Roundtable Session 1 - Table 9 - Best Practices for Predicting, Elucidating and Monitoring Hotspots by MS

Facilitator: Chris Sauer*, Johnson & Johnson*
Scribe: Michelle English, *Genedata*

Abstract:
Protein hotspots continue to be an important facet of biomolecular characterization of protein structure and function. Understanding protein structure and lability is critical in describing sites of aggregation, potential degradation and assessing risk of common PTMs like isomerization, deamidation, glycation, oxidation, etc. This information is necessary for characterizing risks at different stages of drug development, including drug design and engineering, clone selection, and ultimately manufacturing.

Characterization of protein hotspots is a complex problem being tackled by researchers across sectors, combining structural analysis with quantitative risk assessment. Mass spectrometric methods have emerged to meet experimental needs of researchers, offering quantitative, site-specific information with high throughput. These methods have been further improved by advances in hardware, data analysis, and modeling.

This roundtable discussion will be focused on the current state of hotspot analysis and advances in the field. We will discuss the changing role of mass spectrometry in conjunction with novel sample preparation approaches, computational methods, and in silico modeling to better characterize hotspots. Lastly, we will discuss what areas still require development and how the best approaches for hotspot analysis are changing across industries.

Discussion Questions:
What are common practices to characterize hotspots?

- Late development/scale up groups largely rely on peptide mapping. For example, using known peptides as a control. For example, they compare detected CDR oxidation to a known FC oxidation to understand susceptibility. Don't find that they need to focus on prediction as they are so late stage.
- In earlier stages most companies have developed in house databases of sequences and observed hotspots, some even with biophysical data. This data is then used to power custom solutions that vary from Excel database to a ML solution that might even attempt to predict 3-D structure

How are these data used?

- Largely as a guide to focus on where to look for potential problems

How are the in-house models fed?

- All teams seem to feed the models with data collected over time from characterization groups and take that data as the basis for their models. In at least one organization the in-house data sciences team continuous collects as much analytical data as they can and continuously improve their modeling capabilities.

- Other organizations are also beginning to explore in silico or ML or AI structure prediction, but suffer from the ability to get robust training datasets

How do you make the curated data meaningful?

- This really depends on your ability to collect and organize the data in meaningful ways. For example, how to do you relate the MS or other assay data to a structure or liability.
- Largely this will rely on MAM or Dev information
- Noted that in late-stage groups like late-stage cell line expressing dev groups, often the sequence liabilities seen as process conditions change are not the same as those highlighted during earlier development stages. The more abundant modification may be completely different or when the same may be at very different levels. This means they often need to be concerned with different hotspots and often investigate previously modifications that have not bee previously described for the given molecule.

How do you measure the identified liabilities?

- The primary approach for all groups is peptide mapping
- Discussed the idea of preliminary screening before peptide mapping to identify hotspots. While some might run an intact or subunit, when they are examining hotspots, the method of choice is really peptide mapping
- Also noted this is really more of a development stage activity; not really a concern at clone selection time. Don't really dive in until down to one or two candidates
- Though for specific liabilities like oxidation, some groups may also monitor with HIC
- The group also noted the promise of tools like iCIEF, works well for acidic species, and matches well to peptide mapping results, but will not replace peptide mapping.

Side conversation on impacts of liabilities and impacts beside sequence.

- Does oxidation really show a structure function relationship? Most seemed to say they don't always study the specific effect, instead they consider anything in the CDR as meaningful, while others may check for changes to binding.
    - One study showed that while the local environment mattered more, expression related affects can have an impact.
- Does chemistry beyond the local sequence affect hotspots (eg molecular pI or salt conditions)? Yes, likely it does, but it is even more difficult to model.
- Similarly do liabilities vary with the expression time or purification? Likely yes, but not often explored in the context of hotspots by the participants.

Can you really have a standardized method for hot spot screening?

- It is a good idea and most start with a standard method but often have to adapt to specific molecules.
- Noted that the role of peptide mapping in this case is to prove or disprove what we already suspect, and thus standardized methods are often sufficient to answer this question.

How do you develop a good model?

- Much of this work relies on a good model, but this is difficult to achieve. As stated above a comprehensive library of well-organized data is required
- One suggested that perhaps a third-party proteome wide model would be really beneficial to the community, but acknowledged this would be very difficult to achieve.

What advances in hardware and software are needed?

- Would love to have general industry database and model, but acknowledge that the data sharing component would be challenging
- Might also like external software that could take in an organizations existing data and feed that into a private version of generalized ML/structure information software
- Challenges to this might be solved by moving to standardized data formats or at least allowing exports to standardized data formats.
- Would like to see hotspot prediction software similar to existing O-glycan prediction software.

Vendor feedback discussion

- Really would like to see storage of multiple types of analysis data in a single archive. Often while MS data is key, having orthogonal assay data combined really helps
- It was noted that some organizations are starting to do this in house, but many participants are looking for vendors to provide this type of access.
- Noted that moving in this direction might require much more open formats to allow queries across vendor and data types. It was also mentioned that there is a consortium forming to tackle this issue.

How will AI/ML change the need for mass spec analysis?

- Initially we are all the more essential to help build out the data for models
- As time passes, it may become a greater possibility. However, there is so much variability and modalities continue to evolve that it is difficult to see this being an impact in the short term.
- This also assumes the systems are truly deterministic and not all were sure they are on the whole.
- Nonetheless, the participants seem to feel this is a worthwhile path to follow
- Noted that there are similar (maybe simpler) AI systems that are already proving useful such as RT retention prediction
- Also thought incremental steps could prove very useful. For example, an AI that simply models the difference between the MS2 of a modified and unmodified peptide or one that assess the data and automatically processes it in the best way.