Hidden structure in protein energy landscapes

Dengming Ming,[1,2] Marian Anghel,[1] and Michael E. Wall[1,3,*]

[1]*Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[2]*School of Life Sciences, Nanjing University, Nanjing, Jiangsu 210093, People's Republic of China*
[3]*Bioscience Division and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

Inherent structure theory is used to discover strong connections between simple characteristics of protein structure and the energy landscape of a Gō model. The potential energies and vibrational free energies of inherent structures are highly correlated, and both reflect simple measures of networks of native contacts. These connections have important consequences for models of protein dynamics and thermodynamics.

PACS number(s): 87.15.A−

Protein activity is controlled by dynamical transitions among conformational substates [1]; the transitions may be understood in terms of motions on an energy landscape [2]. Substates correspond to local minima in the energy landscape, and transitions correspond to the hurdling of barriers between minima. Interestingly, the protein energy landscape resembles that of glasses [3].

Spin-glass models have yielded insight into properties of protein energy landscapes [4,5] and the kinetics of protein folding [6]. The main motivation for using spin-glass models rather than structural-glass models is that spin-glass models are more analytically tractable; however, it has long been recognized that structural-glass models might be better-suited to describe proteins [5]. Indeed, protein unfolding has been characterized as a rigidity transformation that is similar to those seen in network glasses [7].

Structural-glass-forming liquids have been fruitfully characterized using inherent structure (IS) theory [8,9], which treats the energy landscape as a set of discrete basins that are separated by saddles. Each basin contains a local minimum, called an inherent structure, which is analogous to a protein conformational substate. The dynamics are then naturally described as vibrations about local minima, punctuated by transitions between neighboring basins. A key assumption in IS theory is that vibrations are similar about minima with the same potential energy; however, importantly, IS theory allows for diversity among vibrations that have different potential energies.

Guo and Thirumalai [10] have used IS theory to analyze fluctuations in the neighborhood of the native state of a coarse-grained model of a designed four-helix bundle protein. Baumketner, Shea, and Hiwatari [11] have applied IS theory to study the glass transition in a coarse-grained model of a 16-residue polypeptide; by IS analysis of molecular dynamics trajectories, they demonstrated the ability to rigorously calculate the glass transition temperature due to freezing in the native-state basin. In a more recent study, Nakagawa and Peyrard [12] used IS theory to analyze the energy landscape of the B1 segment of the IgG binding domain of streptococcal protein G (GB1) using a coarse-grained model, finding that the density of minima increases

exponentially with the energy. Importantly, their analysis relied on an assumption that vibrations are the same about all potential energy minima. However, because protein models become less rigid as noncovalent bonds are broken [13], vibrations are expected to be different for different minima, especially for minima with different potential energies. Diversity in vibrations not only would change the density of minima, but also would have important implications for the kinetics of transitions among conformational substates [1]; however, if vibrations are the same for different minima, their role in determining the kinetics of transitions would be trivial.

To characterize the diversity in vibrations among different protein inherent structures, we used IS theory to analyze the same protein fragment considered by Nakagawa and Peyrard [12], GB1 (Protein Data Bank [14] entry 2GB1 [15]). GB1 has 56 amino acids and consists of a four-stranded β sheet packed against a single helix (Fig. 1).

As in Ref. [12], we use a coarse-grained Gō model of GB1. Compared to all-atom models, the Gō model enabled ample sampling for IS analysis and enabled us to compare our results to those in Ref. [12]. Although trivial compared to all-atom molecular mechanics models, Gō models nevertheless can capture notrivial aspects of protein folding and assembly (see, e.g., Refs. [17–19]), emphasizing the important role of topology in determining mechanisms of these processes. In addition, we hoped that analysis of a Gō model would provide clues about the energy landscape of more detailed models.

In the GB1 model, a configuration x was represented by the set of C_α positions, and the potential energy $U(x)$ for C_α configurations x of all proteins was similar to that used for the GB1 studies in Refs. [20] and [12]:

$$U(x) = \sum_{i=1}^{N-1} \frac{K_b}{2}(r_i - r_{i,0})^2 + \sum_{i=1}^{N-2} \frac{K_\theta}{2}(\theta_i - \theta_{i,0})^2$$
$$+ \sum_{i=1}^{N-3} \frac{K_\phi}{2}\left[1 - \cos\left(2\phi_i - \frac{\pi}{2}\right)\right]$$
$$+ \sum_{i>j-3}^{NC} \epsilon\left[5\left(\frac{r_{ij,0}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{ij,0}}{r_{ij}}\right)^{10}\right] + \sum_{i>j-3}^{NNC} \epsilon\left(\frac{C}{r_{ij}}\right)^{12}.$$
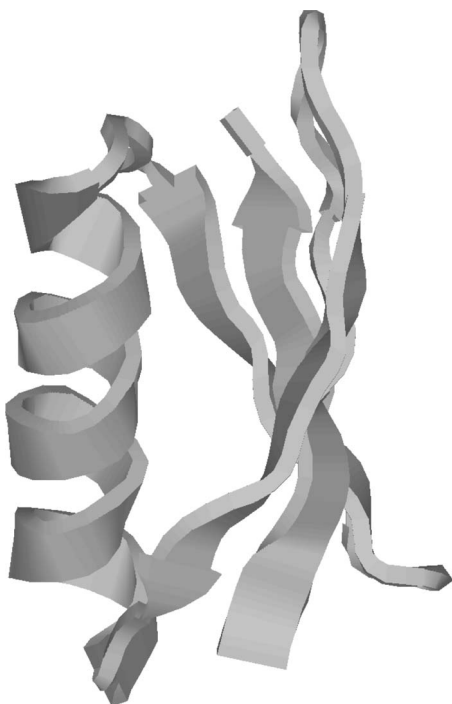
$$(1)$$

*Corresponding author: mewall@lanl.gov

FIG. 1. Schematic illustration of GB1 created using RASMOL [16].

The first term in Eq. (1) is the contribution from neighboring backbone C_α bond distances r_i, the second is from angles θ_i between neighboring bonds, the third is from dihedral angles ϕ_i, the fourth is from noncovalent interactions between atom pairs (i,j) that form a native contact, and the fifth is from noncovalent interactions between other pairs of atoms. The crystal structure was used as the reference structure to determine $r_{i,0}$, $\theta_{i,0}$, and $r_{ij,0}$, with native contacts determined using a cutoff distance of 5.5 Å. Other parameter values are $K_b=200\epsilon_0$ Å^{-2}, $K_\theta=40\epsilon_0$ rad^{-2}, $K_\phi=0.3\epsilon_0$, $\epsilon=0.18\epsilon_0$, and $C=4$ Å. The absolute energy unit $\epsilon_0=1.89$ kcal mol^{-1} was determined as in Ref. [20], assuming a folding temperature $T=350$ K. As in Ref. [12], instead of being defined with respect to the reference structure, the dihedral angle terms were defined to yield energy minima at 45° and −135° [20], adding some frustration in the native state.

Langevin dynamics simulations were performed using a time step 0.0007τ and a friction coefficient of $0.2/\tau$, where $\tau=1.47$ ps (following Ref. [20]). The collapse temperature T_θ, defined by the maximum of the specific heat C_v vs T [12], was located by performing multiple simulations at different temperatures to estimate C_v and locate the value of T at which it is maximal (T_θ is not guaranteed to be the same as the folding temperature as defined, e.g., using the maximum of the fluctuations in the number of native contacts versus T [21]). A trajectory at temperature T_θ with 3×10^8 time steps was sampled every 10^4 steps to obtain an ensemble of 3×10^4 inherent structures for analysis. (Results were similar for independent runs, indicating that this sampling scheme was adequate.) Local minima e_α, corresponding to inherent structures α, were found using conjugate gradient minimization terminated when a step resulted in an energy change less
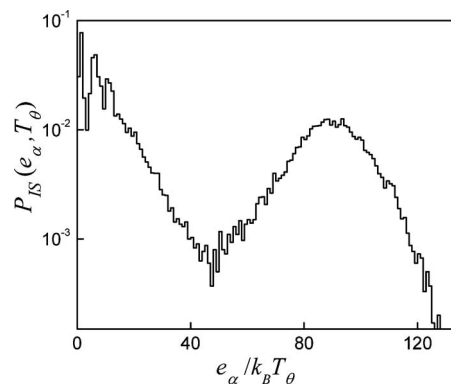

FIG. 2. Estimated probabilities of inherent structures with energy e_α at the collapse temperature T_θ.

than $10^{-12}\epsilon_0$. The protein exhibited multiple transitions between extended and collapsed states during the course of the simulation, and the inherent structure ensembles exhibited a bimodal probability distribution $P_{IS}(e_\alpha,T_\theta)$ of collapsed and extended inherent-structure potential energies e_α (Fig. 2), similar to the distribution in Ref. [12].

Like a previous application of IS theory to proteins by Baumketner, Shea, and Hiwatari [11], we replace the configurational integral in the partition function for an isolated protein with a sum over contributions from individual inherent structures:

$$Z=\left(\prod_{i=1}^{3N}\Lambda_i^{-1}\right)\frac{1}{\sigma}\int_V dx\ e^{-U(x)/k_BT}$$

$$=\sum_\alpha e^{-\beta e_\alpha}(\prod_{i=1}^{3N}\Lambda_i^{-1})\frac{1}{\sigma}\int_{R(\alpha)} dx\ e^{-[U(x)-e_\alpha]/k_BT}$$

$$=\sum_\alpha e^{-[e_\alpha+F_v(\alpha,T)]/k_BT}, \tag{2}$$

which defines the vibrational free energy $F_v(\alpha,T)$. In Eq. (2), $R(\alpha)$ is the basin surrounding inherent structure α, Λ_i is the thermal wavelength of atom i, and σ is a factor to account for symmetries.

Values of $F_v(\alpha,T)$, calculated as differences with respect to the native structure $\alpha=0$ (the same holds for values of e_α), were estimated at the collapse temperature T_θ using a harmonic approximation,

$$F_v(\alpha,T)=\frac{k_BT}{2}\sum_{i=7}^{3N}\ln\frac{\lambda_i^{(\alpha)}}{\lambda_i^{(0)}}, \tag{3}$$

where $\lambda_i^{(\alpha)}$ is the ith eigenvalue of the Hessian $h_{jk}=\partial^2U/\partial x_j\partial x_k$ calculated at the energy minimum corresponding to inherent structure α and $\lambda_i^{(0)}$ is the same for the ground-state inherent structure. The sum is over all modes with nonzero frequency: we neglect a contribution due to changes in the rotational entropy for different inherent structures. Values of F_v are similar for inherent structures with a similar potential energy e_α (Fig. 3). The contribution to F_v from the highest 1/3 of the eigenvalues does not change for
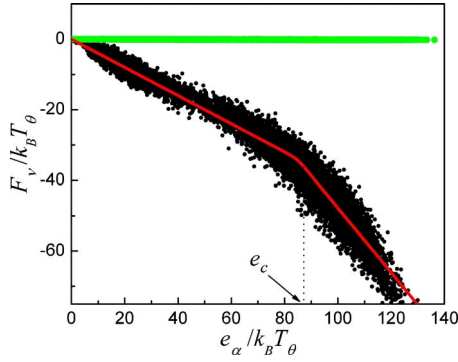
FIG. 3. (Color online) Vibrational free energies F_v of inherent structures vs their potential energies e_α (black points). The dependence is well modeled by Eq. (4) with $e_c=88.4k_BT_\theta$ (piecewise-linear red line). The contribution from the highest 1/3 of the eigenvalues is constant (green points following $y=0$).

different inherent structures. Interestingly, there is a gap in the eigenvalue spectrum between the lowest 2/3 and the highest 1/3 of the eigenvalues (Fig. 4); in addition, only the highest 1/3 of the eigenvalues change when the bond-distance force constant K_b is increased by a factor of 10, indicating that the corresponding modes describe the bond vibrations. Therefore bond vibrations do not change significantly among different inherent structures. However, the total F_v, which includes contributions from the lowest 2/3 of the eigenvalues, changes significantly with e_α (Fig. 3). The assumption of constant F_v by Nakagawa and Peyrard [12] therefore is only valid for modes that involve bond vibrations. This result is consistent with studies of the loss of protein rigidity when local constraints are relaxed [13]—e.g., in protein unfolding [7]—and is also consistent with molecular dynamics studies suggesting that vibrations can be diverse for different protein conformational substates [22,23].

As demonstrated by the fit in Fig. 3, F_v is well modeled using the function
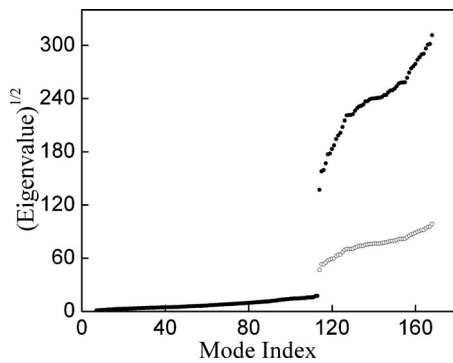

FIG. 4. Eigenvalue spectrum of an arbitrary inherent structure before (open circles) and after (solid circles) increasing the bond-distance spring constant K_b by a factor of 10. The low-frequency spectra are indistinguishable.

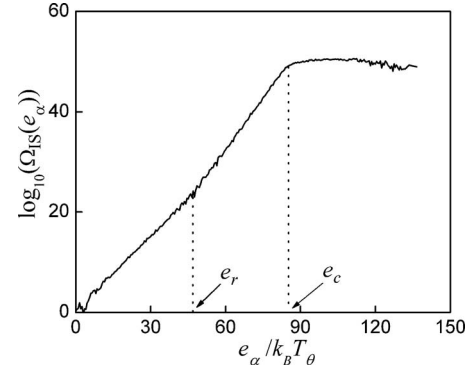FIG. 5. Density of inherent structures $\Omega_{IS}(e_\alpha)$. The knee at $e_r=47.4k_BT_\theta$ is due to a change in stress, and the plateau beginning at roughly $e_c=88.4k_BT_\theta$ is due to a change in rigidity; both are understood in terms of the network of native contacts (Figs. 6 and 7).

$$F_v(e_\alpha,T_\theta) = k_2e_\alpha + (k_2 - k_1)k_BT_\theta \ln(e^{-e_c/k_BT_\theta} + e^{-e_\alpha/k_BT_\theta}). \quad (4)$$

Equation (4) is essentially a piecewise-linear function with slope k_1 for $e_\alpha<e_c$ and slope k_2 for $e_\alpha>e_c$. For GB1, $k_1=-0.40$, $k_2=-0.91$, and $e_c=88.4k_BT_\theta$.

Inherent structure theory [8,9] assumes that $F_v(\alpha,T) = F_v(e_\alpha,T)$ (validated for the present application in Fig. 3) and relates $F_v(e_\alpha,T)$ and $P_{IS}(e_\alpha,T)$ to the density of states $\Omega_{IS}(e_\alpha)$ through

$$P_{IS}(e_\alpha,T) = \frac{1}{Z}\Omega_{IS}(e_\alpha)e^{-e_\alpha/k_BT}e^{-F_v(e_\alpha,T)/k_BT}. \quad (5)$$

Given $\Omega_{IS}(e_0)=1$, $F_v(e_0,T)=0$, and $e_0=0$, $\Omega_{IS}(e_\alpha)$ is given by

$$\Omega_{IS}(e_\alpha) = e^{e_\alpha/k_BT}e^{F_v(e_\alpha,T)/k_BT}\frac{P_{IS}(e_\alpha,T)}{P_{IS}(e_0,T)}, \quad (6)$$

which generalizes a similar equation in Nakagawa and Peyrard [12] to values of $F_v(e_\alpha,T)$ that vary with e_α. Because F_v in Eq. (3) is proportional to T, in a harmonic approximation of F_v, temperature changes are guaranteed only to influence P_{IS} through the Boltzmann factor e^{-e_α/k_BT} in Eq. (5). This behavior was observed in Ref. [12]; however, it was interpreted there as indicating that values of F_v are similar for different inherent structures.

We used Eq. (6) along with the calculated P_{IS} and F_v from Eq. (4) to model the density of inherent structures Ω_{IS}. At energies below e_c, Ω_{IS} exhibits an exponential increase, but with a slight increase in the exponent factor above an energy e_r, giving rise to a knee in the plot of $\log \Omega_{IS}$ vs e_α (Fig. 5). The knee is located at the minimum in P_{IS} between the extended and collapsed states (Fig. 2) and is thus associated with the transition state. Such a knee was also seen in a previous model of Ω_{IS} in which vibrations were assumed to be identical for different inherent structures [12]. Above e_c, Ω_{IS} plateaus and decreases at the highest energies, which is a
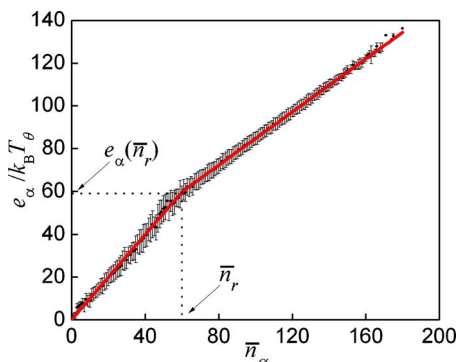
FIG. 6. (Color online) Potential energy e_α vs number of broken contacts, $\bar{n}_\alpha$. The dependence is well modeled by Eq. (8) (red line). The energy required to break a native contact is approximately equal to the binding energy ϵ, with differences due to stress in the structure. The knee corresponds to a change in stress at $\bar{n}_r=60$, where $e_\alpha(\bar{n}_r)=59.6k_BT_\theta$.

consequence of the structure in both P_{IS} and F_v. Rather than being exponential in form [12], from Eqs. (3), (4), and (6), Ω_{IS} in this region has the form

$$\Omega_{IS}(e_\alpha > e_c) = e^{e_\alpha/k_BT}e^{k_2 e_\alpha/k_BT_\theta}e^{(k_1-k_2)e_c/k_BT_\theta}\frac{P_{IS}(e_\alpha,T)}{P_{IS}(e_0,T)}.$$
(7)

Because k_2 is close to -1, the structure of $\Omega_{IS}(e_\alpha)$ closely resembles that of $P_{IS}(e_\alpha,T_\theta)$ above e_c.

We found (Fig. 6) that e_α is closely related to the number of broken native contacts, $\bar{n}_\alpha$, through the piecewise-linear function

$$e_\alpha = h_2\bar{n}_\alpha + (h_2-h_1)\ln(e^{-\bar{n}_r}+e^{-\bar{n}_\alpha}).$$
(8)

The slopes h_1 and h_2 correspond to the amount of energy required to break a native contact below (h_1) and above (h_2) a critical number of broken contacts, $\bar{n}_r$. Data for GB1 are well modeled using $h_1=0.997$, $h_2=0.622$, and $\bar{n}_r=60$. Below $\bar{n}_r$, breaking a native contact requires more potential energy than above $\bar{n}_r$. Therefore, $\bar{n}_r$ is associated with a change in protein stress.

There are interesting connections between the structure of Ω_{IS} below e_c (Fig. 5) and the dependence of e_α on $\bar{n}_\alpha$ (Fig. 6). The change in the slope of Ω_{IS} at e_r is closely related to the change in the slope of $e_\alpha(\bar{n}_\alpha)$ at $\bar{n}_r$, suggesting that Ω_{IS} has a simple exponential dependence on $\bar{n}_r$ below e_c. However, the knee in Ω_{IS} occurs at $e_r=47.4k_BT_\theta$, which is smaller than the value $e_\alpha(\bar{n}_r)=59.6k_BT_\theta$ at the knee in Fig. 6. While the density of inherent structures might truly be enhanced in the gap between e_r and $e_\alpha(\bar{n}_r)$, we note that the shift of e_r with respect to $e_\alpha(\bar{n}_r)$ might indicate that the inherent structure basins associated with the transition state are especially large (as noted above, e_r is associated with the transition state) and that the harmonic approximation might be especially ill suited to estimating their free energies for use in Eq. (6).

The source of the plateau in Ω_{IS} above e_c may be understood in terms of the dependence of the free energy on both
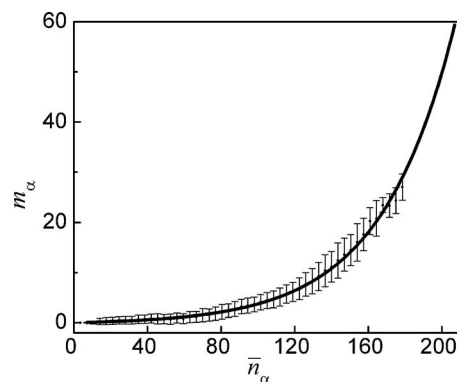

FIG. 7. Number of residues, m_α, for which all native contacts are broken vs the number of broken contacts, $\bar{n}_\alpha$. The dependence is well modeled by Eq. (9) (black line). Note that $m_\alpha=55$ (close to the expected value of 56) at $\bar{n}_\alpha=207$.

$\bar{n}_\alpha$ and the number of residues, m_α, for which all native contacts are broken. As shown in Fig. 7, a plot of m_α vs $\bar{n}_\alpha$ is well modeled by the function

$$m_\alpha = \frac{1}{3}(e^{\bar{n}_\alpha/40}-1),$$
(9)

supporting an expectation that breaking a contact is only likely to create a residue with no native contacts at high $\bar{n}_\alpha$. The following simple model for F_v then successfully captures the structure of F_v in Fig. 3:

$$F_v(\bar{n}_\alpha) = \nu\bar{n}_\alpha + \mu m_\alpha,$$
(10)

with m_α given by Eq. (10). Using $\nu=-0.32$ and $\mu=-1.07$ yields good agreement between values of F_v obtained either directly from the Hessian or using Eq. (10) (Fig. 8), with a correlation coefficient of 0.993 for values calculated from all inherent structures. We conclude that the change in the slope of F_v vs e_α at e_c, and therefore the plateau in $\Omega_{IS}(e_\alpha)$ above e_c, is associated with an increase in the likelihood that breaking a native contact will increase the number of residues with no native contacts.
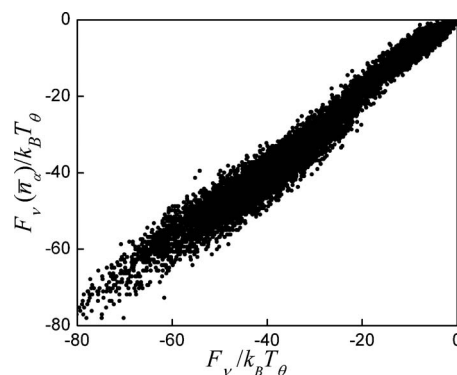

FIG. 8. Comparison of vibrational free energies calculated from Eq. (10) (y axis) and estimated using inherent structure theory (x axis).

We found that protein stress and rigidity are closely tied to the network of native contacts through Eqs. (8) and (10). This finding is reminiscent of analyses of protein rigidity using concepts from graph rigidity [13] and the association of a loss of network rigidity with protein unfolding [7]. It is therefore tempting to associate the region between e_r and e_c in Fig. 5 with the region of the mean coordination number $\langle r \rangle$ where proteins were found in Ref. [7] to become floppy and unfold. However, the present approach differs from that used in Ref. [7] in two key ways. First, because all residue interactions in the present study are lumped into C_α atoms, the coordination numbers are higher and the relation of coordination numbers to protein rigidity might be different than for the all-atom models considered in Ref. [7]. Second, whereas the present results were obtained using a dynamical model, those obtained in Ref. [7] were obtained using a static model of the protein. It will be interesting to further explore connections between the analyses based on IS theory and network rigidity.

The maximum in the density of states above e_c is a consequence of considering diversity in vibrations and is not observed when uniform vibrations are assumed [12]. Interestingly, a similar structure for the density of states, in which an exponential increase is followed by a maximum, has been observed for many structural-glass-forming liquids [9]. To further explore this correspondence, it will be interesting to improve the estimation of the density of states by obtaining more accurate estimates of F_v than are possible using a harmonic approximation [11].

Studies of two other Gō models of proteins yielded results that are qualitatively similar to those found here for GB1 [24], suggesting the possibility that a simple phenomenological relationship between the network of native contacts and the energy landscape might exist for all Gō models. It will be interesting to explore this relationship for a large number of proteins and seek representations in which it is identical for different proteins. Discovery of such "universality" would enable the prediction of important properties of the energy landscapes of Gō models without performing numerical simulations.

It will be important to extend the present results to models whose energy landscapes exhibit more frustration than the model of GB1 considered here. For example, increasing K_ϕ by a factor of 10 changes the detailed location, depth, and vibrations about energy minima; however, importantly, the native contacts and vibrations of inherent structures remain consistent with Eqs. (9) and (10) [24]. The validity of these equations is therefore robust to changes in frustration mediated by the dihedral angle geometry term in Eq. (1). In addition, consider a modified model in which there is a weak attractive interaction for non-native contacts. In contrast to the simple relation illustrated in Fig. 6, in such a model, inherent structures with the same potential energy would likely have diverse numbers of native contacts. However, by extending the parameter space, the energy still might be simply related to a combination of both the number of native contacts and the number of non-native contacts. Similarly, the vibrational free energies might exhibit diversity among inherent structures with the same energy, but might still be simply related to both the number of native contacts and non-native contacts through an equation analogous to Eq. (10).

Ultimately, it will be interesting to incrementally increase the complexity of the model, extending the present results (as far as computationally feasible) to realistic, all-atom models of proteins that include explicit solvent and other effects that are important in controlling protein function. Comparisons with all-atom models will create an avenue for understanding more precisely the limitations of Gō models, which are gaining in popularity. Use of all-atom models will also facilitate the exploration of links between analyses based on IS theory and network rigidity.

The present results demonstrate that simple connections to protein structure are hidden within the energy landscape of a Gō model. The potential energies and vibrational free energies of inherent structures are highly correlated, and both reflect simple measures of networks of native contacts. Through use of IS theory, these regularities should enable significant simplification of thermodynamic models of proteins [8,9,12]. It will be important to determine the relevance of the present results to more realistic models; both the success of Gō models in studying mechanisms of protein folding and the fact that diversity in vibrations about different minima has been observed in all-atom molecular mechanics models of proteins [22,23] suggest that the present results might be meaningfully generalized.

[1] H. Frauenfelder and P. G. Wolynes, Science **229**, 337 (1985).

[2] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, Science **254**, 1598 (1991).

[3] A. Ansari, J. Berendzen, S. F. Bowne, H. Frauenfelder, I. E. Iben, T. B. Sauke, E. Shyamsunder, and R. D. Young, Proc. Natl. Acad. Sci. U.S.A. **82**, 5000 (1985).

[4] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).

[5] D. L. Stein, Proc. Natl. Acad. Sci. U.S.A. **82**, 3670 (1985).

[6] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).

[7] A. J. Rader, B. M. Hespenheide, L. A. Kuhn, and M. F. Thorpe, Proc. Natl. Acad. Sci. U.S.A. **99**, 3540 (2002).

[8] F. H. Stillinger and T. A. Weber, Phys. Rev. A **25**, 978 (1982).

[9] F. H. Stillinger and T. A. Weber, Science **225**, 983 (1984).

[10] Z. Guo and D. Thirumalai, J. Mol. Biol. **263**, 323 (1996).

[11] A. Baumketner, J.-E. Shea, and Y. Hiwatari, Phys. Rev. E **67**, 011912 (2003).

[12] N. Nakagawa and M. Peyrard, Proc. Natl. Acad. Sci. U.S.A. **103**, 5279 (2006).

[13] D. J. Jacobs, L. A. Kuhn, and M. F. Thorpe, *Rigidity Theory and Applications* (Kluwer, New York, 1999), pp. 357–384.

[14] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).

[15] A. M. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore, Science **253**, 657 (1991).

[16] R. A. Sayle and E. J. Milner-White, Trends Biochem. Sci. **20**, 374 (1995).

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

[18] S. Yang, S. S. Cho, Y. Levy, M. S. Cheung, H. Levine, P. G. Wolynes, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **101**, 13786 (2004).

[19] M. S. Li, C. K. Hu, D. K. Klimov, and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **103**, 93 (2006).

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

[21] A. Baumketner and Y. Hiwatari, Phys. Rev. E **66**, 011905 (2002).

[22] D. Janezic, R. M. Venable, and B. R. Brooks, J. Comput. Chem. **16**, 1554 (1995).

[23] H. W. T. van Vlijmen and M. Karplus, J. Phys. Chem. B **103**, 3009 (1999).

[24] D. Ming, M. Anghel, and M. E. Wall (unpublished).