

Computational Proteomics

Rebekah Gundry^{*1}, Lennart Martens^{*2}, and Magnus Palmblad^{*3}

1 University of Nebraska – Omaha, US. rebekah.gundry@unmc.edu

2 Ghent University, BE. lennart.martens@ugent.be

3 Leiden University Medical Center, NL. n.m.palmblad@lumc.nl

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 23301 “Computational Proteomics”. This seminar was built around three topics: the increasingly widespread use, and continuously increasing promise of advanced machine learning approaches in proteomics; the highly exciting, yet fiendishly complicated, field of single cell proteomics, and the development of novel computational methods to analyse the highly challenging data obtained from the glycoproteome. These three topics fuelled three parallel breakout sessions, which ran in parallel at any given time throughout the seminar. A fourth, cross-cutting breakout session was created during the seminar as well, which dealt with the standardisation efforts in proteomics data, and explored the possibilities to upgrade these standards to better cope with the increasing demands being put on the relevant data storage and dissemination formats. This report comprises an Executive Summary of the Dagstuhl Seminar, which describes the overall seminar structure together with the key take-away messages for each of the three topics. This is followed by the abstracts, comprising three introduction talks, one for each topic, which were intended to whet the participants’ appetite for each topic, while also introducing an expert perspective on the current challenges and opportunities in that topic. Along with the topic talks, two ad-hoc talks were presented during the seminar as well, and their abstracts are provided next. Moreover, each breakout session also comes with its own abstract, which provides insights into its discussions and relevant outcomes throughout the seminar.

Seminar July 23–28, 2023 – <https://www.dagstuhl.de/23301>

2012 ACM Subject Classification Applied computing → Bioinformatics

Keywords and phrases bioinformatics, glycoproteomics, machine learning, mass spectrometry, proteomics, single cell proteomics

Digital Object Identifier 10.4230/DagRep.13.7.152

1 Executive Summary

Lennart Martens (Ghent University, BE)

Rebekah Gundry (University of Nebraska – Omaha, US)

Magnus Palmblad (Leiden University Medical Center, NL)

License  Creative Commons BY 4.0 International license
© Lennart Martens, Rebekah Gundry, and Magnus Palmblad

The Dagstuhl Seminar 23301 “Computational Proteomics” was based around three key topics of rapid development in mass spectrometry-based proteomics, which were discussed in-depth in light of their challenges and opportunities. These three topics were: (i) the expanding and highly successful adoption of machine learning (ML) approaches in proteomics; (ii) the varied computational challenges posed by the very recent, but very rapidly evolving field of single cell proteomics; and (iii) the possible paths to adoption of advanced computational

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Computational Proteomics, *Dagstuhl Reports*, Vol. 13, Issue 7, pp. 152–165

Editors: Rebekah Gundry, Lennart Martens, and Magnus Palmblad



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

approaches in the challenging field of glycoproteomics. Each of these topics was introduced by a short lecture, delivered by an expert in the field, and focused on two main goals: (i) to provide an informed opinion of the current state of the field, while highlighting its key challenges; and (ii) to thus entice the participants to contribute their own views on this topic, to help set the agenda for the discussions throughout the remainder of the seminar. Apart from these three invited talks, two ad-hoc talks also emerged during the seminar, and these concerned the specific topic of large scale spectral clustering, and the promise of using the Rust programming language in proteomics applications.

Based on the ideas collected after each of the introductory, topic-specific presentations, a list of discussion points was collated for each topic, and three parallel breakout sessions were then organised in the mornings and afternoons around these discussion points. A final, joint session in the evening of each day served to bring all participants from the different breakout groups together again, and summarized the key points of their respective discussions. Moreover, these joint sessions were also used to update the lists of discussion points for the three topics with any newly emerged points, and to reprioritise discussion points for the next day's breakout sessions.

The Machine Learning Working Group had a lot of topics to explore, mostly focusing on refining current approaches, by, for instance, introducing quality control and explainable AI, as well as setting out new applications for ML in the field. The latter included the possibility of a foundational model, the extended prediction of analyte behaviour in the instrumentation, and the possibility to analyse the resulting models to gain a better understanding of the physico-chemical properties at play in the analytics workflow. Finally, community-building efforts were discussed, and suggestions for improvements of existing initiatives (notably proteomicsML.org), as well for novel community engagements were made.

The Single Cell Proteomics Working Group discussed the applicability of current tools in the new discipline of single cell proteomics. Correspondingly, issues in capacity in current tools also came up, in light of the fast-growing data sizes for single cell experiments, which are currently at hundreds of analytical runs, but likely soon expanding towards thousands of runs; well beyond the capabilities of present-day algorithms. Standardisation of this field, and its metadata, which is even more important given the sheer size and complexity of typical single cell proteomics data sets, was also considered in some detail. This entailed the standardisation of the workflows and algorithm parameters, which are currently very diverse and specific, as well as the standards for data and metadata representation and dissemination.

The Glycomics and Glycoproteomics Working Group saw plentiful opportunities for the field to strengthen their bioinformatics, and put emphasis on adopting machine learning techniques in their field. However, they also saw open challenges regarding the collection and annotation of their data. Some time was also spent on identifying current weaknesses in their field, notably the quantification of glycopeptides, and possible avenues for addressing these.

At the end of the seminar, a critical assessment of the seminar was performed by all participants, highlighting the strengths and improvement points of the overall Seminar organisation, and a list of potential future Dagstuhl Seminar topics was drafted based on the participant's input. The assessment of the Seminar highlighted in particular the extremely fruitful nature of the open and engaging discussions, the unique and highly valuable nature of Schloss Dagstuhl and its unmatched seminars, and the ongoing gratitude of the computational proteomics community for the opportunity to convene in this singular setting. Concerning possible future topics, a plethora of enticing options were put forward, indicating that the field of computational proteomics remains in full expansion and that it continues to brim with both challenges and promise!

2 Table of Contents

Executive Summary

Lennart Martens, Rebekah Gundry, and Magnus Palmblad 152

Overview of Talks

Topic introduction: Machine learning in Proteomics: everything everywhere all at once	
<i>Robbin Bouwmeester</i>	155
Topic introduction: Mass spectrometry-based single-cell proteomics: current challenges	
<i>Laurent Gatto</i>	155
Topic introduction: Glycomics and glycoproteomics – Computational challenges and opportunities	
<i>Sriram Neelamegham</i>	156
Ad-hoc talk: Unsupervised clustering of spectra	
<i>Lukas Käll</i>	156
Ad hoc talk: A community-driven project – Rusteomics	
<i>Dirk Winkelhardt</i>	157

Working groups

Working Group Report: Machine Learning in Proteomics	
<i>Robbin Bouwmeester, Viktoria Dorfer, Laurent Gatto, Arzu Tugce Guler, Tiannan Guo, Michael Hoopmann, Lukas Käll, Ville Koskinen, Lennart Martens, Magnus Palmblad, Tobias Schmidt, Veit Schwämmle, Mathias Wilhelm, and Dirk Winkelhardt</i>	157
Working Group Report: Single Cell Proteomics	
<i>Laurent Gatto, Bernard Delanghe, Melanie Föll, Lukas Käll, Ville Koskinen, Lennart Martens, Sriram Neelamegham, Tobias Schmidt, Veit Schwämmle, Mathias Wilhelm, and Gamze Nur Yapici</i>	160
Working Group Report: Glycosylation and Glycoproteomics	
<i>Rebekah Gundry, Kiyoko Aoki-Kinoshita, Robbin Bouwmeester, Robert Chalkley, Bernard Delanghe, Viktoria Dorfer, Melanie Föll, Arzu Tugce Guler, Catherine Hayes, Michael Hoopmann, Lukas Käll, Ville Koskinen, Karina Martinez, Sriram Neelamegham, Magnus Palmblad, Erdmann Rapp, Tobias Schmidt, Mathias Wilhelm, Bernd Wollscheid, and Gamze Nur Yapici</i>	161

Participants	165
-------------------------------	-----

3 Overview of Talks

3.1 Topic introduction: Machine learning in Proteomics: everything everywhere all at once

Robbin Bouwmeester (Ghent University, BE)

License  Creative Commons BY 4.0 International license
© Robbin Bouwmeester

It can be clear that machine learning has integrated with many (if not outright most) of proteomics data analysis, as it has been a key ingredient in today's proteomics informatics advances. At the same time, this widespread uptake of machine learning in the field can correctly be labelled as only a beginning. Indeed, there are many different new models available in the domain of machine learning, including large language models, transfer learning, and graph convolutional networks, alongside many new applications in proteomics itself, such as the prediction of chromatographic elution profiles, optimal experiment design solutions, and the provision of confidence intervals on predictions, all to name but a few. Moreover, the onset of machine learning models in the field has also led to a need for comparing and evaluating the impact of these models. Of course, the presence of machine learning-interested researchers in proteomics has also led to calls for the establishment of a machine learning in proteomics, community, which in turn also led to education efforts in the form of ProteomicsML. As a result, there are several interesting opportunities regarding the further advancement of machine learning in proteomics, which can be organised into three subtopics: (i) new models and applications; (ii) comparing of models and their downstream impact; and (iii) community-building and support around machine learning in proteomics.

3.2 Topic introduction: Mass spectrometry-based single-cell proteomics: current challenges

Laurent Gatto (University of Louvain, BE)

License  Creative Commons BY 4.0 International license
© Laurent Gatto

Single cell proteomics is a relatively new field within proteomics, as the analysis of the truly minute quantities of proteins from single cells had been regarded as well outside the range of the mass spectrometers' limits of detection. However, instrumentation and methodological advances have allowed these limits to be breached, and have resulted in a rapidly growing interest in, and uptake of, single cell proteomics. This is not to say that analysing the proteome of a single cell is now by any means a routine endeavour. It remains a highly challenging and work-intensive process, and requires dedicated instrumentation and extremely careful sample preparation and handling. However, the challenges in single cell proteomics are most certainly not only practical. Indeed, considerable computational challenges accompany this young field, and these can be summarised in three categories as identification challenges, quantification challenges, and statistical data analysis challenges. Indeed, at all levels of the data processing workflow, single cell proteomics provides unique challenges for present-day computational algorithms and approaches, and thus plentiful opportunities for research into novel methods.

3.3 Topic introduction: Glycomics and glycoproteomics – Computational challenges and opportunities

Sriram Neelamegham (University at Buffalo – SUNY, US)

License  Creative Commons BY 4.0 International license
© Sriram Neelamegham

The field of glycomics/glycoproteomics is a challenging one, due to the combined complexity of requiring analysis of both glycan structures as well as their carrier peptide sequences. Importantly, while a vast array of software tools are available for glycomics (the study of the glycans) on the one hand, and glycoproteomics (the study of the peptides carrying the glycans) on the other hand, many of these tools have been deprecated swiftly after publication, are not open source, or lack modularity. This unfortunately greatly limits the actual availability of computational solutions in the field. A classification of the various computational approaches can be made in four types: peptide-first, glycan-first, O-glycan specific, and glycan database focused algorithms. These approaches vary in their sensitivities, and limited overlap between the results of various tools is the rule rather than the exception, as also highlighted in a recent Human Proteome Organisation (HUPO) study on the topic. However, a very promising new avenue is provided by machine learning approaches, and it is very encouraging to see heightened interest from the field in supporting, and subsequently adopting, such approaches. However, some key challenges remain in the field, notably the lack of large, well-curated data sets. There are interesting opportunities available for exploration, however, especially if curated data sets of sufficient size were to become available, and these include the use of ion mobility information in the identification process, and the possibility to predict retention times from glycan (or even peptido-glycan) structures. In summary, some of the most pressing challenges in the field of glycomics/glycoproteomics are clearly computational in nature, and there is therefore ample opportunity for innovation in this field.

3.4 Ad-hoc talk: Unsupervised clustering of spectra

Lukas Käll (KTH Royal Institute of Technology – Solna, SE)

License  Creative Commons BY 4.0 International license
© Lukas Käll

Tandem mass spectrometry experiments generate large amounts of spectra, generally with high redundancy between and within samples. An exciting way to process such data is to use unsupervised clustering to group spectra with similar appearance that tentatively were generated by the same analyte across samples. By associating the spectra to their corresponding MS1 abundances, the approach enables the quantification of analytes represented by spectra before their identification. The approach also allows the merging of the spectra of the clusters into consensus spectra, potentially increasing the accuracy of further identification. This approach has successfully been applied to proteomics data (<https://doi.org/10.1038/s41467-020-17037-3>), and we discussed if the approach could be further extended into the analysis of glycans and glycopeptides.

3.5 Ad hoc talk: A community-driven project – Rusteomics

Dirk Winkelhardt (Ruhr-Universität Bochum, DE & ELIXIR Germany – Jülich, DE)

License © Creative Commons BY 4.0 International license
© Dirk Winkelhardt

The Rust programming language has been making inroads in recent years, due, amongst others, to its speed of execution, its use of centralised libraries (crates) coupled to an excellent dependency manager (Cargo), and the ability to be cross-compiled to multiple platforms. Moreover, Rust is conceived to be particularly safe, as the memory ownership model protects from memory leaks at compile time. In proteomics as well, Rust has started to gain some initial traction, and this talk documents a Rust in proteomics hackathon that took place at the EuBIC Developers Meeting in early 2023. After a brief introduction to the programming language, a simple exercise was completed by the participants in which the peak intensity of the peak closest to a target mass-over-charge was determined. The hackathon participants then engaged in testing the I/O speed of rust, which was deemed to be extraordinarily high for reading text-formatted mass spectra (in MGF format) and sequence database files (in FASTA format). In addition, efforts were spent on exploring the generation of API documentation (via Sphinx), project structure, and the use of Github-Actions on Rust and Python code. The participants also worked on bridging Microsoft .Net libraires with Rust code, as most vendors write their proprietary libraries (which are essential when reading their proprietary, binary data formats) in .Net. Finally, the participants drafted future plans for Rust in proteomics, which involved setting up a host of core mass spectrometry libraries, and integration of the Rust-written Sage open modification search engine in this overall framework.

4 Working groups

4.1 Working Group Report: Machine Learning in Proteomics

Robbin Bouwmeester (Ghent University, BE), Viktoria Dorfer (University of Applied Sciences Upper Austria, AT), Laurent Gatto (University of Louvain, BE), Arzu Tugce Guler (Leiden, NL), Tiannan Guo (Westlake University – Hangzhou, CN), Michael Hoopmann (Institute for Systems Biology – Seattle, US), Lukas Käll (KTH Royal Institute of Technology – Solna, SE), Ville Koskinen (Matrix Science Ltd. – London, GB), Lennart Martens (Ghent University, BE), Magnus Palmblad (Leiden University Medical Center, NL), Tobias Schmidt (MSAID – Garching, DE), Veit Schwämmle (University of Southern Denmark – Odense, DK), Mathias Wilhelm (TU München – Freising, DE), and Dirk Winkelhardt (Ruhr-Universität Bochum, DE & ELIXIR Germany – Jülich, DE)

License © Creative Commons BY 4.0 International license
© Robbin Bouwmeester, Viktoria Dorfer, Laurent Gatto, Arzu Tugce Guler, Tiannan Guo, Michael Hoopmann, Lukas Käll, Ville Koskinen, Lennart Martens, Magnus Palmblad, Tobias Schmidt, Veit Schwämmle, Mathias Wilhelm, and Dirk Winkelhardt

The Working Group on Machine Learning (ML) in Proteomics convened every day, and even split off into two parallel groups for certain sessions to pursue multiple topics of discussion. This abstract reflects these discussions and their outcomes in chronological order throughout the Seminar.

On the first day, the Working Group began by looking at quality control (QC) and ML, and first of all, the ambiguity in the topic title was unpacked. Indeed, one could look at the use of ML in QC, or conversely, at the QC of ML. On the topic of using ML in quality control (QC), a presentation of data from several labs and instruments using paired features that may be suitable for ML and prediction of the source of experimental problems. Highlighting the discrepancy between observed and expected (ML predicted) features is key for successful QC, whether it's done using internal standards or, for example, relative entropy. On the topic of the QC of ML, one of the key needs is for users to know when they are using a model (too far) out of context. We discussed some very simple ways on how this could be presented, such as a (box) plot of model performance on the current data set(s) in the context of a boxplot of the expected performance of that model. Tools that provide such feedback are much more likely to get used in practice, and should confer higher confidence in the applicability of the model to that data set. From this, a digression into explainable AI was made, where feature relevance could be assessed by looking at training gradients, or looking at discrimination power rather than correlation between predicted and measured features.

The second day began with a discussion centered around the concept of Foundation Models, which are versatile models trained on broad data for various downstream tasks. The concept and its applications were introduced, which included for instance gene expression prediction. Participants questioned the potential for a proteomics-focused Foundation Model, given the challenges of data annotation. The group outlined a multi-step plan involving data collection, processing, and integration into a Foundation Model. Challenges, such as fragmentary and poorly annotated data, and variations between labs and instruments, were acknowledged. The group suggested that separate Foundation Models for sequences, spectra, and other data types could be linked together for a combined output. The conversation shifted to a proposed first spectra-only Foundation Model project. Several data repositories were suggested for sourcing spectral data, with the intent to cover as much sequence diversity as possible. Converting data to tensors for GPU computation was discussed, along with specific spectral encodings. The session resulted in a concrete plan to start by applying for GPU resources and a number of other action points for the coming months.

On the afternoon of the second day, the Working Group focused on two distinct topics. The first of these centered on the machine-learning based prediction of the behaviour of peptide and small molecule analytes on the instruments. We aimed to identify explainable properties, the models to use, and any opportunities for deep learning. Trivial, or already ML-predicted properties like mass, isotopic distributions, peak shapes, diffusion rates, protein localization, and isoelectric point were mentioned, upon which the focus shifted to properties benefitting most from further investment of machine learning techniques, such as ion suppression, ion mobility/collision cross-section (CCS), pH, and chromatographic column temperature. Revisiting early work with molecular dynamics and the application of association mining were raised. The impact of post-translational modifications (PTMs) on retention times and collision energy optimization for improved fragmentation were considered. Applying AI methods for protein structural dynamics, and for exploring enzymatic digestion and protein fractionation were also discussed. Action items include a review manuscript focusing on successful deep learning models like pDeep, Prosit and DeepLC. Throughout, this discussion highlighted machine learning's continuing potential for understanding and predicting physicochemical properties in MS-based proteomics.

The second topic of the afternoon of the second day discussed a project that aims to develop an end-to-end model to predict protein quantities directly from raw data, potentially linking it to a phenotype or disease state. This is a deeply challenging endeavor, but there are

already some preliminary data available for reprocessing. One project, soon to be published, offers around 20k DIA runs of serum samples with quantified proteins. Nevertheless, questions remain regarding the amount of available data. The model architecture is also a crucial consideration. While convolutional neural networks (CNNs) are one option, it is unclear whether the local structure is sufficient to become informative hyper-features. An alternative might involve using a transformer network that processes the output of a scan or window. The model's input is a tensor or matrix, possibly with mass-over-charge and intensity dimensions. However, there are concerns that the matrix might be too high-dimensional for conventional deep learning. The DIA tensor (DIAT) tool has been proposed to convert DIA (SWATH) data into a tensor, effectively creating large "images". Another possible concept to explore involves graph convolutional neural networks. With this approach, peaks are nodes, and correlations or co-occurrences between peaks form the edges. However, this approach could be problematic as calculating correlations between features might go against the idea of an end-to-end model and may complicate feature finding or peak picking. A possible starting point for the project could be focusing on fourteen proteins in serum or blood, to begin with. Incorporating ideas from the field of ion networks might also be beneficial, as this essentially forms a network of ions that could be used as a starting point.

The third day saw the Working Group first explore ways to interrogate deep learning models for their knowledge about physicochemical rules through data-driven exploration. Retention time and fragmentation spectrum prediction models seem to be most suitable to extract rules that are "semi-orthogonal" to what they have been trained with. That could include details about binding energies, structure and the impact of motifs. The approach would be to generate simple hypotheses and then explore their potential impact on the predictions. This approach can be automated into building a Great Hypothesis Tester built on regular expressions for the peptide input.

The afternoon of the third day considered the ProteomicsML.org resource, which was first demonstrated, followed by general discussions on possible additions and modifications. Suggestions include adding recommended reading for ML and proteomics beginners, linking to existing tutorials, and covering relevant frameworks. Currently, the platform uses various ML frameworks and mainly focuses on behavioral predictions. Considerations include simplifying ML frameworks, introducing structural methods like AlphaFold, and clarifying the purpose of ProteomicsML. Outreach should emphasize seeking contributors, offering lessons and tutorials, and addressing platform maintenance.

On the fourth day, finally, the discussions began by considering community aspects of ML model reproducibility through better access to data and code. Platforms such as ProteomicsML, DLOmix and Zenodo were suggested for sharing and comparing models. Ongoing efforts are being coordinated with BioHackathon Europe, and their BioModelsML project. Another point was that model evaluation requires better comparison metrics, possibly focusing on relevant biological information such as coverage of relevant pathways, rather than simply counting peptide-to-spectrum matches. Other topics covered variational autoencoders, probabilistic modeling, and missing value imputation, with suggestions to use noise injection and quantitative data, building on the histone-based "proteomic ruler".

4.2 Working Group Report: Single Cell Proteomics

Laurent Gatto (University of Louvain, BE), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Melanie Föll (Universitätsklinikum Freiburg, DE), Lukas Käll (KTH Royal Institute of Technology – Solna, SE), Ville Koskinen (Matrix Science Ltd. – London, GB), Lennart Martens (Ghent University, BE), Sriram Neelamegham (University at Buffalo – SUNY, US), Tobias Schmidt (MSAID – Garching, DE), Veit Schwämmle (University of Southern Denmark – Odense, DK), Mathias Wilhelm (TU München – Freising, DE), and Gamze Nur Yapici (Koc University – Istanbul, TR)

License © Creative Commons BY 4.0 International license

© Laurent Gatto, Bernard Delanghe, Melanie Föll, Lukas Käll, Ville Koskinen, Lennart Martens, Sriram Neelamegham, Tobias Schmidt, Veit Schwämmle, Mathias Wilhelm, and Gamze Nur Yapici

This abstract documents the discussions of the Working Group on Single Cell Proteomics, which took place throughout the Seminar, and whose subtopics are presented here in chronological order.

On the first day, the Working Group discussed that, whenever a novel technology is proposed, or an existing one is pushed beyond what was considered possible, such as in mass spectrometry-based single-cell proteomics (SCP), one can wonder whether existing software are still applicable. A notable comparison between fragmentation mass spectrometry scans from bulk and single-cell data highlighted possible differences, and thus wondered if existing search engines can be used as they are. Our working hypothesis is that the differences between SCP and bulk fragmentation mass spectra are mainly due to the lower intensity of the precursors, leading to some fragments getting lost, and that current tools can in fact still be used. We plan to test our hypothesis using data readily available from the members of our Working Group, comparing identifications with and without re-scoring of features, exploration of these features, and incorporating precursor intensity into the identification models.

On the second day, the discussion initially focused on SDRF metadata for proteomics. We first discussed which adaptations and challenges arise when using SDRF to annotate single cell proteomics experiments. We furthermore discussed how to generally improve the SDRF format e.g. more automatic collection of metadata and using it to facilitate writing materials & methods sections. Both would likely improve usage of SDRF by the community.

In the afternoon of the second day, the Working Group concentrated on recommendations for single-cell DIA data analysis. Single-cell TMT techniques are already “solved”, because FDR control and batch effects are the same as non-single cell data. We further narrowed down to identification. With DIA, the two main tools are DIA-NN and Spectronaut. Users treat both as “black boxes”, and much more transparency is needed on configuration options as well as the rationale for enabling or disabling various processing steps. We recommend saving the log file and exporting the config options at minimum, and tool developers are strongly recommended to provide sensible defaults and hide less-commonly used advanced options. It is also common to analyse a bulk sample together with single-cell raw files and use options like match-between-runs (MBR) to boost peptide and protein IDs. We will test the validity of this procedure by acquiring human single-cell data and analysing it together with 1) a bulk human sample (say 100 cells) and 2) an *E. coli* sample. The expectation is that the latter has no effect on identification rates. The opposite outcome would demonstrate the procedure is not valid.

The third day saw continued discussion on practical pipeline issues. Scaling of the data processing capability already is a current, and will be an especially important future challenge. Processing 100 raw files is still doable, but no tools currently handle thousands of samples.

Data size poses storage issues as well as data transfer problems, such as that one cannot easily upload a terabyte of data to the cloud. Moreover, many institutions forbid cloud processing of sensitive proteomics data, which differs from transcriptomics and genomics data processing. File formats like mzML are space inefficient, so can a better standard format be selected or developed? Or is it better to use just vendor raw files? The field also needs to clarify the research question at the end of a single-cell study. It is infeasible to interpret a spreadsheet with a thousand columns, so is it sufficient for software to only report protein differences between cells? To answer these questions, we need a concrete example, such as the 100 cell data set already discussed.

On the fourth day, the discussion touched on several different angles on how and whether transcriptomics of single cells could be combined with single-cell proteomics. Due to the extremely low sample amounts, it is difficult to do both on the same cell. Mosaic integration could allow some shared dimensions. Another discussion focused on the importance of the different modalities, and that we should expect to see the same effect. This, and the substantial differences in dimensions (a thousand for SCP, but ten thousand for scRNA-Seq) would also make their integration difficult or even questionable. Even within a single modality, focusing on proteomics, different features will provide different information. Carefully selected surface markers used in flow cytometry, for instance, will not necessarily be recapitulated, at least not as clearly, by hundreds or thousands of proteins; the same large, well differentiated clusters/types are expected to be recovered, but with much more variability. We finally discussed the notion of stable and unstable balance – protein abundances won't change easily if many copies are present in the cell, whereas proteins with only few copies can much more easily suffer changes in intensity. Moreover, (post-translational) modifications could also easily trigger quantification or cellular balances.

4.3 Working Group Report: Glycosylation and Glycoproteomics

Rebekah Gundry (University of Nebraska – Omaha, US), Kiyoko Aoki-Kinoshita (Soka University – Tokyo, JP), Robbin Bouwmeester (Ghent University, BE), Robert Chalkley (University of California – San Francisco, US), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Viktoria Dorfer (University of Applied Sciences Upper Austria, AT), Melanie Föll (Universitätsklinikum Freiburg, DE), Arzu Tugce Guler (Leiden, NL), Catherine Hayes (Swiss Institute of Bioinformatics – Geneva, CH), Michael Hoopmann (Institute for Systems Biology – Seattle, US), Lukas Käll (KTH Royal Institute of Technology – Solna, SE), Ville Koskinen (Matrix Science Ltd. – London, GB), Karina Martinez (George Washington University – Washington, DC, US), Sriram Neelamegham (University at Buffalo – SUNY, US), Magnus Palmblad (Leiden University Medical Center, NL), Erdmann Rapp (MPI – Magdeburg, DE), Tobias Schmidt (MSAID – Garching, DE), Mathias Wilhelm (TU München – Freising, DE), Bernd Wollscheid (ETH Zürich, CH), and Gamze Nur Yapici (Koc University – Istanbul, TR)

License © Creative Commons BY 4.0 International license

© Rebekah Gundry, Kiyoko Aoki-Kinoshita, Robbin Bouwmeester, Robert Chalkley, Bernard Delanghe, Viktoria Dorfer, Melanie Föll, Arzu Tugce Guler, Catherine Hayes, Michael Hoopmann, Lukas Käll, Ville Koskinen, Karina Martinez, Sriram Neelamegham, Magnus Palmblad, Erdmann Rapp, Tobias Schmidt, Mathias Wilhelm, Bernd Wollscheid, and Gamze Nur Yapici

The Glycosylation and Glycoproteomics Working Group convened in several sessions throughout the Seminar, and discussed a variety of topics, which are listed in chronological order in this abstract.

On the first day of the seminar, the Working Group discussed the idea that we do not yet know how or if the presence of glycosylation affects fragmentation of the peptide backbone when acquiring MS/MS data on intact glycopeptides. Does the glycan affect observation and/or intensities of fragment ions? The question is important to answer as, if the presence of glycosylation does affect fragmentation, it stands to reason that fragmentation prediction algorithms and fragmentation interpretation algorithms would need to consider this to ensure accurate interpretation of the data. Major questions to be answered include: 1) Does the presence of a glycan (site occupied vs site not occupied) affect fragmentation? 2) Does the fragmentation of the peptide differ if the glycan composition on the site is varied? 3) Does the fragmentation of the peptide differ if the glycan structure on the site is varied? 4) Does the fragmentation of the glycopeptide differ if the glycan is attached to a glycopeptide that is the result of full tryptic digestion vs. one with ragged (non-tryptic) terminus? 5) Does the presence of N-linked glycosylation have the same or different effect on peptide fragmentation as the presence of O-linked glycosylation?

The second day saw the Working Group discussing current challenges with glycopeptide and released glycan analyses as it relates to best practices for interpreting and reporting data. Major concepts included discussion of ambiguity in glycopeptide and glycan assignments, the use and misuse of oxonium ions in glycopeptide spectra interpretation, the impact of search space on false discovery rate (FDR) calculations and accuracy of assignments, and the call to action for software developers to incorporate strategies to highlight assignments that may be ambiguous and to include GlyYouCan accession numbers, the reference annotation language for glycan compositions, topologies, and structures. Our discussion led to the development of an outline for a perspective or white paper focused on best practices in data analysis and reporting, which will serve as a catalyst to promote scientists to follow MIRAGE guidelines. It will include real-world examples to promote scientists at any level of experience in glycoproteomics or glycomics to become aware of key issues which may otherwise not be obvious to those with limited experience in this discipline.

In the afternoon of the second day, the discussion turned to a variety of topics related to how we can make more and better use of the glycopeptide and glycomic data that we generate. We discussed available tools for interpretation and integration of data, including a tour of tools on GlyConnect website. On the subject of glycopeptide quantification, the opinion of the group was that this has so far remained an unmet goal, despite any claims to the contrary. We decided to add this as point #5 in the best practices manuscript outlined in the morning session of the second day – to outline best practices for anyone publishing in a journal that is not *Nature*. Several studies have been published that compare “quantitative” comparisons of glycoforms of glycopeptides to suggest that some glycan classes are “more abundant” on a site than another. Given the inherent differences in ionization potential for glycans which have different compositions, this is problematic. We reviewed published data that demonstrates the difference in peak abundance for different glycoforms. We will use these data as examples in our best practices manuscript. We need to get more funding to support glycoproteomics and glycomics. While we have a lot of success stories regarding analytical capabilities, applications are less known. There are certainly success stories to demonstrate how our approaches have impacted physiology and disease, including fundamentals of physiological processes and clinical examples. However, not all examples are highly visible. We discussed ideas for creating resources as a community to help educate others of our value. This could be videos, documents, websites that summarize success stories that would resonate with institutional leaders, administrators, benefactors, funding agencies.

The third day focused initially on interactions between proteomics and glycomics, with a team of experts in glycomics and proteomics discussing ideas on how to transfer knowledge from the proteomics field for application in glycomics workflows. The first part of the discussion focussed on current challenges in the analysis of glycopeptides and released glycans, particularly with respect to: 1. Structure assignment, i.e. the challenges in the identification of glycan composition in glycoproteomics studies, and also with respect to topology and structure determination that are common to both the fields of glycomics and glycoproteomics. These challenges stem from, amongst others, the isomeric nature of glycans. 2. Variable experimental workflows: current best practices in glycomics studies and glycoproteomics wetlab studies were discussed with respect to different experimental workflows used by multiple groups. The use of multiple experimental methods, such as derivatization strategies and fragmentation modes results in lack of consensus in the field. 3. Limitations of current software solutions: It was highlighted that glycosciences is a niche area, with relatively few researchers deeply involved. Knowledge was shared about observations on how MS intensity (or relative intensity) may be used for glycan topology assignments. The conclusion of the first part was that the glycan search space is not so large and it may be possible to develop spectral libraries or clustering approaches to find reliable topology solutions. Additionally, already established proteomics tools may be adopted using transfer learning approaches to aid developments in glycoproteomics. Among the methodologies discussed were i) MaRaCluster as it is agnostic to MS/MS spectra and may be used to develop consensus spectra. While there is no perfect method to develop consensus spectra for glycomics, a variety of approaches were also discussed ii) GLEAMS as reference spectra may be populated in this framework, perhaps using Siamese/pairwise inputs so that the system may be trained to find commonalities and distinctions among glycan classes. While there were many advantages to this approach, concerns were raised regarding the ability of proteomics trained datasets to learn glycomics experimental results, and the fact that the reference spectra may vary with collision energy and that the actual experimental results would have to be tightly clustered to a reference dataset. iii) Application of transfer learning, possibly using graph convoluted neural networks, to predict MS/MS intensity. Challenges were discussed here with respect to the representation of branched glycan structures as graphs in such modeling. The final portion discussed the need to either write an independent perspective article related to this subject, or simply to use this discussion as discussion material in the paper that is ‘forthcoming’ from the first day’s session. This will be important to raise awareness and obtain additional funding to support the project.

The afternoon of the third day brought a discussion on the common concepts we consider when trying to connect our glycomics and glycoproteomics data to biology, including a review of available resources to support these efforts. We also reviewed a public repository of bioinformatic tools. We discussed the state of the art of tools for glycopeptide and glycomics analysis, including limitations and what would be important to consider for future versions. Overall, challenges to selecting which tools make sense to focus on include that there is high heterogeneity in the field regarding which approaches to use, so not immediately clear that there is a single solution to focus on as it would risk putting effort into a tool that would not be broadly used in the field.

The fourth day, finally, continued the discussion of available tools and their limitations. We had new discussion of how we can better make use of machine learning and neural network tools to advance the field. An example emerged that centered around how we could use machine learning to help the interpretation of glycosidase reactions. Briefly, we often treat samples with glycosidases to help determine structural details. These reactions cause mass

and RT shifts in the glycans/glycopeptides. This is probably something where ML could help us to automate and interpret the results. We ended by planning details regarding the two manuscripts we anticipate submitting by end of 2023. The first will be a research article that uses existing data to answer the question of whether the glycan/peptide affects peptide fragmentation. The other is a guidelines/tutorial type paper likely to be submitted to MCP.

Participants

- Kiyoko Aoki-Kinoshita
Soka University – Tokyo, JP
- Robbin Bouwmeester
Ghent University, BE
- Robert Chalkley
University of California –
San Francisco, US
- Bernard Delanghe
Thermo Fisher GmbH –
Bremen, DE
- Viktoria Dorfer
University of Applied Sciences
Upper Austria, AT
- Melanie Föll
Universitätsklinikum
Freiburg, DE
- Laurent Gatto
University of Louvain, BE
- Arzu Tugce Guler
Leiden, NL
- Rebekah Gundry
University of Nebraska –
Omaha, US
- Tiannan Guo
Westlake University –
Hangzhou, CN
- Catherine Hayes
Swiss Institute of Bioinformatics –
Geneva, CH
- Michael Hoopmann
Institute for Systems Biology –
Seattle, US
- Lukas Käll
KTH Royal Institute of
Technology – Solna, SE
- Ville Koskinen
Matrix Science Ltd. –
London, GB
- Lennart Martens
Ghent University, BE
- Karina Martinez
George Washington University –
Washington, DC, US
- Sriram Neelamegham
University at Buffalo –
SUNY, US
- Magnus Palmblad
Leiden University Medical
Center, NL
- Erdmann Rapp
MPI – Magdeburg, DE
- Tobias Schmidt
MSAID – Garching, DE
- Veit Schwämmle
University of Southern Denmark –
Odense, DK
- Mathias Wilhelm
TU München – Freising, DE
- Dirk Winkelhardt
Ruhr-Universität Bochum, DE &
ELIXIR Germany – Jülich, DE
- Bernd Wollscheid
ETH Zürich, CH
- Gamze Nur Yapici
Koc University – Istanbul, TR

