Research Article
Open Access
CrossMark

A smart way for talking with proteins; proteomics

Abstract

Proteomics is the large-scale analysis of proteins, contributing for understanding of gene function. Functional genomics, proteomics, and even metabolomics are the footsteps of genomics that are useful tool to expand of our knowledge on the biological hierarchy of the transcription, translation, and production of small molecules. However, proteomics is a method for assessing the wide range of information such as the structure, expression, localization, biochemical activity, interactions, post-translational modifications and cellular roles of proteins following protein isolation, digestion and mass spectrometry. Proteomics, as a significant post-genomic tool in the field of science, allows researchers to decipher underlying molecular mechanisms behind different metabolic pathways. Proteomics studies are mostly based on protein identification as using mainly bottom-up approaches such as DDA or MudPIT methods as examples of shotgun proteomics techniques. By using the high throughput mass spectrometer technology, huge output data of peptide spectra has been generated. And this increasing data day by day is able to be analyzed by using wide range of bioinformatics tools and the results or raw data has been storage and shared on publicly available databases. From point of this view, in this review analysis of proteins will be summarized considering our knowledge of biology and bioinformatics.

Volume 5 Issue 3 - 2017

Talip Zengin,[1,2] Sercan Kılıç,[2] Ufuk Yenigün,[2] Serhat Erdoğmuş,[2] Ömür Baysal[1,2]
[1]Department of Molecular Biology and Genetics, Muğla Sıtkı Koçman University, Turkey
[2]Bioinformatics Graduate Program, Muğla Sıtkı Koçman University, Turkey

Correspondence: Ömür Baysal, Department of Molecular Biology and Genetics, Muğla Sıtkı Koçman University, Turkey, Email omurbaysal@mu.edu.tr

Received: February 20, 2016 | **Published:** March 23, 2017

Introduction

Proteomics is a rapidly tool to identify proteins and map their interactions in a cellular context[1] and detailed study of proteins on a genome-wide scale as a key technology.[2] These studies relies on dates back to the late 1970s to build databases of proteins using two-dimensional gel electrophoresis,[3] the term "proteomics" was first introduced in 1995[4–6] and defined as the large-scale characterization of the entire protein complement of a cell, tissue, or organism.

Today, two different definitions of proteomics are used; the first one is restricting wide range analysis of gene products to only protein studies that is more classical definition. The second one combines protein studies with analyses that is the large-scale study of proteins, usually by biochemical methods.[3,4,7] A proteomics studies can be divided into the three main steps:

i. The isolation of proteins from a cell line, tissue, or organism.
ii. The acquiring of chemical information of protein for the purposes of protein identification and characterization.
iii. Database utilization.

Expression proteomics, structural proteomics and functional proteomics are the known types of proteomics following protein mapping and characterization, protein-protein interactions and protein function.[1] This method can be divided into three main areas including "protein micro" that can be described as characterization for large-scale identification of proteins with their post-translational modifications "differential display" proteomics for comparison of protein levels with potential application in a wide range of diseases; and studies of protein–protein interactions using techniques such as mass spectrometry or the yeast two-hybrid system. The importance of proteomics in all areas of biomedical and bioscience research should be highlighted as strong evident to understand roles of proteins inside the cell.[1] Since it is a powerful set of tools for the large-scale study of gene function directly at the protein level,[3] proteomics provide a clear assessment on many subjects and can give information on protein bio-markers, relation between protein structure and function after and before the interactions [i.e., functional proteomics]. Many different areas, such as protein-protein interaction studies, protein modifications, protein function, and protein localization studies can be classified under the sub-studies of proteomics to obtain a more global and integrated view of biology rather than focusing on one sample individually (Figure 1).

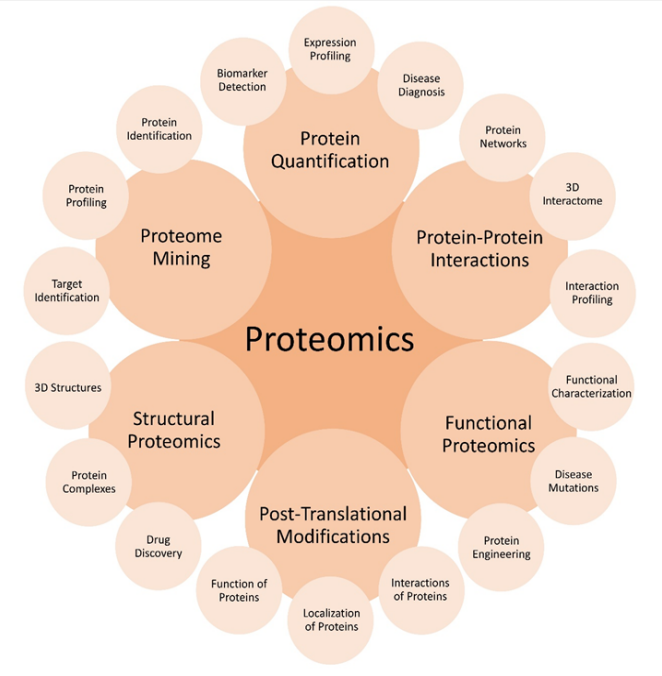

Figure 1 Evaluation fields of Proteomics.

Two different strategies have been used to study the proteome; one is the isolation of specific proteins and analysis of their structure and function by using biochemical and biophysical methods; another is

performing large-scale systematic measurements of proteomes with the computational analysis of proteomic data sets.[8] Both strategies can be performed by powerful mass-spectrometry-based methods which are mostly used for identification and quantification of proteins. It is also possible to identify and localize modified amino acids in peptides; to determine the composition, interaction and topology of sub-units of multi-protein complexes and even their structure by these methods.[8] And present technology can supply the analysis of the complete protein inventory of biological systems, including cell-type-specific proteomes of mammalian organs.[9–11]

The two main approaches to identify and characterize proteins by using mass-spectrometry methods are the "top-down" and "bottom-up" methodologies. At top-down proteomics, proteins are studied as intact by mass spectrometry. Top-down proteomics gives ability to characterize the actual combination of modifications for each proteoform.[12] This approach has the advantage that all modifications on the same molecule can be measured together, enabling identification of the precise proteoform.[13] Although this approach seems to be attractive, top-down mass spectrometry is experimentally and computationally challenging because it is difficult to analyze proteins in comparison with peptides and each protein may have multiple proteoforms that might or might not have same function.[8] However, at bottom-up proteomics, proteins are turned into peptides by the enzymatic digestion. Bottom-up proteomics has been used widely because it is experimentally and computationally more tractable. In all of the techniques of bottom-up proteomics, proteins are extracted from the original source material and then digested into peptides by a sequence-specific protease such as trypsin. The peptides are separated by reverse-phase chromatography and then they are exposed to electrospray ionization. The peptide ions are then transferred to the vacuum of a mass spectrometer, where peptide ions are fragmented in the gas phase to generate MS/MS [MS2] spectra that gives information used for identification and quantification of specific peptides.[8] The resulting data are analyzed by computational pipelines and programs designed for mass-spectrometry.[14] There are many bottom-up techniques which has different purposes, performance profiles and a range of utility. But three main techniques are mostly used: Discovery [or shotgun] proteomics aims to achieve unbiased and complete coverage of the proteome by the method called DDA [Data-Dependent Acquisition]; Targeted proteomics, aims the reproducible, sensitive and streamlined acquisition of a subset of known peptides of interest; and Multiplexed fragmentation of all peptides that are eluted from the high-performance liquid chromatography column by DIA [Data-Independent Acquisition] method, aims to generate comprehensive fragment-ion maps for a sample.[8]

To carry out proteomics study for specific proteins, two-dimensional gel electrophoresis [2D-Gel] has been used and suggests separation method for mass spectrometry to provide analysis of complex protein samples. By this method, proteins within a biological sample are initially resolved by 2D-Gel [first dimension is separation by charge or isoelectric point [pI] and then second dimension obtains separation according to relative molecular mass to fully separate from each other with different physical properties, resulting in reduction of the complexity of protein spots. Individual resolved spots are then extracted from gel and analyzed by mass spectrometry [MS, MS/MS, MALDI-TOF, etc].[15] However, there are many limitations for 2D-Gel; [i] only the most abundant proteins have chance to be identified; [ii] proteins which have extreme pI [<4 or >9] and molecular masses [<15 or >200kDa] cannot be resolved and complex samples are required multiple gels to be resolved; [iii] membrane proteins may not be represented because of poor solubility in sample buffer and the gel; [iv] modified proteins can be visualized as multiple spots; [v] quantification methods are not practical.[16]

Multidimensional protein identification technology [MudPIT][17] as an example of shotgun proteomics, provides a solution to overcome many limitations of 2D-Gel system. MudPIT consists of studies on the resolving of proteins by high-pressure liquid chromatography [HPLC], the peptide analysis by tandem mass spectrometry [MS/MS] and database searching. In this method, a protein sample [either soluble proteins or membrane proteins] is first digested with proteases [trypsin is preferred to produce small peptide mixture]. Then peptides are separated using two different physical characteristics [hydrophobicity and charge] by columns and eluted from the column. They are ionized, isolated according to mass: charge ratio, and fragmented by mass spectrometry. The peptide fragmentation spectrum can be used to identify the protein. Identification realizes on both high- and low-abundance proteins with extremes in pI and molecular mass at low amounts of proteins as well, depending on a sequence database using available softwares.[16]

Applications of proteomics

The aim of proteomics is not only to identify and characterize all expressed proteins in a cell but also to provide a complete atlas of the cell indicating where proteins are located, what functions they perform, which structures they have, how much amount they are existing, which other proteins they interact with and what is the overall phenotype at different time points or situations. Proteome is exceedingly dynamic and continuously variable field in response to cellular or environmental factors, influencing either synthesis or degradation of protein compared to the static genome.[18] Studies on genes alone do not provide much information. Reflection on phenotypes of cells is monitored with focusing on proteins, not genes.

In proteomics area, not all organisms and not whole proteome of them have been able to be studied. Therefore the main studies have been performed on protein mining by using mass spectrometric analysis for a wide range of organisms by targeting specific proteins or whole proteome. The protein mining studies consist of protein identification, protein profiling and target identification. Meanwhile protein quantification studies are also performed by using mass spectrometry and other techniques such as protein microarrays to determine expression profiling of proteins that can also be used for biomarker detection or diagnosis for a disease. Moreover, development of optimal biomarkers for screening and early detection, characterization of the mechanism of disease progression, and predicting the risk of diseases are the positive advantages of ongoing proteomics research[19] (Figure 1).

One of the most important applications of proteomics is the characterization of functions of the proteins -main role players- of the cell. After all functions of proteins are determined, we can understand the total life -screen play- of the cell. But the function and also structure of proteins are affected by static or dynamic post-translational protein modifications therefore modifications on proteins have importance in proteomics research. To elucidate mechanisms of disease, aging, and effects of the environment by studying on the genome is not sufficient. The logic way seems as also studying on proteins with possible modifications for characterization and then targets of drugs can be exactly identified. On the other hand proteins play role in huge interconnectivity creating a protein network, with other proteins in protein complexes, signaling or metabolic pathways

and enzymatic reactions. And the phenotype of the cell is based on actual result of performance of this protein network. In this point, proteomics tools initially help us to determine the interactions between proteins and biomolecules that influence biochemical and physicochemical characteristics of the cell (Figure 1).

These ambitious aims can be realized with the involvement of a large number of different disciplines such as transcriptomics, proteomics, metabolomics and bioinformatics. Even bioinformatics is alone powerful branch using computers to organize the immense amount of information generated from experiments.[7] Inevitably, the advances in proteomics and bioinformatics tools have increased our understanding of the function and metabolic pathways of the molecules playing role in the cell.[2]

Bioinformatics part of proteomics

The widespread usage of mass spectrometer machines not just in proteomics, but also in many other areas of the life sciences, has resulted in rapid developments in hardware, software, and data management. These developments have led to many newly developed instruments, analysis algorithms, softwares, data formats and databases. Therefore, biologists and computational biologists, has tried to manage mass spectrometry data, by using the most up-to-date methods available in order to maximize accurate protein identifications and minimize false identifications.[20] In this manner, the pipeline for management of mass spectrometer data for identification, quantification and characterization of proteins becomes very important and the programs together with databases are needed to be chosen very well for the used mass spectrometer method and approach, and the purpose of the research.

Proteomics tools

Last generations of mass spectrometers can generate large number of MS2 spectra and these need to be used by search algorithms for peptide identification. Search algorithms aim to explain MS2 spectrum generated a peptide sequence by searching against a database including list of peptide sequences which fit to experimental data. The protein databases includes information from translated genomic data, spectral libraries or mRNA databases. And also a final step is needed to assemble the identified peptides into proteins that can be challenging because of redundant peptides or isoforms of proteins. And there are several strategies to reduce the false discovery rate both at peptide identification and protein assembling level.[21]

Peptide mass fingerprinting [PMF] is one of the methods of protein identification using mass spectrometry. This method uses the experimental spectrum consists of the masses of the digested protein fragments detected by the mass spectrometer and the theoretical spectra each generated from the list of masses expected by an enzymatic digestion of each protein sequence in the reference database.[20] Some examples of tools for this method is Mascot,[22] MS-Fit,[23] Profound[24] and PeptIdent with their addresses listed at Table 1.

Table 1 A short list of popular PMF packages

PMF package	URL
Mascot	http://www.matrixscience.com/search_form_select.html
MS-Fit	http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msfitstandard
ProFound	http://prowl.rockefeller.edu/prowl-cgi/profound.exe
PeptIdent	http://iop.vast.ac.vn/theor/conferences/smp/1st/kaminuma/ExPASy/peptident.html

Mascot is one of the most used PMF tool for peptide identification. Mascot integrates different types of searches such as searching sequence database by using mass spectrometer data; searching peptide molecular weights generated from a digestion of protein by an enzyme; and using tandem mass spectrometry [MS/MS] data.[22] The scoring algorithm of Mascot is probability based with advantages: [i] a simple rule can be used to determine whether a result is significant or not [default $p<0.05$]; [ii] scores can be compared with the results from other types of search such as sequence homology; [iii] search parameters can be optimized.[22] There are also some limitations of this tool: [i] It is based on certain assumption that the experimental data are independent measurements. And if the data are not independent, the absolute score becomes unreliable result. [ii] Most commonly seen problem is duplicate mass values depending on different reasons (Table 2 & 3).[22]

Table 2 A short list of popular PFF packages

PFF package	URL
Mascot	http://www.matrixscience.com/search_form_select.html
SEQUEST	http://proteomicsresource.washington.edu/protocols06/SEQUEST.php
X!Tandem	http://prowl.rockefeller.edu/tandem/thegpm_tandem.html

Table 3 Hierarchy of proteomics databases according to the data types

Data type	Proteomics databases
Raw Data	Chorus, Proteome Xchange, PRIDE, Mass IVE, PASSEL,
Peptide/Protein Identification or Quantification	PRIDE, PASSEL, MaxQB, Human Proteome Map, Peptide Atlas, ProteomicsDB, MOPED, Human Proteinpedia, GPMDB, PaxDb
Protein Knowledge-bases	neXtProt, UniProt

Citation: Zengin T, Kılıç S, Yenigün U, et al. A smart way for talking with proteins; proteomics. *MOJ Proteomics Bioinform.* 2017;5(3):86–92. DOI: 10.15406/mojpb.2017.05.00160

MS-Fit first groups the proteins of interest according to their initial [parent] mass weight and within each group, a series of groups are created according to the trypsin digested peptide masses. By calculating the probability of a random tryptic peptide match for the distribution of these tryptic peptide masses for a given parent mass, the size of the search database is reduced. This has the effect of lowering the threshold to determine an identified protein. MS-Fit runs over FASTA format databases. MS-Fit allows the user to pre-filter any likely contaminants from the spectrum, increasing the quality of the spectrum.[20]

Profound[24] uses a Bayesian algorithm to identify proteins from databases using mass spectrometric peptide mapping data. The algorithm ranks the candidate proteins by using individual properties of each protein such as enzyme cleavage information, the knowledge that particular aminoacids are present [or absent] in the sample protein, and previous experiments on the sample protein.[20,24] This program can identify the correct proteins even the data quality is low or the sample consists of mixture of proteins.[24]

PeptIdent calculates theoretical peptides of all proteins in the Swiss-Prot/TrEMBL database by digestion of them with the enzyme of choice and calculates theoretical masses of generated peptide fragments.[25] PeptIdent matches the masses of peptides from experiment against all peptide masses in the index. Best matching proteins in database are ranked by the number of hits against experimental peptides. Unlike other PMF tools PeptIdent takes into account post-translational modifications and alternative splicing events annotated in database. PeptIdent also removes signal sequences and propeptides before computing pI and peptide masses for each mature forms.[25]

These tools can be comparable for similar results by using same reference dataset. By using matrix-assisted laser desorption–ionization time-of-flight [MALDI-TOF] mass spectrometry data, Mascot and Profound showed similar performance while MS-Fit shows low protein identification coverage. On the other hand, Profound performs better with different parameter settings such as taxonomy restriction, mass accuracy variation, variable modifications, and missed cleavages.[26] And all together, each tool has advantages, disadvantages and limitations. Ms-Fit can be used for proteins with known parameters and as advantage pI and MW can be searched at the same time but it is hard to obtain accurate results with too many parameters.[26] Profound can be used for simple mixtures of proteins and it is possible to use amino acid sequence information. Another advantage of Profound is that it enhances sensitivity and selectivity with other types of information such as proteolytic peptide distribution from experience. But it is impossible to use pI values for identification.[26] Mascot is suitable to use with Profound and uses distribution system but it searches too wide range of MW.[25] PeptIdent takes into account post-translational modifications and provides a likely protein identifications but PeptIdent does not use characteristic information of proteins and it is possible to identify annotated proteins.

Peptide fragment fingerprinting [PFF] approaches are the mainstream of high-throughput protein identification. Also in this method, at first proteins are digested with a protease and then selected for further fragmentation to generate PFF spectra. The set of these spectra and the parent mass of these fragmented peptides are used for database search. The scoring systems generally consist of two steps: [i] generating a score for each protein in the database and [ii] calculating confidence level for the top-ranking identified.[20]

Mascot[22] and SEQUEST[27] have been mostly used tools for protein identification by using PFF approach. Mascot is based on probabilistic algorithm and uses the parent masses of the peptides and their abundance for search. Mascot also includes new parameters such as selecting the mass spectrometer type and so the input data type.[20] SEQUEST is a tandem mass spectrometry database search program based on a patented scoring algorithm and uses the data of mass spectrometer three times while selecting peptides which have similar parent ion mass from database; while performing ''closeness-of-fit'' filter to select the top 500 peptide candidates from first list of candidate peptides; and for correlation function to produce final scores.[27] Scores of protein identifications coming from SEQUEST is needed to convert into probabilities by softwares such as PeptideProphet.[28] Although SEQUEST is popular, its performance is slow and varies based on the size of the peptide database especially for the number of candidate peptides per spectrum. Efficient data analysis of experiments needs significant computational resources.[29] UW SEQUEST is no longer being supported but there is open-source version of this tool named Comet.[30] There are a few variants of SEQUEST now: the primary version is supplied as a part of a software package Proteome Discoverer by Thermo Fischer Scientific; there is a high throughput version called SEQUEST Sorcerer of Sage-N Research; the Yates Laboratory has a version of it; University of Washington proteomics community currently uses a version of the SEQUEST linking to Mike Hoopmann's MSToolkit; and the core algorithms in SEQUEST have also been derived in the Crux program from the Noble Laboratory.[31] X! TANDEM is an open source program uses expectation values for both experimental peptides and theoretical peptides, and derives scores for same expectation values. Taxonomy information can be added into the system and so faster searches than other PFF tools, can be performed. X! TANDEM can allow various protein point mutations during the search.[32] Kapp et al.,[33] compared the publicly available PFF packages by using a common dataset and found that SEQUEST showed best performance identifying significantly more proteins and more sensitivity. Mascot and X! Tandem to be better at distinguishing correct and incorrect peptide hits. Resing et al.,[34] reported that Mascot and SEQUEST can validate less than half of the potentially identifiable MS/MS spectra showing similar results from a manually expert verified dataset. Chamrad et al.,[26] reported that SEQUEST shows more obvious separation of correct identifications and random matches than Mascot, also SEQUEST identified more than twice the number of proteins compared to Mascot.

There are also other tools and they can be grouped as three main types: Sequence search tools such as Mascot,[22] SEQUEST,[27] X!Tandem,[32] OMSSA,[35] MyriMatch[36] which matches acquired spectra from mass spectrometer with theoretical spectra generated from possible peptide sequences in a protein database; Spectral library search tools such as SpectraST,[37] X!Hunter,[38] and Bibliospec,[39] which matches acquired spectra with a library of previously observed and identified spectra; and de novo search tools such as PEAKS,[40] PepNovo,[41] and Lutefisk,[42] which derives peptide identifications based on the MS/MS spectrum peak patterns alone, without reference sequences or previous spectra.[43] Additionally, there are hybrid search tools such as InSpecT[44] and PEAKS-DB[45] which have combinations of de novo sequencing and database searching.[46] Other search tools are specifically designed for the analysis of post-translationally modified peptides such as ModifiComb[47] or InsPecT.

Among the many tools used for peptide identification, choosing best tool for the experiment samples and datasets is not easy. The principles, algorithms and parameters of tools should be well known and the pipeline should be designed very well specifically for the used mass spectrometry method, output data and purpose of the experiment. It is hard to compare many tools with the experimental

data set of interest but the combining results of a few tools can avoid disadvantages of individual tools. Shteynberg et al.,[46] showed that the results of a single search engine are improved by combining search engine results. They demonstrated that selecting the engines with the most complementary scoring functions is most beneficial. And also using as many search engines as possible with a combiner will increase the confidence of identified peptide spectrum matches, distinct peptide sequences, and proteins. It can also maximize the coverage of identified proteins which would be missed at searching by single engine.[46]

Proteomics databases

As the proteomics technologies- mainly MS based protein identification and quantification – have been developed through advances in methodologies, instrumentation, computational analysis tools and protein sequence databases. Therefore depending on the development of more powerful and sensitive methods, more protein can be identified and quantified. As a result of this improvement the output data has been increased day by day. According to the other data intensive fields such as genomics, original proteomics data storage has been less common in public resources since the proteomics data are more complex due to the wide variety of MS technologies, proteomics tools and pipelines. The complexity of proteomics data is caused by protein isoforms results of alternative splicing, post-translational modifications, protein degradation events, dynamic inter-connectivity of proteins. And new analytical and computational tools have been developed to solve this complexity and this complicates the data standardization and deposition. And also researchers have different interest and research projects on proteomics data.[48] And one of the major challenges of global proteomic studies that we have faced with is the missing data, because many statistical approaches are not sufficient since they require complete datasets.[49] However, the importance of the data standardization, storage and sharing publicly has been increased. For this purpose, many publicly available protein databases have been developed each with different purposes and formats, such as Global Proteome Machine Database [GPMDB],[50] Peptide Atlas,[51] and the PRIDE database,[52] ProteomicsDB,[53] MassIVE [Mass Spectrometry Interactive Virtual Environment], Chorus, MaxQB,[54] PASSEL [PeptideAtlas SRM Experiment Library],[55] MOPED [Model Organism Protein Expression Database],[56] PaxDb,[57] Human Proteinpedia,[58] and the Human Proteome Map [HPM].[59] Proteome Xchange [PX] consortium[60] has been formed and PRIDE, Peptide Atlas, PASSEL, and MassIVE are the active members of this consortium.[48]

The PX consortium has been created for to generate collaboration and integration of research on MS proteomics repositories, proteomics researchers, and representatives from journals. Two major workflows -MS/MS and SRM methods- are fully supported in PX. PRIDE and Mass IVE acts as the initial submission point for MS/MS data, while PASSEL has the similar role for SRM data. Proteome Central has metadata associated with datasets [PRIDE and MassIVE for MS/MS data, PASSEL for SRM data, or Peptide Atlas for reprocessed original PX datasets].[48] PRIDE[52] has peptide/protein identifications including PTMs, expression values, the analyzed mass spectra, and the related metadata. PRIDE supports both complete and partial submissions and data are stored as originally analyzed by the researchers. PASSEL[55] supports the submission of datasets generated by SRM methods and stores the experimental results and the corresponding raw data. The submitted raw data are automatically reprocessed in a uniform manner and the results are loaded into the database. Peptide Atlas[51] has served as a data reprocessing resource and a research database for the development of spectral libraries and SRM-related tools. Peptide Atlas is one of the biggest and well-curated protein expression data resources.[48]

Proteomics DB[53] is a human protein expression database storing protein and peptide identifications and quantification measurements. It contains information of over 18 000 human genes, representing around 90% of the human proteome. It contains more than 70million spectra from human cancer cell lines, tissues, and body fluids.[48] The HPM [Human Proteome Map][59] has been developed as an output of the human proteome. HPM aims to make possible to review, navigate and visualize the protein expression information related with gene families, protein complexes, signaling pathways and biomarkers. MS/MS data obtained from all different experiments were searched against the Human RefSeq database using SEQUEST and MASCOT. Protein and peptide identifications were converted into MySQL format. NCBI RefSeq annotations were used as additional information about the genes.[59]

UniProtKB[61] is one of the most used databases as protein sequence and functional annotation provider. UniProtKB provides a broad range of protein sequence datasets for a large number of species, with high-quality sequence annotations and mappings to the genomics and proteomics information. UniProtKB use MS proteomics data of the other main public repositories to enrich protein sequence annotations at the level of the evidence.[48] neXtProt[62] is a web-based protein knowledge platform to support research uniquely on human proteins. The set of manually curated annotations taken from UniProtKB/Swiss-Prot for human, which is quality-filtered and carefully selected high-throughput experiments from different scientific research areas for abundance, distribution, subcellular localization, interactions, and cellular functions. All the relevant proteomics MS-related information including such as post-translated modifications, has been integrated into neXtProt and is available via the web interface.[48]

There are many other protein databases with specialized purposes and data types such as protein database of NCBI, SwissProt, PIR [Protein Information Resource], PRF [PRF/SEQDB, Protein/Peptide Sequence Database] and PDB [Protein Data Bank]. For extra information the review of Kumari et al.,[63] is suggested for detailed overview of huge collection of different types of databases along with bioinformatics tools used for protein structure, function prediction, conserved regions in protein families, 3D structure prediction of drug targets as well as new drug discovery and protein-protein interactions.

Conclusion

Proteomics, together with transcriptomics and metabolomics, is major branch within the "-omics" approaches and represents a powerful technique that allows taking of a "snapshot" of the entire repertoire of proteins present in a cell, tissue or organ at a given time under defined conditions. It is a relatively young science and both technical and computational improvement is still needed to "fill the gap" with other -omics techniques in terms of sensitivity and specificity. Also the researchers need more powerful, less cost effective and time consuming MS tools and also standardization of data, databases and pipelines to be able to handle different scientific problems. In this manner, more user friendly techniques and computational tools also should be generated to decrease the hard work and complexity of proteomics research in future. However, proteomics alone may not enough to solve proteome complexity of the living cell or organism. Other-omics techniques share the possibility of amplification of analytical targets and this peculiarity has pushed the progress of these sciences to frontiers that were unimaginable only a few years ago. We

Citation: Zengin T, Kılıç S, Yenigün U, et al. A smart way for talking with proteins; proteomics. *MOJ Proteomics Bioinform.* 2017;5(3):86–92.
DOI: 10.15406/mojpb.2017.05.00160

believe that in future with together of all data provided from –omics techniques, whole biological information will be analyzed deeply and broadly and use of these techniques will open new horizons to understand of cases remained unclear in biological systems.

Acknowledgements

None.

Conflict of interest

The author declares no conflict of interest.

References

1. Yarmush ML, Jayaraman A. Advances in proteomic technologies. *Annu Rev Biomed Eng*. 2002;4:349–373.
2. Khoranhlai KK. Advances in Proteomics and Bioinformatics in Agriculture Research and Crop Improvement. *Journal of Proteomics & Bioinformatics*. 2015;8(3):39–48.
3. Pandey A, Mann M. Proteomics to study genes and genomes. *Nature*. 2000;405(6788):837–846.
4. Kellner R. Proteomics. Concepts and perspectives. *Fresenius J Anal Chem*. 2000;366(6–7):517–524.
5. Landels A, Evans C, Noirel J, et al. Advances in proteomics for production strain analysis. *Curr Opin Biotechnol*. 2015;35:111–117.
6. Humphery–Smith I. The 20th anniversary of proteomics and some of its origins. *Proteomics*. 2015;15(11):1773–1776.
7. Graves PR, Haystead TA. Molecular biologist's guide to proteomics. *Microbiol Mol Biol Rev*. 2002;66(1):39–63.
8. Aebersold R, Mann M. Mass–spectrometric exploration of proteome structure and function. *Nature*. 2016;537(7620):347–355.
9. Azimifar SB, Nagaraj N, Cox J, et al. Cell–type–resolved quantitative proteomics of murine liver. *Cell Metab*. 2014;20(6):1076–1087.
10. Richards AL, Merrill AE, Coon JJ. Proteome sequencing goes deep. *Curr Opin Chem Biol*. 2015;24:11–17.
11. Sharma K, Schmitt S, Bergner CG, et al. Cell type– and brain region–resolved mouse brain proteome. *Nat Neurosci*. 2015;18(12):1819–1831.
12. Meissner F, Scheltema RA, Mollenkopf HJ, et al. Direct proteomic quantification of the secretome of activated immune cells. *Science*. 2013;340(6131):475–478.
13. Smith LM, Kelleher NL. Proteoform: a single term describing protein complexity. *Nat Methods*. 2013;10(3):186–187.
14. Tyanova S, Temu T, Sinitcyn P, et al. The Perseus computational platform for comprehensive analysis of proteomics data. *Nat Methods*. 2016;13(9):731–740.
15. Gundry RL, White MY, Murray CI, et al. Preparation of proteins and peptides for mass spectrometry analysis in a bottom–up proteomics workflow. *Curr Protoc Mol Biol*. 2009; Chapter 10:Unit10.25.
16. Wu CC, MacCoss MJ. Shotgun proteomics: tools for the analysis of complex biological systems. *Curr Opin Mol Ther*. 2002;4(3):242–250.
17. Washburn MP, Wolters D, Yates JR. Large–scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*. 2001;19(3):242–247.
18. Wu W, Fu Y, Therkildsen M, et al. Molecular Understanding of Meat Quality Through Application of Proteomics. *Food Reviews International*. 2014;31(1):13–28.
19. Elzek MA, Rodland KD. Proteomics of ovarian cancer: functional insights and clinical applications. *Cancer Metastasis Rev*. 2015;34(1):83–96.
20. McHugh L, Arthur JW. Computational methods for protein identification from mass spectrometry data. *PLoS Comput Biol*. 2008;4(2):e12.
21. Schmidt A, Forne I, Imhof A. Bioinformatic analysis of proteomics data. *BMC Syst Biol*. 2014;8 Suppl 2:S3.
22. Perkins DN, Pappin DJ, Creasy DM, et al. Probability–based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999;20(18):3551–3567.
23. University of California San Francisco. *UCSF Protein Prospector version 5.19.1 [computer program]*. 1996.
24. Zhang W, Chait BT. ProFound: An expert system for protein identification using mass spectrometric peptide mapping information. *Anal Chem*. 2000;72(11):2482–2489.
25. Joo WA, Lee JB, Park M, et al. Comparison of search engine contributions in protein mass fingerprinting for protein identification. *Biotechnology and Bioprocess Engineering*. 2007;12(2):125–130.
26. Chamrad DC, Korting G, Stuhler K, et al. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics*. 2004;4(3):619–628.
27. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*. 1994;5(11):976–989.
28. Keller A, Eng JK, Hubley R. P*eptide Prophet*. 2002.
29. Diament BJ, Noble WS. Faster SEQUEST serrching for peptide identification from Tandem mass–spectra. *J Proteome Res*. 2012;10(9):3871–3879.
30. Comet MS/MS database search program release 2016.01 rev. 2 (2016.01.2)
31. http://proteomicsresource.washington.edu/protocols06/SEQUEST.php
32. Craig R, Beavis RC. TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics*. 2004;20(9):1466–1467.
33. Kapp EA, Schütz F, Connolly LM, et al. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. *Proteomics*. 2005;5(13):3475–3490.
34. Resing KA, Meyer–Arendt K, Mendoza AM, et al. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal Chem*. 2004;76(13):3556–3568.
35. Geer LY, Markey SP, Kowalak JA, et al. Open mass spectrometry search algorithm. *J Proteome Res*. 2004;3(5):958–964.
36. Tabb DL, Fernando CG, Chambers MC. MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res*. 2007;6(2):654–661.
37. Lam H, Deutsch EW, Eddes JS, et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*. 2007;7(5):655–667.
38. Craig R, Cortens JC, Fenyo D, et al. Using anno– tated peptide mass spectrum libraries for protein identification. *J Proteome Res*. 2006;5:1843–1849.
39. Frewen BE, Merrihew GE, Wu CC, et al. Analysis of peptide MS/MS spectra from large–scale proteomics experiments using spectrum libraries. *Anal Chem*. 2006;78(16):5678–5684.
40. Ma B, Zhang K, Hendrie C, et al. PEAKS: powerful software for peptide de novo se– quencing by tandem mass spectrometry. Rapid Commun. *Mass Spectrom*. 2003;17(20):2337–2342.
41. Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem*. 2005;77(4):964–973.

42. Taylor JA, Johnson RS. Implementation and uses of auto– mated de novo peptide sequencing by tandem mass spectrometry. *Anal Chem*. 2001;73(11):2594–2604.

43. Pevtsov S, Fedulova I, Mirzaei H, et al. Performance evaluation of existing de novo sequencing algorithms. *J Proteome Res*. 2006;5(11):3018–3028.

44. Tanner S, Shu H, Frank A, et al. 'InsPecT: Identification of post translationally modified peptides from tandem mass spectra. *Anal Chem*. 2005;77(14):4626–4639.

45. Zhang J, Xin L, Shan B, et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identifica– tion. *Mol Cell Proteomics*. 2012;11(4):M111.010587.

46. Shteynberg D, Nesvizhskii AI, Moritz RL, et al. Combining Results of Multiple Search Engines in Proteomics. *Mol Cell Proteomics*. 2013;12(9):2383–2393.

47. Savitski MM, Nielsen ML, Zubarev RA. 'ModifiComb, a New Proteomic Tool for Mapping Substoichiometric Post–translational Modifications, Finding Novel Types of Modifications, and Fingerprinting Complex Protein Mixtures. *Mol Cell Proteomics*. 2006;5(5):935–948.

48. Perez–Riverol Y, Alpi E, Wang R, et al. Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. *Proteomics*. 2015;15(5–6):930–950.

49. Webb–Robertson BJ, Wiberg HK, Matzke MM, et al. 'Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry–Based Label–Free Global Proteomics'. *J Proteome Res*. 2015;14(5):1993–2001.

50. Craig R, Cortens JP, Beavis RC. 'Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res*. 2004;3(6):1234–1242.

51. Farrah T, Deutsch EW, Omenn GS, et al. 'State of the Human Proteome in 2013 as Viewed through Peptide Atlas: Comparing the Kidney, Urine, and Plasma Proteomes for the Biology– and Disease–Driven Human Proteome Project. *J Proteome Res*. 2014;13(1):60–75.

52. Vizcaíno JA, Côté RG, Csordas A, et al. 'The Proteomics Identifications (PRIDE) database and associated tools: Status in 2013. *Nucleic Acids Res*. 2013;41(Database issue):D1063–D1069.

53. Wilhelm M, Schlegl J, Hahne H, et al. Mass–spectrometry–based draft of the human proteome. *Nature*. 2014;509(7502):582–587.

54. Schaab C, Geiger T, Stoehr G, et al. Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol Cell Proteomics*. 2012;11(3):M111.014068.

55. Farrah T, Deutsch EW, Kreisberg R, et al. PASSEL: The Peptide Atlas SRM experiment library. *Proteomics*. 2012;12(8):1170–1175.

56. Montague E, Stanberry L, Higdon R, et al. 'MOPED 2.5–an integrated multi–omics resource: multi–omics profiling expression database now includes transcriptomics data'. *OMICS*. 2014;18(6):335–343.

57. Wang M, Weiss M, Simonovic M, et al. 'PaxDb, a Database of Protein Abundance Averages Across All Three Domains of Life. *Mol Cell Proteomics*. 2012;11(8):492–500.

58. Kandasamy K, Keerthikumar S, Goel R, et al. Human Proteinpedia: A unified discovery resource for proteomics research. *N Nucleic Acids Res*. 2009;7(Database issue):D773–D781.

59. Kim MS, Pinto SM, Getnet D, t al. A draft map of the human proteome. *Nature*. 2014;509(7502):575–581.

60. Vizcaíno JA, Deutsch EW, Wang R, et al. 'ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol*. 2014;32(3):223–226.

61. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res*. 2015;43(Database issue):D204–D212.

62. Gaudet P, Argoud–Puy G, Cusin I, et al. 'NeXtProt: Organizing protein knowledge in the context of human proteome projects'. *J Proteome Res*. 2013;12(1):293–298.

63. Kumari A, Kanchan S, Sinha RP, et al. Applications of Bio–molecular Databases in Bioinformatics. *In Medical Imaging in Clinical Applications*. 2016;651:329–351.

Citation: Zengin T, Kılıç S, Yenigün U, et al. A smart way for talking with proteins; proteomics. *MOJ Proteomics Bioinform.* 2017;5(3):86–92.
DOI: 10.15406/mojpb.2017.05.00160