A METHOD FOR DETERMINING REACTION PATHS IN LARGE MOLECULES: 
APPLICATION TO MYOGLOBIN *

R. ELBER and M. KARPLUS 
Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA 

Received 8 July 1987

An algorithm is described for determining reaction paths between two known structures with many degrees of freedom. The method uses first-derivative techniques to optimize the entire path between the two end forms subject to certain constraints. The stability and convergence properties of the method are illustrated by applications to structural transitions in two test systems 
(cyclohexane and dialanine) and to a conformational change involving all degrees of freedom in the protein, myoglobin, with 1531 atoms.

1. Introduction

The problem of finding the transition state or, more generally, a reaction path between two known local minimum-energy structures is of considerable interest [1-4]; methods currently in use have been reviewed recently [1]. For small and intermediate sized molecules with less than 100 degrees of freedom, search methods involving local propagation by means of the second derivative of the potential (modified Newton–Raphson algorithms) have been shown to be effective [2,3]. However, they are difficult to extend to large systems, such as proteins with more than a thousand degrees of freedom, because the computational effort increases at least as the third power of the number of degrees of freedom. Further, for complex potential surfaces, the second-derivative matrix is likely to have many eigenvalues that are negative or close to zero so that the calculations (including matrix inversion) are unstable. We present here an algorithm that uses first derivatives and is based on the refinement of an arbitrary discretized path connecting the two structures of interest.

In section 2 we describe the method. Illustrative applications to small molecules (cyclohexane, dialanine) and to the protein myoglobin are given in section 3.

2. Method

We consider the general problem of minimizing the line integral, $S(R_i, R_j)_L$

$$ S(R_i, R_j)_L = \frac{1}{L} \int_{R_i}^{R_j} [G(R) \cdot d(l(R))]_L, \quad (1) $$

where $G(R)$ is a function of the $N$-dimensional position vector $R$ representing the coordinates of the system of interest and $d(l(R))$ is a line element on the path $L$ of length $L$ between the fixed end points $R_i$ and $R_j$ along which the integral is performed. To find the path $L$ from $R_i$ to $R_j$ that minimizes $S(R_i, R_j)_L$, we introduce a discretized form of eq. (1) which includes $M$ intermediate grid points (the constant contributions from the fixed end point $R_0$ and $R_{M+1}$ are omitted); that is, we write

$$ S(R_0, R_{M+1})_L = \frac{1}{L} \sum_{j=1}^{M} G(R_j) \cdot \Delta l_j, \quad (2) $$

where $R_j$ is the value of $R$ at the endpoint of the interval $\Delta l_j$, defined as

* This work was supported in part by a grant from the National Institutes of Health.
\[ \Delta l_j = \left[ (R_j - R_{j-1})^2 \right]^{1/2} u_{j,j-1} = \Delta l_j \cdot u_{j,j-1}, \]  

with \( u_{j,j-1} \) a unit vector in the direction of \( R_j - R_{j-1} \).

A straightforward minimization of \( S'(R_0, R_{M+1})_L \) as a function of \( N \times M \) variables is not possible without restricting the relative lengths of the line segments, i.e., the lengths \( \Delta l_j \) of different elements could take on such disparate values during the minimization that eq. (2) would no longer be a good approximation to eq. (1). We therefore introduce a quantity \( S'(R_0, R_{M+1})_L \) that includes constraints which ensure that the \( \Delta l_j \) are approximately the same for all \( j \); that is,

\[ S'(R_0, R_{M+1})_L = S(R_0, R_{M+1})_L + \sum_{j=0}^{M} \lambda (\Delta l_j - \Delta l)^2, \]  

where

\[ \Delta l = \left( \sum_{j=0}^{M} (\Delta l_j)^2 / (M+1) \right)^{1/2}. \]

Here \( \lambda \) is a parameter which determines the range of allowed fluctuations in \( \Delta l_j \) relative to the average value, \( \Delta l \). The path \( L' \) that minimizes \( S'(R_0, R_{M+1})_L \) is expected to approach the path \( L \) for \( \Delta l_j \) constant and small. Thus, with a sufficiently large \( \lambda \) and a suitable number of intervals \( (M+1) \), we can approximate \( S'(R_0, R_{M+1})_L \) by a straightforward minimization of \( S'(R_0, R_{M+1})_L \) as a function of the path \( L' \), which is determined by the choice of intermediate points \( j \). There is no restriction on the actual length of the path and the size of \( \Delta l \). It may vary during the optimization provided that all \( \Delta l_j \) vary at the same time to the same extent. One must therefore make certain that \( \Delta l \) is indeed small enough to obtain an accurate description of the path; the linear interpolation used between pairs of structures is sufficient if the addition of an intermediate point does not change the path significantly.

With eq. (4) any standard non-linear optimization method may be applied to \( S' \). We have used the Powell algorithm [5]. It requires only first derivatives, which makes the method suitable for large systems since the value \( M \) is expected to be small relative to \( N \) (e.g., \( M \) is on the order of 10 for \( N \) as large as 4600, as described below). For small systems a second-derivative Newton-Raphson type algorithm could be introduced to minimize \( L' \), although it is not clear that for such problems the present method is better than alternative approaches [2,3].

To employ eq. (4) for determining a "reacting path" on a physical potential-energy surface, the function \( G(R_j) \) is chosen to be \( V(R_j) u_{j,j-1} \), where \( V(R_j) \) is the potential energy of the system at the point \( R_j \), which identified with the Cartesian coordinates of structure \( j \). We then have

\[ S(R_0, R_{M+1})_L = \frac{1}{L-1} \sum_{j=1}^{M} V(R_j) \Delta l_j. \]  

Since a rigid-body translation or rotation will change the distance between two structures in the full Cartesian space, an additional penalty function is introduced. For a given structure \( j \), the form is [6]

\[ \Delta l_j = [R_j^y \times M(R_j - R_{j-1}^y) + M(R_j - R_{j-1}^y)] \],

where \( M \) denotes the diagonal matrix of the atomic masses, \( R_j \) is the current position vector of structure \( j \) and \( R_{j-1}^y \) is the position vector of \( j \) at the beginning of the calculation; the cross denotes a vector product. Thus, the function to be minimized in finding the reaction path has the form

\[ T(R_0, R_{M+1})_L = S'(R_0, R_{M+1})_L + \sum_{j=0}^{M} \lambda' (\Delta l_j)^2, \]  

where the constant \( \lambda' \) should be chosen as large as possible while still maintaining numerical efficiency.

It is important to point out that the method proposed in eqs. (4)-(7) defines the reaction path by minimizing the average value of the potential energy along the path. The more standard approaches are usually concerned with finding the path with the lowest barrier [1-3]. If the potential surface is simple and well behaved (i.e., there is a single sharp col which connects the two low-energy structures), the two types of approaches will give the same or similar results. In more complex or pathological cases, where there are several possible paths, the present method will find one that does not necessarily go over the lowest barrier. For large systems like proteins or other macromolecules, more than one path is likely, in
analogy with the existence of multiple minima on the potential surface in the neighborhood of the average structure [7]. Thus, the transition path may depend on the initial guess and it may be useful to try several alternatives. In spite of such complications, the present method is expected to be useful since it is the only systematic way of treating transitions in large molecules. Once a path has been determined, gradient optimization of the transition state [8] may be useful.

3. Results

To illustrate the method we apply it to conformational transitions in three different molecules. The first is cyclohexane (54 degrees of freedom), in which we examine the transition from the chair to the twist-boat minimum-energy structures; the second is a transition between two minima in dianaline (45 degrees of freedom; the CH₃ and CH groups are treated as extended atoms [9]); and the third is the transition between two of the minima found in a quenched molecular dynamics simulation of metmyoglobin (4593 degrees of freedom; the CH₃, CH₂, and CH and SH groups are treated as extended atoms [9]). Since the first two systems, cyclohexane and dianaline, are molecules which can be studied by other approaches, they serve as tests of the algorithm. Myoglobin is used to demonstrate the relatively weak dependence of the proposed method on dimensionality and its applicability to molecules with several thousands degrees of freedom.

In all cases we use as the initial guess for the transition path a simple straight-line interpolation in Cartesian coordinate space; that is,

$$R_i = R_0 + j\Delta R, \quad \Delta R = (R_0 - R_{M+1})/(M+1).$$

(8)

This is a reasonable choice when a better zero-order guess for the reaction coordinate is difficult to obtain, as is often true for macromolecules. The calculations were performed with a version of the program CHARMM [9] adapted for this problem.

3.1. Cyclohexane

Cyclohexane is known to have two minimum-energy structures, a chair and a twist-boat form. A number of studies of the transition between the two structures have been made by both theoretical and experimental methods [10,11]. The most detailed study is that by Picket and Strauss [10] who used a restricted conformational space, i.e. they expressed the path in terms of three pseudorotation angles, while keeping fixed the other degrees of freedom (bond lengths and bond angles). Here we consider the problem in the full conformational space of the molecule (54 degrees of freedom). The calculated energy of the twist-boat minimum is 6.48 kcal/mol, relative to the chair form. The values used for λ (eq. (4)) and λ' (eq. (7)) were 900 (with distances in Å, masses in amu, and energy in kcal/mol); ten intermediate points were employed. The linear interpolation results in a barrier of 130 kcal/mol (relative to the chair form), which is reduced to 11.5 kcal/mol by optimization. This value is close to the experimental estimate (10 to 11 kcal/mol) and the results of other calculations (10 to 14 kcal/mol); differences with the latter reflect the choice of potential and not the accuracy of the optimization method. The error in the barrier height due to the discretization of the path is, in fact, less than 0.2 kcal. This value was determined by minimizing the square of the potential-energy gradient [8] for the structure found by the reaction path method to be closest to the barrier; the final total potential-energy gradient was less than $10^{-4}$ kcal/mol Å.

To measure the structural difference between the initial guess and the optimized path, the root-mean-square (rms) value for each point $j$ was calculated, i.e.

$$\text{rms}_j = [(1/N)(R_j - R_j^0)^2]^{1/2},$$

(9)

where $R_j$ and $R_j^0$ are the final and initial position vectors, respectively. The maximum value of (rms), was 0.2 Å, and the total summed over all points was 1.48 Å. The rms difference between the structure at the barrier obtained by the path optimization and the potential gradient minimization was 0.07 Å. The transition state has $C_s$ symmetry; the ring dihedral angles are $(-12.40^\circ, -2.52^\circ, -12.84^\circ, 42.26^\circ, -58.15^\circ, 41.93^\circ)$ in the linear interpolation and $(-6.84^\circ, -13.53^\circ, -7.35^\circ, 47.76^\circ, -68.67^\circ, 47.25^\circ)$ in the optimized structure.

The present results are similar to those of Picket and Strauss [10]. However, there is a significant difference due to the fact that they kept bond lengths...
and bond angles constant; the bond lengths and bond angles for structures along the path had maximum deviations of 0.02 Å and 7°, respectively, from the initial (chair) and final (twist-boat) structures.

3.2. Dialanine

As a second example, we consider the transition between two low-energy conformations of a small zwitterionic peptide, dialanine, with the formula NH₃⁺ CH₂CONHCH₂CO²⁻. The conformations can be specified by the values of the central dihedral angles ψ and φ, the relevant soft degrees of freedom for the system. Fig. 1a is a picture of dialanine that defines the coordinates, and fig. 1b shows the energy map for ψ and φ generated by an adiabatic mapping procedure (i.e. the energy of the peptide was minimized for all degrees of freedom subject to constraints on ψ and φ at 10° intervals). There are two stable conformations at (φ, ψ) equal to (−58.25°, 75.33°) and (38.55°, −47.97°); the minimum-energy path determined by interpolating the contour map is also indicated in fig. 1b.

Starting with the two minima and a straight-line interpolation, the optimized path was determined using ten intermediate points; the constraint values were λ = λ’ = 1000. The starting configuration had a barrier of 153 kcal/mol, while the final barrier is 5.8 kcal/mol relative to the deeper well. The barrier estimated from a potential gradient optimization of the best structure is 5.3 kcal/mol. The conformation (ψ, φ) at the rather flat barrier (see fig. 1b) is (−18.68°, −13.92°) from the optimized path, and (−19.08°, −11.31°) from the gradient method. The rms difference between the straight-line interpolation and the optimized path is shown in fig. 1c.

3.3. Myoglobin

Myoglobin, which can be represented as 1531 heavy atoms and polar hydrogens is the protein which stores oxygen in the muscles (fig. 2a). In a recent quenched molecular dynamics study [7], it was shown that the molecule moves on a complex multimimum potential surface in the neighborhood of the native structure; e.g. approximately 2000 minima with minimum rms differences on the order of 0.2 Å were sampled in a 300 ps simulation at 300 K. The structural changes between any two minima cannot be assigned in a few internal coordinates; instead they are distributed over large portions of the molecule and include motions in the loops and the turns which connect the helices and displacements of the helices with respect to each other. To further

Fig. 1. Dialanine. (a) Structure of molecule indicating the dihedral angles ψ and φ. (b) A contour plot of the energy as a function of ψ and φ with minimization of the rest of degrees of freedom. The linear interpolation (- - -) and optimal path (----) between the two stable configurations are shown. The contour lines represent energy differences of 1 kcal/mol. (c) All atom rms difference between linear interpolation and optimized path as a function of the structure index along the path.
characterize the potential surface it is necessary to analyze the pathways and potential-energy barriers between the minima; the latter determine the dynamic behavior of the molecule at low temperatures and govern a possible transition to a glassy state [7,13].

We examine here the path between two of the minimum-energy conformations that differ in energy by 4.4 kcal/mol and in rms coordinate values by 0.36 Å. A straight-line interpolation with ten grid points was used and $\lambda=\lambda'=1000$ was chosen. The linear interpolation path had a barrier of 1398 kcal/mol. This was reduced to 6.4 kcal/mol relative to the lower-energy minimum by the optimization procedure. The rms difference for the path (averaged over the atoms and over the different structures) is 0.15 Å relative to the initial path; individual structures have deviation as large as 0.26 Å.

The reaction path is not local and the changes are distributed over many atoms, though not over the entire protein. The transition is dominated by the motion of Arg 45 (CD3), a disordered residue in the CD loop of the MbCO crystal structure [14]. A 180° rotation of the side chain induces a large displacement of close residues (e.g., Phe 46 (CD4) and Asp (CD2)). Residues which are distant along the peptide chain (e.g., Lys 96 (FG2) and Lys 98 (FG4)) in the FG corner interact strongly with the CD loop and hence the motion of the FG corner follows that of the C helix and CD loop in response to Arg 45 rotation. The motion approximately preserves the distances between the main-chain atoms of FG and CD. Other residues that interact strongly with Arg 45 are Leu 61 (E4) and His 64 (E7) of the E helix, which undergoes a significant displacement. Finally the propionic acid groups of the heme change their position and induce a motion of the heme group and of the proximal histidine. The differences between the initial guess for the transition state and the optimized estimate is shown in fig. 2b; it is a plot of the rms displacement between the two structures for the individual residues averaged over the atoms in each residue. Though the global change is not large (0.26 Å) there are changes in residue positions after minimization which are close to 1 Å (in the CD corner) and changes on the order of 0.5 Å are found in the E helix. The changes tend to be concentrated in the regions of the protein where the difference between the initial and final structure is greatest though significant motions of several hundred atoms occur.

4. Conclusion

The proposed algorithm for the analysis of reac-
tion paths in complex systems has been shown to be applicable to molecules as large as proteins. This is particularly important since structural transitions in such systems are known to play a functional role [15]. The time required for a full optimization for $M$ structures and $N$ degrees of freedom varies approximately as $(M + N)^2$ and the algorithm is suitable for implementation on parallel and vectorized computers. Since the minimization procedure employed in the present work is limited to finding a local minimum for the reaction path, more global minimization techniques may be useful. One such approach is simulated annealing [16], which has been shown recently to be a powerful tool in a macromolecular X-ray structure refinement when implemented with molecular dynamics [17].

Acknowledgement

We wish to thank Jeremy Smith for helpful discussions and for providing the all hydrogen parameters used for cyclohexane.

References


