MS-MS Analysis Programs

Basic Process

- Genome - Gives AA sequences of proteins
- Use this to predict spectra
- Compare data to prediction
- Determine degree of correctness
- Make assignment – Did we see the protein?
  - Correctness
  - Number of sequences
  - Observed in different experiments
    - Repeats
    - Replicates
  - Quantity and Quality of sequences

Data

- Listing of ion masses
- Amplitude
- Precursor ion mass
- Knowledge of modifications
- Knowledge of ionization probabilities
  - y ions will be most obvious
  - b and a ions can be missing
  - some amino acids disrupt cleavage
  - Addition of or removal of water or OH
  - ETC

Parameters that affect searches

- Several parameters are important in searching data bases for peptide matches
  - Data quality
    - Signal to noise
    - Spurious peaks
  - Mass tolerance
    - How much can the mass vary –
    - Usually 0.3 -5 amu, <=1 is most common.
      - Both parent and fragment mass tolerances important
  - Modifications allowed
  - Missed cleavages
  - Data base size

Mass Tolerance / Error

ppm (parts per million) Formula: $\frac{\Delta}{\text{Theoretical value}} \times 10^6 = \text{ppm error}$

OR

% Formula: $\frac{\Delta}{\text{Theoretical value}} \times 100 = \%\ \text{error}$

where Δ = Theoretical *m/z* value - Experimental *m/z* value

Example

$\Delta = 1479.63\ m/z_{(theoretical)} - 1480.10\ m/z_{(experimental)}$
$= 0.47$

ppm error: $\frac{0.47}{1479.63_{(theor)}} \times 10^6 = 317$ ppm error

OR

% error: $= 0.0317\ \%$

Slide provided by the U of Minn Mass Spectrometry Center

Absolute Values of Mass Tolerance As a Function of Mass for Two Different % Tolerance Setting

Mass (Da)	% Tolerance	Error (+/-)
1200	0.03%	0.4
3500	0.03%	1
12000	0.03%	4
25000	0.1%	25
60000	0.1%	60
80000	0.1%	80

Reminder: 0.03% = 300 ppm

Slide provided by the U of Minn Mass Spectrometry Center

Peptide Identification Programs

- Categories
- Precursor approaches: most often used
  - Sequest®
  - Mascot®
  - X!Tandem
- Sequence approaches: newer
  - MS-Tag*
  - Protein Pilot™

Sequest®

- **From a database it will choose peptide candidates based on *PRECURSOR MASS.***
  - **This is the mass of the ion before fragmentation**
  - **Take know sequences and predict precursor mass**
  - **Chooses all the sequences that fit the mass**
  - **Create a model spectrum for each theoretical candidate.**
  - **Determine and sum peaks that overlap between theoretical and experimental data, S_p score.**

Sequest

- **S_p score is based on the number of ions that fit**
- **Value is dependent on the peptide size**
- **Spectra are shifted to get different S_p scores**

Number of matching fragments

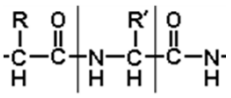
Bonus for consecutive fragmentation ions β= 0.075

Sum of intensities of matching fragments

Bonus for presence of immonium ions ρ=0.15

$$S_p = \frac{\left(\sum i_m\right) n_i \left(1+\beta\right)\left(1+\rho\right)}{n_t}$$

Total Number of predicted sequence ions



Immonium ion

Sequest

- **The Sp scores are tallied**
- **The best 500 Sp score sequences are used**
- **The Sp score is dependent on the size of the peptide.**
- **For 20 aa a score of 1000 is good**
- **For 6 aa a score of 500 is good**
- **The best 500 Sp score sequences are used to calculate the Xcorr value**

Sequest®

- **Shift the experimental spectrum back and forth and sum overlapping peaks**
  - **Generates the Auto-correlation information (background noise calculation)**
- **Calculate cross correlation score (XCorr)**

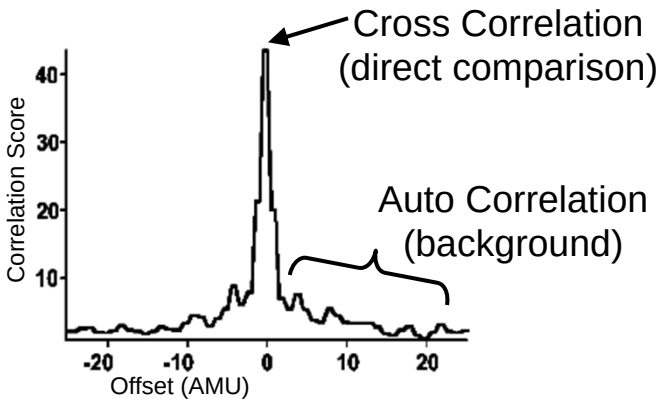
$$C_{xy} = \int_{-\infty}^{+\infty} x(t)y(t+\tau)dt$$

  - **C_{xy} is the correlation score**
  - **τ is the displacement between the spectra**
  - **x(t) and y(t) are the original and predicted spectra**

Sequest®

- **This is a "common procedure"**
- **Is used as measure of correctness**
- **The larger the value the better**
- **Value is dependent on the peptide size**
  - **Bigger the peptide the larger the value**
- **Value is dependent on the charge**

SEQUEST® Scoring: XCorr



$$\text{XCorr} = \frac{CrossCorr}{avg\left(AutoCorr_{\text{offset=-75 to 75}}\right)}$$

Gentzel M. et al
Proteomics **3** (2003) 1597-1610

http://www.proteomesoftware.com/Proteome_software_Proteomics.html
Adapted from: Interpreting_MSMS_results_cartoon.ppt (Brian Searle)

Sequest®

- **Calculate ΔCn – measure of how good the XCorr is relative to the next best match.**
- $$\Delta\text{Cn} = \frac{(Xcorr1 - Xcorr\ 2)}{Xcorr1}$$
  - **Measures how different the top value is from the next one. The larger the better.**
  - **Minimum value is 0.1**
  - **Value depends on the database size**
  - **Xcor values > 1.9, 2.2, or 3.7 for ions of 1, 2, or 3 charges are usually accurate**

Sequest®

- **Other measure of fit**
  - Ion score.
    - Number or percent of ions that match
    - Values of 70 to 80% good

X!Tandem

- Uses precursor ion to pick candidates
- Compares spectrum to predicted spectrum
- Uses only peaks that show up in both
- Calculate y/b value for preliminary assignment

$$y/b = \left[\sum_{i=0}^{n} I_i * P_i \right]$$

  - I is ion intensity
  - P is whether theoretical peak matches 1 or 0

Fragment ion mass tolerance affects identification

- P is a vector of n intensities that represents the masses calculate for all potential sequence specific ions (a, b, y and -17 (OH) and -18 (H_2O) neutral loss products).
- This is the theoretical spectrum
- n is the parent ion mass divided by the mass accuracy.
- P_i has values of 1 for present and 0 for absent
- Vector I_i represents the actual spectrum that correspond to the intensity, normalized to 100 or 0 for missing fragment

X!Tandem

- Calculates hyperscore
  - N_y! is the factorial of the y ions identified
  - N_b! Is the factorial of the b ions identified

$$hyperscore = \left[\sum_{i=0}^{n} I_i * P_i \right] * N_b! * N_y!$$

- Assume that highest hyperscore is the correct assignment
- Makes histogram of hyperscore from candidate peptides
  - Used for fitness calculation

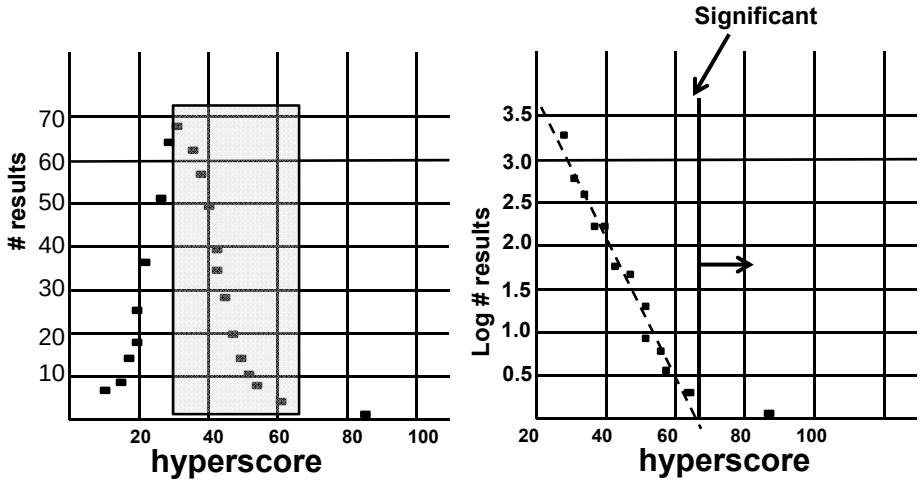
X!Tandem

- Takes log of histogram values greater than peak value and plots log of number versus hyperscore.
- A straight line indicates that these assignments are random
- Takes linear regression of log plot of all random data and extends out to value of chosen assignment. This is the E value.

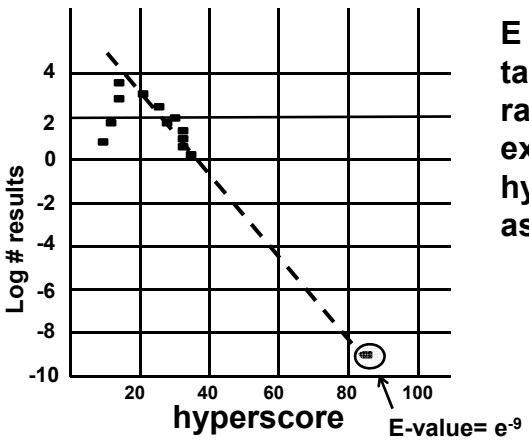
http://www.proteomesoftware.com/pdf_files/XTandem_edited.pdf

http://h.thegpm.org/tandem/thegpm_tandem.html

Hyperscores



E-Value



E value is determined by taking the slope due to random assignments and extrapolating to the hyperscore of the best assignment.

Which Program is better

	How closely does the spectrum match the model	What is the distance from the top match from the other possibilities
X!Tandem	Hyperscore	E-value <0.01 Better measure
Sequest	XCorr >2.5 Better measure	ΔCn >0.1

Neither program has all the best features

Viewers and additional analysis

- Scaffold
  - Protein Prophet
- Mass Sieve
- Progroup
  - Protein pilot
- etc

Scaffold

- Scaffold uses both X!Tandem and Sequest
- Takes identified spectra and adds additional analysis to assign probability of correct protein identification

Protein Prophet Discriminant Function

- Discriminant score for Sequest

$$D = \left(8.4 * \frac{\ln(XCorr)}{\ln(\#AAs)}\right) + (7.4 * \Delta Cn) - (0.2 * \ln(rankSp)) - (0.3 * \Delta Mass) - 0.96$$

  - First segment corrects for peptide length
  - Accounts for distance from next closest assignment. Farther the better.
  - includes top ranked sequence
  - Penalizes for poor mass accuracy

Protein Prophet

- Makes histogram of discriminant scores
- Curve fits histogram for correct and incorrect assignments
- Uses Bayesian statistics to compute probability of correct match
- This gives better sensitivity and lower errors to assignments

- http://www.proteomesoftware.com/Proteome_software_prod_Scaffold_tour.html
- http://appliedbiosystems.cnpg.com/lsca/webinar/proteinpilot/20060516/