JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

LETTER TO THE EDITOR

A Call for Benchmark Data in Mass Spectrometry-Based Proteomics

Jens Allmer

Molecular Biology and Genetics, Izmir Institute of Technology, Urla, Izmir.

Received: 02 October 2012 **Accepted:** 26 October 2012 **Available Online**: 28 October 2012

Abstract

Proteomics is a quickly developing field. New and better mass spectrometers, the platform of choice in proteomics, are being introduced frequently. New algorithms for the analysis of mass spectrometric data and assignment of amino acid sequence to tandem mass spectra are also presented on a frequent basis. Unfortunately, the best application area for these algorithms cannot be established at the moment. Furthermore, even the accuracy of the algorithms and their relative performance cannot be established. This is due to the lack of proper benchmark data. This letter first introduces the field of mass spectrometry-based proteomics and then defines the expectations of a well-designed benchmark dataset. Thereafter, the current situation is compared to this ideal. A call for the creation of a proper benchmark dataset is then placed and it is explained how measurement should be performed. Finally, the benefits for the research community are highlighted.

Keywords: Mass spectrometry; Proteomics; Benchmark data; Database search; De novo sequencing; Fragmentation analysis.

Abbreviations:

MS, Mass Spectrometry; MALDI, Matrix Assisted Laser Desorption Ionization; Da, Dalton; m/z, Mass to charge ratio; TOF, Time of Flight; NCBI, National Center for Biotechnology Information.

1. Introduction

Proteomics is the study of the proteome, the entirety of proteins, their spatial and temporal expression patterns, their modifications, interactions, and of course their functions. To elucidate the proteome of any higher eukaryote is currently a futile endeavor but subsets thereof can be investigated. Mass spectrometry (MS) has become the tool of choice in proteomics [1] and is used to establish protein sequence, quantity, and modification.

The output of any mass spectrometer is a list of mass to charge ratios (m/z) of the peptides in a sample. To derive further information additional stages of MS can be employed following a fragmentation of the peptide precursor for which many methods are available [2]. Tandem-MS (MS/MS, MS^2) spectra contain a list of m/z values of the fragmented peptide.

There are basically two ways to assign a peptide sequence to an MS/MS spectrum from an unknown precursor. One method, termed database search, depends on the availability of a database of either sequences or reference tandem-MS spectra [3]. In contrast to that, *de novo* sequencing derives the sequence solely form the MS/MS spectrum [2]. For both database search and *de novo* sequencing many algorithms have been developed for computational analysis of MS^2 spectra.

***Corresponding author:** Jens Allmer, Phone Number.: 00902327507517, Fax Number: 00902327507509, E -mail address: jens@allmer.de, Postal Address: Assoc. Prof. Dr. Jens Allmer, Molecular Biology and Genetics, Izmir Institute of Technology, Gulbahce Campus, Urla, Izmir, Turkey.

An abundance of different mass spectrometers are now available, which can further be coupled with a large number of fragmentation methods. This leads to a large number of possible measurement methodologies. According to the 'no free lunch' theorem, there is no one algorithm which is best for all instances of a problem [4,5] let alone for all problems posed by varying combinations of mass spectrometers and fragmentation methods. This seems to be a problem widely ignored in biological sciences where a tool if it works on a small number of instances of the general problem will be quickly adopted to solve even problems well beyond its domain [6]. In mass spectrometry this can be exemplified by the ubiquitous usage of Mascot [7] for many types of MS measurements while it was initially intended to be used for peptide mass fingerprinting on matrix assisted laser desorption ionization - time of flight (MALDI-TOF) mass spectrometers. Admittedly, Mascot has since been extended to include fragmentation spectra but it is still targeted towards MALDI. It seems clear that any of the algorithms in database search or *de novo* sequencing perform differently on diverse data.

Unfortunately, it is unclear which algorithm is best for which combination of mass spectrometer and fragmentation method. Some studies have investigated the performance of database search algorithms [3,8–10] and other studies the accuracy of *de novo* sequencing algorithms [11–13]. A mere review of the relevant publications is in this case not possible since all new algorithms are usually developed on different mass spectrometric data sets. Therefore, it is essential to employ the algorithms to be compared on the same dataset before comparing their performance.

Only few benchmark datasets have been published in mass spectrometry-based proteomics which would allow such a comparison. Whether these can truly be called benchmark datasets will be investigated below (Section 3), but before that a general description of a proper benchmark dataset will be given (Section 2).

In order to further the field of mass spectrometry and to develop useful algorithms, this letter is also a call to action. Anyone having access to synthetic peptides should participate in the development of a first benchmark dataset and provide a few hundred spectra per synthetic peptide so that a comprehensive benchmark dataset, based in ground truth, can be created.

2. Benchmark Data

In the widest sense of the word a benchmark is a standardized performance test. In mass spectrometry-based proteomics a benchmark would thus measure the performance of algorithms to assign a sequence to an MS/MS spectrum. A benchmark dataset for mass spectrometry-based proteomics must thus consist of MS/MS spectra and their correct sequence annotation. Additionally, the mass spectrometer used, the fragmentation method, and the measurement settings should be specified.

Aniba and colleagues defined six measures for well-constructed benchmark datasets [14] which will be discussed in respect to mass spectrometry-based proteomics in the following.

2.1. *Relevance*

In mass spectrometry-based proteomics and in accordance with the 'no free lunch' theorem, a benchmark dataset could target a combination of a particular mass spectrometer and fragmentation method. This platform should be used to generate enough spectra from enough different peptides so that the scope of the benchmark dataset can be fulfilled. In practice that means that if the benchmark dataset is targeted to test the performance of database search algorithms, it should consist of spectra similar to the ones that may be expected in experimental studies. For mass spectrometry this is difficult since the peptide sequence seems to have a strong influence on fragmentation [15] and therefore the resulting spectrum. Therefore, measurements from all MS platforms are needed.

2.2. *Solvability*

The benchmark dataset needs to be solvable. It should be neither too hard nor too easy for existing algorithms so that the benchmark can be used to differentiate performance among algorithms. In practice this disqualifies all existing benchmark datasets since all of them present the same "chicken or egg" problem: The MS/MS spectra are usually derived from a protein digest and are then assigned sequences using a database search algorithm. Unfortunately, in many cases algorithms do not agree on the sequence and thus the correctness of the sequence cannot be guaranteed. Since in existing datasets this is unaccounted for, they are not correctly solvable as they are not correct in themselves. Hence, current algorithms are presented with an insurmountable challenge.

2.3. *Scalability*

A benchmark dataset should present problems at various level of difficulty so that it can scale with the maturity of algorithms. Current datasets might contain examples of different difficulty, but they are not properly annotated and thus not applicable for benchmarking. In mass spectrometry, datasets are exceedingly difficult to solve and more simple datasets like repeated measurements of synthetic peptides are needed to present less perplexing problems and establish the accuracy of the foundation of more advanced methodologies. Later, problems like larger tolerances in the m/z measurement, increasing noise level (i.e.: unexplained peaks from currently poorly understood fragmentation pathways such as sequence scrambling [16], and precursor ions of higher charge can be addressed. Peptides, which lead to fewer fragments than expected, can present a difficult challenge for algorithms following successes on simple benchmark datasets. Even more difficulty can be created by measuring co-eluting and co-fragmenting peptides, somewhat rare, but yet significant problems.

2.4. *Accessibility*

The benchmark dataset needs to be accessible so that algorithm developers or users can benchmark the tools they develop or want to use. Mass spectrometric data may have an intrinsic problem since proper benchmark datasets will likely have large file sizes (several gigabytes) which may be difficult to host and/or transfer. Nonetheless, several platforms for data sharing have been established (see Section 3).

2.5. *Independence*

Aniba and colleagues [14] make a very important point: “The methods or approaches to be evaluated should not be used to construct the gold standard tests. Otherwise, the developers could be accused of ‘cheating’, i.e. designing the benchmark to suit the software”. This seems to be obvious but many new algorithms in mass spectrometry-based proteomics are published alongside with their own datasets. This is obviously due to the fact that no proper benchmark dataset is available and should not be used to accuse developers of cheating in this case.

2.6. *Evolution*

The benchmark dataset needs to be modified constantly to prevent researchers from optimizing their algorithms to solve it. New suitable datasets should thus be published frequently to test the performance of existing algorithms. The need for scalability must be taken into account and subsequent datasets should be gradually more challenging.

Currently, none of the available mass spectrometric datasets adhere to more than one of these six requirements. Nonetheless, they are being used as benchmark datasets with all the negative consequences that entail such as low or unknown accuracy of commonly used algorithms.

3. Mass Spectral Data

Although it is not mandatory for most journals publishing in the area of proteomics, some journals and funders make it mandatory that raw data be made publicly available [17]. Currently, not enough is being done to ensure that raw data is made available and a suitable incentive for researchers to comply is yet to be found [18]. Nonetheless, an abundance of mass spectrometric measurements have been made available in a number of public repositories. The major repositories, in no particular order, are Global Proteome Machine Database [19], PeptideAtlas [20], Proteomics IDEntifications Database [21], Proteome Commons’ Tranche (https://proteomecommons.org/tranche/), and NCBI’s Peptidome [22]. Other smaller or more targeted collections are also available and have been reviewed in Mead and colleagues [23] and Riffle and Eng [24].

Proteomics repositories provide large amounts of raw mass spectrometric data which may also be annotated through database search and may be useful for research [17]. However, there is a hidden “chicken or egg” problem present. MS/MS spectra are usually identified using the database search engine of choice of the laboratory that made the measurements. And these assignments are then used to train new algorithms. An example for this is the benchmark dataset created by Keller and colleagues who used Sequest [25] to assign a sequence to the tandem-MS spectra [26]. Later it was shown that the data set contains additional possible assignments [27] which were not given in the initial publication. It is also likely that many assignments for the Keller et al. dataset are wrong since we were not able to reproduce them with other database search engines or *de novo* sequencing tools (data not shown). This leads to the serious problem that new algorithms are trained on a dataset with assignments of another algorithm. Thus all new trained algorithms will likely duplicate errors done by the algorithm used to assign peptides to spectra during the creation of the dataset.

The field of mass spectrometry-based proteomics is large and interfaces with many instruments other than just mass spectrometers. This fact is mirrored in the approaches used to develop benchmark datasets. For example the Keller et al. dataset closely mirrors standard high throughput studies. Two more recent datasets do the same and although they are more elaborate still present the same “chicken or egg” problem. Wessels et al. present a very comprehensive dataset based on the *Escherichia coli* proteome with additional spiked in known protein digests as a real life challenge [28]. Beasley-Green and colleagues also chose a model organism (*Saccharomyces cerevisiae*) to design their dataset [29]. Another approach for designing benchmark datasets is based on simulation [30], but this approach, while offering a ground truth, is synthetic and should itself be benchmarked on real data. All mentioned datasets aim to benchmark the overall process and neglect the fact that each individual process should be properly benchmarked before integration testing can be performed. Therefore it is necessary to develop more targeted and well-designed benchmark datasets to prove the effectiveness of all modules of the overall mass spectrometry-based proteomics workflow.

4. Call for Action

As detailed above, there is no publicly available dataset which is solvable and thus at least two of the measures for well-constructed benchmark datasets are violated. It is the aim of this letter to engage the mass spectrometric community in creating a compliant benchmark dataset.

As current datasets are not solvable due to the “chicken or egg” problem, it is necessary to assure the assigned sequence by a different means than any of the current database search or *de novo* sequencing algorithms. The simplest way that the sequence assignment can be guaranteed is by directly injecting/spotting pure synthetic peptides. The resulting dataset will be solvable in theory and thus would enable true benchmarking of current algorithms and would enable developers to benchmark their new algorithms.

This letter urges anyone with access to synthetic peptides and mass spectrometers to measure them in the following way:

1. Measure the synthetic peptide at high concentration (10-50 MS/MS spectra)
2. Decrease the concentration (e.g.: lower the flow rate) and measure 10-50 MS/MS spectra
3. Keep decreasing the concentration and measure 10-50 MS/MS spectra
4. Stop measuring when the signal is disappearing in the noise.

Please submit the measurements for individual synthetic peptides in mzXML [31] or mzML [32,33] file format to the author. A comprehensive dataset will be created and deposited in Proteomics IDEntifications Database [21] and NCBI's Peptidome [22]. For this dataset an additional website will be created, crediting anyone submitting data, making available the data, and providing additional information about the dataset. A preliminary version of this website can be reached at http://msbenchmark.biolnk.com.

The dataset is relevant as MS/MS spectra are measured as they would be measured in current experimental procedures although the fact that usually mixtures are investigated is ignored. Additional problem due to liquid chromatography are also ignored. It is the aim to have undisturbed MS/MS measurements of known precursors and varying quality that is solvable in the domain of assigning sequence to MS/MS spectra. As different spectral qualities are created it is to some degree scalable. All benchmark datasets created will be accessible but their relevance will decrease with time when more challenging datasets will be prepared. Thus the dataset will evolve and make optimization targeting the dataset only possible for older benchmarks. Finally, as this is a community effort, the independence of the data is guaranteed. The resulting dataset thus adheres to all features of proper benchmark datasets.

5. Community Benefits

Some of the benefits that the mass spectrometric community can gain from proper benchmark datasets are quite obvious. For experimentalists it would be important to know which algorithms are best for their mass spectrometer and fragmentation method combination. There is no silver bullet among algorithms so no one algorithm can perform best on all measurement platforms. Currently, it is not possible to compare algorithms but given a comparison the best suited computational analysis strategy can be pursued. This in turn leads to more accurate data entering public repositories, again benefiting experimentalists even outside of the field of mass spectrometry-based proteomics. It has also been pointed out by Noble and MacCoss that a critical assessment for algorithms in the field is lacking [34] which further underlines the issue raised above.

Clearly, developers of new database search or *de novo* sequencing algorithms will be enabled to find out for which platform their algorithm is best suited and how it compares to the performance of other algorithms on a publicly available well annotated and solvable benchmark dataset presented in a standard format.

The benchmark dataset is not limited to the benchmarking of algorithms; it can also be used in novel ways that cannot be foreseen now. However, some additional benefits are for instance the ability to employ data mining on the benchmark dataset to learn general parameters about for example peptide fragmentation [24]. A large number of synthetic peptides measured with a variety of mass spectrometric platforms will enable theoretical chemists to develop new fragmentation models. This in turn will enhance peptide identification when these models are integrated into database search or *de novo* sequencing algorithms.

Finally, the associated web page which credits laboratories, researchers, and their measurements may foster the exchange of measurements or samples within the resulting community (http://www.biolnk.com/msbenchmark).

6. Concluding Remarks

This letter was meant to prove that there is a current lack of benchmark data for assigning sequence to MS/MS spectra. This influences the speed of development of the field of mass spectrometry-based proteomics. Furthermore, it forces experimentalists to work with computational analysis tools of unknown accuracy. The benefits of making benchmark data available have been briefly mentioned. It will help increase the accuracy of sequence assignments and in turn the accuracy of conclusions from experimental data in literature and in public databases.

This letter is a call for the submission of MS/MS measurements of synthetic peptides and intends to create an initiative to develop a first proper benchmark for the field of mass spectrometry-based proteomics.

Acknowledgements

This work was in part supported by an award received from the Turkish Academy of Sciences (TÜBA, www.tuba.gov.t).

References

1. M. Mann, R.C.C. Hendrickson, A. Pandey, Annual Review of Biochemistry 70 (2001) 437–73. DOI: 10.1146/annurev.biochem.70.1.437
2. J. Allmer, Expert Review of Proteomics 8 (2011) 645–57.

DOI: 10.1586/epr.11.54

3. E. Kapp, F. Schütz, Current Protocols in Protein Science / Editorial Board, John E. Coligan ... [et Al.] Chapter 25 (2007) Unit25.2. DOI: 10.1002/0471140864.ps2502s49

4. D.H. Wolpert, W.G. Macready, IEEE Transactions on Evolutionary Computation 1 (1995) 67–82.

5. H.P. Schwefel, in: T. Back, D.B. Fogel, Z. Michalevicz (Eds.), Evolutionary Computation 1 Bacis Algorithms and Operators, Institute of Physics Publishing, Bristol and Philadelphia, 2000, pp. 20–22.

6. J. Brownlee, On Biologically Inspired Computation, 2005.

7. D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell, Electrophoresis 20 (1999) 3551–67. DOI: 10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2

8. E.A. Kapp, F. Schutz, L.M. Connolly, J.A. Chakel, J.E. Meza, C.A. Miller, D. Fenyo, J.K. Eng, J.N. Adkins, G.S. Omenn, R.J. Simpson, Proteomics 5 (2005) 3475–90.

9. D.C. Chamrad, G. Korting, K. Stuhler, H.E. Meyer, J. Klose, M. Bluggel, Proteomics 4 (2004) 619–28.

10. I. Shadforth, D. Crowther, C. Bessant, Proteomics 5 (2005) 4082–95. DOI: 10.1002/pmic.200402091

11. S. Bringans, T.S. Kendrick, J. Lui, R. Lipscombe, Rapid Communications in Mass Spectrometry 22 (2008) 3450–3454. DOI: 10.1002/rcm.3752

12. S. Pevtsov, I. Fedulova, H. Mirzaei, C. Buck, X. Zhang, J Proteome Res 5 (2006) 3018–28.

13. E. Pitzer, A. Masselot, J. Colinge, Proteomics 7 (2007) 3051–4. DOI: 10.1002/pmic.200700224

14. M.R. Aniba, O. Poch, J.D. Thompson, Nucleic Acids Research 38 (2010) 7353–63. DOI: 10.1093/nar/gkq625

15. A.R. Dongre, J.L. Jones, A. Somogyi, V.H. Wysocki, Journal of the American Chemical Society 118 (1995) 8365–8374.

16. A.E. Atik, T. Yalcin, Journal of the American Society for Mass Spectrometry 22 (2011) 38–48. DOI: 10.1007/s13361-010-0018-3

17. J.A. Vizcaíno, J.M. Foster, L. Martens, Journal of Proteomics 73 (2010) 2136–46. DOI: 10.1016/j.jprot.2010.06.008

18. Editorial, Nature Biotechnology 27 (2009) 579. DOI: 10.1038/nbt0709-579

19. R. Craig, J.P. Cortens, R.C. Beavis, Journal of Proteome Research 3 (2004) 1234–42. DOI: 10.1021/pr049882h

20. E.W. Deutsch, Methods in Molecular Biology (Clifton, N.J.) 604 (2010) 285–96. DOI: 10.1007/978-1-60761-444-9_19

21. J.A. Vizcaíno, R. Côté, F. Reisinger, H. Barsnes, J.M. Foster, J. Rameseder, H. Hermjakob, L. Martens, Nucleic Acids Research 38 (2010) D736–42. DOI: 10.1093/nar/gkp964

22. L. Ji, T. Barrett, O. Ayanbule, D.B. Troup, D. Rudnev, R.N. Muertter, M. Tomashevsky, A. Soboleva, D.J. Slotta, Nucleic Acids Research 38 (2010) D731–5. DOI: 10.1093/nar/gkp1047

23. J.A. Mead, L. Bianco, C. Bessant, Proteomics 9 (2009) 861–81. DOI: 10.1002/pmic.200800553

24. M. Riffle, J.K. Eng, Proteomics 9 (2009) 4653–63. DOI: 10.1002/pmic.200900216

25. J. Eng, A.L. McCormack, J.R. Yates, J Am Soc Mass Spectrom 5 (1994) 976–989.

26. A. Keller, S. Purvine, A.I. Nesvizhskii, S. Stolyar, D.R. Goodlett, E. Kolker, Omics 6 (2002) 207–12.

27. D. Tsur, S. Tanner, E. Zandi, V. Bafna, P.A.P. a Pevzner, Nature Biotechnology 23 (2005) 1562–7. DOI: 10.1038/nbt1168

28. H.J.C.T. Wessels, T.G. Bloemberg, M. van Dael, R. Wehrens, L.M.C. Buydens, L.P. van den Heuvel, J. Gloerich, Proteomics 12 (2012) 2276–81. DOI: 10.1002/pmic.201100284

29. A. Beasley-Green, D. Bunk, P. Rudnick, L. Kilpatrick, K. Phinney, Proteomics 12 (2012) 923–31. DOI: 10.1002/pmic.201100522

30. C. Bielow, S. Aiche, S. Andreotti, K. Reinert, Journal of Proteome Research 10 (2011) 2922–9. DOI: 10.1021/pr200155f

31. P.G.A. Pedrioli, J.K. Eng, R. Hubley, M. Vogelzang, E.W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R.H. Angeletti, R. Apweiler, K. Cheung, C.E. Costello, H. Hermjakob, S. Huang, R.K. Julian, E. Kapp, M.E. McComb, S.G. Oliver, G. Omenn, N.W. Paton, R. Simpson, R. Smith, C.F. Taylor, W. Zhu, R. Aebersold, Nature Biotechnology 22 (2004) 1459–66.

32. E.W. Deutsch, Methods in Molecular Biology (Clifton, N.J.) 604 (2010) 319–331. DOI: 10.1007/978-1-60761-444-9_22

33. L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W.H. Tang, A. Römpp, S. Neumann, A.D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P.-A. Binz, E.W. Deutsch, Molecular & Cellular Proteomics: MCP 10 (2011) R110.000133. DOI: 10.1074/mcp.R110.000133

34. W.S. Noble, M.J. MacCoss, PLoS Computational Biology 8 (2012) e1002296. DOI: 10.1371/journal.pcbi.1002296