Factors governing fibrillogenesis of polypeptide chains using lattice models

Mai Suan Li[1], Nguyen Truong Co[2], Govardhan Reddy[3], C-K. Hu[5,6], J. E. Straub[7], and D. Thirumalai[3,4]
[1]*Institute of Physics, Polish Academy of Sciences, Al. Lotnikow 32/46, 02-668 Warsaw, Poland*
[2]*Saigon Institute for Computational Science and Technology, 6 Quarter, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam*
[3]*Biophysics Program, Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742*
[4]*Department of Chemistry and Biochemistry, University of Maryland, College Park, MD 20742*
[5]*Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan*
[6]*Center for Nonlinear and Complex Systems and Department of Physics, Chung Yuan Christian University, Chungli 32023, Taiwan* [7]*Department of Chemistry, Boston University, Boston, Massachusetts 02215*

Using lattice models we explore the factors that determine the tendencies of polypeptide chains to aggregate by exhaustively sampling the sequence and conformational space. The morphologies of the fibril-like structures and the time scales (τ_{fib}) for their formation depend on a subtle balance between hydrophobic and coulomb interactions. The extent of population of an ensemble of $\mathbf{N}^*$ structures, which are fibril-prone structures in the spectrum of conformations of an isolated protein, is the major determinant of τ_{fib}. This observation is used to determine the aggregation-prone consensus sequences by exhaustively exploring the sequence space, thus providing a basis for genome wide search of fragments that are aggregation prone.

PACS numbers: 87.15.A,87.14.E

Proteins that are unrelated by sequence or structure aggregate to form amyloid-like fibrils with a characteristic cross β-structures, which are linked to a number of deposition diseases such as Alzheimer's and prion-disorders [1](a). The observation that almost any protein could form fibrils seemed to imply that fibril rates can be predicted solely based on sequence composition and the propensity to adopt global secondary structure. Such a conclusion has limited validity because it does not account for fluctuations that populate aggregation-prone structures. Despite the common structural characteristics of amyloid fibrils [1](b)-(e) the factors that determine the fibril formation tendencies are not understood.

Experiments on fibril formation times (τ_{fib}) have been rationalized using global factors such as the hydrophobicity of side chains [2](a), net charge [2](b,c), patterns of polar and non-polar residues [2](d), frustration in secondary structure elements [2](e,f), and aromatic interactions [2](g). However, the inability to sample the sequences and conformational spaces exhaustively [3] has prevented deciphering plausible general principles that govern protein aggregation using limited computations and experiments. The purpose of this letter is to obtain a quantitative correlation between intrinsic properties of polypeptide sequences and their fibril growth rates using lattice models, which have given remarkable insights into the general principles of protein folding and aggregation [4]. Using a modification of the model in [5] we explore the sequence-dependent variations of τ_{fib} on the nature of conformations explored by the monomer. We highlight the role of aggregation-prone ensemble of $\mathbf{N}^*$ structures [6] in the folding landscape of the monomer in determining τ_{fib} and the propensity of sequences to form fibrils.

Lattice model. To explore the dependence of fibril formation rates on the intrinsic properties of monomers and the sequence, we use a lattice model [5] in which each chain consists of M connected beads that are confined to the vertices of a cube. The simulations are done using N identical chains with $M = 8$. The peptide sequence which is used to illustrate the roles of electrostatic and hydrophobic interaction is +HHPPHH- (Fig. 1), where H, P, + and - are hydrophobic, polar, positively charged and negatively charged beads respectively [5].

The energy of N chains is [5] $E = \sum_{l=1}^{N}\sum_{i<j}^{M} E_{sl(i)sl(j)}\delta(r_{ij} - a) + \sum_{m<l}^{N}\sum_{i,j}^{M} E_{sl(i)sm(j)}\delta(r_{ij} - a)$, where r_{ij} is the distance between residues i and j, a is a lattice spacing, $sm(i)$ indicates the type of residue i from m-th peptide, and $\delta(0) = 1$ and zero, otherwise. The first and second terms represent intrapeptide and interpeptide interactions, respectively.

The propensity of polar and charged residues to be "solvated" is mimicked using $E_{P\alpha}$ =-0.2 (in the units of hydrogen bond energy ϵ_H), where α= P,+,or -. To

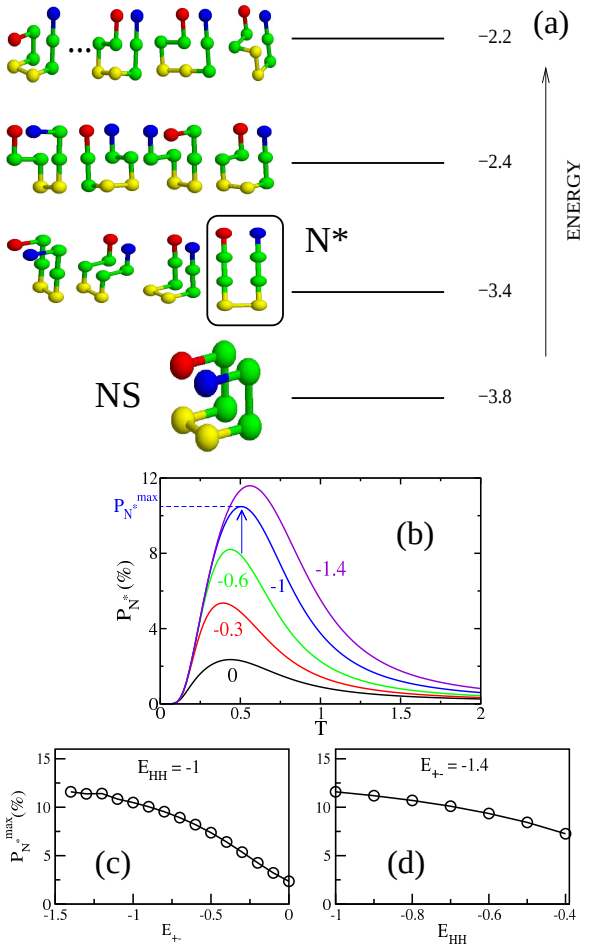

FIG. 1: (a) Spectrum of energies and associated low energy structures of the monomer sequence +HHPPHH-. H, P, + and - are in green, yellow, blue, and red, respectively. We set $E_{HH} = -1$ and $E_{+-} = 1.4$. There are a total 1831 possible conformations that are spread among 17 possible energy values. The conformations in the first excited state represent the ensemble of $\mathbf{N}^*$ structures and the $\mathbf{N}^*$ conformation that coincides with the peptide state in the fibril (see Fig. 2a) is enclosed in a black colored box. (b) The probability P_{N^*} of populating the structure in the box in (a) as a function of T for $E_{+-} = 0, -0.3, 0.6, -1$ and -1.4 keeping $E_{HH} = -1$. The arrow indicates T^*, where $P_{N^*} = P_{N^*}^{max}$. Dependence of $P_{N^*}^{max}$ on E_{+-} for $E_{HH} = -1$ (c), and on E_{HH} for $E_{+-} = -1.4$ (d).

assess the importance of electrostatic and hydrophobic interactions, we vary either E_{+-} in the interval $-1.4 \leq E_{+-} \leq 0$ or E_{HH} between -1 and 0. If E_{+-} is varied, we set $E_{HH} = -1$, while if E_{HH} is varied, then $E_{+-} = -1.4$. We used $E_{++} = E_{--} = -E_{+-}/2$ and all other contact interactions have $E_{\alpha\beta} = 0.2$.

Monomer spectra depends on E_{+-} and E_{HH}. The spectrum of energy states of the monomer for a given sequence is determined by exact enumeration of all possible conformations (Fig. 1). For all sets of contact energies chosen above the native state (NS) of the monomer is compact (lowest energy conformation in Fig. 1a). In anticipation of the role of ensemble of $\mathbf{N}^*$ structures (the conformations in the first excited state in Fig. 1a) play in promoting fibril formation we focus on the change in the rank order of the $\mathbf{N}^*$ energy level as E_{+-} is varied. For $E_{+-} < 0$, the ensemble of $\mathbf{N}^*$ structures are the first excited state (Fig. 1a). However, if $E_{+-} = 0$, the ensemble of $\mathbf{N}^*$ structures are part of the 19-fold degenerate states in the second excited state (see Supplementary Information (SI) Fig. 1).

The population of the putative fibril-prone conformation in the monomeric state is $P_{N^*} = \exp(-E_{N^*})/Z$, where Z is the partition function is obtained by exact enumeration. Fig. 1b, shows the temperature dependence of P_{N^*} for various values of (E_{+-}) interaction, with $E_{HH} = -1$ and other contact energies constant. Depending on E_{+-}, the maximum value of P_{N^*} varies from 2% $< P_{N^*}^{max} <$ 12% (Fig. 1c). $P_{N^*}^{max}$ decreases to a lesser extent as the hydrophobic interaction grows (Fig. 1d). Here we consider only $E_{HH} \leq -0.4$ because the fibril-like structure is not the lowest-energy when $E_{HH} > -0.4$.

Morphology of lowest-energy structures of multi-chain systems depends on sequences. When multiple chains are present in the unit cell, aggregation is readily observed, and in due course they lead to ordered structures. We used the Monte Carlo (MC) [5] annealing protocol, which allows for an exhaustive conformational search, to find the lowest energy conformation. For non-zero values of E_{+-} the chains adopt an antiparallel arrangement in the ordered protofilament, which ensures that the number of salt-bridge and hydrophobic contacts are maximized (Fig. 2a and see also Ref. [5]). If $E_{+-} = 0$ then the lowest energy fibril structure has a vastly different architecture even though they are assembled from $\mathbf{N}^*$ (Fig. 2b). The structure in Fig. 2b, in which a pair of $\mathbf{N}^*$ conformations are stacked by flipping one with respect to the other is rendered stable by maximizing the number of +P and -P contacts. We now set $E_{+-} = -1.4$ and vary E_{HH}. For $E_{HH} < -0.4$, the fibril conformation adopts the same shape as that shown in Fig. 2a, but for $E_{HH} = -0.4$ the energetically more favorable double-layer structure emerges (Fig. 2c). If $E_{HH} \geq -0.3$, then the lowest-energy conformation ceases to have the fibril-like shape (Fig. 2d). The close packed heterogenous structure is stitched together by a mixture of the NS conformation and one of the second excited conformations. Even for this simple model a variety of lowest-energy structures of oligomers and protofilaments with different morphologies emerge, depending on a subtle balance between electrostatic and hydrophobic interactions.

Dependence of τ_{fib} on E_{+-} and E_{HH}. Simulations were performed by enclosing N chains in a box with periodic boundary conditions and move sets described in ref. [5]. The effect of finite size is discussed in SI, Fig. 2. The fibril formation time τ_{fib} is defined as an average of first passage times needed to reach the fibril state with the lowest energy starting from initial random conformations. For a given value of T, we generated 50-100 MC trajectories to obtain reliable estimates of τ_{fib}. We

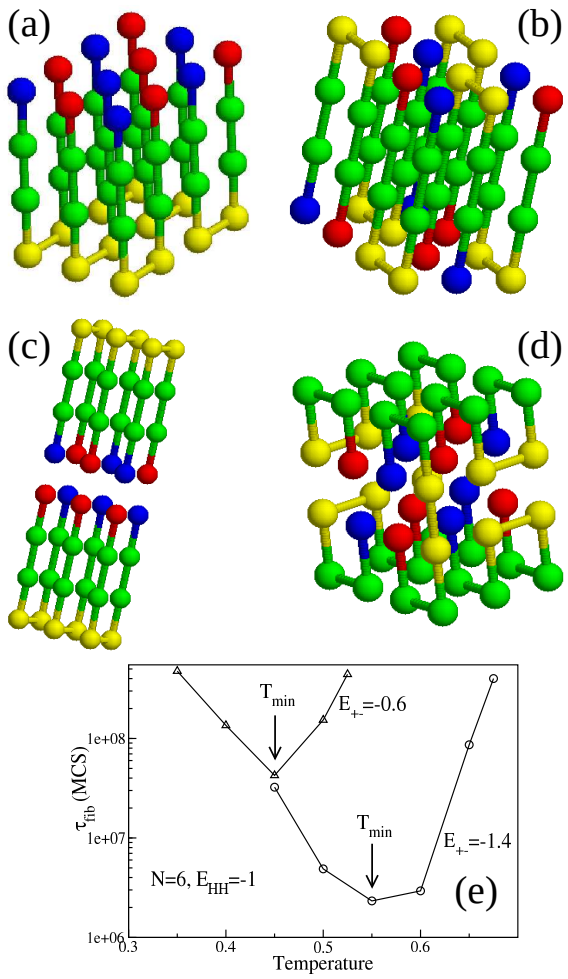

FIG. 2: (a) The lowest energy fibril structure for $E_{+-} = -1.4$ and $E_{HH} = -1$. (b) Same as in (a) but with $E_{+-} = 0$. (c) Double layer structure for $E_{HH} = -0.4$ but keeping $E_{+-} = -1.4$. (d) For $E_{+-} = -1.4$ and $E_{HH} = -0.3$ the fibril structure is entirely altered. (e) Temperature dependence of τ_{fib} for $E_{+-} = -1.4$ (circles) and $E_{+-} = -0.6$ (triangles). $N = 6$ and $E_{HH} = -1$. Arrows show the temperatures at which the fibril formation is fastest.

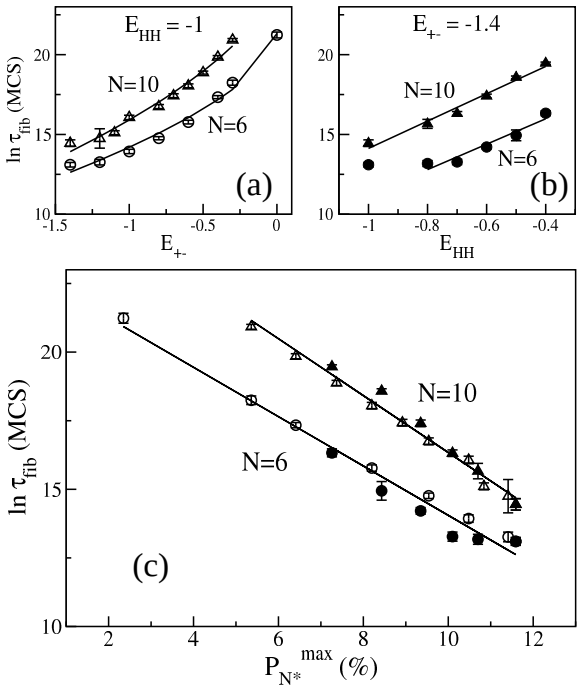

FIG. 3: (a) Dependence of τ_{fib} on E_{+-} for $N = 6$ (circles) and $N = 10$ (triangles) with $E_{HH} = -1$. The solid curves are fits to $y = c_0 + c(-x)^{\alpha}$, where $\alpha \approx 0.59$. $c_0 = 21.32$ and $c = -7.12$ and $c_0 = 25.14$ and $c = -9.23$ for $N = 6$ and 10, respectively. (b) Dependence of τ_{fib} on E_{HH} with $E_{+-} = -1.4$ hold constant for $N = 6$ (solid circle) and $N = 10$ (solid triangles). Lines are fits $y = 19.17 + 7.97x$ and $y = 22.69 + 8.56x$ for $N = 6$ and 10, respectively. For $N = 6$ the first point $E_{HH} = -1$ is excluded from fitting. (c) Dependence of τ_{fib} on $P_{N^*}^{max}$ for $N = 6$ and 10. Symbols are the same as in (a) and (b) τ_{fib} is measured in MCS and $P_{N^*}^{max}$ in %. The correlation coefficient for all fits $R \approx 0.98$.

measure time in units of Monte Carlo steps (MCS). The combination of local and global moves constitutes one MCS.

We performed an exhaustive study of the dependence of τ_{fib} on the number of chains in the simulation box, N (SI, Fig. 2). For highly favorable interaction between the terminal charged residues, $E_{+-} = -1.4$, τ_{fib} scales linearly with the size of the system (SI, Fig. 2a), while for less favorable interactions, $E_{+-} = -0.6$ and -0.8, $\ln(\tau_{fib})$ scales linearly with the size of the system (SI, Fig. 2b). The average energy per monomer, $\langle E \rangle / N$, for $E_{+-} = -1.4$ scales linearly as a function of $1/N$ (SI, Fig. 2c). The temperature dependence of τ_{fib} displays a U-shape (Fig. 2e) and the fastest assembly occurs at T_{min}, which roughly coincides with the temperature, T^*, where P_{N^*} reaches maximum (Fig. 1b). To probe the correlation between τ_{fib} and E_{+-} and E_{HH} we performed simulations at T_{min}. The dependence of τ_{fib} on E_{+-} can be fit using $\tau_{fib} \sim \exp[-c(-E_{+-})^{\alpha}]$ where $\alpha \approx 0.6$ and the constant $c \approx 7.12$ and 9.23 for $N = 6$ and 10 respectively (Fig. 3a). Thus, variation of E_{+-} drastically changes not only the morphology of the ordered protofilament (Fig. 2), but also τ_{fib}. As the strength of the charge interaction between the terminal beads increases, the faster is the fibril formation process. Interestingly, the fibril formation rate at $E_{+-} = 0$ is about four orders of magnitude slower than that at $E_{+-} = -1.4$. Our model suggests that the propensity to fibril assembly strongly depends on the charge states of the polypeptide sequences [1](f).

By fixing $E_{+-} = -1.4$ we calculated the dependence of τ_{fib} on the hydrophobic interaction (Fig. 3b), which may be approximated using $\tau_{fib} \sim \exp(cE_{HH})$. Here constant $c \approx 7.97$ and 8.56 for $N = 6$ and 10, respectively. For $N = 10$, a change in hydrophobicity of $\Delta E_{HH} = 0.6$, leads to self-assembly rates that are more than two orders of magnitude. Thus, enhancement of hydrophobic interactions speeds up fibril formation rates [1, 8].

Fibril formation rates depend on P_{N^*}. The dramatic variations in τ_{fib} on E_{+-} and E_{HH} prompted us to link the underlying spectrum of monomer conformations to τ_{fib}. A plot of the data in Fig. 3a and 3b as a function of $P_{N^*}^{max}$ (Fig. 3c) yields the surprising relation

$$\tau_{fib} = \tau_{fib}^0 \exp(-cP_{N^*}^{max}), \quad (1)$$

where the prefactor $\tau_{fib}^0 \approx 1.014 \times 10^{10}$ MCS and 3.981×10^{11} MCS, and $c \approx 0.9$ and 1.0, for $N = 6$ and 10, respectively. Eq. 1 is also valid for three other degenarate conformations in the N* ensemble, which are structurally similar to the one enclosed in the box in Fig. 1a. There are a few implications of the central result given in Eq. 1. (i) The sequence-dependent spectrum of the monomer is a harbinger of fibril formation. In proteins there are multiple $\mathbf{N}^*$ conformations corresponding to distinct free energy basins of attraction [6](b). Aggregation from each of the structures in the various basins of attraction could lead to fibrils with different morphologies (polymorphism) that cannot be captured using lattice models. (ii) Enhancement of P_{N^*} either by mutation or chemical cross linking should increase fibril formation rates. Indeed, a recent experiment [9] showed that the aggregation rate of Aβ_{1-40}-lactam[D23-K28], in which the residues D23 and K28 are chemically constrained by a lactam bridge, is nearly a 1000 times greater than in the wild-type. Since the salt bridge constraint increases the population of the $\mathbf{N}^*$ conformation in the monomeric state [10], it follows from Eq. 1, τ_{fib} should decrease. (iii) Since $P_{N^*}(T)$ depends on the spectrum of the precise sequence foe a given set of external conditions, it follows that the entire free energy landscape of the monomer [6](b) and not merely the sequence composition as ascertained else where[1](f), should be considered in the predictions of the amyloidogenic tendencies of a particular sequence. (iv) Eq. 1 is suggestive of a fluctuation-driven nucleation mechanism with a complicated temperature dependence. (v) Finally, as a negative control we plot $\ln(\tau_{fib})$ as a function of P_C^{max}, where C represents a conformation from the second or the third excited state (SI, Fig. 3). The plots clearly show that Eq. 1 does not hold for these structures and it holds only for an ensemble of $\mathbf{N}^*$ structures.

Sequence space scanning. We further exploit the result in Eq. 1 to determine the amylome [11], the universe of sequences in the lattice model, that can form fibrils. We posit that aggregation prone sequences are those with a unique native state with a maximum in $P_{N^*}(T)$ in the interval $1.0 \leq T^*/T_F \leq 1.25$. If $T_F = 300K$, which is physically reasonable, $T^* = 375K$ if $\frac{T^*}{T_F} = 1.25$. Thus, for values of $\frac{T^*}{T_F} > 1.25$ T^* would be far too high to be physically relevant. Moreover, our conclusions will not change by increasing $\frac{T^*}{T_F}$ or alternatively by choosing a reasonable threshold value for P_{N^*}. Out of the 65,536 sequences only 217 satisfy these criteria (see Supplementary Information (SI) for details). The sequence space exploration shows that there is a high degree of correlation between the positions of charged and hydrophobic residues leading to a limited number of aggregation prone sequences with +HHPPHH- being an example. In addition, there are substantial variations in T^*/T_F for sequences with identical sequence composition, which reinforces the recent finding [11] that context in which charged and hydrophobic residues are found is important in the tendency to form amyloid-like fibrils.

There is a strong correlation between the extent of population of $\mathbf{N}^*$ structures (in proteins we expect multiple aggregation prone conformations), which depend on the exact sequence and the environment. Due to the strong dependence of τ_{fib} on $P_{N^*}^{max}$ we suggest that only limited number of sequences are aggregation prone (see SI for details). Although the conclusions were obtained using lattice models, we expect them to hold for peptide aggregation. Our study also provides a basis for genome wide search for consensus sequences with propensity to aggregate.

The work was supported by the Ministry of Science and Informatics in Poland (grant No 202-204-234), grants NSC 96-2911-M 001-003-MY3 & AS-95-TP-A07, National Center for Theoretical Sciences in Taiwan, and NIH Grant R01GM076688-05.

[1] (a) F. Chiti, and C. M. Dobson, Annual Rev. Biochemistry **75**, 333 (2006); (b) A. T. Petkova, Y. Ishii, J. Balbach, O. Antzutkin, R. Leapman, F. Delaglio, and R. Tycko, Proc. Natl. Acad. Sci. USA **99**, 16742 (2002); (c) R. Tycko, Quart. Review Biophys. **39**, 1 (2006); (d) D. J. Selkoe, Nature **426**, 900 (2003); (e) M. Sunde, and C. Blake, Adv. Protein Chem. **50**, 123 (1997); (f) F. Chiti, M. Stefani, N. Taddei, G. Ramponi, and C. M. Dobson, Nature **424**, 805 (2003).

[2] (a) D. E. Otzen, O. Kristensen, and M. Oliveberg, Proc. Natl. Acad. Sci. (USA) **97**, 9907 (2000); (b) F. Massi, D. Klimov, D. Thirumalai, J. E. Straub, Prot. Sci. **11**, 1639 (2002); (c) F. Chiti, M. Calamai, N. Taddei, M. Stefani, G. Ramponi, and C. M. Dobson, Proc. Natl. Acad. Sci. (USA) **99**, 16419 (2002); (d) M. W. West, W. X. Wang, J. Patterson, J. D. Mancias, J. R. Beasley, and M. H. Hecht, Proc. Natl. Acad. Sci. (USA) **96**, 11211 (1999); (e) Y. Kallberg, M. Gustafsson, B. Persson, J. Thyberg, and J. Johansson, J. Biol. Chem. **276**, 12945 (2001); (f) R. I. Dima, and D. Thirumalai, Biophys. J. **83**, 1268 (2002); (g) E. Gazit, FASEB **16**, 77 (2002).

[3] (a) D. K. Klimov, and D. Thirumalai, Structure **11**, 295 (2003); (b) G. Bellesia, and J. E. Shea, Biophys. J. **96**, 875 (2009); (c) M. L. de la Paz, G. M. S. de Mori, L. Serrano, and G. Colombo, J. Mol. Biol. **349**, 583 (2005); (d) D. W. Li, S. Mohanty, A. Irback, and S. H. Huo,

PLOS Comp. Biol. **4**, e1000238 (2008).

[4] (a) D. K. Klimov, and D. Thirumalai, J. Chem. Phys. **109**, 4119 (1998); (b) E. Shakhnovich, Chem. Rev. **106**, 1559 (2006); (c) P. Gupta, C. K. Hall, and A. C. Voegler, Prot. Sci. **7**, 2642 (1998); (d) R. I. Dima, and D. Thirumalai, Prot. Sci. **11**, 1036 (2002); (e) M. Maiti, M. Rao, and S. Sastry, Eur. Phys. J. E **32**, 217 (2010); (f) M. Cieplak, T. X. Hoang, and M. S. Li, Phys. Rev. Lett. **83**, 1684 (1999); (g) M. S. Li. D. K. Klimov, and D. Thirumalai, Phys. Rev. Lett. **93**, 268107 (2004).

[5] M. S. Li, D. K. Klimov, J. E. Straub, and D. Thirumalai, J. Chem. Phys. **129**, 175101 (2008).

[6] (a) D. Thirumalai, R. I. Dima, and D. K. Klimov, Curr. Opin. Struct. Biol. **13**, 146 (2003); (b) B. Tarus, J. E. Straub, and D. Thirumalai, J. Am. Chem. Soc. **128**, 16159 (2006).

[7] (a) H. J. Hilhorst, and J. M. Deutch, J. Chem. Phys. **63**, 5153 (1975); (b) M. S. Li, D. K. Klimov, and D. Thirumalai, J. Phys. Chem. B **106**, 8302 (2002).

[8] C. J. Bowerman, D. M. Ryan, D. A. Nissan and B. L. Nilsson, Mol. BioSysystems **5**, 1058 (2009).

[9] K. Sciarretta, D. Gordon, A. Petkova, A. Tycko, and S. Meredith, Biochemistry **44**, 6003 (2005).

[10] G. Reddy, J. E. Straub, and D. Thirumalai, J. Phys. Chem. B **113**, 1162 (2009).

[11] L. Goldschmidt, P. K. Teng, R. Riek, and D. Eisenberg Proc. Natl. Acad. Sci. (USA) **107**, 3487 (2010).