Optimization-Based Peptide Mass Fingerprinting for Protein Mixture Identification

Weichuan Yu, Ph.D.

Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology

The First China Workshop on Computational Proteomics
Beijing, 2010-11-11

Joint work with Zengyou He, Can Yang, Chao Yang, Robert Qi and Jason Tam

Outline

1 Introduction

2 Method

3 Experiments
- Simulation Study
- Real Data

4 Conclusion

Background of Mass Spectrometry Data Analysis

- We like to know:
  - what proteins/peptides are in a sample?
  - What are their expression levels?
- Issues in the analysis:
  - Noise:
    - Many sources: chemical, electrical, instrumental
    - Physics not completely understood
    - Low abundance signal coexists
  - Measurement range:
    Is it enough to record all the information?
  - Dynamic properties of samples:
    Do we obtain the data at the right time?

Motivation of Protein Identification

- MS data describes information directly at the peptide level.
- We need to identify the corresponding proteins to better understand the cellular functions.
- Information at the protein level is probably more robust.

Common Methods for Protein Identification

1. Peptide Sequencing Method (Using MS/MS data).
   - Pros: More accurate
   - Cons: Lower coverage (especially peptides with low intensity values)
2. Peptide Mass Fingerprinting (Using single-stage MS Data)
   - Pros: Higher coverage
   - Cons: Less accurate and cannot handle protein mixtures

- Can we remove the limitations of PMF ?

Our approach: We formulate the identification of protein mixtures as an optimization problem.

Common Methods for Protein Identification

1. Peptide Sequencing Method (Using MS/MS data).
   - Pros: More accurate
   - Cons: Lower coverage (especially peptides with low intensity values)
2. Peptide Mass Fingerprinting (Using single-stage MS Data)
   - Pros: Higher coverage
   - Cons: Less accurate and cannot handle protein mixtures

- Can we remove the limitations of PMF ?

Our approach: We formulate the identification of protein mixtures as an optimization problem.

PMF for Single Protein Identification

PMF method for single protein identification consists of the following steps:

- (1) Protein purification:
  The 2D gel-based separation produces purified protein samples.
- (2) Protein digestion: e.g. trypsin digestion.
- (3) MS data acquisition and peak detection:
  record the masses of resulting peptides.
- (4) PMF scoring:
  match the MS spectrum with respect to the protein database and report the best ones.

PMF for Single Protein Identification

Find a single protein that maximizes the scoring function:

$$\widehat{X} = \arg\max_{X_i \in D} S^{(L)}(Z, X_i), \qquad (1)$$

- $Z = (z_1, z_2, ...z_l)$: Experimental peaks
- $D = (X_1, X_2, ...X_g)$: Protein database
- $S^{(L)}(Z, X_i, \sigma)$: Scoring function,
  σ: mass tolerance threshold.

PMF for Protein Mixture Identification

Replace single protein X_i with a set of proteins Y:

$$\widehat{Y} = \arg\max_{Y \subseteq D} S^{(M)}(Z, Y), \tag{2}$$

- We have the same input Z
- Our objective is to find a set of proteins $\widehat{Y}$ that best "explains" Z

Existing Approaches

- $S^{(L)}(Z, X_{u_j})$
  Directly apply the single protein identification method to protein mixtures.
- The *subtraction* strategy:
  - (1) First identify a protein with the highest score;
  - (2) Remove the peaks associated with this protein from the input data
  - (3) Go back to step (1) until the score is lower than a predefined threshold.

  Suppose the peak subset Z_0 is empty. At each step $j(1 \leq j \leq k)$, calculate the score as:

$$S^{(L)}(Z - \bigcup_{t=0}^{j-1} Z_t, X_{u_j}). \tag{3}$$

Choice of Scoring Function

- (1) Virtual single protein approach.
  $S^{(M)}(Z, Y) = S^{(L)}(Z, \widetilde{V})$.
  $\widetilde{V}$: Use a set of proteins to represent a virtual single protein.
- (2) Peak partition approach.
  $S^{(M)}(Z, Y) = \sum_{j=1}^{k} S^{(L)}(Z_j, X_{u_j})$.
  - Partitioning Z into Z_j is tricky.
  - The *subtraction* strategy: greedy partition
  - Random matching is the major concern
- We choose the virtual single protein approach here.

Scoring Function for Single Protein Identification

The probability that a protein X_i has r_i randomly matched peaks in Z (assuming binomial distribution):

$$Pr(|M_Z(X_i)| = r_i) = C_l^{r_i} p_i^{r_i} (1 - p_i)^{l - r_i} \tag{4}$$

$$\text{Score: } S^{(L)}(Z, X_i) = -\ln C_l^{r_i} - r_i \ln p_i - (l - r_i) \ln(1 - p_i) \tag{5}$$

- $M_Z(X_i)$: subset of Z whose peaks match protein X_i
- l: the number of observed peaks in Z
- r_i: number of peptides in protein X_i.
- p_i: the probability for at least one match

$$p_i = 1 - (1 - 2\sigma/\Delta)^{n_i}, \tag{6}$$

Δ: mass range; σ: mass tolerance.

Scoring Function for Protein Mixture Identification

$$S^{(M)}(Z,Y) = -\ln C_l^{r_Y} - r_Y \ln p_Y - (l - r_Y)\ln(1 - p_Y) \qquad (7)$$

- Y consists of k proteins $X_{u_1}, X_{u_2}, ..., X_{u_k}$
- $r_Y = |\bigcup_{j=1}^{k} M_Z(X_{u_j})|$
  and
  $p_Y = 1 - (1 - 2\sigma/\Delta)^{\sum_{j=1}^{k} n_{u_j}}$.

Maximization of Scoring Function

Two cases:

- The number of ground-truth proteins k is known.
  Losak Algorithm
- The number of ground-truth proteins k is unknown.
  Losau Algorithm

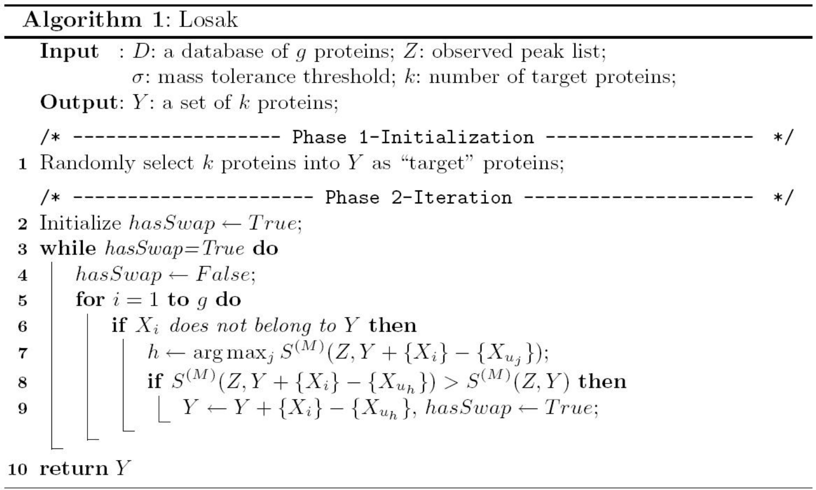
Losak Algorithm

A LOcal Search Algorithm with Known k.

- (1) Randomly select k proteins into Y as "target" proteins
- (2.1) In the iteration process, swap each "non-target" protein with the k target proteins and re-evaluate the scoring function.
- (2.2) keep those "target" proteins that achieve the best scoring values and proceed to the next protein.

Losak Algorithm-Detail

Losak Algorithm-Detail



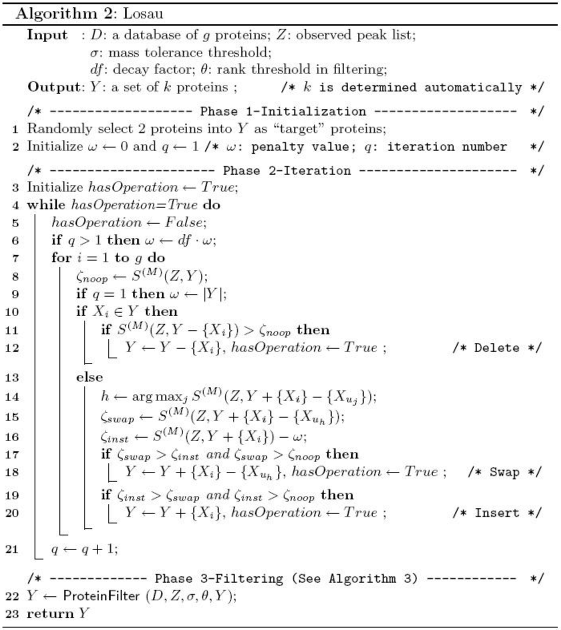
Algorithm 1: Losak

```
Input  : $D$: a database of $g$ proteins; $Z$: observed peak list;
         $\sigma$: mass tolerance threshold; $k$: number of target proteins;
Output: $Y$: a set of $k$ proteins;
/* ------------------- Phase 1-Initialization ------------------- */
Randomly select $k$ proteins into $Y$ as "target" proteins;
/* --------------------- Phase 2-Iteration --------------------- */
Initialize $hasSwap \leftarrow True$;
while $hasSwap = True$ do
    $hasSwap \leftarrow False$;
    for $i = 1$ to $g$ do
        if $X_i$ does not belong to $Y$ then
            $h \leftarrow \arg\max_j S^{(M)}(Z, Y + \{X_i\} - \{X_{u_j}\})$;
            if $S^{(M)}(Z, Y + \{X_i\} - \{X_{u_h}\}) > S^{(M)}(Z, Y)$ then
                $Y \leftarrow Y + \{X_i\} - \{X_{u_h}\}$, $hasSwap \leftarrow True$;
return $Y$
```

Losau Algorithm

A LOcal Search Algorithm with Unknown k.

- (1) Initialize with 2 proteins.
- (2) Iterate with swap, insert and delete operation. (Occam's razor principle,Penalizing insert operation using ω)
- (3) Prune the protein list

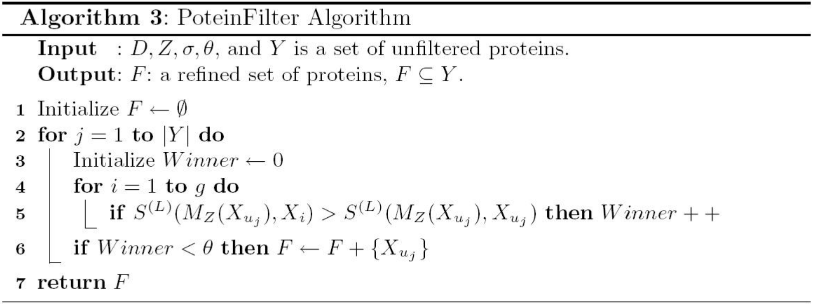


```
Algorithm 2: Losau
  Input : $D$: a database of $g$ proteins; $Z$: observed peak list;
          $\sigma$: mass tolerance threshold;
          $df$: decay factor; $\theta$: rank threshold in filtering;
  Output: $Y$: a set of $k$ proteins ;      /* $k$ is determined automatically */

  /* ------------------- Phase 1-Initialization ------------------- */
  Randomly select 2 proteins into $Y$ as "target" proteins;
  Initialize $\omega \leftarrow 0$ and $q \leftarrow 1$ /* $\omega$: penalty value; $q$: iteration number */
  /* --------------------- Phase 2-Iteration --------------------- */
  Initialize $hasOperation \leftarrow True$;
  while $hasOperation = True$ do
      $hasOperation \leftarrow False$;
      if $q > 1$ then $\omega \leftarrow df \cdot \omega$;
      for $i = 1$ to $g$ do
          $\zeta_{noop} \leftarrow S^{(M)}(Z, Y)$;
          if $q = 1$ then $\omega \leftarrow |Y|$;
          if $X_i \in Y$ then
              if $S^{(M)}(Z, Y - \{X_i\}) > \zeta_{noop}$ then
                  $Y \leftarrow Y - \{X_i\}$, $hasOperation \leftarrow True$ ;      /* Delete */
          else
              $h \leftarrow \arg\max_j S^{(M)}(Z, Y + \{X_i\} - \{X_{u_j}\})$;
              $\zeta_{swap} \leftarrow S^{(M)}(Z, Y + \{X_i\} - \{X_{u_h}\})$;
              $\zeta_{inst} \leftarrow S^{(M)}(Z, Y + \{X_i\}) - \omega$;
              if $\zeta_{swap} > \zeta_{inst}$ and $\zeta_{swap} > \zeta_{noop}$ then
                  $Y \leftarrow Y + \{X_i\} - \{X_{u_h}\}$, $hasOperation \leftarrow True$ ;   /* Swap */
              if $\zeta_{inst} > \zeta_{swap}$ and $\zeta_{inst} > \zeta_{noop}$ then
                  $Y \leftarrow Y + \{X_i\}$, $hasOperation \leftarrow True$ ;      /* Insert */
      $q \leftarrow q + 1$;
  /* ------------ Phase 3-Filtering (See Algorithm 3) ------------ */
  $Y \leftarrow$ ProteinFilter $(D, Z, \sigma, \theta, Y)$;
  return $Y$
```

Filtering Procedure in the Losau Algorithm

Idea: if X_{u_j} is the ground-truth protein, then the chance that other proteins in the database has a better score than X_{u_j} on $M_Z(X_{u_j})$ is very low.
We use the number of “winning proteins” to measure the rank uncertainty and θ as the threshold to remove false positives.



Algorithm 3: PoteinFilter Algorithm

```
Input  : $D, Z, \sigma, \theta$, and $Y$ is a set of unfiltered proteins.
Output: $F$: a refined set of proteins, $F \subseteq Y$.
Initialize $F \leftarrow \emptyset$
for $j = 1$ to $|Y|$ do
    Initialize $Winner \leftarrow 0$
    for $i = 1$ to $g$ do
        if $S^{(L)}(M_Z(X_{u_j}), X_i) > S^{(L)}(M_Z(X_{u_j}), X_{u_j})$ then $Winner{+}{+}$
    if $Winner < \theta$ then $F \leftarrow F + \{X_{u_j}\}$
return $F$
```

Evaluation Criteria and PMF Algorithms

We use standard performance metrics in information retrieval, including *precision*, *recall*, and $F1 - measure$.

- n_{TP}: the number of true positives.
- n_{FP}: the number of false positives.
- n_P: the number of all ground-truth proteins.
- $precision = n_{TP}/(n_{TP} + n_{FP})$
- $recall = n_{TP}/n_P$
- $F1 - measure = \frac{2 \cdot precision \cdot recall}{precision + recall}$

Algorthims

In performance comparison, we use the following algorithms:

- SPA: single protein identification algorithm.
- Subtraction algorithm
- Losak algorithm
- Losau algorithm

Outline

1 Introduction

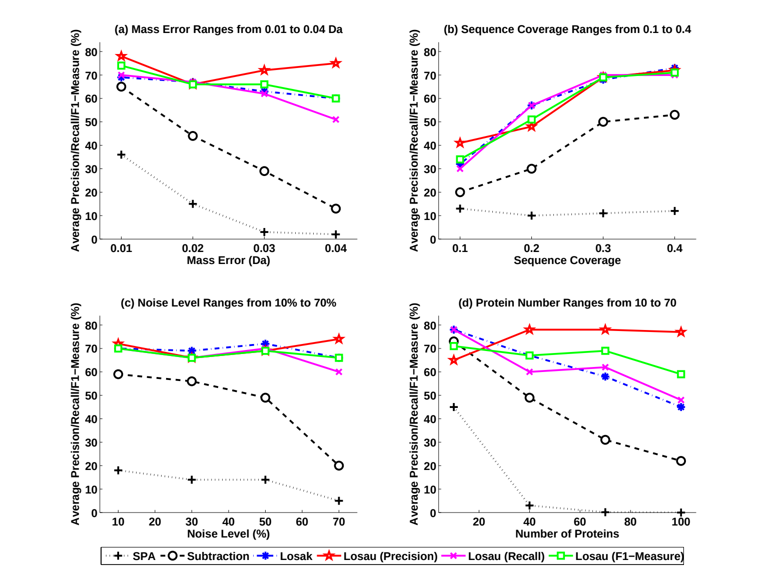
2 Method

3 Experiments
- Simulation Study
- Real Data

4 Conclusion

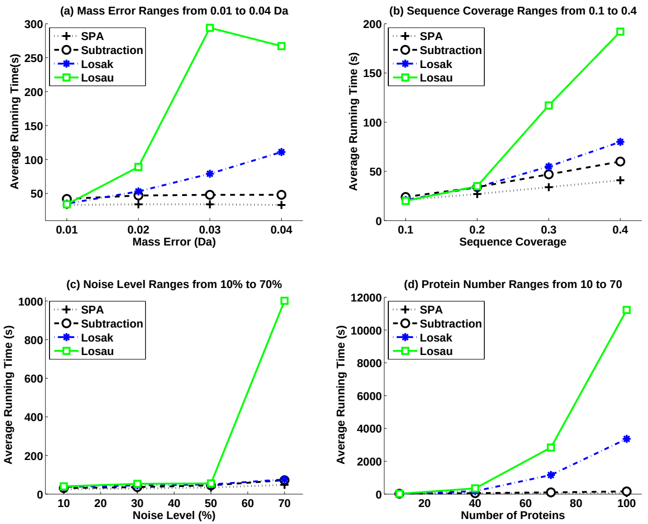
Performance Comparison

Performance Comparison



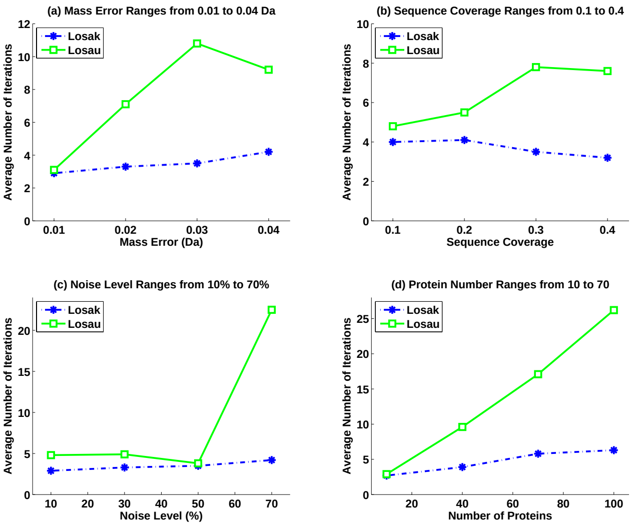
Running Time Comparison

Running Time Comparison



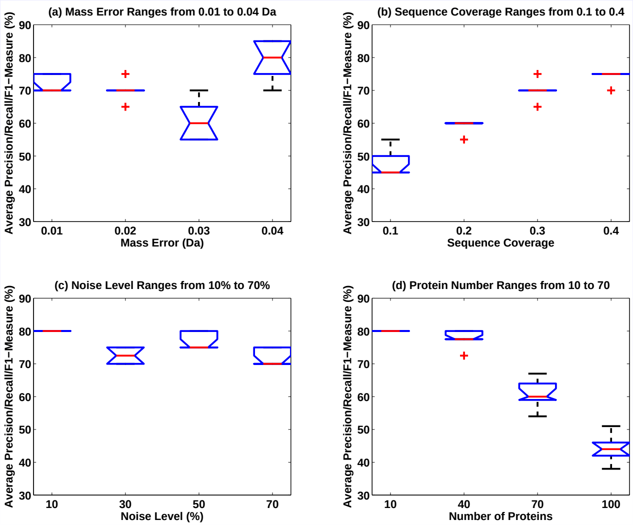
Number of Iterations of Our Algorithms

Number of Iterations of Our Algorithms



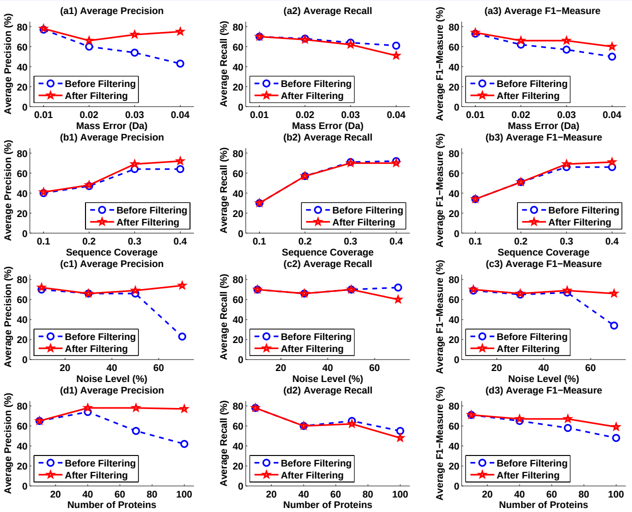
Sensitivity to Initialization

Sensitivity to Initialization



The Effect of Filtering in Losau

The Effect of Filtering in Losau



Outline

1 Introduction

2 Method

3 Experiments
- Simulation Study
- Real Data

4 Conclusion

Results on Real Data

Here we use a mixture of 49 standard human proteins in the ABRF sPRG2006 study.

Table: Identification performance and running time of different algorithms on the real MS data. Here the number of reported proteins for SPA, Subtraction, and Losak is 49, i.e. the number of ground-truth proteins.

Algorithms	Precision	Recall	F1-Measure	Running Time(s)
SPA	24%	24%	24%	7.9
Subtraction	43%	43%	43%	24.0
Losak	67%	67%	67%	21.2
Losau	61%	71%	66%	19.6

Conclusion and Future Work

- Optimization-based PMF methods have great potential for protein mixture identification, especially in the analysis of low-abundance proteins, whose peptide digestion results are less likely to be covered by the peptide sequencing method.
- We like to combine MS and MS/MS data to further improve protein identification accuracy and robustness.

Thank you !