Published in final edited form as:
Nat Methods. 2017 June 29; 14(7): 644. doi:10.1038/nmeth.4339.

Response to “Mass spectrometrists should search for all peptides, but assess only the ones they care about”

William Stafford Noble and
Department of Genome Sciences, Department of Computer Science and Engineering, University of Washington

Uri Keich
School of Mathematics and Statistics F07, University of Sydney

We find much to agree with in the commentary by Clement et al. Overall, it is clear that we are engaged in the same general project: to first ensure the validity of our statistical confidence estimates and thereafter to maximize our statistical power. We also agree that controlling the false discovery rate (FDR) among matches to a large peptide database and then reporting results relative to a selected subset of peptides fails to correctly control the FDR. Indeed, this point has been made previously on multiple occasions [1–3] and is well established in the statistical literature [4]. We also agree that the “sub-sub” strategy—searching a subset database and evaluating the FDR within that subset—necessarily forces some matches between peptides in the subset and spectra that were generated by peptides outside of the database.

This leads to our two points of contention. First, Clement et al. claim that their proposed “all-sub” strategy leads to improved statistical power relative to sub-sub. In support of this claim they report empirical results on two data sets. We contend that all-sub is not always better than sub-sub. Accordingly, we constructed a different setup that allowed us to more accurately characterize false positive spectrum identifications. Specifically, we ran a concatenated set of spectra—from 18 purified proteins (ISB18 [5]) and from the plant *Arabidopsis thaliana* [6]—against a corresponding concatenated database. Contrary to what Clement et al. found, in this setting the relative performance of the two methods is reversed: at a 1% FDR threshold, sub-sub accepts 11,416 PSMs, whereas all-sub accepts only 10,307. All-sub’s loss of statistical power is due to the large size of the *Arabidopsis* database (see Supplementary Note).

Now to the second point of contention. In addition to claiming superior statistical power of the all-sub procedure, Clement et al. imply that the sub-sub strategy leads to invalid FDR control. As evidence, they point to the number of subset PSMs that matched a different peptide sequence in the complete search (all-all) and the subset search (sub-sub). However, their analysis fails to account for the possibility that some of these PSMs may be incorrect in the all-all search and correct in the sub-sub search. Indeed, as the size of the competing,

Data availability All data used in this work is publicly available via the URLs specified in the Supplementary Note.
Competing financial interests The authors declare that they have no competing financial interests.

complement database gets larger, the probability that a correct match to the subset database will receive a lower score than an incorrect match in the complement database goes up. This is precisely the effect that sub-sub aims to avoid. In the context of our simulation, Clement et al. are concerned that by forcing *Arabidopsis* spectra to match against the ISB18 database, we will create many false positive PSMs. Fortunately, in our experimental setup, we can directly observe this rate of false matching: among the 11,416 PSMs accepted by sub-sub, only 41 (0.36%) involve an *Arabidopsis* spectrum. This is well below the 1% FDR threshold. Furthermore, we note that in the subset database search, 1127 of the accepted PSMs involving ISB18 spectra actually switch to matching *Arabidopsis* peptides when we search against the combined database. According to the arguments laid out by Clement et al., this rate of switching implies that that the actual sub-sub FDR is ~10%. However, in our setup, we know that those ISB18 spectra are certainly not better off when matched to *Arabidopsis* peptides.

Thus, though all-sub may indeed provide superior statistical power in some settings, we have shown that this is not always the case. Precisely characterizing the situations in which a given analysis strategy is optimal will require further research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH awards R01 GM121818 and P41 GM103533 to WSN.

References

1. Baker PR, Medzihradszky KF, Chalkley RJ. Improving software performance for peptide electron transfer dissociation data analysis by implementation of charge state- and sequence-dependent scoring. Molecular and Cellular Proteomics. 2010; 9:1795–1803. [PubMed: 20513802]
2. Fu Y, Qian X. Transferred subgroup false discovery rate for rare post-translational modifications detected by mass spectrometry. Molecular and Cellular Proteomics. 2014; 13(5):1359–1368. [PubMed: 24200586]
3. Woo S, Cha SW, Na S, Guest C, Liu T, Smith RD, Rodland KD, Bafna V. Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. Proteomics. 2014; 14(23–24):2719–2730. [PubMed: 25263569]
4. Efron B. Simultaneous inference: When should hypothesis testing problems be combined? The Annals of Applied Statistics, pages. 2008:197–223.
5. Klimek J, Eddes JS, Hohmann L, Jackson J, Peterson A, Letarte S, Gafken PR, Katz JE, Mallick P, Lee H, Schmidt A, Ossola R, Eng JK, Aebersold R, Martin DB. The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. Journal of Proteome Research. 2008; 7(1):96–1003. [PubMed: 17711323]
6. Engineer CB, Ghassemian M, Anderson JC, Peck SC, Hu H, Schroeder JI. Carbonic anhydrases, EPF2 and a novel protease mediate CO2 control of stomatal development. Nature. 2014; 513(7517): 246–250. [PubMed: 25043023]