Report from Dagstuhl Seminar 17421

# Computational Proteomics

**Edited by**

**Bernhard Küster[1], Kathryn Lilley[2], and Lennart Martens[3]**

1 **TU München, DE,** kuster@tum.de
2 **University of Cambridge, GB,** k.s.lilley@bioc.cam.ac.uk
3 **Ghent University, BE,** lennart.martens@ugent.be

## Abstract

The Dagstuhl Seminar 17421 “Computational Proteomics” discussed in-depth the current challenges facing the field of computational proteomics, while at the same time reaching out across the field’s borders to engage with other computational omics fields at the joint interfaces. The ramifications of these issues, and possible solutions, were first introduced in short but thought-provoking talks, followed by a plenary discussion to delineate the initial discussion sub-topics. Afterwards, working groups addressed these initial considerations in great detail.

**Seminar** October 15–20, 2017 – http://www.dagstuhl.de/17421
**1998 ACM Subject Classification** Bioinformatics, Computational proteomics, Metabolomics / metabonomics, Transcriptomics
**Keywords and phrases** computational biology, computational mass spectrometry, integrative bioinformatics, large scale public data, proteomics
**Digital Object Identifier** 10.4230/DagRep.7.10.69

## 1 Executive Summary

*Lennart Martens*

The Dagstuhl Seminar 17421 “Computational Proteomics” discussed in-depth the current challenges facing the field of computational proteomics, while at the same time reaching out across the field’s borders to engage with other computational omics fields at the joint interfaces. The issues that were discussed reflect the emergence of novel applications within the field of proteomics, notably proteogenomics (the identification of proteins based on sequence data obtained from prior genomics and/or transcriptomics analyses), and metaproteomics (the study of the combined proteome across an entire community of (micro-)organisms). These two new proteomics approaches share several challenges, which predominantly revolve around the sensitive identification of proteins from large databases while maintaining an acceptably low false discovery rate (FDR). The ramifications of these issues, and possible solutions, were first introduced in short but thought-provoking talks, followed by a plenary discussion to delineate the initial discussion sub-topics. Afterwards, working groups addressed these initial considerations in great detail.

In addition, both proteogenomics and metaproteomics suffer from coverage issues, as neither is currently capable of providing anywhere near a complete view on the true complexity of the (meta-)proteome. This issue is exacerbated by the fact that the true extent of the proteome remains unknown, and is likely to be time-dependent as well. As a result, a separate

Computational Proteomics, *Dagstuhl Reports*, Vol. 7, Issue 10, pp. 69–83
Editors: Bernhard Küster, Kathryn Lilley, and Lennart Martens
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

working group was created to discuss the issues and possible remedies related to proteome coverage.

The field of proteomics has, however, not only extended into novel application areas, but meanwhile also continues to see a strong development of novel technologies. Over the past few years, the most impactful of these is data-independent acquisition (DIA), which comes with its own unique computational challenges. On the one hand, the analysis of DIA data currently relies heavily on spectral libraries, which have so far been a rather niche product in proteomics (as opposed to, for instance, metabolomics, where spectral libraries have a much longer and much more fruitful history), while on the other hand, FDR estimation remains contested in DIA approaches. As a result, two further working groups were established during the seminar, one on the applications for, and methods to create spectral libraries, and the other on the specific challenge of calculating a reliable FDR when performing spectral library searching.

Another key topic of the seminar was the (orthogonal) re-use of public proteomics data, which focused on the provision of metadata for the assembled proteomics data, as this is the key bottleneck facing researchers who wish to perform large-scale re-analysis of public proteomics data, especially when the objective is to obtain biological knowledge. A working group was therefore created to explore the issues with metadata provision, and to explore means to ameliorate the current suboptimal metadata reporting situation.

Throughout the seminar, the topic of visualizing the acquired data and the obtained results cropped up with regularity. A corresponding working group was therefore set up to delineate the state-of-the-art in proteomics data visualization, and to explore the issues with, and opportunities of advanced visualizations in proteomics.

As a last core topic, a short introductory talk and subsequent working group was dedicated to the education of computational proteomics researchers, with special focus on their ability to work at the interfaces with other omics fields (genomics, transcriptomics, and metabolomics). This working group assembled an extensive list of already available materials, along with an overview of the different roles and specializations that can be found across informaticians, bio-informaticians, and biologists, and how each field should evolve in order to bring these more closely together in the future.

In addition to abovementioned topic introduction talks, and the associated working groups, two talks illustrated specific topics of the seminar. Paul Wilmes showed his recent work in bringing metaproteomics together with advanced metatranscriptomics and metagenomics, showing that the flexible use of sequence assembly graphs at the nucleotide level opens up many highly interesting possibilities at the proteome level through enhanced identification. Nevertheless, it was observed that there is strong enrichment for genes with unknown function at the protein identification level, highlighting quite clearly that we have yet to achieve a more complete biochemical understanding of microbial ecosystems. Finally, Magnus Palmblad delighted the participants with a highly original talk on the exploration of mass spectrometry data (of both peptides as well as small molecules) through the five senses (sight, hearing, touch, smell, and taste).

## 2 Table of Contents

**Executive Summary**

**Overview of Talks**

**Working groups**

# 3 Overview of Talks

## 3.1 Exploration of the key computational interfaces between omics domains

*Frédérique Lisacek (Swiss Institute of Bioinformatics, CH)*

Connecting glycomics with proteomics and interactomics raises many issues. To begin with, protein glycosylation and its impact on structure and function is widely ignored probably due to the lower throughput of glycomics experiments in comparison to other omics. Nonetheless this modification along with many others generates proteoforms and the extent of this repertoire as well as possible cross-talk between modifications remains difficult to study and evaluate. Another obstacle in linking glycomics with other –omics is the independent accumulation of data regarding the constituents of glycoconjugates. Most glycan structures are solved after being cleaved off their natural support and key information on the conjugate is lost. Conversely, protein glycosylation sites are mapped independently of the glycan structure and key information on the attached glycan is lost. Glycoproteomics is on the rise and a promising technology that preserves associated glycan and peptide data though data submission and sharing remains confidential at this stage. Despite the numerous gaps challenging software development, rapid change is expected in this field in the years to come. Furthermore, the role of glycans comes to the forefront in many biomedical applications including for example microbiome studies; glycans mediate specific protein-protein interactions.

## 3.2 Interpretation of proteomics and transcriptomics to model the dynamics of gene expression

*Gerben Menschaert (Ghent University, BE)*

The task of integrating proteomics and transcriptomics data faces several challenges. First, there is the complexity of the proteome (due to the wide variety of proteoforms), which confounds the specifics of the (re)-annotation mechanism to employ, not in the least because of issues with the mapping of sequence information between RNA and protein. This also impacts the correlation of quantification between the different ‘omics data types. Specific issues are encountered in the context of immunopeptides, where the sequences are possibly not directly genomically encoded. Because of these issues, it is clear that specific tools and algorithms need to be developed for proteogenomics. The following topics were described in more detail as indicators of where the field is moving. Novel sequencing technologies that are currently developed (PacBio, SMRT-seq and MinION, Oxford Nanopore) will shed new light on the identification of novel translation products from lncRNAs, sORFs, uORFs. But ot make the most of these, cell-type or tissue-type matching across-omics datasets will be needed. Moreover, these novel findings will also have to be deposited into public repositories, as these can then be used for genome re-annotation. In that context, it is necessary to develop a better integration between PRIDE (proteomics) and Ensembl (genomics/transcriptomics). The issues surrounding such integration are missing metadata, stringent filtering for false positives and a need for robust workflows. From a quantitative correlation perspective, robust implementations are needed to compare sequencing-based semi-quantitative measures with mass spectrometry derived quantitative metrics. FInally, there is a need to further improve upon, or develop de novo, tools for the integration with genomics, for the elaboration of standards such as proBED and proBAM to map proteomics data to genome browsers, and for establishng tighter connectoins with platform interfaces like BioConda/Galaxy.

## 3.3 Assessing and addressing the specific computational challenges of metaproteomics

*Thilo Muth (Robert Koch Institut – Berlin, DE)*

In recent years, the impact of microbial consortia on human health has gained increasing attention due to the acquired knowledge regarding the important role of the intestinal gut microbiome. Metaproteomics, the mass spectrometry based proteomic analysis of an entire microbial community, helps to elucidate enzymatic capabilities and taxonomic origin of key species. Accordingly, this method can also be used for detecting pathogens in samples from which the exact microbial origin is not known. However, several computational challenges exist of which the most severe ones are highlighted here. First, there is the protein inference issue, which is worsened in metaproteomics because of the occurrence of multiple homologous species, with many homologous protein sequences, in complex and heterogeneous samples. Second is the need to select an appropriate database that covers a sufficient amount of relevant species without too strong a bias (e.g., towards clinical relevant strains). Third is the large and constantly growing size of the resulting sequence databases, which affects FDR estimation and/or sensitivity of the searches in target-decoy approaches. Fourth, biologists typically want answers to very specific questions, which require much more than a protein identification list; for instance, identifying a certain pathway, its functional proteins, and its related microbial species. On the computational level, the combination of multiple database search engines is shown as a reasonably straightforward means to increase the detection rate at the taxonomic level. Finally, the current status and performance of de novo sequencing algorithms is demonstrated, along with their potential and limitations when used as alternative approaches to database driven peptide identification.

## 3.4 Exploring mass spectrometry data with the five senses

*Magnus Palmblad (Leiden University Medical Center, NL)*

Another view on mass spectrometry data, exploring the possible applications of sight, hearing, touch, smell, and even taste in the sensory perception and analysis of mass spectrometry data. This is illustrated with anaglyphs, audio of mass spectrometry transients and spectra, and 3D printed mass spectra, complete with chromatographic and isotopic dimensions. A small contest will be held to guess the mass of a compound for which the isotope distribution is given as an image, and another which is provided as a 3D printed model in a closed box. The sense of smell coupled to liquid chromatography and even mass spectrometry is also discussed.

## 3.5 Training of integrative bioinformatics experts

*Hannes Röst (University of Toronto, CA) and Andreas Hildebrandt*

The training of integrative bioinformatics experts will be a key challenge for educators throughout the next decade. This challenge is complicated by the fact that the field is evolving rapidly, and that several types of undergraduate and/or Master's degrees could feed into such a programme. It is therefore likely that there will not be a single such curriculum, but rather a set of courses, from which a choice is made based on the pre-existing knowledge of the trainee. At the same time, the level of education on which this training is to take place is flexible. Basic programming courses in often-used languages such as Python, for instance, should preferentially begin at the undergraduate level at the latest (it would be far better to start much earlier, e.g., in secondary education), while training in advanced mathematical modelling is more likely to take place at the Master's or at the post-graduate level. Overall, online training courses could be a very interesting means to educate people, but care should be taken that the courses stay up-to-date. This is challenging in a fast evolving field, and will require substantial time investment.

## 3.6 Analysis and interpretation of public proteomics data in orthogonal contexts

*Juan Antonio Vizcaino (EBI – Hinxton, GB)*

The availability of public proteomics datasets continues to increase, and a plateau has clearly not been reached yet. Many possibilities for data reuse exist and some of these forms are increasingly popular. Two particularly rewarding but difficult scenarios for reuse of proteomics data are 're-analysis' and 're-purpose'. The difference between the two is subtle: in the case of the former, the analysis settings change compared with the original study, but the goals of the study do not. In the case of the latter, both analysis settings as well as goals can be different from the original. In the case of 're-analysis', examples are found in widely used resources, e.g. Peptide Atlas, MassIVE and ProteomicsDB. 'Re-purpose' examples can be found in proteogenomics studies and in the detection of new PTMs/sequence variants. Existing challenges for the reuse of data were highlighted as well. There is a lack of suitable annotation for many data sets, which prevents re-use to extract biological meaning. Moreover, there is a need for robust computational infrastructure to provide the required calculation power. A special mention is made of the difficulty in matching different datasets coming from different 'omics approaches (where the data is also spread across different 'omics specific data repositories). A slowly emerging issue that should be taken into account already, is access controlled data in the case of clinical samples. Of course, any re-analysis will run into challenges related to false discovery rate estimations in the context of large search spaces (for instance, when searching for single amino acid substitutions) and when false positives are combined across different data sets. At the same time, there are opportunities available that have not been covered so far. The re-analysis of atypical data sets, such as metaproteomics experiments or data independent acquisition (DIA) analyses provides an obvious example, but hinges on the availability of dedicated algorithms, and specialized resources such as spectral libraries. A more future-oriented goal is the integration of proteomics and metabolomics data sets to elucidate metabolite fluxes and influences on the proteome.

## 3.7 Integrated multi-omics for enhanced metaproteomics

*Paul Wilmes (University of Luxembourg, LU)*

Metaproteomics involves analysing the protein complement of microbial consortia. Peptide and protein identification is challenged by the inherent complexity of the samples. The generation of concomitant metagenomic and metatranscriptomic data allows the construction of sample-specific protein databases which facilitates enhanced data usage including for protein identification. Furthermore, exploitation of the de Bruijn metagenomic and metatranscriptomic assembly graphs allows the resolution of variant paths which in turn enables strain-level resolution of peptides and proteins. These approaches greatly enhance peptide and protein identification rates. Consequently, integrated multi-omic analyses of microbial communities overall result in much improved metaproteomic coverage.

# 4 Working groups

## 4.1 False Discovery Rates in Spectral Library Searching and Data Independent Acquisition Identification

*Eric Deutsch (Institute for Systems Biology – Seattle, US), Robert Chalkley (UC – San Francisco, US), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Viktoria Dorfer (University of Applied Sciences Upper Austria, AT), Nico Jehmlich (UFZ – Leipzig, DE), Bernhard Küster (TU München, DE), Hannes Röst (University of Toronto, CA), Timo Sachsenberg (Universität Tübingen, DE), Stephen Tate (SCIEX – Concord, CA), Mathias Wilhelm (TU München, DE), and Paul Wilmes (University of Luxembourg, LU)*

The breakout group on data indepenten acquisition (DIA) spectral library false discovery rate (FDR) in the Dagstuhl Seminar on Computational Proteomics, containing 11 participants, discussed the current issues in estimating and maintaining the reliability during generation of spectral libraries and in subsequent analyses. As far as the generation of spectral libraries is concerned, determining the FDR at the creation of the library is based on data dependent acquisition (DDA) FDR estimates. However, when it comes to extending the spectral libraries for new entries, maintenance and estimation of the FDR within the library is not really solved other than re-searching all the data again. The within library FDR should be propagated and considered in the search results to compensate for the reliability of the library itself.

Estimating the FDR on spectral library search results is a challenge itself, as decoy generation is not as easy as for database searches. Current methods seem to over- or underestimate the true FDR in the data sets. As an action item this breakout group aims to generate one (or more) gold standard spectral library data sets to evaluate current and future approaches for FDR estimation. This could also allow for checking whether decoy generation is the way to estimate FDR for spectral library searching. It was recognized that approaches being applied for DDA may not be valid for DIA. FDR calculation for DIA data is even more complicated than for DDA because of the increased complexity. We may have to come up with new/better solutions to estimate FDR on DIA searches.

The group also discussed a few additional related topics, including how post-translational modification (PTM) site localisation could be handled and how FDR estimation approaches from other fields could be adopted.

## 4.2 Spectral Libraries in Proteomics

*Eric Deutsch (Institute for Systems Biology – Seattle, US), Nuno Bandeira (University of California – San Diego, US), Sebastian Böcker (Universität Jena, DE), Robert Chalkley (UC – San Francisco, US), John Cottrell (Matrix Science Ltd. – London, GB), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Viktoria Dorfer (University of Applied Sciences Upper Austria, AT), Nico Jehmlich (UFZ – Leipzig, DE), Lukas Käll (KTH – Royal Institute of Technology, SE), Hannes Röst (University of Toronto, CA), Timo Sachsenberg (Universität Tübingen, DE), Stephen Tate (SCIEX – Concord, CA), Hans Vissers (Waters Corporation – Wilmslow, GB), Pieter-Jan Volders (Ghent University, BE), Mathias Walzer (EBI – Hinxton, GB), Ana L. Wang (Scripps Research Institute – La Jolla, US), Mathias Wilhelm (TU München, DE), and Dennis Wolan (Scripps Research Institute – La Jolla, US)*

The eighteen participants of the Spectral Libraries Breakout Group of the 2017 Computational Proteomics Dagstuhl Seminar discussed the current state and future directions for the generation and use of peptide tandem mass spectrometry spectral libraries. Their use in proteomics is growing slowly, but there are multiple challenges in the field that must be addressed to further increase the adoption of spectral libraries and related techniques. This Spectral Libraries Breakout Group aims to generate and publish a set of recommendations for addressing these challenges, building on prior work of the Proteomics Standards Initiative (PSI).

The primary bottlenecks are the paucity of high quality and comprehensive libraries, and the general difficulty of adopting spectral library searching into existing workflows. There are several existing spectral library formats, but none of them capture a satisfactory level of metadata, and therefore a logical next advance is to design a more advanced, PSI-approved spectral library format that can encode all of the desired metadata.

The group discussed a series of metadata requirements, organized into three levels of completeness or quality, tentatively dubbed bronze, silver, and gold. The metadata would be encoded at the collection (library) level (e.g., methods details, such as whether library spectra are consensus or representative spectra), at the individual entry (peptide ion) level (e.g., FDR of identification used for inclusion in the library), and at the peak (fragment ion) level (e.g., intensity variance).

The group discussed strategies for encoding mass modifications in a consistent manner (there was agreement that the mzTab specification seems adequate) and the requirement for encoding high quality and commonly-seen but as-yet unidentified spectra. The exact style of the new standard format (e.g., enhancement of the currently most popular MSP format, XML, heavily optimized binary formats, database-based storage, etc.) remains the subject of vigorous debate.

The group also discussed a few additional related topics, including strategies for comparing two spectra, techniques for generating representative spectra for a library, approaches for selection of optimal signature ions for targeted workflows, and issues surrounding the merging of two or more libraries into one.

## 4.3 Assessment of proteome coverage

*Gerben Menschaert (Ghent University, BE), Marco Hennrich (EMBL – Heidelberg, DE), Bernhard Küster (TU München, DE), Kathryn Lilley (University of Cambridge, GB), Frédérique Lisacek (Swiss Institute of Bioinformatics, CH), Lennart Martens (Ghent University, BE), Elien Vandermarliere (Ghent University, BE), Juan Antonio Vizcaino (EBI – Hinxton, GB), and Henrik Zauber (Max-Delbrück-Centrum – Berlin, DE)*

The impact of various biological processes on the coverage of the complete proteoform space varies. The following topics are encountered when considered in order of importance. First is the occurrence of splice isoforms, which can be cell or tissue type specific. In order to study these, it will therefore be important to rely on matching data (e.g., from genomics, transcriptomics, and proteomics). It is also interesting to see that emerging technologies can potentially be used to improve our understanding of isoforms and their annotation (for instance, the nanopore technology).

Another, somewhat related topic is that of ORF delineation. Current approaches do exist, but these are still at a reasonably early stage in development: adding extra unannotated (re-annotated) open reading frames (ORFs) to the search space is probably the most mature. Moreover, these can be supplemented with (potential) alternative start sites. The impact of these additions on identification rate is rather limited, because the database size increase remains modest. At the same time, there is an entire family of potential open reading frames that are currently too small to be picked up by gene prediction algorithms, and these (upstream (uORFs) and/or small ORFs (sORFs)) should also be investigated. An important question is whether these are actually active at the protein level, and if these can thus be picked up by mass spectrometry.

A further expansion of the potential sequence space is conferred by single amino-acid variations, which can even be increased further in the context of disease. While potentially detectable with existing methods (although it will be challenging due to the dependency of detection probability on the abundance of the parent protein, and the ionization potential and possible modification status of the peptide in question), it remains challenging to include the frequency information for these variations. It should also be noted in this context that, while variation is included in databases such as the UniProt KnowledgeBase, frequencies are not included for these variations. When combined with ribosome profiling, the sequence space can be further expanded to include frameshifts as well as stop-codon read-through. Although it should be noted that the occurrence rate of these events may be quite low, and the biological relevance could be quite limited.

Finally, beyond the sequence space, the chemical space can be extended as well, through post-translational modifications (PTMs). Many of these occur frequently, and thus have a large impact on the total proteoform space. Moreover, mass spectrometry remains the primary means of exploring these PTMs, but is in turn hindered by a lack of knowledge on the underlying biological processes and mechanisms that carry out and regulate these modifications, which makes it hard to predict what we could possibly expect. Throughout, a question that remains essentially unanswered, is how to derive biological meaning from results that are obtained from an extension of our coverage.

## 4.4 False Discovery Rate Estimation Issues in Large Database Searches and Proposal of Benchmarking Challenges

*Thilo Muth (Robert Koch Institut – Berlin, DE), Magnus Arntzen (Norwegian University of Life Sciences – As, NO), Sebastian Böcker (Universität Jena, DE), John Cottrell (Matrix Science Ltd. – London, GB), Julien Gagneur (TU München, DE), Laurent Gatto (University of Cambridge, GB), Marco Hennrich (EMBL – Heidelberg, DE), Lukas Käll (KTH – Royal Institute of Technology, SE), Jeroen Krijgsveld (DKFZ – Heidelberg, DE), Phillip Pope (Norwegian University of Life Sciences – As, NO), Hans Vissers (Waters Corporation – Wilmslow, GB), Ana L. Wang (Scripps Research Institute – La Jolla, US), Dennis Wolan (Scripps Research Institute – La Jolla, US), and Henrik Zauber (Max-Delbrück-Centrum – Berlin, DE)*

We first discussed issues of false discovery rate (FDR) estimation for large databases with an emphasis on metaproteomics. Different levels of FDR were recognized, such as peptide, protein, isoform and species FDR. There were concerns regarding the FDR control when using multiple search engine because of different scoring schemes. It was also discussed that the size and completeness of the search space (i.e. spectra, sequences and post-translational modifications (PTMs)) has an influence of FDR at all levels for target-decoy searches. One possibility is to reduce the search space, e.g. by limiting the database to the peptides which can be expected by using custom (metagenome/metatranscriptome) databases. While PSM and peptide FDR have been evaluated thoroughly so far, still no clear consensus can be found on how to assess the protein FDR. Secondly, we proposed to initiate two benchmarking challenges which are open for the community, one for metaproteomics and another for splicing isoforms. For assessing splicing isoforms, two different cell types will be grown and mixture series will be generated. One expects the quantification of isoforms to be proportional to the mixture ratio which allows for benchmarking their linear relationships, e.g. using R-squared of isoform quantity estimates vs. dilution ratio. Participants are requested to estimate isoform quantities for each sample individually. Moreover, transcriptome data are generated for each cell type and methods for MS-based isoform quantification are benchmarked using state-of-the-art RNA-sequencing isoform quantities as ground truth estimates. The metaproteomics challenge consists of three different options, (i) creating a metaproteome of a mock community of known isolates for evaluating peptide/protein/species FDR, (ii) providing a mock communities of different dilution series of variants (rare vs. abundant species) allowing also to assess fold change estimation, (iii) spiking the mock into a complex background community (e.g. with closely related species) to assess the recovery. Sample spectra and database consisting of the whole complex (real + mock) community will be provided to the participant. For the metaproteomics challenge, the connection with CAMI challenge for metagenomics (version 2) will be coordinated. The splicing isoform challenge may be linked to either ABRF (http://www.cosmosid.com/nist-challenge/) or DREAM http://dreamchallenges.org/) challenges.

## 4.5 Visualization of proteomics and multi-omics data

*Magnus Palmblad (Leiden University Medical Center, NL), Magnus Arntzen (Norwegian University of Life Sciences – As, NO), Harald Barsnes (University of Bergen, NO), Ileana M. Cristea (Princeton University, US), Laurent Gatto (University of Cambridge, GB), Lydie Lane (Swiss Institute of Bioinformatics, CH), Bart Mesuere (Ghent University, BE), Thilo Muth (Robert Koch Institut – Berlin, DE), Phillip Pope (Norwegian University of Life Sciences – As, NO), Veit Schwämmle (University of Southern Denmark – Odense, DK), and Olga Vitek (Northeastern University – Boston, US)*

The old saying "a picture is worth a thousand words" probably understates the necessity for appropriate visualization tools in data intensive sciences such as genomics or proteomics. In the breakout session, we contrasted interactive visualizations to explore data with reproducible generation of figures for reports or publications. We discussed the importance of mindful visualization – what is the question to be addressed, is the data available, what kind of transformations are required, and what software should be used? We covered these questions in the contexts of six use cases: (1) influence of PTMs on PPI networks, (2) alignment and visualization of unidentified features across datasets, (3) integrating spatially resolved quantitative omics data, (4) flux analysis integrating time-resolved omics data, (5) metaproteomics with taxonomies down to the strain level, and (6) Mapping PTM crosstalk and proteoforms to structures.

Network visualizations were found to address questions in all use cases. Careful attention should be paid to data representation, including using controlled vocabularies and ontologies for metadata used for the visualizations. Distinction was also made between visualizing many entities (proteins or metabolites) in one experiment versus showing the distribution of few entities across many datasets.

Though many powerful visualization software platforms exist, there is a need for refined tools for displaying PTMs or proteoform information in the context of PPI networks or pathways (use cases 1 and 4), systematic metadata annotation using controlled vocabularies (use cases 3, 4 and 5), and integrating alignment of unidentified LC-MS(/MS) features with study metadata. Network and pathway visualization tools must clearly distinguish between absolute and relative changes in abundance (all use cases) and between no change with no data. Potential pitfalls were also discussed, such as adding information lacking experimental evidence in visualizations and attempting to display too much information in one figure. Sometimes, visualization is more about what to hide than what to show.

## 4.6 Metadata Provision for Public Proteomics Data

*Juan Antonio Vizcaino (EBI – Hinxton, GB), Lydie Lane (Swiss Institute of Bioinformatics, CH), Frédérique Lisacek (Swiss Institute of Bioinformatics, CH), Lennart Martens (Ghent University, BE), Gerben Menschaert (Ghent University, BE), Veit Schwämmle (University of Southern Denmark – Odense, DK), and Mathias Walzer (EBI – Hinxton, GB)*

The value of public data increases with reuse, but such reuse requires proper metadata annotation. Unfortunately, metadata is currently only sparsely available, and mostly remains unstructured. The way to resolve this issue, and thus to add value to public data, revolves around two complementary strategies. The first strategy is to recover the already submitted meta data through post-hoc annotation; this can be achieved by structuring currently unstructured data, or by extracting metadata from in-depth analysis of the data proper. The second startegy is to increase the annotation of submitted proteomics data by the submitter. Importantly, formats already exist that allow metadata to be be structured, and that covers a variety of metadata pertaining to more or less the complete analytical workflow in proteomics. The working group therefore looked into existing solutions and readily available metadata, and reviewed these with respect to tangible applications to put metadata into a structured format.

At the same time, however, data repositories should endeavour to make the added value of metadata availability more visible, and to lower the threshold for entering metadata annotation. This will not only motivate submitters to provide these metadata, but will also enable the efficient annotation of these data.

The overall conclusion was a need for specific tools to annotate metadata, both at the site of the experimentalist, and preferrably in such a way that the user-specified metadata (as opposed to instrument-derived metadata, which tends to be captured more comprehensively and transparently already) is captured even before the project starts. In addition, efforts by 'annotation super users' (researchers who already actively reuse public proteomics data on a large scale) that connect existing data with metadata should be captured for subsequent general reuse.

## 4.7 Computational Proteomics Education

*Pieter-Jan Volders (Ghent University, BE), Harald Barsnes (University of Bergen, NO), Lennart Martens (Ghent University, BE), Bart Mesuere (Ghent University, BE), Magnus Palmblad (Leiden University Medical Center, NL), and Elien Vandermarliere (Ghent University, BE)*

Computational proteomics, and bioinformatics in general, attracts people with different backgrounds such as bio(medical) and computer sciences. We discussed the required skillset and different profiles of people working in the field distinguishing computational scientists, bioinformaticians and biologists. The key aspect is computational thinking. Moreover, bioinformaticians, and scientists in general, need to be taught enough of the neighbouring fields to communicate efficiently with everyone involved in a research project. This is partly reflected in the observation that the boundaries between being a biologist, a bioinformatician and a computer scientist are becoming increasingly vague. Next, we focused on training. Training someone in computational proteomics requires knowledge from (molecular) biology, statistics, computer science, mass spectrometry and general bioinformatics. We thus propose a curriculum with required skills and knowledge from those fields. We compiled a set of guidelines and a repository of online resources for both students and educators that can serve as a basis for designing educational programs in computational proteomics.

## Participants

- Magnus Arntzen
Norwegian University of Life Sciences – As, NO
- Nuno Bandeira
University of California – San Diego, US
- Harald Barsnes
University of Bergen, NO
- Sebastian Böcker
Universität Jena, DE
- Robert Chalkley
UC – San Francisco, US
- John Cottrell
Matrix Science Ltd. – London, GB
- Ileana M. Cristea
Princeton University, US
- Bernard Delanghe
Thermo Fisher GmbH – Bremen, DE
- Eric Deutsch
Institute for Systems Biology – Seattle, US
- Viktoria Dorfer
University of Applied Sciences Upper Austria, AT
- Julien Gagneur
TU München, DE
- Laurent Gatto
University of Cambridge, GB
- Marco Hennrich
EMBL – Heidelberg, DE
- Nico Jehmlich
UFZ – Leipzig, DE
- Lukas Käll
KTH – Royal Institute of Technology, SE
- Oliver Kohlbacher
Universität Tübingen, DE
- Jeroen Krijgsveld
DKFZ – Heidelberg, DE
- Bernhard Küster
TU München, DE
- Lydie Lane
Swiss Institute of Bioinformatics – Genève, CH
- Kathryn Lilley
University of Cambridge, GB
- Frédérique Lisacek
Swiss Institute of Bioinformatics – Genève, CH
- Lennart Martens
Ghent University, BE
- Gerben Menschaert
Ghent University, BE
- Bart Mesuere
Ghent University, BE
- Thilo Muth
Robert Koch Institut – Berlin, DE
- Magnus Palmblad
Leiden University Medical Center, NL
- Phillip Pope
Norwegian University of Life Sciences – As, NO
- Hannes Röst
University of Toronto, CA
- Timo Sachsenberg
Universität Tübingen, DE
- Veit Schwämmle
University of Southern Denmark – Odense, DK
- Stephen Tate
SCIEX – Concord, CA
- Elien Vandermarliere
Ghent University, BE
- Hans Vissers
Waters Corporation – Wilmslow, GB
- Olga Vitek
Northeastern University – Boston, US
- Juan Antonio Vizcaino
EBI – Hinxton, GB
- Pieter-Jan Volders
Ghent University, BE
- Mathias Walzer
EBI – Hinxton, GB
- Ana L. Wang
Scripps Research Institute – La Jolla, US
- Mathias Wilhelm
TU München, DE
- Paul Wilmes
University of Luxembourg, LU
- Dennis Wolan
Scripps Research Institute – La Jolla, US
- Henrik Zauber
Max-Delbrück-Centrum – Berlin, DE