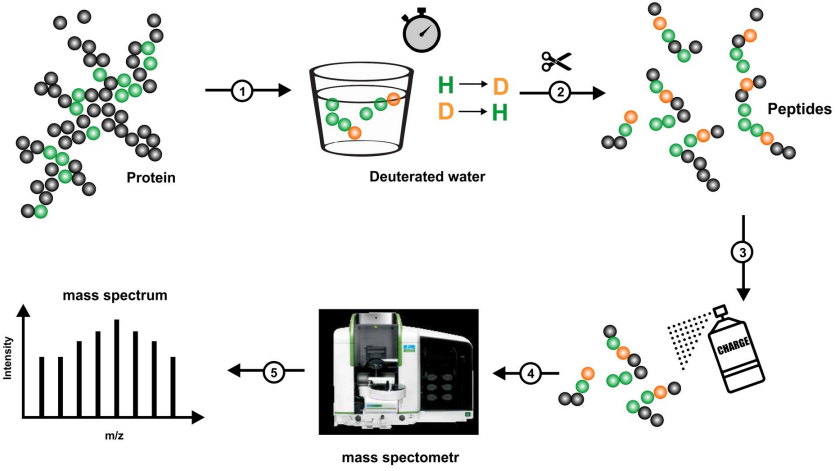
Comparing deuteration curves

Krystyna Grzesiak

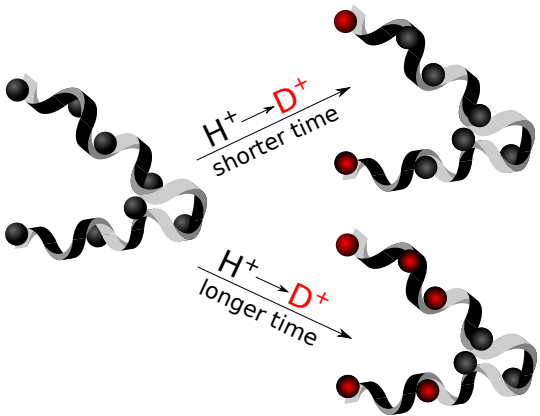
June 1, 2021

HDX: dynamics measurements of protein structure

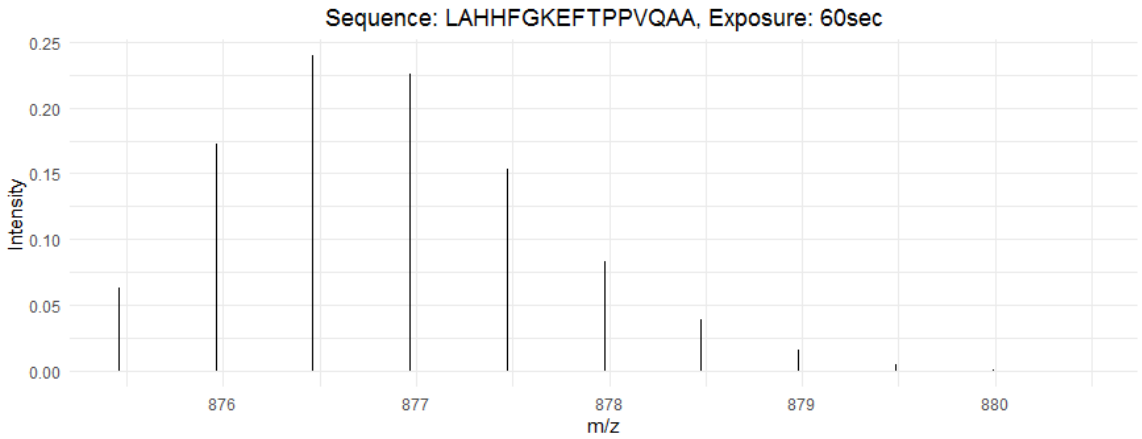


HDX: dynamics measurements of protein structure

- 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



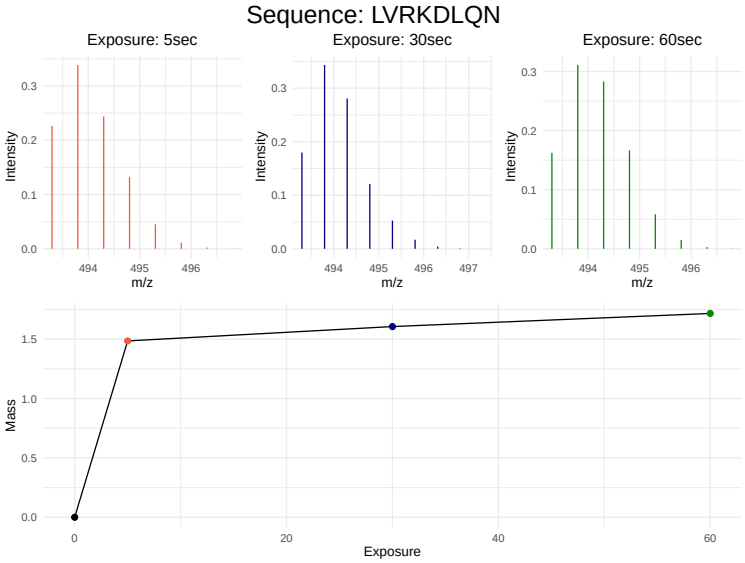
Mass spectrum



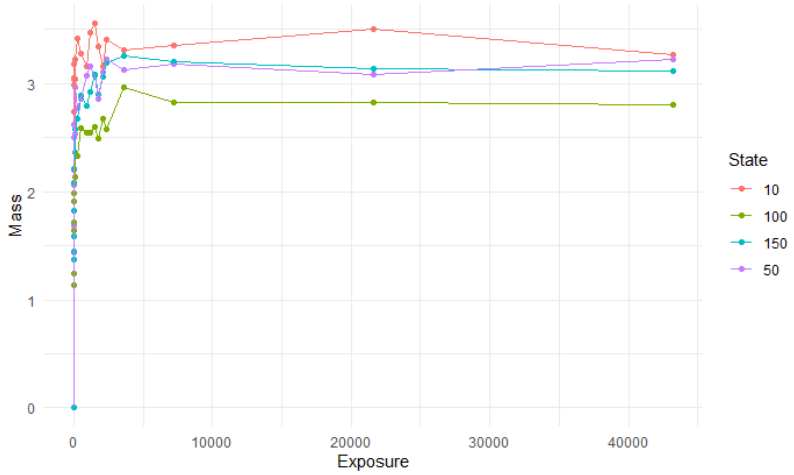
Simulation time

```
$SHCLL
Unit: milliseconds
                                          expr          min           lq         mean       median           uq          max neval
   {     simulate_rcpp(all_params[i, ], times) } 323805.71187 324775.59353 325667.15083 326036.01068 326555.02622 326704.1386    10
 {     simulate_markov(all_params[i, ], times) }     50.06764     51.08157     77.08071     51.46785     54.90911    301.9468    10

$SVKLGHPDTLNQGEF
Unit: milliseconds
                                          expr        min          lq        mean      median          uq         max neval
   {     simulate_rcpp(all_params[i, ], times) } 183544.7983 183847.21832 184394.13997 184246.83311 185043.75039 185778.4020    10
 {     simulate_markov(all_params[i, ], times) }     46.6729     46.74509     60.72676     46.92625     47.10634    180.7228    10

$HIMEDLDTNA
Unit: milliseconds
                                          expr        min         lq       mean     median         uq        max neval
   {     simulate_rcpp(all_params[i, ], times) } 96848.01266 97072.14389 97183.8649 97127.66521 97305.82859 97732.71294    10
 {     simulate_markov(all_params[i, ], times) }    46.50001    46.53818    46.7617    46.63647    46.81681    47.49714    10

$TKTYFPHF
Unit: milliseconds
                                          expr        min         lq       mean     median         uq        max neval
   {     simulate_rcpp(all_params[i, ], times) } 25116.72428 25201.31742 25311.59897 25285.08453 25448.4792 25582.61801    10
 {     simulate_markov(all_params[i, ], times) }    34.78354    34.84255    34.90605    34.89878    34.9277    35.12618    10

$GNVLVCVLAHHFGKEFTPPV
Unit: milliseconds
                                          expr        min         lq       mean     median         uq        max neval
   {     simulate_rcpp(all_params[i, ], times) } 38400.17988 38542.93172 38641.02559 38662.71431 38733.22403 38791.55223    10
 {     simulate_markov(all_params[i, ], times) }    30.53725    30.63669    30.68082    30.67195    30.69526    30.96322    10

$KSAVT
Unit: milliseconds
                                          expr        min         lq       mean     median         uq        max neval
   {     simulate_rcpp(all_params[i, ], times) } 24933.5226 25055.16236 25102.21923 25085.45363 25119.66704 25448.875    10
 {     simulate_markov(all_params[i, ], times) }    37.2085    37.38842    39.26699    37.41472    39.53007    46.163    10
```

Mass spectrum



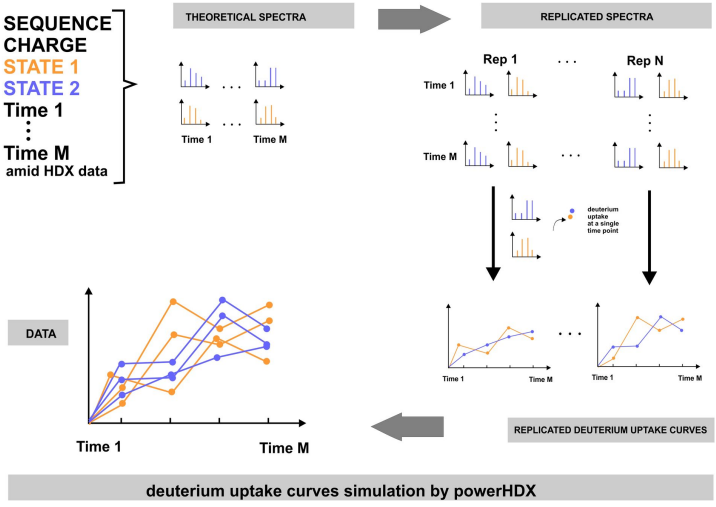
Differences in deuteration levels



powerHDX package

- spectra
- noisy spectra
- noisy curves
- power simulation

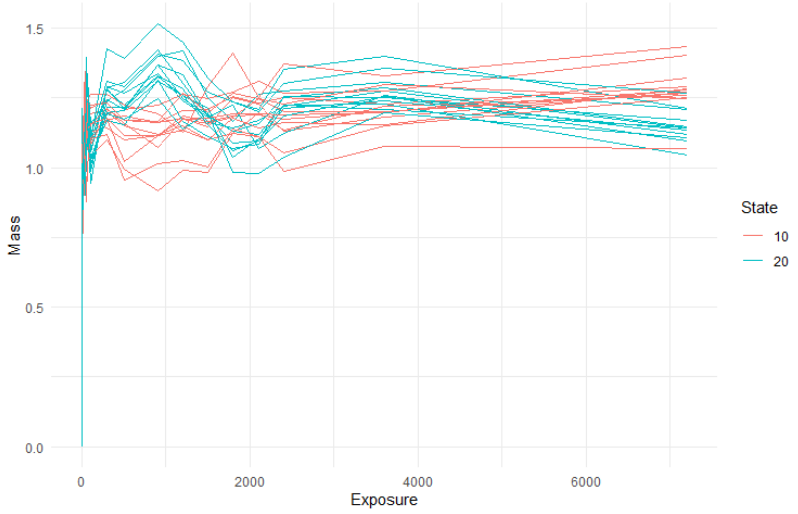
powerHDX package



Data

	Sequence	Rep	State	Exposure	Mass	Charge	Experimental_state
1	PPAQHI	1	10	0.00	0.00	1	A
2	PPAQHI	1	10	0.00	0.00	2	A
3	PPAQHI	1	10	0.00	0.00	3	A
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
502	PPAQHI	4	20	43200.00	1.10	1	B
503	PPAQHI	4	20	43200.00	1.18	2	B
504	PPAQHI	4	20	43200.00	1.01	3	B

Data



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

Existing methods

In general, differences in deuteration levels can be measured using two approaches:

1. based on linear models
   - single time-point analysis with time-point selection or multiple testing correction,
   - multivariate analysis with time as a variable,
2. based on geometrical properties of the curves
   - analysis of the area under the curve,
   - functional data analysis (functional ANOVA).

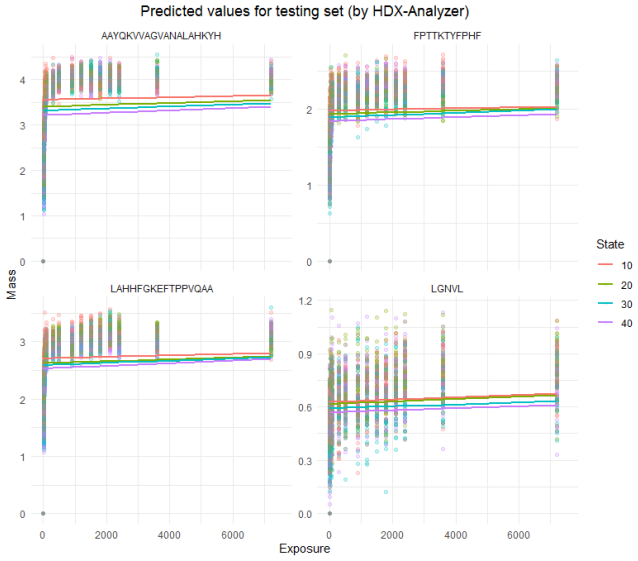
HDX-Analyzer

One of models was introduced in *HDX-Analyzer: a novel package for statistical analysis of protein structure dynamics* (2011) and included an interaction term for state and time. It is defined by the following formula

$$Y = \beta_T X_{Time} + \beta_G X_{Group} + \beta_{TG} X_{Time} \times X_{Group},$$

where Y denotes deuteration level, X_{Time} denotes exposure duration and X_{Group} is a protein state indicator.

HDX-Analyzer fit



GAM - Generalized Additive Model

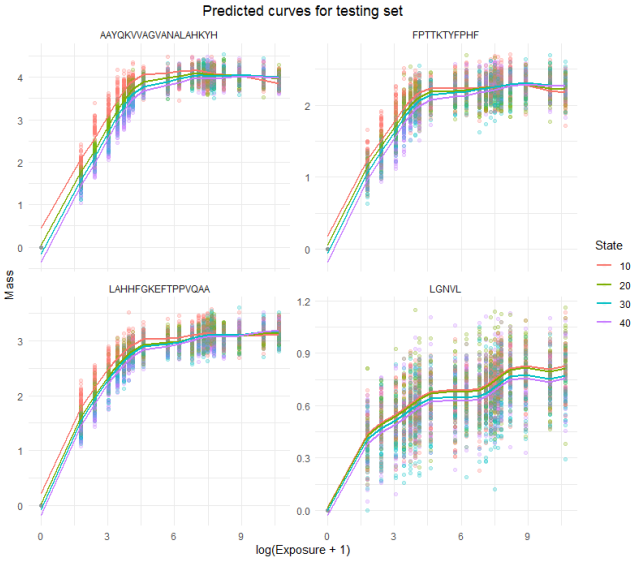
Generalized additive model for response Y (along with link function g) and predictors $x_1, \ldots, x_p$ can be represented by following formula

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

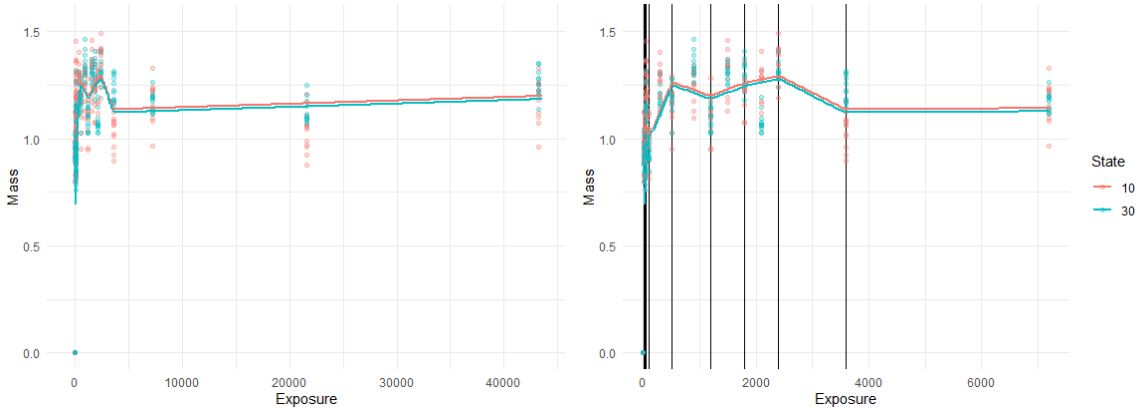
where $f_1, \ldots, f_p$ are some smooth functions.
An example basis for space of smooth functions is K-spline:

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

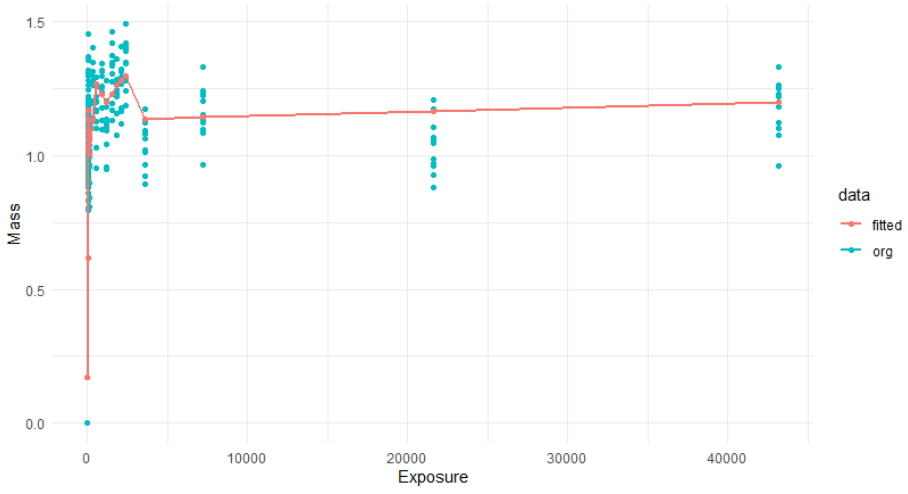
Example semiparametric fit



Regression splines



Regression splines

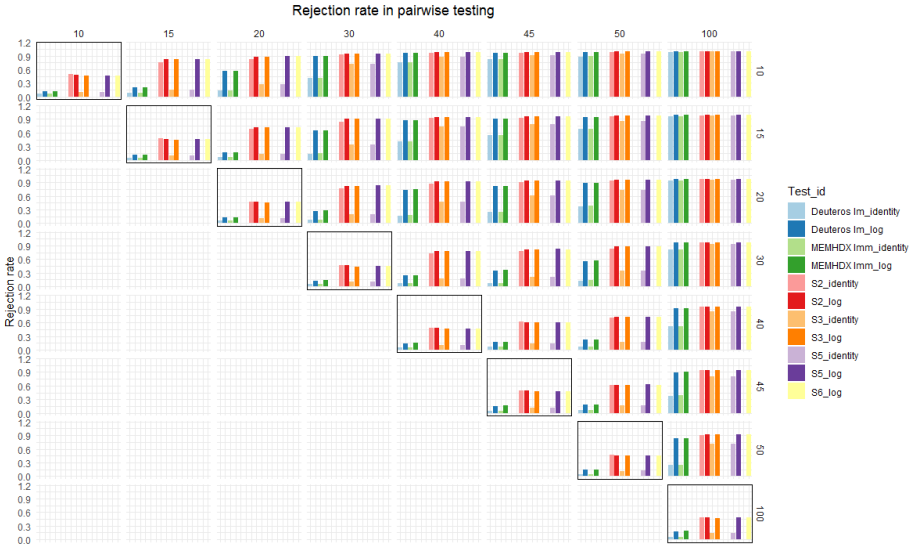


Testing

HDX - Analyzer:

```
model = lm(Mass ~ Exposure*State, data = data)
model_reduced = lm(Mass ~ Exposure, data = data)
result = anova(model, model_reduced)
```

Rejection rate



Papers

- *Simple ftting 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)