JCTC Journal of Chemical Theory and Computation

Dual Grid Methods for Finding the Reaction Path on Reduced Potential Energy Surfaces

Steven K. Burger and Paul W. Ayers*

Department of Chemistry & Chemical Biology, McMaster University,
1280 Main St. West, Hamilton, Ontario, Canada

Received January 7, 2010

Abstract: Two new algorithms are presented for determining the minimum energy reaction path (MEP) on the reduced potential energy surface (RPES) starting with only the reactant. These approaches are based on concepts from the fast marching method (FMM), which expands points outward as a wavefront on a multidimensional grid from the reactant until the product is reached. The MEP is then traced backward to the reactant. Since the number of possible grid points that must be considered grows exponentially with increasing dimensionality of the RPES, interpolation is important for maintaining manageable computational costs. In this work, we use Shepard interpolation, which we have modified to resolve problems in overfitting. In contrast to FMM, which accurately locates the MEP, the new algorithms focus on locating the single rate-limiting transition state and provide only a rough estimate of the MEP. They do this by mapping out the RPES on a coarse grid and then refining a least action path on a finer grid. This is done so that the majority of the interpolation is done on the finer grid, which minimizes the amount of extrapolation inherent in an outward searching algorithm. The first method scans the entire PES before iteratively locating the transition state (TS) for the MEP on the lower bound estimate of the fine PES. The second method explores the coarse grid in a similar manner to FMM and then iteratively locates the rate-limiting TS in the same manner as the first method. Both methods are shown to be capable of rapidly obtaining (in less than 30 constrained optimization cycles) an approximation to the MEP and the rate limiting TS for three example systems: the 4-well potential, the molecule *N*-hydroxymethyl-methylnitrosaminee (HMMN), and a cluster model of DNA-uracil glycosylase.

1. Introduction

For many chemical problems, we are interested in the kinetics of a reaction, which requires knowing the mechanism and the energy barrier. Common kinetically interesting examples are gas and solution phase molecular reactions, enzyme mechanisms,[1] and conformation changes of proteins.[2] When dealing with such systems, we would like to know all of the kinetically accessible minima and transition state (TS) structures. With this information, we can determine the reaction rate with a variety of methods such as transition state theory[3] or the reaction path Hamilton method.[3,4]

Also of interest is the minimum energy path (MEP) between the reactant and product. This is defined as the steepest descent path from the TS to each minimum and it represents the most probable path the system would take at 0 K. It can be obtained relatively easily if the TS structures are known, since it reduces the problem to an initial value problem,[5] solvable with an implicit Runge−Kutta method. However, the MEP is generally less interesting for computational chemists than the TS structures since any thermodynamic path[6] connecting the rate limiting TS to the end points will give the correct kinetics.

Finding TS structures is an optimization problem and the methods used are similar to the methods used for minimization. However, the problem is complicated by the fact that one eigenvector of the Hessian must have a negative eigenvalue. This greatly increases the difficulty of the problem,

* Corresponding author e-mail: ayers@mcmaster.ca.

10.1021/ct100012y © 2010 American Chemical Society
Published on Web 03/30/2010

since it implies that a good guess of the Hessian is available and computing the Hessian analytically is usually computational expensive. If a good initial guess is available though, then there are a number of algorithms[7−14] that can get the TS starting from a point relatively close to the solution.

When the location of the TS is difficult to approximate then, instead of using a single point to find the TS, a string of points can be used to approximate the MEP. The highest point on this path is the best guess of the TS. There are a large number of string methods available,[15−25] all of which take advantage of the fact that the MEP must be a minimum in all directions perpendicular to the path. This class of methods tends to be easy to parallelize and will give a good approximation to the entire path, from which approximate TS structures can be obtained and then refined with TS searching methods.[7,8,25−27] Two major shortcomings of these methods are that many points are needed to accurately represent the path and lower barrier paths may be missed if one starts with a poor initial guess of the path.

The difficulty in finding the MEP and TS is related to the dimensionality of the problem. Fortunately, for many chemical problems a few key coordinates[28] can be identified, such as the interatomic distance involved in bond breaking or forming. The energy can then be formulated in terms of a reduced set of coordinates by minimizing all other degrees of freedom, thus removing their contribution. An extreme case of this is the coordinate driving method, where the mechanism is determined by changing only one coordinate while minimizing all of the other degrees of freedom. However, one dimension is usually not enough to describe a reaction. As a result, discontinuities will often emerge, due to the fact that other important degrees of freedom are sensitive to very small changes in the driven coordinate.

If all of the important degrees of freedom are included in the reduced set of coordinates, then the reduced potential energy surface (RPES) will be smooth and interpolation methods can be used. Shepard interpolation[29,30] has been shown to work well for the fast-marching and string methods[31] and it is our choice for the methods outlined here. Shepard interpolation only requires values for the potential; however, the quality of the interpolation can be dramatically improved by using derivative terms as well. For chemical problems, the gradient can be computed at a similar cost to the potential, so it is usually included. Higher derivatives are usually too costly to compute directly, but they can be approximated by interpolated moving least-squares (IMLS).[32,33]

Given a good error estimate for the interpolated RPES, either a string method or the fast-marching method (FMM)[34−37] can be used to explore the surface.[31] Unlike string methods, FMM has the nice property that it exhaustively searches the RPES to find the MEP and does not require knowledge of the product state. FMM propagates a wavefront and effectively fills up the surface on a grid. Once the product is found, the MEP is obtained by integrating the steepest descent path of an action surface back from the product.

FMM is greatly improved with interpolation, but its main shortcoming is that each new point is obtained by extrapolation rather than interpolation, since the algorithm expands outward from the current set of determined points. To get around this issue, two new algorithms are developed which evaluate the points on two grids: (1) a coarse grid that allows a rough estimate of the MEP to be obtained, (2) a finer grid that is used to determine the rate limiting TS structure. While extrapolation is the only alternative to directly evaluating the potential on the coarser grid, on the finer grid, new points can be approximated with Shepard interpolation. With interpolation we show that these methods work well for three different systems of increasing complexity.

2. Theory and Computational Methods

Least Action Surface. The fast marching method (FMM),[34−37] like all reaction path methods, is based on finding the path of length q which minimizes the integral,

$$S(q) = \int_0^q f(q(\tau))d\tau \tag{1}$$

where τ is a parametrization of the arc length and $f(\tau)$ is a cost function which can be defined as follows:

$$f(q) = \left[\frac{E - V(q)}{E - V(q_{\min})}\right]^{l/2} \tag{2}$$

E is the classically highest allowed energy of the system, $V(q_{\min})$ is the lowest potential value and l determines the cost being considered. If $l = 0$ then solving eq 1 will minimize the distance between two points, while if $l = 1$ we obtain the least time path. In the limit $l \rightarrow \infty$, minimizing eq 1 results in the MEP.[35] If we differentiate eq 1 we obtain the Hamilton−Jacobi differential equation,

$$|\nabla S(q)| = f(q) \tag{3}$$

Equation 3 can be solved practically as a finite difference equation. For the 2D case, this has the form,

$$\max\left(\frac{S_{i,j} - S_{i-1,j}}{\Delta q_1}, \frac{S_{i,j} - S_{i+1,j}}{\Delta q_1}, 0\right)^2 + \max\left(\frac{S_{i,j} - S_{i,j-1}}{\Delta q_2}, \frac{S_{i,j} - S_{i,j+1}}{\Delta q_2}, 0\right)^2 = (f(q))^2 \tag{4}$$

The full details of solving eq 4 are given in ref 35. Once the action has been computed at all points $S_{i,j}$ on the 2D grid, then the MEP can be determined by integrating backward (backtracing) from the product on the least action surface. This can be done only at grid points or on an interpolated surface. For the methods in this work, we do backtracing on the grid points.

Shepard Interpolation. To reduce the number of energy and gradient evaluations, interpolation is used when the error in the interpolant is sufficiently small. How large an error is tolerable is set as a user-defined parameter. Shepard interpolation[29,30,38−42] has been used with FMM and string methods[31] and we use the same basic scheme here as well. Shepard interpolation uses a set of points $\mathbf{X}^{(i)}$ where the Taylor expansion at each point is,

$$T_n^{(i)}(\mathbf{X}) = V(\mathbf{X}^{(i)}) + (\mathbf{X} - \mathbf{X}^{(i)})\cdot\nabla V(\mathbf{X}^{(i)}) + \frac{1}{2}(\mathbf{X} - \mathbf{X}^{(i)})\cdot\nabla\nabla V(\mathbf{X}^{(i)})\cdot(\mathbf{X} - \mathbf{X}^{(i)}) + ... \quad (5)$$

such that n is the highest order of the expansion and $\mathbf{X}$ is the coordinate of the point of interest. For the methods in this work, we calculate the first two terms in the Taylor series ($V(\mathbf{X}^{(i)})$ and $\nabla V(\mathbf{X}^{(i)})$) and use weighted least-squares to determine the $p = (1)/(n!)\prod_{i=0}^{n-1}(d + i)$ components of the higher-order ($n > 1$) derivatives. Rewriting eq 5 as a weighted least-squares equation we get,

$$\min_{\mathbf{x}} \|\mathbf{W}(\mathbf{A}\mathbf{x} - \mathbf{b})\| \quad (6)$$

where $\mathbf{x} = \{\nabla^2 V(\mathbf{X}^{(i)}), \nabla^3 V(\mathbf{X}^{(i)}),...\}$ is a vector of length p containing the unknown higher order terms we are interested in; $\mathbf{b}$ is a linear combination of $V(\mathbf{X}^{(i)})$ and $\nabla V(\mathbf{X}^{(i)})$ of length $M(d + 1)$, where M is the number of neighboring points used and d is the dimensionality of the system; $\mathbf{A}$ is a matrix of the $(\mathbf{X} - \mathbf{X}^{(i)})$ terms in eq 5; and $\mathbf{W}$ is a diagonal weighting matrix. For the weight matrix, if we use the Bettens−Collins isotropic formula[39] or other similar forms,[31] then there can be problems with overfitting when too few points make significant contributions to the Shepard interpolant. Specifically, problems arise unless M is appreciably larger than $p/d - 1$. To get around this, we can change the weighting function to ensure that the weighting is more evenly distributed among the points by using the usual form for the diagonal terms,

$$w^{(i)}(\mathbf{X}) = \frac{v^{(i)}(\mathbf{X})}{\sum_{j=1}^{M} v^{(j)}(\mathbf{X})} \quad (7)$$

with,

$$v^{(i)}(\mathbf{X}) = e^{-1/2\left(\frac{\|X - X^{(i)}\|}{\sigma^{(i)}}\right)^2} \quad (8)$$

But instead of basing the trust radius $\sigma^{(i)}$ on the grid spacing, we sort the neighbors of point $\mathbf{X}^{(i)}$ based on distance and then chose the $k = p/d - 1$ element so that $\sigma^{(i)} = \|\mathbf{X} - \mathbf{X}^{(k)}\|$. This ensures that enough points are within one standard deviation of the weighting function so that overfitting does not occur.

Determining the error in the coefficients requires the residual of the least-squares fit,

$$\sigma^2 = \frac{\mathbf{b}^T\mathbf{b} - (\mathbf{A}\mathbf{x})^T\mathbf{b}}{(d + 1)M - p} \quad (9)$$

from which the covariance matrix can be determined,

$$\mathbf{V} = (\mathbf{A}^T\mathbf{A})^{-1}\sigma^2 \quad (10)$$

The error from the higher order fitted terms in the Taylor series is estimated as follows:

$$\varepsilon_T(\mathbf{X}) = \sqrt{\sum_{i=1}^{M(d+1)} \sum_{j=1}^{p} V_{ij} A(\mathbf{X})_{ij}^2} \quad (11)$$

Of course the residual in the Taylor series is still not accounted for, but usually this term is significantly smaller than the error introduced by fitting the higher order terms.

Equation 5 can be fit to any order so long as $p \le M(d + 1)$. However, overfitting will be a problem if $p/M(d + 1) \approx 1$. To determine if the next order is a good model for the surface, we can check to see if there is a "lack of fit".[43] This is generally done by testing the general linear hypothesis, $\mathbf{b}_{p-q...p} = \mathbf{0}$. To test this hypothesis, the ratio,

$$\frac{\left(\dfrac{\sigma_{n-1}^2 - \sigma_n^2}{q}\right)}{\left(\dfrac{\sigma_n^2}{n - p}\right)} \quad (12)$$

is compared against the $F(q, n - p)$ distribution.[44] In eq 12 σ_n^2 is the standard deviation when derivatives are fit up to the order n. Unfortunately, we found that this method does not work well for the problems we considered. Instead, we used a more practical method, leave one out cross-validation (LOOCV). In this scheme, each of the M neighbor points used in the fitting is left out in sequence while the remaining M-1 points are used to fit $\mathbf{x} = \{\nabla^2 V(\mathbf{X}^{(i)}), \nabla^3 V(\mathbf{X}^{(i)}),...\}$. Each point left out is used to estimate one term of a weighted mean squared error.

$$\mathrm{MSE}(n) = \sum_{k=1}^{M} w^{(i)}(\mathbf{X}^{(k)})(V(\mathbf{X}^{(k)}) - T_n^{(i)}(\mathbf{X}^{(k)}))^2 \quad (13)$$

This is done for each order between 2 and 5. The derivatives are fit up to and including the order which gave the lowest value for eq 13.

Once the higher order terms have been fit, the potential at any point $\mathbf{X}$ on the surface is obtained by the sum,

$$\tilde{V}(\mathbf{X}) = \sum_{i=1}^{M} w^{(i)}(\mathbf{X}) T^{(i)}(\mathbf{X}) \quad (14)$$

where $T^{(i)}(\mathbf{X})$ are given by eq 5, M is the number of neighbor points, and $w^{(i)}(\mathbf{X})$ is the weight function. The error for this sum can be estimated as follows:

$$\varepsilon_{\tilde{V}}(\mathbf{X}) = \sqrt{\sum_{i=1}^{M} w^{(i)}(\mathbf{X})(\tilde{V}(\mathbf{X}) - T_n^{(i)}(\mathbf{X}))^2} \quad (15)$$

where the same weight function from ref 31 is used in both eqs 14 and 15. This works well when one point does not dominate the sum, which we define as $w^{(j)}(\mathbf{X}) > 0.9$. If one term does dominate, then eq 15 will likely underestimate the error, and eq 11 will generally be a better estimate.

Dual Grid−Low Path Methods. Both algorithms are based on a dual grid approach. The methods differ mainly in their treatment of the coarse grid. In the low path method, one evaluates the energy and gradient at every point on the coarse grid. The boundary low-path method, by contrast, evaluates points on the coarse grid in a similar way to FMM, avoiding points that lie outside the boundary where the error in extrapolation is too large.

We denote the upper and lower bounds on the d coordinates under consideration as **ub** and **lb**. The reactant and product configurations are $\mathbf{R}_{\mathrm{react}}$ and $\mathbf{R}_{\mathrm{prod}}$, respectively. The coarse grid consists of the set of points, $X_i^{(k)} = R_{\mathrm{react},i} + c_i \Delta R_{\mathrm{large},i}$, for which $lb_i < X_i^{(k)} < ub_i$. Here $1 \le i \le d$ denotes the particular coordinate of interest, c_i is an integer, and $\Delta\mathbf{R}_{\mathrm{large}}$ is the vector of grid spacings for the coarse grid. Similarly, points on the fine grid are defined by $X_i^{(k)} = R_{react,i} + c_i \Delta R_{small,i}$, where $\Delta\mathbf{R}_{\mathrm{small}}$ is the vector of grid spacings for the small grid. A parameter, α, is used to construct a lower error bound on the true RPES. In keeping with our previous notation, values of the potential energy evaluated using computational chemistry software are given as $V(\mathbf{X}^{(k)})$ and interpolated values of the potential are $\tilde{V}(\mathbf{X}^{(k)})$.

At first, the algorithm constructs an interpolation of the full RPES on the coarse grid and then locates the best guess at the TS. Next we set $\alpha = 1$, so that the error is subtracted from the value of the interpolant; this provides an (approximate) lower bound to the true potential energy. The TS is then located on the lower-bound surface. If the TS is located at a grid point which has already been evaluated (i.e., where the error is zero), then we identify this conformation as the rate limiting TS. Otherwise, the energy and gradient are evaluated at this point, the surface is reinterpolated and the process repeats.

The transition-state estimate will be accurate, up to grid-spacing of the fine grid, as long as: (a) the coarse grid is fine enough that alternative pathways are not missed and (b) the error estimate of the interpolated potential is not underestimated. The full algorithm is as follows:

Algorithm 1: Low Path Method (LPM).

(a) Set $\Delta\mathbf{R}_{\mathrm{large}}$, $\Delta\mathbf{R}_{\mathrm{small}}$, **lb**, **ub**, $\mathbf{R}_{\mathrm{react}}$, $\mathbf{R}_{\mathrm{prod}}$, E_{barr}, and $\alpha = 0$.

(b) Coarsely scan the RPES evaluating $V(\mathbf{X}^{(k)})$ and $\nabla V(\mathbf{X}^{(k)})$ at the points $X_i^{(k)} = R_{\mathrm{react},i} + c_i \Delta R_{\mathrm{large},i}$ which satisfy $lb_i < X_i^{(k)} < ub_i$ for $i = 1...d$ and $c_i \in \mathbb{Z}$.

(c) Fit the higher-order derivatives of the potential to the evaluated points using eq 6.

(d) Interpolate all unevaluated points $X_i = R_{\mathrm{react},i} + c_i \Delta R_{\mathrm{small},i}$, $lb_i < X_i < ub_i$ on the fine grid to get $\{V(\mathbf{X}),\varepsilon(\mathbf{X})\}$ where $\varepsilon(\mathbf{X})$ is the error given by either eq 11 or eq 15. For evaluated points set $\varepsilon(\mathbf{X}) = 0$.

(e) Determine the action S at each point on the fine grid by solving eq 4, with $f(\mathbf{X}) = [(E - V(\mathbf{X}) + \alpha\varepsilon(\mathbf{X}))/(E - V_{\min}(\mathbf{X}))]^{1/2}$, where E is an upper-bound estimate of the potential at the highest TS.

(f) Backtrace from $\mathbf{R}_{\mathrm{prod}}$ on the action surface to get the MEP.

(g) If the highest point X_{TS} on the MEP is evaluated, then i. if α=1 THEN STOP; ELSE set α=1.

(h) Evaluate $\{V(X_{\mathrm{TS}}),\nabla V(X_{\mathrm{TS}})\}$. Set $\varepsilon(X_{\mathrm{TS}}) = 0$. GOTO (c).

The algorithm first iterates until the rate limiting TS is found on the interpolated surface without consideration of the error. Then with $\alpha = 1$, a lower bounded RPES is used to find the TS within the accuracy of the grid. This can be skipped if the coarse grid size is sufficiently small, but otherwise alternate paths with lower energy TS structures may be missed.

The low-path method (LPM) works well when we are interested in the full RPES, **lb** < **R** < **ub**. Often, however, there are large regions of conformation space where the potential energy is too high to be of interest. It would be more efficient not to explore those regions. FMM is particularly good at only exploring the low-energy regions, so we propose a second algorithm, called the boundary low-path method (BLPM) that (a) uses FMM to explore the RPES until the product state is located and then (b) uses the same methodology as LPM to find the TS on the fine grid.

The key new idea in the BLPM is the construction of a boundary set, B, on the fine grid that separates the region of the potential energy surface where the interpolant is sufficiently accurate from the region where the interpolation cannot be trusted. A point on the fine grid, $\mathbf{X}^{(i)}$, is a boundary point if it satisfies our error criterion ($\varepsilon(\mathbf{X}^{(i)}) < \varepsilon_{\max}$), but one of its neighbors on the fine grid does not.

Algorithm 2: Boundary Low Path Method (BLPM).

(a) Set $\Delta\mathbf{R}_{\mathrm{large}}$, $\Delta\mathbf{R}_{\mathrm{small}}$, **lb**, **ub**, $\mathbf{R}_{\mathrm{react}}$, $\mathbf{R}_{\mathrm{prod}}$, E_{barr}, $\varepsilon_{\max}$, and set $\alpha = 0$.

(b) Follow steps (c)−(e) in Algorithm 1 to interpolate the RPES and to get the action values.

(c) Construct the boundary set:

$$B = \{\mathbf{X}^i | \varepsilon(\mathbf{X}^i) < \varepsilon_{\max}, \exists \mathbf{X}^k : |X_j^i - X_j^k| \le \Delta R_{\mathrm{small},j}, \\ j = 1...d, \varepsilon(\mathbf{X}^k) > \varepsilon_{\max}\}$$

(d) Take the element of B which has the lowest action $b_{\min} = \arg\min_b\{S(b), b \in B\}$.

(e) If $S(b_{\min}) > S(\mathbf{R}_{\mathrm{prod}})$, then set α=1 and GOTO to Algorithm 1, starting at step (f).

(f) Evaluate the nearest neighbors of $b_{\min}$ on the course grid to obtain the set, $\{(V(\mathbf{X}^k),\nabla V(\mathbf{X}^k)) | |b_{\min,j} - X_j^k| \le \Delta R_{\mathrm{large},j}, j = 1...d\}$. If all of the surrounding points on the larger grid are evaluated then only evaluate $\{V(b_{\min}),\nabla V(b_{\min})\}$.

(g) GOTO (b).

This method has fewer initial evaluations but has drawbacks. Since BLPM evaluates points outward from the reactant on the RPES in a similar fashion to FMM, the error in the approximated energies tends to be larger during this first step because more extrapolation is used. Also it may be more difficult to parallelize than LPM since the grid points that need to be evaluated are less predictable.

3. Results and Discussion

To compare BLPM and LPM, we examined the 4-well potential, a gas phase molecular dissociation and a cluster model of an enzyme. All systems were done in two-dimensions so they could be visualized. The code is broken up into a number of Fortran 90[45] programs, which communicate with external files. For the energy and gradient calculations, we used system calls to Gaussian03[46] and for the cost function we used $l = 15$.

The 4-Well Potential. Analytic systems can be good at demonstrating flaws in certain methods. The 4-well potential provides an example for how string methods can fail when starting with a linear interpolation as an initial guess of the path. The highest TS on the MEP is located at (−0.274223,

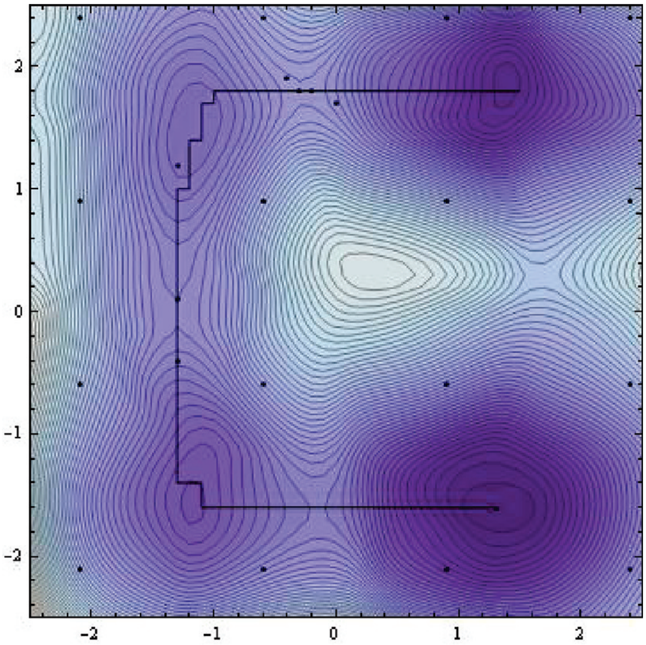

Figure 1. A contour plot of the four-well potential with Shepard interpolation using the fine grid spacing $\Delta\mathbf{R}_{\mathrm{small}} = (0.1,0.1)$. The evaluated points from LPM are shown as black dots. The backtrace path on the grid is the black curve. LPM converges after 24 evaluations to within the accuracy of the fine grid.

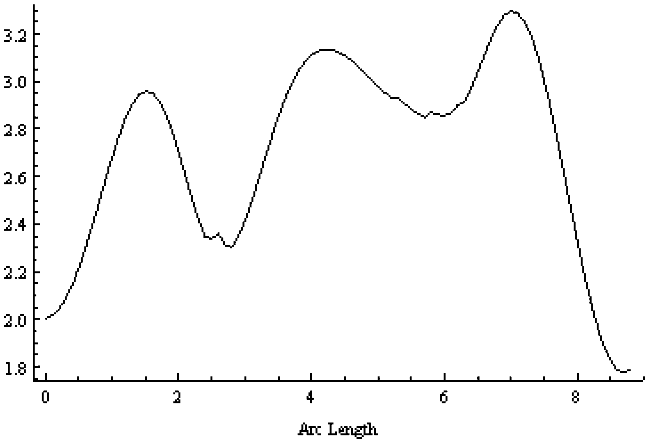

Figure 2. The energy plot of the MEP for the LPM. The intermediates are poorly resolved since they are furthest from the evaluated points.

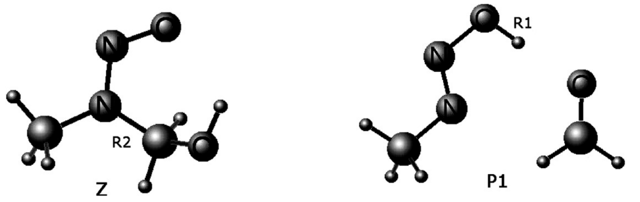

Figure 3. The g03 optimized end points for *N*-hydroxymethyl-methylnitrosamine (HMMN) using HF/3-21G. R1 is the hydroxyl O−H bond distance and R2 is the N−C distance. The Z and P1 labels correspond to the structures from ref 47.

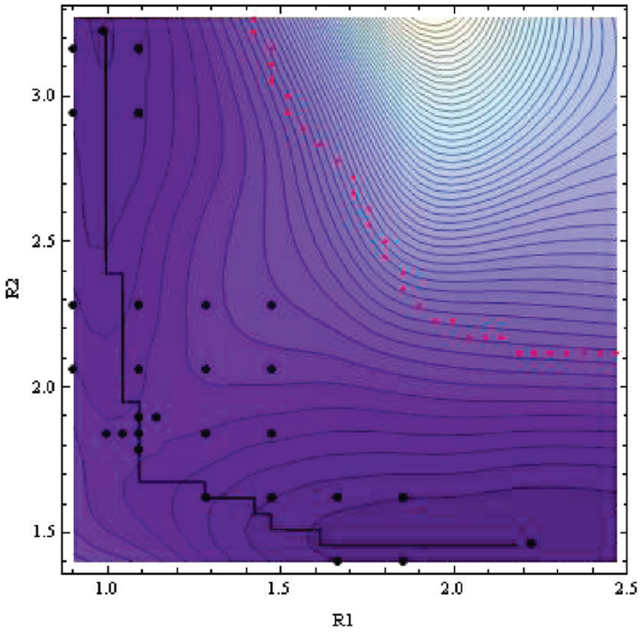

Figure 4. The HF/3-21G potential energy surface for HMMN demethylation after the BLPM has converged, where the distances are in Angstroms. The line of dots at the top of the plot is the boundary for an error tolerance of 0.01 au. The large black dots are points are optimized points on the surface and the curve is the approximate MEP. Most evaluations are clustered near the TS.

1.79308) and has a potential value of 3.2961. We set $\Delta\mathbf{R}_{\mathrm{large}} = (1.5,1.5)$, $\Delta\mathbf{R}_{\mathrm{small}} = (0.1,0.1)$, $\mathbf{lb} = (-2.5, -2.5)$, $\mathbf{ub} = (2.5,2.5)$, $E_{\mathrm{barr}} = 6$, and $\varepsilon_{\mathrm{shep}} = 0.2$ for the LPM. The coarse grid points are shifted by (0.4, 0.4) away from the lower bound. The converged results are shown in Figure 1 with the interpolated MEP shown in Figure 2. LPM converges on the grid to $\mathbf{R}_{\mathrm{TS}} = (-0.3,1.8)$ with $V(\mathbf{R}_{\mathrm{TS}}) = 3.2953$. The method requires 16 evaluations to calculate the potential on the initial grid, 1 evaluation for the starting point and 7 more evaluations on the finer grid for a total of 24. The points in the last step are largely focused on the regions with the highest barriers, with one point located near the second intermediate where the error is particularly large. Operating on the finer grid, it takes FMM 241 evaluations to resolve the TS to the same degree of accuracy, and it takes 75 with a grid spacing (0.5, 0.5).

***N*-Hydroxymethyl-Methylnitrosamine (HMMN).** This system, which is shown in Figure 3, is taken from ref 47. Rather than examine the entire reaction we simply looked at the demethylation step which results in the product methyldiazohydroxide. The compound is interesting since has been shown that it may methylate DNA bases in vivo.[48] Gaussian 03[46] was used to evaluate each point using the keywords OPT and MODRED in the heading with the bond variables R1 and R2, shown in Figure 3, kept frozen (http://www.chemistry.mcmaster.ca/ayers/projects.html). For this system, we use $\Delta\mathbf{R}_{\mathrm{large}} = (0.38,0.44)$, $\Delta\mathbf{R}_{\mathrm{small}} = (0.0475, 0.055)$, $\mathbf{lb} = (0.9,1.4)$, $\mathbf{ub} = (2.5,3.3)$, $\varepsilon_{\mathrm{shep}} = 0.01$ au, $\mathbf{R}_{\mathrm{react}} = (0.98,3.22)$ and $\mathbf{R}_{\mathrm{prod}} = (2.22,1.46)$, starting from the product “P1” rather than “Z”. The potential energy surface is shown in Figure 4. LPM requires 34 constrained geometry optimizations to converge to the TS while BLPM requires just 20.

When the coarser grid is made finer by setting $\Delta\mathbf{R}_{\mathrm{large}} = (0.19,0.22)$, then the LPM takes significantly longer using

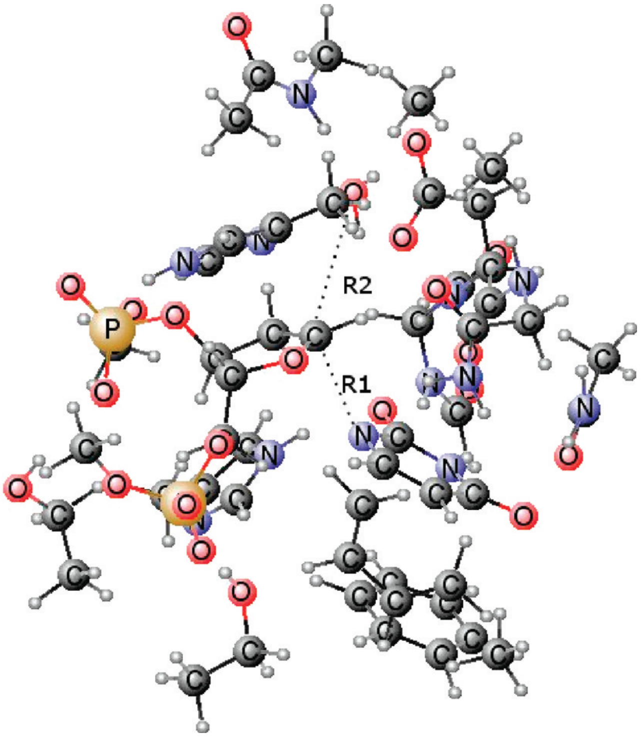

Figure 5. A Uracil-DNA glycosylase cluster model based on ref 50. The two coordinates used for the BLPM are shown as R1 and R2. This oxocarbenium cation/anion intermediate was optimized with AM1 keeping key unreactive atoms fixed.

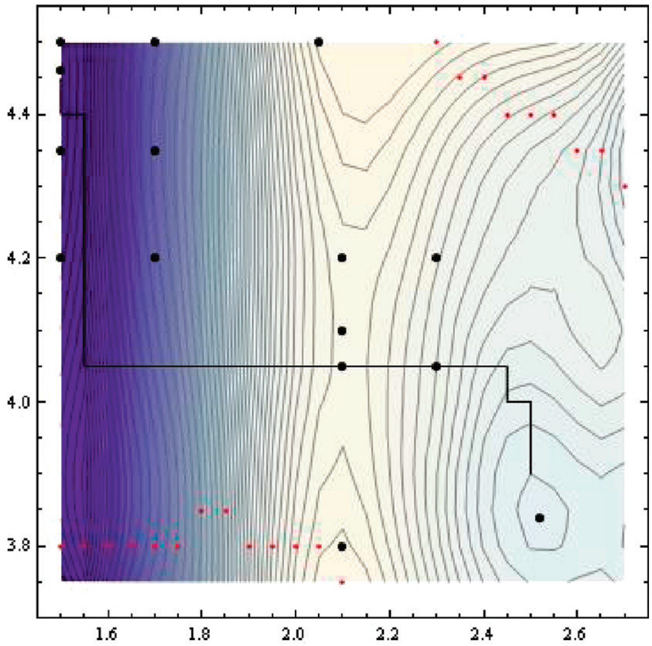

Figure 6. The RPES of the Uracil-DNA glycosylase cluster model shown in Figure 5. The *x*-axis is R1, the length of the glycosylic bond in Angstroms and the *y*-axis is R2, the distance between C1′ on the sugar ring and the oxygen on the water molecule, also in Angstroms. The boundary points are the string of points near the top and bottom of the plot. The dots are evaluated points and the line is the approximate MEP that is obtained by backtracing from the product.

85 constrained optimizations to converge compared to BLPM, which only takes 28. In both cases, the methods converge to within 0.2 kcal/mol of the true barrier.

Cluster Model of Uracil-DNA Glycosylase. To test this method on a larger system, we selected a cluster model of 177 atoms, shown in Figure 5, from Uracil-DNA Glycosylase based on the crystal structure 1EHM.[49] As in ref 50, a select number of atoms were frozen to keep the cluster together and AM1 was used to minimize the end point structures. For this system, we only tested BLPM with the parameters: $\Delta\mathbf{R}_{\text{large}} = (0.2,0.15)$, $\Delta\mathbf{R}_{\text{small}} = (0.05, 0.05)$, $\mathbf{lb} = (1.5,3.75)$, $\mathbf{ub} = (2.7,4.5)$, $\varepsilon_{\text{shep}} = 0.05$ au, $\mathbf{R}_{\text{react}} = (1.5,4.46)$, and $\mathbf{R}_{\text{prod}} = (2.52,3.84)$. The maximum allowed error was set to 0.01au. The exact AM1 TS, (2,12, 4.07), was located with Gaussian03 using the keyword opt(QST3) starting from the grid method's final structure at (2.1, 4.05). However, Gaussian03 had trouble converging for this structure, taking more than 500 steps before finishing. The interpolated RPES is shown in Figure 6 for BLPM, which converged in just 15 optimization cycles to (2.10, 4.05). Each constrained optimization cycle took about 30 iterations using the normal convergence criterion and 15 iterations using a loose criterion. The error was particularly large near the boundary at the top and bottom of the RPES, and the algorithm guessed at paths that would run along the boundaries. After the algorithm placed points near the boundary of the allowed region, it converged relatively quickly to the correct TS region.

4. Conclusions

Two new methods, the low-path method (LPM) and the boundary low-path method (BLPM), are proposed for finding the transition state (TS) structures on a reduced dimensional potential energy surface (RPES). Although it is more expensive to obtain points on the RPES than it is to evaluate points on the full-dimensional PES, the reduced dimensionality simplifies finding TS structures and allows the use of interpolation methods. Specifically for LPM and BLPM, Shepard interpolation was used. To prevent overfitting and to get a better error estimate, new methods were devised for determining the trust radius and the highest order of the interpolant.

The methods were shown to be able to rapidly locate the TS for an analytical function and two molecular systems. We attribute the success of these methods not only to the fact they work on the RPES, but also to the fact that, (a) they attempt to interpolate, rather than extrapolate, and (b) that they focus on providing an accurate description of the TS region, rather than the minimum energy path.

Acknowledgment. The authors thank the Chemistry Department of McMaster University, the Hamilton Foundation, NSERC, and the Canada Research Chairs for financial support.

References

(1) Kraut, D. A.; Carroll, K. S.; Herschlag, D. Challenges in enzyme mechanism and energetics. *Annu. Rev. Biochem.* **2003**, *72*, 517–571.

(2) Hammes, G. G. Multiple conformational changes in enzyme catalysis. *Biochemistry* **2002**, *41* (26), 8221.

(3) Morokuma, K.; Kato, S. *Potential Energy Surfaces and Dynamics Calculatoins*; Plenum: New York, 1981.

(4) Miller, W. H.; Handy, N. C.; Adams, J. E. Reaction Path Hamiltonian for Polyatomic Molecules. *J. Chem. Phys.* **1980**, *72* (1), 99.

(5) Gonzalez, C.; Schlegel, H. B. An improved algorithm for reaction path following. *J. Chem. Phys.* **1989**, *90* (4), 2154–2161.

(6) Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993**, *93* (7), 2395–2417.

(7) Bofill, J. M.; Anglada, J. M. Finding transition states using reduced potential-energy surfaces. *Theor. Chem. Acc.* **2001**, *105* (6), 463–472.

(8) Bofill, J. M. Updated Hessian matrix and the restricted step method for locating transition structures. *J. Comput. Chem.* **1994**, *15* (1), 1–11.

(9) Culot, P.; Dive, G.; Nguyen, V.; Ghuysen, J. A quasi-Newton algorithm for first-order saddle-point location. *Theor. Chim. Acta* **1992**, *82*, 189–205.

(10) Munro, L. J.; Wales, D. J. Defect migration in crystalline silicon. *Phys. Rev. B* **1999**, *59*, 3969–3980.

(11) Kumeda, Y.; Wales, D. J.; Munro, L. J. Transition states and rearrangement mechanisms from hybrid eigenvector-following and density functional theory.: Application to C10H10 and defect migration in crystalline silicon. *Chem. Phys. Lett.* **2001**, *341* (1), 185.

(12) Baker, J. *J. Comput. Chem.* **1986**, *7*, 385.

(13) Henkelman, G.; Jónsson, H. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Chem. Phys.* **1999**, *111*, 7010.

(14) Burger, S. K.; Ayers, P. W., Methods for finding transition states on reduced potential energy surfaces. J. Chem. Phys. 2010, (accepted).

(15) E, W. N.; Ren, W. Q.; Vanden-Eijnden, E. Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *J. Chem. Phys.* **2007**, *126* (16), 164103.

(16) E, W. N.; Ren, W. Q.; Vanden-Eijnden, E. String method for the study of rare events. *Phys. Rev. B* **2002**, *66* (5), 052301.

(17) Henkelman, G.; Jonsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **2000**, *113* (22), 9978–9985.

(18) Henkelman, G.; Uberuaga, B. P.; Jonsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **2000**, *113* (22), 9901–9904.

(19) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. String method in collective variables: Minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.* **2006**, *125* (2), 024106.

(20) Peters, B.; Heyden, A.; Bell, A. T. A growing string method for determining transition states: Comparison to the nudged elastic band and string methods. *J. Chem. Phys.* **2004**, *120* (17), 7877–7886.

(21) Quapp, W. A growing string method for the reaction pathway defined by a Newton trajectory. *J. Chem. Phys.* **2005**, *122* (17), 174106.

(22) Quapp, W. Reaction pathways and projection operators: Application to string methods. *J. Comput. Chem.* **2004**, *25* (10), 1277–1285.

(23) Burger, S. K.; Yang, W. T. Sequential quadratic programming method for determining the minimum energy path. *J. Chem. Phys.* **2007**, *127* (16), 164107.

(24) Burger, S. K.; Yang, W. T. Quadratic string method for determining the minimum-energy path based on multiobjective optimization. *J. Chem. Phys.* **2006**, *124* (5), 054109.

(25) Ayala, P. Y.; Schlegel, H. B. A combined method for determining reaction paths, minima, and transition state geometries. *J. Chem. Phys.* **1997**, *107* (2), 375–384.

(26) Jensen, F. Locating transition structures by mode following: A comparison of six methods on the Ar8 Lennard-Jones potential. *J. Chem. Phys.* **1995**, *102*, 6706.

(27) Peng, C.; Ayala, P.; Schlegel, H.; Frisch, M. Using redundant internal coordinates to optimize equilibrium geometries and transition states. *J. Comput. Chem.* **1996**, *17* (1), 49–56.

(28) Budzelaar, P. Geometry optimization using generalized, chemically meaningful constraints. *J. Comput. Chem.* **2007**, *28* (13), 2226–2236.

(29) Ischtwan, J.; Collins, M. A. Molecular-Potential Energy Surfaces by Interpolation. *J. Chem. Phys.* **1994**, *100* (11), 8080–8088.

(30) Collins, M. A. Molecular potential-energy surfaces for chemical reaction dynamics. *Theor. Chem. Acc.* **2002**, *108* (6), 313–324.

(31) Burger, S. K.; Liu, Y.; Sarkar, U.; Ayers, P. W. Moving least-squares enhanced Shepard interpolation for the fast marching and string methods. *J. Chem. Phys.* **2009**, *130*, 024103.

(32) Dawes, R.; Thompson, D. L.; Guo, Y.; Wagner, A. F.; Minkoff, M. Interpolating moving least-squares methods for fitting potential energy surfaces: Computing high-density potential energy surface data from low-density ab initio data points. *J. Chem. Phys.* **2007**, *126* (18), 084107.

(33) Kawano, A.; Guo, Y.; Thompson, D. L.; Wagner, A. F.; Minkoff, M. Improving the accuracy of interpolated potential energy surfaces by using an analytical zeroth-order potential function. *J. Chem. Phys.* **2004**, *120* (14), 6414–6422.

(34) Dey, B. K.; Ayers, P. W. Computing tunneling paths with the Hamilton-Jacobi equation and the fast marching method. *Mol. Phys.* **2007**, *105* (1), 71–83.

(35) Dey, B. K.; Ayers, P. W. A Hamilton-Jacobi type equation for computing minimum potential energy paths. *Mol. Phys.* **2006**, *104* (4), 541–558.

(36) Dey, B. K.; Bothwell, S.; Ayers, P. W. Fast marching method for calculating reactive trajectories for chemical reactions. *J. Math. Chem.* **2007**, *41* (1), 1–25.

(37) Dey, B. K.; Janicki, M. R.; Ayers, P. W. Hamilton-Jacobi equation for the least-action/least-time dynamical path based on fast marching method. *J. Chem. Phys.* **2004**, *121* (14), 6667–6679.

(38) Crittenden, D. L.; Jordan, M. J. T. Interpolated potential energy surfaces: How accurate do the second derivatives have to be. *J. Chem. Phys.* **2005**, *122* (4).

(39) Bettens, R. P. A.; Collins, M. A. Learning to interpolate molecular potential energy surfaces with confidence: A Bayesian approach. *J. Chem. Phys.* **1999**, *111* (3), 816–826.

(40) Jordan, M. J. T.; Thompson, K. C.; Collins, M. A. The Utility of Higher-Order Derivatives in Constructing Molecular-Potential Energy Surfaces by Interpolation. *J. Chem. Phys.* **1995**, *103* (22), 9669–9675.

(41) Schatz, G. C. The analytical representation of electronic potential-energy surfaces. *Rev. Mod. Phys.* **1989**, *61* (3), 669–688.

(42) Farwig, R. Rate of convergence of shepard global interpolation formula. *Math. Comput.* **1986**, *46* (174), 577–590.

(43) Draper, N. R.; Smith, H. *Applied Regression Analysis*. 2nd ed.; John Wiley & Sons, Inc: New York, 1980.

(44) Neter, J.; Kutner, M.; Wasserman, W.; Nachtsheim, C., *Applied Linear Statistical Models*; McGraw-Hill/Irwin: New York, 1996.

(45) Burger, S.; Liu, Y.; Ayers, P. Fast Marching Method Fortran 90 code v. 1.0, 2009.

(46) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R. ; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al.Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*, Gaussian, Inc.: Wallingford, CT, 2003.

(47) Lu, C. L.; Liu, Y. D.; Zhong, R. G. Theoretical Investigation of mono- and bi-function alkylating agents. *J. Mol. Struct. THEOCHEM* **2009**, *893*, 106–110.

(48) Anderson, L. M.; Souliotis, V. L.; Chhabra, S. K.; Moskal, T. J.; Hargaugh, S. D.; Kyrtopouls, S. A. *Int. J. Cancer* **1996**, (66), 130.

(49) Parikh, S.; Walcher, G.; Jones, G.; Slupphaug, G.; Krokan, H.; Blackburn, G.; Tainer, J. Uracil-DNA glycosylase-DNA substrate and product structures: Conformational strain promotes catalytic efficiency by coupled stereoelectronic effects. *Proc. Natl. Acad. Sci.* **2000**, *97* (10), 5083.

(50) Dinner, A. R.; Blackburn, G. M.; Karplus, M. W. Uracil-DNA glycosylase acts by substrate autocatalysis. *Nature* **2001**, *413*, 752.

CT100012Y