Published in final edited form as:

Comput Biol Chem. 2007 June ; 31(3): 215–221. doi:10.1016/j.compbiolchem.2007.03.001.

A New Protocol of Analyzing Isotope Coded Affinity Tag Data from High Resolution LC-MS Spectrometry

Weichuan Yu^{a,*}, Junfeng Liu^b, Chris Colangelo^c, Erol Gulcicek^d, and Hongyu Zhao^e

^aDepartment of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China, Tel: +852 2358-7054, Fax: +852 2358-1485

^bDepartment of Statistics, West Virginia Univ., Morgantown, WV 26506, USA

^cKeck Biotechnology Laboratory, Yale University, New Haven, CT 06520, USA

^dKeck Biotechnology Laboratory, Yale University, New Haven, CT 06520, USA

^eDepartments of Epidemiology and Public Health and Genetics, Yale University

Abstract

Isotope Coded Affinity Tags (ICAT) is a labeling technique that provides insights into quantitative molecular changes. In this paper, we propose a new protocol to identify and analyze ICAT labeled peak pairs in high-resolution LC-MS data. Our major contributions are: 1. We use isotope distance constraint, ICAT distance constraint, and LC-span constraint to identify ICAT labeled peak pairs; and 2. We propose to trigger tandem MS/MS scanning based on the ratio estimation value of identified ICAT peak pairs instead of the peak intensity values. Compared with current approaches that choose peaks with high intensity values for tandem MS/MS scanning, the new protocol can improve the scanning efficiency and accuracy.

Keywords

Isotope labeling; mass spectrometry data analysis; tandem MS triggering; quantitative proteomics

1 Introduction

The ultimate goal of comparative proteomics using mass spectrometry (MS) techniques is to study the protein/peptide expression differences under different conditions (e.g. disease vs. normal, treatment vs. control, etc.). To this end, efficient identification of peptide related peaks in MS data and accurate quantification of peptide expression levels are two key issues.

*Corresponding author. ; Email: eeyu@ust.hk (Weichuan Yu)

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Supplementary data files, figures, tables, and MATLAB source code are available at: <http://www.ece.ust.hk/~eeyu/icat.htm>

In current MS based proteomics studies, however, these two issues are still very challenging due to the following reasons:

1. Mass spectrometry instruments are very sensitive to experimental variations (including parameter setting of instruments, operational temperature, sample preparation, and many other factors).
2. Biochemical and electronic noise is common in MS data.

In order to reduce the effect of experimental variations, Isotope Coded Affinity Tags (ICAT) was introduced (Gygi et al. (1999)) for comparative proteomics. The attractive property of ICAT is that the use of isotope produces a pair of chemically identical reagents with the heavy reagent weighting a few Dalton more than the light reagent on the m/z axis. When we label one test sample with a light reagent and another test sample with a heavy reagent, both reagents will bind on the same peptides. After we combine two samples together and feed them into the MS scanner simultaneously (to reduce experimental variations), reagent labeled peptides will show up as peak pairs in MS spectra. Their abundance difference revealed in the MS data may indicate the difference between light- and heavy-reagent labeled samples. Moreover, the fixed m/z distance in each peak pair is a very strong constraint for us to distinguish true peptide-related peaks from random noise. This is extremely helpful for us to detect peptides with low expression levels, which is one of the bottle-neck issues in analytic protein techniques (Han et al. (2001); Schmidt et al. (2004)).

The ICAT concept has been widely used after its introduction (e.g. Moseley (2001); Qian et al. (2003); Bottari et al. (2004); Conrads et al. (2004); Johnson et al. (2004)). Different software programs were developed to analyze ICAT labeled MS data (e.g. proICAT from Applied Biosystems, spectrum Mill from Agilent Technologies, Sashimi (including SEQUEST and XPRESS from Thermo Finnigan) from the Institute of System Biology, and ASAP Ratio by Li et al. (2003)). Their performances are also evaluated (Haller et al. (2003a,b); Moulder et al. (2005)), although the details of commercial software programs (such as proICAT) are not publicly available.

Possible applications of ICAT techniques include:

1. **Cancer biomarker discovery.** By labeling cancer samples with light reagents and normal samples with heavy reagents (or vice versa), we can quantitatively identify peptides that are differently expressed and use them as cancer biomarkers for diagnosis purpose.
2. **Drug effect study.** This will be the same scenario as cancer biomarker discovery if we replace cancer group with treatment group and replace normal group with control group.
3. **Development study.** We can compare samples of the same target at different time points by using ICAT reagents.

While ICAT has great potential in comparative proteomics, we also notice that some improvements are needed in current ICAT MS data analysis programs. In our current project, we need to automatically identify ICAT peak pairs from HPLC ESI-MS (High-

performance Liquid Chromatography ElectroSpray Ionization-Mass Spectrometry) data and estimate their ratio values. After applying the ASAP Ratio (Li et al. (2003)) and proICAT in the analysis, we found the following issues need to be addressed:

1. **Peak identification for high-resolution MS data.** Current methods (such as Li et al. (2003)) are based on simple local maximum search algorithm to detect peaks. This may result in many false positive peaks. We should be able to remove some false positive peaks by using additional constraints (such as the isotopic peak distance constraint) in high-resolution MS data.
2. **MS/MS scan triggering.** The MS/MS scanning is currently triggered either manually or by a large intensity value in the precursor MS spectrum. In the practice of quantitative proteomics, however, we are more interested in peptides whose expression levels are different under different conditions (such as disease versus control cases).
3. **ICAT ratio estimation.** Currently, the ICAT ratio is estimated as the ratio of the sums of peak pair intensity values or possibly after some smoothing (proICAT and ASAP Ratio). We like to consider the influence of noise and systematic measurement error from a bivariate analysis point of view.

To deal with these challenges, here we propose a **New Protocol for ICAT labeled LC-MS/MS Data Acquisition and Analysis**.

1. Divide the ICAT labeled test sample into two equal portions.
2. Run the first portion through the LC-MS under the survey mode (i.e. only using one MS scanning setting), identify ICAT labeled peak pairs and estimate their ratio values. One of the advantages of running samples only under the LC-MS survey mode is that we get much greater sampling across the chromatographic peak profile. This allows us to use an elution time check for increased peak identification confidence.
3. Run the second portion through the LC-MS under the survey mode. Switch to MS/MS at (m/z, LC) locations where the ratio values of identified ICAT peak pairs (see step 2) were not 1:1 and further identify the peptide sequences using database search.

The above protocol facilitates the analysis of ICAT MS data, making ICAT techniques closer to real applications. The rest of the paper is organized as follows: Section 2 describes an imaging and filtering method for the identification of ICAT peak pairs. Section 3 discusses ICAT pair quantification methods. Section 4 presents experiment results. Section 5 concludes the paper with some discussions.

2 ICAT Peak Pair Identification

For simplicity, we only take the example of +1 charge state when we discuss about the m/z distance. The principle is the same for multiple charges. The identification of ICAT peak pairs consists of the following five steps (shown in figure 1):

1. **Peak Identification.** For each spectrum, search for maxima in the local neighborhood as peaks.
2. **Isotope Distance Check.** This is accomplished in two steps:
 - a. Extract all peak pairs whose m/z distances satisfying the isotope distance constraint. Ideally, the distance between two neighboring isotope peaks should be 1 Dalton. In practice, we use a small window of $[1 - m_1, 1 + m_1]$ with m_1 denoting our belief on the isotope distance variation.
 - b. Identify three successive peak components with the distance of neighboring peaks satisfying the isotope distance constraint.
3. **Isotope Intensity Check (Optional).** Given intensity values I_1 , I_2 and I_3 for three successive isotope peak components, we consider them belonging to the same peptide only when $I_1 > I_2$ and $I_2 > I_3$ are simultaneously satisfied. Note that this is a relatively specific requirement as the data considered by us have a small m/z range (only up to about 1,600 Da). If the m/z range goes higher, we may have to relax the first condition since $I_2 > I_1$ happens sometimes at a higher m/z value due to the existence of multiple residuals.
4. **ICAT Peak Pair Identification.** Take the left most peak of isotope resolved peaks and identify peak pairs satisfying the ICAT pair distance constraint, i.e. the distance between ICAT pair should lie in the window of $[d_{icat} - d_{icat}^-, d_{icat} + d_{icat}^+]$, where d_{icat} is the ideal ICAT distance and $d_{icat}^\pm$ denotes possible variations.
5. **Elution Time Profile Check.** Under the LC-MS setting, the elution of peptide components takes some time. A true ICAT peak pair should have a span larger than a certain number along the LC direction (i.e. the time axis). In this paper, we used a relatively conservative span number of 10, while a thorough study of the influence of the span number may be of interest in future research.

In the following, we describe an imaging and filtering method to implement the above framework.

2.1 Imaging and Filtering Method

We consider the entire LC-MS data set as a 2-D image with the m/z value as x-axis, LC-index as y-axis, and spectral intensity values as image values. To convert the m/z values from floating numbers to integers, we construct a grid using the same sampling interval on the m/z axis for all spectra. For an arbitrary data point in one MS spectrum, we will keep its LC-index and intensity value, but change its m/z value to its closest m/z grid index (through a closest neighbor interpolation). In this paper, the sampling interval was set as 0.1 Da.

After the 2-D image is formed, we convert the ICAT pair identification problem to a filtering problem. Concretely, this is done through the **ICAT Peak Pair Filtering Algorithm**

1. **Image smoothing along the LC-axis.** We first apply a 1-D Gaussian function to smooth out the data along the LC-direction. The window size of the Gaussian function is determined based on the common knowledge of elution duration. Here we set the window size as 11.

2. **ICAT pair identification using a comb shape filter.** This step carries out the isotope distance check and ICAT distance check. Suppose we have 9 Da of ICAT distance and the charge state of ICAT pairs is +2. Assuming that the left most isotope peak of the light reagent labeled peak is located at i Da on the m/z axis, we will observe a pattern of $(i, i + 0.5, i + 1, i + 4.5, i + 5, i + 5.5)$ for ICAT peak pairs in the spectrum, with the first three components corresponding to the light reagent labeled peak and the last three components corresponding to the heavy one. Using a sampling interval of 0.1 Da, we have a filter mask covering 56 grid points with the 1-st, 5-th, 11-th, 46-th, 51-st, and 56-th components equal to one and the rest equal to zero. The comb shape filter can be used in different ways:
 - a. If we are interested in finding out highly-expressed peptides that are labeled by ICAT reagents (i.e. we are interested in large intensity values), we can directly convolve the filter with the image and search for high filter responses.
 - b. If we are only interested in finding out peptides that are labeled by ICAT reagents (which may not necessarily be highly-expressed), we may convolve the filter with a binary version of the smoothed image and search for filter responses that are equal to six (i.e. having six isotope components using the filter mask mentioned above).
3. **ICAT peak pair grouping** After ICAT pair identification has been carried out in individual spectra, we need to group together multiple ICAT pairs across the LC-axis due to the elution. Then, we give the median LC-axis value. This value together with the m/z value serve as the coordinates of identified ICAT pairs.

3 ICAT Intensity Ratio Estimation

For a certain peptide, suppose we obtain a set of ICAT peak pairs across multiple spectra (x_i, y_i) with $i = 1, \dots, N$, where x_i and y_i denote the intensity values of the left (light) and right (heavy) ICAT peaks, respectively, and N denotes the number of ICAT pairs. Two methods can be used to estimate the intensity ratio values:

1. Estimating ratio in each spectrum: $k_i = \frac{y_i}{x_i}$. The advantage is that we can trigger the ratio-based tandem MS scanning on the fly. The limitation is that the result may be sensitive to noise and we are not using all the information we have from the LC-MS data.
2. Estimating the ratio in a group-wise manner: We assume that all pairs (x_i, y_i) have the same ratio value.

Here we focus on the second option. A simple way to estimate the ratio using multiple spectra is to calculate the ratio of sum (or equivalently ratio of mean),

$$k = \frac{\sum_{i=1}^N y_i}{\sum_{i=1}^N x_i}, \quad (1)$$

where N denotes the number of ICAT peak pairs across multiple spectra. The assumption behind this idea is that noise has zero-mean and is additive to the measurement of x_i and y_i . Besides, there is no systematic measurement error.

In Li et al. (2003), Li et al. considered the possibility of having systematic measurement error and proposed a smoothing method to remove it. While the idea is interesting, it is unclear how the smoothing parameter was determined and how to evaluate the performance of the smoothing method.

In this paper, we reformulate the problem of estimating the ratio as a parameter estimation problem in a structural equation model (SEM) setting (Schumacker and Lomax (2004)):

$$\begin{cases} X = a_x + X_0 + \varepsilon_x \\ Y = a_y + kX_0 + \varepsilon_y \end{cases}, \quad (2)$$

where X and Y are observed intensity values of light and heavy peaks, respectively. We use a_x and a_y to denote the constant baselines (they are referred to as intercepts in SEM), X_0 and kX_0 to denote the true intensity values of light and heavy reagent labeled peaks, respectively, with k denoting their ratio value. The additive noise in both measurements are ε_x and ε_y , respectively. We assume that $\varepsilon_x \sim \mathcal{N}(0, \sigma_x)$ and $\varepsilon_y \sim \mathcal{N}(0, \sigma_y)$ and they are uncorrelated with X_0 and kX_0 .

Given a scatter plot of two variables X and Y (i.e. x_i vs. y_i), our goal is to fit a straight line (representing the linear relation) to the scatter plot and use the slope of the straight line as the estimate of the ratio value. This reminds us of the principal component analysis (PCA) method, which is also called Hotelling transform (Hotelling (1933)) or Karhunen-Love transform in the literature. In PCA, the direction of the principal component with the largest eigenvalue of the covariance matrix corresponds to the dimension that has the strongest correlation in the data set, while correlation is a quantitative measure of the strength of the linear relation between two variables. Thus, we may use a standard singular value decomposition (SVD) method to find the vector pointing to the direction where the scatter plot shows the largest variation (i.e. corresponding to the largest singular value) and then estimate the slope correspondingly as an estimate of the ratio. The success of the PCA method largely depends on the data structure: The more elongated the cloud of data points, the better the correlation (thus the estimation). We can use the ratio between the largest eigenvalue and the sum of all eigenvalues as a quantitative measure of the data quality:

$$r = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \quad (3)$$

The value of r is supposed to be close to 1 if the data has an elongated structure (i.e. has a strong linear relation). If the structure is not elongated enough, we have to switch to other methods. In this sense, the value of r serves as a quality measure of the data.

4 Experiments

4.1 Comparison of different ratio estimation methods

We used the SEM in Eq. (2) to synthesize three examples with known ground truth. The parameters are listed in table 1. We used two normally distributed functions with zero-mean to simulate noise. We repeated the synthesization 100 times and estimated the ratio values accordingly. Figure 2 shows one of the 100 scatter plots of observed peak pairs in each simulation. On top of the scatter plots, we also showed the intercept adjusted true pairs. Table 2 shows the mean and variance values of the estimation results using simple ratio of sum and PCA based method.

In above examples, the ratio of sum method is unable to deal with the existence of intercepts. The PCA based approach provides better performance when the data has an elongated structure (i.e. when the PCA score r is close to 1). But when the PCA score r gets smaller, PCA based method also starts to fail.

4.2 Real Examples

In this subsections, we compared our protocol with the proICAT method with respect to the following criteria: 1. the identification of correct ICAT peak pairs; 2. the number of possible ICAT pairs identified.

For validation purpose, we need data with known ground truth. Here we used the data in Moulder et al. (2005), in which only nine proteins with known ratio values were present. Since each protein can be digested into many peptides in different combinations, we cannot predict the m/z locations of ICAT pairs purely based on the prior information of proteins. Also, it is unknown where the ICAT pairs are located on the LC-axis. Here we approximately inversed the ground truth using the following rules:

1. We do not care about the LC-axis locations of identified ICAT pairs as long as their span on the LC-axis is larger than 10 spectra.
2. For an identified ICAT pair, if the tandem MS scanning shows that the corresponding peptide is from one of nine give proteins, then we consider its location as the ground truth.
3. The intensity ratio of ICAT peak pairs must be close to the protein ratio given in Table 1 in Moulder et al. (2005).

Let us first check the performance of proICAT. This program gives a quality score ranging from 100 (the best quality) to 0 (the worst quality). In total, proICAT identified 1679 peak pairs with positive quality score values. Among them, 111 peak pairs have quality scores higher than 90 and 373 peak pairs with quality scores higher than 80 (see supplementary data for details). In the comparison, we only consider the results with the quality score higher than 90 to guarantee the quality of the results.

We searched the tandem MS parts against the database and derived the relationship between peptides and proteins. Totally, we found 55 peptide-protein relations among these 111 identified peak pairs (The table is submitted as supplementary data). Figure 3 plots the

distribution of these peptides. Other peptides are not necessarily false positive identification results since the database search can only be carried out on those peptides with tandem MS spectra, while the triggering principle of tandem MS scanning in proICAT does NOT guarantee that all peptides labeled by ICAT reagents are scanned by tandem MS.

In our analysis, if a peptide is identified by proICAT as ICAT labeled and by proID as digested from one of the nine known proteins, then we consider it as the ground truth. For those peptides that are not identified by proID, we have to ignore them for the moment. In figure 3, we found 12 such peptides. In the following, we will compare the ratio estimation values of these 12 ICAT labeled peptides using different methods.

In our imaging and filtering method, we set the threshold value first as 10 and then as 5 for peak preservation and sparse matrix construction (step 2 in ICAT Peak Pair Filtering Algorithm). The smoothing filter along the LC-axis is a Gaussian with window size of 11. The ICAT peak pair distance is 9, 4.5, 3, 2.25, 1.8, and 1.5 Da for charge state +1, +2, +3, +4, +5, and +6, respectively.

Our major focus is to check whether our new method can provide higher sensitivity than the proICAT method and whether the ratio estimation in our method is more accurate than the proICAT method. Figure 4 plots the centers of identified ICAT pairs using our method (shown as dots) as well as the proICAT results (left, shown as circles) and the proID results (right, shown as circles). Our method totally found 3992 pairs after thresholding the image with 10. This number further goes up to 17781 in our method if we decrease the threshold value to 5. The Venn diagrams in figure 5 clearly shows that the majority of ICAT peak pairs that are identified by proICAT and proID are also identified by our method. Specially, the 12 ICAT peak pairs that are considered as ground truth in the proICAT method are all identified in our method.

It is possible that our method also introduces false positives since some noisy peak pairs may pass our check of distance constraints. But the point here is that we do NOT worry about false positives. Instead, we are afraid that the tandem MS scanning may ignore interesting peptides with low abundance values. Since we use tandem MS spectrum as the ultimate tool for protein/peptide identification, the influence of false positives can be largely reduced after checking the tandem MS spectra.

Now let us use the 12 ground truth pairs to compare the ratio estimation results. The linear relation between light and heavy peak values is very strong (see the plots in the supplementary data for details). Table 3 shows the ratio estimation results of these twelve peak pairs using three different methods: the ratio of sum, proICAT, and PCA method. Overall, the results using the ratio of sum method are the worst, while the performance of the other two methods are comparable. For non 1:1 cases, the proICAT seems even slightly better, although the relative estimation error still needs to be reduced. This result is consistent with the simulation results shown in the previous subsection (note that the proICAT method cannot be applied to simulation examples).

5 Conclusion and Discussion

In this paper, we proposed a new framework to identify ICAT labeled peak pairs from high resolution LC-MS data. We also proposed to directly trigger the tandem MS/MS scanning based on the ICAT ratio estimation result instead of the current peak intensity-based triggering method.

In order to identify ICAT labeled peak pairs, we proposed an imaging and filtering method. It is computationally more efficient than carrying out five constraints in a step by step manner. But it also requires that the relative shift of peaks between spectra is less than the sampling interval. If the requirement cannot be satisfied, then we have to align peaks together before carrying out the distance check (see Yu et al. (2006) for more details about peak alignment).

Precisely estimating the ratio of expression levels between light and heavy reagent labeled peaks is critical in quantitative proteomics research. While our principal component analysis based method provides improved performance than a simple ratio of sum approach, we also observe that the estimation precision of ratio values needs to be improved, especially for the ratio values that are not 1:1.

In our ICAT peak pair identification method, we make no assumption (or constraint) on the intensity values. In LC-MS data, we can observe a clear first increasing and then decreasing trend of ICAT peak intensity values. In our future work, we will explore using this trend to improve the identification of ICAT peak pairs and the estimation of their ratio values.

The method proposed in this paper is not limited to analyze ICAT data only. Similar labeling techniques such as iTRAQ (isotope Tagging reagents for Relative and Absolute Quantification) and SILAC (Stable Isotope Labeling with Amino acids in Cell culture) are also used in the proteomic quantification studies. We plan to extend the application of the same algorithm into the analysis of iTRAQ and SILAC data in our future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported with Federal Funds from NHLBI/NIH contract N01-HV-28186, NIDA/NIH grant P30 DA 018343-01, and NIGMS grant R01 GM 59507 in the U.S. and a RGC Direct Allocation Grant 2006/07 from the Hong Kong University of Science and Technology.

References

- Bottari P, Aebersold R, Turecek F, Gelb M. Design and synthesis of visible isotope-coded affinity tags for the absolute quantification of specific proteins in complex mixtures. *Bioconjugate Chemistry*. 2004; 15(2):380–388. [PubMed: 15025535]
- Conrads K, Yu L, Lucas D, Zhou M, Chan K, Simpson K, Schaefer C, Issaq H, Veenstra T, Beck G Jr, Conrads T. Quantitative proteomic analysis of inorganic phosphate-induced murine mc3t3-e1 osteoblast cells. *Electrophoresis*. 2004; 25(9):1342–1352. [PubMed: 15174057]

- Gygi S, Rist B, Gerber S, Turecek F, Gelb M, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*. 1999; 17:994–999.
- Haller PV, Yi E, Donohoe S, Vaughn K, Keller A, Nesvizhskii A, Eng J, Li X, Goodlett D, Aebersold R, Watts J. The application of new software tools to quantitative protein profiling via isotope-coded affinity tag (icatt) and tandem mass spectrometry: I. statistically annotated datasets for peptide sequences and proteins identified via the application of icatt and tandem mass spectrometry to proteins copurifying with t cell lipid rafts. *Molecular and Cellular Proteomics*. 2003a; 2(7):426–427. [PubMed: 12832456]
- Haller PV, Yi E, Donohoe S, Vaughn K, Keller A, Nesvizhskii A, Eng J, Li X, Goodlett D, Aebersold R, Watts J. The application of new software tools to quantitative protein profiling via isotope-coded affinity tag (icatt) and tandem mass spectrometry: II. evaluation of tandem mass spectrometry methodologies for large-scale protein analysis, and the application of statistical tools for data analysis and interpretation. *Molecular and Cellular Proteomics*. 2003b; 2(7):428–442. [PubMed: 12832459]
- Han D, Eng J, Zhou H, Aebersold R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnology*. 2001; 19:946–951.
- Hotelling H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*. 1933; 24:417–441. 498–520.
- Johnson M, Yu L, Conrads T, Kinoshita Y, Uo T, Matthews J, Lee S, Smith R, Veenstra T, Morrison R. Proteome analysis of dna damage-induced neuronal death using high throughput mass spectrometry. *Journal of Biological Chemistry*. 2004; 279(25):26685–26697. [PubMed: 15060066]
- Li X, Zhang H, Ranish J, Aebersold R. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Analytical Chemistry*. 2003; 75(23):6648–6657. [PubMed: 14640741]
- Moseley M. Current trends in differential expression proteomics: isotopically coded tags. *Trends in Biotechnology*. 2001; 19(Suppl):S10–S16. [PubMed: 11780964]
- Moulder R, Filen J, Salmi J, Katajamaa M, Nevalainen O, Oresic M, Aittokallio T, Lahesmaa R, Nyman T. A comparative evaluation of software for the analysis of liquid chromatography-tandem mass spectrometry data from isotope coded affinity tag experiments. *Proteomics*. 2005; 5(11): 2748–2760. [PubMed: 15952233]
- Qian W, Goshe M, Camp D II, Yu L, Tang K, Smith R. Phosphoprotein isotope-coded solid-phase tag approach for enrichment and quantitative analysis of phosphopeptides from complex mixtures. *Analytical Chemistry*. 2003; 75(20):5441–5450. [PubMed: 14714534]
- Schmidt F, Donahoe S, Hagens K, Mattow J, Schaible U, Kaufmann S, Aebersold R, Jungblut P. Complementary analysis of the mycobacterium tuberculosis proteome by two-dimensional electrophoresis and isotope-coded affinity tag technology. *Molecular and Cellular Proteomics*. 2004; 3(1):24–42. [PubMed: 14557599]
- Schumacker R, Lomax R. *A Beginner's Guide to Structural Equation Modeling*. Lawrence Erlbaum Associates, Publishers. 2004
- Yu, Li X, Liu J, Wu B, Williams K, Zhao H. Multiple peak alignment in sequential data analysis: A scale-space based approach. *IEEE/ACM Trans. Computational Biology and Bioinformatics*. 2006; 3(3):208–219. [PubMed: 17048459]

ICAT Peak Pair Identification

1 Peak identification

2 Isotope distance check

3 Isotope intensity check

4 ICAT peak pair identification

5 Elution time profile check

Fig. 1.
Overview of the ICAT peak pair identification framework.

Fig. 2.

Scatter plot of simulated peak pairs (from left to right: Example I, II, and III). Every circle represents an observation (x_i, y_i) and every star represents an intercept adjusted true pair $(a_x + x_{0,i}, a_y + y_{0,i})$. The parameters in the simulations are listed in Table 1. Note that the plot scales are different.

Fig. 3.

We plot the left peak parts of ICAT labeled peak pairs found by proICAT as crosses. The quality scores of these ICAT pairs are higher than 90. We also plot 55 peptides that are digested from nine known proteins as circles. The overlapping peptides are considered as ground truth in the comparison of ratio estimation later.

Fig. 4.

Left: We plot ICAT labeled peak pairs found by our method as dots and those found by proICAT as circles. **Right:** The same as the left plot except that we replace the proICAT pairs with peptides identified by proID. Enlarged plots are available as supplementary data.

Fig. 5.

Left: Venn diagram of numbers of identified ICAT peak pairs using proICAT, proID, and our method. The quality score in proICAT is higher than 90. The number in the circle marked as proID denotes the number of peptides identified by database search. In our method, we threshold the image with a value of 10. **Right:** Diagram after decreasing the threshold to 5.

Parameter settings in synthesized examples. The scatter plots of examples are shown in Fig 2. Here σ_x and σ_y are standard deviations of the error terms ε_x and ε_y , respectively. The term $\text{rand}(1,N)$ denotes a vector of N components following a uniform distribution between 0 and 1.

Table 1

Examp	N	k	a_x	a_y	X_0	σ_x	σ_y
I	40	5.0	10	20	$40 \times \text{rand}(1,N)$	4	4
II	30	0.2	30	40	$30 \times \text{rand}(1,N)$	10	2
III	40	2.0	20	20	$\text{rand}(1,N)$	0.5	0.5

Table 2

Estimation results of the ratio value in synthesized examples. We repeat simulation 100 times for each example. The scatter plots of examples are shown in Fig 2. In each column we show mean and variance values of the 100 results. PCA: Principal component analysis based method. The PCA score r is defined in Eq. (3).

Truth	Ratio of Sum	PCA	PCA Score
5.0	(4.00, 0.01)	(5.00, 0.06)	(0.94, 0.0001)
0.2	(0.95, 0.0021)	(0.09, 0.0013)	(0.85, 0.0009)
2.0	(1.02, 0.01)	(11.60, 8783)	(0.62, 0.0012)

Table 3

Comparing results with the ground truth. The columns from left to right: protein name, m/z value of the ICAT-labeled peptide, LC index of the peptide, true ratio value, ratio value using the "ratio of the sum" method (denoted as ROS), ratio value using the proICAT algorithm (denoted as proI), ratio value using the PCA method, and the quality score *r* in PCA method. The 9-th and 12-th peptides are actually the same based on their *m/z* and LC values.

protein	m/z	LC	truth	ROS	proI	PCA	r
Lactoferrin	689.38	1612	1	0.95	0.85	0.94	0.98
Lactoferrin	613.67	1829	1	0.95	1.06	0.92	0.96
Lactoferrin	622.37	927	1	0.96	1.03	0.97	0.98
Lactoferrin	775.41	1176	1	0.99	0.99	0.99	0.98
Lactoferrin	699.41	1267	1	0.90	0.94	0.99	0.97
Lysozyme	748.40	1378	0.5	0.32	0.37	0.29	0.99
Ribonuclease A	798.76	1434	2	2.09	2.01	2.04	0.95
Lactoferrin	551.29	1489	1	1.11	1.01	1.00	0.93
Beta-galactosidase	726.06	1595	0.2	0.11	0.15	0.09	0.99
Lactoferrin	509.30	963	1	0.95	0.95	0.93	0.98
Glycogen phosphorylase	643.05	1442	1.5	1.71	1.65	1.74	0.99
Beta-galactosidase	726.06	1596	0.2	0.03	0.04	0.00	0.99