

Gene expression

In response to 'On *E*-values for tandem MS scoring schemes'

Jainab Khatun and Morgan Giddings*

Received and revised on May 23, 2008; accepted on May 28, 2008

Associate Editor: Alfonso Valencia

Contact: giddings@unc.edu

We thank Mark Segal for raising the issue of interpreting MS/MS scores. As he noted, we used a method proposed by Fenyő and Beavis (FB) (2003) to assess the significance of identification using HMM_Score. In his letter, Segal makes two basic assertions about this use: (1) that the extreme value distribution does not apply for the MS/MS database scoring systems used by FB and our HMM and (2) the linear tail fitting of the log survival function is not robust. He proposes a method that he authored as an alternative for estimating evd parameters that he says may be more robust, and also points to a method by Shen *et al.* that is specific to assessing significance of proteins/peptides identifications using MS/MS data.

While it is valuable to examine whether there exist better ways of statistically interpreting the results of MS/MS search, in his letter, Segal did not provide any clear supporting evidence for his claim that the MS/MS scorers cannot use *E*-values. In our case, we calculate a score distribution for all random matches on-the-fly, then deriving the survival function, *s*, (the cumulative probability distribution) and finally, fitting a line to log of this function for the high-scoring portion of *s*. We verified the methodology for a series of randomly chosen HMM_Score search results, observing that in all cases, the fit had very high correlation values ($R^2 > 0.9$). All subsequent validation of HMM_Score was performed using the *E*-values produced, and as reported the system performs well.

Estimating *E*-values from the tail of the survival function may not be a precise statistical estimator of significance, but in the article we did not claim that it was. We stated about the *E*-value that it will: 'convert a score expressed on an arbitrary value range into a measure of the relative uniqueness of a given match score', a statement that we stand by. This *E*-value approach has the empirical benefit of strongly discriminating true and false positive identifications by estimating the rate of random matches versus score value for a given query. Another approach, like Segal's, may have desirable properties for the precise calculation of statistical significance, but like any approach, it may need extensive real-world testing to determine whether it also improves discrimination capability, which is just as important.

Regarding the framework that Segal cited by Shen *et al.*, it is focused upon the problem of combining multiple peptide identifications to determine whether the protein was present, so does not appear to be applicable to the single-peptide scoring we reported in our own article.

We are encouraged that this kind of discussion and advancement is occurring regarding the statistical interpretation of MS/MS search scores. It was only a few years ago that nearly all MS/MS scoring schemes produced scores of arbitrary numerical range and distribution, which required significant training by a human operator to properly interpret. Improved statistical scoring will provide improved means for end-user interpretation of search results, which is something both Segal and we strongly agree on.

Conflict of Interest: none declared.

*To whom correspondence should be addressed.