Mass spectrometrists should search only for peptides they care about

William Stafford Noble

Analysis pipelines that assign peptides to shotgun proteomics mass spectra often discard identified spectra deemed irrelevant to the scientific hypothesis being tested. To improve statistical power, I propose that researchers remove irrelevant peptides from the database prior to searching rather than assigning these peptides to spectra and then discarding the matches.

Over the past decade, perhaps the most significant trend in the analysis of mass spectrometry–based shotgun proteomics data has been the increasing statistical rigor of most analysis pipelines. No longer are reviewers or editors satisfied with spectrum identifications defined with respect to arbitrary score thresholds or by using 'rules of thumb' based on multiple thresholds. Most proteomics journals now require that statistical confidence estimates be reported. A variety of methods have been devised for assigning confidence estimates to individual matches, using parametric[1] or exact procedures[2,3]. Perhaps most importantly, target-decoy analysis[4,5] provides a straightforward method for estimating the false discovery rate (FDR), defined as the percentage of incorrect identifications, associated with nearly any data set and scoring procedure.

However, despite these advances in statistical confidence estimation, I believe that one problematic protocol remains in common use. The protocol is quite general and involves testing more hypotheses than we are actually interested in. To illustrate the idea, consider the analysis of mass spectra derived from the erythrocytic cycle of the malaria parasite *Plasmodium falciparum*. Because erythrocytic *Plasmodium* parasites inhabit human red blood cells, any *Plasmodium* proteomics experiment will inevitably generate spectra from a mixture of human and *Plasmodium* peptides. It is therefore common to search the observed spectra against a combined database of human and *Plasmodium* peptides and then to discard the spectra that match to human peptides.

From a statistical perspective, I believe that this protocol is suboptimal, in the sense that it needlessly sacrifices statistical power. In particular, at a fixed FDR threshold, we can obtain a larger set of identifications by searching the spectra against only the *Plasmodium* peptides. To understand why this is the case, we need only consider a single spectrum *s* searched against either the *Plasmodium* database or the '*Plasmodium*+human' database. If the spectrum has an associated precursor mass of, say, 2,000 Da, and if we search the databases using a tolerance of 10 p.p.m., then the *Plasmodium* database yields 17 candidate peptides and the *Plasmodium*+human database yields 29 candidate peptides. Let us assume that we are using a method such as MS-GF+[2] that assigns a *P* value to each possible match. After selecting the best-scoring candidate peptide (i.e., the one with the smallest *P* value), a Bonferroni adjustment requires that, to achieve a statistical significance of $P < 0.01$, we must observe a *P* value of $0.01/17 = 5.9 \times 10^{-4}$ from the *Plasmodium* database but a *P* value of $0.01/29 = 3.4 \times 10^{-4}$ from the *Plasmodium*+human database. Thus, whenever our best-scoring match against the *Plasmodium* database receives a *P* value in between these two values, we will fail to identify that match if we choose also to search the human database. This is ironic, because the only potential gain we achieve by searching the human database is an identification that we have no interest in and will ultimately discard.

Many mass spectrometrists, when presented with this idea, would likely express concern that the spectrum identifications produced by this simpler protocol will be contaminated with human spectra. The reasoning goes like this: even if spectrum *s* obtained a score that exceeds a given threshold when we searched it against the *Plasmodium* database, that spectrum might have received an even larger score if we had searched it against the combined *Plasmodium* and human databases. Although this statement is certainly true, it misses the point of our statistical confidence estimation procedure, which is to accurately estimate the FDR associated with a given collection of identified spectra. Well-calibrated statistical confidence estimates should allow us to skip the testing of these extraneous hypotheses.

One potential source of confusion in the assignment of confidence estimates to identified spectra is that the most commonly used assignment method—assigning FDRs using target-decoy competition—does not explicitly make use of *P* values. It is not obvious, therefore, that notions such as Bonferroni adjustment and multiple-hypothesis testing are applicable in this domain. The following two case studies, which I carried out using target-decoy FDR estimation, illustrate that these concepts

William Stafford Noble is in the Department of Genome Sciences and the Department of Computer Science and Engineering, University of Washington, Seattle, Washington, USA.
e-mail: william-noble@uw.edu

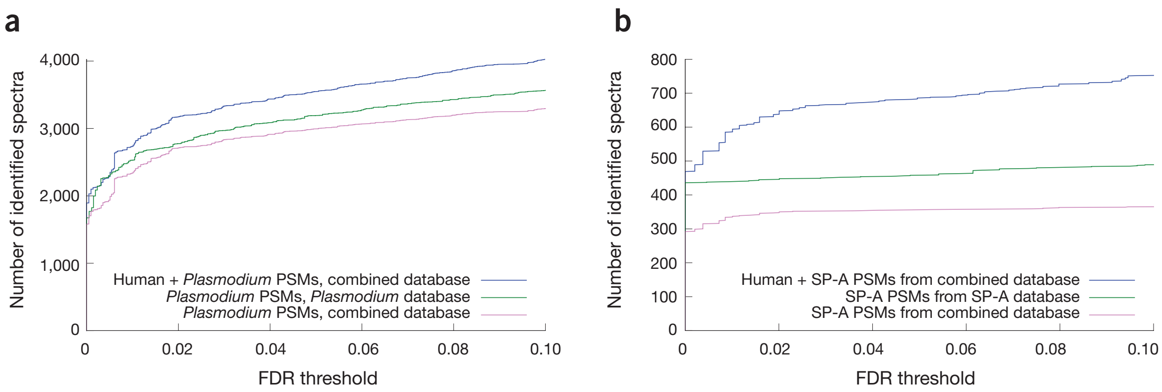

Figure 1 | Boosting statistical power by eliminating irrelevant hypotheses. (**a**) The number of accepted peptide-spectrum matches (PSMs) as a function of FDR threshold for the *Plasmodium* data set. Searching a combined database of *Plasmodium* and human sequences yields more identified spectra overall (blue) but fewer *Plasmodium* identifications (violet) than does searching just the *Plasmodium* database (green). (**b**) Analysis of isoforms of SP-A by searching an SP-A variant database alone or by searching a database consisting of the entire human proteome plus SP-A variants. Source data are available in **Supplementary Data 4**.

do indeed apply. Any valid confidence estimation procedure, even one that does not make explicit use of *P* values, must take into account the large number of peptides in the database and the large number of spectra in a given experiment.

Two case studies

To illustrate the loss of statistical power associated with searching a large database, I analyzed two recently published collections of mass spectra (see also **Supplementary Methods** and **Supplementary Data 1–3**). The first contains 12,478 high-resolution *Plasmodium* spectra[6]. I searched the spectra against a *Plasmodium* database, with and without the human database appended (**Supplementary Data 1**). I used the MS-GF+ search engine[2] and estimated FDRs using target-decoy competition[5]. Searching the combined database always yielded fewer *Plasmodium* identifications across FDR thresholds up to 10% (**Fig. 1a**). In particular, at an FDR threshold of 1%, the combined search assigned *Plasmodium* peptides to 2,339 spectra, whereas the *Plasmodium*-only search assigned 2,530 such spectra, an increase of 8.2%. Furthermore, only 2 of the 2,530 spectra identified in the *Plasmodium*-only search received a better score in the corresponding *Plasmodium*+human search, indicating that the *Plasmodium*-only results are not highly contaminated with spurious matches due to the presence of human-derived spectra.

A potential complication arises when peptides are shared between *Plasmodium* and human. In a tryptic digestion with no missed cleavages and no variable modifications, the *Plasmodium* and human databases contain 221,567 and 432,840 peptides, respectively, with an overlap of 916 peptides. One risk associated with searching only the *Plasmodium* database is that human peptides from this overlap set might be misidentified as *Plasmodium* peptides. I propose that the solution to this problem is to check whether any of the identified *Plasmodium* peptides occur in human, and to handle these identifications accordingly. Indeed, one might wish to eliminate these overlapping peptides from the *Plasmodium* database a priori, thereby further reducing the multiple-testing burden.

I also analyzed a collection of 1,503 fragmentation spectra generated during an investigation of coding variants and isoforms of pulmonary surfactant protein A (SP-A)[7] (**Supplementary Data 2**). SP-A proteins were purified from bronchial lavage fluid and subjected to tandem mass spectrometry analysis. The purification procedure was necessarily imperfect, however. In that study, the resulting spectra were searched against the entire human proteome, augmented with a collection of six known SP-A variants. I repeated this analysis by searching with Tide[3] against the combined 'human+SP-A' database as well as against only the known variants of SP-A. Similar to the *Plasmodium* analysis, my analysis here showed that searching against only the proteins of interest yields much better statistical power (**Fig. 1b**). At a 1% FDR threshold, the combined search assigned 345 spectra to peptides from SP-A, whereas the search against the SP-A database identified 448 spectra, an increase of 29.9%.

This practice of sacrificing statistical power by considering irrelevant hypotheses is common. For example, I propose that any proteomics experiment that targets a particular pathway or set of pathways would benefit from using a protein database consisting of only the proteins of interest. Similarly, any study that aims to identify only phosphorylation sites could gain statistical power by not searching for unphosphorylated peptides. When searching for cross-linked peptides, uninteresting species such as non-cross-linked, self-loop and dead-end peptides should be left out of the database. I believe a good rule of thumb is that if you are going to do a post-filtering step to eliminate some of the matches from your analysis, and if those matches are not relevant to the scientific hypothesis you are testing, then you should consider eliminating those peptides beforehand to avoid having to correct for these extra statistical tests.

A practical follow-up question to both the *Plasmodium* and SP-A examples is how the magnitude of the loss in statistical power varies as a function of the number of irrelevant proteins included in the search. To address this question empirically, I carried out searches using a series of randomly downsampled human databases. For the *Plasmodium* data set, I observed a rapid loss of power even when only 10,000 human proteins are added to the database: at a 1% FDR threshold, the number of spectra assigned to *Plasmodium* peptides drops by 3.2%, from 2,729 to 2,643 (**Fig. 2a**). Power continues to drop relatively smoothly as the database grows larger. In the SP-A case, the initial drop in power is

Table 1 | Viral peptides identified in the honeybee data set

Spectra	Viral	Honeybee	Combined	Species
52	IWHHTFYNELR	Same	Both	Honeybee
46	HKGVMVGMGQK	HQGVMVGMGQK	Both	Honeybee
17	LAVNMVPFPR	Same	Both	Honeybee
7	IIAQVVSSITASLR	LIGQIVSSITASLR	Honeybee	Honeybee
3	SYELPDGQVIKIGSER	SYELPDGQVITIGNER	Honeybee	Honeybee
16	IGPISEVASGVK	Various (6 decoys)	Viral	Kashmir bee virus
6	DYMSYLSYLYR	Various (5 decoys)	Viral	Acute bee paralysis virus
4	IDTPMAQDTSSAR	Various (4 decoys)	Viral	Acute bee paralysis virus and Kashmir bee virus
1	VNNLHEYTK	NVNITFPQGK	Viral	Sacbrood virus

more dramatic. Even adding 1,000 human peptides reduces power by 13.4% (434 to 376 peptides) at 1% FDR (**Fig. 2b**). The reason for this behavior is that, in this case, less than half (709 of 1,503) of the observed spectra have associated precursor *m*/*z* values that lie within ±3 *m*/*z* of the mass of at least one peptide in the SP-A database. The remaining 794 spectra are therefore not matched at all in the SP-A search. When we add human proteins to the database, most of those 794 spectra match at least one candidate peptide. Thus, in this case, adding human peptides to the database increases the multiple-testing burden in two ways: by increasing (i) the number of candidate peptides per spectrum and (ii) the total number of spectra under consideration.

Beyond maximizing statistical power

In practice, simply maximizing statistical power may not be the only or even the primary concern; applying the approach I propose above may not be the best choice for every situation. For example, focusing an experiment on a small, targeted collection of peptides necessarily eliminates the possibility of serendipitous, unexpected discoveries. Indeed, one of the great powers of tandem mass spectrometry is its depth of coverage, with the potential to query a large proportion of the proteome in a single experiment.

A second important consideration is that the analysis of otherwise irrelevant peptides can provide a useful sanity check. For example, common contaminants such as human keratin should probably always be included in any search. Overall, the total number of such contaminants is generally quite small, and their presence in the database will likely have a negligible impact on statistical power.

Irrelevant peptides can also be helpful for identifying problems with the statistical analysis. For example, Bromenshenk *et al.* reported an analysis of honeybee-derived spectra that yielded sequences from two previously unreported RNA viruses—Varroa destructor virus 1 and Kakugo virus—in North American honeybees as well as an invertebrate iridescent virus, which led them to conclude that such viruses are associated with collapsed colonies[8]. The analysis involved searching the spectra against a database containing only bacterial and viral sequences, without including the honeybee proteins that are expected to match the majority of the spectra. A total of 3,000 peptides were confidently identified from over 900 different microbial species. Three critiques of this study were subsequently published by two different research groups, pointing out various problems with the reported results and with the analysis procedures[9–11]. Knudsen and Chalkley[9] hypothesized that the tool PeptideProphet[12] was provided with an input data set containing no correct identifications, and as a result the tool gave invalid results. Both critical reassessments claimed that had Bromenshenk *et al.* included the honeybee proteins in their initial database search, mistakes could have been avoided.

To investigate this claim, I reanalyzed the full set of 262,572 spectra from the original Bromenshenk *et al.* study (**Supplementary Data 3**). As with the *Plasmodium* and SP-A case studies, I searched the spectra against a viral database, a honeybee database and a concatenation of the two. At an FDR threshold of 1%, the Tide search engine identified 36,371 spectra by searching the honeybee database. Searching the viral database, on the other hand, identified only 152 spectra. These viral identifications correspond to only nine distinct peptides (**Table 1**). In agreement with previous reanalyses of this data set[9,11], none of the nine peptides comes from the Varroa destructor 1, Kakugo or invertebrate

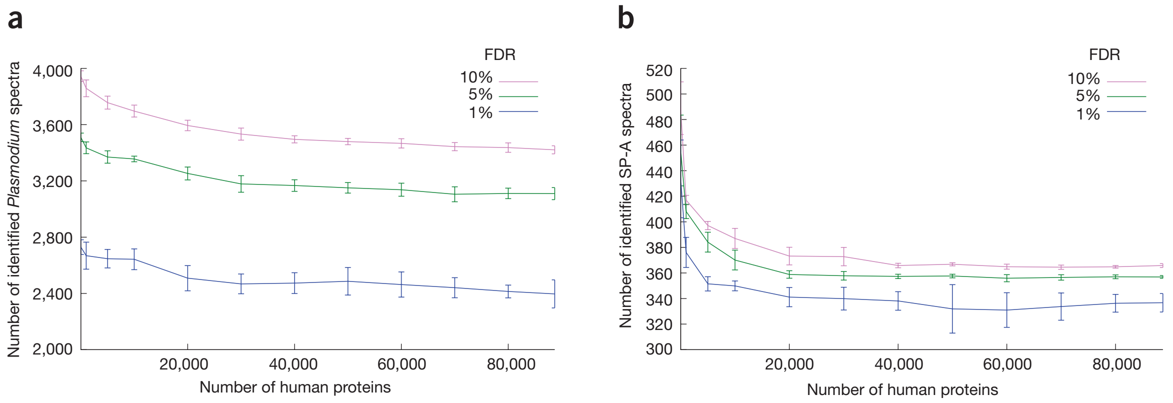

Figure 2 | Loss of statistical power as a function of database size. The *Plasmodium* and SP-A data sets were searched, using Tide, against a series of databases containing various numbers of human proteins. (**a**,**b**) The plots show the number of *Plasmodium* (**a**) and SP-A (**b**) proteins identified at various FDR thresholds as a function of the number of human proteins included in the database. Results are averaged over ten different decoy databases, and error bars correspond to standard deviation. Source data are available in **Supplementary Data 5**.

iridescent viruses. Two of the nine peptides occur in both the honeybee and viral databases, and one occurs in both databases but differs by a substitution of isobaric amino acids (lysine → glutamine). Together, these three shared peptides account for 115 of the 152 spectra. Among the remaining six viral peptides, four appear to be matches to actual viral sequences from the Kashmir bee virus, the acute bee paralysis virus and the sacbrood virus. Only two peptides that were identified as viral in the initial search, corresponding to ten distinct spectra, have a better match to a homologous honeybee peptide in the combined search.

My analysis suggests two conclusions. First, in rare cases, leaving irrelevant peptides out of the database may lead to incorrectly identified spectra owing to homology between the proteins in the database and the left-out proteins: in this case, 10 identifications out of 152. Thus, the risk of such false positive identifications must be weighed against the potential gain in statistical power (loss of false negatives) when deciding upon an analysis strategy. Second, in the particular case of the honeybee data set, the choice of protein database does not explain the incorrect results obtained by Bromenshenk *et al.*[8]. Instead, the problem apparently lay in how the analysis was performed.

Challenges and future directions

One challenge associated with my proposed rule of thumb is that if we use a decoy-based estimation procedure, and if our database consists of only a handful of proteins, then the resulting confidence estimates will likely be inaccurate. Indeed, one published standard explicitly warns about the difficulty of achieving accurate confidence estimates when the database contains fewer than 1,000 proteins[13]. Three points are worth considering with respect to this problem. First, any empirical confidence estimation protocol such as target-decoy competition yields intrinsic variance due to the stochastic generation of decoys. This variance is not removed by using a deterministic decoy-generation scheme such as reversing peptides. Furthermore, as the confidence estimates get small, the relative magnitude of the variance increases relative to the magnitude of the estimate (**Fig. 2**). Second, if we use a shuffling procedure to generate decoys, then it is possible to reduce the variance in our confidence estimates substantially by repeating the target-decoy competition many times[14]. Such an approach is particularly applicable when the protein database is small, because the cost of repeated searches is relatively low. Third, the use of analytic methods to compute exact *P* values with respect to a particular null model avoids the variance associated with empirical confidence-estimation schemes[2,3].

An alternative protocol to the one I propose here involves first searching the spectra against a database containing only the irrelevant peptides and eliminating from the data set all spectra that match this 'garbage' database with high confidence. The remaining spectra could then be searched against the database of interesting peptides. For stringent FDR thresholds, I believe that this approach is likely to yield slightly better power than the simple strategy proposed here, at the expense of being somewhat more complicated to implement.

Thus far, I have focused on assigning confidence estimates to peptide-spectrum matches. Confidence estimation at the level of peptides—taking into account multiple spectra matching to the same peptide—or at the level of proteins—taking into account the many-to-many mapping between peptides and proteins—is considerably more challenging. Nonetheless, existing methods for assigning peptide- and protein-level confidence estimates typically do so by aggregating spectrum-level evidence[15–17]; hence, any gains in statistical power at the spectrum level achieved by leaving out irrelevant peptides should in principle yield concomitant gains in power at the peptide and protein levels.

The relatively simple idea I propose here, of limiting the hypothesis space to only hypotheses of interest, is an expression of a more general statistical goal: to explore the hypothesis space in a fashion that maximizes statistical power. Thus, a clear avenue for future work at the interface of statistics and proteomics lies in the application of existing protocols for stratified multiple-testing control[18].

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.3450).

ACKNOWLEDGMENTS

I thank G. Knudsen for providing access to the honeybee data set. This work was funded by US National Institutes of Health awards R01 GM096306 and P41 GM103533.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

1. Klammer, A.A., Park, C.Y. & Noble, W.S. *J. Proteome Res.* **8**, 2106–2113 (2009).
2. Kim, S., Gupta, N. & Pevzner, P.A. *J. Proteome Res.* **7**, 3354–3363 (2008).
3. Howbert, J.J. & Noble, W.S. *Mol. Cell. Proteomics* **13**, 2467–2479 (2014).
4. Moore, R.E., Young, M.K. & Lee, T.D. *J. Am. Soc. Mass Spectrom.* **13**, 378–386 (2002).
5. Elias, J.E. & Gygi, S.P. *Nat. Methods* **4**, 207–214 (2007).
6. Pease, B.N. *et al. J. Proteome Res.* **12**, 4028–4045 (2013).
7. Foster, M.W. *et al. J. Proteome Res.* **13**, 3722–3732 (2014).
8. Bromenshenk, J.J. *et al. PLoS ONE* **5**, e13181 (2010).
9. Knudsen, G.M. & Chalkley, R.J. *PLoS ONE* **6**, e20873 (2011).
10. Foster, L.J. *Mol. Cell. Proteomics* **10**, M110.006387 (2011).
11. Foster, L.J. *Mol. Cell. Proteomics* **11**, A110.006387-1 (2012).
12. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. *Anal. Chem.* **74**, 5383–5392 (2002).
13. American Society for Biochemistry and Molecular Biology. Revised publication guidelines for documenting the identification and quantification of peptides, proteins, and post-translational modifications by mass spectrometry. *Molecular and Cellular Proteomics* http://www.mcponline.org/site/misc/peptide_and_protein_identification_guidelines.pdf (2015; accessed 6 April 2015).
14. Keich, U. & Noble, W.S. *J. Proteome Res.* **14**, 1147–1160 (2015).
15. Nesvizhskii, A.I., Keller, A., Kolker, E. & Aebersold, R. *Anal. Chem.* **75**, 4646–4658 (2003).
16. Serang, O., MacCoss, M.J. & Noble, W.S. *J. Proteome Res.* **9**, 5346–5357 (2010).
17. Cox, J. & Mann, M. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
18. Dudoit, S. & van der Laan, M.J. *Multiple Testing Procedures with Applications to Genomics* (Springer, 2008).