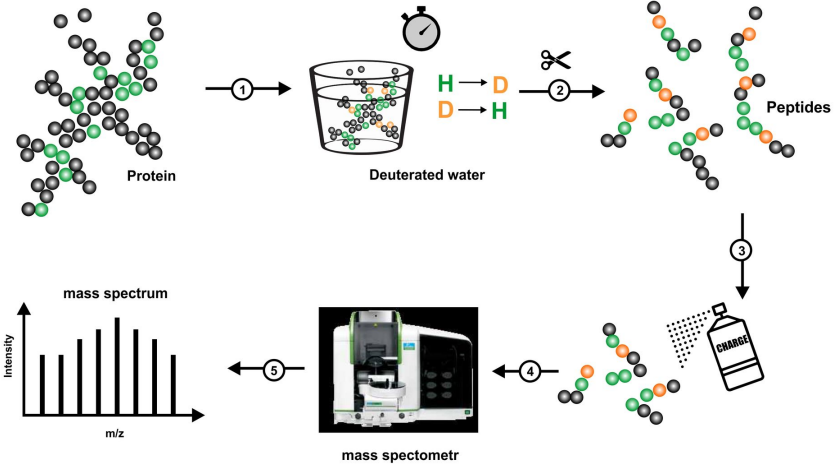
Comparing deuteration curves

Krystyna Grzesiak

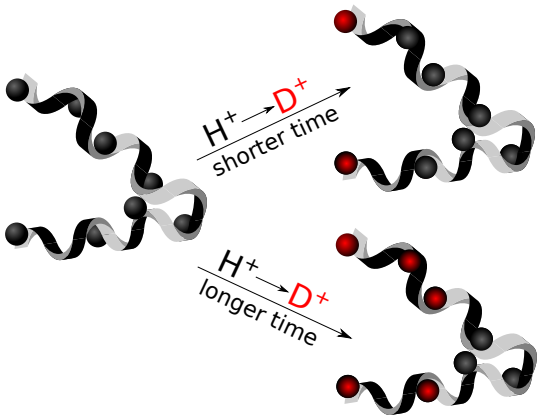
May 19, 2021

HDX: dynamics measurements of protein structure



Closer look at hydrogen-deuterium exchange

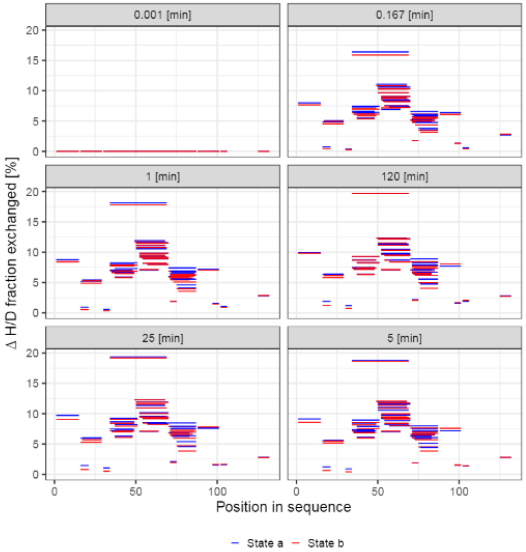
- measurement of the mass of peptides coming from proteins incubated in the deuterated water
- the most exposed amide hydrogens tend to be replaced by deuters
- exchange rate is related to the position of the peptide in the structure of protein



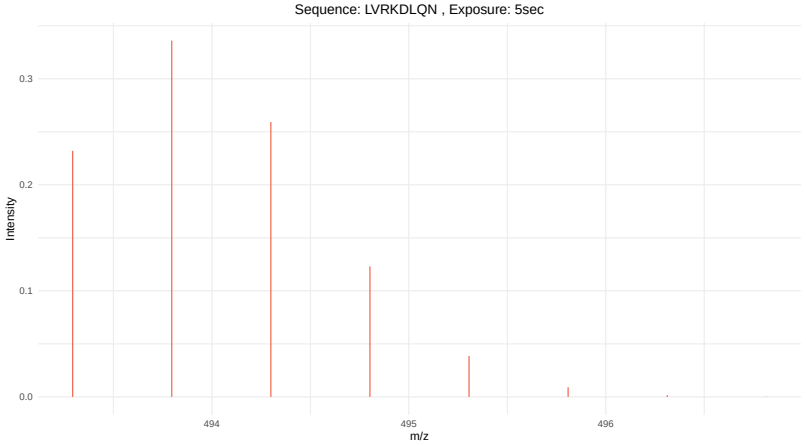
Limitations introduced by the usage of mass spectrometry

- "Randomly" appended charge - another aggregation coming from different values of charge
- Experimenter and his reflex - the variability of the variance of measurement errors among all the time points must be taken into account
- Multidimensionality of assessment - peptide can be of various lengths and their parts can overlap each other. Besides, peptides are generated randomly and in an artificial way. Thus, it can happen that they do not cover or do not sufficiently cover the major regions of interest.

Multidimensionality of assessment



Mass spectrum - example plot



Mass spectrum - data

Exposure [sec]	Intensity	Mz	Charge	Sequence
0.00	0.58	493.29	2.00	LVRKDLQN
0.00	0.30	493.29	2.00	LVRKDLQN
0.00	0.09	493.29	2.00	LVRKDLQN
0.00	0.02	493.29	2.00	LVRKDLQN
0.00	0.00	493.29	2.00	LVRKDLQN
0.00	0.00	493.29	2.00	LVRKDLQN
5.00	0.18	493.29	2.00	LVRKDLQN
5.00	0.36	493.80	2.00	LVRKDLQN
5.00	0.28	494.30	2.00	LVRKDLQN
5.00	0.13	494.80	2.00	LVRKDLQN
5.00	0.04	495.31	2.00	LVRKDLQN
5.00	0.01	495.81	2.00	LVRKDLQN
5.00	0.00	496.31	2.00	LVRKDLQN
5.00	0.00	496.82	2.00	LVRKDLQN

Deuterium uptake calculations

We calculate the averaged experimental mass for the exposure time 5sec as follows:

$$m_5 = \frac{1}{N} \sum_{k=1}^{N} Intensity_k \cdot Charge \cdot (Mz_k - 1.007276)$$

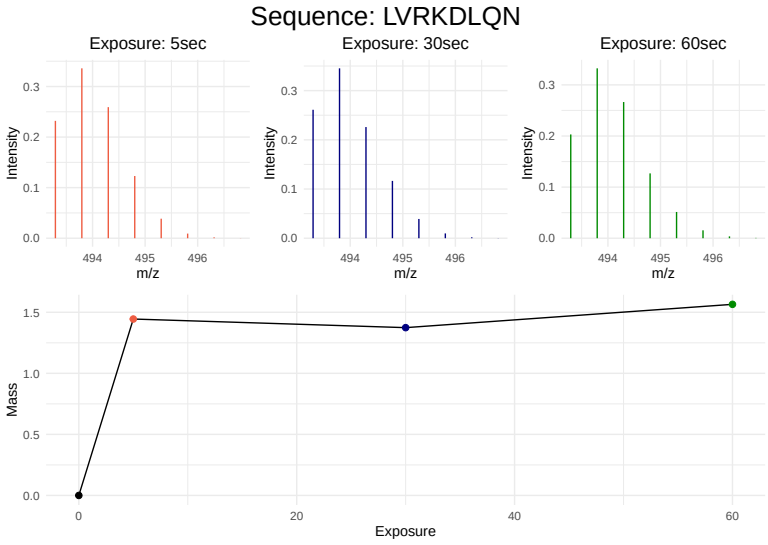
Based on the formula below we calculate the deuterium uptake at time 5

$$D_5 = m_5 - m_0.$$

Analogously we obtain results for different time points:

Exposure [sec]	Charge	Sequence	Mass
0.00	2.00	LVRKDLQN	0.00
5.00	2.00	LVRKDLQN	1.44
30.00	2.00	LVRKDLQN	1.37
60.00	2.00	LVRKDLQN	1.56

Deuterium uptake

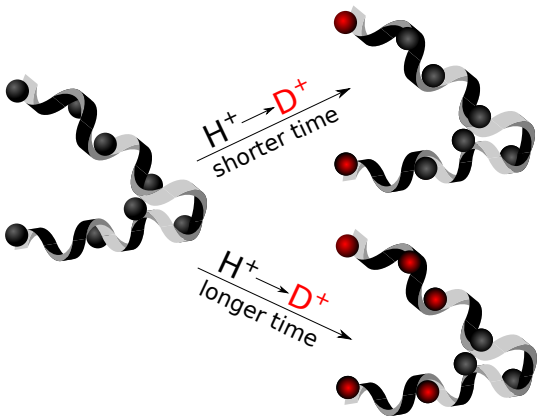


Protection factors - a key to understanding protein structures

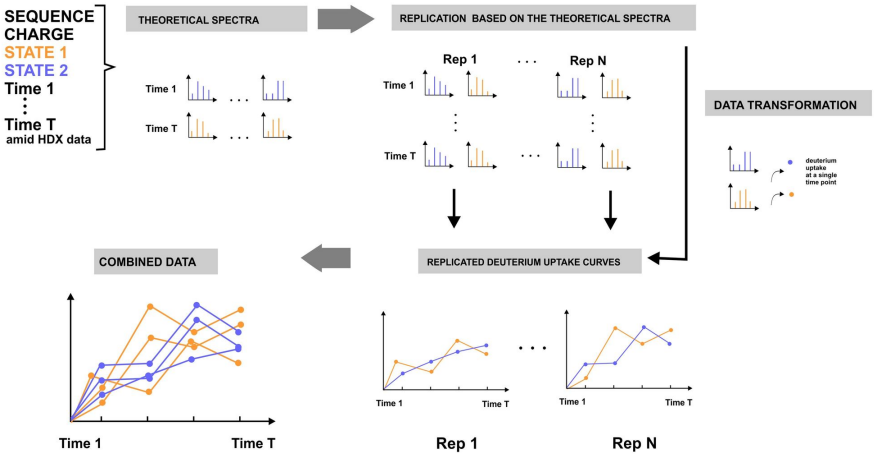
- a kind of upper level of protein description, basically impossible to be obtained from experimental data
- actually affects the exchanges between hydrogens and deuters

$$P\left(H \rightarrow D\right) = 1 - \exp\left(\frac{-kcHD \cdot \Delta t}{Pf}\right),$$

$$P\left(D \rightarrow H\right) = 1 - \exp\left(\frac{-kcDH \cdot \Delta t}{Pf}\right).$$



powerHDX package

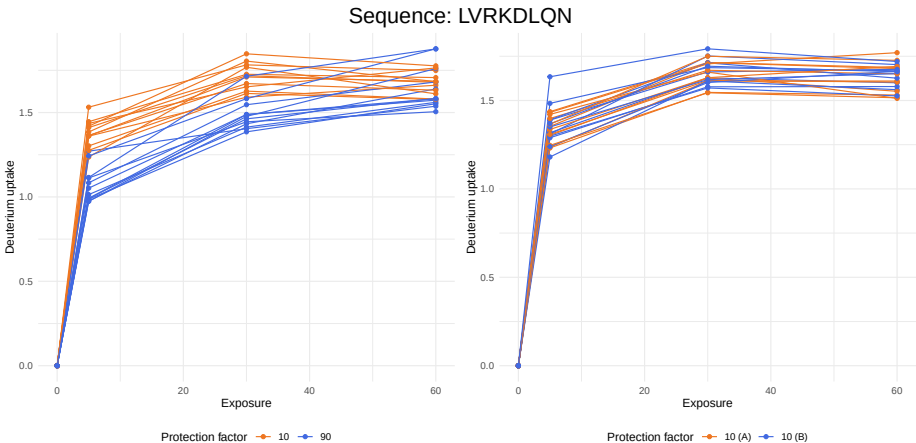


deuterium uptake curves simulation by powerHDX

Data

Sequence	Rep	Protection factor	Exposure[sec]	Mass	Charge	Experimental_state
LVRKDLQN	1	10	0.00	0.00	1	A
LVRKDLQN	1	10	0.00	0.00	2	A
LVRKDLQN	1	10	0.00	0.00	3	A
⋮	⋮	⋮	⋮	⋮	⋮	⋮
LVRKDLQN	4	40	43200.00	1.57	1	B
LVRKDLQN	4	40	43200.00	1.55	2	B
LVRKDLQN	4	40	43200.00	1.65	3	B

Data


Sequence: LVRKDLQN
Deuterium uptake
0.0
0.5
1.0
1.5
0
20
40
60
Exposure
Protection factor
10
90
Deuterium uptake
0.0
0.5
1.0
1.5
0
20
40
60
Exposure
Protection factor
10 (A)
10 (B)

Main issue

What are we looking for?

a test based on semiparametric regression such that it can determine statistical significance of differences in deuteration levels between states. Namely:

$$H_0 : StateA = StateB$$

vs.

$$H_1 : StateA \neq StateB$$

GAM - Generalized Additive Model

Generalized additive model for response Y (along with link function g) and predictors $x_1, \ldots, x_p$ can be represented by following formula

$$g(\mathbb{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p),$$

where $f_1, \ldots, f_p$ are some smooth functions.
An example basis for space of smooth functions is K-spline:

$$f(x_i) = \beta_0 + x_i\beta_1 + \sum_{k=1}^{K} u_k(x_i - \kappa_k)_+,$$

where $(x - a)_+ = max(0, x - a)$.

Regression splines

Let $\kappa \in \{5, 10, 20, 30, 40, 50, 60, 100, 300, 500, 900, 1200, 1500, 1800, 2100, 2400, 3600, 7200, 21600\}$. We consider the following models for the variable *Time* transformed by identity or logarithm

$$g(\mathbb{E}(Y)) = \beta_0 + \beta_1 Time + \beta_2 (Time - \kappa)_+ + \beta_3 State + \beta_4 Time \times State$$

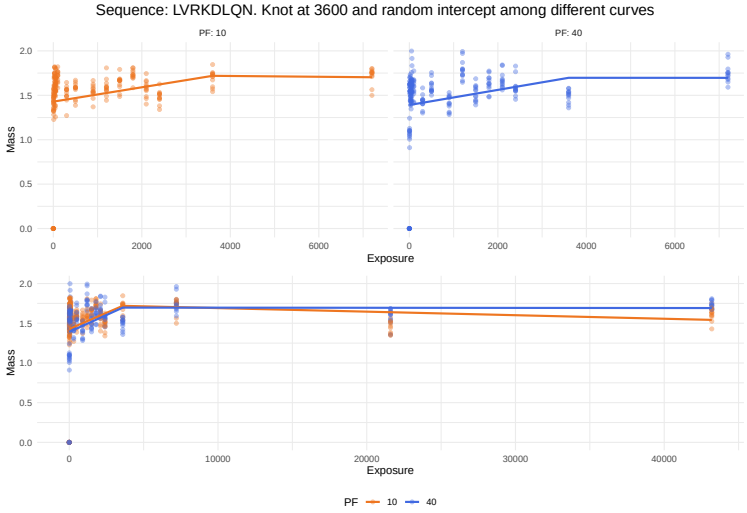
$$g(\mathbb{E}(Y)) = \beta_0 + \beta_1 Time + \beta_2 (Time - \kappa)_+^2 + \beta_3 State + \beta_4 Time \times State$$

$$g(\mathbb{E}(Y)) = \beta_0 + b_{id} + \beta_1 Time + \beta_2 (Time - \kappa)_+ + \beta_3 State + \beta_4 Time \times State$$

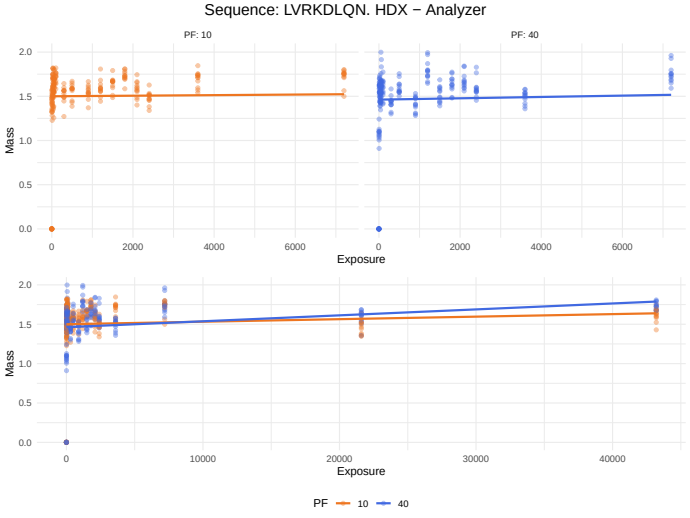
$$g(\mathbb{E}(Y)) = \beta_0 + b_{rep} + \beta_1 Time + \beta_2 (Time - \kappa)_+ + \beta_3 State + \beta_4 Time \times State$$

where b_{id} and b_{rep} are random intercepts for *id* and *rep* respectively.

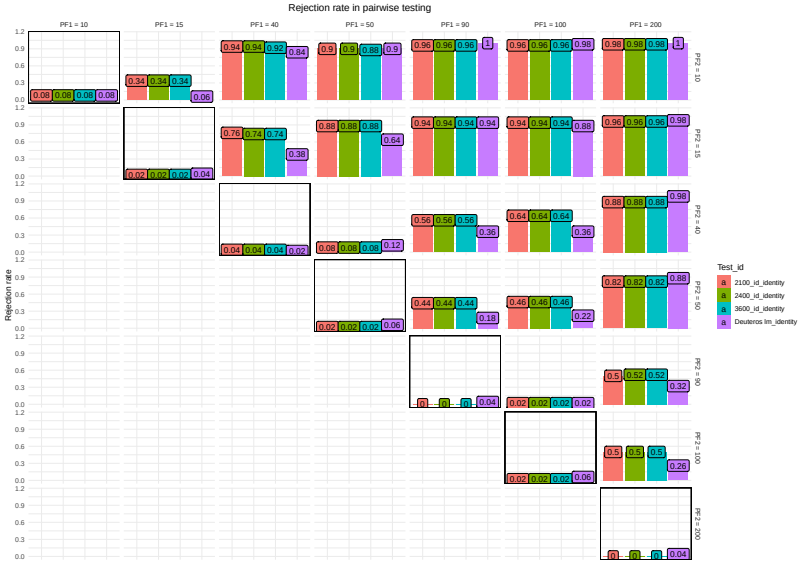
Regression splines



HDX - Analyzer



Rejection rate



Papers

- *Simple fitting of subject-specific curves for longitudinal data* Durban, Harezlak, Wand Carroll (Stat in Med, 2005)
- *HaDeX: an R package and web-server for analysis of data from hydrogen–deuterium exchange mass spectrometry experiments* Weronika Puchała, Michał Burdukiewicz, Michał Kistowski, Katarzyna A Dabrowska, Aleksandra E Badaczewska-Dawid, Dominik Cysewski, Michał Dadlez (Bioinformatics, 2020)
- *HDX-Analyzer: a novel package for statistical analysis of protein structure dynamics* Sanmin Liu, Lantao Liu, Ugur Uzuner, Xin Zhou, Manxi Gu, Weibing Shi, Yixiang Zhang, Susie Y Dai & Joshua S Yuan (Bioinformatics, 2011)