On E-values for Tandem MS Scoring Schemes

In a recent article in this journal, Khatun, Hamlett, and Giddings (2008) (KHG) advance a new scoring scheme for use in conjunction with tandem mass spectrometry (MS/MS) based peptide identification. As they note, such identifications are fundamental to much proteomics research but, due to MS/MS data complexity and the scale of attendant database searches, their accuracy is limited. The scoring technique they propose, which employs a hidden Markov model (HMM) over a set of states that represent key features of MS/MS data, is convincingly motivated and exhibits good performance. The purpose of this brief note is to critique the method chosen for calibrating the HMM scores, rather than the genesis of the scores themselves.

The ubiquity of expectation (E) values, as provided by BLAST sequence based searches and based on type I extreme value (Gumbel) distributions, prompted efforts to produce analogous summaries for the seemingly similar MS/MS database searches. In particular, Fenyö and Beavis (2003) (FB) devise such a summary, and it is this approach that is employed by KHG. The appropriateness of the type I extreme value distribution (evd) for sequence based search stems from the selection of *maximal* scoring segments (Karlin and Althschul, 1990) and has strong theoretic and empiric underpinnings, although low complexity sequence constitutes an exception (Sharon *et al.*, 2005). FB consider four scoring functions and assert that the arguments used to justify the extreme value distribution pertain. However, on theoretic grounds, this does not appear to be the case, nor is it immediate that the evd pertains to KHGs HMM based scores. Further, FB contend that, for the evd, the tail of the log survival function is linear in log score and propose a corresponding estimation scheme for evd parameters. Regardless of the fact that the claimed linearity does not hold for a type I evd (it is approximate for the type II (Frechet) evd), there are existing parameter estimation schemes (e.g., Segal *et al.*, 2000) that enjoy superior efficiency and robustness properties and are readily computable. In their Supplementary Methods KHG state that log survival relates linearly to score. This approximates an exponential distribution. Again, pursuing E-value determination by linear tail fitting can be highly non-robust.

Before proceeding to demonstrate these points some general comments are in order. Firstly, the problem of tail area estimation (underlying E/p value computation) is challenging, and difficulties associated with using parametric extrapolation for such purposes are not confined to MS/MS peptide/protein identification. Nonetheless, I believe it purposeful to showcase concerns in this arena because of the additional limitations of proposed approaches. Secondly, it can be argued that non-robustness in E/p value estimation is not consequential since these values are used for downstream screening or discrimination, rather than direct interpretation in formal probabilistic terms. However, confidence or significance statements based on E/p values are still commonly proffered. And, improved discrimination can potentially be obtained by improved estimation, the shortcomings of the FB schema being avoidable.

To illustrate these concerns with the FB approach, I showcase an example using MS/MS data (kindly provided by Robert Chalkley and Aenoch Lynn), with database matching scores obtained using Protein Prospector (Chalkley *et al.*, 2005). It is important to recognize that the concerns transcend the specific scoring scheme employed but, rather, are fundamental to the estimation approach. Figure 1 shows the (smoothed) empiric density (black curve) for these scores, with superimposed fits of type I evd (red) and gamma (green) densities. As confirmed by Q-Q plots (not shown), these densities both provide good fits to the data, with the gamma proving superior. [In the majority of instances examined the evd did *not* provide a good summary.] The inset p-values under each density pertain to the maximal score, indicated by the blue arrow, that represents a true peptide match per manual verification. Figure 2 depicts the FB estimation process, based on linear fitting of log survival to log score, for two (10%, 1%) candidate tail quantiles. Two features are notable. First, the inherent and theoretically expected non-linearity is evident. Note that this pertains to scores *conforming* to the type I evd. The same phenomenon is apparent for random samples generated according to this distribution (not shown). Secondly, and more importantly, the combination of curvature and differing tail quantile specifications can have a big impact on FB derived p-values. The two upper tail prescriptions result in p-values that differ by two orders of magnitude. The difference between the p-value obtained using the (recommended) upper 10% quantile (p = 5e-06) and that obtained using maximum likelihood or method of moments fitting (p = 2.18e-09) exceeds three orders of magnitude. Further, these differences arise despite large attained R^2 values for the linear fit, exceeding 93% for the 10% quantile. Now while the p-value disparities may not be critical in and of themselves since, as noted above, they are often used for screening or discrimination purposes rather than for formal probability statements, it remains the case that they unnecessarily result from the non-robust FB estimation scheme. Moreover, even discrimination-based p-value usage can be distorted when applied across different settings due to the interplay of distributional lack-of-fit and non-robust estimation. It is important to note that the showcased disparities arose in the favorable, yet infrequent, situation where the score distribution is well approximated by a type I evd. Finally, p-values serve as inputs into multiple testing correction methods, an inherent component of peptide identification via database search, which can further compound problems due to inaccurate estimates.

In summary, I believe that such parametric attempts at inheriting BLAST style E-values for assessing significance of MS/MS database search scoring schemes should proceed with caution in view of both the difficulties associated with tail estimation and the complexities of MS/MS data. The latter point is exemplified by the framework developed by Shen *et al.*, wherein score distributions are but one of several components required to assign significance/confidence to putative peptide identifications. Additionally, a recent special issue of the *Journal of Proteome Research* (2008, Vol 7, Issue 1) was devoted to related concerns. While the focus of several contributions was on competing approaches to multiple testing correction, there was unanimity surrounding the criticality of properly framed null (referent) distributions. Even when so equipped, the illustrations herein demonstrate the need for sound p-value estimation techniques.

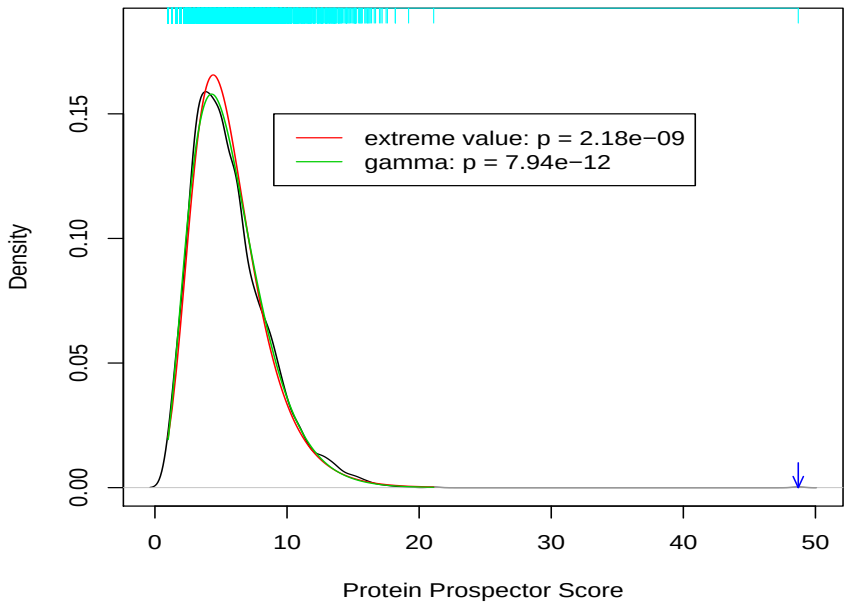

Figure 1: Type I extreme value (red curve) and gamma (green) densities fit to the Protein Prospector database matching scores as shown by the rug (teal) on the top axis and smoothed density (black). The blue arrow pinpoints the maximal score to which the inset p-values pertain.

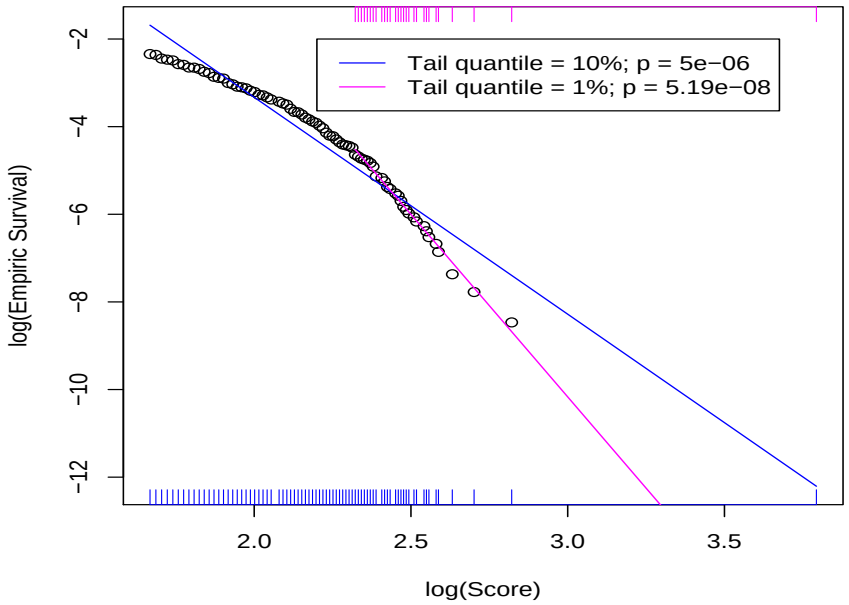

Figure 2: Illustration of the Fenyö and Beavis approach for estimating p-values via linear fitting of log(survival) to log(score) for differing upper tail prescriptions.

Sincerely,

Mark R. Segal, PhD

Professor, Epidemiology and Biostatistics
Director, Center for Bioinformatics and Molecular Biostatistics

Acknowledgments

Helpful comments were provided by Robert Chalkley, the associate editor and three referees. This work was supported by NIH Grant Number 1 UL1 RR024131-01.

References

Chalkley,R.J. *et al.* (2005) Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: II. New developments in Protein Prospector allow for reliable and comprehensive automatic analysis of large datasets. *Mol. Cell Proteomics.*, 4, 1194-1204.

Fenyö,D. and Beavis,R.C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.*, 75, 768-774.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci.*, 87, 2264-2268.

Khatun,J. *et al.* (2008) Incorporating sequence information into the scoring function: a hidden Markov model for improved peptide identification. *Bioinformatics*, 24, 674-681.

Segal,M.R. *et al.* (2000) Comparing DNA fingerprints of infectious organisms. *Statist. Sci.*, 15, 27-45.

Sharon,I. *et al.* (2005) Correcting BLAST e-Values for Low-Complexity Segments *J. Comp. Bio.*, 12, 978-1001.

Shen,C. *et al.* (2007) A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry. *Bioinformatics*, 24, 202-208.