Published in final edited form as:

J Mass Spectrom. 2010 October ; 45(10): 1095–1097. doi:10.1002/jms.1786.

Protein PTMs: post-translational modifications or pesky trouble makers?

Ralph A. Bradshaw, Katalin F. Medzihradszky, and Robert J. Chalkley*

NCRR Mass Spectrometry Facility, Department of Pharmaceutical Chemistry, UCSF, San Francisco, CA 94158, USA

Abstract

Analysis of protein post-translational modifications using mass spectrometry is an intensive area of proteomic research. This perspective discusses the current state of the field with respect to what can be achieved, the challenges encountered, most notably with modification site assignment, the reliability of the published results, consequences of unreliable data and what is needed to be done to improve the situation.

Keywords

PTMs; mass spectrometry; database searching; modification site assignment; phosphorylation

Proteins perform a vast array of biological functions in living organisms. Although the 20 amino acids genetically encoded possess a wide range of chemical properties that are sufficient for some of these activities, such as proper protein folding, many more roles, in particular those connected to control and regulation require that they be supplemented with an extensive array of covalent modifications that are introduced after translation is complete. [Some modifications are actually inserted before protein synthesis is complete; for the purpose of this article both co- and post-translational events will be considered as one group, denoted collectively as post-translational modifications (PTMs).] These modifications can occur either on the residue side chains or the assembled polypeptide backbone, the latter generally being manifested as proteolytic events, and several hundred types of modifications have been reported [this list is considerably larger when modifications introduced in the laboratory are included. In addition, there are a few modifications that also can occur artifactually through sample handling and sometimes it is not possible to determine if the modification occurred in the cell (or tissue) or after isolation. Methionine sulfoxide formation and asparagine/glutamine hydrolysis are illustrative].^[1] Indeed, it is now generally held that there are very few proteins that do not undergo some kind of PTM and most incur multiple alterations during their lifetimes. The net effect is to enormously expand the number of different proteins in an isolate, which in turn, compounds the task of not only identifying the nature of the modification but also where in the sequence of the protein/peptide it has occurred. Although the importance of PTMs has been appreciated for a long time, the magnitude, both qualitatively and quantitatively, of this biological activity has been strikingly revealed by high throughput proteomic analyses based on mass spectrometric technology.^[2] Indeed, by employing selective enrichment strategies, it is not uncommon to detect, identify and record 100s to 1000s of modifications in a single experiment.

Phosphorylation of serine, threonine and tyrosine, intracellular events associated with a plethora of cell functions, and both simple and complex glycosylation that occurs both intra- and extracellularly, are typical but are by no means the only examples.

Proteomic methodologies have also demonstrated how complex the modification patterns of some proteins can be, where many kinds of modifications on the same or multiple sites can occur. Histones, the modification of which helps regulate gene expression, are an excellent example of proteins that undergo extensive regulatory changes, and have been heavily studied using proteomic strategies.^[3,4] Finally, the reversible nature of some PTMs and the sub-stoichiometric occupancy of many sites, both well exemplified by, but not limited to, phosphorylation, add substantially to the complexity of samples. All together, the variety and extent of PTMs offer far greater challenges to the field of proteomics than just the identification of proteins using peptide mass fingerprinting (PMF) or sequence analyses (MS/MS) does. As experiments directed at determining the identification and localization of PTMs proliferate, it is reasonable to ask just how well this is being accomplished. Is the data solid and reliable, or are we filling the literature (and hence various databases and repositories) with an increasing amount of flawed and/or outright incorrect information? The answer, at least for the moment, seems to contain too much of the latter for comfort.

Although there still may be problems with differentiating between some well known PTMs (such as sulfation and phosphorylation), there is also a significant number of additional masses that have been observed, often repeatedly, in proteomic experiments that have not as yet been clearly characterized. In addition, the ability to reliably and accurately assign the *locations* of PTMs remains a major issue and is the principal problem in published results and the greatest source of error. Generally, mass spectrometric data of a higher quality are required to identify the exact modified residue, i.e. its position in the protein sequence, than is needed to identify either the peptide or the modification present. This often leads to ambiguity if there is more than one potential site present, e.g. there are multiple serine and threonine residues present in a phosphorylated peptide and this situation is exacerbated when there is more than one modification to assign, or when the modification may occur on different residue types. For example, an additional oxygen may be located on a cysteine, histidine, lysine, methionine, proline, tyrosine or tryptophan residue.

Proteomic data analysis used to identify and localize PTMs is carried out by search engines, of which there is a significant variety. However, few are optimized for PTM identification and localization, rather being focused for the most part on peptide identification. Basically there are two major flaws in the current search engine software when being used for PTM assignment: (1) they mostly apply an intensity threshold to the peak list to limit 'noise peaks'; this leads to a higher percentage of peaks being matched but can eliminate peaks that contain vital information for the localization process (this problem is closely related to difficulties of reliable peak-picking); and (2) they do not clearly indicate site assignment reliability when the data are consistent with the site reported, but another site is equally possible, i.e. they ignore the ambiguity issue. In most cases, the 'back up' solution is manual verification and this immediately introduces a wide range of variability in the results. Some researchers with extensive experience and a high proficiency in reading mass spectrometric data will be quite adept at culling the 'wheat from the chaff', but others with less skill and/or fortitude will not be. At best, any analyses exposed to human intervention will automatically have a subjective element introduced, accompanied by an unquantifiable error rate. There are programs specifically designed for modification site assignment that have been developed and as these improve and become more sophisticated, they will provide more consistent results. However, at the moment, they are not better than experienced data analyzers. Hence, there is a push, spearheaded by journal editors and funding agencies, to

make the raw data for these results available for independent assessment of the data supporting a site assignment conclusion.

As the field of proteomics expands, producing ever increasing amounts of data in the literature, the collation of these results into readily useable resources has proceeded apace. These various databases range from curated repositories to personal (laboratory) web sites and offer an equally broad range of reliability. Information is added to many of these sites without regard to origin, meaning that some will be derived from data used to support peer reviewed articles and some will simply be datasets that are literally outputs from software used to analyze mass spectrometric experiments, without any measures of reliability attached or any attempt at independent verification of the results. This bioinformatic challenge is of some concern, as downstream analyses and interrogations of these data generally treat the material without regard to origin or previous evaluation. Subsequent iterations and reanalyses will further obscure the origin (and thus the reliability) of the information. Thus, unreliable data will be treated the same as correct data that have been seamlessly integrated through the various treatments. There is only one way to prevent this issue expanding and that is to try to guarantee (as much as one can) the correctness of the data in the first place, before it is made open to the public. For peptide/protein identification results some repositories choose to reinterpret all data submitted to implement a level of quality control. However, there is no equivalent mechanism employed for PTM assignments; indeed there is currently no accurate metric that can report the reliability of a site assignment.

How serious is the error rate in PTM localizations and how realistic is it to assume that this will notably improve in the foreseeable future? The first question is difficult to estimate accurately, but there is reason to believe that it can be quite extensive. In a recent publically conducted test by the Proteome Informatics Research Group (iPRG) of the Association of Biomedical Research Facilities (ABRF), a set of phosphopeptide data derived from a large-scale LC/MS/MS experiment was given out and laboratories were asked to analyze it with respect to phosphopeptide identification and PTM localization. The results, presented at the recent ABRF meeting in Sacramento, CA, (March, 2010) showed a broad spectrum of interpretations. When multiple research groups claimed site assignment was possible for a particular spectrum there was a very high degree of consensus about the site of modification, but there were many modification sites reported by only one of the many participating groups; i.e. there was a much lower agreement about in which spectra a phosphorylation site could be unambiguously assigned from the data available [A report of the iPRG group study presented in Sacramento by Karl Clauser of the Broad Institute, Cambridge, MA is in preparation for submission to a journal and the results are available through the ABRF web-site at:

<http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm>].

While this cannot be directly translated into an error percentage in the locations of PTMs in general, it supports the growing concern in the community that this rate may be disturbingly high, at least in some circumstances.

The question whether this situation can be significantly improved in the near future is equally hard to predict. The onus is on journal editors and database curators to apply as stringent criteria as is feasible and to set as high standards as is reasonable to data of this sort. One obvious solution that can be implemented immediately is the requirement that the raw mass spectrometric data behind any dataset be made available as open data, placed in a publicly available repository, where it can be inspected, interrogated and reanalyzed at will. This will allow ready access to the information behind localization assignments (as well as protein identifications) and thus a means for resolving discrepancies and differences between reported results from different research groups. As more reliable and powerful

bioinformatic tools are developed it will allow the reanalysis of data to improve the interpretation of previously reported studies. This will hopefully have the effect of purging, if slowly, the misinformation already in the literature. A not insignificant fringe benefit of this step would be the added information gained from reanalyses of existing data asking different questions than the original experimenters who generated the data set, leading to new insights for only marginally increased costs. In a time when biomedical research support worldwide is feeling a significant pinch, this 'better bang for the buck' is not necessarily trivial.

Thus, one can expect (perhaps hope?) that with improved technology and software, the increased deposition of open raw data and greater diligence on the part of editors, curators and authors, or at least attention to the problem [several conferences/workshops dealing wholly or in part with these issues have occurred or are in the planning stages: a workshop at the 58th ASMS Conference, May 24, Salt Lake City, UT; a workshop sponsored by MCP (by invitation only) in Boston, July 19, 2010, an NCI sponsored workshop on raw data release (as a satellite meeting to the HUPO Congress, September 19, 2010) and an ASBMB conference at Granlibakken, CA, October 21–24, 2010], that this extremely important part of biological research in general, and proteomics in particular, will move forward into an era of better and more accurate PTM assignments, i.e. the pesky trouble makers will become pertinent trustworthy material.

References

1. <http://www.unimod.org>.
2. Nielsen ML, Savitski MM, Zubarev RA. Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. *Mol Cell Proteomics*. 2006; 5:2384. [PubMed: 17015437]
3. Burlingame AL, Zhang X, Chalkley RJ. Mass spectrometric analysis of histone posttranslational modifications. *Methods*. 2005; 36:383. [PubMed: 16112065]
4. Garcia BA, Shabanowitz J, Hunt DF. Characterization of histones and their post-translational modifications by mass spectrometry. *Curr Opin Chem Biol*. 2007; 11:66. Epub 2006 Dec 6. Review. [PubMed: 17157550]