Master's Thesis

석사 학위논문

A multilayer networks analysis for mining quantification rules from big proteomics data

Suhyoen Jin (진 수 현 晉 秀 賢)

Department of

Information and Communication Engineering

DGIST

2017

Master's Thesis

석사 학위논문

A multilayer networks analysis for mining quantification rules from big proteomics data

Suhyoen Jin (진 수 현 晉 秀 賢)

Department of

Information and Communication Engineering

DGIST

2017

MS/IC

201522022

진 수 현. Suhyoen Jin. A multilayer networks analysis for mining quantification rules from big proteomics data. Department of Information and Communication Engineering. 2017. 39p. Advisor Min-Soo Kim, Co-Advisor Prof. Daehee Hwang.

Abstract

Data modeling is important to understand and obtain the information from the data. Diverse designs can be developed for finding hidden information. Existing research in proteomics is limited in data modeling since only analysis of Protein–protein interaction (PPI) network is usually conducted.

Here, we present a new approach for finding rules and bases to understand mechanisms of protein function. We build the multilayer network for integrating bottom-up proteomics data which is named TLP network. TLP network contains diverse biological information including the peptide expression data, and PTMs as well as Protein–protein interactions (PPIs). TLP network is expected to answer a wide range of questions in proteomics research area.

Keywords: Bioinformatics, Proteomics, Multilayer network

Contents

List of tables & figures

1. Motivation

1.1 Difficulty to identify the Differentially Expressed Proteins

Protein quantification is determining the amount of proteins in a sample. If it is possible to obtain the quantity of proteins accurately, there is definite answer to which proteins are differentially expressed proteins (DEPs) between the different groups (for example, disease vs. control). DEPs play a central role in determining the course of infection. DEPs are also closely related to Primary goal of proteomics, that is the functional annotations for the entire proteome. Therefore, quantitative proteomics is highly relevant for systems biology, biomarker discovery, and many other biomedical applications.

However, protein quantification is difficult since the inherent property of tandem mass spectrometry (MS/MS) data generate subtasks of protein quantification. MS/MS data is given as not the protein level, but the peptide level. Many of these peptides can be found in not only one protein (unique peptide), but also multiple proteins (shared peptide). Furthermore, peptide identification problem makes the protein quantification problem more complex. An experimental MS/MS spectrum is unclean data therefore determining an amino acid sequence of a protein does not have a definitive answer. The implication is that unique peptide is not “unique” peptide in MS/MS proteomics data in fact.

Consequently, protein assembly plays a critical role in identification as it transforms a list of identified peptides into a list of identified proteins [1]. Moreover, grouping of peptides to proteins result in an amplification of error rates going from PSM to protein level [2]. As a results, it is complex problem to determine quantitative information from unique peptides and shared peptides.

Although a number of protein quantification methods have been proposed, identification of DEPs still remains unsolved since protein assembly is conducted in similar way. Ignoring shared peptides and only utilizing the results of unique peptide quantification is widely used way to protein quantification. Utilization the results of unique and razor peptides quantification as a compromise between unequivocal peptide assignment and most-accurate quantification is also used [3]. In this way, significant portion of proteins are discarded, thus it is not available to discover DEPs [4].

1.2 Data integration to find rules from big data

We would like to suggest the solution for this inherent problem by the power of big data. When we focus on informative peptides, which offers the information about DEPs through the big data, we can effectively summarize peptide-level results into protein-level. Furthermore, we can get the solution for other problems in proteomics research area.

In order to find rules from big data, we integrate the proteomics data through noble data model. Data modeling is important to understand and obtain the information from the data. Especially in network analysis, what can be found in the network depends on network design. Diverse designs can be developed for exploiting the relationships, in terms of (node or edge) attribute, data integration, etc.

Existing research is limited in modeling of proteomics data since only analysis of Protein–protein interaction (PPI) network is usually conducted. The novelty of our study lies in the network architecture. There has been few or no research on the PTM itself as well as relationship of similarity in abundances of PTMs. co-occurrence which is defined as edge attribute in modified peptide layer is the first proposed concept to reveal relationship between occurrences of modified peptides.

Furthermore, multilayer network to represent different entities and relationship between them in proteomics research area is a new effort. The interaction between proteins are intimately related with PTMs which provides the information of signaling pathway and the expression of peptides provides the information of biological process. That is, all of those things defined as node in TLP network are closely connected to each other with respect to function of protein. Therefor we have to look into the relationship between those things by data integration. Since graph is appropriate for modeling the relationship between entities, designing a graph data model like as TLP network can answer a wide range of questions in proteomics research area.

2. Introduction

2.1 Bioinformatics

Bioinformatics is an interdisciplinary field that develops computational and statistical methods for understanding biological data. The field of bioinformatics has been rapidly evolving and has drawn attention. Due to the advances in biomedical techniques and applications such as genome sequencing, protein identification, the amount of data exploded. Currently, the rate of data accumulation is much faster than the rate of data interpretation [5]. These data need to be effectively analyzed to discover useful information.

Since uncertainty should be taken into account when analyzing the biomedical data, numerous statistical methods and mining approaches are needed which usually lead to complex computation. As a result, computational approaches for data analysis have become the critical final step of the bioinformatics workflow because of a great deal of uncertainty.

2.2 Proteomics

Proteomics is the large-scale study of proteins which is important because proteins represent the actual functional molecules in the cell. According to the central dogma of molecular biology which represent the flow of genetic information within a biological system, the DNA makes RNA which in turn makes protein. As a result, protein is the

end product which is the primary effectors in cellular function. Primary goal of proteomics is functional annotations for the entire proteome. There are two phase of proteomics data analysis, identification and quantification.

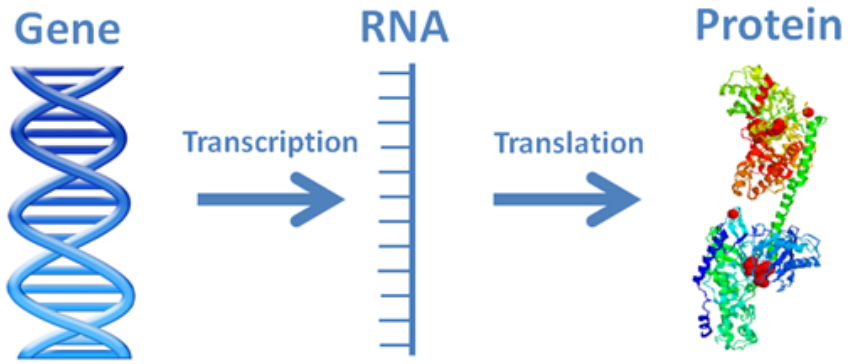

Figure 1. Central dogma.

2.3 Post-translational modification (PTM)

PTM is the chemical modification of a protein after its translation. PTM is important since many of these can regulate protein function and play a key role in many cellular processes such as signaling pathway, regulation of gene expression, and protein-protein interactions. Therefore, the identification of PTMs is critical to gaining insight into biological functions. In MS/MS, PTMs can be detected by PTM-related diagnostic mass shifts of fragment ions in MS/MS spectra [6].

2.4 Bottom-up proteomics

Bottom-up proteomics is a common method which allow the identification and characterization of proteins and their amino acid sequences, including PTMs, by proteolytic digestion prior to mass spectrometry (MS) analysis.

Workflow in bottom-up proteomics is illustrated in Figure 2. Firstly, protein extracted from sample of interest such as cancer patient. Next, complex protein mixture is digested by protease such as trypsin. This peptide mixture goes through the liquid chromatography and is separated prior to MS. Peptides are then ionized and selected ions subjected to fragmentation in the collision cell to produce tandem mass spectra through Multiple steps of mass spectrometry i.e., tandem mass spectrometry (MS/MS). At this stage, database search programs for peptide identification such as SEQUEST [7], and Mascot [8] are used. The list of identified peptides is used to infer which proteins are present in the original sample. [9]

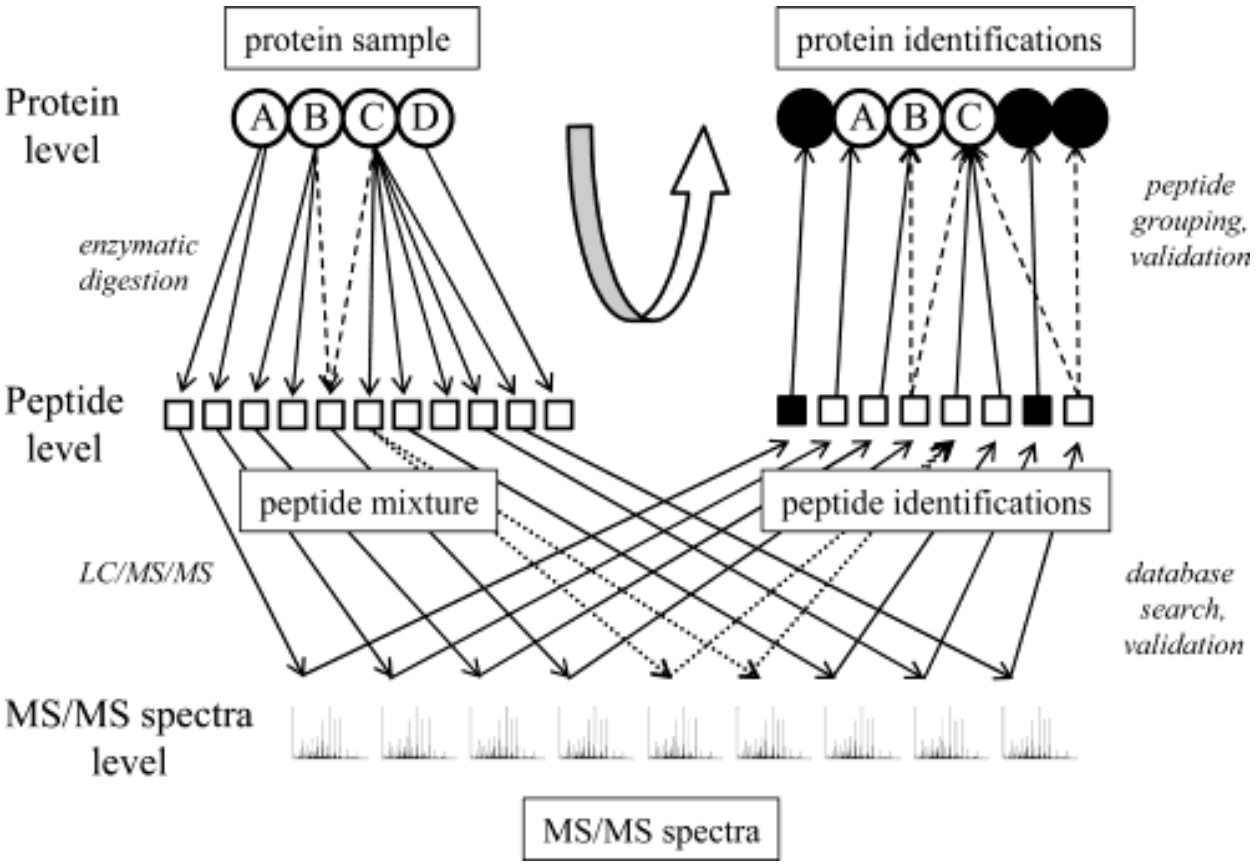

Figure 2. Workflow of Bottom-up proteomics. [9]

2.5 Peptide identification

Peptide identification is deriving correct Peptide-Spectrum Matches (PSMs) from given MS/MS spectrum. There are three major approaches to identifying peptides from MS/MS spectrums: traditional database search approach (Figure. 3 (a)), de novo sequencing approach (Figure. 3 (b)), and hybrid approach (Figure. 3 (c)).

Database search approach tries to find a peptide by comparing an experimental MS/MS spectrum with a theoretical spectrum predicted from a protein database. De

novo sequencing approach directly infers peptide sequences from experimental spectrums without any resort to a database. Also, there are efforts to combine the database search approach and de novo sequencing approach for better interpretation. Hybrid approach first performs simplified de novo sequencing for generating candidate peptides. Then, it performs database search against the results of the simplified de novo sequencing.

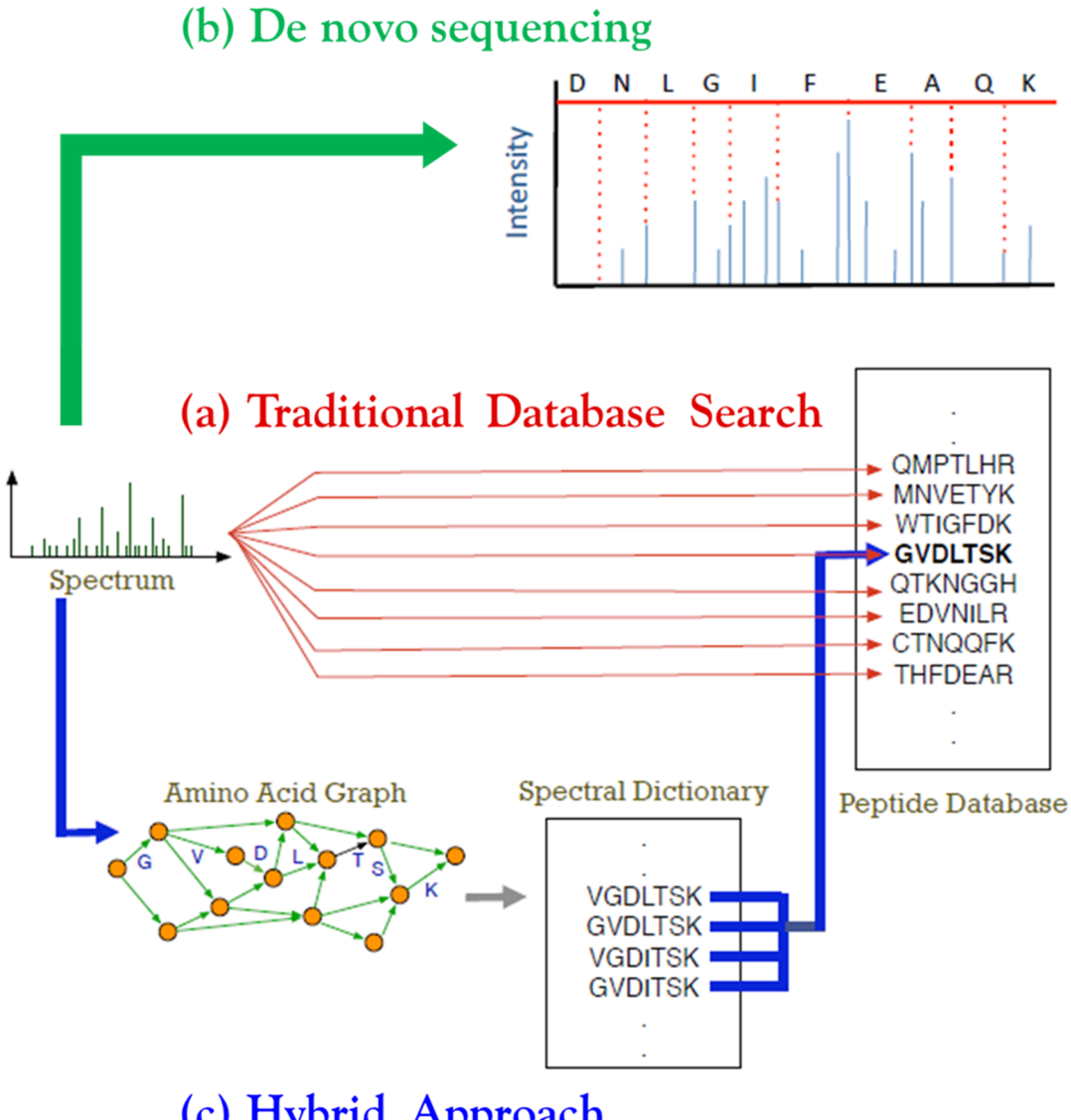

Figure 3. Three major approaches for peptide identification.

2.6 Quantitative proteomics

Quantitative proteomics is important extension to protein identification, determining the amount of proteins in a sample. Quantitative proteomics is highly relevant for systems biology, biomarker discovery, and many other biomedical applications. [10]

There are two kinds of quantification approaches by MS. First, in label-free quantification, all samples are analyzed in separate LC/MS experiments, and the individual peptide properties of the individual measurements are then compared. Label-free quantification is in expensive and can provide the highest flexibility. However, the results of label-free quantification shows relatively poor accuracy compared to the results of labeling quantification.

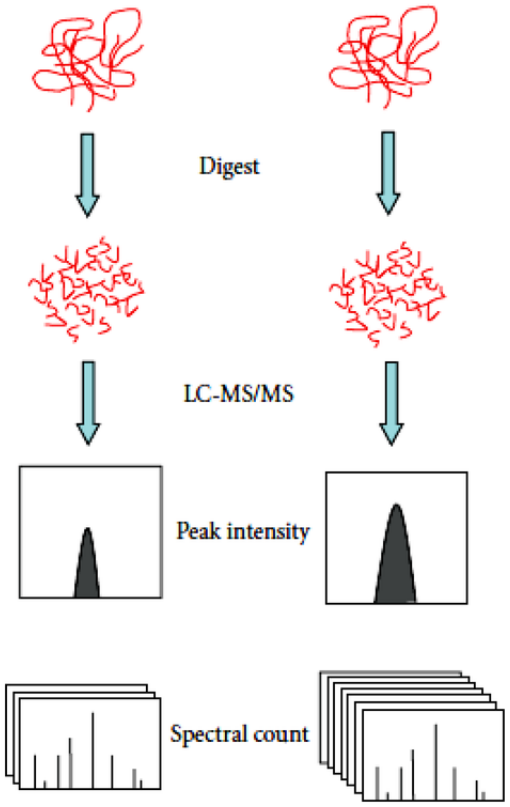

Figure 4. Label-free quantification.

Second, in labeling quantification, proteins from different samples are labeled which allows to distinguish between identical proteins in separate samples. It is relatively easy to map the signals of the same peptide from two samples in labeling quantification and more reliable results of quantification from the peak intensity ratio. However, it is more costly and time-consuming. Sample loss and error introduced by labeling and limitation of number of samples also can be disadvantage of the labeling quantification.

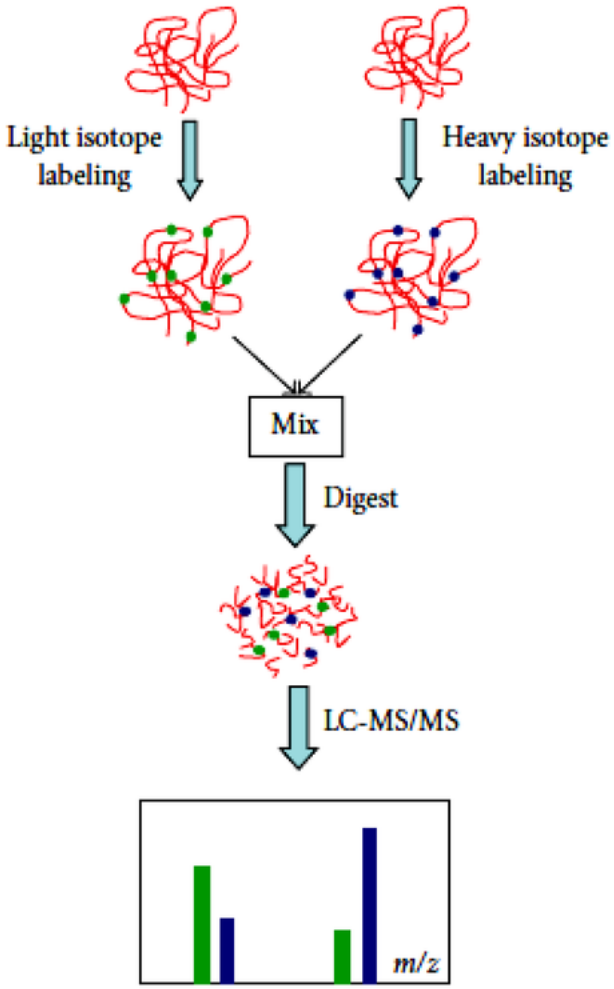

Figure 5. Isotope labeling method.

2.7 Proteomics data repositories

The advent of high-throughput proteomics has enabled the dramatically increased number of publications centered on these protein identifications. The proteomics identifications (PRIDE) database (http://www.ebi.ac.uk/pride) is proposed as a means to finally turn publicly available data into publicly accessible data [11]. PRIDE databases is the repository that European Bioinformatics Institute hosted which is a centre in bioinformatics, and is part of European Molecular Biology Laboratory. By September 1 2015, PRIDE Archive contained 3336 data sets, 52% (1731) of the data sets were publicly available. [12]

The Clinical Proteomic Tumor Analysis Consortium (CPTAC) data portal (https://cptac-data-portal.georgetown.edu/cptacPublic) is also the centralized repository for proteomics data. CPTAC data portal is the repository that National Cancer Institute in NIH launched, which is part of the U.S. Department of Health and Human Services. Especially, CPTAC is a comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer through the application of proteomic technologies and workflows to clinical tumor samples with characterized genomic and transcript profiles. In 2015, the portal hosts 6.3 TB of data and includes proteomic investigations of breast, colorectal, and ovarian tumor tissues from The Cancer Genome Atlas (TCGA) [13].

PRIDE database and CPTAC data portal both databases are world-leading data repositories of MS-based proteomics data and all data are freely available to the public.

2.8 Network analysis

Graph is widely used in network analysis since it offers a convenient way to represent relationships among objects. Graphs and networks are all around us, including protein-protein interaction network, friendship network, and internet map. Graph consists of a set of nodes and a set of edges. Node represents an object, and the edge represents the relationship between the objects.

Power-law distributions occur in many situations of scientific interest and have significant consequences for our understanding of natural and man-made phenomena. The most representative example is a scale-free network. A scale-free network is a network whose degree distribution follows a power law. In real network, hub node means a node which has the number of edges that greatly exceeds the average.

There are lots of algorithms for the network analysis. For example, graph clustering can be used for community detection in networks. General graph clustering algorithms is organized the graph topology into modules commonly called communities or clusters. The essence here is that nodes of the same community are highly similar while on the contrary, nodes across communities present low similarity [14]. For example, SCAN, which is a density-based clustering algorithm, assign densely connected node set as cluster [15].

3. Three Layer proteomics (TLP) network

3.1 Data collection

We collected the data from PRIDE database and CPTAC data portal according to following five criteria.

First criterion is the type of disease. We collected data from human ten major tumors (Lung, Liver, Pancreas, Prostate, Breast, Glioma, GI, Ovarian, Leukemia, Melanoma).

Second criterion is sample class (tissue, plasma/serum/ascites, cell lines, and primary cells).

Third criterion is size of measured proteomes which is related to the quality of data. When the data is not enrichment data for Protein post-translational modification (PTM), we only include the data which has more than 500 global proteins identification result. When the data is enrichment data for PTM, we only include the data which has more than 100 proteins identification results with 500 global proteins identification result.

Fourth criterion is quantification method. We collected the data which conducted isobaric labeling method (i.e., iTRAQ [16] and TMT [17]) or the method which used areas of elusion curves (i.e., SILAC [18] and label-free).

Finally, we check whether the data has samples from more than two conditions. This enables to compare the identification results between different conditions (for example, control and treatment).

According to five criteria, we selected total 86 studies with 16200 vendor raw files. The 72 studies with 7600 raw files are from the PRIDE database, and the 14 studies with 8600 raw files are from the CPTAC data portal. Collected raw data is amount to ten terabytes.

3.2 Process of peptide identification

Before peptide identification, we converted vendor raw files with .RAW extension to open spectrum files with .mgf extension. Mascot Generic Format (MGF) file is the most common text format among open spectrum files. This simple text formats were created with the emergence of search engine, and widely supported by many proteomics search engines.

After the conversion, we conducted MS-GF+ for peptide identification of collected data. MS-GF+ is state-of-the-art database search tool. In MS-GF+, MS/MS spectra are scored against peptides derived from a protein sequence database. [19]

The data files searched against the Uniprot database (September 2015 with 91,797 sequences) allowing a maximum of one missed cleavage sites. UniProt is a freely accessible database of protein sequence and many entries being derived from genome sequencing projects.

Default settings used were the following: The precursor mass tolerance was set to

20 ppm. Carbamidomethylation on cysteine setting as fixed modifications and variable modifications as Deamidation, Oxidation (M), Pyroglutamic acid formation from Gln, and Pyroglutamic acid formation from Glu. In addition to default setting, such as phosphorylation, and acetylation as variable modification were used for PTM enriched experiments. In case of labeling data, reporter ion is also used as variable modification. The peptide is considered as confidently identified when false discovery rate (FDR) is lower than 0.01. Number of peptide matches per spectrum to report was set to one so that we obtain one peptide from each spectrum.

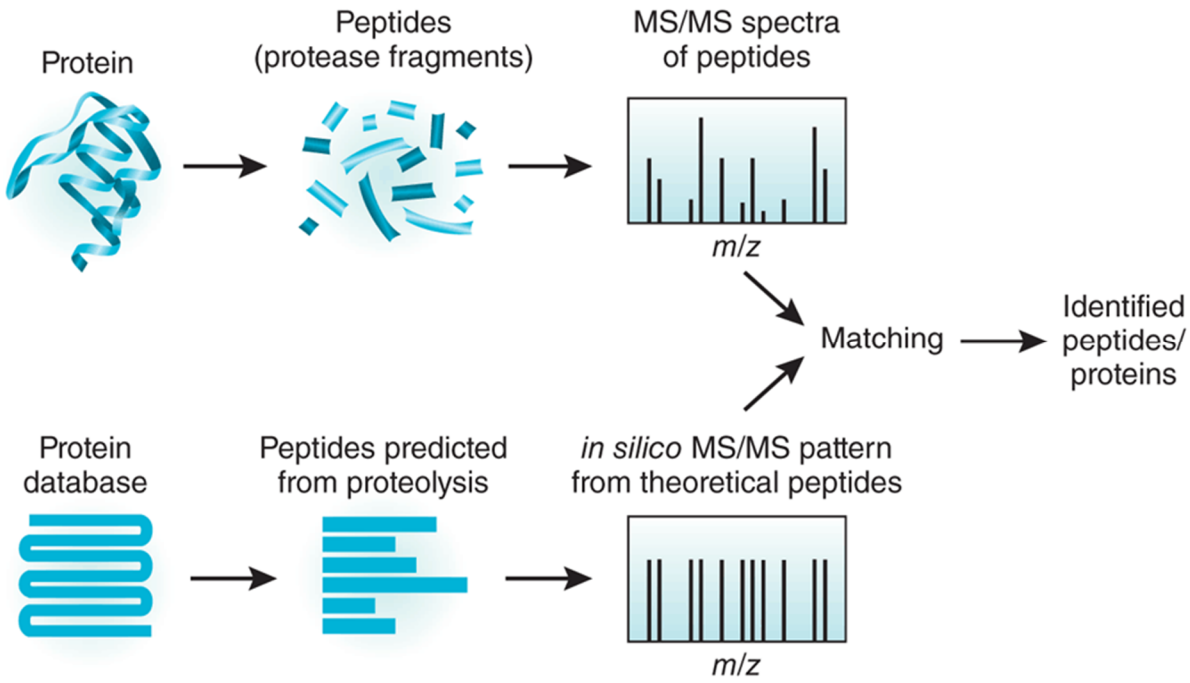

Figure 6. Traditional database search approach for peptide identification.

The number of total Peptide-Spectrum-Matches (PSM) is 160 million and the number of reliable PSMs which has FDR lower than 0.01 is 80 million. Finally, we constructed multi-layer proteomics network based on these 80 million reliable PSMs.

Total search time was amount to 180,000 hours. MS-GF+ search is conducted during four months with multiple machines.

3.3 Process of peptide quantification

Peptide quantification is determining the amount of peptides in a sample. There are a number of quantification method including isobaric labeling method and the method which used areas of elusion curves.

We only quantified the data which has isobaric labeling which can offer highly reliable quantification results through MS/MS data. Peptides or proteins are labeled with various chemical groups that are isobaric (identical masses). Tags are cleaved from the peptides during MS/MS, and MS/MS data is used for both identification and quantification. We can get peptide-spectrum match (PSM) by peptide identification and quantity of the peptide by reporter ion which tag generate from the spectrum.

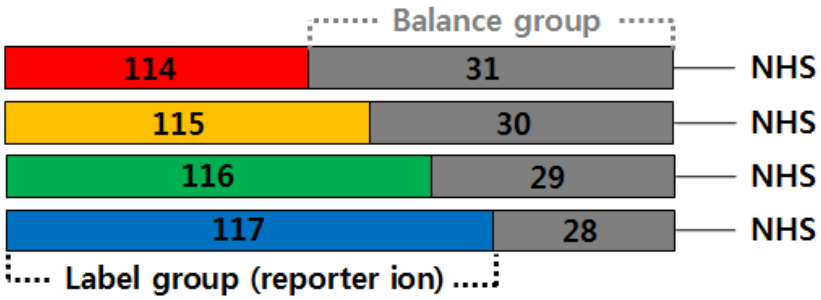

Figure 7. Isobaric tags.

According to MS-GF+ search result, we determine the PSMs list to quantify which confidently identified with FDR lower than 0.01. Max value among the intensities of reporter ions in each PSMs, and quantile normalization is conducted within and between datasets in each study. Quantile normalization is technique for making two distributions identical in statistical properties.

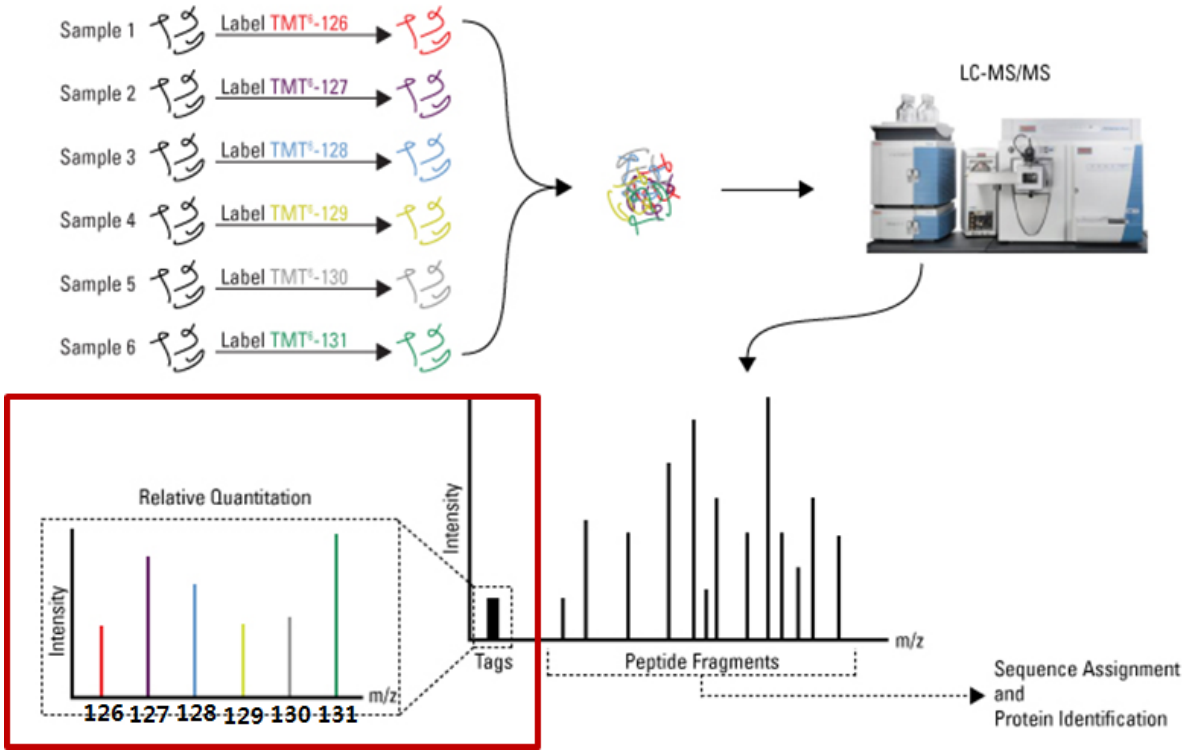

Figure 8. Quantification in iTRAQ and TMT.

After normalization, intensity ratio to reference pool is computed according to labeling data. Labeling is used to distinguish the different groups such as control, subtypes or treatment of tumors. Reference pool usually means control or untreated group for verifying the effect of treatment. Intensity ratio to reference pool is measure of the relative abundance of the same peptide in different samples. In general, relative

abundance is used rather than absolute abundance for comparing peptide quantity between different datasets since intensity distribution of datasets are different from each other. There are 20 studies with 1900 raw files which used isobaric labeling method in our data. The number of datasets is 39 and total 178 intensity ratio is computed.

3.4 Network architecture

Our network consists of three layer which is named Three Layer Proteomics (TLP) network. First layer of TLP network is protein layer, second layer is peptide layer, and third layer is modified peptide layer. This structure followed the workflow of bottom-up proteomics. Node is defined in each layer for representing the entities of results in bottom-up proteomics. Edge is defined in each layer and between layer for representing important information (relationship) between the entities which did not exploited in itself or together in this way. An entity–relationship model (ER model) of TLP network shown in figure 9.

In protein layer, the node is defined as a protein included in Uniprot database which we used in database search. Therefor node attribute of node becomes Uniprot ID.

The edge in protein layer represents the PPIs. We combine the information of PPIs from four reliable protein interaction databases, BioGRID, MINT, IntAct, and HPRD. All interactions these databases provide are derived from literature curation or direct user submissions and are freely available.

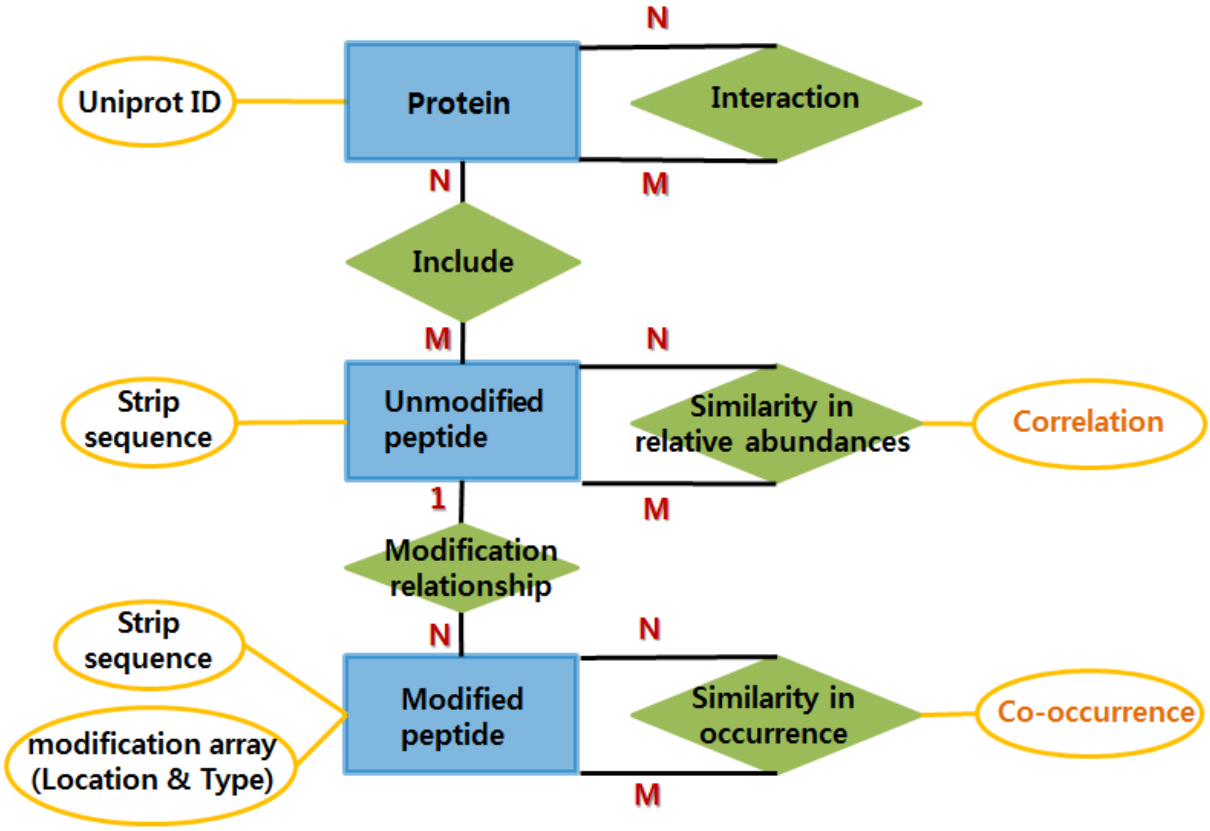

Figure 9. ER model of TLP network.

In peptide layer, the node is defined as an unmodified peptide which is strip sequence of peptide that identified by MS-GF+ search. Strip sequence is a series of amino acids which do not contain mass shift from PTM. Therefor strip sequence itself makes the node unique. The edge in peptide layer represents the similarity in relative abundances. For each peptide node, relative abundance is defined by 178 column vector as described above. Thus similarity in relative abundances is computed as a correlation between two intensity ratio vectors of each peptide node.

In addition to correlation, we use median absolute deviation (MAD) as a measure of the variability of the peptide. The MAD is robust measure of how spread out a set of

data is. The variance and standard deviation are also measures of spread, but they are more affected by extremely high or extremely low values and non-normality. Our data isn't normal, therefor the MAD is suitable statistic we can use instead.

Firstly, MAD cut-off is used to remove the peptide which is not informative. Informative peptide is defined whether its quantity variability is significant between the control group and treatment group. Second, among the informative peptides, correlation is calculated to find the similarity in relative abundances. Then correlation cut-off is used to identify the pairs which has similar pattern in relative abundances across 178 datasets.

Empirical statistics is used to decide the cut-off value of MAD and correlation. In order to obtain null distribution, the virtual intensity ratio is filled up with gaussian random variable from mean and standard deviation (SD) of each dataset. Then distribution of MAD and correlation from the gaussian random variable is used to decide the cut-off value.

Intuitively, a protein is a long string of elements (amino acids), and a peptide is a smaller substring. Therefor the edge between first layer (protein layer) and second layer (peptide layer) represents the intrinsic relationship between protein and peptide, that is to say "including" relationship.

In modified peptide layer, the node is defined as the peptide which has more than one PTM on its sequence. Biologically, peptides which have identical sequence and identical modification type with different modification site are different peptide.

Peptides which have identical sequence and different modification type are different peptide, obviously. Therefor strip sequence with the location and the type of modification makes the node unique.

We have to store two information about modification, the location of modification and the type of modification. These two information is stored as modification array. Modification array contains the information about location of modification as position of element. Value of element indicates the type of modification by a different single letter.

Since MS-GF+ limits the maximum length of the peptide to be considered in the peptide identification as a parameter MaxPepLength and we use 50 for the value of parameter, the maximum length of the modification array also becomes 50 except for last null element.

There is a number of modification which occurs protein N-term such as protein N-term acetylation, additional element before the 50 elements for representing modification on each amino acid. Final length of modification array is 53 with beginning element “M″ for representing the start of the modification array, the 51 element for representing modification, and the last null element.

There are total 16 modification types which identified by MS-GF+ search in our data. Among 16 modification types, several types belong to functional PTM which is important to protein function. The other types are not related to protein function, so

excluded from further analysis. The unique ID assigned to modification type and whether the type is functional modification or not is shown in Table 1.

Mass Shift	Representation in Modification Array	Modification Type	Functional Modification
0.984	d	Deamidation	
0.987	g	Glycosylation	v
15.995	o	Oxidation	
45.988	b	Fixed beta-Methylthiolation by MMTS	
57.021	c	Fixed Carbamidomethyl C	
79.966	p	Phosphorylation	v
17.027	r	Pyroglutamic acid formation from Gln	
18.011	r	Pyroglutamic acid formation from Glu	
42.011	a	Acetylation	v
333.169	s	Sulfenylation	v
339.207	s	Sulfenylation, heavy Dyn-2	v
28.031	i	Dimethylaion-light	
32.056	i	Dimethylation-heavy	
36.076	i	Dimethylation-intermediate	
71.037	n	Fixed Propionamide in C, S-carboxamidoethyl-L-cysteine	
14.016	m	Methylation	v

Table 1. Modification list.

The edge in modified peptide layer represents the similarity in occurrence of modified peptide node. It is named as co-occurrence. Contrary to proteomics, there are lots of research on the simultaneous expression of two or more genes which is named as co-expression.

A gene co-expression network is constructed by looking for pairs of genes which show a similar expression pattern across samples, which means the transcript levels of two co-expressed genes rise and fall together across samples. Gene co-expression networks are of biological interest since co-expressed genes are controlled by the same transcriptional regulatory program, functionally related, or members of the same

pathway. Similar to gene co-expression network, the idea of co-occurrence network can be proposed in proteomics research area and intuitively we can expect biologically significant information from the co-occurrence network.

The co-occurrence network is constructed by looking for pairs of modified peptides which show a similar occurrence pattern across datasets. We defined the individual occurrence of modified peptide as a binary value, whether modified peptide is found or not in the dataset. Since the number of datasets in our data is 900, each modified peptide node has the 900-length binary vector to represent individual occurrence across the datasets. Co-occurrence is defined as the number of datasets which the pair of modified peptides occurs together. Maximum value of the sum of individual occurrence becomes 900 and maximum value of the co-occurrence also does since the number of datasets in our data is 900.

In practice, co-occurrence is calculated on the modified peptide nodes which have at least one of functional PTMs, not entire modified peptide nodes. In terms of dataset, co-occurrence is calculated on the PTM enriched datasets, not entire datasets. The number of PTM enriched datasets is 400, so maximum value of the co-occurrence becomes 400.

Similar to edge attribute in unmodified peptide layer, cut-off value is used to remove the co-occurrence of pairs which may be meaningless. Empirical statistics is used to decide the cut-off value. In order to obtain null distribution, we permutated the individual occurrence of modified peptide nodes. Column permutation is applied on

individual occurrence vector while the sum of individual occurrence on each modified peptide node is maintained. As a results, the sum of occurrence of entire modified peptide nodes in each dataset appear to be uniform distribution, unlike the actual distribution.

Empirical statistics (green color) shows the much lower co-occurrence value compared to actual distribution (red color) in Figure 10. This means that modified peptide nodes in our data shows the statistically significant co-occurrence. Then cut-off value of co-occurrence is decided as nine by comparing empirical statistics and real distribution.

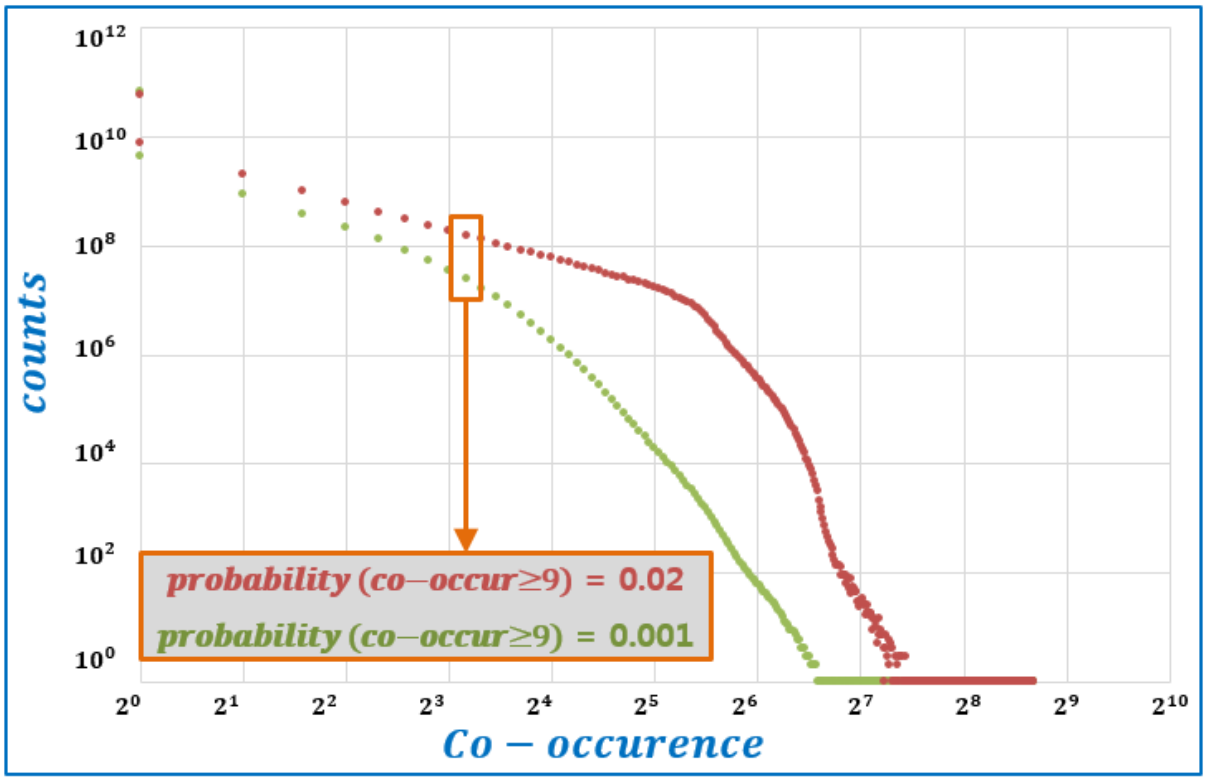

Figure 10. Co-occurrence log-log distribution.

The edge between the unmodified peptide layer and modified peptide layer represents the modification relationship. As it can be inferred from the definition of the unmodified peptide node and the modified peptide node, when the modified peptide is identified from the MS-GF+ search results, edge from the unmodified peptide spread branches to the modified peptide node.

3.5 Example of graph construction process

As an example of illustrating the entities and relationship between them, Figure 11 shows an example of graph construction process with three protein nodes, three unmodified peptide nodes, and four modified peptide nodes.

Since the node in protein layer is defined as the proteins included in Uniprot database, protein nodes is 91,797 finite set. Edges between these protein nodes are generated according to PPIs constructed by four protein interaction databases as described above. Relationship between the protein nodes is many-to-many.

When the unmodified peptide node with sequence **LLL** is identified by MS-GF+ search, unmodified peptide node **LLL** is generated. Then edge between the node and the protein node C, which has the Uniprot id F8VRK9 and sequence M**LLL**HRAVVLRLQQACRLKSIPSRICI-QAC**STN**DSFQPQRPSL, is generated. Since substring **LLL** can be found other protein node B, edge between them is also generated. Therefore, relationship between the protein and peptide is many-to-many.

The edge between unmodified peptide node with sequence **STN** and protein node C and the edge between unmodified peptide node with sequence AALT and protein A are generated in this way. In case of unmodified peptide, node is generated not only when the unmodified peptide itself is found in MS-GF+ search results, but also when the modified peptide is found in MS-GF+ search results.

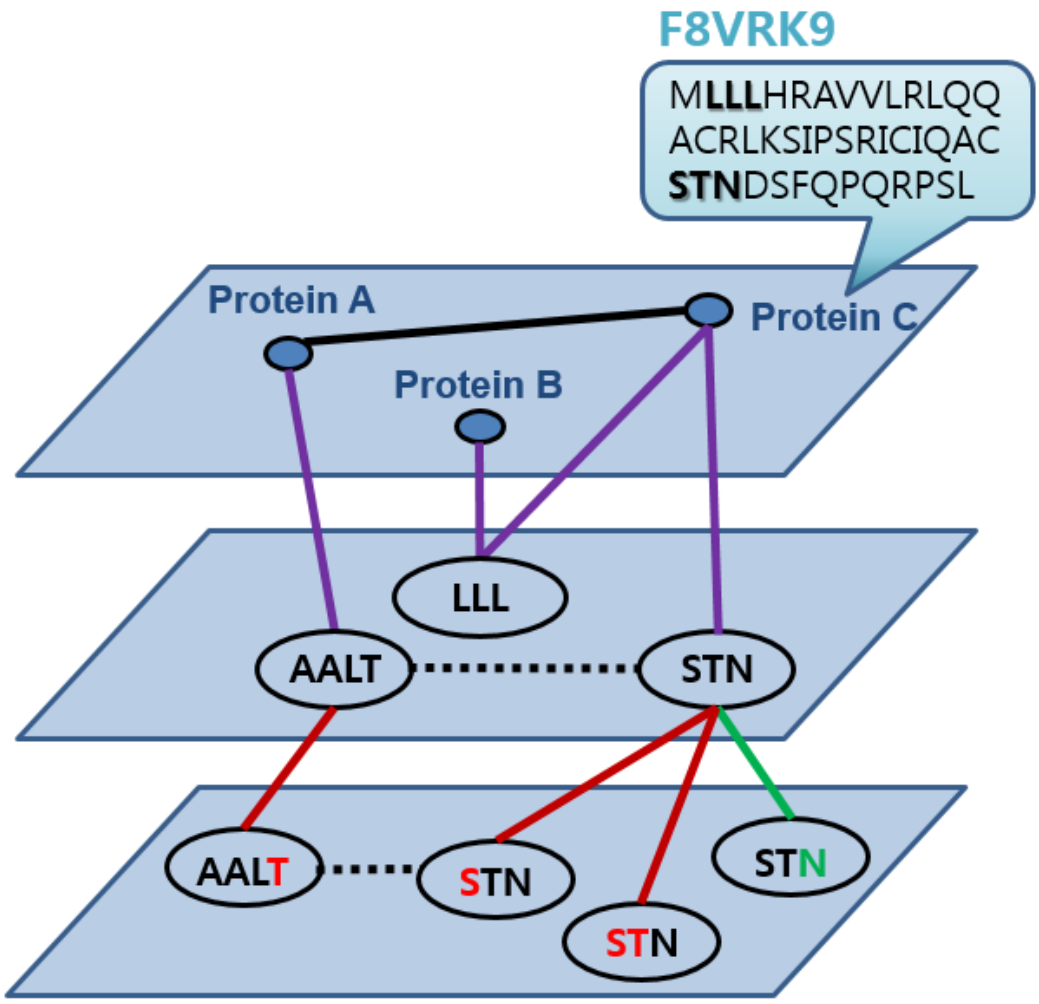

Figure 11. Example of graph construction.

Edge between the unmodified peptide nodes is generated by looking for the value of MAD and correlation of quantities as described above. Figure 12 shows the process of correlation computation from the quantification results. For example, edge between unmodified peptide node **AALT** and **STN** is generated since the value of MAD and correlation passed the cut-off value. According to definition, relationship between unmodified peptide nodes is many-to-many.

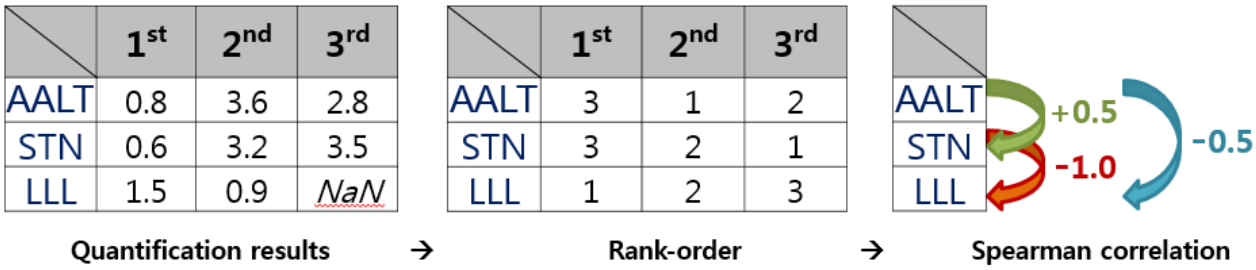

Figure 12. Computation of correlation.

When the four modified peptides showed in Figure 11 is identified by MS-GF+ search, each node is generated since these are different with respect to strip sequence or the location and the type of the modification. Different color on an amino acid indicates different type of modification in this Figure 11.

Edge between the unmodified peptide node and modified peptide node is one-to-many relationship since multiple modification can occurs on the peptide, but each modified peptide is subordinated to one unmodified peptide.

Edge between the modified peptide nodes is generated by looking for the value of co-occurrence as described above. Figure 13 shows the process of co-occurrence

computation from individual occurrence. For example, edge between modified peptide node **AALT** and **STN** is generated since the value of co-occurrence passed the cut-off value. According to definition, relationship between modified peptide nodes is many-to-many.

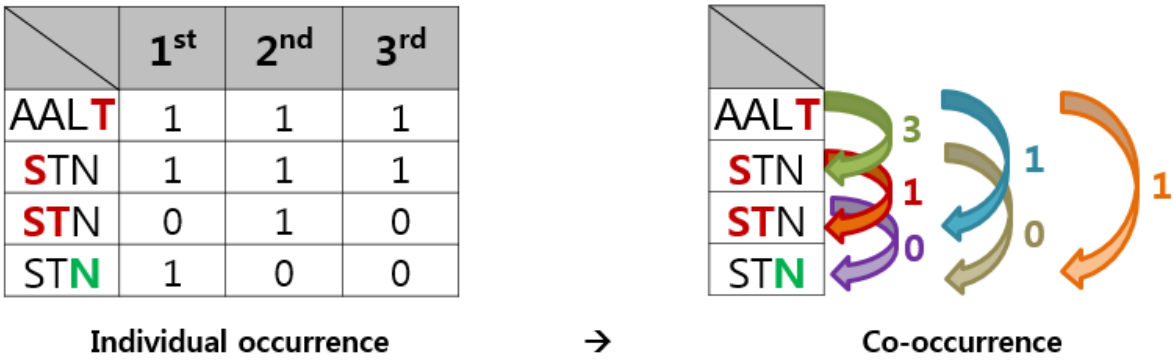

Figure 13. Computation of co-occurrence.

3.6 Network size

Figure 14 shows the size of TLP network. Since protein nodes is 91,797 finite set, the number of nodes in protein layer is also 91,797. The number of edges (PPIs) in protein layer is 115,209. The number of edges between protein layer and unmodified peptide layer is approximately four million. That is, one protein has four peptide nodes on average.

The number of nodes in unmodified peptide layer is approximately one million. This value is proportional to the number of spectrum data to some degree, but the rate of increase become smaller gradually. The number of edges in unmodified peptide layer would be 250 billion before applying the cut-off value of MAD and correlation. After applying the cut-off value, the number of edges would be approximately one billion.

The number of edges between unmodified peptide layer and modified peptide layer equals the number of nodes in modified peptide layer since the modified peptide node is connected with only one unmodified peptide.

The number of the nodes in modified peptide layer is one million. This value is also proportional to the number of spectrum data to some degree, but the rate of increase become smaller gradually. Since edge in modified peptide layer, co-occurrence, is calculated on only functional PTM nodes as described above, the number of edges in modified peptide layer is related to the number of functional PTM nodes. The number of functional PTM nodes is 370 thousand in our data. So the number of edges in modified peptide layer would be 70 billion before applying the cut-off value of co-occurrence. After applying the cut-off value, the number of edges would be one billion.

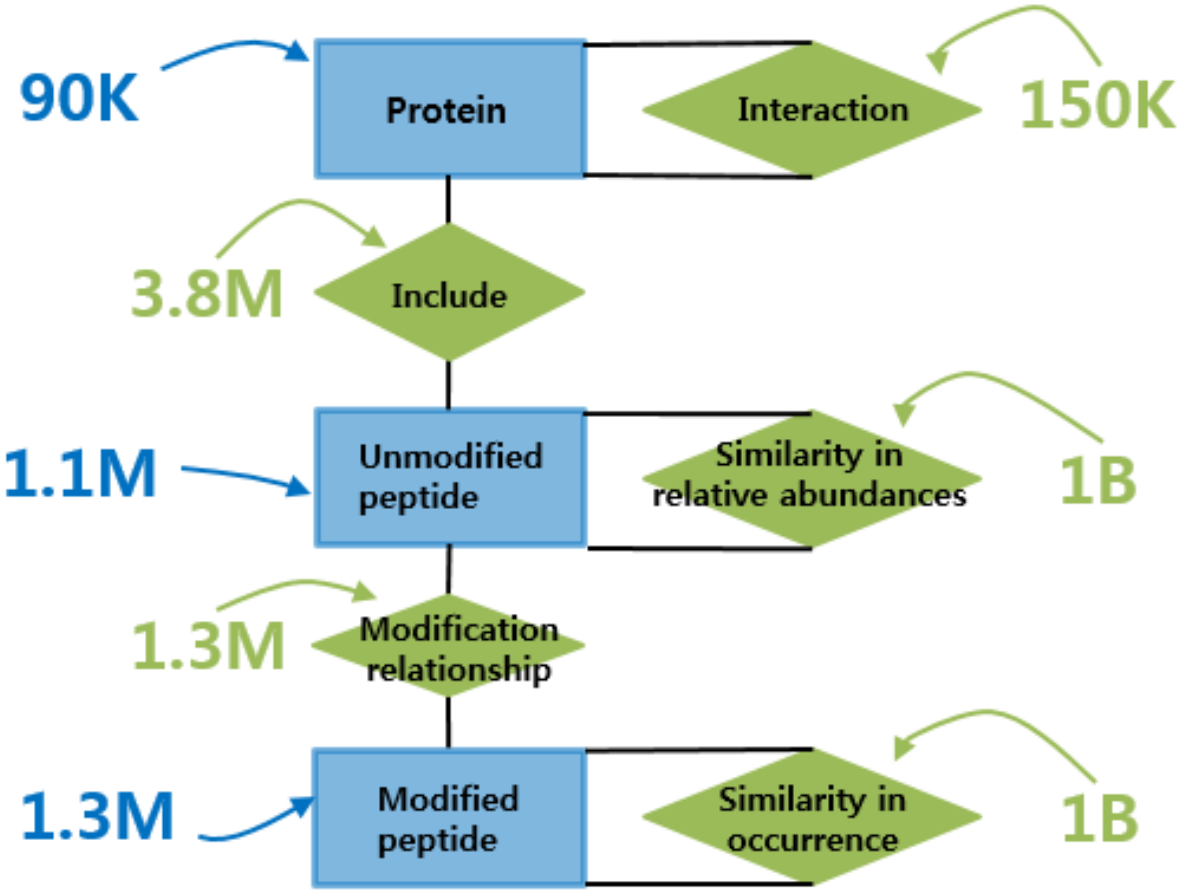

Figure 14. Size of TLP network after applying cut-off.

4. Analysis of TLP network

We find the close relationships across the three layer. As expected, PPIs network is similar to PTM co-occurrence network. Figure 15 shows the four subgraph where PPI network looks like cluster and PTM nodes has high co-occurrence value more than 100. These subgraphs related with DNA repair, protein homeostasis, RNA processing, mitogen-activated protein kinase (MAPK) pathway respectively.

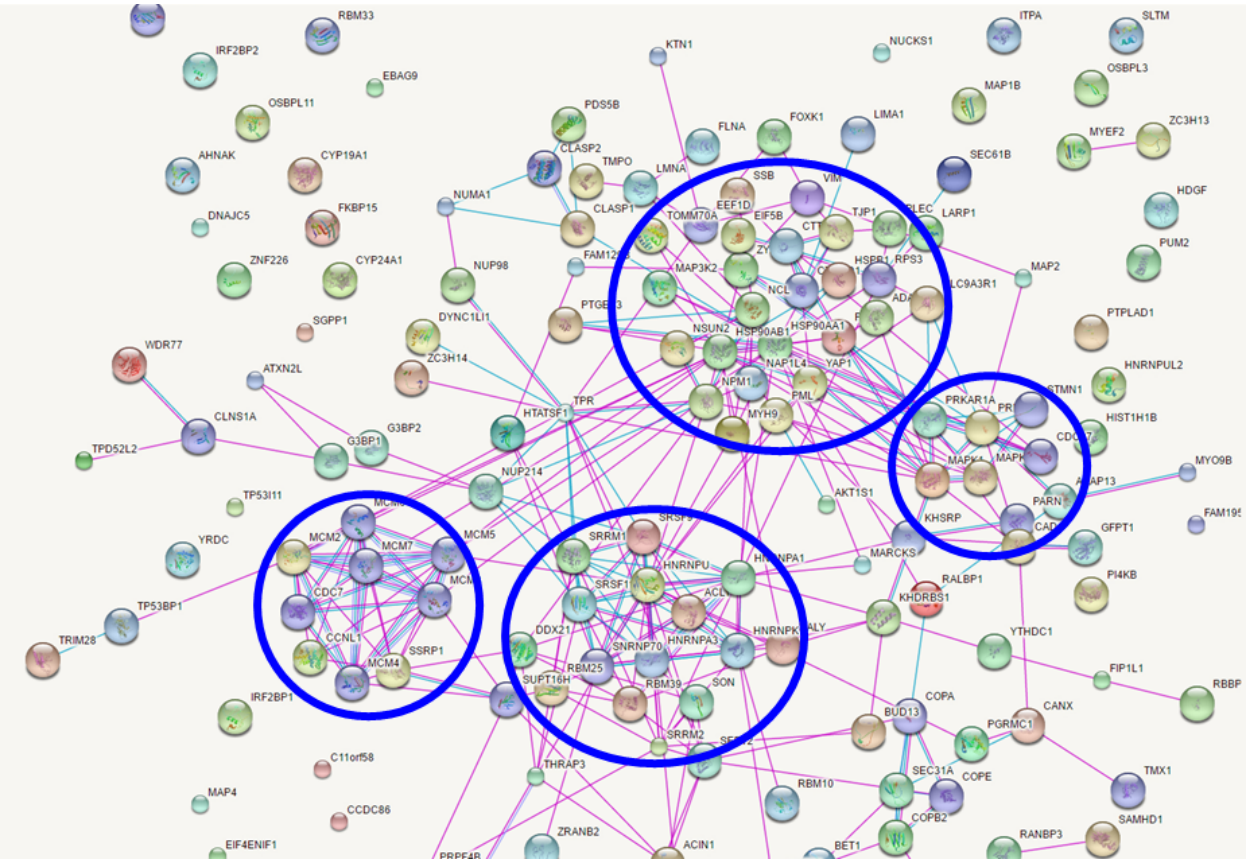

Figure 15. PTM bases found in TLP network.

This kind of subgraph is named as PPI-PTM base. In PPI-PTM bases, PPI network looks like cluster and PTM nodes has high co-occurrence. These bases can be used to interpret related with DEPs.

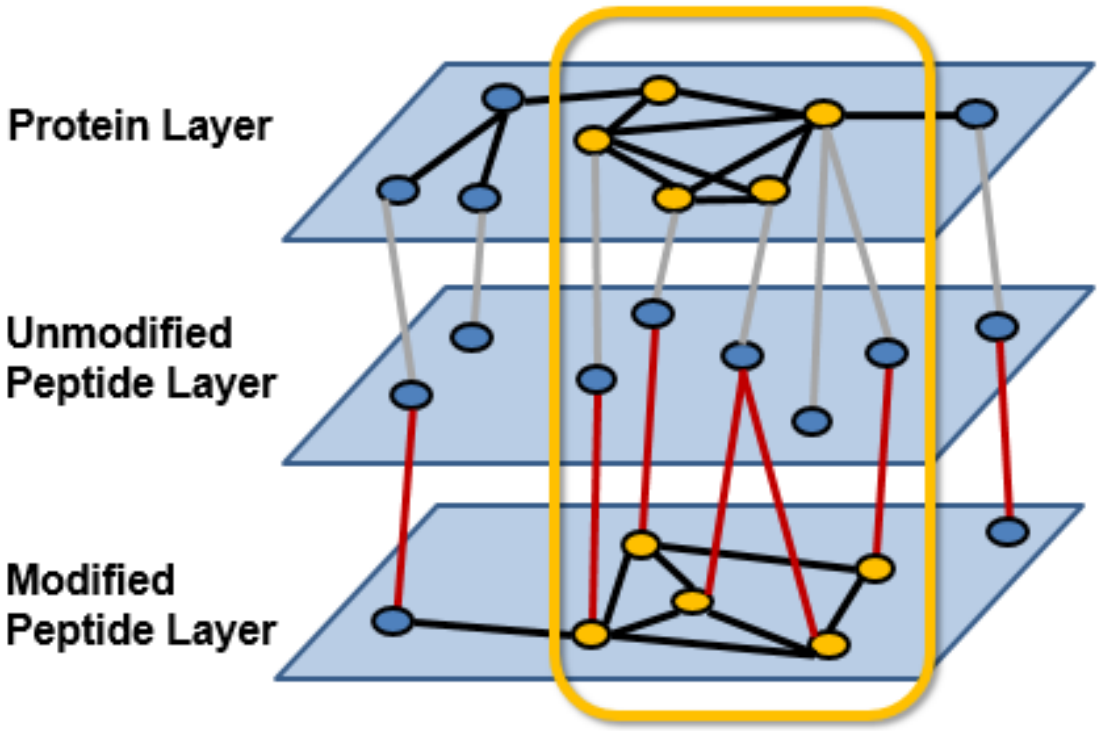

Figure 16. Illustration of PPI-PTM base.

Like as the PPI-PTM base, the subgraph can be found between protein layer and peptide layer where PPI network becomes cluster and peptide nodes has high correlation in abundance pattern. This kind of subgraph is named as peptide abundance base. These kinds of bases also can be used to interpret related with DEPs.

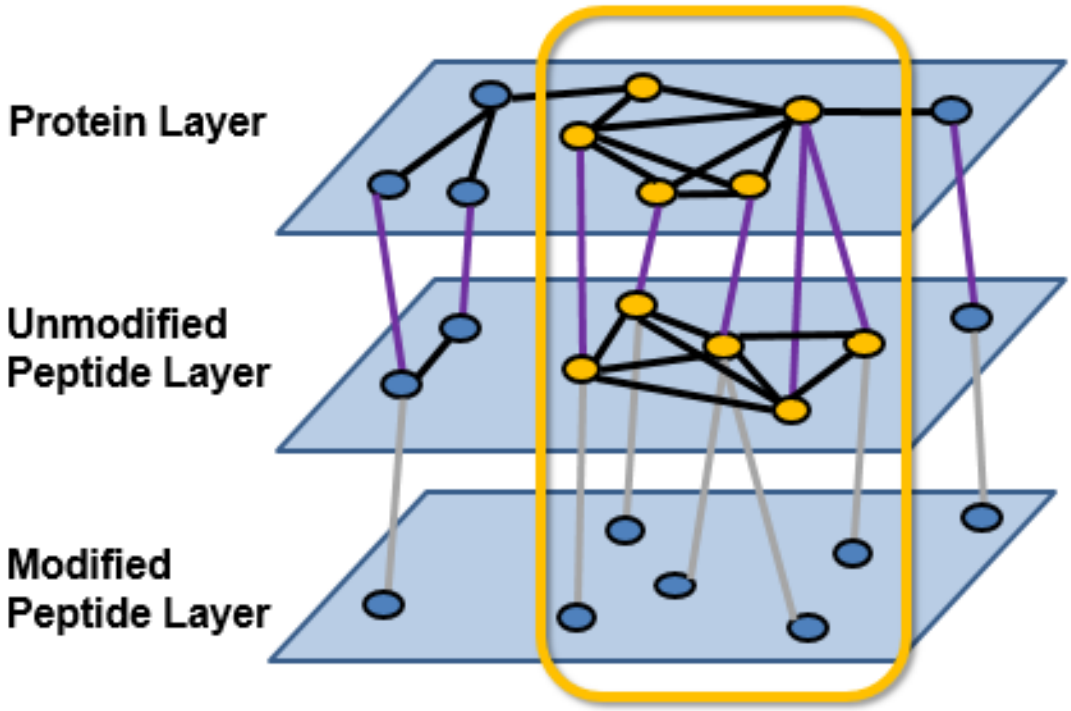

Figure 17. Illustration of peptide abundance base.

First of all, we will confirm the existing knowledge as network verification. For example, looking for PPI network and PTM nodes connected with these proteins of identical pathway can offers the information about how TLP network appears with the existing knowledge about pathway. Then knowledge discovery in TLP network can be conducted in various aspects. There are numerous questions in proteomics including PTM base and peptide abundance base. For example, finding informative peptides for protein quantification also becomes important discovery. Efficient algorithm that can find biologically significant results would be used.

5. Conclusion

We build the multilayer network for integrating bottom-up proteomics data which is named TLP network. TLP network contains diverse biological information including the peptide expression data, and PTMs as well as Protein–protein interactions (PPIs). All those things are closely connected to each other with respect to function of protein. Especially we utilize the edge attributes to exploit the relationship between this information. Weight on edge can represents the strength of relationship, or potential biological significance.

We find close relationships across the three layer. PPI clusters are consistent with PTM high co-occur clusters. PPI-PTM bases and peptide abundance bases would be used to interpret related with DEPs.

As a result, TLP network is expected to answer a wide range of questions in proteomics research area including informative peptides from big data for protein quantification, Co-occurring PTMs in the same pathway, and Frequently co-changing network modules that can be used as bases for interpretation of a new dataset.

REFERENCE

[1] Zhang, Bing, Matthew C. Chambers, and David L. Tabb. "Proteomic parsimony through bipartite graph analysis improves accuracy and transparency." Journal of proteome research 6.9 (2007): 3549-3557.

[2] Nesvizhskii, Alexey I. "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics." Journal of proteomics 73.11 (2010): 2092-2123.

[3] Cox, Jürgen, and Matthias Mann. "MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification." Nature biotechnology 26.12 (2008): 1367-1372.

[4] Dost, Banu, et al. "Accurate mass spectrometry based protein quantification via shared peptides." Journal of Computational Biology 19.4 (2012): 337-348.

[5] Chen, Hsinchun, et al., eds. Medical informatics: knowledge management and data mining in biomedicine. Vol. 8. Springer Science & Business Media, 2006.

[6] Na, Seungjin, Nuno Bandeira, and Eunok Paek. "Fast multi-blind modification search through tandem mass spectrometry." Molecular & Cellular Proteomics 11.4 (2012): M111-010199.

[7] Eng, Jimmy K., Ashley L. McCormack, and John R. Yates. "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database." Journal of the American Society for Mass Spectrometry 5.11 (1994): 976-989.

[8] Cottrell, John S., and U. London. "Probability-based protein identification by

searching sequence databases using mass spectrometry data." electrophoresis 20.18 (1999): 3551-3567.

[9] Nesvizhskii, Alexey I., et al. "A statistical model for identifying proteins by tandem mass spectrometry." Analytical chemistry 75.17 (2003): 4646-4658.

[10] Nahnsen, Sven, et al. "Tools for label-free peptide quantification." Molecular & Cellular Proteomics 12.3 (2013): 549-556.

[11] Martens, Lennart, et al. "PRIDE: the proteomics identifications database." Proteomics 5.13 (2005): 3537-3545.

[12] Vizcaíno, Juan Antonio, et al. "2016 update of the PRIDE database and its related tools." Nucleic acids research 44.D1 (2016): D447-D456.

[13] Edwards, Nathan J., et al. "The CPTAC data portal: a resource for cancer proteomics research." Journal of proteome research 14.6 (2015): 2707-2713.

[14] Malliaros, Fragkiskos D., and Michalis Vazirgiannis. "Clustering and community detection in directed networks: A survey." Physics Reports 533.4 (2013): 95-142.

[15] Xu, Xiaowei, et al. "Scan: a structural clustering algorithm for networks." Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007.

[16] Ross, Philip L., et al. "Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents." Molecular & cellular proteomics 3.12 (2004): 1154-1169.

[17] Thompson, Andrew, et al. "Tandem mass tags: a novel quantification

strategy for comparative analysis of complex protein mixtures by MS/MS." Analytical chemistry 75.8 (2003): 1895-1904.

[18] Ong, Shao-En, et al. "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics." Molecular & cellular proteomics 1.5 (2002): 376-386.

[19] Kim, Sangtae, and Pavel A. Pevzner. "MS-GF+ makes progress towards a universal database search tool for proteomics." Nature communications 5 (2014).

요약문

빅 데이터에 기반하여
단백질체학 데이터에서의 수량화 규칙을 찾기 위한
다층 네트워크 구축과 분석

데이터를 어떻게 모델링 하느냐는 데이터를 이해하고 그 안의 정보를 찾아내고자 할 때 중요한 역할을 한다. 그리고 내재된 정보를 찾아내기 위해서는 다양한 디자인이 가능하다. 현재 단백질체학의 연구들은 주로 단백질-단백질 상호작용에 관한 네트워크만에 관심을 가지며 데이터 모델링 측면에서 한계를 보이고 있다.

이 논문에서 우리는 단백질의 기능 메커니즘을 이해하기 위한 새로운 접근법을 제안하고자 한다. 우리는 TLP 네트워크라고 명명된 단백질체학 데이터를 통합하기 위한 다층 네트워크를 만들었다. TLP 네트워크는 단백질-단백질 상호작용 뿐 아니라 펩타이드 발현, 단백질 변형을 포함하는 다양한 생물학 정보를 담고 있다. 새로운 데이터 모델인 TLP 네트워크를 통하여 단백질체학에서의 여러가지 중요한 질문들에 대한 답을 찾을 수 있을 것으로 예상된다.