PeptideProphet: Validation of Peptide Assignments to MS/MS Spectra

Andrew Keller

Outline

- Need to validate peptide assignments to MS/MS spectra
- Statistical approach to validation
- Running PeptideProphet software
- Interpreting results of PeptideProphet
- Exercises

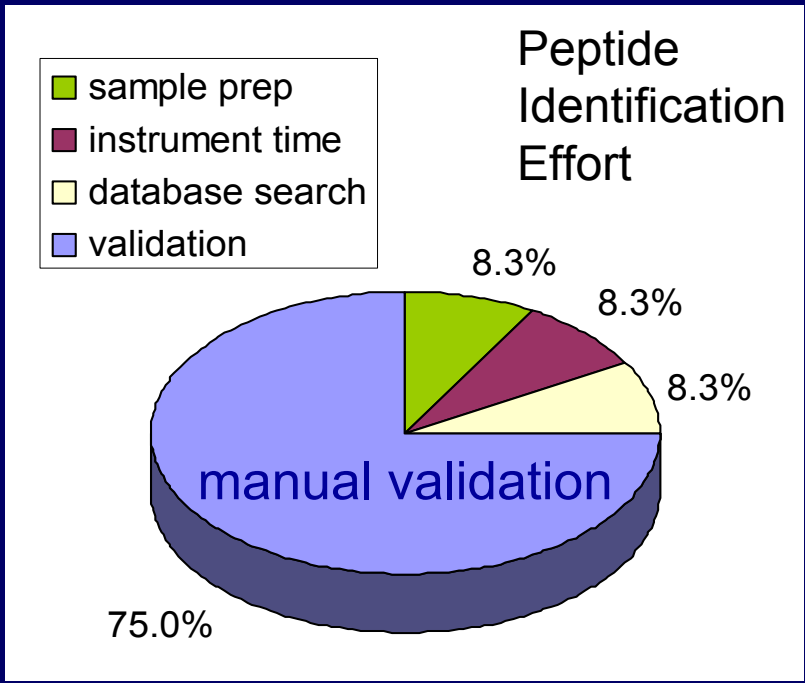
Most search results are wrong

- $[M+2H]^{2+}$/$[M+3H]^{3+}$ uncertainty (LCQ)
- Non-peptide noise
- Incomplete database
  e.g. post-translational modifications
- Limitation of database search algorithm

Validation of peptide assignments

In the past, a majority of analysis time was devoted to identifying the minority of correct search results from the majority of incorrect results

Required manual judgment



Results of 50 Spectrum Test

Consistency among 'Experts':
Of 50 search results
9 had < 67% 'publishable', 'borderline', or 'not pub'

Consistency of Individual 'Experts':
Of 10 duplicated search results, on average
0.4 were assessed 'publishable'/'not publishable'
2 were assessed inconsistently

The true validity of the search results are known.
Accuracy of 'Experts':
of 511 total 'publishable': 95% correct
of 102 total 'borderline': 49% correct
of 387 total 'not publishable': 14% correct

Even 'Experts' are not dependable!

Need for objective criteria

Manual scrutiny of search results is not practical for large datasets common to high throughput proteomics

As an alternative to relying on human judgment, many research groups employ search scores and properties of the assigned peptides to discriminate between correct and incorrect results

Traditional filtering criteria

Each SEQUEST search result has a:

Xcorr, dCn, Sp, NTT (number of tryptic termini)

Accept all results that satisfy:

$[M+2H]^{2+}$: Xcorr $\geq$ 2, dCn $\geq$ 0.1, Sp $\leq$ 50

$[M+3H]^{3+}$: Xcorr $\geq$ 2.5, dCn $\geq$ 0.1, Sp $\leq$ 50

Traditional filtering criteria

Each SEQUEST search result has a:

Xcorr, dCn, Sp, NTT (number of tryptic termini)

Accept all results that satisfy:
$[M+2H]^{2+}$: Xcorr $\geq$ 2, dCn $\geq$ 0.1, Sp $\leq$ 50, NTT $\geq$ 1
$[M+3H]^{3+}$: Xcorr $\geq$ 2.5, dCn $\geq$ 0.1, Sp $\leq$ 50, NTT $\geq$ 1

Traditional filtering criteria

Each SEQUEST search result has a:

Xcorr, dCn, Sp, NTT (number of tryptic termini)

Accept all results that satisfy:
$[M+2H]^{2+}$: Xcorr $\geq$ 2, dCn $\geq$ 0.1, Sp $\leq$ 50 (NTT $\geq$ 1)
$[M+3H]^{3+}$: Xcorr $\geq$ 2.5, dCn $\geq$ 0.1, Sp $\leq$ 50 (NTT $\geq$ 1)

$[M+2H]^{2+}$: Xcorr $\geq$ 2, dCn $\geq$ 0.1, Sp $\leq$ 50
$[M+3H]^{3+}$: Xcorr $\geq$ 2, dCn $\geq$ 0.1, Sp $\leq$ 50

Traditional filtering criteria

Each SEQUEST search result has a:

Xcorr, dCn, Sp, NTT (number of tryptic termini)

Accept all results that satisfy:
$[M+2H]^{2+}$: Xcorr $\geq$ 2, dCn $\geq$ 0.1, Sp $\leq$ 50 (NTT $\geq$ 1)
$[M+3H]^{3+}$: Xcorr $\geq$ 2.5, dCn $\geq$ 0.1, Sp $\leq$ 50 (NTT $\geq$ 1)

$[M+2H]^{2+}$: Xcorr $\geq$ 2, dCn $\geq$ 0.1, Sp $\leq$ 50, NTT $\geq$ 1
$[M+3H]^{3+}$: Xcorr $\geq$ 2, dCn $\geq$ 0.1, Sp $\leq$ 50, NTT $\geq$ 1

Traditional filtering criteria

Each SEQUEST search result has a:

Xcorr, dCn, Sp, NTT (number of tryptic termini)

Accept all results that satisfy:
$[M+2H]^{2+}$: Xcorr $\geq$ 2, dCn $\geq$ 0.1, Sp $\leq$ 50 (NTT $\geq$ 1)
$[M+3H]^{3+}$: Xcorr $\geq$ 2.5, dCn $\geq$ 0.1, Sp $\leq$ 50 (NTT $\geq$ 1)

$[M+2H]^{2+}$: Xcorr $\geq$ 2, dCn $\geq$ 0.1, Sp $\leq$ 50 (NTT $\geq$ 1)
$[M+3H]^{3+}$: Xcorr $\geq$ 2, dCn $\geq$ 0.1, Sp $\leq$ 50 (NTT $\geq$ 1)

Problems with traditional filtering

- Different research groups use different thresholds
- Combines scores in unsatisfactory manner:
  What if Xcorr is just below its threshold, but dCn is far above?
- Divides data into correct and incorrect- no in between
- Unknown error rates (fraction of data passing filter that are incorrect)
- Unknown sensitivity (fraction of correct results passing filter)
- Appropriate threshold may depend on database, mass spectrometer type, sample, etc.

Statistical Approach

Use search scores and properties of the assigned peptides to compute a probability that each search result is correct

Desirable model properties:

- Accurate
- High power to discriminate correct and incorrect results
- Robust

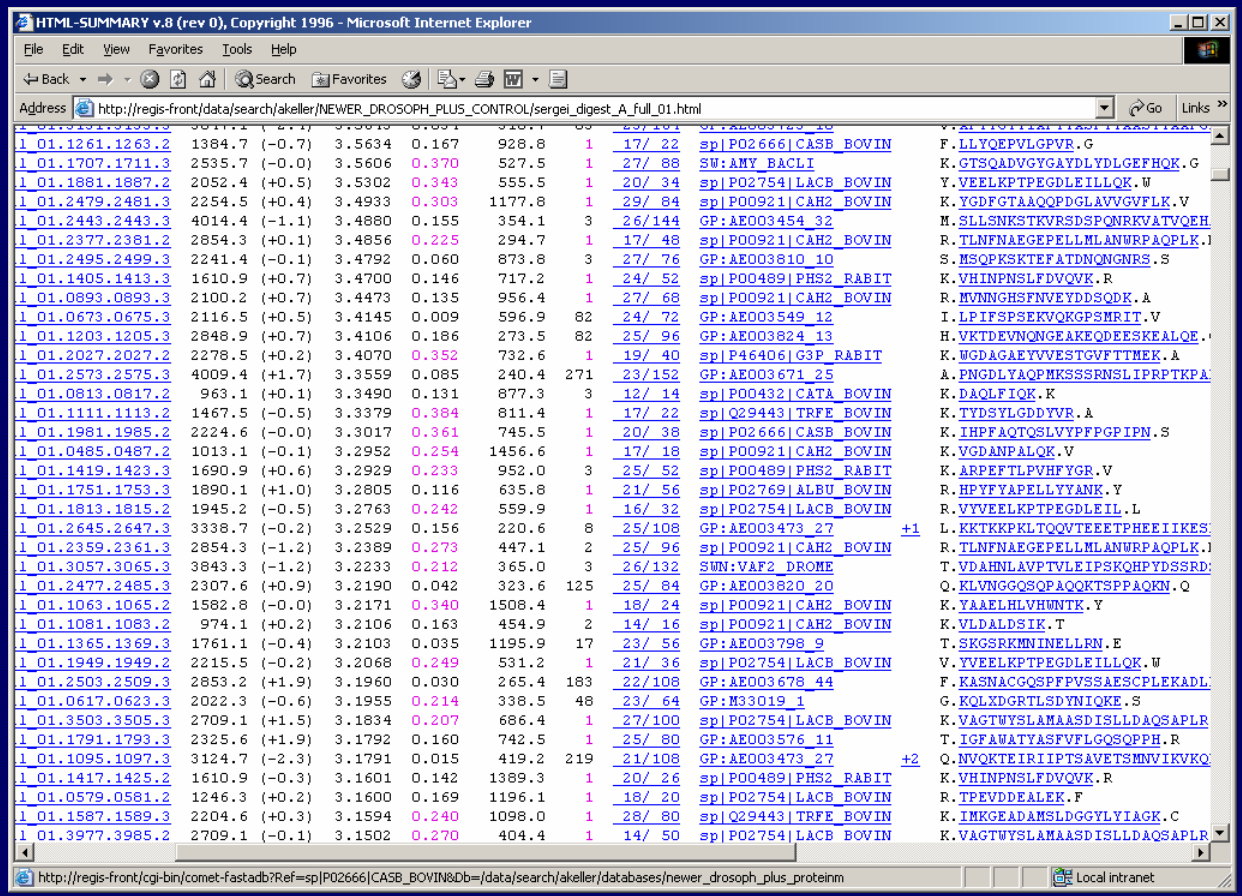
Training dataset

Want dataset of SEQUEST search results for which the true correct and incorrect peptide assignments are known

Sample of 18 control proteins (bovine, yeast, bacterial)

Collect ~40,000 MS/MS spectra, and search using SEQUEST vs. a Drosophila database appended with sequences of 18 control proteins and common sample contaminants

Training dataset



Peptides corresponding to drosophila proteins are incorrect

Peptides corresponding to 18 control proteins or contaminants are correct*

Combine multiple SEQUEST scores into single discrminant score F

Want to combine together Xcorr, dCn, and Sp in a linear manner to produce a new score, F, that maximally separates the correct and incorrect search results in the training dataset:

$$F = c_0 + c_1\text{Xcorr} + c_2\text{dCn} + c_3\text{Sp}$$

Actually first transform Xcorr and Sp:

$[M+2H]^{2+}$: Xcorr′ = log(Xcorr) / (log 2 * peplength)

$[M+3H]^{3+}$: Xcorr′ = log(Xcorr) / (log 4 * peplength)

Sp′ = log Sp

Derive Discriminant Fucntion

Derive F for each precursor ion charge separately:

$F = c_0 + c_1 Xcorr' + c_2 dCn + c_3 Sp'$

For $[M+3H]^{3+}$ search results in training datset,

$c_0 = -2.0$
$c_1 = 10.68$
$c_2 = 11.26$
$c_3 = -0.2$

$F = -2.0 + 10.68 * Xcorr' + 11.26 * dCn - 0.2 * Sp'$

Compute Discriminant Score

$F = -2.0 + 10.68 * Xcorr' + 11.26 * dCn - 0.2 * Sp'$

Example:

Peptide **K.ARPEFTLPVHFYGR.V**

Xcorr = 3.29

dCn = 0.233

Sp = 3

Precursor Ion Charge = 3

Peplength = 14

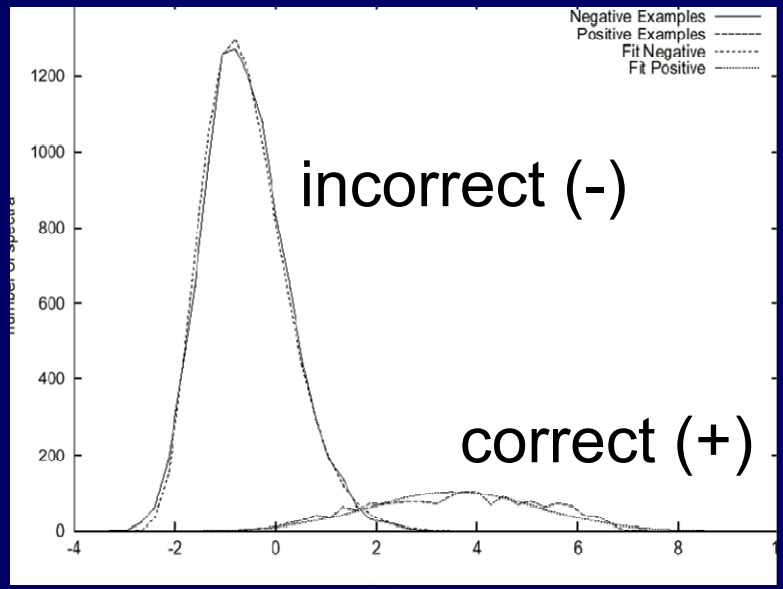
Xcorr′ = log(3.29)/(log 56) = 0.296

Sp′ = log(3) = 1.09

F = -2.0 + 10.68 * 0.296 + 11.26 * 0.233 +
-0.2 * 1.09 = 3.56

Discriminant Score Distributions

no of spectra

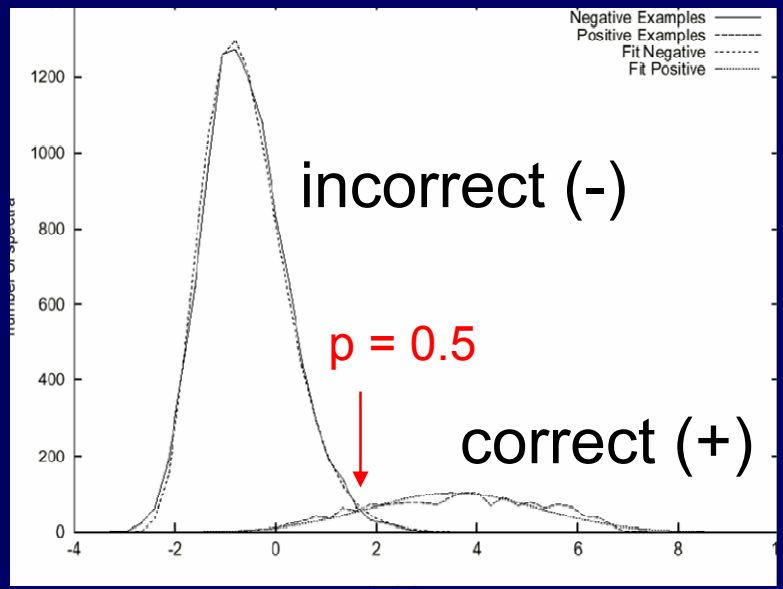


discriminant score (F)

Training dataset $[M+2H]^{2+}$ spectra

Computing probabilities from discriminant score distributions

no of spectra



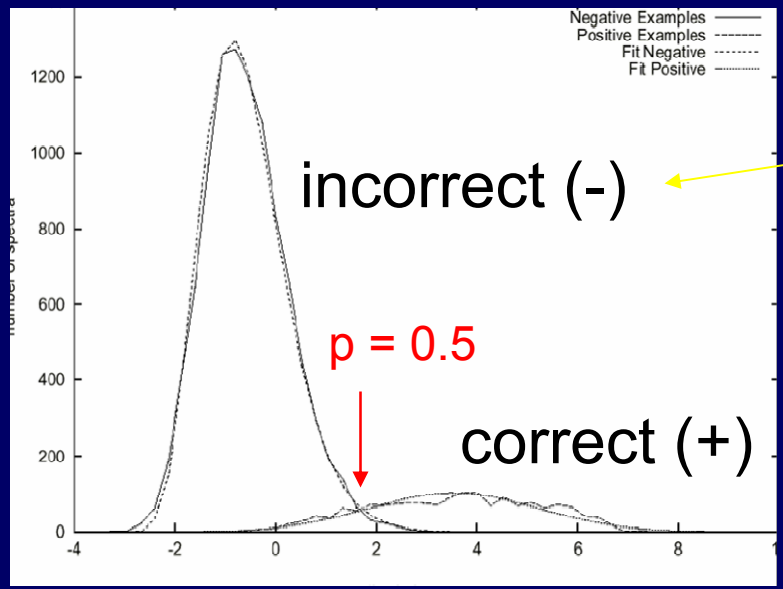
discriminant score (F)

Probability of being correct, given discriminant score F_{obs}, is:

$$p = \frac{\text{Number of correct search results with } F_{obs}}{\text{Total number of search results with } F_{obs}}$$

Computing probabilities from discriminant score distributions

no of spectra



discriminant score (F)

Model Incorrect results as ***Gamma Distribution***

Model Correct results as ***Normal Distribution***

Probability of being correct, given discriminant score F_{obs}, is:

$$p = \frac{\text{Normal}_{\mu,\sigma}(F_{obs}) * \text{Total correct}}{\text{Normal}_{\mu,\sigma}(F_{obs}) * \text{Total correct} + \text{Gamma}_{\alpha,\beta,\text{zero}}(F_{obs}) * \text{Total incorrect}}$$

Employing peptide properties

Properties of the assigned peptides, in addition to search scores, are useful information for distinguishing correct and incorrect results

For example in unconstrained SEQUEST searches for MS/MS spectra collected from trypsinized samples, a majority of correct assigned peptides have 2 tryptic termini (preceded by K,R), whereas a majority of incorrect assigned peptides have 0 tryptic termini

Number of tryptic termini (NTT)

NTT can equal 0, 1, or 2:

G.HVEQLDSSS.D NTT = 0

K.HVEQLDSSS.D NTT = 1
G.HVEQLDSSR.D NTT = 1

K.HVEQLDSSR.D NTT = 2

Number of tryptic termini (NTT)

For the same value of F, assigned peptides with *higher* NTT values are *more* likely to be correct

Example: training dataset
Correct: 0.03 NTT=0, 0.28 NTT=1, 0.69 NTT=2
Incorrect: 0.80 NTT=0, 0.19 NTT=1, 0.01 NTT=2

Probability of being correct, given discriminant score F_{obs} with NTT=2 is:

$$p = \frac{\text{Normal}_{\mu,\sigma}(F_{obs}) * \text{Total corr} * 0.69}{\text{Normal}_{\mu,\sigma}(F_{obs}) * \text{Total corr} * 0.69 + \text{Gamma}_{\alpha,\beta,\text{zero}}(F_{obs}) * \text{Total incorr} * 0.01}$$

F_{obs}: p = 0.5 without NTT becomes p=0.99 using NTT

Number of tryptic termini (NTT)

For the same value of F, assigned peptides with *lower* NTT values are *less* likely to be correct

Example: training dataset
Correct: 0.03 NTT=0, 0.28 NTT=1, 0.69 NTT=2
Incorrect: 0.80 NTT=0, 0.19 NTT=1, 0.01 NTT=2

Probability of being correct, given discriminant score F_{obs} with NTT=0 is:

$$p = \frac{\text{Normal}_{\mu,\sigma}(F_{obs}) * \text{Total corr} * 0.03}{\text{Normal}_{\mu,\sigma}(\text{Fobs}) * \text{Total corr} * 0.03 + \text{Gamma}_{\alpha,\beta,\text{zero}}(F_{obs}) * \text{Total incorr} * 0.80}$$

F_{obs}: p = 0.5 without NTT becomes p=.04 using NTT

Additional peptide properties

Number of missed tryptic cleavages (NMC)

Mass difference between precursor ion and peptide

Presence of light or heavy cysteine (ICAT)

Presence of N-glyc motif (N-glycosylation capture)

Calculated pI (FFE)

Incorporate similar to NTT above, assuming independence of peptide properties and search scores among correct and incorrect results

Computed Probabilities

Given training dataset distributions of F, NTT, NMC, Massdiff, ICAT, N-glyc, and pI among correct and incorrect search results,…

…then the probability of any search result with F_{obs}, NTT_{obs}, NMC_{obs}, $Massdiff_{obs}$, $ICAT_{obs}$, $N\text{-}glyc_{obs}$, and pI_{obs} can be computed as described above, with terms for each piece of information

Accurate

Discriminating

Robust Model

One cannot rely on the ***training dataset*** distributions of F, NTT, NMC, Massdiff, ICAT, N-glyc, and pI among correct and incorrect search results

These distributions are expected to vary depending on:

- Database used for search
- Mass spectrometer
- Spectrum quality
- Sample preparation and purity

EM Algorithm

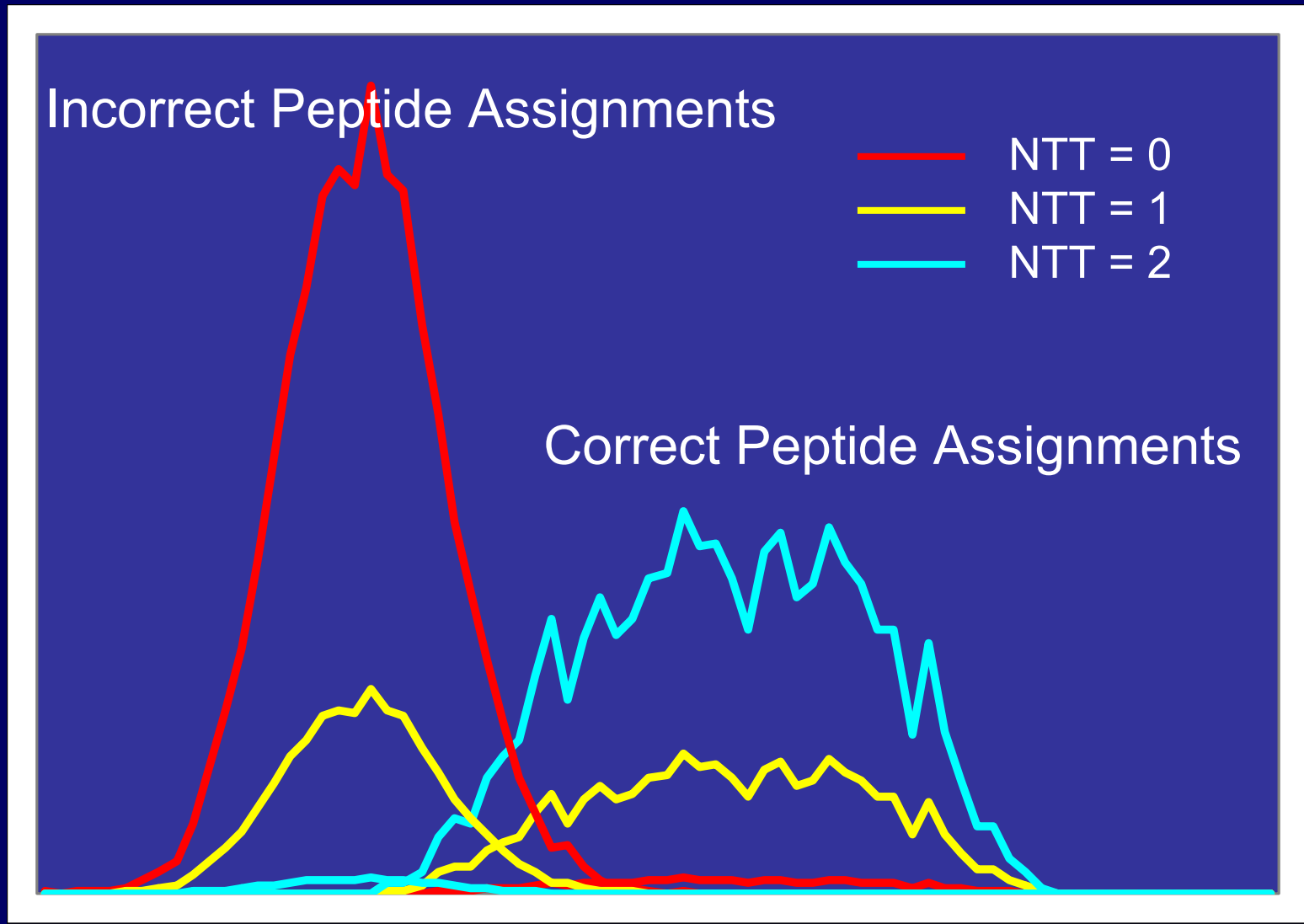
PeptideProphet learns the distributions of F and peptide properties among correct and incorrect search results in each dataset

It then uses the learned distributions to compute probabilities that each search result is correct

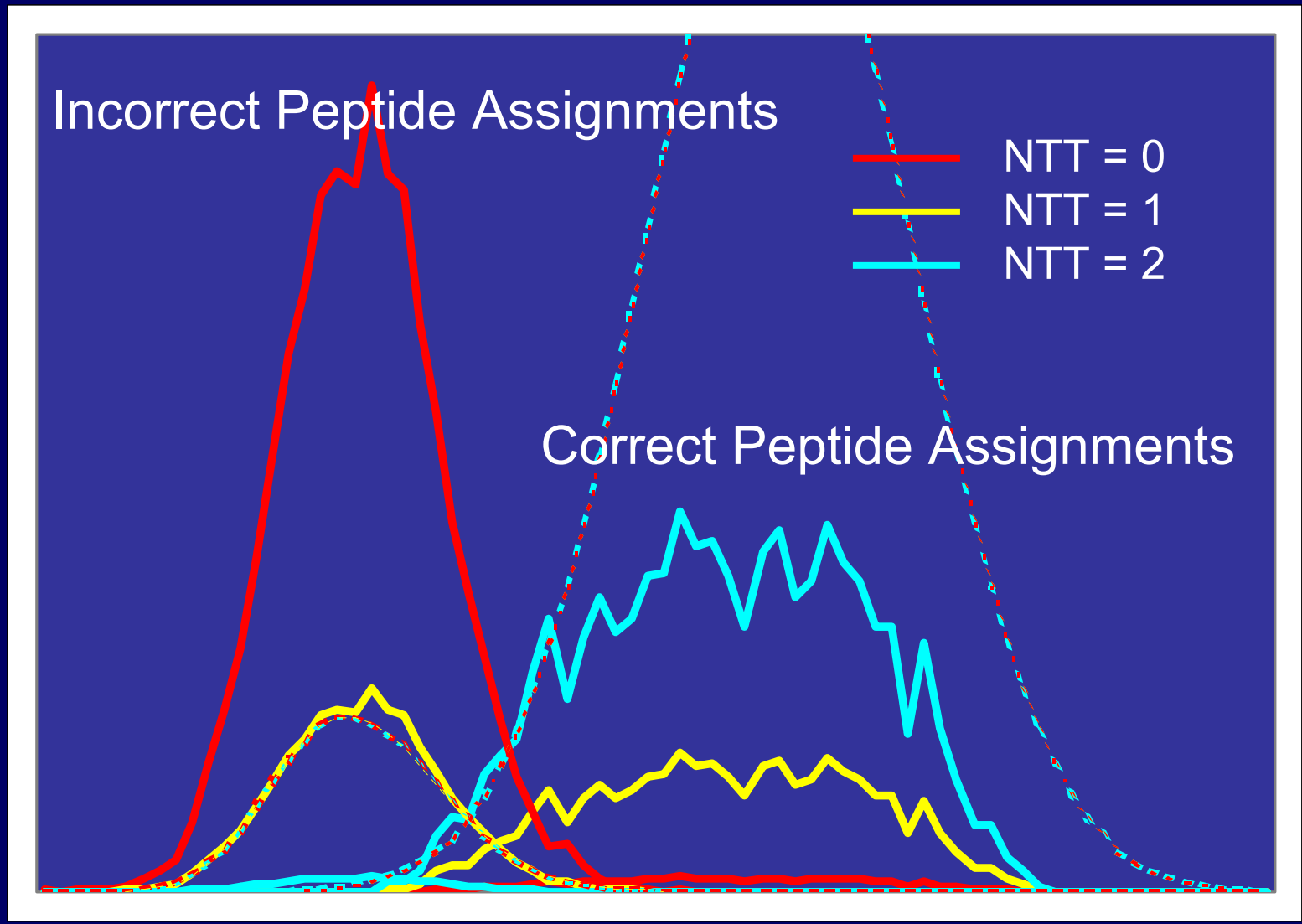
EM algorithm: unsupervised learning method that ***iteratively*** estimates the distributions given probabilities that each search result is correct, and then computes those probabilities given the distributions

Initial settings help guide algorithm to good solution

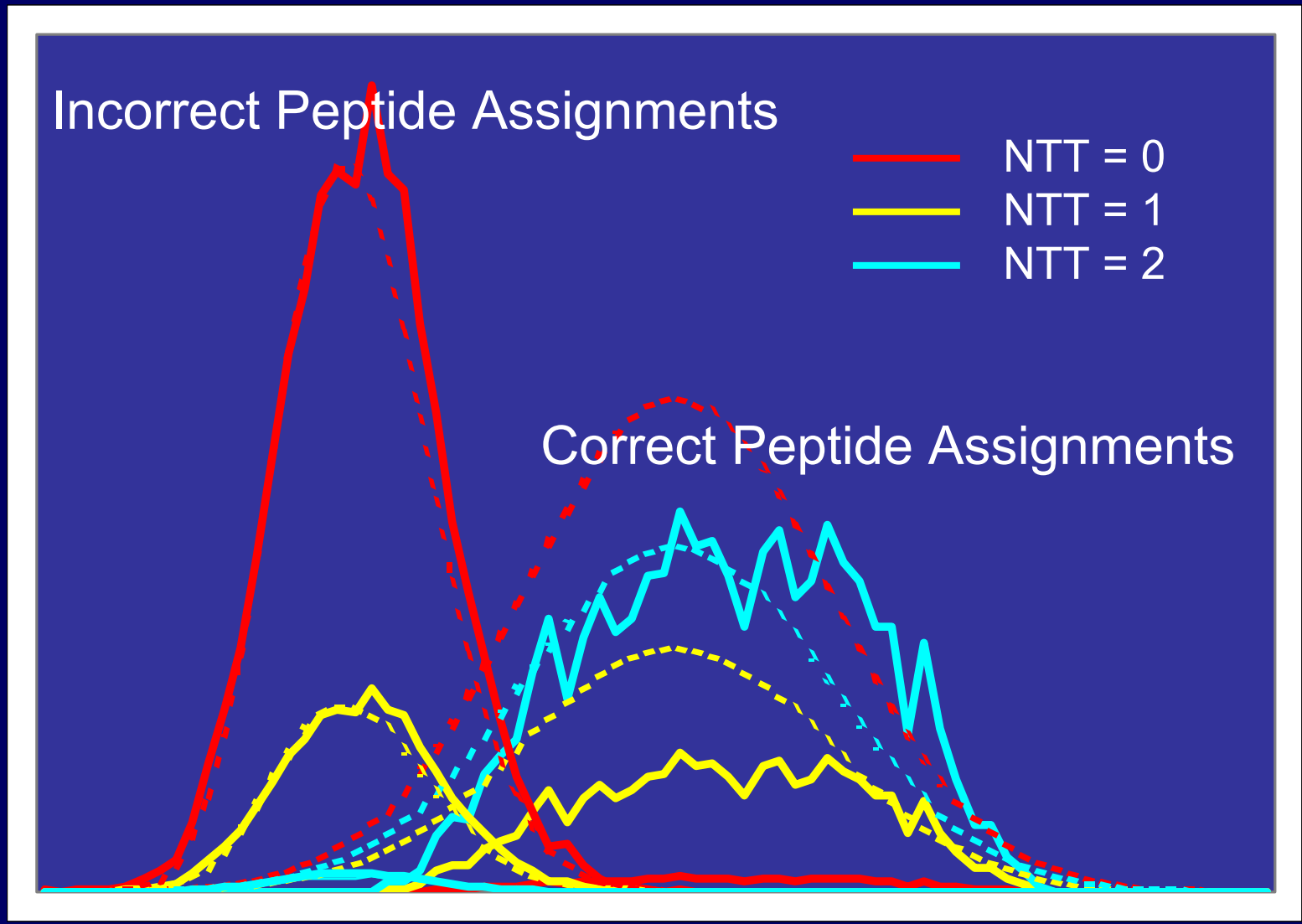
E-M Algorithm learns test data score distributions



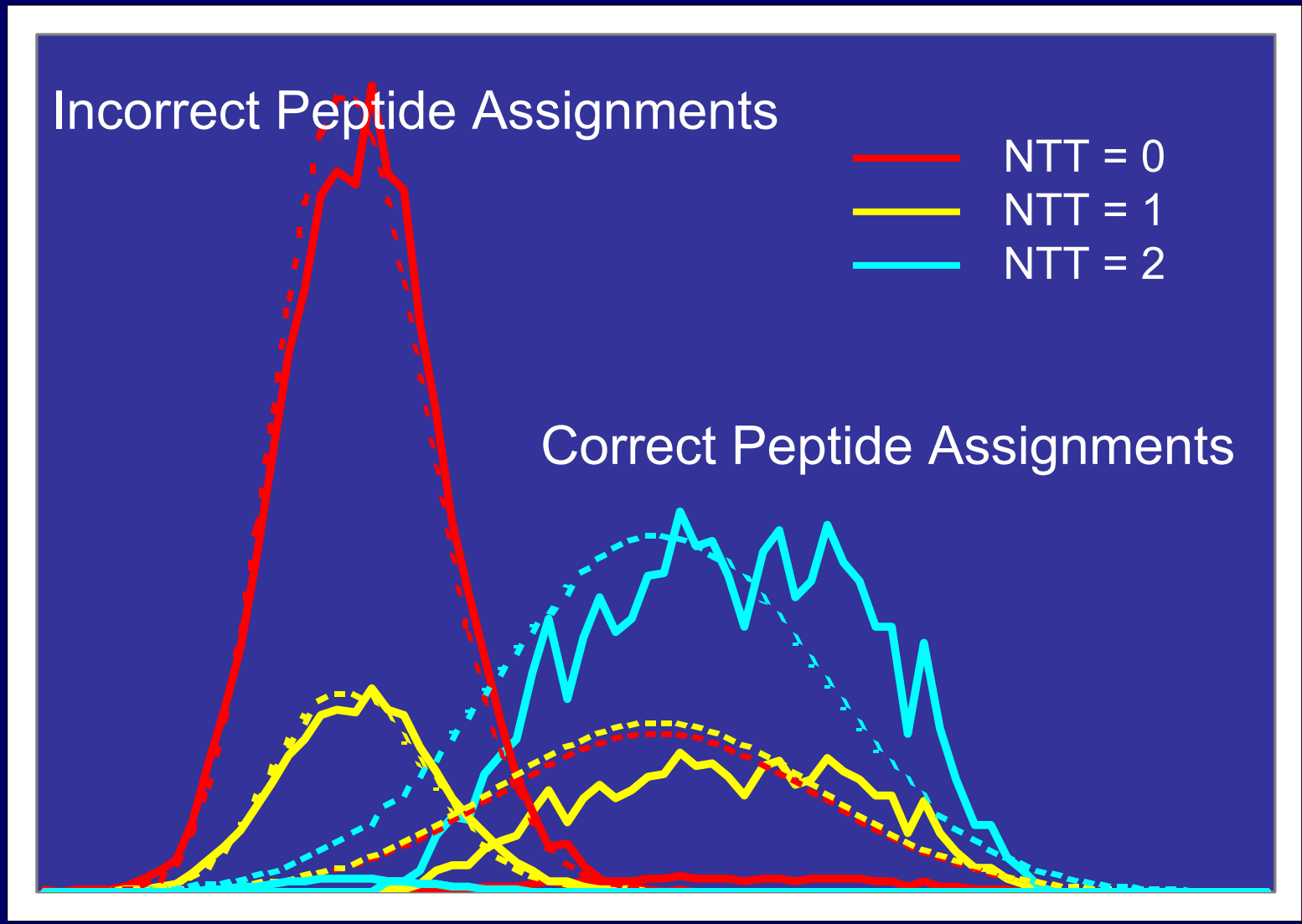
E-M Iteration 0



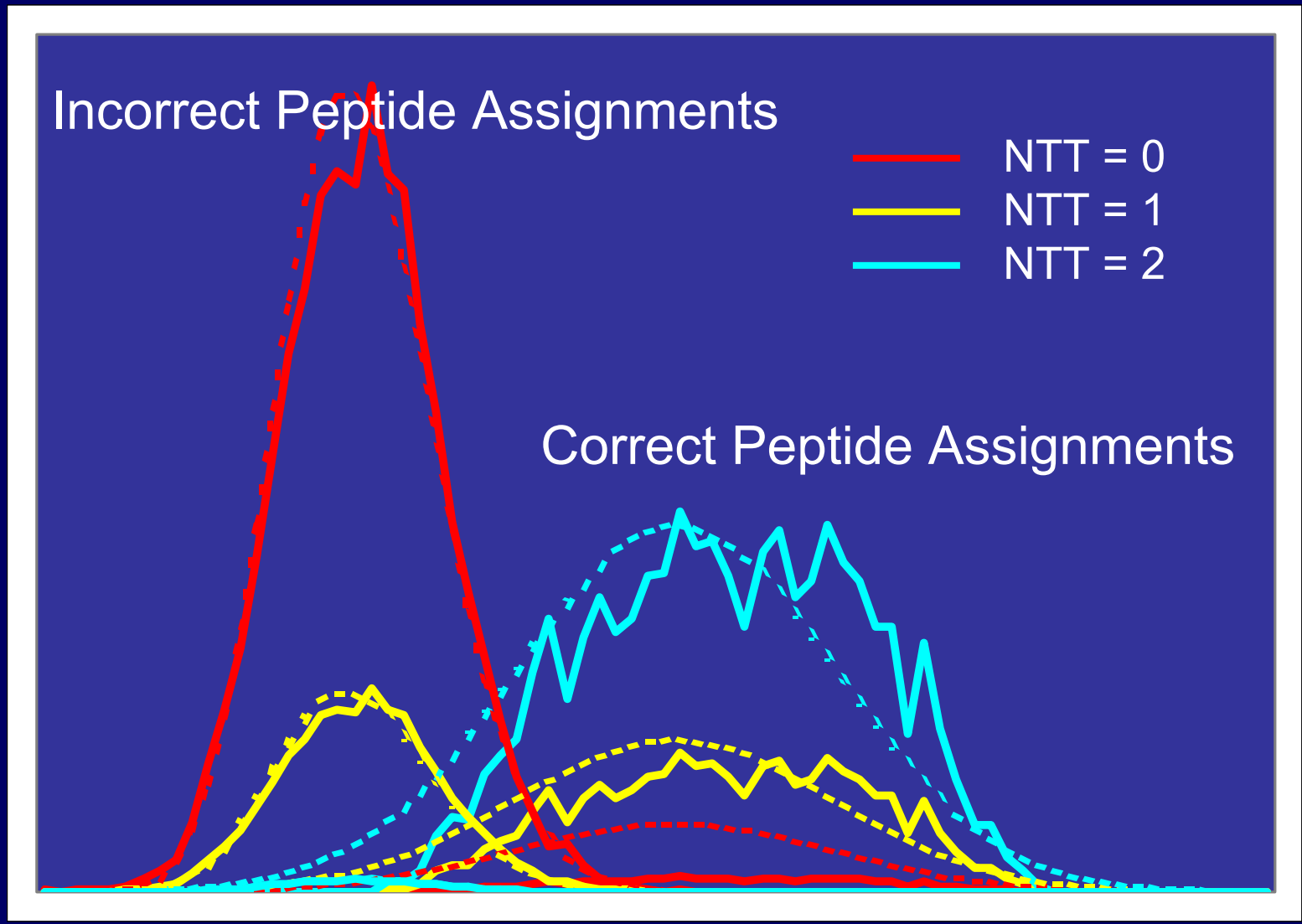
E-M Iteration 1



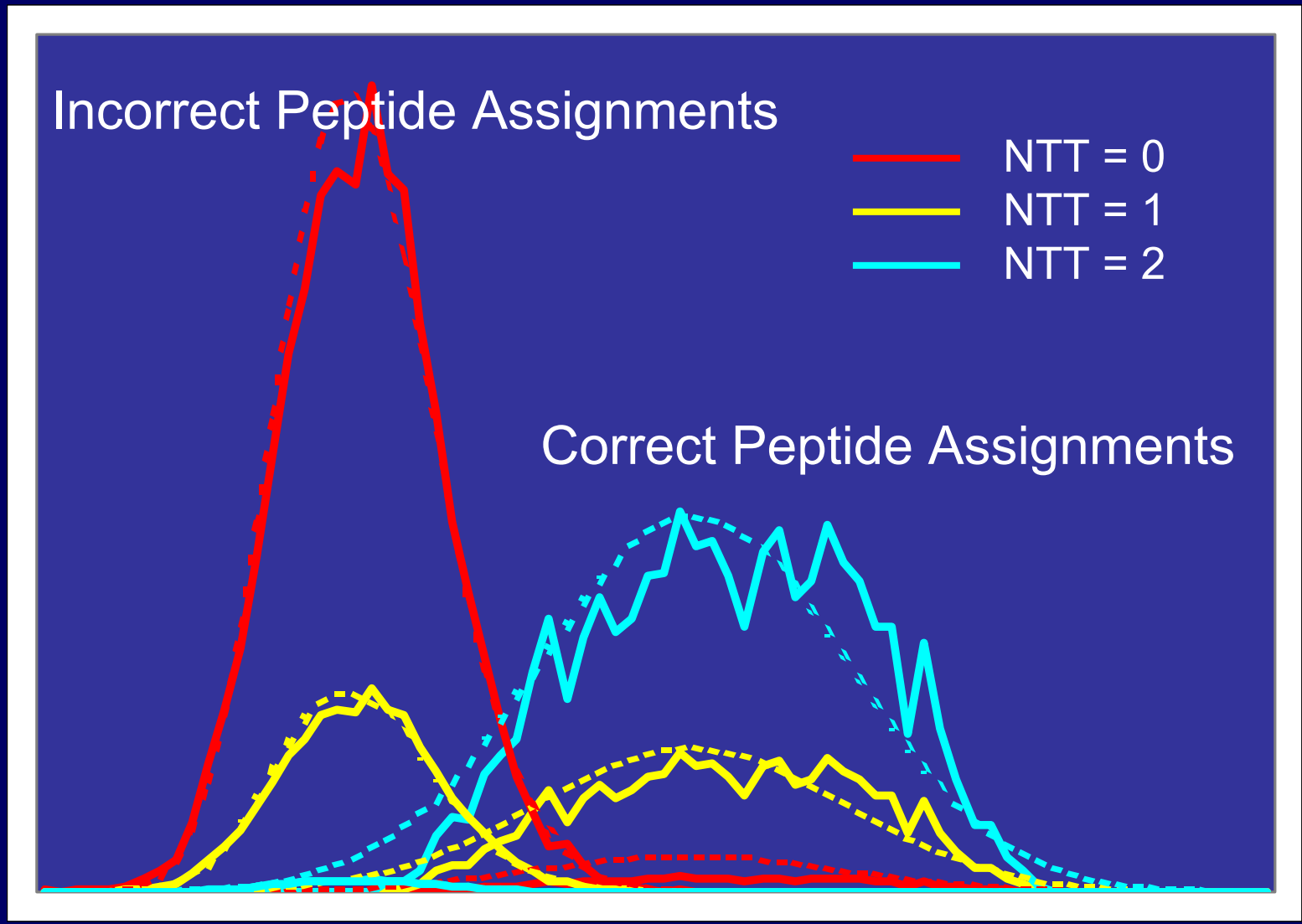
E-M Iteration 2



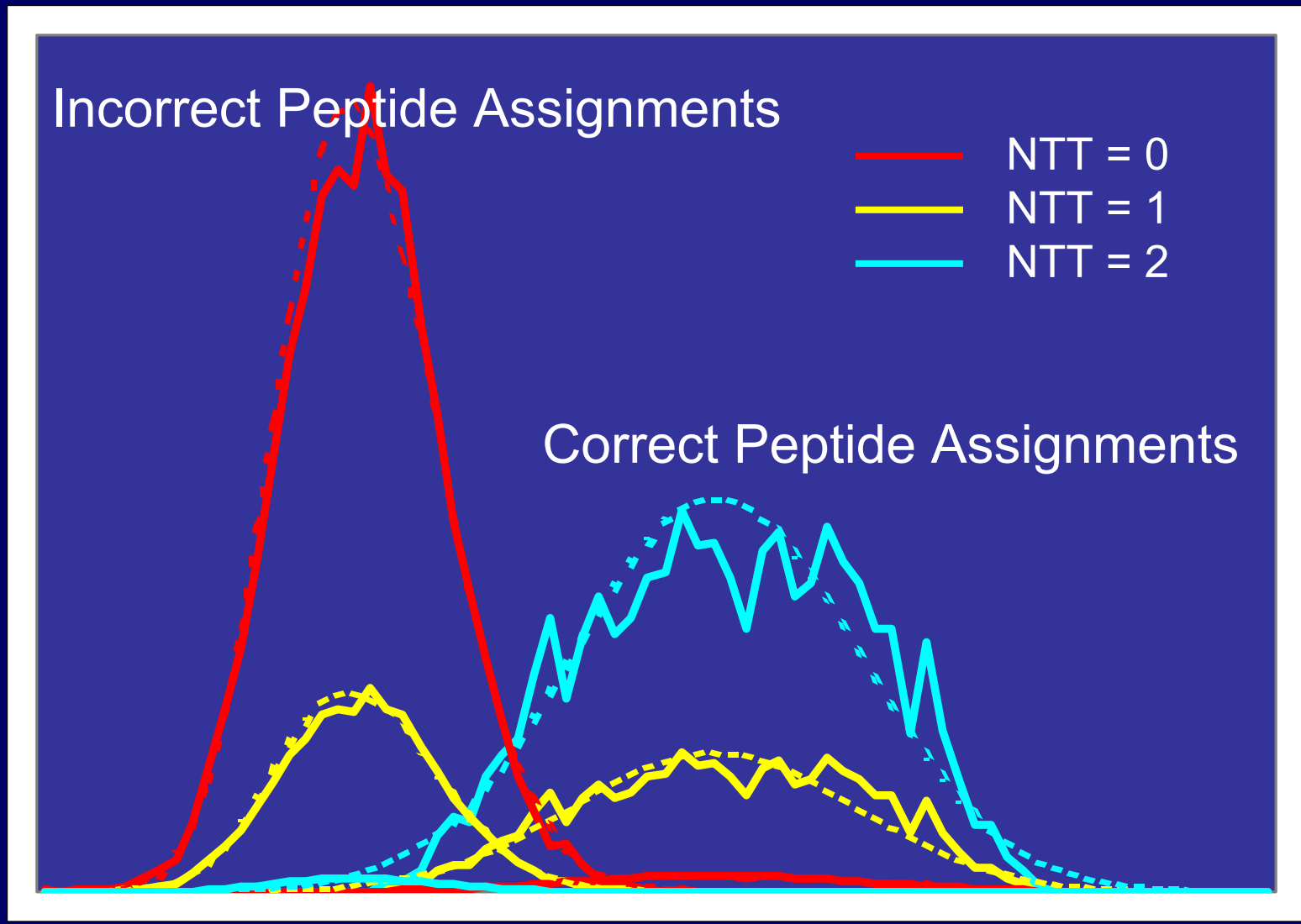
E-M Iteration 3



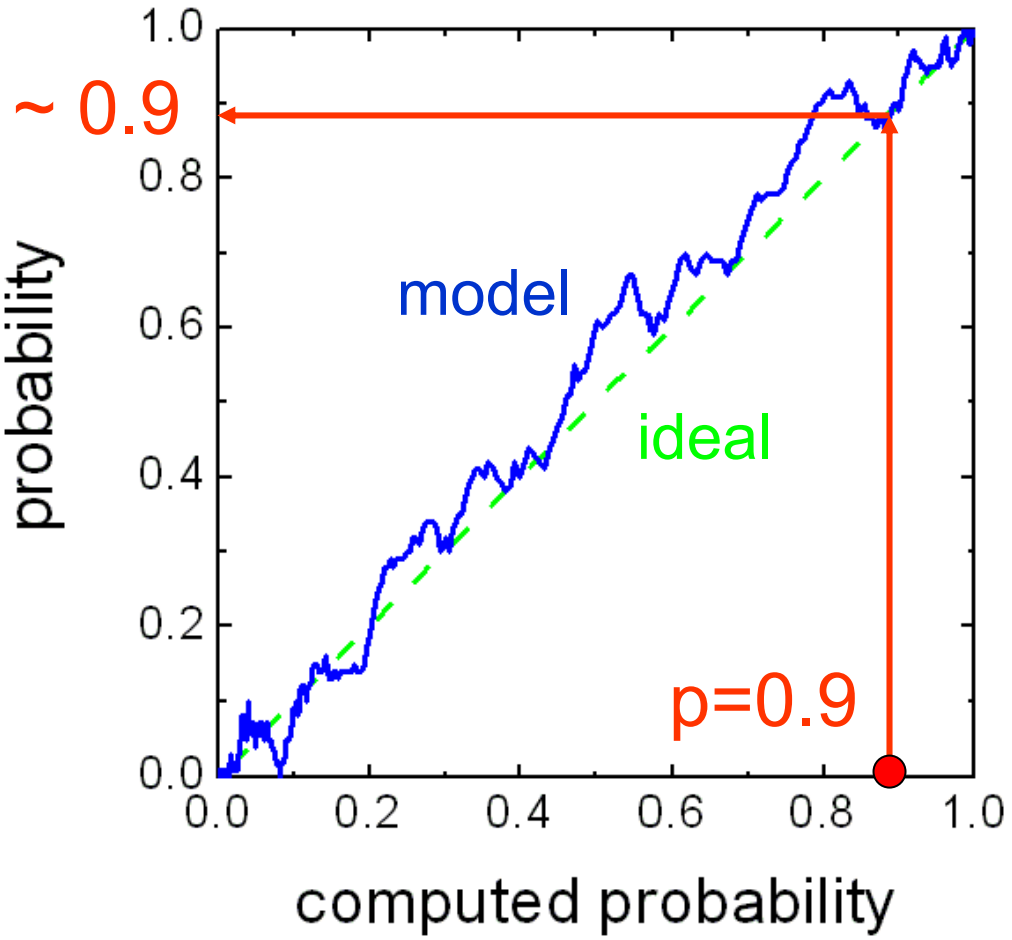
E-M Iteration 4



E-M Iteration 7



Accuracy of the Model



test data: A. Keller *et al*. OMICS 6, 207 (2002)

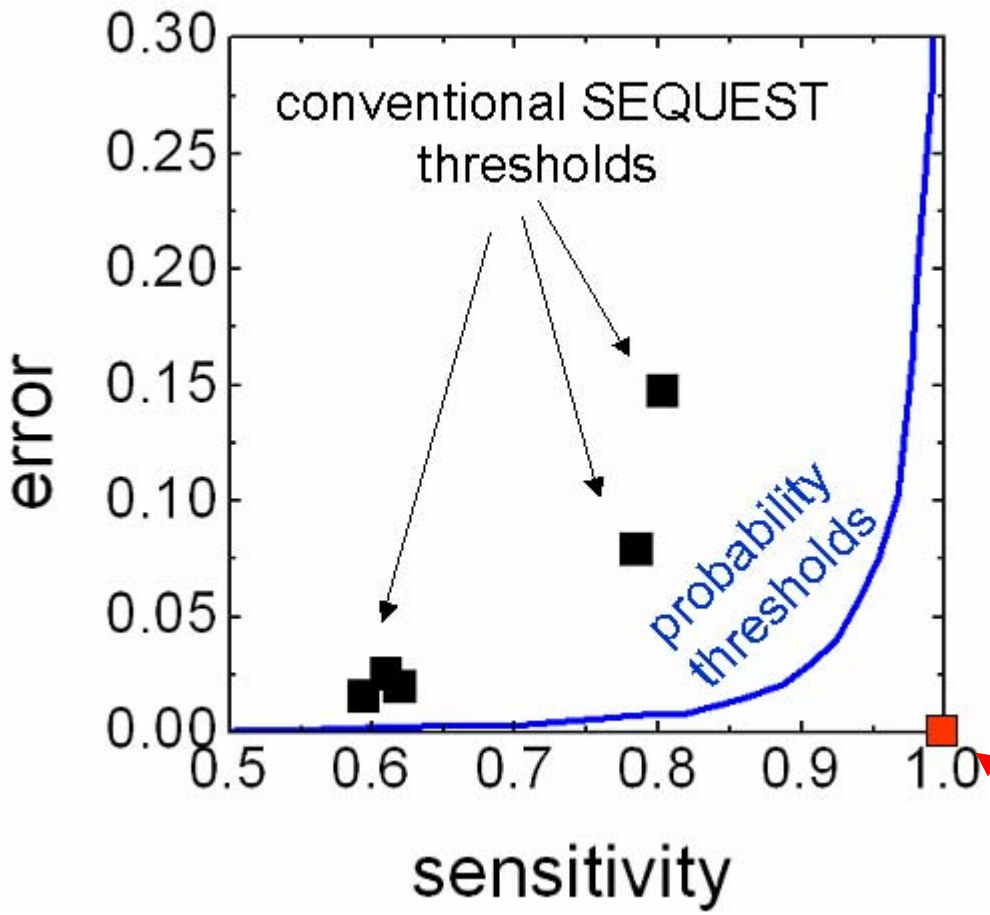
100 spectra with computed p ~ 0.9

90% of them (90) should be correct

observed probability is around 0.9

Model is accurate

Discriminating Power of Computed Probabilities



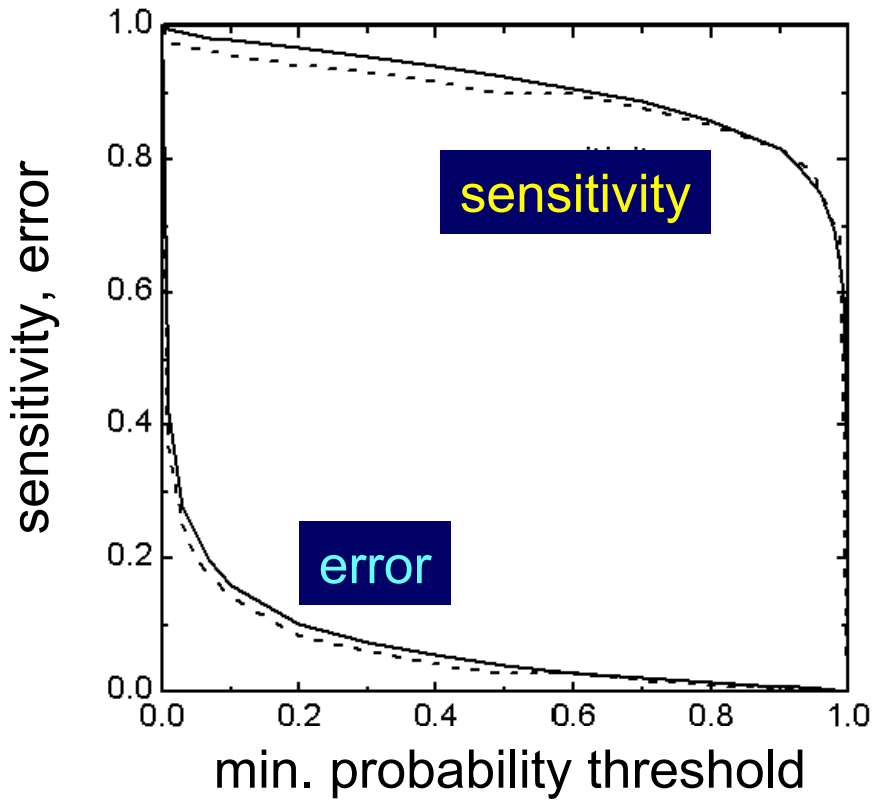
test data: A. Keller *et al*. OMICS 6, 207 (2002)

Sensitivity: fraction of all correct results passing filter

Error: fraction of all results passing filter that are incorrect

Ideal Spot

Discriminating Power of Computed Probabilities

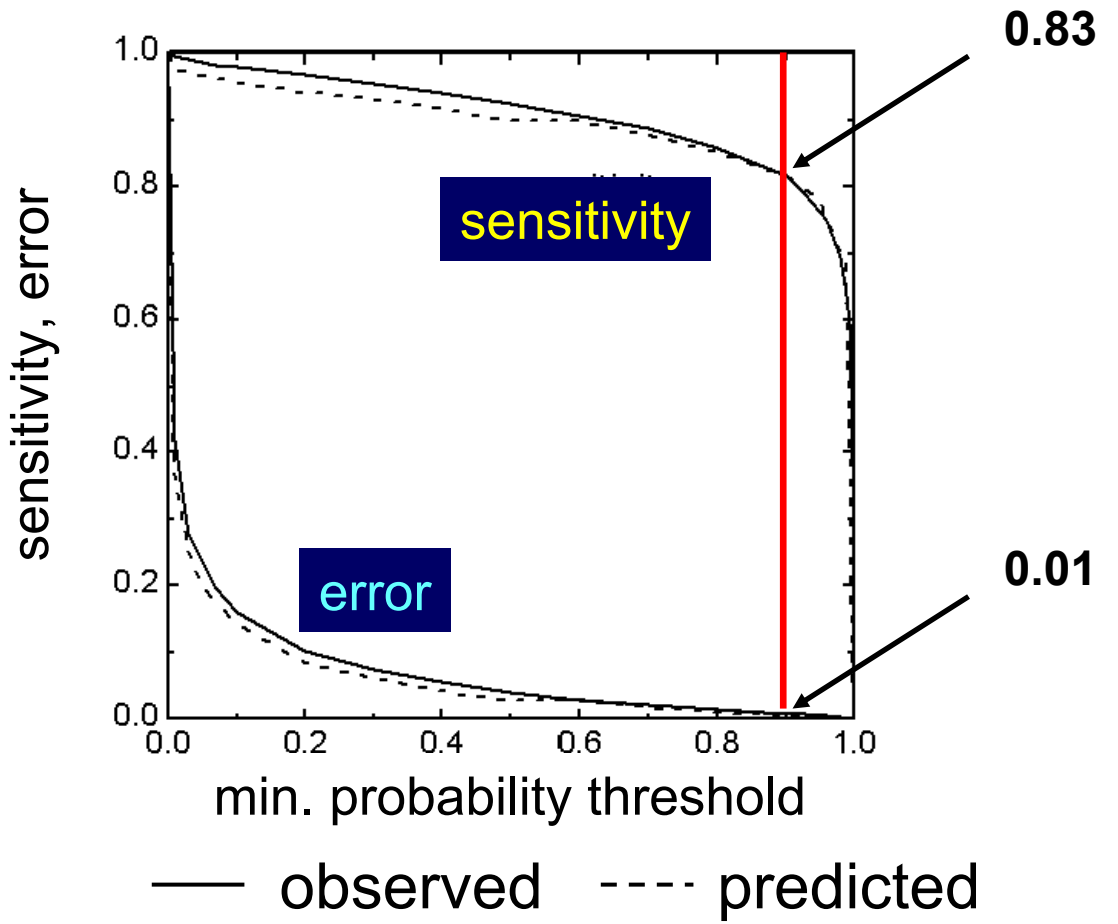


test data: A. Keller *et al*. OMICS 6, 207 (2002)

Sensitivity: fraction of all correct results passing filter

Error: fraction of all results passing filter that are incorrect

Discriminating Power : Example p ≥ 0.9

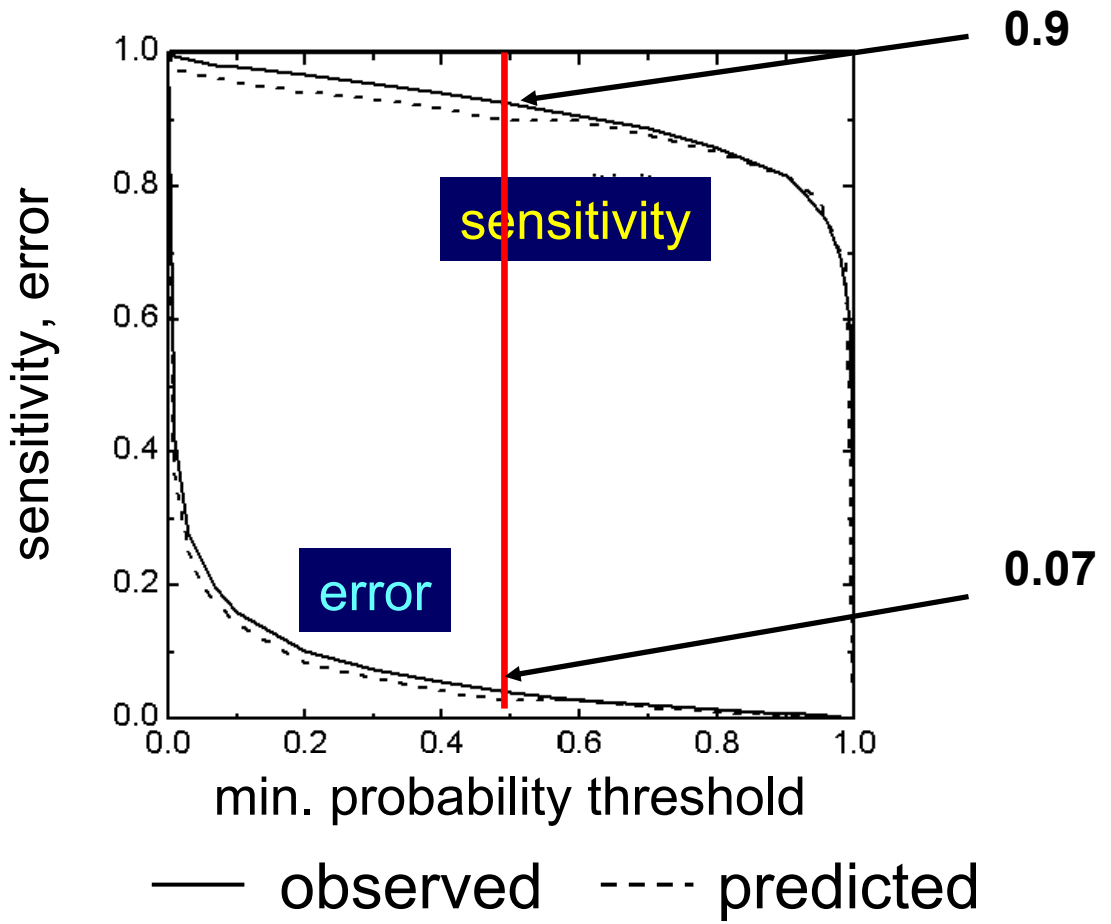


test data: A. Keller *et al*. OMICS 6, 207 (2002)

Sensitivity: fraction of all correct results passing filter

Error: fraction of all results passing filter that are incorrect

Discriminating Power : Example p ≥ 0.5



test data: A. Keller *et al*. OMICS 6, 207 (2002)

Sensitivity: fraction of all correct results passing filter

Error: fraction of all results passing filter that are incorrect

Can experts discriminate better than model?

20 of the test spectra were assigned probabilities close to 0.5 (in complete test dataset)

Of those, 10 were correct:

on average:

51% ‘publishable’, 24% ‘borderline’, 25% ‘not pub’

and 10 were incorrect:

on average:

11% ‘publishable’, 16% ‘borderline’, 74% ‘not pub’

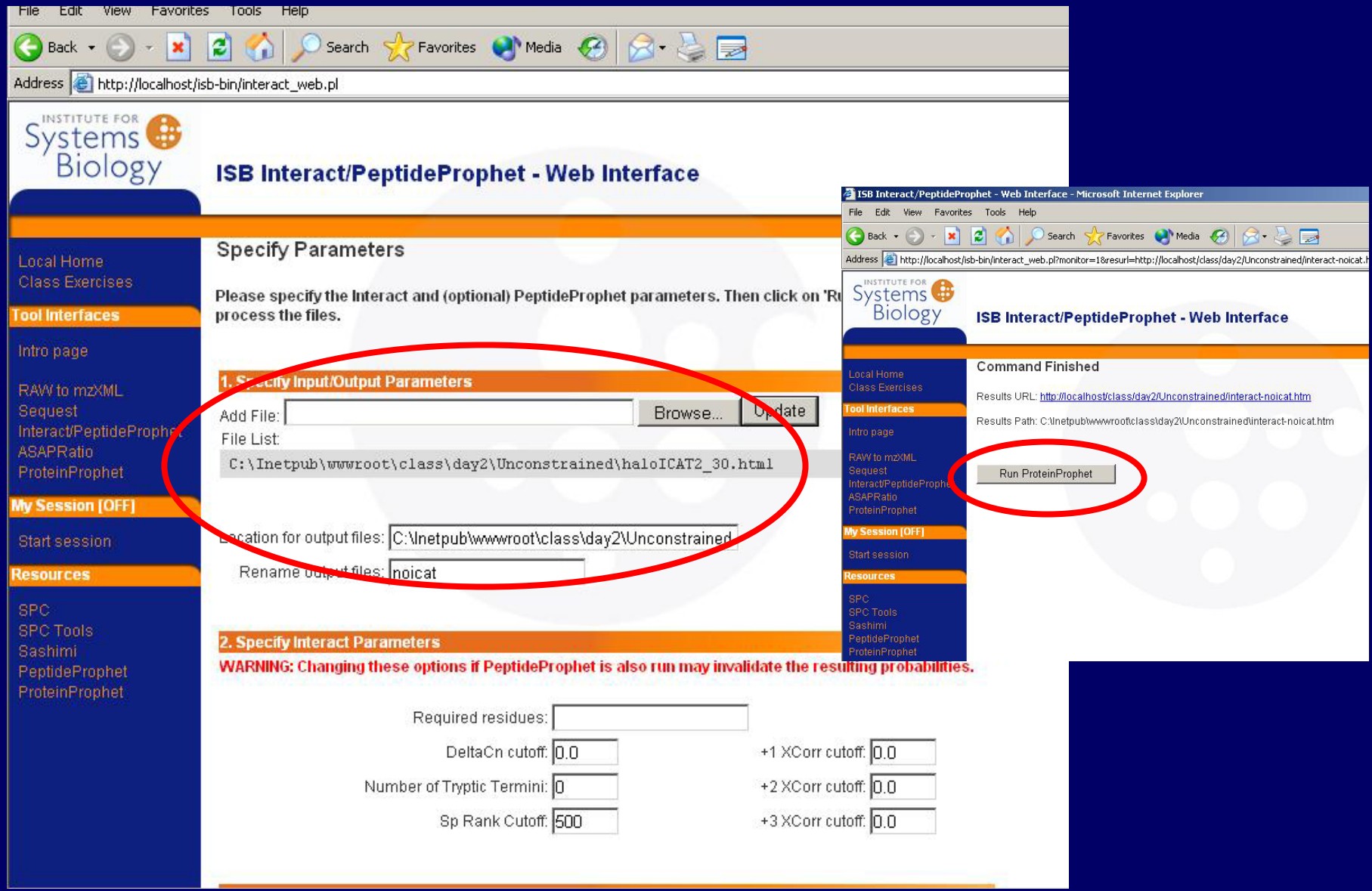
Getting started with PeptideProphet

Input: SEQUEST summary html files (`file.html`)

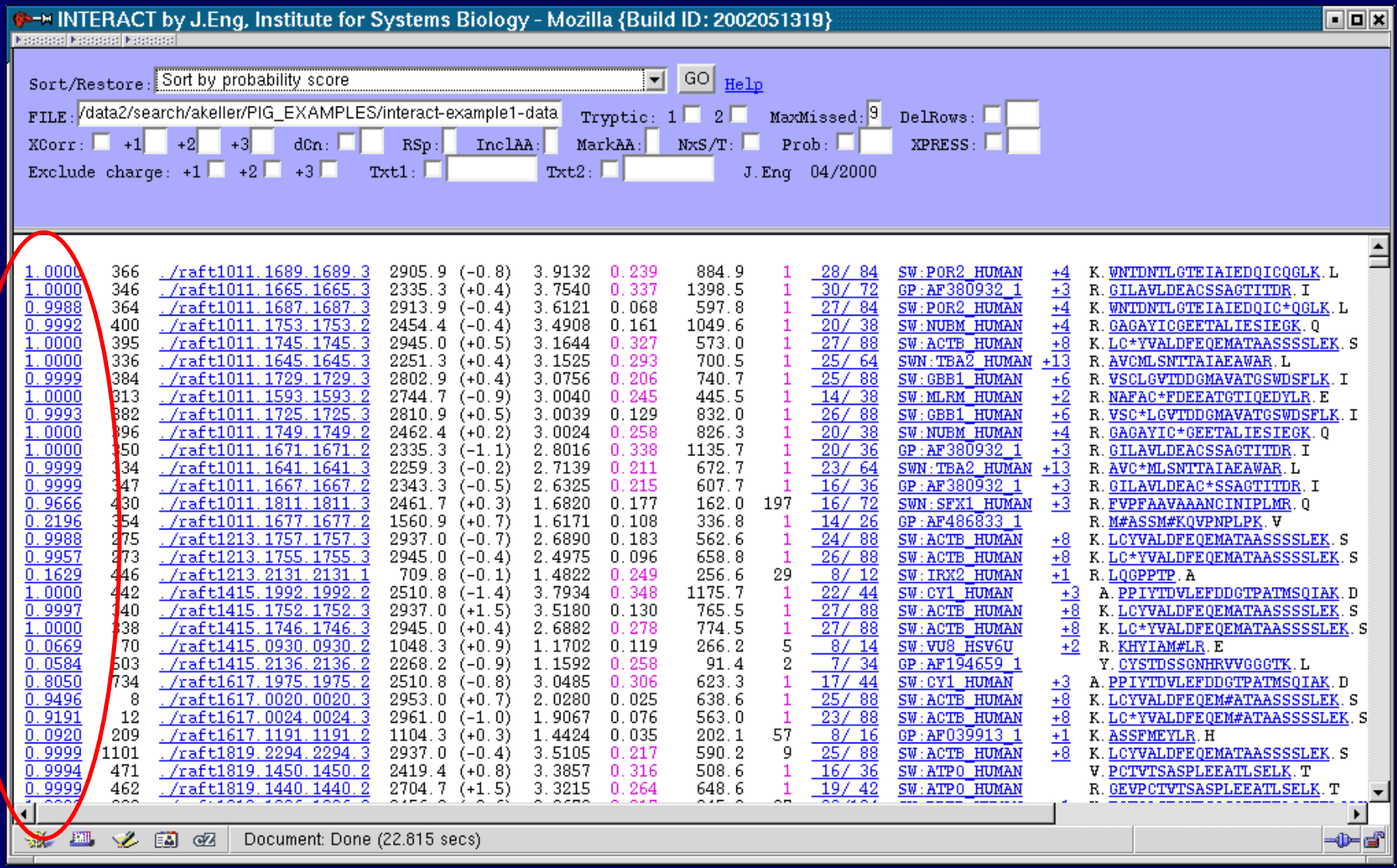
Interact merges files together into `interact.htm`, then PeptideProphet runs model, computes probabilities, and writes probabilities as first column

Combine together runs that are similar (sample, database, search constraints, mass spectrometer)

Getting started with PeptideProphet



PeptideProphet Results



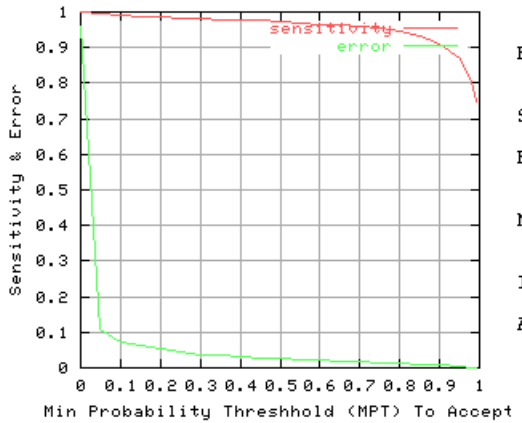
PeptideProphet Results: Model Summary

PLOTMODEL v.1.2, by J.Eng © Institute for Systems Biology, 2002. All rights reserved. - Mozilla {Build ID: 2002051319}

File Edit View Search Go Bookmarks Tasks Help

Back Forward Reload Stop http://regis-front/cgi-bin/plotmodel.cgi?ModelFile=/data2/search/akeller/PIG_EXAMPLE: Search Print

Home Bookmarks Red Hat Network Training Support Software Hardware Developers Embedded Search Documentation Download



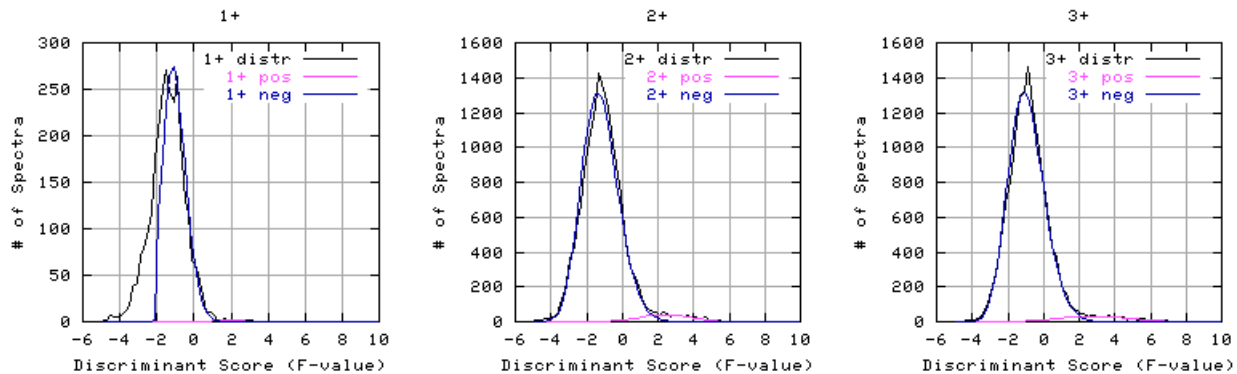
Estimated total number of correct peptide assignments in dataset: 1546

Sensitivity (Red Line): fraction of all correct assignments (1546 total) passing MPT filter

Error (Green Line): fraction of peptide assignments passing MPT filter that are incorrect

MPT = Minimum Probability Threshhold to Accept

If you need an explanation, please contact Alexey Nesvizhskii (nesvi@systemsbiology.org) or Andy Keller (akeller@systemsbiology.org) for more information.



INTERAC

Sort/Restore

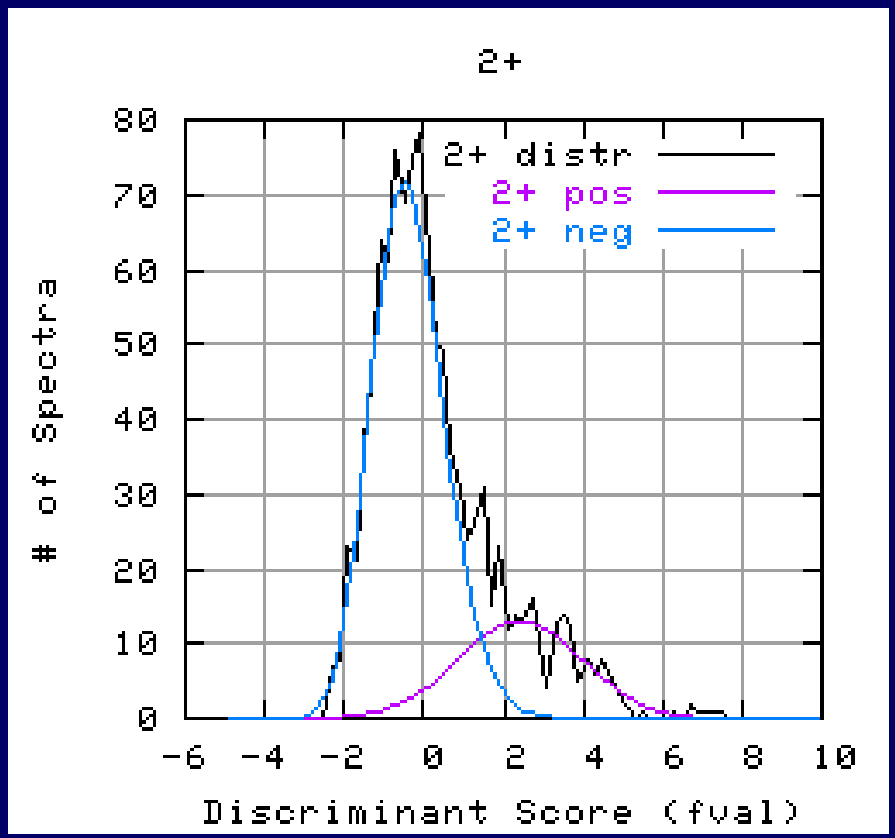
FILE: /data2/s

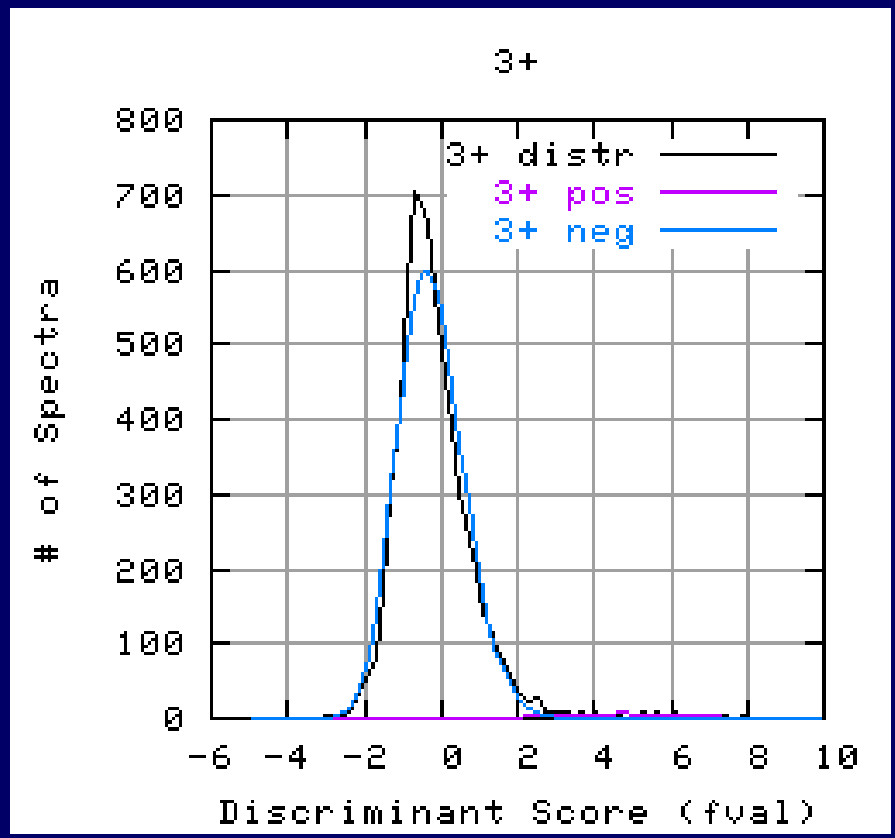
XCorr: +1

Exclude char

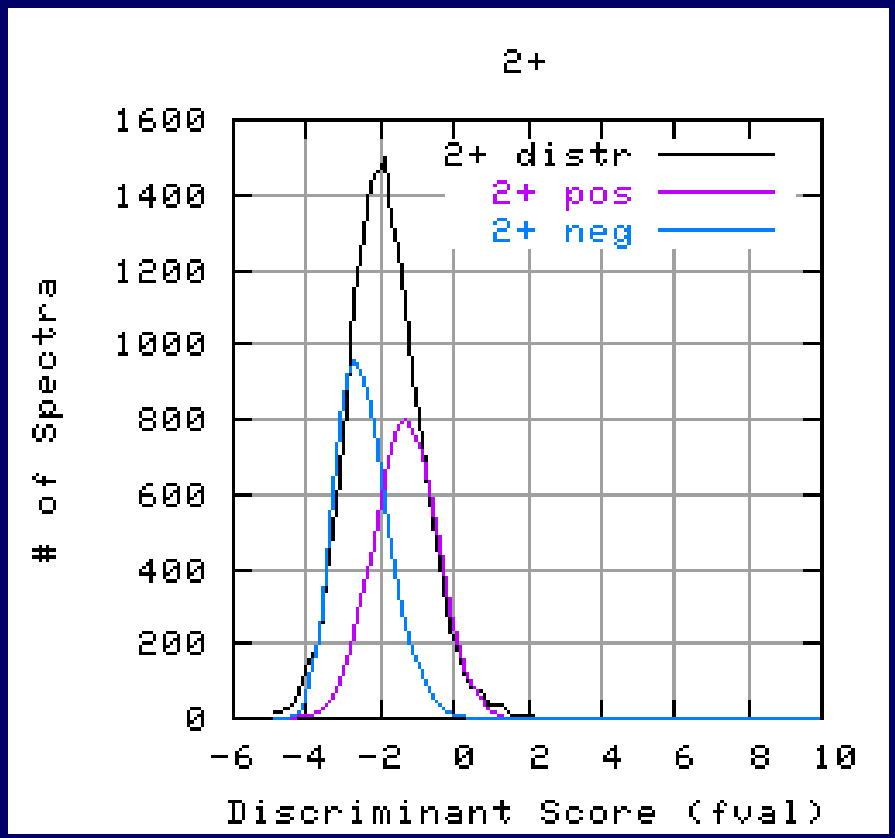
1.0000	366
1.0000	346
0.9988	364
0.9992	400
1.0000	395
1.0000	336
0.9999	384
1.0000	313
0.9993	382
1.0000	396
1.0000	350
0.9999	334
0.9999	347
0.9666	430
0.8196	354
0.9988	275
0.9957	273
0.1629	446
1.0000	442
0.9997	340
1.0000	338
0.0669	70
0.0584	503
0.8050	734
0.9496	8
0.9191	12
0.0920	209

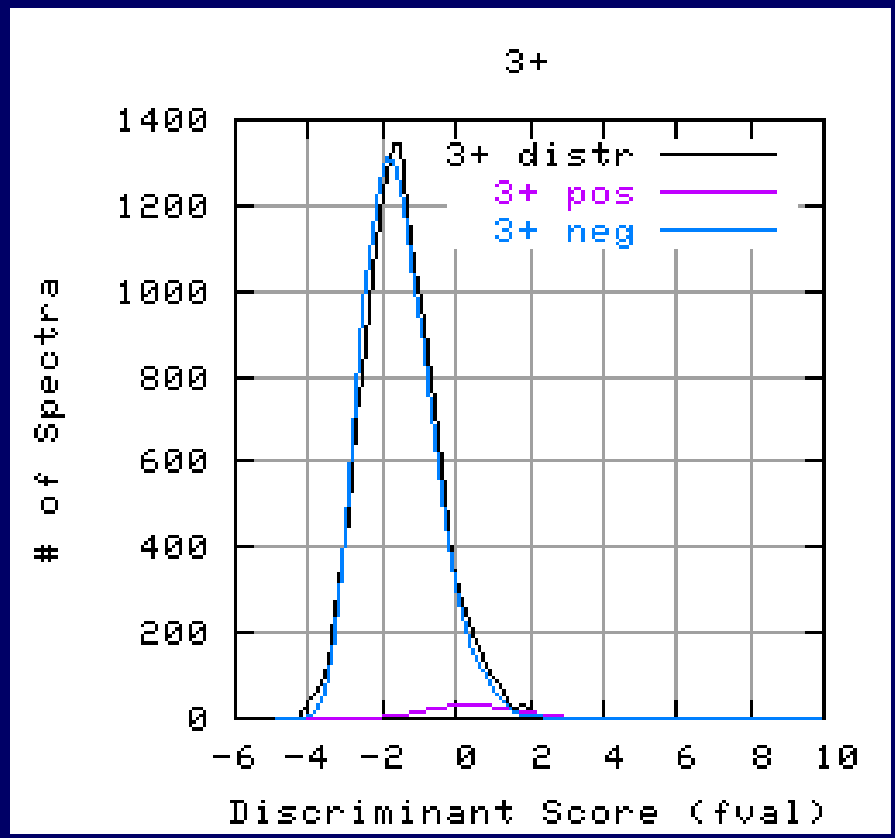
Reasonable Learned Discriminant Score Distributions



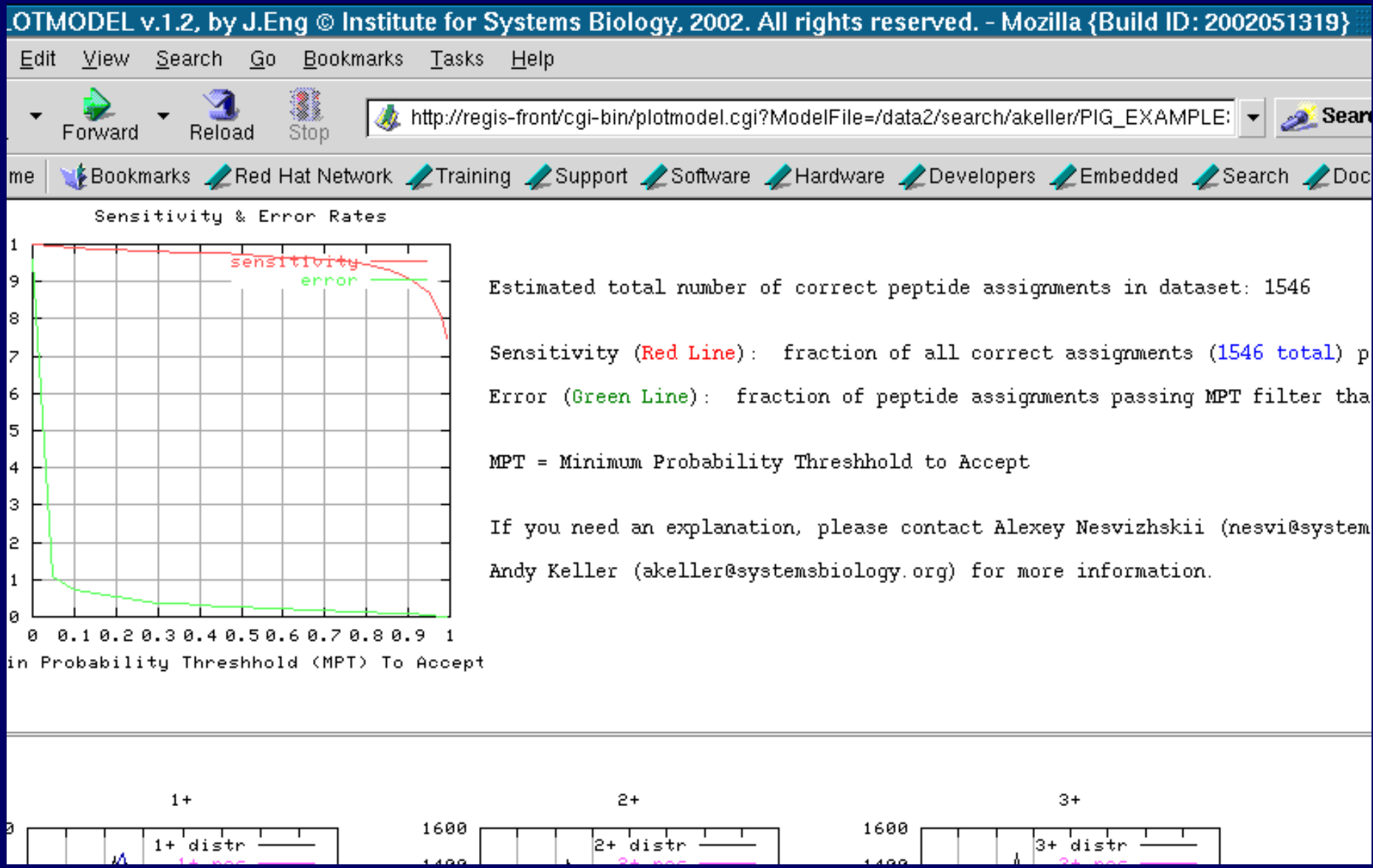


Suspicious Looking Learned Discriminant Score Distributions

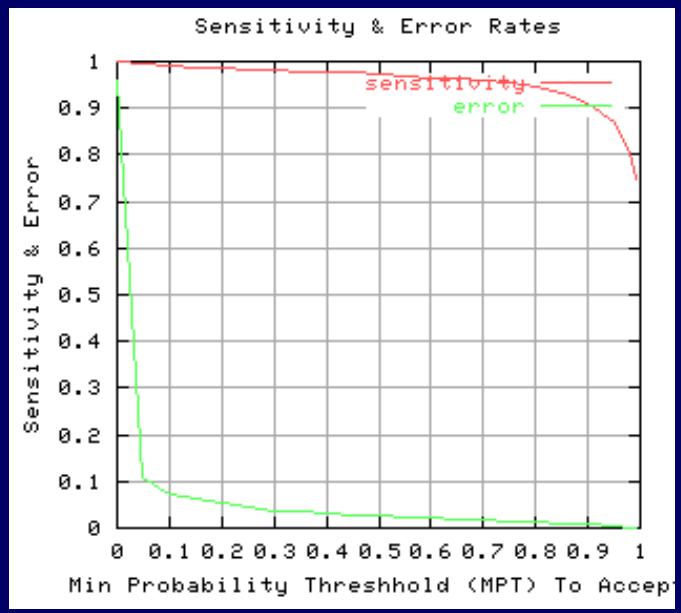




PeptideProphet Results: Model Summary

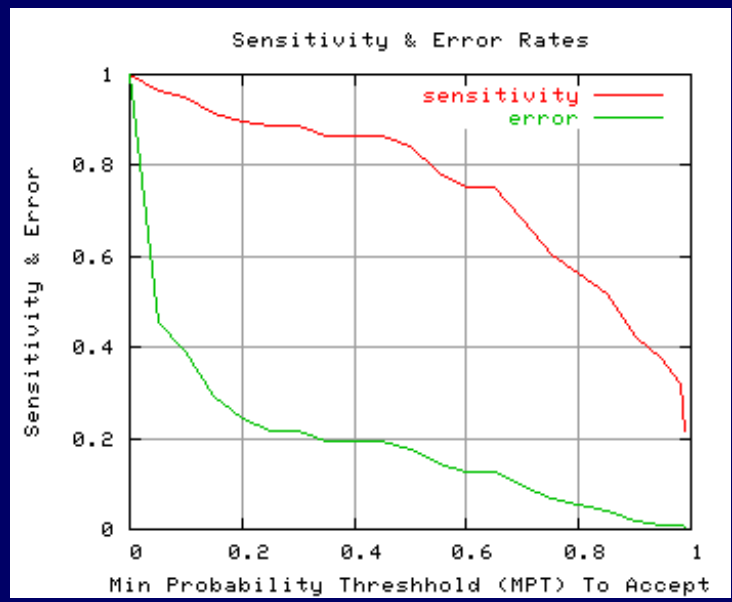


PeptideProphet Results: Model Summary



Good

Not so good



PeptideProphet Results: Model Summary

PLOTMODEL v.1.2, by J.Eng © Institute for Systems Biolo

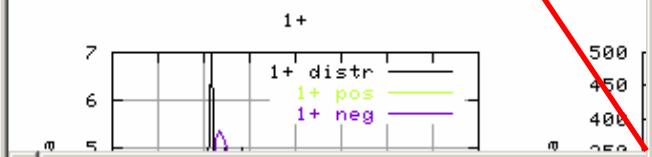
File Edit View Favorites Tools Help

Back Search Favorites

Address http://regis.systemsbiology.net/cgi-bin/plotmodel.cgi?M

PeptideProphet results

Click here to view more details of the ar



Done

Internet

http://regis.systemsbiology.net/data2/search/akeller/TEST/interact-andy-data.model.txt - Microsoft Inte...

File Edit View Favorites Tools Help

Back Search Favorites

Address http://regis.systemsbiology.net/data2/search/akeller/TEST/interact-andy-data.model.txt

Go Links

```
Tue Dec 30 17:03:50 2003

interact -Oix -Nandy glycprot2.html raft20.html

OPTIONS:  ICAT Exclude_DeltaCn*_Entries


PeptideProphet v. 1.0  Andrew Keller  ISB


FINAL 1+ MODEL after 2 iterations:
number of spectra: 35
using training data negative distributions
        prior: 0.000, total: 0.0
discrim score
        pos: (gaussian mean: 0.54, stdev: nan)
        neg: (gamma m1: 1.16, m2: 1.76, alpha: 0.35, beta: 3.30, zero: -2.00)
        negmean: 0.49
no. tryptic term.
        pos: (0.878 ntt=0, 0.122 ntt=1, 0.000 ntt=2)
        neg: (0.886 ntt=0, 0.114 ntt=1, 0.000 ntt=2)
no. missed cl
        pos: (0.972 nmc=0, 0.028 1<=nmc<=2, 0.000 nmc>=3)
        neg: (0.971 nmc=0, 0.029 1<=nmc<=2, 0.000 nmc>=3)
var offset mass diff (offset: 0.00)
        pos: (0.11 massd=-2.0, 0.26 massd=-1.0, 0.37 massd=0.0, 0.26 massd=1.0
             )
        neg: (0.11 massd=-2.0, 0.26 massd=-1.0, 0.37 massd=0.0, 0.26 massd=1.0
             )
icat cys
        pos: (1.000 icat incompatible, 0.000 icat compatible)
        neg: (1.000 icat incompatible, 0.000 icat compatible)
```

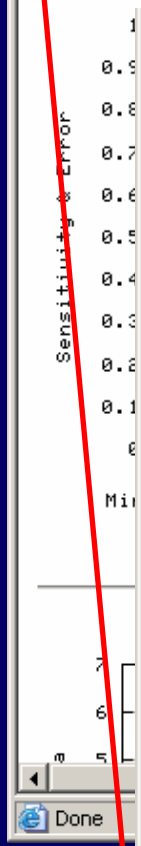
Done

Internet

PeptideProphet Results: Model Summary

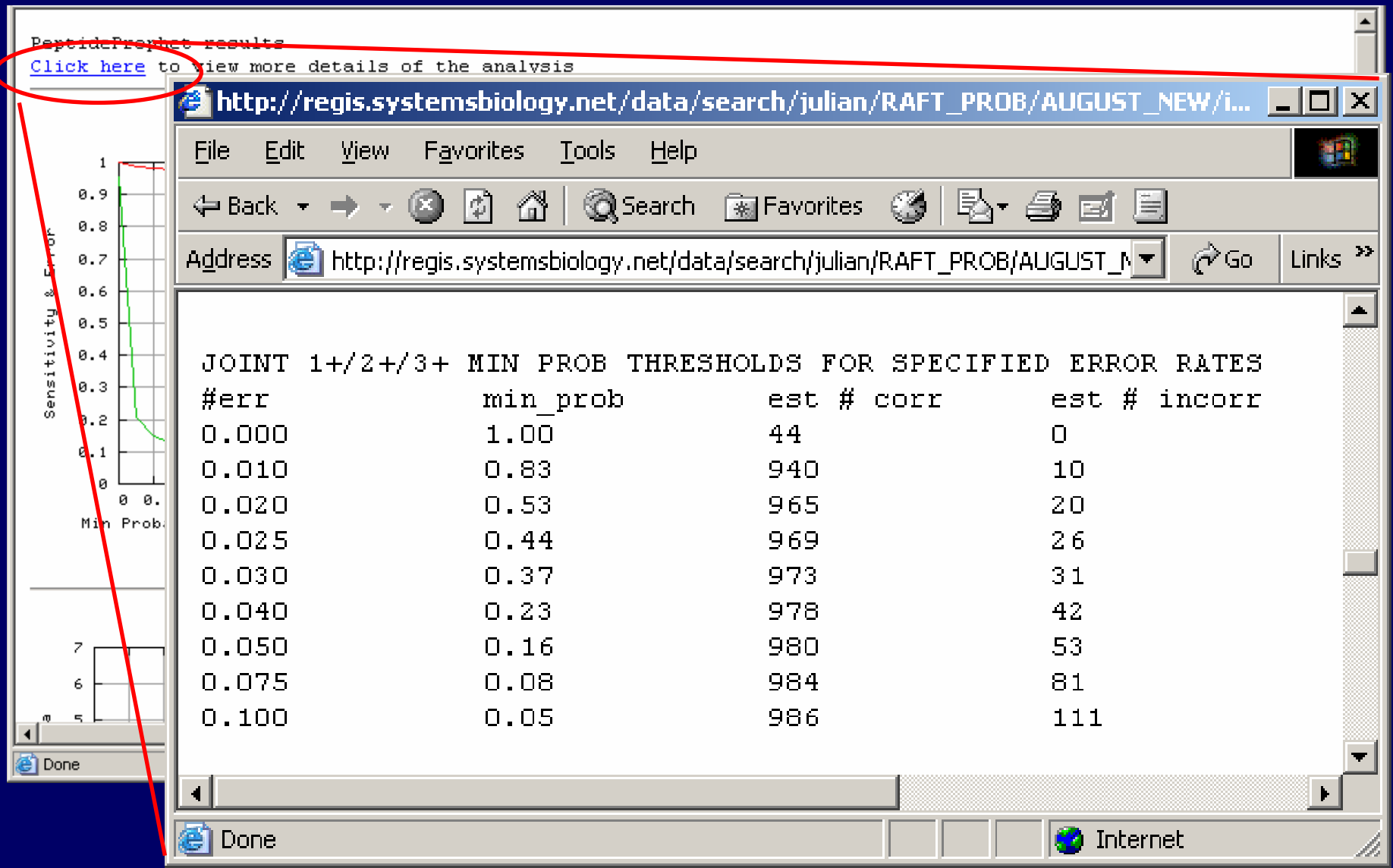
PeptideProphet results

Click here to view more details of the analysis

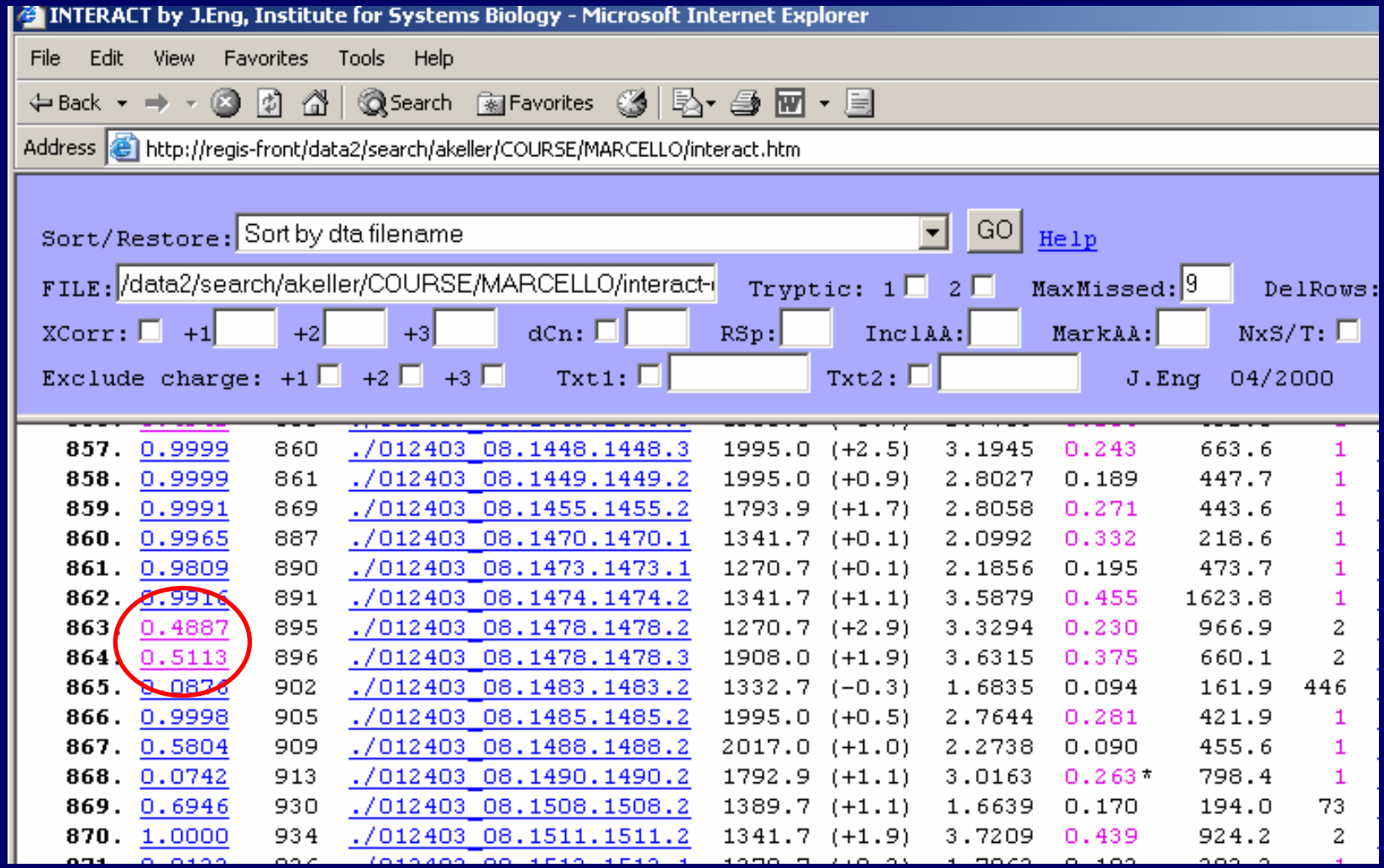



```
FINAL 2+ MODEL after 11 iterations:
number of spectra: 4518
using no. tolerable tryptic term. [ntt] 0 data as pseudonegatives
        prior: 0.057, est. total no. correct: 258.2
SEQUEST discrim score [fval]
        pos: (gaussian mean: 3.27, stdev: 1.36)
        neg: (gamma m1: 4.78, m2: 23.54, alpha: 0.14, beta: 35.31, zero: -5.17)
        negmean: -0.79
no. tolerable tryptic term. [ntt]
        pos: (0.011 ntt=0, 0.136 ntt=1, 0.853 ntt=2)
        neg: (0.752 ntt=0, 0.229 ntt=1, 0.019 ntt=2)
no. missed enz. cleavages [nmc]
        pos: (0.775 nmc=0, 0.224 1<=nmc<=2, 0.001 nmc>=3)
        neg: (0.415 nmc=0, 0.486 1<=nmc<=2, 0.099 nmc>=3)
var offset mass diff [massd] (offset: -0.60)
        pos: (0.00 massd=-2.0, 0.04 massd=-1.0, 0.53 massd=0.0, 0.33 massd=1.0,
              0.05 massd=2.0, 0.04 massd=3.0, 0.00 massd=4.0)
        neg: (0.12 massd=-2.0, 0.13 massd=-1.0, 0.15 massd=0.0, 0.18 massd=1.0,
              0.20 massd=2.0, 0.22 massd=3.0, 0.01 massd=4.0)
icat cys [icat]
        pos: (0.022 icat=0 (incompatible), 0.978 icat=1 (compatible))
        neg: (0.927 icat=0 (incompatible), 0.073 icat=1 (compatible))
```

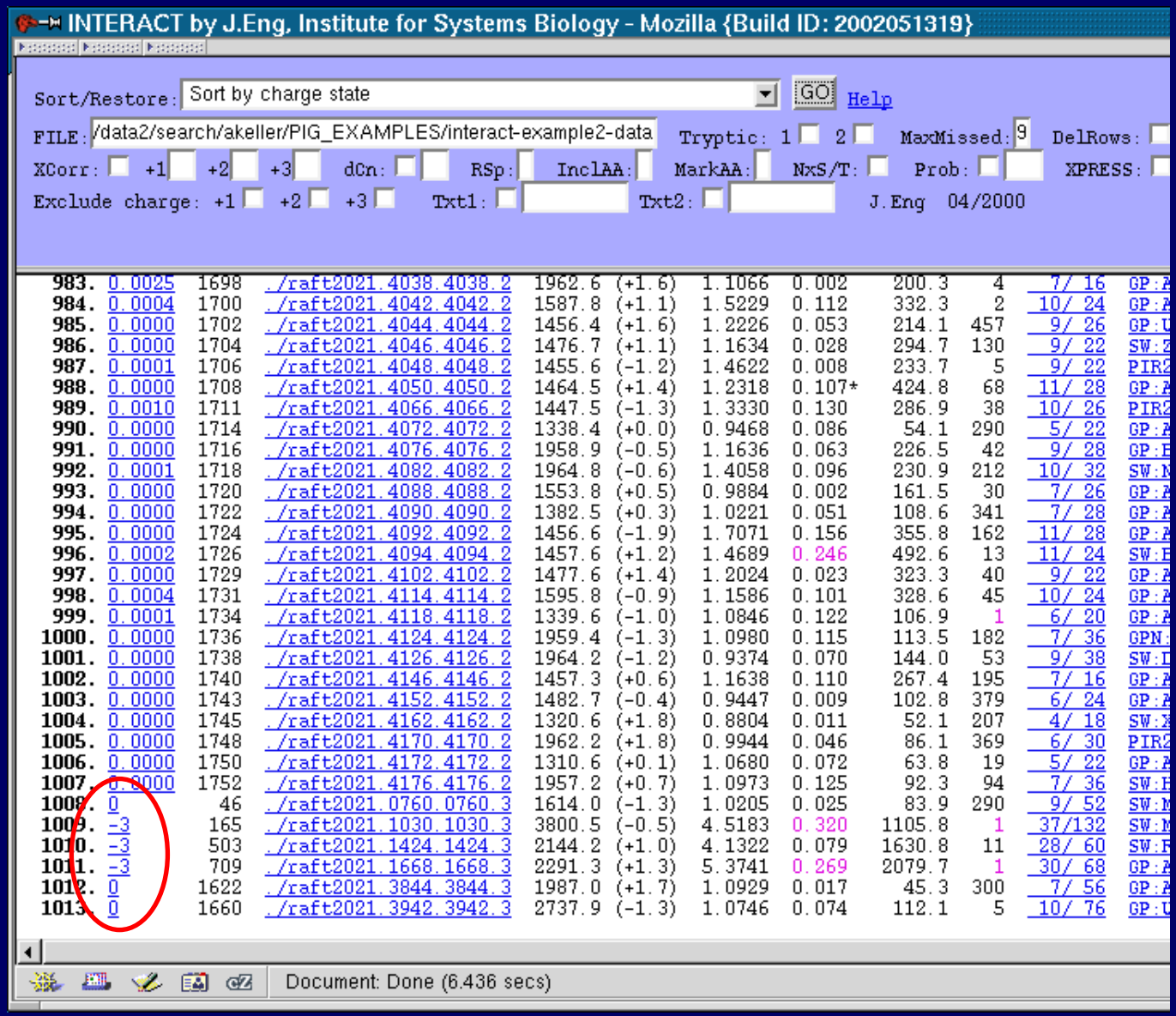
PeptideProphet Results: Predicted Numbers of Correct and Incorrect Peptides



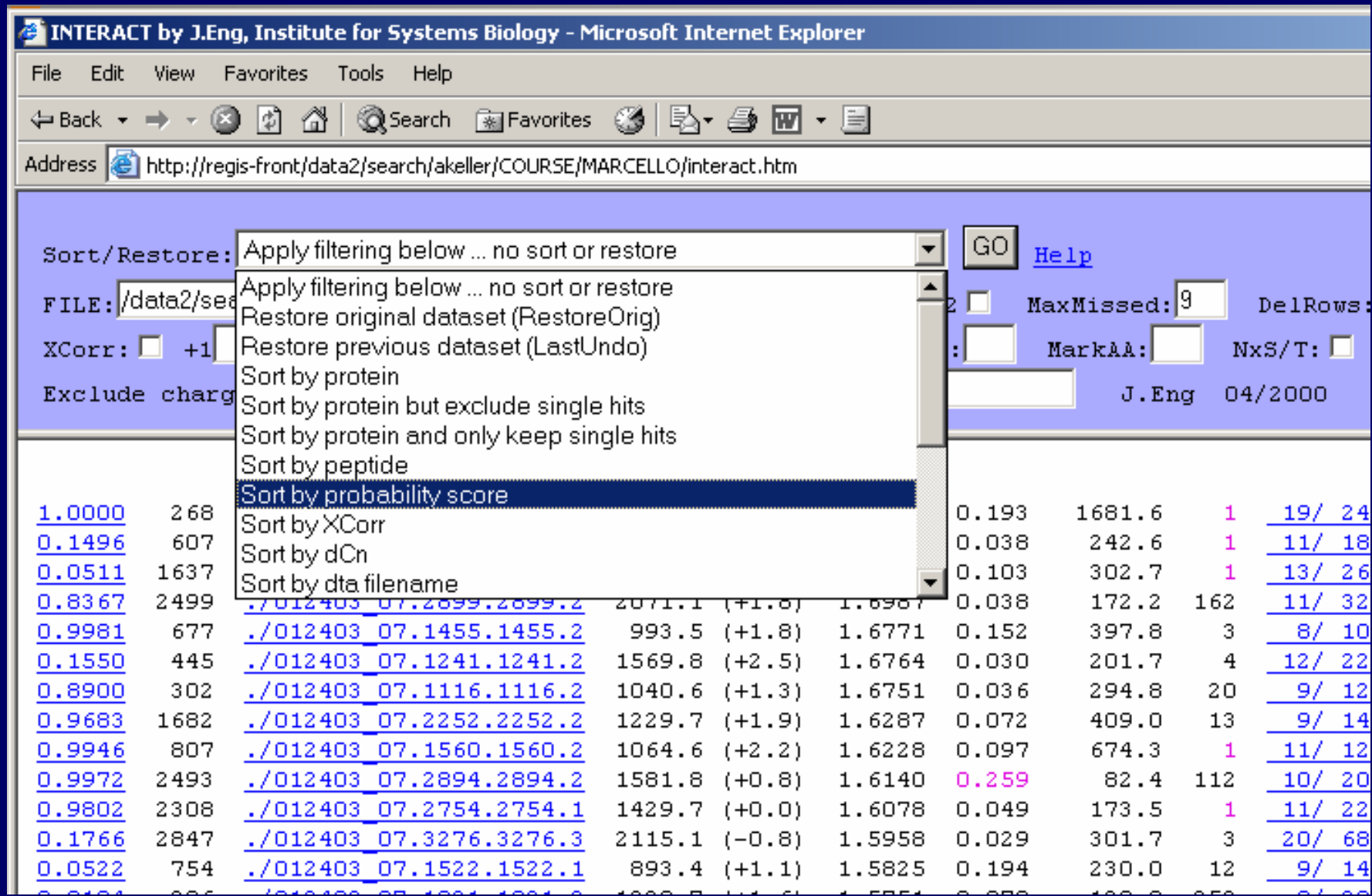
PeptideProphet $[M+2H]^{2+}$ vs $[M+3H]^{3+}$ Precursor Ions



PeptideProphet Results: Incomplete Analysis

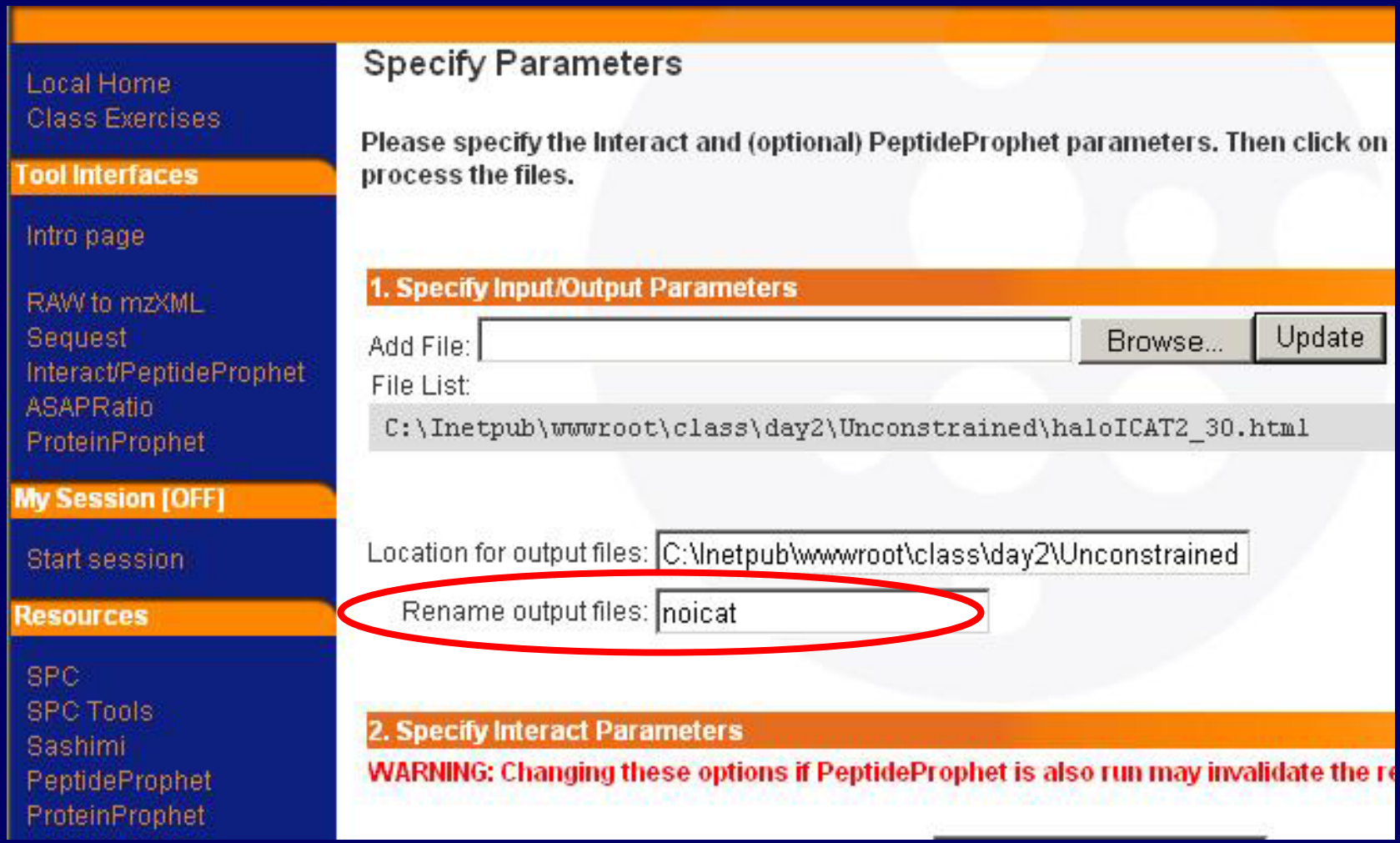


Sort Data by Computed Probability



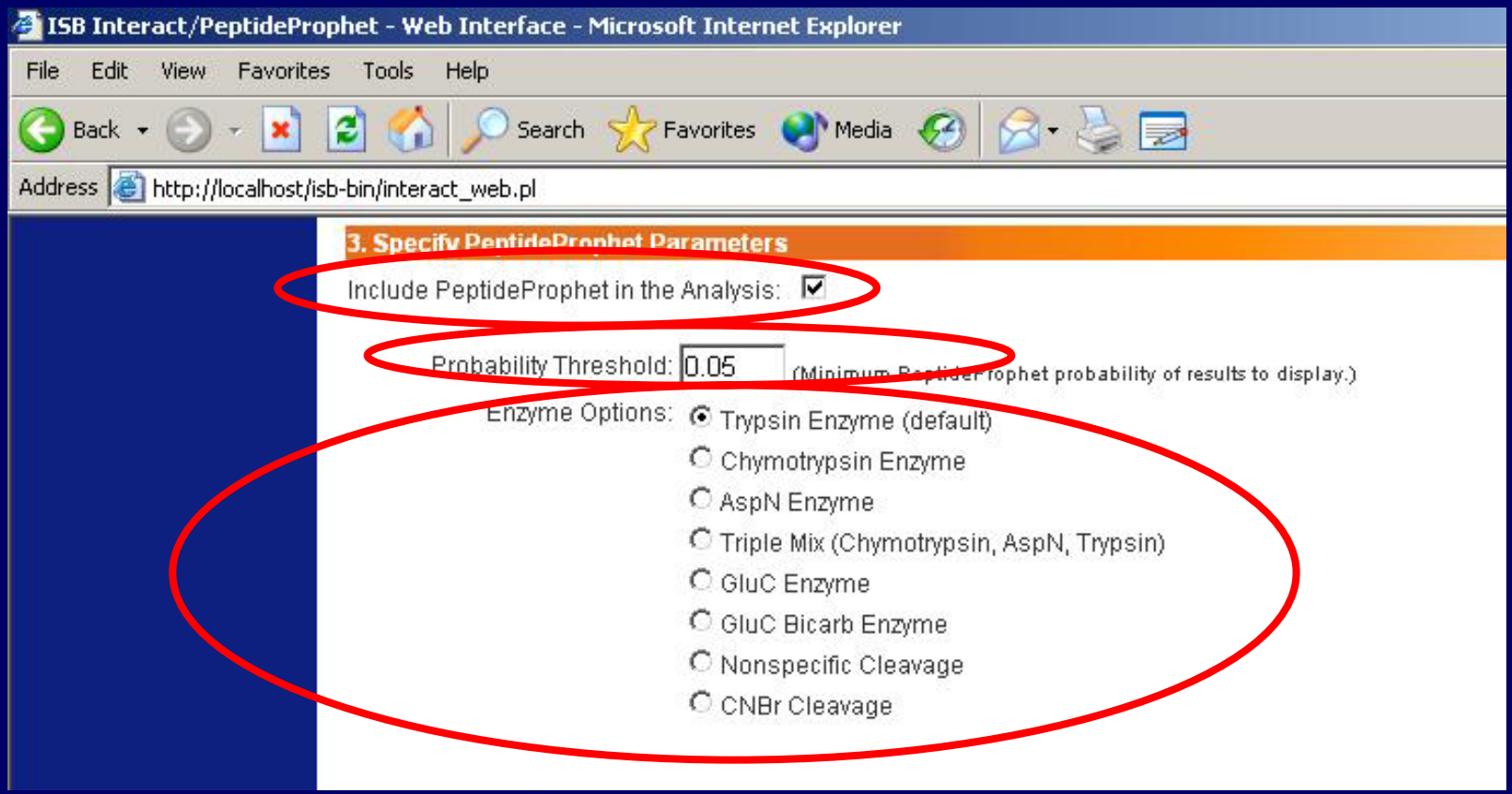
Some Options for Interact

Rename Output File (e.g. to `interact-noicat.htm`):



Some Options for Interact

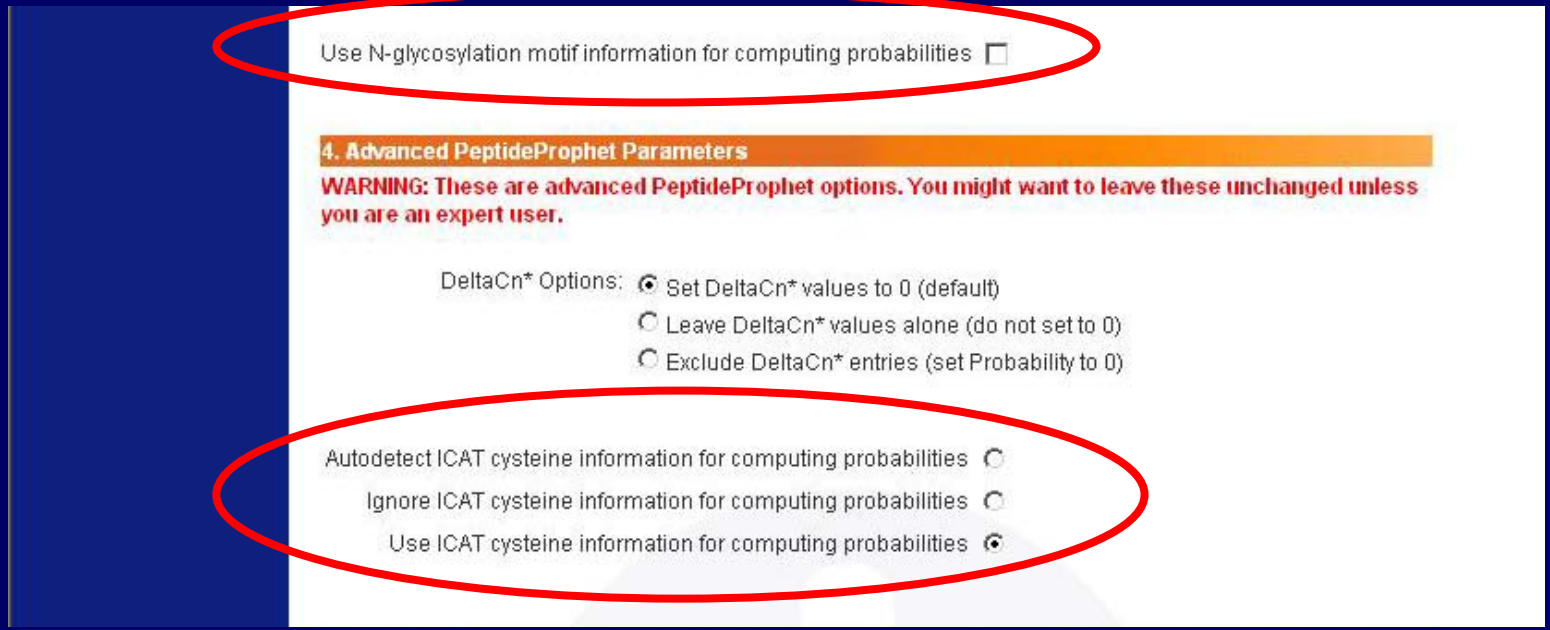
- no PeptideProphet analysis
- alternative minimum probability
- sample enzyme other than trypsin



Use of Supplemental Discriminating Information

Use additional discriminating information, including ICAT or N-glyc, when relevant

PeptideProphet automatically uses ICAT information when it thinks appropriate. Nevertheless, you can explicitly set whether or not ICAT information is utilized



DeltaCn* Example

Exclude charge: +1 ☐ +2 ☐ +3 ☐ Txt1: ☐ Txt2: ☐ J.Eng 04/2000

0.9999	91	./sergei_digest_A_full_01.0607.0609.2	1983.0	(-0.4)	4.6602	0.358	3224.6	1
0.9994	625	./sergei_digest_A_full_01.1523.1525.3	2046.3	(-0.7)	4.6539	0.307	1901.4	1
1.0000	117	./sergei_digest_A_full_01.0663.0665.3	2100.2	(+0.0)	4.5901	0.286	1627.3	1
1.0000	712	./sergei_digest_A_full_01.1675.1681.2	1832.0	(-0.3)	4.5772	0.393	1148.8	1
1.0000	651	./sergei_digest_A_full_01.1569.1571.3	3106.3	(+0.4)	4.5737	0.362	537.5	1
0.9981	942	./sergei_digest_A_full_01.2085.2089.3	2314.7	(+1.4)	4.5707	0.299	1348.2	1
0.9997	1045	./sergei_digest_A_full_01.2277.2279.3	2804.0	(-0.5)	4.5012	0.352	1236.7	1
0.8970	861	./sergei_digest_A_full_01.1951.1953.2	2314.7	(-0.4)	4.4874	0.348	482.4	1
0.9810	626	./sergei_digest_A_full_01.1529.1531.2	1480.7	(-0.5)	4.4192	0.315	1731.6	1
1.0000	176	./sergei_digest_A_full_01.0767.0767.2	1983.0	(+0.7)	4.4109	0.414	1966.2	1
0.9998	328	./sergei_digest_A_full_01.1009.1011.3	1440.7	(+0.8)	4.4102	0.302	2269.3	1
0.9999	1150	./sergei_digest_A_full_01.2487.2489.2	2219.5	(-0.5)	4.3932	0.334	1350.2	1
0.9505	940	./sergei_digest_A_full_01.2083.2087.3	2508.8	(-0.3)	4.3518	0.375	1166.5	1
0.9900	348	./sergei_digest_A_full_01.1027.1029.3	1640.9	(-0.4)	4.3059	0.308*	1451.7	1
0.9770	766	./sergei_digest_A_full_01.1783.1785.3	2119.5	(+1.2)	4.2653	0.356*	1147.4	2
0.9982	863	./sergei_digest_A_full_01.1955.1957.2	2052.4	(+0.4)	4.2588	0.302	453.4	1
0.9996	738	./sergei_digest_A_full_01.1727.1729.2	1531.8	(-0.1)	4.2559	0.427	1676.8	1
0.9999	168	./sergei_digest_A_full_01.0753.0753.2	1983.0	(+0.8)	4.2504	0.308	2035.6	1

#	Rank/Sp	(M+H)+	deltCn	XCorr	Sp	Ions	Reference		Peptide
1.	1 / 2	2119.4870	0.0000	4.2653	1147.4	30/72	SW:PHS3_HUMAN	+2	K.VAIQLNDTHPALSIPELMR.I
2.	2 / 1	2119.4870	0.0283	4.1445	1150.6	29/72	sp|P00489|PHS2_RABIT	+2	K.VAIQLNDTHPSLAIPELMR.V
3.	3 /158	2119.4894	0.3564	2.7454	377.2	22/76	SWN:PRKD_HUMAN	+4	A.VPSAGLRLFALHASQFSTCL.L
4.	4 /461	2119.5767	0.3948	2.5813	318.8	19/68	GP:AY027526_1		S.VAKLLHPQLTCRLLELRD.I

DeltaCn* Options

There are three ways asterisked deltas can be treated by PeptideProphet:

1. Penalize (the default option, sets asterisked deltas to 0)
2. Leave alone (suitable for the context of homologues)
3. Exclude (the most conservative, assigns probability 0)

Ongoing Developments for PeptideProphet

Optimize for various additional mass spectrometers

- New discriminant function

Adapt to additional methods for assigning peptides to tandem mass spectra

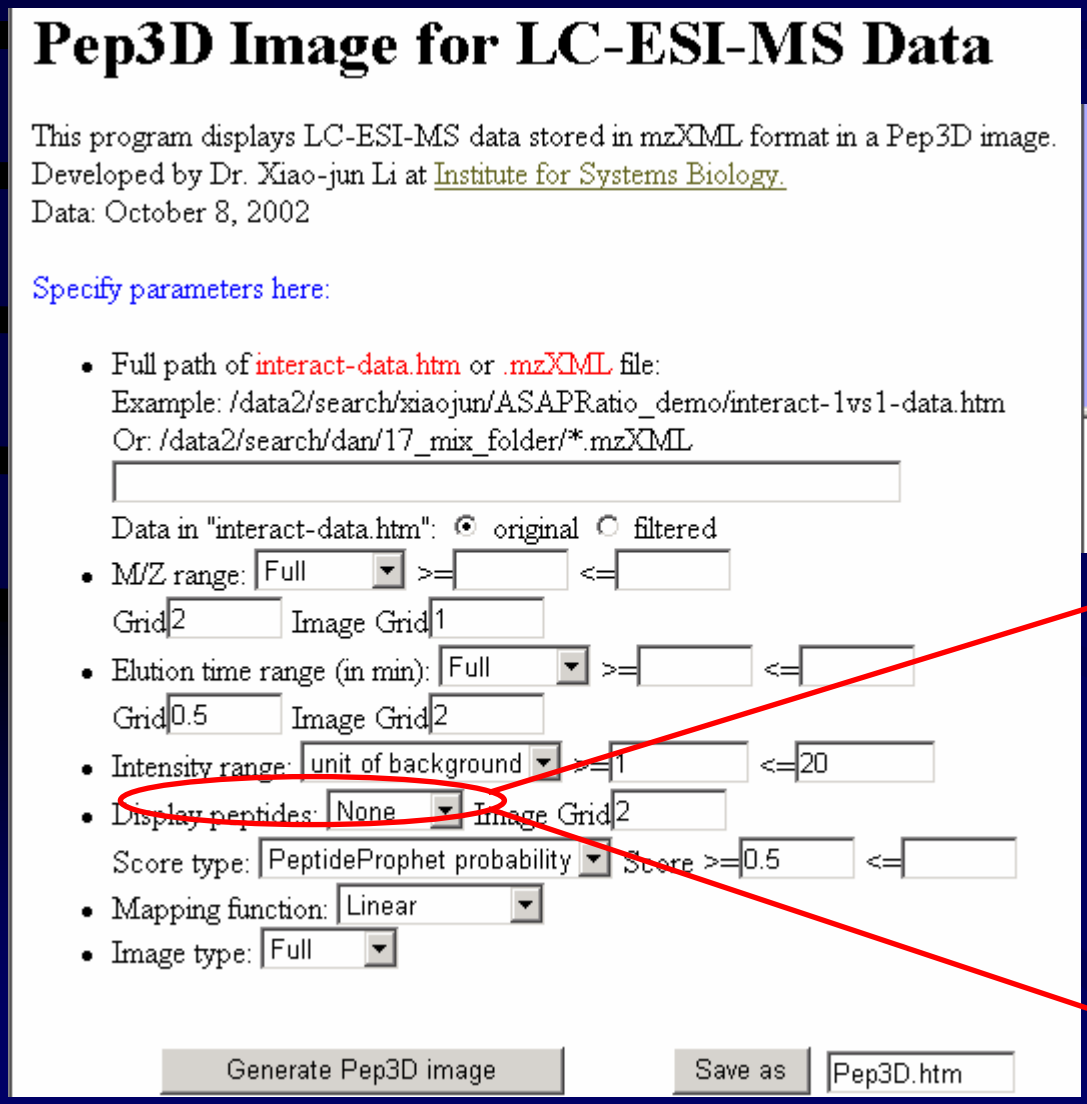
- Mascot ✓
- COMET ✓
- X!Tandem
- Others

Pep3D mzXML Data Viewer

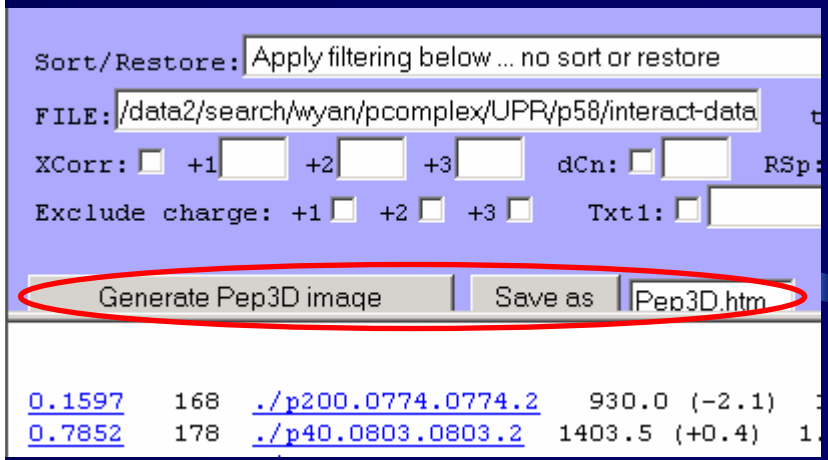
Xiao-jun Li

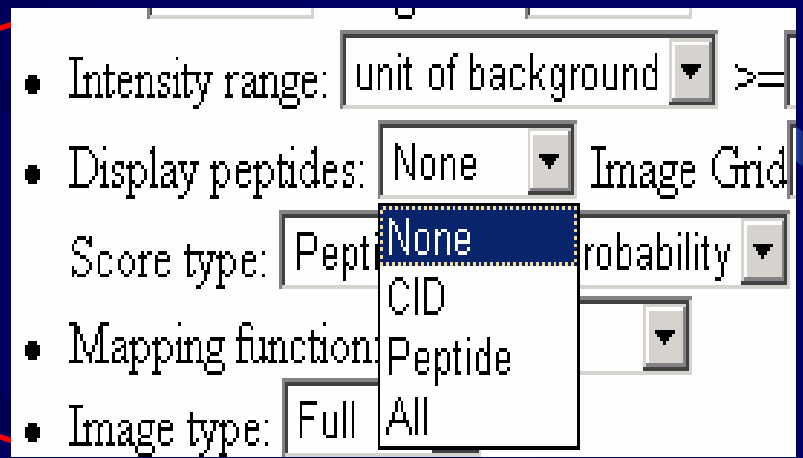
1. Main features of Pep3D
2. Evaluating sample quality
3. Evaluating LC-ESI-MS/MS performance

1. Main Features of Pep3D

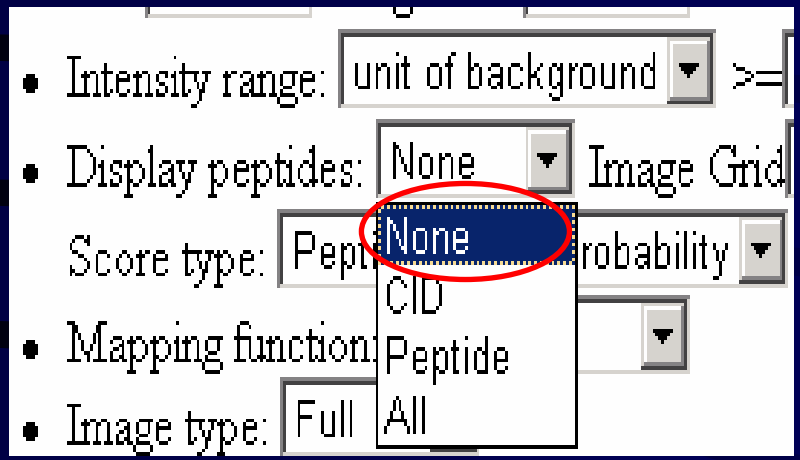


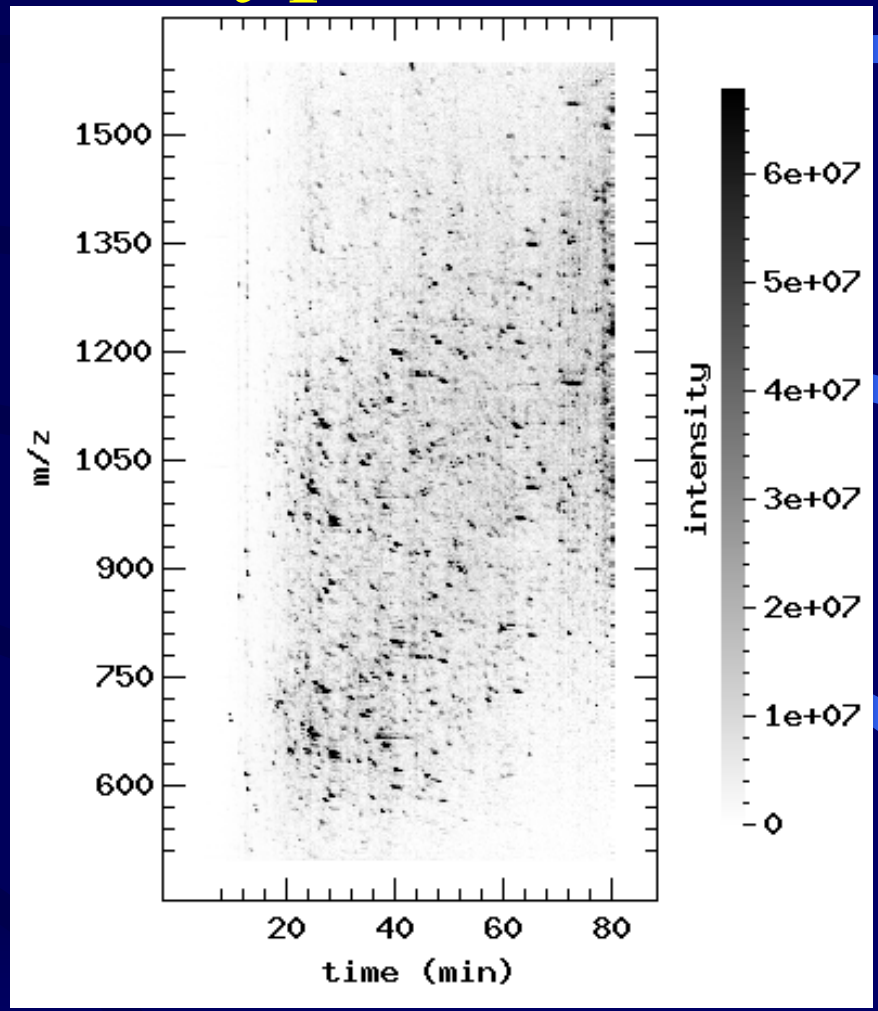
interact



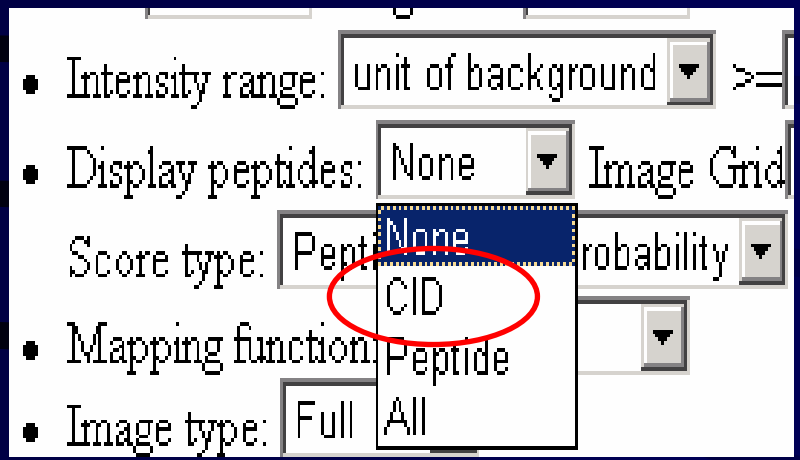


Pep3D Images: Type A

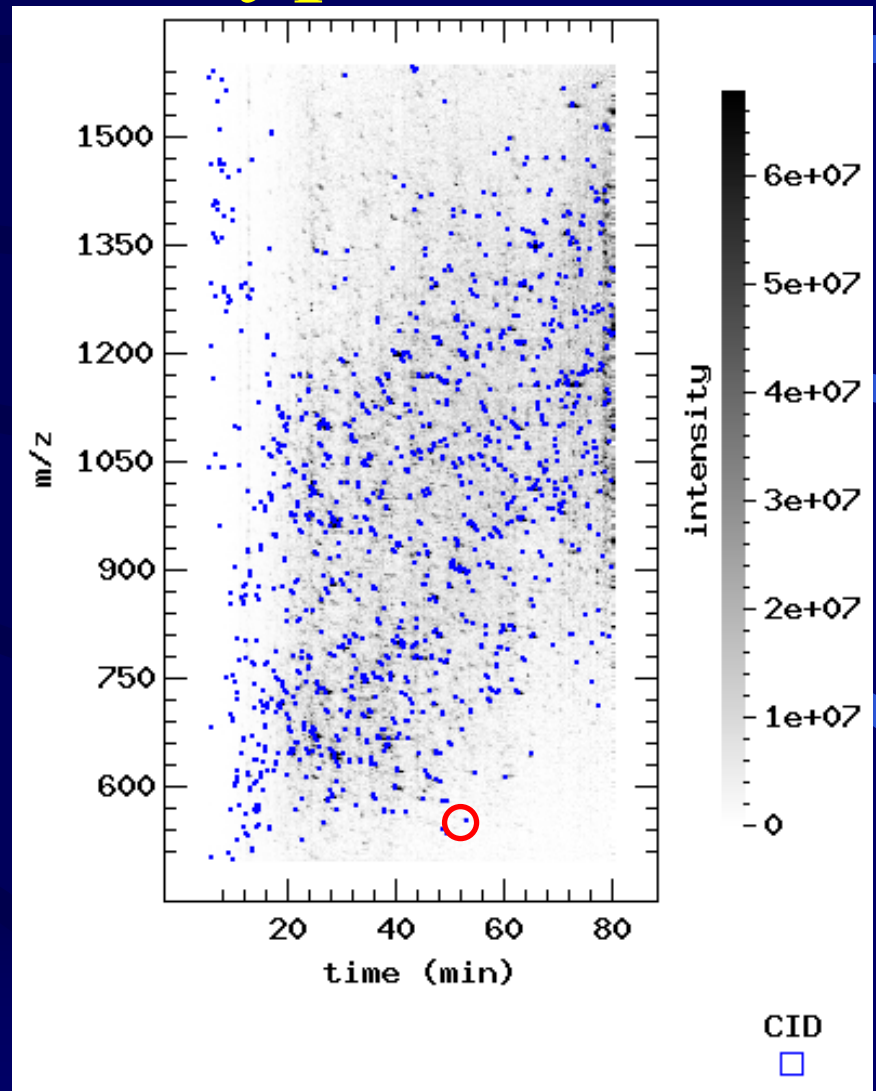




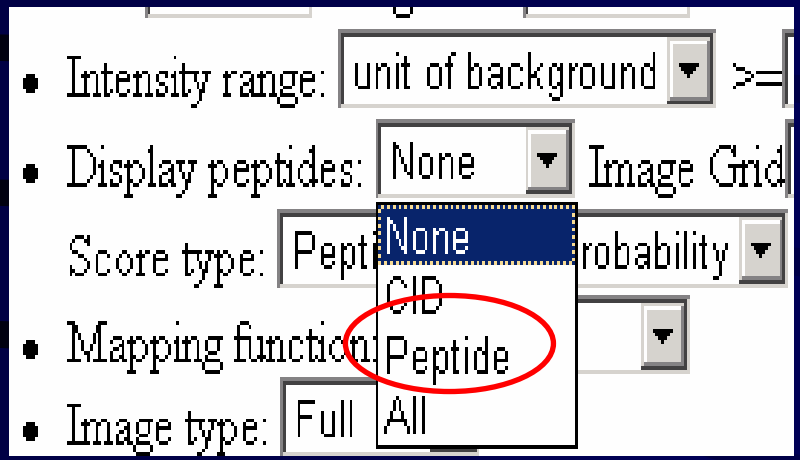
Pep3D Images: Type B



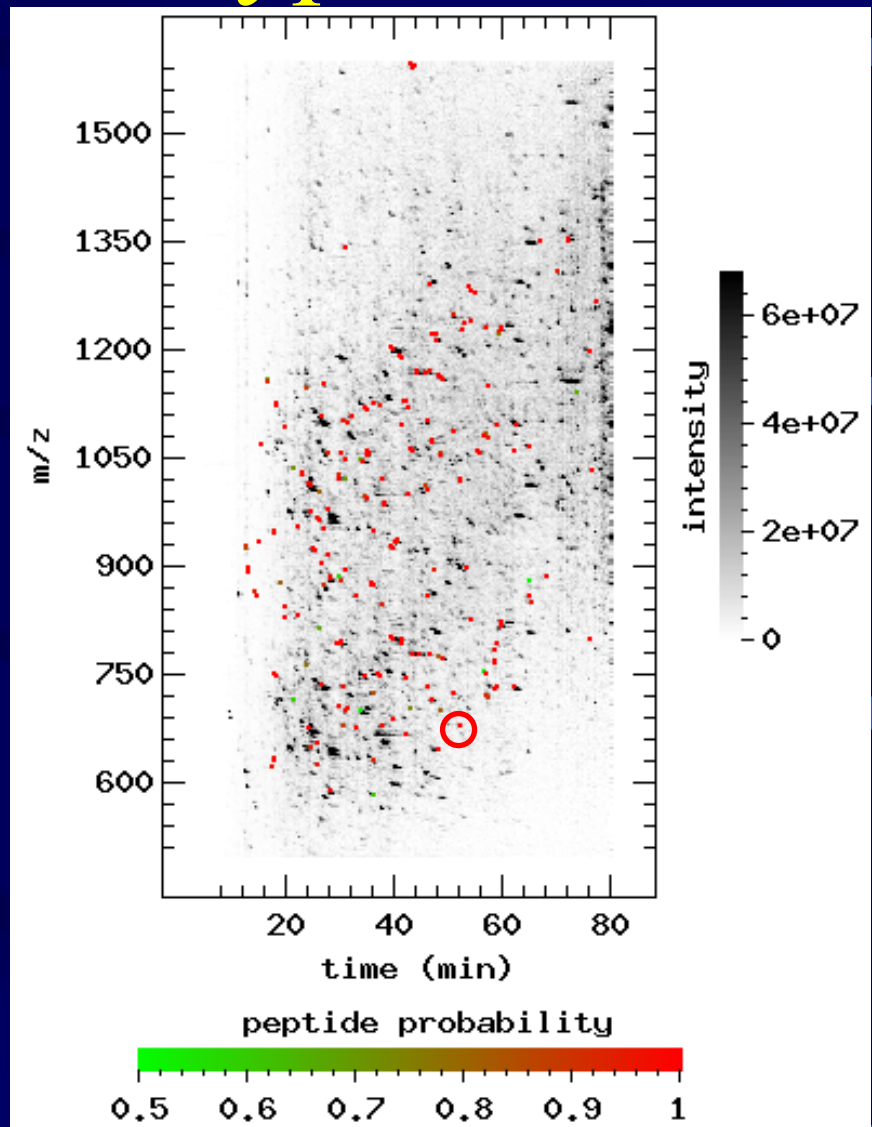
CID (1198)



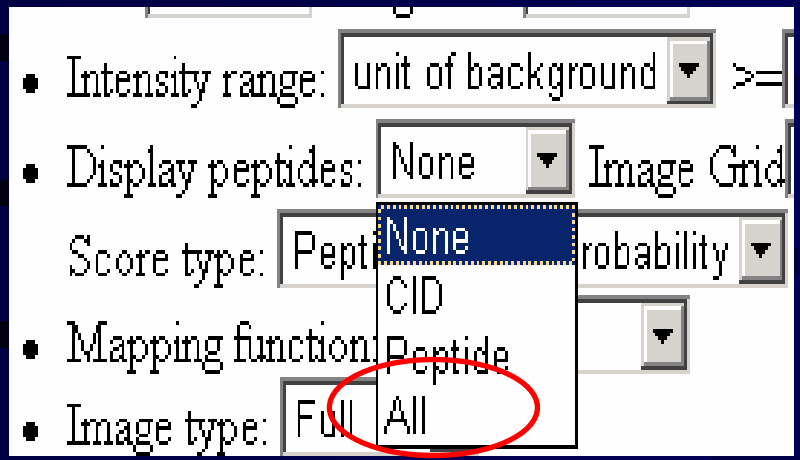
Pep3D Images: Type C



peptide (221)

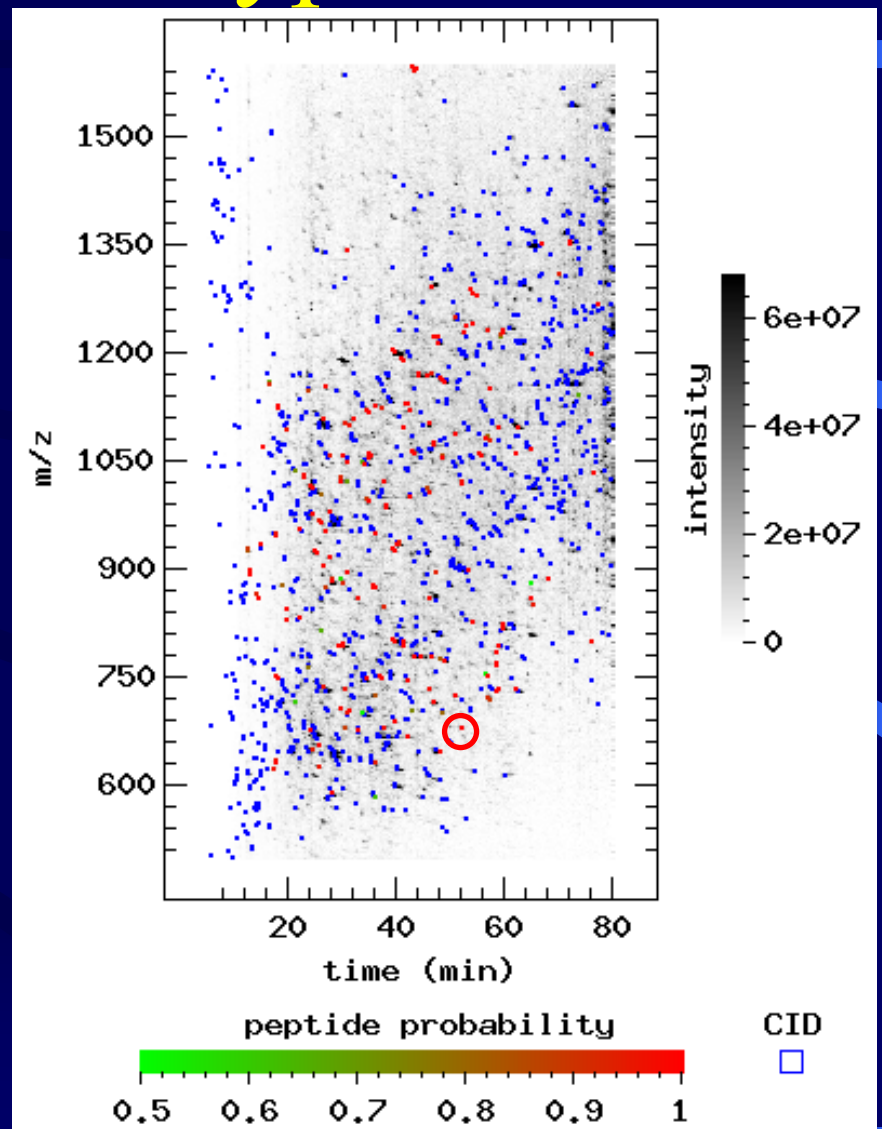


Pep3D Images: Type D

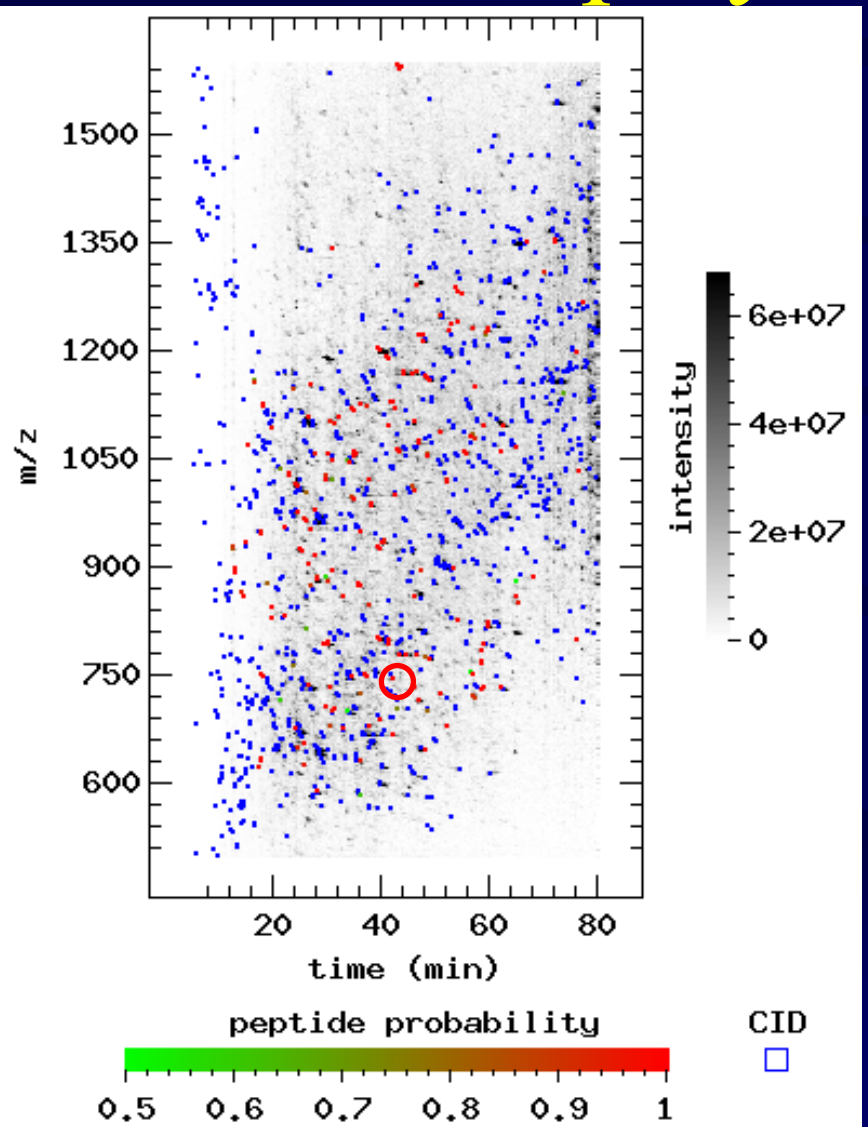


peptide/CID

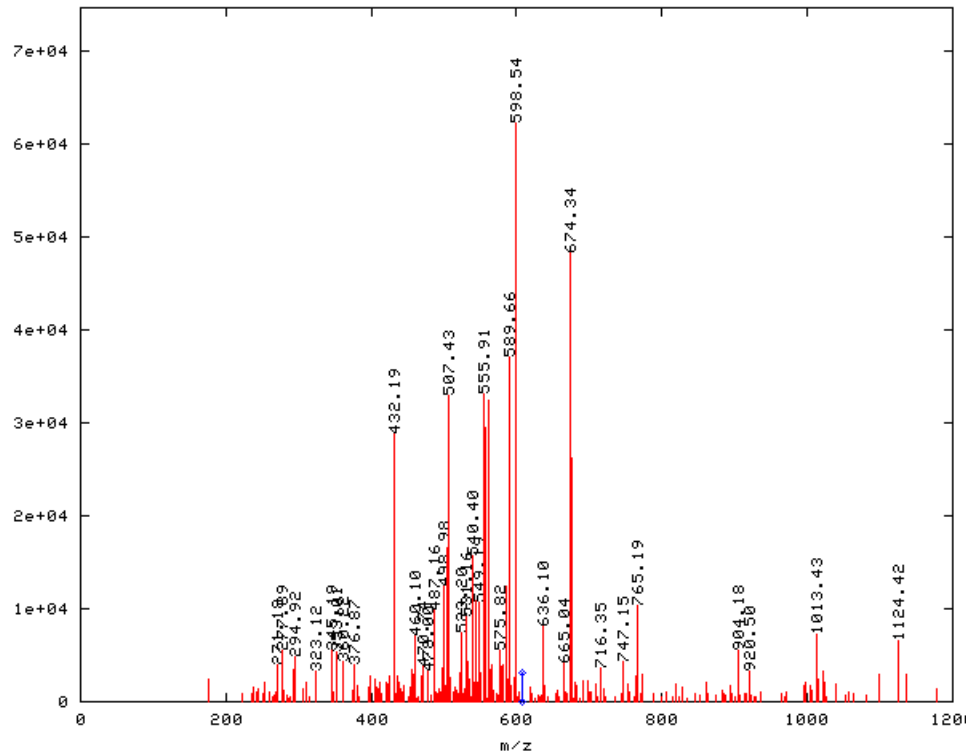
(221/1196 = 0.184783)



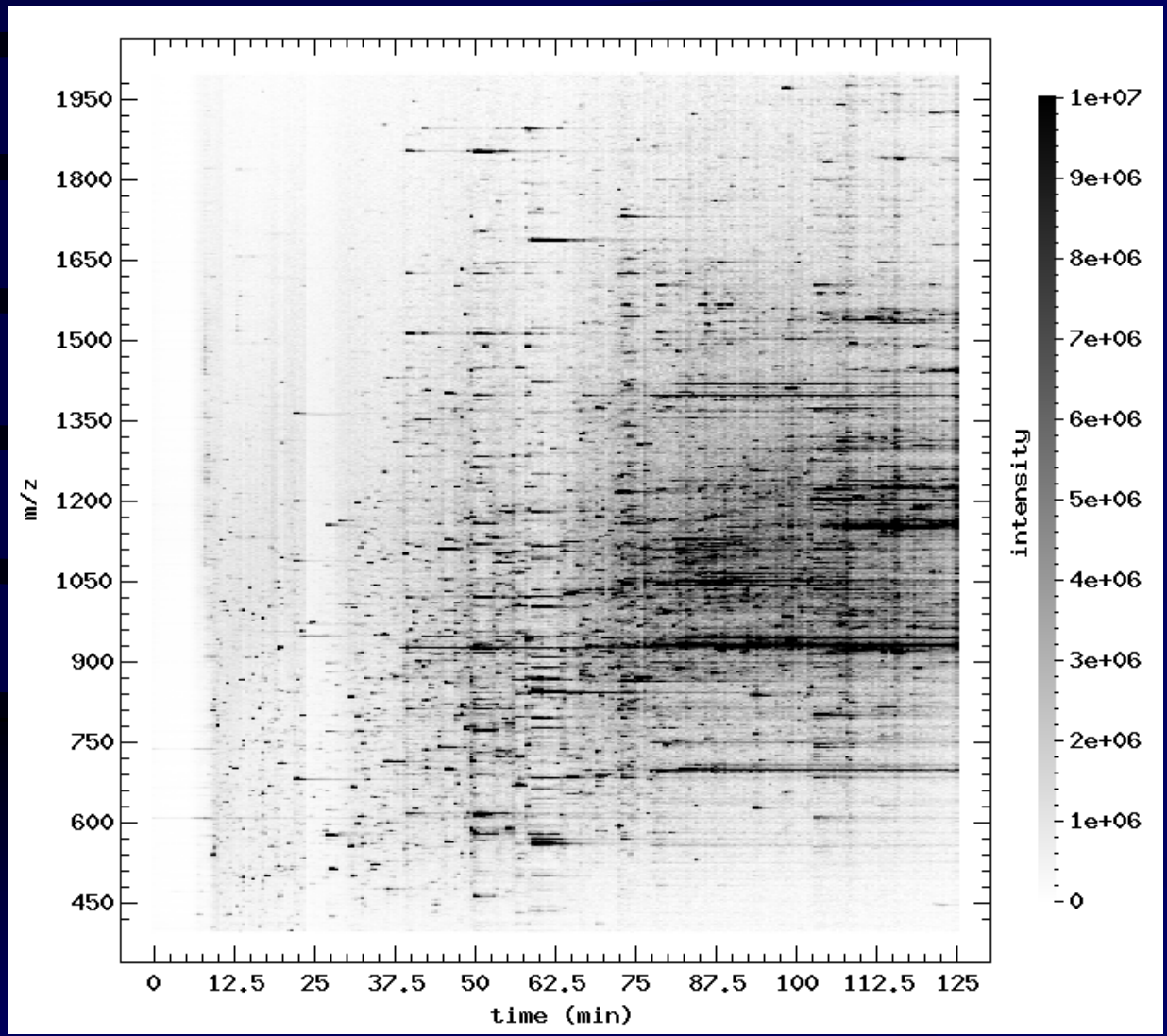
Display CID Spectrum



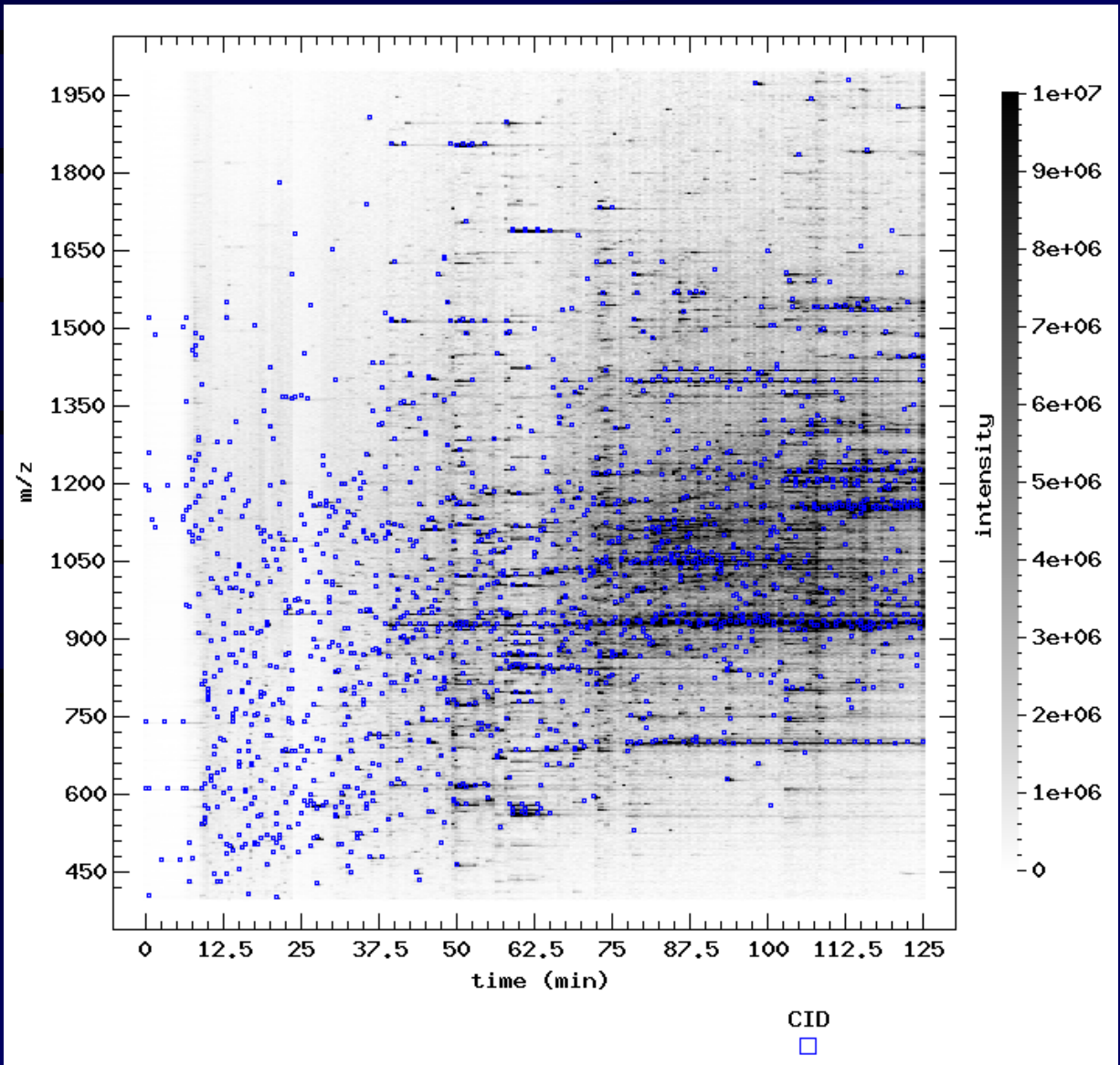
- scan number: 783
  m/z: 607.578003
  time: 40.491000
- ID:
  0.9157 155 ./p40.0783.0783.2 1214.3 (-0.1) 1.6765 0.144 244.8 64 8/ 16 IPI00289535 R.VTC
  0.713 (56.292%) 0.845580 1



Export to dta charge: +1 +2 +3 +4

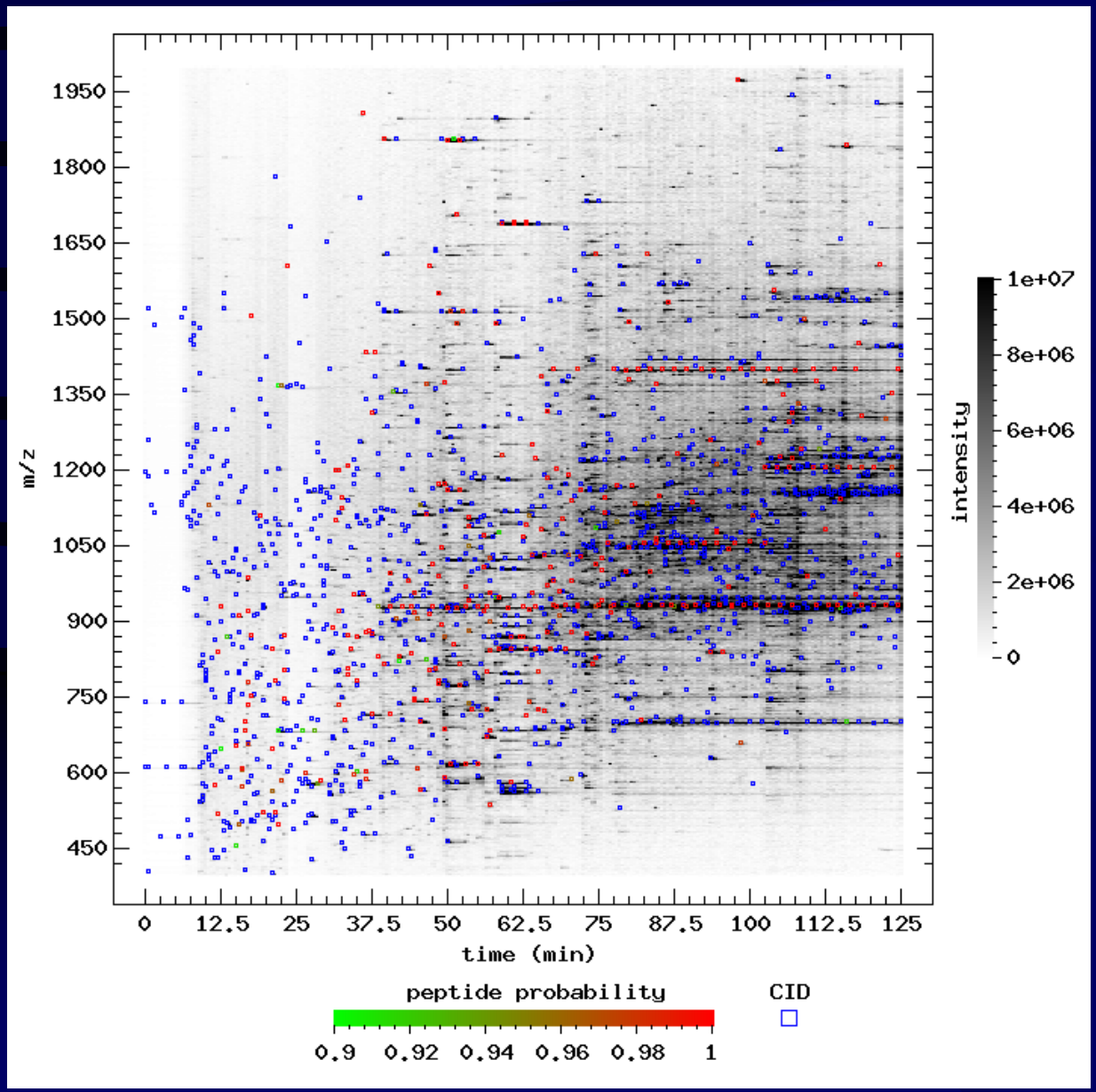


Features: 2720



Features: 2720

CIDs: 1633



Features: 2720

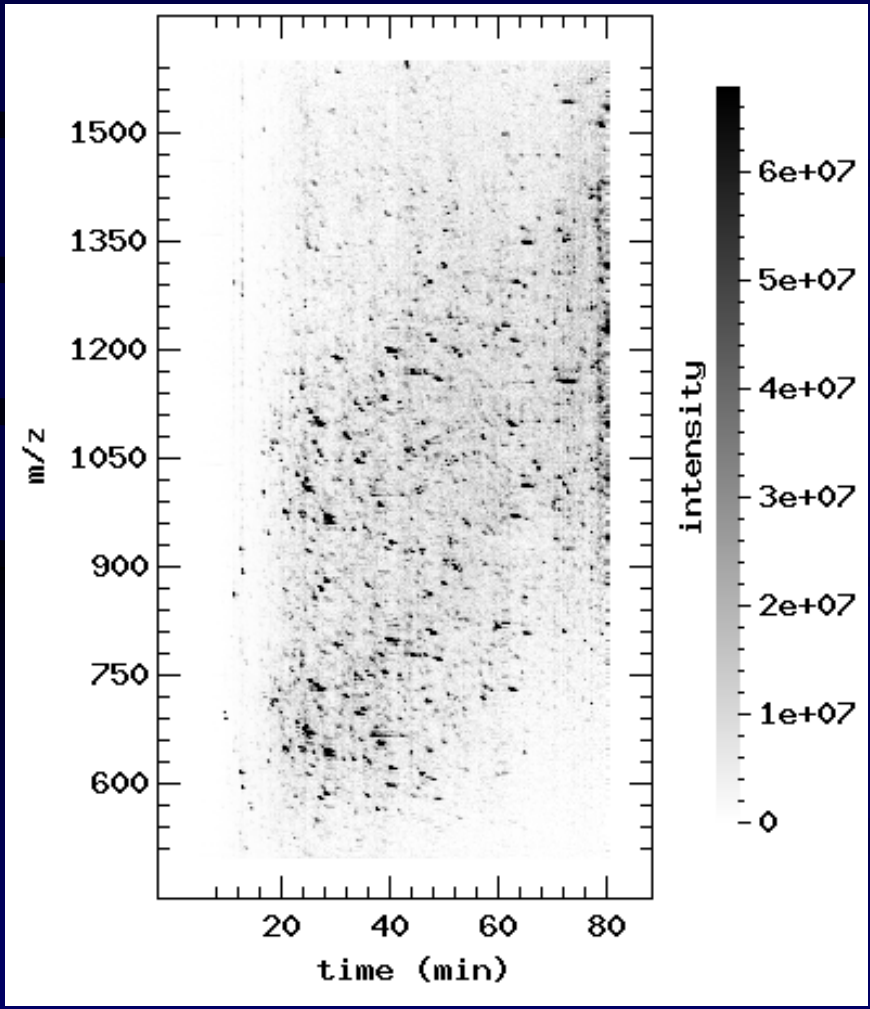
CIDs: 1633

IDs: 363

ID/CID: 22%

ID/feature: 13%

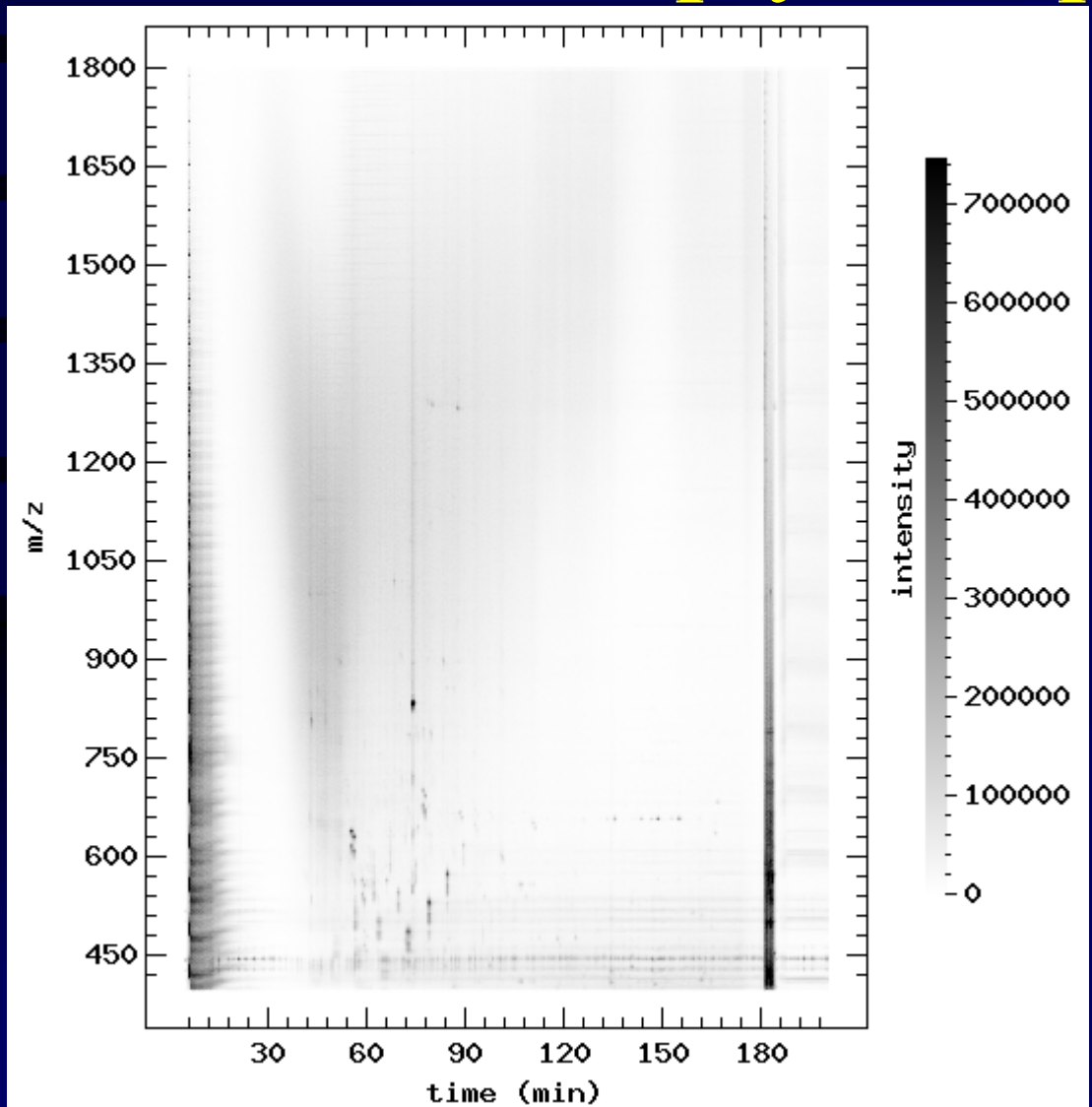
2. Evaluating Sample Quality



Good sample:

Plenty well-localized spots without any particular large-scale pattern

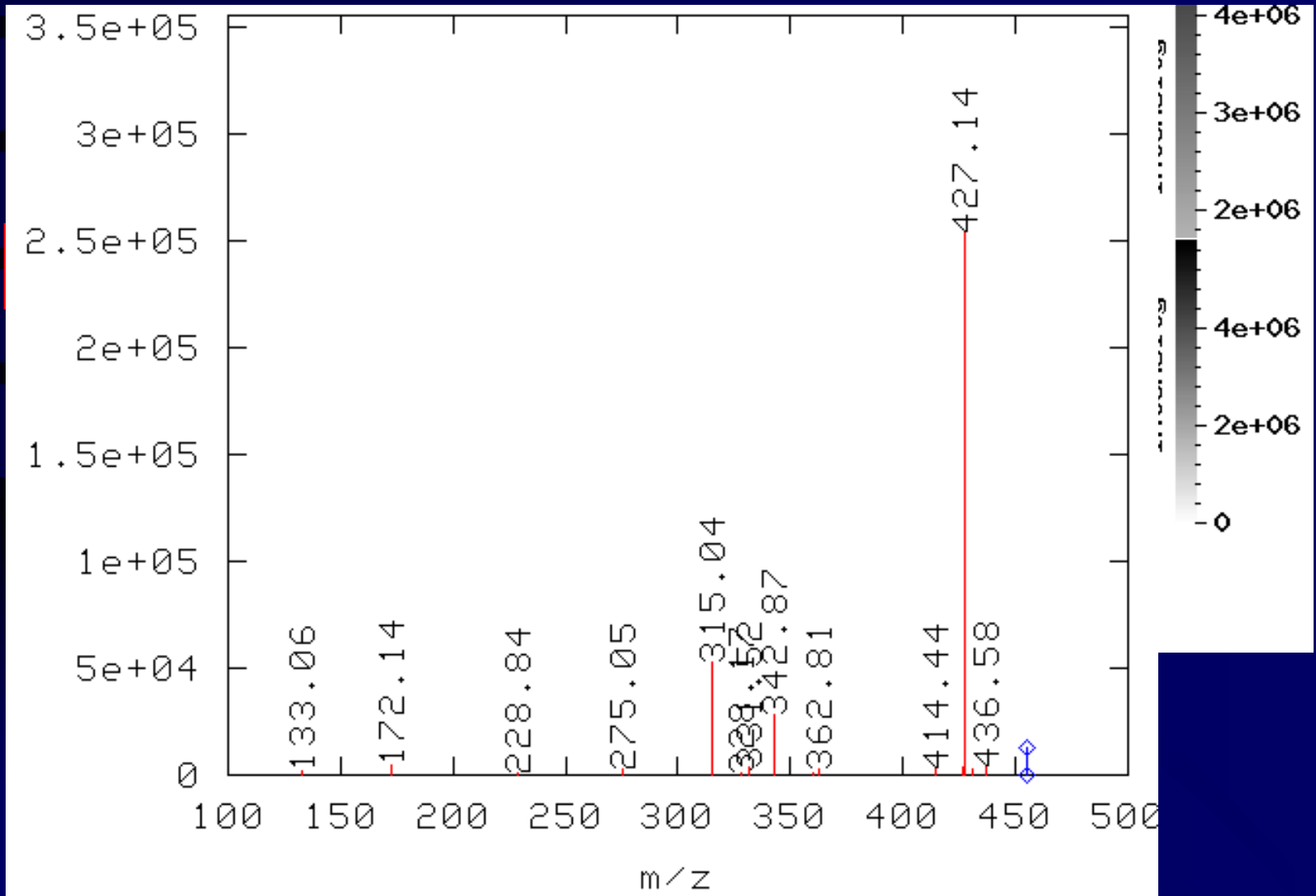
Empty Sample



peptide/CID (0/311 = 0)

- Very few localized spots
- Mainly background noise
- Distinguishable from no-spray

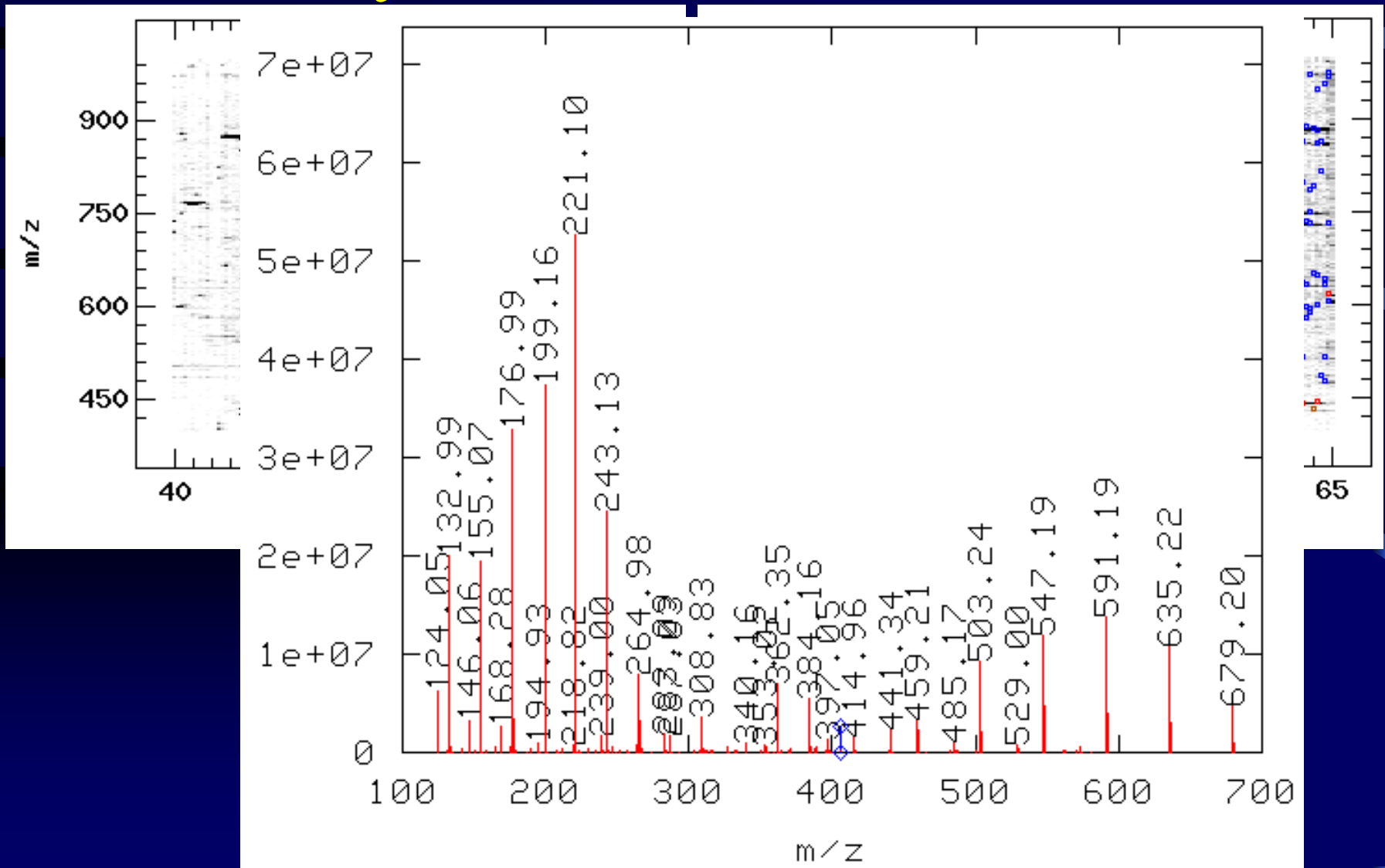
Chemical Contamination



Features of Chemical Contamination

- Long horizontal streaks
- Low m/z values
- Singly charged ions
- Many wasteful CID attempts
- Can be put on CID exclusive list

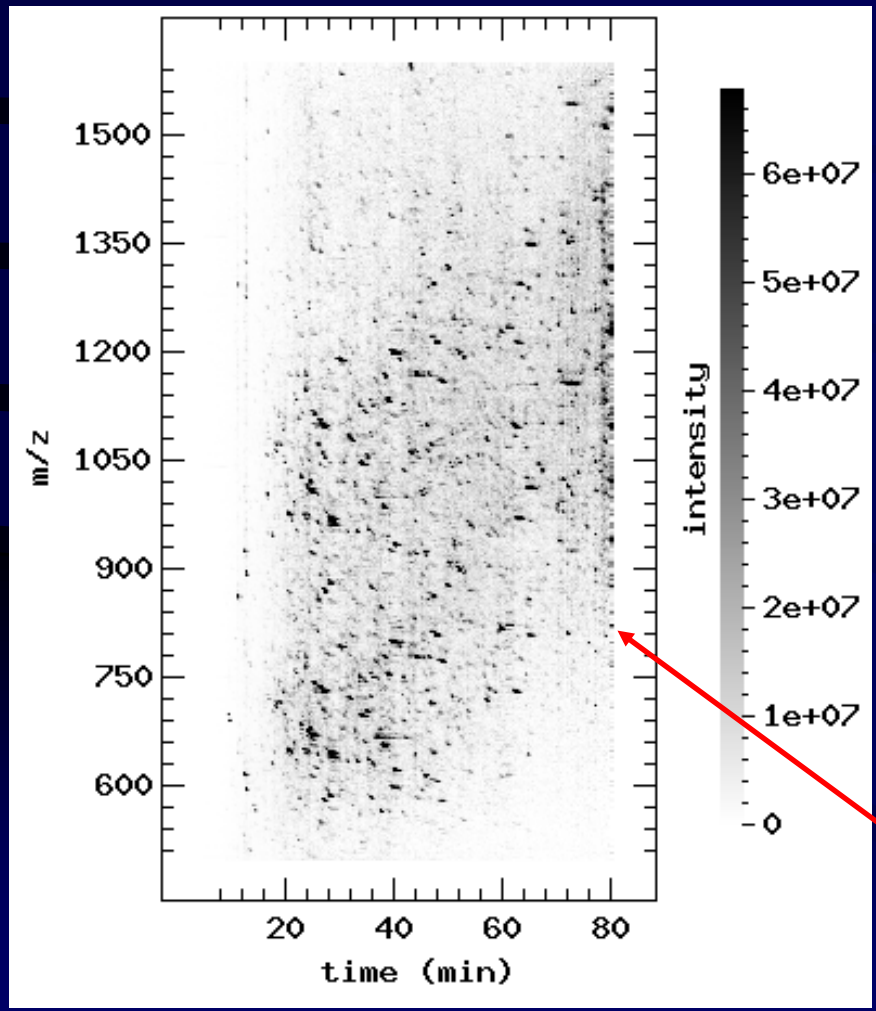
Polymer Contamination



Features of Polymer Contamination

- Localized spots running off-diagonally
- Equal distance in m/z
- Almost equal distance in time
- May be ionized in multiple charge states
- Many wasteful CID attempts
- Eliminated by better washing steps?

3. Evaluating LC-ESI-MS/MS performance

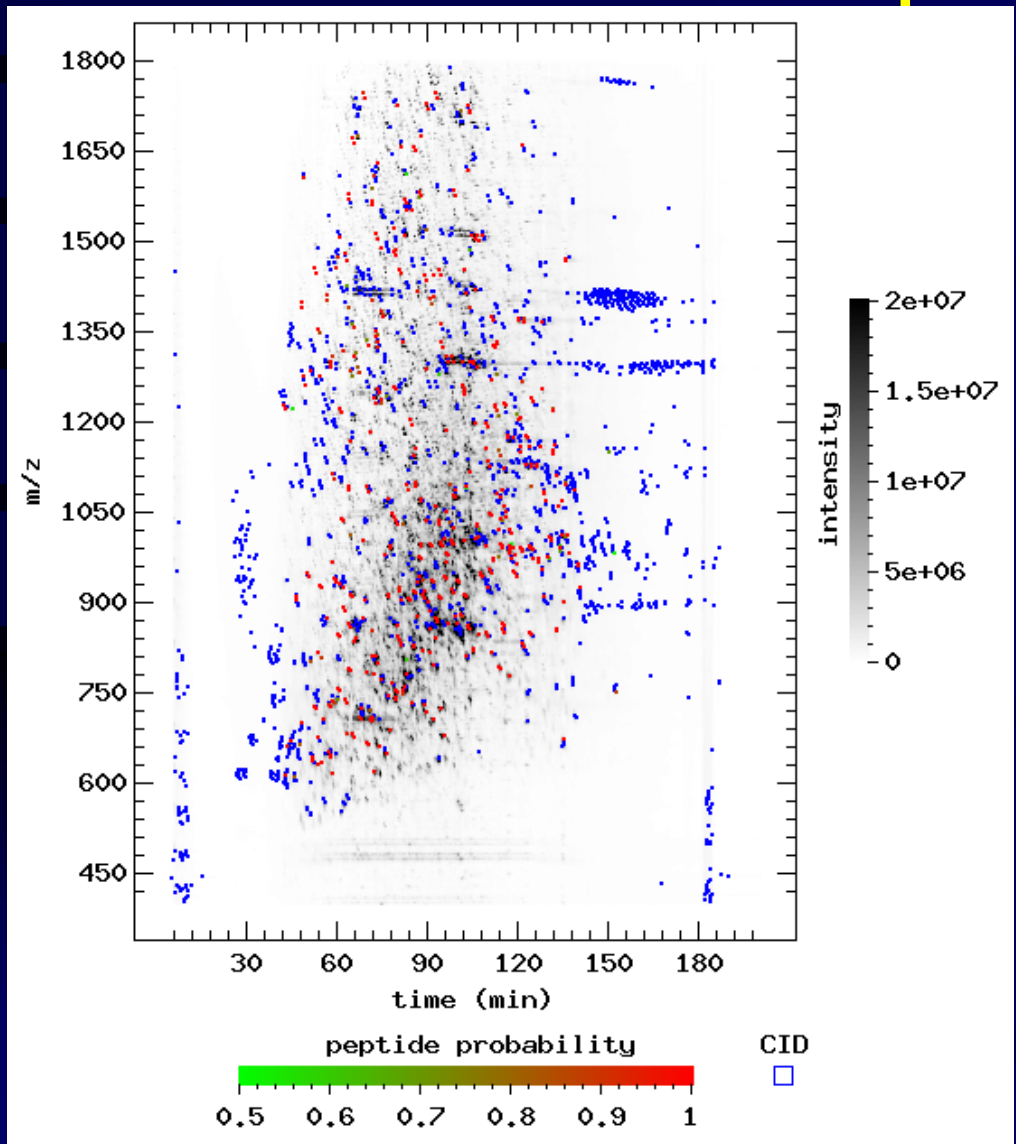


Good performance:

- Peptide ions evenly distributed
- Smooth background
- Majority of intensive ions fragmented

End too soon

Insufficient Sample Separation

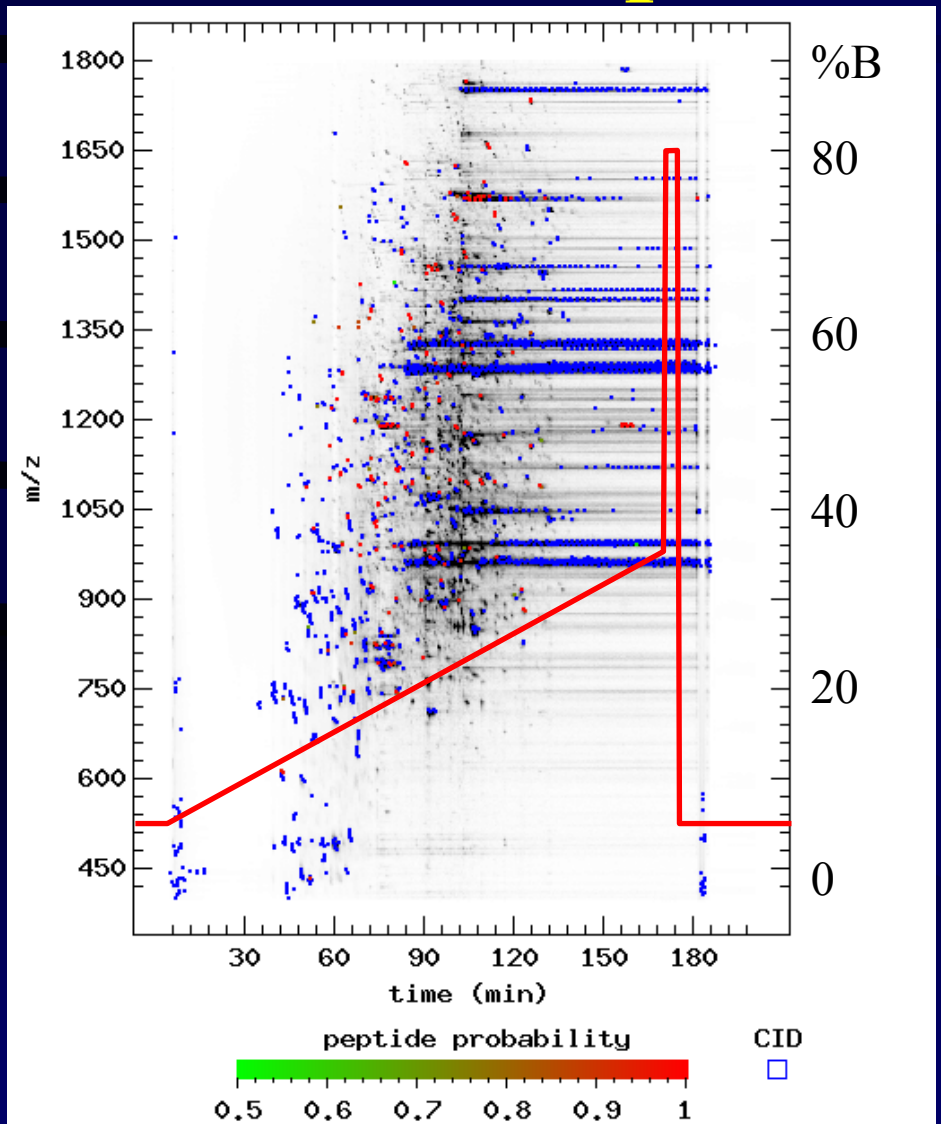


peptide/CID

(527/2150 = 0.245116)

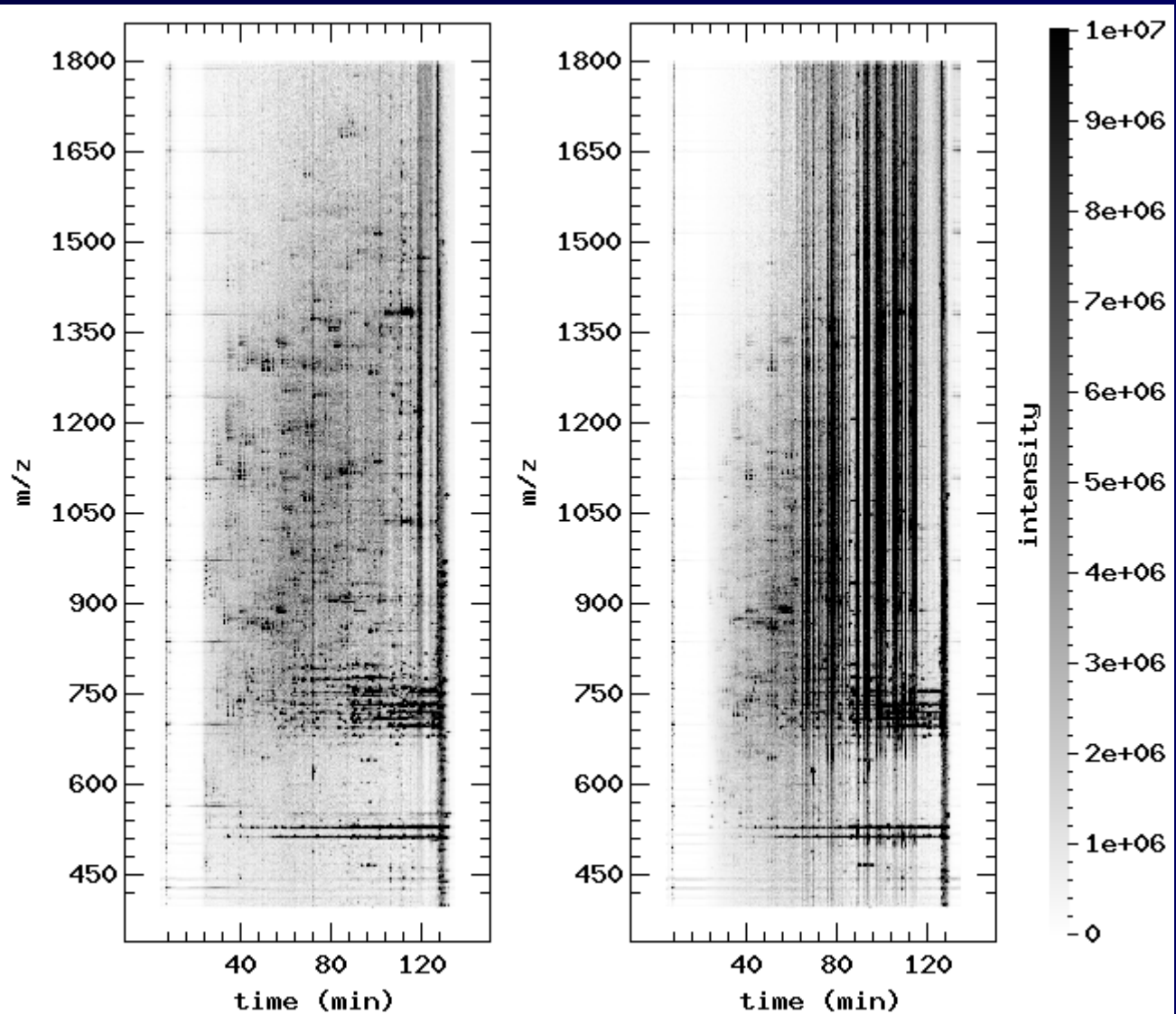
- Large proportion of intensive ions not fragmented
- Insufficient SCX separation
- Non-optimal RP separation

Non-Optimal RP Gradient



- 1st ID: 42 min
- Effective range: 60-120 mins
- Horizontal streaks
- Larger slope at beginning
- Slower slope in middle
- Higher %B at end

Bad RP Column



Same sample, same system, different columns

peptide/CID

(354/3004 = 0.12)
(224/2760 = 0.08)

37% less IDs

Quantification also suffers

Summary

- Pep3D can be used to evaluate sample quality and LC-ESI-MS/MS performance
- Other applications possible
- Suggestion: Use complex standard sample to check system performance

Exercises with PeptideProphet

- Accuracy of computed probabilities
- Utility of conventional SEQUEST score thresholds and PeptideProphet analysis
- Model results for ICAT data analyzed with and without ICAT information
- Model results for unconstrained vs. tryptic constrained search results

Exercise Datasets

Many of the exercises utilize SEQUEST search results generated from datasets for which true results are known:

- HaloICAT: ICAT halobacterium sample searched against a halo_plus_human protein database

PeptideProphet run on these datasets automatically colors all correct corresponding proteins red !