MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **70** (2013) 801-827

ISSN 0340 - 6253

Kinetics of H/D Exchange by Mass Spectrometry. I. Analysis of a Single Exchange Profile

D. Babić, S. Kazazić, D. M. Smith

Institute "Ruđer Bošković", Bijenička 54, HR-10002 Zagreb, P.O.B. 180, Croatia

(dbabic@irb.hr, kazazic@irb.hr, dsmith@irb.hr)

(Received April 15, 2013)

Abstract

An exchange profile is a sequence of fragment fractions with 0, 1, ..., *n* deuterons introduced by hydrogen/deuterium exchange into a given molecular fragment. Derivation of the site-exchange probabilities p_1, p_2, ..., p_n, responsible for the experimentally observed exchange profile, is analyzed. A relation is established between p_1, p_2, ..., p_n and the zeros of the polynomial whose coefficients are equal to the fractions in the exchange profile. When the zeros are complex, the probabilities are usually determined by the least squares fitting or by the maximum likelihood estimation. It is proved that in such cases the calculated probabilities can not be all distinct. Analytical expressions for the gradient and the Hessian of a sum of squared deviations between theoretical and experimental values are given. A case with only two exchanging sites (n=2) is examined in detail. Numerical simulations were performed to estimate the incidence of complex polynomial zeros for n=3-10 with varied number of detected fragment ions N=1000-1000000.

1. Introduction

Knowledge of protein's three dimensional structure is of utmost importance for understanding their biological function. Protein structural studies usually involve application of several complementary methods to obtain insight on the protein 3D structure/dynamics/function relationship.[1-6] Various methods have been developed to determine the protein structure in a solution or to find out its features. One of the promising and already much applied techniques is hydrogen/deuterium exchange (HDX) followed by mass spectrometry.[7] The specific goal within this approach is determination of the kinetic parameters for protein sites at which HDX takes place. Under controlled pH and temperature, susceptibility of amide hydrogen to exchange, quantitatively expressed by the kinetic constant, reflects the protein structure through steric and inductive effects of neighboring amino acid side chains, solvent accessibility and strength of inductive bonding in secondary structure elements. Knowing the distribution of exchange susceptibilities along the protein chain enables one to conclude about protein conformational dynamics, allosteric effects, ligand binding and aggregation.

A comprehensive account of the technique can be found in literature,[8-10] and here the procedure will be only briefly outlined. H/D exchange takes place at the amide bonds which are uniformly distributed along the protein chain and thus represent a good structural probe. H/D exchange is initiated by dissolving the protein in D_2O. After specific period of time the exchange is quenched by decreasing pH to 2.5 and lowering the temperature close to 0°C. The protein is then submitted to enzymatic digestion (by adding pepsine or other suitable enzyme) in order to break it into smaller fragments. This is necessary to localize HDX data, that is, to measure extent of exchange for a particular region of protein structure. Since the fraction of exchanged hydrogens is determined for the peptide fragment as a whole, having smaller fragments increases the resolution by enabling monitoring of the exchange process at smaller protein pieces. The fragments are identified by combination of liquid chromatography and mass spectrometry. By repeating the procedure for variable time periods of deuteration, one obtains insight into temporal dynamics and gets the data from which the site exchange constants can be determined. Normally, a simple kinetic model for unimolecular process, with exponential time dependence, is used to describe time variation of the exchange extent.

A peptide fragment produced by enzyme digestion shows up in the mass spectrum as an *isotopic profile* – represented by a sequence of signals at regularly spaced *m*/*z* values. The signals arise from fragments with different isotopic compositions – all representing the same

chemical species. In the present context, the isotopic profile is a consequence of a natural isotopic variation and of the experimentally induced H/D exchange. The exchange extent for a given fragment is commonly expressed as the average number of hydrogens exchanged by deuterons. It is simply related to the centroid shift of the observed profile from a pure natural isotopic distribution.[11] The time dependence of the average number of exchanged hydrogens is then fit to a kinetic model involving as many rate constants as there were hydrogens available for exchange [12] or the chosen number of rate constants that represent discernible groups (e.g. slow, medium and fast exchanging sites). [13,14] Minimization of the sum of squared distances between the calculated and the experimental values now appears to be a common fitting objective,[12] but for this particular purpose it was frequently combined with maximization of the information entropy.[13,15,16]

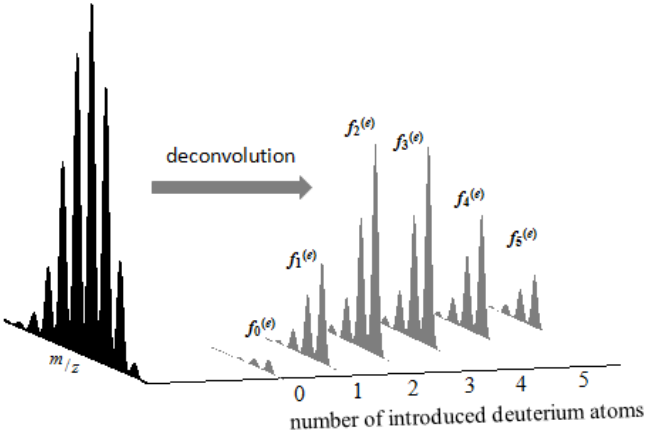

Figure 1. Schematic illustration of deconvolution of an isotopic profile from H/D exchange (in black) into theoretical isotopic profiles (in grey) arising from natural isotopic variation only. The theoretical profiles are equal in shape, shifted along the *m*/*z* axis according to the number of introduced deuterium atoms and scaled by the fractions $f_0^{(e)} - f_n^{(e)}$.

The isotopic profile is a convolution of identical natural isotopic profiles shifted by *m*/*z* values that correspond to the number of artificially deuterated sites (see Figure 1). In other words, the experimental isotopic profile is a linear combination of the natural profiles shifted by *i*/*z*, with *i* going from 0 to *n* = the total number of exchanging hydrogens. The coefficients of the linear combination represent fragment fractions (or percentages) with 0 to *n* exchanged hydrogens. They are easy to obtain although extraction of reasonable values from the data with

experimental error can be nontrivial.[16-18] Collection of the fragment fractions for a given peptide fragment will be termed *exchange profile*. Certainly, the exchange profile contains more information about H/D exchange than the average number of exchanged hydrogens (the last one can be easily calculated from the exchange profile). The theoretical model is a bit more complicated, but more specific data should provide more reliable and more precise results. The reports with a comprehensive treatment of experimental data, involving deconvolution of isotopic profile are relatively rare, [19-21] and the details of the fitting procedure, like the expressions for a gradient or a Hessian of the objective function were not yet presented.

Here we show how a given exchange profile can be analyzed to obtain exchange probabilities at the individual sites. In normal practice, several exchange profiles – recorded after different times of deuteration, are used for fitting the site-exchange rate constants. We show that the kinetic parameters can, at least in principle, be derived from a single exchange profile. Clearly, with more exchange profiles, the results should be more accurate and the theoretical model better verified. Still, the analysis of a single exchange profile can be applied to parts of experimental data for evaluation of their consistency and for recognition of outliers, or to obtain better initial guess for fitting to all the data at once. It will be shown that the results produced by the least squares fitting and by the maximum likelihood estimation exhibit inevitable degeneracy with at least two equal values among the fitted parameters. Numerical results show that the number of equal values is commonly greater than two. This finding may help to explain the results when collection of exchange profiles is fitted. Furthermore, we give analytical expressions for gradient and Hessian of the sum of squared deviations which are needed for proper fitting. These expressions are simple to extend for fitting to multiple exchange profiles.

2. Mathematical background

The object of the present analysis is a single exchange profile, that is, a set of values f_0, f_1, ... f_n representing fractions with 0, 1, ..., n deuterated sites due to the H/D exchange. The experimental values will be denoted by $f_m^{(e)}$ (m=0-n). The set of fractions for a given exchange profile will be denoted as a vector $\boldsymbol{f}$ ($\boldsymbol{f}^{(e)}$ for experimental values).

The exchange probabilities at sites i = 1, ..., n, are denoted by p_i, and the vector $\boldsymbol{p}$ denotes all of them at once. In kinetic modelling the probabilities are a function of time, usually assumed

to be: $p_i = exp(-k_i t)$. Since we are limited here to a single exchange profile, the time dependence is omitted from our consideration and all the results are given in terms of the site probabilities p_i.

Equations (1) represent the basic relations between the fractions f_i and the site exchange probabilities p_i under a common assumption that the exchanges at different sites are random and independent.

$$\left.\begin{aligned} f_0 &= (1-p_1)\cdot(1-p_2)\cdot\ldots\cdot(1-p_n) = \prod_{i=1}^{n}(1-p_i) \\ f_1 &= p_1\cdot(1-p_2)\cdot\ldots\cdot(1-p_n) + (1-p_1)\cdot p_2\cdot\ldots\cdot(1-p_n) + \ldots \\ &\quad \ldots + (1-p_1)\cdot(1-p_2)\cdot\ldots\cdot p_n = \sum_{i=1}^{n} p_i \prod_{j\neq i}(1-p_j) \\ &\ldots \\ f_n &= p_1\cdot p_2\cdot\ldots\cdot p_n = \prod_{i=1}^{n} p_i \end{aligned}\right\} \quad (1)$$

Our goal is to determine the probabilities p_i that best reproduce the experimental data $\boldsymbol{f}^{(e)}$. Basically, there are two approaches to this end: (i) the least squares fitting (LSF) and (ii) the maximum likelihood estimation (MLE). In LSF one seeks $\boldsymbol{p}$ which gives the least sum of squared deviations S between experimental and calculated values:

$$S = \sum_{m=0}^{n}\left[f_m^{(e)} - f_m\right]^2 \quad (2)$$

with f_m, m=0-n, given by (1). The conditions under which LSF and MLE approaches are sensible to apply, can be found in the literature.[22]

An inevitable experimental error is introduced by a finite number of molecular fragments that enter the detector. The isotopic composition of the fragments varies according to a multinomial distribution.[23] The observed fractions in the isotopic profile and those derived from them as the exchange profile, represent only a sample from this distribution. In the MLE approach one looks for the set of probabilities which yield the greatest likelihood of the observed set. A method for finding the site exchange constants that correspond to maximum likelihood of a set of isotopic profiles obtained for different deuteration times has been already described.[21] Here we establish specific properties of the MLE for a single exchange profile. For a given set of probabilities $\boldsymbol{p}$, the probability $P(N_0,N_1,...,N_n)$ that among the total

of $N=N_0+N_1+...+N_n$ detected fragments of a given peptide, N_m of them are with m exchanged hydrogens, is:

$$P(N_0, N_1,..., N_n) = \frac{N!}{N_0!N_1!...\cdot N_n!}(f_0)^{N_0}\cdot(f_1)^{N_1}\cdot\ldots\cdot(f_n)^{N_n} \qquad (3)$$

with f_m, m=0-n, given by (1). Maximization of $P(N_0,N_1,...,N_n)$ will be considered in Section 6. To establish relation between (3) and experimental data, we notice that (i) N_m, m=0-n, are related to the experimental values by $N_m=f_m^{(e)}\cdot N$, and (ii) $\boldsymbol{p}$ where $P(N_0,N_1,...,N_n)$ has the maximum for a given set $\boldsymbol{f}^{(e)}$ does not depend on N (which is normally unknown).

3. Exact derivation of the site exchange probabilities p_i

In principle, for a given data set $f_m^{(e)}$, m=0-n, there is a unique set of probabilities p_i that exactly reproduce the experimental data. To make it clear, we rewrite the eqns. (1) in terms of elementary symmetric polynomials, $s_m(n)$, which are defined by:

$$\left.\begin{aligned} & s_0(\boldsymbol{n}) = 1 \\ & s_1(\boldsymbol{n}) = p_1 + p_2 + \ldots + p_n = \sum_{i=1}^{n} p_i \\ & s_2(\boldsymbol{n}) = p_1p_2 + p_1p_3 + \ldots + p_{n-1}p_n = \sum_{i>j}^{n} p_i p_j \\ & \vdots \\ & s_n(\boldsymbol{n}) = p_1\cdot p_2\cdot\ldots\cdot p_n = \prod_{i=1}^{n} p_i \end{aligned}\right\} \qquad (4)$$

The dependence on n is explicitly designated in $s_m(\boldsymbol{n})$ in order to distinguish them from similar symbols that will be introduced later. After doing the multiplications denoted in eqns. (1) and recognizing the terms with equal order m in p_i as $s_m(\boldsymbol{n})$, the equations for f_m can be rewritten in a simple form:

$$f_m = \sum_{i=m}^{n}(-1)^{m-i}\binom{m}{i}s_i(\boldsymbol{n}) \qquad , \qquad \text{: a binomial coefficient} \qquad (5)$$

The above expression is more convenient for further manipulation. The fractions f_m and the elementary symmetric polynomials $s_m(\boldsymbol{n})$ can be easily interconverted by using the eqns. (5) and (6):

$$s_m(\boldsymbol{n}) = \sum_{i=m}^{n} \binom{i}{m} f_i \tag{6}$$

The eqn. (6) is obtained by starting with the eqn. (5) for m=n, and subsequently solving the eqns. (5) with m=n-1, n-2, ..., 1, for $s_{n-1}(\boldsymbol{n})$, $s_{n-2}(\boldsymbol{n})$, ..., $s_1(\boldsymbol{n})$, respectively. By the elementary fact from algebra, the roots p_1, ..., p_n of the polynomial $F(p)$:

$$F(p) = s_0(\boldsymbol{n})p^n - s_1(\boldsymbol{n})p^{n-1} + s_2(\boldsymbol{n})p^{n-2} - \ldots + (-1)^{n-1} s_{n-1}(\boldsymbol{n})p + (-1)^n s_n(\boldsymbol{n}) \tag{7}$$

satisfy the equations (4). Therefore, one may use (6) to transform $f_m^{(e)}$ to $s_m(\boldsymbol{n})$, and by finding the roots of $F(p)$, one obtains p_i that exactly reproduce the experimental set $f_m^{(e)}$. However, there is a pitfall – the roots p_i are not always acceptable as probabilities, as they are not always real numbers. It frequently happens that the roots of $F(p)$ involve complex numbers, which although satisfying (1), can not be interpreted as probabilities. This is a consequence of experimental error, unavoidable at least due to a finite number of detected molecules.

It turns out that the transformation from $f_m^{(e)}$ to $s_m(n)$ can be avoided by introducing a new variable q defined by:

$$q = \frac{p}{1-p}, \qquad p = \frac{q}{1+q} \tag{8}$$

Substitution of p_i with q_i gives from (1):

$$\left.\begin{aligned}
f_0 &= \frac{1}{1+q_1}\cdot\frac{1}{1+q_2}\cdot\ldots\cdot\frac{1}{1+q_n} = \frac{1}{\prod_{i=1}^{n}(1+q_i)} \equiv f_0 t_0(\boldsymbol{n}) \\
f_1 &= \frac{q_1}{1+q_1}\cdot\frac{1}{1+q_2}\cdot\ldots\cdot\frac{1}{1+q_n} + \frac{1}{1+q_1}\cdot\frac{q_2}{1+q_2}\cdot\ldots\cdot\frac{1}{1+q_n} + \ldots \\
&\quad \ldots + \frac{1}{1+q_1}\cdot\frac{1}{1+q_2}\cdot\ldots\cdot\frac{q_n}{1+q_n} = \frac{\sum_{i=1}^{n} q_i}{\prod_{i=1}^{n}(1+q_i)} = f_0 \sum_{i=1}^{n} q_i \equiv f_0 t_1(\boldsymbol{n}) \\
&\vdots
\end{aligned}\right\} \tag{9}$$

$$\vdots$$

$$f_n = \frac{q_1}{1+q_1}\cdot\frac{q_2}{1+q_2}\cdot\ldots\cdot\frac{q_n}{1+q_n} = \frac{\prod_{i=1}^{n} q_i}{\prod_{i=1}^{n}(1+q_i)} = f_0\prod_{i=1}^{n} q_i \equiv f_0 t_n(\boldsymbol{n})$$

The expressions on the right side are written as products of the common factor f_0 and $t_m(\boldsymbol{n})$, m=0-n, which denote elementary symmetric polynomials in variables q_i, i=1-n (in parallel to $s_m(\boldsymbol{n})$ which denote elementary symmetric polynomials in p_i). Consider the polynomial

$$G(q) = f_0 q^n - f_1 q^{n-1} + f_2 q^{n-2} - \ldots + (-1)^{n-1} f_{n-1} q + (-1)^n f_n \qquad (10)$$

The roots of $G(q)$ are equal to q_i since dividing of $G(q)$ with f_0 converts the coefficients f_m into $t_m(\boldsymbol{n})$ and the division does not change the roots of $G(q)$. Thus to determine the probabilities p_i, one can plug $f_m^{(e)}$ directly into (10) and search for the roots of $G(q)$. Even the alternation of the coefficients' signs can be ignored since leaving them all positive, only changes the roots q_i into $-q_i$. The roots q_i can be transformed to p_i by using (8) or checked for complex values directly if this only is needed.

Since $f_m^{(e)}$ are normally positive, the coefficients of $G(q)$ strictly alternate in sign. By the Decartes rule,[24] $G(q)$ has no negative real roots. It is clear from (8) that real positive q_i correspond to $p_i \in [0,1)$ (p_i=1 corresponds to q_i=+∞). Hence, apart from complex values, the roots of $F(p)$ cannot be nonphysical in any other way, like $p_i < 0$ or $p_i > 1$. This is true under the condition that all $f_m^{(e)}$ are positive. Due to experimental error, various situations can occur in practice. Negative values would be an obvious error that should be automatically corrected by replacing with zero since there is no sense to search for probabilities that reproduce nonphysical conditions. It may frequently happen that some leading or ending $f_m^{(e)}$ show up as zeros due to insufficient sensitivity. How these cases should be best treated, we plan to present in the next sequel [25] in which different approaches for fitting to a single exchange profile will be examined. Here we take that all $f_m^{(e)}$ are positive or that small positive value can be assigned to those that turn up as zeros, just in order to ensure alternation of the coefficients in (10) and applicability of the Kurtz's theorem (see below).

A question arises if there exists a condition which $f_0^{(e)}$ - $f_n^{(e)}$ should satisfy to assure that $G(q)$ has only real zeros. With such a condition at hand, the exchange profiles producing complex

zeros could be detected more directly. It would also enable fitting by adjusting the elements in $\boldsymbol{f}^{(e)}$ until they fulfill the condition for real zeros. Indeed, suitable conditions for real zeros are known: one that is necessary,[26] and one that is sufficient.[27] The necessary condition holds for a general polynomial with real coefficients, but the sufficient condition applies only to the polynomials with strictly positive or strictly alternating coefficients as in (10), which is a fortunate situation. Even more, both conditions are simple to understand and apply. The necessary condition was discovered long time ago by Newton:[26]

$$f_i \geq \sqrt{\frac{(n-i+1)}{n-i}\frac{(i+1)}{i}}\sqrt{f_{i-1}f_{i+1}}, i = 1,\ldots,n-1 \tag{11}$$

The sufficient condition was formulated by D. C. Kurtz,[27] in even more simple form:

$$f_i > 2\sqrt{f_{i-1}f_{i+1}}, i = 1,\ldots,n-1 \tag{12}$$

The condition (12) guarantees not only real but also all distinct zeros. Except for n=2, the condition that would be simultaneously sufficient and necessary is not known; thus the fitting criterion can not be formulated without calculation of the site exchanging probabilities p_i.

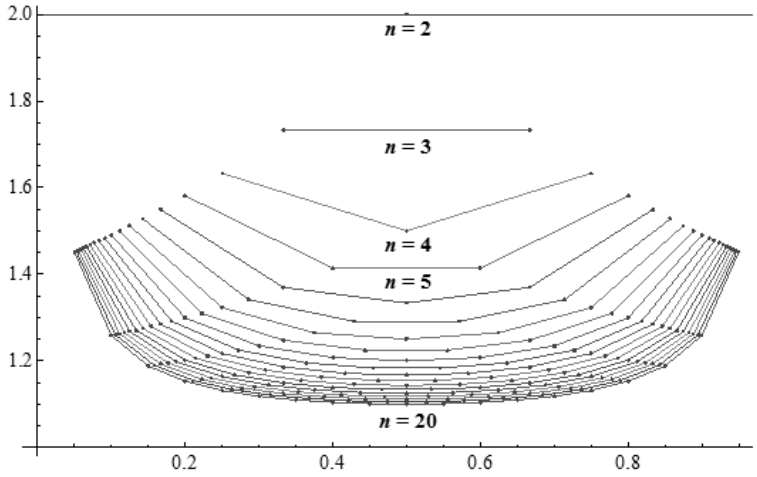

Figure 2. Graphical presentation of the factors $\sqrt{(n-i+1)(i+1)/(i(n-i))}$ (on the ordinate) across the ratio i/n (on the abscissa) for i = 1–n-1 and n = 2–20, from top to bottom. The points with i varied for fixed n are connected by lines.

Kurtz also proved that the factor 2 in (12) cannot be improved since there always exists a polynomial with non-real zeros that satisfies the modified condition (12) in which 2 is replaced by any smaller value. Note that only for n=2, the conditions (11) and (12) coincide. This simplest case is considered separately in Section 7.

The factor from (11) is graphically depicted across the ration i/n in Figure 2 to show how its values change with increasing n. The point at the top represents the case n=2 and at the same time the minimal value of the ratio $r \equiv f_i / \sqrt{f_{i-1} f_{i+1}}$ that guarantees that all polynomial roots are real. When the ratio is below the point for given i and n, non-real roots are present. When r is between the top line at 2 and the point for appropriate i and n, the character of the roots is uncertain. It can be seen how the region with undetermined character of the roots grows with increasing n.

4. Least squares fitting of the site exchange probabilities p_i

Whenever there are non-real roots of $F(p)$ (or equivalently, of $G(q)$), the probabilities p_i can be determined by fitting to the experimental values according to some criterion. In this section we consider the least squares fitting (LSF), in which p_i are determined by minimizing the sum S of squared deviations between the experimental $f_m^{(e)}$ and calculated f_m:

$$S = \sum_{m=0}^{n} \left[f_m^{(e)} - f_m \right]^2 = \sum_{m=0}^{n} \left[f_m^{(e)} - \sum_{k=m}^{n} (-1)^{k-m} \binom{k}{m} s_k(\boldsymbol{n}) \right]^2 \tag{13}$$

Here we give analytical expressions for the gradient and the Hessian since they are needed in various fitting algorithms. Hessian is not necessarily required, but it is always good to have since fitting algorithms generally perform better if Hessian is used. Besides, after finding the stationary point (by any algorithm), it should be checked whether it represents a minimum or perhaps a saddle point and this requires diagonalization of the Hessian. The elements g_i of the gradient $\boldsymbol{g}$ read as:

$$g_i = \frac{\partial S}{\partial p_i} = -2 \sum_{m=0}^{n} \left[f_m^{(e)} - \sum_{k=m}^{n} (-1)^{k-m} \binom{k}{m} s_k(\boldsymbol{n}) \right] \sum_{k=m}^{n} (-1)^{k-m} \binom{k}{m} \frac{\partial s_k(\boldsymbol{n})}{\partial p_i} \tag{14}$$

The derivative of $s_k(\boldsymbol{n})$ is written in a bit cryptic way:

$$\frac{\partial s_k(\boldsymbol{n})}{\partial p_i} = s_{k-1}\left(\{p_1, p_2 \dots p_n\} \setminus p_i\right) \equiv s_{k-1}(\boldsymbol{n} \setminus i) \tag{15}$$

$s_m(\{p_1, p_2 \dots p_n\} \setminus p_i)$ denotes elementary symmetric polynomial of order m in variables p_j, j=1-n, $j \neq i$, that is – in the reduced set of probabilities with p_i omitted. This is shortly designated by $s_m(\boldsymbol{n} \setminus i)$, where boldface $\boldsymbol{n}$ stands for all n probabilities and "\i" denotes omitting p_i. The final expression for g_i reads:

$$g_i = -2\sum_{m=0}^{n}\left[f_m^{(e)} - \sum_{k=m}^{n}(-1)^{k-m}\binom{k}{m} s_k(\boldsymbol{n}) \right] \sum_{k=m}^{n}(-1)^{k-m}\binom{k}{m} s_{k-1}(\boldsymbol{n} \setminus i) \tag{16}$$

In a similar way one obtains the expression for elements H_{ij} of the Hessian $\boldsymbol{H}$:

$$\begin{aligned} H_{ij} &= \frac{\partial^2 S}{\partial p_i \partial p_j} = \frac{\partial g_i}{\partial p_j} = \\ &= -2\sum_{m=0}^{n}\left[\sum_{k=m}^{n}(-1)^{k-m}\binom{k}{m} s_{k-1}(\boldsymbol{n} \setminus j) \right] \sum_{k=m}^{n}(-1)^{k-m}\binom{k}{m} s_{k-1}(\boldsymbol{n} \setminus i) - \\ &\quad - 2\sum_{m=0}^{n}\left[f_m^{(e)} - \sum_{k=m}^{n}(-1)^{k-m}\binom{k}{m} s_k(\boldsymbol{n}) \right] \sum_{k=m}^{n}(-1)^{k-m}\binom{k}{m} s_{k-2}(\boldsymbol{n} \setminus i \setminus j) \end{aligned} \tag{17}$$

The result of $\partial s_{m+1}(\boldsymbol{n} \setminus i)/\partial p_j$, which occurs in derivation of (17), is denoted by $s_m(\boldsymbol{n} \setminus i \setminus j)$ – which stands for the elementary symmetric polynomial of order m in the variables $\{p_k, k=1\text{-}n \mid k \neq i$ and $k \neq j\}$, that is, with omitted p_i and p_j. By definition, $s_m(\boldsymbol{n} \setminus i \setminus j)=0$ when $i=j$. In order to keep simple form of expressions (16) and (17), we extend the notation for s_m by defining $s_m(...)=0$ if $m<0$ or if m is greater than the cardinality of the set in the argument: e.g. $s_n(\boldsymbol{n} \setminus i)=0$ and $s_{n-1}(\boldsymbol{n} \setminus i \setminus j)=0$, by this definition.

One of the simplest and quite efficient fitting algorithms is the Newton method,[28] in which the fitted function (S) is assumed to be sufficiently close to a quadratic function in the parameters $\boldsymbol{p}$. With the gradient $\boldsymbol{g}$ and the Hessian $\boldsymbol{H}$ calculated for some guessed probabilities $\boldsymbol{p}$, the improved values $\boldsymbol{p}_{new}$, are obtained from:

$$\boldsymbol{p}_{new} = \boldsymbol{p} - \boldsymbol{H}^{-1}\boldsymbol{g} \tag{18}$$

The quantities $s_m(\boldsymbol{n})$, $s_m(\boldsymbol{n} \setminus i)$ and $s_m(\boldsymbol{n} \setminus i \setminus j)$, in (16) and (17) may look complicated and difficult for evaluation. Here we describe a simple way to compute them efficiently. Note that these

quantities are sums of the products, each involving m p_i-factors. As m goes from 1 to n in (4), (16) and (17), each of the subsets of $\{p_1, p_2, ..., p_n\}$ will appear as a product. Instead of a given $s_m(...)$, we focus on systematic generation of all subsets, evaluating the product for each one and summing the result into the appropriate array element. The subsets can be coded as binary representations of the numbers from 0 to 2^n-1, with 1/0 at the i-th position meaning that p_i is included/excluded in the product. In the loop going from 0 to 2^n-1, one generates the binary representation of the counting index and evaluates the product of the included p_i. According to the length m of the product and the distribution of 0s in the binary representation, the product is summed into the appropriate elements of 1-, 2- and 3-dimensional arrays representing $s_m(\boldsymbol{n})$, $s_m(\boldsymbol{n}\backslash i)$ and $s_m(\boldsymbol{n}\backslash i\backslash j)$, respectively, for m=0–n and i,j=1–n.

5. Degeneracy in the least squares fitting results

Here we prove that there are always equal values, at least two of them, among the probabilities obtained by LSF. This is so whenever the polynomial $F(p)$ has complex roots, that is, whenever the experimental set $f_m^{(e)}$ cannot be exactly reproduced by a real set of site-exchange probabilities.

We start by rewriting equations that g_i satisfy at the stationary point of S in terms of the variable q introduced in (8), in parallel to the expression (14) written in terms of p:

$$g_i = \frac{\partial S}{\partial q_i} = -2\sum_{m=0}^{n}\left[f_m^{(e)} - f_m\right]\frac{\partial f_m}{\partial q_i} = 0 \quad i = 1, ..., n \tag{19}$$

After dividing with -2 and denoting $[f_m^{(e)}–f_m]$ with δ_m, the above equations become:

$$\sum_{m=0}^{n}\delta_m \frac{\partial f_m}{\partial q_i} = 0 \qquad i = 1, ..., n \tag{20}$$

There is an additional condition from having both sums of $f_m^{(e)}$ and f_m equal to 1:

$$\sum_{m=0}^{n}\delta_m = 0 \tag{21}$$

The eqns. (20-21) can be considered as a homogeneous linear system with δ_m as unknowns. We use it only for purposes of the proof; the system has no practical value for finding the stationary point. This system of (n+1) equations in (n+1) unknowns has a nontrivial solution – with at least one $\delta_m \neq 0$, only if the determinant of the coefficients' matrix is zero. Here we

prove that it may happen only if q_i are not all distinct. By recalling the right hand sides of (9), one obtains for the partial derivatives in (20):

$$\begin{aligned}\frac{\partial f_m}{\partial q_i} &= \frac{\partial f_0}{\partial q_i} t_m(\boldsymbol{n}) + f_0 \frac{\partial t_m(\boldsymbol{n})}{\partial q_i} = -f_0 \frac{t_m(\boldsymbol{n})}{1+q_i} + f_0 t_{m-1}(\boldsymbol{n} \setminus i) = \\ &= \frac{f_0}{1+q_i}\left[t_{m-1}(\boldsymbol{n} \setminus i) + q_i t_{m-1}(\boldsymbol{n} \setminus i) - t_m(\boldsymbol{n})\right] = \frac{f_0}{1+q_i}\left[t_{m-1}(\boldsymbol{n} \setminus i) - t_m(\boldsymbol{n} \setminus i)\right]\end{aligned} \tag{22}$$

After substituting (22) into (20) and multiplying with $(1+q_i)/f_0$, the matrix of coefficients becomes:

$$\begin{bmatrix} -1 & 1-t_1(\boldsymbol{n} \setminus 1) & t_1(\boldsymbol{n} \setminus 1) - t_2(\boldsymbol{n} \setminus 1) & \cdots & t_{n-1}(\boldsymbol{n} \setminus 1) \\ -1 & 1-t_1(\boldsymbol{n} \setminus 2) & t_1(\boldsymbol{n} \setminus 2) - t_2(\boldsymbol{n} \setminus 2) & \cdots & t_{n-1}(\boldsymbol{n} \setminus 2) \\ \vdots & \vdots & & \ddots & \vdots \\ -1 & 1-t_1(\boldsymbol{n} \setminus n) & t_1(\boldsymbol{n} \setminus n) - t_2(\boldsymbol{n} \setminus n) & \cdots & t_{n-1}(\boldsymbol{n} \setminus n) \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix} \tag{23}$$

Note that because $t_{-1}(\boldsymbol{n} \setminus i)=0$ and $t_n(\boldsymbol{n} \setminus i)=0$ by definition, the leftmost and the rightmost columns contain only a single term. Now we examine the rank of the matrix (23). After adding the first (the leftmost) column to the second, then the second column to the third, and proceeding in the same way until the n-th column is added to the last column (the rightmost), the matrix becomes:

$$\begin{bmatrix} -1 & -t_1(\boldsymbol{n} \setminus 1) & -t_2(\boldsymbol{n} \setminus 1) & \cdots & 0 \\ -1 & -t_1(\boldsymbol{n} \setminus 2) & -t_2(\boldsymbol{n} \setminus 2) & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ -1 & -t_1(\boldsymbol{n} \setminus n) & -t_2(\boldsymbol{n} \setminus n) & \cdots & 0 \\ 1 & 2 & 3 & \cdots & n+1 \end{bmatrix} \tag{24}$$

If the determinant of the above matrix is zero, there must be a linearly dependent combination of rows. The last row cannot be in the combination since it contains the only nonzero element in the rightmost column. We proceed by examining the top n rows. Let each of them represents a polynomial with coefficients given in the columns 1 through n. There are n such row-polynomials, each of degree n-1. Linear dependence of rows implies linear dependence of the represented polynomials. Recall that $t_m(\boldsymbol{n} \setminus j)$ are elementary symmetric polynomials in q_i, $i \neq j$. Therefore all the row-polynomials are zero at $-q_j$ except the j-th one. The j-th row cannot be in the linearly dependent combination because it is the only polynomial with nonzero value at $-q_j$. The same reasoning holds for every row, and thus we conclude that the

matrix of the coefficients (23) may not have zero determinant if all q_i are different. In this case the system (20-21) may have only the trivial solution, in contradiction to the assumption that $f_m^{(e)}$ cannot be exactly reproduced by f_m. Therefore not all q_i can be different.

Numerical results obtained by LSF with simulated exchange profiles show that degeneracy is usually greater than minimal, that is, with more than 2 equal values among the fitted probabilities. Their number increases with the number of complex roots of $F(p)$ (or $G(q)$) with certain regularity that has yet to be proven (see Section 9).

6. Maximum likelihood estimation of p_i

In another way of estimating the site exchange probabilities $\boldsymbol{p}$, one maximizes the likelihood (3) of observing the experimental exchange profile $\boldsymbol{f}^{(e)}$, that is, one searches for $\boldsymbol{p}$ where the distribution (3) has maximum value.

Instead of (3) it is more convenient to consider the logarithm which has maximum at the same point:

$$\log\big(P(N_0, N_1, ..., N_n)\big) = \log(N!) - \sum_{m=0}^{n} \log(N_m!) + \sum_{m=0}^{n} N_m \log(f_m) \tag{25}$$

The first two terms do not depend on p_i and can be omitted from further consideration. Only f_m in the last summation depend on p_i as given in (1). The necessary condition for a maximum is given by equating partial derivatives to zero:

$$\frac{\partial \log\big(P(N_0, N_1, ..., N_n)\big)}{\partial q_j} = \sum_{m=0}^{n} N_m \frac{\partial \log(f_m)}{\partial q_j} = 0\,, \quad j = 1, 2, ..., n \tag{26}$$

By dividing with N, N_m is replaced by $f_m^{(e)}$, and then by using (9) and (15), the above set of equations is transformed as follows:

$$\sum_{m=0}^{n} f_m^{(e)} \frac{\partial \log f_m}{\partial q_j} = \sum_{m=0}^{n} f_m^{(e)} \frac{\partial \log f_0 + \partial \log t_m(\boldsymbol{n})}{\partial q_j} = \sum_{m=0}^{n} f_m^{(e)} \left(-\frac{1}{1+q_j} + \frac{t_{m-1}(\boldsymbol{n} \setminus j)}{t_m(\boldsymbol{n})} \right) = 0 \tag{27}$$

By multiplying with $(1+q_j)$ and recalling that $f_m^{(e)}$ sum to 1, one obtains:

$$(1+q_j)\sum_{m=0}^{n} f_m^{(e)} \frac{t_{m-1}(\boldsymbol{n} \setminus j)}{t_m(\boldsymbol{n})} = 1, \quad j = 1, 2, ..., n \tag{28}$$

Equations (28) represent conditions for stationary points of the likelihood density for observing the experimental set $\boldsymbol{f}^{(e)}$. In the rest of the section we show some characteristic solutions of these equations.

First we show that the roots of $G(q)$ after substituting f_m in (10) with $f_m^{(e)}$, that is, those q_i for which $\boldsymbol{f}=\boldsymbol{f}^{(e)}$, satisfy (28), as it was previously with the LSF equations (19). Then $f_m^{(e)} = f_0 \cdot t_m(\boldsymbol{n})$ and (28) becomes:

$$(1+q_j) f_0 \sum_{m=0}^{n-1} t_m(\boldsymbol{n} \setminus j) = 1, \quad j = 1, 2, ..., n \tag{29}$$

By using (9) it is easy to verify that the above equation holds since the summation is equal to $[(1+q_j)f_0]^{-1}$. Thus the site probabilities which exactly reproduce the experimental values, are the most likely ones, as could be expected. However, this is acceptable only if all the roots of $G(q)$ are real.

When there are non-real roots of $G(q)$, the MLE solution exhibits degeneracy as well as the LSF solution. By multiplying the last eqn. in (27) with $(1+q_j)$ and by simple transformation we obtain:

$$\sum_{m=0}^{n} \frac{f_m^{(e)}}{t_m(\boldsymbol{n})} - t_m(\boldsymbol{n}) + (1+q_j) t_{m-1}(\boldsymbol{n} \setminus j) \; = \; f_0 \sum_{m=0}^{n} \frac{f_m^{(e)}}{f_m} \Big(- t_m(\boldsymbol{n}) + (1+q_j) t_{m-1}(\boldsymbol{n} \setminus j) \Big) = 0 \tag{30}$$

By dividing with f_0 and by using the identity: $t_m(\boldsymbol{n}) = q_j \cdot t_{m-1}(\boldsymbol{n}\setminus j) + t_m(\boldsymbol{n}\setminus j)$, one obtains

$$\sum_{m=0}^{n} \frac{f_m^{(e)}}{f_m} \big(t_{m-1}(\boldsymbol{n} \setminus j) - t_m(\boldsymbol{n} \setminus j) \big) = 0 , \quad j = 1, 2, ..., n \tag{31}$$

This set of equations reminds to the set in (20) and (22). We use similar reasoning to show that all distinct probabilities q_j in the MLE solution imply $\boldsymbol{f} = \boldsymbol{f}^{(e)}$. First we regroup the terms in (31):

$$\sum_{m=0}^{n-1} - t_m(\boldsymbol{n} \setminus j) \cdot \left(\frac{f_m^{(e)}}{f_m} - \frac{f_{m+1}^{(e)}}{f_{m+1}} \right) = 0 \tag{32}$$

This is a homogenous system of n equations with the differences in brackets treated as unknowns. Note that the matrix of coefficients is equal to the top left part of the matrix in (24) involving the first n rows and n columns. By the same consideration as for the matrix (24), we conclude that the system of equations (32) has only trivial solution when all q_j are distinct. In the trivial solution all the unknowns are equal to zero, implying:

$$\frac{f_0^{(e)}}{f_0} = \frac{f_1^{(e)}}{f_1} = \cdots = \frac{f_n^{(e)}}{f_n} = const. \tag{33}$$

From $\sum_{m=0}^{n} f_m^{(e)} = \sum_{m=0}^{n} f_m = 1$, one obtains *const*.=1. Thence, if $\boldsymbol{f} \neq \boldsymbol{f}^{(e)}$, all q_j cannot be distinct.

In the rest of the section we point out a special stationary point, with q_j being equal for all j: $q_1 = q_2 = ... = q_n \equiv q$. Then:

$$t_{m-1}(\boldsymbol{n} \setminus j) = \binom{n-1}{m-1} q^{m-1}, \tag{34}$$

and from (28) one obtains:

$$q = \frac{\sum_{m=1}^{n} m f_m^{(e)}}{n - \sum_{m=1}^{n} m f_m^{(e)}} \tag{35}$$

From (6) follows that each summation in (35) is equal to $s_1(\boldsymbol{n})$. By using (8) to express p from q, one finds that q corresponds to:

$$p = \frac{s_1(n)}{n} \tag{36}$$

This is nothing else but the mean value of roots of the polynomial in (7). Since complex roots come in conjugated pairs, the imaginary parts cancel and the values for p and q are real (this is also clear directly from (35) which involves only real numbers).

7. Case $n = 2$

When n=2, H/D exchange takes place at two sites, with probabilities p_1 and p_2. The exchange profile consists of $f_0^{(e)}$, $f_1^{(e)}$ and $f_2^{(e)}$. Since their sum is fixed (to 1), only two values can be varied independently and we choose $f_0^{(e)}$ and $f_2^{(e)}$ because they are symmetrically related. The range of possible values for the pairs ($f_0^{(e)}$, $f_2^{(e)}$) are determined by the conditions $f_0^{(e)}$, $f_2^{(e)} \geq 0$. and $f_0^{(e)} + f_2^{(e)} \leq 1$. The set of all possible pairs ($f_0^{(e)}$, $f_2^{(e)}$) that can be realized in the experiment is called an *experimental region*. Pairs within the experimental region are differentiated by non/existence of probabilities p_1 and p_2 that exactly correspond to a given pair $f_0^{(e)}$ and $f_2^{(e)}$. In other words, the experimental region is split into two parts: one containing the pairs ($f_0^{(e)}$, $f_2^{(e)}$) for which $G(q) = f_0^{(e)} \cdot q^2 - f_1^{(e)} \cdot q + f_2^{(e)}$ has real roots, and the other with all the remaining pairs. The subset with real roots is called the *exact region* as the roots exactly reproduce a given pair ($f_0^{(e)}$, $f_2^{(e)}$). Boundary between the exact and *inexact* regions is determined by the zero value of the discriminant d:

$$\begin{aligned} d &= \left(f_1^{(e)}\right)^2 - 4 f_0^{(e)} f_2^{(e)} = \left(1 - f_0^{(e)} - f_2^{(e)}\right)^2 - 4 f_0^{(e)} f_2^{(e)} = \\ &= \left(1 - f_0^{(e)} - f_2^{(e)} - 2\sqrt{f_0^{(e)} f_2^{(e)}}\right) \cdot \left(1 - f_0^{(e)} - f_2^{(e)} + 2\sqrt{f_0^{(e)} f_2^{(e)}}\right) = \\ &= \left[1 - \left(\sqrt{f_0^{(e)}} + \sqrt{f_2^{(e)}}\right)^2\right] \cdot \left[1 - \left(\sqrt{f_0^{(e)}} - \sqrt{f_2^{(e)}}\right)^2\right] = 0 \end{aligned} \qquad (37)$$

The discriminant is zero only when $\sqrt{f_0^{(e)}} + \sqrt{f_2^{(e)}} = 1$. Figure 3 shows the diagram with delineated experimental region and its exact and inexact parts.

The LSF probabilities for the pairs ($f_0^{(e)}$, $f_2^{(e)}$) in the inexact region are determined by the zero condition for gradients, eqns. (16) and (19). When n=2, by the degeneracy rule established in Section 5, the two LSF probabilities are equal: $p_1 = p_2 \equiv p$, and the stationary point conditions reduce to a single cubic equation (written in two forms: in terms of p and in terms of q):

$$\text{derived from (16):} \quad p^3 - \frac{3}{2} p^2 - \frac{1}{6}\left(3 f_0^{(e)} + 3 f_2^{(e)} - 7\right) p + \frac{1}{6}\left(2 f_0^{(e)} + f_2^{(e)} - 2\right) = 0$$

$$\text{derived from (19):} \quad q^3\left(2 f_2^{(e)} + f_0^{(e)} - 2\right) + q^2\left(3 f_2^{(e)} + 1\right) - q\left(3 f_0^{(e)} + 1\right) - \left(2 f_0^{(e)} + f_2^{(e)} - 2\right) = 0 \qquad (38)$$

We are interested in the number of real solutions, which is determined by the sign of the discriminant of the cubic equation. The discriminant d reads as:

$$d = \frac{1}{46656}\left[216\left(f_0^{(e)} + f_2^{(e)}\right)^3 - 621\left(f_0^{(e)} + f_2^{(e)}\right)^2 + 450\left(f_0^{(e)} + f_2^{(e)}\right) + 324 f_0^{(e)} f_2^{(e)} - 125\right] \quad (39)$$

The subregion with positive values of the discriminant is delineated in Fig. 3. It is relatively small triangular zone, corresponding to high and similar values f_0 and f_2. As the most distant part from the exact region, it involves the greatest experimental error. There are three real solutions of (38), representing three stationary points of the sum of squared deviations. Two of them are minima and the third is a saddle point. It should be understood that the two minima represent two different degenerate solutions $p_1 = p_2 \equiv p$. Fig. 3 demonstrates complexity which is present already in the simplest case n=2. It demonstrates that even in the simplest case, there can be more than one minimal sum of the squared distances and that saddle points can be also present. For n>2, the complexity is certainly bigger, and one should expect increased number of local minima and saddle points.

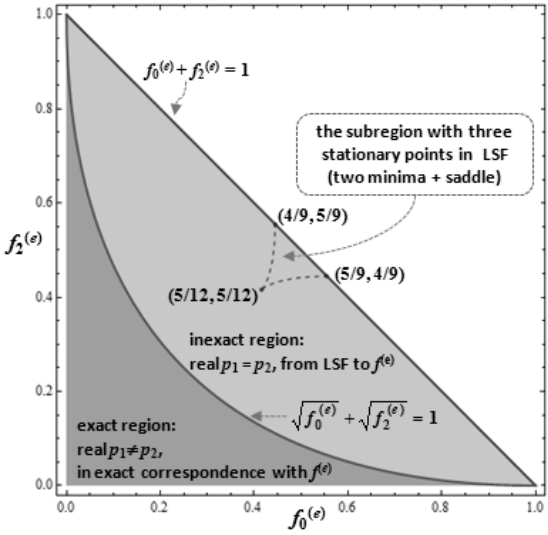

Figure 3. The plane $(f_0^{(e)}, f_2^{(e)})$ with the experimental region (below the straight line $f_0^{(e)}+f_2^{(e)}=1$), divided into the exact and inexact regions.

From Section 6 we know there is only one MLE solution in the inexact region:

$$q_1 = q_2 = \frac{f_1^{(e)} + 2 f_2^{(e)}}{2 - f_1^{(e)} - 2 f_2^{(e)}} \tag{40}$$

It corresponds to $p_1 = p_2$, equal to the real part of the polynomial roots from (7).

The simplest case n=2 allows us to examine the influence of a finite number of molecules detected in the experiment. The pairs ($f_0^{(e)}$, $f_2^{(e)}$) are distributed according to the multinomial distribution:

$$P(f_0^{(e)}, f_2^{(e)}; p_1, p_2, N) = \\ = \frac{N!}{N_0! N_1! N_2!} \left[(1-p_1)(1-p_2)\right]^{N_0} \left[(p_1(1-p_2) + p_2(1-p_1)\right]^{N_1} \left[p_1 p_2\right]^{N_2} \tag{41}$$

P stands for the probability of observing the pair ($f_0^{(e)}$, $f_2^{(e)}$) when the true exchange probabilities are p_1 and p_2, and N is the total number of detected molecules. N_0, N_1 and N_2 represent the number of molecules with zero, one and two hydrogens exchanged by deuterons. Clearly:

$$\begin{aligned} N &= N_0 + N_1 + N_2 \\ f_0^{(e)} &= N_0 / N \\ f_1^{(e)} &= N_1 / N \\ f_2^{(e)} &= N_2 / N \end{aligned} \tag{42}$$

Figure 4 shows the distribution when N=100, p_1=0.5 and p_2=0.7 (N=100 is unrealistic low, but it well illustrates the distribution features). In order to see better how the distribution divides between exact and inexact regions, the inexact part is cut and shifted towards the base centre.

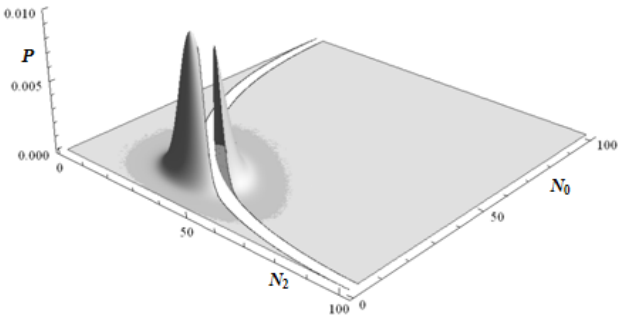

Figure 4. The distribution (41) for N=100, p_1=0.5 and p_2=0.7. The distribution is split into exact and inexact parts and the inexact one is shifted towards the base centre.

It is instructive to observe the distribution of pairs ($f_0^{(e)}$, $f_2^{(e)}$) obtained by integrating the distribution (41) over p_1 and p_2 from 0 to 1:

$$P(f_0^{(e)}, f_2^{(e)}; N) = \\ = \frac{N!}{N_0!N_1!N_2!}\int_0^1 dp_2 \int_0^1 dp_1 [(1-p_1)(1-p_2)]^{N_0} [(p_1(1-p_2)+p_2(1-p_1)]^{N_1} [p_1 p_2]^{N_2} \quad (43)$$

If one takes that p_1 and p_2 are uniformly distributed, then P in (43) indicates general observation incidence of pairs ($f_0^{(e)}$, $f_2^{(e)}$).

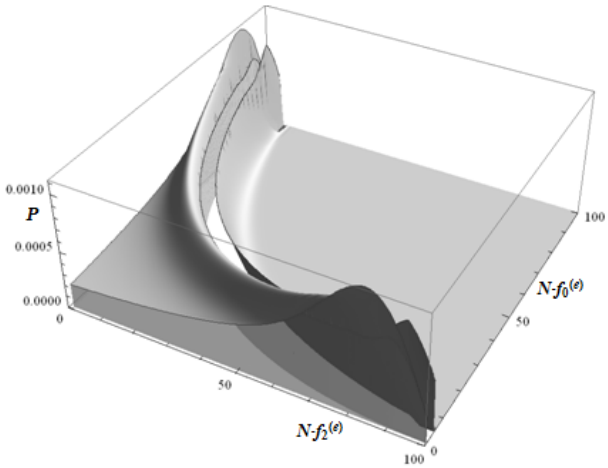

Figure 5. The distribution (43) for N = 100. As in Fig. 4, the distribution is split into exact and inexact parts and the inexact part is shifted towards the base centre.

The distribution (43) for N=100 is showed in Figure 5. It should be noted that the uniform distribution of (p_1, p_2) results in the distribution of ($f_0^{(e)}$, $f_2^{(e)}$) with higher probabilities close to the boundary of the exact region, where $\sqrt{f_0^{(e)}} + \sqrt{f_2^{(e)}} = 1$. Figure 6 shows cumulative distributions of $\sqrt{f_0^{(e)}} + \sqrt{f_2^{(e)}}$ for N increasing from 100 to 1000. The pairs ($f_0^{(e)}$, $f_2^{(e)}$) do not extend far beyond the exact region, with inexact fractions ranging from 0.17 for N=100, to 0.11 for N=1000. In parallel with shrinking of the inexact fraction, the exact fraction increases just near the boundary of the exact region. With an additional experimental error, perturbing

($f_0^{(e)}$, $f_2^{(e)}$) uniformly in all directions, the inexact fraction increases due to the accumulated distribution near the boundary. One would expect improvement in accuracy by increasing the number of detected ions, but the above observation makes the net effect somewhat obscured (see also discussion in Section 8 referring to Table 2). Extrapolation to n>2 of the distribution in Fig. 5 in respect to the consequences for exactness of experimental exchange profiles, is not straightforward since it is not known how the boundary between exact and inexact regions looks like for n>2.

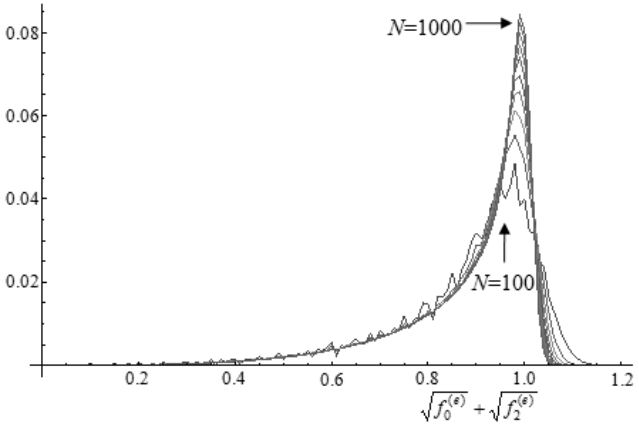

Figure 6. The cumulative distributions of $\sqrt{f_0^{(e)}}+\sqrt{f_2^{(e)}}$ obtained for N = 100-1000 with step=100. The lines for lower N appear broken because the distributions have more pronounced discrete character.

8. Some numerical results

There are many interesting questions in the present context that can be answered only by simulations. One of them is about the incidence of complex roots of polynomials $G(q)$ obtained from realistic mass spectra. To answer this question, at least approximately, we performed numerical simulations in which random effects were simulated by using random numbers. First, n random numbers from the interval [0, 1] were produced to represent H/D exchange probabilities at n exchanging sites. They were converted to probabilities of having 0 to n H/D replacements in a given molecular fragment. The interval [0, 1] was then divided into segments numbered by 0–n and ordered so that the length of the m-th segment was equal to probability of having m hydrogens exchanged. Another random number from [0, 1] was drawn and the ordinal number of the segment containing the drawn number was taken as

equal to the number of H/D exchanges in the simulated fragment. The last drawing was repeated N times, in a sequence N = 1 000, 10 000, 100 000, 1 000 000. N represents the number of detected fragments from which $f_0^{(e)} - f_n^{(e)}$ were determined. No other experimental error was simulated. Simulations for each N were repeated 1 000 000 times and each time the roots of the polynomial $G(q)$ were determined.

Table 1. Proportions (in percentages) of polynomials $G(q)$ (10) with z complex conjugate pairs of zeros. Polynomial coefficients were obtained by simulating random deuteration at n sites in N molecular fragments (see the text for simulation details; the sum of percentages for a given pair (n, N) can be slightly different from 100.0 due to a rounding error).

n	z	N = 1 000	N = 10 000	N = 100 000	N = 1 000 000
3	0	63.1	75.8	85.0	90.9
	1	36.9	24.2	15.0	9.1
4	0	32.8	49.0	64.6	76.7
	1	59.0	46.2	33.0	22.1
	2	8.2	4.8	2.5	1.2
5	0	12.2	23.4	39.5	55.8
	1	55.5	54.1	47.3	37.1
	2	32.2	22.5	13.2	7.1
6	0	3.4	8.0	18.4	33.1
	1	34.8	40.3	45.7	44.8
	2	55.6	46.9	33.2	20.9
	3	6.1	4.9	2.7	1.3
7	0	0.8	2.0	6.0	15.2
	1	16.5	19.8	29.8	38.5
	2	59.5	55.1	48.6	37.6
	3	23.2	23.1	15.6	8.7
8	0	0.2	0.4	1.4	5.1
	1	6.4	7.0	12.9	23.3
	2	48.0	41.4	44.0	44.0
	3	43.1	46.9	38.5	25.9
	4	2.3	4.3	3.2	1.8
9	0	0.0	0.1	0.2	1.2
	1	2.2	2.0	3.9	9.5
	2	32.1	22.7	25.7	33.6
	3	56.6	56.9	52.5	43.9
	4	9.1	18.4	17.7	11.7
10	0	0.0	0.0	0.0	0.2
	1	0.6	0.5	0.9	2.7
	2	19.0	9.9	10.7	17.0
	3	60.0	49.9	45.0	45.3
	4	20.1	37.7	40.3	32.5
	5	0.3	2.0	3.1	2.4

Table 1 displays percentages of complex conjugated pairs among the roots (note that these are only approximate values obtained by simulation). One should observe that the incidence of all the roots being real (z=0 in Table 1) quickly decreases as n grows. At the same time, the average number of complex conjugated pairs also increases (see Table 2). As could be expected, for greater N, the number of complex roots in dependence on n increases slower. Remarkably, the last few rows of Table 2 display increasing average number of complex roots when N increases.

Table 2. Average number of complex conjugated pairs in the roots of polynomials $G(q)$ (10)

n	N = 1 000	N = 10 000	N = 100 000	N = 1 000 000
3	0.37	0.24	0.15	0.09
4	0.75	0.56	0.38	0.24
5	1.20	0.99	0.74	0.51
6	1.64	1.49	1.20	0.90
7	2.05	1.99	1.74	1.40
8	2.41	2.48	2.29	1.96
9	2.73	2.92	2.84	2.55
10	3.00	3.31	3.34	3.14

9. Concluding remarks

Most of the findings were discussed in the previous sections, and here we give only some additional remarks. As the main result we point out the degeneracy of the probabilities obtained by fitting, either within the LSF or the MLE approach, whenever experimental values are inconsistent with the model.

The LSF results obtained in a preliminary investigation (to be reported in detail in the next sequel [25]) support a hypothesis that the number of distinct probabilities does not exceed the number of distinct roots of $G(q)$ when conjugated pairs are counted as single roots. If the number of complex roots is considered as a measure of inconsistency with the model, it appears that the uncertainty, represented by degeneracy of the fitted results, increases as the data deviate more from the physical model. It remains to be verified whether it is also present in the common experimental setup when several exchange profiles, recorded after different deuteration times, are fitted at once.

An important aspect of the fitting procedure, irrespectively of the fitting criterion, is a high symmetry of the objective function, being invariant to any permutation of the site exchange probabilities. As a consequence, every stationary point of the sum of squared distances or of the likelihood for a given exchange profile, is replicated by permuting distinct site-probabilities. It means that there are many equivalent stationary points - minima and various saddle points between them. With a possibility of different minima (local and global), the expected number of stationary points further increases. Even the simplest case with n=2 demonstrates that this is possible and it should be expected with greater incidence for n>2. This emphasizes importance of verifying the character of the stationary point obtained numerically.

In regard to the numerical results in Section 8, in particular to their dependence on N – the number of detected ions in an instrument, it is interesting to know the experimental value of N. Certainly, it can vary a lot depending on the spectrometer type, the characteristics of a sample and the specific setup. It seems that in the instruments used for the H/D experiments with proteins, the number of ions trapped in the FT-ICR mass spectrometer is of the order of 10^5.[29, 30] However, it should be taken into account that this quantity of ions produces an isotopic profile, including some noise in the signals. The precision of the exchange profile obtained by deconvolution reflects not only a finite number of detected ions but also the other sources of experimental error. One could take that the final effect for exactness of the exchange profile can be considered as effective reduction of the number of detected ions which we roughly and deliberately estimate by one order of magnitude. By assuming that the average number of aminoacids in the peptides used for quantitative analysis is 5-6, from the middle part of Table 1 one sees that the roots of $G(q)$ will rarely be all real.

At the end we emphasize that by no means the single exchange profiles discussed here, are proposed as a vehicle for quantitative analysis of HDX. In fitting the data from a single exchange profile, it is important to consider uncertainties arising from experimental error and the curvature at the minimum sum of squared deviations. Comparison of the fitting algorithms and characteristics of the fitting results are elaborated in the next sequel.[25]

Acknowledgement: The presented research was financially supported by the Ministry of Science, Education and Sports of The Republic of Croatia through grants 098-0982915-2942, 098-0982915-2945 and 098-0982933-2937. D.B. is grateful to Dr. Ivan Ljubić (Zagreb) for useful discussions and for pointing to us the paper by D. C. Kurtz.[27]

References

[1] A. Ilari, C. Savino, Protein structure determination by X-ray crystallography, in: J. Keith (Ed.), *Bioinformatics*, Humana Press, Totowa, New Jersey, 2008, pp. 63-87.

[2] M. Billeter, G. Wagner, K. Wüthrich, Solution NMR structure determination of proteins revisited, *J. Biomol. NMR* **42** (2008) 155-158.

[3] I. Jelesarov, H. R. Bosshard, Isothermal titration calorimetry and differential scanning calorimetry as complementary tools to investigate the energetics of biomolecular recognition, *J. Mol. Recognit.* **12** (1999) 3-18.

[4] S. M. Kelly, N. C. Price, The use of circular dichroism in the investigation of protein structure and function, *Curr. Protein Pept. Sci.* **1** (2000) 349-384.

[5] D. W. Piston, G. J. Kremers, Fluorescent protein FRET: the good, the bad and the ugly, *Trends Biochem. Sci.* **32** (2007) 407-414.

[6] K. J. Pacholarz, R. A. Garlish, R. J. Taylor, P. E. Barran, Mass spectrometry based tools to investigate protein-ligand interactions for drug discovery, *Chem. Soc. Rev.* **41** (2012) 4335-4355.

[7] L. Konermann, J. Pan, Y. H. Liu, Hydrogen exchange mass spectrometry for studying protein structure and dynamics, *Chem. Soc. Rev.* **40** (2011) 1224-1234.

[8] A. N. Hoofnagle, K. A. Resing, N. G. Ahn, Protein analysis by hydrogen exchange mass spectrometry, *Annu. Rev. Biophys. Biomol. Struct.* **23** (2003) 1-25.

[9] L. S. Busenlehner, R. N. Armstrong, Minireview: Insights into enzyme structure and dynamics elucidated by amide H/D exchange mass spectrometry, *Arch. Biochem. Biophys.* **433** (2005) 34-46.

[10] A. Brock, Fragmentation hydrogen exchange mass spectrometry: A review of methodology and applications, *Protein Expr. Purif.* **84** (2012) 19-37.

[11] Z. Zhang, D. L. Smith, Determination of amide hydrogen exchange by mass spectrometry: A new tool for protein structure elucidation, *Protein Sci.* **2** (1993) 522-531.

[12] P. G. Fajer, G. M. Bou-Assaf, A. G. Marshall, Improved sequence resolution by global analysis of overlapped peptides in hydrogen/deuterium exchange mass spectrometry, *J. Am. Soc. Mass Spectrom.* **23** (2012) 1202-1208.

[13] H. M. Zhang, S. Kazazić, T. M. Schaub, J. D. Tipton, M. R. Emmett, A. G. Marshall, Enhanced digestion efficiency, peptide ionization efficiency, and sequence resolution for protein hydrogen/deuterium exchange monitored by Fourier transform ion cyclotron resonance mass spectrometry, *Anal. Chem.* **80** (2008) 9034-9041.

[14] E. Althaus, S. Canzar, C. Ehrler, M. R. Emmett, A. Karrenbauer, A. G. Marshall, A. Meyer-Baese, J. D. Tipton, H. M. Zhang, Computing H/D-exchange rates of single residues from data on proteolytic fragments, *BMC Bioinformatics* **11** (2010) 424-435.

[15] Z. Zhang, W. Li, T. M. Logan, M. Li, A. G. Marshall, Human recombinant [C22A] FK506-binding protein amide hydrogen exchange rates from mass spectrometry match and extend those from NMR, *Protein Sci.* **6** (1997) 2203-2217.

[16] Z. Zhang, S. Guan, A. G. Marshall, Enhancement of the effective resolution of mass spectra of high-mass biomolecules by maximum-entropy-based deconvolution to eliminate the isotopic natural abundance distribution, *J. Am. Soc. Mass Spectrom.* **8** (1997) 659-670.

[17] M. Hotchko, G. S. Anand, E. A. Komives, L. F. Ten Eyck, Automated extraction of backbone deuteration levels from amide H/^{2}H mass spectrometry experiments, *Protein Sci.* **15** (2006) 583-601.

[18] J. K. Chik, J. L. Vande Graaf, D. C. Schriemer, Quantitating the statistical distribution of deuterium incorporation to extend the utility of H/D exchange MS Data, *Anal. Chem.* **78** (2006) 207-214.

[19] G. S. Anand, M. Hotchko, S. H. J. Brown, L. F. Ten Eyck, E. A. Komives, S. S. Taylor, R-subunit isoform specificity in protein kinase A: Distinct features of protein interfaces in PKA types I and II by amide H/^{2}H exchange mass spectrometry, *J. Mol. Biol.* **374** (2007) 487-499.

[20] F. He, A. G. Marshall, Weighted quasi-Newton and variable-order, variable-step Adams algorithm for determining site-specific reaction rate constants, *J. Phys. Chem. A* **104** (2000) 562-567.

[21] B. G. Reuben, Y. Ritov, O. Geller, M. A. McFarland, A. G. Marshall, C. Lifshitz, Applying a new algorithm for obtaining site specific rate constants for H/D exchange of the gas phase proton-bound arginine dimer, *Chem. Phys. Lett.* **380** (2003) 88-94.

[22] P. J. Bickel, K. A. Doksum, *Mathematical Statistics. Basic Ideas and Selected Topics*, Prentice Hall, Upper Saddle River, 2001, pp. 99-160.

[23] D. Valkenborg, I. Mertens, F. Lemière, E. Witters, T. Burzykowski, The isotopic distribution conundrum, *Mass Spectrom. Rev.* **31** (2012) 96-109.

[24] X. Wang, A simple proof of Descartes's rule of signs, *Amer. Math. Montly* **111** (2004) 525-526.

[25] D. Babić *et al.*, work in progress.

[26] G. H. Hardy, J. E. Littlewood, G. Polya, *Inequalities*, Cambridge Univ. Press, London, 1934, pp. 51-55 and 104-106.

[27] D. C. Kurtz, A sufficient condition for all the roots of a polynomial to be real, *Amer. Math. Monthly* **99** (1992) 259-263.

[28] Ch. 11.3 in I. Griva, S. G. Nash, A. Sofer, *Linear and Nonlinear Optimization*, SIAM, Philadelphia, 2009, pp. 364-371.

[29] S. Kazazić, H. M. Zhang, T. M. Schaub, M. R. Emmett, C. L. Hendrickson, G. T. Blakney, A. G. Marshall, Automated data reduction for hydrogen/deuterium exchange experiments, enabled by high-resolution Fourier transform ion cyclotron resonance mass spectrometry, *J. Amer. Mass Spectrom.* **21** (2010) 550-558.

[30] P. Sagulenko, V. Frankevich, R. Steinhoff, R. Zenobi, Fluorescence-based method for determining the number of ions trapped in a FT-ICR mass-spectrometer, *Int. J. Mass Spectrom.* **338** (2013) 11-16.