Journal of proteome research

pubs.acs.org/jpr

Article

Protein Probability Model for High-Throughput Protein Identification by Mass Spectrometry-Based Proteomics

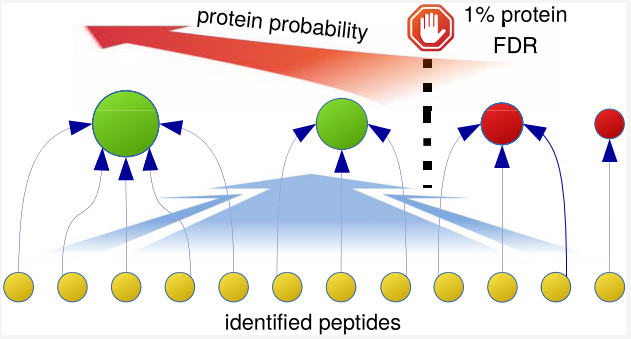
Gorka Prieto* and Jesús Vázquez*

Cite This: *J. Proteome Res.* 2020, 19, 1285−1297

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Shotgun proteomics is the method of choice for high-throughput protein identification; however, robust statistical methods are essential to automatize this task while minimizing the number of false identifications. The standard method for estimating the false discovery rate (FDR) of individual identifications and keeping it below a threshold (typically 1%) is the target-decoy approach. However, numerous works have shown that FDR at the protein level may become much larger than FDR at the peptide level. The development of an appropriate scoring model to identify proteins from their peptides using high-throughput shotgun proteomics is highly needed. In this study, we present a novel protein-level scoring algorithm that uses the scores of the identified peptides and maintains all of the properties expected for a true protein probability. We also present a refinement of the *picked* method to calculate FDR at the protein level. These algorithms can be used together as a robust identification workflow suitable for large-scale proteomics, and we show that the identification performance of this workflow is superior to that of other widely used methods in several samples and using different search engines. Our protein probability model offers the scientific community an algorithm that is easy to integrate into protein identification workflows for the automated analysis of shotgun proteomics data.



KEYWORDS: *FDR, proteomics, protein identification, target-decoy approach*

INTRODUCTION

In shotgun proteomics, MS/MS spectra are matched against the theoretical spectra of peptides resulting from an in silico digestion of a protein database. Each of these matches is called a peptide-to-spectrum match (PSM), and a score is assigned to each PSM depending on the similarity between the measured and theoretical spectra. The validity of identifications is checked by calculating the false discovery rate (FDR), which is usually set to a maximum of 1%. The same concept can be applied at the peptide level to control for the peptide FDR.

The most popular and widely accepted strategy for computing FDRs in shotgun proteomics is the target-decoy approach.[1,2] In this approach, two databases are used: a target database with all of the protein sequences that could be present in the sample, and a decoy database with fictitious protein sequences that should not be detected. If the decoy database is correctly constructed (with the same size and statistical distributions as the target database), the probability of obtaining a false PSM in the target and decoy databases is identical. Several strategies have been proposed for the calculation of FDR by the target-decoy approach. The number of above-threshold matches in the decoy database can be directly used to estimate the number of false matches in the target database. Since spectra yielding high scores in the target database also tend to produce high scores in the decoy database, the search is usually performed against a concatenated target + decoy database, so that target and decoy sequences compete for the spectra.[3] In turn, different formulae can be used to calculate the FDR at PSM or peptide levels using the competition strategy, depending on the population of matches used to estimate FDR.[2] Further refinements have been proposed to calculate peptide FDR avoiding drawbacks associated with the target-decoy competition.[4,5] In practice, the differences between these methods are small, and the competition target-decoy strategy for computing FDR is currently widely accepted in the proteomics community.

Nonetheless, the main goal of high-throughput proteomics is to identify not peptides but the proteins present in a biological sample. The identity of these proteins can be inferred from the peptides identified in the target database, but controlling the FDR at the protein level is not straightforward. Since each protein can be identified by several peptides, true peptide identifications tend to concentrate on the proteins present in the

Received: December 4, 2019
Published: February 10, 2020

https://dx.doi.org/10.1021/acs.jproteome.9b00819
J. Proteome Res. 2020, 19, 1285−1297

sample, whereas false peptide identifications are randomly distributed among the proteins in the database. Therefore, the ratio between the number of peptides and the number of proteins matched by these peptides is higher for target peptides than for decoy peptides, and consequently the protein-level FDR can be much larger than the peptide-level FDR.[2,6,7] Because of this protein FDR buildup problem, it is necessary to control FDR at the protein level and not just at the peptide level. Notably, this effect was not considered in one of the first drafts of the human proteome,[8] which used a peptide-level FDR threshold but did not control the amplification of the error rate at the protein level, resulting in invalid identifications, as pointed out in later analyses.[9,10]

One strategy for minimizing the impact of error-rate amplification is to decrease the FDR threshold value at the PSM or peptide levels until an acceptable protein-level FDR is achieved. For instance, in the Human Plasma Peptide Atlas, a 1% protein-level FDR was achieved by using a stringent PSM-level FDR filter of 0.0002.[11] However, such stringent criteria reduce the sensitivity of the identification workflow.

Another way to control amplification of the error rate is to compute a protein-level score from the scores of the corresponding peptides and estimate the FDR by directly applying the target-decoy approach at the protein level. Several protein-scoring methods have been proposed (reviewed in refs 2, 12), and a considerable effort has been devoted to addressing the identification of proteins that share common peptides.[13−18] The majority of the approaches convert the peptide scores into probabilities and integrate the peptide probabilities into a protein probability. This is the case, for instance, of ProteinProphet,[19] Mascot,[20] PeptideShaker,[21] MaxQuant,[22] PIA,[18] FIDO,[14] and ProteomeDiscoverer. An alternative approach is to assign each protein with the score of its best peptide,[2,23,24] as in Percolator.[24] The widespread implementation of FDR quality control has allowed comparison of the different approaches, and the best peptide scoring method currently seems the most efficient, especially for large data sets.[2,24] However, this is somewhat paradoxical, since from a purely statistical standpoint, one would expect to obtain more information and evidence for correct protein identification by considering all peptides mapping to a protein. This peptide-to-protein paradox illustrates the urgent need for the development of a comprehensive protein-scoring model that effectively integrates all of the existing peptide information. Such a model would facilitate the development of automated workflows with increased quality of protein information and improve annotation in protein databases.

An additional problem is the lack of a widely accepted standard for calculating protein FDR. While the target-decoy approach is accepted by most proteomics researchers, it can be put into practice in several different ways. By defining a decoy protein as one that is not matched by any of the peptides identified in the target database, the number of decoy proteins at a given protein score threshold can be used as a direct estimate of the number of false-positive identifications. Other authors took account of the fact that target proteins can harbor both true- and false-positive (FP) peptides and developed a model based on a complex hypergeometric distribution,[6] which was further simplified in a more compact formula.[25] More recently, a "*picked FDR*" approach has been presented, in which a competition strategy similar to that used at the peptide level was implemented at the protein level.[23] To further complicate the situation, commonly used protein identification workflows combine these FDR methods with different protein-scoring approaches.

In this paper, we propose a protein probability model derived from analytical considerations that integrates the information provided by all identified peptides, accurately predicts the random behavior of decoy proteins, and effectively resolves the peptide-to-protein paradox. We also propose a refined version of the *picked FDR* method for calculating protein-level FDR that rescues proteins that are missed due to the target-decoy competition. Our results were validated by analyzing the results from three search engines for several tissues from the Human Proteome Map (HPM) and two protein standards for the protein inference benchmark.[26−28]

EXPERIMENTAL SECTION

Proteomics Data Set

For this study, the proteomics data set we used to test different algorithms was one of the first drafts of the human proteome, submitted to ProteomeXchange as PXD000561.[8] This data set is known as the Human Proteome Map (HPM) and contains experimental data from different human tissues. We carried out comparisons at this tissue level and validated the reproducibility of the results in three tissues: `Adult_Heart`, `Adult_Liver`, and `Adult_Testis`. We selected these tissues because they span the data-set size range of the HPM: `Adult_Heart` has a lower number of identifications, `Adult_Testis` has one of the largest numbers of identifications, and `Adult_Liver` is an intermediate case. The raw files for each tissue have been converted to the `mgf` file format using `msconvert`[29] with the `peakPicking true 1-` option.

Target and Decoy Databases

The target fasta database was generated from GENCODE[30] version 25, with the addition of 47 common contaminants from UniProt. To account for PSM ambiguities in leucine vs isoleucine assignments, we replaced all leucines in the database with isoleucines.

The decoy database was generated with DecoyPyrat,[31] a software tool that generates decoy sequences with minimal overlap between target and decoy peptides. DecoyPyrat achieves this by first switching proteolytic cleavage sites with the preceding amino acid, reversing the database, and then shuffling any decoy sequences that become identical to target sequences.

Database Search

Three different search engines were used for the database search: X!Tandem (ALANINE 2017.02.01), Comet (2016.01 rev. 3), and MSFragger (build 20170103.0). The results for X! Tandem are shown in the main paper, and the results for Comet and MSFragger are provided as the Supporting Information. To respect the parameters used for the HPM,[8] we used a fragment monoisotopic mass error window of 0.05 Da, a parent monoisotopic mass error window of ±10 ppm, carbamidomethylation of cysteine as a fixed modification, oxidation of methionine as a variable modification, trypsin as the protease, and a maximum of two missed cleavages.

Our algorithm requires target-decoy competition, which can be performed searching against a concatenated target-decoy database or using separate searches against the target and decoy databases followed by target-decoy competition a posteriori. In this work, we followed the second procedure to conserve the option of using different FDR algorithms at the PSM or peptide

https://dx.doi.org/10.1021/acs.jproteome.9b00819
J. Proteome Res. 2020, 19, 1285−1297

Table 1. Summary of the Protein Probability Models and FDR Algorithms Used in This Work

(A) features common to all models

workflow step	description
MS/MS spectrum to PSM assignation	only the PSM with the best score (*rank* = 1) is considered per MS/MS spectrum.
PSM to peptide assignation	each peptide takes the score of the PSM with the best score
peptide to protein ambiguity	peptides shared by two or more proteins are not considered in this comparative to avoid bias produced by the peptide-to-protein assignation algorithm; to minimize ambiguities, genes are used instead of proteins.
decoy database	constructed using DecoyPyrat.
database search	performed separately against target and decoy databases; competition was performed a posteriori.
calculation of peptide probability	decoy and target peptide probabilities are calculated as *p*-values by nonparametric interpolation in the score distribution of decoy peptides.

(B) calculation of protein probability from peptide probabilities

protein probability algorithm	description
LPM	probability of the peptide with the best score.
LPS	product of probabilities of the *n* peptides that are matched.
LPF	product of probabilities of the *m* peptides that are identified at 1% peptide FDR.
LPGM	probability of obtaining a peptide with *LPM* when *n* peptides are matched.
LPGS	probability of obtaining *LPS* when *n* peptides are matched.
LPGF	probability of obtaining *LPF* when *m* peptides are identified at 1% peptide FDR among *n* matched peptides.

(C) calculation of protein FDR from protein probabilities

FDR algorithm	acronym	description
normal FDR	*FDRn*	ratio of decoy to target proteins above the protein probability threshold.
MAYU FDR	*FDRm*	ratio of false positive (FP) to target proteins above the protein probability threshold. FP was estimated assuming that the proportion of false positives is identical in target and decoy protein populations.
picked FDR	*FDRp*	ratio of decoy to target proteins above the protein probability threshold after target-decoy protein competition.
refined FDR	*FDRr*	ratio of refined decoy to target proteins above the protein probability threshold; refined decoy proteins are estimated assuming that decoy and FP proteins have the same distribution after target-decoy protein competition.

level, but the conclusions presented here are not affected by the competition method used.

A scheme of the features common to all models used in this work is presented in Table 1A.

Processing Details

To compute a peptide-level FDR, we considered the best PSM for each spectrum (*rank* = 1) and then the best PSM for each peptide.

The source code and a binary distributable to calculate protein probabilities and FDRs from the results of separated or concatenated database searches are available at https://github.com/akrogp/SpHPP/tree/master/dist/LPGF.

RESULTS AND DISCUSSION

Derivation of a Protein Probability Model

Generic identification workflows start by considering PSM-level scores, which are integrated into peptide-level scores and then into protein- or gene-level scores. In this study, we focused on the problem of computing protein-level probabilities from peptide-level probabilities. Hence, we made two simplifications. First, we avoided the problem of integrating PSM-level scores into peptide-level scores by assigning to each peptide the score of its best PSM; this commonly used simplification is used in most identification workflows.[7]

The second simplification is related to the protein inference problem.[32] Since the method chosen to manage peptides shared by two or more proteins or to perform protein grouping affects the outcome of protein inference algorithms, we decided to test the performance of the protein probability models presented here using only unique peptides. This simplification avoids a potential bias of sharing/grouping algorithms in favor of one or other of the models, allowing a more objective judgment of their comparative performance. Note, however, that the probability models proposed here are fully compatible with any peptide-to-protein assignation algorithm of nonunique peptides provided that each peptide is assigned to only one protein (e.g., the most probable protein, also called "master" or "leading razor" protein[22]). To minimize the effect of this simplification, we considered peptides that were unique at the gene level. Since most of the ambiguities come from different protein products of the same gene, we did not exclude peptides shared by different proteins produced by the same gene, avoiding an excessive decrease in data-set size. The results are thus provided at the gene level, although sometimes we continue to use the protein term for simplicity.

Since the different search engines provide different types of scores, in our model, we first translate the scores into a general-purpose peptide-level *p*-value, in a process we call "calibration". The *p*-value, to which we will also refer here as peptide probability, is defined as the probability of obtaining a peptide with equal or better score by chance alone, following the classical definition by Käll et al.[1] This value corresponds to the false-positive rate, and when expressed as a probability density, it has also been called the likelihood of obtaining a score given that the null hypothesis is true;[1] it corresponds to the score distribution of incorrect peptide assignments in the classical mixture model formulation of PeptideProphet.[2,33] Note that this measure is obtained from the global or average distribution of scores in the entire data set and should not be confounded with the *p*-values

https://dx.doi.org/10.1021/acs.jproteome.9b00819
J. Proteome Res. 2020, 19, 1285−1297

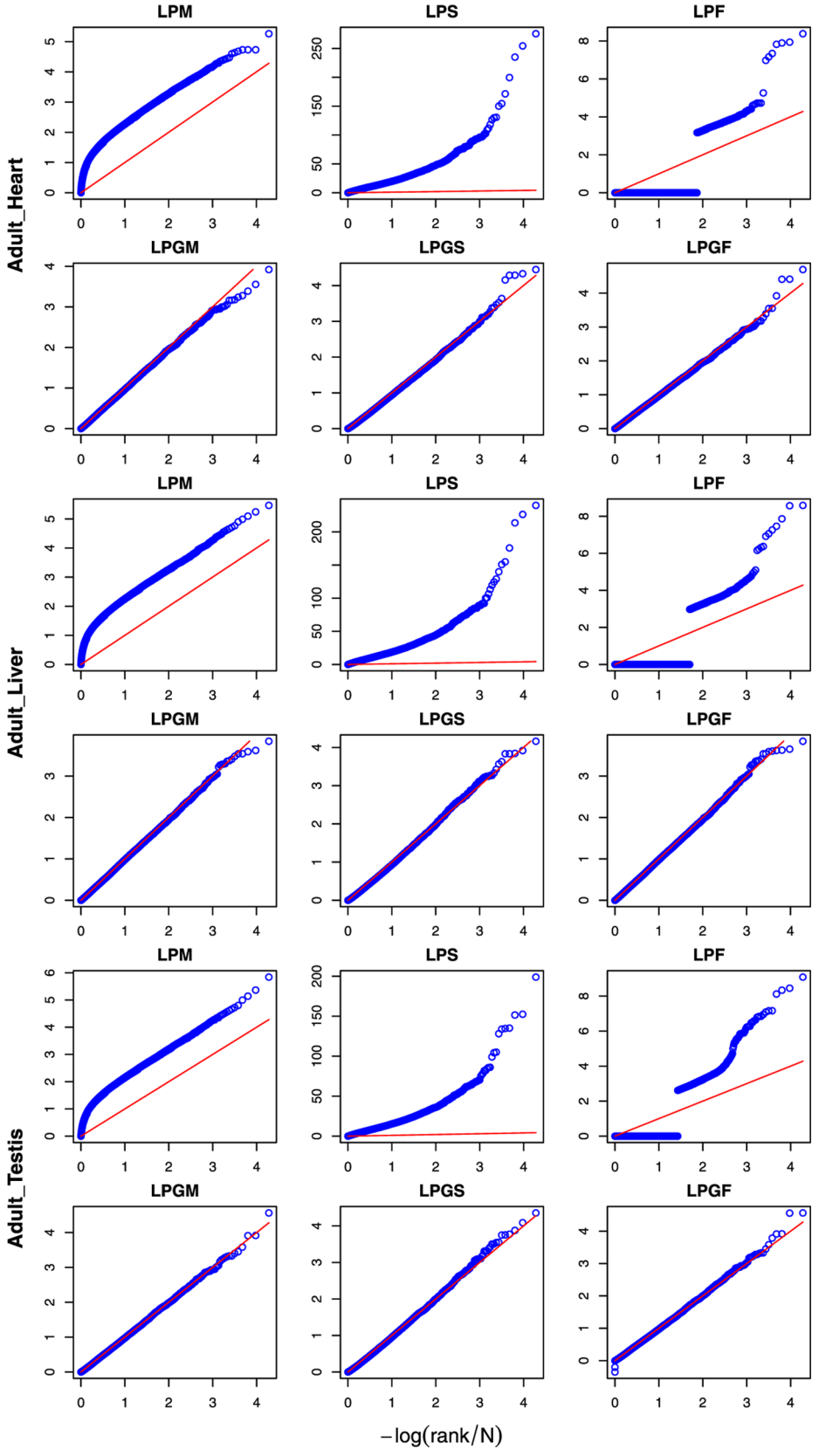


Figure 1. Distribution of different decoy protein scores using three tissues from the Human Protein Map as separate test data sets. The *y*-axis represents the cologarithm of protein probabilities calculated by the methods of Table 1B as indicated by the title of each graph. The *x*-axis represents the cologarithm of the expected uniform protein probabilities. Deviations from the identity line (drawn in red) mean that the calculated probabilities are inaccurate.

https://dx.doi.org/10.1021/acs.jproteome.9b00819
J. Proteome Res. 2020, 19, 1285−1297

assigned to PSMs by some search engines from the analysis of the single-spectrum distributions.[2,34] Note also that by score, we refer not only to a parameter directly produced by the search engine but also to any quality parameter that combines information derived from several scores (such as in PeptideProphet[33]) or includes complementary data usually implemented in postsearch filtering software (such as the machine learning algorithm from Percolator[35]).

In our model, we assume that peptide probabilities are properly calibrated, in the sense that they follow a uniform distribution. Well-calibrated probabilities are produced, for instance, by unsupervised[33] or semisupervised parametric models that fit the distribution of incorrect peptide assignments or of decoy peptides using Gamma or Gumbel distributions.[36] Semiparametric models have also been used for this purpose.[36]

Although any of these formulations may be used in the model we present here, for conceptual simplicity and general applicability, we convert scores into peptide probabilities from the analysis of the population of decoy peptides, which are assumed to be produced by random events, by a simple nonparametric approach, as proposed previously.[7,34,37] After a database search, a single ordered list of decoy and target peptides is produced according to their scores. For each peptide (target or decoy), we count the number of decoy peptides (d) with an equal or better score and divide this number by the total number of decoy peptides (D)

$$p\text{-value}(\text{peptide}_i) = \frac{d + k}{D} \tag{1}$$

where k is a constant offset value. For decoy peptides, $k = -0.5$; this is a continuity correction factor used to center the rank bins. For target peptides, $k = +0.5$, so that they conservatively take the value of the closest decoy with a worse score. Note that this procedure assigns the probability of the best decoy peptide to all of the target peptides whose score is better than the best decoy score, as previously done.[34] This is a conservative approach that avoids extrapolating peptide probabilities in a region where the score behavior of decoy peptides is not known. In any case, extrapolating or not p-values beyond the best decoy peptide does not have appreciable influence on the results obtained at the protein level as shown later. Equation 1 ensures that the p-values for decoy peptides follow a uniform distribution.

Since protein probability functions usually derive from multiplicative operations on their peptide probabilities, it was more convenient to use the cologarithm of p-values

$$LP_i = -\log_{10}(p\text{-value}(\text{peptide}_i)) \tag{2}$$

where LP refers to the co*L*ogarithm of the *P*robability.

Inspection of the algorithms employed by commonly used software tools reveals that, independent of the method used to derive peptide probabilities, the majority of them use simple multiplicative operations to integrate peptide probabilities into a protein probability. This method is followed by ProteinProphet,[19] Mascot,[20] PeptideShaker,[21] MaxQuant,[22] PIA,[18] FIDO,[14] and the latest versions of the ProteomeDiscoverer package. The latest version of Percolator assigns to each protein the probability of the best peptide.[24] Using the notation of eq 2, we can generically represent these approaches as follows:

$$LPM = \max_{i=1,\ldots,n} LP_i \tag{3}$$

$$LPS = \sum_{i=1}^{n} LP_i \tag{4}$$

$$LPF = \sum_{i=1}^{n} LP_i \quad \forall\ i|FDR_i \leq 1\% \tag{5}$$

where *LPM* stands for *LP Maximum* and assigns to the protein the probability of its best peptide (i.e., the peptide with the lowest probability). *LPS* stands for *LP Sum* and estimates the protein probability as the product of the probability of all matched peptides that belong to the protein. The matched peptides are defined as the collection of peptides that are scored as the best candidate per spectrum by the search engine, without any consideration about the PSM quality. Finally, *LPF*, which stands for *LP Filter*, is a variant of *LPS* that only integrates the probability of the identified peptides. The identified peptides are defined as the collection of peptides that are correctly identified according to a specific criterion; in this work, we used a 1% FDR threshold at the peptide level. A scheme of the probability models used in this work is presented in Table 1B.

Equations 3−5 are constructed from calibrated (i.e., uniformly distributed) and independent decoy peptide probabilities; however, when applied to decoy proteins, none of these scores followed the expected uniform distribution (Figure 1). This can be explained in the case of *LPM*, since the probability of the best score is not a true measure of protein probability. However, *LPS* increases much faster than would be expected for a true probability, and this marked deviation from a uniform distribution is somewhat surprising, since proteomics specialists have often assumed that the product of peptide probabilities is a good estimate of protein probability.[15,19,22,38] The deviation from the uniform distribution, though less pronounced, still grows too fast when *LPF* is used, i.e., when only the decoy peptides above the peptide-FDR threshold are considered in the protein score (Figure 1). These results indicate that these protein scores do not reflect the actual behavior of decoy proteins.

In an effort to derive alternative formulas able to predict decoy behavior, we considered each of these cases separately. We tried to reformulate the problem in the form of an appropriate question from which a probability model could be derived, as proposed for the FDR.[39] If we are to use the best peptide as a means of scoring proteins, the question to address is, "What is the probability of getting a decoy protein that contains at least one decoy peptide with an equal or lower peptide probability when the protein has been assigned n matched peptides?" Since this event is complementary to having all n peptides matched with a higher p-value, we arrive at the expression

$$\text{probability}(LPM_{\text{protein}}) = 1 - (1 - p)^n \tag{6}$$

where p is the p-value of the peptide with the lowest probability. This equation is a well-known result of order statistics. Applying cologarithms to the above equation, and since from eq 2 $p = 10^{-LP_i}$, the equation is re-expressed as follows:

$$LPGM = -\log_{10}(1 - (1 - 10^{-LPM})^n) \tag{7}$$

where *LPGM* stands for the gamma formulation of the *M*aximum *LP* value since this value can also be computed using a gamma function. Notably, when *LPGM* was used as a protein score, it followed very accurately the uniform distribution, and this result was reproduced in all of the data sets analyzed (Figure 1). Besides, this finding was also

https://dx.doi.org/10.1021/acs.jproteome.9b00819
J. Proteome Res. 2020, 19, 1285−1297

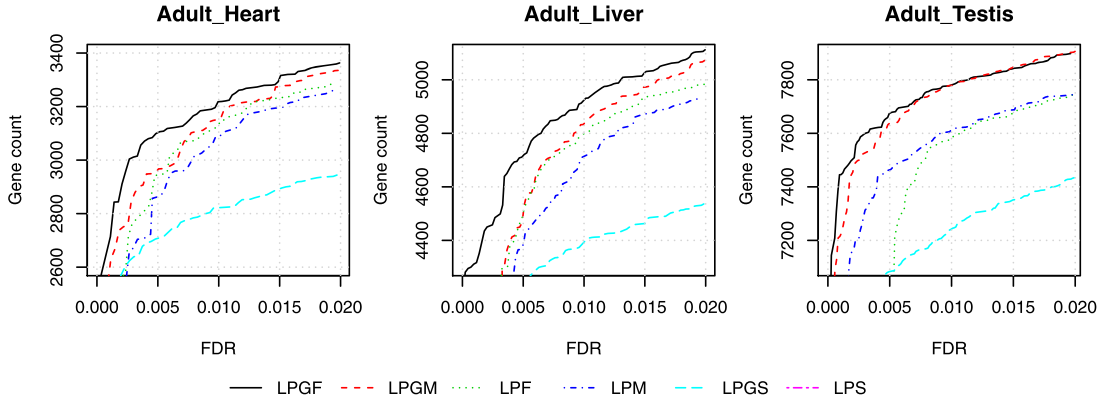


Figure 2. Number of identified genes as a function of the FDR threshold for different protein score types and using as separated tests three tissues from the Human Protein Map. The number of identifications provided by *LPS* is so small (less than 400) that it is not depicted in the figure. For these calculations, the FDR was calculated as the fraction of decoy proteins divided by the number of target proteins that pass the protein score threshold (i.e., the *FDRn* method defined in eq 11).

consistently reproduced when different search engines were used (Figures S1 and S2). We concluded that *LPGM* behaved as a true protein probability.

We used a similar approach for the other two protein scores. If a protein score is to consider all of the matched peptides of a protein, as *LPS* does, we would have to take account of the set of *p*-values for each of the peptides. However, the formulation of a probability question here is not straightforward. Let us assume that a certain decoy protein is matched by a "configuration" comprising *n* matched peptides with a list of the corresponding *p*-values. To calculate the probability of obtaining such a configuration or a better one by chance would require the definition of a criterion to determine whether a specific configuration is better than another, so that the decoy proteins can be ranked accordingly. While there are several possible ways to define such a criterion, we reasoned that the product of peptide *p*-values had some of the properties needed to rank the configurations as expected for a random event. This is evident in some cases. For instance, if several decoy proteins are matched by the same number of peptides but have different *p*-values, we can intuitively affirm that the protein most probably matched by chance is the one having the highest product of *p*-values. A similar argument can be made for the case of two proteins matched with the same number of peptides with the same *p*-values, except that the second protein is matched by an additional peptide; in such a case, the protein with the lower number of peptides is more likely to be detected by chance, which again corresponds to the protein with the highest product of peptide *p*-values. In the most general case, we need to take into account the number of matched peptides and the product of *p*-values. These two parameters can be used to formulate a coherent probability question: "What is the probability of getting a decoy protein with an equal or lower product of peptide *p*-values when it is matched by *n* peptides?" According to statistical theory, the natural logarithm of the product of *n* independent and identically distributed uniform $[0, 1]$ random variables follows a gamma$(n,1)$ distribution.[40] Since peptide *p*-values are uniformly distributed, we can use the gamma function to construct a protein probability as follows:

$$\text{probability}(LPS_{\text{protein}}) = 1 - G(-\ln(P);\, n,\, 1) \tag{8}$$

where P is the product of *p*-values of the n peptides and $G(x; k, \theta)$ is the cumulative distribution function of the gamma distribution with shape k and scale θ. Using the cologarithmic notation, and since from eqs 2 and 4 $LPS = -\log(P) = -\ln(P)/\ln(10)$, we obtain that

$$LPGS = -\log_{10}(1 - G(LPS\cdot\ln\ 10;\, n,\, 1)) \tag{9}$$

LPGS stands for *LP* gamma *of the Sum*. In marked contrast to *LPS*, *LPGS* has the expected protein probability properties and accurately predicts the observed probability of matching a decoy protein in all samples across different search engines (Figures 1, S1, and S2). Note that in the case of MSFragger, a deviation from the expected trend was observed for the top-scoring proteins (Figure S2). A similar deviation, although less accused, was also observed in the case of Comet (Figure S1). A close inspection of the three best matching peptides from these decoy proteins (Table S1) revealed the presence of sequences containing high proportions of repeated amino acids like Gly, Ala, Ser, Pro, and Leu; all of these sequences, typical of proteins like keratins and collagens, are highly homologous to target proteins and produce nonrandom PSM. These decoy-target homologies are difficult, if not impossible, to be avoided. Hence, these deviations are produced by imperfections in the decoy database and not to the protein probability algorithm.

The *LPF* score can be thought of as a refinement of *LPS* that only considers the peptides that have been identified according to the criterion used to select correct identifications. The question to address in this case is, "What is the probability of getting a decoy protein with an equal or lower *p*-value product from *m* identified peptides that are selected among *n* matched peptides?" To construct a probability function using the gamma distribution, we had to take into account the fact that several subsets of identified peptides are possible from the same set of matched peptides. We also had to consider proteins for which none of the matched peptides were positively identified; in these cases, the *LPM* value was assigned to *LPF*. The probability function then takes the form

$$LPGF = \begin{cases} LPGM & m = 0 \\ -\log_{10}\left((1 - G(LPF\cdot\ln 10;\, m,\, 1))\cdot\binom{n}{m}\right) & m \geq 1 \end{cases} \tag{10}$$

Again, *LPGF* accurately predicts the behavior of decoy proteins in all tissues and across all search engines (Figures 1, S1, and S2) and thus acts as a true protein probability score. Note that, in contrast to *LPGS*, *LPGF* stands very well the imperfections in the decoy database, showing negligible deviations in the

https://dx.doi.org/10.1021/acs.jproteome.9b00819
J. Proteome Res. 2020, 19, 1285−1297

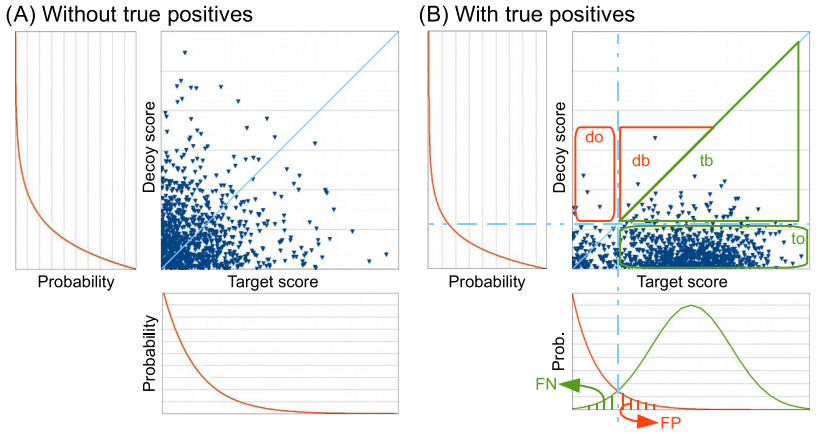


Figure 3. Competition between target and decoy proteins for FDR estimation using a simulated data set. In the dot plots, the scores of the target proteins are plotted against the scores of the corresponding decoy proteins. The plots at the left and below these dot plots represent the distribution of protein probabilities against the target or decoy scores. (A) When no true-positive (TP) target proteins are present, the probability distributions of decoy protein scores and false-positive (FP) target protein scores are identical (red lines), and therefore the points are distributed symmetrically across the diagonal in the dot plots (blue triangles). *FDRp* and *FDRr* are both based on this symmetry. (B) When there are TP target proteins, the cloud of target protein scores is shifted to the right due to the presence of TP scores (TP distribution, in green). The diagonal and the score threshold (dashed blue lines) delimit four regions. The "decoy-only" (*do*) region contains pairs in which the decoy protein has an above-threshold score and the target protein has a below-threshold score. The "decoy-best" (*db*) region contains pairs for which both scores are above threshold, but the decoy protein score is better than the target protein score. The "target-only" (*to*) and "target-best" (*tb*) regions are defined in a similar fashion. The number of points in these regions is used in different manners to compute *FDRp* and *FDRr*.

expected trend for top-scoring proteins with MSFragger and Comet (Figures S1 and S2).

Analysis of the performance of these six scores—measured as the number of target proteins identified at fixed protein FDR thresholds—revealed three consistent patterns that were maintained in all of the tissues and across all search engines (Figures 2, S3, and S4). First, each of the three protein probability scores theoretically derived here (*LPGM*, *LPGS*, and *LPGF*) was more sensitive than its original counterpart (*LPM*, *LPS*, and *LPF*). This may be because *LPM* uses the information from only one peptide, discarding the evidence from the other identified peptides about the presence of the protein; in contrast, *LPGM* selects the best peptide after considering a total of *n* peptides. Similarly, *LPS* and *LPF* do not account for protein length, since they only consider the final product of peptide *p*-values, thus introducing a bias toward larger proteins, which tend to be matched by more decoy peptides. In contrast, *LPGS* takes into account the number of peptides used to calculate the product, and *LPGF* includes not only the number of peptides but also the total number of peptides matching the protein. In general, this finding highlights the importance of using true probabilities that accurately reflect decoy protein behavior instead of empiric protein scores.

The second key observation is that *LPGM* and *LPM* were consistently more sensitive than *LPGS* and *LPS*, respectively. This finding reflects the detrimental accumulation in *LPGS* and *LPS* of false target peptide matches, which only add random noise and contribute nothing to target protein identification. This phenomenon is avoided in *LPGM* and *LPM*, increasing their sensitivity despite the inclusion of only the best peptide in the score.

The third and most important pattern is that *LPGF* was consistently more sensitive than *LPGM*. This contrasts with *LPF*, which was less sensitive than *LPM* in most cases. This finding shows that the inclusion of all relevant peptide information in a well-constructed protein probability score is always preferable to the simplification provided by using only the best peptide. More importantly, it resolves the peptide–protein paradox, demonstrating that the inclusion of relevant information produces better results than ignoring it and establishing the basis for a rational approach to protein identification.

Finally, we analyzed whether the specific method used to calculate peptide probabilities could influence the relative performance of the different algorithms. We specifically analyzed whether using a parametric model to calculate *p*-values, instead of the ranking formula of eq 1, could have a significant impact on the results. For this purpose, we fitted the Comet score distribution of decoy peptides from the `Adult_Liver` data set to a gamma function, following the approach implemented in PeptideProphet.[36] The fitted function was used to estimate *p*-values for decoy and target peptides. We then compared the protein identification performances of *LPM*, *LPF*, and *LPGF* calculated from *p*-values estimated using eq 1 with those obtained using the gamma function. As shown in Figure S5, the results obtained by the two methods were virtually indistinguishable.

Derivation of a Refined Protein FDR

Let us assume a hypothetical experiment in which spectra are searched against a decoy and a target database, but the proteins in the sample are not present in the protein database, so that all protein values provided by a given protein-scoring algorithm come from false assignations. If we calculate a protein score on the basis of the matched peptides belonging to each of the proteins, then the score distribution in the two databases is expected to be identical and the representation of decoy vs target scores would generate a cloud of points symmetrically distributed around the diagonal (Figure 3A). The presence of true protein identifications can be viewed as an increase in the score of proteins harboring true target peptide matches, producing a horizontal displacement of scores to the right. The resulting distribution of target protein scores can thus be viewed as a superposition of the decoy distribution (containing

https://dx.doi.org/10.1021/acs.jproteome.9b00819
J. Proteome Res. 2020, 19, 1285−1297

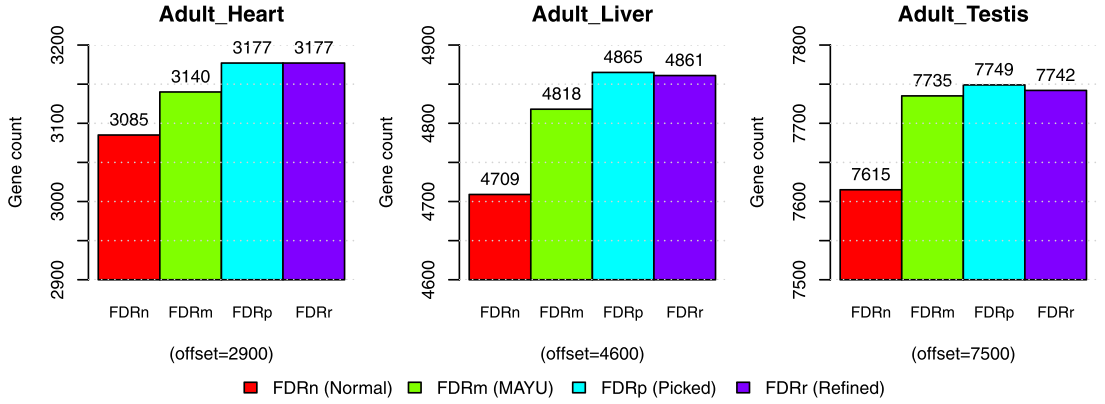


Figure 4. Number of identifications using the different protein-level FDR algorithms for three sample tissues in the Human Proteome Map. For this comparison, the protein score has been calculated simply as the score of its best peptide. Note that the axes have different offset values to better highlight differences.

the false-positive target protein scores) and the distribution of true-positive (TP) protein scores.

Let us also suppose that a protein score threshold is applied to select a population of positively identified target proteins. The FDR of this population can be calculated in several ways. In the simplest approach, we use the number of above-threshold decoy proteins (d) to estimate the fraction of false positives in the population of above-threshold target proteins (t)

$$FDRn = \frac{d}{t} \tag{11}$$

A scheme of the FDR methods used in this work is presented in Table 1C. Note that, depending on the method used to apply the score threshold to the ranked list of peptides, eq 11 may have to be substituted by

$$FDRn = \frac{d + 1}{t} \tag{12}$$

which, as proposed by some authors,[41,42] may produce a more accurate estimation of the FDR. This correction may be particularly useful when filtering small data sets or using very low FDR thresholds. A similar correction should be considered for the other FDR models. For large data sets such as those analyzed in this work, the effect of this correction is negligible, and for simplicity, we will use the uncorrected equations.

Using the notation we proposed previously to compare methods for estimating peptide FDR,[4] we defined a set of regions delimited by the diagonal and the score threshold. This allowed us to express *FDRn* as

$$FDRn = \frac{do + db + tb}{to + db + tb} \tag{13}$$

In the MAYU method,[25] FDR is estimated as

$$FDRm = \frac{d}{t} \cdot \frac{T - t}{D - d} \tag{14}$$

where D and T are the sizes of the decoy and target databases. This equation can be derived under the basic assumption that the proportion of decoy matches above (d) and below ($D - d$) the protein score threshold is identical to the proportion of false-positive (FP) target matches above threshold and the number of target matches below threshold ($T - t$)

$$\frac{d}{D - d} = \frac{FP}{T - t} \tag{15}$$

In this approach, FP is considered a more accurate estimate and replaces d in eq 11. However, as schematized in Figure 3B, the separation between the false and true target protein distributions is, in general, incomplete. Therefore, the population $T - t$ target proteins below the score threshold are not only composed of false protein identifications but also contain some false negatives. Therefore, the denominator in eq 15 is an overestimate to some extent, and FP is also overestimated.

In the *picked FDR* approach, the target and decoy scores of the same protein are treated as pairs rather than as independent entities, and the protein that receives the highest score is selected; the other one is discarded.[23] This kind of competition generates two populations of score pairs on either side of the diagonal, so that the *picked FDR* can be expressed thus

$$FDRp = \frac{do + db}{to + tb} \tag{16}$$

This equation highlights a conceptual problem in this definition of FDR: a subpopulation of target proteins (the *db* region) is discarded, despite having a score above the threshold. In the original target-decoy approach at the peptide level, the decoy and target sequences compete for the same spectrum, and thus it makes sense to discard the target sequence when the decoy gets the higher score because this shows that the target sequence does not achieve a score higher than that obtained by a random sequence. However, at the protein level, the situation is completely different; the decoy counterpart of a target protein is matched by different spectra, and therefore its actual score gives no information about the quality of the target protein identification. Hence, there is no reason to discard the target protein on the basis of the score of its decoy counterpart.

We reasoned that the target-decoy competition strategy at the protein level is a valid way to estimate statistical measures by exploiting the symmetry around the diagonal; however, direct comparison of target and decoy protein pairs should be avoided. With this idea in mind, we derived a "refined FDR" approach, based on a concept proposed previously at the peptide level.[4] The population of target proteins with a score above threshold includes three regions (*db*, *tb*, and *to*) as shown in Figure 3. Because of the symmetry of false-positives and decoy proteins around the diagonal, the expected total number of false positives in these regions is given by $do + 2 \cdot db$. Therefore, the refined FDR is given by

https://dx.doi.org/10.1021/acs.jproteome.9b00819
J. Proteome Res. 2020, 19, 1285−1297

Table 2. Number of Identifications Provided by the Different Workflows Using Three Tissues of the Human Proteome Map as Separate Target Data Sets

	Adult_Heart		Adult_Liver		Adult_Testis	
	genes	geneFDR (%)	genes	geneFDR (%)	genes	geneFDR (%)
1% pepFDR	3 561	7.27	5 384	7.04	8 023	8.81
adjusted pepFDR	3 085	0.97	4 709	0.98	7 615	1.00
FDRn(*LPF*)	3 125	0.99	4 778	0.94	7 573	0.96
FDRp(*LPM*)	3 177	0.98	4 865	0.97	7 749	0.99
FDRr(*LPGF*)	3 268	0.98	5 010	1.00	7 852	0.99

target database size = 20 407 genes

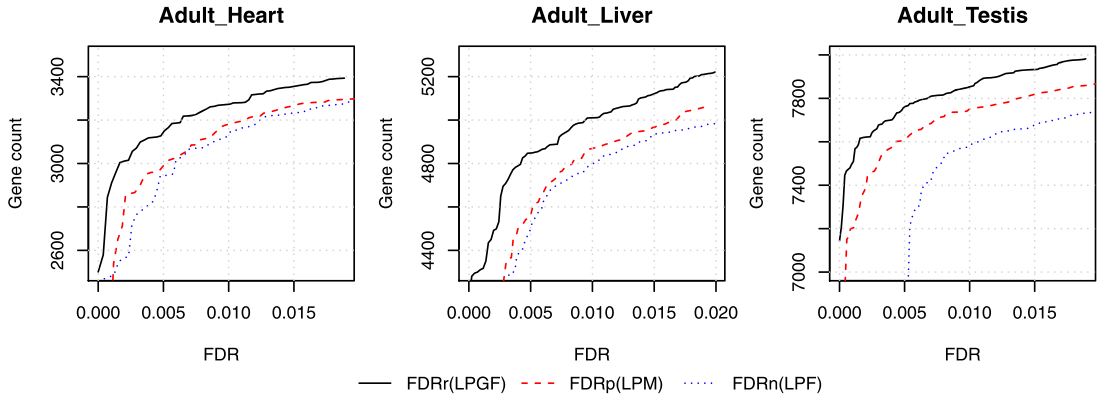


Figure 5. Number of identified genes as a function of the FDR threshold for different protein identification workflows, using three tissues from the Human Protein Map as separate tests.

$$FDRr = \frac{do + 2 \cdot db}{to + tb + db} \quad (17)$$

This equation differs from eq 16 only in the term *db* in the numerator and denominator. Since the denominator is larger than the numerator, $FDRp \leq FDRr$. In other words, the two methods differ only slightly, with *FDRp* being somewhat more sensitive that *FDRr*, since it requires a lower protein score threshold to reach the same FDR. However, the higher sensitivity of the *picked FDR* is obtained by rescuing proteins whose score is just below the threshold, while the refined approach rescues proteins in the *db* region, which may have a score well above the threshold.

The number of identifications obtained with the different protein-level FDR algorithms is shown in Figure 4; the same protein score was used in all cases. As expected, *FDRn* produces the smallest number of identifications, while *FDRp* and *FDRr* yield more identifications than MAYU. Also as expected, *FDRp* yields slightly more identifications than *FDRr*; however, *FDRp* loses some proteins that are clear positive identifications in *FDRr*. For instance, in the `Adult_Liver` tissue results, we found that *FDRp* discarded a protein, GIGYF1, which is identified with a peptide with such a low *p*-value (7.2×10^{-5}) that it should be accepted as a valid identification. This gene is eliminated because its decoy pair contains a decoy peptide with sequence `IIIIEQR` that yields a valid PSM (*p*-value = 8.0×10^{-6}) because it has a high homology with a target protein. Note that these two peptides are matching different spectra, and therefore the decoy match does not give any relevant information.

Comparison with Other Protein Identification Workflows

To analyze the performance of the new formulations in practice, we compared the results obtained from analysis of three tissues from the HPM data set with several commonly used identification workflows. Inspection of the workflows used in the literature reveals three common types. In one kind of workflow, peptide identifications are filtered according to their peptide-level FDR, and the FDR of the resulting list of proteins is calculated using the conventional approach (*FDRn*); this method is frequently used by researchers in the field. In a second type of workflow, protein probabilities are calculated using multiplicative formulations to integrate the probabilities of their identified peptides, and the FDR is calculated using the traditional *FDRn* approach. This approach, which we will call *FDRn*(*LPF*), is used, for instance, by PeptideShaker,[21] MaxQuant,[22] PIA,[18] Mascot,[20] FIDO,[14] and also by the most recent versions of ProteomeDiscoverer. In a third type of workflow, the proteins are assigned the score of their best peptide, and the *picked FDR* is used to validate the results. This workflow, which we will call *FDRp*(*LPM*), is used in the most recent version (3.0) of Percolator.[24] Finally, we used the best performing protein probability model proposed here (*LPGF*) in combination with the refined protein-level FDR; we call this workflow *FDRr*(*LPGF*).

Filtering by 1% FDR at the peptide level produced a list of proteins whose protein *FDRn* was >7% in all cases (Table 2). In this approach, the peptide-level FDR could be adjusted until 1% protein FDR was achieved, but this came at the cost of a marked decrease in protein identification performance. This result highlights the error-rate amplification that occurs when peptides are integrated into proteins in the absence of an appropriate protein-scoring model. Using the product of peptide probabilities (*LPF*) as the protein score allowed direct control of protein FDR; however, this method produced only a moderate increase in the number of proteins identified, and the improvement was not observed in all cases. The *FDRp*(*LPM*) workflow increased the protein identification performance presumably because the use of only the best peptide minimizes

https://dx.doi.org/10.1021/acs.jproteome.9b00819
J. Proteome Res. 2020, 19, 1285−1297

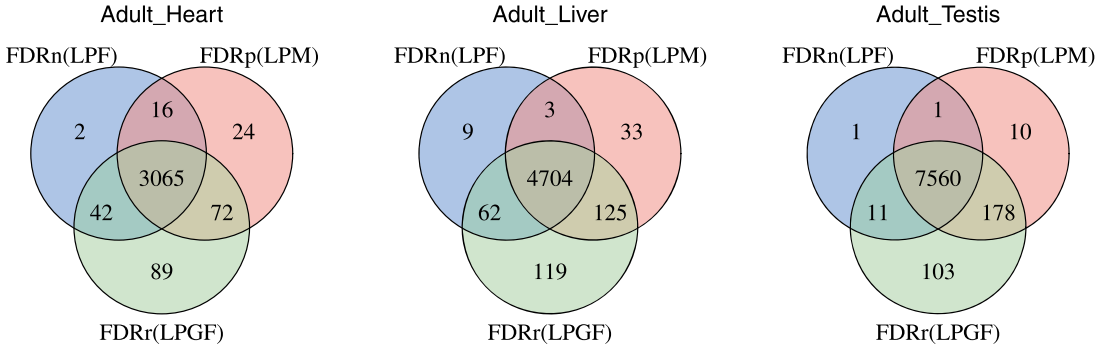


Figure 6. Venn diagrams showing the number of genes identified by different protein identification workflows in three tissues of the Human Protein Map.

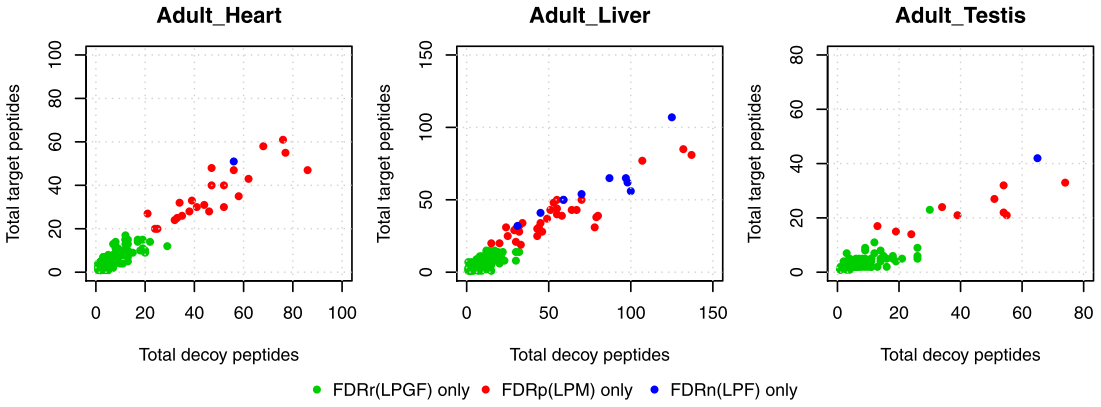


Figure 7. Comparison of target versus decoy peptides for each gene identified exclusively by each of the three protein identification workflows discussed in the text. The total number of peptides is considered, without filtering by FDR. Each point corresponds to a target-decoy pair. The comparison was carried out in three tissues of the Human Protein Map.

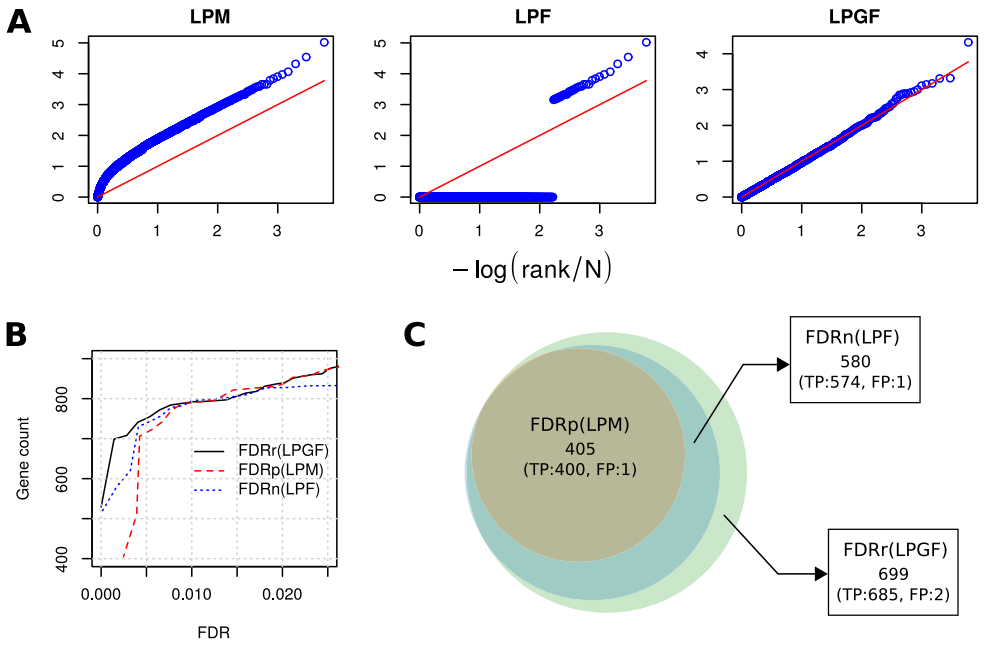


Figure 8. Performance of the different workflows using a yeast gold standard data set. (A) Distribution of the protein decoy scores. The meaning of the points and lines is as in Figure 1. (B) Number of identified genes as a function of the FDR threshold. (C) Venn diagrams showing the number of genes identified at 0.25% FDR. The numbers of true positives (*TPs*) and false positives (*FPs*) are indicated in parenthesis. A protein is considered *TP* when it is included in at least two of the three non-MS yeast data sets or the four MS yeast data sets.[26] Proteins that were absent in the four MS data sets were considered *FP*, as proposed.[26]

the error-rate increase for decoy proteins (Table 2 and Figure 5). The algorithm proposed here (*LPGF*), in combination with the refined FDR, is able to use the information provided by all of the peptides in a more efficient way, outperforming the other

https://dx.doi.org/10.1021/acs.jproteome.9b00819
J. Proteome Res. 2020, 19, 1285−1297

algorithms in all cases (Table 2 and Figure 5). This result was consistently reproduced when the same data were analyzed in other search engines (Figures S6 and S7).

To explore these results in more depth, we analyzed the proteins differentially identified by the protein-scoring workflows. *FDRr*(*LPGF*) included most of the identified proteins, missing only a small number of proteins identified by the other workflows (Figure 6). Closer inspection revealed that all proteins identified only by *FDRp*(*LPM*) were wonder-hits (i.e., proteins identified from only one peptide) that were matched by unexpectedly large numbers of target peptides (more than 50 in most cases). Wonder-hits were the minority of proteins identified by *FDRn*(*LPF*) only; however, these proteins were also matched by an abnormally high number of target peptides. To analyze the random peptide matching behavior of these proteins, we plotted the number of peptides matched to the decoy and target databases separately for the subsets of proteins identified by each one of the three workflows only. Proteins identified only by *FDRp*(*LPM*) or by *FDRn*(*LPF*) were matched by more than 20 peptides and sometimes by more than 100 peptides when searched against either the decoy or target databases; this indicates that these were proteins that, due to their size and sequence, tended to receive a large number of random matches and therefore were more likely to be identified by false peptides (Figure 7). In clear contrast, the proteins identified by *FDRr*(*LPGF*) only were matched in most cases by fewer than a dozen peptides (Figure 7), indicating that their positive identification was less likely due to random matching and therefore that these proteins were most likely to be true positives. These findings were consistently reproduced with other search engines (Figures S8−S11), indicating that this behavior was not specific to any particular search engine. We concluded that the novel protein identification workflow proposed here not only was more sensitive but also provided identifications that were better supported by statistical considerations.

Validation Using a Yeast Gold Standard Data Set

To further validate the results obtained by using the new protein probability method, we analyzed two protein standards recently proposed for the protein inference benchmark. First, we used the PrEST *mixture A*[28] containing true 191 human PrEST proteins and searched it against FASTA files containing these proteins plus the 1000 additional false "entrapment" sequences. This data set, however, contained too few protein sequences to detect differences, and the three models identified 181 proteins at 1% protein FDR, among which only 1 was a false positive.

Second, we used the "Gold Standard of Protein Expression in Yeast" used by Audain et al.,[27] which was generated by Ramakrishnan et al.[26] This data set was significantly smaller than the human tissue data sets but was large enough to capture differences between the protein probability models. As shown in Figure 8A, in this data set, *LPGF* provided an accurate estimation of the behavior of decoy proteins, while *LPF* and *LPM* overestimated the probability, confirming the results obtained with human tissues. Figure 8B also confirmed that *FDRr*(*LPGF*) performed significantly better than the other models below 1% protein FDR. Interestingly, this data set also showed that the majority of extra proteins identified by *FDRr*(*LPGF*) and not by the other two models were true positives (Figure 8C).

CONCLUSIONS

In this paper, we present a novel identification workflow based on a newly developed protein probability model together with a refined version of the *picked FDR* method to calculate FDR at the protein level. The refined FDR protein algorithm is based on the refined FDR method we proposed previously to validate identifications at the peptide level.[4] While the protein *FDRr* model does not offer improved performance over the *picked FDR* method proposed previously,[23] it avoids the loss of a certain proportion of proteins that are discarded by *FDRp* and can thus be viewed as a refinement of the *FDRp* method. In contrast, our protein probability model is completely new and is solely based on analytical considerations, making it, to our knowledge, the first protein probability model to accurately predict the random behavior of decoy proteins. Most importantly, our model takes into account the information of all identified peptides, and our results demonstrate that the integration of all this information into a well-constructed, true probability model yields a superior identification performance to that achieved by using only the information provided by the best peptide. This observation was consistently reproduced in HPM data sets for three tissues and using several search engines.

In this work, we based our protein probability algorithm on the analysis of *p*-values and not on posterior error probabilities (PEPs), which are also commonly used in the scientific literature. We preferred the use of *p*-values by mathematical simplicity; our model only needs to calibrate the distribution of decoy scores, and the proposed formulation (*LPGF* model) leads straightforwardly to the calculation of a protein probability. Note also that integration of peptide PEP values into a protein-level score using either their product or the Fisher method (which is conceptually equivalent to the *LPGS* model) has been shown not to improve the results obtained using the best-scoring peptide for each protein[24] and hence did not resolve the peptide-to-protein paradox.

In conclusion, by providing improved sensitivity and a statistically coherent approach, our protein probability model resolves for the first time the peptide-to-protein paradox and offers the scientific community an algorithm that can be easily integrated in protein identification workflows for the analysis of high-throughput proteomics data obtained by mass spectrometry.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.9b00819.

Figure S1, distribution of different decoy protein scores derived from Comet after searching three tissues of the Human Protein Map as separated test data sets; Figure S2, distribution of different decoy protein scores derived from MSFragger after searching three tissues of the Human Protein Map as separated test data sets; Figure S3, number of identified genes as a function of the FDR threshold for different protein score types derived from Comet after searching three tissues from the Human Protein Map; Figure S4, number of identified genes as a function of the FDR threshold for different protein score types derived from MSFragger after searching three tissues from the Human Protein Map; Figure S5, comparison of the number of identified genes as a function of the FDR threshold when using parametric

https://dx.doi.org/10.1021/acs.jproteome.9b00819
J. Proteome Res. 2020, 19, 1285−1297

peptide scores; Figure S6, number of identified genes as a function of the FDR threshold for different protein identification workflows using as separated tests three tissues from the Human Protein Map searched with Comet; Figure S7, number of identified genes as a function of the FDR threshold for different protein identification workflows using as separated tests three tissues from the Human Protein Map searched with MSFragger; Figure S8, Venn diagrams with the number of identified genes using different protein identification workflows in three tissues of the Human Protein Map searched with Comet; Figure S9, Venn diagrams with the number of identified genes using different protein identification workflows in three tissues of the Human Protein Map searched with MSFragger; Figure S10, comparison of target versus decoy peptides for each gene identified exclusively by any of the three different protein identification workflows discussed using Comet; Figure S11, comparison of target versus decoy peptides for each gene identified exclusively by any of the three different protein identification workflows discussed using MSFragger; Table S1, sequences of the best three peptides for top-scoring decoy proteins using *LPGS* as the protein probability after searching with MSFragger the `Adult_Heart` tissue of the HPM (PDF)

AUTHOR INFORMATION

Corresponding Authors

Gorka Prieto − *Department of Communications Engineering, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain*; orcid.org/0000-0002-6433-8452; Phone: +34 94 601 3994; Email: gorka.prieto@ehu.eus

Jesús Vázquez − *Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), 28049 Madrid, Spain*; Email: jesus.vazquez@cnic.es

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.9b00819

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Simon Bartlett (CNIC) for English editing. This study was supported by competitive grants from the Spanish Ministry of Economy and Competitiveness (MINECO) (BIO2015-67580-P) through the Carlos III Institute of Health-Fondo de Investigación Sanitaria (ISCIII-SGEFI/ERDF, ProteoRed) (PRB2 IPT13/0001 and PRB3 IPT17/0019), the Fundació La Marató TV3 (grant 122/C/2015), and the "La Caixa" Banking Foundation (project code HR17-00247). The CNIC is supported by the Ministry of Economy, Industry and Competitiveness (MEIC) and the Pro-CNIC Foundation and is a Severo Ochoa Center of Excellence (MINECO award SEV-2015-0505).

REFERENCES

(1) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **2008**, *7*, 29−34.

(2) Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **2010**, *73*, 2092−2123.

(3) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4*, 207−214.

(4) Navarro, P.; Vázquez, J. A refined method to calculate false discovery rates for peptide identification using decoy databases. *J. Proteome Res.* **2009**, *8*, 1792−1796.

(5) Keich, U.; Kertesz-Farkas, A.; Noble, W. S. Improved false discovery rate estimation procedure for shotgun proteomics. *J. Proteome Res.* **2015**, *14*, 3148−3161.

(6) Reiter, L.; Claassen, M.; Schrimpf, S. P.; Jovanovic, M.; Schmidt, A.; Buhmann, J. M.; Hengartner, M. O.; Aebersold, R. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* **2009**, *8*, 2405−2417.

(7) Granholm, V.; Navarro, J. F.; Noble, W. S.; Käll, L. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. *J. Proteomics* **2013**, *80*, 123−131.

(8) Kim, M. S.; et al. A draft map of the human proteome. *Nature* **2014**, *509*, 575−581.

(9) Ezkurdia, I.; Vázquez, J.; Valencia, A.; Tress, M. Analyzing the first drafts of the human proteome. *J. Proteome Res.* **2014**, *13*, 3854−3855.

(10) Ezkurdia, I.; Calvo, E.; Del Pozo, A.; Vázquez, J.; Valencia, A.; Tress, M. L. The potential clinical impact of the release of two drafts of the human proteome. *Expert Rev. Proteomics* **2015**, *12*, 579−593.

(11) Farrah, T.; Deutsch, E. W.; Omenn, G. S.; Campbell, D. S.; Sun, Z.; Bletz, J. A.; Mallick, P.; Katz, J. E.; Malmström, J.; Ossola, R.; Watts, J. D.; Lin, B.; Zhang, H.; Moritz, R. L.; Aebersold, R. A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol. Cell. Proteomics* **2011**, *10*, No. M110.006353.

(12) Serang, O.; Noble, W. A review of statistical methods for protein identification using tandem mass spectrometry. *Stat. Interface* **2012**, *5*, 3.

(13) Ma, Z.-Q.; Dasari, S.; Chambers, M. C.; Litton, M. D.; Sobecki, S. M.; Zimmerman, L. J.; Halvey, P. J.; Schilling, B.; Drake, P. M.; Gibson, B. W.; Tabb, D. L. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.* **2009**, *8*, 3872−3881.

(14) Serang, O.; MacCoss, M. J.; Noble, W. S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J. Proteome Res.* **2010**, *9*, 5346−5357.

(15) Huang, T.; He, Z. A linear programming model for protein inference problem in shotgun proteomics. *Bioinformatics* **2012**, *28*, 2956−2962.

(16) Li, Y. F.; Arnold, R. J.; Radivojac, P.; Tang, H. Protein identification problem from a Bayesian point of view. *Stat. Interface* **2012**, *5*, 21.

(17) Alves, G.; Yu, Y.-K. Mass spectrometry-based protein identification with accurate statistical significance assignment. *Bioinformatics* **2014**, *31*, 699−706.

(18) Uszkoreit, J.; Maerkens, A.; Perez-Riverol, Y.; Meyer, H. E.; Marcus, K.; Stephan, C.; Kohlbacher, O.; Eisenacher, M. PIA: an intuitive protein inference engine with a web-based user interface. *J. Proteome Res.* **2015**, *14*, 2988−2997.

(19) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, 4646−4658.

(20) Matrix Science, Mascot database search: MS/MS Results Interpretation. 2019. http://www.matrixscience.com/help/interpretation_help.html#SCORING.

(21) 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*, 22−24.

(22) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367.

(23) Savitski, M. M.; Wilhelm, M.; Hahne, H.; Kuster, B.; Bantscheff, M. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol. Cell. Proteomics* **2015**, *14*, 2394−2404.

(24) The, M.; MacCoss, M. J.; Noble, W. S.; Käll, L. Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. *J. Am. Soc. Mass Spectrom.* **2016**, *27*, 1719−1727.
(25) Higdon, R.; Reiter, L.; Hather, G.; Haynes, W.; Kolker, N.; Stewart, E.; Bauman, A. T.; Picotti, P.; Schmidt, A.; van Belle, G.; Aebersold, R.; Kolker, E. IPM: An integrated protein model for false discovery rate estimation and identification in high-throughput proteomics. *J. Proteomics* **2011**, *75*, 116−121.
(26) Ramakrishnan, S. R.; Vogel, C.; Prince, J. T.; Wang, R.; Li, Z.; Penalva, L. O.; Myers, M.; Marcotte, E. M.; Miranker, D. P. Integrating shotgun proteomics and mRNA expression data to improve protein identification. *Bioinformatics* **2009**, *25*, 1397−1403.
(27) Audain, E.; Uszkoreit, J.; Sachsenberg, T.; Pfeuffer, J.; Liang, X.; Hermjakob, H.; Sanchez, A.; Eisenacher, M.; Reinert, K.; Tabb, D. L.; Kohlbacher, O.; Perez-Riverol, Y. In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *J. Proteomics* **2017**, *150*, 170−182.
(28) The, M.; Edfors, F.; Perez-Riverol, Y.; Payne, S. H.; Hoopmann, M. R.; Palmblad, M.; Forsström, B.; Käll, L. A protein standard that emulates homology for the characterization of protein inference algorithms. *J. Proteome Res.* **2018**, *17*, 1879−1886.
(29) Chambers, M.; et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30*, 918.
(30) Frankish, A.; et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **2018**, *47*, D766−D773.
(31) Wright, J. C.; Choudhary, J. S. DecoyPyrat: Fast Non-redundant Hybrid Decoy Sequence Generation for Large Scale Proteomics. *J. Proteomics Bioinf.* **2016**, *9*, 176.
(32) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **2005**, *10*, 1419−1440.
(33) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383−5392.
(34) Martínez-Bartolomé, S.; Navarro, P.; Martín-Maroto, F.; López-Ferrer, D.; Ramos-Fernández, A.; Villar, M.; García-Ruiz, J. P.; Vázquez, J. Properties of average score distributions of SEQUEST: the probability ratio method. *Mol. Cell. Proteomics* **2008**, *7*, 1135−1145.
(35) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923−925.
(36) Choi, H.; Ghosh, D.; Nesvizhskii, A. I. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J. Proteome Res.* **2008**, *7*, 286−292.
(37) Käll, L.; Storey, J. D.; Noble, W. S. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics* **2008**, *24*, i42−i48.
(38) Feng, J.; Naiman, D. Q.; Cooper, B. Probability model for assessing proteins assembled from peptide sequences inferred from tandem mass spectrometry data. *Anal. Chem.* **2007**, *79*, 3901−3911.
(39) The, M.; Tasnim, A.; Käll, L. How to talk about protein-level false discovery rates in shotgun proteomics. *Proteomics* **2016**, *16*, 2461−2469.
(40) Devroye, L. *Non-Uniform Random Variate Generation*; Springer-Verlag: New York, 1986; p 405.
(41) He, K.; Fu, Y.; Zeng, W.-F.; Luo, L.; Chi, H.; Liu, C.; Qing, L.-Y.; Sun, R.-X.; He, S.-M. A Theoretical Foundation of the Target-Decoy Search Strategy for False Discovery Rate Control in Proteomics. 2015, arXiv preprint arXiv:1501.00537. arXiv.org e-Print archive. https://arxiv.org/abs/1501.00537.
(42) Levitsky, L. I.; Ivanov, M. V.; Lobas, A. A.; Gorshkov, M. V. Unbiased false discovery rate estimation for shotgun proteomics based on the target-decoy approach. *J. Proteome Res.* **2017**, *16*, 393−397.

https://dx.doi.org/10.1021/acs.jproteome.9b00819
J. Proteome Res. 2020, 19, 1285−1297