and HCl^-,[25] but its introduction in the present problem would complexify the procedure since the diabatization would act on more than two eigenstates of the CI problem.

The use of diabatic descriptions may be especially tempting when basis set and CI limitations introduce serious deviations from experiment. Due to the expected constant physical content of the diabatic function, a proper fitting on the asymptotes becomes possible, which would be impossible in adiabatic approaches. This has been done frequently on diatomic problems. Here the diabatic surface matches two different (although related) asymptotes in the entrance and product channel, and a double fitting is necessary on both asymptotes. The LEPS fit of the diabatic surface is especially convenient to proceed to such adjustments on experiment.

Acknowledgment. Thanks are due to Dr. F. X. Gadêa for helpful discussions. One of us (O.K.K.) acknowledges the support of the Action Intégrée Franco-Marocaine.

Registry No. Li_2, 14452-59-6; Li^-, 14808-04-9; H^-, 12184-88-2; H, 12385-13-6; Li_2^-, 11062-41-2; Cl, 16887-00-6; Cl_2, 7782-50-5; CH_3Cl, 74-87-3.

(25) Rajzmann, M.; Spiegelmann, F.; Malrieu, J. P. *J. Chem. Phys.* **1988**, *89*, 433.

Reaction Path Study of Ligand Diffusion in Proteins: Application of the Self Penalty Walk (SPW) Method To Calculate Reaction Coordinates for the Motion of CO through Leghemoglobin

Wieslaw Nowak,[†] Ryszard Czerminski,[‡] and Ron Elber*

Contribution from the Department of Chemistry, M/C 111, University of Illinois at Chicago, P.O. Box 4348, Chicago, Illinois 60680. *Received December 7, 1990*

Abstract: Reaction coordinates for the diffusion of carbon monoxide through leghemoglobin are calculated by using the recently developed SPW algorithm (Self Penalty Walk).[1] The new algorithm makes it possible to study in a systematic way reaction coordinates in molecules with more than 1000 atoms. To explore properties of similar (but distinct) reaction coordinates, three diffusion paths were calculated. The separate coordinates were generated from different initial guesses for the paths, which were obtained from classical trajectories.[2] Analysis of the three calculated paths reveals that the "local" properties of the coordinates in the vicinity of the CO are very similar. The interaction energy profile of the carbon monoxide with the rest of the protein has a similar shape in the three paths, and the structural features of the local transition states are essentially the same. On the other hand, global protein properties vary considerably in the three paths. The macromolecule motions include many fluctuations that are not coupled to the diffusing ligand. It is concluded that while the density of alternative paths for the diffusion process may be very high, the close neighborhood of the ligand appears to be very much alike in the sampled paths. The diffusion process consists of two steps. In the first, the ligand hops from the heme pocket to another cavity in the protein matrix, and in the second it hops to the protein exterior. The first barrier is dominated by a tilt of a single residue (Phe 29 B9). The second barrier is more complex and includes many types of motions. In particular, global translations and rotations of helices C and G are involved.

I. Introduction

The reaction coordinate (RC) approach has been shown to be useful in the study of a variety of chemical processes of small molecules (<10 atoms).[3] The RC is helpful in estimating barrier heights for the reaction, in obtaining structural insight to the properties of the transition state, and finally, in quantitative calculations of the rate.

The RC is usually defined as the steepest descent path (SDP).[3] This definition guarantees a continuous description of the motion between the reactant and the product, with a low energy barrier.

Though successful for small molecules, tools for a few atoms are not necessarily adequate for studying RC in macromolecules (>1000 atoms). There are two serious difficulties in attempting to extrapolate from systems with several atoms to systems with thousands of atoms. The first obstacle is computational: It is far from easy to calculate reaction paths in large systems, and only recently appropriate tools were developed. Moreover, in large systems it is not clear if the concept of a *single* reaction coordinate is valid and if the SDP is the best description of the reaction coordinate.

Obviously in order to study properties of paths in macromolecules, one needs to calculate them first. We initiated a program to develop theoretical tools to study reaction paths in large molecular systems. These tools were tested on small molecules[1] (~20 atoms), and we apply them here to a macromolecule (1471 atoms). To the best of our knowledge, this is the first application of an automatic algorithm to calculate and compare different reaction coordinates in a system of comparable size. Here we address the problem of the diffusion of a small ligand through a protein matrix.

Diffusion of a small ligand from active sites buried in protein matrices attracted considerable theoretical and experimental attention over the years. Of special interest were the investigations of the diffusion of a small ligand (e.g., oxygen, carbon monoxide) in myoglobin, for which a wealth of data is currently available.[4]

Here we consider the diffusion in a "myoglobin variant"—lupine leghemoglobin. Leghemoglobin has the same evolutionary ancestor as myoglobin but binds oxygen significantly faster and more strongly.[5] Currently there is no clear structural or dynamical

[†] On leave from the Institute of Physics, N. Copernicus University, ul. Grudziadzka 5, PL-87-100 Torun, Poland.
[‡] Present address: Polygen, 200 Fifth Ave., Waltham, MA 02254.

(1) Czerminski, R.; Elber, R. *Int. J. Quantum Chem.* **1990**, *24*, 167.
(2) Czerminski, R.; Elber, R. *Proteins* **1991**, *10*, 70.
(3) For reviews, see for instance: Muller, K. *Angew. Chem., Int. Ed. Engl.* **1990**, *1*, 19. Bell, S.; Crighton, J. S. *J. Chem. Phys.* **1990**, *80*, 2464.
(4) For a recent review, see: Brunori, M.; Coletta, M.; Ascenzi, P.; Bolognesi, M. *J. Mol. Struct.* **1989**, *42*, 175.

0002-7863/91/1513-5627$02.50/0

interpretation for the experimentally measured dissimilar binding properties of the two structurally related proteins.

In this paper, we focused on one of the problems, that of diffusion. Rebinding experiments[6] suggest that, in leghemoglobin, the ligand moves in and out of the binding pocket much faster than in myoglobin. Similarly to other globins, the small ligand (e.g., oxygen or carbon monoxide) is blocked at the binding site of the leghemoglobin matrix and is not exposed to the solvent. The ligand escape must be associated with appropriate protein fluctuations, which open the gate.

The aim of the present work is to calculate the reaction coordinates for carbon monoxide escape from fluctuating leghemoglobin. The calculations are based on a crude approximation for the reaction path generated by a special molecular dynamics protocol that was developed recently (Locally Enhanced Sampling, LES).[7] The crude approximation is refined at present to yield the corresponding reaction coordinates.

In the LES approach, the trajectories of a large number of ligands are run in parallel in a single protein matrix, exploring alternative diffusion routes. The trajectories of the ligands that escape from the protein matrix suggest the gates for ligand escape. Some of the LES predictions can be tested by site-directed mutagenesis experiments (such experiments were pursued for myoglobin[8]); i.e., site-directed mutagenesis can be employed to block or to open suggested diffusion routes. However, in order to estimate barriers, rates, and time scales, more calculations (in addition to LES) are required. The LES trajectories are approximate, and it is difficult to extract from them time scales to be compared to kinetic measurements.

Useful techniques that can be based on the data acquired in the LES study are the reaction path approach and the transition-state theory.[9] Application of these techniques enables the estimation of rates and time scales of various processes. The first step in applying the theory is the identification of a reaction coordinate. In this paper, reaction coordinates for the escaping ligand are calculated by using as a starting point the previously calculated LES trajectories. We comment that reaction coordinates provide an estimate for the energy barrier, which is hard to obtain from trajectories. Furthermore, in trajectories, many irrelevant motions that do not influence the reaction coordinate occur. A significant fraction of these motions is quenched in the path calculations, which simplifies the analysis of the mechanism. After the coordinates are identified, it should be possible to estimate the potential of mean force along these paths and then to calculate the rate by using statistical theories. The last step will be pursued in future work.

There are a number of "philosophical" questions associated with the assignment of a reaction coordinate that need to be reexamined in large systems: (a) Is the reaction coordinate in a large polyatomic system unique? (b) If there is an ensemble of reaction coordinates, what are the properties of this ensemble? (c) What can we learn from a partial set of this ensemble or from a single coordinate? These questions will be addressed in the context of the specific problem studied—the diffusion of a ligand in leghemoglobin. We do not give complete and conclusive answers to points a–c. We provide however some data and a discussion based on our (limited) sample of paths.

Statistical properties of transition states (saddle points) were examined in the past on significantly smaller systems. Berry et al.[10] studied saddle points in rare gas clusters. Harris and Stillinger investigated the distribution of saddle points for a chemical reaction in liquid argon.[11] Tanaka and Ohmine studied the transitions between inherent water structures.[12] Nguyen and Case studied conformational transition in a dipeptide.[13] Czerminski and Elber[14] and Choi and Elber[15] examined the reaction coordinate network for conformational transitions in tetrapeptides. The protein of the present problem is much larger, and reaction coordinates for leghemoglobin are significantly more challenging to calculate and to analyze. Furthermore, leghemoglobin has a well-defined heterogeneous structure. The inhomogeneity adds to the complexity of the observed paths and to the difficulty of calculation. The computational aspects of calculating paths in large systems will be examined in detail in this paper.

This text is organized as follows. In the next section (II), we discuss the method used to calculate reaction coordinates.[1] In III, we describe the computational protocols. In IV, the results for the three different paths are presented. Discussion is in V, and conclusions are given in VI.

II. Method

We usually associate the reaction coordinate with a line (curvilinear coordinate) connecting the reactant and the product with a minimal energy barrier, i.e., with a minimum energy path. The steepest descent path (SDP) satisfies this criterion, and the protocol described below (SPW, Self Penalty Walk) satisfies it too, under appropriate choice of computational parameters. This section is devoted to the description of the properties of the SPW.

SPW was introduced recently by Czerminski and Elber[1] to calculate reaction coordinates in large polyatomic systems. SPW was tested on a conformational transition in a dipeptide and in a tetrapeptide. The SPW is a modification of an earlier algorithm (Gaussian chain) by Elber and Karplus.[16] The SPW was shown to be approximately 10 times faster than the Gaussian chain and also (in contrast to the Gaussian chain) to have a simple physical realization. Under certain limiting conditions, the SPW is reduced to a classical trajectory.[1]

A number of alternative approaches to calculate saddle points and reaction paths in large molecular systems was discussed by different people.[1,10,11,16,17] However, to the best of our knowledge, the algorithm of Elber and Karplus (for which SPW is a generalization and an improvement) was the only technique applied to a system with more than 1000 atoms (a side chain flip in myoglobin[16]). In the present paper, we applied the SPW approach to a diffusion problem of biological interest—the escape of a ligand from a buried binding site. The problems of ref 16 and the present study are of a similar size.

We developed recently two additional methods to calculate reaction paths.[14,15,18] Compared to other methods introduced in the researchers' group (constrained energy minimization[14] and locally updated planes[15,18]) the SPW is the most robust and requires the least amount of human intervention when applied to *large molecules*. On the other hand, SPW is relatively slow and the resulting coordinate (in contrast to locally updated planes[15,18]) is not the steepest descent path (the SDP is the "intrinsic reaction coordinate" that was commonly used for small molecular systems). Nevertheless, the SPW provides a reasonable approximation to the SDP as was demonstrated by Czerminski and Elber.[1] Furthermore, the energy barriers estimated by SDP or by SPW are very similar.

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, the leghemoglobin protein with the ligand). The minimum is a discrete representation of the reaction coordinate. Thus, if r_i is the Cartesian coordinate vector of the ith monomer, then the polymer coordinates are $\{r_i\}_{i=1}^{M}$. The two edges of the polymer are connected to fixed points—the "reactant" and the "product". An example for a reactant and a product is the protein with the ligand trapped in the buried cavity and with the ligand outside the protein matrix, respectively. The potential energy of the polymer is a sum of inter- and intramonomer energies. We denote by V_i the potential energy of the ith copy of the physical system, i.e., the internal energy of the ith monomer. The intermonomer interactions include nearest-neighbor

(5) Kellin, D.; Wang, Y. L. *Nature* **1945**, *155*, 227. Wittenberg, J. B.; Appleby, C. A.; Wittenberg, P. A. *J. Biol. Chem.* **1972**, *247*, 527.
(6) Stetzkowski, F.; Banerjee, R.; Marden, M. C.; Beece, D. K.; Bowne, S. F.; Doster, W.; Eisenstein, L.; Frauenfelder, H.; Reinisch, L.; Shyamsunder, E.; Jung, C. *J. Biol. Chem.* **1985**, *260*, 8803. Gibson, Q. H.; Wittenberg, J. B.; Wittenberg, B. A.; Bogusz, D.; Appleby, C. A. *J. Biol. Chem.* **1989**, *264*, 100.
(7) Elber, R.; Karplus, M. *J. Am. Chem. Soc.* **1990**, *112*, 9161.
(8) 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.
(9) Truhlar, D. G.; Garret, B. C. *Acc. Chem. Res.* **1980**, *133*, 440. Pechukas, P. *Annu. Rev. Phys. Chem.* **1981**, *32*, 159.
(10) Berry, R. S.; Davis, H. L.; Beck, T. L. *Chem. Phys. Lett.* **1988**, *147*, 13.
(11) Harris, J. G.; Stillinger, F. H. *Chem. Phys.* **1990**, *149*, 63.
(12) Tanaka, H.; Ohmine, I. *J. Chem. Phys.* **1989**, *91*, 6318.
(13) Nguyen, D. T.; Case, D. A. *J. Phys. Chem.* **1985**, *84*, 4020.
(14) Czerminski, R.; Elber, R. *J. Chem. Phys.* **1990**, *92*, 5580.
(15) Choi, C.; Elber, R. *J. Chem. Phys.* **1991**, *94*, 751.
(16) Elber, R.; Karplus, M. *Chem. Phys. Lett.* **1987**, *139*, 375.
(17) Pratt, L. R. *J. Chem. Phys.* **1986**, *85*, 5045.
(18) Ulitsky, A.; Elber, R. *J. Chem. Phys.* **1990**, *92*, 1510.

bondlike terms, which we denote by B_i, and repulsion terms between the pairs i,j ($j > i + 1$) denoted by $R_{i,j}$. The complete expression for the polymer energy is given below.

$$S = \sum_{i=1}^{M} V(\mathbf{r}_i) + \sum_{i=0}^{M} B_i + \sum_{\substack{i,j=0 \\ j>i+1}}^{M+1} R_{i,j} \qquad (1)$$

$$B_i = \gamma(d_{i,i+1} - \langle d \rangle)^2 \qquad R_{i,j} = \rho \exp\left(-\frac{d_{i,j}^2}{(\lambda\langle d \rangle)^2}\right)$$

$$d_{i,j} = [(\mathbf{r}_i - \mathbf{r}_j)^2]^{1/2} \qquad \langle d \rangle = \left(\frac{1}{M+1}\sum_{i=0}^{M} d_{i,i+1}^2\right)^{1/2}$$

The index 0 refers to the fixed reactant and the index $M + 1$ to the fixed product. The parameter γ is the "force constant" for the bond between the monomers which ensures that the distances between the monomers are approximately equal (but *not* fixed). ρ and λ determine the repulsion between the monomers. Their purpose is to obtain more significant sampling of the transition-state region by increasing the stiffness of the polymer. This avoids aggregation of the polymer in the neighborhood of the minima as sometimes was observed in the Elber and Karplus protocol.[16] In addition to the increase in computational efficiency, Czerminski and Elber derived a physical interpretation for the repulsion and S.[1] They demonstrated that the $R_{i,j}$ can be associated with the kinetic energy. To rigorously transform the repulsion to kinetic energy, a few modifications in S are required and the continuum limit ($M \to \infty$) should be considered. However, the qualitative picture (as demonstrated in the previous paper[1]) of the repulsion as "kinetic energy" remains valid even with the approximate expression above: In the limit of high polymer repulsion (stiffness), the monomers follow a straight line connecting the reactant and the product that resembles the "ballistic" motion limit. In contrast, for very low repulsion, the polymer spends most of its time in the minima and the transition resembles that of diffusive motion. The view of a minimum energy path taken in the present study is the path with minimum repulsion (kinetic energy) that is required to pass the barrier with no aggregation of the polymer in one of the energy minima along the path.

We further comment that since the calculations are pursued in Cartesian coordinates, it is necessary to fix the "rigid body" variables of the different monomers (i.e., the translation and the rotations of the individual monomers). In the Cartesian space, translations and rotations of the system copies would change the distances between them, a change that we should like to avoid. Czerminski and Elber introduced a gradient projection technique to fix this problem that speeds up the computations by a factor of approximately 10 compared to other approaches.[16] The gradient projection method is the one used in the present study.

The minimization of the target function S can be done by using different nonlinear optimization tools, such as conjugate gradient minimization or simulated annealing.[19] The details of the computational approach taken in the present investigation will be discussed in the next section.

III. Computational Protocol

In order to find the minimum of S, an initial guess for the set of monomer coordinates $\{\mathbf{r}_i\}_{i=1}^{M}$ is required. We extracted the initial set of coordinates from LES simulations. A technical description of LES and application to myoglobin was described in ref 7. The application of LES to ligand diffusion in leghemoglobin can be found in ref 2. Briefly LES provides approximate ligand trajectories out of the protein matrix. Approximate ligand trajectories that escape from protein matrices were generated in the past.[20-22]

The protocol to obtain the minimum energy paths (minimum of S) was separated to three major steps: (a) Select a set of structures from a trajectory of a ligand that escapes. (b) Minimize the selected structures along the diffusion route to determine local minima along the escape path. (c) Connect the local minima by minimum energy paths.

Three trajectories by which ligand escaped from the protein matrix were employed. They were all taken from the same LES run—run A of ref 2—which included 120 ligand copies. The intention was to make the comparison between the different paths easier. This is since the fluctuations between alternative paths are expected to be smaller for a single (protein) trajectory. More details on steps a–c are given below.

(a) Selection of Trajectories. The qualitative behavior of the ligand trajectories was essentially the same for the three paths. The trajectories consist of a long "incubation" time in the heme pocket followed by a sudden transition to another spot in the protein interior, which is in the proximity of the B/C bend and the G helix. The final jump was to the exterior of the protein. Since the incubation time is of little interest as far as ligand dynamics is concerned, the "incubating" structures were excluded from the path optimization. If the incubating structures were included, the trajectory would be considerably longer. From the LES simulation, and with the help of computer graphics,[23] a set of protein + ligand structures was selected in which the ligand seems to make spatial progress along the escape path. We eliminated in our selected coordinates any oscillations (aggregation of the polymer) within the protein cavities.

The ligands in the three paths escaped from the protein matrix at the following times: 2.4 ps (A), 2.4 ps (B) and 2.9 ps (C). Each ligand felt distinct protein environment in spite of the fact that a single protein trajectory was used. We denoted the three trajectories (and the resulting three minimum energy paths) by A–C.

(b) Minimization of Selected Structures. Each of the selected structures in the trajectories A–C was subjected to 7500 minimization steps. A typical energy gradient value after the minimization was 2×10^{-4} kcal/(mol Å). Throughout this work, the minimization of the potential energy of the structures and of the polymer energy S was pursued by using the Powell conjugate gradient algorithm.[19] We found this algorithm very stable, efficient, and requiring relatively small amount of computer memory. All the minimizations performed fell to distinct energy minima. In the minimization, a modified CHARMM program was employed and the potential energy was that of CHARMM19.[24] The distance-dependent dielectric (dielectric "constant" proportional to r) was employed. This choice of dielectric constant is expected to reproduce some of the solvent screening effects. No explicit water molecules were included in the simulation. One may expect two solvent effects. The first is a modification of the effective potential that the ligand feels in the protein interior. The distance-dependent dielectric addresses this problem. However, another effect that is not considered in the present work is the penetration of water molecules to the protein. These water molecules may change the diffusion process by "colliding" with the ligand directly.

The 1–4 scaling factor was 1.0, and the cutoff distance (unless specifically stated otherwise in the text) was 9.0 Å. The electrostatic and the van der Waals forces were truncated smoothly over a distance of 1 Å by using the SHIFT cutoff option.[24] The nonbonded list was recalculated each 50 steps of minimization. For each of the paths A–C, we obtained a set of minimized coordinates, which we call "reference structures" (RS).

If the reaction coordinate includes only a single barrier between two energy minima, then the energy minimization of the different structures along the diffusion path should yield only two configurations: one corresponding to a reactant (ligand buried in the pocket) and one corresponding to a product (ligand outside the protein matrix). This is however not the case, and the major fraction of the minimized structures differ significantly from each other. The distinct minimized structures—the RS—were further examined graphically. Since the prime focus of the present investigation is the motion of the ligand, we select only sequential minima in which the ligand makes some progress to its "ultimate goal" of escaping from the protein matrix. Sequential minima with a ligand "at rest" were excluded from the list of the reference structures—RS's—even if other protein segments may have been active (e.g., surface side chains). This procedure yielded 18 RS for path A, 13 RS for path B, and 10 RS for path C.

(c) Calculation of Reaction Paths between the Reference Structures. The intermediate energy minima along the escape pathway were then connected by reaction coordinates by using the SPW approach.[1] The average root mean square difference between the ith and $(i + 1)$th reference structures was 0.28 Å, which is quite small. Counting on the proximity of the intermediate minima, only four interpolating structures between sequential RS's were used. The calculations between the RS's were initiated by using straight line interpolation. This procedure saved a considerable amount of computer time compared to the calculation of the whole path (from the heme pocket outside the protein) in one segment.

The recommendations of the previous computational study on alanine tetrapeptide[1] for SPW parameters were followed. The parameters for

(19) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes*; Cambridge University Press: Cambridge, 1986; Chapter 10.

(20) Case, D. A.; Karplus, M. *J. Mol. Biol.* **1979**, *132*, 343.

(21) 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.

(22) Tilton, R. F.; Singh, U. C.; Kuntz, I. D., Jr.; Kollman, P. A. *J. Mol. Biol.* **1988**, *19*, 195.

(23) The QUANTA molecular graphic package (product of POLYGEN Corporation) was employed.

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

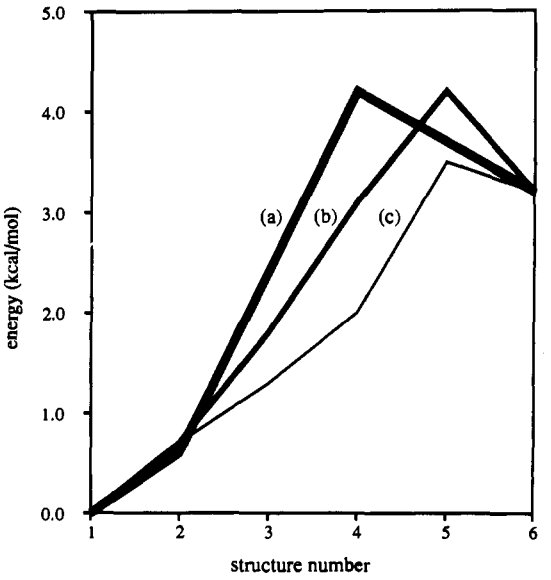

Figure 1. Dependence of the total energy profiles on the stiffness of the chain (the parameter ρ). The paths between the first and the second reference structure of path A are presented. The increasing width of a line corresponds to increasing stiffness: (a) ρ = 128, (b) ρ = 64, (c) ρ = 32 kcal/mol.

the chain optimization were γ = 128 kcal/(mol Å^2) and λ = 2. ρ—the repulsion strength—was varied from 32 kcal/mol (which was used most of the time) to 64 and 128 kcal/mol at two segments of the path. The larger ρ values were used for paths with jumps (see below).

Along the optimized segments, we sometimes found new intermediate minima that were lower in energy than the end points—the original RS's. In these cases, the deeper minima were added to the RS list. They were further minimized (by using 6000 Powell conjugated gradient steps), and new path segments were constructed between the lower energy structures in the same way as for the original RS's.

We argued before[1] that the repulsion can be associated with "kinetic" energy. We found that, in path segments with significant energy barriers, the increase in ρ (kinetic energy) was necessary in order to obtain smooth paths. The typical numerical observation for values of kinetic energy that are too small is an abrupt change in atomic positions along the reaction coordinate. For example, one piece of the polymer is located before the barrier and another piece after the barrier. Therefore proper sampling of the barrier region is not obtained. These sudden changes were corrected either by increasing ρ and/or by a "focusing" procedure to be described below. In practice, the potential energy profiles were found to depend only slightly on ρ (Figure 1). The increase in stiffness in the range considered here did not affect the height of the calculated barriers significantly (the maximum difference was 0.7 kcal/mol).

Abrupt changes along the reaction coordinate occur also as a result of differences in "time scales". For a given time interval, some degrees of freedom move faster and further than others. Or in the "polymer" language, the step taken by the polymer (the displacement between structures r_{i+1} and r_i) can be distributed in a very inhomogeneous way between the different atoms. Some of the specific components ($r'_{i,i+1}$) of the vector—$r_{i,i+1} = r_i - r_{i+1}$—can be very large while other components are practically zero. The differences between the values of the components—$r'_{i,i+1}$—are constrained only by the norm of the vector differences. This is a single constraint on a vector with a number of independent components that is equal to the number of degrees of freedom of the physical system (a few thousand). Hence, the system is very flexible, and the distribution of the coordinate differences between the two vectors can be very spiky and may include large changes at a few hot spots. Ideally we should like to have changes that are small everywhere. Obviously, even more critical is to keep the displacements of the coordinates in which we are most interested (i.e., the ligand position) as small as possible. We found numerically that increasing the chain stiffness is not always helpful for widely inhomogeneous distribution of step sizes. Returning to the kinetic energy view of the repulsion term, we note that increasing the stiffness is similar to a uniform increase of the kinetic energy of all the coordinates. Since the "velocities" of the fast and the slow degrees of freedom are enhanced in a similar way, it is hard to obtain a path with a better resolution if time scale separation is the problem.

Rigorously, one needs to add as many intermediate points as possible to decrease the "time step" and to sample the "fast" motions. This however makes the computations significantly more expensive. We therefore implemented an alternative focusing method. In this approach, the polymer constraints are added only to a selected subset of atoms and the rest of atoms in the system are directly minimized. Physically speaking, the unconstrained atoms are moving adiabatically with respect to the constrained ones. For any displacement of the constrained atoms, the atoms that are constrained adjust instantaneously to a minimum energy configuration. This adjustment is independent of the rate in which the constrained atoms are moving. This is possible only if the unconstrained atoms are moving "infinitely" fast. Hence, this procedure changes the hierarchy of the time scales in the system; the previously slow motions are accelerated, and the fast motions can now be studied in greater detail.

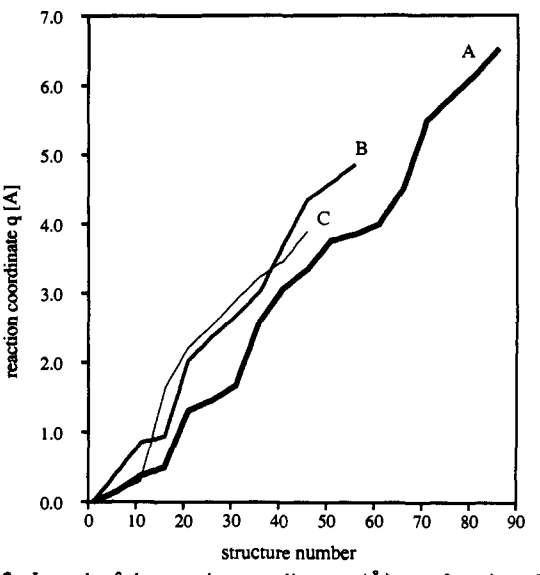

Figure 2. Length of the reaction coordinate q (Å) as a function of the structure number.

Practically, we redefine in formula 1 the distance between two monomers to be $d_{i,j}^{\text{select}}$. The distance with selection—$d_{i,j}^{\text{select}}$—includes only the K selected degrees of freedom:

$$d_{i,j}^{\text{select}} = [\sum_{k=1}^{K} (r_{ik} - r_{jk})^2]^{1/2} \quad (2)$$

where r_{ik} is the selected kth vector element of the coordinate set of the ith monomer. In some path segments, we detected large jumps in the position of the center of mass of the ligand (carbon monoxide), which varied in the range 1.4–2.5 Å. The jumps occur during the transition over energy barriers for the CO motion. We therefore selected a subset of atoms enclosed by a 6-Å sphere around the ligand. This subset was used to calculate $d_{i,j}^{\text{select}}$ (eq 2). The selected domain of 6 Å around the CO was used only for path pieces with jumps. The selection yielded a smooth path, which is referred in the text as the "high-resolution" path. Since the usual protocol that we employed is different, we analyzed the two paths separately. When the focus was on gross features of the protein motion, the "low-resolution" path was employed. To extract detailed interactions of the small ligand with the residues along the path, the high-resolution path was used (when necessary).

Each of the path segments was subjected to direct minimization. The minimization was pursued until the norm of the gradient of S was less than 0.005 kcal/(mol Å). Typically 800 minimization steps were sufficient, but sometimes a larger number of steps (up to 2500) was required. In Figure 2, we plot the length of the reaction coordinate as a function of the number of the structure along the reaction coordinate. The length of the path is defined by the sum of the root mean square differences between the sequential structures. The almost linear dependence of the length of the path on the structure number indicates that the reference structures were chosen properly. The path lengths varied from less than 4.0 Å to more than 6.5 Å. The variations in the length of the different paths are correlated with the length of the CO trajectory: In the longest path (path A), the CO traveled the distance of 29.66 Å and in the shortest path (path C)—20.76 Å. Note that the paths considered in Figure 2 are the low-resolution paths. Along these coordinates, the CO position varied smoothly most of the time but jumps occurred in the neighborhood of the barrier. The largest jump of 5.56 Å in CO center of mass position was detected for path B between structures 18 and 19. All the jumps were eliminated by using the focusing protocol described above.

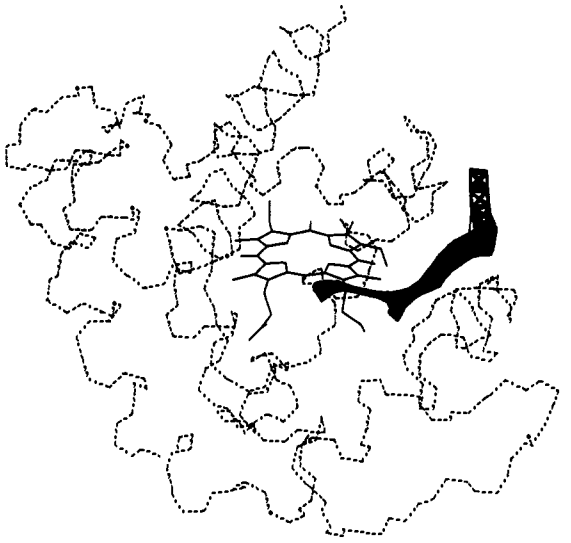

Figure 3. Complete three reaction coordinates for the diffusion of CO through leghemoglobin. The positions of the three ligands at all times are plotted. For clarity, only a single set of a protein backbone and a heme group is shown.

Before proceeding to the results section, we summarize the steps employed in constructing a path: (a) Select structures from a trajectory connecting the reactant and the product. (b) Minimize the energies of the selected structures to obtain the reference structures (RS). (c) Connect the distinct minima by paths obtained from minimization of S. (d) If minima with lower energies than the RS energies for the present path segment are detected, add these minima to the RS and go back to (c). (e) Search for abrupt changes in atomic coordinates. If found, correct by increasing ρ or by using the focusing protocol.

IV. Results

In this section, the results for the three path calculations are described in detail. Most of the analysis is focused on the motion of the ligand through the protein matrix and the protein fluctuations that are coupled to that motion. Analysis of other motions will be discussed more briefly. A schematic view of the pathways of the ligands through the protein matrix is shown in Figure 3.

Path A. This is the longest path studied. Eighty-six structures are used to represent the reaction coordinate. In the first 15 structures, only small "activity" is observed. The rms between the first and the 15th structure is 0.25 Å, and the length of the reaction coordinate at that position—q_{15}—is 0.48 Å. The CO ligand is still in the heme pocket, which is defined in leghemoglobin by the following residues:[2] Phe 29 B9, Phe 44 CE1, His 63 E7, Ala 64 E8, Val 67 E11, Phe 68 E12, Val 110 G8, Ile 114 G12, and Tyr 138 H12. The structural fluctuations in this piece of the path are associated with small side chain fluctuations. For example, the largest χ_1 changes that occurred were of Thr 4 A1 (47°), Leu 32 B12 (10°), and Glu 112 (10°). We also observed significant fluctuations of other side chain dihedrals for the residues Glu 16 A12, Lys 111 G9, Lys 115 G13, Lys 116 G14, Asp 139 H13, and Met 149 H23.

In the following three structures (16–18), some rearrangement of the B and the G helices occurs. For example, the distance between the C_α of Arg 28 B8 and the C_α of Thr 117 G15 increases slightly from 7.11 to 7.26 Å, and at the same time the distance between the C_α's of Leu 32 B12 and Ala 113 G11 decreases from 6.02 to 5.68 Å. Also observed is the reorganization of two pairs of charged residues. The conformations of the following side chains are altered to obtain more favorable electrostatic interactions: (a) Glu 35 B15–Lys 116 G14 (the distance between $O_{\epsilon 1}$ and $H_{\xi 1}$ decreases from 6.38 to 3.61 Å) and (b) Arg 28 B8–Glu 120 G18 (the distance between $H_{\xi 1}$ and $O_{\epsilon 1}$ decreases from 4.29 to 2.69 Å).

We emphasize that we did not find any evidence that these motions are correlated with the ligand motion. In fact, it is most likely that these surface transitions occur during the LES molecular dynamics[2] and the minimization of the reference structures, which are used to guide the path yields minima close to the new conformers.

Between the 18th and the 19th structures, a shift of 4.7 Å in the CO position was observed. This jump is correlated with a tilt of Phe B9 29. To obtain a more detailed description of the ligand motion, we calculated a high-resolution path between the structure before the jump and the structure after the jump. Examination of the structural differences between the starting reactant and the highest energy point of the interaction energy of the CO with the protein reveals that the gate opening is dominated by a tilt of Phe 29 B9. The tilt cannot be described by a single "traditional" internal coordinate of Phe 29 B9 (e.g., bonds, angles, torsions) but rather as a combination of a number of them. χ_1 increases by 6.7° and χ_2 by 9.9°, and the angle C_α–C_β–C_γ changes by 1.8°. In the RS that is the end of this path segment, the torsions relaxed toward the original values but the relaxation was not complete. χ_1 of Phe 29 B9 of the final RS is 4° higher (χ_2 is 4.4° lower) than the corresponding values of the RS at the beginning of the path segment. Also observed is a small rotation of the B helix along its axis. The tilt of the Phe 29 B9 ring resulted in a shift of the position of the C_ξ atoms by 1.27 Å. The shift maximizes the close distances of the CO and the nearby residues at the gate. For example, the distance between C_ξ of Phe 29 B9 and C_{bc} of the heme increased from 5.8 to 7.2 Å at the top of the barrier. This distance decreased to 5.0 Å after the ligand passed the barrier. Other residues that form the gate show similar behavior. The distance $C_{\gamma 2}$ of Val 110 G8 increased from reactant (5.8 Å) to "barrier" (6.8 Å) and is reduced to 6.4 Å at the product. Without the tilt, the distance between the oxygen of the carbon monoxide and the C_ξ of Phe 29 B9 would be 2.9 Å instead of the observed 3.3 Å.

It is also interesting to note that a reaction coordinate for ring motions in proteins was *assumed* in the past to be a combination of several torsions.[25] This is in general agreement with the present calculation.

In Figure 4, the "local energy" of the CO as a function of the reaction coordinate is shown. The local energy includes the internal energy of the ligand and the interaction of the ligand with the rest of the protein (9-Å cutoff for nonbonded interactions was employed). The energy profile for the present path (path A) and the other paths (paths B and C) shows the existence of a clear energy barrier associated with the escape from the heme pocket. We found the concept of local energy advantageous in the description of the diffusion process as compared to the total protein energy. The possible existence of many irrelevant motions (e.g., surface side chain flips) may complicate the analysis of the total protein energy (Figure 5). We were unable to correlate this energy with the details of the diffusion of the ligand through the protein.

After passing the first barrier described above, the ligand is trapped in another cavity in the protein interior, which is surrounded by the following residues: Phe 29 B9, Leu 32 B12, Val 33 B13, Ile 36 B16, Phe 44 CE1, Val 109 G7, Val 110 G8, and Ala 113 G11. All the residues that were within a 5-Å cutoff distance from the position of the CO at the 24th structure were employed to define the cavity. The definition of this cavity remained essentially the same for the three paths. For example, the above residues were also found by using the same cutoff in path C. Two additional residues that were found for path C—Ser 13 A10 (4.8 Å) and Ala 40 C3 (4.3 Å)—are close to the boundary of the 5-Å sphere. Their addition is a consequence of cutoff conditions rather than a significant difference in the properties of the cavity. The distance between the center of the above cavity and the center of the heme pocket is 7.4 Å (6.8 Å in path C). We call this cavity "cavity II".

At structures 20–24, Phe 29 B9 recoiled back to an orientation similar to the coordinate segment in which the ligand was in the heme pocket (closing of the door). At the same time, the B and C helices increased their distances to the heme (for example, the

(25) Northrup, S. H.; Pear, M. R.; Lee, C. Y.; McCammon, J. A.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 4035.

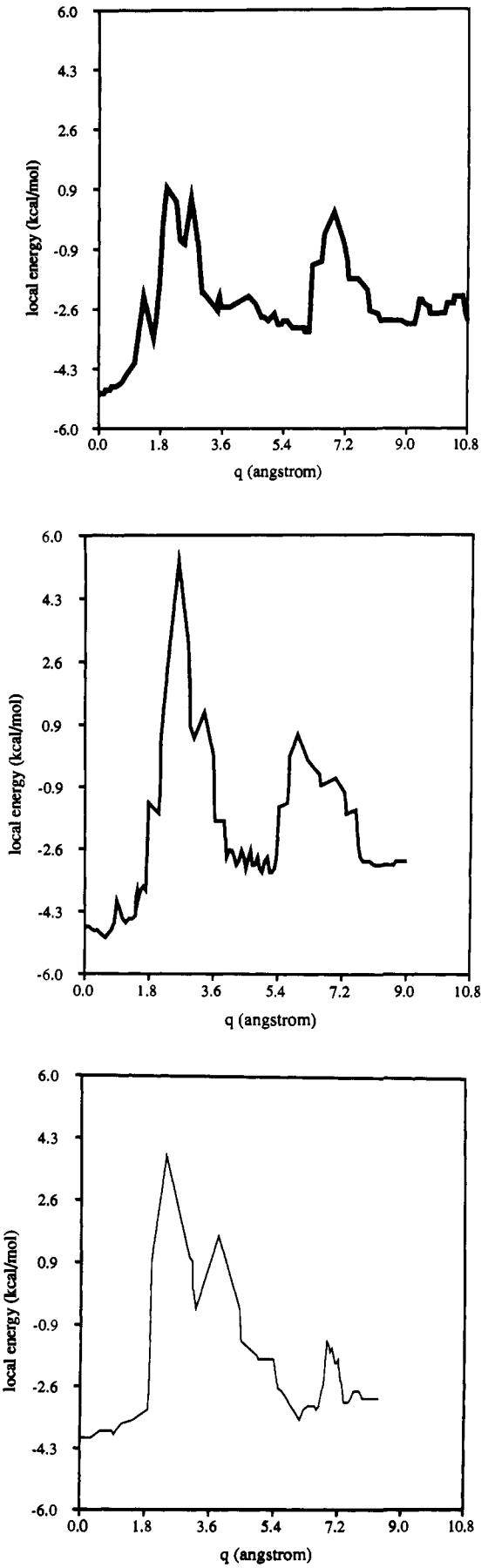

Figure 4. "Local" energy (kcal/mol) of the ligand—carbon monoxide—as a function of the reaction coordinate q. The width of the lines decreases in the order path A, path B, and path C. See text for more details.

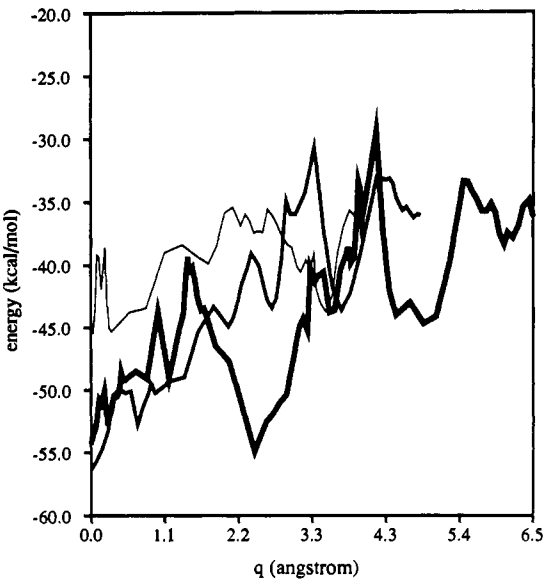

Figure 5. Potential energy profiles (kcal/mol) of path A (thick line), path B (intermediate width), and path C (thin line) along the reaction coordinate q.

distance between the C_α of Ala 40 C3 and the heme varied from 4.84 to 5.11 Å and the distance between the C_α of Ala 37 B17 and the heme changed from 7.80 to 8.02 Å). Structures 24–30 in which the CO "rested" at cavity II had little "relevant" protein structural fluctuations. More significant motions occur in structures 33 and 34. The C helix is translated in a direction approximately parallel to the line connecting the proximal histidine with the iron—W. The C helix translation continues to structure number 40. By overlapping only the backbone of the C helix after overlapping first the whole protein, we estimate the C helix translation between structure 34 and 40 to be 0.44 Å. The helix is also slightly rotated by 2.7°.

In structures 43–63, along the reaction coordinate, small fluctuations of the protein are observed and the CO is moving by only 0.66 Å. In structures 64–67, the backbone of the G helix undergoes a significant fluctuation that is not rigid helix motion. The backbone of the residues 102–111 (the residues of the G helix are 103–123) is shifted in the W direction. In the next path segment, the same fragment of the G helix moves in the backward direction and at the same time the remaining portion of the G helix (residues 112–123) bends toward heme.

The fluctuations of the C and G helices are of crucial importance for opening the final gate, as is evident from the ligand motion. The ligand escape is highly correlated with the helices displacements. In the first fluctuation of the G helix, the CO approaches the barrier, and "positions itself" close to the contact between the G and the C helices. Then, the backward motion of the first segment of the G helix "spits" the CO outside the protein. At the same time, the motion of the C helix minimizes the possible bad contacts of the CO at the gate. In parallel to the large structural fluctuations of the G and the C helix, the ligand jumps by approximately 6 Å to the exterior. The nearest neighbors of the CO just before the abrupt shift in position are Ala 37 B17 (3.15 Å), Ala 40 C3 (3.30 Å), His 106 G4 (3.63 Å), and Val 109 G7 (3.32 Å). The distance between Ala 37 B17 and His 106 G4 increases from 5.60 to 6.98 Å, supporting the evidence that the extensive fluctuations of both the C and the G helices are important in opening the gate.

A more detailed picture of the diffusion path was obtained by the focusing protocol. The focusing protocol reduced the translation of the ligand molecule between sequential structures to a maximum value of 1.0 Å. It is therefore considerably easier to analyze the mechanism of the gate opening. In Figure 6, we show the distance between the C_β of Ala 40 C3 and the $C_{\gamma 1}$ of Val 109 G7 as a function of the reaction coordinate in the high-resolution path. It is evident that when the CO passes through the second

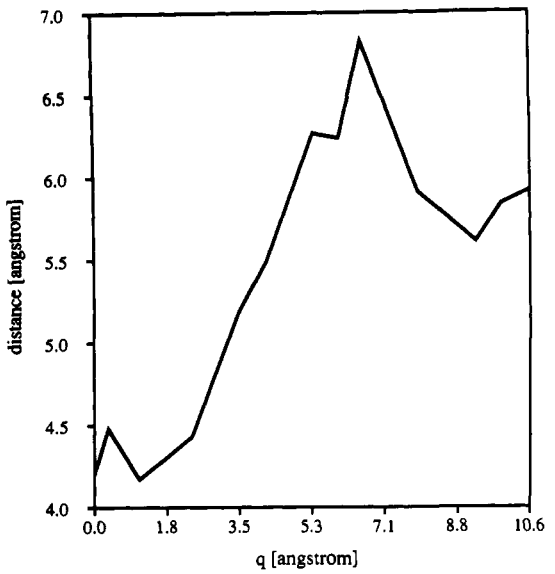

Figure 6. Distance d between Ala 40 C3 (atom C_β) and Val 109 G7 (atom $C_{\gamma 1}$). Note a maximum in d that occurs at the point in which the CO crosses the second barrier. The data are presented for the high-resolution variant of path A.

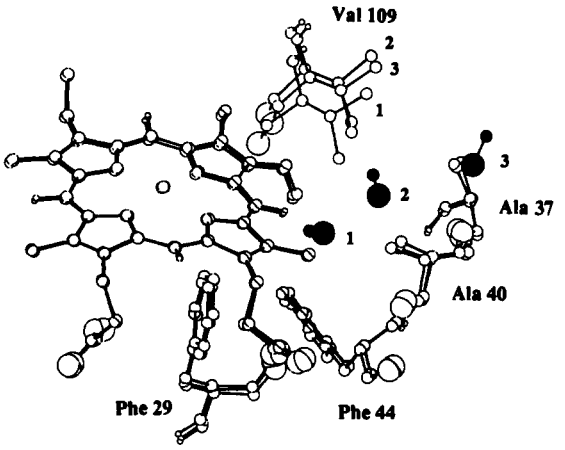

Figure 7. Protein and ligand motions associated with the opening of the second gate (path B). Superimposed are three structures from the high-resolution path corresponding to the location of the carbon monoxide (dark) before, at the top, and after the barrier. Significant motion is observed in this view only for Val 109 G7. The numbers 1, 2, and 3 near the carbon monoxide and the valine 109 describe the path progress; i.e., 1 is the reactant, 2 is the transition state, and 3 is the product.

gate, the distance between the above two residues is the largest. In Figure 7, we show the structural variations of the protein residues that are relevant for the opening of the second gate. Three structures are overlapped: before the transition, after the transition, and at the top of the barrier. This is the final gate after which the CO resides at the surface of the protein.

Path B. This path is described by using a discrete set of 56 structures, which were derived from 13 RS structures. Since the different RS's were selected from the same protein trajectory (but from different ligand trajectories), it is of interest to compare the minimized RS's of the three paths. In Figure 8, we show the set of structures that were chosen from the trajectory for the different paths. We note that minimizing exactly the same protein structure while changing only the ligand yields two protein configurations that differ by ~0.3 Å. This is a significant difference since the maximum deviation between any minimized RS's was only ~1 Å. This suggests that including the same protein structures from the LES trajectory but with different copies of the ligand is sufficient to produce significantly different RS's and therefore different paths.

Similarly to path A, the initial fragment of the path is quiet in the CO coordinates. The CO remains in the heme pocket in

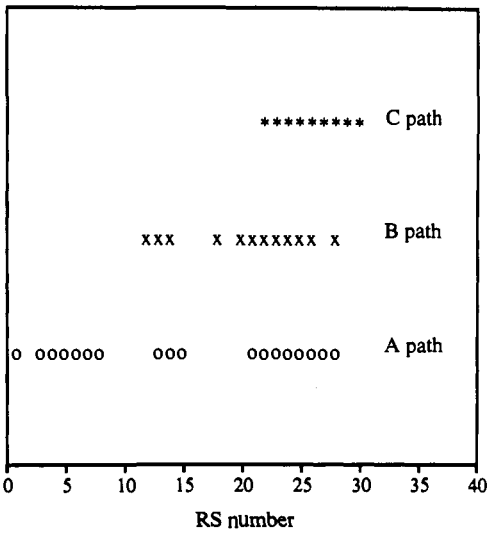

Figure 8. Relationship between the trajectory and the reference structures (RS) that were used to generate the three paths. The axis refers to the structure number from the trajectory. Note that a single trajectory for the protein was used for the three different ligand trajectories. For each path, we marked the structures that were minimized to RS's. See text for more details on the computational protocol.

the first 18 structures, and then in the 19th structure it abruptly changes its position to cavity II. In a comparable way to the previous path, it passes between Val 110 G8 and Phe 29 B9. The CO translation between structures 18 and 19 was 5.9 Å, which necessitates the use of the focusing protocol to study the mechanism of the transition. We therefore constructed a high-resolution path between structures 16 and 21. The high-resolution path consists of 4 path segments and 23 intermediates connecting the end points and the 4 intermediate minima. In Figure 9a and b, we present a picture of the CO and the most relevant residues for the transition between the heme and cavity II. The figure includes the structure before the barrier, on the top of the barrier, and after it. In addition to the ligand, only Phe 29 B9 shows significant displacement. The internal degree of freedom that changed most is χ_2 (10°). Note however that this is not the only internal coordinate that changes, and the addition of slight modifications of many coordinates yields a net effect of tilting the phenyl ring by approximately 10° (Figure 9a). For example, the angle C_α–C_β–C_γ is reduced by 1.2°, and χ_1 is reduced by 5°. The C_ξ of the Phe 29 B9 is displaced by 1.57 Å, and at the same time the C_α of the same residue is shifted by only 0.2 Å due to small helix rotation in the direction of the helix long axis. The tilt is critical for reducing bad contacts. If the displacement of the phenyl ring did not occur, the distance between the ring and the ligand would be 1.9 Å instead of the observed 3.3 Å.

It is also amusing to note that the jump in the CO coordinate is clearly associated with a local energy barrier (Figure 4b). The situation became less clear, however, when the total energy of the protein was examined (Figure 5). It includes many "irrelevant" motions that affect the total energy value but are poorly correlated with the behavior of the carbon monoxide.

In the following path segment (structures 19–37), we find the CO in cavity II. Examination of the CO movement within the cavity reveals an unsuccessful attempt to leave it during structures 20–26. The CO is moving toward Ala 37 B17, which is one of the critical residues at the second gate (see previous path), but it recoils back at structure 30. The CO position in the second cavity at structure 30 is very similar to that of structure 20. The second attempt to leave the protein matrix (structures 30–38) is more successful. The CO jump to the exterior occurs between structures 37 and 38.

The first translation of the CO toward Ala 37 B17 resembles the behavior of a trajectory in which the reactant collides unsuccessfully with the barrier before passing it. The computational scheme that we followed was aimed at minimizing the number

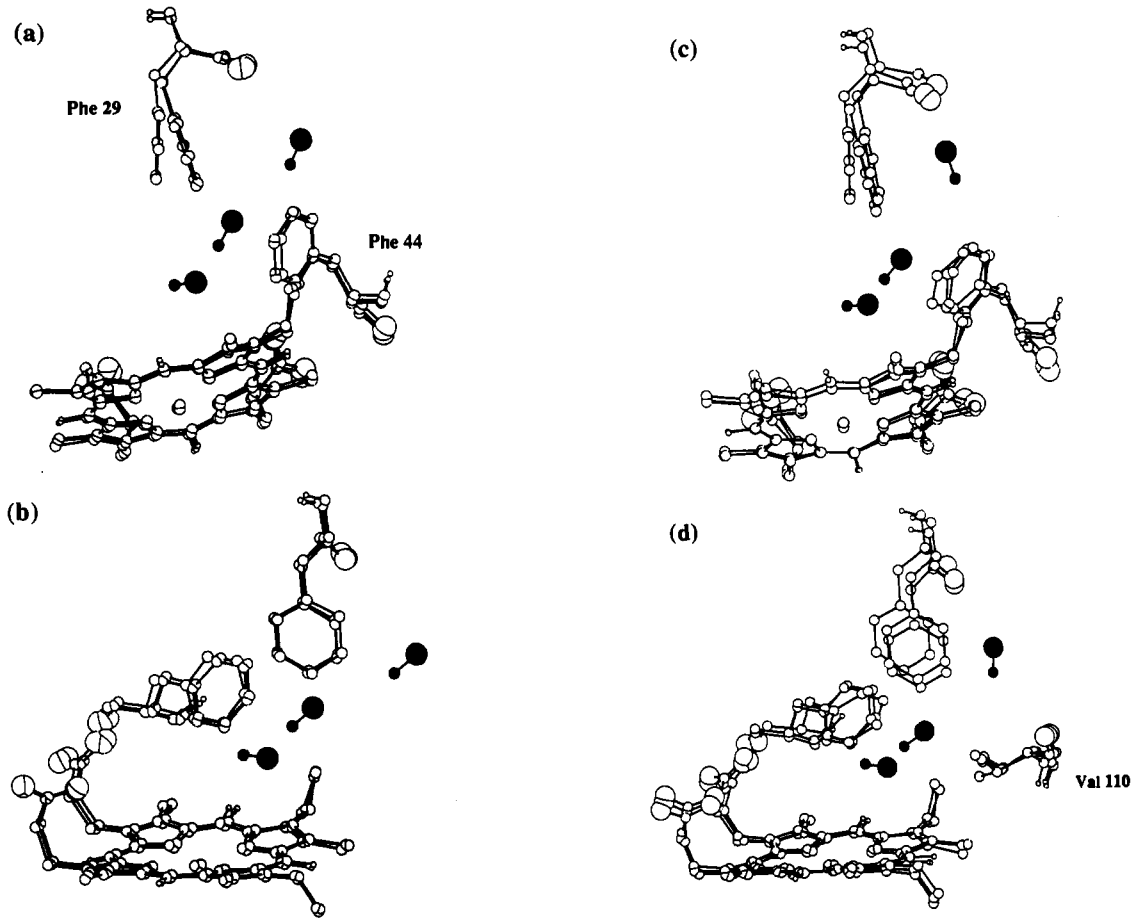

Figure 9. Structural changes connected with the crossing of the first gate between the heme pocket and cavity II. Superimposed are three structures from the high-resolution path before, at the top, and after the barrier. The dark diatom is the ligand—carbon monoxide. (a) and (b) are two different projections of the gate for the path B, while (c) and (d) illustrate the same gate in the path C. Note the significant tilt of the side chain of Phe 29 B9.

of such oscillations. Though we were quite successful ("periodic" motion of the ligand was observed only in this path segment), the success rate was smaller than 100%. Before the final escape, the CO makes hard contacts with Ala 37 B17 and Val 109 G7. This finding is in accord with the analysis of the LES trajectory by Czerminski and Elber.[2] In their analysis, it was suggested that the contacts of the ligand with the C and G helices are important in determining the final barrier for the escape.

The largest fluctuations of the protein were observed at the final steps of the ligand escape (structures 40–50). The most active domain is the B, C, and G helices. In Figure 10, we show a matrix of differences between distances. The distance matrix for the C_α of structure 40 was subtracted from the corresponding matrix of structure 50. The figure shows that the distances between the B, C, and G helices increased. Thus, the protein adoped a somewhat more open structure. Furthermore, examination of the dihedral degrees of freedom—ϕ,Ψ and χ_1—reveals no transitions (a transition is defined by a change of at least 60°). The maximum χ_1 displacement observed is of Asn 19 A16 (59.8°). The maximum of ϕ,Ψ displacement was of 23°. The fluctuations observed are therefore global and arise from a large number of small changes in internal coordinates. Since the CO molecule is almost outside the protein matrix at the beginning of these extensive motions, the global fluctuations do not seem to affect the diffusion process significantly. Some contribution to the opening of the C/G gate is expected from the beginning of the process.

Path C. This is the shortest path of all three (3.9 Å). As in previous paths, at the beginning (the first 0.6 Å and 12 structures), the ligand is almost stationary in the heme pocket. To obtain insight to the characteristics of fluctuations that occur in the protein structure, the rms as a function of residue number is presented in Figure 11. The rms is between the first and the tenth

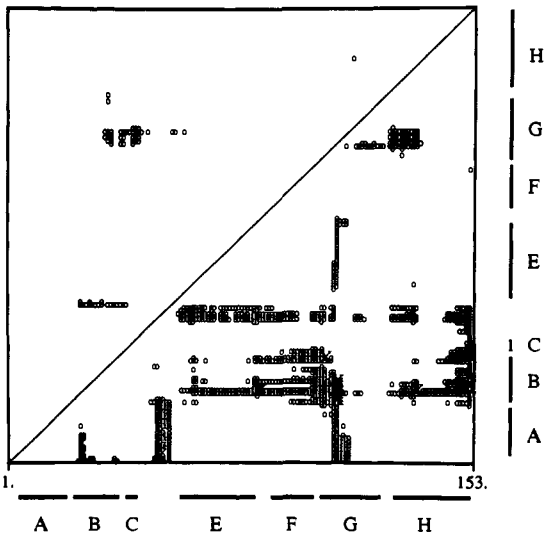

Figure 10. Matrix of differences between distances (d) for structure 40 and 50 of path B. The distances between the C_α atoms of structure 50 were subtracted from the corresponding distances of structure 40. Circles denote 1.0 Å < d < 2.0 Å, crosses d > 2.0 Å; positive differences are marked on the upper triangle of the matrix, while negative differences are on the lower one. The bars on margins indicate ranges of helices.

structures along the reaction coordinate, and it is close to zero for the most of the residues. A similar result is obtained for the rms of the backbone (Figure 11a) and for all atoms' rms (Figure 11b). Only at the C/E and the G/H loops the fluctuations are larger than 0.2 Å. The sharp spikes in Figure 11b correspond

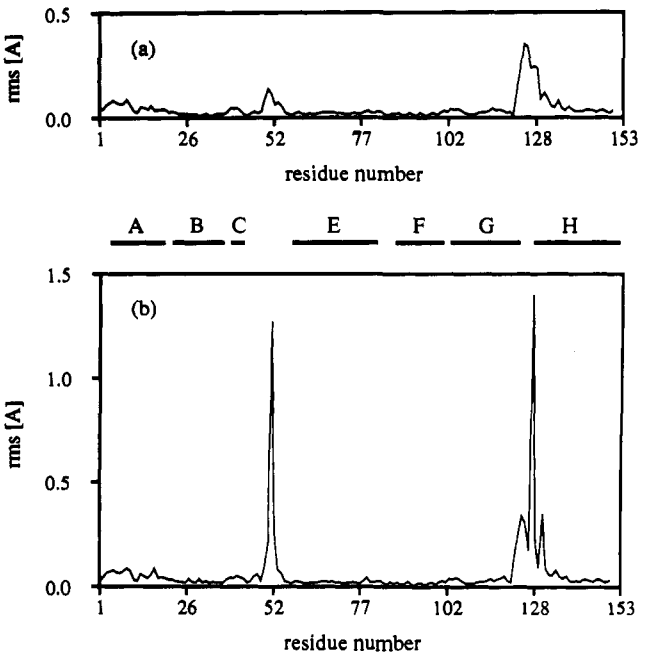

Figure 11. Root mean square between the first and the tenth structures of path C as a function of the residue number: (a) backbone atoms only, (b) all atoms.

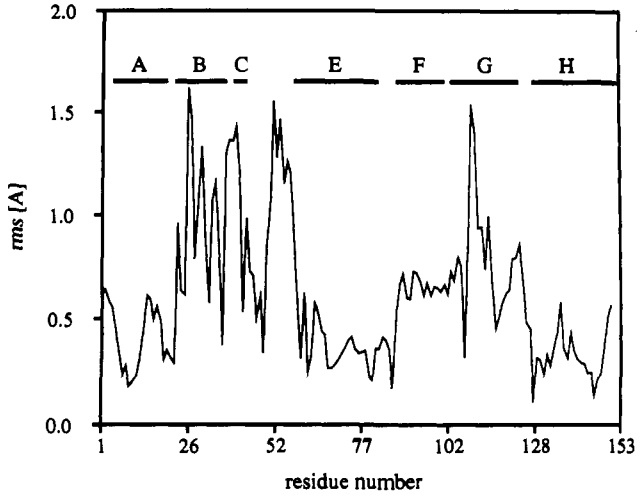

Figure 12. Root mean square between the 10th and the 20th structures of path C as a function of the residue number. Only backbone atoms are shown.

to hydrogen-bonding rearrangement of Ser 51 CE10 and Ser 127 H1 on the protein surface.

In contrast to the first "quiescent" segment of path C, the next set of structures (10–20) show significantly more activity (Figure 12). The protein domain that includes the B and C helices and the C/E loop increases its distances from the rest of the protein and especially of the G helix. These fluctuations are best demonstrated in a distance matrix plot in Figure 13. Furthermore, the B and C helices change their orientations by 11° and 8°, respectively. We obtained these values by overlapping first all the protein atoms and then by overlapping the backbone of each of the helices (B and C). The remaining helices change their orientation by less than 3°. From the overlap of the helices, we also obtained the relative translation, which was the largest for the B and C helices (0.75 and 1.09 Å, respectively). The radius of gyration of the protein changes somewhat from 14.87 Å (structure 10) to 15.03 Å (structure 20). In parallel to *global* structural fluctuations in the protein, we observed a large jump (5.6 Å) in the position of the center of mass of the CO between structures number 13 and 14. This jump is similar of course to the previously studied transition between the heme pocket and

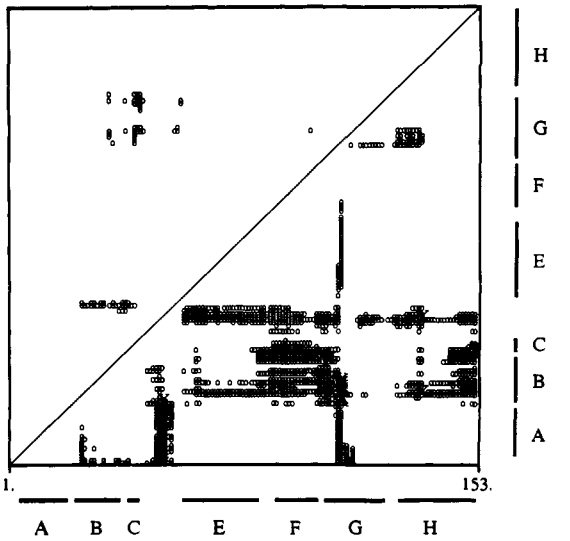

Figure 13. Matrix of differences between distances for structures 10 and 20 of path C. For more details, see Figure 10.

cavity II. Some of the global protein fluctuations observed (but not necessarily all) should be associated with the sudden jump of the ligand.

At this position, we tried to increase the path resolution using a number of alternatives to the focusing protocol. Here we describe the other methods in detail. Similar results were obtained for the other paths too. The first attempt to refine the path was to use 10 intermediates instead of 4 between structures 11 and 16. The second trial was the application of higher stiffness (ρ = 128 kcal/mol). The attempts were not successful. In both cases, large shift in the CO position between sequential structures was observed (in the range 1.5–2.5 Å).

The solution was to return to the protocol that worked: the focusing. The number of selected atoms on which the constraints of the "focused" polymer were employed was 56. These atoms were selected according to the first structure of the path segment. The path with selection was smoother than the path without the selection, and the largest distance between sequential CO positions after the refinement was 0.87 Å. As in paths A and B, this step of the "reaction" corresponds to the ligand escape from the heme pocket to cavity II. The similarity between all three paths suggests narrower distribution of alternative paths at this segment of the reaction coordinate compared to other path segments. At the top of the barrier (structure 24 in the "high-resolution path", Figure 9c,d) we noticed a significant tilt of Phe 29 B9. The tilt cannot be described by a single internal degree of freedom. Instead a combination of several must be used. For example, the angle C_α–C_β–C_γ increases by 2°, χ_2 increases by 10°, and the B helix is slightly rotated along the long axis to further increase the distance between the CO and Phe 29 B9. Other residues that are in the proximity of the CO when it escapes from the heme pocket are Val 110 G8 and Phe 44 CE1. A few more details on the CO motion can be found in the comparable description of the CO motion in paths A and B. The jump brought the CO to cavity II.

In contrast to the escape from the heme pocket, which occurred sharply and was associated with a significant energy barrier (Figure 4), the next phase in which the ligand escaped from the second cavity to the protein surface was continuous and no abrupt changes in the CO position as a function of the reaction coordinate were detected. This behavior is different from the observations in paths A and B. In the first two paths, jumps in CO coordinates were detected in the second phase of the diffusion process too. The final escape of the CO from the protein matrix was between helices G and C; however, it was difficult to correlate the barrier with the fluctuations of a specific residue or with a specific type of motion. The opening was clearly of global nature, but we were not able to assign it to traditional protein coordinates such as torsional angles or even rigid shifts of helices. The fluctuation

of the internal coordinates of the residues that formed the gate (suggested by Czerminski and Elber[2]—Ala 37 B17 and His 106 G4) were too small to be correlated with the gate opening. For example, the maximum χ_1 changes of His 106 G4 in path C were 10.6°. It seems likely that the second barrier was opened in a fluctuation uncorrelated with the CO position. The distance between the B and the G helices increases sharply when the ligand passes its *first* barrier. For example, the distance from the atom C_ξ of Phe 29 B9 and the atom $C_{\gamma 2}$ of Val 110 G8 increases from ~9 to ~14 Å. These structural fluctuations are probably responsible for opening the second gate.

V. Discussion

We presented the results of three different path calculations obtained from quenching LES trajectories to a minimum of the functional S (eq 1). The reaction coordinates are defined as minima of S. Qualitatively speaking, the reaction coordinate is defined as an approximate trajectory with the minimum amount of kinetic energy (the polymer repulsion) that is needed to obtain a continuous nonoscillatory path. This definition does not coincide with the traditional reaction coordinate in small molecular systems (the steepest descent path, SDP). This difference can be viewed in two ways: (a) In a previous study, it was demonstrated that this path is quite close to the SDP, especially in the neighborhood of the transition state, and it is useful as a starting point for further refinement to the SDP.[3] (b) It is not clear if the steepest descent path is the best choice for a reaction coordinate in large molecules, and the present definition may serve as an alternative. For example, the most probable path obtained from stochastic differential equation (which suggests a model for reaction coordinates in condensed phases) was discussed by Onsager and Machlup.[26] Different functionals were reviewed by Wolynes,[27] and Pratt proposed to use Markov chains to sample paths.[17] Elber and Karplus proposed to optimize a line integral.[16] That line integral is the limit of zero kinetic energy (repulsion) of the present protocol. Only the alternative definitions of Elber and Karplus and of Czerminski and Elber were tested on molecular systems.

The present calculation transformed classical trajectory (from the LES protocol) to an approximate trajectory (a minimum of the functional defined by S). A reasonable question is what we can learn from the S trajectory that we could not get from the initial guess—the LES trajectory. An answer is as follows: (a) Some estimates for the barrier energies can be obtained from the low kinetic energy S trajectories, estimates that are much harder to get from an ordinary trajectory. The variations in the system potential energy during regular MD (molecular dynamics) are too large to estimate barrier heights for diffusion. (b) The optimization of S at relatively low repulsions (kinetic energies) quenches a significant fraction of the motions that are not coupled to the ligand diffusion. S trajectory therefore helps to elucidate the structural features of the gate opening. Such analysis is harder to pursue in MD, which includes the irrelevant motions.

Reaction coordinates in biological macromolecules have been investigated in the past. Just to mention a few previous studies, reaction coordinates were studied for ring flips,[25] ligand diffusion in myoglobin,[20] the hemoglobin R to T transition,[28] and the B to Z transition in DNA.[29] The difference between the present study and previous investigations is in the direction that we are taking. The previous researchers used insight, chemical intuition, and ingenuity to propose plausible reaction coordinates in very complex systems. We are trying to design a more automatic approach employing nonlinear optimization. The nonlinear optimization of a general function (in our case, the functional S) is a problem under very intensive investigation in numerical analysis. A number of reasonable solutions are available, one of which we adopted here.

The three paths were calculated by using as initial guesses three ligand trajectories.[2] In accord with the commonly used replacement of time averages by spatial averages, we used protein structures from the same time sequence (trajectory) to generate three ligand paths. We considered the single sequence in time to be similar to spatial average; i.e., ligands that escaped at different times were assumed to be independent, and therefore separate reaction coordinates were calculated.

Another point worthy of discussion is the way in which we generate the reference structures (RS's). They were obtained by direct minimization of trajectory structures. We used previous experience[30] in studying the properties of energy minima in a similar globin (myoglobin). In that study, the potential energy surface was found to be very rough and to include an enormous number of local energy minima in the neighborhood of the native structure. It is likely therefore that the minimization of the trajectory structures will give minima close to the original trajectory structures, simply because of the high density of nearby minimum energy configurations. This was indeed the case in the present investigation as well.

A nontrivial question is what we can learn from a *single* reaction coordinate in these large systems. Czerminski and Elber[1] demonstrated that, for a flexible system (alanine tetrapeptide), the number of steepest descent paths connecting the helix and the extended chain conformations with accessible energy barriers is very large. This may discourage reaction path calculations in molecules larger than short peptides. A point in favor is that the type of ϕ,Ψ transitions obtained in a tetrapeptide is unlikely to be observed in leghemoglobin, which has a well-defined (average) three-dimensional structure. Increased rigidity is expected to reduce the number of paths to a more manageable amount. Furthermore, when the interest is focused only on a small part of the system and the remaining "irrelevant motions" of the macromolecule may be reasonably ignored, a single reaction coordinate can be useful to study the structure and the energy of the local part. The similarity of the local properties of the paths was demonstrated in the Results section.

Qualitatively, all the paths can be separated into two sequential steps; the first step includes the translation of the CO from the initial heme pocket to another pocket—cavity II. The final step is the escape of the ligand from the protein matrix. Below we discuss further the two steps.

Consider the first barrier separating the heme pocket and cavity II (Figure 9). In the three paths, the barrier is dominated by the tilt of Phe 29 B9, while the rest of the protein residues are approximately at rest. The barrier is primarily local and includes a small number of internal coordinates, though some rotation of the B helix seems to be involved as well. On the other hand, the total energy of the protein and of the ligand shows little correlation with the observed local structural similarities. The reason for the large energy variations are other protein motions at positions far from the ligand, for example, the side chain flips of Glu 35 B15 and Lys 116 G14. These transitions do not affect directly the neighborhood of the CO, but their possible effect on the reaction rate is not obvious. The "irrelevant transitions" are history dependent; i.e., they depend on the way in which the path was constructed or on the initial guess, which here is a trajectory. The reference structures were obtained by quenching structures from a trajectory. It is therefore hard to eliminate these fluctuations, which are present in the initial molecular dynamics simulation. In the language of the mean force potential, one assumes that the irrelevant motions such as the side chain flips are much faster than the motion along the reaction coordinate and therefore one should average over all their combinations when estimating the available phase space for the transition state. Unfortunately, such time scale separation between the diffusion and the side chain flips is hard to prove, especially in leghemoglobin in which the diffusion is known to be exceptionally fast and in the nanosecond time

(26) Onsager, L.; Machlup, S. *Phys. Rev.* **1953**, *91*, 1512.
(27) Wolynes, P. G. In *Chemical Reaction Dynamics in Complex Molecular Systems. SFI Studies in the Sciences of Complexity*; Stein, D., Ed.; Addison-Wesley: Reading, MA, 1989; pp 335–387.
(28) Janin, J.; Wodak, S. J. *Biopolymers* **1985**, *24*, 509.
(29) Olson, W. K.; Srinivasan, A. R.; Marky, N. L.; Balaji, V. N. *Cold Spring Harbor Symp. Quant. Biol.* **1983**, *47*, 229. Harvey, S. C. *Nucleic Acid Res.* **1983**, *11*, 4867.
(30) Elber, R.; Karplus, M. *Science* **1982**, *235*, 318.

range.[6] If the time scales cannot be separated, one should consider parallel processes between different conformers. This is essentially the same as the substates model of Austin et al., which leads to parallel paths and to substantial deviation from the "normal" exponential rate law.[31]

It is also of interest to examine the close neighborhood of the ligand, for that purpose we calculated the local energy. It is the interaction of the CO with the rest of the protein, not including the interaction of the protein atoms with each other. The local energy profile is similar in the three paths (Figure 4). An energy barrier associated with the transfer of the ligand to cavity II is clearly seen. Nevertheless, the height of the local barrier varies significantly (up to 6 kcal/mol), which is another source for heterogeneity in the rate and a proposition for substates.

To summarize some of the properties of the first barrier, it seems that, from a single path, one can obtain structural information and yes/no information on the existence of an energy barrier. The *distribution* of barriers, which is important in quantitative estimates of the rate and in understanding the properties of the "reactive" substates, is (of course) impossible to obtain from a single path calculation.

The second step in which the ligand escapes from cavity II to the exterior of the protein shows similar characteristics. The important residues for the opening of the gate are the same in the three different paths, and the local energy profile demonstrates the existence of an energy barrier of an order of a few kilocalories per mole. Similarly to the transition from the heme pocket, the total energy (protein + ligand) changes show little correlation with the diffusion of the small ligand. There are however differences between the two transitions. In contrast to the first barrier, which is dominated by the motion of a single residue (Phe 29 B9), the opening of the second gate is more global and involves the motions of the B and C helices relative to G. Similar observations were made in the LES trajectory,[2] though in the latter only the second barrier was identified. This observation is worth emphasizing since it is not an intuitive result. One would like to think of a reaction coordinate in macromolecules in terms of a single (or at most several) internal coordinates, preferably local. In the first barrier, the reaction coordinate is indeed local; in the second, however, significant coupling to extended coordinates occurs and one must include the protein in the picture. For extensive structural changes (at the second barrier), one may expect that the local energy barrier may be smaller, since the ligand is less involved in the gate opening. As "expected", the local energy barrier for the second step is smaller by approximately a factor of 2 compared to the first barrier. The variations in the height of the local barrier are smaller at the second step compared to the first. This, however, may be accidental.

Possible experimental tests of the proposed diffusion path are finally discussed. At the present level of the theory, it is not possible to give a reliable estimate of the rate. A larger number of reaction coordinates is required, and the problem of weight of the paths should be considered. However, a feature that seems to be converged (at least in our limited sample) is the structural properties. Protein engineering enhanced significantly the range of experiments that can be pursued in investigations of molecular mechanisms in biology. This type of experiments can be helpful here as well. Useful mutations to test the above picture on leghemoglobin are (a) the replacement of Phe 29 B9 by a small residue (to speed up the diffusion) and (b) modifying Val 110 G8 to a larger residue (e.g. Phe) to decrease the diffusion rate. We propose these mutations as a test for the local barrier of escaping from the heme. We comment that the proposed diffusion path is in disagreement with a previous proposition for a path that includes the fluctuations of the distal histidine.[8,20,32] There the ligand leaves the heme pocket directly to the solvent. Here the ligand escapes first to another cavity in the interior of the protein. Site-directed mutagenesis experiments may differentiate between the two paths or at least give them appropriate relative weights.

VI. Summary and Conclusions

A recently developed technique[1] to calculate reaction coordinates in large and flexible systems was employed to study ligand diffusion in a protein. The calculation was shown to be possible even for this 1471 atoms system and to yield some useful results. A systematic and "automatic" protocol was presented of how to construct reaction coordinates in very large systems (>1000 atoms) on the basis of minimization of a functional (eq 1).

We further examined what the quantities are that can be extracted from computations of reaction coordinates in very large molecules. The calculation of the three paths resulted in the following conclusions.

(a) The three paths are similar if the *local properties* are considered. Thus, the energy profile for the interaction of the carbon monoxide with the rest of the protein is similar, and the structural features of the escape pathways are essentially the same for the three calculations.

(b) The paths differ substantially if differences in global protein properties are examined. Since the starting point for the path calculations were trajectories, significant amount of irrelevant motions was added from the initial guess, e.g., transitions of side chains on the surface. These motions changed widely the total energy values (e.g., when charged surface residues formed salt bridges), but they were not correlated with the diffusive motion of the ligand.

As a general conclusion, we note that the most meaningful analysis is obtained by probing local properties. The calculation however needs to be pursued by using all degrees of freedom in order to have a systematic protocol and to describe barriers associated with global protein motions such as the barrier for exit from cavity II.

The calculated paths enable us to explore the feasibility of a previously proposed diffusion pathway that was based on an approximate LES trajectory.[2] The barrier heights calculated for the three paths are reasonable and support the proposed model. Finally suggestions for experimental tests of the theoretically predicted paths by site directed mutagenesis were made.

Acknowledgment. This research was supported by NIH Grant No. GM40698. W.N. thanks the Polish Central Project for Fundamental Research CPBP 01.06.2.03 for the travel grant. R.E. is a Camille and Henry Dreyfus New Faculty Scholar. The calculations were pursued on a TITAN minisupercomputer purchased with an NIH shared instrumentation grant (No. RR04884) to the Department of Chemistry, University of Illinois at Chicago.

Registry No. CO, 630-08-0.

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

(32) Ringe, D.; Petsko, G. A.; Kerr, D.; Ortiz de Montellano, P. R. *Biochemistry* **1984**, *23*, 2.