Quantitative Proteomic Analysis System: *i*-OPAL

Mitsuhiro KANAZAWA^{a,b}, Hisae ANYOJI^a, Hsiao-kun TU^c, Umpei NAGASHIMA^{b,d*},
and Atsushi OGIWARA^c

^aReifys Inc., Level 9, Ariake 3-7-26, Koto-ku, Tokyo 135-0063, Japan

^bGraduate School of Pure and Applied Sciences, University of Tsukuba, Tenodai 1-1-1, Tsukuba, Ibaraki 305-8571, Japan

^cMedical ProteoScope Co., Ltd., Kameido 2-26-10, Koto-ku, Tokyo 136-0071, Japan

^dNanosystem Research Institute, National Institute of Advanced Industrial Science and Technology, Umezono
1-1-1, Tsukuba, Ibaraki 305-8568, Japan

*e-mail: u.nagashima@aist.go.jp

(Received: June 9, 2009; Accepted for publication: July 6, 2010; Advance publication: September 3, 2010)

Recently, significant improvements have been shown in proteomics, an analytical technology recognized as a method to discover biomarkers. Several methods for proteome differential display have been developed to discover biomarkers, and these are based on labeling methods such as ICATTM reagent. These labeling methods are not easily able to detect biomarkers in complex biological samples, and they have several issues such as the limitation of the modification site and the production of artificial errors. We therefore developed a computational quantitative proteomic analysis system *i*-OPAL (internal standard guided Optimal Profile Alignment) based on LC-MS measurements for biomarker discovery. This system can provide extensive quantitative analysis because it does not involve any artificial labels that might cause sampling bias and limitation of quantifiable proteins.

Keywords: Proteomics, LC-MS, Protein quantification, Peptide sequence

1 Introduction

Since the Human Genome Project [1] that decoded the whole human genome sequence as the master blueprint of a human has been completed in 2003, studies for the intravital mechanism, the sources of disease and the responses by medication and similar applications into medicinal research are expected. The main approaches for these researches are known and are described as post-genome. They are based on molecular identification and quantitative analysis for protein or sugar chain. Since better genome sequencing technology is achieved by DNA sequencers and computational algorithm progress, the developments for effective computational algorithm are also required for post-genome analysis that can utilize rapidly-advancing mass spectrometry (MS).

Proteomics is the study of proteins; it is now recognized as an approach for precisely comprehending the intravital

mechanisms in the post-genomic era. Unlike transcriptome analysis that observes the amount of RNA expression from genes, wherein gene transcription may or may not occur in vitro, proteomics directly facilitates qualitative and quantitative analyses of the protein in terms of in vivo gene expression noted at the time of analysis. Proteome analysis has been most commonly accomplished by a combination of two-dimensional gel electrophoresis (2DE) for protein separation and mass spectrometry for measurement of the proteins separated. However, this technique has been under scrutiny, because it failed to detect low-abundance proteins in whole cell lysate or blood plasma analysis. Several limitations of 2DE for proteome analysis can be overcome by combining liquid chromatography (LC) with tandem mass spectrometry (MSⁿ) and applying it to the direct analysis of protein mixtures [2].

LC-MSⁿ instruments generate MS and MSⁿ spectra of the observed molecules. To be precise, the MS spectrum is a snap

shot in the retention time (RT) of liquid chromatography; it consists of the mass-to-charge ratio (m/z) of the observed peptide molecule and the ion peak intensity. LC-MS measurements are represented by the time series of RT , and they construct a three-dimensional ion signal map of RT , m/z , and ion peak intensity throughout the measurement. The signals on this ion signal map are separated as molecular ions by liquid chromatography, based on their different charge states. Since protein mixtures are digested by trypsin in the typical proteomic analysis, the MS signals indicate trypsin digested peptides. MS^n signals indicate the weights of peptide fragments with cleaved internal bond (s) due to the collision energy in mass spectrometry. Here, "n" is the number of measurements recorded with bond cleavage, for example, the MS spectrum is the 1st measurement without a cleavage trial, the MS^2 spectrum is the 2nd measurement with one cleavage trial, and the MS^n spectrum is the n-th measurement with (n-1) cleavage trials.

There are two main methods for proteomic analysis using LC-MS: qualitative analysis and quantitative analysis. Qualitative analysis involves identification of the amino acid sequence from the measured protein by using the MS^n spectrum. Database searching approaches such as SEQUEST [3] or Mascot [4] are often employed for qualitative analysis in order to identify the proteins and peptides. These programs search for the molecular weights of the peptide or peptide fragments against the calculated peptide masses from protein sequence database. Quantitative analysis [5] involves the comparison of more than one protein mixture such as cell lysates or blood plasma samples; peptide ion signals whose expressions are significantly different among comparative groups are then detected as quantitative differences. This analysis searches for peptides derived from a protein in different samples first, and compares the amount of peptide distributed among the groups. Here, the groups could correspond to, for example, healthy vs. diseased individuals for the discovery of disease related proteins, or treated vs. untreated patients for the analysis of drug efficacy, safety or mechanism. Quantitative analysis identifies the target peptide molecules as the biomarker.

Labeling methods such as ICATTM reagent [6] have been generally employed for quantitative analysis. These methods modify a specific amino acid in a peptide with a mono isotopically or isotopically labeled reagent prior to LC-MS measurement, and use the weight differences between the artificial labels of the common peptide ion signals among the compared samples. The comparative samples are combined prior to MS measurement, and one compares the ion-signal intensities of the MS spectrum, which indicates the quantity

of the expressed peptide. However, these methods are only effective for the peptides that have specific amino acids, and they are thus not feasible for quantitative analysis. In addition to this limitation, as artificial bias might be caused by the labeling process, so the quantitative analysis method without any artificial labels should be discussed.

In this study, we introduced a new quantitative proteomic analysis system *i*-OPAL (internal standard guided Optimal Profile Alignment). This system can provide extensive quantitative analysis, because it does not involve the application of any artificial labels.

2 Methods

Our quantitative proteomic analysis system *i*-OPAL is designed to compare more than one group by multiple LC-MS measured samples, and the analysis results are a list of peptide signals whose expressions are significantly different among groups. This system utilizes the alignment technique for ion peaks to identify the common peptide signal in the ion signal map among LC-MS measurements whose data consist of RT , m/z and ion intensity, and then compares the ion-signal intensity based on alignment results. An image based method such as DeCyder MS [7] creates 2DE-like grayscale image by RT - m/z plane projection of ion intensity, and aligns gray color spots among samples. As it approximates high resolution of ion intensity by lower image resolution, its alignment speed is not a major issue. Quantitative accuracy is essential. On the other hand, ion peak alignment among samples requires huge computational calculations, if *i*-OPAL calculates all of the peak combinations among samples. In this alignment process, *i*-OPAL, therefore, utilizes some specific peptides as the Internal Standard (*IS*) to normalize the ion signal intensity of each measurement and to reduce alignment search space. SIEVE [8] is also known as quantitative proteomic analysis method using alignment technique, but the application of *IS* drastically reduces alignment search space and accelerates the alignment process, which is usually time consuming.

The followings are *i*-OPAL analysis phases, and the practice of those phases is to discover peptide ion signals whose quantity is significantly different among groups.

2.1 Signal normalization

The normalization of the signal intensity by each measurement is considerably important for quantitative analysis. The analytical accuracies of chromatography and mass spectrometry have been significantly improved. However, signal normalization

is still necessary to revise both the measurement time series fluctuations and the measurement batch fluctuations.

Labeling methods experimentally combine all the comparative samples to reduce these fluctuations, but experimental combining under the same quantities carries a risk of causing artificial bias. This system therefore observes the signal intensity variation in the *IS* peptides among the measurement samples as *RT* progresses: signal ion intensity is normalized using the appropriate normalization ratio, first by total ion intensity of each measurement, and then by the intervals between the initial *RT* of the measurement and next *IS*, its *IS* and following *IS*, ..., the final *IS* on the *RT* axis and the final *RT* of measurement (See Figure 1).

2.2 Ion signal alignment

Prior to comparing the peptide ion intensities among comparative groups, this system performs two alignments to identify the consensus signal among samples. One is intra-group alignment that involves alignment of ion signal maps in each comparative group on the *RT* - *m/z* plane, and the other is the inter-group alignment that involves the alignment of ion signals between the comparative groups on the *RT* - *m/z* plane. Quantitative analysis based on ion signal intensity compares the ion intensity values at the aligned coordinates, and it is desirable to have the maximum possible aligned.

First of all, intra-group alignment calculates the correlation of all the signals between each sample in the same comparative group to define the alignment order. If the measurement results of two samples indicate considerable variation, it is difficult for the alignment process to identify the consensus signals between the two measurement results. This system, therefore, defines the alignment order by sample similarities that are calculated by solving the minimum spanning tree problem against the calculated correlations between each pair of samples. If signal alignment is performed using considerably different samples, the number of signal coordinates in the aligned signal map could be approximately equal to the sum of all the signal coordinates in two different samples. However, this means that almost all aligned signals are constructed by one of the two samples. On the other hand, if signal alignment is performed using similar samples, the number of signal coordinates in the aligned signal map could be almost equal to the number of signal coordinates in a sample that has more signals than the others. This is the reason why the definition of alignment order among samples with a higher similarity is important for effectively obtaining aligned coordinates. After defining the alignment order, the system starts aligning all signals on a sample-by-sample basis

Figure 1. The conceptual spectrum for the illustration of signal normalization. The peaks in these spectra show the typical *m/z* of *IS* peptides. Signal intensities are normalized by each corresponding interval of *IS*s.

in each group.

Several alignment methods like Dynamic Time Warping (DTW) or Correlation Optimized Warping (COW) based on DP (Dynamic Programming) alignment approach⁹ have been developed and modified for chromatographic data. In biomarker discovery, comprehensive and high throughput method analyses are expected such as transcriptome, and the alignment approaches for LC-MS spectra are focused as one of solutions to realize these analyses. We, therefore, developed a new alignment approach based on DP to expand the chromatogram to three-dimensional data as ion signal maps constructed by all of the *RT*, *m/z* and signal intensity values.

Additionally, this alignment approach is also based on DTW and COW, and it is capable of nonlinear shifts on the *RT* axis. When a minimum distance or a high correlation is obtained between two similar spectra, they are aligned. Consider two univariate spectra *S1* and *S2*. Here, a plot is constructed with the *S1* spectrum in the x axis and *S2* spectrum in the y axis (See Figure 2). This approach constructs the shortest path such that the similar spectra *S1* and *S2* are linked. Construction of this shortest path is indicative of an appropriate signal alignment.

Figure 2 shows an example of a mapping grid and illustrates the alignment path. The alignment path *F* shows the reference time axes and is denoted as

$$F = \{[S1(k), S2(l)] \mid k = 1, \dots, K, l = 1, \dots, L\} \quad (1)$$

where *K* and *L* indicate the length of each *RT* axis, and each point *[S1 (k), S2 (l)]* in the grid describes the comparison

Figure 2. Alignment path between $S1$ and $S2$.

between the spectra $S1$ and $S2$ at $RT = k$ and $RT = l$, respectively. It is noteworthy that $S1(k)$ and $S2(l)$ do not indicate the signal intensity of the chromatogram but the MS spectrum constructed by the ion signal distribution.

When the alignment process searches the grid for the optimal path, it may skip some grid points. The construction of the optimal path is known as synchronization, and one spectrum is considered as the reference. Occasionally, more than one point in the skipped grid is aligned with the same point as the reference spectrum, and the optimal path of alignment is elongated along with the reference.

Most of the alignment programs establish some constraints to reduce the search space for analysis time reduction. As the simplest path constitutes the shortest trajectories from the first [$S1(1), S2(1)$] to the last [$S1(K), S2(L)$] of two extreme points, it is simple to consider the constraint that limits the path from deviating more than the $\pm M$ grid points from its simplest path. (See Figure 3a) This constraint is easily set for search space reduction, but it is difficult to control the alignment whose optimal path is beyond the tolerance M .

In the actual condition, the search space of this constraint is evaluated in the last search grid and the alignment path that exceeds beyond the tolerance limit is avoided to a certain extent (See Figure 3b). However, the essential search path occasionally reaches beyond the limits of these constraints owing to measurement fluctuation.

We therefore devised a novel alignment constraint into i -OPAL by using the IS to reduce the search space. The IS

a) Intuitive constraint by using the distance from the simplest path

b) General constraint of the DP alignment process (using the distance from the last search grid).

Figure 3. The general constraints of alignment search space: (a) Intuitive constraint of the DP alignment process (using the distance from the simplest path); (b) General constraint of the DP alignment process (using the distance from the last search grid).

peptides used in this case were applicable to either the same peptide or to the peptides different from those we used for signal normalization. If the IS peptides in every sample are regarded as the landmarks of alignment, the search space is split by its IS landmarks. This is because the coordinates of the IS peptide can be determined before the alignment phase.

Figure 4. Constraint by the *RT* intervals of *IS* peptides by using general constraint.

Figure 4 shows the reduced search spaces obtained using *IS* peptides with the constraint in terms of the distance from the simplest path between the initial *RT* of measurement, that is, *RT* = 0 and *IS*, *IS*, followed by *IS*, ..., the final *IS* and the final *RT* of measurement. The application of this constraint does not need to consider the alignment path beyond the tolerance limit.

In the case where there was no restriction on the slope of the optimal path, there were three permissive predecessors for the grid [*S1* (*k*), *S2* (*l*)], namely, [*S1* (*k* - 1), *S2* (*l*)], [*S1* (*k* - 1), *S2* (*l* - 1)] and [*S1* (*k*), *S2* (*l* - 1)]. The alignment path of the grid [*S1* (*k*), *S2* (*l*)], that is, is evaluated by these three predecessors, and the highest similarity score obtained between *S1* and *S2* in these predecessors is selected for optimal alignment.

The similarity score $E(S1(k), S2(l))$ on the grid [*S1* (*k*), *S2* (*l*)] is calculated based on the following equation during the alignment process.

$$E(S1(k), S2(k)) = \max \left\{ \begin{array}{l} E(S1(k-1), S2(l)) + e(S1(k), S2(l)) \\ E(S1(k-1), S2(l-1)) + e(S1(k), S2(l)) \\ E(S1(k), S2(l-1)) + e(S1(k), S2(l)) \end{array} \right\} \quad (2)$$

$$E(S1(k), S2(k)) = e(S1(k), S2(k))$$

where $e(S1(k), S2(l))$ is the local similarity score of the grid [*S1* (*k*), *S2* (*l*)], and $E(S1(k), S2(l))$ is the local cumulative similarity score of the grid [*S1* (*k*), *S2* (*l*)].

A noteworthy fact is that this alignment approach needs to manage not just the chromatogram but also the measured

MS spectra according to the chromatogram. Therefore, the similarity scoring function evaluates the signal distribution of each *MS* spectrum and calculates the similarity score $e(S1(k), S2(l))$ taking into account the scoring parameter for applying weight penalty and bonus, similar to the DP alignment algorithm.

After one calculates the cumulative score throughout the entire search space that is limited by *IS* constraints, the optimal path can be constructed from [*S1* (*k*), *S2* (*l*)] to [*S1* (1), *S2* (1)] through all the *IS* landmarks. After construction, this alignment process updates the coordinates of *RT* and *m/z* axes of the original ion map for each sample to create the aligned ion signal map. The consensus ion map construction in each comparative group after alignment is repeated in the comparative group, and it is referred to as intra-group alignment.

Inter-group alignment basically utilizes the same approach as intra-group alignment without respect to alignment order calculation. Since group consensus ion maps are calculated before this alignment, ion maps of each comparative group are aligned and new align-IDs are subsequently assigned as consensus alignment locations of all samples.

After these two alignment phases, all the *RT* - *m/z* coordinates of the aligned signal in all samples are treated as align-IDs, and the align-ID is the reference with original ion signal intensity of all measurement samples.

2.3 Spot detection

The aligned ion map contains multiple signals of the same peptide because a peptide is identified not on a unique snapshot of *RT* but in a continuous time range. The LC gradually starts eluting a peptide along with the gradient of concentration between water and acetonitrile, and it is known that the distribution of signals for a single peptide generally follows a Gaussian distribution.

This system applies the fitting function of its distribution shape to the aligned ion map and detects the signals of every peptide. Here, we refer to this distribution of every peptide signal as a spot due to the shape including *m/z* tolerance. On detecting the single peptide spots as signal distributions, all the align-IDs belonging to a single peptide spot among all samples are assigned a new, unique align-ID. Simultaneously, this system collates the ion intensities of each sample and each comparative group based on the new unique align-IDs. We termed this process spot detection, and after this process, the system prepares to compare the quantities among groups by each align-ID.

Note that the *MS*^{*n*} spectra are measured for a cleaved

peptide, and these are used for qualitative analysis. All the signals of a single peptide are located from all measurement ion maps, implying that MSⁿ spectra are also located based on their align-ids because of the relationship between the ion signal in MS measurement and the MSⁿ spectrum. Based on the relationship between the MS signal and MSⁿ signals, the peptide qualitative information can be identified and the quantitative difference of the peptide can be determined.

3 Results

We prepared two same human plasma samples (protein amount: 1 µg each), and injected six proteins of different concentrations in each of them. Both human plasma and the six injected proteins were purchased from Sigma-Aldrich Japan K.K. The two kinds of samples were measured 5 times each instead of preparing samples of each group. The injected six proteins and their fluctuations between the two sample types (Groups A and B) are shown in Table 1. The fluctuation shows the increase/decrease in the two groups. The base peak chromatogram of the collated LC-MS measurement results for each group and measurement instruments are shown in Figure 5. The

Table 1. Six injected proteins and fluctuations between sample groups.

	Group A	Group B	Fluctuation (A:B)
Lactoperoxidase (Bovine)	2.5 pmol	500 fmol	>
Peroxidase (Horse radish)	500 fmol	100 fmol	>
<i>b</i> -Casein (Bovine)	2 pmol	1 pmol	>
Ribonuclease A (Bovine)	500 fmol	500 fmol	=
Deoxyribonuclease I (Bovine)	500 fmol	1 pmol	<
Ovalbumin (Chicken)	500 fmol	2.5 pmol	<

difference between these base peak chromatograms is easily observed by the naked eyes, but it is difficult for the existing method to identify the protein or peptide that is attributable to

Figure 5. Piled base peak chromatogram of LC-MS measurements and measurement instruments. This chromatogram indicates the difference between Group A and Group B by base peaks. Each value indicates RT(min.) and observed *m/z* at representative peaks.

the difference in protein expression between the two groups. The results of *i*-OPAL were compared with the following two labeling methods. One (Label 1) is the ICATTM labeling method, and the other (Label 2) is the Shimadzu ¹³C NBS Stable Isotope Labeling Kit-N [10]. Both labeling methods are commercially used as general quantitative analysis methods.

3.1 The number of quantifiable proteins

First of all, we examined the total number of identified proteins using each method. This is derived from not only the six injected proteins but also the proteins that are already present in the human plasma sample. As for this protein identification, we used Matrix Science Mascot 2.1 against the Swiss-Prot database and each identified protein was then judged by peptide score ≥ 30 belonging to each protein hit.

Table 2 shows the total number of identified proteins using each method. Here, the results of labeling 1 and 2 indicate the proteins that detected by both the artificial light and heavy label.

The standard experimental protocol for the labeling methods involves the extraction of the non-quantifiable peptides from samples. This protocol could be important when focusing on the quantification of peptides, however, it restricts the variety of quantifiable protein or peptides. Since *i*-OPAL does not any involve peptide extractions in the experimental protocol, 161 proteins were identified and all the identified proteins were quantifiable. Therefore the number of the identified and the number of the quantified is the same. On the other hand, the total number of obtained quantifiable proteins in the labeling methods was relatively few in spite of the existence of the six injected proteins that supposed to be quantified.

3.2 Limitation of the labeling methods

We confirmed the number of quantifiable proteins in six injected proteins (See Table 3). Labeling methods modify a specific amino acid by an artificial chemical label, and the experimental and measurement processes also reduce the number of detectable peptide fragments to focus the quantization of the

Table 2. The total number of identified proteins.

	Label 1	Label 2	<i>i</i> -OPAL
Identified protein	21	92	161
Quantifiable protein	20	16	161

label modified peptide.

On the other hand, *i*-OPAL does not have any limitations in terms of the modification sites, and the number of detected quantifiable proteins mainly depends on the limitation of the experimental samples and LC-MS measurement.

No the quantifiable methods is able to detect the exact quantified ratios between the two groups, but all the directions of fluctuation are exactly matched. On the discussion of quantitative fluctuation, the number of quantifiable protein is important and *i*-OPAL has advantage in this point.

Computer power is a limitation of *i*-OPAL because both *m/z* and intensity of all peaks are considered simultaneously. Computer power, however, was rapidly improved recently.

4 Conclusion

We proposed a computational quantitative analysis approach, namely, *i*-OPAL, which does not use any artificial labels. The application of the existing labeling methods restricted quantifiable peptides to specific amino acids that have specific chemical modification sites. The variety of quantifiable proteins is restricted by the existing labeling methods.

Since *i*-OPAL does not use any artificial labels, it realizes unrestricted peptide detection and can detect peptide molecules that have significant difference of expressed amount between the groups based on peptide signals by LC-MS measurements. We believe that *i*-OPAL will be quite useful tool for biomarker discovery, and it could also facilitate understanding of personalized medicine.

Table 3. The number of quantifiable peptides in each protein.

	Label 1	Label 2	<i>i</i> -OPAL
Lactoperoxidase (Bovine)	6	3	13
Peroxidase (Horse radish)	1	—	2
β -Casein (Bovine)	—	—	1
Ribonuclease A (Bovine)	2	—	2
Deoxyribonuclease I (Bovine)	—	—	1
Ovalbumin (Chicken)	3	—	8
The number of protein quantified	4/6	1/6	6/6

The computational quantitative proteomic analysis system *i-OPAL* has been developed and its patent is held by Medical Proteosope Co., Ltd., Tokyo, Japan. We are grateful to Shimadzu Corporation, Kyoto, Japan for providing us for free the ^{13}C NBS Stable Isotope Labeling Kit-N. We also thank Dr. Takao Kawakami and Dr. Toshihide Nishimura from the Clinical Proteome Center, Tokyo Medical University, for devising the experimental and measurement approach for *i-OPAL*.

References

- [1] F. S. Collins, M. Morgan, A. Patrinos, *Science*, **300**, 286 (2003). [Medline] doi:10.1126/science.1084564
- [2] J. Peng, S. P. Gygi, *J. Mass. Spectrom.*, **36**, 1083 (2001). [Medline] doi:10.1002/jms.229
- [3] J. K. Eng, A. L. McCormack, J. R. I. Yates, III, *J. Am. Soc. Mass. Spectrom.*, **5**, 976 (1994). doi:10.1016/1044-0305(94)80016-2
- [4] D. N. Perkins, D. J. Pappin, D. M. Creasy, J. S. Cottrell, *Electrophoresis*, **20**, 3551 (1999). [Medline] doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2
- [5] G. J. Boulnois, R. Aebersold, *Trends in Biotechnology*, **18**, 31 (2000). [Medline] doi:10.1016/S0167-7799(99)01393-1
- [6] S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, R. Aebersold, *Nat. Biotechnol.*, **17**, 994 (1999). [Medline] doi:10.1038/13690
- [7] M. S. DeCyder, URL: <http://www.gelifesciences.com/>.
- [8] SIEVE, URL: <http://www.thermofisher.com/>.
- [9] V. Praydova, B. Walczak, D. L. Massart, *Anal. Chim. Acta*, **456**, 77 (2002). doi:10.1016/S0003-2670(02)00008-9
- [10] H. Kuyama, M. Watanabe, C. Toda, E. Ando, K. Tanaka, O. Nishimura, *Rapid Commun. Mass Spectrom.*, **17**, 1642 (2003). doi:10.1002/rcm.1100