BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 11 2013, pages 1375–1381
doi:10.1093/bioinformatics/btt168

Structural bioinformatics

Advance Access publication April 24, 2013

A novel web server predicts amino acid residue protection against hydrogen–deuterium exchange

Mikhail Yu. Lobanov[1], Masha Yu. Suvorina[1], Nikita V. Dovidchenko[1], Igor V. Sokolovskiy[1], Alexey K. Surin[1,2] and Oxana V. Galzitskaya[1,*]

[1]Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region 142290, Russia and [2]State Research Center for Applied Microbiology & Biotechnology, Obolensk, Serpukhov district, Moscow Region, 142279, Russia

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: To clarify the relationship between structural elements and polypeptide chain mobility, a set of statistical analyses of structures is necessary. Because at present proteins with determined spatial structures are much less numerous than those with amino acid sequence known, it is important to be able to predict the extent of proton protection from hydrogen–deuterium (HD) exchange basing solely on the protein primary structure.

Results: Here we present a novel web server aimed to predict the degree of amino acid residue protection against HD exchange solely from the primary structure of the protein chain under study. On the basis of the amino acid sequence, the presented server offers the following three possibilities (predictors) for user's choice. First, prediction of the number of contacts occurring in this protein, which is shown to be helpful in estimating the number of protons protected against HD exchange (sensitivity 0.71). Second, probability of H-bonding in this protein, which is useful for finding the number of unprotected protons (specificity 0.71). The last is the use of an artificial predictor. Also, we report on mass spectrometry analysis of HD exchange that has been first applied to free amino acids. Its results showed a good agreement with theoretical data (number of protons) for 10 globular proteins (correlation coefficient 0.73). We pioneered in compiling two datasets of experimental HD exchange data for 35 proteins.

Availability: The H-Protection server is available for users at http://bioinfo.protres.ru/ogp/

Contact: ogalzit@vega.protres.ru

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 7, 2012; revised on April 2, 2013; accepted on April 7, 2013

1 INTRODUCTION

Because flexible sequence regions frequently play an important functional role, they are currently in the focus of numerous studies. The functional properties of a protein molecule are a compromise between flexibility and rigidity. It was demonstrated that loops and β-turns showed a higher mobility than other structural elements (Rose *et al*., 1985). Flexibility is also characteristic of epitopes and binding sites of a number of enzymes modulating the life span of proteins *in vivo* (Fontana, 1988; Westhof *et al*., 1984). The method combining hydrogen–deuterium (HD) exchange and mass spectrometry data (HDMS) is widely used to study conformational changes in protein structures. HD exchange studies are mostly focused on the exchange of protons involved in H-bonding, which contributes to stabilization of secondary and spatial structures of proteins (Englander, 2006). HDMS results have proved to be useful in improving prediction of *de novo* protein structures (Malmstroem *et al*., 2009).

Recently, it was demonstrated that the direct contact of an amino acid residue with solvent did not guarantee a rapid HD exchange (Skinner *et al*., 2012), which pointed to the necessity of taking into account the structure–rate relationships (Skinner *et al*., 2012).

It should be noted that to predict rigid/flexible regions and HD exchange rates, a 3D structure was typically used as the subject matter of, for example, the pebble game/FIRST method (Jacobs *et al*., 2001). The CamP server made for prediction of protection factors solely from an amino acid sequence (Tartaglia *et al*., 2007) and previously available at http://www-almost.ch.cam.ac.uk/camp.php now does not exist, but in 2008 we had a chance to make predictions for 6 of 14 proteins (Dovidchenko and Galzitskaya, 2008). The fraction of amino acid residues with the correctly CamP-predicted degree of protection was 50%. In spite of the fact that CamP was a complex neural network-based method, it heavily overpredicted the number of HD exchange-protected residues, and thus, with 97% accuracy of prediction of protected regions, that of unprotected ones was as low as 13%.

We believe that the absence of protection may be explained mostly by large amplitude of fluctuations of the protein chain regions between packed elements of the secondary structure, which promotes contacts of the fluctuating region with solvent. Hence, prediction of unfolded and flexible regions can reveal those prone to large structural fluctuations, and therefore, to HD exchange. Some structural aspects of local flexibility are outlined in this work. We concentrated on the possibility of prediction of polypeptide chain protection against HD exchange, which resulted in the presented here server made for prediction of residues protected/unprotected against HD exchange. Additionally, we report on the pioneer use of mass spectrometry analysis of HD exchange for free amino acids and good correlation of its results with theoretical data for 10 globular proteins.

*To whom correspondence should be addressed.

2 METHODS

Approaches to prediction of residue protection against HD exchange: The observed packing density and H-bonding statistics were obtained for a database that contained spatial structures of 3769 proteins (Galzitskaya *et al.*, 2006). The database contained proteins that belong to four main structural classification of proteins (SCOP) (Murzin *et al.*, 1995) classes (classes a, b, c and d with all-α, all-β, α/β and $\alpha+\beta$ proteins, respectively). The proteins had <25% sequence identity to one another.

The observed packing density (the observed number of contacts) for each amino acid residue in the database was calculated (Galzitskaya *et al.*, 2006) as an average number of close residues (within a given distance). A residue was considered to be close to another one if any pair of their non-hydrogen atoms was at a distance <8 Å (the neighboring residues were excluded from the consideration). Then, the observed packing densities were averaged for each of 20 types of amino acid residues.

Hydrogen bonds were searched for in the same database. Backbone−backbone hydrogen bonds (that is, hydrogen bonds where the donor is an NH-group of the protein backbone and the acceptor is an O-atom of the protein backbone) were analyzed by a standard program DSSP (Kabsch and Sander, 1983). For each NH-group, only one hydrogen bond (which had the best energy, according to DSSP) was taken into consideration in this case. The criterion of hydrogen bond formation was that recommended by DSSP (the calculated energy lower than −0.5 kcal/mol). Based on the obtained data, the probability of H-bonding by each of 20 types of amino acid residues was calculated. For the calculation, the hydrogen bonds were 'ascribed' to donors.

The expected values (the expected packing density or H-bonding probability) served as predictors for constructing a packing density profile or an H-bonding probability profile. The calculations were made using the sliding-window averaging technique. First, the expected value was determined for each residue (it was equal to the average value observed for this type of residue in a spatial structure); then, these numbers were averaged inside the window, and the average was assigned to the central residue of the window. The 'smoothed' expected value for every position of the polypeptide chain provided the final profile, which was directly used for prediction of residues protected against HD exchange.

Basing either on the predictor of expected contacts or on that of expected number of hydrogen bonds, we predicted protected residues as those with a cut-off–exceeding number of expected contacts or expected hydrogen bonds.

Criterion for estimation of the quality of prediction: To assess the ability of a certain parameter to predict the degree of protection of amino acid residues against HD exchange, we compared the results of our predictions with experimental data. The quality of our predictions was evaluated using the commonly used standard values sensitivity and specificity, which are the fraction of correctly predicted protected residues and the fraction of correctly predicted unprotected residues, respectively.

$$Sensitivity = TP/N_p \tag{1}$$

$$Specificity = TN/N_{np} \tag{2}$$

$$ACC = \frac{Sensitivity + Specificity}{2} \tag{3}$$

where TP (true positives) is the number of amino acid residues correctly predicted as protected against HD exchange, TN (true negatives) is the number of amino acid residues correctly predicted as unprotected ones. N_p is the number of residues experimentally shown as protected, and N_{np} is the number of residues experimentally shown as unprotected against HD exchange. The quality of prediction was estimates as

$$Q_2 = (TP + TN)/N \tag{4}$$

where N is the total number of amino acid residues for which experimental data were available.

The error of averaging over proteins was calculated from

$$\Delta = \frac{\sigma}{\sqrt{n}} \tag{5}$$

where σ is the root-mean-square deviation, and n is the number of proteins. The error of averaging over amino acid residues was calculated using the equation

$$\Delta = \frac{1}{N} * \left(s_{TN}^2 + s_{TP}^2\right)^{1/2} = \frac{1}{N} * \left(\frac{TN * (N_{np} - TN)}{N_{np}} + \frac{TP * (N_p - TP)}{N_p}\right)^{1/2}. \tag{6}$$

Datasets: We used proteins with reported experimental data on protection of their separate amino acid residues against HD exchange. These residues were divided in two groups: protected and unprotected against HD exchange (in accordance with references below).

Amino acid sequences of the selected proteins were taken from UniProt (www.uniprot.org). We constructed three datasets. Dataset I contained 14 proteins and was used to test the methods (Dovidchenko *et al.*, 2009); Dataset II contained 21 proteins and was created to check the quality of our method and to present a novel server for the prediction of residues protected/unprotected against HD exchange from solely the amino acid sequence; Dataset III included 10 proteins and was used to compare the predictions with experimental mass spectrometry results (Suvorina *et al.*, 2012). For proteins from Datasets I and II, we used experimental results obtained for amide backbone protons only, while for Dataset III all exchangeable protons were considered. In case there were several sets of experimental data available for one protein, we preferred the set obtained in conditions closest to native.

Dataset I contained the data for the following proteins (pdb codes are given in parenthesis): α-Lactalbumin from *Bos taurus* (1F6R) (Wijesinha-Bettoni *et al.*, 2001), Cardiotoxin analogue III (CTX III) from *Naja naja atra* (1H0J) (Sivaraman *et al.*, 2000), Antibody Fragment from *Lama glama* (1HCV) (Perez *et al.*, 2001), Cytochrom C from *Equus caballus* (1HRC) (Milne *et al.*, 1998), Ovomucoid Third Domain from *Meleagris gallopavo* (1PPF) (Swint-Kruse and Robertson, 1996), SH3 domain from α-Spectrin from *Gallus gallus* (1SHG) (Sadqi *et al.*, 1999), Staphylococcal Nuclease (1SNO) (Loh *et al.*, 1993), cobrotoxin from *N. naja atra* (1V6P) (Sivaraman *et al.*, 2000), chymotripsin inhibitor 2 from *Hordeum vulgare* (2CI2) (Itzhaki *et al.*, 1997), Lysozyme from *Equus caballus* (2EQL) (Morozova *et al.*, 1995), Ribonuclease H from *Escherichia coli* (2RN2) (Chamberlain *et al.*, 1996), CheY from *E.coli* (3CHY) (Lacroix *et al.*, 1997), ferricytochrome c551 from *Pseudomonas aeruginosa* (451 C) (Russell *et al.*, 2003), Bovine Pancreatic Trypsin Inhibitor (6PTI) (Kim *et al.*, 1993). The sequences with exchange-protected amino acid residues, as shown experimentally, are given in Supplementary Dataset 1 and all predicted data are given in Data_for_Dataset1.

The data for the following proteins were collected in Dataset II: Kinase interaction domain from Arabidopsis thaliana (1MZK) (data taken from a database with the order number 5841, http://www.bmrb.wisc.edu), 15.5KD RNA-binding protein (1E7K) (database order number 15445, http://www.bmrb.wisc.edu), YOJN histidine-phosphotransferase (HPT) domain from *E.coli* (1SR2) (Rogov *et al.*, 2004), cellular retinol-binding protein type-I (1JBH) (Franzoni *et al.*, 2010), cellular retinol-binding protein type-II (1OPA) (Franzoni *et al.*, 2010), Tendamistat (1OK0A) (Qiwen *et al.*, 1987), Protein L (2PTL) (Yi and Baker, 1996), Protein G (1PGB) (Orban *et al.*, 1995), Interleukin-1β from (1HIB) (Varley *et al.*, 1993), IGA-kappa MCP603 FAB Heavy chain from *Mus Musculus* (2MCP) (Freund *et al.*, 1997), IGA-kappa MCP603 FAB Light chain (2MCP) (Freund *et al.*, 1997), Ribonuclease T1 from *Aspergillus oryzae* (9RNT) (Mullins *et al.*, 1997), Barnase from *Bacillus amyloliquefaciens* (1A2P) (Bycroft *et al.*, 1990), human erythrocytic ubiquitin (1UBQ) (Pan and Briggs, 1992), PSE-4 beta-lactamase from *P. aeruginosa* (1G68) (Morin *et al.*, 2009), PTP-BL second PDZ Domain from *M. musculus* (1OZI) (Walma *et al.*, 2004), dihydrofolate reductase (apo-form) from

Lactobacillus casei (2L28) (Feeney *et al.*, 2011), small archaeal modifier protein 1 (SAMP1) from *Methanosarcina acetivorans* (2L52) (Ranjan *et al.*, 2011), c-terminal domain of a multiprotein bridging factor 1 (MBF1) from *Trichoderma reesei* (2JVL) (Salinas *et al.*, 2009), component IV monomeric hemoglobin-CO from *Glycera dibranchiata* (1VRE) (Volkman *et al.*, 1998) and protein A from *Staphylococcus aureus* (1BDD) (Bai *et al.*, 1997). The sequences with exchange-protected amino acid residues, as shown experimentally, are given in Supplementary Dataset 2.

The amino acid sequences of 10 proteins included in Dataset III are listed in Table 3. The identities of proteins used in the different sets are given in the Supplementary material, H-D_exchange.xls.

Mass spectrometry and tandem mass spectrometry analysis of amino acids: HD exchange was initiated by diluting amino acids (Sigma, USA) in 98% D_2O (pD = 3.0). The incubation time was one day. The solution contained 10^{-9} mol/μl amino acid. After HD exchange, the samples were analyzed on an ion trap LCQ Deca XP Plus (Thermo Finnigan, USA) using electrospray ionization. The sample volume varied from 10 to 15 μl. The flow rate did not exceed 2 μl/min. Mass spectra were accumulated at positive ions mode from 50 to 220 m/z, at capillary voltage 1.7–2.3 kV, and entrance capillary temperature of 220°C. Fragmentation of amino acids was performed by collision-activated dissociation. MS/MS–spectra were registered at isolation width of the parent ion of 10 m/z. The normalized collision energy varied from 25% to 30%. The average number of exchangeable protons was calculated using the following equation:

$$<n> = \sum_{i=1}^{8} i \cdot P_i/100 \qquad (7)$$

where *i* is the number of exchangeable protons (the maximal value can be 8 for arginine), P_i is the probability for the state with *i* exchangeable protons.

Mass spectrometric analysis of protein: We used 20 mM solution of ammonium acetate (Panreac), pH 6.8, as a buffer system to prepare protein solutions. All buffer solutions were prepared using deionized water and deuterated water (D_2O content was 98%).

The proteins were desalted by dialysis against a buffer solution of 20 mM ammonium acetate, pH 6.8. We used a dialysis membrane of 10 kDa (Serva, Germany) for dialysis of protein samples.

The reaction of HD exchange in proteins started with dilution (1:5) of D_2O-containing buffer. The remaining non-deuterium water was removed using gel filtration on microcolumns produced in the laboratory. Sefadex G-10 (40–120 µm granules, Pharmacia Fine Chemicals, Sweden) was used as a matrix for microcolumn packing. Protein solution (5 µl) with a concentration from 1 to 3 mg/ml was applied onto a column, equilibrated with 10 mM ammonium acetate, pH 6.8, prepared using deuterium water. The columns were centrifuged at 2000 rpm for 2–5 min (Eppendorf, Germany). The average number of protons that were subject to HD exchange was determined by mass spectrometry.

The mass spectrometric analysis of proteins was performed using an ion trap LCQ Deca XP Plus (Thermo Finnigan, USA). Ionization of samples was carried out by direct infusion of nanoelectrospray. The sample volume varied from 10 to 15 µl. The voltage supplied to capillary varied from 1 to 1.5 kV. The temperature of the entrance capillary varied from 220 to 245°C. The mass spectra were registered at positive ions from 500 to 2000 m/z and from 500 to 4000 m/z. The coarse resolution and low dynamic range of the used mass spectrometer led to increased reading spread for ions with a low relative concentration. Calculation of the average mass using such peaks could result in an increased calculation error. Therefore, to calculate the average mass, we used peaks with amplitudes of >60%. At the same time, we used all significant peaks for calculation of the charge state of ions with the corresponding mass-to-charge ratio (m/z). The error of measurements was ~5%.

3 RESULTS

3.1 Assessment of the method for 35 proteins

On the assumption that an amide proton exchanges for a deuterium, provided it is sufficiently flexible and accessible to solvent, we considered the number of inter-residue contacts and H-bonding energy to be indicative of such a possibility. The number of inter-residue contacts shows how tightly the amino acid residue in question is surrounded by other amino acid residues, thereby revealing the extent of accessibility of the amide proton to solvent (the more contacts, the less accessibility). H-bonding energy points to the presence and strength of inter-residue H-bonds; hydrogen involved there does not participate in HD exchange. Besides, we sought for characteristics of amino acid residues derived not from the protein 3D structure but solely from their sequence. Specifically, we used the expected number of contacts per residue (Galzitskaya *et al.*, 2006) and the probability of H-bonding (Savitski *et al.*, 2007).

We were the first to compile in two datasets HD exchange data for 35 proteins with known rates of the native state out-exchange (Supplementary datasets). One of these (Dataset I) included 14 single-domain proteins or separate domains of multi-domain proteins; it served for finding the threshold of the method used here. The proteins involved were not large in size (from 56 to 155 amino acid residues) and contained 563 amino acid residues protected against HD exchange, 667 unprotected ones and 110 residues for which the degree of protection was not determined (these were mainly proline residues without the amide proton). Solely on the basis of the amino acid sequence, we determined H-bond predictor sensitivity and specificity as 0.59 and 0.65, respectively (Fig. 1a) and those of the contact predictor as 0.69 and 0.54 (Fig. 1c).

Figure 1a shows the threshold-depending changes in H-bonding predictor sensitivity and specificity. It should be noted that we did not choose the optimal value of the threshold. For the H-bonding predictor, the maximal Q_2 coincides with the threshold determined previously for prediction of amyloidogenic regions in the members of another database containing 144 amyloid forming peptides and 263 non-amyloid forming peptides (see our server FoldAmyloid (Garbuzynskiy *et al.*, 2010) at http://bioinfo.protres.ru/fold-amyloid/oga.cgi), which points to applicability of this threshold to different tasks.

For the predictor of expected number of contacts, the maximum is wide (Fig. 1c); a threshold of 20.4 practically coincides not only with this maximum but also with the threshold used in search for protein disordered regions by FoldUnfold (Galzitskaya *et al.*, 2006) (see our server http://bioinfo.protres.ru/ogu/).

According to our assumption, flexible regions, which—due to their flexibility—can contact with solvent, are accessible for HD exchange; regions with irregular secondary structure can be attributed to such flexible regions; therefore, the type of the secondary structure points to amino acid residues that are not protected against HD exchange. Additionally, in water-soluble globular proteins (all proteins in our datasets are like that), the regions with irregular secondary structure, as a rule, are situated on the surface of a protein globule, i.e. are accessible to solvent. When predicting HD exchange-protected/unprotected residues from the secondary structure, we considered those involved in regular structural elements (α-helices, β-strands) as protected,

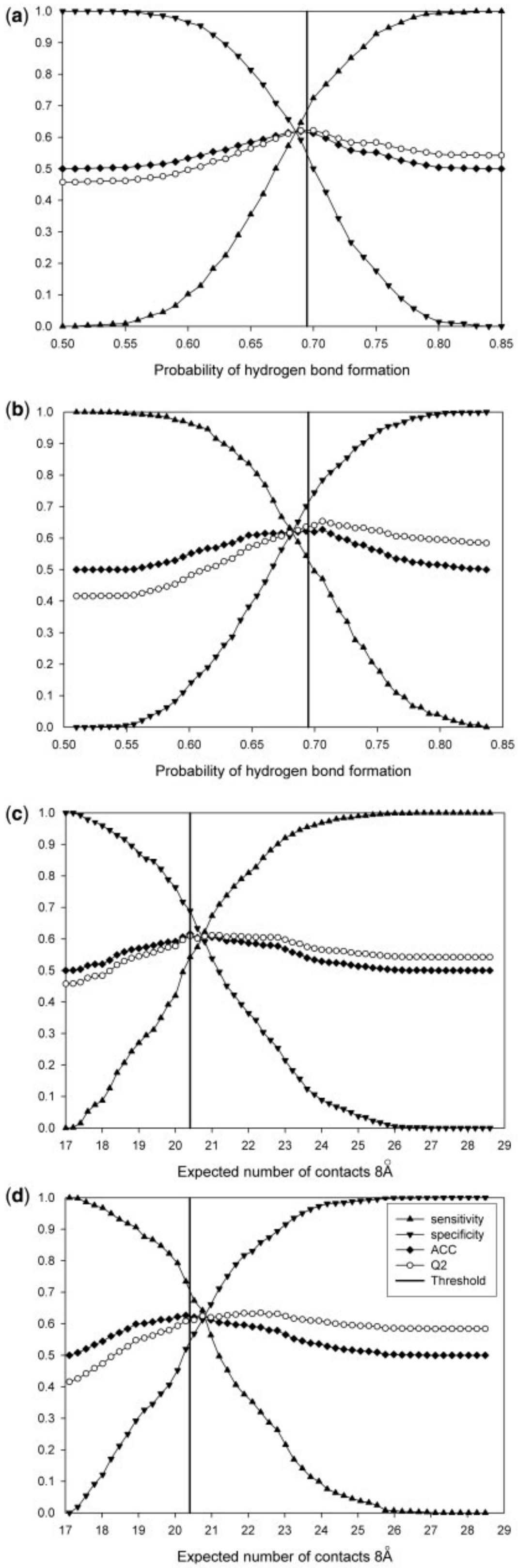

Fig. 1. Threshold-dependent changes in Q_2, ACC, sensitivity and specificity for Dataset I (**a**) and Dataset II (**b**) as predicted by the H-bonding–based method. The vertical line at 0.695 separates protected (upper values) and unprotected residues. The same changes as predicted by the contact-based method: (**c**), Dataset I; (**d**), Dataset II. The vertical line at 20.4 separates protected (upper values) and unprotected residues. The averaging frame size was 3

and the ones from irregular regions as unprotected. The type of the secondary structure was predicted using the standard program PsiPred (Jones, 1999). The sensitivity and specificity for this predictor are 0.65 and 0.62, respectively.

We created an artificial predictor as a fraction of protected residues for each type of amino acid residues using 14 proteins from Dataset I.

In Dataset II, 992 amino acid residues are protected against HD exchange, 1393 amino acid residues are unprotected and for 83 residues the degree of protection was not determined. For the H-bonding predictor, sensitivity and specificity are 0.53 and 0.71, respectively (Fig. 1b). Those for the contact predictor are 0.71 and 0.54, respectively (Fig. 1d). For the artificial predictor, the values are 0.58 and 0.63, respectively. Therefore, for the two datasets we obtained similar results.

We performed the cross validation test for all 35 proteins. The ACC, sensitivity and specificity proved to be practically the same for the whole set of 35 proteins and for each of two subsets of 18 and 17 proteins taken at random (Table 1 and Supplementary Material, H-D_exchange.xls).

A schematic representation of predictions made from the amino acid sequence is shown in Figure 2 for protein CI2. As follows from experimental data, some loop regions are capable of HD exchange, while others are not. Specifically, among the latter we see the N-terminal loop (amino acid residues 27–30) of CI2. Thus, loops can be divided into available (flexible) and unavailable (rigid) for HD exchange.

3.2 Mass spectrometry data for amino acids

Until now there have been no experimental data about HD exchange for free amino acids. We bridged this gap using mass spectrometry. The tandem mass spectrometry technique allowed neutralizing contribution of the N-terminal group -NH_3 to HD exchange. Fragments lacking -NH_3 were obtained for arginine, asparagine, glutamine, lysine, tyrosine and tryptophan (Table 2). For all of them, relative distribution of the number of exchangeable protons was calculated (Table 2). Statistical distribution of HD exchange was calculated taking into account contribution of natural isotopes. Fragments without the C-terminal group were obtained for aspartic and glutamic acids, threonine, serine and histidine. It should be noted that hydrogens belonging to atom N exchange with higher probability than those of atom O. For glutamic and aspartic acids without C-terminus, as well as for lysine without -NH_3, the average number of exchangeable protons is close to zero. Moreover, asparagine showed a smaller average number of exchangeable protons than glutamine (Table 2). This may be explained by the difference in their structures: glutamine has an additional CH_2 group in its side chain. This subtle difference gives considerable distinctions in their biological activities.

For aspartic acid, glutamic acid, histidine, serine and threonine, the average number of exchangeable protons was

Table 1. Accuracy of predictions made by H-protection predictors

Dataset	Predictor	ACC	Sensitivity	Specificity
All (35)	H-bonding predictor	0.62	0.54	0.71
All (35)	Contact predictor	0.62	0.70	0.54
All (35)	Psipred	0.66	0.72	0.61
All (35)	Artificial	0.62	0.60	0.63
Subset I (18)	H-bonding predictor	0.61	0.51	0.72
Subset I (18)	Contact predictor	0.61	0.69	0.54
Subset I (18)	Psipred	0.66	0.71	0.61
Subset I (18)	Artificial	0.60	0.58	0.62
Subset II (17)	H-bonding predictor	0.63	0.56	0.71
Subset II (17)	Contact predictor	0.62	0.71	0.54
Subset II (17)	Psipred	0.65	0.69	0.62
Subset II (17)	Artificial	0.63	0.61	0.64

Table 2. Probability of the state with *i* exchangeable protons, the average number of exchangeable protons, and root-mean-square deviation

Amino acid	$i=0$	$i=1$	$i=2$	$i=3$	$<n>$	$<s>$
Arg/withoutNH_3	0	3	30	67	2.6	0.5
Asn/withoutNH_3	46	40	14	0	0.7	0.7
Asp/$(CO+H_2O)$	97	3	0	0	0.0	0.2
Gln/withoutNH_3	0	33	63	4	1.7	0.5
Glu/$(CO+H_2O)$	98	2	0	0	0.0	0.1
His/$(CO+H_2O)$	36	52	12	0	0.8	0.6
Lys/withoutNH_3	87	13	0	0	0.1	0.3
Ser/$(CO+H_2O)$	48	39	13	0	0.7	0.7
Thr/$(CO+H_2O)$	45	29	25	1	0.8	0.8
Tyr/withoutNH_3	60	38	2	0	0.4	0.5
Trp/withoutNH_3	11	39	41	9	1.5	0.8

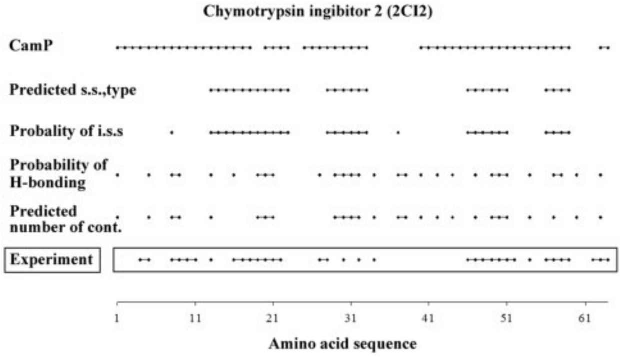

Fig. 2. Comparison of predictions made by different methods based on the amino acid sequence for chymotrypsin inhibitor (2CI2). Regions protected against HD exchange are marked with horizontal lines. S.s, secondary structure; i.s.s, irregular secondary structure

determined in the presence of the terminal group -NH3, although the C-terminus was lacking. To exclude the influence of the both termini, another set of mass spectrometry experiments on fragments with modified termini is required.

3.3 Comparison of mass spectrometry and prediction data for 10 proteins

Our additional goal was to assess whether the obtained bioinformatics data were applicable for predicting the behavior of proteins in solution. For this purpose, we made a mass spectrometry analysis of HD exchange in 10 proteins and compared its results with our predictions.

Proton and deuterium occupations of amide and indole groups were characterized by nuclear magnetic resonance to show the behaviour of separate amino acid residues in terms of HD exchange analyzed by mass spectrometry (Suvorina *et al.*, 2012). The increments of the average mass in D_2O were measured after 15 min and 1 day incubation of 10 protein samples (Table 3).

Here we present the recalculated numbers of protons exchanged for deuterium in side chains of amino acid residues predicted as unprotected. To do so, we compared the predicted numbers with mass spectrometry results first obtained for free amino acids (Table 2). Specifically, our mass spectrometry experiments showed that lysine predicted to exchange three protons had zero exchangeable protons, while asparagine and glutamine had one and two, respectively, instead of predicted two for each of them. Glutamic acid and aspartic acid are believed to exchange one proton each, but the extremely low probability of the event makes us consider the number as zero (Table 2). It should be noted that arginine capable of exchanging five protons in reality exchanges about 3, as shown experimentally. Serine, threonine, tryptophan, tyrosine and histidine are to exchange one proton each, whereas in reality we observed one exchangeable proton per side group.

The number of deuteration sites predicted by our method is in good agreement with the experimentally found protein mass increment in the course of HD exchange (Table 3). For the H-bonding predictor, the correlation coefficient is 0.73; for the contact predictor, it is 0.47; for the artificial predictor, it is 0.66, and prediction from the secondary structure had it equal to 0.33. It should be noted that the correlation coefficients were calculated for normalized protein lengths to exclude the length effect on the correlation. The theoretical results based on H-bonding are in a better agreement with mass spectrometry measurements of increased protein mass after long incubation in D_2O than those based on the expected number of contacts. However, the latter are closer to experimental results obtained after 15 min incubation in D_2O. So, exchangeable protons from the surface of a globule or unstructured regions of the chain are better predicted by the contact predictor. In its turn, prediction based on the H-bonding statistics is more accurate for exchangeable protons from the groups in contact with the solvent during a long-term incubation.

4 THE H-PROTECTION SERVER

To predict the status (protected/unprotected against HD exchange) of a residue and the number of exchangeable protons on the basis of solely the amino acid sequence of a protein, we use three predictors: the expected number of contacts, the probability of H-bonding within the backbone amide group and an

Table 3. Number of exchangeable protons in 10 proteins as predicted by four predictors and obtained from mass spectrometry experiments (Suvorina *et al.*, 2012)

Protein, mass, Da	Predicted: contact predictor	Predicted: H-bonding predictor	Predicted: artificial predictor	Predicted: PsiPred	Experimental number of exchangeable protons: 15 min/24 h
Bovine beta-lactoglobulin, 18361	82	112	92	86	117/130
Bovine carbonic anhydrase B, 29086	166	239	207	190	212/265
Human lactalbumin, 14112	80	103	86	107	84/110
Bovine lactalbumin, 14221	80	103	94	96	90/117
Whale apomyoglobin, 17335	91	119	112	66	96/133
Human proinsulin, 9390	53	71	69	83	62/120
GroES *E.coli*, 10386	72	83	83	71	88/96
Horse Cytochrome c, 12359	77	88	83	77	87/132
GFP *A.Victoria*, 26308	136	206	189	144	192/214
Hen egg lysozyme, 14310	84	131	112	106	134/146

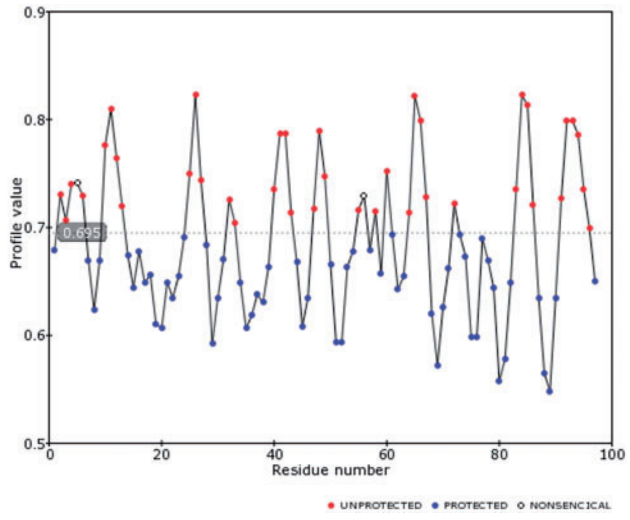

Fig. 3. Profiles of H-bonding probability for GroES. Prediction was made using the H-bonding predictor. The dashed line at 0.695 (y-axis) shows the threshold line for residues to be protected

artificial predictor. When using any of these predictors, the appropriate value was assigned to the amino acid residue in question. Then these values were averaged using a shifting window. The H-protection web server is available at http://bioinfo.protres.ru/ogp/. Using the entered amino acid sequence, the server produces a profile of expected contacts or that of probable H-bonding for each residue to be described in terms of its protection against HD exchange. Residues with a number of predicted contacts above the cut-off point (20.4 intraprotein contacts per residue) are shown as protected, and those with a smaller number as unprotected. Similarly, residues with H-bonding probability above the cut-off point (0.695) are predicted as protected, while those with a probability below this threshold are shown as unprotected against HD exchange (Fig. 3).

To calculate the number of exchangeable protons of unprotected residue side groups, we used the tabulated experimental results obtained for pH = 3 (Table 2). With the effect of the termini on this number taken into account, and only integer numbers used, we found that arginine is able to exchange three protons, glutamine—two, asparagine, histidine, tryptophan, serine, threonine, tyrosine—one, glutamic acid, aspartic acid and lysine—zero. Assessment of prediction accuracy allows concluding that the number of protected protons is better predicted from the expected number of contacts (sensitivity is 0.73), while that of unprotected protons from H-bonding probability (specificity is 0.71). We obtained the number of exchangeable protons for all amino acids at pH = 7 (see our server).

Prediction of unfolded and flexible regions of the protein chain reveals regions that must be prone to strong structural fluctuations, and therefore, subjects to HD exchange. Similar to prediction from the protein secondary structure, prediction of the number of exchangeable protons is somewhat limited by protein topology that has a dominant effect over the exchange kinetics (Craig *et al.*, 2011).

Funding: This study was supported in part by the Russian Foundation for Basic Research (grant no. 11-04-00763), Russian Academy of Sciences (programs ‘Molecular and Cell Biology’ 01200959110 and 01201353567) and ‘Fundamental Sciences to Medicine’.

Conflict of Interest: none declared.

REFERENCES

Bai,Y. *et al.* (1997) Absence of a stable intermediate on the folding pathway of protein A. *Protein Sci.*, **6**, 1449–1457.

Bycroft,M. *et al.* (1990) Detection and characterization of a folding intermediate in Barnase by NMR. *Nature*, **346**, 488–490.

Chamberlain,A.K. *et al.* (1996) Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH. *Nat. Struct. Biol.*, **3**, 782–787.

Craig,P.O. *et al.* (2011) Prediction of native-state hydrogen exchange from perfectly funneled energy landscapes. *J. Am. Chem. Soc.*, **133**, 17463–17472.

Dovidchenko,N.V. and Galzitskaya,O.V. (2008) Prediction of status residue to be protected or not protected from hydrogen exchange using amino acid sequence only. *Open Biochem. J.*, **2**, 77–80.

Dovidchenko,N.V. *et al.* (2009) Prediction of amino acid residues protected from hydrogen-deuterium exchange in a protein chain. *Biochemistry*, **74**, 1091–1102.

Englander,S.W. (2006) Hydrogen exchange and mass spectrometry: a historical perspective. *J. Am. Soc. Spectrom.*, **17**, 1481–1489.

Feeney,J. *et al.* (2011) NMR structures of Apo L. casei dihydrofolate reductase and its complexes with trimethoprim and NADPH: contributions to positive cooperative binding from ligand-induced refolding, conformational changes, and interligand hydrophobic interactions. *Biochemistry*, **50**, 3609–3620.

Freund,C. *et al.* (1997) Comparison of the amide proton exchange behavior of the rapidly formed folding intermediate and the native state of an antibody scFv fragment. *FEBS Lett.*, **407**, 42–46.

Franzoni,L. *et al.* (2010) New insights on the protein-ligand interaction differences between the two primary cellular retinol carriers. *J. Lipid Res.*, **51**, 1332–1343.

Fontana,A. (1988) Structure and stability of thermophilic enzymes. Studies on thermolysin. *Biophys. Chem.*, **29**, 181–193.

Galzitskaya,O.V. *et al.* (2006) FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics*, **22**, 2948–2949.

Garbuzynskiy,S.O. *et al.* (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*, **26**, 326–332.

Itzhaki,S.L. *et al.* (1997) Hydrogen exchange in chymotrypsin inhibitor 2 probed by denaturants and temperature. *J. Mol. Biol.*, **270**, 89–98.

Jacobs,D.J. *et al.* (2001) Protein flexibility predictions using graph theory. *Proteins*, **44**, 150–165.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kim,K.S. *et al.* (1993) Hydrogen exchange identifies native-state motional domains important in protein folding. *Biochemistry*, **32**, 9600–9608.

Lacroix,E. *et al.* (1997) Amide hydrogen exchange and internal dynamics in the chemotactic protein CheY from *Escherichia coli*. *J. Mol. Biol.*, **271**, 472–487.

Loh,S.N. *et al.* (1993) Hydrogen exchange in unligated and ligated staphylococcal nuclease. *Biochemistry*, **32**, 11022–11028.

Malmstroem,L. *et al.* (2009) On the use of hydrogen/deuterium exchange mass spectrometry data to improve de novo protein structure prediction. *Rapid Commun. Mass Spectrom.*, **23**, 459–461.

Milne,J.S. *et al.* (1998) Determinants of protein hydrogen exchange studied in equine cytochrome c. *Protein Sci.*, **7**, 739–745.

Morin,S. *et al.* (2009) NMR dynamics of PSE-4 beta-lactamase: an interplay of ps-ns order and mus-ms motions in the active site. *Biophys. J.*, **96**, 4681–4691.

Morozova,L.A. *et al.* (1995) Structural basis of the stability of a lysozyme molten globule. *Nat. Struct. Biol.*, **2**, 871–875.

Mullins,L.S. *et al.* (1997) Conformational stability of ribonuclease T1 determined by hydrogen-deuterium exchange. *Protein Sci.*, **6**, 1387–1395.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Orban,J. *et al.* (1995) Assessment of stability differences in the protein G Bl and B2 domains from hydrogen-deuterium exchange: comparison with calorimetric data. *Biochemistry*, **34**, 15291–15300.

Pan,Y. and Briggs,M.S. (1992) Hydrogen exchange in native and alcohol forms of ubiquitin. *Biochemistry*, **31**, 11405–11412.

Perez,J.M. *et al.* (2001) Thermal unfolding of a llama antibody fragment: a two-state reversible process. *Biochemistry*, **40**, 74–83.

Qiwen,W. *et al.* (1987) Amide proton exchange in the a-Amylase polypeptide inhibitor tendamistat studied by two-dimensional 1H nuclear magnetic resonance. *Biochemistry*, **26**, 6488–6493.

Ranjan,N. *et al.* (2011) Solution structure and activation mechanism of ubiquitin-like small archaeal modifier proteins. *J. Mol. Biol.*, **405**, 1040–1055.

Rogov,V.V. *et al.* (2004) Solution Structure of the *Escherichia coli* YojN Histidine-phosphotransferase domain and its interaction with cognate phosphoryl receiver domains. *J. Mol. Biol.*, **343**, 1035–1048.

Rose,G.D. *et al.* (1985) Turns in peptides and proteins. *Adv. Protein Chem.*, **37**, 1–109.

Russell,B.S. *et al.* (2003) Backbone dynamics and hydrogen exchange of *Pseudomonas aeruginosa* ferricytochrome c(551). *J. Biol. Inorg. Chem.*, **8**, 156–166.

Sadqi,M. *et al.* (1999) The native state conformational ensemble of the SH3 domain from alpha-spectrin. *Biochemistry*, **38**, 8899–8906.

Salinas,R. *et al.* (2009) Solution structure of the C-terminal domain of multiprotein bridging factor 1 (MBF1) of *Trichoderma reesei*. *Proteins*, **75**, 518–523.

Savitski,M.M. *et al.* (2007) Backbone carbonyl group basicities are related to gas-phase fragmentation of peptides and protein folding. *Angew. Chem. Int. Ed. Engl.*, **46**, 1481–1484.

Sivaraman,T. *et al.* (2000) Comparison of the structural stability of two homologous toxins isolated from the Taiwan cobra (Naja naja atra) venom. *Biochemistry*, **39**, 8705–8710.

Skinner,J.J. *et al.* (2012) Protein hydrogen exchange: testing current models. *Protein Sci.*, **7**, 987–995.

Suvorina,M.Y. *et al.* (2012) Comparison of experimental and theoretical data on hydrogen-deuterium exchange for ten globular proteins. *Biochemistry*, **77**, 758–766.

Swint-Kruse,L. and Robertson,A.D. (1996) Temperature and pH dependences of hydrogen exchange and global stability for ovomucoid third domain. *Biochemistry*, **35**, 171–180.

Tartaglia,G.G. *et al.* (2007) Prediction of local structural stabilities of proteins from their amino acid sequences. *Structure*, **15**, 139–143.

Varley,P. *et al.* (1993) Kinetics of folding of the all-beta sheet protein interleukin-1 beta. *Science*, **260**, 1110–113.

Volkman,B.F. *et al.* (1998) Solution structure and backbone dynamics of component IV glycera dibranchiata monomeric hemoglobin-CO. *Biochemistry*, **37**, 10906–10919.

Walma,T. *et al.* (2004) A closed binding pocket and global destabilization modify the binding properties of an alternatively spliced form of the second PDZ domain of PTP-BL. *Structure*, **12**, 11–20.

Westhof,E. *et al.* (1984) Correlation between segmental mobility and the location of antigenic determinants in proteins. *Nature*, **311**, 123–126.

Wijesinha-Bettoni,R. *et al.* (2001) Comparison of the structural and dynamical properties of holo and apo bovine alpha-lactalbumin by NMR spectroscopy. *J. Mol. Biol.*, **307**, 885–898.

Yi,Q. and Baker,D. (1996) Direct evidence for a two-state protein unfolding transition from hydrogen-deuterium exchange, mass spectrometry, and NMR. *Protein Sci.*, **5**, 1060–1066.