Chapter 10

METHODS FOR FINDING SADDLE POINTS AND MINIMUM ENERGY PATHS

Graeme Henkelman, Gísli Jóhannesson and Hannes Jónsson

Department of Chemistry 351700,
University of Washington,
Seattle, WA 98195-1700

Abstract The problem of finding minimum energy paths and, in particular, saddle points on high dimensional potential energy surfaces is discussed. Several different methods are reviewed and their efficiency compared on a test problem involving conformational transitions in an island of adatoms on a crystal surface. The focus is entirely on methods that only require the potential energy and its first derivative with respect to the atom coordinates. Such methods can be applied, for example, in plane wave based Density Functional Theory calculations, and the computational effort typically scales well with system size. When the final state of the transition is known, both the initial and final coordinates of the atoms can be used as boundary conditions in the search. Methods of this type include the Nudged Elastic Band, Ridge, Conjugate Peak Refinement, Drag method and the method of Dewar, Healy and Stewart. When only the initial state is known, the problem is more challenging and the search for the saddle point represents also a search for the optimal transition mechanism. We discuss a recently proposed method that can be used in such cases, the Dimer method.

I. INTRODUCTION

A common and important problem in theoretical chemistry and in condensed matter physics is the calculation of the rate of transitions, for example chemical reactions or diffusion events. In either case, the configuration of atoms is changed in some way during the transition. The interaction between the atoms can be obtained from an (approximate) solution of the Schrödinger equation describing the electrons, or from an otherwise determined potential energy function. Most often, it is sufficient to treat the motion of the atoms using classical mechanics, but the transitions of interest are typically many orders of magnitude slower than vibrations of the atoms,

so a direct simulation of the classical dynamics is not useful. This 'rare event' problem is best illustrated by an example. We will be describing below a study of configurational changes in a Pt island on a Pt(111) surface, relevant to the diffusion of the island over the surface. The approximate interaction potential predicts that the easiest configurational change has an activation energy barrier of 0.6 eV. This is a typical activation energy for diffusion on surfaces. Such an event occurs many times per second at room temperature and is, therefore, active on a typical laboratory time scale. But, there are on the order of 10^{10} vibrational periods in between such events. A direct classical dynamics simulation which necessarily has to faithfully track all this vibrational motion would take on the order of 10^5 years of computer calculations on the fastest present day computer before a single diffusion event can be expected to occur! It is clear that meaningful studies of these kinds of events cannot be carried out by simply simulating the classical dynamics of the atoms. It is essential to carry out the simulations on a much longer timescale. This time scale problem is one of the most important challenges in computational chemistry, materials science and condensed matter physics.

The time scale problem is devastating for direct dynamical simulations, but makes it possible to obtain accurate estimates of transition rates using purely statistical methods, namely Transition State Theory (TST).[1–5] Apart from the Born-Oppenheimer approximation, TST relies on two basic assumptions: (a) the rate is slow enough that a Boltzmann distribution is established and maintained in the reactant state, and (b) a dividing surface of dimensionality D−1 where D is the number degrees of freedom in the system can be identified such that a reacting trajectory going from the initial state to the final state only crosses the dividing surface once. The dividing surface must, therefore, represent a bottleneck for the transition. The TST expression for the rate constant can be written as

$$k = \frac{\langle |v| \rangle}{2} \frac{Q^{\ddagger}}{Q_R}$$

where $\langle |v| \rangle$ is the average speed, $Q^{\ddagger}$ is the configurational integral for the transition state dividing surface, and Q_R is the configurational integral for the initial state. The bottleneck can be of purely entropic origin, but most often in crystal growth problems it is due to a potential energy barrier between the two local minima corresponding to the initial and final states. It can be shown that TST always overestimates the rate of escape from a given initial state[2,3] (a diffusion constant can be underestimated if multiple hops are not included in the analysis[6]). This leads to a variational principle which can be used to find the optimal dividing surface.[3,7] The TST rate estimate gives an approximation for the rate of escape from the initial state, irrespective of the final state. The possible final states can be determined by short time simulations of the dynamics starting from the dividing surface. This can also give an estimate of the correction to transition state theory due to approximation (b), the so called dynamical corrections.[8,9]

Since atoms in crystals are usually tightly packed and the relevant temperatures are low compared with the melting temperature, the harmonic approximation to TST (hTST) can typically be used in studies of diffusion and reactions in crystals.[9] This greatly simplifies the problem of estimating the rates. The search for the optimal transition state then becomes a search for the lowest few saddle points at the edge of the potential energy basin corresponding to the initial state. The rate constant for transition through the region around each one of the saddle points can be obtained from the energy and frequency of normal modes at the saddle point and the initial state,[10,11]

$$k^{\mathrm{hTST}} = \frac{\prod_i^{3N} \nu_i^{\mathrm{init}}}{\prod_i^{3N-1} \nu_i^{\ddagger}} \, e^{-(E^{\ddagger} - E^{\mathrm{init}})/k_B T}.$$

Here, $E^{\ddagger}$ is the energy of the saddle point, E^{init} is the local potential energy minimum corresponding to the initial state, and the ν_i are the corresponding normal mode frequencies. The symbol ‡ refers to the saddle point. The most challenging part in this calculation is the search for the relevant saddle points. Again, the mechanism of the transition is reflected in the saddle point. The reaction coordinate at the saddle point is the direction of the unstable mode (the normal mode with negative eigenvalue). After a saddle point has been found, one can follow the gradient of the energy downhill, both forward and backward, and map out the Minimum Energy Path (MEP), thereby establishing what initial and final state the saddle point corresponds to. The identification of saddle points ends up being one of the most challenging tasks in theoretical studies of transitions in condensed matter.

The MEP is frequently used to define a 'reaction coordinate'[12] for transitions. It can be an important concept for building in anharmonic effects, or even quantum corrections.[5] The MEP may have one or more minima in between the endpoints corresponding to stable intermediate configurations. The MEP will then have two or more maxima, each one corresponding to a saddle point. Assuming a Boltzmann population is reached for the intermediate (meta)stable configurations, the overall rate is determined by the highest energy saddle point. It is, therefore, not sufficient to find *a* saddle point, but rather one needs to find the *highest* saddle point along the MEP, in order to get an accurate estimate of the rate from hTST.

For systems where one or more atoms need to be treated quantum mechanically, a quantum mechanical extension of TST, so called RAW-QTST, can be used.[13,14] Zero point energy and tunneling are then taken into account by using Feynman Path Integrals.[15] Since RAW-QTST is a purely statistical theory analogous to classical TST, the path integrals are statistical (involve only imaginary time) and are easy to sample in computer simulations even for large systems. The definition of the transition state needs to be extended to higher dimensions, but otherwise the RAW-QTST calculation for quantum systems is quite similar to the TST calculations for classical systems. A central problem is finding a good reaction coordinate and a good transition state surface. In a harmonic approximation to RAW-QTST, the central problem becomes the identification of saddle points on an effective potential energy

surface with higher dimensionality than the regular potential energy surface.[13,14] The saddle points are often referred to as 'instantons' and the harmonic approximation to RAW-QTST is the so called Instanton Theory.[16–18] Any method that can be used to locate saddle points efficiently in high dimension, can, therefore, also be useful for calculating rates in quantum systems.

Many different methods have been presented for finding MEPs and saddle points.[19,20] Since a first order saddle point is a maximum in one direction and a minimum in all other directions, methods for finding saddle points invariably involve some kind of maximization of one degree of freedom and minimization in other degrees of freedom. The critical issue is to find a good and inexpensive estimate of which degree of freedom should be maximized. Below, we give an overview of several commonly used methods in studies of transitions in condensed matter. We then compare their performance on the surface island test problem.

II. THE DRAG METHOD

The simplest and perhaps the most intuitive method of all is what we will refer to as the Drag method. It actually has many names because it keeps being reinvented. One degree of freedom, the drag coordinate, is chosen and is held fixed while all other D-1 degrees of freedom are relaxed, i.e. the energy of the system minimized in a D-1 dimensional hyperplane. In small, stepwise increments, the drag coordinate is increased and the system is dragged from reactants to products. The maximum energy obtained is taken to be the saddle point energy. Sometimes, a guess for a good reaction coordinate is used as the choice for the drag coordinate. This could be the distance between two atoms, for example, atoms that start out forming a bond which ends up being broken. In the absence of such an intuitive choice, the drag coordinate can be simply chosen to be the straight line interpolation between the initial and final state. This is a less biased way and all coordinates of the system then contribute in principle to the drag coordinate. We will follow this second approach, which is illustrated in figure 1. We have implemented the Drag method in such a way that the force acting on the system is inverted along the drag coordinate and the velocity Verlet algorithm[21] with a projected velocity is used to simulate the dynamics of the system. The velocity projection is carried out at each time step and ensures that only the component of the velocity parallel to the force is included in the dynamics. When the force and projected velocity point in the opposite direction (indicating that the system has gone over the energy ridge), the velocity is zeroed. This projected velocity Verlet algorithm has been found to be an efficient and simple minimization algorithm for many of the methods discussed here.

The problem with the Drag method is that both the intuitive, assumed reaction coordinate and the unbiased straight line interpolation can turn out to be bad reaction coordinates. They may be effective in distinguishing between reactants and products, but a reaction coordinate must do more than that. A good reaction coordinate should

give the direction of the unstable normal mode at the saddle point. Only then does a minimization in all other degrees of freedom bring the system to the saddle point. Figure 1 shows a simple case where the drag method fails. As the drag coordinate is incremented, starting from the initial state, **R**, the system climbs up close to the slowest ascent path. After climbing high above the saddle point energy, the energy contours eventually stop confining the system in this energy valley and the system abruptly snaps into an adjacent valley (the product valley in the case of figure 1). The system is never confined to the vicinity of the saddle point because the direction of the drag coordinate is at a large angle to the direction of the unstable normal mode at the saddle point. While there certainly are cases where the drag method works, there are also many examples where it does not work.[22,23] The method failed, for example, on half the saddle points in the surface island test problem described below. What seems to be a more intuitive reaction coordinate, such as the distance between two atoms, can also fail, for example if adjacent atoms also get displaced in going from the initial to final states. As the two atoms get dragged apart, the adjacent atoms can snap from one position to another, never visiting the saddle point configuration. As we will demonstrate below, much more reliable methods exist which are not significantly more involved to implement or costly to use.

III. THE NEB METHOD

In the Nudged Elastic Band (NEB) method[20,24,25] a string of replicas (or 'images') of the system are created and connected together with springs in such a way as to form a discrete representation of a path from the reactant configuration, **R**, to the product configuration, **P**. Initially, the images may be generated along the straight line interpolation between **R** and **P**. An optimization algorithm is then applied to relax the images down towards the MEP. The NEB and the CPR method are unique among the methods discussed here in that they not only give an estimate of the saddle point, but also give a more global view of the energy landscape, for example, showing whether more than one saddle point is found along the MEP.

The string of images can be denoted by $[\mathbf{R}_0,\ \mathbf{R}_1,\ \mathbf{R}_2, \ldots, \mathbf{R}_N]$ where the endpoints are fixed and given by the initial and final states, $\mathbf{R}_0 = \mathbf{R}$ and $\mathbf{R}_N = \mathbf{P}$, but $N-1$ intermediate images are adjusted by the optimization algorithm. The most straightforward approach would be to construct an object function

$$S(\mathbf{R}_1, \ldots, \mathbf{R}_N) \;=\; \sum_{i=1}^{N-1} E(\mathbf{R}_i) \;+\; \sum_{i=1}^{N} \frac{k}{2}(\mathbf{R}_i - \mathbf{R}_{i-1})^2 \qquad (1)$$

and minimize with respect to the intermediate images, $\mathbf{R}_1, \ldots, \mathbf{R}_N$. This mimics an elastic band made up of $N-1$ beads and N springs with spring constant k. The band is strung between the two fixed endpoints. The problem with this formulation is that the elastic band tends to cut corners and gets pulled off the MEP by the spring forces in regions where the MEP is curved. Also, the images tend to slide down

towards the endpoints, giving lowest resolution in the region of the saddle point, where it is most needed.[20] Both the corner-cutting and the sliding-down problems can be solved easily with a force projection. This is what is referred to as 'nudging'. The reason for corner-cutting is the component of the spring force perpendicular to the path, while the reason for the down-sliding is the parallel component of the true force coming from the interaction between atoms in the system. Given an estimate of the unit tangent to the path at each image (which will be discussed later), $\hat{\tau}_i$, the force on each image should only contain the parallel component of the spring force, and perpendicular component of the true force

$$\mathbf{F}_i \;=\; -\nabla E(\mathbf{R}_i)|_\perp \;+\; \mathbf{F}_i^s \cdot \hat{\tau}_i \, \hat{\tau}_i \tag{2}$$

where $\nabla E(\mathbf{R}_i)$ is the gradient of the energy with respect to the atomic coordinates in the system at image i, and $\mathbf{F}_i^s$ is the spring force acting on image i. The perpendicular component of the gradient is obtained by subtracting out the parallel component

$$\nabla E(\mathbf{R}_i)|_\perp \;=\; \nabla E(\mathbf{R}_i) \;-\; \nabla E(\mathbf{R}_i) \cdot \hat{\tau}_\parallel \, \hat{\tau}_\parallel \tag{3}$$

In order to ensure equal spacing of the images (when the same spring constant, k, is used for all the springs), even in regions of high curvature where the angle between $\mathbf{R}_i - \mathbf{R}_{i-1}$ and $\mathbf{R}_{i+1} - \mathbf{R}_i$ deviates significantly from $0°$, the spring force should be evaluated as

$$\mathbf{F}_i^s \,|_\parallel = \; k\,(|\mathbf{R}_{i+1} - \mathbf{R}_i| - |\mathbf{R}_i - \mathbf{R}_{i-1}|)\,\hat{\tau}_i. \tag{4}$$

III.1 ESTIMATE OF THE TANGENT

We now discuss the estimate of the tangent to the path. In the original formulation of the NEB method, the tangent at an image i was estimated from the two adjacent images along the path, $\mathbf{R}_{i+1}$ and $\mathbf{R}_{i-1}$. The simplest estimate is to use the normalized line segment between the two

$$\hat{\tau}_i = \frac{\mathbf{R}_{i+1} - \mathbf{R}_{i-1}}{|\mathbf{R}_{i+1} - \mathbf{R}_{i-1}|} \tag{5}$$

but a slightly better way is to bisect the two unit vectors

$$\tau_i \;=\; \frac{\mathbf{R}_i - \mathbf{R}_{i-1}}{|\mathbf{R}_i - \mathbf{R}_{i-1}|} + \frac{\mathbf{R}_{i+1} - \mathbf{R}_i}{|\mathbf{R}_{i+1} - \mathbf{R}_i|} \tag{6}$$

and then normalize $\hat{\tau} = \tau/|\tau|$. This latter way of defining the tangent ensures the images are equispaced even in regions of large curvature.

These estimates of the tangent have, however, turned out to be problematic in some cases.[26] When the energy of the system changes rapidly along the path, but the

restoring force on the images perpendicular to the path is weak, as when covalent bonds are broken and formed, the paths can get 'kinky' and convergence to the MEP may never be reached. One way to aleviate the problem is to introduce a switching function that introduces a small part of the perpendicular component of the spring force.[20] This, however, can introduce corner-cutting and lead to an overestimate of the saddle point energy. The kinkiness can be eliminated by using a better estimate of the tangent.[26] The tangent of the path at an image i is defined by the vector between the image and the neighboring image with higher energy. That is

$$\tau_i = \begin{cases} \tau_i^+ & \text{if } E_{i+1} > E_i > E_{i-1} \\ \tau_i^- & \text{if } E_{i+1} < E_i < E_{i-1} \end{cases} \tag{7}$$

where

$$\tau_i^+ = \mathbf{R}_{i+1} - \mathbf{R}_i, \text{ and } \tau_i^- = \mathbf{R}_i - \mathbf{R}_{i-1}, \tag{8}$$

and $E_i = E(\mathbf{R}_i)$. If both of the adjacent images are either lower in energy, or both are higher in energy than image i, the tangent is taken to be a weighted average of the vectors to the two neighboring images. The weight is determined from the energy. The weighted average only plays a role at extrema along the MEP and it serves to smoothly switch between the two possible tangents τ_i^+ and τ_i^-. Otherwise, there is an abrupt change in the tangent as one image becomes higher in energy than another and this could lead to convergence problems. If image i is at a minimum $E_{i+1} > E_i < E_{i-1}$ or at a maximum $E_{i+1} < E_i > E_{i-1}$, the tangent estimate becomes

$$\tau_i = \begin{cases} \tau_i^+ \Delta E_i^{\max} + \tau_i^- \Delta E_i^{\min} & \text{if } E_{i+1} > E_{i-1} \\ \tau_i^+ \Delta E_i^{\min} + \tau_i^- \Delta E_i^{\max} & \text{if } E_{i+1} < E_{i-1} \end{cases} \tag{9}$$

where

$$\begin{aligned} \Delta E_i^{\max} &= \max\left(|E_{i+1} - E_i|, |E_{i-1} - E_i|\right), \quad \text{and} \\ \Delta E_i^{\min} &= \min\left(|E_{i+1} - E_i|, |E_{i-1} - E_i|\right). \end{aligned} \tag{10}$$

Finally, the tangent vector needs to be normalized. With this modified tangent, the elastic band is well behaved and converges rigorously to the MEP if sufficient number of images are included.

III.2 MINIMIZATION OF THE FORCE

The implementation of the NEB method in a classical dynamics program is quite simple. First, the energy and gradient need to be evaluated for each image in the elastic band using some description of the energetics of the system (a first principles

calculation or an empirical or semi-empirical force field). Then, for each image, the coordinates and energy of the two adjacent images are required in order to estimate the local tangent to the path, project out the perpendicular component of the gradient and add the parallel component of the spring force. The computation of ∇V for the various images of the system can be done in parallel on a cluster of computers, for example with a separate node handling each one of the images. Each node then only needs to receive coordinates and energy of adjacent images to evaluate the spring force and to carry out the force projections. Various techniques can be used for the minimization. We have used projected velocity Verlet algorithm described above (see the section on Drag method).

To start the NEB calculation, an initial guess is required. We have found a simple linear interpolation between the initial and final point adequate in many cases. When multiple MEPs are present, the optimization leads to convergence to the MEP closest to the initial guess, as illustrated in figure 2. In order to find the optimal MEP in such a situation, some sampling of the various MEPs needs to be carried out, for example a simulated annealing procedure, or an algorithm which drives the system from one MEP to another, analogous to the search for a global minimum on a potential energy surface with many local minima.[27]

It is important to eliminate overall translation and rotation of the system during the optimization of the path. A method for constraining the center of mass and the orientation of the system has been described, for example, by reference 37. Often, it is sufficient to fix six degrees of freedom in each image of the system, for example by fixing one of the atoms (zeroing all forces acting on one of the atoms in the system), constraining another atom to only move along a line (zeroing, for example, the x and y components of the force), and constraining a third atom to move only in a plane (zeroing, for example, the x component of the force).

III.3 INTERPOLATION BETWEEN IMAGES

In order to obtain an estimate of the saddle point and to sketch the MEP, it is important to interpolate between the images of the converged elastic band. In addition to the energy of the images, the force along the band provides important information and should be incorporated into the interpolation. By including the force, the presence of intermediate local minima can often be extracted from bands with as few as three images. The interpolation can be done with a cubic polynomial fit to each segment $[\mathbf{R}_i, \mathbf{R}_{i+1}]$ in which the four parameters of the cubic function can be chosen to enforce continuity in energy and force at both ends. Writing the polynomial as $a_i x^3 + b_i x^2 + c_i x + d_i$, the parameters are[26]

$$\begin{aligned} a_x &= \frac{2E_{i+1}-E_i}{R^3} - \frac{F_i+F_{i+1}}{R^2} \\ b_x &= \frac{3E_{i+1}-E_i}{R^2} + \frac{2F_i+F_{i+1}}{R} \\ c_x &= -F_i \\ d_x &= E_i. \end{aligned} \tag{11}$$

where E_i and E_{i+1} are the values of the energy at the endpoints, and F_i and F_{i+1} are the values of the force along the path. This type of interpolation is usually quite smooth even though the second derivative is not forced to be continuous. A possible improvement is to generate a quintic polynomial interpolation so that the second derivatives can also be matched (and set to zero at the end points for a natural spline). This higher order polynomial can, however, add artificial wiggles in the path.[26]

III.4 APPLICATIONS OF THE NEB METHOD

The NEB method has been applied successfully to a wide range of problems, for example studies of diffusion processes at metal surfaces,[28] multiple atom exchange processes observed in sputter deposition simulations,[29] dissociative adsorption of a molecule on a surface,[25] diffusion of rigid water molecules on an ice Ih surface,[30] contact formation between metal tip and a surface,[31] cross-slip of screw dislocations in a metal (a simulation requiring over 100,000 atoms in the system, and a total of over 2,000,000 atoms in the MEP calculation),[32] and diffusion processes at and near semiconductor surfaces (using a plane wave based Density Functional Theory method to calculate the atomic forces).[33] In the last two applications the calculation was carried out on a cluster of workstations with the force on each image calculated on a separate node.

III.5 OTHER CHAIN-OF-STATES METHODS

The NEB method is an example of what has been called a chain-of-states method.[34] The common feature is that several images of the system are connected together to trace out a path of some sort. The simple object function for a chain (equation 1) is mathematically analogous to a Feynman path integral[15] for an off-diagonal element of a density matrix describing a quantum particle, which was used, for example, by Kuki and Wolynes to study electron tunneling in proteins.[35] Several chain-of-states methods have been formulated for finding transition paths that are optimal in one way or another.[36–43] The NEB method is the only one that converges to the MEP without having to use second derivatives of the energy. Elber and Karplus[36] formulated an object function which is essentially similar to equation 1 although more complex. Czerminski and Elber presented an improved method with the Self-Penalty Walk algorithm (SPW)[37] where a repulsion between images was added to the object function to prevent aggregation of images and crossings of the path with itself in regions near minima. Ulitsky and Elber,[38] and Choi and Elber presented a quite different algorithm, the Locally Updated Planes (LUP).[39] There, the optimization of the chain-of-states involves estimating a local tangent using equation 5 and then minimizing the energy of each image, i, within the hyperplane with normal q_i, i.e.

relaxing the system according to

$$\frac{\partial \mathbf{R}_i}{\partial t} = -\nabla V(\mathbf{R}_i)[1 - \hat{\mathbf{q}}_i \hat{\mathbf{q}}_i]. \tag{12}$$

After every M steps (where M is on the order of 10) in the relaxation, the local tangents $\hat{\mathbf{q}}_i$ are updated. Since there is no interaction between the images (such as the spring force in the NEB), the LUP algorithm gives an uneven distribution of images along the path, and can even give a discontinuous path when two or more MEPs lie between the given initial and final states.[39] Also, the images do not converge rigorously to the MEP, but slide down slowly to the endpoint minima because of kinks that form spontaneously on the path and fluctuate as the minimization is carried out. Choi and Elber point out that it is important to start with a good initial guess to the MEP to minimize these problems. The NEB method is closely related to both the LUP method and the Elber-Karplus method. The NEB method incorporates the strong points of both of these approaches.

Smart[43] modified the Elber-Karplus-Czerminski formulation to get better convergence to the saddle point. The object function in his formulation involves a very high power (on the order of 100 to 1000) of the energy of the images to increase the weight of the highest energy image along the path.

Sevick, Bell and Theodorou[40] proposed a chain of states method for finding the MEP, but their optimization method, which includes explicit constraints for rigidly fixing the distance between images, requires evaluation of the matrix of second derivatives of the potential and is, therefore, not as applicable to large systems and complex interactions.

Chain-of-states methods have also been used for finding classical dynamical paths.[41,42] Gillilan and Wilson[42] suggested using an object function similar to equation 1 for finding saddle points, but this suffers from the corner-cutting and down-sliding problems discussed above.

IV. THE CI-NEB METHOD

Recently, a modification of the NEB method has been developed, the Climbing Image - NEB.[44] There, one of the images, the one that turns out to have the highest energy after one, or possibly a few relaxation steps, is made to move uphill in energy along the elastic band. This is accomplished by zeroing the spring force on this one image completely and including only the inverted parallel component of the true force

$$\mathbf{F}^{\text{climb}}_{\text{imax}} = \nabla V(\mathbf{R}_{\text{imax}}) \cdot \hat{\tau}_{\parallel}\, \hat{\tau}_{\parallel} \tag{13}$$

The climbing image is dragged uphill, analogous to the drag method, but the essential difference is that the drag direction is determined by the location of the adjacent images in the band, not just **R** and **P** (unless the band only consists of one movable image). The tangent to the path is also weighted by the energy of the adjacent images as explained above. This turns out to be important in the surface island test problem.

Figure 2 shows the result of a CI-NEB calculations for the two dimensional test problem. Three movable images are included between the end points, and a straight line interpolation between **R** and **P** is used as a starting guess. The central image becomes the climbing image since it has the highest energy initially. Simultaneously, as the climbing image is pushed uphill, the other two images relax subject to the force projections of the nudging algorithm. After convergence is reached, a crude representation of the MEP has been obtained and one of the images is sitting at the saddle point to within the prescribed tolerance. An important aspect of the algorithm is that all movable images are adjusted simultaneously, and since only the position of adjacent images are needed for each step, the algorithm again parallelizes just as efficiently as the regular NEB.

V. THE CPR METHOD

For the conjugate peak refinement method,[45] (CPR), a set of images is generated, one at a time, between the initial and final configurations, **R** and **P**. After the images are optimized, a line between the images constitutes a path that lies close to (but not at) the MEP. The maxima along the path will be at saddle points. Each point along the path is generated in a cycle of line maximizations and conjugate gradient minimizations. This is illustrated in figure 3. In the first cycle, the maximum along the vector $\mathbf{P}-\mathbf{R}$ is found, $\mathbf{y}_1$. Then, a minimization is carried out along the direction of each of the conjugate vectors (a total of D$-$1 dimensions) to give a new point $\mathbf{x}_1$.

In the second cycle the maximum along an estimated tangent to the $\mathbf{R}-\mathbf{x}_1-\mathbf{P}$ path is found. The tangent is estimated using equation 6. This new maximum is denoted $\mathbf{y}_2$ in figure 3. The energy is then minimized along each of the conjugate vectors to give a new point that could potentially get incorporated into the path, etc. The rules for deciding whether a new point gets added to the path permanently are quite complicated and will not be given here. The cycle of maximization along the tangent and then conjugate gradient line minimizations is repeated until a maximum along the path has a smaller gradient than the given tolerance for saddle points.

A detailed implementation of the CPR method, the TRAVEL algorithm, has been described by Fischer,[46] providing values for all relevant parameters. We have used standard algorithms from reference 47 for bracketing energy extrema and the line-optimizations.

We did not use the algorithm to generate a full path but stopped as soon as a point was found that satisfied our criterion for a saddle point (the magnitude of the gradient of the energy being less than a given tolerance).

VI. THE RIDGE METHOD

The Ridge method of Ionova and Carter[48] involves advancing two images of the system, one on each side of the potential energy ridge, down towards the saddle point. The pair of images is moved in cycles of ‘side steps’ and ‘downhill steps’

in the following way. First, a straight line interpolation between products, **P**, and reactants, **R**, is formed and the maximum of energy along this line is found. The method is illustrated in figure 4, where the maximum is found at point **a**. We used the routine DBRENT from reference 47 to carry out the line maximizations, which makes use of the force, and typically takes a couple of force evaluations to converge to within 0.01 Å of the maximum. Then, two replicas of the system are created on the line, one on each side of the maximum, $\mathbf{x}'_0$ and $\mathbf{x}'_1$ (see figure 4).

The magnitude of the displacement of the two images from the maximum needs to be chosen. This 'side-step' distance is typically chosen to be 0.1 Å in the first cycle. The force is now evaluated at the two images and they are moved in the direction of the force a certain distance, the 'downhill-step'. This generates points $\mathbf{x}''_0$ and $\mathbf{x}''_1$. The downhill distance is typically chosen to be 0.1 Å in the first cycle. This completes the first cycle. Then, a new cycle is started by maximizing along the line $[\mathbf{x}''_0, \mathbf{x}''_1]$ to obtain the point **b**, etc.

The side-step and downhill-step of the images need to gradually decrease as the images get closer to the saddle point. It is possible that the energy of a point (in the sequence $\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$) is higher than at the previous point. In such cases the downhill displacement is reduced by a half. Also, if the ratio of the side-step to downhill-step distance becomes larger than a certain, chosen ratio, the side step distance is also decreased by a half. This ratio is typically chosen to be some number in the range between 1 and 10. We found that the algorithm worked best for a ratio of 1.2 in the test cases we carried out. As the two images move and the size of the side-step to downhill-step is decreased, the sequence of points $\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$ should lead to a saddle point.

If the two images are almost equally displaced from the top of the energy ridge and the ridge is straight, it can be sufficient to evaluate the force only at the central point, rather than at the two images, thereby saving a factor of two in the number of force evaluations. This is implemented in such a way that if a new point in the sequence $\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$ is close to the center of the two images (not within 30% of either image), then the force in the next cycle is only evaluated at the central point and applied to both images in the downhill-step.

It turns out that most of the force evaluations are needed when the two images are rather close to the saddle point. Ionova and Carter [48] have discussed possible ways to improve the performance of the method in this final stage of the search.

VII. THE DHS METHOD

Dewar, Healy and Stewart[49] (DHS) have proposed a method which also involves two images of the system. First, the endpoints **R** and **P** are joined by a line segment. The two images are then systematically drawn toward each other until the distance between them is smaller than a given tolerance for finding the saddle point.

There are two steps in each cycle. First, the energy of both images is calculated. The one at lower energy is then pulled towards the one at higher energy along the line segment, typically about 5% of the way. Second, the energy of the lower energy image is minimized keeping the distance between the two fixed. An application of the method to the two-dimensional test problem is shown in figure 5. In the first cycle, the image at **P** is higher in energy than the one at **R**, so the latter is brought in towards **P** by 5% and the allowed to relax with a fixed distance constraint. This repeats several times, causing the image that starts at **R** to climb up the potential energy valley leading up from **R**. Eventually, the image at **P** becomes lower in energy. The five cycles following that are shown with solid lines in figure 5. Remarkably, the pair of images end up moving past the local maximum and converge on the saddle point on the other side.

The method can locate the neighboring region of the saddle point quite quickly, but does not converge close to the saddle point efficiently. If the images are pulled towards each other too quickly, the probability of both images ending on the same side of the ridge is increased. Eventually, as the pair of images gets close enough to the saddle point, such a slip over the ridge is bound to occur and both images will then settle into one of the minima **R** or **P**.

We chose to use a velocity Verlet type algorithm[21] for the minimization of the position of the lower energy image. At each step only the force perpendicular to the line segment connecting the two images was included. The velocity parallel to the force was included in the dynamics until the two pointed in the opposite direction, at which point the velocity was zeroed. This is the same kind of minimization algorithm we use with the Drag, NEB and CI-NEB methods.

VIII. THE DIMER METHOD

When the final state of a transition is not known, the search for the saddle point is more challenging. A climb up from the initial state to the saddle point is more difficult than might at first appear. It is not sufficient to just follow the direction of slowest ascent – the two-dimensional test problem illustrated in figures 1 to 5 is an example of that. Several methods have been developed where information from second derivatives is built in to guide the climb.[50–55] These methods have become widely used in studies of small molecules and clusters. Their disadvantage is that they require the second derivatives of the energy with respect to all the atomic coordinates, i.e. the full Hessian matrix, and then the matrix needs to be diagonalized to find the normal modes, an operation that scales as D^3. The evaluation of second derivatives is often very costly, for example in plane wave based Density Functional Theory calculations. Also, in large systems where empirical potentials are used, the D^3 scaling becomes a problem. For example, in a very interesting recent study of relaxation processes in Lennard-Jones glasses, a practical limit was reached at a

couple of hundred atoms,[56] while system size effects can be present in such systems even when up to 1000 atoms are included.[57]

A new method for finding saddle points was recently presented which has the essential qualities of the mode following methods, but only requires first derivatives of the energy and no diagonalization.[58] It can therefore be applied to plane wave DFT calculations and it can be applied to large systems with several hundred atoms, as illustrated below. The method involves two replicas of the system, a 'dimer', as illustrated in figure 6. The dimer is used to transform the force in such a way that optimization leads to convergence to a saddle point rather than a minimum. The force acting on the center of the dimer (obtained by interpolating the force on the two images) gets modified by inverting the component in the direction of the dimer. Before translating the dimer, the energy is minimized with respect to orientation. As pointed out by Voter,[59] this gives the direction of the lowest frequency normal mode. This effective force will take the dimer to a saddle point when an optimization scheme is applied, for example conjugate gradients or the velocity Verlet algorithm with velocity damping. A detailed algorithm for finding the optimal orientation in an efficient way is described in reference 58. In a test problem involving Al adatom diffusion on the Al(100) surface, the Dimer method was found to converge preferably on the lowest saddle points (75% of the time the method converged on one of the lowest four saddle points) and the computational effort was found to increase only weakly as the number of degrees of freedom in the system was increased.[58]

Figure 7 shows a Dimer calculation for the two-dimensional test problem. The initial configurations for the dimer searches were taken from the extrema of a short high temperature molecular dynamics trajectory (shown as a dashed line). The three initial points are different enough that the dimer searches converge to separate saddle points. In general the strategy for the Dimer method is to try many different initial configurations around a minimum, in order to find the saddle points that lead out of that minimum basin.

IX. CONFIGURATIONAL CHANGE IN AN ISLAND ON FCC(111)

As a test problem for comparing the various methods described above, we have chosen a heptamer island on the (111) surface of an FCC crystal. Partly, this choice is made because it is relatively easy to visualize the saddle point configurations and partly because there is great interest in the atomic scale mechanism of island diffusion on surfaces (see for example reference 60). The interaction potential is chosen to be a simple function to make it easy for others to verify and extend our results. The atoms interact via a pairwise additive Morse potential

$$V(r) = A\left(e^{-2\alpha(r-r_0)} - 2e^{-\alpha(r-r_0)}\right) \tag{14}$$

with parameters chosen to reproduce diffusion barriers on Pt surfaces[61] ($A = 0.7102$ eV, $\alpha = 1.6047$ Å^{-1}, $r_0 = 2.8970$ Å). The potential was cut and shifted at 9.5 Å. While exchange processes are not well reproduced with such a simple potential, the predicted activation energy for hop diffusion processes is quite similar to the predictions of more complex potential functions and in some cases in quite good agreement with experimental measurements.[28,61]

The surface is simulated with a 6 layer slab, each layer containing 56 atoms. The minimum energy lattice constant for the FCC solid is used, 2.74412 Å. The bottom three layers in the slab are held fixed. A total of 7 + 168 = 175 atoms are allowed to move during the saddle point searches. This is 525 degrees of freedom. The displacements mainly involve some of the island atoms, but relaxation of the substrate atoms can also be important.

The initial configuration of the island is a compact heptamer as shown in figure 8. The question is how the island diffuses. We have focused on the initial stage of such a configurational transition, i.e. saddle points that are at the boundary of the potential basin corresponding to the compact heptamer state. A total of 13 processes were found with saddle point energy less than or equal to 1.513 eV. The lowest energy processes correspond to uniform translation of the island from FCC sites to HCP sites. There are two slightly different directions for the island to hop, and thus two slightly different saddle points, of energy 0.601 eV and 0.620 eV (see figure 8). The next three low energy saddle points, processes 3 to 5, correspond to a pair of edge atoms shifting to adjacent FCC sites. The three processes are quite similar, just three slightly inequivalent directions. Process 6 and 7 are quite interesting. Here, a pair of atoms is again shifted, but now only to the nearby HCP sites. The other 5 atoms in the cluster are also shifted to adjacent HCP sites but in the opposite direction. The final state has all island atoms sitting at HCP sites. Processes 8 and 9 involve a concerted move of three edge atoms. Process 10 and 11 involve an edge dimer where one of the atoms moves in a direction away from the island while the other one takes its place. This is a significantly higher energy final state, because of the low coordination of one of the displaced atoms. Finally, processes 12 and 13 involve the displacement of just one atom away from the island, again resulting in low coordination in the final state.

One common feature of processes 3 to 13 is that the final state is higher in energy than the initial state. The saddle point is typically late, i.e. close to the final state.

X. RESULTS

The results of the calculations are given in tables 1 and 2. The number of force evaluations needed to reach a saddle point is given. We use this unit of computational effort because the evaluation of the force dominates the effort at each step, even with empirical potentials. We are particularly interested in plane wave based DFT calculations where the evaluation of just the energy and not the force presents

insignificant savings. The computational effort is, therefore, simply characterized by the number of force evaluations. Table 1 gives the results obtained with convergence tolerance of 0.01 eV/Å in the magnitude of the force, i.e. the saddle point searches were stopped when the magnitude of the force on each degree of freedom had dropped below this value. This tolerance is small enough to get the saddle point energy to within 0.01 eV. To illustrate how fast the various methods home in on the saddle points, the number of force evaluations needed to satisfy a tighter tolerance, 0.001 eV/Å, is given in table 2 for comparison. In most cases, the saddle point energy obtained is different by less than 0.001 eV as the tolerance is reduced, but in some cases the difference is on the order of 0.01 eV.

The results show that the drag method fails for 7 out of the 13 processes. This is because the MEP has large curvature and the direction of the unstable normal mode at the saddle point is quite different from the direction of the vector **P-R**. The drag method should, therefore, not be used. When the drag method works, however, it is very efficient.

The CI-NEB method with three movable images, CI-NEB(3), is highly reliable, gets all the saddle points, and is less than three times more expensive than the drag method. Since it is easy to paralellize the CI-NEB with one image per node, the number of force evaluations per node, and therefore the elapsed time until the calculation finishes on a three node cluster, would actually be just about the same or even less for CI-NEB(3) than for Drag.

It is interesting to push the elastic band method to the extreme and reduce the number of images to one. This is essentially the same as the Drag method except the direction of the drag is different. If the tangent in the CI-NEB were estimated using equation 5, then the two methods would be identical. The fact that CI-NEB uses an estimate of the tangent, equations 7 and 9, where the weight of the adjacent points is a function of the energy, makes the CI-NEB(1) converge in these cases while the Drag method diverges. The saddle point is closer to the higher energy final state, and the tangent of the path is biased more towards the line segment to the final state than to the initial state. It is interesting that CI-NEB(1) is so successful in these test problems, but it cannot be expected to work in all cases.

The Ridge method is significantly more expensive than CI-NEB(3), a factor of 2.7 for the larger tolerance and a factor of 3.3 for the smaller tolerance. The method has relatively hard time converging rigorously on the saddle point, i.e. it uses a large number of force evaluations towards the end of the search. There are several parameters in the Ridge method that need to be chosen and the performance depends quite strongly on the choice of these parameters. We optimized for one of the saddle point searches and then used the same parameter set for all of them (the parameters are given in the discussion of the method above).

The CPR method is the most difficult method to implement, because of the complex rules for adding or rejecting points on the path. It is also the least efficient of the methods tested. It does, however, converge quickly to the saddle point once it is

close, as is evident from comparing table 1 and 2. This is probably because of the use of the conjugate gradient minimization which is quite efficient.

The DHS method of Dewar and coworkers is easy to implement and it does quite well. It is the second best method at the larger tolerance. But, as the Ridge method, it has hard time converging on the saddle point. A significant improvement in the timing might occur if a switch to a different method, for example the CI-NEB(1), is made once the two images are in the region of the saddle (for example, when the force has dropped to 0.1 eV/Å).

The Dimer method can be started from any point on the potential energy surface. While the method is designed to work without any knowledge of the final state, it is possible to make use of the final state in cases where it is known. Tables 1 and 2 are timings for the Dimer method where a line maximization along the $\mathbf{P} - \mathbf{R}$ line is first carried out, and then the Dimer search is started from the maximum. The dimer method is highly efficient, each saddle point search involves fewer force evaluations than CI-NEB(3). The advantage of CI-NEB(3) is that it gives some picture of the whole MEP in addition to the saddle point, as discussed below. The unique quality of the Dimer method is its ability to climb up the potential surface starting from the minimum. Results of 50 such runs are shown in figure 9. Here, the starting points were generated by random displacements of the atoms about the initial state minimum with maximum amplitude of 0.1 Å. The tighter tolerance, 0.001 eV/Å was used in these runs. It is surprising that the average number of force evaluations is not that much larger than when the search was started from the maximum along $\mathbf{P} - \mathbf{R}$ (590 force evaluations vs. 528). Of course, if one is only interested in a particular final state, the dimer method started from the minimum may converge on the 'wrong' saddle point and then needs to be repeated a few times.

For comparison, we have included in tables 1 and 2 the timings for a simpler algorithm, ART,[27] a method which is mainly used to help equilibrate systems by finding final states rather than saddle points (and has proven to be highly successful in simulations of amorphous materials,[62] for example). The method is analogous to the drag method except no reference is made to the final state, the drag coordinate is taken to be the direction from the initial state to the current location. The force was inverted along the drag coordinate and velocity Verlet algorithm with velocity projections used to home in on the saddle point. The method is very efficient and takes somewhat fewer iterations than the dimer method, but similar to the drag method, it does not find about half the saddle points.

XI. DISCUSSION

It is important to point out that all the timings given above are for a search of a single saddle point. In order to verify that the saddle point found is indeed the highest saddle point on the MEP for the process of interest, a calculation of the MEP needs to be carried out. Given the saddle point, it is rather straightforward to slide

down along the MEP. One stable method is to displace the system downward and then minimize the energy with a fixed distance to the previous point higher up along the path. The CI-NEB(3) method provides three points along the MEP and with the interpolation where forces are included this is typically enough to see whether the path has more than one saddle point. The CI-NEB(3) timings in table 1 and 2 are, therefore, the total number of force evaluations needed to get both the saddle point energy and to get a reasonable idea of what the MEP looks like. If it is evident that additional saddle points are present, additional images can be introduced starting from the best estimate from the interpolation. The Ridge, CPR and DHS methods would all need to be followed by a calculation of the MEP starting from the saddle point. This would typically add a couple of hundred force evaluations to the numbers given for the Drag, Ridge, CPR and DHS methods in table 1 and 2.

XII. SUMMARY

An overview has been given of several methods used to find saddle points on energy surfaces when only the energy and first derivatives with respect to atomic positions are available. Finding saddle points is the most challenging task when estimating rates of transitions within harmonic Transition State Theory. The high dimensionality of condensed matter systems makes this non-trivial. Several commonly used methods have been applied to a test problem involving configurational changes in an island on a crystal surface where the final state of the transition is known. The CI-NEB method turned out to be the most efficient method. In addition to the saddle point, it gives an idea of the shape of the whole MEP. This is necessary to determine whether more than one saddle points are present, and then which one is highest. When the final state is not known, the Dimer method can be used to climb up the potential energy surface starting from the initial state. The average number of force evaluations for a Dimer to converge on a saddle point is similar to a CI-NEB calculation with three movable images in the test problem studied here.

It is our hope that the test problem presented will continue to be a useful standard for comparing methods for finding saddle points. Clearly, other test problems with different qualities should also be added. To make it easier for others to use this test problem, we have made configurations and other supplementary information available on the web at:
http://www-theory.chem.washington.edu/~hannes/paperProgrInThChem

Acknowledgments

This work was funded by the National Science Foundation, grant CHE-9710995. We are grateful to Prof. Emily Carter for sending us a code for carrying out the Ridge method calculations and to Dr. Fischer for making his Ph.D. thesis available.

Appendix: The two-dimensional test problem

This model includes a LEPS[63] potential contribution which mimics a reaction involving three atoms confined to motion along a line. Only one bond can be formed, either between atoms A and B or between atoms B and C. The potential function has the form

$$V^{\mathrm{LEPS}}(r_{AB}, r_{BC}) = \frac{Q_{AB}}{1+a} + \frac{Q_{BC}}{1+b} + \frac{Q_{AC}}{1+c} - \left[\frac{J_{AB}^2}{(1+a)^2} + \frac{J_{BC}^2}{(1+b)^2} + \frac{J_{AC}^2}{(1+c)^2} - \frac{J_{AB}J_{BC}}{(1+a)(1+b)} - \frac{J_{BC}J_{AC}}{(1+b)(1+c)} - \frac{J_{AB}J_{AC}}{(1+a)(1+c)} \right]^{\frac{1}{2}} \quad \text{(A.1)}$$

where the Q functions represent Coulomb interactions between the electron clouds and the nuclei and the J functions represent the quantum mechanical exchange interactions. The form of these functions is

$$Q(r) = \frac{d}{2}\left(\frac{3}{2}e^{-2\alpha(r-r_0)} - e^{-\alpha(r-r_0)}\right)$$

and

$$J(r) = \frac{d}{4}\left(e^{-2\alpha(r-r_0)} - 6e^{-\alpha(r-r_0)}\right).$$

The parameters were chosen to be $a = 0.05$, $b = 0.80$, $c = 0.05$, $d_{AB} = 4.746$, $d_{BC} = 4.746$, $d_{AC} = 3.445$, and for all three pairs we use $r_0 = 0.742$ and $\alpha = 1.942$.

In order to reduce the number of variables, the location of the end point atoms A and C is fixed and only atom B is allowed to move. A 'condensed phase environment' is represented by adding a harmonic oscillator degree of freedom coupled to atom B. This can be interpreted as a fourth atom which is coupled in a harmonic way to atom B

$$V(r_{AB}, x) = V^{\mathrm{LEPS}}(r_{AB}, r_{AC} - r_{AB}) + 2k_c(r_{AB} - (r_{AC}/2 - x/c))^2 \quad \text{(A.2)}$$

where $r_{AC} = 3.742$, $k_c = 0.2025$, and $c = 1.154$. This type of model has frequently been used as a simple representation of an activated process coupled to a medium, such as a chemical reaction in a liquid or in a solid matrix.

In order to create two saddle points rather than just one, a Gaussian function is added to $V(r_{AB}, x)$ to give

$$V^{\mathrm{tot}}(r_{AB}, x) = V(r_{AB}, x) + 1.5\ G(r_{AB} - 2.02083, x + 0.272881) \quad \text{(A.3)}$$

where the Gaussian function is $G(a, b) = \exp(-0.5((a/0.1)^2 + (b/0.35)^2))$. A contour plot of this 2D potential surface is given in figures 1 to 5.

References

[1] H. Eyring, *J. Chem. Phys.* **3**, 107 (1935).

[2] E. Wigner, *Trans. Faraday Soc.* **34**, 29 (1938).

[3] J. C. Keck, *Adv. Chem.* **13**, 85 (1967).

[4] P. Pechukas, in 'Dynamics of Molecular Collisions', part B, ed. W. H. Miller (Plenum Press, N.Y. 1976).

[5] D.G. Truhlar, B.C. Garrett and S.J. Klippenstein, *J. Phys. Chem.* **100**, 12771 (1996).

[6] A. F. Voter and D. Doll, *J. Chem. Phys.* **80**, 5832 (1984).

[7] D. G. Truhlar and B. C. Garrett, *Annu. Rev. Phys. Chem.* **35**, 159 (1984).

[8] J. B. Anderson, *J. Chem. Phys.* **58**, 4684 (1973).

[9] A. F. Voter and D. Doll, *J. Chem. Phys.* **82**, 80 (1985).

[10] C. Wert and C. Zener, *Phys. Rev.* **76**, 1169 (1949).

[11] G. H. Vineyard, *J. Phys. Chem. Solids* **3** 121 (1957).

[12] R. Marcus, *J. Chem. Phys.* **45**, 4493 (1966).

[13] G. Mills, G. K. Schenter, D. Makarov and H. Jónsson *Chem. Phys. Lett.* **278**, 91 (1997).

[14] H. Jónsson, G. Mills and K. W. Jacobsen, 'RAW Quantum Transition State Theory', in 'Classical and Quantum Dynamics in Condensed Phase Simulations', ed. B. J. Berne, G. Ciccotti and D. F. Coker, page 405 (World Scientific, 1998).

[15] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals*, (McGraw Hill, New York, 1965).

[16] W. H. Miller, *J. Chem. Phys.* **62**, 1899 (1975).

[17] S. Coleman, in *The Whys of Subnuclear Physics*, ed. A. Zichichi (Plenum, N.Y., 1979).

[18] V.A. Benderskii, D.E. Makarov and C.A. Wight, *Chemical Dynamics at Low Temperature* (Whiley, New York, 1994).

[19] M.L. McKee and M. Page, *Reviews in Computational Chemistry* Vol. IV, K.B. Lipkowitz and D.B. Boyd, Eds., (VCH Publishers Inc., New York, 1993).

[20] H. Jónsson, G. Mills and K. W. Jacobsen, 'Nudged Elastic Band Method for Finding Minimum Energy Paths of Transitions', in 'Classical and Quantum Dynamics in Condensed Phase Simulations', ed. B. J. Berne, G. Ciccotti and D. F. Coker, page 385 (World Scientific, 1998).

[21] H. C. Andersen, *J. Chem. Phys.* **72**, 2384 (1980).

[22] T. A. Halgren and W. N. Lipscomb, *Chem. Phys. Lett.* **49**, 225 (1977).

[23] M. J. Rothman and L. L. Lohr, *Chem. Phys. Lett.* **70**, 405 (1980).

[24] G. Mills and H. Jónsson, *Phys. Rev. Lett.* **72**, 1124 (1994).

[25] G. Mills, H. Jónsson and G. K. Schenter, *Surf. Sci.* **324**, 305 (1995).

[26] G. Henkelman and H. Jónsson, (submitted to *J. Chem. Phys.*).

[27] N. Mousseau and G. T. Barkema, *Phys. Rev. E* **57**, 2419 (1998).

[28] M. Villarba and H. Jónsson, *Surf. Sci.* **317**, 15 (1994).

[29] M. Villarba and H. Jónsson, *Surf. Sci.* **324**, 35 (1995).

[30] E. Batista and H. Jónsson, *Computational Materials Science* (in press).

[31] M. R. Sørensen, K. W. Jacobsen and H. Jónsson, *Phys. Rev. Lett.* **77**, 5067 (1996).

[32] T. Rasmussen, K. W. Jacobsen, T. Leffers, O. B. Pedersen, S. G. Srinivasan, and H. Jónsson, *Phys. Rev. Lett.* **79**, 3676 (1997).

[33] B. Uberuaga, M. Levskovar, A. P. Smith, H. Jónsson, and M. Olmstead, 'Diffusion of Ge below the Si(100) surface: Theory and Experiment', *Phys. Rev. Lett.* **84**, 2441 (2000).

[34] L. R. Pratt, *J. Chem. Phys.* **85**, 5045 (1986).

[35] A. Kuki and P. G. Wolynes, *Science* **236**, 1647 (1986).

[36] R. Elber and M. Karplus, *Chem. Phys. Lett.* **139**, 375 (1987).

[37] R. Czerminski and R. Elber, *Int. J. Quantum Chem.* **24**, 167 (1990); R. Czerminski and R. Elber, *J. Chem. Phys.* **92**, 5580 (1990).

[38] A. Ulitsky and R. Elber, *J. Chem. Phys.* **92**, 1510 (1990).

[39] C. Choi and R. Elber, *J. Chem. Phys.* **94**, 751 (1991).

[40] E. M. Sevick, A. T. Bell and D. N. Theodorou, *J. Chem. Phys.* **98**, 3196 (1993).

[41] T. L. Beck, J. D. Doll and D. L. Freeman, *J. Chem. Phys.* **90**, 3183 (1989).

[42] R. E. Gillilan and K. R. Wilson, *J. Chem. Phys.* **97**, 1757 (1992).

[43] O. S. Smart, *Chem. Phys. Lett.* **222**, 503 (1994).

[44] G. Henkelman, B. Uberuaga and H. Jónsson, (submitted to *J. Chem. Phys.*).

[45] S. Fischer and M. Karplus, *Chem. Phys. Lett.* **194**, 252 (1992).

[46] Stefan Fischer, *"Curvilinear reaction-coordinates of conformal change in macromolecules: application to rotamase catalysis"*, Ph. D. Thesis, Harvard University, (1992).

[47] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, in *Numerical Recipies* (Cambridge University Press, New York, 1986).

[48] I. V. Ionova and E. A. Carter, *J. Chem. Phys.* **98**, 6377 (1993).

[49] M. J. S. Dewar, E. F. Healy, and J. J. P. Stewart, *J. Chem. Soc., Faraday Trans. 2* **80**, 227 (1984).

[50] C. J. Cerjan and W. H. Miller, *J. Chem. Phys.* **75**, 2800 (1981).

[51] D. T. Nguyen and D. A. Case, *J. Phys. Chem.* **89**, 4020 (1985).

[52] W. Quapp, *Chem. Phys. Lett.* **253**, 286 (1996).

[53] H. Taylor and J. Simons *J. Phys. Chem.* **89**, 684 (1985).

[54] J. Baker, *J. Comput. Chem.* **7**, 385 (1986).

[55] D. J. Wales, *J. Chem. Phys.* **91**, 7002 (1989).

[56] N. P. Kopsias and D. N. Theodorou, *J. Chem. Phys.* **109**, 8573 (1998).

[57] J, D, Honeycutt and H. C. Andersen, *Chem. Phys. Lett.* **108**, 535 (1984); *J. Chem. Phys.* **90**, 1585 (1986).

[58] G. Henkelman and H. Jónsson, *J. Chem. Phys.* **111**, 7010 (1999).

Table I *Number of force evaluations needed to reach saddle point to* 0.01 eV/Å *tolerance in the force.*

saddle	Drag	CI-NEB(3)	CI-NEB(1)	Ridge	CPR	DHS	Dimer	ART
1	47	81	25	189	241	232	80	83
2	37	75	25	288	240	230	76	70
3	-	285	177	1369	1277	788	439	246
4	-	276	179	1129	1464	785	94	236
5	-	333	151	1165	1443	736	354	250
6	-	654	204	1369	2412	2434	449	-
7	-	735	206	1245	2426	2057	430	-
8	146	300	163	772	776	526	262	380
9	149	351	179	781	748	483	281	386
10	-	363	115	734	1551	736	510	-
11	-	282	126	869	2612	706	214	-
12	156	294	48	884	718	521	186	-
13	153	333	105	913	686	478	304	-
Average	115	336	131	901	1276	824	283	236
Std	56	184	64	368	810	662	149	125

[59] A. F. Voter, *Phys. Rev. Lett.* **78**, 3908 (1997).

[60] G. Mills, T. R. Mattsson, I. Mollnitz, and H. Metiu, *J. Chem. Phys.* **111**, 8639 (1999).

[61] D.W. Bassett and P.R. Webber, *Surf. Sci.* **70**, 520 (1978).

[62] G. T. Barkema and N. Mousseau, *Phys. Rev. Lett.* **77**, 4358 (1996).

[63] Polanyi and Wong, *J. Chem. Phys.* **51**, 1439 (1969).

Table II *Number of force evaluations needed to reach saddle point to* 0.001 eV/Å *tolerance in the force.*

saddle	Drag	CI-NEB(3)	CI-NEB(1)	Ridge	CPR	DHS	Dimer	ART
1	324	372	122	3441	653	795	328	332
2	70	192	45	288	433	290	244	146
3	-	597	327	2382	1610	1295	746	336
4	-	585	246	2047	1729	1296	546	366
5	-	675	314	2112	1695	1258	570	377
6	-	999	274	2187	2821	4310	704	-
7	-	978	271	2144	2720	4076	588	-
8	323	573	309	4090	1197	1320	559	742
9	338	855	446	1995	1268	1342	553	754
10	-	648	174	1610	1739	1468	816	-
11	-	447	237	1859	2793	1474	308	-
12	299	687	150	1861	1038	1160	386	-
13	293	738	230	1901	969	1097	562	-
Average	275	642	242	2147	1590	1629	532	436
Std	102	228	103	890	788	1182	173	227

Figure 1 *The 'drag' method. A drag coordinate is defined by interpolating from* **R** *to* **P** *with a straight line (dashed line). Starting from* **R**, *the drag coordinate is increased stepwise and held fixed while relaxing all other degrees of freedom in the system. In a two-dimensional system, the relaxation is along a line perpendicular to the* $\mathbf{P} - \mathbf{R}$ *vector. The solid lines show the first and last relaxation line in the drag calculation. The final location of the system after relaxation is shown with filled circles. As the drag coordinate is increased, the system climbs up the potential surface close to the slowest ascent path, reaching a potential larger than the saddle point, and then, eventually, slipping over to the product well. In this simple test case, the drag method cannot locate the saddle point.*

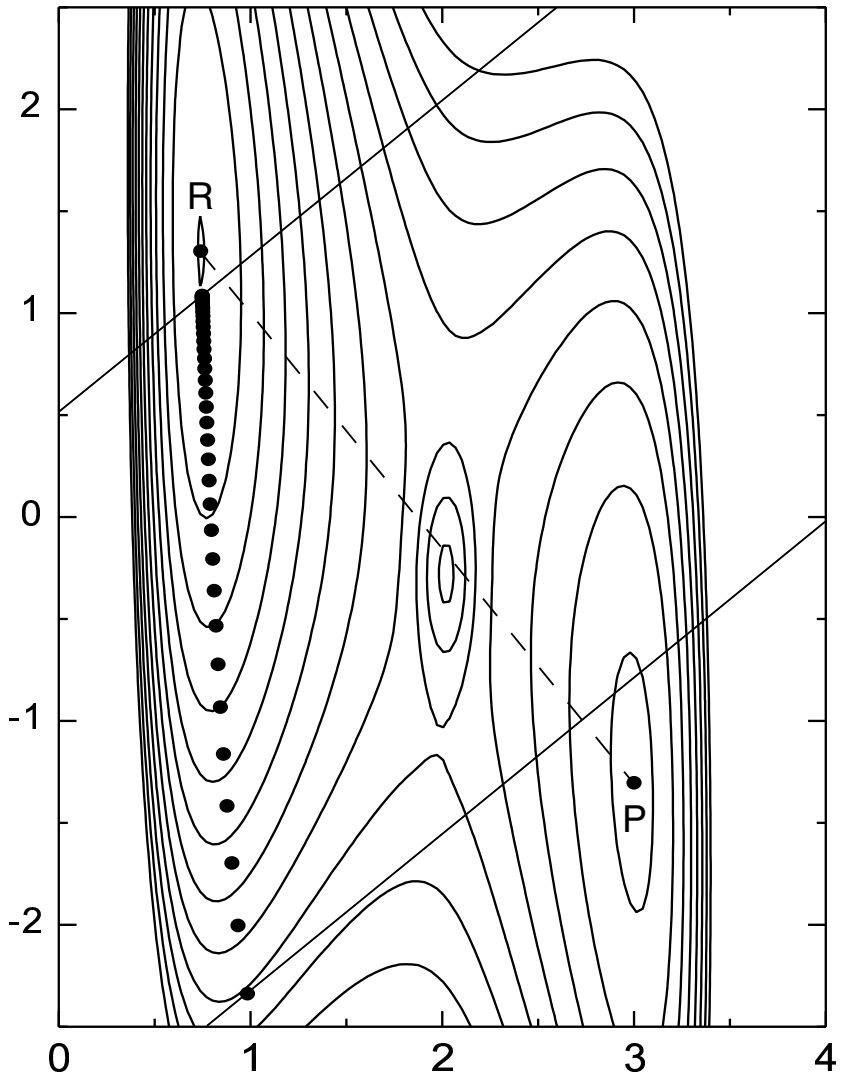

Figure 2 *The Climbing Image Nudged Elastic Band method, CI-NEB. An elastic band is formed with three movable images of the system connected by springs and placed between the fixed endpoints,* **R** *and* **P**. *The calculation is started by placing the three images along a straight line interpolation. The images are then relaxed keeping only the the component of the spring force parallel to the path and the component of the true force perpendicular to the path. The image with the highest energy is also forced to move uphill along the parallel component of the true force to the saddle point.*

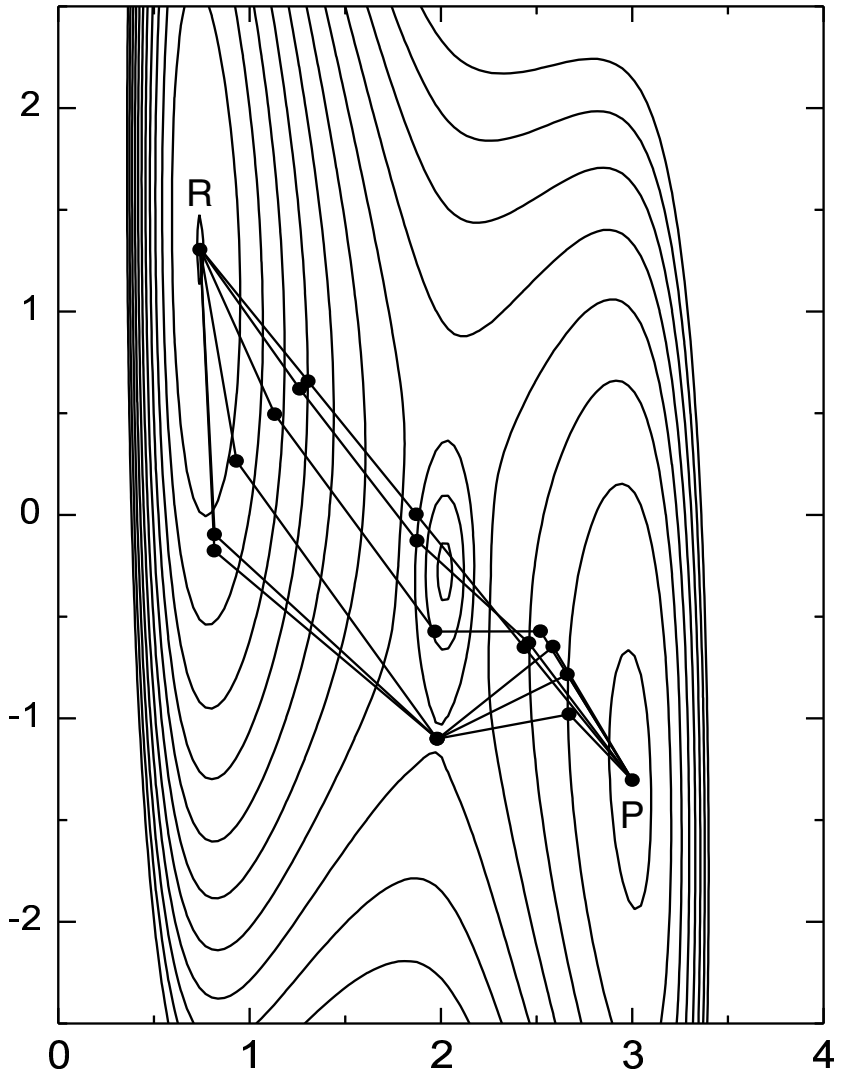

Figure 3 *The conjugate peak refinement (CPR) method. Points along a path connecting* **R** *and* **P** *are generated, one point at a time through a cycle of maximization and then minimization. First, the maximum along the vector* $\mathbf{P} - \mathbf{R}$ *is found,* $\mathbf{y}_1$*. Then, a minimization is carried out along a conjugate vector (small dashed line) to give location* $\mathbf{x}_1$ *on the path. In the second cycle (shown in inset) the maximum along an estimated tangent to the* $\mathbf{R} - \mathbf{x}_1 - \mathbf{P}$ *path (solid line in inset) is found,* $\mathbf{y}_2$*, and then energy is minimized along a conjugate vector (small dashed line in inset) to give a fourth point along the path, etc.*

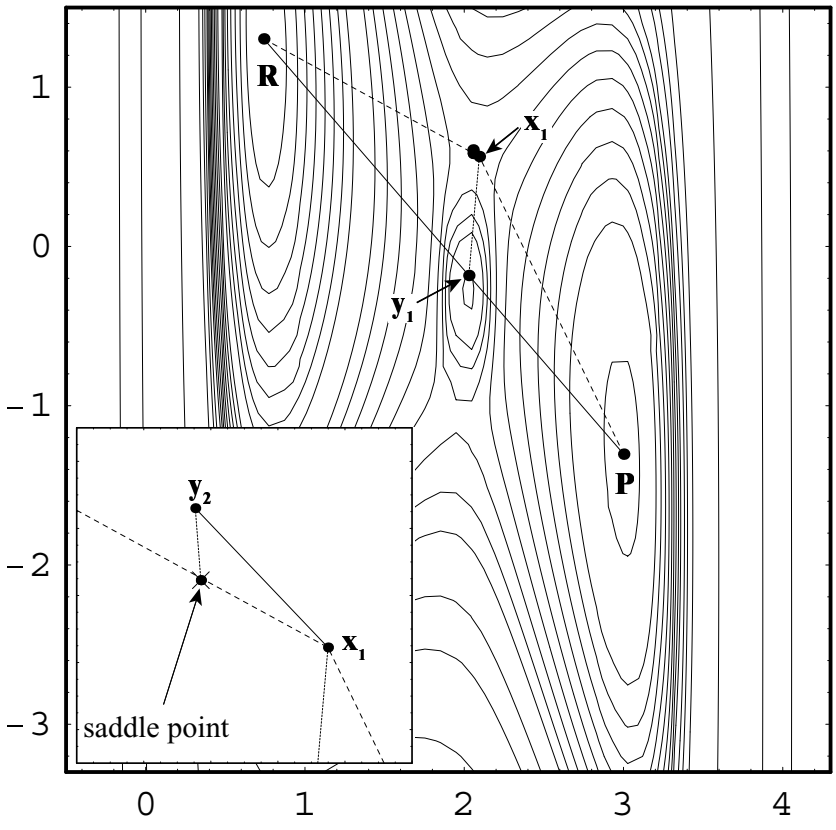

Figure 4 *The Ridge method. A pair of images on each side of the potential energy ridge is moved towards the saddle point. First, the maximum along the vector* $\mathbf{P} - \mathbf{R}$ *is found, point* **a** *in the inset. Then the two images are formed on each side of the maximum, points* $\mathbf{x}_0'$ *and* $\mathbf{x}_1'$*, and are displaced downhill along the gradient to points* $\mathbf{x}_0''$ *and* $\mathbf{x}_1''$*. This cycle of maximization between the two images, and the downhill move of the two images along the gradient is repeated, with smaller and smaller displacements until the saddle point is reached.*

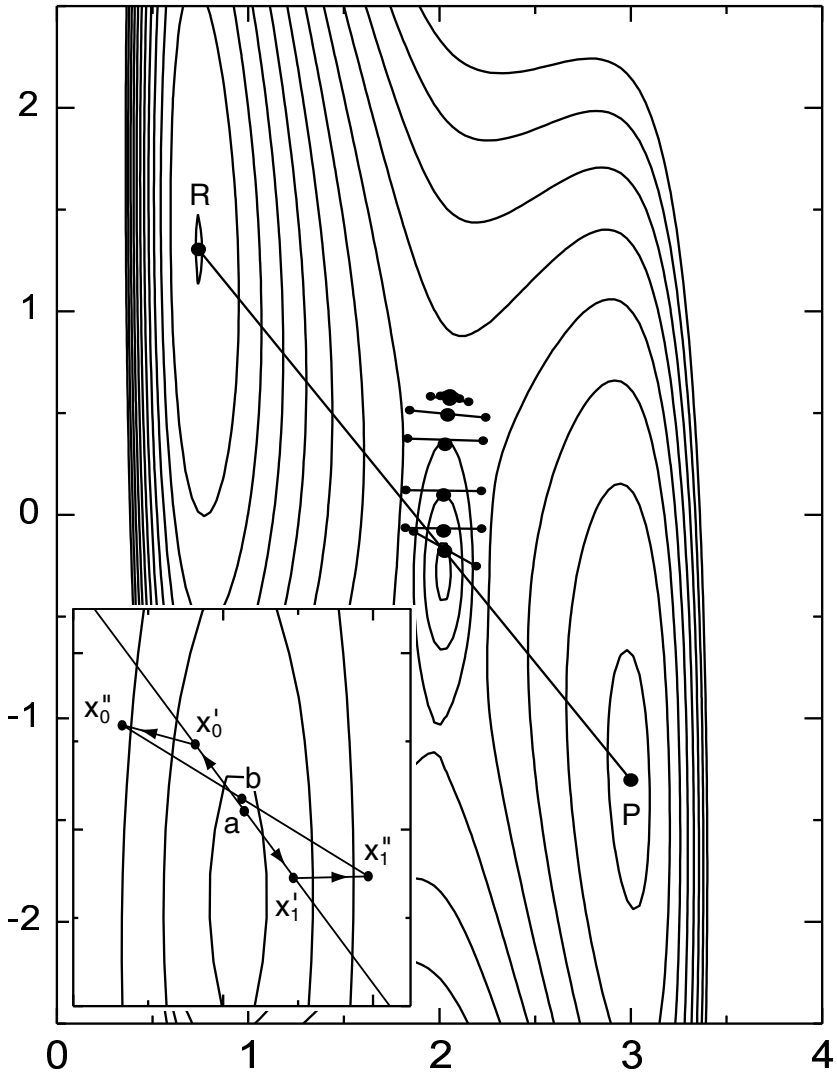

Figure 5 *The method of Dewar, Healy and Stewart (DHS). Initially, a pair of images is created at* **R** *and* **P**. *In each cycle, the lower energy image is pulled towards the higher energy one and then allowed to relax keeping the distance between the two fixed. Eventually, the two images straddle the energy ridge near the saddle point.*

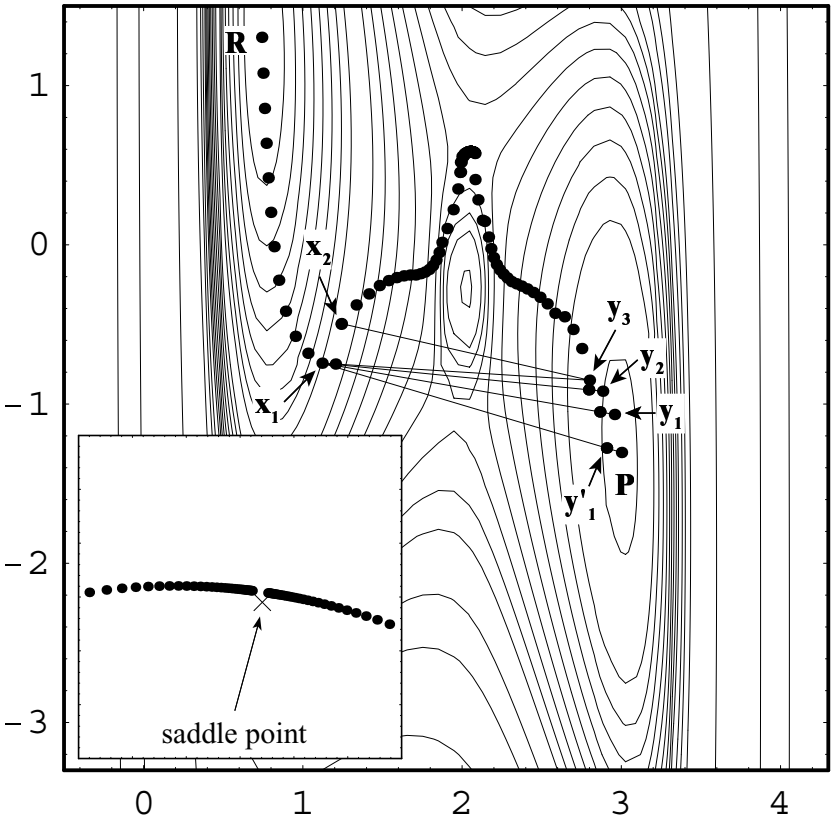

Figure 6 *The calculation of the effective force in the Dimer method. A pair of images, spaced apart by a small distance, on the order of* 0.1 Å*, is rotated to minimize the energy. This gives the direction of the lowest frequency normal mode. The component of the force in the direction of the dimer is then inverted and the minimization of this effective force leads to convergence to a saddle point. No reference is made to the final state.*

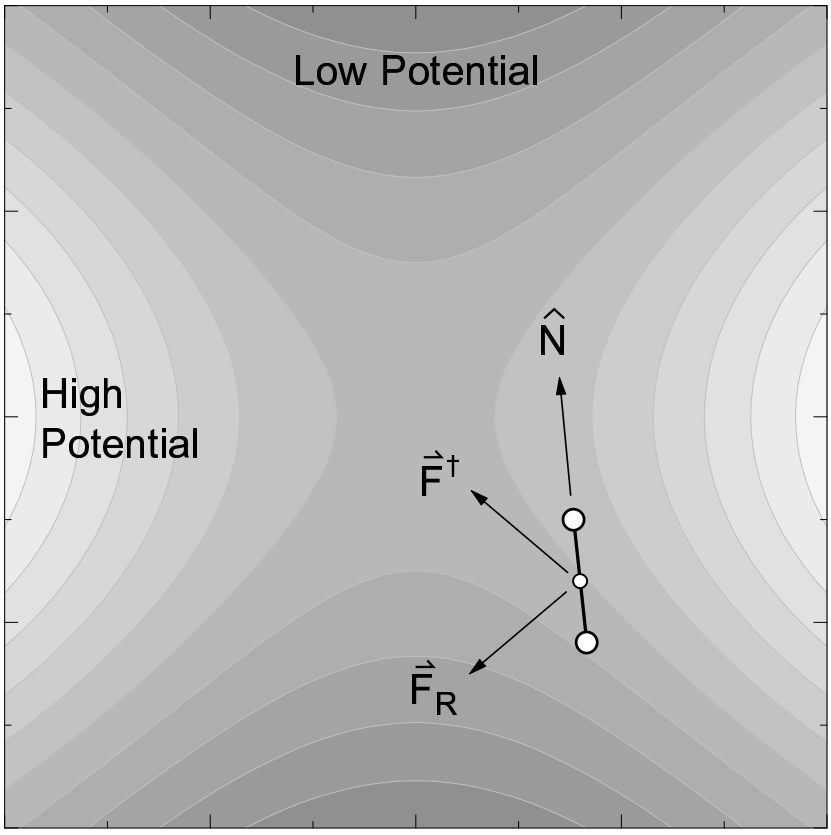

Figure 7 *Application of the dimer method to a two-dimensional test problem. Three different starting points are generated in the reactant region by taking extrema along a high temperature dynamical trajectory. From each one of these, the dimer is first translated only in the direction of the lowest mode, but once the dimer is out of the convex region a full optimization of the effective force is carried out at each step (thus the kink in two of the paths). Each one of the three starting points leads to a different saddle point in this case.*

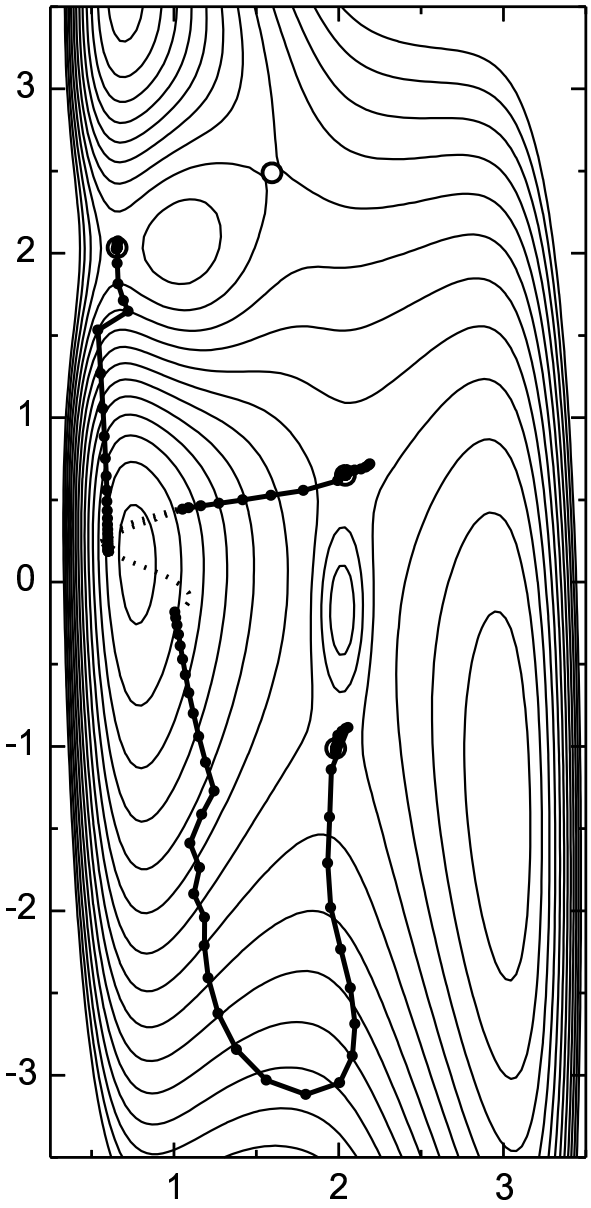

Figure 8 *On-top view of the surface and the seven atom island used to test the various saddle point search methods. The shading indicates the height of the atoms. The initial state is shown on top. The saddle point configuration and the final state of the 13 transitions are also shown, with the energy of the saddle point (in* eV*) indicated to the left. The first two transitions correspond to a uniform translation of the intact island. Transitions 3-5 correspond to a pair of atoms sliding to adjacent FCC sites. In transitions 6 and 7 the pair of atoms slides to the adjacent HCP sites and the remaining 5 atoms slide in the opposite direction to HCP sites. In transitions 8 and 9, a row of three edge atoms slides into adjacent FCC sites. In transitions 10 and 11 a pair of edge atoms moves in such a way that one of the atoms is displaced away from the island while the other atom takes its place. In transitions 12 and 13 a single atom gets displaced.*

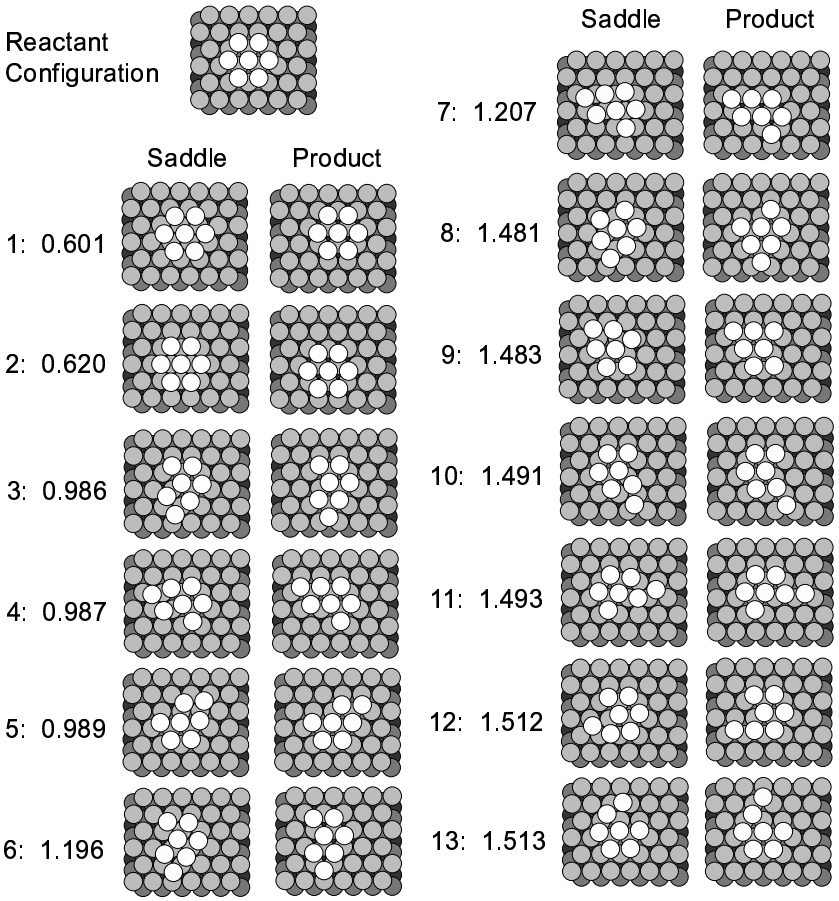

Figure 9 *The frequency at which the various saddle points for the surface island transitions (illustrated in figure 8) are found with the Dimer method. The lowest saddle points are found with the highest frequency. Also shown are the number of iterations required to go from the intial state to the saddle point to within a force tolerance of* 0.001 eV/Å*. For the more practical* 0.01 eV/Å *tolerance, the average number of force evaluations was a little under 300. The error bars show the standard deviation.*