1/16

Protein Prospector and Ways of Calculating Expectation Values

Aenoch J. Lynn; Robert J. Chalkley; Peter R. Baker; Mark R. Segal; and Alma L. Burlingame

University of California, San Francisco, San Francisco, CA 94143-0446.

2/16

Introduction

With the production of large, multidimensional chromatography tandem mass spectrometry datasets it has now become essential to characterize the reliability of results. In addition, publication guidelines for mass spectrometry experiments will require a measure of the statistical significance of a peptide assignment[1]. The most commonly reported statistical measure of significance is an expectation, or e-value, which represents how many random matches would be expected to achieve a given score or greater, in a search of a given size. Conventional e-value choices are 0.05 or 0.01, with 0.05 commonly used by other database search engines (Mascot[2], OMSSA[3] and X!Tandem[4]). We discuss the various methods of calculating an e-value and their relative merits.

3/16

Expectation Values

- The probability value (p-value) is the probability an event will occur at random.
- The expectation value (e-value) is the expected number of times an event will occur at random from a given set of trials.
  (e-value = p-value * number of trials)
- For mass spectrometry, an e-value is the number of times a given peptide score (or greater) will be achieved by incorrect matches from a database search.
- If a peptide assignment has an e-value of 0.01, then one would expect 1 peptide to match at random from a database 100 times in size.
- e-values can be calculated by:
  A. A theoretical calculation of the chances of a given number of peak masses out of a total number of peaks matching at random[2,3].
  B. Fitting the incorrect (null) results from a database search to a distribution and using this distribution to calculate a p-value[4,5].

4/16

How to Calculate e-values

A. Theoretical Chances of Matching Peaks

- What is the probability of 15 out of 25 masses matching to a random (incorrect) assignment?
- Potentially fast.
- Fails to account for database factors (amino acid frequencies).
- Reliable e-values requires understanding and accounting for all factors that contribute to random matching of peaks; not all factors are understood.

B. Modeling the Incorrect Distribution

- Model the incorrect (random) distribution to determine p-values (and thereby e-values) for a corresponding score.
- Applicable to any scoring scheme without requiring an understanding of the factors contributing to peptide fragmentation and measurement error.
- Limited by the distribution family chosen for modeling.

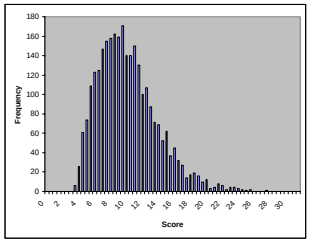
5/16

Linear Tail-Fit

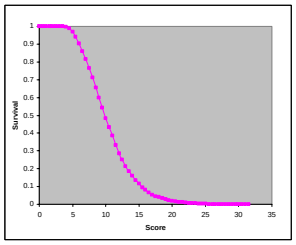
- Use top fraction of the scores (10%) requiring large numbers of incorrect matches to model the distribution[5].
- Plot log(survival) vs log(score) and use linear regression to estimate p-values for a given peptide score.
- Reasonably accurate for e-values between 0.01 and 0.00001.
- log(survival) vs log(score) not always linear; for Prospector, log(survival) vs score is more linear.
- Sensitive to matching homologous peptides and skewing the upper end of the tail.
- Relies upon extrapolation.
- High mass-accuracy spectra or species restricted database searches may not return sufficient numbers of incorrect matches for this method to work.

6/16

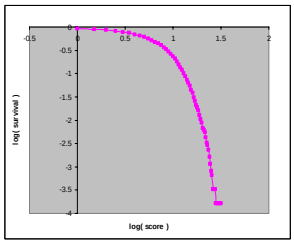
Calculating Linear Tail-Fit



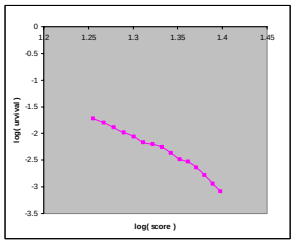
Score vs Frequency



Score vs Cumulative Frequency



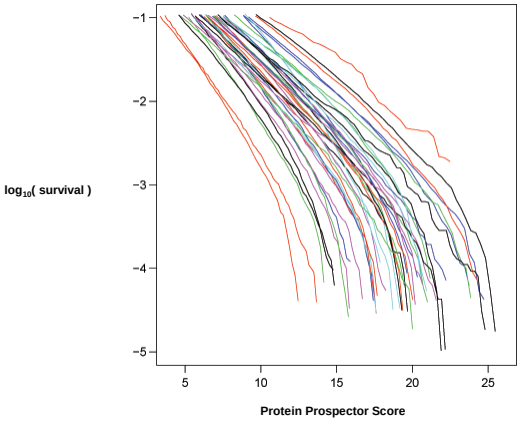
Log Survival vs Log Score



Top 10% Scores

7/16

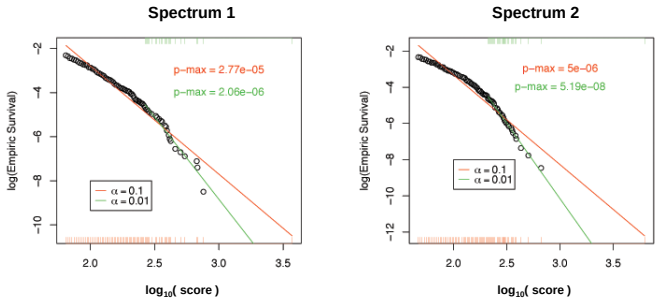
Survival Curves



Survival curves for 44 spectra searched against SwissProt. Using the top 10% of the scores, as used in the linear tail-fit.

8/16

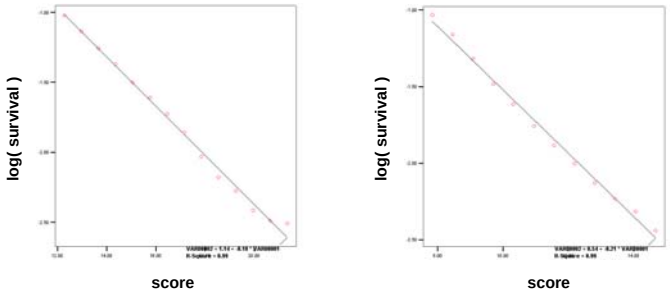
Non-Linear Survival Tails



Plots of log(survival) vs. log(score) for the top 10 (red regression line) and top 1% (green regression line) of scoring peptides for two spectra. For log(score), the tail-fits do not appear very linear, and are sensitive to the percentile selected for the cutoff.

9/16

Linear Survival Tails



Plots of log(survival) vs. score for the top 10% of scoring peptides for two spectra. In this region the survival tails are linear and are less sensitive to percentile selected for the cut-off.

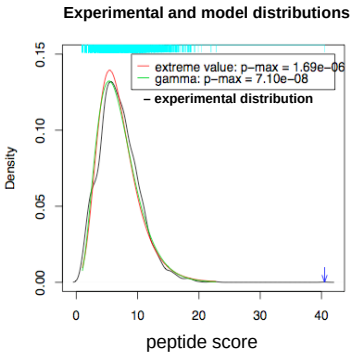
10/16

Model Distributions

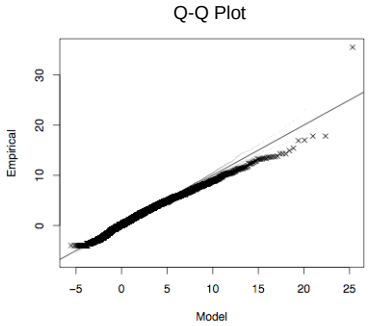
- Fit the null (incorrect) peptide scores to statistical distributions.
  - Extreme value distribution
    - Method of Moments
    - Closed Form Maximum Likelihood
  - Poisson distribution
  - Gamma
- Less sensitive to fewer data points than Tail-Fit method (able to model high mass-accuracy MS data).
- Assumes that the distribution of scores (except for the correct match) is random.
- Use quantile-quantile plots (Q-Q plots) to determine the appropriateness of a model distribution.
  - A quantile is the fraction of points below a given value.
  - Plot the quantiles from the incorrect distribution against the quantiles of the model.
  - If the data are both from the same distribution, they will fall along the 45-degree line for the plot.

11/16

Modeling Using Extreme Value Distribution



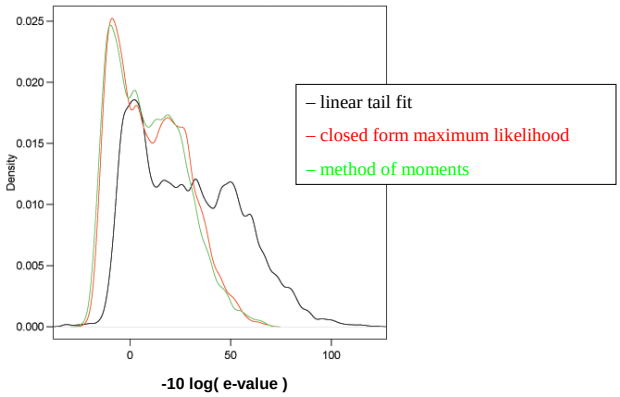
Distribution of peptide scores from a single spectrum in a Prospector search. The red trace is the extreme value model of the experimental data.



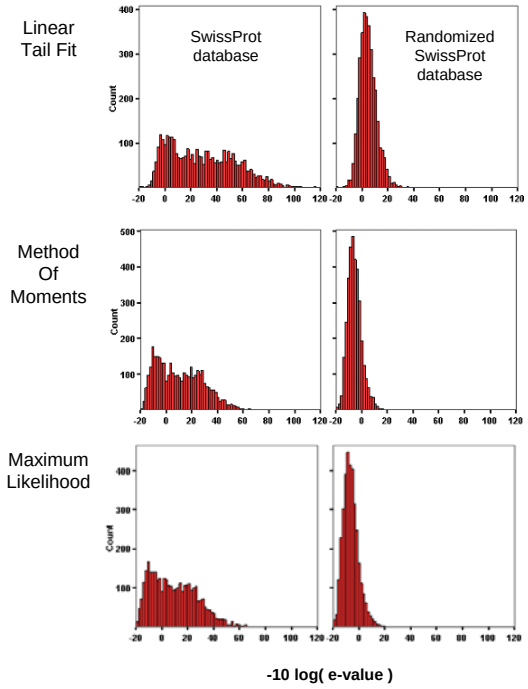
Plot of the peptide score quantiles against the extreme value distribution quantiles shows the appropriateness of using the distribution to model the experimental data.

12/16

Different Estimators



Plot of all e-values from the top scoring peptide assignment from a Protein Prospector search using three methods to calculate the e-values. The left-most peak in each distribution are cases where the top scoring peptide is not significantly different than a random, incorrect match.



13/16

Comparing Methods to Calculate e-values

Results of Protein Prospector database searches comparing three methods for calculating e-values. The plots on the left are the e-values of the top scoring peptide from each of 3270 spectra searched against the SwissProt database. The plots on the right are the same spectra searched against a randomized SwissProt database. For a random database, the –10log(e-value) should be centered around 0.

The Linear Tail-Fit overestimates the reliability of the top peptide assignments, while the Method of Moments and Maximum Likelihood underestimates the reliability.

Using the distribution of e-values from the search of the random database, a false positive rate can be calculated for the database search.

14/16

Conclusions

- The scores of the incorrect peptide assignments follows the extreme value distribution.
- Protein Prospector scores are better suited to linear tail-fits of log(survival) vs score.
- Linear Tail-Fit estimation of e-values overestimates the reliability of the assignment.
- Method of Moments and Maximum Likelihood methods to model extreme value distributions accurately model the observed data, but underestimate the reliability of the assignment.

15/16

Future Work

- Add the ability to calculate e-values into Protein Prospector.
  - Method for calculation to be determined.
- Use e-values to improve Protein Prospector's Discriminant Score.

16/16

Acknowledgements

NIH NCRR grant RR01614
Vincent Coates Foundation

References

1. Bradshaw, R. A. (2005). "Revised draft guidelines for proteomic data publications." **Molecular & Cellular Proteomics 4**(9):1223-5.

2. Perkins, D. N., Pappin, D. J., *et al.* (1999). "Probability-based protein identification by searching sequence databases using mass spectrometry data." **Electrophoresis**. **20**(18):3551-67.

3. Geer, L. Y. *et al.* (2004). "Open Mass Spectrometry Search Algorithm." **J Proteome Research**, **3**:958-964.

4. Craig, R. and Beavis, R. C. (2004). "TANDEM: matching proteins with tandem mass spectra." **Bionformatics 20**(9):1466-7.

5. Fenyo, D. and Beavis, R. C. (2003). "A Method for Assessing the Statistical Significance of MS Based Protein IDs Using General Scoring Schemes." **Anal Chem 75**:768-774.