Supplemental material for

A data processing pipeline for mammalian proteome dynamics studies using stable isotope metabolic labeling

Shenheng Guan¹*, John C. Price², Stanley B. Prusiner², Sina Ghaemmaghami², and Alma L. Burlingame¹

¹Department of Pharmaceutical Chemistry and Mass Spectrometry Facility, University of California, San Francisco, CA 94158-2517 and ²Institute for Neurodegenerative Diseases and Department of Neurology, University of California, San Francisco, CA 94143-0518

* To whom correspondence should be addressed: Shenheng Guan, Mass Spectrometry Facility and Department of Pharmaceutical Chemistry, University of California, San Francisco, 600 16th Street, Genentech Hall, Room N472, Box 2240, San Francisco, CA 94158-2517; Email: sguan@cgl.ucsf.edu
Description of Algorithms

The whole data processing pipeline is illustrated in Figure S1.

Each of the processing modules (circles in Figure S1) performs a discrete function. First, MSMS peaklists are extracted from the raw files corresponding to no $^{15}$N incorporation (t=0) by a module called PAVA. A list of identified peptides is generated from a database search of the MSMS peaklists by the in-house database search engine Protein Prospector. The processing module fitXIC extracts ion chromatograms for each identified peptides from the centroid survey scans of the no incorporation (t=0) raw files. The CrossExtract module uses retention times defined in t=0 to extract $^{15}$N survey scan ions from the raw data files of matching gel slices at later time points (t>0). The NNLS module converts the cross-extracted survey ion peak intensities for each incorporation time into a $^{15}$N...
incorporation distribution based on theoretical isotope distributions and a nonnegative least square algorithm. The genCurve module uses the $^{15}$N distributions at each incorporation time point to generate a time-dependent relative $^{15}$N isotope abundance (RIA) curve for each peptide identified in the $t=0$ sample. The Pep2Prot module aggregates the peptide RIA curves into protein RIA curves. And finally, protein turnover rate constants are extracted from the RIA curves by fitting them with delayed exponential functions with the fitCurve module.

**PAVA.** The peaklist extraction module, written in Visual Basic, accesses the raw data files generated by LTQFT through functions provided by the Xcalibur software (ThermoFisher Scientific, Bremen, Germany). PAVA separates scans according to (1) MS levels (MS, MSMS, or MS3), (2) final detection devices (for LTQFT, either LTQ or FT can be used), and (3) dissociation methods for MSMS or MS3 scans. All current dissociation methods including collision-induced dissociation (CID), electron transfer dissociation (ETD), electron capture dissociation (ECD), infrared multiphoton dissociation (IRMPD), pulsed-q dissociation (PQD), and high collision energy dissociation (HCD) are support in PAVA. PAVA also separates survey scans into full or zoom scans. The output peaklists for MSn (n$\geq$2) scans are formatted as Matrix Science’s (Boston, MA) mgf format, which is compatible with most available database search engines. For spectra acquired in LTQ, data are first centroided if acquired in profile mode. The centroid peaklists are subjected to two stages of noise filtering similar to that employed in a previous study (1). In the first stage, the most intense peak is chosen within a $\pm 1.1$ m/z unit window. This step removes all isotope peaks except the most abundant from an ion’s isotopic distribution. In the second stage, the top six peaks
were chosen within a + 27 unit window. For spectra acquired in FT devices, PAVA centroids the spectra using a five-point routine with no noise filtering. PAVA allows batch processing and is very fast. It requires less than 10 seconds to process a 100MB raw data file containing 8000 MSMS scans acquired in centroid mode on the LTQFT instrument. The PAVA program is executable in a Windows environment, and requires installation of the Xcalibur software.

Protein Prospector(2, 3). This in-house database search engine allows for simultaneous searches of multiple peaklist mgf files against a combined forward/reverse or random database. The peptide identification data were formatted as a tab delimited text file.

fitXIC. This module, developed based on a concept of “identified ion chromatogram”(4), extracts the ion chromatogram for a peptide ion which has been identified by a database search. First, survey scans are centroided by PAVA’s centroid routine described early and the identified ion chromatogram is extracted from the centroid data using theoretical monoisotopic m/z of the peptide and a large retention time window, centered at the MSMS time. The ion chromatogram is then linearly interpolated onto an even-spaced grid and moving window averaging is used to reduce the noise. The approximate peak location is found at the maximum of the smoothed peak and approximate peak width is obtained by measuring the smooth peak width at half peak maximum. The approximate peak height is obtained by centroiding the original data centered at the approximate peak location. These peak parameters are used as initial fitting values and the original data is fitted with a modified Gaussian function with a parabolic variance (PVMG) (5). Peak location, peak height at peak location, peak area, peak width, and coefficient of determination (R2) are obtained from the fit of the raw data.
CrossExtract. This module extracts all $^{15}$N incorporated ions in raw data files of different $^{15}$N incorporation time points, based on the retention times determined from the fitXIC processing and on m/z values from the elemental compositions of the identified ions. For an identified peptide, the number of possible nitrogen atoms ($NNA$) being metabolically labeled can be easily calculated from its sequence or elemental composition. Externally introduced nitrogen atoms, such as those in the chemical thiol alkylation groups, are not counted. The maximal possible $^{15}$N incorporation is determined by the enrichment of the algae in the diet (99.56%).

To determine how many isotope peaks need to be extracted, the maximally incorporated isotope distribution is simulated by the mercury2 algorithm(6). The heaviest isotope peak of the distribution to keep must have an intensity of at least 1% relative to the intensity of the most abundant peak in the distribution. The number of ions needs to be extracted (denoted as $NPX$) is the Dalton units from the monoisotopic peak of the no $^{15}$N incorporation to the heaviest isotope peak of the most incorporated distribution plus one, including the monoisotopic peak. The m/z of each of the ions to be extract is calculated by use of the monoisotopic m/z of the peptide ion plus the product of the m/z difference between $^{15}$N and $^{14}$N atoms and the number of $^{15}$N atoms. Theoretically, the mass difference between the adjacent isotopes can be classified into two types. One type is due to the substitution of an $^{15}$N atom for an $^{14}$N (0.99703489Da) and the other is approximately the difference between an $^{13}$C and an $^{12}$C (1.00335484Da), in a natural isotope distribution. If spectra were acquired under an ultrahigh mass resolving power conditions, those peaks with the same nominal mass could be separated. Under our operating conditions, the fine structures(7) were not resolved.
The *CrossExtract* module averages survey scans in the raw data files within the retention time window with a width on the order of the extracted ion chromatographic peak width, centered at the peptide retention time. A built-in function of the Xcalibur software is called to average scans acquired in profile mode. The averaged spectral segment is then centroided and matched with the theoretical m/z values of possible $^{15}$N labeled ions. A mass tolerance was obtained from the precursor m/z error distribution of identified peptide ions. The direct cross extraction is feasible only if the variation of retention time is small compared to the chromatographic peak width. We avoid the need to consider LC alignment (8) procedures by achieving retention time stability experimentally. As will be discussed in more detail, variation in retention times is minimized (a) by use of constant column temperature, (b) by normalization of injection volume according to total protein concentration, and (c) by proper aging of the column.

**NNLS.** The module first synthesizes $NNA + 1$ theoretical isotope distributions using the mercury2 algorithm (6) (an example is given in Figure 2). The mercury2 algorithm is adapted so that the total abundance of the isotope distribution was normalized to one. The simulated distributions correspond to the peptide’s elemental composition with 0 to $NNA$ numbers of $^{15}$N, at 99.56% enrichment. Recall that the total number of isotope peaks to be extracted, $NPX$, is the sum of the number of possible $^{15}$N incorporations ($NNA$) and the number of isotope peaks with higher masses over the first isotopic peak in the most incorporated distribution. The theoretical distributions for all the numbers of $^{15}$N incorporation formed a matrix $A$ with $NNA + 1$ rows and $NPX$ columns. We can relate the $^{15}$N incorporation distribution and the cross-extracted ion intensities with the following linear equation...
\[ Af_i = p_i \]  

(1)

in which \( f_i \) is the \( ^{15}\)N incorporation distribution of the peptide at the incorporation time point of \( i (=1, 2, ..., NIT) \) and is a \( (NNA + 1) \times 1 \) vector whose elements indicate the contribution of a particular \( ^{15}\)N incorporated elemental composition to the observed ion intensity; \( p_i \) is an \( NPX \times 1 \) vector of the cross-extracted (experimental) ion intensities at the incorporation time point of \( i \). We solve the equation for \( f_i \) by use of an non-negative linear least square (NNLS) algorithm(9).

gencurve. This module implements an effective three-stage noise elimination procedure for each peptide ion. The gencurve module also converts the \( ^{15}\)N distribution matrix into a time vs \( ^{15}\)N relative fraction curve. Although the survey scans were acquired in high resolution, the cross-extracted \( ^{15}\)N incorporated ions might still be contaminated with a significant number of peaks that do not belong. The range of these ions could span tens of Daltons. The average number of possible \( ^{15}\)N incorporations for the peptides within our study was 18.6. The most common source of interference was due to oxidation of the same peptide (mainly on methionine residues). The oxidized form is 15.9949146221 Da heavier than molecular weight of the peptide. The same un-oxidized peptide with 16 \( ^{15}\)N incorporation is 15.9525582912 Da from the monoisotopic mass of the no incorporated peptide. The mass difference between the 16 \( ^{15}\)N ion and the oxidized ion is merely 42milliDa. The contamination of cross-extracted ions becomes unwanted “noise” in the \( ^{15}\)N incorporation distribution and must be minimized. In order to reduce the interference, three levels of noise removal procedures were carried out. First, distribution values (except zero incorporation number) at later incorporation time points which do not increase in mass since \( t=0 \) were recognized as contamination and removed. Second, we
construct a boundary between the signal and noise regions in the $^{15}$N distribution matrix. The boundary pointer starts at the last incorporation time and the highest incorporation peak location, moving towards the opposite corner of the $^{15}$N distribution matrix. Recognized that the $^{15}$N distribution can only move to the smaller incorporation numbers with decreases in incorporation times, the boundary detection pointer starts at the previous $^{15}$N incorporation number and checks the peak intensities at the current incorporation time point. It will move to a lower $^{15}$N incorporation number if the intensity is less than 0.1% of the maximum. To accommodate the signal fluctuation, the pointer is allowed to move two incorporation numbers higher if the intensity is greater than the 0.1% of the maximum. The boundary is the trajectory of the pointer moving from the most incorporated time points and the highest incorporation number toward the zero incorporation time point and the zero incorporation number (an example is given in Figure 4(b)). The two dimensional space of incorporation time and incorporation number is divided by the boundary into the signal region which contains the maximal incorporation time and zero incorporation number and the noise region which contains the corner of the zero incorporation time and the maximal incorporation number. Finally a signal threshold was estimated from the (noise) intensity histogram of peaks in the noise regions for the whole set of data. The noise intensity histogram is indeed a Poisson distribution. The distribution values in the noise region were set to zero. Third, distribution values in the signal region whose intensities below a noise threshold were also removed.

After these noise procedures were performed, genCurve calculated the relative $^{15}$N fraction of the peptide population at each incorporation time point. Let $F = \{ f_{ij} \}$ be the
distribution matrix, in which \( i=1, 2, \ldots, \text{NIT} \), is the index for the incorporation time points and \( j=0, 1, \ldots, \text{NNA} \), is the number of \( ^{15}\text{N} \) atoms incorporated number. \( j=0 \) is the zero \( ^{15}\text{N} \) incorporation number and \( j=\text{NNA} \) is the maximal \( ^{15}\text{N} \) incorporation number. We extracted the relative isotope abundance (RIA) of the newly synthesized peptide from the \( ^{15}\text{N} \) distribution of a given incorporation time point \( i \)

\[
RIA_i = \frac{\sum_{j=1}^{\text{NNA}} f_{ij}}{\sum_{j=0}^{\text{NNA}} f_{ij}}
\]

(2)

Recall that the number, NNA is the number of possible metabolically labeled nitrogen atoms. The \( \frac{1}{\sum_{j=0}^{\text{NNA}} f_{ij}} \) factor normalizes the maximal possible value of RIA to one. A peptide relative isotope abundance (RIA) curve is the function of the relative intensity of the newly synthesized peptide, RIA against the incorporation time.

In the genCurv module, we also calculate the relative number of \( ^{15}\text{N} \) atoms in the peptide (RNN)

\[
RNN_i = \left( \frac{\sum_{j=1}^{\text{NNA}} j \times f_{ij}}{\sum_{j=0}^{\text{NNA}} f_{ij}} \right) / \text{NNA}
\]

(3)

**Pep2Prot.** Since the relative abundance of the newly synthesized peptide defined in Eq.2 does not depend on the number of exchangeable nitrogen atoms in a peptide, we can aggregate or combine the peptide relative abundance curves into a protein curve. It is common that the components in a protein complex share some peptides of the same sequences. The shared peptides were removed before the aggregation process.
To aggregate the peptide relative abundance curves, a Pearson correlation coefficient was calculated for every pair of peptide curves. The peptide curve best connected to the rest (reflected by the maximal sum of Pearson coefficients) was chosen as the seed curve. If the second best connected curves had a Pearson coefficient greater than 0.9 to the seed curve, the second curve was averaged with the first to produce a new seed curve. This process continued until none of the other curves produced a Pearson coefficient greater than 0.9 with the seed curve or all the curves have been exhausted.

Aggregation of peptide relative abundance curves into a protein curve neglects the fact that different types of amino acids may have very different turnover behaviors. The turnover of amino acid has been shown to be very fast with a half life of minutes, therefore our assumption is that amino acid enrichment is not rate limiting (10, 11). The aggregation process produces the “averaged” protein relative abundance curves, of peptides with different amino acid compositions.

**fitCurve.** Protein turnover is a complex process, where multiple steps are involved in both protein synthesis and degradation. The food source needs to be digested before amino acids can go into the circulation system. Also, we analyzed proteins from whole tissue samples, which are collections of many different cell types. Protein synthesis and degradation occur in different cellular and organelle locations. One may attempt to fit experimental data with complicated models in order to include many possible processes associated with protein turnover. But this typically leads to over-fitting of data. On the other hand, a too simplistic model may not capture essential features of the protein turnover processes. For our current purpose to extract global protein turnover information, the protein curves were fit with a single exponential function with a time
delay. In a recent dynamic SILAC experiment (12), the natural amino acids were used to “chase” the labeled amino acids. A single exponential function was used to fit the experimental decay curves. The critical difference between our method and that of (12) is that there is a delay in delivery of $^{15}$N-labeled amino acids to cells in mammalian tissue. We consider that it is appropriate to include a time delay parameter into the fitting function. The enrichment (relative abundance) of our newly synthesized proteins or peptides was model with

$$y(t) = 1 - e^{-k(t-t_0)}$$  \hspace{1cm} (4)

The simple functional form of Equation 4 is able to capture the essential features of our peptide and protein turnover curves. However, some protein curves, obtained from the aggregation of many peptide curves, have detailed features that are not represented by this simple function. More sophisticated compartment modeling (10, 13) requires fitting of the curves with poly-exponential functions. The number of exponentials used corresponds to the number of compartments considered in the model. This will be the subject of our future investigations.

Reference for Description of Algorithms:

1. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *Journal of proteome research* **3**, 958-64.
2. Chalkley, R. J., Baker, P. R., Medzihradszky, K. F., Lynn, A. J., and Burlingame, A. L. (2008) In-depth analysis of tandem mass spectrometry data from disparate instrument types. *Mol Cell Proteomics* **7**, 2386-2398.
3. Clauser, K. R., Baker, P., and Burlingame, A. L. (1999) Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Analytical chemistry* **71**, 2871-2882.
4. Guan, S., Oses-Prieto, J., Chu, F., Huang, Y., Maltby, D. A., and Burlingame, A. L. (2007) A Label-Free Quantitation Method for Ion-Trap/High-Resolution Hybrid
Instruments. *The 55th ASMS Conference on Mass Spectrometry*, p. TPI 311, Indianapolis, Indiana.

5. Baeza-Baeza, J. J., and Garcia-Alvarez-Coque, M. C. (2004) Prediction of peak shape as a function of retention in reversed-phase liquid chromatography. *Journal of chromatography* **1022**, 17-24.

6. Rockwood, A. L., and Van Orden, S. L. (1996) Ultrahigh-Speed Calculation of Isotope Distributions. *Analytical chemistry* **68**, 2027–2030.

7. Marshall, A. G., Hendrickson, C. L., and Jackson, G. S. (1998) Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass spectrometry reviews* **17**, 1-35.

8. Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T. A., Hill, L. R., Norton, S., Kumar, P., Anderle, M., and Becker, C. H. (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Analytical chemistry* **75**, 4818-26.

9. Lawson, C. L., and Hanson, R. J. (1974) *Solving least squares problems*, p. 161, Prentice-Hall, Englewood Cliffs, NJ.

10. Waterlow, J. C. (2006) *Protein Turnover*, CABI, Cambridge, MA.

11. Wolfe, R. R., and Chinkes, D. L. (2005) *Isotope Tracers in metabolic Research*, 2nd Ed., John Wiley & Sons, Hoboken, New Jersey.

12. Doherty, M. K., Hammond, D. E., Clague, M. J., Gaskell, S. J., and Beynon, R. J. (2009) Turnover of the human proteome: determination of protein intracellular stability by dynamic SILAC. *Journal of proteome research* **8**, 104-12.

13. DiStefano, J. J., 3rd, and Landaw, E. M. (1984) Multiexponential, multicompartmental, and noncompartmental modeling. I. Methodological limitations and physiological interpretations. *The American journal of physiology* **246**, R651-64.