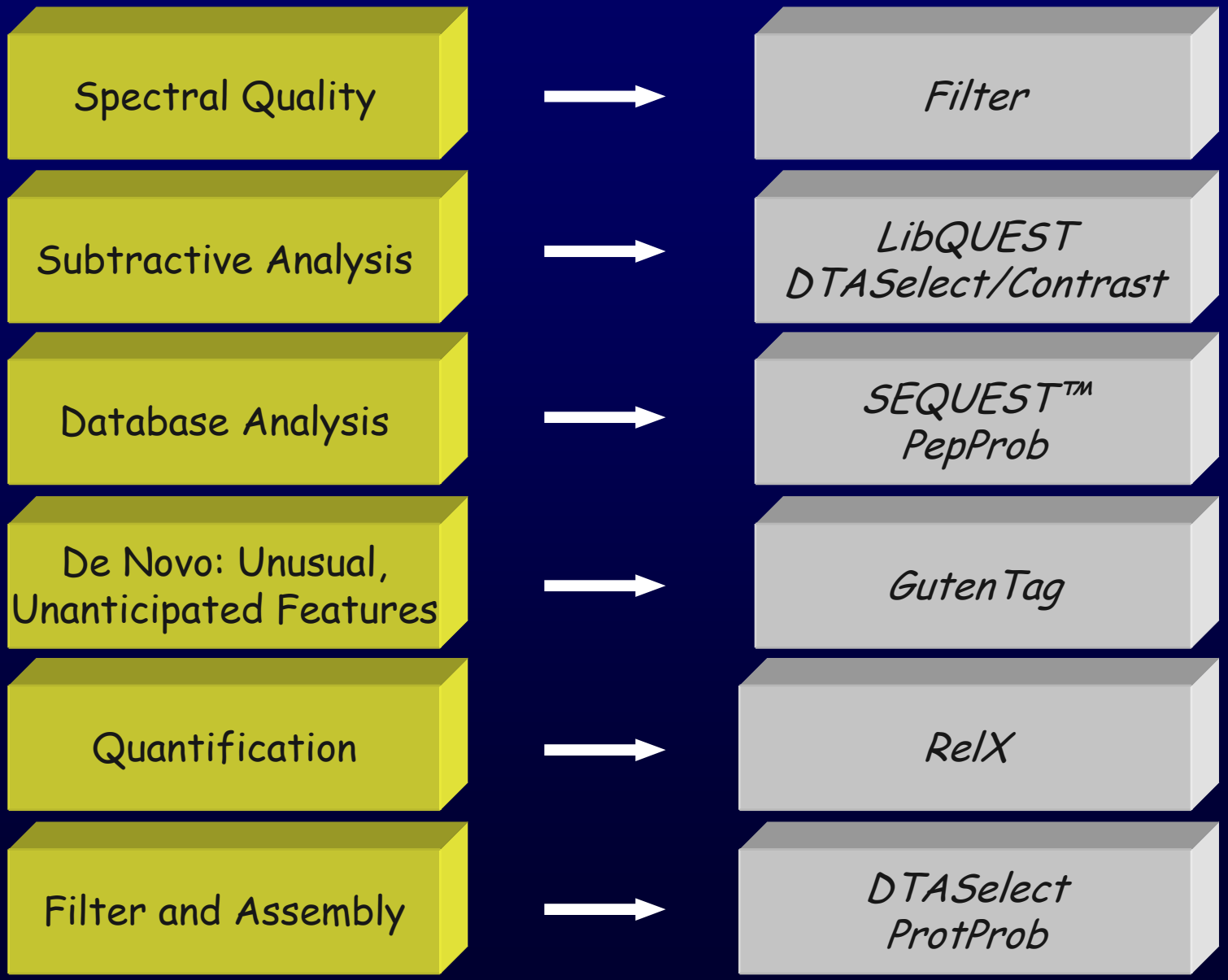
Data Processing Issues with Shotgun Proteomics

1:1 Mixture of Unlabeled/^{15}N-Labeled Yeast
Soluble Proteins Analyzed Using a Single 12h Analysis

	LCQ	LTQ
MS/MS Spectra	18,970	86,950 (4.5 x)
Protein ID's (*1 peptide confirmed w/ RelEx)	559	891 (1.6x)
Protein ID's (*2 peptides confirmed w/ RelEx)	157	304 (1.9x)

*RelEx was used to evaluate the presence of labeled isotopomer

Processing Tandem Mass Spectra



Data Issues

- Data quality
- How is a match determined?
  - Protease issues?
  - Validation issues?
- Posttranslational Modifications?
- Quantification?
- Sampling Issues?

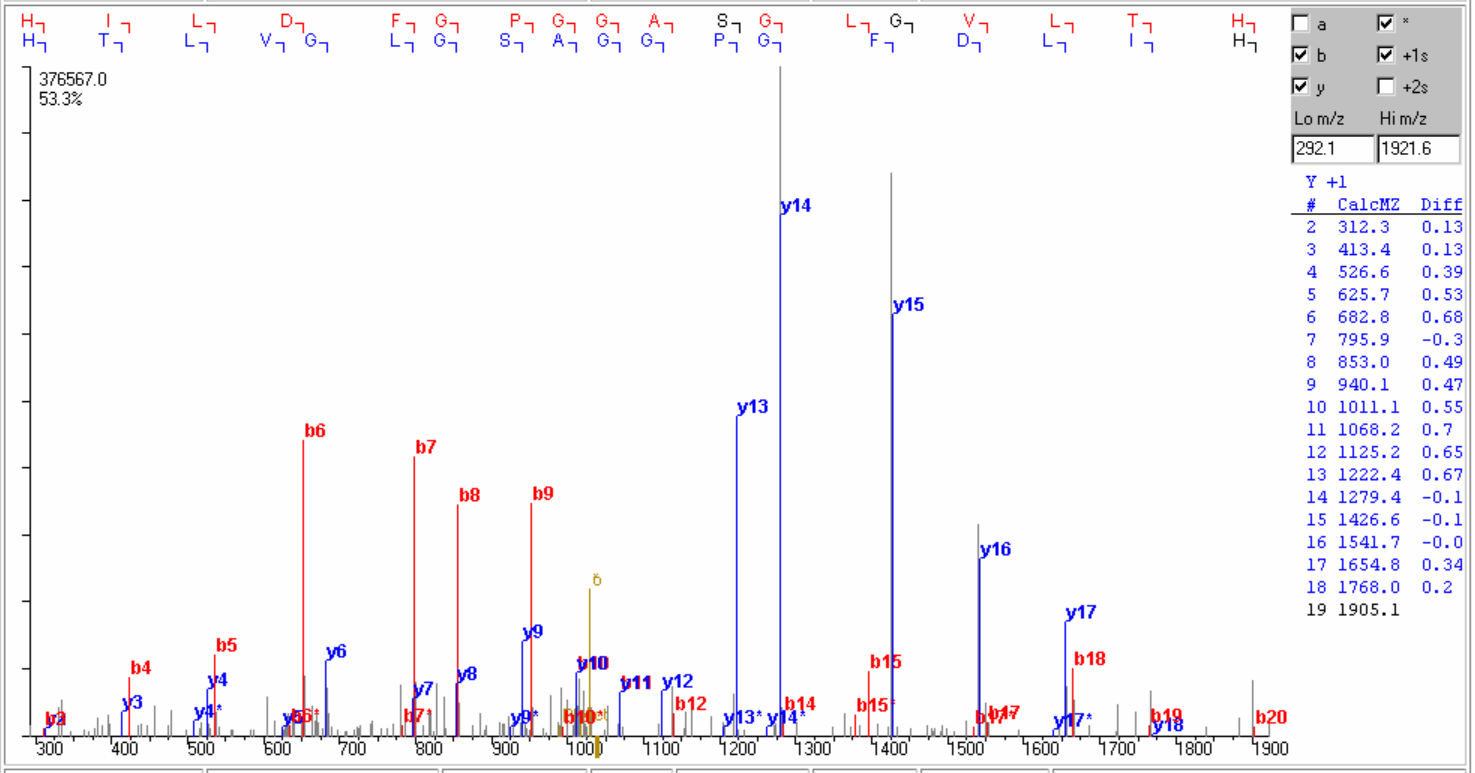
Spectral Filtering with Hand Crafted Features

	Called Good	Called Bad	%Correct
+1 GOOD	671	75	89.9%
+1 BAD	5585	11475	67.3%
+2/+3 GOOD	3166	348	90.1%
+2/+3 BAD	11611	26684	69.7%
All GOOD	3837	423	90.1%
All BAD	17196	38159	68.9%

Bern, Goldberg, MacDonald, Yates *Bioinformatics* (in press)

File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss AIM
Address http://137.131.4.127/cgi-bin/show?Dr=/wfs/bfd/5/rovshan/test_prob/YEAST_DBS&Da=112700proteosomeTCA60ug06&Sc=1766&Sq=ATHILDFGPGGASG Go Links Customize Links Free Hotmail Windows

Mass: 2076.64	Datfile: 112700proteosomeTCA60ug06	Scan number: 1766	Charge: 2	Database:



Xcalib/Probability	Probability	deltCn	XCorr	Confidence	Ions	Reference	Peptide
1/1	19.24	0.00	5.434	96.45	30/40	YKL182W	K.ATHILDFGPGGASGLGVLTHR.N
2/2	0.77	0.40	3.248	89.91	17/36	YGL115W	V.VDDVGRLVGVLTLSDILKY.I
3/12	4.89	0.55	2.455	86.05	15/38	YMR216C	S.HVSIQSDSGPQSDSDLDSDS.S
4/128	3.53	0.60	2.190	80.65	12/34	YDR038C	Y.IPVINDKVFLHKPIGAEW.G
5/456	2.54	0.61	2.142	73.17	11/38	YNL049C	W.VGGDAVPELLIDVFNTDTIS.Q

unevaluated	maybe	YES	NO	Peptide evaluation
◉	○	○	○	Submit Peptide

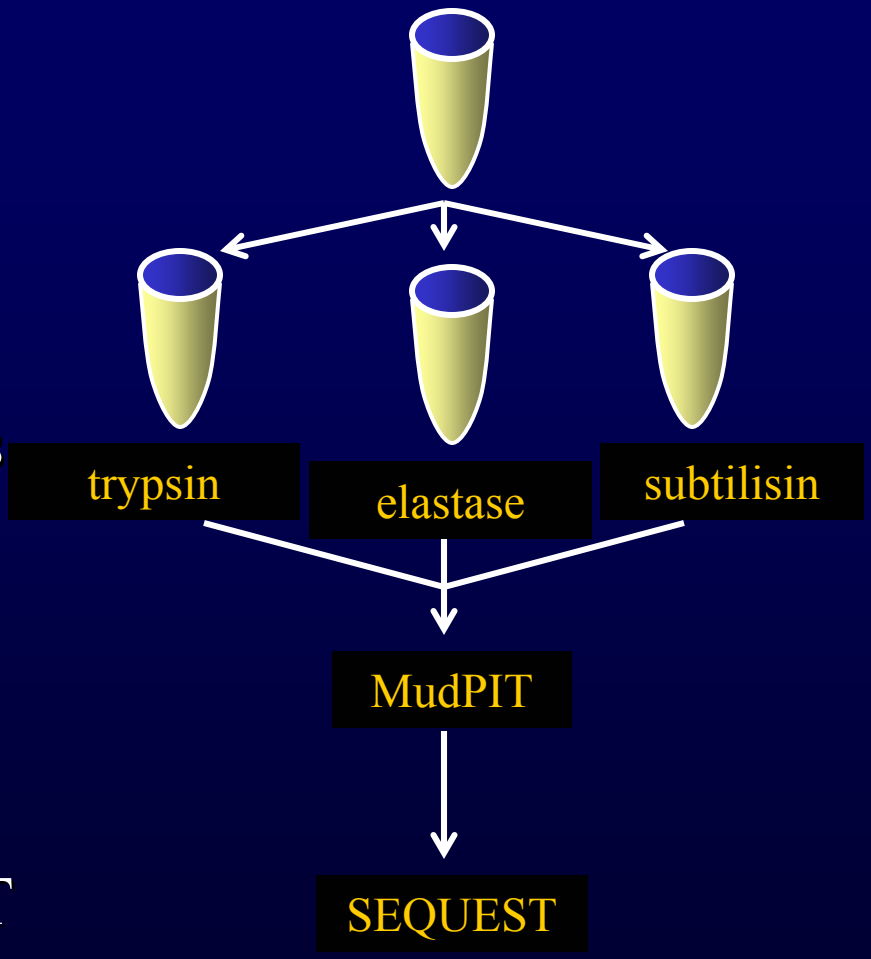
unevaluated	maybe	YES	NO	Locus evaluation
◉	○	○	○	Submit Locus

Refresh DTASelect.html

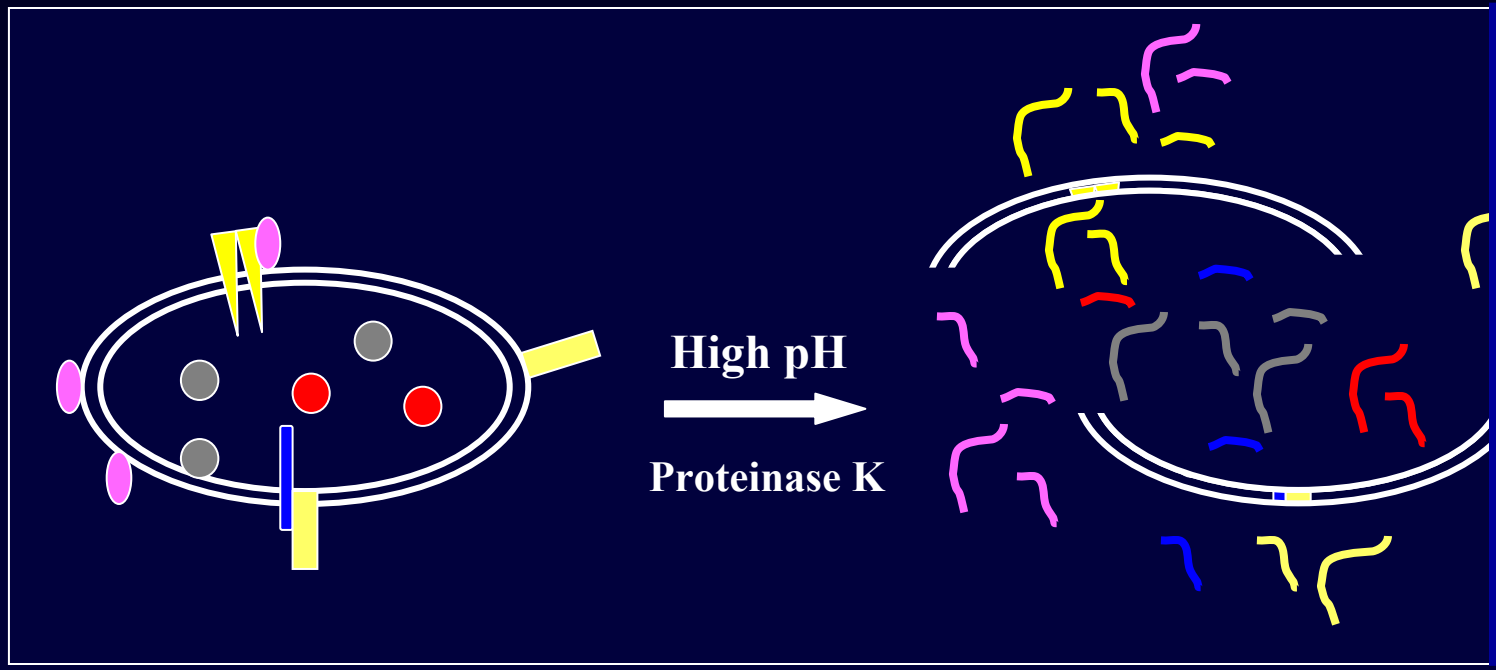
Applet started
Internet

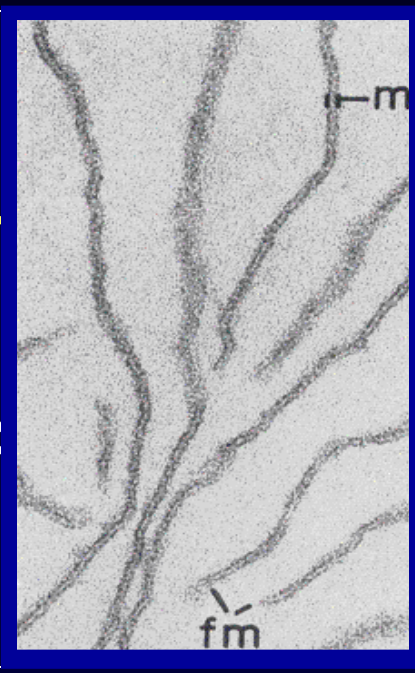
Multi-Enzyme Digestion Procedures

- Identification of PTMs
- Sample is split into 3 aliquots
- Digest using 3 different proteases
- Mix and analyze by LC/LC-MS/MS
- Interpret spectra using SEQUEST



High pH/Proteinase K Method (hpPK Method)





Wu et al., Nat. Biotech. 21:532-538 (2003)

Howell and Palade, J. Cell Biol. 92:822-832 (1982)

Overlapping Peptide Coverage

(TM7) gi|13794265|ref|NP_056312.1| DKFZP564G2022 protein

MAAAAWLQVL PVILLLLGAH PSPLSFFSAG PATVAAADRS
KWHIPIPSGK NYFSFGKILF RNTTIFLKFD GEPCDLSLNI
TWYLKSADCY NEIYNFKAEE VELYLEKLKE KRGLSGNIQT
SSKLFQNCSE LFKTQTFSGD FMHRLPLLGE KQEAKENGTN
LTFIGDKTAM HEPLQTWQDA PYIFIVHIGI SSSKESSKEN
SLSNLFTMTV EVKGPYEYLT LEDYPLMIFF MVMCIVYVLF
GVLWLAWSAC YWRDLLRIQF WIGAVIFLGM LEKAVFYAEF
QNIRYKGESV QGALILAELL SAVKRSLART LVSIVSLGYG
IVKPRLGVTL HKVVVAGALY LLFSGMEGVL RVTGAQTDLA
SLAFIPLAFL DTALCWWIFI SLTQTMKLLK LRRNIVKLSL
YRHFTNTLIL AVAASIVFII WTTMKFRIVT CQSDWRELWV
DDAIWRLLFS MILFVIMVLW RPSANNQRFA FSPLSEEEEE
DEQKEPMLKE SFEGMKMRST KQEPNGNSKV NKAQEDDLKW
VEENVPSSVT DVALPALLDS DEERMITHFE RSKME

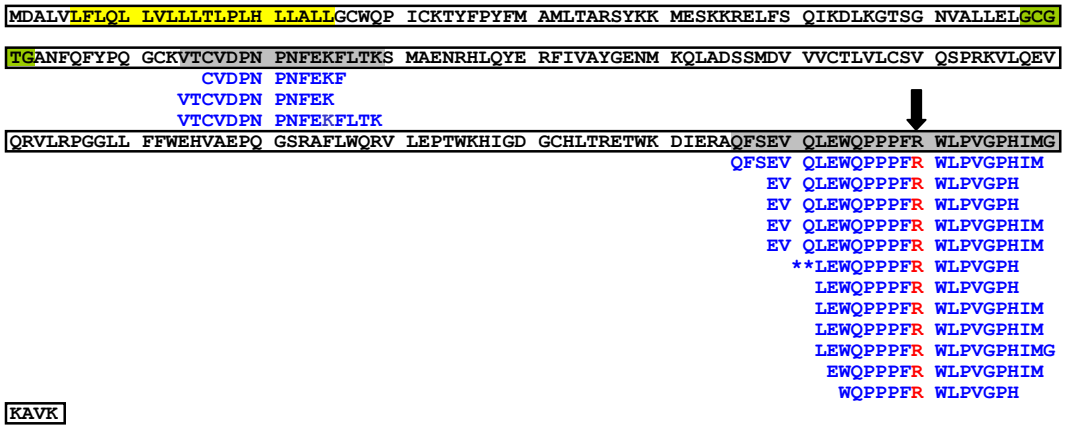
WVEENVPSSVTDVALPALLDS*DEER
VEENVPSSVTDVALPALLDS*DEER
EENVPSSVTDVALPALLDS*DEER
ENVPSSVTDVALPALLDS*DEER
VPSSVTDVALPALLDS*DEER
PSSVTDVALPALLDS*DEER
LPALLDS*DEER
PALLDS*DEER

(TM6) gi|14249524|ref|NP_116213.1| hypothetical protein FLJ14681

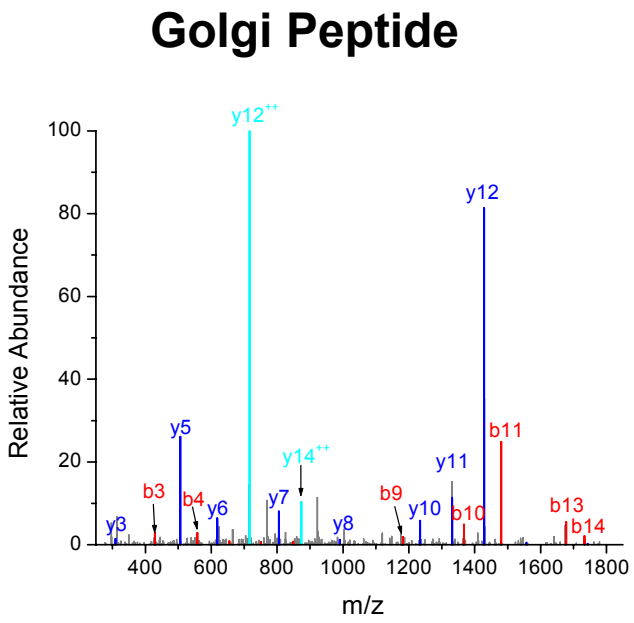
MVAACRSVAG LLPRRRRCFP ARAPLLRVAL CLLCWTPAAV
RAVPELGLWL ETVNDKSGPL IFRKTMFNST DIKLSVKSFH
CSGPVKFTIV WHLKYHTCHN EHSNLEELFQ KHKLSVDEDF
CHYLKNDNCW TTKNENLDCN SDSQVFPSLN NKELINIRNV
SNQERSMDVV ARTQKDGFHI FIVSIKTENT DASWNLNVSL
SMIGPHGYIS ASDWPLMIFY MVMCIVYILY GILWLTWSAC
YWKDILRIQF WIAAVIFLGM LEKAVFYSEY QNISNTGLST
QGLLIFAELI SAIKRTLARL LVIIVSLGYG IVKPRLGTVM
HRVIGLGLLY LIFAAVEGVM RVIGGSNHLA VVLDDIILAV
IDSIFVWFIF ISLAQTMKTL RLRKNTVKFS LYRHFKNTLI
FAVLASIVFM GWTTKTFRIA KCQSDWMERW VDDAFWSFLF
SLILIVIMFL WRPSANNQRY AFMPLIDDSD DEIEEFMVTS
ENLTEGIKLR ASKSVSNGTA KPATSENFDE DLK**WVEENIP**
SSFTDVALPV LVDSDEEIMT RSEMAEKMFS SEKIM

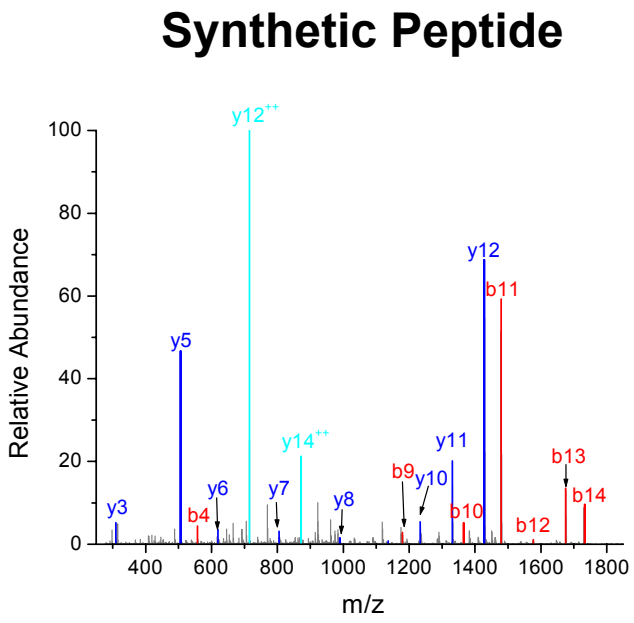
WVEENIPSSFTDVALPVLVDS*DEEIMTR
IPSSFTDVALPVLVDS*DEEIMTR
PSSFTDVALPVLVDS*DEEIMTR
SFTDVALPVLVDS*DEEIMTR
TDVALPVLVDS*DEEIMTR
TDVALPVLVDS*DEEIMTRS
DVALPVLVDS*DEEIMTR
VALPVLVDS*DEEIMTR
ALPVLVDS*DEEIMTR
ALPVLVDS*DEEIMTRS
LPVLVDS*DEEIMTR
PVLVDS*DEEIMTR
PVLVDS*DEEIMTRS

gi|27229118|ref|NP_082129| RIKEN cDNA 0610006F02; S-adenosylmethionine-dependent methyltransferase activity [Mus musculus]



Dimethyl Arginine Containing Peptide





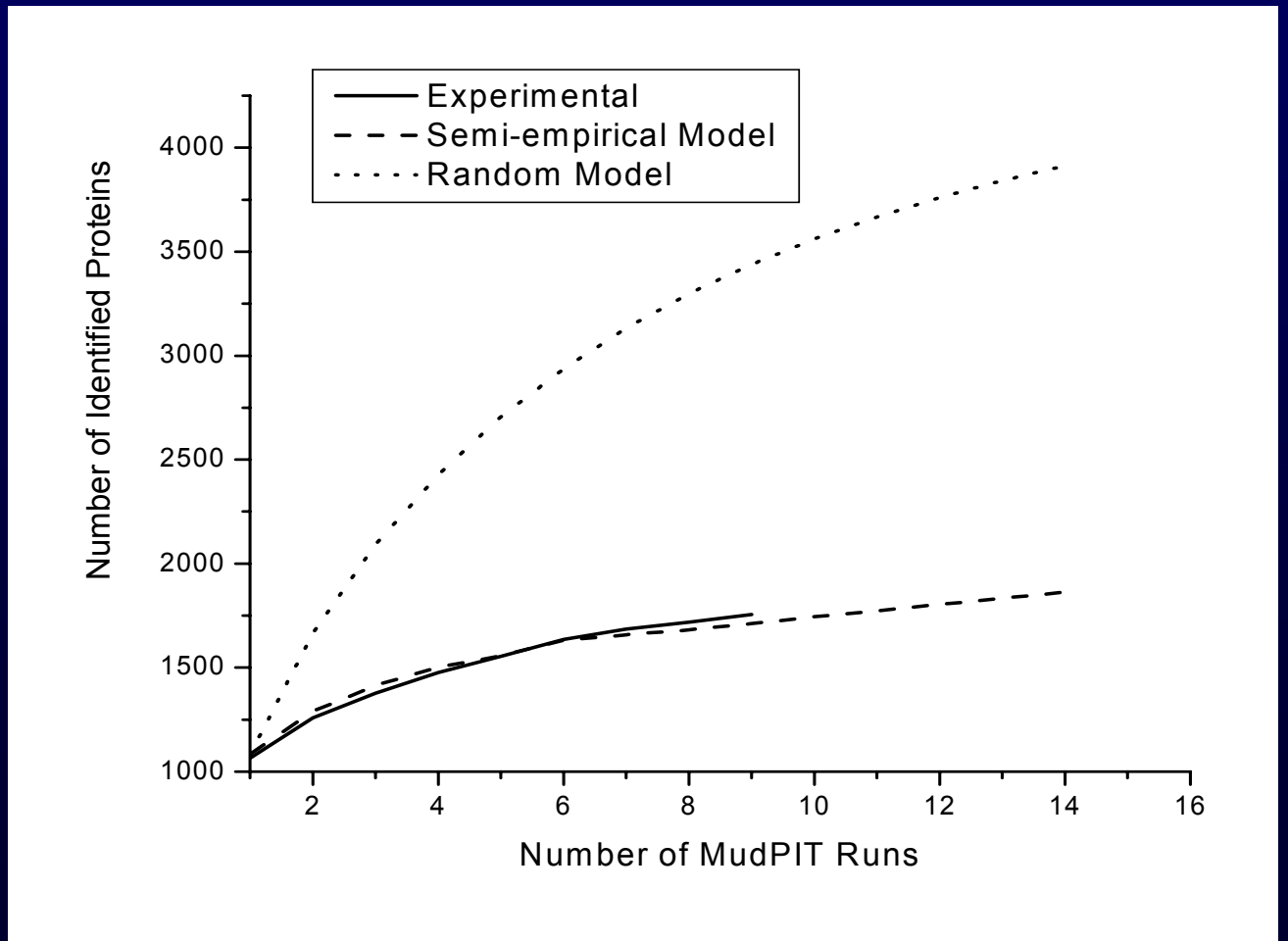
Shotgun Proteomic Experiments and Sampling Issues

- Based on prior studies in yeast, we know not every protein present is id'd
- Reproducibility is good for high abundance proteins 70-80%
- Reproducibility is not as good for low abundance proteins. 20-30%
- Is this predictable?

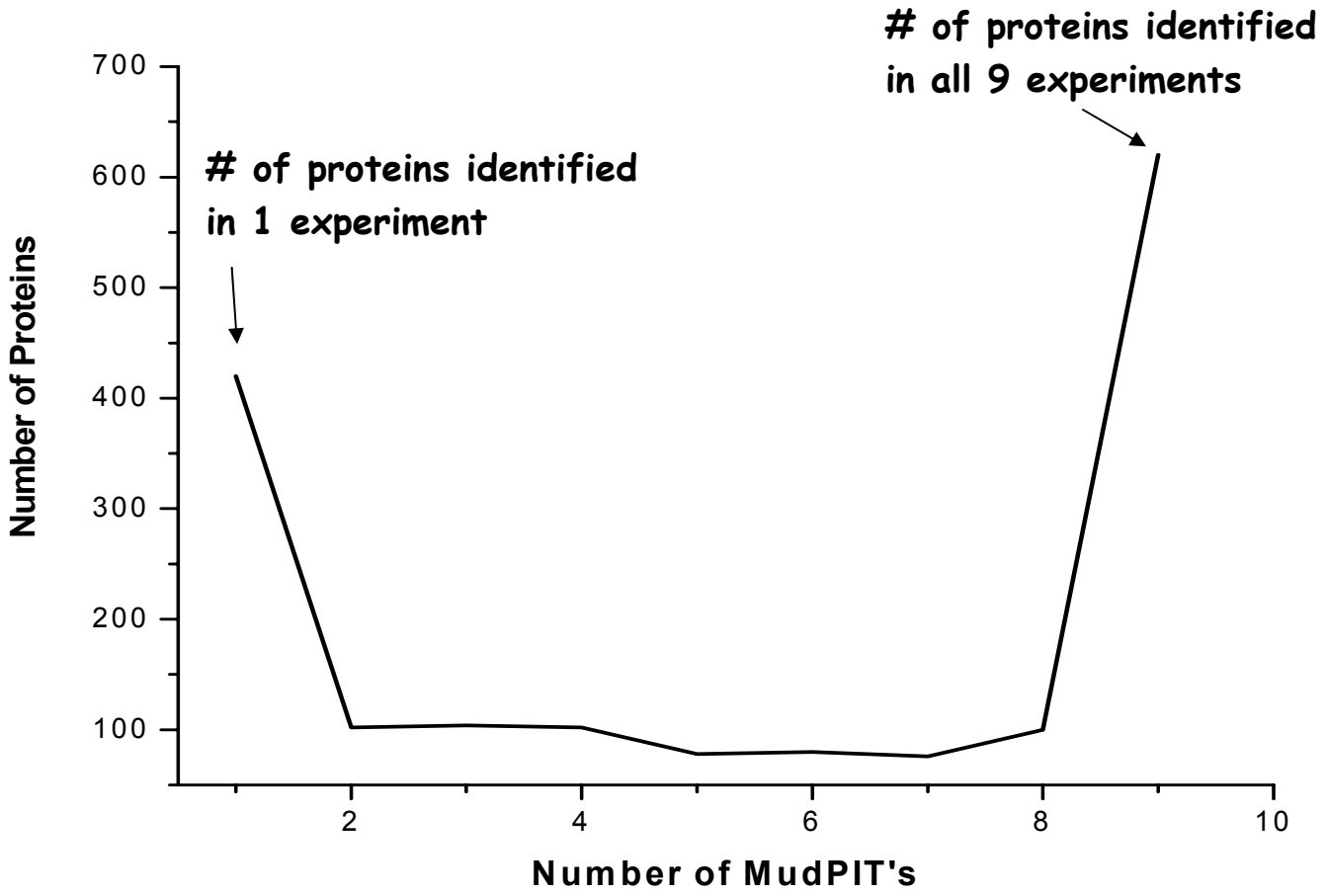
Random Sampling Model for Data Dependent Acquisition

$$K = n_L * (1 - (1 - L/N)^S)$$

- n_L = # of protein species at particular level
- L = abundance level
- N = total number of proteins
- S = experiments



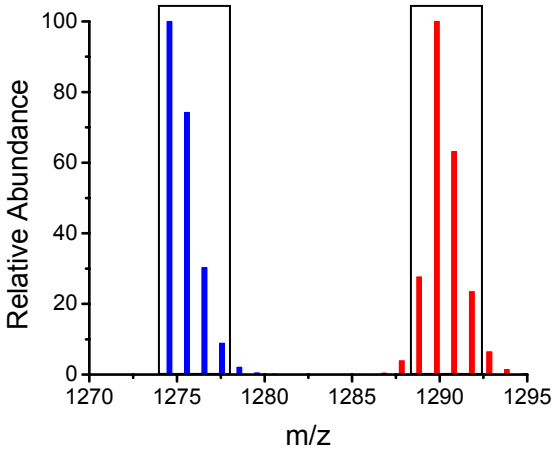
Distribution of Protein Identifications After Repeating Analysis 9 times



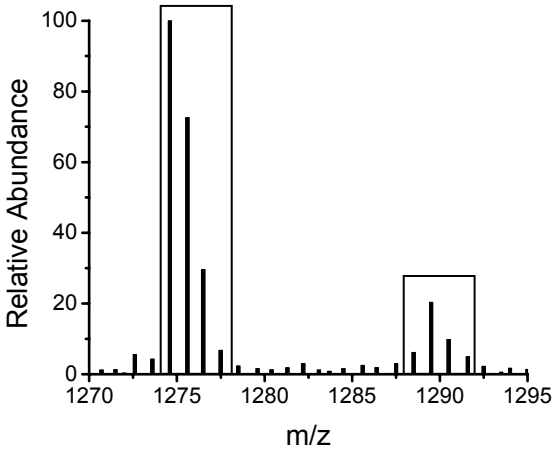
RelX software

DTASelect Output
Peptide Sequence
LVNHFIQEFK

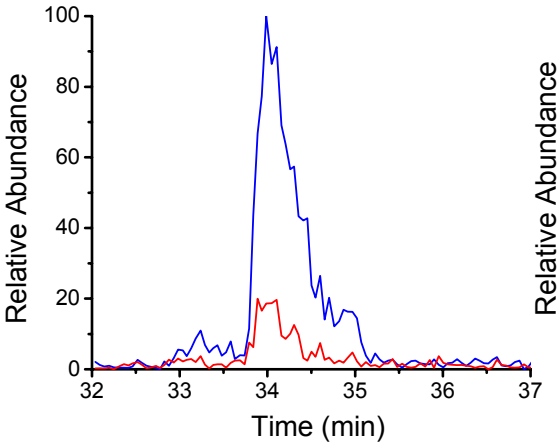
1) Predict m/z Range



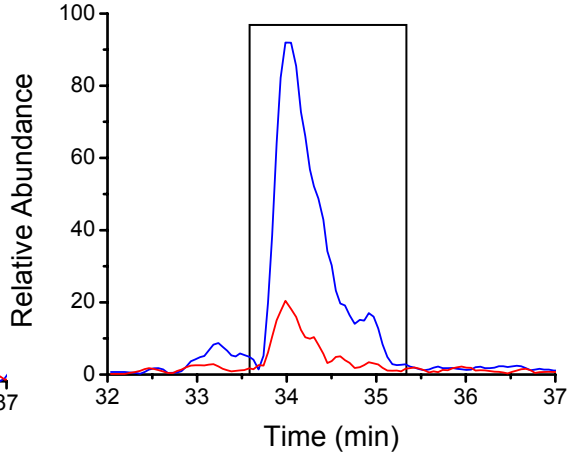
2) Sum Signal in Range



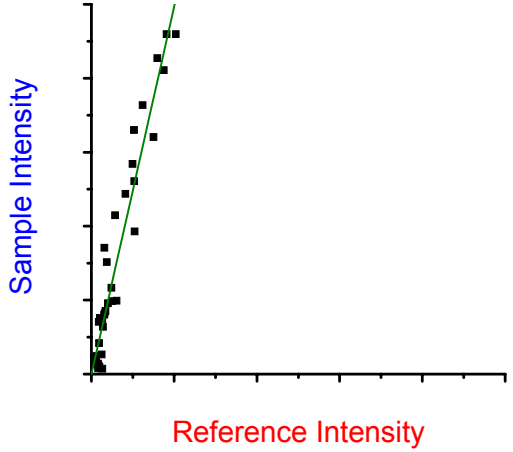
3) Store Mass Chromatograms



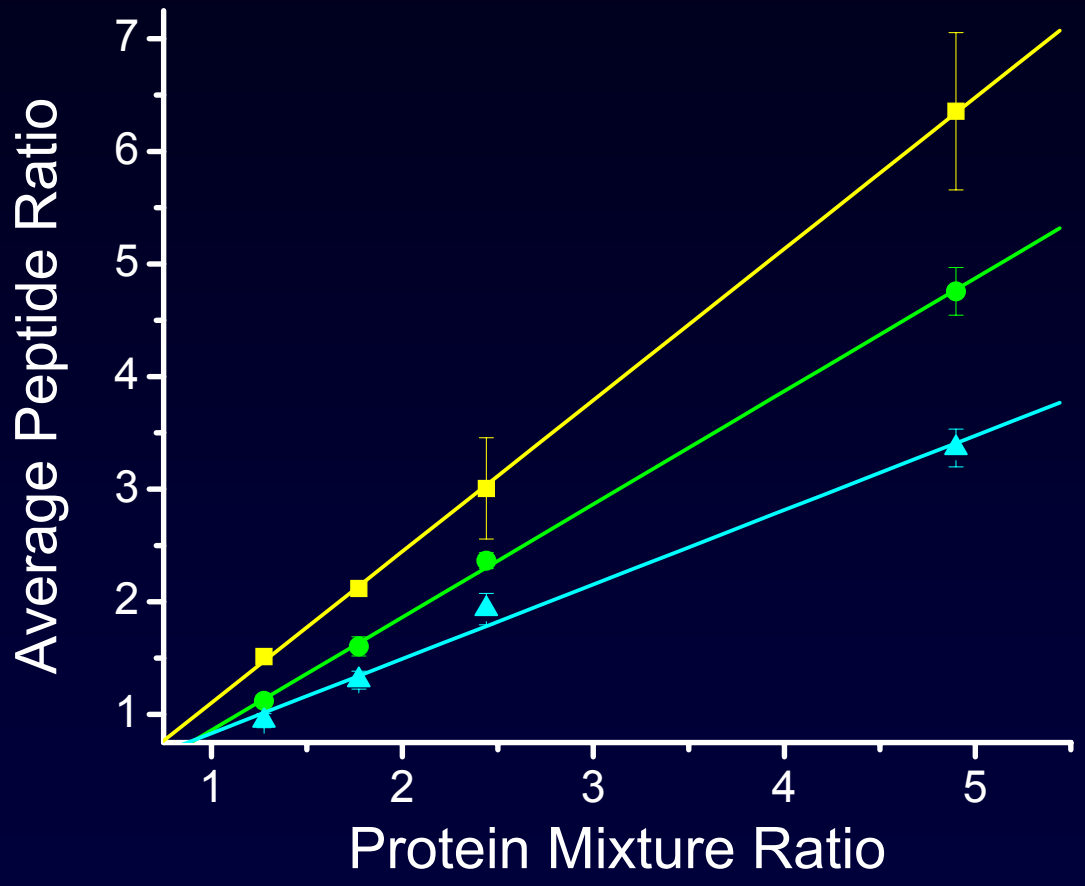
4) Peak Detection



5) Correlation



Systematic Errors are Present in Samples



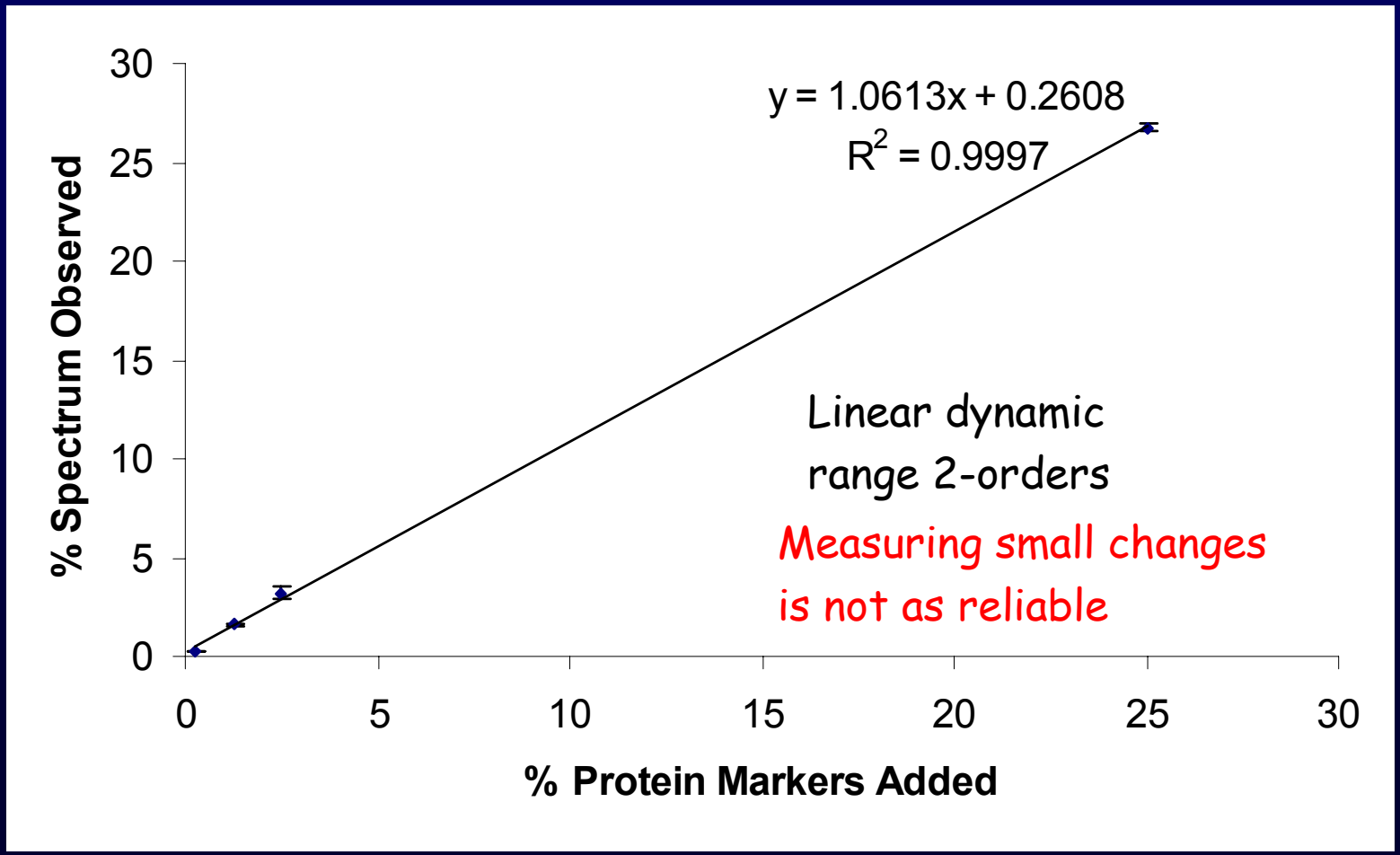
TSA1 = 1.335±0.015

SSA1 = 1.004±0.018

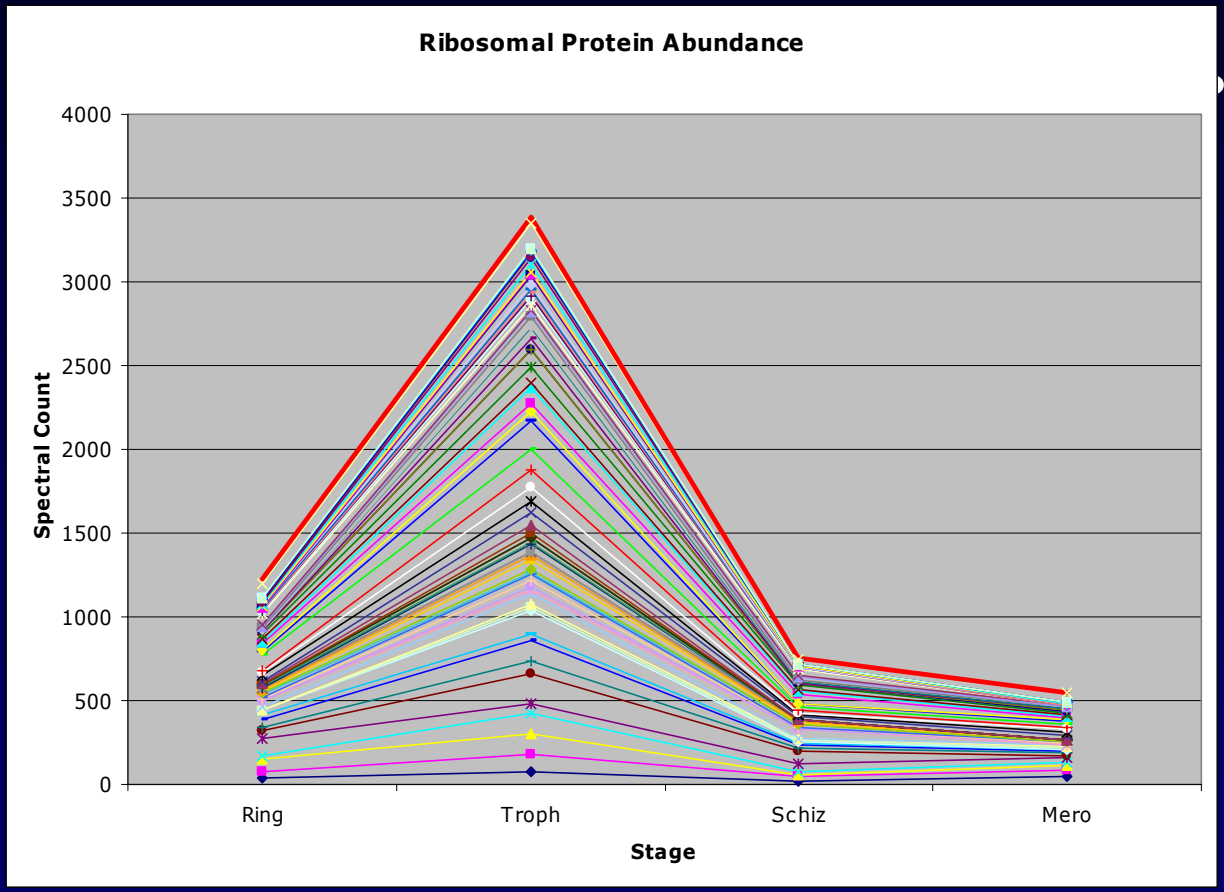
ADH1 = 0.661±0.045

Spectral Sampling for Relative Quantification

Combined Data for 6 proteins added to Yeast Soluble Cell Lysate at 4 different levels



Synthesis of Ribosomal Proteins in *Plasmodium falciparum*



- Striking trend: almost all ribosomal proteins increase over ring to troph transition

Standards

- 1. Data formats: McDonald et al *MS1, MS2, and SQT - three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications*, ***RCM*** 2004, 18, 2162-8.
- Database search standard based on:

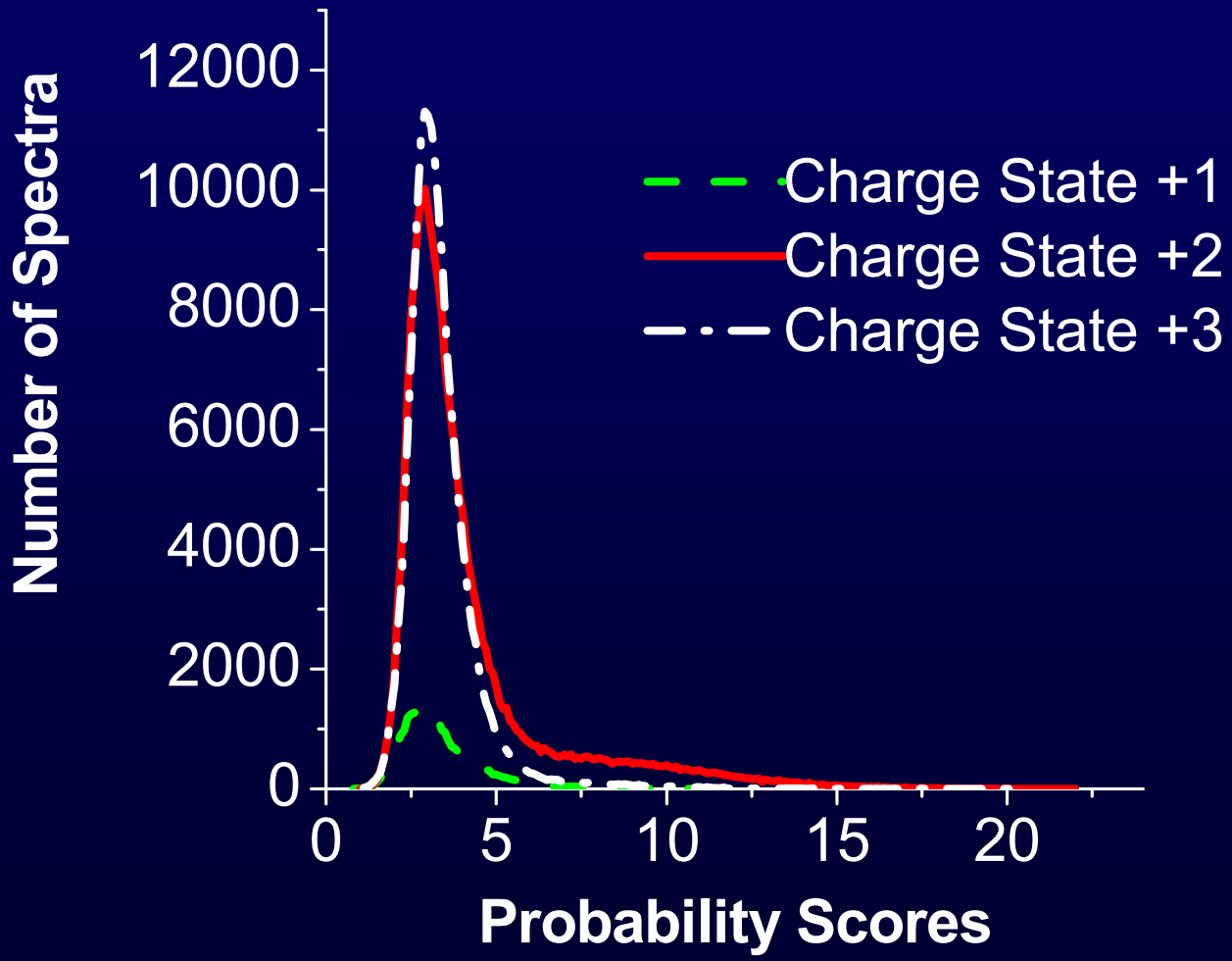
Washburn et al, Large Scale analysis of the yeast proteome via multidimensional protein identification technology *Nature Biotechnology* 19, 242-247 (2001)

MacCoss et al , Probability Based Validation of Protein Identifications Using a Modified SEQUEST Algorithm, *Analytical Chemistry* 74, 5593-5599 (2002). Normalized Scores

Standards

- 4. Standards should not prevent innovation
  - Data formats should be practical e.g. storage space
- 7. Data processing tools should be transparent and validated, e.g. published
- 8. Data for publication: information to support biological conclusions- sequences of peptides id'd.
- 9. Archiving data: biological conclusions should be the most important part of the experiment

Probability Distributions

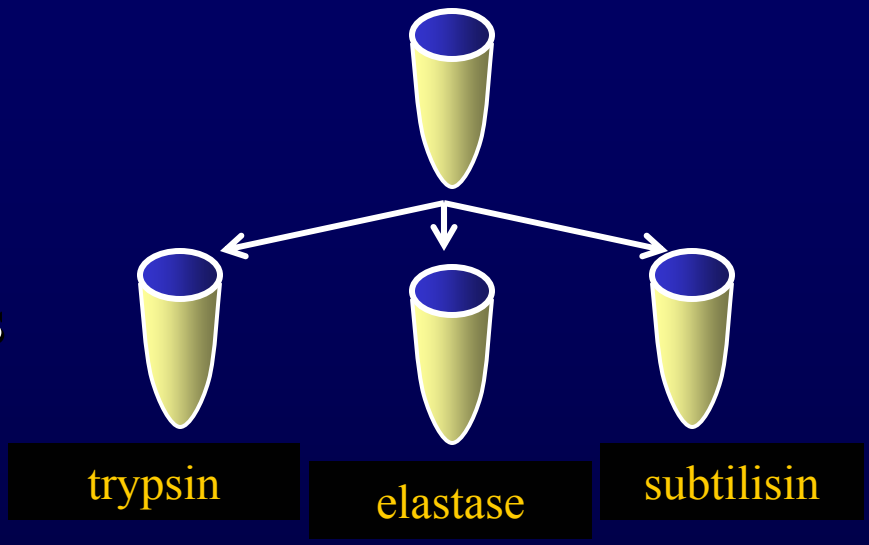


Single Spectral Matches are Problematic: How to tell if they are correct

- Searches determine closeness of fit based on some measure: Compare matches with different programs
- Probability scoring: P = random match based on frequency of fragment ions in database
- SEQUEST: XCorr measures how close the spectrum fits to ideal spectrum
- Manual validation, experimental validation
- *de novo* interpretation

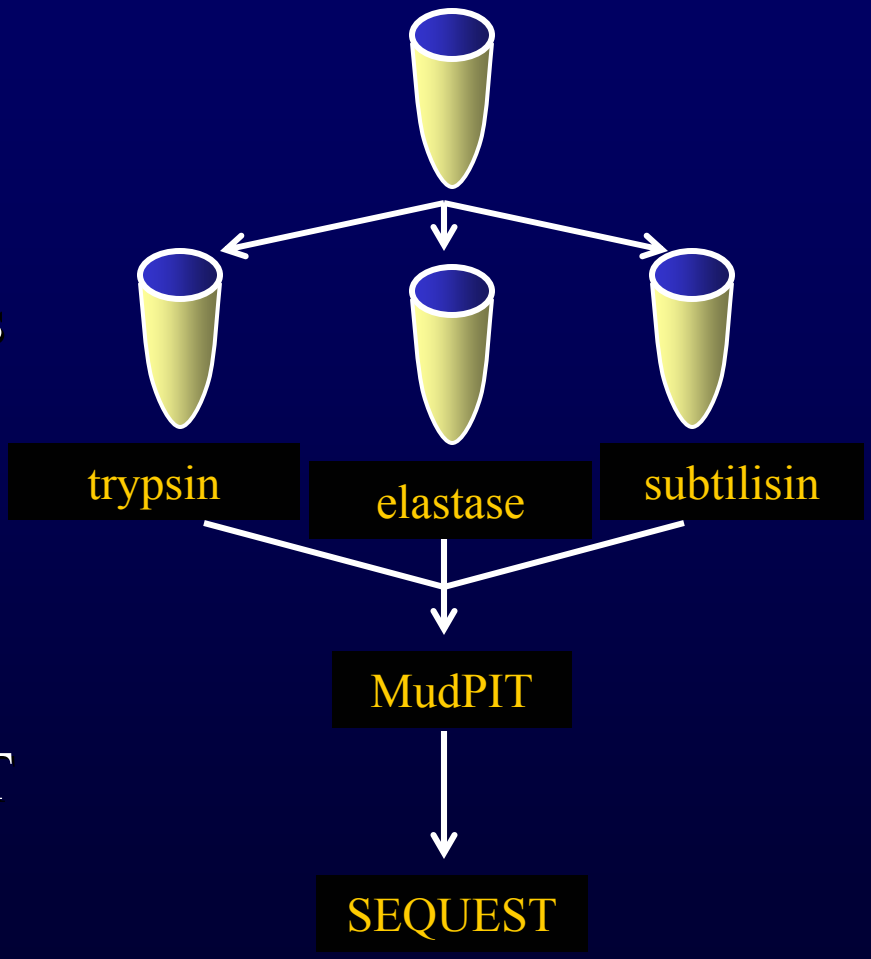
Multi-Enzyme Digest

- Sample is split into 3 aliquots
- Digest using 3 different proteases



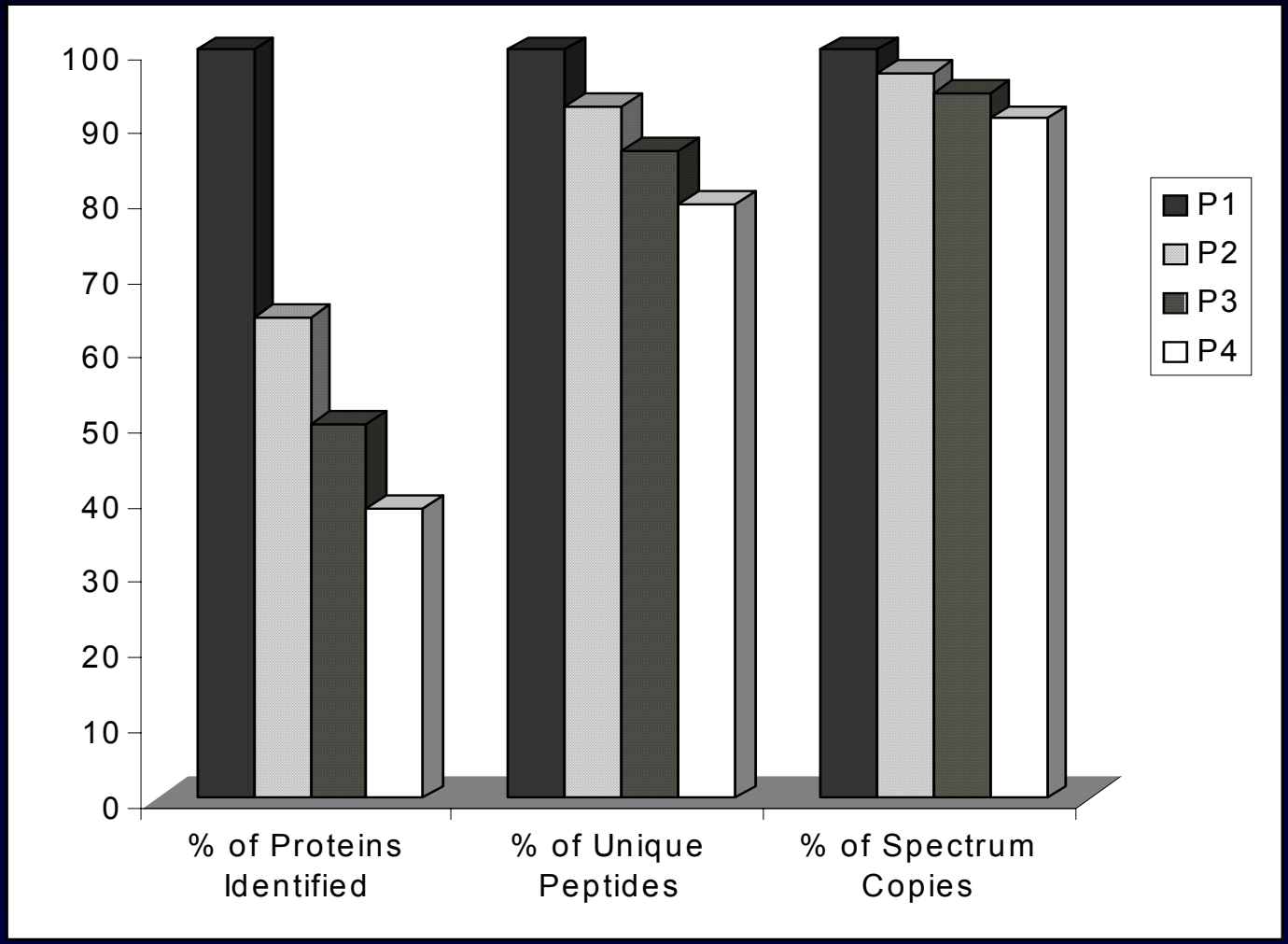
Multi-Enzyme Digest

- Sample is split into 3 aliquots
- Digest using 3 different proteases
- Mix and analyze by LC/LC-MS/MS
- Interpret spectra using SEQUEST



Properties of Data Dependent Data Acquisition

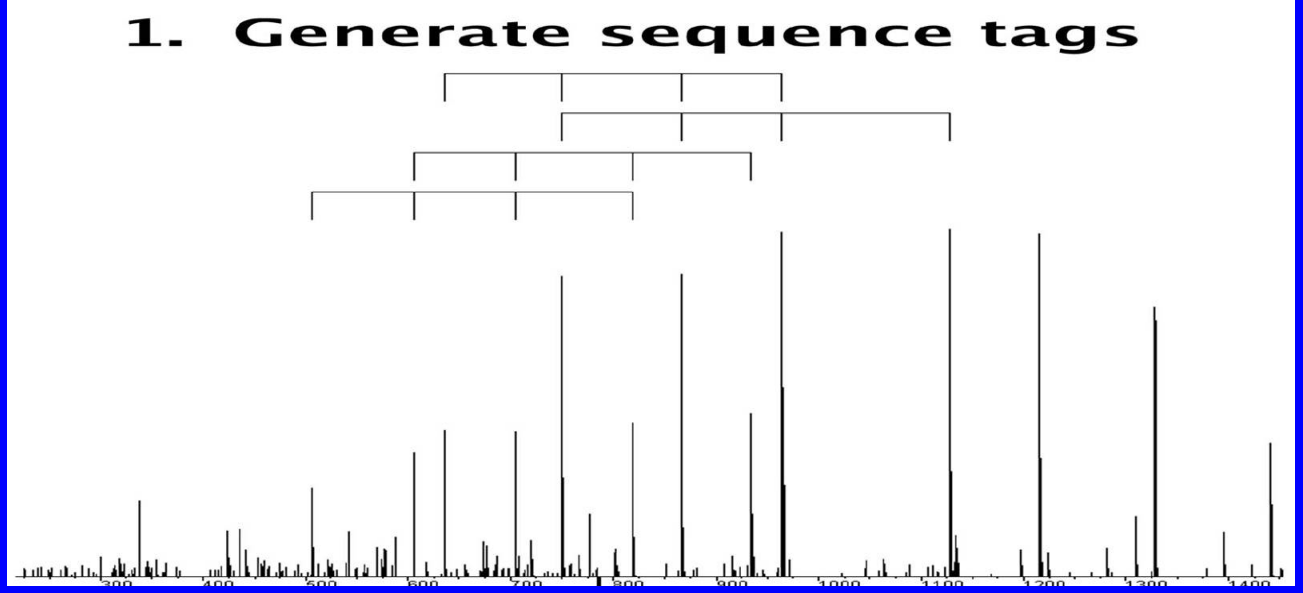
Most invariant property is spectral copy number



GutenTag: Partial de novo Sequencing of Tandem Mass Spectra

- Database searching assumes minimal errors in the database and sequence variations between strains, individuals and species
- Modifications need to be specified in database searches, so unanticipated modifications will be missed.
- Partial *De novo* analysis of tandem mass spectra in large-scale can identify peptides containing sequence variations and unanticipated modifications

1. Generate sequence tags



GutenTag

- 6170 tandem mass spectra: LC/LC/MS/MS analysis of simple digested protein mixture
- 1987 spectra matched by SEQUEST
- 1328 spectra matched by GutenTag
- 766 partial matches suggesting modifications and sequence variations
- Total matching spectra by *GutenTag*: 2,094
- Partial de novo will extend identifications
- Software is Automated and Large-Scale

Improved Spectral Quality Effects Peptide Identification

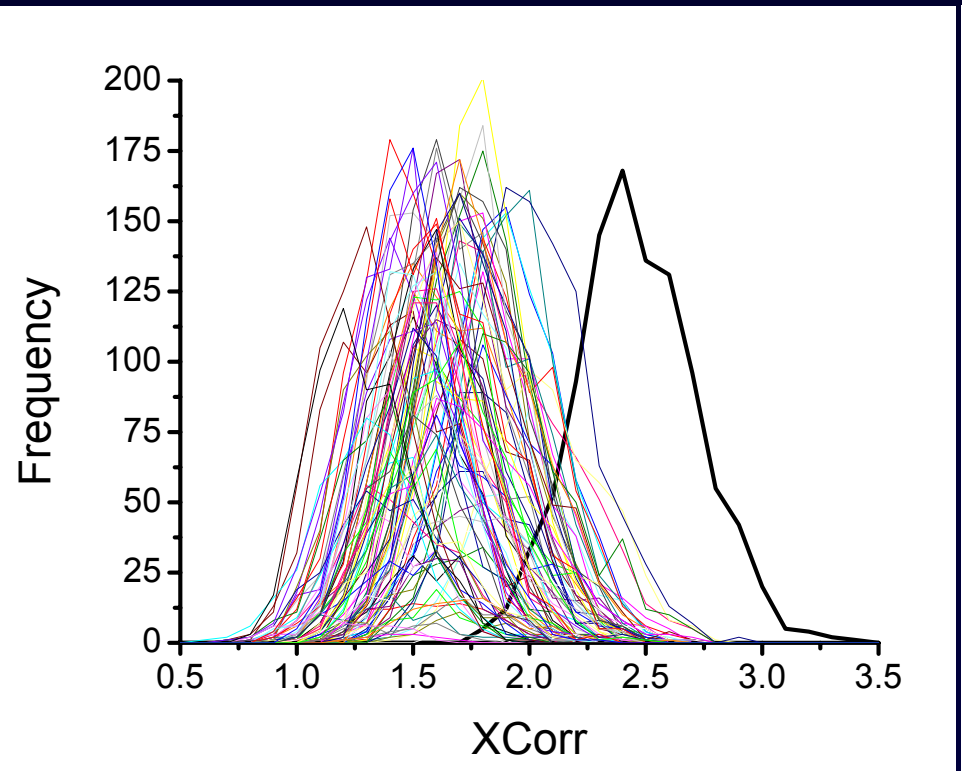
Infusion of 1 pmol/µl Angiotensin I

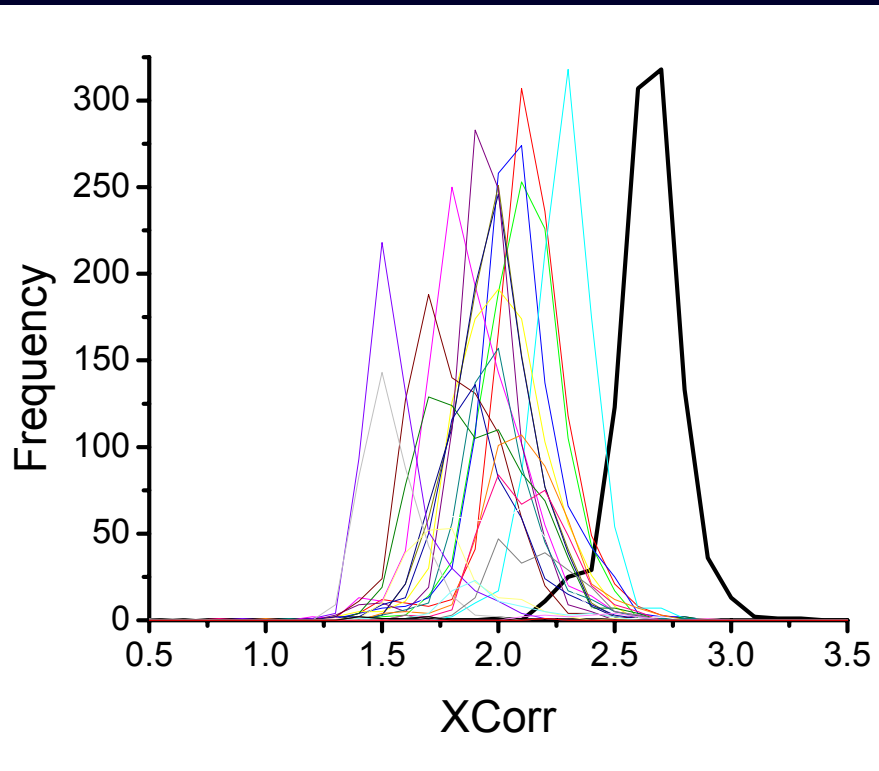
LCQ-Classic

761 of 1000 MS/MS Spectra Matched the Correct Sequence

LTQ

970 of 1000 MS/MS Spectra Matched the Correct Sequence





Database Searching with Tandem Mass Spectra

- The goal is to identify peptides using MS/MS spectra and amino acid sequence databases.
- Develop a probabilistic model that establishes a relationship between the database sequences and the spectrum to complement quantitative measures of closeness-of-fit
- Develop non-empirical probabilistic measures using cross-correlation measurements

Probability Model

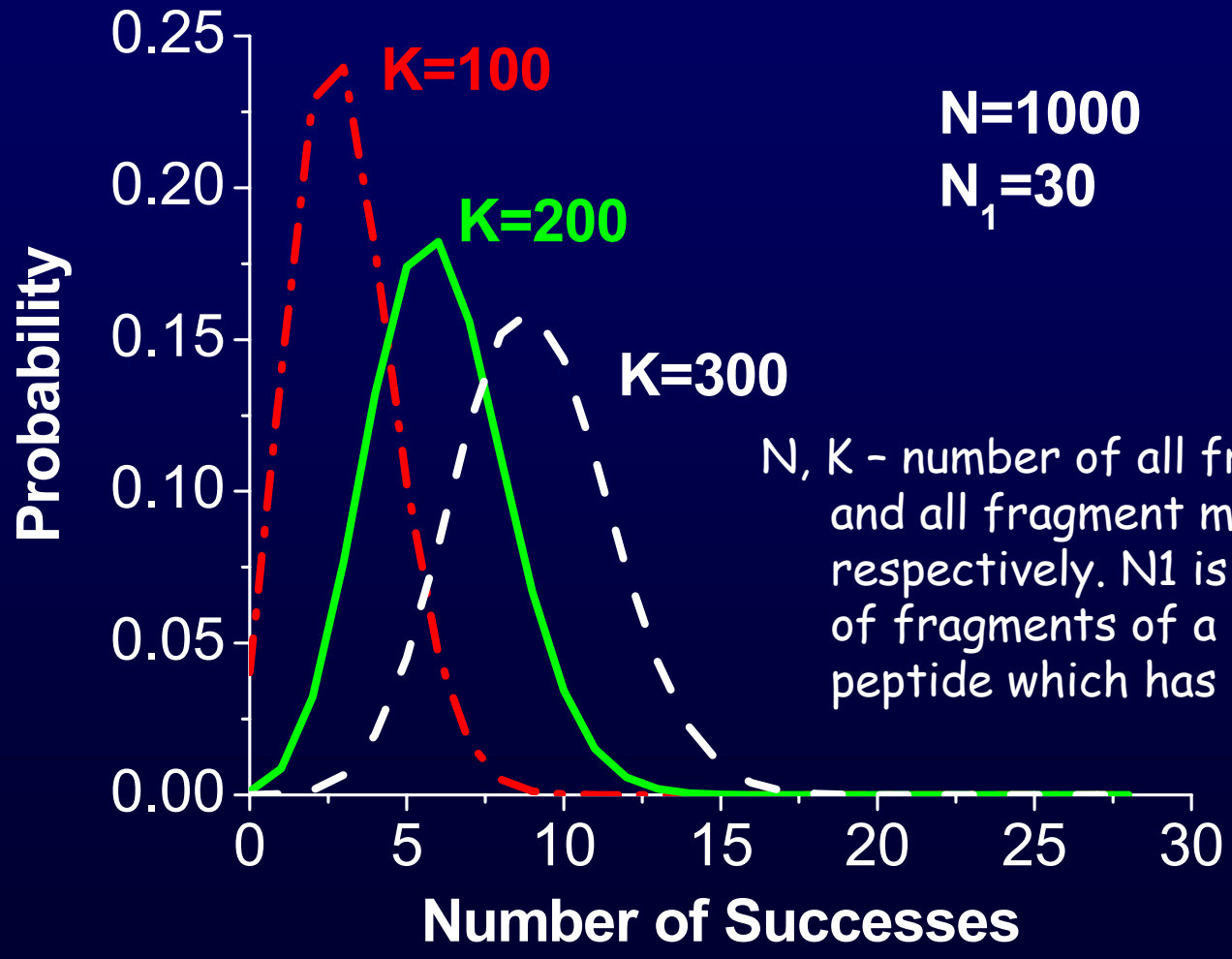
- Null hypothesis: All fragment matches to MS/MS spectrum are by random.

- N, K – number of all fragments and all fragment matches, respectively. N_1 is the number of fragments of a particular peptide which has K_1 matches.

$$P_{K,N}(K_1, N_1) = \frac{C_K^{K_1} * C_{N-K}^{N_1-K_1}}{C_N^{N_1}}$$

- We seek an amino acid sequence that has the smallest probability of being a random match.

Model Hypergeometric Distributions



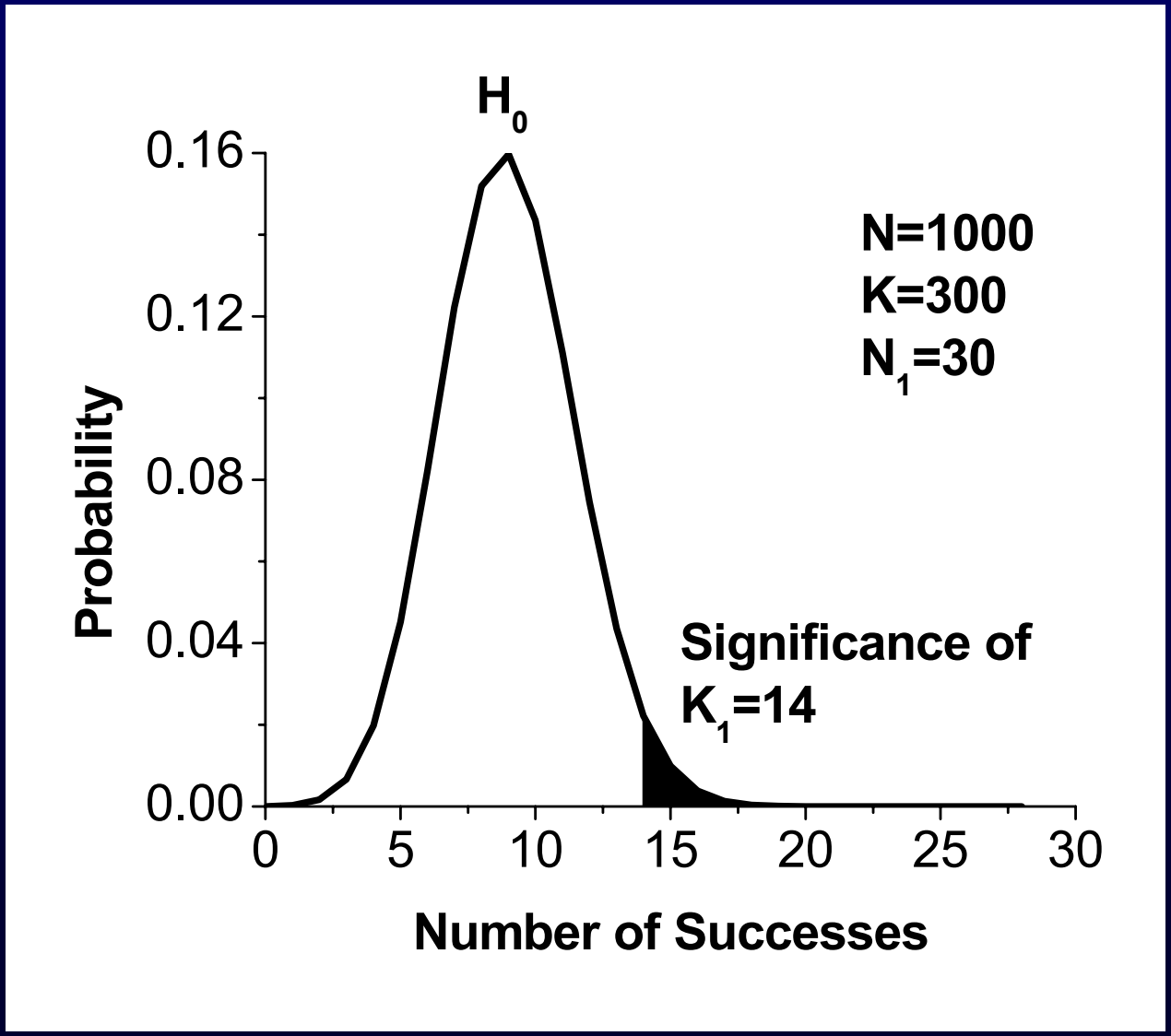
N, K – number of all fragments and all fragment matches, respectively. N1 is the number of fragments of a particular peptide which has K1 matches.

Significance of Peptide Identification

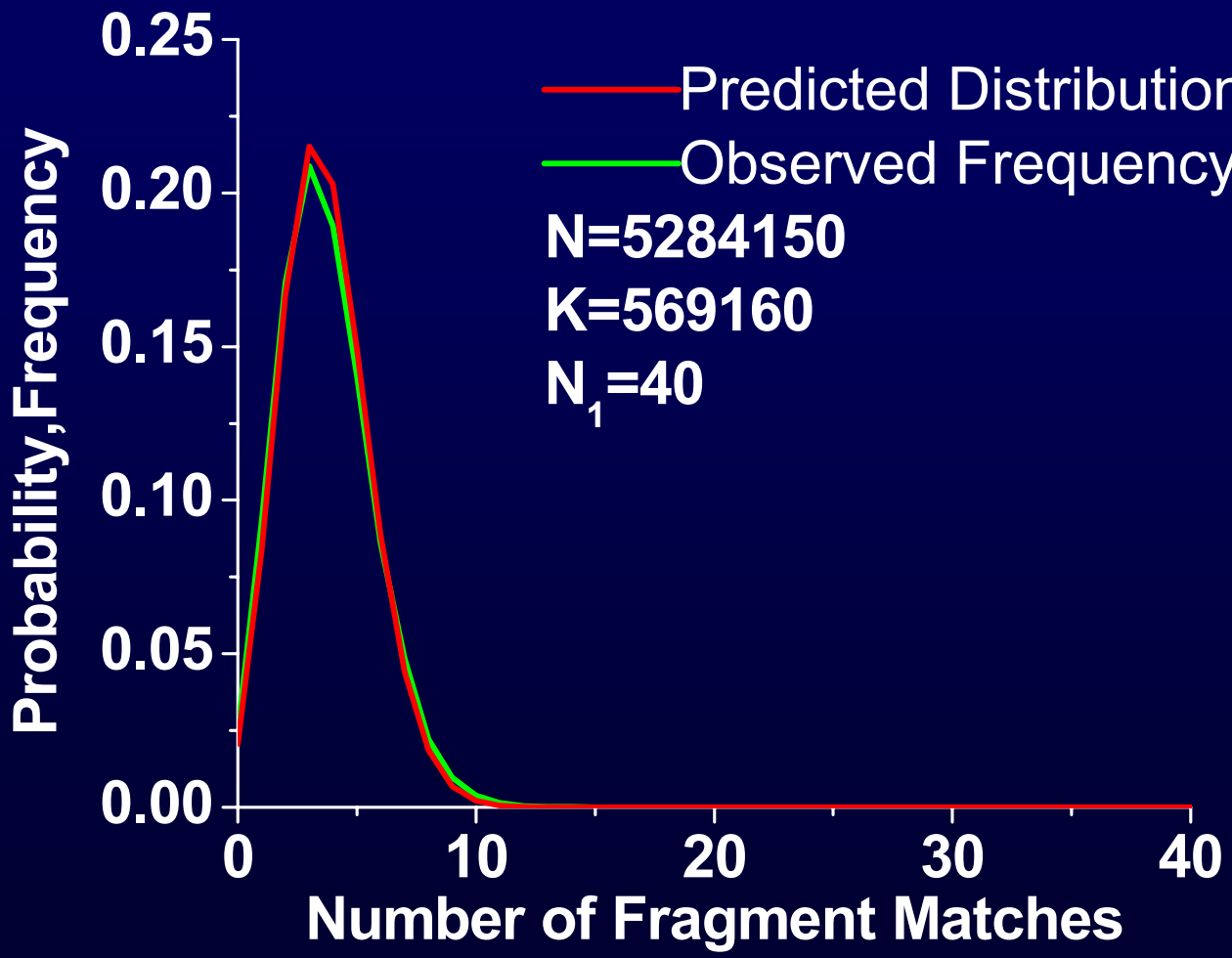
- Not all identifications are significant: poor quality spectra of peptides , incomplete peptide fragmentation, inaccuracies in database, posttranslational modifications, MS/MS of chemical noise and non-peptide molecules.

- Significance of a match (P_value) is also obtained from the hypergeometric distribution.

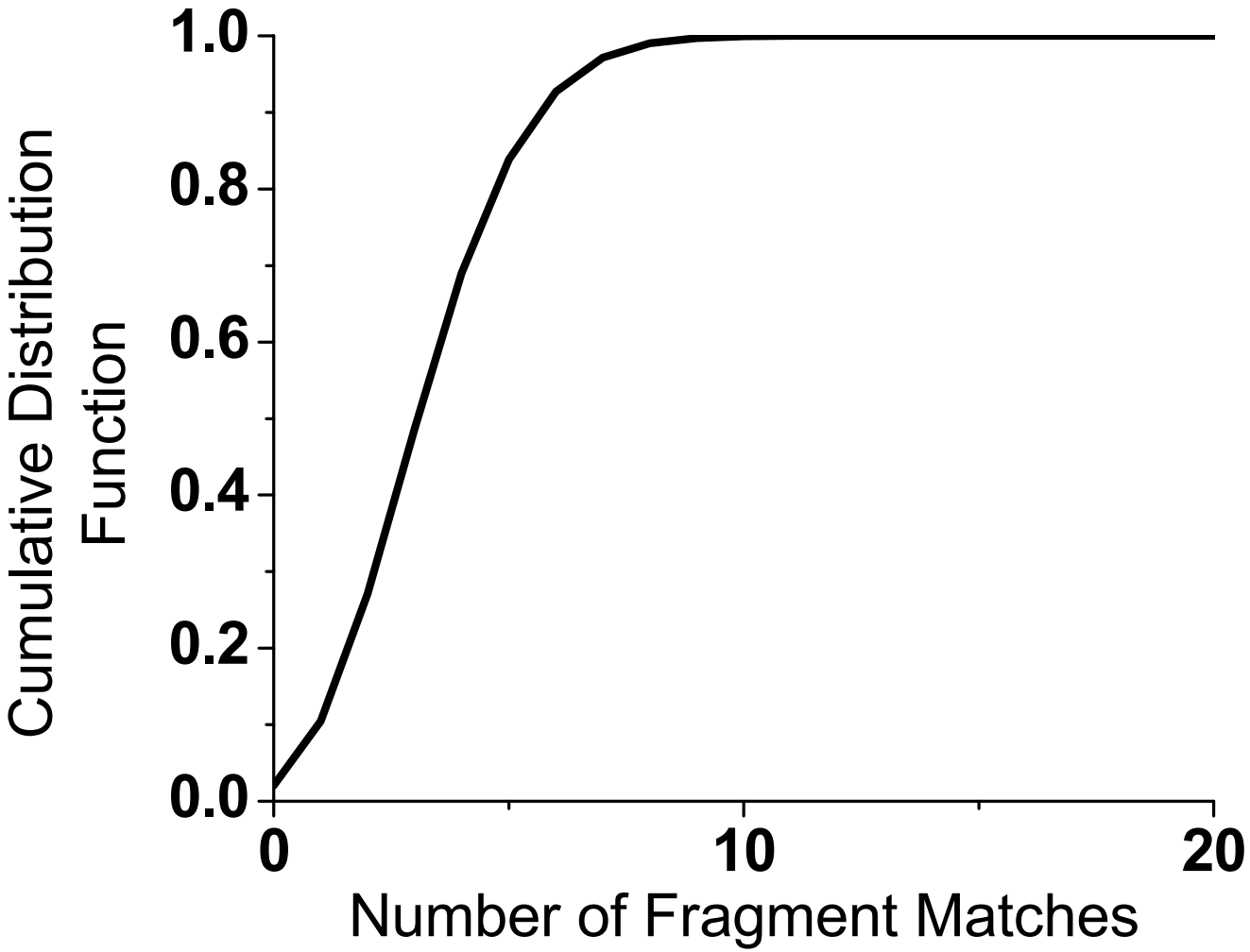
Significance of a Match



Yeast Database



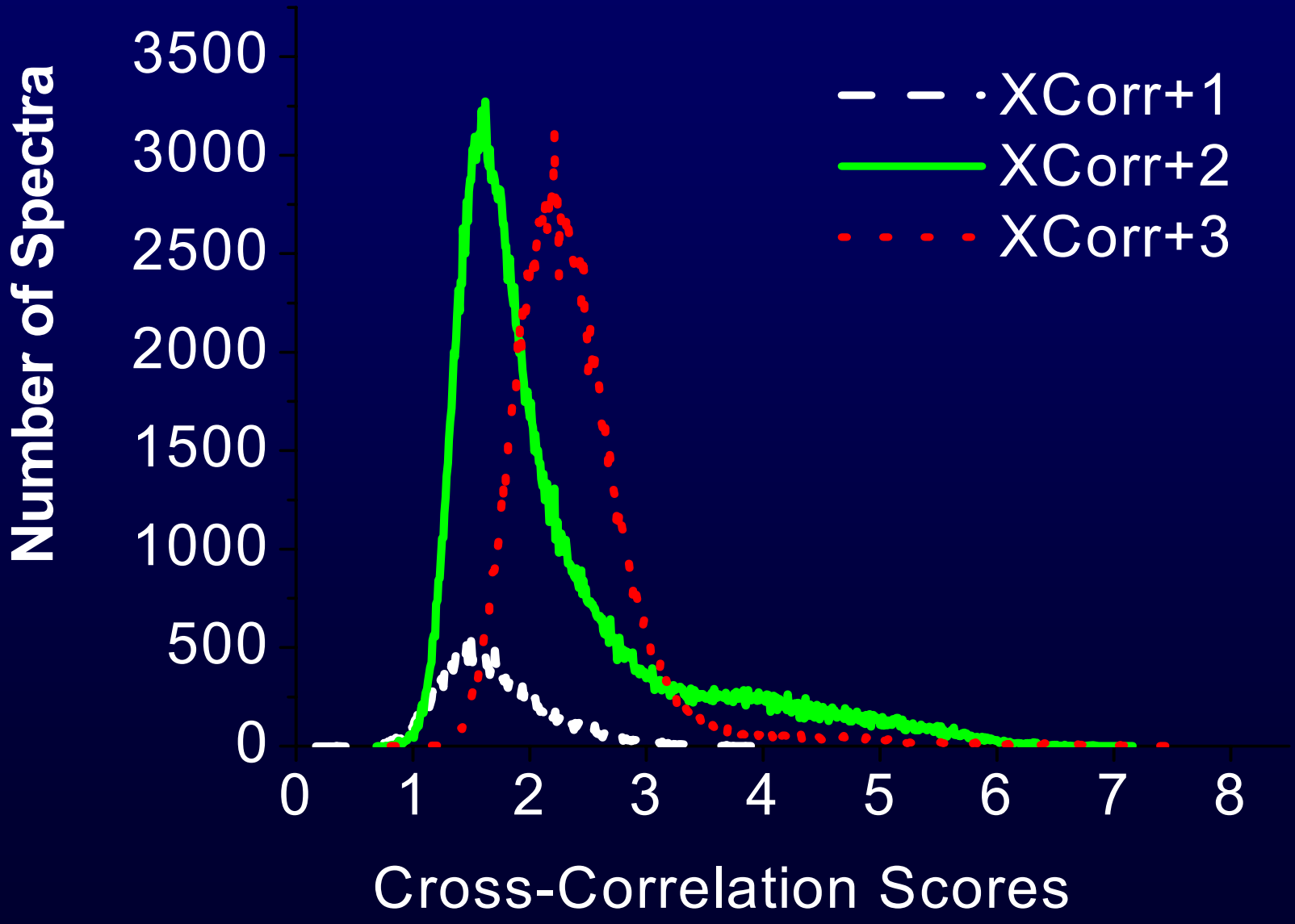
Cumulative Distribution



Mass/Charge State Dependence

- Scores that use closeness of fit measures can artificially inflate with weight/mass.
- This complicates use of uniform criteria for identification.
- Probabilities generated by hypergeometric distribution are charge/weight independent.

Cross-Correlation Score Distribution

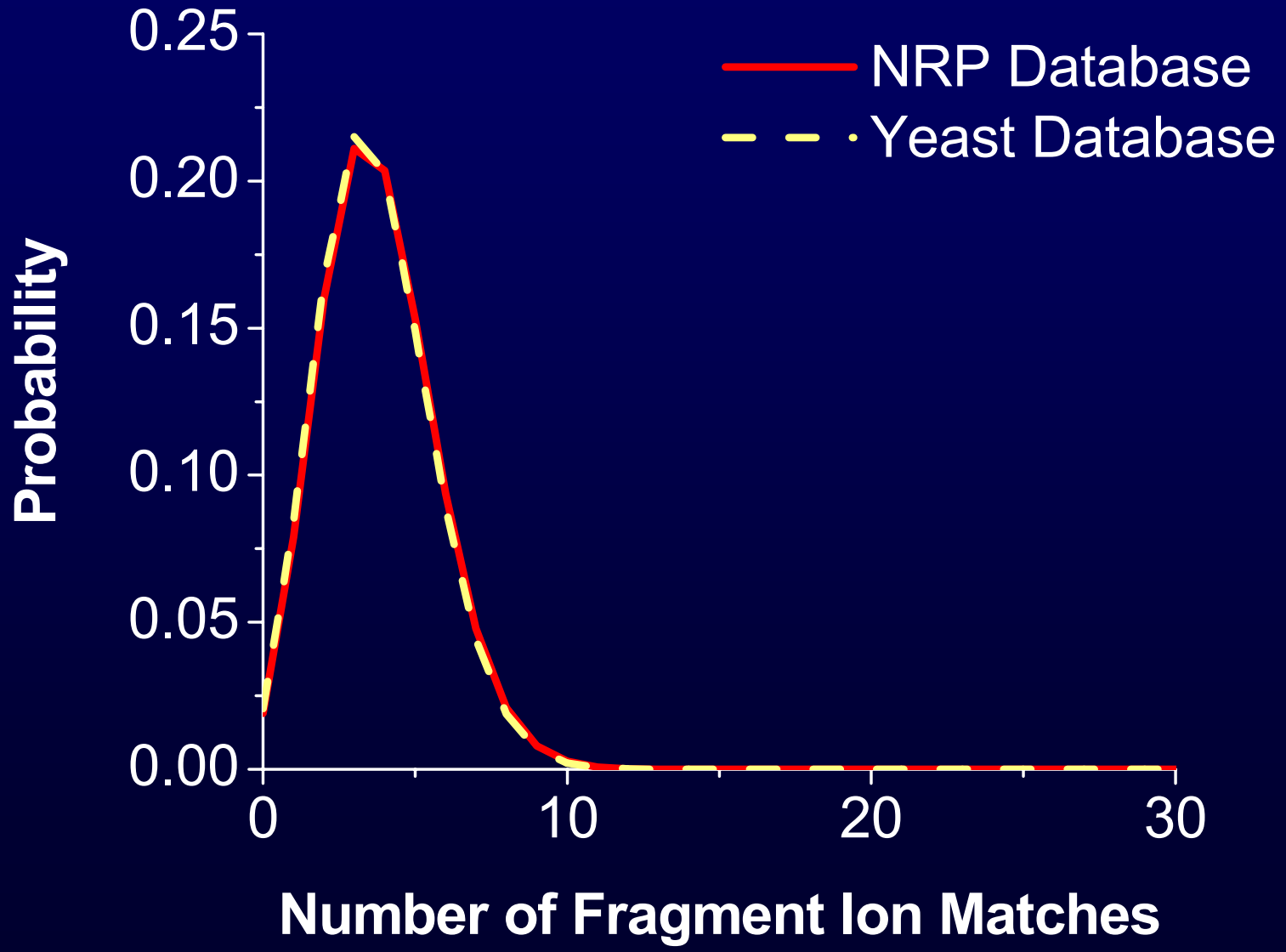


Database Dependence

- The probability inferred from the hypergeometric distribution is in principle database dependent.

- However, the dependency is very weak.

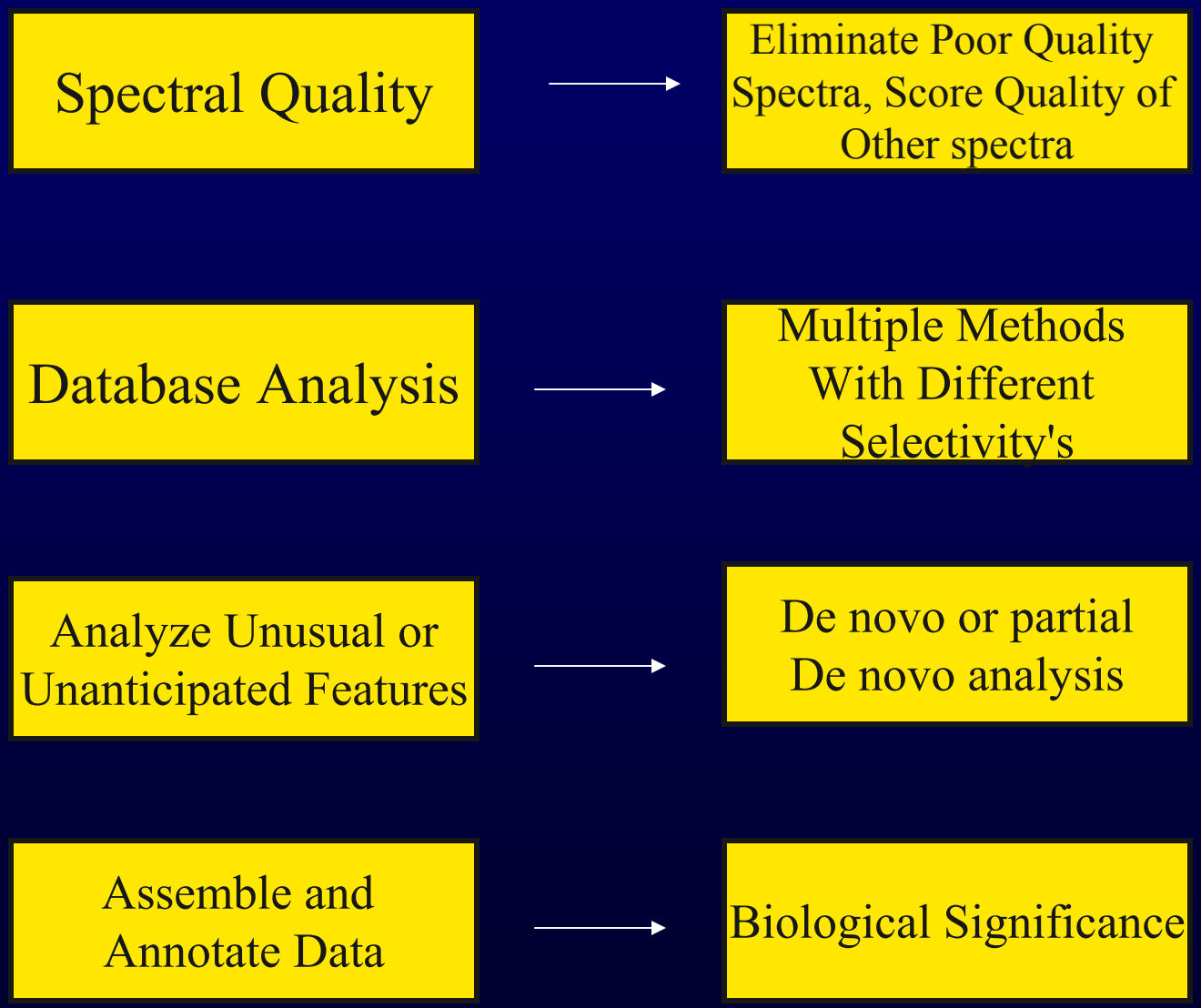
NRP and Yeast Databases



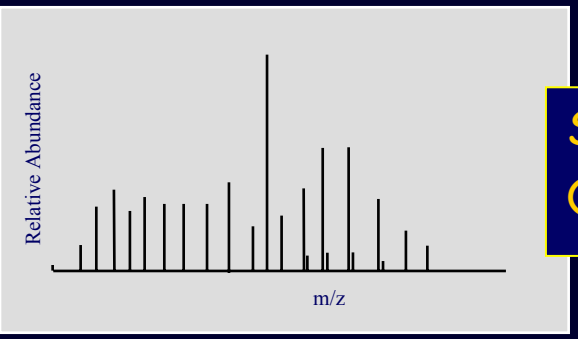
Pep_Probe Summary

- Implements 4 scoring schemes: hypergeometric, poisson, maximum likelihood and cross-correlation. Sorts results either by hypergeometric or cross-correlation scores.
- No enzyme specificity is assumed.
- Reports significance of each match.
- Can search for posttranslational modifications to three different amino acids.
- Has been implemented to run on a standalone or compute clusters.
- Runs on heterogeneous cluster of computers, in WINDOWS and LINUX platforms.

Processing Tandem Mass Spectra



Increased Data Production Requires Automated Data Analysis



Spectral Quality

Bioinformatics (in press)

LIBQUEST

MS/MS Comparison

Library Searching
Comparative Analysis
Subtractive Analysis

Yates et al. *Anal.Chem.* **70, 3557 (1998)**

SEQUEST & Pep_Prob

Database Search

Existing Sequence
PTM

Yates et al., *JASMS* 5, 976 (1994)
Yates et al. *Anal. Chem* 67, 1426 (1995)
Yates et al. *Anal. Chem*. 67, 3205 (1995)
MacCoss et al. *Anal. Chem* (2002)
Sadygov and Yates, *Anal. Chem* (in press)

Probability for Protein Identification

SEQUEST-SNP

Variant Search

Related Sequence
SNP Analysis
Mutation Analysis

Gatlin et al. *Anal.Chem.* 72, 757 (2000).

Link et al. *Nature Biotech.* **17**, 676-682 (1999)
Tabb et. al. *J. Proteome Res.* 1, 26, (2002)

GutenTag

De Novo Sequencing

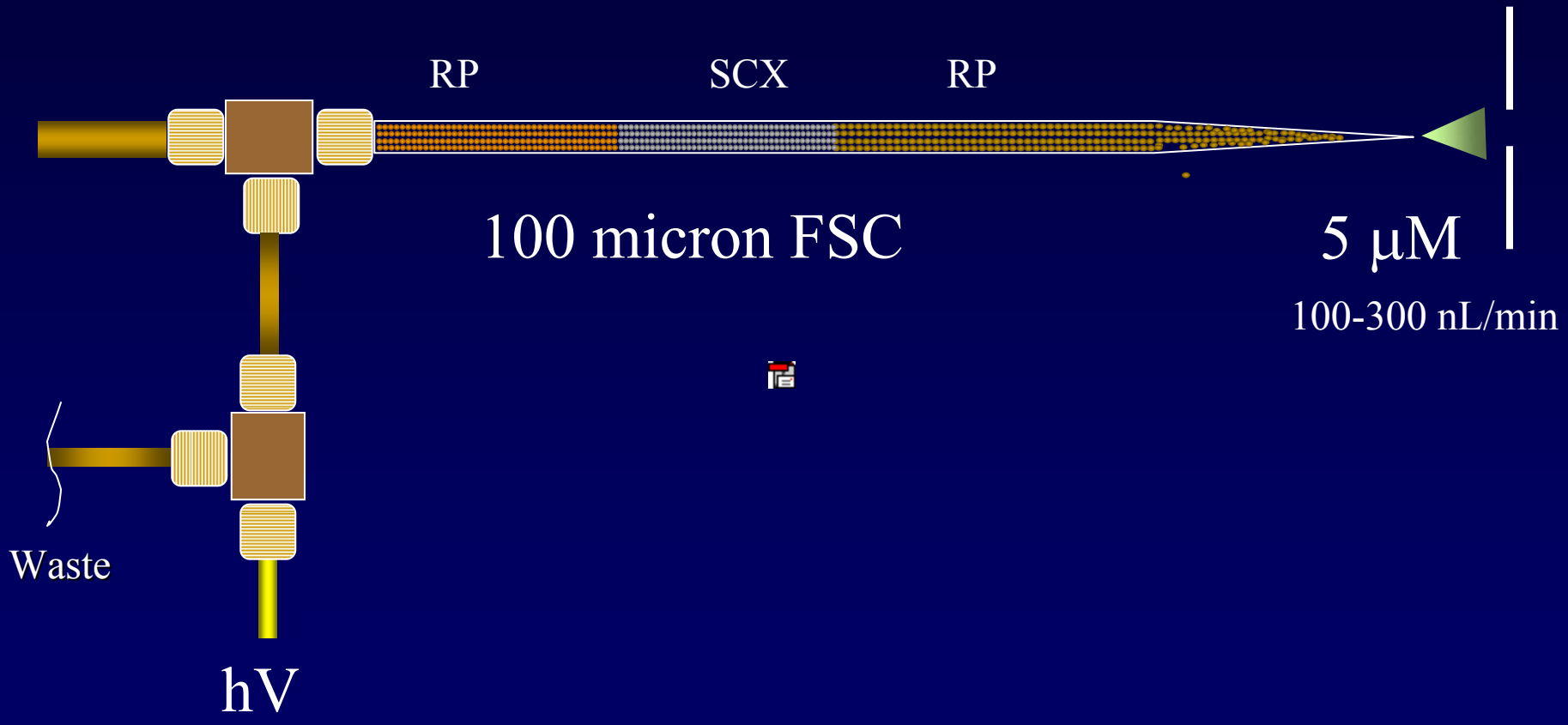
Alternate Splicing
Unantic. Mod.
Unknown ORFs

Tabb et al Anal. Chem

Data Considerations

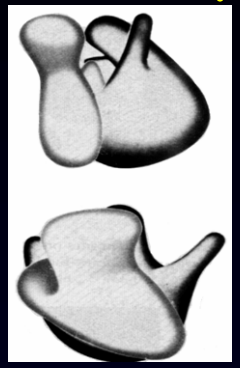
- Different types of experiments
- Different types of data analysis

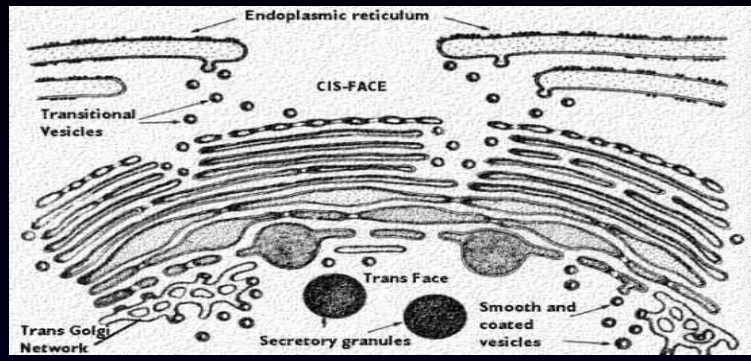
Integrated Multi-Dimensional Liquid Chromatography



Comprehensive Analysis of Complex Protein Mixtures

Cells/Tissues





Multiprotein Complex/Organelle

Total Protein Characterization

- Protein Identification: *What's there*
- Post Translational Modifications: *Regulation*
- Quantification: *Dynamics*
- Proteomic Data to Knowledge: *Genetics, RNAi, siRNA*

Translation of technology development into biological discovery