MATHEMATICA APPLICANDA
Vol. 46(2) 2018, p. 193–206
doi: 10.14708/ma.v43i1.xxx

Krystyna Grzesiak * (Wrocław)
Weronika Puchała * (Warszawa)
Michał Dadlez * (Warszawa)
Małgorzata Bogdan (Wrocław)
Michał Burdukiewicz * (Białystok)

A novel semiparametric model for hydrogen-deuterium exchange monitored by mass spectrometry data.

Abstract The hydrogen-deuterium exchange monitored by mass spectrometry (HDX-MS) is one of the methods for studying the structure of proteins. HDX-MS associates the speed of hydrogen-deuterium exchange with the regional stability of a protein. Such stability is affected by the biological state (e.g., presence of a biological ligand or lack thereof). Therefore, the changes in the protein's molecular structure caused by the biological state are inferred from the differences in the speed of hydrogen-deuterium exchange. In the following paper, we propose a test based on a mixed semiparametric model and ridge regression that allows for the accurate identification of regions with significantly different exchange speeds at two biological states. To assess its performance, we compared it with existing HDX-MS data analysis methods.

2010 Mathematics Subject Classification: Primary: 62J05; Secondary: 92D20.

Key words and phrases: deuterium uptake, hydrogen-deuterium exchange, mass spectrometry, HDX-MS, semiparametric model, spline regression.

1. Introduction

Proteins are common biomolecules facilitating many biological processes such as transport (e.g transport of oxygen, iron), catalysts (hydration of carbon dioxide), storage (ferritin stores iron ions in the liver), and many others. The structure of a protein in its most fundamental level is a chain of amino acid residues bonded by so-called peptide bond. However, in the process known as folding, this chain assumes different spatial structures. It is

* This research was financed by the Foundation of Polish Science (TEAM TECH CORE FACILITY/2016-2/2 Mass Spectrometry of Biopharmaceuticals - improved methodologies for qualitative, quantitative and structural characterization of drugs, proteinaceous drug targets and diagnostic molecules) to Michał Dadlez.

commonly assumed that the protein's function is usually defined by its structure [1]. Thus, to understand how proteins participate in biological process, we need to examine their structures.

The hydrogen-deuterium exchange monitored by mass spectrometry (HDX-MS) recently emerged as a tool for the assessment of protein structure and related properties including flexibility, stability, and affinity towards various interacting compounds, i.e. ligands and other macromolecules. This technique can measure the structural alterations caused by a biological state, an umbrella term for factors ranging from the presence of interacting molecules to the environmental conditions [12].

To do so, the proteins in different biological states are incubated in heavy water (D_2O) for a certain time. Less protected amino acid residues (e.g., more exposed to the solvent) are undergoing the exchange of hydrogens to deuters much faster than better-protected residues [2]. The most common approach to HDX-MS does not rely on measuring the mass of the whole protein, but rather its short fragments, peptides. Since the mass of deuterium is twice the size of hydrogen's, the differences in the masses of peptides allow of the comparison of the protection of given residues along the protein for different biological states.

The HDX experiments are said to be monitored by mass spectrometry as the mentioned masses are calculated based on mass spectra - an intensity versus mass-to-charge ratio (m/z) plots representing the distribution of ions (the occurrence of given m/z value). From the spectrum, we calculate a centroid which is identified as a measurement of the peptide mass after deuteration [13]. The difference between the masses before and after exposure to D_2O is called deuterium uptake. The measurements of deuterium uptake are made for several exposure times and repeated within several replications. A deuterium uptake curve is a longitudinal study of deuterium uptake over time for a single peptide at a given biological state.

HDX data analysis This paper concerns the problem of the inference about the structural (protection) factors of the peptides, based on their masses during the measurement. The challenge is to identify peptides whose deuterium uptake differ significantly between biological states. Such differences imply that biological state affects the structure of peptide's source region in the protein. The testing problem can be written down using the following hypotheses:

$$H_0\text{: } \textit{The deuterium uptake does not differ between states}$$
$$\text{vs.}$$
$$H_1\text{: } \textit{The deuterium uptake differs between states.}$$

2. Semiparametric test

We propose a novel test based on a semiparametric mixed model regularized by ridge regression for HDX-MS data. The test is meant for a specifically designed experiment in which each replication was a sample prepared by incubation of the protein stock (at a particular biological state) in the labeling solution for various time-points i.e. 0s (control), 10 s, 60 s, etc. Thus, under the term *replication* we understand a single deuterium uptake curve.

The model requires the following data:

- deuterium uptake (response) - calculated from spectrum
- Exposure time
- Biological state
- Replication ID

We describe the testing procedure applied in the semiparametric test in detail in four steps:

Step 1: Regression spline The regression spline is an extension of the linear approach that allows for nonlinear relationship between response and the features [see 9, Generalized Additive Models]. Such a model for the response Y along with predictors $X_1, \ldots, X_p$ can be represented by the following formula

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i, \tag{1}$$

where f_j, $j = 1, \ldots, p$ are smooth (nonlinear) functions. The semiparametric part of our model is a spline based on the simplest base functions f, so called truncated lines [see 3]. Namely, for y denoting deuteration level and x_T denoting exposure time we have

$$y_i = \beta_0 + \beta_1 x_{Ti} + \sum_{k=1}^{K} u_k (x_{Ti} - \kappa_k)_+. \tag{2}$$

where u_k for any $k = 1, \ldots, K$ are coefficients corresponding to k^{th} truncated line (exposure time) and for any $x, a \in \mathbb{R}$ we define

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

We create the design matrix $\mathbf{X}$ for all the time points as follows

$$\mathbf{X} = \begin{pmatrix} (x_1 - \kappa_1)_+ & \ldots & (x_1 - \kappa_K)_+ \\ \vdots & & \vdots \\ (x_n - \kappa_1)_+ & \ldots & (x_n - \kappa_K)_+ \end{pmatrix}$$

and go to step 2.

Step 2: Regularization by ridge regression We use l_2-norm penalized regression to select columns of $\mathbf{X}$. To do so, we fit the ridge regression model with response Y (deuterium uptake) and the abovementioned design matrix $\mathbf{X}$, by solving the formula [5, Shrinkage Methods]

$$\hat{\beta}_{ridge} = \underset{b \in \mathbb{R}^p}{\arg\min} \left\{ ||Y - Xb||^2 + \lambda \sum_{i=1}^{p} b_i^2 \right\} \tag{3}$$

where λ is the regularization penalty. Let us notice that the greater value of λ we set, the sparser estimator of β we obtain. At the same time, the sparser estimator of β we get, the fewer knots we select to the spline and the more 'boxy' the target model will be as shown on the figure 1. We denote the set

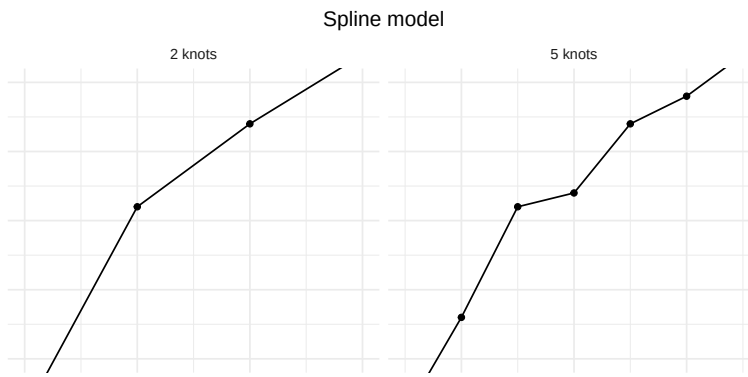


Figure 1: Splines depending on the number of knots.

of chosen knots (time points) as κ_{ridge}, and consequently the semiparametric part of the model as

$$\sum_{\kappa \in \kappa_{ridge}} u_\kappa (x_{Ti} - \kappa)_+$$

and go to step 3.

Step 3: Final model The starting point of our model is a simple linear model with an interaction term proposed in [11]

$$y_i = \beta_0 + \beta_T x_{Ti} + \beta_S x_{iS} + \beta_{TS} x_{Ti} x_{Si} + \epsilon_i \tag{4}$$

where x_S denotes protein state. Since the considered data has a longitudinal characteristic, it is desirable to fit a model with random effects [see 6]. Our model includes two random intercepts - one varying across different curves (b_{id}) and second across different time points (b_T). Next, we add penalized semiparametric part and obtain the final model:

$$y_i = \beta_0 + b_{id} + b_T + \beta_T x_{Ti} + \beta_S x_{iS} + \beta_{TS} x_{Ti} x_{Si} + \sum_{\kappa \in \kappa_{ridge}} u_\kappa (x_{Ti} - \kappa)_+ + \epsilon_i. \quad (5)$$

Step 4: Testing Comparing the differences in deuteration levels in the case of the model-based approaches comes down to determining whether the state indicator is dependent on the measurement of masses. In other words, we test whether the variable describing the protein state should be selected for the model. It can be done via the F test in the context of nested models comparison. In the case of comparing the fixed parts of mixed models, we use the analogous version of the F-test extended by Satterthwaite's method for computing the denominator degrees of freedom and F-statistics [10].

3. Simulation

We simulated the data of 73 peptide sequences using the R-package *powerHaDeX* (available on CRAN [4]). A single data set consists of 4 replications of the experiment and two biological states. Thus, there are 24 different deuterium uptake curves. The measurements of deuterium uptake were collected at the exposure times 5, 10, 20, 30, 40, 50, 60, 100, 300, 500, 900, 1200, 1500, 1800, 2100, 2400, 3600, 7200, 21600, 43200 in seconds (Figure 3).

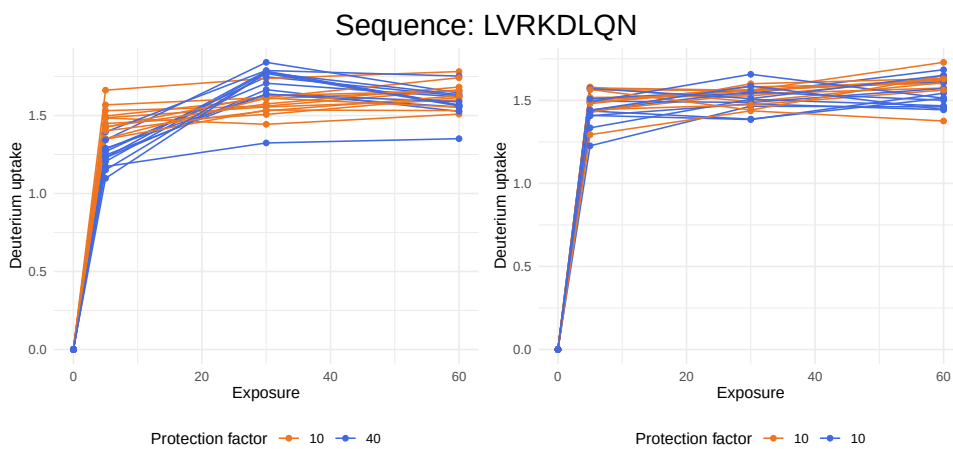


Figure 2: An example testing data set

The simulated data covers two cases. The first happens when we deal with different values of the protection factor, i.e. when the null hypothesis is false and the rejections are true discoveries (the plot on the left). The second is when the protection factors are equal between groups and the null hypothesis is true (the plot on the right). Then, the rejections are false discoveries (type I errors).

4. Results The outcome of our simulation is the set of rejection rates

in the pairwise testing procedure at the significance level 0.05 presented on three charts. Each of them is composed of a grid of average rejection rates depending on the considered hypothesis. On the grid's diagonal (in black frames) the null hypotheses are true ($Pf_1 = Pf_2$) so the rejection rate is the estimator of type I error. Beyond the diagonal, the null hypotheses are false ($Pf_1 \neq Pf_2$) so the rejection rate is the estimator of power. The three following cases are presented on the figure 3:

- Plot 1: small values of ΔPf, Pf_1, Pf_2,
- Plot 2: small values of ΔPf and great values of Pf_1 and Pf_2,
- Plot 3: great values of ΔPf.

We have also included in our comparison three tests addressing the same hypothesis:

- HDX-Analyzer model [11] - models jointly the effect of time and protein state by including regression terms for both quantities and an interaction term

$$y_i = \beta_0 + \beta_T x_{iT} + \beta_S x_{iS} + \beta_{TS} x_{iT} x_{iS} + \epsilon_i. \tag{6}$$

- MEMHDX mixed model [8] - linear mixed model including the effect of a replication:

$$y_{i,r} = \beta_0 + \beta_T x_{iT} + \beta_S x_{iS} + \beta_{TS} x_{iT} x_{iS} + w_{i,r} + \epsilon_i, \tag{7}$$

  where w_r is a random effect associated with the replication r.

- Houde's confidence intervals test [7] - community recognized test derived based on an empirical evaluation of a dataset collected by the author.

Type I error We discuss the results starting with the type I error. As we can see on the figure 3 the Houde's test turns out to be conservative when compared to the others. The null hypothesis is rejected by Houde's test on average in 1% of the cases. However, the other approaches do not fall far behind. The models HDX-Analyzer and MEMHDX generally hold the type I error at 5% level and model fitted after variable selection done by ridge regression mostly reaches 6% of false rejections.

Power The main remark that we can notice first is that in case of small differences between the protection factors all the tests have trouble with accurate rejections. It is not surprising, as for close protection factors we obtain close exchange probabilities, and consequently, close deuteration levels. Thus, such a situation requires a big sensitivity of the test. Since even a small change

of the protection factor affects the whole protein regarding its conformation and dynamics we are interested in finding a test as sensitive as it is possible, which is kind of a challenge while keeping the type I error at a small level.

As we can see on the first plot, in the case of small values of Pf_1 and Pf_2 the test based on the spline model reaches on average better power when compared to the existing tools. When ΔPf is small, the existing methods have a way lower rate of true rejections, except for Houde's intervals. The results of Houde's test are better but still differ from the semiparametric model.

On the second plot we can see that the quality of the tests' performance decreases - we obtain fewer true rejections when compared to the first case. Houde's confidence intervals test generally fared better than the other existing methods. However, in the case when $\Delta Pf = 10$ the rejection rate is close to zero for all of the three approaches. The semiparametric test handle this difficult case better. The results are equal to about 30%.

The last plot presents the case when the differences between the protection factors ΔPf are big. As we can see, the case when $Pf_1 = 300$ and $Pf_2 = 400$ turned out to be problematic for the existing approaches. The spline model fared better there. We can see the similar situation for $Pf_1 = 90$ and $Pf_2 = 50$. However, in this case, confidence intervals reached better results than before. For other pairs of protection factors, we observe an increase in the rate of true rejections. The values generally fluctuate between 85 and 100% except for Houde's test which turned out to be the worse.

Conclusions The performance of linear models leaves a lot to be desired. It is not surprising as the greatest variability of the values of deuterium uptake occurs during the short exposure. Since both HDX-Analyzer (estimated by OLS) and MEMHDX (estimated by MME) are sensitive to outliers we expect them to fit more the measurements of masses obtained using long incubation. The methods of robust regression such as Huber or bisquare weighting (in the case of HDX-Analyzer) may improve the results there.

The fact that the performance of Houde's test is better than the performance of the linear models can be also intuitively explained. Unlike other existing approaches, Houde's confidence intervals test takes into account the variability (averaged over time) of deuterium uptake along with the differential values. However, it is not surprising that its performance is poor in the case of close protection factors and that the test becomes quite conservative - averaging the differences over all the time points yields lower precision.

The regression spline model is not as exposed to the impact of outliers as HDX-Analyzer and MEMHDX (although it depends on the number of knots) because of its local characteristics. Thus, it is capable of modeling the deuterium uptake reasonably well along the entire curve. Moreover, the effects associated with the subjects and exposure time are modeled. It makes

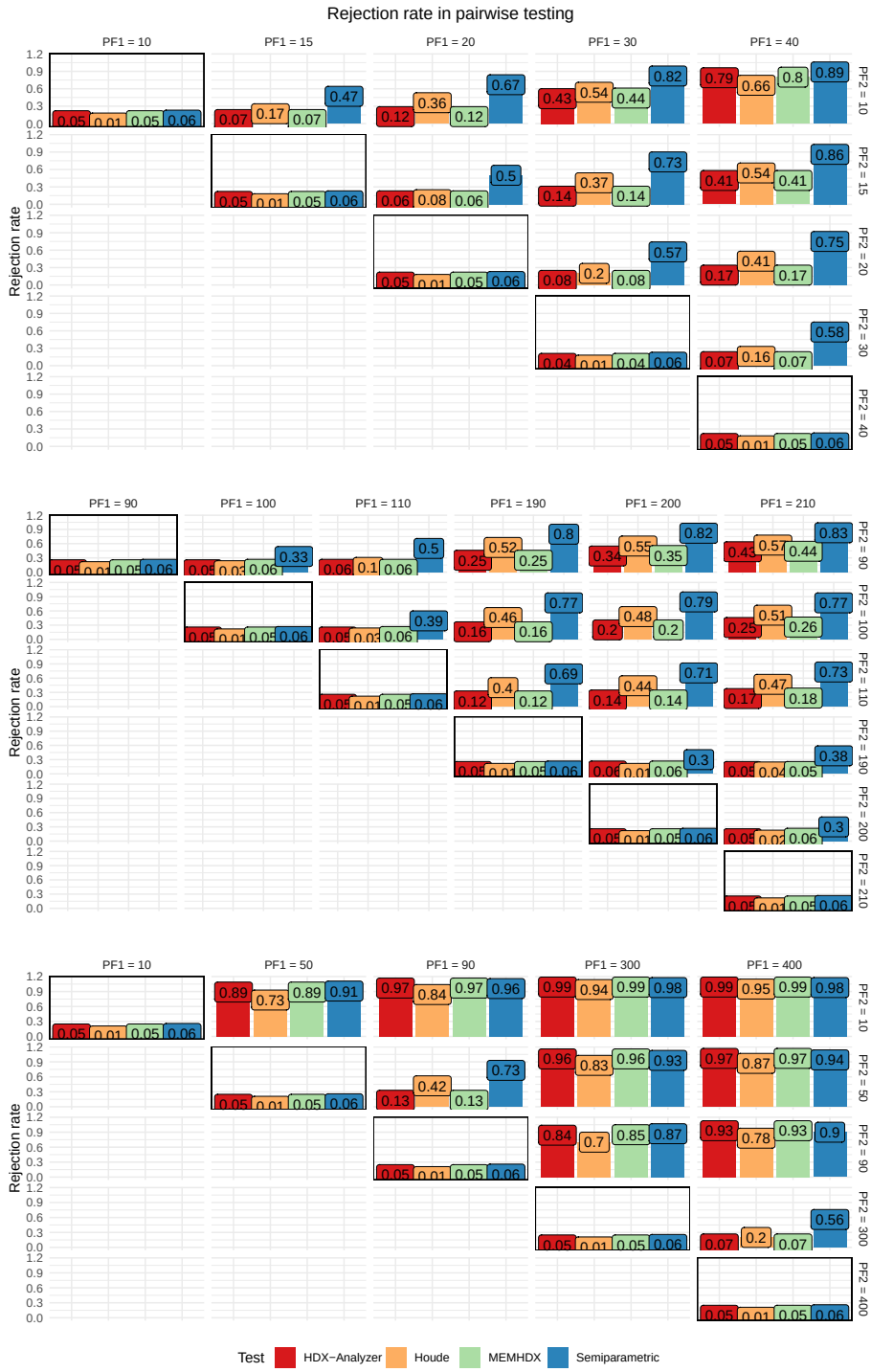


Figure 3: Rejection rate in pairwise testing at the significance level 0.05.

the test more sensitive to the protein protection level changes even when the differences are imperceptible by a human.

The versatility of the proposed model is sufficient to employ it in similar problems, for example deuterium-hydrogen exchange (when we monitor the exact opposite of the HDX) or global HDX (when we focus on the whole protein instead of peptides). Therefore, we hope that our semiparametric model will be a valuable contribution to still too limited array of tools necessary to properly analyze results of HDX-MS

5. References

[1] C. B. Anfinsen. The formation and stabilization of protein structure. *Biochemical Journal*, 128(4):737, 1972. cited on p. 194.

[2] A. Berger and K. Linderstrøm-Lang. Deuterium exchange of poly-dl-alanine in aqueous solution. *Archives of Biochemistry and Biophysics*, 69:106–118, 1957. cited on p. 194.

[3] M. Durbán, J. Harezlak, M. P. Wand, and R. J. Carroll. Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, 24(8):1153–1167, 2005. cited on p. 195.

[4] K. Grzesiak and M. Staniak. *powerHaDeX: Efficient Simulation of HDX-MS Data and Tools for the Statistical Analysis*, 2021. R package version 1.0. cited on p. 197.

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001. cited on p. 196.

[6] C. R. Henderson. SIRE EVALUATION AND GENETIC TRENDS. *Journal of Animal Science*, 1973(Symposium):10–41, 01 1973. cited on p. 196.

[7] D. Houde, S. A. Berkowitz, and J. R. Engen. The utility of hydrogen/deuterium exchange mass spectrometry in biopharmaceutical comparability studies. *Journal of pharmaceutical sciences*, 100(6):2071–2086, 2011. cited on p. 198.

[8] V. Hourdel, S. Volant, D. P. O'Brien, A. Chenal, J. Chamot-Rooke, M.-A. Dillies, and S. Brier. MEMHDX: an interactive tool to expedite the statistical validation and visualization of large HDX-MS datasets. *Bioinformatics*, 32(22):3413–3419, 07 2016. cited on p. 198.

[9] J. Z. Huang. An introduction to statistical learning: With applications in r by gareth james, trevor hastie, robert tibshirani, daniela witten, 2014. cited on p. 195.

[10] A. Kuznetsova, P. B. Brockhoff, R. H. Christensen, et al. lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82(13):1–26, 2017. cited on p. 197.

[11] S. Liu, L. Liu, U. Uzuner, X. Zhou, M. Gu, W. Shi, Y. Zhang, S. Y. Dai, and J. S. Yuan. Hdx-analyzer: a novel package for statistical analysis of protein structure dynamics. *BMC bioinformatics*, 12(1):1–10, 2011. cited on pp. 196 and 198.

[12] G. R. Masson, J. E. Burke, N. G. Ahn, G. S. Anand, C. Borchers, S. Brier, G. M. Bou-Assaf, J. R. Engen, S. W. Englander, J. Faber, et al. Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (hdx-ms) experiments. *Nature Methods*, 16(7):595–602, 2019. cited on p. 194.

[13] W. Puchała, M. Burdukiewicz, M. Kistowski, K. A. Dąbrowska, A. E. Badaczewska-Dawid, D. Cysewski, and M. Dadlez. HaDeX: an R package and web-server for analysis of data from hydrogen–deuterium exchange mass spectrometry experiments. *Bioinformatics*, 36(16):4516–4518, 06 2020. cited on p. 194.

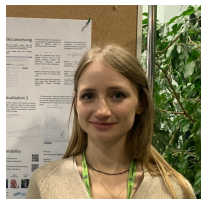
Model semiparametryczny dla danych dotyczących wymiany izotopowej wodoru na deuter monitorowanej spektrometrią mas.

Streszczenie Wymiana wodór-deuter monitorowana spektrometrią mas (HDX-MS) jest jedną z metod badania dynamiki i ukształtowania struktury białek. Podczas inkubacji w ciężkiej wodzie (D2O) reszty aminokwasowe bardziej wystawione na wymianę wodoru na deuter przechodzą ją znacznie szybciej niż inne. Pomiary mas peptydów ujawniają trwałość sieci wiązań wodorowych i regiony z ograniczoną dostępnością rozpuszczalnika. Szybkość wymiany może zależeć od stanu biologicznego białka (np. z lub bez obecnością ligandu) tzn. możebyć powiązana ze zmianami w ukształtowaniu jego struktury. W tym artykule proponujemy test opierający się na mieszanym modelu semiparametrycznym i regresji grzbietowej, który pozwala na dokładną identyfikację peptydów z istotne różnymi prędkościami wymiany w różnych stanach biologicznych. W celu oceny wyników testu, porównaliśmy go z innymi metodami służącymi do analizy danych z eksperymentów HDX-MS.

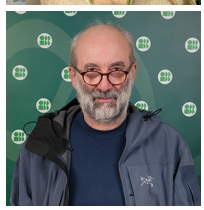
Klasyfikacja tematyczna AMS (2010): 62J05; 92D20.

Słowa kluczowe: deuteracja, wymiana izotopowa wodoru na deuter, spektrometria mas, HDX-MS, model semiparametryczny.

Krystyna Grzesiak has completed her M.Sc Mathematics with the Data Analysis specialization at the University of Wroclaw. She defended her master's thesis in 2021 under the supervision of doctor Michał Burdukiewicz and professor Małgorzata Bogdan.



Weronika Puchała is a physicist persuading her career in science in the field of biohysics and proteomics in the Dadlez Lab. Her PhD project is focused on the analysis of HDX-MS data. She is the main developer of the HaDeX software.



Michał Dadlez is the head of the Mass Spectrometry Laboratory at Institute of Biochemistry and Biophysics PAS since 2001. His scientific interests include proteomics (with a focus on structural proteomics) and metabolomics, including small-molecule quantification.

Małgorzata Bogdan is an Associate Professor at the Institute of of Mathematics of Wroclaw University and Guest Professor at University of Lund. She is interested in the analysis of high dimensional data and its applications in statistical genetics.

Michał Burdukiewicz is a research assistant in the Centre for Clinicial Research at Medical University of Białystok. His research interests cover applications of machine learning in functional analysis of peptides and proteins and proteomics, especially hydrogen-deuterium exchange monitored by mass spectrometry.

Krystyna Grzesiak
University of Wrocław
Faculty of Mathematics and Computer Science, University of Wrocław, Fryderyka Joliot-Curie 15, 50-383 Wrocław
E-mail: krygrz11@gmail.com

Weronika Puchała [1]
Polish Academy of Sciences
Institute of Biochemistry and Biophysics, Polish Academy of Sciences, FAdolfa Pawińskiego 5A, 02-106 Warszawa
E-mail: puchala.weronika@gmail.com

Michał Dadlez [1]
Polish Academy of Sciences
Institute of Biochemistry and Biophysics, Polish Academy of Sciences, FAdolfa Pawińskiego 5A, 02-106 Warszawa
E-mail: michald@ibb.waw.pl

Małgorzata Bogdan
University of Wrocław
Faculty of Mathematics and Computer Science, University of Wrocław, Fryderyka Joliot-Curie 15, 50-383 Wrocław
E-mail: malgorzata.bogdan20@gmail.com

Michał Burdukiewicz [1]
Medical University of Białystok
Jana Kilińskiego 1, 15-089 Białystok

E-mail: michalburdukiewicz@gmail.com

(Received: 11th of January 2017; revised: 5th of June 2017)