Alma Mater Studiorum · University of Bologna

School of Science
Department of Physics and Astronomy
Master Degree in Physics

Hydrogen Deuterium Exchange: Methods to Probe Protein Dynamics at Single Residue Resolution

Supervisor:
Prof. Gastone Castellani

Co-supervisor:
Prof. Emanuele Paci

Submitted by:
Michele Stofella

Academic Year 2019/2020

Ringraziamenti

Credo profondamente che il cammino sia più importante della meta. Per questo ritengo opportuno dedicare qualche riga per ringraziare le persone che hanno contribuito alla realizzazione di questo elaborato.

Ringrazio i professori Gastone Castellani ed Emanuele Paci, non solo per avermi trasmesso nozioni, ma soprattutto per aver creduto in me, nelle mie idee e nelle mie capacità, e per avermi insegnato a districarmi nel labirintico mondo accademico.

D'altro canto, non avrei mai raggiunto questo obiettivo senza il sostegno di Alberto, Cristina, Simone e Lorenzo, senza dimenticare il supporto dei miei compagni di corso, di tutti i miei amici di Lugo e dei miei fratelli nella fede. Grazie per avermi spinto ad andare oltre le mie preoccupazioni e per avermi sostenuto nei momenti più difficili.

Infine, ringrazio la mia famiglia, la costante che mi ha permesso di arrivare fino a questo punto senza perdere il senno. Ringrazio mia mamma Daniela e mio papà Sandro per avermi dato l'opportunità di studiare nella consapevolezza che lo studio non è la vita. Ringrazio i miei fratelli Lorenzo, Chiara ed Anna per avermi supportato nelle scelte e sopportato nella quotidianità. Ringrazio i miei nonni Enzo e Barbara per tutte le volte in cui mi hanno chiesto di spiegare nella semplicità che cosa riguardasse questa tesi. Ringrazio i miei nonni Ezio e Luciana che dal Cielo mi hanno accompagnato, trasmettendomi una profonda curiosità per il mondo che mi circonda. Un sincero grazie anche ai miei zii Francesco, Cecilia e Carla.

La storia non si scrive con i se e con i ma, tuttavia credo che ci sia un'alta probabilità che senza di voi questo elaborato non sarebbe esistito. Per questo vi ringrazio.

Abstract

The purpose of this work is to provide computational methods to fingerprint protein dynamics probed by hydrogen deuterium exchange mass spectroscopy. Hydrogen deuterium exchange consists in the spontaneous exchange of amide hydrogens of amino acids with deuterium contained in solution. The exchange rate (or protection factor) provides a parameter probing protein dynamics at single residue resolution.

In Chapter 1, hydrogen deuterium exchange is introduced as a high throughput experiment in the biophysical context of protein dynamics and the main statistical issues regarding data analysis are reported. Chapter 2 describes the theoretical background of hydrogen deuterium exchange with a focus on the approximations of the model.

The experimental workflow of hydrogen deuterium exchange mass spectrometry is described in Chapter 3 and the state of the art of data analysis in the field is discussed.

The core of the work is represented by the ExPfact algorithm implemented in Chapter 4. Statistical issues are deepened via its application to synthetic data. Chapter 5 focuses on the application of ExPfact to real world data. A first application validates the algorithm via a comparison of extracted protection factors with rates calculated by NMR experiments. A second application shows how structural changes between different states of the same protein can be detected at amino acidic resolution.

Fine-grained information extracted with ExPfact can be coupled with a back exchange correction to reproduce experimental spectra. This correction is discussed in Chapter 6 together with the development of a structural model connecting the structure of a protein to its protection factors.

Achievements and further developments are highlighted in Chapter 7.

Contents

Chapter 1

Introduction

Summary of the chapter. Hydrogen deuterium exchange is introduced in the biophysical context of protein dynamics and the main statistical issues linked to data analysis are listed.

1.1 Proteins

1.1.1 Nomenclature

Proteins are an essential component of life (Finkelstein and Ptitsyn (2002); Ingalls (2013)). They are polymers formed by amino acids linked into a peptide chain. The portion of the free amino acid that remains after polymerization is called residue. There exist 20 different amino acids that can be referred with several types of nomenclature: here they are addressed with their one letter code (Fig. 1.1). Each amino acid has a different functional group attached that is called side chain. The sequence of residues forming the protein is referred as its primary structure.

The physical process through which the sequence of amino acids folds into a 3D structure is known as protein folding. Some motifs can be found in the structure of proteins: $\alpha-$helices, $\beta-$sheets, loops. These are referred as secondary structure of the protein. The packing of secondary structures into a globule is called tertiary structure while the integration of several chains forms the quaternary structure of a protein.

Asp	D	aspartic acid		Ile	I	isoleucine
Thr	T	*threonine*		Leu	L	leucine
Ser	S	serine		Tyr	Y	tyrosine
Glu	E	glutamic acid		Phe	F	phenylalanine
Pro	P	proline		His	H	histidine
Gly	G	glycine		Lys	K	lysine
Ala	A	alanine		Arg	R	arginine
Cys	C	cysteine		Trp	W	tryptophan
Val	V	valine		Gln	Q	glutamine
Met	M	methionine		Asn	N	asparagine

Figure 1.1: Amino acid nomenclature: three letter code, one letter code and full name of every amino acid.

The function of a protein derives from its structure. Being maintained by hydrogen bonds, the 3D configuration of a protein is flexible, a crucial aspect of the functioning of some proteins since it allows the protein to adopt multiple conformations.

1.1.2 Protein dynamics

According to the Anfinsen's dogma (Anfinsen (1973)), the native structure of a protein is completely determined by its primary structure. This means that the native state is the unique, stable and kinetically accessible minimum of the free energy landscape in which the protein lives.

The research question of protein folding can be summarised in finding the pathway through which the protein reaches such a minimum, one of the most challenging and fascinating topics in biophysics. To give a taste of it, we introduce the Levinthal's paradox (Robert et al. (1992)). The number of possible conformations that a protein can adopt is astronomically large: the age of the universe would not be sufficient for a 100-residue protein to explore all of them and choose the most appropriate one. Nevertheless, small proteins fold spontaneously in a time scale ranging from microsecond to millisecond.

Moreover, the energy landscape can be altered by a number of external factors (e.g. a

change of temperature, pH, the introduction of an external force, the bonding to other molecules). The alteration of the energy landscape may lead to the identification of a new minimum, allowing the conformational change of the protein. An important example is given by allosteric regulation: a protein's function is activated or inhibited through the binding to a specific molecule (Fig. 1.2). The choice of a reaction coordinate, i.e. a unique quantity to describe the state in which a protein lies, is not trivial (Krivov (2013)). Drawing a free energy landscape is thus a complicated issue.

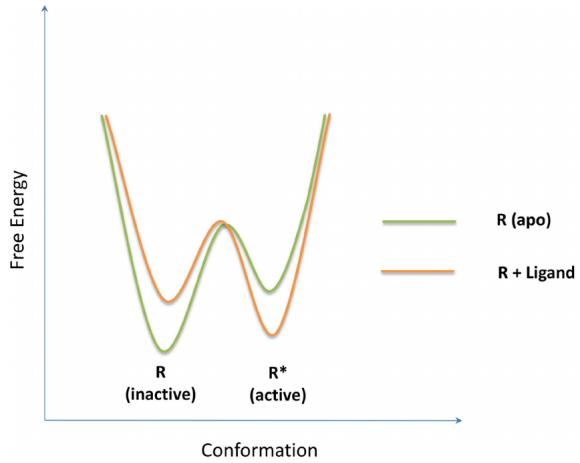

Figure 1.2: Sketch of the alteration of the free energy landscape in allosteric regulation. Different conformational minima are associated to the inactive (apo) and the active state. Figure from Chung-Jung and Ruth (2014).

In the light of this, how can we probe structural and dynamical properties of proteins? Exploiting the phenomenon of hydrogen deuterium exchange, we aim to fingerprint protein dynamics at single residue resolution.

1.2 Hydrogen deuterium exchange

Hydrogen exchange is a unique chemical reaction occurring for certain hydrogens in proteins which are in continuous exchange with hydrogens in solution. This equilibrium reaction takes place even when hydrogens in solution are replaced with heavier isotopes. If a protein is diluted in a solution containing D_2O, its amide hydrogens spontaneously exchange with deuterium contained in solution, consequently increasing the weight of

the protein. This phenomenon is called hydrogen deuterium exchange (HDX) and the pioneering papers exploiting it date back to the middle of the last century (Liotta and Mer (1937); Kwart et al. (1954); Linderstrøm-Lang (1955)). Since then, HDX has been widely used to study protein folding (Clarke and Fersht (1996); Vendruscolo et al. (2003); Krishna et al. (2004)).

1.2.1 HDX-MS

HDX-MS relies on the difference in mass between the protonated and the deuterated poly-peptide chains: as exchange occurs, the increase in mass of the protein can be monitored through mass spectrometry. To obtain more specific information, the reaction is quenched: the protein is fragmented in small peptides (10-30 residues) by proteolysis under conditions that drastically slow down the exchange, namely low pH ($\approx 2-3$) and low temperature ($\approx 0-3^oC$) (Kan et al. (2011); Lam et al. (2002)). This allows the measurement of deuterium uptake of peptides of variable length.

Only the exchange of amide hydrogens shall be considered. As a matter of fact, HDX takes advantage of three different types of hydrogens in proteins: those in amide functional groups, in carbon-hydrogen bonds and in side-chain groups. However, carbon-bond hydrogens have exchange rates so slow that cannot be detected and side-chain hydrogens exchange so fast that they back-exchange rapidly when the reaction is quenched in H_2O-based solution, and the exchange is not registered (Englander et al. (1996)). As a consequence, the exchange rate provides a measure of protein's accessibility to the solvent at single residue resolution. The only exception occurs for prolines where the exchange cannot occur because they lack of the amide hydrogen when in a peptide bond.

HDX was firstly probed through NMR spectroscopy (Dempsey (2001)), but mass spectrometry (MS) has been established as a remarkable alternative (Zhang and Smith (1993); Englander et al. (2003); Masson et al. (2019)). With HDX-NMR, exchange rates can be measured for each residue of the protein. On the other hand, HDX-MS leads in terms of automation of the workflow, costs, dimensions of the protein under analysis and concentrations needed to perform the experiment (only picomoles of protein).

HDX-MS counts various and relevant applications both in academy and industry: to study conformations of individual proteins or large complexes (Harrison and Engen

(2016)), to locate protein sites involved in binding (Chalmers et al. (2011)), to probe allosteric effects (Englander et al. (2003)), intrinsic disorder (Balasubramaniam and Komives (2013)) or to map and characterize biotherapeutics (Deng et al. (2016)). The versatility of the technique together with its high automation and data production rate makes HDX-MS a remarkable exponent of the family of high throughput experiments.

1.2.2 High throughput experiments

One of the most important milestones of high throughput experimentation can be put in 2011 (Liu et al. (2019)) when the USA started the funding of the the Materials Genome Initiative (MGI) (Holdren (2011)), a huge project whose aim was to speed up the discovery of advanced materials and to shorten the time taken to bring them to the market. High throughput experimentation is the set of new experiments, new computational tools and new data needed to promote the efficient development of new materials under the idea of the MGI project.

After about a decade from launching the project, every field in material sciences has been influenced (Green et al. (2017)), from electronics and artificial intelligence to environment and architecture. The pharmaceutical industry, for instance, is currently facing such a great revolution that today every company has a dedicated HTE (high throughput experimentation) group (Mennen et al. (2019)). The high automation involved in high-throughput experimentation promotes the connection of material sciences to apparently unrelated fields like robotics or statistics (Carson (2020)), giving a new genuinely interdisciplinary essence to the world of science.

Concerning the academic research in biology and chemistry, the MGI project brought a wave of enthusiasm with the leading idea of obtaining *rapid results from complex mixtures* (Kempa et al. (2019)). One of the most prominent techniques exploiting this idea is mass spectroscopy (MS): native MS, ion-mobility spectrometry, chemical cross-linking, LS-MS/MS, HDX-MS and FPOP (Johnson et al. (2019)) are just examples of novel MS based techniques.

1.2.3 Other techniques: LiP

Before focusing on HDX-MS for the rest of the manuscript, it is important to stress that any high throughput experiment is able to fingerprint dynamical or structural properties of proteins. Therefore, the best way to capture as much information as possible from a biological system is the coupling of several complementary techniques.

For instance, limited proteolysis (LiP) is a MS-based technique (de Souza and Picotti (2020)) that enables the unbiased and proteome-wide profiling of protein conformational changes. Such changes can be the result of different factors (heat shock, protein-protein interactions, compound binding, ...) and they affect the kinetics of the proteolytic cleavage. A protein is diluted in solution together with the unspecific proteinase K that cleaves it. Smaller peptides are identified in solution at different times and the cleavage kinetics is monitored via mass spectrometry. The data available are the sequence of a cut peptide together with a value proportional to the number of times the specific fragment is found in solution at a certain time.

The underlying phenomenon is different from hydrogen deuterium exchange, but several similarities arise between HDX-MS and LiP:

1. Both techniques provide insights of protein structure looking at its dynamics, overcoming the static pictures provided by NMR or X-ray approaches.
2. Both experiments are (generally) performed in triplicates and only average quantities are analysed (see 1.3.2).
3. Both techniques provide information from different overlapping peptides: in the case of HDX-MS each piece of information contains the time evolution of the deuterium uptake of a specific fragment, in LiP the time evolution of the probability to find a certain fragment.

Because of the nature of high throughput experiments, they provide partial information that can be integrated with data coming from different sources (e.g. HDX-MS and LiP). Furthermore, the similarities between datasets of different experiments leads to challenges in data analysis that can perhaps be solved with similar tools. Development

in HDX-MS analysis could be thus helpful not only for the experiment itself, but for a wide range of other applications.

1.3 Statistical issues

Experimental data associated to HDX-MS are coarse-grained data: they encode the information of large subcomponents of the system, i.e. the uptake of proteolytic fragments measured at different times. The statistical problem here addressed is to extract fine-grained information (regarding single residues) out of coarse data.

1.3.1 Underdetermination

As described in Chapter 2, the statistical model that we can be associate to experimental data y_i at times t_i (with $i = 1, ..., m$) can be written as

$$y_i = \sum_{j=1}^{n} a_{ij} \left(1 - e^{-k_j t_i / P_j}\right) + \epsilon_i \tag{1.1}$$

where a_{ij} and k_j are known constants and the goal is to infere the set of parameters $\{P_j\}$. The experimental errors ϵ_i are assumed to be independent and identically distributed variables with mean 0.

If the error is not present ($\epsilon_i = 0, \forall i$), a unique solution can be calculated if the number of experimental data is greater than the number of parameters to be estimated: $m \geq n$. If instead $m < n$, a unique solution does not exist: there exists an affine space of solutions of dimensions $l = n - m$.

The presence of experimental error worsen the situation because the problem is no longer linear. If $m > n$, the problem is overdetermined and there is no set of $\{P_j\}$ that exactly fits the measurements. On the other hand, if $m < n$, the problem is underdetermined and there exists a plurality of sets exactly reproducing experimental data.

HDX-MS experimental data are characterised by underdetermination.

1.3.2 Replicates

HDX-MS measures the uptake of peptides at several time points, computed as an average value over technical and biological replicates. Technical replicates test the same sample multiple times: the experiment is repeated under the same conditions. Biological replicates probe different samples that are expected to have a similar behaviour within the same experiment. In HDX-MS, the experiment generally consists of three technical replicates (triplicates). The number of biological replicates depends on the sample: mass spectra of peptides can be detected with different charge states.

Despite being widely used, this pre-processing is not statistically robust and gives rise to an open debate. What is the amount of information that we loose by considering only these averages? Are three technical replicates sufficient to provide a proper statistics?

Chapter 2

Theoretical background

Summary of the chapter. The modeling of hydrogen deuterium exchange is introduced, focusing on the approximations of the model, namely the EX1 and EX2 regimes, and on the calculation of intrinsic exchange rates.

2.1 Modeling HDX

When a protein is diluted in solvent containing D_2O, its amide hydrogens spontaneously exchange with deuterium contained in solution. Backbone groups of proteins may be highly protected against this exchange because they are situated within stable elements of secondary structure (like β-sheets) or because they are buried in the protein. The exchange of such groups occurs if small-scale local fluctuations (or local folding) enable the exposure to the solvent. This is why hydrogen deuterium exchange can be modeled as a two step process (Fersht (2017); Hvidt and Nielsen (1966)).

If the amide hydrogen of an amino acid is not exposed to the solvent, it is in a closed state C and the exchange cannot occur. On the other hand, if the amide hydrogen is exposed, the amino acid is in an opened state O and deuteration may occur. If the exchange occurs, the amino acid is in a deuterated state D. Supposing that back exchange cannot

occur, i.e. the residue cannot leave the deuterated state, the model reads

$$C \underset{k_c}{\overset{k_o}{\rightleftharpoons}} O \xrightarrow{k_i} D \tag{2.1}$$

where the transition between these states occurs with opening rate k_o, closing rate k_c and intrinsic exchange rate k_i. The intrinsic rate k_i depends on the sequence of the protein analysed, on the pH and temperature of the solution (see 2.1.3).

The model in equation 2.1 is nothing but a Michaelis Menten model and corresponds to the system of equations

$$\begin{aligned} \dot{C} &= -k_o C + k_c O \\ \dot{O} &= k_o C - (k_c + k_i) O \\ \dot{D} &= k_i O \end{aligned} \tag{2.2}$$

that can be written in matrix form as

$$\dot{\mathbf{X}} = K\mathbf{X} \tag{2.3}$$

where $\mathbf{X} = (C, O, D)$ and

$$K = \begin{pmatrix} -k_o & k_c & 0 \\ k_o & -(k_c + k_i) & 0 \\ 0 & k_i & 0 \end{pmatrix} \tag{2.4}$$

The solution of the equation is

$$\mathbf{X} = \sum_{i=1}^{3} \alpha_i e^{\lambda_i t} \mathbf{u_i} \tag{2.5}$$

where λ_i are the eigenvalues

$$\begin{aligned} \lambda_1 &= 0 \\ \lambda_2 &= -\tfrac{1}{2}\left((k_c + k_i + k_o) + \sqrt{(k_c + k_i + k_o)^2 - 4k_i k_o}\right) \\ \lambda_3 &= -\tfrac{1}{2}\left((k_c + k_i + k_o) - \sqrt{(k_c + k_i + k_o)^2 - 4k_i k_o}\right) \end{aligned} \tag{2.6}$$

and eigenvectors $(\mathbf{u_1}, \mathbf{u_2}, \mathbf{u_3})$ are

$$\begin{aligned}
\mathbf{u_1} &= (0, 0, 1) \\
\mathbf{u_2} &= (-\frac{-k_c+k_i-k_o-\sqrt{(k_c+k_i+k_o)^2-4k_ik_o}}{2k_i}, -\frac{k_c+k_i+k_o+\sqrt{(k_c+k_i+k_o)^2-4k_ik_o}}{2k_i}, 1) \\
\mathbf{u_3} &= (-\frac{-k_c+k_i-k_o+\sqrt{(k_c+k_i+k_o)^2-4k_ik_o}}{2k_i}, -\frac{k_c+k_i+k_o-\sqrt{(k_c+k_i+k_o)^2-4k_ik_o}}{2k_i}, 1)
\end{aligned} \tag{2.7}$$

and the constants α_i are set by the conditions $O + C + D = 1$, $D(0) = 0$ and $D(\infty) = 1$. It is important to stress that no assumption is made over the initial values of the open and closed states (O and C).

The exact solution for $D(t)$ is

$$\begin{aligned}
D(t) = 1 \quad &+ \quad \frac{k_o + k_i + k_c - \sqrt{(k_o + k_i + k_c)^2 - 4k_ik_o} - 2k_ik_o/(k_i + k_o)}{2\sqrt{(k_o + k_i + k_c)^2 - 4k_ik_o}} e^{\lambda_2 t} \\
&- \quad \frac{k_o + k_i + k_c + \sqrt{(k_o + k_i + k_c)^2 - 4k_ik_o} - 2k_ik_o/(k_i + k_o)}{2\sqrt{(k_o + k_i + k_c)^2 - 4k_ik_o}} e^{\lambda_3 t}
\end{aligned} \tag{2.8}$$

2.1.1 EX1 and EX2 limits

Native state $\mathbf{k_c \gg k_o}$

Following the Anfinsen's dogma, the protein in the native state is folded. We can thus assume that it is more probable for a residue of a native protein to be in the closed state C rather than in the opened state O: $k_c \gg k_o$. This assumption is known as native state approximation.

Wagner and Wüthrich (1979) state that under native conditions the solution is a single exponential $D(t) = 1 - \exp(-k_x t)$ with exchange rate

$$k_x = \frac{k_i k_o}{k_c + k_i + k_o} \tag{2.9}$$

Eq. 2.9 derives from the Taylor expansion for $k_c \gg k_o$. In fact, defining

$$\Delta \equiv \sqrt{(k_c + k_i + k_o)^2 - 4k_ik_o} \tag{2.10}$$

and using the Taylor expansion $\sqrt{1-x} \approx 1 - x/2$ (holding when $x \to 0$), we obtain:

$$\begin{aligned} \Delta &= (k_c + k_i + k_o)\sqrt{1 - \tfrac{4k_i k_o}{(k_c+k_i+k_o)^2}} \\ &\approx (k_c + k_i + k_o)(1 - \tfrac{2k_i k_o}{(k_i+k_c+k_o)^2}) \end{aligned} \tag{2.11}$$

Inserting the approximation (Eq. 2.11) in the eigenvalue λ_3 we get Eq. 2.9.

EX1 regime $\mathbf{k_i \gg k_c}$

From the rate in Eq. 2.9, two limits can be identified relating the intrinsic exchange rate with the closing rate. The EX1 regime is approached for $k_i \gg k_c$. The meaning of this approximation is that as soon as the residue gets exposed, its amide hydrogen exchanges with deuterium. In this situation, the exchange rate reduces to

$$k_x = k_o \tag{2.12}$$

This regime is experimentally fingerprinted by a bimodal pattern of isotopic distribution in mass spectra (Adhikary et al. (2017); Ferraro et al. (2004); Zhou et al. (2017)). Such a condition can be reached by increasing the temperature, adding *subdenaturant* concentrations of denaturant (like urea) or shifting the pH towards alkaline values (Lapidus (2017); Malhotra et al. (2017)).

The single exponential approximation and the EX1 regime are incompatible (see 2.1.2).

EX2 regime $\mathbf{k_c \gg k_i}$

On the other hand, the EX2 limit occurs when $k_c \gg k_i$: the residue is expected to fluctuate between the opened and closed state more probably than to acquire a deuterium. In this case, the exchange rate can be written as

$$k_x = \frac{k_i}{P} \tag{2.13}$$

where we introduced the protection factor $P = k_c/k_o$. In the EX2 regime, the kinetics is sensitive to pH only through intrinsic rates k_i and the corresponding isotopic envelope

(or mass spectrum) evolves progressively towards the fully deuterated limit.

The protection factor measures the degree of protection against the exchange. Dealing with native conditions $k_i \gg k_c$, from the definition of P it follows that

$$P \equiv \frac{k_c}{k_o} \gg 1 \tag{2.14}$$

The parameter can thus be interpreted as a measure of the *nativeness* of the residue.

The EX2 regime is generally approached in experiments performed at room temperature and pH>3 and thus the deuterium uptake for the single residue can be written as

$$D(t) = 1 - e^{-\frac{k_i}{P}t} \tag{2.15}$$

As a matter of fact, HDX-MS measures the change in mass due to the deuteration of proteolytic fragments of the protein. The deuterium uptake D_j for a peptide j starting at residue m_j of the sequence of the protein and n_j residue long can be written at time t_k as the sum of the uptakes of its residues:

$$D_j(t_k, \{P_i\}) = \frac{1}{n_j} \sum_{i=m_j+1}^{m_j+n_j-1} 1 - e^{-\frac{k_i}{P_i}t_k} \tag{2.16}$$

where P_i is the protection factor of residue i and k_i is the intrinsic exchange rate of residue i. The sum starts from the second residue because the first one forms the free N-terminus of the peptide (Walters et al. (2012)).

2.1.2 About the single exponential approximation

Let us write the general solution (Eq. 2.8) as

$$D(t) = 1 + \alpha e^{\lambda_2 t} + \beta e^{\lambda_3 t} \tag{2.17}$$

with

$$\begin{aligned} \alpha &= \frac{k_o+k_i+k_c-\Delta-2k_ik_o/(k_i+k_o)}{2\Delta} \\ \beta &= \frac{k_o+k_i+k_c+\Delta-2k_ik_o/(k_i+k_o)}{2\Delta} \end{aligned} \tag{2.18}$$

where Δ is defined in Eq. 2.10.

The single exponential approximation (Eq. 2.16) holds only if both native and EX2 approximations are valid. In fact, it holds if in Eq. 2.17 $\alpha \to 0$ and $\beta \to 1$ simultaneously. Using the native approximation ($k_c \gg k_o$) and thus writing Δ as stated in Eq. 2.11, the numerator of α reduces to

$$\text{num}(\alpha) = k_o + k_i + k_c - \Delta - \frac{2k_ik_o}{k_i + k_o} \approx \frac{2k_ik_o}{k_i + k_c + k_o} - \frac{2k_ik_o}{k_i + k_o} \tag{2.19}$$

and we need to use both native and EX2 conditions ($k_c \ll k_i$) in order to get $\alpha \to 0$. In fact, we have that

$$\text{num}(\alpha) \approx \frac{2k_ik_o}{k_i + k_c + k_o} - \frac{2k_ik_o}{k_i + k_o} = \frac{2k_ik_o}{k_i\left(1 + \frac{k_o}{k_i} + \frac{k_c}{k_i)}\right)} - \frac{2k_ik_o}{k_i\left(1 + \frac{k_o}{k_i}\right)} \to 2k_o - 2k_o = 0 \tag{2.20}$$

Analogously, it can be demonstrated that $\beta \to 1$ only if both native and EX2 limits are satisfied.

Furthermore, the non validity of the single exponential assumption under the EX1 regime can explain the bimodality in the experimental mass spectra under such conditions. To proof this statement, we shall remember that no assumption has been made on the initial population of closed and opened states C(0) and O(0): these values are fixed once the rates k_o, k_c and k_{int} are set. In particular, the solution for the opened population can be written from Eq. 2.5 as:

$$O(t, k_i, k_c, k_o) = \alpha_2(k_i, k_c, k_o)e^{\lambda_2 t}u_2^y + \alpha_3(k_i, k_c, k_o)e^{\lambda_3 t}u_3^y \tag{2.21}$$

where u_2^y and u_3^y are the second component of the eigenvectors in Eq. 2.7 and we defined

$$\begin{aligned} \alpha_2(k_i, k_c, k_o) &= \frac{k_o + k_i + k_c - \Delta - \frac{2k_ik_o}{k_i + k_o}}{2\Delta} \\ \alpha_3(k_i, k_c, k_o) &= -\frac{k_o + k_i + k_c + \Delta - \frac{2k_ik_o}{k_i + k_o}}{2\Delta} \end{aligned} \tag{2.22}$$

Δ is defined in Eq. 2.11. As a consequence, the initial population O(t=0) reads:

$$O(t=0, k_i, k_c, k_o) = \alpha_2(k_i, k_c, k_o)u_2^y + \alpha_3(k_i, k_c, k_o)u_3^y \tag{2.23}$$

If native conditions are fixed (e.g. taking $k_o/k_c = 10^{-3}$), the initial population in the opened state depends on EX1 and EX2 conditions as shown in Fig. 2.1: a pure EX1 system is characterised by O(0)=1 while a pure EX2 system by O(0)=0.

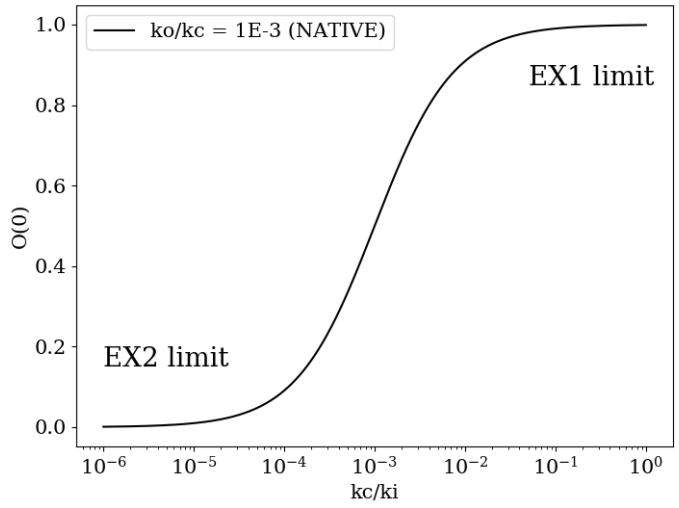

Figure 2.1: Initial population in the opened state O(0) as a function of k_c/k_i evaluated from Eq. 2.23 at fixed native conditions $k_o/k_c = 10^{-3}$. Low values of k_c/k_i represent the EX2 limit, high values the EX1 limit.

As a consequence, a pure EX1 process can be modeled with Eq. 2.1 using O(0)=1 while a pure EX2 process has O(0)=0:

$$\begin{aligned} &\text{Pure EX1} \quad C_{EX1} \rightleftharpoons O_{EX1} \rightarrow D_{EX1} \quad O_{EX1}(0) = 1 \\ &\text{Pure EX2} \quad C_{EX2} \rightleftharpoons O_{EX2} \rightarrow D_{EX2} \quad O_{EX2}(0) = 0 \end{aligned} \tag{2.24}$$

A mixed system can be written with Eq. 2.1 using initial populations $C(0) = C_0$ and $O(0) = O_0 = 1 - C_0$:

$$\text{Mixed} \quad C_{mix} \rightleftharpoons O_{mix} \rightarrow D_{mix} \quad C_{mix}(0) = C_0 \tag{2.25}$$

The uptake of the mixed system can be equivalently written as a weighted combination of two pure systems:

$$D_{mix}(t) = C_0 D_{EX2}(t) + (1 - C_0) D_{EX1}(t) \tag{2.26}$$

Such a decomposition can be used to interpret the bimodal pattern of the isotopic envelope arising under mixed conditions: each component in Eq. 2.26 is responsible for the mode of the spectrum associated to pure EX1 and EX2 conditions.

2.1.3 Intrinsic exchange rates

The intrinsic exchange rate k_i is the exchange rate of an amide hydrogen in fully exposed protein. The dependence of such rates has been widely studied and the values of k_i can be calculated once the sequence of the protein is known (Molday et al. (1972)) together with the temperature and pH of the solution (Bai et al. (1993)). Such relations have been extrapolated from experimental studies probing homo-dimers, homo-oligomers and homo-polypeptides for all the 20 amino acids.

Intrinsic exchange rates have a high dependence on the pH of the solution since exchange is mainly catalysed by water ions. Experimental studies performed by Bai et al. (1993) proof that this relation is a V-shaped curve which reads

$$k_i(pD) = k_A 10^{-pD} + k_B 10^{pD - pK_D} + k_W \tag{2.27}$$

where K_D is the dissociation constant of D_2O, k_A, k_B and k_W are the second order rate constants for catalysis by D_3O^+,DO^- and D_2O respectively. Glasoe and Long (1960) and more recently Krezel and Bal (2004) suggest that pD values can be calculated from pH calibrated electrodes by the application of an empirical correction

$$pD = pH + 0.4 \tag{2.28}$$

which was determined by comparing the pH when the same amount of acid or base was dissolved in H_2O and D_2O.

The temperature dependence of intrinsic exchange rates was originally estimated for poly

alanine peptide at temperature 293 K and was used as a reference. The behaviour of k_i with respect to temperature follows the Arrhenius law

$$k_i(T) = k_i(T_0)exp\Big\{ - \frac{E_i^a}{R}\Big(\frac{1}{T} - \frac{1}{T_0}\Big)\Big\} \tag{2.29}$$

where E_i^a is the activation energy of the catalysis by species, $k_i(T_0)$ is the constant rate at T_0 =293 K, R is the universal gas constant and T is the temperature in K.

The intrinsic exchange rate also depends on the neighbouring side chains. Taking into account the effect of two neighbouring sides, it has been demonstrated (Molday et al. (1972)) that the effects of the neighbours can be both positive or negative and are additive, meaning that the right and left side chains affect exchange independently. Again, using poly alanine as reference, the dependence on the neighbouring side chains on the intrinsic rate can be written as

$$k_i(L, R) = k_i(Alanine)\phi_i(L)\rho_i(R) \tag{2.30}$$

where $\phi_i(L)$ and $\rho_i(R)$ are correction factors for the left and right sides.

Intrinsic exchange rates have been calculated both for in-exchange (protonated protein in deuterated solution) and back-exchange (deuterated protein in protonated solution). Back-exchange is neglected in Eq. 2.1 supposing that in a deuterated buffer under physiological conditions in-exchange is much more probable than back exchange. However, back exchange cannot be neglected when the solution is quenched in H_2O (Walters et al. (2012)).

Chapter 3

Experimental workflow and data

Summary of the chapter Experimental data that can be analysed through HDX-MS are discussed, the experimental workflow of the technique is analysed and isotopic envelopes are introduced.

3.1 Experimental data

As described in Chapter 2, the deuterium uptake at time t_k of a n_j-residue long peptide starting at residue m_j can be written as in Eq. 2.16:

$$D_j(t_k, \{P_i\}) = \frac{1}{n_j} \sum_{i=m_j+1}^{m_j+n_j-1} 1 - e^{-\frac{k_i}{P_i} t_k}$$

As a consequence, it should be possible to determine protection factor P_i of each residue of the peptide. However, such an estimation is not trivial: the length of most peptides (in any dataset) exceeds the number of experimental points available. Thus, the number of parameters to be estimated (protection factors) is greater than the number of experimental data: the problem is statistically underdetermined (see 1.3.1). Moreover, even if enough time points were available, it would not be possible to localize the estimated exchange rates within the peptide: the protection factor of a residue could be arbitrarily switched with any other rate.

3.1.1 HDX-MS workflow

The HDX-MS experimental workflow (Masson et al. (2019)) is summarised in Fig 3.1.

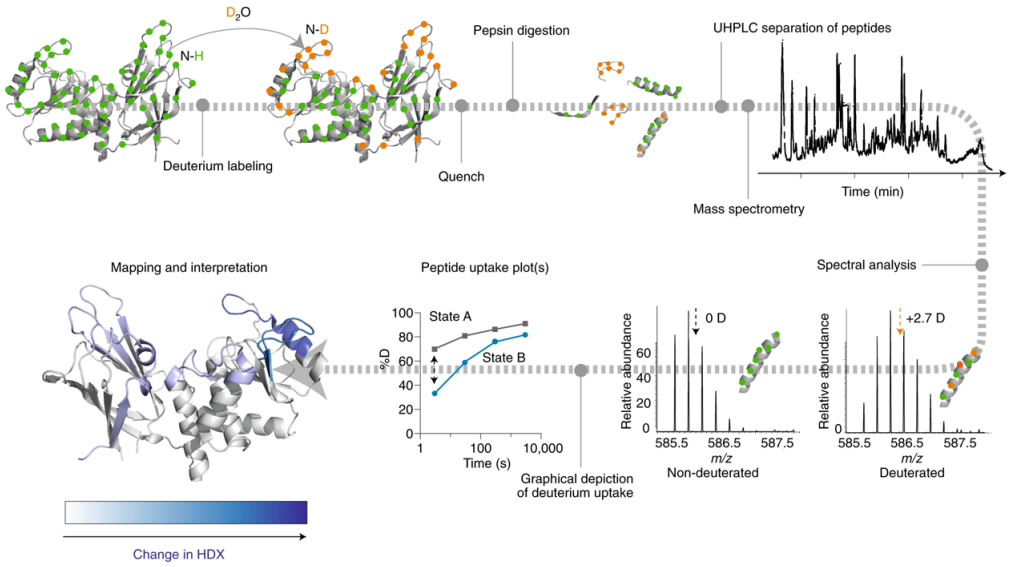

Figure 3.1: HDX-MS workflow (Masson et al. (2019)). Proteins are diluted in a deuterated buffer for several time points, allowing the incorporation of deuterium into the protein backbone. The reaction is then quenched at low pH and low temperature. Proteins are digested by a protease. The proteolytic peptides are desalted and separated using a UHPLC system, ionized by electrospray, inserted in a mass spectrometer and subjected to mass analysis. During spectral analysis, the isotopic envelopes of peptides are visualized and the deuterium uptake is evaluated as the intensity-weighted average mass (arrows) of the peptide. Deuterium uptake is then graphically depicted as a function of time for specific regions of the protein. Differences in deuterium incorporation can be mapped on a 3D representation of the protein.

Proteins are incubated in a deuterated buffer allowing the exchange between amide hydrogen of the protein and deuterium contained in solution. The concentration of D_2O must be precisely maintained during the labeling reaction: HDX-MS experiments can be performed at any concentration, but to speed up the exchange high concentrations $(80 - 90\%)$ are generally used.

To obtain more specific information, the exchange reaction suffers a switch to acidic pH and a temperature drop. Optionally, denaturants can be included to enhance protein

unfolding. These conditions, namely quenching conditions, drastically slow down the exchange, almost stopping it. Proteins are then digested by an acid-functional protease like pepsin.

The proteolytic fragments are desalted and separated using UHPLC: ultra high performance liquid chromatography. Liquid chromatography (Ramos (2013)) is a chemical technique used to separate, identify and quantify each component in a mixture. It relies on pumps to press a liquid solvent containing the mixture through a column filled with an adsorbing material (generally silica or polymers). Each component interacts differently with the material, causing different flow rates and leading to the separation of the components while flowing out of the column. With respect to traditional chromatography, performed at low pressures, high-performance liquid chromatography (HPLC) relies on high pressures (50-350 bar) to speed up the process. In UHPLC pressure is increased at values bigger than 1000 bar.

Finally, the proteolytic fragments are eluted into a mass spectrometer, where they are ionized by electrospray and subjected to mass analysis to determine the increase in mass due to hydrogen deuterium exchange. Electrospray ionization - or ESI (Gross (2017)) - is a technique used in mass spectrometry to ionize molecules and thus to let them be detected by the spectrometer. The application of high voltages to a volatile solvent containing low concentration of ionic analyte ($10^{-6} - 10^{-4}$ M) leads to the transfer of ions from condensed to gas phase. Such phase change starts at atmospheric pressures and increases into the high vacuum of the mass analyser. ESI brings to the formation of multiply charged ions, shifting even heavier peptides into a m/z range accessible to most spectrometers.

3.1.2 Isotopic envelopes

During spectral analysis, the isotopic-related mass spectra of the peptides, also known as isotopic envelopes, are visualized. Isotopic envelopes take into consideration the natural occurrence of isotopic variants which increases the mass of the monoisotopic species of the fully protonated peptide.

If the sequence of a peptide (and thus its elemental composition) is known, the monoisotopic mass of the peptide can be calculated. The fully protonated isotopic envelope of a

peptide can be computed by the convolution of the monoisotopic mass with the natural distribution of isotopes of oxygen, nitrogen and carbon. Such a calculation is performed by many softwares: we used the online freely available MS-Isotope[1].

Because of the importance of isotopic envelopes, it is worth performing an example (adapted by Skinner et al. (2019)). Let us take into consideration a fictitious sequence of amino acids: IDSQVLCGAVKW. Since we know the elemental composition of the peptide, we are able to assert that its monoisotopic mass is 1318.68 Da. Using the MS-Isotope software previously mentioned, we can calculate the fully protonated isotopic envelope of the peptide, shown in Fig 3.2.

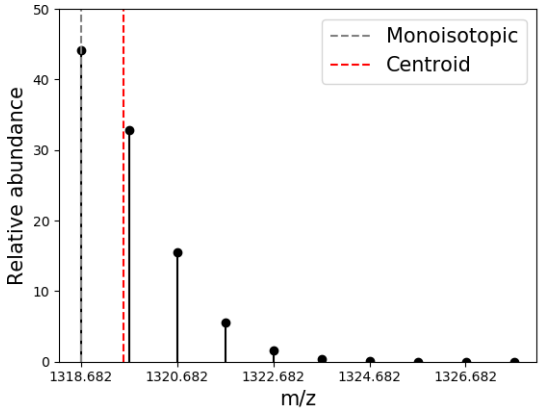

Figure 3.2: Fully protonated isotopic envelope of a peptide with sequence IDSQVLCGAVKW calculated by MS-Isotope. Monoisotopic mass is shown by the left peak (gray dashed line). Other peaks are calculated taking into consideration natural occurrence of isotopic variants. The centroid of the envelope is also shown (red dashed line).

The fully deuterated isotopic envelope, i.e. the spectrum detected when all hydrogens have exchanged into deuterium, is the fully protonated envelope shifted towards higher values of m/z by N-1 units, N being the length of the peptide under analysis (in the example in Fig 3.2, the envelope would be shifted of 11 units).

At intermediate times the envelope may assume different shapes: the intensity of each peak changes with a probability depending on the exchange rates of each residue. For a peptide formed by n exchangeable amides, the probability that k have exchanged

[1] prospector.ucsf.edu/prospector

$(0 \leq k \leq n)$ at time t can be written as

$$\Pi(k,t) = \sum_{A\in\{1,\dots,n\}}^{|A|=k} \prod_{i\in A} D_i(t) \prod_{j\in\{1,\dots,n\}/A} (1 - D_j(t)) \tag{3.1}$$

$D_i(t)$ being the deuterium uptake (Eq. 2.16). Referring to the fully protonated isotopic envelope as π_i, the isotopic envelope of the peptide at time t would be given by $\pi_i\Pi(k,t)$. The fully deuterated isotopic envelope can also be obtained by the application of the time evolution in Eq. 3.1 at infinite time.

Eq. 3.1 introduces an important consequence of the knowledge of protection factors at single residue resolution. In fact, if we were able to evaluate the protection factor of each residue, we could evaluate the deuterium uptake of the peptide at every time using Eq. 2.16 and the isotopic envelope could be predicted at any time.

3.1.3 From envelopes to deuterium uptake

Deuterium incorporation is measured as the centroid of the envelope, i.e. the intensity-weighted mass average. Actually, the uptake associated to an envelope with centroid m at a specific time is normalized using the following formula:

$$D = \frac{m - m_{0\%}}{m_{100\%} - m_{0\%}} \tag{3.2}$$

where $m_{0\%}$ is the centroid of the experimental fully protonated peptide and $m_{100\%}$ is the centroid of the maximally labeled peptide. As a consequence of Eq. 3.2, $0 \leq D \leq 1$.

Because of back exchange occurring under quenching conditions, the maximally labeled envelope $m_{100\%}$ does not coincide with the theoretical fully deuterated envelope. In particular, the former lies at lower values of m/z with respect to the latter (see 6.1).

Repeated measurements of deuterium incorporation are needed in order to ensure the replicability of the experiment and to estimate the precision of the measurements. According to the experimental recommendations provided by Masson et al. (2019), there should be at least three technical replicates of the experiment under the same labeling condition at each time point. The uptake of a peptide is thus given by the mean value of

Eq. 3.2 averaged over the technical and biological replicates with the associated standard deviation. Examples of deuterium uptake curves are shown in Fig 3.3.

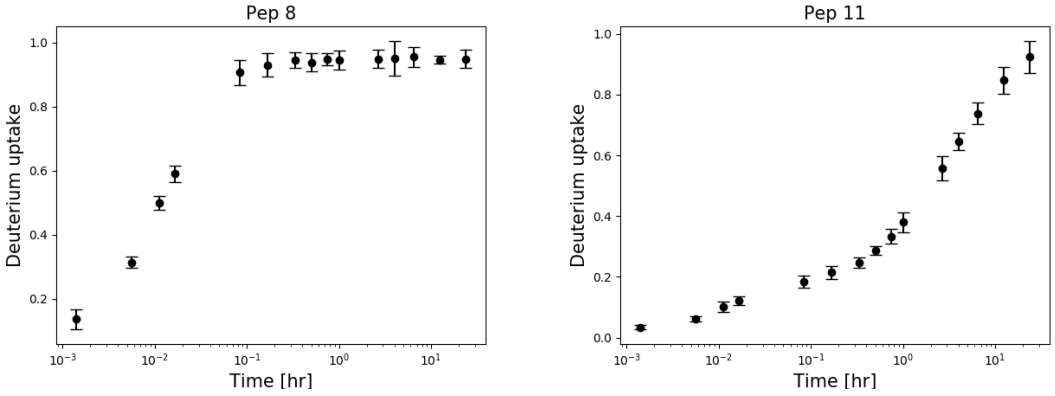

Figure 3.3: Deuterium uptake curves. Experimental deuterium uptake is shown as a function of times for two peptides. Technical replicates: 3. Time points: 15. Data from Moulick et al. (2015).

If the structure of the protein is available, differences in deuterium incorporation due to different states can be mapped on a three-dimensional representation of the protein to facilitate structural interpretation (see Fig. 3.1).

3.2 Estimating protection factors

The information contained in the deuterium uptake of the whole protein prevents us from reaching protection factors at single residue resolution. The time points required to correctly fit Eq. 2.16 for a small protein (50-100 residues) would be acquired in such a long time that the idea of high throughput experimentation would be completely lost. The problem is underdetermined (see 1.3.1). Moreover, even if enough time points were acquired and a thus unique set of protection factors were estimated, we would not be able to associate a value P_i to a specific residue i. We address the latter issue as the problem of switchable residues.

Considering the information encoded in proteolytic peptides does not completely solve these issues. In fact, most datasets do not exceed 15 time points and fragments are

generally 10-30 residues long. As a consequence, underdetermination is not solved.

However, the problem of switchable residues can be faced if the proteolytic fragments show an overlapping pattern (see the peptide map in Fig 3.4). In an ideal case, one peptide differs from the adjacent ones of one only residue. From the subtraction of the information encoded in adjacent peptides, it is possible to extract the uptake of the single residue and protection factors can be thus extracted at single residue resolution. However, such situation can be reached experimentally only for few isolated amino acids.

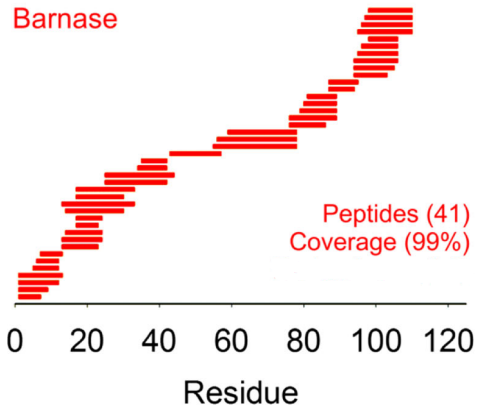

Figure 3.4: Example of peptide map for a dataset by Harris et al. (2019). Each segment encodes a deuterium uptake curve like the ones shown in Fig 3.3; the length of the segment corresponds to the length of the peptide.

According to Masson et al. (2019), the quality of a peptide map should be evaluated in terms of coverage and redundancy. Coverage is calculated as the number of residues analysed divided by the total amount of amides forming a protein and expressed as percentage. Redundancy is evaluated as the number of identified peptides divided by total number of amides.

3.3 State of the art

HDX-MS analysis is often limited to qualitative results (Lisal et al. (2005); Lísal et al. (2006); Kan et al. (2011)): the apparent average rate of exchange of different peptides

is mapped on the structure of the protein and the kinetics of the same fragment under different conditions is compared. One example of such maps is shown in Fig 3.1. In order to estimate the rate, single exponential or single stretched exponential curves are used, returning a value for each peptide that has no physical meaning.

3.3.1 HDSite: an envelope-based approach

Directly using the information contained in the isotopic envelopes, the Englander laboratory pioneered a different approach (Kan et al. (2013)). More recently, a similar method has been implemented to provide single residue resolution exchange rates of equine cytochrome c (Hamuro (2017)).

The algorithm developed by the Englander Lab (Kan et al. (2013)), called HDSite, relies on fitting experimental envelopes in order to determine the uptake level of each amino acid (D-occupancy). Its workflow is summarised in Fig 3.5.

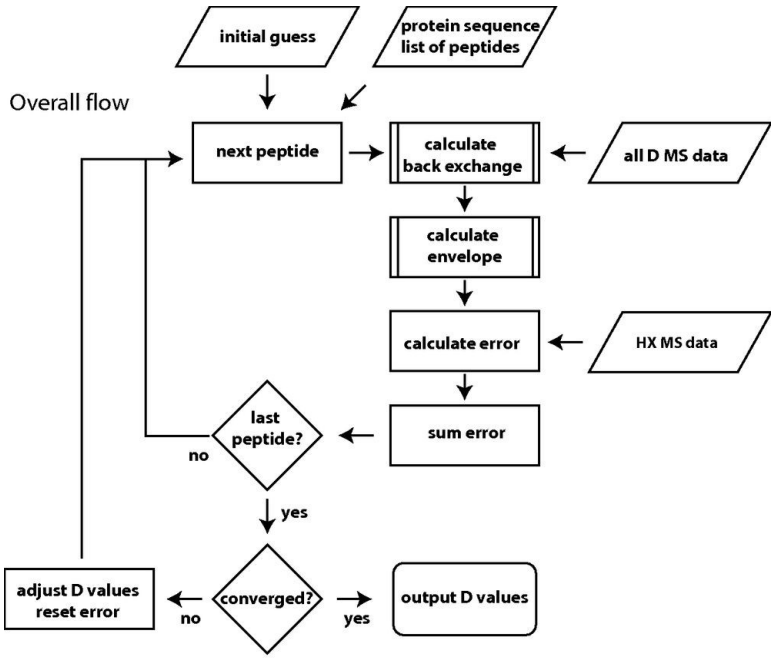

Figure 3.5: Flow of the HDSite method, from Kan et al. (2013)

After setting an initial set of D-occupancies, the isotopic envelopes are evaluated using

a standard binomial that is further convoluted with the natural abundance distribution of elements (Fig 3.6). The predicted envelopes are compared with the experimental ones and an error is calculated through a weighted sum of squared deviations. D-occupancy values are then adjusted and the cycle is repeated until the error function decreases no more or a maximum number of iterations is reached.

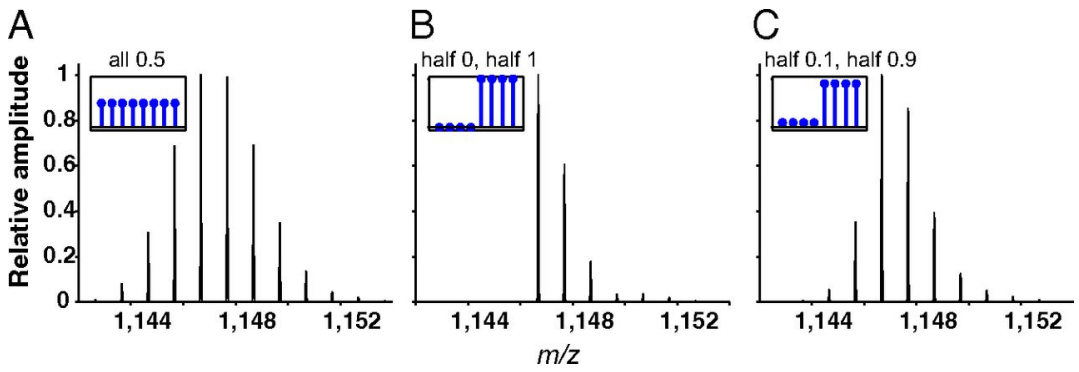

Figure 3.6: Simulated 9-residue peptide spectra with different D-occupancies. Insets specify the D-occupancies used. A) All residue are 50% deuterated; B) Four residues are protonated, the others are deuterated; C) Four residues are 10% deuterated, the others are 90% deuterated. In the HDSite workflow, these simulated spectra are compared with experimental ones and D-occupancies are adjusted in order to minimize an error function. Figure from Kan et al. (2013).

Once the deuterium uptake is known at a single residue resolution at different times, the protection factor of the specific amino acid can be evaluated using the single exponential in Eq. 2.15.

HDSite suffers the problem of switchable residues. In fact, the deuterium uptake curve of each residue cannot be uniquely determined: looking at Fig 3.6 B we could obtain the same spectrum considering any 4 deuterated residues (instead of considering the last 4, we could consider the even ones or the first 4 amino acids). As a consequence, protection factors are uniquely extracted for a small minority of amino acids. Only if an ideal dataset were available, the exchange rate of every amino acid could be evaluated exactly. Adjacent switchable residues are treated as a whole: estimated D-occupancy uptake of each switchable residue is considered and fitted with a single exponential. The evaluated protection factors are averaged and associated to every switchable residue composing the fragment.

HDSite also takes into consideration a back exchange correction on a per amino acid basis, exploiting the difference in deuteration between the experimental and theoretical envelopes (Fig 3.7) due to back exchange occurring under quenching conditions.

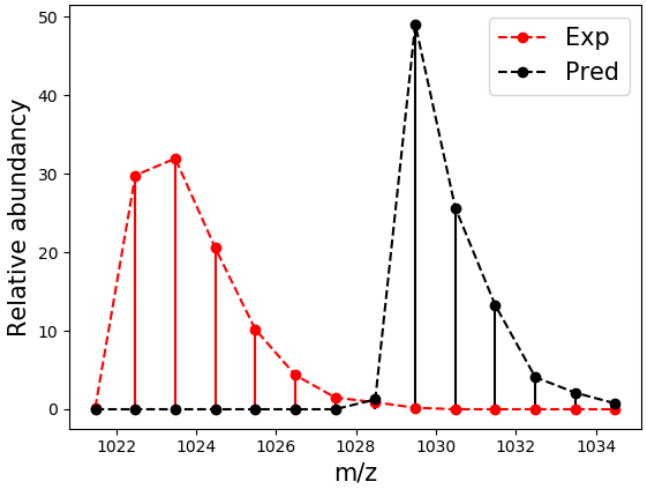

Figure 3.7: Experimental (red) and theoretical (black) fully deuterated isotopic envelopes are compared for a peptide. Data from Moulick et al. (2015).

The difference in terms of centroids of the experimental and theoretical envelopes is evaluated and fitted using a non linear least squares method to the function

$$\text{Back}(t, \{k_{b,i}\}) = \frac{1}{N-1} \sum_{i=2}^{N} 1 - pe^{-k_{b,i}t} \tag{3.3}$$

where an effective back exchange time τ is evaluated from the knowledge of the intrinsic back exchange rates $k_{b,i}$ (tabulated by Bai et al. (1993)) and the percentage of deuterium in solution p, N being the length of the peptide.

Eq. 3.3 assumes that at a single residue level back exchange is not modeled as in-exchange (Eq. 2.1) but as a two state irreversible model,

$$D \xrightarrow{k_b} P \tag{3.4}$$

D and P being the deuterated and protonated states respectively. The solution of such

a model is straightforwardly a single exponential:

$$\begin{aligned} \text{Deuterated} \quad & D(t) = D_0 e^{-k_b t} \\ \text{Back exchanged} \quad & P(t) = 1 - D(t) = 1 - D_0 e^{-k_b t} \end{aligned} \tag{3.5}$$

D_0 being the percentage of deuterium in solution (p in Eq. 3.3). This hypothesis suggests that protection factors play no role in back exchange and that in exchange is completely stopped under quench conditions.

The correction must be estimated for every fragment since different peptides provide different effective back exchange times. In fact, not only does back exchange depend on experimental conditions like pH, temperature and percentage of deuterium, but also on the ionic strength of the peptide involved (Walters et al. (2012)). Finally, the effective back exchange time τ is used to calculate back exchange at single residue resolution using Eq. 3.5.

Despite providing a nice correlation between exchange rates extracted with respect to NMR measurements, the method developed by the Englander laboratory provides single residue resolved protection factors only for a small subset of amino acids. Moreover, most datasets available are in the form of centroids value and show peptide maps far from being optimal, limiting the usage of such method to privileged cases.

Chapter 4

ExPfact algorithm

Summary of the chapter. The ExPfact algorithm aiming to reach single residue resolution is described. The application of the algorithm to synthetic data reveals insights of the advantages and drawbacks of the algorithm. Moreover, ExPfact is compared with HDSite through the generation of a dataset with an ideal overlapping of fragments.

4.1 Workflow

The ExPfact (EXtract Protection FACTors) algorithm was firstly implemented by Skinner et al. (2019) in order to reach single residue resolution of protection factors extracted from HDX-MS data. It relies on the information encoded in the centroids of isotopic envelopes and fits experimental data with Eq. 2.16 in order to get as a result the set of $\{P_i\}$ describing a peptide:

$$D_j(t_k, \{P_i\}) = \frac{1}{n_j} \sum_{i=m_j+1}^{m_j+n_j-1} 1 - e^{-\frac{k_i}{P_i} t_k}$$

As discussed in the previous chapters, the problem suffers underdetermination: the number of time points available is lower than the number of parameters to be estimated. Consequently, a unique solution does not exist.

The idea of the ExPfact algorithm is that despite the degeneracy of sets of $\{P_i\}$ in agree-

ment with experimental data, a finite number of clusters of solutions may be identified, thus reducing the degeneracy of the problem. The workflow of the ExPfact algorithm is summarized in Fig 4.1 and can be divided into three components: random search, least squares minimization and clustering.

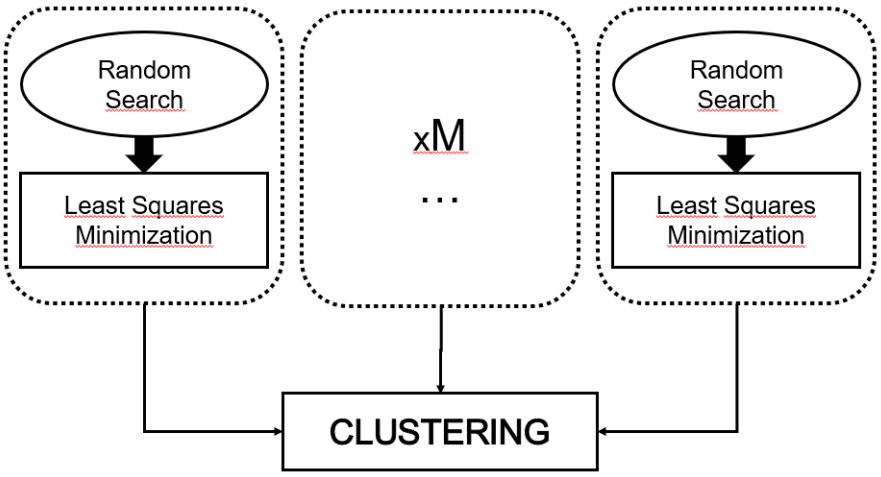

Figure 4.1: ExPfact workflow. Random Search: N random sets of protection factors are randomly initialized and the one with best agreement with experimental data is selected. Least Squares Minimization: the previously selected set is used as initial guess for a least squares minimization. Clustering: the process is repeated M times and the results are clustered together using a model-based clustering algorithm.

Random Search

The first part of the algorithm consists in the random search of a set of protection factors. In particular, N sets of protection factors are randomly initialised with the constraint

$$0 \leq \ln(P_i) \leq 20$$

This boundaries mean that the exchange rate of an amide can be as fast as in a completely unstructured peptide (if $\ln P = 0$, the exchange rate k_x is equal to the intrinsic exchange rate k_i) or up to 5×10^8 times slower ($\ln P = 20$). These values can also be considered as empirical limits that can be deduced by HDX-NMR studies.

For each of these random sets, the value of the cost function is evaluated:

$$C(\{P_i\}) = \sum_j \sum_k w_{jk} \left[D_j^{pred}(t_k, \{P_i\}) - D_j^{exp}(t_k, \{P_i\}) \right]^2 \tag{4.1}$$

where the sum is performed at each time point available t_k and for each peptide j. The predicted deuterium uptake D_j^{pred} is evaluated using Eq. 2.16 and an appropriate choice for the weights w_{jk} is the inverse of the standard deviation of experimental measurements.

Least Squares Minimization

The set of protection factors with the lowest cost function is selected from the random search and used as initial guess for a least squares minimization which adjusts the coefficients $\{P_i\}$ in order to reduce the value of the cost function itself (Eq. 4.1).

Since a correlation between the protection factors and the structure of a protein has been found (Best and Vendruscolo (2006)), ExPfact enables the introduction of a penalization term in order to avoid abrupt changes of protection factors of adjacent residues:

$$\mathrm{Pen}(\lambda, \{P_i\}) = \lambda \sum_{i=2}^{n-1} (P_{i-1} - 2P_i + P_{i+1})^2 \tag{4.2}$$

The value λ has to be properly set depending on the dataset (e.g. using cross validation). If not specified, the penalization term is neglected ($\lambda = 0$). The total cost function to be minimised reads

$$\mathrm{Cost}(\{P_i\}) = C(\{P_i\}) + \mathrm{Pen}(\lambda, \{P_i\}) \tag{4.3}$$

Since the measurements D_j^{exp} suffer experimental uncertainty, the final value of the cost function is not 0 and even if synthetic data without uncertainty are considered, the number of solutions to which the minimization converges is generally not unique because of the underdetermination of the problem. The existence of a unique solution strongly depends on the set of experimental data.

Clustering

Since the solution of the least squares minimization is not unique, repeating the procedure with different initial guesses will return different sets of protection factors. The random search and the consequent minimization is thus repeated a number of times and a model-based clustering approach (Fraley and Raftery (2002)) is used to obtain sets of protection factors for each region of the whole chain that is covered by overlapping peptides. If the random search is applied M times ($M \approx 10^2 - 10^4$), the goal of the clustering algorithm is to reduce the degeneracy of the solutions from M to a smaller number, ideally finding one only cluster of solutions for each region covered by overlapping peptides.

The convergence of the algorithm is reached when the addition of solutions of the least squares minimization does not influence the clustering analysis, i.e. the number of identified clusters does not change.

The implemented clustering algorithm gathers sets of protection factors of areas covered by overlapping peptides. If the random search and minimization steps have been applied M times, M values of protection factors are associated to each residue and a histogram can be built. Considering the N residues forming a region of the protein composed by overlapping peptides, the multivariate distribution of the protection factors of those N amino acids can be acquired. The model based clustering algorithm (Scrucca et al. (2016)) fits such distribution with a mixture of multivariate gaussians and the fit is optimized using the EM algorithm (Bishop (2006)).

A mixture of gaussians is a linear superimposition of gaussians:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{4.4}$$

where $\boldsymbol{\mu}_k$ is the mean of the k-th gaussian and $\boldsymbol{\Sigma}_k$ the covariance matrix associated to the k-th component. The weights π_k associated to the components are called mixing coefficients and are normalized to 1, i.e. $\sum_k \pi_k = 1$.

The log-likelihood function of the mixture of gaussians in Eq. 4.4 for a set of N obser-

vations $\mathbf{X}$ reads

$$\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right] \tag{4.5}$$

Maximizing the derivative with respect to $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ and $\boldsymbol{\pi}$ and defining

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{4.6}$$

an estimate of the means, covariances and mixing coefficients can be obtained:

$$\begin{aligned} \mu_k &= \tfrac{1}{N_k} \textstyle\sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n \\ \boldsymbol{\Sigma}_k &= \tfrac{1}{N_k} \textstyle\sum_{n=1}^{N} \gamma(z_{nk}) (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \\ \pi_k &= \tfrac{N_k}{N} \end{aligned} \tag{4.7}$$

where $N_k = \sum_{i=1}^{N} \gamma(z_{nk})$ can be interpreted as the effective number of points assigned to cluster k.

The solution of the EM algorithm is found by the iterative application of two steps. In the expectation step (or E step), the posterior probabilities (Eq. 4.6) are computed starting from randomly initialised values of means, covariances and mixing coefficients. Consequently, in the maximization step (or M step), the posterior probabilities are used to estimate $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ and $\boldsymbol{\pi}$. The procedure is repeated until the change of the log-likelihood (Eq. 4.5) falls below a certain threshold.

ExPfact applies the EM algorithm with mixtures of gaussians formed by $K = 1 - 99$ components. The optimal number of components is chosen using the BIC.

4.2 Application to synthetic data

The ExPfact algorithm is applied to synthetic data with no error associated to generated experimental measurements. First, a toy model is used to reproduce the results obtained

by Skinner et al. (2019). Furthermore, ExPfact is applied to a dataset with an ideal overlapping and the outcomes are compared with HDSite (Kan et al. (2013)).

4.2.1 Toy model

Dataset

A synthetic dataset adapted from the original article by Skinner et al. (2019) is created. The purpose is to study a fictitious 15-residue peptide with sequence IDSQVLCGAVKWLIL and to give insights on how the algorithm faces underdetermination and the problem of switchable residues. Reference protection factors and a peptide map are assigned (Fig 4.2). The dataset has 100% coverage and redundancy 0.47.

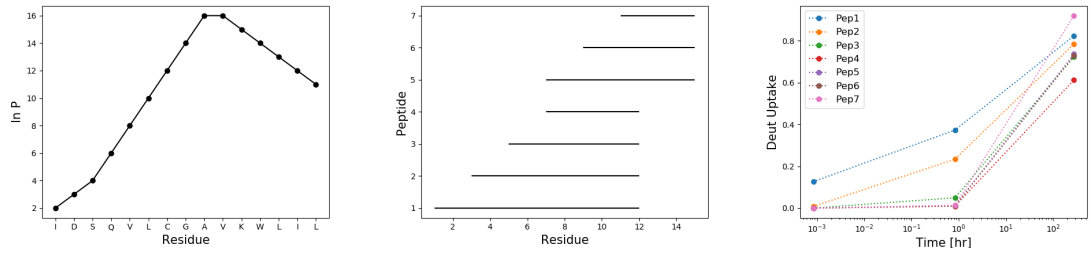

Figure 4.2: Synthetic data.Left: reference protection factors of the fictitious peptide. Centre: peptide map of the 15-residue peptide. Right: data generated from protection factors and assignments calculated at times 0.00083, 0.83330 and 277.78 hours at temperature 300 K and pH 7. Different colors represent different fragments.

Using the reference protection factor and the assignments in Fig 4.2 we can evaluate the deuterium uptake at any time point for each fragment via Eq. 2.16. To calculate the intrinsic exchange rates k_i, we set temperature $T = 300$ K and $pH = 7$. We generate a dataset with 3 time points at times 0.00083, 0.83330 and 277.78 hr and we associate no experimental error to the synthetic data (Fig. 4.2, right).

By the application of the ExPfact algorithm, we aim to reproduce the exact pattern of protection factors shown in Fig. 4.2 starting from the generated synthetic data.

Random Search

10000 random sets of protection factors are initialised with the constraint $0 \leq \ln P_i \leq 20$. For each set, the cost function in Eq. 4.1 is evaluated and the set with the best agreement with experimental data is selected and used as initial guess for a least squares minimization. Performing one run of the algorithm, an optimized pattern of protection factors is obtained. An example is shown in Fig. 4.3. The solutions of three different runs are compared in Fig. 4.4.

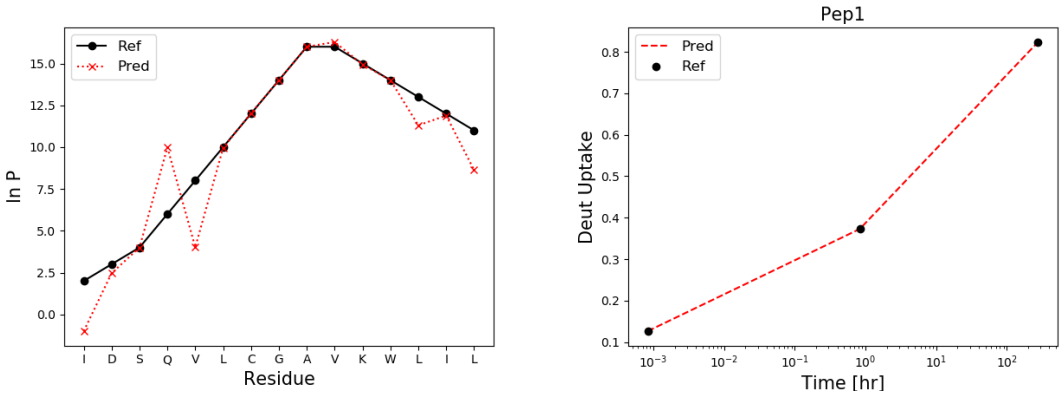

Figure 4.3: Results of the random search. On the left, one optimized solution (red) is compared with the reference pattern of protection factors (black). On the right, the prediction (red line) for peptide 1 is compared with experimental data (black dots).

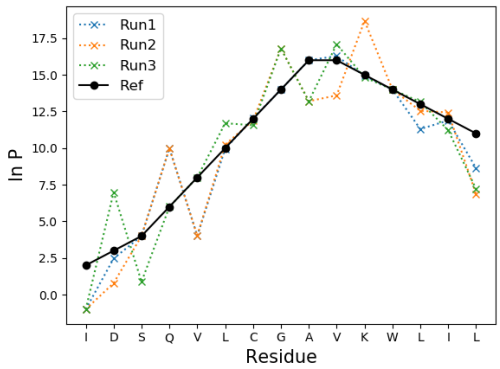

Figure 4.4: Degeneracy of ExPfact. Three extracted sets of protection factors with same agreement with experimental data are compared with the reference pattern.

As can be noticed from Fig. 4.3 and Fig. 4.4, the estimated pattern of protection factors differs from the reference set. On the other hand, the predictions of deuterium uptake differ from experimental data only at the order of machine precision. As a consequence, different patterns of protection factors have the same agreement with experimental data.

Because of the degeneracy of solutions, a proper statistic needs to be accumulated to return robust outcomes. The random search is repeated 100 times, thus finding 100 different patterns of protection factors with similar agreement with protection factors. The results are summarised in Fig. 4.5.

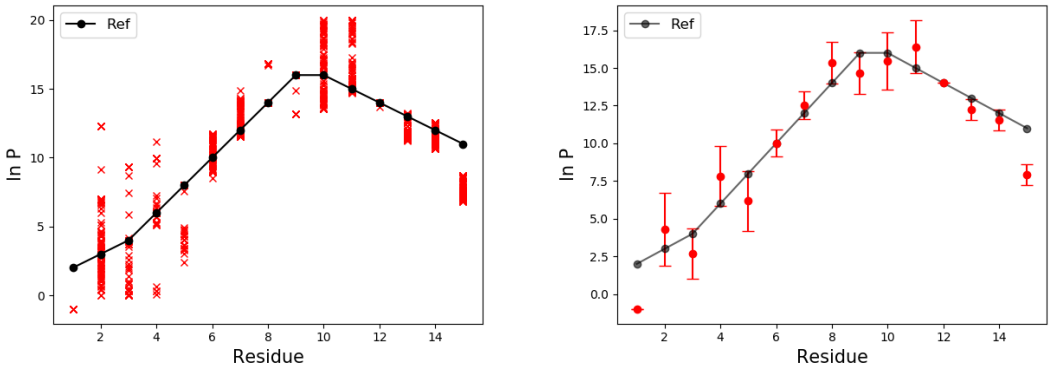

Figure 4.5: Random search applied to synthetic data. On the left, the protection factors find for each residue by each run (red crosses) are compared with the reference pattern. On the right, mean values and relative standard deviations (red bars).

The mean values of protection factors found by the random search (Fig. 4.5) show the average value among the 100 runs of the algorithm. If associated with one standard deviation, they represent the 68% confidence interval, including the reference protection factor for most residues. The predictions of deuterium uptake calculated from the mean values are not necessarily as good as the ones calculated by estimates of single runs. This is due to the fact that several residues show a multimodal behaviour that can be visualized with histograms of the estimated protection factors for different residues (Fig. 4.6). The multimodality shown in Fig. 4.6 justifies the usage of a mixture of gaussians as a fitting model for the clustering algorithm.

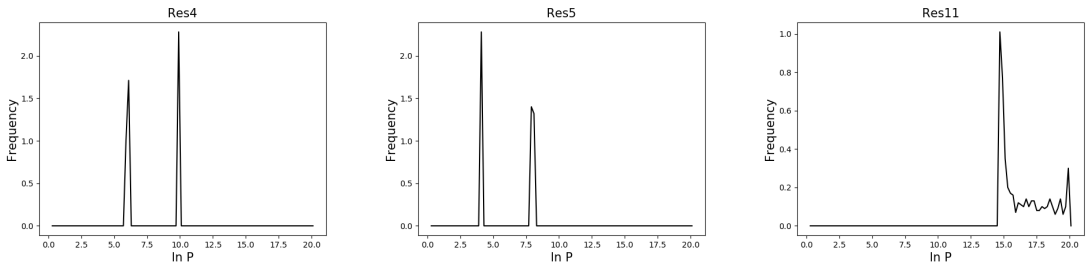

Figure 4.6: Distribution of the protection factors for residues 4, 5 and 11.

Clustering

The histograms in Fig. 4.6 are the marginal distributions of the multivariate probability distribution that is given as input to the clustering algorithm. The model-based approach is implemented in the R library `mclust` (Scrucca et al. (2016)).

Running 1000 times the algorithm and selecting the top 500 solutions (i.e. the half of solution with lowest cost function), the clustering algorithm identifies 4 components, drastically reducing the degeneracy of the sets of protection factors. The number of components is chosen using the BIC (Fig. 4.7).

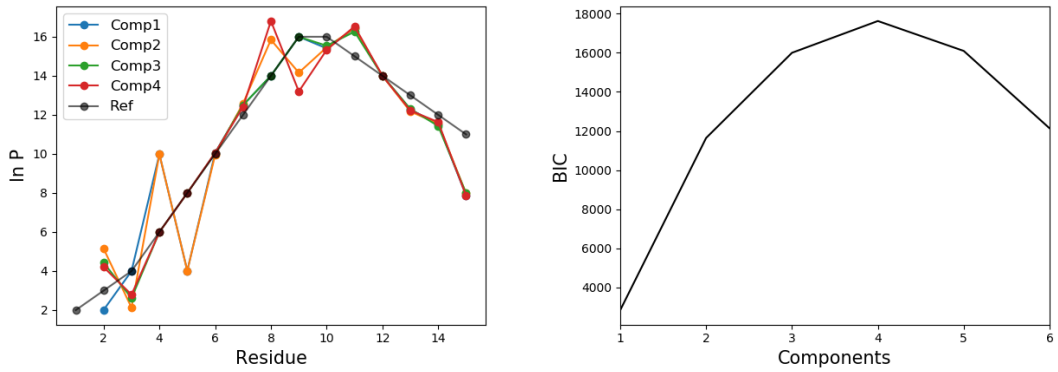

Figure 4.7: Clustering algorithm for synthetic data. On the left, the 4 patterns of protection factors (different colors) identified by the clustering algorithm are compared with the reference set (black). On the right, BIC values associated to 1-6 components.

The clusters identified in Fig. 4.7 unveil the problems of underdetermination and switchable residues. Underdetermination is reduced since the degeneracy of solutions is reduced

from 500 solutions down to 4 components. However, even if synthetic data with no experimental uncertainty have been used, it is not possible to completely solve the issue.

Concerning the problem of switchable residues, some adjacent residues (2-3, 4-5, 8-9) show two different and coupled patterns of protection factor. For instance, the highest protection factor of residue 4 is coupled with the lowest one of residue 5. However, it is interesting to notice that some residues (6, 7, 10, 12, 13, 14) are provided with one only protection factors (the 4 components coincide) compatible with the reference pattern.

4.2.2 The ideal dataset

The development of a software must be accompanied by the comparison with existing methods aiming to reach the same goal. As described in section 3.3.1, the state of the art of HDX-MS data analysis is represented by HDSite (Kan et al. (2013)). The latter is able to reach single residue resolution if all the fragments of the dataset differ of one only residue. A second synthetic dataset is built to proof that ExPfact is also able to reach such resolution if an ideal overlapping of peptides is available.

The peptide used to generate synthetic data has the same sequence used in section 4.2.1: the reference pattern of protection factor is shown together with the peptide map and the generated synthetic data in Fig. 4.8. Three experimental data are generated at time points 0.00083, 0.83330 and 277.78 hours at temperature 300 K and pH 7. The dataset has 100% coverage and redundancy 0.6.

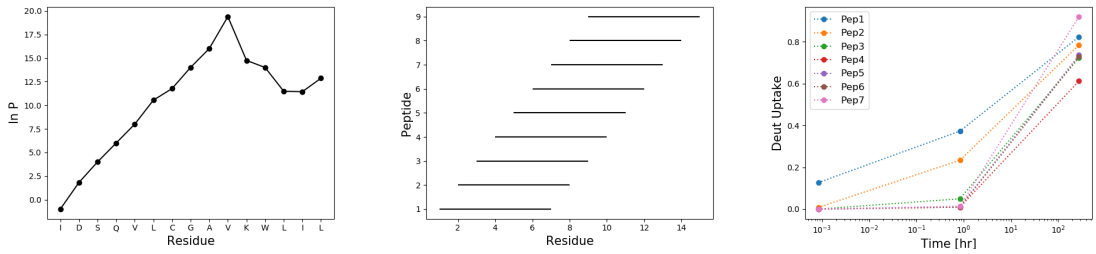

Figure 4.8: Ideal dataset. Synthetic data are generated from a reference pattern of protection factors (left) for a set of fragments with ideal overlapping (centre). Experimental points are generated at three time points at temperature 300 K and pH 7 (right).

The ExPfact algorithm is applied to the ideal dataset shown in Fig. 4.8. The random

search is repeated 1000 times and the top 500 solutions are used as input for the clustering algorithm which identifies one only component which is compatible with the reference pattern (Fig. 4.9).

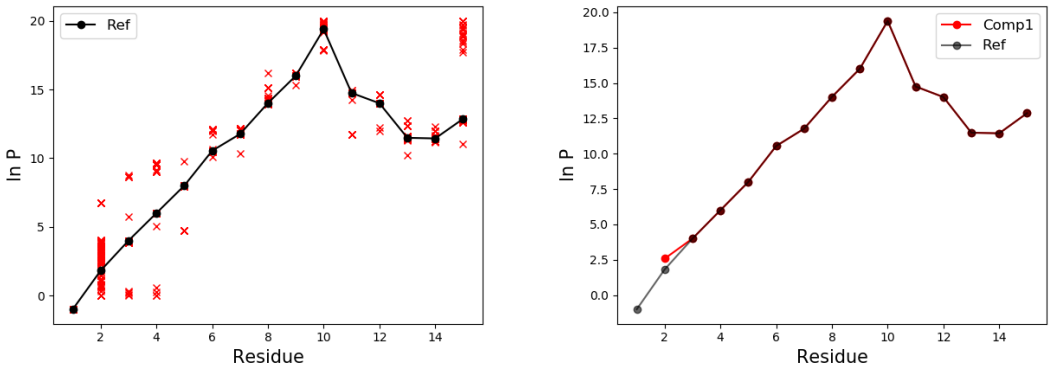

Figure 4.9: ExPfact for the ideal dataset. On the left, the random search is applied 1000 times to the ideal dataset in Fig. 4.8 and a comparison with the reference pattern of protection factors is shown. On the right, the cluster identified by ExPfact and the reference pattern are compared.

The results in Fig. 4.9 show that ExPfact is able to estimate protection factors at single residue resolution if a dataset with ideal overlapping is available.

This unveils the interconnectedness between underdetermination and the problem of switchable residues. The ideal overlapping in the dataset in Fig. 4.8 erases the problem of switchable residues and this is sufficient for ExPfact to reproduce the reference pattern of protection factor.

4.3 Software availability

ExPfact is available at the following github repository:

https://github.com/skinnersp/exPfact

Scripts are accompanied by test data and text files describing how to use the algorithm to replicate the results in the original paper by Skinner et al. (2019). The software is mainly written in Python but includes scripts implemented in R, Bash and Fortran.

Chapter 5

Application to real world data

Summary of the chapter. The ExPfact algorithm is applied to real data. The application to mouse prion protein validates the algorithm through a comparison with protection factors of the same system extracted from HDX-NMR measurements under the same experimental conditions. Moreover, the application to glycogen phosphorylase shows the potentiality of ExPfact when dealing with a high-quality dataset.

5.1 Application to moPrP

As a first application of the ExPfact algorithm to real world data, a dataset studying the mouse prion protein is analysed (Moulick et al. (2015)). Since experimental data are available for the same protein under the same conditions from both HDX-MS and HDX-NMR techniques, a comparison between protection factors extracted by the algorithm and exchange rates estimated by NMR is performed to validate the algorithm.

5.1.1 Prion proteins

The importance of prion proteins is here summarized following Apriola (2001). Back in 1970s, Great Britain altered the process through which animal carcasses are rendered to provide meat and bone meal supplements to sheep, cattle and other livestock. As a consequence, after approximately one decade, a new pathology was found in British cattle

population and it was classified in the group of transmissible spongiform encephalopaty (TSE) and thus named BSE (bovine spongiform encephalopaty). In 1996, a previous unknown form of TSE was detected in human population: the Creutzfeldt-Jacob disease (CJD). The correlation between CJD and BSE has been supported by several studies and more than 90 variants have been identified. CJD represents the emergence of TSE as a potentially widespread health treat to human population: this is the reason why it is crucial to determine mechanisms underlying TSE pathologies.

TSE diseases are characterised by the abnormal accumulation of an anomalous form of the prion protein (PrP). The normal PrP is a glycoprotein formed by approximately 250 amino acids and it is expressed in the cell surface of several tissues. It is both soluble and sensitive to digestion with Proteinase-K. On the other hand, the anomalous PrP forms insoluble aggregates and is partially resistant to digestion with Proteinase-K. Because of their characteristics, TSE diseases are another form of amyloid diseases like Alzheimer's, Huntington's and type II diabetes. The main difference is that TSEs are transmissible. Moreover, since no virus or bacteria associated to the disease have been found, TSE diseases have a unique etiology. The key role of PrPs in TSE diseases is undeniable and studying the unfolding dynamics of such proteins is crucial in understanding the mechanisms behind CJD.

The dataset that we are going to analyse is taken from Moulick et al. (2015). In the paper, HDX-MS and HDX-NMR are coupled to characterize the structural and energetic properties of the native state of the mouse prion protein (moPrP). Various segments are found to undergo subglobal unfolding events at pH 4, condition at which the misfolding to a $\beta-$rich conformation is favoured. In addition, the native state is found at equilibrium with at least two partially unfolded forms (PUFs) that can be accessed through stochastic fluctuations around the native state and have altered surface exposure. Moreover, one of these PUFs resembles a conformation that has been found to be an initial intermediate in the conversion of the monomeric protein into the misfolded oligomer.

5.1.2 Dataset

The dataset used to study HDX of the MoPrP was provided by the authors of the above mentioned paper (Moulick et al. (2015)). The protein structure is available within the

Protein Data Bank database (Berman et al. (2000)) with the PDB code 1AG2. The energetic and structural characterization of the native state of the protein is probed at pH 4.0 and room temperature (25 ^oC), conditions at which the misfolding to the oligomer conformation should be favoured. The sample was inserted in a 95% deuterated buffer. To stop the exchange and obtain more specific information, the solution was quenched at pH 2.4 and temperature 0 ^oC and digested by pepsin.

The experiment leads to the identification of 14 fragments. The peptide map is shown in Fig. 5.1. The dataset here studied has a coverage of 69% and a redundancy of 0.14. No fragments have been identified for a central region involving residues 45-59.

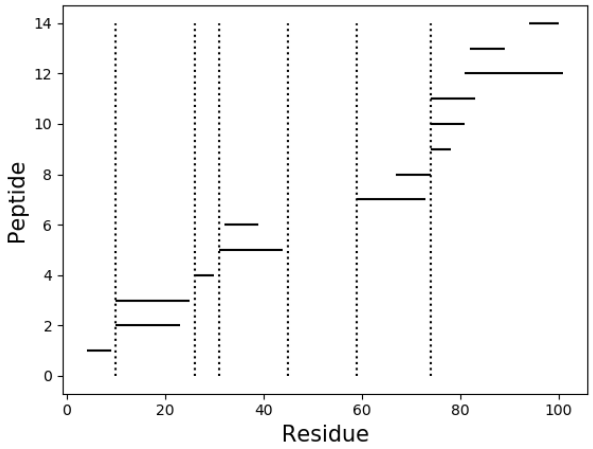

Figure 5.1: Peptide assignments for moPrP dataset. 14 fragments are identified by Moulick et al. (2015) and shown (horizontal solid lines). Dotted lines enhance the regions of the protein covered by overlapping peptides. Coverage: 70 %; redundancy: 0.14.

Despite the low values of coverage and redundancy of the dataset, the exchange of each fragment has been studied with quite a remarkable temporal sampling: 15 time points are available, ranging from 5 s to 24 hr. Some examples of the uptake curves for this dataset are shown in Fig. 3.3.

5.1.3 Results

The ExPfact algorithm is used to extract protection factors out of the dataset with peptide map shown in Fig. 5.1. The algorithm is initialised with 10000 random sets of protection factors and the best one is selected as initial guess for a penalised least squares minimization. The penalization term is set to $\lambda = 2 \times 10^{-5}$ by cross validation.

Cross validation

A penalised least squares minimization is introduced to both reduce overfitting and avoid abrupt changes in the final pattern of protection factors. The functional form of the penalty term is shown in Eq. 4.2:

$$\mathrm{Pen}(\lambda, \{P_i\}) = \lambda \sum_{i=2}^{n-1} (P_{i-1} - 2P_i + P_{i+1})^2$$

The penalization parameter λ must be properly set.

Leave-one-out cross validation was used. The dataset is divided into a train dataset composed by 14/15 time points (for all fragments) and a test dataset composed by the remaining time point. The splitting is repeated by leaving out one time point at a time from the train dataset so that 15 train (and test) datasets are generated (Fig. 5.2).

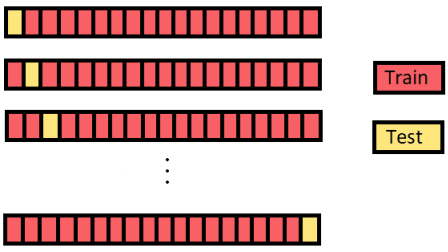

Figure 5.2: Schematic representation of leave-one-out cross validation. Figure from https://aiaspirant.com/cross-validation.

The penalised least squares minimization is performed using the training dataset and protection factors are estimated. The performance of the minimization at a fixed value

of the penalization parameter λ is evaluated by setting a seed for the pseudo-random number generator. The cross validation error

$$\text{CVE} = \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K}\left(D_{i,k}^{pred} - D_{i,k}^{exp}\right)^2 \tag{5.1}$$

is evaluated for the train dataset. Here M is the number of time points forming the specific dataset. For the train dataset, $M = 14$; for the test dataset, $M = 1$. The cross validation error must be calculated for all the K fragments available. The estimated protection factors are used to make predictions for the remaining time point of the test dataset and the test cross validation error is evaluated. The total cross validation error error is the sum of the train and the test validation error.

Leave-one-out cross validation is applied to the moPrP dataset with penalization parameters ranging from 10^{-15} to 10^{0}. The results in Fig. 5.3 show that the total cross validation error reaches a minimum at $\lambda = 2 \times 10^{-5}$. Such a value is used as the penalization parameter for the penalised least squares minimization.

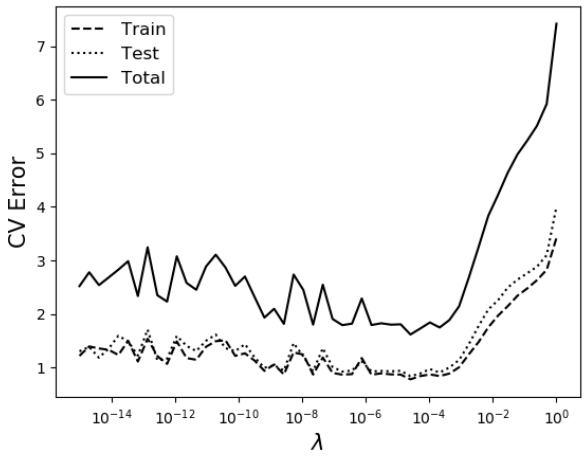

Figure 5.3: Leave-one-out cross validation on the moPrP dataset. Train error (dashed line), test error (dotted line) and total error (solid line) are shown as a function of the penalization term λ.

Predictions

The predictions of the uptake calculated by the penalised minimization for one run of the algorithm are shown in Fig. 5.4 for some fragments.

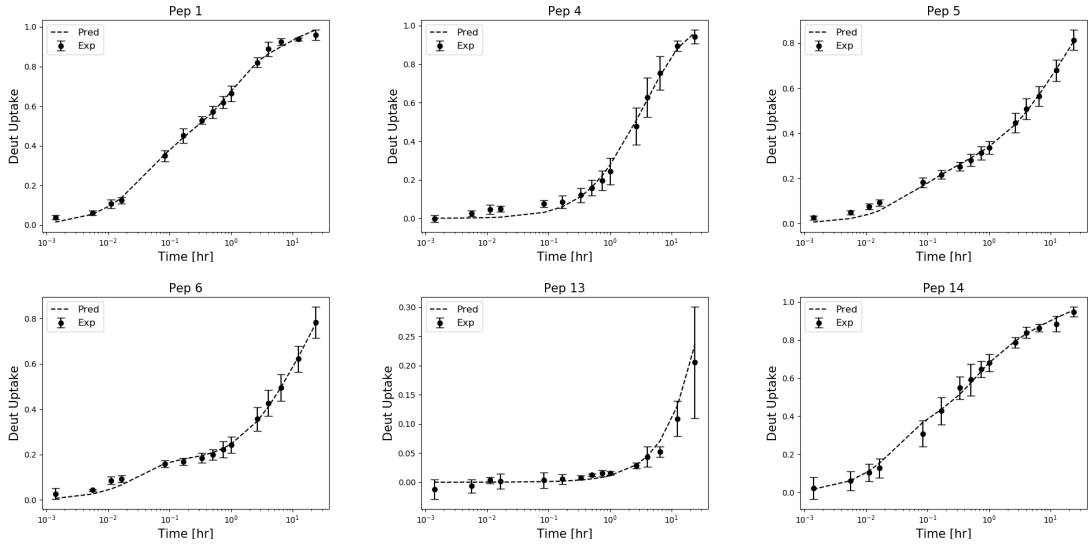

Figure 5.4: Deuterium uptake predictions for the moPrP dataset. Experimental uptake (black dots) are compared with the deuterium uptake curves (dashed line) predicted by one run of the ExPfact algorithm for peptides 1, 4, 5, 6, 13 and 14

The results in Fig. 5.4 show a nice agreement between experimental and predicted curves that can be quantified by looking at the average cost function over all the peptides that is of the order of magnitude of 10^{-3}.

Clustering

Each run of the algorithm generates a different pattern of protection factors. In order to accumulate a proper statistic, 5000 solutions were evaluated and the best 2500 were selected to generate histograms of protection factors for each residue. Such histograms are the marginal probability distributions of protection factors given as input to the clustering algorithm. The outcomes of the clustering algorithm for two specific regions are analysed. First, the region with the best overlapping and redundancy in the dataset (residues 79-84); second, a region with poor quality (residues 5-9).

Residues 79-84

The region with the best overlapping and redundancy of the dataset is formed by fragments 9, 10 and 11 and covers the residues 79-84 (Fig. 5.1). Histograms in this region show one only peak for all adjacent residues (Fig. 5.5).

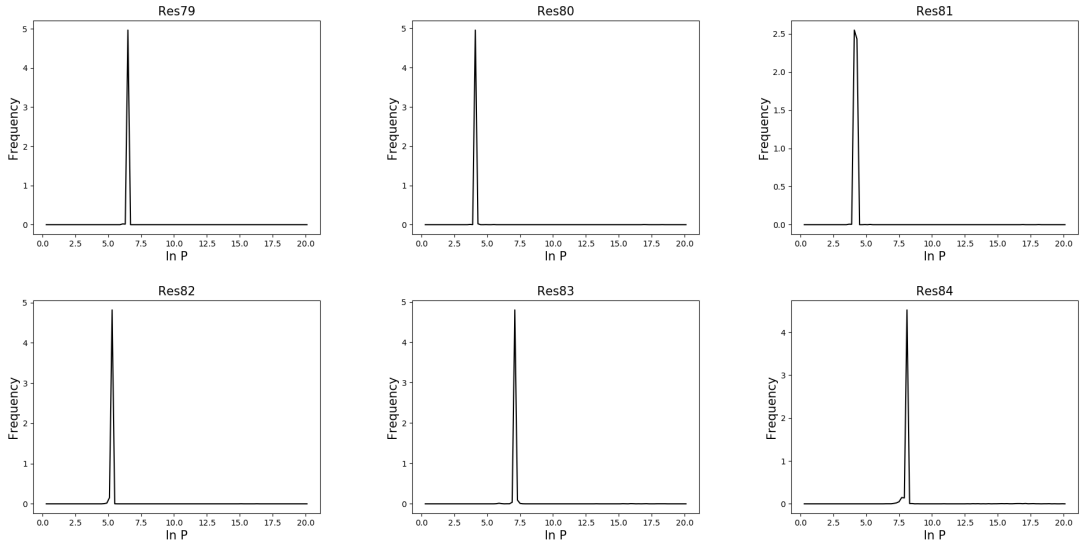

Figure 5.5: Histograms of protection factors for residues 79-84. One only peak is found from 2500/5000 runs of the ExPfact algorithm.

As a consequence, one only cluster is identified in this region (Fig. 5.7A). For regions where one only gaussian component is identified, one only protection factor is identified at single residue resolution.

Residues 5-9

A region with very poor overlapping can be identified in the area of the protein described by the first fragment, covering residues 5-9. Because of the low number of amino acids involved and the high number of time points available, underdetermination should be solved. However, we are not able to estimate protection factors at single residue resolution because of the problem of switchable residues. In fact, if one only peptide is considered, the problem of switching residues cannot be solved: the set of predicted

exchange rates (k_i/P) is the same for different runs, but such rates can be arbitrarily associated to one residue or another (see histograms in Fig. 5.6).

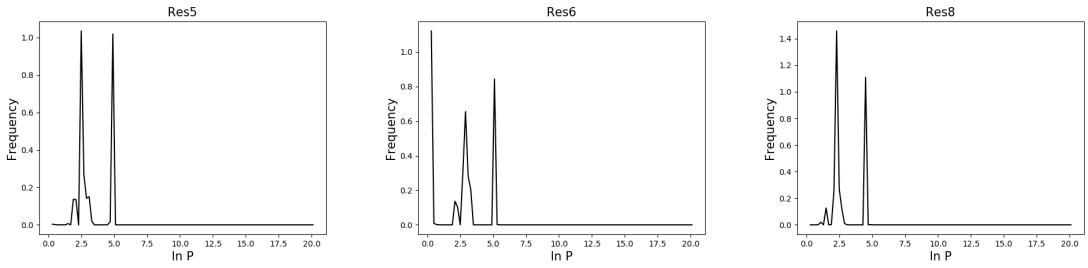

Figure 5.6: Histograms of protection factors for residues 5-9. Different peaks are found from 2500/5000 runs of the ExPfact algorithm.

As a consequence, more components are identified by the clustering algorithm. For the specific region covering residues 5-9, 7 clusters are estimated (Fig. 5.7B). Areas with poor overlapping and/or redundancy are thus characterised by many clusters of protection factors. The ideal solution to the problem would be to repeat the experiment by increasing the quality of data. Most times, this is unfeasible. To tackle the problem and associate one only protection factor to each residue, one could consider mean values only for those regions where more clusters have been identified.

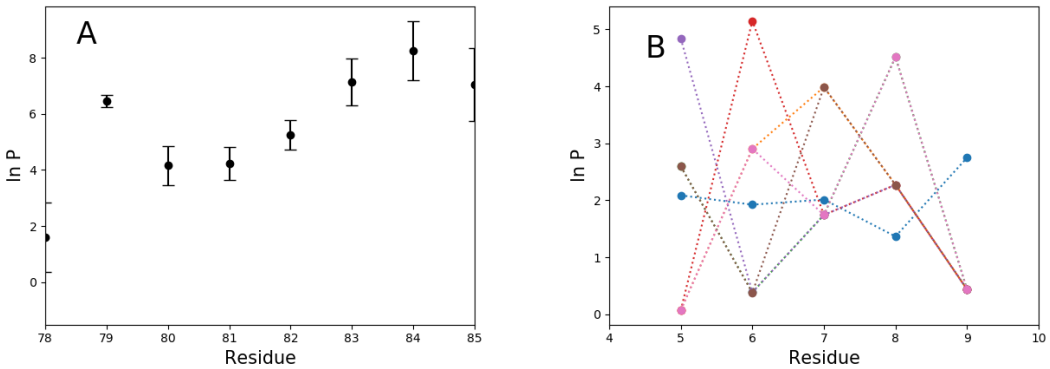

Figure 5.7: A. Unique cluster of protection factors identified for residues 79-84. B. 7 clusters identified for residues 5-9. Standard deviations are not shown.

The results shown in Fig. 5.7 enhance the fact that a dense temporal sampling (15 time points) is not sufficient to cancel out the underdetermination of the problem (Fig. 5.7B).

On the other hand, if a proper number of time points is coupled with a good overlapping of fragments, then single residue resolution can be achieved (Fig. 5.7A).

Validation

The moPrP dataset provides measurements of hydrogen deuterium exchange rates from both mass spectrometry and NMR experiments for the same protein under the same conditions (pH 4.0, temperature 25^oC). A comparison between the outcomes of the two techniques is able to determine if the protection factors extracted by the ExPfact algorithm are the *real* exchange rates of the protein.

The protection factors extracted at single residue resolution by MS and NMR are compared in Fig. 5.8. For the regions where one only cluster is found by the clustering algorithm, the centroid and standard deviation of the cluster is shown. Concerning the areas where more clusters are identified, mean values are considered.

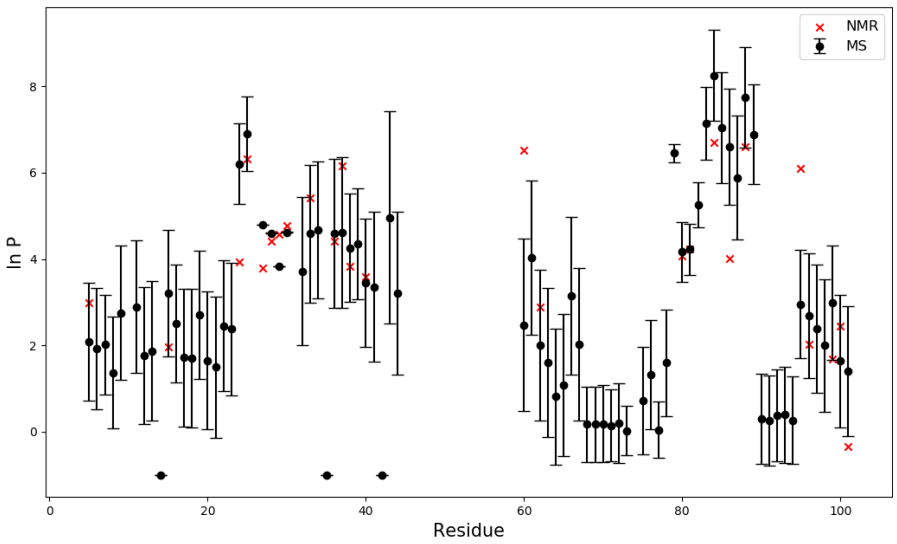

Figure 5.8: Comparison between protection factors extracted by ExPfact (black bars) and by NMR (red crosses) for the moPrP at pH 4 and temperature 25 ^oC.

The results in Fig. 5.8 show that protection factors extracted by NMR and MS experiments are compatible for most residues (for whom measurements are available). Notice

that clusters with protection factor $\ln P = -1$ may identify prolines or the first residue of a region covered by overlapping fragments. Since the clustering algorithm makes use of gaussians distributions, it is possible that the predicted cluster has a standard deviation leading to negative protection factors, which is not possible.

The correlation between the measurements from the two techniques is shown in Fig. 5.9 and is found to be 0.64. All the NMR measurements are compatible within 3σ to the protection factor extracted by ExPfact. Moreover, the 83% of measurements (19/23) is compatible within one standard deviation.

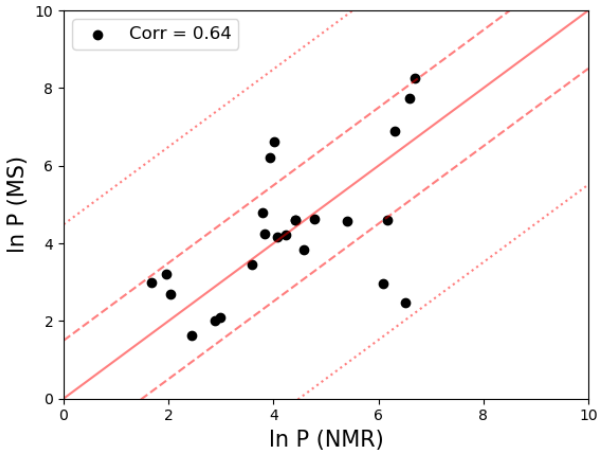

Figure 5.9: Correlation between protection factor extracted by NMR (x axis) and MS experiments (y axis). The red dashed line represents one standard deviation from the line $y = x$ (solid red line), the red dotted line represent the 3σ interval.

The compatibility between protection factors extracted by the ExPfact and calculated by NMR validates the ExPfact algorithm.

5.2 Application to Glycogen Phosphorylase

The second application of the ExPfact algorithm to real world data regards the high quality dataset provided by Kish et al. (2019) to study the allosteric regulation of glycogen phosphorylase. The high quality is guaranteed both by a remarkable overlapping and redundancy and by a dense temporal sampling exploiting the millisecond regime.

5.2.1 Glycogen Phosphorylase

Glycogen phosphorylase (GlyP) is the enzyme that catalyses the sequential removal of glycosyl residues from unbranched glycogen, producing glucose 1-phospate (Lackie (2019)). The historical importance of GlyP is given by the fact that it was the first allosteric enzyme to be discovered. In biochemistry, allosteric regulation is the regulation of an enzyme by binding an effector molecule at a site (allosteric site) other than the enzyme's active site. The effector acts by altering the equilibrium between the active and inactive state.

Since GlyP regulates the glycogen metabolism, understanding in detail its mechanics is fundamental for a number of objectives. In fact, glycogen regulates glycemia in the liver, it provides energy for muscular contraction in muscles and acts as an emergency store in the brain (Mathieu et al. (2017)).

The goal of Kish et al. (2019) in studying the allosteric regulation of glycogen phosphorylase is to quantify changes in local stability between the activated and inhibited forms of the enzyme. Despite being one of the most studied enzymes, the nature of its regulatory mechanism remains ambiguous. In particular, changes in local stability in response to allosteric regulation are detected for the first time in the so-called tower helix, i.e. the 280s loop gating access to the active site.

Moreover, Kish et al. (2019) describe the construction, validation and implementation of a novel HDX-MS apparatus allowing measurements in the millisecond regime.

5.2.2 Dataset

The dataset used to study the allosteric regulation of GlyP through HDX was provided by Kish et al. (2019). GlyP was studied under three different conditions: the apo state, i.e. the inactive and unbound state (PhosA), the active state (PhosB) and the bound inactive state (PhosC). The structure of the protein is available for the active and inactive states with PDB codes 1GPB and 9GPB respectively. These three states were studied under physiological pH 7 and at temperature 23 ^oC. The solution was then quenched at temperature 0^oC and pH 2.5 and digested by pepsin.

The 263 fragments identified in the dataset are shown in the peptide map in Fig. 5.10. The assignments have a coverage of 96% and a redundancy equal to 0.31. Despite the high number of fragments provided, the redundancy is still low because of the dimensions of the enzyme: GlyP is formed by 842 residues. For each fragment, 9 time points were acquired, ranging from 50 ms to 300 s. The experiment was performed in triplicate.

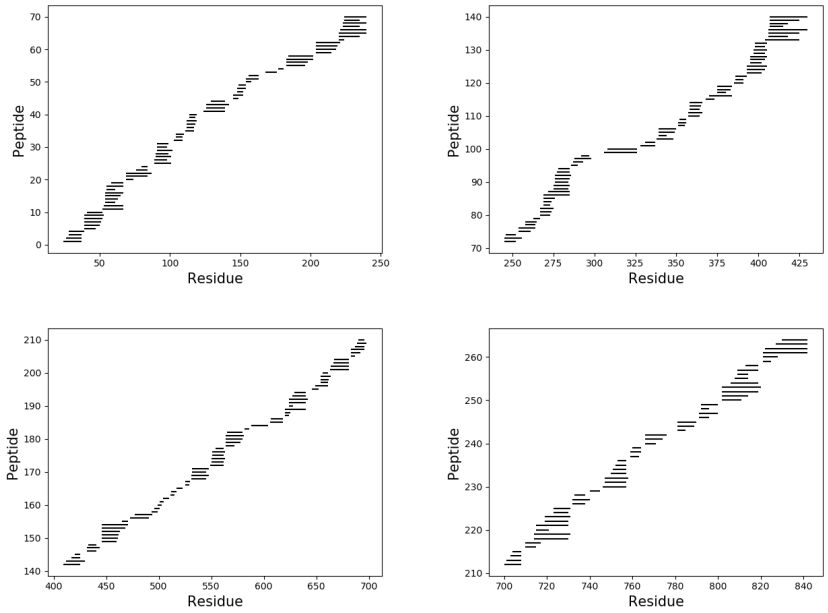

Figure 5.10: Peptide assignments for the GlyP dataset. The 263 fragments identified by Kish et al. (2019) are shown with horizontal solid lines in 4 different plots. Top left: fragments 1-70; top right: 71-140; bottom left: 141-210; bottom right: 211-263. Coverage: 96%; redundancy: 0.31.

5.2.3 Results

The ExPfact algorithm was applied to the dataset with peptide map in Fig. 5.10. 1000 runs of the algorithm were performed and the top 700 were considered.

The predictions for the extensive GlyP dataset started being challenging in terms of computational cost. The average time required to compute one solution is about 30 minutes. The average value of the cost function for the top 700 solutions is shown in Fig. 5.11 and is equal to 0.03 for PhosA and PhosC, while it is slightly higher (0.04) for

PhosB. This degree of agreement with experimental data was a compromise between the quality of the predictions and the time required to compute one solution.

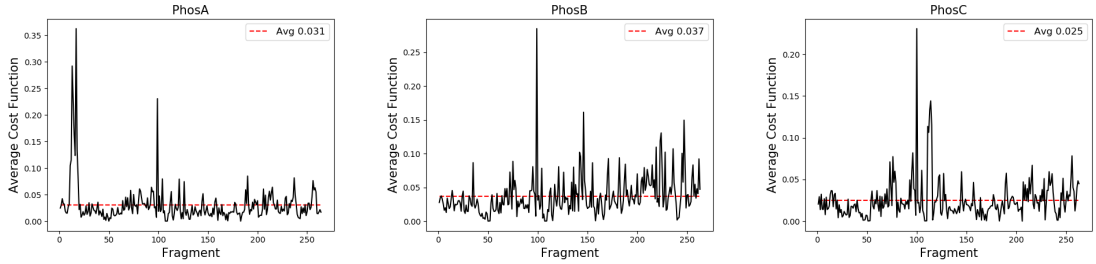

Figure 5.11: Average cost function per fragment for states PhosA (left), PhosB (centre) and PhosC (right) over the top 700/1000 runs. Red dotted lines represent the average cost function over all fragments.

Some peptides have higher average cost function with respect to others because the inverse of the standard deviation was used to weight experimental data. The higher the standard deviation, the lower the weight of a specific measurement. As a consequence, fragments with greater uncertainties are associated with higher cost functions.

When the clustering algorithm identifies one only cluster, the protection factor associated to the residue is the cluster itself. When more clusters are identified, the protection factor is computed as the average value over all the runs considered. The resulting pattern of protection factors for the whole protein is depicted in Fig. 5.12.

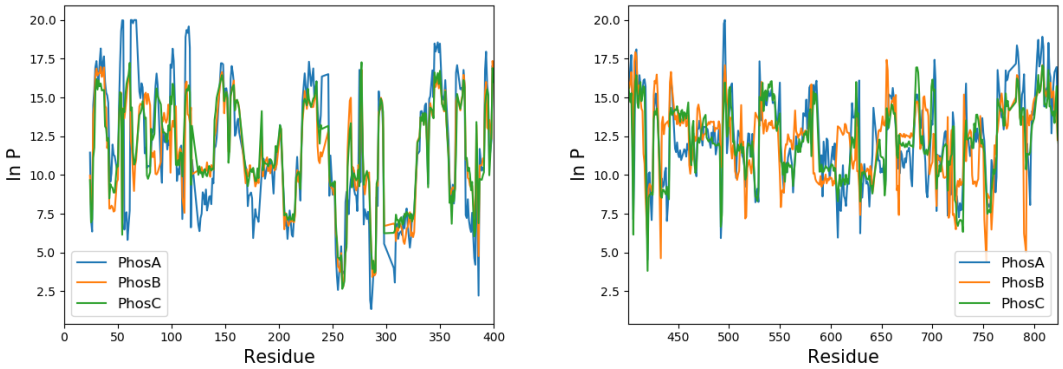

Figure 5.12: Protection factors for states PhosA (blue), PhosB (orange) and PhosC (green). On the left, residues 1-400. On the right, residues 401-823.

The differences between the patterns of protection factors estimated for the different states can be quantified through a t-test. To confirm the findings of Kish et al. (2019), we show the results of the t-test statistic for a comparison between states PhosA (apo state) and PhosC (bound inactive state) in Fig. 5.13.

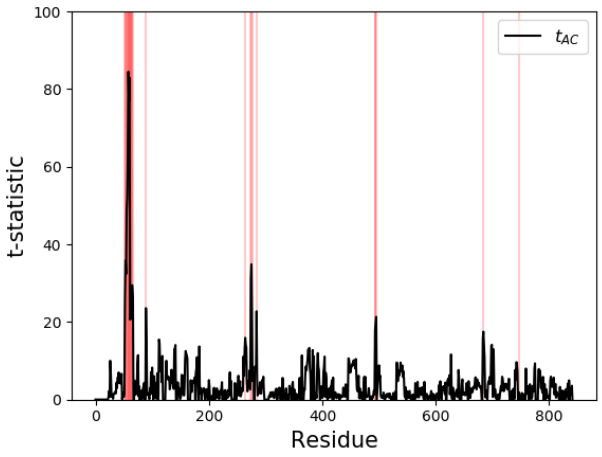

Figure 5.13: Results of t-test analysis of patterns of protection factors for states PhosA and PhosC. Most relevant regions (p value $< 10^{-10}$) are shown with vertical red lines.

As stated by Kish et al. (2019), PhosA and PhosC states are similar. The most significant changes can be identified in the region of the allosteric sites (residues 35-78) since the effector is bound to the enzyme in the state PhosC and is not in state PhosA. Minor differences can be found in the region of the tower helix providing the evidence of an entropic switch regulating the access of the substrate to the active site.

Allosteric sites (residues 35-78)

The first region of interest that we analyse is the area where allosteric sites are situated, formed by residues 35-78. Deuterium uptake curves for 6 fragments partially covering this region are shown in Fig. 5.14. Some of them show a compatible behaviour among the different states (peptides 14 and 17). The majority of the fragments, however, highlights significant differences between the curve for the apo state PhosA with respect to the other states (peptides 15, 16, 18, 19).

The extracted pattern of protection factors is depicted in Fig. 5.15. Again, the state PhosA shows significant differences with respect to states PhosB and PhosC. One advantage of using protection factors (Fig. 5.15) instead of uptake curves (Fig. 5.14) is the ability to encode the information in one only parameter. Moreover, being ExPfact able to reach single residue resolution, local changes can be further localised.

Tower helix (residues 246-286)

A second region that mostly showing minor differences between states PhosA and PhosC (Fig. 5.13) is the area around the 280s loop. The uptake curves in this region are shown in Fig. 5.16 for peptides 72-77. For peptides 72 and 73, the behaviour of states PhosA and PhosB is compatible while significant changes can be visualized for state PhosC. In other fragments (peptides 74, 75 and 76), the uptake is different for each state. Another situation is found in peptide 77 where PhosA differs from PhosB and PhosC.

The patterns of protection factors are shown in Fig. 5.17, highlighting significant differences only for state PhosA with respect to the other states and only in specific regions (residues 260-265 and 275-286). Protection factors for the active and inactive states PhosB and PhosC are compatible. The usage of protection factors enhances that the differences in uptake, whose interpretation differs from peptide to peptide (Fig. 5.16), can be summarized in a difference in terms of protection factor of the state PhosA with respect to states PhosB and PhosC.

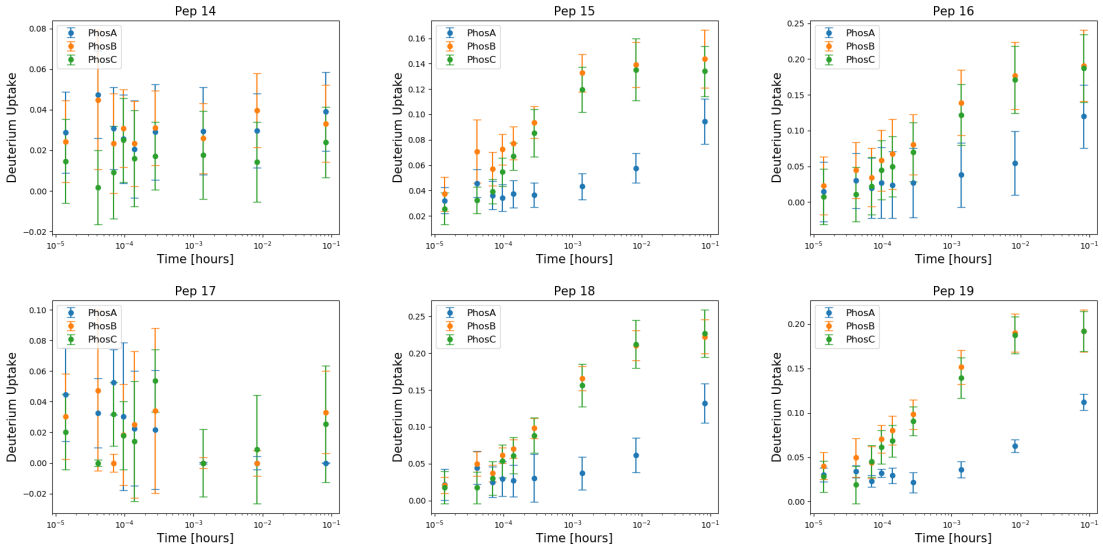

Figure 5.14: Deuterium uptake for peptides 14-19 within the region where allosteric sites are situated (residues 35-78). States PhosA (blue), PhosB (orange) and PhosC (green) are compared.

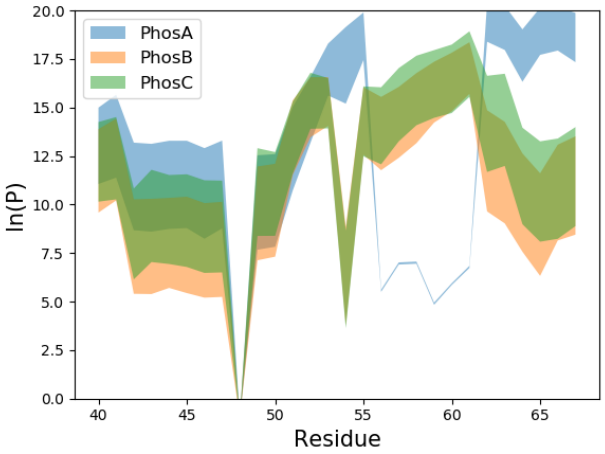

Figure 5.15: Patterns of protection factors of residues 40-67, partially covering the allosteric sites, for states PhosA (blue), PhosB (orange) and PhosC (green). Confidence intervals are shown within one σ.

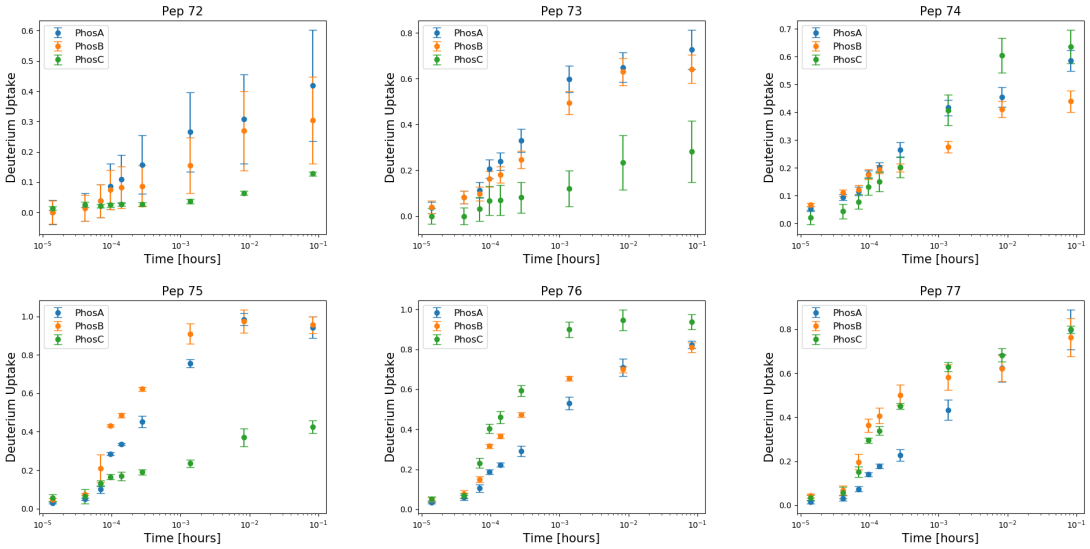

Figure 5.16: Deuterium uptake for peptides 72-77 within the region where the tower helix is situated, close to the 280s loop.. States PhosA (blue), PhosB (orange) and PhosC (green) are compared.

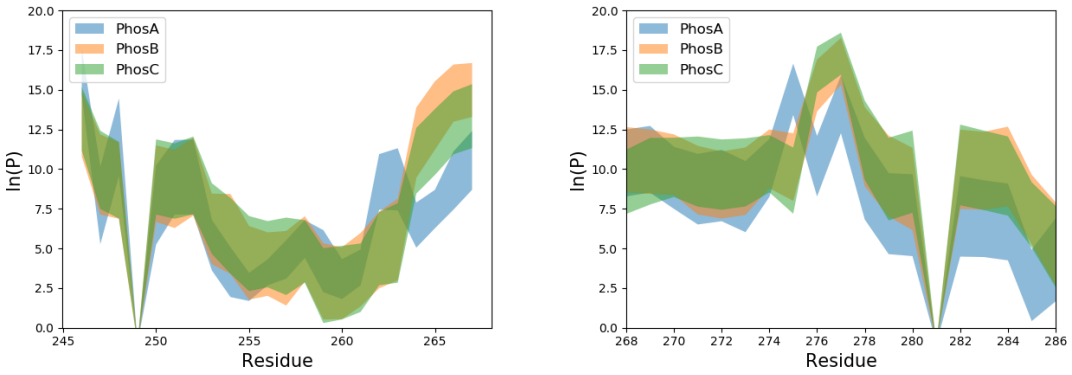

Figure 5.17: Patterns of protection factors of residues 246-286, covering the 280s loop, for states PhosA (blue), PhosB (orange) and PhosC (green). Confidence intervals are shown within one σ.

Chapter 6

Exploiting protection factors

Summary of the chapter. Once protection factors are known at single residue resolution, they can be exploited to further study hydrogen deuterium exchange. A back exchange correction is introduced to reproduce experimental isotopic envelopes. Also, a structural model connecting protection factors to the structure of a protein is presented and developed via the introduction of a dependence on the electric potential.

6.1 Back exchange

The Michaelis Menten model used to calculate the uptake of a peptide (Eq. 2.1) assumes that the deuteration process is irreversible. However, experiments show that back exchange is not negligible especially during the LC/MS step of the experimental workflow (Fig. 3.1) when the solution is quenched. In fact, the intrinsic exchange rates have been tabulated both for in-exchange and back-exchange by Bai et al. (1993): while in-exchange rates k_i are higher than back-exchange rates k_b at room temperature and physiological pH, they have the same order of magnitude at acidic pH and low temperature. As a consequence, back-exchange cannot be neglected.

The problem is tackled by the normalization in Eq. 3.2: the uptake of a peptide is constrained to be 1 for the maximally labeled envelope. This means that the mass $m_{100\%}$ is not the centroid of the fully deuterated envelope, but of the mostly deuterated envelope that has been detected. However, the fully deuterated envelope and the maximally la-

beled envelope do not coincide because of back exchange. In order to properly reproduce experimental envelopes starting from protection factors extracted by ExPfact, a back exchange correction must be introduced.

Finding datasets with isotopic envelopes is not trivial. Because of the dimensions of the files, research groups tend to show only raw data concerning centroids of the envelopes. Also, the format of the files containing information related to envelopes is dependent on the software used by the operator.

The moPrP dataset, provided by Moulick et al. (2015) and already analysed in section 5.1, contains isotopic envelopes at five time points for three different peptides. For a detailed description of the dataset, we address the reader to section 5.1.2. In particular, isotopic envelopes are available for peptides 1, 5 and 13 at time points 1 min, 1 hr, 24 hr; the fully protonated and fully deuterated samples are also given. The sequences and charge states of the peptides are shown in Table 6.1.

Peptide	Charge	Sequence
1	1	YMLGSA
5	3	YRYPNQVYYRPVDQ
13	2	MERVVEQM

Table 6.1: Peptides in the moPrP dataset (Fig. 5.1.2) for which envelopes are available.

We shall note here that the results obtained in this section are partial because of the small dataset provided and should be intended as the starting point for further research.

6.1.1 Isotopic envelope calculation

Fully protonated envelope

The calculation of the fully protonated isotopic envelope does not require the knowledge of protection factors. In fact, such envelope is nothing but the mass spectrum of the protonated peptide. To calculate it, the sequence of the peptide is sufficient: if the amino acids forming the peptide are know, than the monoisotopic mass can be calculated by adding up the masses of the elements forming the sequence. Such value is then convoluted with the natural abundance of elements in order to obtain the isotopic envelope of the

fully protonated peptide. Several tools enable the calculation of the fully protonated envelope. The reference software for this section is MS-Isotope[1].

The charge state of the peptide has to be known. In fact, the positive charge acquired during MS analysis by the peptide is given by an additional proton bound to the peptide that increases its mass. If the charge state is bigger, a bigger mass is detected. Isotopic envelopes are plotted as a distribution of m/z to discriminate different charge states.

The fully protonated envelopes for the peptides listed in Table 6.1 are shown in Fig. 6.1.

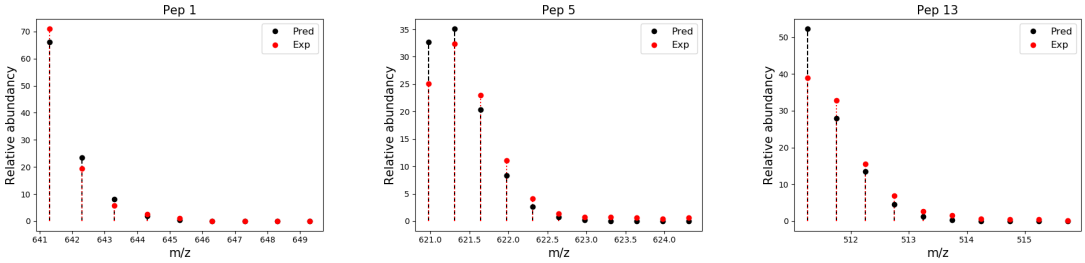

Figure 6.1: Fully protonated isotopic envelopes of peptides 1 (left), 5 (centre) and 13 (right) for the moPrP dataset (Fig. 5.1.2). Experimental envelopes (red) are compared with theoretical envelopes (black) predicted by MS-Isotope.

As Fig. 6.1 shows, experimental and predicted isotopic envelopes present minor differences, mainly due to two experimental artifacts:

- Saturation. In order to further separate peptides, liquid chromatography can be coupled with ion mobility: peptides are separated in their gas phase before reaching the mass spectrometer. This is helpful to identify with higher intensities peptides that were previously found with low probability. On the other hand, peptides that already had high intensities tend to reach the maximum intensity value detectable by the spectrometer, thus deforming the shape of the envelope.

- Carryover effect. It consists in the fact that some peptides can be identified from previous runs. This happens in all LC-MS systems: despite being washed between different runs, the columns of the LC system could still contain some peptides. When performing the next run, such peptides are completely back exchanged and tend to deform the lower peaks of the isotopic envelope.

[1] prospector.ucsf.edu/prospector/mshome.htm

The presence of these artifacts is checked by the operator: envelopes that are saturated or affected by the carryover effect are considered or not depending on a manual choice of the operator.

Fully deuterated envelope

The fully deuterated envelope represents the mass spectrum of a peptide after exchange occurred for every residue. Theoretically, it can be calculated by shifting the fully protonated envelope of N-1 m/z units, N being the length of the peptide. In agreement with the model in Eq. 2.1, this calculation assumes that back exchange is not possible.

Experimental and predicted isotopic envelopes of the fully deuterated samples are shown in Fig. 6.2 for the peptides available. The envelopes in Fig. 6.2 clearly show that back

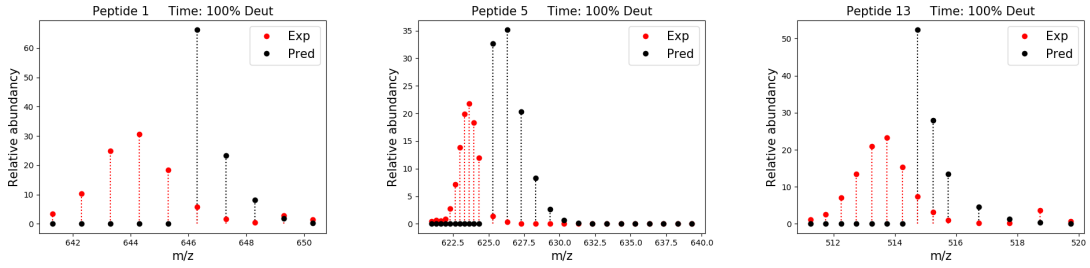

Figure 6.2: Fully deuterated isotopic envelopes for peptides 1 (left), 5 (centre) and 13 (right). Experimental envelopes (red) are compared with predicted ones (black), calculated by shifting the fully protonated envelopes in Fig.6.1.

exchange cannot be neglected. In fact, if back exchange is not considered as in the model in Eq. 2.1, the predicted fully deuterated samples always assume greater values of m/z with respect to experimental spectra. This is true even if intermediate time points are considered.

Evolution of the envelope

At intermediate times, the envelope may assume different shapes. The intensity of each peak depends on the exchange rates of the residues forming the peptide. Within a peptide formed by n residues, the probability that k have exchanged, with $k \leq n$ is expressed by

Eq. 3.1 that we report here for the sake of completeness:

$$\Pi(k,t) = \sum_{A\in\{1,\dots,n\}}^{|A|=k} \prod_{i\in A} D_i(t) \prod_{j\in\{1,\dots,n\}/A} (1 - D_j(t))$$

The time evolution of the isotopic envelope here described follows the assumption of the model in Eq. 2.1 that back exchange is not possible.

A python script was implemented to calculate the evolution of the isotopic envelope at given time points, starting from the fully protonated envelope calculated by MS-Isotope. In particular, the predicted isotopic envelopes are compared with the experimental ones for the peptides available in Fig. 6.3.

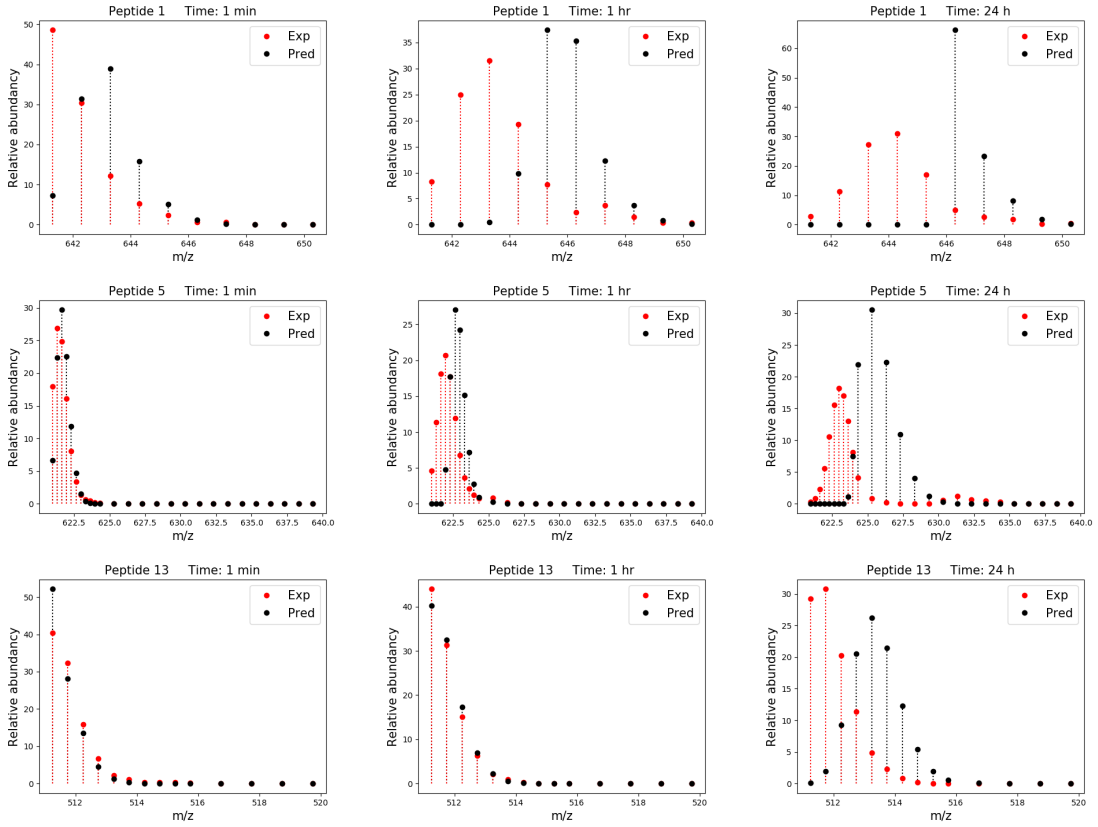

Figure 6.3: Evolution of the isotopic envelopes for peptides 1 (top), 5 (centre) and 13 (bottom) at times 1 min (left), 1 hr (centre) and 24 hr (right). Experimental (red) and predicted (black) envelopes are compared.

As expected from previous considerations (Fig. 6.2), the predicted isotopic envelopes lie at higher values of m/z since the model implemented does not consider back exchange.

To evaluate the time evolution of the isotopic envelope, the deuterium uptake of each residue - and thus protection factors at single residue resolution - must be known. The protection factors used for the calculation of the isotopic envelopes in Fig. 6.3 are the ones that have been estimated by ExPfact for the moPrP dataset (Fig. 5.8).

6.1.2 Back exchange correction

In order to properly reproduce experimental isotopic envelopes, back exchange must be taken into account. The aim of this section is to provide a procedure to obtain isotopic envelope reducing the back exchange problem to the proper setting of one only parameter.

The first assumption is that back exchange occurs only during LC analysis (i.e. under quenching conditions): we can thus assume that the envelopes predicted in Fig. 6.3 are the actual spectra before LC analysis.

The second assumption is that under quenching conditions in-exchange is completely stopped and only back exchange can occur. As a consequence, to reproduce experimental envelopes (red spectra in Fig. 6.3) starting from the predicted envelopes (black spectra in Fig. 6.3), the time evolution shown in Eq. 3.1 can still be exploited, but 1) the evolution must be towards lower m/z values and 2) the intrinsic exchange rates to be considered are the back exchange rates under quenching conditions (pH 2.4 and temperature 0 ^oC for the moPrP dataset).

The aim is to determine the effective back exchange time τ that minimizes the difference between experimental and predicted envelopes. A second python script was developed to perform this back exchange evolution starting from the envelope predicted at a specific experimental time. A number of back exchange effective times is set, ranging from 0 to 100 hours, and the back evolution is performed. For each effective time, a cost function is evaluated and the effective back exchange time associated to the lowest cost function is chosen.

Naming $\Pi_k^{pred}(\tau, t)$ and $\Pi_k^{exp}(t)$ the predicted and experimental envelopes evaluated at

$m/z = k$ respectively, the cost function reads:

$$C(\tau) = \frac{1}{K}\sqrt{\sum_{k=1}^{K}\left(\Pi_k^{pred}(\tau, t) - \Pi_k^{exp}(t)\right)^2} \tag{6.1}$$

K being the number of points at which the envelope is defined.

The results of the application of the back exchange evolution to isotopic envelopes predicted at experimental times (black spectra in Fig. 6.3) are shown in Fig. 6.4 for the peptides available and the goodness of predictions is quantified via R^2 statistic.

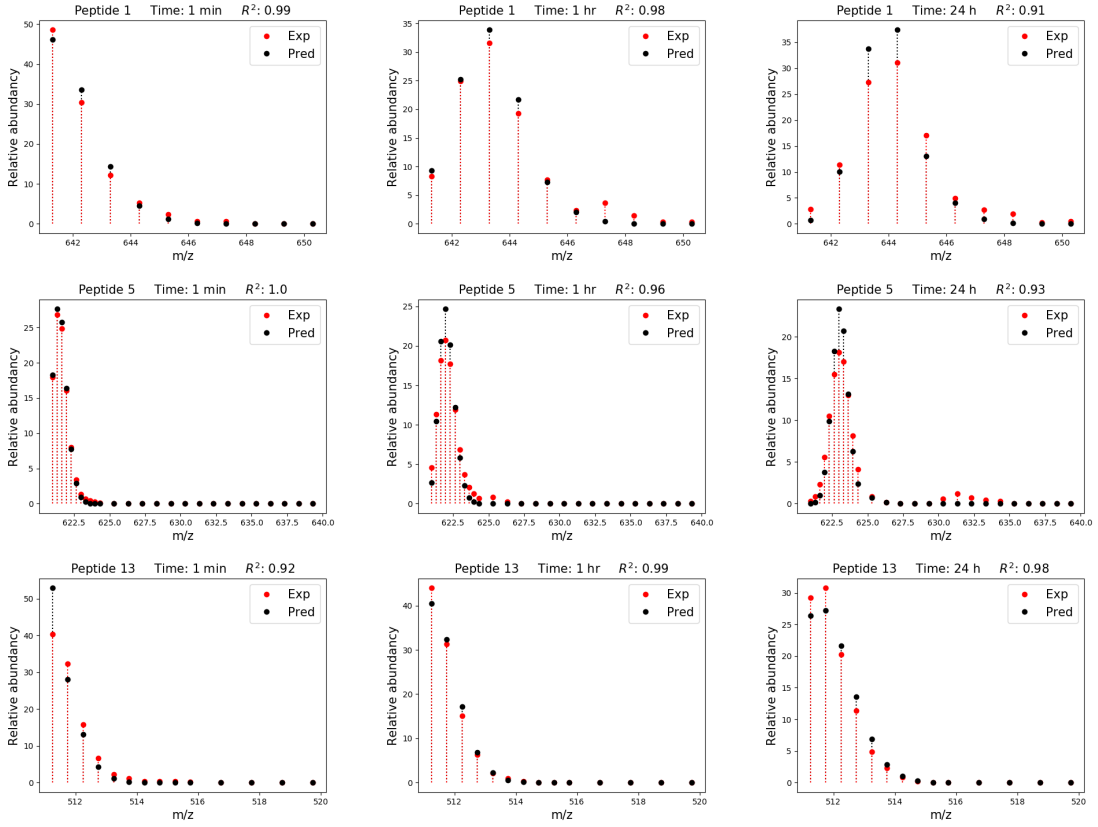

Figure 6.4: Evolution of the isotopic envelopes for peptides 1 (top), 5 (centre) and 13 (bottom) at times 1 min (left), 1 hr (centre) and 24 hr (right). Experimental (red) and predicted (black) envelopes are compared with the application of back exchange evolution. Goodness of fit is quantified through R^2 statistic.

Fig. 6.4 shows that it is possible to reproduce the experimental isotopic envelopes if

the protection factors of the peptide are known. The procedure to obtain them can be summarized as follows:

1. Calculate the fully protonated isotopic envelope of the peptide, e.g. using the software MS-Isotope.
2. Use the temporal evolution of the isotopic envelope (Eq. 3.1) to estimate the envelope at a specific time. Protection factors, in-exchange intrinsic rates k_i and pH of the deuterated buffer are needed.
3. Use the temporal evolution of the isotopic envelope (Eq. 3.1) to apply a back exchange correction to the the envelope at a specific time. Protection factors, back-exchange intrinsic rates and pH of the quenched solution are needed.

The effective back exchange times minimizing the difference between experimental and predicted envelopes (Eq. 6.1) are shown in Fig. 6.5 as a function of the time spent in the deuterated buffer. The different orders of magnitude covered by the effective back

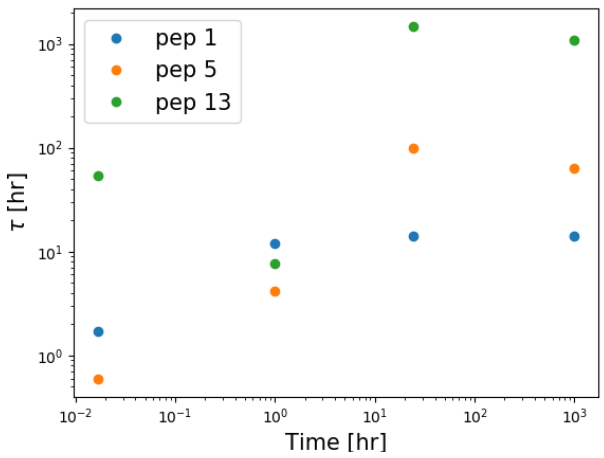

Figure 6.5: Back exchange effective time τ that minimize difference between experimental and predicted envelopes are shown as a function of in-exchange time.

exchange times evaluated at the same in-exchange time for different peptides suggest that the back exchange correction depends on the peptide considered.

Within the same peptide, the effective time τ tends to increase as the time spent in the deuterated buffer increases. This suggests that back exchange is not completely stopped in the deuterated buffer. Regarding this, it is worth noting that the moPrP dataset involves a deuterated buffer with pH 4.0: such an acidic condition could lead back exchange not to be negligible.

To conclude, the back exchange correction here introduced enables to reproduce the experimental envelopes starting from patterns of protection factors estimated by the ExPfact algorithm. Moreover, it lays the foundations for further research about back exchange: the usage of an extensive dataset could lead to the understanding of the main parameters influencing the phenomenon and solve a number of open questions: is back exchange completely stopped in the deuterated buffer? Is back exchange dependent on the analysed peptide? Can the effective back exchange time be related to experimental parameters like the percentage of deuterium in solution?

6.2 Structural model

One of the most ambitious goals of high throughput experimentation is to directly and rapidly fingerprint the structure of a protein. In HXD-MS this can be translated into the development of a structural model connecting protection factors to the structure of the sample. To take some steps towards this ultimate goal, the inverse (and simpler) problem shall be deepened: is it possible to estimate protection factors when the structure of a protein is available?

One attempt to connect the structure of a protein to protection factors was developed by Best and Vendruscolo (2006). In the present section, we describe such a structural model and we test its correlation with a dataset formed by 6 different proteins for which protection factors are known. Moreover, we introduce in the model a dependence on the electric potential of the protein and we show how this term improves the model.

6.2.1 Best model

The model developed by Best and Vendruscolo (2006) aims to connect the structure of a protein with its protection factors. The structure of the protein X is the set of Cartesian coordinates of each atom of the protein. The most common format to store such information is the PDB format (Berman et al. (2000)).

The exchange rate for one amino acid is given by

$$k_{ex} = \frac{k_{int}}{P}$$

where the intrinsic exchange rate encodes the dependence of the exchange rate on both the experimental conditions (pH and temperature) and the sequence of the protein (see section 2.1.3).

Assuming that the protection factor P of a residue only depends on the structure of the protein, Best and Vendruscolo (2006) stated that it can be written as a linear combination of the number of burial contacts N^c (i.e. number of connected heavy atoms) and the number of hydrogen bonds N^h. The protection factor for the residue i can be thus written as

$$\ln P_i(X) = \beta_c N_i^c(X) + \beta_h N_i^h(X) \tag{6.2}$$

where $\beta_c = 0.35$ and $\beta_h = 2.0$ are the weights of the number of heavy contacts and hydrogen bonds respectively. Such parameters were optimized by Best and Vendruscolo (2006) using a dataset formed by 7 proteins for which protection factors are known.

Note that in their paper Best and Vendruscolo (2006) suggest that the protection factor should be evaluated as an average value over a conformational space sampled during Molecular Dynamics simulations (Rapaport (2004)):

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i \tag{6.3}$$

where N is here the total number of frames constituting a trajectory. Some simulation softwares like CHARMM (Brooks et al. (2009)) can automatically implement this calculation. Here we focus only on the PDB structure.

Following the implementation in CHARMM, the number of heavy contacts and hydrogen bonds for a residue i can be calculated as follows:

$$\begin{aligned} N_i^c &= \sum_{j \in H_i} \frac{1}{1+\exp 5(r_{ij}-6.5)} \\ N_i^h &= \sum_{j \in O_i} \frac{1}{1+\exp 10(r_{ij}-2.4)} \end{aligned} \tag{6.4}$$

where H_i is the list of heavy atoms (i.e. all atoms but hydrogens) that are not part of residues i-1, i or i+1. Analogously, O_i is the list of oxygens that are not included in residues i-1, i or i+1. The value r_{ij} is the euclidean distance between the amide hydrogen of residue i and the j-th atom.

Results

Using the Best model (Eq. 6.2), protection factors are calculated for a dataset composed by 6 proteins, probed under different conditions, for which exchange rates are known:

- 1BRN: Barnase (Jane et al. (1993)).
- 1FRC: Horse heart ferrocytochromce c (Chevance et al. (2003)).
- 1MBC: Carbon-monoxy (Fe II)-myoglobin (Uzawa et al. (2008)).
- 2LN3: De novo designed protein IF3-like fold (Basak et al. (2019)).
- 3NPO: Bovine beta lactoglobulin (Forge et al. (2000)).
- 8PTI: Mutant of bovine pancreatic trypsin inhibitor (Key-Sun et al. (1993)).

For the sake of simplicity, we will refer to these proteins using their PDB code.

Protection factors predicted by the Best model in Eq. 6.2 with parameters $\beta_c = 0.35$ and $\beta_h = 2.00$ (i.e. the parameters found by Best and Vendruscolo (2006)), are compared with experimental protection factors in Fig. 6.6.

The Best model shows a poor correlation (0.08) between predicted and experimental protection factors for the considered dataset. This may be only due to the fact that we are only considering the PDB structure without averaging the protection factor over

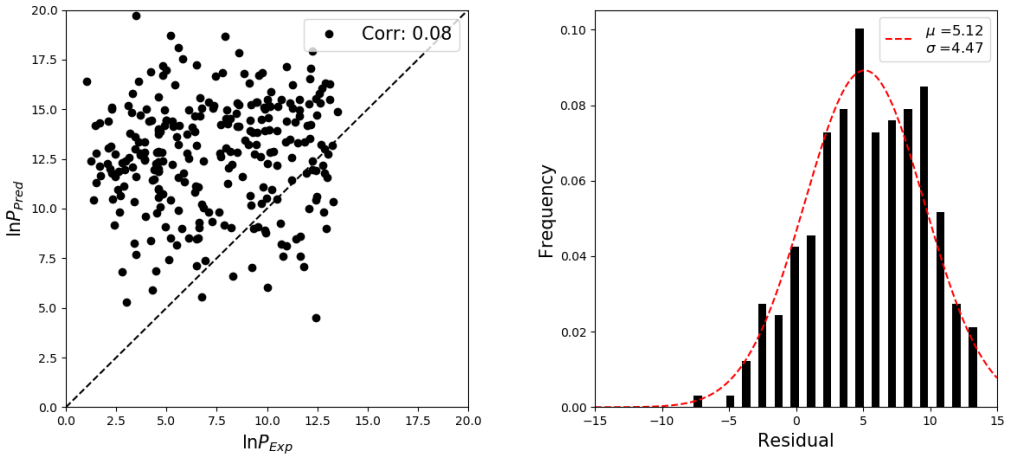

Figure 6.6: Left: correlaton between experimental and predicted protection factors using the Best model (Eq. 6.2); the black dashed line shows the line $y = x$. On the right, residuals are shown and fitted with a gaussian distribution (red dashed line). Parameters: $\beta_c = 0.35$, $\beta_h = 2.00$.

an ensemble of conformations, but Radou (2015) showed that performing Molecular Dynamics simulations only introduce minor changes in the predicted rates. Such a poor correlation is more likely to be linked to the fact that the optimal parameters β_c and β_h change if different datasets are implemented. In fact, as the gaussian fit of the residual plot shows, the parameters $\beta_c = 0.35$ and $\beta_h = 2.00$ optimized by Best and Vendruscolo (2006) introduce an overestimate of the predicted protection factor with respect to the experimental values (from the gaussian fit, $\mu = 5.12$).

To properly assess the quality of the Best model in estimating protection factors directly from the structure of the protein, we optimize the parameters β_c and β_h in Eq. 6.2 on the dataset here used. Since we are dealing with a linear model, an ordinary least squares regression implemented in python (Pedregosa et al. (2011)) can be used to estimate such coefficients. The weights that minimize the residual sum of squares reads:

$$\beta_c = 0.20 \qquad \beta_h = -3.08 \tag{6.5}$$

It is interesting to note that the coefficient associated to hydrogen bonds is negative ($\beta_h = -3.08$): hydrogen bonds lead the protection factor to be lower. The results for

the Best model with parameters optimized over the dataset are shown in Fig. 6.7.

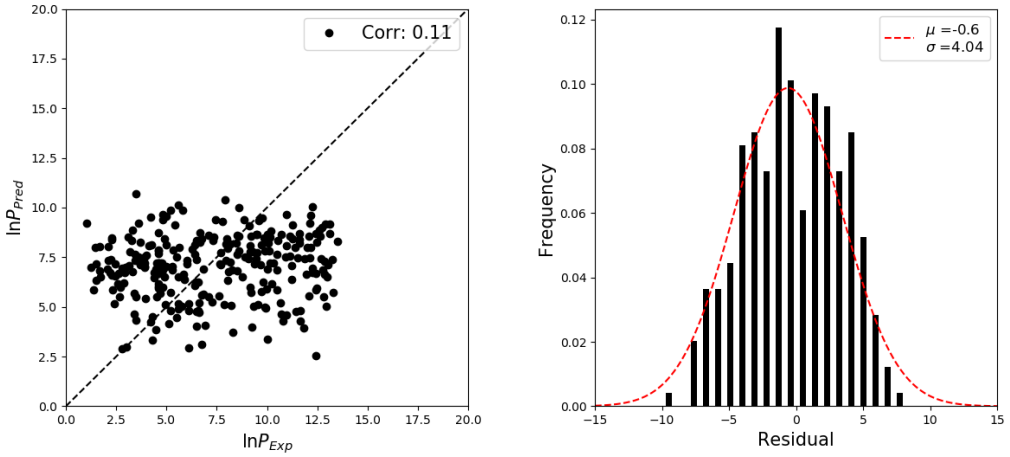

Figure 6.7: Left: correlation between experimental and predicted protection factors using the Best model (Eq. 6.2); the black dashed line shows the line $y = x$. On the right, residuals are shown and fitted with a gaussian distribution (red dashed line). Parameters: $\beta_c = 0.20$, $\beta_h = -3.08$.

The parameters β_c and β_h optimized over the dataset are able to reduce the overestimation of the predicted protection factors (from the gaussian fit of the residuals, $\mu = -0.6$). However, the correlation between predicted and experimental protection factors is still low (0.11), showing that the dependency of protection factors on burial contacts and hydrogen bonds is not sufficient to exhaustively reproduce experimental exchange rates.

6.2.2 Introducing potential dependence

Potential dependence

In order to improve the correlation between protection factors calculated by the Best model (Eq. 6.2) and experimental rates, we add a third variable in the linear model, namely the electrostatic potential U of the protein:

$$\ln P_i(X) = \beta_c N_i^c(X) + \beta_h N_i^h(X) + \beta_u U_i(X) \tag{6.6}$$

More specifically, $U_i(X)$ is the electrostatic potential of the configuration X evaluated at the amide hydrogen of the residue i.

The introduction of the dependence on the electrostatic potential comes from considerations regarding an article by Barnes et al. (2019). The analysed protein (PDB code: 6OBI) is the single α-helical domain of myosin-VI. Being an helix, it is quite symmetric and one could expect protection factors to show a periodic pattern. From HDX-NMR measurements, however, such periodicity is not detected and the electrostatic potential of the protein can be addressed as a responsible.

The introduction of such a variable arises two issues: how to calculate the potential and how to interpolate it at the amide hydrogen of each residue of the protein.

Calculating the potential

The electrostatic potential of a protein can be calculated using the software APBS (Dolinsky et al. (2004)), which solves the equations of continuum electrostatics for biomolecular assemblies. It requires accurate and complete structural data (like a PDB structure) as well as force fields parameters. APBS is coupled with PDB2PQR, a software enabling the automatic preparation of a PDB structure for continuum electrostatics calculations.

Electrostatic models generally belong to two families: explicit or implicit solvent models. Explicit solvent models treat each atom of the solvent with the same accuracy of atoms of the molecule under analysis, thus leading to a high level of detail that must be coupled with an extensive sampling of a conformational ensemble. On the other hand, implicit solvent models treat the solvent as an additional term in the force field: lower level of detail can be reached, but the need of an extensive sampling is eliminated. Implicit solvent methods are widely used to calculate the electrostatic potential of a molecule.

The calculation of polar solvation energy relies on the difference in charges from energies in vacuum and in solvent (Fogolari (2002)). In particular, APBS solves the Poisson Boltzmann equation, that is a non linear elliptic partial differential equation:

$$\nabla \left[\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})\right] = -4\pi\rho_f(\mathbf{r}) - 4\pi \sum_i c_i^\infty z_i q \exp \frac{z_i q \phi(\mathbf{r})}{kT} \lambda(\mathbf{r}) \tag{6.7}$$

where $\phi(\mathbf{r})$ is the electrostatic potential, $\epsilon(\mathbf{r})$ the dielectric function, $\rho_f(\mathbf{r})$ the molecular charge density, c_i^{∞} is the concentration of the ion i at infinite distance from the molecule, q is the proton charge and z_i is the valency of ion i; $\lambda(\mathbf{r})$ is a function describing the accessibility of ions at point $\mathbf{r}$.

The main advantage of Eq. 6.7 is that all the coefficients can be directly related to the structure of the protein. The main limitations of this model is the loss of accuracy when strongly charged systems or high salt concentrations are analysed.

Being a partial differential equation, the numerical solution of the Poisson Boltzmann equation is evaluated at the nodes of a three dimensional grid. In fact, the output of APBS is a file containing the value of the electrostatic potential calculated at the nodes of a grid containing the protein.

The electrostatic potential of the 6 proteins in the dataset here analysed can be visualised with a visualization software like PyMol (Schrödinger, LLC (2015)). The results are shown in Fig. 6.8, where the electric potential is shown on a surface surrounding the protein (Connolly surface).

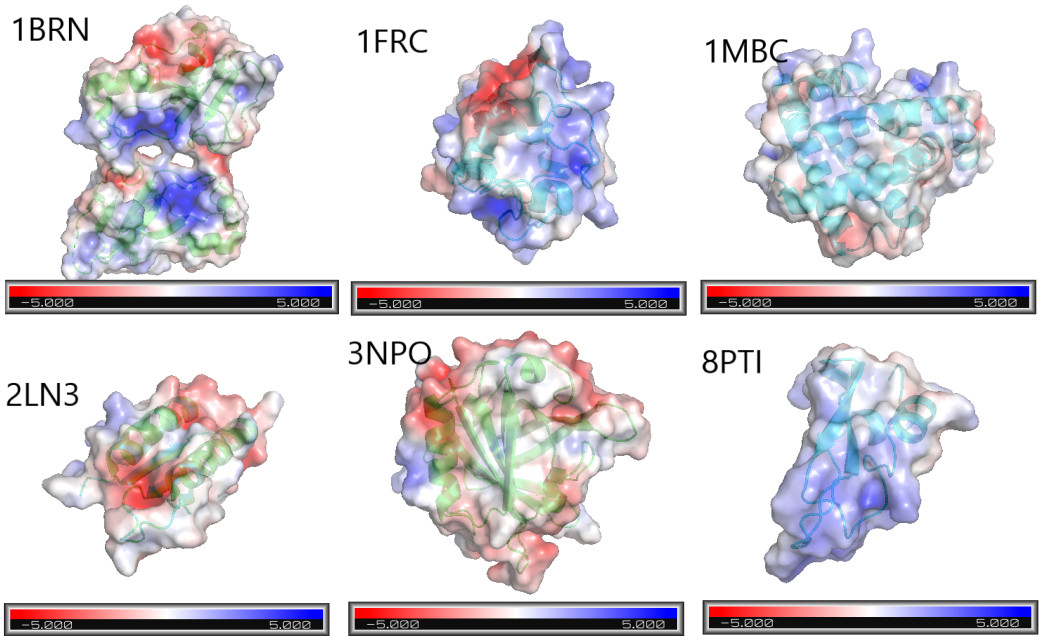

Figure 6.8: Electrostatic potential on the Connolly surface surrounding the proteins in the dataset. Potential values range from -5 kT/e (red) to +5 kT/e (blue).

Interpolation

APBS returns the values of the electrostatic potential evaluated at the nodes of a grid containing the protein under analysis. Since we want to evaluate the protection factor at single residue resolution (Eq. 6.6), we need to calculate the potential for each amino acid of the protein. In order to do so, we need to interpolate the grid in order to find a specific value of potential to be associated to the coordinates of the atoms of the protein.

The first question to be answered is: what is the point at which interpolation shall be performed? For instance, the potential could be evaluated for each atom of the protein and then an average value over all the atoms forming a residue could be used as the potential of the amino acid. Alternatively, the value of the potential interpolated at the alpha carbon of each residue could be used. However, since hydrogen deuterium exchange mainly involves the amide hydrogen of the residue, we opted for an interpolation at the coordinates of such atom.

Nearest Neighbour interpolation was used: the value of the electrostatic potential associated to a specific residue is thus the potential evaluated at the node of the grid nearest to the amide hydrogen of that specific residue.

The electrostatic profiles of the 6 proteins in the dataset are shown in Fig. 6.9.

An interesting remark regards the fact that the potential may assume both positive and negative values that will result in opposite influence on the final value of protection factor (Eq. 6.6). Moreover, it is worth noting that several periodicities may be found in some patterns. The protein 2LN3, for example, has two helices at residues 11-26 and 37-56, both showing a potential decreasing towards a minimum.

Results

After the calculation of the electrostatic potential of a protein and its interpolation at the amide hydrogen of each residue, we optimize the parameters β_c, β_h and β_u of the Best model with the introduction of the potential dependence (Eq. 6.6). As previously, an ordinary least squares can be performed since we are dealing with a linear model.

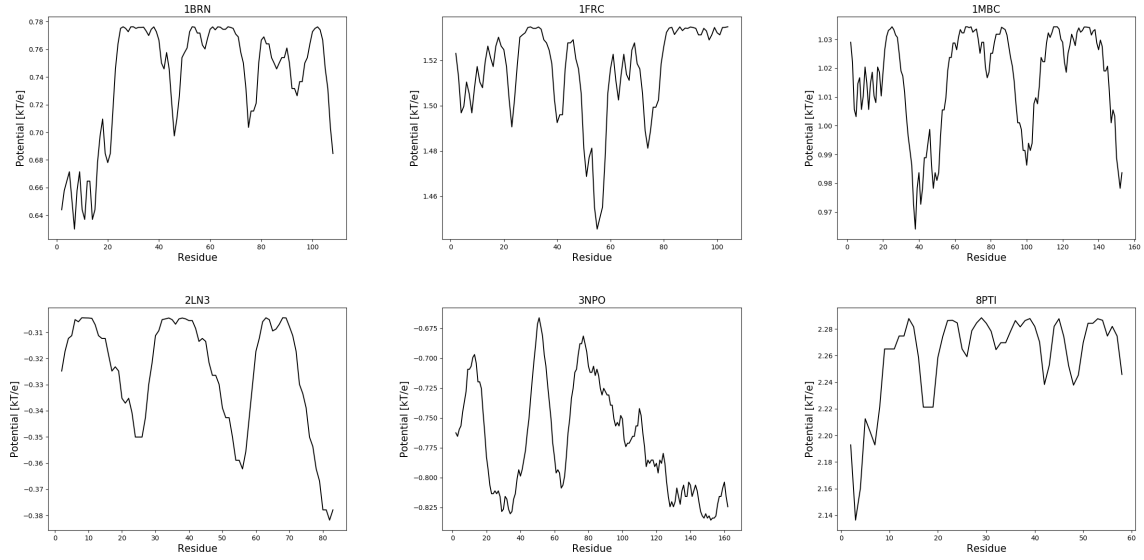

Figure 6.9: Potential profiles of the 6 proteins in the dataset. The electrostatic potential at each residue is evaluated through a Nearest Neighbour interpolation from the grid calculated with APBS (Fig. 6.8).

The optimization leads to the following values for the parameters:

$$\beta_c = 0.21 \qquad \beta_h = -1.33 \qquad \beta_u = -0.94 \tag{6.8}$$

As for the parameters of the Best model optimized over the dataset (Eq. 6.5), hydrogen bonds have a negative influence on the final value of protection factors. At the same time, the coefficient β_u is negative, meaning that positive values of potential have a negative influence on the protection factor and vice versa.

Protection factors estimated with the updated Best model (Eq. 6.6) are compared with experimental values in Fig. 6.10 and show the improvement of the model with respect to the *classic* Best model (Fig. 6.6 and Fig. 6.7).

The introduction of the dependence on the electrostatic potential in the structural model (Eq. 6.6) improves the model developed by Best and Vendruscolo (2006). In fact, correlation increases up to 0.66, six times higher than the value found for the *classic* Best model for the same dataset (Fig. 6.7). Moreover, the gaussian fit of the residuals shows that overestimation of predicted protection factors is further decreased.

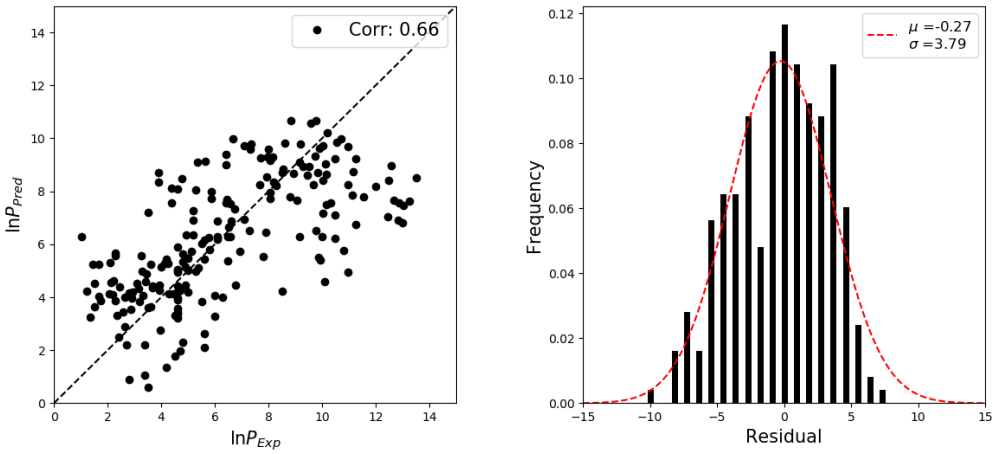

Figure 6.10: Left: correlation between experimental and predicted protection factors using the Best model with the introduction of a dependence on the electrostatic potential (Eq. 6.6); the black dashed line shows the line $y = x$. On the right, residuals are shown and fitted with a gaussian distribution (red dashed line). Parameters: $\beta_c = 0.21$, $\beta_h = -1.33$, $\beta_u = -0.94$

To summarize, the model developed by Best and Vendruscolo (2006) is not able to reproduce experimental protection factors for the dataset here analysed. The low correlation found using the parameters β_c and β_h defined in the original paper (Fig. 6.6) cannot be significantly increased even if such parameters are optimized using the dataset (Fig. 6.7). The introduction of the electrostatic potential of the protein evaluated at the amide hydrogen of the residue leads to an improvement of correlation (Fig. 6.10).

The final value of correlation (0.66) should be considered as an initial step towards further development of the model. Insights of the model could be obtained by considering protection factors averaged over an ensemble of structures calculated through Molecular Dynamics simulations, recovering the original idea of Best and Vendruscolo (2006). Moreover, other variables could be introduced: for instance, the flexibility of the protein could be considered.

Chapter 7

Conclusions

The purpose of this work was to provide computational methods to extract high resolution information from coarse data. In the context of high-throughput experimentation, HDX-MS has been established as a powerful technique to rapidly fingerprint both structural and dynamical properties of a protein.

The theoretical background of the phenomenon introduces several approximations that must be taken into consideration while analysing experimental data. On the other hand, the procedure to monitor the exchange of the amide hydrogens of a protein with deuterium contained in solution shall be accurately known since the presence of artifacts is not negligible. In fact, despite being highly automatized, HDX-MS experiments still rely on a manual correction of the isotopic envelopes performed by operators. Not only is such an operation time consuming, but the results of this procedure are hardly repeatable. Softwares discriminating saturated peptides and correcting the carryover effect in experimental isotopic envelopes would be appreciated by the HDX-MS community.

After introducing the state of the art in data analysis of HDX-MS experiments, the ExPfact algorithm initially developed by Skinner et al. (2019) was described in detail. First, it was applied to synthetic data to show the main issues arising from its usage and to proof that, if an ideal dataset is available, the true values of protection factors can be calculated. ExPfact was then applied to real world data. A first application probing the mouse prion protein enables a comparison of protection factors extracted by the algorithm with experimental values estimated by HDX-NMR, leading to the validation of the algorithm. Protection factors of glycogen phosphorylase in three different states were

then calculated by the algorithm, enabling the localization of conformational changes at single residue resolution.

If high-quality datasets are available, ExPfact is able to calculate protection factors at single residue resolution. Once protection factors are known, the evolution of isotopic envelopes can be calculated and compared with experimental spectra. A back exchange correction was proposed to recover experimental data. Such a correction reduces the problem of back exchange to the estimate of one only parameter, namely the effective back exchange time. The application of this correction to an extensive dataset could relate such a parameter with experimental variables. Finding such a relation would mean to characterize the effective back exchange time directly from experimental data. As a consequence, the envelopes could be reproduced starting from protection factors extracted by uptake curves. This would mean that the information encoded in the centroid of the envelope is sufficient to completely describe experimental data.

Protection factors calculated for several residues of a protein can also be used to improve a structural model connecting the structure of a protein to its exchange rates. A structural model developed by Best and Vendruscolo (2006) has been introduced and improved with the introduction of a dependence of the protection factor on the electrostatic potential of the protein. Such a dependence enables to characterize protection factors if the structure of the protein is known:

$$\text{Structure} \rightarrow \text{Protection Factor}$$

Despite improving correlation between predicted and experimental protection factors, the correlation of the structural model is still low to exhaustively explain the relation between structure and exchange rates. Since ExPfact enables the estimation of protection factors at single residue resolution, it may allow the creation of novel datasets that could be exploited to study other dependencies of the structural model. In addition, a complete characterization of protection factors starting from the structure of a protein would open pathways to address the inverse ambitious problem, i.e. to estimate the structure of a protein starting from the estimated protection factors:

$$\text{Protection Factor} \rightarrow \text{Structure}$$

ExPfact is constantly under development, mainly proceeding towards its computational

optimization and user-friendliness. The importance of the achievements of the algorithm is testified by the interest of pharmaceutical companies like GSK and Astra Zeneca in using ExPfact to find patterns of protection factors at single residue resolution for different purposes.

Bibliography

Adhikary, S., Deredge, D. J., Nagarajan, A., Forrest, L. R., Wintrode, P. L., and Singh, S. K. (2017). Conformational dynamics of a neurotransmitter:sodium symporter in a lipid bilayer. *Proceedings of the National Academy of Sciences*, 114(10):E1786–E1795.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(4096):223–230.

Apriola, S. (2001). Prion protein diversity and disease in the transmissible spongiform encephalopathies. In *Prion Proteins*, volume 57 of *Advances in Protein Chemistry*, pages 1 – 27. Academic Press.

Bai, Y., Milne, J. S., Mayne, L., and Englander, S. W. (1993). Primary structure effects on peptide group hydrogen exchange. *Proteins: Structure, Function, and Bioinformatics*, 17(1):75–86.

Balasubramaniam, D. and Komives, E. A. (2013). Hydrogen-exchange mass spectrometry for the study of intrinsic disorder in proteins. *BBA - Proteins and Proteomics*, 1834(6):1202 – 1209.

Barnes, C. A., Shen, Y., Ying, J., Takagi, Y., Torchia, D. A., Sellers, J. R., and Bax, A. (2019). Remarkable rigidity of the single α-helical domain of myosin-vi as revealed by nmr spectroscopy. *Journal of the American Chemical Society*, 141(22):9004–9017.

Basak, S., Nobrega, R. P., Tavella, D., Deveau, L. M., Koga, N., Tatsumi-Koga, R., Baker, D., Massi, F., and Matthews, C. R. (2019). Networks of electrostatic and hydrophobic interactions modulate the complex folding free energy surface of a designed $\beta\alpha$ protein. *Proceedings of the National Academy of Sciences*, 116(14):6806–6811.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

Best, R. B. and Vendruscolo, M. (2006). Structural interpretation of hydrogen exchange protection factors in proteins: Characterization of the native state fluctuations of ci2. *Structure*, 14(1):97 – 106.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Brooks, B. R., Brooks III, C. L., Mackerell Jr., A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. (2009). Charmm: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614.

Carson, N. (2020). Rise of the robots. *Chemistry – A European Journal*, 26(15):3194–3196.

Chalmers, M. J., Busby, S. A., Pascal, B. D., West, G. M., and Griffin, P. R. (2011). Differential hydrogen/deuterium exchange mass spectrometry analysis of protein-ligand interactions. *Expert review of proteomics*, 8(1):43 – 59.

Chevance, S., Le Rumeur, E., de Certaines, J. D., Simonneaux, G., and Bondon, A. (2003). 1h nmr structural characterization of the cytochrome c modifications in a micellar environment. *Biochemistry*, 42(51):15342–15351. PMID: 14690444.

Chung-Jung, T. and Ruth, N. (2014). A unified view of "how allostery works". *PLoS Computational Biology*, 10(2):e1003394.

Clarke, J. and Fersht, A. R. (1996). An evaluation of the use of hydrogen exchange at equilibrium to probe intermediates on the protein folding pathway. *Folding and design*, 1(4):243 – 254.

de Souza, N. and Picotti, P. (2020). Mass spectrometry analysis of the structural proteome. *Current Opinion in Structural Biology*, 60(Folding and Binding Proteins):57 – 65.

Dempsey, C. E. (2001). Hydrogen exchange in peptides and proteins using nmr spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*, (2):135.

Deng, B., Lento, C., and Wilson, D. J. (2016). Hydrogen deuterium exchange mass spectrometry in biopharmaceutical discovery and development - a review. *Analytica chimica acta*, 940:8 – 20.

Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., and Baker, N. A. (2004). Pdb2pqr: an automated pipeline for the setup of poisson–boltzmann electrostatics calculations. *Nucleic Acids Research*, 32(suppl2):W665–W667.

Englander, J. J., Del Mar, C., Li, W., Englander, S. W., Kim, J. S., Stranz, D. D., Hamuro, Y., and Woods, V. L. (2003). Protein structure change studied by hydrogen-deuterium exchange, functional labeling, and mass spectrometry. *Proceedings of National Academy of Sciences USA*, (12):7057.

Englander, S., Sosnick, T., Englander, J., and Mayne, L. (1996). Mechanisms and uses of hydrogen exchange. *Current Opinion in Structural Biology*, 6(1):18–23.

Ferraro, D. M., Lazo, N. D., and Robertson, A. D. (2004). Ex1 hydrogen exchange and protein folding. *Biochemistry*, 43(3):587–594. PMID: 14730962.

Fersht, A. (2017). *Structure and Mechanism in Protein Science*. World Scientific.

Finkelstein, A. V. and Ptitsyn, O. B. (2002). *Protein physics : a course of lectures.* Academic Press.

Fogolari, F. (2002). The poisson-boltzmann equation for biomolecular electrostatics: a tool for structural biology: A review. *Journal of Molecular Recognition*, (6):377.

Forge, V., Hoshino, M., Kuwata, K., Arai, M., Kuwajima, K., Batt, C. A., and Goto, Y. (2000). Is folding of β-lactoglobulin non-hierarchic? intermediate with native-like β-sheet and non-native α-helix. *Journal of Molecular Biology*, 296(4):1039 – 1051.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631.

Glasoe, P. and Long, F. (1960). Use of glass electrodes to measure acidities in deuterium oxide. *Journal of Physical Chemistry*, 64(1):188–190.

Green, M. L., Choi, C. L., Hattrick-Simpers, J. R., Joshi, A. M., Takeuchi, I., Barron, S. C., Campo, E., Chiang, T., Empedocles, S., Gregoire, J. M., Kusne, A. G., Martin, J., Mehta, A., Persson, K., Trautt, Z., Van Duren, J., and Zakutayev, A. (2017). Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Applied Physics Reviews*, 4(1):011105.

Gross, J. H. (2017). Electrospray ionization. In *Mass Spectrometry: A Textbook*, chapter 12, pages 721–778. Springer International Publishing.

Hamuro, Y. (2017). Determination of equine cytochrome c backbone amide hydrogen/deuterium exchange rates by mass spectrometry using a wider time window and isotope envelope. *Journal of The American Society for Mass Spectrometry: The official journal of The American Society for Mass Spectrometry*, 28(3):486.

Harris, M. J., Raghavan, D., and Borysik, A. J. (2019). Quantitative evaluation of native protein folds and assemblies by hydrogen deuterium exchange mass spectrometry (hdx-ms). *Journal of The American Society for Mass Spectrometry: The official journal of The American Society for Mass Spectrometry*, 30(1):58.

Harrison, R. A. and Engen, J. R. (2016). Conformational insight into multi-protein signaling assemblies by hydrogen–deuterium exchange mass spectrometry. In *Current opinion in structural biology*, page 187.

Holdren, J. P. (2011). Materials genome initiative for global competitiveness. *Report. White House Office of Science and Technology Policy.*

Hvidt, A. and Nielsen, S. O. (1966). Hydrogen exchange in proteins. In Anfinsen, C., Anson, M., Edsall, J. T., and Richards, F. M., editors, *Advances in Protein Chemistry*, volume 21, pages 287 – 386. Academic Press.

Ingalls, B. P. (2013). *Mathematical modeling in systems biology : an introduction.* MIT Press.

Jane, C., Andrea M., H., Mark, B., and Alan R., F. (1993). Local breathing and global unfolding in hydrogen exchange of barnase and its relationship to protein folding pathways. *Proceedings of the National Academy of Sciences of the United States of America*, 90(21):9837.

Johnson, D. T., Di Stefano, L. H., and Jones, L. M. (2019). Fast photochemical oxidation of proteins (fpop): A powerful mass spectrometry-based structural proteomics tool. *Journal of Biological Chemistry*, 294(32):11969 – 11979.

Kan, Z. Y., Mayne, L., Sevugan Chetty, P., and Englander, S. W. (2011). Exms: Data analysis for hx-ms experiments. *Journal of American Society for Mass Spectrometry*, (11):1906.

Kan, Z.-Y., Walters, B. T., Mayne, L., and Englander, S. W. (2013). Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis. *Proceedings of the National Academy of Sciences*, 110(41):16438–16443.

Kempa, E. E., Hollywood, K. A., Smith, C. A., and Barran, P. E. (2019). High throughput screening of complex biological samples with mass spectrometry – from bulk measurements to single cell analysis. *Analyst*, 144:872–891.

Key-Sun, K., Fuchs, J. A., and Woodward, C. K. (1993). Hydrogen exchange identifies native-state motional domains important in protein folding. *Biochemistry (Easton)*, 32(37):9600 – 9608.

Kish, M., Smith, V., Subramanian, S., Vollmer, F., Lethbridge, N., Cole, L., Bond, N. J., and Phillips, J. J. (2019). Allosteric regulation of glycogen phosphorylase solution phase structural dynamics at high spatial resolution. *bioRxiv*.

Krezel, A. . . . and Bal, W. . . . (2004). A formula for correlating pka values determined in d 2o and h2o. *Journal of Inorganic Biochemistry*, 98(1):161–166.

Krishna, M., Hoang, L., Lin, Y., and Englander, S. (2004). Hydrogen exchange methods to study protein folding. *Methods*, 34(1):51–64.

Krivov, S. (2013). On reaction coordinate optimality. *Journal of Chemical Theory and Computation*, 9(1):135–146.

Kwart, H., Kuhn, L., and Bannister, E. (1954). The rates and equilibria of hydrogen-deuterium exchange in hydroxylic compounds. *Journal of the American Chemical Society*, 76(23):5998–6001.

Lackie, J. (2019). Glycogen phosphorylase. *A Dictionary of Biomedicine*.

Lam, T. T., Lanman, J. K., Emmett, M. R., Hendrickson, C. L., Marshall, A. G., and Prevelige, P. E. (2002). Mapping of protein :protein contact surfaces by hydrogen/deuterium exchange, followed by on-line high-performance liquid chromatography-electrospray ionization fourier-transform ion-cyclotron-resonance mass analysis. *Journal of chromatography*, (1):85.

Lapidus, L. (2017). Protein unfolding mechanisms and their effects on folding experiments [version 1; peer review: 2 approved]. *F1000Research*, 6(1723).

Linderstrøm-Lang, K. (1955). The ph-dependence of the deuterium exchange of insulin. *BBA - Biochimica et Biophysica Acta*, 18:308.

Liotta, S. and Mer, V. K. L. (1937). Hydrogen—deuterium exchange in acetate solution. *Journal of the American Chemical Society*, 59(5):946–946.

Lisal, J., Lam, T. T., Kainov, D. E., Emmett, M. R., Marshall, A. G., and Tuma, R. (2005). Functional visualization of viral molecular motor by hydrogen-deuterium exchange reveals transient states. *Nature Structural and Molecular Biology*, (5):460.

Liu, Y., Hu, Z., Suo, Z., Hu, L., Feng, L., Gong, X., Liu, Y., and Zhang, J. (2019). High-throughput experiments facilitate materials innovation: A review. *Science China Technological Sciences*, 62(4).

Lísal, J., Kainov, D. E., Lam, T. T., Emmett, M. R., Wei, H., Gottlieb, P., Marshall, A. G., and Tuma, R. (2006). Interaction of packaging motor with the polymerase complex of dsrna bacteriophage. *Virology - New York*, (1):73.

Malhotra, P., Jethva, P. N., and Udgaonkar, J. B. (2017). Chemical denaturants smoothen ruggedness on the free energy landscape of protein folding. *Biochemistry*, 56(31):4053–4063. PMID: 28714672.

Masson, G., Burke, J., N.G., A., and al. (2019). Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (hdx-ms) experiments. *Nat Methods*, 16:595–602.

Mathieu, C., Dupret, J.-M., and Rodrigues Lima, F. (2017). The structure of brain glycogen phosphorylase-from allosteric regulation mechanisms to clinical perspectives. *FEBS Journal*, (4):546.

Mennen, S. M., Alhambra, C., Allen, C. L., Barberis, M., Berritt, S., Brandt, T. A., Campbell, A. D., Castañón, J., Cherney, A. H., Christensen, M., Damon, D. B., Eugenio de Diego, J., García-Cerrada, S., García-Losada, P., Haro, R., Janey, J., Leitch, D. C., Li, L., Liu, F., Lobben, P. C., MacMillan, D. W. C., Magano, J., McInturff, E., Monfette, S., Post, R. J., Schultz, D., Sitter, B. J., Stevens, J. M., Strambeanu, I. I., Twilton, J., Wang, K., and Zajac, M. A. (2019). The evolution of high-throughput experimentation in pharmaceutical development and perspectives on the future. *Organic Process Research & Development*, 23(6):1213–1242.

Molday, R. S., Englander, S. W., and Kallen, R. G. (1972). Primary structure effects on peptide group hydrogen exchange. *Biochemistry*, 11(2):150 – 158.

Moulick, R., Das, R., and Udgaonkar, J. (2015). Partially unfolded forms of the prion protein populated under misfolding-promoting conditions: characterization by hydrogen exchange mass spectrometry and nmr. *The Journal of biological chemistry*, 290.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Radou, G. (2015). Helicase functional dynamics from low-resolution experimental data and simulation.

Ramos, F. (2013). *Liquid Chromatography : Principles, Technology and Applications.* Nova Science Publishers, Inc.

Rapaport, D. C. (2004). *The Art of molecular dynamics simulation.* Cambridge University press.

Robert, Z., Attila, S., and Biman, B. (1992). Levinthal's paradox. *Proceedings of the National Academy of Sciences of the United States of America*, 89(1):20.

Schrödinger, LLC (2015). The PyMOL molecular graphics system, version 1.8.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1):289.

Skinner, S. P., Radou, G., Tuma, R., Houwing-Duistermaat, J. J., and Paci, E. (2019). Estimating constraints for protection factors from hdx-ms data. *Biophysical Journal*, 116(7):1194 – 1203.

Uzawa, T., Nishimura, C., Akiyama, S., Ishimori, K., Takahashi, S., Dyson, H. J., and Wright, P. E. (2008). Hierarchical folding mechanism of apomyoglobin revealed by ultra-fast h/d exchange coupled with 2d nmr. *Proceedings of the National Academy of Sciences*, 105(37):13859–13864.

Vendruscolo, M., Paci, E., Dobson, C. M., and Karplus, M. (2003). Rare fluctuations of native proteins sampled by equilibrium hydrogen exchange. *Journal of the American Chemical Society*, (51):15686.

Wagner, G. and Wüthrich, K. (1979). Structural interpretation of the amide proton exchange in the basic pancreatic trypsin inhibitor and related proteins. *Journal of Molecular Biology*, 134(1):75 – 94.

Walters, B. T., Riccicuti, A., Mayne, L., and Englander, W. S. (2012). Minimizing back exchange in the hydrogen exchange-mass spectrometry experiment. *Journal of the American Society for Mass Spectrometry*, 23(12):2132 – 2139.

Zhang, Z. and Smith, D. (1993). Determination of amide hydrogen exchange by mass spectrometry: A new tool for protein structure elucidation. *Protein Science*, 2(4):522–531.

Zhou, J., Yang, L., DeColli, A., Freel Meyers, C., Nemeria, N. S., and Jordan, F. (2017). Conformational dynamics of 1-deoxy-d-xylulose 5-phosphate synthase on ligand binding revealed by h/d exchange ms. *Proceedings of the National Academy of Sciences*.