Charge-shift bonding is essentially a new bonding paradigm, albeit with traditional roots, and as such it will be tested by the range of its applicability in chemistry. From preliminary observations[21,23] this bonding feature may turn out to be ubiquitous in bonding of heteropolar bonds, hypercoordinated molecules, and excited states. A task therefore lies ahead to find the chemical consequences of this bonding type on structure and reactivity.

(23) Our preliminary VB computations for the hypercoordinated $(XMX)^-$ species show large RE values (65-93 kcal/mol) for the cases where X = F and M = CH_3, SiH_3, and H respectively, and small RE values for X = H.

Acknowledgment. The research at BGU is supported by the Basic Research Foundation administered by the Israel Academy of Sciences and Humanities.

Microscopic Modeling of Ligand Diffusion through the Protein Leghemoglobin: Computer Simulations and Experiments

Gennady Verkhivker,[†] Ron Elber,*[,†] and Quentin H. Gibson[‡]

Contribution from the Department of Chemistry, M/C 111, University of Illinois at Chicago, P.O. Box 4348, Chicago, Illinois 60680, and Department of Biochemistry, Cornell University, 107 Biotechnology Building, Ithaca, New York 14853. *Received February 10, 1992*

Abstract: The diffusion of carbon monoxide through lupine leghemoglobin was investigated. The potential of mean force, the transition-state theory rate constant, the friction kernel, the transmission coefficient, and the diffusion constant were calculated. The computations are based on our previous exploration of the diffusion dynamics using the mean field method (LES)[12] and on our calculation of the reaction coordinate.[13] The back of the heme pocket (close to phenylalanine 44 and phenylalanine 29) is a shallow free energy minimum for the dissociated ligand. The minimum is directly accessible (without a barrier) from the binding site. The barrier for escaping from the free energy minimum to the CE loop is low (approximately 3 kcal/mol). Once the ligand leaves the pocket, the diffusion is barrierless. The ligand escapes in two steps. In the first (activated) step the ligand is hopping from the heme pocket to the protein interior, and in the second step it diffuses through the protein matrix to the surface of the macromolecule. The transition-state theory (which is appropriate for activated processes) is used for the first part of the process. For the second part a diffusion model is constructed. The calculated friction kernel and its power spectrum strongly depend on the reaction coordinate. The power spectrum is consistent with previous interpretations of the diffusion dynamics. In the first step of the process the barrier is local and the power spectrum shows only high-frequency modes. In the second step significant coupling to low-frequency (extended) modes is observed, and the diffusion coordinate is dominated by motions of the C and the G helices of the protein. Experimental results for ligand rebinding kinetics in lupine leghemoglobin are reported. It is shown that different diatomic ligands have an unusually fast diffusion rate in accord with theory.

I. Introduction

The activated diffusion of a small ligand through a protein matrix attracted considerable attention in the past. Perutz[1] noted that, according to the X-ray structure of hemoglobin, there is no obvious way for the ligand to escape from the protein matrix to the solvent. Since then, thermal fluctuations of the protein, which open transient gates for the ligand diffusion, were the focus of a number of theoretical investigations.[2-12] There are two extreme atomic models of ligand diffusion which one may have in mind. One is of ligand escape along a well-defined and (almost) unique path, and the second is of diffusion through a large number of alternative channels. We call the first the "hole" model and the second the "sponge" model.

The pioneering calculations of Case and Karplus,[2] Case and McCammon,[7] and Kottalam and Case[8] focused on the application of the "hole" model to the protein myoglobin. Tilton et al. addressed the question of alternate paths by simulating the motion of a probe particle through a rigid[9] and flexible[10] myoglobin. Their studies suggest that the single path assumption may be too restrictive. Elber and Karplus[11] provided more support to the existencee of multiple paths in myoglobin. They employed their LES method to investigate the diffusion of carbon monoxide through myoglobin. The LES method was designed to provide efficient search for possible openings in a fluctuating protein structure; therefore, more paths were detected than in previous studies.[2,9,10] The searches were, however, approximate, and the existence of alternative diffusion routes still awaits experimental and theoretical verification.

It is of interest to extend the investigations of diffusion routes to other proteins, especially ones with significantly different binding properties. We investigated recently lupine leghemoglobin which is a "relative" of the protein myoglobin.[12,13] Leghemoglobins are plant proteins with a global fold similar to that of myoglobin. The binding properties of the two protein families are, however, very different. For example, in soybean leghemoglobin the diffusion rate is much faster than in sperm whale myoglobin.[14,15] Only

[†] University of Illinois.
[‡] Cornell University.

(1) Perutz, M. F.; Mathews, F. S. *J. Mol. Biol.* **1966**, *21*, 199.
(2) Case, D. A.; Karplus, M. *J. Mol. Biol.* **1979**, *132*, 343.
(3) Agmon, N.; Hopfield, J. J. *J. Chem. Phys.* **1983**, *79*, 2042.
(4) Goldstein, R. F.; Bialek, W. *Biophys. J.* **1985**, *48*, 1027.
(5) Stein, D. L. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 3670.
(6) Miers, J. B.; Postlewaite, J. C.; Zyung, T.; Chen, S.; Roemig, G. R.; Wen, X.; Dlott, D. D.; Szabo, A. *J. Chem. Phys.* **1990**, *93*, 8771.
(7) Case, D. A.; McCammon, J. A. *Ann. N.Y. Acad. Sci.* **1986**, *482*, 222.
(8) Kottalam, J.; Case, D. A. *J. Am. Chem. Soc.* **1988**, *110*, 7690.
(9) Tilton, R. F., Jr.; Singh, U. C.; Weiner, S. J.; Connolly, M. L.; Kuntz, I. D., Jr.; Kollman, P. A.; Max, N.; Case, D. A. *J. Mol. Biol.* **1986**, *192*, 443.
(10) Tilton, R. F., Jr.; Singh, U. C.; Kuntz, I. D., Jr.; Kollman, P. A. *J. Mol. Biol.* **1988**, *199*, 195.
(11) Elber, R.; Karplus, M. *J. Am. Chem. Soc.* **1990**, *112*, 9161.
(12) Czerminski, R.; Elber, R. *Proteins* **1991**, *10*, 70.
(13) Nowak, W.; Czerminski, R.; Elber, R. *J. Am. Chem. Soc.* **1991**, *113*, 5627.
(14) Stetzkowski, F.; Banerjee, R.; Mardens, M. C.; Breece, D. K.; Bowne, S. F.; Doster, W.; Eisenstein, L.; Frauenfelder, H.; Reinisch, L.; Shyamsunder, E.; Jung, C. *J. Biol. Chem.* **1985**, *260*, 8803.

0002-7863/92/1514-7866$03.00/0

the structure of lupine leghemoglobin is available,[16] and this is the reason why we work on this member of the leghemoglobin family. Czerminski and Elber[12] studied the diffusion of carbon monoxide in lupine leghemoglobin using the LES method. The aim was to compare the theoretical estimates of the diffusion rates for the protein sperm whale myoglobin and lupine leghemoglobin using the same (approximate) computational method (LES). The ligand escape rate in leghemoglobin was found to be faster than in myoglobin, in accord with the experimental differences between myoglobins and leghemoglobins. Furthermore, the LES technique suggested that the mechanism for ligand penetration and escape is different for the two proteins. In myoglobin the "sponge" model (many alternate diffusion routes) seems appropriate;[11] in contrast, for leghemoglobin the "hole" model is supported by a simulation.[12] Thus leghemoglobin is a more convenient candidate for detailed theoretical studies than myoglobin. This is since only one dominant path needs to be investigated.

The LES method provides a rough estimate of the reaction coordinate. This coordinate was refined by Nowak, Czerminski, and Elber to a minimum energy path using the self penalty walk (SPW) method.[13] In this manuscript we use the minimum energy path calculated previously to compute the free energy profile for the diffusion process in lupine leghemoglobin. We also obtain an estimate for the diffusion time scale using a two-step mechanism. In the first step which is activated, the ligand leaves the heme pocket; in the second step, it diffuses through the protein interior to the surface. We extract a diffusion constant for the second step. In addition to the estimate of the diffusion rate, we analyzed the friction kernel as a function of the reaction coordinate. This helps us to understand the nature of the coordinates that are coupled to the reaction path. The interplay between coupling to low-frequency extended coordinates and to high-frequency local coordinates will be demonstrated in connection with the qualitative mechanism presented in refs 12 and 13. From this point of view, the present manuscript is a direct continuation of the work presented in these references.

The prime goal of the present work is to study diffusion through the protein matrix. Our interest is therefore focused on the fluctuations of the atoms in the interior of the protein and not on the penetration through the external solvent molecules. Globin fluctuations were represented reasonably well in a number of vacuum simulations taking into account the solvent effect as dielectric screening.[17] We did not, therefore, include explicit water molecules in our calculations (only one internal water molecule observed in the X-ray structure[16] was included). The transition from the protein to the solvent will be studied in a future work. It should be emphasized that the complete protein molecule was considered. This allows the investigation of the influence of extended motions such as translation and rotation of complete helices on the diffusion dynamics. These motions have been shown to be important in reducing the barrier for the diffusion process significantly.[12,13] We also looked for the influence of the solvent by repeating the calculations for the diffusion constant for a protein covered by a solvation shell. This was done in the flat portion of the potential of mean force where the extended protein motions are most likely to be influenced by the solvent.

The manuscript is organized as follows. In section II we describe the computational methods and present a test case for the specific implementation of the umbrella sampling which we used. In section III we outline the results. Discussion is in section IV, and final remarks are given in section V. The experimental results are given in the Appendix.

II. Method

In this section we describe how the numerically derived minimum energy path is used to calculate the potential of mean force and outline the protocol used to estimate the effective mass along the reaction coordinate, the rate constant, the transmission coefficient, the friction kernel, and the diffusion constant.

The mass is not really required in the diffusive non-activated part of the process, it is still useful even in this domain for qualitative understanding of different contributions to the reaction coordinate. Obviously for the activated part of the process, the mass is essential. The transmission coefficient is employed in correcting the rate constant for multiple crossing of the dividing surface. The friction kernel is used for the qualitative analysis of the diffusion mechanism and in a rough estimate of the diffusion constant. In the second more accurate calculation of the diffusion constant, we employed the Einstein relation.[18]

(i) **The Minimum Energy Path.** We begin by briefly describing the reaction path obtained previously and the methodology that was employed.[13] The reaction path q describes changes in the molecular system as it moves from a "reactant" to a "product" configuration. q is a necessary input for the calculation of the potential of mean force. It is a curvilinear coordinate which we present by a grid of L copies of the system $\{\mathbf{R}_{qi}\}_{i=1,\dots,L}$ along the path.[19]

We used the SPW algorithm[19] that is designed for calculating minimum energy paths in large molecular systems. It provides paths that were demonstrated to be close to the steepest descent path. In SPW calculations, the energy of a linear polymer composed of M monomers is minimized. Each of the monomers is a complete copy of the physical system (in the present case, leghemoglobin, the internal water molecule, and the ligand). The first monomer is the reactant (ligand in the heme pocket) and the last monomer is the ligand outside the protein matrix. The minimum energy of the polymer is a discrete representation of the reaction coordinate. The energy of the polymer, S, is the following sum

$$S = \sum_{i=1}^{M} V(\mathbf{R}_{qi}) + \sum_{i=0}^{M} B_i + \sum_{\substack{i,j=0 \\ j>i+1}}^{M+1} (\text{REPULSION})_{ij}$$

$$B_i = \gamma(d_{i,i+1} - \langle d \rangle)^2$$

$$(\text{REPULSION})_{i,j} = \rho \exp\left(-\frac{\lambda d_{i,j}^2}{\langle d \rangle^2}\right)$$

$$d_{ij} = [(\mathbf{R}_{qi} - \mathbf{R}_{qj})^2]^{1/2}; \quad \langle d \rangle = \left(\frac{1}{M+1}\sum_{i=0}^{M} d_{j,j+1}^2\right)^{1/2}$$

where $\mathbf{R}_0$ refers to the reactants and $\mathbf{R}_{M+1}$ to products. γ, the "bond force constant" between the monomers, ensures that the monomers are equally distributed along the path. γ does not affect the value of the barrier. It should be chosen sufficiently high to satisfy the constraint but not too high, since then the numerical optimization will be difficult. ρ and λ determine the repulsion between the monomers that was shown to minic kinetic energy.[19] Thus, the minimum energy path is obtained as a set of $\mathbf{R}_i$ that minimizes S for a given set of parameters γ, ρ, and λ. Since ρ and λ are correlated, we vary only ρ so that the barrier will be the lowest possible and still give a continuous path.

The R's are given in Cartesian space. We emphasize that we do not have an analytical expression for $q(\mathbf{R})$ but only points equally spaced along it. The direction of q at q_i (in Cartesian space) is defined by a unit vector, $\mathbf{e}_{qi}$

$$\mathbf{e}_{qi} = \frac{\mathbf{R}_{qi+1} - \mathbf{R}_{qi}}{|\mathbf{R}_{qi+1} - \mathbf{R}_{qi}|} \tag{1}$$

In this manuscript q is *defined* as a sum of all the scalar products between the difference vectors of sequential configurations and the path slope.

$$q_j = \sum_{i=1}^{j} (\mathbf{R}_{qi+1} - \mathbf{R}_{qi})\mathbf{e}_{qi} \tag{2}$$

(15) Gibson, Q. H.; Wittenberg, J. B.; Wittenberg, B. A.; Bogusz, D.; Appleby, C. A. *J. Biol. Chem.* **1989**, *264*, 100.
(16) Arutyunyan, E. G.; Kuranova, I. P.; Vainstein, B. K.; Steigmann, W. *Sov. Phys. Crystallogr.* **1980**, *25*, 43.
(17) Kuczera, K.; Kuriyan, J.; Karplus, M. *J. Mol. Biol.* **1990**, *213*, 351.
(18) Kubo, R.; Toda, M.; Hashitsume, N. *Statistical Physics. II. Nonequilibrium Statistical Mechanics*; Springer Verlag: Berlin, 1985.
(19) Czerminski, R.; Elber, R. *Int. J. Quantum Chem.* **1990**, *24*, 167.

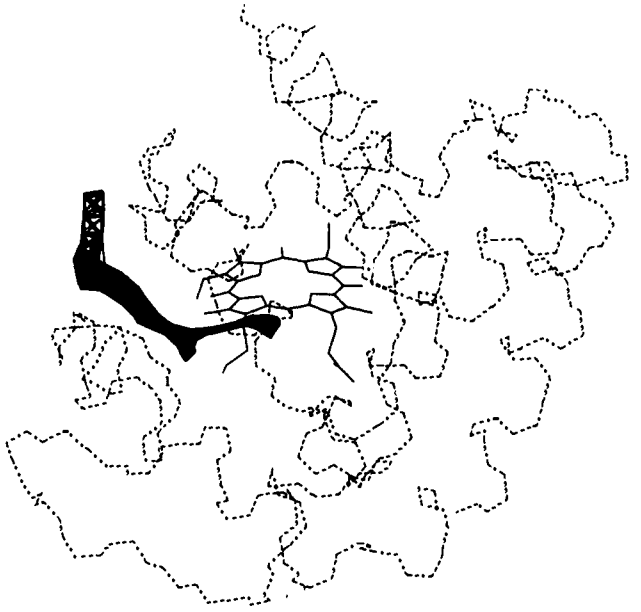

Figure 1. The minimum energy path for the motion of carbon monoxide through leghemoglobin. For clarity only the heme group and the backbone of the protein (dashed line) are shown. The length of the path is approximately 6 Å and the ligand escapes between the C and G helices.

The step size (or the metric) of q cannot be determined unambiguously by the R's alone. For example, the summation in eq 2 that is used to *define* q can be carried out instead over *any* monotonic function of the scalar product. The definition of q cannot (of course) change the dynamics, but it can change (for example) the effective mass. Note that the numerical values of the Cartesian vectors depend (of course) on the choice of the coordinate system. The $\mathbf{R}_q$'s are therefore oriented so that the calculated value of the distance between sequential structures is the smallest possible (for details see ref 13).

q and the set of R's were calculated in ref 13 for the diffusion of carbon monoxide in leghemoglobin. Three similar but distinct paths were investigated. The calculations presented here were pursued for path C. Three new path segments were added to the original path using the SPW algorithm.[14] Four structures were added in the neighborhood of the first local energy barrier, and nine structures were added in the neighborhood of the second barrier. This is to impove the convergence of the free energy calculations by decreasing the distance between sequential structures. The SPW parameters were γ = 128 kcal/mol, ρ = 32 kcal/mol, and λ = 2, which are the same as the ones used in ref 13. One point (at the top of the barrier) was added with a stronger repulsion parameter ρ = 3200 kcal/mol, to ensure better overlap of distributions before and after the barrier. The additional path segments did not change any of the qualitative features of the path; however, the (total) energy barrier was reduced by 4.07 kcal/mol. A schematic view of the trajectory of carbon monoxide moving through the protein frame is shown in Figure 1.

Throughout this text the CHARMm potential energy is employed.[20] The extended atom model (CH_n groups are modeled as spheres) is used. The cutoff distance for nonbonded interactions (electrostatic and van der Waals) was 9 Å with a smooth truncation using the shift cutoff function. The 1–4 scaling factor was 1.

(ii) The Potential of Mean Force. The potential of mean force at a specific position along the reaction path $W(q)$ is given by the following thermal average over the rest of the degrees of freedom

$$W(q) = (-1/\beta) \log [\langle \exp(-\beta U(q,\mathbf{Q}))\rangle_{\mathbf{Q}}] \quad (3)$$

where $\beta = (k_B T)^{-1}$, Q is the coordinate set complementary to q, and U is the microscopic potential. The brackets $\langle ... \rangle_{\mathbf{Q}}$ denote thermal average with respect to the Q coordinates. We focused on the determination of the difference between the potential of mean force along q, e.g., at q_1 and q_2:

$$W(q_1) - W(q_2) = (-1/\beta) \log [\langle \exp(-\beta(U(q_1,\mathbf{Q}) - U(q_2,\mathbf{Q})))\rangle_{\mathbf{Q}}] \quad (4)$$

In ref 21 we described how to obtain the potential of mean force along a numerical presentation of a reaction path using the free energy perturbation method. Below we show how the umbrella procedure can be used as well. Umbrella sampling has some weak and some strong points compared to the free energy perturbation. The umbrella sampling (in contrast to the free energy perturbation) requires "human intervention" in performing the matching of the overlapping distributions.[22] This is, of course, a disadvantage. An advantage of the umbrella procedure is that the *distributions* along q are calculated first and the potential of mean force is extracted from the (already averaged) distributions. In the free energy perturbation we are interested in the direct calculation of the average on the right-hand side of eq 4. Individual points from the molecular dynamics simulation contribute directly to the average. Consider the function $\Delta = -\beta(U(q_1,\mathbf{Q}) - U(q_2,\mathbf{Q}))$ and assume that q_1 and q_2 are sufficiently close to each other so that most of the Δ's are zero. If the trajectory samples a configuration with Δ of order of 10 (which is not unheard of in flexible and large systems), then the weight of this point, $\exp(\Delta)$, is approximately 22 000; i.e., the contribution of this single point to the average is the same as other 22 000 points of the typical value one. This may result in poor statistics. In the umbrella procedure, configurations with low statistical weight are unlikely to contribute to the average, since the matching (see below) is performed only for q's which were sampled significantly during the simulation. We found the umbrella sampling more adequate for systems in which the reaction coordinate consists of more than a few hundred atoms. This is the case of lupine leghemoglobin, which in our model consists of 1474 particles including the carbon monoxide and one internal TIP3P water molecule.[23]

(iii) Umbrella Sampling Method. In principle, one may calculate the potential of mean force (eq 3) by straightforward molecular dynamics. Thus, an ordinary molecular dynamics trajectory (or Monte Carlo simulation) is used to sample configurations in both q and Q. Then, the distribution of the trajectory configurations along $q - P(q)$ is calculated as the average over the trajectory of the function $\delta(q - q(t))$, i.e., $\langle \delta(q - q(t))\rangle_t$. The potential of mean force $W(q)$ is finally given by $W(q) = (-1/\beta) \log (P(q))$.

The problem in the straightforward approach is that near the barrier the number of configurations sampled would be small. This results in poor statistics and low accuracy in the estimate of the free energy barrier.

In the umbrella method[22] the molecular dynamics trajectory is forced to "visit" high-energy positions of the potential energy surface, and a better sampling is obtained.

We consider the following biasing potential for the position q_i along the reaction coordinate. The potential below constrains the system to the plane perpendicular to q at q_i

$$U_{Bi}(\mathbf{R};q_i) = \tfrac{1}{2}K[(\mathbf{R} - \mathbf{R}_{qi})\cdot\mathbf{e}_{qi}]^2 \quad (5)$$

where $\mathbf{e}_{qi}$ is the unit vector parallel to the direction of the reaction coordinate at q_i that was defined in eq 1. K is a constant chosen to provide good sampling at q_i and at the same time to provide overlap between sequential windows (we call a system with a biasing potential at q_i "window i"). U_{Bi} is added to the full microscopic potential and a molecular dynamics simulation is carried out with the new effective potential. We note that, since the reaction coordinate is given in Cartesian space, it is important to eliminate the rigid body motions during the dynamics. Rotations and translations may affect the reaction coordinate value.

(20) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.

(21) Elber, R. *J. Chem. Phys.* **1990**, *93*, 4312.

(22) (a) Patey, G. N.; Valleau, J. P. *Chem. Phys. Lett.* **1973**, *21*, 297. (b) Valleau, J. P.; Torrie, G. M. In *Statistical Mechanics*, Part A; Berne, B. J., Ed.; Plenum Press: New York, 1977; pp 169–194.

(23) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

The elimination is carried out using linear constraints on the system. The application of linear constraints in molecular dynamics simulations was outlined in detail in ref 21. For completeness, we write below the appropriate differential equations of motion

$$m_j \frac{d^2 r_j}{dt^2} = -\nabla_j (U + U_{Bi}) - \sum_{l=1}^{6} \lambda_l \nabla_j \sigma_l \quad (6)$$

where r_j is the coordinate vector of the jth atom, m_j is its mass, and ∇_j is the corresponding gradient. The σ_l are constraints on the rigid body motions

$$\sigma_{l=1,3} = \sum_{j=1}^{N} m_j (r_j - r_{j0}) \quad (7a)$$

$$\sigma_{l=4,6} = \sum_{j=1}^{n} m_j r_{j0} \times r_j \quad (7b)$$

where r_{j0} is the atomic coordinate vector at the beginning of the simulation. As was shown in ref 21, for each time step it is possible to find an explicit expression for the Lagrange multipliers, λ_l's. This reduces eqs 6 and 7 to an ordinary molecular dynamics simulation.

Hence a molecular dynamics trajectory is employed to sample configurations from the effective potential $U + U_{Bi}$ that satisfy the rigid body constraints.[7] From a single simulation, a probability density $P_i(q)$ in the neighborhood of q_i is extracted as a trajectory average of $\delta(q - q(t))$; i.e., $P_i(q) = \langle \delta(q - q(t)) \rangle_i$. The biasing potential U_{Bi} is then shifted to be centered on q_{i+1} and the process is repeated to obtain $P_{i+1}(q)$. The two distributions are then matched by requiring that for overlapping regions the value of the logarithm of the probabilities will be the same.[22] Good overlap is obtained when $P_i(q)$ and $P_{i+1}(q)$ differ appreciably from zero and have similar curves that differ by a constant for q's between q_i and q_{i+1}.

We checked the protocol against a system for which the results are known. The potential of mean force for a conformational transition in a solvated valine dipeptide was computed in ref 21 using the free energy perturbation method. Here we repeated the calculation using the umbrella sampling procedure. The energy parameters were (of course) the same as in ref 21, and the biasing force constant was 5 kcal/(mol Å^2). For each window 5 ps was employed for equilibration and 20 ps for data collection; 25 windows were employed with a typical distance of 0.1 Å between sequential points (the distance between the ith and the jth structures, d_{ij}, is defined by $d_{ij}^2 = (1/N)\sum_k (r_k^i - r_k^j)^2$, where k is the atom index). In Figure 2 we show a segment of the potential of mean force starting from the minimum at the following values of the peptide dihedral angles (ϕ = 61.4°, Ψ = −64.0°) and continuing to the barrier at (ϕ = 5.9°, Ψ = −47.8°). Within the typical error bars of such calculations, the agreement between the two methods is reasonable. We therefore proceed to calculate the potential of mean force for carbon monoxide diffusion in leghemoglobin using the umbrella procedure.

In the study of leghemoglobin, 78 $P_i(q)$ distributions (or 78 windows) were calculated. The average distance between the structures along the minimum energy path was 0.08 Å which ensures good overlap (the largest distance was 0.23 Å, and the shortest was 0.02 Å). At each window 5 ps was used for equilibration and 40 ps for data collection. In 12 windows we found it necessary to increase the simulation time to 80 ps. The biasing force constant that we used was K = 50 kcal/(mol Å^2), but in 20 windows K was reduced to 25 kcal/(mol Å^2). This is in order to obtain better overlap between the distributions.

It cannot be over emphasized that the calculations for leghemoglobin are significantly more difficult than those for the valine dipeptide. The size, the heterogeneity of the system, and the broad range of frequencies make the leghemoglobin calculations converge less reliably, as was also shown by our estimates of the statistical errors. The errors (of order of 0.5 kcal/mol) are significantly larger than in valine dipeptide and were obtained by dividing the 40 ps of data collection into two segments, and repeating the matching for the two smaller data sets. While we cannot prove (of course) that the results are truly converged, we do not expect the statistical error bars to be larger than 1 kcal/mol.

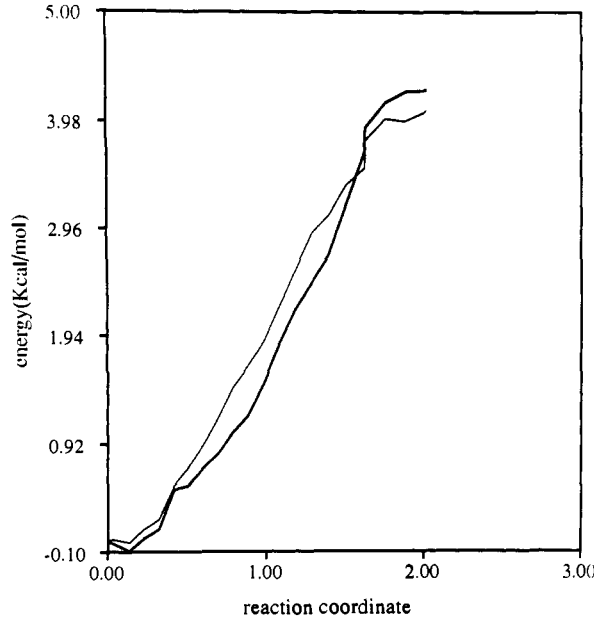

Figure 2. A comparison study between the free energy perturbation method and the umbrella sampling for calculation of the potential mean force. The calculation is done along a numerical reaction coordinate for a conformational transition in valine dipeptide as described in the text. The calculation is from the (ϕ = 61.4°, Ψ = −64.0°) minimum to the barrier at (ϕ = 4.9°, Ψ = −47.8°). The thin line represents the results of the free energy perturbation[21] and the thick line the results of the present study using the umbrella sampling procedure.

We also comment that, since the potential of mean force is calculated at a fixed orientation, there is a need to add to the potential of mean force the centrifugal term. We examined this correction in initial calculations using the method described in ref 21, and we found it to be very small (typical value was 0.002–0.004 kcal/mol). We therefore ignored it in further calculations.

(iv) The Effective Mass, $m(q)$. The effective mass is required only at the top of the barrier, where we calculate the rate constant using the transition-state theory. The mass is not required in the flat portion of the free energy surface in which the process is not activated. However, the dependence of the mass on the reaction coordinate is still of considerable interest. It serves as an indicator of the variation of the coupling between the ligand motion and the protein degrees of freedom, and we therefore calculate the mass for the complete reaction coordinate.

Consider a specific fixed q, say q_i. The coordinates orthogonal to q_i are $\mathbf{Q}_i$ which for simplicity we take to be Cartesian. We can write the kinetic energy T in Cartesian space as

$$T = \frac{1}{2} \mathbf{v} \mathbf{M} \mathbf{v} = \frac{1}{2} \sum_{k=1}^{3N} m_k v_k^2 \quad (8)$$

where $\mathbf{M}$ is the (diagonal) mass matrix and $\mathbf{v}$ is the velocity vector of all atoms in Cartesian space. At q_i the velocity component along q is $(\mathbf{v}, \mathbf{e}_{qi})$ where the (..) denote a scalar product. If we define $|\mathbf{e}_{qi})(\mathbf{e}_{qi}|$ as a projection operator on the q_i direction, we can separate the kinetic energy to q and $\mathbf{Q}$ parts where the T_Q and T_q are given by

$$T_q = \tfrac{1}{2} (\mathbf{v}, \mathbf{e}_{qi})(\mathbf{e}_{qi}|\mathbf{M}|\mathbf{e}_{qi})(\mathbf{e}_{qi}, \mathbf{v}) \quad (9a)$$

$$T_Q = \tfrac{1}{2} (\mathbf{v}|[\mathbf{1}](\mathbf{1}| - |\mathbf{e}_{qi})(\mathbf{e}_{qi}|]\mathbf{M}|[|\mathbf{1})(\mathbf{1}| - |\mathbf{e}_{qi})(\mathbf{e}_{qi}|]|\mathbf{v}) \quad (9b)$$

The effective mass of the coordinate q is therefore identified as

$$m(q) = (\mathbf{e}_q|\mathbf{M}|\mathbf{e}_q) \quad (10)$$

(v) The Transition-State Theory and the Transmission Coefficient. The transition-state theory (TST) is a useful approach

to study the dynamics of activated processes. A fundamental assumption in the TST is that the hypersurface dividing the reactants and the products is crossed by the system only once. This assumption is likely to be correct when the barrier is considerably larger than the average thermal energy. This is because rare events are expected to occur only once. However, when the barrier is very low or non-existent, recrossing of the above surface may occur many times. Deviation from transition-state theory can be also related to "frictional effects". Too low friction yields inertial motion and many recrossings occur, and high friction causes many collisions at the transition domain and therefore multiple recrossing.

The TST rate constant can be corrected to the exact result using the transmission coefficient κ.[24] κ modifies the rate constant by taking into account the presence of recrossing; i.e., the exact rate constant can be formally written as $k = \kappa k^{TST}$. A necessary condition for κ to be a useful computational tool is a separation of time scales; i.e., the time scale of "sliding" down from the barrier, t_s, would be much shorter than the time scale for recrossing, t_r. This makes κ approximately constant at times, t, so that $t_s \ll t \ll t_r$.

Formally $\kappa(t)$ is given by

$$\kappa(t) = \frac{\langle \nu_+ \Theta(q(t) - q^\#) \rangle_{q^\#}}{\langle \nu_+ \rangle_{q^\#}} \tag{11}$$

where the average is for initial configurations sampled at the top of the barrier (denoted by $q^\#$); ν_+ is thermal velocity from Boltzmann distribution at 300 K with a positive component along q. Θ is a step function that measures the "success" of the trajectory. Θ is one if the trajectory is on the product side (including the top of the barrier) and zero if the trajectory is in the reactant domain.

The first step was to choose structures from the hypersurface defined by $q = q^\# = 2.1$ Å that are distributed thermally. A molecular dynamics trajectory that was linearly constrained to be exactly at $q^\#$ [21] was employed to sample the required configurations. The sampling trajectory was of 50 ps and structures were saved every 0.5 ps, providing us with a total of 100 different initial structures to pursue the average of eq 11.

For a structure from the set, initial velocities were chosen at random from the 300 K distribution (with a positive component along q), and a 5-ps molecular dynamics trajectory was computed. This process was repeated 100 times. The length of the individual trajectories (5 ps) is comparable to the value used in previous investigations of κ in other biological systems.[25,26] The results (Figure 3) indicate that the calculation converged by that time.

The rate constant is written as

$$k = \kappa\omega \exp[-\beta(W(q^\#) - W(q^0))] \tag{12a}$$

where ω is a frequency factor and $W(q^0)$ is the value of the potential of mean force at the minimum. We already described how we compute W and κ. ω was estimated as[24]

$$\omega = \frac{(1/2\pi\beta m_q)^{1/2}}{\int_0^{q^\#} dq \exp[-\beta(W(q) - W(q^0))]} \tag{12b}$$

where the numerator is the average thermal velocity in the positive q direction and the denominator is the configurational integral for the reactant part.

(vi) The Memory Function, the Langevin Equation, and the Diffusion Time Scale. A phenomenological equation which has been demonstrated in the past to be useful in simulating molecular processes is the Langevin equation

$$\frac{dp_q}{dt} = -\frac{dW}{dq} - \Gamma p_q + R \tag{13}$$

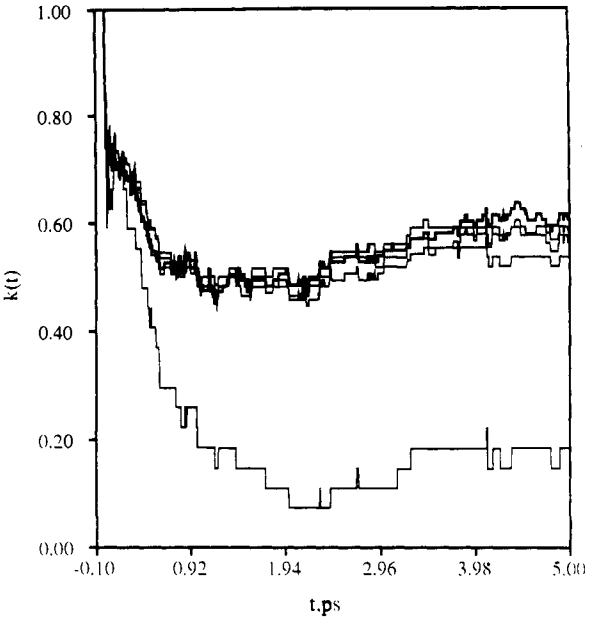

Figure 3. The calculation of the transmission coefficient; 100 trajectories were initiated at the top of the barrier at $q = 2.1$ Å. Four curves are shown that correspond to the different samples of 30, 50, 70, and 100 trajectories. The calculation seems to converge to the value of 0.6 after 50 trajectories. See text for more details.

p_q is the conjugate momentum to q, and W is here an effective potential that is taken to be the potential of mean force, Γ is phenomenological friction, and R is the random force which is related to the friction Γ via the fluctuation–dissipation theorem. One application of eq 13 is for a semiquantitative estimate of the diffusion rate (after the first activated process); a second application is in qualitative and molecular level analysis of the coupling of q to other degrees of freedom. In particular, we should like to extract the functions W and Γ from a microscopic model. In order to estimate the validity of (13), it is useful to examine the "friction" approximation. An intermediate between the Langevin equation and a microscopic model is the generalized Langevin equation. There the constant friction Γ is replaced by the friction kernel to give the generalized Langevin equation:[18,26-29]

$$\frac{dp_q}{dt} = -\frac{dW}{dq} - \int_0^t \gamma(t - \tau) p_q(\tau)\, d\tau + R \tag{14}$$

It is difficult to calculate the friction kernel exactly from a microscopic model. Some approximations are therefore employed. The approximate technique used to calculate the memory function is based on the force fluctuations with a frozen reaction coordinate. This approximation has been employed on other systems by Bergsma et al.,[27] Berne et al.,[28] and Roux et al.[26]

The force–force correlation function for a fixed value of the reaction coordinate is considered. The microscopic force fluctuations, say δF, are associated with the random force R as follows. The forces are calculated from a detailed molecular dynamics trajectory at a fixed value of q. Then the deviation δF from the mean force (the derivative of the potential of mean force) is calculated and projected along q; i.e., δF is given by

$$\delta F = (e_q | \mathbf{F}(t;q) - \langle \mathbf{F}(t';q) \rangle)$$

The friction kernel is then calculated from the fluctuation dissipation theorem[18] where we identify the random force R with δF

$$\gamma(\tau) = (\beta/m_q)\langle \delta F(t) \delta F(t + \tau) \rangle_t \tag{15}$$

(24) Chandler, D. *J. Chem. Phys.* **1978**, *68*, 2959.
(25) Northrup, S. H.; Pear, M. R.; Lee, C. L.; McCammon, J. A.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *79*, 4035.
(26) Roux, B.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 4856.
(27) Bergsma, J. P.; Reimers, J. R.; Wilson, K. R.; Hynes, J. T. *J. Chem. Phys.* **1986**, *85*, 5625.
(28) Berne, B. J.; Tuckerman, M. E.; Straub, J. E.; Bug, A. L. R. *J. Chem. Phys.* **1990**, *93*, 5084.
(29) Straub, J. E.; Berne, B. J.; Roux, B. *J. Chem. Phys.* **1990**, *93*, 6804.

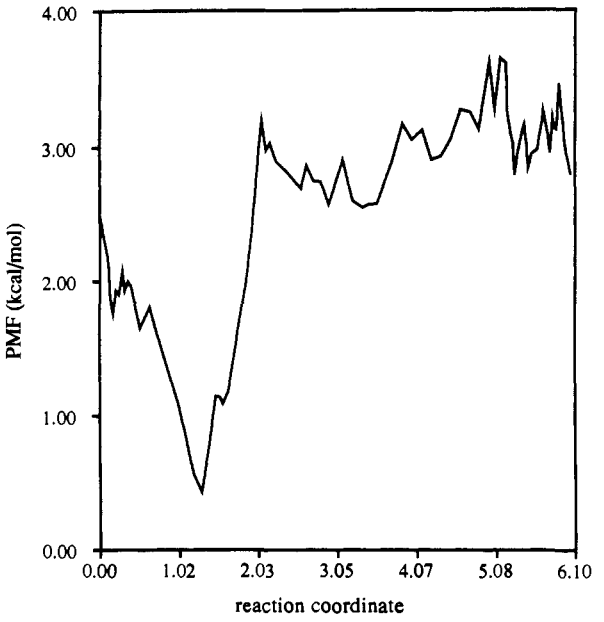

Figure 4. The potential of mean force for the reaction coordinate associated with the diffusion of a diatomic ligand (carbon monoxide) through leghemoglobin. Note the shallow free energy minimum which is at the back of the heme pocket. The coordinate length is 6.1 Å and it is associated with a considerable amount of protein motions and not only with that of the ligand.

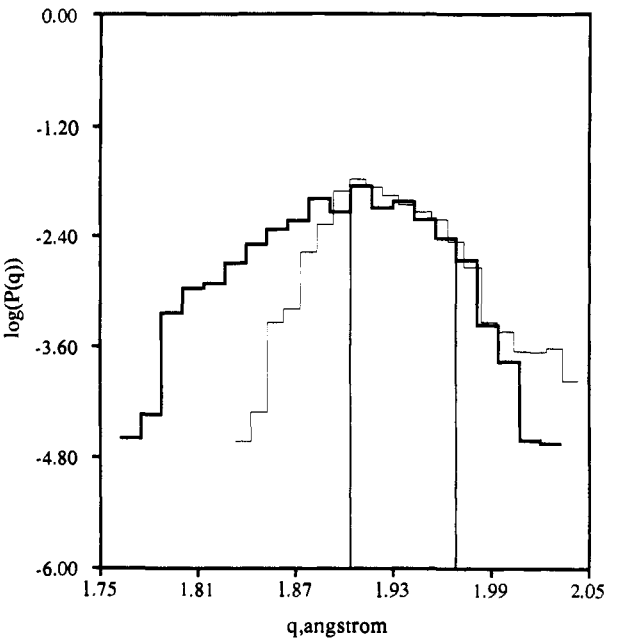

Figure 5. An example for a window matching in the umbrella sampling: a match of the distributions on the top of the barrier. The distributions were calculated at q_i = 1.90 Å and q_{i+1} = 1.96 Å using an 80-ps trajectory. See text for more details.

$F(t;q)$ depends parametrically on q since it is evaluated at a fixed q. Thus, γ may depend on q which is indeed the case as we shall show later. The average $\langle \cdots \rangle_t$ denotes an ensemble average, which we performed using the structures obtained from the molecular dynamics calculations.

$\gamma(t)$ was calculated at three positions along the reaction coordinate (i) close to the free energy minimum (q = 1.00 Å), (ii) on the barrier (q = 2.07 Å), and (iii) on a point in the free energy plateau (q = 3.45 Å). The calculations were performed by averaging over 5-ps trajectories at each position.

From the memory function the static friction Γ which appears in the Langevin equation was extracted:

$$\Gamma = \int_0^\infty dt\, \gamma(t) \tag{16}$$

The friction Γ is (in principle) all that we need in the relatively flat portion of the potential energy surface (after the first barrier). The variations in the potential of mean force in that segment of the free energy curve are probably within the error bars of the calculations. As a side effect we can estimate the diffusion constant, D, from the formula[18]

$$D = 1/\beta m_q \Gamma \tag{17}$$

The calculation of Γ by numerical integration of $\gamma(t)$ is, however, a difficult numerical task. $\gamma(t)$ is a rapidly oscillating function (see, for instance, Figure 7) that probably has a long tail that we missed. Furthermore, eq 17 is approximate, while the second formula which we used to calculate D (see below) is more direct and exact. We calculated D from the following formula:[18]

$$D = \langle [q(t) - q(0)]^2 \rangle / 2t \tag{18}$$

Three 50-ps trajectories starting at different q's (3.12 Å, 3.25 Å, 3.45 Å) in the flat portion of the mean force potential were employed. The trajectories were "normal" with no constraints or biasing potentials. The q's were saved at each step (1 fs) and were employed in the average required to compute D.

III. Results

In Figure 4 we show the potential of mean force calculated by the umbrella sampling procedure that was described in the Methods section. Note that at the last point, q = 6.1 Å, the ligand is at the protein surface. The potential of mean force at that point is not reliable since the solvent was not taken into account in our calculations. The critical step in the calculation is the matching procedure. To obtain insight into the difficulties associated with this process, we show in Figure 5 the match for two distributions at the top of the barrier. The match is shown after the overlap of the log of the distributions, and it is for q_i = 1.90 Å and q_{i+1} = 1.96 Å. The results have considerable error bars, and the accuracy of the calculations is probably bound by ±1 kcal/mol, so that a very accurate analysis of the results is not possible. The maximum energy difference found on the surface constructed is ~3 kcal/mol. This means that the diffusion through the protein matrix (along the specified path) is relatively easy. The potential differences are small, and they do not suggest a model of an activated process for the diffusion.

A better test for an activated process is the use of the transmission coefficient, $\kappa(t)$. We consider a process to be activated if recrossing of the dividing surface occurs only during a short time interval after the system reaches the barrier; i.e., it is unlikely for the system to return to the top of the barrier for the second time if it already "fell" to the side of the reactants or the products. $\kappa(t)$ is a measure of the recrossing (see Methods section), and, if it relaxes *rapidly* to its "asymptotic" value at times considerably shorter than the reaction rate (i.e., no more crossing), we consider the process to be activated. We calculated $\kappa(t)$ at q = 2.1 Å, the position in which the ligand passes phenylalanine 29 and leaves the heme pocket. This position was identified as an energy as well as a free energy barrier. $\kappa(t)$ is plotted in Figure 3. Four curves are shown; the lowest one is of a sample of only 30 trajectories and the other three of 50, 70, and 100 trajectories, respectively. It is evident that the asymptotic value of κ, $\kappa(t \ll 1/k)$ = 0.60, was reached after ~3 ps in the curves of 50–100 trajectories. It is also clear that the time required for fall-off from the top of the barrier is similar in *four* curves and it is of order of 3 ps. This is considerably shorter than the typical time scale for passing this barrier as we show below. Hence according to this test the transition-state theory seems to be adequate regardless of the fact that the barrier is rather low.

The frequency factor ω was calculated according to eq 12b. The integration over the potential of mean force was performed numerically. We obtain ω = 3.9 ps^{-1}. Taking the barrier to be 2.8 kcal/mol and κ to be 0.60, we obtain an estimate for the rate of escape from the heme pocket in leghemoglobin at room temperature:

$$k = (0.60)(3.9)\exp(-2.8/0.6) = 0.022 \text{ ps}^{-1} = 22 \text{ ns}^{-1}$$

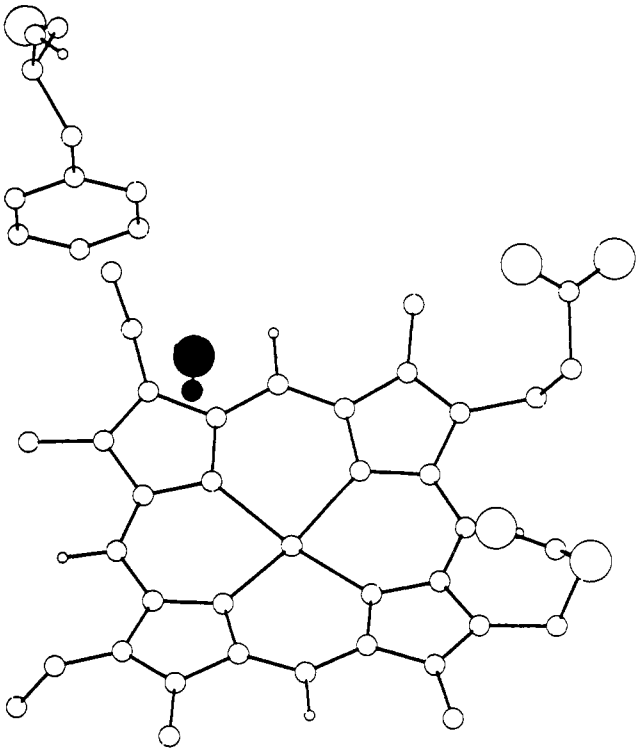

Figure 6. A stick and ball model of the structure along the minimum energy path that corresponds to the free energy minimum. For clarity only the heme, the ligand, and Phe 29 are shown. Note the displaced position of the ligand with respect to the iron in the heme center.

The greatest uncertainty in our calculation is the exact value of the free energy barrier. We estimate *statistical errors* of order of 0.5 kcal/mol, but additional sources of inaccuracy such as incomplete sampling due to trapping at certain portions of phase space are likely to increase the errors even further. We therefore provide the values for the rate constants for barriers of 3.8 and 1.8 kcal/mol as well: $k(3.8) = 4\ ns^{-1}$ and $k(1.8) = 117\ ns^{-1}$. We do not believe the lower bound ($k(1.8)$) since it is in contradiction with the independent calculation of the transmission coefficient. $k(1.8)$ corresponds to an escape time of approximately 9 ps. This time is very close to the relaxation time of the transmission coefficient (3 ps). It is therefore unlikely that, in the 5-ps trajectories which we used to calculate κ, the results would have been converged.

Thus, in spite of the low free energy barrier, we consider the first step in which the ligand leaves the heme pocket to a cavity between the C and G helices and the CE loop to be activated. The time scale for this process is estimated to be between several tens to several hundreds of picoseconds. The time scale associated with $k(2.8)$ is 45 ps and with $k(3.8)$ is 250 ps.

It is interesting to note that the starting point of the ligand (which is the ligand position in the X-ray structure[16]) leads directly (without a barrier) to a low free energy minimum. The position of the ligand, the heme, and phenylalanine 29 in the corresponding minimum energy structure is shown in Figure 6. The minimum is located at the back of the pocket near residues phenylalanine 29, phenylalanine 44, and the internal water molecule. The well depth is approximately 3 kcal/mol.

After the first barrier the free energy profile (within the accuracy of the calculations) is close to a plateau. It is clear that in this segment of the reaction coordinate the process is diffusive, and we did not attempt to identify any specific barriers or to apply the transition-state theory in this domain.

We proceed with the calculation of the parameters of the Langevin equation. An essential ingredient of the stochastic picture is the friction coefficient Γ. We calculated Γ in steps. First we estimated the friction kernel by an approximate procedure based on molecular dynamics calculations with a frozen reaction coordinate (see the Methods section). Then we integrated the friction kernel $\gamma(t)$ to obtain Γ (eq 12). Hence, $\gamma(t)$ is employed as a means to calculate Γ; however, it is also useful in qualitative analysis of the diffusion as we shall demonstrate below.

$\gamma(t)$ is computed at three positions along the reaction coordinate: (i) close to the free energy minimum ($q = 1.00$ Å), (ii) on the barrier ($q = 2.07$ Å), and (iii) at a point in the plateau ($q = 3.45$ Å). For each of the positions along the reaction coordinate, we consider the time dependence of the friction kernel and its power spectrum, $\hat{\gamma}(\omega)$.

$$\hat{\gamma}(\omega) = \left| \int_0^\infty dt\, \exp(i\omega t)\gamma(t) \right|^2 \tag{19}$$

In Figure 7 we show the results for i–iii. The memory function has a rapidly oscillating component that decays relatively slowly compared to another study of memory functions for diffusion through a biological molecule (ions in the gramicidin channel[26]). While previous investigations have focused on coupling to a bath of low-frequency modes, here the reaction coordinate is coupled to a considerable number of high-frequency vibrations. The power spectra i and ii show a peak of the CO stretch at around 2100 cm^{-1}. The assignment of the 2100-cm^{-1} peak was tested by projecting out the CO contribution to $\gamma(t)$, i.e., by calculating the force–force correlation function (eq 15) without including the carbon monoxide contribution to the force fluctuations. This results in the disappearance of that peak in i–ii (see Figure 8). In the flat portion of the potential of mean force (iii), q is associated with the motion of protein atoms and less with the motion of the ligand. Therefore the 2000-cm^{-1} peak is significantly less pronounced (compared to i and ii), and it is not affected by the removal of the CO. This peak is associated with the protein carbonyl vibrations.

Another similar feature of i and ii is the peak at around 3800 cm^{-1} which corresponds to the O–H stretch of the internal water molecule (we are employing a flexible water molecule). This assignment was also confirmed by the removal of the water molecule. Thus, the free energy minimum (position i) is coupled almost entirely to high-frequency vibrations; in contrast to that, low-frequency modes seem to contribute to the friction kernel at the top of the barrier (position ii). The low-frequency motions are of the CO and the internal water molecule. The removal of both molecules (not only one of them) results in an essentially flat spectrum.

Thus, the present reaction coordinate which is relatively flat (or corresponds to a low-frequency mode) is coupled at the beginning (positions i and ii) to high-frequency vibrations that relax very slowly. On the computed time scale the memory function still oscillates rapidly as a function of time. There is a fast decay of the amplitude in the first 100 fs and then rapid oscillations of the remaining amplitude throughout the simulation period.

The power spectrum of $\gamma(t)$ (Figure 7) at the third position shows a qualitatively different picture. The spectrum is considerably more complex and it is *not* affected by the removal of either the internal water or the carbon monoxide. Hence it reflects protein fluctuations (atoms other than CO or water) and not the ligand motion. It should be noted that even this γ has a considerable fraction of high-frequency components and $\gamma(t)$ shows oscillatory motion up to 5 ps with little sign of relaxation.

According to our plan, we should now integrate $\gamma(t)$ to obtain the static friction Γ. This is a nontrivial task, since $\gamma(t)$ changes sign frequently, and the summation of a large number of numbers, most of them canceling each other, is a numerically inaccurate procedure. Nevertheless, we attempted to calculate the static friction at the three positions i–iii using eq 16. The static friction Γ at the three positions is (i) 3 ps^{-1}, (ii) 1 ps^{-1}, and (iii) 10 ps^{-1}. These values are low and reflect the inefficient relaxation of the vibrational modes and/or inaccuracies in the attempted integration. The friction coefficient for translational relaxation is typically 10 times faster and is of order of 100 ps^{-1}. While the reaction coordinate is coupled to the high-frequency motion of the protein, it is not obvious whether we are able to determine the relaxation of these modes accurately. We therefore separate the contributions of the two. Consider the Fourier transform of $\gamma(t)$, that is, $\gamma(\omega)$.

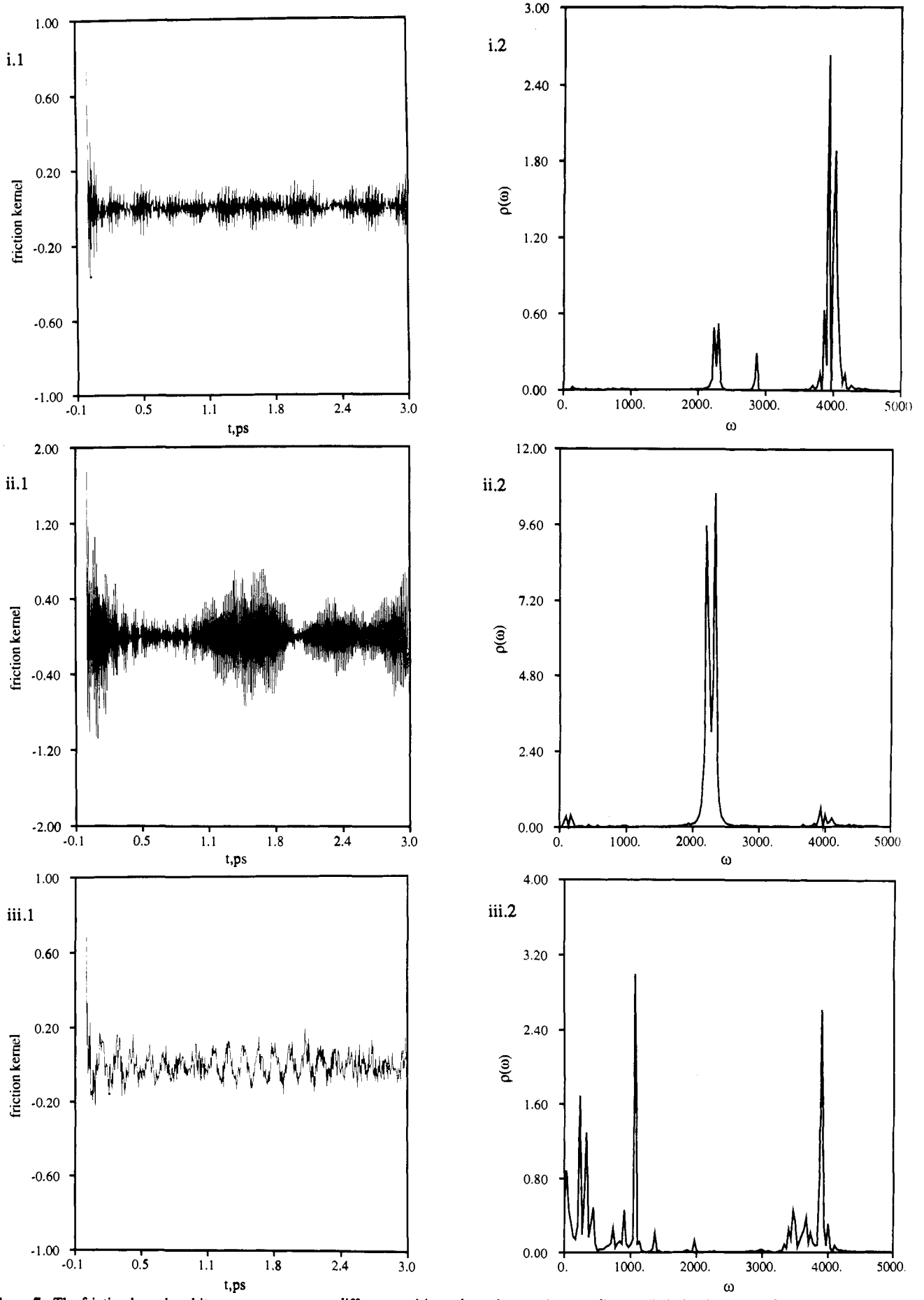

Figure 7. The friction kernel and its power spectrum at different positions along the reaction coordinate: (i.1) the time-dependent friction kernel, $\gamma(t)$, at a specific position along the reaction coordinate, q = 1.0 Å (this is the free energy minimum; (i.2) the power spectrum of the friction kernel at q = 1.0 Å; (ii.1) same as (i.1) except that q is 2.07 Å, the position of the free energy barrier; (ii.2) the power spectrum of (ii.1); (iii.1) same as (i.1) except that now q is 3.45 Å, a position within the free energy plateau; (iii.2) the power spectrum of (iii.1).

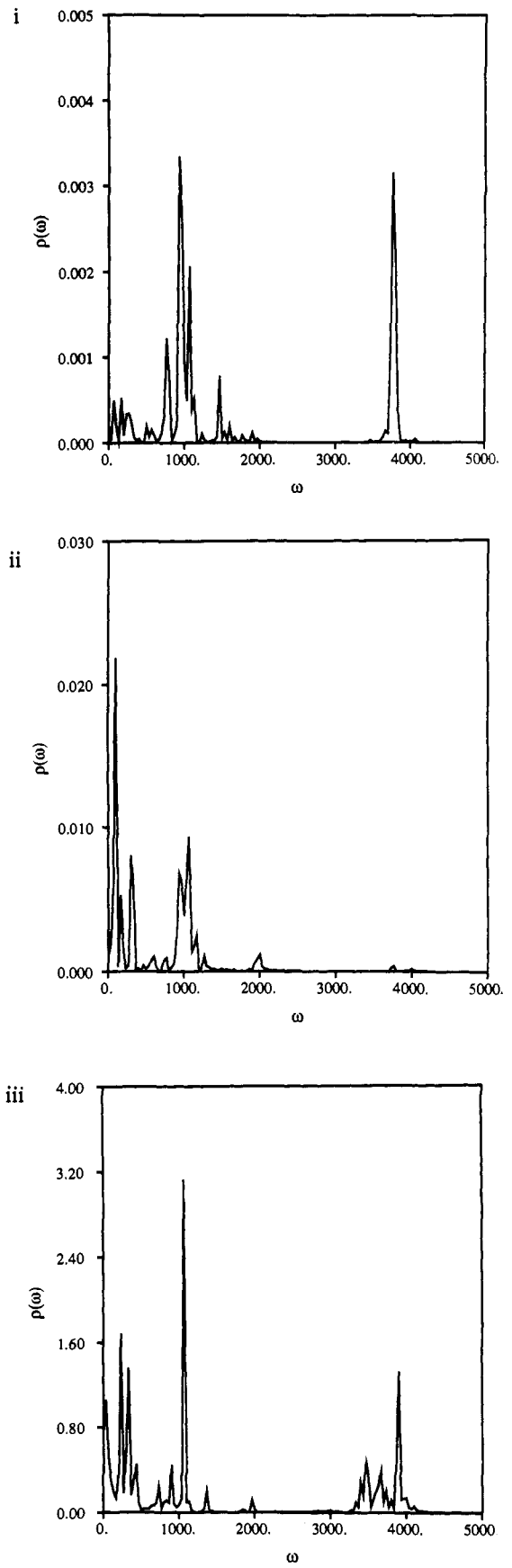

Figure 8. The power spectrum of the friction kernel after excluding the contribution of the carbon monoxide and the internal water molecules: (i) the free energy minimum (q = 1.0 Å), (ii) the free energy barrier (q = 2.07 Å), and (iii) the free energy plateau (q = 3.45 Å). Note that in (i) and (ii) the spectra were scaled by factors of 600 and 400, respectively, compared to Figure 7.

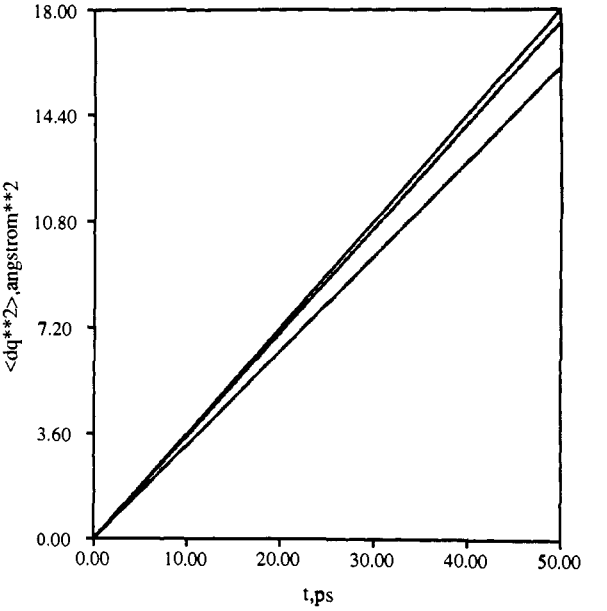

Figure 9. The diffusion constant calculated from three different 50-ps trajectories using the displacement fluctuations. The calculations were initiated at 3.12 Å, 3.25 Å, and 3.45 Å. Note that the displacements are dominated by *protein* atoms and not by the ligand.

We set $\gamma(\omega)$ to zero for all frequencies higher than 1500 cm^{-1} and transformed the result back to the time domain. The resulting function $\gamma(t)$ does not contain the contribution of the high-frequency vibrations (which we cannot estimate accurately anyhow) and makes it possible to obtain static friction with a better numerical accuracy. The result for Γ at position iii was 20 ps^{-1}.

When the result for Γ at position iii (using the spectrum with the frequency cutoff) is plugged in the formula for D (eq 17), a numerical estimate for the diffusion constant D is obtained (D = 1.2 $Å^2$/ps). This value is very large and reflects the low value obtained for the friction Γ. It also differs considerably from the value of the diffusion constant that we calculated in an alternative and more direct way (see below). The lesson from this part of the study is that, if the diffusion includes vibrational degrees of freedom, then it is hard to obtain accurate values of D from the integration of the correlation function.

In our second attempt to calculate the diffusion constant, we computed D more directly from the Einstein relation (eq 18). In Figure 9 we show the average of the fluctuations of $q - \langle (q(t) - q(0))^2 \rangle$ as a function of time. The average was performed over the position of $q(0)$ in three separate trajectories of the plateau regime. The plot of the fluctuations (Figure 9) is remarkably linear with time which suggests that the calculations of the fluctuations are a precise method for determining D. Furthermore, the value of the slope remained practically unchanged when we calculated the diffusion constant for each of the three trajectories: D = 0.16–0.18 $Å^2$/ps. This value for the diffusion constant is an order of magnitude smaller than the estimate based on the static friction. The numerical difficulties and the approximations associated with the calculation of Γ led us to consider the static friction method as unreliable and to accept the lower number obtained from the Einstein relation.

The diffusion process includes large-scale motions of the protein, and therefore the solvent effect may be important. In order to examine possible solvation effects on the diffusion constant, we repeated the calculation of D at one position (q = 3.45 Å) using a solvation shell of 768 TIP3P water molecules. This calculation yielded a diffusion constant of 0.10 $Å^2$/ps.

The effective mass as a function of the reaction coordinate is shown on Figure 10. The effective mass is calculated as defined by eq 7b. According to that equation, it is not possible to have an effective mass which is larger than the weight of the heaviest atom contributing to the reaction coordinate. Even for pure translation of the carbon monoxide molecule, the conjugated mass

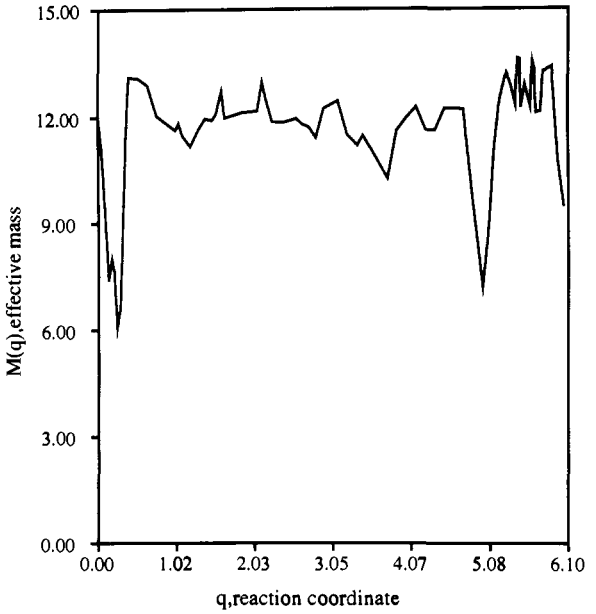

Figure 10. The effective mass as a function of the reaction coordinate. Note the two "dips" in the value of the effective mass which are associated with side-chain flips that include hydrogens (Ser 51 and Asn 19).

(according to our definition of the metric of q) would be the *average* mass of the carbon and the oxygen.

Note the two "dips" in the curve of the mass as a function of the reaction coordinate at $q = 0.23$ Å and at $q = 5.00$ Å. The mass is approximately a constant for most q's excluding these two places. The sharp reduction in the effective mass is a result of side chain transitions (Ser 51 and Asn 19) that occur at the corresponding q's. The hydrogens of these side chains were displaced significantly which results in a large contribution of these hydrogens to m_q.

We consider this observation an example for the global character of the reaction coordinate that includes a large number of protein atoms.

IV. Discussion

The aim of the present work was to understand the diffusion of a small ligand *in the protein interior*. The basic issue addressed in this paper is the time typically required for a diatomic ligand to reach the surface of the protein. According to our previous investigations,[12,13] the ligand route through the leghemoglobin matrix is spatially constrained but quite long. The diffusion is pursued along one dimension for ~6 Å. The identification of the diffusion as one dimensional is obviously an approximation, but it is supported by visual inspection of the trajectories in ref 12 and by the analysis of the different reaction coordinates of ref 13. This situation is in contrast with another (similar) protein, myoglobin, for which computational studies[9-11] demonstrate the availability of alternative pathways. Another important difference between leghemoglobins and myoglobins is in the rate. It is known experimentally that diffusion (and binding) in leghemoglobins is considerably faster than in myoglobins,[14,15] and in the Appendix it is shown that the diffusion in the protein investigated here, lupine leghemoglobin, is especially fast.

Myoglobins have been studied theoretically in detail in the past. Here we focus on lupine leghemoglobin and investigate how the spatially long diffusion coordinate can be translated to fast diffusion kinetics. The translation of structural data on a protein into internal diffusion kinetics is far from trivial. Straightforward calculations (such as the one pursued in this paper) are still of considerable difficulty. An actual derivation of a rate constant has twice been attempted (diffusion of oxygen in myoglobin: Case and McCammon[7] and Kottalam and Case[8]), and a phenomenological treatment (for myoglobin) has also been published.[3-6,30,31] However, it is hard to extract from the phenomenological models detailed atomic information that can be tested by specific chemical modifications or by genetic engineering techniques. Furthermore these models differ considerably from each other (e.g., refs 3, 4, 6, 31).

We propose here (after some labor) a model for ligand diffusion through leghemoglobin based on an atomic detail model. In an earlier work we pursued a search for plausible diffusion pathways[12] and then refined the resulting path to a reaction coordinate.[13] Reaction coordinates are not a simple quantity to test in experiment (though qualitatively they can be tested today using genetic engineering techniques). Therefore, in the present study we further analyzed our data and suggested a stochastic model that consists of two sequential steps: firstly an activated process and secondly a diffusive part. The detailed analysis and the proposed model are based on the following calculations: (i) calculation of the potential of mean force, (ii) evaluation of the transition-state rate constant and of the transmission coefficient, (iii) computation of the friction kernel, and (iv) calculation of the diffusion constant in the flat portion of the potential.

The most striking observation of the present investigation is the small variation of the mean force potential along the reaction coordinate. The variations are bound by less than 3 kcal/mol. This is in contrast to what is suggested by rigid protein structure and even by the calculated minimum energy path.[13] Furthermore, excluding the initial free energy minimum in the back of the heme pocket, we were not able to identify (within the accuracy of the calculations) any other intermediate binding site. The diffusion is spatially constrained to one dimension, but it is not constrained to a neighborhood of a point.

Consider the first activated process. The error bars for the potential of mean force are significant and the barrier is small (below 3 kcal/mol). Nevertheless, as we argued in the Results section, the converged transmission coefficient suggests that this barrier is meaningful. The use of the transition-state theory is therefore justified. Pictorially we expect the dissociated ligand to "sit" in the back of the heme pocket for the period between a few tens to a few hundreds of picoseconds (a few nanoseconds would not be a great surprise to us considering the limited accuracy of the calculations). On this time scale we expect significant direct interaction of the ligand with the heme iron and substantial (first-order) rebinding. The calculated time scale is significantly shorter than what is currently known for myoglobin, and is consistent with the experimental data on the two proteins which suggest considerably faster diffusion in leghemoglobin than in myoglobin.

Our studies of the friction kernel in this domain suggest coupling of the reaction coordinate only to high-frequency local modes (Figure 7); therefore, we predict low sensitivity of the activated step to "external" conditions such as viscosity. The ligand "sees" only its closest neighbors. We also propose that modifying residues in the proximity of the binding site (e.g., Phe 29; see ref 13 for more detailed structural information) may drastically alter the subnanosecond kinetics.

So far we have considered the progress of the ligand up to $q = 2$ Å; from now on we discuss its motion along the reaction coordinate from $q = 2$ Å to $q = 6$ Å. The rebinding kinetics is expected to change markedly once the ligand enters the second phase of escape from the protein.

It is important to emphasize that the second step of the diffusion is not the motion of the carbon monoxide ligand alone but includes a significant number of protein atoms. In the flat portion of the mean force potential a considerable fraction of the motion is the diffusion of the *protein* to an open state. Confirmation to this came from graphical studies of the reaction coordinate and also from the study of e_q, the slope of the reaction coordinate at q. In Figure 11 we show the amplitude of the atomic displacements at e_q as a function of the atomic number. At the extended part of

(30) Austin, R. H.; Beeson, K. W.; Eisenstein, L.; Frauenfelder, H.; Gunsalus, I. C. *Biochemistry* **1975**, *14*, 5355.

(31) For a recent discussion on a model for the reaction coordinate in myoglobin, see: Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598.

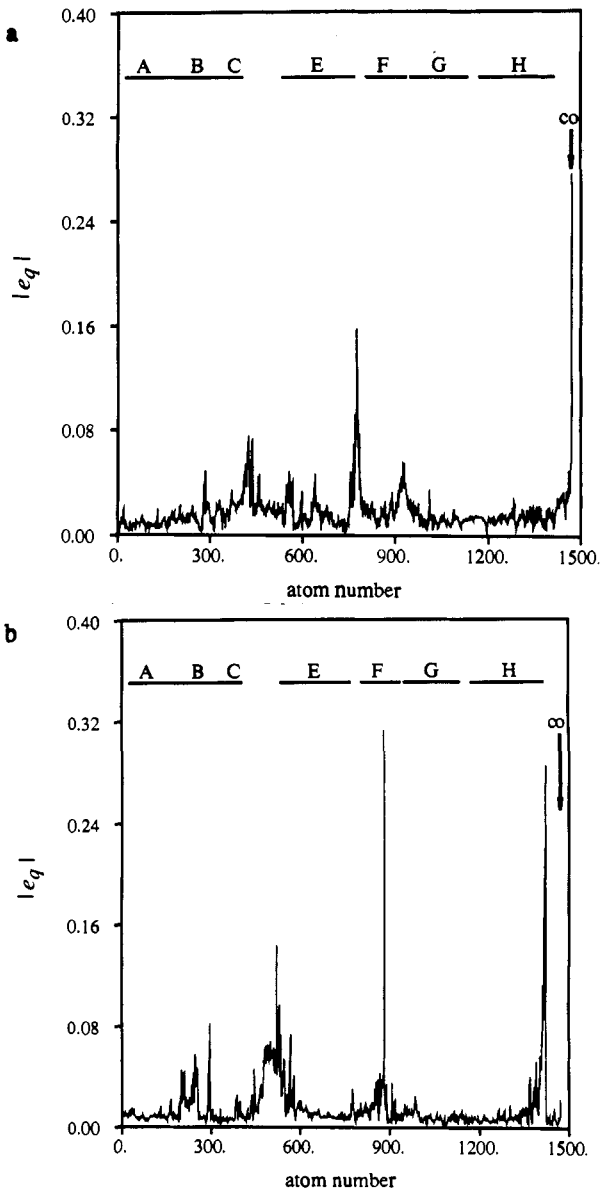

Figure 11. The atomic displacements along the reaction coordinate at specific q's as a function of the atomic number. A–H denotes the protein helices and the arrow indicates the position of the carbon monoxide: (a) the top of the free energy barrier; (b) a point on the free energy plateau.

the potential of mean force, the reaction coordinate has only a small component of the ligand and is dominated by protein motions. This is in contrast to the same type of plot for the top of the barrier (Figure 11b) in which the ligand (and the internal water molecule) plays an important role. Furthermore, analysis of the friction kernel demonstrates that the reaction coordinate is primarily coupled to the protein and not to the CO in the flat portion of the potential of mean force. The removal of the CO and the water contributions to the force fluctuations does not affect the power spectrum $\tilde{\gamma}(\omega)$ (Figure 8).

The protein motions that play an important role in the diffusion process have been discussed qualitatively in our previous studies. In ref 12 approximate mean field trajectories were used to study the diffusion dynamics. There the gate opening was assigned to global motions of the C and the G helices. As the relative distance between the two helices increases by ~1.5 Å, a hole is opened in the interface between the helices and the ligand escapes freely through that hole. In ref 13 the minimum energy path was calculated and a second barrier identified as the global motion of (again) the C and G helices. Here the shift of two helices is shown to be diffusive and not activated (see Figure 9 for the determination of the diffusion constant) with a diffusion constant only slightly different from that of a small molecule (like oxygen) in water (for oxygen in water: $D = 0.17$ Å^2/ps). This is, however, a numerical coincidence since the diffusion examined here is for a significant fraction of the protein molecule and not for a diatomic molecule. The length of the diffusion path is 4 Å after the "activated domain". If the extended protein is treated as a random walk, the typical time for the ligand to reach the protein surface and for the helices to diffuse to an open position is $t \sim q^2_{surface}/D$. Thus $t = 4^2/(0.17) \sim 100$ ps. More sophisticated diffusion models such as solution of the diffusion equation with appropriate boundary conditions were not applied, since the surface condition is not known. We do not have an estimate for the rate of ligand transfer from the protein to the solvent. This issue will be studied in a future work.

Our estimate for the diffusion time has been obtained for the "zero" viscosity limit. In a viscous fluid the low-frequency modes describing the relative fluctuations of the C and the G helices may be different. To explore some of the solvation effects, we repeated the calculations of the diffusion constant for a protein in a solvation shell (see Results section). The diffusion constant for the solvated protein was smaller, 0.1 Å^2/ps; however, the difference was less than a factor of 2. It does not appear likely that increasing the solvation shell will decrease the value of the diffusion constant further. In fact, surface tension effects (in the water solvation shell) may reduce the global motions of the protein as compared to bulk water. Significant differences between the values obtained for D and the value in bulk water are therefore not expected.

For this diffusive non-activated process we therefore make the following predictions. (i) Rebinding still obeys first-order kinetics. (ii) Kinetics of rebinding that occurs in this time domain (to be on the safe side we should consider times longer than a few nanoseconds) is expected to be non-exponential and to obey a power law, a simple result for diffusion processes. (ii) Because of the global nature of the motions involved and the coupling to low-frequency extended modes, this part of the diffusion process is expected to depend on external parameters, such as viscosity, a prediction subject to experimental test.

These predictions for the activated and diffusive process also suggest an experimental way to distinguish between the two different domains of the kinetics. The first process is not expected to depend strongly on the viscosity while the second is.

The model emerging from the above discussion for ligand diffusion, in the *protein interior*, is summarized in the formula below:

$$\text{Fe–CO} \xrightarrow[\text{direct}]{} \text{heme–pocket}\cdots\text{CO}_{\text{localized state}} \xrightarrow[\text{activated}]{} (\cdots\text{G}\cdots\text{C\&CE}\cdots\text{CO})_{\text{diffusive state}}$$

We suggest that these "theoretical" states are related to the experimentally observed intermediates B and C. In the Appendix experimental results for the dissociation and the recombination rates of a diatomic ligand are expressed by a phenomenological rate equation. The computations are relevant to the step B → C (k_3) of the kinetic equation. B and C are phenomenological ligand states that describe a ligand still trapped in the protein matrix. The rate constant extracted from the recombination data (see Figure 12D) is 4.0 ns^{-1}. Considering the error bars of the calculations and the difficulties in the experiment (see Appendix), the calculated rate agrees quite well with the experimental rate.

After dissociation the ligand is released from the heme iron to the nearby free energy minimum in a direct barrierless process (at room temperature). From the free energy minimum the ligand needs to hop over a barrier that is rather small (one to a few kcal/mol) in order to reach the flat portion of the free energy state and the diffusive state. The last "state" is not spatially localized but rather smeared along the diffusion path in the protein matrix. It involves the diffusion of a large number of atoms and it cannot be interpreted as the diffusion of the ligand only.

The idea of a diffusive protein coordinate that is coupled to the ligand motion is not new and was proposed in a number of phenomenological models (for myoglobin). Agmon and Hopfield,[3] in their phenomenological model for ligand recombination in myoglobin, proposed an "abstract" diffusion coordinate with parameters chosen to fit the experimental data of Austin et al.[30] The

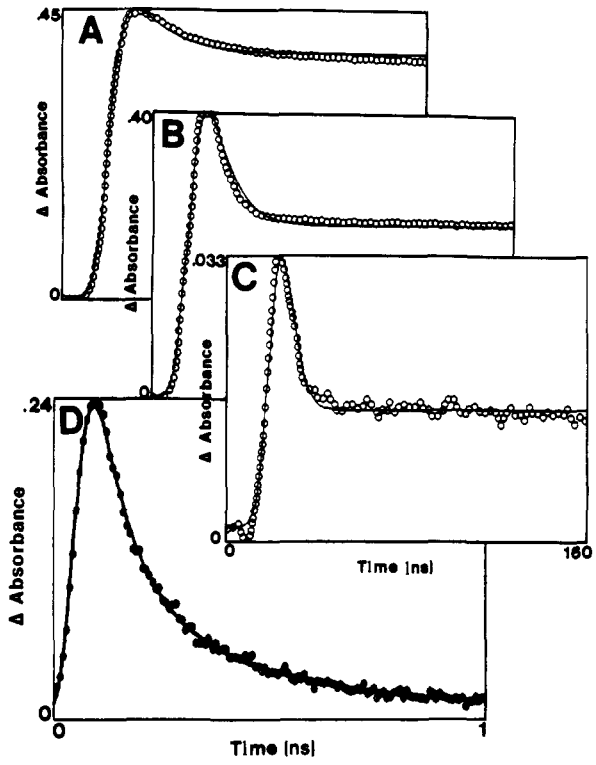

Figure 12. The points represent experimental data, the continuous line is derived from the scheme given in the Appendix, and the values of the parameters refer to it. The values for k_3, k_4, and k_5 were the same for all ligands and were 4.0, 3.5, and 0.08 ns^{-1}, respectively; temperature 10 °C; excursion for 100% photolysis 0.81. Panel A: 1 atm of CO, light 1/16 (light/1 peak rate of work for k_1 3.6 ns^{-1}); total time 160 ns; maximum absorbance excursion 0.45 at 436 nm; estimated value for k_2 0.038 ns^{-1}. Panel B: air equilibrated, light/1; max absorbance excursion 0.402; estimated value of k_2 0.13 ns^{-1}. Panel C: 5% NO gas, light/1; max absorbance excursion 0.033; estimated value of k_5 0.08 ns^{-1}. Panel D: 5% NO gas, picosecond experiment, total time 1 ns; estimated value of k_2 30 ns^{-1}.

same experiment also found significant dependence of the rate on viscosity. It is also intriguing that another phenomenological model (Miers et al.[6]) interprets ligand recombination data (in myoglobin) as a sequence of an activated and a diffusive process. This is the same picture that seems to emerge from the present study for leghemoglobin. Using the atomic detail model, we identify the previously proposed reaction coordinate[13] as partially diffusive.

We devote the last paragraph of the Discussion to possible sources of errors in our simulations. One source of errors is statistical. We invested considerable effort to ensure the statistical convergence of our calculations. The calculations covered more than 4 ns of simulation time and for this size of a protein were quite extensive. However, the error bars in the calculations of the potential of mean force are still significant.

Another source for *systematic* errors is the partial neglect of the solvent. We focused on the motions in the interior of the protein, which are expected to be influenced less by the lack of solvent compared to motions close to the surface. The local nature of the first barrier suggests that this is a sound approximation for this part of the reaction coordinate. This approximation is, however, more questionable for the second part in which global motions of the protein were detected. This is also why we repeated the calculation of the diffusion constant using a protein covered with a solvation shell. The "not too different" values for the diffusion constant in the solvated protein and in the protein in vacuo argue for accepting that approximation. Nevertheless, we did not calculate the potential of mean force with the solvation shell. It is possible that some of the features of the flat portion of the free energy surface would be modified if the solvent were included explicitly. This calculation could not be pursued due to computer resources limitation. We emphasize the global nature of the protein motions that contribute to the second part of the process. Their extensive character makes it difficult to "solvate" them extensively in feasible simulations.

The present study was also limited to the "protein" phase. We did not include in our calculations possible barriers for the transfer of the ligand from the protein to the liquid phase. This issue will be addressed in a future work. Clearly the fact that the process studied is only a part of the complete recombination process that is measured experimentally[14,15] (see also Appendix) makes the comparison with the experiment more difficult.

Summary

We constructed a stochastic model for ligand diffusion in leghemoglobin based on an atomic detail model. The stochastic model consists of two parts: (a) an activated process (with a barrier of approximately 3 kcal/mol) in which the ligand leaves the back of the heme pocket to the contact between B, C and G helices, and (b) a diffusive process in which the *protein* helices C and G diffuse to an open conformation while the ligand moves only a little. We employed the transition-state theory to study the first process and estimated a diffusion constant for the second. In the Appendix experimental measurements for ligand recombination kinetics are given. The theoretical and experimental time scales for the diffusion process are in a reasonable agreement.

Acknowledgment. This research was supported by NIH Grant No. GM40698. The calculations were done on a Stardent 3030 minisupercomputer purchased by an NIH equipment grant, No. RR04884. R.E. is a Camille and Henry Dreyfus New Faculty and University of Illinois West Scholar. R.E. thanks Professor John Straub for useful discussions on the friction kernel.

Appendix: Experimental Measurements of Rebinding Kinetics: Recombination of Diatomic Ligands (CO, NO, and O_2) to Lupine Leghemoglobin

(i) Materials and Methods. Lupine leghemoglobin type I was the generous gift of Dr. Cyril Appleby, Moruya, Australia, and was prepared by him as described in ref 32. Experiments were performed using 0.1 M KP_i buffer pH 7.0.

Flash photolysis experiments were performed as described in ref 33. Picosecond data were obtained in a pulse–probe experiment, with a photolysis pulse at 532 nm and a probe at 436 nm from a Raman shifter using H_2 gas. Nanosecond data were collected using a 9-ns photolysis pulse at 532 nm and a pulsed Xe arc and monochromator for observation.

(ii) Ligand Rebinding to Lupine Leghemoglobin in Short Times. Although data for second-order binding of O_2, CO, and NO have been reported in ref 15, no results for the geminate reactions are available. Accordingly, experiments have been performed with these ligands for comparison with the results of the theoretical calculations. To permit a compact approximate description of the results, diffusion of the ligands away from the iron after photolysis is represented by consecutive first-order reactions. The three ligands differ widely in their rebinding behavior, and, with the additional assumption that their rates of diffusion within the protein are likely to be similar, two distinct steps between photodissociation and escape from the protein may be identified.

Nitric oxide is the only ligand to show extensive ultra-rapid rebinding and, in a probe–pulse experiment, over 90% of the NO which is free at the end of the 35-ps pulse recombines at a rate of 30 ns^{-1}. The remainder combines much less rapidly, at a rate which is poorly determined because of its small amplitude and because the phase is not complete at the longest time of observation (1 ns). It is, however, about one tenth the rate of the fast phase. Longer term experiments performed with a 9-ns flash, and recorded with a 1-ns time constant, show only one relaxation up to 500 ns with a rate of 1.7×10^8 s^{-1}. The amplitude of the absorbance excursion is small, accounting for only 3% of the

(32) Bogusz, D.; Kortt, A. A.; Appleby, C. A. *Arch. Biochem. Biophys.* **1987**, *254*, 263.
(33) Carver, T. E.; Rohlfs, R. J.; Olson, J. S.; Gibson, Q. H.; Blackmore, R. S.; Springer, B. A.; Sligar, S. G. *J. Biol. Chem.* **1990**, *265*, 20007.

possible total (derived from static spectrophotometry) when the maximum light intensity from the laser is used. For comparison, some 10% of carbon monoxide was dissociated under the same conditions with the flash attenuated 256-fold. The observed nanosecond relaxation accounts for only half of the total absorbance change, and no slower recombination is seen. These results are illustrated in Figure 12 which includes a calculated time course of rebinding using the sequential scheme

$$M \rightleftharpoons A \underset{k_2}{\overset{k_1}{\rightleftharpoons}} B \underset{k_4}{\overset{k_3}{\rightleftharpoons}} C \underset{(k_6)}{\overset{k_5}{\rightleftharpoons}} D$$

where k_1 represents the photochemical breakdown of A, the fully liganded species. B and C are species in which ligand is not combined with the heme but remains associated with the protein, and D in unliganded leghemoglobin. The species M has been described in ref 34. It has a lifetime of 3 ps and shows absorbance at 450 nm. In the present experiments its only effect is to influence the apparent quantum yield. For lupine NO 0.85 of the quanta leads to M, for oxygen 0.95, and for CO 0.05. In applying the scheme, k_2 was set at 30 ns^{-1}, and nonlinear least-squares calculations were used to assign the remaining parameters. These are highly correlated, and only a range can be given. The ratios are better determined than the individual values. The least value which can be given to k_3 is 2 to 3 ns^{-1}. This allows sufficient NO to escape from the vicinity of the iron to represent the slower component in the picosecond experiment, identified with the sequence C → B → A. In the nanosecond experiment, values of 4 ns^{-1} for k_4 and 0.08 ns^{-1} for k_5 reproduce both the rate and the amplitude of the relaxation (Figure 12C).

No significant picosecond relaxation was observed with oxygen or carbon monoxide. If diffusion of the three gases is assumed to be similar, i.e., k_3, k_4, and k_5 are given the same values as for NO, the value of k_2 for oxygen is 0.09 ns^{-1}, and for carbon monoxide it is 0.04 ns^{-1}. The results of these experiments are shown in Figure 12, A and B, for CO and O_2, respectively. Values for the standard errors of the parameters k_3, k_4, and k_5 cannot be readily given because fixed values were assigned to other parameters in the scheme, as described in the text, to carry the calculations to longer times. The uncertainty in k_3 may be as much as 50% of itself, with a lower limit, judged by increasingly nonrandom distribution of residuals, of perhaps 3 ns^{-1}. This corresponds to uncertainty of 30% in k_4, but the ratio k_3/k_4 is much better defined, and is probably correct to within 10%. With the values given, k_2 and k_5 are defined to within 10% for each of the ligands. The results are consistent with the idea that diffusion of all three ligands is similar, and the differences in their geminate behavior may be accounted for by changes in k_2 alone. In spite of these uncertainties, the results with lupine leghemoglobin are clearer than earlier work with sperm whale myoglobin,[33] where geminate recombination of NO accounts for too large a part of the total reaction, and CO for too small a part, to allow estimates of the diffusion parameters even with the modest precision reached here.

The key difference experimentally is in the nanosecond geminate region where with the lupine leghemoglobin (panel C, Figure 12) only 50% of NO recombines, whereas in analogous experiments with sperm whale myoglobin, 85% does so at a rate about half of that (0.035 ns^{-1}) for lupine. The implication appears to be that diffusion out of the protein is some five times faster in leghemoglobin than in sperm whale myoglobin, a conclusion that does not depend on detailed analysis of the reactions. It is this difference in the diffusion rate for NO which is responsible for the 6- to 10-fold difference in the overall rates of NO binding in lupine and sperm whale myoglobin.[15,33]

Comparing the theory and the experiment, we consider the very small nanosecond recombination kinetics for NO in leghemoglobin. Thus, the ligand escapes from the protein matrix at times comparable to the time of the nanosecond flash. This is observed directly from the experimental curves (Figure 12) *independently* of the kinetic model that is used to interpret the results. For this observable the agreement between theory and experiment is good. The detailed description of the subnanosecond dynamics is somewhat different. This is since the theory predicts one activated process and one diffusive process, while the experiment employs a model of two activated processes. The uncertainties in the experimental data make it difficult to distinguish between the two, and the more commonly employed scheme of two exponentials was employed. This, however, makes the quantitative comparison between the experiment and the theory (especially k_4) difficult to pursue, and more experiments will be required to address the detailed dynamics of the subnanosecond domain.

In summary, bimolecular rebinding of NO is limited by a diffusion of NO to the heme, while O_2 and CO binding rates are determined by the probability of reaction once a molecule has diffused there.

(34) Petrich, J. W.; Poyart, C.; Martin, J. L. *Biochemistry* **1988**, *27*, 4049; *J. Biol. Chem.* **1986**, *261*, 10228.

Registry No. CO, 630-08-0; NO, 10102-43-9; O_2, 7782-44-7; heme, 14875-96-8.