Data handling and processing in proteomics


Michael Hamacher, PhD
Author for correspondence
Lead Discovery Center GmbH, Emil-Figge-Strasse 76a, D-44227 Dortmund, Germany
Tel.: +49 231 9742 7012
hamacher@lead-discovery.de

Christian Stephan, PhD
Medizinisches Proteom-Center (MPC), Ruhr-Universität Bochum, ZKF E 143, Universitätstrasse 150, D-44801 Bochum, Germany
Tel.: +49 234 322 9288
christian.stephan@rub.de

Helmut E Meyer
Professor, Medizinisches Proteom-Center (MPC), Ruhr-Universität Bochum, ZKF E 143, Universitätstrasse 150, D-44801 Bochum, Germany
Tel.: +49 234 322 9264
helmut.e.meyer@rub.de

Martin Eisenacher, PhD
Medizinisches Proteom-Center (MPC), Ruhr-Universität Bochum, ZKF E 143, Universitätstrasse 150, D-44801 Bochum, Germany
Tel.: +49 234 322 9288
martin.eisenacher@rub.de

“...by providing all suitable proteomics data to public repositories and by subsequent reprocessing of these datasets, one also can obtain more information than from the single experiment, especially when combining datasets...”

Hightech comes with a scourge: large amounts of data

Proteomics, gel-based and, especially, liquid chromatography (LC)-based techniques are becoming more and more high-tech and GxP driven, namely good manufacturing practice (GMP) and good laboratory practice (GLP), for example. Inspired by the leading genomics field, proteomics has caught up rapidly in terms of elaborate strategies and techniques, ranging from Joachim Klose’s 2D gel systems in the early 1970s to recent (quantitative) LC mass spectrometry approaches [1–3]. Ultraprecise mass spectrometers were introduced several years ago and are now used widely in promising proteomics approaches. Regardless of the strategies and aims that these instruments are used for, they primarily produce enormous amounts of data. Depending on the complexity of the sample, the raw data rate of Thermo’s Orbitrap is up to 100 MB/h, whereas Bruker Daltonics’ Micro ToFQ or maXis produces up to 500 MB/h. Compared with these figures, a High-Capacity Trap (HCT) Ultra rate of 50 MB/h is tiny. Assuming that the mass spectra devices are running 24 h per day, 7 days per week, this leads to 17–84 GB of raw data per week.

As can be deduced already from this simple arithmetic example, data storage, processing and interpretation should never be underestimated, although it seems to be a tedious common process. An initial optimal, standardized design of the data management and reprocessing system is one of the key points for successful research.

Standardized formats

Owing to the complexity of data handling, it is always advisable to gather information and guidelines from current standardization initiatives, in addition to avoiding the repetition of work and mistakes. The most well-known representatives of these initiatives are the Human Proteome Organisation (HUPO) and the EU-funded consortium Proteomics Data Collection (ProDaC). HUPO, founded in 2001 [4], aims for the optimization and spreading of proteomics techniques, demonstrating their potential [101]. In the beginning, it became clear that the HUPO Proteomics Standards Initiative would have to define community standards for data representation in proteomics to facilitate data comparison, exchange and verification [5]. The HUPO Proteomics Standards Initiative has already elaborated various reporting guidelines (Minimum Information About a Proteomics Experiment), data-exchange formats and controlled vocabularies, ranging from drafts to community-reviewed published specification documents [102].

“...data storage, processing and interpretation should never be underestimated...”

The close connection of standards to the researchers was the focus of the ProDaC consortium within the 6th EU Framework Programme [103]. The main goals of this consortium, which linked academic researchers with industry and community journals [6,7], were to support the Proteomics Standards Initiative’s efforts in relation to standards definition, to offer end user-friendly,
standards-compliant data converters for laboratory scientists and other nonbioinformaticians [104], and to start a community-based data collection in public proteomics repositories. The results of most or, if possible, all proteomics studies can now easily be stored in these repositories, initially in a private manner. After publication in a journal, they can be made publicly available for the community. This allows further studies and, hopefully, prevents the rediscovery of already known results. Therefore, it is highly advisable to use these freely available tools and standards, as results will become more comparable and the handling of one’s own data will be much easier.

Data storage solutions & tools
The solutions for data handling are as diverse as the distinct approaches. Regarding data storage solutions, strategies range from stand-alone computers to server-based data banks, from local (intranet) databases to central repositories, from passive approaches to intelligent, semi-automated workflow support systems. Examples for integrated laboratory information management systems include ProteinScape (Bruker Daltonik GmbH, Bremen, Germany) [105], Mascot Integra (Matrix Science, London, UK) [106], Computational Proteomics Analysis System [8,107], Proteus (GenoLogics Life Sciences Software, Victoria, BC, Canada) [108], Proteios [9,109], ms_lims [10,110], Protecs (Decodon, Greifswald, Germany) [111] and SBEAMS [112], among others. Composition and origin of central public repositories are also manifold, for example, PRIDE [11,113], PeptideAtlas [12,114], ProteomeCommons Tranche [13,115] and World-2DPAGE [14,116].

“Millions of datasets and hidden treasures are waiting for further exploitation!”

The intriguing chance of a broad data collection lies in the fact that the results of the countless studies offer highly valuable additional information beyond the original study, even retrospective, after several years. Storage in central repositories offers multiple possibilities for data comparison and checking, as lists can be generated concerning proteins that have been detected in specific diseases, tissues and areas and that show high statistic validation owing to the high number of independent experiments. With the help of such clusters, classes of proteins can be defined that are specific for a special kind of disease or – also very importantly – show decreased or increased expression in a target group (ongoing project at the Medizinisches Proteom Center [6th Framework Programme, Project Number LSHG-CT-2006-036814]). In summary, it is possible to optimize a study’s output by applying well-designed guidelines and best-suited strategies for data handling and interpretation (please, refer to the voluntarily driven standardization groups). Moreover, by providing all suitable proteomics data to public repositories and by subsequent reprocessing of these datasets, one also can obtain more information than from the single experiment, especially when combining datasets from different ‘omics. Millions of datasets and hidden treasures are waiting for further exploitation!

Financial & competing interests disclosure
ProDaC was funded as a Coordination Action by the European Commission (6th Framework Programme, Project Number LSHG-CT-2006-036814) and has been coordinated by the Medizinisches Proteom Center. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References
Editorial

Data handling & processing in proteomics


Websites

101 Human Proteome Organisation www.hupo.org

102 The HUPO Proteomics Standards Initiative www.psidev.info

103 Proteomics Data Collection www.fp6-prodac.eu

104 Proteomics Data Collection. List of developments www.fp6-prodac.eu/developments


106 Matrix Science www.matrixscience.com/integra.html


109 Proteios www.proteios.org

110 Ghent University. Mass-spectrometry oriented LIMS project http://genesis.ugent.be/ms_lims

111 Decodon. Protecs www.decodon.com/Solutions/Protecs

112 SBEAMS. Proteomics overview www.sbeams.org/Proteomics

113 PRIDE. Proteomics identifications database www.ebi.ac.uk/pride

114 PeptideAtlas www.peptideatlas.org/

115 ProteomeCommons https://proteomecommons.org/tranche


117 Medizinisches Proteome Center. Protein List Comparator www.medizinisches-proteom-center.de/ result_comparison

118 PICR. Protein Identifier Cross-Reference Service www.ebi.ac.uk/Tools/picr


120 Gene Map Annotator and Pathway Profiler www.genmapp.org

121 Medizinisches Proteome Center. Software www.medizinisches-proteom-center.de/software