FEBS 22212 FEBS Letters 454 (1999) 31–36

Prediction of the maximal stability temperature of monomeric globular proteins solely from amino acid sequence

C. Ganesh[a], Narayanan Eswar[a], Sarika Srivastava[a], Chandrasekharan Ramakrishnan[a], Raghavan Varadarajan[a,b,*]

[a] *Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India*
[b] *Chemical Biology Unit, Jawaharlal Nehru Center for Advanced Scientific Research, Jakkur P.O., Bangalore 560 004, India*

Received 7 May 1999

Abstract Globular protein thermostability is characterized the cold denaturation, maximal stability (T_{ms}) and heat denaturation temperatures. For mesophilic globular proteins, T_{ms} typically ranges from −25°C to +35°C. We show that the indirect estimate of T_{ms} from calorimetry and the direct estimate from chemical denaturation performed in a range of temperatures are in close agreement. The heat capacity change of unfolding per mol residue (ΔC_p) alone is shown to accurately predict T_{ms}. ΔC_p and hence T_{ms} can be predicted solely from the protein sequence. The average difference in free energy of unfolding at the observed and predicted values of T_{ms} is 1.0 kcal mol^{-1}, which is small compared to typical values of the total free energy of unfolding.
© 1999 Federation of European Biochemical Societies.

Key words: Protein; Folding; Stability; Thermodynamics; Heat capacity; Accessible surface area

1. Introduction

It is generally thought that most folded protein structures observed in nature lie at a global minimum in free energy. While the various interactions that stabilize a protein relative to its unfolded state have been qualitatively understood for some time, it is still not possible to predict the free energy of unfolding (ΔG^0) of a globular protein from its three-dimensional structure or from the protein sequence. Since the folded state is generally essential for function, it is important to know the solution conditions under which a specific protein remains folded. The free energy of unfolding of a globular protein depends on temperature as follows (the Gibbs-Helmholtz equation):

$$\Delta G^0(T) = \Delta H^0(T_1) + \Delta C_p^*(T-T_1) - T^*[\Delta S(T_2) + \Delta C_p^* \ln(T/T_2)] \quad (1)$$

where $\Delta H^0(T_1)$ and $\Delta S(T_2)$ are the enthalpy and entropy of unfolding at any two reference temperatures [1–4]. The plot of ΔG^0 as a function of temperature is termed the stability curve [5]. Proteins are characterized by two denaturation temperatures (T_d) at which ΔG^0 is zero. In between these two temperatures, ΔG^0 attains a maximum value (i.e. stability of a folded protein is maximal) at a temperature called the temperature of maximal stability (T_{ms}). For temperatures outside this range, a large fraction of the protein is unfolded and hence inactive. It would therefore be highly desirable to predict T_{ms} from protein structure or better still solely from the protein sequence. The present work shows that is possible to do so, with comparable accuracies. We first show that T_{ms} can be predicted solely from the knowledge of the heat capacity change upon unfolding per mol residue of protein, ΔC_p. We next show that is possible to predict both ΔC_p and T_{ms} (with errors comparable to experimental errors) from either protein structure or protein sequence with comparable accuracies using a dataset of 28 proteins (Table 1).

2. Materials and methods

2.1. Materials

α-Chymotrypsin, lysozyme and ribonuclease A (type XII A from bovine pancreas) were all from Sigma Chemical Co. and used without further purification. *Escherichia coli* thioredoxin (P40S mutant) was expressed under the control of the T7 promoter using plasmid pET20b in strain BL21(DE3) and purified as described [6]. Protein purity was confirmed by SDS-PAGE [7]. Ultrapure grades of HEPES, urea and guanidine hydrochloride were from USB.

2.2. Datasets and fitting procedures

A total of 28 monomeric proteins were chosen for which both crystal structures and thermodynamic stability data were available. We attempted to ensure that the dataset contained a range of proteins of different sizes, number of disulfide bonds and ΔC_ps (per mol residue). In addition, for several proteins in the dataset, values of T_{ms} were directly determined by us, by carrying out chemical denaturation studies as a function of temperature. All fits described in this work were unweighted fits and were carried out using the package Sigmaplot Version 1.02 for Windows (Jandel Scientific).

2.3. Accessible area calculations

The coordinates for the folded state of a given protein were taken from the protein crystal structure. The coordinates for the unfolded state were generated by setting the backbone (ϕ,ψ) dihedral angles to value of −110° and +140° respectively. These values were chosen as they were found to be the maximally populated values in the extended region in a dataset of 156 high resolution, non-homologous, protein structures. All sidechains in the unfolded state were modeled in fully extended conformations. The accessible surface area calculations for folded and unfolded proteins were carried out using the procedure of Connolly [8] with a probe radius of 1.4 Å. All disordered residues were ignored in the calculations. All carbon and sulfur atoms were considered to be non-polar atoms for the purpose of calculating the total non-polar surface area buried upon folding, ΔA_{np}. ΔA_{np} is obtained by subtracting the total non-polar area accessible area of the folded state from that of the unfolded state. The values of ΔA_{np} calculated in this study are very similar to those obtained previously [9]. The mean difference between the two values is 5.3%.

*Corresponding author. Fax: (91) (80) 3341683, 3348535.
E-mail: varadar@mbu.iisc.ernet.in

Abbreviations: ΔA_{np}, non-polar accessible surface area buried upon folding; DSC, differential scanning calorimetry; N_{res}, number of residues; N_{S-S}, number of disulfide bonds; T_{dh}, temperature of heat denaturation; T_{ms}, temperature of maximal stability

0014-5793/99/$20.00 © 1999 Federation of European Biochemical Societies. All rights reserved.
PII: S0014-5793(99)00758-9

2.4. Experimental measurements of T_{ms}

All denaturant induced unfolding experiments were performed at pH 7.1 in 10 mM HEPES buffer containing the appropriate quantity of denaturant. For urea denaturation studies the buffer also contained 150 mM KCl. Samples were incubated at the appropriate temperature in buffer containing denaturant until equilibrium was established under the given set of conditions. Equilibrium unfolding was then measured by monitoring changes in either the fluorescence intensity (α-chymotrypsin, lysozyme and thioredoxin) or circular dichroism (CD) at 222 nm (RNase A) as a function of denaturant concentration. Denaturants used were urea for α-chymotrypsin and guanidine for the other proteins. Fluorescence and CD spectra were measured using a Jasco FP-777 spectrofluorimeter and a Jasco J500 spectropolarimeter respectively. Samples were placed in a thermostatted cell holder and in all cases measurements were made at least in duplicate. The denaturant concentration was obtained by measuring the refractive index of the solution. Data were analyzed in terms of a two-state transition. The fraction of protein in the unfolded state and ΔG^0 as a function of denaturant concentration were determined as described [10]. At each temperature the value of ΔG^0 at zero denaturant was obtained by linear extrapolation and the denaturant concentration C_m at which ΔG^0_D is zero is given by the equation $C_m = -\Delta G^0(H_2O)/m$ [11]. If m is relatively independent of temperature over a small range of temperature (here and earlier reports [12,13]) then the temperature at which C_m is maximal will also be the temperature at which $\Delta G^0(H_2O)$ is maximal. Thus, all the values reported here are in the absence of any denaturant and at near-neutral pH. The value of T_{ms} can be obtained either by fitting the values of $\Delta G^0(H_2O)$ as a function of temperature to Eq. 1 or by fitting the data for C_m to a quadratic function of temperature. Results of duplicate experiments showed that the latter method gave more reliable estimates of T_{ms}. This is probably because the values of m and $\Delta G^0(H_2O)$ obtained from a given fit are correlated and subject to greater experimental uncertainty than C_m.

3. Results

3.1. Convergence temperatures for protein unfolding

It has earlier been shown by Privalov [1,2] that at temperatures denoted by T_H and T_S (of around 110°C) the *specific* unfolding enthalpies and entropies of globular proteins converge to values (denoted by $\Delta H^0(T_H)$ and $\Delta S(T_S)$ respectively) of 1.49 ± 0.05 kcal (mol residue)$^{-1}$ and 4.21 ± 0.14 cal (mol residue)$^{-1}$ K^{-1}. The physico-chemical factors responsible for these convergence temperatures are not well understood and the above conclusions were based on calorimetric measurements of 12 proteins by Privalov and co-workers. In order to check the generality of these conclusions with a larger protein dataset, we calculated specific enthalpies and entropies of unfolding as a function of temperature for the proteins in the dataset in Table 1 from published values of ΔC_p, $\Delta H^0(T_{dh})$ and T_{dh} as described [2]. At each temperature, we also calcu-

Table 1
Data for proteins in dataset

Protein	PDB code	N_{S-S}	N_{res}	ΔA^a_{np} (Å^2)	ΔC^b_p (obs)	ΔC^b_p (calc)	T_{mso} (K)	T_{msc} (K)
Ovomucoid III	1cho	3	56	2712	10.5	10.5	268	257
RNase A	9rsa	4	124	7170	10.6	12.6	256[d]	275
CSP[c]	1csp	0	67	3736	10.7	11.5	280	266
Protein G[c]	1pgb	0	56	2834	11.1	10.5	269	257
CI2[c]	2ci2	0	65	3409	11.1	11.4	262	265
Parvalbumin	5cpv	0	108	6393	12.0	13.5	277	280
Hen lysozyme	6lyz	4	129	7701	12.0	12.8	272[d]	276
HH myoglobin[c]	lymb	0	153	9757	12.2	15.3	291	291
Hu lysozyme	1lz1	4	130	8119	12.2	12.8	276	276
Interleukin 1β	5i1b	0	153	10175	12.4	14.4	285	286
Chymotrypsin	4cha	5	241	15383	12.5	14.4	281[d]	286
Iso-1-cytochrome *c*[e]	lycc	1	108	6220	12.7	14.4	271	286
Barnase	1rnb	0	110	6166	12.8	13.6	255	281
Sac 7d	1sap	0	66	3446	13.0	11.5	296	265
RNase T1	9rnt	2	104	5847	13.0	12.6	259	274
Trypsin	1tld	6	223	14161	13.8	14.1	281	284
α-Lactalbumin	1alc	4	123	7404	14.6	12.6	291	274
CAB[c]	2cab	0	260	16850	14.6	15.3	290	291
S. nuclease	2sns	0	149	8360	14.8	14.3	289	286
Thioredoxin	2trx	1	108	6701	15.4	13.1	298[d]	278
Papain	9pap	3	212	13776	15.6	14.5	290	286
T4 lysozyme	1l63	0	164	10024	15.7	14.5	281	287
Cytochrome *c*[e]	2pcb	1	104	5978	16.1	14.3	293	285
Barstar	1bta	0	89	5770	16.4	12.8	299	276
Hpr[c]	2hpr	0	87	5121	16.7	12.7	290	275
MBP[c]	1omp	0	370	25283	17.6	15.7	306	293
SW myoglobin[c]	5mbn	0	153	9964	18.1	15.3	305	291
PGK[c]	3pgk	0	415	26466	18.1	15.8	301	294

The ΔC_p values in column 7 are calculated from the amino acid sequence using Eq. 4. The observed temperatures of maximal stability (T_{mso}) are obtained from calorimetric data. The calculated temperatures of maximum stability (T_{msc}) are obtained from the calculated values of ΔC_p in column 7 using Eq. 4 with $T_1 = 358$ K and $\Delta S(T_1) = 3.15$ cal (mol residue)$^{-1}$ K^{-1}.

[a]Areas calculated as described in the text using the algorithm of Connolly [8].

[b]Observed and calculated values of ΔC_p in units of cal (mol residue)$^{-1}$ K^{-1}. Values for ΔC_p (obs.) are obtained from the following sources: for PDB codes 1cho, 1pgb, 1lzl, 5ilb, 1ycc, 1rnb, 9rnt, 1l63, 2pcb, 5mbn, 1ymb, 6lyz, 4cha, 2trx, 2ci2 from [9]; for PDB codes 2cab, 5cpv, 1tld, 9pap from [2]; for PDB codes 1alc and 3pgk from [4]; for PDB codes 2sns, 1sap, 1bta, 2hpr, 1csp, 1omp and 9rsa from [26,27,18,12,13,28,29].

[c]Protein G: IgG binding domain of protein G; CI2: chymotrypsin inhibitor 2; HH: horse heart; CSP: cold shock protein; CAB: carbonic anhydrase; SW: sperm whale; PGK: phosphoglycerate kinase; Hu: human; HPr: histidine-containing phosphocarrier protein; MBP: maltose binding protein.

[d]See Table 2.

[e]The indicated S-S linkage for the cytochromes is the thioether linkage between the vinyl groups in the heme prosthetic group to the sulfur atoms in the cysteine side chains in the protein.

lated the mean and standard deviation in the values of specific enthalpies and entropies. Visual inspection of plots of ΔH^0 and ΔS as a function of temperature (Fig. 1A,B) does not reveal a clear-cut convergence temperature for proteins in the dataset.

The convergence temperature can be alternatively defined as the temperature at which the standard deviation in ΔH^0 or ΔS is minimal. Fig. 1C shows that this occurs at about 80–85°C for both ΔH^0 and ΔS. At 85°C the values of ΔH^0 and ΔS are 1.09 ± 0.14 kcal (mol residue)$^{-1}$ and 3.15 ± 0.46 cal (mol residue)$^{-1}$ K^{-1} respectively. However, from the figure it is also apparent that the standard deviations of ΔH^0 and ΔS change very little with temperature and hence the exact values chosen for T_H and T_S are not important. For example, at 110°C the corresponding average values for ΔH^0 and ΔS of 1.44 ± 0.15 and 4.1 ± 0.48 respectively from our data set are very similar to the values of Privalov (see above). Hence in contrast to previous assertions [1,4], there does not appear to be a universal, well defined convergence temperature for all proteins. Also, our results are in good agreement with an earlier report [14].

In the subsequent discussion, we have used the value of $T_S = 85$°C. As the standard deviations in ΔH^0 and ΔS change very little with temperature in the above range, the values of

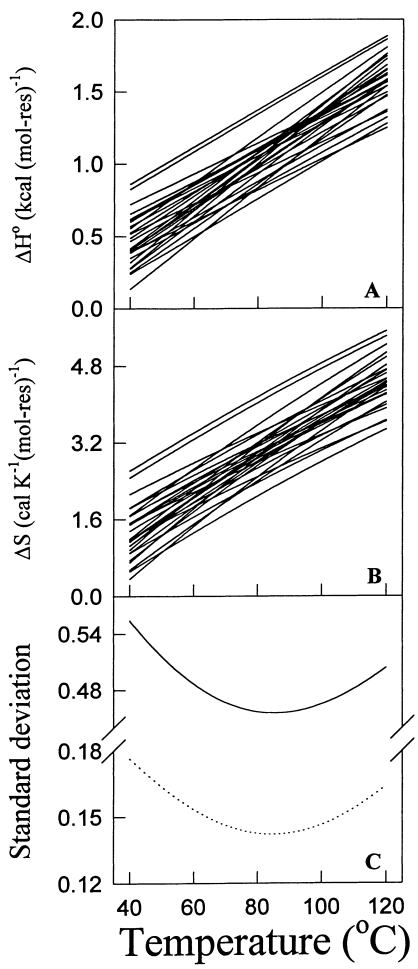

Fig. 1. A, B: ΔH^0 and ΔS as a function of temperature for all proteins in the dataset of Table 1. C: Standard deviations in ΔH^0 (dashed line) and ΔS (solid line) as a function of temperature for the proteins in the dataset. Units are the same as indicated in parts A and B.

T_H, T_S, $\Delta H^0(T_H)$ and $\Delta S(T_S)$ all appear to be similar for all proteins. The principal conclusions of our study, however, are insensitive to the exact values of these parameters and we have checked it by performing detailed analyses (data not shown).

3.2. Heat capacity change and protein stability

ΔC_p is the parameter closely linked to the hydrophobic effect and it has been shown that molar values of ΔC_p can be accurately calculated from the protein crystal structure [9,15,16]. Using differential scanning calorimetry (DSC) it is possible to measure the values of ΔC_p, T_{dh} and $\Delta H^0(T_{dh})$ for protein denaturation. The values of T_{dh} are experimentally determined while the values of T_{dc} are obtained from the stability curve of the protein. The stability curve is constructed from the *experimentally determined* T_{dh}, $\Delta H^0(T_{dh})$, $\Delta S(T_{dh})$ and ΔC_p values, using Eq. 1 and setting $T_1 = T_2 = T_{dh}$. In almost all cases, direct experimental observation of T_{dc} is not possible as these values are typically below 0°C and the sample freezes before denaturation occurs. The error in measurement of T_{dh} is typically less than 1°C. The average error of 7% in ΔC_p estimate and 12% in $\Delta H^0(T_{dh})$ [14] can be used to determine the error in the stability curves for each of the proteins in dataset, as described earlier [17]. For real proteins, the values of ΔC_p lie in the range 10–20 cal (mol residue)$^{-1}$ K^{-1} (Table 1). Also, as already mentioned, values of ΔH^0 and ΔS (on a per mol residue basis) are similar for most proteins.

3.3. ΔC_p can be used to predict T_{ms}

At T_{ms}, the slope of the stability curve, $\delta\Delta G^0/\delta T$ (equal to ΔS) is zero. It can therefore be shown that:

$$T_{ms} = T_1{}^{*}\exp[-\Delta S(T_1)/\Delta C_p] \quad (2)$$

Fig. 2A depicts T_{ms} of proteins from Table 1 as a function of their molar ΔC_p while Fig. 2B shows the form of the above function (Eq. 2) obtained using a reference temperature of $T_1 = T_S$ of 85°C, the value of $\Delta S(T_S)$ listed earlier and per mol residue ΔC_p. It is clear from Fig. 2 that T_{ms} exhibits a clear-cut relationship with per mol residue ΔC_p. It is also important to note that the curves in Fig. 2B are not sensitive to the exact values of T_S and $\Delta S(T_S)$ (legend to Fig. 2B) unlike the denaturation temperatures (data not shown). Hence, T_{ms} can be calculated directly from per mol residue ΔC_p as T_{ms} appears solely to be a function of the heat capacity change.

The calorimetric estimate of T_{ms} for any protein is obtained by choosing $T_1 = T_{dh}$ and using the calorimetric determined value of $\Delta S(T_{dh})$. The difference (ΔT_{ms}) between the calorimetric estimates of T_{ms} for proteins in Table 1 and the values predicted by Eq. 2 are small and the average difference between the observed and predicted values of T_{ms} is 6.6°C. The average experimental error in calorimetric estimation of T_{ms} is estimated to be 10°C, based on the average experimental errors of 7% for ΔC_p and 15% for $\Delta S(T_{dh})$ [14]. The average value of ΔC_p is 13.8 cal (mol residue)$^{-1}$ K^{-1} for proteins in the dataset and corresponds to a T_{ms} value of 285 K. Further analysis (data not shown) led us to the conclusion that to estimate T_{ms} for a protein, it is imperative to use its unique per mol residue ΔC_p value and not just the average value (in which case all proteins would have the same T_{ms}, see above).

The calorimetric estimates of T_{ms} are obtained from the stability curve calculated from calorimetric data and experimental values of T_{ms} are relatively pH-independent (unlike T_{dh}). In the case of a few proteins such as barstar, Hpr and cold shock protein [12,13,18], the value of T_{ms} has been measured directly by carrying out denaturation studies of the protein as a function of temperature in order to determine the temperature at which ΔG^0 is maximal. In order to confirm the accuracy of the calorimetric estimates of T_{ms} in other proteins, we have carried out chemical denaturation studies of chymotrypsin, thioredoxin, lysozyme and RNase A as a function of temperature as described in Section 2 and the measured values of T_{ms} are listed in Table 2. As a representative example, the results of experimental measurements of T_{ms} for thioredoxin are summarized in Fig. 3. In all cases there was excellent agreement between the value of T_{ms} determined by these studies and the values inferred from DSC. The average difference between these two values was 4°C. Fig. 2, Tables 1 and 2 thus demonstrate that knowledge of ΔC_p alone is sufficient to predict T_{ms}.

Calculations were performed to study the effect of the choice of the convergence temperature on the predicted T_{ms} values using Eq. 2. Values of 70°C, 85°C and 100°C for T_S (and corresponding mean $\Delta S(T_S)$ values at these temperatures for proteins in the dataset) were used and the results reveal that the mean error in T_{ms} prediction due to these variations is about 10°C. Hence, the choice of convergence temperature is not critical in predicting the T_{ms} values. Thus, T_{ms} for any protein is a function of its per mol residue ΔC_p alone, and T_{ms} increases with an increase in per mol residue ΔC_p.

3.4. Prediction of ΔC_p (per mol residue) from structure or sequence

Previous work [4,9,15] has shown that ΔC_p of a globular protein (in units of kcal K^{-1} mol^{-1}) is linearly proportional either to the non-polar surface area buried upon protein folding (ΔA_{np}) or simply to the number of residues in the protein (N_{res}). The accuracy of the fit in the latter case was slightly improved if the number of disulfide bonds (N_{S-S}) is also taken into consideration [9]. ΔA_{np} is determined from the protein crystal structure while N_{res} is determined from the protein sequence. In addition to N_{res}, it is also straightforward to determine N_{S-S} either experimentally [19] or computationally [20] if the protein sequence is known. Attempts were also made to make use of the exact amino acid composition to calculate ΔC_p (Eswar and Varadarajan, unpublished). However, the accuracy of such calculations was no greater than those which used just N_{res} and N_{S-S} and hence this approach was not pursued further.

It is shown above that T_{ms} can be predicted from the observed ΔC_p where ΔC_p is in units of cal (mol residue)$^{-1}$ K^{-1} rather than in units of kcal mol^{-1} K^{-1}. Hence we attempted to predict ΔC_p (per mol residue) from protein structure as well as from protein sequence. The observed range of ΔC_p is considerably smaller when this quantity is measured in units of cal (mol residue)$^{-1}$ K^{-1} rather than in units of kcal mol^{-1} and hence accurate prediction of the former quantity is more difficult to achieve.

We have fitted experimentally measured values of ΔC_p (per

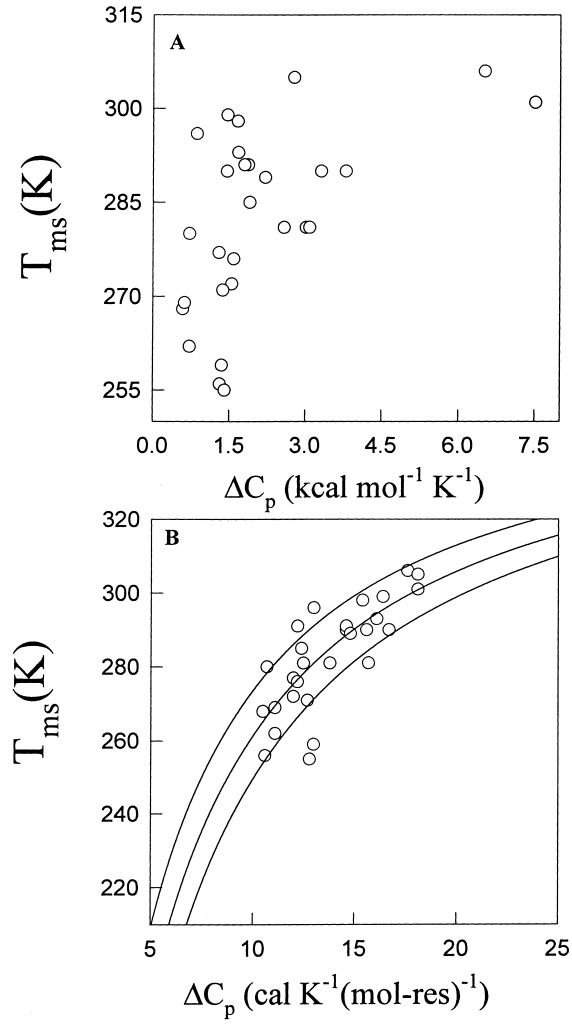

Fig. 2. T_{ms} as a function of molar (A) and per mol residue ΔC_p (B) calculated from Eq. 2 with $T_S = 358$ K. Curves in B (left to right), the values of $\Delta S(T_S)$ (in units of cal (mol residue)$^{-1}$ K^{-1}) used are 2.69, 3.15 and 3.61 respectively. Calorimetric estimates of T_{ms} (○) for proteins in the dataset of Table 1 are also shown.

Table 2
Comparison of experimentally determined and predicted T_{ms} values

Protein	N_{res}	T_{mso} (K)	T^a_{ms-LEM} (K)	T_{msc} (K)
Thioredoxin	108	298	302	278
RNase A	124	256	–[b]	275
Hen lysozyme	129	272	277	276
HH myoglobin	153	291	296	291
α-Chymotrypsin	241	281	279	286
MBP	370	306	302	291

[a] T_{ms-LEM} for all the proteins were calculated as described in Section 2.
[b] T_{ms} of RNase A could not be accurately determined experimentally as it was found to lie below 0°C.

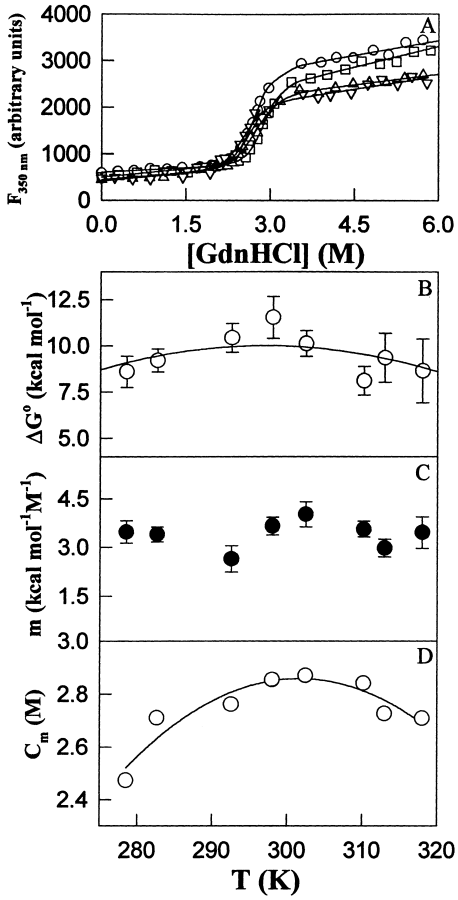

Fig. 3. Experimental determination of T_{ms} for thioredoxin. A: Fluorescence monitored unfolding as a function of denaturant concentration at four representative temperatures of 283 K (○), 298 K (□), 308 K (△) and 318 K (▽). The solid lines indicate fits to a two-state unfolding model in which the free energy of unfolding, ΔG^0_D, at a denaturant concentration [D], is given by $\Delta G^0_D = \Delta G^0 + m*[D]$ where ΔG^0 is the free energy of unfolding at zero denaturant concentration [10,11]. From each fit at a given temperature one obtains ΔG^0, m and C_m, the denaturant concentration at which half the molecules are unfolded. B–D: ΔG^0, m and C_m values as a function of temperature. A fit of ΔG^0 values to Eq. 1 is shown in B. However, the values of m and ΔG^0 obtained from the data are correlated and are subject to greater experimental uncertainty (as determined from duplicate experiments) than C_m. Since m is relatively independent of temperature, the temperature at which ΔG^0 is maximal (T_{ms}) will also be the temperature at which C_m is maximal. Hence the data for C_m are fitted to a quadratic function of temperature to obtain an estimate of $T_{ms} = 302$ K.

mol residue) for the dataset (values and units of ΔC_p, ΔA_{np}, N_{res} and N_{S-S} listed in Table 1) to the following equation:

$$\Delta C_p = a*\Delta A_{np}/N_{res} + b*N_{S-S}/N_{res} \quad (3)$$

The best fit values of a and b are 0.24 ± 0.1 and -42 ± 23 respectively. This fit and all the other fits in this work are unweighted fits. As described previously [9], inclusion of the second term in Eq. 3 results in a small but significant improvement in the accuracy of the fit. The average errors [21] in prediction of ΔC_p was 1.5 cal (mol residue)$^{-1}$ K^{-1} for the proteins in the dataset. This error is only slightly larger than the average experimental errors in ΔC_p of 7% [14].

We next repeated the above analysis using ΔC_ps calculated from protein sequence data. The observed ΔC_ps of proteins in the dataset were fit to the equation:

$$\Delta C_p = c + d/N_{res} + e*N_{S-S}/N_{res} + f*N_{heme}/N_{res} \quad (4)$$

The best fit parameters were $c = 16.6 \pm 1.0$, $d = -274 \pm 87$, $e = -51 \pm 25$ and $f = 76 \pm 129$. The average difference between observed and calculated ΔC_ps was 1.6 cal (mol residue)$^{-1}$ K^{-1} for the dataset. Eq. 4 shows that ΔC_p (per mol residue) increases with an increase in protein size. This is probably because globular proteins have constant packing densities that are relatively independent of protein size and larger proteins tend to bury a greater fraction of the total surface than smaller proteins.

3.5. Prediction of T_{ms} from protein structure or sequence

The predicted value of ΔC_p is substituted into Eq. 2 to obtain the predicted value of T_{ms} using $T_1 = T_S = 85$°C and $\Delta S(T_S) = 3.15$ cal (mol residue)$^{-1}$ K^{-1}. The values of T_{ms} calculated using values of ΔC_p predicted from Eq. 3 were compared with the experimental value listed in Table 1. A histogram of the absolute value of the difference ($|\Delta T_{ms}|$) between the calculated (T_{msc}) and observed (T_{mso}) temperatures of maximal stability is shown in Fig. 4A. In most cases there is good agreement between the two values and the average value of $|\Delta T_{ms}|$ is 10°C. This error is small relative to the observed range of T_{mso} (about 250–310 K (Table 1 and Fig. 3)) and comparable to the experimental error estimates of 3–10°C in determination of T_{ms}. The fact that errors in the prediction of

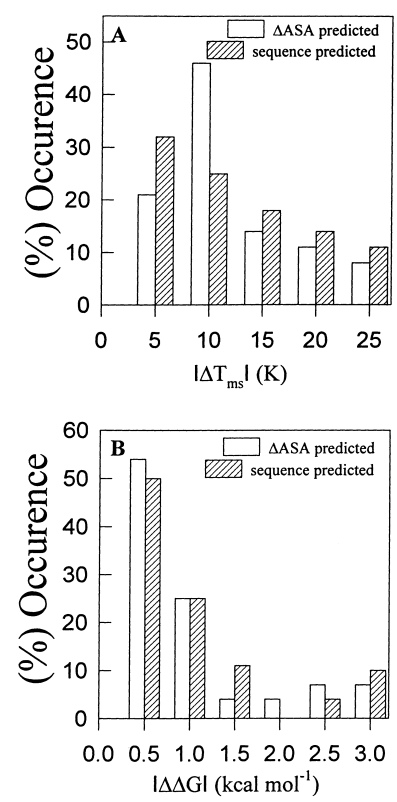

Fig. 4. A: Histogram of $|\Delta T_{ms}|$ obtained using ΔC_ps calculated from protein crystal structure (Eq. 3, open bars) and that from protein sequence (Eq. 4, hatched bars). B: Histograms of $|\Delta\Delta G(T_{ms})|$ obtained using ΔC_ps calculated as in A.

both ΔC_p and T_{ms} are close to the experimental errors indicates that our procedure has significant predictive value.

Fig. 4A also shows the histogram of $|\Delta T_{ms}|$ obtained using the values of ΔC_p calculated from Eq. 4. The average value of $|\Delta T_{ms}|$ now also was 10°C. T_{ms} can thus be calculated with reasonable accuracy solely from protein sequence.

For small values of ΔT_{ms} it can be shown that the difference in the stability of the protein at temperatures T_{mso} and T_{msc} ($\Delta\Delta G^0$) is given by the equation:

$$\Delta\Delta G^0 = \Delta G^0(T_{msc}) - \Delta G^0(T_{mso}) \approx -\Delta C_p * (\Delta T_{ms})^2 / T_{mso} \quad (5)$$

Shown in Fig. 4B is a histogram of the values of $|\Delta\Delta G^0|$ (in kcal mol^{-1}) for proteins in the dataset using values of T_{msc} estimated using ΔC_p calculated from Eqs. 3 and 4. The figure shows that the average error in ΔG^0 that results from an error in the prediction of T_{ms} is small, about 1.0 kcal mol^{-1}.

4. Discussion

It has been previously been shown that the molar values of ΔC_p are linearly proportional either to ΔA_{np} or to N_{res} [9,15]. While our method for ΔC_p is very similar and of comparable accuracy to ones used earlier, the focus of the present work is to show that values of ΔC_p (per molresidue) are significantly different for different proteins and it is these values rather than the molar values of ΔC_p that determine the values of T_{ms} for a given protein. In contrast to earlier views, there does not appear to be a well-defined convergence temperature for either ΔH^0 or ΔS and hence it may not be appropriate to attach any special physical significance to earlier proposed convergence temperature values of 110°C or 130°C [1,22,23]. This illustrates the difficulty of translating observations based on small molecules and free energy of transfer data to protein folding.

It is feasible to predict ΔC_p (per molresidue) and T_{ms} from either protein structure or protein sequence. Surprisingly, the prediction accuracies are similar for both structure and sequence based methods. The primary determinant of ΔC_p and hence of T_{ms} is protein size. The amino acid composition of the protein does not appear to be important. Smaller proteins tend to have lower values of ΔC_p (per mol residue) and hence lower T_{ms}. Inclusion of additional factors such as the number of disulfides [9] and heme groups results in a small increase in the average accuracy of ΔC_p. While the correlation between ΔA_{np} for proteins and ΔC_p has been used to infer that ΔC_p is closely related to the hydrophobic driving force for protein folding, it is worth noting that since ΔC_p is equally well predicted from protein size alone it may not be appropriate to draw such a conclusion. An important and as yet unresolved issue is the validity of using small molecule thermodynamic data to draw conclusions about the relative magnitudes of the various driving forces for protein folding.

Our analysis applies to monomeric globular proteins that show reversible unfolding behavior and a ΔC_p that is relatively independent of temperature. For multimeric proteins, the values of T_{dc}, T_{ms}, and T_{dh} will be concentration-dependent [24]. The analysis is based on a dataset composed primarily of naturally occurring mesophilic proteins and is unlikely to apply to proteins from hyperthermophiles or to site-specific mutants in which severely destabilizing substitutions have been made [25]. These caveats aside, our results show that it is possible to predict T_{ms} for a variety of globular proteins of different sizes, stabilities and secondary and tertiary structures with errors close to those in experimental estimations.

Acknowledgements: Financial support from Grants DST/SP/SO/D-21/93 and CSIR/37(913)96-EMR-II to R.V. We thank A. Surolia and S. Varadarajan for helpful suggestions and S. Boxer for generous gifts of lysozyme and chymotrypsin. We thank Dr. Atis Chakrabarti and Swathi Seshadri for carrying out several denaturant melts of myoglobin and MBP.

References

[1] Privalov, P.L. (1979) Adv. Protein Chem. 33, 167–241.
[2] Privalov, P.L. and Gill, S.J. (1988) Adv. Protein Chem. 39, 191–234.
[3] Murphy, K.P., Privalov, P.L. and Gill, S.J. (1990) Science 247, 559–561.
[4] Murphy, K.P. and Freire, E. (1992) Adv. Protein Chem. 43, 313–361.
[5] Becktel, W.J. and Schellman, J.A. (1987) Biopolymers 26, 1859–1877.
[6] Wynn, R. and Richards, F.M. (1993) Protein Sci. 2, 395–403.
[7] Laemmli, U.K. (1970) Nature 227, 680–685.
[8] Connolly, M. (1983) J. Appl. Crystallogr. 16, 548–558.
[9] Myers, J.K., Pace, C.N. and Scholtz, J.M. (1995) Protein Sci. 4, 2138–2148.
[10] Santoro, M.M. and Bolen, D.W. (1988) Biochemistry 27, 8063–8068.
[11] Greene, R.F. and Pace, C.N. (1974) J. Biol. Chem. 249, 5388–5393.
[12] Nicholson, E.M. and Scholtz, J.M. (1996) Biochemistry 35, 11369–11378.
[13] Schindler, T. and Schmid, F.X. (1996) Biochemistry 35, 16833–16842.
[14] Robertson, A. and Murphy, K.P. (1997) Chem. Rev. 97, 1251–1268.
[15] Spolar, R.S., Livingstone, J.R. and Record Jr., M.T. (1992) Biochemistry 31, 3947–3955.
[16] Gomez, J., Hilser, V.J., Xie, D. and Freire, E. (1995) Proteins Struct. Func. Genet. 22, 404–412.
[17] Ganesh, C., Shah, A.N., Swaminathan, C.P., Surolia, A. and Varadarajan, R. (1997) Biochemistry 36, 5020–5028.
[18] Agashe, V.R. and Udgaonkar, J.B. (1995) Biochemistry 34, 3286–3299.
[19] Creighton, T.E. (1989) Protein Structure: A Practical Approach, IRL Press, Oxford.
[20] Fiser, A., Czerno, M., Tudos, E. and Simon, I. (1992) FEBS Lett. 302, 117–120.
[21] Hogg, R.V. and Tanis, E.A. (1983) Probablity and Statistical Inference, 2nd edn., Macmillan Publishing Co., Inc. New York.
[22] Lee, B. (1991) Proc. Natl. Acad. Sci. USA 88, 5154–5158.
[23] Baldwin, R.L. and Muller, N. (1992) Proc. Natl. Acad. Sci. USA 89, 7110–7113.
[24] Johnson, C.R., Morin, P.E., Arrowsmith, C.H. and Freire, E. (1995) Biochemistry 34, 5309–5316.
[25] Varadarajan, R., Connelly, P.R., Sturtevant, J.M. and Richards, F.M. (1992) Biochemistry 31, 1421–1426.
[26] Calderon, R.O., Stolowich, N.J., Gerlt, J. and Sturtevant, J.M. (1985) Biochemistry 24, 6044–6049.
[27] McCrary, B.S., Edmondson, S.P. and Shriver, J.W. (1996) J. Mol. Biol. 264, 784–805.
[28] Novokhatny, V. and Ingham, K. (1997) Protein Sci. 6, 141–146.
[29] Catanzano, F., Giancola, C., Graziano, G. and Barone, G. (1996) Biochemistry 35, 13378–13385.