X!Tandem Explained

Brian C. Searle
Proteome Software Inc.
www.proteomesoftware.com
1336 SW Bertha Blvd, Portland OR 97219
(503) 244-6027

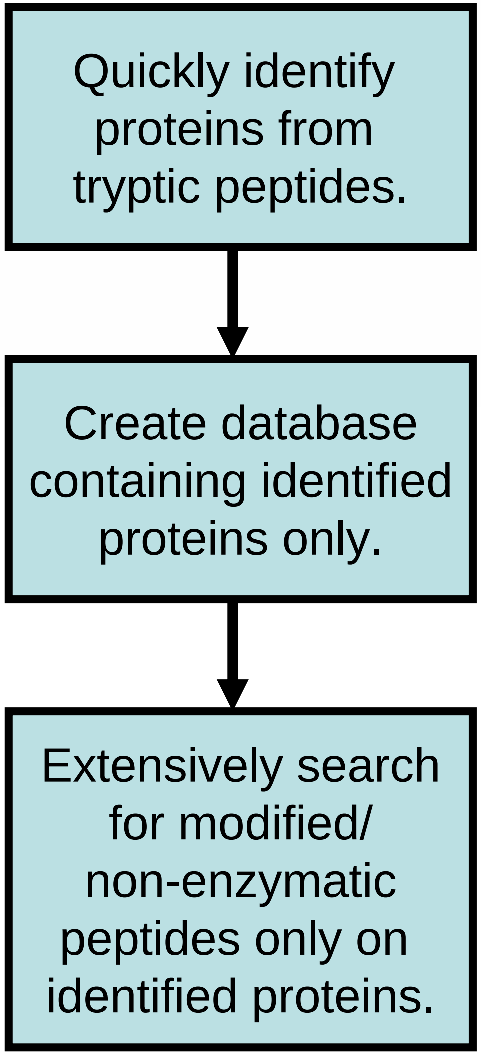
An explanation of the X!Tandem MS/MS spectra search program developed by

Craig,R. and Beavis,R.C. (2003) *Rapid Commun. Mass Spectrom.*, **17**, 2310–2316.

X!Tandem Theme

Central Axiom: For each identifiable protein, there is at least one detectable tryptic peptide.

X!Tandem Workflow



Quickly identify proteins from tryptic peptides.

Create database containing identified proteins only.

Extensively search for modified/ non-enzymatic peptides only on identified proteins.

X!Tandem, like Mascot and SEQUEST, compares each spectrum to all likely candidate peptides in a protein database.

One of X!Tandem’s strengths is its automatic search for modified peptides — but only on proteins it has otherwise identified.

The following pages explain how X!Tandem matches peptides, and how this differs from the way SEQUEST matches them.

X!Tandem’s Chief Advantage

Speed:

- ~200x faster for nonspecific searches
- ~1000x faster if you also look for oxidation, deamidation, and phosphorylation

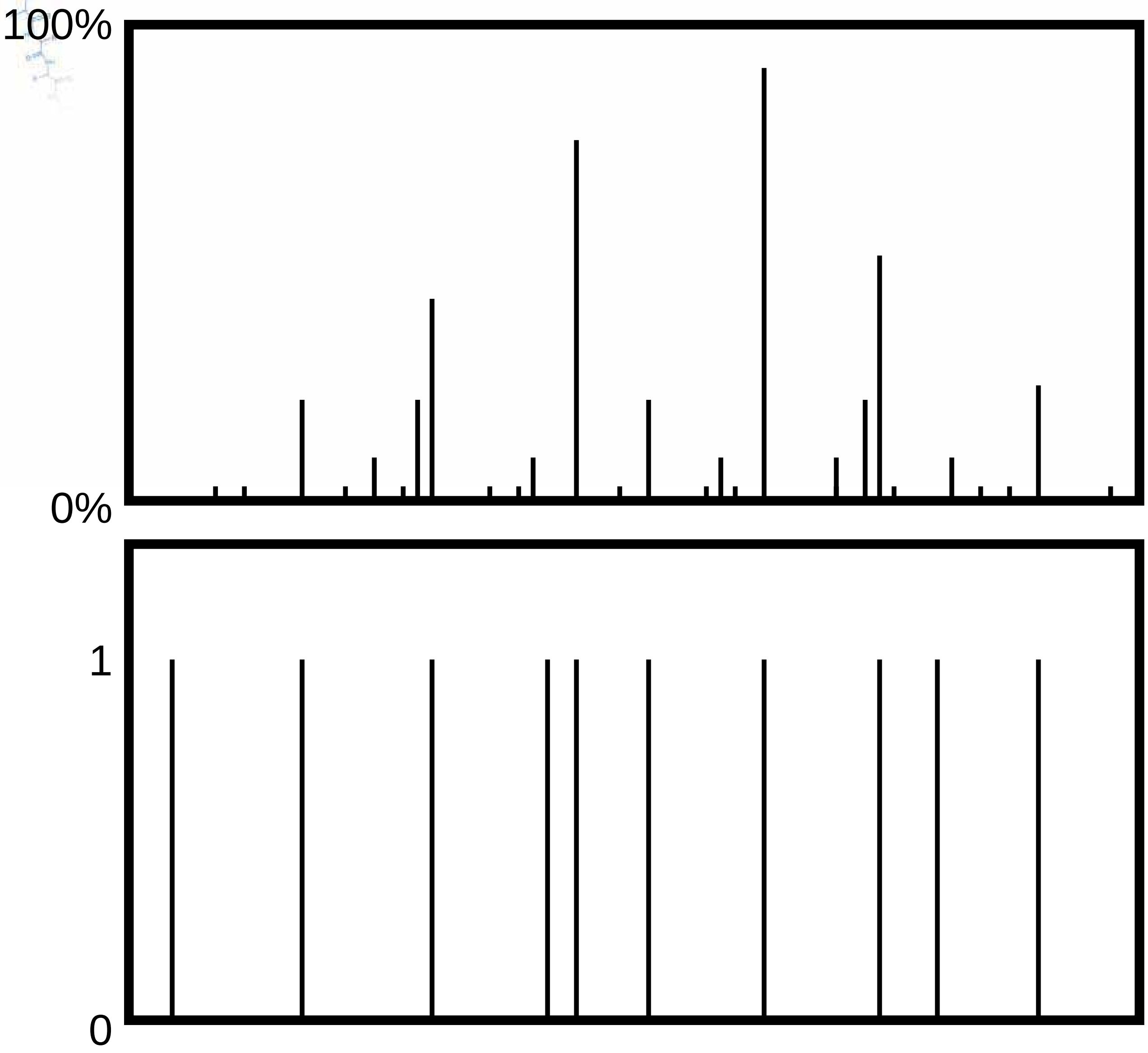
Other X!Tandem Advantages

Modern search techniques:

- Considers semi-tryptic peptides.
- Considers sequence polymorphisms.
- Uses probability-based scoring.

Spectra

X!Tandem works by matching the acquired MS/MS spectra to a model spectrum based on peptides in a protein database. The model spectrum is very simple, based on the presence or absence of y and b ions.

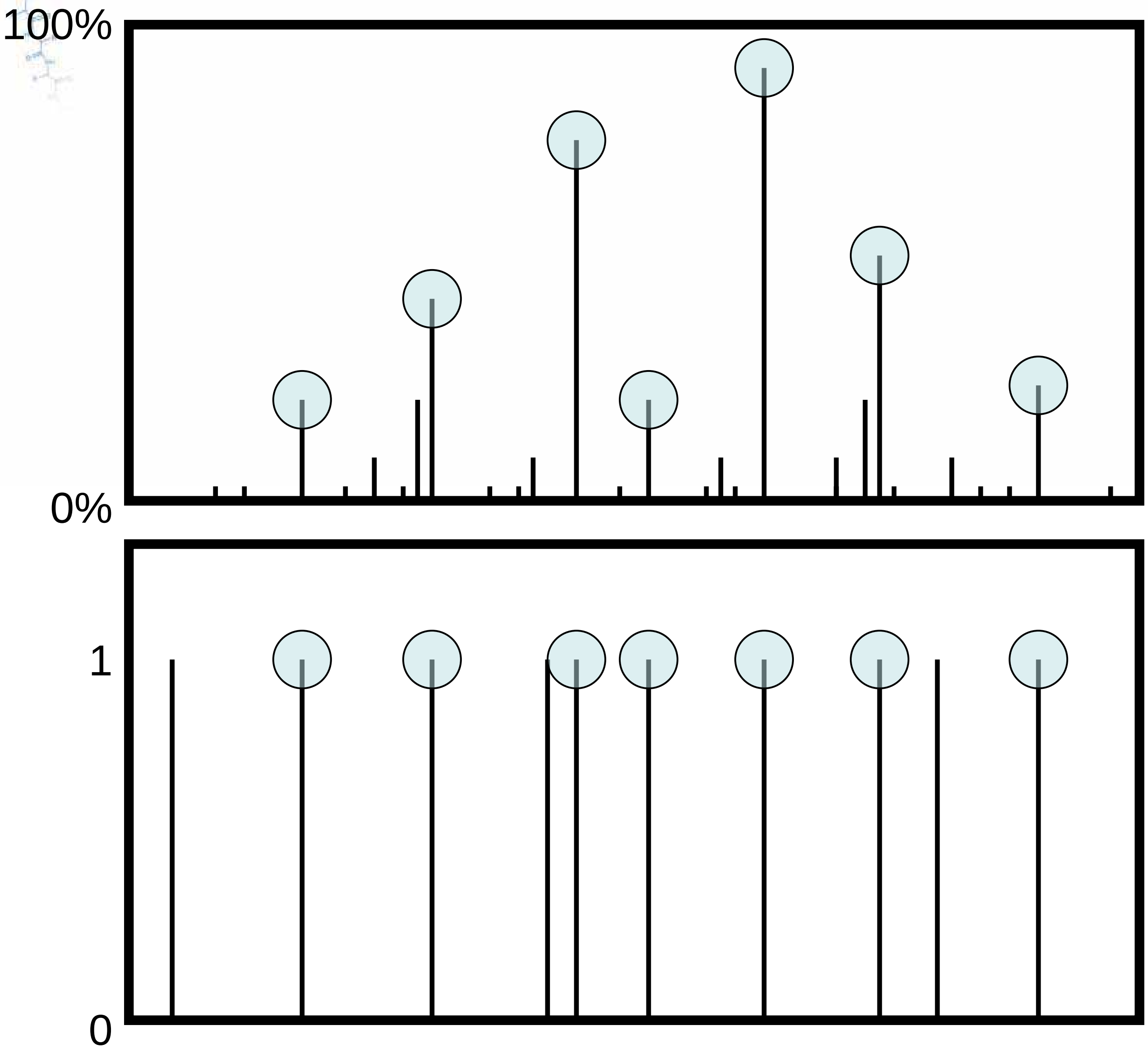


acquired spectrum

model spectrum (y/b ions)

Spectra matched

Only matching spectral peaks — the ones marked in the figure — are considered. Any peaks that don’t match, in either the model or acquired spectra, are not used.



acquired spectrum

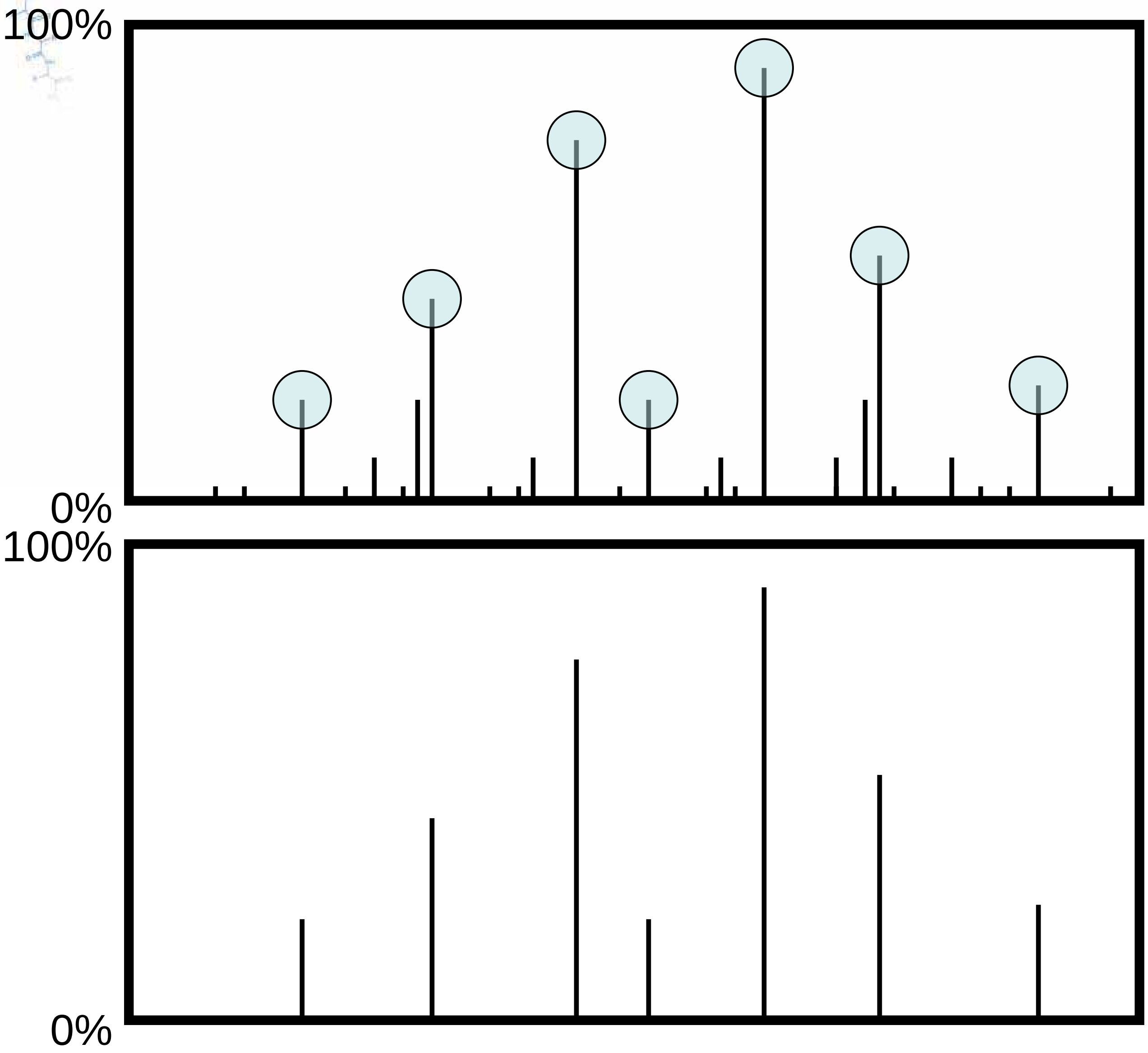
X

hypothetical spectrum (y/b ions)

Similar peaks

The acquired spectrum is simplified to only those peaks that are similar to the peaks in the model spectrum.

The acquired spectrum is simplified to only those peaks that are similar to the peaks in the model spectrum.



acquired spectrum

similar peaks (y/b ions)

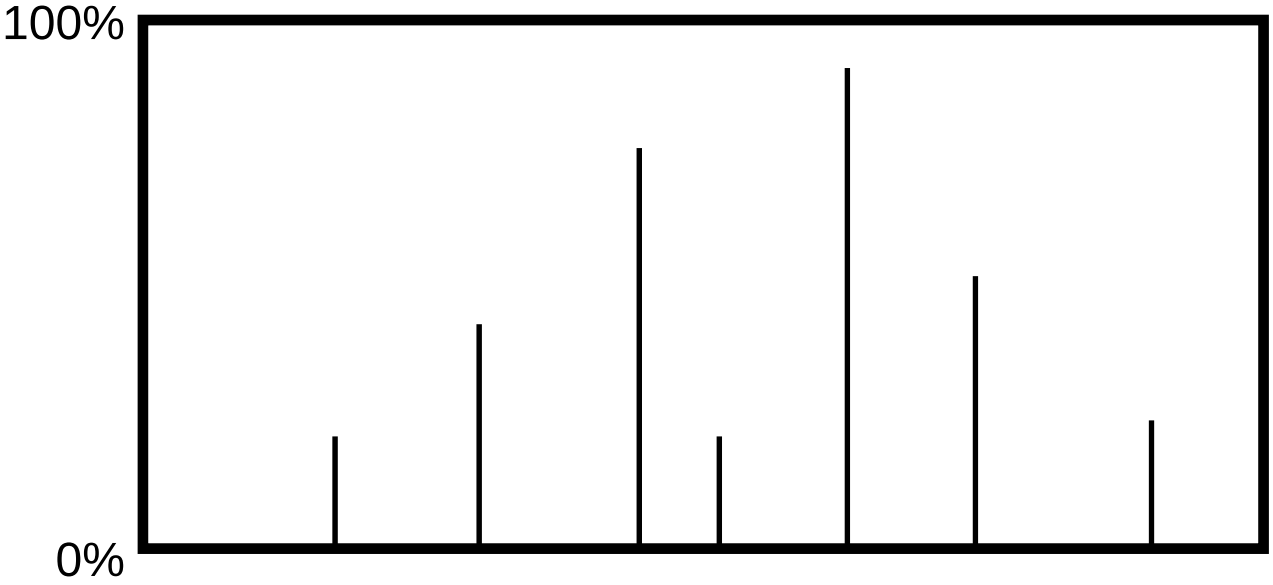
Dot product

X!Tandem's preliminary score is a dot product of the acquired and model spectra. Because only similar peaks are considered, this is the sum of the intensities of the matched y and b ions.

$$y/bScore = \left(\sum_{i=0}^{n} \boxed{I_i} * \boxed{P_i} \right)$$

spectrum intensities

predicted? (1,0)



similar peaks (y/b ions)

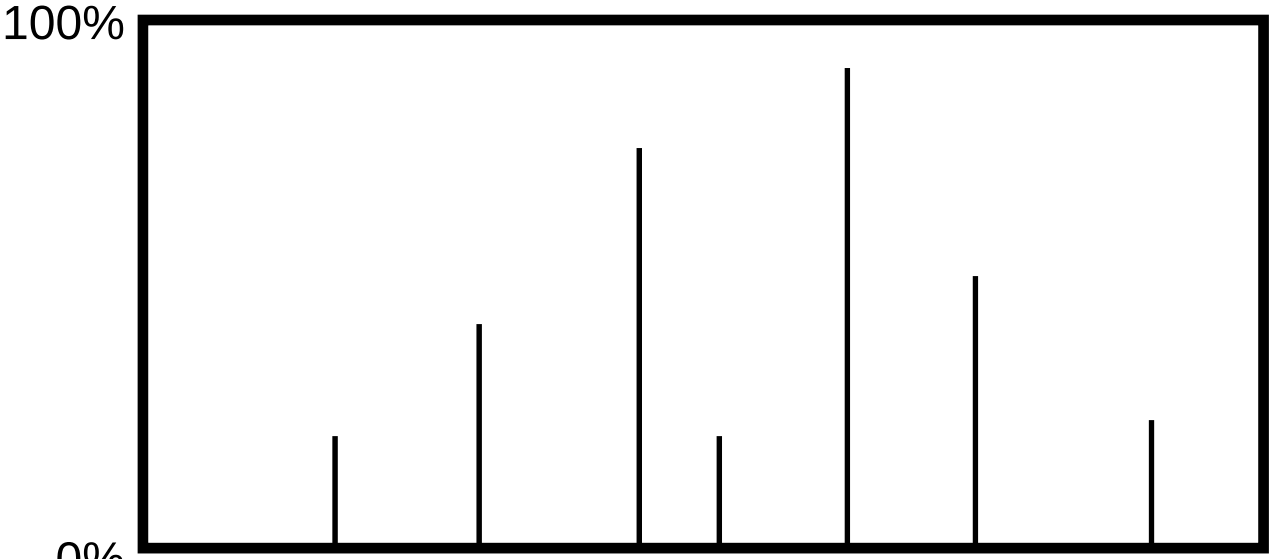
Hyperscore

$$HyperScore = \left(\sum_{i=0}^{n} \boxed{I_i} * \boxed{P_i} \right) * N_b! * N_y!$$

spectrum intensities

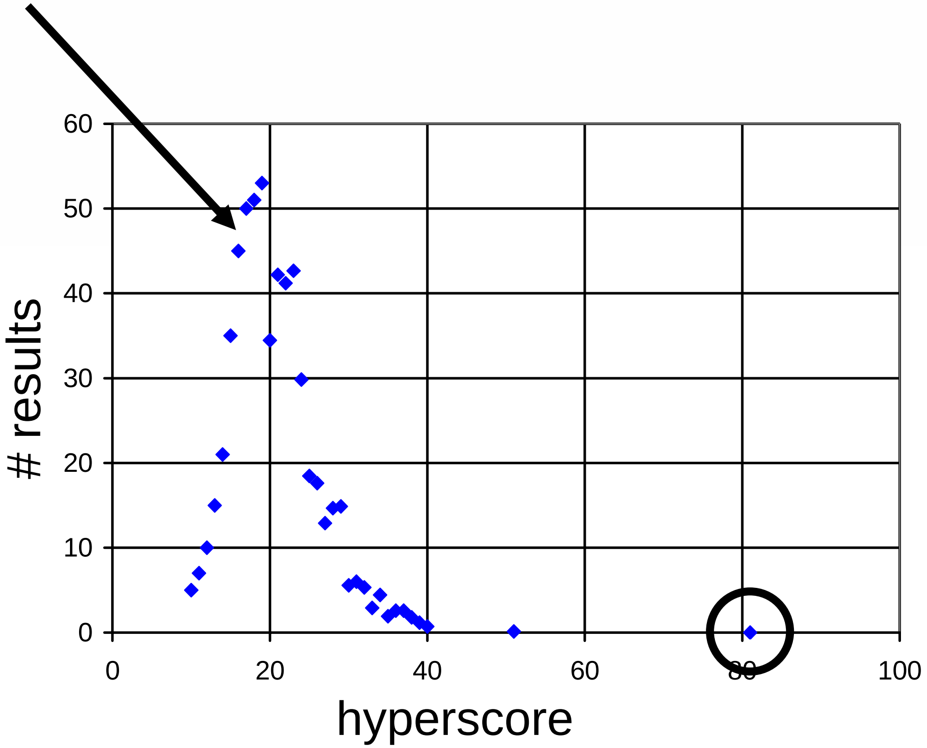
predicted? (1,0)

X!Tandem modifies the preliminary score by multiplying by N factorial for the number of b and y ions assigned. The use of factorials is based on the hypergeometric distribution.



similar peaks (y/b ions)

Histogram of hyperscores



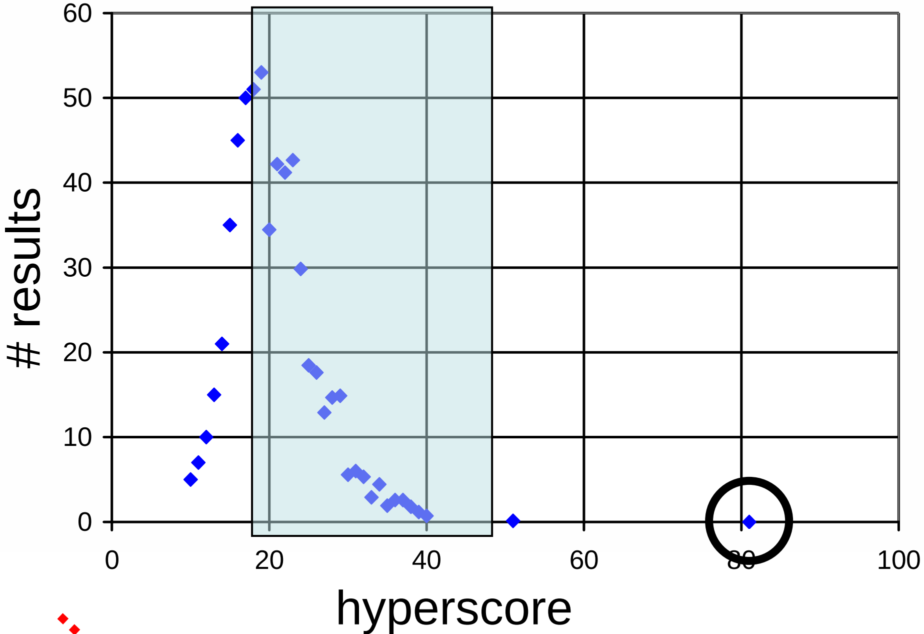
incorrect IDs

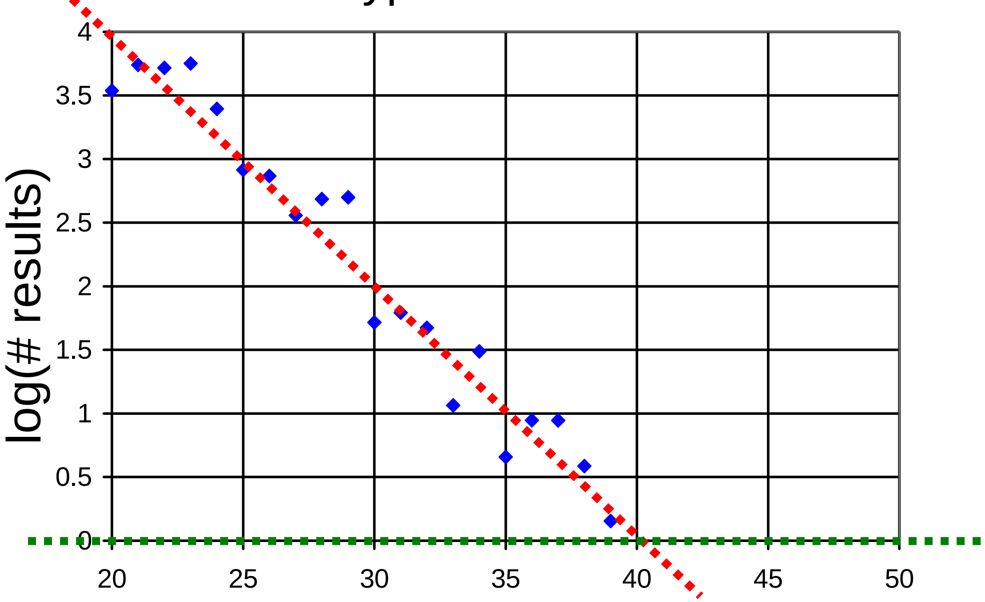
Next, X!Tandem makes a histogram of all the hyperscores for all the peptides in the database that might match this spectrum.

For example, in this figure, 52 peptides were found with a hyperscore of 19, and one peptide with a hyperscore of 83.

X!Tandem assumes that the peptide with the highest hyperscore is correct, and all others are incorrect.

Log histogram



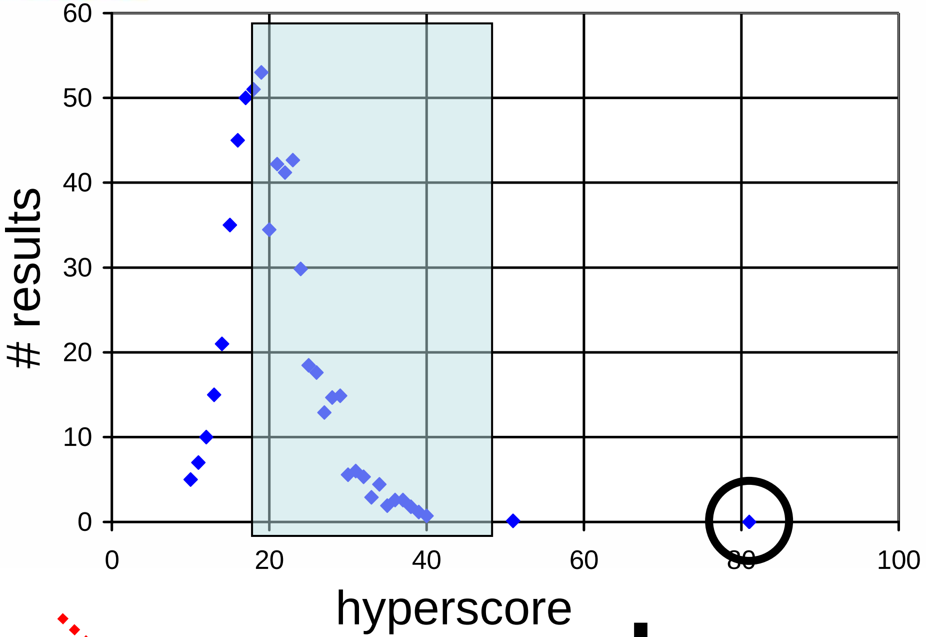


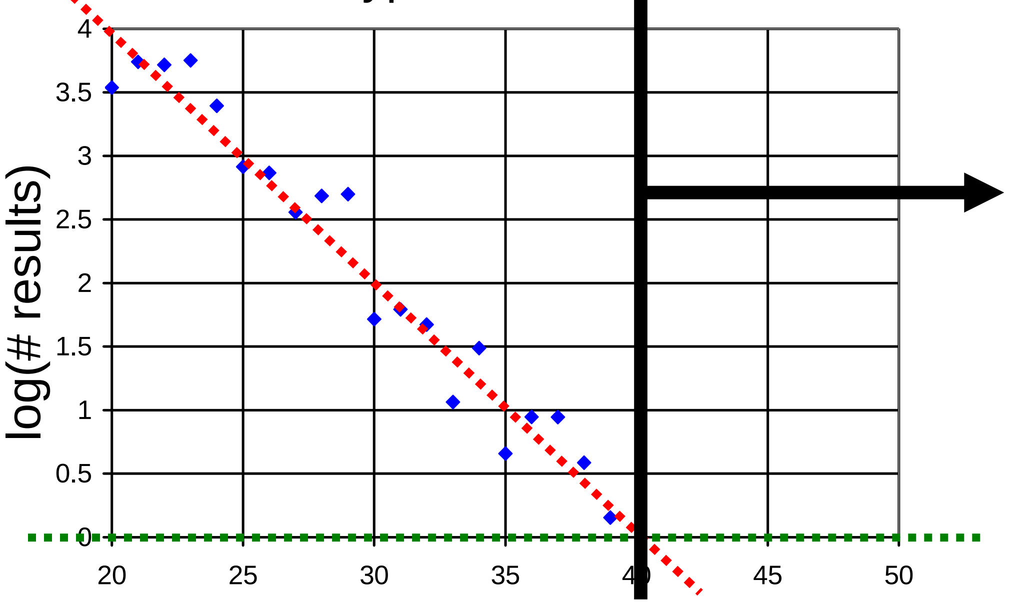
If the data on the right side of the histogram, (colored in upper figure) is taken and log-transformed, the data fall on a straight line.

A straight line is the expected result from a statistical argument that assumes the incorrect results are random.

Note: this histogram is calculated independently for each spectrum.

Significant scores





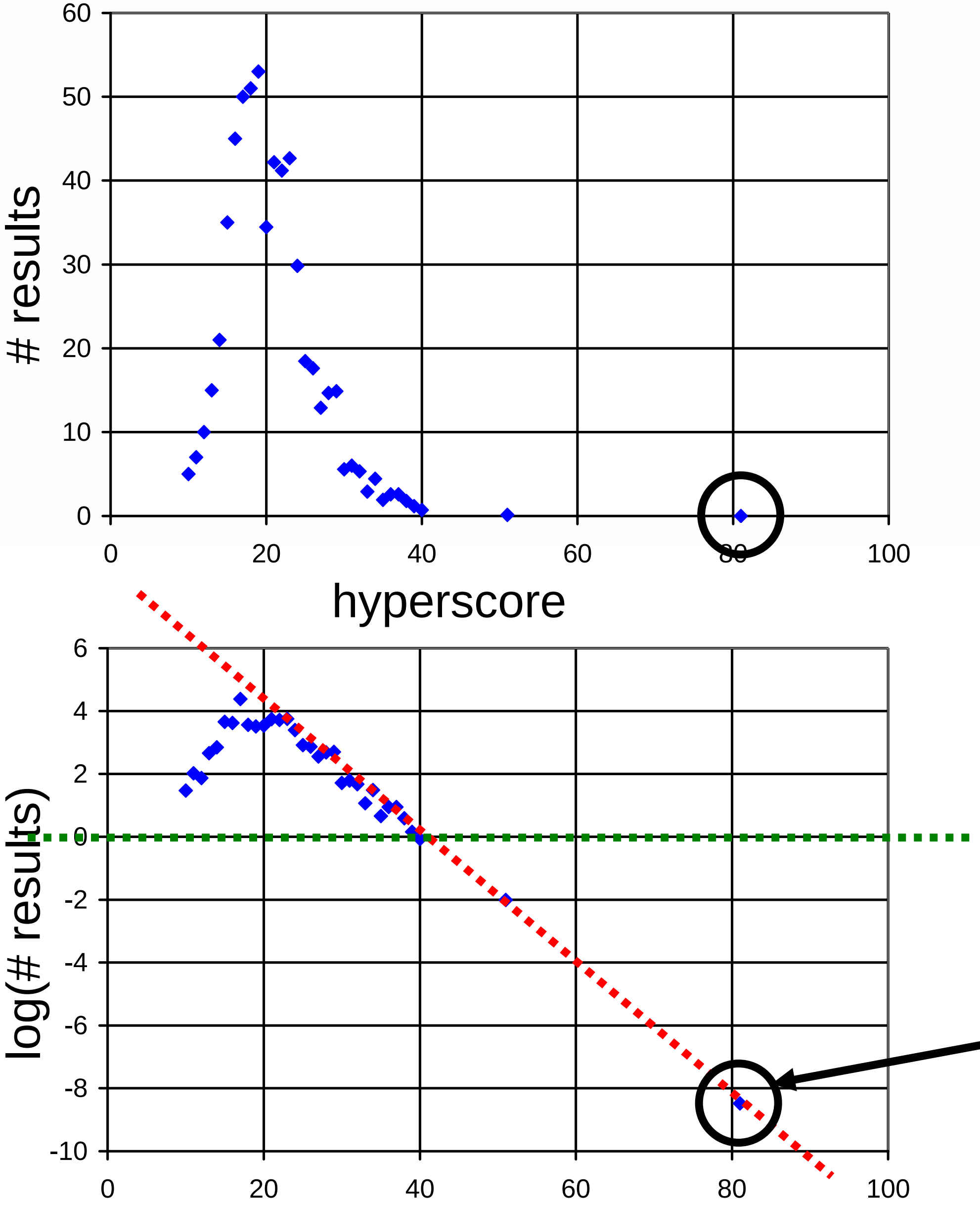
significant

X!Tandem has already assumed that the top hyperscore is the only possible correct match.

This match is significant if it is greater than the point at which the straight line through the log data intersects the log(#results)=0 line.

Any hyperscores greater than this are unlikely to have arisen by chance.

E-value



The E-value expresses just how unlikely a greater hyperscore is.

X!Tandem calculates the E-value by extrapolating the red line of the log histogram.

For the example shown, a hyperscore of 83 would occur by chance where the red line crosses 83. The log of this value — the E-value — is -8.2, as shown.

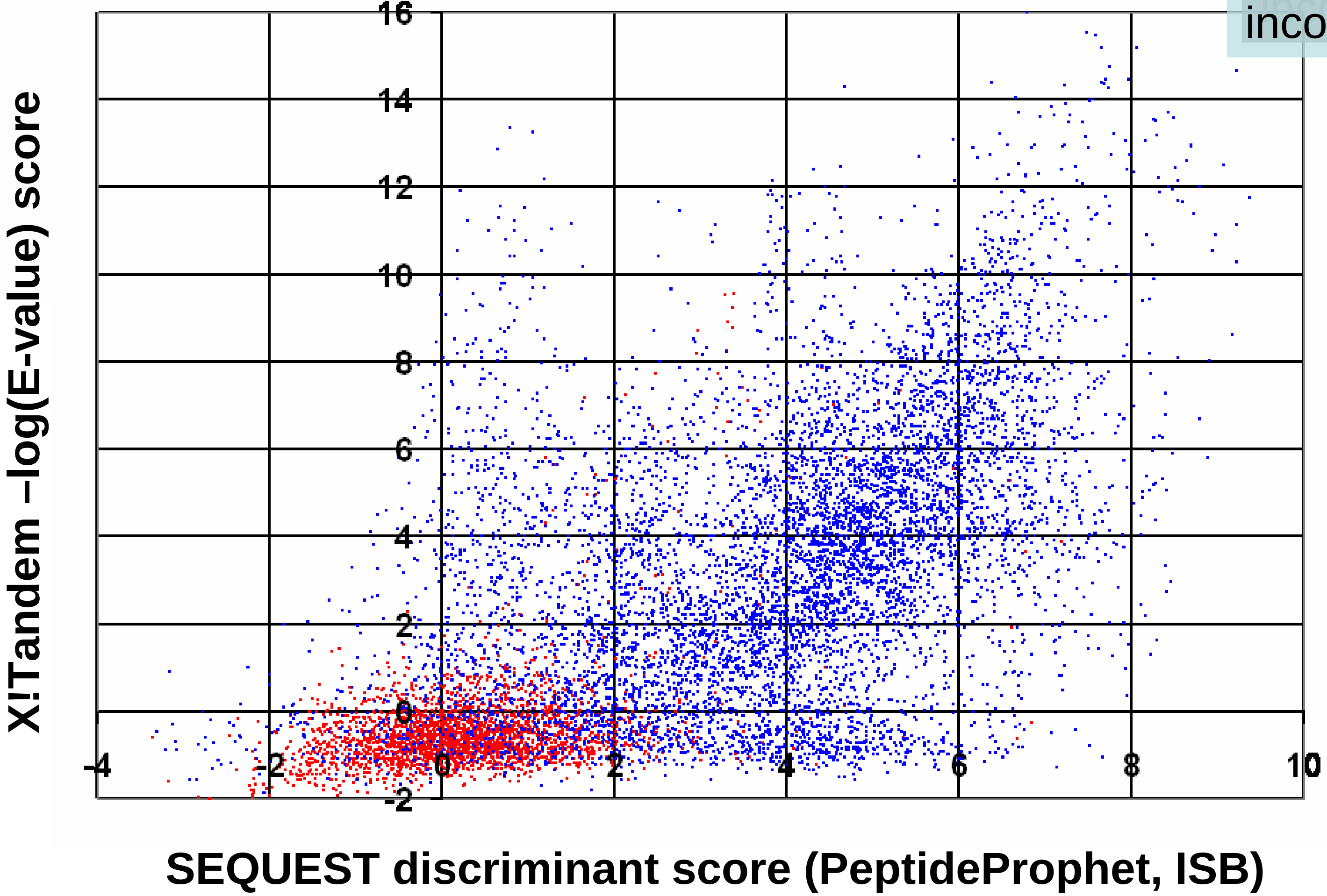
E-value=$e^{-8.2}$

X!Tandem vs. SEQUEST

Which is better?	How close is the spectrum to the peptide model?	How far is the top-scoring match from the rest of the pack?
X!Tandem	hyperscore	E-value **better measure**
SEQUEST	XCorr **better measure**	ΔCn

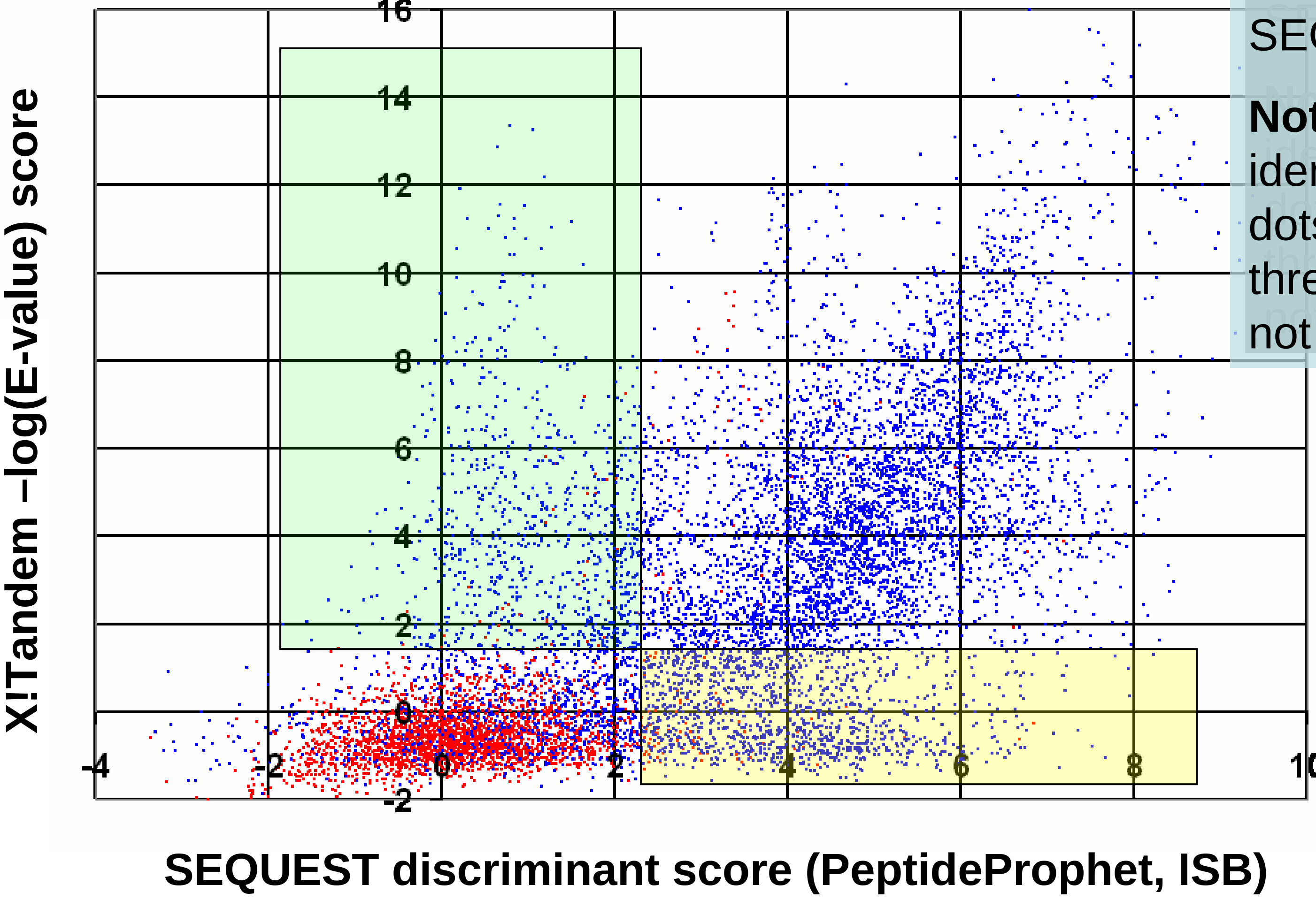
Plotting SEQUEST and X!Tandem scores

The sample in this experiment has only 10 proteins. Spectra identifying one of these 10 are plotted in blue. Spectra in red are incorrect matches.



X!Tandem and SEQUEST scores for all the spectra in a control experiment

Cutoffs for SEQUEST and X!Tandem



The yellow box contains the spectra with scores below X!Tandem’s threshold. The green box shows the spectra with scores below SEQUEST’s threshold.

Note: Many good identifications (blue dots) are below threshold for one, but not both.

X!Tandem Summary

- X!Tandem is fast.

  The hyperscore is much faster to calculate than SEQUEST’s XCorr. This allows X!Tandem to build the histograms that allow it to calculate the E-values.

- X!Tandem grades on a curve.

  It’s the E-value that counts, not the hyperscore.

- X!Tandem and SEQUEST can often match different spectra.