Original Article

In silico Evaluation of Crosslinking Effects on Denaturant m_{eq} values and ΔCp upon Protein Unfolding

Maryam Hamzeh-Mivehroud [1], Ali Akbar Alizade [1, 2], Monire Ahmadifar [1, 2], and Siavoush Dastmalchi [1, 2*]

1. Biotechnology Research Center, Tabriz University of Medical Sciences, Daneshgah Street, Tabriz, Iran
2. School of Pharmacy, Tabriz University of Medical Sciences, Tabriz, Iran

Abstract

Important thermodynamic parameters including denaturant equilibrium m values (m_{eq}) and heat capacity changes (ΔCp) can be predicted based on changes in Solvent Accessible Surface Area (*SASA*) upon unfolding. Crosslinks such as disulfide bonds influence the stability of the proteins by decreasing the entropy gain as well as reduction of *SASA* of unfolded state. The aim of the study was to develop mathematical models to predict the effect of crosslinks on $\Delta SASA$ and ultimately on m_{eq} and ΔCp based on *in silico* methods. Changes of *SASA* upon computationally simulated unfolding were calculated for a set of 45 proteins with known m_{eq} and ΔCp values and the effect of crosslinks on $\Delta SASA$ of unfolding was investigated. The results were used to predict the m_{eq} of denaturation for guanidine hydrochloride and urea, as well as ΔCp for the studied proteins with overall error of 20%, 31% and 17%, respectively. The results of the current study were in close agreement with those obtained from the previous studies.

Avicenna J Med Biotech 2012; 4(1): 23-34

Keywords: Crosslinks, Disulfides, Protein stability, Thermodynamics

*** Corresponding author:**
Siavoush Dastmalchi, Ph.D., Department of Medicinal Chemistry, School of Pharmacy, Tabriz University of Medical Sciences, Tabriz, Iran
Tel: +98 411 3364038
Fax: +98 411 3379420
E-mail: dastmalchi.s@tbzmed.ac.ir
Received: 21 Oct 2011
Accepted: 26 Feb 2012

Introduction

Through the human genome project we now know that a human cell can synthesize about 20,000 to 25,000 different proteins [(1)]. Proteins are an important class of biological macromolecules present in all biological organisms, and constitute high proportion of the dry mass of all cells [(2)]. Most of the biological processes in all cells are executed by proteins. The amino acid sequence of a protein contains all information needed for adopting its three-dimensional structure. However, misfolding does occur, even though help from other molecules, such as chaperons, for correct and fast *in vivo* folding are in place [(3-5)].

Denaturation studies are very useful for investigating the thermodynamic properties of proteins. Transition from native to denatured states can be brought about by changing the properties of protein's environment. In general, this can be done by increasing the temperature, adding chemical denaturants or changing the pH.

Urea and guanidinium ion (used in the form of guanidinium chloride-GdnHCI) favor the denatured state by increasing the solubility of the unfolded chain in an aqueous solution. In comparison to temperature denaturation, chemical denaturation is often a reversible process. This is possible since the hydrophobic groups of the unfolded chain are shielded by the denaturants, which prevent aggregation.

The unfolding free energy (ΔG_U) depends linearly on the denaturant concentration as:

$$\Delta G_U = \Delta G_U^{H2O} - m_U \left[Denaturant\right] \text{ (Eq.1)}$$

Where is $\Delta G_U^{H_2O}$ the free energy of unfolding in the absence of denaturant and m_U denotes the dependency of free energy on denaturant concentration (*i.e.* m_{eq}) [6]. A good linearity is observed at high denaturant concentrations and $\Delta G_U^{H_2O}$ is obtained by extrapolation to the zero concentration of denaturant. $\Delta G_U^{H_2O}$ values calculated from guanidinium chloride and urea denaturation are in very good agreement [7] which gives this relation some further credibility.

One of the major challenges in the field of protein science is to predict the stability and function of proteins from their primary structures. To accomplish this task, efficient algorithms are needed to relate the structure to stability. The availability of about 77,000 protein structures in Protein Data Bank (PDB) [8] and a great deal of experimental works on the thermodynamic stability of proteins have provided a wealth of information which can be used for the development of empirical functions that relate thermodynamic and structural parameters.

The success of such approach in developing structure-based methods to predict various thermodynamic parameters that define the Gibbs energy, *i.e.*, the enthalpy, entropy and heat capacity changes, has been shown previously [9-13]. In the process of unfolding, the major contribution to the enthalpy change arises from the disruption of intramolecular interactions such as van der Waals and hydrogen bonds and also solvation of the interacting groups. Therefore, the change in solvent accessible surface area ($\Delta SASA$) upon unfolding has been used as a mean for predicting the ΔH as presented below:

$$\Delta H = \sum_i \alpha_i\left(\overline{\rho}\right) \cdot \Delta SASA_i \text{ (Eq.2)}$$

Where $\Delta SASA_i$ is the change in *SASA* of atom *i* upon unfolding, and $\alpha_i\left(\overline{\rho}\right)$ is a coefficient that depends on the atom type and the average packing density of that atom within the protein [14].

The heat capacity change (ΔC_p) in protein unfolding largely arises from changes in the hydration of groups that are buried in the native form away from the surrounding aqueous environment. ΔCp is correlated to the changes in *SASA* upon unfolding, as shown in the following equation:

$$\Delta C_p = \sum_i a_i \cdot \Delta SASA_i \text{ (Eq.3)}$$

Where a_i is the contribution of atom *i* per unit area and $\Delta SASA_i$ is as defined above. Using both equations, good correlations were obtained between experimental and calculated ΔH and ΔCp values [14].

The aim of current study is to develop empirical models to account for the effect of crosslinks on $\Delta SASA$ and hence on thermodynamic parameters (*i.e.*, m_{eq} and ΔCp) of protein unfolding based on computational approach.

Materials and Methods

Databases and programs

The experimental m_{eq} values for urea and GdnHCl denaturation, as well as ΔCp denaturations for a set of 45 proteins used in this study were from Myers et al [10]. The three-dimensional (3D) structures of the studied proteins were obtained from Protein Data Bank (http://www.rcsb.org/) at RCSB [8].

The *SASA* of the proteins in folded and unfolded forms were calculated using DSSP program implemented in GROMACS package. The DSSP program was designed by Wolfgang Kabsch and Chris Sander to standardize secondary structure assignment based on a database of secondary structure for protein entries in the PDB [15].

Swiss-Pdb Viewer (SPDBV, version 3.7, Swiss Institute of Bioinformatics), an interactive molecular graphics program was used for viewing and analyzing protein structures [16]. HyperChem (version 7.1; 2002; Hyper-

cube Inc.) is the other molecular modeling software used in this study.

GROMACS (version 3.3, University of Groningen, The Netherlands, currently maintained by ScalaLife), an engine to perform molecular dynamics simulations and energy minimization [17] was used under Linux operating system (Fedora core 5) on a cluster consisting of 8 nodes each with two dual-core Opteron 2212 CPUs and 2 GB RAM.

Unfolding the proteins

The unfolded states of the proteins were achieved by three different approaches; (*i*) building the fully extended conformation of the protein, (*ii*) instantaneously assigning standard bond lengths, bond angles, torsion angles, and stereochemistry properties to the model structure using a given force field method, or (*iii*) molecular dynamics simulation.

Fully extended conformation

SPDBV was used to upload the sequence of the protein saved in FASTA format. Then the sequence was folded into an extended conformation by setting phi (φ) and psi (ψ) angles to those corresponding with β-pleated strand. It is clear that in such a conformation there is no crosslink in the generated model even if the native form of protein consists such constraints.

Instantaneous unfolding using standard bond and angle assignment

HyperChem program was used to open the crystal structure (native form) of protein. In this way, the disulfide bonds are lost. If the re-establishment of crosslinks was desired, first the residues involved in the crosslink were selected and then the necessary bonds were created between the sulfur atoms involved in the disulfide bonds. Subsequently, the structure was forced to unfold into a random coil losing its regular structures while preserving the crosslinks. The unfolded structural model was energy minimized using the molecular mechanics force field.

The minimization protocol employed the steepest descent method using BIO+, the HyperChem implementation of CHARMM (Chemistry at HARvard using Molecular Mechanics) force field [18], until the difference in energy after two consecutive iterations was less than 0.1 *kcal/mol*. The model structures were stored as unfolded states and their *SASA* were calculated as described above. In the case of heme containing proteins, two bonds were built linking the chelating atoms to the central iron atom. This effectively constrains the spatial distance between two residues to which the iron atom of the heme group is linked through coordination of the unpaired electrons of nitrogen or sulfur atoms.

Unfolding using molecular dynamics simulation

In order to unfold proteins using Molecular Dynamics (MD) simulation technique, the following steps were performed. First, the native structure was downloaded from PDB at RCSB and converted into standard Gromacs file format. The positions of all hydrogen atoms were reconstructed. Subsequently, the protein structure was energy minimized in vacuum using steepest descent algorithm until the maximum force was smaller than 1.0 $kJ\ mol^{-1}nm^{-1}$.

GROMOS-96, the officially distributed force field for Gromacs, was used for molecular mechanics simulations as implemented in the software package [19]. Then a simulation box was created and protein was centred into it. The simulation box was filled in by Simple Point Charge (spc216) water and urea molecules. The final concentration of the urea in the box was about 4.4 *M*. Before running the MD simulation, the system was neutralized by adding appropriate number of either Na^+ or Cl^- counter ions to have zero net charge. Ultimately, the solvated protein was subjected to MD simulation for 10 *ns* at 500 $^{\circ}K$ and the trajectories were saved every 0.02 *ns*.

$$Crosslinking\ number = \frac{SASA_{nc} - SASA_{c}}{n} \quad \text{(Eq.4)}$$

$$Crosslinking\ factor = \frac{1}{N}\sum_{i}^{N} Crosslinking\ number \quad \text{(Eq.5)}$$

Crosslinking factor (CLF)

In order to investigate the effect of crosslinking on the unfolding behaviour of a protein, an index named Crosslinking Factor (CLF) was defined as follows:

Where *SASA* refers to solvent accessible surface area of unfolded conformation and the subscripts *c* and *nc* denote whether the crosslinks are preserved or not in the unfolded conformation, respectively. The value of n equals the number of crosslinks present in any of those proteins studied here which have crosslinks in the native form. *N* is the number of proteins with crosslinks, and *i* denotes any of the studied proteins used to derive CLF value.

Statistical treatment

Validation of models: Statistical analyses were performed by SPSS (SPSS for windows version 11.5, IBM) and Excel (Microsoft Office 2007) programs. Predictive power of the mathematical models were evaluated by excluding one of the data points, *i.e.* one of the proteins from the data set of 45 proteins listed in table 1, and training the model based on the remaining proteins and subsequently predicting the value of thermodynamic parameter for the excluded protein. This was continued until all proteins were used for the prediction.

The Standard Deviation of Error of Prediction (SDEP) was calculated to give a measure for the distribution of the errors involved in the predictions using the following equation:

$$SDEP = \sqrt{\frac{\sum_{i=1}^{N} \left(A_{exp} - A_{calc} \right)^2}{N}} \quad \text{(Eq.6)}$$

Here A_{exp} and A_{calc} are predicted values, respectively. N denotes the number of data points.

Mean absolute percentage error (MAPE)

To evaluate the accuracy of predictions, absolute percentage errors were calculated based on the following equations:

$$APE = \frac{\left| A_{calc} - A_{exp} \right|}{A_{exp}} \times 100 \quad \text{(Eq.7)}$$

Where A_{calc} and A_{exp} are the calculated and experimental values for a given parameter of interest, such as ΔC_p, m_{eq} for GdnHCl or urea. The average of *APE* over all data points for each of the above mentioned parameters was calculated and called MAPE.

$$MAPE = \frac{1}{N} \sum_{i}^{N} APE_i \quad \text{(Eq.8)}$$

Where *N* is the number of data points.

Results and Discussion

The changes in solvent accessible surface area ($\Delta SASA$) upon unfolding, as determined by the differences in solvent accessibilities of native form (calculated from the crystal structure) and denatured form (modeled by an extended polypeptide chain) are given for a set of 45 proteins in table 1. The table also shows m_{eq} values from denaturation experiments, ΔCp of unfolding, number of residues as well as crosslinks present in each of these proteins taken from the compilation made by Myers et al [10].

Figures 1A and 1B demonstrate dependencies that exist between the denaturants m_{eq} values and the changes in the solvent accessible surface area upon unfolding. There are significant linear correlations in both cases, with the correlation coefficient (R) values of 0.85 and 0.87 for GdnHCl and urea, respectively. The slopes of the linear regression lines are 0.25 and 0.17 *cal/* ($mol.M.\AA^2$) for GdnHCl and urea, respectively, indicating the stronger denaturing effects of GdnHCl.

Denaturation heat capacity changes (ΔCp) were also correlated with the $\Delta SASA$ strongly with the correlation coefficient of 0.97 as shown in figure 1C. The same linear correlations between m_{eq} values and $\Delta SASA$ have been shown previously by Myers et al [10]. $\Delta SASA$ has been also related linearly to ΔCp by others [20,21].

Table 1. Characteristics of 45 proteins that have m_{eq} values and crystal structures available [a].

Protein name	PDB	Number of Residues	Number of crosslinks	m_{eq}(GdnHcl)	m_{eq}(Urea)	ΔCp	$SASA_{folded}$	$SASA_{unfolded}$ [b]	$\Delta SASA$
				cal/(mol.M)					
Ovomucoid third domain (turkey)	1CHO	53[d]	3	580	250	590	3735	7157	3422
1gG binding domain of protein G	lPGB	56	0	1800	NA	620	3752	7705	3953
BPTl (A30, A51)	7PTI	58	2	1200	NA	NA	3969	8254	4285
BPTl (V30, A51)	1AAL	58	2	1500	NA	NA	3993	8276	4283
SH3 domain of α-spectrin	lSHG	57[d]	0	1880	766	813	3925	8220	4295
Chymotrypsin inhibitor 2	2CI2	65[d]	0	1890	NA	720	4564	9246	4682
Calbindin D9K	1IG5	75	0	NA	1140	NA	4774	10373	5599
Ubiquitin	lUBI	76	0	NA	1140	NA	4911	10758	5847
HPr (B. subtilis)	2HPR	87[d]	0	NA	1050	1160	4751	11688	6937
Barstar	1BTA	89	0	2400	1250	1460	5653	12596	6943
Lambda repressor (N-terminal)	1LMB	102	0	2400	1090	NA	6270	13013	6743
Cytochrome c (tuna)	5CYT	103	1	2800	NA	NA	6087	14382	8295
Cytochrome c (horse heart)	2PCB	104	1	3010	1200	1730	6363	14812	8449
Ribonuclease T1	9RNT	104	2	2560	1210	1270	5467	13651	8184
Arc repressor[c]	1PAR	106	0	3270	1910	1600	6566	15471	8906
FK binding protein (human)	lFKD	107	0	NA	1460	NA	6144	14798	8654
Iso-I-cytochrome c (yeast)	lYCC	108	1	3400	1430	1370	6575	15169	8594
Thioredoxin (*E.coli*)	2TRX	108	1	3310	1300	1660	5847	14776	8929
Barnase	I RNB	109[d]	0	4400	1940	1650	6050	15093	9043
Ribonuclease A	9RSA	124	4	3100	1100	1230	6965	16983	10018
ROP	1 ROP	126	0	2400	NA	1890	6445	16191	9746
Che Y (*E.coli*)	3CHY	128[d]	0	2260	1600	NA	6673	17646	10973
Lysozyme (hen egg white)	1AKI	129	4	2330	1290	1540	6755	17886	11131
Lysozyme (human)	lLZl	130	4	3460	NA	1580	6777	18305	11528
Fatty acid binding protein (rat)	llFC	131	0	4470	1770	NA	7145	18564	11419
Staphylococcal nuclease	2SNS	141[d]	0	6830	2380	2320	8052	20083	12031
Interleukin 1-β	511B	151	0	5580	NA	1890	8209	21188	12979
Apomyoglobin (horse)	IYMB	153	1	3710	2140	1870	8296	21895	13599
Apomyoglobin (sperm whale)	5MBN	153	1	2600	1460	2770	8320	22180	13860
Metmyoglobin (horse)	IYMB	153	0	NA	1800	NA	8296	21042	12746
Metmyoglobin (sperm whale)	5MBN	153	0	NA	2040	NA	8320	21327	13007
Ribonuclease H	2RN2	155	0	4500	1930	NA	8785	21635	12850
Dihydrofolate reductase (*E.coli*)	4DFR	159	0	NA	1900	NA	8717	21945	13228
T4 lysozyme (T54, A97)	1L63	162[d]	0	5500	2000	2570	8553	22913	14360
Gene v protein[c]	1VQB	172[d]	0	3600	NA	NA	13216	26728	13512
Adenylate kinase (porcine)	3ADK	194	0	4800	NA	NA	11051	26978	15927
HIV-1 protease[c]	lHVR	198	0	NA	2050	NA	9865	26784	16919
SIV protease[c]	lSIV	198	0	NA	1880	NA	9962	26576	16614
Trp aporepressor[c]	3WRP	202[d]	0	NA	2900	NA	11388	28583	17195
α-Chymotrypsin	4CHA	239[d]	5	4100	2070	3020	10742	31985	20498
Chymotrypsinogen A	2CGA	245	5	4440	2030	NA	10742	31985	21243
Tryptophan synthase, α-subunit	IBKS	255[d]	0	NA	3750	4600	11585	34271	22686
β-Lactamase	3BLM	257[d]	0	7200	3210	NA	11561	36444	24883
Pepsinogen	2PSG	370	3	NA	7800	6090	14748	48478	33730
Phosphoglycerate kinase (yeast)	3PGK	415	0	9700	NA	7500	18988	53051	34063

NA: Not Available; [a] for each protein, the PDB file code, number of residues, and number of disulfides or covalent heme-protein crosslinks is shown. *SASA* values were calculated by DSSP program as described in the text. The 5, 6 and 7th columns give experimental m_{eq} values for GdnHCl or urea denaturation and the observed ΔC_p, for each protein, taken from reference [10]. $\Delta SASA$ values are in $Å^2$, m_{eq} values in *cal*/(*mol.M*), and ΔCp, in *cal*/(*mol.K*); [b] $SASA_{unfolded}$ values in this table were calculated using the extended β-strand conformation of all proteins; [c] Dimer; [d]These values were checked and corrected based on the number of the residues in the corresponding PDB files and hence are different from those reported in Myers et al. [10].

The main purpose of this study is to re-evaluate the effect of crosslinks on $\Delta SASA$ and also predict the m_{eq} and ΔCp of unfolding based on protein sequence information. These latter two parameters are amongst the important criterion indicative of the stability of proteins. Therefore, prediction or any improvement in the prediction of these values has significant theoretical and practical applications.

The presence of crosslinks such as disulfide bonds and heme groups in a protein (as shown in table 2) will result in a more compact unfolded state, thus reducing the solvent accessi-

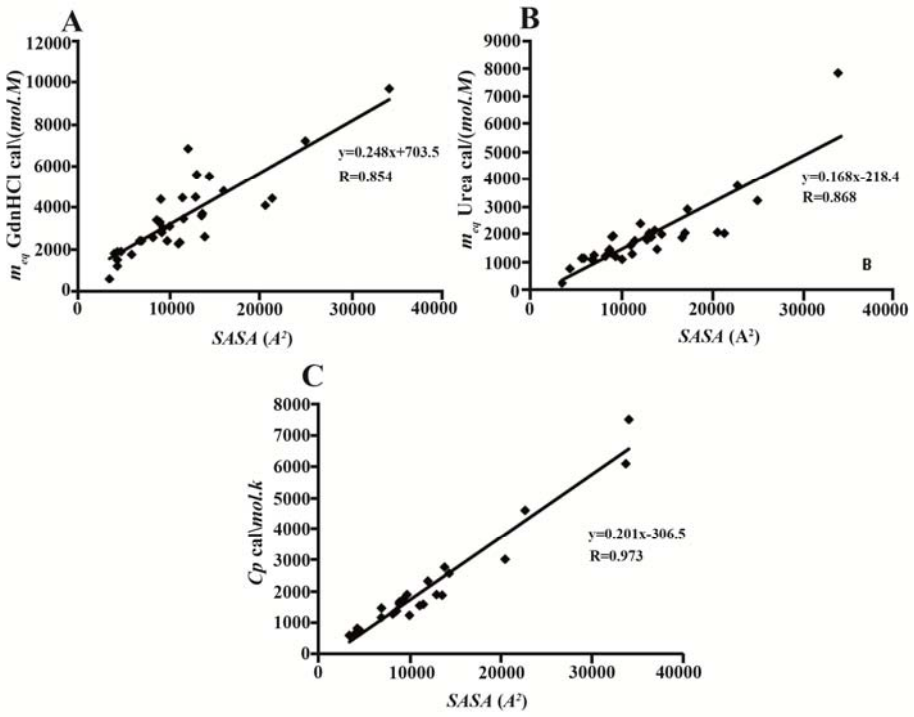

Figure 1. Dependence of A) m_{eq} value for Gdn HCl denaturation, B) m_{eq} value for urea denaturation, and C) heat capacity changes upon unfolding on Δ*SASA* for the 45 proteins shown in table 1

bility of the unfolded polypeptide chain. To compensate for the effects of crosslinks, Myers et al [10] employed the results of different empirical methods [22] to estimate the magnitude of the reduction of solvent accessible surface area (Δ*SASA*) per disulfide bond. The reduction of Δ*SASA* per crosslink was estimated to be about $900\mathring{A}^2$.

In the current study, to find out more about the effect of crosslinking through theoretical and computational methods, the different unfolded models were generated for crosslink-containing proteins while the crosslinks were preserved or removed in the unfolded states generated by instantaneous unfolding method based on assigning standard bond length and angle values. Then the *SASA* values were calculated for the generated unfolded structural models (Table 2).

To quantitatively indicate the effect of crosslinks on Δ*SASA* upon unfolding a new term called Crosslinking Factor (CLF) was introduced (CLF was described in Materials and Methods section.) Effectively, CLF is a measure of reduction in the *SASA* of unfolded protein as a consequence of presence of a single crosslink, such as disulfide bond, and calculated to be equal to 918.5 $\mathring{A}^2$. This value is the average of crosslinking numbers calculated for 16 crosslink-containing proteins listed in table 2 for which the m_{eq} and ΔCp values were available.

In five proteins listed in the table, crosslinks are formed via ligation of central ion atom of heme groups by sulfur or nitrogen atoms of the side chains of the interacting residues. The average of crosslinking numbers for these proteins (759.0 $\mathring{A}^2$) is smaller than the average of the numbers (991.0 $\mathring{A}^2$) for the remaining proteins where the crosslinks are formed by disulfide bounds. However, the difference is not statistically significant (p-value >0.05). None of these values are statistically different from the calculated CLF value of 918.5.

Based on the above findings, the Δ*SASA* values were corrected for the effect of crosslinks on the solvent accessibility of the unfolded state by taking 918.5 $\mathring{A}^2$ per crosslink off the $\Delta SASA$ (called $\Delta SASA_{\text{corrected}}$) and then the corrected values were re-correlated to the m_{eq} and ΔCp values. Linear correlation coef-

Table 2. List of crosslink-containing proteins used in this study. Differences of *SASA* values for the unfolded stats in two different forms, *i.e.*, with and without conserving the crosslinks, have been shown along with the number of crosslinks and crosslinking number for each protein

PDB code	$SASA_{unfolded}$ without Crosslinks[a]	$SASA_{unfolded}$ with Crosslinks[a]	*ΔSASA*	Number of crosslinks (*n*)	Crosslinking number
1CHO	7039	5730	1309	3	436.33
7PTI	8278	6616	1662	2	831.00
1AAL	8315	6539	1776	2	888.00
5CYT[b]	13929	13392	537	1	537.00
2PCB[b]	14410	14118	292	1	292.00
9RNT	13348	11227	2121	2	1060.50
1YCC[b]	15145	14620	525	1	525.00
2TRX	13568	13254	314	1	314.00
9RSA	17125	13015	4110	4	1027.50
1AKI	17597	12575	5022	4	1255.50
1LZ1	18698	12967	5731	4	1430.25
1YMB[b]	22110	20864	1246	1	1246.00
5MBN[b]	22202	21007	1195	1	1195.00
4CHA	26990	25585	1405	5	281.00
2CGA	27707	23693	4014	5	802.80
2PSG	45170	37447	7723	3	2574.33
CLF (equals to the average of crosslinking numbers)±Standard Error					918.5±145.1

[a] In order to be consistent, the results presented in this table were derived from instantaneous unfolding method using standard bond length and angle values for both sets of data labeled "without crosslinks" and "with crosslinks" and then the *SASA* values were calculated using DSSP. [b] The heme containing proteins

ficients improved to 0.90, 0.88 and 0.99 for GdnHCL and urea m_{eq} as well as ΔCp values, respectively as shown in figure 2.

The extent of increase in *SASA* upon unfolding of a protein highly depends on the number of residues (*i.e.* protein size) and the constraints present in the unfolded state. The unfolded state of a protein is populated by an ensemble consisting huge number of conformationally distinct species. The presence of structural constraints limits the conformational space available to be explored by the protein polypeptide chain. Our analyses, in agreement with the results of others [10], show that the amount of area buried in each protein correlates very strongly (R=0.99) with the number of residues in each protein (Eq. 9). The strong correlation between $\Delta SASA$ and the number of residues, makes it possible to estimate the thermodynamic parameters using equations 10 to 12.

Where *k* denotes the number of residues for a given protein. These equations provide means to predict m_{eq} and ΔCp directly based on the primary structure information. The results of experimental studies are in close agreement with the results of our theoretical calculations which indicate the important thermodynamic parameters can be predicted using $\Delta SASA$ upon unfolding and taking into account the presence of crosslinks in the protein.

In a different approach to estimate *SASA* of unfolded state of proteins, we have used MD to simulate the unfolding behavior of proteins in denaturing condition, as stated in Materials and Methods section. Four of the proteins in our dataset (listed in Table 3) were subjected

$$\Delta SASA = 90.01 \times k - 804.55\,, \quad R = 0.99;\ N = 45 \quad \text{(Eq.9)}$$

$$m_{eq(GdnHCl)} = 22.55 \times k - 0.362(n \times \text{CLF}) + 835.25\,, \quad R = 0.89;\ N = 34 \quad \text{(Eq.10)}$$

$$m_{eq(Urea)} = 17.55 \times k - 0.173(n \times CLF) - 517.92\,, \quad R = 0.92\,,\ N = 34 \quad \text{(Eq.11)}$$

$$\Delta Cp = 18.48 \times k - 0.163(n \times CLF) - 327.208\,, \quad R = 0.99,\ N = 25 \quad \text{(Eq.12)}$$

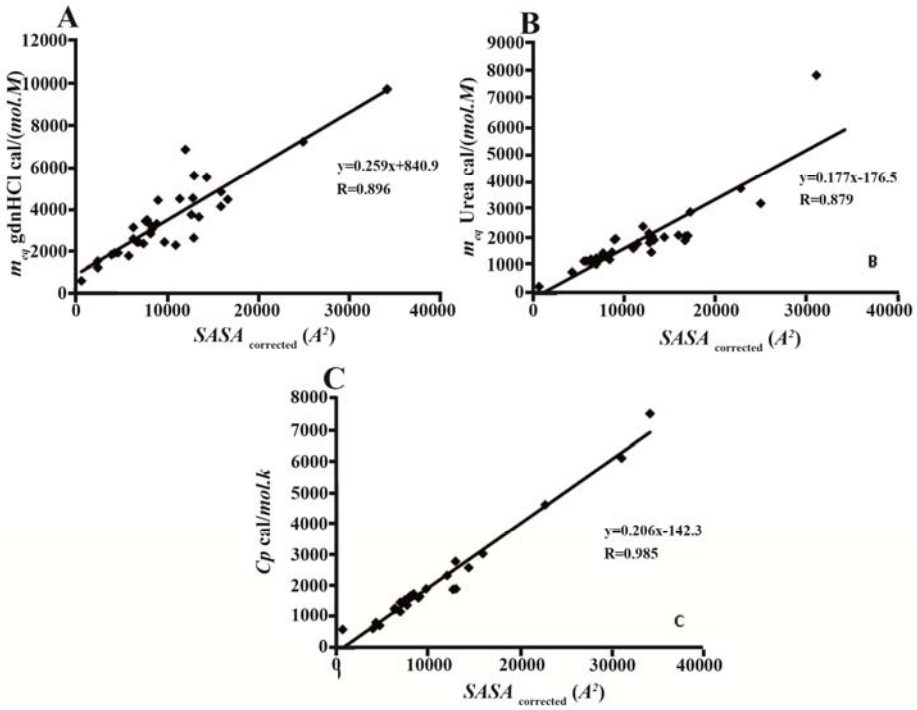

Figure 2. Dependence of A) m_{eq} value for GdnHCI denaturation, B) m_{eq} value for urea denaturation, and C) heat capacity changes upon unfolding on Δ*SASA* after correction for the effect of crosslinks by taking out 918.5 $Å^2$ per crosslink for the 45 proteins in our data set (see text for further explanation)

to MD simulations for 10 *ns* at 500 *°K* while inserted in a solvation box filled by a mixture of water and urea molecules. As can be seen from the table, the maximum *SASA* values for the unfolded conformations of proteins obtained by MD are smaller than that achieved by non-simulation method. Consequently, the *ΔSASA* values are also relatively smaller.

Figure 3 shows the snapshots of conformational changes during unfolding simulation of IgG binding domain of protein G (IBPG) which has 56 residues with no crosslink. As time evolves, both tertiary and secondary structures of IBPG are lost and at the same time its *SASA* increases. The maximum *SASA* achieved during 10 *ns* is 5772 $Å^2$ which is less than that estimated for fully extended conformation (8143 $Å^2$).

The presence of crosslinks in the unfolded state will result in a more compact unfolded form and the higher the number of crosslinks, the more pronounced is this effect. For example, as shown in figure 4, the unfolded conformation of lysozyme (hen egg white), a 129-residue protein with four disulfide bonds, retained more globular shape at the end of MD simulation, although it loses the elements of secondary structures.

As shown in table 3, the *SASAs* of the investigated proteins increased at the end of MD simulation. However, the extent of this increase is bigger for the protein with no crosslink. For example 1PGB which is a 56-residue protein without any crosslink showed

Table 3. Comparison of *SASA* and *ΔSASA* values obtained by different methods used to unfold the proteins

PDB code	*SASA* of native structure	*SASA* of the unfolded model		*ΔSASA* of the unfolding	
		Unfolding method		Unfolding method	
		MD simulation	Instantaneous	MD simulation	Instantaneous
1AKI[a]	6755	10412	12575	3657	5820
1AAL[b]	3993	5579	6539	1586	2546
2TRX[c]	5847	9135	13254	3288	7407
1PGB[d]	3752	5772	8143	2020	4391

a, b, c and d are 4, 2, 1, and zero, respectively and denote the number of crosslinks

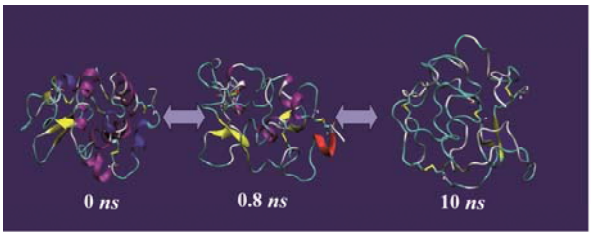

Figure 3. Molecular dynamics simulation of IgG binding domain of protein G (PDB code 1PGB) solvated in 4.4 *M* urea in water at 500 *°K* for 10 *ns* using GROMOS-96 force field parameters. The non-protein molecules (*i.e.* water and urea) are not shown for the sake of clarity

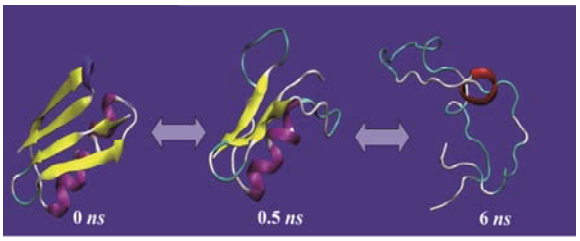

Figure 4. Molecular dynamics simulation of lysozyme (hen egg white) (PDB code 1AKI) solvated in 4.4 *M* urea in water at 500 *°K* for 10 *ns* using GROMOS-96 force field parameters

54% increase in *SASA* upon unfolding using MD method. However, applying the same unfolding condition on 1AAL, a protein with almost equal size (*i.e.* 58 residues) and two disulfide bond has led to only 40% increase in *SASA*. In all studied cases, the maximum *SASA* for unfolded conformations achieved by MD are smaller than that of instantaneous method.

Analyses of MD trajectories showed that the RMSD differences for C atoms increases as time evolves approaching high values in the range of ~14-19 *Å* for the studied proteins during the simulation. The rate of RMSD increase was dramatically fast for 1PGB and 2TRX, with no and one crosslink, respectively. However, the rate was gradual in the case of 1AAl and 1AKI with two and four crosslinks, respectively.

Although MD simulation under the condition used in this study can unfold the proteins and also demonstrates the effect of crosslink, however, using this method the *SASA* of unfolded conformations never reached to the *SASA* values of the unfolded conformations obtained by instantaneously decomposing the protein native structure just by taking into consideration to preserve the standard bond lengths, bond angles and other standard chemical structure geometries. This could be due to insufficient simulation time or entrapment of protein in an ensemble of conformations in a local minimum of energy landscape. However, to find out more about these issues and draw more sensible conclusion, further computational experiments such as hydrodynamic simulation are required.

Crosslinks such as disulfide bonds and heme groups have profound effect on the conformational flexibility and *SASA* values of unfolded state and hence influence the stability of the proteins by decreasing the entropy gain as well as reduction of *ΔSASA* upon unfolding. Studies of proteins with chemical crosslinks have shown clearly that the major effect of the crosslink on the stability results from a decrease in the conformational entropy of the unfolded molecule [23,24].

On the other hand, attempts to increase the stability of proteins through introducing disulfide bonds suggest that the structural restraints in the native state due to the crosslinks may also make an important contribution to the net effect of the crosslinks on the stability [25]. Furthermore, inspection of the model structures of Micro-myoglobin (Mb) revealed a role for heme in stabilizing the folded state [26].

Doig and Williams [27] investigated the effect of disulfide crosslinks on hydrophobicity derived stability of proteins. Based on data obtained from solvent transfer experiment, they calculated the non-polar Δ*SASA* to be 590 and 690 $Å^2$ per disulfide bond according to free energy of hydration and ΔC_p measurements, respectively. Taking into account that the fraction of total area buried which is non-polar is about 0.70, these values correspond to a reduction in the total area change of 850 $Å^2$ and 990 $Å^2$ per disulfide. Using solvent perturbation difference spectroscopy, Pace et al demonstrated that the solvent accessibility of the aromatic residues (Tyr and Trp) in three studied proteins (lysozyme, RNase A and RNase T1) was changed upon unfolding [22].

Table 4. Thermodynamic parameters of proteins predicted based on different methods

PDB code	Myers[a]			Predicted values using equations 13 to 15[b]			Experimental[c]		
	mGdnHCL	mUrea	ΔCp	$mGdnHCL_{pre}$	$mUrea_{pre}$	ΔCp_{pre}	mGdnHCL	mUrea	ΔCp
1CHO	1069.3	-94.1	-71.9	1117.5	-29.4	134.2	580.0	250.0	590.0
7PTI	1502.6	261.3	366.8	1516.4	245.5	444.7	1200.0	NA	NA
1AAL	1478.4	261.0	366.4	1491.8	245.5	444.7	1500.0	NA	NA
5CYT	2982.8	1283.7	1559.3	2834.1	1142.9	1426.2	2800.0	NA	NA
2PCB	3017.1	1315.3	1584.6	2849.9	1158.0	1430.2	3010.0	1200.0	1730.0
9RNT	2490.1	939.1	1168.1	2528.5	1007.8	1296.3	2560.0	1210.0	1270.0
1YCC	2813.1	1174.8	1448.4	2930.0	1218.6	1525.8	3400.0	1430.0	1370.0
2TRX	2906.8	1242.6	1507.0	2933.1	1224.2	1511.7	3310.0	1300.0	1660.0
9RSA	2476.8	951.5	1177.8	2212.8	1077.7	1392.4	3100.0	1100.0	1230.0
1AKI	2790.8	1156.5	1411.0	2442.8	1165.9	1456.8	2330.0	1290.0	1540.0
1LZ1	2874.4	1227.0	1488.9	2315.6	1182.8	1453.9	3460.0	NA	1580.0
1YMB	4147.5	2071.2	2506.9	3992.2	1998.9	2390.2	3710.0	2140.0	1870.0
5MBN	4258.2	2140.5	2523.8	4028.9	2019.7	2351.2	2600.0	1460.0	2770.0
4CHA	5039.8	2685.9	3171.5	4770.5	3152.4	3468.4	4100.0	2070.0	3020.0
2CGA	5230.5	2830.7	3317.1	4838.6	3307.1	3450.4	4440.0	2030.0	NA
2PSG	8892.8	4129.0	6354.9	8203.9	3964.0	6045.6	NA	7800.0	6090.0
Correlation Coefficient[d]	0.7293	0.7082	0.9751	0.6996	0.7519	0.9596			
MAPE	21.5	31.8	17.9	20.5	31.1	16.8			
SDEP	646.5	1128.3	300.9	612.1	1226.0	294.8			

a: Prediction of heat capacity changes and m_{eq} values for GdnHCL and Urea upon unfolding based on Myers' equations [10]. b: Same predictions using equations 13-15. c: Experimental data which are compiled from the literature and taken from reference [10]. d: Correlation coefficient between predicted and experimental values

Myers et al, used these experimental data to estimate an approximate average value of 900 $Å^2$ reduction in $\Delta SASA_{\text{unfolded}}$ per disulfide, assuming a universal change in accessibility across all residue types [10].

The results of these experimental methods have been averaged and used by Myers et al to compensate for the effects of crosslinks on $\Delta SASA$. However, it may suffer from an over simplification by using only the changes in accessibility of just two amino acids and extrapolating these changes to the total area. It should also be mentioned that these results have been concluded from very limited number of experiments performed on only three globular proteins [22].

One of the shortcomings of using either CLF, introduced in this work, or experimentally derived value of 900, introduced by Myers et al to compensate for the effects of crosslinks on $\Delta SASA$ and hence estimation of m_{eq} and ΔCp values is the scarcity of the data used. The correction value close to 900 $Å^2$ (proposed by Myers and here as CLF) can be justified by fitting equations 12 to 14 presented in Myers et al where the disulfide bond corrections that maximize the fits are all close to 900 $Å^2$. However, there is no need to use a correction factors such as 900 $Å^2$ proposed by Myers or CLF to account for the effects of crosslinks on m_{eq} or ΔCp. Although we believe the correction factor most likely is close to 900 $Å^2$, but it is not a magic number and any other value close to that can be used to do the correction and then draw empirical equations to relate m_{eq} or ΔCp to the corrected $\Delta SASA$ (or to the combination of number of amino acids and CLF as we used in here).

The coefficients in the final mathematical equations will be adjusted to balance out any

$$m_{eq(GdnHCl)} = 835.25 - 325.49 \times n + 22.55 \times k \quad R = 0.889;\ N = 34 \text{ (Eq.13)}$$

$$m_{eq(Urea)} = -517.92 - 155.85 \times n + 17.55 \times k \quad R = 0.878,\ N = 34 \text{ (Eq.14)}$$

$$\Delta Cp = -327.21 - 149.49 \times n + 18.48 \times k \quad R = 0.990,\ N = 25 \text{ (Eq.15)}$$

changes in the value of correction factor. In a situation where the ultimate aim is to be able to predict the thermodynamic parameters as precise as possible, one may decide to use different structural descriptors to derive empirical equations for the prediction purposes.

To this end we have furthered our investigation by trying to develop different empirical equations to predict m_{eq} and ΔCp values. We have examined the effects of different structural properties such as number of amino acids, number of crosslinks, size of loops representing the total number of amino acids involved in the loops formed by crosslinks, and the central position of the regions in the loop area on the prediction of the thermodynamic properties. The best statistical Multiple Linear Regression (MLR) models were achieved using variables representing total number of amino acids and the number of crosslinks.

In order to test the predictive power of these models, the Leave One Out (LOO) cross validation method was used. The mean absolute percentage error of predictions (MAPE) of m_{eq}(GdnHcl), m_{eq}(Urea) and ΔCp values for all proteins listed in table 1 (or proteins with crosslinks listed in table 4) based on Myers' models are 19.7 (21.5), 22.9 (31.8) and 13.8 (17.9). The corresponding MAPEs using models presented in equations 13, 14 and 15 are 22.3 (20.5), 23.6 (31.1) and 13.5 (16.8), respectively.

The results show that both methods are not statistically different in predicting the evaluated thermodynamic parameters, either for all data points (proteins in Table 1) or for the proteins with crosslinks (*i.e.* values indicated inside the brackets), and simple MLR equations based on limited number of structural descriptors, *i.e.* number of residues and number of crosslinks, are able to perform equally well. In fact equations 13 to 15 are identical to equations 10 to 12 and the only difference is the way to represent the effect of crosslink on the parameter of interest. For example in equation 13 the coefficient of variable n (*i.e.* number of crosslinks) equals to the coefficient of the second term on the right hand side of the equation 10 multiplied by the value of CLF. These equations are also too close to the equations proposed by Myers et al (eqs. 12 to 14 in reference 10). For instance, the coefficient of variable n in equation 14 above (*i.e.* 155.58) is very close to the 139.3 calculated by multiplying 0.14 and 995 in equation 13 in Myers' study [10].

Conclusion

In summary, it can be concluded that the proposed relationships represent valuable tools for predicting thermodynamic parameters of protein folding using the primary sequence information. The proposed crosslinking factor (CLF; which shows the effect of a single crosslink on $\Delta SASA$ upon unfolding) of 918.5 $Å^2$ obtained based on computational simulation is very close to the previously published experimentally derived value of 900 $Å^2$. Such a correction factor can be used to estimate the $\Delta SASA$ upon unfolding which in turn can be used for the prediction of thermodynamic parameters such as m_{eq} and ΔCp. For the prediction of these parameters, one may also use number of amino acids (k) and number of crosslinks (n) without need to any kind of correction factor.

Although the correction factor for the effect of crosslink on $\Delta SASA$ is a quantitative value describing a fundamental property in protein folding, however, for the prediction purposes, the use of more simple properties taken from the primary structure of proteins gives as well accurate results. In addition, the current work demonstrates an example where theory is capable of reproducing the results obtained from experimental works.

Acknowledgement

Authors would like to thank Research Office of Tabriz University of Medical Sciences and Iran National Science Foundation for providing financial support.

References

1. Uhlen M, Ponten F. Antibody-based proteomics for

human tissue profiling. Mol Cell Proteomics 2005; 4(4):384-393.

2. Berg JM, Tymoczko JL, Stryer L. Biochemistry. 6th ed. New York: W.H. Freeman; 2006.
3. Morris MB, Dastmalchi S, Church WB. Rhodopsin: structure, signal transduction and oligomerisation. Int J Biochem Cell Biol 2009;41(4):721-724.
4. Winklhofer KF, Tatzelt J, Haass C. The two faces of protein misfolding: gain- and loss-of-function in neurodegenerative diseases. EMBO J 2008;27(2): 336-349.
5. Stefani M. Protein misfolding and aggregation: new examples in medicine and biology of the dark side of the protein world. Biochim Biophys Acta 2004; 1739(1):5-25.
6. Whitford D. Proteins structure and function. Chichester: John Wiley & Sons Ltd; 2005.
7. Shirley BA, Urea and guanidine hydrochloride denaturation curves. In: Shirley BA (ed). Protein stability and folding. Iowa City: Humana Press Inc; 1995, 177-190.
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000;28(1):235-242.
9. Bernado P, Blackledge M, Sancho J. Sequence-specific solvent accessibilities of protein residues in unfolded protein ensembles. Biophys J 2006;91 (12):4536-4543.
10. Myers JK, Pace CN, Scholtz JM. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. Protein Sci 1995; 4(10):2138-2148.
11. Godoy-Ruiz R, Ariza F, Rodriguez-Larrea D, Perez-Jimenez R, Ibarra-Molero B, Sanchez-Ruiz JM. Natural selection for kinetic stability is a likely origin of correlations between mutational effects on protein energetics and frequencies of amino acid occurrences in sequence alignments. J Mol Biol 2006;362(5):966-978.
12. Schellman JA. Protein stability in mixed solvents: a balance of contact interaction and excluded volume. Biophys J 2003;85(1):108-125.
13. Geierhaas CD, Nickson AA, Lindorff-Larsen K, Clarke J, Vendruscolo M. BPPred: a Web-based computational tool for predicting biophysical parameters of proteins. Protein Sci 2007;16(1):125-134.
14. Freire E. The termodynamic linkage between protein structure, stability, and function. In: Murphy KP (ed). Protein structure, stability, and folding. Iowa City: Humana Press; 2001, 37-68.
15. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577-2637.
16. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. Electrophoresis 1997;18(15): 2714-2723.
17. van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS: fast, flexible, and free. J Comput Chem 2005;26(16):1701-1718.
18. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy minimization and dynamics calculation. J Comput Chem 1983;4(2):187-217.
19. van Gunsteren WF, Billeter SR, Eising AA, Hünenberger PH, Krüger P, Mark AE, et al. Biomolecular Simulation: The GROMOS96 Manual and User Guide. Zürich: Vdf Hochschulverlag AG an der ETH Zürich. 1996, 41-82.
20. Livingstone JR, Spolar RS, Record MTJ. Contribution to the thermodynamics of protein folding from the reduction in water-accessible nonpolar surface area. Biochemistry 1991;30(17):4237-4244.
21. Spolar RS, Livingstone JR, Record MTJ. Use of liquid hydrocarbon and amide transfer data to estimate contributions to thermodynamic functions of protein folding from the removal of nonpolar and polar surface from water. Biochemistry 1992;31 (16):3947-3955.
22. Pace CN, Laurents DV, Erickson RE. Urea denaturation of barnase: pH dependence and characterization of the unfolded state. Biochemistry 1992;31 (10):2728-2734.
23. Perry LJ, Wetzel R. The role of cysteine oxidation in the thermal inactivation of T4 lysozyme. Protein Eng 1987;1(2):101-105.
24. Pace CN, Grimsley GR, Thomson JA, Barnett B. Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. J Biol Chem 1988;263(24):11820-11825.
25. Wetzel R. Harnessing disulfide bonds using protein engineering. Trends Biochem Sci 1987;12:478-482.
26. Ji HF, Shen L, Grandori R, Müller N. The effect of heme on the conformational stability of micromyoglobin. FEBS J 2008;275(1):89-96.
27. Doig AJ, Williams DH. Is the hydrophobic effect stabilizing or destabilizing in proteins? The contribution of disulphide bonds to protein stability. J Mol Biol 1991;217(2):389-398.