ISSN 0006-2979, Biochemistry (Moscow), 2009, Vol. 74, No. 8, pp. 888-897. © Pleiades Publishing, Ltd., 2009.
Published in Russian in Biokhimiya, 2009, Vol. 74, No. 8, pp. 1091-1102.
Originally published in Biochemistry (Moscow) On-Line Papers in Press, as Manuscript BM09-044, June 28, 2009.

Prediction of Amino Acid Residues Protected from Hydrogen–Deuterium Exchange in a Protein Chain

N. V. Dovidchenko, M. Yu. Lobanov, S. O. Garbuzynskiy, and O. V. Galzitskaya*

Institute of Protein Research, Russian Academy of Sciences, 142290 Pushchino,
Moscow Region, Russia; fax: (495) 632-7871; E-mail: ogalzit@vega.protres.ru

Received February 16, 2009
Revision received March 3, 2009

Abstract—We have investigated the possibility to predict protection of amino acid residues from hydrogen–deuterium exchange. A database containing experimental hydrogen–deuterium exchange data for 14 proteins for which these data are known has been compiled. Different structural parameters related to flexibility of amino acid residues and their amide groups have been analyzed to answer the question whether these parameters can be used for predicting the protection of amino acid residues from hydrogen–deuterium exchange. A method for prediction of protection of amino acid residues, which uses only the amino acid sequence of a protein, has been elaborated.

DOI: 10.1134/S0006297909080100

Key words: hydrogen–deuterium exchange, secondary structure, hydrogen bond, B-factor, regions with irregular secondary structure

Discussion of flexibility of protein molecules was not popular in the 1960s. And only at the end of 1960s, internal protein mobility was determined experimentally: the results of investigation of exchange of an amide proton to deuterium testified to the flexibility of globular proteins [1]; in experiments studying ^{13}C-NMR relaxation, mobility of the polypeptide backbone and side chains of amino acids was shown [2]. On the other hand, a comparison of crystallographic protein structures in different functional states revealed significant differences in conformations of protein molecules [3].

It was demonstrated that loops and β-turns are the most mobile regions in protein structure [4]. Such flexible regions can be epitopes or include binding sites of different enzymes, which regulate the life cycle of proteins *in vivo* [5, 6].

Some loops in protein structure have large amplitudes of fluctuations when the protein is in solution, whereas they can adopt a well-ordered structure (for example, β-hairpins) in a crystal structure. This can be a consequence of formation of additional contacts in the crystal structure (for example, this is observed in the structure of ribosomal protein S7) [7].

Experimental approaches giving information about internal motions of macromolecules have been developed for the last two decades. Among these, there are methods of fluorescence depolarization of tryptophan residues, NMR, inelastic neutron scattering, Mössbauer spectroscopy, infrared spectroscopy, and analysis of Debye–Waller factors obtained from crystallographic data.

To clarify the relationship between structural elements and polypeptide chain mobility, a set of static analyses of structures obtained from NMR experiments was prepared. Goodman et al. [8] demonstrated (using the experimental NMR data) that flexibility of a residue is connected with its size (the smaller an amino acid residue is, the larger is its mobility). In addition, it was shown that fluctuations of a given amide group also depend on the sizes of side chains of the amino acid residues which are its neighbors in the protein sequence [8].

In crystallography, the factors of Debye–Waller (further, B-factors) are used to describe fluctuations of atoms relative to their average positions in the crystal structure. Since the crystallographic data are averaged over time, the amplitudes of fluctuations include contributions of both temperature-dependent thermal motions and deviations from average positions of atoms as a result of disorder in the crystal itself:

$$\left\langle u_i^2 \right\rangle_{\text{total}} = \frac{3B_i}{8\pi^2} = \left\langle u_i^2 \right\rangle_{\text{thermal}} + \left\langle u_i^2 \right\rangle_{\text{disorder}}, \qquad (1)$$

where B_i is the crystallographic B-factor of atom i and u_i is the amplitude of its fluctuation. In practice, it is possi-

* To whom correspondence should be addressed.

ble to distinguish static and dynamic contributions in B-factors by collecting crystallographic data at different temperatures. The contribution of disorder of a crystal does not depend on temperature, whereas the square of thermal amplitudes of harmonic vibrations depends linearly on the absolute temperature [9].

If an experiment on hydrogen–deuterium exchange occurs under high concentration of a denaturing agent (for example, urea), then such an experiment is called global exchange (or experiment of the EX1 type), since practically all atoms of amides of the main chain are accessible for the exchange. If the exchange occurs under "native" conditions or if the concentration of denaturant is not sufficient for denaturation of the investigated protein, such exchange is called local (or experiment of the EX2 type). Such experiments are usually used for measuring of accessibility to a solvent of H-bonded amides of the protein in its native state (Fig. 1). In this case, the protection factor for the given residue from hydrogen–deuterium exchange is the ratio of the predicted exchange rate of the H-bonded amide on the deuterium of water in the unstructured peptide to the observed amide hydrogen exchange rate in the globular protein measured in the experiments [1, 10].

In theoretical works, attempts have been made to find structural characteristics of protein chain connected with its flexibility to predict the protection factor of amide protons. For example, the accessible surface of amino acid residues was considered as such a parameter [11]. The Gaussian network model (GNM) has been also suggested for the interpretation of experimental data obtained under local hydrogen–deuterium exchange [12]. Vendruscolo and coauthors suggested a phenomenological equation for prediction of protection factors [13, 14]. The equation includes two terms, which reflect the contribution of van der Waals contacts and hydrogen bonds. However, knowledge of the three-dimensional structure of the protein is necessary to solve this equation.

Simultaneously with the appearance and improvement of methods for predicting protection factors using

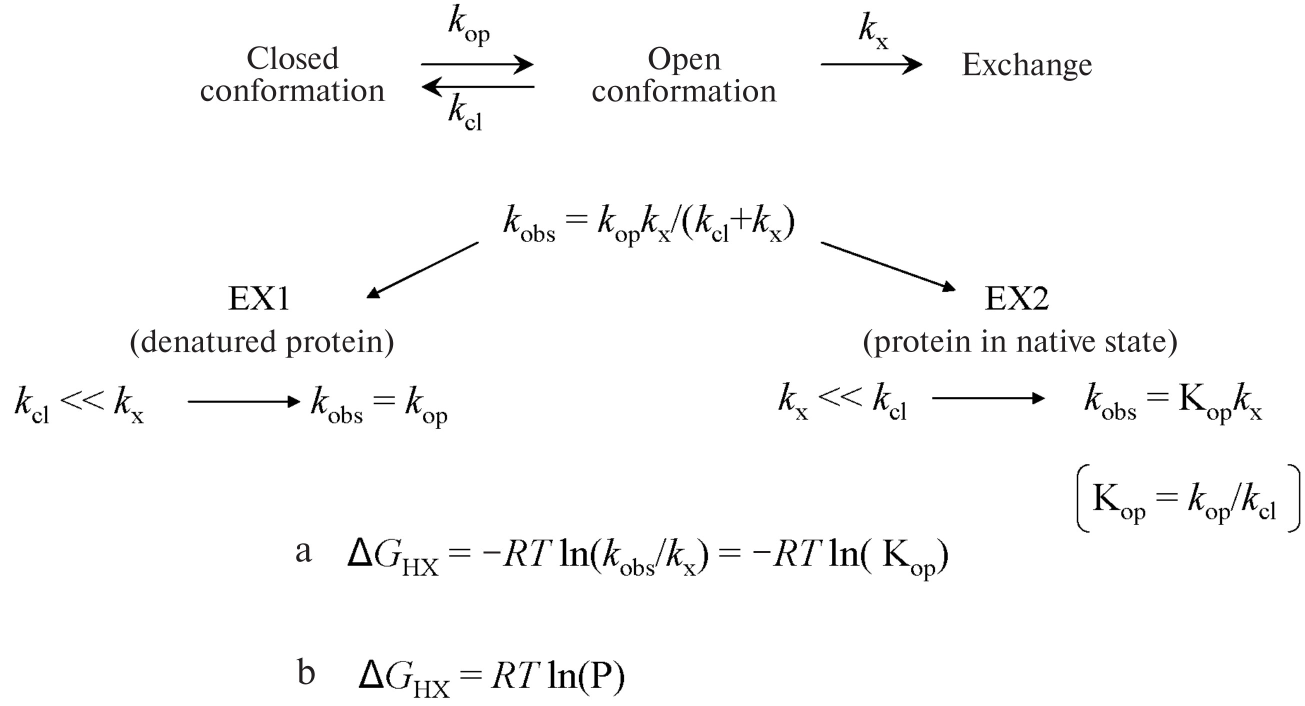
Closed conformation
k_{op}
k_{cl}
Open conformation
k_x
Exchange
$k_{obs} = k_{op}k_x/(k_{cl}+k_x)$
EX1
(denatured protein)
EX2
(protein in native state)
$k_{cl} << k_x \longrightarrow k_{obs} = k_{op}$
$k_x << k_{cl} \longrightarrow k_{obs} = K_{op}k_x$
$\left(K_{op} = k_{op}/k_{cl}\right)$
a $\Delta G_{HX} = -RT\ln(k_{obs}/k_x) = -RT\ln(K_{op})$
b $\Delta G_{HX} = RT\ln(P)$

Fig. 1. Schematic diagram of estimation of the rate of hydrogen–deuterium exchange. The scheme of the process can be presented in the following way: before the event of hydrogen–deuterium exchange *per se* which occurs with the rate k_x, there is equilibrium between conformations of an amide proton. The first conformation is called open (i.e. accessible for hydrogen–deuterium exchange), and the second is called closed (the exchange of hydrogen is hindered). Transition from the closed conformation to the open one proceeds with the rate k_{op}, and backwards with the rate k_{cl}. Then for the observed rate of hydrogen–deuterium exchange it is possible to write an equation based on the scheme. There are two kinds of experiments on hydrogen–deuterium exchange: EX1 and EX2. The experiment of type EX1 occurs with a sufficient amount of denaturant and deals with proteins in denatured state. In this case, the rate of the exchange is much higher than the rate of the transition to the closed conformation, whence it follows that the observed rate is equal to the rate of transition to the open conformation. In the experiment of type EX2, the rate of transition to the closed conformation is larger than the rate of the exchange, whence it follows that the observed rate equals to the product of the rate of the exchange and the constant of equilibrium of transition to the open conformation. From the experimental data, it is possible to calculate the change in the free energy (a), where the constant of equilibrium of transition to the open conformation will be under the sign of logarithm. If to put the minus under the sign of logarithm (b), one will obtain under the logarithm the value called a protection factor.

three-dimensional structure, methods in which only the amino acid sequence of the studied protein is necessary for the prediction have also been suggested. For example, the method CamP allows prediction of protection factors starting solely from the amino acid sequence of a protein [15]. This method is based on using a neural network.

We suggest that the absence of protection can be mostly explained by a large amplitude of fluctuations of the regions of a protein chain between packed elements of secondary structure that promotes the contact of the fluctuating region with solvent. Thus, the prediction of unfolded and loop regions should reveal regions prone to strong structural fluctuations and therefore subjected to hydrogen–deuterium exchange.

METHODS OF INVESTIGATION

Experimental data. According to literature data, we chose proteins for which the protection of separate amino acid residues from hydrogen–deuterium exchange was investigated experimentally. The amino acid residues with experimentally determined protection factors were divided (according to the corresponding papers) into two groups: protected and non-protected from hydrogen–deuterium exchange. If for some protein there were several sets of experimental data on protection of residues from hydrogen–deuterium exchange, the preference was made for the experiment carried out under conditions closest to native ones. In this work, we used experimental data for backbone amide protons only.

The amino acid sequences of the chosen proteins were taken from the UniProt database (www.uniprot.org).

The three-dimensional structures of these proteins in the native state were taken from the Protein Data Bank (PDB [16]). If there were several PDB-files for a given protein (the structure was determined several times), we preferred the structure determined by X-ray analysis with the best resolution, containing no ligands. If not the whole protein but a separate domain was investigated in the experiment of hydrogen–deuterium exchange, we preferred the structure in which this domain was also crystallized separately.

The data for the following proteins were entered in the created database (the PDB-codes of the spatial structures of these proteins are indicated in parentheses): α-lactalbumin from *Bos taurus* (1F6R) [17], cardiotoxin III (CTX III) analog from *Naja naja atra* (1H0J) [18], antibody fragment from *Lama glama* (1HCV) [19], cytochrome *c* from *Equus caballus* (1HRC) [20], ovomucoid third domain from *Meleagris gallopavo* (1PPF) [21], SH3 domain of α-spectrin from *Gallus gallus* (1SHG) [22], staphylococcal nuclease (1SNO) [23], cobrotoxin from *Naja naja atra* (1V6P) [18], chymotrypsin inhibitor 2 from *Hordeum vulgare* (2CI2) [24], lysozyme from *Equus caballus* (2EQL) [25], ribonuclease H from *Escherichia coli* (2RN2) [26], CheY from *E. coli* (3CHY) [27], ferricytochrome c_{551} from *Pseudomonas aeruginosa* (451C) [28], and bovine pancreatic trypsin inhibitor (6PTI) [29].

Parameters for predicting protection of a residue from hydrogen–deuterium exchange. For the prediction of protection of amino acid residues from hydrogen–deuterium exchange, the following structural parameters determined from the protein spatial structure were used: number of residue–residue contacts, hydrogen bond energy, type of secondary structure, and B-factors. For the prediction of protection starting from the amino acid sequence, the following parameters were used: predicted number of residue–residue contacts, predicted probability of hydrogen bond formation, type of predicted secondary structure, and probability of predicted irregular secondary structure.

Parameters obtained from the known spatial structure. *Number of contacts.* According to the protein three-dimensional structure, the number of contacts with other amino acid residues was calculated for each amino acid residue. Two residues are considered to have a contact if the distance between at least one pair of their non-hydrogen atoms is less than 8 Å.

Type of secondary structure. The type of secondary structure was determined using a standard program, DSSP [30]. Regions with regular secondary structure (α-helices, symbol H according to DSSP, and β-strands, symbol E according to DSSP) were considered as protected. The other amino acid residues, which had irregular secondary structure, were predicted as non-protected from hydrogen–deuterium exchange.

Energy of hydrogen bonds. The energy of hydrogen bonds formed by hydrogen atoms of amide groups of the main chain was also calculated using the DSSP program.

B-factors. B-factors for each atom were taken directly from the PDB-files. For prediction of protection of amino acid residues from hydrogen–deuterium exchange, we used B-factors either of all atoms of the amino acid residue or of only nitrogen atoms of the main chain (the hydrogen atom exchanged by deuterium in experiments is directly bonded to this nitrogen atom).

Parameters obtained from amino acid sequence. *Prediction of the type of secondary structure.* The type of secondary structure for each residue was predicted using PsiPred, a standard program for prediction of secondary structure by the amino acid sequence [31]. As in the case of the "real" secondary structure, we predicted a residue as protected from hydrogen–deuterium exchange if regular secondary structure (α or β, symbols H and E, respectively) was predicted for it; a residue was predicted as non-protected from hydrogen–deuterium exchange if irregular secondary structure was predicted for it (symbol C according to PsiPred).

Prediction of probability of irregular secondary structure. The probability of irregular secondary structure for each amino acid residue was also predicted using PsiPred [31].

Prediction of the probability of hydrogen bond formation based on amino acid sequence. Starting from the database of spatial structures (PDB), we collected a database of spatial structures of protein domains with amino acid sequence identity less than 25%, belonging to four general classes (α-proteins, β-proteins, α/β-proteins, α+β-proteins) according to the Structural Classification of Proteins (SCOP) [32]. For each amino acid residue, we determined (using DSSP) the presence of hydrogen bonds within the protein backbone (only the hydrogen bonds with energy below –0.5 kcal/mol were taken into account) [33]. Since we are interested in protection from hydrogen–deuterium exchange for the NH-group of the main chain, the hydrogen bond has been "ascribed" to the donor (i.e. to the residue possessing the NH-group which formed the given hydrogen bond). Since most inter-protein hydrogen bonds are formed within the main chain, we did not consider possible hydrogen bonds with side chains. The cases when an NH-group forms two or more hydrogen bonds ("fork") were not taken into account either; in such a case, we considered only the hydrogen bond with the best energy (according to DSSP). Further, the data were averaged over 20 types of amino acid residues, and the average probability with which a residue of a given type forms a hydrogen bond by its NH-group of the main chain with the main chain of protein was calculated. The calculated averaged probabilities of hydrogen bond formation by the NH-group of the residue are presented in Fig. 2. Under prediction based on the amino acid sequence of the protein, the average probability of hydrogen bond formation for the given type of residue was taken for each type of amino acid residues.

Prediction of expected number of contacts per residue. The scale of the expected number of contacts per residue is the statistics of the number of contacts per residue [34] obtained from the spatial structures. The process of creation of this scale was described in detail elsewhere [34, 35]. For creating this scale, the same database was used as for making the scale of probability of hydrogen bond formation. For each amino acid residue, the number of contacts with other residues was calculated. Two residues are considered as having a contact if at least one pair of their atoms is situated at a distance less than 8 Å. Further on, the average number of contacts for each of the 20 types of amino acid residues was calculated. The obtained 20 values are given in Fig. 3. Further, these 20 values were used by us as the scale under the prediction based on the amino acid sequence. Under prediction based on the amino acid sequence, the average number of contacts for the given type of residues was taken for each type of amino acid residues.

Criterion of evaluation of the quality of the prediction. To determine the accuracy with which a given parameter allows us to predict the protection of amino acid residues from hydrogen–deuterium exchange, we compared the results of our predictions with the experimental data. For estimation of the quality of the predictions, we used the following score:

$$Q_2 = (TP + TN)/N, \qquad (2)$$

where N is the number of all amino acid residues for which the experimental data on protection from hydrogen–deuterium exchange are available, TP ("true positives") is the number of amino acid residues correctly predicted as protected, and TN ("true negatives") is the number of amino acid residues correctly predicted as non-protected from hydrogen–deuterium exchange.

The error of averaging over proteins was calculated using the following equation:

$$\Delta = \frac{\sigma}{\sqrt{n}}\ , \qquad (3)$$

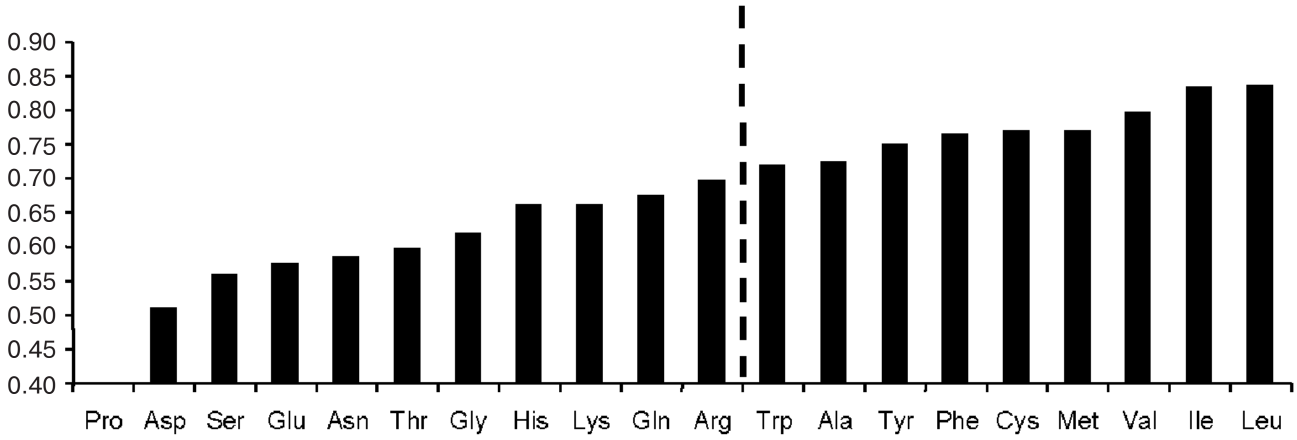

Fig. 2. Probability of hydrogen bond formation by an amino acid residue of each of the 20 types in globular protein. The dashed line corresponds to the threshold value, which is optimal for the prediction of protection of residues from hydrogen–deuterium exchange.

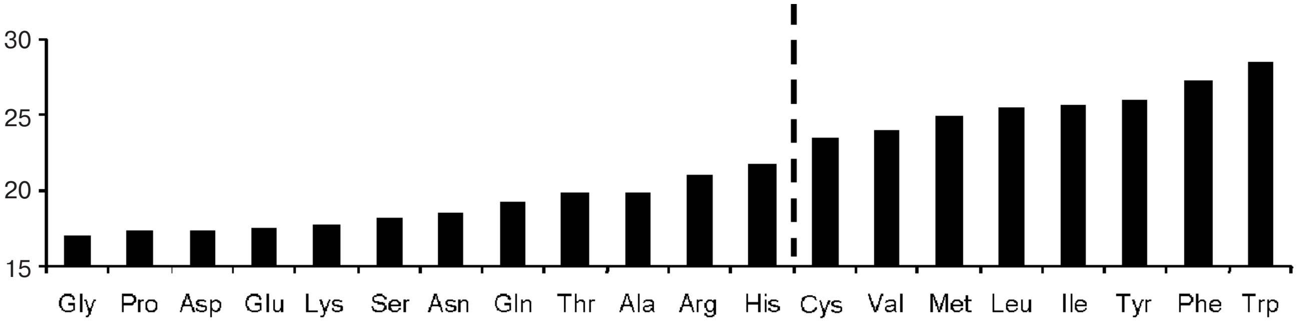

Fig. 3. Average number of contacts per amino acid residue in a globular protein. The dashed line corresponds to the threshold value, which is optimal for the prediction of protection of residues from hydrogen–deuterium exchange.

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 following equation:

$$\Delta = \frac{1}{N} \cdot \left(s_{TN}^2 + s_{TP}^2\right)^{1/2} = \frac{1}{N} \cdot$$

$$\cdot \left(\frac{TN \cdot (Nnp - TN)}{Nnp} + \frac{TP \cdot (Np - TP)}{Np} \right)^{1/2}, \quad (4)$$

where *Np* is the number of residues protected from hydrogen–deuterium exchange in the experiments and *Nnp* is the number of residues non-protected from hydrogen–deuterium exchange in the experiments.

RESULTS AND DISCUSSION

Database. Based on published experimental data, we selected such proteins for amino acid residues of which there are experimental data on hydrogen–deuterium exchange of amide protons of the main chain. The database composed by us contains information (experimental data on hydrogen–deuterium exchange for each amino acid residue, spatial structure of the protein in its native state, as well as the amino acid sequence of the protein) on 14 proteins. All of them are not large (from 56 to 155 amino acid residues long) single-domain proteins or separate domains of multi-domain proteins. In the whole database, 563 amino acid residues are protected from hydrogen–deuterium exchange, 667 amino acid residues are not protected, and for 110 residues protection is not determined (these are mainly proline residues which have no amide proton). Names of the proteins, which are included into our database, references to the corresponding experimental works, as well as PDB-entries of the spatial structures are listed in the "Methods of Investigation" section.

Choice of parameters allowing predicting protection of amide proton from hydrogen–deuterium exchange. According to our assumption, amide proton will exchange for deuterium if it is sufficiently flexible and is accessible to the solvent. Among the parameters that can indicate the possibility of a given hydrogen atom to exchange for deuterium, we chose the following four parameters: number of inter-residue contacts, energy of hydrogen bond, type of secondary structure, and B-factors. So, the number of inter-residue contacts shows how tightly the given amino acid residue is surrounded by other amino acid residues, of which one can presume the extent of accessibility of the amide proton to the solvent (the more is the number of inter-residue contacts – the less is the accessibility of the given residue to the solvent). Energy of hydrogen bond points to the presence and strength of inter-residue hydrogen bonds; here, being involved in hydrogen bond, the hydrogen is not exchanged for deuterium. It should be noted that the presence of intraprotein hydrogen bonds depends on the type of secondary structure (amide protons in the main chain of amino acid residues, which are included in regular secondary structure, form mainly intraprotein hydrogen bonds – in fact, regular secondary structure is stabilized also namely by these bonds [30, 33], – while amide protons in the main chain of amino acid residues with irregular secondary structure mainly form hydrogen bonds with solvent molecules). Besides, according to our assumption, flexible regions, which (due to their flexibility) can contact with solvent, are accessible for hydrogen–deuterium exchange; regions with irregular secondary structure can be attributed to such flexible regions; therefore, the type of secondary structure can indicate amino acid residues that are not protected from hydrogen–deuterium exchange. In addition, in water-soluble globular proteins (all proteins in our database are such), the regions with irregular secondary structure, as a rule, are situated on the surface of a protein globule [33], i.e. are contacting the solvent. B-factors indicate the amplitude of fluctuations of atoms, also showing their mobility as well as – indirectly – density of their surroundings and availability of the amide proton to the solvent.

Predicting protection of amino acid residues from hydrogen–deuterium exchange using information on protein spatial structure. Based on the spatial structure of a protein, the following parameters were determined for each amino acid residue: the number of intraprotein contacts (at a distance below 8 Å), hydrogen bond energy (for amide proton of the main chain), type of secondary structure (regular or irregular), and B-factor for each atom. Further on, each of these parameters was used for predicting protection of amino acid residues from hydrogen–deuterium exchange.

Upon predicting protection of an amino acid residue from hydrogen–deuterium exchange by hydrogen bond energy, residues which form (by the amide group of the main chain) strong hydrogen bonds were considered as protected from hydrogen–deuterium exchange, and the other residues were considered as non-protected ones. The cutoff value of hydrogen bond energy was optimized to obtain the best quality of predictions (the maximum of Q_2 parameter, see “Methods of Investigation”). The optimal cutoff value of hydrogen bond energy appeared to be –1.4 kcal/mol. In this case, the quality of predictions (averaged over all proteins) is 0.72 ± 0.02 (see table); in other words, the extent of protection from hydrogen–deuterium exchange was predicted correctly for 72% of amino acid residues. It should be noted right away that the hydrogen bond energy provides the best (among the parameters considered by us) quality of predicting protection of amino acid residues from hydrogen–deuterium exchange.

When predicting protection of amino acid residues from hydrogen–deuterium exchange by the number of intraprotein contacts, the amino acid residues with a greater number of intraprotein contacts than the cutoff value were considered as protected from hydrogen–deuterium exchange, and amino acid residues with a smaller number of intraprotein contacts were considered as non-protected ones. The cutoff value was optimized to obtain the best quality of predictions. When the cutoff value was 19.5 intraprotein contacts per residue, the quality of predictions was the best (the extent of protection from

Comparison of quality of predictions (Q_2) for methods of predicting the extent of protection of amino acid residues from hydrogen–deuterium exchange

PDB-entry	Methods that use information on spatial structure				Methods that use information on amino acid sequence only			
	hydrogen bond energy (DSSP)	number of contacts	type of the observed secondary structure (DSSP)	normalized B-factor of nitrogen atoms of the main chain	type of predicted secondary structure (PsiPred)	probability of irregular secondary structure (PsiPred)	probability of hydrogen bond formation	predicted number of contacts
1F6R	0.60	0.57	0.60	0.48	0.54	0.53	0.63	0.65
1H0J	0.58	0.55	0.56	0.53	0.45	0.47	0.58	0.58
1HCV	0.83	0.71	0.69	0.70	0.68	0.69	0.67	0.61
1HRC	0.73	0.75	0.60	0.55	0.60	0.60	0.62	0.60
1PPF	0.74	0.68	0.57	0.55	0.66	0.66	0.42	0.45
1SHG	0.78	0.84	0.73	0.67	0.78	0.78	0.78	0.73
1SNO	0.79	0.63	0.75	0.60	0.76	0.76	0.63	0.60
1V6P	0.65	0.70	0.72	0.60	0.70	0.70	0.58	0.58
2CI2	0.77	0.73	0.67	0.75	0.70	0.70	0.62	0.57
2EQL	0.76	0.68	0.63	0.61	0.55	0.59	0.58	0.59
2RN2	0.67	0.59	0.63	0.61	0.63	0.65	0.68	0.63
3CHY	0.69	0.60	0.60	0.50	0.61	0.59	0.59	0.68
451C	0.73	0.76	0.65	0.63	0.63	0.61	0.53	0.55
6PTI	0.79	0.63	0.67	0.63	0.63	0.62	0.65	0.67
Average over proteins	0.72 ± 0.02	0.67 ± 0.02	0.65 ± 0.02	0.60 ± 0.02	0.64 ± 0.02	0.64 ± 0.02	0.61 ± 0.02	0.61 ± 0.02
Average over residues	0.72 ± 0.01	0.66 ± 0.01	0.64 ± 0.01	0.59 ± 0.01	0.63 ± 0.01	0.63 ± 0.01	0.62 ± 0.01	0.61 ± 0.01

hydrogen–deuterium exchange was correctly predicted for 67% of amino acid residues, see the table).

Upon predicting protection of amino acid residues from hydrogen–deuterium exchange by the type of secondary structure, amino acid residues involved in regular secondary structure (α-helices, β-strands) were considered as protected, and amino acid residues involved in irregular secondary structure were considered as non-protected. The quality of prediction Q_2 for this method is 0.65 ± 0.02.

When predicting protection of amino acid residues from hydrogen–deuterium exchange by B-factors, we used several variants of prediction: by the average B-factor (averaged over the whole amino acid residue) and by B-factor of nitrogen of the main chain (since the exchanging hydrogen *per se* is usually not observed in the structures resolved by X-ray crystallography, we took B-factor of the covalently bound nitrogen atom that is observed in the structure), by B-factors directly taken from the PDB-entry, and by B-factors normalized to the average value of B-factor in the protein (the averaging was done either over all protein atoms or over all nitrogen atoms of the main chain). If the B-factor was greater than a cutoff value, the amino acid residue was predicted as protected; if the B-factor was below the cutoff value, the residue was predicted as non-protected. The cutoff value was also optimized. The best quality of prediction (the fraction of correctly predicted residues was 60%, see the table) was obtained when normalized B-factors for nitrogen atoms of the main chain were used, and the cutoff value was 1 (if the B-factor was smaller than the average value over all nitrogen atoms of the main chain, the residue was predicted as protected; if the B-factor was larger than the average, the residue was predicted as non-protected). However, one can see in the table that even the optimal prediction of protection of amino acid residues from hydrogen–deuterium exchange based on B-factors is somewhat worse compared to the predictions based on other parameters obtained from the protein spatial structure.

As seen from the table, the best quality of prediction is if the predictions are done based on the hydrogen bond energy. For other examined parameters (number of contacts, type of the observed secondary structure, B-factor), the quality of the predictions is somewhat worse (0.67 ± 0.02, 0.65 ± 0.02, and 0.60 ± 0.02, respectively).

Interestingly, B-factor has shown the worst results, which may be connected with different inaccuracies arising during its determination. For example, B-factors of nitrogen atoms of the main chain are shown in Fig. 4 for three different spatial structures of the same protein (the third domain of ovomucoid); one can see that even for the same protein, B-factor can considerably vary.

Predicting protection of amino acid residues from hydrogen–deuterium exchange using amino acid sequence only. As seen from the data described above, the knowledge of the hydrogen bond energy (or rather, determination of the hydrogen bond energy from the protein spatial structure) allows predicting the extent of protection of proton from hydrogen–deuterium exchange (the Q_2

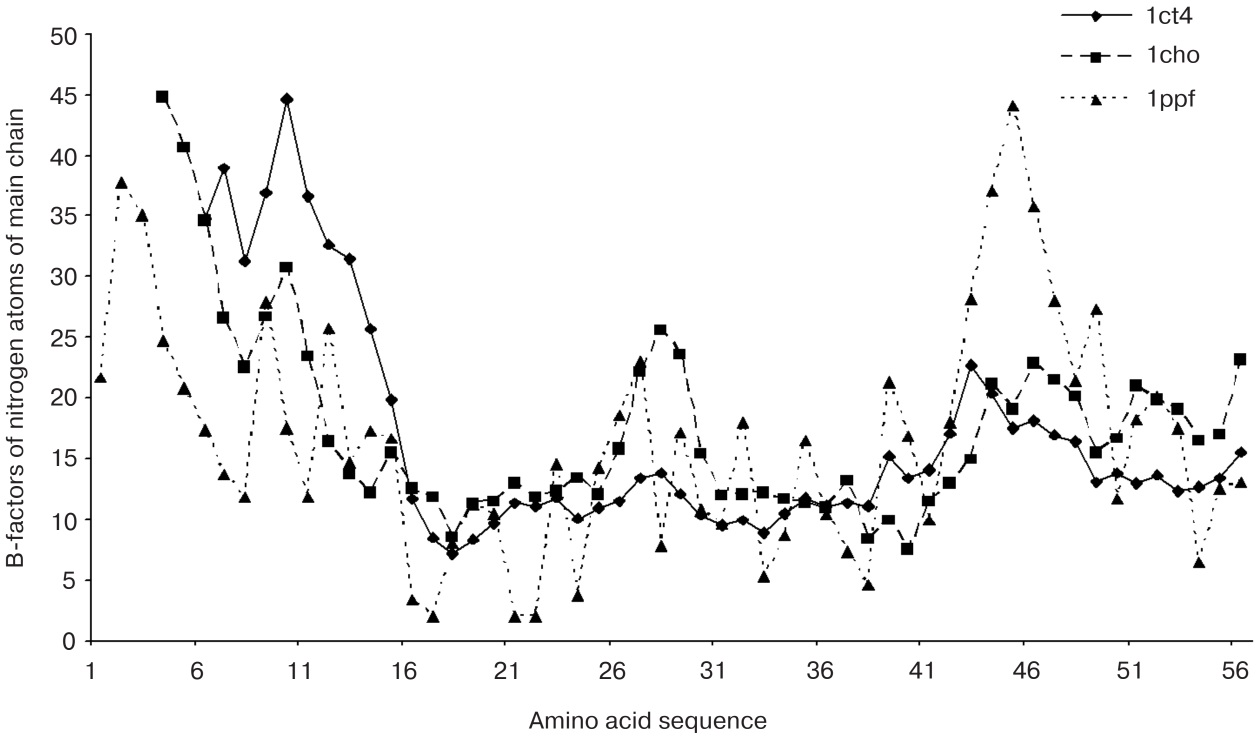

Fig. 4. Comparison of B-factors of nitrogen atoms of the main chain for ovomucoid third domain.

parameter, which is the fraction of correctly predicted amino acid residues, is 0.72 ± 0.02 on average). However, it is necessary to know the spatial structure of the predicted protein to make such a prediction. Since at present proteins with known spatial structures are much less numerous than proteins for which only the amino acid sequence is known, it is very important to be able to predict the extent of protection of proton from hydrogen–deuterium exchange starting only from the amino acid sequence of the protein. That is why we will consider parameters, for determination of which the knowledge of the spatial structure is not required, and only the knowledge of the amino acid sequence is necessary. As such parameters we took "analogs" of the parameters that were determined from the spatial structure: the predicted number of contacts per residue [34], the predicted probability of hydrogen bond formation, and the predicted secondary structure.

For predicting the type of secondary structure, we used the standard program PsiPred [31]. The extent of the protection of amino acid residues from hydrogen–deuterium exchange was predicted in two ways. In the first variant residues predicted by PsiPred as involved in the α-helix or β-strand were predicted as protected from hydrogen–deuterium exchange, and residues predicted as situated in irregular secondary structure were predicted as non-protected ones. In the second variant, the protection from hydrogen–deuterium exchange was predicted according to the probability of irregular structure, which was obtained using PsiPred for each residue. If the probability of the irregular structure was greater than a cutoff value, the residue was predicted as non-protected from hydrogen–deuterium exchange; if the probability was below the cutoff value, the residue was predicted as protected. The best results of the prediction of the extent of protection were obtained at the cutoff value 0.5.

For the prediction of the extent of protection of amino acid residues from hydrogen–deuterium exchange by the amino acid sequence of the protein, we collected statistics of intraprotein contacts and of hydrogen bonds in the known spatial structures of proteins [33, 34]. For this purpose, we composed a database of spatial structures (based on the SCOP [32] database). The database included proteins or protein domains with the amino acid sequence identity below 25% and belonging to the most widespread four structural classes (all-α proteins, all-β proteins, α/β proteins, and $\alpha+\beta$ proteins). The resulting database includes 3769 proteins and protein domains. For each of the 20 types of amino acid residues we calculated the average (for a given type) number of intraprotein contacts (Fig. 3) as well as the probability of intraprotein hydrogen bond formation (Fig. 2) by the amide group of the main chain of amino acid residues of the given type. Upon prediction of number of contacts and probability of hydrogen bond formation by protein amino acid sequence, these average values were attributed to each residue of the given type. The average number of intraprotein contacts was minimal for glycine (17.11 contacts) and maximal for tryptophan (28.48 contacts). Large hydrophobic amino acid residues had (on average) more intraprotein contacts compared to small and hydrophilic residues (see Fig. 3). The probability of hydrogen bond formation by the NH-group of a residue varied from 0 (in the case of proline in which NH-group is absent) to 0.84 (for leucine). It should be noted that hydrophilic amino acid residues had smaller probabilities of hydrogen bond formation inside the main chain of the protein compared to hydrophobic residues (see Fig. 2), which probably reflects a larger involvement of hydrophobic residues in regular secondary structure elements.

During prediction of protection of amino acid residues from hydrogen–deuterium exchange, residues that had the number of predicted contacts above a cutoff value were predicted as protected, and residues that had a smaller number of predicted contacts were predicted as non-protected. Similarly, residues that had the probability of hydrogen bond formation above a cutoff value were predicted as protected, and residues that had smaller probability of hydrogen bond formation were predicted as non-protected. In both cases, the cutoff values were varied to obtain the best quality of predictions. The results of the prediction of protection of a residue from hydrogen–deuterium exchange were the best if the cutoff value was 22 intraprotein contacts per residue. Consequently (see Fig. 3), residues of tryptophan, phenylalanine, tyrosine, isoleucine, leucine, methionine, valine, and cysteine were predicted as protected residues while the other types of residues were predicted as non-protected ones in the optimal case. Thus, mainly large hydrophobic residues are predicted as protected from hydrogen–deuterium exchange. A similar situation is observed in the case of predicting protection of amino acid residues by the predicted probability of hydrogen bond formation. As seen from Fig. 2, residues of leucine, isoleucine, valine, methionine, cysteine, phenylalanine, tyrosine, alanine, and tryptophan are predicted as protected ones, while the other types of residues are predicted as non-protected. The best quality of predictions is in the case when the cutoff value is 0.7 (amino acid residues which have probability of hydrogen bond formation greater than 0.7 are predicted as protected from hydrogen–deuterium exchange, and the other residues are predicted as non-protected ones). It is interesting that the set of amino acid residues predicted as protected ones is practically the same in both variants of prediction. The only exception is alanine, which is predicted as non-protected in predictions by contacts and as protected in predictions by hydrogen bonds.

As seen in the table, the quality of predictions for methods that predict the protection based on the amino acid sequence only is almost the same (the fraction of correctly predicted residues is from 61 to 64%). At the same time, the quality of the prediction by the hydrogen bond energy, which is obtained from spatial structure, is

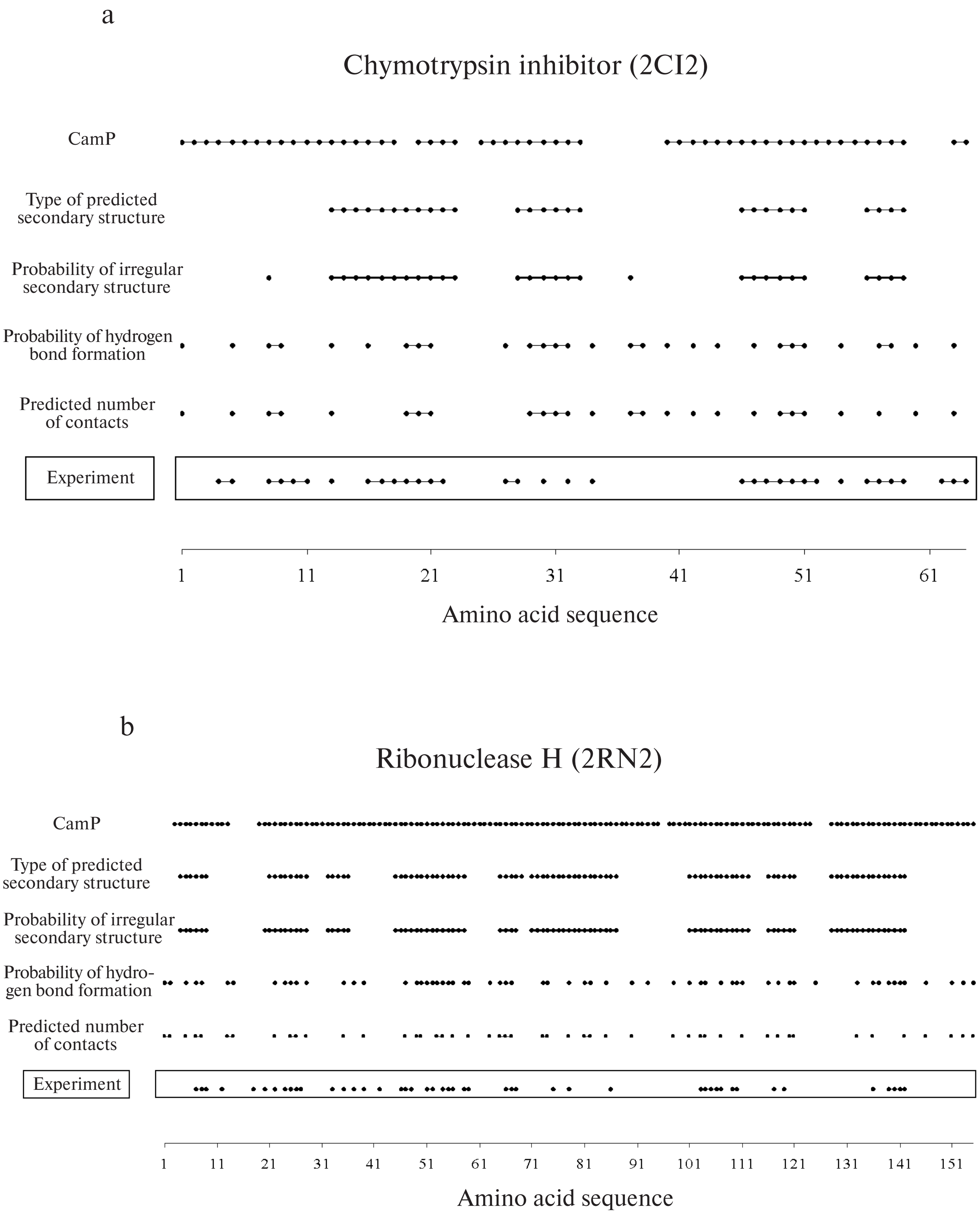

Fig. 5. Comparison of predictions made by different methods based on the amino acid sequence for two proteins: chymotrypsin inhibitor (2CI2) (a) and ribonuclease H (2RN2) (b). Regions protected from hydrogen–deuterium exchange are marked with horizontal lines.

somewhat more precise (the fraction of correctly predicted residues is 72%).

We have made predictions of protection from hydrogen–deuterium exchange using the CamP method [15] that also predicts protection of residues from hydrogen–deuterium exchange starting from the amino acid sequence. The fraction of amino acid residues the extent of protection of which was correctly predicted by the CamP method is 50%. It should be noted that the data for only six of the 14 proteins were available to us; for the same proteins, the fraction of residues correctly predicted by other methods described by us is 61% on average. In spite of the fact that CamP is a complex method based on neural network, the method heavily overpredicts the number of residues protected from hydrogen–deuterium exchange and thus, at 97% accuracy of prediction of protected regions, the accuracy of prediction of non-protected regions is as low as 13%.

A schematic representation of predictions made from the amino acid sequence is shown in Fig. 5 for two proteins. From the analysis of experimental data, it is seen that some loop regions are capable of hydrogen–deuterium exchange while other loop regions are unavailable for the exchange. For example, the N-terminal loop (amino acid residues 27-30) of the CI2 protein is unavailable for the exchange. This allows sorting the loops into available for hydrogen–deuterium exchange (flexible) and unavailable (rigid) ones.

In the present work, we have demonstrated by means of what parameters one can estimate the extent of protection of a residue (protected/non-protected) during the process of hydrogen–deuterium exchange. Moreover, our results based on hydrogen bond energy and residue–residue contacts are comparable with NMR experiments [8] where it was shown that amino acid residues with not large side chains have on average larger fluctuation of the main chain than amino acid residues possessing a massive side chain.

This work was done under financial support of the Russian Foundation for Basic Research (grant No. 08-04-00561), the Russian Academy of Sciences (programs "Molecular and Cell Biology" and "Fundamental Sciences to Medicine"), Howard Hughes Medical Institute (grant No. 55005607), and the Russian Science Support Foundation.

REFERENCES

1. Hvidt, A., and Nielsen, S. O. (1966) *Adv. Protein Chem.*, **21**, 287-386.
2. Allerhand, A., Doddrell, D., Glushko, V., Cochran, D. W., Wenkert, E., Lawson, P. J., and Gurd, F. R. N. (1971) *J. Am. Chem. Soc.*, **93**, 544-546.
3. Serdyuk, I. N., Zaccai, N., and Zaccai, J. (2007) *Methods in Molecular Biophysics*, Cambridge University Press, Cambridge.
4. Rose, G. D., Gierasch, L. M., and Smith, J. A. (1985) *Adv. Protein Chem.*, **37**, 1-109.
5. Westhof, E., Altschuh, D., Moras, D., Bloomer, A. C., Mondragon, A., Klug, A., and van Regenmortel, M. H. V. (1984) *Nature*, **311**, 123-127.
6. Fontana, A. (1988) *Biophys. Chem.*, **29**, 181-193.
7. Wimberly, B. T., White, S. W., and Ramakrishnan, V. (1997) *Structure*, **5**, 1187-1198.
8. Goodman, J. L., Pagel, M. D., and Stone, M. J. (2000) *J. Mol. Biol.*, **295**, 963-978.
9. Perutz, M. F., Kilmartin, J. V., Nagai, K., Szabo, A., and Simon, S. R. (1976) *Biochemistry*, **15**, 378-387.
10. Bai, Y. W., Milne, J. S., Mayne, L., and Englander, S. W. (1993) *Proteins*, **17**, 75-86.
11. Sheinerman, F. B., and Brooks, C. L. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 1562-1567.
12. Bahar, I., Wallqvist, A., Covell, D. G., and Jernigan, R. L. (1998) *Biochemistry*, **37**, 1067-1075.
13. Vendruscolo, M., Paci, E., Dobson, C. M., and Karplus, M. (2003) *J. Am. Chem. Soc.*, **125**, 15686-15687.
14. Best, R. B., and Vendruscolo, M. (2006) *Structure*, **14**, 97-106.
15. Tartaglia, G. G., Cavalli, A., and Vendruscolo, M. (2007) *Structure*, **15**, 139-143.
16. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) *Nucleic Acids Res.*, **28**, 235-242.
17. Wijesinha-Bettoni, R., Dobson, C. M., and Redfield, C. (2001) *J. Mol. Biol.*, **307**, 885-898.
18. Sivaraman, T., Kumar, T. K. S., Hung, K. W., and Yu, C. (2000) *Biochemistry*, **39**, 8705-8710.
19. Perez, J. M. J., Renisio, J. G., Prompers, J. J., van Platerink, C. J., Cambillau, C., Darbon, H., and Frenken, L. G. J. (2001) *Biochemistry*, **40**, 74-83.
20. Milne, J. S., Mayne, L., Roder, H., Wand, A. J., and Englander, S. W. (1998) *Protein Sci.*, **7**, 739-745.
21. Swint-Kruse, L., and Robertson, A. D. (1996) *Biochemistry*, **35**, 171-180.
22. Sadqi, M., Casares, S., Abril, M. A., Lopez-Mayorga, O., Conejero-Lara, F., and Freire, E. (1999) *Biochemistry*, **38**, 8899-8906.
23. Loh, S. N., Prehoda, K. E., Wang, J., and Markley, J. L. (1993) *Biochemistry*, **32**, 11022-11028.
24. Itzhaki, S. L., Neira, J. L., and Fersht, A. R. (1997) *J. Mol. Biol.*, **270**, 89-98.
25. Morozova, L. A., Haynie, D. T., Arico-Muendel, C., van Dael, H., and Dobson, C. M. (1995) *Nature Struct. Biol.*, **2**, 871-875.
26. Chamberlain, A. K., Handel, T. M., and Marqusee, S. (1996) *Nature Struct. Biol.*, **3**, 782-787.
27. Lacroix, E., Bruix, M., Lopez-Hernandez, E., Serrano, L., and Rico, M. (1997) *J. Mol. Biol.*, **271**, 472-487.
28. Russell, B. S., Zhong, L., Bigotti, M. G., Cutruzzola, F., and Bren, K. L. (2003) *J. Biol. Inorg. Chem.*, **8**, 156-166.
29. Kim, K. S., Fuchs, J. A., and Woodward, C. K. (1993) *Biochemistry*, **32**, 9600-9608.
30. Kabsch, W., and Sander, C. (1983) *Biopolymers*, **22**, 2577-2637.
31. Jones, D. T. (1999) *J. Mol. Biol.*, **292**, 195-202.
32. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) *J. Mol. Biol.*, **247**, 536-540.
33. Savitski, M. M., Kjeldsen, F., Nielsen, M. L., Garbuzynskiy, S. O., Galzitskaya, O. V., Surin, A. K., and Zubarev, R. A. (2007) *Angew. Chem. Int. Ed. Engl.*, **46**, 1481-1484.
34. Galzitskaya, O. V., Garbuzynskiy, S. O., and Lobanov, M. Y. (2006) *Mol. Biol.* (Moscow), **40**, 341-348.
35. Galzitskaya, O. V., Garbuzynskiy, S. O., and Lobanov, M. Y. (2006) *PLoS Comput. Biol.*, **2**, 177.