Finding pathways between distant local minima

Joanne M. Carr, Semen A. Trygubenko, and David J. Wales

University Chemical Laboratories, Lensfield Road, Cambridge CB2 1EW, United Kingdom

(Received 2 March 2005; accepted 13 April 2005; published online 17 June 2005)

We report a new algorithm for constructing pathways between local minima that involve a large number of intervening transition states on the potential energy surface. A significant improvement in efficiency has been achieved by changing the strategy for choosing successive pairs of local minima that serve as endpoints for the next search. We employ Dijkstra’s algorithm [E. W. Dijkstra, Numer. Math. 1, 269 (1959)] to identify the “shortest” path corresponding to missing connections within an evolving database of local minima and the transition states that connect them. The metric employed to determine the shortest missing connection is a function of the minimized Euclidean distance. We present applications to the formation of buckminsterfullerene and to the folding of various biomolecules: the B1 domain of protein G, tryptophan zippers, and the villin headpiece subdomain. The corresponding pathways contain up to 163 transition states and will be used in future discrete path sampling calculations. © 2005 American Institute of Physics.

[DOI: 10.1063/1.1931587]

I. INTRODUCTION

The discrete path sampling (DPS) approach provides a means to sample pathways corresponding to “rare events” using a coarse-grained framework based on the underlying potential energy surface (PES). The PES can be formally partitioned into the catchment basins of local minima, while rate constants for transitions between these basins can be estimated using the statistical rate theory once transition states on the basin boundaries have been found. Here we adopt the geometrical definition of a transition state as a stationary point on the PES with a single negative Hessian eigenvalue, following Murrell and Laidler. Minima are stationary points with no negative Hessian eigenvalues.

Two minima may be connected via a single transition state if they are sufficiently close to each other in configuration space. More generally, local minima may interconvert via any number of multistep paths, and the minimum number of elementary steps will depend upon the isomers in question. Here we consider an elementary step as a rearrangement that involves a single transition state. Locating multistep pathways on a complex PES can be a difficult task. Double-ended methods usually do not reveal all the transition states at once, mainly because of the existence of multiple barrier height and path length scales. Consecutive double-ended searches have proved to be an efficient way of working around the above problem, and this strategy was adopted in our previous work. An essential part of this approach is a mechanism to incorporate the information obtained in all the previous searches into the next one. Various strategies can be adopted, the most general and effective being the one based on the Euclidean separation, and we refer the reader to the original publications for details.

A guess for the initial pathway is required for most double-ended searches. Linear interpolation is a straightforward way to automate this part of the calculation. For large endpoint separations guessing the initial pathway can be difficult, and there is a large probability of finding many irrelevant stationary points at the beginning of the calculation.

Due to the discrete nature of the double-ended methods, such as nudged elastic band (NEB) and doubly nudged elastic band (DNEB), which use a series of images to represent the path, the candidate transition states obtained from such searches usually need to be optimized further. For this purpose we employ eigenvector-following methods. The connectivity is obtained by following the two unique steepest-descent (SD) paths downhill from each transition state. In our calculations the refinement of transition states and the calculation of approximate steepest-descent paths are the most time-consuming steps.

The connection algorithm described in Ref. 37 uses one double-ended search per cycle. However, we have found that this approach can be overwhelmed by the abundance of stationary points and pathways for complicated rearrangements. In the present contribution we introduce the idea of an unconnected pathway and make the connection algorithm more focused by allowing more than one double-ended search per cycle. This approach, in combination with some other modifications described in Sec. II, greatly reduces the computational demands of the method and has allowed us to tackle more complicated problems. The new algorithm has been implemented within our OPTIM package, and a public domain version will be made available for download from the internet. The examples presented below each involve the search for an initial connection between distant endpoints for use in subsequent DPS calculations, which will be described in future publications. The DPS approach is a coarse-grained analog of dynamical path sampling based upon discrete

---

1Electronic mail: dw34@cam.ac.uk
paths of connected minima and the intervening transition states. The initial paths considered in the present work provide a starting point for such calculations, where new paths are generated by successively replacing segments of an evolving reference path. The final result is a database of local minima and transition states, which can then be subjected to kinetic analysis in various ways to extract phenomenological rate constants.

II. METHODS
A. Outline of the connection algorithm

A detailed description of the previous version of our connection algorithm was provided elsewhere. Each iteration consists of one double-ended transition state search, followed by refinement of transition state candidates using eigenvector-following, and calculation of approximate steepest-descent paths for each transition state to establish the connectivity. Double-ended transition state searches employed the doubly nudged elastic band algorithm to provide a guess for the transition state between a given pair of minima. For multistep paths there will be more than one such structure, and each of these geometries is subsequently used as the starting point for a hybrid eigenvector-following calculation to provide tightly converged transition states. The paths connecting each transition state to the two corresponding minima were calculated by energy minimization using a modified version of Nocedal’s L-BFGS algorithm.

We classify all the known local minima into three categories: \( S \) minima, which are connected to the starting structure, \( F \) minima, which are connected to the final structure, and \( U \) minima, which are so far unconnected to the \( S \) and \( F \) sets. \( U \) minima can be connected between themselves but not to any \( S \) or \( F \) minima. When a connection is established between members of the \( U \) and the \( S \) or \( F \) sets, the unconnected minimum and all the minima connected to it become members of \( S \) or \( F \), respectively. If a connection is found between members of the \( S \) and \( F \) sets then the algorithm terminates.

Before each cycle a decision must be made as to which minima to try and connect next. Various strategies can be adopted, for example, selection based on the order in which transition states were found or selection of minima with the smallest separation in Euclidean distance space. However, when the endpoints are very distant in configuration space, neither of these approaches is particularly efficient. The number of possible connections that might be tried simply grows too quickly if the \( S \), \( F \), and \( U \) sets become large. However, the new algorithm described below seems to be very effective.

B. A Dijkstra-based selector

The modified connection algorithm we have used in the present work is based on a shortest path method proposed by Dijkstra. We can describe the minima that are known at the beginning of each connection cycle as a complete graph, \( G=(M,E) \), where \( M \) is the set of all minima and \( E \) is the set of all the edges between them. Edges are considered to exist between every pair of minima \( u \) and \( v \), even if they are in different \( S \), \( F \), or \( U \) sets, and the weight of the edge is chosen to be a function of the minimum Euclidean distance between them.

\[
w(u, v) = \begin{cases} 0, & \text{if } u \text{ and } v \text{ are connected via a single transition state,} \\ \infty, & \text{if } n(u, v) = n_{\text{max}}, \\ f(D(u, v)), & \text{otherwise,} \end{cases}
\]

where \( n(u, v) \) is the number of times a pair \((u, v)\) was selected for a connection attempt, \( n_{\text{max}} \) is the maximal number of times we may try to connect any pair of minima, and \( D(u, v) \) is the minimum Euclidean distance between \( u \) and \( v \). \( f \) should be a monotonically increasing function, such as \( f(D(u, v)) = D(u, v)^2 \). We denote the number of minima in the set \( M = S \cup U \cup F \) as \( m \) and the number of edges in the set \( E \) as \( e = m(m-1)/2 \).

Using the Dijkstra algorithm and the weighted graph representation described above, it is possible to determine the shortest paths between any minima in the database. The source is selected to be one of the endpoints. Upon termination of the Dijkstra algorithm, a shortest path from one endpoint to the other is extracted. If the weight of this pathway is nonzero, it contains one or more “gaps.” Connection attempts are then made for every pair \((u, v)\) of adjacent minima in the pathway with nonzero \( w(u, v) \) using the DNEB approach.

The computational complexity of the Dijkstra algorithm is at worst \( O(m^2) \), and the memory requirements scale in a similar fashion. The most appropriate data structure is a weighted adjacency matrix. For the calculations presented in this paper, the single-source shortest path problem was solved at the beginning of each cycle, which altogether took less than 10% of the total execution time for the largest database encountered.

We emphasize here that once an initial path has been found, the perturbations considered in typical DPS calculations will generally involve attempts to connect minima that are separated by far fewer elementary rearrangements than
the endpoints. It is also noteworthy that the initial path is unlikely to contribute significantly to the overall rate constant. Nevertheless, it is essential to construct such a path to begin the DPS procedure.

The nature of the definition of the weight function allows the Dijkstra algorithm to terminate whenever a second endpoint, or any minimum connected to that endpoint via a series of elementary rearrangements, is reached. This observation reduces the computational requirements by an amount that depends on the distribution of the minima in the database among the S, U, and F sets. One of the endpoints is always a member of the S set, while the other is a member of the F set. Either one can be chosen as the source, and we have found it most efficient to select the one from the set with fewest members. However, this choice does not improve the asymptotic bounds of the algorithm.

### III. RESULTS

#### A. Buckminsterfullerene

Various suggestions have been made for the formation mechanism of buckminsterfullerene in the gas phase. In total, there are 1812 different C60 fullerene isomers, which probably interconvert via the “pyracylene” or “Stone–Wales” (SW) rearrangement. This process has been investigated in several previous studies, and the most accurate calculations yield a picture of the energy landscape that is quite similar to that obtained using a tight-binding potential. The latter model was therefore used in this initial pathway calculation to minimize the computational expense, which is significantly greater than for analytical empirical potentials. The present calculation therefore has most in common with the annealing study of Xu and Scuseria, although the latter work did not involve transition state calculations.

An initial high-energy starting point was constructed by simply placing 60 carbon atoms in a container and minimizing the energy. The resulting structure, which contains a number of large rings and chains, including polyyactylene fragments, is shown in Fig. 1. Although the structure of the buckminsterfullerene endpoint is well known, it is not at all clear which would be the best permutational isomer to attempt a connection with, since the endpoints are so different. The distance between these structures was minimized with respect to permutation-inversion isomers, center of mass, and orientational coordinates. However, this connection still presents a significant challenge and required 383 cycles of the Dijkstra-based connection algorithm, including 1620 DNEB searches. V is the potential energy in hartree and s is the integrated path length in bohr.

The nature of the definition of the weight function allows the Dijkstra algorithm to terminate whenever a second endpoint, or any minimum connected to that endpoint via a series of elementary rearrangements, is reached. This observation reduces the computational requirements by an amount that depends on the distribution of the minima in the database among the S, U, and F sets. One of the endpoints is always a member of the S set, while the other is a member of the F set. Either one can be chosen as the source, and we have found it most efficient to select the one from the set with fewest members. However, this choice does not improve the asymptotic bounds of the algorithm.

### B. GB1 hairpin

The GB1 hairpin consists of residues 41–56 from the C-terminal fragment of the B1 domain of protein G. It forms a β hairpin both in the complete protein and for the isolated fragment in solution. A number of previous experimental and theoretical studies have been conducted for this system, including a DPS investigation. Obtaining an initial path between unfolded and hairpin conformations in the latter study was not an easy task, and the new algorithm speeded up this part of the calculation by at least a couple of orders of magnitude.

The CHARMM program has previously been interfaced to our OPTIM code, which includes a wide variety of algorithms for locating stationary points and characterizing pathways, and now includes the Dijkstra-based connection approach. As for the previous DPS study we employed the investigation of paths involving nonfullerene C60 isomers may be worthwhile, although we note once again that this initial path may not be dynamically significant.

![Fig. 1. Top: connected path between a high-energy C60 cluster and buckminsterfullerene. The path contains 82 transition states, and required 383 cycles of the Dijkstra-based connection algorithm, including 1620 DNEB searches.](image-url)
CHARMM19 force field with the EEF1 implicit solvation model. The path illustrated in Fig. 2 consists of 163 transition states and was found in 126 cycles of the Dijkstra-based algorithm. It connects a partially collapsed non-native structure, with two turns, to a conformation from the native hairpin ensemble.

C. Tryptophan zippers

Tryptophan zippers are stable fast-folding β hairpins designed by Cochran et al., which have recently generated considerable interest. In the present work we have obtained native to denatured state rearrangement pathways for five tryptophan zippers: trpzip 1, trpzip 2, trpzip 3, trpzip 3-I, and trpzip 4. The notation is adopted from the work of Du et al. All these peptides contain 12 residues, except for trpzip 4, which has 16. Tryptophan zippers 1, 2, 3, and 3-I differ only in the sequence of the turn. Experimental measurements of characteristic folding times for these peptides have shed some light on the significance of the turn sequence in determining the stability and folding kinetics of peptides with the β-hairpin structural motif.

To model these molecules we used a modified CHARMM19 force field, with symmetrized asparagine (ASN), glutamine (GLN) and tyrosine (TYR) dihedral angle and CTER improper dihedral angle terms, to ensure that rotamers of these residues have the same energies and geometries. Another small modification concerned the addition of a nonstandard amino acid, D-proline, which was needed to model trpzip 3. The implicit solvent model EEF1 was used to account for solvation, with a small change to the original implementation to eliminate discontinuities.

We have used the Dijkstra-based connection algorithm to obtain folding pathways for all five trpzip peptides. In each case the first endpoint was chosen to be the native state structure, which, for 1, 2, and 4 trpzips, was taken from the Protein Data Bank (PDB). There are no experimental structures available for 3 and 3-I, so for these peptides the first endpoint was chosen to be the putative global minimum obtained using the basin-hopping method. The second endpoint was chosen to be an extended structure, which was obtained by simply minimizing the energy of a conformation with all the backbone dihedral angles set to 180°. All the stationary points (including those obtained during the connection procedure) were tightly converged to reduce the root-mean-squared force below 10⁻¹⁰ kcal mol⁻¹ Å⁻¹.

Each of the five trpzip pathway searches was conducted on a single Xeon 3.0-GHz CPU and required less than 24 h of CPU time. The timings could certainly be improved by optimizing the various parameters employed throughout the searches. However, it is more important that the connections actually succeed in a reasonably short time. It only requires one complete path to seed a DPS run, and we expect the DPS...
states and required 142 cycles of the Dijkstra-based algorithm. In this rearrangement, which connects a partially folded minimum to the locally minimized PDB structure, helices 2 and 3 are completed, the C-terminal residues pack correctly between helices 1 and 3, and the three helices adopt their native relative orientations.

IV. CONCLUSIONS

Discrete path sampling calculations of effective two-state rate constants require an initial path consisting of local minima and the transition states that connect them.\(^1\)\(^3\)\(^116\)\(^147\)\(^148\) For complex rearrangements the number of elementary steps involved may be rather large, and new methods for constructing an initial path are needed. This path does not need to be the shortest, or the fastest, but it does need to be fully connected. In the present work we have described a connection procedure based upon Dijkstra’s shortest path algorithm, which enables us to select the most promising paths that include missing connections for subsequent double-ended searches. We have found that this approach, which is now implemented within the OPTIM package,\(^52\) enables initial paths containing more than a hundred steps to be calculated automatically for a variety of systems. Some typical results have been presented for buckminsterfullerene, trpzip peptides, the GB1 hairpin, and the villin headpiece subdomain. These paths will be employed to seed future discrete path sampling calculations.

ACKNOWLEDGMENTS

We gratefully acknowledge Dr. David Evans for his Dijkstra-based utilities that locate shortest paths in connected databases of minima and transition states. One of the authors (S.A.T.) is a Cambridge Commonwealth Trust/Cambridge Overseas Trust scholar. Most of the tryptophan zippers calculations were performed using computational facilities funded by the Isaac Newton Trust. Another author (J.M.C.) is grateful to the EPSRC for financial support.

\(^{1}\)D. J. Wales, Mol. Phys. 100, 3285 (2002).