Identifying crosstalks among post-translational modifications in lung cancer proteomic data

Shengzhi Lai[1], Shuaijian Dai[1], Peize Zhao[2], Chen Zhou[1], Ning Li[3,4,*] and Weichuan Yu[1,3,*]

[1] Department of Electronic and Computer Engineering,
The Hong Kong University of Science and Technology, Hong Kong 000000, China

[2] Interdisciplinary Programs Office,
The Hong Kong University of Science and Technology, Hong Kong 000000, China

[3] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute,
Futian, Shenzhen 518000, China

[4] Division of Life Science,
The Hong Kong University of Science and Technology, Hong Kong 000000, China

* Correspondence: N.L. (boningli@ust.hk, Telephone: (852) 2358-7335) and W.Y. (eeyu@ust.hk, Telephone: (852) 2358-7054)

Abstract

In a lung squamous cell carcinoma data set containing over 20 million tandem mass spectra, we identified 860 peptides with post-translational modifications (PTMs) that were significantly upregulated in lung cancer samples as compared to normal samples using our new search engine named PIPI3. Among the modified peptides related to upregulated gene ontology terms, about 50% carried multiple PTMs. PIPI3 demonstrated its enabling power to provide insight into PTM crosstalk research.

Post-translational modifications (PTMs) are important in regulating cellular activities[1]. While extensive research has focused on the mechanisms of single PTMs[2], increased attention has been di-

rected towards understanding the combined effects of multiple PTMs, referred to as PTM crosstalk[3]. Notably, certain PTM crosstalk events may occur within the same peptide. For example, García-Moreno et al.[4] showed that peptide CFFCHAP bearing three PTMs related to rheumatoid arthritis (RA) functions as an antigen associated with RA. Remarkably, the prevalence of severe joint damage was statistically significant for the IgG isotype only when working with this peptide bearing three PTMs. However, this peptide and the three PTMs were selected and synthesized based on the authors' prior knowledge of RA. Identifying crosstalks among PTMs on the proteome scale remains challenging since no high-performing search method is available.

Currently, most methods[5–8] identify peptides with PTMs based on dynamic programming or enumeration. From a combinatorial perspective, identifying peptides with multiple PTMs is equivalent to finding the optimal PTM pattern that produces theoretical spectra most closely resembling the experimental spectra[9]. However, existing methods resort to approximate solutions only, due to the challenge of the exponentially increasing number of PTM combinations as the number of PTMs in one peptide increases, resulting in unsatisfactory performance when identifying peptides with multiple PTMs[10].

To address the limitations of existing approaches, we propose PIPI3, a new search engine incorporating a novel Mixed Integer Linear Programming (MILP) model to find the optimal peptide sequence with the optimal PTM pattern. PIPI3 uses a fuzzy and bidirectional matching approach to locating short amino acid sequences in protein sequences. It then constructs and solves the MILP model to determine the peptide sequences and the PTM patterns. To the best of our knowledge, PIPI3 is the first method from a combinatorial optimization angle for peptide identification and PTM characterization. We comprehensively validated PIPI3's performance using various types of data sets, demonstrating its superiority over state-of-the-art methods when identifying peptides with multiple PTMs, hence the enabling power for discovering PTM crosstalk among PTMs (**Supplementary Note 1** to **Supplementary Note 4**).

We applied PIPI3 to a proteome-scale data set of lung squamous cell carcinoma (LSCC)[11]. The data set contains 22,121,363 tandem mass spectra (MS2 spectra) collected from 22 experiments

using treatment-naive, primary LSCC tumors and paired normal adjacent tissues labeled with TMT-11 for quantification analysis. At a peptide-level false discovery rate[12] (FDR) of 0.01, PIPI3 identified 10,616,210 peptide-spectrum matches (PSMs) with TMT-11 labels, among which over 50% carried PTM(s) besides the TMT-11 labels. Unique modified peptides followed Zipf's law (**Fig.1A**), and common PTMs such as Oxidation, Deamidation, and Formylation were abundantly identified (**Fig.1B**).

The search results were then quantified and underwent motif analysis and gene ontology (GO) analysis (**METHODS**). The quantification results (q-value < 0.05) show that 4,593 unique PTM site patterns (UPSPs, groups of peptides containing the same PTM sites) have at least two PTMs, corresponding to 171,960 PSMs in total (**Fig.2A**). Among all UPSPs, 860 were significantly upregulated (q-value < 0.05, fold change > 3), and 20 were significantly downregulated (q-value < 0.05, fold change $< 1/3$) in the lung cancer sample as compared to normal samples (**Fig.2B**). From the 860 upregulated UPSPs, we analyzed the motifs on Phosphorylated S or T. We found several significantly enriched motifs, namely, p[S/T]PK, p[S/T]P, [S]xp[S/T] with q-value < 0.01 (**Fig.2C**). Based on the GO analysis of the upregulated UPSPs, we found significant upregulations (q-value $<$ 0.01) of biological process, cellular component, and molecular function (**Fig.2D**). Most of these upregulations are related to lung functions and the respiratory system and have been reported in the literature, such as hemoglobin alpha binding[13], protein binding[14], carbon dioxide transport[15], oxygen transport[16], AnxA2[17], extracellular vesicle[18], focal adhesion[19], and endoplasmic reticulum[20]. Notably, the upregulation of extracellular exosomes had high statistical significance. An example of the correspondence between UPSPs and the upregulated GO terms is shown in **Fig.2E**. About 50% of the involved UPSPs contain at least two PTMs, indicating extensive potential intra-peptide PTM crosstalks in LSCC. Among these PTMs, 63 have been identified in the literature as being associated with lung cancer (see the link in **Data Availability**).

In conclusion, we showed that PTM crosstalk exists extensively in lung cancer by reanalyzing a proteome-scale public data set with our new search engine, PIPI3. More importantly, PIPI3 can be widely applied to other proteome-scale mass spectrometry data and other disease studies to enable more findings in PTM crosstalk.

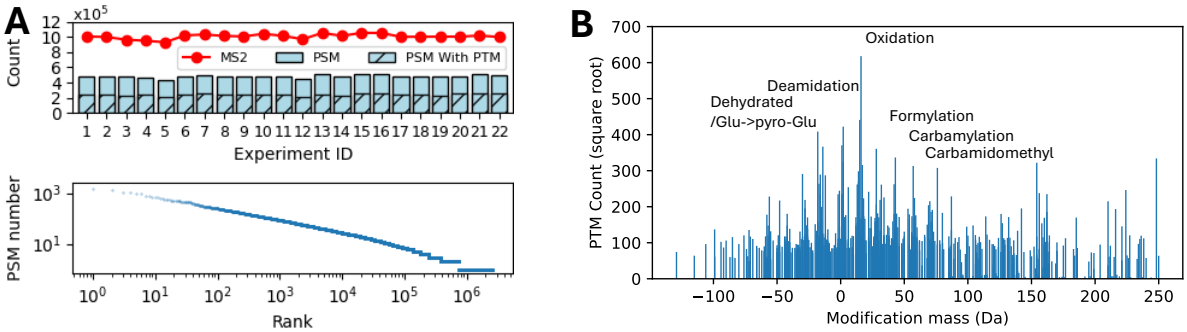
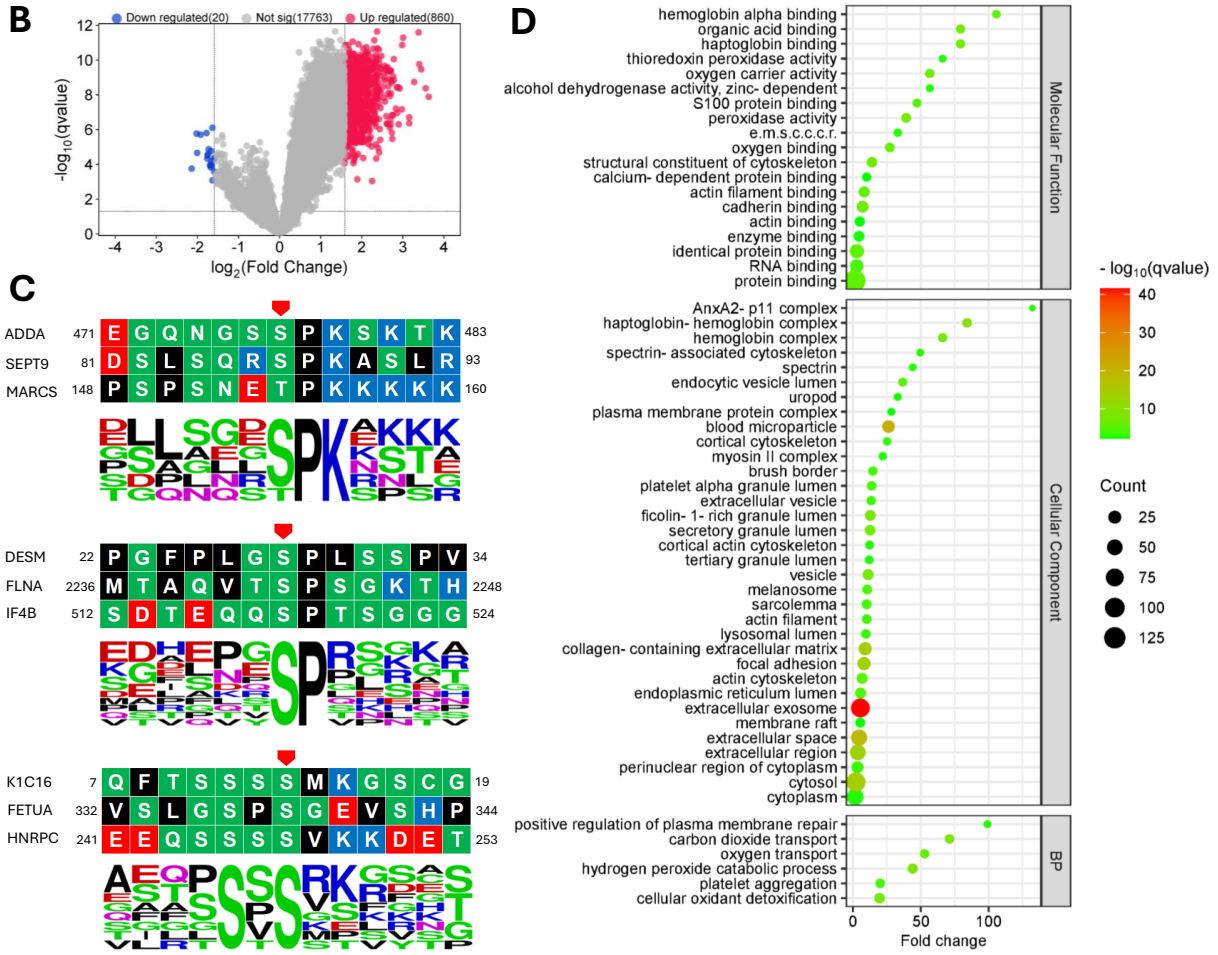


Figure 1: Search results of the LSCC data using PIPI3 at a peptide-level FDR < 0.01. (A) Identification results: upper, the number of PSMs and those with PTMs, and the number of MS2 spectra in each experiment; lower, PSMs with different modified peptides ranked by the PSM number. (B) PTM mass distribution with some abundant ones annotated.

A

PTM number in UPSP	Total number of UPSPs	Total number PSMs	Average experiment number
0	29326	2350669	19.2
1	9311	418681	17.5
2	4456	165632	16.8
3	108	4399	16.4
4	29	1929	16.5



E

Examples of the upregulated GO terms	Number of involved UPSPs with k PTMs				Examples of UPSPs with multiple PTMs
	k=1	k=2	k=3	k=4	
Hemoglobin alpha binding	146	115	0	0	V(28.031)L(53.972)GAFSDGLAHLDNLK
Protein binding	367	376	6	1	T(85.032)YFPH(142.111)FDLSHG(14.016)S(60.036)AQVK
Carbon dioxide transport	273	246	5	1	SA(57.058)V(31.990)TALWGK
Oxygen transport	273	244	5	1	L(27.995)LSHCLLVTLAAHLPAEFTPAVH(142.111)AS(87.051)LDK
Extracellular exosome	289	278	6	1	SH(72.021)CLAEVENDEM(15.995)P(2.016)ADLPSLAADFVESK
Endoplasmic reticulum lumen	22	35	1	0	W(34.063)E(37.956)RPFEVK

Figure 2: Quantification results and analysis of the LSCC data set, q-value < 0.05. (A) UPSPs with different numbers of PTMs. 4593 UPSPs have at least two PTMs. (B) 860 upregulations (fold change > 3) and 20 downregulations (fold change $< 1/3$). (C) Sequence logos of upregulated Phosphorylated S/T, q-value < 0.01. (D) GO analysis of the proteins with significantly regulated UPSPs. BP: biological process, BH-FDR < 0.01. (E) Examples of the correspondence between the upregulated GO terms and UPSPs with different numbers of PTMs.

References

1. Chicooree, N., Unwin, R. D. & Griffiths, J. R. The application of targeted mass spectrometry-based strategies to the detection and localization of post-translational modifications. *Mass Spectrometry Reviews* **34,** 595–626 (2015).

2. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnology* **24,** 1285–1292 (2006).

3. Venne, A. S., Kollipara, L. & Zahedi, R. P. The next level of complexity: crosstalk of post-translational modifications. *Proteomics* **14,** 513–524 (2014).

4. García-Moreno, C., Gómara, M. J., Castellanos-Moreira, R., Sanmartí, R. & Haro, I. Peptides bearing multiple post-translational modifications as antigenic targets for severe rheumatoid arthritis patients. *International Journal of Molecular Sciences* **22,** 13290 (2021).

5. Chi, H. *et al.* Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nature Biotechnology* **36,** 1059–1061 (2018).

6. Na, S., Kim, J. & Paek, E. MODplus: Robust and unrestrictive identification of post-translational modifications using mass spectrometry. *Analytical Chemistry* **91,** 11324–11333 (2019).

7. Devabhaktuni, A. *et al.* TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. *Nature Biotechnology* **37,** 469–479 (2019).

8. Han, X., He, L., Xin, L., Shan, B. & Ma, B. PeaksPTM: Mass spectrometry-based identification of peptides with unspecified modifications. *Journal of Proteome Research* **10,** 2930–2936 (2011).

9. Matthiesen, R., Azevedo, L., Amorim, A. & Carvalho, A. S. Discussion on common data analysis strategies used in MS-based proteomics. *Proteomics* **11,** 604–619 (2011).

10. Lai, S., Zhao, P., Zhou, C., Li, N. & Yu, W. PIPI2: Sensitive tag-based database search to identify peptides with multiple post-translational modifications. *Journal of Proteome Research* **23,** 1960–1969 (2024).

11. Satpathy, S. *et al.* A proteogenomic portrait of lung squamous cell carcinoma. *Cell* **184,** 4348–4371 (2021).

12. 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. *Journal of Proteome Research* **16,** 393–397 (2017).

13. Salem, A. *et al.* Targeting hypoxia to improve non–small cell lung cancer outcome. *JNCI: Journal of the National Cancer Institute* **110,** 14–30 (2018).

14. Mi, L. *et al.* The role of protein binding in induction of apoptosis by phenethyl isothiocyanate and sulforaphane in human non–small lung cancer cells. *Cancer Research* **67,** 6409–6416 (2007).

15. Potter, C. & Harris, A. Diagnostic, prognostic and therapeutic implications of carbonic anhydrases in cancer. *British Journal of Cancer* **89,** 2–7 (2003).

16. Keppner, A. *et al.* Lessons from the post-genomic era: Globin diversity beyond oxygen binding and transport. *Redox Biology* **37,** 101687 (2020).

17. Wang, Y. *et al.* TIM-4 orchestrates mitochondrial homeostasis to promote lung cancer progression via ANXA2/PI3K/AKT/OPA1 axis. *Cell Death & Disease* **14,** 141 (2023).

18. Shan, C. *et al.* Emerging function and clinical significance of extracellular vesicle noncoding RNAs in lung cancer. *Molecular Therapy-Oncolytics* **24,** 814–833 (2022).

19. Havel, L. S., Kline, E. R., Salgueiro, A. M. & Marcus, A. I. Vimentin regulates lung cancer cell adhesion through a VAV2–Rac1 pathway to control focal adhesion kinase activity. *Oncogene* **34,** 1979–1990 (2015).

20. Tsai, H. *et al.* Endoplasmic reticulum ribosome-binding protein 1 (RRBP1) overexpression is frequently found in lung cancer patients and alleviates intracellular stress-induced apoptosis through the enhancement of GRP78. *Oncogene* **32,** 4921–4931 (2013).

METHODS

Overview of PIPI3 method

PIPI3 takes a protein database and raw files of MS2 spectra as input and outputs a PSM list. A raw file undergoes pre-processing using MSConverter[21], and short amino acid sequences (called tags) are extracted from the MS2 spectra with scores equal to the total intensity of the involved peaks (**Supplementary Note 6**). These tags are subsequently used to retrieve candidates from an FM-indexed[22] protein database with a fuzzy and bidirectional matching approach. The mass difference between a peptide sequence and the precursor of an MS2 spectrum, denoted as Δm, is considered as the total mass of potential PTMs, which are then used to construct MILP models that can be solved using Gurobi. Subsequently, peptide candidates with characterized PTM patterns are collected and reranked by a protein feedback module[10,23]. Finally, PIPI3 outputs a PSM list with FDR controlled by the target-decoy strategy.

Fuzzy and bidirectional match with FM-index

While many methods employ fixed-length tags to retrieve candidates from a database, PIPI3 combines tags of various lengths to obtain a better quality candidate list, thus resulting in better performance, an approach described and validated in our previous work[10,24]. This is enabled by the FM-index, which allows efficient querying of sequences of any length in the template text. Nevertheless, it is worth noting that even a single erroneous amino acid in a tag can cause the retrieval to fail. Thus, we use a fuzzy and bidirectional matching strategy with the FM-index to accommodate erroneous amino acids at either the N-term or C-term of a tag.

The matching process is demonstrated in **Algorithm 1**. All target and decoy protein sequences are concatenated using dot signs into a long sequence, which is used to generate the original FM-indexed database (o-FMDB). The matching between a tag and the long sequence is performed in the

backward direction of the tag sequence. For instance, the occurrence of a tag TAGABC is found by checking all suffixes of the tag in the following: C, CB, CBA, ..., CBAGAT. If CBAGAT is finally matched, all the relative locations of TAGABC in the long sequence are output; otherwise, the retrieval fails. This situation is common, particularly with long tags that are prone to containing erroneous amino acids extracted from noise peaks. In this example, if the tag is XTAGABC, where X represents an incorrect amino acid, the matching process will fail when checking CBAGATX. As a result, the previously matched TAGABC will be discarded. To address this issue, we use a fuzzy matching strategy capable of accommodating error amino acids on the N-term of the tag. During the matching process, we record the matched amino acids and output the longest substring that matched successfully. The output for XTAGABC would be TAGABC.

However, fuzzy matching alone is insufficient where erroneous amino acids appear at the C-term of a tag, such as TAGABCX, because the initial matching is incorrect. In such cases, we employ a bidirectional match approach along with an additional reversed FM-indexed database (r-FMDB) generated from the reversed long sequence. In this approach, we attempt to match the reversed tag XCBAGAT against the r-FMDB. The backward matching process of the reversed tag is equivalent to a forward matching of the original tag. By combining this fuzzy and bidirectional matching strategy, we consistently obtain the output TAGABC, regardless of whether the erroneous amino acid X appears on the N-term or C-term of the tag, which enhances the candidate retrieval sensitivity.

One might consider manually splitting the tag sequence into substrings and matching them individually to avoid generating the r-FMDB. However, such an approach would be more time-consuming. For example, if we split TAGABCX into TAGABCX, TAGABC, TAGAB, TAGA, AGABCX, GABCX, and ABCX, we would need to perform seven separate matches to the o-FMDB to find the correct one. In contrast, with the fuzzy and bidirectional matching approach, we only perform two matches: one for TAGABCX against the o-FMDB and another for XCBAGAT against the o-FMDB. Furthermore, we can generate and store the two FMDBs for a protein database and reuse them whenever needed. Compared to the overall running time, the time used for the r-FMDB generation is negligible.

Algorithm 1 Fuzzy and bidirectional candidate retrieval

Require: t: tag, DB protein database

Ensure: $protPosList$: list of all proteins and positions where t occurs

```
longSeq ← getLongSeqFromDB(DB)                          ▷ join protein sequences
o-FMDB ← getFMDB(longSeq)                               ▷ generate o-FMDB, refer to FM-index paper[22]
r-FMDB ← getFMDB(longSeq.reverse())                     ▷ generate r-FMDB
protPosList ← ∅                                         ▷ initialize protPosList as empty

tmpProtPosList ← fuzzyMatch(t,o-FMDB)                   ▷ match t to o-FMDB
protPosList.extend(tmpProtPosList)
tmpProtPosList ← fuzzyMatch(t.reverse(),r-FMDB)         ▷ match reversed t to r-FMDB
protPosList.extend(tmpProtPosList)
return protPosList

function getLongSeqFromDB(DB)
    longSeq ← ""                                        ▷ initialize longSeq as an empty string
    for prot ∈ DB do
        longSeq.append(prot+"."+prot.shuffle()+".")     ▷ join all target and decoy sequences
    end for
    return longSeq
end function

function fuzzyMatch(t,FMDB)
    tmpProtPosList ← ∅
    for i ← t.length-1 do                               ▷ try every suffix of t
        suffix ← t.substring(i, t.length)
        tmpProtPosList ← FMDB.match(suffix)             ▷ refer to FM-index paper[22]
        if tmpProtPosList.isEmpty() then
            break                                       ▷ output current tmpProtPosList if match fails
        end if
        i ← i − 1
        if i < 0 then
            break                                       ▷ finish trying all suffixes
        end if
    end for
    return tmpProtPosList
end function
```

Protein candidates scoring

The relationship between the protein sequences and the corresponding tags located on them is a many-to-many matching. Occasionally, one protein sequence may contain multiple tags in close proximity, indicating a higher chance that the underlying peptide is digested in this region

(**Figure 3B**). Thus, we consider the proteins and either a single tag or multiple tags in close proximity as candidates for subsequent peptide identification and PTM characterization. Overlapping tags in close proximity are merged into one if their peaks align; otherwise, tags with lower scores are discarded as noise. Finally, the candidates are ranked by the sum of scores of the located tag(s). This aggregates the evidence from multiple tags that support the same peptide sequence, thereby avoiding redundant examination of individual tags.

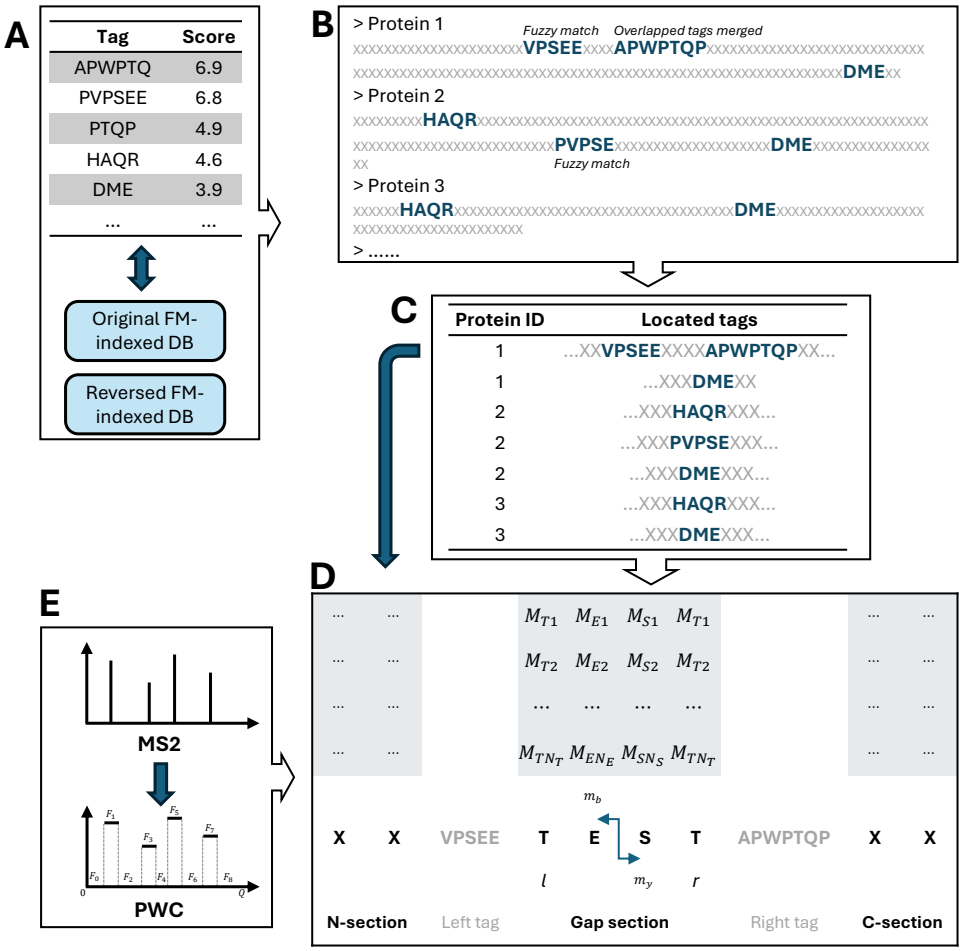


Figure 3: A schematic workflow of protein candidates generation. (A) Extracted tags are ranked by scores and used to retrieve protein candidates from the FM-indexed databases. (B) Tags are located on the retrieved protein sequence using a fuzzy and bidirectional matching strategy. (C) Protein sequences with located tags are considered as candidates and ranked by the total score of the located tags. (D) The peaks of an MS2 spectrum are represented by a piece-wise constant function. (E) In either the N-section or C-section, amino acids are grouped into fixed ones, optional ones, and infeasible ones for MILP model construction.

MILP model

Given a candidate pair, the protein sequence can be segmented into an N-section, a C-section, or gap sections by the tag(s). These sections may or may not be present in every candidate. Without loss of generality, we illustrate the MILP model using a protein sequence along with two tags that segment the protein into an N-section, a C-section, and a gap section, as shown in **Figure 3D**. The mass shift Δm in each section is known based on the peak locations of the tags, indicating that the PTM patterns in these sections are mutually independent. The assumption is that the maximal total intensity of experimental peaks, which match to the theoretical peaks, is achieved when the theoretical peaks are generated from the peptide sequence with the optimal PTM pattern. The notations of the symbols are presented in Table 1.

MILP for the gap section

We start with the simpler case within the gap between a left tag and a right tag. Δm is computed as the mass difference between the start peak of the right tag and the end peak of the left tag, i.e., $\Delta m = M_{r,start} - M_{l,end}$. The set of amino acids within the gap is denoted as $\mathcal{A}$. The set of potential PTMs on amino acid a is denoted as $\mathcal{P}_a$. For each $p \in \mathcal{P}_a$, where $a \in \mathcal{A}$, a binary variable x is assigned to indicate the presence of p. The masses of b- and y-ions, b and y, are expressed as linear combinations of x and other constants. The objective is to maximize the total intensity of

Table 1: Symbols used in the MILP model

x	binary variable, the presence of a PTM (0/1=absent/present)
z	binary variable, the presence of an optional amino acid (0/1=absent/present)
b, y	continuous variable, the mass of a b-, y-ion
a, a'	amino acids
p	PTM
m	mass of a PTM
w	mass of an amino acid
Δm	total mass of PTMs to be settled in any section
$\mathcal{A}$	set of all amino acids in any section
$\mathcal{F}$	set of all fixed amino acids in an N-/C-section
$\mathcal{O}$	set of all optional amino acids in an N-/C-section
$\mathcal{P}_a$	set of potential PTMs in amino acid a
$\mathcal{L}_a$	set of amino acids to the left (N-term) of amino acid a in any section
$\mathcal{R}_a$	set of amino acids to the right (C-term) of amino acid a in any section
M, I	mass and intensity of an MS2 peak
E	number of MS2 peaks
C	mass of the precursor plus two protons
τ_1, τ_2	MS1 and MS2 tolerance
N_{max}	maximum allowed number of PTMs in any section
P_{min}, P_{max}	minimum and maximum allowed total mass of PTMs
d, f, t	assistant variables
F, G, Q, ϵ	assistant parameters

the matched experimental peaks by all b- and y-ions. The model is formulated as follows:

$$\max_{x_1,x_2,\ldots,\in\{0,1\}} \sum_{a\in\mathcal{A}} [PWC(b_a) + PWC(y_a) - 0.1 \sum_{p\in\mathcal{P}_a} x_p] \tag{1}$$

$$\text{s.t. } |\sum_{a\in\mathcal{A}}(w_a + \sum_{p\in\mathcal{P}_a} m_p x_p) - \Delta m| \leq \tau_1, \tag{2}$$

$$b_a = M_{l,end} + \sum_{a'\in\mathcal{L}_a} (w_{a'} + \sum_{p\in\mathcal{P}_{a'}} m_p x_p),\ \forall\ a \in \mathcal{A}, \tag{3}$$

$$y_a = M_{r,start} + \sum_{a'\in\mathcal{R}_a} (w_{a'} + \sum_{p\in\mathcal{P}_{a'}} m_p x_p),\ \forall\ a \in \mathcal{A}, \tag{4}$$

$$|b_a + y_a - C| \leq \tau_1,\ \forall\ a \in \mathcal{A}, \tag{5}$$

$$\sum_{p\in\mathcal{P}_a} x_p \leq 1,\ \forall\ a \in \mathcal{A}, \tag{6}$$

$$\sum_{a\in\mathcal{A}} \sum_{p\in\mathcal{P}_a} x_p \leq N_{max}, \tag{7}$$

where $|\cdot|$ is the absolute value of a scalar or the cardinality of a set, and $\mathcal{L}_a$ and $\mathcal{R}_a$ are the sets of amino acids to the left or right of amino acid a. Constraint (2) enforces that the difference between the total mass of PTMs and Δm is within τ_1; (3) and (4) express b and y with PTM masses, amino acid masses, and peaks of the left and right tags; (5) indicates the relationship between each pair of b and y at each a; (6) enforces that at most one PTM can occur on each amino acid; (7) is an upper limit of the total number of PTMs.

The objective function is the summation of the intensities of peaks matched by b- and y-ions, plus a penalty term on the total number of PTMs. $PWC(\cdot)$ is a piece-wise constant (PWC) function that represents the MS2 peak intensities and zeros between two MS2 peaks. As shown in **Fig. 3E**, these values are denoted as $(F_0, F_1, \ldots, F_{2E})$, where E is the number of MS2 peaks. Thus we have

$$(F_0, F_1, F_2, F_3, F_4, \ldots, F_{2E-2}, F_{2E-1}, F_{2E}) = (0, I_1, 0, I_2, 0, \ldots, 0, I_E, 0). \tag{8}$$

As shown in (9), if the input mass t matches the mass of any peak M_e, the function returns the

corresponding intensity I_e; otherwise, it returns 0.

$$PWC(t) = \begin{cases} I_e, & M_e - \tau_2 < t \leq M_e + \tau_2, \ \forall e = 1, 2, \ldots, E, \\ 0, & otherwise. \end{cases} \tag{9}$$

$PWC(\cdot)$ is not inherently linear but can be linearized using auxiliary variables. We define binary variables $(d_0, d_1, \ldots, d_{2E})$, one for each F_i, and a continuous variable f for the return value of $PWC(t)$. Then, we apply the constraints as follows:

$$\sum_{i=0}^{2E} d_i = 1, \tag{10}$$

$$m \geq \sum_{e=1}^{E} (M_e - \tau_2 + \epsilon) d_{2e-1} + \sum_{e=1}^{E} (M_e + \tau_2 + \epsilon) d_{2e} + 0 d_0, \tag{11}$$

$$m \leq \sum_{e=1}^{E} (M_e - \tau_2) d_{2e-2} + \sum_{e=1}^{E} (M_e + \tau_2) d_{2e-1} + Q d_{2E}, \tag{12}$$

$$f \leq F_i + G_i(1 - d_i), \ \forall \ i = 0, 1, 2, \ldots, 2E, \tag{13}$$

$$d_i \in \{0, 1\}, \ \forall \ i = 0, 1, 2, \ldots, 2E, \tag{14}$$

where Q is a constant that upper bounds all M_e, G_i is a big constant such that $F_i + G_i \geq G_j, \forall j \neq i$, and ϵ is a small tolerance constant. Constraint (10) enforces that only one F_i can be selected; (11) and (12) enforce that $M_{i-1} \leq m \leq M_i \implies d_i = 1$; (13) imposes an upper bound on the return value f, restricting it to be F_i only when $d_i = 1$ because other inequalities where $d_i = 0$ become redundant because of G_i. Since the objective function is maximum-wise, f will automatically assume the value of F_i. As a result, the objective function can be replaced with

$$\max_{x_1, x_2, \ldots, \in \{0,1\}} \sum_{a \in \mathcal{A}} (f_{b_a} + f_{y_a} - 0.1 \sum_{p \in \mathcal{P}_a} x_p). \tag{15}$$

MILP for the N- and C-section

The models for the N- and C-sections are similar. Thus, without loss of generality, we only present the one for the N-section here. Taking the first candidate in **Figure 3C** as an example, VPSEE is the right tag for the N-section, and no left tag is available. The N-section ends at the right tag, but the starting amino acid site is unknown without a left tag, i.e., the peptide sequence needs to be determined. To achieve this, we could extend the N-section towards the protein N-term, building and solving one MILP model at every potential position, which is quite time-consuming and involves repetitive work. Instead, we incorporate the extension step into a single model, which simultaneously enables the determination of the peptide sequence and the characterization of the PTM pattern.

In the N-section, Δm equals the start peak of the right tag $M_{r_s tart}$. Given P_{min} and P_{max}, the amino acids on the N-term side of the right tag can be classified into fixed, optional, and infeasible categories. From the right tag towards the protein N-term, the first several amino acids are in the fixed amino acids set $\mathcal{F}$; without these amino acids, the total mass of present PTMs will be beyond P_{max}. The following several amino acids are in the optional amino acids set $\mathcal{O}$, which may or may not be present in the sequence. The remaining amino acids are infeasible, with which the total mass of present PTMs will be smaller than P_{min}. In summary, the amino acids in $\mathcal{F}$ and $\mathcal{O}$ are selected and added to the sets from the ones in the N-section in the direction from the right tag to the N-term, satisfying the following conditions:

$$\Delta m - \sum_{a \in \mathcal{F}} w_a \leq P_{max}, \tag{16}$$

$$\Delta m - \sum_{a \in \mathcal{F} \cup \mathcal{O}} w_a \geq P_{min}. \tag{17}$$

b_a and y_a are calculated similarly to that in the gap section, except that $M_{l,end} = 0$ for the N-section.

Then, the model is written as follows:

$$\max_{\substack{x_1,x_2,\ldots,\in\{0,1\},\\ z_1,z_2,\ldots,\in\{0,1\}}} \sum_{a\in\mathcal{A}} [PWC(b_a) + PWC(y_a) - 0.1 \sum_{p\in\mathcal{P}_a} x_p] \tag{18}$$

$$\text{s.t. } |\sum_{a\in\mathcal{F}} w_a + \sum_{a\in\mathcal{O}} w_a z_a + \sum_{a\in\mathcal{A}} \sum_{p\in\mathcal{P}_a} m_p x_p - \Delta m| \leq \tau_1, \tag{19}$$

$$b_a = \sum_{a'\in\mathcal{L}_a} (w_{a'} + \sum_{p\in\mathcal{P}_{a'}} m_p x_p),\ \forall\ a \in \mathcal{A}, \tag{20}$$

$$y_a = M_{r,start} + \sum_{a'\in\mathcal{R}_a} (w_{a'} + \sum_{p\in\mathcal{P}_{a'}} m_p x_p),\ \forall\ a \in \mathcal{A}, \tag{21}$$

$$|b_a + y_a - C| \leq \tau_1,\ \forall\ a \in \mathcal{A}, \tag{22}$$

$$z_a \leq z_b,\ \forall\ \textit{adjacent amino acids}\ a, b \in \mathcal{O}, \tag{23}$$

$$\sum_{p\in\mathcal{P}_a} x_p \leq 1,\ \forall\ a \in \mathcal{F}, \tag{24}$$

$$\sum_{p\in\mathcal{P}_a} x_p \leq z_a,\ \forall\ a \in \mathcal{O}, \tag{25}$$

$$\sum_{a\in\mathcal{A}} \sum_{p\in\mathcal{P}_a} x_p \leq N_{max}. \tag{26}$$

The objective function (18) can be linearized using the same approach as in the gap section. The differences between the model for the gap section and the one here are as follows: for $a \in \mathcal{O}$, the mass is only added when $z_a = 1$ as constrained by (19); $M_{l,end}$ in (20) is omitted for being zero; (23) enforces that, for every two adjacent optional amino acids, the left one can be present only if the right one is present; and constraints (25) enforces that if an optional amino acid is absent, no PTM can occur on it.

Analysis results of the LSCC data set

We applied PIPI3 to uncover the PTM profile underlying the LSCC data. The data sets were generated by the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC) and downloaded from the Proteomic Data Commons[25] with the data identifier PDC000234. The data set consists of 22,121,363 MS2 spectra prospectively collected from 22 TMT-11 plex experi-

ments from 108 treatment-naive, primary LSCC tumors, and 99 paired normal adjacent tissues. The TMT label was set as a variable modification on the peptide N-term and all Serines and set as a fixed modification on all Lysines. Full parameter settings are recorded in **Supplementary Table S5**. Original parameter files are deposited on Zenodo at https://doi.org/10.5281/zenodo.13234739.

Quantification

PSMs without a TMT label were excluded from the analysis. Peptides with more than one PSM (PSM number > 1) were selected as quantifiable. The intensity of the TMT reporter ion for each PSM was extracted from the original mgf data. It was assumed that each sample would contain a set of unregulated peptides with abundances comparable to the common reference (CR) sample, meaning these peptides should have a log TMT ratio centered at zero in the normalized sample. Consequently, two-component normalization of each TMT label's intensity was applied against the CR[11].

PSM-level log ratios were calculated as the mean log values for each patient within the corresponding PSM. Peptides containing the same PTM sites were grouped into a UPSP, and the UPSP ratio was calculated as the mean value of all contributing PSM-level ratios. Each pool of the LSCC dataset was treated as an individual experiment, and a UPSP was required to have different mean log ratios in each pool. Only UPSPs present in more than 60% of the pools were selected for further quantification. A one-sample t-test was conducted to generate the p-value for each UPSP, followed by BH-FDR correction to produce the q-value.

Motif analysis

Motifs were extracted by Motif-X[26] via the MEME suite[27], and plotted by WebLogo[28]. The foreground data set was obtained from all the peptides with Phosphorylation@S, with fold change > 3. In total, 43 sequences were input for motif discovery. These peptide sequences were centered

on Phosphorylation@S and extended to 13 amino acids (±6 on both the peptide N- and C-term). When the extension was unavailable on the term, "X" would be added to keep the same sequence length. Background peptides were extracted from all peptide sequences of length 13 in the Uniprot Human protein database, which were centered on S. The minimum number of motif occurrences was set to 8, and the significance threshold was set to q-value < 0.01.

Gene ontology enrichment analysis

All proteins that contain quantified and modified peptides (q-value < 0.05 and fold change > 3) were further analyzed for GO enrichment against the human genome as the background data set. The statistical significance of GO analysis was calculated by a two-tailed Fisher exact test and followed by BH-FDR correction, and GO terms with q-value < 0.01 were output and plotted.

Code availability

PIPI3 was developed in the Java programming language. The source code can be accessed on GitHub at https://github.com/lsz1994024/PIPI3.

Data availability

The parameter and result files are deposited on Zenodo at https://doi.org/10.5281/zenodo.13234739. The LSCC data set can be downloaded from the Proteomic Data Commons with the mentioned data identifier.

Acknowledgements

The study was partially supported by grants from the Hong Kong Research Grants Council (RGC): 16102422, 16103621, T12-101/23-N, and R4012-18; MHP/033/20 from the Innovation and Technology Commission (ITC) of the Hong Kong S.A.R., the Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone project (HZQB-KCZYB-2020083), and internal grants 3030_009 and BGF.001.2023 from HKUST.

References

21. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* **30,** 918–920 (2012).

22. Ferragina, P. & Manzini, G. *Opportunistic data structures with applications* in *Proceedings 41st Annual Symposium on Foundations of Computer Science* (2000), 390–398.

23. Zhou, C. *et al.* Exhaustive cross-linking search with protein feedback. *Journal of Proteome Research* **22,** 101–113 (2022).

24. Lai, S., Li, N. & Yu, W. *Combining tags of various lengths benefits peptide identification in bottom-up proteomics* in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2023), 241–246.

25. Thangudu, R. R. *et al.* Abstract LB-242: Proteomic Data Commons: A resource for proteogenomic analysis. *Cancer Research* **80,** LB–242 (2020).

26. Schwartz, D. & Gygi, S. P. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nature Biotechnology* **23,** 1391–1398 (2005).

27. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME suite. *Nucleic Acids Research* **43,** W39–W49 (2015).

28. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: A sequence logo generator. *Genome research* **14,** 1188–1190 (2004).