# Comparing deuteration curves

Krystyna Grzesiak
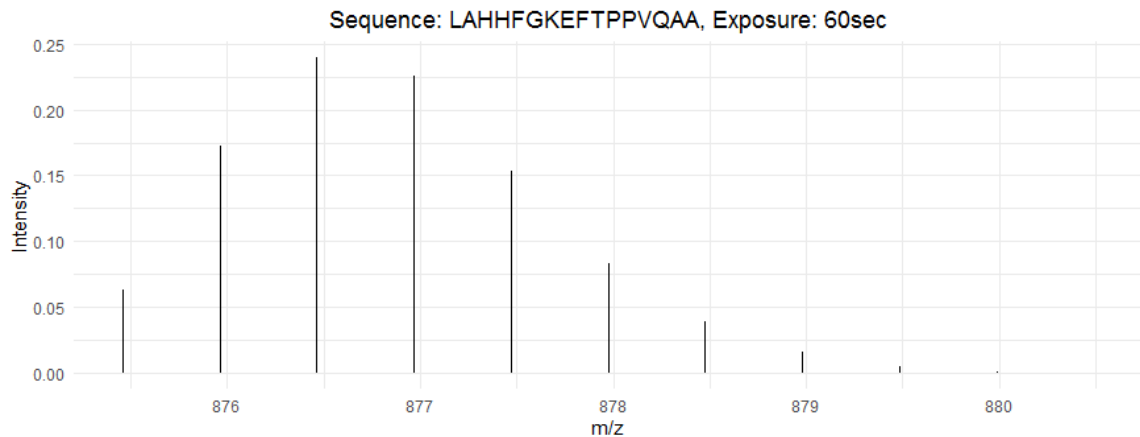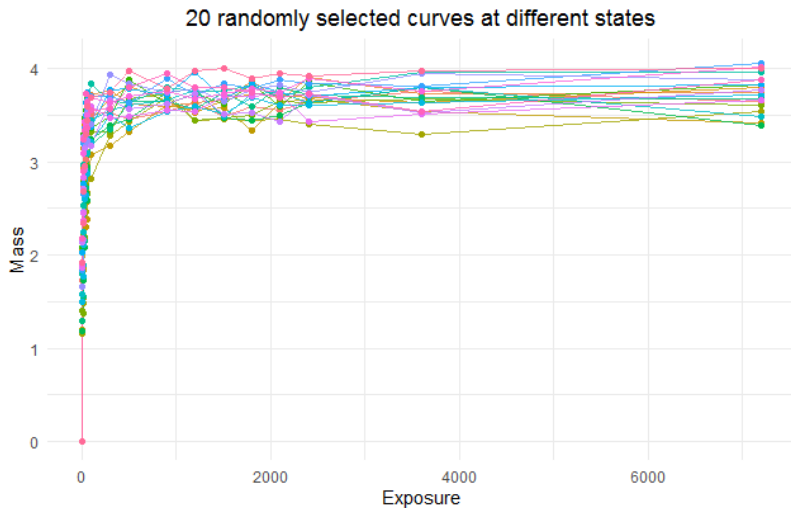
February 17, 2021

- 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
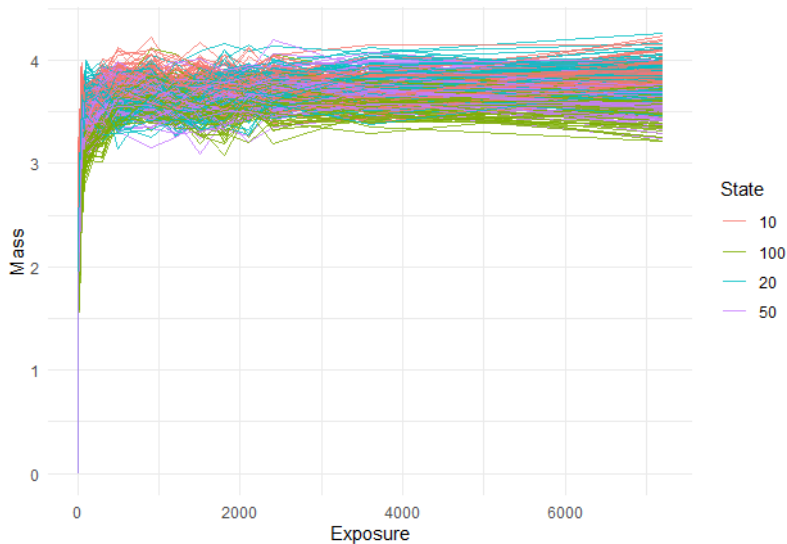


$H^+ \rightarrow D^+$
shorter time

$H^+ \rightarrow D^+$
longer time

# Mass spectrum



Sequence: LAHHFGKEFTPPVQAA, Exposure: 60sec

# Mass spectrum



20 randomly selected curves at different states
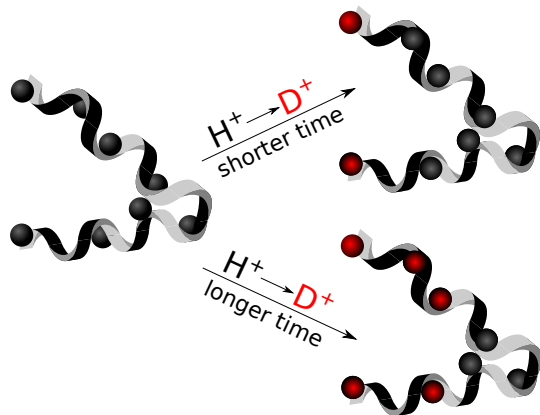
# Differences in deuteration levels

# 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

# 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.

Predicted values for testing set (by HDX-Analyzer)
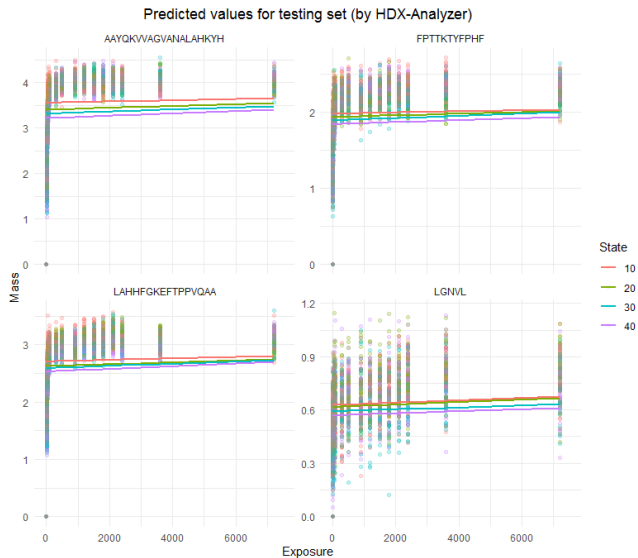
# GAM - Generalized Additive Model

Generalized additive model for response $Y$ (along with link function $g$) and predictors $x_1, \ldots, x_p$ can be represented by following formula
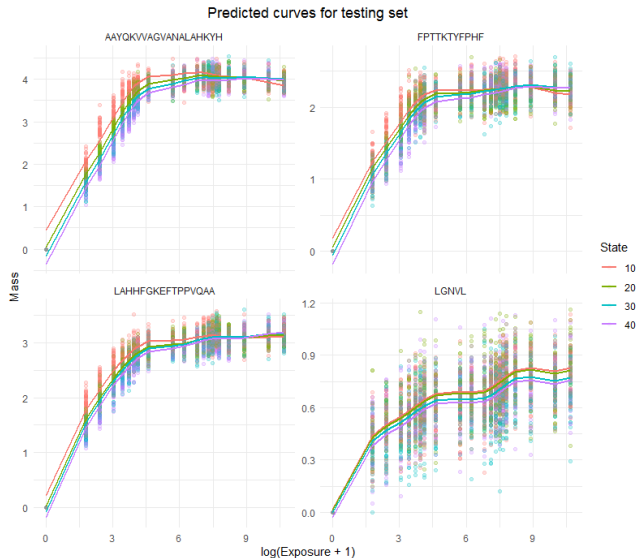
$$g(\mathbb{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p),$$
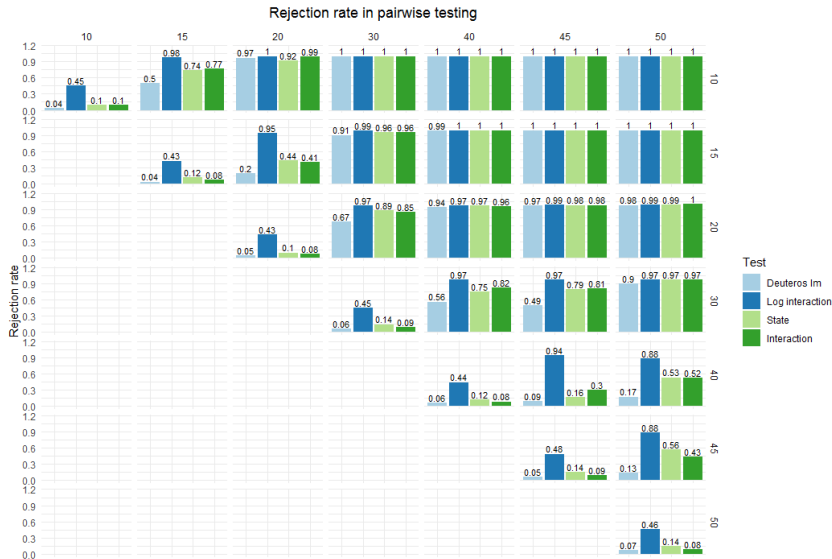
where $f_1, \ldots, f_p$ are some smooth functions.
An example basis for space of smooth functions is K-spline:

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

# Example semiparametric fit

Rejection rate in pairwise testing

# Summary

Is there a room for improvement?

- choice of basis functions
- choice of K
- transformations making data more "semiparametric-friendly"
- random effects for parametric or non-parametric terms or both
- taking covariance structure into account

# 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)