Published in final edited form as:

Nat Methods. ; 9(5): 455–456. doi:10.1038/nmeth.1983.

Toward Objective Evaluation Of Proteomic Algorithms

John R. Yates III, Sung Kyu Robin Park, Claire M. Delahunty, Tao Xu, Daniel Cociorva, and Paulo Costa Carvalho¹

10550 North Torrey Pines Road, SR11, The Scripps Research Institute, LaJolla, CA 92037

¹Instituto Carlos Chagas, Fiocruz, Rua Prof. Algacyr Munhoz Mader, 3775 – CIC, Curitiba, PR, Brazil, 81350-010

Proteomic analysis of peptides and proteins using mass spectrometry was enabled by the development of software tools that can identify^{1–5} and quantify peptides^{6–8} and their modifications^{9,10} on a large scale. After publication of an original algorithm to analyze proteomic mass spectrometry data, there is often a proliferation of similar software tools reported in the literature. The development and optimization of software tools for widespread use in proteomics has been complicated by the availability of an array of mass spectrometers which employ different mass analyzers and collision cells for peptide analysis. These instruments vary in their fragmentation processes and collision energies and analysis tools are frequently written with data from a single kind of instrument. Software that is developed for a particular type of tandem mass spectrometer may be inadvertently or intentionally optimized for data from that instrument and may be less well suited for more general use. Dissemination of software to other laboratories can validate tools for use on other instrument platforms or in other types of experiments.

The creation of new software tools that claim to be novel or to improve on published algorithms frequently compels reviewers to request comparisons with existing software tools. Such comparisons can be useful as a way to illustrate performance differences inherent in different algorithms. However, for comparisons to be instructive for the proteomics field, they must be fair and objective. Often, the comparisons of tools are unintentionally biased simply because the user is inexperienced with foreign software and a detailed operating manual is unavailable. Additionally, comparisons may be biased because a user fails to put forth his best effort to optimize foreign software for a particular dataset. These issues can lead to confusing comparisons which claim to be fair and objective analyses, but instead raise doubts about the veracity of the analyses or the expertise of the authors. For example, Balgley, et.al., conducted a comparison of the protein search engines Mascot, OMSSA, SEQUEST and X!Tandem and claimed that OMSSA identified the largest number of tandem mass spectra¹¹. In contrast, Nahnsen, et. al. evaluated the search engines Mascot, X!Tandem and OMSSA and found that OMSSA generated the lowest number of identified peptides from a digested 18-protein mixture run on an LTQ-Orbitrap instrument¹². Our own experience suggested that the results of Balgley, et.al. were incorrect, but when we attempted to access their test data set to verify their results, the authors claimed that the data set was proprietary and not available, making it impossible to confirm or refute their results. Such conflicting results illustrate the point that as search engines have become more sophisticated, more expert knowledge has become necessary to benchmark the capabilities of the software and accepted standards are necessary to make reliable comparisons. For example, a set false discovery rate (FDR) has often been used when comparing search results from different algorithms, but recent papers argue that FDR is not a dependable benchmark^{13,14}. Without a consensus for an accepted standard, published comparisons^{15,16} may be confusing and viewed with skepticism.

There have been fewer publications comparing protein quantification analysis tools, but a recent paper illustrates that similar problems can occur. Colaert, et. al., used a stable isotope labeled data set (labeled with heavy Lys and Arg) from human neuroblastoma SHEP cells to compare four quantitative analysis tools: Census, MaxQuant, MsQuant and Mascot Distiller¹⁷. A technique known as “COFRADIC” was used to analyze the sample and identify only the methionine-containing tryptic peptides by LC-MS/MS on an Orbitrap XL mass spectrometer. After protein identification with the Mascot algorithm, the labeled samples were analyzed with the four different quantification algorithms. The authors reported that in their hands the quantification algorithms yielded a very wide range of quantified proteins: Census (454), MaxQuant (1452), MsQuant (2066) and Mascot Distiller (2092). After obtaining the dataset from the authors, we found strikingly different numbers in our re-analysis of the data. Using their database search results and the same Census version (1.54, which is not the most recent version) as Colaert, et al, we found: 4752 redundant proteins (RD), 4431 non-redundant proteins (NR), and 2555 non redundant proteins with 3 peptides or more (UN) (Table 1). The “forward labeling” Census results were consistent with the “reverse labeling” experiment: 4694 redundant proteins, 4371 non-redundant proteins and 2516 non-redundant proteins with 3 or more peptides. A log ratio plot of the quantified peptides from both the forward and reverse labeling samples showed a tight distribution, which was same quality control metric used by Colaert et al. The error made in the original analysis was that the authors did not specify a modification mass shift in Census parameters, which caused selection of incorrect precursors for all modified peptides. As COFRADIC intentionally oxidizes methionine residues, the lack of such a parameter caused the program to miss most of the modified peptides.

The examples cited here demonstrate some of the pitfalls of comparing software for proteomic analysis. Comparisons are easily influenced by the familiarity and expertise of the authors with the programs being compared, and results may be biased unless great effort is made to achieve the same level of competence with all algorithms being compared. A high level of expertise is difficult to obtain since detailed user manuals are rarely provided for academic software, and experience is gained only in a trial and error learning process. Unless a benchmarking data set and results are provided by the program’s developers, it is difficult to know when expert-level use has been achieved. In addition, different programs may use different data file formats and the conversion of data between formats provides another source of error which can result in false comparisons.

How, then, do we create confidence in the accuracy in analyses performed to compare software tools? A few general problems can be easily overcome, and we propose four steps to provide uniformity in comparisons:

1. **Provide Benchmark Data Sets.** To provide a uniform and reproducible comparison between algorithms, a positive control with a standard data set should be provided and the results presented along with the parameter file. The standard or calibration data set should return the same answers for an algorithm regardless of who uses it so there will be confidence that analyses were correctly performed. A benchmark dataset provides a “chain of legitimacy” for all subsequent comparisons. Such an analysis establishes that the user has a level of expertise with programs being compared. Ideally, one data set could be used by all, but this is unlikely to happen. Thus, when a new algorithm is published it should be accompanied by a benchmark data set and results. New users can establish their expertise with an algorithm through the use of this data set.
2. **Make All Test Sets Available.** Additional or unusual data sets can still be used to document special performance characteristics of algorithms, but they should be made freely available so exact results can be replicated by other laboratories or

accessed for future comparisons. A current problem with data repositories, with the exception of PRIDE, is that sites for data sharing appear to be collapsing. PRIDE does not allow the deposition of raw data. Raw data provides background noise and dynamic range that can create a fair test of sensitivity and specificity. In the absence of a repository in a government organization, we have established a site to host test sets for proteomic software at <http://www.mstestdatasets.org>. This site currently contains a few test sets.

3. **Provide File Conversion Information.** When using raw MS data for comparisons, there are several steps in file conversion where errors or variations can occur that will affect the mass spectrometry data available for the analysis. This can be particularly true when comparing quantitative analysis algorithms that go through different pipelines and may require extraction of different data types. Simple errors can occur converting from one format to another, so data should be provided to establish that information is not lost during conversions. These measurements should provide general information on a data set, i.e. how many total scans, how many MS1 scans and MS2 scans are present in the file? If QC filters were employed on the data set, what did they do and how many spectra were removed? This information can establish that differences observed between algorithms aren't simply differences in data preprocessing.
4. **Software version.** Software tools evolve over time, so it is important for evaluators to compare up-to-date tools for fair comparisons. For example, an older version could have been designed only for low resolution mass spectrometric data, while a newer version of the same software tool supports high resolution data. Using an old version of software with improper input data sets can lead to different or poor results.

Conclusion

By creating a uniform path for the testing of algorithms with common datasets and procedures, we can instill confidence that comparisons have been objectively and honestly performed. A benefit is that real advances and innovation will be easily recognized and the field will move forward more quickly.

References

1. Eng JK, McCormack AL, Yates JR. J Am Soc Mass Spectrom. 1994; 5:976.
2. Sadygov R, Yates JR III. Analytical Chemistry. 2003; 75:3792. [PubMed: 14572045]
3. Geer LY, Markey SP, Kowalak JA, et al. J Proteome Res. 2004; 3(5):958. [PubMed: 15473683]
4. Craig R, Beavis RC. Bioinformatics. 2004; 20(9):1466. [PubMed: 14976030]
5. Perkins DN, Pappin DJ, Creasy DM, et al. Electrophoresis. 1999; 20(18):3551. [PubMed: 10612281]
6. MacCoss MJ, Wu CC, Liu H, et al. Anal Chem. 2003; 75(24):6912. [PubMed: 14670053]
7. Park SK, Venable JD, Xu T, et al. Nat Methods. 2008; 5(4):319. [PubMed: 18345006]
8. Cox J, Matic I, Hilger M, et al. Nat Protoc. 2009; 4(5):698. [PubMed: 19373234]
9. Beausoleil SA, Villen J, Gerber SA, et al. Nat Biotechnol. 2006; 24(10):1285. [PubMed: 16964243]
10. Lu B, Ruse C, Xu T, et al. Anal Chem. 2007; 79(4):1301. [PubMed: 17297928]
11. Balgley BM, Laudeman T, Yang L, et al. Mol Cell Proteomics. 2007; 6(9):1599. [PubMed: 17533222]
12. Nahnsen S, Bertsch A, Rahnenfuhrer J, et al. J Proteome Res. 2011; 10(8):3332. [PubMed: 21644507]

13. Gupta N, Bandeira N, Keich U, et al. *Journal of the American Society for Mass Spectrometry*. 2011; 22:1111. [PubMed: 21953092]
14. Barboza R, Cociorva D, Xu T, et al. *Proteomics*. 2011; 11(20):4105. [PubMed: 21834134]
15. Kapp EA, Schutz F, Connolly LM, et al. *Proteomics*. 2005; 5(13):3475. [PubMed: 16047398]
16. Ramos-Fernandez A, Paradela A, Navajas R, et al. *Mol Cell Proteomics*. 2008; 7(9):1748. [PubMed: 18515861]
17. Colaert N, Vandekerckhove J, Martens L, et al. *Methods Mol Biol*. 2011; 753:373. [PubMed: 21604136]

\$watermark-text

\$watermark-text

\$watermark-text

Table 1
Reanalysis of the Colaert et al data set with correct parameters for Census. Results for the other software tools are from Colaert et al.

Software	Forward labeling				Reverse labeling			
	Census	MsQuant	MaxQuant	Mascot distiller	Census	MsQuant	MaxQuant	Mascot distiller
Quantified proteins	4,752(RD) 4,431(NR) 2,555(UN)	2,066	1,452	2,092	4,694(RD) 4,371(NR) 2,516(UN)	2,425	1,689	2,135
validated proteins with up regulated ratio ¹	57	38	45	35	93	94	55	71
validated proteins with down regulated ratio ¹	127	5	29	37	138	1	59	63

¹ Colaert et al provided biological validation for proteins present in the sample.