Peptide Validation

2020.06

김현우

Peptide Assignment by MS/MS

Database searching approach

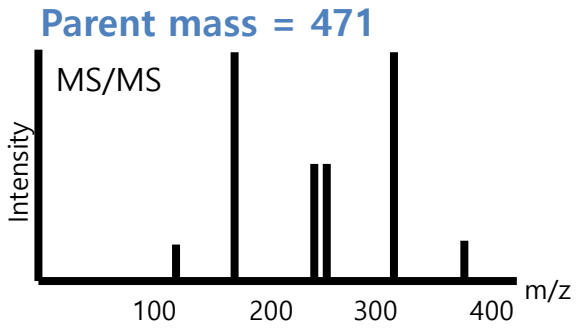
- SEQUEST(Comet), Mascot

Candidate peptides

READ
DVGAE
DEAR
GAGVDA
EGDVA
...

Comparison

Experimental MS/MS spectrum

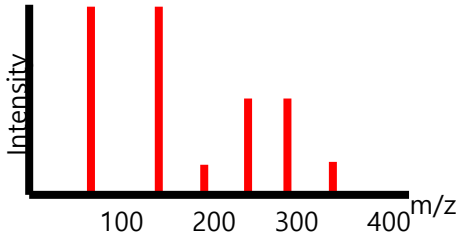


Peptide Assignment by MS/MS

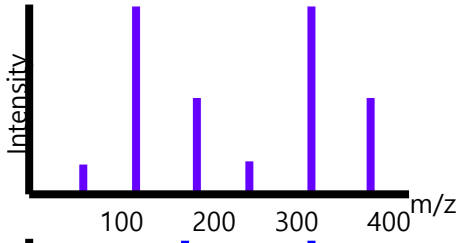
Candidate peptides

Theoretical MS/MS spectrum

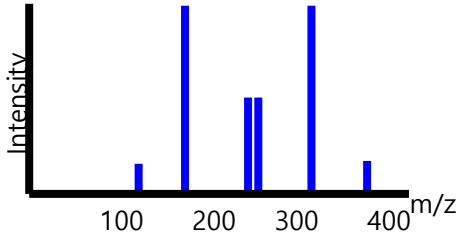
READ



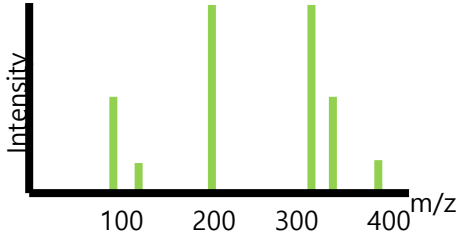
DVGAE



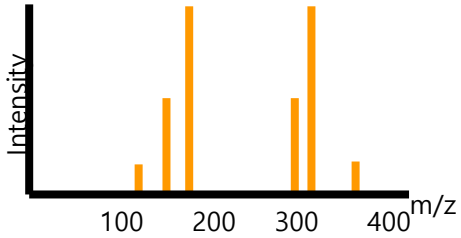
DEAR



GAGVDA

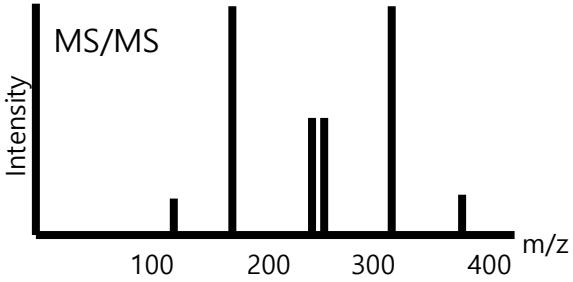


EGDVA

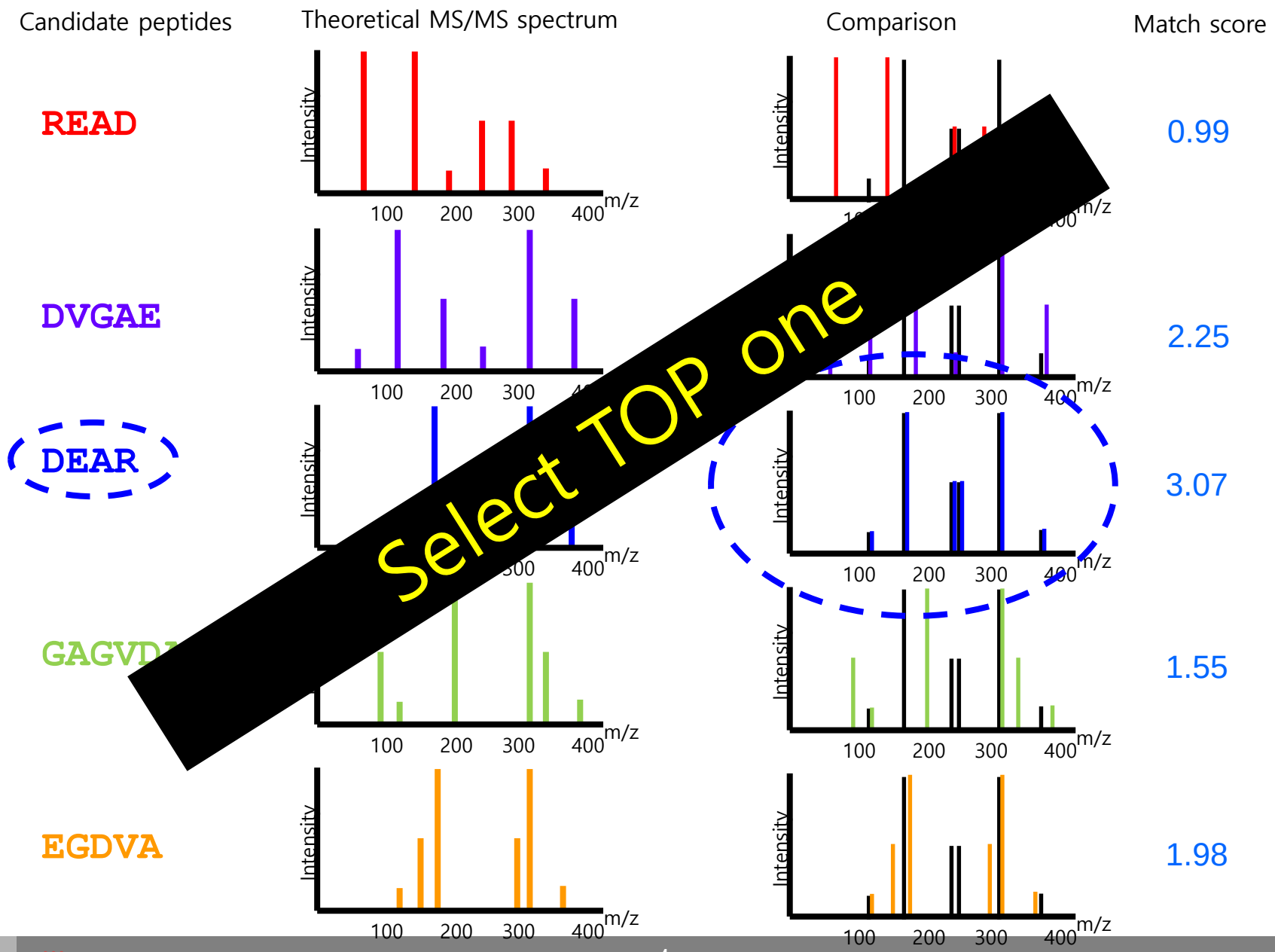


Experimental MS/MS spectrum

Parent mass = 471



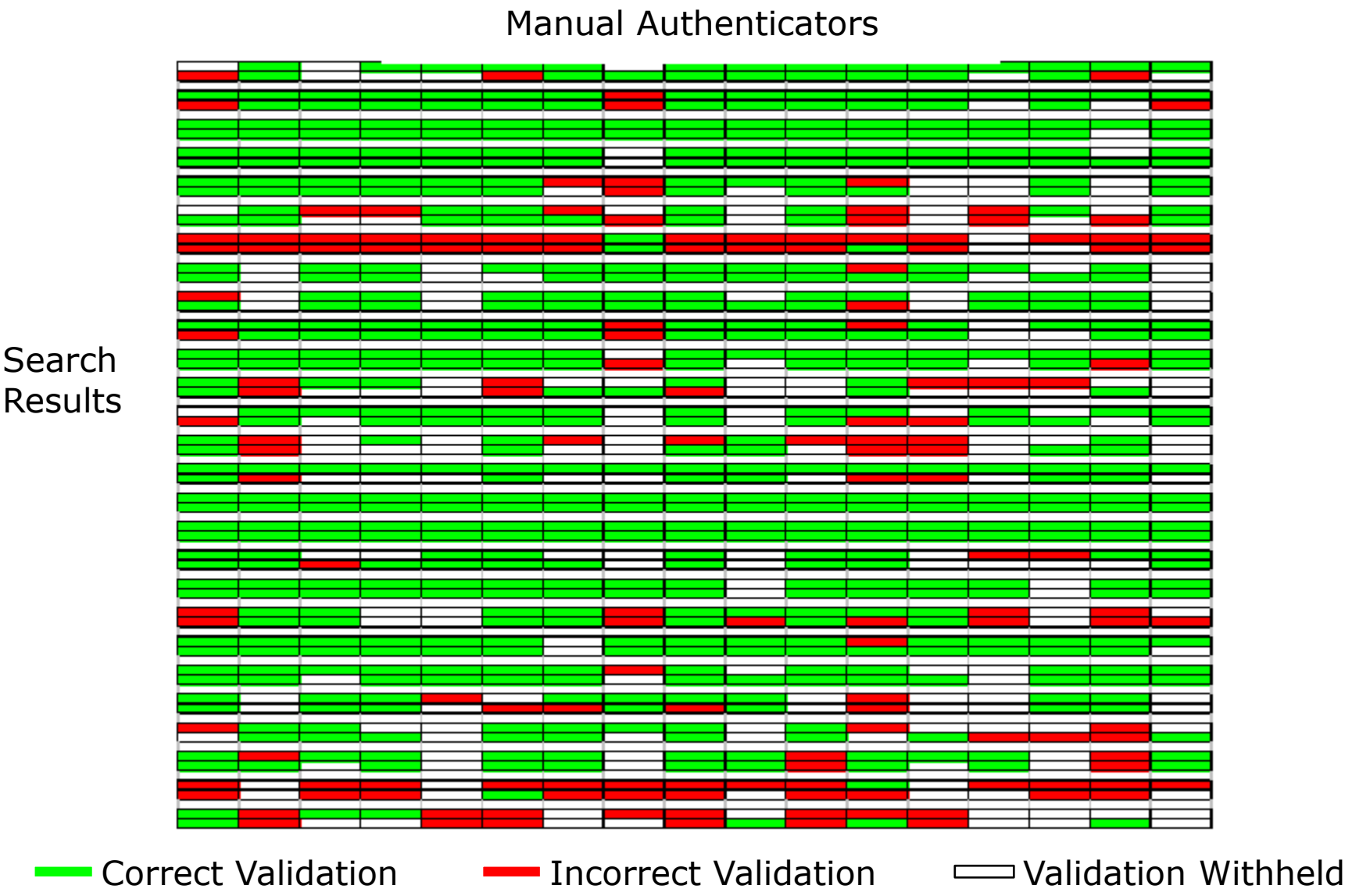
Peptide Assignment by MS/MS



Peptide Validation

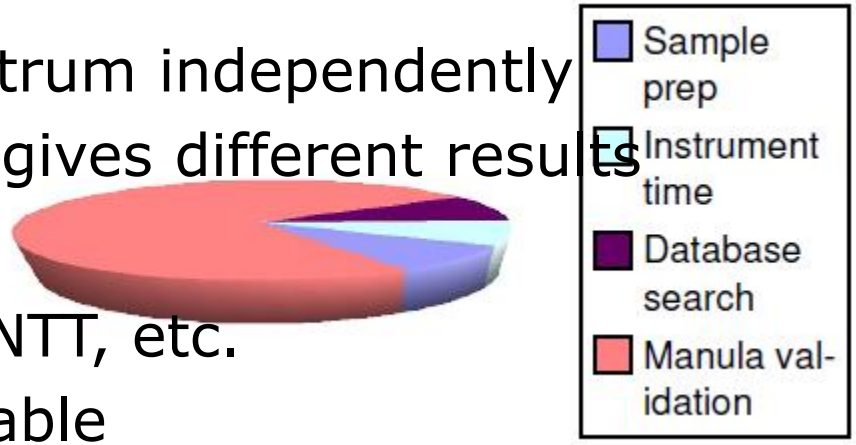
- Peptide assignment
  - interpret each MS/MS spectrum independently
  - different analysis software gives different results
- Manual validation
  - filtering by search scores, NTT, etc.
  - subjective opinion is inevitable
  - hard to estimate the error rate
  - when dataset gets large?
- Statistical validation
  - model-based validation: probabilistic model for score distribution
  - target-decoy: estimate false discovery rate based on the match to the “decoy” database

(Un)reliability of manual validation

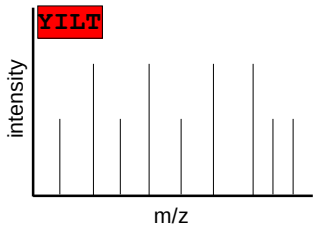


Peptide Validation

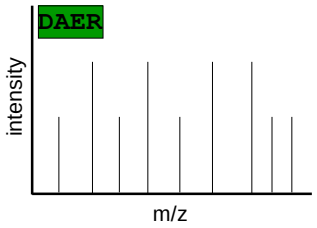
- Peptide assignment
  - interpret each MS/MS spectrum independently
  - different analysis software gives different results
- Manual validation
  - filtering by search scores, NTT, etc.
  - subjective opinion is inevitable
  - hard to estimate the error rate
  - when dataset gets large?
- Statistical validation
  - model-based validation: probabilistic model for score distribution
  - target-decoy: estimate false discovery rate based on the match to the "decoy" database



Target/Decoy method



3.15

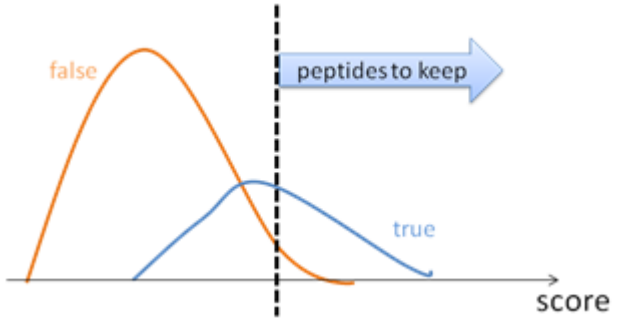


2.47

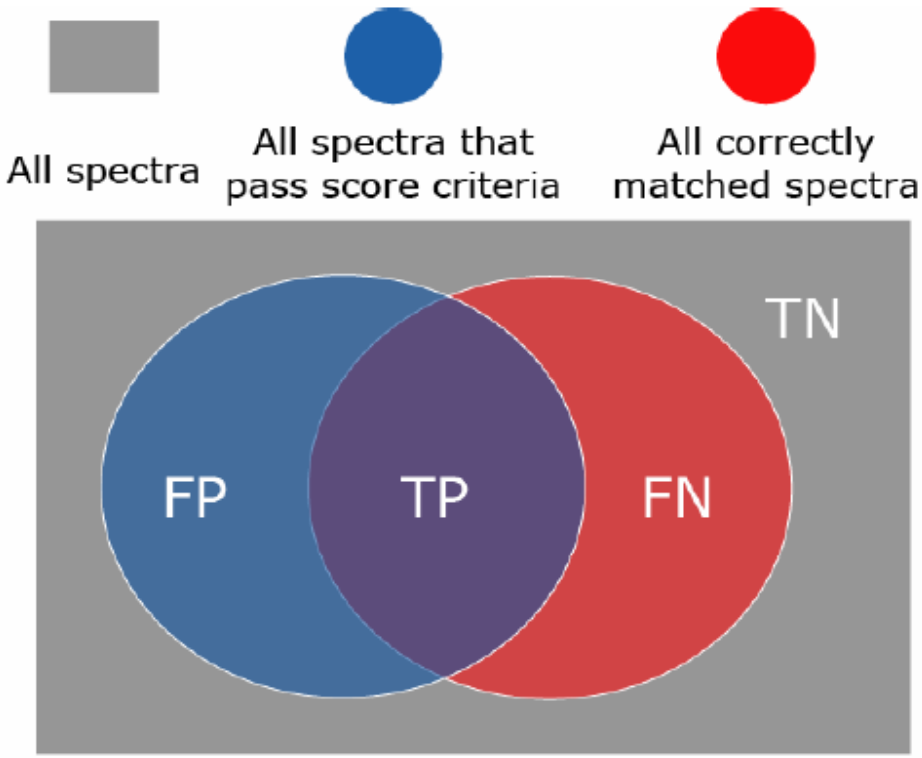
>Protein A (Target Sequence)
MEMEKEFEQIDKSGSWAAIYQDIDVGAEDFPCRVAKLPK
NKNRNRYRDVSPFDHSRKREADDNDYINASLIKMEEAQR
SYILTQQIDKSGSWAAIYQDIRHEASDFHEASDFPCRVA
KLPKNKDEARYMEKEFEQIDKGAGVDADIRHEMEKEFEQ
IDKSGSWAAIYQDIRHE

>Reversed Protein A (Decoy Sequence)
EHRIDQYIAAWSGSKDIQEFEKEMEHRIDADVGAGKDIQ
EFEKEMYRAEDKNKPLKAVRCPFDSAEHFDSAEHRIDQY
IAAWGSGKDIQQTLIYSRQAEEMKILSANIYDNDDAERK
RSHDFPSVDRYRNRNKNKPLKAVRCPFDEAGVDIDQYIA
AWSGSKDIQEFEKEMEM

- T = 1000 # of matches to the target sequence (above score threshold)
- D = 20 # of matches to the decoy sequence (above score threshold)
- False Discovery Rate = ?



Target/Decoy method



TP = True Positive
FP = False Positive

TN = True Negative
FN = False Negative

$$\text{FP rate} = \frac{FP}{TP+FP}$$

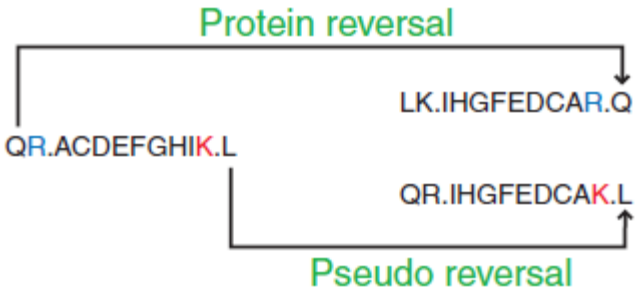
$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

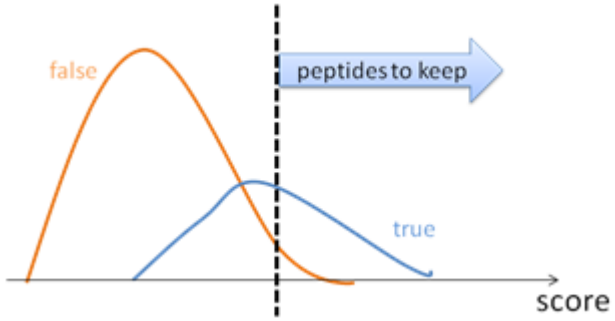
$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

Target/Decoy method

- Target/Decoy method is meaningful only for large datasets.
- Decoy database? (database size, amino acid composition, peptide length distribution, precursor mass distribution)
  - Reversed sequence
  - Pseudo-reverse sequence
  - Random (Shuffled) sequence
  - Pseudo-shuffle

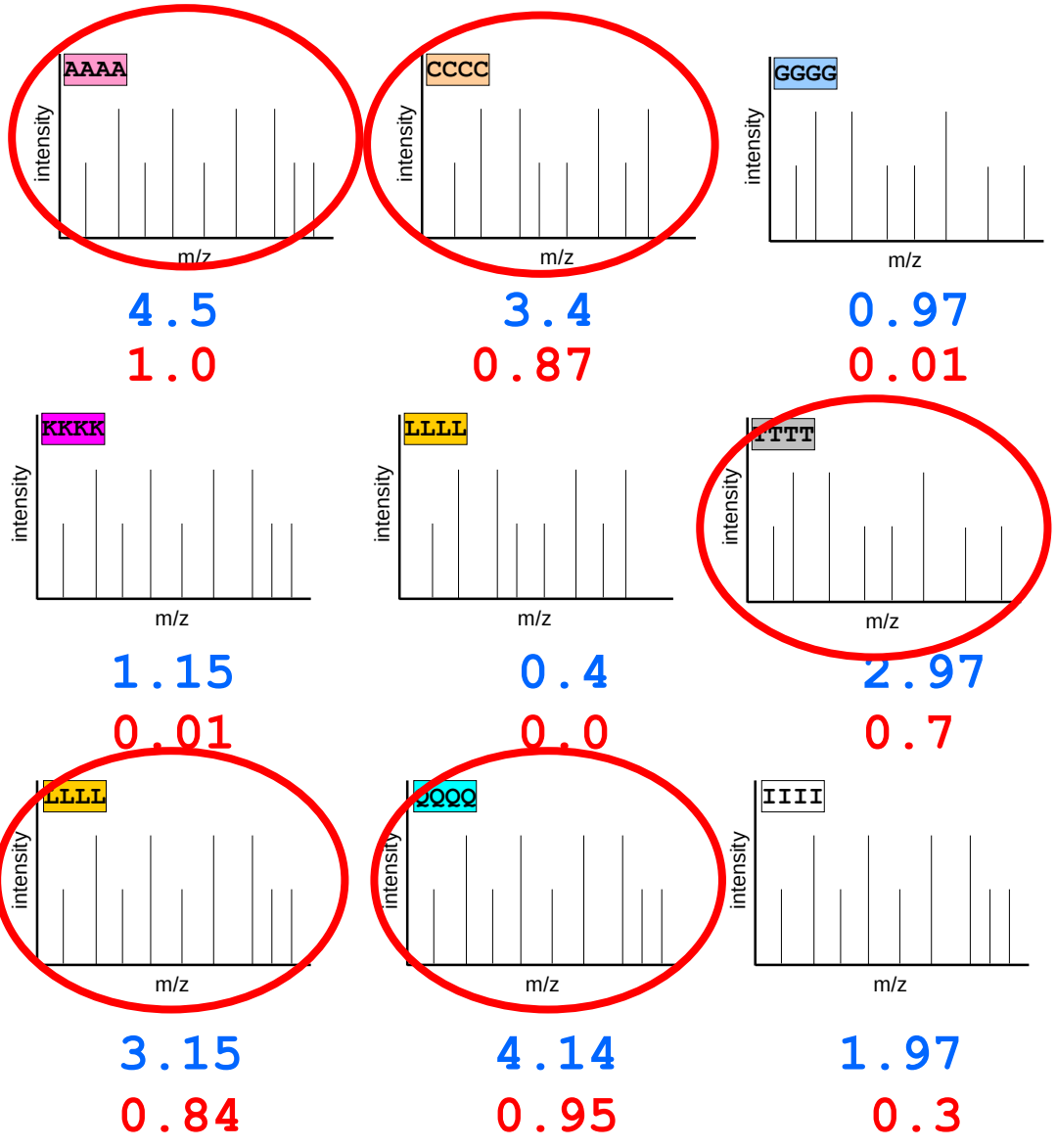


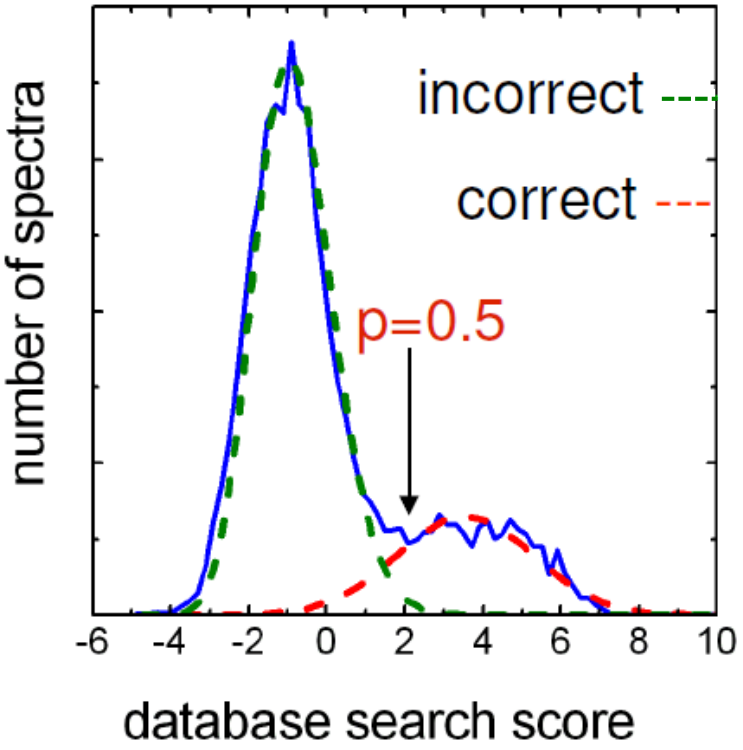
- Separated or concatenated?
  - Threshold score 30
  - Match to the target score 50
  - Match to the decoy score 40
  - Is this counted as a false positive?



- FDR calculated as: # of decoy hits / # of target hits above a certain threshold.

Probabilistic model-based: PeptideProphet





Probabilistic model-based: PeptideProphet

- Combine search scores into a single discriminant score.
- The discriminant score, *F*, can be computed as

$$F(x_1, x_2, \ldots, x_S) = c_0 + \sum_{i=1}^{S} c_i x_i$$

  - with constant c_0 and weights c_i derived such that the ratio of between-class variation to within-class variation is maximized.
- The discriminant score can be substituted to enable tractable calculation.

$$p(+|F) = \frac{p(F|+)p(+)}{p(F|+)p(+) + p(F|-)p(-)}$$

$$p(+|F, NTT) = \frac{p(F|+)p(NTT|+)p(+)}{p(F|+)p(NTT|+)p(+) + p(F|-)p(NTT|-)p(-)}$$

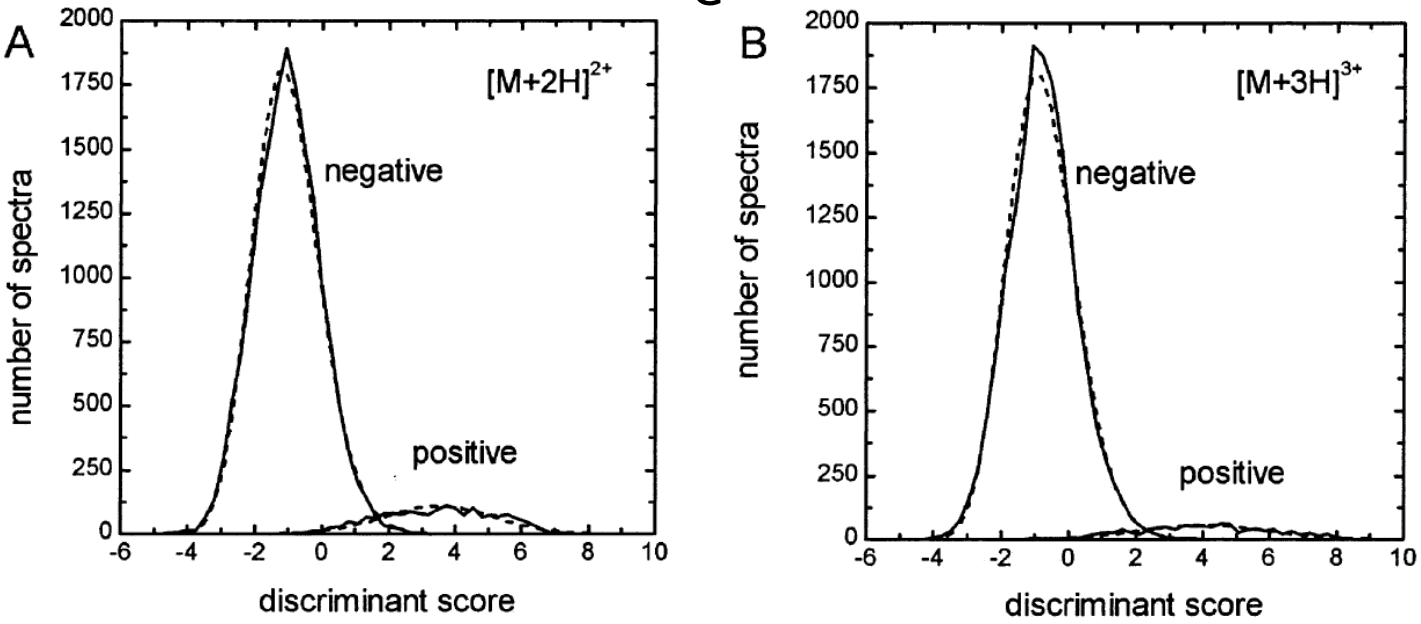
Probabilistic model-based: PeptideProphet

Table 1. Discriminant Functions Derived from Training Dataset Spectra of $[M + 2H]^{2+}$ and $[M + 3H]^{3+}$ Precursor Ions[a]

	$[M + 2H]^{2+}$		$[M + 3H]^{3+}$	
variable	coefficient	correlation	coefficient	correlation
Xcorr′	8.362	0.798	9.933	0.698
ΔC_n	7.386	0.746	11.149	0.806
ln SpRank	−0.194	−0.510	−0.201	−0.491
d_M	−0.314	−0.306	−0.277	−0.251
constant	−0.959		−1.460	

Probabilistic model-based: PeptideProphet

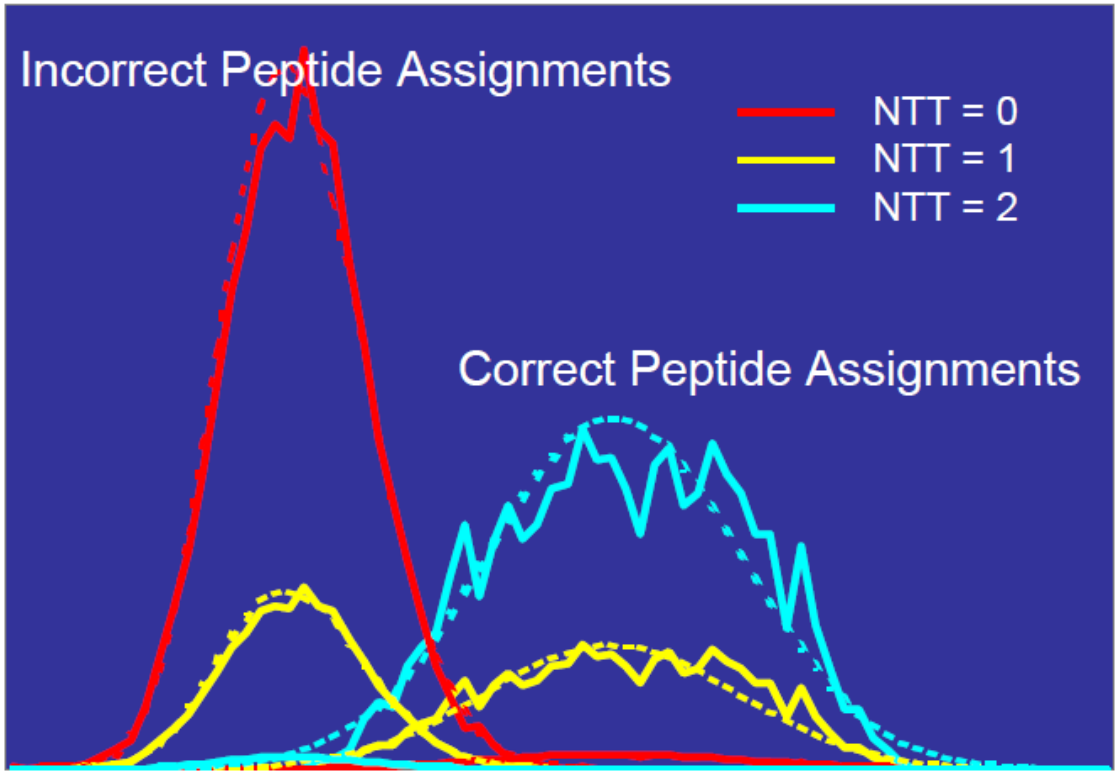
- Among spectra of $[M + 2H]^{2+}$
  - 84% of correct had discriminant scores of 1.7 or greater,
  - whereas 99% of incorrect had scores below that



Training dataset(solid line) and Gaussian and gamma distributions(dashed line).

Gaussian: $p(F|+) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(F-\mu)^2/2\sigma^2}$ **gamma:** $p(F|-) = \frac{(F-\gamma)^{\alpha-1} e^{-(F-\gamma)/\beta}}{\beta^{\alpha}\Gamma(\alpha)}$

Semi-parametric PeptideProphet



- Decoy search results => distribution for incorrect assignments
- EM algorithm to estimate distributions of correct assignments
  - NTT (number of tryptic termini)

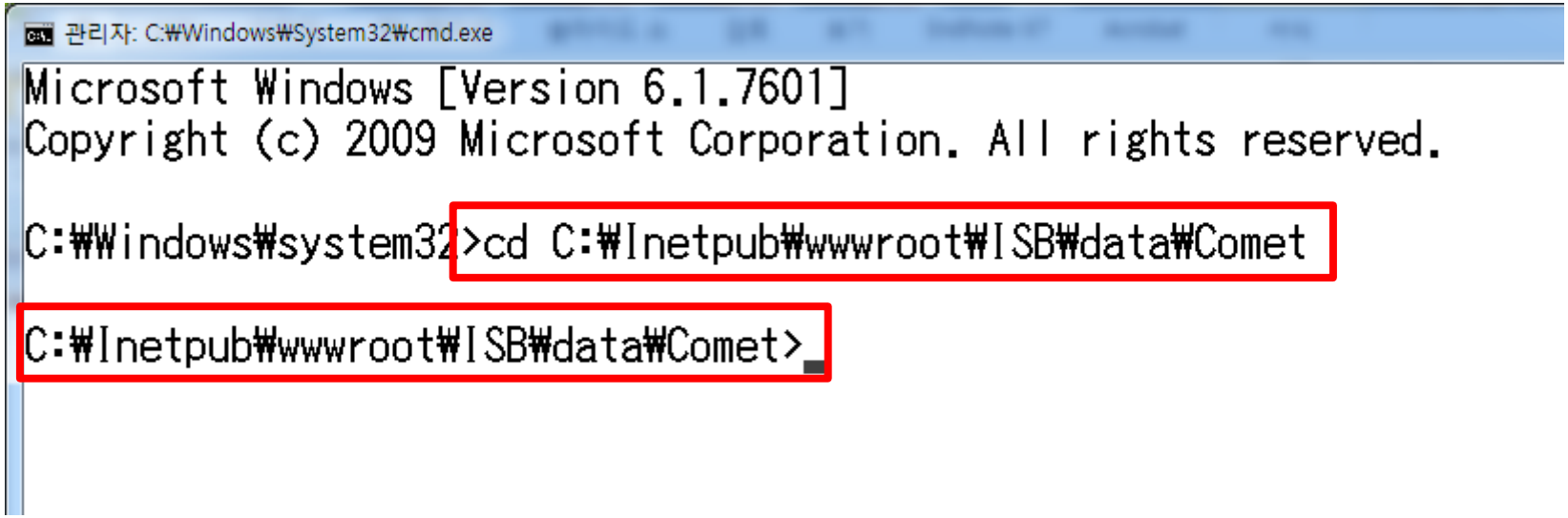
Target/Decoy and PeptideProphet

Target/Decoy 방법

- Cmd 창 실행
  - 시작메뉴
  - Cmd 입력

Target/Decoy 방법

- Cmd 창
  - cd C:\Inetpub\wwwroot\ISB\data\Comet



Target/Decoy 방법

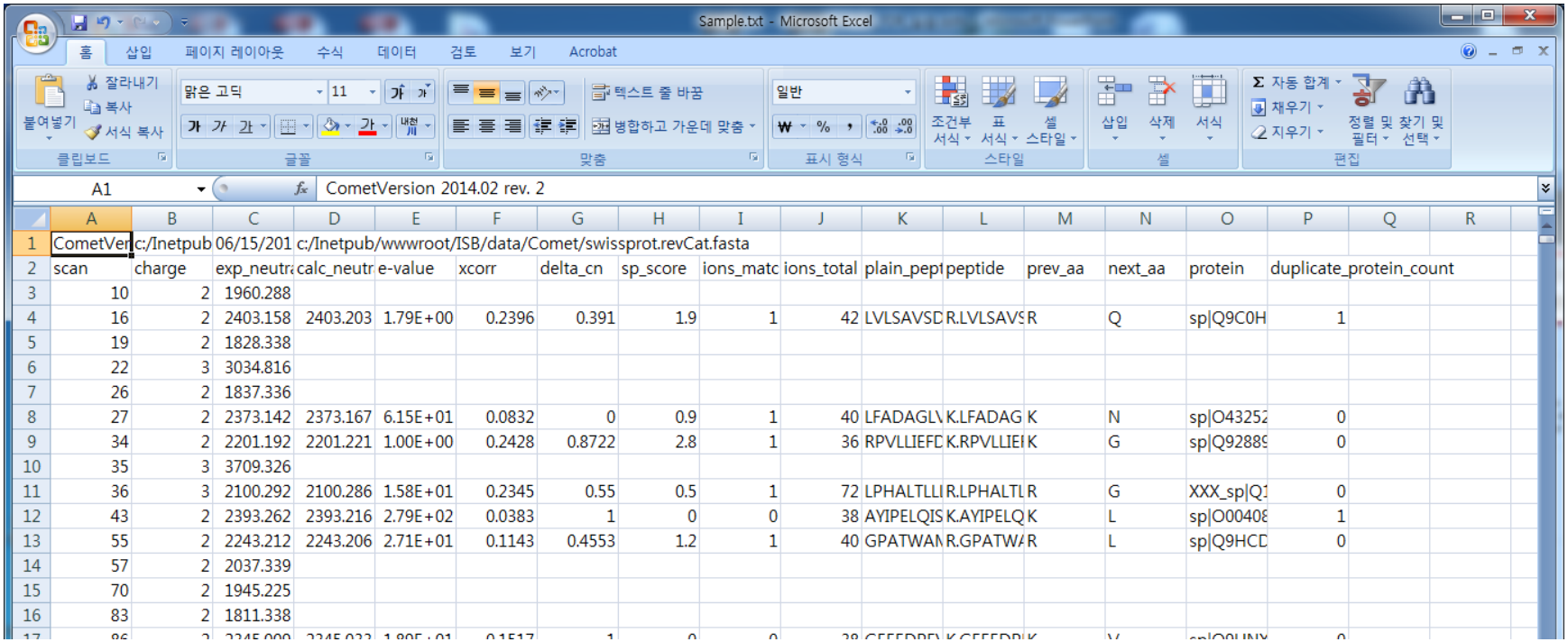
- Cmd 창
  - java -jar CometTD.jar -i Sample.txt –fdr 0.01 –d XXX_

```
C:₩Inetpub₩wwwroot₩ISB₩data₩Comet>java -jar CometTD.jar -i Sample.txt -fdr 0.01
-d XXX_
To Analyze     : Sample.txt
Designated FDR : 0.01
Decoy Proteins starting with XXX_

_TargetDecoy for Charge State: 2, No. of Enzymatic Termini: 2
Threshold=90.2000 | FDR= 0.96% | Target= 418 | Decoy= 4

_TargetDecoy for Charge State: 3, No. of Enzymatic Termini: 2
Threshold=43.0000 | FDR= 0.00% | Target= 31 | Decoy= 0

_TargetDecoy for Charge State: equal to or more than 4, No. of Enzymatic Termini
: 2
Threshold=17.5000 | FDR= 0.00% | Target= 1 | Decoy= 0

_Analysis Report_
At designated FDR 1.00%, Overall No. Identifications: 450
Actual FDR= 0.89% | Target= 450 | Decoy= 4
```

Target/Decoy 방법

```
_TargetDecoy for Charge State: 2, No. of Enzymatic Termini: 2
Threshold=90.2000 | FDR= 0.96% | Target= 418 | Decoy= 4
```

- Threshold 확인
  - C:\Inetpub\wwwroot\ISB\data\Comet 폴더 이동
  - Sample.txt 파일 excel에서 열기

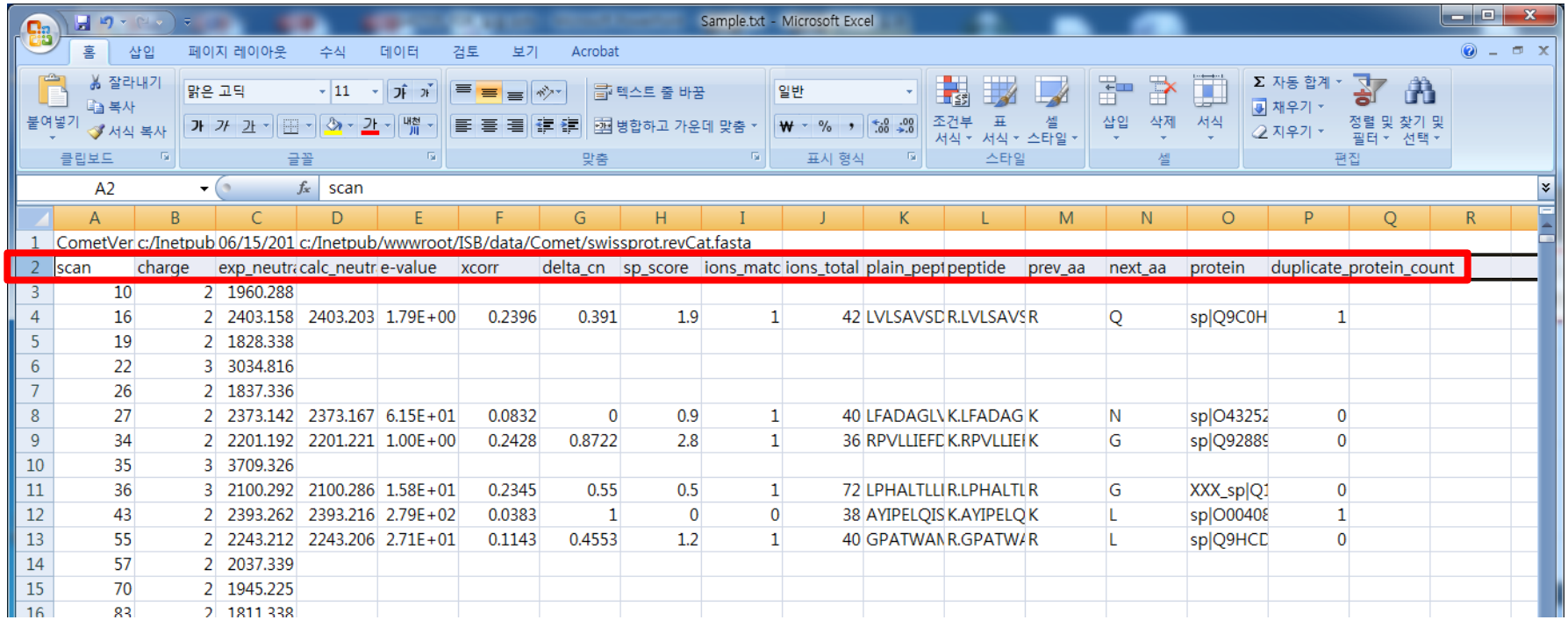


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	CometVer	c:/Inetpub	06/15/201	c:/Inetpub/wwwroot/ISB/data/Comet/swissprot.revCat.fasta														
2	scan	charge	exp_neutr	calc_neutr	e-value	xcorr	delta_cn	sp_score	ions_matc	ions_total	plain_pept	peptide	prev_aa	next_aa	protein	duplicate_protein_count		
3	10	2	1960.288															
4	16	2	2403.158	2403.203	1.79E+00	0.2396	0.391	1.9	1	42	LVLSAVSD	R.LVLSAVS	R	Q	sp|Q9C0H	1		
5	19	2	1828.338															
6	22	3	3034.816															
7	26	2	1837.336															
8	27	2	2373.142	2373.167	6.15E+01	0.0832	0	0.9	1	40	LFADAGL\	K.LFADAG	K	N	sp|O4325	0		
9	34	2	2201.192	2201.221	1.00E+00	0.2428	0.8722	2.8	1	36	RPVLLIEFD	K.RPVLLIE	K	G	sp|Q9288	0		
10	35	3	3709.326															
11	36	3	2100.292	2100.286	1.58E+01	0.2345	0.55	0.5	1	72	LPHALTLLI	R.LPHALTL	R	G	XXX_sp|Q1	0		
12	43	2	2393.262	2393.216	2.79E+02	0.0383	1	0	0	38	AYIPELQIS	K.AYIPELQ	K	L	sp|O00408	1		
13	55	2	2243.212	2243.206	2.71E+01	0.1143	0.4553	1.2	1	40	GPATWAN	R.GPATW/	R	L	sp|Q9HCD	0		
14	57	2	2037.339															
15	70	2	1945.225															
16	83	2	1811.338															

Target/Decoy 방법

_TargetDecoy for Charge State: 2, No. of Enzymatic Termini: 2
Threshold=90.2000 | FDR= 0.96% | Target= 418 | Decoy= 4

- Threshold 확인
  - 두 번째 행 선택
  - Shift + Ctrl + L

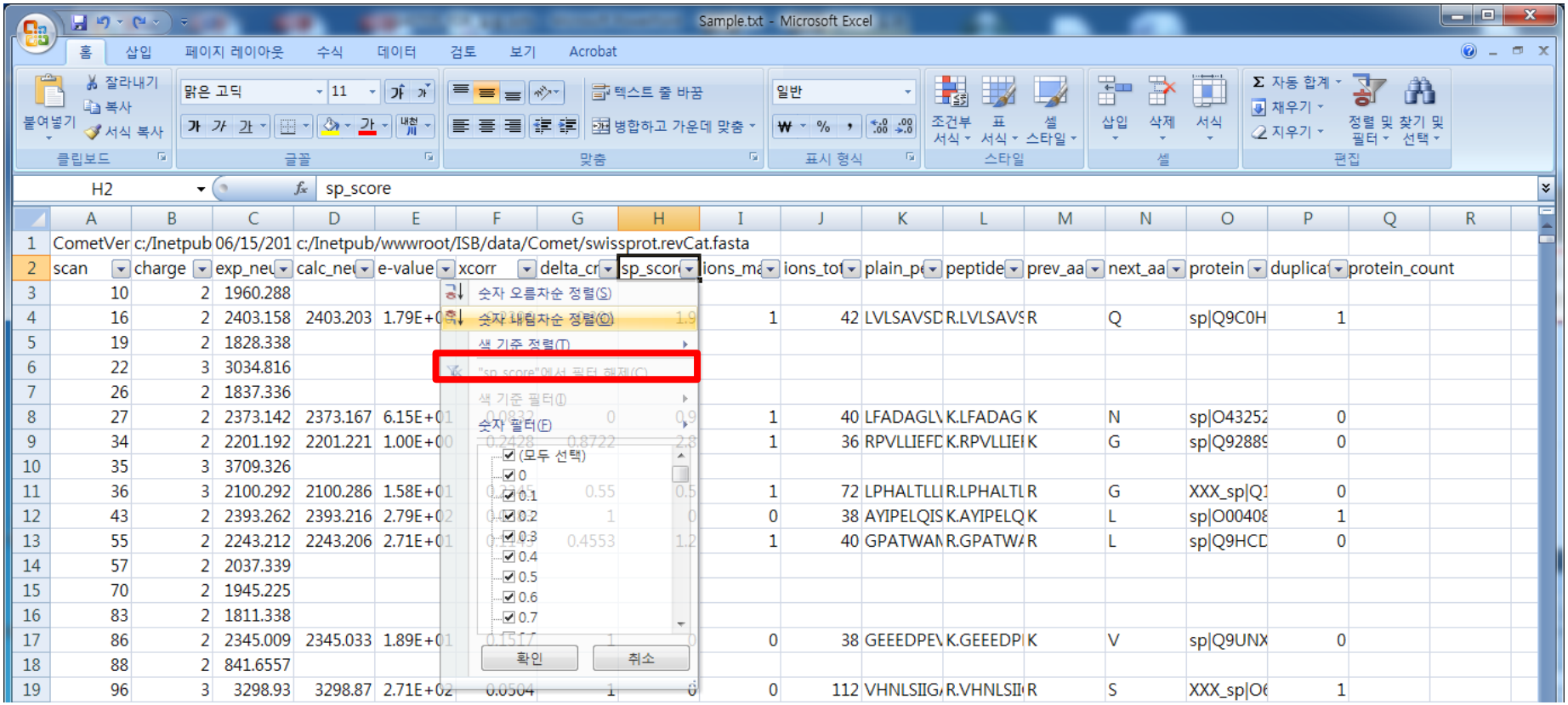


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	CometVer	c:/Inetpub	06/15/201	c:/Inetpub/wwwroot/ISB/data/Comet/swissprot.revCat.fasta														
2	scan	charge	exp_neutr	calc_neutr	e-value	xcorr	delta_cn	sp_score	ions_matc	ions_total	plain_pep	peptide	prev_aa	next_aa	protein	duplicate_protein_count		
3	10	2	1960.288															
4	16	2	2403.158	2403.203	1.79E+00	0.2396	0.391	1.9	1	42	LVLSAVSD	R.LVLSAVS	R	Q	sp|Q9C0H	1		
5	19	2	1828.338															
6	22	3	3034.816															
7	26	2	1837.336															
8	27	2	2373.142	2373.167	6.15E+01	0.0832	0	0.9	1	40	LFADAGLV	K.LFADAG	K	N	sp|O43252	0		
9	34	2	2201.192	2201.221	1.00E+00	0.2428	0.8722	2.8	1	36	RPVLLIEFD	K.RPVLLIE	K	G	sp|Q92889	0		
10	35	3	3709.326															
11	36	3	2100.292	2100.286	1.58E+01	0.2345	0.55	0.5	1	72	LPHALTLLI	R.LPHALTL	R	G	XXX_sp|Q1	0		
12	43	2	2393.262	2393.216	2.79E+02	0.0383	1	0	0	38	AYIPELQIS	K.AYIPELQ	K	L	sp|O00408	1		
13	55	2	2243.212	2243.206	2.71E+01	0.1143	0.4553	1.2	1	40	GPATWAN	R.GPATWA	R	L	sp|Q9HCD	0		
14	57	2	2037.339															
15	70	2	1945.225															
16	83	2	1811.338															

Target/Decoy 방법

```
_TargetDecoy for Charge State: 2, No. of Enzymatic Termini: 2
Threshold=90.2000 | FDR= 0.96% | Target= 418 | Decoy= 4
```

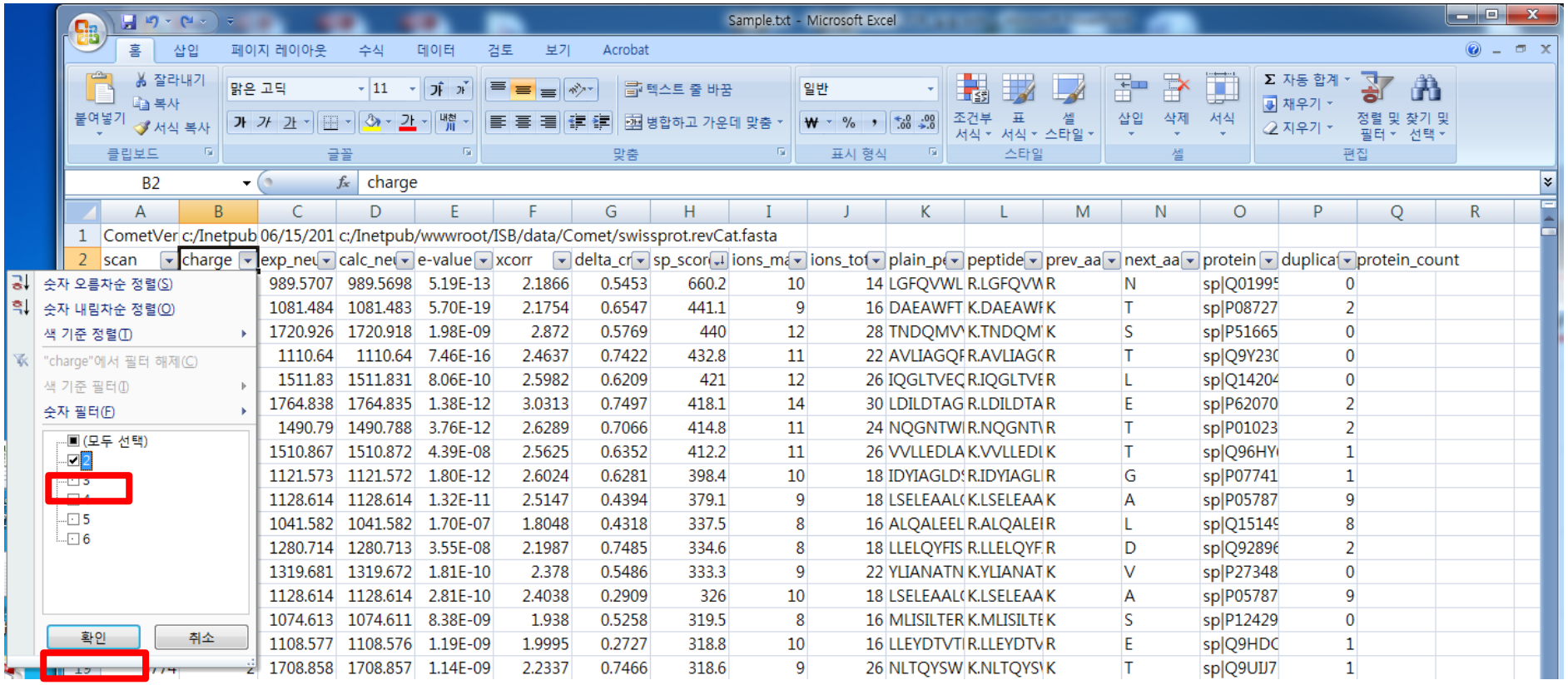
- Threshold 확인
  - sp_score에서 숫자 내림차순 정렬



Target/Decoy 방법

```
_TargetDecoy for Charge State: 2, No. of Enzymatic Termini: 2
Threshold=90.2000 | FDR= 0.96% | Target= 418 | Decoy= 4
```

- Threshold 확인
  - Charge에서 2 선택, 확인 클릭



Target/Decoy 방법

_TargetDecoy for Charge State: 2, No. of Enzymatic Termini: 2
Threshold=90.2000 | FDR= 0.96% | Target= 418 | Decoy= 4

- Threshold 확인
  - Threshold 확인 = 90.2

462	7077	2	1444.721	1444.72	7.36E-05	1.3655	0.5741	92.8	6	24	WDDSGNI	K.WDDSG	K	Q	sp|P26232	7
463	10466	2	1385.77	1385.767	5.75E-05	1.7048	0.5607	92.8	6	22	VGWEQLL	R.VGWEQ	R	T	sp|P12814	3
464	1412	2	685.3932	685.3911	3.71E-05	1.4022	0.0689	92.6	4	10	HVVFGK	K.HVVFGK	K	V	sp|P45877	3
465	4484	2	1027.498	1027.497	1.17E-05	1.1676	0.4557	92.6	6	14	NSLFEYQK	K.NSLFEYQ	K	N	sp|P02671	1
466	3518	2	748.4239	748.4232	1.32E-05	1.1051	0.0448	92.5	5	12	VAFSVAR	R.VAFSVA	R	A	sp|P39023	0
467	7581	2	1151.609	1151.609	2.78E-04	1.7439	0.5957	92.5	6	18	IANVFTNA	R.IANVFTN	R	Y	sp|P05164	2
468	8489	2	989.5718	989.5698	4.56E-07	1.7155	0.4594	92.5	5	14	LGFQVWL	R.LGFQVW	R	N	sp|Q01995	0
469	1843	2	784.4544	784.4443	1.02E-02	1.3165	0.2394	92	5	14	VLNPSGAI	R.VLNPSG	R	T	sp|Q14315	1
470	11395	2	1320.671	1320.671	6.84E-05	1.6685	0.6403	92	7	26	STSGGTA/	K.STSGGT/	K	D	sp|P01857	1
471	4086	2	1295.628	1295.626	2.50E-05	1.8041	0.6137	91.9	6	22	GSYVSIHS	K.GSYVSIH	K	D	sp|Q13838	2
472	7256	2	1418.743	1418.741	7.42E-08	1.6258	0.5878	91.8	10	22	QAQEYEA	R.QAQEYE	R	V	sp|P05783	3
473	2180	2	1030.567	1030.566	8.55E-06	1.4848	0.6746	91.7	5	18	LLEAAAQS	R.LLEAAA(	R	G	sp|Q15149	8
474	6194	2	1091.565	1091.561	4.04E-07	1.6977	0.5773	91.5	5	16	AELNEFLTI	K.AELNEFL	K	E	sp|P23396	0
475	7813	2	1734.893	1734.879	1.47E-04	2.0746	0.6146	91.3	7	34	LGNTISSLF	K.LGNTISS	K	E	sp|Q9Y4L1	0
476	6965	2	965.6288	965.6274	5.55E-06	1.5159	0.6849	91.1	6	16	LPLILVGNI	R.LPLILVG	R	S	sp|Q8IXI2|	5
477	9717	2	1352.832	1352.839	8.60E-05	1.825	0.82	91.1	5	22	NLQNLLIL	R.NLQNLL	R	A	sp|Q00610	3
478	7836	2	1083.667	1083.665	1.25E-01	1.0124	0.1871	90.7	6	18	LVLPSLISS	K.LVLPSLI	K	I	sp|Q13228	1
479	9962	2	1360.692	1360.691	3.76E-07	1.7782	0.6347	90.7	6	20	SLFDYFLTI	R.SLFDYFL	R	C	sp|Q9NQ8	0
480	4987	2	1409.625	1409.625	2.17E-06	1.6197	0.589	90.6	6	18	YFYNQEEY	R.YFYNQEI	R	F	sp|Q30134	1
481	3018	2	741.45	741.4497	3.08E-04	1.0032	0.1117	90.3	4	10	LNQLVR	K.LNQLVR	K	R	XXX_sp|Q8	0
482	3485	2	837.5083	837.496	1.00E+00	0.9294	0.028	90.3	6	14	LPEGVVPK	R.LPEGVV	R	S	sp|Q9NVS	1
483	6356	2	1399.638	1399.641	1.33E-08	1.6915	0.2732	90.3	7	22	YGGDEIPF	K.YGGDEI	K	V	sp|P21333	1
484	8920	2	1518.828	1518.823	1.01E-02	1.5423	0.5992	90.3	6	28	FGANAILG	K.FGANAI	K	A	sp|P06733	5
485	9353	2	1502.789	1502.79	8.83E-09	2.221	0.6328	90.3	6	24	LVSDFMV	K.LVSDFM	K	N	sp|P54819	5
486	6793	2	1176.638	1176.633	1.84E-02	1.0946	0.1355	90.2	5	18	VLYCAAQ	R.VLYCAA	R	A	XXX_sp|Q8	1
487	8234	2	1239.684	1239.684	3.30E-02	1.2782	0.334	89.7	4	18	HRPYQVIT	R.HRPYQV	R	V	sp|Q96MI9	2
488	3884	2	1549.825	1549.821	8.18E-06	2.1249	0.6669	89.6	8	30	LVGGPVA	K.LVGGPV	K	G	sp|O15230	0
489	10963	2	1070.548	1070.547	3.73E-06	1.1392	0.4069	89.6	5	16	MFLSFPTT	R.MFLSFP	R	T	sp|P69905	0
490	8866	2	1557.806	1557.804	6.01E-07	2.0982	0.4725	89.4	7	26	FLSSSLYTA	K.FLSSSLY	K	R	sp|Q16706	0
491	2192	2	1235.507	1235.505	1.73E-07	1.8023	0.6682	89.3	5	22	DGEEAGA	R.DGEEAG	R	T	sp|P30101	0
492	4105	2	972.5608	972.5604	4.75E-06	1.2805	0.4463	89.3	5	16	NLQVALTS	K.NLQVAL	K	Q	sp|P12111	2

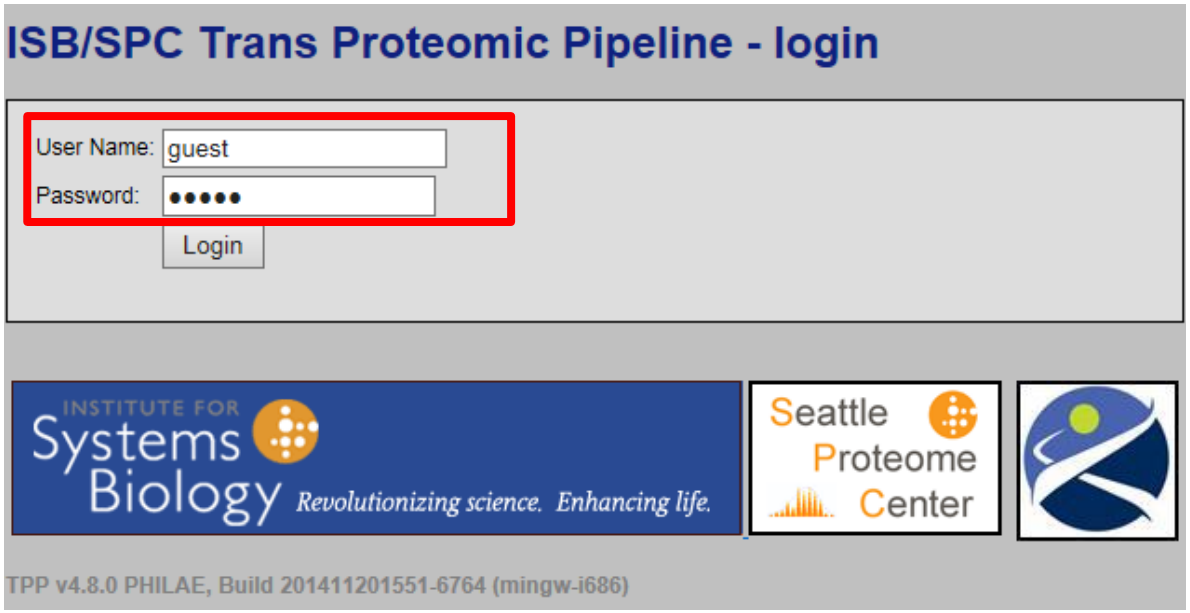
TPP 실행 및 로그인

- TPP 실행
  - 바탕화면에 아이콘 실행

TPP 실행 및 로그인

- TPP 실행
  - 바탕화면에 아이콘 실행

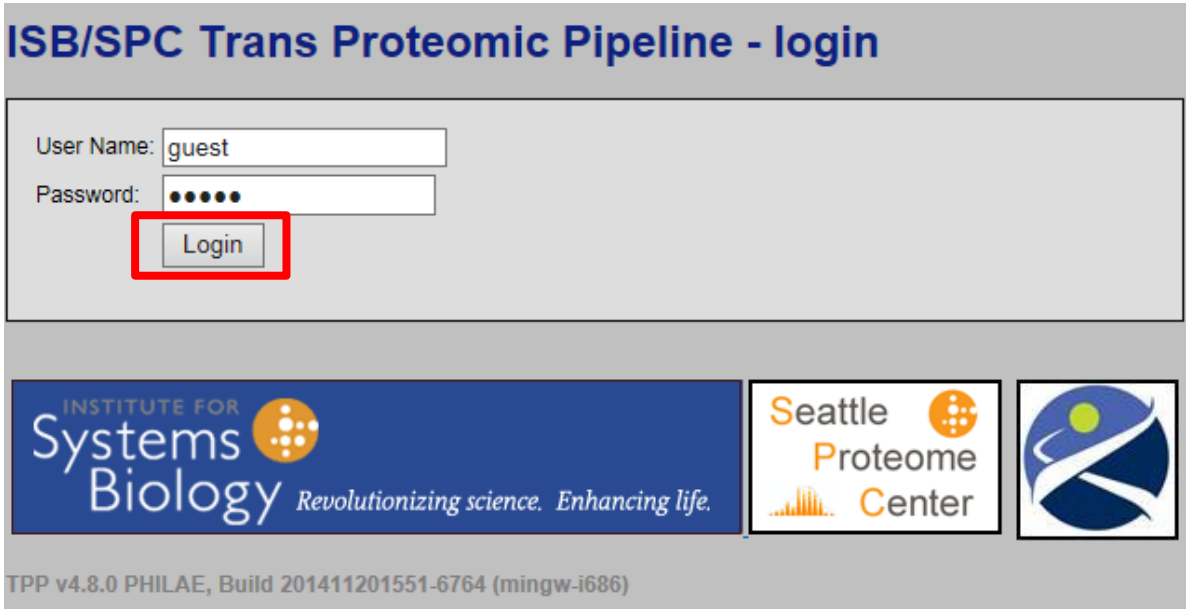
- TPP 웹페이지에서 로그인
  - User Name: guest
  - Password: guest
  - Login 클릭



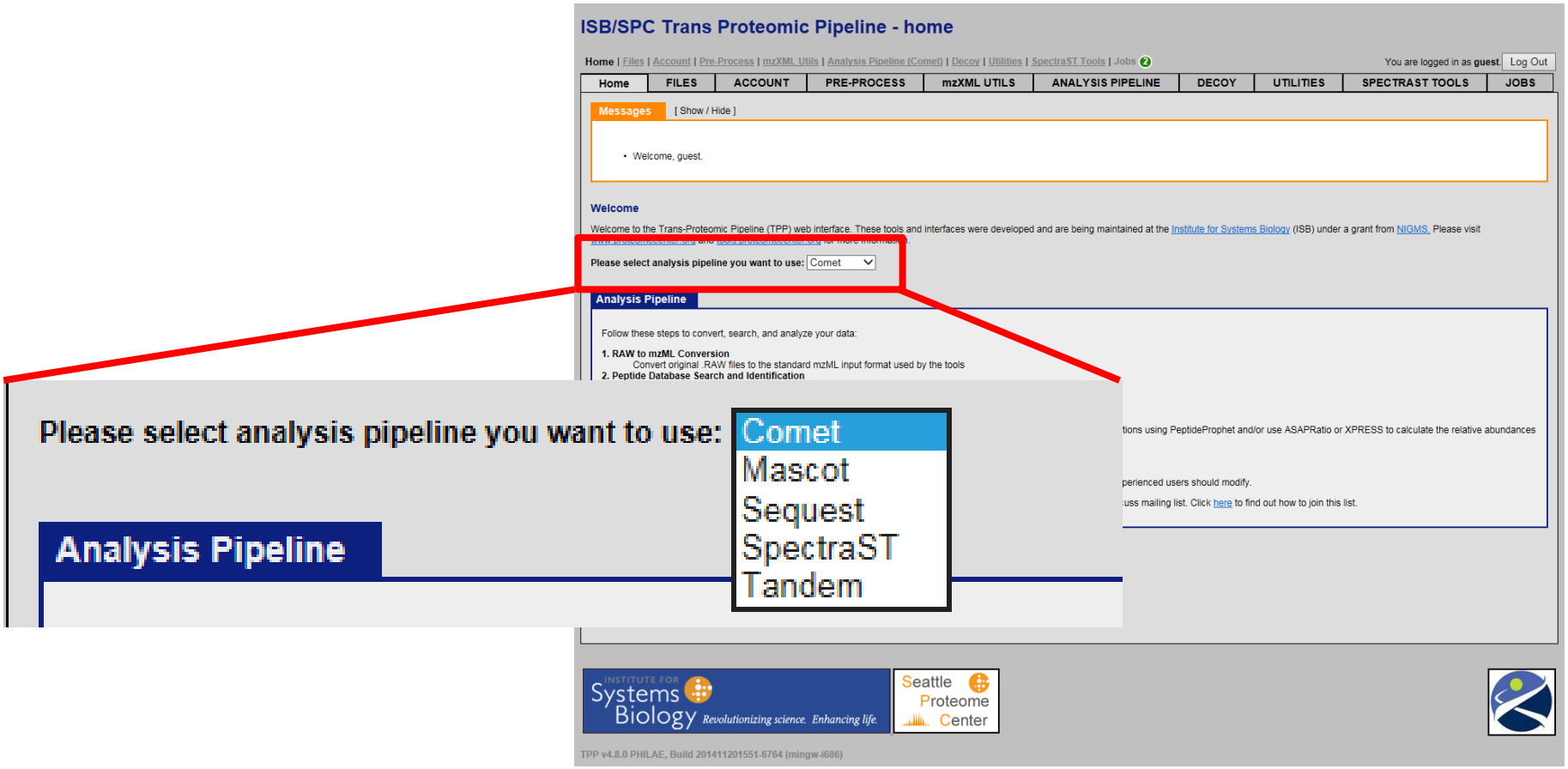
TPP 실행 및 로그인

- TPP 실행
  - 바탕화면에 아이콘 실행

- TPP 웹페이지에서 로그인
  - User Name: guest
  - Password: guest
  - Login 클릭

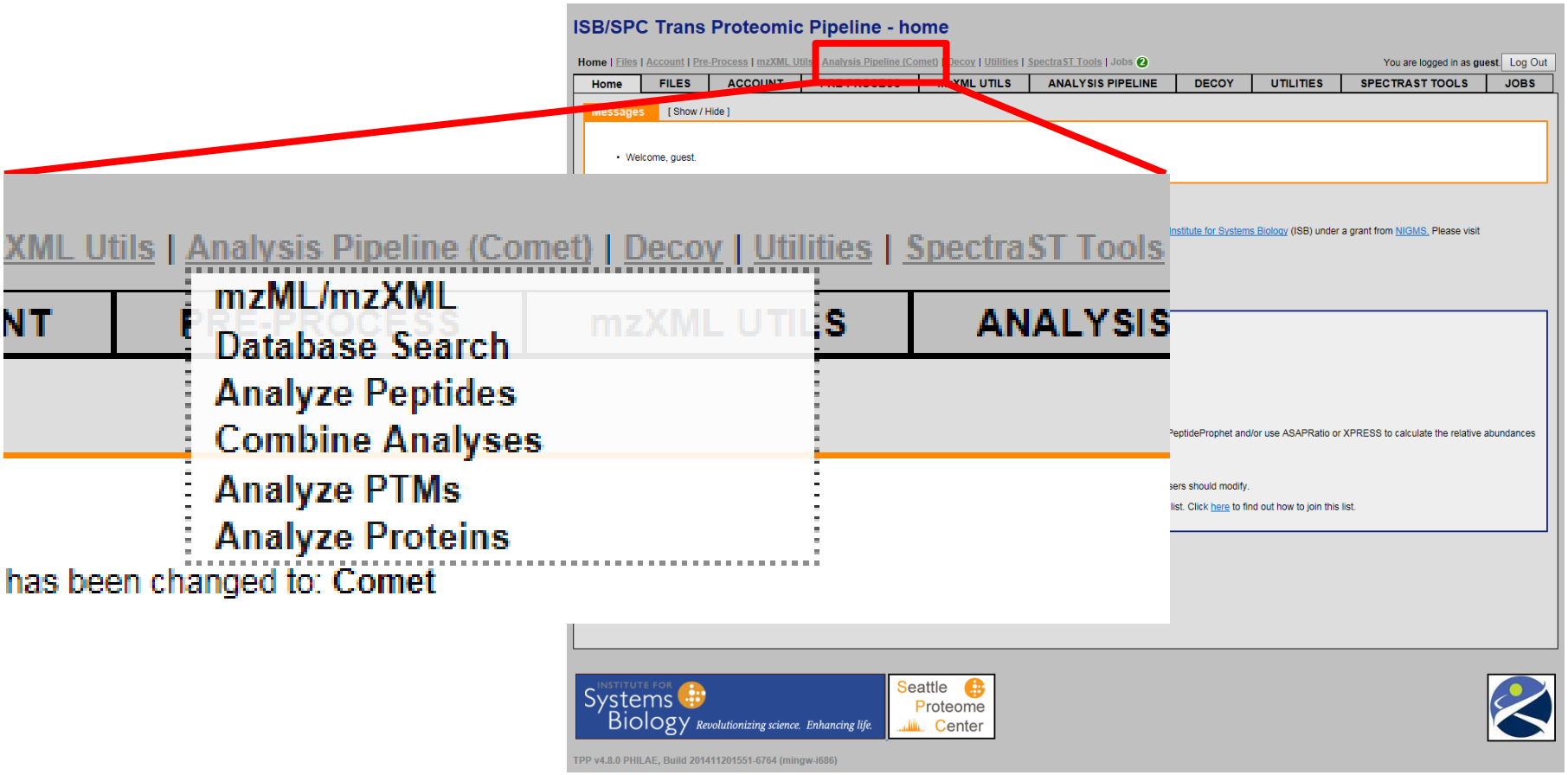


- Analysis pipeline 확인
  - Comet으로 선택



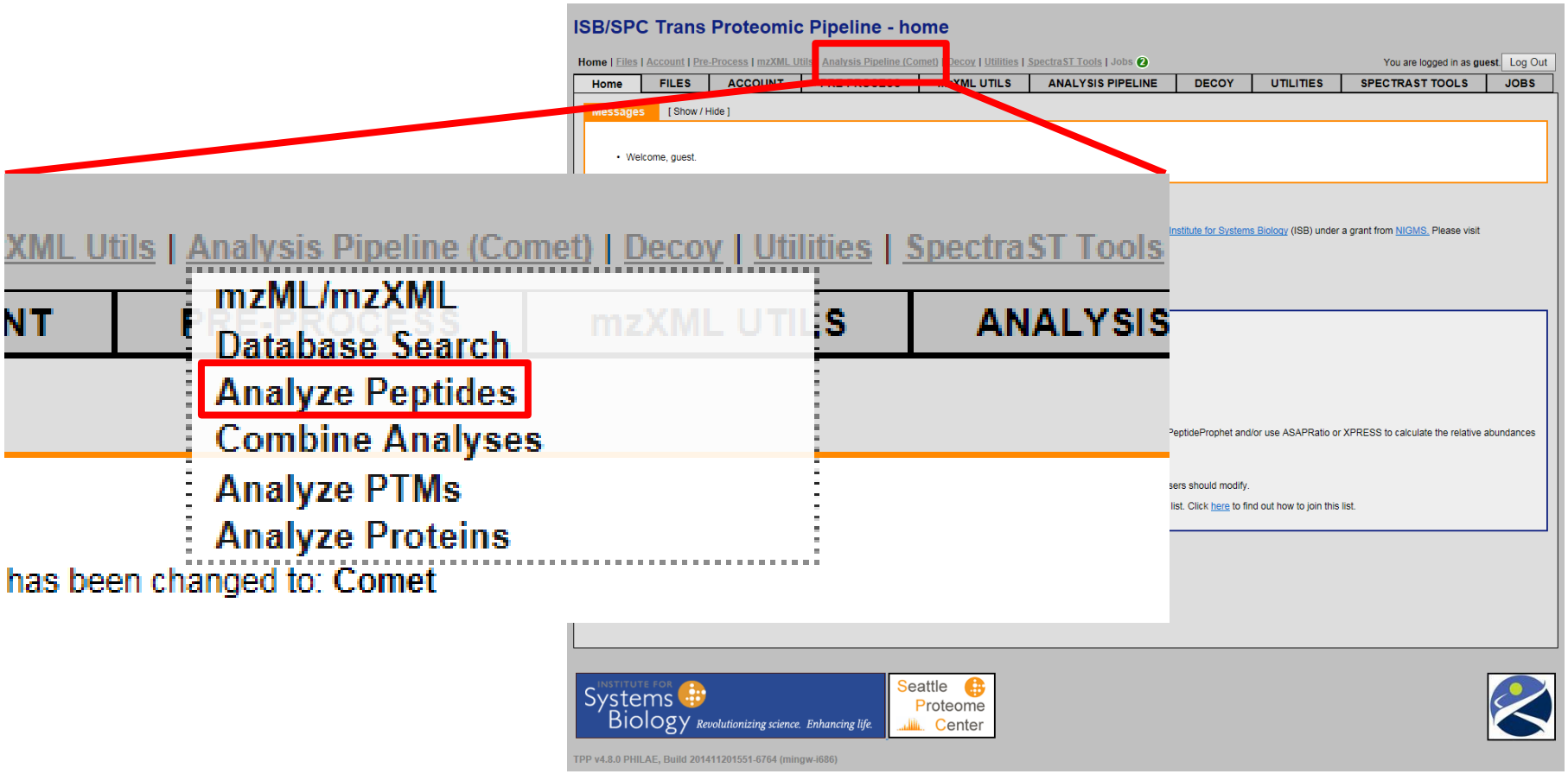
PeptideProphet 실행

- Analysis pipeline 확인
  - Analysis Pipeline(Comet)



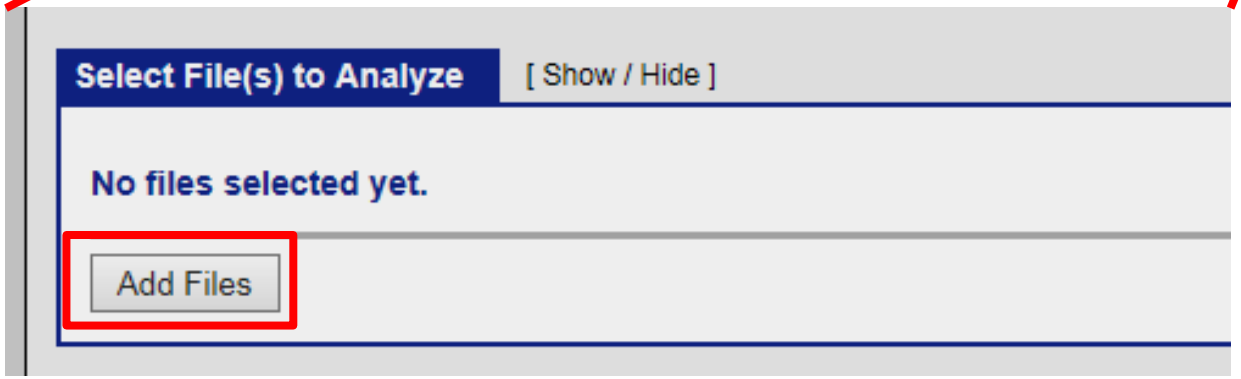
PeptideProphet 실행

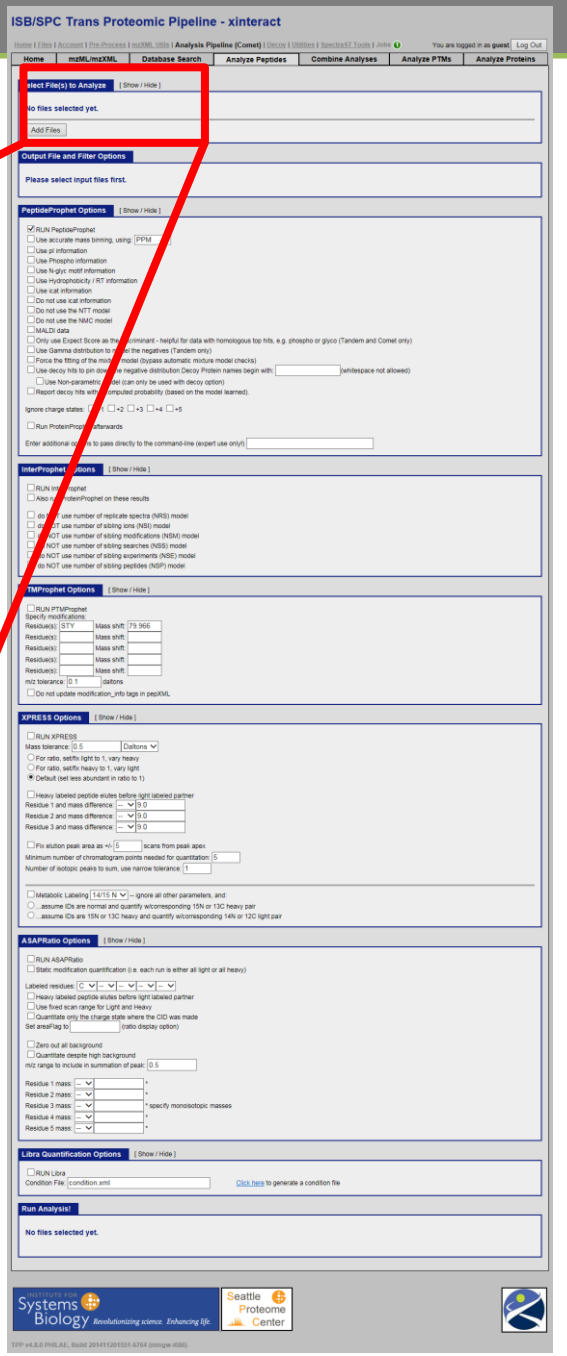
- Analysis pipeline 확인
  - Analysis Pipeline(Comet)
  - Analyze Peptides 클릭



PeptideProphet 실행

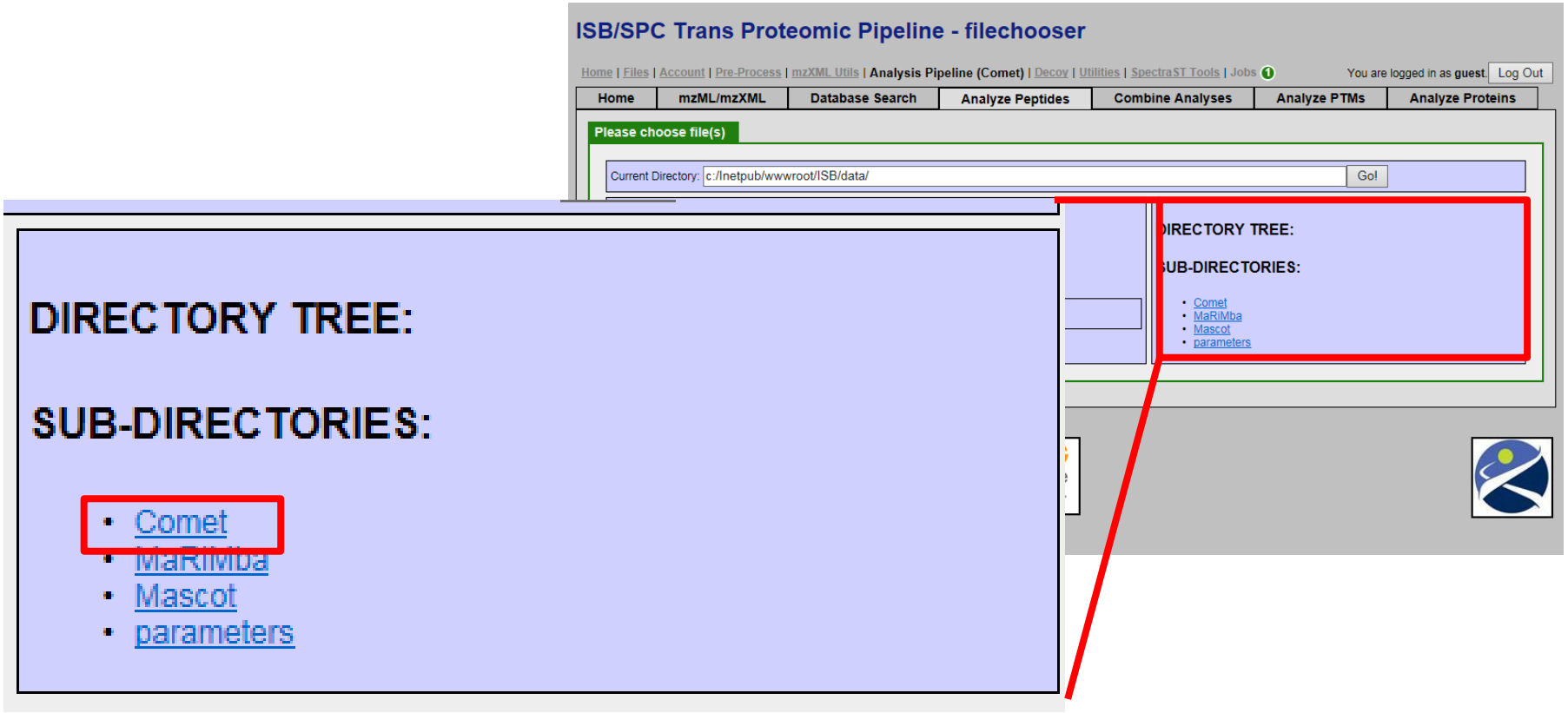
- Peptide identification 결과 선택
  - Add files 클릭





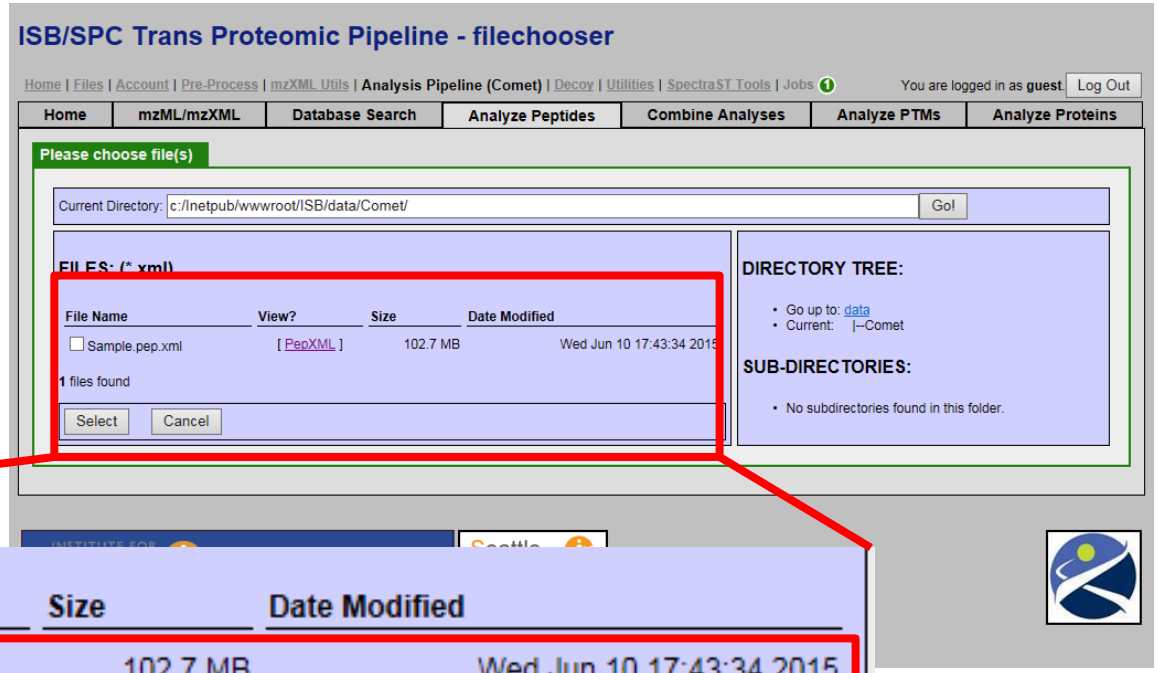
PeptideProphet 실행

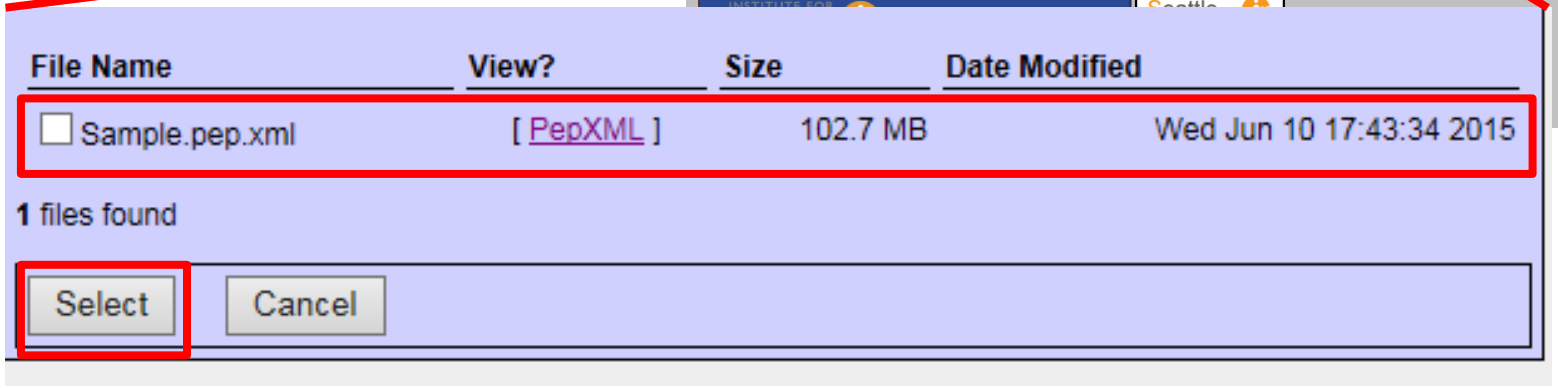
- Peptide identification 결과 선택
  - Comet 클릭



PeptideProphet 실행

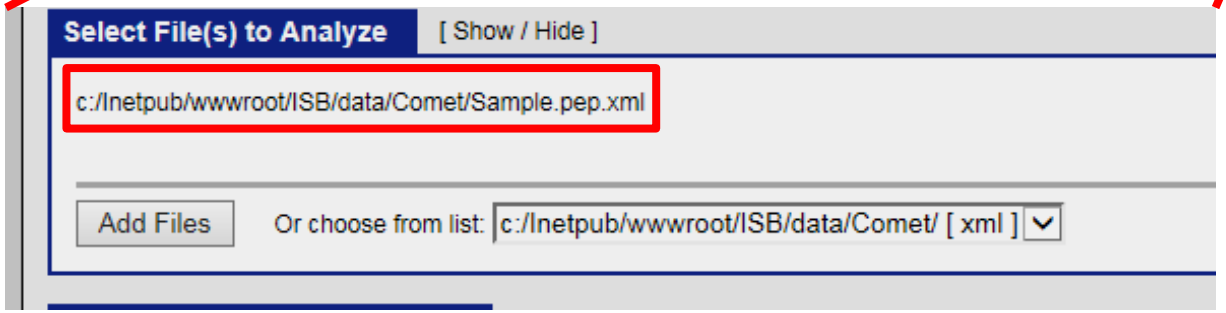
- Peptide identification 결과 선택
  - Sample.pep.xml 선택
  - Select 클릭

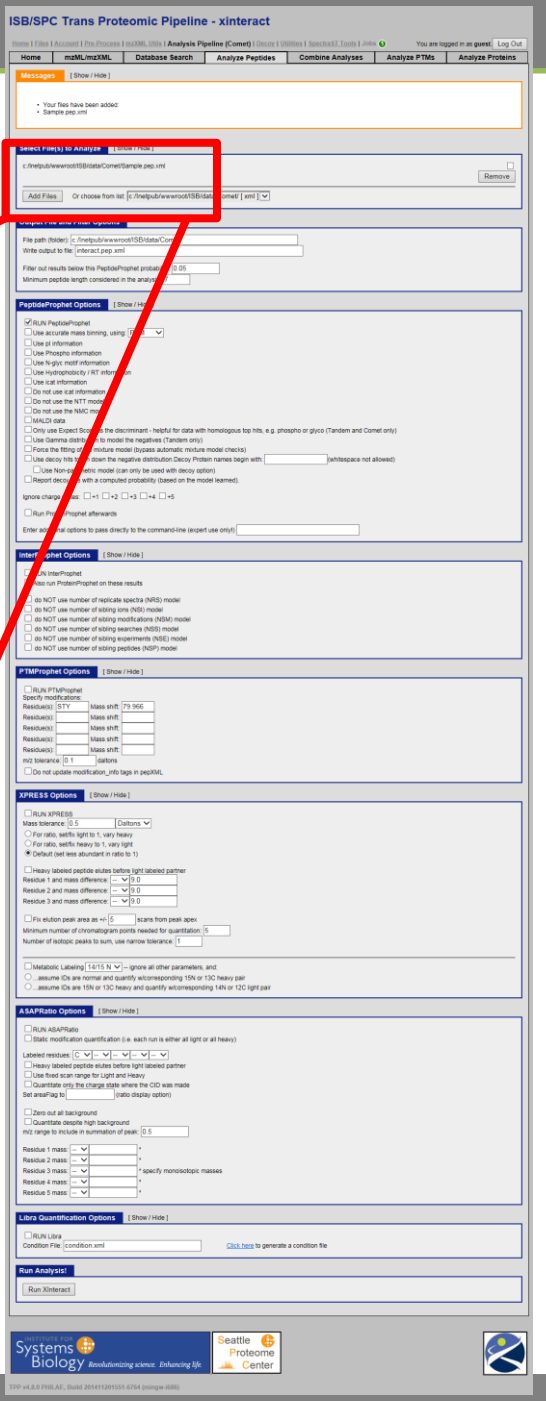




PeptideProphet 실행

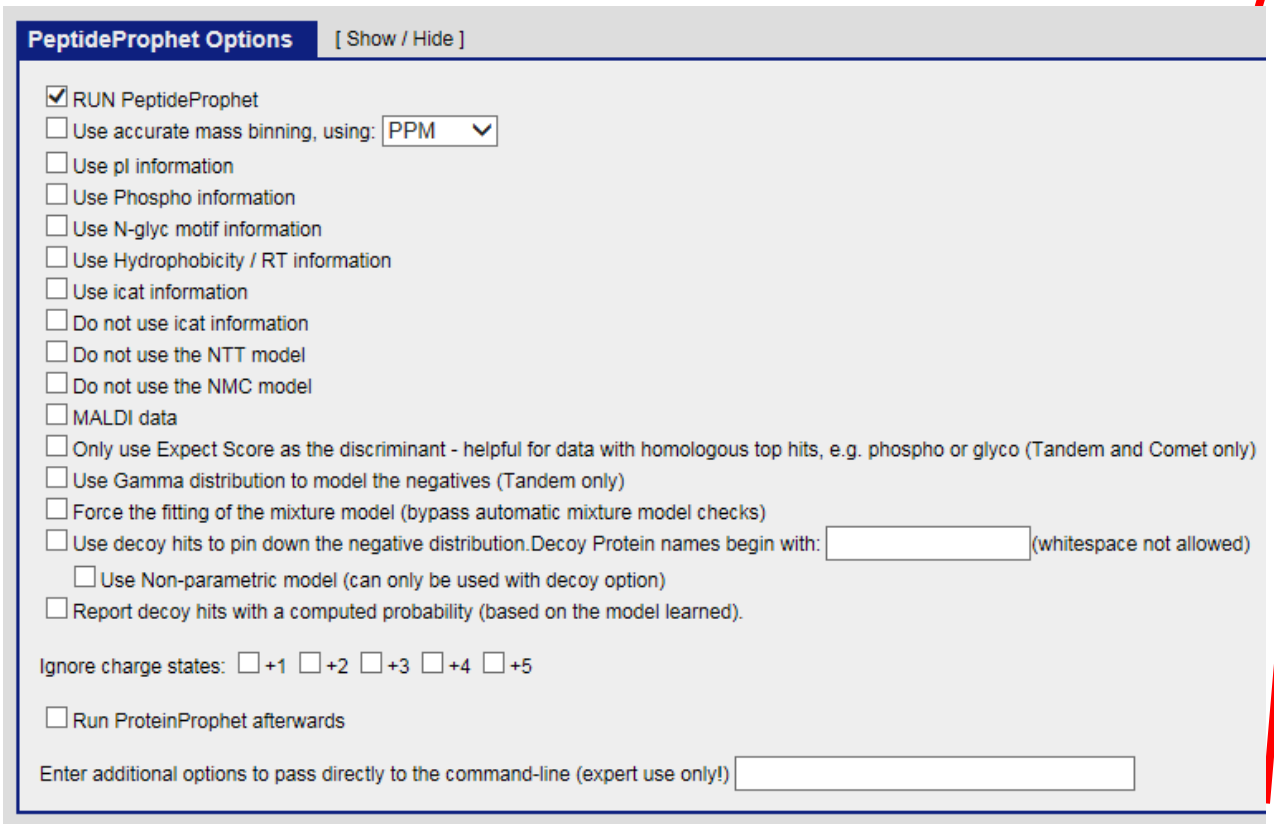
- Peptide identification 결과 확인

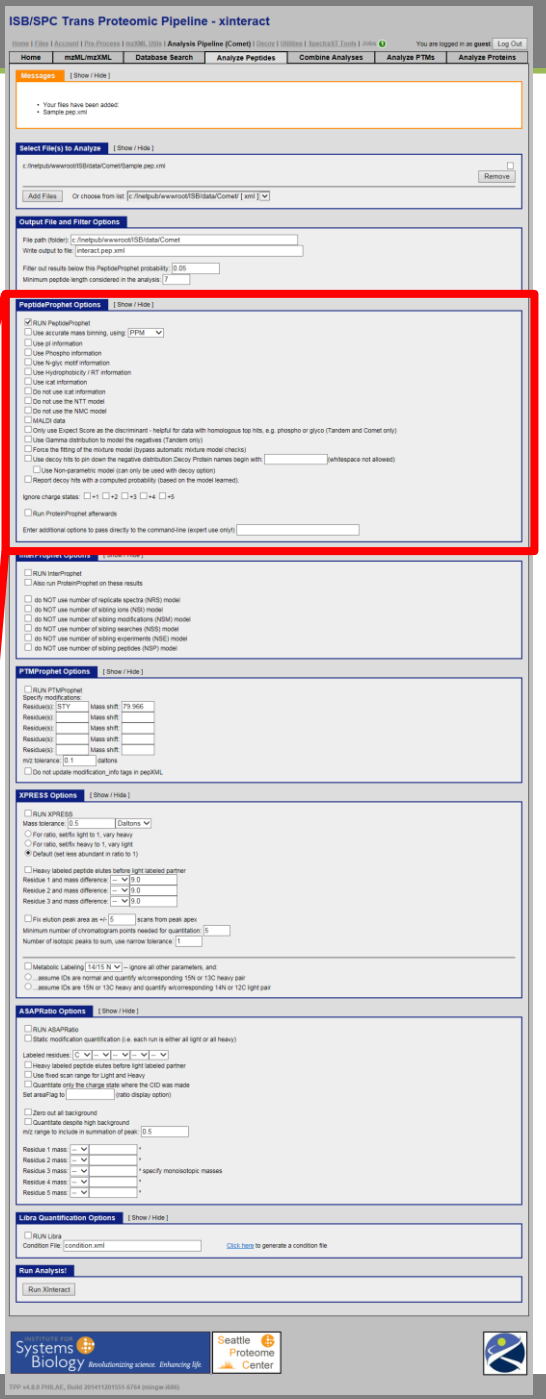




PeptideProphet 실행

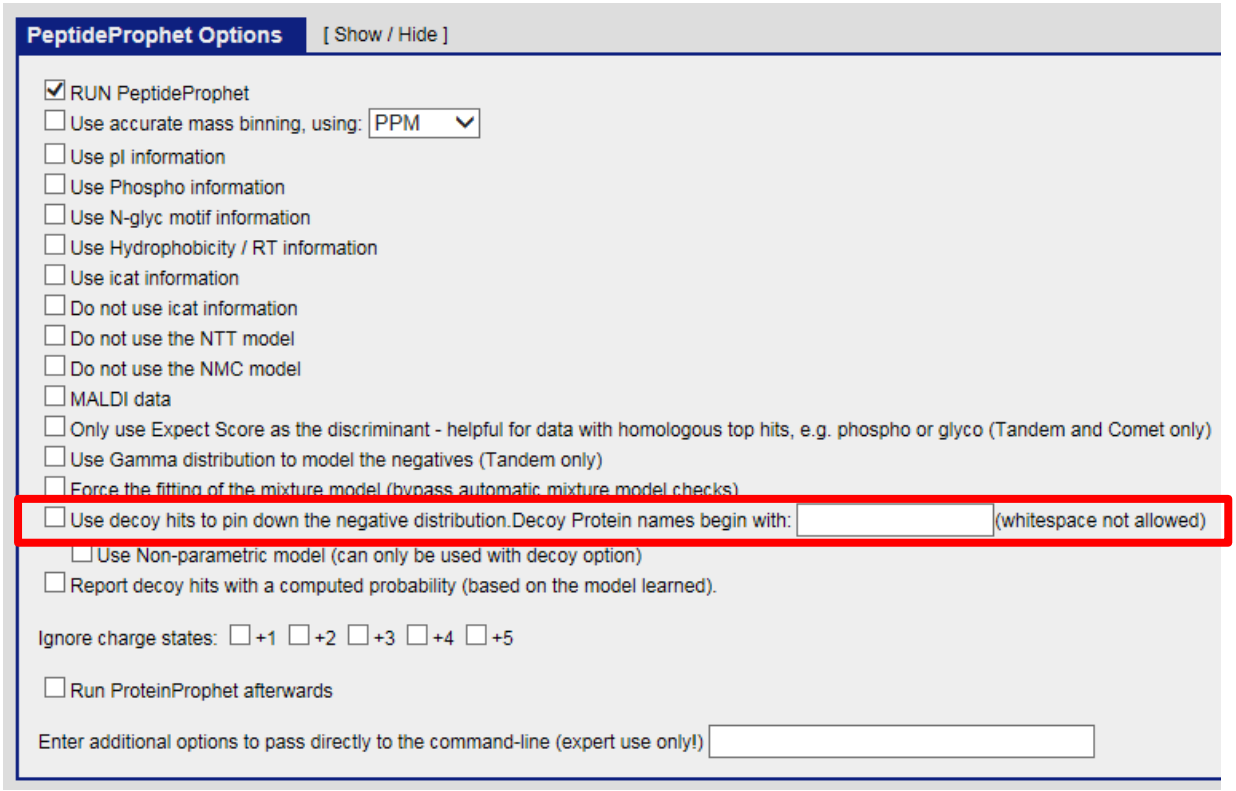
- PeptideProphet option





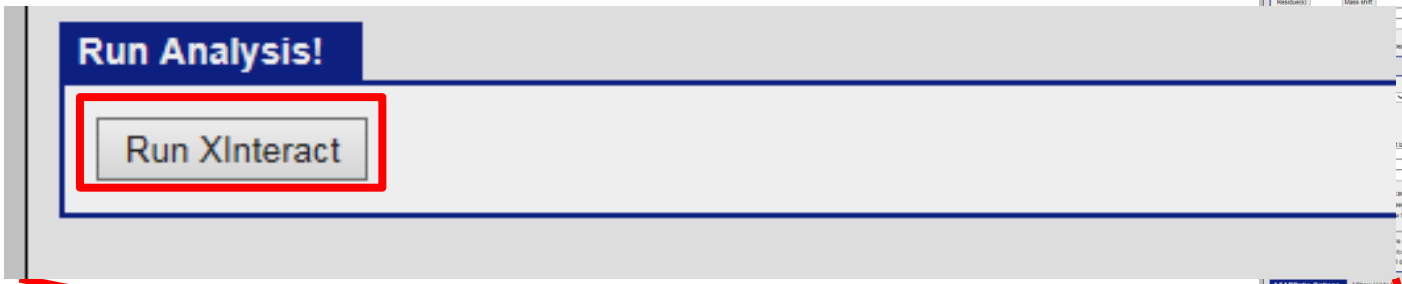
PeptideProphet 실행

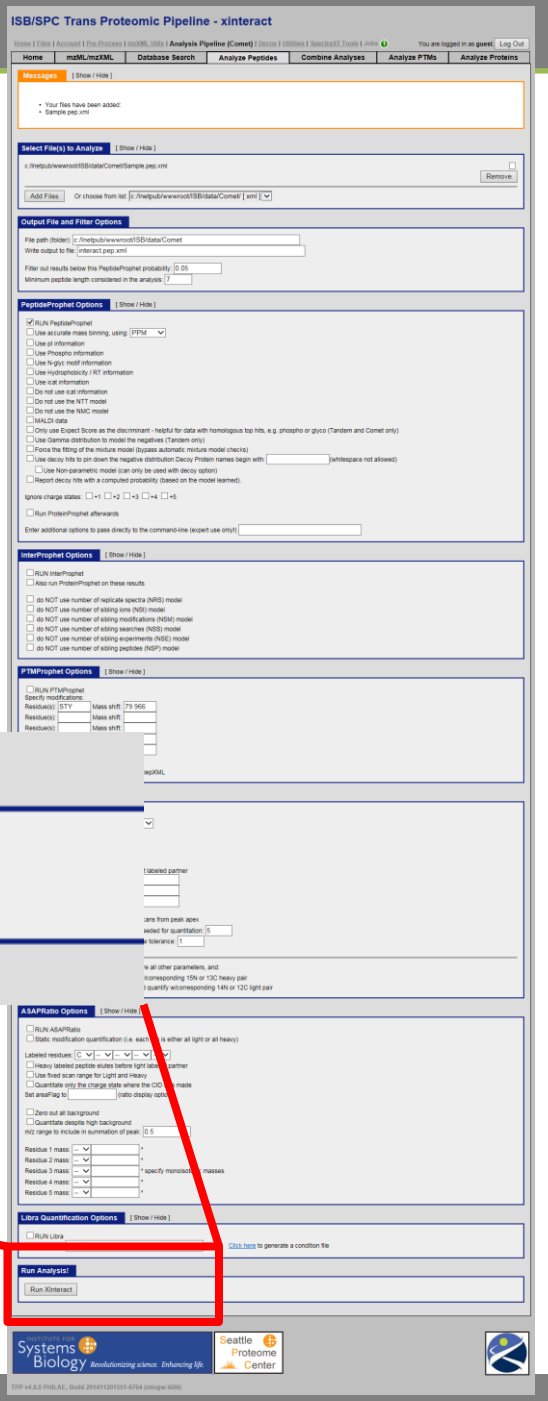
- PeptideProphet option
  - Use decoy hits to pin down the negative distribution.
  - Decoy Protein names begin with: XXX_



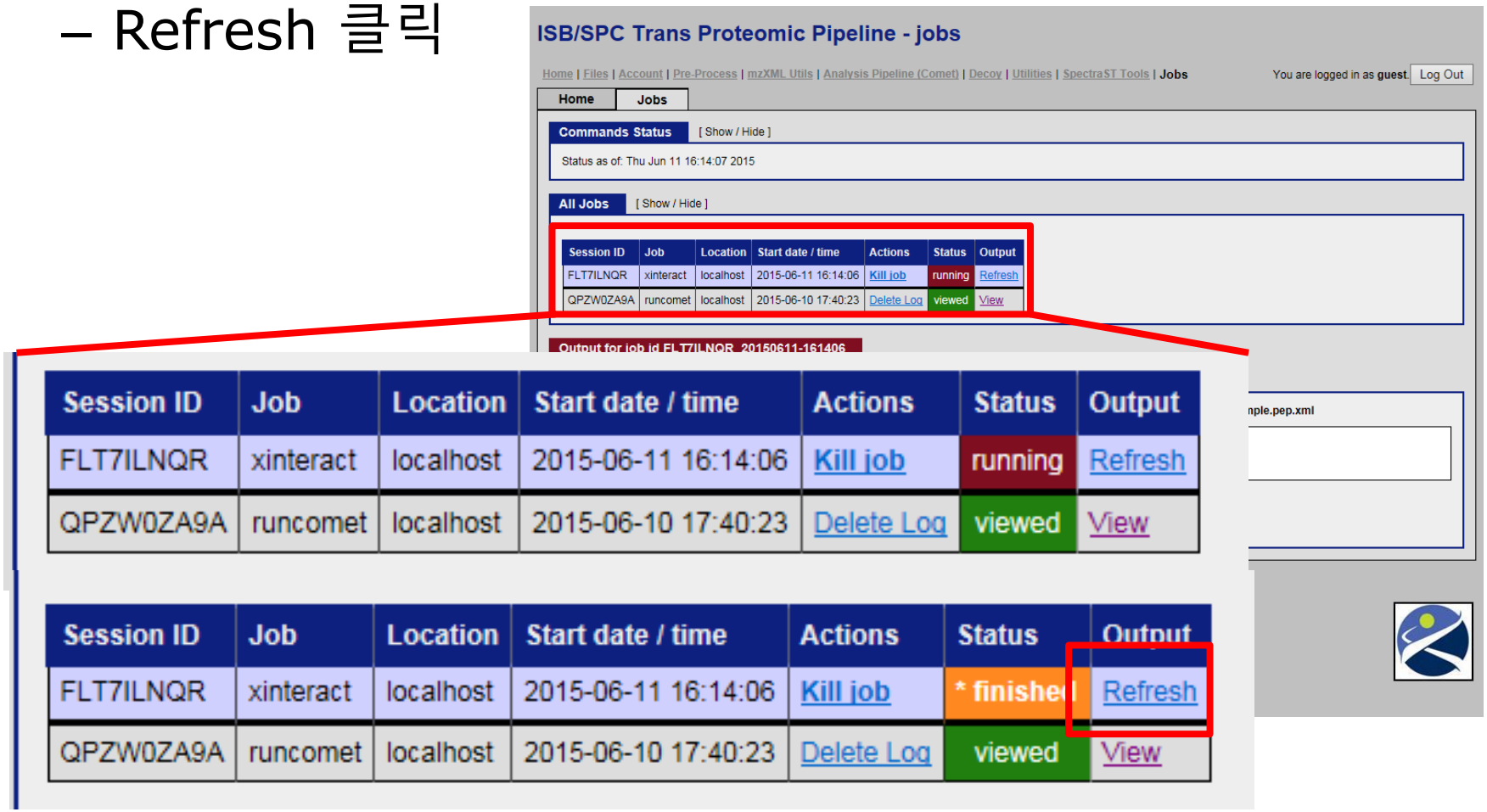
PeptideProphet 실행

- PeptideProphet 실행
  - Run XInteract 클릭

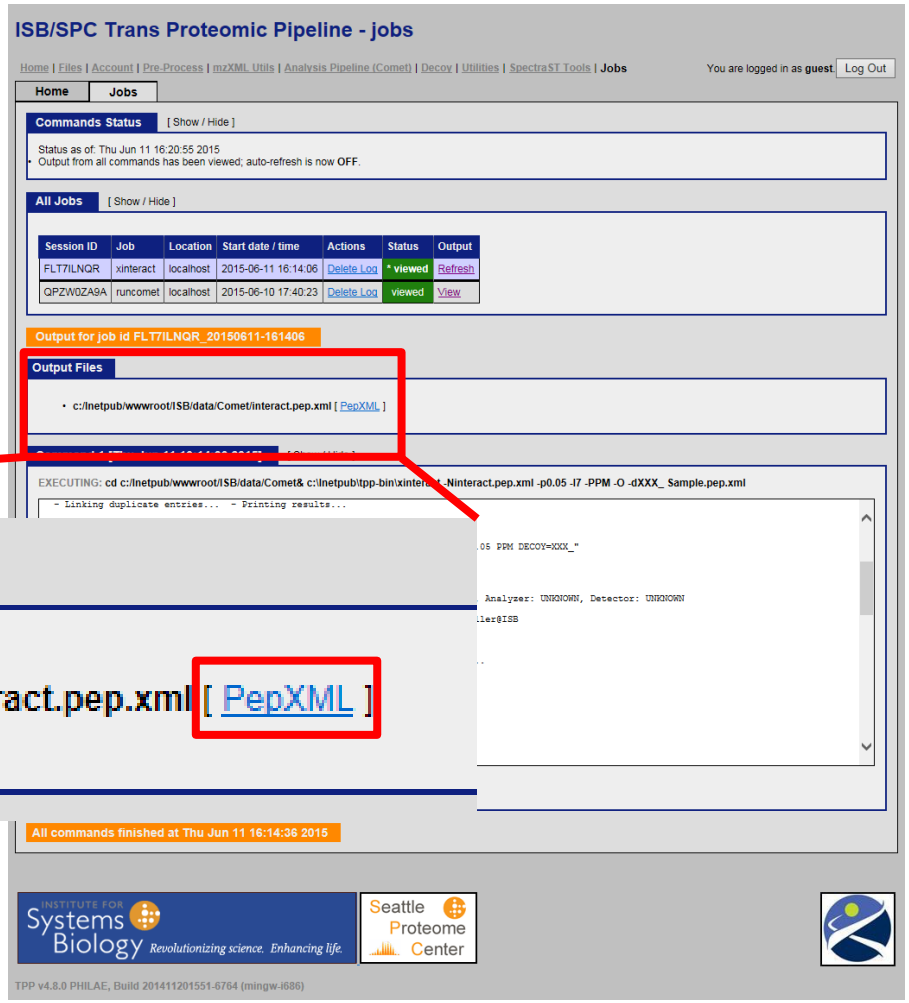




- PeptideProphet 실행
  - running ➔ finished
  - Refresh 클릭



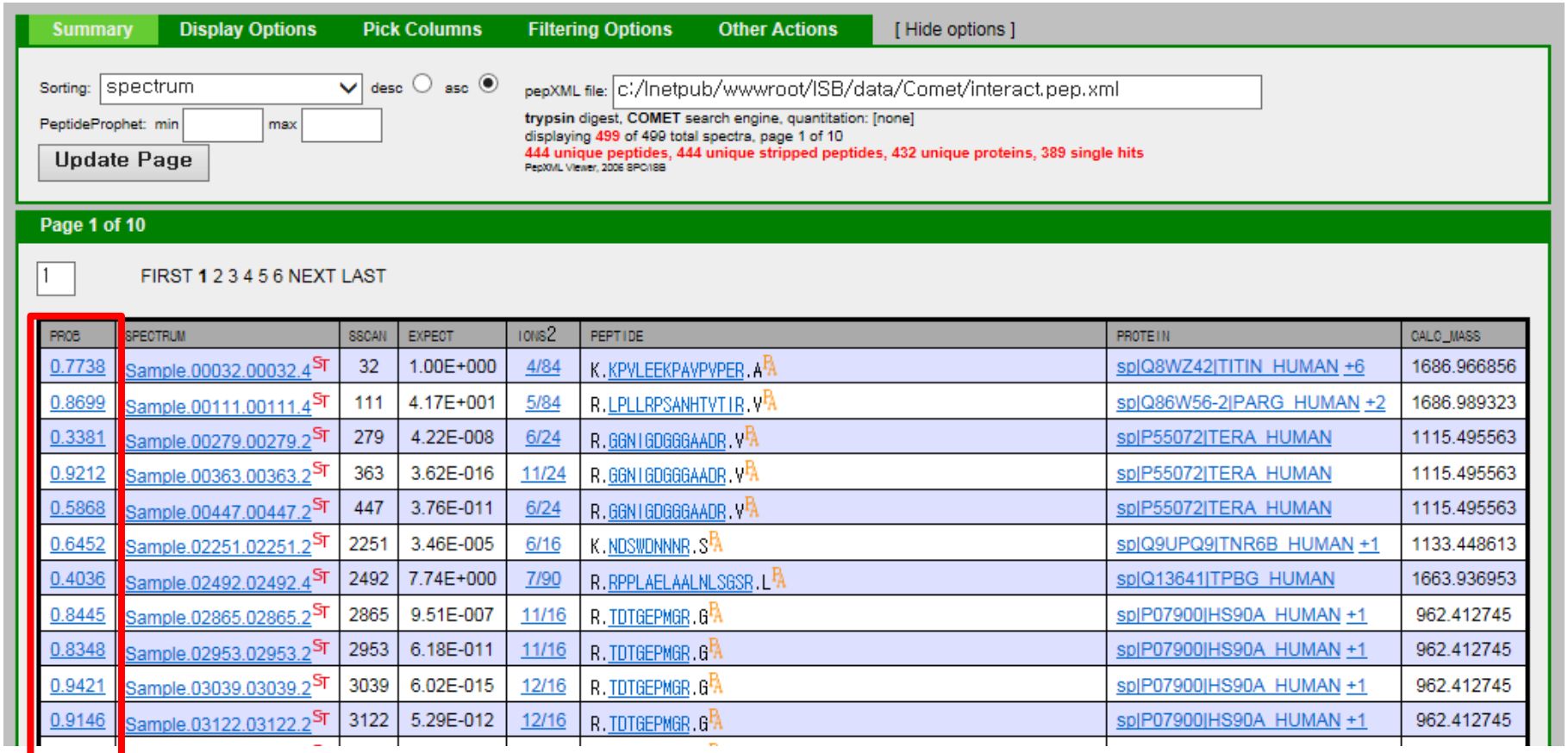
- PeptideProphet 실행 결과 확인
  - PepXML 클릭



Output Files
c:/Inetpub/wwwroot/ISB/data/Comet/interact.pep.xml [PepXML]

PeptideProphet 실행

- PeptideProphet 실행 결과 확인
  - PROB 열



Summary | Display Options | Pick Columns | Filtering Options | Other Actions | [Hide options]

Sorting: spectrum desc asc

PeptideProphet: min max

Update Page

pepXML file: c:/Inetpub/wwwroot/ISB/data/Comet/interact.pep.xml

trypsin digest, **COMET** search engine, quantitation: [none]
displaying 499 of 499 total spectra, page 1 of 10
444 unique peptides, 444 unique stripped peptides, 432 unique proteins, 389 single hits
PepXML Viewer, 2006 SPC/ISB

Page 1 of 10

1 FIRST **1** 2 3 4 5 6 NEXT LAST

PROB	SPECTRUM	SSCAN	EXPECT	IONS2	PEPTIDE	PROTEIN	CALC_MASS
0.7738	Sample.00032.00032.4 ST	32	1.00E+000	4/84	K.KPVLEEKPAVPVPER.A PA	sp|Q8WZ42|TITIN_HUMAN +6	1686.966856
0.8699	Sample.00111.00111.4 ST	111	4.17E+001	5/84	R.LPLLRPSANHTVTIR.V PA	sp|Q86W56-2|PARG_HUMAN +2	1686.989323
0.3381	Sample.00279.00279.2 ST	279	4.22E-008	6/24	R.GGNIGDGGGAADR.V PA	sp|P55072|TERA_HUMAN	1115.495563
0.9212	Sample.00363.00363.2 ST	363	3.62E-016	11/24	R.GGNIGDGGGAADR.V PA	sp|P55072|TERA_HUMAN	1115.495563
0.5868	Sample.00447.00447.2 ST	447	3.76E-011	6/24	R.GGNIGDGGGAADR.V PA	sp|P55072|TERA_HUMAN	1115.495563
0.6452	Sample.02251.02251.2 ST	2251	3.46E-005	6/16	K.NDSWDNNNR.S PA	sp|Q9UPQ9|TNR6B_HUMAN +1	1133.448613
0.4036	Sample.02492.02492.4 ST	2492	7.74E+000	7/90	R.RPPLAELAALNLSGSR.L PA	sp|Q13641|TPBG_HUMAN	1663.936953
0.8445	Sample.02865.02865.2 ST	2865	9.51E-007	11/16	R.TDTGEPMGR.G PA	sp|P07900|HS90A_HUMAN +1	962.412745
0.8348	Sample.02953.02953.2 ST	2953	6.18E-011	11/16	R.TDTGEPMGR.G PA	sp|P07900|HS90A_HUMAN +1	962.412745
0.9421	Sample.03039.03039.2 ST	3039	6.02E-015	12/16	R.TDTGEPMGR.G PA	sp|P07900|HS90A_HUMAN +1	962.412745
0.9146	Sample.03122.03122.2 ST	3122	5.29E-012	12/16	R.TDTGEPMGR.G PA	sp|P07900|HS90A_HUMAN +1	962.412745