Supporting Information for

The Area between Exchange Curves as a Measure of Conformational Differences in Hydrogen-Deuterium Exchange Mass Spectrometry Studies

Sharlyn J. Mazur[1,*] and Daniel P. Weber[2]

[1]Laboratory of Cell Biology, National Cancer Institute,
National Institutes of Health, Bethesda, Maryland 20892 U.S.A.

[2]Interclypse, Hanover, Maryland 21076 U.S.A.

*corresponding author: sm328j@nih.gov

Theoretical Background

A conceptual framework for describing amide hydrogen exchange rates in globular proteins under native conditions was developed over forty years ago and has been summarized in recent reviews [1-4]. The native structure of globular proteins protects amide hydrogens from exchange with solvent deuterium. Exchange becomes possible during transient unfolding of a microdomain and is usually characterized by its behavior under two limiting conditions, EX1 and EX2 [1-3]. Under EX1 conditions, the rate of microdomain refolding is much slower than the rate of exchange, resulting in correlated exchange at neighboring residues. Under EX2 conditions, which usually apply to globular proteins under native conditions, the rate of microdomain refolding is much higher than the rate of exchange.

We will focus exclusively on the exchange behavior of native proteins in the context of EX2 conditions for the purpose of interpreting peptide-level hydrogen-deuterium-mass spectrometry (HDX-MS) data. Following Hvidt and Wallevik [4], the fraction of the j^{th} amide hydrogen available for exchange, $\rho^{(j)}$, can be related to the equilibrium constant for microdomain unfolding, $K^{(j)}$:

$$\rho^{(j)} = [P_o H^{(j)}]/[P_{tot}] = 1/(1+K^{(j)}) \qquad (1)$$

where $[P_{tot}]$ and $[P_o H^{(j)}]$ indicate the concentrations of the total protein and of the protein in which the j^{th} amide is available for exchange, respectively. Note that the inverse of ρ has often been termed the protection factor [1]. The rate constant for exchange of a fully exposed amide hydrogen, k_o, is affected by solution pH, temperature, and the identities of the side chain and neighboring residues [4-6]. After exchange time, t, the remaining amide hydrogen fraction is:

$$y(t) = \frac{1}{m}\sum_{j=1}^{m} \exp(-\rho^{(j)} k_o t) \qquad (2)$$

Hvidt and Wallevik [4] showed that in plots of exchange curves versus log time, the reference

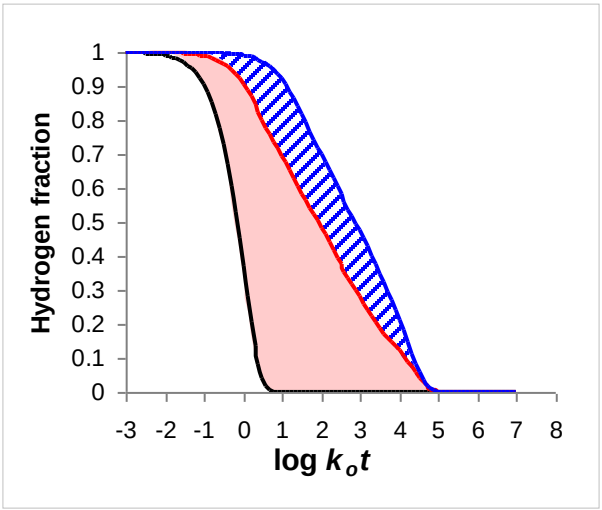

Fig. S1 Simulated exchange curves and areas between exchange curves. Curves representing remaining hydrogen fractions of fully exposed peptides (black) and of peptides under native conditions in state ***a*** (red) and state ***b*** (blue) are plotted against the $\log_{10}$ of the exchange time scaled by the intrinsic exchange rate constant k_o. The reference area (pink) between the exchange curves for the native state ***a*** and the fully exposed peptide is related to the peptide-averaged fraction of exchange-competent amide hydrogens in the native state. The area (blue hatched) between the exchange curves for states ***a*** and ***b*** is related to the ratio of the exchange-competent amide hydrogens for the native conditions ***a*** and ***b***.

area defined between the exchange curve for a protein under native conditions and the hypothetical exchange curve for the protein in a fully exposed state is related to the average fraction of amide hydrogens in the exchange-competent state as:

$$A_{ref} = \langle -\ln \rho \rangle \tag{3}$$

The reference area is depicted schematically in Figure S1. We can extend this formalism to interpret the area between exchange curves for peptides from native proteins in states ***a*** (*e.g.*, free) and ***b*** (*e.g.*, ligand-bound). The reference area for native state ***a*** can be expressed as the difference of the integrals:

$$A_{ref,a} = \int_{t=0}^{\infty} y_a(t) d\ln t - \int_{t=0}^{\infty} \exp(-k_o t) d\ln t \tag{4}$$

After substitution with equation (2), this becomes:

$$A_{ref,a} = \tfrac{1}{m} \sum_{j=1}^{m} \int_{t=0}^{\infty} \exp(-\rho_a^{(j)} k_o t) d\ln t - \int_{t=0}^{\infty} \exp(-k_o t) d\ln t \tag{5}$$

The reference area for state b is similarly expressed. After cancelling terms reflecting the common reference state of the fully exposed peptide, the area between exchange curves for peptides from native proteins in states ***a*** and ***b*** can be expressed as:

$$A_{bec} = \tfrac{1}{m} \sum_{j=1}^{m} \int_{t=0}^{\infty} \exp(-\rho_b^{(j)} k_o t) d\ln t - \tfrac{1}{m} \sum_{j=1}^{m} \int_{t=0}^{\infty} \exp(-\rho_a^{(j)} k_o t) d\ln t \tag{6}$$

After rearranging, and then evaluating the definite integral, this yields:

$$A_{bec} = \tfrac{1}{m} \sum_{j=1}^{m} \ln\left(\rho_a^{(j)} / \rho_b^{(j)}\right) = \langle \ln(\rho_a / \rho_b) \rangle \tag{7}$$

The area between exchange curves can thus be interpreted as an estimate of average log ratio of the exchange competent fractions in the two states. When $y_a(t)$ and $y_b(t)$ are plotted against the $\log_{10}$ of the exchange time, A_{bec} values of ~0.3 unit correspond to a two-fold increase in the exposed fraction in the less stable state, and an A_{bc} value of ~1 unit corresponds to a 10-fold increase in the exposed fraction.

Under EX2 conditions and when ρ is small, the average exposed fraction can be related to the average equilibrium constant for microdomain unfolding as:

$$\langle -\log \rho \rangle \approx \langle \log K \rangle = \langle \Delta G \rangle / 2.303RT \tag{8}$$

where R is the gas constant and T is the absolute temperature (see [4] for details). Similarly, the average ratio of the exposed fractions in the two states (eq. 7) can be related to the average difference in the free energies of microdomain unfolding between the two states:

$$A_{bec} \approx \langle \Delta\Delta G \rangle / 2.303RT \tag{9}$$

As the fully exposed state is the same for the bound and free states of the protein, A_{bec} provides an estimate of the average difference in stability of the observed protein segment between the two states. Thus, the area between exchange curves, when plotted as remaining hydrogen fraction versus logtime, admits a molecular interpretation of interest without requiring resolution of the individual exchange rate constants.

Analytical and Numerical Methods

HDX-MS data can be categorized as time series data with a relatively small number of time points. Analysis of short time series data is challenging because observations taken at different time points are not independent, yet the number of data points is too small to support methods expressly developed to analyze longitudinal or periodic data. Functional Data Analysis (FDA) has been usefully applied to short time series data in fields such as gene expression analysis and metabolomics [7-9]. The general concept in the FDA approach is that observations represent noisy measurements of an underlying function that describes the dynamic behavior of an entity. The dynamic behavior of two entities can be judged to be similar if a suitable distance measure comparing the two curves is small. For curves with simple shapes, the area between the curves is a suitable measure, whereas in the case of curves with complex shapes, a distance measure that includes shape information may be necessary. In general, curve shapes for HDX-MS data are not complex. When expressed as remaining hydrogen fraction, the curves are monotonically decreasing, with various degree of structure. For most applications of the HDX-MS method to native proteins, in which EX2 conditions hold, the area between exchange curves is a suitable measure of exchange curve dissimilarity.

The area between exchange curves is a useful measure of dissimilarity in dynamic behavior. We provide two non-parametric, geometric methods and two parametric curve-fitting methods for estimating the area between exchange curves. In most applications of the HDX-MS method, the number of replicate runs for the conditions being compared are the same and identical exchange intervals are sampled. The methods described here do not necessarily require identical numbers of replicate runs or identical exchange intervals for the two conditions.

We are interested in using HDX-MS methods to define differences in the conformational mobility of a protein between states ***a*** (*e.g.*, free) and ***b*** (*e.g.*, ligand-bound) [1-3]. For $n_{rep,a}$ and $n_{rep,b}$ replicated experiments, samples of the protein in the two conditions are exposed to high deuterium content buffers for n exchange intervals, $t_1 \ldots t_n$. Samples are quenched and digested into peptides that are chromatographically separated and analyzed by mass spectrometry. Data are most commonly reported as mean relative percent deuterium incorporated, $D_{a,i}$, for the i^{th} exchange interval for condition ***a***, and are often presented as plots of D vs. log time. We used a plot digitizer utility [10] to extract deuterium uptake values from published HDX data [11] and calculated the mean remaining hydrogen fractions, $y_{a,i} = (100\text{-}D_{a,i})/100$ and $y_{b,i} = (100\text{-}D_{b,i})/100$, and the corresponding standard deviations, $s_{a,i}$ and $s_{b,i}$.

Scaled difference of averaged incorporation. An established method for representing HDX results compactly is to calculate the average deuterium uptake [12]. This method reports the progress of the exchange reaction averaged over all exchange intervals and does not explicitly retain temporal relationships between timepoints. In the usual practice, the difference in the averaged deuterium uptake between the two conditions, $\overline{D}_b - \overline{D}_a$, is used as the basis for assessing differences in dynamic behavior [12]. For our purpose, we calculated the difference in remaining hydrogen fractions and scaled the result by the span of the log of the exchange intervals:

$$A_{bec}^{averaged} = \log(t_n/t_1)(\bar{y}_b - \bar{y}_a) = \log(t_n/t_1)(\tfrac{1}{n})\left(\sum_{i=1}^{n} y_{b,i} - y_{a,i}\right) \qquad (10)$$

Expressing the HDX-MS data as remaining hydrogen fractions and multiplying by a constant (the span in log time) converts the difference in averaged reaction progress to an estimate of the area between exchange curves, allowing direct comparison with the other methods and supporting the potentially useful molecular interpretation. Assuming homogeneity of variance, then the pooled standard deviation is:

$$S_{bec}^{averaged} = \left(((n_{rep,a}-1)\bar{s}_a^2 + (n_{rep,b}-1)\bar{s}_b^2)/(n_{rep,a}+n_{rep,b}-2)\right)^{1/2} \quad (11)$$

where the calculated standard deviation of the scaled average remaining hydrogen fraction for state ***a*** is

$$\bar{s}_a = \sqrt{\log(t_n/t_1)\cdot(\tfrac{1}{n})\cdot\sum_{i=1}^{n}(s_{a,i})^2} \quad (12)$$

and similarly for $\bar{s}_b$.

Trapezoidal approximation. The area between the curves can be estimated from the areas of n - 1 trapezoids as depicted in Figure S2(a):

$$A_{bec}^{trapezoid} = \tfrac{1}{2}\sum_{i=1}^{n-1}(y_{b,i} - y_{a,i} + y_{b,i+1} - y_{a,i+1})\cdot\log(t_{i+1}/t_i) \quad (13)$$

Note that some terms in the above expression cancel. The area of the n-1 trapezoids is exactly equal to the area of n rectangles as depicted in Figure S2(b):

$$A_{bec}^{trapezoid} = \tfrac{1}{2}\left((y_{b,1}-y_{a,1})\log\frac{t_2}{t_1} + (y_{b,n}-y_{a,n})\log\frac{t_n}{t_{n-1}} + \sum_{i=2}^{n-1}(y_{b,i}-y_{a,i})\log\frac{t_{i+1}}{t_{i-1}}\right) \quad (14)$$

This expression makes explicit the interpretation that intervals in the log time dimension provide weighting factors for differences in deuterium uptake between the two states observed at each timepoint. The standard deviation for the area under the curve for state ***a*** is:

$$s_{a,trapezoid} = \sqrt{\tfrac{1}{2}\left(s_{a,1}^2\log(t_2/t_1) + s_{a,n}^2\log(t_n/t_{n-1}) + \sum_{i=2}^{n-1}s_{a,i}^2\log(t_{n+1}/t_{n-1})\right)} \quad (15)$$

and similarly for state ***b***. The pooled standard deviation of the area between the curves is given by:

$$S_{bec}^{trapezoid} = \left(((n_{rep,a}-1)s_{a,trapezoid}^2 + (n_{rep,b}-1)s_{b,trapezoid}^2)/(n_{rep,a}+n_{rep,b}-2)\right)^{1/2} \quad (16)$$

In addition to the two geometric methods described above, we provide two parametric, curve-fitting approaches as heuristic methods for estimating the area between exchange curves. The functional forms used

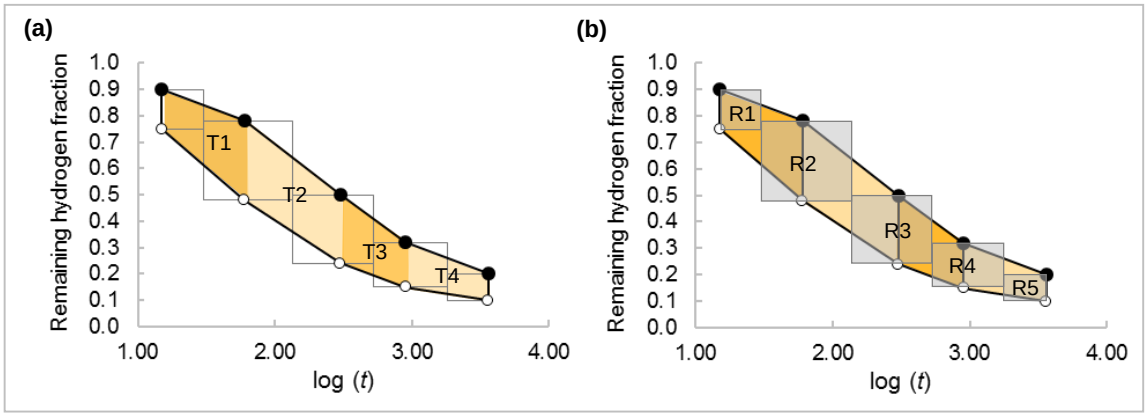

Fig. S2 Schematic showing remaining hydrogen fraction for a peptide from states ***a*** (○) and *b* (●) plotted against the log of the exchange time, *t*, sampled after 5 discrete exchange intervals. The area between the curves is equivalently represented as **(a)** four trapezoids (T1 to T4, dark or light orange) or **(b)** five rectangles (R1 to R5, grey).

in both cases are obviously incorrect for describing amide hydrogen exchange dynamics. Other methods, such as spline fitting, may also be appropriate. However, the use of these methods provides an illustration of the robustness of the area between the curves as a metric for exchange curve dissimilarity and of the sensitivity of significance calculations to the uncertainites derived from parametric modeling.

Weighted second order polynomial. The remaining hydrogen fraction for state ***a*** can be modeled as a function of log time as a second order polynomal:

$$y_a(\log t) = a_2 \cdot (\log t)^2 + a_2 \cdot (\log t) + a_0 \tag{17}$$

where the independent variable, log*t*, is the base 10 log of the time in seconds. Weighted linear least squares methods can be used to determine the best estimates of the parameters and their standard deviations. In particular, we used the *lm*() function in R with the model y_a ~ log*t* +I(log*t*^2) and with weights taken as inversely proportional to the observed variance. Fitting to state ***b*** was performed similarly. The area under the curve for state ***a*** can be evaluated analytically:

$$A_{a,poly} = a_2((\log t_n)^3 - (\log t_1)^3)/3 + a_1((\log t_n)^2 - (\log t_1)^2)/2 + a_0(\log t_n - \log t_1) \tag{18}$$

and with the corresponding uncertainty in the area under the curve for state ***a*** being given by:

$$s_{a,poly} = \sqrt{\sigma_{a2}^2((\log t_n)^3 - (\log t_1)^3)/3 + \sigma_{a1}^2((\log t_n)^2 - (\log t_1)^2)/2 + \sigma_{a0}^2(\log t_n - \log t_1)} \tag{19}$$

where σ_{a2}, σ_{a1}, and σ_{a0} are the standard deviations of the model parameters. Calculations for state ***b*** were performed similarly. The area between exchange curves can be estimated from the difference in the areas under the curves:

$$A_{bec}^{poly2w} = A_{b,poly} - A_{a,poly} \tag{20}$$

with the associated pooled standard deviation:

$$S_{bec}^{poly2w} = \left(\left((n_{rep,a} - 1)s_{a,poly}^2 + (n_{rep,b} - 1)s_{b,poly}^2\right)/(n_{rep,a} + n_{rep,b} - 2)\right)^{1/2} \tag{21}$$

Three parameter logistical function. When plotted against log time, exchange curves often appear sigmoidal, suggesting that a logistical function may provide a useful model. In practice, however, fitting HDX-MS data to logistical functions is limited by incompletely defined curves. This may result from exchange processes that occur at times shorter or longer than the tested exchange intervals or because the number of timepoints is small. We attempted fitting a logistical function to HDX data for 31 peptides for a protein in each of two states, obtained for five exchange intervals spanning 1s to 3600 s, with four replicates for each condition [8]. For a sizeable fraction of the peptides, nonlinear least squares fitting of the logistical function with four parameters failed to converge. This was not surprising because the exchange curves were not well defined by the small number of timepoints. We were able to fit a three parameter logistical function for 29 of the 31 peptides, while keeping the parameter b_{fixed} fixed for the two states for a given peptide. In this case, the remaining hydrogen fraction for state ***a*** can be modeled as a function of time by

$$y_a(t) = d_a + (a_a - d_a)/(1 + \exp(b_{fixed} \cdot \ln(t/c_a))) \tag{22}$$

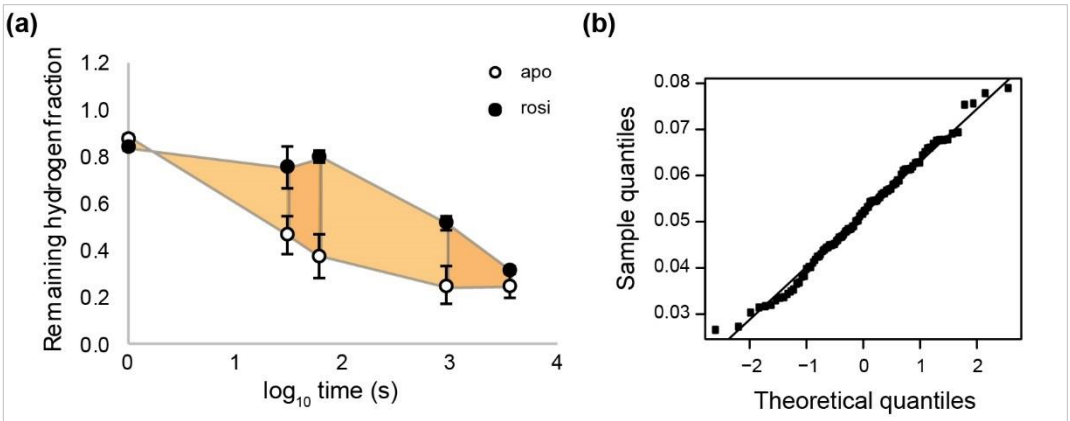

Figure S3 Trapezoidal estimation of the area between exchange curves as a measure of hydrogen deuterium exchange curve dissimilarity. (a) Remaining hydrogen fractions for Peroxisome proliferator-activated receptor gamma (PPARγ) peptide *aa*445-452 in the absence (apo, ○) and presence of rosiglitazone (rosi, ●). HDX data are from [11]. The area between the exchange curves, A_{bec}, was estimated by trapezoidal approximation (orange trapezoids) alonw with the pooled standard deviation, S_{bec}. (b) QQ plot of simulated S_{bec} values compared against a normal distribution. Simulated data were prepared from 100 runs using normally distributed values with means as in (a) and standard deviations of 0.02. A_{bec} and S_{bec} values were calculated by trapezoidal approximation and propagation of errors, respectively.

where d_a is the lower asymptote, a_a is the upper asymptote, and c_a is the characteristic time for state ***a***. The coefficient b_{fixed}, which influences the steepness of the curve, was held fixed for states ***a*** and ***b*** of a given peptide. Nonlinear least squares methods can be used to determine the best estimates of the parameters and their standard deviations. In particular, we used the R *nls*() function with the model $y_a \sim d_a + (a_a - d_a)/(1 + \exp(b_{fixed} \cdot \ln(t/c_a)))$ to estimate the three parameters and their standard deviations [13]. Fitting to state ***b*** was performed similarly. The integrated areas under the curves, $A_{a,logi}$ and $A_{b,logi}$, and the associated standard deviations, $s_{a,logi}$ and $s_{b,logi}$, were evaluated numerically. The area between the exchange curves can be estimated as:

$$A_{bec}^{logistical} = A_{b,logi} - A_{a,logi} \tag{23}$$

with the associated pooled standard deviation:

$$S_{bec}^{logistical} = \left(\left((n_{rep,a} - 1)s_{a,logi}^2 + (n_{rep,b} - 1)s_{b,logi}^2\right)/(n_{rep,a} + n_{rep,b} - 2)\right)^{1/2} \tag{24}$$

Statistical tests. The area between curves is a measure of curve dissimilarity. The A_{bec} value is a single measure of dynamic behavior over the entire time course of the experiment. In usual practice, the standard deviation of deuterium incorporation does not change markedly between samples, suggesting homogeneity in the variances, and supporting the use of a pooled standard deviation for statistical tests. As calculated by the trapezoidal approximation, A_{bec} values are well-behaved, with an associated standard deviation, S_{bec} (Figure S3a). We performed a simulation with estimated random variation for 100 runs for the two conditions. By quantile-quantile analysis, the distribution of calculated S_{bec} values was approximately normal with small deviations (Figure S3b). The Shapiro-Wilke normality test (R: shapiro.test, W = 0.993, p = 0.686) also supports an approximately normal distribution. In this case, the use of Student's t test to assess the significance of non-zero values of A_{bec} is suitable. The *t* statistic is given by:

$$t = A_{bec} \Big/ \left(S_{bec} \cdot (n_{rep,a}^{-1} + n_{rep,b}^{-1})^{1/2}\right) \tag{25}$$

with $(n_{rep,a} + n_{rep,b} - 2)$ degrees of freedom.

References

1. Englander, S.W.: Hydrogen exchange and mass spectrometry: a historical perspective. J. Am. Soc. Mass Spectrom. **17**, 1481-1489 (2006)

2. Engen, J.R.: Analysis of protein conformation and dynamics by hydrogen-deuterium exchange MS. Anal. Chem. **81**, 7870-7875 (2009)

3. Gallagher, E.S., and Hudgens, J.W.: Mapping protein-ligand interactions with proteolytic fragmentation, hydrogen/deuterium exchange-mass spectrometry. Meth. Enzym. **566**, 357-404 (2016)

4. Hvidt, A., Wallevik, K.: Conformational changes in human serum albumen as revealed by hydrogen-deuterium exchange studies. J. Biol. Chem. **247**, 1530-1535 (1972)

5. Engen, J.R., and Wales, T.E.: Analytical aspects of hydrogen exchange mass spectrometry. Ann. Rev. Anal. Chem. **8**, 127-148 (2015)

6. Bai, Y., Milne, J.S., Mayne, L., Englander, S.W.: Primary structure effects on peptide group hydrogen exchange. Proteins **17**, 75-86 (1993)

7. Coffey, N., Hinde, J.: Analyzing time-course microarray data using functional data analysis. Stat. Appl. Genom. Mol. Biol. **10**: Article 23 (2011)

8. Minas, C., Waddell, S.J., Montana, G.: Distance-based differential analysis of gene curves. Bioinformatics **27**, 3135-3141 (2011)

9. Montana, G., Berk, M., Ebbels, T.: Modelling short time series in metabolomics: a functional data analysis approach. Adv. Exp. Med. Biol. **696**, 307-315 (2011)

10. Rohatgi, A., WebPlotDigitizer. arohatgi.info/WebPlotDigitizer (2016)

11. Bruning, J.B., Chalmers, M.J., Prasad, S., Busby, S.A., Kamemecka, T.M., He, Y., Nettles, K.W., Griffin, P.R.: Partial agonists activate PPARγ using a helix 12 independent mechanism. Structure **15**, 1258-1271 (2007)

12. Pascal, B.D., Chalmers, M.J., Busby, S.A., Griffin, P.R.: HD Desktop: an integrated platform for the analysis and visualization of H/D exchange data. J. Am. Soc. Mass Spectrom. **20**, 601-610 (2009)

13. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2016)