Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

Gene Expression

Characterization of peptide-protein relationships in protein ambiguity groups via bipartite graphs

Karin Schork [1,2,3], Michael Turewicz [1,2], Julian Uszkoreit [1,2], Jörg Rahnenführer [3,†], Martin Eisenacher [1,2,*,†]

[1] Medizinisches Proteom-Center, Medical Faculty, Ruhr-University Bochum, 44801 Bochum, Germany
[2] Medical Proteome Analysis, Center for Protein Diagnostics (PRODI), Ruhr-University Bochum, 44801 Bochum, Germany
[3] Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany

[*] To whom correspondence should be addressed.
[†] These authors contributed equally to this work.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: In bottom-up proteomics, proteins are enzymatically digested before measurement with mass spectrometry. The relationship between proteins and peptides can be represented by bipartite graphs. This representation is useful to aid protein inference and quantification, which is complex due to the occurrence of shared peptides. We conducted a comprehensive analysis of bipartite graphs using theoretical peptides from *in silico* digestion of protein databases as well as quantified peptides quantified from real data sets.
Results: The graphs based on quantified peptides are smaller and have less complex structures compared to graphs using theoretical peptides. The proportion of protein nodes without unique peptides and of graphs that contain such proteins are considerably greater for real data. Large differences between the two analyzed organisms (mouse and yeast) on database as well as quantitative level have been observed. Insights of this analysis may be useful for the development of protein inference and quantification algorithms.
Availability: R Code available at https://github.com/mpc-bioinformatics/bipartite-peptide-protein-graphs
Contact: martin.eisenacher@rub.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The digestion of intact proteins to peptides via enzymes like trypsin is a requirement for high-troughput bottom-up proteomics based on mass spectrometry (MS) (Zhang *et al.*, 2013; Gillet *et al.*, 2016; Dupree *et al.*, 2020). Because of this, peptides are identified and quantified directly from the MS measurements instead of proteins. The protein ambiguity problem describes the challenge to built a list of proteins for which evidence exists that they are present in the respective sample using protein inference methods (Audain *et al.*, 2017; Huang *et al.*, 2012). Many identified peptides cannot unambiguously be assigned to a single protein, as the respective peptide sequence is part of multiple protein sequences in the underlying database. These shared peptides lead to the need to form protein ambiguity groups, which consist of proteins whose presence in a sample cannot be completely decided using the peptides identified in the sample. Beyond this, for protein quantification peptide quantities need to be summarized to protein quantities. This is often done by using only unique peptides, although including shared peptides would be desirable to exploit the full potential of the data and to avoid inaccuracies (Blein-Nicolas *et al.*, 2012; Dost *et al.*, 2012; Zhang *et al.*, 2015). The worst case, a valid quantification of proteins without any quantified unique peptide, is very challenging and not solved yet despite the large variety of different protein quantification methods that exist (Blein-Nicolas and Zivy, 2016; Ahrné *et al.*, 2013).

To illustrate and facilitate the protein inference and quantification steps, bipartite graph representations of the protein-peptide relationships have been used (Pavlopoulos *et al.*, 2018; Zhang *et al.*, 2007; Gerster *et al.*, 2010; Bamberger *et al.*, 2018). A bipartite graph is an undirected graph G (with node set $N(G)$ and edge set $E(G)$) whose nodes are grouped into two sets $N_1(G)$ and $N_2(G)$, so that each edge connects a node of the first set with a node of the second set. In the context of protein-peptide-relationships, proteins are represented by nodes from $N_1(G)$ and peptides

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

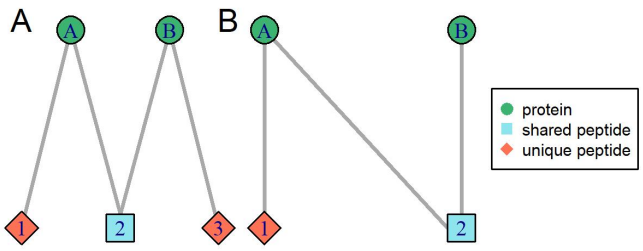

Fig. 1. Examples of bipartite graphs depicting protein-peptide relationships. Protein nodes are shown as green circles and peptide nodes as red diamonds (unique to a protein node) or blue squares (shared between protein nodes). Each node may represent multiple protein accessions or peptide sequences.

by nodes from $N_2(G)$. One straightforward approach would be to add an edge if and only if the considered peptide sequence is part of the protein sequence based on database knowledge. Another possibility is to draw an edge only if the corresponding peptide was quantified in a specific data set (see Section 2.3). As not all proteins are connected to each other via chains of shared peptides, the bipartite graph for the whole data set can be divided into smaller connected components. These components can be analysed and handled separately during protein inference or quantification.

The structure of the connected components is determined by the underlying protein sequence database and contains valuable information on how difficult the protein inference or quantification problem is in this specific case. The M-shaped example in Fig. 1A shows two protein nodes that each have one unique peptide node and there is a shared peptide node connecting both protein nodes. In this case, the protein inference step is quite simple, as both proteins are likely to be present in the underlying samples unique peptides exists. For protein quantification however, the shared peptides contain valuable information that should not be discarded and on the other hand could also lead to contradictions, if the shared peptides' quantities do not fit to the quantities from the unique peptides. In the second example (N-shape, Fig. 1B), protein B does not have a unique peptide. In this case it is unclear, if this protein exists in the sample. Even if it does, it is not straightforward to quantify this protein, as it only has shared peptides whose quantities are also influenced by protein node A. Although bipartite graphs have been used to display the relationships between proteins and peptides and to assist protein inference or quantification (Zhang *et al.*, 2007; Gerster *et al.*, 2010; Bamberger *et al.*, 2018), a comprehensive analysis and characterization of the occurring graph types has yet not been conducted. In this paper, we generate bipartite graphs of peptide-protein relationships using peptides quantified by mass spectrometric proteomics in two different real data sets as well as theoretical peptides from an *in silico* digestion of the related protein databases. We characterize common types of graphs as well as aggregated metrics over all graphs of a data set and compare results for theoretical and quantified peptides.

2 Methods

2.1 Data sets and corresponding protein databases

Two quantitative peptide-level data sets were used the corresponding protein databases. The samples of the first data set (D1) consist of 13 non-mouse proteins that were spiked into a mouse cell background (C2C12 cell lysate) in different concentrations leading to five different states (Barkovits *et al.*, 2020). Each state was measured in three replicates leading to 15 samples in total. The raw files and the protein database were taken from the PRIDE repository (Perez-Riverol *et al.*, 2019), identifier PXD012986. Data were re-analyzed using a KNIME workflow employing the search engines Mascot 2.5 (Perkins *et al.*, 1999), MS-GF+ (Kim *et al.*, 2010) and X!Tandem (Craig and Beavis, 2004). Peptide identifications were combined using PIA (Uszkoreit *et al.*, 2015, 2019). The workflows were similar to the original publication (Barkovits *et al.*, 2020), except for a peptide identification filter directly after the spectrum identification leaving only peptides which are strictly tryptic (i.e. cutting after arginine (R) and lysine (K), if not followed by proline (P)). The resulting peptide quantities (quantified_peptides-featureFinderCentroided.csv) were used for further analysis. The corresponding protein sequence database (in total 52,824 entries) consists of 52,548 mouse proteins (UniProt reference mouse proteome version 2017_12), 13 spike-in proteins, 147 spike-in contaminants and 115 contaminants from the cRAP database.

For the second data set (D2) the UPS1 standard (containing 48 human proteins of equal molarity) were spiked into a yeast background in 10 different concentrations (Ramus *et al.*, 2016). Raw files were taken from the PRIDE repository, identifier PXD001819, and re-analyzed with MaxQuant version 1.6.3.4 (Cox and Mann, 2008) and Andromeda (Cox *et al.*, 2011). Settings were taken from Workflow 7 of the corresponding publication (Ramus *et al.*, 2016), except that Trypsin (no cutting after proline) was used instead of Trypsin/P. The peptides.txt file was used for further processing of the quantitative peptide-level data. The corresponding database (in total 6,371 entries) consists of the UniProt yeast proteome with isoforms (version 2019_11, 6,078 entries), the UPS1 fasta file[1] (48 entries), as well as the contaminants database[2] provided by Andromeda (245 entries).

For each of the two protein databases, all protein sequences were digested *in silico* using the digestion rules for trypsin (cut after R and K, if not followed by P) to obtain all theoretical tryptic peptides. Peptides with a maximum of two missed cleavage sites were allowed. Only peptides with a length between five and 50 amino acids (AA) were considered further, as longer or shorter peptides are unlikely to be found via bottom-up MS (Swaney *et al.*, 2010). Different values for the lower bound for the peptide length (five, six, seven and nine) were evaluated (see Section 3.1).

2.2 Processing of quantitative peptide-level data

For data set D1, peptides exclusively stemming from decoy proteins were filtered out. Entries for the same peptide sequence but with different post-translational modifications (PTMs) were aggregated by summing up the corresponding intensities. Peptide sequences with a length larger than 50 AA or lower than the chosen minimal peptide length from Section 3.1 were removed, to be directly comparable with the theoretical peptides. The remaining sequences were mapped to the proteins in the corresponding database. The three replicates for each state were combined by averaging the peptide intensities, while zero values were treated as missing values (NA). If two or more of the three values were NA, the combined peptide intensity for this peptide was set to NA. For each possible comparison of the five different states, these aggregated values were used to calculate peptide ratios, which were set to NA if one of the corresponding values was NA. For each comparison, only peptides with a valid ratio were considered in the construction of the bipartite graphs (see Section 2.3). For data set D2, peptide LFQ values were extracted from the peptides-txt table and values of zero were re-assigned by NA. Further processing was the same as explained for data set D1.

[1] https://www.sigmaaldrich.com/content/dam/sigma-aldrich/life-science/proteomics-and-protein/ups1-ups2-sequences.fasta

[2] http://www.coxdocs.org/doku.php?id=maxquant:start_downloads.htm

Table 1. Influence of different minimal peptide lengths on the bipartite graphs.

	D1_fasta				D2_fasta			
	min 5 AA	min 6 AA	min 7 AA	min 9 AA	min 5 AA	min 6 AA	min 7 AA	min 9 AA
protein accessions	52,760	52,752	52,741	52,679	6,365	6,364	6,362	6,362
protein nodes	52,405	52,388	52,364	52,250	6,292	6,291	6,290	6,290
peptide sequences	2,872,496	2,777,410	2,642,464	2,365,974	752,585	718,527	680,043	603,459
peptide nodes	133,137	103,517	96,241	92,346	13,008	8,178	7,609	7,412
edges	444,272	277,656	241,601	221,186	26,917	15,108	13,730	12,754
graphs	2,916	10,356	15,402	17,114	1,649	4,908	5,471	5,604
graphs with 1 protein node	2,004	4,809	6,355	7,147	1,499	4,276	5,023	5,242
isomorphism classes	206	1,474	2,583	2,692	19	66	43	38
largest graph*								
protein nodes	47,525	25,006	1,632	939	4,433	117	69	69
peptide nodes	126,315	57,763	3,971	2,415	10,941	358	264	253
edges	433,174	185,147	14,093	8,751	24,540	1,296	3,329	3,033
second largest graph*								
protein nodes	26	91	1,104	529	21	70	52	50
peptide nodes	75	177	2,940	1,046	37	267	203	181
edges	288	486	10,541	2,484	131	3,359	743	613

* In terms of number of protein nodes.

2.3 Construction of bipartite graphs

For D1 and D2, the bipartite protein-peptide graphs were constructed using two different sets of peptides. First, all theoretical peptides obtained from *in silico* digestion of the corresponding protein database were used (D1_fasta and D2_fasta, see Section 2.1). Second, for each comparison of two states only those peptides with a valid ratio are used (D1_quant and D2_quant, see Section 2.2). The connected components that arise from the pairwise comparisons of the different states within one data set were put together and analyzed jointly.

To construct the bipartite graphs, a matrix with peptides in rows and all corresponding proteins in columns is created, with entries of 1 if the peptide belongs to the respective protein and 0 otherwise. This matrix is used as an biadjacency matrix (Pavlopoulos *et al.*, 2018) to build the bipartite graph of the relationship between proteins and peptides. Proteins that have the same set of peptides are indistinguishable and are therefore collapsed to a single node, i.e. a protein group. Similarly, all peptides sequences that belong to exactly the same set of proteins are collapsed to one single node to further simplify the graph structure. Accordingly, in Fig. 1, one node may stand for multiple protein accessions or peptide sequences. In the resulting bipartite graph, protein nodes can be connected to each other via chains of shared peptide nodes. As not all protein nodes are connected in this way, the graph can be decomposed into connected components, where all nodes are connected to each other, but not to any other node outside of it. The connected components (called "graphs" in the following for simplicity) can be treated separately for protein inference or protein quantification provided that each MS spectrum was assigned to only one peptide sequences or the data have been filtered accordingly.

Two graphs G and H are called isomorphic if there exists a bijective function $f : V(G) \rightarrow V(H)$ that maps the node set of G to the one of H, so that the edge structure remains the same $((u, v) \in E(G) \Leftrightarrow (f(u), f(v)) \in E(H)\ \forall u, v \in N(G))$ (Junttila and Kaski, 2007). Additionally, we require that the function f preserves the node type, i.e. that protein and peptide nodes are distinguished in terms of isomorphisms. Groups of isomorphic graphs reflect identical structures of protein-peptide relationships and contain the same information needed for inference or quantification, except the specific quantities of the peptides. Sets of isomorphic bipartite graphs (isomorphism classes) are built for the two data sets, based theoretical (D1_fasta, D2_fasta) and quantified peptides with a valid ratio (D1_quant, D2_quant).

2.4 Software

Calculations were performed and graphics created using R 4.0.1 (R Core Team, Vienna, Austria). The package seqinr (4.2-4) was used to import the fasta files and OrgMassSpecR (0.5-3, modified version of the Digest() function) to perform the *in silico* digestion. For dealing with (bipartite) graphs, the igraph package (1.2.6) was used, and for further graphics ggplot2 (3.3.2). Additionally, function from the following R packages were used: reshape2 1.4.4, BBmisc 1.11, pbapply 1.4-3, limma 3.44.3, Matrix 1.2-18, openxlsx 4.2.2, xtable 1.8-4, tidyverse 1.3.0, cowplot 1.1.0, ggpubr 0.4.0. R code that was used for this work is available at: https://github.com/mpc-bioinformatics/bipartite-peptide-protein-graphs.

3 Results

3.1 General overview and impact of minimal peptide length

Very small and very large peptides cannot be measured by MS for high-throughput bottom-up proteomics (Swaney *et al.*, 2010). During peptide identification, the length of peptides is usually restricted by limiting the number of amino acids as a setting in the search engine. Even if small peptides are allowed during database search, often only a few are identified and subsequently quantified in the end (see Fig. S1/S2 B and D). However, on database level, many small peptides exist after tryptic digestion. Especially peptides with a length of five amino acids (which are usually not identified or quantified) show a larger proportion of shared peptides compared to longer peptides (see Fig. S1/S2 A and B). We analyzed the influence of different minimal peptide lengths on the overall bipartite graph structures (Table 1). We chose six and seven, as common default values for search engines, five as an extreme lower bound, as well as nine, as the required length by the Human Proteome Project of the Human Proteome Organization (HUPO HPP) [3].

[3] https://www.hupo.org/HPP-Data-Interpretation-Guidelines

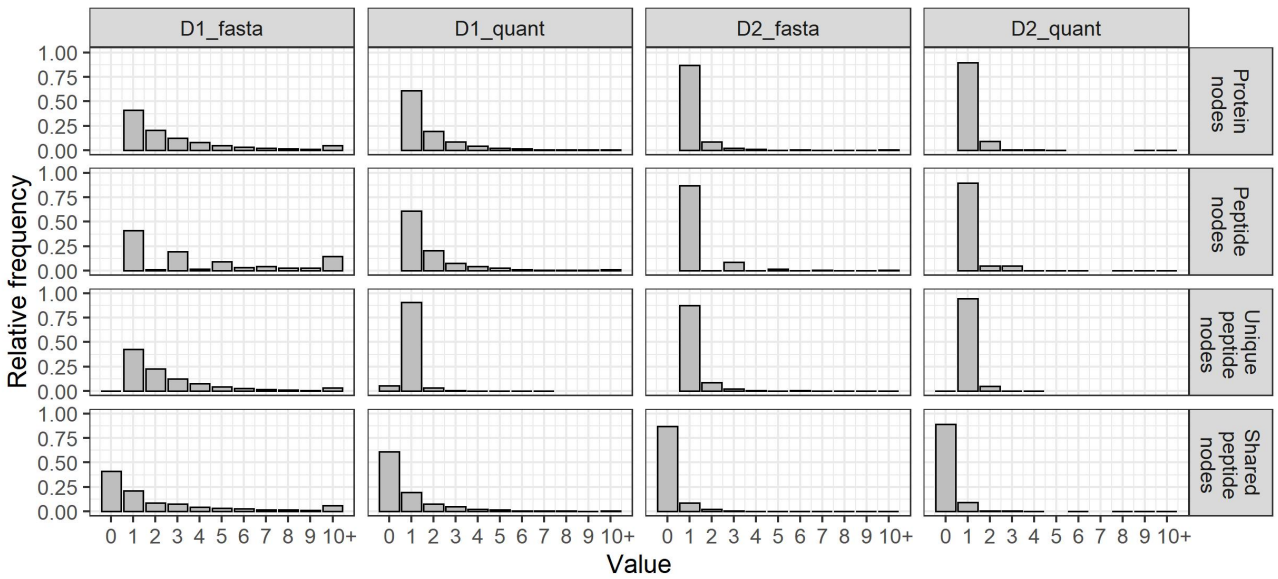

Fig. 2. Distribution of numbers of different node types (protein nodes, peptide nodes, unique peptide nodes and shared peptide nodes) for the data sets D1 and D2, on database level (*_fasta) and on the level of quantified peptides (*_quant). The rightmost bar with the label "10+" comprises the values of 10 and above.

For D1_fasta, 52,824 protein sequences from the database were *in silico* digested to around 2.87 million peptides, including peptides with a length between five and 50 amino acids as well as up to two missed cleavages. 64 protein sequences for which no tryptic peptide within the desired length was found were excluded from the analysis. The number of protein nodes is only slightly lower than the number of protein accessions (this is expected, as only few database proteins should be indistinguishable considering their tryptic peptides). However, the number of peptide nodes is notably lower than the number of peptide sequences (only about 130,000). The bipartite graph covering all these nodes consists of 2,916 connected components that form 206 different isomorphism classes. The largest class consists of the smallest possible graph with only one protein and one unique peptide node, which is an easy case for protein inference or quantification. This system occurs 2,004 times, making up around 69% of all graphs and covering 3.8% of all protein nodes. The largest graph (in terms of the number of protein nodes) contains over 47,000 protein nodes and therefore around 90% of all proteins nodes and almost 95% of all peptide nodes. All these proteins are connected somehow via chains of shared peptides. The second largest graph is much smaller with only 26 protein nodes and 75 peptide nodes.

Almost 100,000 short peptides of length five are included in the above analysis, but in database searches these are rarely identified with an acceptable score (see Fig S1B). When these are omitted (minimum peptide length six), the number of graphs more than triples to over 10,000. A major influence on this is the fragmentation of the former largest graph, which splits into 7,412 graphs. With minimal peptide length six, still about half of all protein and peptide nodes are connected within one single graph. Seven protein nodes had to be removed as they only had connections to peptide sequences of length five. The number of different isomorphism classes also increases dramatically to almost 1,500. If the minimal peptide length is further increased to seven, there are around 15,000 graphs that belong to more than 2,500 different isomorphism classes. The largest graph now only contains around 1,600 protein nodes and has a similar size as the second largest graph. For a minimal peptide length of nine, the graphs decompose further, but this process is slowed down. For example, the number of graphs rises only by around 1,700 when omitting peptides of length seven and eight. Going from six to seven amino acids as a minimal length, this increase is much higher, with over 5,000 additional graphs.

The second database (D2_fasta) contains 6,371 protein accessions, much less than D1_fasta. When using five AA as minimal length, only six protein accessions did not produce at least one peptide with the given criteria. More than 750,000 peptides are created during the *in silico* digestion. After collapsing of nodes, around 6,300 protein and 13,000 peptide nodes remain. There are 1,649 graphs, but 1,499 (90.1 %) of them contain only one peptide node and protein node. There are only 19 different types of graphs. The largest graph contains 70.5% of all protein nodes and 84.1% of all peptide nodes. Increasing the minimal peptide length to six decreases the number of peptides by about 34,000 and the number of peptide nodes by almost 5,000. The number of protein nodes is only reduced by one. The number of graphs almost triples to 4,908 while the number of different isomorphism classes rises to 66. The fraction of graphs with only one protein node decreases slightly to 87.1%. The size of the largest graph decreases considerably to 117 protein nodes and 358 peptide nodes, which is more similar to the size of the second largest graph. For a minimum of seven amino acids the number of peptide nodes further decreases to around 7,600, while the number of protein nodes only decreases by one. The number of graphs increases further to almost 5,500. The number of different types of graphs decreases to 43, while the percentage of graphs with only one protein node increases to 91.8%. The size of the largest graph also decreases further. By increasing the minimal peptide length further to nine amino acids, the number of graphs rises to over 5,600, which is a slower increase than before. The number of protein nodes in the largest and second largest graphs are not influenced by omitting peptides of length seven and eight, however, they now contain fewer peptide nodes and edges.

In summary, the minimal peptide length has a huge impact on the graph structures and especially on the size of the largest graph. For the following analyses, a minimal peptide length of seven for D1_fasta and of six amino acids for D2_fasta is used. This ensures that the largest graph is decomposed far enough to yield a reasonable size comparable to the second largest graph. 370 quantified peptides with a length below seven or above 50 are excluded from D1_quant to ensure direct comparability

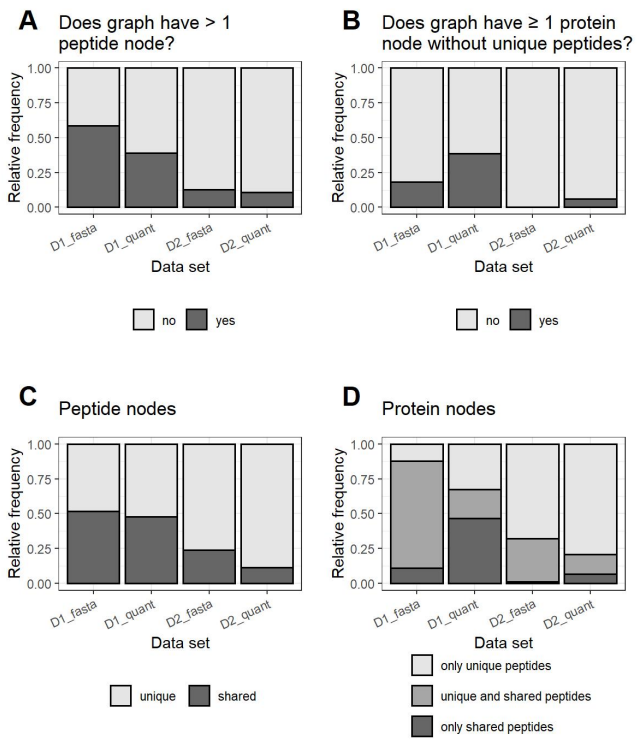

Fig. 3. A: Relative frequency of graphs with one or with more than one peptide node (most simple graph type vs. all other). B: Relative frequency of graphs with none vs. one or more protein nodes, that are not connected to any unique peptide node. C: Relative frequency of unique vs. shared peptide nodes. D: Relative frequency of protein nodes with only unique peptides vs. unique and shared peptides vs. only shared peptides.

with the theoretical peptides, leaving 22,940 sequences. In D2_quant none of the 8,078 peptides needed to be removed, as they were all in the range between six and 50 amino acids.

3.2 Occurring types of graphs

In the following, common isomorphism classes occuring from theoretical or quantified peptides of D1 and D2 are described. For D1_fasta, 15,402 bipartite graphs are constructed which fall into 2,583 different isomorphism classes. The graphs of the ten largest isomorphism classes are shown in Fig. S3. The largest class (6,355 occurrences) consists of the the smallest possible graph with only one protein and one peptide node. About 41% of all graphs are of this type and containing 12.1% of all protein nodes. The second largest class (almost 3,000 occurrences, i.e. around 19% of the graphs) contains graphs with two protein nodes and three peptide nodes. In these M-shapes graphs there are two peptide nodes that are unique for one of the protein groups each, and a peptide node that is shared by the two protein nodes. As each protein group has at least one unique peptide, both should be reported during protein inference. For quantification, however, it is not clear how to handle shared peptides. They may be omitted (as done in some currently used quantification algorithms), but this would mean to delete valuable information.

The third, fourth, fifth and tenth largest isomorphism classes have three protein nodes that are each connected to a unique peptide node. The graphs differ in number and connections to shared peptide nodes. From an inference and quantification point of view, they are similar to the second largest isomorphism class (M-shaped). The same holds for the eighth most common isomorphism class, which contains four proteins, each connected to a unique peptide node. The sixth, seventh and ninth largest isomorphism classes have in common, that one of the protein nodes is not connected to a unique peptide node. This means that from the identified peptides it is not clear if this protein is in the sample or not. However, the quantities of the shared peptides in combination with the unique peptides can help to estimate a quantity for this protein node. E.g., in case of the seventh largest isomorphism class with only two protein nodes and one unique and one shared peptide node (shaped like the letter N), this is possible. The ratio for protein A cannot be given exactly, but a lower or upper limit can be calculated, depending on the values of the two peptide ratios. This opportunity is not exploited by most currently applied protein quantification methods. Of the remaining isomorphism classes, 2,307 only occur once in D1_fasta. These are often relatively large systems that are uniquely found in the data due to their complex combination of protein and peptide nodes and the corresponding edges. However, these graphs contain almost 26,700 protein nodes which corresponds to almost 51% of all protein nodes.

The data set D1_quant contains five states of different combinations of concentrations of 13 spike-in proteins, leading to ten possible pairwise comparisons. Over the ten comparisons 28,802 graphs are formed in total, which belong to 354 different isomorphism classes. The largest isomorphism class (Fig. S4) contains again the most simple type with only one protein and peptide node each. The proportion of this graph type increases to almost 61% of all graphs when compared to the database level, containing 32.5% of the protein nodes. In comparison to the database level, the N-shaped graph type occurs more often and is the second most common graph type with almost 19% of all graphs. In contrast, the M-shaped graph that was the second most common on the database level is now on position eight with only 0.62%. From the top ten graph types, eight contain at least one protein node without a unique peptide node. The fifth and tenth largest isomorphism classes do not contain any unique peptide node.

For D2_fasta, 4,908 graphs are formed that fall into 66 different isomorphism classes. Again, the most common graph type is the most simple one (Fig. S5). It is more common than in D1_fasta with more than 87% of all subgraphs that cover in total 68% of all protein nodes. This can be explained by the lower complexity of the analyzed yeast compared to the mouse proteome, which leads to fewer connections between proteins via shared peptides. The second largest class also is the same as for in D1_fasta, with 8.72% of all graphs. The remaining classes are each relatively seldom but nevertheless contain in total more than 18% of all protein nodes. In almost all of the top 10 graph types all proteins nodes are connected to a unique peptide node, with the exception of the N-shapes graph type on place seven.

On quantitative level (D2_quant, Fig. S6), the percentage of the most simple graph type only increases slightly compared to the database level. However, the N-shaped graph type is now the second most common, like for D1_quant, with almost 5%. The M-shaped graph type moves from place 2 to 3 and still covers 4.5% of the graph types. All except three of the graph types in the top 10 now contain proteins without any unique peptide, which was not seen on database level.

3.3 Aggregrated characteristics of graphs

To characterize the graphs that occur in different situations, the distribution of numbers of different node types is shown in Fig. 2. The two main node types are protein and peptide nodes, which can be further divided into unique and shared peptide nodes. The most simple graph type is the only one with just one protein node. This is the most frequent class in all four situations (D1 and D2, on database and quantitative level), and therefore the corresponding bar is the highest. Going from the database to the quantitative level, the proportion of this simple graph type increases due to the disappearance of shared peptides that do not fulfill the criteria (see Section 2.2). This increase is less pronounced for data set D2, since

the proportion of the simple graph type is already high on database level. In all four cases, the frequency of graph types with higher numbers of protein nodes decreases steadily. This decrease is stronger for D2 and especially for the quantitative level, where the distribution is shifted to smaller numbers, as graphs split up due to disappearing shared peptides.

For peptide nodes, the first bar again belongs to the most simple graph type with only one peptide node. On database level, graph types with two peptide nodes (N-shaped, like 1B) are very uncommon and much less present than those with three peptide nodes. This graph type forms the second largest class for both D1_quant and D2_quant, but almost never appears on the database level. The M-shaped graph type (see Fig. 1A) is the second most common for D1_fasta and D2_fasta, which causes a higher bar for a peptide number of three. This causes "gaps" for even numbers in the peptide node distributions for theoretical peptides, while these are not present for quantified peptides. A lot of M-shaped graph types fall apart from database to quantitative level to the most simple or N-shaped graphs or vanish completely. In general, for the database level, graph types with even numbers of peptide nodes seem to be less common than the following odd numbers. For unique or shared peptides these gaps are not existing. In this case also the number zero is possible for the smallest graph type. Zero unique peptides occur rarely, but a little more often on the quantitative than on the database level. These graphs are complex during protein quantification, as there is no unique peptide as an "anchor".

In Fig. 3A, the proportions of graphs with exactly one (smallest possible graph, which can easily be handled during inference and quantification) compared to more than one peptide node (all other graphs) is shown. Graph types with more than one peptide node contain at least one shared peptide node and therefore call out for a strategy that uses the potential of these. In general, the proportion of the smallest graph type increases when going from database to quantitative level. This makes sense, because larger graphs fall apart into smaller ones when non-quantified peptides are missing. For D2, the difference is rather small, as the proportion of the smallest graphs is already high on database level (87.1% vs. 89.4%). For D1, this proportion increases from 41.3% to 60.9%, leaving a large number of systems (with an even larger number of affected protein nodes) that are influenced by shared peptides. Protein nodes without a unique peptide node are the most difficult to quantify and are ignored completely or subsumed into protein subgroups by many quantification methods. Fig. 3B shows the proportion of graphs that contain at least one of these protein nodes. It increases on the quantitative level in comparison to the database level, possibly explained by the disappearance of non-quantified unique peptides. For D2, the increase is on a lower level (from 0.3% to 5.8%) compared to D1 (18.1% to 38.4%). For D1, the proportion of graphs that contain at least one protein node without unique peptides is high. Fig. 3C and D show the distribution of uniqueness for peptide and protein nodes. The proportion of shared peptides decreases a bit from database to quantitative level and is lower in D2 than in D1. While the proportion of protein nodes with unique and shared peptides both decreases from database to quantitative level, the proportion of nodes with only one of the two peptide node types increases. Especially for D1, the frequency of protein nodes with only shared peptides strongly increases. These proteins are hard to quantify, as they are missing unique peptides and the shared peptides are also influenced by other proteins. On the quantitative level, 46.6% of the protein nodes fall into this category for D1, while this proportion is only 6.4% for D2.

4 Discussion and Conclusion

In this paper, we analyzed bipartite graphs of protein-peptide relationships using two different quantitative data sets as well as the corresponding protein databases. We did this with the aim to better understand the relationship between proteins and peptides and the implications for protein inference and quantification. We saw that the minimum peptide length is a crucial setting for the *in silico* digestion of proteins and has a huge impact on the structures of the bipartite graphs. Allowing too small peptides leads to large graphs that cover a majority of the proteins inside a database which is not informative. It also does not reflect the situation for real measurements, as small peptides are rarely found or even searched for in database searches. The minimal peptide length that fits best to the data also varies by database and organism. Here, we chose the smallest peptide length with leads to a comparable size if the largest and second largest graph (seven for D1 and six for D2) for further investigation. Smaller peptides were not or very rarely quantified in the data sets and lead to a largest graph that covers the majority of protein nodes.

In all four different settings, the most simple graph type with only one peptide and protein node is the most frequent one. These systems are an easy case for protein inference or quantification. As there are unique peptides, there is strong evidence for the presence of the corresponding protein (or at least one of the proteins within the node). A simple strategy for quantification, like averaging the peptide quantities, may work well in these cases. However, one has to keep in mind that no matter which quantification method is used, this will only lead to a common quantity for the members of a protein node if it consists of multiple protein accessions.

The composition of graph types differs between database to quantitative level as well as between the two data sets. Larger bipartite graphs split up into smaller ones when going from the database to the quantitative level, while at the same time graph types containing proteins without any unique peptide become more prominent. Aggregated characteristics over the whole set of bipartite graphs confirm these findings. There are two opposing effects when comparing database level and quantitative level. First, the proportion of the smallest possible graph (situation with simple solution for quantification) becomes larger. Second, the proportion of proteins without unique peptides was considerably higher on quantitative than on the database level. For these proteins, inference and quantification becomes more challenging. This indicates the importance of considering the quantified peptides instead of assessing the difficulty of quantification only based on the corresponding databases. The reason for the huge differences between database and quantitative level is that many theoretical peptides remain unquantified. These missing peptides cause the graphs to fall apart if these are shared peptides. Missing unique peptides cause some protein nodes to completely lose their unique peptides. These protein nodes may collapse with another node with the exactly same set of quantified peptides leading to a smaller graph. In other cases, they remain in a protein node without unique peptides, which is a challenge for protein quantification algorithms.

There are large differences in the characteristics of the bipartite graphs between the two different data sets for theoretical as well as quantified peptides. For example, D2 includes a higher proportion of unique peptide nodes and a more frequent occurrence of the most simple graph type. This makes the overall protein inference and quantification considerably easier than for D1. A possible explanation could be the different complexities of the underlying organisms, mouse (D1) and yeast (D2). The yeast genome is clearly smaller than the mouse genome (around 12[4] vs. over 2,600 Mb[5]) and the same holds for the proteome (around 6,000 vs. over 50,000 protein entries without isoforms in the UniProt reference proteomes, respectively). Already on the database level, the percentage of unique peptide nodes and of protein nodes with only unique peptides for D2 is much higher than for D1. This indicates that the complexity of protein quantification can vary a lot for different organisms and data sets, which may also have an impact

[4] https://www.ncbi.nlm.nih.gov/genome/?term=Saccharomyces+cerevisiae

[5] https://www.ncbi.nlm.nih.gov/genome/?term=mus+musculus

on the choice of a suitable protein quantification method. For D2 a simple quantification method, like averaging the unique peptides, might work well for most proteins. For D1, however, this would leave many proteins unquantifiable (46.6%) because of the lack of unique peptides and would also ignore a lot of information from the large number of shared peptides (47.9%) (see 3). In this case a more appropriate method that takes into account the shared peptides and the graph structures is needed.

In general, looking at bipartite peptide-protein graphs and aggregated characteristics is helpful to judge how complex the protein inference and quantification will be for a particular data set. Independent from the peptide quantification method applied, these characteristics can easily be generated and interpreted. In the future, different MS techniques might turn out to generate graphs more close to the database level than others, which may have a huge impact on the quality and difficulty of protein quantification. One has to be careful though, as different specifications (e.g., enzyme for digestion, mass spectrometer type, fragmentation and ionization techniques) may have an impact on the resulting graphs on the quantitative level, although this was not investigated here. Additionally, on the database level the complexity may appear higher or lower dependent on small variations of the settings (e.g. minimal peptide length, allowed missed cleavages, consideration of isoforms). This again highlights the need to generate the bipartite graphs also for only the quantified peptides and not only on all possible peptides from the database and to compare these two scenarios. As a consequence, there is an urgent need for protein quantification methods that take into account the graph structures and that can use this information to reliably quantify proteins without unique peptides.

Data Availability

The data underlying this article are available in the ProteomeXChange Consortium at http://proteomecentral.proteomexchange.org via the PRIDE partner repository, and can be accessed with the dataset identifiers PXD024684 and PXD024305.

Funding

This work was supported by the German Network for Bioinformatics Infrastructure (de.NBI), a project of the German Federal Ministry of Education and Research (BMBF) [FKZ 031 A 534A to K.S. and M.T.]. The funding of M.E. relates to PURE and VALIBIO, projects of Northrhine-Westphalia. J.U. is funded by the research building Center for Protein Diagnostics (PRODI), funded by North Rhine-Westphalia state and German Federal funds.

Acknowledgements

The authors thank Anika Frericks-Zipper for help with the PRIDE upload.

References

Ahrné, E., Molzahn, L., Glatter, T., and Schmidt, A. (2013). Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics*, **13**(17), 2567–2578.

Audain, E., Uszkoreit, J., Sachsenberg, T., Pfeuffer, J., Liang, X., Hermjakob, H., Sanchez, A., Eisenacher, M., Reinert, K., Tabb, D. L., Kohlbacher, O., and Perez-Riverol, Y. (2017). In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *J. Proteomics*, **150**, 170–182.

Bamberger, C., Martínez-Bartolomé, S., Montgomery, M., Pankow, S., Hulleman, J. D., Kelly, J. W., and Yates III, J. R. (2018). Deducing the presence of proteins and proteoforms in quantitative proteomics. *Nat. Commun.*, **9**, 2320.

Barkovits, K., Pacharra, S., Pfeiffer, K., Steinbach, S., Eisenacher, M., Marcus, K., and Uszkoreit, J. (2020). Reproducibility, specificity and accuracy of relative quantification using spectral librarybased data-independent acquisition. *Mol. Cell. Proteomics*, **19**(1), 181–197.

Blein-Nicolas, M. and Zivy, M. (2016). Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics. *Biochim. Biophys. Acta*, **1864**(8), 883–895.

Blein-Nicolas, M., Xu, H., de Vienne, D., Giraud, C., Huet, S., and Zivy, M. (2012). Including shared peptides for estimating protein abundances: A significant improvement for quantitative proteomics. *Proteomics*, **12**(18), 2797–2801.

Cox, J. and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**(12), 1367–1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011). Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.*, **10**(4), 1794–1805.

Craig, R. and Beavis, R. C. (2004). TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics*, **20**(9), 1466–1467.

Dost, B., Bandeira, N., Li, X., Shen, Z., Briggs, S. P., and Bafna, V. (2012). Accurate mass spectrometry based protein quantification via shared peptides. *J. Comput. Biol.*, **19**(4), 337–348.

Dupree, E. J., Jayathirtha, M., Yorkey, H., Mihasan, M., Petre, B. A., and Darie, C. C. (2020). A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field. *Proteomes*, **8**(3), 14.

Gerster, S., Qeli, E., Ahrens, C. H., and Bühlmann, P. (2010). Protein and gene model inference based on statistical modeling in k-partite graphs. *Proc. Natl. Acad. Sci. U. S. A.*, **107**(27), 12101–12106.

Gillet, L. C., Leitner, A., and Aebersold, R. (2016). Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annu. Rev. Anal. Chem.*, **9**(1), 449–472.

Huang, T., Wang, J., Yu, W., and He, Z. (2012). Protein inference: A review. *Brief. Bioinform.*, **13**(5), 586–614.

Junttila, T. and Kaski, P. (2007). Engineering an Efficient Canonical Labeling Tool for Large and Sparse Graphs. In *Proc. Meet. Algorithm Eng. Exp.*, pages 135–149.

Kim, S., Mischerikow, N., Bandeira, N., Navarro, J. D., Wich, L., Mohammed, S., Heck, A. J., and Pevzner, P. A. (2010). The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: Applications to database search. *Mol. Cell. Proteomics*, **9**(12), 2840–2852.

Pavlopoulos, G. A., Kontou, P. I., Pavlopoulou, A., Bouyioukos, C., Markou, E., and Bagos, P. G. (2018). Bipartite graphs in systems biology and medicine: A survey of methods and applications. *Gigascience*, **7**(4), 1–31.

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Pérez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yilmaz, S., Tiwary, S., Cox, J., Audain, E., Walzer, M., Jarnuczak, A. F., Ternent, T., Brazma, A., and Vizcaíno, J. A. (2019). The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res.*, **47**(D1), D442–D450.

Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**(18), 3551–3567.

Ramus, C., Hovasse, A., Marcellin, M., Hesse, A. M., Mouton-Barbosa, E., Bouyssié, D., Vaca, S., Carapito, C., Chaoui, K., Bruley, C., Garin, J., Cianférani, S., Ferro, M., Dorssaeler, A. V., Burlet-Schiltz, O., Schaeffer, C., Couté, Y., and Gonzalez de Peredo, A. (2016). Spiked proteomic standard dataset for testing label-free quantitative software and statistical methods. *Data Brief*, **6**, 286–294.

Swaney, D. L., Wenger, C. D., and Coon, J. J. (2010). Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J. Proteome Res.*, **9**(3), 1323–1329.

Uszkoreit, J., Maerkens, A., Perez-Riverol, Y., Meyer, H. E., Marcus, K., Stephan, C., Kohlbacher, O., and Eisenacher, M. (2015). PIA: An Intuitive Protein Inference Engine with a Web-Based User Interface. *J. Proteome Res.*, **14**(7), 2988–2997.

Uszkoreit, J., Perez-Riverol, Y., Eggers, B., Marcus, K., and Eisenacher, M. (2019). Protein Inference Using PIA Workflows and PSI Standard File Formats. *J. Proteome Res.*, **18**(2), 741–747.

Zhang, B., Chambers, M. C., and Tabb, D. L. (2007). Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.*, **6**(9), 3549–3557.

Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C., and Yates, J. R. (2013). Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.*, **113**(4), 2343–2394.

Zhang, Y., Wen, Z., Washburn, M. P., and Florens, L. (2015). Improving label-free quantitative proteomics strategies by distributing shared peptides and stabilizing variance. *Anal. Chem.*, **87**(9), 4749–4756.