PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

ProInfer: An interpretable protein inference tool leveraging on biological networks

Hui Peng[1,2], **Limsoon Wong**[3]*, **Wilson Wen Bin Goh**[1,2,4]*

1 Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore, **2** School of Biological Sciences, Nanyang Technological University, Singapore, Singapore, **3** Department of Computer Science, National University of Singapore, Singapore, Singapore, **4** Center for Biomedical Informatics, Nanyang Technological University, Singapore, Singapore

* wongls@comp.nus.edu.sg (LW); wilsongoh@ntu.edu.sg (WWBG)

Citation: Peng H, Wong L, Goh WWB (2023) ProInfer: An interpretable protein inference tool leveraging on biological networks. PLoS Comput Biol 19(3): e1010961. https://doi.org/10.1371/journal.pcbi.1010961

Editor: Arne Elofsson, Stockholm University: Stockholms Universitet, SWEDEN

Received: September 5, 2022

Accepted: February 20, 2023

Published: March 17, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1010961

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This work was supported by the Ministry of Education Singapore via an AcRF Tier 2 award

Abstract

In mass spectrometry (MS)-based proteomics, protein inference from identified peptides (protein fragments) is a critical step. We present ProInfer (Protein Inference), a novel protein assembly method that takes advantage of information in biological networks. ProInfer assists recovery of proteins supported only by ambiguous peptides (a peptide which maps to more than one candidate protein) and enhances the statistical confidence for proteins supported by both unique and ambiguous peptides. Consequently, ProInfer rescues weakly supported proteins thereby improving proteome coverage. Evaluated across THP1 cell line, lung cancer and RAW267.4 datasets, ProInfer always infers the most numbers of true positives, in comparison to mainstream protein inference tools Fido, EPIFANY and PIA. ProInfer is also adept at retrieving differentially expressed proteins, signifying its usefulness for functional analysis and phenotype profiling. Source codes of ProInfer are available at https://github.com/PennHui2016/ProInfer.

Author summary

Protein inference is a key step in proteomics data analysis. However, this procedure suffers from coverage issues due to high statistical stringency requirement and noise. Integration of prior knowledge to guide protein assembly can be a powerful approach. Hence, we developed a novel protein inference tool ProInfer that incorporates a length-adjusted and weighted-accumulated posterior error probability score with protein-complex networks. Compared against existing tools, ProInfer achieves the highest recall and F1 score in protein inference and also identifies novel differentially expressed proteins not reported by any other tool.

This is a *PLOS Computational Biology* Methods paper.

(MOE2019-T2-1-042 to WWBG and LW) and a AcRF Tier 1 award RT11/21 to WWBG). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Contemporary mass spectrometry (MS)-based proteomics is characterized by advanced high-throughput technologies for identifying proteins from complex mixtures [1–2]. MS proteomics measures the mass-to-charge (m/z) ion ratios, retention times, and ion intensities of precursor ions and peptide fragments [3–4]. A complex multi-step analytical procedure is then performed to reverse-engineer spectral information into peptide sequences (peptide-spectrum matching), followed by re-assembly of peptides to constituent proteins. The process of estimating the optimal set of proteins given acquired spectra is known as the protein inference problem [5–7]. Quantitative analysis is then performed to identify phenotype-specific proteins [8], obtain their function annotation [9] and determine potential applications in clinical [10] and medical settings [11] (**Fig 1A**).

In peptide-spectrum matching, a spectrum is matched against peptides in both a reference and a decoy protein sequence database, producing a score for each peptide-spectrum match (PSM) [12]. Given these PSMs, one can perform peptide identification, i.e., distinguishing correct PSM from incorrect ones, with well-known tools such as PeptideProphet [13] and Percolator [14]. Percolator was shown to identify more PSMs under similar q-value thresholds [14]. It can output q-values and posterior error probability scores (PEP) for identified peptides. Briefly, a q-value is the minimal false discovery rate (FDR) threshold needed for a positive identification of a given peptide while a PEP score indicates the probability of incorrectly identifying a non-existing peptide [15]. PEP scores are reported in popular protein inference tools such as Fido [16], PIA (Protein Inference Algorithms) [17] and EPIFANY [18].

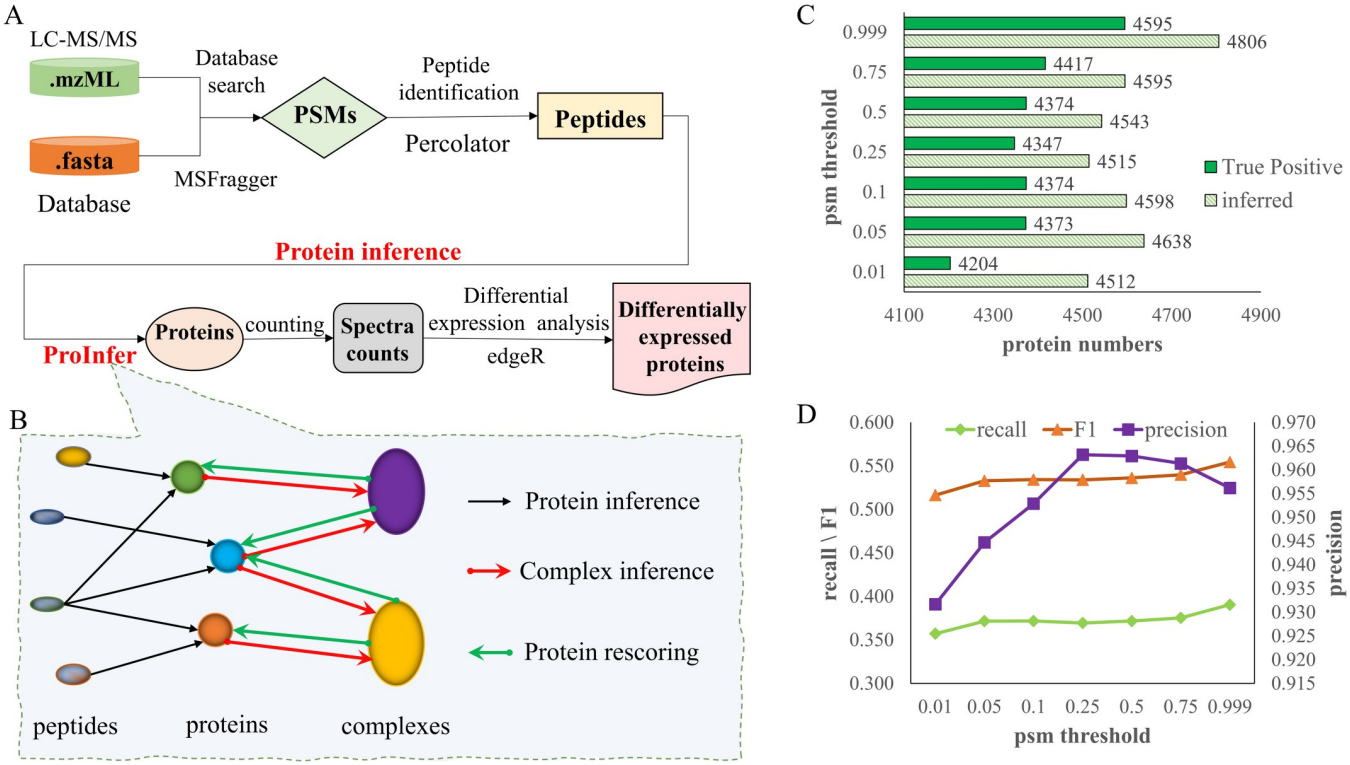


Fig 1. Workflow for protein inference and differential expression analysis with proteomic data and hyperparameter optimization results of proposed ProInfer. A shows a simple workflow of proteomics data-based protein inference and differential expression analysis. B gives a schema diagram of how ProInfer works. C presents the average inferred protein numbers and numbers of true positives found by ProInfer under different psm filtering thresholds (y-axis). The curves in D illustrate the performances of ProInfer indicated by recall, F1 (left-y axis) and precision (right-y axis) affected by various psm filtering thresholds (x-axis).

https://doi.org/10.1371/journal.pcbi.1010961.g001

Protein inference is concerned with protein identification from identified peptides. Protein inference is a demanding task and is affected by experimental/biological and analytical challenges. Example experimental/biological challenges include incomplete proteome coverage due to high dynamic range of protein abundances, limitations in digestion and protein separability under experimental running conditions, and detector sensitivity and resolution [18], whereas the algorithm design and assumption validity, parameter values, and sequence library completeness are example analytical challenges. A particularly cumbersome problem is dealing with peptide ambiguity where one peptide can be mapped to two or more proteins [19]. Proteome coverage issues can be eased by leveraging on alternative data acquisition strategies, e.g., parallel accumulation–serial fragmentation combined with data-independent acquisition (diaPASEF) [20], which increases precursor identification specificity. Peptide ambiguity problem is solved by either discarding ambiguous peptides (peptides which map to > 1 protein), e.g., Percolator [14]; or conducting network analysis on peptide-protein bipartite networks, e.g., EPIFANY [18]. Those proteins sharing the same constituent peptides may be reported as protein groups [17] in which case, we can be assured that at least 1 member in the protein group exists. Protein inference is a critical problem for proteomics, and so, many such methods have been developed (please see Huang et al [21] for details regarding some early methods). Newer (and popular) protein inference methods include Percolator [14], Fido [16], PIA [17], EPIFANY[18] and ProteinProphet [22].

Percolator [14] was originally designed for post-processing of peptide-spectrum matching results using semi-supervised learning. When protein inference is required, users may opt for a protein-level FDR threshold to output inferred proteins and their respective probabilities [23]. Perhaps to improve precision, Percolator does not consider ambiguous peptides, which may reduce proteome coverage. Fido [16] is a Bayesian probabilistic method for addressing ambiguous peptide problems and computing protein probabilities using graph-transforming algorithms. Fido creators claimed their method outperforms the heuristic posterior probability models based on expectation-maximization such as ProteinProphet [22]. PIA [17] is a consensus tool for integrating results of different search engines and different protein inference tools, e.g., ProteinProphet [22], Scaffold [19], and IDPicker [24]. PIA addresses ambiguous peptides by employing maximum parsimony principles and finding a minimal set of proteins explaining found peptides or PSMs [18]. EPIFANY [18] is a recently published protein inference tool that applies a loopy belief propagation algorithm (LBP) with convolution trees to process Bayesian networks. Via a peptide-protein bipartite graph, EPIFANY adopts convolution trees to propagate probabilities between peptides and proteins even for ambiguous peptides.

While these popular tools play significant roles in protein inference, there is room for improvement, especially in reported protein accuracy and proteome coverage. Current protein inference methods generally report proteins with lower q-values (q-values in this scenario, is a rank-based metric for comparing confidences of inferred proteins being present [15]). Consequently, proteins that are in fact present but have lower peptide support are ignored. Unless approaches exist to exhaustively mine low quality spectra for peptide spectra matches (PSMs), we do not expect peptide information supporting each protein to change drastically. Given such constraints, we believe most current tools are reliant only on direct peptide-to-protein information, to attain an upper limit (albeit incomplete) on the observable proteome [21]. To overcome these information barriers, we believe incorporating prior knowledge via data integration, e.g., drawing from independent data sources such as biological networks, is essential. Hence, we propose a protein-length adjusted posterior error probability accumulation method ***ProInfer*** (short for Protein Inference), which features a comprehensive yet simple rule towards protein scoring (including how it handles peptide ambiguity). To help users, ProInfer's methodology is easily understood; involving no complex calculations while possessing reasonable

assumptions. Additionally, ProInfer borrows similar principles from our missing protein prediction method PROTREC [25] that leverages on the phenomenon that proteins forming a stable protein complex (or constituting part of a tightly clustered network module) are more likely to be co-expressed [26]. Specifically, it incorporates protein complex information to rescue proteins with weak signals themselves but have neighbors with strong evidence. ProInfer achieves excellent performance in both protein inference and downstream quantitative analysis.

Materials and methods

ProInfer

A schematic of ProInfer is shown in **Fig 1B**. ProInfer comprises three stages: protein inference from the peptide-protein network, protein complex inference and protein rescoring from the protein-complex network. To define these terms: A peptide-protein network refers to the connections of identified peptides to their host proteins. This information is vital to demonstrating existence of these proteins in the sample (see an example of a peptide-protein network in left side of **Fig 1B**). Peptide-protein networks are useful for performing protein assembly from constituent peptides. Proteins work together as aggregates known as biological networks. These in turn, can be expressed as protein complexes or pathways, and is information rich. Thus, a protein-complex network (see the right-side bidirectional network in **Fig 1B**) is composed of proteins aggregating into protein complexes (defined as groups of polypeptide chains linked by noncovalent protein-protein interactions [27]). An arrow from a protein pointing towards a protein complex denotes the protein belongs to that complex. And thus, the protein's existence adds evidence supporting the presence of the complex in a tissue (i.e., protein complex inference process based on observed constituent proteins). Alternatively, an arrow from a protein complex to a protein propagates the existence probability of this complex to its constituent proteins, helping to re-evaluate our confidence in the protein's existence in a tissue (i.e., protein rescoring).

Protein inference from peptide-protein network

Our idea for protein inference from peptide-protein network stems originally from a need to address peptide ambiguity issues. Suppose an identified peptide Q from spectra is mappable to N proteins $\{P_1, P_2, \ldots, P_N\}$. We may reasonably assume that Q has an equal chance to come from N candidate proteins.

The false-reporting probability of using Q to support reporting P_1 is the chance that Q is an incorrect identification (with PEP score of pep_Q) plus the chance that Q is not incorrect but does not come from P_1 (i.e., it is from one of $P_2, \ldots, P_N$). This may be expressed as:

$$pep_Q + (1 - pep_Q) \cdot (N - 1)/N \tag{1}$$

where pep_Q is the PEP score of peptide Q and N is the number of mappable proteins.

Eq 1 can be rewritten as

$$1 - (1 - pep_Q) \cdot (1/N) \tag{2}$$

Here, the term $(1 - pep_Q) \cdot (1/N)$ means each candidate protein receives equal support, i.e., Q's posterior probability $(1 - pep_Q)$. In practice, Q should not arise equally from each candidate parent protein as "longer proteins are more likely to generate spurious matches than shorter ones" [28]. In other words, the support attributed by an ambiguous peptide towards the existence of a protein also depends on the protein's length. Thus, to express this idea, we normalize

the probability of Q from P_1 by accounting for protein length. The error of using Q to support P_1 can be modified to

$$1 - (1 - pep_Q) \cdot \phi \tag{3}$$

where ϕ is a length-based adjustment and is computed as:

$$\phi = \frac{\sum_{i=1}^{N} len(P_i)/(len(P_1))}{\sum_{h=1}^{N}(\sum_{i=1}^{N} len(P_i)/(len(P_h)))} \tag{4}$$

where $len(P_i)$ is a function computing the length of protein P_i.

Now we consider the problem of reporting a protein P_1 to be present given supporting peptides $\{Q_1, Q_2, \ldots Q_j, \ldots, Q_m\}$. If a peptide Q_j supports a protein P_1 uniquely, this peptide is considered "unique". If a peptide Q_j supports n > 1 proteins $\{P_1, P_2, \ldots P_i, \ldots, P_n\}$, this peptide is considered "ambiguous".

The error of peptide Q_j supporting P_1 can be context-driven (unique or ambiguous) as follows:

$$p(P_1 \textit{ is a false report}|Q_j) = \begin{cases} pep_{Q_j}, & Q_j \textit{ is a unique peptide} \\ 1 - (1 - pep_{Q_j}) \cdot \phi, & Q_j \textit{ is a ambiguous peptide} \end{cases} \tag{5}$$

where ϕ is computed as shown in Eq (4).

We assume that peptide Q_1 supporting P_1 is independent of other peptides $Q_{j,j\neq 1}$ supporting P_1. The total errors of using $\{Q_1, Q_2, \ldots Q_j, \ldots, Q_m\}$ to support P_1 to be present can be computed by:

$$p(P_1 \textit{ is a false report } |\{Q_1, Q_2, \ldots Q_j, \ldots, Q_m\}) = \prod_{j=1}^{m} p(P_1 \textit{ is a false report}|Q_j) \tag{6}$$

To calculate the FDR, we transform the cumulative PEP score from (6) (denoted as *accPEP*) to a confidence score:

$$S = -10 \cdot \log_{10}(accPEP + 1e - 14) \tag{7}$$

In (7), we spike a small value of 1e−14 to avoid errors where S becomes undefined (NaN) when accPEP is 0. We compute the FDR derived q-value for reported proteins in the same way as EPIFANY [18]. Firstly, the reported L proteins are ranked by confidence scores (S) in descending order, i.e., $\{P^{r1}, P^{r2}, \ldots, P^{rk}, \ldots, P^{rL}\}$. Then, the q-value of P^{rL} is the FDR with the threshold of its confidence score calculated by

$$qvalue(P^{rL}) = FDR(x = S(P^{rL})) = \frac{|\{y \geq x, y \in D\}| + 1}{|\{y \geq x, y \in T\}| + 1} \tag{8}$$

where x is the threshold, and $|\{y \geq x, y \in D\}|$ or $|\{y \geq x, y \in T\}|$ counts the number of decoy (D) or target proteins (T) with confidence scores no less than the threshold.

For $k \in [L-1, 1]$,

$$qvalue(P^{rk}) = min\{FDR(S(P^{rk})), qvalue(P^{r(k+1)})\} \tag{9}$$

Given these q-values, we can select an appropriate FDR, e.g., 1%, to report the proteins that qualify under this threshold. At FDR 1%, we expect 1 decoy protein (False Positive) per 100 correct target proteins (True Positive).

Protein complex inference and protein rescoring with protein-complex network

Given only PSM information, we may easily reach the upper boundary of reportable proteins no matter how good the protein inference tool is. This is because these protein inference methods are ultimately dependent on spectra completeness and quality [21]. In a typical experiment setting, due to limitations in instrument sensitivity, protein abundance and protein sequence uniqueness, some proteins are only supported by weak signals (e.g., few supporting peptides and/or low confidence peptides), and thus, are difficult to observe. To rescue such proteins, and improve proteome coverage, we may "borrow" information from other important modalities (protein-protein interaction network [29], gene expression profiles [30], etc.) Biological network information encapsulated in the form of protein complexes is particularly valuable, possessing high biological information value [31], improving statistical reproducibility [32] and improving phenotype characterization [33]. Using protein complexes, we developed PROTREC [25], a tool for missing protein recovery, which outperforms other missing protein prediction methods. We hypothesize that protein complexes can also be useful for improving protein inference from peptide information. To test this idea, we use protein complex information in ProInfer.

Suppose ProInfer (refers to the part described in the above section) outputs L candidate proteins and their accPEP scores and q-values denoted by:

$$Pros = \{(P_i, accPEP_i, q_i)|i \in [1, L]\} \tag{10}$$

We collected C reliable protein complexes and generated C decoy protein complexes by replacing the protein ids (e.g., sp|P41182) in each real complex with corresponding decoy protein ids (e.g., DECOY_sp|P41182). The complexes are denoted by:

$$Cpxs = \{(c_j, cP^j)|j \in [1, 2C]\} \tag{11}$$

where c_j is the jth known protein complex, and it contains X constituent proteins denoted by $cP^j = \{cP^j_1, cP^j_2, \ldots, cP^j_X\}$.

The following procedures describe the protein complex inference and integration of protein complex information in *Cpxs* with ProInfer's outputs *Pros*:

Step1. Initialization. We initialize the probability of protein P_i being present in the sample as $p(Pi) = 1 - accPEP_i$. For a given FDR f, ProInfer reports *num0* numbers of target proteins.

Step2. Protein complex inference. Calculate the probability of a protein complex c_j being present in the sample as the maximum probability of the subset proteins in this complex and in *Pros*, denoted by:

$$p(c_j) = \max\{1 - accPEP_a|a \in [1, z]\} \tag{12}$$

where $accPEP_a$ is the accPEP score of the ath protein in the subset $cP^j \cap Pros$.

Step3. Calculate the probability of protein P_i in *Pros* being present in the sample according to protein complex information. Let $\{c_1, c_2, \ldots, c_Q\}$ be the Q complexes containing P_i, then the probability of P_i is computed as the maximum probability of $\{c_1, c_2, \ldots, c_Q\}$:

$$p\_cpx(P_i) = \max(p(c_1), p(c_2), \ldots, p(c_Q)) \tag{13}$$

If no protein complex contains P_i, then $p\_cpx(P_i) = 0$.

Step4. Update the probability of P_i being present in the sample. By comparing $p(P_i)$ with $p\_cpx(P_i)$, we update the probability of P_i being present in the sample as:

$$p(P_i)' = \max(p(P_i), p\_cpx(P_i)) \tag{14}$$

Step5. Check whether we can now report new target proteins with given FDR *f*. From Step4, we get $accPEP_i\prime = 1 - p(P_i)'$. We transform $accPEP_i'$ to its confidence score via above Formula (7). Then, via Formulas (8) and (9), we compute the q-value of P_i as q_i'. *Pros* is updated as following formula:

$$Pros = \{(P_i, accPEP_i', q_i')|i \in [1, L]\} \tag{15}$$

With FDR *f*, *num* target proteins are reported. If $num > num0$, then turn to **Step1**, otherwise output *Pros* and stop.

Unlike PROTREC, ProInfer does not compute the probability of a protein complex being present in a biological sample based on the weighted probability of all observed constituent proteins [25]. ProInfer's approach is based on calculating the maximum of constituent proteins posterior probability (PP) expressed as PP = 1—*accPEP*. We used protein complexes downloaded from CORUM 3.0 [34]. We compute the probability of a protein complex being present in the sample as the maximum probability of its proteins' posterior probability measured by 1—*accPEP*. Then, we update the posterior probability of a protein being present with the higher value compared between the protein's original PP value and its parent complexes' posterior probabilities, i.e., max(original PP, complexes' PPs). For decoy proteins, PP derived from corresponding decoy complexes will be used (similarly, calculated by max(original PP, complexes' PPs)). A decoy complex is constructed by replacing target proteins in its twin true complex with decoy ones. The introduction of decoy complexes is to make the inference of both target proteins and decoy proteins in a similar way to avoid bias in estimating FDR. This propagation procedure is iterated until no additional target proteins are reported under a given FDR.

Hyperparameter optimization and datasets

Protein inference is conducted following peptide identification, where PSMs are evaluated and then filtered by a given PEP threshold. Retained PSMs are regarded as reliable. A strict PEP threshold retains high confidence PSMs but also produces many false negatives. Conversely, relaxed PEP thresholds alleviate the false negative problem but at the cost of introducing more false positives. Different tools adopt different strategies for threshold optimization.

Tools such as EPIFANY, Fido and PIA have some tool-specific hyperparameters to be tuned, e.g., greedy group resolution for EPIFANY [18] and Fido [16], regularization type for EPIFANY [18], and input score type and scoring method for PIA [17]. For a fair comparison, we optimized their hyperparameters with the same Hela cell line dataset initially. The Hela cell line, derived from cervical cancer cells, is the oldest and most used human cell line [35]; it is well-documented and widely applied in biochemical, biological, and medical experiments [36]. 4-replicates Hela cell line raw data of Mehta et al [37] were downloaded from PRIDE [38] with Project ID PXD022448 (see **Table 1**).

For performance benchmarking, the lung cancer data (lung cancer) of Li et al [39] (PXD000853), the THP1 cell line and RAW264.7 mouse macrophage cell line of Li et al [40] (PXD019800) were used. THP1 is a human leukemia monocytic cell line and is commonly studied for estimating modulation of monocyte and macrophage activities [41]. RAW264.7 is a mouse leukemia cell line of monocyte macrophage, where it has been extensively used to

Table 1. Summary of datasets used for hyperparameter optimization and performance evaluation.

Dataset	Condition	Replicates/Samples	PRIDE ID	Purpose
Hela	-	DDA1,DDA2,DDA3,DDA3	PXD022448	hyperparameter optimization
THP1	M0	M0_1, M0_2, M0_3	PXD019800	Performance benchmarking
THP1	M1	M1_1, M1_2, M1_3	PXD019800	Performance benchmarking
RAW264.7	M0	M0_1, M0_2, M0_3	PXD019800	Performance benchmarking
RAW264.7	M1	M1_1, M1_2, M1_3	PXD019800	Performance benchmarking
lung cancer	Normal	N24742,N31945,N32813_r,N35480	PXD000853	Performance benchmarking
lung cancer	Patient	T24742,T31945,T32813_r,T35480	PXD000853	Performance benchmarking

https://doi.org/10.1371/journal.pcbi.1010961.t001

study macrophage functions, mechanisms, and signaling pathways [37,42]. The lung cancer data was adopted to discover new anticancer therapeutic targets [39] (see **Table 1**).

Proteomic dataset processing

Raw data were converted to.mzML format with MSConvert [43] and processed as per flowchart in **Fig 1A**. MSFragger-3.4 [44] was used to conduct database search. A target-decoy searching strategy [12] was adopted where the protein database contains human reviewed proteins from UniProt [45] (UP000005640, downloaded in 5/5/2022) and known contaminants from the common Repository of Adventitious Proteins (cRAP, https://www.thegpm.org/crap/, added by FragPipe-17.1 [44], https://fragpipe.nesvilab.org/) database together with the decoy proteins generated by sequence reversal. Search parameters are as follows: precursor mass tolerance (PMT) of 20ppm, fragment mass tolerance (FMT) of 0.05Da, and peptide length of 7 to 50 (remaining parameters are left as default). Prior to inputting to different protein inference tools, we performed peptide indexing and feature extraction with OpenMS (version 2.7.0) [46] for Percolator, which was then used to conduct peptide identification with PEP scores computed for each PSM. With these scored PSMs, hyperparameters were optimized with grid search: PSM filtering thresholds (PSMs with PEP scores bigger than the threshold are dropped) were ranged among [0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.999], while for other tool-specific hyperparameters, all possible values are tested.

Validation and performance evaluation

The protein-expression tissue database, Human Protein Atlas (HPA) [47], is used for protein validation. HPA (https://www.proteinatlas.org/) is a manually curated database that collects human proteins in cells, tissues, and organs via integrating various omics technologies such as antibody-based imaging, MS-based proteomics, transcriptomics, and systems biology [47]. Positive proteins for Hela cell line were downloaded (data were downloaded from https://www.proteinatlas.org/search/NOT+celline_category_rna%3AHeLa%3BNot+detected). Proteins having UniProt ids [45] were retained. For Hela cell line, there are 11806 validated proteins (See detail protein list in **S4 Table**). We label the proteins predicted by different protein inference tools, e.g., our ProInfer, EPIFANY, etc., and validated by the Human Protein Atlas as true positives, otherwise they are considered false positives. We calculate several metrics for evaluating competing tools and optimizing their hyperparameters including inferred protein numbers, numbers of true positives, recall, precision and F1 score. Recall, precision and F1 score are given by:

$$\text{recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{16}$$

$$\text{precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{17}$$

$$\text{F1} = 2 \cdot \text{recall} \cdot \text{precision}/(\text{recall} + \text{precision}) \tag{18}$$

where, TP (true positive) refers to a protein reported by an inference tool, e.g., our ProInfer and validated by HPA, FN (false negative) means a protein in HPA but has not been reported by a given protein inference tool, and FP is a protein not in HPA but has been reported.

The final performance evaluation of a tool is determined by the average performance across the 4-replicates' Hela cell line data and varying protein reporting FDR among 0.005, 0.01, 0.025 and 0.05. F1 score is one of the most widely used metrics for measuring performance of a classifier and is used to select optimal hyperparameters.

Downstream differential expression analysis

A simple differential expression analysis workflow is shown in Fig 1A. This was used to benchmark protein inference tools by evaluating their ability to identify differentially expressed proteins. Taking the lung cancer data as an example, proteins differentially expressed in patient samples (4 biological replicates) when compared against normal samples, are expected to be identified. In each of the 8 samples, proteins are inferred by different tools with their matched spectra numbers being counted. An expression matrix for this lung cancer data is formed by integrating the 8 samples' protein (final protein list is a union of all 8 samples) spectra counts where missing proteins are filled with counts of 0. We used this expression matrix as input to edgeR, a widely used differential expression analysis tool [48], to identify differentially expressed proteins. Those proteins with less than 2 non-missing values in samples of each condition are dropped. We define a differentially expressed protein (DEP) as the protein with absolute log2FC ≥ 0.585 (FC means fold change, equals to $|\text{FC}| \geq 1.5$) and Benjamini & Hochberg adjusted p-value (adj.pvalue) ≤ 0.05 [49].

Results

Summary of optimized hyperparameters with Hela cell line data

We identified the optimal running conditions (or settings) for each tool (ProInfer, EPIFANY, Fido, Percolator and PIA) given data of a particular nature. This would allow us to compare the best outcomes possible for each tool.

For each method, during hyperparameter optimization, we ranked their hyperparameters (or combinations of hyperparameters if more than one hyperparameter needs to be tuned) by corresponding average F1 scores across the 4-replicates Hela cell line data. And returned the best hyperparameter/combination.

Fig 1C and 1D shows the hyperparameter optimization results of ProInfer. Only PSM filtering threshold (PEP) needs to be selected for ProInfer. Ostensibly, ProInfer has good resistance to low reliability PSMs: When a loose PSM filtering threshold is used, ProInfer achieves higher recall and F1 score with small decrease in precision. For example, when we set PSM filtering threshold as PEP ≤ 0.999, we increase coverage by ~400 more correct proteins in Hela cell line than filtering with PEP ≤ 0.01 (Fig 1C, 4595–4204). However, this comes at the cost of introducing ~300 more false positive proteins (Fig 1C, 4806–4512). When filtering with PSM PEP ≤ 0.25, the highest precision of 0.963 is obtained, which is 0.007 bigger than PEP ≤ 0.999 (0.956), but 0.02 less (Fig 1D, 0.554–0.534) for recall and 0.021 less (Fig 1D, 0.390–0.369) for F1 score. Notably, stricter filtering condition also eradicated many target peptides, resulting in

the loss of signals that are potentially rescuable via integration with network information. Hence, a looser PSM filtering threshold for ProInfer is preferred. Accordingly, PSM PEP $\leq$ 0.999 is set as default hyperparameter for ProInfer in following tests.

We tuned hyperparameters for EPIFANY, Fido, Percolator and PIA accordingly. EPIFANY works best when setting PSM PEP $\leq$ 0.05 and parameter greedy_group_resolution to be "remove_associations_only" and without regularize. For Fido, its optimal hyperparameters are: PSM PEP $\leq$ 0.1 and greedy_group_resolution setting as "true". Like ProInfer, Percolator also works best with PSM PEP $\leq$ 0.999. For PIA, inference method of Spectrum Extractor, multiplicative scoring method and PSM PEP score are chosen, and PEP $\leq$ 0.999 is used. More details are found in **S2 Table**.

Benchmarking competing tools with THP1 cell line data

We used the technical replicates of M0 THP1 (**Table 1**) for conducting independent benchmarking. Positive proteins were obtained from HPA (11584 proteins, see **S4 Table**, data can be downloaded via https://www.proteinatlas.org/search/NOT+celline_category_rna%3ATHP-1%3BNot+detected). For each tool, optimal hyperparameters were determined as described above. The average performance across the 3 M0 THP1 replicates were used. In addition, an alternative positive protein set generated by filtering out proteins in HPA but without protein level evidence (with 11483 proteins) were also tested, minor performance differences were obtained, see **S7 Table** for more details.

In **Fig 2A**, we showed the proportions of true positives (deep colors) against false positives (light colors). Protein reporting FDRs were set as 0.005, 0.01, 0.025 and 0.05 respectively. Regardless of FDR threshold, ProInfer reports the most numbers of true positives (albeit, with correspondingly more false positives as well). For instance, given FDR 0.01, ProInfer reports 5390 proteins in total, of which, 5041 are true positives. Compared against Percolator, 640 (**Fig 2A**, 5041–4401) more true positives were reported with just 230 (**Fig 2A**, 349–119) more false positives, achieving a 1:2.78 ratio of false positives:true positives gain (230:640). Similarly, the false positives:true positives gains comparing to EPIFANY, Fido and PIA are 1:2.90, 1:6.1 and 1:9.37. In addition, from **Fig 2A**, even with looser FDR thresholds, e.g., 0.025 and 0.05, ProInfer reports more true positives without incurring great changes to the presence of false positives (about 10 more false positives comparing to FDR 0.005 or 0.01). From **Fig 2B**, ProInfer produces stable precisions while other tools acquire lower precisions as FDRs relaxes. In **Fig 2C and 2D**, the line plots show that ProInfer always achieves the highest recall and F1 score. Notably, all methods obtain better recalls and F1 scores when FDR relaxes from 0.005 to 0.05. Amongst the methods, Percolator always gets highest precisions but lowest recalls and F1 scores.

In **Fig 2E**, we also used an upset plot to investigate overlaps among different methods' reported proteins at FDR 0.01 in an example replicate of THP1 M0 cell line (refers to M0_1). We identified 4093 proteins commonly reported by all 5 tools. The overlap amongst competing tools is deep, making up at least 75% of total reported proteins (from 77% for ProInfer to 92% for Percolator). Almost all EPIFANY reported proteins are also reported by at least one tool. Notably, each of the remaining four tools can identify some proteins missed by others. Thus, we added an additional Venn diagram on top of the upset plot to show the reliability of these tool-unique proteins (**Fig 2E** inset). Amongst the 4093 commonly reported proteins, 4010 (98%) is validated by Human Protein Atlas (HPA detectable). Importantly, ProInfer identified the biggest number (1014) of uniquely reported proteins of which 776 out of 1014 (76.5%) were validated. PIA uniquely reported just 166 proteins, with 57.8% (96 out of 166) validated. Fido uniquely reported 83 proteins with less than half (38 out of 83) validated. For

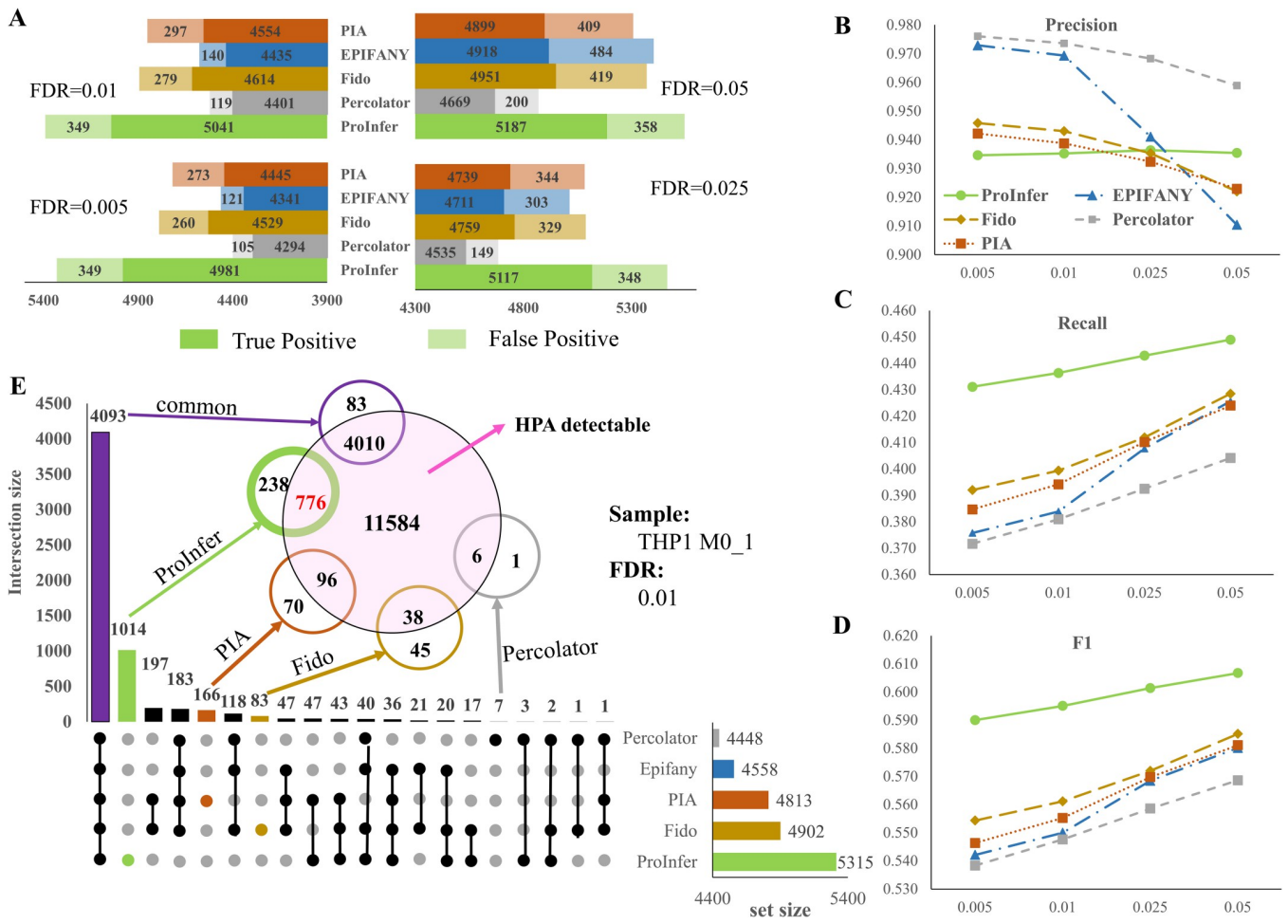


Fig 2. Performance evaluation of various tools tested on 3 replicates of human THP1 dataset. A shows average numbers of true positives (in deep colors) and false positives (in light colors) reported in 3 replicates of THP1 sample with FDR of 0.005, 0.01, 0.025 and 0.05 respectively with hyperparameters optimized with Hela data. **B, C & D** show the changes of precision, recall and F1 values of competing tools under different protein reporting FDRs. **E** is a Venn diagram revealing the overlap of proteins reported by different tools and validation status of proteins that reported only by a specific tool.

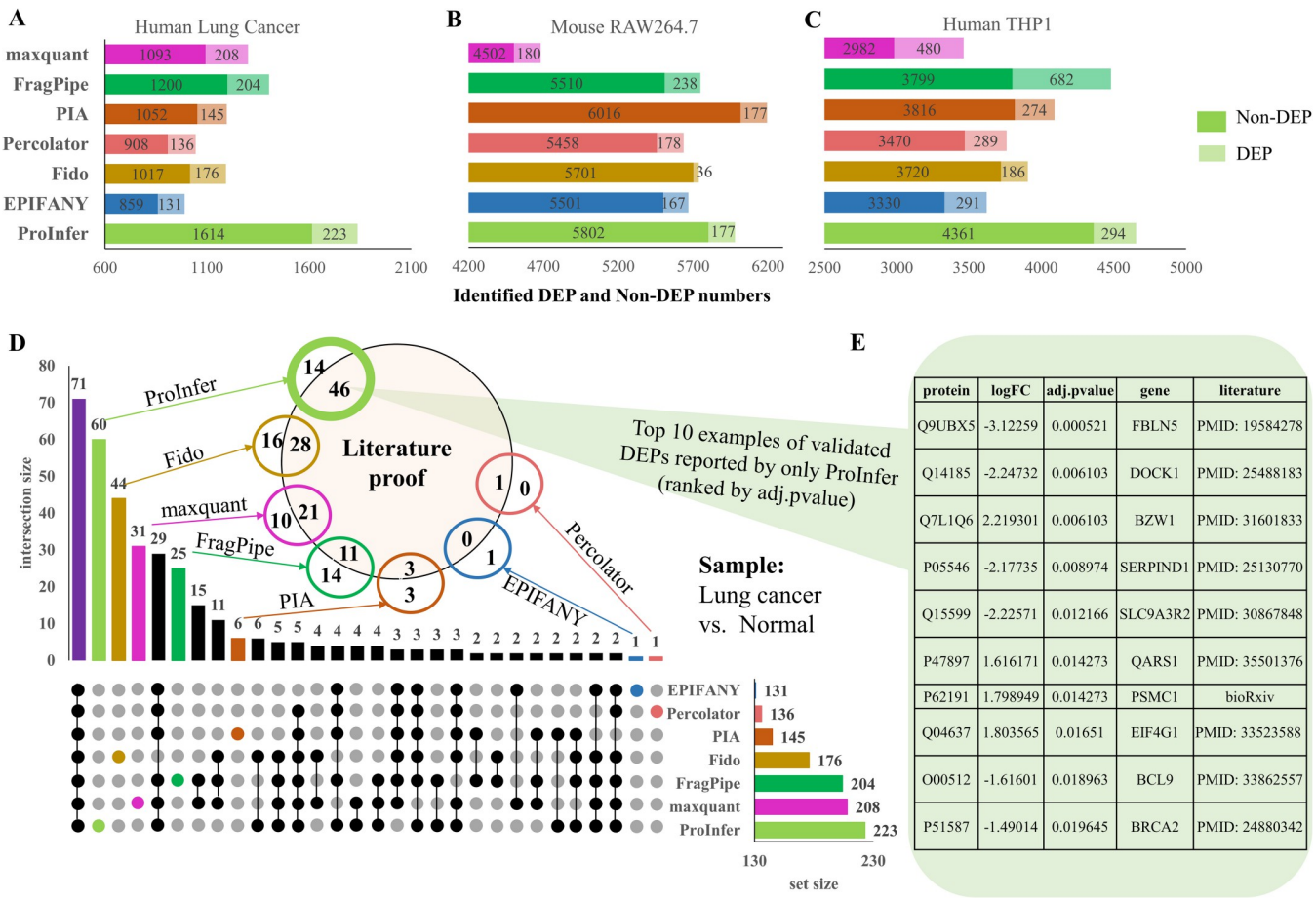
https://doi.org/10.1371/journal.pcbi.1010961.g002

Percolator, the validation rate of its uniquely reported proteins is 85.7% (6/7), however the list size is very small in comparison. Hence, the superior performance of ProInfer to report many reliable proteins, not found by any other tool, makes it a promising tool to find novel protein biomarkers.

Evaluations of protein inference tools in differential expression analysis

Following protein inference, downstream quantitative analysis is a key step for proteomics data analysis. To investigate how protein inference tools affect differential expression analysis, we used 3 published datasets for testing. FragPipe [44] and maxquant [50] are two popular platforms for proteome quantification, thus are tested for comparison.

Bar plots in Fig 3A–3C present found DEP (in light colors) and non-DEP (in deep colors) numbers in human lung cancer, mouse RAW264.7 and human THP1 data by five protein inference tools, FragPipe and maxquant-based workflows. Given lung cancer and THP1 data, ProInfer reports the most proteins (including DEPs and Non-DEPs) comparing to other four protein inference tools and FragPipe or maxquant. In RAW264.7 data, PIA inferred the most



protein	logFC	adj.pvalue	gene	literature
Q9UBX5	-3.12259	0.000521	FBLN5	PMID: 19584278
Q14185	-2.24732	0.006103	DOCK1	PMID: 25488183
Q7L1Q6	2.219301	0.006103	BZW1	PMID: 31601833
P05546	-2.17735	0.008974	SERPIND1	PMID: 25130770
Q15599	-2.22571	0.012166	SLC9A3R2	PMID: 30867848
P47897	1.616171	0.014273	QARS1	PMID: 35501376
P62191	1.798949	0.014273	PSMC1	bioRxiv
Q04637	1.803565	0.01651	EIF4G1	PMID: 33523588
O00512	-1.61601	0.018963	BCL9	PMID: 33862557
P51587	-1.49014	0.019645	BRCA2	PMID: 24880342

Fig 3. Evaluation of methods on differential expression analysis. A, B & C show the numbers of differentially expressed proteins (DEPs) and non-differentially expressed proteins (Non-DEPs) found by different protein inference tool-based, FragPipe and maxquant-based workflows from lung cancer, RAW264.7 and THP1 cell line data. D shows overlapping of DEPs reported by different workflows from lung cancer data and the validation status of those uniquely found DEPs. E gives the top10 example validation proofs of DEPs uniquely found by ProInfer.

https://doi.org/10.1371/journal.pcbi.1010961.g003

proteins compared to other tools, however, ProInfer still found more proteins than other workflows except for PIA. For DEPs, FragPipe and maxquant always report more DEPs than these five protein inference tools except for the lung cancer data, where ProInfer identified about 20 more DEPs. This may be due to both FragPipe and maxquant adopting the match-between-runs (MBR) method to mitigate the missing value problem [51], where smaller missing rates will be obtained, e.g., about 15% for both FragPipe (15.11%) and maxquant (15.41%), but more than 19% for ProInfer (19.43%) and PIA (22.77%) averagely in THP1 M0 and M1 samples. ProInfer performs stably in protein reporting and DEP identification compared to other methods given the three tests.

We are interested in how DEPs identified by different workflows overlap. In Fig 3D, an upset plot displays the intersections of DEPs found in lung cancer data by 7 workflows. 71 DEPs were reported by all 7 workflows, comprising 54% of all DEPs found by EPIFANY-based workflow (highest) and 32% of ProInfer-based workflow (lowest). All 7 workflows reported some unique DEPs, which is interesting. These unique DEPs may be valuable, e.g., novel biomarkers or drug targets. We drew an additional Venn diagram (Fig 3D inset) to check the validation status of these unique DEPs, where literature proofs were searched to prove they really associate with lung cancer. We searched literature by key words of given

protein and lung cancer at first. Then we read the literature to check whether they reported any correlations between the protein and lung cancer. ProInfer found the largest number of uniquely reported DEPs, where 46 out of 60 of these DEPs can be validated by literature. Similarly, 44 DEPs were uniquely reported by Fido and 28 of them can be proved to play significant roles in lung cancer. For FragPipe and maxquant, 31 and 25 unique DEPs were confirmed and more than a half of them are supported by literature. In comparison, EPIFANY, PIA and percolator reported few unique DEPs. We listed the top 10 DEPs unique to ProInfer ranked by their adj.pvalues in **Fig 3E**. We can find published papers or recent preprints to confirm that these DEPs are lung cancer related. For example, the protein Fibulin-5 (Uniport id: Q9UBX5) is a product of gene FBLN5, which was reported to suppress lung cancer invasion by inhibiting matrix metalloproteinase-7 expression [52]. Pan et al. found that Dedicator of cytokinesis protein 1 (Uniport id: Q14185, product of gene DOCK1) plays significant role in cell migration, Akt expression, and vimentin phosphorylation and it's a drug target for lung cancer [53]. However, these two important DEPs were missed by all other tools. More details about the literature proofs for these uniquely reported DEPs can be found in **S3 Table**. In addition, those unique DEPs with no existing evidence may be novel lung cancer related proteins. These examples show that ProInfer improves protein inference and DEP identification. It is thus useful for biomarker or drug target identification.

Discussion

Most protein assembly methods have limited coverage of underlying proteomes

In HPA, there are 11584 confirmed proteins in THP1. Most protein inference methods except ProInfer, identified fewer than 5000 true positives even at a loose protein FDR of 0.05 (**Fig 2A**). Percolator has the worst proteome coverage. This may be due to elimination of ambiguous peptides alongside a simple approach towards protein inference. Though, both ProInfer and other protein inference tools such as Percolator apply a loose PSM filtering threshold, i.e., PEP $\leq$ 0.999, ProInfer always achieves excellent performance in protein inference, where highest recalls and F1 scores are always obtained. This may means dropping low confidence peptides (including ambiguous peptides) too early adversely impact proteome coverage. Using biological networks, ProInfer is a successful method that can make good use of peptides with weaker signals (or with lower confidence that may be dropped by stricter filtering conditions) to achieve good proteome coverage.

While ProInfer dominates in our benchmarks, it does have drawbacks

ProInfer works well on protein inference, especially when a looser peptide filtering criterion, e.g., PSM PEP $\leq$ 0.999, was applied. A strict filtering criterion may stave off more decoy peptides, which benefits some tools e.g., EPIFANY and Fido (see **Results**). However, such conservatism also results in widespread loss of informative target peptides, reducing proteome coverage. To manage noise from looser criteria, ProInfer integrates biological network information (e.g., protein complexes) to make good use of those peptides possessing relatively lower confidences to rescue more target proteins. In our evaluations, this strategy has proven effective in identifying more true positives than existing tools.

However, ProInfer has several drawbacks:

Firstly, ProInfer cannot manifest its full potential should biological networks be inapplicable or unavailable in the analytical context (e.g., when we don't have enough reliable protein complexes or there is no complex network formable in the sample). In the current version of

CORUM 3.0, there are 2916 curated human protein complexes, which is quite small, and does not account for all possible networks and complexes partook by all human proteins. The lack of a tissue-specific complexome database is a further limiting factor that we hope can be overcome eventually. Moreover, there are limited number of other species with well-characterized and extensive protein complex lists in this database; thus, unless homology mapping is an option, ProInfer may not work well on samples from other species (e.g., mouse, where PIA reports more proteins from mouse RAW264.7 cell line data).

Secondly, we see potential for further optimization: The parameters may yet be further explored for ProInfer. For example, previously in a related study, we only used the complexes with size ≥ 5 to reduce instability issues [25]. Here, all curated complexes were used; if a threshold of 5 were used, many complexes would be unavailable, resulting in loss of many weak signal proteins. In future optimizations, we may study the impact of protein complex filtering on ProInfer performance. Moreover, during the calculation of protein complex existence probabilities, only a subset of proteins in a complex and in the candidate protein list ($cP^j \cap Pros$) are considered. The low coverages of complexes may cause overestimation of the probabilities of complexes being present. However, if we filter out these low coverage proteins, then only a few complexes are usable, thus the performance of ProInfer also decreases. Other settings to be tuned includes how to better determine the probability of a complex to be present from its proteins and the signal propagation approach from complexes back to other same-complex proteins. Here, we simply assume the probability of a complex to be present equals to the **maximum** posterior probability of its proteins while the probability of a protein to be present from the complex side is measured as the **maximum** probability of all complexes containing it. In **S5 Table**, we tested setting a complex's probability to be present as the **mean** posterior probability of its proteins. However smaller F1 scores are always achieved especially when a higher psm threshold is configured. Though using mean helps reduce false positive rates (within 1%), much more true positives are also dropped (~10%). Using maximum is currently an optimal selection, more advanced methods that help reduce false positives but keep true positives could be tried, e.g., calculate a prior probability of a complex to be present with enrichment test or our previous weighted probability method [25].

Thirdly, for paralogous proteins that share the same peptides and can participate as mutually exclusive partners in protein complexes [54], ProInfer may possibly infer them as either simultaneously present or absent. This is because ProInfer is dependent on prior knowledge captured in the protein complex databases. It is possible to extend ProInfer by enriching protein complex data with information on gene expression and paralogs. This may reduce potential false positives.

Future work

Our future work will focus on three aspects. Firstly, we may also try to incorporate tissue-specific expression gene information, e.g., from database TissGDB [55] and housekeeping gene information, e.g., from HRT Atlas [56], to help the identification of proteins with higher confidence of existence based on their biological functions. Such proteins can be accorded higher confidence scores even if their observable peptides present with low signals. Secondly, to cater for big data, we may implement ProInfer in more efficient programming languages, e.g., Scala [57] or C [58]. Finally, while we have evaluated across selective yet high-quality data, there are many new technological advances. Hence, we may further evaluate ProInfer on exciting new data such as single-cell proteomics [59] and spatial proteomics [60].

Conclusion

We propose a novel biological network-guided method ProInfer for performing protein inference. ProInfer maximizes use of peptide information (including ambiguous peptides) via a simple yet logical assignment rule. More importantly, biological networks, in the form of protein complexes, is integrated with ProInfer to rescue proteins with weak signals. In our evaluations, ProInfer is robust, and stable even across a wide range of conditions. This is in stark contrast to most other protein assembly tools which are sensitive to adjustments of filtering parameters (especially important since the optimal cutoff is often unknown). Critically, ProInfer can identify large numbers of validated novel proteins not found by any other tool. In our evaluations, we find that these novel proteins are phenotype relevant. Thus, ProInfer is promising for functional profiling and discovering novel biomarkers or drug targets. Source codes of ProInfer are publicly accessible at https://github.com/PennHui2016/ProInfer.

Supporting information

S1 Table. Data for generating figures in the main text. Sheets in S1_Table.xlsx with names "Fig 1C" and "Fig 1D" show the tables containing the data for generating **Fig 1C** and **Fig 1D** in our main text. Similarly, sheets "Fig 2A" to "Fig 2E" and "Fig 3A" to "Fig 3E" show the corresponding data for generating our **Fig 2A** to **Fig 2E** and **Fig 3A** to **Fig 3E** in main text respectively.
(XLSX)

S2 Table. Results of parameter optimization for competing protein inference tools. The five Sheets with names "supp.tab1" to "supp.tab5" show the parameter optimization results for EPIFANY, Fido, Percolator, PIA and ProInfer.
(XLSX)

S3 Table. Literature proofs for validating uniquely found differentially expressed proteins based on different protein inference tools. The seven sheets with names "supp.tab1" to "supp.tab7" show the literature proofs for validating uniquely found differentially expressed proteins based on five protein inference tools ProInfer, EPIFANY, Fido, Percolator, PIA and two quantification analysis platforms FragPipe and maxquant.
(XLSX)

S4 Table. Proteins obtained from the Human Protein Atlas for protein inference validation. Sheets "Hela detectable Protein in HPA" and "THP1 detectable Protein in HPA" contain the lists of detectable proteins in Hela and THP1 from the Human Protein Atlas for validating protein inference performances.
(XLSX)

S5 Table. Comparison of methods "mean" and "max" for calculating protein complex confidence scores. Sheet 1 shows the comparison results of using "mean" and "max" to calculate protein complex confidence scores.
(XLSX)

S6 Table. Results of testing the robustness of proposed ProInfer by removing pre-inferred proteins. Sheet 1 shows the testing of robustness of ProInfer by removing 5~50% of pre-inferred proteins.
(XLSX)

S7 Table. The alternative validation data and the comparisons of validating inferred proteins with original validation data and the alternative validation data. Sheets

"HPA_Hela_protein_level" and "HPA_THP1_protein_level" give the two alternative validation data by removing proteins without protein level evidence. Sheet "parameter optimization" shows the detail parameter optimization results of different protein inference tools based on the alternative validation data in sheet "HPA_Hela_protein_level". The Sheets "Fig 1C", and "Fig 1D" show the minor changes in result data for generating our **Fig 1C** and **Fig 1D** in main text when using the alternative validation data in sheet "HPA_Hela_protein_level". Sheets with names "Fig 2A" to "Fig 2E" show the minor changes when using the alternative validation data in sheet "HPA_THP1_protein_level" for performance tests of different protein inference tools.
(XLSX)

Author Contributions

Conceptualization: Hui Peng.

Data curation: Hui Peng.

Formal analysis: Hui Peng.

Investigation: Hui Peng.

Methodology: Hui Peng.

Supervision: Limsoon Wong, Wilson Wen Bin Goh.

Writing – original draft: Hui Peng.

Writing – review & editing: Hui Peng, Limsoon Wong, Wilson Wen Bin Goh.

References

1. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003; 422: 198–207. https://doi.org/10.1038/nature01511 PMID: 12634793
2. Ong SE, Mann M. Mass spectrometry–based proteomics turns quantitative. Nat Chem Biol. 2005; 1: 252–262. https://doi.org/10.1038/nchembio736 PMID: 16408053
3. Jemal M. High-throughput quantitative bioanalysis by LC/MS/MS. Biomed Chromatogr. 2000; 14: 422–429. https://doi.org/10.1002/1099-0801(200010)14:6<422::AID-BMC25>3.0.CO;2-I PMID: 11002279
4. Wu CC, MacCoss MJ. Shotgun proteomics: tools for the analysis of complex biological systems. Curr Opin Mol Ther. 2002; 4: 242–250. PMID: 12139310
5. Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data. Mol Cell Proteom. 2005; 4: 1419–1440.
6. Webb-Robertson BJM, Cannon WR. Current trends in computational inference from mass spectrometry-based proteomics. Brief Bioinform. 2007; 8: 304–317. https://doi.org/10.1093/bib/bbm023 PMID: 17584764
7. The M, Edfors F, Perez-Riverol Y, Payne SH, Hoopmann MR, Palmblad M, et al. A protein standard that emulates homology for the characterization of protein inference algorithms. J Proteome Res. 2018; 17: 1879–1886. https://doi.org/10.1021/acs.jproteome.7b00899 PMID: 29631402
8. de Lima-Souza RA, Scarini JF, Lavareze L, Emerick C, Crescencio LR, Domingues RR, et al. Discovery proteomics reveals potential protein signature associated with malignant phenotype acquisition in pleomorphic adenoma. Oral Dis. 2021; 00: 1–11. https://doi.org/10.1111/odi.14102 PMID: 34902207
9. Kustatscher G, Grabowski P, Schrader TA, Passmore JB, Schrader M, Rappsilber J. Co-regulation map of the human proteome enables identification of protein functions. Nat Biotechnol. 2019; 37: 1361–1371. https://doi.org/10.1038/s41587-019-0298-5 PMID: 31690884
10. Bhawal R, Oberg AL, Zhang S, Kohli M. Challenges and opportunities in clinical applications of blood-based proteomics in cancer. Cancers. 2020; 12: 2428. https://doi.org/10.3390/cancers12092428 PMID: 32867043
11. Uzozie AC, Aebersold R. Advancing translational research and precision medicine with targeted proteomics. J Proteomics. 2018; 189: 1–10. https://doi.org/10.1016/j.jprot.2018.02.021 PMID: 29476807

12. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007; 4: 207–214. https://doi.org/10.1038/nmeth1019 PMID: 17327847
13. Keller A, Nesvizhskii AI, 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. https://doi.org/10.1021/ac025747h PMID: 12403597
14. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods. 2007; 4: 923–925. https://doi.org/10.1038/nmeth1113 PMID: 17952086
15. Käll L, Storey JD, MacCoss MJ, Noble WS. Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res. 2008; 7: 40–44. https://doi.org/10.1021/pr700739d PMID: 18052118
16. Serang O, MacCoss MJ, Noble WS. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. J Proteome Res. 2010; 9: 5346–5357. https://doi.org/10.1021/pr100594k PMID: 20712337
17. Uszkoreit J, Maerkens A, Perez-Riverol Y, Meyer HE, Marcus K, Stephan C, et al. PIA: an intuitive protein inference engine with a web-based user interface. J Proteome Res. 2015; 14: 2988–2997. https://doi.org/10.1021/acs.jproteome.5b00121 PMID: 25938255
18. Pfeuffer J, Sachsenberg T, Dijkstra TMH, Serang O, Reinert K, Kohlbacher O. EPIFANY: A Method for Efficient High-Confidence Protein Inference. J Proteome Res. 2020; 19: 1060–1072. https://doi.org/10.1021/acs.jproteome.9b00566 PMID: 31975601
19. Searle BC. Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. Proteomics. 2010; 10: 1265–1269. https://doi.org/10.1002/pmic.200900437 PMID: 20077414
20. Meier F, Brunner AD, Frank M, Ha A, Bludau I, Voytik E, et al. diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. Nat Methods. 2020; 17: 1229–1236. https://doi.org/10.1038/s41592-020-00998-0 PMID: 33257825
21. Huang T, Wang J, Yu W, He Z. Protein inference: a review. Brief Bioinform. 2012; 13: 586–614. https://doi.org/10.1093/bib/bbs004 PMID: 22373723
22. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 2003; 75: 4646–4658. https://doi.org/10.1021/ac0341261 PMID: 14632076
23. Savitski MM, Wilhelm M, Hahne H, Kuster B, Bantscheff M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. Mol Cell Proteom. 2015; 14: 2394–2404. https://doi.org/10.1074/mcp.M114.046995 PMID: 25987413
24. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, et al. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. J Proteome Res. 2009; 8: 3872–3881. https://doi.org/10.1021/pr900360j PMID: 19522537
25. Kong W, Wong BJH, Gao H, Guo T, Liu X, Du X, et al. PROTREC: A probability-based approach for recovering missing proteins based on biological networks. J Proteomics. 2022; 250: 104392. https://doi.org/10.1016/j.jprot.2021.104392 PMID: 34626823
26. Fraser HB, Hirsh AE, Wall DP, Eisen MB. Coevolution of gene expression among interacting proteins. Proc Natl Acad Sci USA. 2004; 101: 9033–9038. https://doi.org/10.1073/pnas.0402591101 PMID: 15175431
27. Tolani P, Gupta S, Yadav K, Aggarwal S, Yadav AK. Big data, integrative omics and network biology. Adv Protein Chem Struct Biol. 2021; 127: 127–160. https://doi.org/10.1016/bs.apcsb.2021.03.006 PMID: 34340766
28. Gupta N, Pevzner PA. False discovery rates of protein identifications: a strike against the two-peptide rule. J Proteome Res. 2009; 8: 4173–4181. https://doi.org/10.1021/pr9004794 PMID: 19627159
29. Ramakrishnan SR, Vogel C, Kwon T, Penalva LO, Marcotte EM, Miranker DP. Mining gene functional networks to improve mass-spectrometry-based protein identification. Bioinformatics. 2009; 25: 2955–2961. https://doi.org/10.1093/bioinformatics/btp461 PMID: 19633097
30. Price TS, Lucitt MB, Wu W, Austin DJ, Pizarro A, Yocum AK, et al. EBP, a program for protein identification using multiple tandem mass spectrometry datasets. Mol Cell Proteom. 2007; 6: 527–536. https://doi.org/10.1074/mcp.T600049-MCP200 PMID: 17164401
31. Fraser HB, Plotkin JB. Using protein complexes to predict phenotypic effects of gene mutation. Genome Biol. 2007; 8: 1–9. https://doi.org/10.1186/gb-2007-8-11-r252 PMID: 18042286
32. Goh WWB, Wong L. Evaluating feature-selection stability in next-generation proteomics. J Bioinform Comput Biol. 2016; 14: 1650029. https://doi.org/10.1142/S0219720016500293 PMID: 27640811

33. Goh WWB, Guo T, Aebersold R, Wong L. Quantitative proteomics signature profiling based on network contextualization. Biol Direct. 2015; 10: 1–19.

34. Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, et al. CORUM: the comprehensive resource of mammalian protein complexes—2019. Nucleic Acids Res. 2019; 47: D559–D563. https://doi.org/10.1093/nar/gky973 PMID: 30357367

35. Rahbari R, Sheahan T, Modes V, Collier P, Macfarlane C, Badge RM. A novel L1 retrotransposon marker for HeLa cell line identification. Biotechniques. 2009; 46: 277–284. https://doi.org/10.2144/000113089 PMID: 19450234

36. Fountoulakis M, Tsangaris G, Oh J, Maris A, Lubec G. Protein profile of the HeLa cell line. J Chromatogr A. 2004; 1038: 247–265. https://doi.org/10.1016/j.chroma.2004.03.032 PMID: 15233540

37. Mehta D, Scandola S, Uhrig RG. BoxCar and Library-Free Data-Independent Acquisition Substantially Improve the Depth, Range, and Completeness of Label-Free Quantitative Proteomics. Anal Chem. 2022; 94: 793–802. https://doi.org/10.1021/acs.analchem.1c03338 PMID: 34978796

38. Vizcaíno JA, Csordas A, Del-Toro N, Dianes JA, Griss J, Lavidas I, et al. 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. 2016; 44: D447–D456. https://doi.org/10.1093/nar/gkv1145 PMID: 26527722

39. Li L, Wei Y, To C, Zhu CQ, Tong J, Pham NA, et al. Integrated omic analysis of lung cancer reveals metabolism proteome signatures with prognostic impact. Nat Commun. 2014; 5: 5469. https://doi.org/10.1038/ncomms6469 PMID: 25429762

40. Li P, Hao Z, Wu J, Ma C, Xu Y, Li J, et al. Comparative proteomic analysis of polarized human THP-1 and mouse RAW264. 7 macrophages. Front immunol. 2021; 12: 700009. https://doi.org/10.3389/fimmu.2021.700009 PMID: 34267761

41. Chanput W, Mes JJ, Wichers HJ. THP-1 cell line: an in vitro cell model for immune modulation approach. Int Immunopharmacol. 2014; 23: 37–45. https://doi.org/10.1016/j.intimp.2014.08.002 PMID: 25130606

42. Hartley JW, Evans LH, Green KY, Naghashfar Z, Macias AR, Zerfas PM, et al. Expression of infectious murine leukemia viruses by RAW264. 7 cells, a potential complication for studies with a widely used mouse macrophage cell line. Retrovirology. 2008; 5: 1–6. https://doi.org/10.1186/1742-4690-5-1 PMID: 18177500

43. Holman JD, Tabb DL, Mallick P. Employing ProteoWizard to convert raw mass spectrometry data. Curr Protoc Bioinform. 2014; 46: 13–24. https://doi.org/10.1002/0471250953.bi1324s46 PMID: 24939128

44. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. Nat Methods. 2017; 14: 513–520. https://doi.org/10.1038/nmeth.4256 PMID: 28394336

45. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019; 47: D506–D515. https://doi.org/10.1093/nar/gky1049 PMID: 30395287

46. Röst HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods. 2016; 13: 741–748. https://doi.org/10.1038/nmeth.3959 PMID: 27575624

47. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015; 347: 1260419.

48. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26: 139–140. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308

49. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995; 57: 289–300.

50. Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc. 2016; 11: 2301–2319. https://doi.org/10.1038/nprot.2016.136 PMID: 27809316

51. Lim MY, Paulo JA, Gygi SP. Evaluating false transfer rates from the match-between-runs algorithm with a two-proteome model. J Proteome Res. 2019; 18: 4020–4026. https://doi.org/10.1021/acs.jproteome.9b00492 PMID: 31547658

52. Yue W, Sun Q, Landreneau R, Wu C, Siegfried JM, Yu J, et al. Fibulin-5 suppresses lung cancer invasion by inhibiting matrix metalloproteinase-7 expression. Cancer Res. 2009; 69: 6339–6346. https://doi.org/10.1158/0008-5472.CAN-09-0398 PMID: 19584278

53. Pan Y, Li X, Duan J, Yuan L, Fan S, Fan J, et al. Enoxaparin Sensitizes Human Non–Small-Cell Lung Carcinomas to Gefitinib by Inhibiting DOCK1 Expression, Vimentin Phosphorylation, and Akt Activation. Mol Pharmacol. 2015; 87: 378–390. https://doi.org/10.1124/mol.114.094425 PMID: 25488183

54. Ori A, Iskar M, Buczak K, Kastritis P, Parca L, Andrés-Pons A, et al. Spatiotemporal variation of mammalian protein complex stoichiometries. Genome Biol. 2016; 17: 1–15.

55. Kim P, Park A, Han G, Sun H, Jia P, Zhao Z. TissGDB: tissue-specific gene database in cancer. Nucleic Acids Res. 2018; 46: D1031–D1038. https://doi.org/10.1093/nar/gkx850 PMID: 29036590

56. Hounkpe BW, Chenou F, de Lima F, De Paula EV. HRT Atlas v1. 0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. Nucleic Acids Res. 2021; 49: D947–D955. https://doi.org/10.1093/nar/gkaa609 PMID: 32663312

57. Odersky M, Altherr P, Cremet V, Emir B, Maneth S, Micheloud S, et al. An overview of the Scala programming language. 2004. Available from: https://infoscience.epfl.ch/record/52656.

58. Kernighan BW, Ritchie DM. The C programming language. 2nd. Englewood Cliffs(NJ): Prentice Hall; 1988.

59. Schoof EM, Furtwängler B, Üresin N, Rapin N, Savickas S, Gentil C, et al. Quantitative single-cell proteomics as a tool to characterize cellular hierarchies. Nat Commun. 2021; 12: 3341. https://doi.org/10.1038/s41467-021-23667-y PMID: 34099695

60. Gatto L, Breckels LM, Wieczorek S, Burger T, Lilley KS. Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata. Bioinformatics. 2014; 30: 1322–1324. https://doi.org/10.1093/bioinformatics/btu013 PMID: 24413670