In silico Evaluation of Crosslinking Effects on Denaturant \( m_{eq} \) values and \( \Delta C_P \) 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 \( m_{eq} \) values and heat capacity changes (\( \Delta C_P \)) 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 \( \Delta \text{SASA} \) and ultimately on \( m_{eq} \) and \( \Delta C_P \) 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 \( \Delta C_P \) values and the effect of crosslinks on \( \Delta \text{SASA} \) of unfolding was investigated. The results were used to predict the \( m_{eq} \) of denaturation for guanidine hydrochloride and urea, as well as \( \Delta C_P \) 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.

Keywords: Crosslinks, Disulfides, Protein stability, Thermodynamics

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-GdnHCl) 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 (\( \Delta G_U \)) depends linearly on the denaturant concentration as:
$\Delta G_U = \Delta G_U^{H_2O} - m_U [\text{Denaturant}]$ (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 ($\Delta SASA_i$) upon unfolding has been used as a mean for predicting the $\Delta H$ as presented below:

$$\Delta H = \sum_i \alpha_i \left( \bar{r} \right) \cdot \Delta SASA_i$$ (Eq.2)

Where $\Delta SASA_i$ is the change in SASA of atom $i$ upon unfolding, and $\alpha_i \left( \bar{r} \right)$ 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 ($\Delta C_p$) 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. $\Delta C_p$ is correlated to the changes in SASA upon unfolding, as shown in the following equation:

$$\Delta C_p = \sum_i a_i \cdot \Delta SASA_i$$ (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 $\Delta H$ and $\Delta C_p$ 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 $\Delta C_p$) 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 $\Delta C_p$ 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 K and the trajectories were saved every 0.02 ns.

\[
\text{Crosslink ng number} = \frac{SASA_{nc} - SASA_c}{n} \quad (\text{Eq.4})
\]

\[
\text{Crosslink ng factor} = \frac{1}{N} \sum_{i=1}^{N} \text{Crosslink ng number} \quad (\text{Eq.5})
\]
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 $\text{SASA}$ 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:

$$SDEP = \sqrt{\frac{\sum_{i=1}^{N} (A_{exp} - A_{calc})^2}{N}}$$ (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:

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

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

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

Where $N$ is the number of data points.

Results and Discussion

The changes in solvent accessible surface area ($\Delta \text{SASA}$) 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, $\Delta C_p$ 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.Å²) for GdnHCl and urea, respectively, indicating the stronger denaturing effects of GdnHCl.

Denaturation heat capacity changes ($\Delta C_p$) were also correlated with the $\Delta \text{SASA}$ strongly with the correlation coefficient of 0.97 as shown in figure 1C. The same linear correlations between $m_{eq}$ values and $\Delta \text{SASA}$ have been shown previously by Myers et al (10). $\Delta \text{SASA}$ has been also related linearly to $\Delta C_p$ by others (20,21).
The main purpose of this study is to re-evaluate the effect of crosslinks on ΔSASA and also predict the \( m_{eq} \) and \( ΔC_p \) 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 accessibility of the molecule.
Effect of Crosslinking on Protein Unfolding

myers et al (10) employed the results of different empirical methods (22) to estimate the magnitude of the reduction of solvent accessible surface area ($\Delta SASA$) per disulfide bond. The reduction of $\Delta SASA$ per crosslink was estimated to be about 900Å$^2$.

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 $\Delta SASA$ 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 Å$^2$. This value is the average of crosslinking numbers calculated for 16 crosslink-containing proteins listed in table 2 for which the $m_{eq}$ and $\Delta Cp$ 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 Å$^2$) is smaller than the average of the numbers (991.0 Å$^2$) 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 $\Delta SASA$ values were corrected for the effect of crosslinks on the solvent accessibility of the unfolded state by taking 918.5 Å$^2$ per crosslink off the $\Delta SASA$ (called $\Delta SASA_{corrected}$) and then the corrected values were re-correlated to the $m_{eq}$ and $\Delta Cp$ values. Linear correlation coef-

![Figure 1](http://www.ajmb.org)

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 $\Delta SASA$ for the 45 proteins shown in table 1.
coefficients improved to 0.90, 0.88 and 0.99 for GdnHCl and urea $m_{eq}$ as well as $\Delta C_p$ 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 $\Delta C_p$ 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

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

<table>
<thead>
<tr>
<th>PDB code</th>
<th>$SASA_{unfolded}$ without Crosslinks $^a$</th>
<th>$SASA_{unfolded}$ with Crosslinks $^a$</th>
<th>$\Delta SASA$</th>
<th>Number of crosslinks ($n$)</th>
<th>Crosslinking number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1CHO</td>
<td>7039</td>
<td>5730</td>
<td>1309</td>
<td>3</td>
<td>436.33</td>
</tr>
<tr>
<td>7PTI</td>
<td>8278</td>
<td>6616</td>
<td>1662</td>
<td>2</td>
<td>831.00</td>
</tr>
<tr>
<td>1AAI</td>
<td>8315</td>
<td>6539</td>
<td>1776</td>
<td>2</td>
<td>888.00</td>
</tr>
<tr>
<td>5CYT $b$</td>
<td>13929</td>
<td>13392</td>
<td>537</td>
<td>1</td>
<td>537.00</td>
</tr>
<tr>
<td>2PCB $b$</td>
<td>14410</td>
<td>14118</td>
<td>292</td>
<td>1</td>
<td>292.00</td>
</tr>
<tr>
<td>9RNT</td>
<td>13348</td>
<td>11227</td>
<td>2121</td>
<td>2</td>
<td>1060.50</td>
</tr>
<tr>
<td>1YCC $b$</td>
<td>15145</td>
<td>14620</td>
<td>525</td>
<td>1</td>
<td>525.00</td>
</tr>
<tr>
<td>2TRX</td>
<td>13568</td>
<td>13254</td>
<td>314</td>
<td>1</td>
<td>314.00</td>
</tr>
<tr>
<td>9RSA</td>
<td>17125</td>
<td>13015</td>
<td>4110</td>
<td>4</td>
<td>1027.50</td>
</tr>
<tr>
<td>1AKI</td>
<td>17597</td>
<td>12575</td>
<td>5022</td>
<td>4</td>
<td>1255.50</td>
</tr>
<tr>
<td>1LZI</td>
<td>18698</td>
<td>12967</td>
<td>5731</td>
<td>4</td>
<td>1430.25</td>
</tr>
<tr>
<td>1YMB $b$</td>
<td>22110</td>
<td>20864</td>
<td>1246</td>
<td>1</td>
<td>1246.00</td>
</tr>
<tr>
<td>5MBN $b$</td>
<td>22202</td>
<td>21007</td>
<td>1195</td>
<td>1</td>
<td>1195.00</td>
</tr>
<tr>
<td>4CHA</td>
<td>26990</td>
<td>25585</td>
<td>1405</td>
<td>5</td>
<td>281.00</td>
</tr>
<tr>
<td>2CGA</td>
<td>27707</td>
<td>23693</td>
<td>4014</td>
<td>5</td>
<td>802.80</td>
</tr>
<tr>
<td>2PSG</td>
<td>45170</td>
<td>37447</td>
<td>7723</td>
<td>3</td>
<td>2574.33</td>
</tr>
</tbody>
</table>

CLF (equals to the average of crosslinking numbers)$\pm$Standard Error $918.5\pm145.1$

$^a$ 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

$$\Delta SASA = 90.01 \times k - 804.55$$

$$m_{eq}(\text{GdnHCl}) = 22.55 \times k - 0.362(n \times CLF) + 835.25$$

$$m_{eq}(\text{Urea}) = 17.55 \times k - 0.173(n \times CLF) - 517.92$$

$$\Delta C_p = 18.48 \times k - 0.163(n \times CLF) - 327.208$$

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 \( \text{SASA} \) values for the unfolded conformations of proteins obtained by MD are smaller than that achieved by non-simulation method. Consequently, the \( \Delta \text{SASA} \) 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 \( \text{SASA} \) increases. The maximum \( \text{SASA} \) 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 \( \text{SASAs} \) of the investigated proteins increased at the end of MD simulation. However, the extent of this increase is bigger for the protein without any crosslink. For example 1PGB which is a 56-residue protein without any crosslink showed

---

Table 3. Comparison of \( \text{SASA} \) and \( \Delta \text{SASA} \) values obtained by different methods used to unfold the proteins

<table>
<thead>
<tr>
<th>PDB code</th>
<th>( \text{SASA} ) of native structure</th>
<th>( \text{SASA} ) of the unfolded model</th>
<th>( \Delta \text{SASA} ) of the unfolding</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unfolding method</td>
<td>Unfolding method</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MD simulation</td>
<td>Instantaneous</td>
<td>MD simulation</td>
</tr>
<tr>
<td>1AKI(^a)</td>
<td>6755</td>
<td>10412</td>
<td>12575</td>
</tr>
<tr>
<td>1AAL(^b)</td>
<td>3993</td>
<td>5579</td>
<td>6539</td>
</tr>
<tr>
<td>2TRX(^c)</td>
<td>5847</td>
<td>9135</td>
<td>13254</td>
</tr>
<tr>
<td>1PGB(^d)</td>
<td>3752</td>
<td>5772</td>
<td>8143</td>
</tr>
</tbody>
</table>

\( a, b, c \) and \( d \) are 4, 2, 1, and zero, respectively and denote the number of crosslinks.
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 ΔSASA to be 590 and 690 Å² 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 buried which is non-polar is about 0.70, these values correspond to a reduction in the total area change of 850 Å² and 990 Å² 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).
Myers et al. used these experimental data to estimate an approximate average value of 900 Å² reduction in ΔSASA 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 ΔSASA. However, it may suffer from an oversimplification 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 ΔSASA and hence estimation of $m_{eq}$ and ΔCp values is the scarcity of the data used. The correction value close to 900 Å² (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 Å². However, there is no need to use a correction factors such as 900 Å² 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 Å², 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 ΔSASA (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

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

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

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

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

<table>
<thead>
<tr>
<th>PDB code</th>
<th>Myers⁶</th>
<th>Predicted values using equations 13 to 15b</th>
<th>Experimental⁷</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GdnHCl</td>
<td>Urea</td>
<td>ΔCp</td>
</tr>
<tr>
<td>1CHO</td>
<td>1069.3</td>
<td>-94.1</td>
<td>-71.9</td>
</tr>
<tr>
<td>7PTI</td>
<td>1502.6</td>
<td>261.3</td>
<td>366.8</td>
</tr>
<tr>
<td>1AAL</td>
<td>1478.4</td>
<td>261.0</td>
<td>366.4</td>
</tr>
<tr>
<td>5CVT</td>
<td>2982.8</td>
<td>1283.7</td>
<td>1559.3</td>
</tr>
<tr>
<td>2PCB</td>
<td>3017.1</td>
<td>1315.3</td>
<td>1584.6</td>
</tr>
<tr>
<td>9RNT</td>
<td>2490.1</td>
<td>939.1</td>
<td>1168.1</td>
</tr>
<tr>
<td>1YCC</td>
<td>2813.1</td>
<td>1174.8</td>
<td>1448.4</td>
</tr>
<tr>
<td>2TRX</td>
<td>2906.8</td>
<td>1242.6</td>
<td>1507.0</td>
</tr>
<tr>
<td>9RSA</td>
<td>2476.8</td>
<td>951.5</td>
<td>1177.8</td>
</tr>
<tr>
<td>1AKI</td>
<td>2790.8</td>
<td>1156.5</td>
<td>1411.0</td>
</tr>
<tr>
<td>1LZI</td>
<td>2874.4</td>
<td>1227.0</td>
<td>1488.9</td>
</tr>
<tr>
<td>1YMB</td>
<td>4147.5</td>
<td>2071.2</td>
<td>2506.9</td>
</tr>
<tr>
<td>5MBN</td>
<td>4125.8</td>
<td>2140.5</td>
<td>2523.8</td>
</tr>
<tr>
<td>4CHA</td>
<td>5039.8</td>
<td>2685.9</td>
<td>3171.5</td>
</tr>
<tr>
<td>2CGA</td>
<td>5230.5</td>
<td>2830.7</td>
<td>3317.1</td>
</tr>
<tr>
<td>2PSG</td>
<td>8892.8</td>
<td>4129.0</td>
<td>6354.9</td>
</tr>
<tr>
<td>3: MAPE</td>
<td>0.7293</td>
<td>0.7082</td>
<td>0.9751</td>
</tr>
<tr>
<td>3: SDEP</td>
<td>21.5</td>
<td>31.8</td>
<td>17.9</td>
</tr>
<tr>
<td>3: Correlation Coefficient</td>
<td>646.5</td>
<td>1128.3</td>
<td>300.9</td>
</tr>
</tbody>
</table>
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 $\Delta C_p$ 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 $\Delta C_p$ 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 Å² obtained based on computational simulation is very close to the previously published experimentally derived value of 900 Å². 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 $\Delta C_p$. 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


