Michele Stofella

Estimating Protection Factors from HDX-MS data

Contents

1 Introduction

Hydrogen deuterium exchange probed by mass spectrometry (HDX-MS) is a promising technique to highlight both structural and dynamical properties of proteins. If a protein is put in a solution containing deuterium oxide $^{2}H_2O_2$, a spontaneous exchange of amide hydrogens with deuterium from solvent occurs [2]. Since deuterium is heavier than hydrogen, the mass of the protein changes and the time-dependent change in mass can be detected by mass spectrometry.

The change in mass of any fragment of the polypeptide chain depends uniquely on the rate of exchange of its amide hydrogens. However, determining the exchange rate from the change in mass is generally non possible, unless time-resolved measurements are available for several overlapping peptides that cover the whole sequence. In this case, exchange rates (or equivalently, as we will see, protection factors) can be extracted and the uniqueness of the solution depends on the degree of peptide overlapping. In most cases, the solution is not unique and several alternatives should be considered.

In the present article, we replicate the statistical method implemented by Skinner [1] that aims to cluster the multiple sets of solutions in order to reduce their number. As reported by Skinner, "the degeneracy of the solution depends on the number of peptides and overlap (the more the better), on their lenght (the shorter the better) and on the range of times at which the measurement has been measured (the broader the better)".

The goal of the project is to get familiar with HDX-MS data, replicating the statistical method implemented by Skinner [1].

1.1 Hydrogen deuterium exchange

Hydrogen deuterium exchange is the spontaneous exchange of the amide hydrogens of a protein with deuterium from solvent containing deuterium oxide. Exchange occurs faster for amides that are solvent-exposed and/or amides that are not involved in hydrogen bonds.

When the protein is fully exposed, the exchange of the amide follows a first order kinetics with rate

k_{int}, that depends on temperature, solution pH and side chains of the two neighboring residues [3]. Instead, if the protein is folded, a local opening of the structure is required by the exchange of the amide hydrogen, thus the process can be schematised in two steps:

$$NH_{cl} \xleftrightarrow{k_{cl}/k_{op}} NH_{op} \xrightarrow{k_{int}} ND_{op} \tag{1}$$

where k_{cl} and k_{op} are the closing and opening rates. The observed deuterium uptake rate can be expressed as follows:

$$k_{obs} = \frac{k_{int}k_{op}}{k_{int} + k_{op} + k_{cl}} \tag{2}$$

Considering the protein in the native state, i.e. $k_{cl} >> k_{op}$, two regimes are generally used to describe the hydrogen deuterium exchange kinetics:

- EX1 limit: $k_{int} >> k_{cl}$, thus $k_{obs} = k_{op}$. In this regime, the exchange is limited by slow conformational changes due to golbal unfolding or cooperative changes in quaternary structure;
- EX2 limit: $k_{cl} >> k_{int}$, thus

  $$k_{obs} = \frac{k_{int}}{P} \tag{3}$$

  where $P = k_{cl}/k_{op}$ is the protection factor for a particular amide hydrogen. This limit governs the exchage under native conditions and is sensitive to local stability.

Protection factors contain both structural and dynamic information: the degree of protection of amide hydrogen from solvent deuterium correlates to the degree of involvement in secondary and tertiary structure [4].

1.2 HDX-MS Data

HDX-MS data rely on the difference in mass between the deuterated and the non-deuterated polypeptide chains. As mentioned above, it is not possible to determine a finite number of sets of protection factors from data coming from the whole chain. In fact, to obtain specific information, the polypeptide chain is fragmented through proteolysis at low pH and low temperature.
Thus, the data we are dealing with are measurements for deuterium uptake for specific fragments covering the whole chain. Note that these data do not provide any information about the exchange rate (or protection factors) of individual residues.
The problem we aim to solve is to find the set of protection factors $\{P_i\}$ such that the deuterium uptake D_j for the peptide j, starting at residue m_j and n_j residues long, can be written at time t_k as

$$D_j(t_k, \{P_i\}) = \frac{1}{n_j} \sum_{i=m_j+1}^{m_j+n_j-1} \left(1 - e^{-\frac{k_{int}}{P_i}t_k}\right) \tag{4}$$

where P_i is the protection factor of the i-th residue. Note that the first residue is ignored in the sum because it becomes amine during proteolysis and exchanges rapidly during MS analysis.

2 Methods

In order to estimate protection factors from HDX-MS data, we follow the method implemented by Skinner [1], assuming deuterium uptake follows equation (4). Let us write the problem in the following way: we look for the set of protection factors $\{P_i\}$ such that

$$D_j^{pred}(t_k) = D_j^{exp}(t_k) + \epsilon_{j,k} \tag{5}$$

for each available fragment j and time k, where $\epsilon_{j,k}$ is the difference between experimental and theoretical value.
This problem corresponds to find the set of protection factors that minimizes the cost function

$$C(t_k, \{P_i\}) = \sum_j \sum_k w_{jk} \left[D_j^{pred}(t_k, \{P_i\}) - D_j^{exp}(t_k)\right]^2 \tag{6}$$

where w_{jk} are weights that we will not take into account since we will deal with *in silico* data; for real-world data, if at time t_k several measurements are available, an appropriate choice of the weights could be the inverse of stardard deviations of such measurements.

2.1 Extracting protection factors from HDX-MS data

In order to extract protection factors out of experimental data, we follow the following procedure:

- Extract 5000 random sets of protection factors: random protection factors are bounded to $0 \leq ln(P) \leq 20$. This means assuming that the exchange rate of an amide can be as fast as in a completely unstructured peptide and up to 5×10^8 times slower;
- Evaluate the cost function (equation 6) for each set;
- Select the set with the lowest cost function as initial condition for a least-squares minimization, performed through a squential programming approach implemented in SciPy [5].

The whole procedure is repeated 200 times and only the final sets with cost function lower than a tolerance threshold, set at 2×10^{-5}, are considered[1].

2.2 Descriptive statistics

As mentioned in the **Introduction**, several solutions may occur. In order to analyze the results, we compute the median and the interquantile range of the predicted $ln(P)$ for each residue and we associate these values to the estimated protection factors and their variances.

2.3 Clustering

To further reduce the number of solutions, we apply the model-based clustering method [6] implemented in the Mclust package [7]. We give here a short description of the algorithm.
Let Λ be the number of clusters and suppose the polypeptide chain of lenght L is fully covered by resolved peptides. We are looking for solutions

$$\vec{\mu} = \begin{pmatrix} \mu_{\lambda_1} \\ ... \\ \mu_{\lambda_L} \end{pmatrix}$$

with $\lambda = 1, ..., \Lambda$. To find the solutions, the following likelihood has to be maximized:

$$\mathcal{L}(\pi_1, ..., \pi_\Lambda, \vec{\mu}_1, ..., \vec{\mu}_\Lambda, \Sigma_1, ..., \Sigma_\Lambda | \{P_i\}) = \prod_{j=1}^{J} \sum_{\lambda=1}^{\Lambda} \pi_\lambda f_{\vec{\mu}_\lambda, \Sigma_\lambda}(\{P\}_j) \tag{7}$$

where π_λ is the fraction of data vectors belonging to cluster λ and $f_{\vec{\mu}_\lambda, \Sigma_\lambda}$ is a multivariate normal distribution. The number of solutions Λ is chosen according to the lowest Bayesian Information Crtierion. Notice that the estimate $\vec{\mu}_{\lambda_i} \pm \sigma_{\lambda_{ii}}$ corresponds to the $\approx 70\%$ confidence interval. See [6] and [7] for further details.

[1] In the original article, Skinner uses only the sets with exact agreement with experimental data, i.e. the ones with null cost function (C=0). We used a tolerance threshold since, setting a fixed value of repetitions of the algorithm at 200, we obtain very few points with cost function equal to zero. The time occured for a simulation with 200 repetitions on the PC used is approximately 3 hours for the analysis of all the 7 peptides (see section **Test data**). To summarize, the threshold is here applied for lack of computational power, but is not necessary for the analysis of *in silico* data without experimental errors.

3 Test data

Many HDX-MS experimental data are available, but real-world examples consider proteins too large to be analyze with this method on a personal computer. However, *in silico* data can be generated strating from a fictitious sequence. The sequence here analyzed is the same used as test case by Skinner [1].
The synthetic data are exctracted from a 15-residue sequence (IDSQVLCGAVKWLIL) with a reference set of protection factors (figure 1, left). 7 partially overlapping fragments are arbitrarily considered: taking into account all the peptides, the whole sequence is covered (figure 1, right).

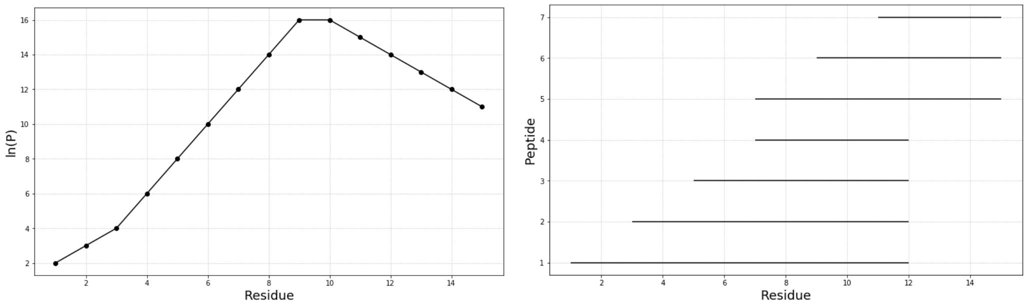

Figure 1: On the left, the fixed set of protection factors for the 15-residue sequence used to generate *in silico* data; on the right, the 7 partially overlapping fragments used for the analysis.

Following Bai [3] and the scripts implemented by Skinner [1], we calculated the intrinsic exchange rate k_{int} for each residue at pH 7 and temperature 300 K. Once the intrinsic exchange rates are known, using equation (4) it is possible to evaluate the deuterium uptake for each one of the seven fragments (figure 2). To generate experimental data, we set three time points, namely 3 s, 50 min and 280 h and we evaluate the deuterium uptake of each peptide at those points (dots in figure 2).

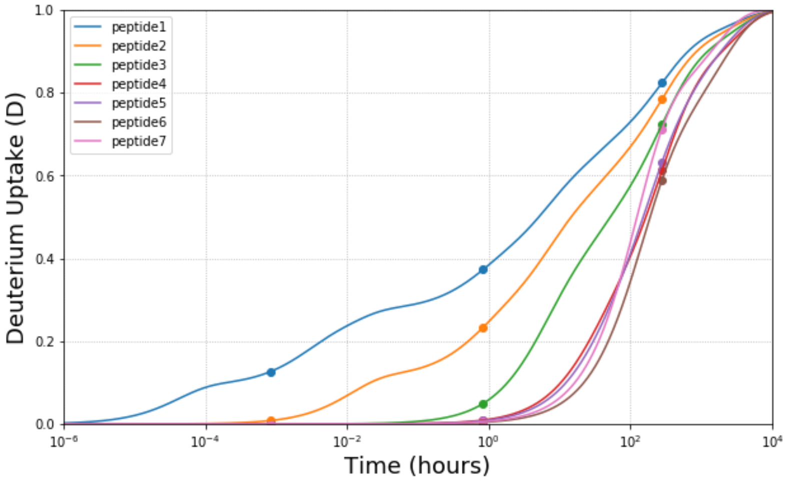

Figure 2: Deuterium uptake for each peptide evaluated using equation (4). The dots represent experimental data evaluated at time 3ss, 50 min and 280 h.

The experimental data (dots in figure 2) are the starting point of the analysis. In principle, we should be able to extract from these points the protection factors in figure 1. We will see that even is a simple case like this, it is not possible to find a unique solution.

4 Results

In order to analyze the outcomes of the method, we applied it in three different cases, considering respectively: peptides 1 and 3; peptides 1, 2, 3 and 5; all the peptides (1 to 7). See figure 1 to look at the peptides taken into account. The results are shown in figure 3.

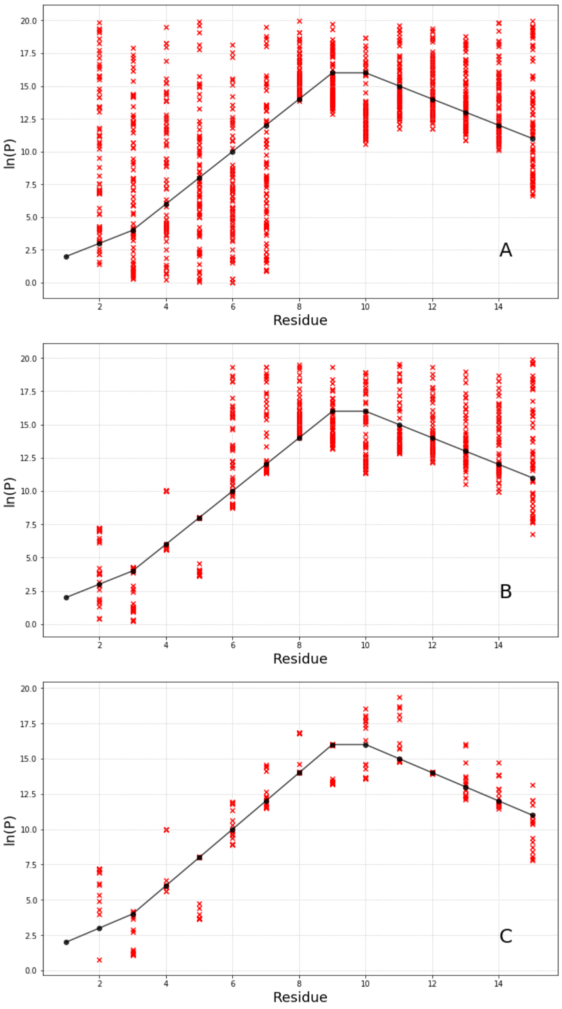

Figure 3: Predicted protection factors with deuterium uptake evaluated for peptides 1 and 5 (A), 1-3 and 5 (B) and all 7 peptides (C). Reference protection factors are shown in black. Red crossed represent the estimated protection factors with cost function lower than 2×10^{-5}.

As shown in figure 3, it is possible to evaluate protection factors even if only two peptides are available, in this case peptide 1 and 5 (A). However, the evaluated protection factors have very high degeneracy: notice that for residues 2-7, the whole range $0 \leq ln(P) \leq 20$ is almost covered. When the number of overlapping peptides increases, the degeneracy progressively decreases (B and C). In addition, some discrete values of predicted protection factors seem to arise: for residues 4 and 5, two values are available, while for residue 12 a unique protection factor can be evaluated. These results show the general feature that the estimate of the protection factor for a given residue is linked to the ones of its neighbours and, as a consequence, the number of solutions progressively decreases if neighbouring residues are taken into account simultaneously.

In order to estimate the possible solutions, we use the two ways described in the **Methods** section, namely the descriptive statistics and clustering approaches.

4.1 Results of descriptive statistics

We evaluated the median and interquantile range for each residue in the three cases described above. The median is associated to the estimate of the protection factor, while the interquantile range represent the error associated to the estimate. As a result, we obtain the estimated protection factors in figure 4.

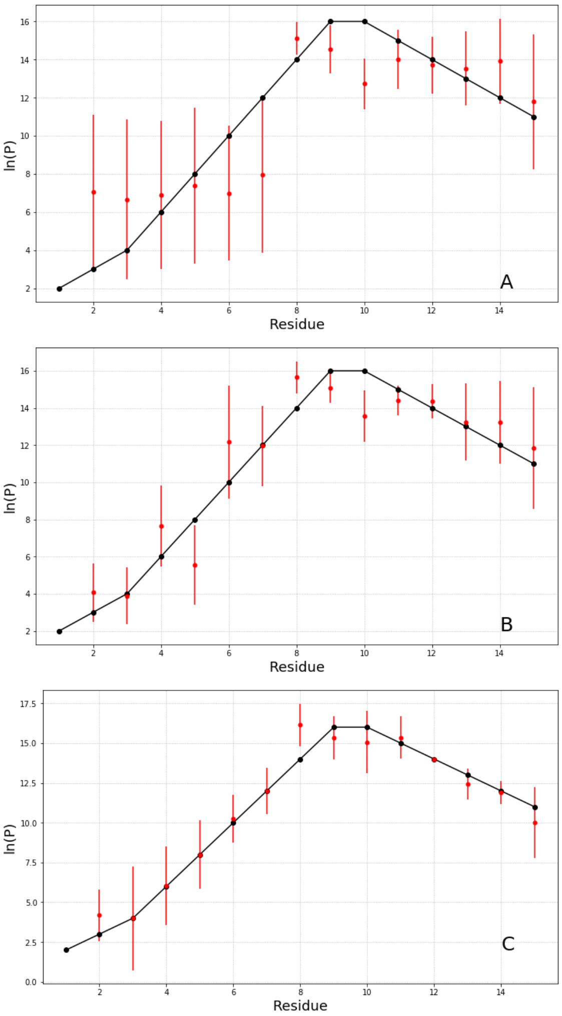

Figure 4: Estimated protection factors via descriptive statistics while considering peptides 1 and 5 (A), peptides 1-3 and 5 (B) and peptides 1-7 (C). Reference protection factors are shown in black. Red dots represent the medians of each residues, while red lines are the interquantile ranges.

Via increasing the number of overlapping peptides, we notice that the error bars (i.e. the interquantile ranges) progressively reduce. This is mostly relevant in residues 2-7 in plot (A) and (B): for example, for residue 2 the confidence interval is reduced from 8.13 to 3.18 (≈ 60%), while for residue 3 from 8.37 to 3.07 (≈ 65%). In addition, the estimated value (i.e. the median) get closer and closer to the true value.
Descriptive statistics helps understanding a possible range of values in which the protection factors lay, but it does not consider some features highlighted before, like the fact that a set of discrete values may arise for some residue.

4.2 Results of clustering

In order to further reduce the number of solutions and to overcome the failings of descriptive statistics, the clustering method described in section **Methods** can be performed. The method, implemented in the R package Mclust, chooses the number of solutions Λ through BIC: in this case $\Lambda = 8$. This number may depend on the number of points considered, i.e. the red crosses in figure 3, and on the threshold imposed on the cost function (in this case, we considered protection factors with cost function lower than 2×10^{-5}). Thus it is not a surprise that Skinner [1] found for the same data $\Lambda = 7$[2].

For each cluster, the means $\vec{\mu}_\lambda$ and the variances $\vec{\sigma}_\lambda$ were computed. The results of the clustering approach are shown in figure 5. As mentioned in the **Methods** section, the interval $\mu_{i,\lambda} \pm \sigma_{i,\lambda}$ represent the $\approx 70\%$ confidence interval.

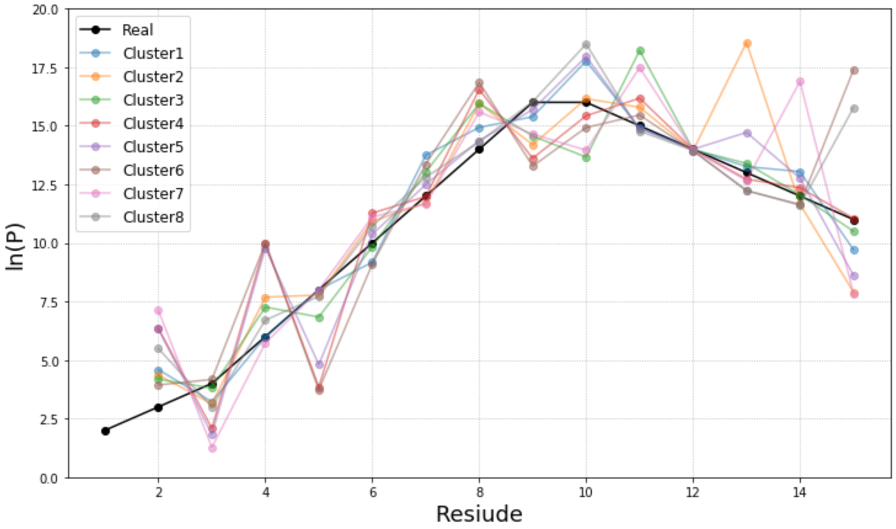

Figure 5: Means of the $\Lambda = 8$ clusters identified by the clustering approach. Each cluster represent a possible solution for protection factors considering all the peptides (1-7). Reference protection factors are shown in black.

The results show several interesting features of the algorithm. While descriptive statistics returned a continuous range of values for the protection factors of each residue, here we get $\Lambda = 8$ solutions for each residue, thus the degeneracy of the solution decreases.

Furthermore, we have seen that for residues 4 and 5 two discrete values arise, thus it could be easy to conclude that the solutions for protection factors of residues 4 and 5 are four different solution (i.e. the combinations of the two solutions of each residue). On the contrary, the clustering approach shows that only two of the solutions have to be considered: the others do not reproduce the expected deuterium uptake. This highlights the dependence of the protection factor of a residue with the protection factors of the adjacent residues and this property could not be seen through the descriptive statistics approach.

5 Conclusions

HDX-MS is a promising technique to obtain structural and dynamical information of proteins. Results of the studies often deal only with qualitative information at the resolution level of several amino acids. Here we show the method implemented by Skinner [1] through which it is possible to estimate constraints for protection factors at the resolution level of single residues.

In order to handle experimental data, we used a reference set of protection factors and a 15-peptide

[2] We decided to take 200 sets of protection factors and from these we extracted the ones with cost function lower than 2×10^{-5}. Skinner [1] used a more accurate approach: the convergence of the random search is reached when the number of cluster decreases no more by adding more sets of protection factors.

sequence (figure 1) from which we evaluated the deuterium uptake at three time points (figure 2). We then applied the Skinner method to these data points, highlighting several features of the results.
First of all, a random search of the estimate of the set of protection factors leads to a degeneracy that decreases if we consider more overlapping fragments of the whole peptide chain, as can be seen from figure 3. The random search highlights the possibility to have a finite number of discrete solutions for the protection factor of a given residue.
To summarize the results, the median and interquantile range were calculated for the protection factor of each residue (figure 4). The interquantile ranges decrease if we consider more overlapping peptides, as expected from the properties already shown by the random search.
In order to reduce the number of solutions, the clustering approach was applied. It highlighted the dependence of the protection factor of the single residue on the protection factors of its neighbours. Most of all, the method represents a search for a self-consistent solution to the problem. In fact, it gives as result 8 alternative sets of protection factors that fit the experimental data (figure 5).

The purpose of the present project was to get in touch with HDX-MS data and the Skinner method to analyze these kind of data. The implementation from scratch of the scripts shows results in agreement with the ones in the original article [1].
The work has been performed on a PC, thus the results are less accurate than the ones proposed by Skinner; for the same reason, we limited our study to the simple case of synthetic data, without dealing with real-world examples. The usage of more computing power should give more accurate results and should expand the range of treatable data. Of course, the application of the method to real data should be the natural continuation of the project.

Supplementary Material

Codes (implemented in Python and R) and images are available in the following Github repository:

https://github.com/michelestofella/skinner

Notice that the scripts used for the analysis in the present article are mostly implemented from scratch. For the original codes, see the original article by Skinner [1].

Bibliography

[1] Skinner et al., Estimating Constraints for Protection Factors from HDX-MS Data. Biophysical Journal (2019), https://doi.org/10.1016//j.bpj.2019.02.024
[2] Linderstrom-Lang, K., 1955. The pH-dependence of the deuterium exchange of insulin. Biochim. Biophys. Acta. 18:308.
[3] Bai, Y., J. S. Milne, ..., S. W. Englander, 1993. Primary structure effects on peptide group hydrogen exchange. Proteins. 17:75-86.
[4] Clarke, J., L. S. Itzhaki, and A. R. Fersht, 2006. Hydrogen exchange at equilibrium; a short cut for analysing protein-folding pathways? Trends Biochem. Sci. 22:284-287.
[5] Jones, E., O. Travis, and P. Peterson. 2001. SciPy: open source scientific tools for Python. http://www.scipy.org/.
[6] Fraley, C., and A. E. Raftery. 2002. Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. 97:611-631.
[7] Scrucca, L., M. Fop, ..., A. E. Raftery. 2016. Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J. 8:289-317.