A Practical Beginner's Guide to Proteomics

This manuscript (permalink) was automatically generated from jessegmeyerlab/proteomics-tutorial@fed79f4 on January 26, 2022.

Authors

- **Dina Schuster**
  0000-0001-6611-8237 · dschust-r · dina_sch
  Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich 8093, Switzerland; Department of Biology, Institute of Molecular Biology and Biophysics, ETH Zurich, Zurich 8093, Switzerland; Laboratory of Biomolecular Research, Division of Biology and Chemistry, Paul Scherrer Institute, Villigen 5232, Switzerland

- **Jesse G. Meyer**
  0000-0003-2753-3926 · jessegmeyerlab · j_my_sci
  Department of Biochemistry, Medical College of Wisconsin · Funded by Grant R21 AG074234; Grant R35 GM142502

Abstract

Proteomics is the large-scale study of protein structure and function from biological systems. “Shotgun proteomics” or “bottom-up proteomics” is the prevailing strategy, in which proteins are hydrolyzed into peptides that are analyzed by mass spectrometry. Proteomics studies can be applied to diverse studies ranging from simple protein identification to studies of protein-protein interactions, post-translational modifications, and protein stability. To enable this range of different experiments, there are diverse strategies for proteome analysis. The nuances of how proteomics workflows differ may be challenging to understand for new practitioners. Here, we provide a comprehensive tutorial of different proteomics methods. Our tutorial covers all necessary steps starting from protein extraction and ending with biological interpretation. We expect that this work will serve as a basic resource for new practitioners of the field of shotgun or bottom-up proteomics.

Introduction

Proteomics is the large scale study of protein structure and function. Proteins are translated from mRNAs that are transcribed from the genome. Although the genome encodes potential cellular functions and states, the study of proteins is necessary to truly understand biology. Currently, proteomic studies are facilitated by mass spectrometry, although alternative methods are being developed.

Modern proteomics started around the year 1990 with the introduction of soft ionization methods that enabled, for the first time, transfer of large biomolecules into the gas phase without destroying them [[1]][2]. Shortly afterward, the first computer algorithm for matching peptides to a database was introduced [3]. Another major milestone that allowed identification of over 1000 proteins were actually improvements to chromatography [4]. As the volume of data exploded, methods for statistical analysis transitioned use from the wild west to modern informatics based on statistical models [5] and the false discovery rate [6].

Two strategies of mass spectrometry-based proteomics differ fundamentally by whether proteins are cleaved into peptides before analysis: “top-down” and “bottom-up”. Bottom-up proteomics (also refered to as shotgun proteomics) is defined by the hydrolysis of proteins into peptide pieces [7]. Therefore, bottom-up proteomics does not actually measure proteins, but must infer their presence [8]. Sometimes proteins are infered from only one peptide sequence representing a small fraction of the total protein sequence predicted from the genome. In contrast, top-down proteomics attempts to measure all proteins intact [9]. The potential benefit of top-down proteomics is the ability to measure proteoforms [10]. However, due to analytical challenges, the depth of protein coverage that is achievable by top-down proteomics is less than the depth that is achievable by bottom-up proteomics.

In this tutorial we focus on the bottom-up proteomics workflow. The most common version of this workflow is generally comprised of the following steps. First, proteins in a biological sample must be extracted. Usually this is done by denaturing and solubilizing the proteins while disrupting DNA and tissue. Next, proteins are hydrolyzed into peptides, usually using a protease like trypsin. Peptides from proteome hydrolysis must be purified. Most often this is done with reversed phase chromatography cartridges or tips. The peptides are then almost always separated by liquid chromatography before they are ionized and introduced into a mass spectrometer. The mass spectrometer then collects precursor and fragment ion data from those peptides. The data analysis is usually the rate limiting step. Peptides must be identified, and proteins are inferred and quantities are assigned. Changes in proteins across conditions are determined with statistical tests, and results must be interpreted in the context of the relevant biology.

There are many variations on this workflow. The wide variety of experimental goals that are achievable with proteomics technology leads to a wide variety of potential proteomics workflows. Even choice is important and every choice will affect the results. In this tutorial, we cover all of the required steps in detail to serve as a tutorial for new proteomics practioners.

1. Types of experiments enabled by proteomics
2. Protein extraction
3. proteolysis
4. Isotopic Labeling
5. Enrichments
6. Peptide purification
7. Mass Spectrometry
8. Peptide Ionization

9. Data Acquisition
10. Basic Data Analysis
11. Biological Interpretation
12. Experimental considerations and design

Types of Experiments

A wide range of questions are addressable with proteomics technology, which translates to a wide range of variations of proteomics workflows. Sometimes identifying what proteins are present is desired, and sometimes the quantities of as many proteins as possible are desired. Proteomics experiments can be both qualitative and quantitative.

Qualitative experiments

- Identifying proteins
- Identifying post translational modifications
- Identifying protein isoforms

Quantitative experiments

- Protein abundance changes
- Phosphoproteomics
- Glycoproteomics
- Structural techniques (XL-MS, HDX-MS, FPOP, protein-painting, LiP-MS, radical footprinting, ion mobility)
- Protein stability and small molecule binding (Thermal proteome profiling, TPP, or cellular thermal shift assay, CETSA)
- Protein-protein interactions (PPIs): AP-MS, APEX, BioID

Protein Extraction

First, proteins must be isolated from the sample matrix. Because some proteins alter other proteins, the goal is to simultaneously solubilizw and denature proteins. This is achieved with a combination of salt and chaotropic agent.

1. Choice of Lysis buffer

- Urea can cause chemical modifications

2. Sample type and homogenisation methods

- specialised sample preparation protocols for non-denaturing protein isolation (i.e. for LiP-MS, HDMX-MS etc)

4. chemicals to avoid: PEGs, detergents etc
5. removal of contaminations, Protein Precipitation

- detergent reomoval resins, S-TRAP (Protifi) etc

7. protein alkylation

- choices of reduction and alkylation reagents, TCEP/DTT/2BME, Chloroacetamide/iodoacetamide, n-ethyl maleimide

Proteolysis

Proteolysis is the defining step that differentiates bottom-up or shotgun proteomics from top-down proteomics. Hydrolysis of proteins is extremely important because it defines the population of potentially identifyable peptides. Generally peptides between a length of 7-35 amino acid are considered useful for mass spectrometry analysis. Peptides that are too long are difficult to identify by tandem mass spectrometry, or may be lost during sample preparation due to Peptides that are too short are less likely to uniquely match to a single protein. There are many choices of enzymes and chemicals that hydrolyze proteins into peptides. This section summarizes potential choices and their strengths and weaknesses.

Trypsin is the most common choice of protease for proteome hydrolysis [11].

3. theoretical studies of proteolysis enzymes [12]
4. Challenges associated with alternative enzyme choices (non-specific and semi-specific enzymes)
5. Alternative enzyme choices (one paragraph each?) - LysC
6. GluC
7. AspN
8. Alpha-lytic protease [13] and how it enables mapping human SUMO sites [14].
9. others?

Peptide and Protein Labeling

Discussion of methods to isotopically label peptides or proteins that enable quantification

1. SILAC/SILAM
2. iTRAQ
3. TMT
4. dimethyl labeling

Peptide or Protein Enrichment

Protein enrichment (e.g. for protein protein interactions)

- coIP
- APEX
- bioID
- bioplex

Peptide enrichment

- antibody enrichments of modifications, e.g. lysine acetylation [15].
- TiO2 and Fe enrichment of phosphorylation
- Glycosylation
- SISCAPA

Methods for Peptide Purification

1. Reverse phase including tips and cartridges
2. stage tips
3. in stage tip (iST)
4. SP2, SP3
5. s traps

Types of Mass Spectrometers used for Proteomics

1. QQQ
2. Q-TOF
3. Q-Orbitrap
4. LTQ-Orbitrap
5. TOF/TOF
6. FT-ICR
7. types of ion mobility

- SLIM
- FAIMS
- traveling wave
- tims

Peptide Ionization

The 2002 Nobel Prize in Chemistry was awarded to partially to John Fenn and Koichi Tanaka “for their development of soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules” [16/].

MALDI

Electrospray Ionization

Data Acquisition

Data acquisition strategies for proteomics fall generally within targeted or untargeted, and they can depend on the data (data dependent acquisition or DDA) or be data independent (data-independent acquisition or DIA).

DDA

Targeted DDA

Untargeted DIA

DIA

Targeted DIA

Untargeted DIA

Analysis of Raw Data

The goal of basic data analysis is to convert raw spectral data into identities and quantities of peptides and proteins that can be used for biologically-focused analysis. This step may often include measures of quality control, cross-run data normalization, quantification on different levels (precursor, peptide, protein), protein inference, PTM (post translational modification) localization and also first steps of data analysis, such as statistical hypothesis tests.

In typical bottom-up proteomics experiments, proteins are digested into peptides and further analyzed with LC-MS/MS systems. Peptides can have different PTMs and ionize differently depending on their length and amino acid distributions. Therefore, mass spectrometers often record different charge and modification states of one single peptide. The entity that is recorded on a mass spectrometer is usually referred to as a precursor ion (peptide with its modification and charge state). This precursor ion is fragmented and the precursor or peptide sequences are obtained though spectral matching. The quantity of a precursor is estimated with various methods. The measured precursor quantities are combined to generate a peptide quantity. Peptides are also often combined into a protein group through protein inference, which combines multiple peptide identifications into a single protein identification [17] [18]. Protein inference is still a challenge in bottom-up proteomics.

Due to the inherent differences in the data structures of DDA and DIA measurements, there exist different types of software that can facilitate the steps mentioned above. The existing software for DDA and DIA analysis can be further divided into freeware and non-freeware:

DDA freeware

Name	Publication	Website
MaxQuant	Cox and Mann, 2008[19]	MaxQuant
MSFragger	Kong et al., 2017[20]	MSFragger
Mascot	Perkins et al., 1999[21]	Mascot
MS-GF+	Kim et al., [22]	MS-GF+

DIA freeware:

Name	Publication	Website
MaxDIA	Cox and Mann, 2008[19]	MaxQuant
Skyline	MacLean et al., 2010[23]	Skyline
DIA-NN	Demichev et al., 2019[24]	DIA-NN

Targeted proteomics freeware:

Name	Publication	Website
Skyline	MacLean et al., 2010[23]	Skyline

DDA non-freeware:

Name	Publication	Website
ProteomeDiscoverer		ProteomeDiscoverer
Mascot	Perkins et al., 1999[21]	Mascot
Spectromine		Spectromine
PEAKS	Tran et al., 2018[25]	PEAKS

DIA non-freeware:

Name	Publication	Website
Spectronaut	Bruderer et al., 2015[26]	ProteomeDiscoverer
PEAKS	Tran et al., 2018[25]	PEAKS

Analysis of DDA data

Strategies for analysis of DIA data

Targeted proteomics data analysis

Quality control

Statistical hypothesis testing

Biological Interpretation

1. term enrichment analysis (KEGG, GO)
2. network analysis methods
3. structure analysis
4. isoform analysis
5. follow-up experiments

Experiment Design

This section should discuss trade offs and balancing them to design an experiment. 1. constraints: Each experiment will have different constraints, which may include the number of samples needed for analysis, or desire to quantify a specific subset of proteins within a sample. 2. sample size 3. statistics 4. costs

References

1. **Electrospray Ionization for Mass Spectrometry of Large Biomolecules**
John B Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, Craig M Whitehouse
Science (1989-10-06) https://doi.org/cq2q43
DOI: 10.1126/science.2675315 · PMID: 2675315

2. **Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry**
Koichi Tanaka, Hiroaki Waki, Yutaka Ido, Satoshi Akita, Yoshikazu Yoshida, Tamio Yoshida, T Matsuo
Rapid Communications in Mass Spectrometry (1988) https://doi.org/ffbwwr
DOI: https://doi.org/10.1002/rcm.1290020802

3. **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.**
JK Eng, AL McCormack, JR Yates
Journal of the American Society for Mass Spectrometry (1994-11)
https://www.ncbi.nlm.nih.gov/pubmed/24226387
DOI: 10.1016/1044-0305(94)80016-2 · PMID: 24226387

4. **An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics**
Dirk A Wolters, Michael P Washburn, John R Yates
Analytical Chemistry (2001-10-25) https://doi.org/bn4kq6
DOI: 10.1021/ac010617e · PMID: 11774908

5. **A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry**
Alexey I Nesvizhskii, Andrew Keller, Eugene Kolker, Ruedi Aebersold
Analytical Chemistry (2003-07-15) https://doi.org/b2xv45
DOI: 10.1021/ac0341261 · PMID: 14632076

6. **Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry**
Joshua E Elias, Steven P Gygi
Nature Methods (2007-03) https://doi.org/djz7fz
DOI: https://doi.org/10.1038/nmeth1019

7. **Mass-spectrometric exploration of proteome structure and function**
Ruedi Aebersold, Matthias Mann
Nature (2016-09) https://doi.org/f83zqm
DOI: 10.1038/nature19949 · PMID: 27629641

8. **A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry**
* Alexey I. Nesvizhskii, * Andrew Keller, ‡ and Eugene Kolker, Ruedi Aebersold
ACS Publications (2003-07-15) https://pubs.acs.org/doi/abs/10.1021/ac0341261

9. **High-throughput quantitative top-down proteomics**
Kellye A Cupp-Sutton, Si Wu
Molecular Omics (2020) https://doi.org/gnx98p
DOI: 10.1039/c9mo00154a · PMID: 31932818 · PMCID: PMC7529119

10. **Proteoforms as the next proteomics currency**
Lloyd M Smith, Neil L Kelleher

Science (2018-03-09) https://doi.org/gn6p4x
DOI: 10.1126/science.aat1884 · PMID: 29590032 · PMCID: PMC5944612

11. **Getting intimate with trypsin, the leading protease in proteomics**
Elien Vandermarliere, Michael Mueller, Lennart Martens
Mass Spectrometry Reviews (2013-06-15) https://doi.org/gn64qb
DOI: 10.1002/mas.21376 · PMID: 23775586

12. **In Silico Proteome Cleavage Reveals Iterative Digestion Strategy for High Sequence Coverage**
Jesse G Meyer
ISRN Computational Biology (2014-04-22) https://doi.org/gb6s2r
DOI: https://doi.org/10.1155/2014/960902

13. **Expanding Proteome Coverage with Orthogonal-specificity α-Lytic Proteases**
Jesse G Meyer, Sangtae Kim, David A Maltby, Majid Ghassemian, Nuno Bandeira, Elizabeth A Komives
Molecular & Cellular Proteomics (2014-03) https://doi.org/f5vgcg
DOI: https://doi.org/10.1074/mcp.m113.034710

14. **Site-specific identification and quantitation of endogenous SUMO modifications under native conditions.**
Ryan J Lumpkin, Hongbo Gu, Yiying Zhu, Marilyn Leonard, Alla S Ahmad, Karl R Clauser, Jesse G Meyer, Eric J Bennett, Elizabeth A Komives
Nature communications (2017-10-27) https://www.ncbi.nlm.nih.gov/pubmed/29079793
DOI: 10.1038/s41467-017-01271-3 · PMID: 29079793 · PMCID: PMC5660086

15. **Simultaneous Quantification of the Acetylome and Succinylome by 'One-Pot' Affinity Enrichment**
Nathan Basisty, Jesse G Meyer, Lei Wei, Bradford W Gibson, Birgit Schilling
PROTEOMICS (2018-08-19) https://doi.org/gn4cmb
DOI: 10.1002/pmic.201800123 · PMID: 30035354 · PMCID: PMC6175148

16. **The Nobel Prize in Chemistry 2002**
NobelPrize.org
https://www.nobelprize.org/prizes/chemistry/2002/summary/

17. **Interpretation of Shotgun Proteomic Data**
Alexey I Nesvizhskii, Ruedi Aebersold
Molecular & Cellular Proteomics (2005-10) https://doi.org/cm99cj
DOI: https://doi.org/10.1074/mcp.r500012-mcp200

18. **In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics**
Enrique Audain, Julian Uszkoreit, Timo Sachsenberg, Julianus Pfeuffer, Xiao Liang, Henning Hermjakob, Aniel Sanchez, Martin Eisenacher, Knut Reinert, David L Tabb, … Yasset Perez-Riverol
Journal of Proteomics (2017-01) https://doi.org/f9r8r6
DOI: 10.1016/j.jprot.2016.08.002 · PMID: 27498275

19. **MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification**
Jürgen Cox, Matthias Mann
Nature Biotechnology (2008-11-30) https://doi.org/crn24x
DOI: 10.1038/nbt.1511 · PMID: 19029910

20. **MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics**
Andy T Kong, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu, Alexey I Nesvizhskii
Nature Methods (2017-04-10) https://doi.org/f9z6p7
DOI: 10.1038/nmeth.4256 · PMID: 28394336 · PMCID: PMC5409104

21. **Probability-based protein identification by searching sequence databases using mass spectrometry data.**
DN Perkins, DJ Pappin, DM Creasy, JS Cottrell
Electrophoresis (1999-12) https://www.ncbi.nlm.nih.gov/pubmed/10612281
DOI: 10.1002/(sici)1522-2683(19991201)20:18<3551::aid-elps3551>3.0.co;2-2 · PMID: 10612281

22. **MS-GF+ makes progress towards a universal database search tool for proteomics**
Sangtae Kim, Pavel A Pevzner
Nature Communications (2014-10-31) https://doi.org/ggkdq8
DOI: 10.1038/ncomms6277 · PMID: 25358478 · PMCID: PMC5036525

23. **Skyline: an open source document editor for creating and analyzing targeted proteomics experiments**
Brendan MacLean, Daniela M Tomazela, Nicholas Shulman, Matthew Chambers, Gregory L Finney, Barbara Frewen, Randall Kern, David L Tabb, Daniel C Liebler, Michael J MacCoss
Bioinformatics (2010-02-09) https://doi.org/bqx9rq
DOI: 10.1093/bioinformatics/btq054 · PMID: 20147306 · PMCID: PMC2844992

24. **DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput**
Vadim Demichev, Christoph B Messner, Spyros I Vernardis, Kathryn S Lilley, Markus Ralser
Nature Methods (2019-11-25) https://doi.org/gj9xgj
DOI: 10.1038/s41592-019-0638-x · PMID: 31768060 · PMCID: PMC6949130

25. **Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry**
Ngoc Hieu Tran, Rui Qiao, Lei Xin, Xin Chen, Chuyi Liu, Xianglilan Zhang, Baozhen Shan, Ali Ghodsi, Ming Li
Nature Methods (2018-12-20) https://doi.org/gftvmn
DOI: 10.1038/s41592-018-0260-3 · PMID: 30573815

26. **Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues**
Roland Bruderer, Oliver M Bernhardt, Tejas Gandhi, Saša M Miladinović, Lin-Yang Cheng, Simon Messner, Tobias Ehrenberger, Vito Zanotelli, Yulia Butscheid, Claudia Escher, … Lukas Reiter
Molecular & Cellular Proteomics (2015-05) https://doi.org/f7b76h
DOI: 10.1074/mcp.m114.044305 · PMID: 25724911 · PMCID: PMC4424408