Death by data – condensing data for proteins

The communication of knowledge – especially in the life sciences and medicine – is driven by the accessible graphical representation of concepts, data, and results. With the advancements in technology for protein measurements the amount of data to handle and present continues to grow exponentially. Many **representations** have established themselves as de facto standards while new ways to handle, condense and visualize the ever-increasing complexity of the data are emerging. Even a whole research field has developed around the optimization of data visualization.

In this project, we are going to use and build computational tools to analyse, compare, condense, and visualize molecular data. Combining the different methods of an analysis workflow into a comprehensive, easy-to-use toolbox and expanding it with new analysis techniques with machine learning approaches or methods commonly applied in different research fields is the focus of this project. We will largely work in R and Python and refine, expand, and develop methods to handle, analyse, condense, and represent.

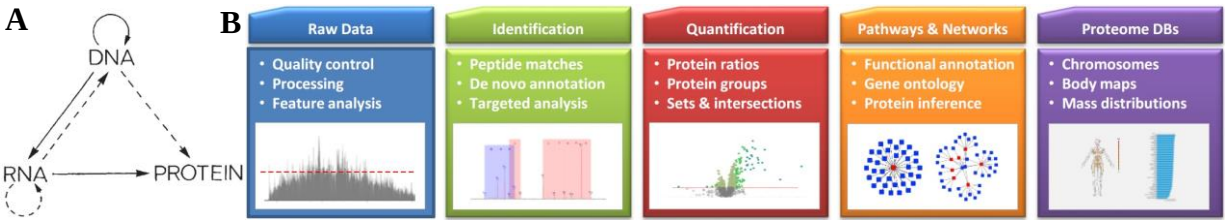


Figure 1. A) Central dogma for molecular biology. DNA as the blueprint translation into proteins through the intermediate RNA. (Adapted from Crick F., 1970, Nature 227 (5258): p. 562) B) Areas of visualization in a proteomic mass spectrometry workflow. (Adapted from Overland E. et al., 2015, Proteomics 15: p. 1341)

The basis of the analysis and visualization are proteomics in-house and public datasets from the PRIDE data repository alongside common and specialized graphical representations. The project starts with compiling visualization ideas and identifying the associated data handling methods. These methods range from data normalization, imputation, analysis, and machine learning approaches (see Figure 1B). This is followed by applying the chosen methods and visualizations to the example datasets before streamlining them into an easy-to-use library/R-package and an adding a graphical user interface to reach a wide userbase.

We will deal with proteomic mass spectrometry at different levels of granularity and at different stages of the analysis process. Therefore, different additional perspectives could be explored.

1. Bides the standard analysis approaches to streamline the visualization the toolbox can be expanded to incorporate enhanced algorithms and **machine learning** for various aspects such as sped-up peptide-protein assignment, advanced classification of healthy and diseased samples using lower and higher dimensional data.
2. **Integration** with data from other molecular levels, e.g., **genomics and transcriptomics** measurements, is an emerging field of research focusing on modelling the cell as a whole.

Using visualization methods from these other research fields enables communication across fields easier. Bringing proteomics data into formats meant for other omics tools allows for a whole set of new analysis methods and visualizations to be added to the toolbox.

Infobox: Biological background

In human body cells, different types of molecules act together and determine cellular behaviour. While DNA containing genetic information gives the overall blueprint, RNA provides a specialized set of instructions on how proteins should be built. This is called the **central dogma** of molecular biology (see Figure 1A). Proteins then can interact in different ways to determine health and disease. Proteomic mass spectrometry is a technology to analyse proteins. Whole proteins are digested into smaller entities called peptides. These peptides are fragmented, and the fragments are measured using mass spectrometers, million dollar highly sensitive scales. Many thousands of proteins are measured and used to compare disease tissue with healthy tissue to identify biomarkers that can help with diagnostics and treatment.
With ever increasing ways to analyse the resulting measurements and proteins, graphical representations have been amended and evolved over the last 30 years while others are deeply ingrained in the standard sets of graphics that help biologists understand these data and draw their conclusions from them (see Figure 2A). To push forward easy communication of proteomic mass spectrometry data, we aim to compare build a comprehensive toolbox for analyses and graphical representations.

3. Finally, how data can be **presented in new** and easily understandable ways is an open and ever developing question that could be pursued. Thereby, different options how information can be prepared and presented through novel still and interactive processes can be developed (see Figure 2B).

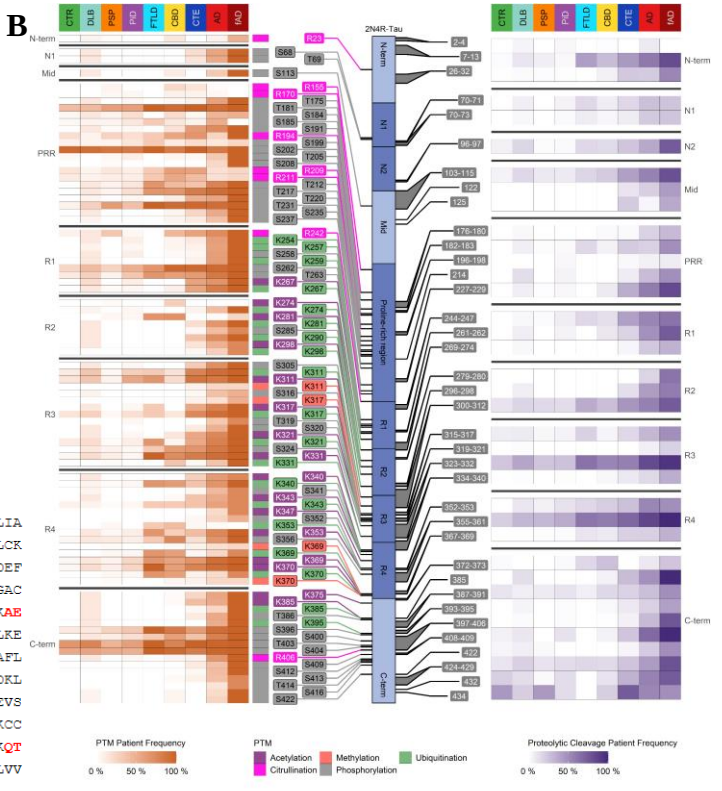
Figure 2. A) Standard representation of the coverage of a protein sequence through peptides identified in an experiment. It highlights the identified peptides in red while representing the overall protein sequence. B) Advanced representation of protein domain structure in the middle in blue alongside identities (color coded PTMs) and sites within the protein sequence across different diseases. This highlights the prevalence of these changes in each disease group. © Christoph Schlaffner, unpublished work.

A

Protein sequence coverage: 8%

Matched peptides shown in ***bold red***.

```
  1 MKWVTFISLL LLFSSAYSRG VFRRDTHKSE IAHRFKDLGE EHFKGLVLIA
 51 FSQYLQQCPF DEHVKLVNEL TEFAKTCVAD ESHAGCEKSL HTLFGDELCK
101 VASLRETYGD MADCCEKQEP ERNECFLSHK DDSPDLPKLK PDPNTLCDEF
151 KADEKKFWGK YLYEIARRHP YFYAPELLYY ANKYNGVFQE CCQAEDKGAC
201 LLPKIETMRE KVLASSARQR LRCASIQKFG ERALKAWSVA RLSQKFPKAE
251 FVEVTKLVTD LTKVHKECCH GDLLECADDR ADLAKYICDN QDTISSKLKE
301 CCDKPLLEKS HCIAEVEKDA IPENLPPLTA DFAEDKDVCK NYQEAKDAFL
351 GSFLYEYSRR HPEYAVSVLL RLAKEYEATL EECCAKDDPH ACYSTVFDKL
401 KHLVDEPQNL IKQNCDQFEK LGEYGFQNAL IVRYTRKVPQ VSTPTLVEVS
451 RSLGKVGTRC CTKPESERMP CTEDYLSLIL NRLCVLHEKT PVSEKVTKCC
501 TESLVNRRPC FSALTPDETY VPKAFDEKLF TFHADICTLP DTEKQIKKQT
551 ALVELLKHKP KATEEQLKTV MENFVAFVDK CCAADDKEAC FAVEGPKLVV
601 STQTALA
```



Why should you join this project?

You will deal with a wide range of datasets, their research application, subject specific algorithms, and their graphical representation. There are many potential avenues this project can take, and we can adapt to your interest.

You will work with real-world molecular data from proteomic mass spectrometry and provide the graphical tools so questions and solutions on prevailing molecular mechanisms can be widely and easily communicated. You will not only provide a useful tool set for biologists across the globe but will also likely have a lasting impact on the scientific way of communicating results. Exploring new ways to let the data tell their story is always an option!

If you are familiar with R and Python and proficient in at least one of the languages, and interested in data visualization, data analysis methods and helping biologists with a useful tool set, this project is for you.

Look forward to an engaged supervision by Prof. Dr. Bernhard Renard (bernhard.renard@hpi.de) and Dr. Christoph Schlaffner (christoph.schlaffner@hpi.de) in your upcoming master project! Please do not hesitate to contact us in case of questions.

Further reading

Representation of data and the basics (Midway 2020)
Central dogma of molecular biology (Crick 1970)
Machine learning in proteomics e.g. (Swan et al. 2013; Suvarna et al. 2021)
Integration with other omics data e.g. (Schlaffner et al. 2017; Schlaffner 2018)
Representing proteomics data e.g (Overland et al. 2015; Gatto et al. 2015; Cheng et al. 2018)

Midway S.R. (2020). „Principles of Effective Data Visualization." Patterns **1(9)**.
Crick F. (1970). "Central Dogma of Molecular Biology." Nature **227(5258)**.
Swan A.L., Mobasheri A., Allaway D., Liddell S. and Bacardit J. (2013). "Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology." OMICS **17(12)**.
Suvarna K., Biswas D., Pai M.G.J., Acharjee A., Bankar R., Palanivel V., Salkar A., Verma A., Mukherjee A., Choudhary M., Ghantasala S., Ghosh S., Singh A., Banerjee A., Badaya A., Bihani S., Loya G., Mantri K., Burli A., Roy J., Srivastava A., Agrawal S., Shrivastav O., Shastri J. and Srivastava S. (2021). "Proteomics and Machine Learning Approaches Reveal a Set of Prognostic Markers for COVID-19 Severity With Drug Repurposing Potential." Frontiers in Physiology **12**.
Schlaffner C.N., Pirklbauer G.J., Bender A. and Choudhary J.S. (2017). "Fast, Quantitative and Variant Enabled Mapping of Peptides to Genomes." Cell Systems **5(2)**.
Schlaffner C.N. (2018). "Proteogenomics for Personalised Molecular Profiling." Doctoral thesis
Overland E., Muth T., Rapp E., Martens L., Berven F.S. and Barsens H. (2015). "Viewing the proteome: How to visualize proteomcis data?." Proteomics **15**.
Gatto L., Breckels L.M., Naake T. and Gibb S. (2015). "Visualization of proteomics data using R and Bioconductor." Proteomics **15(8)**.
Cheng C., Xu K., Guo C., Wang J., Yan Q., Zhang J., He F. and Zhu Y. (2018). "PANDA-view: and easy-to-use tool for statistical analysis and visualization of quantitative proteomics data." Bioinformatics **34(20)**.