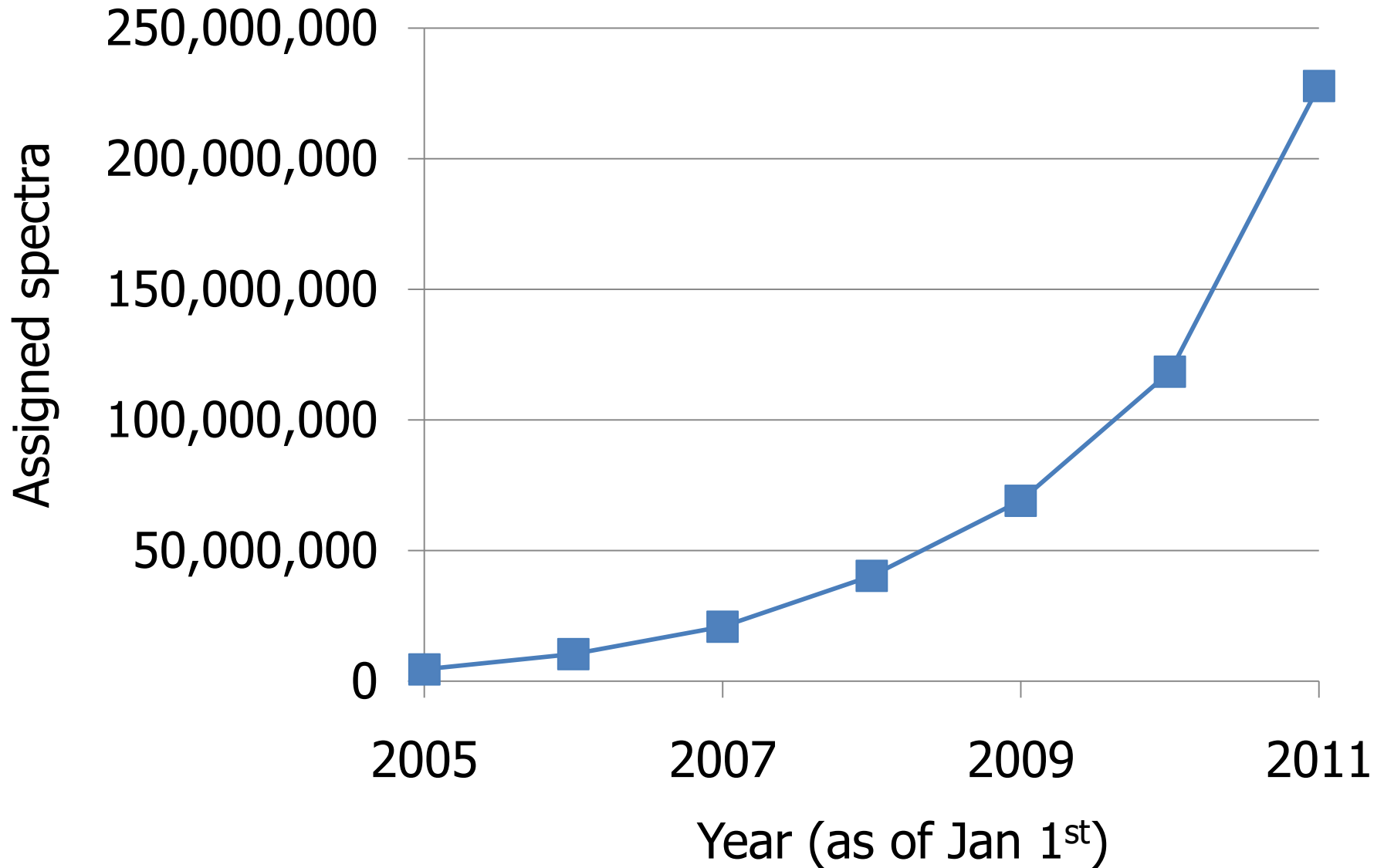


**Quality assessment of large data sets  
for use in the construction of  
spectrum libraries and data  
repositories.**

**Dan Evans  
David Fenyö  
Ron Beavis**

# Sequence-spectrum assignments in GPMDB



## Breakdown of assignments by organism

Species	Assigned spectra
<i>H. sapiens</i>	108,673,621
<i>M. musculus</i>	31,370,220
<i>S. cerevisiae</i>	11,628,207
<i>C. elegans</i>	7,344,749
<i>A. thaliana</i>	3,888,470
<i>D. rerio</i>	3,026,628
<i>R. norvegicus</i>	2,726,900
<i>G. gallus</i>	2,526,070
<i>D. melanogaster</i>	2,239,002
<i>A. gambiae</i>	2,079,653

# Spectrum library construction

List all peptides observed from a proteome

Assemble spectra from qualified data sets

Group/discard spectra (z, PTMs, patterns)

Create composite spectra from groups

Validate with fragmentation mechanisms

Align peptide sequences in proteome coordinates

# Methods for validating peptide assignments

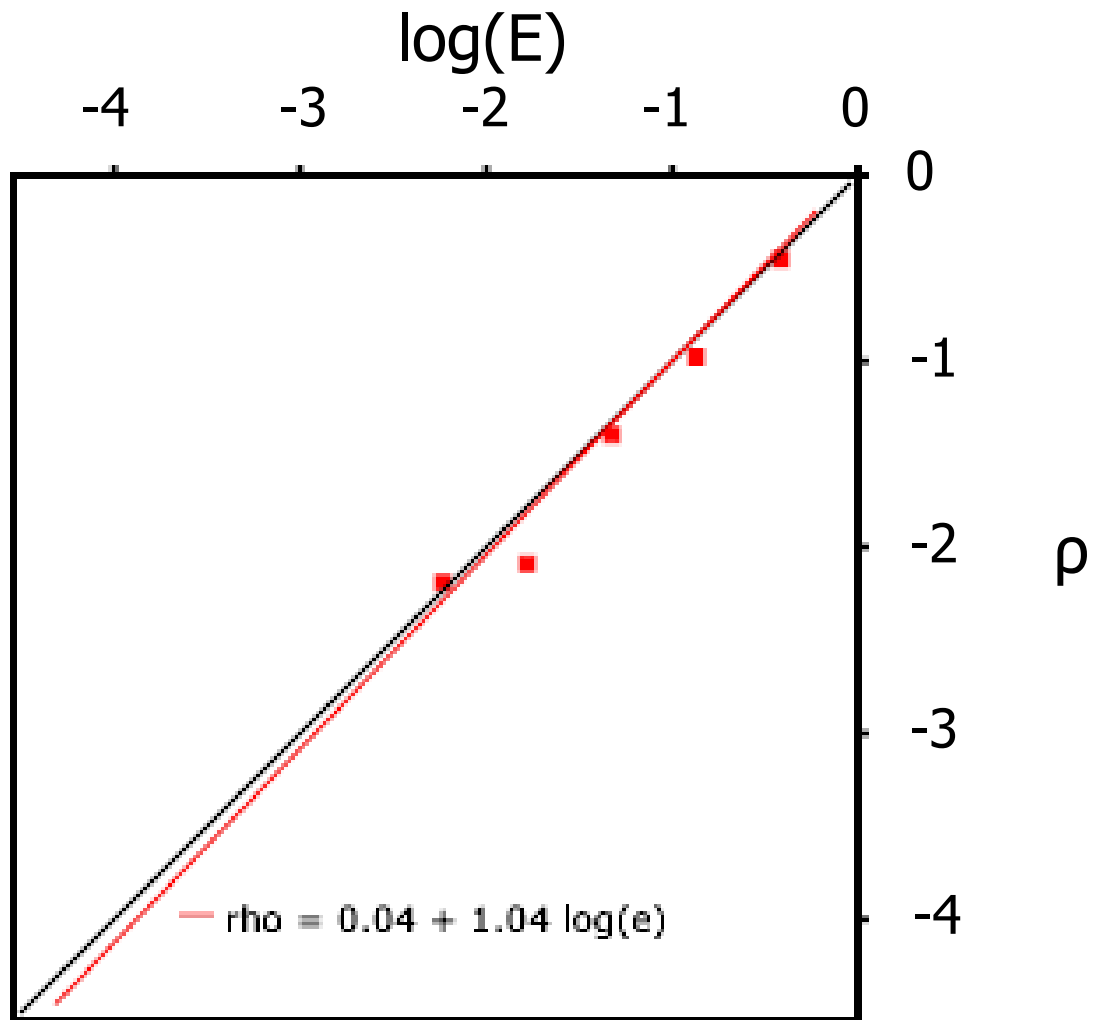
1. Statistical analysis of individual data sets.
2. Analysis of the physical properties of assigned peptide sequences.
3. Statistical analysis of multiple, unrelated data sets.

**What to expect for purely  
random (false) assignments?**

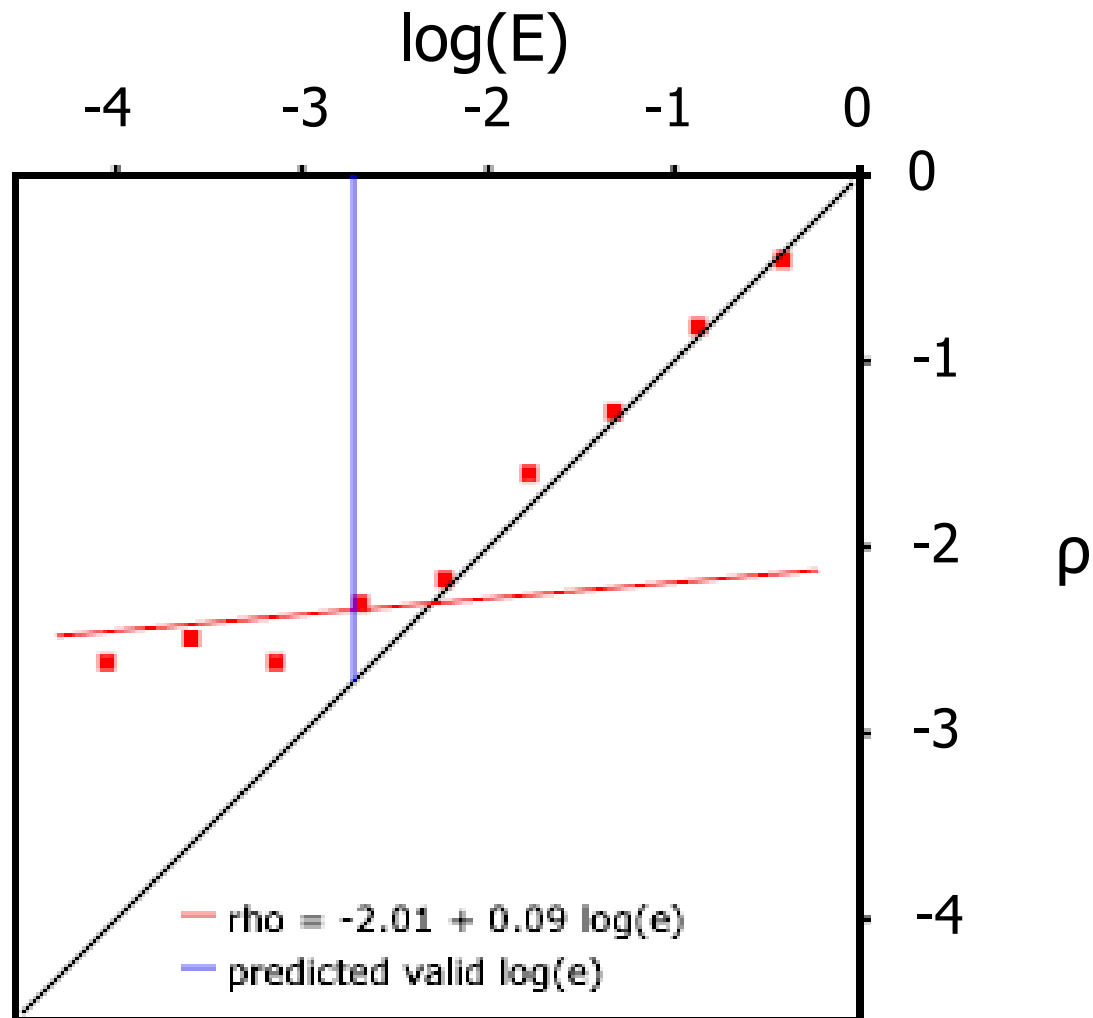
## $\rho$ -diagram construction

Expectation	Assignments (N)	$N/N_0$	$\rho = \log(N/N_0)$
1 to 0.1	1000 ( $N_0$ )	1.000	0
0.1 to 0.01	100	0.100	-1
0.01 to 0.001	10	0.010	-2
0.001 to 0.0001	1	0.001	-3

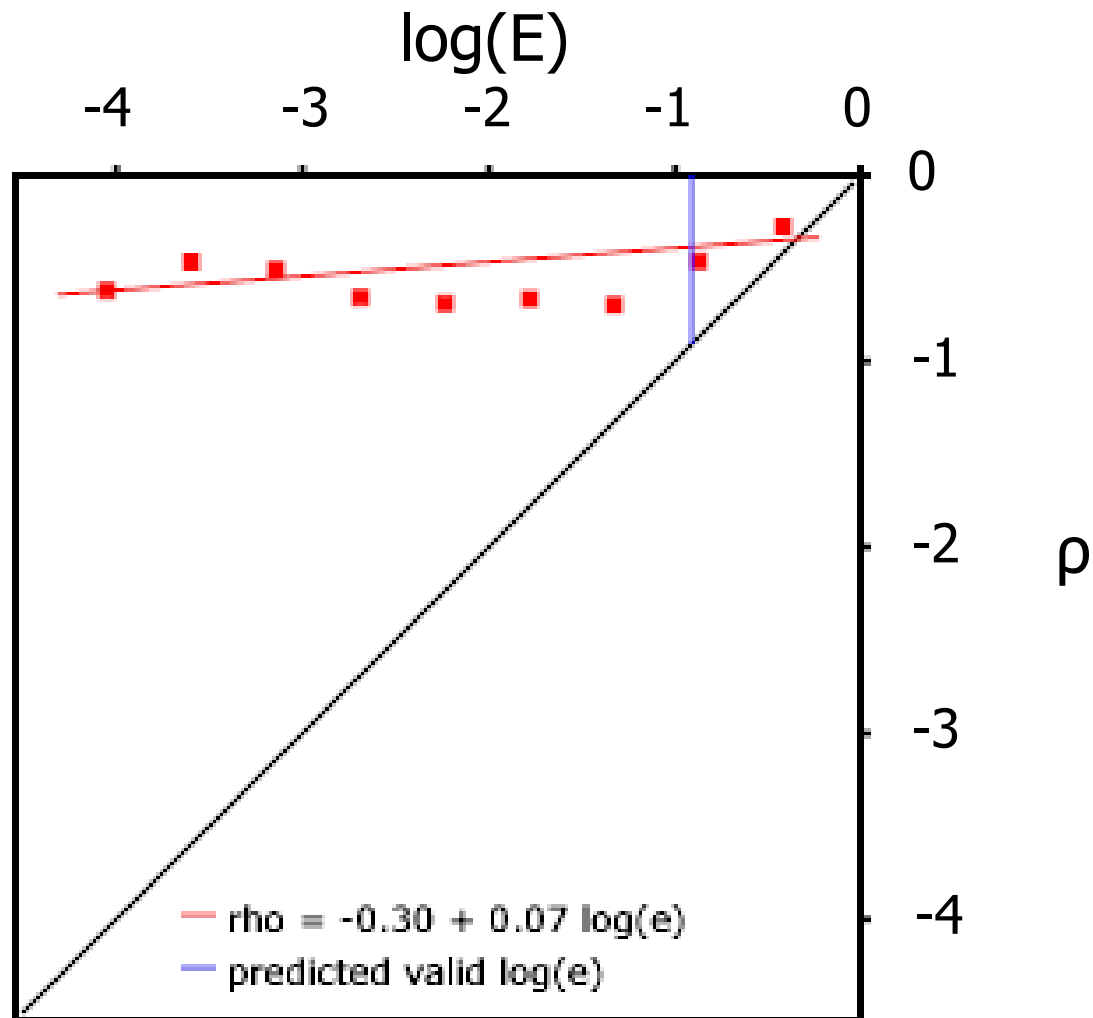
# $\rho$ -diagram: no correct assignments available



# $\rho$ -diagram: some correct assignments available



# $\rho$ -diagram: all correct assignments available



# **Physical properties of assigned peptide sequences:**

- 1. Charge**
- 2. Parent ion mass**
- 3. Amino acid analysis**

## Zeta ( $\zeta$ ) charge ratio

$$\zeta = z_{\text{measured}} / z_{\text{canonical}}$$

$z_{\text{measured}}$  – measured ion charge

$z_{\text{canonical}}$  – number of basic sites (H,K,R+nt)

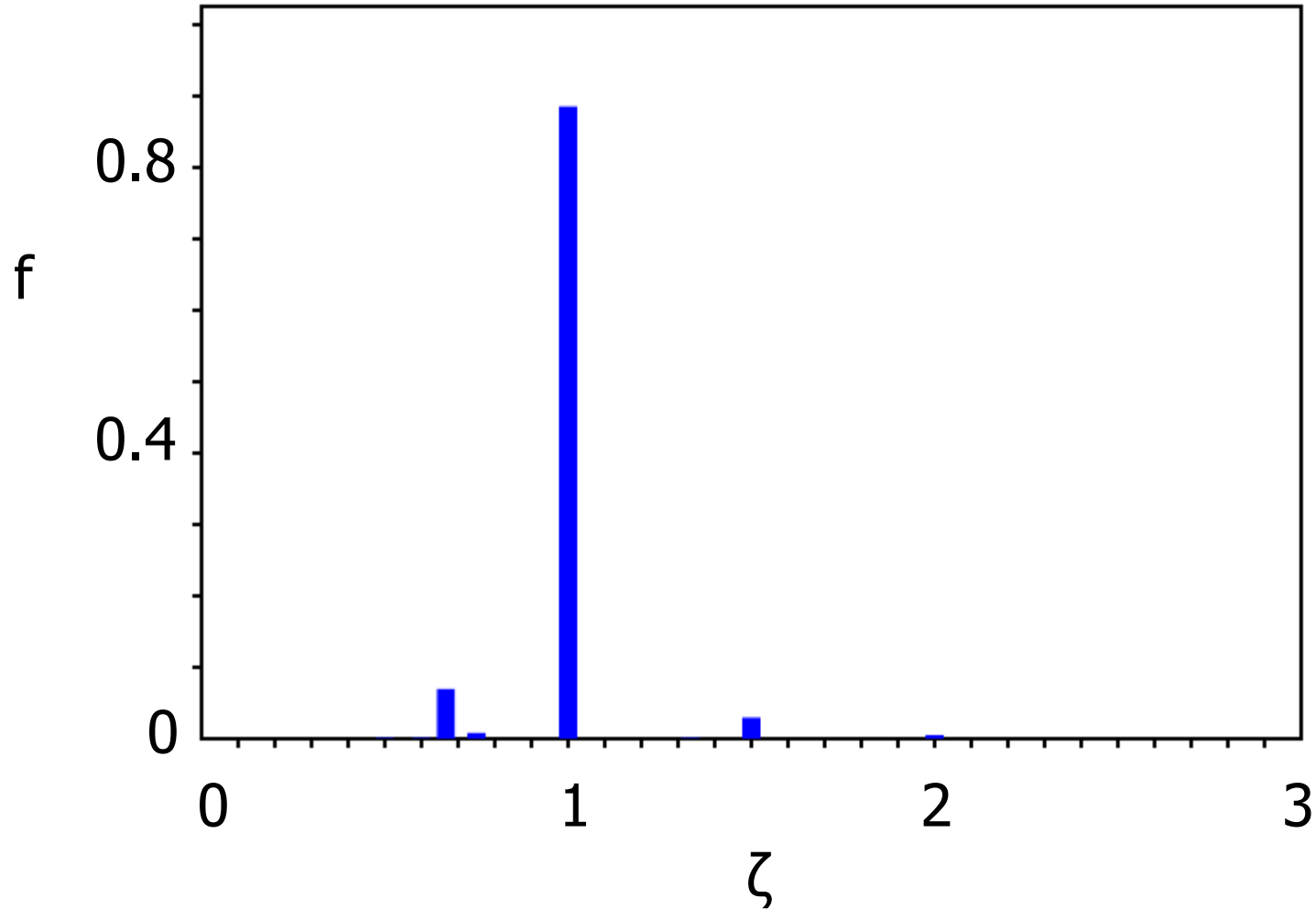
## Zeta ( $\zeta$ ) charge ratio: example

Measured ion charge = 2

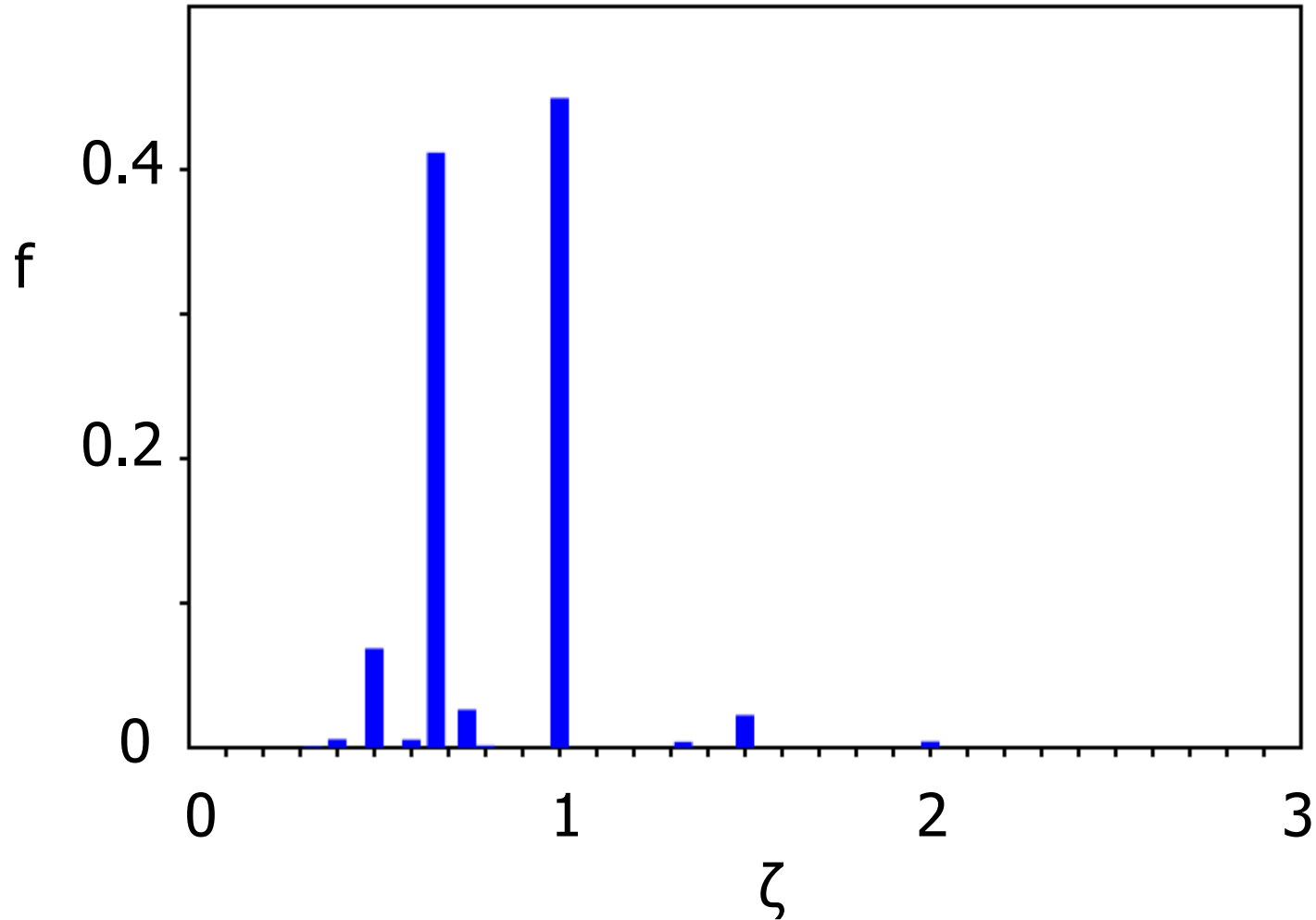
Peptide sequence = **NFTIEADK**

$$\zeta = 2/2 = 1$$

# Correct assignments available ( $\zeta$ vs. $f$ )



# No correct assignments available ( $\zeta$ vs. $f$ )

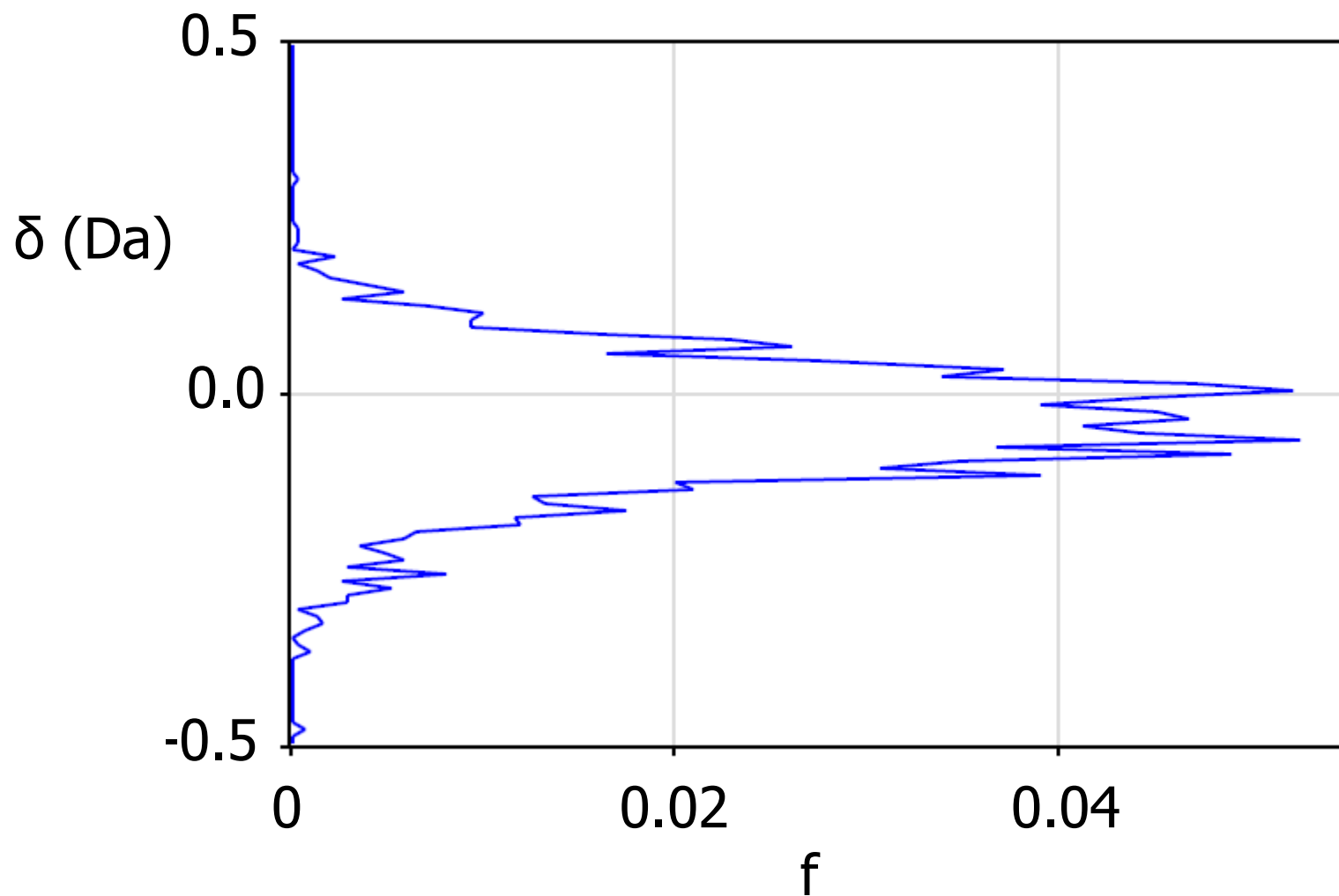


## Parent ion mass error ( $\delta$ )

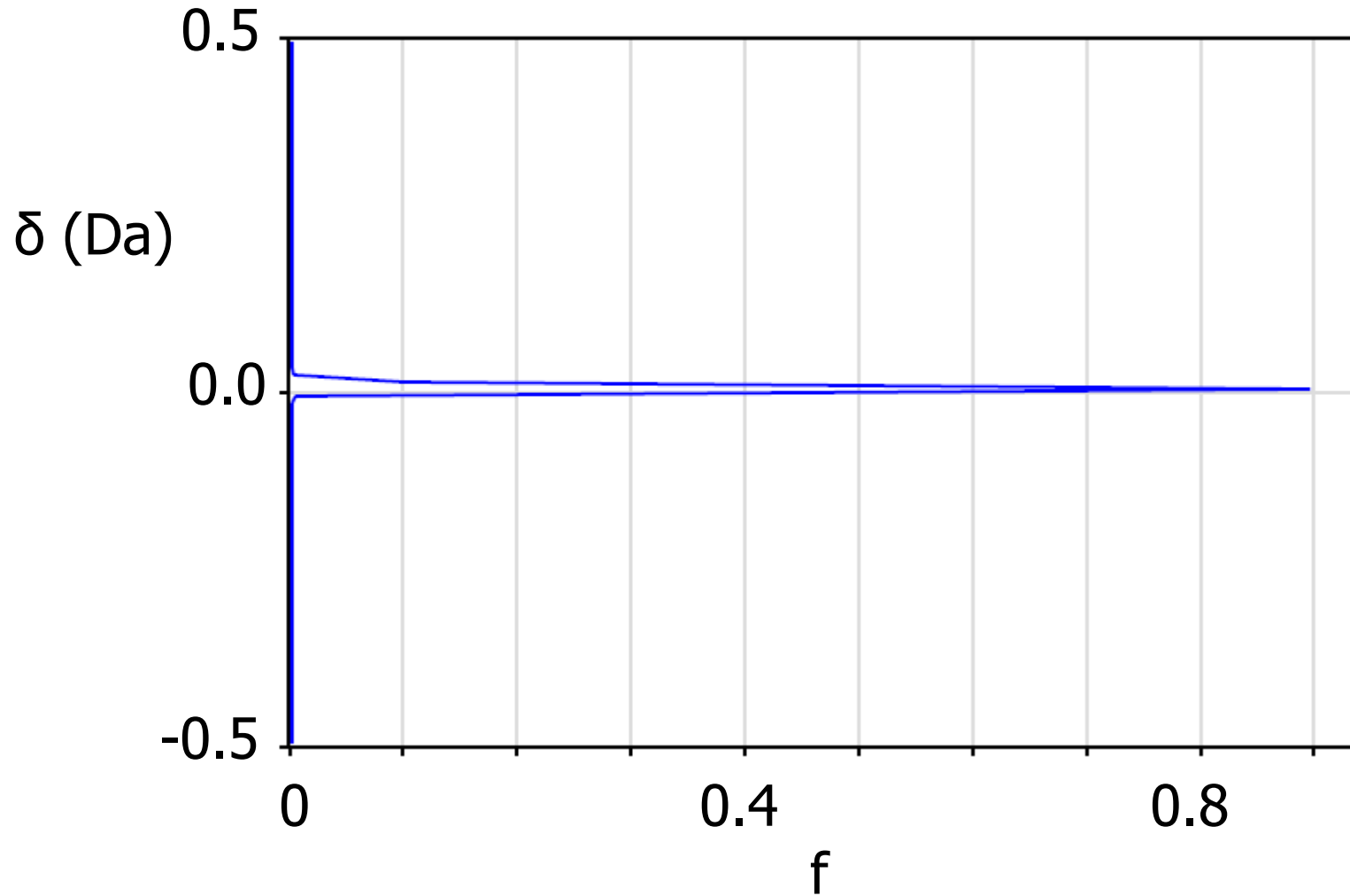
$$\delta = m_{\text{measured}} - m_{\text{calculated}}$$

( $\delta$  may be in Daltons or ppm)

No correct assignments available (f vs.  $\delta$ )



# Correct assignments available (f vs. $\delta$ )

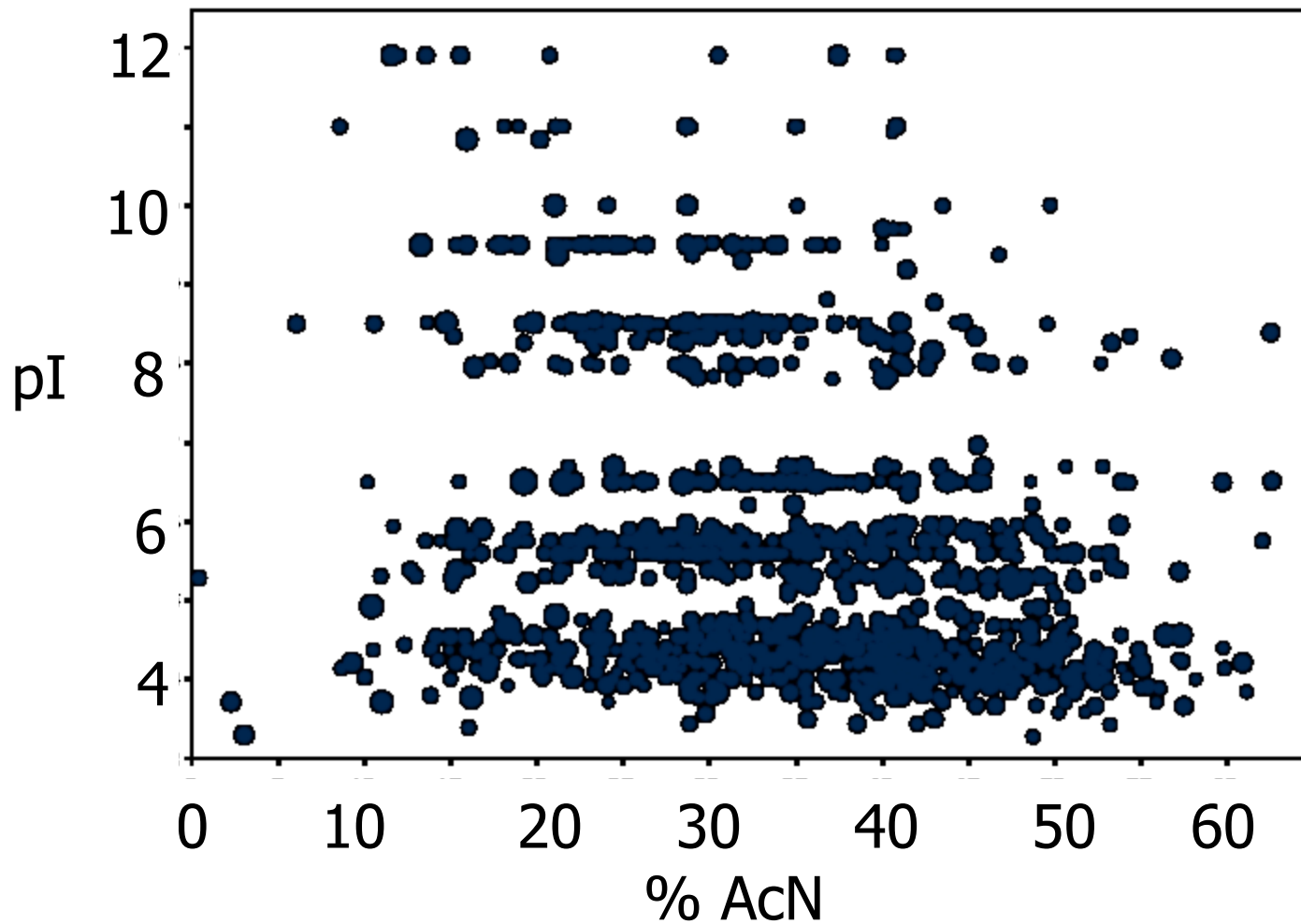


# Amino Acid Analysis (AAA)

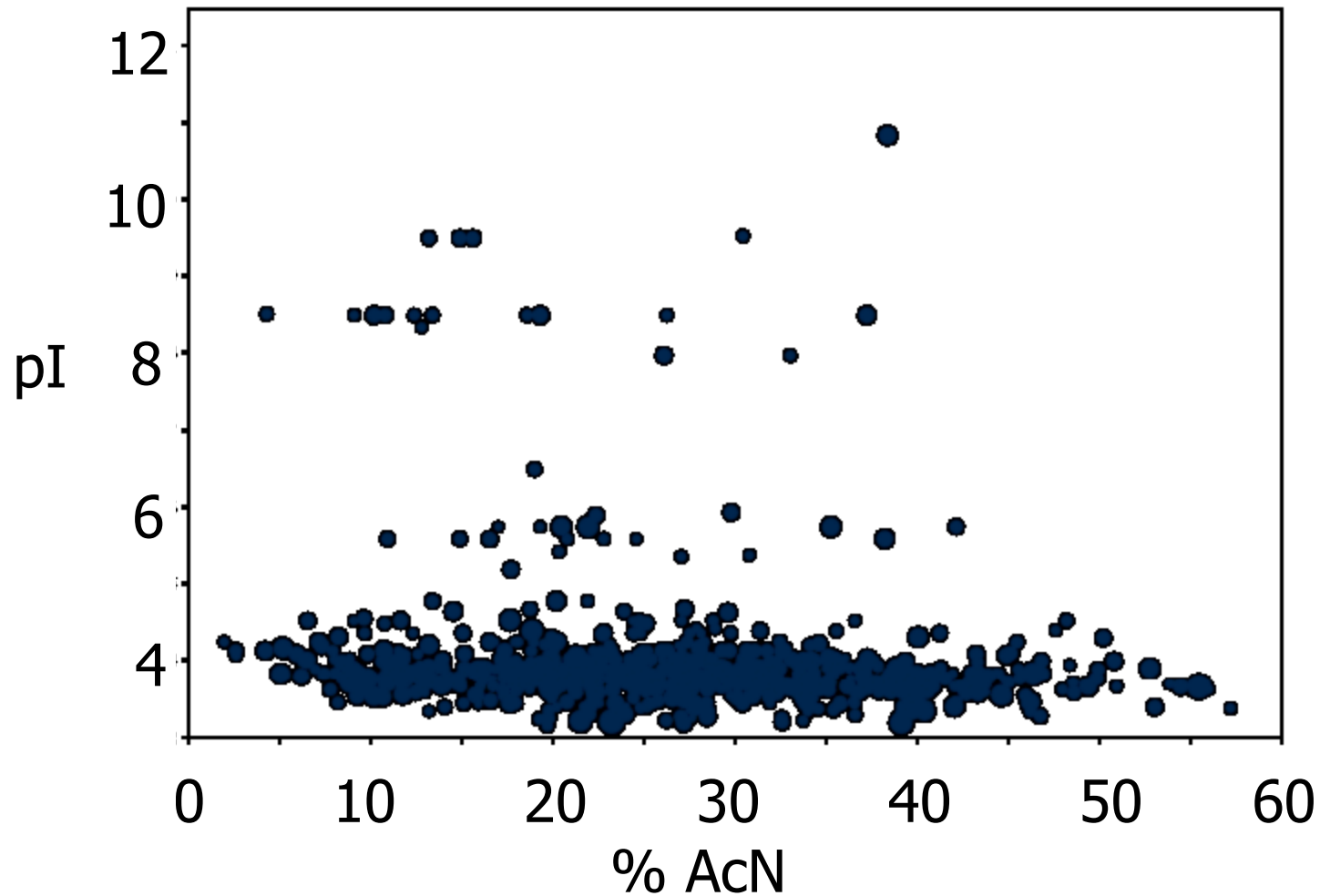
1. Total for proteins ID'd
2. Total for peptides ID'd
3. By residue position for peptides ID'd:
  - N-terminal
  - C-terminal
  - Before/after peptide

<b>Residue</b>	<b>Pre</b>	<b>N-term.</b>	<b>C-term.</b>	<b>Post</b>	<b>All</b>	<b>Protein</b>	<b>Unique</b>	<b>(U-P)%</b>
A	0	11.3	0.1	8.5	9.5	7.4	9	21
C	0	0.5	0	1.7	1.3	1.6	1.4	-13
D	0	7.2	0.1	5.9	6.7	5.3	6.6	24
E	0	7.4	0.1	8	9.1	7.9	9.1	15
F	0	4.6	0.1	3.3	3.2	3.5	3.3	-6
G	0	8	0	6.3	8.2	6.6	7.8	19
H	0	1.7	0	2	0.8	2.2	1	-55
I	0	6	0	5.5	4.5	4.6	4.5	-2
K	55.4	1.6	53.2	7.7	4.3	6.6	4.4	-33
L	0	10.7	0.1	9.8	8.4	9.7	8.8	-9
M	0.5	2.6	0	2.3	2.2	2.3	2	-13
N	0	5.2	0.1	3.9	4.3	3.6	4.2	16
P	0.1	0.2	0	0	4.8	5.6	4.8	-15
Q	0	1.4	0.1	4.6	5	4.9	5.2	6
R	43.7	1.7	45.4	6.1	3.6	5.7	3.7	-34
S	0	8.7	0.1	6.3	7.5	7.5	7.7	2
T	0	7	0.1	5.7	6.2	5.1	6	16
V	0	9.1	0.1	7.5	7.1	6.2	6.9	11
W	0	0.7	0	0.9	0.5	1	0.6	-40
Y	0	4.5	0.1	2.6	2.9	2.6	2.9	13

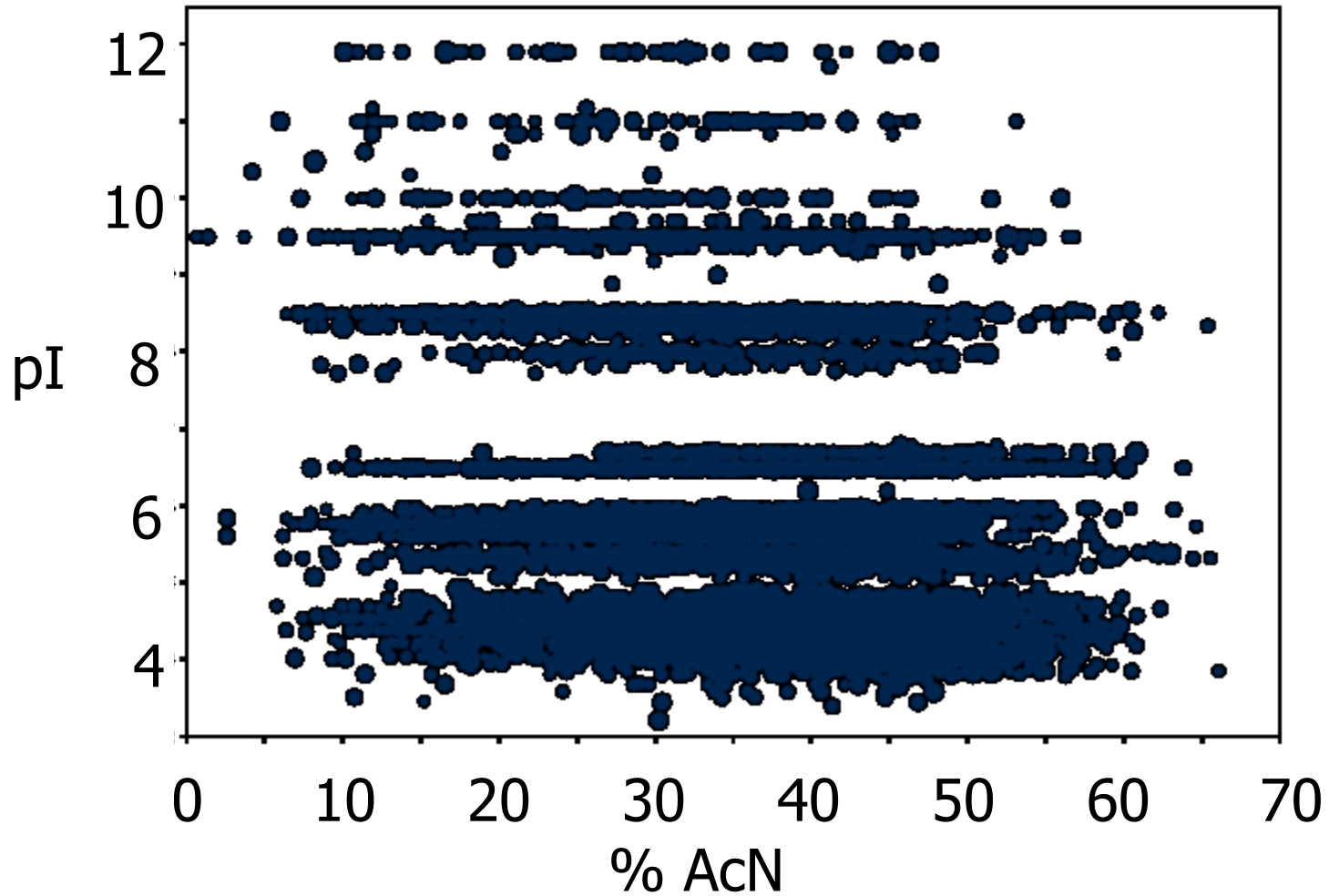
# Normal HPLC data set (%AcN vs. pI)



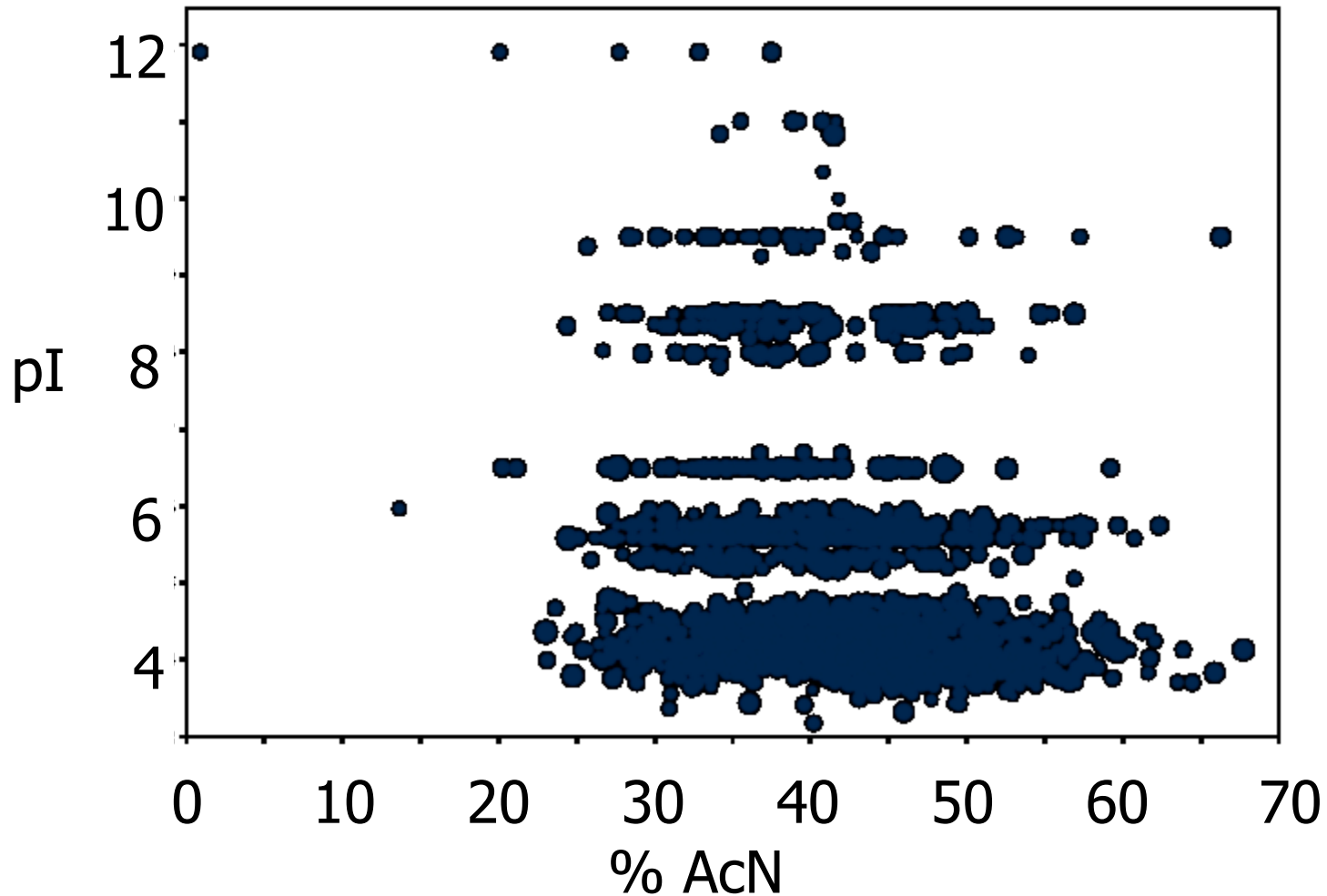
# TiO<sub>2</sub> + HPLC data set (%AcN vs. pI)



# Normal HPLC data set (%AcN vs. pI)

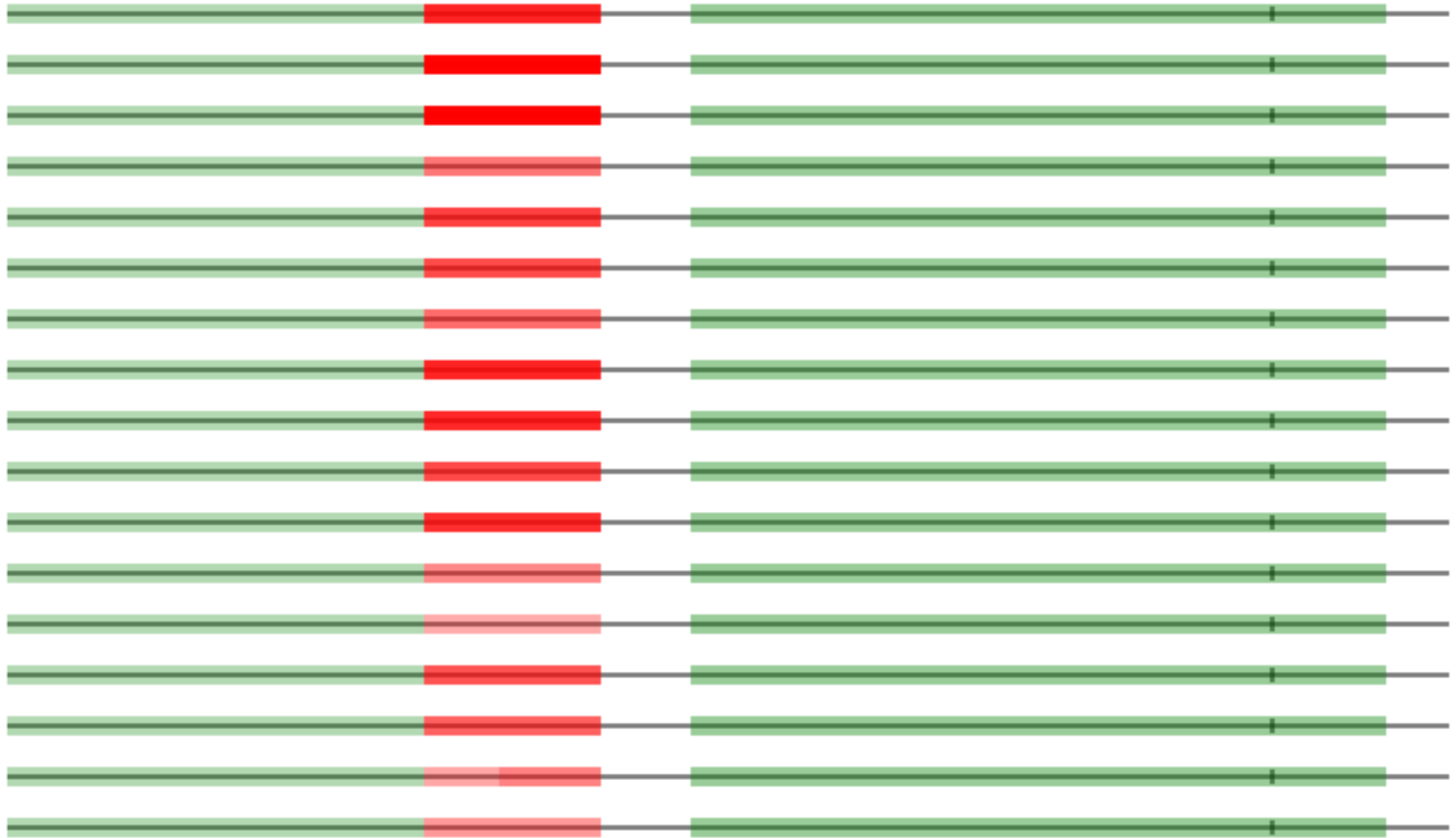


# HPLC data set – spray problems (%AcN vs. pI)



**Comparing a specific result  
with pools of results using  
objective measures.**

# MT-ND3 (human): Only one peptide observed



## Global frequency of observing a peptide

Peptide Sequence	Observations
FSTVAGESGSADTVR	2633
FNTANDDNVTQVR	2432
AFYVNVLNEEQR	1722
LVNANGEAVYCK	1701
GPLLVQDVVFTDEMAHFDR	1637
LSQEDPDYGIR	1560
LFAYPDTHR	1499
NLSVEDAAR	1400
FYTEDGNWDLVGNNTPIFFIR	1386
ADVLTTGAGNPVGDK	1338

## Global frequency of observing a peptide

If the number of times a peptide sequence ( $i$ ) has been observed is  $n_i$ , then for a particular protein:

$$N_{total} = \sum_i n_i$$

## Global frequency of observing a peptide ( $\omega$ )

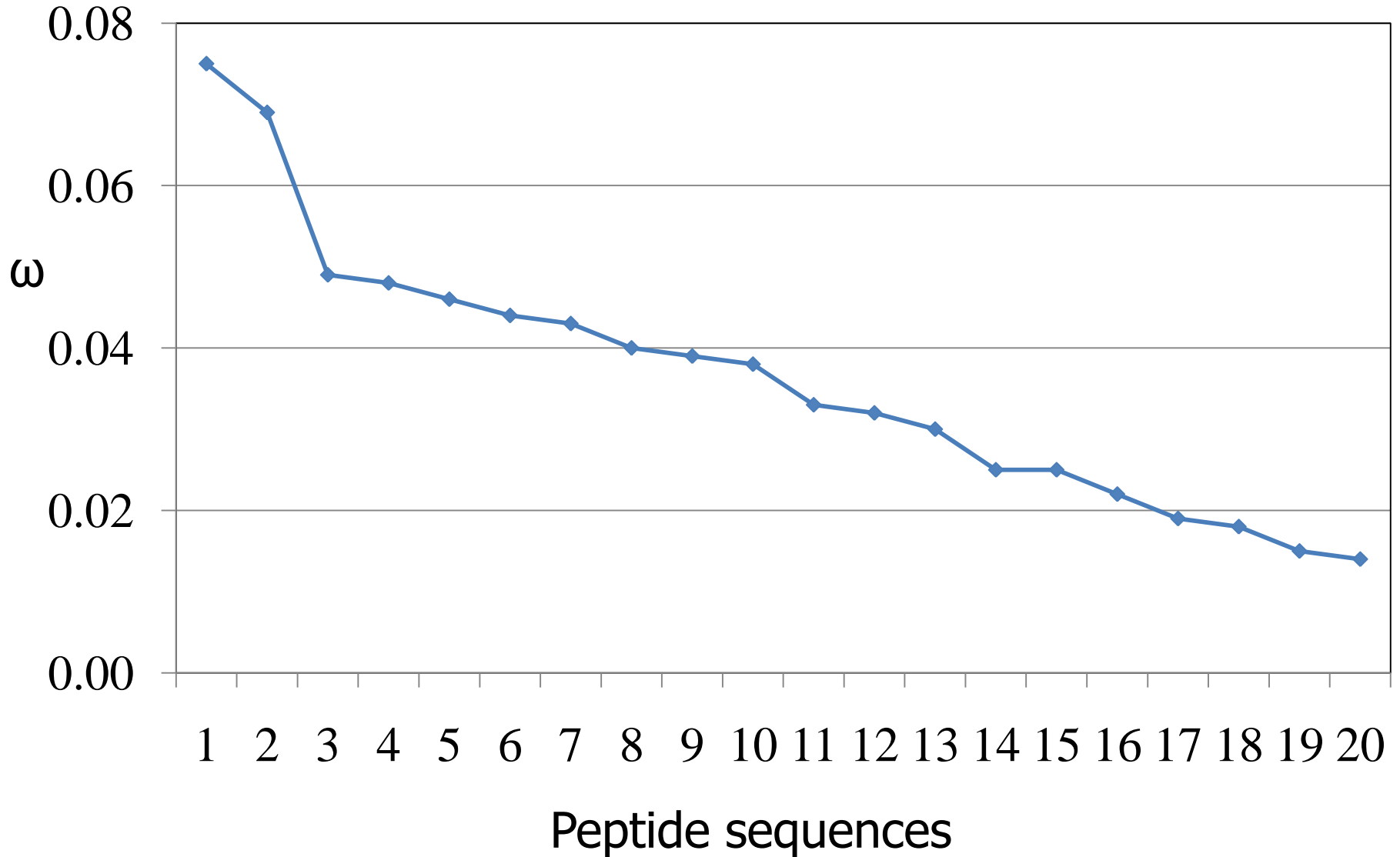
Define a normalized global frequency of observation for a particular peptide sequence from a particular protein as:

$$\omega_i = \frac{n_i}{N_{total}}$$

## Global frequency of observation ( $\omega$ ), catalase

Peptide Sequence	$\omega$
FSTVAGESGSADTVR	0.08
FNTANDDNVTQVR	0.07
AFYVNVLNNEEQR	0.05
LVNANGEAVYCK	0.05
GPLLVDVFTDEMAHFDR	0.05
LSQEDPDYGIR	0.04
LFAYPDTHR	0.04
NLSVEDAAR	0.04
FYTEDGNWDLVGNNTPIFFIR	0.04
ADVLTGAGNPVGDK	0.04

# Global frequency of observation ( $\omega$ ), catalase



## Omega ( $\Omega$ ) value for a protein identification

For any set peptides observed in an experiment assigned to a particular protein (*1 to j*):

$$\Omega(\textit{protein}) = \sum_j \omega_j$$

$$\Omega(\textit{protein}) \leq 1$$

## Protein $\Omega$ 's for a set of identifications

Protein ID	$\Omega$ (z=2)	$\Omega$ (z=3)
SERPINB1	0.88	0.82
SNRPD1	0.88	0.59
CFL1	0.81	0.87
SNRPE	0.8	0.81
PPIA	0.79	0.64
CSTA	0.79	0.36
PFN1	0.76	0.61
CAT	0.71	0.78
GLRX	0.66	0.8
CALM1	0.62	0.76
FABP5	0.57	0.17

## In conclusion

1. There are objective methods of testing the quality of a set of proteomics identifications:
  - independent of the ID algorithm.
2. Peptide physical characteristics can be used for “tuning” of algorithm-specific parameters.
3. These methods can be used to improve the overall consistency of annotated spectrum libraries:
  - they make “crowd-sourced” proteomics data more reliable.