PeptideProphet: Validation of Peptide Assignments to MS/MS Spectra

Andrew Keller

Day 2

October 17, 2006

Andrew Keller
Rosetta Bioinformatics, Seattle

INSTITUTE FOR Systems Biology

Outline

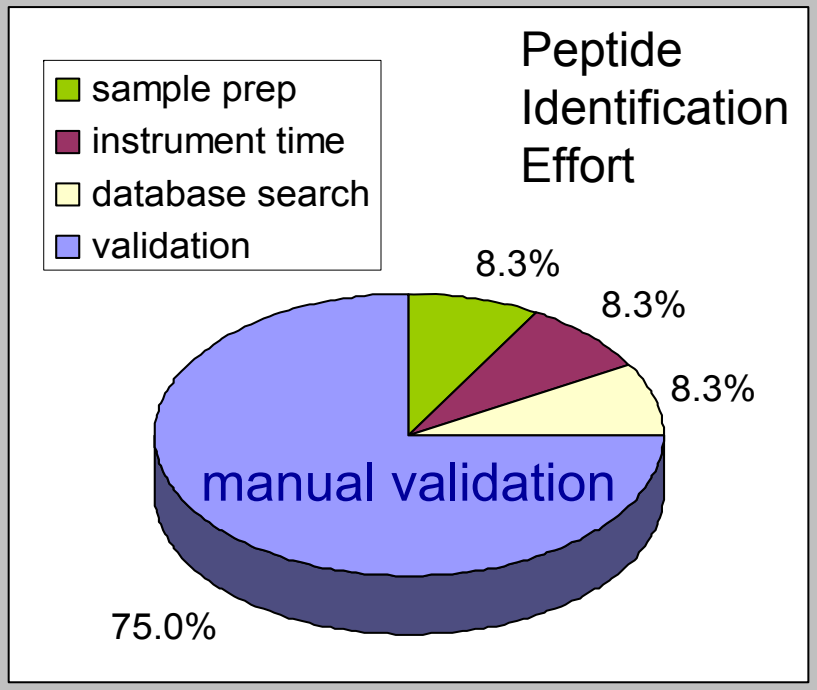
- Need to validate peptide assignments to MS/MS spectra
- Statistical approach to validation
- Running PeptideProphet software
- Interpreting results of PeptideProphet
- Exercises

Most search results are wrong

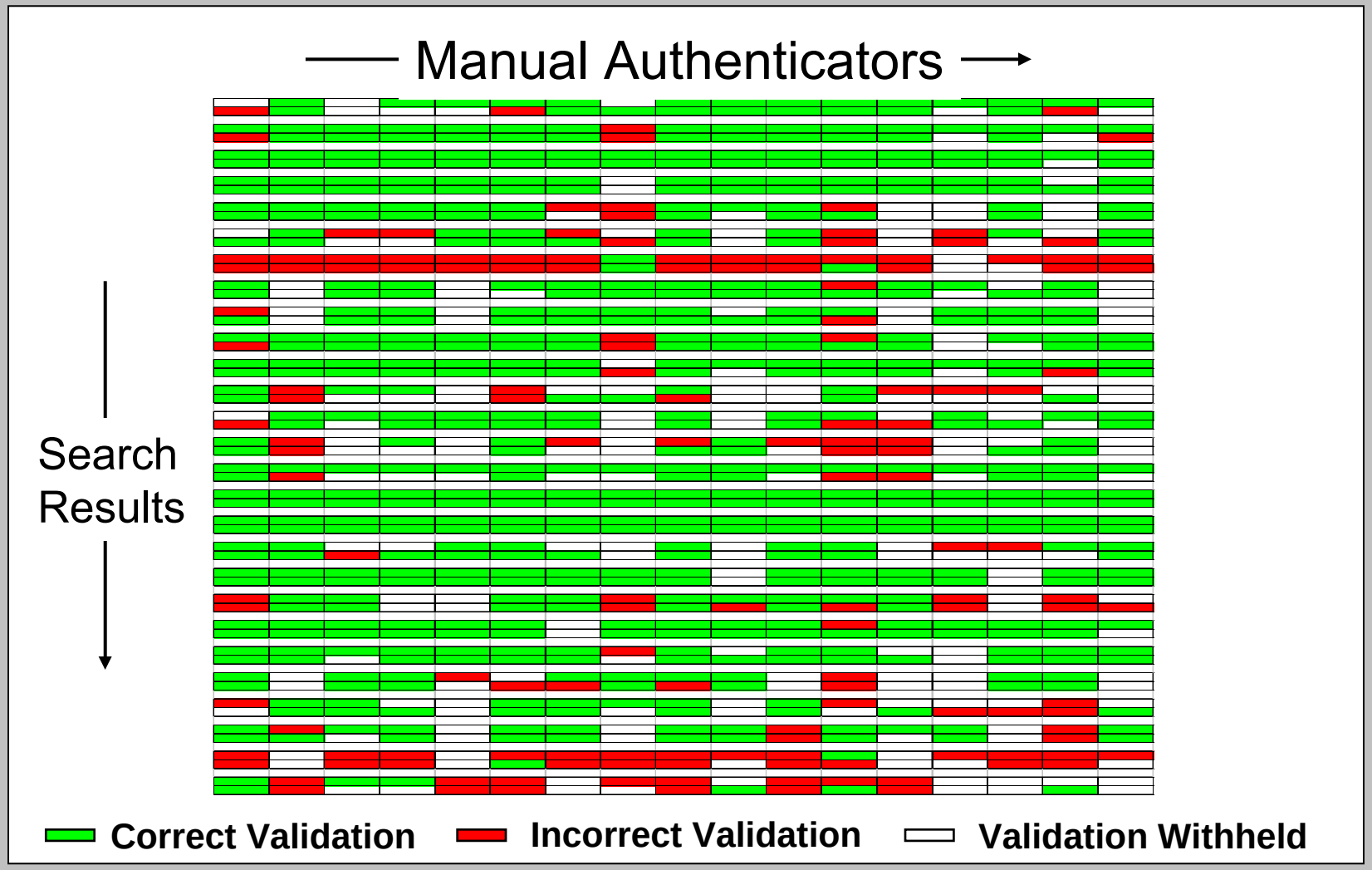
- [M+2H]2+/[M+3H]3+ uncertainty (LCQ)
- Non-peptide noise
- Incomplete database
  - *e.g.* post-translational modifications
- Multiple precursors
- Limitation of database search algorithm

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



(Un)reliability of Manual Validation



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 Mascot search result has:
  - Ionscore, Identityscore, Homologyscore, NTT (number of tryptic termini)
- Accept all results that satisfy:

  Ionscore > Identityscore

Traditional Filtering Criteria

- Each Mascot search result has:
  - Ionscore, Identityscore, Homologyscore, NTT (number of tryptic termini)
- Accept all results that satisfy:

  Ionscore > Identityscore (NTT = 2)

Traditional Filtering Criteria

- Each Mascot search result has:
  - Ionscore, Identityscore, Homologyscore, NTT (number of tryptic termini)
- Accept all results that satisfy:

  Ionscore > Identityscore (NTT = 2)

  Ionscore > Homologyscore

Traditional Filtering Criteria

- Each Mascot search result has:
  - Ionscore, Identityscore, Homologyscore, NTT (number of tryptic termini)
- Accept all results that satisfy:

  Ionscore > Identityscore (NTT = 2)

  Ionscore > Homologyscore (NTT = 2)

Traditional Filtering Criteria

- Each Mascot search result has:
  - Ionscore, Identityscore, Homologyscore, NTT (number of tryptic termini)
- Accept all results that satisfy:

  Ionscore > Identityscore (NTT = 2)

  Ionscore > Homologyscore (NTT = 2)

  Ionscore ≥ 30

Traditional Filtering Criteria

- Each Mascot search result has:
  - Ionscore, Identityscore, Homologyscore, NTT (number of tryptic termini)
- Accept all results that satisfy:

  Ionscore > Identityscore (NTT = 2)

  Ionscore > Homologyscore (NTT = 2)

  Ionscore ≥ 30 (NTT = 2)

Problems with Traditional Filtering

- Different research groups use different thresholds
- 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.

Use of Forward/Reverse Database to Estimate False Positive Error Rates

- Do search against single Forward/Reverse database containing usual entries along with their sequence-reversed counterparts
- Forward and Reverse protein sequences each comprise 50% of the database peptides
- Incorrect results, taken at random from the database, are predicted to correspond with Reverse protein sequences on average 50% of the time
- Number of incorrect results passing any score filter calculated as twice the number of accepted results corresponding to Reverse proteins
- Search takes twice as long

Use of Separate Forward and Reverse Database Searches

- Do searches against Forward and Reverse databases separately
- Number of incorrect results in Forward search passing any score filter calculated as the number of results passing the same filter applied to the Reverse search
- Gives an overestimate of the number of incorrect results passing a filter since compares the Reverse search which has no correct results with the Forward search which may have up to 100% correct results
- Results of 2 searches must be analyzed in parallel

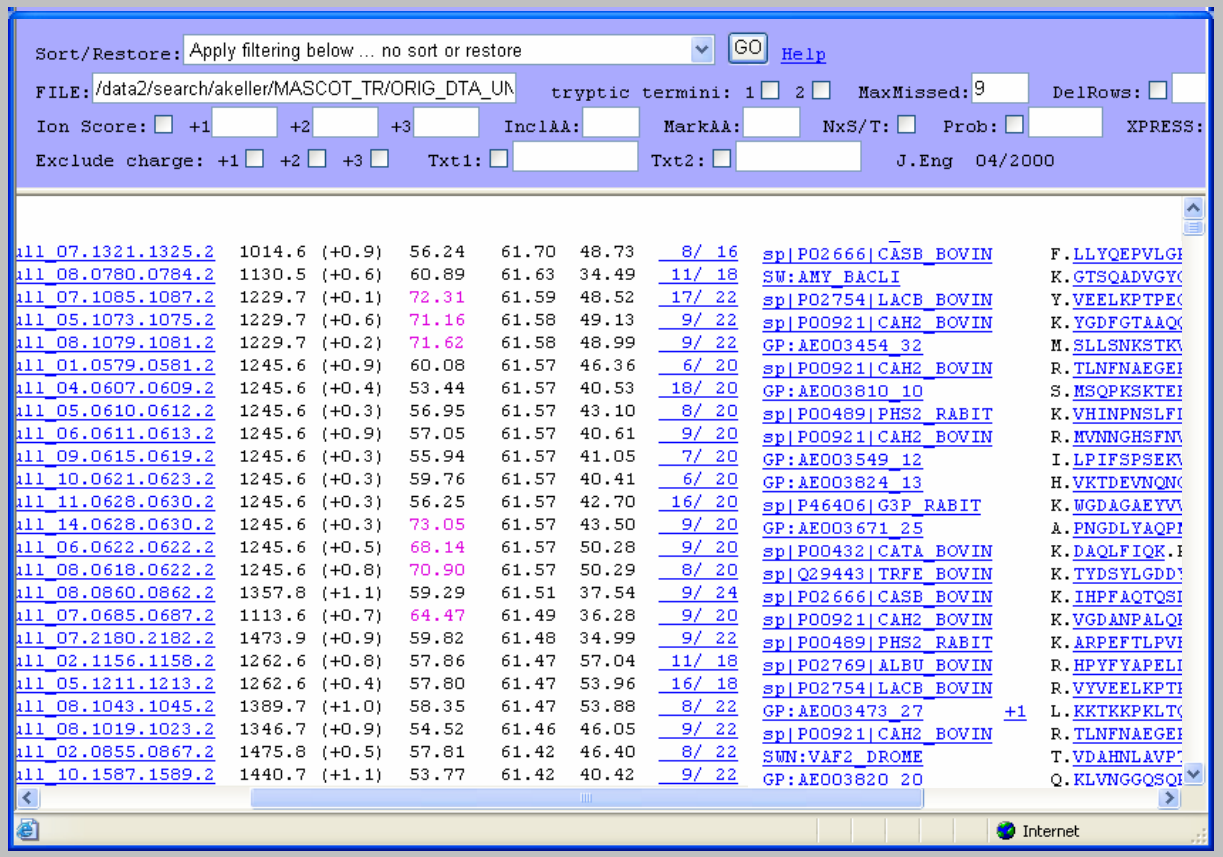
Statistical Approach

- 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 Mascot 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 Mascot vs. a *Drosophila* database appended with sequences of 18 control proteins and common sample contaminants

Training Dataset



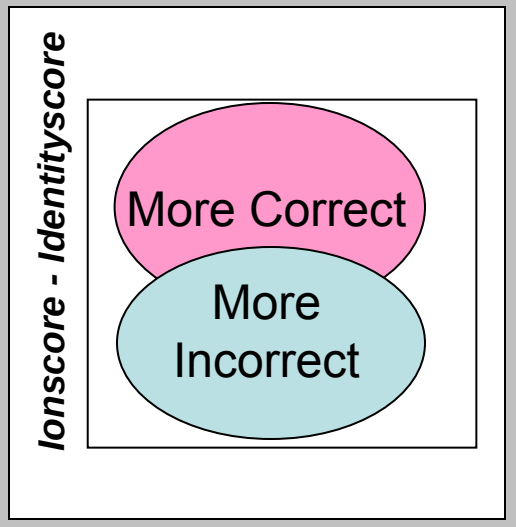
- Peptides corresponding to *Drosophila* proteins are incorrect
- Peptides corresponding to 18 control proteins or contaminants are correct*

Derive Discriminant Function

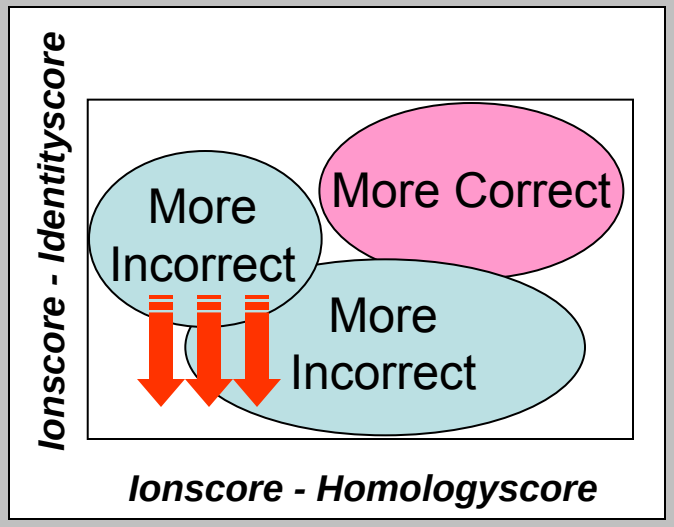
- Derive single search score best at discriminating correct from incorrect search results
  - Generally, can combine together multiple search engine scores, when available, into single linear combination score using Linear Discriminant Function Analysis (*e.g.* SEQUEST's Xcorr, DeltaCn, and SpRank)
  - Use search engine score directly if only one
- Derive separately for search results of each parent ion charge (1+, 2+, and 3+)

Mascot Discriminant Function

- Use (Ionscore – Identityscore) difference

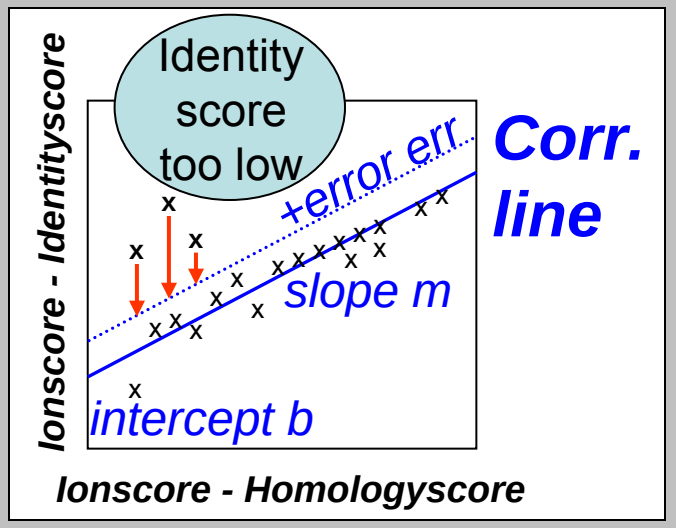


- Secondarily, use (Ionscore – Homologyscore) difference to penalize some predominantly incorrect results and improve discrimination



Mascot Discriminant Function

- In particular, use the (Ionscore – Identityscore) difference adjusted for the Average Identityscore in the dataset for given parent ion charge
- Require (Ionscore – Identityscore) not exceed *m**(Ionscore – Homologyscore) + *b* + *err*, where *m*, *b*, and *err* are correlation parameters learned from the data for each parent ion charge
- Discriminant Function, F = 0.1 * {(Ionscore – Identityscore) + Average Identityscore} – 3.0



Compute Discriminant Score

Example:

Peptide = **LSISGTYDLK**

Precursor Ion Charge = 2

Ionscore = 50.91 Identityscore = 46

Homologyscore = 37 Ave. Identityscore = 47

Corr. Slope = 0.53, Intercept = -6.99, Error = 10

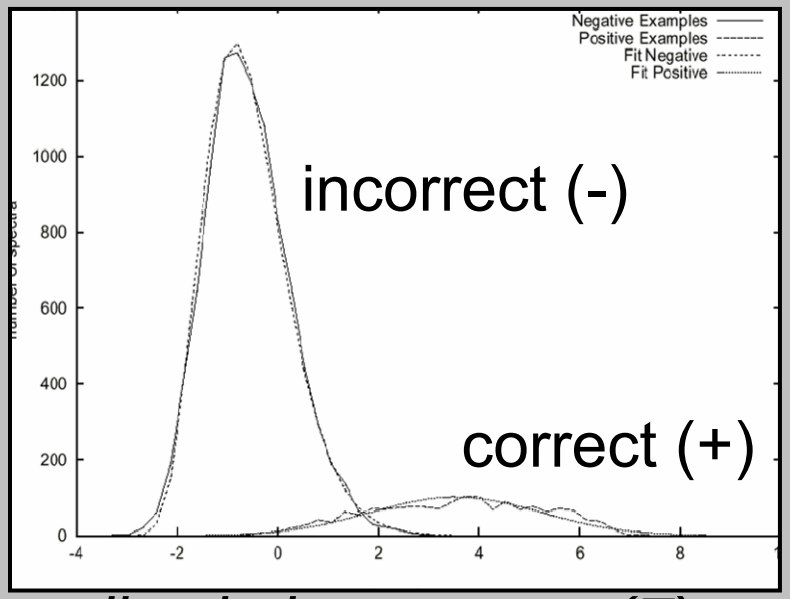
(Ionscore – Identityscore) = 50.91 – 46 = 4.91

(Ionscore – Identityscore) not allowed to exceed
0.53 * (Ionscore – Homologyscore) – 6.99 + 10,
or 0.53 * (50.91 – 37) – 6.99 + 10 = 10.38

F = 0.1 * {4.91 + 47} – 3.0 = 2.19

Discriminant Score Distributions

no. of spectra

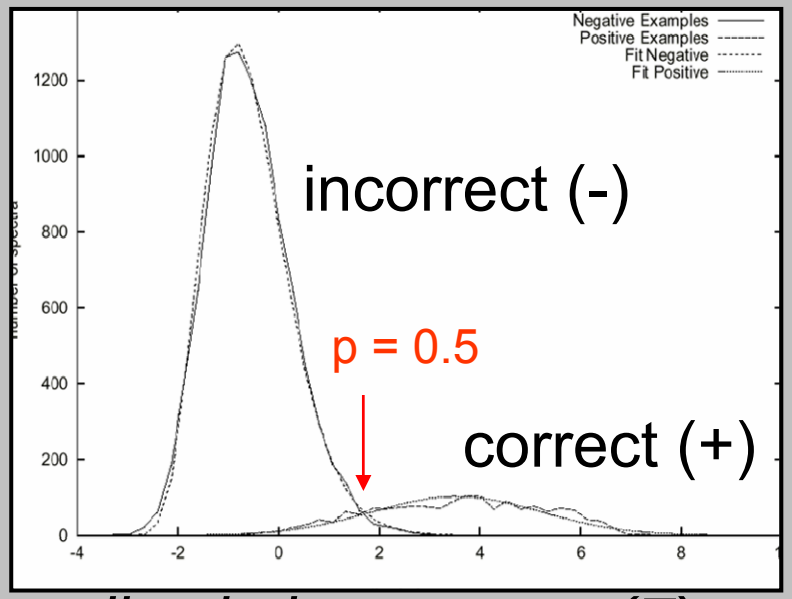


discriminant score (F)

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

Computing probabilities from discriminant score distributions

no. of spectra



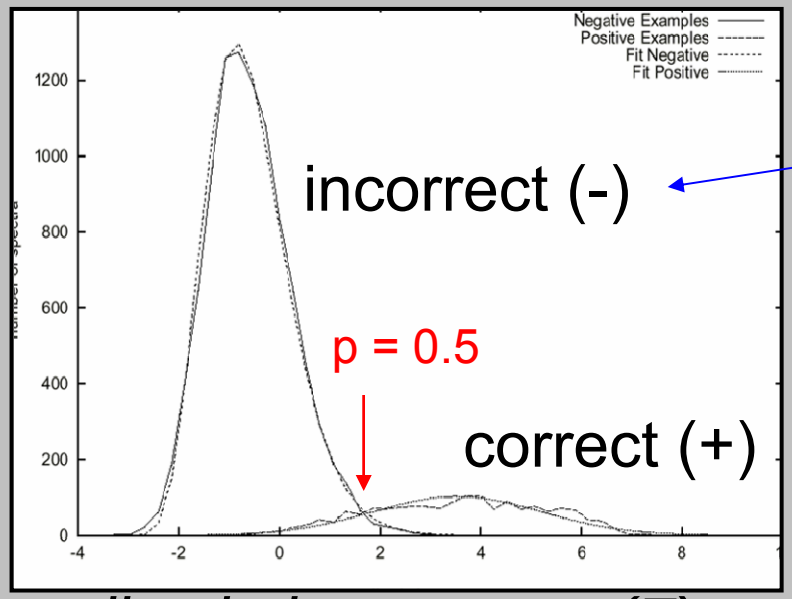
discriminant score (F)

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 ***Extreme Value Distribution (EVD)***

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

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

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

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 Mascot searches with 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:

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

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:

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

F_{obs}: p = 0.5 without NTT becomes p=0.04 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

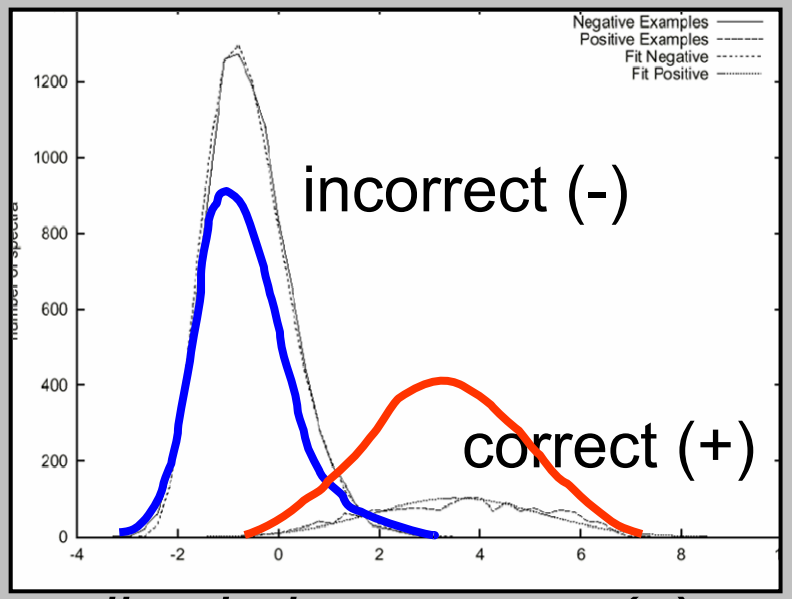
These distributions are expected to vary depending upon:

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

Variations in Discriminant Score Distributions

Different proportion of correct results in dataset

no of spectra



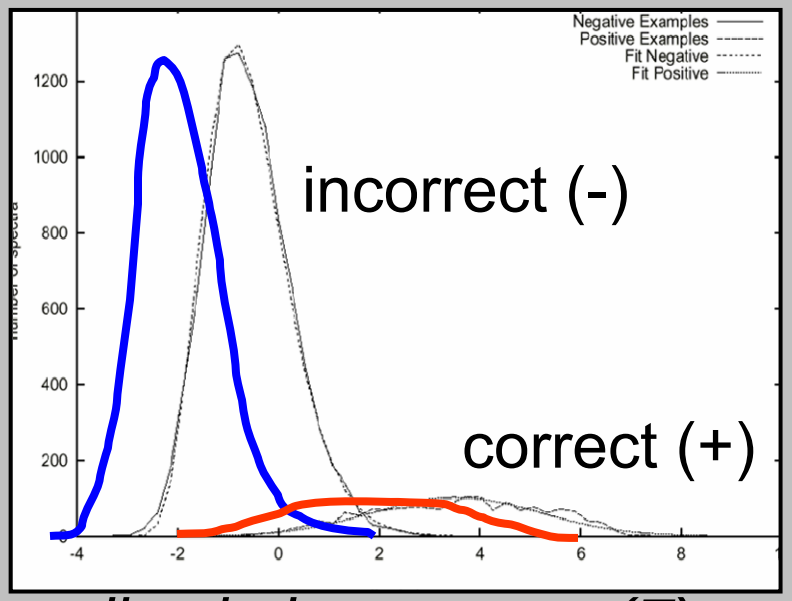
discriminant score (F)

vs. training dataset $[M+2H]^{2+}$ spectra

Variations in Discriminant Score Distributions

Different distribution means

no of spectra



discriminant score (F)

vs. training dataset $[M+2H]^{2+}$ spectra

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
- Expectation-Maximization (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

EM Algorithm Details

1. Initial estimates of result probabilities

Search Result	*F*	*NTT*	*prob*	*1-prob*
A	3.0	2	1.0	0.0
B	2.0	1	0.5	0.5
C	1.0	1	0.5	0.5
D	0.0	0	0.0	1.0

2. Update F value distributions among correct and incorrect results

$$P(F|+):\ m = \frac{(3.0)(1.0) + (2.0)(0.5) + (1.0)(0.5) + (0.0)(0.0)}{1.0 + 0.5 + 0.5 + 0.0} = 2.25$$

$$P(F|-):\ m = \frac{(3.0)(0.0) + (2.0)(0.5) + (1.0)(0.5) + (0.0)(1.0)}{0.0 + 0.5 + 0.5 + 1.0} = 0.75$$

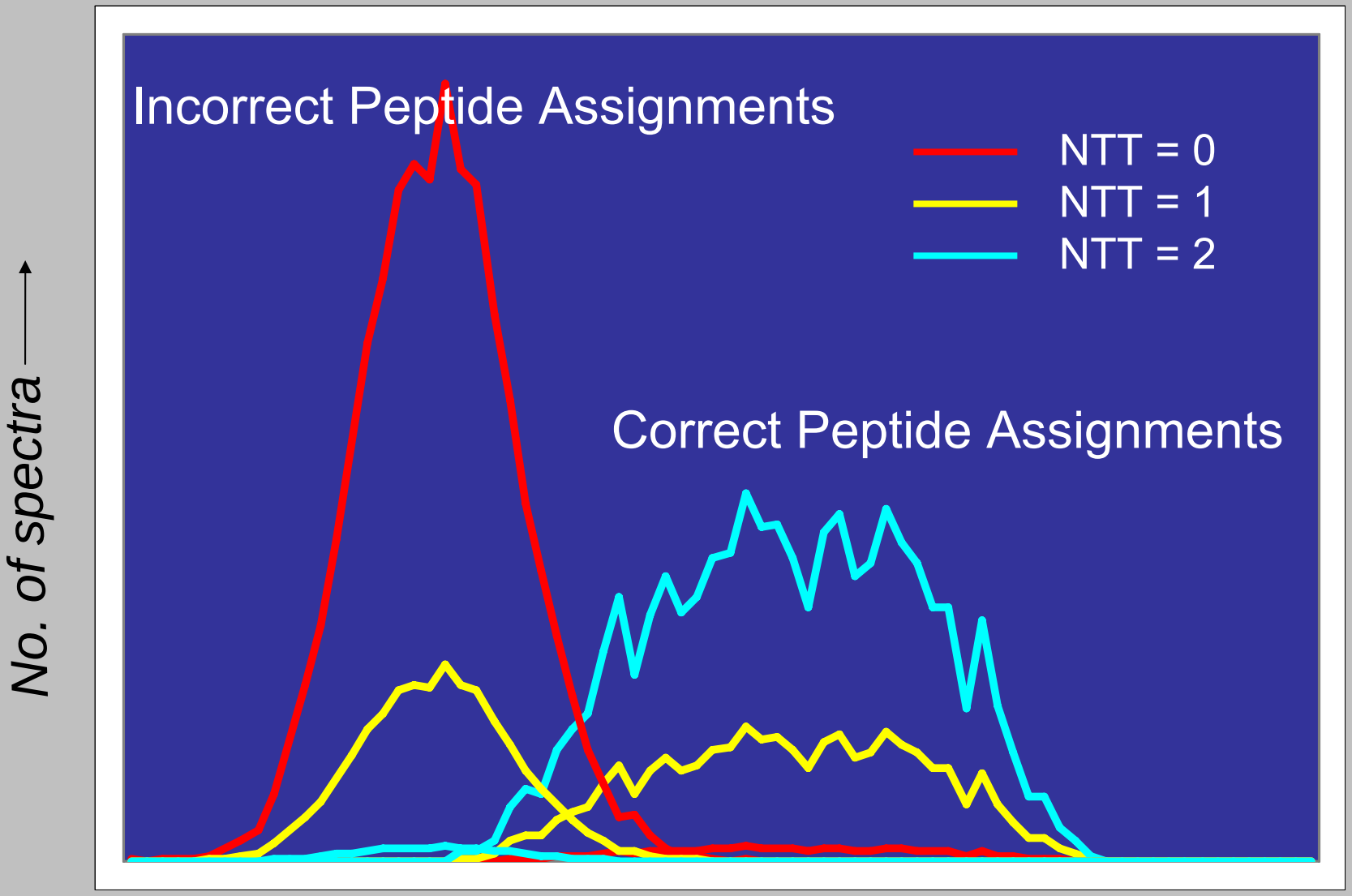
3. Update NTT distributions among correct and incorrect results

P(NTT=0|+) = (0.0) / (1.0 + 0.5 + 0.5 + 0.0) = 0.0
P(NTT=1|+) = (0.5 + 0.5) / (1.0 + 0.5 + 0.5 + 0.0) = 0.5
P(NTT=2|+) = (1.0)/(1.0 + 0.5 + 0.5 + 0.0) = 0.5

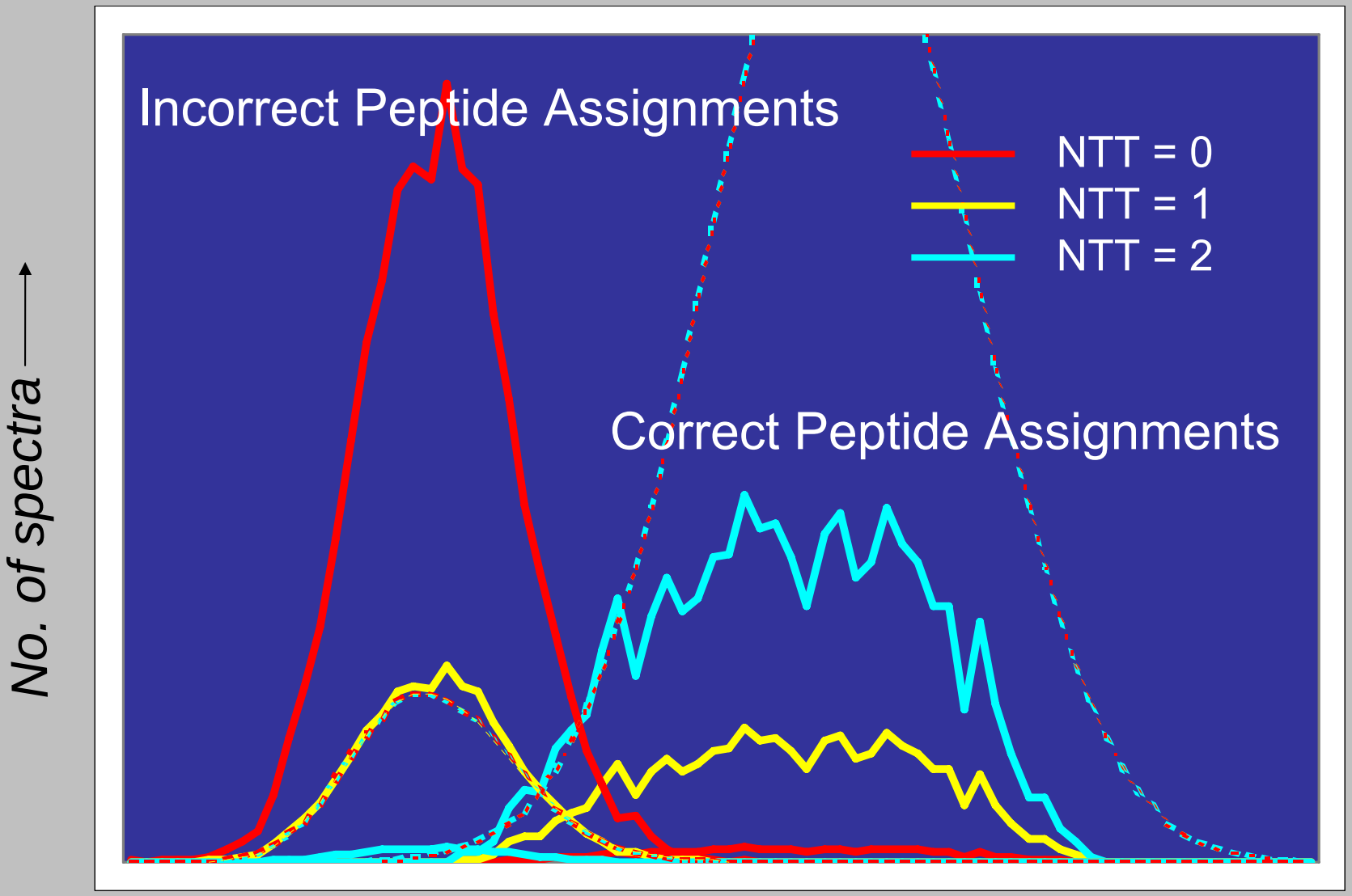
P(NTT=0|-) = (1.0) / (0.0 + 0.5 + 0.5 + 1.0) = 0.5
P(NTT=1|-) = (0.5 + 0.5) / (0.0 + 0.5 + 0.5 + 1.0) = 0.5
P(NTT=2|-) = (0.0)/(0.0 + 0.5 + 0.5 + 1.0) = 0.0

4. Recompute result probabilities using updated distributions, and iterate

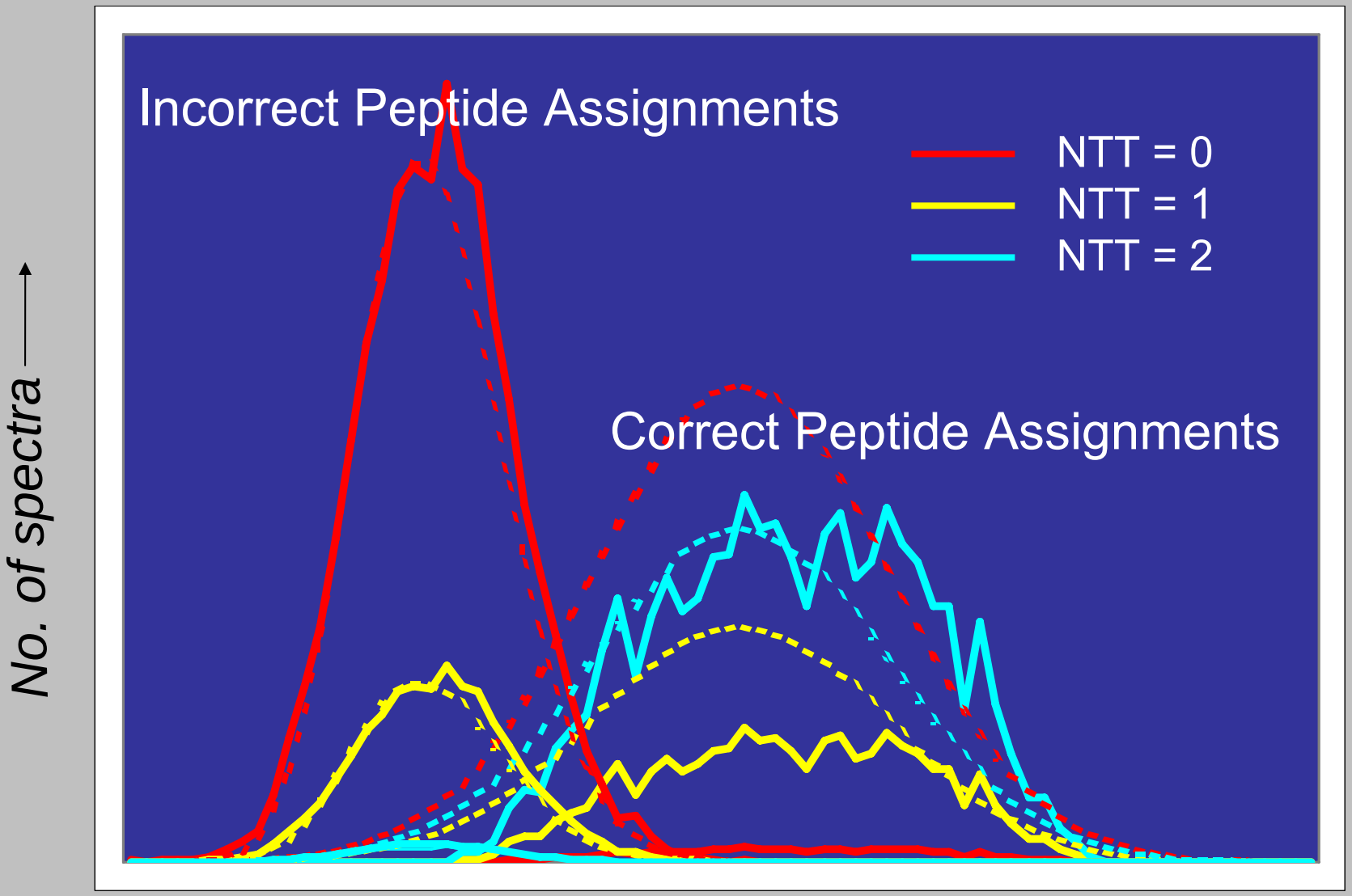
EM Algorithm learns test data score distributions



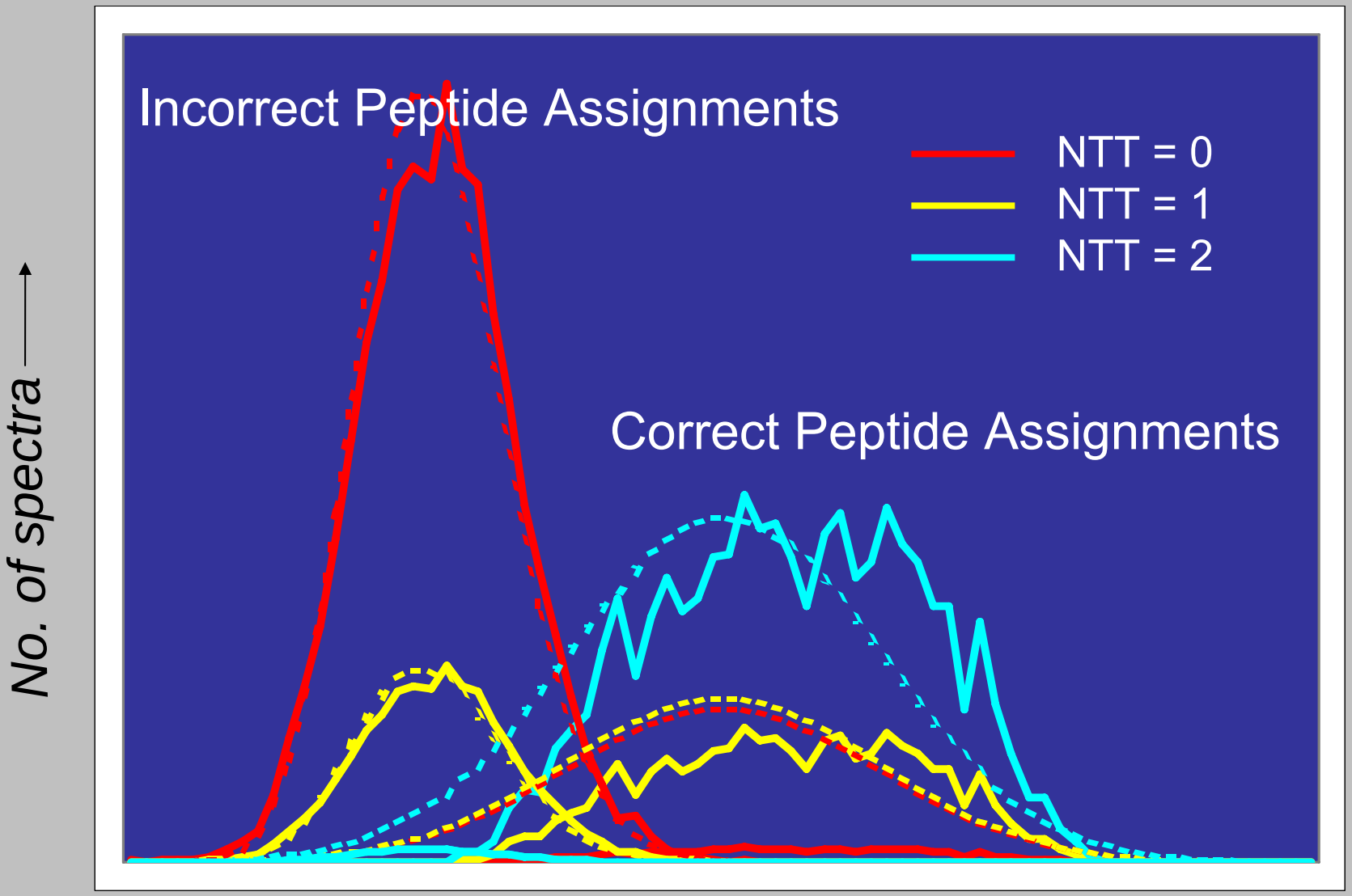
EM Iteration 0



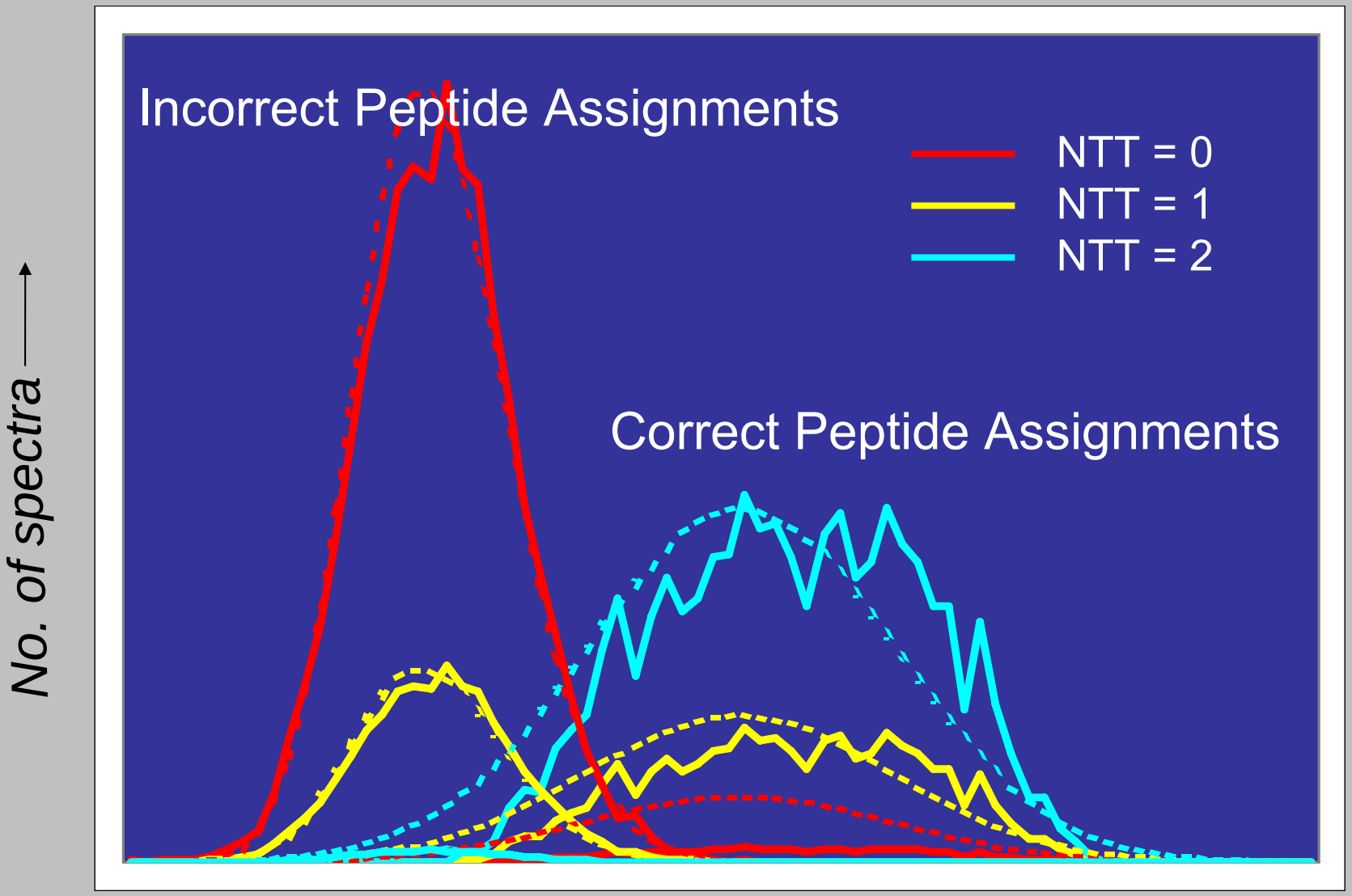
EM Iteration 1



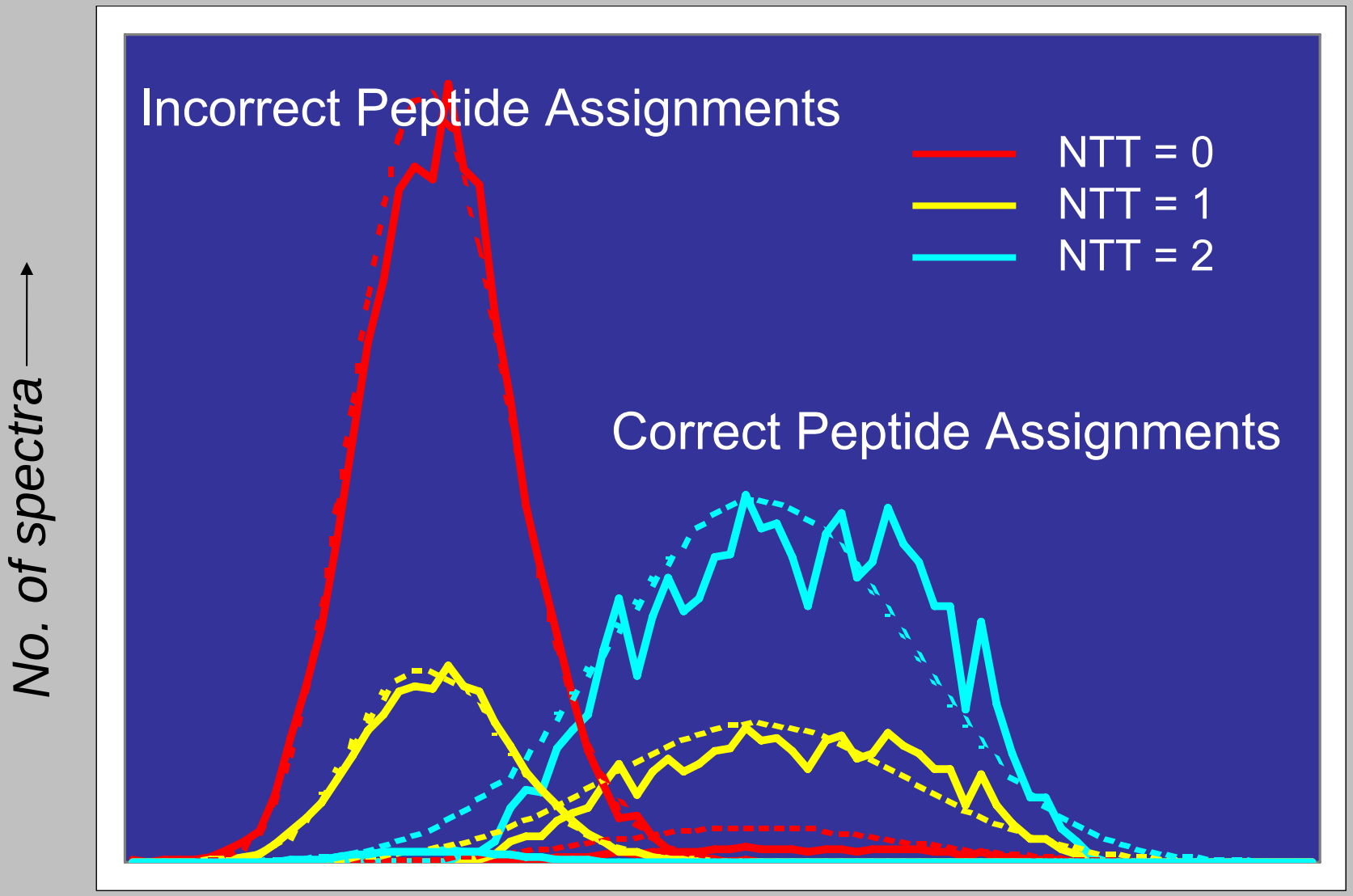
EM Iteration 2



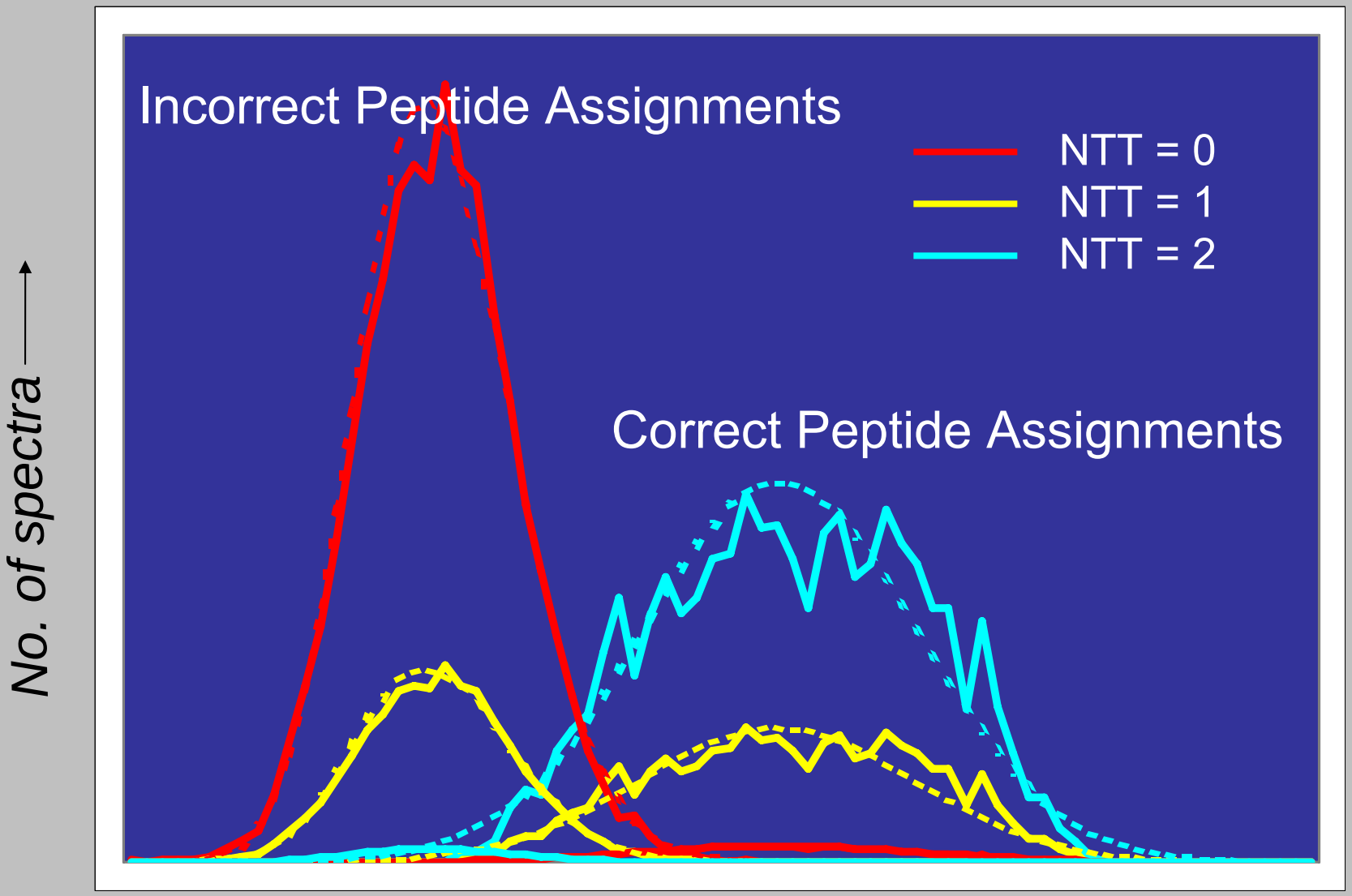
EM Iteration 3



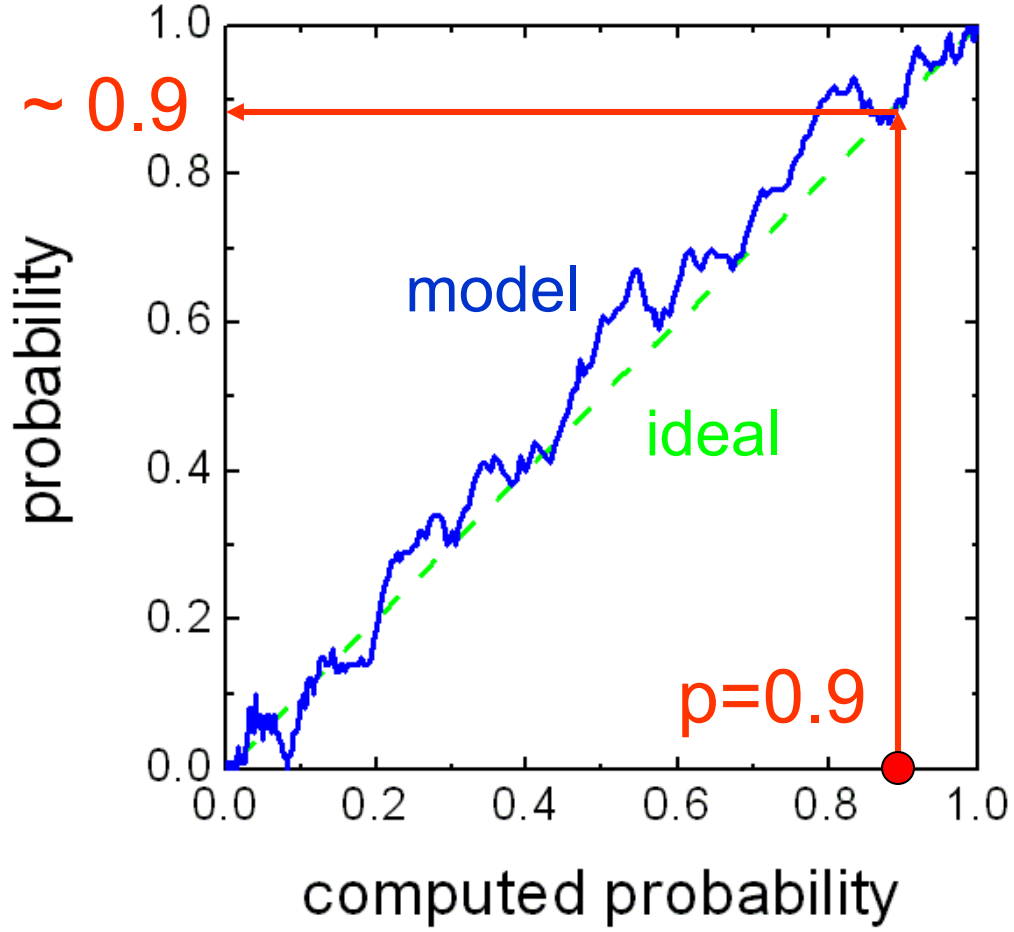
EM Iteration 4



EM Iteration 7



Accuracy of the Model



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

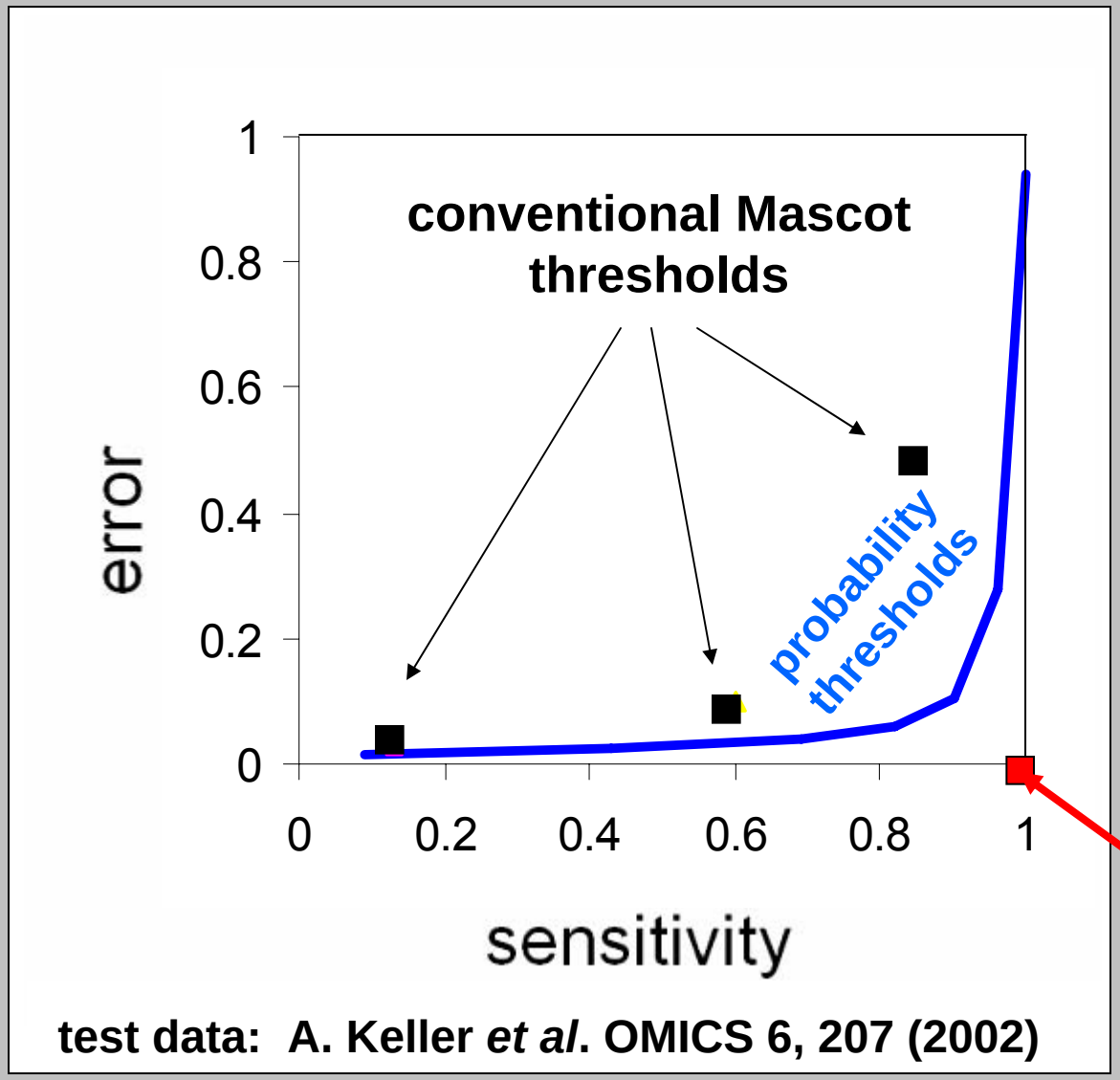
100 spectra with computed p ~ 0.9

90% of them (90) should be correct

Observed probability is around 0.9

Model is accurate

Discriminating Power of Computed Probabilities

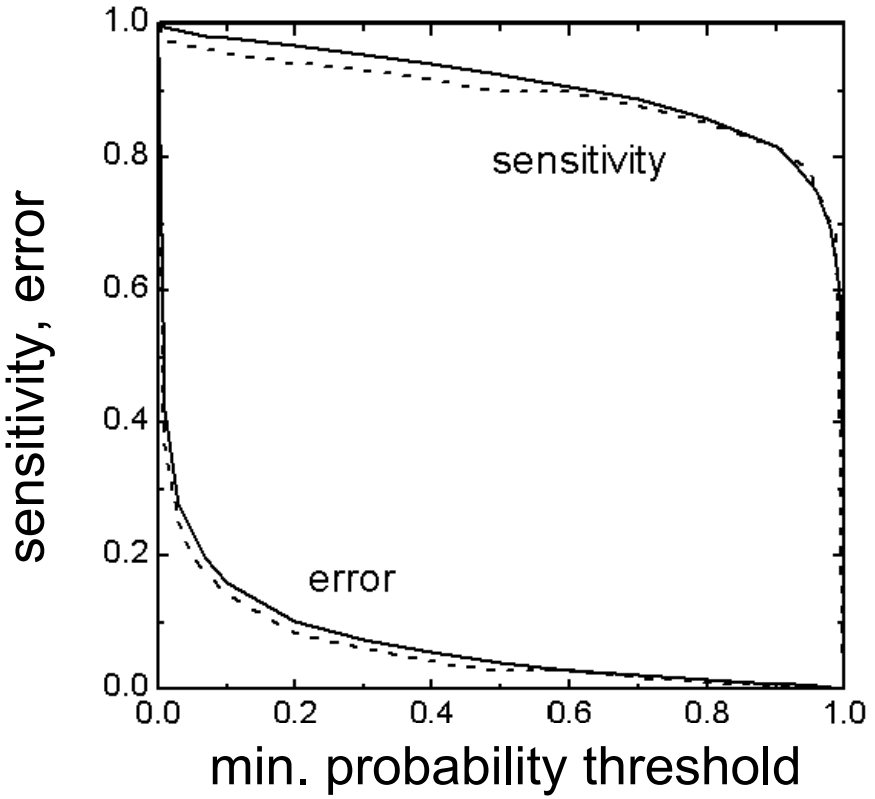


Sensitivity: fraction of all correct results passing filter

Error: fraction of all results passing filter that are incorrect

Ideal Spot

Discriminating Power of Computed Probabilities

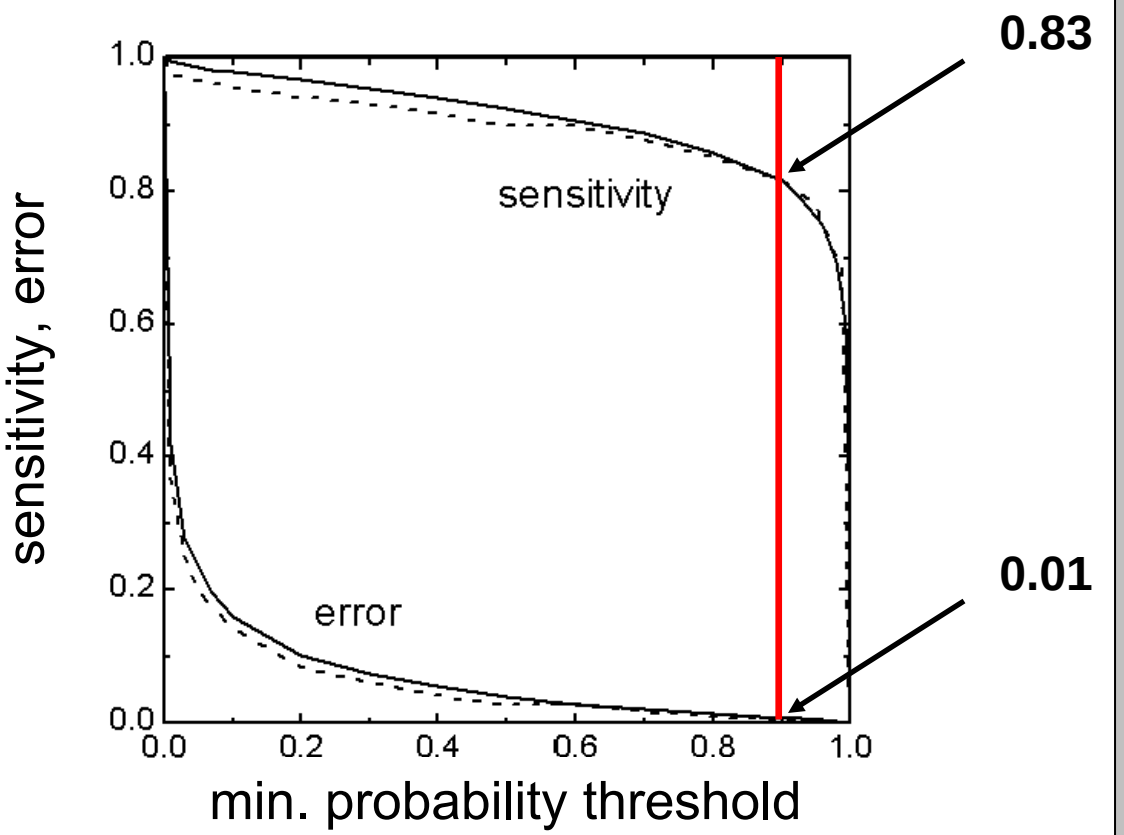


test data: A. Keller *et al*. OMICS 6, 207 (2002)

Sensitivity:
fraction of all correct results passing filter

Error:
fraction of all results passing filter that are incorrect

Discriminating Power Example: p ≥ 0.9

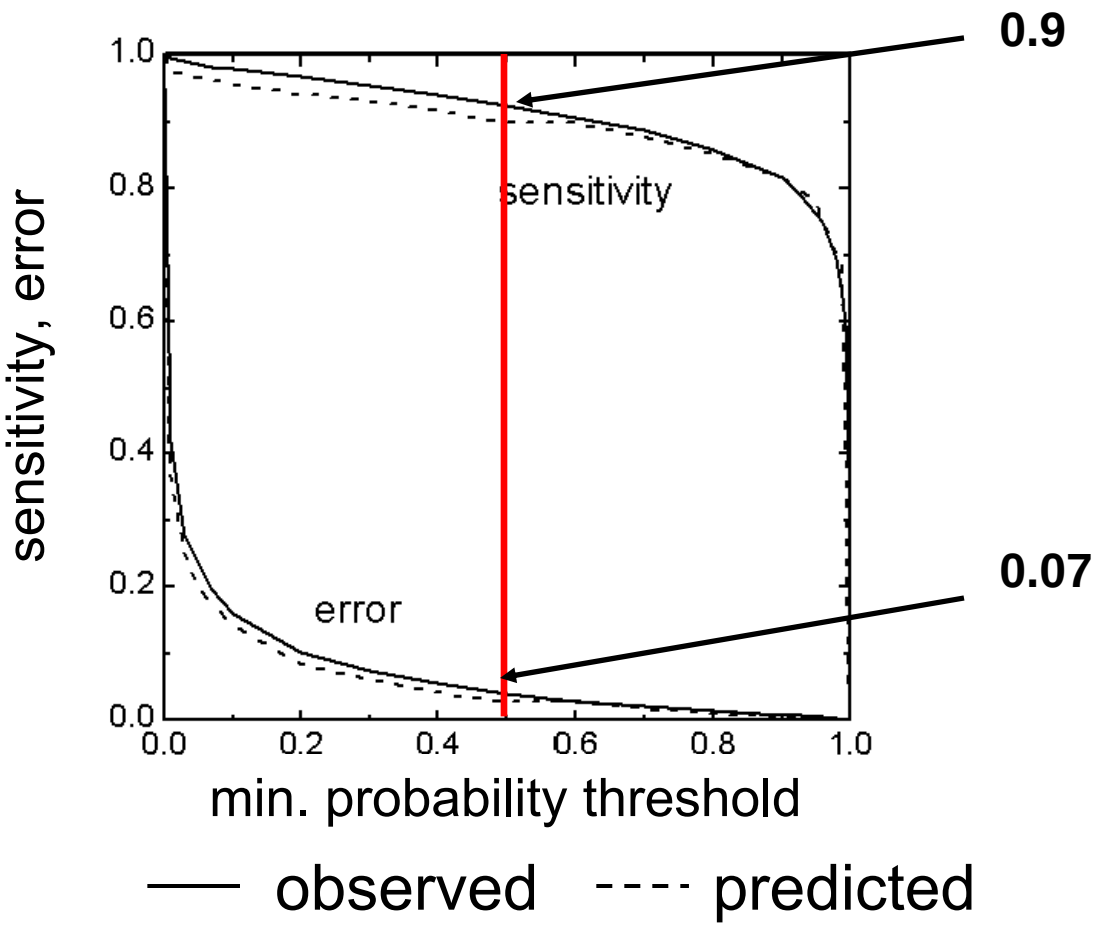


test data: A. Keller *et al*. OMICS 6, 207 (2002)

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

Use of PeptideProphet Probabilities to Compare Searches

- False positive error rate predicted by PeptideProphet is an objective criterion for comparing different searches
  - Sample preparation and LC/MS/MS
  - Search conditions
  - Search engine
- Compare the number of results of each search passing its minimum probability threshold to achieve a fixed predicted false positive error rate
  - Reflects both search engine and PeptideProphet performance

From Peptide to Protein Level Analysis

- When the identification of proteins rather than peptides is of interest, it is unnecessary in practice to filter search results based on probabilities
- Instead, all search results and their computed probabilities are passed to the ProteinProphet program which infers sample proteins by combining together the peptide evidence for each protein
  - Initially adjusts the PeptideProphet probabilities based on whether a peptide corresponds to a single-hit or multi-hit protein
  - Then apportions shared peptides among all their corresponding proteins in such a way to derive the simplest list of proteins that explain the observed peptides
  - Computes accurate protein probabilities

PeptideProphet Software Tutorial

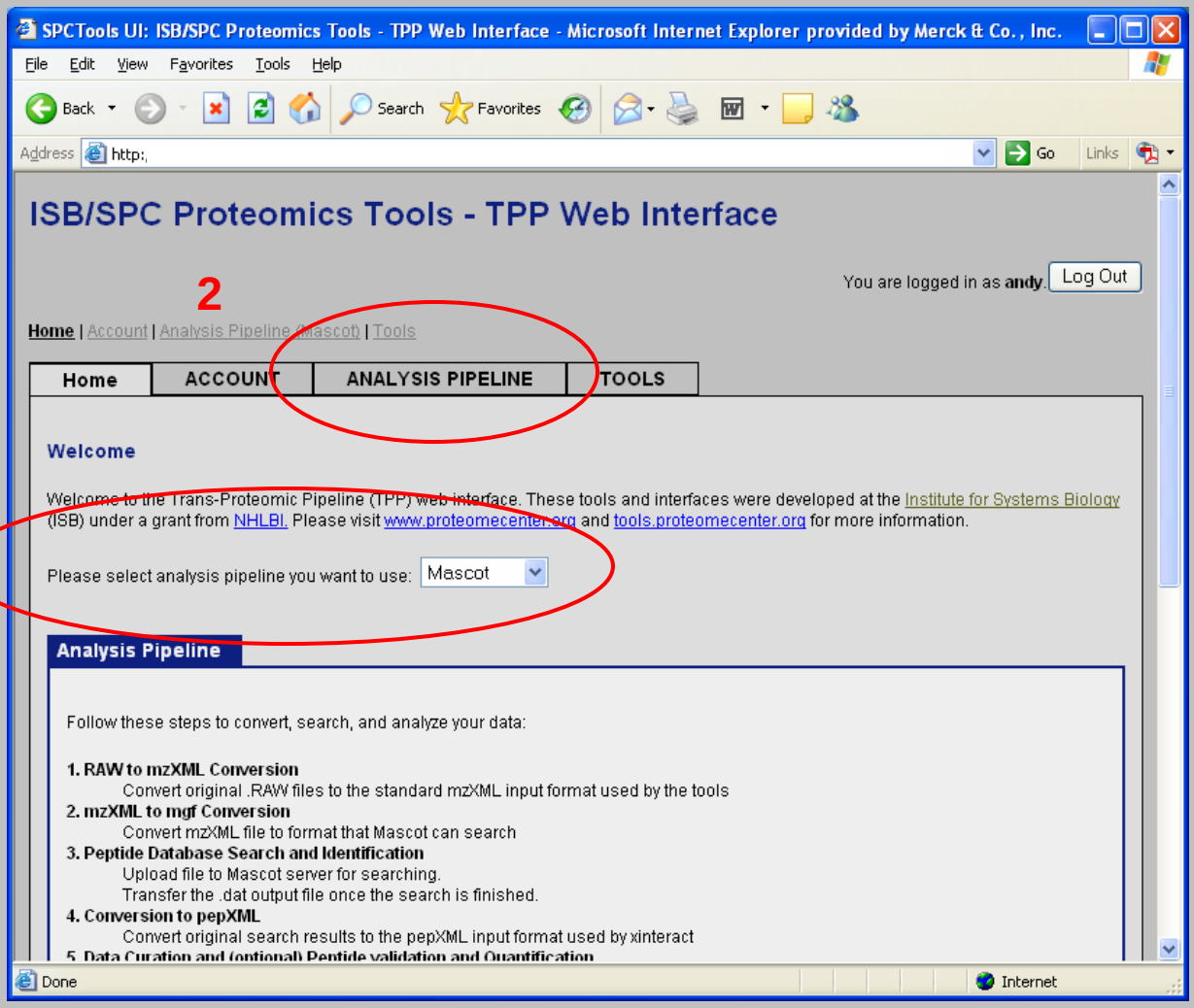
- How to run PeptideProphet through the TPP Web Interface
- Interpretation of analysis results
- User options

Getting started with PeptideProphet

- Input: pepXML files (**`file1.xml, file2.xml`**…)
- XInteract program first merges files together into single file **`interact.xml`**, then PeptideProphet runs model, computes probabilities, and writes probabilities as first column
- Combine together runs that are similar (sample, database, search constraints, mass spectrometer)

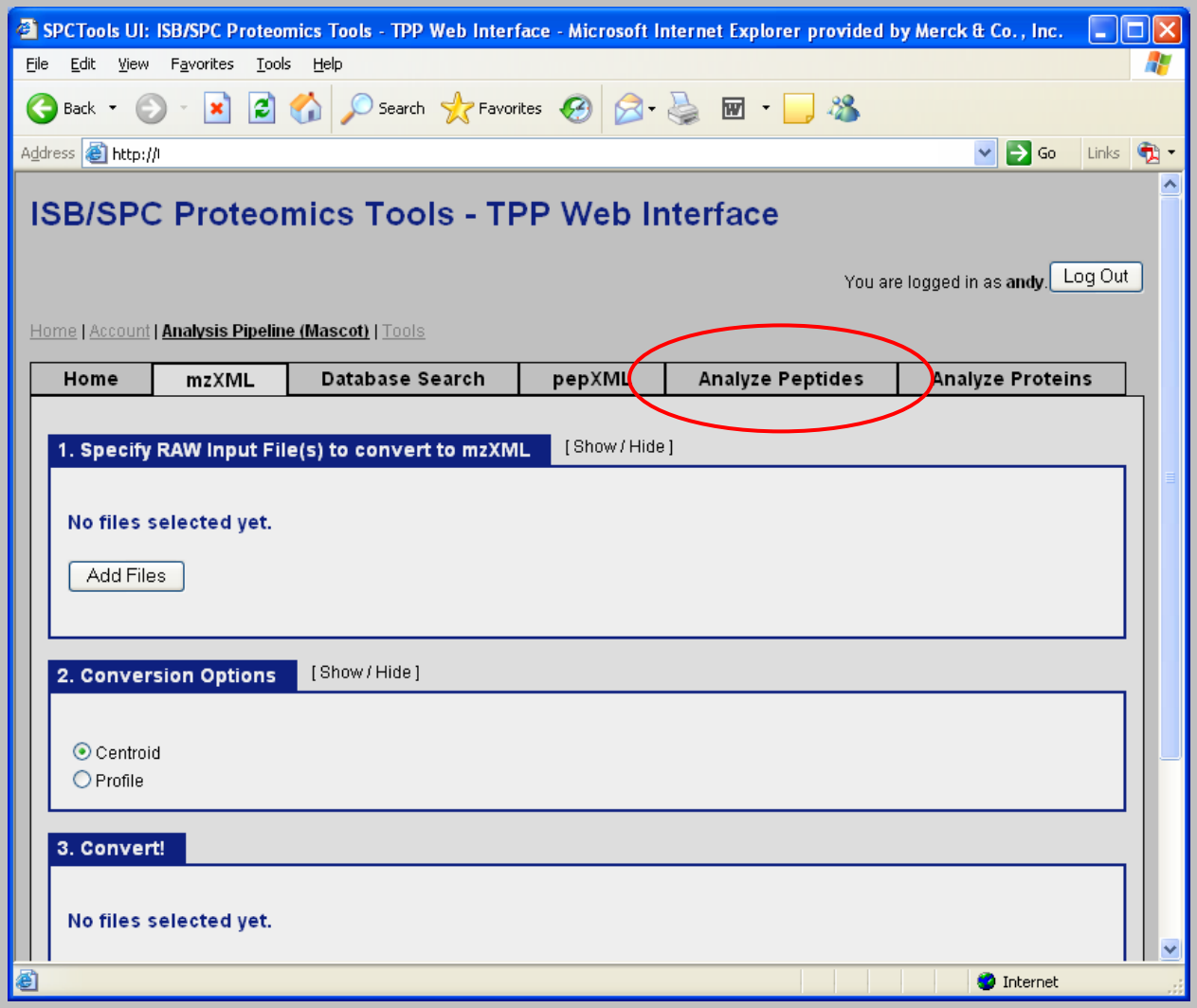
Getting started with PeptideProphet

Specify search engine and select Analysis Pipeline



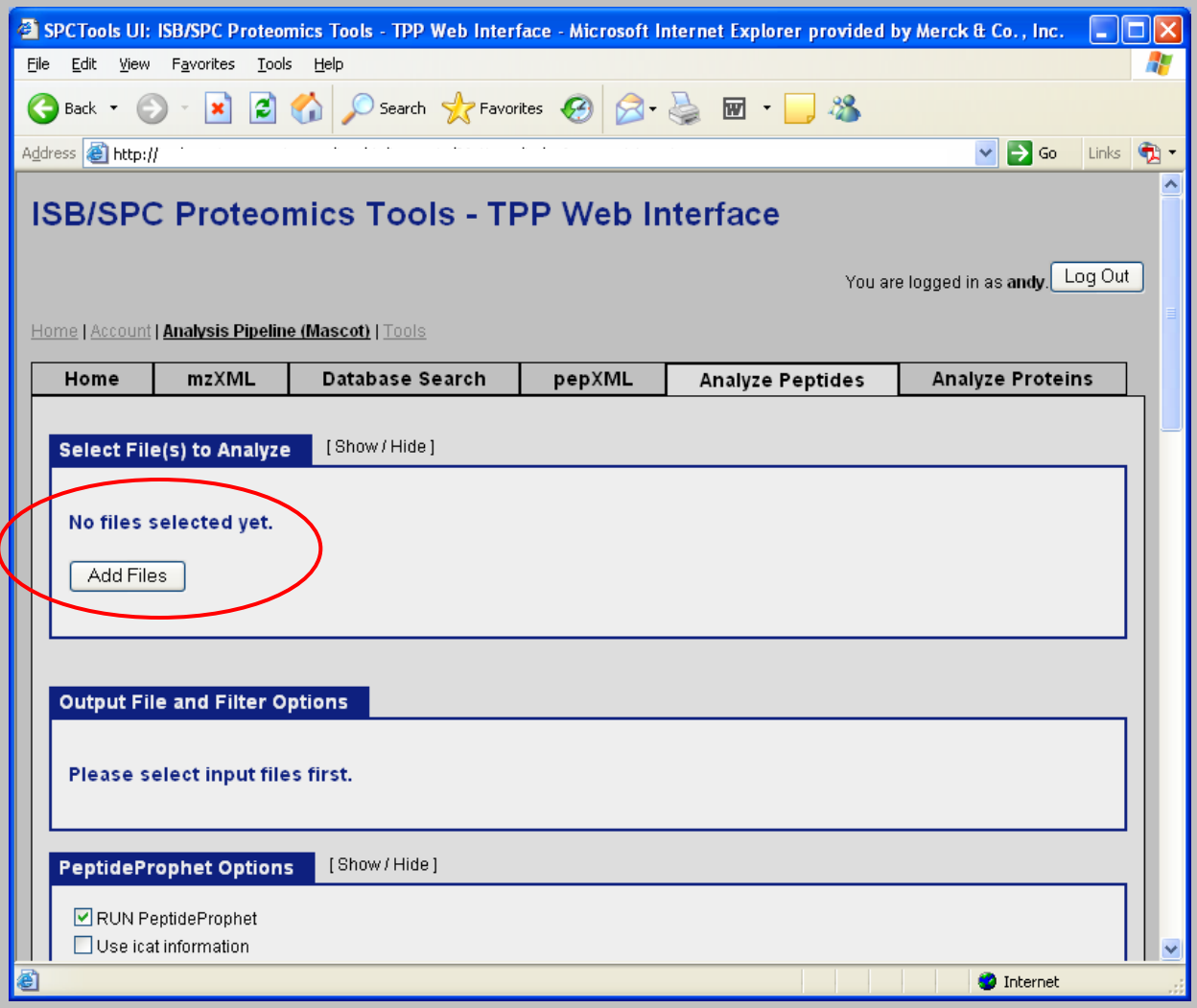
Getting started with PeptideProphet

Select peptide level analysis



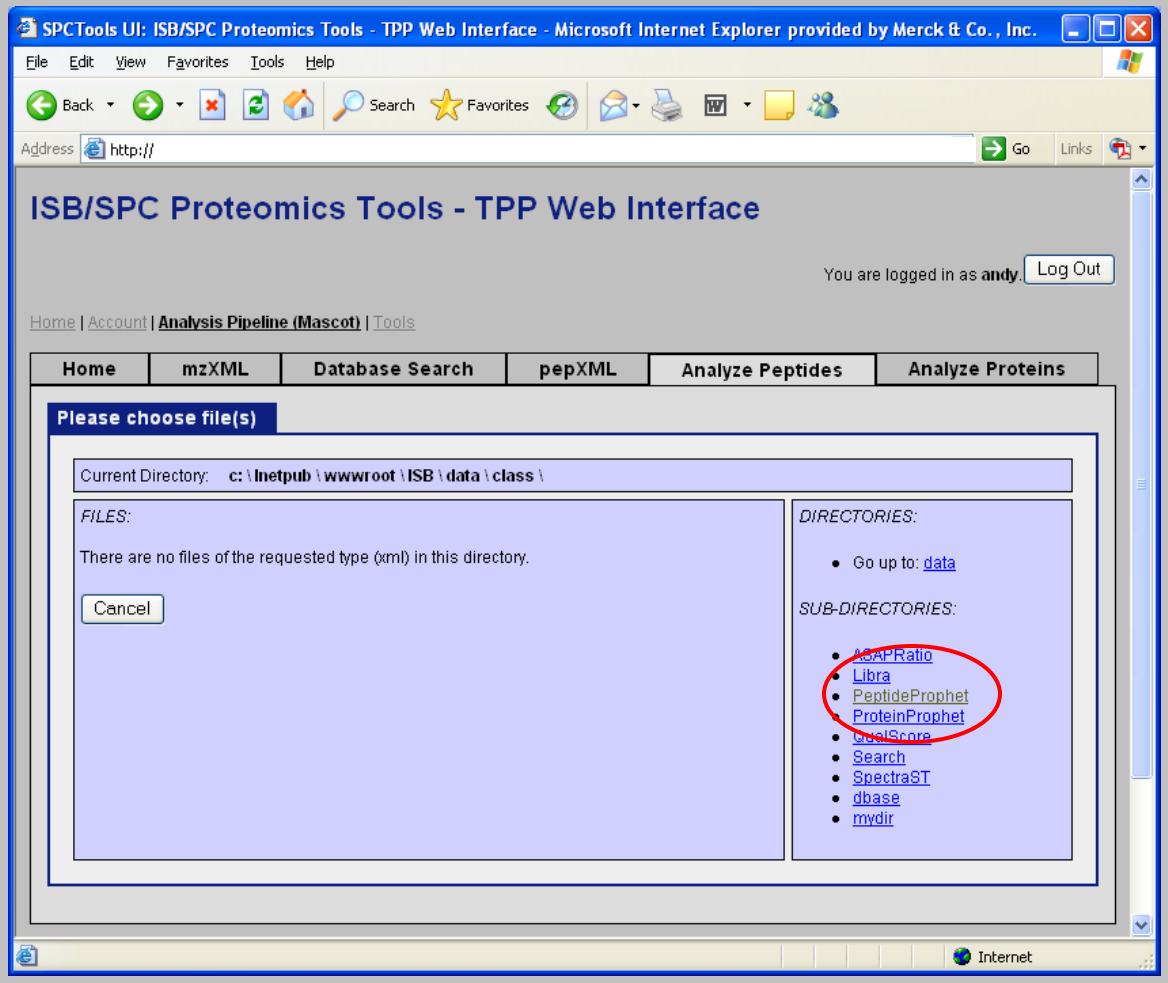
Getting started with PeptideProphet

Specify search results to analyze



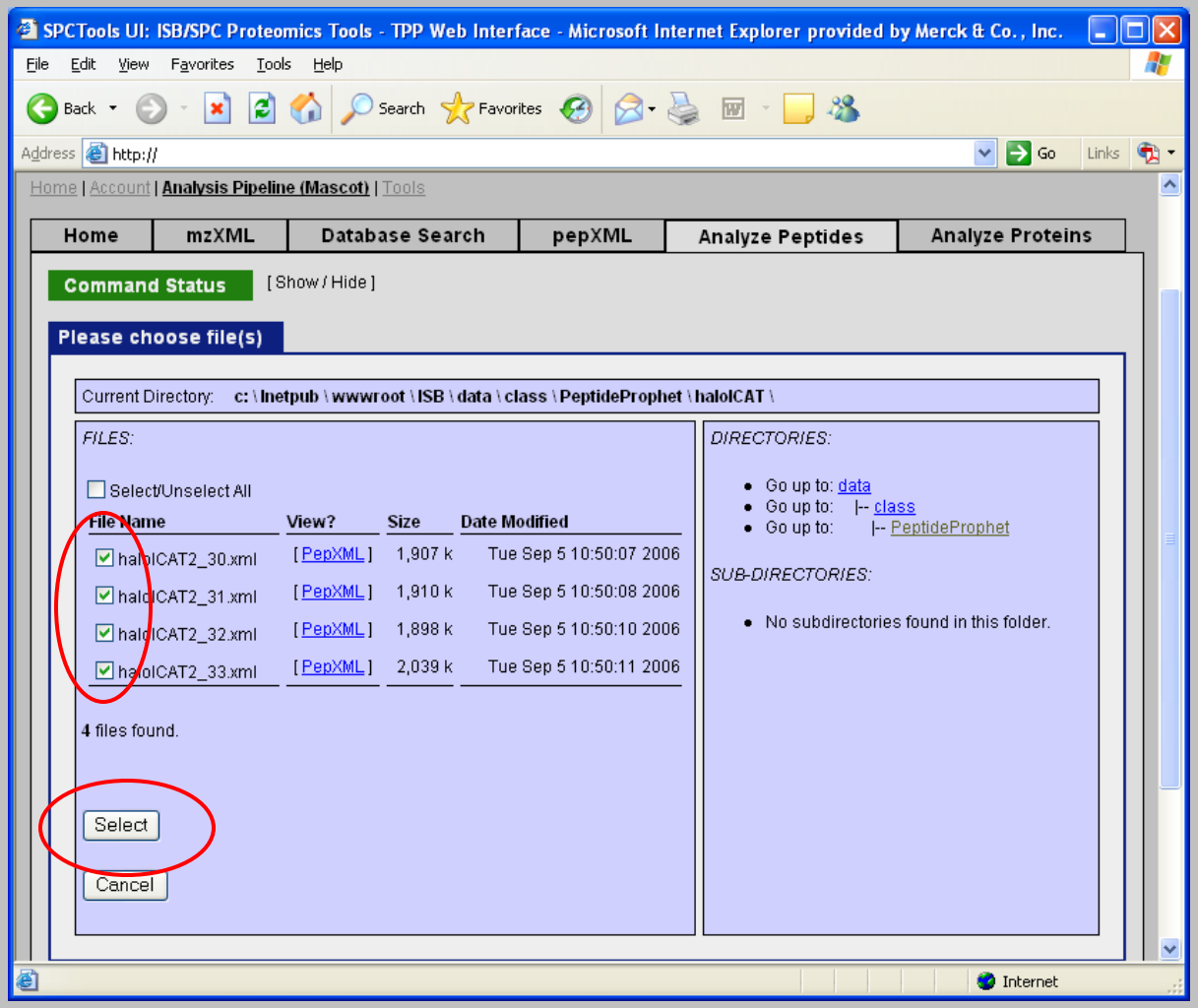
Getting started with PeptideProphet

Navigate data directories



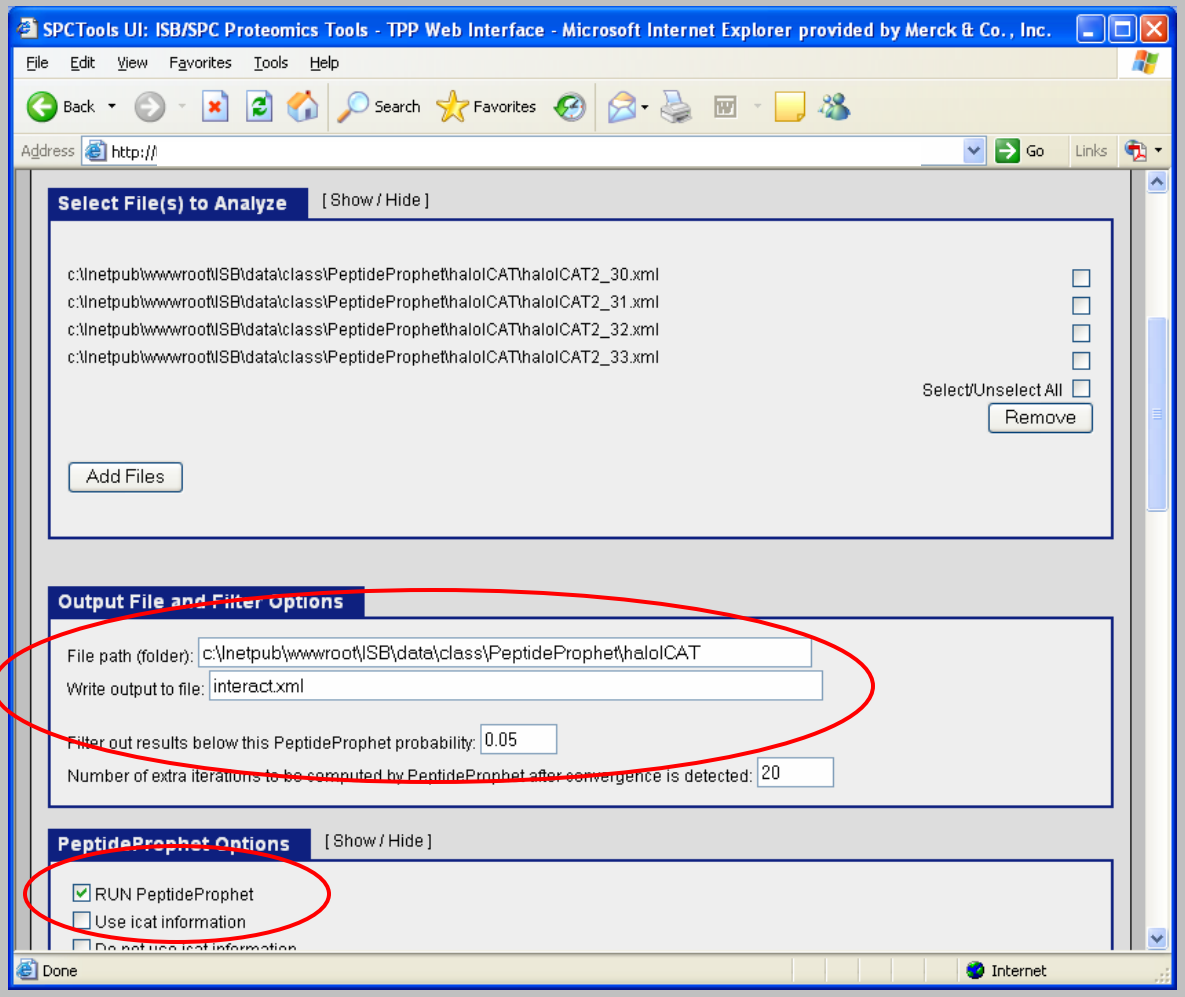
Getting started with PeptideProphet

Add each search run pepXML included in analysis



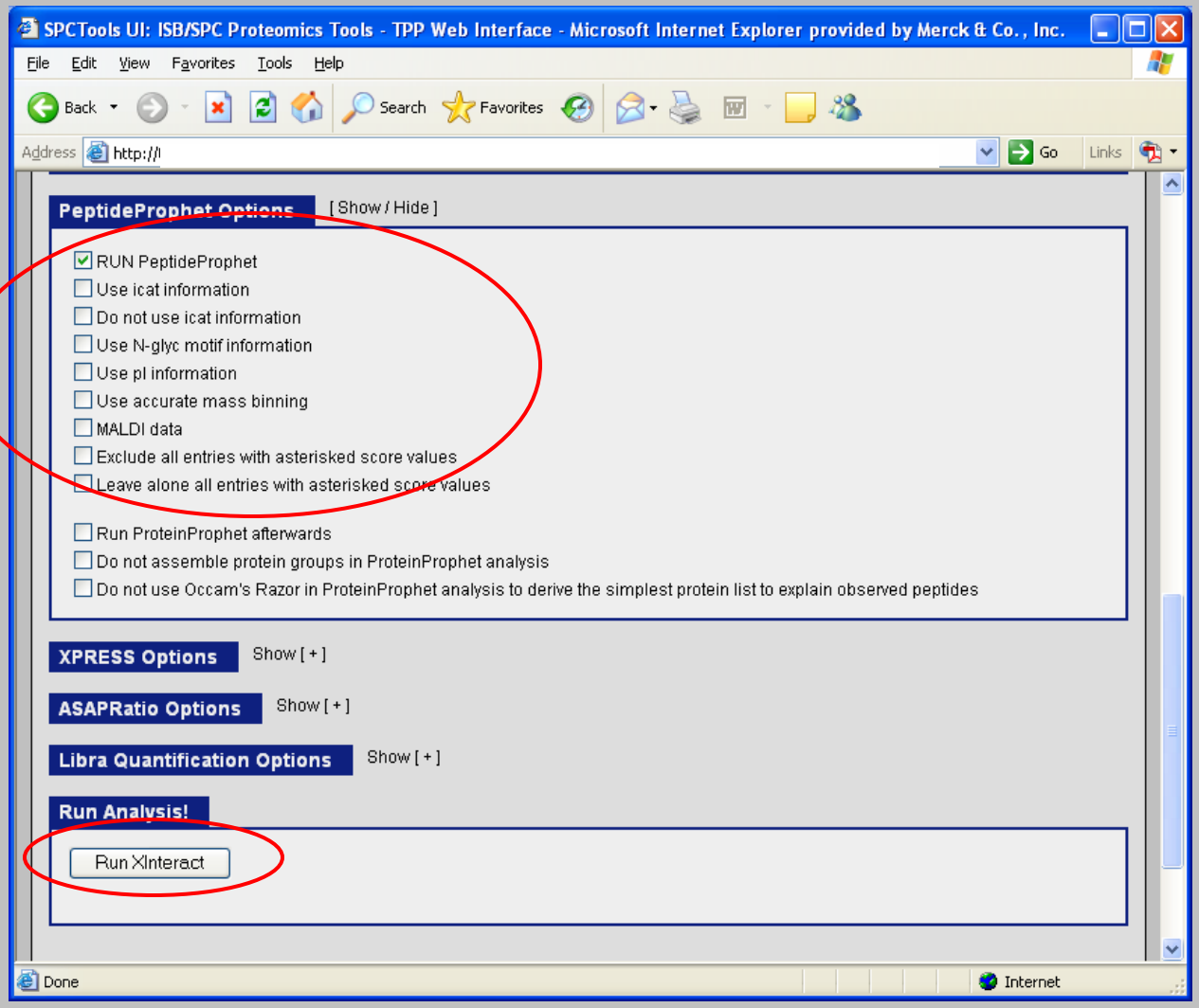
Getting started with PeptideProphet

Specify output file name and minimum probability filter, opt to run PeptideProphet



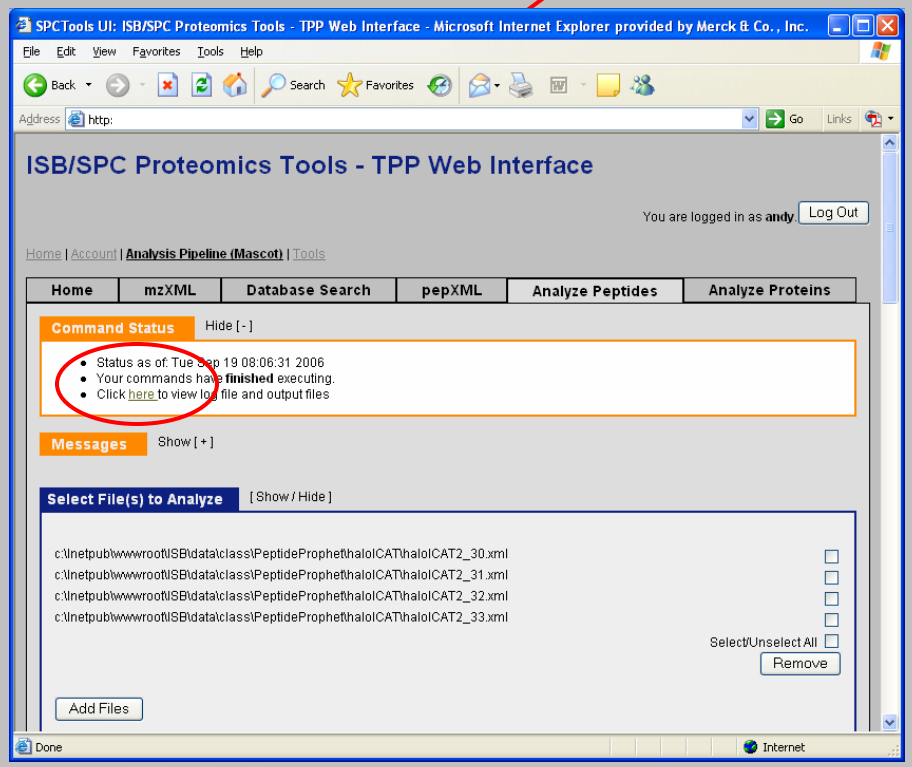
Getting started with PeptideProphet

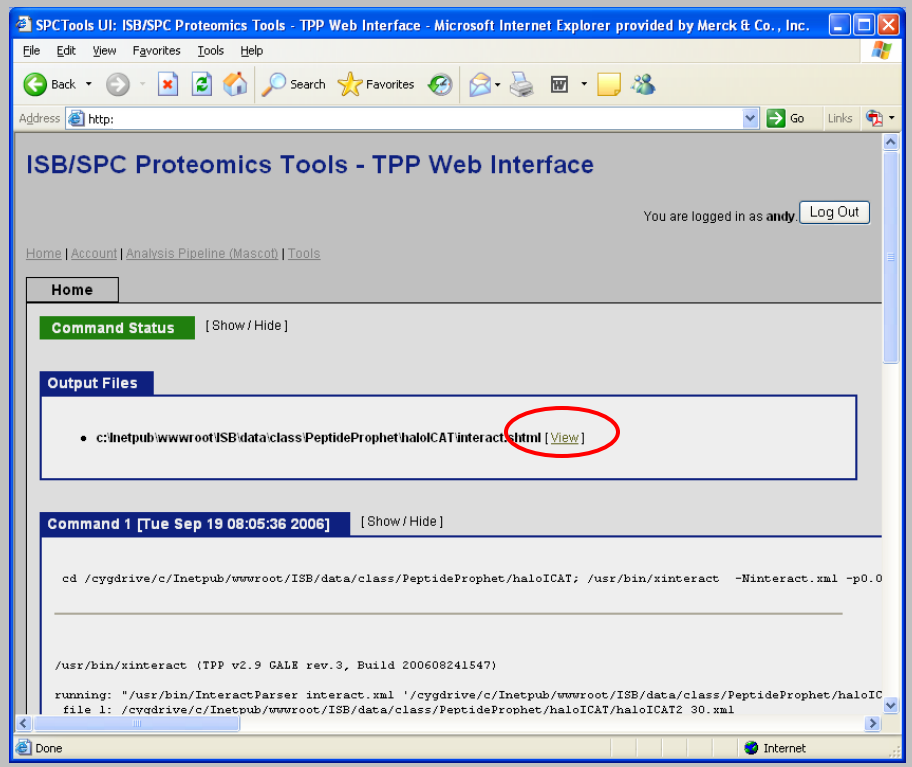
Specify PeptideProphet optional parameters and run analysis



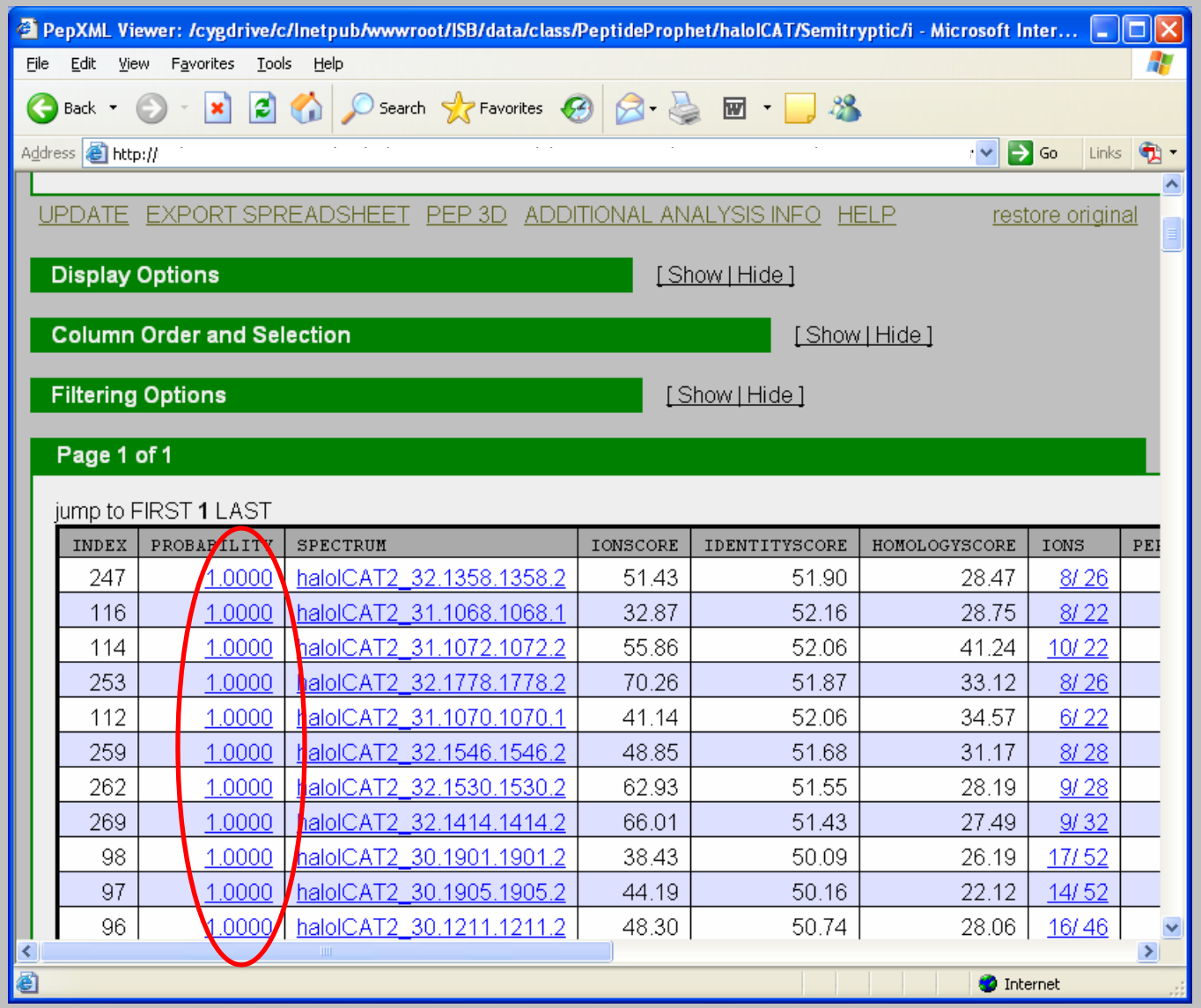
Getting started with PeptideProphet

Click on links to view results of analysis

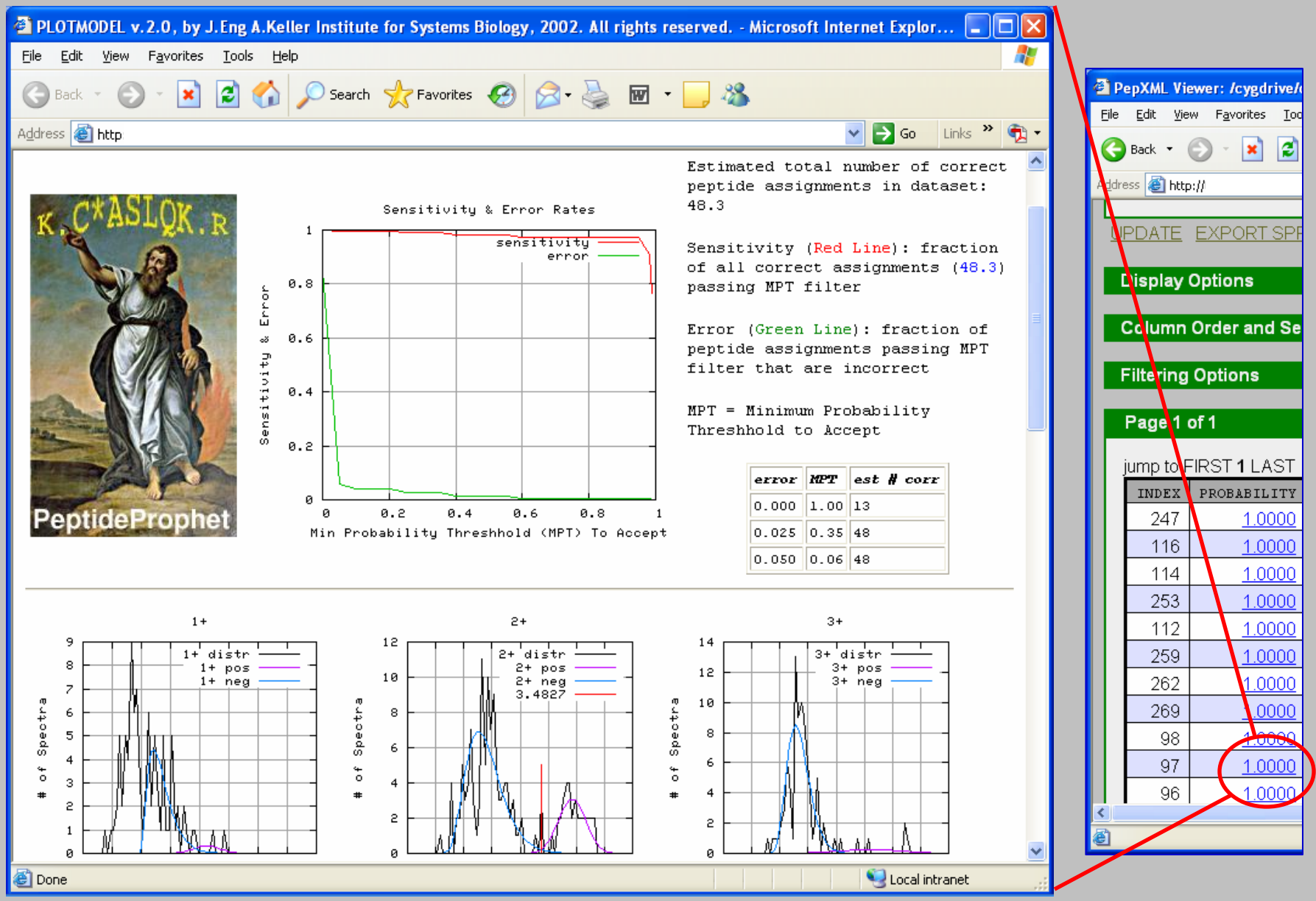




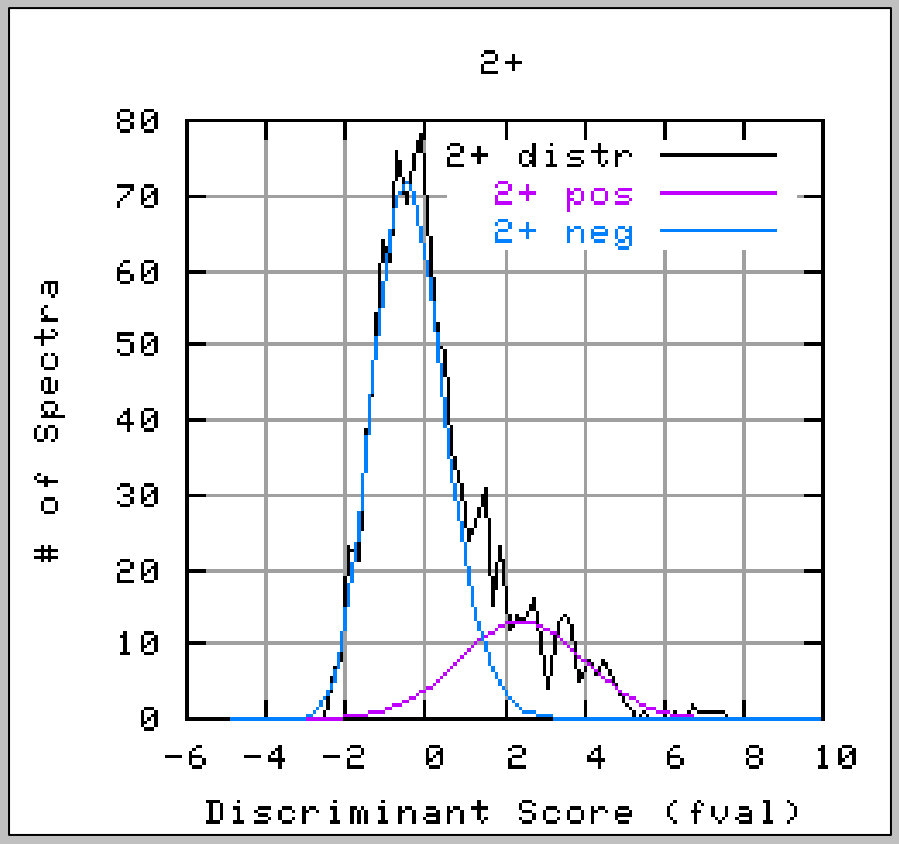
PeptideProphet Results

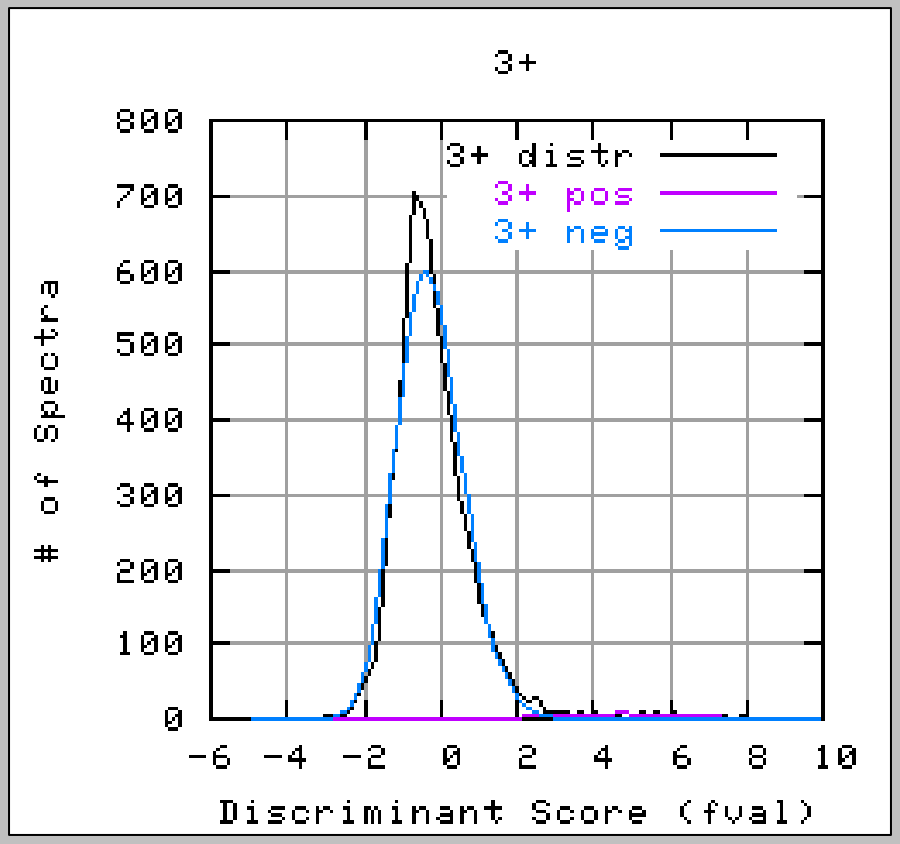


PeptideProphet Results: Model Summary

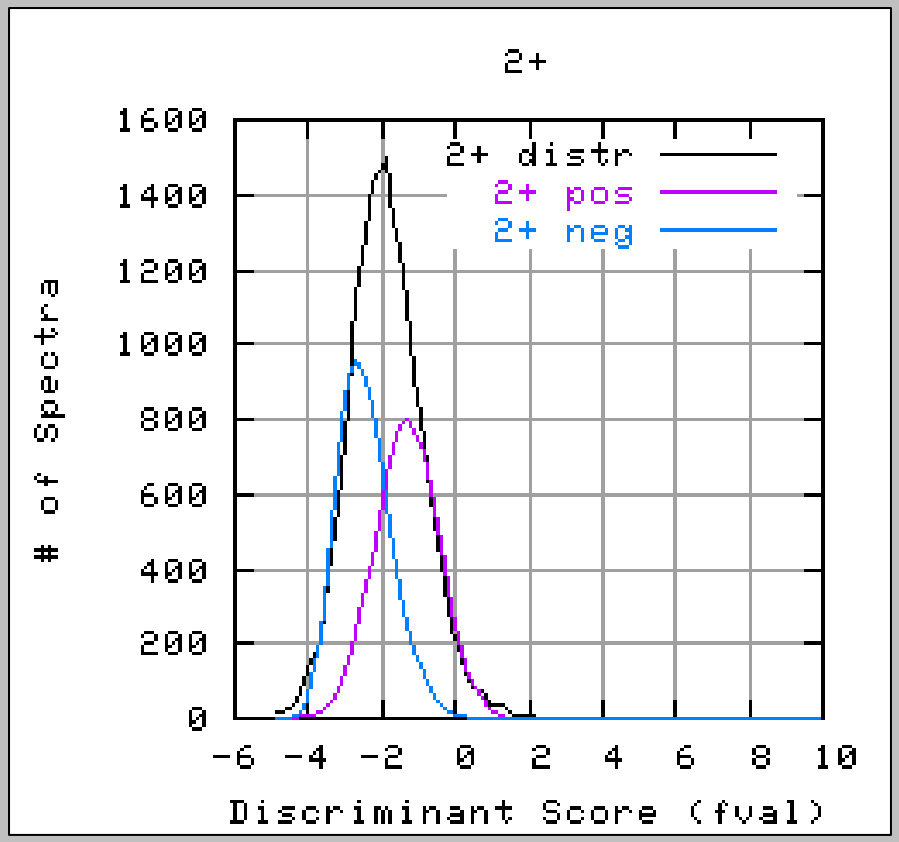


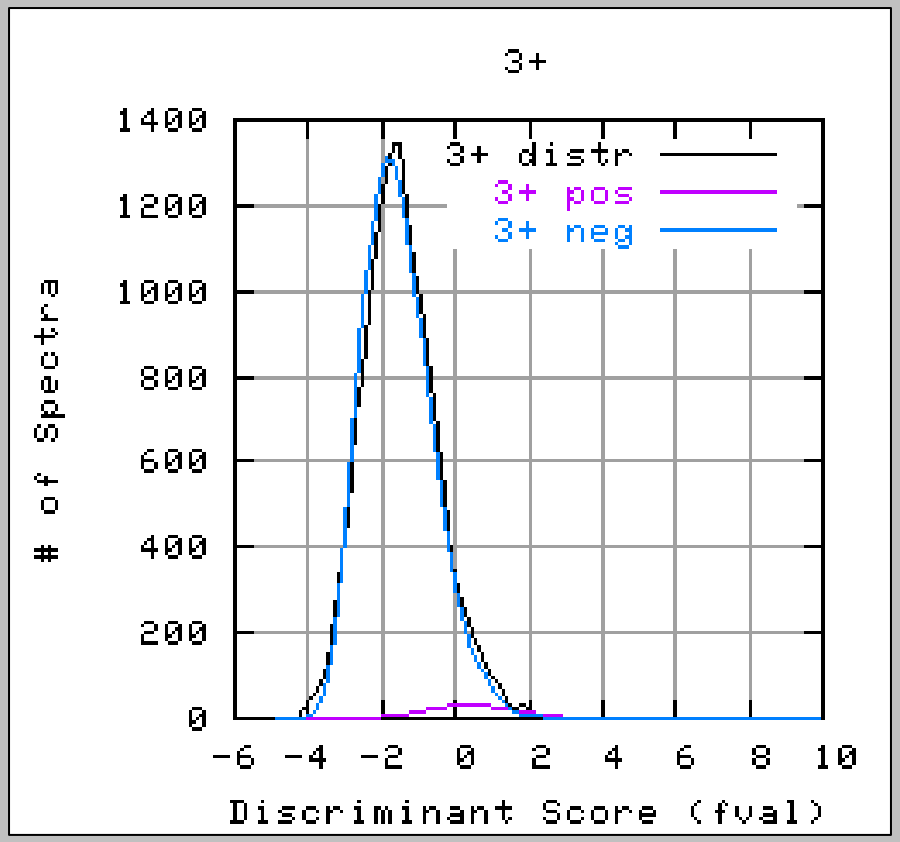
Reasonable Learned Discriminant Score Distributions





Suspicious Looking Learned Discriminant Score Distributions





PeptideProphet Results: Model Summary

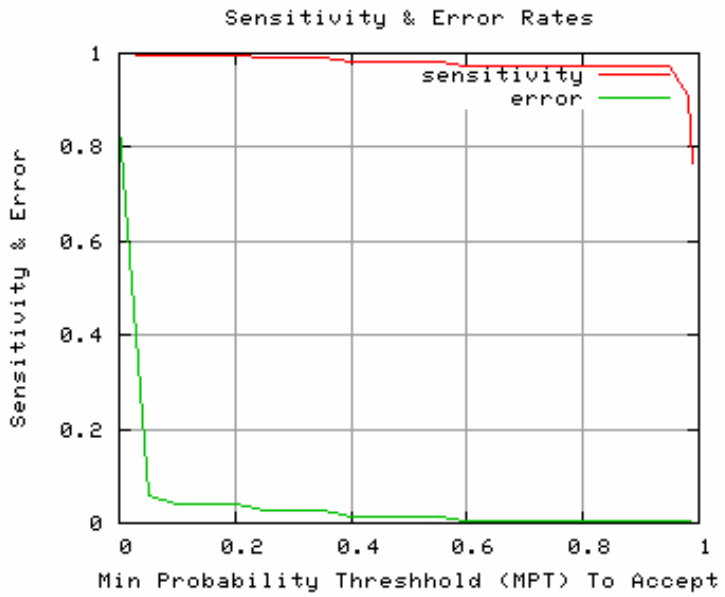
PLOTMODEL v.2.0, by J.Eng A.Keller Institute for Systems Biology, 2002. All rights reserved. - Microsoft Internet Explor...

File Edit View Favorites Tools Help

Back Search Favorites

Address http://

Go Links



Estimated total number of correct peptide assignments in dataset: 48.3

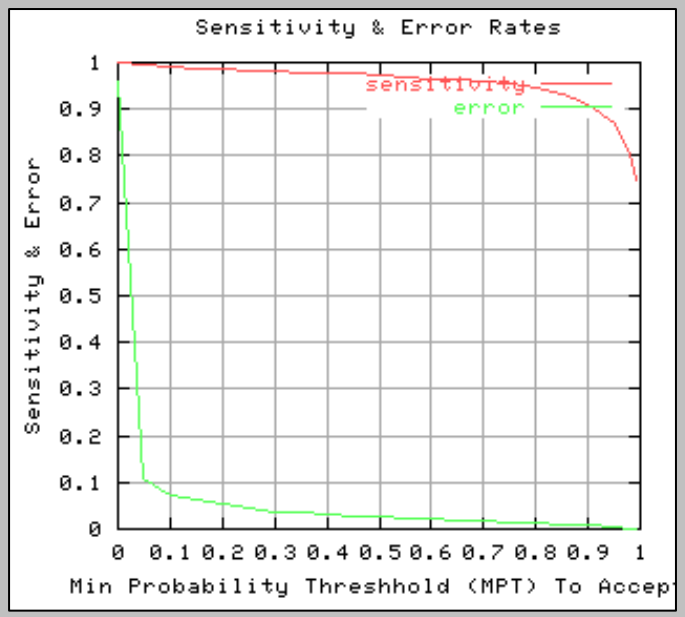
Sensitivity (Red Line): fraction of all correct assignments (48.3) passing MPT filter

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

MPT = Minimum Probability Threshhold to Accept

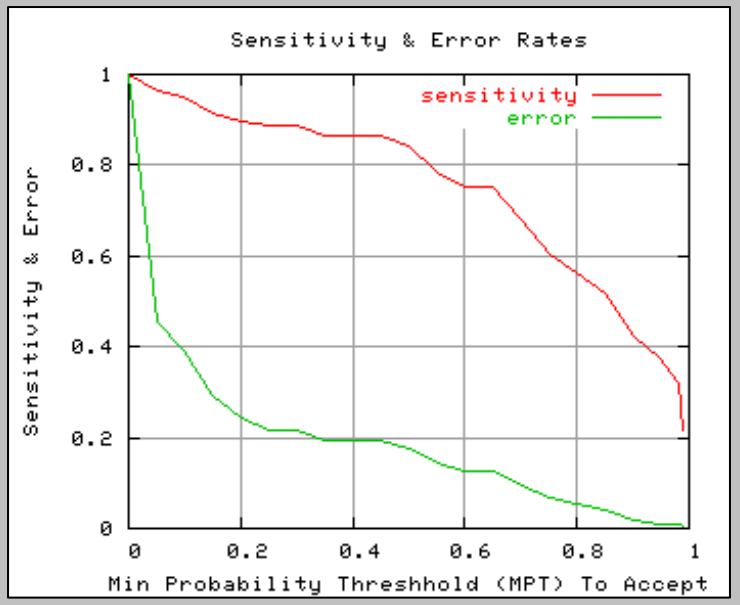
error	MPT	est # corr
0.000	1.00	13
0.025	0.35	48
0.050	0.06	48

PeptideProphet Results: Model Summary



Good

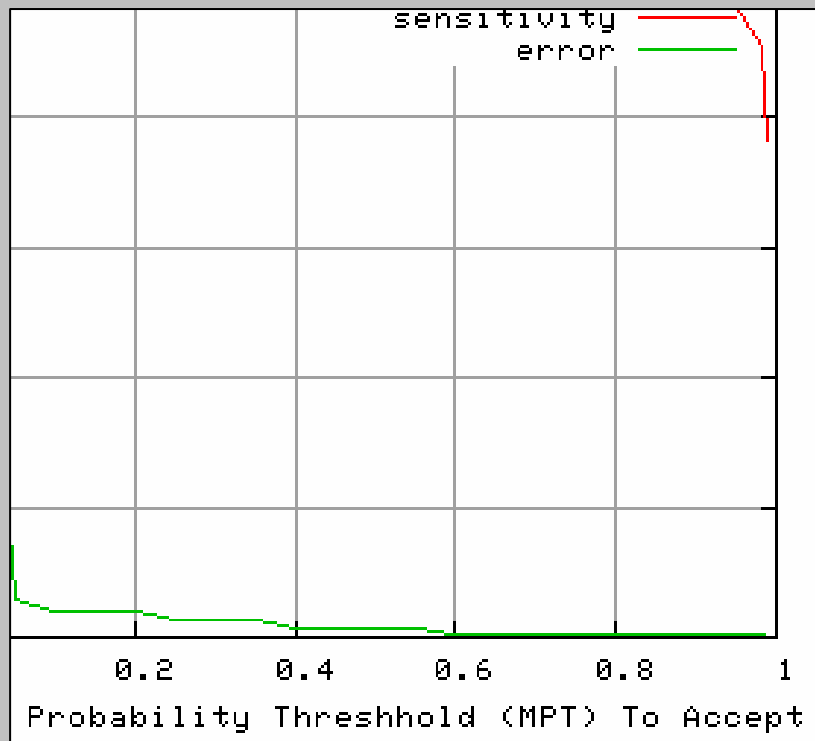
Not so good



PeptideProphet Results: Model Summary

```
FINAL 2+ MODEL after 6 iterations;
number of spectra: 18495
using no. tolerable trypsin term. [ntt] 0 data as pseudonegatives
  prior: 0.087, est. total correct: 1612.2
MASCOT discrim score [fval]   slope: 1.04 intercept: -22.88
  regression error: 6.64 negmean: -1.26
  pos: (gaussian mean 1.43, stdev 1.52)
  neg: (evd mean -1.29, stdev 0.52, mu -1.52, beta 0.41)
no. tolerable trypsin term. [ntt]
  pos: (ntt=0 0.093, ntt=1 0.278, ntt=2 0.628)
  neg: (ntt=0 0.796, ntt=1 0.190, ntt=2 0.015)
no. missed enz. cleavages [nmc]
  pos: (nmc=0 0.949, 1<=nmc<=2 0.051, nmc>=3 0.000)
  neg: (nmc=0 0.385, 1<=nmc<=2 0.537, nmc>=3 0.077)
var offset mass diff [massd] (offset: 0.70)
  pos: (massd=-4.0 0.00, massd=-3.0 0.00, massd=-2.0 0.01, massd=-1.0 0.04,
        massd=0.0 0.80, massd=1.0 0.14, massd=2.0 0.02)
  neg: (massd=-4.0 0.03, massd=-3.0 0.15, massd=-2.0 0.16, massd=-1.0 0.17,
        massd=0.0 0.18, massd=1.0, 0.17, massd=2.0 0.13)
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 Peptides



Sensitivity (Red Line): fraction of all correct assignments (48.3) passing MPT filter

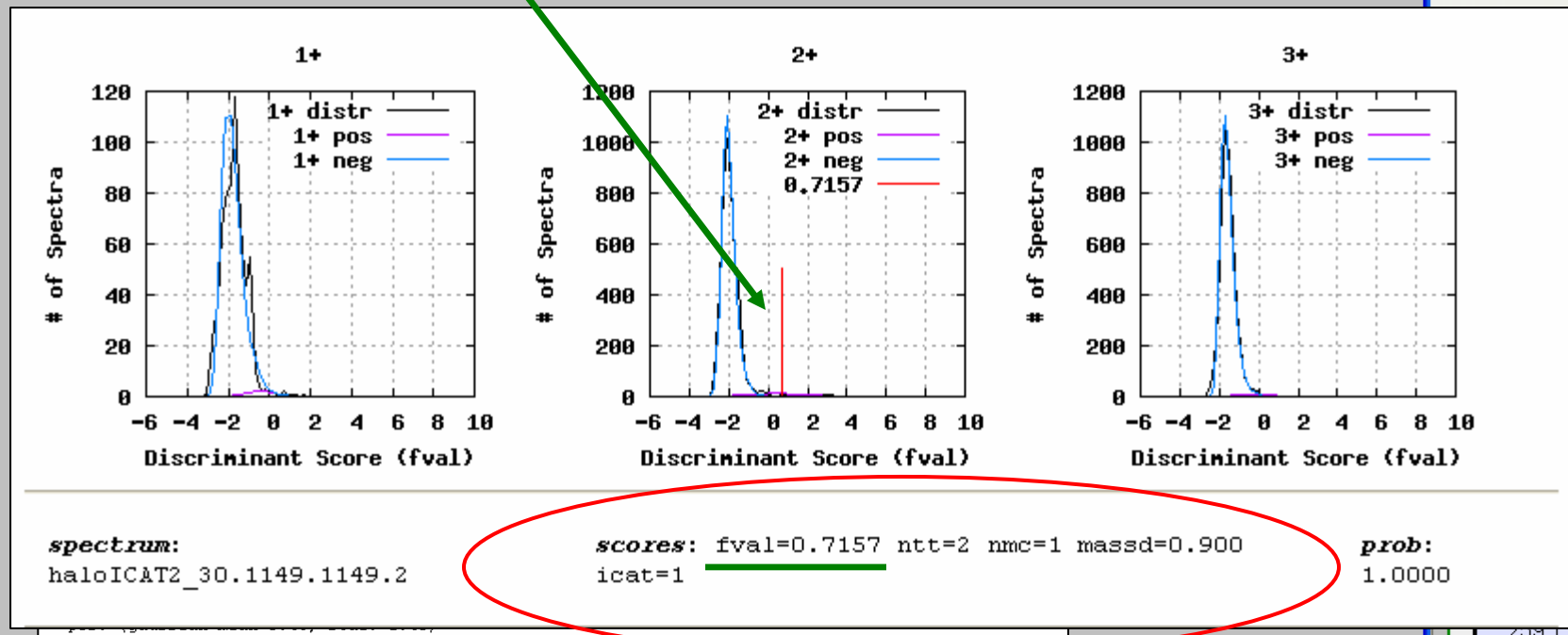
Error (Green Line): fraction of peptide assignments passing MPT filter that are incorrect

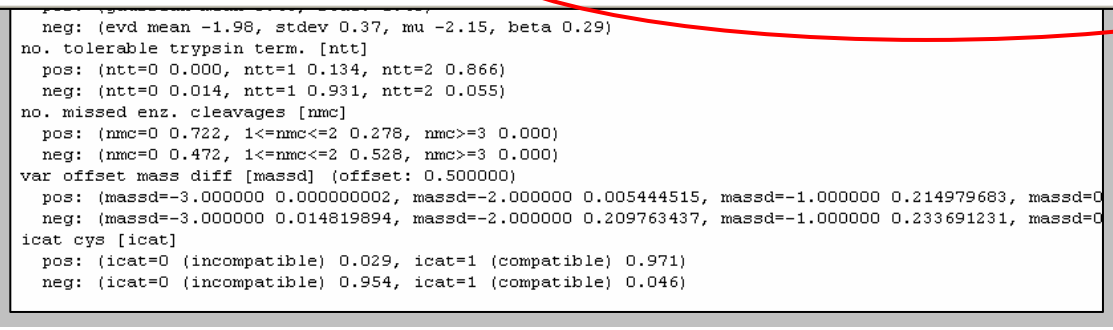
MPT = Minimum Probability Threshhold to Accept

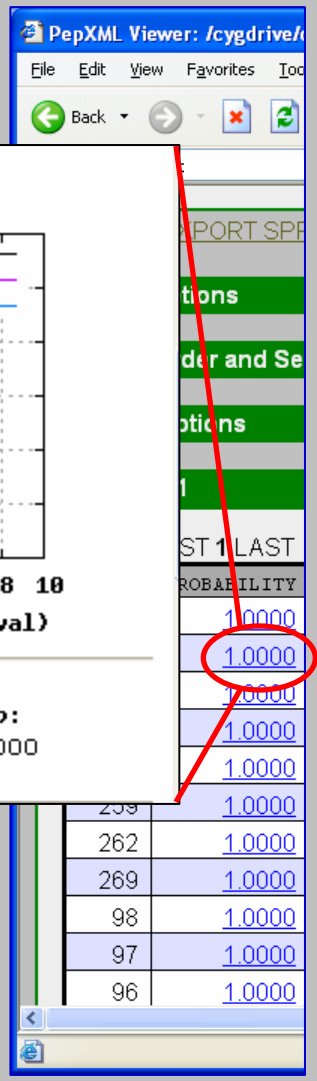
error	MPT	est # corr
0.000	1.00	13
0.025	0.35	48
0.050	0.06	48

PeptideProphet Results: Contributing Score and Peptide Properties

2+ search result discriminant score value





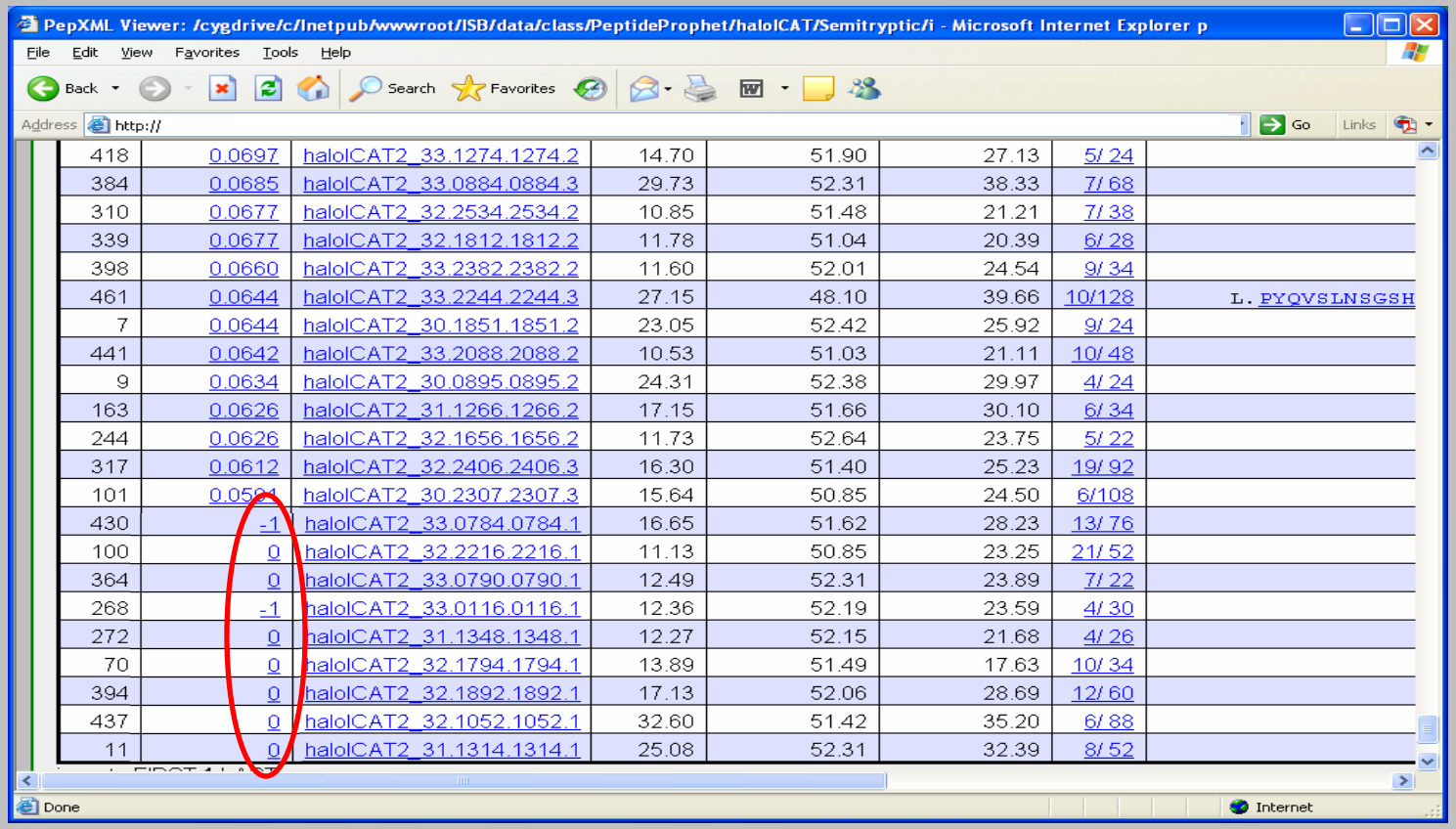


PeptideProphet $[M+2H]^{2+}$ vs $[M+3H]^{3+}$ Precursor Ions

334	1.0000	haloICAT2_33.1062.1062.3	51.06	51.94	36.29	10/ 44	
338	0.9984	haloICAT2_33.1042.1042.2	33.62	51.98	33.58	6/ 22	
357	0.9996	haloICAT2_33.1034.1034.2	27.70	51.85	26.77	7/ 18	
331	0.9596	haloICAT2_33.1024.1024.3	24.96	52.07	32.23	12/ 44	
386	0.4275	haloICAT2_33.1014.1014.3	17.19	49.67	27.34	6/ 96	
372	0.5725	haloICAT2_33.1014.1014.2	41.36	51.37	27.11	15/ 36	
373	0.9992	haloICAT2_33.1012.1012.2	34.49	51.39	35.29	11/ 36	
312	0.6340	haloICAT2_33.1004.1004.1	27.62	52.49	33.39	7/ 22	
330	0.9992	haloICAT2_33.1002.1002.2	26.84	52.06	29.37	8/ 22	

Spectrum searched as both 2+ and 3+ precursor received significant probability

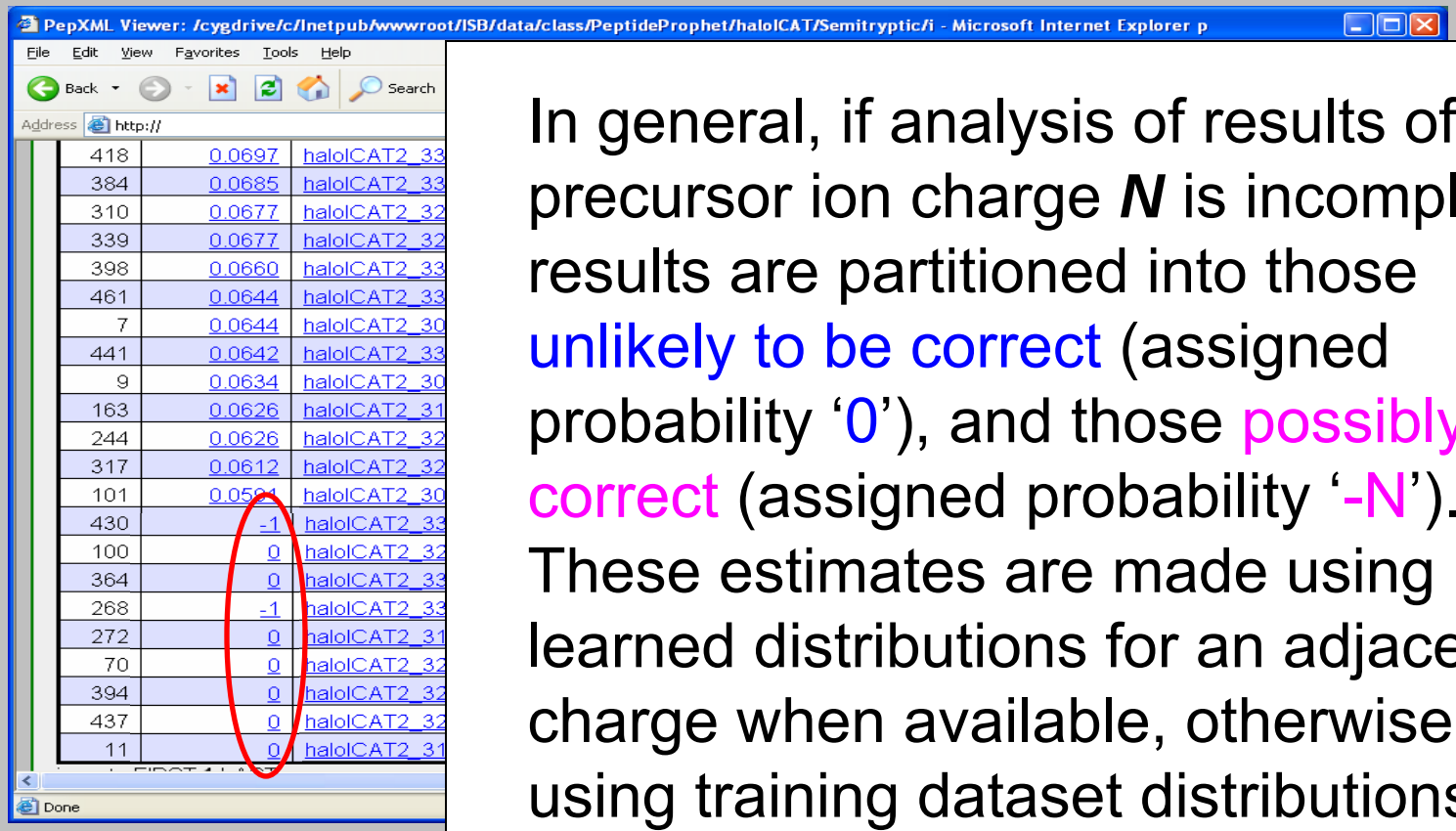
PeptideProphet Results: Incomplete Analysis



418	0.0697	haloICAT2_33.1274.1274.2	14.70	51.90	27.13	5/ 24	
384	0.0685	haloICAT2_33.0884.0884.3	29.73	52.31	38.33	7/ 68	
310	0.0677	haloICAT2_32.2534.2534.2	10.85	51.48	21.21	7/ 38	
339	0.0677	haloICAT2_32.1812.1812.2	11.78	51.04	20.39	6/ 28	
398	0.0660	haloICAT2_33.2382.2382.2	11.60	52.01	24.54	9/ 34	
461	0.0644	haloICAT2_33.2244.2244.3	27.15	48.10	39.66	10/128	L.PYQVSLNSGSH
7	0.0644	haloICAT2_30.1851.1851.2	23.05	52.42	25.92	9/ 24	
441	0.0642	haloICAT2_33.2088.2088.2	10.53	51.03	21.11	10/ 48	
9	0.0634	haloICAT2_30.0895.0895.2	24.31	52.38	29.97	4/ 24	
163	0.0626	haloICAT2_31.1266.1266.2	17.15	51.66	30.10	6/ 34	
244	0.0626	haloICAT2_32.1656.1656.2	11.73	52.64	23.75	5/ 22	
317	0.0612	haloICAT2_32.2406.2406.3	16.30	51.40	25.23	19/ 92	
101	0.0591	haloICAT2_30.2307.2307.3	15.64	50.85	24.50	6/108	
430	-1	haloICAT2_33.0784.0784.1	16.65	51.62	28.23	13/ 76	
100	0	haloICAT2_32.2216.2216.1	11.13	50.85	23.25	21/ 52	
364	0	haloICAT2_33.0790.0790.1	12.49	52.31	23.89	7/ 22	
268	-1	haloICAT2_33.0116.0116.1	12.36	52.19	23.59	4/ 30	
272	0	haloICAT2_31.1348.1348.1	12.27	52.15	21.68	4/ 26	
70	0	haloICAT2_32.1794.1794.1	13.89	51.49	17.63	10/ 34	
394	0	haloICAT2_32.1892.1892.1	17.13	52.06	28.69	12/ 60	
437	0	haloICAT2_32.1052.1052.1	32.60	51.42	35.20	6/ 88	
11	0	haloICAT2_31.1314.1314.1	25.08	52.31	32.39	8/ 52	

Model incomplete for results of 1+ precursor ions

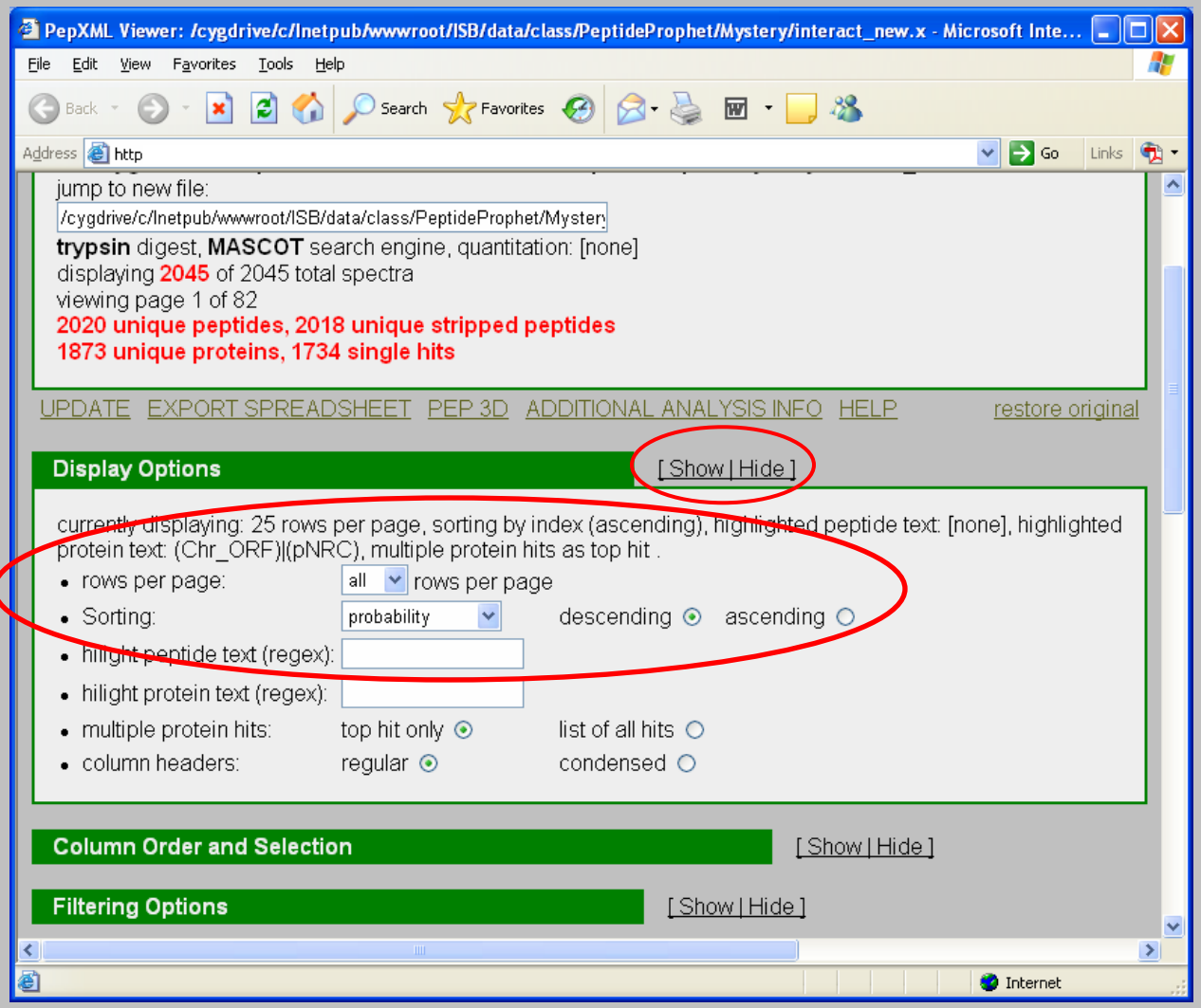
PeptideProphet Results: Incomplete Analysis



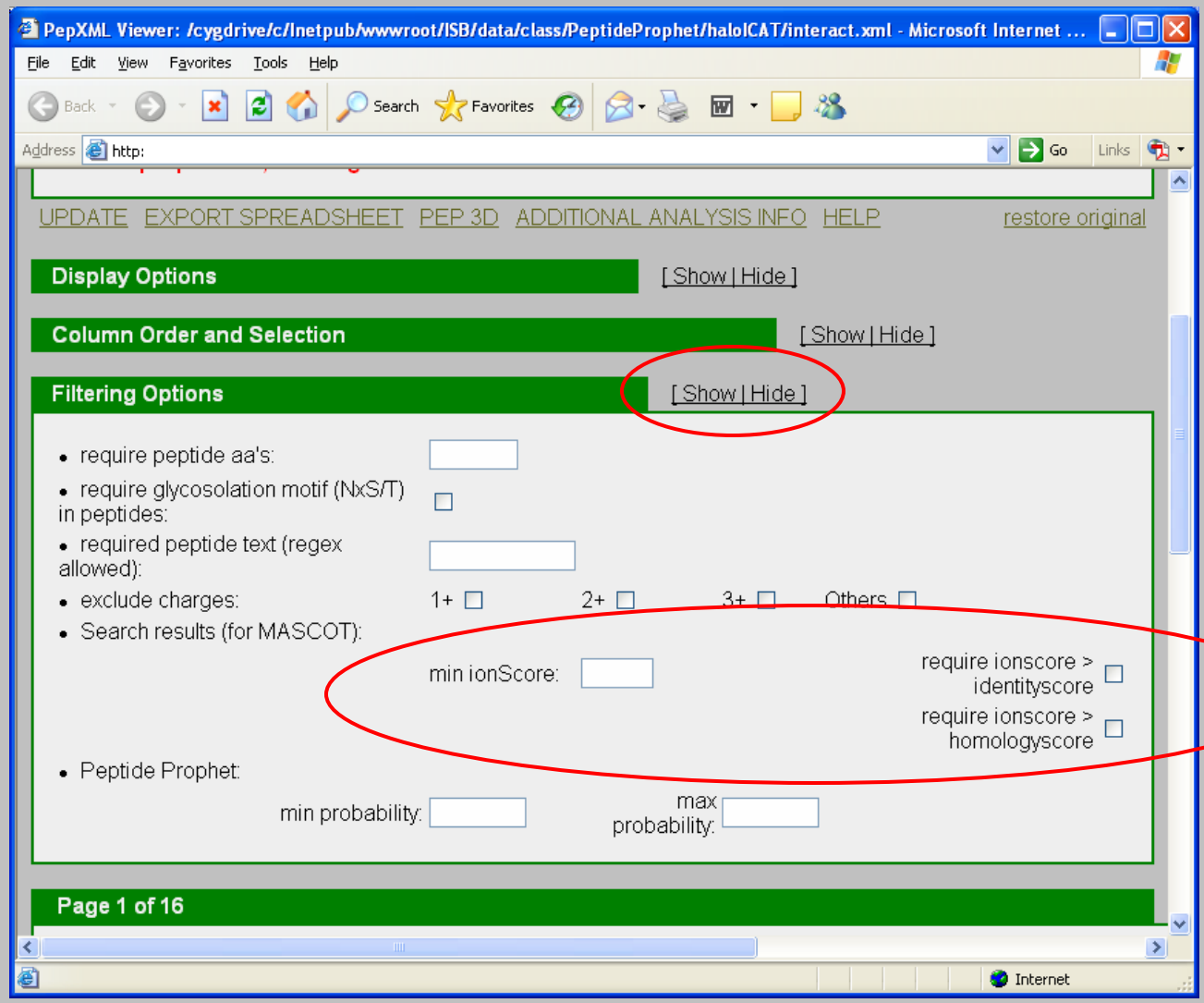
In general, if analysis of results of precursor ion charge ***N*** is incomplete, results are partitioned into those unlikely to be correct (assigned probability '0'), and those possibly correct (assigned probability '-N'). These estimates are made using learned distributions for an adjacent charge when available, otherwise using training dataset distributions

Model incomplete for results of 1+ precursor ions

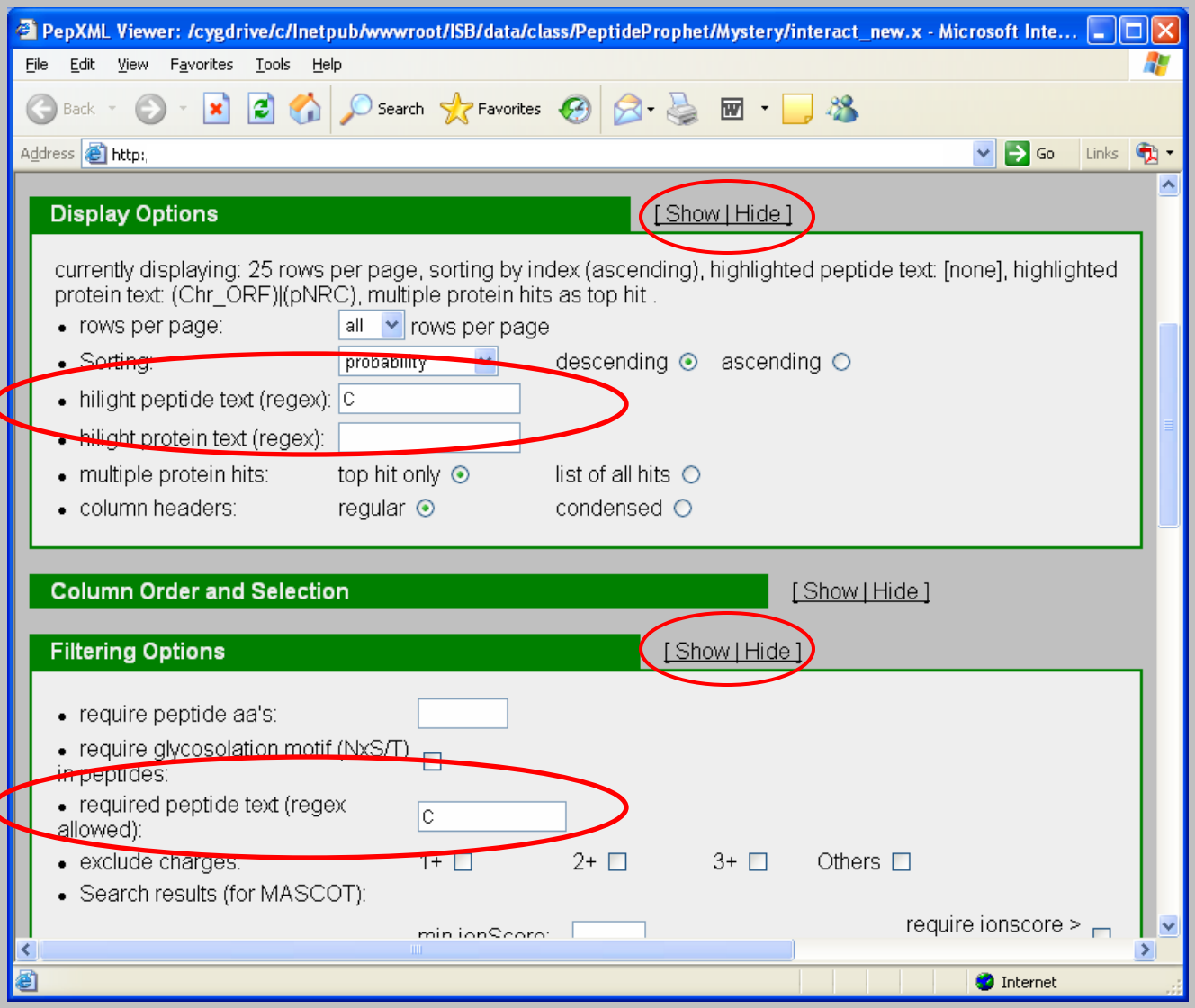
Sort Data by Computed Probability



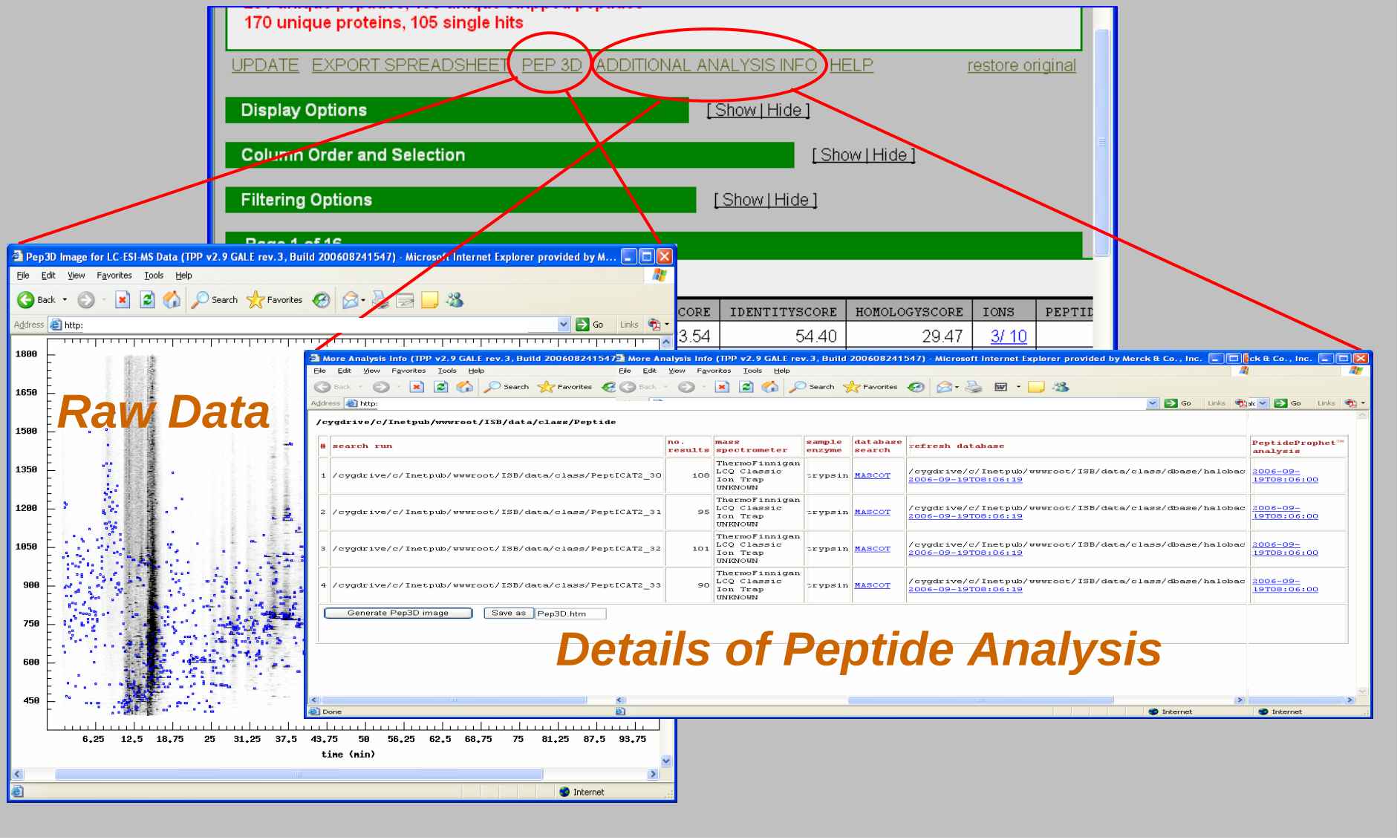
Filter Data by Mascot Ionscore



Select and Color Specified AA's

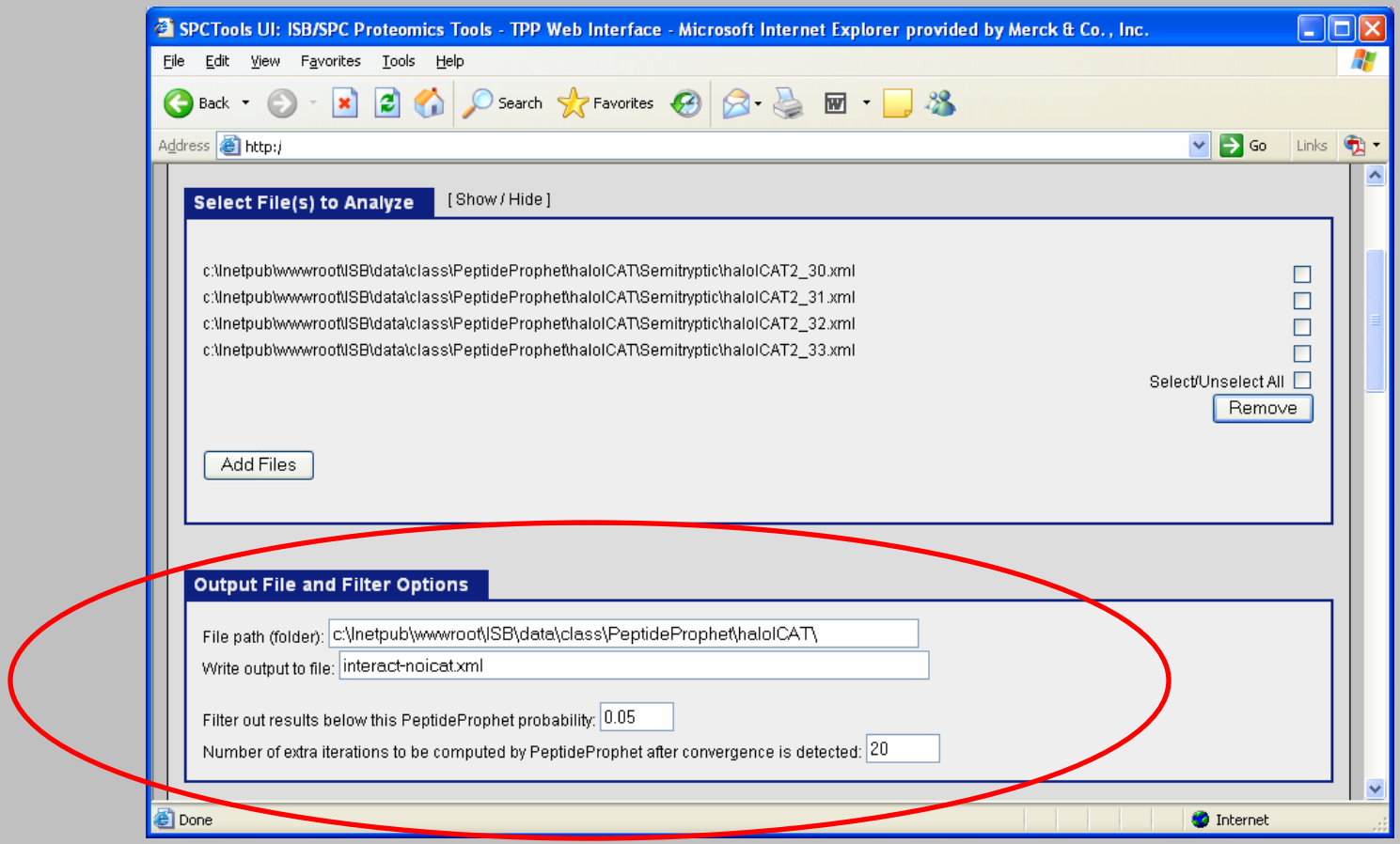


Pep3D and Analysis Summary Links



User Options for PeptideProphet

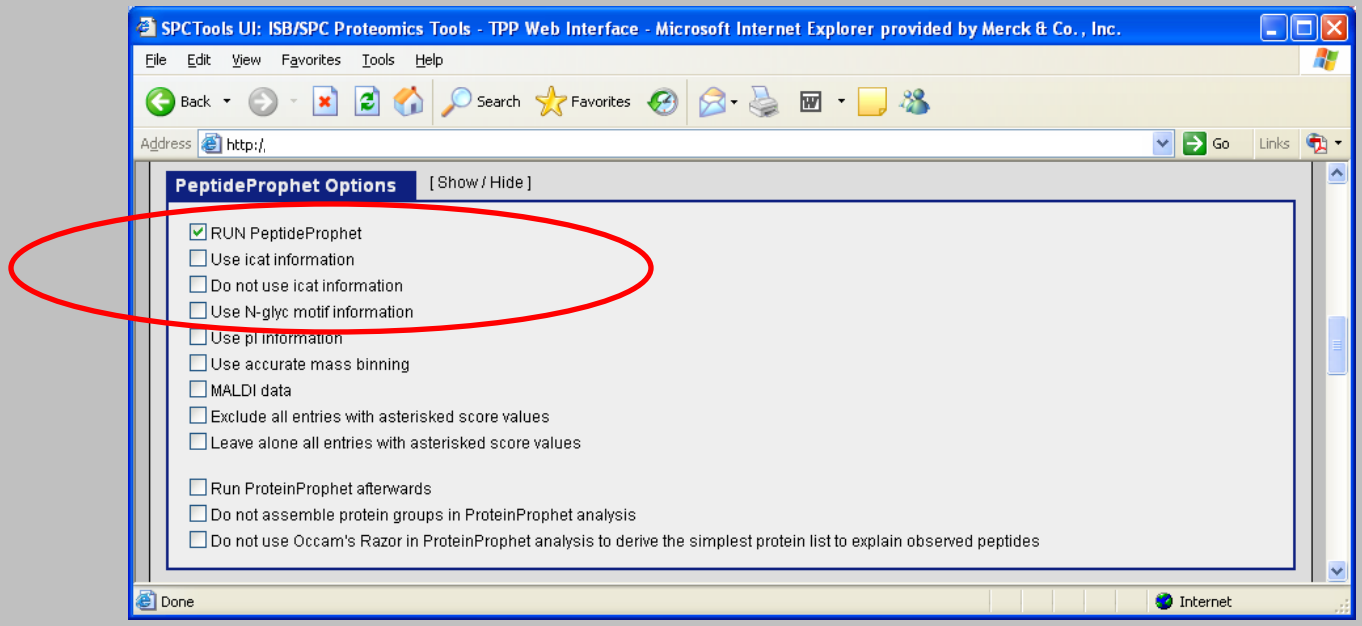
Rename Output File (*e.g.* to **`interact-noicat.xml`**):



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



Ionscore* Example

- Search results are marked with asterisked Ionscore when runner up peptide(s) share at least 75% sequence identity with top peptide

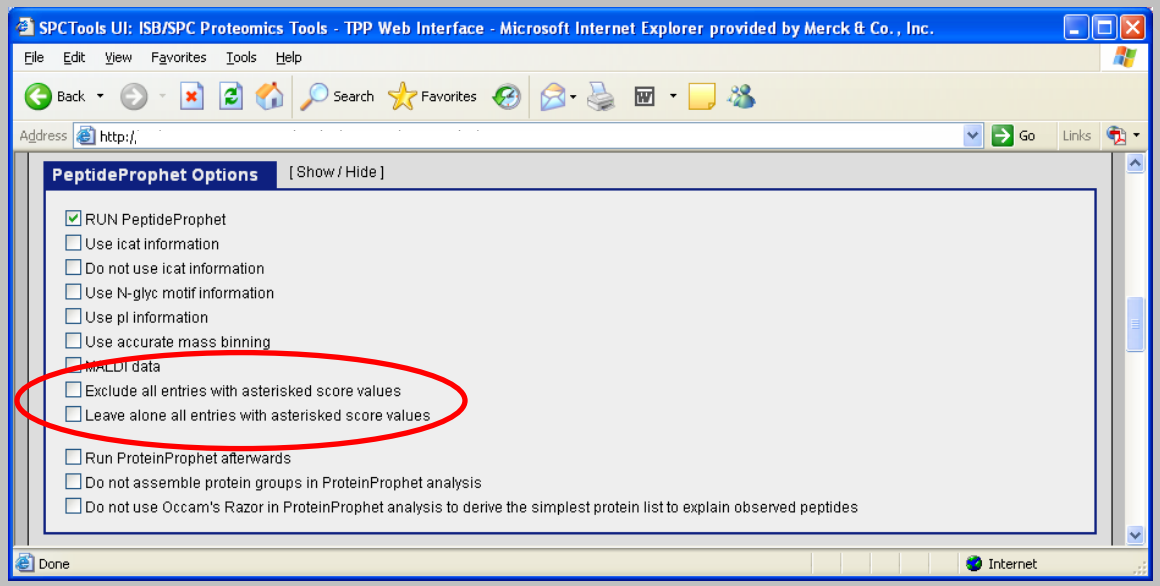
6779	0.0000	halolCAT2_32.1630.1630.3	17.17	62.53	29.81	32/116	
6778	0.0000	halolCAT2_32.1282.1282.3	16.15*	62.53	27.83	8/ 88	
6777	0.0000	halolCAT2_32.2442.2442.3	18.98	62.53	30.02	29/112	
7349	0.0000	halolCAT2_32.0013.0013.3	19.66	63.49	30.44	17/160	
6776	0.0000	halolCAT2_32.1122.1122.3	18.83	62.55	29.99	9/100	
6759	0.0000	halolCAT2_32.2730.2730.3	19.43*	62.53	31.44	10/108	
6775	0.0000	halolCAT2_32.1194.1194.3	14.71	62.55	26.12	8/104	
6774	0.0000	halolCAT2_32.2836.2836.3	22.91	62.55	29.73	18/108	
6773	0.0000	halolCAT2_32.2562.2562.3	23.14	62.55	35.85	6/100	
7350	0.0000	halolCAT2_32.3348.3348.3	16.86	63.49	29.49	10/152	
6830	0.0000	halolCAT2_32.2410.2410.3	19.88	62.64	32.38	22/120	

#	MH+	IonSc	Ions	Ref		Sequence
1	2956.4 (-0.4)	19.43	10/108	SWN:RN19_HUMAN		F.STNTSSDNGLTSISKQIGDFIECPLCLL.R
2	2956.4 (-0.4)	17.93	10/ 92	SWN:RN19_HUMAN		F.STNTSSDNGLTSISKQIGDFIEC[545]P.L
3	2954.6 (+1.4)	17.79	14/100	GP:AY038599_2		T.AQSEVALLRFVNPDTGRVLFESKLHK.Q
4	2955.5 (+0.4)	17.73	11/104	SW:ELS_HUMAN	+6	F.PLGGVAARPGFGLSPIFPGGACLGKAC[545].G
5	2954.6 (+1.4)	16.24	5/108	SWN:Y450_HUMAN		G.LFLRGPKPGSLDSHAAGRPPARPSVSQR.I
6	2955.6 (+0.4)	15.79	12/ 92	pNRC100_ORF5058	+3	T.DVQPWRLLVGGVFVGIGTRVGKGC[545].T
7	2954.6 (+1.4)	15.44	22/124	GP:M24766_1		P.KGDPGFPGAPGTVGAPGIAGIPQKIAVQPGTV.G
8	2955.4 (+0.5)	15.19	8/112	SW:GGT5_HUMAN	+2	P.CGPQAFAHAAVAADSKVCSDIGRAILQQQ.G
9	2956.3 (-0.3)	14.84	19/112	SW:VWF_HUMAN	+2	E.CCGRCLPSACEVVTGSPRGDSQSSWKSVG.S
10	2954.4 (+1.6)	14.78	22/124	GP:S79774_1		P.PTGDSGPPPVPPTGDSGAPPVTPTGDSETAPV.P

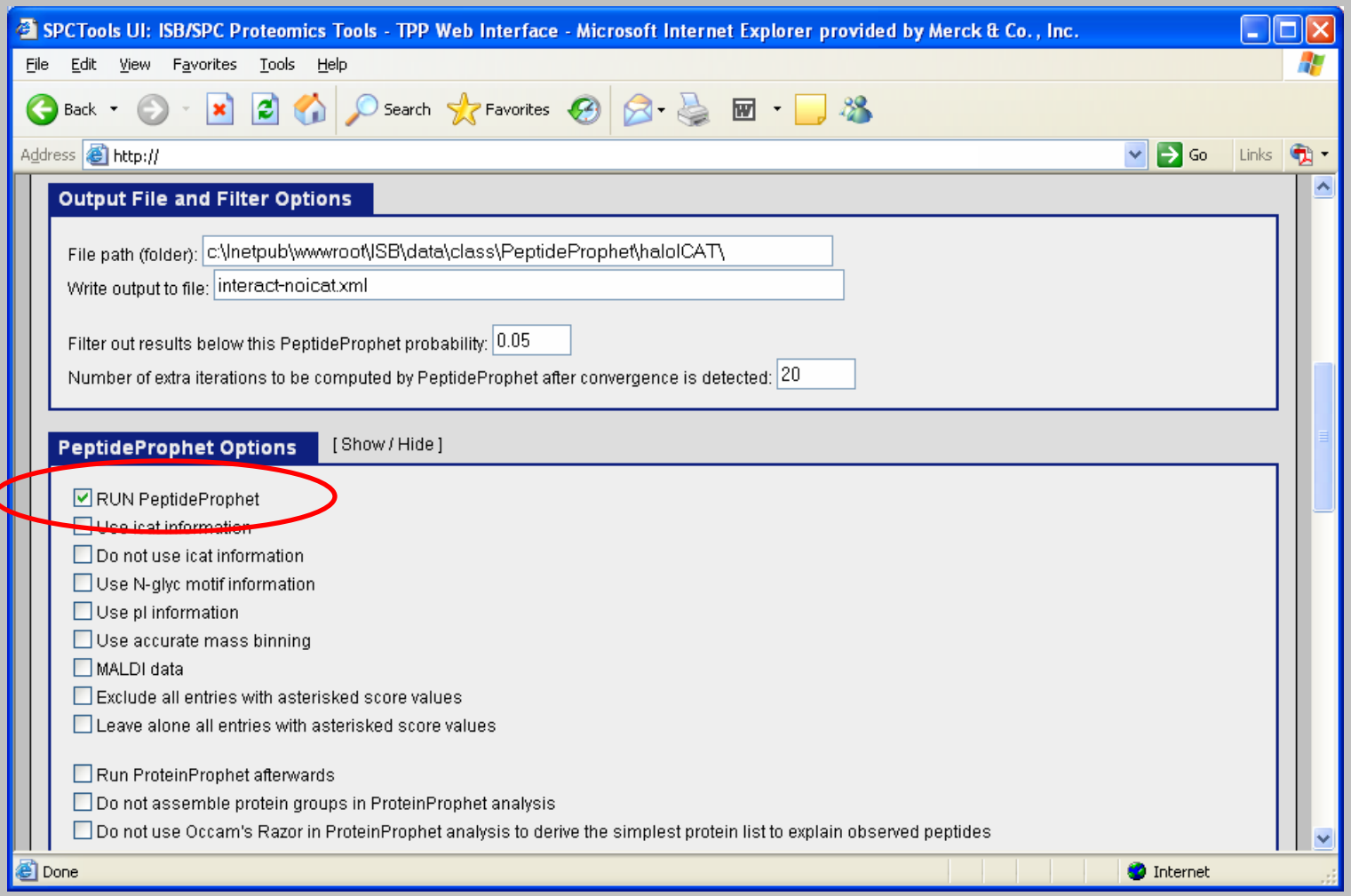
Ionscore* Options

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

- Penalize (the default option, halves Ionscore values)
- Leave alone (suitable for the context of homologues)
- Exclude (the most conservative, assigns probability 0)



Run/Don't Run PeptideProphet



Ongoing Developments for PeptideProphet

- Optimize for various additional mass spectrometers
  - New discriminant function
- Adapt to additional methods for assigning peptides to tandem mass spectra
  - SEQUEST
  - COMET
  - ProbID
  - SpectraST
  - X!Tandem
  - Others

Exercises with PeptideProphet

- Accuracy of computed probabilities
- Utility of conventional Mascot score thresholds and PeptideProphet analysis
- Model results for ICAT data analyzed with and without ICAT information
- Model results for Mystery dataset

Exercise Datasets

Many of the exercises utilize Mascot search results of ***HaloICAT*** datasets for which correct results are independently known:

- MS/MS spectra generated from *Halobacterium* ICAT sample searched against a halobacterium_plus_human protein sequence database

The pepXML Viewer is pre-configured for this class to automatically color all ***HaloICAT*** correct corresponding proteins **red!**