PeptideProphet: Validation of Peptide Assignments to MS/MS Spectra

Andrew Keller
Day 2
October 17, 2006
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

![Pie chart showing Peptide Identification Effort]

- Manual validation: 75.0%
- Sample prep: 8.3%
- Instrument time: 8.3%
- Database search: 8.3%
- Validation: 8.3%
(Un)reliability of Manual Validation

--- Manual Authenticators ---

Search Results

Correct Validation Incorrect Validation Validation Withheld
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, Identitiescore, Homologyscore, NTT (number of tryptic termini)

- Accept all results that satisfy:
  Ionscore > Identitiescore
## Traditional Filtering Criteria

<table>
<thead>
<tr>
<th>Each Mascot search result has:</th>
</tr>
</thead>
<tbody>
<tr>
<td>– Ionscore, Identityscore, Homologyscore, NTT (number of tryptic termini)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accept all results that satisfy:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ionscore &gt; Identityscore (NTT = 2)</td>
</tr>
</tbody>
</table>
Traditional Filtering Criteria

- Each Mascot search result has:
  - Ionscore, Identitiescore, Homologyscore, NTT (number of tryptic termini)

- Accept all results that satisfy:
  Ionscore > Identitiescore (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 \((\text{Ionscore} - \text{Identityscore})\) difference

- Secondarily, use \((\text{Ionscore} - \text{Homologyscore})\) difference to penalize some predominantly incorrect results and improve discrimination
Mascot Discriminant Function

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

\[ F = 0.1 \times \{4.91 + 47\} – 3.0 = 2.19 \]
Discriminant Score Distributions

Training dataset \([\text{M}+2\text{H}]^{2+}\) spectra
Computing probabilities from discriminant score distributions

No. of spectra

discriminant score (F)

Incorrect (-)

Correct (+)

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

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

$p = 0.5$
Computing probabilities from discriminant score distributions

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

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

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

$F_{obs}$: 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:

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

$F_{obs}$: 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 pl among correct and incorrect search results,…

…then the probability of any search result with $F_{\text{obs}}$, $NTT_{\text{obs}}$, $NMC_{\text{obs}}$, $\text{Massdiff}_{\text{obs}}$, $\text{ICAT}_{\text{obs}}$, $\text{N-glyc}_{\text{obs}}$, and $\text{pl}_{\text{obs}}$ 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 pl 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

vs. training dataset [M+2H]^{2+} spectra
Variations in Discriminant Score Distributions

Different distribution means

no of spectra

incorrect (-)
correct (+)

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

<table>
<thead>
<tr>
<th>Search Result</th>
<th>F</th>
<th>NTT</th>
<th>prob</th>
<th>1-prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3.0</td>
<td>2</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>B</td>
<td>2.0</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>C</td>
<td>1.0</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>D</td>
<td>0.0</td>
<td>0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

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

\[
P(NTT=0|+) = \frac{0.0}{1.0 + 0.5 + 0.5 + 0.0} = 0.0
\]

\[
P(NTT=1|+) = \frac{0.5 + 0.5}{1.0 + 0.5 + 0.5 + 0.0} = 0.5
\]

\[
P(NTT=2|+) = \frac{1.0}{1.0 + 0.5 + 0.5 + 0.0} = 0.5
\]

\[
P(NTT=0|-) = \frac{1.0}{0.0 + 0.5 + 0.5 + 1.0} = 0.5
\]

\[
P(NTT=1|-) = \frac{0.5 + 0.5}{0.0 + 0.5 + 0.5 + 1.0} = 0.5
\]

\[
P(NTT=2|-) = \frac{0.0}{0.0 + 0.5 + 0.5 + 1.0} = 0.0
\]

4. Recompute result probabilities using updated distributions, and iterate
EM Algorithm learns test data score distributions

Incorrect Peptide Assignments

Correct Peptide Assignments

No. of spectra

Discriminant search score
Incorrect Peptide Assignments

Correct Peptide Assignments

No. of spectra

Discriminant search score

NTT = 0
NTT = 1
NTT = 2

EM Iteration 0
EM Iteration 1

Incorrect Peptide Assignments

Correct Peptide Assignments

No. of spectra

Discriminant search score
EM Iteration 2

Incorrect Peptide Assignments

Correct Peptide Assignments

No. of spectra

Discriminant search score

NTT = 0
NTT = 1
NTT = 2
EM Iteration 3

Incorrect Peptide Assignments

Correct Peptide Assignments

No. of spectra

Discriminant search score

NTT = 0
NTT = 1
NTT = 2
EM Iteration 7

Incorrect Peptide Assignments

Correct Peptide Assignments

No. of spectra

Discriminant search score

Em Iteration 7

Incorrect Peptide Assignments

Correct Peptide Assignments

No. of spectra

Discriminant search score

NTT = 0
NTT = 1
NTT = 2
Accuracy of the Model

100 spectra with computed $p \sim 0.9$

90% of them (90) should be correct

Observed probability is around 0.9

Model is accurate

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

*test data: A. Keller et al. OMICS 6, 207 (2002)*
**Discriminating Power of Computed Probabilities**

**Sensitivity:**
fraction of all correct results passing filter

**Error:**
fraction of all results passing filter that are incorrect

---

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

Example: \( p \geq 0.9 \)

**Sensitivity**: fraction of all correct results passing filter

**Error**: fraction of all results passing filter that are incorrect

Test data: A. Keller et al. OMICS 6, 207 (2002)
Discriminating Power
Example: $p \geq 0.5$

**Sensitivity:** fraction of all correct results passing filter

**Error:** fraction of all results passing filter that are incorrect

Sensitivity, error

- observed
- predicted

Test data: A. Keller et al. OMICS 6, 207 (2002)
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...)
- XIInteract 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

<table>
<thead>
<tr>
<th>INDEX</th>
<th>PROBABILITY</th>
<th>SPECTRUM</th>
<th>IONSCORE</th>
<th>IDENTITYSCORE</th>
<th>HOMOLOGYSCORE</th>
<th>IONS</th>
<th>PEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>247</td>
<td>1.0000</td>
<td>haloCAT2_32.1358 1358 2</td>
<td>51.43</td>
<td>51.90</td>
<td>28.47</td>
<td>8/26</td>
<td></td>
</tr>
<tr>
<td>118</td>
<td>1.0000</td>
<td>haloCAT2_31.1068 1068 1</td>
<td>32.87</td>
<td>52.16</td>
<td>28.75</td>
<td>8/22</td>
<td></td>
</tr>
<tr>
<td>114</td>
<td>1.0000</td>
<td>haloCAT2_31.1072 1072 2</td>
<td>55.86</td>
<td>52.06</td>
<td>41.24</td>
<td>10/22</td>
<td></td>
</tr>
<tr>
<td>253</td>
<td>1.0000</td>
<td>haloCAT2_32.1778 1778 2</td>
<td>70.26</td>
<td>51.87</td>
<td>33.12</td>
<td>8/26</td>
<td></td>
</tr>
<tr>
<td>112</td>
<td>1.0000</td>
<td>haloCAT2_31.1070 1070 1</td>
<td>41.14</td>
<td>52.06</td>
<td>34.57</td>
<td>6/22</td>
<td></td>
</tr>
<tr>
<td>259</td>
<td>1.0000</td>
<td>haloCAT2_32.1546 1546 2</td>
<td>48.85</td>
<td>51.68</td>
<td>31.17</td>
<td>8/28</td>
<td></td>
</tr>
<tr>
<td>282</td>
<td>1.0000</td>
<td>haloCAT2_32.1530 1530 2</td>
<td>62.93</td>
<td>51.55</td>
<td>28.19</td>
<td>8/28</td>
<td></td>
</tr>
<tr>
<td>289</td>
<td>1.0000</td>
<td>haloCAT2_32.1414 1414 2</td>
<td>66.01</td>
<td>51.43</td>
<td>27.49</td>
<td>8/32</td>
<td></td>
</tr>
<tr>
<td>98</td>
<td>1.0000</td>
<td>haloCAT2_30.1901 1901 2</td>
<td>38.43</td>
<td>50.09</td>
<td>26.19</td>
<td>17/52</td>
<td></td>
</tr>
<tr>
<td>97</td>
<td>1.0000</td>
<td>haloCAT2_30.1905 1905 2</td>
<td>44.19</td>
<td>50.16</td>
<td>22.12</td>
<td>14/52</td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>1.0000</td>
<td>haloCAT2_30.1211 1211 2</td>
<td>48.30</td>
<td>50.74</td>
<td>28.06</td>
<td>16/46</td>
<td></td>
</tr>
</tbody>
</table>
PeptideProphet Results: Model Summary

- Estimated total number of correct peptide assignments in dataset: 48.3
- Sensitivity (Red Line): fraction of all correct assignments (48.3) passing NPT filter
- Error (Green Line): fraction of peptide assignments passing NPT filter that are incorrect
- NPT - Minimum Probability Threshold to Accept

<table>
<thead>
<tr>
<th>m/z</th>
<th>m/z</th>
<th>est. # count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>1.000</td>
<td>13</td>
</tr>
<tr>
<td>0.025</td>
<td>0.16</td>
<td>48</td>
</tr>
<tr>
<td>0.050</td>
<td>0.06</td>
<td>48</td>
</tr>
</tbody>
</table>

- Distribution of spectra and number of spectra per criterion.
Reasonable Learned Discriminant Score Distributions

2+

- Discriminant Score (fval)
- # of Spectra
- 2+ distr
- 2+ pos
- 2+ neg

3+

- Discriminant Score (fval)
- # of Spectra
- 3+ distr
- 3+ pos
- 3+ neg
Suspicious Looking Learned Discriminant Score Distributions
PeptideProphet Results: Model Summary

Estimated total number of correct peptide assignments in dataset: 48.3

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

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

MPT = Minimum Probability Threshold to Accept

<table>
<thead>
<tr>
<th>error</th>
<th>MPT</th>
<th>est # corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>1.00</td>
<td>13</td>
</tr>
<tr>
<td>0.025</td>
<td>0.35</td>
<td>48</td>
</tr>
<tr>
<td>0.050</td>
<td>0.06</td>
<td>48</td>
</tr>
</tbody>
</table>
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

\[ MPT = \text{Minimum Probability Threshold to Accept} \]

<table>
<thead>
<tr>
<th>error</th>
<th>MPT</th>
<th>est # corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>1.00</td>
<td>13</td>
</tr>
<tr>
<td>0.025</td>
<td>0.35</td>
<td>48</td>
</tr>
<tr>
<td>0.050</td>
<td>0.06</td>
<td>48</td>
</tr>
</tbody>
</table>
PeptideProphet Results: Contributing Score and Peptide Properties

2+ search result discriminant score value

\[ \text{spectrum: } \text{haloICAT2}_30.1149.1149.2 \]

\[ \text{scores: } fval=0.7157 \quad \text{ntt}=2 \quad \text{nmc}=1 \quad \text{massd}=0.900 \quad \text{prob: } 1.0000 \]

\[ \text{ntt: } [\text{ntt=0 } 0.000, \text{ntt=1 } 0.994, \text{ntt=2 } 0.006] \]

\[ \text{nmc: } [\text{nmc=0 } 0.992, \text{nmc=1 } 0.042, \text{nmc=2 } 0.006] \]

\[ \text{massd: } [\text{massd=-1.0000000 } 0.00000000002, \text{massd=0.0000000 } 0.005444515, \text{massd=0.0000000 } 0.214975881, \text{massd=0.0000000 } 0.005444515, \text{massd=-1.0000000 } 0.214975881, \text{massd=0.0000000 } 0.214975881] \]

\[ \text{icat=1 (compatible) } 0.971 \]

\[ \text{icat=0 (incompatible) } 0.029, \text{ compatible } 0.971 \]

\[ \text{icat=0 (incompatible) } 0.954, \text{ compatible } 0.046 \]
**PeptideProphet \([M+2H]^{2+}\) vs \([M+3H]^{3+}\)**

**Precursor Ions**

<table>
<thead>
<tr>
<th>334</th>
<th>1.0000</th>
<th>haloCAT2_33.1062.1063</th>
<th>51.06</th>
<th>51.94</th>
<th>36.29</th>
<th>10/44</th>
</tr>
</thead>
<tbody>
<tr>
<td>338</td>
<td>0.9984</td>
<td>haloCAT2_33.1042.1042</td>
<td>33.62</td>
<td>51.98</td>
<td>33.58</td>
<td>6/22</td>
</tr>
<tr>
<td>357</td>
<td>0.9996</td>
<td>haloCAT2_33.1034.1034</td>
<td>27.70</td>
<td>51.85</td>
<td>26.77</td>
<td>7/18</td>
</tr>
<tr>
<td>331</td>
<td>0.9596</td>
<td>haloCAT2_33.1024.1024</td>
<td>24.96</td>
<td>52.07</td>
<td>32.23</td>
<td>12/44</td>
</tr>
<tr>
<td>386</td>
<td>0.4275</td>
<td>haloCAT2_33.1014.1014</td>
<td>17.19</td>
<td>49.67</td>
<td>27.34</td>
<td>6/22</td>
</tr>
<tr>
<td>372</td>
<td>0.5725</td>
<td>haloCAT2_33.1014.1014</td>
<td>41.36</td>
<td>51.37</td>
<td>27.11</td>
<td>15/36</td>
</tr>
<tr>
<td>373</td>
<td>0.9992</td>
<td>haloCAT2_33.1014.1012</td>
<td>34.49</td>
<td>51.39</td>
<td>35.29</td>
<td>11/36</td>
</tr>
<tr>
<td>312</td>
<td>0.6340</td>
<td>haloCAT2_33.1004.1004</td>
<td>27.62</td>
<td>52.49</td>
<td>33.39</td>
<td>7/22</td>
</tr>
<tr>
<td>330</td>
<td>0.9992</td>
<td>haloCAT2_33.1002.1002</td>
<td>26.84</td>
<td>52.06</td>
<td>29.37</td>
<td>8/22</td>
</tr>
</tbody>
</table>

Spectrum searched as both 2+ and 3+ precursor received significant probability
PeptideProphet Results: Incomplete Analysis

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.
Sort Data by Computed Probability
Filter Data by Mascot Ionscore
Select and Color Specified AA’s
Pep3D and Analysis Summary Links

Details of Peptide Analysis

170 unique proteins, 106 single hits

Raw Data
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.

<table>
<thead>
<tr>
<th>#</th>
<th>MH+</th>
<th>IonSc</th>
<th>Ions Ref</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2956.4 (-0.4)</td>
<td>19.43</td>
<td>10/108 SWN:RN9_HUMAN</td>
<td>STNTSSDNLTSISKQIGDFIECPLCLL.R</td>
</tr>
<tr>
<td>2</td>
<td>2956.4 (-0.4)</td>
<td>17.93</td>
<td>10/ 92 SWN:RN9_HUMAN</td>
<td>STNTSSDNLTSISKQIGDFIECPLCLL.R</td>
</tr>
<tr>
<td>3</td>
<td>2954.6 (+1.4)</td>
<td>17.79</td>
<td>14/100 GP:AYO38359.2</td>
<td>T.AQSELALLAVHVNPTQKVLFSKLHK.Q</td>
</tr>
<tr>
<td>4</td>
<td>2955.5 (+0.4)</td>
<td>17.73</td>
<td>11/104 SW:ELS_HUMAN</td>
<td>+6 P.PLGGVAAARPFGPGLSPIFPFGACLGKAC...G</td>
</tr>
<tr>
<td>5</td>
<td>2954.6 (+1.4)</td>
<td>16.24</td>
<td>5/108 SWN:Y450_HUMAN</td>
<td>G.LFLRGPRPGSLDSHAAGRPPARPSVSQR.I</td>
</tr>
<tr>
<td>6</td>
<td>2955.6 (+0.4)</td>
<td>15.79</td>
<td>12/ 92 pNRC100.ORF5058</td>
<td>+3 T.DVPQWRLLVGGVGIVGTVRKG...T</td>
</tr>
<tr>
<td>7</td>
<td>2954.6 (+1.4)</td>
<td>15.44</td>
<td>22/124 GP:M24766_1</td>
<td>P.KGDPFPGAPGTVAPGIAQPKIAYQPGTV.G</td>
</tr>
<tr>
<td>8</td>
<td>2955.4 (+0.5)</td>
<td>15.19</td>
<td>8/112 SW:GGT5_HUMAN</td>
<td>+2 P.CGPQAFHAHAAADSKCSPDGRATLQQQ.G</td>
</tr>
<tr>
<td>9</td>
<td>2956.3 (-0.3)</td>
<td>14.84</td>
<td>19/112 SW:VUF_HUMAN</td>
<td>+2 E.CCGRCLPSACEVVTGSPRGSQSSDWSKVVG.S</td>
</tr>
<tr>
<td>10</td>
<td>2954.4 (+1.6)</td>
<td>14.78</td>
<td>22/124 GP:S79774_1</td>
<td>P.PTGDSGPPVVPTGDSGAPPVTPTGDSSETAPV.P</td>
</tr>
</tbody>
</table>
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

**PeptideProphet Options**
- [x] RUN PeptideProphet
- [ ] Use rest information
- [ ] Do not use rest information
- [ ] Use N-glyco motif information
- [ ] Use pl information
- [ ] Use accurate mass binning
- [ ] MALDI data
- [ ] Exclude all entries with asterisked score values
- [ ] Leave alone all entries with asterisked score values
- [ ] Run ProteinProphet afterwards
- [ ] Do not assemble protein groups in ProteinProphet analysis
- [ ] Do not use Occam’s Razor in ProteinProphet analysis to derive the simplest protein list to explain observed peptides
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**!