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, Fenyo 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 Altschul, 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 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 (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. First, 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. Second, 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 scheme being avoidable.

To illustrate these concerns with the FB approach, I showcase an example using MS/MS data kindly provided by Robert Chalkley and Archon 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 empirical 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). Second, and more important, 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 ($P = 5e^{-9}$) 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
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. (2007) 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, Volume 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 (referred) distributions. Even when so equipped, the illustrations herein demonstrate the need for sound P-value estimation techniques.

ACKNOWLEDGEMENTS

Helpful comments were provided by Robert Chalkley, the associate editor and three referees.

Funding: This work was supported by NIH Grant Number 1 UL1 RR024131-01.

Conflict of Interest: none declared.

REFERENCES


