Beyond the model organism – Mechanistic analysis of protein post-translational modifications is ready for all challengers

Benjamin C. Orsburn*
The Department of Pharmacology and Molecular Sciences
The Johns Hopkins University School of Medicine, Baltimore, MD, USA, 21205
*Corresponding: borsbur1@jhmi.edu

Abstract

A historic challenge for shotgun proteomics has been the requirement for high quality, simple and nonredundant curated protein sequences in small .fasta text files. Due to the intrinsic informatic challenges and time required to assemble these files, proteomics has struggled to expand beyond the confines of a few model organisms. When considering post-translational modifications that may or may not be present on a specific peptide sequence, these factors inevitably compound. A study on how mangos continue to ripen on the shelf may not be the first thing you'd think of as proof of a scientific discipline shedding historic limitations. However, Bautiste-Valle *et al.,* may be just that. These authors present a quantitative comparison of both peptide and glycopeptide alterations through the complexity of the fruit ripening process and in this we see the present state of a field that no longer needs to wait on genomics to obtain deep mechanistic insights.

Keywords

Post-translational modifications, non-model organisms, glycosylation, multi-omics

Life on our planet can be described, in nearly all cases, as bags made of lipids that hold in water and assorted useful classes of molecules. One of the useful classes of molecules are oligonucleotides which, in the form of DNA, serve as the blueprint for taking a bunch of other useful molecules - amino acids - and eventually making them into proteins. The proteins are then stuck with the unenviable task of doing basically everything else When the proteins are doing all the work and the system is self-sustaining, that lipid bag of molecules does enough things to be considered a living organism. Obviously, it is more complex than that in most cases, but this is the very root of it. Now, you might think that because we have the ability today to figure out the sequence of billions of nucleotides with something approaching accuracy in any organism we'd know all of those DNA sequences for all of the organisms. Despite projects with amazing marketing budgets and lofty titles that have been going on since Michael Jordan was still a Chicago Bull, this clearly isn't the case. The genetic sequences from projects as seemingly ancient as the 1,000 Genome Project simply haven't trickled down to the broader scientific community in usable forms, such as the nice curated .fasta files we need for proteomics. This gets even worse as you step outside of a few model organisms, even when these sequences are of critical importance.

If you are reading this on a more or less average day over at least the last 100,000 years, malaria will kill more people today than anything else. A lot of scientists thought they'd fix that by sequencing the malaria parasites and launched the PlasmoDB in 2001.[1] You'd think those sequences would be pretty good, right? They aren't. Well, I'm spoiled, but they aren't even close to as nice as the crisp, clean human FASTA libraries you find in SwissProt.[2]

And quality in input sequences drops rapidly as you move away from the two or three most studied organisms. And that is a big problem for proteomics technologies of any sort, including the mass spectrometry sorts. In shotgun proteomics we typically chop our proteins up into little bits with enzymes that let us know what amino acid should be on one end of each of those little chunks. Then we take those little chunks, turn them into gas and accelerate them or shake them or shake them with nasty chemicals and shake them some more to break them into even smaller pieces. The resulting fragments can then be matched against peptide sequences that we believe to be present. You fragmented something that has the same mass as this? Great! Do the observed fragments line up with what you'd expect if you did this terrible thing to that peptide? Even better! You've sequenced this peptide successfully.

Things get a lot more challenging for everyone because a lot of proteins don't actually do anything until another protein chemically modifies them. In classical examples, Protein A is in an open configuration until it is phosphorylated by Protein B. When that phosphate goes on, Protein A snaps closed like a mouse trap and does some magic, like forcing a chloride ion into a room where it is absolutely convinced is too crowded with other soluble ions. In the lab we now have to consider two possibilities, that our experimentally observed fragment sequence could be our nice theoretical sequence or a sequence with a phosphate on it somewhere. We've now made the work of our computers twice as hard, if not more. Mass spectrometrists love the phosphorylation example because we're generally pretty good at identifying those because they break in exactly one place, have one mass, and are typically found on 2 or 3 amino acids main amino acids. We've generally been a lot less fond of things like glycosylation because that PTM may have hundreds of different masses with the uniform annoyance of fragmenting in multiple places. In addition, the same PTM mass of a trisaccharide, for example, may have completely different biological functions depending on the order those tri (three) saccharides are assembled in.

That was a lot of background for why you don't see a lot of glycoproteomics in nonmodel organisms. When your database quality isn't perfect, do you really want to add that uncertainty to the uncertainty of searching for thousands of potentially present glycan chains? Is that something you want to take on? We tried it once and gave up on being quantitative. We just counted how many spectra we saw with oxonium ions in them and called it a day.[3]

And that brings us to a study of what happens to a mango as it is sitting on a shelf. Bautiste-Valle *et al.,* went all the way through. What should catch your attention first is the quantitative proteomics comparison, in which ripening tissue was studied with label free proteomics and using multiplexed reagents with the SPS MS3 method. Over 1,200 proteins were identified as significantly changing, with around 250 of those detected by both methods. To identify 1,200 proteins as altered should suggest to you the degree of coverage they obtained in these fruits. Where this work makes the big step, however, is in going into quantitative glycoproteomics. The thoroughness of this part of the study is truly impressive. Glycopeptides are quantified by label free methods and again by SPS MS3 based multiplexing, supplemented with electron transfer dissociation (ETD) to aid in both identification and localization.[4] The complete workflow specifically shines in opening up the ability to identify N-linked glycopeptides which, as these authors note, has been a considerable challenge in applying glycoproteomics to plant biology.

This isn't the first study to demonstrate this level of quantitative proteomic coverage in non-model organisms,[5] but I dare you to find a glycoproteomic study that has succeeded in generating this level of depth and overall insight.

While these researchers clearly sought out to understand a particular agricultural phenomenon, this is a sign of what is to come. The big genomics projects and consortia may never provide us with the beautifully curated protein databases that we have for our few model organisms. Maybe, just maybe, we can stop waiting for them to.

Conflict of interest statement

The author declares no relevant conflicts of interest

References

(1) Collaborative, T. P. G. D. PlasmoDB: An Integrative Database of the Plasmodium Falciparum Genome. Tools for Accessing and Analyzing Finished and Unfinished Sequence Data. *Nucleic Acids Res* **2001**, *29* (1), 66–69.

(2) Bateman, A.; Martin, M. J.; O'Donovan, C.; Magrane, M.; Apweiler, R.; Alpi, E.; Antunes, R.; Arganiska, J.; Bely, B.; Bingley, M.; Bonilla, C.; Britto, R.; Bursteinas, B.; Chavali, G.; Cibrian-Uhalte, E.; Da Silva, A.; De Giorgi, M.; Dogan, T.; Fazzini, F.; Gane, P.; Castro, L. G.; Garmiri, P.; Hatton-Ellis, E.; Hieta, R.; Huntley, R.; Legge, D.; Liu, W.; Luo, J.; Macdougall, A.; Mutowo, P.; Nightingale, A.; Orchard, S.; Pichler, K.; Poggioli, D.; Pundir, S.; Pureza, L.; Qi, G.; Rosanoff, S.; Saidi, R.; Sawford, T.; Shypitsyna, A.; Turner, E.; Volynkin, V.; Wardell, T.; Watkins, X.; Zellner, H.; Cowley, A.; Figueira, L.; Li, W.; McWilliam, H.; Lopez, R.; Xenarios, I.; Bougueleret, L.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M. C.; Boeckmann, B.; Bolleman, J.; Boutet, E.; Breuza, L.; Casal-Casas, C.; De Castro, E.; Coudert, E.; Cuche, B.;

Doche, M.; Dornevil, D.; Duvaud, S.; Estreicher, A.; Famiglietti, L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Jungo, F.; Keller, G.; Lara, V.; Lemercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T.; Nouspikel, N.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pozzato, M.; Pruess, M.; Rivoire, C.; Roechert, B.; Schneider, M.; Sigrist, C.; Sonesson, K.; Staehli, S.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Veuthey, A. L.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Garavelli, J. S.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Suzek, B. E.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Yeh, L. S.; Yerramalla, M. S.; Zhang, J. UniProt: A Hub for Protein Information. *Nucleic Acids Res* **2015**.

(3) Jenkins, C.; Orsburn, B. The Cannabis Proteome Draft Map Project. *Int J Mol Sci* **2020**.

(4) Kim, S.; Mischerikow, N.; Bandeira, N.; Navarro, J. D.; Wich, L.; Mohammed, S.; Heck, A. J. R.; Pevzner, P. A. The Generating Function of CID, ETD, and CID/ETD Pairs of Tandem Mass Spectra: Applications to Database Search. *Molecular and Cellular Proteomics* **2010**.

(5) Heck, M.; Neely, B. A. Proteomics in Non-Model Organisms: A New Analytical Frontier. *J Proteome Res* **2020**, *19* (9), 3595–3606.