Proceedings of the XLII Zakopane School of Physics, Zakopane 2008

Tests of the Structure-Based Models of Proteins

M. Cieplak[a] and J.I. Sułkowska[a,b]

[a]Institute of Physics, Polish Academy of Sciences
al. Lotników 32/46, 02-668 Warsaw, Poland

[b]CTBP, University of California, San Diego
Gilman Dr. 9500, La Jolla 92037, USA

The structure-based models of proteins are defined through the condition that their ground state coincides with the native structure of the proteins. There are many variants of such models and they yield different properties. Optimal variants can be selected by making comparisons to experimental data on single-molecule stretching. Here, we discuss the 15 best performing variants and focus on fine tuning the selection process by adjusting the velocity of stretching to match the experimental conditions. The very best variant is found to correspond to the 10-12 potential in the native contacts with the energies modulated by the Miyazawa–Jernigan statistical potential and variable length parameters. The second best model incorporates the Lennard–Jones potential with uniform amplitudes. We then make a detailed comparison of the two models in which theoretical surveys of stretching properties of 7510 proteins were made previously.

PACS numbers: 87.80.Nj, 87.15.ap, 87.14.E–

1. Introduction

All-atom molecular dynamics simulations of proteins are capable of generating a detailed picture of the dynamics of a protein on short timescales, usually not exceeding a fraction of a μs. However, the expected tasks of molecular level modeling extend currently to much longer timescales, as when considering protein folding, and to much larger system sizes, such as assemblies of proteins and other biomolecules. Virus capsids and ribosomes [1] offer examples of such assemblies. These new tasks require developing new methods of modeling. A natural route of action is to develop effective models in which the number of degrees of freedom is reduced. One way to do it is to represent amino acids by single beads located at the C^{α} atoms and tether them together into chains by harmonic interactions with minima at 3.8 Å. In order to prevent interpenetration of those beads which are not sequential neighbors, they are endowed with repulsive cores with a radius which is often taken to be around 4 Å [2–4].

Such chain molecules represent homopolymers and adopting them as models of proteins requires adding two steps: introducing a local backbone stiffness, so that sharp twists in the backbone are unlikely to take place, and bringing in attractive interactions between certain beads, so that the system may acquire globular forms, including the one corresponding to the native state of a protein. It is not an easy task to derive the attractive interactions at the coarse-grained level from atomic level considerations. However, a successful phenomenological way out has been proposed by Go and his collaborators [5] as aptly summarized by Takada [6]. The Go-like approach takes the structure of the native state as the defining experimental characteristic of a protein and introduces the attractive couplings so that the resulting ground state of the chain coincides with the native state of the protein. This description is structure-based and not sequence based.

Clearly, there are many ways to implement the Go-like prescription and in Ref. [7] we consider 62 variants of the resulting Hamiltonians. These variants can be characterized by attributes summarized by

$$\text{model} = \left\{V^{\text{NAT}}, S, \mathcal{M}, E\right\}, \tag{1}$$

where the first term denotes selection of a pairwise potential, the second identifies the nature of the local backbone stiffness, the third defines the contact map (i.e. the list of pairs of beads that can attract mutually), and the fourth determines the nature of selection of the energy parameters in the potentials.

The various variants of the Go-like models lead to a broad spectrum of physical properties and it is desirable that this ambiguity in the choice is reduced. Since these models should work in the vicinity of the native state the best, a proper test of the properties should be provided by making comparisons to experimental data on stretching of proteins. The stretching process starts in or near the native state and is usually accomplished by an atomic force microscope. Table lists the proteins that have been studied experimentally and it specifies the maximum values, $F_{\max}$, of the force of resistance to stretching at constant speed.

TABLE

Comparison between experimentally measured values $F_{\max}$ with theoretical predictions in $\{6\text{–}12, C, M3, E_0\}$ and $\{6\text{–}12, C, M3, E_0\}$ Go-like models. 10 trajectories are considered. When two results are given, two distinct trajectories are observed and the averages are split into groups.

PDB	$F^{\rm e}_{\max}$ [pN]	v_p [nm/s]	$M3$:$F^{\rm t}_{\max}$ [ε/Å]	$M2$:$F^{\rm t}_{\max}$ [ε/Å]		Refs.
1tit	204±30	600	2.15	2.04	I27*8	[12, 13]
1nct	210±10	500	2.4±0.2	2.38	I54-I59	[14, 15]
1g1c	127±10	600	2.3±0.2	2.17	I5 titin	[16]
1b6i	64±30	1000	1.2	1.52	T4 lysozyme(21-141)	[17]
1aj3	68±20	3000	1.23	1.57	spectrin R16	[18]
1qjo	15±10	600	1.2	1.4±0.2	eE2lip3(N-C)	[19]
1qjo	177±10	600	2.0	2.1	E2lip3(N-41)	[19]
1dqv	60±15	600	1.5	1.5	calcium binding C2A	[20]
1rsy	60±15	600	1.7±0.2	1.5	calcium binding C2A	[20]
1byn	60±15	600	1.4	1.45	calcium binding C2A	[20]
1cfc	<20	600	0.55	0.79	calmodulin	[20]
1n11	37±9	0.2	0.4	0.59	ankyrin*1	[21, 22]
1bni	70±15	300	1.4, 1.7	1.45	barnase/i27	[23]
1bnr	70±15	300	1.05	1.45	barnase/i27	[23]
1bny	70±15	300	1.1, 1.3	1.25±0.2	barnase/i27	[23]
1hz6	152±10	700	3.5	3.05	protein L	[24]
1hz5	152±10	700	2.8	2.5	protein L	[24]
2ptl	152±10	700	2.2±0.2	2.1	protein L	[24]
1ksr	45±20	350	2.0±0.3	1.9±0.3	DdFLN −4	[25, 26]
2rn2	19±10	700	1.8±0.2	1.8±0.2	ribonuclease H	[27]
1ubq	230±34	1000	2.32	2.4	ubiquitin	[28]
1ubq	203±35	410	2.32	2.4	ubiqutin(N-C)*9	[28, 29]
1ubq	85±20	300	0.9	0.9	ubiquitin(K48-C)*(2-7)	[28, 29]
1emb	350±30	3600	5.15±0.4	4.28	GFP(3-132)	[30]
1emb	130±30	3600	2.3, 4.3	2.48	GFP(3-212)	[30]
1emb	120±30	3600	2.2±0.2	2.52	GFP(132-212)	[30]
1emb	104±40	3600	2.3±0.2	2.33	GFP(N-C)	[31]
1fnf	75±20	3000	1.6, 1.8	1.4, 1.7	Fniii-10	[32, 33]
1ttf	75±20	600	0.7, 1.2	0.6, 1.19	Fniii-10	[34]
1ttg	75±20	600	0.7, 1.0	0.6, 1.27	Fniii-10	[34]
1fnh	124±18	600	1.8	1.75±0.2	Fniii- 12	[33]
1fnh	89±18	600	1.4, 1.7	1.87±0.2	Fniii-13	[33]
1oww	220±31	600	2.1±0.2	2.15	FNiii-1	[33]
1ten	135±40	500	1.7	1.85	TNFNiii-3	[33, 35]
1pga	190±20	400	2.4±0.2	1.8, 2.1	protein G	[36]
1gb1	190±20	400	1.65±0.2	2.1	protein G	[36]

2. The models

In order to select the optimal variants of the models, we have performed statistical tests based on evaluation of two parameters: the Pearson correlation coefficient, R^2, and Theil's U coefficient. They are defined by

$$R^2 = 1 - \sqrt{\frac{1}{D}\sum_{\lambda=1}^{D}\left(\frac{F^{\rm t}_\lambda - F^{\rm e}_\lambda}{F^{\rm e}_\lambda}\right)^2} \qquad (2)$$

and

$$U = \sqrt{\frac{\sum_{\lambda=1}^{D-1}(W_{\lambda+1} - w_{\lambda+1})^2}{\sum_{\lambda=1}^{D-1}(w_{\lambda+1})^2}}, \tag{3}$$

where

$$W_{\lambda+1} = \frac{F^{\rm t}_{\lambda+1} - F^{\rm e}_{\lambda}}{F^{\rm e}_{\lambda}}, \qquad w_{\lambda+1} = \frac{F^{\rm e}_{\lambda+1} - F^{\rm e}_{\lambda}}{F^{\rm e}_{\lambda}}. \tag{4}$$

Here, $F^{\rm t}_{\lambda}$ denotes a theoretical value of $F_{\max}$ for protein λ whereas $F^{\rm e}_{\lambda}$ — the corresponding experimental value. $W_{\lambda+1}$ and $w_{\lambda+1}$ denote the predicted and actual relative single-step changes, respectively. In the definition of U, the proteins are assumed to be rank ordered from the smallest experimental value to the largest and U measures deviations in the local slopes.

The best models should have R^2 close to 1 and U close to zero. Out of the 62 variants considered, 15 were found to perform substantially better than the other and here we focus on this set here. Their full list is contained in Table 5 of Ref. [7]. The original assessment has singled out model

$$\left\{\text{6-12}, C, M3, E^0\right\} \tag{5}$$

as the best performer. Here, 6-12 denotes the Lennard–Jones potential in a contact

$$V^{\text{6-12}} = 4E_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], \tag{6}$$

where r_{ij} is the distance between the C^{α}'s in amino acids i and j whereas σ_{ij} is determined pair-by-pair so that the minimum is located at the experimentally established native distance $r^{\rm n}_{ij}$, i.e., $\sigma_{ij} = r^{\rm n}_{ij}/\sqrt[6]{2}$. E^0 means that the energy parameters E_{ij} are chosen to be uniform and equal to ε which is of order 1–1.5 kcal/mole. The symbol C means that the local backbone stiffness is described by a term which favors the native sense of the local chirality [2, 8]. The chirality-based stiffness is approximately equivalent to favoring of the native values of the dihedral angles [7]. An alternative is to have an angular stiffness, denoted as A, in which also the bonds tend to take the native values as in Ref. [9]. Finally, $M3$ means that the contact map is determined in the following fashion. The heavy atoms are assigned by van der Waals radii as in Refs. [10, 11]. The radii are multiplied by 1.24 to account for attraction. In this way, each amino acid is represented by a cluster of spheres. If these spheres overlap in the native state, the corresponding pair of amino acids is said to have a native contact. This procedure may select some $i, i+2$ pairs as forming a contact. However, such contacts are usually weak as being due to van der Waals dispersion interactions. In the $M3$ map such $i, i+2$ contacts are discarded whereas in an $M2$ map they are kept. Models with the angular stiffness usually come together with the $M4$ contact map in which contacts which are more local than $i, i+4$ are not considered.

The assessment that chooses $\left\{\text{6-12}, C, M3, E^0\right\}$ as the optimal variant has involved pulling theoretically at the same speed of about 0.005 Å/ns, i.e. 500,000 nm/s, which is 3 orders of magnitude faster than the typical experimental speeds of 600–700 nm/s as listed in Table (the "steered" all atom simulations [37] usually involve speeds which are 7 orders of magnitude bigger than the experimental). Table also indicates that certain proteins have been studied at higher speeds (e.g. the speed used to study 1emb was 6 times larger than the one used to study 1tit). The values of $F_{\max}$ depend on the pulling speed in a logarithmic fashion (see, e.g. Ref. [28]) so it is appropriate to account for this dependence. When we rescale the theoretical pulling speeds to imitate the experimental variations in the conditions of stretching, then the statistical assessment is changed [7]. Its results for the 15 models are shown in Fig. 1. Five of these models are highlighted by providing the lists of their attributes. The $\left\{\text{6-12}, C, M3, E^0\right\}$ becomes the second best model whereas the very top place goes to

$$\left\{\text{10-12}, A, M4, E^{\rm HB,MJ}\right\}, \tag{7}$$

in which the potential has the 10-12 form,

$$V^{\text{10-12}} = E_{ij}\left[5\left(\frac{r^{\rm n}_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{r^{\rm n}_{ij}}{r_{ij}}\right)^{10}\right] \tag{8}$$

and $E^{\rm HB,MJ}$ assigns the amplitudes in this potential in a non-uniform way. Specifically an energy of ε is assigned to hydrogen-bonded amino acids (the bond between the N and C atoms; the details and modifications are discussed in Ref. [7]), an energy related to the Miyazawa–Jernigan couplings [38], $\varepsilon^{\rm MJ}_{ij}$, to non-hydrogen-bond side chain–side chain contacts, and again ε to all other contacts. The relation of the energy to $\varepsilon^{\rm MJ}_{ij}$ is through $\varepsilon\varepsilon^{\rm MJ}_{ij}/\langle\varepsilon^{\rm MJ}_{ij}\rangle$, where the energy is over 210 different possible pairs of amino acids.

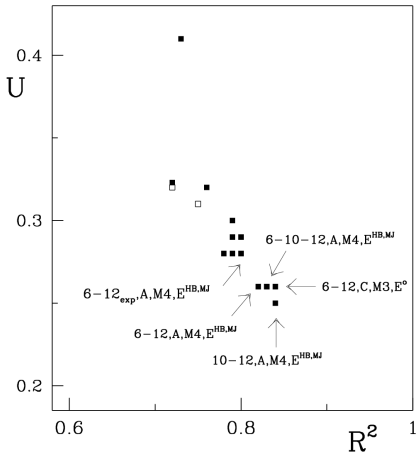

Fig. 1. Correlation between R^2 and U coefficients for the top 15 models studied. This is an analog of Fig. 6 in Ref. [7], but with an implementation of adjustment in the pulling speeds to imitate the experimental conditions. The open squares correspond to models with poor folding properties. These are $\left\{\text{6–12}, C, M3, E^{\rm HB,MJ}\right\}$ and $\left\{\text{6-10-12}, C, M3, E^{\rm HB,MJ}\right\}$.

This best model is very closely related to another well performing model $\left\{\text{6-10-12}, A, M4, E^{\rm HB,MJ}\right\}$ that

has been introduced by Karanicolas and Brooks [39] in which the following potential is used:

$$V^{\text{6-10-12}} = 4E_{ij}\left[13\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 18\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} + 4\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]. \quad (9)$$

The fourth best model is $\{\text{6-12}, A, M4, E^{\text{HB,MJ}}\}$. The fifth top model, also highlighted in Fig. 1, with the $6\text{–}12_{\text{exp}}$ potential containing an exponential hump [7]. Thus, generally, the models with the Miyazawa–Jernigan couplings are performing the best. However, the simplest model with the uniform couplings and the Lennard–Jones potentials is the most economical in description and the second best in the stretching tests.

3. Results and discussion

All of the results quoted here have been obtained at a fixed temperature which is close to optimal folding temperature and is expected to be close to the room temperature (at $0.3\varepsilon/k_{\text{B}}$ in the case of $\{\text{6-12}, C, M3, E^0\}$) as discussed in Ref. [7]. The fifth-order predictor-corrector integration method, the thermostating, and enhancement of the disulphide bonds are explained in more detail in Ref. [40]. It should be noted that some of the models may perform well in the stretching tests and yet lead to poor folding properties. These models are shown by the open symbols in Fig. 1.

We now focus on one application of the structure--based models discussed here: making a survey of the elastic properties of proteins whose structures are deposited in the Protein Data Bank. The survey was based on 7510 proteins comprising not more than 150 amino acids and it involved pulling by the terminal amono acids. It was done by using two variants of the models: $\{\text{6-12}, C, M2, E^0\}$ in Ref. [40] and $\{\text{6-12}, C, M3, E^0\}$ in Ref. [41]. The former variant has R^2 of 0.85 and U of 0.23 in the test involving stretching, compared to 0.89 and 0.22 for the latter variant (when the pulling speed is not adjusted). The object of the survey has been to determine theoretical values of F_{max}, correlate the values of F_{max} with structural classes, architectures, and topologies, identify proteins which should be particularly strong in their resistance to pulling, and understand what makes them so. We refer the reader to the original articles. Here, we consider only the differences between the two models as seen in the set of the strongest proteins. 134 proteins were designated as strong in Ref. [41] (the top 1.8% forces) and 137 proteins in Ref. [40] (the top 2%). Here, we consider the common core of 134 proteins. Even though the two models select the same 134 proteins as the strongest of all considered proteins, the values of F_{max} found are different because the contact map is the most important attribute that defines a structure-based model. Figure 2 shows the values of F_{max} obtained with the $M3$ contact map versus the corresponding values for the $M2$ contact map. The two sets of the data points are correlated with an average slope of 0.92 meaning that the $M3$ contact map yields lower forces more often than the $M2$ one, especially in the α class of the proteins. The top strongest four proteins have the same identity in both models: 1c4p, 1qqr, 1g1k, and 1aoh consecutively when counting from the strongest. These proteins have long stretches of parallel β strands that stretch under shearing. These strands experience additional stabilization from other strands in the neighborhood [40, 41]. The ranking of remaining strong proteins is otherwise not related in the two models as testified by the scattered nature of the data points shown in Fig. 3. Of course, if all 7510 proteins were correlated in terms of their ranking

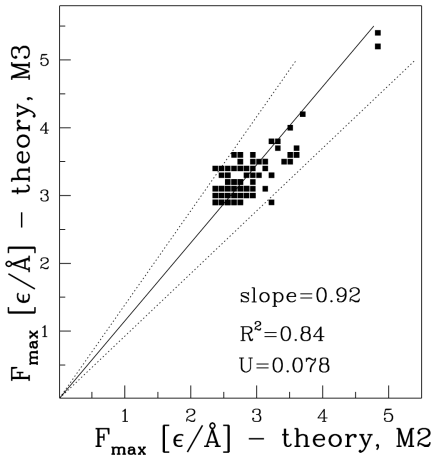

Fig. 2. The values of F_{max} obtained in two uniform Lennard–Jones models with the chirality-based stiffness: one with the $M2$, and the other with the $M3$ contact map. The unit of force, $\varepsilon/\text{Å}$ is of order 71 pN [7]. Only the strongest 134 proteins are shown. The theoretical resolution is usually of order 0.1 so many data points overlap.

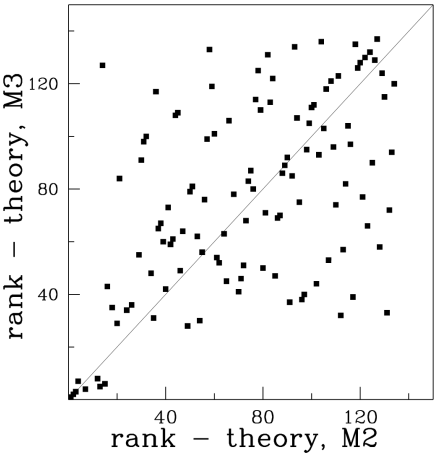

Fig. 3. Similar to Fig. 2 but for the ranking numbers associated with a protein in the two models. Rank number 1 corresponds to the very strongest protein: 1c4p.

number, then the top strongest 134 proteins would appear correlated in a distinct way, since they would all be at the top.

It is customary to think that a structure-based model is just one theoretical object that leads to unique results when applied. This paper illustrates the fact that there are many of its variants and that they may perform differently.

Acknowledgments

This work has been supported by the grant N N202 0852 33 from the Ministry of Science and Higher Education in Poland.

References

[1] V. Tozzini, J. Trylska, C. Chang, J.A. McCammon, *J. Struct. Biol.* **157**, 606 (2007).

[2] M. Cieplak, T.X. Hoang, M.O. Robbins, *Proteins: Struct. Funct. Bio.* **49**, 114 (2002).

[3] M. Cieplak, T.X. Hoang, *Biophys. J.* **84**, 475 (2003).

[4] M. Cieplak, T.X. Hoang, M.O. Robbins, *Proteins: Struct. Funct. Bio.* **56**, 285 (2004).

[5] H. Abe, N. Go, *Biopolymers* **20**, 1013 (1981).

[6] S. Takada, *Proc. Natl. Acad. Sci. USA* **96**, 11698 (1999).

[7] J.I. Sułkowska, M. Cieplak, *Biophys. J.* **95**, 3174 (2008).

[8] J.I. Kwiecinska, M. Cieplak, *J. Phys., Condens. Matter* **17**, S1565 (2005).

[9] C. Clementi, H. Nymeyer, J.N. Onuchic, *J. Mol. Biol.* **298**, 937 (2000).

[10] J. Tsai, R. Taylor, C. Chothia, M. Gerstein, *J. Mol. Biol.* **290**, 253 (1999).

[11] G. Settanni, T.X. Hoang, C. Micheletti, A. Maritan, *Biophys. J.* **83**, 3533 (2002).

[12] M. Rief, M. Gautel, F. Oesterhelt, J.M. Fernandez, H.E. Gaub, *Science* **276**, 1109 (1997).

[13] M. Carrion-Vasquez, A.F. Oberhauser, S.B. Fowler, P.E. Marszalek, S.E. Broedel, J. Clarke, J.M. Fernandez, *Proc. Natl. Acad. Sci. USA* **96**, 3694 (1999).

[14] K. Watanabe, C. Muhle-Goll, M.S.Z. Kellermayer, S. Labeit, H.L. Granzier, *Struct. Biol. J.* **137**, 248 (2002).

[15] K. Watanabe, P. Nair, D. Labeit, M.S.Z. Kellermayer, M. Greaser, S. Labeit, H.L. Granzier, *J. Biol. Chem.* **277**, 11549 (2002).

[16] H.B. Li, J.M. Fernandez, *J. Mol. Biol.* **334**, 75 (2003).

[17] G. Yang, C. Cecconi, W.A. Baase, I.R. Vetter, W.A. Breyer, J.A. Haack, B.W. Matthews, F.W. Dahlquist, C. Bustamante, *Proc. Natl. Acad. Sci. USA* **97**, 139 (2000).

[18] P.F. Lenne, A.J. Raae, S.M. Altmann, M. Saraste, J.K.H. Horber, *FEBS Lett.* **476**, 124 (2000).

[19] D.J. Brockwell, E. Paci, R.C. Zinober, G. Beddard, P.D. Olmsted, D.A. Smith, R.N. Perham, S.E. Radford, *Nat. Struct. Biol.* **10**, 731 (2003).

[20] M. Carrion-Vazquez, A.F. Oberhauser, T.E. Fisher, P.E. Marszalek, H. Li, J.M. Fernandez, *Prog. Biophys. Mol. Biol.* **74**, 63 (2000).

[21] G. Lee, K. Abdi, Y. Jiang, P. Michaely, V. Bennett, P.E. Marszalek, *Nature* **440**, 246 (2006).

[22] L.W. Li, S. Wetzel, A. Pluckthun, J.M. Fernandez, *Biophys. J.* **90**, 30 (2006).

[23] R.B. Best, B. Li, A. Steward, V. Daggett, J. Clarke, *Biophys. J.* **81**, 2344 (2001).

[24] D.J. Brockwell, S. Godfrey, S. Beddard, E. Paci, Dan K. West, P.D. Olmsted, D. Alastair Smith, S.E. Radford, *Biophys. J.* **89**, 506 (2005).

[25] I. Schwaiger, A. Kardinal, M. Schleicher, A.A. Noegel, M. Rief, *Nat. Struct. Mol. Biol.* **11**, 81 (2004).

[26] M. Schlierf, M. Rief, *J. Mol. Biol.* **345**, 497 (2005).

[27] C. Cecconi, E.A. Shank, C. Bustamante, S. Marqusee, *Science* **309**, 2057 (2005).

[28] C.L. Chyan, F.C. Lin, H. Peng, J.M. Yuan, C.H. Chang, S.H. Lin, G. Yang, *Biophys. J.* **87**, 3995 (2003).

[29] M. Carrion-Vazquez, H. Li, H. Lu, P.E. Marszalek, A.F. Oberhauser, J.M. Fernandez, *Nat. Struct. Biol.* **10**, 738 (2003).

[30] H. Dietz, M. Rief, *Proc. Natl. Acad. Sci. USA* **103**, 1244 (2006).

[31] H. Dietz, M. Rief, *Proc. Natl. Acad. Sci. USA* **101**, 16192 (2004).

[32] L. Li, H. Han-Li Huang, C.L. Badilla, J.M. Fernandez, *J. Mol. Biol.* **345**, 817 (2005).

[33] A.F. Oberhauser, C. Badilla-Fernandez, M. Carrion-Vazquez, J.M. Fernandez, *J. Mol. Biol.* **319**, 433 (2002).

[34] Y. Oberdorfer, H. Fuchs, A. Janshoff, *Langmuir* **16**, 9955 (2000).

[35] A.F. Oberhauser, P.E. Marszalek, H.P. Erickson, J.M. Fernandez, *Nature* **14**, 181 (1998).

[36] Y. Cao, H. Li, *Nature Mater.* **6**, 109 (2007).

[37] H. Lu, K. Schulten, *Chem. Phys.* **247**, 141 (1999).

[38] S. Miyazawa, R.L. Jernigan, *J. Mol. Biol.* **256**, 623 (1996).

[39] J. Karanicolas, C.L. Brooks III, *Protein Sci.* **11**, 2351 (2002).

[40] J.I. Sułkowska, M. Cieplak, *J. Phys., Condens. Matter* **19**, 283201 (2007).

[41] J.I. Sułkowska, M. Cieplak, *Biophys. J.* **94**, 6 (2008).