Journal of proteome research

Article

pubs.acs.org/jpr

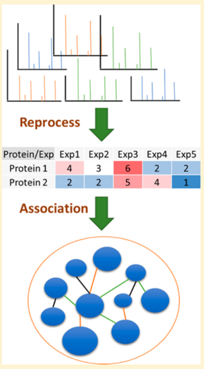
Unbiased Protein Association Study on the Public Human Proteome Reveals Biological Connections between Co-Occurring Protein Pairs

Surya Gupta,[†,‡,§] Kenneth Verheggen,[†,‡,§] Jan Tavernier,[†,‡] and Lennart Martens*[,†,‡,§]

[†]VIB-UGent Center for Medical Biotechnology, VIB, A. Baertsoenkaai 3, B-9000 Ghent, Belgium
[‡]Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium
[§]Bioinformatics Institute Ghent, Ghent University, B-9000 Ghent, Belgium

Supporting Information

ABSTRACT: Mass-spectrometry-based, high-throughput proteomics experiments produce large amounts of data. While typically acquired to answer specific biological questions, these data can also be reused in orthogonal ways to reveal new biological knowledge. We here present a novel method for such orthogonal data reuse of public proteomics data. Our method elucidates biological relationships between proteins based on the co-occurrence of these proteins across human experiments in the PRIDE database. The majority of the significantly co-occurring protein pairs that were detected by our method have been successfully mapped to existing biological knowledge. The validity of our novel method is substantiated by the extremely few pairs that can be mapped to existing knowledge based on random associations between the same set of proteins. Moreover, using literature searches and the STRING database, we were able to derive meaningful biological associations for unannotated protein pairs that were detected using our method, further illustrating that as-yet unknown associations present highly interesting targets for follow-up analysis.



Protein/Exp	Exp1	Exp2	Exp3	Exp4	Exp5
Protein 1	4	3	6	2	2
Protein 2	2	2	5	4	1

KEYWORDS: *mass spectrometry, protein co-occurrence, pathways, computational analysis, proteomics, protein−protein interaction, protein complex*

1. INTRODUCTION

Proteins associate with other proteins, nucleic acids, lipids, or metabolites to regulate the cellular and molecular mechanisms of the cell.[1] These associations can be of various types, including interactions, complex formation, or different roles in a single pathway.[2] A deeper understanding of the various roles and functions of proteins therefore requires the study of their relations to other proteins or molecules. As a result, a multitude of in vivo and in vitro experiments have been designed to determine protein associations, with typically only partial overlap in the results.[3] However, thanks to efforts toward greater transparency and data sharing in science,[4] it is now possible to reuse and reprocess publicly available proteomics data with computational approaches to obtain new knowledge in silico.[5] One such type of reuse has focused on the comparison of entire experiments, as pioneered for proteomics data by Klie et al.[6] and lately taken to an unprecedented level of sophistication by the online OmicsDI tool[7] to discover similar data sets across different omics domains, but so far no studies have been done to analyze the relations between proteins that are identified across many independent shotgun proteomics data sets.

One way to study such a relation between two entities is to study their co-occurrence across various observations. This phenomenon of co-occurrence has already been explored in many fields to determine the significance between such co-occurring entities; for instance, domain−domain co-occurrence has been used to determine the function of proteins,[8] while short polypeptide co-occurrence is used to predict global protein interactions.[9,10] Similarly, this concept of co-occurrence could also be applied to study biological association in proteins that co-occur across many different mass-spectrometry (MS)-based proteomics experiments. One of the examples of the use of co-occurrence in MS data is the detection of direct protein−protein interactions through special-purpose affinity purification-mass spectrometry (AP−MS) experiments,[11] yet, so far, the use of protein co-occurrence across heterogeneous public MS data sets to detect relevant biological protein associations (e.g., protein−protein interaction, complex formation, or co-occurrence in a given pathway) has not been demonstrated. For such co-occurrence approaches one needs a large number of different data sets to ensure the significance of the hypothesized relationships. Interestingly, there is a large number of public proteomics data sets available in the PRIDE[12] database.

On the basis of this large number of independent data sets in PRIDE, we applied this concept of co-occurrence detection to protein pairs across MS experiments. Ultimately, our method is based on the assumption that consistent protein co-occurrence across a large collection of different, unrelated MS experiments is not merely due to chance, but instead this reveals an underlying biological association between these co-occurring proteins.

Received: December 24, 2016
Published: May 8, 2017

DOI: 10.1021/acs.jproteome.6b01066

To verify the fundamental validity of our approach, we mapped protein pairs with a high degree of association to existing biological knowledgebases. The majority of these pairs were found to indeed be biologically associated. Furthermore, when we compared the level of biological association for the original results with that of protein pairs from randomized associations, we found an extremely significant drop in matches with existing knowledge, lending further support to the validity of our approach.

2. METHODOLOGY

The human MS data from the PRIDE database (downloaded in May 2015) was used in this study and analyzed in three major steps (Figure 1a): (i) reprocess step, (ii) associations step, and (iii) annotations step. For the reprocess step, the downloaded mass spectra from PRIDE were matched to the human proteome using an automated pipeline to obtain uniformly validated peptide and protein identifications (Figure 1b). In the subsequent associations step, these identified proteins were analyzed for co-occurrence across PRIDE experiments using the Jaccard similarity coefficient (Figure 1c). In the final annotations step, all protein pairs with a similarity score above threshold were mapped to existing knowledgebases to retrieve any known biological association (Figure 1d). These three steps are explained in detail in Sections 2.1, 2.2, and 2.3, respectively.

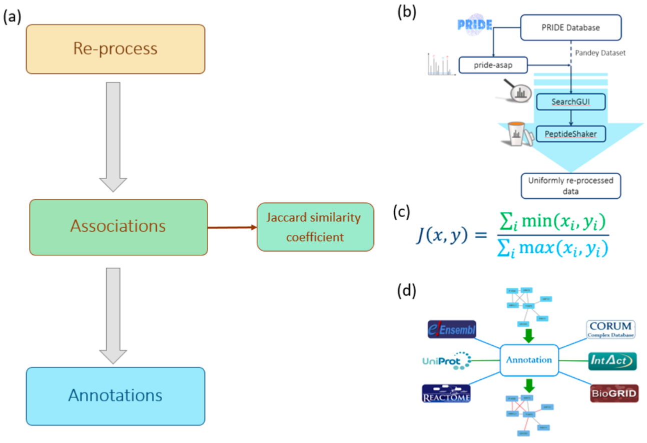


Figure 1. (a) Outline of the workflow to calculate and annotate protein pairs generated from MS-based proteomics experiments. (b) Identification is performed using a pipeline built from three existing tools, pride-asap, SearchGUI and PeptideShaker, all automated on the Pladipus backend. (c) Identified proteins are then analyzed for co-occurrence using the Jaccard similarity coefficient. (d) Protein pairs with a similarity coefficient above threshold are then mapped to existing knowledgebases to validate our findings.

Moreover, we validated our method by applying this pipeline to 1000 iterations of an equivalent number of randomly assigned protein pairs, as explained in Section 2.4. All relevant Python scripts are available on GitHub (https://github.com/compomics/ProteinAssociationPair).

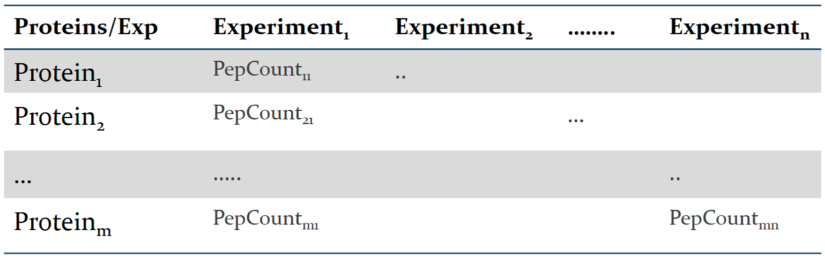
2.1. Reprocessing of Public Proteomics Data

All projects annotated to be of human origin were retrieved from the PRIDE database (release May_2015) using the PRIDE web service.[13] The retrieved projects were filtered for complete projects, which contain both identifications and fragmentation mass spectra. For future reference, these retrieved projects will be referred to as the *pride-data set*. In addition to the pride-data set, we also used the data from the draft human proteome by Kim et al., which contains more than 1000 experiments from a variety of human tissue samples.[14] For future reference, the Kim et al. project will be referred to as the *Pandey-data set*. Unlike the pride-data set, which contains both spectra and identifications for each experimental data set, the Pandey-data set contained only fragmentation mass spectra without identifications.

The experiment-specific identifications present in the pride-data set are used for automatic preprocessing by pride-asap[15] (Figure 1b), which infers optimal search parameters for the downstream analysis of each individual experiment. These inferred parameters include the precursor ion and fragment ion mass tolerances, the choice of digestion enzyme, and the most relevant variable and fixed modifications. However, because the Pandey-data set does not come with identifications, search parameter settings were applied as defined in the original Kim et al. publication.[14]

The obtained search parameter settings were used for the reprocessing of all experimental data using a pipeline composed of SearchGUI[16] and PeptideShaker[17] built on the Pladipus[18] platform (Figure 1b). The sequence database searches were conducted using SearchGUI with three different search engines: MS-GF+,[19,20] MyriMatch,[21] and X!tandem,[22] and matches were made against the human proteome complement of the UniProt[23] Swiss-Prot, which comprises only canonical sequences (release May_2015, with 20 887 protein sequence). The sequence database was expanded with all commonly encountered contaminant proteins listed in the common Repository of Adventitious Proteins (cRAP) from The Global Proteome Machine (GPM) database.[24] This expanded database was then automatically extended with its reversed decoy sequences by SearchGUI.

The results of these searches were processed by PeptideShaker to integrate the output of the different search engines and to control the local false discovery rate (FDR) for the



Proteins/Exp	$Experiment_1$	$Experiment_2$	....	$Experiment_n$
$Protein_1$	$PepCount_{11}$	..		
$Protein_2$	$PepCount_{21}$		...	
...	.....			..
$Protein_m$	$PepCount_{m1}$			$PepCount_{mn}$

Figure 2. Protein−experiment matrix obtained after reprocessing human PRIDE data. Columns represent experiments and rows proteins, with values representing distinct peptide counts for a protein in an experiment.

$$J(x,y) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}$$

Protein/Exp	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6
Protein 1	4	3	6	2	2	3
Protein 2	2	2	5	4	1	4

$$J(Protein1, Protein2) = \frac{2+2+5+2+1+3}{4+3+6+4+2+4} = \frac{15}{23} = 0.65$$

Figure 3. Example of Jaccard similarity coefficient calculation, where similarity between Protein 1 and Protein 2 is calculated by using peptide counts for each across six different experiments.

integrated results at the PSM level. Moreover, PeptideShaker also performed the protein inference.

Proteins that were not identified as human proteins or that were not obtained from UniprotKB/Swiss-Prot were filtered out. The retained results were then collated in a protein−experiment matrix with experiments as columns, proteins as rows and distinct peptide counts for that protein in that experiment as values (Figure 2).

2.2. Detecting Associations between Proteins

The peptide counts in the protein−experiment matrix were used to determine the co-occurrence between two proteins. These co-occurrence values were calculated using the Jaccard similarity coefficient (Figure 1c), where $J(x,y)$ represents the similarity between two proteins x and y, where x_i and y_i represent the number of distinct and nonshared peptide count between protein x and y in the ith experiment. Figure 3 shows an example of a Jaccard similarity calculation between two proteins (Protein 1 and Protein 2), with the distinct and nonshared peptide count listed in the Table for a given protein in a given experiment. The Jaccard similarity coefficient (0.65) is the ratio of the total minimum (marked in blue) and maximum (marked in green) peptide count in each experiment. The number of distinct, nonshared peptides is calculated as the total number of distinct peptide sequences of a protein in an experiment minus the peptides that are shared by the paired protein in that same experiment. Furthermore, the similarity is only calculated for protein pairs found together in at least 10 different experiments (SI Section S1).

2.3. Annotation with Existing Knowledge of Proposed Protein Pairs

Protein pairs with a Jaccard similarity coefficient of 0.4 and above were mapped to several publicly available knowledgebases to find known biological associations. The selection for the Jaccard similarity coefficient of 0.4 is explained in SI Section S1. The Reactome[25,26] database (V56) was used to determine the presence of a protein pair in the same pathway. Because Reactome structures pathways in a hierarchical way, we use the lowest level of pathways (called leaf pathways) that only contain a series of reactions and that do not divide further into subpathways. The IntAct[27] (release 2016_01) and BioGRID[28] (version 3.4.145) databases were used to find known binary protein−protein interactions between proteins in a pair, and the COmprehensive ResoUrce of Mammalian protein complexes (CORUM)[29,30] database (release 2012_02) was used to detect whether the proteins in a pair were both part of a known protein complex. The Ensembl[31] database (Ensembl 83 version) was used to detect paralog proteins in a protein pair.

Furthermore, cRAP from GPM was used to label human-derived common contaminant proteins in protein pairs. For the remaining unannotated pairs, the Gene Ontology (GO) annotation from UniProt was used to label pairs in which the proteins share same GO biological Process or GO molecular Function.

For a select number of protein pairs from each data set that did not yield a match against any of the above knowledgebases, a detailed manual investigation was conducted using literature search and the STRING[32] database (version 10.0) to obtain a possible explanation for the suggested biological relation.

2.4. Validation of the Approach

To validate our approach, we compared the amount of known biological associations for proteins found in pairs according to our Jaccard similarity threshold of 0.4, with the amount of known biological associations for randomly assigned protein pairs. This comparison was done in three steps. First, proteins in the original protein−experiment matrix were randomly assigned to pairs, that is, without actually calculating a Jaccard coefficient. Then, out of all of these random pairs, we randomly selected an equal number of protein pairs as were originally found with a Jaccard coefficient of 0.4 or above. These randomly selected pairs were then mapped to five knowledgebases: Reactome, IntAct, BioGRID, CORUM, and Ensembl. These three steps were repeated 1000 times, and the number of annotated pairs was calculated each time.

3. RESULTS

We present the results generated by our method for the two data sets (pride-data set and Pandey-data set) as well as the results of an investigation in more detail of five unannotated protein pairs, two from each data set.

3.1. Results from Reprocessing, Association, and Annotation of Pairs

After reprocessing of the pride-data set, we had analyzed 1063 experiments that yielded 12 085 identified proteins in total. Of these 12 085 proteins, 4562 proteins were contained in the human complement of UniProtKB/Swiss-Prot and were present in at least 10 experiments. Of these selected 4562 proteins, all possible protein pairs were scored using the Jaccard similarity coefficient. Of all scored pairs, 2325 protein pairs (see Table S1), comprising 749 unique proteins, passed the Jaccard coefficient threshold of 0.4.

DOI: 10.1021/acs.jproteome.6b01066
J. Proteome Res. 2017, 16, 2204−2212

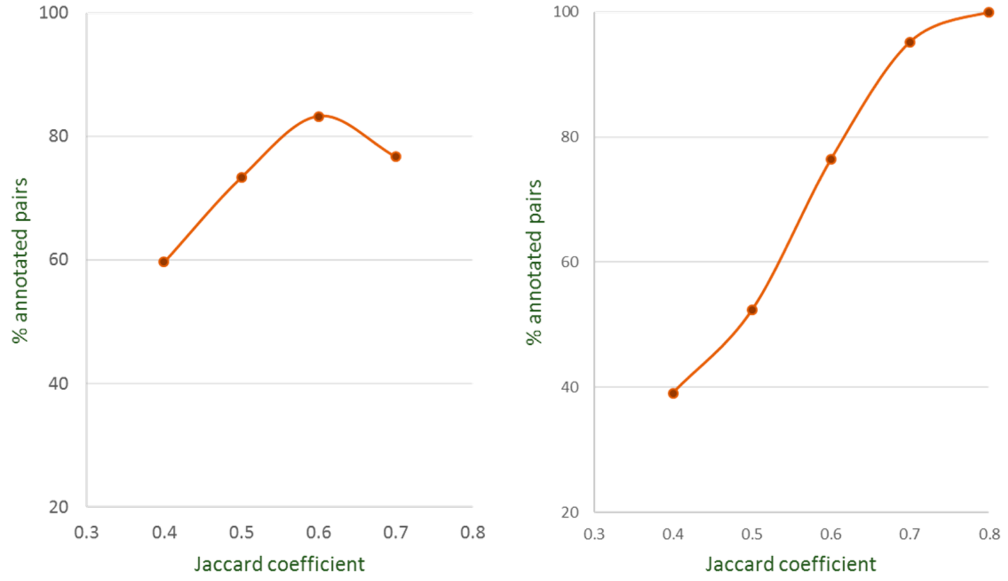


Figure 4. Distribution of percentage of annotated pairs versus similarity score for (a) pride-data set and (b) Pandey-data set.

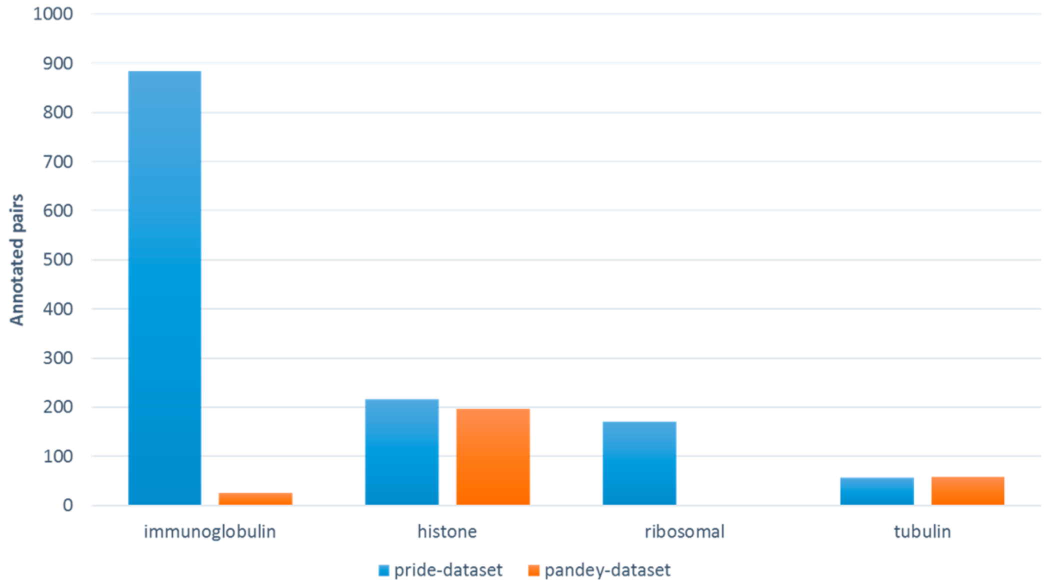


Figure 5. Abundance of four co-occurring proteins in annotated protein pairs for both data sets.

These 2325 selected pairs were then mapped to Reactome, CORUM, IntAct, BioGRID, and Ensembl to determine whether the proteins in each pair were already known to have biological associations. Of the 2325 pairs, 1645 (71%) had known biological associations in at least one of these databases, while 680 had no known associations (see Figure S1 for a network representation of the associated protein pairs). Of the 1645 verified protein pairs, 1527 (93%) pairs had both proteins mapped to the same pathway. These pairs mapped to a total of 275 of Reactome's 1990 leaf pathways (14% of all pathways) (see Table S5). These 275 pathways cover 21 out of the 24 top-level pathways in Reactome (SI Section S4). For 90 pairs (5%), either a direct protein−protein interaction was found in IntAct (25 pairs) or in BioGRID (58 pairs) or the proteins were known to be in a complex according to CORUM (7 pairs). The remaining 28 pairs (2%) were found to be paralogs of each other according to Ensembl.

To remove possible associations with commonly encountered contaminant proteins, we also filtered out pairs in which at least one protein was included in the human proteins included in the cRAP database. Only 56 such pairs were found across the 2325 pairs that passed the Jaccard threshold of 0.4, and 40 of these cRAP protein -containing pairs were found in the 680 unannotated pairs that passed the Jaccard coefficient of 0.4, leaving 640 unexplained, unannotated pairs (28%).

For the Pandey-data set from Kim et al., reprocessing of 1842 experiments resulted in 16 597 identified proteins. Of these 16 597 proteins, 10 121 proteins were found to be contained in the human complement of UniProtKB/Swiss-Prot and were present in at least 10 experiments. Out of all possible pairs scored using these 10 121 proteins, 988 protein pairs passed the Jaccard similarity threshold of 0.4, comprising 454 unique proteins (see Table S2).

Among these 988 protein pairs, 475 (48%) protein pairs were found to be biologically associated, while 513 had no known biological association (see Figure S2 for a network representation of the associated protein pair). Of the 475 protein pairs with known biological associations, 401 (40%) were found to share the same Reactome pathway. These 401 protein pairs were mapped to 165 of Reactome's leaf pathways (see Table S6) and in turn to 20 of the 24 top-level pathways (SI Section S5). 47 (5%) had a direct protein−protein interaction according to IntAct (14) or BioGRID (33), 10 (1%) were found in CORUM to be in the same protein complex, and 17 (2%) were paralogues of each other according to Ensembl. 43 (4%) protein pairs out of the 513 unannotated protein pairs were found to contain one or both proteins from the human proteins in the cRAP database. This left a total of 470 (47%) unexplained, unannotated protein pairs in the Pandey-data set.

3.2. Existing Biological Relations between Proteins in Annotated Pairs

The co-occurrence analysis of the pride-data set resulted in a total of 1685 annotated protein pairs and 640 unannotated protein pairs. Interestingly, the percentage of protein pairs for which a biological association is known, tends to rise with increased Jaccard similarity coefficient, except for the highest Jaccard coefficients, where the correlation paradoxically reverses again (Figure 4a). For unannotated protein pairs with Jaccard coefficients above 0.6, it was typically the case that at least one of the proteins in the pair did not occur at all in any of the knowledgebases.

For the Pandey-data set, a total of 518 protein pairs were found to be biologically associated, while 470 protein pairs did not have any known existing association. Unlike the pride-data set, the percentage of annotated protein pairs for this data set continues to rises with increasing Jaccard similarity coefficient (Figure 4b).

The annotated protein pairs in both data sets consist of various different proteins. Of these various proteins, in the pride-data set, immunoglobulins (Ig), histones, ribosomal proteins, and cytoskeleton proteins were found to pair with proteins of the same family, constituting the majority of annotated protein pairs. In contrast, in the Pandey-data set, only immunoglobulins (Ig), histones, and cytoskeleton proteins were found to co-occur with protein of the same family, and co-occurring ribosomal proteins were missing. Figure 5 shows the absolute abundance of the four proteins found in the annotated protein pairs, for both the pride-data set and the Pandey-data set.

3.3. Possible Biological Relation between Proteins in Unannotated Pairs

In MS experiments, keratin proteins are highly abundant compared with other proteins as they do not only come from the biological sample but also enter the sample as a common environmental contaminant from hair or skin.[33] Therefore, we expected to see keratin proteins co-occurring the high Jaccard coefficients. Interestingly, however, for the pride-data set, we found only 40 protein pairs in which both proteins are a keratin. Similarly, in the Pandey-data set, we also found only 29 co-occurring keratin protein pairs, but for the Pandey-data set, we also found 7 protein pairs where keratin proteins were found to co-occur with a nonkeratin protein, while in the pride-data set, we only ever found keratin proteins to co-occur together. As such, the majority of keratin proteins in both data sets are found to co-occur together, which suggests that our method is capable of retrieving protein pairs with high co-occurrence across different experiments, but keratins are only rarely (if at all) found to be associated with other proteins, which indicates the specificity of our method, even for such frequently occurring proteins.

In the remaining 600 (26%) unannotated, non-keratin-containing protein pairs from the pride-data set, 274 (12%) protein pairs were found to share either same GO biological process (144) or GO molecular function (130). Similarly, for the Pandey-data set, the 434 (44%) unannotated, nonkeratin containing protein pairs yielded 170 (17%) protein pairs with same GO biological process (62) or GO molecular function (108) annotation. Of the remaining 326 unannotated protein pairs in the pride-data set, 257 were found to consist of proteins that share the same GO cellular component. Similarly, for the remaining 264 unannotated protein pairs in the Pandey-data set, 207 protein pairs contained proteins that shared the same GO cellular component.

3.4. Case Study of Unannotated Pairs Based on Associations Inferred from the Literature or from the STRING Database

We selected two cases from each data set to establish the potential biological relation between unannotated protein pairs using a literature search and the STRING database. It should be noted that protein pairs that contained one or two immunoglobulin proteins were also excluded from this manual analysis.

3.4.1. Association between Kininogen-1, Vitamin D-Binding Protein, and Hemopexin. The first example from the pride-data set concerns three proteins that together make up three detected protein pairs. Vitamin-D-binding protein and hemopexin showed a Jaccard similarity of 0.58. In the study by Pawlik et al., it has been shown that vitamin-D-binding protein and hemopexin, together with two other proteins, show overexpression in breast tumor samples.[34] Interestingly, most of the PRIDE projects in which these proteins were detected as co-occurring concern cancer tissue samples, including breast cancer, which may indicate that the cancer-related association of these proteins can be generalized more broadly. However, Vitamin-D-binding protein and hemopexin also show strong association with Kininogen-1, with similarity coefficients of 0.54 and 0.45, respectively. Rithidech et al. found 12 proteins significantly up-regulated which include kininogen, vitamin-D-binding protein, and hemopexin along with nine other proteins in their study of pediatric multiple sclerosis.[35] This study shows that these three proteins have some form of biological relation to each other in this disease. According to STRING the putative homologues for kininogen-1 and hemopexin and for hemopexin and vitamin-D-binding proteins are found to be coexpressed in other species as well. A further disease-related association between kininogen-1 and Hemopexin can be found in Ghafouri et al., where it was shown that the levels of kininogen and hemopexin were both higher in the plasma of a farmer with systemic inflammation caused by musculoskeletal disorder (MSD) than in a reference farmer.[36]

3.4.2. Cystatin-SA and Carbonic Anhydrase 6. The second example from the pride-data set is the pair of cystatin-SA with carbonic anhydrase 6, with a Jaccard coefficient of 0.58. According to a study of diabetic patients by Bencharit et al., cystatin-SA and carbonic anhydrase 6 are both biomarkers for the disease.[37] According to STRING, these two proteins are involved in the same GO biological process "detection of chemical stimulus involved in sensory perception of bitter taste".

3.4.3. Talin-1 and Myosin-9.

The first example case from the Pandey-data set is talin-1 and myosin-9 with a Jaccard similarity coefficient of 0.53. In a study of the platelet proteomes of patients with myelodysplastic syndrome by Fröbel et al., talin-1 and myosin-9 were both found to have a reduced concentration, along with three other proteins.[38] The study furthermore shows that talin-1 and myosin-9, along with the three other regulated proteins, need to be expressed in platelets for adequate integrin $\alpha_{\mathrm{IIb}}\beta_3$ function and hemostasis. These proteins are also found to share the GO biological process "platelet aggregation".

3.4.4. Peripherin-2 and M-Opsin.

The second example from the Pandey-data set is the pair formed by Peripherin-2 and medium-wave-sensitive-opsin 1 (also called M-opsin[39]), with a Jaccard similarity coefficient of 0.49. According to a study from Nguyen et al., it is shown that peripherin-2 differentially interacts with M-opsin.[40]

3.5. Validation Results

For the validation of our approach, we compared the number of annotated protein pairs obtained from the original protein set with the number of annotated protein pairs that were obtained from randomly generated protein sets. The random pairs were mapped to the same five knowledgebases as the original pairs: Reactome, IntAct, CORUM, and Ensembl. This randomized validation was repeated 1000 times, and the number of annotated pairs obtained was calculated each time. Interestingly, in the pride-data set, when we mapped the 2325 randomly selected pairs to the five knowledgebases, we found only 40 annotated protein pairs (on average) and 69 annotated protein pairs (at maximum) over the 1000 iterations (see Table S3). This is in striking contrast with the 1645 annotated pairs we obtained for the original protein pairs. As a result, the difference between real and random data is extremely significant, with a p value of effectively 0. Similarly, for the Pandey-data set, the mapping of 988 randomly selected protein pairs to the five knowledgebases resulted in only 6 annotated protein pairs (on average) and 17 annotated protein pairs (at maximum) over the 1000 iterations (see Table S4). This again in striking contrast with the 475 annotated pairs for the original protein pairs. Here, too, the difference between real and random data is therefore extremely significant, again with a p value of effectively 0.

4. DISCUSSION

We have presented a novel method to determine biologically relevant protein associations between co-occurring proteins and applied it to two different types of MS-based proteomics data sets: the highly heterogeneous pride-data set and the draft human proteome of the Pandey-data set. For the pride-data set, 83% of protein pairs were mapped based on existing biological knowledge, 71% were mapped using Reactome, IntAct, BioGRID, CORUM, and Ensembl, while 12% were mapped using GO (biological process and molecular function) annotations. Similarly, for the Pandey-data set, 65% of protein pairs were mapped with existing biological knowledge, 48% were mapped using the five knowledgebases, while 17% were mapped using GO (biological process and molecular function) annotation. More proteins were identified in the Pandey-data set than in the pride-data set, but the number of pairs that passed the Jaccard coefficient threshold of 0.4 in the Pandey-data set was much lower than for the pride-data set. This is likely due to the fact that, unlike the pride-data set, the samples in the Pandey-data set were chosen to provide maximal complementarity toward the elucidation of the entire human proteome, thus resulting in a lower overall overlap in proteins between samples.

At the same time, the percentage of annotated pairs in the pride-data set is possibly also higher than the Pandey-data set because the pride-data set contains a majority of projects built around disease-related samples, and proteins involved in disease are typically much more studied than other proteins. This, in turn, increases the available level of annotation for these proteins. Conversely, the Pandey-data set with its focus on elucidating the complete human proteome will inevitably include many proteins that have not been studied in detail and that therefore lack knowledge in existing databases.

Moreover, the annotated protein pairs in both data sets are built from only a select number of individual proteins. As shown in Figure 5, co-occurring Immunoglobulin (Ig), tubulin, histones, and ribosomal proteins constitute the majority of annotated protein pairs in the pride-data set, while co-occurring tubulin and histones proteins form the majority in the Pandey-data set. However, while tubulin and histone proteins were found to be abundant in both the pride-data set as well as the Pandey-data set, only a few Ig proteins are found in the Pandey data set. This contrasts sharply with the pride-data set where Ig is involved in the vast majority of annotated pairs. Similarly, ribosomal proteins were only found in pairs in the pride-data set and are missing entirely from the Pandey-data set.

Histones and tubulins are housekeeping genes, meaning they are involved in basic cellular processes and are found to be present in almost all cells and tissues.[14] It is therefore logical to find these proteins as highly co-occuring in the two data sets. However, even though ribosomal proteins are housekeeping proteins as well,[14] we only found them as co-occurring in the pride-data set. A detailed analysis showed that we also find ribosomal protein in pairs in the Pandey-data set but with a Jaccard coefficient below the threshold of 0.4. A possible reason for this could be the difference in the number of peptide counts across experiments (SI Section S2).

The reason for the abundance of Ig-containing pairs in the pride-data set could be that the majority of projects in the pride-data set were focused on disease-related samples. Because Ig proteins serve a very important role in the immune response, this could indeed explain the high co-occurrence of Igs in pride-data set. The Pandey-data set, in contrast, focused on proteins acquired from normal tissue samples, which may well explain the much lower number of co-occurring Ig pairs.

The different amount of co-occurring proteins in both data sets therefore suggests that while housekeeping proteins can be found to co-occur in high abundance irrespective of the data set used, the co-occurrence of more specialized proteins will depend on the type of samples that were studied in the data set. Note that this also indicates that a direct comparison of the pairs obtained in one set of studies against those obtained in another set of studies is likely to not be biologically meaningful as any sample bias is necessarily carried forward in the analysis. The corollary is that it is likely to be especially rewarding to search for protein pairs in specific sample types of particular interest to the researcher (e.g., cancer-related samples) if such sufficient samples can be obtained from the public domain.

We also showed that the percentage of annotated protein pairs with strong biological association tends to rise with increasing the Jaccard coefficient (Figure 4). For the Pandey-data set, the percentage of annotated pairs increases continuously with increasing Jaccard coefficient (Figure 4b); however, for the pride-data set the curve displays an intriguing drop-off above a Jaccard coefficient of 0.6 (Figure 4a). This drop-off could, in part, be attributable to Ig proteins, as most of the Ig proteins are not represented in pathway or interaction databases. Indeed, as stated in the results, the absence of annotation is usually due to one or both proteins in a pair being absent altogether from the five studied knowledgebases. Nevertheless, a manual search of the literature did turn up possible relations between such unannotated Ig pairs in the pride-data set, as detailed in SI Section S3. We would therefore propose that the unannotated, high-scoring pairs in which one protein is as yet mostly unknown provide very interesting targets for follow-up studies to researchers who are interested in the functions and roles of the other protein in such a pair.

Importantly, we have validated our approach by contrasting its performance in obtaining annotated protein pairs with random protein associations. The extremely significant difference in annotated pairs between these two approaches strongly suggests that our selection for highly co-occurring protein pairs yields biological associations that are far from random.

Our method is thus capable of analyzing a large amount of (public) proteomics data sets to detect potential biological association between proteins and can as such function as a hypothesis generator for researchers who are interested in further investigating the roles and functions of (possibly as yet poorly annotated) proteins. We also show that a judicious selection of data sets and samples (e.g., around a given topic such as cancer) can be used to focus the search for biological association to a given disease or tissue.

At the same time, we are aware that the Jaccard similarity metric used here tends to err on the side of caution and therefore only reports protein pairs that show very high co-occurrence among various experiments. This most likely results in a substantial number of false-negatives, implying that protein pairs that have a small but significant co-occurrence are currently missed. A key aspect for potential improvement of our method is therefore the similarity metric, and we can envision that additional computational work can further improve the sensitivity of the overall approach without affecting its precision.

5. CONCLUSIONS

We here present an approach to reuse large amounts of publicly available data to determine possible protein associations using a simple but effective Jaccard similarity coefficient. We have shown that the majority of the protein pairs detected by our method are substantiated by biological annotations from at least one of five established knowledgebases. Furthermore, we have been able to manually discover biological rationales for the association between proteins in unannotated pairs through literature searches. We also showed that the associations we detect are highly significantly different from random protein associations. Interestingly, a substantial fraction of human protein pairs recovered with high Jaccard coefficients from the pride-data set contain at least one partner for which little to no biological knowledge is present in the existing databases. Such proteins, in particular, can be considered low-hanging fruit for targeted biological studies by interested researchers.

We thus show that the compendium of publicly available proteomics data can be considered as a proteome-wide association study and that we can extract various biologically meaningful protein associations from these data. Moreover, we believe that refinements to the association metric will be able to increase sensitivity, allowing even more information to be recovered from these data in the future, a promise that is complemented by the ever-increasing number of relevant data sets in the public domain.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.6b01066.

S1. Distribution of Jaccard coefficient for pride and Pandey-data set. S2. The abundance of ribosomal proteins in Pandey-data set. S3. Possible biological relation between Ig and co-occurring protein in pride-data set. S4. List of 21 Reactome's main pathway found to be mapped with protein pairs in pride-data set. S5. List of 20 Reactome's main pathway found to be mapped with protein pairs in Pandey-data set. (PDF)
Table S1. Protein pairs found for pride-data set. (XLSX)
Table S2. Protein pairs found for Pandey-data set. (XLSX)
Table S3. Randomize output for pride-data set. (XLSX)
Table S4. Randomize output for Pandey-data set. (XLSX)
Table S5. Leaf pathway found to be mapped in pride-data set. (XLSX)
Table S6. Leaf pathway found to be mapped in Pandey-data set. (XLSX)
Figure S1. Associated protein pairs for pride-data set. (PDF)
Figure S2. Associated protein pairs for Pandey-data set. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: lennart.martens@vib-ugent.be. Tel: +3292649358.

ORCID

Surya Gupta: 0000-0002-6290-6161
Lennart Martens: 0000-0003-4277-658X

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded in part by the Concerted Research Action BOF12/GOA/014 of Ghent University. J.T. holds an ERC Advanced Grant (no. 340941, CYRE). We thank Navratan Bagwan and Adriaan Sticker for useful discussion.

■ REFERENCES

(1) Cho, S.; Park, S. G.; Lee, D. H.; Park, B. C. Protein-protein interaction networks: from interactions to networks. *J. Biochem. Mol. Biol.* **2004**, *37* (1), 45−52.

(2) Chautard, E.; Thierry-Mieg, N.; Ricard-Blum, S. Interaction networks: From protein functions to drug discovery. A review. *Pathol. Biol.* **2009**, *57* (4), 324−333.

(3) Venkatesan, K.; Rual, J.-F.; Vazquez, A.; Stelzl, U.; Lemmens, I.; Hirozane-Kishikawa, T.; Hao, T.; Zenkner, M.; Xin, X.; Goh, K.-I.;

et al. An empirical framework for binary interactome mapping. *Nat. Methods* **2009**, *6* (1), 83−90.

(4) Vizcaíno, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ríos, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223−226.

(5) Vaudel, M.; Verheggen, K.; Csordas, A.; Raeder, H.; Berven, F. S.; Martens, L.; Vizcaíno, J. A.; Barsnes, H. Exploring the potential of public proteomics data. *Proteomics* **2016**, *16* (2), 214−225.

(6) Klie, S.; Martens, L.; Vizcaíno, J. A.; Côté, R.; Jones, P.; Apweiler, R.; Hinneburg, A.; Hermjakob, H. Analyzing Large-Scale Proteomics Projects with Latent Semantic Indexing. *J. Proteome Res.* **2008**, *7* (1), 182−191.

(7) Perez-Riverol, Y.; Bai, M.; da Veiga Leprevost, F.; Squizzato, S.; Park, Y. M.; Haug, O. K.; Carroll, A. J.; Spalding, D.; Paschall, J.; Wang, M.; et al. Discovering and linking public omics data sets using the Omics Discovery Index. *Nat. Biotechnol.* **2017**, *35*, 406.

(8) Cohen-Gihon, I.; Nussinov, R.; Sharan, R.; et al. Comprehensive analysis of co-occurring domain sets in yeast proteins. *BMC Genomics* **2007**, *8* (1), 161.

(9) Pitre, S.; Hooshyar, M.; Schoenrock, A.; Samanfar, B.; Jessulat, M.; Green, J. R.; Dehne, F.; Golshani, A.; et al. Short Co-occurring Polypeptide Regions Can Predict Global Protein Interaction Maps. *Sci. Rep.* **2012**, *2*, 4286−4294.

(10) Schoenrock, A.; Samanfar, B.; Pitre, S.; Hooshyar, M.; Jin, K.; Phillips, C. a; Wang, H.; Phanse, S.; Omidi, K.; Gui, Y.; et al. Efficient prediction of human protein-protein interactions at a global scale. *BMC Bioinf.* **2014**, *15* (1), 1−22.

(11) Titeca, K.; Meysman, P.; Gevaert, K.; Tavernier, J.; Laukens, K.; Martens, L.; Eyckerman, S. SFINX: Straightforward Filtering Index for Affinity Purification−Mass Spectrometry Data Analysis. *J. Proteome Res.* **2016**, *15* (1), 332−338.

(12) Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R. PRIDE: the proteomics identifications database. *Proteomics* **2005**, *5* (13), 3537−3545.

(13) Vizcaíno, J. A.; Csordas, A.; del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **2016**, *44* (D1), D447−56.

(14) Kim, M.-S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; et al. A draft map of the human proteome. *Nature* **2014**, *509* (7502), 575−581.

(15) Hulstaert, N.; Reisinger, F.; Rameseder, J.; Barsnes, H.; Vizcaíno, J. A.; Martens, L. Pride-asap: automatic fragment ion annotation of identified PRIDE spectra. *J. Proteomics* **2013**, *95* (100), 89−92.

(16) Vaudel, M.; Barsnes, H.; Berven, F. S.; Sickmann, A.; Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **2011**, *11* (5), 996−999.

(17) Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.; Martens, L.; Barsnes, H. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **2015**, *33* (1), 22−24.

(18) Verheggen, K.; Maddelein, D.; Hulstaert, N.; Martens, L.; Barsnes, H.; Vaudel, M. Pladipus Enables Universal Distributed Computing in Proteomics Bioinformatics. *J. Proteome Res.* **2016**, *15* (3), 707−712.

(19) Kim, S.; Gupta, N.; Pevzner, P. A. Spectral Probabilities and Generating Functions of Tandem Mass Spectra: A Strike against Decoy Databases. *J. Proteome Res.* **2008**, *7* (8), 3354−3363.

(20) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*, 5277.

(21) Tabb, D. L.; Fernando, C. G.; Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **2007**, *6* (2), 654−661.

(22) Fenyö, D.; Beavis, R. C. A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes. *Anal. Chem.* **2003**, *75* (4), 768−774.

(23) The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **2015**, *43* (D1), D204−D212.

(24) Craig, R.; Cortens, J. P.; Beavis, R. C. *Open Source System for Analyzing, Validating, and Storing Protein Identification Data* **2004**, *3*, 1234−1242.

(25) Croft, D.; Mundo, A. F.; Haw, R.; Milacic, M.; Weiser, J.; Wu, G.; Caudy, M.; Garapati, P.; Gillespie, M.; Kamdar, M. R.; et al. The Reactome pathway knowledgebase. *Nucleic Acids Res.* **2014**, *42*, D472−D477.

(26) Fabregat, A.; Sidiropoulos, K.; Garapati, P.; Gillespie, M.; Hausmann, K.; Haw, R.; Jassal, B.; Jupe, S.; Korninger, F.; McKay, S.; et al. The Reactome pathway Knowledgebase. *Nucleic Acids Res.* **2016**, *44* (D1), D481−7.

(27) Orchard, S.; Ammari, M.; Aranda, B.; Breuza, L.; Briganti, L.; Broackes-Carter, F.; Campbell, N. H.; Chavali, G.; Chen, C.; del-Toro, N.; et al. The MIntAct project−IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **2014**, *42* (Database issue), D358−D363.

(28) Chatr-Aryamontri, A.; Oughtred, R.; Boucher, L.; Rust, J.; Chang, C.; Kolas, N. K.; O'Donnell, L.; Oster, S.; Theesfeld, C.; Sellam, A.; et al. The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **2017**, *45* (D1), D369−D379.

(29) Ruepp, A.; Brauner, B.; Dunger-Kaltenbach, I.; Frishman, G.; Montrone, C.; Stransky, M.; Waegele, B.; Schmidt, T.; Doudieu, O. N.; Stümpflen, V.; et al. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.* **2007**, *36* (Database), D646−D650.

(30) Ruepp, A.; Waegele, B.; Lechner, M.; Brauner, B.; Dunger-Kaltenbach, I.; Fobo, G.; Frishman, G.; Montrone, C.; Mewes, H.-W. CORUM: the comprehensive resource of mammalian protein complexes−2009. *Nucleic Acids Res.* **2010**, *38* (suppl_1), D497−D501.

(31) Yates, A.; Akanni, W.; Amode, M. R.; Barrell, D.; Billis, K.; Carvalho-Silva, D.; Cummins, C.; Clapham, P.; Fitzgerald, S.; Gil, L.; et al. Ensembl 2016. *Nucleic Acids Res.* **2016**, *44* (D1), D710−D716.

(32) Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K. P.; et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **2015**, *43* (D1), D447−D452.

(33) Hodge, K.; Have, S. T.; Hutton, L.; Lamond, A. I. Cleaning up the masses: exclusion lists to reduce contamination with HPLC-MS/MS. *J. Proteomics* **2013**, *88*, 92−103.

(34) Pawlik, T. M.; Hawke, D. H.; Liu, Y.; Krishnamurthy, S.; Fritsche, H.; Hunt, K. K.; Kuerer, H. M. Proteomic analysis of nipple aspirate fluid from women with early-stage breast cancer using isotope-coded affinity tags and tandem mass spectrometry reveals differential expression of vitamin D binding protein. *BMC Cancer* **2006**, *6*, 68.

(35) Rithidech, K. N.; Honikel, L.; Milazzo, M.; Madigan, D.; Troxell, R.; Krupp, L. B. Protein expression profiles in pediatric multiple sclerosis: potential biomarkers. *Mult. Scler.* **2009**, *15* (4), 455−464.

(36) Ghafouri, B.; Carlsson, A.; Holmberg, S.; Thelin, A.; Tagesson, C. Biomarkers of systemic inflammation in farmers with musculoskeletal disorders; a plasma proteomic study. *BMC Musculoskeletal Disord.* **2016**, *17*, 206.

(37) Bencharit, S.; Baxter, S. S.; Carlson, J.; Byrd, W. C.; Mayo, M. V.; Border, M. B.; Kohltfarber, H.; Urrutia, E.; Howard-Williams, E. L.; Offenbacher, S.; et al. Salivary proteins associated with hyperglycemia in diabetes: a proteomic analysis. *Mol. BioSyst.* **2013**, *9* (11), 2785−2797.

(38) Fröbel, J.; Cadeddu, R.-P.; Hartwig, S.; Bruns, I.; Wilk, C. M.; Kündgen, A.; Fischer, J. C.; Schroeder, T.; Steidl, U. G.; Germing, U.; et al. Platelet proteome analysis reveals integrin-dependent aggregation defects in patients with myelodysplastic syndromes. *Mol. Cell. Proteomics* **2013**, *12* (5), 1272−1280.

(39) Swaroop, A.; Kim, D.; Forrest, D. Transcriptional regulation of photoreceptor development and homeostasis in the mammalian retina. *Nat. Rev. Neurosci.* **2010**, *11* (8), 563−576.

(40) Nguyen, O. N. P.; Böhm, S.; Gießl, A.; Butz, E. S.; Wolfrum, U.; Brandstätter, J. H.; Wahl-Schott, C.; Biel, M.; Becirovic, E. Peripherin-2 differentially interacts with cone opsins in outer segments of cone photoreceptors. *Hum. Mol. Genet.* **2016**, ddw103.