Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 809–813, February 1993
Biophysics

Exploring the energy landscape in proteins

JOHN E. STRAUB* AND D. THIRUMALAI†

*Department of Chemistry, Boston University, Boston, MA 02215; and †Department of Chemistry and Biochemistry, Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742

Communicated by Peter G. Wolynes, October 15, 1992 (received for review July 10, 1992)

ABSTRACT **We present two methods to probe the energy landscape and motions of proteins in the context of molecular dynamics simulations of the helix-forming S-peptide of RNase A and the RNase A–3′-UMP enzyme–product complex. The first method uses the generalized ergodic measure to compute the rate of conformational space sampling. Using the dynamics of nonbonded forces as a means of probing the time scale for ergodicity to be obtained, we argue that even in a relatively short time (<10 psec) several different conformational substates are sampled. At longer times, barriers on the order of a few kcal/mol (1 cal = 4.184 J) are involved in the large-scale motion of proteins. We also present an approximate method for evaluating the distribution of barrier heights $g(E_B)$ using the instantaneous normal-mode spectra of a protein. For the S-peptide, we show that $g(E_B)$ is adequately represented by a Poisson distribution. By comparing with previous work on other systems, we suggest that the statistical characteristics of the energy landscape may be a "universal" feature of all proteins.**

The dynamics of the internal motions in proteins spanning several decades in time is thought to be a direct consequence of the "complexity" of the underlying energy landscape (1). The wide distribution of time scales for protein motion and its consequences has been described in the experiments of Frauenfelder and coworkers on ligand binding in heme proteins (2). Based on the relaxation characteristics over a wide temperature range, they have argued that a single monomeric protein can possess many nearly isoenergetic conformations, which are referred to as conformational substates (1, 2). The individual substates are believed to be separated by barriers of differing heights. There appears to be a wide distribution of barrier heights ranging from a few hundredths to many kcal/mol (1 cal = 4.184 J). The existence of the multivalley structure of the energy landscape together with the notion of a distribution of barriers has been used to analyze quantitatively the kinetics of binding of small ligands to myoglobin over a wide temperature range (2, 3). Frauenfelder and coworkers (1, 2) have argued that it is necessary to postulate the existence of conformational substates (CS) in which the CS have structure on several energy and length scales. This idea explains naturally the presence of an array of internal motions in proteins as well as fluctuations on several length scales (1). The existence of many time scales implies that, in general, relaxation functions should exhibit nonexponential kinetics. In addition, the rate constant for rebinding apparently exhibits marked deviation from the Arrhenius law; non-Arrhenius temperature dependence is obtained only when the rebinding constant is examined over a range of temperatures.

Molecular dynamics simulations have also provided some evidence for the existence of conformational substates. Following the analysis of a 300-psec trajectory generated by solving Newton's equation for myoglobin, Elber and Karplus (4, 5) have suggested that there are many thermally accessible minima in the neighborhood of the native structure. Structural differences between distinct CS corresponded to relative orientations of the helices, which seem to be initiated by side-chain rearrangements. The measure used by Elber and Karplus is approximate and simple arguments can be used to show that these authors have overestimated the number of distinct minima explored by myoglobin in 300 psec at room temperature. More recently, a much more ambitious study of the topography of the energy landscape in flexible systems has been completed by Czerminski and Elber (6, 7). They were able to map out the distinct minima in a tetrapeptide and obtain the location of the transition states separating the minima in the small peptide molecule. The protocol used by Czerminski and Elber yields a set of "optimized" reaction pathways separating the minima. Perhaps the most important aspect of their study is that the general features (see below) of the distribution of barrier heights in this small molecule seems to be quite similar to that postulated for a much larger protein molecule—namely, myoglobin (8). Finally, Czerminski and Elber showed by using an approximate kinetic model for conformational transitions that as a consequence of the distribution of saddle points there are several relaxation times. The study by Czerminski and Elber, while interesting, is computationally intensive and a similar characterization of the energy landscape in larger polypeptides will be far more difficult.

Although the studies cited above have gone a long way in establishing the existence and importance of CS, relatively little has been done to elucidate the time scales involved in exploration of the underlying energy landscape. The ultimate goal, of course, is to reveal the relationship between protein function and the associated energy landscape (9). Our goals in this paper are the following. (*i*) To devise simple methods that can be readily used to assess the time scales on which a protein samples the available conformational space. We use recently developed techniques for quantitatively measuring the approximate rate at which proteins search the energy landscape as a function of temperature. In the process, we gain insight into the rate of sampling for the conformational degrees of freedom. (*ii*) To compute the distribution of barrier heights in the energy landscape of proteins. We show that the distribution of saddle point energies is approximately described by the Poisson law for energies greater than a certain cut-off value. We wish to stress that our objective is to devise methods that can be used in conjunction with standard molecular dynamics simulations to obtain qualitative information about the dynamics of the energy surface that is explored at a specified temperature. The complete characterization of the energy landscape along the lines attempted by Czerminski and Elber, while important, may not be as relevant for finite temperature dynamics.

METHODS

The analysis presented in this paper is based on finite duration molecular dynamics (MD) simulations in vacuum. We carried out calculations on the S peptide of bovine pancreatic RNase

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: CS, conformational substate(s); MD, molecular dynamics.

A and the RNase A–3′-UMP enzyme–inhibitor complex. The protein parameters are based on the version 19 CHARMM parameter set. The inhibitor and histidine parameters were developed separately and will be described elsewhere (J.E.S., C. Lim, and M. Karplus, unpublished data). The dynamics were performed at constant energy using a modified version 20 of the CHARMM simulation program (10).

As emphasized in the Introduction, the generic feature of the complex energy landscape that dictates protein dynamics is the multivalley structure with a rather wide spectrum of barrier heights. A question of interest is to obtain the time scales for the transition between different conformational substates that are being sampled in a given observation time. Let us assume that there are two distinct CS labeled α and β separated by an effective free energy of activation $\Delta F^{\ddagger}_{\alpha\beta}$. Conformations belonging to α and β would mix only if in the process of time evolution the molecule makes a transition from α to β. Such an activated transition requires times on the order of $\tau_0 \exp(\Delta F^{\ddagger}_{\alpha\beta}/k_B T)$, where τ_0 is typically on the order of a vibrational period. Therefore, to assess whether there are distinct minima separated by barriers, it is necessary to follow the time development of two trajectories starting from independent initial conditions. Let

$$f^{\alpha}_{i\mathrm{N}}(t) = \frac{1}{t}\int_0^t ds F^{\alpha}_{i\mathrm{N}}(s) \qquad [1]$$

be the time average of the force due to the nonbonded potential experienced by the ith atom. Superscript α indicates that the starting conformation maps onto CS α. Similarly, let $f^{\beta}_{i\mathrm{N}}(t)$ be the time averaged nonbonded force whose initial configuration maps onto CS β. Since the initial conformations have to be independent, sufficient care must be used to generate them in any numerical computation. The force metric $d_{\mathrm{FN}}(t)$ is defined as

$$d_{\mathrm{FN}}(t) = \frac{1}{N}\sum_{i=1}^{N} \|f^{\alpha}_{i\mathrm{N}}(t) - f^{\beta}_{i\mathrm{N}}(t)\|^2. \qquad [2]$$

The properties of $d_{\mathrm{FN}}(t)$ are easy to discuss. If the initial conformations mix on the time scale of the simulations, then $d_{\mathrm{FN}}(t)$ should vanish for long times. On the other hand, if the initial conformations belong to distinct CS then $d_{\mathrm{FN}}(t) \to 0$ only for times longer than $\tau_0 \exp(\Delta F^{\ddagger}_{\alpha\beta}/k_B T)$, which can be very long even for moderate values of $\Delta F^{\ddagger}_{\alpha\beta}/k_B T$. If $d_{\mathrm{FN}}(t)$ for long times does not vanish, then it follows that the two trajectories explore distinct conformational substates. The most important aspect of the force metric is that in cases in which there is a bottleneck between the initial conformations α and β that the protein overcomes in the process of simulation, the time scale for such a transition can be obtained from the scaling relation (11, 12)

$$d_{\mathrm{FN}}(t)/d_{\mathrm{FN}}(0) \approx 1/D_{\mathrm{FN}}(T)t. \qquad [3]$$

The parameter D_{FN} (analogous to the maximal Lyapunov exponent) gives the temperature-dependent rate at which the conformational space is being explored. Estimates of barrier heights separating two conformational substates may be obtained from Eq. 3. Since D_{FN} gives roughly the rate for sampling conformations belonging to a distinct CS, it follows that $\Delta F^{\ddagger}_{\alpha\beta} \approx k_B T \ln(D_{\mathrm{FN}}\tau_0)$. According to the CS picture, there is a wide range of barriers in proteins and, consequently, $\Delta F^{\ddagger}_{\alpha\beta}$ obtained from this relation should be viewed as the average free energy barrier between the two CS α and β. This interpretation assumes that the distinct CS are dilutely connected. It should be noted that other quantities, such as the energy of the ith residue, local density, etc., may also be used to monitor the rate of sampling of conformational space.

The temperature-dependent parameter D_{FN} gives an estimate of the average barrier height separating the various CS that are sampled during a MD simulation. It is, however, desirable to obtain a direct expression for the distribution of barrier heights connecting the various CS. The distribution of barrier heights can in principle be obtained by using the force metric. To do this, one has to compute $D_{\mathrm{FN}}(T)$ for a large number of pairs of initial conformations. This is not only cumbersome but it also does not offer any physical insight. Here we use an approximate method for computing $g(E_B)$ from the analysis of the temperature dependence of the fraction of unstable instantaneous normal modes. Typically, normal mode calculations are carried out at zero temperature by expanding the potential function about a mechanically stable energy minimum. Originally in the study of glasses Rahman *et al.* (13) and more recently in the context of melting (14) and the liquid to glass transition (15–17), a number of studies have suggested that the fluidity of the system may in fact be associated with the fraction of unstable modes. The unstable modes, characterized by imaginary frequencies, are associated with motion over regions of the potential surface with negative curvature—namely, the saddle points. Here we show that by using a few simple approximations $g(E_B)$ may be directly estimated from a molecular dynamics trajectory. Our method is easily used for any protein for which the potential function is known. A more accurate determination of $g(E_B)$ can be made by extending the time scale of the simulation. However, we will show that the robustness of the CS picture becomes apparent even within our limited simulation time scales.

We use the following method to obtain $g(E_B)$. If the potential energy function is known, then the equilibrium fraction of unstable modes can be computed as a function of temperature. For example, we can divide the configuration integral at a fixed temperature into regions of the potential corresponding to stable and unstable motion. This decomposition allows us to identify the fraction of imaginary frequencies as

$$f_u(T) = Z_{\mathrm{unstable}}(T)/Z(T), \qquad [4]$$

where $Z_{\mathrm{unstable}}(T)$ is the configuration integral for the unstable motion and $Z(T)$ is the total configuration integral. In general, the required configuration integrals for proteins cannot be performed analytically and one requires numerical methods for evaluating $f_u(T)$. Nevertheless, a simple normal-mode model can be used to obtain $g(E_B)$ from the calculation of $f_u(T)$. We make two assumptions. (*i*) The nonlinear modes of the protein can be adequately described by the corresponding $3N$ one-dimensional normal modes. This appears to be reasonable at least near one of the local minima of the conformational substates as well as near the saddle point. (*ii*) We approximate the true potential separating two CS by a piecewise local harmonic potential. The local harmonic approximation to the free energy surface is likely to be valid (18), even though the dynamics of the protein as a whole is highly anharmonic. These two approximations, which enable us to reduce the complex landscape into a series of local harmonic wells separated by various barrier heights, can be rationalized if the fluctuations of the protein about an instantaneous position are due to barrier crossing and not simply stable anharmonic motion. Within this approximate characterization of the actual free energy landscape, the equilibrium fraction of unstable modes $\bar{f}_u(T, E_B)$, for a particular value of the barrier height E_B, can be readily calculated. The result (unpublished data) is

$$\bar{f}_u(T, E_B) = \frac{e^{-2\xi^2}\int_0^{\xi} dx e^{x^2}}{\int_0^{\xi} dt e^{-t^2} + e^{-2\xi^2}\int_0^{\xi} dx e^{x^2}}, \qquad [5]$$

where $\xi^2 = \beta E_B/2$. If we assume a continuous distribution of barrier heights, the fraction of unstable modes measured at a particular temperature can be written as

$$f_u(T) = \int_0^\infty dE_B\, g(E_B)\tilde{f}_u(T, E_B). \quad [6]$$

By calculating $f_u(T)$ from a molecular dynamics trajectory (or a Monte Carlo simulation) at a series of temperatures, one can in principle extract an approximate form for the distribution of barrier heights $g(E_B)$ for the system. The above equation, which is a Fredholm integral equation of the first kind, is a general way of obtaining $g(E_B)$. The analytic representation of the kernel $\tilde{f}_u(T, E_B)$ that we have obtained using an idealized caricature of the energy landscape is perhaps the simplest one imaginable. The knowledge of $g(E_B)$ together with $D_{FN}(T)$ gives us a way of quantifying the dynamics of exploration of the complex energy landscape in proteins. The above computational methods were used to analyze molecular dynamics trajectories at several temperatures for the S-peptide of RNase A and the RNase A–3′-UMP enzyme–product complex.

RESULTS AND DISCUSSION

We have calculated the nonbonded force metric at several temperatures for the enzyme–product complex by using a number of molecular dynamics trajectories. The reason for computing $d_{FN}(t)$ is the following: proteins adopt well-defined three-dimensional structures because residues far apart in the protein (nonbonded residues) would prefer to be close in the configuration space. Thus, the initiation of folding is determined by the time scale required for establishing the nonbonded contacts (19). The relaxation of nonbonded forces, probed by $d_{FN}(t)$ for times on the order of 100 psec, can yield useful insights into the dynamics of early events in the establishment of these contacts. Fig. 1 shows a plot of $d_{FN}(t)$ as a function of time for the enzyme–product complex at several temperatures. Initial configurations for the α and β states were taken as endpoints of a 50-psec trajectory at a higher temperature. This figure shows that there is a rapid initial convergence followed by a slow, long time decay. We have analyzed the initial decay that occurs at times less than ≈10 psec. This analysis shows that there is a spectrum of relaxation times as indicated by the presence of several distinct slopes in the plot of the reciprocal of $d_{FN}(t)$. This shows that even in the 10-psec regime the protein samples several different CS that are separated by small barriers. These CS would correspond to the substructure in the tier organization of substates proposed by Frauenfelder *et al.* (2). The short time exploration of substates is due to the peptide exploring an essentially local harmonic potential minimum and the relatively constant interactions along the main chain, which are nearly independent of the peptide conformation.

For times greater than ≈15 psec (see Fig. 1), we notice a rather slow decay in $d_{FN}(t)$. Because $d_{FN}(t)$ probes the dynamics of two independent trajectories that map onto distinct CS, this implies that there is a bottleneck that is overcome only on time scales greater than those explored by our simulations. This is suggestive of the existence of an organization of CS into tiers of minima separated by barriers of increasing height. The exploration of various substates depends on the temperature, and this is illustrated by analyzing $d_{FN}(t)$ in terms of the properties of the individual trajectories. The decomposition of the force metric into fluctuation metric and the cross terms provides an interesting contrast between the low- and high-temperature dynamics. We find that at 40 K the independent trajectories are frozen

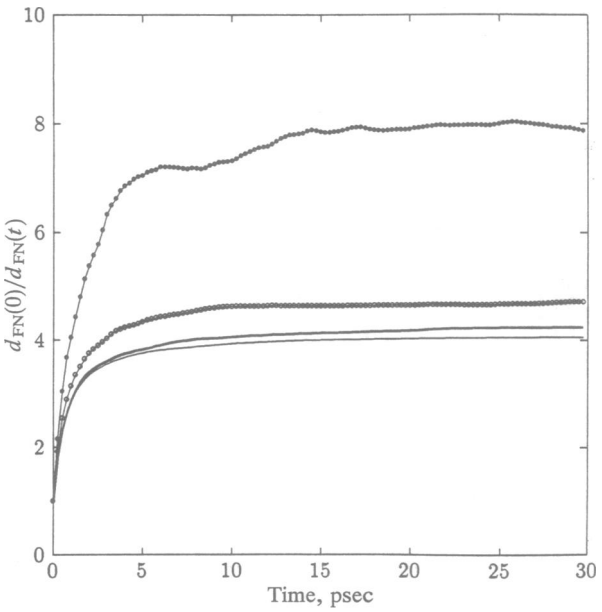

FIG. 1. Reciprocal of the normalized nonbonded force metric, $d_{FN}(0)/d_{FN}(t)$, as a function of time for the RNase A–3′-UMP enzyme–product complex. Thin line, data at 40 K; thick line, data at 120 K; open circles, data at 240 K; solid circles, data at 300 K.

into a particular CS and fluctuate about the average structure without making any significant dihedral angle transitions. This implies that the initial conditions for the two trajectories reside in separate CS and transitions between these and other substates occur on a much longer time scale than explored by our simulations. On the other hand, at 300 K we find that the two trajectories mix and the barriers between the CS can indeed be overcome on time scales on the order of 75 psec. As a quantitative estimate of the average barriers between the CS that are sampled, we have computed the temperature dependence of $D_{FN}(T)$ using the scaling relation given in Eq. 3. The average effective barrier height can be estimated from the relation $\tau_0 \exp(\Delta F^{\ddagger}_{\alpha\beta}/k_B T) \approx D_{FN}^{-1}$. If τ_0 is taken to be ≈1 psec we estimate $\Delta F^{\ddagger}_{\alpha\beta}$ to be 2.2–3.5 kcal/mol. These values fall in the range of activation energies computed by Czerminski and Elber for model peptides (7).

The results for the force metric can also be used to show that the long time relaxation in the protein is strongly tied to infrequent events such as dihedral angle relaxation (unpublished data). To further explore the nature of the relaxation process, we have computed the distribution of barrier heights $g(E_B)$ for the S-peptide by calculating the instantaneous normal-mode spectrum for eight temperatures ranging from 40 K to 500 K. At low temperatures, the eigenfrequencies are real and the density of states agrees with the 0 K normal-mode spectrum. As the temperature is increased, the number of imaginary eigenfrequencies (corresponding to unstable modes) increases and the center of the imaginary lobe of the spectrum shifts to higher values. Fig. 2 shows the fraction of unstable modes $f_u(T)$ computed as a function of temperature. The fraction of unstable modes is zero at 0 K and quickly rises through 100 K, at which there is a break leading to a weaker slope through higher temperatures.

The behavior of $f_u(T)$ can be interpreted in terms of the model leading to Eqs. 4–6. A large number of modes in the peptide will remain stable even at the highest temperature. Other modes, such as those localized in dihedral angle space and the low frequency, continuum-like global fluctuations, are imagined to move over a roughly periodic potential with barriers distributed according to $g(E_B)$. Physically, we know that the quick rise in $f_u(T)$ even at low temperatures indicates

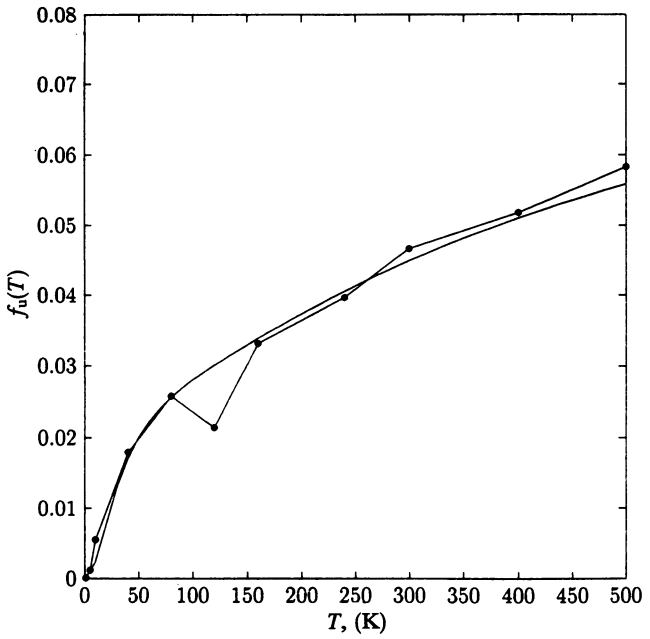

FIG. 2. Fraction of unstable modes as a function of temperature calculated from a 75-psec constant energy trajectory of the S peptide. Shown for comparison is the fit to $f_u(T)$ using Eq. 6 (solid circles).

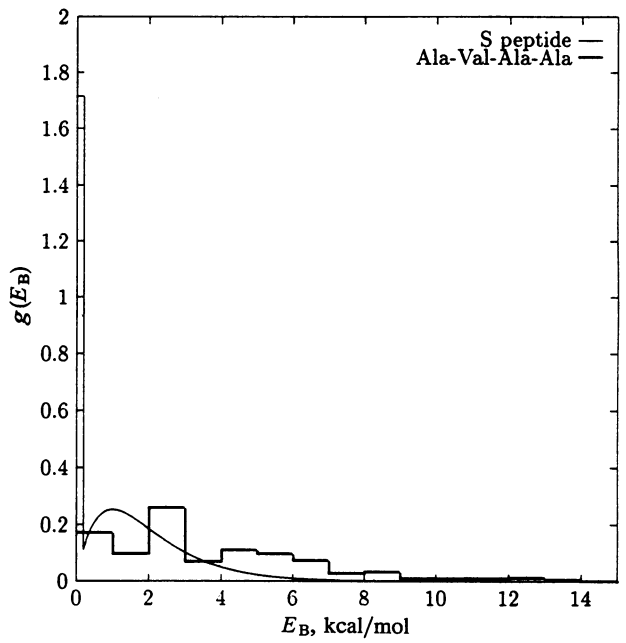

FIG. 3. Distribution of barrier heights $g(E_B)$ extracted from a fit of the temperature dependence of $f_u(T)$. For comparison, we also show the distribution of adiabatic barrier heights calculated directly for the tetrapeptide Ala-Val-Ala-Ala by Choi and Elber (20).

significant anharmonicity in the lowest frequency normal modes and the presence of many low barriers where $E_B \approx 100$ K. The slower rise at higher temperatures indicates the presence of many high barriers and the rate of increase in $f_u(T)$ at these temperatures will depend on the nature of $g(E_B)$. The physical insight provided by the changes in $f_u(T)$ with temperature leads us to postulate the following equation for $g(E_B)$:

$$g(E_B) = a\Theta(E_B - E_{low}) + bE_B e^{-E_B/E_0}. \quad [7]$$

There is a constant density of low energy barriers for $E_B < E_{low}$ [written in terms of the Heavyside function $\Theta(E)$] and Poisson distribution of higher energy barriers. The parameters for our model are $a = 0.5$ and $E_{low} = 0.2$ kcal/mol, and $b = 0.2$ and $E_0 = 1$ kcal/mol, respectively, which we derived from fitting $f_u(T)$ by Eqs. 5, 6, and 7 for a periodic piecewise parabolic potential. Our equation for $g(E_B)$ with the above choice of parameters solves the integral Eq. 5 adequately over the entire temperature range. This is shown in Fig. 2, where a comparison between $f_u(T)$ computed by MD simulations and that computed by evaluating the right-hand side of Eq. 5 is given.

A plot of $g(E_B)$ for the S-peptide along with a distribution of energy barriers for the tetrapeptide Ala-Val-Ala-Ala computed by Choi and Elber (20) is shown in Fig. 3. The tetrapeptide data are plotted only with a resolution of 1 kcal/mol and show a large number of low energy barriers, $E_B < 1$ kcal/mol, in agreement with our result. The two curves for larger values of E_B show qualitatively similar behavior, although our results indicate a significantly smaller number of high energy barriers. In all likelihood, this is because we have computed $f_u(T)$ only for limited values of T, and inversion of $f_u(T)$ to obtain $g(E_B)$, for larger E_B would require data at significantly higher temperatures. Nevertheless, given the differences in the two systems, it appears that the Poisson distribution of barrier heights is a robust feature of the energy landscape in proteins. This is further corroborated by noting that similar distributions have been found experimentally in heme proteins (8). The major consequences of the dynamics of proteins (i.e., nonexponential relaxation) readily follow from this distribution of activation energies.

The calculations presented here show the importance of the CS picture in the dynamics of proteins. The major observation is that even on relatively short time scales there are several CS that are explored, which implies the existence of small barrier heights (a few hundredths kcal/mol). The behavior of $d_{FN}(t)$ at the longest time examined here shows that barriers on the order of 2–3 kcal/mol are encountered. These results together with the normal-mode analysis for determination of $g(E_B)$ show that the statistical characteristics of the energy landscape responsible for the complex dynamics in proteins are manifested in the presence of a wide distribution of activation barriers. The results presented here, when combined with similar previous theoretical findings on model peptides (6, 7) and experimental studies on heme proteins (8), can be used to assert that the Poisson distribution for $g(E_B)$, which is one of the characteristics of a complex energy landscape, may be a "universal" aspect of all proteins. The range of barrier heights may in fact be necessary for the flexibility and enzymatic function of proteins and nucleic acids.

The authors are grateful to Peter G. Wolynes and the anonymous referees for useful comments. We want to thank Ron Elber for providing unpublished results on the tetrapeptide. This work was supported in part by a grant from the National Science Foundation and by the Camille and Henry Dreyfuss Foundation.

1. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. (1991) *Science* **254**, 1598–1603.
2. Frauenfelder, H., Parak, F. & Young, R. D. (1988) *Annu. Rev. Biophys. Chem.* **17**, 451–479.
3. Berendzen, J. & Braunstein, D. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 1–5.
4. Elber, R. & Karplus, M. (1987) *Science* **235**, 318–321.
5. Noguti, T. & Go, N. (1989) *Proteins Struct. Funct. Genet.* **5**, 97–103.
6. Czerminski, R. & Elber, R. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 6963–6967.
7. Czerminski, R. & Elber, R. (1990) *J. Chem. Phys.* **92**, 5580–5601.
8. Austin, R. H., Beeson, K. W., Eisenstein, L., Frauenfelder, H. & Gunsalus, J. C. (1975) *Biochemistry* **14**, 5355–5373.
9. Steinbach, P. J., Ansari, A., Berendzen, J., Braunstein, D., Chu, K., Cowen, B. R., Ehrenstein, D., Frauenfelder, H.,

Johnson, J. B., Lamb, D. C., Luck, S., Mourant, J. R., Nienhaus, G. U., Ormos, P., Phillips, R., Xie, Z. & Young, R. D. (1991) *Biochemistry* **30**, 3988–4001.
10. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983) *J. Comp. Chem.* **4**, 187–217.
11. Thirumalai, D., Mountain, R. D. & Kirkpatrick, T. R. (1989) *Phys. Rev. A* **39**, 3563–3574.
12. Thirumalai, D. & Mountain, R. D. (1990) *Phys. Rev. A Gen. Phys.* **42**, 4574–4587.
13. Rahman, A., Mandell, M. & McTague, J. P. (1976) *J. Chem. Phys.* **64**, 1564–1568.
14. La Violette, R. A. & Stillinger, F. H. (1985) *J. Chem. Phys.* **83**, 4079–4085.
15. Rosenberg, R. O., Thirumalai, D. & Mountain, R. D. (1989) *J. Phys. Cond. Matter* **1**, 2109–2114.
16. Seeley, G. & Keyes, T. (1989) *J. Chem. Phys.* **91**, 5581–5586.
17. Xu, B. C. & Stratt, R. M. (1990) *J. Chem. Phys.* **92**, 1923–1935.
18. Hall, R. W. & Wolynes, P. G. (1987) *J. Chem. Phys.* **86**, 2943–2948.
19. Guo, Z., Thirumalai, D. & Honeycutt, J. D. (1992) *J. Chem. Phys.* **97**, 525–535.
20. Choi, C. (1992) Ph.D. thesis (Univ. of Illinois, Chicago).