iTRAQPak: an R based analysis and visualization package for 8-plex isobaric protein expression data

Mark D’Ascenzo, Leila Choe and Kelvin H. Lee
Advance Access publication date 13 February 2008

Abstract
The field of high throughput proteomics has spawned a number of mass spectrometry-based technologies, which enable the quantitative analysis of protein expression. One of these technologies is iTRAQ (trademarked by Applied Biosystems), which through the use of isobaric tags, enables the quantitation of up to eight complex protein samples in a single multiplexed analysis. Isobaric tagging methods are emerging as an important tool to study protein expression dynamics. In this report, we describe iTRAQPak, a free software package developed in the R statistical and visualization environment that can be applied to the analysis of 8-plex expression data. The utility of this package is demonstrated through its application to the analysis of 8-plex iTRAQ protein expression data obtained from cerebrospinal fluid samples from Alzheimer’s disease subjects involved in a Phase I drug trial.

Keywords: iTRAQ; data analysis; shotgun proteomics; Alzheimer’s disease; immunotherapy; cerebrospinal fluid

INTRODUCTION
The rapidly expanding field of proteomics is advancing the ways in which protein expression dynamics can be studied. Contributing to these advances is the proliferation of mass spectrometry (MS) based shotgun proteomics methods that have been introduced in recent years. Shotgun methods allow the rapid profiling of complex protein mixtures by coupling high-resolution separation methods, such as HPLC, with the accurate quantitation and identification capacity of MS-based technologies. There are a number of methods to quantify protein expression from shotgun experiments including those based on isotope-coded affinity tag and isobaric tag for relative and absolute quantitation (iTRAQ™) technologies [1].

One of the more recently introduced methods is iTRAQ (trademarked by Applied Biosystems), which permits multiplex quantitation of up to eight complex protein samples in a single analysis. The experimental workflow for isobaric tagging based quantitation is similar to traditional single sample HPLC-MS based quantitation methods [2], however multiplexing is achieved through the use of isobaric labelling reagents that allow multiple samples to be pooled and quantitated independently. The first version of the technology permitted four samples to be multiplexed and the newer versions permit eight sample multiplexing [3]. The 4-plex reagents consist of four reporter ions (isobaric tags) which are designated as: 114, 115, 116 and 117, and the 8-plex reagents consist of these four, plus four additional tags, designated as: 113, 118, 119 and 121. Briefly, unlabelled protein samples are trypsin-digested, labelled using isobaric tags, pooled, then separated by liquid chromatography (perhaps multiple dimensions), and finally, peptides are quantified and sequenced by tandem MS (MS/MS). The
isobaric tags covalently bind to the N-terminus and lysine (Lys) side chain of peptides during labelling and enable multiplexing because they each have the same charge and overall mass, but produce different low mass signatures upon MS/MS [4, 5]. This unique characteristic allows otherwise identical peptides from different samples to be detected as a single peak by MS and produce a single set of sequencing ions in MS/MS, while maintaining the quantitation information from the different samples. Absolute and relative quantitation is achieved by determining the MS/MS spectra peak areas associated with each of the reporter ions and comparing them.

As shotgun methods continue to be applied to scientific research studies, it is also important to develop tools capable of analysing the data they generate. In this report, we describe the features, and demonstrate the application, of one such tool, which is intended for use with 8-plex expression data. The tool is developed as a package for use in the R statistical programming environment, and is called: iTRAQPak. It performs routine data transformation tasks associated with isobaric tag-based shotgun proteomics expression data analysis, and also implements more complex analytical, statistical and visualization functionality that may allow important biological relationships to be identified. We apply this tool to the analysis of 8-plex expression data collected in association with a longitudinal study of two Alzheimer’s disease (AD) patients undergoing a passive immunization treatment as part of a Phase I drug trial. The results of this analysis are used to demonstrate the utility of several iTRAQPak functions and highlight its visualization approach, which provides a novel view of 8-plex expression data.

IMPLEMENTATION AND OVERVIEW

The iTRAQPak package was developed in the R programming language [6]. R is a rich statistical, data analysis and visualization environment that is widely used and freely available under the GNU General Public License. iTRAQPak has been developed and tested under the Windows computing environment (Microsoft Windows Server 2003), however R is available under a variety of operating system environments. Package functions are run from the R command line.

Expression data are imported into R using the iTRAQPak function `LoadData`. The function expects data to contain a number of required variables; these are shown in Table 1. Expression data with corresponding variable columns can be generated through the use of software such as GPS Explorer (Applied Biosystems) with Mascot (Matrix Science) integration. While the set of required variables is fixed, the columnar order of the input variables is customizable through command line parameter options, which are accessible during data import.

The package provides a number of data transformation options, which can be selectively applied to imported datasets using user supplied parameter options. Transformation options include: isotope impurity correction, sample normalization, peak scaling and log transformation. Some form of impurity correction is recommended by the manufacturer for iTRAQ data analysis because the labelling reagents contain trace levels of isotopic impurities that cause variations in the MS peak intensities. Applying the impurity correction transformation to a dataset adjusts peak intensities as specified in the ‘Certificate of Analysis’ by the reagent manufacturer. Table 2 defines the default peak area correction factor applied to each peak type; however these values are customizable by modifying parameter input supplied to the package. Various normalization

<table>
<thead>
<tr>
<th>Table 1: Required columns and corresponding R data types</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Column</strong></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>21</td>
</tr>
</tbody>
</table>
procedures can be applied to sample data, and these are intended to correct between sample variation that may arise due to experimental procedure, rather than from biological differences. Normalization methods that can be applied are based on: (i) sample mean, (ii) sample median and (iii) Lowes regression [7]. Applying peak scaling expresses peak areas as a relative quantitation rather than absolute, and parameter settings allow up to two of the tags to be specified as a baseline for relative comparison. For example, with the default peak scaling parameters, areas associated with tags 115, 117 and 119 are expressed relative to tag 113 associated areas, and the remaining areas are expressed relative to 114 associated areas. Finally, scaled data can be Log2 transformed in preparation for statistical analysis or to represent expression increases and decreases as positive and negative values to allow for a more intuitive visual interpretation when viewed graphically.

When several peptides with identical sequences are identified during quantitative shotgun data analysis, one strategy is to express their peak areas as a single averaged area. The advantage of averaging peak areas is that expression values are derived from more than one observation and, additionally, dataset complexity is also reduced, thus allowing for a more simplified interpretation. However, condensing data using such an approach over-simplifies the complexity related to post-transcriptional and post-translational gene expression [3]; and as a consequence, important biological information can be lost. While peptides may be indistinguishable at the amino acid (AA) sequence level, they may be distinguished when post-translational modifications (PTMs) or other modifications are considered. Ignoring this difference may result in important biological relationships being overlooked. Further, two peptides, identical at the sequence level, may actually be from two highly homologous genes that have different biological roles. Averaging, in this case, makes little sense from a biological perspective.

With this in mind, iTRAQPak implements a multitiered strategy to avoid such loss of information by representing data at an uncondensed level plus three condensed levels: (i) the peptide level, (ii) the peptide-modifications level and (iii) the protein level. Data are condensed by averaging grouped peak areas defined by these levels. At the peptide level, peak areas are distinguished by AA sequence, and similarly, at the peptide-modifications level, peptides are distinguished by sequence and modification. At the protein level, peak areas are distinguished by accession number (GI). Finally, at the uncondensed level, as the name implies, no averaging is performed.

Peptide expression maps are generated using the iTRAQPak PlotExpression function. The expression maps present several different views of expression data, where the data are condensed to different levels within several of the views. Uncondensed expression data are displayed using two different view styles. The first displays expression data using heat maps and the second, using expression plots. Heat-maps are widely used to represent expression change, for example, red/green heat-maps are commonly used to represent microarray gene expression data, where red typically represents an increase in expression and green, a decrease in expression [8]. A similar red/green strategy is used to visually represent relative log-fold changes in peptide expression within iTRAQPak peptide expression maps. Expression plots are displayed as an alternative to the heat-map representation, but rather than using colour to represent expression change, two dimensional line plots are used, where the y-axis represents relative log-fold changes in peptide expression. For both formats, the uncondensed view displays expression values of every peptide detected. Expression plots are also used to represent condensed views for both peptide and protein level expression data and are a standard display within expression maps. Because there are a number of modifications that may be of interest in a research study, condensed peptide-modifications level data can be optionally displayed within the heat-map view, using user supplied parameters.

Expression map functions also allow peptide-modifications level expression data to be presented in an uncondensed form. Here, the uncondensed peptides displayed in the heat-map view are simply highlighted. Specifying modifications of interest is

<table>
<thead>
<tr>
<th>TAG</th>
<th>-2</th>
<th>-1</th>
<th>+1</th>
<th>+2</th>
</tr>
</thead>
<tbody>
<tr>
<td>113</td>
<td>0</td>
<td>2.5</td>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>114</td>
<td>0</td>
<td>1</td>
<td>5.9</td>
<td>0.2</td>
</tr>
<tr>
<td>115</td>
<td>0</td>
<td>2</td>
<td>5.6</td>
<td>0.1</td>
</tr>
<tr>
<td>116</td>
<td>0</td>
<td>3</td>
<td>4.5</td>
<td>0.1</td>
</tr>
<tr>
<td>117</td>
<td>0.1</td>
<td>4</td>
<td>3.5</td>
<td>0.1</td>
</tr>
<tr>
<td>118</td>
<td>0.1</td>
<td>2</td>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>119</td>
<td>0.1</td>
<td>2</td>
<td>4</td>
<td>0.1</td>
</tr>
<tr>
<td>121</td>
<td>0.1</td>
<td>2</td>
<td>3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 2: Default correction factors
achieved by supplying plotting functions with a parameter list of modification-codes. Parameter functions also allow peptides that lack a particular modification to be highlighted. As an example, it is possible to highlight peptides that possess a Lys residue but lack a [K]-iTRAQ tag. In the current implementation (1.0.7), it is possible to highlight peptides with these modifications: Oxidation, MMTS, [K]-iTRAQ and [N-Term]-iTRAQ. Additionally, because the plotting functions search the Modifications column of the input dataset, modifications that can be specified are limited to those present in this column.

The use of heat-maps in the peptide expression maps is a method implemented to reduce the visual complexity of the underlying expression data. Another method that is implemented is self-organizing map (SOM) clustering, which is widely used as a data visualization method for mapping high-dimensional non-linear data into a lower dimensional visualization space. Applied to expression plots, it allows complex expression patterns to be colour coded based on similarity. The package implements SOM clustering to colour code peptide expression patterns using the SOM package available in R.

Several data quality functions are implemented in the iTRAQPak package, two of which are: TagSummary and LabelingEfficiency. When applied to the raw expression data, TagSummary applies several transformation steps to the expression data, groups it by isobaric tag, and plots the transformed data to PDF. The PDF output consists of five plots, the first shows Log2 raw data (i.e. the data are not normalized, scaled, nor corrected, but are Log2 transformed) peptide expression values grouped by isobaric tag, the second shows the effects of scaling and the remaining three show the effects of mean, median and Lowes regression normalization. The LabelingEfficiency function uses peptide sequence and modification information to evaluate the efficiency of labelling reactions and then plots the results. It is assumed that under ideal conditions, the labelling reactions will go to completion such that all N-terminal and Lys residues are labelled with an iTRAQ tag. This may not always be the case under normal experimental conditions, thus it is possible to quantify the abundance of partially and fully labelled peptides. A partially labelled peptide is defined as one that has been labelled with at least one isobaric tag, but contains other possible labelling sites that have not been labelled. As an example, a peptide may contain a Lys residue in addition to its N-terminal residue; if it were partially labelled, then perhaps only the Lys group would be labelled. Using peptide sequence information, it is possible to assess which peptides contain Lys, in addition to N-terminal, residues, and using high confidence search results, it is possible to assess which peptide residues have been labelled by an isobaric tag. Using this information, a relative calculation can be made by summing the areas of all partially labelled peptides and expressing this value as fraction of the summed area of fully labelled peptides.

METHODS

To assess the utility of the iTRAQPak package, cerebrospinal fluid (CSF) samples from two probable AD subjects were analysed using 8-plex technology and methods as described previously [3]. Both of these subjects were enrolled in a Phase I drug trial investigating the effects of intravenous immunoglobulin as an AD treatment. As part of a longitudinal study, both patients received ongoing treatment with intravenous immunoglobulin (IVIg) and CSF samples were collected (with appropriate consent) by lumbar puncture from each patient over the course of the study.

In this study, four CSF samples from each AD patient were collected for protein expression analysis. These samples were collected at four different time points: at a baseline timepoint at the beginning of the drug treatment regimen, after 3 months and 6 months of treatment and then after a subsequent three month drug washout period (i.e. after 0, 3, 6 and 9 months). CSF was collected and analysed as described previously [3].

In the R programming environment, the iTRAQPak package was imported and several package functions applied to the isobaric tag-based protein and peptide expression data. As the first step of the analysis, the expression dataset was imported into R using iTRAQPak’s LoadData function. Next, using the TransformData function, peptide expression data were corrected for isotopic impurities, median normalized and scaled. Using scaling parameter options, peptide expression values were scaled relative to the ‘0’ time point (TI) for both patients. Also using the TransformData function, scaled values were then expressed as Log2 values and subsequently used for statistical and visualization analyses.
Statistical analyses applied in this study consisted of three-way analysis of variance (ANOVA) and was performed in R using the `lm` and `anova` functions. Peptide expression maps were generated using the `PlotExpression` function. To highlight peptides containing modifications of interest, the `pepMod` function was applied to expression data and expression maps were regenerated. Labeling Efficiency plots were created using the `LabelingEfficiency` function. Details on how to invoke package functions and parameter details are supplied within the iTRAQPak help files.

RESULTS AND DISCUSSION

The 8-plex iTRAQ dataset contained expression values for 1187 peptides (GPS ion score confidence interval 95% or greater) derived from 167 proteins (false positive rate of 0.8%). A complete list of proteins and peptides is provided in [3]. Here, we report on the development of the iTRAQPak software package and its application to this dataset.

Expression data were imported into R, and several transformation steps were applied to the imported data as described in the Methods section. Mean peptide expression levels for each of the labelled samples were inspected before and after normalization (Figure 1). Before normalization, the mean expression values for peptides labelled with the 113 tag were noticeably lower than the mean expression values for the peptides labelled by the other 7 tags. After normalization, this difference was less noticeable. However, when inspecting the scaled values, it was observed that the expression values for samples scaled to the 113 tag labelled sample were more variable than the expression values scaled to the 114 tag labelled sample. This was also observed for mean and Lowes regression normalization methods (data not shown). While normalization corrected for some of the between sample variability, it did not correct for all, suggesting a more complex level of variability exists for this data.

Using iTRAQPak functions, peptide expression maps were created for each unique protein identified in the normalized and transformed dataset. As an example, the map for albumin is shown in Figure 2. Maps were output from R as high-resolution PDF files, which allowed them to be easily viewed using Acrobat Reader (Adobe Systems). Acrobat Reader’s search functions were useful to quickly search for GI numbers of interest, and additionally, the zoom and scroll features allowed the features of the plots to be more closely inspected. Detailed views of heat-map and expression plots for albumin are shown in Figures 3 and 4, respectively.

The heat-map was found to be an effective method for viewing longitudinal expression changes. For larger proteins (>1200 AA) or very small peptides (<10 AA), it was found that plotting space was noticeably limited, causing peptide blocks to be somewhat compacted. It was possible to compensate for this by plotting the PDF output with wider page dimensions (>13 inches) or by significantly increasing zoom factor. Both the peptide heat-maps and paired expression plots allowed expression profiles to be easily compared. Additionally, the SOM...
colouring allowed otherwise complex expression patterns to be more easily interpreted.

The importance of incorporating multilevel views into peptide expression plots was made evident after inspecting the heat-map and expression plots. Figure 3 shows several albumin peptides with several opposing expression patterns, especially for patient A. The N-terminal peptides (in the 10–50 AA region), for example, show an increase in expression over time ($T_1$–$T_2$), while others show a strong decrease between $T_1$ and $T_2$. When averaged at the peptide level, these opposing expression trends are lost, favouring the decreased expression pattern. Figure 4 shows a more detailed view of the observed albumin peptides with data shown as expression plots for individual peptides. The expression change is plotted log base 2 such that increases in expression have a positive slope and decreases in expression have a negative slope.

The `pepMod` function was applied to the expression data using parameters to highlight peptides lacking an N-terminus tag in a newly generated expression map. Shown in Figure 5 is a zoomed and cropped section of the image of the albumin protein result that was generated. Visual inspection of the output showed a number of highlighted peptides allowing us to determine which of the peptides shown in the heat-map were lacking an N-terminus tag. Interestingly, it was observed that many of the peptides lacking an N-terminus tag also showed a strong decrease in expression over time (0–9 months) for patient A, but this was not observed for patient B.
To determine if this relationship was statistically significant, a three-way ANOVA was applied to the expression data for the peptides highlighted in Figure 5. A highly significant interaction was found between the patient and time terms ($P < 0.0001$), and a highly significant three-way interaction ($P < 0.0001$) between the patient, time and modification terms. These results show that the expression pattern differences between Subject A and Subject B are significantly correlated with the lack of an N-terminus tag. This strongly suggests that the different expression patterns seen between Subject A and Subject B are due to a labelling effect, rather than a biological effect.

To investigate this further, the LabelingEfficiency function was applied to the expression dataset to determine which if any of the tags may be contributing to the observed trends in Figure 5. The resulting plot from the application of this function is shown in Figure 6. The plot shows a noticeable difference in the relative area of unlabelled/labelled peptides associated with the 113 labelled sample as compared with the samples labelled by the other seven tags. From this, we can determine that there are relatively more partially labelled peptides in the 113 labelled sample as compared with the other seven samples. Experimentally, all steps were performed similarly among the different samples and labels. However, this observation adds support to the possibility that the expression patterns observed in Figure 5 may be due in part to a labelling effect rather than a biological effect. Further, the observed drop in expression between 0 and 3 months, for patient A, in peptides lacking an N-terminus tag, also supports the observation that unlabelled peptides are more abundantly represented in the 113 labelled sample.
CONCLUSION

We developed the iTRAQPak software package to aid in the visualization and interpretation of 8-plex shotgun proteomics expression data generated from a four time-point longitudinal study involving two patients. The isotope impurity and transformation functions available in this package facilitate rapid analysis of complex expression data. The iTRAQPak expression maps enable expression data to be viewed on multiple levels using methods that reduce visual complexity. We used the peptide expression maps to reveal complex patterns of expression between peptides identified as being fragments of same protein and even peptides with identical sequences. This observed complexity of expression further emphasizes the importance of interpreting condensed expression data with caution. Using the iTRAQPak modifications highlighting feature we observed a visual correlation among expression patterns from peptides lacking an N-terminal tag and also found this correlation to be statistically significant and possibly related to the labelling efficiency of 113 isobaric tags in these experiments. While we have used this package’s functionality to identify likely factors contributing to experimental variation within our dataset, the approaches demonstrated here are equally applicable to the identification of expression trends that are of important biological significance. Finally, although iTRAQPak was developed for 8-plex datasets, it may be applied to 4-plex datasets as well (one can input a 1 into the data column of unused tags which appears as a 0 on a log axis). Data from electrospray based analyses could be incorporated if formatted appropriately in Microsoft Excel. iTRAQPak is freely available for non-commercial purposes and can be downloaded from the Comprehensive R Archive Network (CRAN): http://cran.r-project.org/. This is most easily achieved from the R interface, first by selecting ‘Install Packages’ from the ‘Packages’ menu bar, and then selecting iTRAQPak from the packages list.

**Key Points**

- A visualization tool is presented for the analysis of 4-plex and 8-plex shotgun proteomics data.
- The visualization tool is extended to facilitate the quantitative analysis of modified peptides.
- The visualization tool is applied to the analysis of CSF proteins from Alzheimer’s patients undergoing passive immunization.
Acknowledgements
The authors thank Dr Norman Relkin for organizing the clinical trial and for sharing CSF. The authors would like to thank support from the Institute for the Study of Aging and NIH.

References